Cuprins

Prefaţă

Aspecte primare ale statisticii se pierd în negura timpului. Astăzi, putem spune, fără a greşi, că statistică facem fiecare dintre noi, cu ştiinţă sau fără. Gândul la ziua de mâine, cum ne organizăm şi cum folosim timpul şi mijloacele materiale de care dispunem, are la bază date (statistice) reţinute (culese) şi prelucrate în mod simplu sau mai complex. Ne dăm seama, prin urmare, că statistica are o largă răspândire, de la utilizarea ei în mod empiric, până la apelarea la metodele fundamentate matematic.

Statistică făceau chinezii antici, care dispuneau de date relative la populaţie, pământuri şi recolte, dar şi egiptenii, care efectuau cadastrări (operaţii de determinare a unor proprietăţi agricole şi imobiliare cu toate caracteristicile lor) şi numărări de populaţie. Numărarea populaţiei era cunoscută şi la evrei, de exemplu, în texte biblice sunt menţionate numărări de populaţie, ca cea efectuată de Moise, privind bărbaţii buni de oaste. La grecii antici erau, de asemenea, efectuate numărări de populaţie, ale bunurilor necesare pentru scopuri militare, pentru aşezarea impozitelor şi evaluarea bogăţiilor. Toate acestea au culminat cu recensămintele şi anchetele romane .

Culegerea de date privind resursele umane şi materiale aveau la bază o gândire practică, folosirea acestora în scopuri fiscale, militare sau administrative. Putem spune că toate acestea erau utilizate în descrierea *statului*.

Tratarea ştiinţifică a datelor relative la descrierea statului se întâlneşte în Germania (sec. XVII–XVIII), iar denumirea de *statistică* apare în cursul lui Martin Schmeitzel (1679–1757), intitulat ”Collegium politico-*statisticum*” (Universitatea din Halle). Unii istorici atribuie întâietate privind această denumire lui Gottfried Achenwall (1719-1772), care a introdus învăţământul ”*Staatskunde*” la Universitatea din Göttingen.

În Anglia, în aceeaşi perioadă, în afara universităţilor, exista ca disciplină descriptivă a statului ceea ce se numea *aritmetică politică*. Aritmetica politică se ocupa în special de cercetarea fenomenelor demografice.

Un moment important în dezvoltarea statisticii îl reprezintă cristalizarea calculului probabilităţilor.

Calculul probabilităţilor şi fundamentarea riguroasă din punct de vedere matema-

tic a teoriei probabilităţilor reprezintă o problemă fundamentală în stabilirea locului şi rolului acestei discipline matematice în evantaiul larg al matematicii, precum şi al fundamentării teoretice a statisticii matematice.

Desigur, faptul că cineva, cu diferite ocazii, foloseşte cuvântul *probabil*, nu înseamnă că este o persoană iniţiată în calculul probabilităţilor. Am putea, eventual, accepta că sunt percepute anumite aspecte empirice ale calculului probabilităţilor, când, relativ la un anumit fenomen care prezintă un anumit grad de nedeterminare, se emit afirmaţii de forma: *este puţin probabil ca...*, *este foarte probabil ca...*, *este improbabil ca...*. De la astfel de afirmaţii şi până la cunoaşterea şi înţelegerea în profunzime a teoriei probabilităţilor este un drum lung, care este de un larg interes atât din punct practic cât şi teoretic, şi care prezintă o atracţie deosebită, nu numai pentru matematicieni, dar şi pentru specialişti din alte discipline ale cunoaşterii umane: fizică, chimie, biologie, economie, medicină, inginerie etc. Aceste aspecte, de-a lungul timpului, se regăsesc în apariţia, dezvoltarea şi fundamentarea axiomatică a teoriei probabilităţilor, care se îmbină şi se continuă cu diferite domenii de aplicabilitate, dintre care statistica ocupă un rol privilegiat.

Este unanim acceptat că noţiunea de probabilitate îşi are originea în jocurile de noroc. În antichitate, la egipteni, greci şi romani erau cunoscute jocuri de noroc cu zaruri, care s-au perpetuat până în zilele de astăzi. Se pare că cel mai vechi zar datează din jurul anilor 3500 î.e.n. şi avea feţele numerotate la întâmplare de la 1 la 6, faţă de zarul actual, care are feţele numerotate astfel ca suma numerelor de pe feţele opuse să fie 7. Proliferarea jocurilor de noroc s-a produs nu numai prin aria de răspîndire, dar şi prin diversitatea acestora. Ajunge să amintim doar jocurile de ruletă care sunt instalate în marile metropole ale lumii.

Problemele teoretice relative la jocurile de noroc, în special privind şansele partenerilor, au atras atenţia savanţilor timpului. Astfel, fundamentarea elementară a calculului probabilităţilor poate fi atribuită francezilor P. de Fermat (1601–1665) şi B. Pascal (1623–1662). Corespondenţa din anul 1654, purtată de cei doi, conţine primele aspecte privind calculul şanselor în jocurile de noroc. O parte din problemele considerate au fost formulate şi prezentate lui Pascal de către Cavalerul de Méré, un renumit amator al jocurilor de noroc din acea vreme.

Prima carte de introducere în calculul probabilităţilor este *De Ratiociniis in Aleae Ludo* (*Cum se raţionează în jocurile cu zarul*), apărută în anul 1657 şi are ca autor pe olandezul Ch. Huygens (1629–1695).

Totuşi, se remarcă faptul că italianului G. Cardano (1501–1576) i se datorează lucrarea intitulată *Liber de Ludo Aleae* (*Cartea despre jocurile cu zarul*), care a fost cunoscută şi publicată la mai bine de o sută de ani de la dispariţia sa.

Un destin asemănător a avut şi lucrarea lui James Bernoulli (1654–1705), *Ars Conjectandi* (*Arta de a face presupuneri*), care a fost publicată de către fratele său

John Bernoulli în anul 1713. Se remarcă în conţinutul cărţii, ceea ce se numeşte azi legea numerelor mari, unul din rezultatele strălucitoare ale teoriei probabilităţilor.

Un alt rezultat de aceeaşi dimensiune, teorema limită centrală, se găseşte în cartea *Doctrine of Chances*, apărută în anul 1718 şi care este opera lui A. de Moivre (1667–1754), ceea ce, se cunoaşte, este strâns legată de legea normală de probabilitate.

Descoperirea legii normale de probabilitate este atribuită, de cele mai multe ori, lui K. F. Gauss (1777–1855), şi care este considerată ca fiind specifică erorilor de măsurare. Trebuie remarcat faptul că în anul 1808, cu un an înainte de descoperirea lui Gauss, topograful american Robert Adrain (1775–1843) publicase o lucrare în care a sugerat utilizarea legii normale pentru descrierea erorilor de măsurare.

Rezultatele lui Moivre, privind teorema limită centrală, sunt extinse de către P. - S. de Laplace (1770–1820), stabilind de asemenea legătura cu legea normală de probabilitate, în cartea *Théorie Analytique des Probabilités* (*Teoria analitică a probabilităţilor*), apărută în anul 1812.

Un statistician belgian de seamă al sec. XIX, unanim recunoscut pentru rezultatele de pionerat în domeniul statisticii matematice este A. Quetelet (1796–1874). De numele lui se leagă noţiuni ale statisticii ca: repartiţie, medie, dispersie, observare de masă, regularitate. Pentru el, statistica reprezenta singura metodă ce se poate aplica fenomenelor de masă.

Un salt remarcabil în dezvoltarea teoria probabilităţilor a fost făcut prin contribuţia matematicienilor ruşi L. P. Cebîşev (1821–1884), A. A. Markov (1856–1922) şi A. M. Leapunov (1857–1918). Primul a introdus noţiunea de variabilă aleatoare, reformulând şi generalizând legea numerelor mari în limbajul variabilelor aleatoare. Folosind noţiunea de variabilă aleatoare, A. A. Markov şi A. M. Leapunov au obţinut apoi alte rezultate privind legea numerelor mari şi teorema limită centrală.

Pasul decisiv privind fundamentarea modernă a teoriei probabilităţilor este făcut prin lucrarea matematicianului rus A. N. Kolmogorov (1903–1989), apărută în anul 1933. A. N. Kolmogorov, folosind teoria măsurii, reuşeşte să construiască modelul axiomatic al teoriei probabilităţilor.

Se remarcă de asemenea preocupări şi rezultate importante obţinute de alţi matematicieni ca: S. D. Poisson (1781–1840), É. Borel (1871–1956), S. N. Bernstein (1880–1968), R. von Mises (1883–1953), A. I. Hincin (1894–1959).

Sfârşitul sec. XIX şi începutul sec. XX se consideră a fi începutul statisticii moderne. Este momentul când se trece de la etapa descriptivă, la interpretarea analitică a fenomenelor de masă şi obţinerea de concluzii inductive pe baza observaţiilor empirice. Statisticieni de renume, care au pus bazele statisticii matematice sunt consideraţi a fi englezii F. Galton (1822–1911) şi K. Pearson (1857–1936). Opera acestora este continuată de celebrul statistician R. A. Fisher (1890–1962). Dintre reprezentanţii şcolii engleze de statistică amintim pe G. U. Yule (1871–1951), W. S. Gosset (1876–

1937), C. E. Spearman (1883–1945), M. Kendall (1907–1983).

Pentru detalii privind istoricul statisticii recomandăm lucrarea [38].

Astăzi, statistica matematică, având un suport teoretic al teoriei probabilităţilor, are o dezvoltare şi aplicabilitate deosebită. Care este motivul aplicării statisticii matematice pe o scară largă şi în atâtea domenii ale cunoaşterii umane? Am putea da două răspunsuri principale. În primul rând, fiecare cercetător sau persoană care analizează fenomene ale lumii reale, modelează astfel de fenomene, caută să pună la baza metodelor sale de cercetare ştiinţifică, de evaluare a fenomenelor reale, un instrument de investigare obiectiv. Şi este cvasiunanim recunoscut că matematica are astfel de instrumente de investigare, inclusiv prin metodele statisticii matematice. În al doilea rând, metodele statisticii matematice sunt uşor de aplicat. Greutatea constă numai în fundamentarea riguroasă din punct de vedere matematic a acestora. Acest aspect este ascuns celui care aplică statistica matematică. În materialul pe care îl prezentăm, considerăm noţiunile (conceptele) de bază în iniţierea studiului şi aplicării statisticii matematice. Se consideră partea "ascunsă" a statisticii matematice, adică fundamentarea teoretică, cât de cât riguroasă, a statisticii matematice, dar şi partea care prezintă modul de aplicare a metodelor statisticii matematice.

Dacă în tot acest demers avem la dispoziţie şi mijlocul modern de lucru, reprezentat astăzi printr–un calculator performant sau mai puţin performant, atunci statistica devine foarte atractivă. De aceea au fost elaborate o serie de produse soft, care prezintă într–o concepţie unitară reprezentarea, prelucrarea şi studiul datelor statistice. Amintim aici câteva astfel de produse soft: *Statgraphics*, *Statistica*, *S–Plus*, *SAS*, *StatXact*, *SPSS*, *Stata*, *GraphPad*, *ViSta* etc. Remarcăm de asemenea că unele produse soft elaborate în alte scopuri decât cel al prelucrării datelor statistice, conţin proceduri privind prelucrarea mai elementară sau mai complexă a datelor statistice. În această categorie se încadrează şi produsul *Matlab*, care va fi invocat în prezenta lucrare, dar tot în această categorie putem aminti şi *Mathematica*, *Maple* etc.

Capitolul 1

Introducere în Matlab

MATLAB (MATrix LABoratory) este un sistem interactiv de nivel şi performanţă înalte pentru efectuarea de calcule ştiinţifice şi tehnice, precum şi de analiză şi vizualizare a datelor şi care dispune de un limbaj propriu de programare.

Prima versiune a fost realizată în anii 70 de către *Cleve Moler*, iar versiunea utilizată în cele ce urmează este *Matlab 6.0*.

Se pare că acest produs are o largă răspândire din mai multe motive:

- interactivitatea sistemului;
- elementele de bază sunt matricele, după cum spune şi denumirea, ale căror tipuri şi dimensiuni nu se declară;
- grafica puternică şi foarte uşor de utilizat;
- posibilitatea integrării unor proceduri proprii în limbaje de programare cum ar fi *Fortran, C*;
- arhitectura modulară, care permite adăugarea unor proceduri şi funcţii proprii, ceea ce a permis deja elaborarea unor pachete de proceduri şi funcţii, numite *toolbox*–uri, orientate pe anumite domenii, cum ar fi
  - Statistica – *Statistics Toolbox*;
  - Funcţii spline – *Spline Toolbox*;
  - Analiză wavelet – *Wavelet Toolbox*;
  - Procesarea semnalelor – *Signal Processing Toolbox*;
  - Procesarea imaginilor – *Image Processing Toolbox*;
  - Modelare, simulare şi analiza sistemelor dinamice – *Simulink*;

– Calcul simbolic – *Symbolic/Extended Symbolic Math Toolbox*;
– Ecuaţii cu derivate parţiale – *PDE Toolbox*;
– Optimizare – *Optimization Toolbox*.

1.1 Gestionarea unei sesiuni Matlab

1.1.1 Lansarea şi închiderea unei sesiuni Matlab

Trebuie să remarcăm de la început că sistemul Matlab dispune de o puternică bază de *help*–uri, dar care poate fi activată doar după ce sistemul Matlab a fost lansat.

Lansarea sistemului Matlab se efectuează prin comenzi ce sunt specifice sistemului de operare pe care este instalat. Vom considera în continuare sistemul de operare Windows, caz în care lansarea se face prin activarea

- *icon*-ului *Matlab* sau
- *aplicaţiei Matlab* din directorul în care a fost instalat sistemul Matlab.

În urma efectuării acestei comenzi se deschide o fereastră în care apare prompterul specific sistemului Matlab

```
>>
```

Sistemul Matlab devine în acest fel interactiv, adică la fiecare comandă sau funcţie tastată şi acceptată de Matlab (editarea comenzii sau funcţiei încheindu–se prin `Enter`), sistemul o execută şi afişează pe ecran rezultatul, dacă este cazul.

Există posibilitatea de trecere de la modul interactiv la cel programat, prin tastarea numelui unui fişier (cu extensia `.m`), care conţine un program în sistemul Matlab (succesiune de instrucţiuni Matlab) şi care mod se încheie odată cu terminarea executării programului respectiv, după care se revenie la modul interactiv.

Încheierea unei sesiuni Matlab se poate face, fie prin comenzi specifice sistemului de operare Windows, fie prin tastarea uneia din comenzile

```
>>quit
>>exit
```

1.1.2 Comenzi `help`

Comanda `help` permite utilizatorilor să aibă acces la o mare cantitate de informaţie (în limba engleză) relativă la întreg sistemul Matlab.

Există diferite variante pentru lansarea comenzii `help`, anume prin:

- tastarea comenzii `help` urmată de domeniul (`topic`) în care suntem interesaţi, adică

```
>>help topic
```

Remarcăm faptul că parametrul `topic` poate lipsi. Astfel, lansarea comenzii

```
>>help
```

are ca efect afişarea pe ecran a unei liste ce cuprinde domeniile care operează în versiunea curentă Matlab

```
HELP topics:

matlab\general   -  General purpose commands.
matlab\ops       -  Operators and special characters.
matlab\lang      -  Programming language constructs.
matlab\elmat     -  Elementary matrices and matrix manipulation.
matlab\elfun     -  Elementary math functions.
matlab\specfun   -  Specialized math functions.
matlab\matfun    -  Matrix functions - numerical linear algebra.
matlab\datafun   -  Data analysis and Fourier transforms.
matlab\audio     -  Audio support.
matlab\polyfun   -  Interpolation and polynomials.
matlab\funfun    -  Function functions and ODE solvers.
matlab\sparfun   -  Sparse matrices.
matlab\graph2d   -  Two dimensional graphs.
matlab\graph3d   -  Three dimensional graphs.
matlab\specgraph -  Specialized graphs.
matlab\graphics  -  Handle Graphics.
matlab\uitools   -  Graphical user interface tools.
matlab\strfun    -  Character strings.
matlab\iofun     -  File input/output.
matlab\timefun   -  Time and dates.
matlab\datatypes -  Data types and structures.
matlab\verctrl   -  Version control.
matlab\winfun    -  Windows Operating System Interface Files
                    (DDE/ActiveX)
matlab\demos     -  Examples and demonstrations.
toolbox\local    -  Preferences.
toolbox\stats    -  Statistics Toolbox.

For more help on directory/topic, type "help topic".
```

- tastarea comenzii `lookfor` urmată de un cuvânt cheie (`key`), în care suntem interesaţi, adică

```
>>lookfor key
```

va lista prima linie de comentariu din fiecare fişier cu extensia `.m`, care conţine cuvântul cheie `key`, precedată de numele fişierului.

- tastarea comenzii `helpwin` urmată de un domeniu (`topic`) în care suntem interesaţi, adică

```
>>helpwin topic
```

produce deschiderea unei ferestre cu informaţii relative la domeniul precizat.

- tastarea comenzii `helpdesk`, adică

```
>>helpdesk
```

produce deschiderea unei ferestre prin care se permite căutarea în baza `help`–ului Matlab.

1.1.3 Comenzi sistem

Sistemul Matlab operează cu o mulţime de comenzi pe care le întâlnim şi în alte sisteme. Enumerăm o parte dintre acestea:

- `mkdir` – deschiderea unui subdirector nou în directorul curent;
- `cd` – schimarea directorului curent;
- `dir` – lista conţinutului directorului curent;
- `version` – afişarea versiunii sistemului Matlab instalat;
- `!com` – ieşire temporară din sistemului Matlab instalat şi execuţia comenzii Windows precizată prin `com`;
- `demo` – lansarea execuţiei de programe demonstrative.

Înainte de începerea unei sesiuni este de dorit să se creeze un director prin comanda `mkdir`, pentru salvarea fişierelor, iar apoi trecerea în acest director prin comanda `cd`.

1.2 Constante. Variabile. Expresii aritmetice

Având în vedere că sistemul Matlab operează numai cu un singur tip de obiect, matrice numerică, care ar putea avea elemente numere complexe, rezultă că orice scalar este considerat ca o matrice cu o linie şi o coloană. În plus, reprezentarea internă a elementelor matricelor şi operaţiile cu acestea se fac în dublă precizie. Deoarece reprezentarea internă este prestabilită, avem doar posibilitatea de precizare a tipului (formatului) de afişare (ieşire) respectiv de intrare a datelor.

1.2.1 Constante

Introducerea constantelor se poate face:

- de la tastatură;
- prin generarea cu ajutorul unor instrucţiuni Matlab;
- dintr–un fişier creat cu ajutorul unui editor.

La intrare constantele numerice pot fi de tipul:

- *întreg*: `2002, -2002;`
- *real*: `3.14159, 0.314e1, -.314e+01, 31.4e-1, 31.4e-01;`
- *complex*: `a`+`bi`, unde `a` şi `b` sunt două constante întregi sau reale, iar `i` reprezintă unitatea imaginară (`i` $= \sqrt{-1}$).

1.2.2 Instrucţiunea `format`

Sistemul Matlab dispune de câteva şabloane (formatări) standard pentru afişarea (extragera) datelor numerice. Instrucţiunea `format` prin care se precizează formatul standard are sintaxa

```
>>format tip
```

unde `tip` poate lua una din valorile `short`, `long`, `short e`, `long e`, `short g`, `long g`, `hex`, `+`, `bank`, `rat`, `loose`, `compact`.

Formatele `short` şi `long` afişează pe ecran datele numerice în virgulă fixă, respectiv cu 5 cifre zecimale şi 15 cifre zecimale (precedate de semnul minus, dacă sunt negative).

Formatele `short e` şi `long e` afişează pe ecran datele numerice în virgulă flotantă, respectiv cu 5 cifre zecimale şi 15 cifre zecimale (precedate de semnul minus, dacă sunt negative).

Formatele `short g` şi `long g` afişează pe ecran datele numerice cu cea mai potrivită reprezentare, virgulă fixă sau virgulă flotantă, respectiv cu 5 cifre zecimale şi 15 cifre zecimale (precedate de semnul minus, dacă sunt negative).

Formatele `rat`, `bank`, `hex` şi `+` afişează pe ecran respectiv datele numerice sub formă de fracţie ordinară, sub formă bancară (cu două zecimale), sub formă hexazecimală, iar în ultimul caz, prin `+` dacă numărul este pozitiv, respectiv `-`, dacă numărul este negativ şi `spaţiu` dacă este zero.

Formatele `loose` şi `compact` permite afişare datelor la două rânduri (format implicit), respectiv la un rând în cazul al doilea.

Pentru a afla la un moment dat, care format este în funcţie se poate apela la comanda

```
>>get(0,'Format')
```

1.2.3 Constante speciale

Sistemul Matlab are câteva constante speciale, care au denumiri specifice (identificatori specifici):

- `inf` – rezultatul împărţirii la 0;
- `NaN` (Not a Number) – rezultatul operaţiei de tipul $0/0$;
- `realmax` – cel mai mare număr real pozitiv reprezentat pe dublu cuvânt (`1.7977e+308`);
- `realmin` – cel mai mic număr real pozitiv reprezentat pe dublu cuvânt (`2.2251e-308`);
- `pi` – π=`3.14159...` (valoarea reprezentată pe dublu cuvânt);
- `eps` – epsilonul maşinii (`eps`=`2.2204e-16`=2^{-54});
- `date` – data curentă;
- `clock` – timpul curent;
- `calendar` – calendarul lunii curente;
- `i, j` – unitatea imaginară (`i,j`=$\sqrt{-1}$);
- `zeros(n)` şi `zeros(m,n)` – generează matricea pătratică de ordin `n`, respectiv matricea cu `m` linii şi `n` coloane, având toate elementele nule;
- `ones(n)` şi `ones(m,n)` – generează matricea pătratică de ordin `n`, respectiv matricea cu `m` linii şi `n` coloane, având toate elementele 1;
- `eye(n)` şi `eye(m,n)` – generează matricea pătratică de ordin `n`, respectiv matricea cu `m` linii şi `n` coloane, având pe diagonala principală 1 şi 0 în rest (când matricea este pătratică avem matricea unitate);
- `magic(n)` – generează matrice magică pătratică de ordin `n`.

Pentru introducerea elementelor unei matrice, linie cu linie, elemente ce pot fi exprimate şi prin expresii aritmetice, acestea sunt incluse între paranteze drepte. Elementele fiecărei linii sunt separate prin spaţiu sau virgulă, iar liniile sunt separate prin punct–virgulă `(;)` sau prin trecerea la rândul următor.

Astfel matricea

$$\begin{pmatrix} 3 & 5 & 2 \\ 9 & 1 & 4 \end{pmatrix}$$

se poate introduce de la tastatură prin

```
>>[3 5 2;9 1 4]
```

sau

```
>>[3,5,2;9,1,4]
```

Dacă elementele matricei prezintă regularitatea, că elementele unei linii sunt date printr–o progresie aritmetică, atunci avem o introducere prescurtată. De exemplu, introducerea matricei

$$\begin{pmatrix} 3 & 5 & 7 & 9 \\ 9 & 6 & 3 & 0 \end{pmatrix}$$

se poate face de la tastatură prin

```
>>[3:2:9;9:-3:0]
```

1.2.4 Variabile

Variabilele sunt matrice ce sunt utilizate într–un program Matlab sau în mod interactiv şi ale căror valori pot fi modificate. Numele lor fiind precizate prin *identificatori*, care sunt daţi de orice combinaţie de litere, cifre zecimale şi liniuţa de subliniere, din care primul caracter este literă: Matrice1, mat, Cluj_Napoca.

Apelarea şi vizualizarea unui element al unei variabile (matrice) se face prin specificarea numelui variabilei urmat de lista indicilor corespunzători, separaţi prin virgulă, închisă între paranteze. De exemplu, `A(i,j )` pune în evidenţă elementul din linia `i` şi coloana `j` al matricei `A`.

Facem observaţia că matricele în sistemul Matlab nu acceptă decât indici strict pozitivi.

Utilizând operatorul două puncte (:) se pot obţine submatrice ale unei matrice. De exemplu, `A(1:3,4:5)` extrage submatricea formată din primele trei linii şi coloanele 4 şi 5 ale matricei `A`, iar `A(:,4)` şi `A(1,:)` extrag respectiv coloana a patra a matricei `A` şi prima linie a matricei `A`. O formă mai generală de obţinere a unei submatrice se realizează prin apelul de forma `A([l],[c])`, unde `l` şi `c` reprezintă liste de indici de linie şi indici de coloane, separaţi prin virgulă sau spaţiu. Astfel va fi obţinută submatricea formată din liniile şi coloanele precizate prin `l`, respectiv `c` ale matricei `A`. Dacă se introduce comanda `A(:)`, atunci matricea `A` este transformată într–un vector coloană, ce conţine coloanele matricei.

Alte combinaţii ale acestor tipuri de specificare ale elementelor unei matrice pot fi considerate.

Mai remarcăm aici comanda

```
>>reshape(A,m,n)
```

care transformă matricea `A` într–o matrice cu `m` linii şi `n` coloane. Dacă `A` nu are `m`$\times$`n` elemente, atunci apare eroare la executare.

1.2.5 Operatori aritmetici

Expresiile aritmetice se construiesc după cunoscutele reguli ale limbajelor de programare evoluate, folosind operatorii aritmetici cunoscuţi, desigur cu unele operaţii speciale specifice sistemului Matlab.

Ordinea execuţiilor operaţiilor aritmetice este de asemenea cea cunoscută şi pe care o reamintim aici, începând cu cel mai înalt nivel :

- prima dată sunt executate operaţiile aritmetice dintre paranteze;
- urmează ridicarea la putere, iar dacă sunt mai multe operaţii consecutive de acest tip, atunci ordinea execuţiei este de la dreapta la stânga;
- următorul nivel de prioritate include operaţiile de înmulţire şi împărţire, iar dacă sunt mai multe operaţii consecutive de acest tip, atunci ordinea execuţiei este de la stânga la dreapta;
- ultimul nivel de prioritate conţine operaţiile de adunare şi scădere, iar dacă sunt mai multe operaţii consecutive de acest tip, atunci ordinea execuţiei este de la stânga la dreapta.

Operatorii aritmetici matriceali de care dispune sistemul de bază Matlab sunt:

- `A+B, A-B` sau `plus(A,B), minus(A,B)` – adună, respectiv face diferenţa matricelor `A` şi `B` de aceleaşi dimensiuni, iar dacă unul dintre operanzi e un scalar, acesta este extins la matricea constantă dată prin scalarul respectiv, având dimensiunile celeilalte matrice, după care se efectuează operaţia;
- `A*B` sau `mtimes(A,B)` – înmulţeşte matricele `A` şi `B`, pentru care numărul coloanelor matricei `A` este acelaşi cu numărul liniilor matricei `B`, iar dacă una din matrice e un scalar, acesta va înmulţi toate elementele celeilalte matrice;
- `A^B` sau `mpower(A,B)` – dacă `B` este un scalar `p` întreg pozitiv, atunci se obţine puterea `p` a matricei pătratice `A`, iar dacă `A` este un scalar α, atunci se ridică scalarul α la puterea `B`, matricea `B` fiind pătratică, adică, pentru exemplificare, dacă $\alpha = e$, atunci $e^{B} = \sum_{n=0}^{\infty} \frac{B^n}{n!}$;
- `A.*B`, `A.^B`, `A.\B`, `A./B`, sau formele echivalente, respectiv `times(A,B)`, `power(A,B)`, `ldivide(A,B)`, `rdivide(A,B)` – efectuează operaţiile de înmulţire, de ridicare la putere, împărţire inversă, împărţire directă, toate de tipul element cu element (ca la operaţiile de adunare şi scădere), pentru matricele de aceleaşi dimensiuni, iar dacă una dintre matrice e un scalar, aceasta este extinsă la matricea constantă dată prin scalarul respectiv, având dimensiunile celeilalte matrice, după care este efectuată operaţia;

- `A\B` sau `mldivide(A,B)` – are ca rezultat matricea ce reprezintă soluţia ecuaţiei matriceale `A*X=B`, când matricea pătratică `A` este inversabilă, iar dacă `B` este un vector coloană se obţine soluţia sistemului liniar corespunzător;

- `A/B` sau `mrdivide(A,B)` – are ca rezultat matricea ce reprezintă soluţia ecuaţiei matriceale `X*B=A`, când matricea pătratică `B` este inversabilă;

- `[A B], [A;B]` (*concatenarea*) – produce concatenarea matricelor `A` şi `B` pe orizontală, respectiv pe verticală, cu condiţia ca la concatenarea pe orizontală trebuie ca numărul liniilor celor două matrice să fie acelaşi, iar la concatenarea pe verticală se impune ca numărul coloanelor celor două matrice să fie acelaşi.

1.2.6 Funcţii predefinite

Sistemul Matlab dispune de numeroase funcţii matematice. Remarcăm faptul că toate acestea, având în vedere că matricele sunt obiectele de bază în Matlab, acţionează asupra matricelor.

Încercăm o clasificare a acestora:

- `abs` (modul), `sqrt` (radical), `sign` (semn), `conj` (conjugat), `angle` (argument al numărului complex), `imag` (parte imaginară), `real` (parte reală);

- `sin` (sinus), `cos` (cosinus), `tan` (tangentă), `cot` (cotangentă), `csc` (cosecantă), `sec` (secantă);

- `asin` (arcsinus), `acos` (arccosinus), `atan` (arctangentă), `acot` (arccotangentă), `acsc` (arccosecantă), `asec` (arcsecantă), `sinh` (sinus hiperbolic), `cosh` (cosinus hiperbolic), `tanh` (tangentă hiperbolică), `coth` (cotangentă hiperbolică), `csch` (cosecantă hiperbolică), `sech` (secantă hiperbolică), `asinh` (arcsinus hiperbolic), `acosh` (arccosinus hiperbolic), `atanh` (arctangentă hiperbolică), `acoth` (arccotangentă hiperbolică), `acsch` (arccosecantă hiperbolică), `asech` (arcsecantă hiperbolică);

- `exp` (exponenţială), `log` (logaritm natural), `log10` (logaritm zecimal), `log2` (logaritm în baza doi);

- `round` (rotunjire la cel mai apropiat întreg), `floor` (rotunjire spre $-\infty$), `fix` (rotunjire spre zero), `ceil` (rotunjire spre $+\infty$).

Facem observaţia că pentru funcţiile trigonometrice se impune ca argumentele să fie date în radiani.

Dacă argumentul funcţiilor predefinite este matrice, atunci funcţia se aplică fiecărui element al matricei.

- `atan2(A,B)` (are acelaşi efect cu `atan(A.\B)`), `rem(A,B)` (restul împărţirii lui `A` la `B`).

Dacă `A` şi `B` sunt matrice, atunci trebuie să fie de aceeaşi dimensiune, iar funcţia se aplică element cu element.

- `size(A)` (dimensiunea matricei `A`), `length(A)` (lungimea vectorului `A`), `ndims(A)` (numărul dimensiunilor matricei `A`), `disp(A)` (afişează elementele matricei `A` fără numele matricei), `det(A)` (determinantul matricei pătratice `A`), `inv(A)` (inversa matricei pătratice `A`), `rank(A)` (rangul matricei `A`), `A'` sau `transpose(A)` (transpusa matricei `A`, dacă matricea este cu elemente numere reale, dacă are elemente numere complexe, atunci se efectuează matricea transpusă a conjugatelor elementelor), `A.'` (transpusa matricei `A`), `diag(A)` (diagonala matricei pătratice `A`, iar dacă `A` este vector, generează matricea pătratică diagonală cu elementele diagonalei date de vector), `trace(A)` (urma matricei pătratice `A`, adică $\sum_i a_{ii}$).

Trebuie să remarcăm că dacă `A` este un vector linie, atunci A' este un vector coloană şi invers, dacă `A` este un vector coloană, atunci A' este un vector linie.

- `factorial(n)` (factorialul lui `n`), `perms(1:n)` sau `perms(v)` (generează toate pemutările primelor `n` numere naturale, respectiv generează toate permutările componentelor vectorului `v`), `nchoosek(n,k)` sau varianta `nchoosek(v,k)` (calculează numărul combinărilor de `n` elemente luate câte `k`, respectiv generează toate combinările elementelor vectorului `v` luate câte `k`), `combnk(v,k)` (generează toate combinările elementelor vectorului `v` luate câte `k`), `factor(n)` (factorii primi ai lui `n` repetaţi de atâtea ori cât reprezintă puterea acestora), `primes(n)` (numerele prime mai mici decât `n`), `gcd` (cmmdc), `lcm` (cmmmc).

1.2.7 Comenzi pentru gestiunea variabilelor şi a spaţiului de lucru

Sistemul Matlab se poate lansa din orice director, acesta devenind directorul curent. Desigur este de dorit ca utilizatorul să lucreze într–un director propriu. Aşa cum am văzut, un astfel de director poate fi creat până nu se intră în sistemul Matlab, ca apoi acesta să fie lansat din acest director sau se poate lansa la început sistemul Matlab, iar apoi prin comenzile `cd` şi `mkdir` să se ajungă într–un director al utilizatorului sau să se creeze un director propriu nou.

Toate operaţiile sunt executate în directorul curent.

Dacă se doresc informaţii relative la variabilele existente la un moment dat în directorul curent se pot lansa una din comenzile:

- `who` – produce lista variabilelor curente;
- `whos` – produce lista variabilelor curente, împreună cu informaţii privind aceste variabile (dimensiune, număr de elemente, memorie ocupată, tipul);
- `clear` – şterge variabilele şi funcţiile temporare din directorul curent (ştergerea unei singure variabile se realizează prin comanda `clear` urmată de numele variabilei);
- `diary file` – salvează toate informaţiile apărute pe ecran în timpul derulării unei sesiuni în fişierul `file`, cu excepţia graficelor;
- `save` – salvează variabilele de lucru pe disc;
- `load` – încarcă variabilele de lucru de pe disc.

Comenzile

```
>>save file
>>save file x y z
```

vor salva în fşierul `file.mat` toate variabilele din sesiune curentă, respectiv numai variabilele `x`, `y`, `z`.

Diferenţa dintre comanda `diary` şi comanda `save` este că fişierul obţinut prin `diary` se poate modifica (edita) cu ajutorul unui editor, pe când celalalt nu se poate edita.

De asemenea, remarcăm faptul că în fişierul creat prin comanda `diary` conţine toate informaţiile până la întâlnirea comenzii

```
>>diary off
```

Dacă înainte de închiderea unei sesiuni Matlab nu se execută o comandă `save`, atunci toate variabilele împreună cu conţinutul lor se pierd.

Într–o altă sesiune Matlab, se pot recupera variabilele din sesiunea precedentă, dacă au fost salvate prin comanda `save`, prin lansarea comenzii `load`. Anume, dacă fişierul salvat este `file.mat` atunci comanda

```
>>load file
```

va încărca varaibilele cu conţinutul lor pentru sesiunea curentă.

Comanda `load` poate fi utilizată şi pentru încărcarea unui fişier ASCII creat cu alt editor. Astfel comanda

```
>>load A.dat -ascii
```

va avea ca efect pentru sesiunea curentă a generării unei variabile cu numele `A` şi care conţine ca valori datele din fişierul `A.dat`.

1.3 Instrucţiuni de atribuire

Sistemul Matlab dispune de trei tipuri de instrucţiuni de atribuire:

```
expresie
variabila = expresie
[lista de variabile] = func
```

Expresia se compune, după reguli cunoscute în programare, cu ajutorul operaţiilor aritmetice şi al funcţiilor, prin operarea asupra constantelor şi asupra variabilelor.

Dacă se consideră prima formă a instrucţiunii de atribuire, atunci rezultatul este păstrat într–o variabilă de lucru numită `ans` şi care rămâne nemodificată pâna la execuţia unei alte instrucţiuni de atribuire de această formă. Forma a doua face această atribuire, variabilei din partea stângă a semnului de egalitate.

Forma a treia este specifică pentru cazul în care se apelează o *procedură* (*funcţie* Matlab), care returnează mai mult de o valoare.

Există expresii specifice sistemului Matlab, care nu au corespondent în alte limbaje evoluate. Dacă vrem să atribuim unei variabile valori care sunt obţinute ca şi elemente ale unei progresii aritmetice, atunci putem folosi una din instrucţiunile de atribuire

```
>>vi:pas:vf
>>x = vi:pas:vf
```

unde `vi`, `pas`, `vf` sunt expresii aritmetice, care pot fi evaluate la momentul executării instrucţiunii. În urma executării acestor instrucţiuni variabila `ans` respectiv `x` va conţine valorile numerice începând cu `vi` pâna la `vf` cu pasul `pas`.

Dacă valoarea lui `vf` este mai mică decât cea a lui `vi` şi `pas` are valaore negativă, atunci rezultatul instrucţiunii de atribuire este *matricea vidă*, având sintaxa [].

Dacă `pas=1`, atunci se pot considera formele echivalente

```
>>vi:vf
>>x = vi:vf
```

Un alt tip specific de instrucţiune de atribuire a sistemului Matlab este acela când variabila primeşte valoarea unei submatrice. De exemplu, instrucţiunile de atribuire

```
>>x = A(:,2)
>>y = A(end,1:10)
```

au ca efect obţinerea în `x` a vectorului coloană format din coloana a doua a matricei `A`, iar `y` va fi un vector linie format din primele 10 elemente ale ultimei linii a matricei `A`.

Orice instrucţiune poate fi continuată pe linia următoare prin trei puncte (...).

O linie poate să conţină mai multe instrucţiuni separate prin virgulă sau prin punct-virgulă (;).

Dacă o instrucţiune este urmată de punct-virgulă (;), atunci valoarea variabilei nu este afişată după executarea instrucţiunii, în caz contrar valoarea curentă a variabilei va fi afişată.

1.4 Instrucţiuni de citire şi scriere

1.4.1 Instrucţiunea `input`

Sintaxa comenzii `input` este:

```
>>v = input('s')
```

Prin executarea acestei instrucţiuni se afişează pe ecran şirul de caractere dat prin `s`, aşteptându–se tastarea datelor ce se doresc a fi atribuite pentru variabila `v`. Când se încheie operaţiunea de tastarea datelor, se tastează `Enter`, drept consecinţă se efectuează operaţia de atribuire.

1.4.2 Instrucţiunea `ginput`

Instrucţiunea `ginput` permite introducerea de date cu ajutorul *mouse*-ului dintr-o fereastră ce are trasate axele de coordonate.

Există următoarele forme ale instrucţiunii `ginput`:

```
>>[x,y] = ginput(n)
>>[x,y] = ginput
>>[x,y,buton] = ginput(n)
```

Executarea unei astfel de instrucţiuni produce introducerea coordonatelor unor puncte din figura curentă, `n` când se folosesc prima şi ultima variantă, respectiv un număr nedefinit, când se foloseşte varianta a doua. La început apare pe ecran în fereastra figurii curente cursorul realizat prin două drepte verticale. Prin poziţionări succesive, cu ajutorul *mouse*-ului, ale cursorului pe anumite puncte, urmate de *click*-uri, coordonatele acestora sunt introduse respectiv în vectorii `x` şi `y`. Pentru prima şi ultima formă trebuie efectuate astfel de operaţii succesive în număr de `n`, iar pentru a doua succesiunea de operaţii se încheie prin tastarea comenzii `Enter`. Forma a treia va returna şi vectorul `buton`, care specifică ce buton al *mouse*–lui a fost folosit la fiecare operaţie (1, 2 sau 3, începând de la stânga).

1.4.3 Instrucţiunea `fprintf`

Afişarea pe ecran şi nu numai, folosind un format propriu, se face cu ajutorul instrucţiunii `fprintf`. Vom prezenta în continuare numai modul de afişare pe ecran folosind această instrucţiune.

Cea mai simplă sintaxă a acestei instrucţiuni este

```
>>fprintf('f',x)
```

prin care valoarea variabilei `x` este afişată pe ecran conform formatului `f`.

Formatul `f` poate fi, de exemplu: `%md`, `%md\n`, `%m.zf`, `%m.zf\n`, `%m.ze`, `%m.ze\n`, unde `md`, `m.zf` şi `m.ze` exprimă respectiv faptul că valoarea lui `x` se afişează pe `m` poziţii ca un întreg, ca un real cu `z` cifre zecimale, ca un real normalizat

cu z cifre zecimale. Parametrul \n specifică faptul că după afişare se trece la rândul următor (ţine locul comenzii Enter).

1.5 Operatori relaţionali şi operatori logici

1.5.1 Operatori relaţionali

- A==B sau eq(A,B), A~=B sau ne(A,B), A<B sau lt(A,B), A>B sau gt(A,B), A<B sau lt(A,B), A>B sau gt(A,B), A<=B sau le(A,B), A>=B sau ge(A,B) – compară element cu element matricele de aceleaşi dimensiuni A şi B şi produce o matrice ce are valorile 1 şi 0, când condiţia este adevărată respectiv falsă, iar dacă unul din operanzi este un scalar, acesta este extins la matricea constantă dată prin scalarul respectiv, având dimensiunile celeilalte matrice, după care se efectuează operaţia relaţională.

1.5.2 Operatori logici

- A&B sau and(A,B) (*conjuncţia*) – compară element cu element matricele de aceleaşi dimensiuni A şi B şi produce o matrice ce are valorile 1 şi 0, când ambele elemente sunt diferite de zero, respectiv cel puţin unul din cele două elemente este zero;

- A|B sau or(A,B) (*disjuncţia*) – compară element cu element matricele de aceleaşi dimensiuni A şi B şi produce o matrice ce are valorile 1 şi 0, când cel puţin unul din cele două elemente este zero, respectiv ambele elemente sunt zero;

- xor(A,B) (*disjuncţia exclusivă*) – compară element cu element matricele de aceleaşi dimensiuni A şi B şi produce o matrice ce are valorile 1 şi 0, când unul şi numai unul din cele două elemente este zero, respectiv ambele elemente sunt zero sau ambele sunt diferite de zero;

Dacă unul din operanzi este un scalar, acesta este extins la matricea constantă dată prin scalarul respectiv, având dimensiunile celeilalte matrice, după care se efectuează operaţia logică.

- ~A sau not(A) (*negaţia*) – produce o matrice de aceleaşi dimensiuni cu cele ale matricei A având valorile 1 şi 0, când elementul corespunzător al matricei A este zero, respectiv când este diferit de zero.

- any(A) – dacă A este vector, returnează valoarea 1, dacă există componente ale vectorului diferite de zero şi 0 în caz contrar, iar dacă A este o matrice, operează

asupra fiecărei coloane a matricei şi returnează corespunzător un vector de `1` şi `0` cu lungimea egală cu cea a numărului coloanelor matrcei `A`;

- `all(A)` – dacă `A` este vector, returneză valoarea `1`, dacă toate componentele vectorului sunt diferite de zero şi `0` în caz contrar, iar dacă `A` este o matrice, operează asupra fiecărei coloane a matricei şi returnează corespunzător un vector de `1` şi `0` cu lungimea egală cu cea a numărului coloanelor matrcei `A`.

1.6 Fişiere script (m–file)

Pentru a executa o secvenţă de instrucţiuni Matlab (program), fără intervenţia utilizatorului, se editează un fişier ASCII cu extensia `.m`, numit (*fişier script*) sau *m–file*, care să conţină instrucţiunile programului. Lansarea execuţiei programului se face prin tastarea numelui (fără extensia `.m`), după care se tastează `Enter`. Astfel, dacă utilizatorul şi–a creat un fişier script cu numele `exemplu.m`, atunci linia de comandă prin care se lansează execuţia este

```
>>exemplu
```

Fşierul `exemplu.m` ar putea avea următorul conţinut

```
m = input('m (nr. natural > 0):');
A = magic(m);
d = ndims(A)
[l c] = size(A)
B = A*ones(m,1);
C = B'
```

În urma lansării execuţiei acestui program, fără a putea interveni până la încheierea execuţiei, se produc următoarele operaţii:

- execută instrucţiunea din prima linie, prin afişarea pe ecran a mesajului

  ```
  m (nr. natural > 0):
  ```

  şi se aşteapă tastarea unui număr întreg pozitiv, după care prin comanda `Enter` se trece la instrucţiunea din linia următoare;

- instrucţiunea din linia a doua va genera matricea magică pătratică `A` de ordin `m`, care nu va fi afişată pe ecran deoarece instrucţiunea se termină cu simbolul punct–virgulă;

- se execută funcţia `ndim`, iar rezultatul este atribuit variabilei `d`, care va avea valoarea `2` şi va fi afişat pe ecran, deoarece instrucţiunea nu se termină cu punct–virgulă, pe două rânduri

  ```
  d =
     2
  ```

- următoarea instrucţiune atribuie variabilelor `l` şi `c` numărul liniilor şi coloanelor matricei `A`, care în cazul de faţă coincid cu valoarea lui `m`, iar valorile lui `l` şi `c` vor fi afişate pe ecran;

- `B` va fi un vector coloană cu `m` componente şi conţine produsul dintre matricea `A` şi vectorul coloană de `m` componente, toate având valoarea `1`, iar dacă avem în vedere că `A` este o matrice magică, toate componentele lui `B` vor fi egale;

- ultima instrucţiune ajută la afişarea pe ecran a componentelor vectorului linie `C`, care este vectorul transpus al vectorului coloană `B`.

Editarea unui fişier script se poate face cu orice editor de text, dar sistemul Matlab dispune de un editor propriu.

Editorul Matlab se poate lansa din fereastra de lucru Matlab prin comenzi specifice sistemului Windows, alegând din meniul `File` opţiunea `New` sau `Open`, dacă se doreşte crearea uni nou fişier, respectiv se doreşte editare unui fişier deja existent. Acelaşi rezultat se obţine prin comanda

```
edit file.m
```

în care `file.m` este opţional. Dacă lipseşte acest argument atunci urmează crearea (editarea) unui fişier nou, iar dacă acest argument este prezent, fişierul cu acest nume urmează să fie reactualizat (editat).

Editorul sistemului Matlab poate fi lansat din orice altă parte prin activarea unui fişier cu extensia `.m`.

Editorul Matlab are capacitatea de a efectua unele verificări sintactice, de exemplu corespondenţa paranteză deschisă – paranteză închisă, dar este prevăzut şi cu un set de comenzi (*debuger*) pentru depanarea programului prin evaluarea şi modificarea unor variabile şi expresii, crearea şi ştergerea unor puncte de întrerupere etc.

Comenzile specifice sistemului de depanare Matlab sunt:

- `dbstop` – crearea unui punct de întrerupere (*breakpoint*);

- `dbclear` – ştergerea unui punct de întrerupere;

- `dbstatus` – listează punctele de întrerupere;

- `dbstack` – listează stiva apelurilor;

- `dbcont` – reluarea execuţiei programului;

- `dbstep` – execuţia uneia sau mai multe instrucţiuni;

- `dbtype` – generează lista fişierelor cu extensia `.m`, împreună cu numărul liniilor acestora;

- `dbup` – schimbă spaţiul de lucru curent, cu cel al unui fişier cu extensia `.m`, care a fost apelat;
- `dbdown` – efectuează operaţia inversă comenzii `dbup`;
- `dbmex` – execuă depanarea unui fişier cu extensia `.mex` (creat cu unul din limbajele Fortran şi C);
- `dbquit` – încheierea depanării.

1.6.1 Comenzi pentru gestionarea fişierelor

- `type file` – listează conţinutul fişierului `file` din directorul curent;
- `delete file` – şterge fişierului `file` din directorul curent;
- `what` – listează fişierele cu extensiile `.m` din directorul curent;
- `which numef` – listează calea directorului în care se află funcţia `numef`;

Comanda

```
>>what director
```

extrage lista fişierelor directorului specificat `...\toolbox\matlab\director` Lista acestor directoare poat fi vizualizată prin comanda `help` fără nici un parametru.

1.7 Grafică bidimensională

Sistemul Matlab dispune de o largă gamă de instrucţiuni sau proceduri (funcţii) pentru reprezentarea grafică a datelor, care permit o apelare simplă şi eficientă.

1.7.1 Instrucţiunea `plot`

Formele acceptate de sistemul Matlab pentru instrucţiunea `plot`, privind reprezentarea grafică în plan sunt

```
plot(y)
plot(x,y)
plot(x1,y1,x2,y2,...)
plot(y,s)
plot(x,y,s)
plot(y1,s1,y2,s2,...)
plot(x1,y1,s1,x2,y2,s2,...)
```

În cazul primei forme, dacă `y` este un vector de lungime `m`, se reprezintă grafic linia poligonală ce trece prin punctele de coordonate (i, y_i), $i = \overline{1, \texttt{m}}$.

Programul 1.7.1. Să considerăm programul, care reprezintă grafic linia poligonală ce uneşte punctele de coordonate (i, y_i), $i = \overline{1, \mathtt{m}}$, unde `y` reprezintă elementele diagonalei unei matrice magice de ordin `m`:

```
m = input('m:');
y = diag(magic(m));
plot(y)
```

În urma execuţiei programului, pentru `m=10`, se obţine graficul din Figura 1.1.

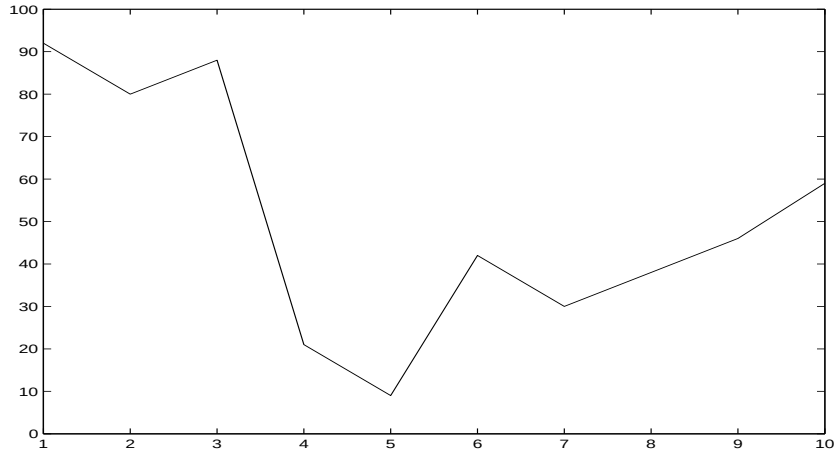

Figura 1.1: Linie poligonală

Dacă `y` este o matrice de tipul `(m,n)`, atunci se reprezintă grafic `n` linii poligonale, corespunzând celor `n` coloane ale matricei `y`, anume $(i, y_{ij})_{i=1}^{\mathtt{m}}$, $j = \overline{1, \mathtt{n}}$.

Programul 1.7.2. Să considerăm programul, care reprezintă grafic liniile poligonale ce unesc respectiv punctele de coordonate $(i, y_{ij})_{i=1}^{m}$, pentru fiecare $j = \overline{1, \mathtt{n}}$, unde `y` este o matrice magică de ordin `m`, iar `n` este un număr întreg pozitiv cel mult `m`:

```
m = input('m:');
n = input('n (n<=m):');
y = magic(m);
plot(y(:,1:n))
```

În urma execuţiei programului, pentru `m=10` şi `n=2`, se obţine graficul din Figura 1.2.

Dacă `y` este de tip complex, atunci prima formă este echivalentă cu a doua formă a instrucţiunii `plot`, adică cu

```
plot(real(y),imag(y))
```

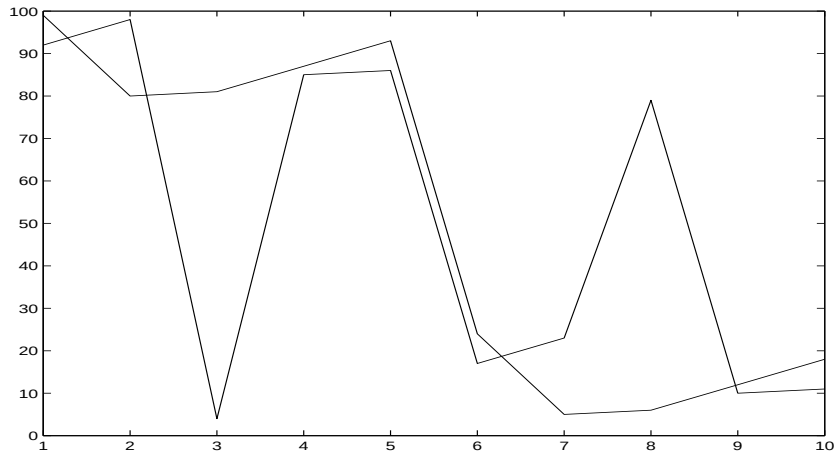

Figura 1.2: Linii poligonale

Programul 1.7.3. Să considerăm programul, care reprezintă grafic funcţia cu valori complexe $z = x\cos(2\pi x) + ix^2\sin(2\pi x), \quad x \in [0,1]$.

```
m = input('m:');
h =1/m; x = 0:h:1;
z = x.*cos(2*pi*x)+i*x.^2.*sin(2*pi*x);
plot(z)
```

În urma execuţiei programului, pentru `m=100`, se obţine graficul din Figura 1.3.

Remarcăm faptul că pentru toate celelalte forme, dacă avem parametri complecşi, partea imaginară este ignorată.

Dacă se utilizează a doua formă, iar `x` şi `y` sunt vectori de aceaşi lungime `m`, atunci se reprezintă grafic linia poligonală ce trece prin punctele de coordonate (x_i, y_i), $i = \overline{1,\mathtt{m}}$.

Programul 1.7.4. Fie programul care reprezintă grafic funcţia $\sin(2\pi x)$ pe intervalul $[0,1]$:

```
m = input('m:');
h = 1/m; x = 0:h:1;
plot(x,sin(2*pi*x))
```

Execuţia programului, pentru `m=200`, generează graficul din Figura 1.4.

În cazul în care `x` şi `y` sunt matrice având acelaşi tip `(m,n)`, se reprezintă grafic `n` linii poligonale, corespunzând celor `n` coloane ale matricelor, adică $(x_{ij}, y_{ij})_{i=1}^{\mathtt{m}}$, pentru fiecare $j = \overline{1,\mathtt{n}}$. Dacă `x` este un vector de lungime `m`, iar `y` este un vector de lungime `m` sau matrice de tipul `(m,n)`, se reprezintă grafic liniile poligonale ce trec

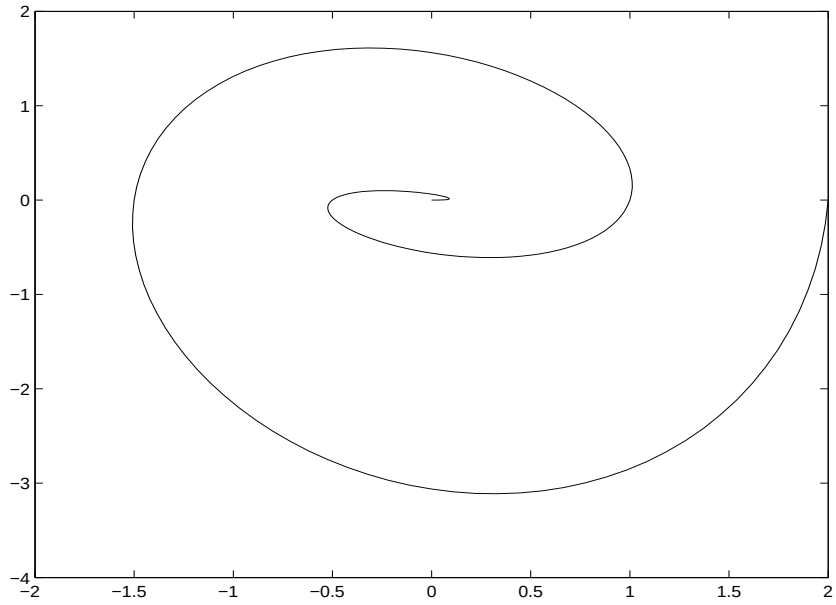

Figura 1.3: $z = x\cos(2\pi\, x) + ix^2 \sin(2\pi\, x)$

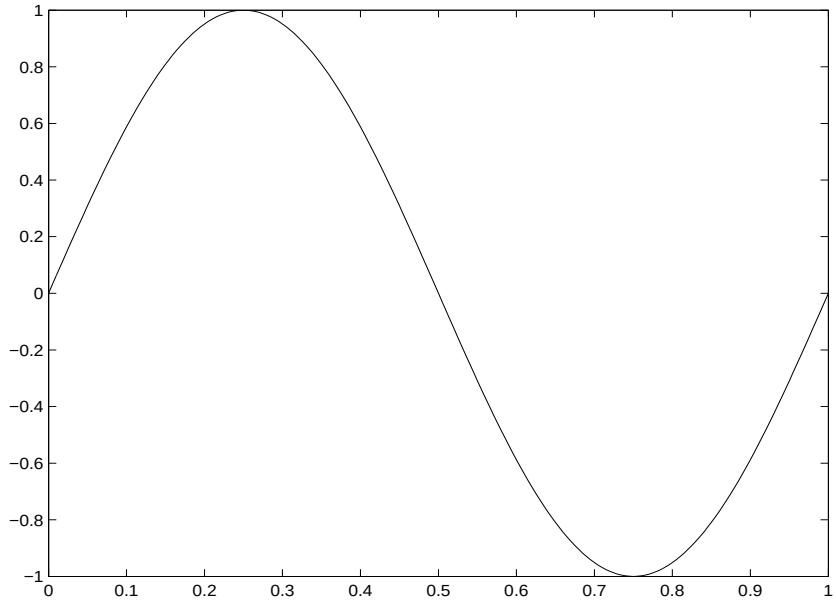

Figura 1.4: $y = \sin(2\pi x)$

prin punctele de coordonate $(x_i, y_i)_{i=1}^{\mathtt{m}}$, respectiv `n` linii poligonale ce trec respectiv prin punctele date prin $(x_i, y_{ij})_{i=1}^{\mathtt{m}}$, pentru fiecare $j = \overline{1,\mathtt{n}}$.

Forma a treia a instrucţiunii `plot` implică reprezentarea grafică pe aceeaşi figură, conform formei precedente, a grupelor de date `(x1,y1)`, `(x2,y2)` ş.a.m.d.

Programul 1.7.5. Vom prezenta două variante de programe, care reprezintă pe aceeaşi figură graficele funcţiilor $\sin(2\pi x)$ şi $\cos(2\pi x)$ pe intervalul $[0, 1]$:

```
m = input('m:');
h = 1/m; x = 0:h:1;
y = 2*pi*x;
plot(x,sin(y),x,cos(y))
```

respectiv

```
m = input('m:');
h = 1/m; x = 0:h:1;
y = 2*pi*x;
y = [sin(y)' cos(y)']; x=[x' x'];
plot(x,y)
```

Executând una din cele două variante, cu `m=200`, se generează graficul din Figura 1.5.

Remarcăm faptul că prima variantă ar fi potrivită dacă cele două funcţii se reprezintă grafic pe acelaşi domeniu, iar a doua variantă când cele două funcţii sunt reprezentate grafic pe domenii diferite.

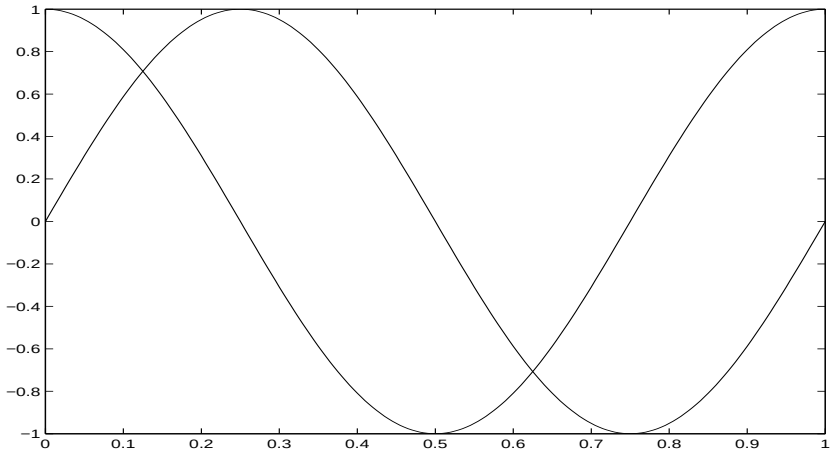

Figura 1.5: $y_1 = \sin(2\pi x)$, $y_2 = \cos(2\pi x)$

Pentru o reprezentare netedă este de dorit ca numărul punctelor ce generează liniile poligonale să fie suficient de mare.

Pe de altă parte, observăm că prin primele trei tipuri de instrucţiune `plot` nu se specifică nici o caracteristică a liniilor reprezentate grafic. Sistemul Matlab face o alegere implicită a acestor caracteristici.

Ultimele patru forme ale instrucţiunii `plot` permit utilizatorului specificarea unor caracteristici ale liniilor reprezentate grafic cu ajutorul parametrului `s`.

Parametrul de tipul `s` este un şir de cel mult trei caractere, prin care se precizează culoarea, tipul de marcaj al punctelor şi tipul de linie.

Culoarea este specificată cu unul din simbolurile următoare: `y` – galben (yellow), `m` – violet (magenta), `c` – ciclamen (cyan), `r` – roşu (red), `g` – verde (green), `b` – albastru (blue), `w` – alb (white), `k` – negru (black).

Tipul de marcaj este dat prin unul din simbolurile: `.` – punct (point), `o` – cerculeţ (circle), `x` – cruciuliţă diagonală (x–mark), `+` – cruciuliţă (plus), $*$ – steluţă (star), `s` – pătrăţel (square), `d` – romb (diamond), `v` – triunghi cu vârf în jos (triangle down), `^` – triunghi cu vârf în sus (triangle up), $<$ – triunghi cu vârf spre stânga (triangle left), $>$ – triunghi cu vârf spre dreapta (triangle right), `p` – steluţă în cinci colţuri (pentagram), `h` – steluţă în şase colţuri (hexagram).

Următoarele patru tipuri de linie sunt disponibile: continuă (solid) (`-`), punctată (dotted) (`:`), întreruptă (dashed) (`--`), linie-punct (dashdot) (`-.`).

De exemplu, dacă `s='c:'`, graficul este trasat cu linie punctată ciclamen, iar dacă `s='bd'` punctele sunt marcate prin semnul romb de culoare albastră, dar nu sunt unite între ele.

Programul 1.7.6. Vom reprezenta grafic funcţiile $\sin(2\pi x)$ şi $\cos(2\pi x)$ pe intervalul $[0,1]$, prima funcţie prin linie continuă, iar a doua prin linie întrerupă, iar punctele de pe cele două curbe, care au abscisele multipli de $\frac{1}{6}$, să fie marcate prin cerculeţe, respectiv prin steluţe.

```
m = input('m:');
h = 1/m; x = 0:h:1; y = 2*pi*x;
y = [sin(y)' cos(y)'];
h = 1/m; p = 0:1/6:1; pp = 2*pi*p;
plot(x,y,p,sin(pp),'o',p,cos(pp),'*')
```

Pentru `m=200`, graficul este cel din Figura 1.6.

1.7.2 Comenzi şi instrucţiuni de gestionare a graficelor

Sistemul Matlab dispune de comenzi şi instrucţiuni prin care sunt gestionate şi specificate anumite caracteristici ale graficelor produse privind titlul, etichetarea axelor, inserarea unor texte, marcarea unor reţele, delimitarea graficului etc:

- `hold on` – păstrează graficul curent, un nou grafic va fi supraimprimat;
- `hold off` – graficul curent este abandonat, fără a fi şters;

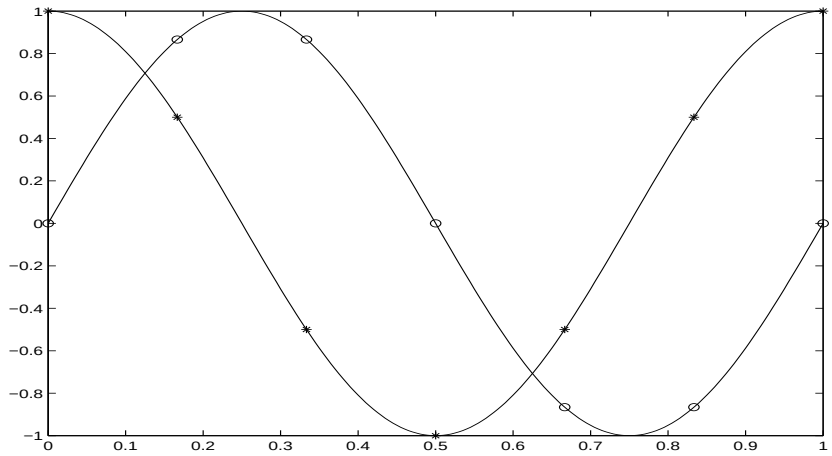

Figura 1.6: $y_1 = \sin(2\pi x)$ şi $y_2 = \cos(2\pi x)$

- `hold` – trece la modul `off`, dacă sistemul se afla în modul `on`, respectiv la modul `on`, când sistemul se află în modul `off`;
- `clf` – şterge graficul curent;
- `cla` – şterge graficul curent, păstrând axele;
- `figure(n)` – activează fereastra cu graficul `n` anterior obţinut;
- `axis([xmin xmax ymin ymax])` – fixează limitele pentru axele absciselor şi ordonatelor pentru figura curentă;
- `v = axis` – întoarce un vector linie ce conţine cele patru limite pentru axele de coordonate, care pot fi şi `-Inf`, `Inf`;
- `axis auto` – consideră forma implicită a limitelor axelor de coordonate;
- `axis manual` – păstrează limitele curente, astfel dacă este utilizată instrucţiunea `hold`, figurile următoare vor avea aceleaşi limite pentru axele de coordonate;
- `axis tight` – atribuie pentru limitele axelor cele mai apropiate valori ce rezultă din datele ce se reprezintă grafic;
- `axis ij` – configurează axele în modul `matrix`, adică originea axelor de coordonate este colţul din stânga sus, axa `i` fiind verticală şi marcată de sus în jos, iar axa `j` este orizontală şi este marcată de la stânga la dreapta;

- `axis xy` – configurează axele în modul `cartezian`, mod implicit;
- `axis equal` – unităţile de măsură pe cele două axe au aceeaşi mărime;
- `axis image` – la fel cu `axis equal`, cu excepţia că figura este restrânsă la domeniul datelor;
- `axis square` – pune unităţile de măsura pe cele două axe de coordoante astfel încât figura să fie prezentată într–un pătrat;
- `axis normal` – rearanjează axele în forma implicită;
- `axis off` – elimină axele de coordoante;
- `axis on` – inserează axele de coordoante;
- `title('text')` – afişează textul specificat în partea de sus a graficului;
- `legend('str1','str2',...)` – afişează legenda prin care se efectuează corespondenţa dintre fiecare tip de curbă a graficului şi textele specificate prin `str1`, `str2` etc.;
- `legend('str1','str2',...,p)` – afişează legenda de tipul precizat în poziţiile respectiv înafara axelor (`p=-1`), în interiorul axelor, dar pe cât e posibil să nu acopere figurile trasate (`p=0`), în partea dreaptă sus (`p=1`), care este poziţia implicită, în partea stângă sus (`p=2`), în partea stângă jos (`p=3`), în partea dreaptă jos (`p=4`);
- `legend off` – elimină legenda;
- `xlabel('text')` – etichetează axa abscielor cu textul precizat;
- `ylabel('text')` – etichetează axa ordonatelor cu textul precizat;
- `text(x,y,'text')` – inserează textul precizat prin `text`, în punctul de coordonate `(x,y)`, iar dacă `x` şi `y` sunt vectori (de aceeaşi lungime), atunci inserează textul respectiv în toate punctele de coordoante (x_i, y_i), $i = 1, 2, \ldots$, pe când dacă textul este dat printr-o matrice cu acelaşi număr de linii cu cel al lungimilor vectorilor `x` şi `y`, atunci textul din linia `i` este inserat în punctul de coordonate (x_i, y_i), $i = 1, 2, \ldots$
- `gtext('text')` – afişează graficul, împreună cu cursorul plus, unde urmează să fie inserat textul, cursorul putând fi poziţionat cu ajutorul *mouse*-ului sau tastele cu săgeţi, după care prin apăsarea oricărei taste (*mouse* sau *keyboard*), se obţine inserarea textului;

- `grid` – adaugă pe grafic o reţea rectangulară;
- `grid off` – grafic fără reţea rectangulară;
- `zoom` – permite mărirea unei părţi a figurii pentru urmărirea anumitor detalii.

Se fixează cursorul *mouse*-ului pe poziţia în care suntem interesaţi a fi mărită. Prin *dublu–click* pe butonul din stânga se produce o mărire a zonei de două ori, iar prin *dublu–click* pe butonul din dreapta se produce o micşorare a zonei de două ori. Dacă se păstrează apăsat butonul din dreapta şi se deplasează cursorul, se obţine un dreptunghi, care prin eliberarea butonului *mouse*–ului, va umple întreaga figură.

Prin instrucţiunile

```
zoom off
zoom out
```

se dezactivează instrucţiunea `zoom`, respectiv se revine la dimensiunile iniţiale ale figurii.

Comenzile de tip `axis` se impun a fi inserate după instrucţiunea `plot`.

Menţionăm că textele introduse pe grafic admit sintaxa sistemului LaTeX simplificat.

Programul 1.7.7. Vom considera un exemplu de reprezentare grafică prin *mascarea* unor părţi a graficului. Să reprezentăm grafic cu linie continuă pe intervalul `[0,6]` funcţia care coincide cu $\sin(\pi x)$, când valorile acestei funcţii sunt pozitive, iar în rest să fie zero. Pe aceaşi figură, să reprezentăm prin linie întrerupă şi $\sin(\pi x)$.

```
m = input('m:');
h = 1/m; x = 0:h:6;
y = sin(pi*x); Y=ge(y,0).*y;
plot(x,y,'k--',x,Y,'k-')
```

Pentru `m=200`, graficul este cel din Figura 1.7.

1.7.3 Instrucţiunile `polar` şi `ezpolar`

Instrucţiunea `polar` este analogă cu instrucţiunea `plot`, numai că este considerată reprezentarea curbei în coordoante polare, iar instrucţiunea `ezpolar` are sintaxa

```
ezpolar('f',[a,b])
```

şi are ca efect reprezentarea grafică a funcţiei în coordonate polare, dată prin expresia algebrică `f`, pe intervalul `[a,b]`. Al doilea parametru este opţional. Valoarea implicită este `[0,2`π`]`.

De exemplu, comanda

```
ezpolar('1+cos(theta)')
```

va produce graficul din Figura 1.8. Mai remarcăm faptul că parametrul `f` poate fi numele unei funcţii Matlab predefinite sau a utilizatorului.

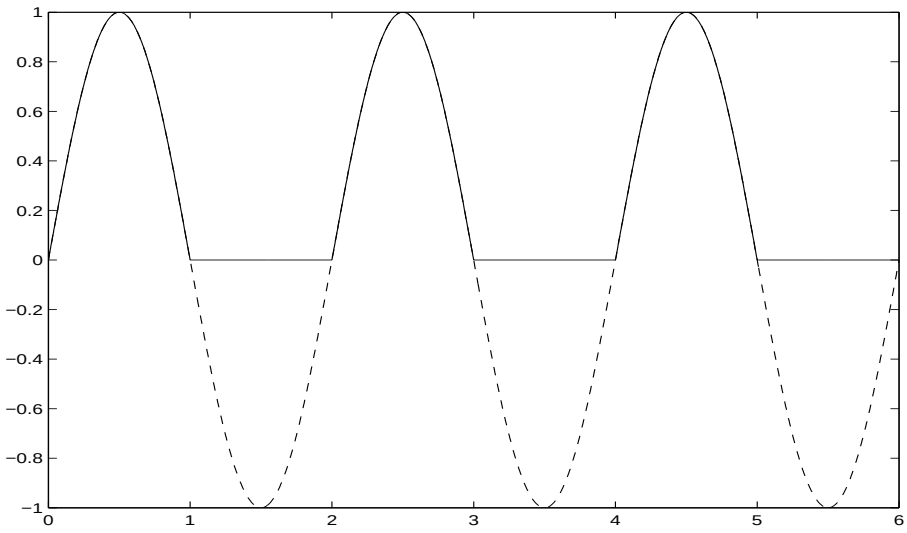

Figura 1.7: $y = \sin(\pi x)$

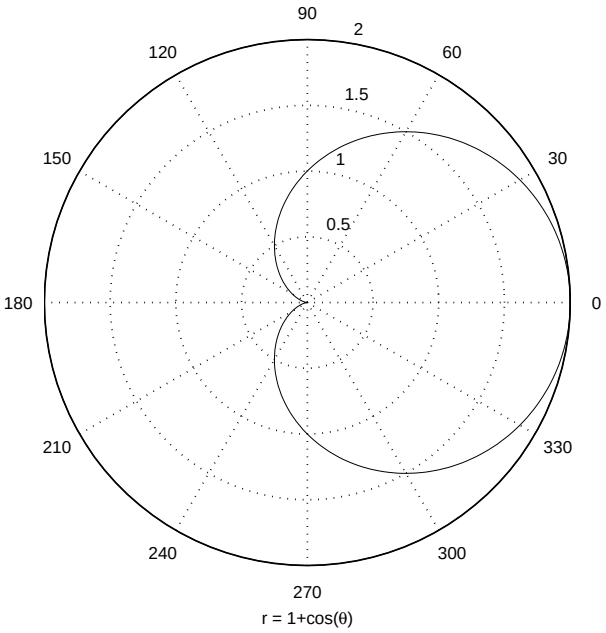

Figura 1.8: $r = 1 + \cos(\theta), \quad \theta \in [0, 2\pi]$

Programul 1.7.8. Următorul program

```
clf; t=0:0.01:pi;
r=sin(3*t).*exp(-0.3*t);
polar(t,r,'k-'), grid
```

are ca efect graficul din Figura 1.9.

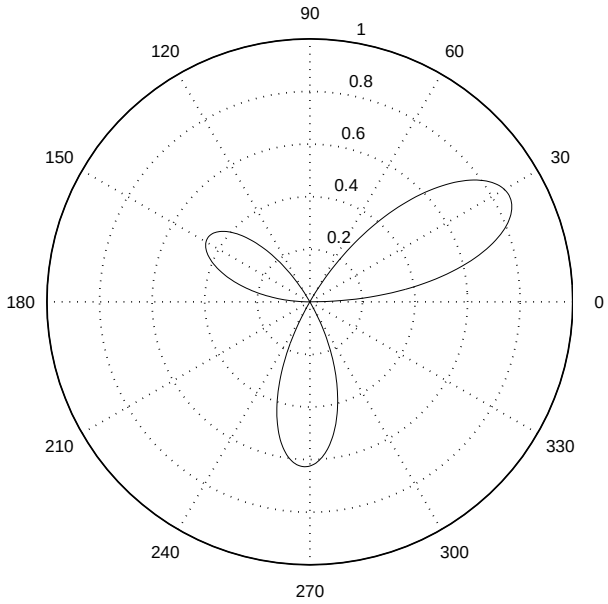

Figura 1.9: $r = \sin(3t)\, \mathrm{e}^{-0.3t}$

Instrucţiunea `colormap`

Sistemul Matlab dispune de un set de gestiune a culorilor în reprezentările grafice. Intrucţiunea `colormap` permite definirea unui set propriu pentru gestiunea culorilor. Lansarea unei instrucţiunii se face prin comanda

```
colormap(map)
colormap('default')
```

A doua comandă setează setul de culori la forma standard (implicită).

Parametrul `map` este o matrice cu trei coloane, fiecare linie conţinând elemente numere din intervalul `[0,1]`, prin care se defineşte câte o culoare. Modul de definire a culorii, după această regulă, se numeşte RGB (Red-Green-Blue), şi specifică în cadrul culorii, respectiv intensităţile culorilor roşu, verde şi albastru.

Sistemul Matlab dispune de asemenea de unele seturi de culori, care pot fi selectate şi specificate de utilizator. O parte din acestea sunt obţinute prin:

```
colormap spring
```

generează nuanţe de culori între violet (magenta) şi galben;

```
colormap summer
```

generează nuanţe de culori între verde şi galben;

```
colormap autumn
```

generează nuanţe de culori între roşu, trecând prin portocaliu spre galben;

```
colormap winter
```

generează nuanţe de culori între albastru şi verde;

```
colormap gray
```

generează nuanţe de culori gri.

1.7.4 Instrucţiunea `stairs`

Sistemul Matlab dispune de procedura `stairs` pentru reprezentarea grafică a funcţiilor în scară şi care acceptă formele instrucţiunii `plot`.

Instrucţiunea de forma

```
stairs(y)
```

reprezintă grafic funcţia în scară cu valorile precizate prin vectorul `y` de lungime `m`, iar abscisele sunt considerate numerele de la `1` la `m`. Dacă `y` este o matrice de tipul `(m,n)` vor fi reprezentate grafic `n` funcţii în scară, câte una pentru fiecare coloană a matricei `y`.

Dacă se consideră instrucţiunea de forma

```
stairs(x,y)
```

se va reprezenta grafic funcţia în scară cu valorile precizate prin vectorul `y` şi cu abscisele corespunzătoare date prin vectorul `x` (componentele lui `x` trebuie să fie crescătoare).

Dacă `y` şi `x` sunt matrice de acelaşi tip `(m,n)`, atunci se vor reprezenta grafic `n` funcţii în scară, câte una pentru fiecare din coloanele celor două matrice (elementele coloanelor matricei `x` trebuie să fie crescătoare). Dacă matricea `x` este un vector coloană cu `m` componente crescătoare, atunci funcţiile în scară corespunzătoare celor `n` coloane ale matricei `y` se consideră că au aceleaşi abscise precizate prin vectorul `x`.

Instrucţiunea de tipul

```
stairs(x,y,s)
```

permite prezentarea tipului liniei cu care se trasează graficul cu ajutorul parametrului `s` şi care are sintaxa cea precizată la instrucţiunea `plot`.

Prin urmare, putem spune că toate variantele folosite în cadrul instrucţiunii `plot` au corespondente pentru instrucţiunea `stairs`.

Programul 1.7.9. Fie *funcţia lui Haar*, numită în analiza wavelet *funcţie wavelet mamă*

$$\psi(x) = \begin{cases} 1, & \text{dacă } 0 \leqslant x < \dfrac{1}{2}, \\ -1, & \text{dacă } \dfrac{1}{2} \leqslant x < 1, \\ 0, & \text{în rest.} \end{cases}$$

Folosind funcţia ψ se defineşte familia de *funcţii wavelet*

$$\psi_{j,k}(x) = 2^{\frac{j}{2}} \psi\left(2^j - k\right)$$

unde j şi k se numesc *indice de dilatare*, respectiv *indice de translatare*. Dacă se are în vedere formula de definiţie pentru funcţia ψ, putem scrie

$$\psi_{j,k}(x) = \begin{cases} 2^{\frac{j}{2}}, & \text{dacă } \dfrac{1}{2^j}k \leqslant x < \dfrac{1}{2^j}k + \dfrac{1}{2^{j+1}}, \\ -2^{\frac{j}{2}}, & \text{dacă } \dfrac{1}{2^j}k + \dfrac{1}{2^{j+1}} \leqslant x < \dfrac{1}{2^j}k + \dfrac{1}{2^j}, \\ 0, & \text{în rest.} \end{cases}$$

Vom scrie un program care să reprezinte grafic funcţia f definită segementar cu ajutorul funcţiilor $\psi_{j,k}$, $k = \overline{r,s}$, unde j, r şi s sunt fixaţi ($r < s$).

Funcţia f este o funcţie în scară, având punctele de discontinuitate $\frac{k}{2^{j+1}}$, pentru $k = \overline{2r, 2s+2}$, cu valorile corespunzătoare acestor puncte, $2^{j/2}$, pentru k par, respectiv $-2^{j/2}$, pentru k impar.

Cu aceste precizări, avem următorul program Matlab pentru reprezentarea grafică a funcţiei f:

```
j = input('j=');
r = input('r=');
s = input('s (>r)=');
rs = 2*s-2*r+3;
fprintf('rs=2s-2r+3=%2d\n',rs);
y = input('rs valori alternative [1, -1 ... 1]:');
y = 2^(j/2)*y;
x = 1/2^j*[r:1/2:s+1];
stairs(x,y)
```

Execuţia programului, pentru `j=1, s=1, r=-2`, produce graficul din Figura 1.10.

Singurul inconvenient la execuţia acestui program este legat de instrucţiunea `input` prin care se iniţializează vectorul `y`. Când numărul funcţiilor $\psi_{j,k}$ prin care se defineşte funcţia f este mare, atunci `rs` este mare şi introducerea listei lui `y` este anevoioasă.

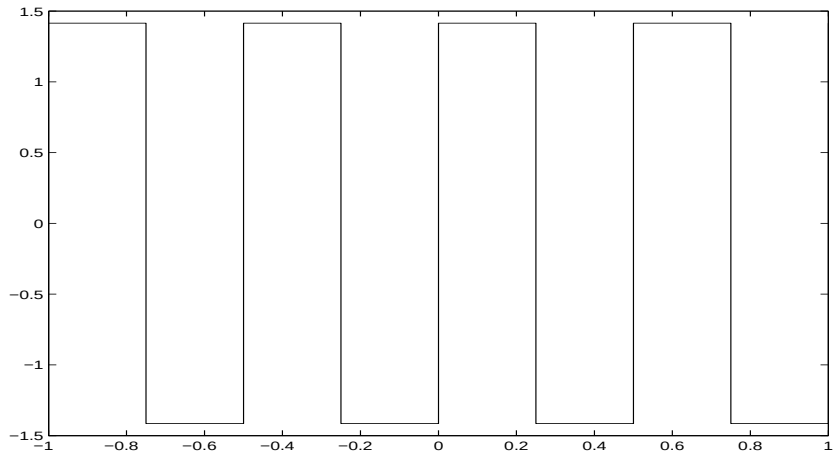

Figura 1.10: $f = (\psi_{1,-2}, \psi_{1,-1}, \psi_{1,0}, \psi_{1,1}, \psi_{1,2})$

O variantă a programului precedent, pentru care nu este necesară iniţializarea lui `y` de la tastatură, ar putea fi următorul program:

```
j = input('j=');
r = input('r=');
s = input('s (>r)=');
rs = 2*s-2*r+3;
y = (-ones(1,rs)).^[2*r:2*s+2];
y = 2^(j/2)*y;
x = 1/2^j*[r:1/2:s+1];
stairs(x,y)
```

Programul 1.7.10. Să reluăm reprezentarea grafică a funcţiilor wavelet Haar şi să scriem un program care să reprezinte pe aceeaşi figură două secvenţe de funcţii wavelet, pentru două valori consecutive ale indicelui de dilatare `j`, iar secvenţele alese să acopere intervalul $[-1, 1]$. Pentru aceasta trebuie ca `r=`-2^j, iar `s=`$2^j - 1$. Totodată să trasăm graficul primei secvenţe cu linie continuă şi punctele să le marcăm prin cerculeţe, iar pentru a doua secvenţă să alegem linie punctată pentru reprezentarea grafică, iar punctele să fie marcate cu steluţe în cinci colţuri:

```
clf, hold on, axis off
plot([-1.2 1.2],[0 0],'k:')
j = input('j=');
r = -2\^{}j; s=-r-1;
rs = 2*s-2*r+3;
y = (-ones(1,rs)).^[2*r:2*s+2];
y = 2^(j/2)*y;
x = 1/2^j*[r:1/2:s+1];
stairs(x,y,'ko-')
```

```
j = j+1; r = -2^j; s=-r-1;
rs = 2*s-2*r+3;
y = (-ones(1,rs)).^[2*r:2*s+2];
y = 2^(j/2)*y;
x = 1/2^j*[r:1/2:s+1];
stairs(x,y,'kp--')
```

Execuţia programului, pentru `j=0`, produce graficele din Figura 1.11. Se observă

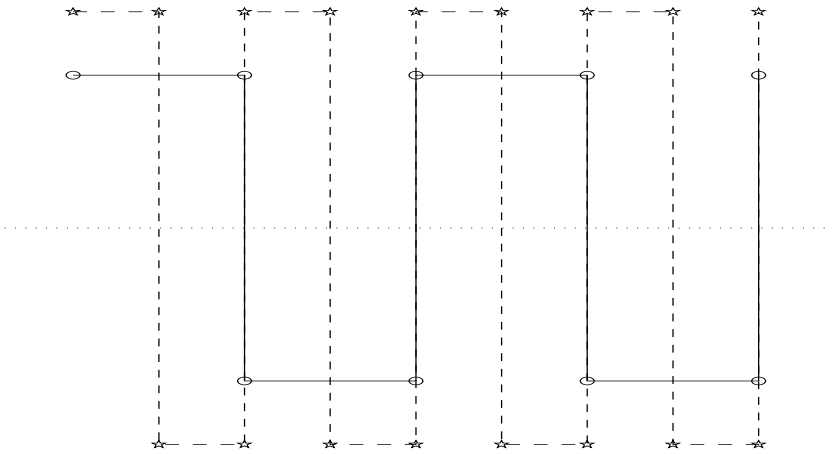

Figura 1.11: $(\psi_{0,-1}, \psi_{0,0})$, $(\psi_{1,-2}, \psi_{1,-1}, \psi_{1,0}, \psi_{1,1})$

că programul mai elimină axele sistemului Matlab, dar adaugă axa absciselor trasată punctat.

1.7.5 Instrucţiunile `bar` şi `barh`

Aceste două instrucţiuni, după cum spun şi denumirile, reprezintă anumite date numerice cu ajutorul barelor (batoanelor) verticale şi respectiv orizontale.

Forme ale acestor instrucţiuni sunt:

```
bar(y,w,'tip','color')
bar(x,y,w,'tip','color')
barh(y,w,'tip','color')
barh(x,y,w,'tip','color')
```

argumentele `w`, `'tip'` şi `'color'` sunt opţionale, dar când sunt specificate, trebuie să păstreze această ordine.

Parametrul `x` este un vector de lungime `m`.

Dacă `y` este un vector de aceeaşi lungime ca şi `x`, atunci vor fi reprezentate grafic în fiecare din punctele de abscise $\mathtt{x}_i$, $i = \overline{1, \mathtt{m}}$, câte o bară de înălţimile date prin $\mathtt{y}_i$,

$i = \overline{1,\mathtt{m}}$, care pot să fie şi negative. Dacă parametrul `x` lipseşte, atunci se consideră implicit că `x` are componentele ca fiind primele `m` numere naturale.

Dacă `y` este o matrice de tipul `(m,n)`, reprezentările sunt efectuate pentru fiecare din cele `n` coloane ale matricei, adică în fiecare punct $\mathtt{x}_i$, $i = \overline{1,m}$, de pe axa absciselor vor fi reprezentate câte `n` bare (batoane).

Parametrul `w` specifică lăţimea barelor, implicit fiind $\mathtt{w} = 0.8$, iar pentru $\mathtt{w} > 1$ barele se suprapun.

Dacă parametrul `tip=grouped`, care este valoarea implicită, atunci avem reprezentările precizate mai sus, iar dacă `tip=stacked` barele (batoanele) vor fi stivuite pe verticală.

Parametrul `color` specifică culoarea barelor (batoanelor) şi are una din următoarele valori: `r` (roşu), `g` (verde), `b` (albastru), `y` (galben), `m` (violet), `c` (ciclamen), `k` (negru), `w` (alb).

Tot ce s–a spus despre `bar` se poate spune şi pentru `barh`, numai că orientarea barelor este pe orizontală.

1.7.6 Instrucţiunea `subplot`

Sistemul Matlab are posibilitatea de a împărţi o fereastră, pentru reprezentare grafică, în $\mathtt{m}\times\mathtt{n}$ zone (ferestre), cu ajutorul instrucţiunii `subplot`.

Forma generală a instrucţiunii este

```
subplot(m,n,p)
```

unde `m`, `n` şi `p` sunt numere naturale, care exprimă faptul că fereastra iniţială se împarte în $\mathtt{m}\times\mathtt{n}$ ferestre aranjate pe `m` linii şi `n` coloane, iar reprezentarea grafică, la momentul execuţiei instrucţiunii, se va face în fereastra `p`, numerotarea ferestrelor făcându–se pe linii.

Programul 1.7.11. Reprezentarea grafică poate fi făcută şi cu ajutorul reprezentărilor parametrice ale curbelor.

Pentru reprezentarea grafică a unui cerc cu centrul în originea axelor de coordoante am putea folosi secvenţa de instrucţiuni:

```
r = input('raza=');
t = 0:0.01:2*pi;
x = r*sin(t); y=r*cos(t);
subplot(2,1,1), plot(x,y)
axis normal
subplot(2,1,2), plot(x,y)
axis square
```

Pentru `r=2`, graficul din partea stângă a Figurii 1.12, se obţine când nu s–a impus ca unităţile de măsură pe cele două axe de coordoanate să fie aceleaşi, iar graficul din partea dreaptă s–a obţinut după ce s–a înlăturat acest neajuns cu instrucţiunea `axis square`.

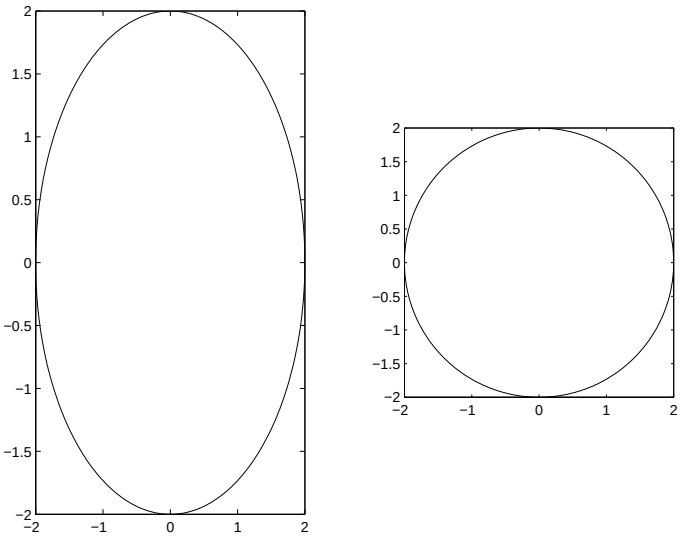

Figura 1.12: $x = r\cos t$, $y = r\sin t$, $r = 2$, $t \in [0, 2\pi]$

Programul 1.7.12. Pentru exemplificarea instrucţiunilor `bar`, `barh` şi `subplot`, facem din nou apel la o matrice magică de ordin `m` şi să reprezentăm în patru ferestre de tipul 2×2, prin bare verticale grupate şi prin bare verticale stivuite, valorile primelor două coloane ale matricei, respectiv prin bare orizontale grupate şi prin bare orizontale stivuite valorile ultimelor două coloane:

```
m=input('m:'); A=magic(m);
subplot(221), bar(A(:,[1 2]),'grouped')
title('Primele doua coloane - grupate')
subplot(222), bar(A(:,[1 2]),'stacked')
title('Primele doua coloane - stivuite')
subplot(223), barh(A(:,[m-1 m]),'grouped')
title('Ultimele doua coloane - grupate')
subplot(224), barh(A(:,[m-1 m]),'stacked')
title('Ultimele doua coloane - stivuite')
colormap summer
```

Reprezentarea grafică, pentru `m=4`, este dată în Figura 1.13.

1.7.7 Instrucţiunea `fplot`

Pentru reprezentarea grafică a unei funcţii definite cu ajutorul unei proceduri Matlab, se pot folosi una din următoarele variante ale instrucţiunii `fplot:`

```
fplot('numef',lim)
fplot('numef',lim,s)
fplot('numef',lim,tol)
```

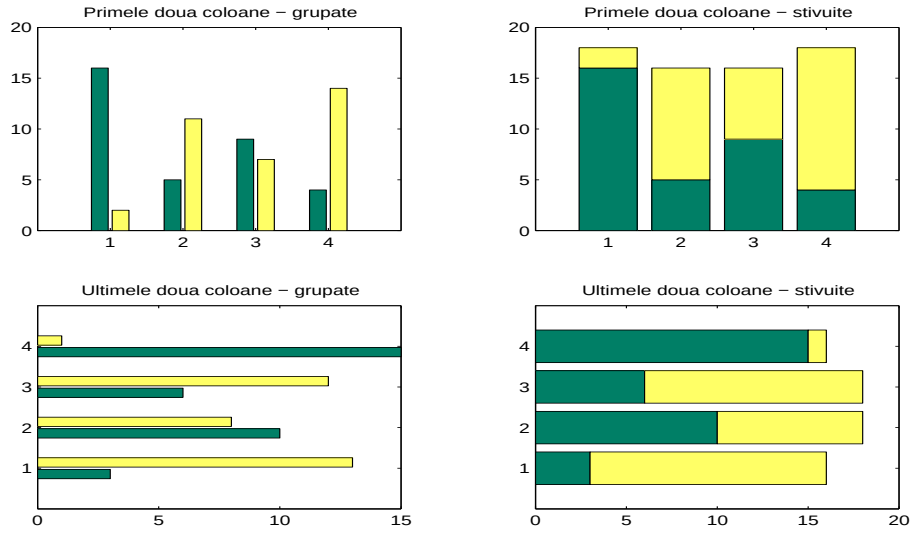

Figura 1.13: Bare verticale şi bare orizontale

```
fplot('numef',lim,tol,s)
fplot('numef',lim,n)
[X,Y] = fplot('numef',lim,...)
```

În toate cazurile se reprezintă grafic funcţia `numef`, cu limitele pentru axa abscise-lor specificate prin `lim=[xmin xmax]` sau `lim=[xmin xmax ymin ymax]`. Tipul liniei folosit în reprezentarea grafică este specificat prin parametrul `s`, ca şi în cazul instrucţiunii `plot`, iar eroarea relativă folosită în reprezentarea grafică este precizată prin parametrul `tol`, care are valoarea implicită `tol=2e-3`. Parametrul `n` specifică faptul că graficul se trasează cu ajutorul a `n+1` puncte, valoarea implicită a lui `n` este `n=1`, iar valoarea maximă a pasului nu depăşeşte $\frac{\texttt{xmax}-\texttt{xmin}}{\texttt{n}}$.

Remarcăm faptul că parametrii opţionali `s`, `tol` şi `n` pot fi luaţi în orice ordine.

Funcţia `numef` poate să întoarcă, pentru un vector `x`, un vector `y` de aceeaşi lungime ca şi `x` sau o matrice cu mai multe coloane şi număr de linii ca şi lungimea lui `x`. În primul caz, se va reprezenta grafic o singură curbă, iar în al doilea caz, pentru fiecare coloană a matricei `y` câte o curbă. Parametrul `numef` poate fi înlocuit şi cu un şir de caractere de tipul `'[f(x),g(x),...]'`, unde `f`, `g`,... sunt nume de funcţii Matlab, caz în care, pe aceeaşi figură, se vor reprezenta grafic funcţiile respective.

De exemplu, instrucţiunea

```
fplot('[sin(2*pi*x), cos(2*pi*x)]',[0 1])
```

va reprezenta grafic, pe aceeaşi figură, funcţiile $\sin(2\pi x)$ şi $\cos(2\pi x)$, pe intervalul $[0, 1]$.

Ultima formă a instrucţiunii `fplot` nu produce nici un grafic, ci păstrează punctele graficului în `X` şi `Y`, care pot fi apoi folosite efectiv, prin instrucţiunea `plot(X,Y)`, la reprezentarea grafică.

1.7.8 Instrucţiunea `ezplot`

Sintaxa instrucţiunii `ezplot` poate fi:

```
ezplot('f')
ezplot('f',[a,b])
ezplot('f',[a,b,c,d])
ezplot('x','y')
ezplot('x','y',[a,b])
```

Parametrul `f` reprezintă o expresie algebrică de una sau două variabile.

Dacă `f` este o expresie algebrică de o singură variabilă, atunci este reprezentată grafic funcţia definită de această expresie, pe intervalul `[a,b]`. Valoarea implicită a acestui parametru este `[-2`π`,2`π`]`.

Dacă `f` este o expresie algebrică de două variabile, atunci este reprezentată grafic funcţia implicită definită prin `f=0`, pe domeniul `[-2`π`,2`π`]`$\times$`[-2`π`,2`π`]`, dacă se foloseşte prima formă de apel, pe domeniul `[a,b]`$\times$`[a,b]`, pentru a doua formă, respectiv pe `[a,b]`$\times$`[c,d]`, pentru a treia formă. Cele două variabile sunt considerate ca fiind ordonate lexicografic.

Ultimele două forme sunt folosite pentru reprezentările parametrice, `x` reprezentând abscisa, iar `y` ordonata. Argumentul acestor parametri variază în intervalul `[a,b]`, iar dacă acesta lipseşte, atunci se consideră intervalul implicit `[0,2`π`]`.

De exemplu, instrucţiunea

```
ezplot('sin(3*t)*cos(t)','sin(3*t)*sin(t)',[0,pi])
```

produce graficul din Figura 1.14. Mai remarcăm faptul că parametrii `f`, `x` şi `y`, pot fi numele unor funcţii Matlab predefinite sau ale utilizatorului.

1.7.9 Instrucţiunea `fill`

Instrucţiunea `fill` este folosită pentru umbrirea interiorului unui poligon specificat prin vârfurile sale:

```
fill(x,y,c)
```

unde `x` şi `y` specifică vârfurile poligonului, iar `c` reprezintă culoarea dorită pentru interiorul poligonului.

Instrucţiunea poate fi utilizată pentru umbrirea unei părţi dintr–o figură.

Programul 1.7.13. Să scriem un program care să umbrească aria cuprinsă între axa absciselor, curba lui Gauss şi dreptele perpendiculare pe axa absciselor în punctele `a` şi `b` (`-3<a<b<3`).

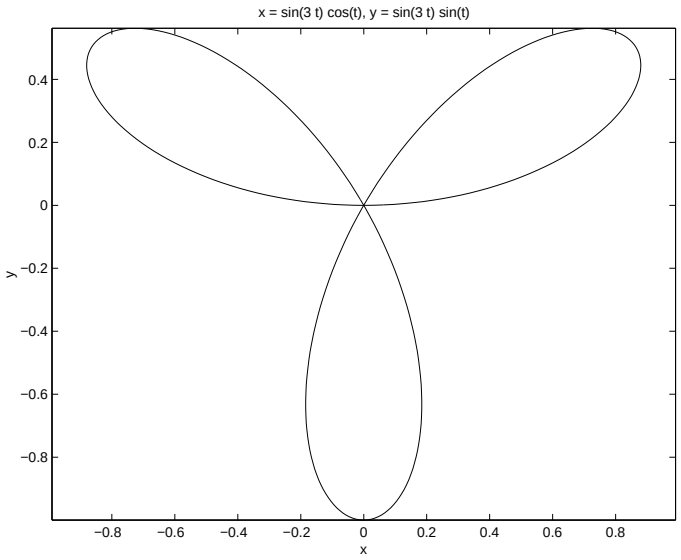

Figura 1.14: $x = \sin(3t)\cos(t)$, $y = \sin(3t)\sin(t)$, $t \in [0, \pi]$

Curba lui Gauss este dată prin funcţia

$$f(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}.$$

```
a = input{'a='}; b = input('b=');
xmin = -3; xmax = 3;
fplot{'1/sqrt(2*pi)*exp(-x^2/2)',[xmin xmax]}
x = a; y = 0;
t = a:0.01:b;
x = [x,t]; y = [y,1/sqrt(2*pi)*exp(-t.^2/2)];
x = [x,b]; y = [y,0];
fill(x,y,'c')
```

Pentru `a=-2` şi `b=1.5`, se obţine graficul din Figura 1.15.

1.8 Instrucţiuni de ciclare şi control

Ca şi în celelalte limbaje de programare evoluate, sistemul Matlab dispune de instrucţiuni, care permit repetarea unei secvenţe de instrucţiuni de un număr definit sau indefinit de ori, precum şi de instrucţiuni, care permit ieşirea din executarea secvenţială a instrucţiunilor unui program.

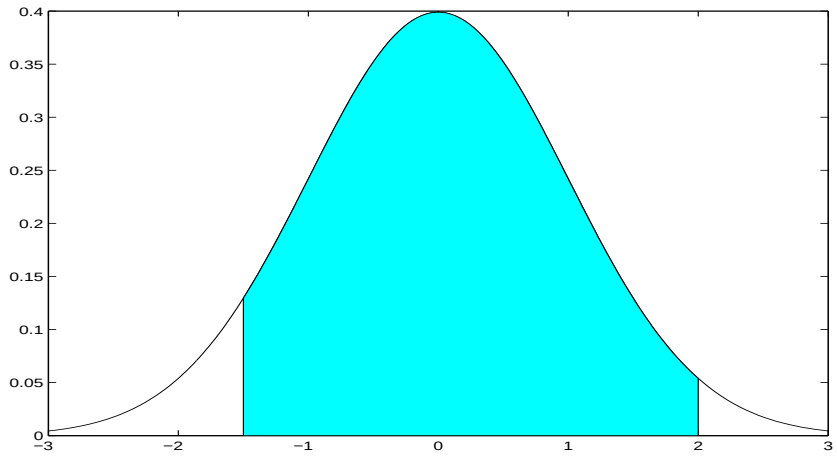

Figura 1.15: Curba lui Gauss

1.8.1 Instrucţiunea `if`

Trei forme de bază ale instrucţiunii `if` sunt:

```
if exp
    secv
end
%%%%%%%%%%%%%%
if exp
    secv
else
    secv1
end
%%%%%%%%%%%%%%
if exp
    secv
elseif exp1
    secv1
else
    secv2
end
```

unde `exp` şi `exp1` sunt expresii logice ce pot fi evaluate în acest punct, iar `secv`, `secv1` şi `secv2` sunt secvenţe de instrucţiuni Matlab.

Instrucţiunea `if` sub prima formă face să se execute secvenţa de instrucţiuni `secv` numai dacă valoarea expresiei logice `exp` este adevărată, după care se trece la prima instrucţiune ce urmează liniei `end`.

A doua formă de instrucţiune `if` are ca efect executarea uneia dintre secvenţele de instrucţiuni `secv` şi `secv1`, în funcţie de faptul că expresia logică `exp` este adevărată, respectiv falsă, după care se trece la instrucţiunea ce urmează după linia `end`.

Ultima formă are ca rezultat executarea uneia dintre secvenţele de instrucţiuni `secv`, `secv1` şi `secv2`, după cum valoarea lui `exp` este adevărată, a lui `exp` este falsă şi a lui `exp1` este adevărată, respectiv valorile expresiilor logice `exp` şi `exp1` sunt false şi `exp2` are valoare adevărată.

De exemplu secvenţa de instrucţiuni ce urmează stabileşte dacă elementele unui vector `v` cu `m` componente întregi sunt toate numere pare, toate impare sau sunt şi pare şi impare:

```
if rem(v,2)==0
    p=0;
elseif rem(v,2)==1
    p=1;
else
    p=2;
end
p
```

Rezultatul execuţiei acestei secvenţe atribuie lui `p` una din valorile `0`, `1`, `2`, în funcţie de faptul că vectorul `v` are toate componentele numere pare, toate impare, respectiv şi pare şi impare.

1.8.2 Instrucţiunea `switch`

Această instrucţiune are ca rezultat trecerea execuţiei, în funcţie de valoarea unei expresii, în diferite puncte ce urmează.

Forma generală a instrucţiunii `switch` este:

```
switch exp
case val1
    secv1
case val2
    secv2
. . . . .
otherwise
    secv
end
```

unde `exp` este o expresie aritmetică sau de tip caracter, iar `secv1`, `secv2`, ..., `secv` sunt secvenţe de instrucţiuni Matlab. Dacă valoarea expresiei `exp` coincide cu cu una din valorile `val1`, `val2`, ... atunci se execută secvenţa de instrucţiuni corespunzătoare `secv1`, `secv2`, ..., iar în caz contrar se execută secvenţa de instrucţiuni `secv`.

De exemplu, secvenţa de instrucţiuni

```
m=input('m=');
switch expr
case 'ones'
    A=ones(m)
case 'eye'
    A=eye(m)
case 'magic'
    A=magic(m)
otherwise
    A=zeros(m)
end
```

generează matricea pătratică `A` de ordin `m`, care poate fi de la caz la caz, respectiv matrice cu toate elementele `1`, matrice unitate, matrice magică, matrice cu toate elementele zero.

1.8.3 Instrucţiunea `while`

Instrucţiunea `while` este o instrucţiune de ciclare prin care se repetă execuţia unei secvenţe de instrucţiuni Matlab până când o condiţie este satisfăcută. Forma generală a instrucţiuni este:

```
while exp
    secv
end
```

şi are ca efect executarea secvenţei de instrucţiuni `secv` atâta timp cât valoarea expresiei logice `exp` este adevărată.

De exemplu, dacă vrem să calculăm valoarea aproximativă a soluţiei ecuaţiei $x = \cos(x)$, folosind metoda iterativă, impunând condiţia de oprire prin aceea ca numărul iteraţiilor să nu depăşească o valoare dată `n` şi distanţa dintre două iterate consecutive să nu fie mai mare decât un `epsilon`$>$`0` dat, atunci avem programul

```
n=input('n=');
epsilon=input('epsilon=');
xv=pi/4; k=1; d=1;
while d>epsilon & k<n
    k=k+1; xn=cos(xv);
    d=abs(xn-xv); xv=xn;
end
fprintf('k=%2d\n',k)
fprintf('d=%10.4f\n',d)
fprintf('x=%10.4f\n',xn)
```

Astfel, pentru valorile la intrare `n=50` şi `epsilon=1e-5`, sunt afişate rezultatele:

```
k=25
d=    0.0000
x=    0.7391
```

1.8.4 Instrucţiunea `for`

Faţă de instrucţiunea `while`, instrucţiunea de ciclare `for` are fixat numărul repetărilor unei secvenţe de instrucţiuni.

Forma generală a instrucţiunii este:

```
for v=vi:pas:vf
    secv
end
```

unde `v` este variabila de ciclare, care ia toate valorile, începând cu valoarea expresiei aritmetice `vi` până la valoarea expresiei `vf`, folosind pasul dat prin expresia aritmetică `pas`. Pentru fiecare din aceste valori ale lui `v` se execută succesiv secvenţa `secv` de instrucţiuni Matlab.

Dacă `pas=1`, atunci putem folosi următoarea formă a instrucţii `for`:

```
for v=vi:vf
    secv
end
```

Mai remarcăm faptul că în interiorul unui ciclu `for` poate exista un alt ciclu `for` ş.a.m.d.

De exemplu, secvenţa următoare va genera matricea `h` de tipul `(m,n)`, numită *matricea lui Hilbert*:

```
m=input('m=');
n=input('n=');
h=[];
for i=1:m
    for j=1:n
        h(i,j)=1/(i+j-1);
    end
end
format rat
disp(h)
```

În urma executării acestui program, cu valorile `m=3` şi `n=6`, se obţin rezulatele

```
m=3
n=4
      1           1/2          1/3          1/4
     1/2          1/3          1/4          1/5
     1/3          1/4          1/5          1/6
```

Mai considerăm un exemplu de folosire a instrucţiunii `for`. Vrem să construim *matricea Vandermonde* `V`, corespunzătoare numerelor a_i, $i = \overline{1,m}$, care sunt date

prin vectorul a. Reamintim că forma generală a aceastei matrice este

$$\mathsf{V} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ a_1 & a_2 & \dots & a_m \\ a_1^2 & a_2^2 & \dots & a_m^2 \\ \dots & \dots & \dots & \dots \\ a_1^{m-1} & a_2^{m-1} & \dots & a_m^{m-1} \end{pmatrix}.$$

Matricea V poate fi generată prin secvenţa de program următoare:

```
m=input('m=');
for i=1:m
    a(i)=input('a(i)=');
end
V=[];
for i=0:m-1
    V=[V;a.^i];
end
disp(V)
```

Se observă că matricea V a fost inţializată cu *matricea vidă*, care este specificată prin [], iar construcţia efectivă s–a făcut prin operaţia de concatenare.

Executând programul pentru m=4 şi $\mathsf{a}(i) = i$, $i = \overline{1,4}$, se obţine

```
m=4
a(i)=1
a(i)=2
a(i)=3
a(i)=4
     1     1     1     1
     1     2     3     4
     1     4     9    16
     1     8    27    64
```

1.9 Instrucţiuni de întrerupere

1.9.1 Instrucţiunea try...catch

Forma generală a instrucţiunii este:

```
try
    secv
catch
    secv1
end
```

şi are ca efect execuţia primei secvenţe de instrucţiuni, secv, până la apariţia unei erori, după care se execută a doua secvenţă de instrucţiuni, secv1. Dacă nici o

eroare nu apare în execuţia secvenţei `secv`, după executarea acesteia, se trece la prima instrucţiune după linia `end`. Dacă la execuţia secvenţei de instrucţiuni `secv1` apare o eroare, execuţia programului este oprită, dacă o altă instrucţiune de tipul `try...catch` nu există.

1.9.2 Instrucţiunea `pause`

Instrucţiunea `pause` are ca efect întreruperi temporare ale execuţiei programelor.

Instrucţiunea `pause` are următoarele forme:

```
pause(n)
pause
pause off
pause on
```

Prima formă are ca efect întreruperea executării programului timp de `n` secunde (`n` este număr pozitiv), când se ajunge cu executarea în acest punct.

A doua formă, `pause` fără argument, întrerupe executarea programului, reluarea executării făcându–se prin apăsarea oricărei taste.

Instrucţiunea `pause` cu argumentul `off` are ca efect inhibarea tuturor instrucţiunilor `pause` şi `pause(n)` ce urmează, iar dacă are argumentul `on`, atunci se produce reactivarea acestora.

1.9.3 Instrucţiunea `return`

Instrucţiunea

```
return
```

are ca efect întoarcerea în programul de pe nivelul imediat superior (programul care apelează unitatea în care se află instrucţiunea). Nivelul cel mai înalt este cel care corespunde modului interactiv de lucru, prin urmare instrucţiunea `return`, la acest nivel nu are niciun efect.

1.9.4 Instrucţiunea `break`

Instrucţiunea

```
break
```

are efecte diferite în funcţie de locul unde este poziţionată.

Dacă instrucţiunea `break` se află într–un ciclu `while`, se produce ieşirea forţată din ciclul respectiv, iar dacă se află într–un ciclu `for` se produce la fel o ieşire forţată din ciclu, cu observaţia că dacă există un ciclu `for` superior, se continuă cu executarea acestuia.

Dacă instrucţiunea `break` se află într–o instrucţiune de tipul `if`, `switch` sau `try...catch`, atunci se abandonează execuţia acestei instrucţiuni, iar dacă este poziţionată în oricare altă parte, se încheie executarea programului.

1.9.5 Instrucţiunea `error`

Instrucţiunea

```
error('mesaj')
```

are ca efect abandonarea execuţiei şi afişarea mesajului precizat prin `mesaj`.

1.10 Funcţii (proceduri) în Matlab

Toate programele scrise până aici, numite *fişiere script*, ar putea fi considerate ca fiind *programe principale*. Un astfel de program principal poate să apeleze alte programe, pe care le numim *funcţii* sau *proceduri*, care la rândul lor pot face apel la alte proceduri.

Un program principal (fişier script), nu poate fi apelat de o altă unitate de program. Lansarea execuţiei unui astfel de program, se poate face numai când sistemul Matlab este în mod interactiv, adică avem pe ecran cursorul $>>$ al sistemului Matlab. Lansarea se face în două moduri:

- prin tastarea, după cursorul $>>$, a numelui său, fără extensia `.m`;
- prin copierea, după cursorul $>>$, a conţinutului fişierului ce conţine programul, un exemplu fiind cuplul de comenzi `Copy` – `Paste`,

după care se tastează comanda `Enter`.

Remarcăm faptul că toată istoria unei sesiuni se află pe ecran, chiar dacă unele secvenţe mai vechi nu sunt vizibile la un momnet dat, dar cu ajutorul *mouse*–ului, folosind *bara* din dreapta a ecranului se poate aduce secvenţa dorită în partea vizibilă a ecranului. Mai mult, cu ajutorul sistemului de săgeţi pot fi readuse comenzi mai vechi sau mai noi, care pot fi reeditate (modificate) şi lansate după o astfel de reeditare. Există chiar o căutare mai rapidă a unor instrucţiuni din lista de instrucţiuni ce au fost utlizate într–o sesiune Matlab, anume se tastează începutul instrucţiunii urmată de săgeată în sus.

Încă ceva, toate variabilele folosite în timpul unei sesiuni Matlab, fie în mod interactiv, fie prin execuţia unui program principal (*script–file*) au un caracter global şi sunt păstrate într–un spaţiu de lucru (`workspace`), care la încheierea sesiunii Matlab sunt pierdute. De aceea, avem la dispoziţie comenzile pe care le-am prezentat, `diary` şi `save` care permit, prin lansarea lor, salvarea acestor informaţii.

1.10.1 Definirea şi structura unei funcţii Matlab

Colecţia de programe ale sistemului Matlab este constituită, în mare parte, din funcţii (proceduri) scrise în limbajul propriu. Acestea sunt fişiere cu extensia `.m` având următoarea structură:

```
function [lista1] = numef(lista2)
% linia H1 (help line)
% comentariu
%
corp
```

unde `lista1` şi `lista2` reprezintă respectiv lista variabilelor de ieşire (despărţite prin spaţii sau virgule) şi lista parametrilor de intrare (despărţiţi prin spaţii sau virgule), `corp` este secvenţa de instrucţiuni ale funcţiei cu numele `numef`.

Variabilele din corpul procedurii au un caracter local, prin urmare la întoarcerea în unitatea care face apelul valorile acestora se pierd, desigur cu excepţia valorilor din `lista1`. Există posibilitatea schimbării caracterului local, prin instrucţiunea

```
global lista
```

care are ca efect obţinerea caracterului global pentru variabilele din `lista`.

În corpul unei funcţii Matlab variabilele `nargin` şi `nargout` specifică numărul parametrilor de intrare, adică lungimea listei `lista2`, respectiv a parametrilor de ieşire, adică lungimea listei `lista1`.

Dacă într–un fişier script sau de la tastatură se dă una din comenzile

```
nargin('numef')
nargout('numef')
```

atunci se obţin numărul parametrilor de intrare şi respectiv numărul parametrilor de ieşire pentru funcţia Matlab cu numele `numef`.

Liniile dinaintea corpului procedurii, care încep cu % sunt linii de comentariu şi vor fi listate când se lansează comanda

```
help numef
```

Din acest motiv, partea de comentariu de la începutul procedurii este bine să conţină informaţii, care să descrie pe scurt cum trebuie să fie apelată şi folosită funcţia respectivă.

Mai remarcăm prima linie de comentariu, notată cu `H1`, care are o însuşire aparte. Când se lansează comanda

```
lookfor key
```

care caută cuvântul cheie `key`, aceasta se face numai în liniile `H1` ale fişierelor cu extensia `.m`, după care se listează numele fişierelor ce conţin cuvântul cheie precizat.

În cadrul liniilor ce alcătuiesc `corp`–ul funcţiei pot exista de asemenea linii de comentariu, care încep cu simbolul %, dar care nu vor fi afişate odată cu lansarea comenzii `help`. Mai mult apariţia simbolului % în interiorul unei linii face ca tot ce urmează pe linia respectivă să fie considerat comentariu.

Numele unei funcţii Matlab, `numef`, scrisă de utilizator, este de dorit să fie diferit de numele funcţiilor deja existente în sistemul Matlab, iar numele fişierului, cu extensia `.m`, ce va conţine funcţia să fie `numef.m`.

Dacă lista parametrilor de intrare, `lista2`, este vidă, atunci linia de definiţie a funcţiei este de forma

```
function [lista1] = numef
```

Dacă lista variabilelor de ieşire, `lista1`, este formată dintr–o singură variabilă, atunci se poate renunţa la parantezele drepte, iar dacă este vidă se poate elimina complet.

Cu aceste precizări, cea mai simplă linie de definţie a unei funcţii (proceduri) este

```
function numef
```

1.10.2 Apelul unei funcţii (proceduri) Matlab

Apelarea unei funcţii se poate face în două moduri.

Primul mod este cel interactiv, de exemplu, prin una din formele

```
[apel1] = numef(apel2)
numef(apel2)
```

unde `apel2` este lista parametrilor actuali, care poate fi vidă sau formată din cel mult atâtea elemente cât are lista parametrilor formali, `lista2`, din linia de definiţie a funcţiei `numef`, de fiecare dată făcându–se legătura de tipul: primul parametru din `apel2` corespunde primului parametru din `lista2`, al doilea element din `apel2` corespunde celui de al doilea parametru din `lista2`,..., corespondenţa încheindu–se când se termină lista `apel2`. Desigur, dacă `apel2` are mai puţine elemente decât `lista2`, atunci trebuie gestionată această situaţie în corpul funcţiei `numef` cu ajutorul varaibilei `nargin`.

Rezultatele executării funcţiei `numef`, se păstrează în lista `apel1` în primul caz, respectiv în `ans` în al doilea caz. O discuţie analogă asupra corespondenţei dintre `apel2` şi `lista2`, se face şi în cazul `apel1` şi `lista1`.

O altă cale de apel a unei funcţii este dintr–un alt program sau procedură Matlab, prin apariţia într–o instrucţiune Matlab a numelui funcţiei, împreună cu lista parametrilor actuali de tipul listei `apel2`, mai sus precizată.

1.10.3 Subfuncţii

Funcţiile pot avea în `corp`–ul lor apeluri la alte funcţii. Totdeauna întoarcerea în unitatea care cheamă se face, când se ajunge la linia finală a `corp`–ului unităţii chemate sau când se întâlneşte instrucţiunea `return`.

Există două situaţii mai speciale în construcţia funcţiilor în Matlab.

În primul rând, există posibilitatea ca după `corp`–ul unei funcţii să fie definite una sau mai multe funcţii, pe care le numim *subfuncţii* şi care pot fi apelate din `corp`–ul funcţiei. Forma unei astfel de funcţii ar fi:

```
function [lista1] = numef(lista2)
%
% comentariu
%
corp
function [list1] = numef1(list2)
%
% comentariu1
%
corp1
```

în secvenţa de instrucţiuni `corp` făcându–se apel la funcţia `numef1`. Funcţiile de tipul funcţiei `numef1` nu se pot apela din exteriorul fişierului `numef.m`. Drept consecinţă nici `comentariu1` nu este vizibil din exteriorul funcţiei `numef`.

Funcţia 1.10.1. Să scriem o procedură, care reprezintă grafic pe intervalul $[-1,1]$ o mulţime finită precizată de polinoame ale lui Cebîşev de speţa I. Dacă numărul acestora este par, atunci să fie reprezentate pe două coloane, iar în caz contrar pe o singură coloană.

```
function cebisev(lista)
% Polinomul Cebisev
% lista -- contine gradele polinoamelor Cebisev
%          care vor fi reprezentate grafic;
% daca lista are un numar par de elemente,
% polinoamele se vor reprezenta pe doua coloane;
% daca lista are un numar impar de elemente,
% polinoamele se vor reprezenta pe o coloana;
%
clf, x=-1:0.001:1; r=length(lista);
if rem(r,2)==0
    m=fix(r/2); n=2;
else
    m=r; n=1;
end
for i=1:r
    y=cos(lista(i)*acos(x));
    splots(m,n,i,x,y,lista(i))
end
function splots(m,n,k,u,v,gr)
subplot(m,n,k), plot(u,v,'k-')
title(['Polinom Cebisev de gradul n=',num2str(gr)])
```

Dacă se apelează funcţia `cebisev` prin

```
cebisev([2:5])
```

se obţin graficele din Figura 1.16.

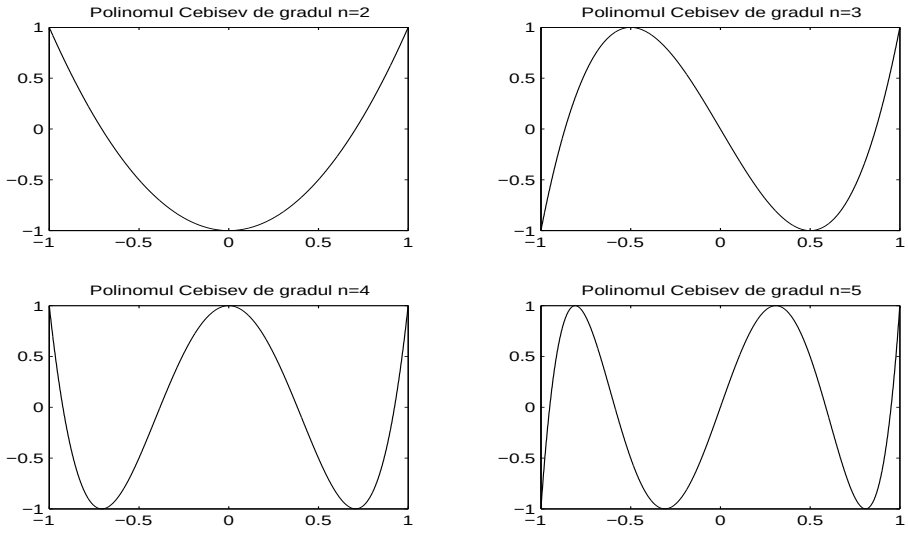

Figura 1.16: Polinoamele lui Cebîşev de grad $n = 2, 3, 4, 5$

1.10.4 Funcţia `feval`

Un alt caz special este acela când printre parametrii formali de intrare din `lista2` există şi parametri ce pot lua valori nume ale unor funcţii. La apelul unei astfel de funcţii pe poziţia parametrului actual corespunzător din lista `apel2` trebuie să se afle un şir de caractere ce denumeşte funcţia actuală.

Funcţia 1.10.2. Funcţia cu numele `opmatrix` operează prin operatorii Matlab cunoscuţi asupra a două matrice:

```
function Z = opmatrix(op,X,Y)
%
% Efectueaza operatia definita prin op
% asupra matricelor X si Y.
% Rezultatul este returnat in Z
%
Z = feval(op,X,Y);
```

Folosind această funcţie, de exemplu prin comenzile

```
opmatrix('plus',A,B)
opmatrix('minus',A,B)
opmatrix('and',A,B)
```

sau

```
opmatrix('+',A,B)
opmatrix('-',A,B)
opmatrix('&',A,B)
```

se obţin suma, diferenţa şi respectiv conjuncţia matricelor A şi B.

1.10.5 Comanda echo

Instrucţiunea echo are un comportament puţin diferit pentru cele două tipuri de fişiere cu extensia .m.

Dacă suntem în cazul unui *fişier script*, atunci instrucţiunile

```
echo on
echo off
echo
```

au respectiv următoarele efecte: prima are ca efect afişarea pe ecran a fişierului în derulare, a doua anulează acest efect, iar a treia are efectul primei forme, dacă sistemul se află în mod off, iar dacă se află în mod off se va face trecerea la modul on.

Dacă suntem în cazul funcţiilor Matlab cu extensia .m, atunci avem formele

```
echo numef on
echo numef off
echo numef
echo on all
echo off all
```

Primele trei forme sunt analoge celor trei din cazul *script*, numai că se mai specifică şi numele funcţiei la care se referă. Celelalte două acţionează asupra tuturor funcţiilor.

1.11 Instrucţiuni de evaluare a eficienţei

Compararea eficienţei algoritmului este posibilă în sistemul Matlab fie prin numărarea operţiilor în virgulă flotantă, fie prin determinarea duratei, în secunde, a execuţiei unei secvenţe de instrucţiuni, CPU (Central Processor Unit).

1.11.1 Instrucţiunea flops

Instrucţiunea flops este încorporată în Matlab 6 prin pachetul LAPACK. Prin urmare este funcţională numai dacă sistemul conţine şi LAPACK.

Obţinerea numărului operaţiilor în virgulă flotantă efectuate pentru execuţia unei secvenţe secv de instrucţiuni Matlab se realizează prin:

```
flops(0)
   secv
flops
```

La încheierea execuţiei acestei succesiuni de instrucţiuni în variabila flops se află numărul acestor operaţii.

1.11.2 Instrucţiunile `tic` şi `toc`

Pentru obţinerea duratei execuţiei unei secventţe `secv` de instrucţiuni Matlab avem următoarea succesiune de instrucţiuni:

```
tic
  secv
toc
```

unde `tic` marchează începutul contorizării timpului, iar `toc` anunţă încheierea contorizării, variabila `toc` conţinând timpul măsurat în secunde.

1.12 Grafică tridimensională

Grafica tridimensională sau altfel numită `3D` este axată în sistemul Matlab pe două direcţii: reprezentarea curbelor în spaţiu şi reprezentarea suprafeţelor.

1.12.1 Instrucţiunea `plot3`

Instrucţiunea `plot3` este utilizată pentru reprezentarea curbelor în spaţiu şi în principiu nu se deosebeşte mult de instrucţiunea `plot` din grafica bidimensională.

Formele întâlnite pentru această instrucţiune sunt:

```
plot3(x,y,z)
plot3(x,y,z,s)
plot3(x1,y1,z1,x2,y2,z2,...)
plot3(x1,y1,z1,s1,x2,y2,z2,s2,...)
```

Dacă `x`, `y`, `z` sunt vectori de acceaşi lungime `m`, atunci se unesc punctele de coordonate $(\mathtt{x}_i, \mathtt{y}_i, \mathtt{z}_i)$, $i = \overline{1, \mathtt{m}}$, printr–o linie poligonală de tipul precizat prin parametrul `s`, de la instrucţiunea `plot`, iar în lipsa acestui parametru, sistemul Matlab efectuează o alegere implicită a tipului curbei. Desigur că netezimea curbei este cu atât mai bună cu cât numărul punctelor este mai mare.

Dacă `x`, `y`, `z` sunt matrice de acelaşi tip `(m,n)`, atunci pentru fiecare coloană se obţine câte o linie poligonală, adică `n` linii poligonale. Acestea sunt generate respectiv de punctele de coordonate $\left(\mathtt{x}_{ij}, \mathtt{y}_{ij}, \mathtt{z}_{ij}\right)_{i=1}^{\mathtt{m}}$, pentru fiecare $j = \overline{1, \mathtt{n}}$, cu aceeaşi observaţie privind parametrul `s`, care precizează tipul curbei. În acest caz pe aceeaşi figură vor fi obţinute mai multe grafice.

Ultimele două forme sunt extinderi ale primelor două, pentru a putea controla mai bine în special tipul curbelor cu ajutorul parametrilor de tip `s`.

Remarcăm de asemenea extinderea acţiunilor comenzilor ce controlează un grafic din cazul bidimensional pentru cazul tridimensional.

Programul 1.12.1. Prezentăm un program care ilustrează mişcarea unei particule din punctul `A` pe o spirală până în punctul `B`.

```
pas=input('pas='); h=input('h=');
t=0:pas:h; r=exp(-0.2*t); th=pi*t*0.5;
x=r.*cos(th); y=r.*sin(th); z=t;
plot3(x,y,z), hold on
plot3([1,1],[-0.5,0],[0,0])
text(1,-0.7,0,'A'), n=length(t);
text(x(n),y(n),h+2,'B')
xlabel('x'), ylabel('x'), zlabel('z')
```

Să explicăm pe scurt o parte a instrucţiunilor acestui program.

Figura este obţinută în două etape, din cauza aceasta a fost introdusă instrucţiunea `hold on`. Prima instrucţiune `plot3` trasează efectiv graficul spiralei, care se mai completează prin segmentul din planul `xoy` delimitat de punctele de coordonate `(1,-0.5,0)` şi `(1,0,0)`, trasat cu a doua instrucţiune `plot3`. Cele două instrucţiuni au acelaşi rezultat ca şi instrucţiunea

```
plot3(x,y,z,[1,1],[-0.5,0],[0,0])
```

doar că spirala va fi trasată cu o anumită culoare, iar segmentul din planul `xoy` cu alta. Dacă vrem ca întreaga curbă să fie trasată cu aceeaşi culoare se poate apela la parametrul de tip `s`.

Se mai observă că poziţia iniţială `A` a punctului şi poziţia finală `B` sunt afişate cu instrucţiunile `text`, iar etichetele celor trei axe de coordoante sunt date prin instrucţiunile de tip `label`.

După executarea programului, cu `h=20` şi `pas=0.01`, s–a obţinut graficul din Figura 1.17.

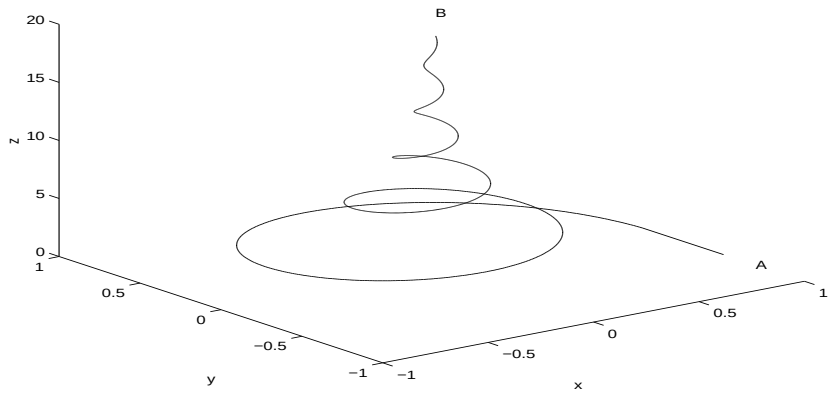

Figura 1.17: Mişcarea pe spirală

1.12.2 Instrucţiunea `ezplot3`

Pentru reprezentarea grafică a curbelor în spaţiu, poate fi folosită instrucţiunea `ezplot3`, care se lansează prin una din comenzile:

```
ezplot3('x','y','z')
ezplot3('x','y','z',[a,b])
ezplot3('x','y','z','animate')
ezplot3('x','y','z',[a,b],'animate')
```

Parametrii `x`, `y` şi `z`, conţin expresii algebrice ale reprezentărilor curbei în funcţie de un parametru `t`, din `[a,b]`. Valoarea implicită a acestui interval este `[0,2`π`]`.

Prezenţa parametrului `animate` produce pe figură un buton cu numele `repeat`, care prin activare prezintă modul de deplasare a unui punct pe curba reprezentată grafic.

De exemplu, dacă se consideră comanda

```
ezplot3('sin(t)','cos(t)','t',[0,10*pi],'animate')
```

se va obţine graficul din Figura 1.18, în care se observă şi butonul `repeat`.

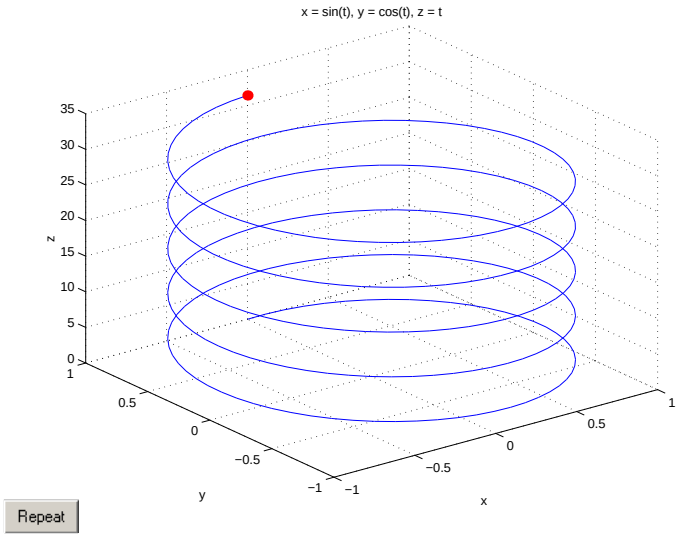

Figura 1.18: $x(t) = \sin(t)$, $y(t) = \cos(t)$, $z(t) = t$, $t \in [0, 10\pi]$

1.12.3 Instrucţiunile `meshgrid`, `mesh` şi `surf`

Reprezentarea grafică a unei suprafeţe, dată prin funcţia $z = f(x, y)$, pe domeniul $\boldsymbol{D} \subset \mathbb{R}^2$, cuprinde două etape.

Prima dată se realizează o reţea rectangulară de puncte, cu ajutorul instrucţiunii `meshgrid`, care are sintaxa

```
[X,Y]=meshgrid(x,y)
```

unde `x` şi `y` sunt vectori, în general de lungimi diferite, `m` şi `n`, care conţin puncte de pe axele `ox` şi `oy` şi care sunt folosite la generarea reţelei rectangulare (x_i, y_j) $i = \overline{1,m}$, $j = \overline{1,n}$.

Rezultatul executării instrucţiunii `meshgrid`, are ca efect generarea matricelor `X` şi `Y` având acelaşi tip `(n,m)`. Fiecare linie a matricei `X` este formată din componentele vectorului `x`, iar fiecare coloană a matricei `Y` este formată din componentele vectorului `y`. În acest fel punctul de coordonate (x_i, y_j) al reţelei rectangulare, se poate exprima, cu ajutorul matricelor `X` şi `Y`, prin perechea `(X(j,i),Y(j,i))`.

Înainte de a trece la a doua etapă, să facem observaţia că dacă `x` coincide cu `y`, atunci se poate utiliza

```
[X,Y]=meshgrid(x)
```

Reprezentarea grafică a funcţiei $z = f(x,y)$, se face cu ajutorul instrucţiunii `mesh`, care are una din formele:

```
mesh(Z)
mesh(Z,C)
mesh(x,y,Z)
mesh(x,y,Z,C)
mesh(X,Y,Z)
mesh(X,Y,Z,C)
```

De la început, să remarcăm faptul că parametrul `C` precizează scara culorilor. Când lipseşte acest parametru, scara culorilor este dată de parametrul `Z`, care reprezintă valorile funcţiei $z = f(x,y)$ pe o reţeaua rectangulară. În acest ultim caz culoarea va fi proporţională cu înălţimea z din punctul (x,y).

De asemenea, trebuie să remarcăm faptul că reprezentarea grafică a suprafeţei este de forma unei reţele curbilinii sprijinită pe înălţimile date prin matricea `Z`.

Parametrii `X`, `Y` şi `Z` sunt matrice de aceleaşi dimensiuni, dacă avem în vedere cum au fost obţinute matricele `X` şi `Y`, atunci toate aceste trei matrice sunt de tipul `(n,m)`.

Parametrii `x` şi `y` sunt vectori de tipul celor ce generează matricele `X` şi `Y` prin instrucţiunea `meshgrid`, adică dacă aceştia au lungimile `m` şi respectiv `n`, matricea `Z` va fi de tipul `(n,m)`, iar punctele pe care se sprijină reţeaua curbilinie, care reprezintă graficul funcţiei, au coordonatele `(x(j),y(i),Z(i,j))`.

Dacă primii doi parametri lipsesc, adică suntem în cazul primelor două forme ale instrucţiunii `mesh`, acestea sunt echivalente respectiv cu următoarele două, prin considerarea valorilor implicite `x=1:m` şi `y=1:n`.

Instrucţiunea `surf` are aceeaşi sintaxă ca instrucţiunea `mesh`, adică

```
surf(Z)
surf(Z,C)
```

```
surf(x,y,Z)
surf(x,y,Z,C)
surf(X,Y,Z)
surf(X,Y,Z,C)
```

unde parametrii au aceleaşi caracteristici. Deosebirea în reprezentarea grafică este că prin instrucţiunea `surf` se produce o colorare a patrulaterelor obţinute prin curbele care generează suprafaţa corespunzătoare funcţiei $z = f(x, y)$.

Programul 1.12.2. Vom scrie un program care să reprezinte grafic funcţia

$$z = f(x, y) = \frac{1}{2\pi} \mathrm{e}^{-\frac{1}{2}\left(x^2+y^2\right)},$$

pe domeniul $\boldsymbol{D} = [-1, 1] \times [-1, 1]$, în două figuri alăturate, odată cu instrucţiunea `mesh` şi apoi cu instrucţiunea `surf`.

```
clear, clf
m=input('m=');
h=1/m; x=-1:h:1;
[X,Y]=meshgrid(x);
Z=1/(2*pi)*exp(-0.5*(X.^2+Y.^2));
subplot(1,2,1), mesh(X,Y,Z)
title('Grafica  prin MESH')
xlabel('x'), ylabel('y'), zlabel('z')
subplot(1,2,2), surf(X,Y,Z)
title('Grafica prin SURF')
xlabel('x'), ylabel('y'), zlabel('z')
```

Executarea programului, cu $m = 6$, conduce la reprezentările grafice din Figura 1.19.

În încheierea acestui paragraf, amintim că există posibilitatea de umbrire a unei suprafeţe cu ajutorul comenzii

```
shading tip
```

unde `tip` poate fi:

- `faceted` (implicit) – umbreşte fiecare patrulater de pe suprafaţa trasată cu o intensitate fixă şi trasează laturile patrulaterelor;
- `flat` – umbreşte fiecare patrulater de pe suprafaţa trasată cu o intensitate fixă, fără trasarea laturilor patrulaterelor;
- `interp` – umbrirea fiecărui patrulater de pe suprafaţa trasată se face în mod gradat, printr–un procedeu de interpolare, care asigură o trecere netedă de la un patrulater la altul, fără trasarea laturilor patrulaterelor.

Programul 1.12.3. Să scriem un program care efectuează umbrirea suprafeţei reprezentată grafic prin Programul 1.12.2:

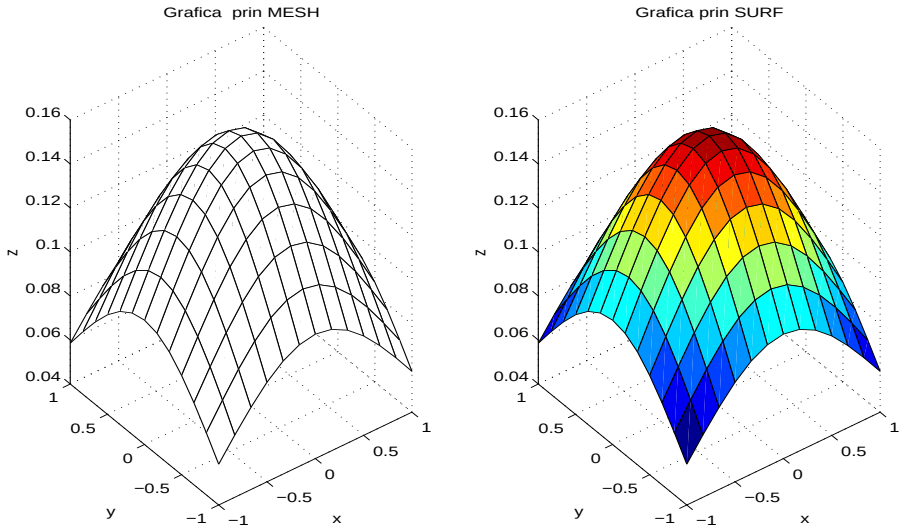

Figura 1.19: $z = \frac{1}{2\pi} e^{-\frac{1}{2}\left(x^2+y^2\right)}$

```
m=input('m=');
h=1/m; x=-1:h:1;
[X,Y]=meshgrid(x);
Z=1/(2*pi)*exp(-0.5*(X.^2+Y.^2))
subplot(1,3,1), surf(X,Y,Z),
shading faceted
title('Umbrire - faceted')
subplot(1,3,2), surf(X,Y,Z),
shading flat
title('Umbrire - flat')
xlabel('x'), ylabel('y'), zlabel('z')
subplot(1,3,3), surf(X,Y,Z),
shading interp
title('Umbrire - interp')
```

Rezultatul executării programului este în Figura 1.20.

1.12.4 Instrucţiunile `contour` şi `contourf`

Reprezentarea curbelor de nivel ale suprafeţei, dată prin funcţia $z = f(x, y)$, se poate face cu una din următoarele forme ale instrucţiunii `contour`:

```
contour(Z)
contour(Z,n)
contour(Z,v)
contour(X,Y,Z)
```

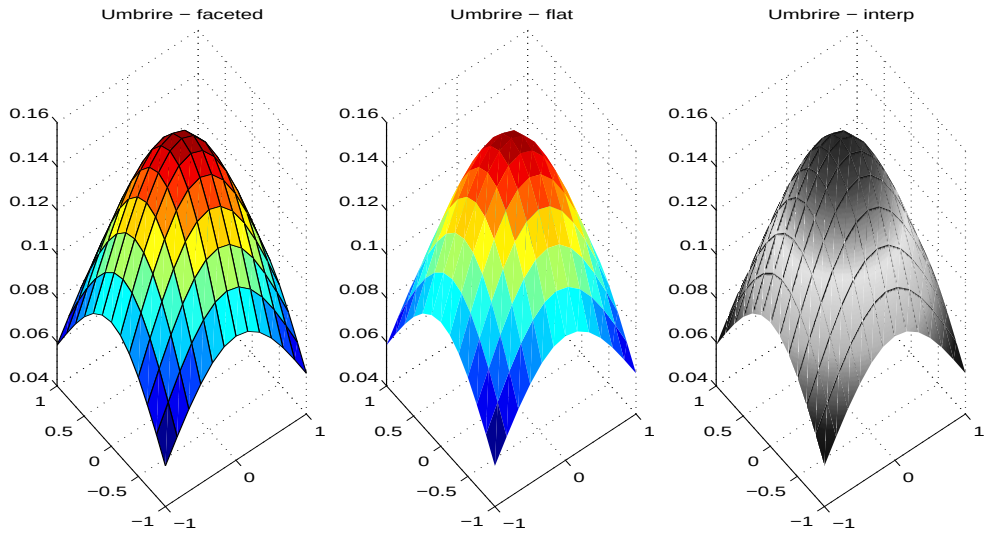

Figura 1.20: Umbrirea unei suprafeţe

```
contour(X,Y,Z,v)
contour(Z,[w w])
contour(X,Y,Z,[w w])
```

unde parametrii `X`, `Y` şi `Z` au aceleaşi interpretări ca şi în cazul instrucţiunilor `mesh` şi `surf`.

Parametrul `v` este un vector prin care se specifică nivelurile curbelor de nivel ce urmează a fi reprezentate grafic, iar în absenţa sa, valorile acestui parametru sunt automat calculate. Numărul curbelor de nivel se poate preciza prin parametrul `n`.

Formele instrucţiunii `contour` cu parametrul de tipul `[w w]` are ca efect trasarea curbei de nivel precizată prin `w`. Aceste forme permit reprezentarea grafică a funcţiilor date sub formă implicită. Astfel, dacă vrem să reprezentăm grafic funcţia $y = y(x)$ dată prin $f(x, y) = 0$, atunci graficul funcţiei $y = y(x)$ va fi dat de curba de nivel pentru funcţia $z = f(x, y)$, considerând `w=0`.

Instrucţiunea `contourf` diferă de `contour` doar prin faptul că ariile delimitate de curbele de nivel sunt umbrite.

Programul 1.12.4. Să reprezentăm `n` curbe de nivel pentru funcţia considerată în Programul 1.12.2. Reprezentările acestor curbe de nivel să fie făcute atât cu instrucţiunea `contour`, cât şi cu `contourf`.

```
clear, clf
m=input('m='); n=input('n=');
h=1/m; x=-1:h:1;
```

```
[X,Y]=meshgrid(x);
Z=1/(2*pi)*exp(-0.5*(X.^2+Y.^2));
subplot(1,2,1), contour(Z,n)
title(['n=',num2str(n),' curbe de nivel'])
subplot(1,2,2), contourf(Z,n)
title(['n=',num2str(n),' curbe de nivel umbrite'])
```

Executarea programului, cu $m = 6$ şi `n=5`, conduce la reprezentările grafice din Figura 1.21.

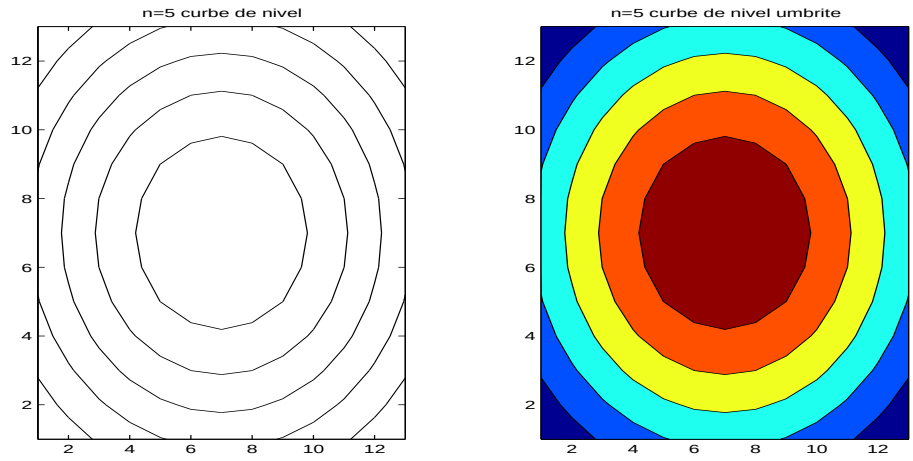

Figura 1.21: Curbe de nivel

1.12.5 Instrucţiunile `ezcontour` şi `ezcontourf`

O formă simplificată, pentru reprezentarea curbelor de nivel, pentru o funcţie dată prin $z = f(x, y)$, se obţine cu ajutorul instrucţiunii `ezcontour`, având una din formele:

```
ezcontour(f)
ezcontour(f,lim)
ezcontour(f,n)
```

unde `f` este expresia matematică a unei funcţii de două variabile, privită ca un şir de caractere.

Parametrul `lim` este fie un vector de forma `[xmin,xmax]`, fie de forma `[xmin,xmax,ymin,ymax]`. Dacă lipseşte, se ia implicit `[-2`π`,2`π`,-2`π`,2`π`]`, iar dacă este de prima formă se ia `[xmin,xmax,xmin,xmax]`.

Parametrul `n` specifică faptul că în reprezentarea grafică se foloseşte o reţea de `n`$\times$`n` puncte (implicit se consideră `n=60`).

De exemplu,

```
f='sqrt(x^2+y^2)'
ezcontour(f,[-2,2])
```

are acelaşi efect cu

```
ezcontour('sqrt(x^2+y^2)',[-2,2])
```

şi produc curbele de nivel pentru funcţia $z = f(x,y) = \sqrt{x^2+y^2}$, pe domeniul $\boldsymbol{D} = [-2,2] \times [-2,2]$.

Instrucţiunea `ezcontourf`, faţă de `ezcontour` execută şi umbrirea ariilor generate de curbele de nivel.

1.12.6 Instrucţiunile `ezmesh`, `ezsurf`, `ezmeshc` şi `ezsurfc`

Sintaxa apelurilor funcţiilor `ezmesh`, `ezsurf`, `ezmeshc` şi `ezsurfc` este aceeaşi, iar efectul executării lor este reprezentarea grafică a unei suprafeţe, respectiv a unei suprafeţe, împreună cu sau fără curbele de nivel corespunzătoare. De aceea prezentăm modurile de apelare, numai pentru una dintre ele.

```
ezmesh('f')
ezmesh('f',lim)
ezmesh('x','y','z')
ezmesh('x','y','z',lim)
```

Parametrul `f` reprezintă o expresie algebrică, ce defineşte suprafaţa, care urmează să fie reprezentată grafic, pe domeniul definit prin `lim`. Implicit, în absenţa parametrului `lim`, domeniul este `[-2`π`,2`π`]`$\times$`[-2`π`,2`π`]`. Dacă `lim=[a,b]`, atunci domeniul va fi `[a,b]`$\times$`[a,b]`, iar dacă `lim=[a,b,c,d]`, atunci domeniul va fi `[a,b]`$\times$`[c,d]`.

Parametrii `x`, `y`, `z`, reprezintă expresii algebrice ce definesc suprafaţa în modul parametric. În acest caz, corespunzător, parametrul `lim`, precizează domeniul în care iau valori parametrii cu ajutorul cărora se defineşte în mod parametric suprafaţa ce urmează a fi reprezentată.

Remarcăm faptul că formele mai sus prezentate, mai pot avea doi parametri. Unul, `n`, de tip întreg pozitiv, care precizează numărul `n`$\times$`n` al punctelor reţelei. Implicit se consideră `n=60`. Un altul este `'circ'`, care prin prezenţa sa atrage reprezentarea lui `f` pe un disc centrat în domeniul mai sus precizat.

Pentru exemplificare, dacă se execută programul Matlab format din următoarele instrucţiuni:

```
subplot(1,2,1)
ezmesh('sin(x)*sin(y)',[-pi,pi],'circ')
subplot(1,2,2)
ezmeshc('sin(x)*sin(y)',[-pi,pi],'circ')
colormap spring
```

se obţin graficele din Figura 1.22.

Mai remarcăm faptul că `f`, `x`, `y`, `z` pot fi numele unor funcţii predefinite sau scrise de utilizator.

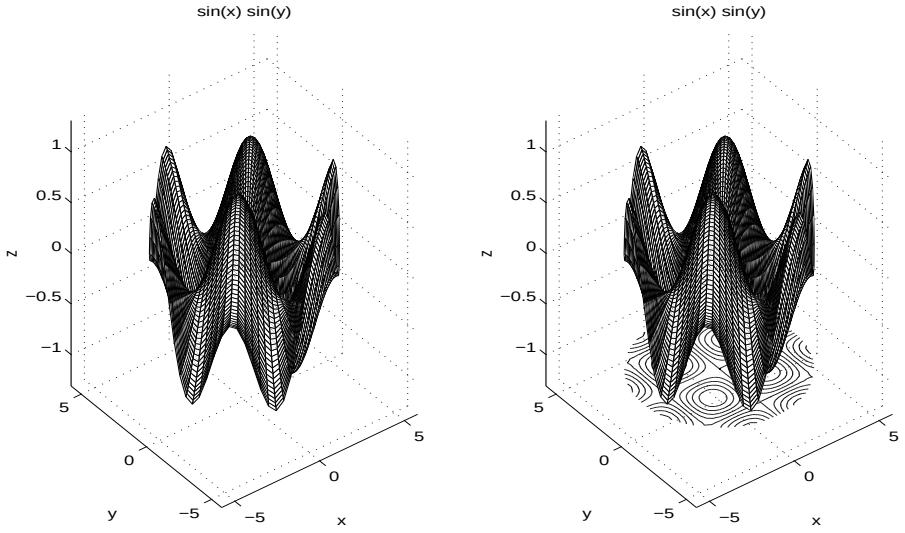

Figura 1.22: Suprafaţă fără şi cu curbe de nivel

Pentru a ilustra că cele două instrucţiuni de tipul `ezmesh` şi `ezsurf` au de regulă acelaşi efect, programul următor reprezintă grafic aceeaşi suprafaţă, respectiv cu `ezmesh` şi `ezsurf`.

```
clf
subplot(1,2,1),
ezsurf('x*exp(-x^2 - y^2)',[-2,2],20)
title('Reprezentare cu ezsurf')
subplot(1,2,2)
ezsurf('x*exp(-x^2 - y^2)',[-2,2],20)
colormap autumn
title('Reprezentare cu ezmesh')
```

Graficele sunt prezentate în Figura 1.23.

1.12.7 Instrucţiunile `bar3` şi `bar3h`

Instrucţiunile `bar3` şi `bar3h` corespund instrucţiunilor `bar` şi respectiv `barh` din grafica bidimensională.

Forme ale acestor instrucţiuni sunt:

```
bar3(z,w,'tip','color')
bar3(y,z,w,'tip','color')
bar3h(y,w,'tip','color')
bar3h(y,z,w,'tip','color')
```

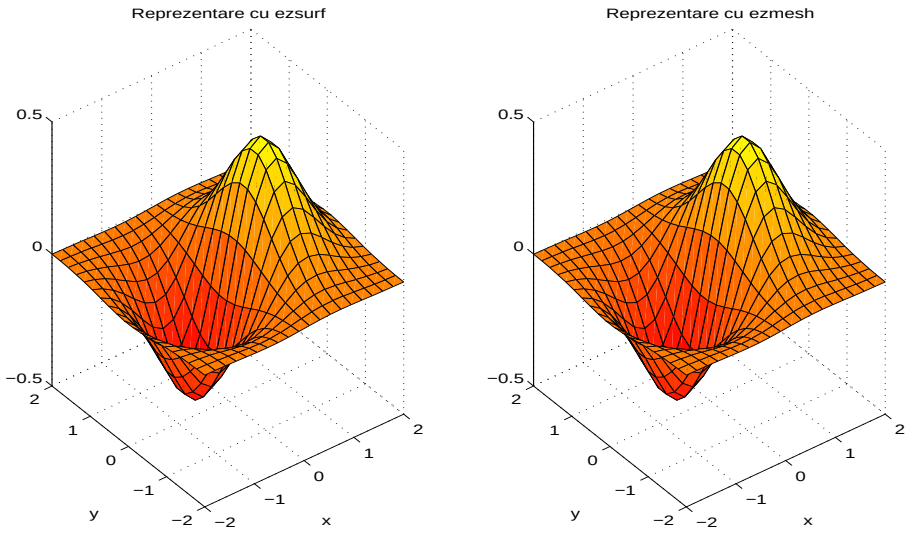

Figura 1.23: Suprafaţă reprezentată $z = f(x, y) = x\mathrm{e}^{-\left(x^2+y^2\right)}$

argumentele `w`, `'tip'` şi `'color'` sunt opţionale, dar când sunt specificate, trebuie să păstreze această ordine.

Parametrul `y` este un vector de lungime `m`, având componentele ordonate crescător sau descrescător. Dacă `y` lipseşte, atunci se consideră implicit `y=1:m`.

Dacă `z` este un vector de aceeaşi lungime cu `y`, atunci vor fi reprezentate grafic `m` bare de lungimi $\mathtt{z}_i$ în dreptul punctelor $\mathtt{y}_i$, $i = \overline{1, \mathtt{m}}$, de pe axa `oy`. Dacă `z` este o matrice de tipul `(m,n)`, reprezentările sunt efectuate pentru fiecare din cele `n` coloane ale matricei, adică în fiecare punct $\mathtt{y}_i$, $i = \overline{1, \mathtt{m}}$, de pe axa `oy`, vor fi reprezentate câte `n` bare (batoane).

Parametrul `w` specifică grosimea barelor, implicit fiind $\mathtt{w} = 0.8$, iar pentru $\mathtt{w} > 1$ barele se suprapun.

Dacă parametrul `tip=detached`, care este valoarea implicită, atunci barele din punctele $\mathtt{y}_i$ de pe axa `oy` vor fi detaşate în adâncime după axa `ox`. Dacă `tip=grouped`, barele (batoanele) vor fi grupate în punctele $\mathtt{y}_i$, de pe axa `oy`, iar dacă `tip=stacked`, atunci barele vor fi stivuite în aceste puncte.

Parametrul `color` specifică culoarea barelor (batoanelor) şi are una din următoarele valori: `r` (roşu), `g` (verde), `b` (albastru), `y` (galben), `m` (violet), `c` (ciclamen), `k` (negru), `w` (alb).

Deosebirea între `bar3` şi `bar3h` este că prima instrucţiune acţionează pe verticală, iar a doua pe orizontală.

Programul 1.12.5. Să scriem un program care generează o matrice magică de ordin $m > 2$. Folosind primele trei coloane ale matricei magice, să reprezentăm prin bare verticale valorile acestora cu fiecare din tipurile `detached`, `grouped` şi `stacked`. Folosind ultimele trei coloane să se facă astfel de reprezentări cu ajutorul barelor orizontale.

```
m=input('m:'); A=magic(m);
subplot(2,3,1), bar3(A(:,1:3),.6,'detached')
title('Bar3 - detached')
subplot(2,3,2), bar3(A(:,1:3),'grouped')
title('Bar3 - grouped')
subplot(2,3,3), bar3(A(:,m-2:m),.6,'stacked')
title('Bar3 - stacked')
subplot(2,3,4), bar3h(A(:,m-2:m),.6,'detached')
title('Bar3h - detached')
subplot(2,3,5), bar3h(A(:,m-2:m),'grouped')
title('Bar3h - grouped')
subplot(2,3,6), bar3h(A(:,1:3),.6,'stacked')
title('Bar3h - stacked')
colormap spring
```

Executarea acestui program, pentru `m=4`, produce graficele din Figura 1.24.

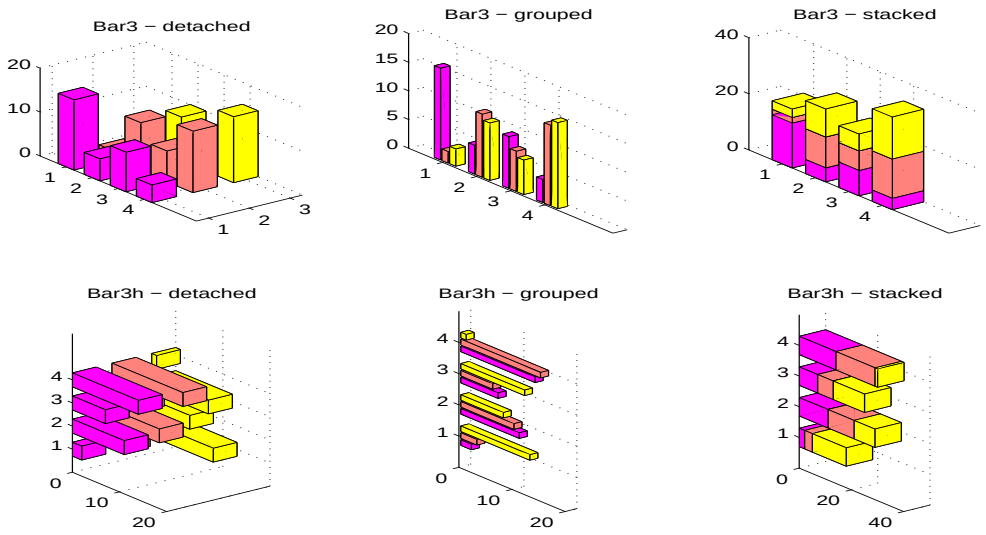

Figura 1.24: Bare verticale şi bare orizontale în 3D

Capitolul 2

Elemente de teoria probabilităţilor

Teoria probabilităţilor are ca obiect de studiu fenomene nedeterministe sau aleatoare (întâmplătoare). Un fenomen nedeterminist $\mathcal{E}$ este rezultatul imposibilităţii cunoaşterii exacte şi complete a tuturor elementelor care concură la realizarea acestuia. Astfel de fenomene aleatoare ar putea fi apelurile dintr-o centrală telefonică, accidentele de circulaţie dintr-o intersecţie, extragerile numerelor loto, rezultatele de la ruletă, noii născuţi la o maternitate, numărul particulelor dezintegrate dintr-o substanţă radioactivă, mişcarea browniană a particulelor unei substanţe dizolvată într-un fluid, sosirea clienţilor într-o staţie de servire, răspândirea erorilor de tipar dintr-o carte, vânzarea pâinii la o unitate alimentară şi în fine răspândirea stafidelor în prăjitura cumpărată de la o cofetărie. Toate aceste fenomene sunt fenomene aleatoare prin faptul că nu sunt cunoscute toate condiţiile în care sunt realizate, sau dacă o parte din aceste condiţii pot fi precizate, ele nu sunt cunoscute cu exactitate. Factorii necunoscuţi, care se regăsesc în fenomenul real, conduc la astfel de fenomene aleatoare (nedeterministe).

Teoria probabilităţilor vine să dea metode unitare pentru analiza şi studiul unor astfel de fenomene nedeterministe şi care provin dintr–o gamă largă a activităţii şi cunoaşterii umane. Pentru aceasta este necesar să fie formalizat limbajul prin care se tratează fenomenele aleatoare, ca apoi prin particularizări să fie obţinute rezultate specifice fenomenului considerat.

Vom presupune că avem anumite cunoştinţe de relative la teoria probabilităţilor. Reamintim în acest capitol doar o parte din noţiunilor de bază ale teoriei probabilităţilor, pentru stabilirea unui limbaj comun în abordarea capitolelor următoare. Ţinând seama de această motivaţie recomandăm celor neiniţiaţi în această disciplină a teoriei probabilităţilor lucrările [10], [6], [9], [8].

2.1 Câmp de probabilitate

Dacă se consideră fenomenul aleator (*experimentul*) $\mathcal{E}$, vom numi *probe* rezultatele experimentului, iar mulţimea probelor relative la experimentul $\mathcal{E}$ o numim *spaţiul probelor* şi pe care îl notăm prin Ω.

Definiţia 2.1.1. *Spunem că submulţimea nevidă* $\mathcal{K} \subset \mathcal{P}(\Omega)$ *este un* σ–corp (corp borelian, σ-algebră) *dacă satisface următoarele condiţii:*

$$\forall A \in \mathcal{K} \Longrightarrow \bar{A} \in \mathcal{K}; \quad \forall A_i \in \mathcal{K},\ i \in I \Longrightarrow \bigcup_{i \in I} A_i \in \mathcal{K}, \tag{2.1.1}$$

iar perechea $(\Omega, \mathcal{K})$ *se numeşte* câmp de evenimente.

Mulţimea de indici I este o mulţime cel mult numărabilă.

Definiţia 2.1.2. *Fiind dat câmpul de evenimente* $(\Omega, \mathcal{K})$*, se numeşte* probabilitate *o aplicaţie* $P\colon \mathcal{K} \longrightarrow \mathbb{R}$*, care verifică următoarele trei axiome:*

(i) $\forall A \in \mathcal{K} \Longrightarrow P(A) \geqslant 0$,

(ii) $P(\Omega) = 1, \quad \Omega$ *fiind* evenimentul sigur (cert),

(iii) $\forall A_i \in \mathcal{K},\ i \in I,\ A_i \cap A_j = \varnothing,\ i \neq j \Longrightarrow P\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} P(A_i)$,

iar tripletul $(\Omega, \mathcal{K}, P)$ *se numeşte* câmp de probabilitate.

2.2 Variabile aleatoare

Fie câmpul de probabilitate $(\Omega, \mathcal{K}, P)$ şi σ–algebrele mulţimilor Borel $\mathcal{B}(\mathbb{R})$, $\mathcal{B}(\mathbb{R}^n)$, respectiv pe $\mathbb{R}$ şi $\mathbb{R}^n$.

Definiţia 2.2.1. *Numim* variabilă aleatoare*, aplicaţia* $X\colon \Omega \longrightarrow \mathbb{R}$*, dacă este* $\mathcal{K}$*–măsurabilă, adică*

$$\forall B \in \mathcal{B}(\mathbb{R}) \Longrightarrow X^{-1}(B) = \left\{\omega \in \Omega \;\middle|\; X(\omega) \in B\right\} \in \mathcal{K}. \tag{2.2.1}$$

Observaţia 2.2.2. Având în vedere că $\mathcal{B}(\mathbb{R})$ poate fi generat de familia de intervale $\{(-\infty, x]\}_{x \in \mathbb{R}}$, condiţia (2.2.1) de $\mathcal{K}$–măsurabilitate poate fi înlocuită prin

$$X^{-1}((-\infty, x]) = (X \leqslant x) = \left\{\omega \in \Omega \;\middle|\; X(\omega) \leqslant x\right\} \in \mathcal{K}, \quad \forall x \in \mathbb{R}. \tag{2.2.2}$$

Observaţia 2.2.3. Cardinalul mulţimii $X(\Omega)$ a valorilor unei variabile aleatoare X, conduce la o clasificarea a mulţimii variabilelor aleatoare:

- *variabile aleatoare de tip discret,* dacă $|X(\Omega)| \leqslant \aleph_0$, adică X are o mulţime cel mult numărabilă de valori;
- *variabile aleatoare simple,* dacă $|X(\Omega)| < \aleph_0$, adică X are o mulţime finită de valori;
- *variabile aleatoare de tip continuu,* dacă $|X(\Omega)| = \aleph$, adică X are o mulţime valori de puterea continuumului.

Definiţia 2.2.4. *Numim* vector aleator n–(dimensional), *aplicaţia* $\boldsymbol{X}\colon \Omega \longrightarrow \mathbb{R}^n$, *dacă este* $\mathcal{K}$*–măsurabilă, adică*

$$\forall \boldsymbol{B} \in \mathcal{B}(\mathbb{R}^n) \Longrightarrow \boldsymbol{X}^{-1}(\boldsymbol{B}) = \left\{\omega \in \Omega \;\middle|\; \boldsymbol{X}(\omega) \in \boldsymbol{B}\right\} \in \mathcal{K}$$

sau având în vedere Observaţia 2.2.2

$$\begin{aligned}(\boldsymbol{X} \leqslant \boldsymbol{x}) &= (X_1 \leqslant x_1, \ldots, X_n \leqslant x_n)\\ &= \left\{\omega \in \Omega \;\middle|\; X_1(\omega) \leqslant x_1, \ldots, X_n(\omega) \leqslant x_n\right\} \in \mathcal{K}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n,\end{aligned}$$

unde X_i *şi* x_i, $i = \overline{1,n}$, *reprezintă respectiv componentele vectorului aleator* $\boldsymbol{X}$ *şi ale vectorului* $\boldsymbol{x}$.

2.3 Funcţie de repartiţie

Fie câmpul de probabilitate $(\Omega, \mathcal{K}, P)$ şi variabila aleatoare $X\colon \Omega \to \mathbb{R}$.

Definiţia 2.3.1. *Numim* funcţie de repartiţie *ataşată variabilei aleatoare* X, *aplicaţia* $F\colon \mathbb{R} \longrightarrow \mathbb{R}$, *definită prin*

$$F(x) = P(X \leqslant x), \quad \forall x \in \mathbb{R}. \tag{2.3.1}$$

Observaţia 2.3.2. Marea majoritate a tratatelor de teoria probabilităţilor definesc funcţia de repartiţie prin formula

$$F(x) = P(X < x), \quad \forall x \in \mathbb{R},$$

dar având în vedere că sistemul Matlab foloseşte formula (2.3.1), în definirea funcţiei de repartiţie, vom considera această abordare în cele ce urmează.

Definiţia 2.3.3. *Numim* funcţie de repartiţie *ataşată vectorului aleator n–dimensional $\boldsymbol{X}$, aplicaţia $F\colon \mathbb{R}^n \longrightarrow \mathbb{R}$, definită prin*

$$F(\boldsymbol{x}) = F(x_1, \ldots, x_n) = P(X_1 \leqslant x_1, \ldots, X_n \leqslant x_n), \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

Fie vectorul aleator $\boldsymbol{X}$ cu funcţia de repartiţie F şi fie notată respectiv prin F_i funcţia de repartiţie a variabilei aleatoare X_i, care este componenta i a vectorului aleator, pentru fiecare $i = \overline{1,n}$.

Definiţia 2.3.4. *Spunem că variabilele aleatoare X_i, $i = \overline{1,n}$, sunt* independente *dacă*

$$F(\boldsymbol{x}) = F(x_1, \ldots, x_n) = F_1(x_1) \ldots F_n(x_n), \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

2.4 Legi de probabilitate de tip discret

Dacă variabila aleatoare X este de tip discret, adică are un număr cel mult numărabil de valori, fie acestea $x_i \in \mathbb{R}$, $i \in I$, atunci funcţia de repartiţie F ataşată este o funcţie în scară şi este dată prin:

$$F(x) = \sum_{\substack{i \in I \\ x_i \leqslant x}} p_i, \quad \forall x \in \mathbb{R}, \tag{2.4.1}$$

unde $p_i = P(X = x_i)$.

Definiţia 2.4.1. *Numim* distribuţia *sau* repartiţia *variabilei aleatoare X de tip discret, tabloul*

$$X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I},$$

unde $x_i \in \mathbb{R}$, $i \in I$, sunt valorile pe care le ia variabila aleatoare X, iar p_i este probabilitatea cu care variabila aleatoare X ia valoarea x_i, adică $p_i = P(X = x_i)$, pentru fiecare $i \in I$.

Deoarece evenimentele $(X = x_i)$, $i \in I$, formează un sistem complet de evenimente, avem că $\sum_{i \in I} p_i = 1$. Remarcăm de asemenea că p_i, reprezintă mărimea saltului funcţiei de repartiţie în punctul de discontinuitate x_i, pentru fiecare $i \in I$.

În teoria probabilităţilor şi în aplicaţiile sale, se întâlnesc clase de variabile aleatoare de tip discret. Forma cea mai generală a unei variabile aleatoare aparţinând unei clase se numeşte *lege de probabilitate de tip discret*.

2.4.1 Funcţiile Matlab `pdf` şi `cdf`

Distribuţia variabilei aleatoare X poate fi precizată prin ceea ce numim *funcţie de probabilitate* (`pdf` – probability distribution function, în Matlab) definită prin:

$$f(x_i) = p_i, \quad i \in I,$$

sau prin *funcţia de repartiţie* (`cdf` – cumulative distribution function).

Dacă valorile p_i, $i \in I$, sunt calculate, atunci folosind instrucţiunile `plot`, `bar` şi `stairs`, se pot reprezenta grafic funcţia de probabilitate (`pdf`) şi funcţia de repartiţie (`cdf`). Anume, dacă vectorul `x` conţine valorile variabilei aleatoare X, iar `p` probabilităţile corespunzătoare, atunci intrucţiunile

```
plot(x,p,'s')
bar(x,p)
stairs(x,p)
```

vor reprezenta grafic respectiv funcţia de probabilitate prin simbolul precizat prin `s`, funcţia de probabilitate prin bare şi funcţia de repartiţie (funcţie în scară).

Remarcăm faptul că dacă X ia o infinitate numărabilă de valori, atunci trebuie să ne limităm la un număr finit de valori ale variabilei aleatoare, iar reprezentările grafice se vor face pe domeniul cuprins între valorile minimă şi maximă ale acestora.

Programul 2.4.2. Să considerăm variabila aleatoare X care are distribuţia

$$X \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \end{pmatrix}$$

şi vrem să reprezentăm grafic funcţia de probabilitate (prin puncte şi bare) şi funcţia de repartiţie pe aceeaşi figură.

Am putea presupune că variabila aleatoare X se referă la aruncarea unui zar. Anume, dacă în urma aruncării zarului se obţine un număr compus (4 sau 6), atunci se pierde o miză ($X = -1$), dacă se obţine un număr prim (2, 3 sau 5) se câştigă o miză ($X = 1$), altfel nu se câştigă şi nu se pierde nimic ($X = 0$).

Următorul program

```
x = [-1:1]; p = [1/3,1/6,1/2];
pc = [1/3,1/2,1];
subplot(1,3,1), plot(x,p,'o')
axis([-1.5 1.5 0 1])
title('Functia de probabilitate')
subplot(1,3,2), bar(x,p)
axis([-1.5 1.5 0 1])
title('Functia de probabilitate')
subplot(1,3,3), stairs(x,pc)
title('Functia de repartitie')
```

produce reprezentările grafice din Figura 2.1.

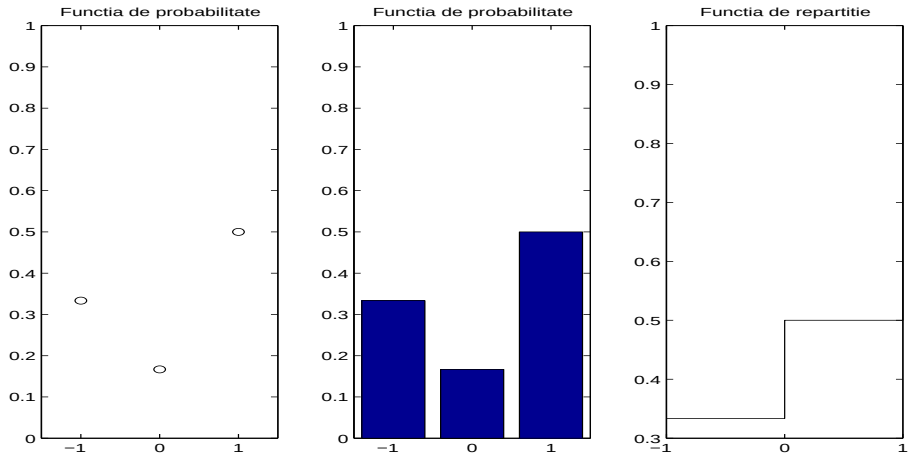

Figura 2.1: Funcţia de frecvenţă şi funcţia de repartiţie

Sistemul Matlab dispune de funcţii (instrucţiuni) care calculează valorile funcţiei de probabilitate şi ale funcţiei de repartiţie pentru legile de probabilitate implementate prin *Statistics toolbox*.

Funcţia `pdf`

Două moduri de apel pentru calculul valorilor funcţiei `pdf` avem:

```
p=pdf('legea',x,par1,par2,...)
p=numef(x,par1,par2,...)
```

unde `legea` este un şir de caractere predefinit pentru fiecare din legile de probabilitate disponibile în *Statistics toolbox*, `numef` este un şir de caractere din care ultimele trei sunt `pdf`, iar cele ce le preced sunt cele care dau numele predefinit al legii de probabilitate (ca şi cele din parametrul `legea`).

În urma executării uneia din cele două instrucţiuni, se calculează matricea `p` a probabilităţilor legii precizată prin parametrii `legea`, respectiv `numef`, corespunzătoare valorilor date prin matricea `x` şi având parametrii daţi prin matricele `par1`, `par2`,... Aceste matrice trebuie să fie de aceleaşi dimensiuni, cu excepţia că dacă unele sunt scalari, aceştia se extind la matricele constante de aceleaşi dimensiuni cu celelalte şi care iau valorile scalarilor corespunzători.

Funcţia `cdf`

Şi pentru calculul valorilor funcţiei de repartiţie funcţia `cdf` avem două forme de apel

```
cp=cdf('legea',x,par1,par2,...)
cp=numef(x,par1,par2,...)
```

unde `legea` este un şir de caractere predefinit pentru fiecare din legile de probabilitate disponibile în *Statistics toolbox*, `numef` este un şir de caractere din care ultimele trei sunt `cdf`, iar cele ce le preced sunt cele care dau numele predefinit al legii de probabilitate (ca şi cele din parametrul `legea`).

În urma executării uneia din cele două instrucţiuni, se calculează matricea `cp` a probabilităţilor cumulate (valorile funcţiei de repartiţie) a legii precizată prin `legea` respectiv `numef`, corespunzătoare valorilor date prin matricea `x` şi având parametrii daţi prin matricele `par1`, `par2`,... Aceste matrice trebuie să fie de aceleaşi dimensiuni, cu excepţia că dacă unele sunt scalari, aceştia se extind la matricele constante de aceleaşi dimensiuni cu celelalte şi care iau valorile scalarilor corespunzători.

2.4.2 Legi de probabilitate de tip discret clasice

Vom prezenta în continuare legile de probabilitate de tip discret recunoscute de sistemul Matlab, prin pachetul de programe *Statistics toolbox*. În paranteză, pentru fiecare lege de probabilitate sunt trecute şi denumirile acceptate de sistemul Matlab. Facem observaţia că pentru unele funcţii sunt acceptate şi alte denumiri.

Legea lui Bernoulli

Variabila aleatoare X urmează *legea lui Bernoulli*, pe care o notăm $\mathcal{B}e\,(p)$, dacă are distribuţia

$$X \begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix}, \quad \text{unde } p \in (0,1),\ q = 1-p. \tag{2.4.2}$$

Legea lui Bernoulli modelează un fenomen aleator în desfăşurarea căruia suntem interesaţi în apariţia unui eveniment fixat A, pe care îl vom numi *succes*, respectiv a lui $\bar{A}$ numit *insucces*. Probabilitatea obţinerii succesului este p, iar valoarea corespunzătoare a variabilei aleatoare X este `1`, respectiv $q = 1-p$ este probabilitatea producerii insuccesului, cu valoarea `0` corespunzătoare pentru X.

Funcţia de probabilitate este

$$f\,(x \,|\, p) = p^x q^{1-x}, \quad x = 0, 1.$$

Legea uniformă discretă (`unid`)

Variabila aleatoare X urmează *legea uniformă discretă*, notăm $\mathcal{U}(N)$, dacă are distribuţia

$$X\begin{pmatrix} k \\ \frac{1}{N} \end{pmatrix}_{k=\overline{1,N}}, \quad \text{unde } N \in \mathbb{N}. \tag{2.4.3}$$

Funcţia de probabilitate este

$$f(x \mid N) = \frac{1}{N}, \quad x = \overline{1, N}.$$

Legea binomială (`bino`)

Spunem că variabila aleatoare X urmează *legea binomială*, notăm $\mathcal{B}(n, p)$, dacă are distribuţia

$$X\begin{pmatrix} k \\ P(n,k) \end{pmatrix}_{k=\overline{0,n}}, \quad \text{unde} \quad P(n,k) = \binom{n}{k} p^k q^{n-k}, \tag{2.4.4}$$

iar $p \in (0,1)$ şi $q = 1 - p$.

Variabila aleatoare X reprezintă numărul succeselor obţinute în n repetări independente ale unui experiment.

Descoperirea legii binomială este atribuită lui James Bernoulli şi care se află în cartea *Ars Conjectandi* (1713). Pascal a considerat cazul particular $p = \frac{1}{2}$.

Funcţia de probabilitate este

$$f(x \mid n, p) = \binom{n}{x} p^x q^{n-x}, \quad x = \overline{0, n}.$$

Remarcăm faptul că variabila aleatoare X se obţine ca suma a n variabile aleatoare independente, ce urmează fiecare legea lui Bernoulli $\mathcal{B}e(p)$. Prin urmare, putem spune şi faptul că legea lui Bernoulli $\mathcal{B}e(p)$ se obţine din legea binomială $\mathcal{B}(n, p)$, pentru $n = 1$.

Programul 2.4.3. Să scriem un program Matlab care să reprezinte grafic funcţia de probabilitate (prin puncte şi prin bare) şi funcţia de repartiţie ale legii de probabilitate $\mathcal{B}(n, p)$.

Executarea programului

```
n = input('n=');
p = input('p='); x = 0:n;
f = pdf('bino',x,n,p);
subplot(1,3,1), plot(x,f,'o')
axis([-0.5 n+0.5 0 max(p)+0.02])
title('Functia de probabilitate')
subplot(1,3,2), bar(x,f)
axis([-.5 n+0.5 0 max(p)+0.02])
title('Functia de probabilitate')
f = cdf('bino',x,n,p);
subplot(1,3,3), stairs(x,f)
title('Functia de repartitie')
axis([0 n 0 1])
```

pentru `n=7` şi `p=0.3`, realizează graficele din Figura 2.2.

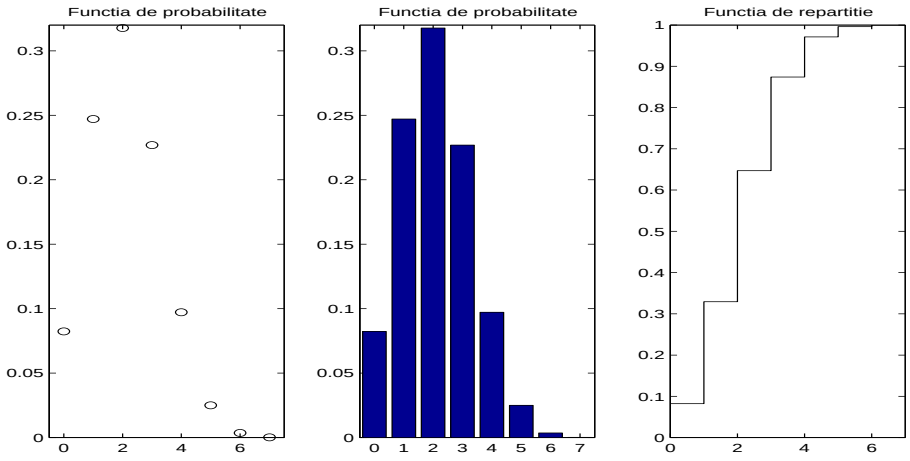

Figura 2.2: Legea binomială $\mathcal{B}(7, 0.3)$

Legea hipergeometrică (`hyge`)

Spunem că variabila aleatoare X urmează *legea hipergeometrică* şi vom nota prin $\mathcal{H}(n, M, K)$, dacă are distribuţia

$$X \begin{pmatrix} k \\ P(n,k) \end{pmatrix}_{k=\overline{0,n}}, \text{ unde } P(n,k) = \frac{\binom{K}{k}\binom{M-K}{n-k}}{\binom{M}{n}}, \quad \text{iar } n \leqslant K \leqslant M. \tag{2.4.5}$$

Variabila aleatoare X reprezintă numărul succeselor obţinute în n extrageri dintr–o populaţie de volum M, fără întoarcere, dacă numărul indivizilor cu proprietatea cercetată este K.

Funcţia de probabilitate corespunzătoare este:

$$f\left(x \mid n, M, K\right) = \frac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}}, \quad x = \overline{0, n}.$$

Observaţia 2.4.4. Dacă se notează $p = \frac{K}{M}$ şi $q = \frac{M-K}{M}$, adică probabilităţile ca la prima extragere să se obţină succes, respectiv insucces, iar $M \to \infty$, atunci

$$P\left(n, k\right) \to \binom{n}{k} p^k q^{n-k},$$

adică se obţine distribuţia binomială.

Programul 2.4.5. Pentru a ilustra afirmaţia precedentă, să scriem un program Matlab care să reprezinte grafic prin bare funcţiile de probabilitate pentru $\mathcal{H}\left(n, M, K\right)$ şi $\mathcal{B}\left(n, p\right)$, cu $p = \frac{K}{M}$.

Executarea programului

```
clf;
M = input('M:');
K = input('K(K<=M):');
n = input('n(n<=K):')
x = 0:n; p=K/M;
fh = pdf('hyge',x,M,K,n);
fb = pdf('bino',x,n,p);
bar(x',[fh',fb'])
colormap winter
```

pentru `M=100`, `K=40` şi `n=10`, realizează graficul din Figura 2.3.

Legea lui Poisson (`poiss`)

Variabila aleatoare X urmează *legea lui Poisson*, vom nota $\mathcal{P}o\left(\lambda\right)$, dacă are distribuţia

$$X \begin{pmatrix} k \\ p_k\left(\lambda\right) \end{pmatrix}_{k=0,1,2,\ldots}, \text{ unde } \quad p_k\left(\lambda\right) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}, \text{ iar } \lambda > 0. \tag{2.4.6}$$

Funcţia de probabilitate corespunzătoare este

$$f\left(x \mid \lambda\right) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}, \quad x = 0, 1, \ldots$$

Variabila aleatoare X, care urmează legea lui Poisson, numără de câte ori apare un anumit eveniment într–un interval de tip, pe o distanţă, pe o suprafaţă etc. Poisson

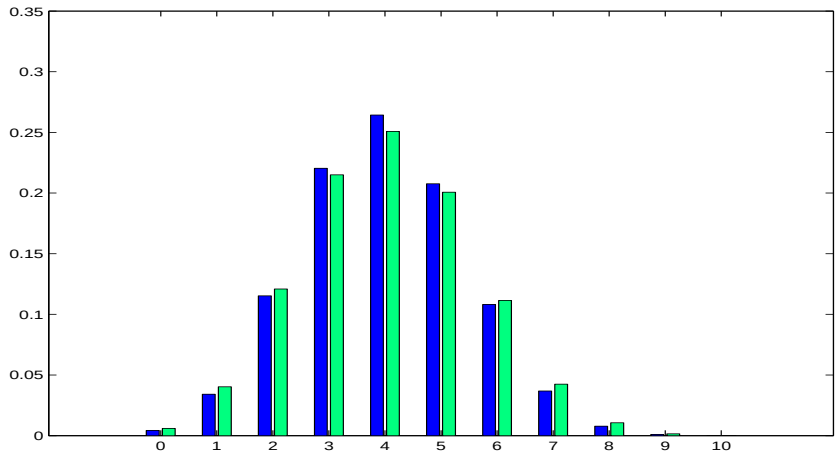

Figura 2.3: Legea hipergeometrică $\mathcal{H}(10, 100, 40)$ şi legea binomială $\mathcal{B}(10, 0.4)$

(1837) a arătat că legea, care–i poartă numele, este un caz limită a legii binomiale, anume când $np \to \lambda$, pentru $n \to \infty$. Altfel spus, când n este mare, p trebuie să fie mic, adică probabilitatea producerii evenimentului considerat este mică, de aici denumirea de *lege a evenimentelor rare*.

Mai remarcăm un rezultat cunoscut, care stabileşte legătura dintre legea lui Poisson şi *legea exponenţială*, care va fi prezentată mai încolo. Anume, pe când legea lui Poisson numără evenimentele ce apar într–un interval de timp, legea exponenţială dă lungimea intervalului dintre două apariţii consecutive ale evenimentelor.

Programul 2.4.6. Vom scrie un program Matlab, care să reprezinte grafic prin bare funcţiile de probabilitate pentru $\mathcal{B}(n, p)$ şi $\mathcal{P}o(\lambda)$ cu $\lambda = np$.

Având în vedere comportarea legii $\mathcal{B}(n, p)$, când $n \to \infty$, graficele funcţiilor de probabilitate pentru cele două legi de probabilitate trebuie să se apropie.

Mai remarcăm faptul că legea $\mathcal{P}o(\lambda)$ are o infinitate (numărabilă) de valori, prin urmare în reprezentarea grafică pentru această lege de probabilitate ne vom limita la un număr finit de valori. Anume, vom considera numai valorile întregi din intervalul $\left[\lambda - 3\sqrt{\lambda}, \lambda + 3\sqrt{\lambda}\right]$, alegere care va fi justificată mai târziu.

Programul

```
clf;
n = input('n='); p = input('p=');
lambda = n*p;
vi = fix(lambda-3*sqrt(lambda));
vf = fix(lambda+3*sqrt(lambda));
x = vi:vf;
```

```
fb = binopdf(x,n,p);
fl = poisspdf(x,lambda);
colormap autumn
bar(x',[fb',fl'])
```

pentru `n=100` şi `p=0.05`, realizează graficul din Figura 2.4.

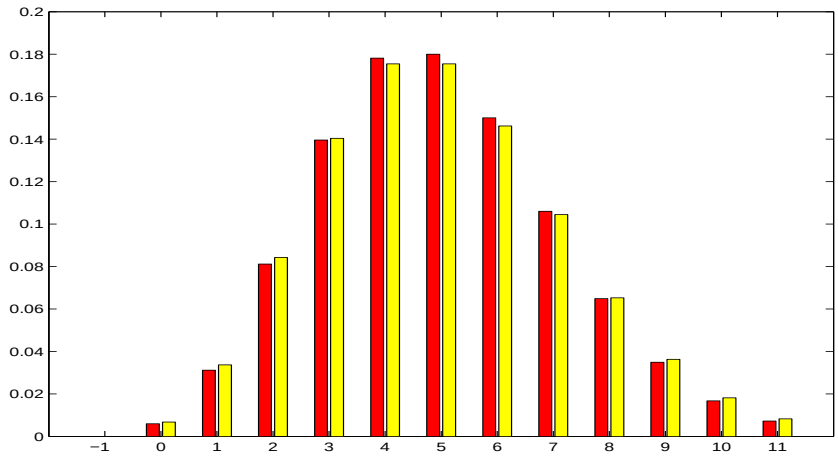

Figura 2.4: Legea binomială $\mathcal{B}(100, 0.05)$ şi legea Poisson $\mathcal{P}o(5)$

Legea binomială negativă (`nbin`)

Spunem că variabila aleatoare X urmează *legea binomială negativă* sau *legea lui Pascal*, notată $\mathcal{BN}(r, p)$, dacă are distribuţia

$$X \begin{pmatrix} k \\ P(r,k) \end{pmatrix}_{k=0,1,2,\ldots}, \text{ cu } \quad P(r,k) = \binom{r+k-1}{k} p^r q^k,$$

iar $p \in (0,1)$ şi $q = 1 - p$. Variabila aleatoare X ce urmează legea $\mathcal{BN}(r, p)$ reprezintă numărul insucceselor până la apariţia succesului de rang r, dacă repetările sunt considerate independente.

Funcţia de probabilitate corespunzătoare legii $\mathcal{BN}(r, p)$ este

$$f(x \,|\, r, p) = \binom{r+x-1}{k} p^r q^x, \quad x = 0, 1, \ldots$$

Observaţia 2.4.7. Unele tratate consideră că o variabilă aleatoare X ce urmează legea lui Pascal reprezintă rangul repetării la care se produce succesul cu numărul r.

Dacă se are în vedere această definiţie se obţine distribuţia

$$X\begin{pmatrix} k \\ P(r,k) \end{pmatrix}_{k=1,2,\ldots}, \text{ cu } \quad P(r,k) = \binom{k-1}{r-1} p^r q^{k-r}. \tag{2.4.7}$$

Legea geometrică (geo)

Legea geometrică, notată $\mathcal{G}e(p)$, se obţine ca şi caz particular, $r = 1$, al legii $\mathcal{BN}(r,p)$. Prin urmare variabila aleatoare X urmează legea $\mathcal{G}e(p)$, dacă are distribuţia

$$X\begin{pmatrix} k \\ pq^k \end{pmatrix}_{k=0,1,2,\ldots}, \quad p \in (0,1),\ p+q=1,$$

şi reprezintă numărul insucceselor până la apariţia primului succes.

Funcţia de probabilitate pentru legea $\mathcal{G}e(p)$ este

$$f(x \mid p) = pq^x, \quad x = 0, 1, \ldots$$

De remarcat faptul că dacă se adună r variabile aleatoare independente, fiecare urmând legea $\mathcal{G}e(p)$, atunci se obţine o variabilă aleatoare ce urmează legea $\mathcal{BN}(r,p)$.

Observaţia 2.4.8. Corespunzător observaţiei de la legea lui Pascal, avem că unele tratate consideră că o variabilă aleatoare X ce urmează legea geometrică reprezintă rangul repetării la care se produce primul succes. Dacă se are în vedere această definiţie se obţine distribuţia

$$X\begin{pmatrix} k \\ pq^{k-1} \end{pmatrix}_{k=1,2,\ldots} \tag{2.4.8}$$

2.5 Legi de probabilitate continue

Definiţia 2.5.1. *Fie variabila aleatoare X având funcţia de repartiţie F. Vom spune că X este* variabilă aleatoare (absolut) continuă, *dacă funcţia de repartiţie F este absolut continuă, sau echivalent se poate reprezenta sub forma*

$$F(x) = \int_{-\infty}^{x} f(t)\,dt, \quad \textit{pentru orice } x \in \mathbb{R},$$

funcţia $f\colon \mathbb{R} \longrightarrow \mathbb{R}$, numindu–se densitatea de probabilitate *a variabilei aleatoare X.*

Definiţia 2.5.2. *Fie vectorul aleator $\boldsymbol{X}$ având funcţia de repartiţie F. Spunem că $\boldsymbol{X}$ este* vector aleator de tip continuu, *dacă funcţia de repartiţie F se poate reprezenta sub forma*

$$F(\boldsymbol{x}) = F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_n} f(t_1, \ldots, t_n)\, dt_1 \ldots dt_n, \quad \forall \boldsymbol{x} \in \mathbb{R}^n,$$

funcţia $f \colon \mathbb{R}^n \longrightarrow \mathbb{R}$, numindu–se densitatea de probabilitate *a vectorului aleator $\boldsymbol{X}$.*

Ca la variabilele aleatoare de tip discret şi la cele de tip continuu există clase de variabile aleatoare. Forma cea mai generală a densităţii de probabilitate a unei variabile aleatoare din clasa respectivă, ne dă o *lege de probabilitate de tip continuu.*

Funcţiile Matlab `pdf` şi `cdf`

Distribuţia variabilei aleatoare X de tip continuu poate fi precizată prin *funcţie densitate de probabilitate* (`pdf` – `p`robability `d`ensity `f`unction) sau prin *funcţia de repartiţie* (`cdf` – `c`umulative `d`istribution `f`unction).

Reprezentările grafice ale densităţii de probabilitate şi funcţiei de repartiţie în Matlab se fac folosind instrucţiunea de tip `plot`.

Valorile densităţii de probabilitate f şi ale funcţiei de repartiţie F, pentru legile de probabilitate implementate prin *Statistics toolbox*, ca şi în cazul discret, se obţin cu ajutorul funcţiilor Matlab `pdf` şi respectiv `cdf`. Trebuie însă să ţinem seama că f este continuă (eventual cu o mulţime cel mult numărabilă de puncte de discontinuitate), iar F nu mai este o funcţie în scară. Prin urmare, pentru asigurarea netezimii graficelor se impune să fie considerat un număr suficient de puncte ale graficelor acestor funcţii.

2.5.1 Legi de probabilitate continue clasice

Prezentăm şi în acest caz legile de probabilitate de tip continue recunoscute de sistemul Matlab, prin pachetul de programe *Statistics toolbox.*

Legea uniformă (`unif`)

Spunem că variabila aleatoare X urmează *legea uniformă* pe intervalul $[a, b]$, şi vom nota $\mathcal{U}(a, b)$, dacă are densitatea de probabilitate

$$f(x \mid a, b) = \begin{cases} \dfrac{1}{b-a}, & \text{dacă } x \in [a, b], \\ 0, & \text{dacă } x \notin [a, b]. \end{cases}$$

Funcţia de repartiţie pentru variabila aleatoare X, ce urmează legea uniformă pe intervalul $[a, b]$, este

$$F(x \mid a, b) = \int_{-\infty}^{x} f(t \mid a, b)\, dt = \begin{cases} 0, & \text{dacă } x < a, \\ \dfrac{x-a}{b-a}, & \text{dacă } a \leqslant x \leqslant b, \\ 1, & \text{dacă } x > b. \end{cases} \tag{2.5.1}$$

Denumirea de lege uniformă este legată de faptul că dacă se consideră subintervale ale intervalului $[a, b]$ de lungimi egale, să zicem că au lungimea ℓ, atunci probabilitatea ca variabila aleatoare X să ia valori într–un astfel de interval este $\ell/(b-a)$. Adică, această probabilitate nu depinde de poziţia subintervalului, ci numai de lungimea lui. Intuitiv, am putea spune că valorile întâmplătoare (aleatoare) ale funcţiei X sunt "uniform" răspândite pe intervalul $[a, b]$. Acest lucru nu se întâmplă la alte legi de probabilitate, când aceste valori sunt "grupate" în jurul uneia sau mai multor valori pe care le ia variabila aleatoare X.

Programul 2.5.3. Să reprezentăm grafic densitatea de probabilitate f şi funcţia de repartiţie F pentru legea $\mathcal{U}(a, b)$.

Programul Matlab care urmează

```
clf;
a = input('a='); b = input('b=');
x = a-1:0.01:b+1;
f = pdf('unif',x,a,b);
F = cdf('unif',x,a,b);
subplot(1,2,1), plot(x,f,'k-')
axis([a-1 b+1 -0.1 1/(b-a)+0.1])
title('Densitatea de probabilitate')
subplot(1,2,2), plot(x,F,'k-')
axis([a-1 b+1 -0.1 1.1])
title('Functia de repartitie')
```

pentru `a=1` şi `b=4`, realizează graficele din Figura 2.5.

Putem remarca faptul că funcţia de repartiţie este continuă. Astfel valorile argumentului $x = a$ şi $x = b$ în formula (2.5.1) pot fi ataşate respectiv la cazurile $F(x) = 0$ şi $F(x) = 1$. Prin urmare intervalul pe care densitatea de probabilitate este diferită de zero poate fi interval închis la ambele capete, închis la un capăt şi deschis la celălalt, respectiv deschis la ambele capete. În toate aceste cazuri vom considera că avem legea uniformă de probabilitate $\mathcal{U}(a, b)$.

Proprietatea 2.5.4. *Dacă variabila aleatoare U urmează legea uniformă pe intervalul $[0, 1]$, adică $\mathcal{U}(0, 1)$, numită şi* lege uniformă standard, *şi dacă $a < b$, atunci variabila aleatoare*

$$X = (b-a)\, U + a,$$

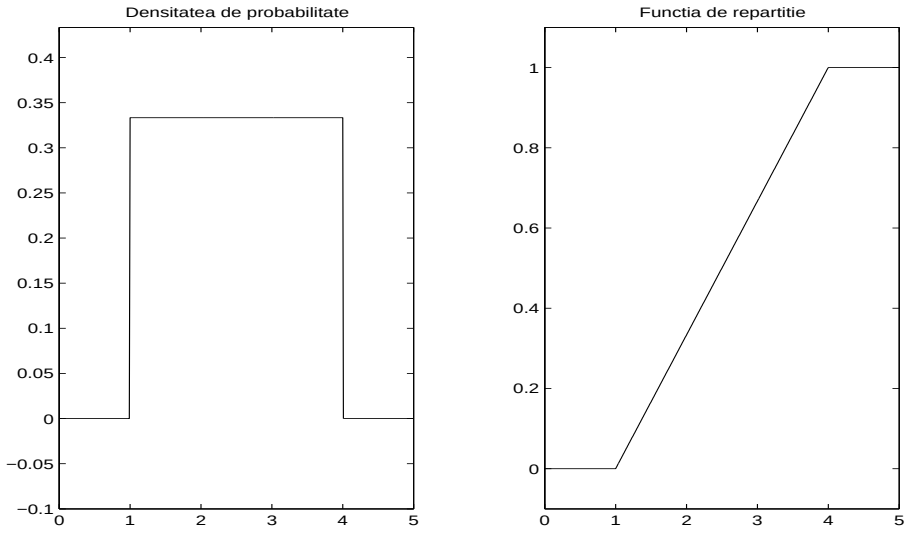

Figura 2.5: Legea uniformă pe $[1, 4]$

urmează legea uniformă $\mathcal{U}(a, b)$.

Legea normală (`norm`)

Spunem că variabila aleatoare X urmează *legea normală* (*legea Gauss–Laplace*) de parametri $\mu \in \mathbb{R}$ şi $\sigma > 0$, notăm aceasta prin $\mathcal{N}(m, \sigma)$, dacă are densitatea de probabilitate

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{pentru orice } x \in \mathbb{R}.$$

Funcţia de repartiţie pentru *legea normală standard* $\mathcal{N}(0, 1)$ este

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{t^2}{2}} dt, \quad x \in \mathbb{R},$$

şi se numeşte *funcţia lui Laplace*.

În Matlab (şi nu numai) există o funcţie `erf` (*funcţia eroare*):

$$\texttt{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} \mathrm{e}^{-t^2} dt,$$

care verifică relaţia

$$\texttt{erf}(x) = 2\Phi\left(x\sqrt{2}\right) - 1.$$

Remarcăm că funcţia

$$\tilde{\Phi}(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \mathrm{e}^{-\frac{t^2}{2}} dt, \quad x \in \mathbb{R},$$

se numeşte de asemenea *funcţia lui Laplace*. Între cele două funcţii Laplace există relaţia

$$\Phi(x) = \frac{1}{2} + \tilde{\Phi}(x).$$

Folosind cele două funcţii ale lui Laplace avem:

$$\begin{aligned} F(x \mid \mu, \sigma) &= \int_{-\infty}^{x} f(t \mid \mu, \sigma)\, dt = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2} + \tilde{\Phi}\left(\frac{x-\mu}{\sigma}\right). \end{aligned}$$

Programul 2.5.5. Vom reprezenta grafic densitatea de probabilitate f pentru legea $\mathcal{N}(\mu, \sigma)$ şi funcţia de repartiţie F corespunzătoare.

Execuţia programului

```
clf;
m = input('mu='); s = input('sigma=');
x = m-3*s:0.01:m+3*s;
f = pdf('norm',x,m,s);
F = cdf('norm',x,m,s);
subplot(1,2,1), plot(x,f,'k-')
title('Densitatea de probabilitate')
subplot(1,2,2), plot(x,F,'k-')
title('Functia de repartitie')
```

pentru `mu=0` şi `sigma=1`, produce graficele din Figura 2.6.

Observaţia 2.5.6. Dacă se consideră variabilele aleatoare X_k, $k = \overline{1,n}$, independente, fiecare urmând respectiv legile normale $\mathcal{N}(\mu_k, \sigma_k)$, atunci combinaţia liniară

$$Z = \sum_{k=1}^{n} a_k X_k, \quad a_k \in \mathbb{R}, \quad \sum_{k=1}^{n} a_k^2 > 0,$$

este o variabilă aleatoare ce urmează legea normală $\mathcal{N}(\mu, \sigma)$, unde

$$\mu = \sum_{k=1}^{n} a_k \mu_k, \quad \sigma^2 = \sum_{k=1}^{n} a_k^2 \sigma_k^2.$$

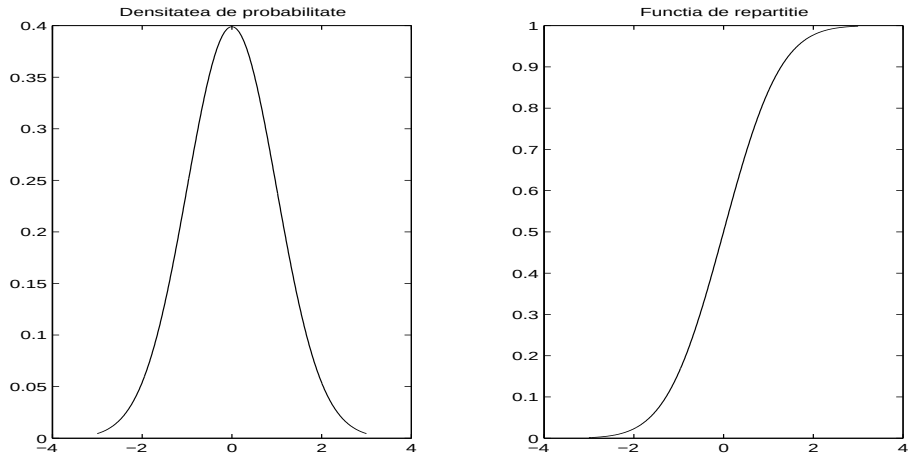

Figura 2.6: Legea normală standard

Legea lognormală (`logn`)

Variabila aleatoare X urmează *legea lognormală de probabilitate* şi vom nota aceasta prin $\mathcal{LN}(\mu, \sigma)$, dacă are densitatea de probabilitate

$$f(x \mid \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \quad x > 0; \quad \mu \in \mathbb{R}, \quad \sigma > 0.$$

Dacă X urmează legea lognormală de parametri μ şi σ, atunci variabila aleatoare $Y = \ln X$ urmează legea normală ($X = \mathrm{e}^Y$).

Legea gamma (`gam`)

Variabila aleatoare X urmează *legea gamma*, notăm aceasta prin $\mathcal{G}a(a, b)$, dacă are densitatea de probabilitate

$$f(x \mid a, b) = \frac{1}{b^a \Gamma(a)} x^{a-1} \mathrm{e}^{-\frac{x}{b}}, \quad x > 0; \quad a, b > 0,$$

unde $\Gamma(a)$ este *funcţia lui Euler de speţa a doua*, adică

$$\Gamma(a) = \int_0^\infty x^{a-1} \mathrm{e}^{-x} dx, \quad a > 0.$$

Când $a \to \infty$, legea gamma tinde la legea normală. Adică, dacă X urmează legea $\mathcal{G}a(a, b)$, atunci pentru $a \to \infty$, variabila aleatoare poate fi considerată ca urmând legea $\mathcal{N}(\mu, \sigma)$, cu $\mu = ab$ şi $\sigma = b\sqrt{a}$.

Programul 2.5.7. Pentru a ilustra această ultimă afirmaţie, să reprezentăm grafic, pe aceeaşi figură, densităţile de probabilitate pentru cele două legi, parametrii acestora satisfăcând relaţiile precizate mai sus.

Executând programul

```
clf;
a = input('a='); b = input('b=');
m = a*b; s = b*sqrt(a);
x = m-3*s:0.01:m+3*s;
fn = pdf('norm',x,m,s);
fg = pdf('gam',x,a,b);
plot(x,fn,'k-.',x,fg,'k-')
legend('Legea normala','Legea gamma')
```

pentru `a=100` şi `b=10`, acesta produce graficele din Figura 2.7.

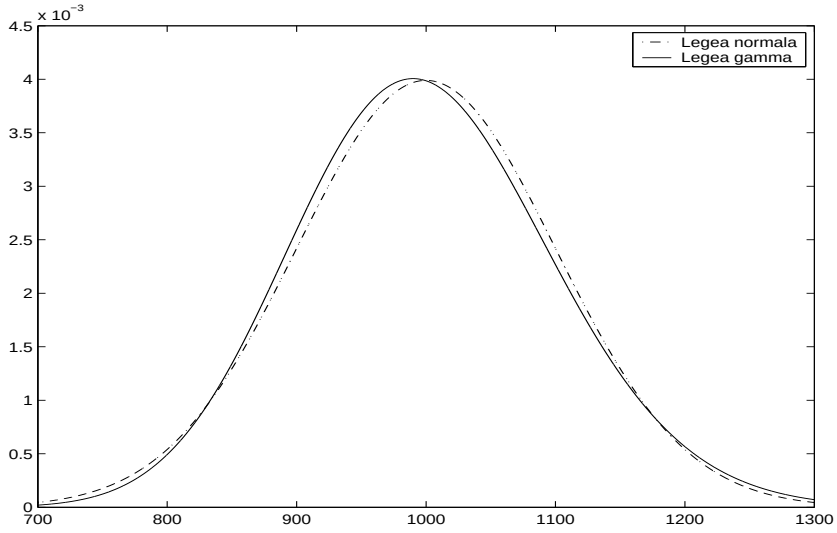

Figura 2.7: Legile $\mathcal{G}a\,(100, 10)$ şi $\mathcal{N}\,(1000, 100)$

Legea exponenţială (`exp`)

Variabila aleatoare X urmează *legea exponenţială*, notăm $\mathcal{E}xp\,(\mu)$, dacă are densitatea de probabilitate

$$f\,(x\,|\,\mu) = \frac{1}{\mu}\mathrm{e}^{-\frac{x}{\mu}}, \quad x > 0; \quad \mu > 0.$$

Observăm că se obţine ca şi caz particular al legii $\mathcal{G}a\,(a, b)$, când se ia $a = 1$ şi $b = \mu$.

Legea exponenţială se aplică la modelarea evenimentelor ce apar aleator în timp. De asemenea are aplicaţii în studiile privind durata de viaţă.

Legea beta (`beta`)

Variabila aleatoare X urmează *legea beta*, notată $\mathcal{B}eta\,(a,b)$, dacă are densitatea de probabilitate:

$$f\,(x\,|\,a,b) = \frac{1}{B\,(a,b)} x^{a-1}\,(1-x)^{b-1}\,, \quad x \in (0,1)\,; \quad a,b>0,$$

unde

$$B\,(a,b) = \int_0^1 x^{a-1}\,(1-x)^{b-1}\,,$$

este *funcţia lui Euler de speţa întâi*.

Este cunoscut faptul că dacă două variabile aleatoare independente U şi V urmează legile $\mathcal{G}a\,(a,1)$ şi $\mathcal{G}a\,(b,1)$, atunci variabila aleatoare $X = U/(U+V)$ urmează legea $\mathcal{B}eta\,(a,b)$.

Dacă se consideră legea $\mathcal{B}eta\,(a,b)$, cu $a = b = \frac{1}{2}$, se obţine *legea de probabilitate arcsin*, care are densitatea de probabilitate

$$f\,(x) = \frac{1}{\pi\sqrt{(x\,(1-x))}}, \quad x \in (0,1)\,.$$

Legea Weibull (`weib`)

Variabila aleatoare X urmează *legea lui Weibull*, notăm $\mathcal{W}\,(a,b)$, dacă are densitatea de probabilitate

$$f\,(x\,|\,a,b) = abx^{b-1}\mathrm{e}^{-ax^b}, \quad x>0; \quad a,b>0.$$

Legea a fost descoperită de Weibull (1939) şi modelează rezistenţa la rupere a materialelor. De asemenea se pare că modelează durata de viaţă mai bine decât legea exponenţială. Se observă că legea exponenţială este un caz particular, anume se obţine când $a = \frac{1}{\mu}$ şi $b = 1$.

Legea Rayleigh (`rayl`)

Variabila aleatoare X urmează *legea lui Rayleigh de probabilitate*, notată prin $\mathcal{R}\,(b)$, dacă are densitatea de probabilitate:

$$f\,(x\,|\,b) = \frac{x}{b^2}\mathrm{e}^{-\frac{x^2}{2b^2}}, \quad x>0; \quad b>0.$$

Se observă că este caz particular al legii Weibull, $b = 2$, $a := 1/\left(2b^2\right)$.

Dacă viteza unei particule în direcţiile x şi y sunt două variabile aleatoare independente, care urmează legi normale de probabilitate, cu mediile zero şi dispersiile egale, atunci distanţa parcursă de particulă pe unitate de timp urmează legea lui Rayleigh.

2.5.2 Legi de probabilitate continue statistice

Prezentăm legile de probabilitate de tip continuu, denumite astfel în sistemul Matlab, prin pachetul de programe *Statistics toolbox*, deoarece utilizarea lor este strâns legată de statistică.

Legea t (Student) (`t`)

Variabila aleatoare X urmează *legea t* (*Student*) *de probabilitate*, notată $\mathcal{T}(n)$, dacă are densitatea de probabilitate

$$f(x \mid n) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad x \in \mathbb{R},$$

$n \in \mathbb{N}$ (*numărul gradelor de libertate*).

Gosset (1908) a descoperit această lege de probabilitate. Nu i-a fost permis să publice rezultatul, drept urmare a publicat–o sub pseudonimul Student.

Pentru $n = 1$, se obţine *legea de probabilitatea a lui Cauchy*.

Dacă $X_0, X_1, \ldots, X_n$ sunt independente, fiecare urmând legea normală de parametri $\mu = 0$ şi $\sigma = 1$, atunci

$$T = \frac{X_0}{\sqrt{\frac{1}{n}\sum_{k=1}^{n} X_k^2}},$$

urmează legea $\mathcal{T}(n)$.

Dacă Y urmează legea $\mathcal{T}(n)$, atunci

$$X = \frac{1}{2} + \frac{1}{2}\frac{Y}{\sqrt{n+Y^2}}$$

urmează legea $\mathcal{B}eta\left(\frac{n}{2}, \frac{n}{2}\right)$.

Când $n \to \infty$, se ajunge la legea normală standard $\mathcal{N}(0,1)$.

Programul 2.5.8. Pentru ilustrarea ultimei afirmaţii, vom scrie un program care să reprezinte pe aceeaşi figură graficele densităţilor de probabilitate pentru legile $\mathcal{T}(n)$ şi $\mathcal{N}(0,1)$.

Prin executarea programului Matlab

```
clf;
n = input('n='); x = -3:0.01:3;
fn = normpdf(x,0,1); fs = tpdf(x,n);
plot(x,fn,'k-.',x,fs,'k-')
legend('Legea normala', 'Legea Student')
```

pentru `n=10`, acesta produce graficele din Figura 2.8.

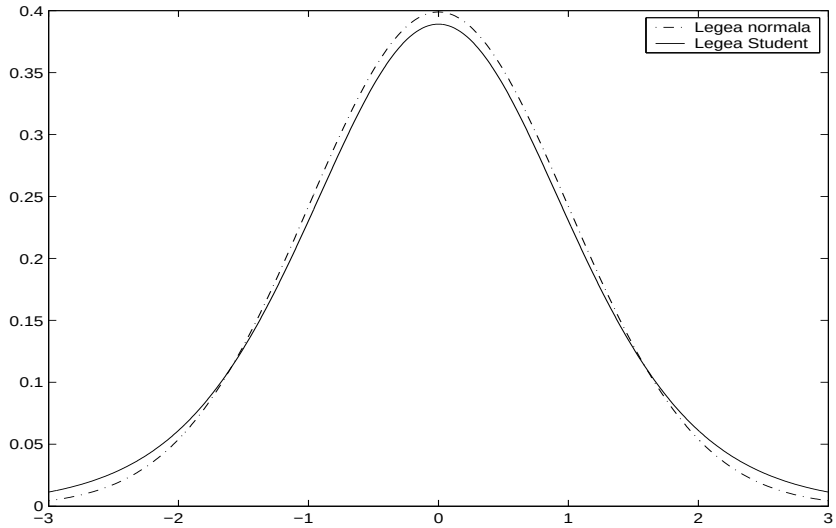

Figura 2.8: Legea $\mathcal{T}(10)$ şi legea $\mathcal{N}(0,1)$

Legea t (Student) necentrată (`nct`)

Legea t (*Student*) *necentrată*, notată $\mathcal{T}nc(n,\delta)$, generalizează legea $\mathcal{T}(n)$ şi are densitatea de probabilitate

$$f(x \mid n,\delta) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \mathrm{e}^{-\frac{\delta^2}{2}} \left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}} h(x \mid n,\delta), \quad x \in \mathbb{R},$$

unde

$$h(x \mid n\delta) = \sum_{k=0}^{\infty} \left(\frac{\delta t\sqrt{(2)}}{n}\right)^k \frac{\Gamma\left(\frac{n+k+1}{2}\right)}{k!\Gamma\left(\frac{n+1}{2}\right)} \left(1+\frac{x^2}{n}\right)^{-\frac{k}{2}}.$$

Programul 2.5.9. Programul Matlab ce urmează reprezintă grafic funcţiile de repartiţie pentru legile $\mathcal{T}(n)$ şi $\mathcal{T}nc(n,\delta)$:

```
clf;
n = input('n='); d = input('delta=');
x = -5:0.01:5;
Fc = tcdf(x,n); Fnc = nctcdf(x,n,d);
plot(x,Fc,'k-.',x,Fnc,'k-')
legend('Legea Student',..
       'Legea Student necentrata',2)
```

Pentru `n=10` şi `d=1`, se obţin graficele din Figura 2.9.

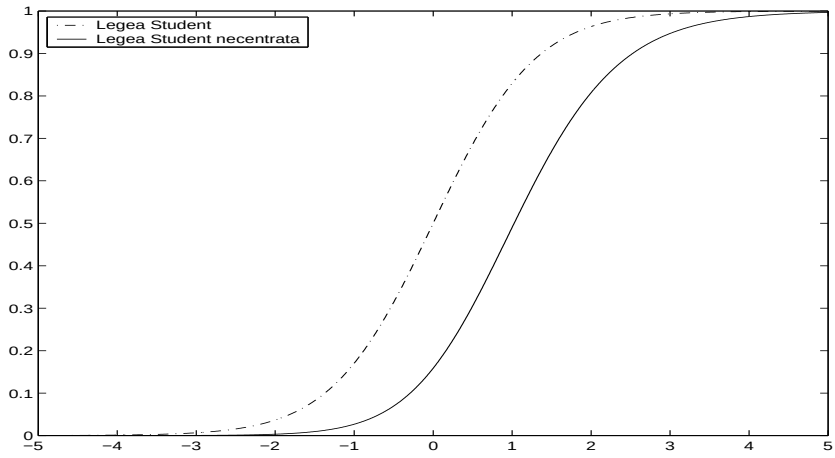

Figura 2.9: Legea $\mathcal{T}nc\,(10,1)$ şi legea $\mathcal{T}\,(10)$

Legea χ^2 (`chi2`)

Variabila aleatoare X urmează *legea* χ^2 sau *legea Helmert–Pearson*, vom nota prin $\chi^2\,(n)$, dacă are densitatea de probabilitate:

$$f\,(x\,|\,n) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1}\mathrm{e}^{-\frac{x}{2}}, \quad x > 0,$$

$n \in \mathbb{N}$ (*numărul gradelor de libertate*).

Legea $\chi^2\,(n)$ este un caz particular al legii $\mathcal{G}a\,(a,b)$, când $a = \frac{n}{2}$, iar $b = 2$.

Dacă $X_1, X_2, \ldots, X_n$ sunt independente, fiecare urmând legea normală de parametri $\mu = 0$ şi $\sigma = 1$, atunci variabila aleatoare

$$\chi^2 = \sum_{k=1}^{n} X_k^2,$$

urmează legea $\chi^2(n)$, iar dacă variabilele aleatoare independente X şi Y urmează respectiv legile $\mathcal{N}(0,1)$ şi $\chi^2(n)$, atunci variabila aleatoare

$$T = \frac{X}{\sqrt{\frac{Y}{n}}}$$

urmează legea $\mathcal{T}(n)$.

Dacă variabila aleatoare X urmează legea $\chi^2(n)$, atunci, pentru $n \to \infty$, variabila aleatoare urmează legea $\mathcal{N}(\mu, \sigma)$, cu $\mu = n$ şi $\sigma = \sqrt{2n}$.

Programul 2.5.10. Pentru ilustrarea afirmaţiei, scriem un program care să reprezinte pe aceeaşi figură graficele densităţilor de probabilitate pentru legile $\chi^2(n)$ şi $\mathcal{N}\left(n, \sqrt{2n}\right)$.

Prin executarea programului Matlab

```
clf;
n = input('n='); s = sqrt(2*n);
x = n-3*s:0.01:n+3*s;
fn = normpdf(x,n,s); fchi = chi2pdf(x,n);
plot(x,fn,'k-.',x,fchi,'k-')
legend('Legea normala', 'Legea chi2',2)
```

pentru `n=10`, se obţin graficele din Figura 2.10.

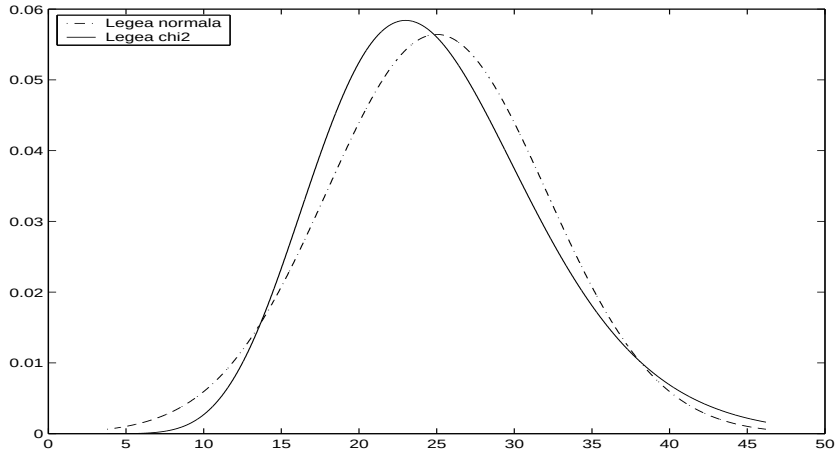

Figura 2.10: Legea $\chi^2(10)$ şi legea $\mathcal{N}\left(10, 2\sqrt{5}\right)$

Legea χ^2 necentrată (ncx2)

Să notăm prin χ^2_n o variabilă aleatoare ce urmează legea $\chi^2(n)$. Spunem că variabila aleatoare X urmează *legea* χ^2 *necentrată*, dacă are funcţia de repartiţie:

$$F(x \mid n, \delta) = \sum_{k=0}^{\infty} \frac{\left(\frac{\delta}{2}\right)^k \mathrm{e}^{-\frac{\delta}{2}}}{k!} P\left(\chi^2_{n+2k} \leqslant x\right).$$

Programul 2.5.11. Programul Matlab ce urmează reprezintă grafic densităţile de probabilitate pentru legea $\chi^2(n)$ şi legea $\chi^2(n, \delta)$ necentrată corespunzătoare:

```
clf;
n = input('n='); d = input('delta=');
x = 0:0.01:15;
c = chi2pdf(x,n); fnc = ncx2pdf(x,n,d);
plot(x,fc,'k-.',x,fnc,'k-')
legend('Legea chi2', 'Legea chi2 necentrata')
```

Pentru `n=4` şi `d=2`, se obţin graficele din Figura 2.11.

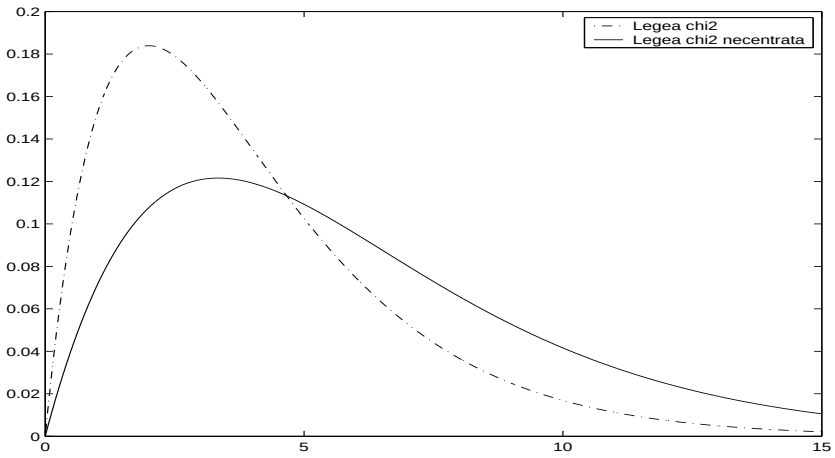

Figura 2.11: Legea $\chi^2(4, 2)$ şi legea $\chi^2(4)$

Legea F (Fisher–Snedecor) (f)

Variabila aleatoare X urmează legea *F* (*Fisher–Snedecor*), notată $\mathcal{F}(m, n)$, dacă are densitatea de probabilitate:

$$f(x \mid m, n) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1} \left(1 + \frac{m}{n}x\right)^{-\frac{m+n}{2}}, \quad x > 0;$$

$m, n \in \mathbb{N}$ (*numărul gradelor de libertate*)

Dacă variabilele aleatoare independente X şi Y urmează legile $\chi^2(m)$ şi $\chi^2(n)$, atunci

$$F = \frac{X}{m} \Big/ \frac{Y}{n}$$

urmează legea $\mathcal{F}(m, n)$.

Dacă variabila aleatoare T urmează legea $\mathcal{T}(n)$, atunci T^2 va urma legea $\mathcal{F}(1, n)$.

Dacă variabila aleatoare X urmează legea $\mathcal{F}(m, n)$, atunci *transformata lui Fisher*, adică $Z = \frac{\log X}{2}$, urmează legea $\mathcal{N}(\mu, \sigma)$, când $m, n \to \infty$, unde

$$\mu = \frac{1}{2}\left(\frac{1}{n} - \frac{1}{m}\right), \quad \sigma = \sqrt{\frac{1}{2}\left(\frac{1}{n} + \frac{1}{m}\right)}. \tag{2.5.2}$$

Programul 2.5.12. Programul care urmează reprezintă pe aceeaşi figură graficele densităţilor de probabilitate pentru transformata lui Fisher Z şi legea $\mathcal{N}(\mu, \sigma)$, unde μ şi σ au valorile date prin (2.5.2).

Remarcăm faptul că densitatea de probabilitate a variabilei aleatoare Z este

$$f_Z(x) = 2\mathrm{e}^{2x} f_X\left(\mathrm{e}^{2x} \mid m, n\right).$$

```
clf
m = input('m='); n = input('n=');
mu = (1/n-1/m)/2; s = sqrt((1/n+1/m)/2);
x = mu-3*s:0.01:mu+3*s;
f1 = normpdf(x,mu,s);
f2 = 2*exp(2*x).*fpdf(exp(2*x),m,n);
plot(x,f1,'k-.',x,f2,'k-')
legend('Legea normala','Transformata Fisher',2)
```

Pentru `m=20` şi `n=30`, se obţin graficele din Figura 2.12.

Legea F (Fisher–Snedecor) necentrată (`ncf`)

Dacă variabilele aleatoare independente X şi Y urmează respectiv legile $\chi^2(m, \delta)$ necentrată şi $\chi^2(n, \delta)$ necentrată, atunci

$$F = \frac{X}{m} \Big/ \frac{Y}{n}$$

urmează *legea F* (*Fisher–Snedecor*) *necentrată*, notată prin $\mathcal{F}nc(m, n, \delta)$.

Funcţia de repartiţie pentru variabila aleatoare legea $\mathcal{F}nc(m, n, \delta)$ este

$$F(x \mid m, n, \delta) = \sum_{k=0}^{\infty} \frac{\left(\frac{\delta}{2}\right)^k \mathrm{e}^{-\frac{\delta}{2}}}{k!} I\left(\frac{mx}{n + mx} \;\middle|\; \frac{m}{2} + k, \frac{n}{2}\right),$$

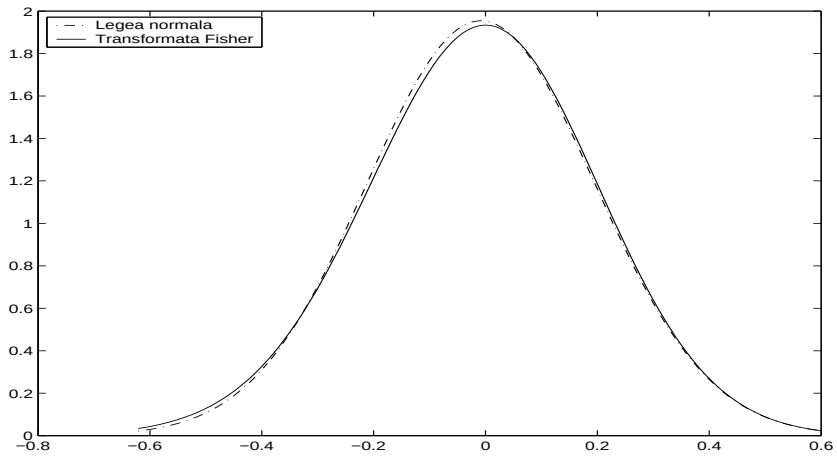

Figura 2.12: Transformata Fisher şi legea normală corespunzătoare

unde $I(x \mid a, b)$ este *funcţia Beta incompletă* de parametrii $a, b > 0$.

Programul 2.5.13. Programul Matlab ce urmează reprezintă grafic densităţile de probabilitate pentru legile $\mathcal{F}(m, n)$ şi $\mathcal{F}nc(m, n, \delta)$:

```
clf;
m = input('m='); n = input('n=');
d = input('delta=');
x = 0:0.01:10.01;
Fc = fpdf(x,m,n); Fnc = ncfpdf(x,m,n,d);
plot(x,Fc,'k-.',x,Fnc,'k-')
legend('Legea F', 'Legea F necentrata')
```

Pentru `m=5`, `n=20` şi `d=10`, se obţin graficele din Figura 2.13.

2.5.3 Funcţia Matlab `normspec`

Având în vedere că legea normală joacă un rol important în teoria probabilităţilor şi în statistică, sistemul Matlab îi acordă la rândul său o atenţie specială. Astfel, în sistemul Matlab de bază există funcţia `normspec`, cu următoarele moduri de apel:

```
normspec(sp,mu,sigma)
p = normspec(sp,mu,sigma)
```

Funcţia `normspec` reprezintă grafic densitatea de probabilitate a legii normale de parametrii `mu` şi `sigma` şi umbreşte aria mărginită de dreptele perpendiculare pe axa absciselor ce trec prin punctele axei precizate prin cele două componente ale vectorului `sp`, curba ce reprezintă graficul densităţii de probabilitate şi axa absciselor.

Parametrul `p`, după execuţia funcţiei `normspec`, conţine valoarea ariei umbrite, adică probabilitatea ca variabila aleatoare să ia valori din intervalul precizat prin `sp`.

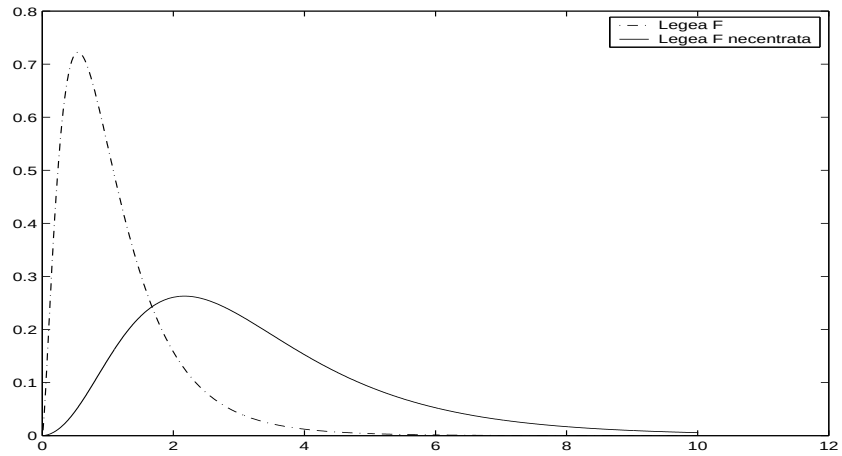

Figura 2.13: Legea $\mathcal{F}nc\,(5, 20, 10)$ şi legea $\mathcal{F}\,(5, 20)$

Programul 2.5.14. Programul Matlab ce urmează va reprezenta grafic densitatea de probabilitate a legii $\mathcal{N}\,(\mu, \sigma)$ şi va umbri aria cuprinsă între dreptele de ecuaţii $x = a$ şi $x = b$ $(a < b)$, curba ce reprezintă graficul densităţii de probabilitate şi axa absciselor.

```
m = input('mu='); s = input('sigma=');
a = input('a:'); b = input('b (b>a):');
p = normspec([a,b],m,s);
xlabel(['a= ',num2str(a),'    b= ',num2str(b)])
ylabel('Densitatea de probabilitate')
title(['Probabilitatea intre a si b : ',num2str(p)])
```

Executând acest program cu valorile de intrare `mu=0`, `sigma=1`, `a=-1.5` şi `b=2`, se obţine graficul din Figura 2.14.

Se observă că instrucţiunea `title` conţine parametrul `num2str(p)`, care transformă valoarea numerică a lui `p` într–un şir de caractere.

2.5.4 Distribuţie marginală

Fie vectorul aleator $\boldsymbol{X} = (X_1, \ldots, X_n)$ şi vectorul aleator $\tilde{\boldsymbol{X}} = (X_{i_1}, \ldots, X_{i_k})$, unde $1 \leqslant k \leqslant n$, iar $1 \leqslant i_1 < \cdots < i_k \leqslant n$.

Definiţia 2.5.15. *Numim* distribuţie marginală *a vectorului distribuţia vectorului* $\tilde{\boldsymbol{X}}$.

Dacă F este funcţia de repartiţie a vectorului $\boldsymbol{X}$, iar $F_{i_1 \ldots i_k}$ este funcţia vectorului aleator $\tilde{\boldsymbol{X}}$, numindu–se *funcţie de repartiţie marginală*, atunci

$$F_{i_1 \ldots i_k}\,(x_{i_1}, \ldots, x_{i_k}) = F\,(+\infty, \ldots, x_{i_1}, \ldots, x_{i_k}, \ldots, +\infty)\,,$$

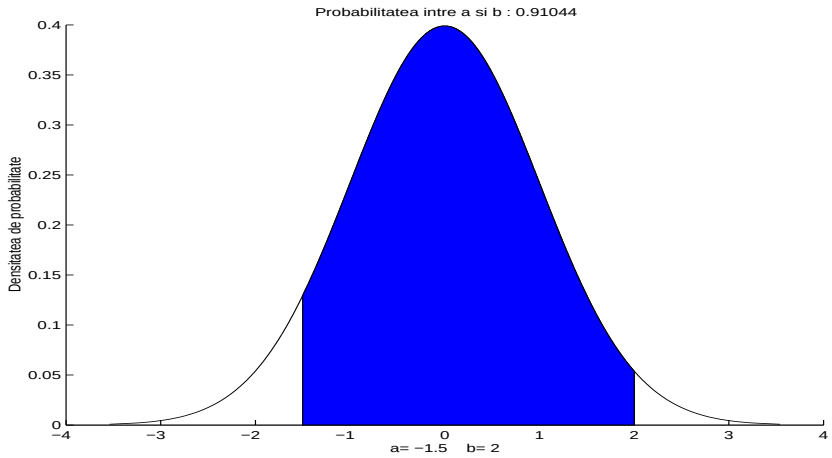

Figura 2.14: Legea $\mathcal{N}(0,1)$

adică, pentru obţinerea funcţiei de repartiţie a vectorului $\tilde{\boldsymbol{X}}$ se fac să tindă la ∞ toate argumentele funcţiei de repartiţie F, cu excepţia argumentelor $x_{i_1}, \ldots, x_{i_k}$.

Dacă vectorul aleator $\boldsymbol{X}$ este de tip continuu, având densitatea de probabilitate f, iar vectorul aleator $\tilde{\boldsymbol{X}}$ are densitatea de probabilitate $f_{i_1 \ldots i_k}$, care se numeşte *densitate de probabilitate marginală*, atunci

$$f_{i_1 \ldots i_k}(x_{i_1}, \ldots, x_{i_k}) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, dx_1 \ldots / \ldots / \ldots dx_n,$$

adică $f_{i_1 \ldots i_k}$ se obţine integrând densitatea de probabilitate f pe $\mathbb{R}^{n-k}$, în raport cu toate variabilele independente, cu excepţia variabilelor $x_{i_1}, \ldots, x_{i_k}$.

2.5.5 Funcţie de repartiţie condiţionată

Fie vectorul aleator bidimensional (X, Y), având funcţia de repartiţie F.

Definiţia 2.5.16. *Numim* funcţie de repartiţie condiţionată *a variabilei aleatoare X de către variabila aleatoare Y, funcţia $F_{X|Y} : \mathbb{R} \to \mathbb{R}$, dată prin*

$$F_{X|Y}(x|y) = P(X \leqslant x \,|\, Y = y), \quad \forall x \in \mathbb{R},$$

pentru fiecare $y \in \mathbb{R}$ fixat.

Trebuie să remarcăm faptul că formula de definiţie se poate scrie sub forma

$$F_{X|Y}(x|y) = \lim_{d \searrow 0} P(X \leqslant x \,|\, y - d < Y \leqslant y)$$

sau

$$F_{X|Y}(x|y) = \lim_{d \searrow 0} \frac{P(X \leqslant x, y - d < Y \leqslant y)}{P(y - d < Y \leqslant y)} = \lim_{d \searrow 0} \frac{F(x,y) - F(x, y-d)}{F_Y(y) - F_Y(y-d)}.$$

Dacă vectorul aleator (X, Y) este de tip discret, adică are distribuţia dată prin tabloul bidimensional

$X \setminus Y$	$\dots$	y_j	$\dots$
$\vdots$		$\vdots$	
x_i	$\dots$	p_{ij}	$\dots$
$\vdots$		$\vdots$	

atunci distribuţia variabilei aleatoare X condiţionată de $(Y = y_j)$ este dată prin

$$P(X = x_i \,|\, Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_{\cdot j}},$$

unde $p_{\cdot j} = P(Y = y_j) = \sum_{i \in I} p_{ij}$.

Remarcăm de asemenea că are loc formula

$$p_{i\cdot} = P(X = x_i) = \sum_{j \in J} p_{\cdot j} p_{i|j}, \tag{2.5.3}$$

unde $p_{i|j} = P(X = x_i \,|\, Y = y_j)$.

Dacă vectorul aleator (X, Y) este de tip continuu, adică are densitatea de probabilitate f, atunci densitatea de probabilitate a variabilei aleatoare X condiţionată de $(Y = y)$ este dată prin

$$f_{X|Y}(x \,|\, y) = \frac{f(x,y)}{f_Y(y)}, \quad \text{pentru } f_Y(y) \neq 0.$$

Are loc formula (2.5.3) în variantă continuă:

$$f_X(x) = \int_{-\infty}^{+\infty} f_Y(y)\, f_{X|Y}(x|y)\, dy.$$

Exemplul 2.5.17 (Legea normală multidimensională). Vectorul aleator, notat prin $\boldsymbol{X}$, urmează *legea normală multidimensională*, notăm aceasta prin $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{V})$, dacă are densitatea de probabilitate

$$\begin{aligned} f(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \boldsymbol{V}) &= \frac{[\det(\boldsymbol{A})]^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}} \times \\ &\times \exp\left[-\frac{1}{2} \sum_{i,j=1}^{n} a_{ij} (x_i - \mu_i)(x_j - \mu_j)\right], \quad \boldsymbol{x} \in \mathbb{R}^n,\ \boldsymbol{\mu} \in \mathbb{R}^n, \end{aligned} \tag{2.5.4}$$

unde $\boldsymbol{V}$ este o matrice pătratică de ordin n pozitiv definită, iar $\boldsymbol{A}$ este inversa matricei $\boldsymbol{V}$.

Dacă $n = 2$ şi notăm prin (X, Y) vectorul aleator bidimensional corespunzător, densitatea de probabilitate devine

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} \times \\ \times \exp\left\{ -\frac{1}{2(1-r^2)} \left[\frac{(x-\mu_1)^2}{\sigma_1^2} - 2r\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\}, \tag{2.5.5}$$

unde $\sigma_1, \sigma_2 > 0$ şi $|r| < 1$.

Se cunoaşte că fiecare din componentele X şi Y ale vectorului aleator urmează legea normală respectiv $\mathcal{N}(\mu_1, \sigma_1)$ şi $\mathcal{N}(\mu_2, \sigma_2)$.

Densitatea de probabilitate condiţionată se exprimă în acest fel prin:

$$f_{X|Y}(x|y) = \frac{1}{\sigma_1\sqrt{2\pi(1-r^2)}} \mathrm{e}^{-\frac{(x-m(y))^2}{2(1-r^2)\sigma_1^2}}, \tag{2.5.6}$$

unde $m(y) = \mu_1 + r\frac{\sigma_1}{\sigma_2}(y - \mu_2)$, adică se obţine densitatea de probabilitate a legii normale $\mathcal{N}\left(m(y), \sigma_1\sqrt{1-r^2}\right)$.

Programul 2.5.18. Scriem în cele ce urmează un program Matlab, care reprezintă grafic densitatea de probabilitate a legii normale bidimensionale.

```
clf, clear
m1 = input('mu1='); m2 = input('mu2=');
s1 = input('sigma1='); s2 = input('sigma2=');
r = input('r(-1<r<1):');
x = m1-3*s1:.2:m1+3*s1; y = m2-3*s2:.2:m2+3*s2;
[X,Y] = meshgrid(x,y);
Z = 1/(2*pi*s1*s2*sqrt(1-r^2))*exp(-1/(2*(1-r^2))...
      *((X-m1).^2/s1^2-2*r*(X-m1).*(Y-m2)/(s1*s2)...
      +(Y-m2).^2/s2^2));
mesh(X,Y,Z)
title('Legea normala bidimensionala')
```

Prin executarea acestui program, pentru `m1=m2=0`, `s1=1`, `s2=2` şi `r=-0.5`, se obţine graficul din Figura 2.15.

Programul 2.5.19. Programul Matlab care urmează reprezintă grafic densitatea de probabilitate condiţionată în cazul unui vector aleator ce urmează legea normală bidimensională.

```
clf, clear
m1 = input('mu1='); m2 = input('mu2=');
s1 = input('sigma1='); s2 = input('sigma2=');
```

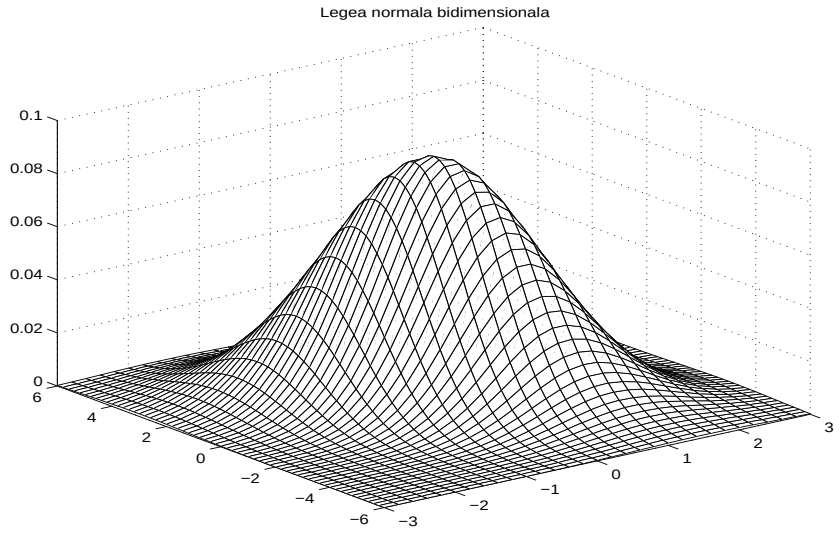

Figura 2.15: Legea $\mathcal{N}(0,0;1,2;-0.5)$

```
r = input('r(-1<r<1):');
Y(1:3) = input('Y(trei valori):');
m = m1+s1*r*(Y-m2)/s2; s = sqrt(1-r^2)*s1;
MM = max(m); mm = min(m);
X = []; X = mm-3*s:0.01:MM+3*s;
hold on
for i=1:3
    P = []; P = normpdf(X,m(i),s);
    if i==1  plot(X,P,'k-.'), end
    if i==2  plot(X,P,'k-'),  end
    if i==3  plot(X,P,'k--'), end
end
title('Legea normala conditionata')
```

Prin executarea acestui program, pentru `m1=1`, `m2=0`, `s1=1`, `s2=2`, `r=-0.5` şi `Y=[-2,0,2]`, se obţin graficele din Figura 2.16. Graficul marcat prin punct–linie este pentru prima valoare a lui `Y`, cel cu linie continuă, pentru a doua valoare, iar prin linie întreruptă, pentru a treia.

2.5.6 Funcţie de supravieţuire. Funcţie hazard

În aplicaţii tehnice şi nu numai, sunt utilizate şi alte funcţii ataşate variabilelor aleatoare, pentru exprimarea mai potrivită a fenomenului aleator cercetat.

Definiţia 2.5.20. *Dacă variabila aleatoare X are funcţia de repartiţie F, se numeşte*

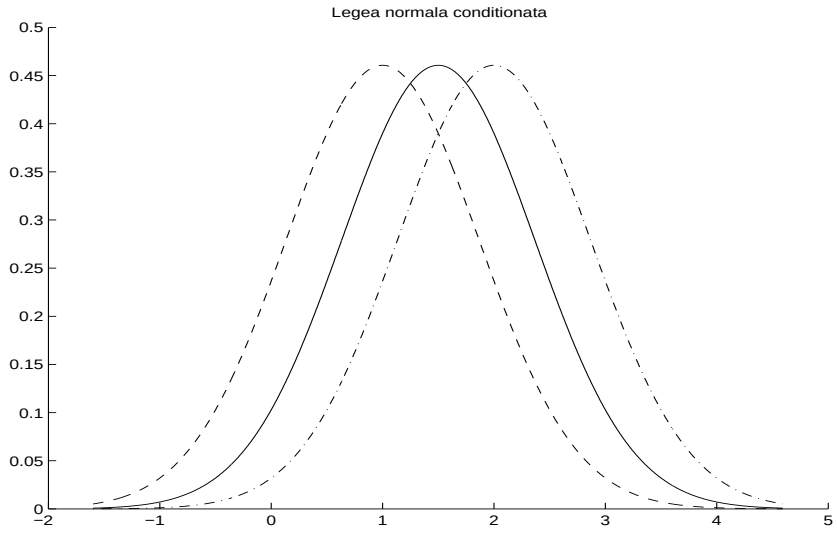

Figura 2.16: Legea normală condiţionată pentru $y = -2, 0, 2$

funcţie de supravieţuire *ataşată variabilei aleatoare X funcţia definită prin:*

$$S(x) = 1 - F(x) = 1 - P(X > x), \quad \forall x \in \mathbb{R}.$$

Se observă că dacă X reprezintă durata de viaţă a unui individ, atunci $S(x)$ reprezintă probabilitatea ca durată de viaţă să depăşească un prag x dat.

Definiţia 2.5.21. *Dacă variabila aleatoare de tip continuu X are funcţia de repartiţie F şi densitatea de probabilitate f, se numeşte* funcţie hazard *sau* funcţie de risc instantaneu *ataşată variabilei aleatoare X funcţia definită prin:*

$$h(x) = \frac{f(x)}{S(s)} = -\frac{S'(x)}{S(x)}, \quad \forall x \in \mathbb{R}.$$

Având în vedere definiţia densităţii de probabilitate putem scrie

$$f(x)\,dx \approxeq P(x \leqslant X < x + dx),$$

de unde

$$h(x)\,dx \approxeq \frac{P(x \leqslant X < x + dx)}{P(X > x)},$$

adică

$$h(x)\,dx \approxeq P(x \leqslant X < x + dx \,|\, X > x).$$

Prin urmare $h(x)$ ar reprezenta probabilitatea ca valoarea variabilei aleatoare X să rămână neschimbată într–un interval imediat următor, de lungime mică, dacă valoarea ei a depăşit pragul x.

2.6 Caracteristici numerice

2.6.1 Valoare medie. Dispersie (varianţă). Covarianţă

Fie variabila aleatoare X având funcţia de repartiţie F.

Definiţia 2.6.1. *Numim* valoare medie *sau* speranţă matematică, *caracteristica numerică*

$$E(X) = \int_{-\infty}^{+\infty} x dF(x), \tag{2.6.1}$$

unde integrala Stieltjes *se impune să fie absolut convergentă pentru ca valoarea medie să existe.*

Observaţia 2.6.2. Dacă variabila aleatoare este de tip discret, adică are distribuţia $X\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I}$, atunci integrala Stieltjes se reduce la o sumă, anume

$$E(X) = \sum_{i\in I} x_i p_i,$$

care există pentru cazul când X este variabilă aleatoare simplă, altfel se impune ca seria ce apare în această formulă să fie absolut convergentă.

Dacă variabila aleatoare este de tip continuu, adică are densitatea de probabilitate f, atunci

$$E(X) = \int_{-\infty}^{+\infty} x f(x)\, dx.$$

Mai remarcăm faptul că dacă $\boldsymbol{X}$ este un vector aleator n–dimensional, iar funcţia $h\colon \mathbb{R}^n \to \mathbb{R}$ este astfel încât $Y = h(\boldsymbol{X})$ să fie o variabilă aleatoare, atunci

$$E(Y) = \int_{-\infty}^{+\infty} x dF_Y(x) = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} h(\boldsymbol{x})\, dF_{\boldsymbol{X}}(\boldsymbol{x}),$$

care în cazul continuu devine

$$E(Y) = \int_{-\infty}^{+\infty} x f_Y(x)\, dx = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} h(\boldsymbol{x})\, f_{\boldsymbol{X}}(\boldsymbol{x}).$$

Definiţia 2.6.3. *Numim* dispersie *sau* varianţă *caracteristica numerică ataşată variabilei aleatoare X definită prin:*

$$Var\left(X\right)=E\left[\left(X-E\left(X\right)\right)^{2}\right]. \tag{2.6.2}$$

Remarcăm formula de calcul $Var\left(X\right)=E\left(X^{2}\right)-\left[E\left(X\right)\right]^{2}$.

Definiţia 2.6.4. *Dacă se consideră vectorul aleator bidimensional (X,Y), numim* covarianţă *sau* corelaţie*, caracteristica numerică*

$$Cov\left(X,Y\right)=E\left[\left(X-E\left(X\right)\right)\left(Y-E\left(Y\right)\right)\right], \tag{2.6.3}$$

respectiv coeficient de corelaţie*, raportul*

$$r\left(X,Y\right)=\frac{Cov\left(X,Y\right)}{\sqrt{Var\left(X\right)Var\left(Y\right)}}. \tag{2.6.4}$$

Observaţia 2.6.5. Pentru suma unui număr finit de variabile aleatoare avem:

$$Var\left(\sum_{k=1}^{n}X_{k}\right)=\sum_{k=1}^{n}Var\left(X_{k}\right)+2\sum_{i<k}Cov\left(X_{i},X_{k}\right),$$

iar dacă variabilele aleatoare sunt independente două câte două

$$Var\left(\sum_{k=1}^{n}X_{k}\right)=\sum_{k=1}^{n}Var\left(X_{k}\right).$$

Definiţia 2.6.6. *Dacă avem vectorul aleator $\boldsymbol{X}=\left(X_{1},\ldots,X_{n}\right)^{\intercal}$, numim* valoare medie*, vectorul*

$$E\left(\boldsymbol{X}\right)=\left(E\left(X_{1}\right),\ldots,E\left(X_{n}\right)\right)^{\intercal},$$

respectiv matrice a covarianţelor*, matricea*

$$V(\boldsymbol{X})=\left(Cov\left(X_{i},X_{k}\right)\right)_{i,k=\overline{1,n}}.$$

Observaţia 2.6.7. Din cauza simetriei covarianţei, avem că matricea covarianţelor este simetrică.

Dacă componentele vectorului aleator $\boldsymbol{X}$ sunt independente două câte două, atunci matricea covarianţelor este o matrice diagonală, pe diagonala principală aflându–se dispersiile componentelor vectorului aleator.

Având în vedere extinderea naturală a valorii medii la matrice aleatoare, putem scrie:

$$V(\boldsymbol{X})=E\left[\left(\boldsymbol{X}-E\left(\boldsymbol{X}\right)\right)\left(\boldsymbol{X}-E\left(\boldsymbol{X}\right)\right)^{\intercal}\right].$$

Teorema 2.6.8. *Fie vectorul aleator* $\boldsymbol{X} = (X_1, \ldots, X_n)^\intercal$ *şi matricea* $\boldsymbol{A}$*, de tipul* (m, n)*, cu elemente numere reale. Dacă se consideră vectorul* $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$*, atunci*

(i) $\boldsymbol{Y}$ *este un vector aleator;*

(ii) *valoarea medie a vectorului aleator* $\boldsymbol{Y}$ *este*

$$E(\boldsymbol{Y}) = \boldsymbol{A}E(\boldsymbol{X}) \quad \text{sau} \quad E(\boldsymbol{Y}^\intercal) = E(\boldsymbol{X}^\intercal)\,\boldsymbol{A}^\intercal;$$

(iii) *matricea covarianţelor vectorului aleator* $\boldsymbol{Y}$ *este dată prin*

$$V(\boldsymbol{Y}) = \boldsymbol{A}\,V(\boldsymbol{X})\,\boldsymbol{A}^\intercal. \tag{2.6.5}$$

Demonstraţie. (i) Transformarea liniară a unui vector aleator este vector aleator.

(ii) Având în vedere proprietăţi ale valorii medii, pentru o componentă a vectorului $E(\boldsymbol{Y})$, se poate scrie

$$E(Y_i) = E\left(\sum_{k=1}^{n} a_{ik}X_k\right) = \sum_{k=1}^{n} a_{ik}E(X_k),$$

de unde rezultă $E(\boldsymbol{Y}) = \boldsymbol{A}E(\boldsymbol{X})$.

În mod analog se arată şi a doua relaţie.

(iii) Se poate proceda ca şi pentru valoarea medie, dacă se ţine seama de proprietăţi ale corelaţiei. Se scrie succesiv:

$$\begin{aligned}
Cov(Y_i, Y_k) &= Cov\left(\sum_{p=1}^{n} a_{ip}X_p, \sum_{q=1}^{n} a_{kq}X_q\right) \\
&= E\left\{\left[\sum_{p=1}^{n} a_{ip}\left(X_p - E(X_p)\right)\right]\left[\sum_{q=1}^{n} a_{kq}\left(X_q - E(X_q)\right)\right]\right\} \\
&= \sum_{p=1}^{n}\sum_{q=1}^{n} a_{ip}E\left[\left(X_p - E(X_p)\right)\left(X_q - E(X_q)\right)\right]a_{kq} \\
&= \sum_{p=1}^{n}\sum_{q=1}^{n} a_{ip}Cov(X_p, X_q)\,a_{kq},
\end{aligned}$$

de unde se obţine relaţia (2.6.5).

Acelaşi lucru se poate obţine şi prin folosirea relaţiei de la punctul (ii). Anume,

se poate scrie:

$$\begin{aligned} V(\boldsymbol{Y}) &= E\left[(\boldsymbol{Y}-E(\boldsymbol{Y}))(\boldsymbol{Y}-E(\boldsymbol{Y}))^{\intercal}\right] \\ &= E\left[(\boldsymbol{AX}-\boldsymbol{A}E(\boldsymbol{X}))(\boldsymbol{AX}-\boldsymbol{A}E(\boldsymbol{X}))^{\intercal}\right] \\ &= E\left[\boldsymbol{A}(\boldsymbol{X}-E(\boldsymbol{X}))(\boldsymbol{X}-E(\boldsymbol{X}))^{\intercal}\boldsymbol{A}^{\intercal}\right] \\ &= \boldsymbol{A}E\left[(\boldsymbol{X}-E(\boldsymbol{X}))(\boldsymbol{X}-E(\boldsymbol{X}))^{\intercal}\right]\boldsymbol{A}^{\intercal} \\ &= \boldsymbol{A}\,V(\boldsymbol{X})\,\boldsymbol{A}^{\intercal}, \end{aligned}$$

de unde se obţine din nou relaţia (2.6.5). □

Proprietatea 2.6.9. *Dacă $V(\boldsymbol{X})$ este matricea covarianţelor vectorului aleator $\boldsymbol{X}$, atunci aceasta este o matrice pozitiv semi–definită, adică*

$$\boldsymbol{a}^{\intercal}V(\boldsymbol{X})\,\boldsymbol{a} \geqslant 0, \quad \forall \boldsymbol{a} \in \mathbb{R}^n.$$

Demonstraţie. Pornind din membrul drept al relaţiei ce se doreşte a fi demonstrată, putem scrie succesiv:

$$\begin{aligned} \boldsymbol{a}^{\intercal}V(\boldsymbol{X})\,\boldsymbol{a} &= \sum_{i=1}^{n}\sum_{k=1}^{n} a_i a_k Cov\,(X_i, X_k) = Cov\left(\sum_{i=1}^{n} a_i X_i, \sum_{k=1}^{n} a_k X_k\right) \\ &= Var\left(\sum_{i=1}^{n} a_i X_i\right) \geqslant 0, \end{aligned}$$

de unde se obţine relaţia din enunţ. □

Proprietatea 2.6.10. *Dacă matricea $\boldsymbol{W}$ este o matrice pătratică pozitiv definită, adică*

$$\boldsymbol{a}^{\intercal}V\,(\boldsymbol{X})\,\boldsymbol{a} > 0, \quad \forall \boldsymbol{a} \in \mathbb{R}^n \setminus \{\,\mathbf{0}\,\},$$

atunci $\boldsymbol{W}$ poate fi considerată matricea covarianţelor unui vector aleator.

Demonstraţie. Din algebra liniară este cunoscut faptul că o matrice pozitiv definită admite decompunerea

$$\boldsymbol{W} = \boldsymbol{U}^{\intercal}\boldsymbol{\Lambda}\boldsymbol{U} \quad \text{sau} \quad \boldsymbol{\Lambda} = \boldsymbol{U}\boldsymbol{W}\boldsymbol{U}^{\intercal},$$

unde $\boldsymbol{U}^{\intercal}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^{\intercal} = \boldsymbol{I}$, $\boldsymbol{I}$ fiind matricea unitate, iar $\boldsymbol{\Lambda}$ este o matrice diagonală cu elementele numere pozitive. Putem face precizarea că diagonala matricei $\boldsymbol{\Lambda}$ conţine valorile proprii ale matricei $\boldsymbol{W}$, iar matricea $\boldsymbol{U}$ este formată din vectorii proprii ortonormaţi corespunzători.

Considerăm vectorul aleator $\boldsymbol{Z}$ având matricea covarianţelor $\boldsymbol{I}$. Un astfel de vector aleator există, având în vedere de exemplu vectorul aleator care are ca şi componente variabile aleatoare independente ce urmează fiecare legea normală $\mathcal{N}(0,1)$.

Vectorul aleator $\boldsymbol{X} = \boldsymbol{U}^{\intercal}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{Z}$ va avea ca matrice a covarianţelor chiar matricea $\boldsymbol{W}$.

Întradevăr, dacă se ţine seama de relaţia (2.6.5), se obţine:

$$V(\boldsymbol{X}) = \boldsymbol{U}^{\intercal}\boldsymbol{\Lambda}^{\frac{1}{2}}V(\boldsymbol{Z})\left(\boldsymbol{U}^{\intercal}\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{\intercal} = \boldsymbol{U}^{\intercal}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U} = \boldsymbol{U}^{\intercal}\boldsymbol{\Lambda}\boldsymbol{U} = \boldsymbol{W}.$$

Din aceeastă succesiune de egalităţi se obţine afirmaţia făcută în enunţ. □

Corolarul 2.6.11. *Dacă matricea covarianţelor $V(\boldsymbol{X})$ a unui vector aleator $\boldsymbol{X}$ este pozitiv definită, atunci există o transformare $\boldsymbol{X} \mapsto \boldsymbol{Y}$ astfel încât pentru noul vector aleator $\boldsymbol{Y}$ să avem $E(\boldsymbol{Y}) = \boldsymbol{0}$ şi $V(\boldsymbol{Y}) = \boldsymbol{I}$.*

Demonstraţie. Folosind descompunerea $V(\boldsymbol{X}) = \boldsymbol{U}^{\intercal}\boldsymbol{\Lambda}\boldsymbol{U}$ şi alegând $\boldsymbol{A} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{U}$, obţinem că $\boldsymbol{Y} = \boldsymbol{A}[\boldsymbol{X} - E(\boldsymbol{X})]$ este vectorul aleator căutat.

Pe de o parte avem

$$E(\boldsymbol{Y}) = \boldsymbol{A}[E(\boldsymbol{X}) - E(\boldsymbol{X})] = \boldsymbol{A}\boldsymbol{0} = \boldsymbol{0}.$$

Pe de altă parte

$$\begin{aligned} V(\boldsymbol{Y}) &= \boldsymbol{A}\,V(\boldsymbol{X} - E(\boldsymbol{X}))\,\boldsymbol{A}^{\intercal} = \boldsymbol{A}\,V(\boldsymbol{X})\,\boldsymbol{A}^{\intercal} \\ &= \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{U}V(\boldsymbol{X})\,\boldsymbol{U}^{\intercal}\boldsymbol{\Lambda}^{-\frac{1}{2}} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-\frac{1}{2}} = \boldsymbol{I}. \end{aligned}$$

Din cel două succesiuni de relaţii se obţine că vectorul $\boldsymbol{Y}$ satisface condiţiile dorite. □

Corolarul 2.6.12. *Dacă matricea covarianţelor $V(\boldsymbol{X})$ a unui vector aleator $\boldsymbol{X}$ este pozitiv definită, atunci variabila aleatoare*

$$S^2 = [\boldsymbol{X} - E(\boldsymbol{X})]^{\intercal}\,[V(\boldsymbol{X})]^{-1}\,[\boldsymbol{X} - E(\boldsymbol{X})],$$

are valoarea medie $E\left(S^2\right) = n$.

Demonstraţie. Cu notaţiile din demonstraţia corolarului precedent avem

$$\boldsymbol{A}^{\intercal}\boldsymbol{A} = \boldsymbol{U}^{\intercal}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{U} = \boldsymbol{U}^{\intercal}\boldsymbol{\Lambda}^{-1}\boldsymbol{U} = [V(\boldsymbol{X})]^{-1}.$$

Prin urmare, avem succesiv:

$$S^2 = [\boldsymbol{X} - E(\boldsymbol{X})]^{\intercal}\,\boldsymbol{A}^{\intercal}\boldsymbol{A}\,[\boldsymbol{X} - E(\boldsymbol{X})] = \boldsymbol{Y}^{\intercal}\boldsymbol{Y} = \sum_{i=1}^{n} Y_i^2.$$

Dar avem că $E(Y_i) = 0$ şi $Var(Y_i) = 1$, astfel că $E\left(Y_i^2\right) = Var(Y_i) = 1$, pentru fiecare $i = \overline{1,n}$, ceea ce conduce la $E\left(S^2\right) = n$. □

2.6.2 Funcţii Matlab pentru valoare medie şi dispersie

Sistemul Matlab dispune de proceduri pentru calulul valorilor medii şi ale dispersiilor legilor de probabilitate implementate prin *Statistics toolbox*.

Pentru calculul valorilor acestor caracteristici numerice se foloseşte apelul

```
[m,v]=numef(x,par1,par2,...)
```

unde `numef` este un şir de caractere, care definesc legea de probabilitate, din care ultimele patru sunt `stat`, iar cele ce le preced sunt cele care dau numele legii.

În urma executării instrucţiunii, se calculează matricele `m` şi `v` ale valorilor medii şi ale dispersiilor pentru legea considerată, având parametrii precizaţi prin matricele `par1`, `par2`,... Aceste matrice trebuie să fie de aceleaşi dimensiuni, cu excepţia că dacă unele sunt scalari, aceştia se extind la matricele constante de aceleaşi dimensiuni cu celelalte şi care iau valorile scalarilor corespunzători.

Valorile medii şi dispersiile se calculează, pentru fiecare lege de probabilitate în parte, folosindu–se formulele din Tabelul 2.1.

2.6.3 Valoare medie condiţionată. Dispersie (varianţă) condiţionată

Fie vectorul aleator bidimensional (X,Y) şi $F_{X|Y}$ funcţia de repartiţie a lui X condiţionată de Y.

Definiţia 2.6.13. *Numim* valoare medie condiţionată *a variabilei aleatoare X de către Y, variabila aleatoare*

$$E(X \mid Y) = \int_{-\infty}^{+\infty} x dF_{X|Y}(x|Y),$$

o realizare a variabilei aleatoare $E(X \mid Y)$ fiind

$$E(X \mid y) = \int_{-\infty}^{+\infty} x dF_{X|Y}(x|y), \quad y \in \mathbb{R}.$$

Dacă vectorul aleator (X,Y) este de tip discret, cu distribuţia

$X \setminus Y$	$\ldots$	y_j	$\ldots$
$\vdots$		$\vdots$	
x_i	$\ldots$	p_{ij}	$\ldots$
$\vdots$		$\vdots$	

atunci variabila aleatoare $E(X|Y)$ va avea distribuţia

$$E(X|Y) \begin{pmatrix} E(X|y_j) \\ p_{\bullet j} \end{pmatrix}_{j \in J},$$

Legea	Denumirea	Valoarea medie şi dispersia
$\mathcal{U}(N)$	`unid`	$E(X)=\frac{N+1}{2}$, $Var(X)=\frac{N^2-1}{12}$
$\mathcal{B}(n,p)$	`bino`	$E(X)=np$, $Var(X)=np(1-p)$
$\mathcal{H}(n,M,K)$	`hyge`	$E(X)=n\frac{K}{M}$, $Var(X)=n\frac{K}{M}\frac{M-K}{M}\frac{M-n}{M-1}$
$\mathcal{P}o(\lambda)$	`poiss`	$E(X)=\lambda$, $Var(X)=\lambda$
$\mathcal{BN}(r,p)$	`nbin`	$E(X)=\frac{r(1-p)}{p}$, $Var(X)=\frac{r(1-p)}{p^2}$
$\mathcal{G}e(p)$	`geo`	$E(X)=\frac{1-p}{p}$, $Var(X)=\frac{1-p}{p^2}$
$\mathcal{U}(a,b)$	`unif`	$E(X)=\frac{a+b}{2}$, $Var(X)=\frac{(b-a)^2}{12}$
$\mathcal{N}(\mu,\sigma)$	`norm`	$E(X)=\mu$, $Var(X)=\sigma^2$
$\mathcal{LN}(\mu,\sigma)$	`logn`	$E(X)=\mathrm{e}^{\mu+\frac{\sigma^2}{2}}$, $Var(X)=\mathrm{e}^{2\mu+2\sigma^2}-\mathrm{e}^{2\mu+\sigma^2}$
$\mathcal{G}a(a,b)$	`gam`	$E(X)=ab$, $Var(X)=ab^2$
$\mathcal{E}xp(\mu)$	`exp`	$E(X)=\mu$, $Var(X)=\mu^2$
$\mathcal{B}eta(a,b)$	`beta`	$E(X)=\frac{a}{a+b}$, $Var(X)=\frac{ab}{(a+b+1)(a+b)^2}$
$\mathcal{W}(a,b)$	`weib`	$E(X)=a^{-\frac{1}{b}}\Gamma\left(1+\frac{1}{b}\right)$, $Var(X)=a^{-\frac{2}{b}}\left[\Gamma\left(1+\frac{2}{b}\right)-\Gamma^2\left(1+\frac{1}{b}\right)\right]$
$\mathcal{R}(b)$	`rayl`	$E(X)=b\sqrt{\frac{\pi}{2}}$, $Var(X)=\frac{4-\pi}{2}b^2$
$\mathcal{T}(n)$	`t`	$E(X)=0$, $Var(X)=\frac{n}{n-2}$, $n>2$
$\mathcal{T}nc(n,\delta)$	`nct`	$E(X)=\frac{\delta\sqrt{\frac{n}{2}}\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$, $Var(X)=\frac{n\left(1+\delta^2\right)}{n-2}-\frac{n\delta^2}{2}\left[\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\right]^2$, $n>2$
$\chi^2(n)$	`chi2`	$E(X)=n$, $Var(X)=2n$
$\chi^2(n,\delta)$	`ncx2`	$E(X)=n+\delta$, $Var(X)=2(n+2\delta)$
$\mathcal{F}(m,n)$	`f`	$E(X)=\frac{n}{n-2},\ n>2$, $Var(X)=\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)},\ n>4$
$\mathcal{F}nc(m,n,\delta)$	`ncf`	$E(X)=\frac{n(m+\delta)}{m(n-2)},\ n>2$, $Var(X)=2\left(\frac{n}{m}\right)^2\frac{(m+\delta)^2+(m+2\delta)(n-2)}{(n-2)^2(n-4)},\ n>4$

Tabelul 2.1: Tabelul valorilor medii şi dispersiilor

unde

$$E(X|y_j) = \sum_{i\in I} x_i P(X = x_i \,|\, Y = y_j) = \sum_{i\in I} x_i \frac{p_{ij}}{p_{\cdot j}}, \quad j \in J.$$

Dacă vectorul aleator (X, Y) este de tip continuu, având densitatea de probabilitate f, atunci o realizare a variabilei aleatoare $E(X|Y)$ este dată prin

$$E(X|y) = \int_{-\infty}^{+\infty} x f_{X|Y}(x \,|\, y)\, dx = \int_{-\infty}^{+\infty} x \frac{f(x,y)}{f_Y(y)} dy.$$

Proprietatea 2.6.14. $E[E(X|Y)] = E(X)$.

Demonstraţie. Demonstrăm proprietatea în cazul continuu. Cazul discret se demonstrează analog.

Folosind definiţia valorii medii avem succesiv:

$$\begin{aligned} E[E(X|Y)] &= \int_{-\infty}^{+\infty} E(X|y)\, f_Y(y)\, dy = \int_{-\infty}^{+\infty} f_Y(y) \left[\int_{-\infty}^{+\infty} x \frac{f(x,y)}{f_Y(y)} dx\right] dy \\ &= \int_{-\infty}^{+\infty} x \left[\int_{-\infty}^{+\infty} f(x,y)\, dy\right] dx = \int_{-\infty}^{+\infty} x f_X(dx) = E(X). \end{aligned}$$

Demonstraţia se încheie dacă se reţin extremităţile acestui şir de egalităţi. □

Definiţia 2.6.15. *Fie vectorul aleator* (X, Y). *Numim* dispersie condiţionată *sau* varianţă condiţionată *a variabilei aleatoare* X *de către* Y*, variabila aleatoare*

$$Var(X \,|\, Y) = E\left[(X - E(X|Y))^2 \,|Y\right].$$

Proprietatea 2.6.16. $Var(X) = E[Var(X|Y)] + Var[E(X|Y)]$.

Demonstraţie. Pornind de la definiţia varianţei şi folosind proprietăţi ale ei, putem scrie:

$$\begin{aligned} Var(X) =& E\left[(X - E(X))^2\right] = E\left[(X - E(X|Y) + E(X|Y) - E(X))^2\right] \\ =& E\left[(X - E(X|Y))^2\right] + E\left[(E(X|Y) - E(X))^2\right] \\ &+ 2E\left[(X - E(X|Y))(E(X|Y) - E(X))\right]. \end{aligned}$$

Vom calcula pe rând cei trei termeni din partea dreaptă a acestei relaţii.

Pentru primul termen, dacă se notează $Z = (X - E(X|Y))^2$ şi se aplică proprietatea valorii medii condiţionate avem:

$$E(Z) = E[E(Z|Y)] = E\left\{E\left[(X - E(X|Y))^2 \mid Y\right]\right\},$$

de unde se obţine că

$$E\left[\left(X-E\left(X|Y\right)\right)^{2}\right]=E\left[Var\left(X|Y\right)\right].$$

Pentru al doilea termen, dacă se notează $Z=E\left(X|Y\right)$, se poate scrie:

$$\begin{aligned}E\left[\left(E\left(X|Y\right)-E\left(X\right)\right)^{2}\right] &= E\left[\left(E\left(X|Y\right)-E\left(E\left(X|Y\right)\right)\right)^{2}\right]\\ &= E\left[\left(Z-E\left(Z\right)\right)^{2}\right]=Var\left(Z\right)=Var\left[E\left(X|Y\right)\right],\end{aligned}$$

de unde

$$E\left[\left(E\left(X|Y\right)-E\left(X\right)\right)^{2}\right]=Var\left[E\left(X|Y\right)\right].$$

A mai rămas să arătăm ca al treilea termen este zero.

Fixăm $Y=y$ şi având în vedere că $E\left(X|y\right)-E\left(X\right)=$ const., putem scrie:

$$\begin{aligned}E\left[\left(X-E\left(X|y\right)\right)\left(E\left(X|y\right)-E\left(X\right)\right)\right] &= \left[E\left(X|y\right)-E\left(X\right)\right]E\left[X-E\left(X|y\right)\right]\\ &= \left[E\left(X|y\right)-E\left(X\right)\right]\left[E\left(X\right)-E\left(X\right)\right]=0.\end{aligned}$$

Ceea ce a mai rămas de arătat. □

Exemplul 2.6.17. Să considerăm vectorul aleator (X,Y), care urmează legea normală, având densitatea dată prin (2.5.5). Este cunoscut că parametrii acestei legi de probabilitate au următoarele interpretări probabilistice:

$$\mu_1=E\left(X\right),\quad \mu_2=E\left(Y\right),\quad \sigma_1^2=Var\left(X\right),\quad \sigma_2^2=Var\left(Y\right),$$

iar $r\in(-1,1)$ este coeficientul de corelaţie dintre X şi Y.

Mai mult, pentru legea normală multidimensională, având densitatea de probabilitate dată prin (2.5.4), se arată că $\boldsymbol{\mu}$ este valoarea medie a vectorului aleator ce urmează legea normală multidimensională, iar $\boldsymbol{V}$ este matricea covarianţelor.

Să revenim la cazul bidimensional şi să constatăm că

$$\begin{aligned}&E\left(X|Y=y\right)=m\left(y\right)=\mu_1+\frac{\sigma_1}{\sigma_2}r\left(y-\mu_2\right),\\ &Var\left(X|Y=y\right)=\sigma_1^2\left(1-r^2\right).\end{aligned}$$

În mod analog, avem că

$$\begin{aligned}&E\left(Y|X=x\right)=m\left(x\right)=\mu_2+\frac{\sigma_2}{\sigma_1}r\left(x-\mu_1\right),\\ &Var\left(Y|X=x\right)=\sigma_2^2\left(1-r^2\right).\end{aligned}$$

Curbele de ecuaţii

$$x(y) = \mu_1 + \frac{\sigma_1}{\sigma_2} r(y - \mu_2),$$
$$y(x) = \mu_2 + \frac{\sigma_2}{\sigma_1} r(x - \mu_1),$$

se numesc *curbele de regresie* a lui X în raport cu Y, respectiv a lui Y în raport cu X.

2.6.4 Funcţie caracteristică

Fie variabila aleatoare X şi vectorul aleator n–dimensional $\boldsymbol{X}$.

Definiţia 2.6.18. *Numim* funcţie caracteristică *ataşată variabilei aleatoare X, respectiv vectorului aleator $\boldsymbol{X}$, funcţia definită prin*

$$\varphi(t) = M\left(\mathrm{e}^{itX}\right), \quad \forall t \in \mathbb{R},$$
$$\varphi(\boldsymbol{t}) = M\left(\mathrm{e}^{i\sum_{k=1}^{n} t_k X_k}\right), \quad \forall \boldsymbol{t} \in \mathbb{R}^n.$$

Utilitatea funcţiei caracteristice este dată de netezimea ei, care este superioară funcţiei de repartiţie, această din urmă putând fi discontinuă, pe când funcţia caracteristică este uniform continuă. În plus, pe baza formulei de inversiune, există o caracterizare completă a distribuţiei unei variabile aleatoare sau a unui vector aleator cu ajutorul funcţiei caracteristice.

2.6.5 Momente

Definiţia 2.6.19. *Fie variabila aleatoare X. Numim* moment (iniţial) *şi respectiv* moment centrat *de ordin n, caracteristicile numerice*

$$\nu_n = E(X^n), \quad \mu_n = E[(X - E(X))^n] = E[(X - \nu_1)^n].$$

Remarcăm că în statistică sunt folosite în primul rând primele patru momente.

Astfel momentele centrate de ordinele doi şi trei se folosesc pentru definirea *asimetriei* (*skewness*)

$$a = \frac{\mu_3}{\mu_2^{\frac{3}{2}}},$$

iar momentele centrate de ordinele doi şi patru se folosesc pentru definirea *excesului* (*kurtosis*)

$$e = \frac{\mu_4}{\mu_2^2} \quad \text{sau} \quad e = \frac{\mu_4}{\mu_2^2} - 3.$$

În sistemul Matlab, excesul este definit cu prima formulă.

2.6.6 Mediana. Cuartile. Cuantile

Definiţia 2.6.20. *Fie variabila aleatoare X, care are funcţia de repartiţie F. Numim* mediană, *caracteristica numerică m, care satisface condiţiile*

$$P(X \geqslant m) \geqslant \frac{1}{2} \leqslant P(X \leqslant m). \tag{2.6.6}$$

Având în vedere proprietăţi ale funcţiei de repartiţie, condiţiile (2.6.6) se pot scrie echivalent sub forma

$$F(m-0) \leqslant \frac{1}{2} \leqslant F(m).$$

Dacă variabila aleatoare este de tip continuu, relaţiile (2.6.6) se reduc la relaţia $F(m) = \frac{1}{2}$, adică $m = F^{-1}\left(\frac{1}{2}\right)$.

În cazul discret, folosind (2.6.6) s–ar putea ca mediana să nu fie determinată în mod unic, de aceea se alege m ca fiind cel mai mic număr pentru care $F(m) \geqslant \frac{1}{2}$.

Definiţia 2.6.21. *Fie variabila aleatoare X, care are funcţia de repartiţie F. Numim* cuantilă *de ordin $\gamma \in (0,1)$, caracteristica numerică x_γ, care satisface condiţiile*

$$P(X \geqslant x_\gamma) \geqslant 1-\gamma \quad \text{şi} \quad P(X \leqslant m) \geqslant \gamma. \tag{2.6.7}$$

Ca şi în cazul medianei, aceste condiţii se pot scrie echivalent sub forma

$$F(x_\gamma - 0) \leqslant \gamma \leqslant F(x_\gamma),$$

care se reduce la relaţia $F(x_\gamma) = \gamma\cdot$, adică $x_\gamma = F^{-1}(\gamma)$, în cazul continuu.

În cazul discret, cuantila x_γ se ia ca fiind cel mai mic număr pentru care are loc inegalitatea $F(x_\gamma) \geqslant \gamma$.

Unele cuantile au denumiri speciale. Iată de exemplu, cuantila de ordin $\gamma = \frac{1}{2}$ este chiar mediana.

Dacă $\gamma = \frac{1}{4}, \frac{2}{4}, \frac{3}{4}$, se obţin ceea ce se numesc *cuartile*, respectiv *cuartila inferioară*, *mediana* şi *cuartila superioară*.

Dacă $\gamma = \frac{i}{10}$, $i = \overline{1,9}$, $\gamma = \frac{k}{100}$, $k = \overline{1,99}$, se obţin ceea ce poartă denumirile *decile* şi respectiv *centile*.

2.6.7 Funcţia Matlab `icdf`

Am văzut că pentru calculul cuantilelor este necesară inversarea funcţiei de repartiţie. Sistemul Matlab prin *Statistics toolbox* dispune de funcţii pentru inversarea funcţiilor de repartiţie ale legilor de probabilitate implementate.

Apelarea acestor funcţii se face cu una din următoarele forme:

```
x = icdf('legea',P,p1,p2,...)
x = numef(P,p1,p2,...)
```

unde `legea` este un şir de caractere predefinit pentru fiecare din legile de probabilitate disponibile în *Statistics toolbox*, `numef` este un şir de caractere din care ultimele trei sunt `inv`, iar cele ce le preced sunt cele care dau numele predefinit al legii de probabilitate (ca şi cele din parametrul `legea`).

În urma executării uneia din cele două instrucţiuni, se calculează matricea `x` a cuantilelor legii precizată prin parametrii `legea`, respectiv `numef`, corespunzătoare valorilor date prin matricea `P` şi având parametrii daţi prin matricele `par1`, `par2`,... Aceste matrice trebuie să fie de aceleaşi dimensiuni, cu excepţia că dacă unele sunt scalari, aceştia se extind la matricele constante de aceleaşi dimensiuni cu celelalte şi care iau valorile scalarilor corespunzători.

Programul 2.6.22. Pentru a ilustra grafic modul de calcul al cuantilelor, vom da un program care calculează mediana pentru legea uniformă discretă şi reprezintă grafic acest lucru pentru două valori distincte ale parametrului `N`, precum şi cuartilele legii normale.

```
clf, clear
N1 = input('N1='); N2 = input('N2=');
m = input('mu='); s = input('sigma=');
xu1 = 0:N1+1; yu1 = unidcdf(xu1,N1);
xu2 = 0:N2+1; yu2 = unidcdf(xu2,N2);
xn = m-3*s:0.01:m+3*s; yn = normcdf(xn,m,s);
me1 = icdf('unid',1/2,N1);
me2 = icdf('unid',1/2,N2);
Q = icdf('norm',[1/4,2/4,3/4],m,s);
subplot(3,1,1), stairs(xu1,yu1)
set(gca,'Xlim',[0,N1+1]), set(gca,'xtick',[me1])
set(gca,'xticklabel',[me1]), hold on
plot([0,me1],[1/2,1/2],'k:',me1,1/2,'o')
plot([me1,me1],[0,1/2],'k-.')
subplot(3,1,2), stairs(xu2,yu2)
set(gca,'Xlim',[0,N2+1]), set(gca,'xtick',[me2])
set(gca,'xticklabel',[me2]), hold on
plot([0,me2],[1/2,1/2],'k:',me2,1/2,'o')
plot([me2,me2],[0,1/2],'k-.')
subplot(3,1,3),
plot(xn,yn,[Q(1),Q(2),Q(3)],[1/4,2/4,3/4],'o')
set(gca,'xtick',[Q(1),Q(2),Q(3)])
set(gca,'xticklabel',[Q(1),Q(2),Q(3)])
hold on
X = [m-3*s,Q(1);m-3*s,Q(2);m-3*s,Q(3)];
plot(X',[1/4,1/4;1/2,1/2;3/4,3/4]','k:')
plot([Q(1),Q(1)],[0,1/4],'k-.')
plot([Q(2),Q(2)],[0,2/4],'k-.')
plot([Q(3),Q(3)],[0,3/4],'k-.')
```

Executarea programului pentru `N=5` şi `N=6`, respectiv pentru `mu=0` şi `sigma=2`, are ca rezulatat graficele din Figura 2.17.

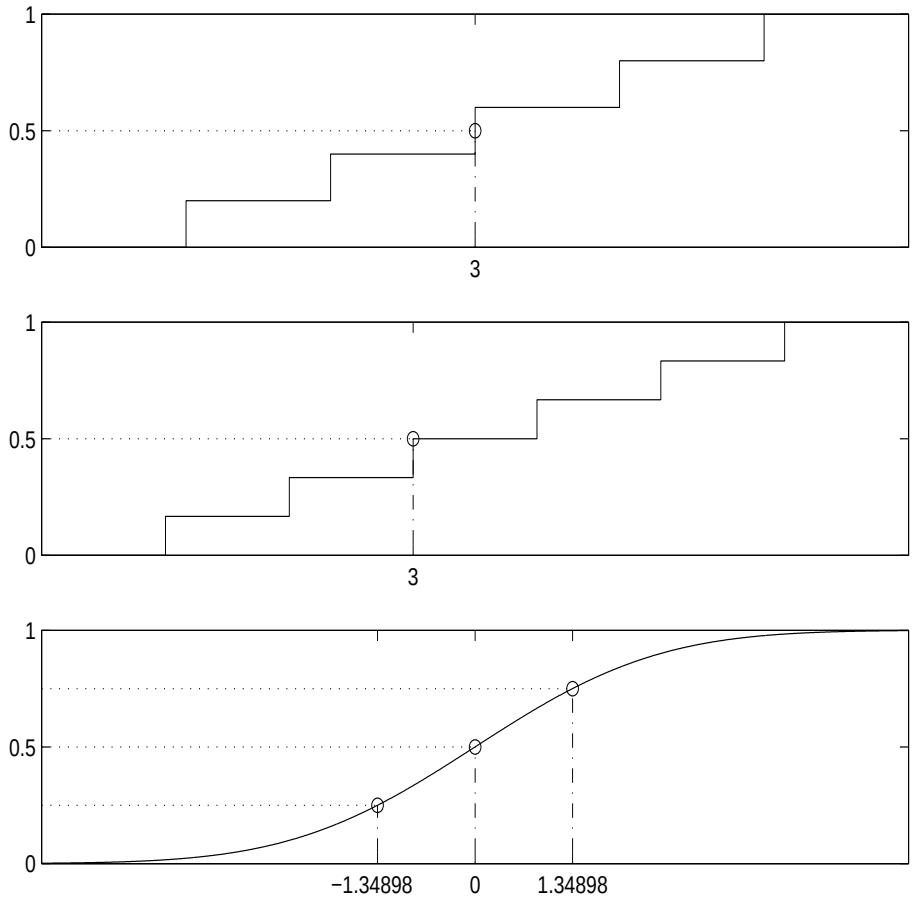

Figura 2.17: Medianele pentru $\mathcal{U}(5)$ şi $\mathcal{U}(6)$ şi cuartilele pentru $\mathcal{N}(0,2)$

2.6.8 Funcţia Matlab `disttool`

Funcţia (comanda) `disttool` este un program de demonstrativ. Lansarea acestuia se face prin:

```
>>disttool
```

în urma căreia se produce o fereastră grafică interactivă (demonstrativă) privind funcţiile de repartiţie (`cdf`) şi funcţiile de probabilitate (`pdf`).

Fixarea (stabilirea) legii de probabilitate în scop demonstrativ se face prin alegerea din meniul legilor de probabilitate situat în partea stângă sus a ferestrei, iar una din alternativele `pdf` şi `cdf` se face folosind meniul din partea dreaptă sus a ferestrei. Pentru stabilirea valorilor parametrilor legii de probabilitate considerate se poate proceda în două moduri. Fie prin introducerea în ferestrele corespunzătoare ale valorilor dorite, fie prin deplasarea barelor ataşate acestora. În plus, limite pentru parametrii legii de probabilitate considerate pot fi precizate prin introducerea acestora în ferestrele considerate.

Determinarea valorii funcţiei `pdf` respectiv `cdf` într–un punct, se poate obţine prin introducerea argumentului funcţiei în fereastra de pe axa absciselor sau prin deplasarea dreptei verticale afişată pe grafic, cu ajutorul *mouse*–lui, până când aceasta trece prin punctul respectiv. Inversa funcţiei de repartiţie (`icdf`) se poate obţine de asemenea, prin introducerea valorii funcţiei de repartiţie (`cdf`) în fereastra de pe axa ordonatelor sau prin deplasarea dreptei orizontale afişată pe grafic, cu ajutorul *mouse*–lui, până când aceasta trece prin punctul respectiv.

2.7 Şiruri de variabile aleatoare

2.7.1 Inegalitatea lui Cebîşev

Fie variabila aleatoare X, pentru care există valoarea medie şi dispersia.

Inegalitatea lui Cebîşev este cunoscută sub una din formele

$$P\left(|X - E(X)| < \varepsilon\right) \geqslant 1 - \frac{Var(X)}{\varepsilon^2},$$
$$P\left(|X - E(X)| \geqslant \varepsilon\right) \leqslant \frac{Var(X)}{\varepsilon^2}, \quad \forall \varepsilon > 0.$$

Observaţia 2.7.1. Dacă în inegalitatea lui Cebîşev se ia $\varepsilon = k\sigma$, unde s–a notat $\sigma = \sqrt{Var(X)}$ *abaterea standard*, atunci aceasta devine

$$P\left(|X - E(X)| < k\sigma\right) \geqslant \frac{k^2 - 1}{k^2}.$$

Astfel, pentru $k = 3$ se obţine $P(|X - E(X)| < 3\sigma) \geqslant \frac{8}{9} = 0,88...$, ceea ce înseamnă că probabilitatea ca valorile lui X să se abată de la valoarea medie $E(X)$ mai puţin de trei ori abaterea standard este mai mare decât $\frac{8}{9}$.

Observaţia 2.7.2. Fie variabilele aleatoare X_k, $k = \overline{1,n}$, independente şi identic repartizate, având valorile medii şi dispersiile $E(X_k) = \mu$, $Var(X_k) = \sigma^2$, pentru fiecare $k = \overline{1,n}$, cu ajutorul cărora construim variabila aleatoare

$$S_n = \sum_{k=1}^{n} X_k, \quad \text{cu} \quad E(S_n) = n\mu \quad \text{şi} \quad Var(S_n) = n\sigma^2.$$

Dacă se aplică inegalitatea lui Cebîşev se obţine

$$P\left(\left|\frac{S_n}{n} - \mu\right| < \varepsilon\right) \geqslant 1 - \frac{\sigma^2}{n\varepsilon^2}, \quad P\left(\left|\frac{S_n}{n} - \mu\right| \geqslant \varepsilon\right) \leqslant \frac{\sigma^2}{n\varepsilon^2}, \quad \forall \varepsilon > 0.$$

Cele două inegalităţi dau evaluări ale probabilităţilor ca media aritmetică a variabilelor aleatoare să se abată de la valoarea medie μ mai puţin, respectiv mai mult, decât $\varepsilon > 0$ fixat.

2.7.2 Tipuri de convergenţă

Să considerăm şirul de variabile aleatoare $(X_n)_{n\geqslant 1}$ cu şirul funcţiilor de repartiţie corespunzătoare $(F_n)_{n\geqslant 1}$.

Definiţia 2.7.3. *Spunem că şirul de variabile aleatoare $(X_n)_{n\geqslant 1}$* converge în probabilitate *la variabila aleatoare X, şi vom nota $X_n \xrightarrow{p} X$, dacă pentru orice $\varepsilon > 0$, avem că*

$$\lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1.$$

Definiţia 2.7.4. *Spunem că şirul de variabile aleatoare $(X_n)_{n\geqslant 1}$* converge în repartiţie *la variabila aleatoare X, şi vom nota $X_n \xrightarrow{r} X$, dacă avem că*

$$\lim_{n\to\infty} F_n(x) = F(x),$$

unde F este funcţia de repartiţie a variabilei aleatoare X, iar limita considerată există pentru orice punct de continuitate $x \in \mathbb{R}$ al funcţiei F.

Observaţia 2.7.5. Dacă $X_n \xrightarrow{p} X$, atunci $X_n \xrightarrow{r} X$. Implicaţia inversă nu are loc, dar dacă X este o constantă a, atunci afirmaţia inversă are loc de asemenea, adică dacă $X_n \xrightarrow{r} a$, pentru $a \in \mathbb{R}$, atunci $X_n \xrightarrow{p} a$.

2.7.3 Legea numerelor mari

Să considerăm şirul de varaibile aleatoare $(X_n)_{n\geqslant 1}$.

Definiţia 2.7.6. *Spunem că şirul de variabile aleatoare* $(X_n)_{n\geqslant 1}$ *urmează* legea (slabă) a numerelor mari, *dacă pentru orice* $\varepsilon > 0$ *avem că*

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n}\sum_{k=1}^{n} M\left(X_k\right)\right| < \varepsilon\right) = 1,$$

adică

$$\frac{1}{n}\sum_{k=1}^{n} X_k - \frac{1}{n}\sum_{k=1}^{n} M\left(X_k\right) \xrightarrow{p} 0.$$

Observaţia 2.7.7. Dacă variabilele aleatoare ale şirului $(X_n)_{n\geqslant 1}$ sunt independente şi identic repartizate, iar $E\left(X_n\right) = \mu$, atunci

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| < \varepsilon\right) = 1, \quad \text{adică} \quad \frac{1}{n}S_n \xrightarrow{p} \mu,$$

unde $S_n = \sum_{k=1}^{n} X_k$.

Prin urmare, media aritmetică a primelor n variabile aleatoare din şirul considerat îşi pierde caracterul aleator, pentru $n \to \infty$.

Evident, deoarece convergenţa în probabilitate implică convergenţa în repartiţie, avem şi $\frac{1}{n}S_n \xrightarrow{r} \mu$, dar vom vedea imediat că există chiar un rezultat mai bun în ceea ce priveşte convergenţa în repartiţie.

2.7.4 Teoreme limită

Am văzut că legea numerelor mari este în strânsă legătură cu convergenţa în probabilitate, chiar şi cu aşa numita convergenţa tare, dar care nu a fost prezentată în secţiunea precedentă.

Teoremele limită abordează problematica comportării asimptotice a şirurilor de varaibile aleatoare din perspectiva convergenţei în repartiţie.

Teorema 2.7.8 (Lindeberg–Levy). *Fie şirul de variabile aleatoare independente şi identic repartizate* $(X_n)_{n\geqslant 1}$, *pentru care se introduc notaţiile*

$$E\left(X_n\right) = \mu, \quad Var\left(X_n\right) = \sigma^2, \quad S_n = \sum_{k=1}^{n} X_k.$$

Dacă se construieşte şirul de variabile aleatoare $(Z_n)_{n\geqslant 1}$*, cu termenul general*

$$Z_n = \frac{\frac{1}{n}S_n - \mu}{\frac{\sigma}{\sqrt{n}}},$$

atunci $Z_n \xrightarrow{r} Z$*, unde* Z *este o variabilă aleatoare ce urmează legea normală* $\mathcal{N}(0,1)$*, adică*

$$\lim_{n\to\infty} F_n(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{t^2}{2}} dt, \quad \forall x \in \mathbb{R},$$

F_n *fiind funcţia de repartiţie a variabilei aleatoare* Z_n*.*

Observaţia 2.7.9. Teorema limită de acest tip, în care legea de probabilitate limită este legea normală se numeşte *teoremă limită centrală.*

Teoremele limită au un rol aparte în statistică, unde sunt privite în următoarea manieră. Dacă n este mare ($n \to \infty$), atunci variabila aleatoare Z_n poate fi considerată ca urmând o lege de probabilitate cunoscută, în cazul teoremei precedente legea normală $\mathcal{N}(0,1)$.

Mai observăm că variabila aleatoare $\frac{1}{n}S_n$, când $n \to \infty$, urmează legea normală $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Teorema 2.7.10 (Moivre–Laplace). *Dacă variabila aleatoare* Z_n *urmează legea binomială, adică are distribuţia*

$$S_n \begin{pmatrix} x \\ f(x \mid n,p) \end{pmatrix}_{x=\overline{0,n}}, \quad \text{unde } f(x \mid n,p) = \binom{n}{x} p^x q^{n-x},\ q = 1-p,$$

atunci

$$\frac{S_n - np}{\sqrt{npq}} \xrightarrow{r} Z,$$

unde Z *este o variabilă aleatoare ce urmează legea normală* $\mathcal{N}(0,1)$*.*

Rezultatul teoremei se exprimă de regulă sub forma

$$P\left(a \leqslant \frac{x-np}{\sqrt{npq}} < b\right) \approxeq \frac{1}{\sqrt{2\pi}} \int_a^b \mathrm{e}^{-\frac{t^2}{2}} dt. \tag{2.7.1}$$

Observaţia 2.7.11. *Teorema limită Moivre–Laplace* este o consecinţă imediată a Teoremei Lindeberg–Levy, dacă se are în vedere că variabila aleatoare S_n este suma

a n variabile aleatoare independente X_k, $k = \overline{1,n}$, fiecare urmând aceeaşi lege a lui Bernoulli, adică având distibuţia

$$X_k \begin{pmatrix} x \\ f\left(x \mid p\right) \end{pmatrix}_{x=1,0}, \quad \text{unde } f\left(x \mid p\right) = p^x q^{1-x}.$$

Mai remarcăm faptul că variabila aleatoare S_n, care urmează legea binomială $\mathcal{B}\left(n,p\right)$, poate fi privită, pentru $n \to \infty$, ca urmând legea normală $\mathcal{N}\left(np, \sqrt{npq}\right)$.

Observaţia 2.7.12 (Regula celor trei σ). Dacă în formula Moivre–Laplace se consideră $b = -a = 3$, atunci

$$P\left(-3 \leqslant \frac{k-np}{\sqrt{npq}} < 3\right) \approxeq \Phi\left(3\right) - \Phi\left(-3\right) = 0.997.$$

Prin urmare, rezultă că $P\left(|k-np| < 3\sqrt{npq}\right) \geqslant 0.99$. Deoarece, abaterea medie pătratică a variabilei aleatoare X ce urmează legea binomială este $\sigma = \sqrt{npq}$, avem regula următoare cunoscută sub denumirea de *regula celor trei* σ: probabilitatea ca frecvenţa absolută k să se abată de la valoarea medie $E(X) = np$ mai puţin de trei ori abaterea medie pătratică este mai mare decât 0.99.

Observaţia 2.7.13 (Corecţie de continuitate). Având în vedere procesul de aproximare a unei legi de tip discret (legea binomială) cu una de tip continuu (legea normală), pentru obţinerea unor rezultate mai bune în aplicarea formulei (2.7.1), se impune aplicarea unei corecţii de continuitate, dată prin

$$\begin{aligned} f\left(x \mid n,p\right) = P\left(S_n = x\right) &\approxeq P\left(\frac{x-\frac{1}{2}-np}{\sqrt{npq}} \leqslant Z < \frac{x+\frac{1}{2}-np}{\sqrt{npq}}\right) \\ &= \Phi\left(\frac{x+\frac{1}{2}-np}{\sqrt{npq}}\right) - \Phi\left(\frac{x-\frac{1}{2}-np}{\sqrt{npq}}\right), \end{aligned}$$

unde

$$\Phi\left(x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{t^2}{2}} dt,$$

reprezintă funcţia lui Laplace.

Folosind această corecţie de continuitate avem următoarele formule de aproxi-

mare:

$$F_n(x) = P(S_n \leqslant x) \approxeq P\left(Z \leqslant \frac{x+\frac{1}{2}-np}{\sqrt{npq}}\right) = \Phi\left(\frac{x+\frac{1}{2}-np}{\sqrt{npq}}\right),$$

$$1-F_n(x) = P(S_n > x) \approxeq P\left(Z > \frac{x-\frac{1}{2}-np}{\sqrt{npq}}\right) = 1-\Phi\left(\frac{x-\frac{1}{2}-np}{\sqrt{npq}}\right),$$

$$P(k_1 \leqslant S_n \leqslant k_2) \approxeq P\left(\frac{k_2+\frac{1}{2}-np}{\sqrt{npq}} \leqslant Z \leqslant \frac{k_1-\frac{1}{2}-np}{\sqrt{npq}}\right)$$
$$= \Phi\left(\frac{k_2+\frac{1}{2}-np}{\sqrt{npq}}\right) - \Phi\left(\frac{k_1-\frac{1}{2}-np}{\sqrt{npq}}\right).$$

Programul 2.7.14. Pentru a ilustra rezultatele Teoremei Moivre–Laplace, programul Matlab care urmează, reprezintă grafic prin bare funcţia de probabilitate a legii binomiale $\mathcal{B}(n,p)$ împreună cu densitatea de probabilitate a legii normale corespunzătoare, adică $\mathcal{N}\left(np, \sqrt{npq}\right)$.

```
clf
n = input('n='); p = input('p=');
m = n*p; s = sqrt(n*p*(1-p));
x = 0:n; xx = -1:0.01:n+1;
P = binopdf(x,n,p); y = normpdf(xx,m,s);
bar(x,P,0.3), hold, plot(xx,y)
```

Executarea programului pentru `n=15` şi `p=0.4`, produce graficul din Figura 2.18.

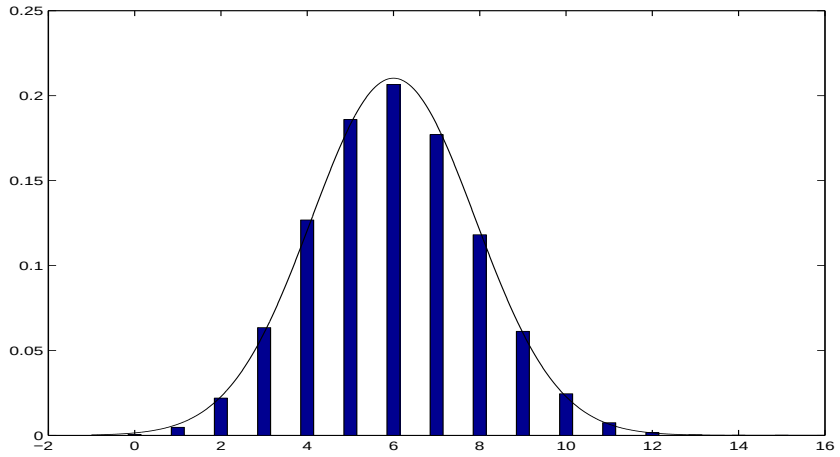

Figura 2.18: Legea $\mathcal{B}(15, 0.4)$ şi legea $\mathcal{N}\left(6, \sqrt{3.6}\right)$

Capitolul 3

Statistică descriptivă

Statistica se ocupă cu descrierea şi analiza numerică a fenomenelor de masă, dezvăluind particularităţile lor de volum, structură, dinamică, conexiune, precum şi regularităţile sau legile care le guvernează.

Etapele principale ce se disting în cercetarea statistică a unui fenomen aleator se pot considera ca fiind:

a). *Definirea (concepţia)* obiectului studiat, care conţine definirea unităţilor statistice, conceperea chestionarului (întrebărilor), planificarea culegerii datelor.

b). *Observarea fenomenului* (*culegerea datelor*) conform criteriilor stabilite la etapa precedentă.

c). *Descrierea statistică*, ce cuprinde reprezentarea grafică a datelor statistice, sistematizarea acestora, precum şi calcularea indicatorilor numerici pentru punerea în evidenţă a unor proprietăţi şi pentru sugerarea unor ipoteze referitoare la legile care guvernează fenomenul cercetat.

d). *Modelarea probabilistică* a fenomenului cercetat, care are ca obiectiv principal cercetarea fenomenului folosind ca instrument de lucru teoria probabilităţilor relativ la datele statistice obţinute conform etapelor precedente.

Dacă primele două etape pot fi considerate ca preliminare, etapa a treia conduce la ceea ce numim *statistică descriptivă*, iar a patra etapă conduce la *statistica matematic*ă sau *statistica inferenţială*.

Desigur, nu se poate face o delimitare exactă între statistica descriptivă şi statistica matematică, mai mult o descriere statistică fundamentată matematic va conduce la realizarea unei modelări a fenomenului, care pune în evidenţă proprietăţile esenţiale ale acestuia şi legăturile dintre acestea. Ca o componentă de bază a statisticii descriptive, care s-a dezvoltat în mod impresionant odată cu explozia utilizării calculatoarelor în statistică, o reprezintă *analiza datelor*. Printre capitolele mai importante ale analizei datelor amintim *metodele de clasificare* şi *metodele factoriale*.

Pe de altă parte, statistica matematică cuprinde ca şi capitole principale *teoria selecţiei*, *teoria estimaţiei* şi *verificarea ipotezelor statistice*, strâns legat de *teoria deciziilor*.

Din cele prezentate putem spune că *statistica descriptivă* are ca obiectiv principal culegerea şi clasificarea datelor statistice în vederea descrierii numerice şi grafice a fenomenului cercetat. *Statistica matematică* elaborează modelul matematic, folosind datele statistice, estimează parametrii modelului, stabileşte metode pentru validarea ipotezelor statistice, toate acestea având la baza teoria probabilităţilor.

Amintim că studiul statistic al fenomenelor economice face obiectul *statisticii economice*, iar studiul statistic al fenomenelor ce pot constitui riscuri asigurate reprezintă obiectul *statisticii actuariale*.

3.1 Concepte de bază ale statisticii

Definiţia 3.1.1. *Numim* colectivitate (populaţie) *o mulţime $\mathcal{C}$ de elemente cercetată din punct de vedere a uneia sau mai multor proprietăţi, elementele componente numindu-se* indivizi *sau* unităţi statistice, *iar numărul elementelor lui $\mathcal{C}$ se numeşte* volumul colectivităţii.

Definiţia 3.1.2. *Numim* caracteristică *sau* variabilă *a colectivităţii $\mathcal{C}$ proprietatea supusă investigării statistice relativă la $\mathcal{C}$. Când o caracteristică poate fi măsurată o numim* caracteristică cantitativă *sau* numerică, *iar dacă aceasta se exprimă printr-o însuşire o numim* caracteristică calitativă *sau* atribut.

În unele tratate de statistică variabilele (caracteristicile) calitative sunt numite *variabile nominale*. O clasă intermediară de variabile (caracteristici) este formată de aşa numitele *variabile* (*caracteristici*) *ordinale*.

Observaţia 3.1.3. Dacă se consideră colectivitatea $\mathcal{C}$ a studenţilor dintr-un an de studiu, atunci înălţimea, greutatea, media obţinută la sfârşitul anului sunt caracteristici cantitative (numerice), pe când sexul, culoarea ochilor, grupa sanguină sunt caracteristici calitative (nominale). Un exemplu de caracteristică ordinală ar fi gradul de pregătire (foarte bun, bun, satisfăcător, nesatisfăcător). Se observă că relativ la o variabilă ordinală se poate face ordonarea indivizilor, ca şi pentru o variabilă numerică, ceea ce nu se întâmplă pentru o variabilă nominală, dar nu se pot face operaţii aritmetice, cum se întâmplă şi la variabilele nominale, pe când pentru o variabilă numerică astfel de operaţii sunt posibile.

Observaţia 3.1.4. Din punct de vedere al teoriei probabilităţilor o caracteristică a unei populaţii $\mathcal{C}$ este o variabilă aleatoare X. Scopul principal al cercetării statistice

este de a stabili legea de probabilitate pe care o urmează caracteristica X, utilizând observaţiile (datele statistice) relative la colectivitatea cercetată.

Dacă avem în vedere mai multe caracteristici $X_1,\ldots,X_n$, acestea reprezintă componentele unui vector aleator $\boldsymbol{X}$, iar cercetarea fenomenului aleator modelat prin $\boldsymbol{X}$, se poate extinde şi la determinarea legăturilor ce există între componentele vectorului aleator.

Definiţia 3.1.5. *O caracteristică X ce ia o mulţime cel mult numărabilă de valori o numim* caracteristică de tip discret*, iar dacă ia valori dintr-un interval finit sau infinit o numim* caracteristică de tip continuu*.*

Observaţia 3.1.6. Dacă se consideră colectivitatea $\mathcal{C}$ a bolnavilor externaţi dintr–un spital pe parcursul unei săptămâni, atunci caracteristica X ce reprezintă numărul zilelor de internare este o caracteristică de tip discret, pe când greutatea Y a bolnavilor externaţi reprezintă o caracteristică de tip continuu. Remarcăm faptul că pentru caracteristica Y caracterul continuu este estompat prin imposibilitatea măsurării exacte a acestei caracteristici, de aceea, practic, se consideră ca fiind de asemenea de tip discret.

3.2 Culegerea, prezentarea şi prelucrarea datelor statistice

Modurile de culegere (observare) a datelor statistice cele mai des utilizate sunt:

a). *Observarea totală* (*exhaustivă*, *recensământ*), când toţi indivizii colectivităţii $\mathcal{C}$ sunt înregistraţi.

b). *Observare parţială* (*sondaj*, *selecţie*), când după criterii bine stabilite sunt înregistraţi o parte din indivizii colectivităţii $\mathcal{C}$, numită *eşantion* sau *selecţie*.

c). *Observare curentă*, când înregistrarea indivizilor se efectuează odată cu apariţia (producerea) lor.

d). *Observarea periodică*, când înregistrarea fenomenului se efectuează la intervale de timp stabilite.

3.2.1 Generarea numerelor aleatoare în Matlab

Matematica, şi prin teoria probabilităţilor şi statistica matematică, printre multiplele şi larg răspânditele aplicaţii, posedă, poate, una legată de *modelarea matematică*, care se distinge în mod special prin vastul domeniu de aplicabilitate. Modelarea matematică, după cum spune şi numele, are drept scop construirea de *modele matematice* asociate unor fenomene sau sisteme pe care le întâlnim în diferite domenii: fizică, chimie, biologie, demografie, medicină, geografie, geologie, economie etc. Desigur,

astfel de modele matematice depind de nişte parametri, fie aleatori, fie nealeatori. Odată modelul matematic construit se are în vedere realizarea practică a acestuia. Dar, apare o mare problemă, este acest model matematic (teoretic) construit viabil? Avem două alternative, sau realizăm practic modelul respectiv şi urmărim cum funcţionează acesta (ceea ce, în general, nu e de dorit), sau simulăm funcţionarea modelului respectiv prin diferite particularizări ale parametrilor modelului şi alegem varianta convenabilă. Dacă relativ la parametrii nealeatori care intervin în modelul teoretic nu se ridică probleme speciale, în privinţa parametrilor aleatori apare o nouă problemă, anume obţinerea valorilor acestor parametri. Deoarece parametrii aleatori sunt, din punct de vedere al teoriei probabilităţilor, variabile aleatoare, apare aşadar problema generării de valori numerice ale acestor variabile aleatoare şi care vor purta denumirea de *numere aleatoare*. Desigur că într-un fel vor fi generate aceste numere aleatoare dacă parametrul urmează, să zicem, legea uniformă pe un anumit interval şi altfel când urmează legea normală.

Deoarece pentru prezentarea noţiunilor şi rezultatelor teoretice ale statisticii matematice sunt necesare exemplificări ale acestora, iar culegerea unor date statistice reale nu este întodeauna la îndemână, de multe ori vom considera numerele aleatoare ca fiind date statistice cărora li se vor aplica metodele cercetării statisticii.

Putem să amintim de asemenea că testarea programelor pe calculator, determinarea statistică a caracteristicilor numerice ale variabilelor aleatoare, care ne conduce la metoda Monte Carlo, fac de asemenea necesară generarea numerelor aleatoare.

Metodele de generare a numerelor aleatoare se împart în trei categorii:

- tabelele cu numere aleatoare (uniforme) obţinute, de exemplu, din aruncarea monedei, aruncarea zarului, jocul de ruletă, tabele matematice, tabele de recensământ etc.

- procedeele fizice, care au la baza fenomene fizice, cum ar fi emiterea particulelor de o sursă radioactivă, intensitatea curentului la diferite momente, zgomotul electronic etc.

- procedeele aritmetice (analitice), care utilizează o formulă de calcul de forma

$$x_{n+1} = f\left(x_n, x_{n-1}, \ldots, x_{n-m}\right), \quad n \geqslant m \geqslant 0.$$

Deoarece, de regulă, se foloseşte un volum mare de numere aleatoare, s-a impus generarea acestora cu ajutorul procedeelor aritmetice. Inconvenientul acestor procedee analitice este legat de faptul că nu posedă caracterul strict aleator, din moment ce există o legătură funcţională între acestea. Dar există posibilitatea de a alege astfel de procedee analitice care produc şiruri de *numere pseudoaleatoare*, foarte apropiate din punct de vedere statistic de numerele aleatoare propriu–zise.

În lucrările [3] şi [5] sunt prezentate proceduri pentru generarea de numere aleatoare pentru diferite legi de probabilitate.

Sistemul Matlab are implementate astfel de proceduri pentru generarea numerelor aleatoare, fie prin *Statistics toolbox*, fie prin sistemul de bază din Matlab.

Funcţia `random`

Instrucţiunile prin care sunt generate numere aleatoare ce urmează una din legile de probabilitate recunoscute de sistemul Matlab prin *Statistics toolbox* sunt de forma:

```
x=random('legea',par1,par2,...,m,n)
x=numef(par1,par2,...,m,n)
```

unde `legea` este un şir de caractere predefinit pentru fiecare din legile de probabilitate disponibile în *Statistics toolbox*, `numef` este un şir de caractere din care ultimele trei sunt `rnd`, iar cele ce le preced sunt cele care dau numele predefinit al legii de probabilitate (ca şi cele din parametrul `legea`).

Executarea uneia din cele două forme ale instrucţiunii `random`, are ca efect generarea matricei `x`, cu `m` linii şi `n` coloane, de numere aleatoare ce urmează legea de probabilitate corespunzătoare cu parametrii precizaţi prin matricele `p1`, `p2`, ... Aceste matrice trebuie să aibă aceleaşi dimensiuni cu matricea `x`, cu excepţia când unele sunt scalari, caz în care acestea sunt extinse la matricele constante de aceleaşi dimensiuni cu celelalte şi care iau valorile corespunzătoare scalarilor. Dacă toţi parametrii legii de probabilitate sunt scalari, când lipsesc parametrii `m` şi `n` se generează un număr aleator, când lipseşte `n`, atunci trebuie ca `m` să fie un vector cu două componente, care conţine dimensiunile matricei `x`. Dacă cel puţin unul din parametrii legii probabilitate nu este scalar şi dacă parametrii `m` şi `n` sunt prezenţi, aceştia trebuie să coincidă cu dimensiunile matricelor.

Funcţiile `mvnrnd` şi `mvtrnd`

Funcţiile `mvnrnd` şi `mvtrnd` sunt singurele funcţii din *Statistics toolbox*, prin care se generează legi de probabilitate multidimensionale, respectiv legile normală şi t (Student).

Apelurile acestor funcţii se pot face respectiv prin:

```
x=mvnrnd(mu,v,m)
x=mvtrnd(r,n,,m)
```

În primul caz, sunt generaţi `m` vectori aleatori ce urmează legea normală multidimensională, având vectorul valorilor medii `mu` şi matricea covarianţelor `v` (matrice pătratică pozitiv definită). Cei `m` vectori aleatori generaţi sunt obţinuţi în matricea `x`, care în consecinţă va avea `m` linii, iar numărul coloanelor este dat de dimensiunea legii de probabilitate şi care trebuie să coincidă cu lungimea vectorului `mu` şi cu ordinul matricei pătratice `v`.

În al doilea caz, vor fi generaţi `m` vectori aleatori ce urmează legea Student multidimensională.

Parametrul `r` reprezintă matricea coeficienţilor de corelaţie, care este matrice pozitiv definită având pe diagonala principală toate elementele 1. Dacă această ultimă condiţie nu este îndeplinită, adică dacă este matricea covarianţelor, atunci sistemul Matlab calculează prima dată, din matricea `r`, matricea coeficienţilor de corelaţie.

Parametrul `n` reprezintă numărul gradelor de libertate, putând fi un scalar sau un vector de lungime `m`.

Cei `m` vectori aleatori sunt obţinuţi în cele `m` linii ale matricei `x` şi care are atâtea coloane cât este dimensiunea legii de probabilitate.

Trebuie să remarcăm faptul că o linie a matricei `x` se obţine dintr–un vector aleator ce urmează legea normală multidimensională, având valoarea medie `0` şi matricea covarianţelor dată prin parametrul `r`, ale cărui componente se împart la un număr aleator ce urmează legea χ^2 cu `n` grade de libertate.

Funcţiile `rand` şi `randn`

Sistemul Matlab de bază conţine două funcţii pentru generarea de numere aleatoare ce urmează legea uniformă $\mathcal{U}(0,1)$, prin `rand`, respectiv legea normală standard $\mathcal{N}(0,1)$, prin `randn`.

Apelul acestor funcţii se poate face prin instrucţiuni de forma:

```
x=rand
x=rand(m)
x=rand(m,n)
x=randn
x=randn(m)
x=randn(m,n)
```

În urma executării unor astfel de instrucţiuni este generat un număr aleator ce urmează legea $\mathcal{U}(0,1)$ respectiv $\mathcal{N}(0,1)$, când nu este folosit nici un argument, o matrice pătratică de ordin `m` de numere aleatoare, când apare un argument şi o matrice de numere aleatoare de tipul `(m,n)`, când sunt folosite ambele argumente.

Unele situaţii impun regăsirea unui anumit şir de numere aleatoare generat prin una din cele două funcţii `rand` şi respectiv `randn`. Pentru aceasta avem la dispoziţie ceea ce se numeşte *starea generatorului*.

Obţinerea stării generatorului `rand` şi `randn` la un moment dat se poate face prin:

```
s=rand('state')
s=randn('state')
```

Pentru reiniţializarea sau schimbarea stării generatorului, se pot utiliza următoarele instrucţiuni:

```
rand('state',s)
randn('state',s)
rand('state',0)
randn('state',0)
rand('state',sum(100*clock))
randn('state',sum(100*clock))
```

Parametrul `s` precizează starea la care urmează să fie resetat generatorul, dacă acesta este `0`, generatorul va fi resetat la starea iniţială. Dacă apare parametrul `sum(100*clock)`, generatorul este resetat la fiecare apel în funcţie de timpul calculatorului.

Instrucţiuni moştenite din Matlab 4 şi Matlab 5 sunt de asemenea active:

```
rand('seed')
randn('seed')
rand('seed',0)
randn('seed',0)
rand('seed',r)
randn('seed',r)
```

Funcţia `randperm`

Generarea unei permutări a primelor `n` numere naturale se realizează prin instrucţiunea

```
perm=randperm(n)
```

şi va avea ca efect obţinerea vectorului `perm` ce conţine o astfel de permutare.

3.2.2 Tabele statistice

Definiţia 3.2.1. *Numim* tabel statistic (simplu, nesistematizat) *un tablou în care înregistrările sunt trecute în ordinea apariţiei lor.*

Observaţia 3.2.2. Dacă observaţiile statistice s-au făcut relativ la p caracteristici, fie acestea X_1, X_2, ...,X_p, asupra a n indivizi, atunci tabelul statistic simplu este matricea $\boldsymbol{X} = (x_{ij})_{i=\overline{1,n},j=\overline{1,p}}$, unde x_{ij} este valoarea caracteristicii X_j pentru individul i.

Definiţia 3.2.3. *Numim* tabel statistic (sistematizat) *relativ la caracteristica X de tip discret, tabloul în care sunt trecute valorile distincte ale caracteristicii şi frecvenţele cu care au apărut aceste valori.*

Observaţia 3.2.4. Având caracteristica X de tip discret, care ia valorile distincte $x_i, i = \overline{1,n}$, pentru care s-au obţinut *datele primare* $x'_1, x'_2, \ldots, x'_N$, tabelul statistic sistematizat are forma

x	f
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_n	f_n

unde f_i este *frecvenţa absolută* a apariţiei valorii x_i în datele primare $x'_k, k = \overline{1, N}$. Prin urmare are loc relaţia $\sum_{i=1}^{n} f_i = N$.

Definiţia 3.2.5. *Fie caracteristica de tip continuu X, care ia valori din intervalul (a, b), descompus în intervale disjuncte, numite* clase, *prin punctele care satisfac relaţiile*

$$a = a_0 < a_1 < \cdots < a_n = b.$$

Numim tabel statistic (sistematizat) *relativ la caracteristica X, tabloul ce conţine clasele caracteristicii şi frecvenţele cu care au apărut aceste clase.*

Observaţia 3.2.6. Dacă datele primare ale caracteristicii continue X, care ia valori în intervalul (a, b), sunt $x'_1, x'_2, \ldots, x'_N$, atunci tabelul statistic sistematizat are forma

x	f
$[a_0, a_1)$	f_1
$[a_1, a_2)$	f_2
$\vdots$	$\vdots$
$[a_{n-1}, a_n)$	f_n

sau

x	f
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_n	f_n

unde f_i este *frecvenţa absolută* a apariţiei clasei $[a_{i-1}, a_i)$ printre datele primare $x'_k, k = \overline{1, N}$, iar $x_i = \frac{a_{i-1}+a_i}{2}$. Dacă se consideră varianta din dreapta pentru tabelul sistematizat, se obţine aceeaşi formă de tabel ca şi în cazul caracteristicii discrete. În acest fel s-a ajuns la o unificare a prezentării prin tabele sistematizate a tipurilor discrete şi continue de caracteristici.

Observaţia 3.2.7. Prin stabilirea claselor unei caracteristici continue (care pot fi de lungimi egale sau nu) s-a efectuat o *grupare* a datelor statistice primare, şi care se efectuează şi în cazul caracteristicilor de tip discret, dacă au un număr mare de valori distincte.

Definiţia 3.2.8. *Numim* amplitudinea *clasei definită de intervalul $[a_{i-1}, a_i)$ lungimea acestui interval, $d_i = a_i - a_{i-1}$.*

Observaţia 3.2.9. Când amplitudinile claselor sunt egale, două reguli de stabilire a numărului claselor sunt utilizate mai des:

$$n = 1 + \frac{10}{3} \lg N \quad (\textit{regula lui Sturges}) \quad \text{şi} \quad d = \frac{1}{100} \cdot 8 \left(x_{\max} - x_{\min} \right),$$

unde

$$x_{\max} = \max\{ x'_1, x'_2, \ldots, x'_N \} \quad \text{şi} \quad x_{\min} = \min\{ x'_1, x'_2, \ldots, x'_N \}.$$

În acest fel, din prima regulă, se obţine că

$$d = \frac{b-a}{n} \quad \text{şi} \quad a_i = a + id, \quad i = \overline{0, n}.$$

Când (a, b) este infinit, atunci

$$d = \frac{x_{\max} - x_{\min}}{n} \quad \text{şi} \quad a_i = x_{\min} + id, \quad i = \overline{0, n}.$$

Exemplul 3.2.10. Se consideră un lot de 70 becuri din punct de vedere al caracteristicii X, ce reprezintă durata de viaţă în mii ore. Datele statistice obţinute sunt:

1.318	3.128	2.758	1.583	2.517	2.304	1.155	3.156	2.807	2.879
3.426	1.690	3.537	2.214	2.219	2.072	2.726	1.403	2.493	1.560
3.972	2.637	0.842	2.256	1.708	1.628	2.345	1.855	1.546	3.852
2.128	2.465	2.316	2.262	1.962	1.802	2.230	3.460	1.493	3.093
1.548	2.298	1.875	3.394	1.931	1.179	1.946	1.355	3.006	2.455
1.937	1.977	2.206	1.681	1.960	3.281	2.838	2.525	1.553	2.676
2.500	2.641	1.631	1.864	2.015	2.502	2.444	2.636	2.337	1.966

Vrem să scriem tabelul sistematizat al datelor statistice, considerând clase de amplitudini egale.

Folosind regula lui Sturges avem că numărul n al claselor poate fi calculat cu ajutorul numărului N al datelor statistice prin

$$n = 1 + \frac{10}{3} \lg N = 1 + \frac{10}{3} \times 1.8451,$$

de unde rezultă că $n = 7$. În acest fel avem că amplitudinea d a claselor este

$$d = \frac{x_{\max} - x_{\min}}{n} = \frac{3.972 - 0.842}{7} = 0.45.$$

Când se foloseşte formula

$$d = \frac{1}{100} \cdot 8 \left(x_{\max} - x_{\min} \right)$$

pentru calculul amplitudinii claselor, se obţine $d = 0.25$. În acest caz numărul n al claselor va fi

$$n = \frac{x_{\max} - x_{\min}}{d} = \frac{3.972 - 0.842}{0.25},$$

de unde $n = 12$.

Se observă că folosind cele două metode s-au obţinut valori distincte pentru numărul claselor. Aceste formule au mai mult rolul de a da o primă informaţie relativă la numărul claselor.

În continuare vom considera numărul claselor $n = 11$, $a_i = 0.7 + i \times d$, $i = \overline{0, 11}$, şi $d = 0.3$.

Folosind această grupare se obţine tabelul sistematizat

x	f
$[0.7; 1.0)$	1
$[1.0; 1.3)$	2
$[1.3; 1.6)$	9
$[1.6; 1.9)$	9
$[1.9; 2.2)$	10
$[2.2; 2.5)$	15
$[2.5; 2.8)$	10
$[2.8; 3.1)$	5
$[3.1; 3.4)$	4
$[3.4; 3.7)$	3
$[3.7; 4.0)$	2

sau

x	f
0.85	1
1.15	2
1.45	9
1.75	9
2.05	10
2.35	15
2.65	10
2.95	5
3.25	4
3.55	3
3.85	2

Programul 3.2.11. În urma executării următorului program Matlab

```
load x.dat -ascii
N=length(x);
ns=fix(1+10/3*log10(N));
ds=(max(x)-min(x))/ns;
fprintf('    ns= %2d,    ds= %4.2f\n',ns,ds)
da=8*(max(x)-min(x))/100;
na=fix((max(x)-min(x))/da);
a=0.7:0.3:4; n=length(a)-1;
fprintf('    na= %2d,    da= %4.2f\n',na,da)
for i=1:N
    for j=1:n
        if (a(j)<=x(i))&(x(i)<a(j+1))
            cl(i)=j;
        end
    end
end
t=tabulate(cl);
fprintf('  Tabelul sistematizat\n')
for j=1:n
    t(j,1)=(a(j)+a(j+1))/2;
    fprintf('%10.2f  |  %2d\n',t(j,1),t(j,2))
end
```

se obţin rezultatele:

```
    ns=  7,    ds= 0.45
    na= 12,    da= 0.25
  Tabelul sistematizat
```

```
0.85  |   1
1.15  |   2
1.45  |   9
1.75  |   9
2.05  |  10
2.35  |  15
2.65  |  10
2.95  |   5
3.25  |   4
3.55  |   3
3.85  |   2
```

Se observă că prima comandă citeşte datele statistice din fişierul `x.dat`, care este un fişier `ascii`.

Definiţia 3.2.12. *Numim* frecvenţă relativă *a clasei* x_i *raportul* $p_i = \frac{f_i}{N}$.

Definiţia 3.2.13. *Numim* frecvenţe cumulate ascendente*, respectiv* frecvenţe cumulate descendente *frecvenţele date de relaţiile*

$$F_k = \sum_{i=1}^{k} f_i, \qquad F'_k = \sum_{i=k+1}^{n} f_i, \quad k = \overline{0,n},$$

unde $F_0 = 0$ *şi* $F'_n = 0$.

Observaţia 3.2.14. Pentru frecvenţele relative are loc relaţia

$$\sum_{i=1}^{n} p_i = 1,$$

iar pentru cele cumulate au loc relaţiile $F_k + F'_k = N$, $F_n = N$ şi $F'_0 = N$.

Definiţia 3.2.15. *Numim* distribuţie statistică *a caracteristicii* X *tabloul de forma*

$$X \begin{pmatrix} x_i \\ f_i \end{pmatrix}_{i=\overline{1,n}} \qquad \textit{sau} \qquad X \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i=\overline{1,n}},$$

unde x_i, $i = \overline{1,n}$, *sunt clasele considerate, iar* f_i *şi* p_i, $i = \overline{1,n}$, *sunt respectiv frecvenţele absolute şi frecvenţele relative.*

Exemplul 3.2.16. Considerăm datele statistice de la Exemplul 3.2.10. Atunci distribuţia statistică a caracteristicii X poate fi scrisă, fie cu ajutorul frecvenţelor absolute, fie cu ajutorul frecvenţelor relative. Astfel avem că

$$X \begin{pmatrix} 0.85 & 1.15 & 1.45 & 1.75 & 2.05 & 2.35 & 2.65 & 2.95 & 3.25 & 3.55 & 3.85 \\ 1 & 2 & 9 & 9 & 10 & 15 & 10 & 5 & 4 & 3 & 2 \end{pmatrix}$$

$X \setminus Y$	y_1	y_2	$\dots$	y_n	
x_1	f_{11}	f_{12}	$\dots$	f_{1n}	$f_{1\cdot}$
x_2	f_{21}	f_{22}	$\dots$	f_{2n}	$f_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_m	f_{m1}	f_{m2}	$\dots$	f_{mn}	$f_{m\cdot}$
	$f_{\cdot 1}$	$f_{\cdot 2}$	$\dots$	$f_{\cdot n}$	$f_{\cdot\cdot} = N$

Tabelul 3.1: Tabel de contingenţă

sau

$$X \begin{pmatrix} 0.85 & 1.15 & 1.45 & 1.75 & 2.05 & 2.35 & 2.65 & 2.95 & 3.25 & 3.55 & 3.85 \\ \frac{1}{70} & \frac{2}{70} & \frac{9}{70} & \frac{9}{70} & \frac{10}{70} & \frac{15}{70} & \frac{10}{70} & \frac{5}{70} & \frac{4}{70} & \frac{3}{70} & \frac{2}{70} \end{pmatrix}$$

Definiţia 3.2.17. *Fie colectivitatea $\mathcal{C}$ relativ la care sunt cercetate două caracteristici X şi Y. Numim* tabel de contingenţă*, un tablou care conţine clasele caracteristicilor X şi respectiv Y, împreună cu frecvenţele absolute ale acestor clase.*

Observaţia 3.2.18. Dacă pentru cele două caracteristici X şi Y avem respectiv clasele date prin x_i, $i = \overline{1,m}$, şi y_j, $j = \overline{1,n}$, iar datele primare sunt date prin perechile $(x'_1, y'_1), (x'_2, y'_2), \dots, (x'_N, y'_N)$, atunci tabelul de contingenţă este Tabelul 3.1 unde f_{ij} este *frecvenţa absolută* a apariţiei clasei (x_i, y_j) în datele primare (x'_k, y'_k), $k = \overline{1,N}$. După cum se vede tabelul a fost completat cu o linie şi o coloană ale căror elemente sunt date prin formulele

$$f_{\cdot j} = \sum_{i=1}^{m} f_{ij}, \quad j = \overline{1,n}, \quad f_{i\cdot} = \sum_{j=1}^{n} f_{ij}, \quad i = \overline{1,m},$$

$$f_{\cdot\cdot} = \sum_{j=1}^{n} f_{\cdot j} = \sum_{i=1}^{m} f_{i\cdot} = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = N.$$

Observaţia 3.2.19. Când caracteristicile X şi Y sunt caracteristici cantitative şi între ele există o relaţie de dependenţă, tabelul de contingenţă se numeşte *tabel de corelaţie*.

Exemplul 3.2.20. Un astfel de tabel de corelaţie este prezentat pentru datele statistice ce reprezintă 425 de copii de 10 ani cercetaţi din punct de vedere al înălţimii X (în centimetri) şi al greutăţii Y (în kilograme). Datele au fost trecute în tabelul de corelaţie care urmează

$X \backslash Y$	24	25	26	27	28	29	30	31	32	33	34	35	$f_{i.}$
126	1	1	3	1									6
127		3	4	1		1							9
128	1	4	5	7	1	2	1						21
129		1	2	6	9	6	4	1					29
130			1	7	36	17	6	2	1				70
131			1	6	27	56	39	18	4	1			152
132				2	10	16	26	7	2	1			64
133					1	7	10	11	6	2	1		38
134						1	2	4	7	3	1		18
135							1	2	3	1	1		8
136									1	2	2	1	6
137										1	2	1	4
$f_{.j}$	2	9	16	30	84	106	89	45	24	11	7	2	425

3.2.3 Funcţiile `tabulate` şi `crosstab`

Obţinerea de tabele sistematizate, în Matlab, este posibilă cu ajutorul acestor două funcţii din *Statistics toolbox*.

Funcţia `tabulate`

Apelul funcţiei `tabulate` se poate face prin:

```
tabulate(x)
t=tabulate(x)
```

unde `x` este un vector cu valori pozitive. Prin executarea unei astfel de instrucţiune, se obţine pe ecran respectiv în `t` un tabel statistic sistematizat având trei coloane. Prima coloană conţine valorile distincte ale lui `x`, a doua conţine frecvenţele absolute ale acestor valori distincte, iar ultima reprezintă frecvenţele relative în procente.

Programul 3.2.21. Vom scrie un program Matlab, care genereaă `N` numere aleatoare ce urmează legea uniformă $\mathcal{U}(a,b)$, după care construieşte tabelul sistematizat, considerând numărul claselor `n` dat prin regula lui Sturges, iar clasele fiind identificate prin mijloacele lor.

```
clear
a=input('a:'); b=input('b (a<b):');
N=input('N='); n=fix(1+10/3*log10(N));
xx=unifrnd(a,b,1,N); c=a:(b-a)/n:b;
for i=1:N
    for j=1:n
        if c(j)<=xx(i) & xx(i)<c(j+1)
           x(i)=j;
        end
    end
end
```

```
t=tabulate(x);
for j=1:n
    t(j,1) =(c(j)+c(j+1))/2;
end
fprintf('  Clasa     f      frel\n')
fprintf(' _____________________\n')
fprintf('%7.3f %5d %8.2f%%\n',t')
```

Executarea programului, cu datele de intrare `a=0`, `b=5` şi `N=200`, produce tabel sistematizat:

```
  Clasa     f      frel
 ____________________
  0.313    22    11.00%
  0.938    27    13.50%
  1.563    27    13.50%
  2.188    32    16.00%
  2.813    27    13.50%
  3.438    23    11.50%
  4.063    25    12.50%
  4.688    17     8.50%
```

Funcţia `crosstab`

Apelul funcţiei `crosstab` se poate face prin:

```
crosstab(x,y,z,...)
t=crosstab(x,y,z,...)
```

unde `x`, `y`, `z`, ... sunt vectori cu valori înntregi pozitive având aceleaşi lungimi. Prin executarea unei astfel de instrucţiune, se obţine pe ecran respectiv în `t` un tabel statistic sistematizat `t`, cu atâtea intrări câţi parametri (vectori) există. Dacă sunt doi parametri, `x` şi `y`, se obţine tabelul de contingenţă.

Programul 3.2.22. Programul Matlab care urmează, generează `N` vectori aleatori ce urmează legea normală bidimensională, după care construieşte tabelul de corelaţie cu `m` clase în raport cu prima variabilă, respectiv `n` clase în raport cu a doua variabilă.

```
clear, mu(1)=input('m1='); mu(2)=input('m2=');
v(1,1)=input('sigma1^2=');
v(2,2)=input('sigma2^2=');
v(1,2)=input('Cov(X,Y)='); v(2,1) =v(1,2);
N=input('N='); m=input('m='); n=input('n=');
if det(v) <= 0
   error('Matricea v nu e pozitiv definita!')
end
X=mvnrnd(mu,v,N); xx=X(:,1); yy=X(:,2);
xmin=min(xx); xmax=max(xx);
ymin=min(yy); ymax=max(yy);
cx =xmin:(xmax-xmin)/m:xmax;
cy =ymin:(ymax-ymin)/n:ymax;
```

```
for k=1:N
   for i=1:m-1
     if (cx(i)<=xx(k)) & (xx(k)<cx(i+1))
        x(k)=i;
     end
   end
   if (cx(m)<=xx(k)) & (xx(k)<=cx(m+1))
       x(k)=m;
   end
   for j=1:n-1
     if (cy(j)<=yy(k)) & (yy(k)<cy(j+1))
        y(k)=j;
     end
   end
   if (cy(n)<=yy(k)) & (yy(k)<=cy(n+1))
       y(k)=n;
   end
end
t=crosstab(x,y);
disp(t)
```

Pentru `N=100`, `mu=(5,10)`, `v` $= \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}$, `m=6` şi `n=4`, se obţine tabelul de corelaţie:

```
     1     2     0     0
     2     7     0     0
     5    14     6     1
     2    17     8     1
     1    12    12     3
     1     1     1     3
```

3.2.4 Funcţiile `caseread`, `casewrite`, `tblread` şi `tblwrite`

Pentru completarea tabelului de contingenţă construit prin funcţia `crosstab`, se poate face apel la aceste funcţii.

Funcţia `casewrite`

Apelul funcţiei `casewrite` are sintaxa

```
casewrite(obs,'file')
```

şi are ca efect scrierea datelor din matricea `obs` de tip caracter în fişierul cu numele `file`.

Funcţia `caseread`

Apelul funcţiei `caseread` are sintaxa

```
obs=caseread('file')
```

şi are ca efect citirea datelor din fişierul cu numele `file` în matricea `obs` de tip caracter.

Funcţia `tblwrite`

Sintaxa funcţiei `tblwrite` este:

```
tblwrite(t,va,obs,'file')
```

şi are ca efect scrierea în fişierul cu numele `file` a matricelor `va`, `obs` de tip caracter şi a datelor din matricea numerică `t`.

Matricele `va` şi `obs` trebuie să aibă atâtea linii câte coloane, respectiv câte linii are matricea `t`. Liniile celor două matrice de tip caracter vor reprezenta etichete pentru coloanele, respectiv liniile matricei `t`, şi drept urmare vor fi aşezate în dreptul coloanelor şi liniilor corespunzătoare.

Funcţia `tblread`

Sintaxa funcţiei `tblread` este:

```
[t,va,obs]=tblread('file')
```

şi are ca efect citirea din fişierul cu numele `file` a matricelor `va`, `obs` de tip caracter şi a datelor din matricea numerică `t`.

Programul 3.2.23. Programul care urmează completează Programul 3.2.22, pentru obţinerea în tabelul de contingenţă şi a mijloacelor claselor pentru cele două variabile.

```
. . . . . .
t=crosstab(x,y);
fy=fopen('ly.txt','w+');
for j=1:n
   laby(j)=(cy(j)+cy(j+1))/2;
   ly=num2str(laby(j));
   fprintf(fy,'%s\n',ly);
end
fx=fopen('lx.txt','w+');
for i=1:m
   labx(i)=(cx(i)+cx(i+1))/2;
   lx=num2str(labx(i));
   fprintf(fx,'%s\n',lx);
end
va=caseread('ly.txt'); obs=caseread('lx.txt');
tblwrite(t,va,obs,'clase.txt');
type clase.txt, fclose('all');
```

În urma executării programului cu aceleaşi date de intrare folosite în Programul 3.2.22, prin instrucţiune `type` este afişat pe ecran tabelul de contingenţă:

```
          7.5244    9.7152    11.906    14.0967
2.5895    4         6         2         0
3.8424    4         16        7         0
5.0952    1         13        12        0
6.348     2         12        9         1
7.6008    1         1         4         3
8.8537    0         0         1         1
```

Se observă că datele din tabelul obţinut nu sunt aceleaşi cu cele obţinute prin executarea Programului 3.2.22, deşi se dau aceleaşi date de intrare. Lucru de aşteptat, având în vedere că se generează un alt set de numere aleatoare.

Se observă de asemenea, în noul program, că apar instrucţiunile `fopen`, `fclose`, `caseread`, `tblwrite`.

Instrucţiunea `fopen` este folosită pentru deschiderea fişierelor `lx.txt` şi `ly.txt`, care vor conţine etichetele claselor celor două variabile. Acestea sunt date de mijloacele claselor şi sunt transformate în şiruri de caractere prin comanda `num2str`. Parametrul `'w+'`, are ca efect crearea unui fişier nou sau actualizarea unuia deja existent, după ce a fost anulat conţinutul său, în vederea scrierii sau citirii în respectivul fişier. Evident instrucţiunea `fclose` este folosită în vederea închiderii fişierelor.

Prin instrucţiunile `caseread` sunt citite etichetele claselor din fişierele `lx.txt` şi `ly.txt`, care apoi sunt folosite pentru scrierea acestora în fişierul `clase.txt`, cu ajutorul comenzii `tblwrite`, împreună cu elementele tabelului de contingenţă.

3.2.5 Reprezentări grafice

Definiţia 3.2.24. *Numim* diagramă prin batoane (bare) *a unei distribuţii statistice X de tip discret, reprezentarea grafică într-un sistem de axe rectangulare a segmentelor (batoanelor) date prin $\{ (x_i, y) \mid 0 \leqslant y \leqslant \alpha\, f_i \}$, $i = \overline{1, n}$, $\alpha > 0$ fiind un factor de proporţionalitate, iar f_i este frecvenţa absolută a valorii x_i.*

Definiţia 3.2.25. *Numim* diagramă cumulativă (ascendentă) *a unei distribuţii statistice X de tip discret, linia poligonală care uneşte punctele de coordonate $(x_1, \alpha F_0)$, $(x_1, \alpha F_1)$, $(x_2, \alpha F_1)$, $(x_2, \alpha F_2)$, $(x_3, \alpha F_2)$, . . ., $(x_n, \alpha F_n)$, unde F_i este frecvenţa cumulată (ascendentă) ataşată valorii x_i, iar $\alpha > 0$ este un factor de proporţionalitate.*

Definiţia 3.2.26. *Numim* histograma *unei distribuţii statistice X de tip continuu, diagrama obţinută prin construirea de dreptunghiuri având drept baze clasele distribuţiei statistice şi înălţimile astfel considerate încât ariile dreptunghiurilor să fie proporţionale cu frecvenţele claselor.*

Observaţia 3.2.27. Când clasele distribuţiei statistice sunt de amplitudini egale, atunci înălţimile dreptunghiurilor histogramei sunt proporţionale cu frecvenţele claselor. Dacă factorul de proporţionalitate este $\frac{1}{N}$, atunci se obţine histograma frecvenţelor relative.

Se ştie că aria delimitată de curba ce reprezintă graficul unei densităţi de probabilitate şi axa absciselor este 1. Din acest motiv, este de preferat ca factorul de proporţionalitate în construirea histogramei frecvenţelor relative să fie astfel ales încît

aria totală a dreptunghiurilor histogramei să fie de asemenea 1. Se verifică uşor că dacă N este volumul datelor, iar clasa $[a_{i-1}, a_i)$ are amplitudinea $h_i = a_i - a_{i-1}$, pentru $i = \overline{1,n}$, atunci funcţia în scară ce delimitează aceste dreptunghiuri este dată prin:

$$\hat{f}(x) = \frac{f_i}{Nh_i}, \quad x \in [a_{i-1}, a_i), \quad i = \overline{1,n}.$$

Exemplul 3.2.28. Dacă se consideră datele statistice din Exemplul 3.2.10, şi având în vedere sistematizarea acestora, atunci histograma frecvenţelor absolute este prezentată în Figura 3.1.

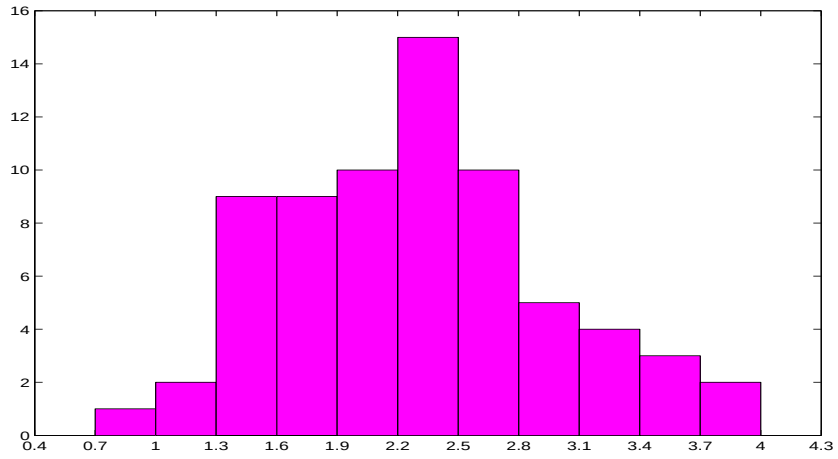

Figura 3.1: Histograma frecvenţelor absolute

Metoda ferestrei mobile

Histograma frecvenţelor relative a distribuţiei statistice reprezintă o aproximare destul de rudimentară a graficului densităţii de probabilitate a caracteristicii X. O ameliorare a acestei aproximări se obţine dacă se aplică *metoda ferestrei mobile*.

Metoda ferestrei mobile constă în construirea clasei $\left(x - \frac{h}{2}, x + \frac{h}{2}\right)$, $h > 0$, pentru fiecare $x \in (a,b)$, şi determinarea frecvenţei f_x a acestei clase, cu ajutorul căreia se obţine curba dată prin $\hat{f}(x) = \frac{f_x}{Nh}$, $\quad x \in (a,b)$.

Dacă se consideră funcţia indicatoare K a intervalului $\left(-\frac{1}{2}, \frac{1}{2}\right)$, adică

$$K(x) = \begin{cases} 1, & \text{dacă } x \in \left(-\frac{1}{2}, \frac{1}{2}\right), \\ 0, & \text{dacă } x \notin \left(-\frac{1}{2}, \frac{1}{2}\right), \end{cases} \tag{3.2.1}$$

atunci

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i'}{h}\right).$$

Pentru a obţine o formă mai regulară pentru $\hat{f}$ se pot considera şi alte tipuri de *funcţie nucleu* K, care de regulă este o densitate de probabilitate simetrică.

Mai des se utilizează *nucleul Gaussian*

$$K(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}, \tag{3.2.2}$$

sau *nucleul parabolic* (*Epanechnikov*)

$$K(x) = \begin{cases} \dfrac{3}{4\sqrt{5}} \left(1 - \dfrac{x^2}{5}\right), & \text{dacă } |x| < \sqrt{5}, \\ 0, & \text{dacă } |x| \geqslant \sqrt{5}. \end{cases} \tag{3.2.3}$$

Alte nuclee ce s–ar putea folosi sunt:

$$K(x) = \begin{cases} \dfrac{15}{16} \left(1 - x^2\right)^2, & \text{dacă } |x| < 1, \\ 0, & \text{dacă } |x| \geqslant 1, \end{cases} \tag{3.2.4}$$

$$K(x) = \begin{cases} 1 + \cos(2\pi x) & \text{dacă } 0 < x < 1, \\ 0, & \text{dacă } |x| \geqslant 1, \end{cases} \tag{3.2.5}$$

$$K(x) = \begin{cases} 1 - |x| & \text{dacă } |x| < 1, \\ 0, & \text{dacă } |x| \geqslant 1. \end{cases} \tag{3.2.6}$$

Programul 3.2.29. Programul ce urmează generează `N` numere aleatoare ce urmează legea normală $\mathcal{N}(0,1)$, după care, folosind acelaşi pas `h` pozitiv, reprezintă grafic funcţia $\hat{f}$ împreună cu densitatea de probabilitate a legii normale, pentru fiecare din cele şase nuclee, mai sus definite.

Calculul valorilor celor şase nuclee se face cu următoarea funcţie Matlab:

```
function y = K(k,x)
switch k
case 1
    y=(abs(x)<1/2);
case 2
    y=1/(sqrt(2*pi))*exp(-x.^2/2);
case 3
    y=(abs(x)<sqrt(5));
```

```
    y=3/(4*sqrt(5))*(1-x.^2/5).*y;
case 4
    y=(abs(x)<1);
    y=15/16*(1-x.^2).^2.*y;
case 5
    y=((x>0)&(x<1));
    y=(1+cos(2*pi*x)).*y;
case 6
    y=(abs(x)<1); y=abs(x).*y;
otherwise
    error('Eroare')
end
```

iar programul Matlab este

```
clf,clear
N=input('N='); h=input('h=');
X=randn(1,N); x=-3:0.01:3;
n=length(x); p=pdf('norm',x,0,1);
for ka=1:6
s=zeros(1,n);
for i=1:N
    s=s+K(ka,(x-X(i))/h);
end
f=s/(N*h);
subplot(3,2,ka), plot(x,f,'-',x,p,'--')
end
```

În urma executării programului, cu valorile `N=500` şi `h=1`, se obţin graficele din Figura 3.2.

Definiţia 3.2.30. *Numim* poligonul frecvenţelor *al unei distribuţii statistice X de tip continuu, poligonul obţinut prin unirea punctelor de coordonate $(x_i, \alpha_i f_i)$, $i = \overline{1,n}$, unde α_i este factor de proporţionalitate, iar f_i este frecvenţa clasei x_i.*

Exemplul 3.2.31. Dacă se consideră datele statistice din Exemplul 3.2.10, şi având în vedere sistematizarea acestora, atunci poligonul frecvenţelor absolute este prezentată în Figura 3.3.

Definiţia 3.2.32. *Numim* diagrame integrale (cumulative) *ale frecvenţelor cumulate ascendente, respectiv descendente, relative la distribuţia statistică X de tip continuu, liniile poligonale obţinute prin unirea punctelor de coordonate (a_k, F_k), $k = \overline{0,n}$, şi respectiv (a_k, F'_k), $k = \overline{0,n}$.*

Exemplul 3.2.33. Dacă se consideră datele statistice din Exemplul 3.2.10, şi având în vedere sistematizarea acestora, atunci diagramele integrale ale frecvenţelor cumulate (ascendente şi descendente) sunt prezentate în Figura 3.4.

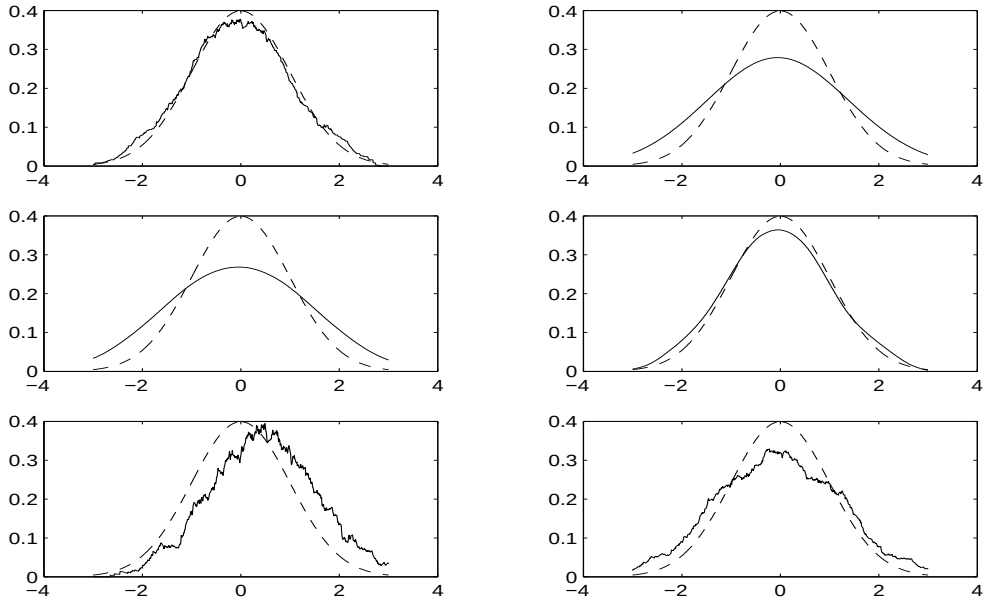

Figura 3.2: Grafice pentru $\hat{f}\,(x)$

Pentru trasarea diagramelor integrale, în tabelul următor calculăm frecvenţele absolute ascendente şi respectiv descendente, conform formulelor

$$F_k = \sum_{i=1}^{k} f_i, \quad F'_k = \sum_{i=k+1}^{n} f_i, \quad k = \overline{0,n}.$$

x	a_0-a_1	a_1-a_2	a_2-a_3	a_3-a_4	a_4-a_5	a_5-a_6	a_6-a_7	a_7-a_8	a_8-a_9	a_9-a_{10}	$a_{10}-a_{11}$
f	1	2	9	9	10	15	10	5	4	3	2
F	1	3	12	21	31	46	56	61	65	68	70
F'	69	67	58	49	39	24	14	9	5	2	0

Definiţia 3.2.34. *Numim* nor statistic *ataşat caracteristicilor X şi Y, punctele din plan obţinute prin reprezentarea grafică a datelor primare (x'_k, y'_k), $k = \overline{1,N}$.*

Observaţia 3.2.35. Norul statistic este utilizat pentru observarea formei legăturii funcţionale care există între cele două caracteristici. Amintim câteva legături mai des întâlnite:

$$y = ax + b, \quad x \in \mathbb{R}, \quad a, b \in \mathbb{R},$$
$$y = ax^b, \quad x > 0, \quad a, b \in \mathbb{R},$$

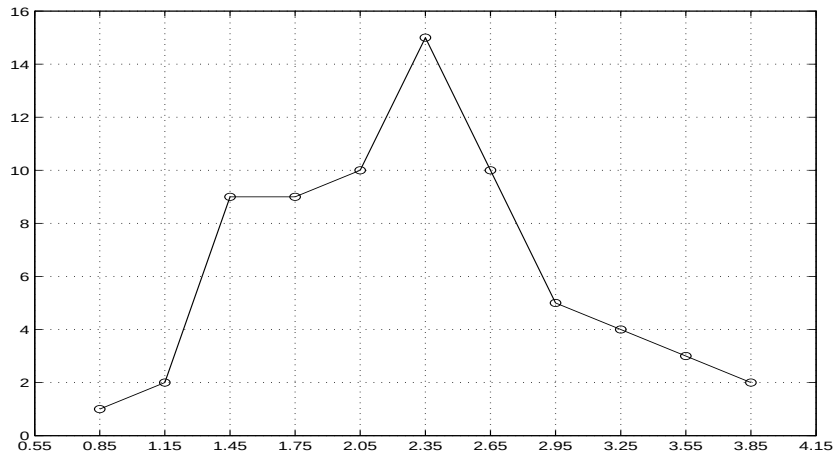

Figura 3.3: Poligonul frecvenţelor absolute

$$
\begin{aligned}
y =& ab^x, \quad x > 0, \quad a > 0,\ b > 0,\\
y =& ax^2 + bx + c, \quad x \in \mathbb{R}, \quad a, b, c \in \mathbb{R},\\
y =& a + \frac{b}{x}, \quad x \neq 0, \quad a, b \in \mathbb{R}.
\end{aligned}
$$

Dacă legătura liniară $y = ax + b$ este uşor de observat din norul statistic, pentru celelalte se încearcă liniarizarea. De exemplu, dacă $y = ax^b$, prin logaritmare se ajunge la $\log y = \log a + b \log x$. Se notează $\log y = v$, $\log a = A$, $\log x = u$ şi se obţine legătura liniară $v = A + bu$, pentru $\log X$ şi $\log Y$. Pentru aceasta se va reprezenta grafic norul statistic dat de punctele $(\log x'_k, \log y'_k)$, $k = \overline{1, N}$. Dacă punctele din acest nor statistic sunt situate în jurul unei drepte, atunci se poate considera că legătura dintre X şi Y este de forma $y = ab^x$.

3.2.6 Funcţii Matlab pentru reprezentarea grafică a datelor statistice

Sistemul Matlab, atât prin sistemul de bază, cât şi prin *Statistics toolbox*, dispune de funcţii pentru reprezentarea grafică a datelor statistice.

Cazul discret

Dacă în cercetare se consideră o caracteristică (variabilă) X de tip discret, funcţiile Matlab, deja prezentate, `bar` şi `stairs` se pot utiliza cu succes.

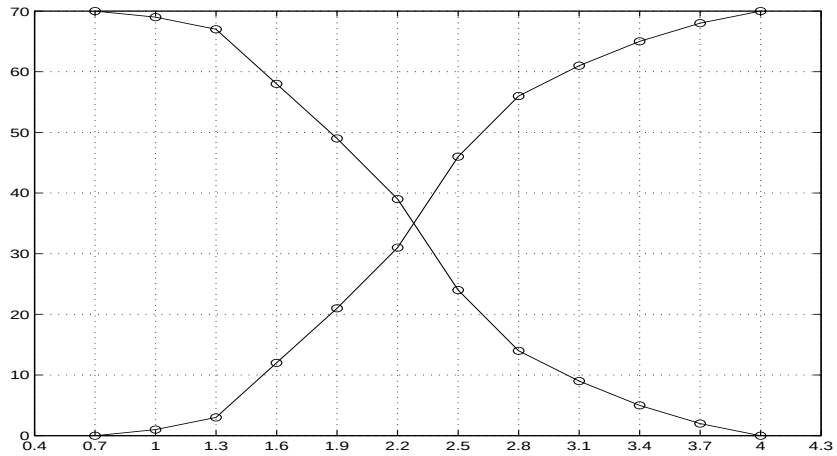

Figura 3.4: Diagramele integrale ale frecvenţelor absolute

Pentru exemplificarea modului de utilizare a acestor funcţii Matlab, vom scrie două funcţii Matlab pentru construirea diagramei prin batoane şi respectiv pentru construirea diagramei cumulative ascendente.

Funcţia 3.2.36. Funcţia Matlab pe care o scriem, va genera N numere aleatoare ce urmează o lege de probabilitate de tip discret, va construi diagrama prin batoane pentru datele astfel obţinute, iar pe aceeaşi figură va reprezenta, pentru comparaţie, tot prin batoane, funcţia de probabilitate a legii de probabilitate considerate. Pentru o vizualizare corectă a comparaţiei, se impune, fie amplificarea funcţiei de probabilitate cu factorul N, fie împărţirea frecvenţelor distribuţiei statistice cu N, adică prin considerarea frecvenţelor relative. În cele ce urmează vom considera prima variantă.

```
  function diag1(N,lege)
% Functia diag1 produce diagrama prin batoane
% N - reprezinta volumul datelor ce urmeaza a fi generate
% lege - reprezinta denumirea Matlab a legii de probabilitate
clf, colormap spring
switch lege
case 'unid'
    n = input('n (parametrul legii uniforme discrete):');
    x = random(lege,n,1,N);
    t = tabulate(x);
    P =N*pdf(lege,t(:,1),n);
    bar(t(:,1),[t(:,2),P])
    legend('Frecvente observate', 'Frecvente teoretice',1)
    return
case 'bino'
    n = input('n='); p = input('p=');
    x = random(lege,n,p,1,N);
```

```
        t = tabulate(x+1);
        P =N*pdf(lege,t(:,1)-1,n,p);
    case 'hyge'
        M = input('M='); K = input('K=');
        n = input('n=');
        x = random(lege,M,K,n,1,N);
        t = tabulate(x+1);
        P =N*pdf(lege,t(:,1)-1,M,K,n);
    case 'poiss'
        lambda = input('lambda=');
        x = random(lege,lambda,1,N);
        t = tabulate(x+1);
        P =N*pdf(lege,t(:,1)-1,lambda);
    case 'nbin'
        r = input('r='); p = input('p=');
        x = random(lege,r,p,1,N);
        t = tabulate(x+1);
        P =N*pdf(lege,t(:,1)-1,r,p);
    case 'geo'
        p = input('p=');
        x = random(lege,p,1,N);
        t = tabulate(x+1);
        P =N*pdf(lege,t(:,1)-1,p);
    otherwise
        error('Lege discreta necunoscuta')
    end
        bar(t(:,1),[t(:,2),P])
        legend('Frecvente observate', 'Frecvente teoretice',1)
```

Apelul funcţiei `diag1` se face prin

```
>>diag1(N,'lege')
```

unde valorile pentru `N` şi `lege`, fie că sunt precizate în acest apel, fie sunt precizate înainte. De exemplu, comanda

```
>>diag1(50,'poiss')
```

are ca efect apelul funcţiei pentru legea lui Poisson, iar pe ecran se va cere introducerea parametrului `lambda`, după care pe ecran va fi reprezentat graficul din Figura 3.5, în cazul în care `lambda=7`. Să remarcăm totuşi că la un nou apel, cu aceeaşi parametri, graficul diferă, deoarece sunt generate alte numere aleatoare.

Funcţia 3.2.37. Următoarea funcţie Matlab, va genera N numere aleatoare ce urmează o lege de probabilitate de tip discret, va construi diagrama cumulativă ascendentă pentru datele astfel obţinute, iar pe aceeaşi figură va reprezenta, pentru comparaţie, funcţia de repartiţie a legii de probabilitate considerate. Pentru o vizualizare corectă a comparaţiei, se impune, fie amplificarea funcţiei de repartiţie cu factorul `N`, fie împărţirea frecvenţelor cumulate ale distribuţiei statistice cu `N`, adică prin considerarea frecvenţelor relative cumulate. În cele ce urmează vom considera a doua variantă.

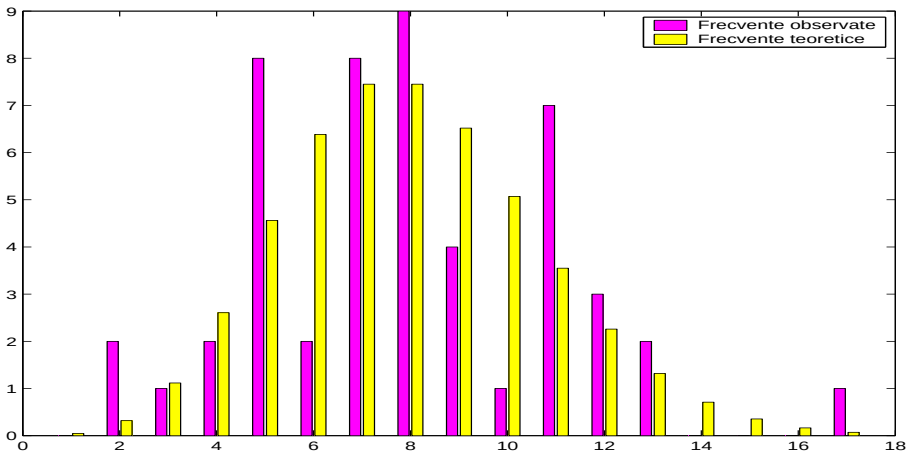

Figura 3.5: Legea $\mathcal{P}o\,(7)$

```
function diag2(N,lege)
% Functia diag2 produce diagrama cumulativa
% N - reprezinta volumul datelor generate
% lege - reprezinta denumirea legii de probabilitate
clf
switch lege
case 'unid'
  n=input('n (parametrul legii uniforme discrete):');
  x=random(lege,n,1,N);
  t=tabulate(x);
  xx=[t(1,1)-1;t(:,1);t(end,1)+1];
  cf=[0;cumsum(t(:,2));N]/N;
  P=cdf(lege,xx,n);
  stairs(xx,cf,'k-'),hold on,
  stairs(xx,P,'k-.'),return
case 'bino'
  n=input('n='); p=input('p=');
  x=random(lege,n,p,1,N);
  t=tabulate(x+1);
  xx=[t(1,1)-2;t(:,1)-1;t(end,1)];
  cf=[0;cumsum(t(:,2));N]/N;
  P=[0;cdf(lege,t(:,1)-1,n,p);1];
case 'hyge'
  M=input('M='); K=input('K=');
  n=input('n=');
  x=random(lege,M,K,n,1,N);
  t=tabulate(x+1);
  xx=[t(1,1)-2;t(:,1)-1;t(end,1)];
```

```
    cf=[0;cumsum(t(:,2));N]/N;
    P=[0;cdf(lege,t(:,1)-1,M,K,n);1];
  case 'poiss'
    lambda=input('lambda=');
    x=random(lege,lambda,1,N);
    t=tabulate(x+1);
    xx=[t(1,1)-2;t(:,1)-1;t(end,1)];
    cf=[0;cumsum(t(:,2));N]/N;
    P=[0;cdf(lege,t(:,1)-1,lambda);1];
  case 'nbin'
    r=input('r='); p=input('p=');
    x=random(lege,r,p,1,N);
    t=tabulate(x+1);
    xx=[t(1,1)-2;t(:,1)-1;t(end,1)];
    cf=[0;cumsum(t(:,2));N]/N;
    P=[0;cdf(lege,t(:,1)-1,r,p);1];
  case 'geo'
    p=input('p=');
    x=random(lege,p,1,N);
    t=tabulate(x+1);
    xx=[t(1,1)-2;t(:,1)-1;t(end,1)];
    cf=[0;cumsum(t(:,2));N]/N;
    P=[0;cdf(lege,t(:,1)-1,p);1];
  otherwise
      error('Lege discreta necunoscuta')
  end
  stairs(xx,cf,'k-'),hold on,
  stairs(xx,P,'k-.')
```

Apelul funcţiei `diag2` se face analog apelului funcţiei `diag1` prezentată în Funcţia 3.2.36:

```
>>diag2(N,'lege')
```

De exemplu, comanda

```
>>diag2(50,'bino')
```

are ca efect apelul funcţiei pentru legea binomială, iar pe ecran se vor cere introducerea parametrilor `n` şi `p`, după care pe ecran va fi reprezentat graficul din Figura 3.6, în cazul în care `n=7` şi `p=0.4`. Diagrama cumulativă este reprezentată prin line continuă, iar funcţia de repartiţie teoretică prin linie-punct. Şi aici remarcăm că la un nou apel, cu aceeaşi parametri, graficul diferă, deoarece sunt generate alte numere aleatoare.

Cazul continuu

Reprezentări grafice pentru distribuţia statistică a unei caracteristici (variabile) de tip continuu se obţin, fie prin utilizarea funcţiei `plot` şi `bar`, fie prin funcţiile `hist`,

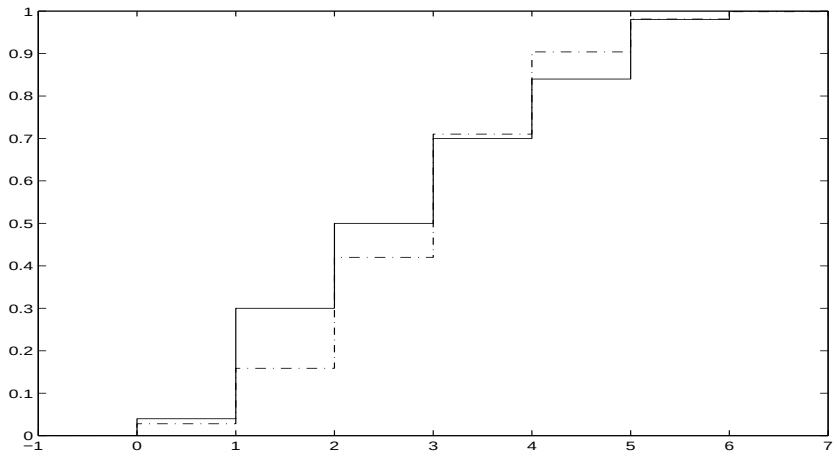

Figura 3.6: Legea $\mathcal{B}(7, 0.4)$

`histc`. Remarcăm faptul că `hist` şi `histc` sunt funcţii specifice sistemului de bază Matlab.

Funcţia `hist`

Modurile de apel ale funcţiei `hist` sunt:

```
hist(y)
hist(y,nb)
hist(y,x)
n=hist(y)
n=hist(y,nb)
n=hist(y,x)
[n,x]=hist(y)
[n,x]=hist(y,nb)
```

Primele trei forme au ca efect reprezentarea grafică a histogramei corespunzătoare datelor conţinute de vectorul `y`. Parametrul `nb` reprezintă numărul claselor de amplitudini egale, iar dacă acesta lipseşte, se consideră `nb=10`. Amplitudinea acestor clase este obţinută prin împărţirea lungimii intervalului definit prin valoarea minimă şi valoarea maximă ale componentelor vectorului `y` la `nb`. Când se utilizează parametrul `x`, acesta trebuie să conţină mijloacele claselor.

Celelalte forme nu produc reprezentări grafice pentru histograme, ci doar calculează frecvenţele absolute şi respectiv mijloacele claselor în vectorii `n` şi `x`, care au lungimea dată prin `nb`.

Toate formele acceptă parametrul `y` ca fiind matrice, caz în care operaţia se execută pentru fiecare coloană a matricei.

Funcţia `histc`

Apelul funcţiei `histc` se poate face prin:

```
n=histc(y,a)
[n,nb]=hist(y,a)
```

care au ca efect, pentru datele vectorului `y`, obţinerea în `n` a frecvenţelor claselor precizate prin componentele vectorului `a`, care se impune a fi ordonate crescător. Valorile lui `y` care sunt înafara componentelor lui `a` nu sunt luate în considerare. De aceea e de dorit ca să fie folosite valorile `-inf` şi `inf` pentru a fi incluse toate valorile lui `y`.

Dacă `y` este o matrice, atunci funcţia operează pentru fiecare coloană în parte.

Parametrul `nb` va conţine numerele de ordine ale claselor în care au intrat fiecare din elementele lui `y`. Pentru elementele care nu au fost clasificate se obţine valoarea `0`.

Să remarcăm că Figura 3.1, Figura 3.3 şi Figura 3.4 au fost realizate, respectiv cu programele Matlab:

```
clear, clf, load x.dat -ascii2
N=length(x);
a=0.7:0.3:4; n=length(a)-1;
for j=1:n
    mij(j)=(a(j)+a(j+1))/2;
end
fr=histc(x,a); fr=fr(1:end-1);
bar(mij,fr,1)
axis([0.4 4.3 0 16])
set(gca,'xtick',[0.4:0.3:4.3]), colormap spring

clear, clf, load x.dat -ascii
N=length(x);
a=0.7:0.3:4; n=length(a)-1;
for j=1:n
    mij(j)=(a(j)+a(j+1))/2;
end
fr=histc(x,a); fr=fr(1:end-1);
plot(mij,fr,'k-o')
axis([0.55 4.15 0 16])
set(gca,'xtick',[0.55:0.3:4.15]), grid on

clear, clf, load x.dat -ascii
N=length(x);
a=0.7:0.3:4; n=length(a)-1;
fr=histc(x,a); fr=fr(1:end-1);
FA=[0;cumsum(fr)], FD=N-FA,
```

```
plot(a,FA,'k-o',a,FD,'k-o')
axis([0.4 4.3 0 70])
set(gca,'xtick',[0.4:0.3:4.3]), grid on
```

care utilizează funcţiile `histc` şi `bar`.

Funcţia 3.2.38. Următoarea funcţie Matlab, va genera N numere aleatoare ce urmează o lege de probabilitate de tip continuu şi va construi histograma pentru datele astfel obţinute. Într–o figură alăturată se va construi histograma folosind funcţia bar, împreună cu densitatea de probabilitate a legii de probabilitate considerate. Pentru o vizualizare corectă a comparaţiei, se impune, împărţirea frecvenţelor absolute cu `Nh`, unde `h` este amplitudinea claselor.

```
function hist0(N,lege)
% Functia hist0 produce histograma
% N - reprezinta volumul datelor generate
% lege - reprezinta denumirea legii de probabilitate
clf, colormap spring
n=fix(1+10/3*log10(N));
switch lege
case 'unif'
    a=input('a:'); b=input('b(a<b):');
    x=random(lege,a,b,1,N);
    [f,c]=hist(x,n); h=c(2)-c(1);
    t=c(1)-h:0.01:c(end)+h;
    y=pdf(lege,t,a,b);
case 'norm'
    mu=input('mu='); s=input('sigma=');
    x=random(lege,mu,s,1,N);
    [f,c]=hist(x,n); h=c(2)-c(1);
    t=c(1)-h:0.01:c(end)+h;
    y=pdf(lege,t,mu,s);
case 'logn'
    mu=input('mu='); s=input('sigma=');
    x=random(lege,mu,s,1,N);
    [f,c]=hist(x,n); h=c(2)-c(1);
    t=c(1)-h:0.01:c(end)+h;
    y=pdf(lege,t,mu,s);
case 'gam'
    a=input('a='); b=input('b=');
    x=random(lege,a,b,1,N);
    [f,c]=hist(x,n); h=c(2)-c(1);
    t=c(1)-h:0.01:c(end)+h;
    y=pdf(lege,t,a,b);
case 'exp'
    mu=input('mu=');
    x=random(lege,mu,1,N);
    [f,c]=hist(x,n); h=c(2)-c(1);
    t=c(1)-h:0.01:c(end)+h;
    y=pdf(lege,t,mu);
```

```
case 'beta'
    a=input('a='); b=input('b=');
    x=random(lege,a,b,1,N);
    [f,c]=hist(x,n); h=c(2)-c(1);
    t=c(1)-h:0.01:c(end)+h;
    y=pdf(lege,t,a,b);
case 'weib'
    a=input('a='); b=input('b=')
    x=random(lege,a,b,1,N);
    [f,c]=hist(x,n); h=c(2)-c(1);
    t=c(1)-h:0.01:c(end)+h;
    y=pdf(lege,t,a,b);
case 'rayl'
    b=input('b=');
    x=random(lege,b,1,N);
    [f,c]=hist(x,n); h=c(2)-c(1);
    t=c(1)-h:0.01:c(end)+h;
    y=pdf(lege,t,a);
otherwise
    error('Lege continua necunoscuta')
end
subplot(1,2,1), hist(x,n)
subplot(1,2,2), bar(c,f/(N*h),1)
hold on, plot(t,y,'k-')
```

Apelul funcţiei `hist0` se face prin

```
>>hist0(N,'lege')
```

De exemplu, comanda

```
>>hist0(500,'norm')
```

are ca efect apelul funcţiei pentru legea normală, iar pe ecran se va cere introducerea parametrilor `mu` şi `sigma`, după care pe ecran va fi reprezentat graficul din Figura 3.7, în cazul în care `mu=0` şi `sigma=1`.

Funcţia `scatter`

Sistemul Matlab este prevăzut cu funcţia `scatter` pentru reprezentarea norului statistic în cazul bidimensional. Apelul funcţiei se face prin:

```
scatter(x,y)
scatter(x,y,s)
scatter(x,y,s,c)
scatter(x,y,s,c,m)
scatter(x,y,s,c,m,'filled')
```

Efectul executării unei astfel de instrucţinui este producerea norului statistic precizat prin vectorii `x` şi `y`, care au aceeaşi lungime.

Parametrul opţional `s` specifică mărimea marcajelor (cerculeţe) punctelor norului. Acesta poate să fie un scalar, în care caz toate marcajele vor avea aceleaşi mărimi

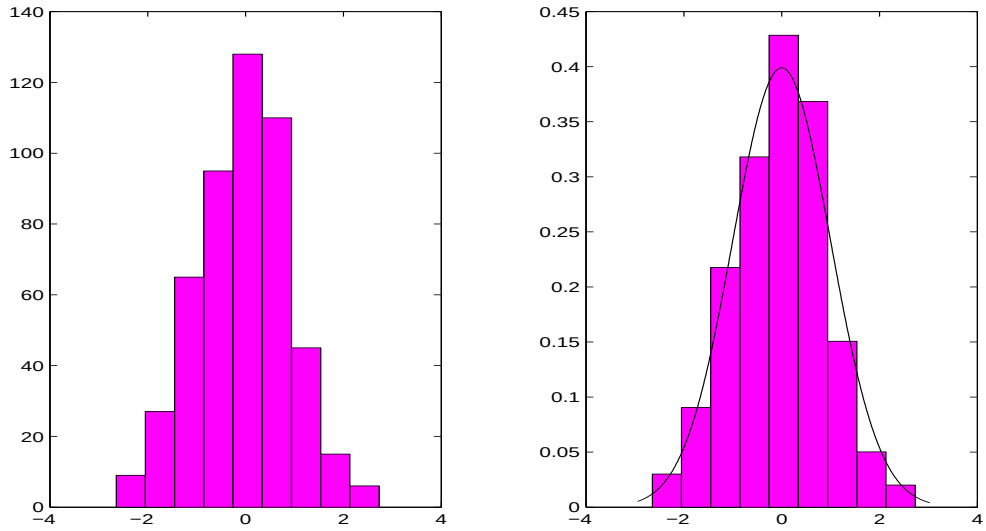

Figura 3.7: Legea $\mathcal{N}(0,1)$

sau poate să fie un vector de aceeaşi lungime cu `x` şi `y`, caz în care se pot specifica mărimi diferite pentru puncte diferite.

Parametrul `c` specifică culoarea marcajelor şi poate lua una din cele opt culori folosite şi la funcţia `plot`. Dacă `c` este vector, atunci trebuie să aibă aceeaşi lungime cu ceilalţi parametri de tip vector, caz în care se pot specifica culorile fiecărui marcaj în parte.

Parametrul `m` permite înlocuirea marcajului implicit (cerculeţ), prin marcajul precizat prin `m`.

Prezenţa opţiunii `'filled'` atrage după sine umbrirea interiorului marcajului.

Remarcăm faptul că se poate folosi cu succes funcţia `plot`, dacă se foloseşte un singur tip de marcaj.

Funcţia `scatter3`

Sistemul Matlab este prevăzut cu funcţia `scatter3` pentru reprezentarea norului statistic în cazul tridimensional. Apelul funcţiei se face cu aceleaşi instrucţiuni ca şi în cazul funcţiei `scatter`, doar că se mai inserează încă un parametru `z`, a treia dimensiune, după primii doi parametri `x` şi `y`.

Funcţia `gscatter`

Statistics toolbox dispune de funcţia `gscatter` pentru reprezentarea norului statistic în cazul bidimensional pe grupe de puncte. Apelul funcţiei se face prin:

```
gscatter(x,y,g)
gscatter(x,y,s)
gscatter(x,y,'c','m',s)
gscatter(x,y,'c','m',s,'leg')
gscatter(x,y,'c','m',s,'leg','xl','yl')
```

Efectul executării unei astfel de instrucţinui este producerea norului statistic precizat prin vectorii `x` şi `y`, care au aceeaşi lungime, folosind gruparea specificată prin vectorul `g` de aceeaşi lungime. Punctele norului vor fi marcate la fel pentru valori identice corespunzătoare ale lui `g`.

Parametrul `'c'` specifică culorile marcajelor şi are valoarea implicită dată prin`'bgrcmyk'`.

Parametrul opţional `'m'` este un tablou de tip caractere recunoscute de funcţia `plot`, valoarea implicită fiind caracterul `'.'` (punct).

Parmetrul `s` specifică mărimea marcajelor punctelor norului. Acesta poate să fie un scalar, în care caz toate marcajele vor avea aceleaşi mărimi sau poate să fie un vector.

Dacă nu sunt specificate valori suficiente pentru toate grupele, sistemul Matlab efectuează o ciclare a valorilor specificate.

Parametrul `leg` precizează dacă se doreşte afişarea unei legende (implicit acesta este `'on'`), sau nu se doreşte legendă, când se va specifica valoarea `'off'`.

Parametrii `'xl'` şi `'yl'` precizează etichetele pentru cele două axe de coordoante. Dacă aceşti parametri lipsesc, sistemul Matlab etichetează axele de coordoante cu denumirile variabilelor `x` şi respectiv `y`.

Programul 3.2.39. Programul ce urmează generează `n` vectori aleatori ce urmează legea normală bidimensională şi `n` numere aleatoare ce urmează legea uniformă discretă. Folosind aceste date, se va reprezenta norul vectorilor aleatori bidimensionali, efectuând gruparea conform numerelor aleatoare uniforme generate.

```
clear, mu(1)=input('m1='); mu(2)=input('m2=');
v(1,1)=input('sigma1^2=');
v(2,2)=input('sigma2^2=');
v(1,2)=input('Cov(X,Y)='); v(2,1) =v(1,2);
if det(v) <= 0
   error('Matricea v nu e pozitiv definita!')
end
N=input('N(parametrul legii uniforme):');
n=input('n='); X=mvnrnd(mu,v,n);
x=X(:,1); y=X(:,2); g=unidrnd(N,n,1);
gscatter(x,y,g,'k','o*.+')
```

Pentru `n=25`, `mu=(5,10)`, $v = \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}$ şi `N=4`, se obţine Figura 3.8.

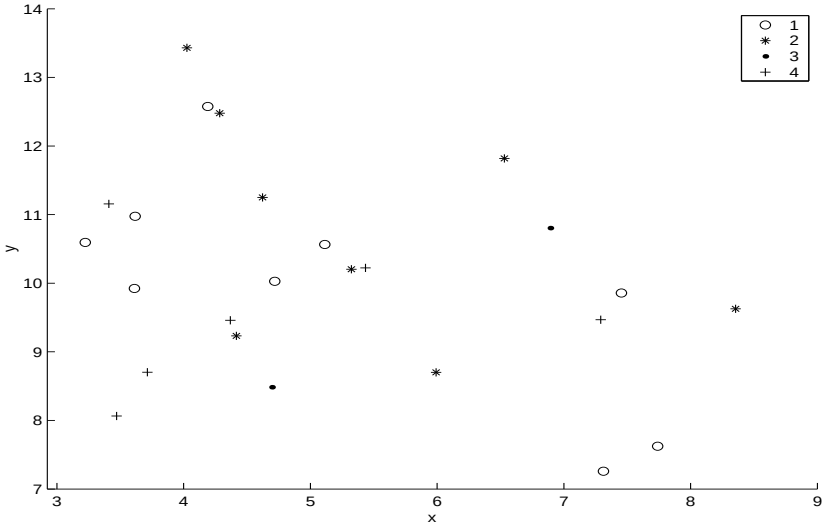

Figura 3.8: Nor statistic pe grupe

Funcţia `plotmatrix`

Sistemul de bază Matlab conţine funcţia `plotmatrix`, care reprezintă pe aceeaşi figură mai mulţi nori statistici. Formele de apel ale funcţiei sunt

```
plotmatrix(x)
plotmatrix(x,y)
plotmatrix(x,y,s)
```

unde `x` şi `y` sunt matrice cu aceleaşi număr de linii, dar numărul coloanelor poate fi diferit, fie acesta `m` şi respectiv `n`. Executarea unei instrucţini de acest tip, produce `m`×`n` nori statistici, pentru fiecare coloană a matricei `x` cu fiecare coloană a matricei `y`, cu marcajele specificate prin parametrul opţional `s`, care are sintaxa de la funcţia `plot`.

Prima formă este echivalentă cu

```
plotmatrix(x,x)
```

cu excepţia că pe diagonala principală a matricei norilor statistici vor fi reprezentate histogramele coloanelor matricei `x`.

Programul 3.2.40. Programul următor va genera o matrice `x` de tipul `(N,3)`, care conţine vectori aleatori ce urmează legea normală tridimensională, după care se va aplica funcţia `plotmatrix`.

```
clear, mu(1)=input('m1=');
mu(2)=input('m2='); mu(3)=input('m3=');
v(1,1)=input('sigma1^2=');
v(2,2)=input('sigma2^2=');
v(3,3)=input('sigma3^2=');
v(1,2)=input('Cov(X,Y)='); v(2,1) =v(1,2);
v(1,3)=input('Cov(X,Z)='); v(3,1) =v(1,3);
v(2,3)=input('Cov(Y,Z)='); v(3,2) =v(1,3);
N=input('N='); X=mvnrnd(mu,v,N);
plotmatrix(X,'o')
```

Pentru `N=10`, `mu=(5,10,15)` şi $\mathtt{v} = \begin{pmatrix} 2 & -1 & 1 \\ -1 & 3 & 0 \\ 1 & 0 & 1 \end{pmatrix}$, se obţine Figura 3.9.

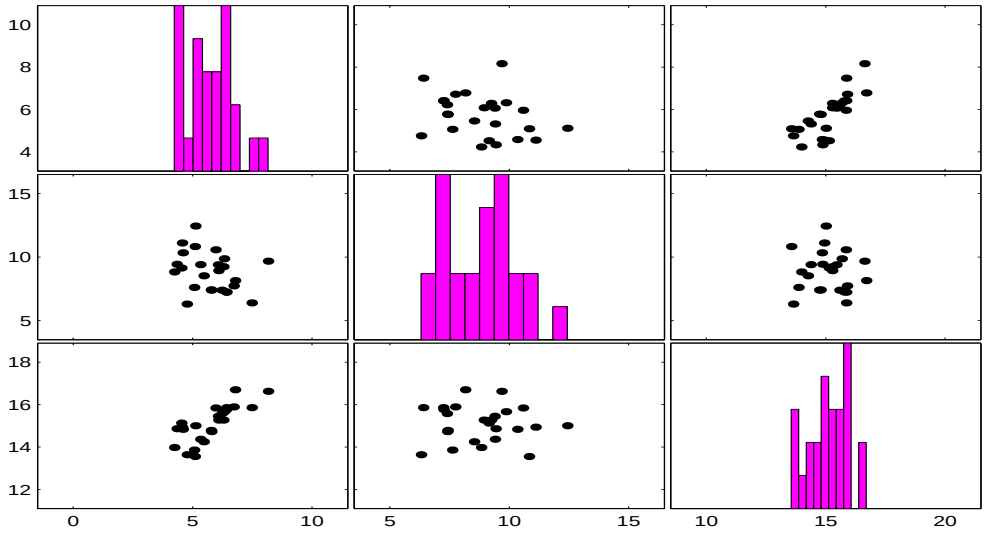

Figura 3.9: Funcţia `plotmatrix`

Funcţia `gplotmatrix`

Statistics toolbox conţine funcţia `gplotmatrix`, care are în principal acelaşi efect ca şi funcţia `plotmatrix`, având însă posibilitatea afişării punctelor norului statisitic în funcţie de grupele la care aparţin.

Formele de apel sunt cele de la funcţia `gscatter`, doar că `x` şi `y`, în acest caz, sunt matrice cu acelaşi număr de linii, iar numărul coloanelor putând fi diferite.

3.2.7 Funcţia `randtool`

Având la dispoziţie funcţiile `random` şi `hist`, putem să prezentăm funcţia demonstrativă `randtool`. Lansarea acestei funcţii se face prin:

```
>>randtool
```

în urma căreia se produce o fereastră grafică interactivă (demonstrativă) privind generarea numerelor aleatoare şi ilustrarea acestora cu ajutorul histogramelor.

Fixarea (stabilirea) legii de probabilitate în scop demonstrativ se face prin alegerea din meniul legilor de probabilitate situat în partea stângă sus a ferestrei.

Volumul numerelor aleatoare, ce urmează a fi generate, se precizează prin introducerea acestuia în fereastra din partea dreaptă sus.

Pentru stabilirea valorilor parametrilor legii de probabilitate considerate se poate proceda în două moduri. Fie prin introducerea în ferestrele corespunzătoare ale valorilor dorite, fie prin deplasarea barelor ataşate acestora. În plus, limite pentru parametrii legii de probabilitate considerate pot fi precizate prin introducerea acestora în ferestrele considerate.

Activarea butonului `output` are ca efect salvarea numerelor aleatoare curente în varaibila `ans` sau în variabila precizată de utilizator, iar butonul `resample` permite repetarea generării de numere aleatoare cu acelaşi volum şi aceeaşi parametri.

3.2.8 Funcţiile `pie` şi `pie3`

Reprezentări ale datelor prin sectoare circulare în plan, respectiv în spaţiu, se obţin prin funcţiile `pie` şi `pie3`, care se apelează prin formele:

```
pie(x,ex,label)
pie3(x,ex,label)
```

unde `x` este un vector având componentele pozitive şi normalizate, adică astfel încât suma acestora să fie 1. Dacă suma acestora va fi mai mică decât 1, atunci numai o parte a cercului va fi reprezentată grafic.

Parametrul `ex` este opţional şi specifică sectoarele de cerc ce urmează să fie detaşate pe figură, acesta, dacă este prezent, trebuie să fie un vector numeric de aceeaşi lungime cu vectorul `x`.

Parametrul opţional `label`, specifică etichetele sectoarele de cerc, acesta, dacă este prezent, trebuie să fie un vector de aceeaşi lungime cu vectorul `x`, dar care conţin numai şiruri de caractere.

Programul 3.2.41. Programul ce urmează generează `n` numere aleatoare ce urmează legea uniformă discretă $\mathcal{U}(N)$, construieşte tabelul sistematizat, după care reprezintă datele sistematizate folosind funcţiile `pie` şi `pie3`.

```
n=input('n='); N=input('N=');
x=unidrnd(N,1,n);
t=tabulate(x); d=length(t(:,3));
```

```
ex=zeros(1,d); ex(1)=1;
subplot(2,1,1), pie(t(:,3)'/100,ex)
subplot(2,1,2), pie3(t(:,3)'/100,ex)
colormap summer
```

Prin execuţia acestui program, cu datele `n=50` şi `N=4`, se obţin reprezentările grafice din Figura 3.10.

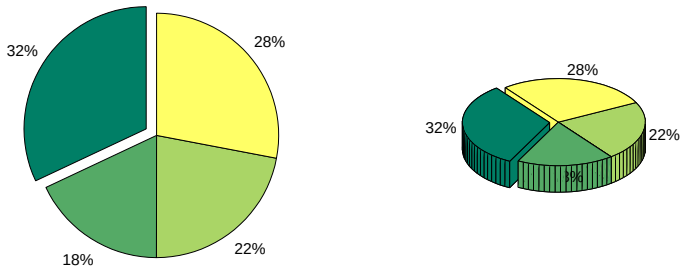

Figura 3.10: Diagrame circulare

3.3 Parametrii distribuţiilor statistice

Se consideră datele statistice primare x'_k, $k = \overline{1, N}$, relative la caracteristica X, din care se obţine distribuţia statistică

$$X \begin{pmatrix} x_i \\ f_i \end{pmatrix}_{i=\overline{1,n}}.$$

3.3.1 Parametri statistici ce măsoară tendinţa

Definiţia 3.3.1. Media (aritmetică) *a distribuţiei statistice a caracteristicii X este dată prin*

$$\bar{x}_a = \frac{1}{N}\sum_{k=1}^{N} x'_k = \frac{1}{N}\sum_{k=1}^{n} f_k x_k = \sum_{k=1}^{n} p_k x_k.$$

Definiţia 3.3.2. Media geometrică *a distribuţiei statistice a caracteristicii pozitive X este dată prin*

$$\bar{x}_g = \sqrt[N]{\prod_{k=1}^{N} x'_k} = \sqrt[N]{\prod_{k=1}^{n} x_k^{f_k}}.$$

Observaţia 3.3.3. În aplicaţii se lucrează mai uşor cu

$$\log \bar{x}_g = \frac{1}{N}\sum_{k=1}^{N} \log x'_k = \frac{1}{N}\sum_{k=1}^{n} f_k \log x_k = \sum_{k=1}^{n} p_k \log x_k.$$

Definiţia 3.3.4. Media armonică *a distribuţiei statistice a caracteristicii nenule X este dată prin*

$$\bar{x}_h = \frac{N}{\sum_{k=1}^{N} \frac{1}{x'_k}} = \frac{N}{\sum_{k=1}^{n} f_k \frac{1}{x_k}} = \frac{1}{\sum_{k=1}^{n} p_k \frac{1}{x_k}}.$$

Observaţia 3.3.5. Între cele trei medii definite mai înainte există relaţiile cunoscute $\bar{x}_h \leqslant \bar{x}_g \leqslant \bar{x}_a$.

3.3.2 Funcţiile `mean`, `geomean` şi `harmean`

Sistemul de bază Matlab conţine funcţia `mean`, iar *Statistics toolbox* dispune de celelalte două funcţii, `geomean` şi `harmmean`.

Apelul acestor funcţii se face prin:

```
ma=mean(x)
mg=geomean(x)
mh=harmmean(x)
```

şi au ca efect calculul mediilor corespunzătoare, pentru componentele vectorului `x`, iar dacă `x` este matrice, atunci operaţia se execută pentru fiecare coloană a matricei. Prin urmare, în acest ultim caz, rezultatul este un vector având atâtea componente câte coloane are matricea `x`.

3.3.3 Funcţia `trimmean`

Sistemul Matlab, prin *Statistics toolbox*, dispune de funcţia `trimmean`, care se poate apela prin

```
m=trimmean(x,p)
```

având ca efect calculul mediei aritmetice a vectorului `x`, după ce au fost eliminate cele mai mici $\frac{\mathtt{p}}{2}\%$ componente şi cele mai mari $\frac{\mathtt{p}}{2}\%$ componente, iar dacă `x` este matrice, această operaţie se face pentru fiecare coloană în parte.

Definiţia 3.3.6. *Numim* mediana *distribuţiei statistice a caracteristicii X, valoarea numerică $\overline{m}$ care împarte datele statistice, ordonate crescător, în două părţi egale.*

Observaţia 3.3.7. Fie datele statistice primare ordonate în mod nedescrescător

$$x'_{(1)} \leqslant x'_{(2)} \leqslant \cdots \leqslant x'_{(N)},$$

atunci mediana va fi dată prin

$$\overline{m} = \begin{cases} x'_{(k)}, & \text{dacă } N = 2k-1, \\ \dfrac{x'_{(k)} + x'_{(k+1)}}{2}, & \text{dacă } N = 2k. \end{cases}$$

Observaţia 3.3.8. Când datele statistice sunt grupate, atunci se determină prima dată *intervalul median* $[a_{j-1}, a_j)$ astfel încât pentru frecvenţele cumulate F_{j-1} şi F_j să fie satisfăcute inegalităţile $F_{j-1} < \frac{N}{2}$ şi $F_j > \frac{N}{2}$. Folosind apoi interpolarea liniară se ia ca mediană

$$\overline{m} = a_{j-1} + d_j \frac{\frac{N}{2} - F_{j-1}}{f_j},$$

unde d_j este amplitudinea intervalului median.

Când caracteristica cercetată este de tip discret, mediana este $\overline{m} = x_j$.

3.3.4 Funcţia `median`

Sistemul de bază Matlab dispune de funcţia `median`. Apelarea acesteia se face prin

```
m=median(x)
```

şi are ca efect calculul medianei pentru componentele vectorului `x`, iar dacă `x` este matrice, atunci operaţia se execută pentru fiecare coloană a matricei. Prin urmare, în acest ultim caz, rezultatul este un vector având atâtea componente câte coloane are matricea `x`.

3.3.5 Parametri statistici ce măsoară dispersarea

Definiţia 3.3.9. *Numim* cuartile *ale distribuţiei statistice a caracteristicii X, valorile numerice $\overline{Q}_1$ (*cuartila inferioară*), $\overline{Q}_2 = \overline{m}$, $\overline{Q}_3$ (*cuartila superioară*), care împart datele statistice, ordonate crescător, în patru părţi egale.*

Observaţia 3.3.10. Când datele statistice sunt grupate, ca şi în cazul medianei, folosind interpolarea liniară considerăm

$$\overline{Q}_1 = a_{i-1} + d_i \frac{\frac{N}{4} - F_{i-1}}{f_i},$$
$$\overline{Q}_3 = a_{k-1} + d_k \frac{\frac{3N}{4} - F_{k-1}}{f_k},$$

după ce s-a determinat *intervalul cuartilic inferior* $[a_{i-1}, a_i)$, astfel încât să aibă loc $F_{i-1} < \frac{N}{4}$ şi $F_i > \frac{N}{4}$, respectiv *intervalul cuartilic superior* $[a_{k-1}, a_k)$, astfel încât $F_{k-1} < \frac{3N}{4}$ şi $F_k > \frac{3N}{4}$.

Când caracteristica cercetată este de tip discret, avem că $\overline{Q}_1 = x_i$ şi $\overline{Q}_3 = x_k$.

Observaţia 3.3.11. În mod analog se definesc parametri numerici cum ar fi *decilele* şi *centilele*.

3.3.6 Funcţia `prctile`

Statistics toolbox conţine funcţia `prctile`, care se poate apela prin

```
c=prctile(x,p)
```

şi care calculează pentru componentele vectorului `x`, după ce acestea au fost ordonate crescător, centilele precizate prin parametrul `p`, care conţine unul sau mai numere întregi de la 1 la 99. De exemplu, `p=[25, 50, 75]` produce respectiv cuartila inferioară ($\overline{Q}_1$), mediana, cuartila superioară ($\overline{Q}_3$):

$$\overline{Q}_1 = \frac{x_{(i)} + x_{(j)}}{2}, \quad i = \texttt{floor}\left(\frac{n+1}{4}\right), \quad j = \texttt{ceil}\left(\frac{n}{4}\right),$$
$$\overline{Q}_3 = \frac{x_{(i)} + x_{(j)}}{2}, \quad i = \texttt{floor}\left(\frac{3\,(n+1)}{4}\right), \quad j = \texttt{ceil}\left(\frac{3n}{4}\right).$$

Dacă `x` este matrice, atunci se operează pentru fiecare coloană în parte.

Definiţia 3.3.12. *Numim* mod *al distribuţiei statistice a caracteristicii X orice punct $\overline{mo}$ de maxim local al distribuţei statistice.*

Observaţia 3.3.13. Când distribuţia statistică are un singur mod spunem că avem distribuţie statistică *unimodală*. Dacă există două sau mai multe moduri spunem că avem distribuţii statistice *bimodale*, respectiv *multimodale*.

Observaţia 3.3.14. Când datele statistice sunt grupate, pentru determinarea modului, se determină *intervalul modal*, adică intervalul cu frecvenţa maximă locală. Dacă intervalul modal este $[a_{k-1}, a_k)$, atunci se consideră

$$\overline{mo} = a_{k-1} + d_k \frac{\Delta f_k}{\Delta f_k - \Delta f_{k+1}},$$

unde $d_k = a_k - a_{k-1}$, $\Delta f_k = f_k - f_{k-1}$, $\Delta f_{k+1} = f_{k+1} - f_k$. Formula se obţine uşor dacă se intersectează interpolantul liniar al punctelor (a_{k-1}, f_{k-1}), (a_k, f_k) cu interpolantul liniar al punctelor (a_{k-1}, f_k), (a_k, f_{k+1})

Când caracteristica cercetată este de tip discret, avem că $\overline{mo} = x_k$.

Definiţia 3.3.15. *Numim* moment de ordin k *al distribuţiei statistice a caracteristicii X, valoarea numerică*

$$\overline{\nu}_k = \frac{1}{N}\sum_{i=1}^{N} x_i'^k = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^k = \sum_{i=1}^{n} p_i x_i^k.$$

Definiţia 3.3.16. *Numim* amplitudinea (interval de variaţie) *distribuţiei statistice a caracteristicii X, valoarea numerică*

$$\overline{\omega} = \max\{\, x'_1, x'_2, \ldots, x'_N \,\} - \min\{\, x'_1, x'_2, \ldots, x'_N \,\} = x_{\max} - x_{\min}.$$

Definiţia 3.3.17. *Numim* abatere cuartilică (interval intercuartilic) *a distribuţiei statistice a lui X, diferenţa dintre cuartila superioară şi cuartila inferioară, adică diferenţa $\overline{Q}_3 - \overline{Q}_1$.*

Observaţia 3.3.18. Mai întâlnim, în aplicaţii, *variaţia intercuartilică* dată prin formula

$$\overline{Q} = \frac{\left(\overline{Q}_3 - \overline{m}\right) + \left(\overline{m} - \overline{Q}_1\right)}{2} = \frac{\overline{Q}_3 - \overline{Q}_1}{2},$$

respectiv *abaterea cuartilă relativă*

$$\overline{Q}_r = \frac{\overline{Q}_3 - \overline{Q}_1}{\overline{Q}_2} = \frac{\overline{Q}_3 - \overline{Q}_1}{\overline{m}}.$$

Observaţia 3.3.19. Dacă $\overline{Q}_3 - \overline{Q}_1 < \frac{\overline{\omega}}{2}$, atunci distribuţia se consideră intens concentrată, iar în caz contrar, intens dispersată.

Definiţia 3.3.20. *Numim* abatere medie (absolută) *a distribuţiei statistice X, valoarea numerică*

$$\overline{\delta} = \frac{1}{N}\sum_{k=1}^{N} |x'_k - \overline{x}| = \frac{1}{N}\sum_{k=1}^{n} f_k |x_k - \overline{x}| = \sum_{k=1}^{n} p_k |x_k - \overline{x}|,$$

unde $\overline{x} = \overline{x}_a$.

Definiţia 3.3.21. *Numim* moment centrat de ordin k *al distribuţiei statistice X, valoarea numerică*

$$\overline{\mu}_k = \frac{1}{N}\sum_{i=1}^{N} \left(x'_i - \overline{x}\right)^k = \frac{1}{N}\sum_{i=1}^{n} f_i \left(x_i - \overline{x}\right)^k = \sum_{i=1}^{n} p_i \left(x_i - \overline{x}\right)^k.$$

Definiţia 3.3.22. *Momentul centrat de ordinul doi al distribuţiei statistice X se numeşte* dispersie *şi se notează $\overline{\sigma}^2 = \overline{\mu}_2$, iar $\overline{\sigma} = \sqrt{\overline{\mu}_2}$ se numeşte* abatere medie pătratică *sau* abatere standard.

Observaţia 3.3.23. Alte formule de calcul pentru dispersie sunt

$$\overline{\sigma}^2 = \frac{1}{N}\left[\sum_{i=1}^{n} f_i x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{n} f_i x_i\right)^2\right],$$

$$\overline{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{n} f_i \left(x_i - a\right)^2 - \left(\overline{x} - a\right)^2, \quad a \in \mathbb{R}, \quad \text{(formula lui König)}.$$

Observaţia 3.3.24. Se observă că pentru calculul dispersiei ar fi necesar să fie parcurse datele statistice de două ori, odată pentru calculul mediei aritmetice $\overline{x}$, iar apoi pentru calculul lui $\overline{\sigma}^2$, care foloseşte pe $\overline{x}$. Prezentăm o metodă pentru calculul dispersiei printr-o singură parcurgere a datelor statistice.

Pentru aceasta introducem notaţiile

$$t_j = \sum_{i=1}^{j} x_i' \qquad \text{şi} \qquad s_j^2 = \sum_{i=1}^{j} \left(x_i' - \frac{t_j}{j} \right)^2, \quad j = \overline{1, N}.$$

Vom arăta, prima dată, că are loc relaţia de recurenţă

$$s_j^2 = s_{j-1}^2 + \frac{1}{j\,(j-1)} \left(j\, x_j' - t_j \right)^2, \quad j = \overline{2, N}.$$

Pentru stabilirea relaţiei de recurenţă se scrie succesiv

$$\begin{aligned} s_j^2 - s_{j-1}^2 &= \sum_{i=1}^{j} \left(x_i' - \frac{t_j}{j} \right)^2 - \sum_{i=1}^{j-1} \left(x_i' - \frac{t_j - x_j'}{j-1} \right)^2 \\ &= \sum_{i=1}^{j} x_i'^2 - \frac{2t_j}{j} \sum_{i=1}^{j} x_i' + \frac{t_j^2}{j} - \sum_{i=1}^{j-1} x_i'^2 + \frac{2\left(t_j - x_j' \right)}{j-1} \sum_{i=1}^{j-1} x_i' - \frac{\left(t_j - x_j' \right)^2}{j-1} \\ &= x_j'^2 - \frac{2t_j^2}{j} + \frac{t_j^2}{j} + \frac{2\left(t_j - x_j' \right)^2}{j-1} - \frac{\left(t_j - x_j' \right)^2}{j-1} = x_j'^2 - \frac{t_j^2}{j} + \frac{\left(t_j - x_j' \right)^2}{j-1}. \end{aligned}$$

Dacă reţinem extremităţile şirului de egalităţi avem:

$$s_j^2 - s_{j-1}^2 = \frac{1}{j\,(j-1)} \left(j^2 x_j'^2 - 2j\, t_j x_j' + t_j^2 \right) = \frac{1}{j\,(j-1)} \left(j\, x_j' - t_j \right)^2,$$

ce trebuie arătat.

Observăm că $\overline{x} = \frac{1}{N} t_N$, iar $\overline{\sigma}^2 = \frac{1}{N} s_N^2$. Prin urmare formula de recurenţă de la punctul precedent permite calculul lui $\overline{\sigma}^2$ odată cu calculul lui $\overline{x}$, deci printr-o parcurgere a datelor statistice.

Algoritmul de calcul poate fi descris prin următoarele etape:

1. $s := 0$, $t := x_1'$.
2. Pentru $j = \overline{2, N}$: $t := t + x_j'$, $s := s + \dfrac{1}{j\,(j-1)} \left(j\, x_j' - t \right)^2$.
3. $\overline{x} = \dfrac{t}{N}$; $\overline{\sigma}^2 = \dfrac{s}{N}$.

Definiţia 3.3.25. *Numim* coeficient de variaţie *al distribuţiei statistice X, raportul*

$$\overline{v} = \frac{\overline{\sigma}}{\overline{x}},$$

care de regulă se exprimă în procente ($100\,\overline{v}\,\%$).

Definiţia 3.3.26. *Numim* coeficient de variaţie intercuartil *al distribuţiei statistice X, raportul*

$$\overline{q} = \frac{\overline{Q}}{\overline{m}} = \frac{\frac{\overline{Q}_3 - \overline{Q}_1}{2}}{\overline{Q}_2} = \frac{\overline{Q}_3 - \overline{Q}_1}{2\overline{Q}_2}.$$

Definiţia 3.3.27. *Numim* coeficientul de asimetrie intercuartil (Yule) *al distribuţiei statistice X, raportul*

$$\overline{Y} = \frac{\left(\overline{Q}_3 - \overline{Q}_2\right) - \left(\overline{Q}_2 - \overline{Q}_1\right)}{\left(\overline{Q}_3 - \overline{Q}_2\right) + \left(\overline{Q}_2 - \overline{Q}_1\right)} = \frac{\overline{Q}_3 + \overline{Q}_1 - 2\overline{m}}{\overline{Q}_3 - \overline{Q}_1}.$$

Observaţia 3.3.28. Coeficientul lui Yule satisface relaţiile $-1 \leqslant \overline{Y} \leqslant 1$, iar dacă $\overline{Q}_3 - \overline{Q}_2 > \overline{Q}_2 - \overline{Q}_1$ ($\overline{Y} > 0$), atunci $\overline{mo} < \overline{m} < \overline{x}$, în caz contrar $\overline{mo} > \overline{m} > \overline{x}$.

Definiţia 3.3.29. *Numim* coeficienţi ai lui Pearson *relativ la distribuţia statistică X, rapoartele:*

$$\begin{aligned} \overline{s} &= \frac{\overline{x} - \overline{mo}}{\overline{\sigma}} \quad \text{(coeficient de asimetrie)}, \\ \overline{\beta}_1 &= \frac{\overline{\mu}_3^2}{\overline{\mu}_2^3} \quad \text{(skewness)}, \\ \overline{\beta}_2 &= \frac{\overline{\mu}_4}{\overline{\mu}_2^2} \quad \text{(kurtosis)}. \end{aligned} \tag{3.3.1}$$

Definiţia 3.3.30. *Numim* coeficienţi ai lui Fisher *relativ la distribuţia statistică a lui X, valorile numerice*

$$\begin{aligned} \overline{\gamma}_1 &= \sqrt{\overline{\beta}_1} = \frac{\overline{\mu}_3}{\overline{\sigma}^3} \quad \text{(asimetria)}, \\ \overline{\gamma}_2 &= \overline{\beta}_2 - 3 = \frac{\overline{\mu}_4}{\overline{\mu}_2^2} - 3 = \frac{\overline{\mu}_4}{\overline{\sigma}^4} - 3 \quad \text{(excesul)}. \end{aligned} \tag{3.3.2}$$

Observaţia 3.3.31. Pentru curba lui Gauss avem că $\overline{\beta}_1 = 0$, deci $\overline{\gamma}_1 = 0$. Când $\overline{\gamma}_1 < 0$, maximul curbei distribuţiei este deplasat spre dreapta în raport cu valoarea medie, iar când $\overline{\gamma}_1 > 0$, maximul este deplasat spre stânga.

De asemenea, pentru curba lui Gauss avem că $\overline{\beta}_2 = 3$, deci $\overline{\gamma}_2 = 0$. Astfel, dacă $\overline{\beta}_2 > 3$ $(\overline{\gamma}_2 > 0)$ atunci curba distribuţiei este mai îngustată decât curba lui Gauss, distribuţia numindu-se *leptocurtică*. Dacă $\overline{\beta}_2 < 3$ $(\overline{\gamma}_2 < 0)$ curba distribuţiei este mai turtită, distribuţia numindu-se *platicurtică*.

O parte din aceşti parametri statistici există în sistemul de bază Matlab, iar alţii se găsesc în *Statistics toolbox*. Facem prezentarea lor în cele ce urmează.

3.3.7 Funcţia `moment`

Apelul funcţiei `moment` se poate face prin

```
s=moment(x,k)
```

şi are ca efect calculul, pentru componentele vectorului `x`, a momentului centrat de ordin `k`, iar dacă `x` este matrice, atunci se operează pe fiecare coloană în parte.

3.3.8 Funcţiile `var` şi `std`

Apelurile funcţiilor `var` şi `std` se pot face prin:

```
v=var(x)
s=std(x)
v=var(x,w)
v=var(x,1)
```

Primele două forme calculează, pentru componentele vectorului `x`, dispersia definită prin

$$\mathtt{v} = \frac{1}{m-1}\sum_{i=1}^{m}(\mathtt{x}_i - \bar{\mathtt{x}})^2,$$

unde m este lungimea vectorului, şi respectiv abaterea standard $\mathtt{s} = \sqrt{\mathtt{v}}$.

Forma a doua pentru `var`, când parametrul opţional `w` este prezent, acesta trebuie să fie un vector cu ponderi pozitive a căror sumă este 1, şi are ca efect calculul lui `v` după formula

$$\mathtt{v} = \sum_{i=1}^{m}\mathtt{w}_i(\mathtt{x}_i - \bar{\mathtt{x}})^2, \quad \bar{\mathtt{x}} = \sum_{i=1}^{m}\mathtt{w}_i\mathtt{x}_i.$$

Forma a treia pentru `var`, când parametrul opţional `1` este prezent, are ca efect calculul lui `v` după formula momentului centrat de ordinul 2.

În toate cazurile, dacă `x` este matrice, se operează pe fiecare coloană în parte.

3.3.9 Funcţiile `range`, `iqr`, `mad`, `skewness` şi `kurtosis`

Apelurile funcţiilor `range`, `iqr`, `mad`, `skewness`, `kurtosis` se fac prin

```
r=range(x)
iq=iqr(x)
md=mad(x)
sk=skewness(x)
k=kurtosis(x)
```

şi calculează pentru componentele vectorului `x`, respectiv amplitudinea, abaterea cuartilică, abaterea absolută, asimetria (cu formula (3.3.2)) şi excesul (cu formula (3.3.1)).

Dacă `x` este matrice, se operează pe fiecare coloană în parte.

3.3.10 Funcţiile `max` şi `min`

Prezentăm numai modul de apel pentru funcţia `max`, deoarece funcţia `min` se utilizează la fel.

Pentru `max` avem următoarele forme de apel:

```
r=max(x)
r=max(x,y)
[r,k]=max(x)
[r,k]=max(x,y)
```

Prin acestea se calculează valoarea maximă a componentelor vectorului `x`, iar în `k` se obţin şi indicii corespunzători componentelor. Dacă este prezent parametrul opţional `y`, care este un vector de aceaşi lungime cu cea a lui `x`, atunci `r` este un vector ce va conţine valorile maxime, prin compararea celor doi vectori pe componente. Dacă `x` este matrice, respectiv când apare `y` matrice de aceaşi dimensiune, atunci comparaţiile sunt efectuate pe coloane în primul caz, iar în al doilea caz se face compararea element cu element.

3.3.11 Funcţiile `sort` şi `sortrows`

Funcţia `sort` are următoarele forme de apel

```
y=sort(x)
y=sort(x,k)
[y,s]=sort(x)
[y,s]=sort(x,k)
```

şi `sortrows` aceleaşi forme de apel.

Funcţia `sort` ordonează crescător elementele fiecărei coloane a matricei `x`, când `k=1` (care este valoarea implicită), respectiv ordonarea crescătoare pe linii, când `k=2`. Vectorul `s` păstrează indicii elementelor dinainte de ordonare. Dacă `x` este vector, atunci parametrul opţional lipseşte şi este efectuată ordonarea crescătoare a componentelor vectorului, eventual cu păstrarea indicilor componentelor dinainte de ordonare în `s`.

Funcţia `sortrows` ordonează lexicografic liniile matricei `x`, având în vedere toate elementele liniilor matricei (când parametrul `k` lipseşte) sau numai elementele ce aparţin coloanelor precizate prin vectorul `k`, cu posibilitatea păstrării în `s` a indicilor liniilor înainte de ordonare.

3.3.12 Funcţiile `sum`, `prod`, `cumsum` şi `cumprod`

Apelurile funcţiilor `sum`, `prod`, `cumsum`, `cumprod`, se pot realiza prin:

```
y=sum(x)
y=sum(x,k)
y=prod(x)
y=prod(x,k)
y=cumsum(x)
y=cumsum(x,k)
y=cumprod(x)
y=cumprod(x,k)
```

Execuţia acestor instrucţiuni calculează respectiv sumele, sumele parţiale, produsele, produsele parţiale pentru fiecare coloană a matricei `x`, când `k=1` (valoare implicită), iar când `k=2`, se execută aceste operaţii pe linii. Dacă `x` este vector, atunci parametrul opţional `k` lipseşte, iar operaţiile se execută asupra componentelor vectorului.

3.3.13 Funcţia `diff`

Apelul funcţiei `diff` se realizează prin:

```
y=diff(x)
y=diff(x,r)
y=diff(x,r,k)
```

În urma executării unei astfel de instrucţiune, se calculează diferenţele de ordin `r` ale elementelor liniilor consecutive ale matricei `x`, când `k=1` (valoare implicită), respectiv diferenţele de ordin `r` ale elementelor coloanelor consecutive ale matricei `x`, când `k=2`, adică

$$\begin{aligned} &\mathtt{y}_{i,j} = \mathtt{x}_{i+1,j} - \mathtt{x}_{i,j}, \quad i = \overline{1, m-1},\ j = \overline{1, n}, \quad (\mathtt{k=1}) \\ &\mathtt{y}_{i,j} = \mathtt{x}_{i,j+1} - \mathtt{x}_{i,j}, \quad i = \overline{1, m},\ j = \overline{1, n-1}, \quad (\mathtt{k=2}). \end{aligned}$$

Diferenţele de ordin doi pentru `x` sunt diferenţele de ordinul întâi pentru `y`. Dacă `x` este vector, atunci parametrul `k` lipseşte, iar operaţia se execută asupra componentelor vectorului.

3.3.14 Funcţiile `nanmean`, `nanmedian`, `nanstd`, `nanmin`, `nanmax` şi `nansum`

Aceste funcţii din *Statistics toolbox* sunt definite ca şi funcţiile corespunzătoare `mean`, `median`, `std`, `min`, `max`, `sum`, cu excepţia faptului că sunt excluse din calcule valorile `NaN`.

3.3.15 Corecţiile lui Sheppard

Dacă datele statistice relative la caracteristica X de tip continuu au fost grupate, în unele cazuri, se cere să fie aplicate *corecţiile lui Sheppard* pentru momentele $\overline{\nu}_k$. Dacă se notează cu $\overline{\nu}_k{}'$ momentul corectat de ordinul k, atunci

$$\overline{\nu}_k = \frac{1}{k+1}\sum_{i=0}^{\left[\frac{k}{2}\right]} \binom{k+1}{2i+1} \left(\frac{d}{2}\right)^{2i} \overline{\nu}'_{k-2i}, \quad k = 1, 2, \ldots$$

unde d este amplitudinea claselor.

Folosind formula precedentă scriem primele patru momente corectate:

$$\overline{\nu}_1{}' = \overline{\nu}_1, \quad \overline{\nu}_2{}' = \overline{\nu}_2 - \frac{d^2}{12}, \quad \overline{\nu}_3{}' = \overline{\nu}_3 - \frac{d^2}{4}\overline{\nu}_1, \quad \overline{\nu}_4{}' = \overline{\nu}_4 - \frac{d^2}{2}\overline{\nu}_2 + \frac{7d^4}{240}.$$

Având în vedere că

$$\overline{\mu}_2 = \overline{\nu}_2 - \overline{\nu}_1^2, \quad \overline{\mu}_3 = \overline{\nu}_3 - 3\overline{\nu}_1\overline{\nu}_2 + 2\overline{\nu}_1^3, \quad \overline{\mu}_4 = \overline{\nu}_4 - 4\overline{\nu}_1\overline{\nu}_3 + 6\overline{\nu}_1^2\overline{\nu}_2 - 3\overline{\nu}_1^4,$$

obţinem, de asemenea, momentele centrate corectate

$$\overline{\mu}_2{}' = \overline{\mu}_2 - \frac{d^2}{12}, \quad \overline{\mu}_3{}' = \overline{\mu}_3, \quad \overline{\mu}_4{}' = \overline{\mu}_4 - \frac{d^2}{2}\overline{\mu}_2 + \frac{7d^4}{240}.$$

Corecţiile lui Sheppard se aplică, de regulă, când avem distribuţie unimodală şi pentru $N > 1000$, $d > \frac{1}{20}\overline{\omega}$.

Exemplul 3.3.32. Să revenim la datele statistice din Exemplul 3.2.10 şi să calculăm parametrii statistici definiţi până aici.

Printre parametrii statistici care măsoară tendinţa îi calculăm pe următorii.

Media (aritmetică), care este dată prin

$$\overline{x}_a = \overline{x} = \frac{1}{N}\sum_{k=1}^{N} x'_k = \frac{1}{70}\left(1.318 + 3.128 + \cdots + 1.966\right) = 2.27077.$$

Mediana, care se obţine după ordonarea crescătoare a datelor statistice. Anume, avem că

$$0.842 < 1.155 < 1.179 < 1.318 < 1.355 < 1.403 < 1.493 < 1.546 < 1.548 < 1.553$$
$$< 1.560 < 1.583 < 1.628 < 1.631 < 1.681 < 1.690 < 1.708 < \boxed{1.802} < 1.855 < 1.864$$
$$< 1.875 < 1.931 < 1.937 < 1.946 < 1.960 < 1.962 < 1.966 < 1.977 < 2.015 < 2.072$$
$$< 2.128 < 2.206 < 2.214 < 2.219 < \boxed{\boxed{2.230 < 2.256}} < 2.262 < 2.298 < 2.304 < 2.316$$
$$< 2.337 < 2.345 < 2.444 < 2.455 < 2.465 < 2.493 < 2.500 < 2.502 < 2.517 < 2.525$$
$$< 2.636 < 2.637 < \boxed{2.641} < 2.676 < 2.726 < 2.758 < 2.807 < 2.838 < 2.879 < 3.006$$
$$< 3.093 < 3.128 < 3.156 < 3.281 < 3.394 < 3.426 < 3.460 < 3.537 < 3.852 < 3.972$$

Datele statistice dublu încadrate fiind la mijlocul acestei ordonări (număr par de date statistice) obţinem mediana ca fiind

$$\overline{m} = \frac{2.230 + 2.256}{2} = 2.243.$$

Cuartila inferioară se obţine din aceeaşi ordonare a datelor statistice, luând data statistică cu numărul de ordine 18, prima încadrată, adică $\overline{Q}_1 = 1.802$, iar cuartila superioară, luând data statistică cu numărul de ordine 53, a doua încadrată, adică $\overline{Q}_3 = 2.641$.

Printre parametrii statistici care măsoară gradul de împrăştiere îi calculăm pe cei care urmează.

Intervalul de variaţie este

$$\overline{\omega} = x_{\max} - x_{\min} = 3.972 - 0.842 = 3.130.$$

Abaterea cuartilă sau intervalul intercuartilic se obţine prin

$$\overline{Q}_3 - \overline{Q}_1 = 2.641 - 1.802 = 0.839.$$

Deoarece $\overline{Q}_3 - \overline{Q}_1 = 0.839 < 1.565 = \dfrac{\overline{\omega}}{2}$, avem că distribuţia statistică este intens concentrată.

Abaterea cuartilă relativă este

$$\overline{Q}_r = \frac{\overline{Q}_3 - \overline{Q}_1}{\overline{m}} = \frac{0.839}{2.243} = 0.374.$$

Abaterea medie (absolută) are valoarea dată prin

$$\overline{\delta} = \frac{1}{N} \sum_{k=1}^{N} | \, x'_k - \overline{x} \, | =$$

$$= \frac{1}{70}\left(\left|1.318-2.271\right| + \left|3.128-2.271\right| + \cdots + \left|1.966-2.271\right|\right)$$
$$= 0.5277.$$

Primele patru momente sunt

$$\overline{\nu}_1 = \overline{x} = 2.27077,$$
$$\overline{\nu}_2 = \frac{1}{N}\sum_{k=1}^{N} x_k'^2 = \frac{1}{70}\left(1.318^2 + 3.128^2 + \cdots + 1.966^2\right) = 5.59435,$$
$$\overline{\nu}_3 = \frac{1}{N}\sum_{k=1}^{N} x_k'^3 = \frac{1}{70}\left(1.318^3 + 3.128^3 + \cdots + 1.966^3\right) = 14.808,$$
$$\overline{\nu}_4 = \frac{1}{N}\sum_{k=1}^{N} x_k'^4 = \frac{1}{70}\left(1.318^4 + 3.128^4 + \cdots + 1.966^4\right) = 41.7278.$$

Folosind legăturile dintre momentele obişnuite şi momentele centrate se obţin acestea din urmă cu formulele

$$\overline{\mu}_2 = \overline{\nu}_2 - \overline{\nu}_1^2 = 0.437946,$$
$$\overline{\mu}_3 = \overline{\nu}_3 - 3\overline{\nu}_1\overline{\nu}_2 + 2\overline{\nu}_1^3 = 0.115571,$$
$$\overline{\mu}_4 = \overline{\nu}_4 - 4\overline{\nu}_1\overline{\nu}_3 + 6\overline{\nu}_1^2\overline{\nu}_2 - 3\overline{\nu}_1^4 = 0.540172.$$

Se pot calcula apoi coeficienţii lui Pearson

$$\overline{\beta}_1 = \frac{\overline{\mu}_3^2}{\overline{\mu}_2^3} = 0.159012 \quad \text{(skewness)}, \quad \overline{\beta}_2 = \frac{\overline{\mu}_4}{\overline{\mu}_2^2} = 2.81637 \quad \text{(kurtosis)},$$

respectiv coeficienţii lui Fisher

$$\overline{\gamma}_1 = \sqrt{\overline{\beta}_1} = 0.3988 \quad \text{(asimetria)}, \quad \overline{\gamma}_2 = \overline{\beta}_2 - 3 = -0.18363 \quad \text{(excesul)}.$$

Să recalculăm parametrii statistici pe baza datelor statistice sistematizate (grupate).

Media (aritmetică) va fi

$$\overline{x} = \frac{1}{N}\sum_{k=1}^{n} f_k x_k = \frac{1}{70}\left(1\times 0.85 + 2\times 1.15 + \cdots + 2\times 3.85\right) = 2.29.$$

Pentru a calcula mediana folosim formula

$$\overline{m} = a_{j-1} + d_j \frac{\frac{N}{2} - F_{j-1}}{f_j}.$$

Aici F_{j-1} este cea mai mare frecvenţă cumulată (ascendentă) ce nu depăşeşte pe $\frac{N}{2}$, deci $F_5 = 31 < \frac{70}{2} < 46 = F_6$. Astfel se obţine $[a_5; a_6) = [2.2; 2.5)$, intervalul median, care are amplitudinea $d_6 = d = 0.3$ şi frecvenţa absolută $f_6 = 15$. Prin urmare rezultă că

$$\overline{m} = 2.2 + 0.3 \cdot \frac{35 - 31}{15} = 2.28.$$

Pentru a calcula cuartila inferioară se determină prima dată intervalul cuartilic inferior $[a_{i-1}; a_i)$ astfel încât F_{i-1} să fie cea mai mare frecvenţă cumulată ascendentă ce nu depăşeşte pe $\frac{N}{4} = 17.5$. Aceasta este $F_3 = 12$. Prin urmare $[a_3; a_4) = [1.6; 1.9)$, iar $d_4 = d = 0.3$. Se foloseşte apoi formula

$$\overline{Q}_1 = a_{i-1} + d_i \cdot \frac{\frac{N}{4} - F_{i-1}}{f_i} = 1.6 + 0.3 \cdot \frac{17.5 - 12}{9} = 1.783.$$

În mod analog, pentru cuartila superioară

$$\overline{Q}_3 = 2.5 + 0.3 \cdot \frac{3 \times 17.5 - 46}{10} = 2.695.$$

Intervalul modal, cel cu frecvenţa maximă, este $[a_5; a_6) = [2.2; 2.5)$. Prin urmare modul va fi calculat cu formula

$$\overline{mo} = a_5 + d_6 \cdot \frac{\Delta f_5}{\Delta f_5 - \Delta f_6} = 2.2 + 0.3 \cdot \frac{15 - 10}{(15 - 10) - (10 - 15)} = 2.35.$$

Abaterea (absolută) este

$$\begin{aligned}
\overline{\delta} &= \frac{1}{N} \sum_{k=1}^{n} f_k \, | \, x_k - \overline{x} \, | \\
&= \frac{1}{70} \left(1 \times \, | \, 0.85 - 2.29 \, | \, + 2 \times \, | \, 1.15 - 2.29 \, | \, + \cdots + 2 \times \, | \, 3.85 - 2.29 \, | \, \right) \\
&= 0.5297.
\end{aligned}$$

Primele patru momente se calculează prin

$$\begin{aligned}
&\overline{\nu}_1 = \overline{x} = 2.29, \\
&\overline{\nu}_2 = \frac{1}{N} \sum_{k=1}^{n} f_k x_k^2 = \frac{1}{70} \left(1 \times 0.85^2 + 2 \times 1.15^2 + \cdots + 1 \times 3.85^2 \right) = 5.68792, \\
&\overline{\nu}_3 = \frac{1}{N} \sum_{k=1}^{n} f_k x_k^3 = \frac{1}{70} \left(1 \times 0.85^3 + 2 \times 1.15^3 + \cdots + 1 \times 3.85^3 \right) = 15.1467,
\end{aligned}$$

$$\overline{\nu}_4 = \frac{1}{N}\sum_{k=1}^{n} f_k x_k^4 = \frac{1}{70}\left(1\times 0.85^4 + 2\times 1.15^4 + \cdots + 1\times 3.85^4\right) = 42.8038,$$

iar apoi se obţin momentele centrate

$$\begin{aligned}
\overline{\mu}_2 &= \overline{\nu}_2 - \overline{\nu}_1^2 = 0.443829,\\
\overline{\mu}_3 &= \overline{\nu}_3 - 3\overline{\nu}_1\overline{\nu}_2 + 2\overline{\nu}_1^3 = 0.088591,\\
\overline{\mu}_4 &= \overline{\nu}_4 - 4\overline{\nu}_1\overline{\nu}_3 + 6\overline{\nu}_1^2\overline{\nu}_2 - 3\overline{\nu}_1^4 = 0.526822.
\end{aligned}$$

Coeficienţii lui Pearson sunt

$$\overline{\beta}_1 = \frac{\overline{\mu}_3^2}{\overline{\mu}_2^3} = 0.08977 \quad \text{(skewness)}, \quad \overline{\beta}_2 = \frac{\overline{\mu}_4}{\overline{\mu}_2^2} = 2.67444 \quad \text{(kurtosis)},$$

iar coeficienţii lui Fisher sunt

$$\overline{\gamma}_1 = \sqrt{\overline{\beta}_1} = 0.2996 \quad \text{(asimetria)}, \quad \overline{\gamma}_2 = \overline{\beta}_2 - 3 = -0.3255 \quad \text{(excesul)}.$$

Programul 3.3.33. Să scriem un program, care să calculeze o parte din aceşti parametri statistici din exemplul precedent. Ne vom opri numai la acei parametri, care sunt definiţi prin funcţii Matlab, dar evident că odată deteminaţi aceşti parametri, se pot calcula uşor şi ceilalţi parametri din exemplul precedent.

```
load x.dat -ascii% fisierul x.dat contine datele
                 % primare din exemplul precedent
ma=mean(x);     fprintf('ma=     %6.4f\n',ma)
mg=geomean(x);  fprintf('mg=     %6.4f\n',mg)
mh=harmmean(x); fprintf('mh=     %6.4f\n',mh)
m=median(x);    fprintf('mediana=%6.4f\n',m)
Q=prctile(x,[25,75]);
                fprintf('Q1=     %6.4f\n',Q(1)),
                fprintf('Q3=     %6.4f\n',Q(2))
md=mad(x);      fprintf('md=     %6.4f\n',md)
mu=moment(x,2); fprintf('mu2=    %6.4f\n',mu)
mu=moment(x,3); fprintf('mu3=    %6.4f\n',mu)
mu=moment(x,4); fprintf('mu4=    %6.4f\n',mu)
r=range(x);     fprintf('range=  %6.4f\n',r)
iq=iqr(x);      fprintf('iq=     %6.4f\n',iq)
s=std(x);       fprintf('s=      %6.4f\n',s)
v=var(x);       fprintf('v=      %6.4f\n',v)
sk=skewness(x); fprintf('sk=     %6.4f\n',sk)
k=kurtosis(x);  fprintf('k=      %6.4f\n',k)
```

Prin executarea programului se obţin rezultatele

```
ma=     2.2708
mg=     2.1725
```

```
mh=     2.0704
mediana=2.2430
Q1=     1.8020
Q3=     2.6410
md=     0.5277
mu2=    0.4379
mu3=    0.1156
mu4=    0.5402
range=  3.1300
iq=     0.8390
s=      0.6666
v=      0.4443
sk=     0.3988
k=      2.8164
```

3.3.16 Funcţia `grpstats`

Funcţia `grpstats`, din *Statistics toolbox*, face posibilă determinarea unor parametri statistici pentru date clasificate (grupate) după o anumită variabilă de grupare.

Apelul funcţiei se face prin

```
ma=grpstats(x,g)
[ma,s,fr,nume]=grpstats(x,g)
```

Parametrul `g` este un vector numeric sau de tip caracter, de aceeaşi lungime cu numărul liniilor matricei `x`, şi care permite gruparea datelor conţinute de `x`. O astfel de grupare este formată din acele date din `x`, pentru fiecare coloană în parte, pentru care valorile corespunzătoare din `g` sunt aceleaşi.

Executarea primei instrucţiuni are ca efect calculul mediilor aritmetice pentru fiecare grupă a fiecărei coloane din matricea `x`.

A doua instrucţiune, faţă de prima instrucţiune, mai calculează abaterile standard `s`, pentru fiecare grupă din fiecare coloană a matricei `x`, precum şi numărul datelor `fr` şi etichetele `nume`, pentru fiecare grupă.

Programul 3.3.34. Programul care urmează generează 100 de numere aleatoare ce urmează legea uniformă discretă, cu $N = 4$, în vectorul `g`, care va reprezenta un vector de grupare.

În matricea `x` cu 100 de linii şi 3 coloane, vor fi generate numere aleatoare ce urmează legea normală $\mathcal{N}(\mu, \sigma)$, unde parametrul $\sigma = 1$, iar parametrul μ, este respectiv 1, 2 şi 3 pentru cele trei coloane.

Folosind vectorul de grupare `g`, se calculează valorile medii, abaterile standard, frecvenţele absolute şi se precizează numele grupelor, pentru coloanele matricei `x`.

```
g=unidrnd(4,100,1);
t=1:3;
t=t(ones(100,1),:);
```

```
x=normrnd(t,1);
[ma,s,fr,nume]=grpstats(x,g);
fprintf('    Mediile pe grupe \n')
disp(ma)
fprintf('    Abaterile standard pe grupe \n')
disp(s)
fprintf('    Frecventele grupelor \n')
disp(fr)
fprintf('    Etichetele grupelor \n')
disp(nume')
```

În urma executării programului se obţin următoarele rezultate

```
Mediile pe grupe
1.2486    1.8376    3.0692
0.6895    2.0323    3.2620
1.0393    1.9893    3.0223
1.0034    1.8834    3.0419

Abaterile standard pe grupe
0.2024    0.1701    0.2033
0.2076    0.1793    0.2763
0.2051    0.1723    0.1574
0.1875    0.1670    0.1859

Frecventele grupelor
24    24    24
23    23    23
27    27    27
26    26    26

Etichetele grupelor
'1'    '2'    '3'    '4'
```

3.3.17 Funcţia `boxplot`

Sistemul Matlab, prin *Statistics toolbox*, dispune de funcţia `boxplot`, care se poate apela cu una din următoarele forme:

```
boxplot(x)
boxplot(x,cr)
boxplot(x,cr,'s')
boxplot(x,cr,'s',v)
boxplot(x,cr,'s',v,w)
```

Efectul executării acestor instrucţiuni este reprezentarea grafică prin dreptunghiuri cu prelungiri de segmente pentru fiecare coloană a matricei `x`. Laturile de jos şi de sus ale dreptunghiurilor au ordonatele date de cuartilele inferioare şi superioare corespunzătoare, medianele fiind şi ele marcate la rândul lor prin segmente paralele situate în interioarele dreptunghiurilor, poziţionate prin ordonatele date de valorile acestora.

În mijloacele laturilor de jos şi de sus ale dreptunghiurilor sunt trasate spre exterior segmente verticale de lungimi date prin

$$\min\left[\overline{Q}_1 - x_{(1)}, \mathtt{w}\right], \quad \text{respectiv} \quad \min\left[x_{(n)} - \overline{Q}_3, \mathtt{w}\right],$$

unde avem valoarea implicită $\mathtt{w} = \frac{3}{2}\left(\overline{Q}_3 - \overline{Q}_1\right)$. Datele care se află înafara acestor prelungiri se numesc `outliers` (erori grosolane) şi sunt marcate prin simbolul precizat prin parametrul `s` (implicit `s=+`). Dacă nu există astfel de date într–o parte sau alta, atunci segmentul din partea de jos se termină cu un marcaj.

Dacă `cr=1`, atunci dreptunghiurile sunt crestate pe laturile verticale, având mijloacele crestăturilor în dreptul medianelor, iar deschiderile crestăturilor prezintă estimaţii robuste pentru valorile medii teoretice. Valoarea `cr=0` (valoare implicită) implică faptul că dreptunghiurile nu sunt crestate.

Când `v`=0 (valoare implicită este `v=1`) reprezentările prezentate pâna aici sunt rotite cu 90^o, adică sunt prezentate pe orizontală.

Remarcăm faptul că dacă `x` este un vector, atunci poate apare un parametru suplimentar `g`, imediat după parametrul `x`. Parametrul `g` este un vector numeric sau de tip caracter, de aceeaşi lungime cu `x`, şi care permite gruparea datelor conţinute de `x`. O astfel de grupare este formată din acele date din `x`, pentru care valorile corespunzătoare din `g` sunt aceleaşi. Funcţia `boxplot` în acest caz va produce dreptunghiuri pentru fiecare grupare în parte.

Programul 3.3.35. Programul următor generează `n` vectori aleatori, care urmează legea normală bidimensională, în matricea `X` de tipul `n`$\times$`2`, după care, folosind funcţia `boxplot`, reprezintă grafic datele pentru prima variabilă şi respectiv pentru a doua variabilă.

```
clear, mu(1)=input('m1='); mu(2)=input('m2=');
v(1,1)=input('sigma1^2=');
v(2,2)=input('sigma2^2=');
v(1,2)=input('Cov(X,Y)='); v(2,1) =v(1,2);
if det(v) <= 0
   error('Matricea v nu e pozitiv definita!')
end
n=input('n='); X=mvnrnd(mu,v,n); boxplot(X)
xlabel('Numarul variabilelor'),
ylabel('Valorile variabilelor')
```

În urma executării programului, cu valorile `mu1=10`, `mu2=10`, $\mathtt{v} = \begin{pmatrix} 2 & -2 \\ -2 & 3 \end{pmatrix}$ şi `n=50`, se obţine Figura 3.11.

3.4 Corelaţie şi regresie

Corelaţia (într-un sens larg) poate fi înţeleasă ca legătura care există între o caracteristică dependentă şi una sau mai multe caracteristici independente, iar *regresia* este

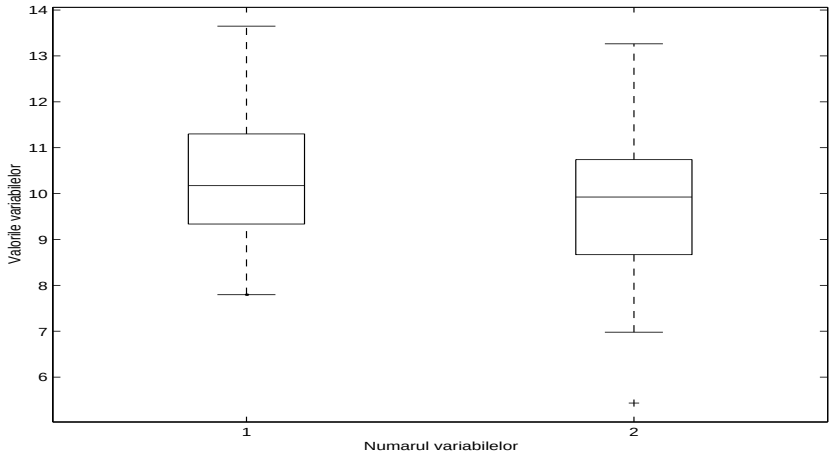

Figura 3.11: Reprezentare de date cu `boxplot`

$X \setminus Y$	y_1	y_2	$\dots$	y_n	
x_1	f_{11}	f_{12}	$\dots$	f_{1n}	$f_{1\cdot}$
x_2	f_{21}	f_{22}	$\dots$	f_{2n}	$f_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_m	f_{m1}	f_{m2}	$\dots$	f_{mn}	$f_{m\cdot}$
	$f_{\cdot 1}$	$f_{\cdot 2}$	$\dots$	$f_{\cdot n}$	$f_{\cdot\cdot} = N$

Tabelul 3.2: Tabel de corelaţie

metoda prin care se stabileşte acestă legătură.

Să considerăm caracteristicile cantitative X şi Y relative la colectivitatea $\mathcal{C}$. Datele statistice primare sunt $(x'_k, y'_k)\,,\ k = \overline{1, N}$, care după grupare sunt prezentate în Tabelul de corelaţie 3.2. unde f_{ij} reprezintă frecvenţa absolută a clasei (x_i, y_j). De asemenea, avem că

$$\sum_{j=1}^{n} f_{ij} = f_{i\cdot}, \quad \sum_{i=1}^{m} f_{ij} = f_{\cdot j}, \quad \sum_{i=1}^{m} f_{i\cdot} = \sum_{j=1}^{n} f_{\cdot j} = f_{\cdot\cdot} = N.$$

Definiţia 3.4.1. *Numim* moment de ordinul (k_1, k_2) *al distribuţiei statistice a caracteristicii bidimensionale* (X, Y)*, valoarea numerică*

$$\overline{\nu}_{k_1 k_2} = \frac{1}{N} \sum_{i=1}^{N} x_i'^{k_1} y_i'^{k_2} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} x_i^{k_1} y_j^{k_2} = \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} x_i^{k_1} y_j^{k_2},$$

unde $p_{ij} = \frac{f_{ij}}{N}$ *este frecvenţa relativă a clasei* (x_i, y_j).

Definiţia 3.4.2. *Numim* moment centrat de ordinul (k_1, k_2) *al distribuţiei statistice bidimensionale* (X, Y)*, valoarea numerică*

$$\overline{\mu}_{k_1 k_2} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i' - \overline{x}\right)^{k_1}\left(y_i' - \overline{y}\right)^{k_2} = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\left(x_i - \overline{x}\right)^{k_1}\left(y_j - \overline{y}\right)^{k_2}$$
$$= \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}\left(x_i - \overline{x}\right)^{k_1}\left(y_j - \overline{y}\right)^{k_2},$$

unde

$$\overline{x} = \overline{\nu}_{10} = \frac{1}{N}\sum_{i=1}^{m} f_{i\cdot} x_i, \quad \overline{y} = \overline{\nu}_{01} = \frac{1}{N}\sum_{j=1}^{n} f_{\cdot j} y_j.$$

Observaţia 3.4.3. Remarcăm, printre momentele centrate, dispersiile pentru distribuţiile statistice ale caracteristicilor X şi respectiv Y, anume

$$\overline{\sigma}_X^2 = \overline{\mu}_{20} = \frac{1}{N}\sum_{i=1}^{m} f_{i\cdot}\left(x_i - \overline{x}\right)^2, \quad \overline{\sigma}_Y^2 = \overline{\mu}_{02} = \frac{1}{N}\sum_{j=1}^{n} f_{\cdot j}\left(y_j - \overline{y}\right)^2.$$

Definiţia 3.4.4. *Numim* coeficient de corelaţie *(al lui Pearson) al distribuţiei statistice bidimensionale* (X, Y)*, raportul*

$$\overline{r} = \frac{\overline{\mu}_{11}}{\sqrt{\overline{\mu}_{20}}\sqrt{\overline{\mu}_{02}}} = \frac{\overline{\nu}_{11} - \overline{\nu}_{10}\overline{\nu}_{01}}{\overline{\sigma}_X \overline{\sigma}_Y} = \frac{\overline{\nu}_{11} - \overline{x}\,\overline{y}}{\overline{\sigma}_X \overline{\sigma}_Y}.$$

Observaţia 3.4.5. Folosind datele statistice negrupate avem formulele de calcul pentru coeficientul $\overline{r}$ de corelaţie

$$\overline{r} = \frac{\sum_{i=1}^{N}\left(x_i' - \overline{x}\right)\left(y_i' - \overline{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_i' - \overline{x}\right)^2}\sqrt{\sum_{i=1}^{N}\left(y_i' - \overline{y}\right)^2}}$$

sau

$$\overline{r} = \frac{\sum_{i=1}^{N} x_i' y_i' - \frac{1}{N}\left(\sum_{i=1}^{N} x_i'\right)\left(\sum_{i=1}^{N} y_i'\right)}{\sqrt{\sum_{i=1}^{N} x_i'^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i'\right)^2}\sqrt{\sum_{i=1}^{N} y_i'^2 - \frac{1}{N}\left(\sum_{i=1}^{N} y_i'\right)^2}}.$$

Observaţia 3.4.6. Folosind inegalitatea Cauchy–Schwarz–Buniakovski

$$\left[\sum_{i=1}^{N}\left(x_i'-\overline{x}\right)\left(y_i'-\overline{y}\right)\right]^2 \leqslant \left[\sum_{i=1}^{N}\left(x_i'-\overline{x}\right)^2\right]\left[\sum_{i=1}^{N}\left(y_i'-\overline{y}\right)^2\right]$$

se stabileşte că $|\,\overline{r}\,| \leqslant 1$. De asemenea, dacă $|\,\overline{r}\,| = 1$, atunci între cele două caracteristici există o legătură liniară, şi invers. Când $\overline{r} = 0$, spunem că cele două caracteristici sunt (liniar) *necorelate*. Prin urmare, coeficientul de corelaţie poate fi folosit pentru a stabili dacă între cele două caracteristici există sau nu o legătură liniară. În cazul în care caracteristica bidimensională (X, Y) urmează legea normală bidimensională $\overline{r} = 0$ implică faptul că cele două caracteristici sunt independente.

Exemplul 3.4.7. Să calculăm câţiva din parametrii definiţi pentru datele statistice de la Exemplul 3.2.20.

Pentru a calcula coeficientul de corelaţie, calculăm prima dată valorile medii ale caracteristicilor X şi Y. Remarcăm, pentru aceasta, că distribuţiile statistice ale caracteristicilor X şi Y sunt respectiv

$$X\begin{pmatrix} 126 & 127 & 128 & 129 & 130 & 131 & 132 & 133 & 134 & 135 & 136 & 137 \\ 6 & 9 & 21 & 29 & 70 & 152 & 64 & 38 & 18 & 8 & 6 & 4 \end{pmatrix},$$

$$Y\begin{pmatrix} 24 & 25 & 26 & 27 & 28 & 29 & 30 & 31 & 32 & 33 & 34 & 35 \\ 2 & 9 & 16 & 30 & 84 & 106 & 89 & 45 & 24 & 11 & 7 & 2 \end{pmatrix}.$$

Obţinem astfel că

$$\overline{x} = \frac{1}{425}\left(6\cdot 126 + 9\cdot 127 + \cdots + 4\cdot 137\right) = 131.054\,\text{cm},$$
$$\overline{y} = \frac{1}{425}\left(2\cdot 24 + 9\cdot 25 + \cdots + 2\cdot 35\right) = 29.2447\,\text{kg}.$$

De asemenea, calculând dispersiile caracteristicilor se obţine

$$\begin{aligned}\overline{\sigma}_X^2 =& \frac{1}{N}\left[\sum_{i=1}^{m} f_i.x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{m} f_i.x_i\right)^2\right]\\ =& \frac{1}{425}\left[\left(6\cdot 126^2 + 9\cdot 127^2 + \ldots + 4\cdot 137^2\right) - \frac{1}{425}\left(6\cdot 126 + \ldots + 4\cdot 137\right)^2\right]\\ =& \frac{1}{425}\left[7300920 - \frac{1}{425}\cdot 55698^2\right] = 3.45354,\end{aligned}$$

$$\begin{aligned}\overline{\sigma}_Y^2 =& \frac{1}{N}\left[\sum_{j=1}^{n} f_{\cdot j} y_j^2 - \frac{1}{N}\left(\sum_{j=1}^{n} f_{\cdot j} y_j\right)^2\right] \\ =& \frac{1}{425}\left[\left(2\cdot 24^2 + 9\cdot 25^2 + \ldots + 2\cdot 35^2\right) - \frac{1}{425}\left(2\cdot 24 + 9\cdot 25 + \ldots + 2\cdot 35\right)^2\right] \\ =& \frac{1}{425}\left[364907 - \frac{1}{425}\cdot 12429^2\right] = 3.35188.\end{aligned}$$

Covarianţa dintre X şi Y se calculează cu formula

$$\begin{aligned}\mathrm{Cov}\,(X,Y) =& \frac{1}{N}\left[\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} x_i y_j - \frac{1}{N}\left(\sum_{i=1}^{m} f_{i\cdot} x_i\right)\left(\sum_{j=1}^{n} f_{\cdot j} y_j\right)\right] \\ =& \frac{1}{425}\left[1629961 - \frac{1}{425}\cdot 55698\cdot 12429\right] = 2.56323.\end{aligned}$$

Astfel se ajunge la coeficientul de corelaţie

$$\overline{r} = \frac{\mathrm{Cov}\,(X,Y)}{\sqrt{\overline{\sigma}_X^2\,\overline{\sigma}_Y^2}} = \frac{2.56323}{\sqrt{3.45354\times 3.35188}} = 0.753374.$$

Programul 3.4.8. Programul Matlab, care urmează, efectuează aceste calcule

```
f=[1,1,3,1,zeros(1,8);
   0,3,4,1,0,1,zeros(1,6);
   1,4,5,7,1,2,1,zeros(1,5);
   0,1,2,6,9,6,4,1,zeros(1,4);
   zeros(1,2),1,7,36,17,6,2,1,zeros(1,3);
   zeros(1,2),1,6,27,56,39,18,4,1,zeros(1,2);
   zeros(1,3),2,10,16,26,7,2,1,zeros(1,2);
   zeros(1,4),1,7,10,11,6,2,1,0;
   zeros(1,5),1,2,4,7,3,1,0; zeros(1,6),1,2,3,1,1,0;
   zeros(1,8),1,2,2,1; zeros(1,9),1,2,1];
x=[126:137]; y=[24:35]; fx=sum(f'); fy=sum(f);
xa=sum(fx.*x)/(sum(fx)); ya=sum(fy.*y)/(sum(fy));
vx=(sum(fx.*x.^2)-(sum(fx.*x))^2/sum(fx))/sum(fx);
vy=(sum(fy.*y.^2)-(sum(fy.*y))^2/sum(fy))/sum(fy);
cov=(x*f*y'-sum(fx.*x)*sum(fy.*y)/sum(fx))/sum(fx);
r=cov/(sqrt(vx*vy));
fprintf(' xa=%8.3f\n ya=%8.4f\n',xa,ya)
fprintf(' vx=%8.5f\n vy=%8.5f\n',vx,vy)
fprintf(' cov=%7.5f\n r=%9.6f',cov,r)
```

iar rezultatele sunt următoarele:

```
 xa= 131.054
 ya= 29.2447
 vx= 3.45354
```

```
vy= 3.35188
cov=2.56323
r= 0.753374
```

Definiţia 3.4.9. *Numim* valoare medie condiţionată *a distribuţiei statistice a caracteristicii Y în raport cu $X = x_j$, valoarea numerică*

$$\overline{y}_i = \overline{y}\left(x_i\right) = \frac{1}{f_{i\cdot}} \sum_{j=1}^{n} f_{ij} y_j = \sum_{j=1}^{n} \frac{f_{ij}}{f_{i\cdot}} y_j, \quad i = \overline{1,m},$$

şi respectiv valoare medie condiţionată *a distribuţiei statistice a caracteristicii X în raport cu $Y = y_j$, valoarea numerică*

$$\overline{x}_j = \overline{x}\left(y_j\right) = \frac{1}{f_{\cdot j}} \sum_{i=1}^{m} f_{ij} x_i = \sum_{i=1}^{m} \frac{f_{ij}}{f_{\cdot j}} x_i, \quad j = \overline{1,n}.$$

Definiţia 3.4.10. *Curba de ecuaţie $y = f\left(x\right)$ pe care se situează punctele de coordonate $\left(x_i, \overline{y}_i\right)$, $i = \overline{1,m}$, se numeşte* curba de regresie *a lui Y în raport cu X, iar curba de ecuaţie $x = g\left(y\right)$ pe care se situează punctele de coordonate $\left(\overline{x}_j, y_j\right)$, $j = \overline{1,n}$, se numeşte* curba de regresie *a lui X în raport cu Y.*

Definiţia 3.4.11. *Numim* dispersie condiţionată *a distribuţiei statistice a caracteristicii Y în raport cu $X = x_i$, valoarea numerică*

$$\overline{\sigma}^2_{Y|x_i} = \frac{1}{f_{i\cdot}} \sum_{j=1}^{n} f_{ij} \left(y_j - \overline{y}_i\right)^2 = \sum_{j=1}^{n} \frac{f_{ij}}{f_{i\cdot}} \left(y_j - \overline{y}_i\right)^2, \quad i = \overline{1,m},$$

şi respectiv dispersie condiţionată *a distribuţiei statistice a caracteristicii X în raport cu $Y = y_j$, valoarea numerică*

$$\overline{\sigma}^2_{X|y_j} = \frac{1}{f_{\cdot j}} \sum_{i=1}^{m} f_{ij} \left(x_i - \overline{x}_j\right)^2 = \sum_{i=1}^{m} \frac{f_{ij}}{f_{\cdot j}} \left(x_i - \overline{x}_j\right)^2, \quad j = \overline{1,n}.$$

Definiţia 3.4.12. *Numim* dispersia condiţionată *a distribuţiei statistice a lui Y în raport cu distribuţia statistică a lui X, valoarea numerică*

$$\overline{\sigma}^2_{Y|X} = \frac{1}{N} \sum_{i=1}^{m} f_{i\cdot} \overline{\sigma}^2_{Y|x_i} = \sum_{i=1}^{m} p_{i\cdot} \overline{\sigma}^2_{Y|x_i}$$

şi respectiv dispersia condiţionată *a distribuţiei statistice a lui X în raport cu distribuţia statistică a lui Y, valoarea numerică*

$$\overline{\sigma}^2_{X|Y} = \frac{1}{N} \sum_{j=1}^{n} f_{\cdot j} \overline{\sigma}^2_{X|y_j} = \sum_{j=1}^{n} p_{\cdot j} \overline{\sigma}^2_{X|y_j},$$

unde $p_{i\cdot} = \frac{f_{i\cdot}}{N}$ *este frecvenţa relativă a clasei* x_i*, iar* $p_{\cdot j} = \frac{f_{\cdot j}}{N}$ *este frecvenţa relativă a clasei* y_j*.*

Proprietatea 3.4.13. *Dispersiile condiţionate definite mai înainte satisfac relaţiile*

$$\overline{\sigma}_Y^2 = \overline{\sigma}_{Y|X}^2 + \overline{\sigma}_{\overline{Y}|X}^2, \quad \overline{\sigma}_X^2 = \overline{\sigma}_{X|Y}^2 + \overline{\sigma}_{\overline{X}|Y}^2,$$

unde

$$\overline{\sigma}_{\overline{Y}|X}^2 = \frac{1}{N}\sum_{i=1}^{m} f_{i\cdot}\left(\overline{y}_i - \overline{y}\right)^2, \quad \overline{\sigma}_{\overline{X}|Y}^2 = \frac{1}{N}\sum_{j=1}^{n} f_{\cdot j}\left(\overline{x}_j - \overline{x}\right)^2,$$

adică sunt dispersiile valorilor medii condiţionate.

Demonstraţie. Se porneşte de la identitatea

$$\left(y_j - \overline{y}\right)^2 = \left(y_j - \overline{y}_i\right)^2 + \left(\overline{y}_i - \overline{y}\right)^2 + 2\left(y_j - \overline{y}_i\right)\left(\overline{y}_i - \overline{y}\right),$$

care se înmulţeşte cu $\frac{f_{ij}}{f_{i\cdot}}$ şi se însumează după $j = \overline{1,n}$. Astfel se obţine

$$\begin{aligned}\sum_{j=1}^{n}\frac{f_{ij}}{f_{i\cdot}}\left(y_j - \overline{y}\right)^2 =& \sum_{j=1}^{n}\frac{f_{ij}}{f_{i\cdot}}\left(y_j - \overline{y}_i\right)^2 + \sum_{j=1}^{n}\frac{f_{ij}}{f_{i\cdot}}\left(\overline{y}_i - \overline{y}\right)^2 + \\ &+ 2\sum_{j=1}^{n}\frac{f_{ij}}{f_{i\cdot}}\left(y_j - \overline{y}_i\right)\left(\overline{y}_i - \overline{y}\right) \\ =& \overline{\sigma}_{Y|x_i}^2 + \left(\overline{y}_i - \overline{y}\right)^2 + 2\left(\overline{y}_i - \overline{y}\right)\left(\overline{y}_i - \overline{y}_i\right) = \overline{\sigma}_{Y|x_i}^2 + \left(\overline{y}_i - \overline{y}\right)^2.\end{aligned}$$

Reţinem extremităţile acestui şir de egalităţi, le înmulţim cu $\frac{f_{i\cdot}}{N}$ şi le însumăm după $i = \overline{1,m}$. Rezultă în acest mod

$$\begin{aligned}\overline{\sigma}_Y^2 =& \frac{1}{N}\sum_{j=1}^{n} f_{\cdot j}\left(y_j - \overline{y}\right)^2 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\left(y_j - \overline{y}\right)^2 \\ =& \frac{1}{N}\sum_{i=1}^{m} f_{i\cdot}\overline{\sigma}_{Y|x_i}^2 + \frac{1}{N}\sum_{i=1}^{m} f_{i\cdot}\left(\overline{y}_i - \overline{y}\right)^2 = \overline{\sigma}_{Y|X}^2 + \overline{\sigma}_{\overline{Y}|X}^2,\end{aligned}$$

de unde se obţine prima relaţie. Analog se obţine şi cealaltă relaţie. □

Definiţia 3.4.14. *Numim* raport de corelaţie *al distribuţiei statistice a caracteristicii* Y *faţă de distribuţia statisticii* X*, valoarea numerică*

$$\overline{\eta}_{Y|X} = \sqrt{1 - \frac{\overline{\sigma}_{Y|X}^2}{\overline{\sigma}_Y^2}} = \sqrt{\frac{\overline{\sigma}_{\overline{Y}|X}^2}{\overline{\sigma}_Y^2}}$$

şi în mod analog avem

$$\overline{\eta}_{X|Y}=\sqrt{1-\frac{\overline{\sigma}_{X|Y}^2}{\overline{\sigma}_X^2}}=\sqrt{\frac{\overline{\sigma}_{\overline{X}|Y}^2}{\overline{\sigma}_X^2}}.$$

Observaţia 3.4.15. Când $\overline{y}_1=\overline{y}_2=\cdots=\overline{y}_m$, atunci $\overline{y}=\overline{y}_i$, prin urmare $\overline{\sigma}_{\overline{Y}|X}^2=0\left(\overline{\eta}_{Y|X}=0\right)$, ceea ce înseamnă absenţa dependenţei în medie. Pe de altă parte, $\overline{\eta}_{Y|X}=1$ când $\overline{\sigma}_{Y|x_i}^2=0$, $i=\overline{1,m}$, adică pentru fiecare clasă x_i valorile caracteristicii Y sunt aceleaşi.

În cazul în care, relativ la caracteristica X, există numai două clase avem

$$\overline{y}=\frac{f_{1.}}{N}\overline{y}_1+\frac{f_{2.}}{N}\overline{y}_2.$$

Dacă se înlocuieşte această expresie în formula

$$\overline{\sigma}_{\overline{Y}|X}^2=\frac{1}{N}\left[f_{1.}\left(\overline{y}_1-\overline{y}\right)^2+f_{2.}\left(\overline{y}_2-\overline{y}\right)^2\right]$$

se obţine că

$$\overline{\sigma}_{\overline{Y}|X}^2=\frac{f_{1.}f_{2.}}{N^2}\left(\overline{y}_1-\overline{y}_2\right)^2,$$

deci

$$\overline{\eta}_{Y|X}^2=\frac{f_{1.}f_{2.}\left(\overline{y}_1-\overline{y}_2\right)^2}{N^2\overline{\sigma}_Y^2}.$$

Aplicaţia 3.4.16. *(Determinarea curbelor de regresie).* Prezentăm în continuare *metoda celor mai mici pătrate* pentru determinarea ecuaţiilor curbelor de regresie.

Presupunem că din reprezentarea în plan a punctelor $(x_i,\overline{y}_i)$, $i=\overline{1,m}$, curba de regresie a lui Y în raport cu X este de forma $y=y(x)=f(x;a_1,a_2,\ldots,a_s)$. Vom determina parametrii $a_k,\ k=\overline{1,s}$, astfel încât

$$\begin{aligned}S(a_1,a_2,\ldots,a_s)&=\sum_{i=1}^{N}\left[y_i'-y\left(x_i'\right)\right]^2=\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}\left[y_j-y\left(x_i\right)\right]^2\\&=\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}\left[y_j-f\left(x_i;a_1,a_2,\ldots,a_s\right)\right]^2\end{aligned}$$

să fie minimă.

Punctul de minim $(\overline{a}_1, \overline{a}_2, \ldots, \overline{a}_s)$ al funcţiei S se obţine prin rezolvarea *sistemului normal de ecuaţii*, rezultat din

$$\frac{\partial S}{\partial a_k} = -2 \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} \left[y_j - f\left(x_i; a_1, a_2, \ldots, a_s\right) \right] \frac{\partial f\left(x_i; a_1, a_2, \ldots, a_s\right)}{\partial a_k} = 0$$

pentru $k = \overline{1, s}$. Ecuaţia curbei de regresie va fi

$$y = f\left(x; \overline{a}_1, \overline{a}_2, \ldots, \overline{a}_s\right).$$

La fel se determină şi ecuaţia curbei de regresie a lui X în raport cu Y.

Aplicaţia 3.4.17. *(Drepte de regresie).* Considerăm cazul liniar, adică ecuaţia curbei de regresie este $y = y(x) = ax + b$. Expresia minimizată este

$$S(a, b) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} \left(y_j - a x_i - b\right)^2,$$

care conduce la sistemul normal de ecuaţii

$$\begin{cases} \left(\sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} f_{ij} x_i^2 \right) a + \left(\sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} f_{ij} x_i \right) b = \sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} f_{ij} x_i y_j \\ \left(\sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} f_{ij} x_i \right) a + \left(\sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} f_{ij} \right) b = \sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} f_{ij} y_j, \end{cases}$$

sau

$$\begin{cases} \overline{\nu}_{20} a + \overline{\nu}_{10} b = \overline{\nu}_{11} \\ \overline{\nu}_{10} a + \overline{\nu}_{00} b = \overline{\nu}_{01}. \end{cases}$$

Soluţia acestui sistem liniar este

$$\overline{a} = \frac{\overline{\nu}_{11} - \overline{\nu}_{10}\overline{\nu}_{01}}{\overline{\nu}_{20} - \overline{\nu}_{10}^2} = \frac{\overline{\nu}_{11} - \overline{\nu}_{10}\overline{\nu}_{01}}{\sqrt{\overline{\nu}_{20} - \overline{\nu}_{10}^2}\sqrt{\overline{\nu}_{02} - \overline{\nu}_{01}^2}} \frac{\sqrt{\overline{\nu}_{02} - \overline{\nu}_{01}^2}}{\sqrt{\overline{\nu}_{20} - \overline{\nu}_{10}^2}} = \overline{r}\, \frac{\overline{\sigma}_Y}{\overline{\sigma}_X},$$
$$\overline{b} = \overline{\nu}_{01} - \overline{a}\,\overline{\nu}_{10} = \overline{y} - \overline{a}\,\overline{x}.$$

Se obţine în acest fel ecuaţia dreptei de regresie a lui Y în raport cu X ca fiind

$$y - \overline{y} = \overline{r}\, \frac{\overline{\sigma}_Y}{\overline{\sigma}_X} \left(x - \overline{x}\right).$$

În mod analog se ajunge la ecuaţia dreptei de regresie a lui X în raport cu Y

$$x - \overline{x} = \overline{r}\,\frac{\overline{\sigma}_X}{\overline{\sigma}_Y}\,(y - \overline{y})\,. \tag{3.4.1}$$

Punctul de intersecţie al celor două drepte de regresie are coordonatele $(\overline{x}, \overline{y})$ şi se numeşte *centrul de greutate* al distribuţiei statistice a caracteristicii bidimensionale (X, Y).

Dacă se notează

$$\overline{a}_{Y|X} = \overline{r}\,\frac{\overline{\sigma}_Y}{\overline{\sigma}_X}$$

coeficientul unghiular al dreptei de regresie a lui Y în raport cu X (numit *coeficientul de regresie* al lui Y în raport cu X) şi cu

$$\overline{a}_{X|Y} = \overline{r}\,\frac{\overline{\sigma}_X}{\overline{\sigma}_Y}$$

coeficientul de regresie al lui X în raport cu Y, atunci

$$|\overline{r}| = \sqrt{\overline{a}_{Y|X}\overline{a}_{X|Y}}.$$

De asemenea se constată că sign $\left(\overline{a}_{X|Y}\right)$ = sign $\left(\overline{a}_{Y|X}\right)$. Se arată că unghiul α format de cele două drepte de regresie este dat prin relaţia

$$\operatorname{tg}\alpha = \frac{1-\overline{r}^2}{\overline{r}^2}\,\frac{\overline{\sigma}_X\overline{\sigma}_Y}{\overline{\sigma}_X^2 + \overline{\sigma}_Y^2}.$$

Folosind această relaţie se pot trage următoarele concluzii:

a) dacă $|\overline{r}| = 1$ atunci $\alpha = 0$, deci dreptele de regresie se confundă, cu specificaţia că pentru $\overline{r} = -1$ dreptele au panta (coeficientul unghiular) negativă, iar pentru $\overline{r} = 1$ panta este pozitivă.

b) dacă X şi Y sunt independente atunci $\overline{r} = 0$, deci $\alpha = \frac{\pi}{2}$ (dreptele de regresie sunt perpendiculare).

Observaţia 3.4.18. Alte tipuri de curbe de regresie mai des întâlnite şi care pot fi liniarizate sunt:

1. $y = ab^x$ (exponenţială), care prin logaritmare se liniarizează după cum urmează $\log y = \log a + x\,\log b$, luând $z = \log y$, $A = \log a$, $B = \log b$;
2. $y = \frac{a}{x} + b$ (hiperbolică), care se liniarizează dacă se notează $z = \frac{1}{x}$;
3. $\frac{1}{y} = \frac{a}{x} + b$ sau $y = \frac{1}{\frac{a}{x}+b}$, care se liniarizează dacă se notează $u = \frac{1}{x}$, $v = \frac{1}{y}$;

4. $y = a \log x + b$ (logaritmică), care se liniarizează dacă se notează $z = \log x$;
5. $y = b\mathrm{e}^{ax}$ (exponenţială), care prin logaritmare se liniarizează $\ln y = \ln b + ax$, luând $z = \ln y$;
6. $y = b\mathrm{e}^{\frac{a}{x}}$, care prin logaritmare se liniarizează $\ln y = \ln b + \frac{a}{x}$, luând $u = \frac{1}{x}$, $v = \ln y$;
7. $y = bx^a$, care prin logaritmare se liniarizează, $\log y = \log b + a \log x$, luând $u = \log x$, $v = \log y$;
8. $\frac{1}{y} = a\mathrm{e}^{-x} + b$ sau $y = \frac{1}{a\mathrm{e}^{-x}+b}$, care se liniarizează dacă se ia $u = \mathrm{e}^{-x}$ şi $v = \frac{1}{y}$.

Această ultimă legătură este un caz particular al *legăturii logistice* definită prin

$$y = \frac{1}{ae^{-cx} + b}.$$

Printre legăturile ce nu pot fi liniarizate amintim:

$y = ax^b + c \log x$,

$y = ax^b e^{cx}$,

$y = a + bx + ce^{dx}$,

a doua putând fi adusă la forma polinomială.

Exemplul 3.4.19. Să considerăm din nou datele statistice de la Exemplul 3.2.20, să calculăm noile caracteristici numerice definite şi să determinăm ecuaţiile dreptelor de regresie.

Valorile medii condiţionate şi dispersiile condiţionate ale lui Y în raport cu X se calculează cu formulele

$$\overline{y}_i = \frac{1}{f_{i\cdot}} \sum_{j=1}^{n} f_{ij} y_j, \quad i = \overline{1, m},$$

şi

$$\overline{\sigma}^2_{Y \mid x_i} = \frac{1}{f_{i\cdot}} \sum_{j=1}^{n} f_{ij} \left(y_j - \overline{y}_i\right)^2, \quad i = \overline{1, m}.$$

În mod analog se obţin valorile medii condiţionate şi dispersiile condiţionate ale lui X in raport cu Y.

Aceste valori calculate sunt trecute în tabelele următoare.

i	$\overline{y}_i$	$\overline{\sigma}^2_{Y\mid x_i}$
1	25.67	0.889
2	26.11	1.432
3	26.62	2.141
4	28.14	1.843
5	28.43	1.045
6	29.32	1.363
7	29.56	1.340
8	30.63	1.706
9	31.67	1.444
10	31.88	1.359
11	33.50	0.917
12	34.00	0.500

şi

j	$\overline{x}_j$	$\overline{\sigma}^2_{X\mid y_j}$
1	127.00	1.000
2	127.56	0.691
3	127.81	1.902
4	129.43	2.112
5	130.46	0.844
6	130.94	1.204
7	131.44	1.280
8	132.00	1.778
9	133.12	2.276
10	134.09	2.992
11	135.43	1.959
12	136.50	0.250

Dispersia condiţionată a lui Y în raport cu X va fi

$$\overline{\sigma}^2_{Y|X}=\frac{1}{N}\sum_{i=1}^{m} f_{i\cdot}\overline{\sigma}^2_{Y|x_i}=\frac{1}{425}\left(6\times 0.889+9\times 1.432+\cdots+4\times 0.500\right)=1.39279.$$

În mod analog dispersia condiţionată a lui X în raport cu Y este

$$\overline{\sigma}^2_{X|Y}=\frac{1}{N}\sum_{j=1}^{n} f_{\cdot j}\overline{\sigma}^2_{X|y_j}=\frac{1}{425}\left(2\times 1.000+9\times 0.691+\cdots+2\times 0.250\right)=1.40291.$$

Dispersia condiţionată a lui Y în raport cu X se poate calcula şi cu formula

$$\overline{\sigma}^2_{Y\mid X}=\overline{\sigma}^2_{Y}-\overline{\sigma}^2_{\overline{Y}\mid X},$$

unde

$$\overline{\sigma}^2_{\overline{Y}\mid X}=\frac{1}{N}\sum_{i=1}^{m} f_{i\cdot}\left(\overline{y}_i-\overline{y}\right)^2$$

este dispersia valorilor medii. Anume, avem că

$$\overline{\sigma}^2_{\overline{Y}\mid X}=\frac{1}{425}\left[6\left(25.67-29.2447\right)^2+\cdots+4\left(34.00-29.2447\right)^2\right]=1.95909,$$

de unde

$$\overline{\sigma}^2_{Y\mid X}=3.35188-1.95909=1.39279.$$

În mod analog,

$$\overline{\sigma}^2_{X\mid Y}=\overline{\sigma}^2_{X}-\overline{\sigma}^2_{\overline{X}\mid Y},$$

unde

$$\overline{\sigma}^2_{\overline{X}\,|\,Y} = \frac{1}{N}\sum_{j=1}^{n} f_{\cdot j}\left(\overline{x}_j - \overline{x}\right)^2,$$

adică

$$\overline{\sigma}^2_{\overline{X}\,|\,Y} = \frac{1}{425}\left[2\left(127 - 131.054\right)^2 + \cdots + 2\left(136.5 - 131.054\right)^2\right] = 2.05063.$$

În acest fel se obţine că

$$\overline{\sigma}^2_{X\,|\,Y} = 3.45354 - 2.05063 = 1.40291.$$

Raportul de corelaţie al lui Y faţă de X este

$$\overline{\eta}_{Y\,|\,X} = \sqrt{1 - \frac{\overline{\sigma}^2_{Y\,|\,X}}{\overline{\sigma}^2_Y}} = \sqrt{\frac{\overline{\sigma}^2_{\overline{Y}\,|\,X}}{\overline{\sigma}^2_Y}} = \sqrt{\frac{1.95909}{3.35188}} = 0.764509,$$

iar raportul de corelaţie al lui X faţă de Y este

$$\overline{\eta}_{X\,|\,Y} = \sqrt{1 - \frac{\overline{\sigma}^2_{X\,|\,Y}}{\overline{\sigma}^2_X}} = \sqrt{\frac{\overline{\sigma}^2_{\overline{X}\,|\,Y}}{\overline{\sigma}^2_X}} = \sqrt{\frac{2.05063}{3.45354}} = 0.770568.$$

Ecuaţiile dreptelor de regresie respectiv a lui Y în raport cu X şi a lui X în raport cu Y au ecuaţiile

$$y - \overline{y} = \overline{r}\,\frac{\overline{\sigma}_Y}{\overline{\sigma}_X}\left(x - \overline{x}\right), \quad x - \overline{x} = \overline{r}\,\frac{\overline{\sigma}_X}{\overline{\sigma}_Y}\left(y - \overline{y}\right).$$

Având în vedere că

$$\overline{r}\,\frac{\overline{\sigma}_Y}{\overline{\sigma}_X} = 0.753374\sqrt{\frac{3.35188}{3.45354}} = 0.742203,$$
$$\overline{r}\,\frac{\overline{\sigma}_X}{\overline{\sigma}_Y} = 0.753374\sqrt{\frac{3.45354}{3.35188}} = 0.764713,$$

avem ecuaţiile

$$y - 29.2447 = 0.742203\left(x - 131.054\right), \quad x - 131.054 = 0.764713\left(y - 29.2447\right).$$

Centrul de greutate al distribuţiei statistice (X, Y) este punctul de coordonate $(\overline{x}; \overline{y}) = (131.05; 29.24)$, adică punctul de intersecţie a celor două drepte de regresie.

Coeficientul de regresie al lui Y în raport cu X este coeficientul unghiular al dreptei de regresie a lui Y în raport cu X, adică

$$\overline{a}_{Y|X} = \overline{r}\,\frac{\overline{\sigma}_Y}{\overline{\sigma}_X} = 0.742203.$$

În mod analog, coeficientul de regresie al lui X în raport cu Y este

$$\overline{a}_{X|Y} = \overline{r}\,\frac{\overline{\sigma}_X}{\overline{\sigma}_Y} = 0.764713.$$

Pentru determinarea unghiului α format de cele două drepte de regresie avem formula

$$\operatorname{tg}\alpha = \frac{1-\overline{r}^2}{\overline{r}^2}\,\frac{\overline{\sigma}_X\overline{\sigma}_Y}{\overline{\sigma}_X^2+\overline{\sigma}_Y^2},$$

adică

$$\operatorname{tg}\alpha = \frac{1-0.5675}{0.5675}\cdot\frac{\sqrt{3.45354\times 3.35188}}{3.45354+3.35188} = 0.381,$$

de unde $\alpha \cong 21^0 = 0.364$ rad.

Programul 3.4.20. Calculele de mai sus se pot face cu programul Matlab, care urmează:

```
f=[1,1,3,1,zeros(1,8);
   0,3,4,1,0,1,zeros(1,6);
   1,4,5,7,1,2,1,zeros(1,5);
   0,1,2,6,9,6,4,1,zeros(1,4);
   zeros(1,2),1,7,36,17,6,2,1,zeros(1,3);
   zeros(1,2),1,6,27,56,39,18,4,1,zeros(1,2);
   zeros(1,3),2,10,16,26,7,2,1,zeros(1,2);
   zeros(1,4),1,7,10,11,6,2,1,0;
   zeros(1,5),1,2,4,7,3,1,0; zeros(1,6),1,2,3,1,1,0;
   zeros(1,8),1,2,2,1; zeros(1,9),1,2,1];
x=[126:137]; y=[24:35]; fx=sum(f'); fy=sum(f);
xa=sum(fx.*x)/(sum(fx)); ya=sum(fy.*y)/(sum(fy));
vx=(sum(fx.*x.^2)-(sum(fx.*x))^2/sum(fx))/sum(fx);
vy=(sum(fy.*y.^2)-(sum(fy.*y))^2/sum(fy))/sum(fy);
cov=(x*f*y'-sum(fx.*x)*sum(fy.*y)/sum(fx))/sum(fx);
r=cov/(sqrt(vx*vy));
yb=f*y'./fx'; xb=f'*x'./fy';
for i=1:12
    syb(i)=f(i,:)*(y'-yb(i)).^2/fx(i);
end
for j=1:12
    sxb(j)=f(:,j)'*(x'-xb(j)).^2/fy(j);
end
syx=sum(fx.*syb)/sum(fx);
```

```
sxy=sum(fy.*sxb)/sum(fx);
syxb=sum(fx'.*(yb-ya).^2)/sum(fx);
sxyb=sum(fy'.*(xb-xa).^2)/sum(fx);
eyx=sqrt(syxb/vy); exy=sqrt(sxyb/vx);
ayx=r*sqrt(vy/vx); axy=r*sqrt(vx/vy);
tang=(1-r^2)/r^2*sqrt(vx*vy)/(vx+vy);
arctan=atan(tang); arc=180*arctan/pi;
fprintf('  i  |  yb   |  syb   |      j  |   xb   |  sxb\n')
for i=1:12
     fprintf('%3d  |%6.2f |%6.3f  |%7d  |%6.3f |%6.3f\n',...
         i,yb(i),syb(i),i,xb(i),sxb(i))
end
fprintf('\n syx=%8.5f,  sxy=%8.5f',syx,sxy)
fprintf('\n eyx=%8.6f,  exy=%8.6f',eyx,exy)
fprintf('\n ayx=%8.6f,  axy=%8.6f',ayx,axy)
fprintf('\n tg=%9.3f,  alpha=%6.0f',tang,arc)
```

În urma executării programului precedent, se obţin următoarele rezultate:

```
  i  |  yb   |  syb   |      j  |   xb   |  sxb
  1  | 25.67 | 0.889  |      1  |127.000 | 1.000
  2  | 26.11 | 1.432  |      2  |127.556 | 0.691
  3  | 26.62 | 2.141  |      3  |127.813 | 1.902
  4  | 28.14 | 1.843  |      4  |129.433 | 2.112
  5  | 28.43 | 1.045  |      5  |130.464 | 0.844
  6  | 29.32 | 1.363  |      6  |130.943 | 1.204
  7  | 29.56 | 1.340  |      7  |131.438 | 1.280
  8  | 30.63 | 1.706  |      8  |132.000 | 1.778
  9  | 31.67 | 1.444  |      9  |133.125 | 2.276
 10  | 31.88 | 1.359  |     10  |134.091 | 2.992
 11  | 33.50 | 0.917  |     11  |135.429 | 1.959
 12  | 34.00 | 0.500  |     12  |136.500 | 0.250

 syx= 1.39279,  sxy= 1.40291
 eyx=0.764509,  exy=0.770568
 ayx=0.742203,  axy=0.764713
 tg=    0.381,  alpha=    21
```

3.4.1 Funcţiile `cov` şi `corrcoef`

Aceste două funcţii sunt specifice sistemului Matlab de bază şi pot fi apelate prin:

```
C=cov(x)
C=cov(x,y)
r=corrcoef(x)
```

Funcţia `cov` calculează matricea covarianţelor pentru matricea `x`, unde fiecare coloană este considerată a fi o variabilă:

$$\mathtt{C}_{ij} = \frac{1}{m-1}\sum_{k=1}^{m}\left(\mathtt{x}_{ki} - \overline{\mathtt{x}}_i\right)\left(\mathtt{x}_{kj} - \overline{\mathtt{x}}_j\right), \quad i,j = \overline{1,m}.$$

Dacă x este vector este returnată varianţa vectorului, caz în care se poate folosi şi parametrul opţional y, care este vector de aceeaşi dimensiune cu x, când este calculată covarianţa dintre x şi y, pe lângă varianţele lui x respectiv y, adică

$$\mathtt{c} = \frac{1}{m-1}\sum_{k=1}^{m}(\mathtt{x}_k - \overline{\mathtt{x}})\,(\mathtt{y}_k - \overline{\mathtt{y}})\,.$$

Prin urmare, în acest caz, cov(x,y) produce acelaşi rezultat ca şi cov([x y]).

Remarcăm faptul că instrucţiunea diag(cov(x)) generează vectorul varianţelor fiecărei coloane a matricei x, iar sqrt(diag(cov(x))) este vectorul abaterilor standard pentru fiecare coloană a matricei x.

Funcţia corrcoef calculează matricea coeficienţilor de corelaţie pentru matricea x, unde fiecare coloană este considerată o variabilă:

$$\mathtt{r}_{ij} = \frac{\mathtt{c}_{ij}}{\sqrt{\mathtt{c}_{ii}\mathtt{c}_{jj}}}, \quad i,j = \overline{1,m}.$$

Programul 3.4.21. Programul următor generează n vectori aleatori, care urmează legea normală bidimensională, în matricea X de tipul $\mathtt{n}\times 2$, după care, folosind funcţiile cov şi corrcoef, calculează matricele covarianţelor şi a coeficienţilor de corelaţie, pentru cele două coloane ale matricei X.

```
clear, mu(1)=input('m1='); mu(2)=input('m2=');
v(1,1)=input('sigma1^2=');
v(2,2)=input('sigma2^2=');
v(1,2)=input('Cov(X,Y)='); v(2,1) =v(1,2);
if det(v) <= 0
   error('Matricea v nu e pozitiv definita!')
end
n=input('n='); X=mvnrnd(mu,v,n);
v=cov(X); r=corrcoef(X);
fprintf('Matricea covariantelor\n'), disp(v)
fprintf('Matricea coeficientilor de corelatie\n'), disp(r)
```

În urma executării programului, cu valorile mu1=100, mu2=-100, $\mathtt{v} = \begin{pmatrix} 5 & 4 \\ 4 & 4 \end{pmatrix}$ şi n=1000, se obţin rezultatele

```
Matricea covariantelor
    5.1327    4.0419
    4.0419    4.0487

Matricea coeficientilor de corelatie
    1.0000    0.8867
    0.8867    1.0000
```

3.4.2 Funcţiile `lsline`, `refline` şi `gline`

Funcţiile `lsline`, `refline` şi `gline` aparţin la *Statistics toolbox*, apelurile lor făcându–se prin:

```
lsline
refline
refline(m)
refline(m,n)
gline
gline(f)
```

Funcţia `lsline` reprezintă grafic în figura curentă dreptele de regresie, obţinute cu metoda celor mai mici pătrate, pentru fiecare curbă a figurii, care a fost reprezentată grafic cu instrucţiunea `plot`, dar nu prin linie continuă, întreruptă sau de tipul punct–linie.

Funcţia `refline` reprezintă grafic în figura curentă dreapta de ecuaţie `y=mx+n`. Dacă lipseşte parametrul `n`, se impune ca parametrul `m` să fie un vector având două componente, `m(1)` reprezentând panta dreptei, iar `m(2)=n`. Adică în acest caz dreapta, ce se va reprezenta grafic, va avea ecuaţia `y=m(1)x+m(2)`. Dacă lipsesc ambele argumente, `m` şi `n`, efectul funcţiei `refline` coincide cu cel al funcţiei `lsline`.

Funcţia `gline` reprezintă grafic în figura precizată prin parametrul `f`, iar dacă acesta lipseşte în figura curentă, segmentul de dreaptă prin marcarea cu ajutorul *mouse*-lui a capetelor segmentului.

Programul 3.4.22. Vom prezenta un program Matlab, care generează `N` vectori aleatori ce urmează legea normală bidimensională, reprezintă grafic norul statistic pentru aceste date, precum şi cele două drepte de regresie, a lui `Y` în raport cu `X`, respectiv a lui `X` în raport cu `Y`.

Dreapta de regresie a lui `Y` în raport cu `X` se va reprezenta grafic cu funcţia `lsline`, iar dreapta de regresie a lui `X` în raport cu `Y` se va face cu funcţia `refline`, având în vedere parametrii `m` şi `n` ai dreptei de regresie definită prin (3.4.1), anume

$$\mathtt{m} = \frac{1}{\bar{r}} \frac{\bar{\sigma}_Y}{\bar{\sigma}_X}, \quad \mathtt{n} = \bar{y} - \mathtt{m}\bar{x}.$$

```
clear,clf, mu(1)=input('m1='); mu(2)=input('m2=');
v(1,1)=input('sigma1^2=');
v(2,2)=input('sigma2^2=');
v(1,2)=input('Cov(X,Y)='); v(2,1) =v(1,2);
if det(v) <= 0
   error('Matricea v nu e pozitiv definita!')
end
N=input('N='); Z=mvnrnd(mu,v,N);
X=Z(:,1); Y=Z(:,2);
```

```
ma=mean(Z); s=std(Z);r=corrcoef([X,Y]);
m=s(2)/(r(1,2)*s(1)); n=ma(2)-m*ma(1);
plot(X,Y,'.'), lsline, refline(m,n)
text(ma(1),ma(2),'o Centrul de greutate')
```

Pentru datele de intrare `N=20`, `mu=(0,0)`, $\mathtt{v} = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$, se obţin graficele din Figura 3.12.

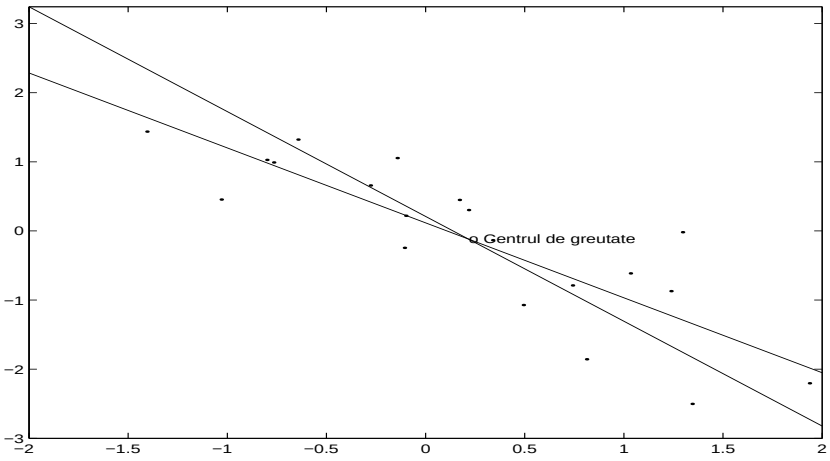

Figura 3.12: Drepte de regresie

3.4.3 Funcţiile `polyfit` şi `refcurve`

Funcţia `polyfit` este specifică sistemului de bază Matlab, iar funcţia `refcurve` este conţinută de *Statistics toolbox*, fiind folosite în ajustarea datelor prin funcţii polinomiale.

Apelurile acestor funcţii se pot face prin:

```
p=polyfit(x,y,n)
refcurve(p)
```

Funcţia `polyfit` determină coeficienţii polinomului de ajustare de grad `n`, cu metoda celor mai mici pătrate, corespunzător datelor statistice $(\mathtt{x}_i, \mathtt{y}_i)_{i=\overline{1,m}}$. Prima componentă a vectorului `p` este coeficientul monomului de grad `n`. Dacă `n=1`, se obţin coeficienţii dreptei de regresie a variabilei `Y` în raport cu variabila `X`.

Funcţia `refcurve` reprezintă grafic în figura curentă funcţia polinomială precizată prin coefcienţii `p`, cu precizarea că `p(1)` reprezintă coeficientul monomului de grad maxim. Dacă parametrul `p` are două componente, funcţia este echivalentă cu funcţia `refline`.

Calculul valorilor unei funcţii polinomiale se face cu ajutorul funcţiei `polyval`, pe care o prezentăm în continuare, împreună cu alte funcţii specifice sistemului Matlab de bază, din acest grup de funcţii.

3.4.4 Funcţiile `polyval`, `polyvalm`, `poly` şi `roots`

Forme de apel pentru funcţiile `polyval`, `polyvalm`, `poly` şi `roots` sunt:

```
y=polyval(p,x)
y=polyvalm(p,x)
p=poly(a)
p=roots(a)
```

Funcţia `polyval` calculează valorile `y` ale polinomului precizat prin coeficienţii conţinuţi de vectorul `p`, pe fiecare punct al matricei sau vectorului `x`.

Funcţia `polyvalm` calculează matricea pătratică `y`, care reprezintă valoarea matriceală a polinomului precizat prin coeficienţii conţinuţi de vectorul `p`, pentru matricea pătratică `x`, adică

$$\mathtt{y} = \mathtt{p}_1\mathtt{x}^n + \mathtt{p}_2\mathtt{x}^{n-1} + \cdots + \mathtt{p}_{n+1},$$

dacă `p` are $n+1$ componente.

Funcţia `poly` calculează coeficienţii polinomului caracteristic al matricei pătratice `a`:

$$P(\lambda) = \begin{vmatrix} \lambda - a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & \lambda - a_{22} & \ldots & a_{1m} \\ \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & \lambda - a_{mm} \end{vmatrix},$$

iar când `a` este un vector calculează coeficienţii polinomului care are ca rădăcini elementele vectorului, adică ai lui

$$P(x) = \prod_{i=1}^{n} (x - \mathtt{a}_i).$$

Funcţia `roots` calculează rădăcinile polinomului cu coeficienţii daţi prin vectorul `p`.

Programul 3.4.23. Programul următor va genera matricea `Z`, care conţine `N` vectori aleatori ce urmează legea normală bidimensională. Se transformă apoi componenta a doua a vectorilor aleatori, prin ridicarea acestora la puterea a treia, iar apoi se determină polinomul de ajustare de grad trei, prin folosirea datelor transformate. Polinomul de ajustare obţinut se reprezintă grafic împreună cu norul statistic al datelor transformate.

```
clf, clear, mu(1)=input('m1='); mu(2)=input('m2=');
v(1,1)=input('sigma1^2=');
v(2,2)=input('sigma2^2=');
v(1,2)=input('Cov(X,Y)='); v(2,1) =v(1,2);
if det(v) <= 0
   error('Matricea v nu e pozitiv definita!')
end
N=input('N='); Z=mvnrnd(mu,v,N);
X=Z(:,1); Y=Z(:,2); p=polyfit(X,Y.^3,3);
scatter(X,Y.^3,7,'filled'), hold on,
m=mu(1); s=sqrt(v(1,1));
x=m-3*s:0.01:m+3*s;
y=polyval(p,x); plot(x,y)
```

Pentru datele de intrare `N=25`, `mu=(0,0)`, $\texttt{v} = \begin{pmatrix} \texttt{1} & \texttt{-0.9} \\ \texttt{-0.9} & \texttt{1} \end{pmatrix}$, se obţine graficul din Figura 3.13.

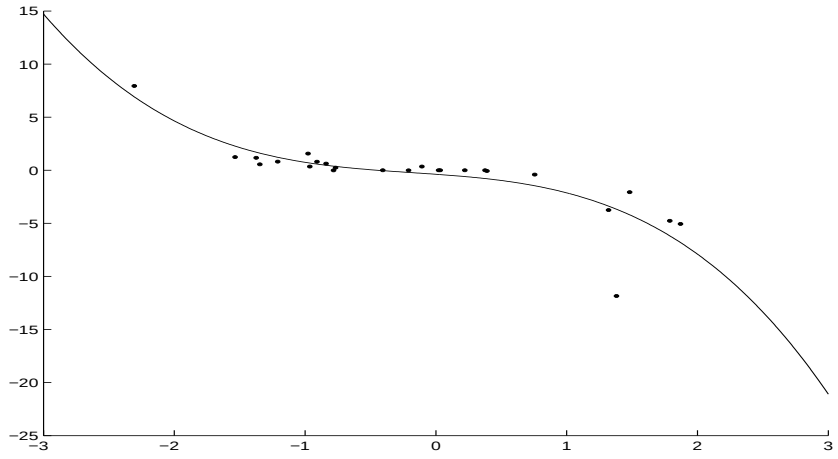

Figura 3.13: Polinom de ajustare de grad 3

3.4.5 Funcţia `polytool`

Funcţia demonstrativă `polytool` se lansează prin:

```
>>polytool(x,y)
>>polytool(x,y,n)
```

În urma lansării se produce un grafic interactiv (demonstrativ) privind polinomul de ajustare pentru datele conţinute în vectorii coloană `x` şi `y`, având gradul precizat

prin parametrul `n` (valoarea implicită este `n=1`). Schimbarea gradului polinomului de ajustare se face prin introducerea noii valori în fereastra din partea de sus.

Pe figură este reprezentat norul statistic pentru datele `x` şi `y`, împreună cu graficul polinomului de ajustare (cu linie continuă).

Prin introducerea valorii unui argument x în fereastra de jos sau prin deplasarea dreptei verticale pe poziţia abscisei x, pe axa ordonatelor se obţine valoarea corespunzătoare.

Pe lângă aceste elemente, graficul demonstrativ conţine şi alte elemente, care se referă la noţiuni ce urmează a fi prezentate în capitolele ce urmează, cum ar fi *probabilitate de încredere* şi *intervale de încredere*.

De asemenea, există posibilitatea ajustării datelor cu alte metode decât metoda celor mai mici pătrate. Alegerea unei astfel de metodă se face prin activarea butonului `method` din meniu.

3.4.6 Funcţia `nlinfit`

Pentru ajustarea neliniară, *Statistics toolbox* dispune de funcţia `nlinfit`, bazată pe metoda celor mai mici pătrate. Apelul funcţiei se poate face prin:

```
a=nlinfit(x,y,numef,a0)
[a,r]=nlinfit(x,y,f,a0)
```

unde `numef` reprezintă numele funcţiei, care defineşte legătura căutată dintre variabila dependentă, cu valorile conţinute de vectorul coloană `y`, variabilele independente, cu valorile conţinute în coloanele matricei `x`. Matricea `x` trebuie să aibă acelaşi număr de linii ca şi lungimea vectorului `y`.

Parametrul `a0` reprezintă valorile de pornire (iniţiale) ale parametrilor `a`, care urmează să fie calculaţi.

Funcţia `numef` poate fi o funcţie Matlab proprie, care are linia de definiţie

```
function z=numef(a,x)
```

În urma executării primei instrucţiuni, se obţin coeficienţii `a` ai funcţiei `numef`, iar dacă se foloseşte a doua instrucţiune se mai obţin şi erorile `r=y-numef(a,x)`.

Programul 3.4.24. Vom considera `N` vectori aleatori ce urmează legea normală bidimensională, păstraţi în matricea `Z` de tipul `(N,2)`. Asupra elementelor coloanei a doua se efectuează transformarea dată prin funcţia `exp`. Dacă a doua coloană a lui `Z` astfel transformată o considerăm că reprezintă valorile unei variabile dependente `Y`, iar prima coloană considerăm că reprezintă valorile unei variabile independente `X`, vrem să determinăm legătura de forma

$$\mathtt{Y} = \mathtt{b}e^{\mathtt{a}\mathtt{X}}.$$

Procedăm în două moduri, cu ajutorul funcţiei `nlinfit` (linie continuă pe grafic), respectiv prin liniarizarea acestei relaţii (linie înreruptă pe grafic), adică prin logaritmare: $\ln \mathtt{Y} = \ln \mathtt{b} + \mathtt{a}\mathtt{X}$.

```
clf, mu(1)=input('m1='); mu(2)=input('m2=');
v(1,1)=input('sigma1^2=');
v(2,2)=input('sigma2^2=');
v(1,2)=input('Cov(X,Y)='); v(2,1) =v(1,2);
if det(v) <= 0
   error('Matricea v nu e pozitiv definita!')
end
N=input('N='); Z=mvnrnd(mu,v,N);
X=Z(:,1); Y=exp(Z(:,2));
a=nlinfit(X,Y,'expo',[1,1]);
p=polyfit(X,log(Y),1);
a1=p(1); b1=exp(p(2));
scatter(X,Y,7,'filled'), hold on
m=mu(1); s=sqrt(v(1,1));
x=m-3*s:0.01:m+3*s;
y=expo(a,x); y1=b1*exp(a1*x);
plot(x,y,'-',x,y1,'--')
```

Considerând datele de intrare `N=20`, `mu=(0,0)`, $\mathtt{v} = \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}$, se obţin graficele din Figura 3.14.

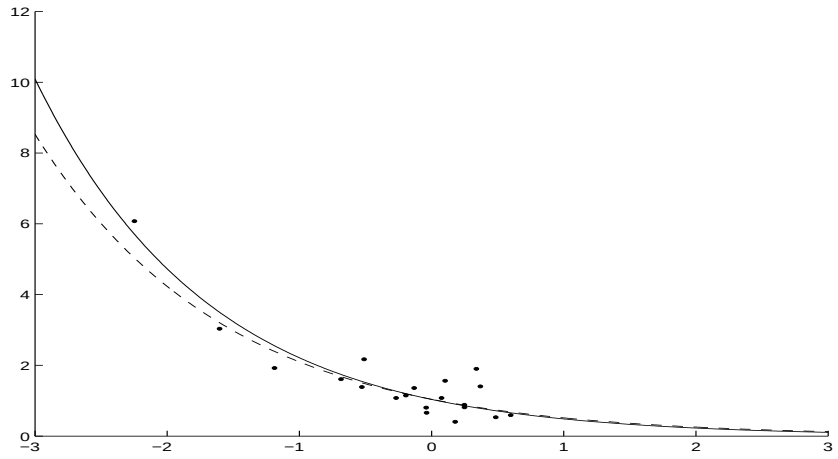

Figura 3.14: Ajustare neliniară

3.4.7 Funcţia `nlintool`

Funcţia demonstrativă `polytool` se lansează prin:

```
>>nlintool(x,y,numef,a0)
```

În urma lansării se produce un grafic interactiv (demonstrativ) privind ajustarea datelor conţinute în vectorul coloană `y` şi matricea `x`, având tipul legăturii precizat prin parametrul `numef`, care poate fi o funcţie Matlab proprie.

Pe figură este reprezentat graficul de ajustare (cu linie continuă).

Prin introducerea valorii unui argument x în fereastra de jos sau prin deplasarea dreptei verticale pe poziţia abscisei x, pe axa ordonatelor se obţine valoarea corespunzătoare pentru y.

Elementele obţinute, prin lansarea acestei funcţii, pot fi salvate în spaţiul de lucru, `workspace`, parţial sau în totalitate, prin utilizarea butonului `export`.

Pe lângă aceste elemente, graficul demonstrativ conţine şi alte elemente, care se referă la noţiuni ce urmează a fi prezentate în capitolele ce urmează, cum ar fi *intervale de încredere*.

3.4.8 Coeficienţii Spearman şi Kendall

Definiţia 3.4.25. *Fie* (u_k, v_k), $k = \overline{1,N}$, *rangurile datelor statistice primare* (x'_k, y'_k), $k = \overline{1,N}$, *obţinute prin ordonarea crescătoare după prima, respectiv a doua componentă. Numim* coeficient de corelaţie al rangurilor *sau* coeficientul lui Spearman, *valoarea numerică*

$$\overline{s} = \overline{r}\,(U, V)\,,$$

unde U *şi* V *sunt noile caracteristici care definesc rangurile datelor statistice, respectiv pentru* X *şi* Y.

Proprietatea 3.4.26. *Dacă notăm* $d_k = |u_k - v_k|$ *diferenţa rangurilor aceluiaşi individ, atunci*

$$\overline{s} = 1 - \frac{6}{N\left(N^2-1\right)} \sum_{k=1}^{N} d_k^2.$$

Demonstraţie. Conform definiţiei lui $\overline{s}$ avem că

$$\overline{s} = \frac{\sum_{i=1}^{N} (u_i - \overline{u})\,(v_i - \overline{v})}{\sqrt{\sum_{i=1}^{N} (u_i - \overline{u})^2}\sqrt{\sum_{i=1}^{N} (v_i - \overline{v})^2}}\cdot$$

Deoarece $\{\, u_1, u_2, \ldots, u_N \,\} = \{\, v_1, v_2, \ldots, v_N \,\} = \{\, 1, 2, \ldots, N \,\}$, rezultă că

$$\overline{u} = \overline{v} = \frac{1}{N} \sum_{k=1}^{N} k = \frac{N+1}{2}$$

şi

$$\sum_{k=1}^{N} (u_k - \overline{u})^2 = \sum_{k=1}^{N} (v_k - \overline{v})^2 = \sum_{k=1}^{N} \left(k - \frac{N+1}{2}\right)^2$$

$$= \frac{N\,(N+1)\,(2N+1)}{6} - \frac{N\,(N+1)^2}{2} + \frac{N\,(N+1)^2}{4} = \frac{N\left(N^2-1\right)}{12},$$

deci numitorul din expresia lui $\overline{s}$ este calculat.

Pentru calcularea numărătorului pornim de la identitatea

$$\begin{aligned} 2\,(u_k-\overline{u})\,(v_k-\overline{v}) &= (u_k-\overline{u})^2 + (v_k-\overline{v})^2 - [(u_k-\overline{u}) - (v_k-\overline{v})]^2 \\ &= (u_k-\overline{u})^2 + (v_k-\overline{v})^2 - d_k^2, \end{aligned}$$

care se însumează pentru $k = \overline{1,N}$. Obţinem în acest fel

$$\begin{aligned} 2\sum_{k=1}^{N}(u_k-\overline{u})\,(v_k-\overline{v}) &= \sum_{k=1}^{N}(u_k-\overline{u})^2 + \sum_{k=1}^{N}(v_k-\overline{v})^2 - \sum_{k=1}^{N} d_k^2 \\ &= \frac{N\left(N^2-1\right)}{12}\cdot 2 - \sum_{k=1}^{N} d_k^2, \end{aligned}$$

adică

$$\sum_{k=1}^{N}(u_k-\overline{u})\,(v_k-\overline{v}) = \frac{N\left(N^2-1\right)}{12} - \frac{1}{2}\sum_{k=1}^{N} d_k^2.$$

Dacă se înlocuiesc aceste expresii în formula lui $\overline{s}$ se obţine relaţia din enunţul proprietăţii. □

Corolarul 3.4.27. *Coeficientul lui Spearman verifică relaţiile* $-1 \leqslant \overline{s} \leqslant 1$.

Demonstraţie. Valoarea maximă $\overline{s} = 1$ se obţine când $d_k = 0,\ k = \overline{1,N}$, adică toate rangurile corespunzătoare celor două caracteristici coincid. Valoarea minimă $\overline{s} = -1$ se va obţine când diferenţele rangurilor sunt maxime, adică rangurile sunt inverse pentru cele două caracteristici. În acest caz avem

$$\sum_{k=1}^{N} d_k^2 = 2\left[(N-1)^2 + (N-3)^2 + \cdots\right] = 2\cdot\frac{N\left(N^2-1\right)}{6},$$

de unde se obţine, într-adevăr, $\overline{s} = -1$.

Acelaşi rezultat se obţine şi din faptul că $\overline{s}$ este definit cu ajutorul coeficientului (pentru ranguri) Pearson, despre care ştie ca ia valori în intervalul $[-1,1]$. □

Observaţia 3.4.28. Pe baza corolarului precedent, vom spune că pentru $\overline{s} = 1$ cele două clasamente pentru caracteristicile X şi Y sunt identice, pentru $\overline{s} = -1$ sunt inverse unul celuilalt, iar pentru $\overline{s} = 0$ sunt independente.

Observaţia 3.4.29. Când există două sau mai multe date statistice primare care au aceeaşi valoare, atunci rangurile acestora se consideră toate egale (*ex-aequo*) cu media aritmetică a rangurilor pe care le ocupă aceste date în ordonarea crescătoare. corespunzător vom face modificarea lui $\overline{s}$.

Dacă în ordonarea după caracteristica X avem u grupe de date statistice care coincid, fiecare grupă conţinând $g_i,\ i=\overline{1,u}$, date statistice şi dacă în ordonarea după caracteristica Y avem v grupe de date statistice ce coincid, fiecare grupă conţinând $h_j,\ j=\overline{1,v}$, date statistice, atunci $\overline{s}$ se modifică după cum urmează

$$\overline{s}=\frac{\frac{1}{6}\left[N\left(N^2-1\right)-(t_x+t_y)\right]-\sum_{k=1}^{N}d_k^2}{\sqrt{\frac{1}{6}N\left(N^2-1\right)-2t_x}\sqrt{\frac{1}{6}N\left(N^2-1\right)-2t_y}},$$

unde

$$t_x=\frac{1}{12}\sum_{i=1}^{u}g_i\left(g_i^2-1\right)\quad \text{şi}\quad t_y=\frac{1}{12}\sum_{j=1}^{v}h_j\left(h_j^2-1\right).$$

Definiţia 3.4.30. *Numim* coeficientul lui Kendall *relativ la distribuţia statistică a caracteristicii bidimensionale* (X,Y)*, raportul*

$$\overline{k}=\frac{2\overline{t}}{N\left(N-1\right)},\quad \text{unde}\quad \overline{t}=\sum_{\substack{i,j=1\\ i<j}}^{N}\text{sign}\left\{\left(x_j'-x_i'\right)\left(y_j'-y_i'\right)\right\}.$$

Proprietatea 3.4.31. *Coeficientul lui Kendall satisface relaţiile* $-1\leqslant\overline{k}\leqslant 1$.

Demonstraţie. Valoarea maximă a lui $\overline{k}$ se obţine când $\overline{t}$ este maxim, adică atunci când $\left(x_j'-x_i'\right)\left(y_j'-y_i'\right)>0$, pentru fiecare $i<j$. În acest caz avem că

$$\overline{t}=\sum_{\substack{i,j=1\\ i<j}}^{N}1=\binom{N}{2}=\frac{N\left(N-1\right)}{2},$$

deci $\overline{k}=1$.

Valoarea minimă a lui $\overline{k}$ se obţine când $\overline{t}$ este minim, $\left(x_j'-x_i'\right)\left(y_j'-y_i'\right)<0$, pentru fiecare $i<j$, caz în care

$$\overline{t}=-\sum_{\substack{i,j=1\\ i<j}}^{N}1=-\binom{N}{2}=-\frac{N\left(N-1\right)}{2},$$

deci $\overline{k}=-1$. Prin urmare avem $-1\leqslant\overline{k}\leqslant 1$. □

Observaţia 3.4.32. Din proprietatea precedentă, avem că pentru $\overline{k} = 1$ cele două clasamente ale celor două caracteristici X şi Y sunt identice, pentru $\overline{k} = -1$ sunt inverse unul celuilalt, iar pentru $\overline{k} = 0$ sunt independente.

Observaţia 3.4.33. Pentru calculul rapid al lui $\overline{k}$, se poate proceda după cum urmează. Se ordonează datele primare (x'_k, y'_k), $k = \overline{1, N}$, în mod crescător după prima componentă. Fie aceasta $\left(x'_{(k)}, y'_{(k)}\right)$, $k = \overline{1, N}$, cu $x'_{(1)} \leqslant x'_{(2)} \leqslant \cdots \leqslant x'_{(N)}$. Se calculează apoi numărul

$$\overline{t} = \sum_{\substack{u,v=1 \\ u<v}} \text{sign}\left(y'_{(v)} - y'_{(u)}\right),$$

obţinându-se astfel $\overline{k}$.

Remarcăm de asemenea că toate formulele relative la coeficientul lui Kendall rămân adevărate dacă se lucrează cu rangurile (u_i, v_i), corespunzătoare datelor statistice primare (x'_i, y'_i), $i = \overline{1, N}$, aşa cum s-au introdus pentru definirea coeficientului lui Spearman.

Observaţia 3.4.34. Când în cele două clasamente sunt valori egale (*ex-aequo*) se procedează ca la Observaţia 3.4.29, adică se înlocuiesc toate rangurile pentru valorile egale prin media aritmetică a rangurilor pe care le ocupă în ordonare. Corespunzător se face modificarea lui $\overline{k}$ prin formula

$$\overline{k} = \frac{\overline{t}}{\sqrt{\frac{N(N-1)}{2} - u_x}\sqrt{\frac{N(N-1)}{2} - u_y}},$$

unde $u_x = \frac{1}{2}\sum_{i=1}^{u} g_i\,(g_i - 1)$ şi $u_y = \frac{1}{2}\sum_{j=1}^{v} h_j\,(h_j - 1)$, cu $g_i,\ h_j$ definiţi la Observaţia 3.4.29. Pentru calculul lui $\overline{t}$ pentru fiecare pereche ex-aequo valoarea lui $\overline{t}$ rămâne neschimbată.

Exemplul 3.4.35. Echipele diviziei naţionale de fotbal sunt clasificate după numărul de goluri primite (X) şi după numărul cartonaşelor galbene primite (Y) în primele 10 meciuri ale campionatului. Datele statistice sunt trecute în tabelul următor

Echipa	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}	E_{13}	E_{14}	E_{15}	E_{16}	E_{17}	E_{18}
X	15	13	10	13	12	11	16	11	13	9	10	14	15	9	8	9	10	11
Y	13	9	11	11	12	8	7	6	9	8	8	10	11	9	5	10	13	14

Să calculăm coeficienţii lui Spearman respectiv al lui Kendall. Pentru aceasta ordonăm crescător echipele după caracteristica X, după care folosind regula ex-aequo stabilim rangurile datelor statistice.

Calculele sunt efectuate în tabelul următor.

Echipa	X	Y	Rang(X)	Rang(Y)	d	d^2	-	+
E_{15}	8	5	1	1	0	0	0	17
E_{10}	9	8	3	5	2	4	2	12
E_{14}	9	9	3	8	5	25	4	9
E_{16}	9	10	3	10.5	7.5	56.25	6	7
E_3	10	11	6	13	7	49	7	4
E_{11}	10	8	6	5	1	1	2	9
E_{17}	10	13	6	16.5	10.5	110.25	9	1
E_6	11	8	9	5	4	16	2	8
E_8	11	6	9	2	7	49	0	9
E_{18}	11	14	9	18	9	81	8	0
E_5	12	12	11	15	4	16	6	1
E_2	13	9	13	8	5	25	1	4
E_4	13	11	13	13	0	0	3	1
E_9	13	9	13	8	5	25	1	3
E_{12}	14	10	15	10.5	4.5	20.25	1	2
E_1	15	13	16.5	16.5	0	0	2	0
E_{13}	15	11	16.5	13	3.5	12.25	1	0
E_7	16	7	18	3	15	225	0	0
Total						715	55	87

Numărul grupelor ex-aequo este cinci atât pentru X cât şi pentru Y. Avem astfel

$$\begin{aligned}
t_x =& \frac{1}{12}\left[3\left(9-1\right)+3\left(9-1\right)+3\left(9-1\right)+3\left(9-1\right)+2\left(4-1\right)\right]=\frac{17}{2},\\
t_y =& \frac{1}{12}\left[3\left(9-1\right)+3\left(9-1\right)+2\left(4-1\right)+3\left(9-1\right)+2\left(4-1\right)\right]=7,\\
u_x =& \frac{1}{2}\left[3\left(3-1\right)+3\left(3-1\right)+3\left(3-1\right)+3\left(3-1\right)+2\left(2-1\right)\right]=13,\\
u_y =& \frac{1}{2}\left[3\left(3-1\right)+3\left(3-1\right)+2\left(2-1\right)+3\left(3-1\right)+2\left(2-1\right)\right]=11,\\
\overline{t} =& 87-55=32.
\end{aligned}$$

Dacă se înlocuiesc aceste valori în formule se obţin coeficienţii

$$\begin{aligned}
\overline{s} =& \frac{\frac{1}{6}\left[18\cdot 19\cdot 17-\left(\frac{17}{2}+7\right)\right]-715}{\sqrt{\frac{1}{6}\cdot 18\cdot 17\cdot 19-2\cdot\frac{17}{2}}\sqrt{\frac{1}{6}\cdot 18\cdot 17\cdot 19-2\cdot 7}}=0.263,\\
\overline{k} =& \frac{3}{2}\sqrt{\frac{18\cdot 17}{2}-13}\sqrt{\frac{18\cdot 17}{2}-11}=0.227.
\end{aligned}$$

Programul 3.4.36. Programul Matlab, care urmează, calculează cei doi coeficienţi, Spearman şi Kendall, fără tratarea cazurilor ex-aequo.

```
x=[15,13,10,13,12,11,16,11,13,9,10,14,15,9,8,9,10,11];
y=[13,9,11,11,12,8,7,6,9,8,8,10,11,9,5,10,13,14];
[xord,ind]=sort(x); [yord,jnd]=sort(y);
for i=1:18
    rx(ind(i))=i; ry(jnd(i))=i;
end
d=abs(rx-ry); d2=d.^2;
sp=1-6*sum(d2)/18/(18^2-1);
ke=0;
for i=1:18
    for j=1:18
        if i<j
        ke=ke+sign((rx(i)-rx(j))*(ry(i)-ry(j)));
    end
    end
end
ke=2*ke/(18*17);
fprintf('  x  |   y  |rang(x)|rang(y)|   d  | d^2 \n')
for i=1:18
    fprintf('%3d  |%4d  |%5d  |%5d  |%4d  |%4d\n',...
        x(i),y(i),rx(i),ry(i),d(i),d2(i))
end
fprintf('\n sp=%5.3f,   ke=%5.3f',sp,ke)
```

În urma executării acestui program, se obţin următoarele rezultate:

```
  x  |   y  |rang(x)|rang(y)|   d  | d^2
 15  |  13  |   16  |   16  |   0  |   0
 13  |   9  |   12  |    7  |   5  |  25
 10  |  11  |    5  |   12  |   7  |  49
 13  |  11  |   13  |   13  |   0  |   0
 12  |  12  |   11  |   15  |   4  |  16
 11  |   8  |    8  |    4  |   4  |  16
 16  |   7  |   18  |    3  |  15  | 225
 11  |   6  |    9  |    2  |   7  |  49
 13  |   9  |   14  |    8  |   6  |  36
  9  |   8  |    2  |    5  |   3  |   9
 10  |   8  |    6  |    6  |   0  |   0
 14  |  10  |   15  |   10  |   5  |  25
 15  |  11  |   17  |   14  |   3  |   9
  9  |   9  |    3  |    9  |   6  |  36
  8  |   5  |    1  |    1  |   0  |   0
  9  |  10  |    4  |   11  |   7  |  49
 10  |  13  |    7  |   17  |  10  | 100
 11  |  14  |   10  |   18  |   8  |  64

 sp=0.269,   ke=0.203
```

Observaţia 3.4.37. *(Formula lui Daniels)*. Coeficientul $\overline{r}$ de corelaţie (al lui Pearson), coeficientul $\overline{s}$ al lui Spearman şi coeficientul $\overline{k}$ al lui Kendall se pot exprima prin formula unică

$$\overline{d} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}b_{ij}}{\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij}^2}\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N} b_{ij}^2}}.$$

Dacă se ia

$$a_{ij} = x_i' - x_j', \qquad b_{ij} = y_i' - y_j'$$

se obţine coeficientul de corelaţie $\overline{r}$.

Într–adevăr, prin calcule algebrice simple se poate scrie

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\left(x_i' - x_j'\right)\left(y_i' - y_j'\right) = 2N\sum_{i=1}^{N} x_i'y_i' - 2\left(\sum_{i=1}^{N} x_i'\right)\left(\sum_{i=1}^{N} y_i'\right),$$

iar

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\left(x_i' - x_j'\right)^2 = 2N\sum_{i=1}^{N} x_i'^2 - 2\left(\sum_{i=1}^{N} x_i'\right)^2,$$
$$\sum_{i=1}^{N}\sum_{j=1}^{N}\left(y_i' - y_j'\right)^2 = 2N\sum_{i=1}^{N} y_i'^2 - 2\left(\sum_{i=1}^{N} y_i'\right)^2,$$

astfel că

$$\overline{d} = \frac{\sum_{i=1}^{N} x_i'y_i' - \frac{1}{N}\left(\sum_{i=1}^{N} x_i'\right)\left(\sum_{i=1}^{N} y_i'\right)}{\sqrt{\sum_{i=1}^{N} x_i'^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i'\right)^2}\sqrt{\sum_{i=1}^{N} y_i'^2 - \frac{1}{N}\left(\sum_{i=1}^{N} y_i'\right)^2}} = \overline{r}.$$

Dacă se ia

$$a_{ij} = u_i - u_j, \qquad b_{ij} = v_i - v_j$$

se obţine, ca mai înainte, coeficientul $\overline{s}$ al lui Spearman, iar dacă se consideră

$$a_{ij} = \operatorname{sign}\left(x_i' - x_j'\right), \qquad b_{ij} = \operatorname{sign}\left(y_i' - y_j'\right)$$

se obţine coeficientul $\overline{k}$ al lui Kendall.

Definiţia 3.4.38. *Fie colectivitatea $\mathcal{C}$ cercetată din punct de vedere a p caracteristici $X_1, X_2, \ldots, X_p$ şi fie datele statistice primare relative la aceste caracteristici scrise în matricea*

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ x_{N1} & x_{N2} & \ldots & x_{Np} \end{pmatrix},$$

cu matricea $\mathbf{R}$ *a rangurilor corespunzătoare pentru fiecare caracteristică dată prin*

$$\mathbf{R} = \begin{pmatrix} r_{11} & r_{12} & \ldots & r_{1p} \\ r_{21} & r_{22} & \ldots & r_{2p} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ r_{N1} & r_{N2} & \ldots & r_{Np} \end{pmatrix}.$$

Numim coeficientul de concordanţă al lui Kendall *raportul*

$$\overline{w} = \frac{12\overline{s}}{p^2\left(N^3 - N\right)},$$

unde

$$\overline{s} = \sum_{i=1}^{N}\left(r_{i\cdot} - \frac{r_{\cdot\cdot}}{N}\right)^2, \qquad r_{i\cdot} = \sum_{j=1}^{p} r_{ij}, \qquad r_{\cdot\cdot} = \sum_{i=1}^{N} r_{i\cdot}.$$

Observaţia 3.4.39. Fiecare coloană a tabloului rangurilor este o permutare a numerelor $1, 2, \ldots, N$, deci suma elementelor din fiecare coloană este $\frac{N(N+1)}{2}$ şi prin urmare

$$r_{\cdot\cdot} = p\frac{N\left(N+1\right)}{2}.$$

Când cele p clasamente sunt identice, deci coloanele din matricea rangurilor sunt identice, totalurile $r_{1\cdot}, r_{2\cdot}, \ldots, r_{N\cdot}$ formează o permutare a elementelor $p, 2p, \ldots, Np$. În acest caz valoarea lui $\overline{s}$ va fi maximă, anume

$$\overline{s} = \sum_{i=1}^{N}\left(p\,i - \frac{p\left(N+1\right)}{2}\right)^2 = \frac{p^2\left(N^3 - N\right)}{12}.$$

prin urmare $\overline{w} \leqslant 1$.

Când $r_{1\cdot} = r_{2\cdot} = \cdots = r_{N\cdot}$ atunci se obţine valoarea minimă $\overline{w} = 0$, caz în care avem independenţă între caracteristicile cercetate.

Observaţia 3.4.40. Cazurile ex-aequo se tratează ca şi pentru două caracteristici, iar formula modificată este

$$\overline{w} = \frac{12\overline{s}}{p^2\left(N^3 - N\right) - p\sum_{j=1}^{p}\left(t_j^3 - t_j\right)}$$

unde t_j este numărul datelor ex-aequo pentru caracteristica X_j.

Exemplul 3.4.41. Se consideră 12 familii pentru care sunt cercetate următoarele caracteristici X_1, X_2, X_3, X_4 reprezentând cheltuielile (în sute de lei) zilnice pentru pâine, legume, fructe, carne respectiv. S-au obţinut următoarele date statistice

X_1	X_2	X_3	X_4
332	428	354	1437
293	559	388	1527
372	767	562	1948
406	563	341	1509
386	608	396	1501
438	843	689	2345
534	660	367	1620
460	699	483	1856
385	789	621	2366
655	776	423	1848
584	995	548	2056
515	1097	887	2630

Vrem să calculăm coeficientul de concordanţă al lui Kendall.

Pentru a calcula coeficientul de concordanţă al lui Kendall, se determină rangurile datelor statistice pentru fiecare caracteristică în parte. Matricea rangurilor este

$$\boldsymbol{R} = \left(r_{ij}\right)_{\substack{i=\overline{1,12}\\ j=\overline{1,4}}} = \begin{pmatrix} 2 & 1 & 2 & 1 \\ 1 & 2 & 4 & 4 \\ 3 & 7 & 9 & 8 \\ 6 & 3 & 1 & 3 \\ 5 & 4 & 5 & 2 \\ 7 & 10 & 11 & 10 \\ 10 & 5 & 3 & 5 \\ 8 & 6 & 7 & 7 \\ 4 & 9 & 10 & 11 \\ 12 & 8 & 6 & 6 \\ 11 & 11 & 8 & 9 \\ 9 & 12 & 12 & 12 \end{pmatrix}.$$

De asemenea, avem că sumele rangurilor de pe fiecare line a matricei rangurilor sunt $6,\ 11,\ 27,\ 13,\ 16,\ 38,\ 23,\ 28,\ 34,\ 32,\ 39,\ 45$, deci

$$\overline{s} = \sum_{i=1}^{N}\left(r_{i\cdot} - \frac{r_{\cdot\cdot}}{N}\right)^2 = \left(6 - \frac{312}{12}\right)^2 + \left(11 - \frac{312}{12}\right)^2 + \cdots + \left(45 - \frac{312}{12}\right)^2$$

$$=(6-26)^2+(11-26)^2+\cdots+(45-26)^2=1682.$$

Coeficientul de concordanţă al lui Kendall va fi

$$\overline{w}=\frac{12\,\overline{s}}{p^2(N^3-N)}=\frac{12\cdot 1682}{4^2(12^3-12)}=\frac{841}{1441}\cong 0.7351.$$

Programul 3.4.42. Un program Matlab, care efectuează aceste calcule, este prezentat în continuare

```
x=[332,428,354,1437;293,559,388,1527;372,767,562,1948;
   406,563,341,1509;386,608,396,1501;438,843,689,2345;
   534,660,367,1620;460,699,483,1856;385,789,621,2366;
   655,776,423,1848;584,995,548,2056;515,1097,887,2630];
N=12; p=4;
[x1,i1]=sort(x(:,1)); [x2,i2]=sort(x(:,2));
[x3,i3]=sort(x(:,3)); [x4,i4]=sort(x(:,4));
for i=1:12
    r1(i1(i))=i; r2(i2(i))=i; r3(i3(i))=i; r4(i4(i))=i;
end
r=[r1',r2',r3',r4']; sb=sum((sum(r')-sum(sum(r))/12).^2);
w=12*sb/(p^2*(N^3-N));
fprintf('   Matricea rangurilor\n')
for i=1:12
    fprintf('%6d |%3d |%3d |%3d\n',r1(i),r2(i),r3(i),r4(i))
end
fprintf('\n       w=%5.3f',w)
```

iar rezultatele obţinute, în urma executării programului, sunt

```
   Matricea rangurilor
     2 |  1 |  2 |  1
     1 |  2 |  4 |  4
     3 |  7 |  9 |  8
     6 |  3 |  1 |  3
     5 |  4 |  5 |  2
     7 | 10 | 11 | 10
    10 |  5 |  3 |  5
     8 |  6 |  7 |  7
     4 |  9 | 10 | 11
    12 |  8 |  6 |  6
    11 | 11 |  8 |  9
     9 | 12 | 12 | 12

       w=0.735
```

Capitolul 4

Teoria selecţiei

Studiul statistic al unei colectivităţi $\mathcal{C}$ din punct de vedere al uneia sau mai multor caracteristici prin considerarea tuturor indivizilor colectivităţii de multe ori este imposibil de efectuat sau foarte costisitor. Dacă ne gândim, de exemplu, la studiul calităţii produselor unei fabrici de conserve, ne dăm seama că nu are sens considerarea tuturor acestor produse pentru efectuarea controlului, ceea ce ar duce la distrugerea întregii producţii. De asemenea, dacă se are în vedere recensământul populaţiei unei ţări, ne dăm seama de costul ridicat al acestei operaţii, motiv pentru care astfel de recensăminte sunt destul de rare.

Şi numai din cele două exemple amintite înainte, ne dăm seama că ar fi potrivit ca să fie considerată pentru studiu numai o parte a colectivităţii cercetate, iar apoi rezultatele obţinute relative la această parte să fie extinse la întreaga colectivitate.

4.1 Tipuri de selecţie

Definiţia 4.1.1. *Numim* eşantion (selecţie, sondaj) *relativ la colectivitatea $\mathcal{C}$ o submulţime de indivizi $\mathcal{E}$ a lui $\mathcal{C}$, care urmează să fie cercetaţi din punct de vedere a uneia sau mai multor caracteristici, iar numărul indivizilor din eşantionul $\mathcal{E}$ se numeşte* volumul *eşantionului.*

Observaţia 4.1.2. Modurile de obţinere a eşantionului $\mathcal{E}$ ne conduc la *metode nealeatoare* şi respectiv *metode aleatoare* de selecţie.

Dintre metodele nealeatoare amintim:

- *selecţia sistematică,* când indivizii care intră în eşantion sunt consideraţi după o anumită regulă, de exemplu din 10 în 10;

- *selecţie tipică*, când, cunoscându-se informaţii anterioare referitoare la colectivitate, sunt consideraţi indivizi cu valori medii apropiate de valoarea medie a întregii colectivităţi;

- *selecţie stratificată*, când colectivitatea este clasificată (stratificată) după anumite criterii, cunoscându–se proporţia indivizilor pentru fiecare strat. Eşantionul se ia astfel încât să fie respectate aceste proporţii pentru fiecare strat.

Pentru metodele aleatoare, fiecare individ al colectivităţii $\mathcal{C}$ poate să intre în eşantion cu aceaşi probabilitate (*selecţie cu probabilităţi egale*) sau cu probabilităţi diferite. Metodele aleatoare de selecţie sunt:

- *repetate* (*bernoulliene*), când individul, ce intră în eşantion, după ce a fost cercetat, este reintrodus în colectivitate;

- *nerepetate*, când individul ce intră în eşantion, după ce a fost cercetat, nu este reintrodus în colectivitate.

Observaţia 4.1.3. Dacă volumul colectivităţii este mult mai mare decât volumul eşantionului, atunci o selecţie nerepetată poate fi considerată ca fiind de tip repetat. Aceasta are la bază rezultatul privind comportarea asimptotică a legii hipergeometrice ca şi o lege binomială (a se vedea Observaţia 2.4.4).

4.2 Funcţii de selecţie

În cele ce urmează vom considera că avem de fiecare dată o selecţie repetată.

Definiţia 4.2.1. *Fie colectivitatea $\mathcal{C}$ cercetată din punct de vedere al caracteristicii X. Numim* date de selecţie *relative la caracteristica X datele statistice $x_1, x_2, \ldots, x_n$ privind indivizii care intră în eşantion.*

Observaţia 4.2.2. Datele de selecţie pot fi considerate ca fiind valorile unor variabile aleatoare $X_1, X_2, \ldots, X_n$, numite *variabile de selecţie* şi care în cazul unei selecţii repetate sunt variabile aleatoare independente, identic repartizate cu X.

Definiţia 4.2.3. *Numim* funcţie de selecţie *sau* statistică *variabila aleatoare*

$$Z_n = h_n\left(X_1, X_2, \ldots, X_n\right),$$

unde $h_n \colon \mathbb{R}^n \longrightarrow \mathbb{R}$ este o funcţie măsurabilă, iar $z_n = h_n\left(x_1, x_2, \ldots, x_n\right)$ se numeşte valoarea funcţiei de selecţie.

4.2.1 Media de selecţie

Definiţia 4.2.4. *Numim* medie de selecţie *funcţia de selecţie*

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k, \quad \textit{iar} \quad \bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

se numeşte valoarea mediei de selecţie.

Proprietatea 4.2.5. *Fie caracteristica X având valoarea medie $m = E(X)$ şi dispersia $\sigma^2 = Var(X)$, atunci*

$$E\left(\bar{X}\right) = m \qquad \textit{şi} \qquad Var\left(\bar{X}\right) = \frac{1}{n}\sigma^2.$$

Demonstraţie. Folosind proprietăţile valorii medii şi ale dispersiei şi având în vedere că selecţia este repetată avem succesiv

$$E\left(\bar{X}\right) = \frac{1}{n}\sum_{k=1}^{n} E(X_k) = \frac{1}{n}\sum_{k=1}^{n} E(X) = \frac{1}{n} n\, m = m,$$
$$Var\left(\bar{X}\right) = \frac{1}{n^2}\sum_{k=1}^{n} Var(X_k) = \frac{1}{n^2}\sum_{k=1}^{n} Var(X) = \frac{1}{n^2} n\, \sigma^2 = \frac{1}{n}\sigma^2.$$

□

Observaţia 4.2.6. În cazul în care caracteristica X urmează legea normală $\mathcal{N}(\mu, \sigma)$, atunci $\bar{X}$, fiind o combinaţie liniară de variabile aleatoare independente, ce urmează fiecare aceaşi lege normală, va urma de asemenea legea normală (a se vedea Observaţia 2.5.6). Pe baza proprietăţii precedente $\bar{X}$ va urma aşadar legea normală $\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Proprietatea 4.2.7. *Fie caracteristica X având valoarea medie $m = E(X)$ şi dispersia $\sigma^2 = Var(X)$, atunci statistica*

$$Z_n = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \tag{4.2.1}$$

converge în repartiţie la legea normală $\mathcal{N}(0,1)$, când $n \to \infty$, iar când X urmează legea normală $\mathcal{N}(m, \sigma)$ afirmaţia are loc pentru orice valoare a lui n.

Proprietatea nu reprezintă altceva decât Teorema 2.7.8.

Funcţia 4.2.8. Pentru ilustrarea proprietăţii precedente, scriem o funcţie Matlab, care generează o matrice de tipul `(n,m)`, ale cărei elemente sunt numere aleatoare ce urmează o lege de probabilitate precizată. Pentru fiecare coloană a matricei generate, se calculează valoarea z_n, folosind formula (4.2.1), după care se reprezintă grafic histograma valorilor z_n, împreună cu densitatea de probabilitate a legii normale $\mathcal{N}(\bar{z}, \bar{\sigma})$.

```
function hist2(n,m,lege)
% Comportare asimptotica a mediei de selectie
% lege - legea de probabilitate
% (n,m) - tipul matricei generate
switch lege
case 'unid'
  N=input('N=');
  X=unidrnd(N,n,m); [med,var]=unidstat(N);
case 'bino'
  nn=input('n='); p=input('p=');
  X=binornd(nn,p,n,m); [med,var]=binostat(nn,p);
case 'hyge'
  M=input('M='); K=input('K='); nn=input('n=');
  X=hygernd(M,K,nn,n,m); [med,var]=hygestat(M,K,nn);
case 'poiss'
  la=input('lambda=');
  X=poissrnd(la,n,m); [med,var]=poisstat(la);
case 'nbin'
  r=input('r='); p=input('p=');
  X=nbinrnd(r,p,n,m); [med,var]=nbinstat(r,p);
case 'geo'
  p=input('p=');
  X=geornd(p,n,m); [med,var]=geostat(p);
case 'unif'
  a=input('a:'); b=input('b(a<b):');
  X=unifrnd(a,b,n,m); [med,var]=unifstat(a,b);
case 'norm'
  mu=input('mu='); s=input('sigma=');
  X=normrnd(mu,s,n,m); [med,var]=normstat(mu,s);
case 'logn'
  mu=input('mu='); s=input('sigma=');
  X=lognrnd(mu,s,n,m); [med,var]=lognstat(mu,s);
case 'gam'
  a=input('a='); b=input('b=');
  X=gamrnd(a,b,n,m); [med,var]=gamstat(a,b);
case 'exp'
  mu=input('mu=');
  X=exprnd(mu,n,m); [med,var]=expstat(mu);
case 'beta'
  a=input('a='); b=input('b=');
  X=betarnd(a,b,n,m); [med,var]=betastat(a,b);
case 'weib'
  a=input('a='); b=input('b=');
```

```
    X=weibrnd(a,b,n,m); [med,var]=weibstat(a,b);
  case 'rayl'
    b=input('b=');
    X=raylrnd(b,n,m); [med,var]=raylstat(b);
  case 't'
    nn=input('n=');
    X=trnd(nn,n,m); [med,var]=tstat(nn);
  case 'chi2'
    nn=input('n=');
    X=chi2rnd(nn,n,m); [med,var]=chi2stat(nn);
  case 'f'
    mm=input('m='); nn=input('n=');
    X=frnd(mm,nn,n,m); [med,var]=fstat(mm,nn);
  otherwise
    error('Lege necunoscuta')
  end
     Z=mean(X); Z=(Z-med)/sqrt(var/n);
     hist(Z,fix(1+10/3*log10(m)), hold on
     x=-3:0.01:3; f=normpdf(x,0,1);
     plot(x,f,'k-'), colormap spring
```

Apelul funcţiei `hist2` se face prin

```
>>hist2(n,m,'lege')
```

unde valorile pentru `n`, `m` şi `lege`, fie că sunt precizate în acest apel, fie sunt precizate înainte. De exemplu, comanda

```
>>hist2(25,35,'unif')
```

are ca efect apelul funcţiei pentru legea uniformă, iar pe ecran se va cere introducerea parametrilor `a` şi `b` (`a`$<$`b`), după care pe ecran va fi reprezentat graficul din Figura 4.1, în cazul în care `a=-1` şi `b=1`. Să remarcăm totuşi că la un nou apel, cu aceeaşi parametri, graficul diferă, deoarece sunt generate alte numere aleatoare.

4.2.2 Momente de selecţie

Definiţia 4.2.9. *Numim* moment de selecţie de ordin k *funcţia de selecţie*

$$\bar{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad \text{iar} \quad \bar{\nu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k,$$

se numeşte valoarea momentului de selecţie de ordin k.

Observaţia 4.2.10. Se vede imediat că $\bar{\nu}_1 = \bar{X}$.

Proprietatea 4.2.11. *Fie caracteristica X pentru care există momentul teoretic de ordin $2k$, $\nu_{2k} = E\left(X^{2k}\right)$, atunci*

$$E\left(\bar{\nu}_k\right) = \nu_k \quad \text{şi} \quad Var\left(\bar{\nu}_k\right) = \frac{1}{n}\left(\nu_{2k} - \nu_k^2\right),$$

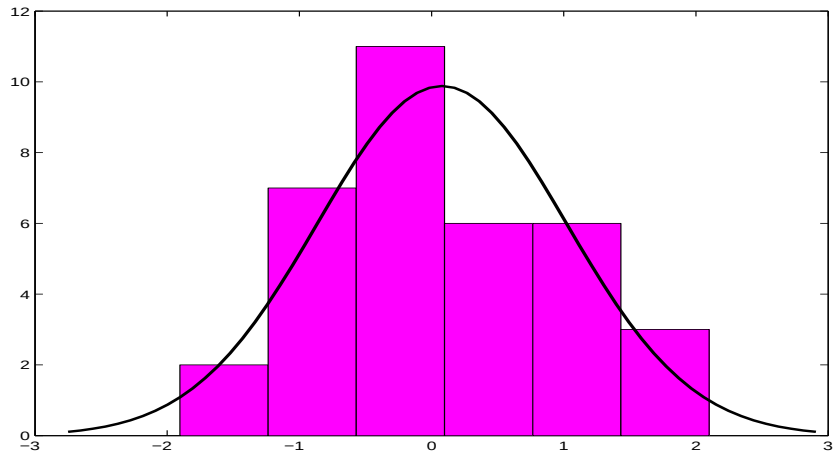

Figura 4.1: Legea $\mathcal{N}(0,1)$

iar pentru $n \to \infty$*, avem că*

$$Z_n = \frac{\bar{\nu}_k - \nu_k}{\frac{\sqrt{\nu_{2k}-\nu_k^2}}{\sqrt{n}}}$$

converge în repartiţie la legea normală $\mathcal{N}(0,1)$.

Demonstraţie. Deoarece selecţia este repetată putem scrie succesiv

$$\begin{aligned} E(\bar{\nu}_k) &= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i^k\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X^k\right) = \frac{1}{n} n\nu_k = \nu_k, \\ Var(\bar{\nu}_k) &= \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X_i^k\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X^k\right) = \frac{1}{n^2} n\left(\nu_{2k}-\nu_k^2\right) \\ &= \frac{1}{n}\left(\nu_{2k}-\nu_k^2\right). \end{aligned}$$

Ultima afirmaţie rezultă prin aplicarea Teoremei 2.7.8 şirului $\left(X_n^k\right)_{n\geqslant 1}$ de variabile aleatoare independente şi identic repartizate. Anume, deoarece $E\left(X_n^k\right) = \nu_k$ şi $Var\left(X_n^k\right) = \nu_{2k} - \nu_k^2$, avem

$$\begin{aligned} Z_n =& \sum_{i=1}^{n} \frac{X_i^k - \nu_k}{\sqrt{\nu_{2k}-\nu_k^2}\sqrt{n}} = \frac{1}{\sqrt{\nu_{2k}-\nu_k^2}\sqrt{n}}\left(\sum_{i=1}^{n} X_i^k - n\nu_k\right) \\ =& \frac{1}{\sqrt{\nu_{2k}-\nu_k^2}\sqrt{n}}\left(n\bar{\nu}_k - n\nu_k\right) = \frac{\bar{\nu}_k - \nu_k}{\frac{\sqrt{\nu_{2k}-\nu_k^2}}{\sqrt{n}}}, \end{aligned}$$

dar Z_n converge în repartiţie la legea normală $\mathcal{N}(0,1)$, ceea ce încheie demonstraţia. □

Definiţia 4.2.12. *Numim* moment centrat de selecţie de ordin k *funcţia de selecţie*

$$\bar{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^k, \quad \text{iar} \quad \bar{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^k,$$

se numeşte valoarea momentului centrat de selecţie de ordin k.

Observaţia 4.2.13. Se vede imediat că $\bar{\mu}_1 = 0$, iar momentul centrat de ordinul doi se poate scrie $\bar{\mu}_2 = \bar{\nu}_2 - \bar{\nu}_1^2$.

Proprietatea 4.2.14. *Fie caracteristica X pentru care există momentul teoretic ν_4, atunci pentru momentul centrat de ordinul doi avem*

$$E\left(\bar{\mu}_2\right) = \frac{n-1}{n}\sigma^2 \quad \text{şi} \quad Var\left(\bar{\mu}_2\right) = \frac{n-1}{n^3}\left[(n-1)\mu_4 - (n-3)\sigma^4\right],$$

unde $\sigma^2 = Var(X)$, şi de asemenea

$$Cov\left(\bar{X}, \bar{\mu}_2\right) = \frac{n-1}{n^2}\mu_3.$$

Demonstraţie. Pentru prima relaţie scriem succesiv

$$\begin{aligned}
E\left(\bar{\mu}_2\right) =& E\left(\bar{\nu}_2\right) - E\left(\bar{\nu}_1^2\right) = \nu_2 - \frac{1}{n^2}E\left(\sum_{k=1}^{n}X_k^2 + 2\sum_{\substack{i,j=1\\ i<j}}^{n}X_iX_j\right)\\
=& \nu_2 - \frac{1}{n^2}\left[\sum_{k=1}^{n}E\left(X_k^2\right) + 2\sum_{\substack{i,j=1\\ i<j}}^{n}E\left(X_i\right)E\left(X_j\right)\right]\\
=& \nu_2 - \frac{1}{n^2}\left[n\nu_2 + n(n-1)\nu_1^2\right] = \nu_2 - \frac{1}{n}\nu_2 - \frac{n-1}{n}\nu_1^2\\
=& \frac{n-1}{n}\left(\nu_2 - \nu_1^2\right) = \frac{n-1}{n}\sigma^2.
\end{aligned}$$

Dacă se reţin extremităţile şirului de egalităţi avem prima relaţie.

Pentru a calcula dispersia, considerăm variabilele aleatoare reduse, notate prin $Y_k = X_k - E(X_k)$, $k = \overline{1,n}$, care sunt independente şi identic repartizate şi pentru care $E(Y_k) = 0$, $Var(Y_k) = \sigma^2$, $k = \overline{1,n}$.

Se arată uşor că

$$\bar{\mu}_2 = \frac{1}{n}\sum_{k=1}^{n}\left(Y_k - \bar{Y}\right)^2, \quad \text{unde} \quad \bar{Y} = \frac{1}{n}\sum_{k=1}^{n}Y_k.$$

Pe de altă parte avem că

$$Var\left(\bar{\mu}_2\right) = E\left(\bar{\mu}_2^2\right) - \left(E\left(\bar{\mu}_2\right)\right)^2 = E\left(\bar{\mu}_2^2\right) - \left(\frac{n-1}{n}\right)^2 \sigma^4.$$

A rămas, prin urmare, de calculat $E\left(\bar{\mu}_2^2\right)$. Pentru aceasta avem

$$\begin{aligned} E\left(\bar{\mu}_2^2\right) =& \frac{1}{n^2} E\left[\left(\sum_{k=1}^{n}\left(Y_k - \bar{Y}\right)^2\right)^2\right] = \frac{1}{n^2} E\left[\left(\sum_{k=1}^{n} Y_k^2 - 2\bar{Y}\sum_{k=1}^{n} Y_k + n\bar{Y}^2\right)^2\right] \\ =& \frac{1}{n^2} E\left[\left(\sum_{k=1}^{n} Y_k^2 - n\bar{Y}^2\right)^2\right] = \frac{1}{n^2} E\left[\left(\sum_{k=1}^{n} Y_k^2\right)^2 - 2n\bar{Y}^2\sum_{k=1}^{n} Y_k^2 + n^2\bar{Y}^4\right], \end{aligned}$$

de unde se obţine

$$E\left(\bar{\mu}_2^2\right) = \frac{1}{n^2} E\left[\left(\sum_{k=1}^{n} Y_k^2\right)^2\right] - \frac{2}{n} E\left(\bar{Y}^2\sum_{k=1}^{n} Y_k^2\right) + E\left(\bar{Y}^4\right).$$

Calculăm pe rând termenii din membrul drept al aceste relaţii. Astfel avem

$$E\left[\left(\sum_{k=1}^{n} Y_k^2\right)^2\right] = \sum_{k=1}^{n} E\left(Y_k^4\right) + 2\sum_{\substack{i,j=1 \\ i<j}} E\left(Y_i^2\right) E\left(Y_j^2\right) = n\mu_4 + n\left(n-1\right)\mu_2^2,$$

apoi

$$\begin{aligned} E\left(\bar{Y}^2\sum_{k=1}^{n} Y_k^2\right) =& \frac{1}{n^2} E\left[\left(\sum_{i=1}^{n} Y_i^2 + 2\sum_{\substack{i,j=1 \\ i<j}}^{n} Y_i Y_j\right)\left(\sum_{k=1}^{n} Y_k^2\right)\right] \\ =& \frac{1}{n^2} E\left[\sum_{k=1}^{n} Y_k^4 + 2\sum_{\substack{i,j=1 \\ i<j}}^{n} Y_i^2 Y_j^2 + 2\sum_{\substack{i,j,k=1 \\ i<j}}^{n} Y_i Y_j Y_k^2\right] \\ =& \frac{1}{n^2}\left[n\mu_4 + n\left(n-1\right)\mu_2^2\right] = \frac{1}{n}\mu_4 + \frac{n-1}{n}\mu_2^2, \end{aligned}$$

deoarece $E\left(Y_i Y_j Y_k^2\right) = 0$, pentru orice $i, j, k = \overline{1,n}$, $i \neq j$. Pentru ultimul termen avem

$$\begin{aligned} E\left(\bar{Y}^4\right) =& \frac{1}{n^4} E\left[\sum_{k=1}^{n} Y_k^4 + 6\sum_{\substack{i,j=1 \\ i<j}}^{n} Y_i^2 Y_j^2 + \cdots\right] \\ =& \frac{1}{n^4}\left[n\mu_4 + 3n\left(n-1\right)\mu_2^2\right] = \frac{1}{n^3}\mu_4 + \frac{3\left(n-1\right)}{n^3}\mu_2^2. \end{aligned}$$

Termenii lui $E\left(\bar{Y}^4\right)$, care nu au fost luaţi în considerare, sunt nuli deoarece conţin ca factor pe $E\left(Y_i\right)=0$. Obţinem în acest fel că

$$\begin{aligned}E\left(\bar{\mu}_2^2\right)=&\frac{1}{n^2}\left[n\mu_4+n\left(n-1\right)\mu_2^2\right]-\frac{2}{n}\left[\frac{1}{n}\mu_4+\frac{n-1}{n}\mu_2^2\right]+\frac{1}{n^3}\mu_4+\frac{3\left(n-1\right)}{n^3}\mu_2^2\\=&\frac{\left(n-1\right)^2}{n^3}\mu_4+\frac{\left(n-1\right)\left(n^2-2n+3\right)}{n^3}\mu_2^2,\end{aligned}$$

deci pentru dispersie avem succesiv

$$\begin{aligned}Var\left(\bar{\mu}_2\right)=&\frac{\left(n-1\right)^2}{n^3}\mu_4+\frac{\left(n-1\right)\left(n^2-2n+3\right)}{n^3}\mu_2^2-\frac{\left(n-1\right)^2}{n^2}\mu_2^2\\=&\frac{\left(n-1\right)^2}{n^3}\mu_4-\frac{\left(n-1\right)\left(n-3\right)}{n^3}\mu_2^2=\frac{n-1}{n^3}\left[\left(n-1\right)\mu_4-\left(n-3\right)\mu_2^2\right],\end{aligned}$$

adică

$$Var\left(\bar{\mu}_2\right)=\frac{n-1}{n^3}\left[\left(n-1\right)\mu_4-\left(n-3\right)\mu_2^2\right].$$

Pentru ultima relaţie putem considera $E\left(X\right)=0$, deci $E\left(\bar{X}\right)=0$, caz în care avem

$$Cov\left(\bar{X},\bar{\mu}_2\right)=E\left(\bar{X}\ \bar{\mu}_2\right)=\frac{1}{n^2}E\left[\left(\sum_{k=1}^{n}X_k\right)\left(\sum_{k=1}^{n}X_k^2-n\bar{X}^2\right)\right]$$

$$=\frac{1}{n^2}E\left[\left(\sum_{k=1}^{n}X_k\right)\left(\sum_{i=1}^{n}X_i^2\right)\right]-E\left(\bar{X}^3\right)=\frac{1}{n^2}\sum_{k=1}^{n}E\left(X_k^3\right)-\frac{1}{n^3}\sum_{k=1}^{n}E\left(X_k^3\right),$$

deoarece $E\left(X_iX_j\right)=E\left(X_iX_j^2\right)=0$, când $i\neq j$. Aşadar

$$Cov\left(\bar{X},\bar{\mu}_2\right)=\frac{\mu_3}{n}-\frac{\mu_3}{n^2}=\frac{n-1}{n^2}\mu_3.$$

□

Observaţia 4.2.15. Din proprietatea precedentă avem că

$$Var\left(\bar{\mu}_2\right)=\frac{1}{n}\left(\mu_4-\mu_2^2\right)+O\left(\frac{1}{n^2}\right),\quad\text{deci}\quad Var\left(\bar{\mu}_2\right)\cong\frac{1}{n}\left(\mu_4-\mu_2^2\right).$$

Deoarece $Cov\left(\bar{X},\bar{\mu}_2\right)=\frac{n-1}{n^2}\mu_3$, avem că statisticile $\bar{X}$ şi $\bar{\mu}_2$ sunt necorelate pentru $n\to\infty$, iar dacă X are distribuţia simetrică ($\mu_3=0$), atunci $\bar{X}$ şi $\bar{\mu}_2$ sunt necorelate pentru orice n.

De asemenea se poate arăta că statistica

$$Z_n = \frac{\bar{\mu}_2 - \mu_2}{\frac{\sqrt{\mu_4 - \mu_2^2}}{\sqrt{n}}}$$

converge în repartiţie la legea normală $\mathcal{N}(0,1)$, când $n \to \infty$.

Definiţia 4.2.16. *Numim* dispersie de selecţie *funcţia de selecţie*

$$\bar{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2, \text{ iar valoarea numerică } \quad \bar{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \bar{x}\right)^2,$$

se numeşte valoarea dispersiei de selecţie.

Observaţia 4.2.17. Între momentul centrat de selecţie de ordinul doi şi dispersia de selecţie există relaţia

$$\bar{\sigma}^2 = \frac{n}{n-1}\bar{\mu}_2. \tag{4.2.2}$$

Ca urmare avem că

$$E\left(\bar{\sigma}^2\right) = \mu_2 = \sigma^2, \quad Cov\left(\bar{X}, \bar{\sigma}^2\right) = \frac{1}{n}\mu_3,$$
$$Var\left(\bar{\sigma}^2\right) = \frac{1}{n(n-1)}\left[(n-1)\mu_4 - (n-3)\mu_2^2\right].$$

Proprietatea 4.2.18. *Fie caracteristica X pentru care există momentul centrat teoretic $\mu_k = E\left[\left(X - E(X)\right)^k\right]$, atunci $E(\bar{\mu}_k) = \mu_k + O\left(\frac{1}{n}\right)$.*

Demonstraţie. Folosind formula binomului avem că

$$\bar{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^k = \frac{1}{n}\sum_{i=1}^{n}\left[X_i^k - \binom{k}{1}X_i^{k-1}\bar{X} + \cdots + (-1)^k\bar{X}^k\right]$$
$$= \bar{\nu}_k - \binom{k}{1}\bar{X}\,\bar{\nu}_{k-1} + \cdots + (-1)^k\bar{X}^k.$$

Fără a restrânge generalitatea, se poate considera că $E(X) = 0$, deci au loc egalităţile $\mu_i = \nu_i$, $i = \overline{1,k}$. Putem scrie atunci

$$E(\bar{\mu}_k) = E(\bar{\nu}_k) - \binom{k}{1}E\left(\bar{X}\,\bar{\nu}_{k-1}\right) + \cdots + (-1)^k E\left(\bar{X}^k\right)$$

$$= \mu_k - \binom{k}{1} E\left(\bar{X}\,\bar{\nu}_{k-1}\right) + \cdots + (-1)^k E\left(\bar{X}^k\right).$$

Pe de altă parte, din inegalitatea lui Schwarz, se obţine că

$$E^2\left(\bar{X}^i\,\bar{\nu}_{k-i}\right) \leqslant E\left(\bar{X}^{2i}\right) E\left(\bar{\nu}_{k-i}^2\right), \quad i = \overline{2,k}.$$

Dacă se are în vedere Proprietatea 4.2.11, rezultă că

$$\begin{aligned} E\left(\bar{\nu}_{k-i}^2\right) &= Var\left(\bar{\nu}_{k-i}\right) + E^2\left(\bar{\nu}_{k-i}\right) = \frac{1}{n}\left[\nu_{2(k-i)} - \nu_{k-i}^2\right] + \nu_{k-i}^2 \\ &= \mu_{k-i}^2 + \frac{1}{n}\left[\mu_{2(k-i)} - \mu_{k-i}^2\right]. \end{aligned}$$

Se poate arăta, de asemenea, că $E\left(\bar{X}^{2i}\right) = O\left(\frac{1}{n^i}\right)$. Astfel din inegalitatea lui Schwarz se obţine

$$E^2\left(\bar{X}^i\,\bar{\nu}_{k-i}\right) \leqslant O\left(\frac{1}{n^i}\right) \quad \text{sau} \quad E\left(\bar{X}^i\,\bar{\nu}_{k-i}\right) \leqslant O\left(\frac{1}{n^{\frac{i}{2}}}\right), \quad i = \overline{2,k}.$$

În cazul $i = 1$, se calculează direct

$$E\left(\bar{X}\,\bar{\nu}_{k-1}\right) = \frac{1}{n^2} E\left[\left(\sum_{i=1}^n X_i\right)\left(\sum_{j=1}^n X_j^{k-1}\right)\right] = \frac{1}{n^2}\sum_{i=1}^n E\left(X_i^k\right) = \frac{1}{n}\mu_k.$$

Luând în considerare aceste evaluări, se obţine în final

$$E\left(\bar{\mu}_k\right) = \mu_k - \frac{k}{n}\mu_k + O\left(\frac{1}{n}\right) = \mu_k + O\left(\frac{1}{n}\right).$$

□

Observaţia 4.2.19. Printr–o cale analogă, se poate obţine că

$$Var\left(\bar{\mu}_k\right) = \frac{\mu_{2k} - 2k\mu_{k-1}\mu_{k+1} - \mu_k^2 + k^2\mu_k\mu_{k-1}^2}{n} + O\left(\frac{1}{n^2}\right).$$

4.2.3 Coeficient de corelaţie de selecţie

Definiţia 4.2.20. *Fie caracteristica bidimensinală (X, Y) şi o selecţie repetată de volum n, cu datele de selecţie (x_k, y_k), $k = \overline{1,n}$, şi respectiv variabilele de selecţie (X_k, Y_k), $k = \overline{1,n}$. Numim* coeficient de corelaţie de selecţie *funcţia de selecţie*

$$\bar{r} = \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2}},$$

iar valoarea numerică

$$\bar{r}=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}},$$

se numeşte valoarea coeficientului de corelaţie de selecţie.

Observaţia 4.2.21. Fisher a arătat că pentru o caracteristică bidimensională (X,Y), care urmează legea normală $\mathcal{N}\left(m_1,m_2;\sigma_1,\sigma_2;r\right)$, densitatea de probabilitate a coeficientului de corelaţie de selecţie este

$$\begin{aligned}f_n\left(x\right)&=\frac{n-2}{\pi}\sqrt{\left(1-r^2\right)^{n-1}}\sqrt{\left(1-x^2\right)^{n-4}}\int_0^1\frac{u^{n-2}}{\left(1-rxu\right)^{n-1}}\frac{du}{\sqrt{1-u^2}}\\&=\frac{2^{n-3}}{\pi\left(n-3\right)!}\sqrt{\left(1-r^2\right)^{n-1}}\sqrt{\left(1-x^2\right)^{n-4}}\times\\&\qquad\times\sum_{k=0}^{\infty}\Gamma^2\left(\frac{n+k-1}{2}\right)\frac{\left(2rx\right)^k}{k!},\ x\in\left(-1,1\right).\end{aligned}$$

De asemenea, dacă se consideră transformarea (lui Fisher) $Z=\frac{1}{2}\ln\frac{1+\bar{r}}{1-\bar{r}}$ şi dacă $\rho=\frac{1}{2}\ln\frac{1+r}{1-r}$, atunci Z urmează aproximativ legea normală $\mathcal{N}\left(\rho+\frac{r}{2(n-1)},\frac{1}{\sqrt{n-3}}\right)$.

În cazul particular $r=0$, avem densitatea de probabilitate a lui $\bar{r}$ exprimată prin

$$f_n\left(x\right)=\frac{1}{\sqrt{\pi}}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n-2}{2}\right)}\sqrt{\left(1-x^2\right)^{n-4}},\quad x\in\left(-1,1\right),$$

iar statistica

$$T=\sqrt{n-2}\,\frac{\bar{r}}{\sqrt{1-\bar{r}^2}} \tag{4.2.3}$$

urmează legea Student cu $n-2$ grade de libertate.

Programul 4.2.22. Vom scrie un program, care generează de `N` ori câte `n` vectori aleatori, ce urmează legea normală bidimensională, având coeficientul de corelaţie dintre cele două componente nul, adică `r=0`. Folosind aceste date se vor calcula `N` numere aleatoare după regula precizată prin statistica (4.2.3), iar histograma corespunzătoare acestor noi date va fi reprezentată grafic împreună cu densitatea de probabilitate a legii Student cu `n-2` grade de libertate.

```
clear,clf,
mu(1)=input('m1='); mu(2)=input('m2=');
v(1,1)=input('sigma1^2=');
v(2,2)=input('sigma2^2=');
v(1,2)=0; v(2,1)=0;
```

```
n=input('n='); N=input('N=');
for k=1:N
Z=mvnrnd(mu,v,n); r=corrcoef(Z); r12=r(1,2);
t(k)=sqrt(n-2)*r12/sqrt(1-r12^2);
end
x=-3:0.01:3; f=tpdf(x,n-2);
nn=fix(1+10/3*log10(N));
[fr,cl]=hist(t,nn); h=cl(2)-cl(1);
bar(cl,fr/(h*N),1),hold on, plot(x,f,'k-')
colormap spring
```

Pentru `N=50`, `n=20`, `mu=(5,10)`, `v` $= \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$, se obţin graficele din Figura 4.2.

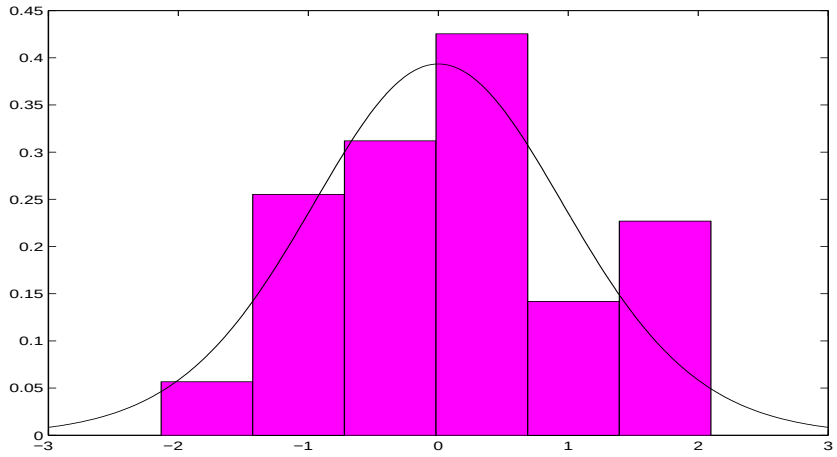

Figura 4.2: Legea $\mathcal{T}(18)$

Lema 4.2.23 (Fisher). *Dacă variabilele aleatoare $X_1, X_2, \ldots, X_n$ sunt independente, fiecare urmând legea normală $\mathcal{N}(0,1)$ şi dacă se consideră matricea ortonormată $\boldsymbol{A} = (a_{ij})_{i,j=\overline{1,n}}$, atunci variabilele aleatoare*

$$Y_i = \sum_{k=1}^{n} a_{ik} X_k, \quad i = \overline{1,n},$$

sunt independente, fiecare urmând legea normală $\mathcal{N}(0,1)$.

Demonstraţie. Pentru a arăta că variabilele aleatoare Y_i, $i = \overline{1,n}$, sunt independente determinăm funcţia caracteristică φ a vectorului aleator $(Y_1, Y_2, \ldots, Y_n)$. Pornind de

la definiţia funcţiei caracteristice avem

$$\begin{aligned}\varphi(t_1,t_2,\ldots,t_n) =& E\left[\exp\left(i\sum_{k=1}^{n}t_kY_k\right)\right] = E\left[\exp\left(i\sum_{k=1}^{n}t_k\sum_{j=1}^{n}a_{kj}X_j\right)\right]\\ =& E\left[\exp\left(i\sum_{j=1}^{n}\left(\sum_{k=1}^{n}t_ka_{kj}\right)X_j\right)\right].\end{aligned}$$

Variabilele aleatoare X_j, $j=\overline{1,n}$, fiind independente se obţine

$$\varphi(t_1,t_2,\ldots,t_n)=\prod_{j=1}^{n}E\left[\exp\left(i\sum_{k=1}^{n}t_ka_{kj}\right)X_j\right)\right].$$

Fiecare factor din membrul drept este valoarea funcţiei caracteristice a legii normale $\mathcal{N}(0,1)$, pe punctele, respectiv $\sum_{k=1}^{n}t_ka_{kj}$, $j=\overline{1,n}$, adică

$$\begin{aligned}\varphi(t_1,t_2,\ldots,t_n) &= \prod_{j=1}^{n}\exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}t_ka_{kj}\right)^2\right] = \exp\left[-\frac{1}{2}\sum_{j=1}^{n}\left(\sum_{k=1}^{n}t_ka_{kj}\right)^2\right]\\ &= \exp\left[-\frac{1}{2}\sum_{j=1}^{n}\left(\sum_{k=1}^{n}t_k^2a_{kj}^2+2\sum_{\substack{i,s=1\\ i<s}}^{n}t_it_sa_{ij}a_{sj}\right)\right]\\ &= \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}t_k^2\sum_{j=1}^{n}a_{kj}^2+2\sum_{\substack{i,s=1\\ i<s}}^{n}t_it_s\sum_{j=1}^{n}a_{ij}a_{sj}\right)\right].\end{aligned}$$

Folosind condiţiile de ortonormalitate relative la coeficienţii a_{ij}, se obţine că

$$\varphi(t_1,t_2,\ldots,t_n)=\exp\left(-\frac{1}{2}\sum_{k=1}^{n}t_k^2\right),$$

adică faptul că $Y_1,Y_2,\ldots,Y_n$ sunt variabile aleatoare independente, ce urmează fiecare legea normală $\mathcal{N}(0,1)$. □

Proprietatea 4.2.24. *Fie caracteristica X, ce urmează legea normală $\mathcal{N}(0,1)$ şi variabilele de selecţie $X_1,X_2,\ldots,X_n$, ce corespund unei selecţii repetate de volum n, atunci statisticile*

$$U_n=\sqrt{n}\,\bar{X}=\frac{1}{\sqrt{n}}\sum_{k=1}^{n}X_k,\quad V_n=\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2,$$

sunt variabile aleatoare independente, ce urmează legea normală $\mathcal{N}(0,1)$ şi respectiv legea χ^2 cu $n-1$ grade de libertate.

Demonstraţie. Se consideră matricea ortonormată

$$\boldsymbol{A} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \dots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & \dots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \dots & -\frac{n-1}{\sqrt{n(n-1)}} \end{pmatrix}$$

pentru care se aplică Lema 4.2.23.

Pe de o parte avem că

$$Y_1 = \sum_{k=1}^{n} a_{1k} X_k = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k = U_n$$

şi

$$\begin{aligned} Y_2^2 + \dots + Y_n^2 &= \left(Y_1^2 + Y_2^2 + \dots + Y_n^2\right) - Y_1^2 = \left(X_1^2 + \dots + X_n^2\right) - U_n^2 \\ &= \sum_{k=1}^{n} \left(X_k - \bar{X}\right)^2 = V_n. \end{aligned}$$

Deoarece $Y_1, Y_2, \dots, Y_n$ sunt independente rezultă că U_n şi V_n sunt independente.

Pe de altă parte $U_n = Y_1$ urmează legea normală $\mathcal{N}(0,1)$, iar V_n fiind suma pătratelor a $n-1$ variabile aleatoare independente, fiecare urmând legea normală $\mathcal{N}(0,1)$, obţinem că V_n urmează legea χ^2 cu $n-1$ grade de libertate (a se vedea legea χ^2 prezentată în Capitolul 2). □

Proprietatea 4.2.25. *Fie caracteristica X, ce urmează legea normală $\mathcal{N}(m,\sigma)$ şi variabilele de selecţie $X_1, X_2, \dots, X_n$, ce corespund unei selecţii repetate de volum n, atunci statisticile*

$$U_n = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}, \quad V_n = \frac{1}{\sigma^2} \sum_{k=1}^{n} \left(X_k - \bar{X}\right)^2,$$

sunt variabile aleatoare independente, ce urmează respectiv legea normală $\mathcal{N}(0,1)$ şi legea χ^2 cu $n-1$ grade de libertate.

Demonstraţie. Se consideră variabilele aleatoare

$$Z_k = \frac{X_k - m}{\sigma}, \quad k = \overline{1,n},$$

care sunt variabile aleatoare independente ce urmează fiecare legea normală $\mathcal{N}(0,1)$. Prin aplicarea Proprietăţii 4.2.24 pentru variabilele aleatoare Z_k, $k=\overline{1,n}$, se obţine afirmaţia făcută. Într-adevăr avem că

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n} Z_k = \frac{1}{\sqrt{n}}\sum_{k=1}^{n}\frac{X_k-m}{\sigma} = U_n$$

şi

$$\sum_{k=1}^{n}\left(Z_k-\bar{Z}\right)^2 = \sum_{k=1}^{n}\left(\frac{X_k-m}{\sigma}-\frac{1}{n}\sum_{i=1}^{n}\frac{X_i-m}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{k=1}^{n}\left(X_k-\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2$$
$$= \frac{1}{\sigma^2}\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2 = V_n.$$

□

Proprietatea 4.2.26. *Fie caracteristica X, ce urmează legea normală $\mathcal{N}(m,\sigma)$ şi variabilele de selecţie $X_1, X_2, \ldots, X_n$, ce corespund unei selecţii repetate de volum n, atunci statistica*

$$T = \frac{\bar{X}-m}{\frac{\tilde{\sigma}}{\sqrt{n}}} = \frac{\bar{X}-m}{\sqrt{\frac{\tilde{\mu}_2}{n-1}}}, \tag{4.2.4}$$

urmează legea Student cu $n-1$ grade de libertate.

Demonstraţie. Cu notaţiile de la Proprietatea 4.2.25, arătăm că

$$T = \frac{U_n}{\sqrt{\frac{V_n}{n-1}}}.$$

Într-adevăr, avem succesiv

$$\frac{U_n}{\sqrt{\frac{V_n}{n-1}}} = \frac{\bar{X}-m}{\frac{\sigma}{\sqrt{n}}}\cdot\frac{\sqrt{n-1}}{\frac{1}{\sigma}\sqrt{\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2}} = \frac{\bar{X}-m}{\frac{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2}}{\sqrt{n}}} = \frac{\bar{X}-m}{\frac{\tilde{\sigma}}{\sqrt{n}}} = T.$$

Din teoria probabilităţilor se ştie că raportul dintre o variabilă aleatoare, ce urmează legea normală $\mathcal{N}(0,1)$ şi radicalul unei variabile aleatoare ce urmează legea χ^2, raportată la numărul gradelor de libertate, în cazul în care cele două variabile aleatoare sunt independente, este o variabilă aleatoare, ce urmează legea Student cu acelaşi număr al gradelor de libertate ca legea χ^2 considerată (a se vedea legile χ^2 şi t prezentate în Capitolul 2). Având în vedere cine sunt U_n şi V_n rezultă afirmaţia din enunţul proprietăţii. □

Programul 4.2.27. Pentru ilustrarea acestui rezultat, programul Matlab, care urmează, generează o matrice de tipul `(n,m)` de numere aleatoare, ce urmează legea normală $\mathcal{N}(\mu, \sigma)$, după care pentru fiecare coloană a matricei generate, se construiesc datele aleatoare conform statisticii (4.2.4). Pentru aceste date se reprezintă grafic histograma corespunzătoare, împreună cu densitatea de probabilitate a legii Student cu `n-1` grade de libertate.

```
clear,clf, mu=input('mu='); sigma=input('s=');
n=input('n='); m=input('m=');
Z=normrnd(mu,sigma,n,m);
t=(mean(Z)-mu)./sqrt(var(Z)/n);
x=-3:0.01:3; f=tpdf(x,n-1);
nn=fix(1+10/3*log10(m));
[fr,clasa]=hist(t,nn);
h=clasa(2)-clasa(1);
bar(clasa,fr/(h*m),1)
hold on, plot(x,f,'k-')
colormap([.7,.7,.7])
```

Pentru `n=15`, `m=100`, `mu=5` şi `sigma=2`, se obţin graficele din Figura 4.3.

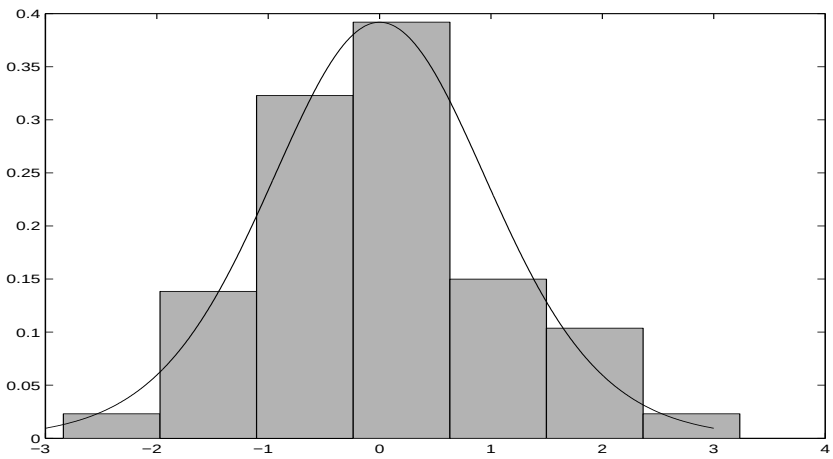

Figura 4.3: Legea $\mathcal{T}(14)$

Proprietatea 4.2.28. *Fie caracteristicile independente X' şi X'', fiecare urmând legea normală, respectiv $\mathcal{N}(m', \sigma)$ şi $\mathcal{N}(m'', \sigma)$ şi variabilele de selecţie $X'_1, \ldots, X'_{n'}$, respectiv $X''_1, \ldots, X''_{n''}$, ce corespund unei selecţii repetate de volum n' pentru caracteristica X' şi unei selecţii repetate de volum n'' pentru caracteristica*

X'', atunci statistica

$$T = \frac{\left(\bar{X}' - \bar{X}''\right) - (m' - m'')}{\sqrt{(n'-1)\,\bar{\sigma}'^2 + (n''-1)\,\bar{\sigma}''^2}} \sqrt{\frac{n' + n'' - 2}{\frac{1}{n'} + \frac{1}{n''}}}, \tag{4.2.5}$$

unde

$$\bar{X}' = \frac{1}{n'} \sum_{k=1}^{n'} X'_k, \quad \bar{X}'' = \frac{1}{n''} \sum_{k=1}^{n''} X''_k,$$

$$\bar{\sigma}'^2 = \frac{1}{n'-1} \sum_{k=1}^{n'} \left(X'_k - \bar{X}'\right)^2, \quad \bar{\sigma}''^2 = \frac{1}{n''-1} \sum_{k=1}^{n''} \left(X''_k - \bar{X}''\right)^2,$$

urmează legea Student cu $n' + n'' - 2$ grade de libertate.

Demonstraţie. Conform Observaţiei 4.2.6, avem că mediile de selecţie $\bar{X}'$ şi $\bar{X}''$ urmează fiecare legea normală respectiv $\mathcal{N}\left(m', \frac{\sigma}{\sqrt{n'}}\right)$ şi $\mathcal{N}\left(m'', \frac{\sigma}{\sqrt{n''}}\right)$. Prin urmare statistica

$$U = \frac{\left(\bar{X}' - \bar{X}''\right) - (m' - m'')}{\sigma \sqrt{\frac{1}{n'} + \frac{1}{n''}}},$$

urmează legea normală $\mathcal{N}(0,1)$.

Pe de altă parte, folosind Proprietatea 4.2.25, se obţine că statistica

$$V = \frac{1}{\sigma^2} \sum_{k=1}^{n'} \left(X'_k - \bar{X}'\right)^2 + \frac{1}{\sigma^2} \sum_{k=1}^{n''} \left(X''_k - \bar{X}''\right)^2,$$

urmează legea χ^2 cu $n' + n'' - 2$ grade de libertate, fiind suma a două variabile aleatoare independente, ce urmează legea χ^2 cu $n' - 1$ grade de libertate şi respectiv $n'' - 1$ grade de libertate.

Ca şi la demonstraţia Proprietăţii 4.2.26, statistica $U \Big/ \sqrt{\frac{V}{n'+n''-2}}$ urmează legea Student cu $n' + n'' - 2$ grade de libertate. Mai rămâne de arătat că această statistică este chiar statistica T. Într-adevăr, avem succesiv

$$\frac{U}{\sqrt{\frac{V}{n'+n''-2}}} = \frac{\left(\bar{X}' - \bar{X}''\right) - (m' - m'')}{\sigma \sqrt{\frac{1}{n'} + \frac{1}{n''}}} \times$$

$$\times \frac{\sqrt{n' + n'' - 2}}{\frac{1}{\sigma} \sqrt{\sum_{k=1}^{n'} \left(X'_k - \bar{X}'\right)^2 + \sum_{k=1}^{n''} \left(X''_k - \bar{X}''\right)^2}}$$

$$=\frac{\left(\bar{X}' - \bar{X}''\right) - (m' - m'')}{\sqrt{(n'-1)\,\bar{\sigma}'^2 + (n''-1)\,\bar{\sigma}''^2}} \sqrt{\frac{n' + n'' - 2}{\frac{1}{n'} + \frac{1}{n''}}} = T,$$

ceea ce trebuie demonstrat. □

Observaţia 4.2.29. Dacă se consideră caracteristicile independente X' şi X'', fiecare urmând legea normală $\mathcal{N}(m', \sigma')$ şi respectiv $\mathcal{N}(m'', \sigma'')$ şi dacă avem variabilele de selecţie $X'_1, X'_2, \ldots, X'_{n'}$, ce corespund unei selecţii repetate de volum n' relativă la caracteristica X' şi variabilele de selecţie $X''_1, X''_2, \ldots, X''_{n''}$, ce corespund unei selecţii repetate de volum n'' relativă la caracteristica X'', atunci statistica

$$Z = \frac{\left(\bar{X}' - \bar{X}''\right) - (m' - m'')}{\sqrt{\frac{\sigma'^2}{n'} + \frac{\sigma''^2}{n''}}},$$

urmează legea normală $\mathcal{N}(0, 1)$.

Proprietatea 4.2.30. *Fie caracteristicile independente X' şi X'', fiecare urmând legea normală, respectiv $\mathcal{N}(m', \sigma')$ şi $\mathcal{N}(m'', \sigma'')$ şi variabilele de selecţie $X'_1, \ldots, X'_{n'}$, respectiv $X''_1, \ldots, X''_{n''}$, ce corespund unei selecţii repetate de volum n' pentru caracteristica X' şi unei selecţii repetate de volum n'' pentru caracteristica X'', atunci statistica*

$$F = \frac{\bar{\sigma}'^2}{\sigma'^2} \Bigg/ \frac{\bar{\sigma}''^2}{\sigma''^2} \tag{4.2.6}$$

urmează legea Fisher–Snedecor cu $m = n' - 1$ şi $n = n'' - 1$ grade de libertate.

Demonstraţie. Din Proprietatea 4.2.25, avem că funcţiile de selecţie

$$V' = \left(n' - 1\right) \frac{\bar{\sigma}'^2}{\sigma'^2}, \quad V'' = \left(n'' - 1\right) \frac{\bar{\sigma}''^2}{\sigma''^2}$$

urmează fiecare legea χ^2 cu $m = n' - 1$ şi $n = n'' - 1$ grade de libertate.

Pe de altă parte, deoarece X' şi X'' sunt independente, V' şi V'' sunt independente. Din calculul probabilităţilor (a se vedea legea f prezentată în Capitolul 2) este cunoscut că raportul a două variabile aleatoare independente, ce urmează legea χ^2, raportate fiecare la numărul gradelor de libertate corespunzător, este o variabilă aleatoare, ce urmează legea Fisher–Snedecor cu numărul gradelor de libertate date de numerele gradelor de libertate ale celor două legi χ^2. Aşadar avem că

$$F = \frac{\bar{\sigma}'^2}{\sigma'^2} \Bigg/ \frac{\bar{\sigma}''^2}{\sigma''^2} = \frac{V'}{n' - 1} \Bigg/ \frac{V''}{n'' - 1}$$

urmează legea Fisher–Snedecor cu $m = n' - 1$ şi $n = n'' - 1$ grade de libertate. □

Observaţia 4.2.31. Dacă F urmează legea Fisher–Snedecor cu m şi n grade de libertate, atunci $\frac{1}{F}$ urmează legea Fisher–Snedecor cu n şi m grade de libertate. Prin urmare, dacă notăm cu $f_{m,n;1-\alpha}$ cuantila de ordin $1-\alpha$ a lui F, adică, numărul pentru care are loc $P\left(F \leqslant f_{m,n;1-\alpha}\right) = 1-\alpha$, atunci avem că $\frac{1}{f_{n,m;\alpha}} = f_{m,n;1-\alpha}$, unde $f_{n,m;\alpha}$ este cuantila de ordin α a lui $\frac{1}{F}$. Într-adevăr avem că relaţia care defineşte pe $f_{n,m;\alpha}$, $P\left(\frac{1}{F} \leqslant f_{n,m;\alpha}\right) = \alpha$, este echivalentă cu $P\left(F \geqslant \frac{1}{f_{n,m;\alpha}}\right) = \alpha$ sau cu relaţia $P\left(F \leqslant \frac{1}{f_{n,m;\alpha}}\right) = 1-\alpha$. Comparând ultima relaţie cu relaţia care defineşte cuantila $f_{m,n;1-\alpha}$, se obţine relaţia dintre cele două cuantile mai sus precizată.

Programul 4.2.32. Pentru ilustrarea Proprietăţilor 4.2.28 şi 4.2.30, vom genera două matrice de tipurile `(n1,m)` şi `(n2,m)`, ale căror elemente sunt numere aleatoare, ce urmează respectiv legile normale $\mathcal{N}(\mu_1, \sigma_1)$ şi $\mathcal{N}(\mu_2, \sigma_2)$. Folosind cele două matrice se vor determina, pentru fiecare pereche de coloane cu acelaşi număr de ordine, din cele două matrice, datele construite conform definiţiilor statisticilor (4.2.5) şi (4.2.6). Datele astfel obţinute se reprezintă grafic folosind histogramele corespunzătoare, împreună cu densitatea legii Student cu `n1+n2-2` grade de libertate, respectiv a legii Fisher–Snedecor cu `n1-1` şi `n2-1` grade de libertate.

```
clear,clf
mu1=input('mu1='); s1=input('sigma1=');
mu2=input('mu2='); s2=input('sigma2=');
n1=input('n1='); n2=input('n2='); m=input('m=');
X1=normrnd(mu1,s1,n1,m); X2=normrnd(mu2,s2,n2,m);
t=(mean(X1)-mean(X2)-mu1+mu2);
t=t./sqrt((n1-1)*var(X1)+(n2-1)*var(X2));
t=t*sqrt((n1+n2-2)/(1/n1+1/n2));
f=(var(X1)/s1^2)./(var(X2)/s2^2);
x1=-3:0.01:3; x2=0:0.01:6;
f1=tpdf(x1,n1+n2-2); f2=fpdf(x2,n1-1,n2-1);
nn=fix(1+10/3*log10(m));
[fr1,c1]=hist(t,nn); [fr2,c2]=hist(f,nn);
h1=c1(2)-c1(1); h2=c2(2)-c2(1);
subplot(1,2,1)
bar(c1,fr1/(m*h1),1), hold on, plot(x1,f1,'k-')
hold off
subplot(1,2,2)
bar(c2,fr2/(m*h2),1), hold on, plot(x2,f2,'k-')
colormap([.7,.7,.7])
```

Pentru `n1=15`, `n2=20`, `m=50`, `mu1=5`, `mu2=10`, `sigma1=2` şi `sigma2=3`, se obţin graficele din Figura 4.4.

4.2.4 Funcţie de repartiţie de selecţie

Definiţia 4.2.33. *Fie caracteristica X cercetată, datele de selecţie $x_1, x_2, \ldots, x_n$ şi variabilele de selecţie $X_1, X_2, \ldots, X_n$. Numim* funcţie de repartiţie de selecţie,

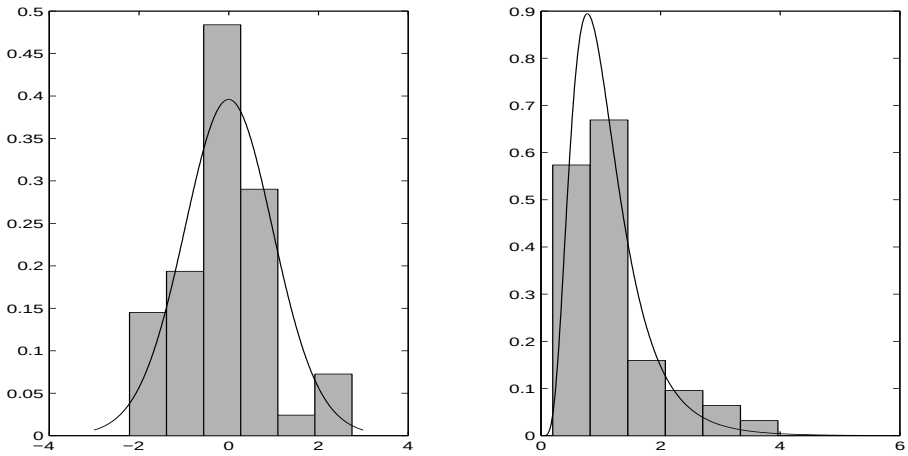

Figura 4.4: Legea $\mathcal{T}(33)$ şi legea $\mathcal{F}(14, 19)$

funcţia de selecţie definită prin

$$\bar{F}_n(x) = \frac{\nu_n(x)}{n}, \quad x \in \mathbb{R},$$

unde

$$\nu_n(x) = \text{card}\{\, X_i \,|\, X_i \leqslant x,\ i = \overline{1,n}\,\},$$

iar

$$\bar{F}_n(x) = \frac{\text{card}\{\, x_i \,|\, x_i \leqslant x,\ i = \overline{1,n}\,\}}{n}, \quad x \in \mathbb{R},$$

se numeşte valoarea funcţiei de repartiţie de selecţie.

Observaţia 4.2.34. Valoarea funcţiei de repartiţie selecţie este o funcţie în scară de o variabilă reală. Dacă datele de selecţie sunt ordonate crescător, atunci

$$\bar{F}_n(x) = \begin{cases} 0, & \text{dacă } x < x_1, \\ \dfrac{k}{n}, & \text{dacă } x_{k-1} \leqslant x < x_k, \quad k = \overline{1,n}, \\ 1, & \text{dacă } x \geqslant x_n. \end{cases}$$

De asemenea, funcţia de repartiţie de selecţie, se vede că este o variabilă aleatoare de tip discret şi care are distribuţia

$$\bar{F}_n(x) \begin{pmatrix} \frac{k}{n} \\ \binom{n}{k} (F(x))^k (1 - F(x))^{n-k} \end{pmatrix}_{k=\overline{0,n}}.$$

Pentru demonstraţia unui rezultat fundamental al statisticii matematicii, descoperit de matematicianul rus Glivenko, vom prezenta două leme.

Lema 4.2.35. *Fie şirul de evenimente* $(A_n)_{n\geqslant 1}$ *astfel încât* $P(A_n) = 1$, $n \geqslant 1$, *atunci, pentru evenimentul* $A = \bigcap_{n=1}^{\infty} A_n$, *avem* $P(A) = 1$.

Demonstraţie. Arătăm la început că proprietatea enunţată are loc pentru un număr finit m de evenimente. Demonstraţia o facem prin inducţie după m.

Pentru $m = 2$, avem $P(A_1 \cup A_2) \geqslant P(A_1) = 1$, de unde $P(A_1 \cup A_2) = 1$, care înlocuită în relaţia

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2),$$

conduce la $P(A_1 \cap A_2) = 1$.

Presupunem că proprietatea are loc pentru m şi arătăm că este adevărată şi pentru $m+1$. Din nou avem că $P\left(\bigcup_{n=1}^{m+1} A_n\right) \geqslant P(A_{m+1}) = 1$, deci $P\left(\bigcup_{n=1}^{m+1} A_n\right) = 1$. Scriem formula lui Poincaré şi ţinem seama de ipoteza inducţiei, ceea ce conduce la

$$1 = P\left(\bigcup_{n=1}^{m+1} A_n\right) = \sum_{i=1}^{m+1} P(A_i) - \sum_{\substack{i,j=1 \\ i<j}}^{m+1} P(A_i \cap A_j) + \cdots + (-1)^m P\left(\bigcap_{n=1}^{m+1} A_n\right)$$

$$= \binom{m+1}{1} - \binom{m+1}{2} + \cdots + (-1)^{m-1}\binom{m+1}{m} + (-1)^m P\left(\bigcap_{n=1}^{m+1} A_n\right).$$

Reţinem extremităţile şi vom obţine

$$(-1)^m P\left(\bigcap_{n=1}^{m+1} A_n\right) = \binom{m+1}{0} - \binom{m+1}{1} + \binom{m+1}{2} - \cdots + (-1)^m \binom{m+1}{m}$$

sau

$$(-1)^m P\left(\bigcap_{n=1}^{m+1} A_n\right) = (1-1)^{m+1} - (-1)^{m+1}\binom{m+1}{m+1},$$

adică $P\left(\bigcap_{n=1}^{m+1} A_n\right) = 1$.

Când şirul de evenimente este infinit, avem în vedere scrierea lui A ca următoarea intersecţie de evenimente

$$A = A_1 \cap (A_1 \cap A_2) \cap (A_1 \cap A_2 \cap A_3) \cap \cdots$$

unde evenimentele intersecţiei satisfac relaţiile

$$A_1 \supset (A_1 \cap A_2) \supset (A_1 \cap A_2 \cap A_3) \supset \cdots$$

Conform teoremei de continuitate din teoria probabilităţilor şi având în vedere prima parte a demonstraţiei putem încheia prin $P(A) = \lim_{m\to\infty} P\left(\bigcap_{n=1}^{m} A_n\right) = 1$. $\square$

Din calculul probabilităţilor este cunoscută forma tare a teoremei lui Bernoulli sau teorema lui Borel şi care o dăm sub forma unei leme.

Lema 4.2.36. *Se consideră n repetări independente ale unui experiment. La fiecare repetare evenimentul A apare cu probabilitatea p. Fie k frecvenţa absolută a apariţiei evenimentului A în cele n repetări, atunci*

$$P\left(\lim_{n\to\infty}\left|\frac{k}{n} - p\right| = 0\right) = 1,$$

adică frecvenţa relativă a apariţiei evenimentului A converge tare (aproape sigur) la probabilitatea p de apariţie a evenimentului A.

Teorema 4.2.37 (Glivenko). *Fie caracteristica X care are funcţia de repartiţie teoretică F şi fie o selecţie repetată de volum n relativă la caracteristica X cu variabilele de selecţie $X_1, X_2, \ldots, X_n$ şi funcţia de repartiţie de selecţie $\bar{F}_n$ corespunzătoare, atunci*

$$P\left(\lim_{n\to\infty}\sup_{x\in\mathbb{R}}\left|\bar{F}_n(x) - F(x)\right| = 0\right) = 1,$$

adică funcţia de repartiţie de selecţie converge aproape sigur la funcţia de repartiţie teoretică.

Demonstraţie. Se consideră $r \in \mathbb{N}$ oarecare, dar fixat, şi considerăm numerele $x_{r,k}$, $k = \overline{0,r}$, definite prin relaţia

$$x_{r,k} = \inf\{\, x \in \mathbb{R} \,|\, F(x) \leqslant \frac{k}{r} \leqslant F(x+0) \,\}.$$

Geometric, punctul $x_{r,k}$ se obţine după cum este arătat în Figura 4.5 şi reprezintă cuantila de ordin $\frac{k}{r}$.

Dacă se consideră evenimentul $B_{r,k} = (X \leqslant x_{r,k})$, atunci avem

$$P\left(B_{r,k}\right) = P\left(X \leqslant x_{r,k}\right) = F\left(x_{r,k}\right)$$

şi

$$\bar{F}_n\left(x_{r,k}\right) = \frac{\nu_n\left(x_{r,k}\right)}{n} = \frac{\text{card}\{\, X_i \,|\, X_i \leqslant x_{r,k},\ i = \overline{1,n}\,\}}{n}.$$

Prin urmare $\bar{F}_n(x_{r,k})$ este frecvenţa relativă a apariţiei evenimentului $B_{r,k}$ şi conform Lemei 4.2.36 rezultă că

$$P\left(\lim_{n\to\infty}\left|\bar{F}_n\left(x_{r,k}\right) - F\left(x_{r,k}\right)\right| = 0\right) = 1,$$

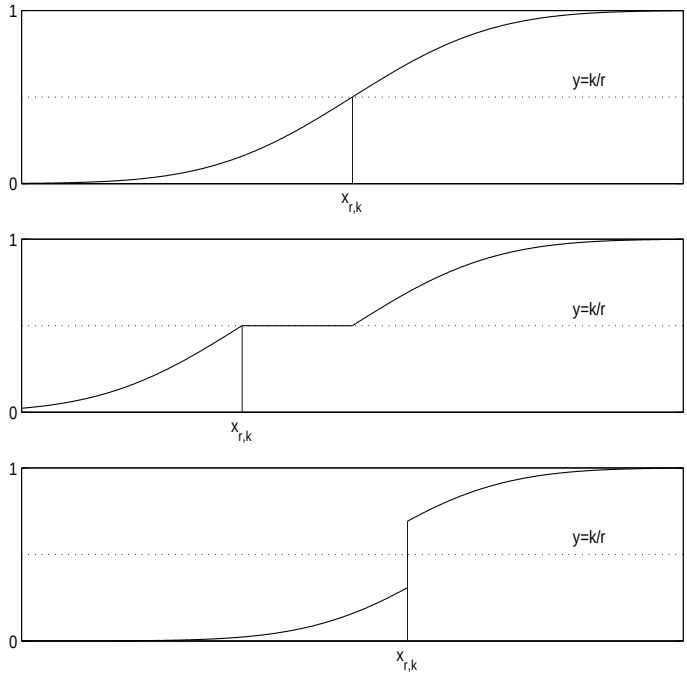

Figura 4.5: Determinarea cuantilei $x_{r,k}$

pentru fiecare $k = \overline{1,r}$. Se poate renunţa la valoarea $k = 0$, deoarece $x_{r,0} = -\infty$ şi $\bar{F}_n(-\infty) = F(-\infty) = 0$.

Notăm prin $A_{r,k}$ evenimentul

$$\left|\bar{F}_n\left(x_{r,k}\right) - F\left(x_{r,k}\right)\right| \longrightarrow 0, \quad \text{când } n \to \infty.$$

Deoarece $P\left(A_{r,k}\right) = 1$, $k = \overline{1,r}$, conform Lemei 4.2.35, rezultă că

$$P\left(A_r\right) = P\left(\bigcap_{k=1}^{r} A_{r,k}\right) = 1,$$

unde evenimentul A_r înseamnă

$$\max_{k=\overline{1,r}} \left|\bar{F}_n\left(x_{r,k}\right) - F\left(x_{r,k}\right)\right| \longrightarrow 0, \quad \text{când } n \to \infty.$$

În mod analog, considerând evenimentul $C_{r,j} = (X < x_{r,j})$, se obţine că

$$P\left(\lim_{n\to\infty} \left|\bar{F}_n\left(x_{r,j} - 0\right) - F\left(x_{r,j} - 0\right)\right| = 0\right) = 1,$$

pentru fiecare $j = \overline{1,r}$. Dacă se notează prin $D_{r,j}$ evenimentul

$$\left|\bar{F}_n\left(x_{r,j}-0\right)-F\left(x_{r,j}-0\right)\right| \longrightarrow 0, \quad \text{când } n \to \infty.$$

şi $D_r = \bigcap_{j=1}^{r} D_{r,j}$, care înseamnă

$$\max_{j=\overline{1,r}} \left|\bar{F}_n\left(x_{r,j}-0\right)-F\left(x_{r,j}-0\right)\right| \longrightarrow 0, \quad \text{când } n \to \infty,$$

atunci $P(D_r) = 1$.

Dacă se consideră evenimentul $E_r = A_r \cap D_r$, care înseamnă

$$\max_{k,j=\overline{1,r}} \left\{ \left|\bar{F}_n\left(x_{r,k}\right)-F\left(x_{r,k}\right)\right|, \left|\bar{F}_n\left(x_{r,j}-0\right)-F\left(x_{r,j}-0\right)\right| \right\} \longrightarrow 0,$$

atunci $P(E_r) = 1$. În acest fel, dacă $E = \bigcap_{n=1}^{\infty} E_r$, conform Lemei 4.2.35, avem $P(E) = 1$, adică

$$P\left(\sup_{r\in\mathbb{N}} \max_{k,j=\overline{1,r}} \left\{ \left|\bar{F}_n(x_{r,k})-F(x_{r,k})\right|, \left|\bar{F}_n(x_{r,j}-0)-F(x_{r,j}-0)\right| \right\} \longrightarrow 0\right) = 1.$$

Arătăm în continuare implicaţia

$$E \subset \left(\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left|\bar{F}_n\left(x\right)-F\left(x\right)\right| = 0\right),$$

din care se va obţine

$$1 = P\left(E\right) \leqslant P\left(\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left|\bar{F}_n\left(x\right)-F\left(x\right)\right| = 0\right) \leqslant 1,$$

deci

$$P\left(\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left|\bar{F}_n\left(x\right)-F\left(x\right)\right| = 0\right) = 1.$$

Pentru aceasta să considerăm $x \in \mathbb{R}$ oarecare. Pentru orice $r \in \mathbb{N}$ va exista $k \in \mathbb{N}$ astfel încât $x_{r,k} < x \leqslant x_{r,k+1}$. Având în vedere că $\bar{F}_n$ şi F sunt funcţii nedescrescătoare, putem să scriem următoarele inegalităţi:

$$\bar{F}_n\left(x_{r,k}\right) \leqslant \bar{F}_n(x) \leqslant \bar{F}_n\left(x_{r,k+1}-0\right)$$

şi

$$F\left(x_{r,k}\right) \leqslant F(x) \leqslant F\left(x_{r,k+1}-0\right),$$

de unde se obţine

$$\bar{F}_n\left(x_{r,k}\right)-F\left(x_{r,k+1}-0\right) \leqslant \bar{F}_n\left(x\right)-F\left(x\right) \leqslant \bar{F}_n\left(x_{r,k+1}-0\right)-F\left(x_{r,k}\right).$$

Pe de altă parte, din modul de definire a numerelor $x_{r,k}$, $k=\overline{0,r}$, avem că

$$0 \leqslant F\left(x_{r,k+1}-0\right)-F\left(x_{r,k}\right) \leqslant \frac{1}{r},$$

de unde

$$F\left(x_{r,k}\right) \geqslant F\left(x_{r,k+1}-0\right)-\frac{1}{r} \quad \text{şi} \quad F\left(x_{r,k+1}-0\right) \leqslant F\left(x_{r,k}\right)+\frac{1}{r}\cdot$$

Utilizate în dubla inegalitate de mai înainte, conduc la

$$\bar{F}_n\left(x_{r,k}\right)-F\left(x_{r,k}\right)-\frac{1}{r} \leqslant \bar{F}_n(x)-F(x) \leqslant \bar{F}_n\left(x_{r,k+1}-0\right)-F\left(x_{r,k+1}-0\right)+\frac{1}{r}\cdot$$

Din prima inegalitate se obţine

$$F(x)-\bar{F}_n(x) \leqslant F\left(x_{r,k}\right)-\bar{F}_n\left(x_{r,k}\right)+\frac{1}{r},$$

de unde

$$\begin{aligned}\left|\bar{F}_n(x)-F(x)\right| &\leqslant \left|\bar{F}_n\left(x_{r,k}\right)-F\left(x_{r,k}\right)\right|+\frac{1}{r}\\ &\leqslant \max_{j=\overline{1,r}}\left\{\left|\bar{F}_n\left(x_{r,j}\right)-F\left(x_{r,j}\right)\right|\right\}+\frac{1}{r}\cdot\end{aligned}$$

Din a doua inegalitate se obţine

$$\begin{aligned}\left|\bar{F}_n(x)-F(x)\right| &\leqslant \left|\bar{F}_n\left(x_{r,k+1}-0\right)-F\left(x_{r,k+1}-0\right)\right|+\frac{1}{r}\\ &\leqslant \max_{k=\overline{1,r}}\left\{\left|\bar{F}_n\left(x_{r,k}-0\right)-F\left(x_{r,k}-0\right)\right|\right\}+\frac{1}{r}\cdot\end{aligned}$$

Folosind cele două inegalităţi obţinute rezultă că $\left|\bar{F}_n\left(x\right)-F\left(x\right)\right| \leqslant M_{n,r}$, unde

$$M_{n,r}=\max_{k,j=\overline{1,r}}\left\{\left|\bar{F}_n\left(x_{r,k}\right)-F\left(x_{r,k}\right)\right|,\left|\bar{F}_n\left(x_{r,j}-0\right)-F\left(x_{r,j}-0\right)\right|\right\}+\frac{1}{r},$$

de unde $\left|\bar{F}_n\left(x\right)-F\left(x\right)\right| \leqslant \sup_{r\in\mathbb{N}} M_{n,r}$, pentru orice $x\in\mathbb{R}$.

În final avem că $\sup_{x\in\mathbb{R}}\left|\bar{F}_n\left(x\right)-F\left(x\right)\right| \leqslant \sup_{r\in\mathbb{N}} M_{n,r}$, care conduce la implicaţia care trebuia demonstrată. □

Teorema 4.2.38 (Kolmogorov). *Fie caracteristica X care are funcţia de repartiţie teoretică F continuă şi fie o selecţie repetată de volum n relativă la caracteristica X*

cu variabilele de selecţie $X_1, X_2, \ldots, X_n$ şi corespunzător funcţia de repartiţie de selecţie $\bar{F}_n$, atunci

$$\lim_{n\to\infty} P\left(\sqrt{n}D_n \leqslant x\right) = K\left(x\right), \quad x > 0,$$

unde

$$D_n = \sup_{x\in\mathbb{R}} \left|\bar{F}_n\left(x\right) - F\left(x\right)\right|, \quad \text{iar} \quad K\left(x\right) = \sum_{k=-\infty}^{+\infty} \left(-1\right)^k e^{-2k^2x^2}, \quad x > 0,$$

este funcţia lui Kolmogorov.

Observaţia 4.2.39. Teorema lui Kolmogorov reprezintă comportarea asimptotică a distanţei dintre funcţia de repartiţie de selecţie şi funcţia de repartiţie teoretică. Demonstraţia poate fi găsită în [56].

Observaţia 4.2.40. Funcţia lui Kolmogorov este tabelată, pentru anumite valori, în *Anexa V*.

Observaţia 4.2.41. Pentru calculul valorilor aproximative ale funcţiei lui Kolmogorov se pot utiliza următoarele formule de aproximare

$$K\left(x\right) \approx \begin{cases} 0, & \text{dacă } x \leqslant 0.27 \\ \dfrac{\sqrt{2\pi}}{x} \displaystyle\sum_{i=1}^{3} e^{-(2i-1)^2\pi^2/\left(8x^2\right)}, & \text{dacă } 0.27 < x < 1, \\ 1 - 2\displaystyle\sum_{i=1}^{4} \left(-1\right)^{i-1} e^{-2i^2x^2}, & \text{dacă } 1 \leqslant x < 3.1, \\ 1, & \text{dacă } x \geqslant 3.1. \end{cases}$$

4.2.5 Funcţia `cdfplot`

Statistics toobox dispune de funcţia `cdfplot`, care reprezintă grafic funcţia de repartiţie de selecţie. Apelul funcţiei se poate face prin:

```
cdfplot(x)
h=cdfplot(x)
[h,stats]=cdfplot(x)
```

Prima formă reprezintă grafic funcţia de repartiţie de selecţie corespunzătoare datelor conţinute de vectorul `x`, caracteristici cum ar fi culoarea, tipul liniei etc. sunt considerate implicit de sistemul Matlab. Pentru modificarea unor astfel de caracteristici, adică alegerea acestora de către utilizator, se recomandă a doua formă, prin care parametrul `h` poate fi folosit în comanda `set`.

Parametrul `stats` conţine, după executarea funcţiei, o parte din parametrii statsitici pentru `x`. Aceştia sunt: valoarea minimă (`stats.min`), valoarea maximă (`stats.max`), valoarea mediei de selecţie (`stats.mean`), valoarea medianei (`stats.median`) şi abaterea standard (`stats.std`).

Funcţia 4.2.42. Funcţia Matlab care urmează generează `n` numere aleatoare, care urmează o lege de probabilitate precizată, după care reprezintă grafic, pe aceeaşi figură, funcţia de repartiţie teoretică prin linie-punct, împreună cu funcţia de repartiţie de selecţie.

```
function compar(n,lege)
% Functie  de repartitie  de selectie
% lege - legea de probabilitate
% n - volumul datelor
clf
switch lege
case 'unid'
  N=input('N=');
  X=unidrnd(N,n,1); [med,var]=unidstat(N);
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,N);
case 'bino'
  m=input('n='); p=input('p=');
  X=binornd(m,p,n,1); [med,var]=binostat(m,p);
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,m,p);
case 'hyge'
  M=input('M='); K=input('K='); m=input('n=');
  X=hygernd(M,K,m,n,1); [med,var]=hygestat(M,K,m);
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,M,K,m);
case 'poiss'
  la=input('lambda=');
  X=poissrnd(la,n,1); [med,var]=poisstat(la);
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,la);
case 'nbin'
  r=input('r='); p=input('p=');
  X=nbinrnd(r,p,n,1); [med,var]=nbinstat(r,p);
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,r,p);
case 'geo'
  p=input('p=');
  X=geornd(p,n,1); [med,var]=geostat(p);
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,p);
case 'unif'
  a=input('a:'); b=input('b(a<b):');
  X=unifrnd(a,b,n,1); [med,var]=unifstat(a,b);
```

```
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,a,b);
 case 'norm'
  mu=input('mu='); s=input('sigma=');
  X=normrnd(mu,s,n,1); [med,var]=normstat(mu,s);
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,mu,s);
case 'logn'
  mu=input('mu='); s=input('sigma=');
  X=lognrnd(mu,s,n,1); [med,var]=lognstat(mu,s);
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,mu,s);
case 'gam'
  a=input('a='); b=input('b=');
  X=gamrnd(a,b,n,1); [med,var]=gamstat(a,b);
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,a,b);
case 'exp'
  mu=input('mu=');
  X=exprnd(mu,n,1); [med,var]=expstat(mu);
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,mu);
case 'beta'
  a=input('a='); b=input('b=');
  X=betarnd(a,b,n,1); [med,var]=betastat(a,b);
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,a,b);
case 'weib'
  a=input('a='); b=input('b=');
  X=weibrnd(a,b,n,1); [med,var]=weibstat(a,b);
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,a,b);
case 'rayl'
  b=input('b=');
  X=raylrnd(b,n,1); [med,var]=raylstat(b);
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,b);
case 't'
  m=input('n=');
  X=trnd(m,n,1); [med,var]=tstat(m);
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,m);
case 'chi2'
  m=input('n=');
  X=chi2rnd(m,n,1); [med,var]=chi2stat(m);
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,m);
case 'f'
  m1=input('m='); m2=input('n=');
```

```
  X=frnd(m1,m2,n,1); [med,var]=fstat(m1,m2);
  x=med-3*sqrt(var):0.01:med+3*sqrt(var);
  f=cdf(lege,x,m1,m2);
otherwise
  error('Lege necunoscuta')
end
plot(x,f,'k-.'), hold on, cdfplot(X), grid off
title('Functia de repartitie de selectie')
```

Apelul funcţiei `compar` se face prin

```
>>compar(n,'lege')
```

unde valorile pentru `n` şi `lege`, fie că sunt precizate în acest apel, fie sunt precizate înainte. De exemplu, comanda

```
>>compar(25,'bino')
```

are ca efect apelul funcţiei pentru legea binomială, iar pe ecran se va cere introducerea parametrilor `n` şi `p`, după care pe ecran va fi reprezentat graficul din Figura 4.6, în cazul în care `n=7` şi `p=0.4`.

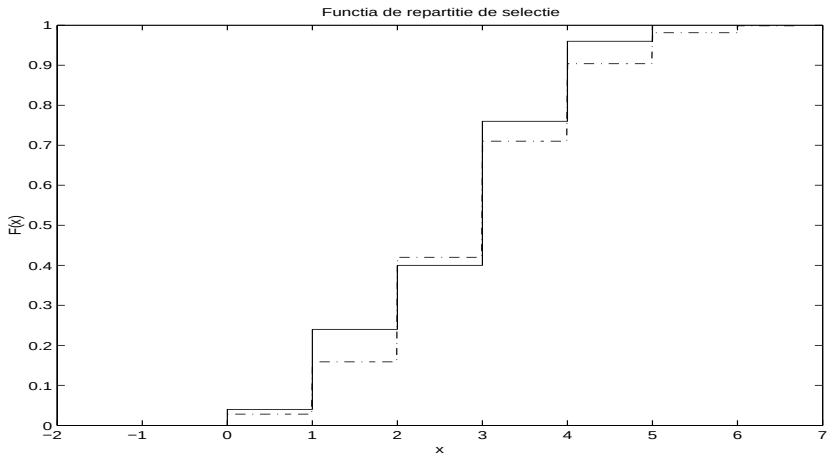

Figura 4.6: Legea $\mathcal{B}(7, 0.4)$

Comanda

```
>>compar(15,'norm')
```

are ca efect apelul funcţiei pentru legea normală, iar pe ecran se va cere introducerea parametrilor `mu` şi `sigma`, după care pe ecran va fi reprezentat graficul din Figura 4.7, în cazul în care `mu=5` şi `sigma=2`.

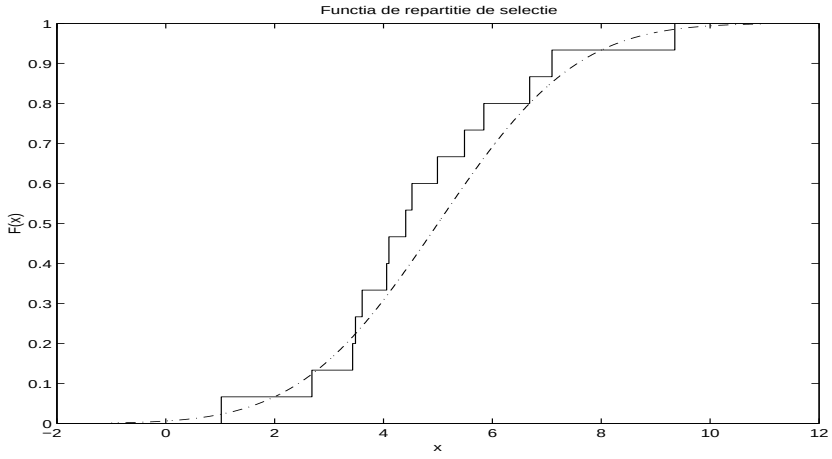

Figura 4.7: Legea $\mathcal{N}(5, 2)$

Capitolul 5

Teoria estimaţiei

Relativ la colectivitatea $\mathcal{C}$ este cercetată caracteristica X, care urmează legea de probabilitate dată prin funcţia de probabilitate $f(x;\theta)$, ce reprezintă funcţia de frecvenţă dacă X este de tip discret, respectiv densitatea de probabilitate în cazul continuu, iar θ este un parametru real necunoscut. Se consideră o selecţie repetată de volum n având corespunzător variabilele de selecţie $X_1, X_2, \ldots, X_n$.

5.1 Funcţie de verosimilitate

Definiţia 5.1.1. *Numim* funcţie de verosimilitate, *funcţia de selecţie*

$$g(X_1, X_2, \ldots, X_n; \theta) = \prod_{k=1}^{n} f(X_k; \theta).$$

Observaţia 5.1.2. Dacă se consideră funcţia de n variabile reale

$$g(x_1, x_2, \ldots, x_n; \theta) = \prod_{k=1}^{n} f(x_k; \theta),$$

aceasta reprezintă funcţia frecvenţă (în cazul discret), respectiv densitatea de probabilitate (în cazul continuu) a vectorului aleator $(X_1, X_2, \ldots, X_n)$.

Definiţia 5.1.3. *Statistica $S = S(X_1, X_2, \ldots, X_n)$ este* statistică suficientă (exhaustivă) *pentru parametrul θ, dacă există funcţia măsurabilă $\varphi : \mathbb{R}^n \to \mathbb{R}$ nenegativă şi funcţia măsurabilă $h_\theta : \mathbb{R} \to \mathbb{R}$ nenegativă, astfel încât*

$$g(x_1, x_2, \ldots, x_n; \theta) = \varphi(x_1, x_2, \ldots, x_n)\, h_\theta(s) = \varphi(x_1, x_2, \ldots, x_n)\, h(s; \theta),$$

unde $s = S(x_1, x_2, \ldots, x_n)$.

Observaţia 5.1.4. Echivalentă cu condiţia precedentă pentru suficienţa statisticii S este condiţia ca funcţia de frecvenţă (în cazul discret), respectiv densitatea de probabilitate (în cazul continuu) $f(x_1, \ldots, x_n; \theta \mid s)$ a vectorului aleator $(X_1, \ldots, X_n)$ condiţionată de evenimentul $S(X_1, \ldots, X_n) = s$, să nu depindă de parametrul θ.

Exemplul 5.1.5. Fie caracteristica X ce urmează legea lui Poisson cu parametrul $\lambda > 0$ necunoscut, adică are funcţia de frecvenţă

$$f(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}, \ x = 0, 1, 2, \ldots.$$

Considerăm statistica $S = \sum_{k=1}^{n} X_k$. Deoarece variabilele aleatoare $X_1, X_2, \ldots, X_n$ sunt independente şi identic repartizate cu X avem că S urmează legea lui Poisson de parametru $n\lambda$, adică

$$P(S = s) = \frac{(n\lambda)^s}{s!} e^{-n\lambda}, \quad s = 0, 1, 2, \ldots$$

Pe de o parte avem:

$$g(x_1, x_2, \ldots, x_n; \lambda) = \prod_{k=1}^{n} f(x_k; \theta) = \prod_{k=1}^{n} \left(\frac{\lambda^{x_k}}{x_k!} e^{-\lambda} \right) = \frac{\lambda^{\sum_{k=1}^{n} x_k}}{\prod_{k=1}^{n} x_k!} \mathrm{e}^{-n\lambda}$$
$$= \frac{\lambda^s}{x_1! x_2! \ldots x_n!} \mathrm{e}^{-n\lambda}.$$

Se poate considera

$$\varphi(x_1, x_2, \ldots, x_n) = \frac{1}{x_1! x_2! \ldots x_n!}, \quad h(s; \lambda) = \lambda^s \mathrm{e}^{-n\lambda}.$$

Având în vedere că putem scrie şi

$$g(x_1, x_2, \ldots, x_n; \lambda) = \frac{(n\lambda)^s}{s!} \mathrm{e}^{-n\lambda} \cdot \frac{s!}{n^s x_1! x_2! \ldots x_n!},$$

se poate considera de asemenea

$$\varphi(x_1, x_2, \ldots, x_n) = \frac{s!}{n^s x_1! x_2! \ldots x_n!}, \quad h(s; \lambda) = \frac{(n\lambda)^s}{s!} \mathrm{e}^{-n\lambda}.$$

Prin urmare, unicitatea acestei scrieri nu se impune.

Pe de altă parte, având în vedere că S urmează legea $\mathcal{P}(n\lambda)$, avem succesiv

$$f(x_1, x_2, \ldots, x_n; \lambda \mid s) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid S = s)$$

$$
\begin{aligned}
&= \frac{P\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, S = s\right)}{P\left(S = s\right)} \\
&= \frac{P\left(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = s - \sum_{k=1}^{n-1} x_k\right)}{P\left(S = s\right)} \\
&= \frac{P\left(X_n = s - \sum_{k=1}^{n-1} x_k\right) \prod_{i=1}^{n-1} P\left(X_i = x_i\right)}{P\left(S = s\right)} \\
&= \frac{\frac{\lambda^{s-\sum_{k=1}^{n-1} x_k}}{\left(s-\sum_{k=1}^{n-1} x_k\right)!} e^{-\lambda} \prod_{i=1}^{n-1} \left(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda}\right)}{\frac{(n\lambda)^s}{s!} e^{-n\lambda}} = \frac{s!}{n^s x_1! x_2! \ldots x_n!}.
\end{aligned}
$$

Aşadar $f\left(x_1, x_2, \ldots, x_n; \lambda \,|\, s\right)$ nu depinde de parametrul λ, decât prin intermediul valorii statisticii S.

Exemplul 5.1.6 (Familia exponenţială). Se consideră caracteristica X, care are funcţia de probabilitate de forma

$$f\left(x; \theta\right) = \exp\Big\{a\left(x\right)\alpha\left(\theta\right) + b\left(x\right) + \beta\left(\theta\right)\Big\}.$$

Vom arăta că, în acest caz, statistica

$$S = S\left(X_1, X_2, \ldots, X_n\right) = \sum_{k=1}^{n} a\left(X_k\right)$$

este suficientă pentru parametrul θ. Într-adevăr avem că

$$
\begin{aligned}
g\left(x_1, x_2, \ldots, x_n; \theta\right) &= \prod_{k=1}^{n} f\left(x_k; \theta\right) \\
&= \exp\left\{\alpha\left(\theta\right) \sum_{k=1}^{n} a\left(x_k\right) + n\beta\left(\theta\right) + \sum_{k=1}^{n} b\left(x_k\right)\right\} \\
&= \exp\left\{\alpha\left(\theta\right) \sum_{k=1}^{n} a\left(x_k\right) + n\beta\left(\theta\right)\right\} \exp\left\{\sum_{k=1}^{n} b\left(x_k\right)\right\}.
\end{aligned}
$$

Dacă se consideră

$$\varphi\left(x_1, x_2, \ldots, x_n\right) = \exp\left\{\sum_{k=1}^{n} b\left(x_k\right)\right\} \quad \text{şi} \quad h\left(s; \theta\right) = \exp\left\{\alpha\left(\theta\right) s + n\beta\left(\theta\right)\right\},$$

atunci

$$g\left(x_1, x_2, \ldots, x_n; \theta\right) = \varphi\left(x_1, x_2, \ldots, x_n\right) h\left(s; \theta\right),$$

deci condiţia suficienţei este îndeplinită.

Definiţia 5.1.7. *Numim* cantitate de informaţie (Fisher) *a unei selecţii de volum* n *relativă la parametrul* $\theta \in \mathbb{R}$ *necunoscut, valoarea medie*

$$I_n(\theta) = E\left[\left(\frac{\partial \ln g(X_1, X_2, \ldots, X_n; \theta)}{\partial \theta}\right)^2\right],$$

când funcţia de verosimilitate este derivabilă în raport cu θ*.*

Observaţia 5.1.8. Dacă parametrul $\boldsymbol{\theta}$ este p–dimensional, matricea

$$\boldsymbol{F} = \left(Cov\left(\frac{\partial \ln g(X_1, \ldots, X_n; \boldsymbol{\theta})}{\partial \theta_i}, \frac{\partial \ln g(X_1, \ldots, X_n; \boldsymbol{\theta})}{\partial \theta_j}\right)\right)_{i,j=\overline{1,p}}$$

se numeşte *matricea informaţiei lui Fisher*.

Teorema 5.1.9. *Dacă domeniul valorilor caracteristicii* X *nu depinde de parametrul* θ*, iar funcţia de verosimilitate este derivabilă de două ori în raport cu* θ*, atunci*

$$I_n(\theta) = -E\left(\frac{\partial^2 \ln g(X_1, X_2, \ldots, X_n; \theta)}{\partial \theta^2}\right).$$

Demonstraţie. Se porneşte de la relaţia cunoscută

$$\int \ldots \int_{\mathbb{R}^n} g(x_1, x_2, \ldots, x_n; \theta)\, dx_1 dx_2 \ldots dx_n = 1,$$

pe care o satisface densitatea de probabilitate. Se derivează această relaţie în raport cu θ şi se ţine seama de faptul că

$$\frac{\partial g(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta} = g(x_1, x_2, \ldots, x_n) \frac{\partial \ln g(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta},$$

obţinându-se

$$\int \ldots \int_{\mathbb{R}^n} \frac{\partial \ln g(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta} g(x_1, x_2, \ldots, x_n; \theta)\, dx_1 dx_2 \ldots dx_n = 0.$$

Derivând încă odată în raport cu θ rezultă

$$\int \ldots \int_{\mathbb{R}^n} \frac{\partial^2 \ln g(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta^2} g(x_1, x_2, \ldots, x_n; \theta)\, dx_1 dx_2 \ldots dx_n$$
$$+ \int \ldots \int_{\mathbb{R}^n} \frac{\partial \ln g(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta} \frac{\partial g(x_1, x_2, \ldots, x_n; \theta)}{\partial \theta} dx_1 dx_2 \ldots dx_n = 0,$$

sau, având în vedere relaţia de mai înainte, rezultă că

$$\int\ldots\int_{\mathbb{R}^n}\frac{\partial^2 \ln g\left(x_1,x_2,\ldots,x_n;\theta\right)}{\partial\theta^2}g\left(x_1,x_2,\ldots,x_n;\theta\right)dx_1dx_2\ldots dx_n$$
$$+\int\ldots\int_{\mathbb{R}^n}\left[\frac{\partial \ln g\left(x_1,x_2,\ldots,x_n;\theta\right)}{\partial\theta}\right]^2 g\left(x_1,x_2,\ldots,x_n;\theta\right)dx_1dx_2\ldots dx_n=0.$$

Aşadar am obţinut că

$$E\left[\frac{\partial^2 \ln g\left(X_1,X_2,\ldots,X_n;\theta\right)}{\partial\theta^2}\right]+E\left[\left(\frac{\partial \ln g\left(X_1,X_2,\ldots,X_n;\theta\right)}{\partial\theta}\right)^2\right]=0,$$

de unde avem relaţia dorită. □

Observaţia 5.1.10. În demonstraţia teoremei precedente am considerat cazul unei caracteristici X de tip continuu. În mod analog se procedează şi în cazul discret, integrala multiplă este înlocuită cu o sumă multiplă.

Corolarul 5.1.11. *Când domeniul de definiţie al caracteristicii X nu depinde de parametrul θ, atunci $I_n\left(\theta\right)=nI_1\left(\theta\right)$.*

Demonstraţie. Deoarece selecţia este repetată, avem că

$$\frac{\partial^2 \ln g\left(x_1,x_2,\ldots,x_n;\theta\right)}{\partial\theta^2}=\sum_{k=1}^{n}\frac{\partial^2 \ln f\left(x_k;\theta\right)}{\partial\theta^2}.$$

Folosind Teorema 5.1.9, se obţine

$$I_n\left(\theta\right)=-\sum_{k=1}^{n}E\left[\frac{\partial^2 \ln f\left(X_k;\theta\right)}{\partial\theta^2}\right]=\sum_{k=1}^{n}I_1\left(\theta\right)=nI_1\left(\theta\right),$$

deoarece

$$I_1\left(\theta\right)=-E\left[\frac{\partial^2 \ln f\left(X;\theta\right)}{\partial\theta^2}\right].$$

□

Observaţia 5.1.12. Din demonstraţia Teoremei 5.1.9 avem de asemenea că

$$I_n\left(\theta\right)=Var\left[\frac{\partial \ln g\left(X_1,X_2,\ldots,X_n;\theta\right)}{\partial\theta}\right],$$

deoarece

$$E\left[\frac{\partial \ln g\left(X_1,X_2,\ldots,X_n;\theta\right)}{\partial\theta}\right]=0.$$

Exemplul 5.1.13. Să considerăm caracteristica X, ce urmează legea normală $\mathcal{N}(\mu, \sigma)$, unde $\mu \in \mathbb{R}$ este necunoscut, iar $\sigma > 0$ este cunoscut.

Deoarece

$$f(x;\mu) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R},$$

avem că

$$I_1(\mu) = E\left[\left(\frac{\partial \ln f(X;\mu)}{\partial \mu}\right)^2\right] = E\left[\frac{(X-\mu)^2}{\sigma^4}\right] = \frac{1}{\sigma^2}.$$

Prin urmare, cantitatea de informaţie conţinută (adusă) de o observaţie, relativ la parametrul μ, este cu atât mai mare cu cât dispersia este mai mică.

Teorema 5.1.14. *Fie caracteristica X, cu funcţia de probabilitate $f(x;\theta)$ derivabilă de două ori în raport cu θ şi statistica $S = S(X_1, X_2, \ldots, X_n)$ relativă la caracteristica X, cu funcţia de probabilitate $h(s;\theta)$, atunci cantitatea de informaţie*

$$I_S(\theta) = -E\left[\frac{\partial^2 \ln h(S;\theta)}{\partial \theta^2}\right],$$

relativă la parametrul θ conţinută în statistica S nu depăşeşte cantitatea de informaţie $I_n(\theta)$ conţinută de selecţia considerată, adică $I_S(\theta) \leqslant I_n(\theta)$.

Demonstraţie. Dacă $h(s;\theta)$ este funcţia de frecvenţă (în cazul discret), respectiv densitatea de probabilitate (în cazul continuu) pentru statistica S, atunci

$$g(x_1, x_2, \ldots, x_n; \theta) = h(s;\theta)\, f(x_1, x_2, \ldots, x_n; \theta \mid s),$$

unde $f(x_1, x_2, \ldots, x_n; \theta \mid s)$ este funcţia de frecvenţă (în cazul discret), respectiv densitatea de probabilitate (în cazul continuu) pentru vectorul aleator $(X_1, \ldots, X_n)$ condiţionată de evenimentul $(S = s)$.

Aşadar avem că

$$\frac{\partial^2 \ln g(x_1, x_2, \ldots, x_n;\theta)}{\partial \theta^2} = \frac{\partial^2 \ln h(s;\theta)}{\partial \theta^2} + \frac{\partial^2 \ln f(x_1, x_2, \ldots, x_n; \theta \mid s)}{\partial \theta^2}, \tag{5.1.1}$$

deci

$$I_n(\theta) = I_S(\theta) - E\left[\frac{\partial^2 \ln f(X_1, X_2, \ldots, X_n; \theta \mid s)}{\partial \theta^2}\right] \geqslant I_S(\theta),$$

ceea ce trebuia arătat. □

Exemplul 5.1.15. Se consideră caracteristica X ce urmează legea normală $\mathcal{N}(\mu, \sigma)$ şi o selecţie repetată de volum n. Vom compara informaţia dispersiei de selecţie

$$\bar{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2,$$

relativă la dispersia teoretică $Var(X) = \sigma^2$, cu informaţia selecţiei relativă la acelaşi parametru.

Calculăm prima dată informaţia selecţiei de volum n relativă la parametrul notat cu $\theta = \sigma^2 = Var\,(X)$. Având în vedere că densitatea de probabilitate a caracteristicii X este

$$f\,(x;\theta) = \frac{1}{\sqrt{2\pi\theta}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\theta}}, \quad x \in \mathbb{R},$$

se obţine

$$\frac{\partial \ln f\,(x;\theta)}{\partial\,\theta} = -\frac{1}{2\theta} + \frac{1}{2\theta^2}(x-\mu)^2 \quad \text{şi} \quad \frac{\partial^2 \ln f\,(x;\theta)}{\partial\,\theta^2} = \frac{1}{2\theta^2} - \frac{1}{\theta^3}(x-\mu)^2.$$

Rezultă astfel că

$$I_1\left(\sigma^2\right) = I_1(\theta) = -\frac{1}{2\sigma^4} + \frac{1}{\sigma^4}E\left[\left(\frac{X-\mu}{\sigma}\right)^2\right] = \frac{1}{2\sigma^4},$$

deci

$$I_n\left(\sigma^2\right) = \frac{n}{2\sigma^4}.$$

Pe de altă parte, se ştie că statistica

$$\chi^2 = \frac{n-1}{\sigma^2}\bar{\sigma}^2 = \frac{1}{\sigma^2}\sum_{k=}^{n}\left(\,X_k - \bar{X}\,\right)^2$$

urmează legea χ^2 cu $n-1$ grade de libertate. Având în vedere legătura dintre statisticile χ^2 şi $\bar{\sigma}^2$ rezultă că densitatea de probabilitate pentru $\bar{\sigma}^2$ este dată prin

$$h\,(s;\theta) = \frac{1}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}\left(\frac{n-1}{\theta}\right)^{\frac{n-1}{2}} s^{\frac{n-1}{2}-1}\mathrm{e}^{-\frac{(n-1)s}{2\theta}}, \quad s > 0.$$

Astfel se obţine că

$$\ln h\,(s;\theta) = \mathrm{C} + \frac{n-1}{2}\ln\frac{n-1}{\theta} + \frac{n-3}{2}\ln s - \frac{(n-1)s}{2\theta},$$

de unde

$$\frac{\partial \ln h\left(s;\theta\right)}{\partial\theta}=-\frac{n-1}{2\theta}+\frac{(n-1)s}{2\theta^2}\quad\text{şi}\quad\frac{\partial^2\ln h\left(s;\theta\right)}{\partial\theta^2}=\frac{n-1}{2\theta^2}-\frac{(n-1)s}{\theta^3}.$$

Calculăm acum informaţia dispersiei de selecţie relativă la σ^2. Pentru aceasta ţinem seama de faptul că

$$E\left(\bar{\sigma}^2\right)=\frac{\sigma^2}{n-1}E\left(\chi^2\right)=\frac{\sigma^2}{n-1}\left(n-1\right)=\sigma^2.$$

Astfel obţinem:

$$I_{\bar{\sigma}^2}\left(\sigma^2\right)=-\frac{n-1}{2\sigma^4}+\frac{n-1}{\sigma^6}E\left(\bar{\sigma}^2\right)=-\frac{n-1}{2\sigma^4}+\frac{n-1}{\sigma^4},$$

deci

$$I_{\bar{\sigma}^2}\left(\sigma^2\right)=\frac{n-1}{2\sigma^4}<\frac{n}{2\sigma^4}=I_n\left(\sigma^2\right).$$

Observaţia 5.1.16. Deoarece, în cazul unei statistici S suficiente, avem că funcţia $f\left(x_1,x_2,\ldots x_n;\theta\mid s\right)$ nu depinde de θ rezultă că $I_n\left(\theta\right)=I_S\left(\theta\right)$ în acest caz. Acest lucru se obţine imediat, dacă se are în vedere relaţia (5.1.1).

5.2 Funcţii de estimaţie

Definiţia 5.2.1. *Fie caracteristica X cu funcţia de probabilitate $f\left(x;\theta\right)$, parametrul $\theta\in A$ necunoscut şi o selecţie repetată de volum n. Numim* funcţie de estimaţie (estimator) *pentru parametrul θ, funcţia de selecţie*

$$\widehat{\Theta}=\widehat{\theta}\left(X_1,X_2,\ldots,X_n\right),$$

care ia valori în domeniul A, iar valoarea numerică $\widehat{\theta}=\widehat{\theta}\left(x_1,x_2,\ldots,x_n\right)$ se numeşte estimaţia *lui θ.*

Definiţia 5.2.2. *Estimatorul $\widehat{\Theta}=\widehat{\theta}\left(X_1,X_2,\ldots,X_n\right)$ este* estimator (funcţie de estimaţie) nedeplasat *pentru parametrul necunoscut θ, dacă $E\left(\widehat{\Theta}\right)=\theta$, iar valoarea numerică $\widehat{\theta}=\widehat{\theta}\left(x_1,x_2,\ldots,x_n\right)$ se numeşte* estimaţie nedeplasată *pentru parametrul θ.*

Definiţia 5.2.3. *Spunem că estimatorul $\widehat{\Theta}=\widehat{\theta}\left(X_1,X_2,\ldots,X_n\right)$ este* estimator consistent*, pentru parametrul necunoscut θ, dacă $\widehat{\Theta}\xrightarrow{p}\theta$, adică*

$$\lim_{n\to\infty}P\left(\left|\widehat{\Theta}-\theta\right|<\varepsilon\right)=1,$$

pentru orice $\varepsilon>0$, iar valoarea numerică $\widehat{\theta}=\widehat{\theta}\left(x_1,x_2,\ldots,x_n\right)$ se numeşte estimaţie consistentă *pentru parametrul θ.*

5.2.1 Funcţii de estimaţie absolut corecte

Definiţia 5.2.4. *Numim* funcţie de estimaţie (estimator) absolut corectă *pentru parametrul θ, funcţia de selecţie $\widehat{\Theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$, care satisface condiţiile*

(i) $E\left(\widehat{\Theta}\right) = \theta,$

(ii) $\lim\limits_{n\to\infty} Var\left(\widehat{\Theta}\right) = 0,$

iar valoarea numerică $\widehat{\theta} = \widehat{\theta}(x_1, x_2, \ldots, x_n)$ se numeşte estimaţie absolut corectă *pentru parametrul θ.*

Proprietatea 5.2.5. *Un estimator absolut corect este un estimator consistent.*

Demonstraţie. Fie estimatorul $\widehat{\Theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$, un estimator absolut corect pentru parametrul θ. Din inegalitatea lui Cebîşev avem

$$1 \geqslant P\left(\left|\widehat{\Theta} - \theta\right| < \varepsilon\right) \geqslant 1 - \frac{Var\left(\widehat{\Theta}\right)}{\varepsilon^2},$$

pentru orice $\varepsilon > 0$. Făcând pe $n \to \infty$, se obţine

$$\lim_{n\to\infty} P\left(\left|\widehat{\Theta} - \theta\right| < \varepsilon\right) = 1,$$

pentru orice $\varepsilon > 0$, ceea ce trebuie demonstrat. □

Exemplul 5.2.6. Fie caracteristica X, ce urmează legea normală $\mathcal{N}(\mu, \sigma)$, unde parametrul $\mu \in \mathbb{R}$ este cunoscut, iar $\sigma > 0$ este necunoscut. Considerând o selecţie repetată de volum n, vom determina constanta κ_n astfel încât statistica

$$\bar{S} = \kappa_n \sum_{k=1}^{n} |X_k - \mu|$$

să fie estimator absolut corect pentru abaterea standard σ.

Având în vedere că variabilele de selecţie $X_1, X_2, \ldots, X_n$ sunt variabile aleatoare identic repartizate cu X, putem scrie succesiv

$$\begin{aligned} E\left(\bar{S}\right) &= \kappa_n \sum_{k=1}^{n} E\left(|X_k - \mu|\right) = \kappa_n \sum_{k=1}^{n} E\left(|X - \mu|\right) \\ &= n\kappa_n E\left(|X - \mu|\right). \end{aligned} \tag{5.2.1}$$

Ţinem seama de faptul că X urmează legea normală $\mathcal{N}(\mu,\sigma)$ şi obţinem

$$E(|X-\mu|)=\int_{-\infty}^{+\infty}|x-\mu|\,f(x;\sigma)\,dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}|x-\mu|\,\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}dx.$$

Facem schimbarea de variabilă $x-\mu=t$ şi ţinem seama de faptul că funcţia obţinută este pară, iar intervalul de integrare este simetric faţă de origine:

$$\begin{aligned}E(|X-\mu|)=&\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty}|t|\,e^{-\frac{t^2}{2\sigma^2}}dt=\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\int_0^{+\infty}te^{-\frac{t^2}{2\sigma^2}}dt\\=&\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\left(-\sigma^2e^{-\frac{t^2}{2\sigma^2}}\right)\Big|_0^{+\infty}=\sigma\sqrt{\frac{2}{\pi}}.\end{aligned}$$

Din condiţia de nedeplasare şi din formula (5.2.1) avem

$$\sigma=E\left(\bar{S}\right)=n\kappa_n\sigma\sqrt{\frac{2}{\pi}},$$

adică $\kappa_n=\frac{1}{n}\sqrt{\frac{\pi}{2}}$.

Până aici am arătat că statistica

$$\bar{S}=\frac{1}{n}\sqrt{\frac{\pi}{2}}\sum_{k=1}^{n}|X_k-\mu|$$

este un estimator nedeplasat pentru parametrul σ.

Deoarece selecţia considerată este repetată rezultă că variabilele de selecţie sunt independente, prin urmare se poate scrie

$$Var\left(\bar{S}\right)=\frac{\pi}{2n^2}\sum_{k=1}^{n}Var\left(|X_k-\mu|\right)=\frac{\pi}{2n^2}\sum_{k=1}^{n}Var\left(|X-\mu|\right)=\frac{\pi}{2n}Var\left(|X-\mu|\right).$$

Astfel este îndeplinită şi condiţia (ii) din Definiţia 5.2.4, anume

$$Var\left(\bar{S}\right)=\frac{\pi}{2n}Var\left(|X-\mu|\right)\longrightarrow 0,\quad\text{când}\quad n\to\infty.$$

Remarcăm că

$$Var\left(|X-\mu|\right)=E\left[(X-\mu)^2\right]-\left[E\left(|X-\mu|\right)\right]^2=\sigma^2-\sigma^2\frac{2}{\pi}=\frac{\pi-2}{\pi}\sigma^2.$$

Proprietatea 5.2.7. *Fie caracteristica X pentru care există momentul teoretic de ordinul $2k$, $\nu_{2k}=E\left(X^{2k}\right)$, şi fie o selecţie repetată de volum n, atunci momentul de selecţie de ordin k*

$$\bar{\nu}_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k$$

este funcţie de estimaţie absolut corectă pentru parametrul ν_k.

Demonstraţie. Din Proprietatea 4.2.11 avem că $E(\bar{\nu}_k) = \nu_k$ şi de asemenea

$$\lim_{n\to\infty} Var(\bar{\nu}_k) = \lim_{n\to\infty} \frac{\nu_{2k} - \nu_k^2}{n} = 0.$$

Condiţiile Definiţiei 5.2.4 sunt satisfăcute, deci proprietatea este demonstrată. □

Observaţia 5.2.8. Media de selecţie $\bar{X}$ $(= \bar{\nu}_1)$ este funcţie de estimaţie absolut corectă pentru media teoretică $E(X)$ $(= \nu_1)$.

5.2.2 Funcţii de estimaţie corecte

Definiţia 5.2.9. *Numim* funcţie de estimaţie (estimator) corectă, *pentru parametrul necunoscut θ, funcţia de selecţie $\widehat{\Theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$, care satisface condiţiile*

(i) $\lim_{n\to\infty} E\left(\widehat{\Theta}\right) = \theta,$

(ii) $\lim_{n\to\infty} Var\left(\widehat{\Theta}\right) = 0,$

iar valoarea numerică $\widehat{\theta} = \widehat{\theta}(x_1, x_2, \ldots, x_n)$ se numeşte estimaţie corectă *pentru parametrul θ.*

Proprietatea 5.2.10. *Un estimator corect este un estimator consistent.*

Demonstraţie. Fie estimatorul $\widehat{\Theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$ corect pentru parametrul θ, atunci din condiţiile (i) şi (ii) avem că pentru orice $\varepsilon > 0$ şi $\delta > 0$ există numărul natural $N = N(\varepsilon, \delta)$ astfel încât

$$\left| E\left(\widehat{\Theta}\right) - \theta \right| < \frac{\varepsilon}{2} \quad \text{şi} \quad Var\left(\widehat{\Theta}\right) < \frac{\varepsilon^2 \delta}{4},$$

pentru $n > N$. Aşadar, putem scrie

$$\left| \widehat{\Theta} - \theta \right| \leqslant \left| \widehat{\Theta} - E\left(\widehat{\Theta}\right) \right| + \left| E\left(\widehat{\Theta}\right) - \theta \right| < \left| \widehat{\Theta} - E\left(\widehat{\Theta}\right) \right| + \frac{\varepsilon}{2},$$

pentru $n > N$, de unde dacă $\left| \widehat{\Theta} - E\left(\widehat{\Theta}\right) \right| < \frac{\varepsilon}{2}$, vom avea că $\left| \widehat{\Theta} - \theta \right| < \varepsilon$, pentru $n > N$. Prin urmare avem

$$\left(\left| \widehat{\Theta} - E\left(\widehat{\Theta}\right) \right| < \frac{\varepsilon}{2} \right) \subset \left(\left| \widehat{\Theta} - \theta \right| < \varepsilon \right), \quad n > N,$$

care conduce la inegalitatea

$$P\left(\left| \widehat{\Theta} - E\left(\widehat{\Theta}\right) \right| < \frac{\varepsilon}{2} \right) \leqslant P\left(\left| \widehat{\Theta} - \theta \right| < \varepsilon \right), \quad n > N.$$

Pe de altă parte, folosind inegalitatea lui Cebîşev,

$$P\left(\left|\widehat{\Theta}-E\left(\widehat{\Theta}\right)\right|<\frac{\varepsilon}{2}\right)\geqslant 1-\frac{4Var\left(\widehat{\Theta}\right)}{\varepsilon^2}.$$

Deoarece $Var\left(\widehat{\Theta}\right)<\frac{\varepsilon^2\delta}{4}$, pentru $n>N$, rezultă că

$$P\left(\left|\widehat{\Theta}-E\left(\widehat{\Theta}\right)\right|<\frac{\varepsilon}{2}\right)\geqslant 1-\delta,\quad n>N.$$

Prin urmare se ajunge la

$$P\left(\left|\widehat{\Theta}-\theta\right|<\varepsilon\right)\geqslant P\left(\left|\widehat{\Theta}-E\left(\widehat{\Theta}\right)\right|<\frac{\varepsilon}{2}\right)\geqslant 1-\delta,\quad n>N.$$

de unde $\widehat{\Theta}\xrightarrow{\text{p}}\theta$, ceea ce trebuie arătat. □

Proprietatea 5.2.11. *Fie caracteristica X pentru care există momentul teoretic de ordin $2k$, $\nu_{2k}=E\left(X^{2k}\right)$, şi fie o selecţie repetată de volum n, atunci momentul centrat de selecţie de ordin k*

$$\bar{\mu}_k=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^k$$

este funcţie de estimaţie corectă pentru momentul centrat teoretic de ordin k, adică pentru $\mu_k=E\left[\left(X-E\left(X\right)\right)^k\right]$.

Demonstraţie. Conform Observaţiei 4.2.18 avem

$$\lim_{n\to\infty}E\left(\bar{\mu}_k\right)=\lim_{n\to\infty}\left[\mu_k+O\left(\frac{1}{n}\right)\right]=\mu_k.$$

De asemenea, avem

$$\lim_{n\to\infty}Var\left(\bar{\mu}_k\right)=\lim_{n\to\infty}\left[\frac{\mu_{2k}-2k\mu_{k-1}\mu_{k+1}-\mu_k^2+k^2\mu_k\mu_{k-1}^2}{n}+O\left(\frac{1}{n^2}\right)\right]=0.$$

Aşadar, condiţiile Definiţiei 5.2.9 sunt satisfăcute. □

Observaţia 5.2.12. Momentul centrat de selecţie de ordinul 2, $\bar{\mu}_2$, este funcţie de estimaţie corectă pentru dispersia teoretică $Var\left(X\right)\ \left(=\mu_2\right)$.

Având în vedere Observaţia 4.2.17 (formula (4.2.2)) avem că dispersia de selecţie

$$\bar{\sigma}^2=\frac{1}{n-1}\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2$$

este funcţie de estimaţie absolut corectă pentru dispersia teoretică $Var\left(X\right)\ \left(=\mu_2\right)$.

5.2.3 Funcţii de estimaţie eficiente

Teorema 5.2.13 (Inegalitatea Rao-Cramer). *Se consideră caracteristica X având funcţia de probabilitate $f(x;\theta)$, $\theta \in (a,b)$, pentru care există derivata parţială de ordinul întâi în raport cu θ şi fie o statistică $\widehat{\Theta} = \widehat{\theta}(X_1,\ldots,X_n)$ absolut corectă pentru parametrul θ, atunci*

$$Var\left(\widehat{\Theta}\right) \geqslant \frac{1}{I_n(\theta)}.$$

Demonstraţie. Deoarece estimatorul $\widehat{\Theta}$ este nedeplasat, avem că $E\left(\widehat{\Theta}\right) = \theta$, adică

$$\int\ldots\int_{\mathbb{R}^n} \widehat{\theta}(x_1,\ldots,x_n)\, g(x_1,\ldots,x_n;\theta)\, dx_1\ldots dx_n = \theta,$$

unde $g(x_1,\ldots,x_n;\theta) = \prod_{k=1}^{n} f(x_k;\theta)$ este funcţia de verosimilitate.

Prin derivarea în raport cu θ a acestei relaţii se obţine că

$$\int\ldots\int_{\mathbb{R}^n} \widehat{\theta}(x_1,\ldots,x_n) \sum_{k=1}^{n} f(x_1;\theta)\cdots\frac{\partial f(x_k;\theta)}{\partial\theta}\cdots f(x_n;\theta)\, dx_1\ldots dx_n = 1$$

sau

$$\int\ldots\int_{\mathbb{R}^n} \widehat{\theta}(x_1,\ldots,x_n)\left[\sum_{k=1}^{n}\frac{\partial \ln f(x_k;\theta)}{\partial\theta}\right]\left[\prod_{i=1}^{n} f(x_i;\theta)\right] dx_1\ldots dx_n = 1.$$

Pe de altă parte, deoarece

$$\int_{\mathbb{R}} \frac{\partial \ln f(x;\theta)}{\partial\theta} f(x;\theta)\, dx = 0,$$

avem

$$\theta\sum_{k=1}^{n}\int\ldots\int_{\mathbb{R}^n}\frac{\partial \ln f(x_k;\theta)}{\partial\theta}\left[\prod_{i=1}^{n} f(x_i;\theta)\right] dx_1\ldots dx_n = 0,$$

care scăzută din egalitatea obţinută înainte conduce la

$$\int\ldots\int_{\mathbb{R}^n}\left[\widehat{\theta}(x_1,\ldots,x_n)-\theta\right]\left[\sum_{k=1}^{n}\frac{\partial \ln f(x_k;\theta)}{\partial\theta}\right] g(x_1,\ldots,x_n;\theta)\, dx_1\ldots dx_n = 1,$$

adică

$$E\left[\left(\widehat{\theta}(X_1,\ldots,X_n)-\theta\right)\left(\sum_{k=1}^{n}\frac{\partial \ln f(X_k;\theta)}{\partial\theta}\right)\right] = 1.$$

Se aplică acum inegalitatea lui Schwarz şi se obţine

$$\begin{aligned}1=&\left[E\left[\left(\widehat{\theta}\left(X_1,\ldots,X_n\right)-\theta\right)\left(\sum_{k=1}^{n}\frac{\partial \ln f\left(X_k;\theta\right)}{\partial\theta}\right)\right]\right]^2\\ \leqslant&E\left[\left(\widehat{\theta}\left(X_1,\ldots,X_n\right)-\theta\right)^2\right]E\left[\left(\sum_{k=1}^{n}\frac{\partial \ln f\left(X_k;\theta\right)}{\partial\theta}\right)^2\right]\\ =&Var\left(\widehat{\Theta}\right)Var\left[\sum_{k=1}^{n}\frac{\partial \ln f\left(X_k;\theta\right)}{\partial\theta}\right]=Var\left(\widehat{\Theta}\right)\cdot nVar\left[\frac{\partial \ln f\left(X;\theta\right)}{\partial\theta}\right],\end{aligned}$$

adică

$$Var\left(\widehat{\Theta}\right)\geqslant\frac{1}{nVar\left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right]}.$$

Dar avem că

$$I_n\left(\theta\right)=nI_1\left(\theta\right)=n\,E\left[\left(\frac{\partial \ln f\left(X;\theta\right)}{\partial\theta}\right)^2\right]=nVar\left[\frac{\partial \ln f\left(X;\theta\right)}{\partial\theta}\right],$$

care înlocuită în inegalitatea dinainte conduce la inegalitatea Rao–Cramer. □

Definiţia 5.2.14. *Se numeşte* eficienţă *a unei funcţii de estimaţie absolut corecte,* $\widehat{\Theta}=\widehat{\theta}\left(X_1,X_2,\ldots,X_n\right)$*, pentru parametrul* θ*, raportul*

$$\mathrm{e}\left(\widehat{\Theta}\right)=\frac{I_n^{-1}\left(\theta\right)}{Var\left(\widehat{\Theta}\right)}.$$

Definiţia 5.2.15. *Spunem că funcţia de estimaţie absolut corectă pentru parametrul* θ*,* $\widehat{\Theta}=\widehat{\theta}\left(X_1,X_2,\ldots,X_n\right)$*, este* eficientă *dacă inegalitatea Rao–Cramer este verificată prin egalitate, adică* $e\left(\widehat{\Theta}\right)=1$.

Exemplul 5.2.16. Se consideră caracteristica X ce urmează legea normală $\mathcal{N}\left(\mu,\sigma\right)$ şi o selecţie repetată de volum n relativă la această caracteristică. Vrem să determinăm eficienţa estimatorului lui $\sigma=\sqrt{Var\left(X\right)}$, construit în Exemplul 5.2.6, dat prin

$$\bar{S}=\frac{1}{n}\sqrt{\frac{\pi}{2}}\sum_{i=1}^{n}\left|\,X_i-\mu\,\right|.$$

Ştim că $E\left(\bar{S}\right)=\sigma$ şi $Var\left(\bar{S}\right)=\frac{\pi-2}{2n}\sigma^2$, deci $\bar{S}$ este estimator absolut corect pentru σ.

Pe de altă parte avem

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

de unde

$$\frac{\partial^2 \ln f(x;\sigma)}{\partial\sigma^2}=\frac{1}{\sigma^2}-\frac{3}{\sigma^4}(x-\mu)^2.$$

Se calculează uşor

$$I_n(\sigma)=nI_1(\sigma)=-nE\left[\frac{1}{\sigma^2}-\frac{3(X-\mu)^2}{\sigma^4}\right]=\frac{2n}{\sigma^2}\cdot$$

Avem astfel eficienţa estimatorului $\bar{S}$

$$\mathrm{e}\left(\bar{S}\right)=\frac{I_n^{-1}(\sigma)}{Var\left(\bar{S}\right)}=\frac{\sigma^2}{2n}\cdot\frac{2n}{(\pi-2)\sigma^2}=\frac{1}{\pi-2}<1,$$

adică estimatorul $\bar{S}$ nu este estimator eficient pentru parametrul σ.

Teorema 5.2.17 (Rao-Cramer). *Se consideră caracteristica X cu funcţia de probabilitate $f(x;\theta)$, $\theta\in(a,b)$, care satisface condiţiile Teoremei 5.2.13 şi fie funcţia de estimaţie absolut corectă $\widehat{\Theta}=\widehat{\theta}(X_1,X_2,\ldots,X_n)$ pentru parametrul θ. Condiţia necesară şi suficientă ca $\widehat{\Theta}$ să fie funcţie de estimaţie eficientă pentru parametrul θ este ca să aibă loc reprezentarea*

$$\ln f(x;\theta)=A'(\theta)[L(x)-\theta]+A(\theta)+N(x),$$

în plus are loc formula

$$\widehat{\Theta}=\widehat{\theta}(X_1,X_2,\ldots,X_n)=\frac{1}{n}\sum_{k=1}^{n}L(X_k).$$

Demonstraţie. Pentru necesitate, din demonstraţia inegalităţii Rao–Cramer, avem că egalitatea în această inegalitate are loc dacă inegalitatea lui Schwarz este verificată prin egalitate. Aceasta are loc dacă şi numai dacă variabilele aleatoare considerate depind în mod liniar, adică

$$K(\theta)\left[\widehat{\theta}(X_1,X_2,\ldots,X_n)-\theta\right]=\sum_{k=1}^{n}\frac{\partial \ln f(X_k;\theta)}{\partial\theta},\quad K\neq 0.$$

Considerând $X_1=x$ (x oarecare), $X_2=\cdots=X_n=$ const., (=0 de exemplu) avem

$$K(\theta)\left[\widehat{\theta}(x,0,\ldots,0)-\theta\right]=\frac{\partial \ln f(x;\theta)}{\partial\theta}+(n-1)\frac{\partial \ln f(0;\theta)}{\partial\theta},$$

de unde

$$\frac{\partial \ln f(x;\theta)}{\partial \theta} = U(\theta) Q(x) + V(\theta).$$

Astfel s-a obţinut că

$$\sum_{k=1}^{n} \frac{\partial \ln f(x_k;\theta)}{\partial \theta} = U(\theta) \sum_{k=1}^{n} Q(x_k) + nV(\theta)$$

şi prin urmare

$$K(\theta) \left[\widehat{\theta}(x_1, x_2, \ldots, x_n) - \theta\right] = U(\theta) \sum_{k=1}^{n} Q(x_k) + nV(\theta),$$

de unde

$$\widehat{\theta}(x_1, x_2, \ldots, x_n) = \frac{U(\theta)}{K(\theta)} \sum_{k=1}^{n} Q(x_k) + \frac{nV(\theta)}{K(\theta)} + \theta,$$

pentru orice $x_1, x_2, \ldots, x_n$. Rezultă de aici că

$$h = \frac{U(\theta)}{K(\theta)}, \quad g = \frac{nV(\theta)}{K(\theta)} + \theta$$

sunt constante (nu depind de θ), deoarece $\widehat{\theta}(x_1, x_2, \ldots, x_n)$ nu depinde de θ. Prin urmare

$$\widehat{\theta}(x_1, x_2, \ldots, x_n) = h \sum_{k=1}^{n} Q(x_k) + g,$$

şi dacă se notează $L(x) = nhQ(x) + g$, atunci

$$\widehat{\theta}(x_1, x_2, \ldots, x_n) = h \sum_{k=1}^{n} \frac{L(x_k) - g}{nh} + g = \frac{1}{n} \sum_{k=1}^{n} L(x_k).$$

Aşadar s-a ajuns la

$$\widehat{\Theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n) = \frac{1}{n} \sum_{k=1}^{n} L(X_k).$$

Pe de altă parte, deoarece

$$\frac{\partial \ln f(x;\theta)}{\partial \theta} = U(\theta) Q(x) + V(\theta),$$

obţinem

$$\frac{\partial \ln f(x;\theta)}{\partial \theta} = hK(\theta)\frac{L(x)-g}{nh} + \frac{K(\theta)(g-\theta)}{n},$$

de unde

$$\frac{\partial \ln f(x;\theta)}{\partial \theta} = \frac{K(\theta)}{n}[L(x)-\theta].$$

Integrăm, după ce am notat $A''(\theta) = \frac{K(\theta)}{n}$, ultima relaţie în raport cu θ şi obţinem

$$\int \frac{\partial \ln f(x;\theta)}{\partial \theta} d\theta = \int A''(\theta)[L(x)-\theta]\,d\theta = A'(\theta)[L(x)-\theta] + \int A'(\theta)\,d\theta$$
$$= A'(\theta)[L(x)-\theta] + A(\theta) + N(x),$$

de unde

$$\ln f(x;\theta) = A'(\theta)[L(x)-\theta] + A(\theta) + N(x).$$

Pentru suficienţă se porneşte de la relaţia cunoscută

$$\int_{\mathbb{R}} f(x;\theta)\,dx = 1,$$

adică

$$\int_{\mathbb{R}} \exp\left[A'(\theta)[L(x)-\theta] + A(\theta) + N(x)\right] dx = 1,$$

pe care o derivăm în raport cu θ. Obţinem astfel că

$$\int_{\mathbb{R}} A''(\theta)[L(x)-\theta] f(x;\theta)\,dx = 0,$$

de unde, deoarece $A'' \neq 0$, rezultă că

$$\int_{\mathbb{R}} [L(x)-\theta] f(x;\theta)\,dx = 0.$$

Derivăm din nou în raport cu θ şi se ajunge la

$$-\int_{\mathbb{R}} f(x;\theta)\,dx + \int_{\mathbb{R}} [L(x)-\theta]\frac{\partial f(x;\theta)}{\partial \theta} dx = 0,$$

de unde

$$\int_{\mathbb{R}} [L(x)-\theta]^2 A''(\theta) f(x;\theta)\,dx = 1.$$

Deoarece $E[L(X)] = \theta$, am obţinut astfel că

$$Var[L(X)] = \frac{1}{A''(\theta)},$$

deci

$$Var\left(\widehat{\Theta}\right) = Var\left[\frac{1}{n}\sum_{k=1}^{n} L\left(X_k\right)\right] = \frac{1}{nA''\left(\theta\right)}\cdot$$

Pe de altă parte

$$\begin{aligned} I_n\left(\theta\right) =& nI_1\left(\theta\right) = n\int_{\mathbb{R}}\left[\frac{\partial \ln f\left(x;\theta\right)}{\partial\theta}\right]^2 f\left(x;\theta\right)dx \\ =& n\int_{\mathbb{R}}\left[A''\left(\theta\right)\right]^2\left[L\left(x\right)-\theta\right]^2 f\left(x;\theta\right)dx = n\left[A''\left(\theta\right)\right]^2\frac{1}{A''\left(\theta\right)} = nA''\left(\theta\right). \end{aligned}$$

Prin urmare putem scrie

$$e\left(\widehat{\Theta}\right) = \frac{I_n^{-1}\left(\theta\right)}{Var\left(\widehat{\Theta}\right)} = \frac{\frac{1}{nA''(\theta)}}{\frac{1}{nA''(\theta)}} = 1,$$

deci $\widehat{\Theta} = \widehat{\theta}\left(X_1, X_2, \ldots, X_n\right)$ este estimator eficient pentru parametrul θ. □

Exemplul 5.2.18. Se consideră caracteristica X ce urmează legea $\mathcal{B}\left(m,p\right)$, cu m cunoscut şi p necunoscut. Vrem să arătăm că media de selecţie este estimator eficient pentru parametrul necunoscut $\theta = E\left(X\right) = mp$. Pentru aceasta vom considera o selecţie repetată de volum n relativă la această caracteristică.

Funcţia de frecvenţă a caracteristicii X este

$$f\left(x;\theta\right) = \binom{m}{x}\left(\frac{\theta}{m}\right)^x\left(1-\frac{\theta}{m}\right)^{m-x},$$

de unde

$$\begin{aligned} \ln f\left(x;\theta\right) =& \ln\binom{m}{x} + x\ln\frac{\theta}{m} + \left(m-x\right)\ln\left(1-\frac{\theta}{m}\right) \\ =& \left(x-\theta\right)\ln\frac{\theta}{m-\theta} + \theta\ln\frac{\theta}{m-\theta} + m\ln\left(m-\theta\right) + \ln\binom{m}{x} - m\ln m. \end{aligned}$$

Considerând

$$A\left(\theta\right) = \theta\ln\theta + \left(m-\theta\right)\ln\left(m-\theta\right), \quad \text{avem} \quad A'\left(\theta\right) = \ln\frac{\theta}{m-\theta},$$

deci

$$f\left(x;\theta\right) = \left[L\left(x\right)-\theta\right]A'\left(\theta\right) + A\left(\theta\right) + N\left(x\right),$$

unde $L\left(x\right) = x$ şi $N\left(x\right) = \ln\binom{m}{x} - m\ln m$. Pe baza teoremei Rao–Cramer se obţine că

$$\bar{\Theta} = \frac{1}{n}\sum_{k=1}^{n} L\left(X_k\right) = \frac{1}{n}\sum_{k=1}^{n} X_k$$

este estimator eficient pentru parametrul $\theta = mp$.

5.2.4 Estimatori optimali

Definiţia 5.2.19. *Estimatorul nedeplasat* $\widehat{\Theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$*, pentru parametrul* θ*, este* optimal *dacă are dispersia minimă dintre toţi estimatorii nedeplasaţi ai lui* θ*.*

Observaţia 5.2.20. Un estimator eficient este optimal, invers proprietatea nu are loc.

Proprietatea 5.2.21. *Estimatorul optimal al unui parametru este unic.*

Demonstraţie. Fie $\widehat{\Theta}_1 = \widehat{\theta}_1(X_1, X_2, \ldots, X_n)$ şi $\widehat{\Theta}_2 = \widehat{\theta}_2(X_1, X_2, \ldots, X_n)$ doi estimatori optimali distincţi pentru parametrul θ. Dacă se consideră funcţia de selecţie $\widehat{\Theta} = \frac{1}{2}\left(\widehat{\Theta}_1 + \widehat{\Theta}_2\right)$, atunci $\widehat{\Theta}$ este estimator nedeplasat pentru θ, deoarece

$$E\left(\widehat{\Theta}\right) = \frac{1}{2}E\left(\widehat{\Theta}_1\right) + \frac{1}{2}E\left(\widehat{\Theta}_2\right) = \frac{1}{2}(\theta + \theta) = \theta.$$

Pe de altă parte

$$\begin{aligned} Var\left(\widehat{\Theta}\right) &= \frac{1}{4}\left[Var\left(\widehat{\Theta}_1\right) + Var\left(\widehat{\Theta}_2\right) + 2Cov\left(\widehat{\Theta}_1, \widehat{\Theta}_2\right)\right] \\ &= \frac{1}{2}\left[\sigma^2 + Cov\left(\widehat{\Theta}_1, \widehat{\Theta}_2\right)\right], \end{aligned}$$

unde $\sigma^2 = Var\left(\widehat{\Theta}_1\right) = Var\left(\widehat{\Theta}_2\right)$.

Din inegalitatea lui Schwarz se obţine

$$Cov^2\left(\widehat{\Theta}_1, \widehat{\Theta}_2\right) = E^2\left[\left(\widehat{\Theta}_1 - \theta\right)\left(\widehat{\Theta}_2 - \theta\right)\right] \leqslant E\left[\left(\widehat{\Theta}_1 - \theta\right)^2\right] E\left[\left(\widehat{\Theta}_2 - \theta\right)^2\right] = \sigma^4,$$

deci $Cov\left(\widehat{\Theta}_1, \widehat{\Theta}_2\right) \leqslant \sigma^2$. Prin urmare, putem scrie că $Var\left(\widehat{\Theta}\right) \leqslant \sigma^2$, ceea ce contrazice faptul că $\widehat{\Theta}_1$, $\widehat{\Theta}_2$ sunt estimatori optimali, afară de cazul în care $Var\left(\widehat{\Theta}\right) = \sigma^2$, adică $Cov\left(\widehat{\Theta}_1, \widehat{\Theta}_2\right) = \sigma^2$. Având în vedere aceste rezultate, putem scrie:

$$Var\left(\widehat{\Theta}_1 - \widehat{\Theta}_2\right) = Var\left(\widehat{\Theta}_1\right) + Var\left(\widehat{\Theta}_2\right) - 2Cov\left(\widehat{\Theta}_1, \widehat{\Theta}_2\right) = \sigma^2 + \sigma^2 - 2\sigma^2 = 0,$$

adică $Var\left(\widehat{\Theta}_1 - \widehat{\Theta}_2\right) = 0$. Dar avem că $E\left(\widehat{\Theta}_1\right) = E\left(\widehat{\Theta}_2\right)$, ceea ce implică faptul că $E\left[\left(\widehat{\Theta}_1 - \widehat{\Theta}_2\right)^2\right] = 0$. Prin urmare se ajunge la $\widehat{\Theta}_1 = \widehat{\Theta}_2$, în contradicţie cu afirmaţia că cei doi estimatori sunt distincţi. □

Teorema 5.2.22 (Rao-Blackwell). *Fie caracteristica X cu funcţia de probabilitate $f(x;\theta)$ şi fie $\widehat{\Theta} = \widehat{\theta}(X_1,\ldots,X_n)$ un estimator nedeplasat pentru parametrul θ. Dacă statistica $S = S(X_1,\ldots,X_n)$ este o statistică suficientă pentru parametrul θ, atunci estimatorul $\bar{\Theta} = \bar{\theta}(X_1,\ldots,X_n) = E\left(\widehat{\Theta} \mid S\right)$ este un estimator nedeplasat pentru θ şi are loc relaţia $Var\left(\bar{\Theta}\right) \leqslant Var\left(\widehat{\Theta}\right)$.*

Demonstraţie. Deoarece statistica S este suficientă rezultă că funcţia de probabilitate $f(x_1,\ldots,x_n;\theta \mid s)$ condiţionată a vectorului aleator $(X_1,\ldots,X_n)$ de evenimentul $(S=s)$ nu depinde de θ, deci

$$E\left(\widehat{\Theta} \mid S=s\right) = \int\ldots\int_{\mathbb{R}^n} \widehat{\theta}(x_1,\ldots,x_n)\, f(x_1,\ldots,x_n;\theta \mid s)\, dx_1\ldots dx_n$$

nu depinde de θ, adică este o funcţie de selecţie ce nu depinde de θ. Pe de altă parte, folosind proprietăţi ale mediei condiţionate avem

$$E\left(\bar{\Theta}\right) = E\left[E\left(\widehat{\Theta} \mid S\right)\right] = E\left(\widehat{\Theta}\right) = \theta,$$

deci nedeplasarea are loc.

Pentru a stabili inegalitatea dintre dispersiile celor doi estimatori avem că

$$\begin{aligned} Var\left(\widehat{\Theta}\right) =& E\left[\left(\widehat{\Theta}-\theta\right)^2\right] = E\left[\left(\widehat{\Theta} - E\left(\widehat{\Theta} \mid S\right) + E\left(\widehat{\Theta} \mid S\right) - \theta\right)^2\right] \\ =& E\left[\left(\widehat{\Theta} - E\left(\widehat{\Theta} \mid S\right)\right)^2\right] + E\left[\left(E\left(\widehat{\Theta} \mid S\right) - \theta\right)^2\right] \\ &+2E\left[\left(\widehat{\Theta} - E\left(\widehat{\Theta} \mid S\right)\right)\left(E\left(\widehat{\Theta} \mid S\right) - \theta\right)\right]. \end{aligned}$$

Dar avem că

$$E\left[Var\left(\widehat{\Theta} \mid S\right)\right] = E\left[E\left(\left(\widehat{\Theta} - E\left(\widehat{\Theta} \mid S\right)\right)^2 \mid S\right)\right] = E\left[\left(\widehat{\Theta} - E\left(\widehat{\Theta} \mid S\right)\right)^2\right],$$

adică primul termen din expresia lui $Var\left(\widehat{\Theta}\right)$. Al doilea termen din aceeaşi expresie este

$$E\left[\left(E\left(\widehat{\Theta} \mid S\right) - \theta\right)^2\right] = E\left[\left(\bar{\Theta}-\theta\right)^2\right] = Var\left(\bar{\Theta}\right),$$

iar ultimul termen este nul deoarece pentru $S=s$ fixat, avem $E\left(\widehat{\Theta}|s\right) - \theta = \text{const.}$, deci

$$\begin{aligned} E\left[\left(\widehat{\Theta} - E\left(\widehat{\Theta} \mid s\right)\right)\left(E\left(\widehat{\Theta} \mid s\right) - \theta\right)\right] &= \left[E\left(\widehat{\Theta} \mid s\right) - \theta\right] E\left[\widehat{\Theta} - E\left(\widehat{\Theta} \mid s\right)\right] \\ &= \left[E\left(\widehat{\Theta} \mid s\right) - \theta\right]\left[E\left(\widehat{\Theta}\right) - E\left(\widehat{\Theta}\right)\right] = 0. \end{aligned}$$

Am stabilit astfel relaţia $Var\left(\widehat{\Theta}\right) = E\left[Var\left(\widehat{\Theta} \mid S\right)\right] + Var\left(\bar{\Theta}\right)$, de unde $Var\left(\bar{\Theta}\right) \leqslant Var\left(\widehat{\Theta}\right)$. □

Definiţia 5.2.23. *Statistica $S = S(X_1, X_2, \ldots, X_n)$ este* completă *pentru familia de legi de probabilitate $f(x;\theta)$, $\theta \in A$, dacă $E[\varphi(S)] = 0$, pentru orice $\theta \in A$, implică faptul că $\varphi = 0$ a.s.*

Exemplul 5.2.24. Statistica $S = \sum_{k=1}^{n} X_k$ este completă pentru familia de legi Poisson cu parametrul $\lambda > 0$.

Prima dată avem că S urmează legea lui Poisson de parametru $n\lambda$, deci

$$E[\varphi(S)] = \sum_{s=0}^{\infty} \varphi(s) \frac{(n\lambda)^s}{s!} e^{-n\lambda} = e^{-n\lambda} \sum_{s=0}^{\infty} \frac{\varphi(s)\, n^s}{s!} \lambda^s.$$

Pe de altă parte $E[\varphi(S)] = 0$, pentru orice $\lambda > 0$, conduce la

$$\sum_{s=0}^{\infty} \frac{\varphi(s)\, n^s}{s!} \lambda^s = 0,$$

pentru orice $\lambda > 0$, ceea ce are loc numai dacă fiecare coeficient este nul, adică $\varphi(s) = 0$, când $s \in \mathbb{N}$.

Teorema 5.2.25 (Lehmann–Scheffé). *În condiţiile Teoremei Rao–Blackwell, dacă statistica $S = S(X_1, X_2, \ldots, X_n)$ este completă, atunci estimatorul*

$$\bar{\Theta} = \bar{\theta}(X_1, X_2, \ldots, X_n) = E\left(\widehat{\Theta} \mid S\right)$$

este estimator optimal.

Demonstraţie. Fie $\widetilde{\Theta} = \widetilde{\theta}(X_1, X_2, \ldots, X_n)$ un estimator nedeplasat pentru parametrul θ. Folosind Teorema Rao–Blackwell, avem că estimatorul

$$\Theta_1 = \theta_1(X_1, X_2, \ldots, X_n) = E\left(\widetilde{\Theta} \mid S\right)$$

este nedeplasat pentru parametrul θ, adică $E(\Theta_1) = \theta$, şi $Var(\Theta_1) \leqslant Var\left(\widetilde{\Theta}\right)$. Aşadar avem că $E(\Theta_1) = E\left(\bar{\Theta}\right) = \theta$ sau $E\left[E\left(\widehat{\Theta} \mid S\right)\right] = E\left[E\left(\widetilde{\Theta} \mid S\right)\right] = \theta$, de unde $E\left[E\left(\widehat{\Theta} \mid S\right) - E\left(\widetilde{\Theta} \mid S\right)\right] = 0$. Având în vedere că statistica S este completă rezultă că $E\left(\widehat{\Theta} \mid s\right) = E\left(\widetilde{\Theta} \mid s\right)$ a.s., deci $\bar{\Theta} = \Theta_1$ a.s., de unde $Var\left(\bar{\Theta}\right) = Var(\Theta_1)$.

În final $Var\left(\bar{\Theta}\right) = Var\left(\Theta_1\right) \leqslant Var\left(\widetilde{\Theta}\right)$, adică $Var\left(\bar{\Theta}\right) \leqslant Var\left(\widetilde{\Theta}\right)$, pentru orice estimator nedeplasat $\widetilde{\Theta}$, ceea ce este chiar condiţia din definiţia optimalităţii. □

Exemplul 5.2.26. Fie caracteristica X, care urmează legea lui Poisson de parametru $\lambda > 0$. Vrem să determinăm un estimator optimal pentru parametrul $\theta = e^{-\lambda}$.

Funcţia de frecvenţă a caracteristicii X este

$$f\left(x;\theta\right) = \frac{\lambda^x}{x!}e^{-\lambda} = \theta\frac{1}{x!}\left(\ln\frac{1}{\theta}\right)^x.$$

În Exemplul 5.2.24 am văzut că statistica suficientă $S = \sum_{k=1}^{n} X_k$ este completă pentru familia de legi Poisson şi urmează legea lui Poisson de parametru $n\lambda = n\ln\frac{1}{\theta}$.

Dacă se consideră funcţia de selecţie

$$\widehat{\Theta} = \widehat{\theta}\left(X_1, X_2, \ldots, X_n\right) = \frac{1}{n}\,\text{card}\left\{\, X_i \,\middle|\, X_i = 0,\ i = \overline{1,n}\,\right\},$$

atunci $\widehat{\Theta}$ are distribuţia $\widehat{\Theta}\begin{pmatrix} \frac{k}{n} \\ \binom{n}{k}\theta^k\left(1-\theta\right)^{n-k} \end{pmatrix}_{k=\overline{0,n}}$, deci $E\left(\widehat{\Theta}\right) = \frac{1}{n}n\theta = \theta$, adică $\widehat{\Theta}$ este estimator nedeplasat pentru parametrul θ.

Dacă se introduc variabilele aleatoare $Y_1, Y_2, \ldots, Y_n$ cu distribuţiile date prin

$$Y_i = \begin{cases} 1, & \text{dacă } X_i = 0, \\ 0, & \text{dacă } X_i \neq 0, \end{cases}$$

deci $P\left(Y_i = 1\right) = \theta$, $P\left(Y_i = 0\right) = 1 - \theta$, atunci avem că $\widehat{\Theta} = \frac{1}{n}\sum_{k=1}^{n} Y_k$.

Aplicăm Teorema Rao–Blackwell şi obţinem

$$\bar{\Theta} = E\left(\widehat{\Theta} \,\middle|\, S\right) = \frac{1}{n}E\left(\sum_{k=1}^{n} Y_k \,\middle|\, S\right) = E\left(Y_1 \,\middle|\, S\right).$$

Pentru $S = s$, avem

$$E\left(Y_1 \,\middle|\, S = s\right) = P\left(Y_1 = 1 \,\middle|\, S = s\right)\cdot 1 = P\left(X_1 = 0 \,\middle|\, S = s\right).$$

Folosind formula lui Bayes avem

$$P\left(X_1=0 \mid S=s\right)=\frac{P\left(X_1=0\right)P\left(S=s|X_1=0\right)}{P\left(S=s\right)}$$
$$=\frac{P\left(X_1=0\right)P\left(X_2+\cdots+X_n=s\right)}{P\left(S=s\right)}$$
$$=\frac{\exp\left(-\lambda\right)\frac{((n-1)\lambda)^s}{s!}\exp[-\left(n-1\right)\lambda]}{\frac{(n\lambda)^s}{s!}\exp\left(-n\lambda\right)}=\left(\frac{n-1}{n}\right)^s,$$

de unde se obţine că

$$\bar{\Theta}=E\left(\widehat{\Theta}\mid S\right)=E\left(Y_1|S\right)=\left(1-\frac{1}{n}\right)^S.$$

Am ajuns astfel la estimatorul optimal pentru parametrul $\theta=e^{-\lambda}$, care este dat prin formula

$$\bar{\Theta}=\bar{\theta}\left(X_1,X_2,\ldots,X_n\right)=\left(1-\frac{1}{n}\right)^{n\bar{X}}.$$

5.3 Metode pentru estimarea parametrilor

5.3.1 Metoda momentelor

Se consideră caracteristica X, care are funcţia de probabilitate $f\left(x;\boldsymbol{\theta}\right)$ cu parametrul necunoscut $\boldsymbol{\theta}=(\theta_1,\theta_2,\ldots,\theta_p)\in\boldsymbol{A}\subset\mathbb{R}^p$ şi o selecţie repetată de volum n.

Definiţia 5.3.1. *Numim* estimator *pentru parametrul $\boldsymbol{\theta}$ obţinut prin* metoda momentelor *soluţia $\bar{\boldsymbol{\Theta}}=\left(\bar{\Theta}_1,\bar{\Theta}_2,\ldots,\bar{\Theta}_p\right)$ a sistemului*

$$\nu_k=\bar{\nu}_k,\quad k=\overline{1,p},$$

unde ν_k este momentul teoretic de ordin k, $\nu_k=E\left(X^k\right)$, iar $\bar{\nu}_k$ este momentul de selecţie de ordinul k, adică

$$\bar{\nu}_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k.$$

Observaţia 5.3.2. Metoda momentelor este una din cele mai vechi metode de estimare a parametrilor, folosită iniţial în mod empiric. Este fundamentată teoretic pe faptul că momentele de selecţie sunt estimatori absolut corecţi pentru momentele teoretice corespunzătoare. Este cunoscut de asemenea că $\bar{\boldsymbol{\Theta}}\xrightarrow{\text{a.s.}}\boldsymbol{\theta}$, când $n\to\infty$.

Exemplul 5.3.3. Se consideră caracteristica X, care urmează legea gamma cu parametrii $a, b > 0$ necunoscuţi. Densitatea de probabilitate pentru X este

$$f(x; a, b) = \frac{1}{\Gamma(a)\, b^a} x^{a-1} \mathrm{e}^{-\frac{x}{b}}, \quad x > 0.$$

Vrem să estimăm parametrii a şi b prin metoda momentelor.

Se ştie că

$$\nu_1 = E(X) = \int_{-\infty}^{+\infty} x f(x; a, b) = \frac{1}{\Gamma(a)\, b^a} \int_0^{+\infty} x^a e^{-\frac{x}{b}} dx = ab \quad \text{şi}$$

$$\nu_2 = E\left(X^2\right) = \int_{-\infty}^{+\infty} x^2 f(x; a, b) = \frac{1}{\Gamma(a)\, b^a} \int_0^{+\infty} x^{a+1} e^{-\frac{x}{b}} dx = ab^2 (a+1).$$

Prin urmare, avem sistemul de ecuaţii

$$\begin{cases} ab = \bar{\nu}_1 = \bar{X} \\ ab^2 (a+1) = \bar{\nu}_2 = \bar{X}^2 + \bar{\mu}_2, \end{cases}$$

care are soluţia

$$\bar{a} = \frac{\bar{X}^2}{\bar{\mu}_2}, \quad \bar{b} = \frac{\bar{\mu}_2}{\bar{X}}.$$

5.3.2 Metoda verosimilităţii maxime

Se consideră caracteristica X cu funcţia de probabilitate $f(x; \boldsymbol{\theta})$, unde parametrul necunoscut $\boldsymbol{\theta} \in \boldsymbol{A} \subset \mathbb{R}^p$. Relativ la caracteristica X se consideră o selecţie repetată de volum n.

Definiţia 5.3.4. *Numim* estimator de verosimilitate maximă, *pentru parametrul* $\boldsymbol{\theta}$, *statistica*

$$\widehat{\boldsymbol{\Theta}} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$$

pentru care se obţine maximul funcţiei de verosimilitate

$$g(X_1, X_2, \ldots, X_n; \boldsymbol{\theta}) = \prod_{k=1}^{n} f(X_k; \boldsymbol{\theta}),$$

iar $\widehat{\boldsymbol{\theta}} = \widehat{\theta}(x_1, x_2, \ldots, x_n)$ *se numeşte* estimaţie de verosimilitate maximă, *pentru parametrul* $\boldsymbol{\theta}$.

Observaţia 5.3.5. În definiţia estimatorului de verosimilitate maximă $\widehat{\boldsymbol{\Theta}}$ nu este necesar ca $f(x; \boldsymbol{\theta})$ să fie diferenţiabilă în raport cu $\boldsymbol{\theta}$. De asemenea, nu este neapărat unic şi nedeplasat.

Observaţia 5.3.6. Dacă funcţia de verosimilitate este diferenţiabilă de două ori în raport cu $\boldsymbol{\theta}$, atunci estimatorul de verosimilitate maximă se obţine ca soluţie a sistemului de ecuaţii

$$\frac{\partial g\left(X_1, X_2, \ldots, X_n; \boldsymbol{\theta}\right)}{\partial \theta_k} = 0, \quad k = \overline{1, p}.$$

Sistemul de ecuaţii este echivalent cu

$$\frac{\partial \ln g\left(X_1, X_2, \ldots, X_n; \boldsymbol{\theta}\right)}{\partial \theta_k} = \sum_{i=1}^{n} \frac{\partial \ln f\left(X_i; \boldsymbol{\theta}\right)}{\partial \theta_k} = 0, \quad k = \overline{1, p},$$

numit *sistemul ecuaţiilor de verosimilitate maximă*.

Exemplul 5.3.7. Să determinăm estimatorii de verosimilitate maximă pentru valoarea medie şi abaterea standard, dacă se consideră caracteristica X, care urmează legea normală $\mathcal{N}\left(\mu, \sigma\right)$.

Se ştie că $E\left(X\right) = \mu$ şi $\sigma\left(X\right) = \sigma$, iar densitatea de probabilitate a lui X este

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

Pentru a scrie sistemul de verosimilitate maximă, avem că

$$\ln f\left(x; \mu, \sigma\right) = -\ln\sqrt{2\pi} - \ln\sigma - \frac{\left(x-\mu\right)^2}{2\sigma^2},$$

de unde

$$\frac{\partial \ln f\left(x; \mu, \sigma\right)}{\partial \mu} = \frac{x-\mu}{\sigma^2}, \quad \frac{\partial \ln f\left(x; \mu, \sigma\right)}{\partial \sigma} = -\frac{1}{\sigma} + \frac{\left(x-\mu\right)^2}{\sigma^3}.$$

În acest mod se obţine

$$\begin{cases} \dfrac{\partial \ln g}{\partial \mu} = \displaystyle\sum_{k=1}^{n} \frac{\partial \ln f\left(X_k; \mu, \sigma\right)}{\partial \mu} = \sum_{k=1}^{n} \frac{X_k - \mu}{\sigma^2} = 0 \\ \dfrac{\partial \ln g}{\partial \sigma} = \displaystyle\sum_{k=1}^{n} \frac{\partial \ln f\left(X_k; \mu, \sigma\right)}{\partial \sigma} = \sum_{k=1}^{n} \left[-\frac{1}{\sigma} + \frac{\left(X_k - \mu\right)^2}{\sigma^3}\right] = 0, \end{cases}$$

sau

$$\begin{cases} \displaystyle\sum_{k=1}^{n} \left(X_k - \mu\right) = 0 \\ \displaystyle\sum_{k=1}^{n} \left[-\sigma^2 + \left(X_k - \mu\right)^2\right] = 0, \end{cases}$$

de unde se rezultă estimatorii de verosimilitate maximă

$$\mu^\star = \frac{1}{n}\sum_{k=1}^{n} X_k = \bar{X}, \quad \sigma^\star = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2} = \sqrt{\bar{\mu}_2},$$

pentru parametrii μ şi σ.

Exemplul 5.3.8. Caracteristica X urmează legea uniformă pe intervalul $(0, \theta)$, unde parametrul $\theta > 0$ este necunoscut. Relativ la caracteristica X se consideră o selecţie repetată de volum n. Vrem să se determinăm estimatorul de verosimilitate maximă $\widehat{\Theta}$, pentru parametrul necunoscut θ.

Estimatorul $\widehat{\Theta}$ de verosimilitate maximă pentru θ se determină astfel încât funcţia de verosimilitate

$$g(X_1, X_2, \ldots, X_n; \theta) = \prod_{k=1}^{n} f\left(x_k; \theta\right) = \frac{1}{\theta^n}$$

să fie maximă pentru $\widehat{\Theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$. Deoarece

$$\max\left\{ \frac{1}{\theta^n} \;\middle|\; X_1 \leqslant \theta, \ldots, X_n \leqslant \theta \right\} = \left(\frac{1}{\max\left\{X_i,\ i = \overline{1,n}\right\}} \right)^n,$$

se obţine că $\widehat{\Theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n) = \max\left\{X_i,\ i = \overline{1,n}\right\}$.

Se observă că în acest exemplu nu s-a putut folosi ecuaţia de verosimilitate maximă, deoarece domeniul valorilor caracteristicii X, care este $(0, \theta)$, depinde de parametrul estimat.

Vom arăta, în continuare, că estimatorul astfel construit este estimator corect pentru parametrul θ. Apoi vom folosi acest estimator pentru obţinerea unui estimator absolut corect pentru θ.

Funcţia de repartiţie a statisticii $\widehat{\Theta}$ este

$$F_{\widehat{\Theta}}\left(x; \theta\right) = P\left(\widehat{\Theta} \leqslant x\right) = \prod_{i=1}^{n} P\left(X_i \leqslant x\right) = \left[F_X\left(x; \theta\right)\right]^n,$$

deci $\widehat{\Theta}$ are densitatea de probabilitate

$$f_{\widehat{\Theta}}\left(x; \theta\right) = \frac{\partial F_{\widehat{\Theta}}\left(x; \theta\right)}{\partial x} = n f\left(x; \theta\right)\left[F_X\left(x; \theta\right)\right]^{n-1} = \frac{n x^{n-1}}{\theta^n}, \quad \text{când} \quad x \in (0, \theta).$$

Se calculează uşor

$$E\left(\widehat{\Theta}\right) = \frac{n}{\theta^n}\int_0^\theta x \cdot x^{n-1} dx = \frac{n}{n+1}\,\theta, \quad E\left(\widehat{\Theta}^2\right) = \frac{n}{\theta^n}\int_0^\theta x^{n+1} dx = \frac{n}{n+2}\,\theta^2.$$

Astfel se obţine că

$$\lim_{n\to\infty} E\left(\widehat{\Theta}\right) = \theta$$

$$\lim_{n\to\infty} Var\left(\widehat{\Theta}\right) = \lim_{n\to\infty}\left[\frac{n}{n+2}\,\theta^2 - \frac{n^2}{(n+1)^2}\,\theta^2\right] = \lim_{n\to\infty}\frac{n}{(n+1)^2\,(n+2)}\,\theta^2 = 0.$$

Prin urmare, $\widehat{\Theta}$ este estimator corect pentru θ.

Punând condiţia $E\left(\bar{\Theta}\right) = \theta$, rezultă că

$$\theta = E\left(\bar{\Theta}\right) = \kappa_n E\left(\widehat{\Theta}\right) = \kappa_n \frac{n}{n+1}\,\theta,$$

de unde se obţine $\kappa_n = \frac{n+1}{n}$ şi în final

$$\bar{\Theta} = \frac{n+1}{n}\,\widehat{\Theta} = \frac{n+1}{n}\,\max\left\{\,X_i,\ i = \overline{1,n}\,\right\}.$$

Deoarece

$$Var\left(\bar{\Theta}\right) = \left(\frac{n+1}{n}\right)^2 Var\left(\widehat{\Theta}\right) = \frac{n\,(n+1)^2}{n^2\,(n+1)^2\,(n+2)}\,\theta = \frac{1}{n(n+2)}\,\theta \to 0,$$

când $n \to \infty$, rezultă că $\bar{\Theta}$ este estimator absolut corect pentru parametrul θ.

Proprietatea 5.3.9. *Dacă $S = S\,(X_1, X_2, \ldots, X_n)$ este statistică suficientă pentru parametrul θ, iar $\widehat{\Theta}$ este estimator de verosimilitate maximă pentru θ, atunci $\widehat{\Theta}$ este funcţie de S.*

Demonstraţie. Deoarece statistica S este suficientă rezultă că

$$g\,(x_1, x_2, \ldots, x_n; \theta) = \varphi\,(x_1, x_2, \ldots, x_n)\,h\,(s, \theta)\,,$$

deci maximul lui g, după θ, se obţine atunci şi numai atunci când se obţine maximul lui h după θ. Astfel că $\widehat{\Theta}$ se exprimă în funcţie de S. □

Teorema 5.3.10. *Dacă $\widehat{\Theta} = \widehat{\theta}\,(X_1, X_2, \ldots, X_n)$ este funcţie de estimaţie eficientă pentru parametrul θ, atunci $\widehat{\Theta}$ este estimator de verosimilitate maximă pentru θ.*

Demonstraţie. Deoarece $\widehat{\Theta}$ este estimator eficient pentru θ, din inegalitatea Rao–Cramer (cu egalitate), avem că

$$\frac{\partial \ln g\,(X_1, X_2, \ldots, X_n; \theta)}{\partial\theta} = K\,(\theta)\left(\widehat{\Theta} - \theta\right).$$

Astfel că

$$\frac{\partial \ln g\left(X_1, X_2, \ldots, X_n; \widehat{\Theta}\right)}{\partial\theta} = K\left(\widehat{\Theta}\right)\left(\widehat{\Theta} - \widehat{\Theta}\right) = 0,$$

deci $\widehat{\Theta}$ verifică ecuaţia verosimilităţii maxime. □

Observaţia 5.3.11. Dacă valoarea adevărată a parametrului θ este θ_0, se arată că funcţia de verosimilitate maximă $\widehat{\Theta}$ pentru parametrul θ are următoarele comportări asimptotice: $\widehat{\Theta} \xrightarrow{\text{a.s.}} \theta_0$, când $n \to \infty$, iar $\sqrt{n}\left(\widehat{\Theta} - \theta_0\right)$ converge în repartiţie la legea normală $\mathcal{N}\left(0, \frac{1}{\sqrt{I(\theta_0)}}\right)$.

5.3.3 Metoda minimului χ^2

Se consideră colectivitatea $\mathcal{C}$ şi caracteristica X cercetată, cu legea de probabilitate dată prin funcţia $f(x;\boldsymbol{\theta})$, unde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s) \in \boldsymbol{A} \subset \mathbb{R}^s$. Domeniul valorilor lui X îl considerăm compus din clasele C_i, $i = \overline{1,k}$. Vom introduce notaţiile următoare $p_i = p_i(\boldsymbol{\theta}) = P(X \in C_i)$, adică probabilitatea ca un individ luat la întâmplare din colectivitatea $\mathcal{C}$ să aparţină clasei C_i.

Dacă se consideră o selecţie repetată de volum n cu datele de selecţie $x_1, \ldots, x_n$, respectiv variabilele de selecţie $X_1, \ldots, X_n$, notăm prin n_i frecvenţa absolută a datelor de selecţie din clasa C_i. Fie N_i variabila aleatoare (de selecţie) corespunzătoare lui n_i, atunci vectorul aleator $N = (N_1, \ldots, N_k)$ urmează legea multinomială de parametri $p_i = p_i(\boldsymbol{\theta})$, $i = \overline{1,k}$.

Definiţia 5.3.12. Estimatorul cu χ^2 minim *pentru parametrul $\boldsymbol{\theta}$ este estimatorul (funcţia de selecţie) $\bar{\boldsymbol{\Theta}} = \bar{\theta}(X_1, X_2, \ldots, X_n)$, care realizează valoarea minimă a expresiei*

$$\chi^2 = \sum_{i=1}^{k} \frac{[N_i - np_i(\boldsymbol{\theta})]^2}{np_i(\boldsymbol{\theta})}$$

în raport cu parametrul $\boldsymbol{\theta}$, iar $\bar{\boldsymbol{\theta}} = \bar{\theta}(x_1, x_2, \ldots, x_n)$ se numeşte estimaţie cu χ^2 minim *pentru parametrul $\boldsymbol{\theta}$.*

Observaţia 5.3.13. Dacă notăm $\widehat{p}_i = \frac{n_i}{n}$, atunci valoarea lui χ^2 se poate scrie în felul următor:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(n\widehat{p}_i - np_i)^2}{np_i} = n\left(\sum_{i=1}^{k} \frac{\widehat{p}_i^2}{p_i} - 1\right),$$

prin urmare, minimul lui χ^2 se obţine când este minimă expresia $\sum_{i=1}^{k} \frac{\widehat{p}_i^2}{p_i}$.

Observaţia 5.3.14. Dacă $p_i(\boldsymbol{\theta})$, $i = \overline{1,s}$, sunt diferenţiabile de două ori în raport cu $\boldsymbol{\theta}$, atunci $\bar{\boldsymbol{\Theta}}$ se obţine ca soluţie a sistemului

$$\frac{\partial}{\partial \theta_j}\left[\sum_{i=1}^{k} \frac{(N_i - np_i(\boldsymbol{\theta}))^2}{np_i(\boldsymbol{\theta})}\right] = 0, \quad j = \overline{1,s}.$$

Observaţia 5.3.15. În locul expresiei lui χ^2 de mai sus, numit *indicatorul lui Pearson*, se pot utiliza alte expresii ca:

$$D^2(\boldsymbol{\theta}) = \sum_{i=1}^{k} \frac{[N_i - np_i(\boldsymbol{\theta})]^2}{N_i}, \quad \textit{indicatorul lui Neyman},$$

$$D^2(\boldsymbol{\theta}) = 2\sum_{i=1}^{k} N_i \ln\left[\frac{N_i}{np_i(\boldsymbol{\theta})}\right], \quad \textit{indicatorul de verosimilitate},$$

$$D^2(\boldsymbol{\theta}) = 2n\sum_{i=1}^{k} p_i(\boldsymbol{\theta}) \ln\left[\frac{np_i(\boldsymbol{\theta})}{N_i}\right], \quad \textit{indicatorul lui Kullbach},$$

$$D^2(\boldsymbol{\theta}) = \sum_{i=1}^{k} n\widehat{p}_i(1-\widehat{p}_i)\left[\ln\frac{\widehat{p}_i}{1-\widehat{p}_i} - \ln\frac{p_i(\boldsymbol{\theta})}{1-p_i(\boldsymbol{\theta})}\right],$$

indicatorul lui Berkson.

5.3.4 Metoda intervalelor de încredere

Se consideră caracteristica X cu legea de probabilitate $f(x;\theta)$, unde parametrul necunoscut $\theta \in A \subset \mathbb{R}$. Fie o selecţie repetată de volum n şi numărul $\alpha \in (0,1)$, numit *probabilitate de risc*, $1-\alpha$ numindu-se *probabilitate de încredere*.

Definiţia 5.3.16. *Numim* interval de încredere *pentru parametrul θ, intervalul aleator*

$$\left(\bar{\Theta}_1, \bar{\Theta}_2\right) = \left(\bar{\theta}_1(X_1, X_2, \ldots, X_n), \bar{\theta}_2(X_1, X_2, \ldots, X_n)\right),$$

unde statisticile $\bar{\Theta}_1$ şi $\bar{\Theta}_2$ sunt astfel încât $P\left(\bar{\Theta}_1 < \theta < \bar{\Theta}_2\right) = 1-\alpha$, iar intervalul numeric $\left(\bar{\theta}_1, \bar{\theta}_2\right) = \left(\bar{\theta}_1(x_1, x_2, \ldots, x_n), \bar{\theta}_2(x_1, x_2, \ldots, x_n)\right)$, se numeşte valoarea intervalului de încredere *pentru parametrul θ*.

Observaţia 5.3.17. Pentru determinarea intervalului de încredere se consideră o statistică $S = S(X_1, X_2, \ldots, X_n)$, care urmează o lege de probabilitate cunoscută, dar în expresia căreia apare parametrul θ. Se determină apoi intervalul numeric (s_1, s_2) astfel încât $P(S \in (s_1, s_2)) = 1-\alpha$, relaţie care se va scrie, în mod echivalent, $P\left(\bar{\Theta}_1 < \theta < \bar{\Theta}_2\right) = 1-\alpha$.

Observaţia 5.3.18. Cu cât α este mai mic şi intervalul de încredere are lungimea mai mică cu atât estimaţia parametrului necunoscut este mai bună.

Observaţia 5.3.19. Din relaţia $P\big(S \in (s_1, s_2)\big) = 1-\alpha$, intervalul (s_1, s_2) nu este determinat în mod unic.

De la problemă la problemă se mai adaugă o condiţie suplimentară, de exemplu fixarea valorii fie a lui s_1, fie a lui s_2. De asemenea, se poate considera o legătură între s_1 şi s_2 dată de anumite ipoteze de lucru.

Interval de încredere pentru media teoretică – dispersie cunoscută

Se consideră caracteristica X ce urmează legea normală $\mathcal{N}(m, \sigma)$, unde $m \in \mathbb{R}$ este necunoscut, iar $\sigma > 0$ este cunoscut.

Pentru construirea unui interval de încredere pentru media teoretică m necunoscută efectuăm o selecţie repetată de volum n şi considerăm probabilitatea de încredere $1 - \alpha$ dată, $\alpha \in (0, 1)$.

Se construieşte statistica

$$Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}},$$

care urmează legea normală $\mathcal{N}(0, 1)$ (a se vedea Observaţia 4.2.6). Prin urmare, pentru α dat determinăm intervalul numeric (z_1, z_2) astfel încât

$$P\big(Z \in (z_1, z_2)\big) = \Phi(z_2) - \Phi(z_1) = 1 - \alpha,$$

unde

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt,$$

este *funcţia lui Laplace*. Funcţia lui Laplace este tabelată în *Anexa I* pentru valori pozitive ale argumentului, având în vedere că $\Phi(-x) = -\Phi(x)$.

Deoarece dubla inegalitate

$$z_1 < \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} < z_2$$

este echivalentă cu

$$\bar{m}_1 = \bar{X} - z_2 \frac{\sigma}{\sqrt{n}} < m < \bar{X} - z_1 \frac{\sigma}{\sqrt{n}} = \bar{m}_2$$

rezultă că $P(\bar{m}_1 < m < \bar{m}_2) = 1 - \alpha$, adică $(\bar{m}_1, \bar{m}_2)$ este un interval de încredere pentru media teoretică m.

Desigur că intervalul numeric (z_1, z_2) nu este în mod unic determinat. Dacă nu există nici o informaţie suplimentară relativă la valoarea medie, atunci se va considera intervalul de încredere de lungime minimă, pentru α fixat, şi se obţine când $z_1 = -z_2$. În acest caz $z_2 = z_{1-\frac{\alpha}{2}}$, va fi dat prin relaţia $\Phi\left(z_{1-\frac{\alpha}{2}}\right) - \Phi\left(-z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$, ceea ce este echivalent cu $\Phi\left(z_{1-\frac{\alpha}{2}}\right) = \frac{1-\alpha}{2}$.

Când se foloseşte funcţia lui Laplace definită prin

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt,$$

atunci $z_{1-\frac{\alpha}{2}}$ se determină din relaţia $\Phi\left(z_{1-\frac{\alpha}{2}}\right) = 1-\frac{\alpha}{2}$ şi reprezintă de fapt cuantila de ordin $1-\frac{\alpha}{2}$.

Intervalul de încredere pentru parametrul m are extremităţile

$$\begin{aligned} \bar{m}_1 =& \bar{m}_1\left(X_1, X_2, \ldots, X_n\right) = \bar{X} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \\ \bar{m}_2 =& \bar{m}_2\left(X_1, X_2, \ldots, X_n\right) = \bar{X} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}. \end{aligned} \tag{5.3.1}$$

Observaţia 5.3.20. Geometric, dacă se consideră graficele funcţiei de repartiţie şi densităţii de probabilitate a legii normale $\mathcal{N}(0,1)$, modul de obţinere a cuantilei z_γ este prezentat în Figura 5.1. Adică, fie intersectând graficul funcţiei de repartiţie a legii normale $\mathcal{N}(0,1)$ cu dreapta de ecuaţie $y=\gamma$, fie alegându–l pe z_γ astfel ca aria umbrită de sub graficul densităţii de probabilitate să fie γ.

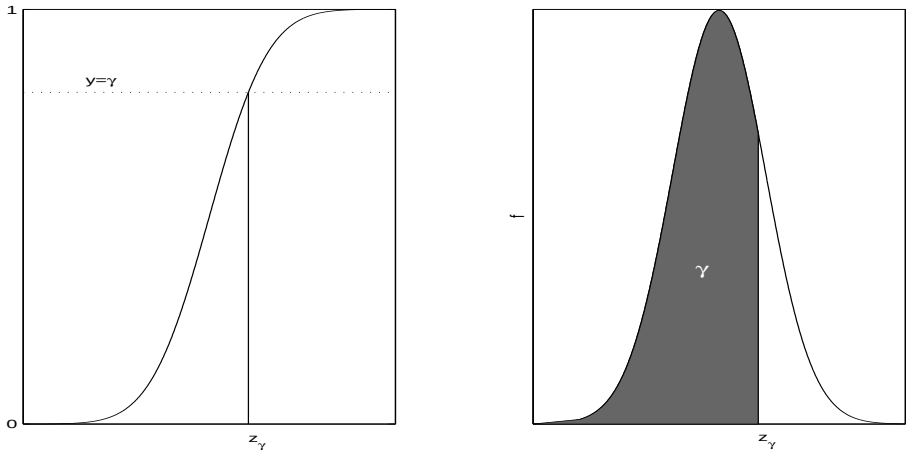

Figura 5.1: Determinarea cuantilei z_γ

Observaţia 5.3.21. Dacă există o informaţie relativă la valoarea medie de forma că aceasta nu este limitată superior, atunci intervalul numeric $(z_1, z_2) = (-\infty, z_{1-\alpha})$, care conduce la intervalul de încredere

$$(\bar{m}_1, \bar{m}_2) = \left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}, \infty\right).$$

În mod analog, dacă se cunoaşte că valoarea medie nu este limitată inferior, adică are tendinţa de a avea valori mici în raport cu ceea ce ne aşteptăm, atunci intervalul

numeric $(z_1, z_2) = (z_\alpha, \infty)$, care conduce la intervalul de încredere

$$(\bar{m}_1, \bar{m}_2) = \left(-\infty, \bar{X} - \frac{\sigma}{\sqrt{n}} z_\alpha \right).$$

Pe baza teoremei limită centrală avem că rezultatul obţinut se menţine când X urmează o lege de probabilitate oarecare, pentru $n > 30$ (a se vedea Proprietatea 4.2.7).

Exemplul 5.3.22. Relativ la populaţia $\mathcal{C}$ se cercetează caracteristica X privind media teoretică $E(X) = m$. Ştiind că dispersia teoretică a caracteristicii X este $Var(X) = 0.35$, să se stabilească un interval de încredere pentru media teoretică m cu probabilitatea de încredere $1 - \alpha = 0.95$, utilizând distribuţia empirică de selecţie

$$X \begin{pmatrix} 22.7 & 22.8 & 22.9 & 23.0 & 23.1 & 23.2 & 23.3 & 23.4 \\ 1 & 3 & 7 & 4 & 6 & 7 & 5 & 2 \end{pmatrix}.$$

Deoarece volumul selecţiei este $n = 35 > 30$, putem considera că statistica

$$Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}, \qquad \text{unde} \quad \sigma = \sqrt{Var(X)},$$

urmează legea normală $\mathcal{N}(0, 1)$. Aşadar, extremităţile intervalului de încredere pentru media teoretică m sunt date prin (5.3.1). Calculăm valorile acestor extremităţi pe baza datelor de selecţie.

Valoarea mediei de selecţie $\bar{X}$ este

$$\bar{x} = \frac{1}{35}(1 \cdot 22.7 + 3 \cdot 22.8 + 7 \cdot 22.9 + 4 \cdot 23 + 6 \cdot 23.1 + 7 \cdot 23.2 + 5 \cdot 23.3 + 2 \cdot 23.4) = 23.077,$$

iar din *Anexa I*, pentru $\frac{1-\alpha}{2} = 0.475$, se găseşte $z_{1-\frac{\alpha}{2}} = 1.96$.

De asemenea, avem că

$$\frac{\sigma}{\sqrt{n}} = \sqrt{\frac{Var(X)}{n}} = \sqrt{\frac{0.35}{35}} = \sqrt{0.01} = 0.1.$$

Obţinem, în acest fel, intervalul de încredere pentru media teoretică $m = E(X)$

$$\begin{aligned} \left(\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \,;\, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \right) &= (23.077 - 0.196 \,;\, 23.077 + 0.196) \\ &= (22.881 \,;\, 23.273). \end{aligned}$$

Programul 5.3.23. Calculele pot fi făcute cu următorul program Matlab:

```
x=[22.7,22.8*ones(1,3),22.9*ones(1,7),23*ones(1,4),...
  23.1*ones(1,6),23.2*ones(1,7),23.3*ones(1,5),23.4*ones(1,2)];
s=sqrt(0.35)*norminv(0.975)/sqrt(35);ma=mean(x);
m1=ma-s; m2=ma+s;
fprintf(' (m1,m2)=(%6.3f,%6.3f)',m1,m2)
```

şi care, în urma executării, afişează rezultatul

```
(m1,m2)=(22.881,23.273)
```

Interval de încredere pentru media teoretică – dispersia necunoscută

În condiţiile secţiunii precedente, dar cu $\sigma > 0$ necunoscut, se va considera statistica

$$T = \frac{\bar{X} - m}{\frac{\bar{\sigma}}{\sqrt{n}}} = \frac{\bar{X} - m}{\sqrt{\frac{\bar{\mu}_2}{n-1}}},$$

care, conform Proprietăţii 4.2.26, urmează legea Student cu $n-1$ grade de libertate.

Se determină intervalul numeric (t_1, t_2) astfel încât

$$P\big(T \in (t_1, t_2)\big) = F_{n-1}(t_2) - F_{n-1}(t_1) = 1 - \alpha,$$

unde

$$F_m(x) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi}\,\Gamma\left(\frac{m}{2}\right)} \int_{-\infty}^{x} \left(1 + \frac{t^2}{m}\right)^{-\frac{m+1}{2}} dt, \quad x \in \mathbb{R},$$

este funcţia de repartiţie a legii Student cu m grade de libertate şi care este tabelată, pentru anumite valori, în *Anexa II.*

Luând $t_2 = t_{n-1,1-\frac{\alpha}{2}}$, $t_1 = -t_2$, avem că $F_{n-1}\left(t_{n-1,1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$, iar $P\left(\bar{m}_1 < m < \bar{m}_2\right) = 1 - \alpha$, deci, intervalul de încredere pentru media teoretică m are extremităţile date prin

$$\begin{aligned} \bar{m}_1 &= \bar{X} - t_{n-1,1-\frac{\alpha}{2}} \frac{\bar{\sigma}}{\sqrt{n}}, \\ \bar{m}_2 &= \bar{X} + t_{n-1,1-\frac{\alpha}{2}} \frac{\bar{\sigma}}{\sqrt{n}}. \end{aligned} \tag{5.3.2}$$

Observaţia 5.3.24. Dacă există o informaţie relativă la valoarea medie de forma că aceasta nu este limitată superior, adică are tendinţa de a avea valori mai mari decât cele aşteptate, atunci intervalul numeric $(t_1, t_2) = (-\infty, t_{n-1,1-\alpha})$, care conduce la intervalul de încredere pentru m:

$$(\bar{m}_1, \bar{m}_2) = \left(\bar{X} - \frac{\bar{\sigma}}{\sqrt{n}} t_{n-1,1-\alpha}, \infty\right).$$

În mod analog, dacă se cunoaşte că valoarea medie nu este limitată inferior, atunci intervalul numeric $(t_1, t_2) = (t_{n-1,\alpha}, \infty)$, care conduce la intervalul de încredere

$$(\bar{m}_1, \bar{m}_2) = \left(-\infty, \bar{X} - \frac{\bar{\sigma}}{\sqrt{n}} t_{n-1,\alpha} \right).$$

Observaţia 5.3.25. Geometric, dacă se consideră graficele funcţiei de repartiţie şi densităţii de probabilitate a legii Student cu $n-1$grade de libertate, modul de obţinere a cuantilei $t_{n-1,\gamma}$ este prezentat în Figura 5.2.

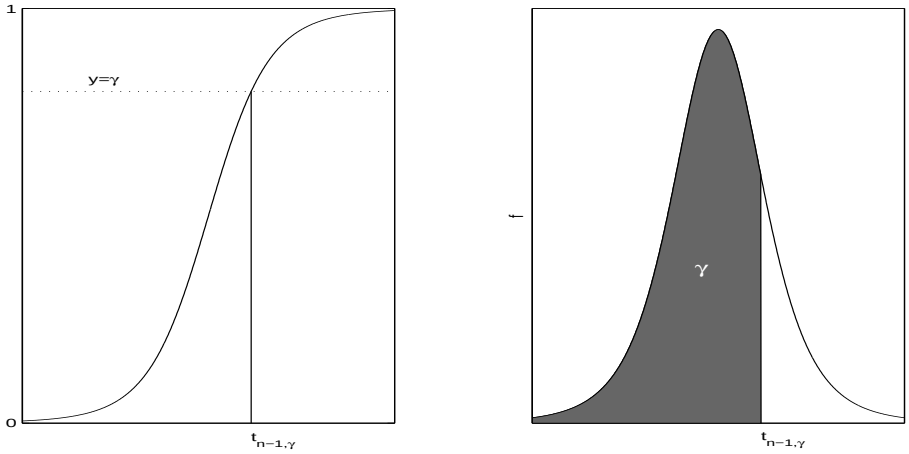

Figura 5.2: Determinarea cuantilei $t_{n-1,\gamma}$

Din teorema limită centrală avem că rezultatele pot fi aplicate pentru o caracteristică X ce urmează o lege de probabilitate oarecare, pentru $n > 30$.

Exemplul 5.3.26. Pentru recepţionarea unei mărfi ambalată în cutii, se efectuează un control, prin sondaj, privind greutatea X a cutiilor. Pentru 22 de cutii cântărite s-a obţinut distribuţia empirică de selecţie, relativ la caracteristica X:

$$X \begin{pmatrix} 2.7 & 2.8 & 2.9 & 3.0 & 3.1 & 3.2 & 3.3 \\ 1 & 2 & 5 & 3 & 5 & 4 & 2 \end{pmatrix}.$$

Folosind probabilitatea de încredere 0.98, să determinăm un interval de încredere pentru valoarea medie a greutăţii cutiilor, presupunând că X urmează legea normală $\mathcal{N}(m, \sigma)$.

Deoarece abaterea standard $\sigma = \sqrt{Var\,(X)}$ este necunoscută, se consideră statistica

$$T = \frac{\bar{X} - m}{\frac{\bar{\sigma}}{\sqrt{n}}},$$

care urmează legea Student cu $n-1$ grade de libertate.

Extremităţile intervalului de încredere pentru valorea medie teoretică $m = E\,(X)$ sunt date prin (5.3.2).

Pentru $n - 1 = 21$ şi $1 - \alpha = 0.98$ $(\alpha = 0.02)$, din *Anexa II* se determină $t_{n-1,1-\frac{\alpha}{2}} = 2.518$.

De asemenea, folosind datele de selecţie, obţinem valoarea $\bar{x}$ a mediei de selecţie $\bar{X}$, anume

$$\bar{x} = \frac{1}{22}\,(1{\cdot}2.7 + 2{\cdot}2.8 + 5{\cdot}2.9 + 3{\cdot}3 + 5{\cdot}3.1 + 4{\cdot}3.2 + 2{\cdot}3.3) = 3.032$$

şi valoarea abaterii standard de selecţie

$$\bar{\sigma} = \sqrt{\frac{1}{21}\sum_{k=1}^{7} f_k\,(x_k - \bar{x}\,)^2} = \sqrt{\frac{0.587728}{21}} = 0.167.$$

Putem scrie atunci intervalul (numeric) de încredere

$$\left(\bar{x} - t_{n-1,1-\frac{\alpha}{2}}\,\frac{\bar{\sigma}}{\sqrt{n}}; \bar{x} + t_{n-1,1-\frac{\alpha}{2}}\,\frac{\bar{\sigma}}{\sqrt{n}}\right)$$
$$= \left(3.032 - \frac{2.518 \cdot 0.167}{\sqrt{22}}\,;\, 3.032 + \frac{2.518 \cdot 0.167}{\sqrt{22}}\right) = (2.942\,;\, 3.122)\,.$$

Programul 5.3.27. Dacă se execută programul Matlab

```
x=[2.7,2.8*ones(1,2),2.9*ones(1,5),3*ones(1,3),...
       3.1*ones(1,5),3.2*ones(1,4),3.3*ones(1,2)];
ma=mean(x); va=var(x);
s=tinv(0.99,21)*sqrt(va)/sqrt(22);
m1=ma-s; m2=ma+s;
fprintf(' (m1,m2)=(%6.3f,%6.3f)',m1,m2)
```

relativ la problema precedentă, se obţine intervalul de încredere:

```
 (m1,m2)=( 2.942, 3.122)
```

Interval de încredere pentru dispersie

Considerăm caracteristica X ce urmează legea normală $\mathcal{N}\,(m, \sigma)$ cu $m \in \mathbb{R}$ necunoscut şi $\sigma > 0$ necunoscut. Determinăm un interval de încredere pentru dispersia teoretică σ^2 a caracteristicii X.

Statistica ce se consideră în acest caz este

$$\chi^2 = \frac{(n-1)\,\bar{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2,$$

care, conform Proprietăţii 4.2.25, urmează legea χ^2 cu $n-1$ grade de libertate.

Se determină intervalul numeric $\left(\chi_1^2, \chi_2^2\right)$ astfel înât

$$P\left(\chi^2 \in \left(\chi_1^2, \chi_2^2\right)\right) = F_{n-1}\left(\chi_2^2\right) - F_{n-1}\left(\chi_1^2\right) = 1-\alpha,$$

unde

$$F_m(x) = \frac{1}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)}\int_0^x t^{\frac{m}{2}-1}e^{-\frac{t}{2}}dt, \quad x>0,$$

este funcţia de repartiţie a legii χ^2 cu m grade de libertate şi este tabelată, pentru anumite valori, în *Anexa III*.

Dacă se alege $\chi_1^2 = \chi^2_{n-1,\frac{\alpha}{2}}$ şi $\chi_2^2 = \chi^2_{n-1,1-\frac{\alpha}{2}}$, adică astfel încât

$$F_{n-1}\left(\chi^2_{n-1,\frac{\alpha}{2}}\right) = \frac{\alpha}{2} \quad \text{şi} \quad F_{n-1}\left(\chi^2_{n-1,1-\frac{\alpha}{2}}\right) = 1-\frac{\alpha}{2},$$

se obţine $P\left(\bar{\sigma}^2 < \sigma^2 < \bar{\sigma}^2\right) = 1-\alpha$, unde

$$\begin{aligned}
\bar{\sigma}_1^2 &= \bar{\sigma}_1^2(X_1, X_2, \ldots, X_n) = \frac{(n-1)\,\bar{\sigma}^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}},\\
\bar{\sigma}_2^2 &= \bar{\sigma}_2^2(X_1, X_2, \ldots, X_n) = \frac{(n-1)\,\bar{\sigma}^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\cdot
\end{aligned} \tag{5.3.3}$$

Observaţia 5.3.28. Modul geometric de obţinere a cuantilei $\chi^2_{n-1,\gamma}$ este prezentat în Figura 5.3, dacă se consideră graficele funcţiei de repartiţie şi densităţii de probabilitate a legii χ^2 cu $n-1$ grade de libertate.

Exemplul 5.3.29. Fie X caracteristica ce reprezintă timpul de producere a unei reacţii chimice, măsurat în secunde. Dacă X urmează legea normală $\mathcal{N}(m,\sigma)$ şi având o selecţie repetată de volum $n=11$, cu datele de selecţie 4.21, 4.03, 3.99, 4.05, 3.89, 3.98 4.01, 3.92, 4.23, 3.85, 4.20, vom determina intervalul de încredere pentru dispersia $\sigma^2 = Var(X)$ şi pentru abaterea standard $\sigma = \sqrt{Var(X)}$, cu probabilitatea de încredere 0.95.

Se consideră statistica

$$\chi^2 = \frac{(n-1)\,\bar{\sigma}^{\,2}}{\sigma^2}, \quad \text{unde} \quad \bar{\sigma}^{\,2} = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2, \quad \bar{X} = \frac{1}{n}\sum_{k=1}^{n}X_k,$$

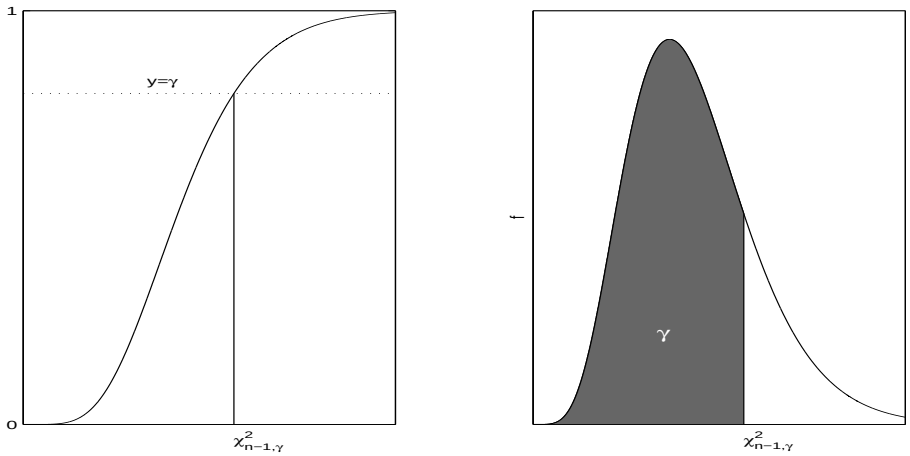

Figura 5.3: Determinarea cuantilei $\chi^2_{n-1,\gamma}$

care urmează legea χ^2 cu $n-1$ grade de libertate. Extremităţile intervalului de încredere pentru σ^2 vor fi date prin (5.3.3).

Pentru determinarea valorilor numerice ale acestor intervale de încredere, calculăm:

$$\bar{x} = \frac{1}{11}\sum_{k=1}^{11} x_k = \frac{1}{11}\left(4.21 + 4.03 + \cdots + 4.20\right) = 4.033;$$

$$\bar{\sigma}^{\,2} = \frac{1}{10}\sum_{k=1}^{11}\left(x_k - \bar{x}\,\right)^2 = 0.017; \quad \chi^2_{10;0.975} = 20.5 \quad \text{şi} \quad \chi^2_{10;0.025} = 3.25.$$

Aşadar, intervalele de încredere pentru σ^2 şi σ sunt

$$\left(\frac{10\cdot 0.017}{20.5}\,;\,\frac{10\cdot 0.017}{3.25}\right) = \left(0.008\,;\,0.052\right),$$

respectiv

$$\left(\sqrt{\frac{10\cdot 0.017}{20.5}}\,;\,\sqrt{\frac{10\cdot 0.017}{3.25}}\right) = \left(0.091\,;\,0.229\right).$$

Programul 5.3.30. Un program Matlab, care să rezolve această problemă, ar putea fi următorul:

```
x=[4.21,4.03,3.99,4.05,3.89,3.98,4.01,3.92,4.23,3.85,4.20];
va=var(x); c1=chi2inv(0.025,10); c2=chi2inv(0.975,10);
```

```
v1=10*va/c2; v2=10*va/c1; s1=sqrt(v1); s2=sqrt(v2);
fprintf(' (v1,v2)=(%6.3f,%6.3f)\n',v1,v2)
fprintf(' (s1,s2)=(%6.3f,%6.3f)',s1,s2)
```

În urma executării, se obţin intervalele de încredere pentru dispersie şi respectiv pentru abaterea standard:

```
(v1,v2)=( 0.008, 0.052)
(s1,s2)=( 0.091, 0.229)
```

Interval de încredere pentru raportul dispersiilor

Se consideră caracteristicile independente X' şi X'' fiecare urmând legea normală, respectiv $\mathcal{N}(m', \sigma')$ şi $\mathcal{N}(m'', \sigma'')$. Relativ la cele două caracteristici se consideră câte o selecţie repetată, respectiv de volume n' şi n''. Vom determina un interval de încredere pentru raportul $\sigma'^2 \Big/ \sigma''^2$ corespunzător probabilităţii de încredere $1-\alpha$ dată.

Pentru aceasta se consideră statistica

$$F = \frac{\bar{\sigma}''^2}{\sigma''^2} \Bigg/ \frac{\bar{\sigma}'^2}{\sigma'^2},$$

care, conform Proprietăţii 4.2.30, urmează legea Fisher–Snedecor cu $m = n'' - 1$ şi $n = n' - 1$ grade de libertate.

Se determină intervalul numeric (f_1, f_2) astfel încât

$$P\big(F \in (f_1, f_2)\big) = F_{m,n}(f_2) - F_{m,n}(f_1) = 1 - \alpha,$$

unde

$$F_{m,n}(x) = \left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \int_0^x t^{\frac{m}{2}-1} \left(1 + \frac{m}{n}t\right)^{-\frac{m+1}{2}} dt, \quad x > 0,$$

este funcţia de repartiţie a legii Fisher–Snedecor cu m şi n grade de libertate şi care este tabelată pentru anumite valori în *Anexa IV.*

Dacă se alege $f_1 = f_{m,n;\frac{\alpha}{2}}$ şi $f_2 = f_{m,n;1-\frac{\alpha}{2}}$, adică astfel încât

$$F_{m,n}\left(f_{m,n;\frac{\alpha}{2}}\right) = \frac{\alpha}{2} \quad \text{şi} \quad F_{m,n}\left(f_{m,n;1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2},$$

atunci

$$P\left(f_{m,n;\frac{\alpha}{2}} \frac{\bar{\sigma}'^2}{\bar{\sigma}''^2} < \frac{\sigma'^2}{\sigma''^2} < f_{m,n;1-\frac{\alpha}{2}} \frac{\bar{\sigma}'^2}{\bar{\sigma}''^2}\right) = 1 - \alpha. \tag{5.3.4}$$

Această relaţie pune în evidenţă intervalul de încredere pentru raportul celor două dispersii.

Observaţia 5.3.31. Modul geometric de obţinere a cuantilei $f_{m,n;\gamma}$ este prezentat în Figura 5.4, dacă se consideră graficele funcţiei de repartiţie şi densităţii de probabilitate a legii Fisher–Snedecor cu m şi n grade de libertate.

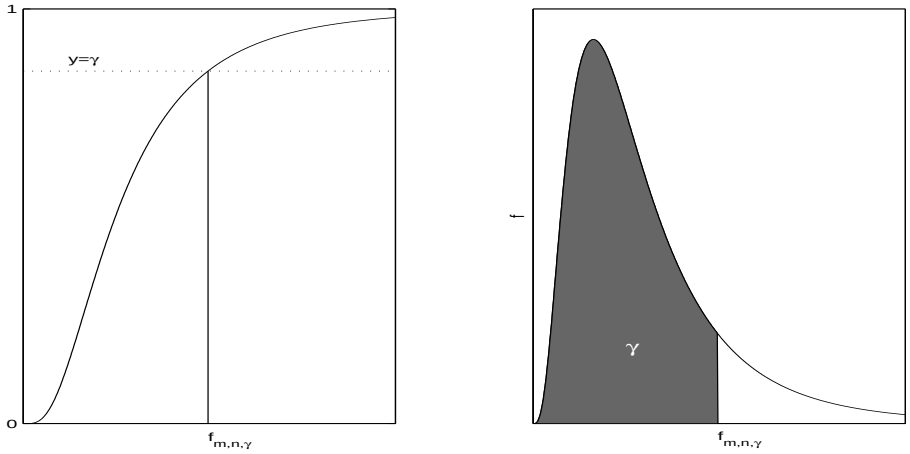

Figura 5.4: Determinarea cuantilei $f_{m,n;\gamma}$

Exemplul 5.3.32. Se cercetează precizia cu care două maşini produc conserve de acelaşi tip. Pentru aceasta, se consideră câte un eşantion din conservele produse de cele două maşini şi se măsoară greutatea acestora. Fie X' greutatea în grame a unei conserve produsă de prima maşină, respectiv X'' pentru a doua maşină. Măsurătorile obţinute sunt:

X' :	1021	980	988	1017	1005	998	1014	985	995	1004	1030
		1015	995	1023	1008	1013					
X'' :	1003	988	993	1013	1006	1002	1014	997	1002	1010	975

Să se calculeze un interval de încredere pentru raportul abaterilor standard, adică pentru $\frac{\sigma'}{\sigma''} = \frac{\sqrt{Var(X')}}{\sqrt{Var(X'')}}$, folosind probabilitatea de risc $\alpha = 0.05$.

Vom considera că cele două caracteristici X' şi X'' sunt independente şi că urmează legile normale $\mathcal{N}(m', \sigma')$ şi respectiv $\mathcal{N}(m'', \sigma'')$.

Statistica ce se utilizează pentru stabilirea intervalului de încredere pentru raportul σ' / σ'' este

$$F = \frac{\bar{\sigma}''^2}{\sigma''^2} \Big/ \frac{\bar{\sigma}'^2}{\sigma'^2},$$

care urmează legea Fisher–Snedecor cu $m = n'' - 1 = 10$ şi $n = n' - 1 = 15$ grade de libertate.

Intervalul de încredere este dat prin (5.3.4).

Pentru aceasta avem că

$$\bar{x}\,' = \frac{1}{n'}\sum_{k=1}^{n'} x'_k = \frac{1}{16}\,(1021 + 980 + \cdots + 1013) = 1005.7,$$

$$\bar{x}\,'' = \frac{1}{n''}\sum_{k=1}^{n''} x''_k = \frac{1}{11}\,(1003 + 988 + \cdots + 975) = 1000.3,$$

$$\bar{\sigma}\,'^2 = \frac{1}{n'-1}\sum_{k=1}^{n'}\left(x'_k - \bar{x}\,'\right)^2 = \frac{1}{15}\left[(1021 - 1005.7)^2 + \cdots + (1013 - 1005.7)^2\right]$$
$$= 210.63,$$

$$\bar{\sigma}\,''^2 = \frac{1}{n''-1}\sum_{k=1}^{n''}\left(x''_k - \bar{x}\,''\right)^2 = \frac{1}{10}\left[(1003 - 1000.3)^2 + \cdots + (975 - 1000.3)^2\right]$$
$$= 134.42.$$

Pe de altă parte, se obţin cuantilele

$$f_{m,n;\frac{\alpha}{2}} = f_{10,15;0.025} = 0.28, \quad f_{m,n;1-\frac{\alpha}{2}} = f_{10,15;0.975} = 3.06,$$

pe baza *Anexei IV*. Facem observaţia că în acest calcul s-a avut în vedere că

$$f_{m,n;\frac{\alpha}{2}} = \frac{1}{f_{n,m;1-\frac{\alpha}{2}}}, \quad \text{adică} \quad f_{10,15;0.025} = \frac{1}{f_{15,10;0.975}} = \frac{1}{3.52} = 0.28.$$

Se ajunge astfel la faptul că

$$\frac{\sigma'}{\sigma''} \in \left(\sqrt{0.28 \cdot \frac{210.63}{134.42}}, \sqrt{3.06 \cdot \frac{210.63}{134.42}}\right) = (0.66, 2.19)\,.$$

La acelaşi rezultat se ajunge şi dacă se consideră statistica $\frac{1}{F}$.

Programul 5.3.33. Programul Matlab, care urmează, calculează intervalele de încredere pentru raportul dispersiilor, respectiv al abaterilor standard, pentru datele considerate în problema precedentă.

```
x1=[1021,980,988,1017,1005,998,1014,985,995,1004,...
    1030,1015,995,1023,1008,1013];
x2=[1003,988,993,1013,1006,1002,1014,997,1002,1010,975];
v1=var(x1); v2=var(x2); r=v1/v2;
```

```
f1=finv(0.025,10,15); f2=finv(0.975,10,15);
r1=f1*r; r2=f2*r; s1=sqrt(r1); s2=sqrt(r2);
fprintf(' (r1,r2)=(%6.3f,%6.3f)\n',r1,r2)
fprintf(' (s1,s2)=(%6.3f,%6.3f)',s1,s2)
```

În urma executării programului, s–a obţinut:

```
(r1,r2)=( 0.445, 4.795)
(s1,s2)=( 0.667, 2.190)
```

Interval de încredere pentru diferenţa mediilor

Caracteristicile independente X' şi X'' urmează respectiv legile normale $\mathcal{N}(m';\sigma')$ şi $\mathcal{N}(m'';\sigma'')$. Folosind câte o selecţie repetată de volum n' şi n'' pentru cele două caracteristici, vrem să determinăm un interval de încredere pentru diferenţa $m'-m''$.

Distingem următoarele trei situaţii:

A. abaterile standard ale celor două caracteristici sunt cunoscute,

B. abaterile standard sunt necunoscute, dar se ştie că sunt egale,

C. abaterile standard sunt necunoscute şi diferite,

pe care le tratăm pe rând în continuare.

A. *Abaterile standard σ' şi σ'' sunt cunoscute.*

Se consideră statistica

$$Z=\frac{\left(\bar{X}'-\bar{X}''\right)-(m'-m'')}{\sqrt{\frac{\sigma'^2}{n'}+\frac{\sigma''^2}{n''}}}, \tag{5.3.5}$$

care urmează legea normală $\mathcal{N}(0,1)$ (a se vedea Observaţia 4.2.29). Astfel, pentru probabilitatea de risc $\alpha\in(0,1)$ dată, se poate determina intervalul numeric $(z_1,z_2)=\left(-z_{1-\frac{\alpha}{2}},z_{1-\frac{\alpha}{2}}\right)$ astfel încât $P(z_1<Z<z_2)=1-\alpha$. Anume, $z_{1-\frac{\alpha}{2}}$ se calculează din relaţia $\Phi\left(z_{1-\frac{\alpha}{2}}\right)=\frac{1-\alpha}{2}$, unde

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x \mathrm{e}^{-\frac{t^2}{2}}dt,$$

este funcţia lui Laplace, tabelată în *Anexa I*. Se ajunge astfel la relaţia

$$P\left(\bar{X}'-\bar{X}''-\kappa_\alpha<m'-m''<\bar{X}'-\bar{X}''+\kappa_\alpha\right)=1-\alpha, \tag{5.3.6}$$

unde

$$\kappa_\alpha = z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma'^2}{n'} + \frac{\sigma''^2}{n''}},$$

care pune în evidenţă extremităţile intervalului de încredere pentru diferenţa celor două medii.

B. *Abaterile standard σ' şi σ'' sunt egale cu σ, dar necunoscut.*

Se consideră statistica

$$T = \frac{\left(\bar{X}' - \bar{X}''\right) - (m' - m'')}{\sqrt{(n'-1)\,\bar{\sigma}'^2 + (n''-1)\,\bar{\sigma}''^2}} \sqrt{\frac{n' + n'' - 2}{\frac{1}{n'} + \frac{1}{n''}}}, \tag{5.3.7}$$

unde

$$\bar{\sigma}'^2 = \frac{1}{n'-1} \sum_{k=1}^{n'} \left(X'_k - \bar{X}'\right), \quad \bar{\sigma}''^2 = \frac{1}{n''-1} \sum_{k=1}^{n''} \left(X'_k - \bar{X}'\right)$$

şi care urmează legea Student cu $m = n' + n'' - 2$ grade de libertate (a se vedea Proprietatea 4.2.28).

Ca la punctul precedent se obţin extremităţile intervalului de încredere

$$\bar{m}_{1,2} = \bar{X}' - \bar{X}'' \pm t_{m,1-\frac{\alpha}{2}} \sqrt{\frac{\frac{1}{n'} + \frac{1}{n''}}{n' + n'' - 2}} \sqrt{(n'-1)\,\bar{\sigma}'^2 + (n''-1)\,\bar{\sigma}''^2}, \tag{5.3.8}$$

unde $t_{m,1-\frac{\alpha}{2}}$ este cuantila de ordin $1-\frac{\alpha}{2}$ pentru legea Student cu m grade de libertate.

C. *Abaterile standard σ' şi σ'' sunt diferite şi necunoscute.*

Se consideră statistica

$$T = \frac{\left(\bar{X}' - \bar{X}''\right) - (m' - m'')}{\sqrt{\frac{\bar{\sigma}'^2}{n'} + \frac{\bar{\sigma}''^2}{n''}}}, \tag{5.3.9}$$

care urmează legea Student cu n grade de libertate şi care se calculează prin formula

$$\frac{1}{n} = \frac{c^2}{n'-1} + \frac{(1-c)^2}{n''-1}, \quad \text{unde} \quad c = \frac{\bar{\sigma}'^2}{n'} \Bigg/ \left(\frac{\bar{\sigma}'^2}{n'} + \frac{\bar{\sigma}''^2}{n''}\right). \tag{5.3.10}$$

Ca la punctul precedent se ajunge la extremităţile intervalului de încredere pentru diferenţa celor două medii:

$$\bar{m}_{1,2} = \bar{X}' - \bar{X}'' \pm t_{n,1-\frac{\alpha}{2}} \sqrt{\frac{\bar{\sigma}'^2}{n'} + \frac{\bar{\sigma}''^2}{n''}}. \tag{5.3.11}$$

Exemplul 5.3.34. Maşinile M_1 şi M_2 ambalează carne în pachete de 1000 grame. Greutatea cutiilor este o caracteristică X', ce urmează legea normală $\mathcal{N}(m', \sigma')$ şi respectiv o caracteristică X'', ce urmează legea normală $\mathcal{N}(m'', \sigma'')$.

Cântărind 100 de pachete din cele produse de maşina M_1 s–a obţinut valoarea mediei de selecţie $\bar{x}\,' = 1007$ grame, iar din cântărirea a 150 de pachete de la maşina M_2 s–a obţinut $\bar{x}\,'' = 1002$ grame.

Folosind probabilitatea de încredere 0.98, să determinăm intervalul de încredere pentru diferenţa $m' - m''$, dacă se ştie că abaterile standard sunt $\sigma' = 3$ şi $\sigma'' = 4$.

Se foloseşte statistica (5.3.5), care urmează legea normală $\mathcal{N}(0, 1)$. Astfel, intervalul de încredere pentru diferenţa $m' - m''$ este dat prin (5.3.6), unde $z_{1-\frac{\alpha}{2}}$ se determină astfel ca $\Phi\left(z_{1-\frac{\alpha}{2}}\right) = \frac{1-\alpha}{2} = 0.49$. Folosind *Anexa I*, obţinem $z_{1-\frac{\alpha}{2}} = 2.33$.

De asemenea, avem că

$$\sqrt{\frac{\sigma'^{\,2}}{n'} + \frac{\sigma''^{\,2}}{n''}} = \sqrt{\frac{9}{100} + \frac{16}{150}} = \sqrt{\frac{59}{300}} = 0.4435.$$

Astfel, intervalul de încredere pentru diferenţa $m' - m''$ este

$$\left(\left(\bar{x}\,' - \bar{x}\,''\right) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma'^{\,2}}{n'} + \frac{\sigma''^{\,2}}{n''}}\,;\ \left(\bar{x}\,' - \bar{x}\,''\right) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma'^{\,2}}{n'} + \frac{\sigma''^{\,2}}{n''}}\right)$$
$$= (5 - 2.33 \cdot 0.4435\,;\ 5 + 2.33 \cdot 0.4435) = (3.97\,;\ 6.03)\,.$$

Programul 5.3.35. Executarea programului Matlab:

```
z1=norminv(0.01,0,1); z2=norminv(0.99,0,1);
s=sqrt(9/100+16/150); d=1007-1002;
m1=d+z1*s; m2=d+z2*s;
fprintf(' (m1,m2)=(%6.3f,%6.3f)',m1,m2)
```

conduce la rezultatul:

```
 (m1,m2)=( 3.968, 6.032)
```

Exemplul 5.3.36. Fie caracteristica X' ce urmează legea normală $\mathcal{N}(m', \sigma)$ şi care reprezintă vânzările în milioane lei pe săptămână la magazinele alimentare în oraşul $\mathcal{A}$ şi X'' vânzările în milioane lei la magazinele alimentare din oraşul $\mathcal{B}$ şi care urmează legea normală $\mathcal{N}(m'', \sigma)$. S–au efectuat două sondaje, respectiv pentru X' şi X'' şi s–au obţinut următoarele date de selecţie:

X' : $226.5,\ 224.1,\ 218.6,\ 220.1,\ 228.8,\ 229.6,\ 222.5$;

X'' : $221.5,\ 230.2,\ 223.4,\ 224.3,\ 230.8,\ 223.8$.

Cu probabilitatea de încredere 0.95, vrem să se construim un interval de încredere pentru diferenţa $m' - m''$, dacă $\sigma > 0$ este necunoscut.

Folosind statistica (5.3.7), care urmează legea Student cu $n = n' + n'' - 2 = 11$ grade de libertate, se va construi intervalul de încredere pentru $m' - m''$. Anume, acest interval de încredere este precizat prin (5.3.8).

Pentru a determina valoarea numerică a intervalului de încredere, se calculează pe rând

$$\bar{x}' = \frac{1}{7}(226.5 + 224.1 + 218.6 + 220.1 + 228.8 + 229.6 + 222.5) = 224.314,$$

$$\bar{x}'' = \frac{1}{6}(221.5 + 230.2 + 223.4 + 224.3 + 230.8 + 223.8) = 225.667,$$

$$\bar{\sigma}'^2 = \frac{1}{6}\sum_{k=1}^{7}\left(x'_k - \bar{x}'\right)^2 = 17.765, \quad \bar{\sigma}''^2 = \frac{1}{5}\sum_{k=1}^{6}\left(x''_k - \bar{x}''\right)^2 = 14.951,$$

$$s = \sqrt{\frac{\frac{1}{7}+\frac{1}{6}}{11}(6 \cdot 17.765 + 5 \cdot 14.951)} = 2.259.$$

De asemenea, din *Anexa II*, pentru $1 - \alpha = 0.95$ şi $n = 11$, obţinem $t_{n,1-\frac{\alpha}{2}} = 2.201$, astfel că intervalul de încredere pentru $m' - m''$ va fi

$$\left(\left(\bar{x}' - \bar{x}''\right) - t_{n,1-\frac{\alpha}{2}} \cdot s\,;\, \left(\bar{x}' - \bar{x}''\right) + t_{n,1-\frac{\alpha}{2}} \cdot s\right)$$
$$= (-1.353 - 2.201 \cdot 2.259\,;\, -1.353 + 2.201 \cdot 2.259) = (-6.32\,;\, 3.62)\,.$$

Programul 5.3.37. Prin executarea programului Matlab:

```
x1=[226.5,224.1,218.6,220.1,228.8,229.6,222.5];
x2=[221.5,230.2,223.4,224.3,230.8,223.8];
ma1=mean(x1); ma2=mean(x2); md=ma1-ma2;
v1=var(x1); v2=var(x2);
t1=tinv(0.025,11); t2=tinv(0.975,11);
s=sqrt((1/7+1/6)/11)*sqrt(6*v1+5*v2);
m1=md+t1*s; m2=md+t2*s;
fprintf(' (m1,m2)=(%6.3f,%6.3f)\n',m1,m2)
```

regăsim intervalul de încredere pentru diferenţa valorilor medii:

```
(m1,m2)=(-6.324, 3.619)
```

Exemplul 5.3.38. Se consideră două aparate de îmbuteliat vin în sticle de 750 ml. Fie caracteristica X' ce reprezintă cantitatea de vin (în ml) îmbuteliată de primul aparat într-o sticlă şi respectiv X'' aceeaşi caracteristică pentru al doilea aparat. Pentru compararea modului de îmbutliere pentru cele două aparate, se consideră câte o selecţie din sticlele îmbuteliate de cele două aparate, respectiv

X' : 746 743 748 748 750 745 744 753 750 751 743 747 749 742
X'' : 749 748 748 753 750 749 747 745 744 751

Vom construi un interval de încredere pentru diferenţa $m' - m'' = E(X') - E(X'')$, folosind probabilitatea de risc $\alpha = 0.05$. Vom considera că cele două caracteristici X' şi X'' sunt independente şi că urmează legile normale $\mathcal{N}(m', \sigma')$ şi $\mathcal{N}(m'', \sigma'')$.

Distingem două cazuri, unul când se ştie că $\sigma' = \sigma''$, caz în care se foloseşte statistica (5.3.7), care urmează legea Student cu $n = n' + n'' - 2$ grade de libertate, respectiv când se ştie că $\sigma' \neq \sigma''$, caz în care se foloseşte statistica (5.3.9), care urmează legea Student cu n grade de libertate. Numărul n al gradelor de libertate se calculează, în acest caz, cu formula (5.3.10)

În cazul $\sigma' = \sigma''$, extremităţile intervalului de încredere pentru diferenţa $m' - m''$ sunt date prin formulele

$$\bar{m}_1 = \left(\bar{x}' - \bar{x}''\right) - t_{n,1-\frac{\alpha}{2}} \sqrt{\frac{\frac{1}{n'} + \frac{1}{n''}}{n' + n'' - 2}} \cdot \sqrt{(n'-1)\,\bar{\sigma}'^2 + (n''-1)\,\bar{\sigma}''^2},$$

$$\bar{m}_2 = \left(\bar{x}' - \bar{x}''\right) + t_{n,1-\frac{\alpha}{2}} \sqrt{\frac{\frac{1}{n'} + \frac{1}{n''}}{n' + n'' - 2}} \cdot \sqrt{(n'-1)\,\bar{\sigma}'^2 + (n''-1)\,\bar{\sigma}''^2}.$$

Cuantila $t_{n,1-\frac{\alpha}{2}} = t_{22,0.975} = 2.07$, s-a determinat, conform *Anexei II*, din relaţia $F_{22}(t_{22,0.975}) = 1 - \frac{\alpha}{2} = 0.975$.

Pe de altă parte

$$\bar{x}' = \frac{1}{14}(746 + 743 + \cdots + 742) = 747.1,$$

$$\bar{x}'' = \frac{1}{10}(749 + 748 + \cdots + 751) = 748.4,$$

$$(n'-1)\,\bar{\sigma}'^2 = \sum_{k=1}^{n'} \left(x'_k - \bar{x}'\right) = (746 - 747.1)^2 + \cdots + (742 - 747.1)^2 = 146.94,$$

$$(n''-1)\,\bar{\sigma}''^2 = \sum_{k=1}^{n''} \left(x''_k - \bar{x}''\right) = (749 - 748.4)^2 + \cdots + (751 - 748.4)^2 = 64.4.$$

Astfel avem că

$$\bar{m}_1 = (747.1 - 748.4) - 2.07\sqrt{\frac{\frac{1}{14} + \frac{1}{10}}{14 + 10 - 2}}\sqrt{146.94 + 64.4} = -4.0,$$

$$\bar{m}_2 = (747.1 - 748.4) + 2.07\sqrt{\frac{\frac{1}{14} + \frac{1}{10}}{14 + 10 - 2}}\sqrt{146.94 + 64.4} = 1.3.$$

Pentru cazul $\sigma' \neq \sigma''$, calculăm prima dată numărul n al gradelor de libertate. Astfel avem că

$$c = \frac{\bar{\sigma}'^2}{n'} \Bigg/ \left(\frac{\bar{\sigma}'^2}{n'} + \frac{\bar{\sigma}''^2}{n''} \right) = \frac{11.30}{14} \Bigg/ \left(\frac{11.3}{14} + \frac{7.16}{10} \right) = \frac{0.807}{0.807 + 0.716} = 0.53,$$

iar apoi

$$\frac{1}{n} = \frac{c^2}{n' - 1} + \frac{(1-c)^2}{n'' - 1} = \frac{0.281}{13} + \frac{0.221}{9} = 0.046,$$

de unde avem că $n = 22$.

Extremităţile intervalului de încredere sunt date prin

$$\bar{m}_1 = \left(\bar{x}' - \bar{x}'' \right) - t_{n,1-\frac{\alpha}{2}} \sqrt{\frac{\bar{\sigma}'^2}{n'} + \frac{\bar{\sigma}''^2}{n''}},$$

$$\bar{m}_2 = \left(\bar{x}' - \bar{x}'' \right) + t_{n,1-\frac{\alpha}{2}} \sqrt{\frac{\bar{\sigma}'^2}{n'} + \frac{\bar{\sigma}''^2}{n''}}.$$

Având în vedere că $t_{22,0.975} = 2.07$ rezultă că

$$\bar{m}_1 = (747.1 - 748.4) - 2.07\sqrt{0.807 - 0.716} = -3.9,$$
$$\bar{m}_2 = (747.1 - 748.4) + 2.07\sqrt{0.807 - 0.716} = 1.2.$$

Programul 5.3.39. Programul Matlab, care urmează, rezolvă problema precedentă, adică, va determina intervale de încredere pentru diferenţa mediilor, în cele două cazuri, când se consideră dispersii egale şi necunoscute, respectiv dispersii diferite necunoscute.

```
x1=[746,743,748,748,750,745,744,753,750,751,743,747,749,742];
x2=[749,748,748,753,750,749,747,745,744,751];
ma1=mean(x1); ma2=mean(x2); md=ma1-ma2;
v1=var(x1); v2=var(x2);
t1=tinv(0.025,22); t2=tinv(0.975,22);
s=sqrt((1/14+1/10)/22)*sqrt(13*v1+9*v2);
m1=md+t1*s; m2=md+t2*s;
fprintf(' (m1,m2)=(%6.3f,%6.3f)\n',m1,m2)
c=(v1/14)/(v1/14+v2/10);n=c^2/13+(1-c)^2/9;
n=ceil(1/n); t1=tinv(0.025,n); t2=tinv(0.975,n);
s=sqrt(v1/14+v2/10); m1=md+t1*s; m2=md+t2*s;
fprintf(' (m1,m2)=(%6.3f,%6.3f)\n',m1,m2)
```

După executarea programului, se obţin cele două intervale de încredere:

```
(m1,m2)=(-3.990, 1.333)
(m1,m2)=(-3.888, 1.231)
```

5.3.5 Metoda intervalelor de încredere pentru selecţii mari

Fie caracteristica X cercetată, cu legea de probabilitate $f(x;\theta)$, unde $\theta \in A \subset \mathbb{R}$ este un parametru necunoscut. Considerăm o selecţie repetată de volum n relativă la caracteristica X, pentru care avem variabilele de selecţie $X_1, X_2, \ldots, X_n$.

Proprietatea 5.3.40. *Dacă se consideră variabilele aleatoare $Y_1, Y_2, \ldots, Y_n$ definite prin relaţia*

$$Y_k = \frac{\partial \ln f(X_k;\theta)}{\partial \theta},$$

şi pentru care există dispersia $Var(Y_k) = d^2 > 0$, atunci statistica

$$Z = \frac{1}{d\sqrt{n}} \sum_{k=1}^{n} Y_k = \frac{1}{d\sqrt{n}} \sum_{k=1}^{n} \frac{\partial \ln f(X_k;\theta)}{\partial \theta},$$

pentru $n \to \infty$, urmează legea normală $\mathcal{N}(0,1)$.

Demonstraţie. Variabilele aleatoare X_k, $k = \overline{1,n}$, fiind independente şi identic repartizate, rezultă că şi variabilele aleatoare Y_k, $k = \overline{1,n}$, sunt independente şi identic repartizate. Conform teoremei limită centrală, avem că

$$Y_{(n)} = \frac{1}{d\sqrt{n}} \sum_{k=1}^{n} \big(Y_k - E(Y_k)\big)$$

converge în repartiţie la legea normală $\mathcal{N}(0,1)$.

Deoarece

$$E(Y_k) = E\left(\frac{\partial \ln f(X_k;\theta)}{\partial \theta}\right) = 0,$$

rezultă că $Y_{(n)} = Z$, ceea ce încheie demonstraţia. □

Pentru probabilitatea de încredere $1-\alpha$ dată se va determina intervalul numeric $\left(-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}\right)$ astfel încât $P\left(Z \in \left(-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}\right)\right) = 2\Phi\left(z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$. Ceea ce revine la determinarea cuantilei $z_{1-\frac{\alpha}{2}}$ astfel încât $\Phi\left(z_{1-\frac{\alpha}{2}}\right) = \frac{1-\alpha}{2}$ · Funcţia Φ a lui Laplace este aici, cea tabelată în *Anexa I.*

Prin operaţii algebrice se înlocuieşte inegalitatea $|Z| < z_{1-\frac{\alpha}{2}}$, cu dubla inegalitate echivalentă de forma $\bar{\theta}_1(X_1, X_2, \ldots, X_n) < \theta < \bar{\theta}_2(X_1, X_2, \ldots, X_n)$, care defineşte intervalul de încredere pentru parametrul θ.

Aplicaţia 5.3.41. Fie caracteristica X ce ia numai valorile 1 şi 0 cu probabilităţile p şi respectiv $1-p$, adică are funcţia de frecvenţă $f(x;p) = p^x(1-p)^{1-x}$, $x = 0, 1$, unde $p \in (0,1)$ este un parametru necunoscut.

Vom folosi metoda intervalelor de încredere pentru selecţii mari în vederea estimării parametrului p. Considerăm o selecţie repetată de volum (mare) n şi probabilitatea de încredere $1-\alpha$. Deoarece $\ln f(x;p) = x\ln p + (1-x)\ln(1-p)$, avem că

$$\frac{\partial \ln f(x;p)}{\partial p} = \frac{x}{p} - \frac{1-x}{1-p} = \frac{x-p}{p(1-p)},$$

şi prin urmare se obţine statistica

$$Z = \frac{1}{d\sqrt{n}} \sum_{k=1}^{n} \frac{X_k - p}{p(1-p)} = \frac{\sqrt{n}}{p(1-p)\,d}\left(\bar{X} - p\right).$$

Ştiind că $Var(X) = p(1-p)$ rezultă

$$\begin{aligned} d^2 =& Var\left(\frac{\partial \ln f(X_k;p)}{\partial p}\right) = Var\left(\frac{X_k - p}{p(1-p)}\right) = \frac{1}{p^2(1-p)^2} Var(X_k) \\ =& \frac{1}{p^2(1-p)^2}\, p(1-p) = \frac{1}{p(1-p)}. \end{aligned}$$

Statistica Z devine aşadar

$$Z = \frac{\sqrt{n}}{\sqrt{p(1-p)}}\left(\bar{X} - p\right)$$

şi care urmează legea normală $\mathcal{N}(0,1)$, când $n \to \infty$.

Pentru α dat, se determină $z_{1-\frac{\alpha}{2}} = z$ astfel încât $P(|Z| < z) = 1-\alpha$.

Putem scrie că $|Z| < z$ este echivalentă cu

$$\frac{n\left(\bar{X} - p\right)^2}{p(1-p)} < z^2 \quad \text{sau} \quad \left(n + z^2\right)p^2 - \left(2n\bar{X} + z^2\right)p + n\bar{X}^2 < 0.$$

Discriminantul trinomului este pozitiv, anume $\Delta = z^2\left[z^2 + 4n\bar{X}\left(1-\bar{X}\right)\right] > 0$, deci inecuaţia în p are soluţia de forma unui interval $(\bar{p}_1, \bar{p}_2)$ şi care va reprezenta intervalul de încredere pentru parametrul p.

Astfel extremităţile intervalului de încredere au următoarele expresii:

$$\bar{p}_1 = \frac{\left(2\bar{X} + \frac{z^2}{n}\right) - \sqrt{\frac{z^4}{n^2} + 4\bar{X}\frac{z^2}{n} - 4\bar{X}^2\frac{z^2}{n}}}{2\left(1 + \frac{z^2}{n}\right)},$$

$$\bar{p}_2 = \frac{\left(2\bar{X} + \frac{z^2}{n}\right) + \sqrt{\frac{z^4}{n^2} + 4\bar{X}\frac{z^2}{n} - 4\bar{X}^2\frac{z^2}{n}}}{2\left(1 + \frac{z^2}{n}\right)}.$$

Aceste formule de calcul sunt complicate. Având în vedere că aceste formule au fost deduse pentru n mare, să zicem $n > 30$, când se obţin rezultate bune, putem face simplificări ale acestor formule. În primul rând putem folosi următoarea scriere asimptotică

$$\frac{2\bar{X} + \frac{z^2}{n}}{2\left(1 + \frac{z^2}{n}\right)} \approxeq \bar{X},$$

iar apoi

$$\sqrt{\frac{z^4 + 4\bar{X}nz^2 - 4\bar{X}^2nz^2}{4\left(n + z^2\right)^2}} = \sqrt{\frac{z^4 + 4\bar{X}nz^2 - 4\bar{X}^2nz^2}{4n^2 + 8nz^2 + 4z^4}}$$
$$\approxeq \sqrt{\frac{\bar{X}nz^2 - \bar{X}^2nz^2}{n^2}} = z\sqrt{\frac{\bar{X}\left(1 - \bar{X}\right)}{n}}.$$

S-a ajuns, în acest mod, la intervalul de încredre pentru p

$$(\bar{p}_1, \bar{p}_2) = \left(\bar{X} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}\left(1 - \bar{X}\right)}{n}}, \bar{X} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}\left(1 - \bar{X}\right)}{n}}\right).$$

Observaţia 5.3.42. Dacă se doreşte să se determine parametrul p cu o incertitudine $\pm\Delta p$, pentru o probabilitate de încredere $1 - \alpha$, atunci volumul selecţiei se determină în felul următor. Având în vedere că

$$z_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{X}\left(1 - \bar{X}\right)}{n}} \leqslant \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}},$$

condiţia $\frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}} \leqslant \Delta p$, echivalentă cu $n \geqslant \frac{z^2_{1-\frac{\alpha}{2}}}{4(\Delta p)^2}$, conduce la volumul optim al selecţiei.

Mai jos prezentăm un tabel cu valorile optime ale volumului selecţiei pentru diferite valori ale nivelului de încredere şi ale lui Δp:

$\Delta p \setminus 1 - \alpha$	0.90	0.95	0.98
0.01	6760	9600	13530
0.02	1700	2400	3380
0.05	270	380	540

5.3.6 Funcţii Matlab privind estimaţia

Sistemul Matlab, prin *Statistics toolbox*, dispune de funcţii corespunzătoare teoriei estimaţiei, pe care le prezentăm în continuare.

Funcţia `histfit`

Funcţia `histfit` are drept rezultat reprezentarea pe aceeaşi figură a histogramei şi a densităţii legii normale $\mathcal{N}(\bar{x}, \bar{\sigma})$.

Lansarea executării funcţiei se face prin una din formele

```
histfit(x)
histfit(x,nc)
```

unde `x` este un vector ce conţine datele ce urmează a fi prelucrate, iar `nc` este numărul claselor. Dacă parametrul `nc` este absent, atunci numărul claselor se consideră ca fiind radicalul volumului datelor, adică radicalul lungimii vectorului `x`.

Programul 5.3.43. Programul următor generează `n` numere aleatoare, ce urmează legea normală $\mathcal{N}(\mu, \sigma)$, după care apelează la funcţia `histfit`, iar graficul obţinut, pentru `n=100`, $\mu = 10$ şi $\sigma = 2$, este prezentat în Figura 5.5.

```
n=input('n='); mu=input('mu=');
s=input('sigma=');x=normrnd(mu,s,1,n);
histfit(x), colormap spring
```

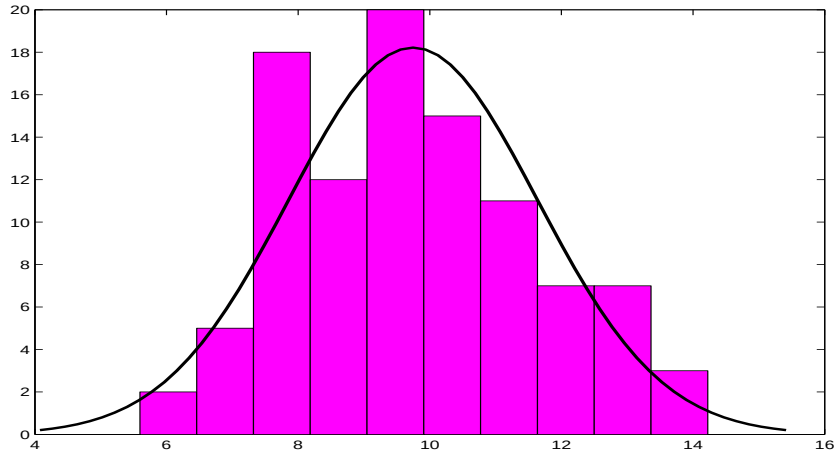

Figura 5.5: Legea $\mathcal{N}(10, 2)$

Funcţia `mle`

Pentru estimarea parametrilor, folosind metoda verosimilităţii maxime, precum şi metoda intervalelor de încredere, poate fi utilizată funcţia `mle`. Apelul funcţiei se poate face prin una din instrucţinile:

```
P=mle('lege',x)
[P,int]=mle('lege',x)
[P,int]=mle('lege',x,alpha)
[P,int]=mle('lege',x,alpha,n)
```

unde `lege` specifică una din legile: `bernoulli`, `unid`, `bino`, `poiss`, `geo`, `unif`, `norm`, `gam`, `exp`, `beta`, `weib` şi `rayl`. Parametrul `x` este un vector, iar în urma executării unei astfel de instrucţiune, `P` şi `int` vor conţine estimatori de verosimilitate maximă, respectiv intervale de încredere, pentru parametrii legii considerate, obţinute pe baza datelor conţinute de `x`.

Parametrul opţional `alpha`, având valoarea implicită `alpha=0.05`, specifică probabilitate de încredere, `1-alpha`, în construirea intervalelor de încredere.

Ultima formă este specifică numai legii binomiale, adică `lege=bino`, parametrul `n` fiind parametrul cel de la legea $\mathcal{B}(\mathtt{n}, p)$. Dacă `x` şi `n` sunt scalari, se returnează în `P` raportul acestora, dacă `x` este vector şi `n` este scalar, se va returna în `P` fiecare element a lui `x` împărţit la `n`, iar dacă `x` şi `n` sunt vectori de aceleaşi dimensiuni, `P` va conţine rapoartele pe componente ale lui `x` şi `n`.

Mai remarcăm că pentru prima dată întâlnim legea lui Bernoulli, care reprezintă cazul particular al legii binomiale, $\mathcal{B}(1, p)$. Dacă `lege=bernoulli`, atunci `x`, desigur, trebuie să conţină numai valorile `0` şi `1`.

Funcţia 5.3.44. Vom scrie o funcţie, care generează `n` numere aleatoare ce urmează una din legile de probabilitate: uniformă discretă, Poisson, uniformă, normală, Gamma, exponenţială, Weibull şi Rayleigh. Se va apela această funcţie pentru obţinerea estimaţiilor de verosimilitate maximă ale parametrilor legii de probabilitate considerate. De asemenea, se va reprezenta grafic, pe aceeaşi figură, funcţia de repartiţie a legii de probabilitate considerată, împreună cu funcţia de repartiţie a aceleaşi legi, dar având parametrii precizaţi înlocuiţi cu estimaţiile acestora.

```
function veros(lege,n)
switch lege
case 'unid'
    P1=input('N=');
case 'poiss'
    P1=input('lambda=');
case 'unif'
    P1(1)=input('a='); P1(2)=input('b=');
case 'norm'
    P1(1)=input('mu='); P1(2)=input('sigma=');
case 'gam'
    P1(1)=input('a='); P1(2)=input('b=');
case 'exp'
    P1=input('mu=');
case 'beta'
    P1(1)=input('a='); P1(2)=input('b=');
case 'weib'
    P1(1)=input('a='); P1(2)=input('b=');
```

```
case 'rayl'
    P1=input('b=');
otherwise
    error('Eroare')
end
if length(P1)==1
    x=random(lege,P1,1,n);
else
    x=random(lege,P1(1),P1(2),1,n);
end
P2=mle(lege,x); xx=min(x):0.01:max(x);
if length(P1)==1
    F1=cdf(lege,xx,P1); F2=cdf(lege,xx,P2);
else
    F1=cdf(lege,xx,P1(1),P1(2)); F2=cdf(lege,xx,P2(1),P2(2));
end
plot(xx,F1,'k-.',xx,F2,'k-')
legend('Functia de repartitie teoretica',...
       'Functia de repartitie estimata',2)
```

Apelul funcţiei `veros` se face prin

```
>>veros('lege',n)
```

unde valorile pentru `n` şi `lege`, fie că sunt precizate în acest apel, fie sunt precizate înainte. De exemplu, comanda

```
>>veros('poiss',50)
```

are ca efect apelul funcţiei pentru legea lui Poisson, iar pe ecran se va cere introducerea parametrului `lambda`, după care pe ecran va fi reprezentat graficul din Figura 5.6, în cazul în care `lambda=7`. Să remarcăm totuşi că la un nou apel, cu aceeaşi parametri, graficul diferă, deoarece sunt generate alte numere aleatoare.

Dacă se consideră comanda

```
>>veros('norm',50)
```

are ca efect apelul funcţiei pentru legea normală, iar pe ecran se va cere introducerea parametrilor `mu` şi `sigma`, după care pe ecran va fi reprezentat graficul din Figura 5.7, în cazul în care `mu=10` şi `sigma=2`.

Funcţiile `normlike`, `gamalike`,`betalike` şi `weiblike`

Funcţiile `normlike`, `gamlike`, `betalike` şi `weiblike` sunt specifice pentru *Statistics toolbox* şi se pot apela cu una din comenzile

```
L=numef(par,x)
[L,cov]=numef(par,x)
```

unde `numef` reprezintă numele uneia din cele patru funcţii.

Parametrul `par` conţine respectiv parametrii legilor normală, gamma, beta şi Weibull, iar `x` este un vector, care conţine datele de selecţie.

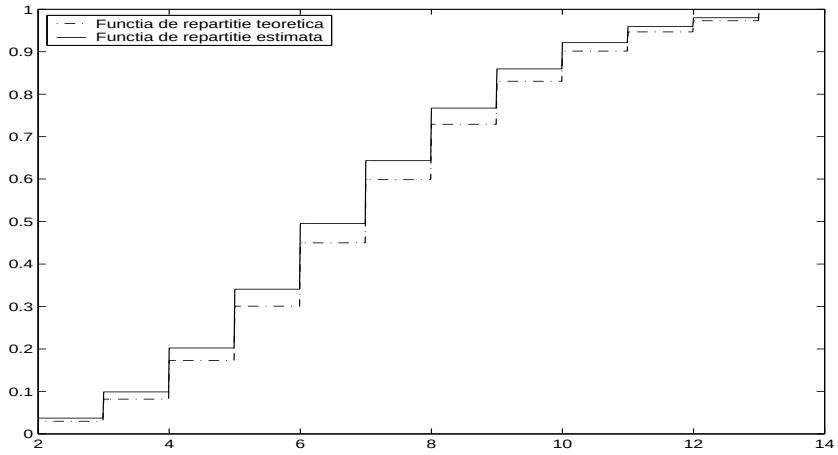

Figura 5.6: Legea $\mathcal{P}o\,(7)$

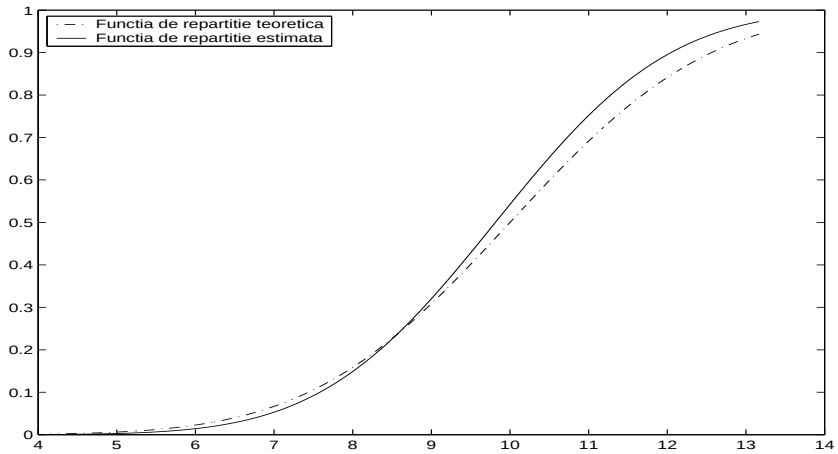

Figura 5.7: Legea $\mathcal{N}\,(10,2)$

Prima formă returnează în `L` valoarea logaritmului funcţiei de verosimilitate de parametrii precizaţi prin `par`, înmulţită cu -1.

Dacă se foloseşte a doua formă, se mai returnează matricea `var`, care reprezintă valoarea asimptotică a matricei covarianţelor estimatorilor parametrilor legii considerate, când s–au considerat parametrii de intrare `par`, ca fiind estimaţiile de verosimilitate maximă ale acestora. Remarcăm faptul că `var` reprezintă inversa matricei informaţiei lui Fisher, când volumul selecţiei tinde la infinit.

Funcţiile `binofit`, `poissfit`, `unifit`, `normfit`, `gamfit`, `expfit`, `betafit`, `raylfit` şi `weibfit`

Funcţiile Matlab din această categorie returnează estimaţii punctuale şi intervale de încredere pentru parametrii respectiv ai legilor de probabilitate binomială , Poisson, uniformă, normală, Gamma, exponenţială, Beta, Rayleigh şi Weibull.

Apelul acestor funcţii nu are o sintaxă generală.

Pentru funcţia `binofit` se pot folosi comenzile

```
P=binofit(x,n)
[P,int]=binofit(x,n)
[P,int]=binofit(x,n,alpha)
```

parametrul `x` fiind un vector, iar `n` precizează parametrul întreg pozitiv al legii binomiale $\mathcal{B}(\mathtt{n}, p)$. Pentru funcţiile `poiss` şi `exp`, respectiv pentru `gam`, `beta`, `rayl` şi `weib`, apelurile sunt de forma:

```
P=numef(x)
[P,int]=numef(x,alpha)
```

unde parametrul `numef` reprezintă numele funcţiei, iar `x` este fie un vector, fie o matrice, pentru prima grupă de funcţii, iar pentru a doua, numai vector. Dacă `x` este matrice, atunci funcţia operează pentru fiecare coloană în parte.

Pentru legile `unif` şi `norm`, apelurile sunt:

```
[par1,par2,int1,int2]=unifit(x)
[par1,par2,int1,int2]=normfit(x,alpha)
```

unde `x` poate fi vector sau matrice. În al doilea caz, operaţiunea se execută pentru fiecare coloană în parte.

De regulă, estimaţiile punctuale returnate, sunt obţinute prin metoda verosimilităţii maxime, folosindu–se datele conţinute în parametrul `x`, în parametrii `P`, `par1` şi `par2`. Intervalele de încredere sunt returnate în parmetrii `int`, `int1` şi `int2`, şi sunt determinate folosind probabilitatea de încredere `1-alpha`. Valoarea implicită pentru `alpha` fiind `alpha=0.05`.

Pentru legile `unif` şi `norm`, în comenzile de apel pot lipsi `int2`, `int1` şi `par2`, caz în care sunt returnate numai valorile parametrilor rămaşi.

Mai remarcăm faptul că `int`, `int1` şi `int2` au câte două componente, extremităţile intervalului de încredere corespunzător.

Funcţia 5.3.45. Vom scrie o funcţie, care generează `n` numere aleatoare, ce urmează una din legile Gamma, Beta, Weibull şi Rayleigh. Folosind aceste numere, se vor da estimaţii punctuale pentru parametrii acestora, precum şi intervale de încredere, considerând o probabilitate de încredere `1-alpha` specificată. Folosind estimaţiile punctuale obţinute pentru parametri, se vor reprezenta grafic densităţile de probabilitate, folosind parametrii estimaţi şi respectiv parametrii consideraţi la generarea numerelor aleatoare. De asemenea, se vor afişa intervalele de încredere obţinute.

```
function estimat(lege,n,alpha)
switch lege
case 'gam'
    a=input('a='); b=input('b=');
    x=gamrnd(a,b,n,1); [P,I]=gamfit(x,alpha);
case 'beta'
    a=input('a='); b=input('b=');
    x=betarnd(a,b,n,1); [P,I]=betafit(x,alpha);
case 'weib'
    a=input('a='); b=input('b=');
    x=weibrnd(a,b,n,1); [P,I]=weibfit(x,alpha);
case 'rayl'
    b=input('b='); x=raylrnd(b,n,1);
    [P,I]=raylfit(x,alpha);
otherwise
    error('Eroare')
end
t=min(x):0.01:max(x);
if length(P)==1
    y=pdf(lege,t,P); yy=pdf(lege,t,b);
    fprintf('  bbarat=%6.3f\n',P)
    fprintf('  (b1,b2)=(%6.3f,%6.3f)\n',I)
else
    y=pdf(lege,t,P(1),P(2)); yy=pdf(lege,t,a,b);
    fprintf('  abarat=%6.3f,  bbarat=%6.3f\n',P)
    fprintf('  (a1,a2)=(%6.3f,%6.3f)\n',I(1,1),I(2,1))
    fprintf('  (b1,b2)=(%6.3f,%6.3f)\n',I(1,2),I(2,2))
end
plot(t,y,'k-',t,yy,'k-.'), plot(t,y,'k-',t,yy,'k-.')
legend('Legea estimata', 'Legea teoretica')
```

Apelul funcţiei `estimat` se face prin

```
>>estimat('lege',n,alpha)
```

unde valorile pentru `n`, `lege` şi `alpha`, fie că sunt precizate în acest apel, fie sunt precizate înainte. De exemplu, comanda

```
>>estimat('beta',500,0.01)
```

are ca efect apelul funcţiei pentru legea Beta, iar pe ecran se va cere introducerea parametrilor `a` şi `b`, după care pe ecran va fi reprezentat graficul din Figura 5.8, în cazul în care `a=3` şi `b=7`. Să remarcăm totuşi că la un nou apel, cu aceeaşi parametri, graficul diferă, deoarece sunt generate alte numere aleatoare. De asemenea, se vor afişate şi estimaţiile punctuale, respectiv intervalele de încredere:

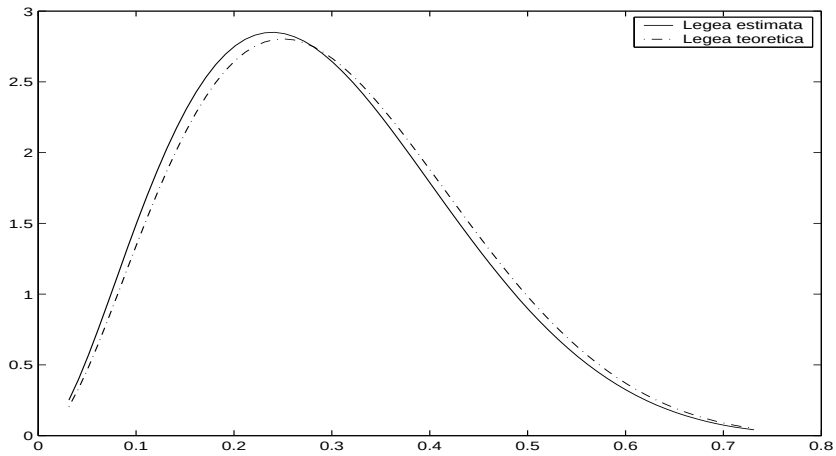

Figura 5.8: Legea $\mathcal{B}eta\,(3,7)$

```
abarat= 2.925,  bbarat= 7.130
(a1,a2)=( 2.431, 3.418)
(b1,b2)=( 5.928, 8.333)
```

Capitolul 6

Verificarea ipotezelor statistice

6.1 Concepte de bază

Fie colectivitatea $\mathcal{C}$ cercetată din punct de vedere al caracteristicii X, care are legea de probabilitate dată prin funcţia de probabilitate $f(x;\theta)$, ce reprezintă funcţia de frecvenţă în cazul discret, respectiv densitatea de probabilitate în cazul continuu, pentru variabila aleatoare X.

Definiţia 6.1.1. *Numim* ipoteză statistică *o presupunere relativă la legea de probabilitate pe care o urmează caracteristica* X.

Definiţia 6.1.2. *Metoda de stabilire a veridicităţii unei ipoteze statistice, se numeşte* test (criteriu) de verificare *a ipotezei statistice.*

Definiţia 6.1.3. *Când ipoteza statistică se referă la parametri de care depinde legea de probabilitate a caracteristicii* X *se obţine un* test parametric*, iar în caz contrar se obţine un* test neparametric.

Observaţia 6.1.4. Pentru testele parametrice vom considera că $\theta \in A = A_0 \cup A_1$, unde $A_0 \cap A_1 = \varnothing$. Ipoteza $H_0 : \theta \in A_0$ o vom numi *ipoteză nulă*, iar ipoteza $H_1 : \theta \in A_1$ o vom numi *ipoteză alternativă*.

Definiţia 6.1.5. *O ipoteză parametrică se numeşte* ipoteză simplă*, dacă mulţimea la care se presupune că aparţine parametrul necunoscut este formată dintr-un singur element, iar în caz contrar se numeşte* ipoteză compusă.

Observaţia 6.1.6. Ipoteza nulă este aceea care o intuim a fi cea apropriată de realitate, intuiţie pe care o obţinem din informaţiile pe care le deţinem referitoare la caracteristica cercetată X.

Observaţia 6.1.7. Construirea unui test revine la obţinerea unui domeniu $\mathcal{U} \subset \mathbb{R}^n$, numit *regiune critică*, pentru un *nivel de semnificaţie (probabilitate de risc)* α dat, astfel încât

$$P\left((X_1, X_2, \ldots, X_n) \in \mathcal{U} \mid H_0\right) = \alpha,$$

unde $X_1, X_2, \ldots, X_n$ sunt variabilele de selecţie corespunzătoare unei selecţii de volum n considerată.

Folosind datele de selecţie şi regiunea critică $\mathcal{U}$ astfel determinată, avem că ipoteza nulă H_0 va fi admisă (acceptată) dacă $(x_1, x_2, \ldots, x_n) \notin \mathcal{U}$, iar în caz contrar va fi respinsă, altfel spus ipoteza alternativă H_1 va fi admisă (acceptată) în al doilea caz.

6.2 Testul Z privind media teoretică

Se consideră caracteristica X care urmează legea normală $\mathcal{N}(m, \sigma)$, unde $m \in \mathbb{R}$ este necunoscut, iar $\sigma > 0$ este cunoscut.

Relativ la media teoretică $m = E(X)$ facem ipoteza nulă $H_0 : m = m_0$, cu una din alternativele:

$H_1 : m \neq m_0$ *(testul Z bilateral),*

$H_1 : m > m_0$ *(testul Z unilateral dreapta),*

$H_1 : m < m_0$ *(testul Z unilateral stânga).*

Pentru verificare ipotezei nule H_0 cu una din alternativele precizate mai înainte, considerăm o selecţie repetată de volum n şi un nivel de semnificaţie $\alpha \in (0, 1)$.

Se cunoaşte că statistica

$$Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}},$$

urmează legea normală $\mathcal{N}(0, 1)$. Prin urmare, pentru $\alpha \in (0, 1)$, putem determina un interval numeric (z_1, z_2) astfel încât

$$P(z_1 < Z < z_2 \mid H_0) = \Phi(z_2) - \Phi(z_1) = 1 - \alpha.$$

Intervalul (z_1, z_2) nu este determinat în mod unic, dar având în vedere alternativa H_1 considerată, adăugăm condiţia suplimentară:

$z_1 = -z_2$, dacă $H_1 : m \neq m_0$, adică $\Phi\left(z_{1-\frac{\alpha}{2}}\right) = \frac{1-\alpha}{2}$, unde $z_{1-\frac{\alpha}{2}} = z_2$;

$z_1 = -\infty,\ z_2 = z_{1-\alpha}$, unde $\Phi(z_{1-\alpha}) = \frac{1}{2} - \alpha$, dacă $H_1 : m > m_0$;

$z_1 = z_\alpha,\ z_2 = +\infty$, unde $\Phi(-z_\alpha) = \frac{1}{2} - \alpha$, dacă $H_1 : m < m_0$.

Corespunzător celor trei alternative definim regiunea critică respectiv prin:

$$\mathcal{U} = \left\{ (u_1, u_2, \ldots, u_n) \in \mathbb{R}^n \;\middle|\; \frac{|\,\bar{u} - m_0\,|}{\frac{\sigma}{\sqrt{n}}} \geqslant z_{1-\frac{\alpha}{2}} \right\},$$

$$\mathcal{U} = \left\{ (u_1, u_2, \ldots, u_n) \in \mathbb{R}^n \;\middle|\; \frac{\bar{u} - m_0}{\frac{\sigma}{\sqrt{n}}} \geqslant z_{1-\alpha} \right\},$$

$$\mathcal{U} = \left\{ (u_1, u_2, \ldots, u_n) \in \mathbb{R}^n \;\middle|\; \frac{\bar{u} - m_0}{\frac{\sigma}{\sqrt{n}}} \leqslant z_{\alpha} \right\}.$$

Aici s-a folosit notaţia $\bar{u} = \frac{1}{n}\sum_{k=1}^{n} u_k$.

Într-adevăr, pentru fiecare din cele trei regiunui critice scrise mai înainte, avem că $P\big((X_1, X_2, \ldots, X_n) \in \mathcal{U} \mid H_0\big) = \alpha$. Modul în care am ales regiunea critică ne conduce respectiv la testul Z bilateral, unilateral dreapta şi unilateral stânga.

Mai târziu va fi fundamentată riguros modalitatea de construire a regiunii critice pentru alternativele considerate.

Odată construită regiunea critică $\mathcal{U}$, ipoteza nulă va fi admisă dacă datele de selecţie satisfac condiţia $(x_1, x_2, \ldots, x_n) \notin \mathcal{U}$, iar în caz contrar va fi respinsă.

Remarcăm de asemenea că regiunea critică $\mathcal{U}$ corespunde mulţimii complementare intervalului (z_1, z_2).

Etapele aplicării testului Z

1. Se dau: $\alpha;\ x_1, x_2, \ldots, x_n;\ m_0;\ \sigma$;
2. Se calculează intervalul (z_1, z_2) astfel încât $\Phi(z_2) - \Phi(z_1) = 1 - \alpha$ (după cum s-a precizat mai înainte);
3. Se calculează $z = \dfrac{\bar{x} - m_0}{\frac{\sigma}{\sqrt{n}}}$, unde $\bar{x} = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} x_k$;
4. Concluzia: dacă $z \in (z_1, z_2)$ ipoteza H_0 este admisă, în caz contrar ipoteza este respinsă.

Observaţia 6.2.1. Testul Z se poate aplica şi în cazul unei caracteristici X care nu urmează legea normală, dacă volumul selecţiei este mare ($n > 30$), considerându-se media teoretică $m = E(X)$ necunoscută şi abaterea standard $\sigma = \sqrt{Var(X)}$ cunoscută. Această consideraţie se bazează pe Teorema 2.7.8.

Exemplul 6.2.2. Caracteristica X reprezintă cheltuielile lunare în mii lei pentru abonamentele la ziare şi reviste ale unei familii. Să se verifice, cu nivelul de semnificaţie $\alpha = 0.01$, dacă media acestor cheltuieli lunare pentru o familie este de 16 mii lei, ştiind că abaterea standard $\sigma = 3$ mii lei şi având o selecţie repetată de volum $n = 40$, care ne dă distribuţia empirică de selecţie

$$X \begin{pmatrix} 11 & 13 & 15 & 17 & 20 \\ 4 & 6 & 12 & 10 & 8 \end{pmatrix}.$$

Deoarece $n = 40 > 30$ şi abaterea standard $\sigma = 3$ este cunoscută, vom folosi testul Z pentru verificarea ipotezei nule

$$H_0 : m = E(X) = 16, \quad \text{cu ipoteza alternativă } H_1 : m \neq 16.$$

Pentru $\alpha = 0.01$, folosind *Anexa I*, se determină $z_{1-\frac{\alpha}{2}} = z_{0.995}$, astfel încât

$$\Phi(z_{0.995}) = \frac{1-\alpha}{2} = 0.495.$$

Anume, se obţine că $z_{0.995} = 2.58$, care ne dă intervalul numeric $(-2.58\,;\, 2.58)$ pentru statistica notată prin:

$$Z = \frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}}.$$

Calculăm succesiv

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{40}\left(4 \cdot 11 + 6 \cdot 13 + 12 \cdot 15 + 10 \cdot 17 + 8 \cdot 20\right) = 15.8;$$

$$z = \frac{\bar{x} - m_0}{\frac{\sigma}{\sqrt{n}}} = \frac{15.8 - 16}{\frac{3}{\sqrt{40}}} = \frac{-0.2 \cdot \sqrt{40}}{3} = -0.422.$$

Deoarece $z = -0.422 \in (-2.58\,;\, 2.58)$, rezultă că se acceptă ipoteza că cheltuielile medii lunare ale unei familii pentru abonamentele la ziare şi reviste sunt de 16 mii lei, cu nivelul de semnificaţie 0.01.

Programul 6.2.3. Programul Matlab, care urmează, execută aceste calcule şi afişează intervalul de încredere, împreună cu valoarea calculată a statsiticii Z.

```
x=[11 13 15 17 20]; f=[4 6 12 10 8];
z=(sum(x.*f)/40-16)/(3/sqrt(40));
z1=norminv(0.005,0,1); z2=norminv(0.995,0,1);
fprintf(' (z1,z2)=(%6.3f,%6.3f)\n',z1,z2)
fprintf('       z=%6.3f,%6.3f)',z)
```

În urma executării programului, s–au obţinut rezultatele:

```
(z1,z2)=(-2.576, 2.576)
      z=-0.422,
```

Observaţia 6.2.4. Având în vedere că funcţia Φ a lui Laplace este monoton crescătoare şi că pentru determinarea intervalului (z_1, z_2) este necesară inversarea funcţiei lui Laplace, se poate renunţa la inversarea acesteia. Pentru aceasta se determină ceea ce se numeşte *valoare critică* şi pe care o notăm prin c. Vom analiza pe rând cele trei alternative.

Dacă se consideră testul bilateral, făcând notaţia

$$1-\frac{c}{2}=\Phi\left(|z|\right), \quad \text{adică} \quad c=1-2\Phi\left(|z|\right),$$

ipoteza nulă va fi respinsă dacă $1-\frac{c}{2}\geqslant 1-\frac{\alpha}{2}$, adică $c\leqslant\alpha$.

Pentru testul unilateral dreapta, se notează

$$1-c=\frac{1}{2}+\Phi\left(z\right), \quad \text{adică} \quad c=\frac{1}{2}-\Phi\left(z\right),$$

iar ipoteza nulă este respinsă, dacă $1-c\geqslant 1-\alpha$, adică, la fel ca mai înainte, dacă $c\leqslant\alpha$.

Un raţionament analog, pentru testul unilateral stânga ne conduce la valoarea critică, calculată după formulă $c=\frac{1}{2}+\Phi\left(z\right)$. Drept urmare, ipoteza nulă este respinsă, dacă $c\leqslant\alpha$.

În rezumat, ipoteza nulă este respinsă, dacă $c\leqslant\alpha$, unde

$$c=\begin{cases}1-2\Phi\left(|z|\right), & \text{dacă } H_1: m\neq m_0,\\ \dfrac{1}{2}-\Phi\left(z\right), & \text{dacă } H_1: m>m_0,\\ \dfrac{1}{2}+\Phi\left(z\right), & \text{dacă } H_1: m<m_0.\end{cases}$$

Pentru datele din Exemplul 6.2.2, se obţine

$$c=\begin{cases}0.673, & \text{pentru testul bilateral},\\ 0.663, & \text{pentru testul unilateral dreapta},\\ 0.337, & \text{petru testul unilateral stânga}.\end{cases}$$

6.2.1 Funcţiile `zscore` şi `ztest`

Sistemul Matlab, prin *Statistics toolbox*, dispune de funcţiile `zscore` şi `ztest`, cu aplicabilitate la testul Z. Apelarea acestor funcţii se face prin:

```
z=zscore(d)
h=ztest(x,m0,sigma)
h=ztest(x,m0,sigma,alpha)
h=ztest(x,m0,sigma,alpha,tail)
[h,c,ci,zc]=ztest(x,m0,sigma,alpha,tail)
```

În urma executării primei instrucţiuni, dacă `d` este un vector, se calculează abaterile componentelor vectorului de la valorea medie a componentelor, raportate la abaterea standard a componentelor, adică $\mathtt{x}_i = \frac{\mathtt{d}_i - \bar{\mathtt{d}}}{\sigma_{\mathtt{d}}}$. Dacă `d` este matrice, atunci operaţia este executată pentru fiecare coloană în parte.

Celelalte instrucţiuni se referă la funcţia `ztest` şi efectuează testul Z asupra datelor conţinute în vectorul `x`, folosind nivelul de semnificaţie `alpha`, care are valoarea implicită `alpha=0.05`.

Parametrul `tail` specifică una din cele trei alternative, care conduc la testul bilateral (`tail=0`, implicit), unilateral dreapta (`tail=1`) şi unilateral stânga (`tail=-1`). Dacă `h=1`, atunci ipoteza nulă va fi respinsă, respectiv dacă `h=0`, ipoteza nu poate fi respinsă.

Ultima formă de apel permite, de asemenea, obţinerea valorii critice `c`, a valorii `zc` a statisticii Z, precum şi a intervalului de încredere pentru media teoretică, corespunzător probabilităţii de încredere `1-alpha`, obţinut în vectorul cu două componente `ci`.

Programul 6.2.5. Programul Matlab, ce urmează, rezolvă problema din Exemplul 6.2.2. Mai mult, se consideră şi testele unilateral dreapta şi unilateral stânga, precum şi obţinerea intervalelor de încredere.

```
x=[11*ones(1,4),13*ones(1,6),15*ones(1,12),...
   17*ones(1,10),20*ones(1,8)];
fprintf(' h    c           ci           z\n')
fprintf('_____________________________________\n')
for i=-1:1
  [h,c,ci,zc]=ztest(x,16,3,0.01,i);
  fprintf(' %d  %5.4f   (%4.2f,%4.2f)   %6.4f\n',...
           h,c,ci,zc)
end
```

În urma executării programului, se obţin rezultatele:

```
 h    c           ci           z
_____________________________________
 0  0.3366   (-Inf,16.90)   -0.4216
 0  0.6733   (14.58,17.02)   -0.4216
 0  0.6634   (14.70, Inf)   -0.4216
```

Se observă că pentru `tail=-1` şi `tail=1`, se construiesc intervale nemărginite, respectiv la stânga şi la dreapta.

6.3 Puterea unui test

Definiţia 6.3.1. *Dacă se consideră un test relativ la ipoteza nulă H_0 cu alternativa H_1, se numeşte* eroare de genul (speţa) întâi *respingerea unei ipoteze adevărate, iar probabilitatea acestei erori se numeşte* risc de speţa întâi (risc al furnizorului) *şi este dată de nivelul α de semnificaţie, adică*

$$\alpha = P\left((X_1, X_2, \ldots, X_n) \in \mathcal{U} \mid H_0\right).$$

Definiţia 6.3.2. *Se numeşte* eroare de genul (speţa) al doilea *admiterea unei ipoteze false, iar probabilitatea acestei erori se numeşte* risc de speţa a doua (risc al beneficiarului) *şi este notată β,*

$$\beta = P\left((X_1, X_2, \ldots, X_n) \notin \mathcal{U} \mid H_1\right).$$

Observaţia 6.3.3. Producerea unei erori de genul al doilea este mai gravă decât a unei erori de genul întâi, când verificăm concentraţia unui medicament, care peste un anumit grad de concentraţie devine dăunător. Pe de altă parte, producerea unei erori de genul întâi este mai gravă decât cea a unei erori de genul doi, când verificăm calitatea unui articol de îmbrăcăminte.

Definiţia 6.3.4. *Numim* puterea unui test *probabilitatea respingerii unei ipoteze false, adică*

$$\pi\left(\widetilde{\theta}\right) = \pi\left(\mathcal{U}; \widetilde{\theta}\right) = P\left((X_1, X_2, \ldots, X_n) \in \mathcal{U} \mid \theta = \widetilde{\theta}\right),$$

când θ este parametrul asupra căruia se face ipoteza statistică, iar $\mathcal{U}$ este regiunea critică construită sub ipoteza nulă cu nivelul de semnificaţie $\alpha \in (0,1)$ fixat.

Observaţia 6.3.5. Dacă testul considerat se referă la ipoteza nulă $H_0 : \theta = \theta_0$ cu ipoteza alternativă $H_1 : \theta = \theta_1$, atunci $\pi(\theta_0) = \alpha$ şi $\pi(\theta_1) = 1 - \beta$.

Aplicaţia 6.3.6. Calculăm în continuare puterea testului Z, când avem ipoteza alternativă $H_1 : m = m_1 \neq m_0$.

După cum am văzut la testul Z regiunea critică în acest caz este

$$\mathcal{U} = \left\{(u_1, u_2, \ldots, u_n) \in \mathbb{R}^n \;\middle|\; \frac{|\,\bar{u} - m_0\,|}{\frac{\sigma}{\sqrt{n}}} \geqslant z_{1-\frac{\alpha}{2}}\right\}.$$

Pentru calculul puterii $\pi(m_1)$ a testului avem

$$\pi(m_1) = P\left((X_1, X_2, \ldots, X_n) \in \mathcal{U} \mid H_1\right) = P\left(\frac{\left|\bar{X} - m_0\right|}{\frac{\sigma}{\sqrt{n}}} \geqslant z_{1-\frac{\alpha}{2}} \;\middle|\; H_1\right).$$

Dacă se consideră evenimentul contrar, vom putea scrie succesiv

$$\begin{aligned}
\beta =& 1 - \pi(m_1) = P\left(\frac{\left|\bar{X} - m_0\right|}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}} \;\middle|\; H_1\right) \\
=& P\left(-z_{1-\frac{\alpha}{2}} < \frac{\bar{X} - m_1}{\frac{\sigma}{\sqrt{n}}} - \frac{m_0 - m_1}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}} \;\middle|\; H_1\right) \\
=& P\left(\frac{m_0 - m_1}{\frac{\sigma}{\sqrt{n}}} - z_{1-\frac{\alpha}{2}} < \frac{\bar{X} - m_1}{\frac{\sigma}{\sqrt{n}}} < \frac{m_0 - m_1}{\frac{\sigma}{\sqrt{n}}} + z_{1-\frac{\alpha}{2}} \;\middle|\; H_1\right),
\end{aligned}$$

de unde avem că

$$\beta = \Phi\left(\frac{m_0 - m_1}{\frac{\sigma}{\sqrt{n}}} + z_{1-\frac{\alpha}{2}}\right) - \Phi\left(\frac{m_0 - m_1}{\frac{\sigma}{\sqrt{n}}} - z_{1-\frac{\alpha}{2}}\right), \tag{6.3.1}$$

adică

$$\pi(m_1) = 1 - \Phi\left(\frac{m_0 - m_1}{\frac{\sigma}{\sqrt{n}}} + z_{1-\frac{\alpha}{2}}\right) + \Phi\left(\frac{m_0 - m_1}{\frac{\sigma}{\sqrt{n}}} - z_{1-\frac{\alpha}{2}}\right). \tag{6.3.2}$$

Deoarece $\Phi(+\infty) = 0.5$ şi $\Phi(-\infty) = -0.5$, rezultă că pentru $n \to \infty$, vom avea $\beta \to 0$, adică $\pi(m_1) \to 1$. Aşadar, putem determina, din relaţia precedentă valoarea lui n (volumul selecţiei) astfel încât puterea testului (corespunzător riscul de speţa a doua) să fie atinsă pentru acel n.

Exemplul 6.3.7. Relativ la caracteristica X s-a efectuat o selecţie de volum $n = 24$, obţinându-se datele de selecţie

3.59	4.27	2.83	2.28	3.69	2.81	3.16	3.03	3.65	3.84	4.32	3.21
3.20	3.32	3.82	3.27	3.77	3.56	3.29	4.08	2.97	3.29	3.78	3.58.

Ştiind că X urmează legea normală $\mathcal{N}(m, \sigma)$ cu abaterea standard teoretică cunoscută $\sigma = \sqrt{Var(X)} = 0.5$, vrem să verificăm ipoteza nulă $H_0 : m = 3.5$, când se consideră ipoteza alternativa $H_1 : m \neq 3.5$, cu nivelul de semnificaţie $\alpha = 0.01$, iar apoi să calculăm puterea testului, când $m = 3.4,\ 3.5,\ 3.6,\ 3.7$, şi să determinăm volumul n optim al selecţiei când se consideră alternativa $H_1 : m = 3.4$, şi riscul de speţa a doua $\beta = 0.05$.

Când se consideră alternativa $H_1 : m \neq 3.5$ se foloseşte testul Z bilateral. Se determină cuantila

$$z_{1-\frac{\alpha}{2}} = 2.58 \quad \text{din relaţia} \quad \Phi\left(z_{1-\frac{\alpha}{2}}\right) = \frac{1-\alpha}{2} = \frac{0.99}{2} = 0.495.$$

Pe de altă parte avem că

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{24}\left(3.59 + 4.27 + \cdots + 3.58\right) = 3.44,$$

astfel că

$$z = \frac{\bar{x} - m_0}{\frac{\sigma}{\sqrt{n}}} = \frac{3.44 - 3.5}{\frac{0.5}{\sqrt{24}}} = -0.57.$$

Deoarece $|z| = 0.57 < 2.58 = z_{1-\frac{\alpha}{2}}$, ipoteza $H_0 : m = 3.5$ este admisă.

Formula de calcul a puterii testului Z bilateral este (6.3.2). Prin urmare, se obţine

$$\begin{aligned}
\pi(3.4) &= 1 - \Phi\left(0.4\sqrt{6} + 2.58\right) + \Phi\left(0.4\sqrt{6} - 2.58\right) = 0.06,\\
\pi(3.5) &= 1 - \Phi(2.58) + \Phi(-2.58) = 0.01,\\
\pi(3.6) &= 1 - \Phi\left(-0.4\sqrt{6} + 2.58\right) + \Phi\left(-0.4\sqrt{6} - 2.58\right) = 0.05,\\
\pi(3.7) &= 1 - \Phi\left(-0.8\sqrt{6} + 2.58\right) + \widetilde{\Phi}\left(-0.8\sqrt{6} - 2.58\right) = 0.27.
\end{aligned}$$

Pentru determinarea volumului optim al selecţiei, când se consideră alternativa $H_1 : m = 3.4$, se foloseşte formula (6.3.1). Astfel se obţine

$$\begin{aligned}
0.05 = \beta =& \Phi\left(\frac{3.5 - 3.4}{\frac{0.5}{\sqrt{n}}} + 2.58\right) - \Phi\left(\frac{3.5 - 3.4}{\frac{0.5}{\sqrt{n}}} - 2.58\right)\\
=& \Phi\left(\frac{1}{5}\sqrt{n} + 2.58\right) - \Phi\left(\frac{1}{5}\sqrt{n} - 2.58\right).
\end{aligned}$$

Când n este astfel încât $\frac{1}{5}\sqrt{n} - 2.58 < 0$, se observă că valoarea $\beta = 0.05$ nu poate fi atinsă. Dacă $\frac{1}{5}\sqrt{n} - 2.58 > 0$, atunci $n > 166$, astfel avem $\Phi\left(\frac{1}{5}\sqrt{n} + 2.58\right) \approx 0.5$. Prin urmare, pentru determinarea lui n optim avem relaţia $\Phi\left(\frac{1}{5}\sqrt{n} - 2.58\right) = 0.45$, şi folosind *Anexa I* se obţine $\frac{1}{5}\sqrt{n} - 2.58 = 1.65$, adică $\sqrt{n} = 5 \times 4.23 = 21.15$, de unde se obţine $n = 448$.

Programul 6.3.8. Programul Matlab, ce urmează, pe lângă faptul că va efectua calculele din exemplul precedent, va reprezenta grafic şi funcţia putere. Punctele de pe curba puterii, ale căror valori au fost calculate în exemplul de mai sus, sunt marcate prin cerculeţe în Figura 6.1.

```
clear,clf
x=[3.59,4.27,2.83,2.28,3.69,2.81,3.16,3.03,3.65,...
   3.84,4.32,3.21,3.20,3.32,3.82,3.27,3.77,3.56,...
   3.29,4.08,2.97,3.29,3.78,3.58];
z=(mean(x)-3.5)/(0.5/sqrt(24));
z1=norminv(0.005,0,1); z2=norminv(0.995,0,1);
fprintf(' z=%6.3f\n z1=%6.3f\n z2=%6.3f\n',z,z1,z2)
m=3:0.01:4; m1=3.4:0.1:3.7;
c1=norminv(0.005,0,1); c2=norminv(0.995,0,1);
pib=1-normcdf((3.5-m)./(0.5/sqrt(24))+c2,0,1)...
     +normcdf((3.5-m)./(0.5/sqrt(24))+c1,0,1);
pib1=1-normcdf((3.5-m1)./(0.5/sqrt(24))+c2,0,1)...
      +normcdf((3.5-m1)./(0.5/sqrt(24))+c1,0,1);
plot(m,pib,'k-',m1,pib1,'o'), grid on, len=length(m1);
for i=1:len
    fprintf(' pi(%3.1f)=%5.3f\n',m1(i),pib1(i))
end
za=norminv(0.005,0,1); zb=norminv(0.95,0,1);
n=ceil(0.5^2*(zb-za)^2/(3.5-3.4)^2);
fprintf(' nopt=%3d',n)
```

Rezultatele obţinute, în urma executării programului, sunt următoarele:

```
z=-0.567
z1=-2.576
z2= 2.576
pi(3.4)=0.055
pi(3.5)=0.010
pi(3.6)=0.055
pi(3.7)=0.269
nopt=446
```

Lema 6.3.9 (Neyman–Pearson). *Fie caracteristica X cu legea de probabilitate dată prin $f(x;\theta)$, unde $\theta \in A \subset \mathbb{R}$ este parametrul necunoscut asupra căruia se face ipoteza nulă simplă $H_0 : \theta = \theta_0$, cu ipoteza alternativă simplă $H_1 : \theta = \theta_1 \neq \theta_0$. Se consideră o selecţie repetată de volum n relativă la caracteristica X şi nivelul de semnificaţie dat $\alpha \in (0,1)$, atunci*

$$\max\left\{ P\left((X_1,\ldots,X_n)\in\mathcal{U} \mid H_1\right) \middle| \mathcal{U}\subset\mathbb{R}^n, P\left((X_1,\ldots,X_n)\in\mathcal{U} \mid H_0\right)=\alpha \right\} = P\left((X_1,\ldots,X_n)\in\widetilde{\mathcal{U}} \mid H_0\right),$$

regiunea critică $\widetilde{\mathcal{U}}$, fiind definită prin

$$\widetilde{\mathcal{U}} = \left\{ (u_1,\ldots,u_n)\in\mathbb{R}^n \;\middle|\; \frac{g(u_1,\ldots,u_n;\theta_1)}{g(u_1,\ldots,u_n;\theta_0)} \geqslant k_\alpha > 0 \right\},$$

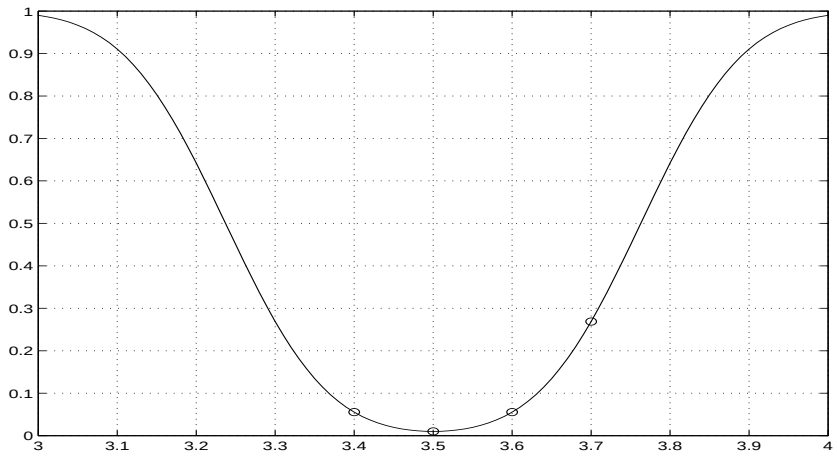

Figura 6.1: Curba funcţiei putere

unde

$$g\left(u_1,\ldots,u_n;\theta\right)=\prod_{k=1}^{n}f\left(u_k;\theta\right).$$

Demonstraţie. Facem demonstraţia în două etape.

În prima etapă, considerăm că se cunoaşte constanta $k_\alpha>0$ şi arătăm că

$$1-\beta=P\left((X_1,\ldots,X_n)\in\mathcal{U}\;\middle|\;H_1\right)\leqslant P\left((X_1,\ldots,X_n)\in\widetilde{\mathcal{U}}\;\middle|\;H_1\right)=1-\widetilde{\beta}.$$

Pentru aceasta, dacă notăm $\mathcal{W}=\mathcal{U}\cap\widetilde{\mathcal{U}}$, obţinem succesiv

$$\begin{aligned}&\int\ldots\int_{\mathcal{U}}g\left(u_1,\ldots,u_n;\theta_0\right)du_1\ldots du_n=P\left((X_1,\ldots,X_n)\in\mathcal{U}\;\middle|\;H_0\right)=\alpha\\&=P\left((X_1,\ldots,X_n)\in\widetilde{\mathcal{U}}\;\middle|\;H_0\right)=\int\ldots\int_{\widetilde{\mathcal{U}}}g\left(u_1,\ldots,u_n;\theta_0\right)du_1\ldots du_n,\end{aligned}$$

de unde avem că

$$\begin{aligned}\int\ldots\int_{\mathcal{U}\setminus\mathcal{W}}g\left(u_1,\ldots,u_n;\theta_0\right)du_1\ldots du_n+\int\ldots\int_{\mathcal{W}}g\left(u_1,\ldots,u_n;\theta_0\right)du_1\ldots du_n\\=\int\ldots\int_{\widetilde{\mathcal{U}}\setminus\mathcal{W}}g\left(u_1,\ldots,u_n;\theta_0\right)du_1\ldots du_n+\\+\int\ldots\int_{\mathcal{W}}g\left(u_1,\ldots,u_n;\theta_0\right)du_1\ldots du_n,\end{aligned}$$

adică

$$\int\ldots\int_{\mathcal{U}\setminus\mathcal{W}} g\,(u_1,\ldots,u_n;\theta_0)du_1\ldots du_n = \int\ldots\int_{\widetilde{\mathcal{U}}\setminus\mathcal{W}} g\,(u_1,\ldots,u_n;\theta_0)\,du_1\ldots du_n. \tag{6.3.3}$$

Putem trece acum la calculul diferenţei

$$\begin{aligned}\beta-\widetilde{\beta} &= (1-\widetilde{\beta}\,)-(1-\beta)\\ &= P\left((X_1,\ldots,X_n)\in\widetilde{\mathcal{U}}\,\middle|\,H_1\right)-P\big((X_1,\ldots,X_n)\in\mathcal{U}\,\big|\,H_1\big)\\ &= \int\ldots\int_{\widetilde{\mathcal{U}}} g\,(u_1,\ldots,u_n;\theta_1)\,du_1\ldots du_n-\int\ldots\int_{\mathcal{U}} g\,(u_1,\ldots,u_n;\theta_1)\,du_1\ldots du_n\\ &= \int\ldots\int_{\widetilde{\mathcal{U}}\setminus\mathcal{W}} g\,(u_1,\ldots,u_n;\theta_1)\,du_1\ldots du_n\\ &\qquad -\int\ldots\int_{\mathcal{U}\setminus\mathcal{W}} g\,(u_1,\ldots,u_n;\theta_1)\,du_1\ldots du_n.\end{aligned}$$

Scriem aceată diferenţă sub forma

$$\begin{aligned}\beta-\widetilde{\beta} &= \int\ldots\int_{\widetilde{\mathcal{U}}\setminus\mathcal{W}}\frac{g\,(u_1,\ldots,u_n;\theta_1)}{g\,(u_1,\ldots,u_n;\theta_0)}\,g\,(u_1,\ldots,u_n;\theta_0)\,du_1\ldots du_n\\ &\quad -\int\ldots\int_{\mathcal{U}\setminus\mathcal{W}}\frac{g\,(u_1,\ldots,u_n;\theta_1)}{g\,(u_1,\ldots,u_n;\theta_0)}\,g\,(u_1,\ldots,u_n;\theta_0)\,du_1\ldots du_n.\end{aligned}$$

Pentru fiecare integrală din membrul drept aplicăm formula de medie. Prin urmare există $(\xi_1,\xi_2,\ldots,\xi_n)\in\widetilde{\mathcal{U}}\setminus\mathcal{W}$ şi $(\eta_1,\eta_2,\ldots,\eta_n)\in\mathcal{U}\setminus\mathcal{W}$ astfel încât

$$\begin{aligned}\beta-\widetilde{\beta} =& \frac{g\,(\xi_1,\ldots,\xi_n;\theta_1)}{g\,(\xi_1,\ldots,\xi_n;\theta_0)}\int\ldots\int_{\widetilde{\mathcal{U}}\setminus\mathcal{W}} g\,(u_1,\ldots,u_n;\theta_0)\,du_1\ldots du_n\\ &-\frac{g\,(\eta_1,\ldots,\eta_n;\theta_1)}{g\,(\eta_1,\ldots,\eta_n;\theta_0)}\int\ldots\int_{\mathcal{U}\setminus\mathcal{W}} g\,(u_1,\ldots,u_n;\theta_0)\,du_1\ldots du_n.\end{aligned}$$

Folosind relaţia (6.3.3), stabilită mai înainte, avem că

$$\begin{aligned}\beta-\widetilde{\beta} =& \left[\frac{g\,(\xi_1,\ldots,\xi_n;\theta_1)}{g\,(\xi_1,\ldots,\xi_n;\theta_0)}-\frac{g\,(\eta_1,\ldots,\eta_n;\theta_1)}{g\,(\eta_1,\ldots,\eta_n;\theta_0)}\right]\times\\ &\times\int\ldots\int_{\mathcal{U}\setminus\mathcal{W}} g\,(u_1,\ldots,u_n;\theta_0)\,du_1\ldots du_n.\end{aligned}$$

Dar integrala din membrul drept este pozitivă, iar din modul cum s-a definit regiunea critică $\widetilde{\mathcal{U}}$ avem că

$$\frac{g\left(\xi_1,\ldots,\xi_n;\theta_1\right)}{g\left(\xi_1,,\ldots,\xi_n;\theta_0\right)} \geqslant k_\alpha > \frac{g\left(\eta_1,\ldots,\eta_n;\theta_1\right)}{g\left(\eta_1,\ldots,\eta_n;\theta_0\right)},$$

deci $\beta-\widetilde{\beta}>0$ şi ca urmare $1-\widetilde{\beta}>1-\beta$. Ceea ce încheie partea întâi a demonstraţiei.

În partea a doua arătăm existenţa constantei $k_\alpha>0$.

Notăm prin $A\left(k\right)$ următoarea regiune

$$A\left(k\right)=\left\{\left(u_1,\ldots,u_n\right)\in\mathbb{R}^n \;\middle|\; g\left(u_1,\ldots,u_n;\theta_1\right)\geqslant k\, g\left(u_1,\ldots,u_n;\theta_0\right)\right\},$$

şi prin $s\left(k\right)$, următoarea probabilitate

$$s\left(k\right)=P\big(\left(X_1,\ldots,X_n\right)\in A\left(k\right)\mid H_0\big)=\int\ldots\int_{A(k)} g\left(u_1,\ldots,u_n;\theta_0\right)du_1\ldots du_n.$$

Se observă că $s\left(+\infty\right)=0$, iar

$$s\left(0\right)=\int\ldots\int_{\mathbb{R}^n} g\left(u_1,\ldots,u_n;\theta_0\right)du_1\ldots du_n=1.$$

De asemenea, avem că funcţia $s\left(k\right)$ este monoton descrescătoare. Prin urmare, pentru $\alpha\in\left(0,1\right)$ fixat, există $k_\alpha>0$ astfel încât $s\left(k_\alpha\right)=\alpha$, adică

$$P\big(\left(X_1,\ldots,X_n\right)\in A\left(k_\alpha\right)\mid H_0\big)=\alpha.$$

Ceea ce încheie demonstraţia lemei. □

Observaţia 6.3.10. Demonstraţia lemei Neyman–Pearson a fost făcută în cazul continuu. În mod analog se demonstrează şi pentru cazul discret, integralele multiple care apar fiind înlocuite cu sume multiple.

Definiţia 6.3.11. *Testul pentru care puterea este maximă îl numim* cel mai puternic test.

Definiţia 6.3.12. *Testul pentru care are loc inegalitatea*

$$1-\beta=P\big(\left(X_1,X_2,\ldots,X_n\right)\in\mathcal{U}\,|\,H_1\big)>P\big(\left(X_1,X_2,\ldots,X_n\right)\in\mathcal{U}\,|\,H_0\big)=\alpha,$$

adică puterea testului este mai mare decât riscul de speţa întâi se numeşte test neplasat.

Proprietatea 6.3.13. *Cel mai puternic test dat de lema Neyman–Pearson este un test nedeplasat, adică* $1-\widetilde{\beta}>\alpha$.

Demonstraţie. Cu notaţiile de la lema Neyman–Pearson, avem că

$$\int\ldots\int_{\widetilde{\mathcal{U}}} g\left(u_1,\ldots,u_n;\theta_1\right)du_1\ldots du_n \geqslant k_\alpha\int\ldots\int_{\widetilde{\mathcal{U}}} g\left(u_1,\ldots,u_n;\theta_0\right)du_1\ldots du_n.$$

Distingem două cazuri.

Dacă avem $k_\alpha>1$, atunci obţinem

$$\begin{aligned}1-\widetilde{\beta}&=\int\ldots\int_{\widetilde{\mathcal{U}}} g\left(u_1,\ldots,u_n;\theta_1\right)du_1\ldots du_n\\&\geqslant k_\alpha\int\ldots\int_{\widetilde{\mathcal{U}}} g\left(u_1,\ldots,u_n;\theta_0\right)du_1\ldots du_n\\&>\int\ldots\int_{\widetilde{\mathcal{U}}} g\left(u_1,\ldots,u_n;\theta_0\right)du_1\ldots du_n=\alpha,\end{aligned}$$

de unde $1-\widetilde{\beta}>\alpha$.

În cazul când $k_\alpha\leqslant 1$, avem

$$\begin{aligned}\widetilde{\beta}&=\int\ldots\int_{\complement\widetilde{\mathcal{U}}} g\left(u_1,\ldots,u_n;\theta_1\right)du_1\ldots du_n\\&<k_\alpha\int\ldots\int_{\complement\widetilde{\mathcal{U}}} g\left(u_1,\ldots,u_n;\theta_0\right)du_1\ldots du_n\\&\leqslant\int\ldots\int_{\complement\widetilde{\mathcal{U}}} g\left(u_1,\ldots,u_n;\theta_0\right)du_1\ldots du_n=1-\alpha,\end{aligned}$$

de unde $1-\widetilde{\beta}>\alpha$, ceea ce trebuie demonstrat. □

Observaţia 6.3.14. Prin micşorarea riscului de speţa întâi, pentru n fixat, creşte constanta k_α şi prin urmare scade puterea testului, $1-\widetilde{\beta}$, deci creşte riscul de speţa a doua.

Aplicaţia 6.3.15. Fie caracteristica X ce urmează legea normală $\mathcal{N}\left(m,\sigma\right)$, unde $m\in\mathbb{R}$ este necunoscut şi $\sigma>0$ este cunoscut. Fie ipoteza nulă $H_0: m=m_0$, cu ipoteza alternativă $H_1: m=m_1\neq m_0$. Vrem să determinăm cel mai puternic test pentru verificarea acestei ipoteze nule.

Pentru aceasta aplicăm lema Neyman-Pearson.

Prima dată determinăm regiunea critică $\widetilde{\mathcal{U}}$.

Deoarece

$$g\left(u_1,u_2,\ldots,u_n;m\right)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}\left(u_k-m\right)^2\right\},$$

rezultă că

$$\frac{g\left(u_1,u_2,\ldots,u_n;m_1\right)}{g\left(u_1,u_2,\ldots,u_n;m_0\right)}=\exp\left\{\frac{1}{2\sigma^2}\left[\sum_{k=1}^{n}\left(u_k-m_0\right)^2-\sum_{k=1}^{n}\left(u_k-m_1\right)^2\right]\right\}.$$

Condiţia care defineşte regiunea critică $\widetilde{\mathcal{U}}$

$$\frac{g\left(u_1,u_2,\ldots,u_n;m_1\right)}{g\left(u_1,u_2,\ldots,u_n;m_0\right)}\geqslant k_\alpha,$$

devine prin logaritmare

$$\sum_{k=1}^{n}\left(u_k-m_0\right)^2-\sum_{k=1}^{n}\left(u_k-m_1\right)^2\geqslant 2\sigma^2\ln k_\alpha,$$

sau

$$2\left(m_1-m_0\right)\sum_{k=1}^{n}u_k-n\left(m_1-m_0\right)\left(m_1+m_0\right)\geqslant 2\sigma^2\ln k_\alpha,$$

adică

$$2\left(m_1-m_0\right)\sum_{k=1}^{n}u_k\geqslant 2\sigma^2\ln k_\alpha+n\left(m_1-m_0\right)\left(m_1+m_0\right).$$

Distingem în continuare două cazuri.

Dacă $m_0>m_1$, avem

$$\frac{1}{n}\sum_{k=1}^{n}u_k=\bar{u}\leqslant\frac{\sigma^2\ln k_\alpha}{\left(m_1-m_0\right)n}+\frac{m_1+m_0}{2},$$

şi notând constanta din membrul drept prin K_α, avem că $\bar{u}\leqslant K_\alpha$. Aşadar regiunea critică a celui mai puternic test este

$$\widetilde{\mathcal{U}}=\left\{\left(u_1,u_2,\ldots,u_n\right)\in\mathbb{R}^n\ \mid\ \bar{u}\leqslant K_\alpha\right\}.$$

Dacă $m_0<m_1$, procedând analog, se ajunge la regiunea critică

$$\widetilde{\mathcal{U}}=\left\{\left(u_1,u_2,\ldots,u_n\right)\in\mathbb{R}^n\ \mid\ \bar{u}\geqslant K_\alpha\right\}.$$

Pentru determinarea constantei K_α (cazul $m_0>m_1$), se scrie relaţia

$$\alpha=P\left(\left(X_1,X_2,\ldots,X_n\right)\in\widetilde{\mathcal{U}}\ \mid\ H_0\right)=P\left(\bar{X}\leqslant K_\alpha\ \mid\ H_0\right)$$

sau

$$P\left(\frac{\bar{X}-m_0}{\frac{\sigma}{\sqrt{n}}}\leqslant\frac{K_\alpha-m_0}{\frac{\sigma}{\sqrt{n}}}\ \middle|\ H_0\right)=\alpha.$$

Prin urmare, K_α se va determina astfel încât $\Phi\left(\frac{K_\alpha - m_0}{\frac{\sigma}{\sqrt{n}}}\right) = \alpha$, unde funcţia Φ a lui Laplace este cea definită prin

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt.$$

Aşadar avem cuantila z_α, ca fiind $z_\alpha = \frac{K_\alpha - m_0}{\frac{\sigma}{\sqrt{n}}}$, de unde $K_\alpha = m_0 + \frac{\sigma}{\sqrt{n}} z_\alpha$.

În acest fel, regiunea critică se poate scrie

$$\begin{aligned}\widetilde{\mathcal{U}} =& \left\{ (u_1, u_2, \ldots, u_n) \in \mathbb{R}^n \;\middle|\; \bar{u} \leqslant m_0 + \frac{\sigma}{\sqrt{n}} z_\alpha \right\} \\ =& \left\{ (u_1, u_2, \ldots, u_n) \in \mathbb{R}^n \;\middle|\; \frac{\bar{u} - m_0}{\frac{\sigma}{\sqrt{n}}} \leqslant z_\alpha \right\}.\end{aligned}$$

În cazul $m_0 < m_1$, procedând la fel, se ajunge la

$$\begin{aligned}\widetilde{\mathcal{U}} =& \left\{ (u_1, u_2, \ldots, u_n) \in \mathbb{R}^n \;\middle|\; \bar{u} \geqslant m_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} \right\} \\ =& \left\{ (u_1, u_2, \ldots, u_n) \in \mathbb{R}^n \;\middle|\; \frac{\bar{u} - m_0}{\frac{\sigma}{\sqrt{n}}} \geqslant z_{1-\alpha} \right\}.\end{aligned}$$

Observaţia 6.3.16. În aplicaţia precedentă, regiunea critică $\widetilde{\mathcal{U}}$ nu depinde de m_1 numai prin faptul că $m_1 < m_0$, respectiv $m_1 > m_0$. Drept urmare, ipoteza H_1 se poate înlocui, pentru cele două situaţii prin $H_1 : m > m_0$, ceea ce ne conduce la testul Z unilateral dreapta, respectiv $H_1 : m < m_0$, care ne conduce la testul Z unilateral stânga. Aceasta confirmă justeţea alegerii regiunilor critice în cazul testului Z unilateral dreapta şi respectiv stânga.

Observaţia 6.3.17. Când se consideră ipoteza alternativă $H_1 : m \neq m_0$ (testul Z bilateral), utilizând lema Neyman–Pearson, nu putem construi un cel mai puternic test. Altă metodă va fi prezentată pentru construirea acestuia în acest caz.

Aplicaţia 6.3.18. Vrem să determinăm în continuare puterea testului Z stabilit la Aplicaţia 6.3.15.

Din nou distingem cele două cazuri.

Dacă $m_0 > m_1$, avem succesiv

$$\begin{aligned}\pi(m_1) &= 1 - \widetilde{\beta} = P\left((X_1, X_2, \ldots, X_n) \in \widetilde{\mathcal{U}} \mid m = m_1\right) \\ &= P\left(\bar{X} \leqslant m_0 + z_\alpha \frac{\sigma}{\sqrt{n}} \;\middle|\; m = m_1\right) \\ &= P\left(\frac{\bar{X} - m_1}{\frac{\sigma}{\sqrt{n}}} \leqslant \frac{m_0 - m_1}{\frac{\sigma}{\sqrt{n}}} + z_\alpha \mid m = m_1\right) = \Phi\left(\frac{m_0 - m_1}{\frac{\sigma}{\sqrt{n}}} + z_\alpha\right),\end{aligned}$$

unde funcţia Φ a lui Laplace este cea din Aplicaţia 6.3.15.

Dacă $m_0 < m_1$, de asemenea, avem

$$\begin{aligned}\pi(m_1) &= 1-\widetilde{\beta} = P\left((X_1, X_2, \ldots, X_n) \in \widetilde{\mathcal{U}} \mid m = m_1\right)\\ &= P\left(\bar{X} \geqslant m_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \mid m = m_1\right)\\ &= P\left(\frac{\bar{X}-m_1}{\frac{\sigma}{\sqrt{n}}} \geqslant \frac{m_0-m_1}{\frac{\sigma}{\sqrt{n}}} + z_{1-\alpha} \mid m = m_1\right) = 1-\Phi\left(\frac{m_0-m_1}{\frac{\sigma}{\sqrt{n}}}+z_{1-\alpha}\right).\end{aligned}$$

Se vede, încă odată, în ambele cazuri, că $\pi(m_1) \to 1$, când $n \to \infty$, adică $\widetilde{\beta} \to 0$.

Să încheiem prin modul de determinare a volumului n al selecţiei, pentru a se atinge o putere a testului (sau un risc de speţa a doua) apriori fixată. Desigur că α este considerat de asemenea dat.

În cazul $m_0 > m_1$, din $P\left((X_1, X_2, \ldots, X_n) \in \widetilde{\mathcal{U}} \mid H_0\right) = \alpha$, am determinat $K_\alpha = m_0 + \frac{\sigma}{\sqrt{n}} z_\alpha$. Pe de altă parte, $P\left((X_1, X_2, \ldots, X_n) \in \widetilde{\mathcal{U}} \mid H_1\right) = 1-\widetilde{\beta}$ se rescrie în felul următor

$$P\left(\frac{\bar{X}-m_1}{\frac{\sigma}{\sqrt{n}}} \leqslant \frac{K_\alpha - m_1}{\frac{\sigma}{\sqrt{n}}} \mid H_1\right) = 1-\widetilde{\beta}, \quad \text{adică} \quad \Phi\left(\frac{K_\alpha - m_1}{\frac{\sigma}{\sqrt{n}}}\right) = 1-\widetilde{\beta},$$

de unde se obţine o altă relaţie ce îl conţine pe K_α, anume $\frac{K_\alpha - m_1}{\frac{\sigma}{\sqrt{n}}} = z_{1-\widetilde{\beta}}$. Comparând cele două relaţii ce îl conţin pe K_α, avem că $m_0 + \frac{\sigma}{\sqrt{n}} z_\alpha = m_1 + \frac{\sigma}{\sqrt{n}} z_{1-\widetilde{\beta}}$, de unde se obţine

$$n = \frac{\sigma^2\left(z_{1-\widetilde{\beta}} - z_\alpha\right)^2}{(m_0-m_1)^2}. \tag{6.3.4}$$

În cazul $m_0 < m_1$, avem următoarele două relaţii, care îl conţin pe K_α,

$$K_\alpha = m_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} \quad \text{şi} \quad \frac{K_\alpha - m_1}{\frac{\sigma}{\sqrt{n}}} = z_{\widetilde{\beta}},$$

de unde se obţine

$$n = \frac{\sigma^2\left(z_{\widetilde{\beta}} - z_{1-\alpha}\right)^2}{(m_0-m_1)^2}. \tag{6.3.5}$$

Exemplul 6.3.19. Să considerăm aceleaşi date de selecţie de la Exemplul 6.3.7. Vom verifica ipoteza nulă $H_0 : m = 3.5$, când se consideră respectiv ipotezele alternative $H_1 : m < 3.5$, $H_1 : m > 3.5$, iar nivelul de semnificaţie este $\alpha = 0.01$. Vom calcula puterea pentru fiecare din testele considerate, când $m = 3.4,\ 3.5,\ 3.6,\ 3.7$, iar apoi vom determina volumul n optim al selecţiei când se consideră alternativa $H_1 : m = 3.4$, respectiv $H_1 : m = 3.7$, cu riscul de speţa a doua fixat $\beta = 0.05$.

Când se consideră alternativa $H_1 : m > 3.5$ se foloseşte testul Z unilateral dreapta. În acest caz, cuantila $z_{1-\alpha} = 2.33$ se determină din $\Phi(z_{1-\alpha}) = \frac{1}{2} - \alpha = 0.49$. Deoarece $z = -0.57 < 2.33 = z_{1-\alpha}$, rezultă că ipoteza nulă este admisă.

Pentru alternativa $H_1 : m < 3.5$ se foloseşte testul Z unilateral stânga. Se determină $z_\alpha = -2.33$ din relaţia $\Phi(-z_\alpha) = \frac{1}{2} - \alpha = 0.49$. Deoarece are loc inegalitatea $z = -0.57 > -2.33 = z_\alpha$, se admite ipoteza H_0 şi în acest ultim caz.

Pentru testul Z unilateral dreapta, formula de calcul a puterii este

$$\pi(m) = \frac{1}{2} - \Phi\left(\frac{m_0 - m}{\frac{\sigma}{\sqrt{n}}} + z_{1-\alpha}\right),$$

şi prin urmare

$$\pi(3.4) = \frac{1}{2} - \Phi\left(0.4\sqrt{6} + 2.33\right) = 0, \quad \pi(3.6) = \frac{1}{2} - \Phi\left(-0.4\sqrt{6} + 2.33\right) = 0.09,$$
$$\pi(3.5) = \frac{1}{2} - \Phi(2.33) = 0.01, \qquad \pi(3.7) = \frac{1}{2} - \Phi\left(-0.8\sqrt{6} + 2.33\right) = 0.36.$$

În cazul testului Z unilateral stânga, avem

$$\pi(m) = \frac{1}{2} + \Phi\left(\frac{m_0 - m}{\frac{\sigma}{\sqrt{n}}} + z_\alpha\right),$$

deci

$$\pi(3.4) = \frac{1}{2} + \Phi\left(0.4\sqrt{6} - 2.33\right) = 0.09, \quad \pi(3.6) = \frac{1}{2} + \Phi\left(-0.4\sqrt{6} - 2.33\right) = 0,$$
$$\pi(3.5) = \frac{1}{2} + \Phi(-2.33) = 0.01, \qquad \pi(3.7) = \frac{1}{2} + \Phi\left(-0.8\sqrt{6} - 2.33\right) = 0.$$

Pentru determinarea volumului optim al selecţiei, când folosim testul unilateral stânga, cu alternativa $H_1 : m = 3.4$, se utilizează formula (6.3.4):

$$n = \frac{\sigma^2 (z_{1-\beta} - z_\alpha)^2}{(m_0 - m_1)^2} = \frac{0.25\,(1.65 + 2.33)^2}{(3.5 - 3.4)^2},$$

de unde se obţine $n = 396$.

Pentru determinarea volumului optim al selecţiei, când folosim testul unilateral dreapta, cu alternativa $H_1 : m = 3.7$, se utilizează formula (6.3.5). Astfel volumul optim n al selecţiei se calculează cu formula:

$$n = \frac{\sigma^2 \left(z_\beta - z_{1-\alpha}\right)^2}{\left(m_0 - m_1\right)^2} = \frac{0.25 \left(-1.65 - 2.33\right)^2}{\left(3.5 - 3.7\right)^2} = \frac{1584}{16},$$

de unde se obţine $n = 99$.

Programul 6.3.20. Programul Matlab următor, efectuează calculele din exemplul precedent, iar în plus, reprezintă grafic curbele putere, în cazul celor două alternative. Punctele de pe curbele putere, care au fost calculate în exemplul precedent, sunt marcate pe Figura 6.2 prin cerculeţe.

```
clear,clf
x=[3.59,4.27,2.83,2.28,3.69,2.81,3.16,3.03,3.65,...
   3.84,4.32,3.21,3.20,3.32,3.82,3.27,3.77,3.56,...
   3.29,4.08,2.97,3.29,3.78,3.58];
z=(mean(x)-3.5)/(0.5/sqrt(24));
fprintf(' z=%6.3f\n',z)
z1=norminv(0,0,1); z2=norminv(0.99,0,1);
fprintf('    m gt 3.5\n')
fprintf(' z1=%6.3f, z2=%6.3f\n',z1,z2)
z1=norminv(0.01,0,1); z2=norminv(1,0,1);
fprintf('    m lt 3.5\n')
fprintf(' z1=%6.3f, z2=%6.3f\n',z1,z2)
m=3:0.01:4; m1=3.4:0.1:3.7;
c2=norminv(0.99,0,1);
pib=1-normcdf(c2+(3.5-m)./(0.5/sqrt(24)),0,1);
pib1(1,:)=1-normcdf(c2+(3.5-m1)./(0.5/sqrt(24)),0,1);
subplot(2,1,1)
plot(m,pib,'k-',m1,pib1(1,:),'o'), grid on
c1=norminv(0.01,0,1);
pib=normcdf(c1+(3.5-m)./(0.5/sqrt(24)),0,1);
pib1(2,:)=normcdf(c1+(3.5-m1)./(0.5/sqrt(24)),0,1);
subplot(2,1,2)
plot(m,pib,'k-',m1,pib1(2,:),'o'), grid on
len=length(m1);
for i=1:len
    for j=1:2
        fprintf(' pi(%3.1f)=%5.3f    ',m1(i),pib1(j,i))
    end
    fprintf('\n')
end
zb=norminv(0.95,0,1); za=norminv(0.01,0,1);
nr=ceil(0.5^2*(zb-za)^2/(3.5-3.4)^2);
zb=norminv(0.05,0,1); za=norminv(0.99,0,1);
nl=ceil(0.5^2*(zb-za)^2/(3.5-3.7)^2);
fprintf(' nr=%3d,  nl=%3d',nr,nl)
```

În urma executării programului, pe lângă graficele curbelor de putere, sunt afişate şi următoarele rezultate:

```
z=-0.567
    m gt 3.5
 z1=  -Inf, z2= 2.326
    m lt 3.5
 z1=-2.326, z2=   Inf
 pi(3.4)=0.000     pi(3.4)=0.089
 pi(3.5)=0.010     pi(3.5)=0.010
 pi(3.6)=0.089     pi(3.6)=0.000
 pi(3.7)=0.357     pi(3.7)=0.000
 nr=395,  nl= 99
```

care reprezintă intervalele numerice `(z1,z2)` şi volumul optim, când se consideră cele două alternative.

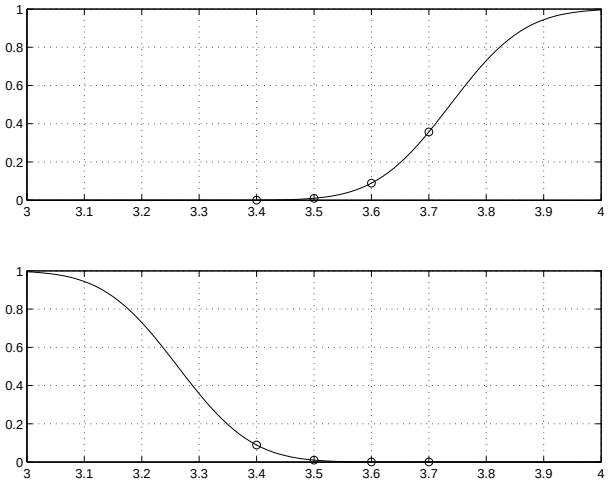

Figura 6.2: Curbele funcţiilor putere

6.4 Testul T (Student) privind media teoretică

Fie caracteristica X ce urmează legea normală $\mathcal{N}(m, \sigma)$, cu ambii parametri $m \in \mathbb{R}$ şi $\sigma > 0$ necunoscuţi. Relativ la valoarea medie a acestei caracteristici, se face ipoteza nulă $H_0 : m = m_0$, cu una din alternativele

$H_1 : m \neq m_0$, când obţinem *testul T bilateral*,

$H_1 : m > m_0$, când obţinem *testul T unilateral dreapta*,

$H_1 : m < m_0$, când obţinem *testul T unilateral stânga*.

Pentru verificarea acestei ipoteze se consideră o selecţie repetată de volum n, cu datele de selecţie $x_1, \ldots, x_n$ şi corespunzător variabilele de selecţie $X_1, \ldots, X_n$. Conform Proprietăţii 4.2.26, statistica

$$T = \frac{\bar{X} - m}{\frac{\bar{\sigma}}{\sqrt{n}}} = \frac{\bar{X} - m}{\sqrt{\frac{\bar{\mu}_2}{n-1}}},$$

unde

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k, \quad \bar{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2 = \frac{n}{n-1}\bar{\mu}_2,$$

urmează legea Student cu $n-1$ grade de libertate.

Prin urmare, pentru nivelul de semnificaţie $\alpha \in (0,1)$ dat, se poate determina intervalul numeric (t_1, t_2) astfel încât

$$P\big(T \in (t_1, t_2) \;\big|\; H_0\big) = F_{n-1}(t_2) - F_{n-1}(t_1) = 1 - \alpha,$$

unde

$$F_m(t) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi}\,\Gamma\left(\frac{m}{2}\right)} \int_{-\infty}^{t} \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}} dx, \quad t \in \mathbb{R},$$

este funcţia de repartiţie pentru legea Student cu m grade de libertate (tabelată în *Anexa II*, pentru anumite valori).

Intervalul numeric (t_1, t_2) pentru statistica T nu este determinat în mod unic din condiţia de mai sus. În funcţie de alternativa H_1 aleasă, se consideră suplimentar:

$t_1 = -t_2,\ t_2 = t_{n-1,1-\frac{\alpha}{2}}$, dacă $H_1 : m \neq m_0$,

$t_1 = -\infty,\ t_2 = t_{n-1,1-\alpha}$, dacă $H_1 : m > m_0$,

$t_1 = t_{n-1,\alpha},\ t_2 = +\infty$, dacă $H_1 : m < m_0$,

unde $t_{m,\gamma}$ este cuantila de ordin γ a legii Student cu m grade de libertate, adică

$$F_m(t_{m,\gamma}) = \gamma.$$

Corespunzător intervalului (t_1, t_2) se consideră respectiv regiunea critice $\mathcal{U}$ definite prin:

$$\mathcal{U} = \left\{(u_1, u_2, \ldots, u_n) \in \mathbb{R}^n \;\middle|\; \frac{|\,\bar{u} - m_0\,|}{\frac{\bar{\sigma}_u}{\sqrt{n}}} \geqslant t_{n-1,1-\frac{\alpha}{2}}\right\},$$

$$\mathcal{U}=\left\{(u_1,u_2,\ldots,u_n)\in\mathbb{R}^n \;\middle|\; \frac{\bar{u}-m_0}{\frac{\bar{\sigma}_u}{\sqrt{n}}}\geqslant t_{n-1,1-\alpha}\right\},$$

$$\mathcal{U}=\left\{(u_1,u_2,\ldots,u_n)\in\mathbb{R}^n \;\middle|\; \frac{\bar{u}-m_0}{\frac{\bar{\sigma}_u}{\sqrt{n}}}\leqslant t_{n-1,\alpha}\right\}.$$

Aici s-au utilizat notaţiile $\bar{u}=\frac{1}{n}\sum_{k=1}^{n}u_k$, respectiv $\bar{\sigma}_u^2=\frac{1}{n-1}\sum_{k=1}^{n}\left(u_k-\bar{u}\right)^2$. Se verifică imediat că $P\left((X_1,X_2,\ldots,X_n)\in\mathcal{U}\,|\,H_0\right)=\alpha$, iar cele trei moduri de definire a regiunii critice $\mathcal{U}$ ne conduc respectiv la testul T bilateral, unilateral dreapta şi unilateral stânga.

Odată construită regiunea critică $\mathcal{U}$, folosind datele de selecţie $x_1,x_2,\ldots,x_n$, ipoteza nulă H_0 va fi admisă dacă $(x_1,x_2,\ldots,x_n)\notin\mathcal{U}$, iar în caz contrar va fi respinsă. Remarcăm de asemenea că regiunea critică $\mathcal{U}$ corespunde mulţimii complementare intervalului (t_1,t_2).

Etapele aplicării testului T

1. Se dau: $\alpha;\ x_1,x_2,\ldots,x_n;\ m_0$;
2. Se calculează intervalul (t_1,t_2) astfel încât $F_{n-1}(t_1)-F_{n-1}(t_2)=1-\alpha$, (după cum s-a prezentat înainte);
3. Se calculează
$$t=\frac{\bar{x}-m_0}{\frac{\bar{\sigma}}{\sqrt{n}}},\quad \text{unde}\quad \bar{x}=\frac{1}{n}\sum_{k=1}^{n}x_k,\quad \bar{\sigma}^2=\frac{1}{n-1}\sum_{k=1}^{n}\left(x_k-\bar{x}\right)^2;$$
4. Concluzia: dacă $t\in(t_1,t_2)$ ipoteza H_0 este admisă, în caz contrar ipoteza este respinsă.

Observaţia 6.4.1. Când numărul gradelor de libertate tinde la infinit, conform teoremei limită centrală, avem că legea Student converge în repartiţie la legea normală $\mathcal{N}(0,1)$. Prin urmare, dacă volumul n al selecţiei este mare $(n>30)$ se poate utiliza testul Z pentru verificarea ipotezei nule $H_0:m=m_0$, prin utilizarea statisticii T în loc de statistica Z. Toate rezultatele de la testul Z rămân, aşadar, adevărate în acest caz.

Observaţia 6.4.2. Având în vedere că funcţia de repartiţie a legii Student este strict crescătoare, iar pentru determinarea intervalului (t_1,t_2) este necesară inversarea acestei funcţii, se poate renunţa la operaţia de inversare. Pentru aceasta se determină ceea ce se numeşte *valoare critică*, ca şi la testul Z, şi pe care o notăm prin c. Vom analiza pe rând cele trei alternative.

Dacă se consideră testul bilateral, făcând notaţia

$$1-\frac{c}{2}=F_{n-1}\left(|t|\right),\quad \text{adică}\quad c=2\left(1-F_{n-1}\left(|t|\right)\right),$$

ipoteza nulă va fi respinsă dacă $1-\frac{c}{2}\geqslant 1-\frac{\alpha}{2}$, adică $c\leqslant\alpha$.

Pentru testul unilateral dreapta, se notează

$$1-c=F_{n-1}\left(t\right),\quad \text{adică}\quad c=1-F_{n-1}\left(t\right),$$

iar ipoteza nulă este respinsă, dacă $1-c\geqslant 1-\alpha$, adică, la fel ca mai înainte, dacă $c\leqslant\alpha$.

Un raţionament analog, pentru testul unilateral stânga ne conduce la valoarea critică, calculată după formulă $c=F_{n-1}\left(t\right)$. Drept urmare, ipoteza nulă este respinsă, dacă $c\leqslant\alpha$.

În rezumat, ipoteza nulă este respinsă, dacă $c\leqslant\alpha$, unde

$$c=\begin{cases}2\left(1-F_{n-1}\left(|t|\right)\right), & \text{dacă } H_1: m\neq m_0,\\ 1-F_{n-1}\left(t\right), & \text{dacă } H_1: m>m_0,\\ F_{n-1}\left(t\right), & \text{dacă } H_1: m<m_0.\end{cases}$$

6.4.1 Funcţia `ttest`

Sistemul Matlab, prin *Statistics toolbox*, dispune de funcţia `ttest`, cu aplicabilitate la testul T. Apelarea acestei funcţii se face prin:

```
h=ttest(x,m0)
h=ttest(x,m0,alpha)
h=ttest(x,m0,alpha,tail)
[h,c,ci]=ttest(x,m0,alpha,tail)
```

În urma executării acestor instrucţiuni, se efectuează testul T asupra datelor conţinute în vectorul `x`, folosind nivelul de semnificaţie `alpha`, care are valoarea implicită `alpha=0.05`.

Parametrul `tail` specifică una din cele trei alternative, care conduc la testul bilateral (`tail=0`, implicit), unilateral dreapta (`tail=1`) şi unilateral stânga (`tail=-1`). Dacă `h=1`, atunci ipoteza nulă va fi respinsă, respectiv dacă `h=0`, ipoteza nu poate fi respinsă.

Ultima formă de apel permite, de asemenea, obţinerea valorii critice `c`, precum şi a intervalului de încredere pentru media teoretică, corespunzător probabilităţii de încredere `1-alpha`, obţinut în vectorul cu două componente `ci`.

Programul 6.4.3. Programul Matlab, ce urmează, aplică testul T, pentru datele din Exemplul 6.2.2, considerând că dispersia este necunoscută. Ipoteza nulă este

$H\colon m = 16$, iar ca ipoteze alternative se consideră, pe rând, cele trei, care conduc, respectiv la testele bilateral, unilateral stânga şi unilateral dreapta. Programul va determina, de asemenea, intervalele de încredere pentru media teoretică, în cele trei cazuri, folosind `alpha=0.01`.

```
x=[11*ones(1,4),13*ones(1,6),15*ones(1,12),...
   17*ones(1,10),20*ones(1,8)];
fprintf(' h    c           ci     \n')
fprintf('_____________________________\n')
for i=-1:1
  [h,c,ci]=ttest(x,16,0.01,i);
  fprintf(' %d  %5.4f   (%4.2f,%4.2f) \n',h,c,ci)
end
```

În urma executării programului, se obţin rezultatele:

```
 h    c           ci
_____________________________
 0  0.3261   (-Inf,16.87)
 0  0.6522   (14.61,16.99)
 0  0.6739   (14.73, Inf)
```

Se observă că pentru `tail=-1` şi `tail=1`, se construiesc intervale nemărginite, respectiv la stânga şi la dreapta.

6.5 Testul raportului verosimilităţilor

Fie caracteristica X cu legea de probabilitate $f(x;\boldsymbol{\theta})$, unde parametrul necunoscut $\boldsymbol{\theta} \in \boldsymbol{A} \subset \mathbb{R}^p$. Relativ la parametrul $\boldsymbol{\theta}$, se consideră ipoteza nulă $H_0 : \boldsymbol{\theta} \in \boldsymbol{A}_0$ cu alternativa $H_1 : \boldsymbol{\theta} \in \boldsymbol{A} \setminus \boldsymbol{A}_0$. Se consideră o selecţie repetată de volum n cu ajutorul căreia se construieşte statistica

$$\widehat{\Lambda} = \widehat{\Lambda}(X_1, \ldots, X_n) = \frac{\sup_{\boldsymbol{\theta} \in \boldsymbol{A}_0} g(X_1, \ldots, X_n; \boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \boldsymbol{A}} g(X_1, \ldots, X_n; \boldsymbol{\theta})},$$

unde

$$g(u_1, \ldots, u_n; \boldsymbol{\theta}) = \prod_{k=1}^{n} f(u_k; \boldsymbol{\theta})$$

este funcţia de verosimilitate. Din felul cum a fost definită statistica $\widehat{\Lambda}$, avem că $0 < \widehat{\Lambda} < 1$, iar ipoteza H_0 va putea fi acceptată dacă $\widehat{\Lambda}$ este apropiată de 1. Folosind această observaţie, pentru un nivel de semnificaţie $\alpha \in (0, 1)$ dat, se determină regiunea critică din relaţia $P\left(\widehat{\Lambda} \leqslant \lambda_\alpha \mid H_0\right) = \alpha$, unde λ_α este cuantila de ordin α, pentru legea de probabilitate a statisticii $\widehat{\Lambda}$.

Exemplul 6.5.1. Fie caracteristica X ce urmează legea normală $\mathcal{N}(m,\sigma)$, unde parametrii $m \in \mathbb{R}$ şi $\sigma > 0$ sunt necunoscuţi. Vrem să verificăm urmatoarea ipoteză nulă $H_0 : (m = m_0, \sigma > 0)$, cu alternativa $H_1 : (m \neq m_0, \sigma > 0)$. Pentru aceasta considerăm o selecţie repetată de volum n şi nivelul de semnificaţie $\alpha \in (0,1)$ fixat. Funcţia de verosimilitate este dată prin

$$g\left(u_1,\ldots,u_n;m,\sigma\right) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{n}\left(u_k-m\right)^2\right].$$

Când ipoteza H_0 este adevărată, atunci estimatorul de verosimilitate maximă pentru σ este dat prin

$$\widehat{\sigma}_1 = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(X_k-m_0\right)^2}.$$

Obţinem astfel, în acest caz, că

$$\begin{aligned}\sup_{\sigma>0} g(X_1,\ldots,X_n;m_0,\sigma) &= \left(\frac{1}{\widehat{\sigma}_1\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\widehat{\sigma}_1^2}\sum_{k=1}^{n}\left(X_k-m_0\right)^2\right]\\ &= \left(\frac{n}{2\pi \mathrm{e}\sum_{k=1}^{n}\left(X_k-m_0\right)^2}\right)^{\frac{n}{2}}.\end{aligned}$$

Pe de altă parte, pentru determinarea numitorului statisticii $\widehat{\Lambda}$, vom căuta estimatorii de verosimilitate maximă pentru parametrii $m \in \mathbb{R}$ şi $\sigma > 0$ necunoscuţi.

Sistemul de verosimilitate maximă

$$\begin{cases}\dfrac{\partial \ln g(X_1,X_2,\ldots,X_n;m,\sigma)}{\partial m} = 0\\[2ex] \dfrac{\partial \ln g(X_1,X_2,\ldots,X_n;m,\sigma)}{\partial \sigma} = 0,\end{cases}$$

ne dă estimatorii de verosimilitate maximă pentru m şi σ, anume

$$\begin{aligned}\widehat{m} &= \frac{1}{n}\sum_{k=1}^{n} X_k = \bar{X},\\ \widehat{\sigma} &= \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2} = \sqrt{\bar{\mu}_2}.\end{aligned}$$

Obţinem astfel că

$$\sup_{m\in\mathbb{R},\sigma>0} g(X_1,X_2,\ldots,X_n;m,\sigma)=\left(\frac{1}{\sqrt{2\pi\bar{\mu}_2}}\right)^n \exp\left[-\frac{1}{2\bar{\mu}_2}\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2\right]$$
$$=\left(\frac{n}{2\pi\mathrm{e}\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2}\right)^{\frac{n}{2}}.$$

Putem scrie acum statistica raportului verosimilităţilor

$$\widehat{\Lambda}=\left(\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2\Big/\sum_{k=1}^{n}\left(X_k-m_0\right)^2\right)^{\frac{n}{2}}.$$

Folosind formula lui König (Observaţia 3.3.23):

$$\sum_{k=1}^{n}(X_k-m_0)^2=\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2+n(\bar{X}-m_0)^2,$$

avem că

$$\widehat{\Lambda}=\left(1+\frac{n(\bar{X}-m_0)^2}{\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2}\right)^{-\frac{n}{2}}.$$

Dacă se introduce statistica T definită prin

$$T=\frac{\bar{X}-m_0}{\frac{\tilde{\sigma}}{\sqrt{n}}}=\frac{\bar{X}-m_0}{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2}\Big/\sqrt{n}},$$

atunci

$$\widehat{\Lambda}=\left(\frac{1}{1+\frac{T^2}{n-1}}\right)^{\frac{n}{2}}.$$

Regiunea critică $\mathcal{U}$, pentru $\alpha\in(0,1)$, se obţine din $P\left(\widehat{\Lambda}\leqslant\lambda_\alpha\;\middle|\;H_0\right)=\alpha$, care este acelaşi lucru cu $P\left(|\,T\,|<t\;\middle|\;m=m_0\right)=1-\alpha$. Altfel spus, determinarea cuantilei λ_α revine la determinarea lui $t>0$, astfel încât $P\left(|\,T\,|<t\;\middle|\;m=m_0\right)=1-\alpha$. Dar statistica T, în ipoteza H_0, urmează legea Student cu $n-1$ grade de libertate (Proprietatea 4.2.26), deci $t=t_{n-1,1-\frac{\alpha}{2}}$, adică cuantila de ordin $1-\frac{\alpha}{2}$ a legii Student cu $n-1$ grade de libertate. Putem scrie în final regiunea critică:

$$\mathcal{U}=\left\{(u_1,u_2,\ldots,u_n)\in\mathbb{R}^n\;\middle|\;\frac{|\bar{u}-m_0|}{\frac{\tilde{\sigma}}{\sqrt{n}}}\geqslant t_{n-1,1-\frac{\alpha}{2}}\right\},$$

adică chiar regiunea critică de la testul T bilateral. Prin urmare, cu această metodă obţinem testul T bilateral.

Teorema 6.5.2. *Fie statistica $\widehat{\Lambda}$ a raportului verosimilităţilor, atunci statistica $-2\ln\widehat{\Lambda}$ urmează legea χ^2 cu p grade de libertate, când $n\to\infty$, p fiind numărul parametrilor necunoscuţi.*

Demonstraţie. Vom considera numai cazul $p=1$.

Folosind formula lui Taylor cu doi termeni avem că

$$\ln g(X_1,\ldots,X_n;\theta_0)-\ln g(X_1,\ldots,X_n;\widehat{\Theta})$$
$$=(\theta_0-\widehat{\Theta})\frac{\partial\ln g(X_1,\ldots,X_n;\widehat{\Theta})}{\partial\theta}+\frac{1}{2}(\theta_0-\widehat{\Theta})^2\frac{\partial^2\ln g(X_1,\ldots,X_n;\xi)}{\partial\theta^2},$$

unde $\xi\in(\theta_0,\widehat{\Theta})$ sau $\xi\in(\widehat{\Theta},\theta_0)$. Dar $\widehat{\Theta}$ este estimator de verosimilitate maximă pentru θ, prin urmare

$$\frac{\partial\ln g(X_1,\ldots,X_n;\widehat{\Theta})}{\partial\theta}=0$$

deci avem că

$$-2\ln\widehat{\Lambda}=-(\theta_0-\widehat{\Theta})^2\,\frac{\partial^2\ln g(X_1,\ldots,X_n;\xi)}{\partial\theta^2}.$$

Pe de altă parte, dacă ipoteza $H_0:\theta=\theta_0$ este adevărată, se ştie că $\widehat{\Theta}\xrightarrow{\text{a.s.}}\theta_0$, deci $\xi\xrightarrow{\text{a.s.}}\theta_0$, când $n\to\infty$. Putem scrie atunci că

$$\frac{\partial^2\ln g(X_1,\ldots,X_n;\xi)}{\partial\theta^2}\cong\frac{\partial^2\ln g(X_1,\ldots,X_n;\theta_0)}{\partial\theta^2}$$
$$=\sum_{k=1}^{n}\frac{\partial^2\ln f(X_k;\theta_0)}{\partial\theta^2}=n\cdot\frac{1}{n}\sum_{k=1}^{n}\frac{\partial^2\ln f(X_k;\theta_0)}{\partial\theta^2}.$$

Folosind legea numerelor mari avem că

$$\frac{1}{n}\sum_{k=1}^{n}\frac{\partial^2\ln f(X_k;\theta_0)}{\partial\theta^2}\xrightarrow{\text{a.s.}}E\left[\frac{\partial^2\ln f(X;\theta_0)}{\partial\theta^2}\right]=-I_1(\theta_0),$$

de unde

$$\frac{\partial^2\ln g(X_1,\ldots,X_n;\xi)}{\partial\theta^2}\cong-nI_1(\theta_0)=-I_n(\theta_0),$$

sau

$$-2\ln\widehat{\Lambda}\cong\left(\theta_0-\widehat{\Theta}\right)^2I_n\left(\theta_0\right).$$

Conform Observaţiei 5.3.11, statistica $\frac{\widehat{\Theta}-\theta_0}{\sqrt{I_n^{-1}(\theta_0)}}$ urmează legea normală $\mathcal{N}(0,1)$, când $n\to\infty$, de unde avem că statistica $(\theta_0-\widehat{\Theta})^2I_n(\theta_0)$ urmează legea χ^2 cu un grad de libertate, ceea ce trebuia arătat. □

6.6 Testul χ^2 privind dispersia teoretică

Fie caracteristica X ce urmează legea normală $\mathcal{N}(m, \sigma)$, unde $\sigma^2 = Var(X)$ este necunoscută şi $m \in \mathbb{R}$, de asemenea necunoscut.

Relativ la dispersia teoretică se face ipoteza nulă $H_0 : \sigma^2 = \sigma_0^2$, cu una din alternativele:

$H_1 : \sigma^2 \neq \sigma_0^2$, când obţinem *testul* χ^2 *bilateral*,

$H_1 : \sigma^2 > \sigma_0^2$, când obţinem *testul* χ^2 *unilateral dreapta*,

$H_1 : \sigma^2 < \sigma_0^2$, când obţinem *testul* χ^2 *unilateral stânga*.

Pentru verificarea ipotezei nule H_0, cu una din alternativele H_1 precizate, se consideră o selecţie repetată de volum n, cu datele de selecţie $x_1, x_2, \ldots, x_n$ şi corespunzător variabilele de selecţie $X_1, X_2, \ldots, X_n$. Conform Proprietăţii 4.2.25, avem că statistica

$$\chi^2 = \frac{1}{\sigma^2} \sum_{k=1}^{n} \left(X_k - \bar{X} \right)^2 = \frac{(n-1)\,\bar{\sigma}^2}{\sigma^2},$$

urmează legea χ^2 cu $n-1$ grade de libertate.

Folosind un nivel de semnificaţie $\alpha \in (0,1)$ dat, se poate determina un interval numeric $\left(\chi_1^2, \chi_2^2\right)$ astfel încât

$$P\left(\chi^2 \in \left(\chi_1^2, \chi_2^2\right) \mid H_0\right) = F_{n-1}\left(\chi_2^2\right) - F_{n-1}\left(\chi_1^2\right) = 1 - \alpha,$$

unde

$$F_m(x) = \frac{1}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)} \int_0^x t^{\frac{m}{2}-1} e^{-\frac{t}{2}} dt, \quad x > 0,$$

este funcţia de repartiţie pentru legea χ^2 cu m grade de libertate (tabelată pentru anumite valori în *Anexa III)*.

Intervalul numeric $\left(\chi_1^2, \chi_2^2\right)$ pentru statistica χ^2 nu este determinat în mod unic din condiţia de mai sus. În funcţie de alternativa H_1 aleasă se consideră suplimentar:

$\chi_1^2 = \chi_{n-1,\frac{\alpha}{2}}^2,\ \chi_2^2 = \chi_{n-1,1-\frac{\alpha}{2}},$ dacă $H_1 : \sigma^2 \neq \sigma_0^2$,

$\chi_1^2 = 0,\ \chi_2^2 = \chi_{n-1,1-\alpha}^2,$ dacă $H_1 : \sigma^2 > \sigma_0^2$,

$\chi_1^2 = \chi_{n-1,\alpha}^2,\ \chi_2^2 = +\infty,$ dacă $H_1 : \sigma^2 < \sigma_0^2$,

unde $\chi_{m,\gamma}^2$ este cuantila de ordin γ a legii χ^2 cu m grade de libertate, adică $F_m\left(\chi_{m,\gamma}^2\right) = \gamma$.

Cu ajutorul intervalului numeric $\left(\chi_1^2, \chi_2^2\right)$, astfel determinat, se consideră respectiv regiunile critice:

$$\mathcal{U} = \left\{ (u_1, u_2, \ldots, u_n) \in \mathbb{R}^n \;\middle|\; \frac{1}{\sigma_0^2} \sum_{k=1}^{n} (u_k - \bar{u})^2 \notin \left(\chi^2_{n-1,\frac{\alpha}{2}}, \chi^2_{n-1,1-\frac{\alpha}{2}} \right) \right\},$$

$$\mathcal{U} = \left\{ (u_1, u_2, \ldots, u_n) \in \mathbb{R}^n \;\middle|\; \frac{1}{\sigma_0^2} \sum_{k=1}^{n} (u_k - \bar{u})^2 \geqslant \chi^2_{n-1,1-\alpha} \right\},$$

$$\mathcal{U} = \left\{ (u_1, u_2, \ldots, u_n) \in \mathbb{R}^n \;\middle|\; \frac{1}{\sigma_0^2} \sum_{k=1}^{n} (u_k - \bar{u})^2 \leqslant \chi^2_{n-1,\alpha} \right\}.$$

Uşor se verifică faptul că $P\big((X_1, X_2, \ldots, X_n) \in \mathcal{U} \,|\, H_0\big) = \alpha$, iar cele trei moduri de definire a regiunii critice $\mathcal{U}$ ne conduc respectiv la testul χ^2 bilateral, unilateral dreapta şi unilateral stânga.

Odată construită regiunea critică $\mathcal{U}$, folosind datele de selecţie, ipoteza nulă H_0 va fi admisă dacă $(x_1, x_2, \ldots, x_n) \notin \mathcal{U}$, iar în caz contrar va fi respinsă. Remarcăm de asemenea că regiunea critică $\mathcal{U}$ corespunde mulţimii complementare intervalului $\left(\chi_1^2, \chi_2^2\right)$.

Etapele aplicării testului χ^2

1. Se dau: $\alpha;\ x_1, x_2, \ldots, x_n;\ \sigma = \sigma_0$;
2. Se determină intervalul $\left(\chi_1^2, \chi_2^2\right)$ astfel încât $F_{n-1}\left(\chi_2^2\right) - F_{n-1}\left(\chi_1^2\right) = 1 - \alpha$, (după cum s-a precizat înainte);
3. Se calculează $\chi^2 = \dfrac{1}{\sigma_0^2} \displaystyle\sum_{k=1}^{n} (\, x_k - \bar{x}\,)^2\,, \quad$ unde $\quad \bar{x} = \dfrac{1}{n} \displaystyle\sum_{k=1}^{n} x_k$;
4. Concluzia: dacă $\chi^2 \in \left(\chi_1^2, \chi_2^2\right)$ ipoteza H_0 este admisă, în caz contrar este respinsă.

Exemplul 6.6.1. Se consideră caracteristica X ce urmează legea normală $\mathcal{N}(m, \sigma)$ cu parametrii $m \in \mathbb{R}$ şi $\sigma > 0$ necunoscuţi. Relativ la caracteristica X se consideră o selecţie repetată de volum $n = 15$. Fie datele de selecţie

24.03	25.92	26.92	23.04	25.44	26.31	26.13	25.97	24.44	24.82
24.17	25.53	23.78	26.06	24.29.					

Vrem să verificăm ipoteza nulă $H_0 : \sigma = 1.8$, cu fiecare din următoarele alternative $H_1 : \sigma \neq 1.8$, $H_1 : \sigma > 1.8$, $H_1 : \sigma < 1.8$, când se consideră $\alpha = 0.05$. Vom calcula apoi valorile funcţiei putere pentru $\sigma = 1.5,\ 1.6,\ 1.7,\ 1.8,\ 1.9$, pentru fiecare din alternativele considerate.

Când $H_1 : \sigma \neq 1.8$ avem testul χ^2 bilateral. Se calculează cuantilele

$$\chi^2_{n-1,\frac{\alpha}{2}} = \chi^2_{14,0.025} = 5.63 \quad \text{şi} \quad \chi^2_{n-1,1-\frac{\alpha}{2}} = \chi^2_{14,0.975} = 26.12.$$

Pe de altă parte avem

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{15}\left(24.03 + 25.92 + \cdots + 24.29\right) = 25.11,$$

$$\bar{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \bar{x}\right)^2 = \frac{1}{14}\left[(24.03-25.11)^2 + \ldots + (24.29-25.11)^2\right] = 1.27.$$

Deoarece

$$\chi^2 = \frac{(n-1)\,\bar{\sigma}^2}{\sigma_0^2} = \frac{14 \cdot 1.27}{1.8^2} = 5.49 \notin (5.63, 26.12)\,,$$

ipoteza H_0 este respinsă.

Când $H_1 : \sigma > 1.8$, avem testul χ^2 unilateral dreapta şi se calculează cuantila $\chi^2_{n-1,1-\alpha} = \chi^2_{14,0.95} = 23.69$. Astfel, având în vedere că $\chi^2 = 5.49 < 23.69$, ipoteza nulă este admisă.

Când $H_1 : \sigma < 1.8$, avem testul χ^2 unilateral stânga. În acest caz se calculează cuantila $\chi^2_{n-1,\alpha} = \chi^2_{14,0.05} = 6.57$ şi deoarece $\chi^2 = 5.49 < 6.57$, ipoteza nulă este respinsă.

Când se consideră testul χ^2 bilateral, puterea se obţine din

$$\begin{aligned} 1 - \pi(\sigma) = \beta &= P\left(\chi^2_{n-1,\frac{\alpha}{2}} < \frac{(n-1)\,\bar{\sigma}^2}{\sigma_0^2} < \chi^2_{n-1,1-\frac{\alpha}{2}} \;\middle|\; H_1\right) \\ &= P\left(\frac{\sigma_0^2}{\sigma^2}\,\chi^2_{n-1,\frac{\alpha}{2}} < \frac{(n-1)\,\bar{\sigma}^2}{\sigma^2} < \frac{\sigma_0^2}{\sigma^2}\,\chi^2_{n-1,1-\frac{\alpha}{2}} \;\middle|\; H_1\right), \end{aligned}$$

deci

$$\pi(\sigma) = 1 - F_{n-1}\left(\frac{\sigma_0^2}{\sigma^2}\,\chi^2_{n-1,1-\frac{\alpha}{2}}\right) + F_{n-1}\left(\frac{\sigma_0^2}{\sigma^2}\,\chi^2_{n-1,\frac{\alpha}{2}}\right).$$

În cazul de faţă, obţinem

$$\pi(1.5) = 1 - F_{14}\left(\frac{1.8^2}{1.5^2}\cdot 26.14\right) + F_{14}\left(\frac{1.8^2}{1.5^2}\cdot 5.63\right) = 0.117,$$

$$\pi(1.6) = 1 - F_{14}\left(\frac{1.8^2}{1.6^2}\cdot 26.14\right) + F_{14}\left(\frac{1.8^2}{1.6^2}\cdot 5.63\right) = 0.073,$$

$$\pi(1.7) = 1 - F_{14}\left(\frac{1.8^2}{1.7^2}\cdot 26.14\right) + F_{14}\left(\frac{1.8^2}{1.7^2}\cdot 5.63\right) = 0.052,$$

$$\pi(1.8) = 1 - F_{14}\left(\frac{1.8^2}{1.8^2} \cdot 26.14\right) + F_{14}\left(\frac{1.8^2}{1.8^2} \cdot 5.63\right) = 0.050,$$
$$\pi(1.9) = 1 - F_{14}\left(\frac{1.8^2}{1.9^2} \cdot 26.14\right) + F_{14}\left(\frac{1.8^2}{1.9^2} \cdot 5.63\right) = 0.068.$$

Pentru testul unilateral dreapta, în mod analog, se ajunge la

$$\pi(\sigma) = 1 - F_{n-1}\left(\frac{\sigma_0^2}{\sigma^2}\,\chi^2_{n-1,1-\alpha}\right),$$

de unde

$$\pi(1.5) = 1 - F_{14}\left(\frac{1.8^2}{1.5^2} \cdot 23.69\right) = 0.002,$$
$$\pi(1.6) = 1 - F_{14}\left(\frac{1.8^2}{1.6^2} \cdot 23.69\right) = 0.008,$$
$$\pi(1.7) = 1 - F_{14}\left(\frac{1.8^2}{1.7^2} \cdot 23.69\right) = 0.022,$$
$$\pi(1.8) = 1 - F_{14}\left(\frac{1.8^2}{1.8^2} \cdot 23.69\right) = 0.050,$$
$$\pi(1.9) = 1 - F_{14}\left(\frac{1.8^2}{1.9^2} \cdot 23.69\right) = 0.095.$$

Când se consideră testul χ^2 unilateral stânga, se obţine

$$\pi(\sigma) = F_{n-1}\left(\frac{\sigma_0^2}{\sigma^2}\,\chi^2_{n-1,\alpha}\right),$$

deci

$$\pi(1.5) = F_{14}\left(\frac{1.8^2}{1.5^2} \cdot 6.57\right) = 0.200, \quad \pi(1.8) = F_{14}\left(\frac{1.8^2}{1.8^2} \cdot 6.57\right) = 0.050,$$
$$\pi(1.6) = F_{14}\left(\frac{1.8^2}{1.6^2} \cdot 6.57\right) = 0.128, \quad \pi(1.9) = F_{14}\left(\frac{1.8^2}{1.9^2} \cdot 6.57\right) = 0.031.$$
$$\pi(1.7) = F_{14}\left(\frac{1.8^2}{1.7^2} \cdot 6.57\right) = 0.080,$$

Programul 6.6.2. Programul următor, efectuează calculele cerute în exemplul precedent. În plus, reprezintă grafic funcţia putere, cea din Figura 6.3, în cazul testului bilateral, marcând prin cerculeţe punctele curbei putere, care au fost determinate în exemplu.

```
clf
x=[24.03,25.92,26.92,23.04,25.44,26.31,...
   26.13,25.97,24.44,24.82,24.17,25.53,...
   23.78,26.06,24.29];
x2=14*var(x)/1.8^2;
c1=chi2inv(0.025,14); c2=chi2inv(0.975,14);
fprintf(' x2=%6.3f\n',x2)
fprintf('    sigma ne 1.8\n')
fprintf(' c1=%6.3f, c2=%6.3f\n',c1,c2)
c1=chi2inv(0,14); c2=chi2inv(0.95,14);
fprintf('    sigma gt 1.8\n')
fprintf(' c1=%6.3f, c2=%6.3f\n',c1,c2)
c1=chi2inv(0.05,14); c2=chi2inv(1,14);
fprintf('    sigma lt 1.8\n')
fprintf(' c1=%6.3f, c2=%6.3f\n',c1,c2)
s=1.2:0.01:2.4; s1=1.5:0.1:1.9;
c1=1.8^2*chi2inv(0.025,14); c2=1.8^2*chi2inv(0.975,14);
pib=1-chi2cdf(c2./s.^2,14)+chi2cdf(c1./s.^2,14);
pib1(1,:)=1-chi2cdf(c2./s1.^2,14)+chi2cdf(c1./s1.^2,14);
plot(s,pib,'k-',s1,pib1(1,:),'o'), grid on
c1=1.8^2*chi2inv(0,14); c2=1.8^2*chi2inv(0.95,14);
pib=1-chi2cdf(c2./s.^2,14)+chi2cdf(c1./s.^2,14);
pib1(2,:)=1-chi2cdf(c2./s1.^2,14)+chi2cdf(c1./s1.^2,14);
c1=1.8^2*chi2inv(0.05,14); c2=1.8^2*chi2inv(1,14);
pib=1-chi2cdf(c2./s.^2,14)+chi2cdf(c1./s.^2,14);
pib1(3,:)=1-chi2cdf(c2./s1.^2,14)+chi2cdf(c1./s1.^2,14);
len=length(s1);
for i=1:len
    for j=1:3
        fprintf(' pi(%3.1f)=%5.3f    ',s1(i),pib1(j,i))
    end
fprintf('\n')
end
```

În urma executării programului s–au obţinut rezultatele:

```
 x2= 5.446
    sigma ne 1.8
 c1= 5.629, c2=26.119
    sigma gt 1.8
 c1= 0.000, c2=23.685
    sigma lt 1.8
 c1= 6.571, c2=   Inf
 pi(1.5)=0.117     pi(1.5)=0.002     pi(1.5)=0.200
 pi(1.6)=0.073     pi(1.6)=0.008     pi(1.6)=0.128
 pi(1.7)=0.052     pi(1.7)=0.022     pi(1.7)=0.080
 pi(1.8)=0.050     pi(1.8)=0.050     pi(1.8)=0.050
 pi(1.9)=0.068     pi(1.9)=0.095     pi(1.9)=0.031
```

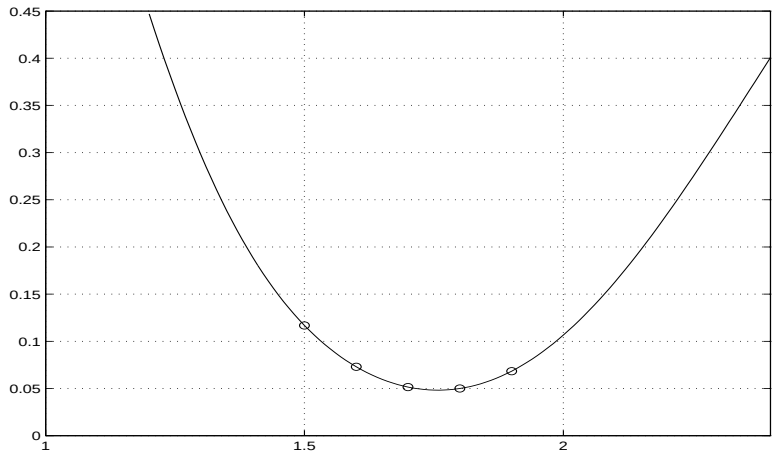

Figura 6.3: Curba funcţiei putere - cazul bilateral

Observaţia 6.6.3. Când caracteristica X nu urmează legea normală, pentru a verifica ipoteza nulă de forma dinainte, cu una din alternativele precizate, unde dispersia teoretică este $\sigma^2 = Var\,(X)$, se ţine seama de faptul că statistica

$$S^2 = \frac{\bar{\sigma}^2 - \sigma^2}{\frac{\sigma^2\sqrt{2}}{n}}, \quad \text{unde} \quad \bar{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2,$$

urmează legea normală $\mathcal{N}\,(0,1)$, când $n \to \infty$.

De exemplu, dacă ipoteza alternativă este $H_1 : \sigma^2 \neq \sigma_0^2$, se va ajunge la regiunea critică

$$\mathcal{U} = \left\{ (u_1, u_2, \ldots, u_n) \in \mathbb{R}^n \;\middle|\; \sqrt{n}\frac{|\,\bar{\sigma}_u^2 - \sigma_0^2\,|}{\sigma_0^2\sqrt{2}} \geqslant z_{1-\frac{\alpha}{2}} \right\},$$

unde $z_{1-\frac{\alpha}{2}}$ este cuantila de ordin $1 - \frac{\alpha}{2}$ de la legea normală $\mathcal{N}\,(0,1)$.

Observaţia 6.6.4. Dacă se consideră statistica

$$H^2 = \frac{\bar{\sigma}^2}{\sigma^2} = \frac{1}{(n-1)\,\sigma^2}\sum_{k=1}^{n}\left(X_k - \bar{X}\right)^2,$$

atunci între statisticile H^2 şi χ^2 există relaţia $\chi^2 = (n-1)\,H^2$.

Deoarece se cunoaşte legea de probabilitate pentru statistica χ^2(legea χ^2 cu $n-1$ grade de libertate), se poate determina şi legea de probabilitate a statisticii H^2. Sunt manuale care conţin tabelată legea de probabilitate a lui H^2. Descrierea testului χ^2, folosind statistica H^2, urmează aceeaşi cale ca cea când se foloseşte statistica χ^2.

Observaţia 6.6.5. Când se cunoaşte parametrul $m \in \mathbb{R}$ (ceea ce se întâmplă mai rar) se poate considera statistica

$$\chi^2 = \frac{1}{\sigma^2} \sum_{k=1}^{n} (X_k - m)^2 = \sum_{k=1}^{n} \left(\frac{X_k - m}{\sigma} \right)^2 .$$

Deoarece $\frac{X_k - m}{\sigma}$ urmează legea normală $\mathcal{N}(0,1)$, avem că statistica χ^2 urmează legea χ^2 cu n grade de libertate. Cele prezentate mai înainte pot fi rescrise cu această statistică.

6.7 Testul F (Fisher–Snedecor) pentru compararea dispersiilor

Se consideră două populaţii independente $\mathcal{C}'$ şi $\mathcal{C}''$ cercetate din punct de vedere al aceleaşi caracteristici. Această caracteristică este X' pentru $\mathcal{C}'$ şi urmează legea normală $\mathcal{N}(m', \sigma')$ şi respectiv X'' pentru $\mathcal{C}''$ şi urmează legea normală $\mathcal{N}(m'', \sigma'')$. Relativ la dispersiile teoretice ale celor două caracteristici se face ipoteza nulă compusă $H_0 : \sigma'^2 = \sigma''^2$, cu una din alternativele:

$H_1 : \sigma'^2 \neq \sigma''^2$, când obţinem *testul F bilateral*,

$H_1 : \sigma'^2 > \sigma''^2$, când obţinem *testul F unilateral dreapta*,

$H_1 : \sigma'^2 < \sigma''^2$, când obţinem *testul F unilateral stânga*.

Pentru verificarea ipotezei nule H_0, cu una din alternativele H_1 considerate, se efectuează câte o selecţie repetată de volume respectiv n' şi n'' din cele două populaţii $\mathcal{C}'$ şi $\mathcal{C}''$. Notăm datele de selecţie prin $x'_1, x'_2, \ldots, x'_{n'}$ şi respectiv $x''_1, x''_2, \ldots, x''_{n''}$, cu variabilele de selecţie $X'_1, X'_2, \ldots, X'_{n'}$ şi $X''_1, X''_2, \ldots, X''_{n''}$. Conform Proprietăţii 4.2.30, statistica

$$F = \frac{\bar{\sigma}'^2}{\sigma'^2} \Bigg/ \frac{\bar{\sigma}''^2}{\sigma''^2},$$

unde

$$\bar{\sigma}'^2 = \frac{1}{n'-1} \sum_{k=1}^{n'} \left(X'_k - \bar{X}' \right)^2, \qquad \bar{X}' = \frac{1}{n'} \sum_{k=1}^{n'} X'_k,$$

$$\bar{\sigma}''^2 = \frac{1}{n''-1} \sum_{k=1}^{n''} \left(X''_k - \bar{X}'' \right)^2, \quad \bar{X}'' = \frac{1}{n''} \sum_{k=1}^{n''} X''_k,$$

urmează legea Fisher–Snedecor cu $m = n' - 1$ şi $n = n'' - 1$ grade de libertate.

Pentru un nivel de semnificaţie $\alpha \in (0,1)$ fixat, se poate determina un interval numeric (f_1, f_2), astfel ca $P\left(F \in (f_1, f_2) \mid H_0\right) = F_{m,n}(f_2) - F_{m,n}(f_1) = 1 - \alpha$, unde

$$F_{m,n}(f) = \left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \int_0^f x^{\frac{n}{2}-1}\left(1 + \frac{m}{n}x\right)^{-\frac{m+n}{2}} dx, \quad f > 0,$$

este funcţia de repartiţie pentru legea Fisher–Snedecor cu m şi n grade de libertate (tabelată pentru anumite valori în *Anexa IV*). Intervalul de încredere (f_1, f_2) pentru statistica F nu este unic determinat. În funcţie de alternativa H_1 aleasă, se consideră suplimentar:

$$f_1 = f_{m,n;\frac{\alpha}{2}},\ f_2 = f_{m,n;1-\frac{\alpha}{2}}, \quad \text{dacă } H_1 : \sigma'^2 \neq \sigma''^2,$$

$$f_1 = 0,\ f_2 = f_{m,n;1-\alpha}, \quad \text{dacă } H_1 : \sigma'^2 > \sigma''^2,$$

$$f_1 = f_{m,n;\alpha},\ f_2 = +\infty, \quad \text{dacă } H_1 : \sigma'^2 < \sigma''^2.$$

Cu ajutorul intervalului numeric (f_1, f_2) astfel determinat, se consideră respectiv regiunile critice

$$\mathcal{U} = \left\{ (u_1, \ldots, u_{n'}; v_1, \ldots, v_{n''}) \in \mathbb{R}^{n'+n''} \;\middle|\; \frac{\bar{\sigma}_u^2}{\bar{\sigma}_v^2} \notin \left(f_{m,n;\frac{\alpha}{2}}, f_{m,n;1-\frac{\alpha}{2}}\right) \right\},$$

$$\mathcal{U} = \left\{ (u_1, \ldots, u_{n'}; v_1, \ldots, v_{n''}) \in \mathbb{R}^{n'+n''} \;\middle|\; \frac{\bar{\sigma}_u^2}{\bar{\sigma}_v^2} \geqslant f_{m,n;1-\alpha} \right\},$$

$$\mathcal{U} = \left\{ (u_1, \ldots, u_{n'}; v_1, \ldots, v_{n''}) \in \mathbb{R}^{n'+n''} \;\middle|\; \frac{\bar{\sigma}_u^2}{\bar{\sigma}_v^2} \leqslant f_{m,n;\alpha} \right\}.$$

S-au folosit notaţiile

$$\bar{\sigma}_u^2 = \frac{1}{n'-1}\sum_{k=1}^{n'} (u_k - \bar{u})^2, \quad \bar{u} = \frac{1}{n'}\sum_{k=1}^{n'} u_k,$$

$$\bar{\sigma}_v^2 = \frac{1}{n''-1}\sum_{k=1}^{n''} (v_k - \bar{v})^2, \quad \bar{v} = \frac{1}{n''}\sum_{k=1}^{n''} v_k.$$

Se verifică imediat că $P\left((X'_1, X'_2, \ldots, X'_{n'}; X''_1, X''_2, \ldots, X''_{n''}) \in \mathcal{U} \mid H_0\right) = \alpha$, iar cele trei alternative ne conduc la cele trei regiuni critice, care definesc respectiv testul F bilateral, unilateral dreapta şi unilateral stânga.

Odată construită regiunea critică $\mathcal{U}$, folosind datele de selecţie, ipoteza nulă va fi admisă dacă $\left(x'_1, x'_2, \ldots, x'_{n'}; x''_1, x''_2, \ldots, x''_{n''}\right) \notin \mathcal{U}$, iar în caz contrar va fi respinsă. Remarcăm de asemenea că regiunea critică $\mathcal{U}$ corespunde mulţimii complementare intervalului (f_1, f_2).

Etapele aplicării testului *F*

1. Se dau: $\alpha;\ x'_1, x'_2, \ldots, x'_{n'};\ x''_1, x''_2, \ldots, x''_{n''}$;
2. Se determină intervalul (f_1, f_2) astfel încât $F_{m,n}(f_2) - F_{m,n}(f_1) = 1 - \alpha$ (după cum s-a prezentat înainte);
3. Se calculează $f = \dfrac{\bar{\sigma}'^2}{\bar{\sigma}''^2}$, unde
$$\bar{\sigma}'^2 = \frac{1}{n'-1}\sum_{k=1}^{n'}\left(x'_k - \bar{x}'\right)^2, \qquad \bar{x}' = \frac{1}{n'}\sum_{k=1}^{n'} x'_k,$$
$$\bar{\sigma}''^2 = \frac{1}{n''-1}\sum_{k=1}^{n''}\left(x''_k - \bar{x}''\right)^2, \quad \bar{x}'' = \frac{1}{n''}\sum_{k=1}^{n''} x''_k;$$
4. Concluzia: dacă $f \in (f_1, f_2)$ ipoteza H_0 este admisă, în caz contrar este respinsă.

Observaţia 6.7.1. Dacă se notează prin $\theta = \frac{\sigma'}{\sigma''}$, atunci ipoteza nulă se poate rescrie $H_0 : \theta^2 = 1$, iar ipotezele alternative se scriu corespunzător $H_1 : \theta^2 \neq 1$, $H_1 : \theta^2 > 1$, respectiv $H_1 : \theta^2 < 1$, iar statistica F, se scrie sub forma $F = \frac{1}{\theta^2}\frac{\bar{\sigma}'^2}{\bar{\sigma}''^2}$.

Exemplul 6.7.2. Se cercetează precizia cu care două maşini produc conserve de acelaşi tip. Pentru aceasta, se consideră câte un eşantion din conservele produse de cele două maşini şi se măsoară greutatea acestora. Fie X' greutatea în grame a unei conserve produsă de prima maşină, respectiv X'' pentru a doua maşină. Măsurătorile obţinute sunt:

X' :	1021	980	988	1017	1005	998	1014	985	995	1004	1030
		1015	995	1023	1008	1013					
X'' :	1003	988	993	1013	1006	1002	1014	997	1002	1010	975

Considerând nivelul de semnificaţie $\alpha = 0.05$, vrem să verificăm ipoteza nulă compusă $H_0 : \sigma' = \sigma''$, respectiv cu fiecare din alternativele $H_1 : \sigma' \neq \sigma''$, $H_1 : \sigma' > \sigma''$, $H_1 : \sigma' < \sigma''$. Vom considera că cele două caracteristici X' şi X'' sunt independente şi că urmează legile normale $\mathcal{N}(m', \sigma')$ şi respectiv $\mathcal{N}(m'', \sigma'')$.

Se consideră statistica
$$F = \frac{\bar{\sigma}'^2}{\sigma'^2} \Big/ \frac{\bar{\sigma}''^2}{\sigma''^2},$$

ce urmează legea Fisher–Snedecor cu $m = n' - 1$ şi $n = n'' - 1$ grade de libertate, care ne conduce la testul F. Pentru aceasta avem

$$\bar{x}\,' = \frac{1}{n'}\sum_{k=1}^{n'} x'_k = \frac{1}{16}\,(1021 + 980 + \cdots + 1013) = 1005.7,$$

$$\bar{x}\,'' = \frac{1}{n''}\sum_{k=1}^{n''} x''_k = \frac{1}{11}\,(1003 + 988 + \cdots + 975) = 1000.3,$$

$$\bar{\sigma}\,'^2 = \frac{1}{n'-1}\sum_{k=1}^{n'}\left(x'_k - \bar{x}\,'\right)^2 = \frac{1}{15}\left[(1021 - 1005.7)^2 + \cdots + (1013 - 1005.7)^2\right]$$
$$= 210.63,$$

$$\bar{\sigma}\,''^2 = \frac{1}{n''-1}\sum_{k=1}^{n''}\left(x''_k - \bar{x}\,''\right)^2 = \frac{1}{10}\left[(1003 - 1000.3)^2 + \cdots + (975 - 1000.3)^2\right]$$
$$= 134.42.$$

Valoarea calculată a statisticii F este

$$f = \frac{\bar{\sigma}\,'^2}{\bar{\sigma}\,''^2} = \frac{210.63}{134.42} = 1.57.$$

De asemenea, avem calculate, conform *Anexei IV*, cuantilele

$$f_{m,n;\frac{\alpha}{2}} = \frac{1}{f_{n,m;1-\frac{\alpha}{2}}} = \frac{1}{f_{10,15;0.975}} = \frac{1}{3.06} = 0.33,$$
$$f_{m,n;1-\frac{\alpha}{2}} = f_{15,10;0.975} = 3.52, \quad f_{m,n;1-\alpha} = f_{15,10;0.95} = 2.85,$$
$$f_{m,n;\alpha} = \frac{1}{f_{n,m;1-\alpha}} = \frac{1}{f_{10,15;0.95}} = \frac{1}{2.54} = 0.39.$$

Când se consideră alternativa $H_1 : \sigma' \neq \sigma''$, avem testul F bilateral. Deoarece

$$f = 1.57 \in (0.33, 3.52) = \left(f_{m,n;\frac{\alpha}{2}}, f_{m,n;1-\frac{\alpha}{2}}\right),$$

ipoteza H_0 este admisă.

Pentru alternativa $H_1 : \sigma' > \sigma''$, avem testul F unilateral dreapta. Astfel că

$$f = 1.57 < 2.85 = f_{m,n;1-\alpha},$$

prin urmare ipoteza H_0 este admisă.

De asemenea, pentru alternativa $H_1 : \sigma' < \sigma''$, se obţine testul F unilateral stânga. Deoarece avem $f = 1.57 > 0.39 = f_{m,n;\alpha}$, ipoteza H_0 este admisă.

Programul 6.7.3. Programul Matlab, care urmează, efectuează calculele din exemplul precedent, după care afişează valoarea `f` a statisticii F, precum şi intervalele numerice `(f1,f2)`, pentru fiecare din cele trei alterantive.

```
x1=[1021,980,988,1017,1005,998,1014,985,995,...
    1004,1030,1015,995,1023,1008,1013];
x2=[1003,988,993,1013,1006,1002,1014,997,...
    1002,1010,975];
f=var(x1)/var(x2);
fprintf('   Bilateral\n')
f1=finv(0.025,15,10); f2=finv(0.975,15,10);
fprintf(' f=%7.3f, (f1,f2)=(%5.3f,%5.3f)\n',f,f1,f2)
fprintf('   Unilateral dreapta\n')
f1=finv(0,15,10); f2=finv(0.95,15,10);
fprintf(' f=%7.3f, (f1,f2)=(%5.3f,%5.3f)\n',f,f1,f2)
fprintf('   Unilateral stanga\n')
f1=finv(0.05,15,10); f2=finv(1,15,10);
fprintf(' f=%7.3f, (f1,f2)=(%5.3f,%5.3f)\n',f,f1,f2)
```

În urma executării programului se obţin rezultatele:

```
   Bilateral
 f=  1.567, (f1,f2)=(0.327,3.522)
   Unilateral dreapta
 f=  1.567, (f1,f2)=(0.000,2.845)
   Unilateral stanga
 f=  1.567, (f1,f2)=(0.393,  Inf)
```

6.8 Teste pentru compararea mediilor

Se consideră două populaţii independente $\mathcal{C}'$ şi $\mathcal{C}''$, cercetate din punct de vedere al aceleaşi caracteristici. Această caracteristică este X' pentru $\mathcal{C}'$ şi urmează legea normală $\mathcal{N}(m', \sigma')$ şi respectiv X'' pentru $\mathcal{C}''$ şi urmează legea normală $\mathcal{N}(m'', \sigma'')$.

Relativ la mediile teoretice ale celor două caracteristici independente, se face ipoteza nulă $H_0 : m' = m''$, cu una din alternativele:

$H_1 : m' \neq m''$, când obţinem un *test bilateral*,

$H_1 : m' > m''$, când obţinem un *test unilateral dreapta*,

$H_1 : m' < m''$, când obţinem un *test unilateral stânga*.

Ca şi în cazul testului F se consideră câte o selecţie repetată de volum n' şi respectiv n''. Păstrăm în continuare notaţiile de la testul F.

Distingem trei cazuri.

6.8.1 Dispersii cunoscute

Dispersiile σ'^2 şi σ''^2 sunt cunoscute. În acest caz se consideră statistica

$$Z = \frac{\left(\bar{X}' - \bar{X}''\right) - (m' - m'')}{\sqrt{\frac{\sigma'^2}{n'} + \frac{\sigma''^2}{n''^2}}}, \tag{6.8.1}$$

care urmează legea normală $\mathcal{N}(0,1)$.

Se aplică prin urmare testul Z pentru compararea celor două medii teoretice.

Pentru nivelul de semnificaţie $\alpha \in (0,1)$ dat, se obţin regiunile critice corespunzătoare celor trei alternative:

$$\mathcal{U} = \left\{ (u_1, \ldots, u_{n'}; v_1, \ldots, v_{n''}) \in \mathbb{R}^{n'+n''} \;\middle|\; \frac{|\,\bar{u} - \bar{v}\,|}{\sqrt{\frac{\sigma'^2}{n'} + \frac{\sigma''^2}{n''}}} \geqslant z_{1-\frac{\alpha}{2}} \right\},$$

$$\mathcal{U} = \left\{ (u_1, \ldots, u_{n'}; v_1, \ldots, v_{n''}) \in \mathbb{R}^{n'+n''} \;\middle|\; \frac{\bar{u} - \bar{v}}{\sqrt{\frac{\sigma'^2}{n'} + \frac{\sigma''^2}{n''}}} \geqslant z_{1-\alpha} \right\},$$

$$\mathcal{U} = \left\{ (u_1, \ldots, u_{n'}; v_1, \ldots, v_{n''}) \in \mathbb{R}^{n'+n''} \;\middle|\; \frac{\bar{u} - \bar{v}}{\sqrt{\frac{\sigma'^2}{n'} + \frac{\sigma''^2}{n''}}} \leqslant z_{\alpha} \right\}.$$

Etapele aplicării testului Z

1. Se dau: α; $x'_1, x'_2, \ldots, x'_{n'}$; $x''_1, x''_2, \ldots, x''_{n''}$; σ', σ'';
2. Se determină intervalul (z_1, z_2) astfel încât $\Phi(z_2) - \Phi(z_1) = 1 - \alpha$, unde $\Phi(x)$ este funcţia lui Laplace tabelată în *Anexa I*. Intervalul numeric (z_1, z_2) este respectiv pentru cele trei alternative considerate: $\left(-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}\right)$, $(-\infty, z_{1-\alpha})$, $(z_\alpha, +\infty)$;
3. Se calculează $z = \dfrac{\bar{x}' - \bar{x}''}{\sqrt{\frac{\sigma'^2}{n'} + \frac{\sigma''^2}{n''^2}}}$, unde $\bar{x}' = \dfrac{1}{n'} \displaystyle\sum_{k=1}^{n'} x'_k$, $\bar{x}'' = \dfrac{1}{n''} \displaystyle\sum_{k=1}^{n''} x''_k$;
4. Concluzia: dacă $z \in (z_1, z_2)$ ipoteza H_0 este admisă, în caz contrar este respinsă.

Exemplul 6.8.1. La o unitate de îmbuteliere a laptelui există două maşini care efectuează această operaţie în sticle de 1 litru. Pentru a cerceta reglajul de îmbuteliere la cele două maşini s–au efectuat două selecţii relative la sticlele îmbuteliate de cele două maşini şi s–au obţinut datele de selecţie:

x_k' (în ml)	990	995	1000	1005	1010
f_k'	7	9	11	8	5

,

x_k'' (în ml)	985	990	995	1000	1005	1010
f_k''	5	5	6	7	6	4

.

Folosind nivelul de semnificaţie $\alpha = 0.01$, vrem să verificăm dacă mediile de umplere a sticlelor de către cele două maşini sunt aceleaşi, în cazul în care abaterile standard sunt $\sigma' = 6$ ml şi $\sigma'' = 7.5$ ml.

Caracteristicile X' şi X'', ce reprezintă cantitatea de lapte (în ml) conţinută de o sticlă îmbuteliată de prima maşină, respectiv de a doua maşină, se consideră ca urmând legile de probabilitate normale $\mathcal{N}(m', 6)$ şi $\mathcal{N}(m'', 7.5)$.

Verificarea ipotezei nule $H_0 : m' = m''$, cu alternativa $H_1 : m' \neq m''$, se va face cu testul Z, deoarece sunt cunoscute abaterile standard.

Folosind nivelul de semnificaţie $\alpha = 0.01$, se determină din *Anexa I* valoarea

$$z_{1-\frac{\alpha}{2}} = z_{0.995}, \quad \text{astfel încât} \quad \Phi\left(z_{1-\frac{\alpha}{2}}\right) = \frac{1-\alpha}{2} = 0.495.$$

Anume, se obţine că $z_{0.995} = 2.58$, care ne dă intervalul $(-2.58\,;\,2.58)$ pentru statistica Z, dată prin formula (6.8.1).

Se calculează succesiv:

$$\bar{x}' = \frac{1}{n'}\sum_{k=1}^{n'} x_k' = \frac{1}{40}\left(7\cdot 990 + 9\cdot 995 + \cdots + 5\cdot 1010\right) = 999.375;$$

$$\bar{x}'' = \frac{1}{n''}\sum_{k=1}^{n''} x_k' = \frac{1}{33}\left(5\cdot 985 + 5\cdot 990 + \cdots + 4\cdot 1010\right) = 997.424;$$

$$z = \frac{\bar{x}' - \bar{x}''}{\sqrt{\frac{\sigma'^2}{n'} + \frac{\sigma''^2}{n''}}} = \frac{999.375 - 997.424}{\sqrt{\frac{36}{40} + \frac{56.25}{33}}} = \frac{1.951}{\sqrt{2.6046}} = 1.209\cdot$$

Deoarece $z = 1.209 \in (-2.58\,;\,2.58)$, rezultă că mediile de umplere a sticlelor nu diferă semnificativ pentru cele două maşini.

Programul 6.8.2. Programul Matlab, care urmează, efectuează calculele din exemplul precedent.

```
x1=[990*ones(1,7),995*ones(1,9),1000*ones(1,11),...
    1005*ones(1,8),1010*ones(1,5)];
x2=[985*ones(1,5),990*ones(1,5),995*ones(1,6),...
    1000*ones(1,7),1005*ones(1,6),1010*ones(1,4)];
ma1=mean(x1); ma2=mean(x2);
z=(ma1-ma2)/sqrt(6^2/40+7.5^2/33);
z2=norminv(0.995); z1=-z2;
fprintf(' z=%7.3f\n z1=%6.3f\n z2=%6.3f',z,z1,z2)
```

Rezultatele obţinute în urma executării programului sunt:

```
z=  1.209
z1=-2.576
z2= 2.576
```

6.8.2 Dispersii egale necunoscute

Dispersiile σ'^2 *şi* σ''^2 *sunt necunoscute şi* $\sigma'^2 = \sigma''^2 = \sigma^2$. Se consideră statistica

$$T = \frac{\left(\bar{X}' - \bar{X}''\right) - (m' - m'')}{\sqrt{(n'-1)\,\bar{\sigma}'^2 + (n''-1)\,\bar{\sigma}''^2}} \sqrt{\frac{n' + n'' - 2}{\frac{1}{n'} + \frac{1}{n''}}}, \tag{6.8.2}$$

care urmează, conform Proprietăţii 4.2.28, legea Student cu $n = n' + n'' - 2$ grade de libertate.

În acest caz se va aplica testul T. Pentru nivelul de semnificaţie $\alpha \in (0,1)$ dat, se obţin regiunile critice corespunzătoare celor trei alternative:

$$\mathcal{U} = \left\{ (u_1, \ldots, u_{n'}; v_1, \ldots, v_{n''}) \in \mathbb{R}^{n'+n''} \;\middle|\; K \,|\, \bar{u} - \bar{v} \,| \geqslant t_{n,1-\frac{\alpha}{2}} \right\},$$

$$\mathcal{U} = \left\{ (u_1, \ldots, u_{n'}; v_1, \ldots, v_{n''}) \in \mathbb{R}^{n'+n''} \;\middle|\; K\,(\bar{u} - \bar{v}) \geqslant t_{n,1-\alpha} \right\},$$

$$\mathcal{U} = \left\{ (u_1, \ldots, u_{n'}; v_1, \ldots, v_{n''}) \in \mathbb{R}^{n'+n''} \;\middle|\; K\,(\bar{u} - \bar{v}) \leqslant t_{n,\alpha} \right\},$$

unde

$$\bar{u} = \frac{1}{n'} \sum_{k=1}^{n'} u_k, \quad \bar{v} = \frac{1}{n''} \sum_{k=1}^{n''} v_k, \quad K = \frac{1}{\sqrt{(n'-1)\,\bar{\sigma}_u^2 + (n''-1)\,\bar{\sigma}_v^2}} \sqrt{\frac{n}{\frac{1}{n'} + \frac{1}{n''}}},$$

$$\bar{\sigma}_u^2 = \frac{1}{n'-1} \sum_{k=1}^{n'} (u_k - \bar{u})^2, \quad \bar{\sigma}_v^2 = \frac{1}{n''-1} \sum_{k=1}^{n''} (v_k - \bar{v})^2.$$

Etapele aplicării testului T

1. Se dau: α; $x'_1, x'_2, \ldots, x'_{n'}$; $x''_1, x''_2, \ldots, x''_{n''}$;
2. Se determină intervalul (t_1, t_2) astfel încât $F_n(t_2) - F_n(t_1) = 1 - \alpha$, unde $F_n(x)$ este funcţia de repartiţie pentru legea Student cu $n = n' + n'' - 2$ grade de libertate, iar intervalul numeric (t_1, t_2) este respectiv, pentru cele trei alternative considerate, $\left(-t_{n,1-\frac{\alpha}{2}}; t_{n,1-\frac{\alpha}{2}}\right)$, $(-\infty; t_{n,1-\alpha})$, $(t_{n,\alpha}; +\infty)$;

3. Se calculează $t = \dfrac{\bar{x}\,' - \bar{x}\,''}{\sqrt{(n'-1)\,\bar{\sigma}\,'^2 + (n''-1)\,\bar{\sigma}\,''^2}} \sqrt{\dfrac{n'+n''-2}{\frac{1}{n'}+\frac{1}{n''}}}$, unde

$$\bar{\sigma}\,'^2 = \frac{1}{n'-1}\sum_{k=1}^{n'}\left(x'_k - \bar{x}\,'\right)^2, \qquad \bar{x}\,' = \frac{1}{n'}\sum_{k=1}^{n'} x'_k,$$

$$\bar{\sigma}\,''^2 = \frac{1}{n''-1}\sum_{k=1}^{n''}\left(x''_k - \bar{x}\,''\right)^2, \quad \bar{x}\,'' = \frac{1}{n''}\sum_{k=1}^{n''} x''_k;$$

4. Concluzia: dacă $t \in (t_1, t_2)$ ipoteza H_0 este admisă, în caz contrar este respinsă.

Exemplul 6.8.3. Se cercetează două loturi de ulei pentru automobile, din punct de vedere al vâscozităţii, obţinându–se datele de selecţie

x'_k	10.27	10.28	10.29	10.30	10.32
f'_k	3	2	1	1	1

pentru primul lot, respectiv

x''_k	10.26	10.27	10.29	10.30	10.31
f''_k	2	1	1	1	3

pentru al doilea lot.

Analizele facându–se cu acelaşi aparat, se consideră că abaterile standard sunt aceleaşi. Considerând nivelul de semnificaţie $\alpha = 0.05$, să se verifice dacă mediile de vâscozitate pentru cele două loturi nu diferă semnificativ.

Caracteristicile X' şi X'', ce reprezintă vâscozităţile pentru cele două loturi de ulei, se consideră că urmează fiecare legea normală, respectiv $\mathcal{N}(m', \sigma)$ şi $\mathcal{N}(m'', \sigma)$, cu abaterea standard $\sigma > 0$ necunoscută.

Verificarea ipotezei nule $H_0 : m' = m''$, cu alternativa $H_1 : m' \neq m''$, se va face cu testul T, deoarece abaterea standard σ este necunoscută.

Folosind nivelul de seminficaţie $\alpha = 0.05$, se determină, din *Anexa II*, valoarea $t_{n,1-\frac{\alpha}{2}}$, astfel încât $F_n\left(t_{n,1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$, unde numărul gradelor de libertate este $n = n' + n'' - 2 = 8 + 8 - 2 = 14$. Adică, se determină $t_{14;0.975}$ astfel încât $F_{14}\left(t_{14;0.975}\right) = 0.975$, obţinându–se $t_{14;0.975} = 2.145$. În acest mod, s–a obţinut intervalul $(-2.145\,;\, 2.145)$ pentru statistica dată prin formula (6.8.2), care urmează legea Student cu $n = n' + n'' - 2$ grade de libertate.

Se calculează pe rând

$$\bar{x}\,' = \frac{1}{n'}\sum_{k=1}^{n'} x'_k = \frac{1}{8}\left(3 \cdot 10.27 + 2 \cdot 10.28 + \cdots + 1 \cdot 10.32\right) = 10.285;$$

$$\bar{x}'' = \frac{1}{n''}\sum_{k=1}^{n''} x_k'' = \frac{1}{8}\left(2\cdot 10.26 + 1\cdot 10.27 + \cdots + 3\cdot 10.31\right) = 10.289;$$

$$\bar{\sigma}'^2 = \frac{1}{n'-1}\sum_{k=1}^{n'}\left(x_k' - \bar{x}'\right)^2 = 3.14\cdot 10^{-4};$$

$$\bar{\sigma}''^2 = \frac{1}{n''-1}\sum_{k=1}^{n''}\left(x_k'' - \bar{x}''\right)^2 = 4.98\cdot 10^{-4};$$

$$t = \frac{\bar{x}' - \bar{x}''}{\sqrt{(n'-1)\ \bar{\sigma}'^2 + (n''-1)\ \bar{\sigma}''^2}}\cdot\sqrt{\frac{n'+n''-2}{\frac{1}{n'}+\frac{1}{n''}}}$$

$$= \frac{10.285 - 10.289}{\sqrt{(22.001+34.881)\cdot 10^{-4}}}\sqrt{\frac{14}{\frac{1}{8}+\frac{1}{8}}} = \frac{-4\cdot 10^{-1}}{\sqrt{56.882}}\cdot\sqrt{56} = -0.37.$$

Deoarece $t = -0.37 \in (-2.145\,;\ 2.145)$, rezultă că vâscozităţile medii ale celor două loturi de ulei nu diferă semnificativ.

Programul 6.8.4. Programul Matlab următor efectuează calculele de mai sus, după care afişează valoarea `t` a statisticii T, împreună cu capetele intervalului `(t1,t2)`.

```
x1=[10.27*ones(1,3),10.28*ones(1,2),10.29,10.3,10.32];
x2=[10.26,10.26,10.27,10.29,10.3,10.31*ones(1,3)];
ma1=mean(x1); ma2=mean(x2);
v1=var(x1); v2=var(x2);
t=(ma1-ma2)/sqrt(7*v1+7*v2)*sqrt(14/(1/8+1/8));
t2=tinv(0.975,14); t1=-t2;
fprintf(' t=%7.3f\n t1=%6.3f\n t2=%6.3f',t,t1,t2)
```

În urma executării programului, se obţin rezultatele:

```
t= -0.372
t1=-2.145
t2= 2.145
```

6.8.3 Funcţia `ttest2`

Sistemul Matlab, prin *Statistics toolbox*, dispune de funcţia `ttest2`, cu aplicabilitate la testul T pentru compararea a două medii, când dispersiile sunt egale, dar necunoscute. Apelarea acestei funcţii se face prin:

```
h=ttest2(x,y)
h=ttest2(x,y,alpha)
h=ttest2(x,y,alpha,tail)
[h,c,ci]=ttest(x,y,alpha,tail)
```

În urma executării acestor instrucţiuni, se efectuează testul T asupra datelor conţinute în vectorii `x` şi `y`, folosind nivelul de semnificaţie `alpha`, care are valoarea implicită `alpha=0.05`.

Parametrul `tail` specifică una din cele trei alternative, care conduc la testul bilateral (`tail=0`, implicit), unilateral dreapta (`tail=1`) şi unilateral stânga (`tail=-1`). Dacă `h=1`, atunci ipoteza nulă va fi respinsă, respectiv dacă `h=0`, ipoteza nu poate fi respinsă.

Ultima formă de apel permite, de asemenea, obţinerea valorii critice `c`, precum şi a intervalului de încredere pentru diferenţa mediilor teoretice, corespunzător probabilităţii de încredere `1-alpha`, obţinut în vectorul cu două componente `ci`. Valoarea critică `c` are aceeaşi semnificaţie cu cea precizată la Observaţia 6.4.2 şi se calculează cu formulele prezentate acolo.

Programul 6.8.5. Programul Matlab, ce urmează, aplică testul T, pentru datele din Exemplul 6.8.3. Mai mult, se consideră şi testele unilateral dreapta şi unilateral stânga, precum şi obţinerea intervalelor de încredere, când `alpha=0.05`.

```
x=[10.27*ones(1,3),10.28*ones(1,2),10.29,10.3,10.32];
y=[10.26*ones(1,2),10.27,10.29,10.3,10.31*ones(1,3)];
fprintf(' h    c           ci     \n')
fprintf('_____________________________\n')
for i=-1:1
  [h,c,ci]=ttest2(x,y,0.05,i);
  fprintf(' %d  %5.4f   (%4.2f,%4.2f) \n',h,c,ci)
end
```

În urma executării programului, se obţin rezultatele:

```
 h    c           ci
_____________________________
 0  0.3577   (-Inf,0.01)
 0  0.7154   (-0.03,0.02)
 0  0.6423   (-0.02, Inf)
```

Se observă că pentru `tail=-1` şi `tail=1`, se construiesc intervale nemărginite, respectiv la stânga şi la dreapta.

6.8.4 Dispersii diferite necunoscute

Dispersiile σ'^2 şi σ''^2 sunt necunoscute şi diferite. Se va considera statistica

$$T = \frac{\left(\bar{X}' - \bar{X}''\right) - (m' - m'')}{\sqrt{\frac{\bar{\sigma}'^2}{n'} + \frac{\bar{\sigma}''^2}{n''^2}}}, \tag{6.8.3}$$

care urmează legea Student cu n grade de libertate. Numărul n al gradelor de libertate se calculează cu formula

$$\frac{1}{n}=\frac{c^2}{n'-1}+\frac{(1-c)^2}{n''-1},\quad \text{unde}\quad c=\frac{\bar{\sigma}'^2}{n'}\Bigg/\left(\frac{\bar{\sigma}'^2}{n'}+\frac{\bar{\sigma}''^2}{n''}\right). \tag{6.8.4}$$

S–a ajuns la testul T, care pentru nivelul de semnificaţie $\alpha\in(0,1)$ dat, conduce la regiunile critice corespunzătoare alternativelor considerate, respectiv:

$$\mathcal{U}=\left\{(u_1,\dots,u_{n'};v_1,\dots,v_{n''})\in\mathbb{R}^{n'+n''}\;\middle|\;\frac{|\,\bar{u}-\bar{v}\,|}{\sqrt{\frac{\bar{\sigma}_u^2}{n'}+\frac{\bar{\sigma}_v^2}{n''}}}\geqslant t_{n,1-\frac{\alpha}{2}}\right\},$$

$$\mathcal{U}=\left\{(u_1,\dots,u_{n'};v_1,\dots,v_{n''})\in\mathbb{R}^{n'+n''}\;\middle|\;\frac{\bar{u}-\bar{v}}{\sqrt{\frac{\bar{\sigma}_u^2}{n'}+\frac{\bar{\sigma}_v^2}{n''}}}\geqslant t_{n,1-\alpha}\right\},$$

$$\mathcal{U}=\left\{(u_1,\dots,u_{n'};v_1,\dots,v_{n''})\in\mathbb{R}^{n'+n''}\;\middle|\;\frac{\bar{u}-\bar{v}}{\sqrt{\frac{\bar{\sigma}_u'^2}{n'}+\frac{\bar{\sigma}_v''^2}{n''}}}\leqslant t_{n,\alpha}\right\}.$$

Etapele aplicării testului T

1. Se dau: $\alpha;\ x'_1,x'_2,\dots,x'_{n'};\ x''_1,x''_2,\dots,x''_{n''}$;
2. Se determină intervalul (t_1,t_2) astfel încât $F_n(t_2)-F_n(t_1)=1-\alpha$, unde $F_n(x)$ este funcţia de repartiţie pentru legea Student cu n grade de libertate calculat cu formula (6.8.4), iar intervalul numeric (t_1,t_2) este respectiv, pentru cele trei alternative considerate, $\left(-t_{n,1-\frac{\alpha}{2}};t_{n,1-\frac{\alpha}{2}}\right)$, $(-\infty;t_{n,1-\alpha})$, $(t_{n,\alpha};+\infty)$;
3. Se calculează $t=\dfrac{\bar{x}'-\bar{x}''}{\sqrt{\frac{\bar{\sigma}'^2}{n'}+\frac{\bar{\sigma}''^2}{n''^2}}}$, unde

$$\bar{\sigma}'^2=\frac{1}{n'-1}\sum_{k=1}^{n'}\left(x'_k-\bar{x}'\right)^2,\qquad \bar{x}'=\frac{1}{n'}\sum_{k=1}^{n'}x'_k,$$

$$\bar{\sigma}''^2=\frac{1}{n''-1}\sum_{k=1}^{n''}\left(x''_k-\bar{x}''\right)^2,\quad \bar{x}''=\frac{1}{n''}\sum_{k=1}^{n''}x''_k;$$

4. Concluzia: dacă $t\in(t_1,t_2)$ ipoteza H_0 este admisă, în caz contrar este respinsă.

Exemplul 6.8.6. Se cercetează capacitatea fiolelor farmaceutice de 100 ml, care provin de la două fabrici. În acest scop, se consideră câte o selecţie pentru două loturi de fiole provenite respectiv de la cele două fabrici. Selecţiile obţinute au distribuţiile empirice de selecţie

$$X' \begin{pmatrix} 100 & 101 & 102 & 103 & 104 & 105 & 106 & 107 & 108 & 109 \\ 1 & 1 & 2 & 3 & 4 & 5 & 4 & 1 & 3 & 1 \end{pmatrix},$$

respectiv, pentru X'': 110, 101, 112, 120, 117, 105, 109, 111, 118, 113, 106, 108, 115, 113, 112, 100, 116, 112, 114, 112.

Folosind nivelul de semnificaţie $\alpha = 0.02$, vom compara dispersiile celor două caracteristici. Apoi, pe baza acestui rezultat, vom compara mediile celor două caracteristici, utilizând acelaşi nivel de semnificaţie $\alpha = 0.02$.

Vom considera că cele două caractersitici X' şi X'' sunt repartizate normal, respectiv $\mathcal{N}(m'; \sigma')$ şi $\mathcal{N}(m''; \sigma'')$. Se poate aplica testul F, pentru compararea dispersiilor teoretice ${\sigma'}^2$ şi ${\sigma''}^2$.

Calculăm pe rând:

$$\bar{x}' = \frac{1}{n'} \sum_{k=1}^{n'} x'_k = \frac{1}{25} (1 \cdot 100 + \cdots + 1 \cdot 109) = 104.76; \ \bar{x}'' = \frac{1}{n''} \sum_{k=1}^{n''} x''_k = 111.2;$$

$$\bar{\sigma}'^2 = \frac{1}{n' - 1} \sum_{k=1}^{n'} \left(x'_k - \bar{x}'\right)^2 = 5.19; \quad \bar{\sigma}''^2 = \frac{1}{n'' - 1} \sum_{k=1}^{n''} \left(x''_k - \bar{x}''\right)^2 = 27.537.$$

Deoarece $\bar{\sigma}'^2 < \bar{\sigma}''^2$, se consideră statistica $F = \frac{\bar{\sigma}''^2}{{\sigma''}^2} \Big/ \frac{\bar{\sigma}'^2}{{\sigma'}^2}$, care urmează legea Fisher–Snedecor cu $(m, n) = (n'' - 1, n' - 1) = (19\,,\, 24)$ grade de libertate.

Dacă se consideră ipoteza nulă $H_0 : {\sigma'}^2 = {\sigma''}^2$, cu alternativa $H_1 : {\sigma'}^2 \neq {\sigma''}^2$, avem că $f = \frac{\bar{\sigma}''^2}{\bar{\sigma}'^2} = \frac{27.537}{5.19} = 5.31$.

Pe de altă parte, pentru $\alpha = 0.02$, avem din *Anexa IV* că

$$f_{m,n;1-\frac{\alpha}{2}} = f_{19.24;0.99} = 2.76, \quad f_{m,n;\frac{\alpha}{2}} = f_{19.24;0.01} = \frac{1}{f_{24.19;0.99}} = \frac{1}{2.92} = 0.34.$$

În acest fel, am obţinut intervalul $(0.34\,;\, 2.76)$, pentru statistica F.

Deoarece $f = 5.31 \notin (0.34\,;\, 2.76)$, respingem ipoteza că ${\sigma'}^2 = {\sigma''}^2$.

Având în vedere că dispersiile teoretice ${\sigma'}^2$ şi ${\sigma''}^2$ sunt necunoscute, iar conform rezultatului precedent diferă în mod semnificativ, folosim testul T pentru compararea mediilor m' şi m''. Statistica T, ce se consideră, în acest caz, este cea dată prin formula (6.8.3), care urmează legea Student cu n grade de libertate, unde n se calculează din relaţia (6.8.4).

Astfel, pentru determinarea lui n, avem succesiv

$$c=\frac{\frac{5.19}{25}}{\frac{5.19}{25}+\frac{27.537}{20}}=0.131 \quad \text{şi} \quad \frac{1}{n}=\frac{0.131^2}{24}+\frac{(1-0.131)^2}{19}=0.0404604,$$

de unde $n=25$. Folosind *Anexa II*, se obţine că $t_{25;0.99}=2.485$, prin urmare intervalul pentru statistica T este $(-2.485\,;\,2.485)$.

Pe de altă parte, avem că

$$t=\frac{\bar{x}'-\bar{x}''}{\sqrt{\frac{\bar{\sigma}'^2}{n'}+\frac{\bar{\sigma}''^2}{n''}}}=\frac{104.76-111.2}{\sqrt{0.2076+1.3768}}=-\frac{6.44}{1.26}=-5.11.$$

Deoarece $t=-5.11\notin(-2.485\,;\,2.485)$, respingem ipoteza că mediile teoretice pentru fiolele produse de cele două fabrici nu diferă semnificativ.

Programul 6.8.7. Calculele din exemplul de mai sus, se pot efectua cu următorul program Matlab:

```
x1=[100,101,102*ones(1,2),103*ones(1,3),...
   104*ones(1,4),105*ones(1,5),106*ones(1,4),...
   107,108*ones(1,3),109];
x2=[110,101,112,120,117,105,109,111,118,113,...
    106,108,115,113,112,100,116,112,114,112];
f=var(x2)/var(x1);
c1=finv(0.01,19,24); c2=finv(0.99,19,24);
fprintf(' f=%6.2f,  c1=%5.2f,  c2=%5.2f\n',f,c1,c2)
ma1=mean(x1); ma2=mean(x2);
v1=var(x1); v2=var(x2); c=(v1/25)/(v1/25+v2/20);
n=c^2/24+(1-c)^2/19; n=ceil(1/n);
t=(ma1-ma2)/sqrt(v1/25+v2/20);
t2=tinv(0.99,n); t1=-t2;
fprintf(' t=%7.3f, t1=%6.3f, t2=%6.3f',t,t1,t2)
```

Rezultatele obţinute, în urma executării programului, sunt:

```
 f=  5.31,  c1= 0.34,  c2= 2.76
 t= -5.116, t1=-2.485, t2= 2.485
```

Observaţia 6.8.8. Dacă se notează prin $\theta=m'-m''$, atunci ipoteza compusă nulă devine $H_0:\theta=0$, iar ipotezele alternative se rescriu după cum urmează: $H_1:\theta\neq 0$, $H_1:\theta>0$ şi respectiv $H_1:\theta<0$, iar statisticile care au fost date mai înainte se pot rescrie cu acest nou parametru necunoscut θ.

Observaţia 6.8.9. Când selecţiile sunt de volum mare, $n',n''>50$, pentru legi de probabilitate oarecari, respectiv $n',n''>20$, pentru legi de probabilitate normale, statistica

$$Z=\frac{\left(\bar{X}'-\bar{X}''\right)-(m'-m'')}{\sqrt{\frac{\bar{\sigma}'^2}{n'}+\frac{\bar{\sigma}''^2}{n''^2}}},$$

poate fi considerată ca urmând legea normală $\mathcal{N}(0,1)$. Aşadar în acest caz se poate aplica testul Z.

6.8.5 Observaţii perechi

Dacă cele două selecţii sunt de tip pereche, adică variabilele de selecţie sunt perechile (X'_k, X''_k), $k = \overline{1,n}$, atunci se pot considera diferenţele $D_k = X'_k - X''_k$, pentru care $E(D_k) = m = m' - m''$, oricare a fi $k = \overline{1,n}$.

Problema poate fi reformulată.

Se consideră o populaţie $\mathcal{C}$, pentru care se cercetează caracteristica D, care urmează legea normală $\mathcal{N}(m,\sigma)$, cu $\sigma > 0$ necunoscut.

Vrem să verificăm ipoteza nulă $H_0 : m = 0$, cu una din alternativele

$H_1 : m \neq 0$, când obţinem *testul T bilateral*,

$H_1 : m > 0$, când obţinem *testul T unilateral dreapta*,

$H_1 : m < 0$, când obţinem *testul T unilateral stânga*.

Conform Proprietăţii 4.2.26, statistica

$$T = \frac{\bar{D} - m}{\frac{\bar{\sigma}_D}{\sqrt{n}}},$$

unde

$$\bar{D} = \frac{1}{n}\sum_{k=1}^{n} D_k = \sum_{k=1}^{n}\left(X'_k - X''_k\right), \quad \bar{\sigma}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(D_k - \bar{D}\right)^2,$$

urmează legea Student cu $n-1$ grade de libertate.

Prin urmare, pentru nivelul de semnificaţie $\alpha \in (0,1)$ dat, se poate determina intervalul numeric (t_1, t_2) astfel încât

$$P\big(T \in (t_1,t_2) \;\big|\; H_0\big) = F_{n-1}(t_2) - F_{n-1}(t_1) = 1 - \alpha,$$

unde

$$F_m(t) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\sqrt{m\pi}\,\Gamma\left(\frac{m}{2}\right)} \int_{-\infty}^{t} \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}} dx, \quad t \in \mathbb{R},$$

este funcţia de repartiţie pentru legea Student cu m grade de libertate (tabelată în *Anexa II*, pentru anumite valori).

Intervalul numeric (t_1, t_2) pentru statistica T nu este determinat în mod unic din condiţia de mai sus. În funcţie de alternativa H_1 aleasă, se consideră suplimentar:

$t_1 = -t_2,\ t_2 = t_{n-1,1-\frac{\alpha}{2}}$, dacă $H_1 : m \neq 0$,

$t_1 = -\infty,\ t_2 = t_{n-1,1-\alpha}$, dacă $H_1 : m > 0$,

$t_1 = t_{n-1,\alpha},\ t_2 = +\infty$, dacă $H_1 : m < 0$,

unde $t_{n-1,\gamma}$ este cuantila de ordin γ a legii Student cu $n-1$ grade de libertate.

Etapele aplicării testului T

1. Se dau: $\alpha;\ x'_1, x'_2, \ldots, x'_n;\ x''_1, x''_2, \ldots, x''_n$;
2. Se calculează intervalul (t_1, t_2) astfel încât $F_{n-1}(t_1) - F_{n-1}(t_2) = 1 - \alpha$, (după cum s-a prezentat înainte);
3. Se calculează
$$t = \frac{\bar{d}}{\frac{\bar{\sigma}_D}{\sqrt{n}}}, \quad \bar{d} = \frac{1}{n}\sum_{k=1}^{n} d_k = \frac{1}{n}\sum_{k=1}^{n}\left(x'_k - x''_k\right),\ \bar{\sigma}_D^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(d_k - \bar{d}\right)^2;$$
4. Concluzia: dacă $t \in (t_1, t_2)$ ipoteza H_0 este admisă, în caz contrar ipoteza este respinsă.

Programul 6.8.10. Vom scrie un program, care generează `n` vectori aleatori, care urmează legea normală bidimensională. Folosind testul T pentru perechi, vom verifica ipoteza nulă, că mediile celor două variabile sunt egale, precum şi intervalul de încredere pentru diferenţa mediilor, respectiv când se aplică testele bilateral, unilateral dreapta şi unilateral stânga. Nivelul de semnificaţie folosit este `alpha=0.05`.

```
mu(1)=input('m1='); mu(2)=input('m2=');
v(1,1)=input('sigma1^2=');
v(2,2)=input('sigma2^2=');
v(1,2)=input('Cov(X,Y)='); v(2,1) =v(1,2);
if det(v) <= 0
   error('Matricea v nu e pozitiv definita!')
end
n=input('n='); Z=mvnrnd(mu,v,n);
d=diff(Z,1,2);
fprintf(' h    c           ci     \n')
fprintf('_____________________________\n')
for i=-1:1
  [h,c,ci]=ttest(d,0,0.05,i);
  fprintf(' %d  %5.4f   (%4.2f,%4.2f) \n',h,c,ci)
end
```

Rezultatele obţinute, pentru `mu1=mu2=10`, `v` $= \begin{pmatrix} 2 & 2 \\ 2 & 3 \end{pmatrix}$ şi `n=30`, sunt

```
h    c          ci
_____________________________
0  0.6715   (-Inf,0.35)
0  0.6570   (-0.26,0.41)
0  0.3285   (-0.20, Inf)
```

Deoarece `h=0`, pentru fiecare alternativă în parte, testul T nu respinge ipoteza că cele două medii sunt egale. Acest lucru se vede şi din faptul că valoarea critică satisface inegalitatea `c>alpha=0.05`.

Programul mai calculează intervalul de încredere pentru diferenţa celor două medii. Se observă că acesta diferă, în funcţie de parametrul `tail=-1,0,1`. De asemenea, se observă că am considerat nivelul de semnificaţie `alpha=0.05`, dar acesta poate fi uşor modificat în apelul funcţiei `ttest`.

6.9 Testul χ^2 pentru concordanţă

Se consideră colectivitatea $\mathcal{C}$ cercetată din punct de vedere al caracteristicii X cantitativă sau calitativă. Fie k numărul claselor caracteristicii X şi E_i evenimentul ca un individ luat la întâmplare din colectivitatea $\mathcal{C}$ să aparţină clasei cu numărul de ordine i.

Notând $p_i = P(E_i)$, $i = \overline{1,k}$, atunci $\sum_{i=1}^{k} p_i = 1$.

Relativ la caracteristica X facem ipoteza nulă $H_0 : p_i = p_i^{(0)}$, $i = \overline{1,k}$, cu ipoteza alternativă H_1: există i_0 astfel încât $p_{i_0} \neq p_{i_0}^{(0)}$.

Pentru a verifica această ipoteză se consideră o selecţie repetată de volum n. Fie datele de selecţie $x_1, x_2, \ldots, x_n$. Folosind aceste date de selecţie se obţin frecvenţele absolute ale claselor caracteristicii X. Vom nota prin n_i frecvenţa absolută a clasei cu numărul de ordine i. Altfel spus, n_i numără de câte ori a apărut evenimentul E_i în selecţia considerată.

Corespunzător frecvenţelor absolute n_i, $i = \overline{1,k}$, avem variabilele aleatoare (de selecţie) N_i, $i = \overline{1,k}$, ce iau aceste valori.

Aşadar, pornind de la caracteristica X, care poate fi şi calitativă, s-a ajuns la *variabilele de selecţie* N_i, $i = \overline{1,k}$, care sunt componentele unui vector aleator k-dimensional $N = (N_1, N_2, \ldots, N_k)$, ce urmează legea multinomială. Anume, avem:

$$P(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k},$$

unde $n = n_1 + n_2 + \cdots + n_k$, $n_i = \overline{0,n}$, $i = \overline{1,k}$, $p_1 + p_2 + \cdots + p_k = 1$.

Dacă privim ipoteza nulă H_0, constatăm că aceasta se referă la *parametrii unei legi multinomiale*.

Proprietatea 6.9.1. *Cu notaţiile dinainte avem că statistica*

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i},$$

urmează legea χ^2 cu $k-1$ grade de libertate, când $n \to \infty$.

Demonstraţie. Se porneşte de la formula lui Stirling $n! \cong \sqrt{2\pi n}\, n^n \mathrm{e}^{-n}$. Astfel, avem că

$$P\left(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k\right) \cong \frac{\sqrt{2\pi n}\, n^n \mathrm{e}^{-n}}{\prod_{i=1}^{k} \left(\sqrt{2\pi n_i}\, n_i^{n_i} \mathrm{e}^{-n_i}\right)} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k},$$

sau

$$P\left(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k\right) \cong \prod_{i=1}^{k} \left(\frac{np_i}{n_i}\right)^{n_i + \frac{1}{2}} \cdot K,$$

unde K este o constantă pozitivă. Logaritmând această relaţie obţinem

$$\ln P\left(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k\right) \cong \ln K + \sum_{i=1}^{k} \left(n_i + \frac{1}{2}\right) \ln \frac{np_i}{n_i},$$

şi notând

$$x_i = \frac{n_i - np_i}{\sqrt{np_i}}, \quad \text{adică} \quad \frac{n_i}{np_i} = 1 + \frac{x_i}{\sqrt{np_i}},$$

rezultă că

$$\ln P\left(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k\right) \cong \ln K - \sum_{i=1}^{k} \left(n_i + \frac{1}{2}\right) \ln \left(1 + \frac{x_i}{\sqrt{np_i}}\right).$$

Păstrând primii doi termeni din dezvoltarea în serie a logaritmului avem

$$\ln \left(1 + \frac{x_i}{\sqrt{np_i}}\right) \cong \frac{x_i}{\sqrt{np_i}} - \frac{x_i^2}{2np_i},$$

deci

$$\begin{aligned}
\ln P\left(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k\right) &\cong \ln K - \sum_{i=1}^{k} \left(n_i + \frac{1}{2}\right) \left(\frac{x_i}{\sqrt{np_i}} - \frac{x_i^2}{2np_i}\right) \\
&= \ln K - \sum_{i=1}^{k} \left(np_i + x_i\sqrt{np_i} + \frac{1}{2}\right) \left(\frac{x_i}{\sqrt{np_i}} - \frac{x_i^2}{2np_i}\right) \\
&\cong \ln K - \sum_{i=1}^{k} \left(x_i\sqrt{np_i} + \frac{x_i^2}{2}\right).
\end{aligned}$$

Dar avem că

$$\sum_{i=1}^{k} x_i \sqrt{np_i} = \sum_{i=1}^{k} (n_i - np_i) = n - n = 0,$$

deci

$$\ln P\left(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k\right) \cong \ln K - \frac{1}{2} \sum_{i=1}^{k} x_i^2$$

sau

$$P\left(N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k\right) \cong K \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^{k} x_i^2}.$$

Dacă se notează $X_i = \frac{N_i - np_i}{\sqrt{np_i}}$, obţinem

$$P\left(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k\right) \cong K \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^{k} x_i^2},$$

adică vectorul aleator

$$(X_1, X_2, \ldots, X_k) = \left(\frac{N_1 - np_1}{\sqrt{np_1}}, \frac{N_2 - np_2}{\sqrt{np_2}}, \ldots, \frac{N_k - np_k}{\sqrt{np_k}}\right),$$

pentru $n \to \infty$, urmează o lege normală k dimensională degenerată, deoarece fiecare componentă X_i se poate exprima ca şi combinaţie liniară a celorlalte componente.

Din teoria probabilităţilor se cunoaşte că suma pătratelor componentelor unui vector aleator ce urmează legea normală şi între care există o legătură liniară este o variabilă aleatoare ce urmează legea χ^2 cu numărul gradelor de libertate dat de numărul componentelor vectorului mai puţin unu. Ceea ce trebuie demonstrat. □

Pe baza proprietăţii precedente, pentru nivelul de semnificaţie $\alpha \in (0, 1)$ dat, se poate determina cuantila $\chi^2_{k-1,1-\alpha}$ astfel încât

$$F_{k-1}\left(\chi^2_{k-1,1-\alpha}\right) = P\left(\chi^2 \leqslant \chi^2_{k-1,1-\alpha} \mid H_0\right) = 1 - \alpha,$$

unde F_{k-1} notează funcţia de repartiţie a legii χ^2 cu $k-1$ grade de libertate.

Regiunea critică $\mathcal{U}$ se poate scrie şi în acest caz, anume

$$\mathcal{U} = \left\{(u_1, u_2, \ldots, u_k) \in \mathbb{R}^k \;\middle|\; \sum_{i=1}^{k} \frac{\left(u_i - np_i^{(0)}\right)^2}{np_i^{(0)}} \geqslant \chi^2_{k-1,1-\alpha}\,,\ \sum_{i=1}^{k} u_i = n\right\}.$$

Astfel, pentru n mare, are loc relaţia $P\big((N_1, N_2, \ldots, N_k) \in \mathcal{U} \,|\, H_0\big) = \alpha$, deci ipoteza H_0 va fi admisă dacă $(n_1, n_2, \ldots, n_k) \notin \mathcal{U}$, respectiv va fi respinsă în caz contrar.

Etapele aplicării testului χ^2

1. Se dau: α; $x_1, x_2, \ldots, x_n$; $p_i = p_i^{(0)}$, $i = \overline{1,k}$. (şirul datelor de selecţie $x_1, x_2, \ldots, x_n$ este un şir având ca elemente evenimentele E_i, $i = \overline{1,k}$, din care se obţin frecvenţele absolute $n_1, n_2, \ldots, n_k$);
2. Se calculează $\chi^2_{k-1,1-\alpha}$ astfel încât $F_{k-1}\left(\chi^2_{k-1,1-\alpha}\right) = 1-\alpha$;
3. Se calculează $x^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}}$;
4. Concluzia: dacă $x^2 < \chi^2_{k-1,1-\alpha}$ ipoteza H_0 este admisă, în caz contrar ipoteza este respinsă.

Observaţia 6.9.2. La utilizarea testului χ^2 privind concordanţa trebuie ca să fie îndeplinite condiţiile $np_i > 4$, când $k > 4$, iar dacă numărul claselor $k \leqslant 4$, atunci np_i să fie mult mai mare decât 4. Când aceste condiţii nu sunt îndeplinite, se efectuează o regrupare a datelor de selecţie.

Exemplul 6.9.3. S-a aruncat un zar de 60 de ori şi s-au obţinut următoarele rezultate:

Faţa	1	2	3	4	5	6
Frecvenţa	15	7	4	11	6	17

Folosind nivelul de semnificaţie $\alpha = 0.05$, să verificăm dacă zarul respectiv este fals sau nu.

Se aplică testul χ^2, privind parametrii legii multinomiale. Într-adevăr, evenimentul E_i reprezintă apariţia feţei cu numărul i, $i = \overline{1,6}$. Se face ipoteza nulă

$$H_0 : p_i = \frac{1}{6},\ i = \overline{1,6}, \quad \text{adică zarul nu este fals},$$

cu alternativa

$$H_1 : \exists\, i_0 \text{ astfel încât } p_{i_0} \neq \frac{1}{6}, \quad \text{adică zarul este fals}.$$

Se calculează valoarea statisticii χ^2, care are $k-1 = 5$ grade de libertate, anume

$$x^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}} = \frac{\left(15 - 60\cdot\frac{1}{6}\right)^2}{60\cdot\frac{1}{6}} + \frac{\left(7 - 60\cdot\frac{1}{6}\right)^2}{60\cdot\frac{1}{6}} + \cdots + \frac{\left(17 - 60\cdot\frac{1}{6}\right)^2}{60\cdot\frac{1}{6}} = 13.6.$$

Pe de altă parte, din *Anexa III*, avem cuantila $\chi^2_{k-1,1-\alpha} = \chi^2_{5,0.95} = 11.1$, deci rezultă $x^2 = 13.6 > 11.1 = \chi^2_{5,0.95}$, ceea ce conduce la respingerea ipotezei nule, adică acceptăm ipoteza că zarul este fals.

Programul 6.9.4. Programul Matlab, care urmează, efectuează calculele din exemplul precedent:

```
x=1:6; f=[15,7,4,11,6,17]; p=1/6*ones(1,6);
x2=sum((f-60*p).^2./(60*p));
cuant=chi2inv(0.95,5);
fprintf(' x^2=  %6.1f\n cuant= %5.2f',x2,cuant)
```

iar rezultatele obţinute sunt:

```
 x^2=    13.6
 cuant= 11.07
```

Aplicaţia 6.9.5. (*Testul χ^2 neparametric privind concordanţa*). Fie caracteristica X, care urmează legea de probabilitate cu funcţia de repartiţie F necunoscută. Relativ la legea de probabilitate, se face ipoteza nulă $H_0 : F = F_0$, cu ipoteza alternativă $H_1 : F \neq F_0$. Când considerăm testul χ^2 neparametric, funcţia de repartiţie F_0 nu depinde de nici un parametru necunoscut.

Dacă domeniul valorilor caracteristicii X este, să zicem, intervalul (a, b) şi dacă vom considera clasele obţinute prin punctele $a = a_0 < a_1 < \cdots < a_k = b$, atunci apar parametrii necunoscuţi

$$p_i = P\left(a_{i-1} < X \leqslant a_i\right) = F\left(a_i\right) - F\left(a_{i-1}\right), \quad i = \overline{1,k}.$$

De asemenea, evenimentul E_i va fi evenimentul ca un individ luat la întâmplare din colectivitatea cercetată să aparţină clasei $[a_{i-1}, a_i)$. Prin urmare, s-a ajuns la aplicarea testului χ^2 privind concordanţa. Ipoteza nulă, mai sus precizată, devine în acest fel $H_0 : p_i = p_i^{(0)}$, $i = \overline{1,k}$, iar ipoteza alternativă se va rescrie H_1: există i_0 astfel încât $p_{i_0} \neq p_{i_0}^{(0)}$, unde $p_i^{(0)} = P\left(a_{i-1} < X \leqslant a_i \,|\, H_0\right) = F_0\left(a_i\right) - F_0\left(a_{i-1}\right)$.

Etapele aplicării testului χ^2 neparametric

1. Se dau: $\alpha;\ x_1, x_2, \ldots, x_n;\ a_0, a_1, \ldots, a_k;\ F_0$;
2. Se calculează frecvenţele absolute n_i, $i = \overline{1,k}$, şi probabilităţile
$$p_i^{(0)} = F_0\left(a_i\right) - F_0\left(a_{i-1}\right), \quad i = \overline{1,k};$$
3. Se calculează $\chi^2_{k-1,1-\alpha}$ astfel încât $F_{k-1}\left(\chi^2_{k-1,1-\alpha}\right) = 1 - \alpha$, F_{k-1} fiind funcţia de repartiţie a legii χ^2 cu $k-1$ grade de libertate;
4. Se calculează $x^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}}$;

5. Concluzia: dacă $x^2 < \chi^2_{k-1,1-\alpha}$ se admite ipoteza că legea de probabilitate a caracteristicii X este dată de funcţia de repartiţie F_0, în caz contrar ipoteza este respinsă.

Exemplul 6.9.6. Fie caracteristica X ce reprezintă numărul fetelor dintr–o familie cu patru copii. Pentru verificarea ipotezei că X urmează o lege binomială de parametru $p = \frac{1}{2}$, s-a efectuat o selecţie de volum $n = 32$. Distribuţia selecţiei lui X este

$$X \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 4 & 10 & 8 & 7 & 3 \end{pmatrix}.$$

Folosind nivelul de semnificaţie $\alpha = 0.05$, să verificăm dacă X urmează legea binomială. Dacă X urmează legea binomială cu $p = \frac{1}{2}$, atunci are distribuţia teoretică

$$X \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ \frac{1}{16} & \frac{1}{4} & \frac{3}{8} & \frac{1}{4} & \frac{1}{16} \end{pmatrix}.$$

Ipoteza privind faptul că X urmează legea binomială, se scrie sub forma

$$H_0 : p_0 = p_4 = \frac{1}{16},\ p_1 = p_3 = \frac{1}{4},\ p_2 = \frac{3}{8}.$$

Se va aplica testul χ^2 neparametric, unde numărul gradelor de libertate este dat de $k - 1 = 5 - 1 = 4$. Valoarea calculată a caracteristicii χ^2 este

$$x^2 = \sum_{i=0}^{4} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}} = \frac{\left(4 - 32 \cdot \frac{1}{16}\right)^2}{32 \cdot \frac{1}{16}} + \frac{\left(10 - 32 \cdot \frac{1}{4}\right)^2}{32 \cdot \frac{1}{4}} + \cdots + \frac{\left(3 - 32 \cdot \frac{1}{16}\right)^2}{32 \cdot \frac{1}{16}} = 4.5.$$

Cum $x^2 = 4.5 < 9.49 = \chi^2_{4,0.95} = \chi^2_{k-1,1-\alpha}$, avem că ipoteza nulă este admisă.

Programul 6.9.7. Prin executarea programului Matlab:

```
x=0:4; f=[4,10,8,7,3];
p=[1/16,1/4,3/8,1/4,1/16];
x2=sum((f-32*p).^2./(32*p));
cuant=chi2inv(0.95,4);
fprintf(' x^2=  %6.1f\n cuant= %5.2f',x2,cuant)
```

se obţine

```
x^2=     4.5
cuant=  9.49
```

Aplicaţia 6.9.8. (*Testul χ^2 neparametric privind exponenţialitatea*). Relativ la caracteristica X se face ipoteza nulă $H_0 : F = F_0$, unde $F_0(x) = 1 - e^{-\frac{x}{\mu}}$, $x > 0$ şi parametrul $\mu > 0$ cunoscut, adică X urmează legea exponenţială.

Deoarece domeniul valorilor caracteristicii X este intervalul $(0, +\infty)$, se consideră clasele caracteristicii X date prin $0 = a_0 < a_1 < \cdots < a_k = +\infty$. Astfel avem:

$$p_i^{(0)} = F_0(a_i) - F_0(a_{i-1}) = e^{-\frac{1}{\mu}a_{i-1}} - e^{-\frac{1}{\mu}a_i},\ i = \overline{1, k-1} \quad \text{şi } p_k^{(0)} = e^{-\frac{1}{\mu}a_{k-1}}.$$

Etapele aplicării testului

1. Se dau: $\alpha;\ x_1, x_2, \ldots, x_n;\ a_1, a_2, \ldots, a_{k-1};\ \mu$;
2. Se calculează frecvenţele absolute n_i, $i = \overline{1, k}$, şi probabilităţile
$$p_1^{(0)} = 1 - e^{-\frac{1}{\mu}a_1},\ p_i^{(0)} = e^{-\frac{1}{\mu}a_{i-1}} - e^{-\frac{1}{\mu}a_i},\ i = \overline{2, k-1},\ p_k^{(0)} = e^{-\frac{1}{\mu}a_{k-1}};$$
3. Se calculează $\chi^2_{k-1,1-\alpha}$ astfel încât $F_{k-1}\left(\chi^2_{k-1,1-\alpha}\right) = 1 - \alpha$;
4. Se calculează $x^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}}$;
5. Concluzia: dacă $x^2 < \chi^2_{k-1,1-\alpha}$ se acceptă ipoteza că X urmează legea exponenţială de parametru λ, în caz contrar ipoteza este respinsă.

Aplicaţia 6.9.9. (*Testul χ^2 parametric privind concordanţa*). Se consideră caracteristica X care urmează legea de probabilitate cu funcţia de repartiţie F necunoscută. Relativ la legea de probabilitate se face ipoteza nulă $H_0 : F = F_0$ cu ipoteza alternativă $H_1 : F \neq F_0$. Faţă de cazul neparametric, funcţia de repartiţie F_0 se consideră că depinde de s parametri necunoscuţi. Fie aceşti parametri $\theta_1, \theta_2, \ldots, \theta_s$, adică $F_0 = F_0(x; \theta_1, \theta_2, \ldots, \theta_s)$.

În acest caz, faţă de cazul neparametric, la început se estimează parametrii necunoscuţi, folosind datele de selecţie considerate, cu ajutorul metodei verosimilităţii maxime. Fie estimaţiile de verosimilitate maximă ale parametrilor mai sus precizaţi, respectiv $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_s$.

În continuare, cele expuse la cazul neparametric rămân neschimbate cu două observaţii. În primul rând, probabilităţile claselor considerate se calculează prin formula

$$p_i^{(0)} = P(a_{i-1} < X \leqslant a_i \mid H_0) = F_0\left(a_i; \widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_s\right) - F_0\left(a_{i-1}; \widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_s\right),$$

pentru $i = \overline{1, k}$. În al doilea rând, legea χ^2, în acest caz, are $k - s - 1$ grade de libertate.

Etapele aplicării testului χ^2 parametric

1. Se dau: $\alpha;\ x_1, x_2, \ldots, x_n;\ a_0, a_1, \ldots, a_k;\ F_0 = F_0\left(x; \theta_1, \theta_2, \ldots, \theta_s\right)$;
2. Se determină estimaţiile de verosimilitate maximă $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_s$ pentru parametrii $\theta_1, \theta_2, \ldots, \theta_s$;
3. Se calculează frecvenţele absolute n_i, $i = \overline{1,k}$, şi probabilităţile
$$p_i^{(0)} = F_0\left(a_i; \widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_s\right) - F_0\left(a_{i-1}; \widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_s\right), \quad i = \overline{1,k};$$
4. Se calculează $\chi^2_{k-s-1,1-\alpha}$ astfel încât $F_{k-s-1}\left(\chi^2_{k-s-1,1-\alpha}\right) = 1 - \alpha$;
5. Se calculează $x^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}}$;
6. Concluzia: dacă $x^2 < \chi^2_{k-s-1,1-\alpha}$ se acceptă ipoteza că X urmează legea de probabilitate dată prin funcţia de repartiţie $F_0 = F_0\left(x; \widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_s\right)$, în caz contrar ipoteza este respinsă.

Exemplul 6.9.10. O substanţă radioactivă este observată în 2608 intervale de timp de lungimi egale (7.5 secunde). Pentru fiecare interval de timp este înregistrat numărul particulelor emise. Rezultatele obţinute sunt:

Nr. part.	0	1	2	3	4	5	6	7	8	9	$\geqslant 10$
n_i	57	203	383	525	532	408	273	139	45	27	16

Notând cu X caracteristica ce reprezintă numărul de particule emise într-un interval de timp de lungime 7.5 secunde, vrem să verificăm dacă X urmează legea lui Poisson, folosind nivelul de semnificaţie $\alpha = 0.01$.

Folosim testul χ^2 parametric privind concordanţa, deoarece parametrul legii lui Poisson, $\lambda = E(X)$, este necunoscut. Aşadar, numărul gradelor de libertate va fi $k - s - 1 = 11 - 1 - 1 = 9$. Avem că estimatorul de verosimilitate maximă pentru λ este media de selecţie. Prin urmare avem

$$\widehat{\lambda} = \frac{1}{n}\sum_{i=0}^{k} n_i x_i = \frac{1}{2608}\left(0 \cdot 57 + 1 \cdot 203 + \cdots + 10 \cdot 16\right) = 3.87.$$

Se calculează apoi probabilităţile de la legea lui Poisson, cu parametrul $\widehat{\lambda} = 3.87$, adică $p_i^{(0)} = \frac{\widehat{\lambda}^i}{i!} e^{-\widehat{\lambda}}$, $i = 0, 1, 2, \ldots$. Astfel se obţin valorile:

i	0	1	2	3	4	5	6	7	8	9	10
$p_i^{(0)}$	0.021	0.081	0.156	0.202	0.195	0.151	0.097	0.054	0.026	0.011	0.006

de unde

$$x^2 = \sum_{i=0}^{10} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}} = 14.9.$$

Pe de altă parte, folosind *Anexa III*, avem cuantila $\chi^2_{9,0.99} = 21.7$ şi prin urmare $x^2 = 14.9 < 21.7 = \chi^2_{9,0.99}$, deci ipoteza că X urmează legea lui Poisson este acceptată.

Programul 6.9.11. Ilustrarea calculelor este făcută prin programul Matlab

```
x=0:10;
f=[57,203,383,525,532,408,273,139,45,27,16];
la=sum(f.*x)/2608; p=poisspdf(x,la);
x2=sum((f-2608*p).^2./(2608*p));
cuant=chi2inv(0.99,9);
fprintf(' x^2=  %6.1f\n cuant= %5.2f',x2,cuant)
```

care prin executare conduce la următoarele rezultate:

```
 x^2=    14.9
 cuant= 21.67
```

Aplicaţia 6.9.12. (*Testul χ^2 parametric privind normalitatea*). Relativ la caracteristica X se consideră ipoteza nulă $H_0 : F = F_0$, unde

$$F_0(x) = F_0(x; m, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} dt, \quad x \in \mathbb{R},$$

$m \in \mathbb{R}$ şi $\sigma > 0$ sunt parametri necunoscuţi, adică faptul că X urmează legea normală cu cei doi parametri necunoscuţi.

Folosind metoda verosimilităţii maxime se determină, pe baza datelor de selecţie $x_1, x_2, \ldots, x_n$, estimaţiile de verosimilitate maximă pentru m şi respectiv σ, anume

$$\widehat{m} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \quad \text{şi} \quad \widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\bar{\mu}_2}$$

(a se vedea Exemplul 5.3.7). Deoarece domeniul valorilor caracteristicii X este $\mathbb{R}$ se consideră clasele date prin $-\infty = a_0 < a_1 < \cdots < a_k = +\infty$.

Pe de altă parte probabilităţile corespunzătoare acestor clase se obţin cu formulele

$$p_i^{(0)} = F_0\left(a_i; \bar{x}, \sqrt{\bar{\mu}_2}\right) - F_0\left(a_{i-1}; \bar{x}, \sqrt{\bar{\mu}_2}\right), \quad i = \overline{1,k}.$$

Având în vedere că

$$F_0(x; m, \sigma) = \frac{1}{2} + \Phi\left(\frac{x-m}{\sigma}\right),$$

unde Φ este funcţia lui Laplace, adică

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \mathrm{e}^{-\frac{t^2}{2}} dt,$$

rezultă că

$$p_i^{(0)} = \Phi\left(\frac{a_i - \bar{x}}{\sqrt{\bar{\mu}_2}}\right) - \Phi\left(\frac{a_{i-1} - \bar{x}}{\sqrt{\bar{\mu}_2}}\right), \quad i = \overline{1,k}.$$

Etapele aplicării testului

1. Se dau: $\alpha;\ x_1, x_2, \ldots, x_n;\ a_1, a_2, \ldots, a_{k-1}$;
2. Se calculează frecvenţele absolute n_i, $i = \overline{1,k}$, estimaţiile

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

şi probabilităţile

$$p_1^{(0)} = \frac{1}{2} + \Phi\left(\frac{a_1 - \bar{x}}{\sqrt{\bar{\mu}_2}}\right), \quad p_k^{(0)} = \frac{1}{2} - \Phi\left(\frac{a_{k-1} - \bar{x}}{\sqrt{\bar{\mu}_2}}\right),$$
$$p_i^{(0)} = \Phi\left(\frac{a_i - \bar{x}}{\sqrt{\bar{\mu}_2}}\right) - \Phi\left(\frac{a_{i-1} - \bar{x}}{\sqrt{\bar{\mu}_2}}\right), \quad i = \overline{2, k-1};$$

3. Se calculează $\chi^2_{k-3,1-\alpha}$ astfel încât $F_{k-3}\left(\chi^2_{k-3,1-\alpha}\right) = 1 - \alpha$;
4. Se calculează $x^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}}$;
5. Concluzia: dacă $x^2 < \chi^2_{k-3,1-\alpha}$ se acceptă ipoteza că X urmează legea normală $\mathcal{N}(\bar{x}, \sqrt{\bar{\mu}_2})$, în caz contrar ipoteza este respinsă.

Exemplul 6.9.13. Se consideră caracteristica X, ce reprezintă rezistenţa, în $K\Omega$, a unor tronsoane de ceramică acoperite cu carbon. Să se verifice normalitatea lui X, folosind o selecţie de volum $n = 124$, pentru care s–au obţinut datele de selecţie

clasa	$(-\infty; 1.65)$	$(1.65; 1.70]$	$(1.70; 1.75]$	$(1.75; 1.80]$	$(1.80; 1.85]$
frecv.	11	14	17	17	18

$(1.85; 1.90]$	$(1.90; 1.95]$	$(1.95; 2.00]$	$(2.0; +\infty)$
16	13	10	8

utilizând testul de concordanţă χ^2, cu nivelul de semnificaţie $\alpha = 0.05$.

Prima dată se estimează parametrii de la legea normală $\mathcal{N}(m, \sigma)$, adică media teoretică $m = E(X)$ şi abaterea standard teoretică $\sigma = \sqrt{Var(X)}$, folosind metoda de verosimilitate maximă. Se cunoaşte că estimaţiile de verosimilitate maximă pentru m şi σ sunt respectiv

$$\widehat{m} = \frac{1}{n} = \sum_{i=1}^{k} x_i = \bar{x}, \quad \widehat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{k} (x_i - \bar{x})^2} = \sqrt{\bar{\mu}_2},$$

(a se vedea Exemplul 5.3.7). Avem distribuţia empirică de selecţie pentru caracteristica X

$$X \begin{pmatrix} 1.625 & 1.675 & 1.725 & 1.775 & 1.825 & 1.875 & 1.925 & 1.975 & 2.125 \\ 11 & 14 & 17 & 17 & 18 & 16 & 13 & 10 & 8 \end{pmatrix}$$

de unde calculăm

$$\widehat{m} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{124} (11 \cdot 1.625 + 14 \cdot 1.675 + \cdots + 8 \cdot 2.125) = 1.82;$$

$$\widehat{\sigma} = \sqrt{\bar{\mu}_2} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} = 0.129.$$

Se consideră valoarea numerică

$$x^2 = \sum_{i=1}^{k} \frac{(f_i - n\widehat{p}_i)^2}{n\widehat{p}_i},$$

unde k este numărul claselor ($k = 9$, în cazul de faţă), f_i este frecvenţa clasei i, iar $\widehat{p}_i$ este dat prin

$$\widehat{p}_i = \Phi\left(\frac{a_i - \bar{x}}{\widehat{\sigma}}\right) - \Phi\left(\frac{a_{i-1} - \bar{x}}{\widehat{\sigma}}\right),$$

subintervalul $[a_{i-1}, a_i)$ definind clasa i. După cum este cunoscut, x^2 este valoarea unei variabile aleatoare care urmează legea χ^2 cu $k - s - 1$ grade de libertate, s fiind numărul parametrilor estimaţi. În cazul de faţă avem că $s = 2$, deci avem $9 - 2 - 1 = 6$ grade de libertate.

Se determină intervalul $\left(0, \chi^2_{k-3,1-\alpha}\right)$, pentru statistica χ^2, folosind *Anexa III*. Anume, se obţine că $\chi^2_{k-3,1-\alpha} = \chi^2_{6;0.95} = 12.59$, adică intervalul numeric pentru statistica χ^2 este $(0\,;\,12.59)$.

Calculele pentru valoarea numerică x^2 se aranjează în următorul tabel:

a_i	f_i	$\frac{a_i-\bar{x}}{\widehat{\sigma}}$	$\Phi\left(\frac{a_i-\bar{x}}{\widehat{\sigma}}\right)$	$\widehat{p}_i$	$n\widehat{p}_i$	$\frac{(f_i-n\widehat{p}_i)^2}{n\widehat{p}_i}$
1.65	11	-1.40	-0.4188	0.0812	10.0647	0.0869
1.70	14	-0.97	-0.3238	0.0861	10.6711	1.0385
1.75	17	-0.53	-0.2030	0.1298	16.0891	0.0516
1.80	17	-0.10	-0.0402	0.1628	20.1848	0.5025
1.85	18	0.33	0.1297	0.1699	21.0714	0.4477
1.90	16	0.76	0.2773	0.1476	18.3036	0.2899
1.95	13	1.20	0.3840	0.1067	13.2299	0.0040
2.00	10	1.63	0.4482	0.0642	7.9568	0.5247
$+\infty$	8	$+\infty$	0.5000	0.0518	6.4286	0.3841
	$n = 124$					$x^2 = 3.3$

Valorile funcţiei lui Laplace Φ se iau din *Anexa I* şi se are în vedere că funcţia Φ este impară, adică $\Phi(-x) = -\Phi(x)$. De asemenea, facem observaţia că

$$\widehat{p}_1 = \Phi\left(\frac{a_1-\bar{x}}{\widehat{\sigma}}\right) - \Phi\left(\frac{a_0-\bar{x}}{\widehat{\sigma}}\right) = \Phi(-1.4) - \Phi(-\infty)$$
$$= -0.4188 + 0.5 = 0.0812.$$

Deoarece $x^2 = 3.3 \in (0\,;\,12.59)$, rezultă că se acceptă ipoteza normalităţii caracteristicii X.

Programul 6.9.14. Toate calculele din tabelul precedent pot fi efectuate prin programul Matlab:

```
x=1.625:0.05:2.025; f=[11,14,17,17,18,16,13,10,8];
ma=sum(f.*x)/124; s=sqrt(sum(f.*(x-ma).^2)/124);
a=[-inf,1.65:0.05:2,inf];
F=normcdf(a,ma,s); p=diff(F);
x2=sum((f-124*p).^2./(124*p));
cuant=chi2inv(0.95,6);
fprintf(' x^2=  %6.1f\n cuant= %5.2f',x2,cuant)
```

Rezultatele obţinute în urma executării programului sunt:

```
x^2=     3.3
cuant= 12.59
```

6.10 Testul χ^2 pentru compararea mai multor caracteristici

Se consideră colectivităţile $\mathcal{C}_i$, $i = \overline{1,r}$, independente, cercetate din punct de vedere al aceleaşi caracteristici (calitative sau cantitative). Fie această caracteristică X_i pentru colectivitatea $\mathcal{C}_i$. Relativ la caracteristicile X_i, $i = \overline{1,r}$, se face ipoteza nulă H_0, că acestea urmează aceeaşi lege de probabilitate. Se notează cu s numărul claselor caracteristicii cercetate şi E_j evenimentul ca un individ luat la întâmplare din una din colectivităţile cercetate să aparţină clasei cu numărul de ordine j.

Dacă se notează prin $p_{ij} = P(E_j \mid \mathcal{C}_i)$, adică probabilitatea ca un individ luat la întâmplare din colectivitatea $\mathcal{C}_i$ să aparţină clasei cu numărul de ordine j, atunci ipoteza nulă se poate rescrie $H_0 : p_{ij} = p_j,\ i = \overline{1,r},\ j = \overline{1,s}$.

Probabilităţile astfel definite satisfac relaţiile

$$\sum_{j=1}^{s} p_{ij} = 1, \quad i = \overline{1,r},$$

iar când ipoteza H_0 este satisfăcută, avem de asemenea $\sum_{j=1}^{s} p_j = 1$.

Pentru a verifica această ipoteză se consideră câte o selecţie repetată de volum n_i, $i = \overline{1,r}$. Datele de selecţie sunt trecute în tabelul următor.

Selecţia	Vol. selecţiei	E_1	E_2	$\cdots$	E_s	Total
$\mathcal{S}_1$	n_1	n_{11}	n_{12}	$\ldots$	n_{1s}	$n_{1\cdot} = n_1$
$\mathcal{S}_2$	n_2	n_{21}	n_{22}	$\ldots$	n_{2s}	$n_{2\cdot} = n_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
$\mathcal{S}_r$	n_r	n_{r1}	n_{r2}	$\ldots$	n_{rs}	$n_{r\cdot} = n_r$
Total	n	$n_{\cdot 1}$	$n_{\cdot 2}$	$\ldots$	$n_{\cdot s}$	$n_{\cdot\cdot} = n$

Elementul n_{ij} al tabelului reprezintă frecvenţa absolută a apariţiei clasei i în selecţia $\mathcal{S}_j$, iar

$$n_{i\cdot} = \sum_{j=1}^{s} n_{ij},\ i = \overline{1,r}, \quad n_{\cdot j} = \sum_{i=1}^{r} n_{ij},\ j = \overline{1,s}, \quad n_{\cdot\cdot} = \sum_{i=1}^{r} n_{i\cdot} = \sum_{j=1}^{s} n_{\cdot j}.$$

Proprietatea 6.10.1. *Estimaţiile de verosimilitate maximă pentru parametrii necunoscuţi p_j, $j = \overline{1,s}$, sunt date prin*

$$\widehat{p}_j = \frac{n_{\cdot j}}{n}, \quad j = \overline{1,s}.$$

Demonstraţie. Funcţia de verosimilitate are expresia

$$g=\prod_{i=1}^{r}\prod_{j=1}^{s}p_{ij}^{n_{ij}}=\prod_{i=1}^{r}\prod_{j=1}^{s}p_{j}^{n_{ij}}=\prod_{j=1}^{s}\Big(\prod_{i=1}^{r}p_{j}^{n_{ij}}\Big)=\prod_{j=1}^{s}p_{j}^{n_{\cdot j}},$$

când ipoteza nulă H_0 este adevărată. Având în vedere relaţia dintre probabilităţile p_j, $j=\overline{1,s}$, se obţine că

$$g=\Big(1-\sum_{j=1}^{s-1}p_j\Big)^{n_{\cdot s}}\prod_{j=1}^{s-1}p_j^{n_{\cdot j}}$$

sau prin logaritmare, se ajunge la

$$\ln g=n_{\cdot s}\ln\Big(1-\sum_{j=1}^{s-1}p_j\Big)+\sum_{j=1}^{s-1}n_{\cdot j}\ln p_j.$$

Sistemul ecuaţiilor de verosimilitate maximă va fi

$$\frac{\partial\ln g}{\partial p_k}=-\frac{n_{\cdot s}}{1-\sum_{j=1}^{s-1}p_j}+\frac{n_{\cdot k}}{p_k}=0,\quad k=\overline{1,s-1},$$

de unde rezultă că

$$\frac{n_{\cdot k}}{p_k}=\frac{n_{\cdot s}}{p_s},\quad k=\overline{1,s},$$

sau, având în vedere proprietăţi ale şirului de rapoarte egale, se obţine

$$\frac{n_{\cdot k}}{p_k}=\frac{n}{1},\quad k=\overline{1,s}.$$

S-a ajuns astfel la estimaţiile

$$\widehat{p}_k=\frac{n_{\cdot k}}{n},\quad k=\overline{1,s}.$$

□

Observaţia 6.10.2. Dacă se are în vedere că n_{ij} sunt valorile unor variabile aleatoare binomiale N_{ij}, rezultă că statistica

$$\chi^2=\sum_{i=1}^{r}\Big[\sum_{j=1}^{s}\frac{(N_{ij}-N_{i\cdot}\widehat{p}_j)^2}{N_{i\cdot}\widehat{p}_j}\Big]=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\big(N_{ij}-\frac{N_{i\cdot}N_{\cdot j}}{n}\big)^2}{\frac{N_{i\cdot}N_{\cdot j}}{n}},$$

cu

$$\sum_{j=1}^{s}N_{ij}=N_{i\cdot}\ i=\overline{1,r},\quad \sum_{i=1}^{r}N_{ij}=N_{\cdot j},\ j=\overline{1,s},\quad \sum_{j=1}^{s}N_{\cdot j}=\sum_{i=1}^{r}N_{i\cdot}=n_{\cdot\cdot}=n.$$

urmează o lege χ^2. Pentru a stabili numărul m al gradelor de libertate se ţine seama de faptul că numărul variabilelor aleatoare N_{ij} este rs, numărul legăturilor între p_{ij} este r, iar numărul parametrilor estimaţi este $s-1$. Prin urmare, numărul gradelor de libertate este $m = rs - r - (s-1) = (r-1)(s-1)$.

Etapele aplicării testului

1. Se dau: α; n_{ij}, $i = \overline{1,r}$, $j = \overline{1,s}$;
2. Se calculează $n_{i\cdot} = \sum_{j=1}^{s} n_{ij}$, $i = \overline{1,r}$, $n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$, $j = \overline{1,s}$;
3. Se calculează $\chi^2_{m,1-\alpha}$, astfel încât $F_m\left(\chi^2_{m,1-\alpha}\right) = 1-\alpha$;
4. Se calculează $x^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}$;
5. Concluzia: dacă $x^2 < \chi^2_{m,1-\alpha}$ se acceptă ipoteza nulă H_0, în caz contrar se respinge.

Observaţia 6.10.3. Când probabiliăţile p_j, $j = \overline{1,r}$, sunt cunoscute (caz mai rar întâlnit), deci nu se cere a fi estimate, atunci statistica

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(N_{ij} - N_{i\cdot}p_j\right)^2}{N_{i\cdot}p_j}$$

urmează legea χ^2 cu $m = rs - r = r(s-1)$ grade de libertate. Aplicarea testului, în acest caz, se adaptează în mod corespunzător.

Exemplul 6.10.4. S-au considerat trei loturi de bolnavi de emfizem pulmonar, în funcţie de numărul ţigărilor fumate zilnic: mai puţin de un pachet, unul sau două pachete şi respectiv mai mult de două pachete. Pentru emfizemul pulmonar sunt considerate 4 stadii notate de la I la IV. Rezultatele cercetărilor sunt trecute în următorul tabel sistematizat

Categoria \ Stadiul	I	II	III	IV	
<1 pachet	41	28	25	6	100
$1-2$ pachete	24	116	46	14	200
>2 pachete	4	50	34	12	100
	69	194	105	32	400

Vrem să să verificăm, cu nivelul de semnificaţie $\alpha = 0.01$, ipoteza că numărul de ţigări fumate zilnic nu influenţează repartizarea bolnavilor în cele patru stadii ale bolii. Verificăm ipoteza făcută cu ajutorul testului χ^2 privind compararea aceluiaşi atribut, dar pentru populaţii diferite. Numărul gradelor de libertate este dat prin formula $(r-1)(s-1) = (3-1)(4-1) = 6$, pentru care se obţine cuantila $\chi^2_{k,1-\alpha} = \chi^2_{6,0.99} = 16.8$.

Se calculează apoi valoarea

$$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}$$
$$= \frac{\left(41 - \frac{100\cdot 69}{400}\right)^2}{\frac{100\cdot 69}{400}} + \frac{\left(28 - \frac{100\cdot 194}{400}\right)^2}{\frac{100\cdot 194}{400}} + \cdots + \frac{\left(12 - \frac{100\cdot 32}{400}\right)^2}{\frac{100\cdot 32}{400}} = 64.41.$$

Deoarece $x^2 = 64.41 > 16.8 = \chi^2_{k,1-\alpha}$, se respinge ipoteza că repartizarea pe cele patru stadii ale bolii nu depinde de numărul ţigărilor fumate zilnic.

Programul 6.10.5. Programul Matlab

```
f=[41,28,25,6;24,116,46,14;4,50,34,12];
fip=sum(f'); fpj=sum(f); x2=0;
for i=1:3
    for j=1:4
        t=fip(i)*fpj(j)/400;
        x2=x2+(f(i,j)-t)^2/t;
    end
end
cuant=chi2inv(0.99,6);
fprintf(' x^2=  %6.1f\n cuant= %5.2f',x2,cuant)
```

în urma executării, conduce la rezultatele din exemplul precedent:

```
 x^2=    64.4
 cuant= 16.81
```

6.11 Testul χ^2 pentru tabele de contingenţă

Se consideră colectivitatea $\mathcal{C}$ cercetată din punct de vedere a două caracteristici X şi Y (calitative sau cantitative). Se pune problema independenţei celor două caracteristici. Fie r numărul claselor pentru caracteristica X şi respectiv s pentru caracteristica Y. Notăm cu A_i evenimentul ca un individ luat la întâmplare din colectivitatea $\mathcal{C}$ să aparţină clasei cu numărul de ordine i în raport cu X şi notăm cu B_j evenimentul ca un individ luat la întâmplare din $\mathcal{C}$ să aparţină clasei cu numărul de ordine j în raport

cu Y. Fie $p_{ij} = P(A_i \cap B_j)$, $i = \overline{1,r}$, $j = \overline{1,s}$. Dacă se notează

$$p_{i\cdot} = \sum_{j=1}^{s} p_{ij},\ i = \overline{1,r} \quad \text{şi} \quad p_{\cdot j} = \sum_{i=1}^{r} p_{ij},\ j = \overline{1,s},$$

atunci ipoteza nulă relativă la independenţa celor două caracteristici se scrie

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}, \quad i = \overline{1,r},\ j = \overline{1,s},$$

cu ipoteza alternativă H_1, că H_0 este falsă.

Pentru verificarea acestei ipoteze se consideră datele de selecţie (x_i, y_i), $i = \overline{1,n}$. Datele de selecţie sunt sistematizate în tabelul de contingenţă

$X \setminus Y$	B_1	B_2	$\dots$	B_s	
A_1	n_{11}	n_{12}	$\dots$	n_{1s}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$\dots$	n_{2s}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\dots$	n_{rs}	$n_{r\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot s}$	$n_{\cdot\cdot} = n$

unde n_{ij} reprezintă frecvenţa absolută a clasei (i, j), iar

$$n_{i\cdot} = \sum_{j=1}^{s} n_{ij}, \quad i = \overline{1,r}, \quad n_{\cdot j} = \sum_{i=1}^{r} n_{ij}, \quad j = \overline{1,s}.$$

Proprietatea 6.11.1. *Estimaţiile de verosimilitate maximă pentru parametrii necunoscuţi $p_{i\cdot}$, $i = \overline{1,r}$, şi $p_{\cdot j}$, $j = \overline{1,s}$, sunt date prin formulele:*

$$\widehat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}, \quad i = \overline{1,r}, \quad \widehat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}, \quad j = \overline{1,s}.$$

Demonstraţie. Funcţia de verosimilitate are expresia

$$\begin{aligned} g &= \prod_{i=1}^{r}\prod_{j=1}^{s} (p_{ij})^{n_{ij}} = \prod_{i=1}^{r}\prod_{j=1}^{s} \Big(p_{i\cdot}p_{\cdot j}\Big)^{n_{ij}} = \Big(\prod_{i=1}^{r} p_{i\cdot}^{n_{i\cdot}}\Big)\Big(\prod_{j=1}^{s} p_{\cdot j}^{n_{\cdot j}}\Big) \\ &= \Big(1 - \sum_{i=1}^{r-1} p_{i\cdot}\Big)^{n_{r\cdot}} \Big(\prod_{i=1}^{r-1} p_{i\cdot}^{n_{i\cdot}}\Big)\Big(\prod_{j=1}^{s} p_{\cdot j}^{n_{\cdot j}}\Big), \end{aligned}$$

când ipoteza nulă H_0 este adevărată. Prin logaritmare se obţine

$$\ln g = n_{r\cdot}\ln\Big(1 - \sum_{i=1}^{r-1} p_{i\cdot}\Big) + \sum_{i=1}^{r-1} n_{i\cdot}\ln p_{i\cdot} + \ln\Big(\prod_{j=1}^{s} p_{\cdot j}^{n_{\cdot j}}\Big),$$

de unde se ajunge la sistemul ecuaţiilor de verosimilitate maximă

$$\frac{\partial \ln g}{\partial p_k} = -\frac{n_{r\cdot}}{1-\sum_{i=1}^{r-1} p_{i\cdot}} + \frac{n_{k\cdot}}{p_{k\cdot}} = 0, \quad k = \overline{1, r-1},$$

sau

$$\frac{n_{r\cdot}}{p_{r\cdot}} = \frac{n_{k\cdot}}{p_{k\cdot}}, \quad k = \overline{1, r}.$$

Având în vedere proprietăţile unui şir de rapoarte egale se obţine că

$$\frac{n_{k\cdot}}{p_{k\cdot}} = \frac{n}{1}, \quad k = \overline{1, r}, \quad \text{deci} \quad \widehat{p}_{k\cdot} = \frac{n_{k\cdot}}{n}, \quad k = \overline{1, r}.$$

În mod analog se obţine şi faptul că

$$\widehat{p}_{\cdot k} = \frac{n_{\cdot k}}{n}, \quad k = \overline{1, s}.$$

□

Observaţia 6.11.2. Valoarea numerică

$$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - n_{i\cdot}\widehat{p}_{ij}\right)^2}{n\widehat{p}_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}},$$

este cea a unei variabile aleatoare ce urmează legea χ^2 cu $m = (r-1)(s-1)$ grade de libertate. Numărul m al gradelor de libertate s-a determinat având în vedere că au fost estimaţi $(r-1)+(s-1)$ parametri, iar între $p_{i\cdot}$ şi $p_{\cdot j}$ există o singură legătură

$$\sum_{i=1}^{r} p_{i\cdot} = \sum_{j=1}^{s} p_{\cdot j} = 1, \quad \text{adică} \quad \sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij} = 1.$$

Aşadar $m = rs - (r-1) - (s-1) - 1 = (r-1)(s-1)$.

Etapele aplicării testului

1. Se dau: α; n_{ij}, $i = \overline{1, r}$, $j = \overline{1, s}$;
2. Se calculează: $n_{i\cdot} = \sum_{j=1}^{s} n_{ij}$, $i = \overline{1, r}$, $n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$, $j = \overline{1, s}$;
3. Se calculează $\chi^2_{m,1-\alpha}$, astfel încât $F_m\left(\chi^2_{m,1-\alpha}\right) = 1-\alpha$;

4. Se calculează $x^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}$;

5. Concluzia: dacă $x^2 < \chi^2_{m,1-\alpha}$ se acceptă ipoteza nulă H_0, în caz contrar se respinge.

Observaţia 6.11.3. Când probabilităţile $p_{i\cdot}$ şi $p_{\cdot j}$ sunt cunoscute (caz mai rar întâlnit), avem că

$$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - np_{ij})^2}{np_{ij}} - \sum_{i=1}^{r} \frac{(n_{i\cdot} - np_{i\cdot})^2}{np_{i\cdot}} - \sum_{j=1}^{s} \frac{(n_{\cdot j} - np_{\cdot j})^2}{np_{\cdot j}}$$

este valoarea unei variabile aleatoare ce urmează legea χ^2 cu $m = (r-1)(s-1)$ grade de libertate. Aceasta rezultă din faptul că prima sumă corespunde unei legi χ^2 cu $rs-1$ grade de libertate, a doua sumă corespunde unei legi χ^2 cu $r-1$ grade de libertate, iar a treia unei legi χ^2 cu $s-1$ grade de libertate. Prin urmare

$$m + (r-1) + (s-1) = rs - 1, \quad \text{de unde} \quad m = (r-1)(s-1).$$

Aplicarea testului, în acest caz, se face prin adaptarea corespunzătoare a cazului precedent.

Observaţia 6.11.4. Testul χ^2 pentru tabele de contingenţă şi respectiv pentru compararea mai multor caracteristici, pare să fie acelaşi . Dar, remarcăm faptul că problemele de la care se porneşte sunt complet diferite.

Exemplul 6.11.5. Pentru cercetarea dependenţei dintre culoarea A a ochilor şi culoarea B a părului, s-a considerat un eşantion format din 6800 indivizi. Rezultatele sunt trecute în tabelul sistematizat următor

A \ B	Blonzi	Şateni	Bruneţi	Roşcaţi	
Albaştri	1768	807	189	47	2811
Gri sau verzi	946	1387	746	53	3132
Căprui	115	438	288	16	857
	2829	2632	1223	116	6800

Folosind nivelul de semnificaţie $\alpha = 0.01$, vom verifica ipoteza că cele două caracteristici (atribute) sunt independente.

Se foloseşte testul χ^2 pentru tabele de contingenţă cu numărul gradelor de libertate $k = (r-1)(s-1) = (3-1)(4-1) = 6$.

Pe de o parte avem că

$$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}$$
$$= \frac{\left(1768 - \frac{2811\cdot 2829}{6800}\right)^2}{\frac{2811\cdot 2829}{6800}} + \frac{\left(807 - \frac{2811\cdot 2632}{6800}\right)^2}{\frac{2811\cdot 2632}{6800}} + \cdots + \frac{\left(16 - \frac{857\cdot 116}{6800}\right)^2}{\frac{857\cdot 116}{6800}} = 1075.$$

Pe de altă parte, avem cuantila $\chi^2_{6,0.99} = 16.8 < 1075 = x^2$, deci se respinge ipoteza că cele două atribute sunt independente.

Programul 6.11.6. Calculele din exemplul precedent pot fi efectuate cu programul Matlab

```
f=[1768,807,189,47;946,1387,746,53;115,438,288,16];
fip=sum(f'); fpj=sum(f); x2=0;
for i=1:3
    for j=1:4
        t=fip(i)*fpj(j)/6800;
        x2=x2+(f(i,j)-t)^2/t;
    end
end
cuant=chi2inv(0.99,6);
fprintf(' x^2=  %6.1f\n cuant= %5.2f',x2,cuant)
```

În urma executării acestui program, se obţin rezultatele:

```
 x^2=  1073.5
 cuant= 16.81
```

6.12 Testul de concordanţă al lui Kolmogorov

Se consideră caracteristica X de tip continuu cu funcţia teoretică de repartiţie F necunoscută. Relativ la funcţia F se face ipoteza nulă $H_0 : F = F_0$, cu una din

(1) $H_1 : F \neq F_0$ (*testul lui Kolmogorov bilateral*),

(2) $H_1 : F > F_0$ (*testul lui Kolmogorov unilateral dreapta*),

(3) $H_1 : F < F_0$ (*testul lui Kolmogorov unilateral stânga*).

Pentru verificarea ipotezei nule H_0, cu una din alternativele precizate, se consideră o selecţie repetată de volum n, cu datele de selecţie $x_1, x_2, \ldots, x_n$, respectiv variabilele de selecţie $X_1, X_2, \ldots, X_n$ corespunzătoare. Se consideră statisticile

$$D_n = \sup_{x\in\mathbb{R}} \left\{ \mid \bar{F}_n(x) - F_0(x) \mid \right\},$$

$$D_n^+ = \sup_{x\in\mathbb{R}} \left\{ \bar{F}_n(x) - F_0(x) \right\},$$
$$D_n^- = \sup_{x\in\mathbb{R}} \left\{ F_0(x) - \bar{F}_n(x) \right\},$$

unde $\bar{F}_n(x)$ este funcţia de repartiţie de selecţie.

Conform teoremei lui Kolmogorov (Teorema 4.2.38) statistica

$$D_n = \sup_{x\in\mathbb{R}} |\, \bar{F}_n(x) - F_0(x) \,|,$$

urmează o lege de probabilitate dată prin legea lui Kolmogorov, când $n \to \infty$. Anume, avem că

$$\lim_{n\to\infty} P\left(\sqrt{n}\, D_n \leqslant x \;\middle|\; H_0\right) = K(x) = \sum_{k=-\infty}^{k=+\infty} (-1)^k \, \mathrm{e}^{-2k^2x^2}, \quad x > 0,$$

funcţia $K(x)$ a lui Kolmogorov fiind tabelată pentru anumite valori în *Anexa V.*

De asemenea,

$$\lim_{n\to\infty} P\left(\sqrt{n}\, D_n^+ \leqslant x\right) = \lim_{n\to\infty} P\left(\sqrt{n}\, D_n^- \leqslant x\right) = K^{\pm}(x) = 1 - e^{-2x^2}, \; x > 0,$$

ce poartă numele de lege χ cu două grade de libertate.

Pentru un nivel de semnificaţie $\alpha \in (0,1)$ fixat, se pot determina cuantilele $k_{1-\alpha}$ şi $k_{1-\alpha}^{\pm}$ astfel încât

$$P\left(\sqrt{n}\, D_n \leqslant k_{1-\alpha}\right) = 1-\alpha, \quad \text{adică} \quad K(k_{1-\alpha}) = 1-\alpha,$$

pentru testul bilateral al lui Kolmogorov,

$$P\left(\sqrt{n}\, D_n^+ \leqslant k_{1-\alpha}^{\pm}\right) = 1-\alpha \quad \text{şi} \quad P\left(\sqrt{n}\, D_n^- \leqslant k_{1-\alpha}^{\pm}\right) = 1-\alpha,$$

adică $K^{\pm}\left(k_{1-\alpha}^{\pm}\right) = 1-\alpha$, pentru testele unilaterale dreapta şi stânga ale lui Kolmogorov.

Corespunzător celor trei alternative, ipoteza H_0 va fi admisă când valorile calculate d_n, d_n^+, d_n^-, pe baza datelor de selecţie $x_1, x_2, \ldots, x_n$ ale statisticilor D_n, D_n^+, D_n^-, satisfac respectiv condiţiile

(1) $\sqrt{n}\, d_n < k_{1-\alpha}$, când $H_1 : F = F_0$,

(2) $\sqrt{n}\, d_n^+ < k_{1-\alpha}^{\pm}$, când $H_1 : F > F_0$,

(3) $\sqrt{n}\, d_n^- < k_{1-\alpha}^{\pm}$, când $H_1 : F < F_0$,

iar în caz contrar va fi respinsă.

Prezentăm, în cele ce urmează, împreună, cele trei teste (bilateral, unilateral dreapta, unilateral stânga) ale lui Kolmogorov. Remarcăm că cele trei teste sunt teste distincte, alegerea unuia dintre ele se face apriori.

Etapele aplicării testului lui Kolmogorov

1. Se dau: α; $x_1, x_2, \ldots, x_n$; $F = F_0$;

2. Se determină

   (1) $k_{1-\alpha}$ astfel încât $K\left(k_{1-\alpha}\right) = 1-\alpha$, când $H_1 : F \neq F_0$,

   (2) $k^{\pm}_{1-\alpha}$ astfel încât $K^{\pm}\left(k^{\pm}_{1-\alpha}\right) = 1-\alpha$, când $H_1 : F > F_0$,

   (3) $k^{\pm}_{1-\alpha}$ astfel încât $K^{\pm}\left(k^{\pm}_{1-\alpha}\right) = 1-\alpha$, când $H_1 : F < F_0$;

3. Se calculează

   (1) $k = \sqrt{n}\, d_n$, când $H_1 : F \neq F_0$,

   (2) $k = \sqrt{n}\, d^+_n$, când $H_1 : F > F_0$,

   (3) $k = \sqrt{n}\, d^-_n$, când $H_1 : F < F_0$;

4. Concluzia: dacă

   (1) $k < k_{1-\alpha}$ ipoteza H_0 este admisă, când $H_1 : F \neq F_0$,

   (2) $k < k^{\pm}_{1-\alpha}$ ipoteza H_0 este admisă, când $H_1 : F > F_0$,

   (3) $k < k^{\pm}_{1-\alpha}$ ipoteza H_0 este admisă, când $H_1 : F < F_0$,

Observaţia 6.12.1. Având în vedere că funcţiile (de repartiţie) K şi $K^{\pm}$ sunt monoton crescătoare şi că pentru determinarea cuantilelor $k_{1-\alpha}$ şi $k^{\pm}_{1-\alpha}$ este necesară inversarea acestor funcţii de repartiţie, se poate evita această operaţie. Pentru aceasta se determină valorile funcţiilor K şi $K^{\pm}$ pe valorile calculate ale statisticilor D_n, D^+_n, D^-_n, adică se calculează respectiv $1-c = K\left(\sqrt{n}d_n\right)$, $1-c = K^{\pm}\left(\sqrt{n}d^+_n\right)$, $1-c = K^{\pm}\left(\sqrt{n}d^-_n\right)$. Astfel, se va respinge ipoteza nulă H_0, dacă $c \leqslant \alpha$, c numindu–se *valoare critică*.

Observaţia 6.12.2. Pentru calculul valorilor statisticilor D_n, D^+_n, D^-_n putem considera că datele de selecţie sunt ordonate crescător, adică $x_1 < x_2 < \ldots < x_n$, altfel se poate face o astfel de ordonare. De asemnea, remarcăm faptul că are loc $D_n = \max\left\{D^+_n, D^-_n\right\}$.

Folosind aceste precizări, precum şi faptul că atât funcţia de repartiţie de selecţie $\bar{F}_n$, cât şi funcţia de repartiţie F_0 sunt funcţii nedescrescătoare, obţinem următoarele

formule de calcul

$$\begin{aligned} d_n^+ &= \sup_{x\in\mathbb{R}} \left\{ \bar{F}_n(x) - F_0(x) \right\} = \max_{k=\overline{1,n}} \left\{ \bar{F}_n(x_k) - F_0(x_k) \right\} \\ &= \max_{k=\overline{1,n}} \left\{ \frac{k}{n} - F_0(x_k) \right\}, \\ d_n^- &= \sup_{x\in\mathbb{R}} \left\{ F_0(x) - \bar{F}_n(x) \right\} = \max_{k=\overline{1,n}} \left\{ F_0(x_k) - \bar{F}_n(x_k - 0) \right\} \\ &= \max_{k=\overline{1,n}} \left\{ F_0(x_k) - \frac{k-1}{n} \right\}, \\ d_n &= \max\left\{ d_n^+, d_n^- \right\}. \end{aligned}$$

Observaţia 6.12.3. Când datele sunt grupate, având clasele date prin punctele

$$a = a_0 < a_1 < \cdots < a_m = b,$$

atunci se folosesc formulele

$$\begin{aligned} d_n &= \max_{k=\overline{0,m}} \left| \bar{F}_n(a_k) - F_0(a_k) \right|, \\ d_n^+ &= \max_{k=\overline{0,m}} \left\{ \bar{F}_n(a_k) - F_0(a_k) \right\}, \\ d_n^- &= \max_{k=\overline{0,m}} \left\{ F_0(a_k) - \bar{F}_n(a_k) \right\}. \end{aligned}$$

Exemplul 6.12.4. Rezultatele măsurătorilor asupra diametrului X, pentru 1000 de piese de acelaşi tip (în mm), sunt cele ce urmează

Diam.	$97.75-98.25$	$98.25-98.75$	$98.75-99.25$	$99.25-99.75$	$99.75-100.25$
Frecv.	21	47	87	158	181

$100.25-100.75$	$100.75-101.25$	$101.25-101.75$	$101.75-102.25$	$102.25-102.75$
201	142	97	41	25

Folosind nivelul de semnificaţie $\alpha = 0.05$, ştiind $m = E(X) = 100.25\,\text{mm}$ şi abaterea standard $\sigma = \sqrt{Var(X)} = 1\,\text{mm}$, se cere verificarea normalităţii caracteristicii X, cu ajutorul testului lui Kolmogorov bilateral, iar apoi cu ajutorul testului χ^2 neparametric.

Ipoteza care se face este că funcţia de repartiţie a lui X este

$$F_0(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} dt, \quad x \in \mathbb{R},$$

unde $m = 100.25$, $\sigma = 1$. Pentru verificarea acestei ipoteze, folosind criteriul lui Kolmogorov, se calculează

$$d_n = \sup_{x \in \mathbb{R}} \left| \bar{F}_n(x) - F_0(x) \right| = \sup_{i = \overline{0,10}} \left| \bar{F}_n(a_i) - F_0(a_i) \right|,$$

unde $a_i = 97.75 + 0.5\, i$, $i = \overline{0,10}$, iar $\bar{F}_n(x) = \bar{F}_{1000}(x)$ este funcţia de repartiţie de selecţie.

Pentru calculul valorilor funcţiei de repartiţie F_0 se ţine seama de faptul că

$$F_0(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} dt = \frac{1}{2} + \Phi\left(\frac{x-m}{\sigma}\right) = \frac{1}{2} + \Phi(x - 100.25),$$

cu funcţia lui Laplace $\Phi(x)$ tabelată în *Anexa I.*

Pe de altă parte, valorile funcţiei de repartiţie de selecţie se calculează cu formula

$$\bar{F}_{1000}(a_i) = \frac{1}{1000} \sum_{j=1}^{i} n_j, \quad \text{pentru } i = \overline{0,10}.$$

Calculele sunt aranjate în următorul tabel.

a_i	n_i	$a_i - m$	$F_0(a_i)$	$\bar{F}_{1000}(a_i)$	$\lvert \bar{F}_{1000}(a_i) - F_0(a_i) \rvert$
97.75	—	−2.5	0.0062	0.000	0.0062
98.25	21	−2.0	0.0228	0.021	0.0018
98.75	47	−1.5	0.0668	0.068	0.0012
99.25	87	−1.0	0.1587	0.155	0.0037
99.75	158	−0.5	0.3085	0.313	0.0045
100.25	181	0.0	0.5000	0.494	0.0060
100.75	201	0.5	0.6915	0.695	0.0035
101.25	142	1.0	0.8413	0.837	0.0043
101.75	97	1.5	0.9332	0.934	0.0008
102.25	41	2.0	0.9772	0.975	0.0022
102.75	25	2.5	0.9938	1.000	0.0062

Din calculele făcute avem că $d_{1000} = 0.0062$, de unde rezultă că

$$\sqrt{n}\, d_n = \sqrt{1000}\, d_{1000} = 10\sqrt{10} \cdot 0.0062 = 0.196.$$

Din *Anexa V*, se determină cuantila $k_{1-\alpha} = k_{0.95} = 1.36$ şi admitem ipoteza de normalitate pentru caracteristica X, deoarece $\sqrt{n}\, d_n = 0.196 < 1.36 = k_{1-\alpha}$.

Pentru a verifica normalitatea caracteristicii X, cu testul χ^2 neparametric, calculăm prima dată probabilităţile

$$p_i^{(0)} = P\left(a_{i-1} < X \leqslant a_i \,|\, F = F_0\right) = \Phi\left(a_i - m\right) - \Phi\left(a_{i-1} - m\right), \quad i = \overline{1,10},$$

unde $a_0 = -\infty$ şi $a_{10} = +\infty$. Astfel avem

X	$(-\infty, a_1]$	$(a_1, a_2]$	$(a_2, a_3]$	$(a_3, a_4]$	$(a_4, a_5]$	$(a_5, a_6]$	$(a_6, a_7]$
$p_i^{(0)}$	0.0228	0.0440	0.0919	0.1498	0.1915	0.1915	0.1498

$(a_7, a_8]$	$(a_8, a_9]$	$(a_9, +\infty)$
0.0919	0.0440	0.0228

de unde

$$x^2 = \sum_{i=1}^{10} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}} = \frac{(21-22.8)^2}{22.8} + \frac{(47-44)^2}{44} + \cdots + \frac{(25-22.8)^2}{22.8} = 3.21.$$

Pe de altă parte, avem cuantila $\chi^2_{k-1,1-\alpha} = \chi^2_{9,0.95} = 16.9$, conform *Anexei III*. Deoarece $x^2 = 3.21 < 16.9 = \chi_{k-1,1-\alpha}$, rezultă că ipoteza de normalitate este acceptată.

Programul 6.12.5. Vom scrie un program Matlab, pentru aplicarea testului lui Kolmogorov şi a testului χ^2, în cazul datelor din exemplul precedent.

```
% Testul lui Kolmogorov
a=97.75:0.5:102.75;
n=[0,21,47,87,158,181,201,142,97,41,25];
F0=normcdf(a,100.25,1); Fn=cumsum(n)/1000;
dn=max(abs(F0-Fn)); ks=sqrt(1000)*dn;
fprintf(' ks=   %6.4f\n',ks)
% Testul chi^2
F0(1)=0; F0(end)=1;n=n(2:end);
p0=diff(F0); x2=sum((n-1000*p0).^2./(1000*p0));
cuant=chi2inv(0.95,9);
fprintf(' x^2=  %6.4f\n',x2)
fprintf(' cuant=%6.2f',cuant)
```

Rezultatele obţinute în urma executării programului sunt:

```
ks=   0.1964
x^2=  3.2117
cuant= 16.92
```

Ceea ce confirmă rezultatele mai sus obţinute.

Exemplul 6.12.6. Se ţin sub observaţie $n = 50$ motoare electrice până la defectarea ultimului dintre ele. Se consideră caracteristica X, ce reprezintă numărul miilor de ore de funcţionare până la defectare. Rezultatele observaţiilor sunt date în tabelul următor

Durata	$0-30$	$30-60$	$60-90$	$90-120$	$120-150$	$150-180$	$180-300$
Frecv.	13	10	5	5	5	5	7

Să se cerceteze exponenţialitatea caracteristicii X, folosind testul χ^2 cu nivelul de semnificaţie $\alpha = 0.05$, iar apoi folosind testul lui Kolmogorov cu acelaşi nivel de semnificaţie.

Legea exponenţială are funcţia de repartiţie

$$F(x;\mu) = 1 - \mathrm{e}^{-\frac{x}{\mu}}, \quad x > 0, \quad \mu > 0 \text{ parametru necunoscut.}$$

Folosind metoda verosimilităţii maxime avem estimaţia pentru μ

$$\widehat{\mu} = \bar{x} = \frac{1}{50}(13 \cdot 15 + 10 \cdot 45 + \cdots + 7 \cdot 240) = 94.5.$$

Se calculează prima dată probabilităţile:

$$p_i^{(0)} = P\left(a_{i-1} < X \leqslant a_i \mid F = F_0\right) = F_0(a_i;\widehat{\mu}) - F_0(a_{i-1};\widehat{\mu}), \quad i = \overline{1,7},$$

unde

$$F_0(x;\widehat{\mu}) = 1 - \mathrm{e}^{-\frac{x}{\widehat{\mu}}} = 1 - \mathrm{e}^{-\frac{x}{94.5}}.$$

Rezultatele sunt trecute în tabelul următor

X	$(0,a_1]$	$(a_1,a_2]$	$(a_2,a_3]$	$(a_3,a_4]$	$(a_4,a_5]$	$(a_5,a_6]$	(a_6,∞)
$p_i^{(0)}$	0.2720	0.1980	0.1442	0.1049	0.0764	0.0556	0.1489

şi care conduce la

$$x^2 = \sum_{i=1}^{k} \frac{\left(n_i - np_i^{(0)}\right)^2}{np_i^{(0)}} = \frac{(13-50\cdot 0.2720)^2}{50\cdot 0.2720} + \cdots + \frac{(7-50\cdot 0.1489)^2}{50\cdot 0.1489} = 2.877.$$

Din *Anexa III* obţinem cuantila $\chi^2_{k-2,1-\alpha} = \chi^2_{5,0.95} = 11.07$. Deoarece avem valoarea calculată $x^2 = 2.877 < 11.07 = \chi^2_{k-2,1-\alpha}$ acceptăm ipoteza exponenţialităţii.

Pentru aplicarea testului lui Kolmogorov, din *Anexa V* se determină cuantila $k_{1-\alpha} = k_{0.95} = 1.36$. Ipoteza exponenţialităţii lui X va fi acceptată dacă

$$\sqrt{n}\, d_n = \sqrt{n} \sup_{i=\overline{0,7}} \left|\bar{F}_n(a_i) - F_0(a_i;\widehat{\mu})\right| < k_{1-\alpha}.$$

Aici $a_7 = +\infty$, iar $\bar{F}_n$ este funcţia de repartiţie de selecţie. Calculele sunt efectuate în tabelul următor.

a_i	30	60	90	120	150	180	300
n_i	13	10	5	5	5	5	7
$\bar{F}_n(a_i)$	0.26	0.46	0.56	0.66	0.76	0.86	1
$F_0(a_i;\widehat{\mu})$	0.2720	0.4700	0.6142	0.7191	0.7955	0.8511	0.9582
$\lvert\bar{F}_n(a_i)-F_0(a_i;\widehat{\mu})\rvert$	0.0120	0.0100	0.0542	0.0591	0.0355	0.0089	0.0418

Aşadar avem că $\sqrt{n}\,d_n = \sqrt{50}\cdot 0.0591 = 0.418 < 1.36 = k_{1-\alpha}$, deci exponenţialitatea lui X este admisă.

Programul 6.12.7. Vom scrie un program Matlab, care efectuează calculele din exemplul precedent:

```
% Testul chi^2
clear
a=[0:30:180,300]; f=[13,10,5*ones(1,4),7];
mij=[15:30:165,240]; mu=sum(mij.*f)/50;
F0=expcdf(a,mu);F0(end)=1; p0=diff(F0);
x2=sum((f-50*p0).^2./(50*p0));
cuant=chi2inv(0.95,5);
fprintf(' x^2=  %6.4f\n',x2)
fprintf(' cuant= %4.2f\n',cuant)
% Testul lui Kolmogorov
F0=F0(2:end);F0(end)=expcdf(a(end),mu);
Fn=cumsum(f)/50;
dn=max(abs(F0-Fn)); ks=sqrt(50)*dn;
fprintf(' ks=   %6.4f',ks)
```

În urma executării programului se obţine:

```
 x^2=  2.8770
 cuant= 11.07
 ks=   0.4181
```

6.12.1 Funcţia `kstest`

Statistics toolbox conţine funcţia `kstest`, care efectuează testul lui Kolmogorov. Apelul funcţiei se poate face cu una din instrucţiunile:

```
h=kstest(x)
h=kstest(x,cdf)
h=kstest(x,cdf,alpha)
h=kstest(x,cdf,alpha,tail)
[h,c,ks]=kstest(x,cdf,alpha,tail)
```

Comenzile de acest tip lansează execuţia testului lui Kolmogorov bilateral (când `tail=0`, valoare implicită), unilateral dreapta (când `tail=1`), respectiv unilateral stânga (când `tail=-1`), prin considerarea datelor conţinute de vectorul `x`, iar funcţia

de repartiţie ipotezată fiind precizată prin prametrul `cdf`. Implicit se consideră `cdf` corespunzând legii normale standard.

Dacă parametrul `cdf` este prezent, trebuie să fie o matrice cu două coloane, care conţine în coloana a doua valorile funcţiei de repartiţie ipotezate pe punctele precizate în prima coloană. Ar fi de dorit ca prima coloană a matricei `cdf` să coincidă cu vectorul `x`, altfel funcţia va efectua un proces de interpolare pentru calculul valorilor funcţiei ipotezate pe punctele lui `x`, folosind punctele precizate prin matricea `cdf`. Din acest motiv, se impune ca toate componentele lui `x` să aibă valorile conţinute în intervalul determinat de valoare minimă şi valoarea maximă a componentelor primei coloane a lui `cdf`.

Parametrul `alpha` (implicit `alpha=0.05`) reprezintă nivelul de semnificaţie.

Rezultatele obţinute au următoarele semnificaţii. Dacă `h=1` ipoteza nulă se respinge, ceea ce corespunde faptului că valoarea critică `c` satisface relaţia $\texttt{c} \leqslant \texttt{alpha}$, iar dacă `h=0`, ipoteza nulă nu poate fi respinsă. Mai remarcăm faptul că `ks` va conţine respectiv valoarea calculată a statisticilor D_n, D_n^+ şi D_n^-, şi verifică relaţiile $\texttt{c} = 1 - K\left(\sqrt{\texttt{n}} d_n\right)$, pentru testul bilateral, respectiv $\texttt{c} = 1 - K^{\pm}\left(\sqrt{\texttt{n}} d_n^+\right)$ şi, $\texttt{c} = 1 - K^{\pm}\left(\sqrt{\texttt{n}} d_n^-\right)$ pentru testele unilaterale.

Programul 6.12.8. Să aplicăm testele lui Kolmogorov (bilateral, unilateral dreapta şi unilateral stânga) pentru datele `x=-2:1:4`, considerând legea ipotezată ca fiind legea normală standard şi nivelul de semnificaţie `alpha=0.05`. Programul ce urmează, să reprezinte grafic şi funcţia de repartiţie de selecţie, împreună cu funcţia de repartiţie ipotezată.

```
x=-2:4;
fprintf(' h     c      ks\n')
fprintf('_____________________\n')
for i=-1:1
    [h,c,ks]=kstest(x,[],0.05,i);
    fprintf(' %d  %5.4f   %4.2f  \n',h,c,ks)
end
cdfplot(x), hold on, title([])
t=-3:0.01:3; plot(t,normcdf(t,0,1),'k--')
```

În urma executării programului, se obţin rezultatele:

```
 h     c      ks
_____________________
 0  0.0682   0.41
 0  0.1363   0.41
 0  0.7753   0.13
```

respectiv graficele din Figura 6.4 Se observă, în mod surprinzător, ipoteza nulă nu poate fi respinsă. Acest lucru se datorează faptului că testul lui Kolmogorov este bazat pe un rezultat asimptotic, adică se aplică pentru valori mari ale volumului selecţiei.

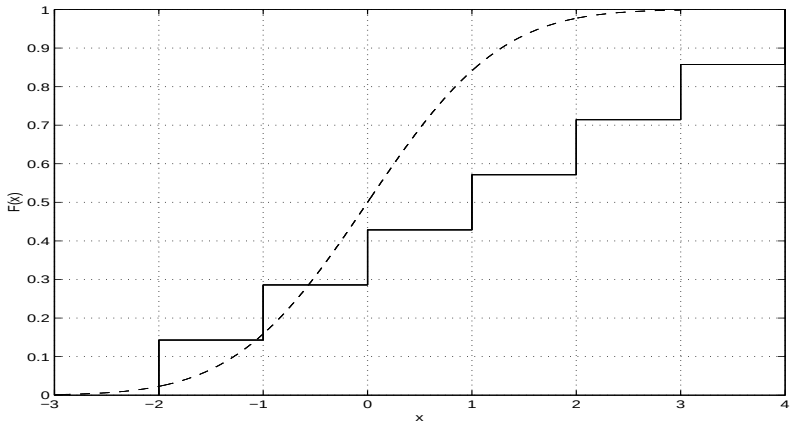

Figura 6.4: Testul lui Kolmogorov

6.12.2 Funcţia `lillietest`

Funcţia `lillietest` efectuează testul Lilliefors, care reprezintă o modificare a testului lui Kolmogorov, în care ipoteza nulă presupune că funcţia de repartiţie este cea de la legea normală, dar parametrii nu mai sunt precizaţi, ci sunt estimaţi folosind datele de selecţie.

Apelul funcţiei se poate face cu una din comenzile

```
h=lillietest(x)
h=lillietest(x,alpha)
[h,c,ls]=lillietest(x,alpha)
```

Comenzile de acest tip lansează execuţia testului Lilliefors, prin considerarea datelor conţinute de vectorul `x`, iar funcţia de repartiţie ipotezată fiind cea de la legea normală $\mathcal{N}(\bar{x}, \bar{\sigma})$.

Parametrul `alpha` (implicit `alpha=0.05`) reprezintă nivelul de semnificaţie.

Rezultatele obţinute au următoarele semnificaţii. Dacă `h=1` ipoteza nulă se respinge, ceea ce corespunde faptului că valoarea critică `c` satisface relaţia `c`$\leqslant$`alpha`, iar dacă `h=0`, ipoteza nulă nu poate fi respinsă.

Mai remarcăm faptul că `ls` va conţine respectiv valoarea calculată a statisticii D_n, iar `c` este *valoarea critic*ă, care se calculează cu ajutorul lui `ls`, folosindu–se tabelele construite de Lilliefors. Dacă în aceste tabele nu se află valoarea corespunzătoare, atunci `c=NaN`, dar `h` indică încă dacă ipoteza nulă este respinsă sau nu.

Programul 6.12.9. Să aplicăm testul lui Lilliefors pentru `n` numere aleatoare ce urmează legea normală $\mathcal{N}(\mu, \sigma)$, conţinute de vectorul `x`, cu nivelul de semnificaţie

`alpha=0.05`. Programul ce urmează să reprezinte grafic şi funcţia de repartiţie de selecţie, împreună cu funcţia de repartiţie ipotezată, adică cea de la legea normală $\mathcal{N}(\bar{x}, \bar{\sigma})$.

```
clf
n=input('n='); mu=input('mu='); s=input('sigma=');
x=normrnd(mu,s,1,n);
[h,c,ls]=lillietest(x);
fprintf( ' h=%d\n c= %5.4f\n ls=%5.4f',h,c,ls)
ma=mean(x); va=sqrt(var(x));
cdfplot(x), hold on, title([])
t=ma-3*va:0.01:ma+3*va;
plot(t,normcdf(t,ma,va),'k--')
```

În urma executării programului pentru `n=10`, `mu=100` şi `sigma=2`, se obţin rezultatele:

```
h=0
c= 0.1433
ls=0.1093
```

şi graficele din Figura 6.5

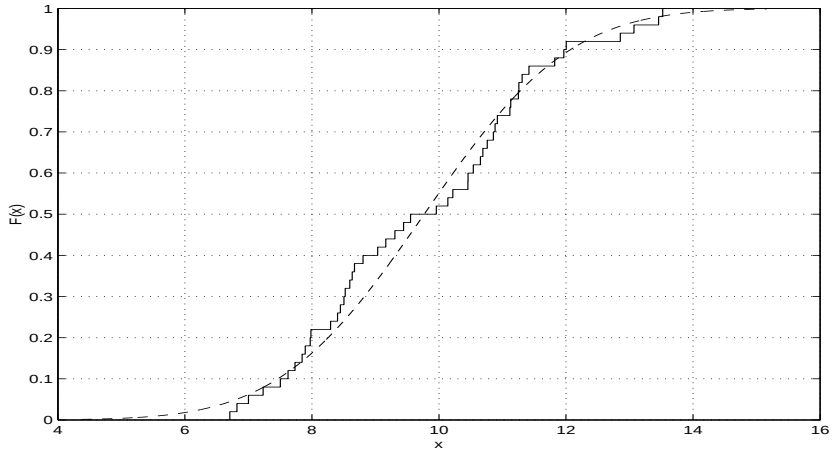

Figura 6.5: Testul Lilliefors

6.13 Testul Kolmogorov–Smirnov

Se consideră două caracteristici independente X şi Y de tip continuu cu funcţiile teoretice de repartiţie F_X şi F_Y necunoscute. Privind cele două caracteristici vrem să verificăm dacă sunt identic repartizate. Astfel, considerăm ipoteza nulă dată prin $H_0 : F_X = F_Y$, cu una din alternativele:

(1) $H_1 : F_X \neq F_Y$ (*testul Kolmogorov–Smirnov bilateral*),

(2) $H_1 : F_X > F_Y$ (*testul Kolmogorov–Smirnov unilateral dreapta*),

(3) $H_1 : F_X < F_Y$ (*testul Kolmogorov–Smirnov unilateral stânga*).

Pentru verificarea ipotezei nule H_0, cu una din alternativele precizate, se consideră câte o selecţie repetată de volum n_1 şi respectiv n_2, cu datele de selecţie $x_1, x_2, \ldots, x_{n_1}$ şi variabilele de selecţie $X_1, X_2, \ldots, X_{n_1}$, corespunzătoare pentru caracteristica X şi datele de selecţie $y_1, y_2, \ldots, y_{n_2}$ şi variabilele de selecţie $Y_1, Y_2, \ldots, Y_{n_2}$, corespunzătoare pentru caracteristica Y.

Se consideră statisticile

$$\begin{aligned} D_{n_1 n_2} &= \sup_{x \in \mathbb{R}} \left\{ \mid \bar{F}_X(x) - \bar{F}_Y(x) \mid \right\}, \\ D^+_{n_1 n_2} &= \sup_{x \in \mathbb{R}} \left\{ \bar{F}_X(x) - \bar{F}_Y(x) \right\}, \\ D^-_{n_1 n_2} &= \sup_{x \in \mathbb{R}} \left\{ \bar{F}_Y(x) - \bar{F}_X(x) \right\}, \end{aligned}$$

unde $\bar{F}_X$ şi $\bar{F}_Y$ sunt funcţiile de repartiţie de selecţie.

Pentru $n_1, n_2 \to \infty$, funcţiile de repartiţie ale acestor funcţii de selecţie au următoarele comportări asimptotice:

$$\lim_{n_1, n_2 \to \infty} P\left(\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\, D_{n_1 n_2} \leqslant x \right) = K(x) = \sum_{k=-\infty}^{k=+\infty} (-1)^k \, \mathrm{e}^{-2k^2 x^2}, \quad x > 0,$$

funcţia lui Kolmogorov $K(x)$, fiind tabelată pentru anumite valori în *Anexa V*, respectiv

$$\begin{aligned} \lim_{n_1, n_2 \to \infty} P\left(\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\, D^+_{n_1 n_2} \leqslant x \right) &= \lim_{n_1, n_2 \to \infty} P\left(\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\, D^-_{n_1 n_2} \leqslant x \right) \\ &= K^{\pm}(x) = 1 - \mathrm{e}^{-2x^2}, \quad x > 0, \end{aligned}$$

ce poartă numele de lege χ cu două grade de libertate.

Pentru un nivel de semnificaţie $\alpha \in (0,1)$ fixat, se pot determina cuantilele $k_{1-\alpha}$, $k^+_{1-\alpha}$ şi $k^-_{1-\alpha}$ astfel încât

$$P\left(\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\, D_{n_1 n_2} \leqslant k_{1-\alpha} \;\middle|\; H_0 \right) = 1 - \alpha, \quad \text{adică} \quad K(k_{1-\alpha}) = 1 - \alpha,$$

pentru testul bilateral Kolmogorov–Smirnov,

$$P\left(\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\, D^+_{n_1 n_2} \leqslant k^+_{1-\alpha} \;\middle|\; H_0 \right) = 1 - \alpha, \text{ adică } K^{\pm}\left(k^+_{1-\alpha}\right) = 1 - \alpha,$$

pentru testul unilateral dreapta Kolmogorov–Smirnov,

$$P\left(\sqrt{\frac{n_1n_2}{n_1+n_2}}\,D^-_{n_1n_2}\leqslant k^-_{1-\alpha}\;\middle|\;H_0\right)=1-\alpha,\ \text{adică}\ K^{\pm}\left(k^-_{1-\alpha}\right)=1-\alpha,$$

pentru testul unilateral stânga Kolmogorov–Smirnov,

Corespunzător celor trei alterantive, ipoteza H_0 va fi admisă când valorile calculate $d_{n_1n_2}$, $d^+_{n_1n_2}$, $d^-_{n_1n_2}$ pe baza datelor de selecţie $x_1,\ldots,x_{n_1}$ şi $y_1,\ldots,y_{n_2}$, ale statisticilor $D_{n_1n_2}$, $D^+_{n_1n_2}$, satisfac respectiv condiţiile

(1) $\sqrt{\frac{n_1n_2}{n_1+n_2}}\,d_{n_1n_2}<k_{1-\alpha}$, când $H_1:F_X\neq F_Y$,

(2) $\sqrt{\frac{n_1n_2}{n_1+n_2}}\,d^+_{n_1n_2}<k^+_{1-\alpha}$, când $H_1:F_X>F_Y$,

(3) $\sqrt{\frac{n_1n_2}{n_1+n_2}}\,d^-_{n_1n_2}<k^-_{1-\alpha}$, când $H_1:F_X<F_Y$,

iar în caz contrar va fi respinsă.

Prezentăm, în cele ce urmează, împreună cele trei teste (bilateral, unilateral dreapta şi unilateral stânga) Kolmogorov–Smirnov. Remarcăm că cele trei teste sunt teste distincte, alegerea unuia dintre ele se face apriori.

Etapele aplicării testului lui Kolmogorov–Smirnov

1. Se dau: α; $x_1,x_2,\ldots,x_{n_1}$; $y_1,y_2,\ldots,y_{n_2}$;
2. Se determină
   (1) $k_{1-\alpha}$ astfel încât $K\left(k_{1-\alpha}\right)=1-\alpha$, când $H_1:F_X\neq F_Y$,
   (2) $k^+_{1-\alpha}$ astfel încât $K^{\pm}\left(k^+_{1-\alpha}\right)=1-\alpha$, când $H_1:F_X>F_Y$;
   (3) $k^-_{1-\alpha}$ astfel încât $K^{\pm}\left(k^-_{1-\alpha}\right)=1-\alpha$, când $H_1:F_X<F_Y$;
3. Se calculează
   (1) $k=\sqrt{\frac{n_1n_2}{n_1+n_2}}\,d_{n_1n_2}$, când $H_1:F_X\neq F_Y$,
   (2) $k=\sqrt{\frac{n_1n_2}{n_1+n_2}}\,d^+_{n_1n_2}$, când $H_1:F_X>F_Y$;
   (3) $k=\sqrt{\frac{n_1n_2}{n_1+n_2}}\,d^-_{n_1n_2}$, când $H_1:F_X<F_Y$;
4. Concluzia: dacă

(1) $k < k_{1-\alpha}$ ipoteza H_0 este admisă, când $H_1 : F_X \neq F_Y$,

(2) $k < k^+_{1-\alpha}$ ipoteza H_0 este admisă, când $H_1 : F_X > F_Y$,

(3) $k < k^-_{1-\alpha}$ ipoteza H_0 este admisă, când $H_1 : F_X < F_Y$,

Observaţia 6.13.1. Având în vedere că funcţiile (de repartiţie) K şi $K^{\pm}$ sunt monoton crescătoare şi că pentru determinarea cuantilelor $k_{1-\alpha}$, $k^+_{1-\alpha}$ şi $k^-_{1-\alpha}$ este necesară inversarea acestor funcţii de repartiţie, se poate evita această operaţie. Pentru aceasta se determină valorile funcţiilor K şi $K^{\pm}$ pe valorile calculate ale statisticilor $D_{n_1n_2}$, $D^+_{n_1n_2}$, $D^-_{n_1n_2}$, adică se calculează, respectiv $1-c = K\left(\sqrt{\frac{n_1n_2}{n_1+n_2}}d_{n_1n_2}\right)$, $1-c = K^{\pm}\left(\sqrt{\frac{n_1n_2}{n_1+n_2}}d^+_{n_1n_2}\right)$, $1-c = K^{\pm}\left(\sqrt{\frac{n_1n_2}{n_1+n_2}}d^-_{n_1n_2}\right)$. Astfel, se va respinge ipoteza nulă H_0 dacă *valoarea critică* c satisface inegalitatea $c \leqslant \alpha$.

6.13.1 Funcţia `kstest2`

Statistics toolbox conţine funcţia `kstest2`, care efectuează testul Kolmogorov–Smirnov. Apelul funcţiei se poate face cu una din instrucţiunile:

```
h=kstest(x,y)
h=kstest(x,y,alpha)
h=kstest(x,y,alpha,tail)
[h,c,ks]=kstest(x,y,alpha,tail)
```

Comenzile de acest tip lansează execuţia testului Kolmogorov–Smirnov bilateral (când `tail=0`, valoare implicită), unilateral dreapta (când `tail=1`), respectiv unilateral stânga (când `tail=-1`), prin considerarea datelor conţinute în vectorii `x` şi `y`.

Parametrul `alpha` (implicit `alpha=0.05`) reprezintă nivelul de semnificaţie.

Rezultatele obţinute au următoarele semnificaţii. Dacă `h=1`, ipoteza nulă se respinge, ceea ce corespunde faptului că valoarea critică `c` satisface relaţia `c`$\leqslant$`alpha`, iar dacă `h=0`, ipoteza nulă nu poate fi respinsă. Mai remarcăm faptul că `ks` va conţine respectiv valoarea calculată a statisticilor D_{n_1,n_2}, $D^+_{n_1,n_2}$ şi $D^-_{n_1,n_2}$, şi verifică relaţiile `c` $= 1 - K\,(\sqrt{\mathtt{n}}d_n)$, pentru testul bilateral, respectiv `c` $= 1 - K^{\pm}\,(\sqrt{\mathtt{n}}d^+_n)$ şi `c` $= 1 - K^{\pm}\,(\sqrt{\mathtt{n}}d^-_n)$, pentru testele unilaterale.

Programul 6.13.2. Să aplicăm testele Kolmogorov–Smirnov (bilateral, unilateral dreapta şi unilateral stânga) pentru datele `x=-1:1:5` şi `y` reprezentând `20` de numere aleatoare, ce urmează legea normală standard, iar nivelul de semnificaţie să fie `alpha=0.05`. Programul ce urmează, să reprezinte grafic şi cele două funcţii de repartiţie de selecţie.

```
clf
x=-1:5; y=randn(1,20);
fprintf(' h     c      ks\n')
fprintf('_________________\n')
for i=-1:1
```

```
    [h,c,ks]=kstest2(x,y,[],i);
    fprintf(' %d  %5.4f   %4.2f\n',h,c,ks)
end
h1=cdfplot(x); hold on, h2=cdfplot(y);
title('Functiile de repartitie de selectie')
set(h1,'linestyle','--')
set(h2,'linestyle','-')
```

În urma executării programului, se obţin rezultatele:

```
 h     c      ks
_____________________
 1  0.0110   0.61
 1  0.0219   0.61
 0  0.9783   0.04
```

respectiv graficele din Figura 6.6. Testul Kolmogorov–Smirnov este bazat pe un re-

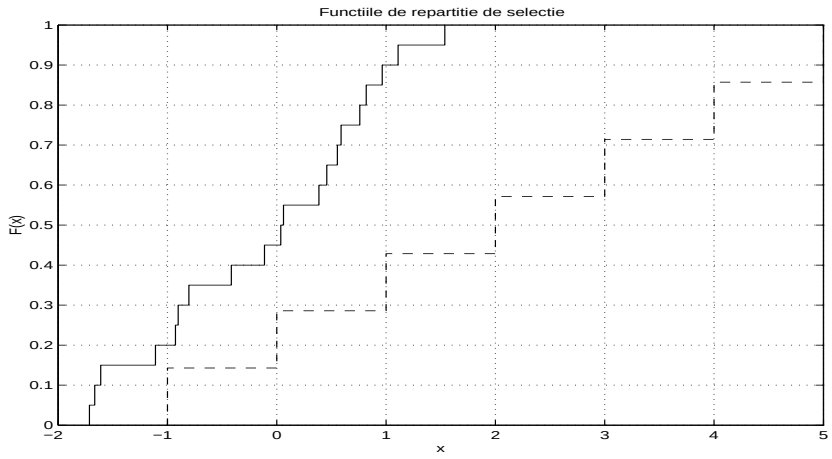

Figura 6.6: Testul Kolmogorov–Smirnov

zultat asimptotic, adică se aplică pentru valori mari ale volului selecţiei, dar aici s–au folosit selecţii de volum mic, în mod special pentru a urmări pe figură modul de calcul a celor trei statistici d_{n_1,n_2}, $d^+_{n_1,n_2}$ şi $d^-_{n_1,n_2}$.

Anexa I (Funcţia lui Laplace) $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x \mathrm{e}^{-\frac{t^2}{2}} dt,\ \Phi(-x) = -\Phi(x)$

	Sutimi ale lui x									
x	0	1	2	3	4	5	6	7	8	9
0.0	0.0000	0040	0080	0120	0160	0199	0239	0279	0319	0359
0.1	0.0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0.2	0.0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0.3	0.1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0.4	0.1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0.5	0.1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0.6	0.2257	2291	2324	2357	2389	2422	2454	2486	2517	2549
0.7	0.2580	2611	2642	2673	2703	2734	2764	2794	2823	2852
0.8	0.2881	2910	2939	2967	2995	3023	3051	3078	3106	3137
0.9	0.3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1.0	0.3413	3437	3461	3485	3508	3531	3554	3577	3599	3621
1.1	0.3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1.2	0.3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1.3	0.4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1.4	0.4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1.5	0.4332	4345	4357	4370	4382	4394	4406	4418	4429	4441
1.6	0.4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1.7	0.4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1.8	0.4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1.9	0.4713	4719	4726	4732	4738	4744	4750	4756	4761	4767
2.0	0.4772	4778	4783	4788	4793	4798	4803	4808	4812	4817
2.1	0.4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2.2	0.4861	4864	4868	4871	4875	4878	4881	4884	4887	4890
2.3	0.4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2.4	0.4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2.5	0.4938	4931	4941	4943	4945	4946	4948	4949	4951	4952
2.6	0.4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2.7	0.4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2.8	0.4974	4075	4976	4977	4977	4978	4979	4980	4980	4981
2.9	0.4981	4982	4982	4983	4984	4984	4985	4985	4986	4986
3.0	0.4986	4987	4987	4988	4988	4989	4989	4989	4990	4990
3.1	0.4990	4991	4991	4991	4992	4992	4992	4992	4993	4993
3.2	0.4993	4993	4994	4994	4994	4994	4994	4995	4995	4995
3.3	0.4995	4995	4996	4996	4996	4996	4996	4996	4996	4996
3.4	0.4997	4997	4997	4997	4997	4997	4997	4997	4998	4998
3.5	0.4998	4998	4998	4998	4998	4998	4998	4998	4998	4998
3.6	0.4998	4998	4999	4999	4999	4999	4999	4999	4999	4999

Anexa II (Legea Student) $F_n\left(t_{n;\gamma}\right) = \dfrac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \int_{-\infty}^{t_{n;\gamma}} \left(1+\dfrac{x^2}{n}\right)^{-\frac{n+1}{2}} dx = \gamma$

γ	0.70	0.75	0.80	0.85	0.90	0.95	0.975	0.99	0.995
n	$t_{n;\gamma}$								
1	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	0.617	0.817	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.584	0.765	0.979	1.250	1.638	2.353	3.182	4.541	5.841
4	0.569	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707
7	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.500
8	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.897	3.355
9	0.544	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10	0.542	0.700	0.879	1.093	1.372	1.813	2.228	2.764	3.169
11	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12	0.539	0.696	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13	0.538	0.694	0.870	1.080	1.350	1.771	2.160	2.650	3.012
14	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.625	2.977
15	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.603	2.947
16	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.584	2.921
17	0.534	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898
18	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878
19	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.540	2.861
20	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845
21	0.533	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831
22	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23	0.532	0.685	0.858	1.060	1.320	1.714	2.069	2.500	2.807
24	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787
26	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.530	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750
35	0.529	0.682	0.852	1.052	1.306	1.690	2.030	2.438	2.724
40	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.705
50	0.528	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678
60	0.527	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660
80	0.526	0.679	0.847	1.044	1.291	1.671	2.000	2.390	2.660
120	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617
∞	0.524	0.674	0.842	1.036	1.282	1.645	1.956	2.326	2.576

Anexa III (Legea χ^2) $$F_n\left(\chi^2_{n;\gamma}\right) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_0^{\chi^2_\gamma} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} dx = \gamma$$

γ	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995
n	$\chi^2_{n;\gamma}$									
1	0.000	0.000	0.001	0.004	0.016	2.71	3.84	5.02	6.63	7.88
2	0.010	0.020	0.051	0.103	0.211	4.61	5.99	7.38	9.21	10.60
3	0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.06	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.1	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.9	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.3	11.6	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	11.0	12.3	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.2	11.7	13.1	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.9	12.4	13.8	15.66	33.20	36.42	39.36	42.98	45.56
25	10.5	11.5	13.1	14.6	16.47	34.38	37.65	40.65	44.31	46.93
26	11.2	12.2	13.8	15.4	17.29	35.56	38.89	41.92	45.64	48.29
27	11.8	12.9	14.6	16.2	18.11	36.74	40.11	43.19	46.96	49.64
28	12.5	13.6	15.3	16.9	18.94	37.92	41.34	44.46	48.28	50.99
29	13.1	14.3	16.0	17.7	19.77	39.09	42.56	45.72	49.59	52.34
30	13.8	15.0	16.8	18.5	20.60	40.26	43.77	46.98	50.89	53.67
35	17.2	18.5	20.6	22.5	24.8	46.1	49.8	53.2	57.3	60.3
40	20.7	22.2	24.4	26.5	29.1	51.8	55.8	59.3	63.7	6.68
60	35.5	37.5	40.5	43.2	46.5	74.4	79.1	83.3	88.4	91.9

Anexa IV (Legea Fisher–Snedecor)

$$F_{m,n}\left(f_{m,n;\gamma}\right)=\left(\frac{m}{n}\right)^{\frac{m}{2}}\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\,\Gamma\left(\frac{n}{2}\right)}\int_0^{f_{m,n;\gamma}}x^{\frac{m}{2}-1}\left(1+\frac{m}{n}\,x\right)^{-\frac{m+n}{2}}dx=\gamma$$

	m	1	2	3	4	5	6	7	8
γ	n	$f_{m,n;\gamma}$							
0.95		161.4	199.5	216	225	230	234	237	239
0.975	1	648	800	864	900	922	937	948	957
0.99		4052	4999	5403	5625	5764	5859	5930	5981
0.95		18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37
0.975	2	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4
0.99		98.49	99.00	99.17	99.25	99.30	99.33	99.35	99.36
0.95		10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85
0.975	3	17.4	16.0	15.1	15.4	14.9	14.7	14.6	14.5
0.99		34.12	30.84	29.46	28.71	28.24	27.91	27.7	27.49
0.95		17.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04
0.975	4	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98
0.99		21.20	18.00	16.69	15.98	15.52	15.21	15.0	14.80
0.95		6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82
0.975	5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76
0.99		16.26	13.27	12.06	11.39	10.97	10.67	10.5	10.27
0.95		5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15
0.975	6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60
0.99		13.74	10.91	9.78	9.15	8.75	8.47	8.26	8.10
0.95		5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73
0.975	7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90
0.99		12.25	9.55	8.45	7.85	7.45	7.19	6.99	6.84
0.95		5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44
0.975	8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43
0.99		11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03
0.95		5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23
0.975	9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10
0.99		10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47
0.95		4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07
0.975	10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85
0.99		10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06
0.95		4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95
0.975	11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66
0.99		9.65	7.20	6.22	5.67	5.32	5.07	4.89	4.74

Anexa IV (continuare)

	m	1	2	3	4	5	6	7	8
γ	n	$f_{m,n;\gamma}$							
0.95		4.75	3.88	3.49	3.26	3.11	3.00	2.91	2.85
0.975	12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51
0.99		9.33	6.93	5.95	5.41	5.06	4.82	4.54	4.50
0.95		4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77
0.975	13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39
0.99		9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30
0.95		4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70
0.975	14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29
0.99		8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14
0.95		4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64
0.975	15	6.20	4.76	4.15	3.80	3.58	3.41	3.29	3.20
0.99		8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00
0.95		4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59
0.975	16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12
0.99		8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89
0.95		4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55
0.975	17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06
0.99		8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79
0.95		4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51
0.975	18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01
0.99		8.29	7.01	5.09	4.58	4.25	4.01	3.84	3.71
0.95		4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48
0.975	19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96
0.99		8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63
0.95		4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45
0.975	20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91
0.99		8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56
0.95		4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42
0.975	21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87
0.99		8.02	5.78	4.87	4.37	4.04	3.84	3.64	3.51
0.95		4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40
0.975	22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84
0.99		7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45
0.95		4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37
0.975	23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81
0.99		7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41
0.95		4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36
0.975	24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78
0.99		7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36

Anexa IV (continuare)

γ	m / n	9	10	11	12	13	14	15	16
		$f_{m,n;\gamma}$							
0.95		19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4
0.975	2	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4
0.99		99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4
0.95		8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.69
0.975	3	14.5	14.4	14.4	14.3	14.3	14.3	14.3	14.2
0.99		27.3	27.1	27.1	27.1	27.0	26.9	26.9	26.8
0.95		6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.84
0.975	4	8.90	8.84	8.79	8.75	8.72	8.69	8.66	8.64
0.99		14.7	14.5	14.4	14.4	14.3	14.2	14.2	14.2
0.95		4.77	4.74	4.70	4.68	4.66	4.64	4.62	4.60
0.975	5	6.68	6.62	6.57	6.52	6.49	6.46	6.43	6.41
0.99		10.2	10.1	9.96	9.89	9.82	9.77	9.72	9.68
0.95		4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.92
0.975	6	5.52	5.46	5.41	5.37	5.33	5.30	5.27	5.25
0.99		7.98	7.89	7.79	7.72	7.66	7.60	7.56	7.52
0.95		3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.49
0.975	7	4.82	4.76	4.71	4.67	4.63	4.60	4.57	4.54
0.99		6.72	6.62	6.54	6.47	6.41	6.36	6.31	6.27
0.95		3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.20
0.975	8	4.36	4.30	4.24	4.20	4.16	4.13	4.10	4.08
0.99		5.91	5.81	5.73	5.67	5.61	5.56	5.52	5.48
0.95		3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.99
0.975	9	4.03	3.96	3.91	3.87	3.83	3.80	3.77	3.74
0.99		5.35	5.26	5.18	5.11	5.05	5.00	4.96	4.92
0.95		3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.83
0.975	10	3.78	3.72	3.66	3.62	3.58	3.55	3.52	3.50
0.99		4.94	4.85	4.77	4.71	4.65	4.60	4.56	4.52
0.95		2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.70
0.975	11	3.59	3.53	3.47	3.43	3.39	3.36	3.33	3.30
0.99		4.63	4.54	4.46	4.40	4.34	4.29	4.25	4.21
0.95		2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.60
0.975	12	3.44	3.37	3.32	3.28	3.24	3.21	3.18	3.15
0.99		4.39	4.30	4.22	4.16	4.10	4.05	4.01	3.97
0.95		2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.51
0.975	13	3.31	3.25	3.20	3.15	3.12	3.08	3.05	3.03
0.99		4.19	4.10	4.02	3.96	3.91	3.86	3.82	3.78
0.95		2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.44
0.975	14	3.21	3.15	3.09	3.05	3.01	2.98	2.95	2.92
0.99		4.03	3.94	3.86	3.80	3.75	3.70	3.66	3.62

Anexa IV (continuare)

	m	9	10	11	12	13	14	15	16
γ	n	$f_{m,n;\gamma}$							
0.95		2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.38
0.975	15	3.12	3.06	3.01	2.96	2.92	2.89	2.86	2.84
0.99		3.89	3.80	3.73	3.67	3.61	3.56	3.52	3.49
0.95		2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.33
0.975	16	3.05	2.99	2.93	2.89	2.85	2.82	2.79	2.76
0.99		3.78	3.69	3.62	3.55	3.50	3.45	3.41	3.37
0.95		2.49	2.45	2.41	2.38	2.35	2.33	2.29	2.27
0.975	17	2.98	2.92	2.87	2.82	2.79	2.75	2.72	2.70
0.99		3.68	3.59	3.52	3.46	3.40	3.35	3.31	3.27
0.95		2.46	2.41	2.37	2.34	2.31	2.29	2.27	2.25
0.975	18	2.93	2.87	2.81	2.77	2.73	2.70	2.67	2.64
0.99		3.60	3.51	3.43	3.37	3.32	3.27	3.23	3.19
0.95		2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.21
0.975	19	2.88	2.82	2.76	2.72	2.68	2.65	2.62	2.59
0.99		3.52	3.43	3.36	3.30	3.24	3.19	3.15	3.12
0.95		2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.19
0.975	20	2.84	2.77	2.72	2.68	2.64	2.60	2.57	2.55
0.99		3.46	3.37	3.29	3.23	3.18	3.13	3.09	3.05
0.95		2.37	2.32	2.28	2.25	2.22	2.20	2.18	2.16
0.975	21	2.80	2.73	2.68	2.64	2.60	2.56	2.53	2.51
0.99		3.40	3.31	3.24	3.17	3.12	3.07	3.03	2.99
0.95		2.34	2.30	2.26	2.23	2.20	2.17	2.15	2.13
0.975	22	2.76	2.70	2.65	2.60	2.56	2.53	2.50	2.47
0.99		3.35	3.26	3.18	3.12	3.07	3.02	2.98	2.94
0.95		2.32	2.27	2.23	2.20	2.18	2.15	2.13	2.11
0.975	23	2.73	2.67	2.62	2.57	2.53	2.50	2.47	2.44
0.99		3.30	3.21	3.14	3.07	3.02	2.97	2.93	2.89
0.95		2.30	2.25	2.21	2.18	2.15	2.13	2.11	2.09
0.975	24	2.70	2.64	2.59	2.54	2.50	2.47	2.44	2.41
0.99		3.26	3.17	3.09	3.03	2.98	2.93	2.89	2.85

Anexa V (Funcţia lui Kolmogorov) $K(x) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2x^2}$

x	Sutimi ale lui x									
	0	1	2	3	4	5	6	7	8	9
0.3	0.0000	0000	0000	0001	0002	0003	0005	0008	0013	0019
0.4	0.0028	0039	0054	0073	0097	0125	0160	0200	0246	0300
0.5	0.0360	0428	0503	0585	0674	0771	0875	0986	1103	1227
0.6	0.1357	1492	1632	1777	1926	2079	2236	2395	2557	2721
0.7	0.2887	3054	3222	3391	3559	3727	3896	4063	4230	4395
0.8	0.4558	4720	4880	5038	5193	5346	5497	5645	5790	5933
0.9	0.6072	6209	6342	6473	6600	6725	6896	6964	7079	7191
1.0	0.7300	7405	7508	7607	7704	7797	7888	7976	8061	8143
1.1	0.8222	8299	8373	8445	8513	8580	8644	8706	8765	8822
1.2	0.8877	8930	8981	9029	9076	9121	9164	9205	9245	9282
1.3	0.9319	9353	9386	9418	9448	9477	9505	9531	9556	9580
1.4	0.9603	9624	9645	9665	9683	9701	9718	9734	9749	9764
1.5	0.9777	9790	9803	9814	9825	9836	9846	9855	9864	9872
1.6	0.9880	9887	9894	9901	9907	9913	9919	9924	9929	9933
1.7	0.9938	9942	9946	9949	9953	9956	9959	9962	9964	9967
1.8	0.9969	9971	9973	9975	9977	9979	9980	9981	9983	9984
1.9	0.9985	9986	9987	9988	9989	9990	9991	9991	9992	9993
2.0	0.9993	9994	9994	9995	9995	9996	9996	9996	9996	9997
2.1	0.9997	9997	9998	9998	9998	9998	9998	9998	9999	9999

Bibliografie

[1] Abramowitz, M., Stegun, I. A., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, tenth ed., Dover Publications, Inc., New York, 1972.

[2] Blaga, P., *Calculul probabilităţilor. Culegere de probleme*, Universitatea "Babeş–Bolyai", Cluj–Napoca, 1984.

[3] _____, *Metode statistice în modelarea cu calculatorul. Lucrări de laborator*, Universitatea "Babeş–Bolyai", Cluj–Napoca, 1993.

[4] _____, *Calculul probabilităţilor şi statistică matematică. Vol. II. Curs şi culegere de probleme*, Universitatea "Babeş–Bolyai", Cluj–Napoca, 1994.

[5] _____, *Statistică matematică. Lucrări de laborator*, Universitatea "Babeş–Bolyai", Cluj–Napoca, 1999.

[6] _____, *Statistică matematică*, Universitatea "Babeş–Bolyai", Cluj–Napoca, 2000.

[7] _____, *Statistică matematică (ediţia II)*, Universitatea "Babeş–Bolyai", Cluj–Napoca, 2001.

[8] Blaga, P., Lupaş, A., Mureşan A. S., *Matematici aplicate, Vol.I–II*, Promedia Plus, Cluj–Napoca, 1999.

[9] Blaga, P., Mureşan A. S., *Matematici aplicate în economie, Vol.I–II*, Transilvania Press, Cluj–Napoca, 1996.

[10] Blaga, P., Rădulescu, M., *Calculul probabilităţilor*, Universitatea "Babeş–Bolyai", Cluj–Napoca, 1987.

[11] Ciucu, G., *Elemente de teoria probabilităţilor şi statistică matematică*, Editura didactică şi pedagogică, Bucureşti, 1963.

[12] Ciucu, G., Craiu, V., *Probleme de statistică matematică*, Editura didactică şi pedagogică, Bucureşti, 1968.

[13] _____, *Teoria estimaţiei şi verificarea ipotezelor statistice*, Editura didactică şi pedagogică, Bucureşti, 1968.

[14] _____, *Introducere în teoria probabilităţilor şi statistică matematică*, Editura didactică şi pedagogică, Bucureşti, 1971.

[15] _____, *Inferenţă statistică*, Editura didactică şi pedagogică, Bucureşti, 1974.

[16] Ciucu, G., Craiu, V., Ştefănescu, A., *Statistică matematică şi cercetări operaţionale, Vol. I*, Editura didactică şi pedagogică, Bucureşti, 1974.

[17] Ciucu, G., Tudor, C., *Probabilităţi şi procese stochastice. Vol. I*, Editura Academiei, Bucureşti, 1978.

[18] Craiu, V., *Verificarea ipotezelor statistice*, Editura ştiinţifică şi enciclopedică, Bucureşti, 1972.

[19] Craiu, V., Enache, R., Bâscă, O., *Teste de concordanţă cu programe în FORTRAN*, Editura ştiinţifică şi enciclopedică, Bucureşti, 1974.

[20] Cramér, H., *Mathematical methods of statistics*, Princeton University Press, Princeton, 1951.

[21] Dacunha–Castelle, D., Duflo, M., *Exercises de probabilités et statistiques. 1. Problèmes à temps fixe. Tome 1*, Masson, Paris, 1990.

[22] _____, *Probabilités et statistiques. 1. Problèmes à temps fixe. Tome 1*, Masson, Paris, 1990.

[23] Deák, I., *Random number generators and simulation*, Akadémiai Kiadó, Budapest, 1990.

[24] Dumitrescu, M., Florea, D., Tudor, C., *Probleme de teoria probabilităţilor şi statistică matematică*, Editura Tehnică, Bucureşti, 1985.

[25] Enders, W., *Applied econometric. Time series*, John Wiley & Sons, 1995.

[26] Ermakov, S. M., *Metoda Monte Carlo şi probleme înrudite*, Editura Tehnică, Bucureşti, 1976.

[27] Feller, W., *An introduction to probability theory and its applications. Vol. I*, John Wiley, New York, 1957.

[28] ———, *An introduction to probability theory and its applications. Vol. II*, John Wiley, New York, 1966.

[29] Gnedenko, B. V., *The theory of probability*, Mir Publishers, Moscow, 1976.

[30] Good, Ph. I., *Resampling Methods. A Practical Guide to Data Analysis*, Birkhäuser, Boston • Basel • Berlin, 1999.

[31] Griffiths, D. F., *An Introduction to Matlab, Version 2.1*, University of Dundee, 1997.

[32] Grinstead, Ch. M., Snell, J. L., *Introduction in probability, Second edition*, American Mathematical Society, 1997.

[33] Gurskiĭ, E. I., *Culegere de probleme pentru teoria probabilităţilor şi statistica matematică (l. rusă)*, Edit. Înv. superior Minsk, Minsk, 1984.

[34] Hoel, P. J., *Introduction to mathematical statistics, Fourth edition*, John Wiley & Sons, New York–London–Sydney–Toronto, 1971.

[35] ———, *Statistique mathématique, Tome II*, Armand Colin, Paris, 1991.

[36] Hoffmann–Jørgensen, J., *Probability with a view toward statistics, Vol. I-II*, Chapman & Hall, New York–London, 1994.

[37] Iosifescu, M., Mihoc, Gh., Theodorescu, R., *Teoria probabilităţilor şi statistica matematică*, Editura Tehnică, Bucureşti, 1966.

[38] Iosifescu, M., Moineanu, C., Trebici, V., Ursianu, E., *Mică enciclopedie de statistică*, Editura ştiinţifică şi enciclopedică, Bucureşti, 1985.

[39] Jansson, B., *Random number generators*, Victor Petterson Bokindustri Aktiebolag, Stockholm, 1966.

[40] Johnson, N. L., Kotz, S., *Distributions in statistics. Discrete distributions*, Houghton Mifflin Company, Boston, 1969.

[41] Kendall, M. G., Stuart, A., *The advanced theory of statistics. Vol. 1. Distribution theory, Second edition*, Charles Griffin & Company Limited, London, 1961.

[42] ———, *The advanced theory of statistics. Vol. 2. Inference and relationship*, Charles Griffin & Company Limited, London, 1963.

[43] ———, *The advanced theory of statistics. Vol. 3. Design and analysis, and time–series*, Charles Griffin & Company Limited, London, 1966.

[44] Knuth, D. E., *The art of computer programming, Vol. II. Seminumerical algorithms*, Addison-Wesley, Mass., 1969.

[45] Lebart, L., Morineau, M. G., Fénelon, J.-P., *Traitment des données statistiques. méthodes et programmes*, Dunod, Paris, 1982.

[46] Lecoutre, J.-P., Tassi, Ph., *Statistique non paramétrique et robustesse*, Economica, Paris, 1987.

[47] Lehmann, E. L., *Testing statistical hypotheses, Second edition*, Springer, New York–Berlin, 1997.

[48] Maliţa, M., Zidăroiu, C., *Incertitudine şi decizie*, Editura ştiinţifică şi enciclopedică, Bucureşti, 1980.

[49] Mihoc, Gh., Ciucu, G., Craiu, V., *Teoria probabilităţilor şi statistică matematică*, Editura didactică şi pedagogică, Bucureşti, 1970.

[50] Mihoc, Gh., Ciucu, G., Muja, A., *Modele matematice ale aşteptării*, Editura Academiei, Bucureşti, 1973.

[51] Mihoc, Gh., Craiu, V., *Tratat de statistică matematică. Vol. I. Selecţie şi estimaţie*, Editura Academiei, Bucureşti, 1976.

[52] ———, *Tratat de statistică matematică. Vol. II. Verificarea ipotezelor statistice*, Editura Academiei, Bucureşti, 1977.

[53] Mihoc, Gh., Firescu, D., *Statistică matematică*, Editura didactică şi pedagogică, Bucureşti, 1966.

[54] Mihoc, I., *Calculul probabilităţilor şi statistică matematică. Part. I, II*, lito. Univ. "Babeş–Bolyai", Cluj–Napoca, 1994,1995.

[55] Nakamura, S., *Numerical Analysis and Graphic Visualization with MATLAB*, Prentice Hall PTR, Upper Saddle River, New Jersey, 1996.

[56] Oancea, E., Rădulescu, M., *Calculul probabilităţilor şi statistică matematică*, lito. Univ. "Babeş–Bolyai", Cluj–Napoca, 1974.

[57] Ogden, R. T., *Essential wavelets for statistical applications and data analysis*, Birkhäuser, Boston • Basel • Berlin, 1997.

[58] Onicescu, O., *Numere şi sisteme aleatoare*, Editura Academiei, Bucureşti, 1962.

[59] Onicescu, O., Botez, M. C., *Incertitudine şi modelare economică (econometrie informaţională)*, Editura ştiinţifică şi enciclopedică, Bucureşti, 1985.

[60] Parzen, E., *Modern probability theory and its applications*, John Wiley, New York–London, 1960.

[61] Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T., *Numerical Recipes. The Art of Scientific Computing*, Cambridge University Press, Cambridge, 1986.

[62] Rancu, N., Tövissi, L., *Statistica matematică cu aplicaţii în producţie*, Editura Academiei, Bucureşti, 1963.

[63] Rao, M. M., *Probability theory with applications*, Academic Press, New York, 1984.

[64] Rényi, A., *Probability theory*, Akadémiai Kiadó, Budapest, 1970.

[65] Rumşiski, L. Z., *Prelucrarea matematică a datelor experimentale*, Editura Tehnică, Bucureşti, 1974.

[66] Saporta, G., *Probabilités, analyse des données et statistique*, Éditions Technip, Paris, 1990.

[67] Schervish, M. J., *Theory of statistics*, Springer, New York–Berlin, 1995.

[68] Shiryaev, A. N., *Probability*, Springer, New York–Berlin, 1995.

[69] Sigmon, K., *Matlab Primer, Third edition*, University of Florida, 1993.

[70] Snedecor, G. W, Cochran, W. G., *Statistical methods*, Iowa State University Press, 1989.

[71] Stapleton, J. H., *Linear statistical models*, John Wiley & Sons, New York–Chichester–Brisbane, 1995.

[72] Stark, H., Woods, J. W., *Probability, random processes, and estimation theory for engineers*, Prentice-Hall, New Jersey, 1986.

[73] Stoyanov, J., Mirazchiiski, I., Ignatov, Z., Tanushev, M., *Exercise manual in probability theory*, Kluwer Academic Publishers, Dordrecht–Boston–London, 1988.

[74] Săcuiu, I., Zorilescu, D., *Numere aleatoare. Aplicaţii în economie, industrie şi studiul fenomenelor naturale*, Editura Academiei, Bucureşti, 1978.

[75] Sveshnikov, A. A., *Problems in probability theory, mathematical statistics and theory of random functions*, W. B. Saunders Company, Philadelphia–London–Toronto, 1968.

[76] Tassi, Ph., *Méthodes statistiques*, 2^e *édition*, Economica, Paris, 1989.

[77] Văduva, I., *Analiză dispersională*, Editura Tehnică, Bucureşti, 1970.

[78] ———, *Modele de simulare cu calculatorul*, Editura Tehnică, Bucureşti, 1977.

[79] Wahba, G., *Spline models for observational data*, Society for Industrial and Applied Mathematics, Philadelphia, 1990.

[80] Yule, G. U., Kendall, M. G., *Introducere în teoria statisticii*, Editura ştiinţifică, Bucureşti, 1969.

Index