4. ANALIZA EXPLORATORIE A UNEI SERII DE DATE

4.1. INTRODUCERE

Analiza exploratorie a datelor (EDA – Exploratory Data Analysis) este o tehnică de abordare a prelucrării datelor de dată mai recentă, ce constă într-o colecţie de tehnici preponderent grafice, ce permit evidenţierea unor structuri în date. John Tukey a lansat această abordare în 1977.

Conform acestei metode, datele trebuie explorate fără a presupune apriori anumite modele statistice, distribuţii de erori, relaţii între variabile etc. Scopul principal este evidenţierea caracteristicilor datelor pentru ca analistul să înţeleagă cât mai bine procesul, să-l poată analiza şi modela. După depistarea caracteristicilor setului de date, modelarea datelor se poate face în condiţii mult mai bune.

Într-o analiză statistică clasică, etapele analizei sunt:
Problemă $\Rightarrow$ Date $\Rightarrow$ Model $\Rightarrow$ Analiză $\Rightarrow$ Concluzii.
În abordarea exploratorie a datelor, modelarea se face doar după depistarea principalelor caracteristici ale setului de date, moment în care modelul are şanse mult mai mari să fie unul corect. Etapele analizei exploratorii sunt:
Problemă $\Rightarrow$ Date $\Rightarrow$ Analiză $\Rightarrow$ Model $\Rightarrow$ Concluzii.

Tehnicile EDA şi cele clasice nu sunt mutual exclusive şi se recomandă utilizarea lor complementară. În faza iniţială a analizei se abordează tehnici EDA, care evidenţiază caracteristicile setului de date, pe baza acestora se stabileşte modelul corespunzător al datelor, iar validarea modelului se face cu metode cantitative: testarea ipotezelor, ANOVA etc.

Reducerea datelor experimentale la indicatori cantitativi, chiar dacă aceştia se bazează pe toate valorile cuprinse în setul de date, fără o reprezentare grafică prealabilă poate conduce la erori. Un exemplu clasic îl constituie următorul set de date [10]:

X: 10.00; 8.00; 13.00; 9.00; 11.00; 14.00; 6.00; 4.00; 12.00; 7.00; 5.00.
Y: 8.04; 6.95; 7.58; 8.81; 8.33; 9.96; 7.24; 4.26; 10.84; 4.82; 5.68.
În urma efectuării calculelor se obţin următoarele valori:

Volumul setului de date: N = 11
Media lui x: $\bar{x}$ = 9.0
Media lui y: $\bar{y}$ = 7.5
Coeficientul de corelaţie liniară: r = 0.816
Dreapta de regresie a lui y în raport cu x: y = x/2+3
Deviaţia standard a erorilor: σ_ε = 1.237

Aceste informaţii cantitative, au importanţa lor, dar nu oferă o caracterizare completă a datelor. Reprezentarea datelor raportate la dreapta de regresie se poate observa în fig. 4.1.

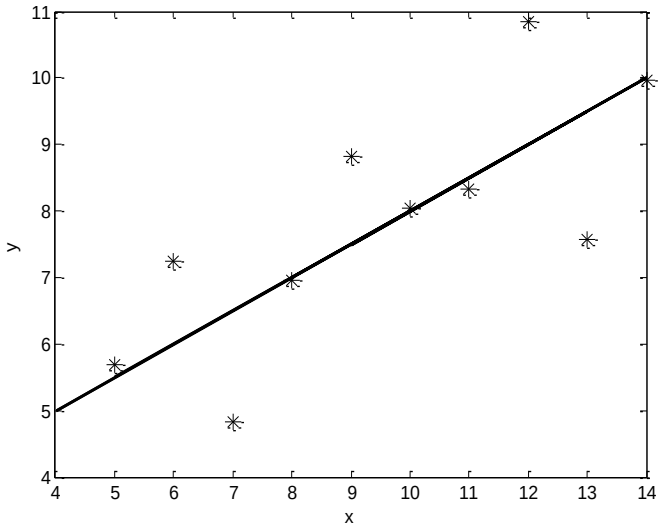


Fig. 4.1. Reprezentarea datelor şi a dreptei de regresie corespunzătoare

Din fig. 4.1 se observă că:

- Datele prezintă o dependenţă liniară;
- Nu este justificată alegerea unui model mai complicat;
- Nu apar valori aberante în setul de date;
- Variabilitatea lui y în raport cu x este aproximativ constantă, deci nu este cazul să se utilizezez ponderi în modelare.

Acest mod de abordare a prelucrării datelor, unde reprezentările grafice se îmbină cu determinarea unor indicatori cantitativi este abordarea corectă, pentru că se evită eventualele interpretări greşite.

Pentru a ilustra pierderea de informaţii, în cazul când nu se fac reprezentări grafice asociate cu calculele, se prezintă următorul exemplu [10] constând din patru seturi de date. Primul set este cel prezentat anterior, iar celelalte trei sunt

conţinute în tabelel 4.1. Se observă că valorile şirului X sunt identice în primele trei cazuri, iar X4 conţine o aceeaşi valoare, 8.00, cu excepţia unui singur element egal cu 19.

Tabel 4.1. Seturi de date

X2	10.00;	8.00	13.00	9.00	11.00	14.00	6.00	4.00	12.00	7.00	5.00
Y2	9.14;	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
X3	10.00;	8.00	13.00	9.00	11.00	14.00	6.00	4.00	12.00	7.00	5.00
Y3	7.46;	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
X4	8.00	8.00	8.00	8.00	8.00	8.00	8.00	19.0	8.00	8.00	8.00
Y4	6.58;	5.76	7.71	8.84	8.47	7.04	5.25	12.5	5.56	7.91	6.89

Tabel 4.2. Indicatori cantitativi

	Set 1	Set 2	Set 3	Set 4
N	11	11	11	11
$\bar{x}$	9.00	9.00	9.00	9.00
$\bar{y}$	7.50	7.50	7.50	7.50
Y=bx+a	Y=0.5x+3	Y=0.5x+3	Y=0.5x+3	Y=0.5x+3
r	0.816	0.816	0.816	0.817
σ_ε	1.237	1.237	1.236	1.236

După cum se observă în tabelul 4.2, cele patru seturi de date au indicatorii cantitativi de valori foarte apropiate. Dar în urma reprezentării grafice a acestor seturi de date (fig. 4.2) se observă că există diferenţe mari în ceea ce priveşte distribuţia lor în jurul dreptei de regresie.

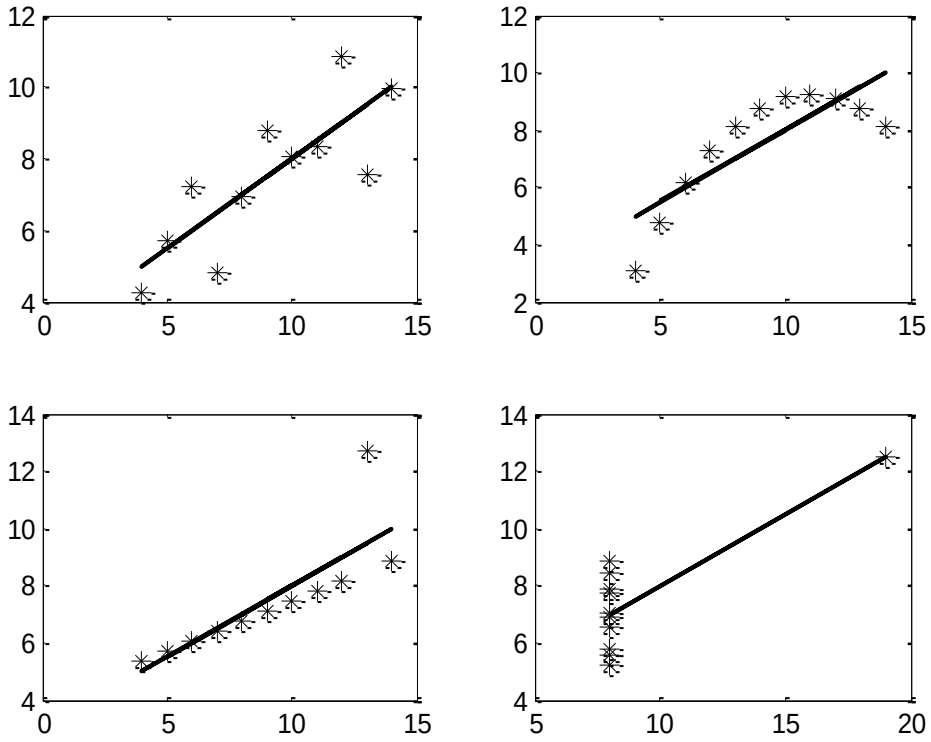


Fig. 4.2. Seturi de date având aceeaşi dreaptă de regresie

Analizând reprezentările grafice din figura 4.2 se pot trage următoarele concluzii:

- Primul set de date are un comportament liniar, fără valori aberante, la care modelul dreptei de regresie este corespunzător;
- Al doilea set de date are un comportament pătratic, fără valori aberante, modelul ar trebui să fie un polinom de gradul doi;
- Al treilea set de date prezintă o valoare aberantă, care ar trebui eliminată; în acest context ecuaţia dreptei de regresie se modifică, dreapta de regresie care ar modela corespunzător setul de date ar avea o pantă mai mică;
- Ultimul set de date este rezultatul unui experiment prost organizat, deoarece conţine o singură valoare mult distanţată de restul valorilor.

Din acest exemplu de prelucrare a datelor se poate observa importanţa analizei exploratorii, care amână alegerea modelului până la cunoaşterea comportamentului datelor. În cazul seturilor 2 – 4, eroarea apare din faptul că se presupune un model liniar, fără a face o analiză preliminară a datelor. Importanţa determinărilor cantitative nu trebuie subestimată, calcularea unor

indicatori: medie, dispersie, coeficient de corelaţie etc. permit o prezentare sintetică a datelor, ceea ce oferă un mare avantaj, dar calculele trebuie asociate cu reprezentări grafice.

La baza analizării unei serii de date stau câteva ipoteze fundamentale:

- Caracterul aleator al datelor;
- Datele au o distribuţie de probabilitate;
- Localizarea constantă;
- Variabilitatea constantă.

Previziunea este un ţel important în inginerie. Dacă ipotezele sunt valabile, se pot face previziuni asupra unui proces, inclusiv afirmaţii legate de evoluţia anterioară, se spune în acest caz că procesul este "în control statistic". Dacă cele patru ipoteze nu sunt valabile, procesul este în derivă (în raport cu locaţia, variabilitatea sau distribuţia), imprevizibil şi necontrolabil. Orice caracterizare a unui astfel de proces va conduce la concluzii eronate.

În cazul unei serii de date, cel mai frecvent prin analiza statistică se urmăreşte înlocuirea seriei de date cu o valoare, la care se asociază un interval de incertitudine. În acest context, modelul matematic asociat unei serii de date este:

$$y_i = a + \varepsilon_i, \tag{4.1}$$

unde a este o constantă, iar ε_i este eroarea aleatoare asociată valorii i a şirului. În funcţie de indicatorul statistic utilizat pentru determinarea constantei a, se stabileşte mărimea intervalului de incertitudine asociat. Aceasta este abordarea ce se efectuează pentru a se aprecia, de exemplu, dacă un lot de repere îndeplinesc condiţiile de calitate, respectiv dacă se încadrează în limitele impuse de intervalele de toleranţă.

Pentru ca modelul matematic asociat să fie corect este necesar să fie îndeplinite toate cele patru ipoteze fundamentale. Testarea ipotezelor asigură valabilitatea concluziilor. Verificarea ipotezelor se face cu ajutorul unor tehnici grafice. Tehnicile utilizate sunt:

- graficul secvenţial al punctelor $Y_i(i)$
- graficul punctelor succesive $Y_i(Y_{i-1})$
- histograma
- graficul probabilităţii normale $Y_{exp}(Y_{estimat\ normal})$ – valorile experimentale se reprezintă în raport cu valorile estimate în cazul când datele au o distribuţie normală.

Pentru a deduce concluzii corecte din analiză este necesară găsirea unui

model corespunzător şi a unor estimări corecte ale parametrilor. În cazul când unele ipoteze nu sunt respectate pot apare probleme:

- Neglijarea caracterului aleator conduce la: invalidarea majorităţii testelor statistice, lipsa de semnificaţie a incertitudinilor calculate, invalidarea modelului $y_i = a + \varepsilon_i$ etc.
- Un aspect specific al caracterului nealeator este autocorelaţia, adică legătura dintre y_t şi y_{t-k}, unde k este un număr întreg ce defineşte decalajul pentru autocorelaţie. Acest lucru este valabil la seriile cronologice. Fenomenul se depistează din graficul de autocorelaţie.
- Nerespectarea localizării constante. Estimarea uzuală pentru localizare este media. Dacă acest lucru nu este real atunci centrarea poate fi deplasată, estimarea centrării nu mai are semnificaţie, iar formula incertitudinii mediei ($S_{\bar{y}} = \frac{S}{\sqrt{n}}$ - S fiind abaterea standard) indică valori mult mai mici decât cele reale.
- Aceleaşi aspecte apar în cazul nerespectării variabilităţii constante.

Uzual, media este estimatorul localizării. Variabilitatea mediei este legată de ipotezele admise ale distribuţiei de probabilitate a datelor. Pentru anumite tipuri de distribuţii media nu este estimatorul cel mai bun. Se poate utiliza media, mediana, mijlocul amplitudinii datelor, modul. Din acest motiv trebuie stabilită întâi distribuţia şi după aceea estimatorul.

4.2. TEHNICI GRAFICE UTILIZATE ÎN ANALIZA EXPLORATORIE A DATELOR UNIDIMENSIONALE

4.2.1. GRAFICUL SECVENŢIAL AL PUNCTELOR (RUN SEQUENCE PLOT)

Scop: verifică deplasări ale localizării, variabilităţii şi prezenţa valorilor aberante.

Reprezentare: $y_i(i)$.

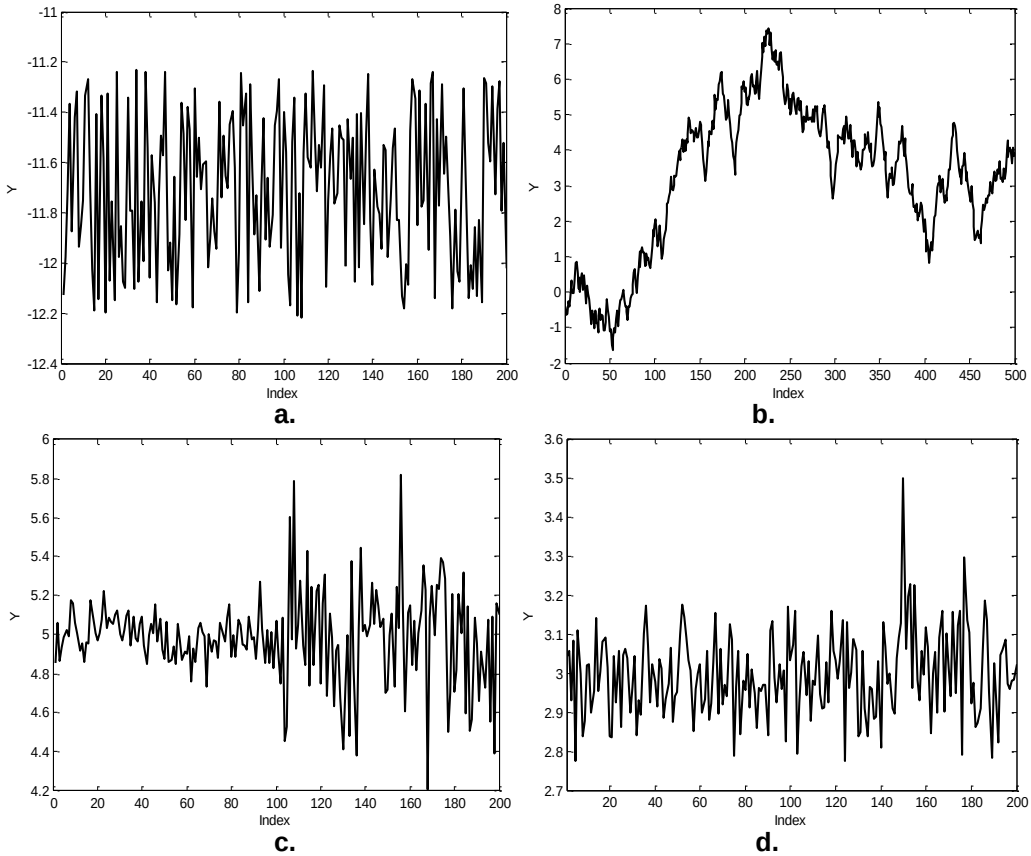


Fig. 4.2. Graficul secvenţial al punctelor

În fig. 4.2 se prezintă două exemple de astfel de grafice. În primul caz, fig. 4.2.a, se observă că localizarea setului de date este constantă – dacă s-ar înlocui graficul printr-o dreaptă, aceasta ar fi o dreaptă paralelă cu axa 0x; variabilitatea şirului este de asemenea constantă – variaţiile de o parte şi alta în jurul valorii medii este aproximativ aceeaşi, adică proiecţia pe axa 0y în diferite zone ale graficului au aceeaşi valoare; nu apar valori aberante. În fig. 4.2.b setul de date prezintă modificări ale localizării, în prima jumătate apare o tendinţă crescătoare, iar în partea a doua una descrescătoare, nu apar modificări de variabilitate sau valori aberante. În fig. 4.2.c se remarcă modificarea de variabilitate care apare în a doua jumătate a setului de date, aproximativ de la jumătate. Localizarea este constantă, iar din punctul de vedere al valorilor aberante, există câteva valori situate la distanţă mai mare de valoarea medie, valori ce pot apare datorită creşterii variabilităţii setului de date. În fig. 4.2.d localizarea setului de date este constantă, variabilitatea de

asemenea, apare însă suspiciunea unei valori aberante, cea situată în jurul valorii 3.5.

Interpretare: graficul trebuie să poată fi aproximat cu o dreaptă paralelă cu axa absciselor, să aibă amplitudinea în direcţia axei 0y aproximativ constantă, să nu apară valori situate la distanţă mare de restul valorilor. Această interpretare a graficului se poate face şi prin metode cantitative, în sensul că se aproximează graficul cu dreapta de regresie (x este numărul de ordine al elementelor, y este seria de date). Panta dreptei de regresie trebuie să fie 0. Se verifică valoarea pantei dacă diferă semnificativ de 0.

4.2.2. GRAFICUL PUNCTELOR SUCCESIVE (LAG PLOT)

Scop: verificarea caracterului aleator al datelor.
Reprezentare: Se reprezintă $Y_i(Y_{i-1})$, adică perechea de puncte (Y_{i-1}, Y_i).
Interpretare: În cazul datelor aleatoare reprezentarea nu evidenţiază nicio structură. Aspectul datelor este asemănător cu cel al unei ţinte dintr-un poligon de tragere.

În fig. 4.3 se prezintă câteva exemple de astfel de grafice. În primul caz (fig. 4.3.a) se remarcă aspectul nestructurat al graficului, ceea ce indică un caracter aleator. Nu par să fie valori aberante neexistând puncte izolate. În graficul 4.3.b se observă un model liniar, ceea ce indică un caracter puternic nealeator. În acest caz un model de autoregresie pare mai potrivit. Datele se grupează de-a lungul unei drepte. Pentru o anumită valoare a unui punct, de ex. $y_{i-1} = 0$, se observă că $y_i \in (-3; 2)$, deci se pot face predicţii asupra valorilor succesive ale şirului de date. Se sugerează un model de forma:

$$y_i = a_0 + a_1 \cdot y_{i-1} + \varepsilon_i \qquad (4.2)$$

În fig. 4.3.c modelul de autoregresie este prezent din nou, de data aceasta cu o autocorelaţie mai puternică. În acest caz, cât şi în cel precedent nu apar valori aberante. Se remarcă gruparea clară a datelor de-a lungul unei drepte. Fenomenul de autocorelare se poate datora fenomenului studiat, variaţiilor condiţiilor de mediu, distorsionării datelor de la sistemul de achiziţie.

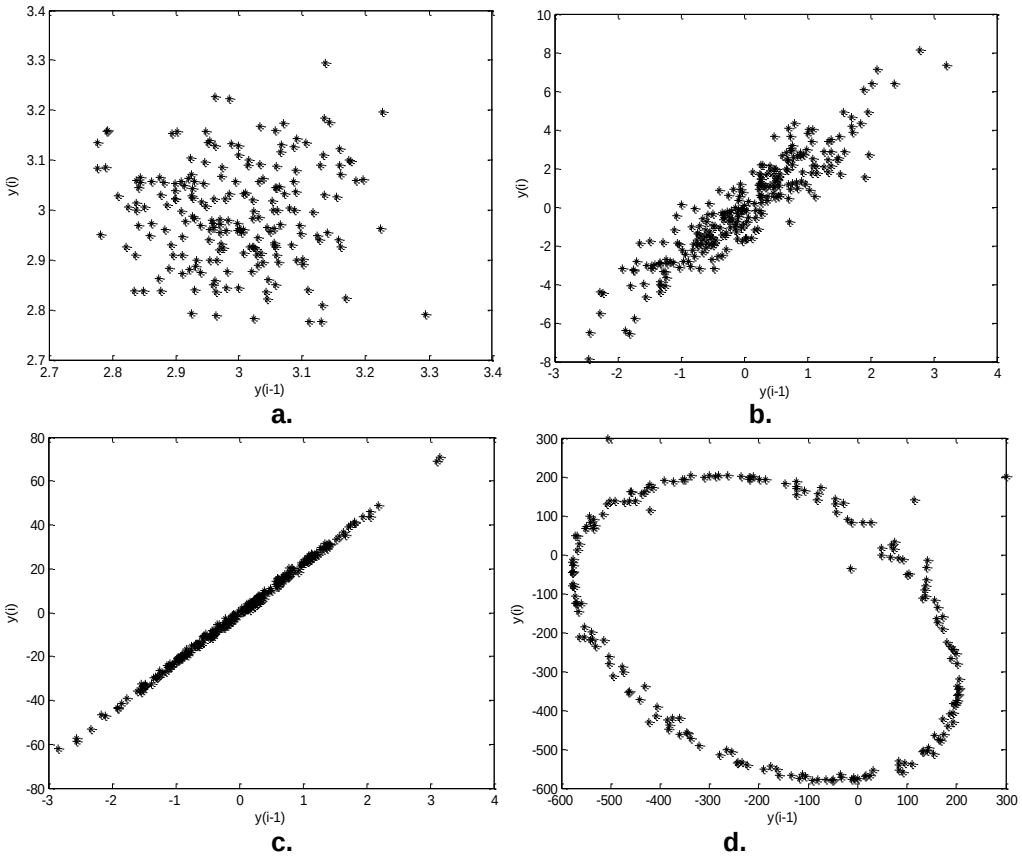


Fig. 4.3. Graficul punctelor succesive

În fig. 4.3.d apare un model de formă eliptică. Acest tip de grupare este caracteristica modelelor sinusoidale. Datele provin dintr-un model periodic de tip armonic; apar şi valori aberante.Modelul corect este:

$$y_i = C + \alpha sin(2\pi\omega t_i + \varphi) + \varepsilon_i. \quad (4.3)$$

iar parametrii modelului se pot estima cu metoda celor mai mici pătrate, [1]. Punctele mai îndepărtate de elipsă indică posibila prezenţă a valorilor aberante.

În concluzie, dacă datele au caracter aleator, graficul puctelor succesive are un aspect nestructurat, nu se pot face predicţii legate de valorile următoare ale setului de date, deci între y_{i-1} şi y_i nu există nici o relaţie.

4.2.3. HISTOGRAMA

Scop: reprezentarea distribuţiei datelor în intervale de lungime constantă.
Histograma indică: localizarea datelor, variabilitatea acestora, asimetria, prezenţa valorilor aberante, caracterul uni- sau multimodal al repartiţiei. Aceste caracteristici furnizează indicaţii clare referitoare la modelul corespunzător distribuţiei de probabilitate a datelor. Graficul de probabilitate sau un test de concordanţă poate fi utilizat pentru verificarea modelului.

În fig. 4.4 se prezintă mai multe histograme. În fig. 4.4.a se observă: simetria repartiţiei, existenţa unor extremităţi de anvergură moderată (cozi de lungime moderată), clasica formă de clopot. Aceast tip de repartiţie apare cel mai frecvent în natură. Dacă histograma este simetrică, cu anvergură moderată la extremităţi, nu există motive să nu se încerce modelarea setului de date cu o repartiţie normală. Pentru confirmarea provenienţei datelor din distribuţia normală se construieşte graficul probabilităţii normale. În cazul că acest grafic este liniar, atunci modelul este corespunzător.

În fig. 4.4.b histograma indică o repartiţie diferită de cea normală având extremităţile fără coadă (short tail).
Pentru o repartiţie simetrică ”corpul repartiţiei” este centrul acesteia, regiunea unde apar probabilităţile maxime de apariţie. Lungimea extermităţilor indică cât de repede aceste probabilităţi se apropie de zero. Repartiţiile ce au o astfel de caracteristică au un caracter trunchiat, cum este repartiţia uniformă – probabilitate constantă pe un domeniu şi zero în rest. Setul de date care a generat această histogramă aparţine unei repartiţii uniforme.

Dacă cozile sunt moderate, probabilitatea ca datele să aparţină unor intervale îndepărtate de medie scade progresiv. Modelul clasic de acest tip este repartiţia normală. Dacă coada este mare, ca în fig. 4.4.c, probabilitatea de apariţie scade lent, există probabilitate de apariţie la distanţă mare de corpul repartiţiei. Modelul clasic pentru un astfel de set de date este repartiţia Cauchy.

Estimarea optimă (nedeplasată şi consistentă) pentru localizarea centrului repartiţiei depinde în mare măsură de anvergura extremităţilor. Alegerea uzuală a N observaţii şi calcularea mediei acestora ca un estimator al localizării este bună în cazul repartiţiei normale (cozi medii), este o alegere nepotrivită pentru pentru repartiţii asemănătoare cu cea uniformă şi total nepotrivită, eronată chiar, în cazul datelor provenind din distribuţia Cauchy (cu anvergură mare). Media este un estimator corespunzător numai în cazul repartiţiei normale, [10].

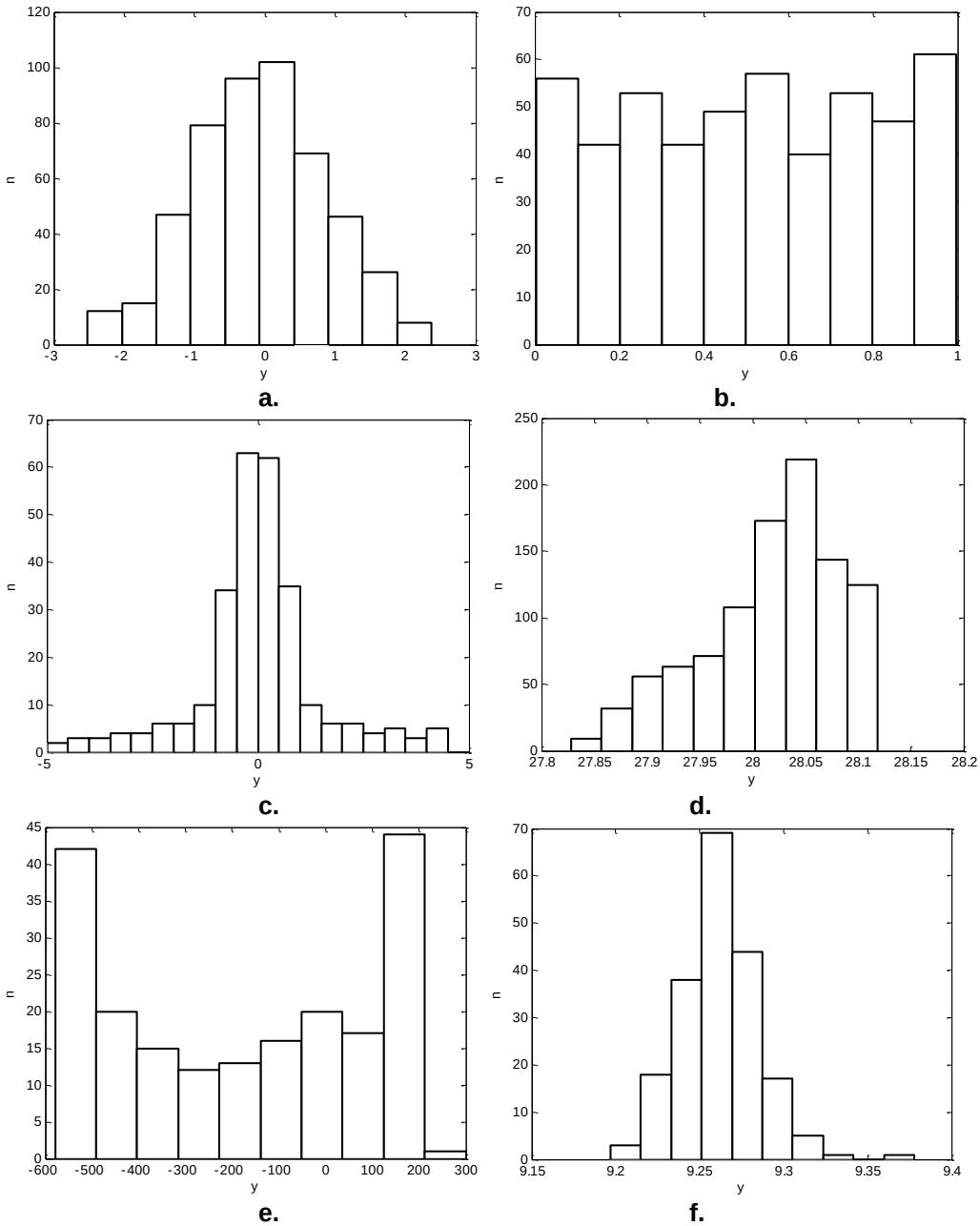


Fig. 4.4. Tipuri de histograme

Pentru repartiţia uniformă cel mai bun indicator al localizării este mijlocul amplitudinii (semisuma valorii minime şi maxime). Pentru repartiţii tip Cauchy,

mediana este cel mai bun estimator al valorii centrale.

În fig. 4.4.d se prezintă o histogramă asimetrică. Pentru repartiţiile asimetrice este uzual să aibe o extremitate considerabil mai lungă decât cealaltă. Asimetria dreapta are extremitatea cu anvergură mai mare în dreapta, respectiv asimetria stânga are extremitatea din stânga cu anvergură mai mare. Repartiţiile asimetrice ridică probleme de estimare, media nu mai are consistenţă. Pentru aceste repartiţii nu mai apare “centrul” în sensul uzual al cuvântului. Dintre indicatorii statistici, nici modulul nu prezintă semnificaţie deosebită.

La repartiţiile simetrice media, mediana şi modulul sunt identice, la cele asimetrice ele sunt distincte. În practică, repartiţiile asimetrice se caracterizează cel mai frecvent prin medie, în unele cazuri prin mediană şi cel mai rar prin modul. Cel mai corect este să se caracterizeze prin cel puţin doi indicatori, preferabil toţi trei. Asimetria poate apare datorită limitării inferioare sau superioare a datelor. Limitarea inferioară generează o asimetrie dreapta, iar limitarea superioară o asimetrie stânga. Asimetria poate fi cauzată de efectul de pornire a unor procese. În aplicaţiile de fiabilitate, unele procese pot avea un număr ridicat de căderi iniţiale.

Dacă histograma indică asimetrie dreapta sau stânga se recomandă următorii paşi:

- rezumarea datelor prin calcularea indicatorilor: medie, mediană şi modul;
- să se determine cea mai bună aproximare din familia repartiţiilor asimetrice, cum ar fi Weibull, gamma, exponenţială etc.
- să se ia în considerare o schimbare de variabilă pentru normalizare.

Pentru multe repartiţii este uzual ca răspunsul să fie grupat în jurul unei singure valori (modulul) şi să se distribuie cu frecvenţe mai scăzute spre extremităţi. Repartiţia normală este exemplul clasic de repartiţie unimodală. Modelul de histogramă din fig. 4.4.e prezintă clar aspectul de repartiţie bimodală. Pentru a găsi explicaţii trebuie continuată analiza datelor:

- se construieşte graficul secvenţial al punctelor pentru a verifica eventualele tendinţe ce apar;
- se construieşte graficul punctelor succesive pentru a verifica caracterul armonic.

Histograma din fig. 4.4.f indică prezenţa valorilor aberante. Acestea trebuie investigate pentru a se găsi cauza apariţiei acestora. Eliminarea valorilor

aberante prin aplicarea unor algoritmi poate fi înşelătoare. Se recomandă verificarea prezenţei valorilor aberante cu box-plot precum şi verificarea cantitativă prin testul Grubbs.

Reprezentarea histogramelor în Matlab se face utilizând frecvenţa absolută a seriei de date (fig. 4.5).

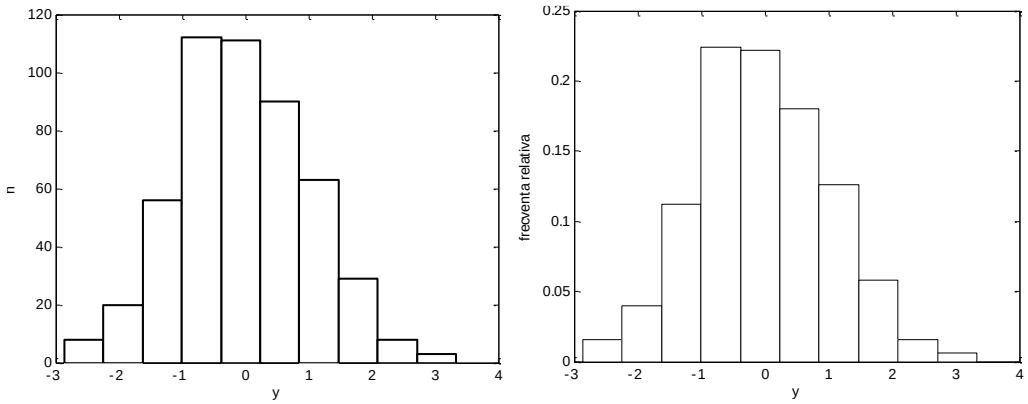


Fig. 4.5 Histograma frecvenţelor absolute şi relative

Pentru a obţine histograma frecvenţelor relative, frecvenţele absolute ale fiecărui interval trebuie împărţite cu numărul total de valori al şirului. Cele două grafice prezentate în fig. 4.4 se obţin cu următoarea secvenţă de program:

```
hist(y),xlabel('y'),ylabel('n')
%setare contur cu negru si dreptunghiuri albe
h = findobj(gca,'Type','patch');
set(h,'FaceColor','w','EdgeColor','k')
[n,limite]=hist(y);
figure,bar(limite,n/length(y),1),
h = findobj(gca,'Type','patch');
set(h,'FaceColor','w','EdgeColor','k')
```

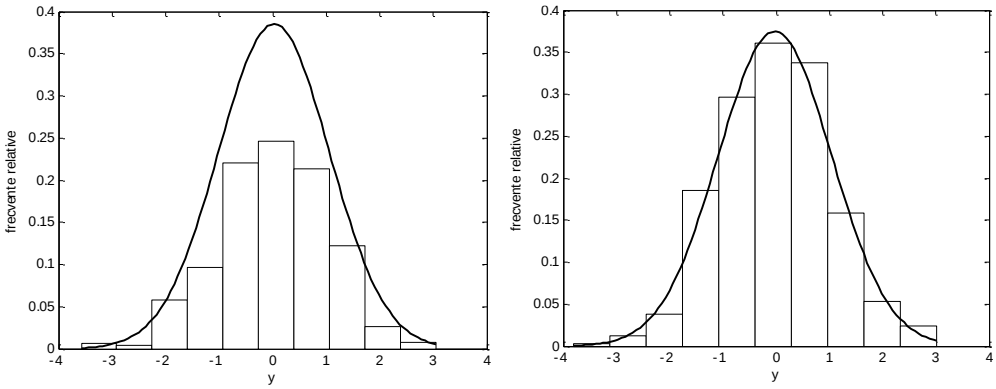


Fig. 4.5 Histograma frecvenţelor relative şi normalizate împreună cu funcţia densităţii de probabilitate

Pentru ca reprezentarea histogramei să poată fi comparată cu fucţia densitate de probabilitate este necesar să se facă normalizarea histogramei, adică suma ariilor dreptunghiurilor trebuie să fie egală cu 1. În fig. 4.5 sunt reprezentate histogramele aceluiaşi set de date, având suprapusă funcţia densitate de probabilitate normală, estimată pe baza parametrilor setului de date. Se remarcă în primul caz, când nu este făcută normalizarea, că funcţia densitate de probabilitate este mult mai înaltă decât histograma, în timp ce în al doilea caz, cele două funcţii sunt mult mai apropiate, histograma reprezentând mult mai veridic funcţia densitate de probabilitate.

Construirea histogramei normalizate se face cu ajutorul următoarei secvenţe de program:

```
% histograma normalizata
miu=mean(y);v=var(y);
xp=linspace(min(y),max(y));
yp=normpdf(xp,miu,v);
[n,limite]=hist(y);
hh=limite(2)-limite(1);
figure,bar(limite,n/(length(y)*hh),1),
h = findobj(gca,'Type','patch');
set(h,'FaceColor','w','EdgeColor','k'),hold on
plot(xp,yp)
```

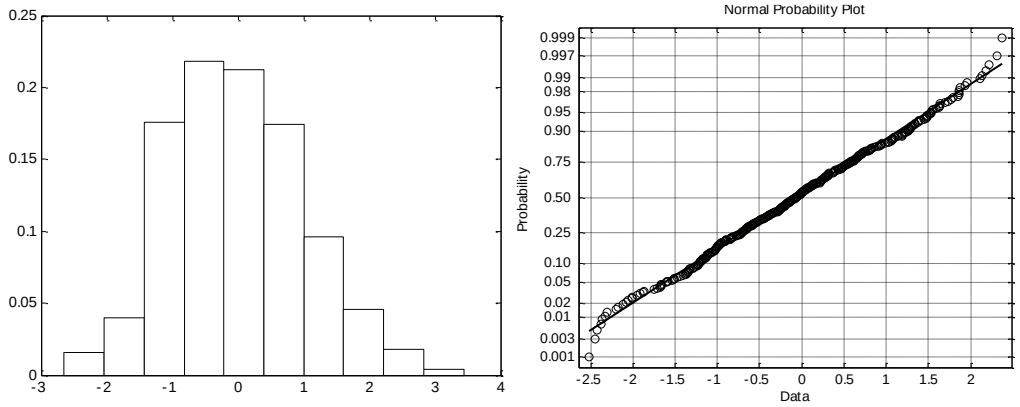
4.2.4. GRAFICUL PROBABILITĂŢII NORMALE (NORMAL PROBABILITY PLOT)

Scop: verificarea provenienţei datelor dintr-o distribuţie normală.

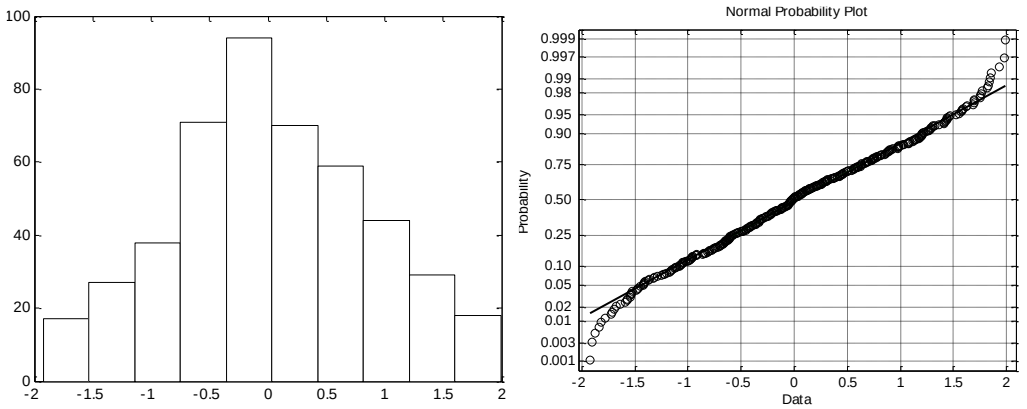
Interpretare: Datele se reprezintă faţă de repartiţia normală teoretică într-un astfel de mod încât punctele ar trebui să se situeze pe o dreaptă. Îndepărtarea de dreaptă indică abateri de la repartiţia normală. Graficul probabilităţii normale este un caz particular al graficului de probabilitate.

Reprezentare: Graficul se generează prin reprezentarea pe axa orizontală a setului de date ordonat, iar pe axa verticală a cvantilei corespunzătoare din repartiţia normală. Similar se procedează şi pentru alte repartiţii, putându-se verifica apartenenţa la distribuţiaţia respectivă. În Matlab acest tip de grafic se construieşte cu comanda: **`normplot(y)`**, y fiind setul de date.

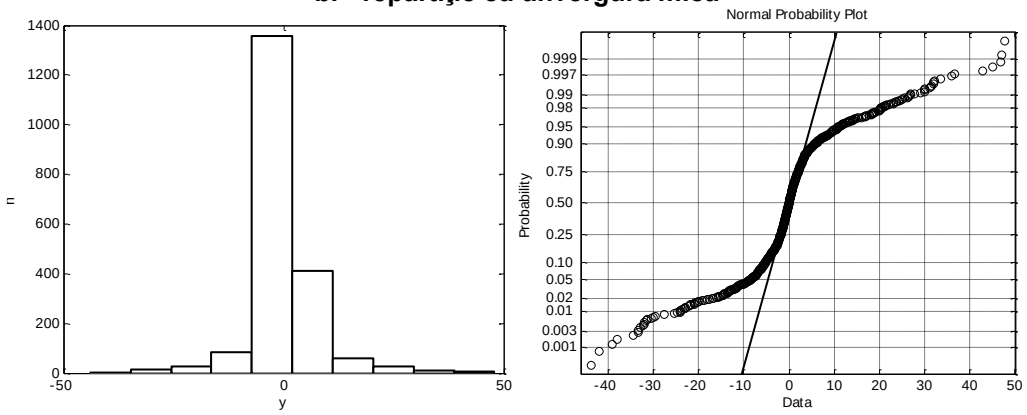
Un avantaj al acestei metode este faptul că panta şi intersecţia în origine a dreptei generate este o estimare a indicatorilor de variabilitate, respectiv de localizare. Acest lucru nu este important la repartiţia normală, dar devine semnificativ în cazul altor repartiţii. Coeficientul de corelaţie al punctelor reprezentate pe grafic se poate compara cu valorile critice pentru a testa provenienţa dintr-o populaţie repartizată normal.



a. anvergură medie



b. repartiţie cu anvergură mică



c. repartiţie cu anvergură mare

Fig. 4.6. Graficul probabilităţii normale pentru diferite tipuri de anverguri ale extremităţilor

Metoda furnizează răspunsuri la următoarele întrebări:

- sunt datele repartizate normal;
- care este natura îndepărtării de la normalitate (asimetria, extremităţi de anvergură mică sau prea mare)

Această metodă grafică oferă răspuns la ipoteza apartenenţei la o anumită repartiţie. Majoritatea modelelor statistice sunt de forma:

$$y_i = a + \varepsilon_i,$$

unde valoarea deterministă, a, se determină prin estimare, iar valoarea aleatoare este eroarea. Această componentă se ipotezează că este repartizată normal, cu localizarea şi variabilitatea constantă. Aceasta este aplicaţia cea mai frecventă a metodei. Se aproximează un model şi se generează graficul probabilităţii normale pentru erori. Dacă aceste erori nu sunt repartizate normal, înseamnă că una din ipotezele de bază a fost încălcată.

În fig. 4.6 sunt prezentate graficele probabilităţii normale pentru diferite anverguri ale extremităţilor asociate cu histogramele aferente. Se observă alinierea la dreaptă în cazul a., setul de date fiind generat din distribuţia normală. În cazul când anvergura extremităţilor este mică, graficul probabilităţii normale are o formă de „S”, situaţie ce se poate observa în fig. 4.6.b (forma de „S” este mai atenuată) şi în 4.7, unde această formă este foarte clar evidenţiată. Acest aspect de „S” apare şi în cazul când setul de date provine dintr-o repartiţie cu anvergură mare (fig. 4.6.c – set de date provenind dintr-o repartiţie Cauchy). Diferenţa de amplitudine se manifestă prin aspectul punctelor de la extremităţile graficului:

- la anvergură mică: punctele de început sunt situate sub dreapta corespunzătoare repartiţiei normale, iar punctele de sfârşit sunt situate deasupra dreptei;
- la anvergură mare: punctele de început sunt situate deasupra, iar cele de sfârşit sunt sub dreaptă.

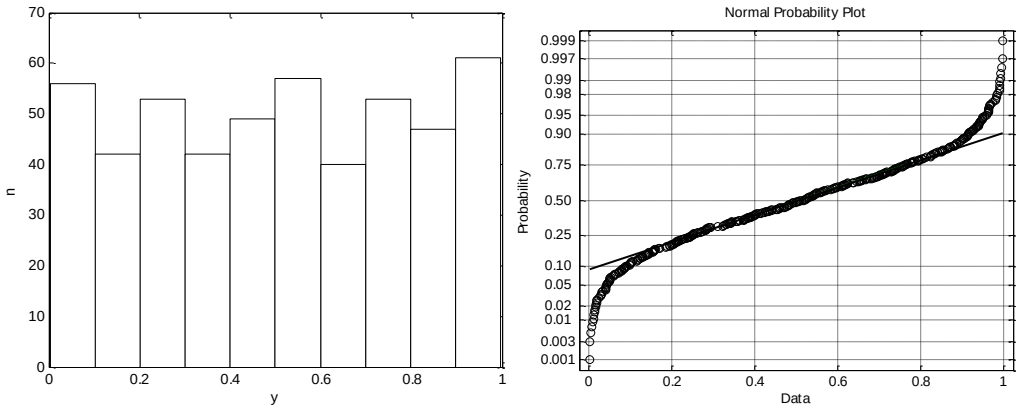
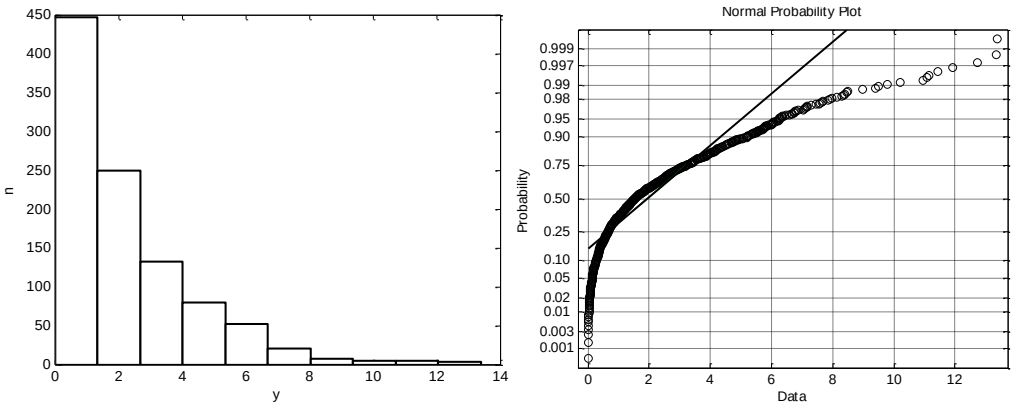


Fig. 4.7 Graficul probabilităţii normale pentru repartiţia uniformă

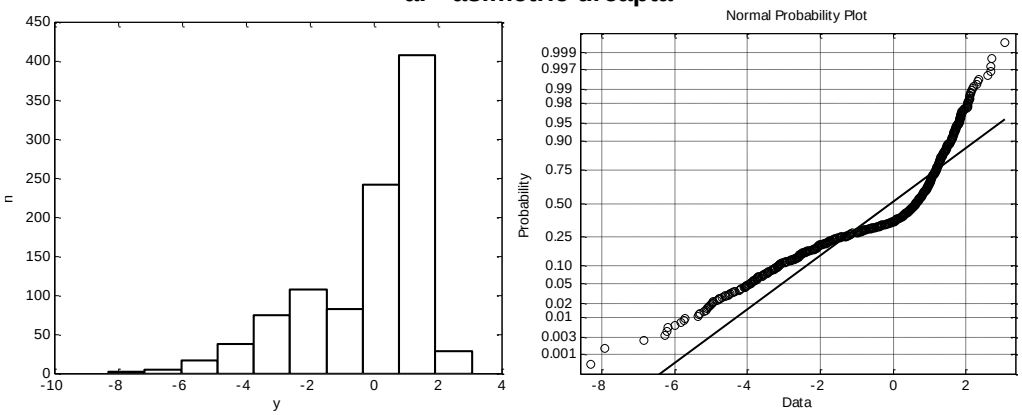
În fig. 4.8 sunt prezentate graficele probabilităţii normale pentru două seturi de date ce provin din repartiţii asimetrice. Caracteristica după care se poate identifica asimetria este aspectul pătratic al graficului. În cazul unei asimetrii dreapta (anvergura repartiţiei este amplă în partea dreaptă a axei), fig. 4.8.a, curba probabilităţii normale are punctele de început şi sfârşit situate în partea inferioară dreptei corespunzătoare repartiţiei normale. Dacă setul de date prezintă o asimetrie stânga, ca în fig. 4.8.b, aspectul pătratic se păstrează, dar punctele de început şi sfârşit ale curbei sunt situate în partea superioară dreptei.

În cazul când la analiza unui set de date se observă un grafic al probabilităţii normale care indică asimetrie, este necesară modelarea setului de date printr-

o distribuţie asimetrică, cum ar fi distribuţia exponenţială, distribuţia Weibull etc.



a. asimetrie dreapta



b. asimetrie stânga

Fig. 4.8 Graficul probabilităţii normale pentru repartiţii asimetrice

4.2.5. GRAFICUL CELOR PATRU

Scop: verificarea ipotezelor statistice fundamentale: caracterul aleator, provenienţa dintr-o anumită distribuţie de probabilitate, localizarea şi variabilitatea constantă.

Acest grafic este, de fapt o colecţie de 4 tehnici grafice:

- graficul secvenţial al punctelor (Run Sequence Plot);
- graficul punctelor succesive (Lag Plot);
- histograma;

- graficul probabilităţii normale.

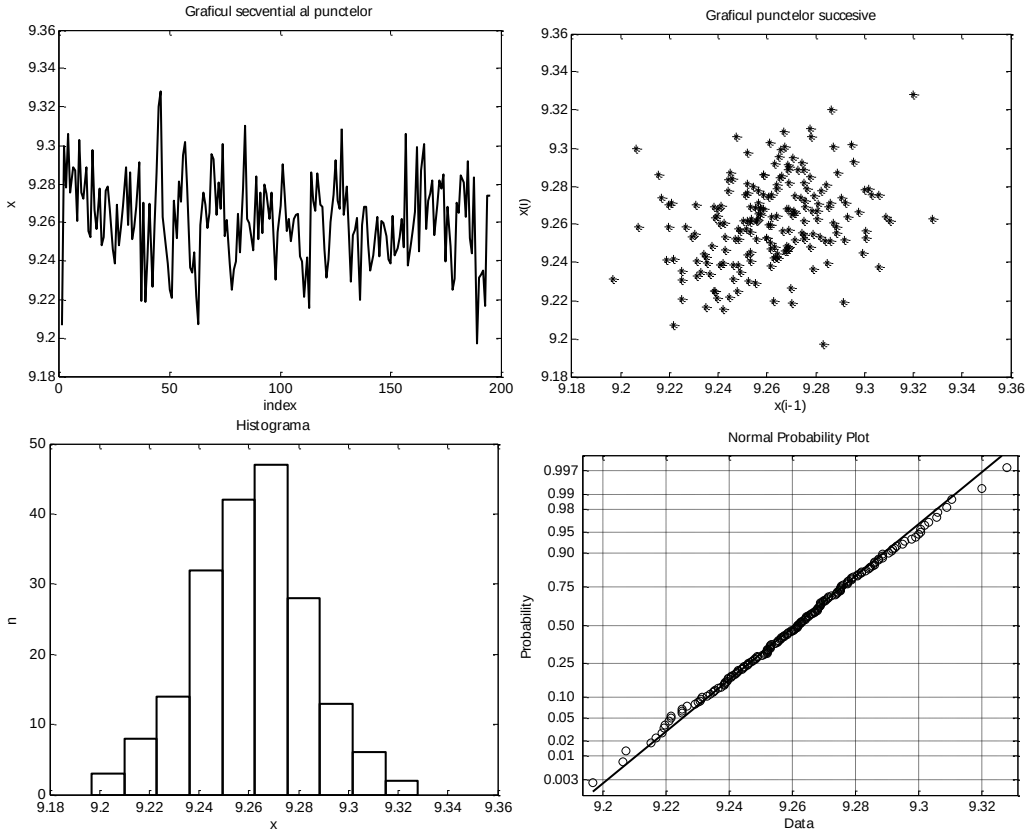


Fig. 4.9. Graficul celor 4 în cazul unui set de date provenind dintr-o repartiţie normală

Dacă cele 4 ipoteze fundamentale ale unui proces de măsurare sunt verificate, cele 4 grafice vor avea o alură caracteristică. Dacă una din ipoteze nu este îndeplinită, atunci acestă anomalie va fi evidenţiată de unul sau mai multe grafice.

Deşi cele 4 grafice se utilizează la şiruri unidimensionale şi serii cronologice, este util să fie extinse şi mai departe. Multe modele statistice se modelează prin:

$$y_i = f(x_1, x_2, \dots x_n) + \varepsilon_i, \tag{4.4}$$

adică cele 4 ipoteze trebuie să fie valabile pentru erorile ε_i, deci metoda se utilizează la validarea modelului prin analiza erorilor.

Se recomandă ca cele patru grafice să fie reprezentate pe aceeaşi pagină pentru a putea fi interpretate cu uşurinţă.

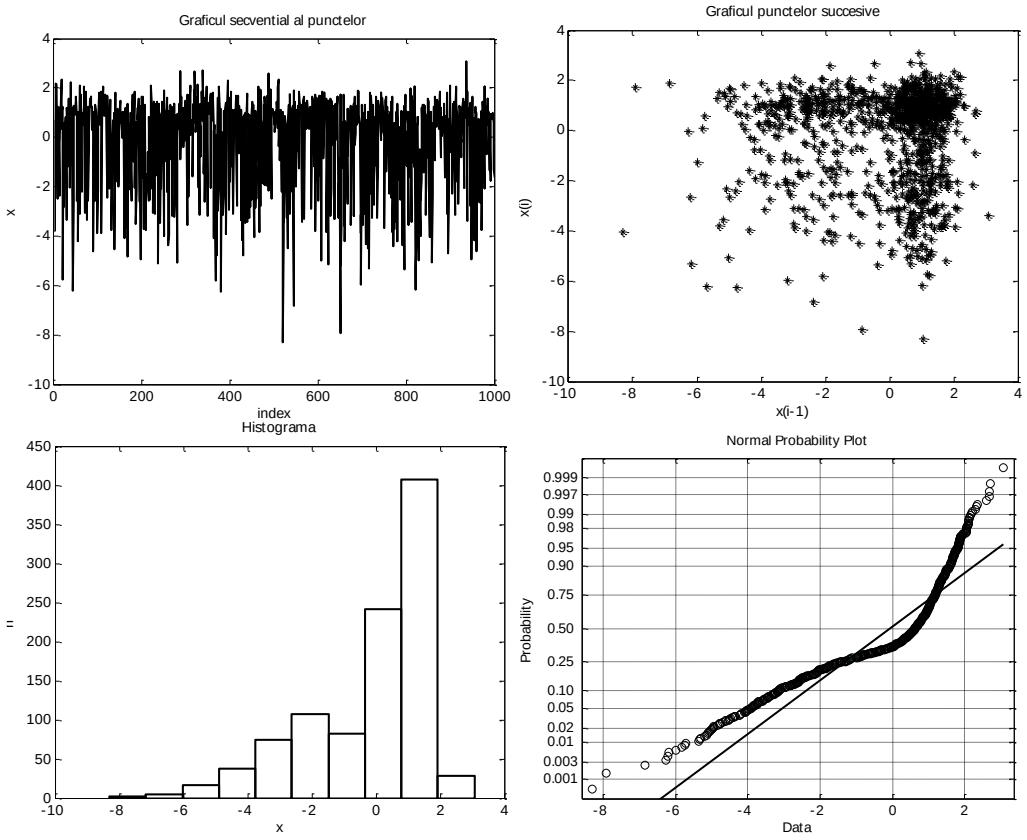


Fig. 4.10. Graficul celor patru pentru un set de date experimentale

Analizând cele patru grafice din fig. 4.9 se poate constata:

- graficul secvenţial al punctelor indică variabilitate şi localizare constantă, nu apr valori aberante;
- graficul punctelor succesive indică existenţa unui caracter aleator;
- histograma are forma clasică de clopot, deci repartiţia de proveninţă este simetrică, are anvergură moderată a extremităţilor, nu apr valori aberante;
- graficul probabilităţii normale are un caracter liniar, indicând faptul că datele provin dintr-o distribuţie normală.

În concluzie, nu sunt motive să nu se considere setul de date ca provenind dintr-o repartiţie normală, caz în care localizarea setului de date se poate

caracteriza prin media şirului, iar variabilitatea se poate caracteriza pe baza abaterii standard.

În fig. 4.10 este prezentat un alt exemplu pentru graficul celor patru pentru un set de date experimentale. Se poate constata:

- localizarea şi variabilitatea setului de date sunt aproximativ constante;
- datele au caracter aleator;
- repartiţia datelor prezintă o asimetrie stânga;
- repartiţia nu este o repartiţie normală.

4.2.6. GRAFICUL DE AUTOCORELARE

Scop: verificarea caracterului aleator.
Caracterul aleator se verifică prin calcularea autocorelaţiei pentru setul de date corespunzătoare la diferite valori ale decalajului. Dacă procesul este aleator, atunci autocorelaţia trebuie să fie 0 pentru orice decalaj. Dacă setul de date nu este aleator pentru unul sau mai multe decalaje, atunci autocorelaţia este semnificativ diferită de 0.

Reprezentare: Graficul se construieşte prin reprezentarea pe axa verticală a mărimii R_h, unde:

$$R_h = \frac{C_h}{C_0}, \tag{4.5}$$

C_h este autocovarianţa:

$$C_h = \frac{1}{N}\sum_{t=1}^{N-h}\left(Y_t - \overline{Y}\right)\left(Y_{t+h} - \overline{Y}\right), \tag{4.6}$$

iar C_0 este dispersia de selecţie necentrată:

$$C_0 = \frac{\sum_{t=1}^{N}\left(Y_t - \overline{Y}\right)^2}{N}. \tag{4.7}$$

Pe axa orizontală se reprezintă decalajul h = 1, 2,..., n-1. Se mai trasează şi 5 segmente de referinţă: un segment central la cota 0, iar celelalte 4 segmente delimitează intervalele de încredere de 95% şi 99%. Limitele de încredere se recomandă să fie calculate cu formula:

$$\pm \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{N}},$$

unde N este volumul eşantionului, z cvantila repartiţiei normale, iar α nivelul de semnificaţie.

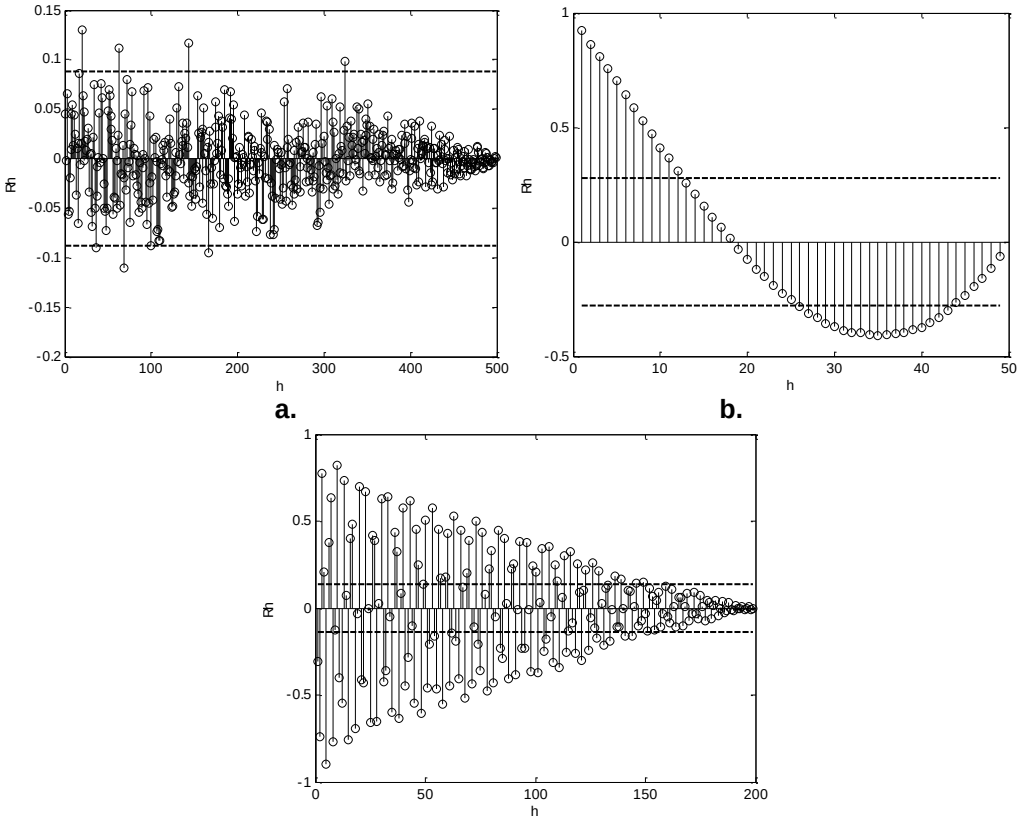


Fig. 4.11. Grafice de autocorelare

Construirea acestui grafic şi interpretarea corectă asigură validitatea concluziilor. Caracterul aleator este una din caracteristicile fundamentale ale proceselor de măsurare. Carcaterul aleator are o importanţă critică din următoarele motive:

- Formula utilizată cel mai frecvent pentru deviaţia standard a mediei de selecţie este $S_{\bar{y}} = S/\sqrt{N}$. Aceasta este adevărată numai în condiţiile caracterului aleator .
- Pentru date unidimensionale, modelul implicit este Y = a + ε, a fiind o constantă (estimarea valorii centrale) şi ε eroarea. Acest model

devine incorect şi nu mai este valid în cazul datelor nealeatoare, iar estimarea parametrilor nu mai are semnificaţie.

În fig. 4.11 sunt prezentate trei exemple de grafice de autocorelare, pe care sunt marcate cu linie întreruptă limitele intervalului de incertitudine corespunzător unui nivel de încredere de 95%. În primul caz se observă că nu apare o autocorelaţie semnificativă. Datele sunt aleatoare. Toate valorile sunt în intervalul de incertitudine de 95%, nu apare nici un tipar. Lipsa unei structurări a graficului indică caracterul aleator. Câteva valori sunt situate puţin în afara intervalului de încredere de 95%. Pentru o probabilitate de 95% este normal ca o valoare din 20 să fie în afara intervalului de încredere. Nu apare nici o asociativitate între Y_i şi $Y_{i+1.}$ Acest lucru este chiar esenţa caracterului aleator.

În fig. 4.11.b datele provin dintr-un model de autoregresie cu autocorelaţie puternică. Graficul indică autocorelare mare la decalaj 1 şi scade până la valori negative. Descreşterea este aproximativ liniară cu zgomot scăzut (unirea punctelor reprezentate generează o curbă netedă). Un astfel de tipar indică o autocorelare puternică, adică o previziune ridicată. Pentru pasul următor, se indică estimarea parametrilor pentru modelul de autoregresie:

$$Y_i = A_0 + A_1 \cdot Y_{i-1} + \varepsilon_i .$$

Această estimare se face cu metoda celor mai mici pătrate, mai precis pe baza dreaptei de regresie.

În cel de-al treilea exemplu, fig. 4.11.c, datele provin din model armonic. Apare o secvenţă alternativă de vârfuri pozitive şi negative, care manifestă tendinţă de scădere.

Construirea graficului de autocorelare se face cu următoarea secvenţă de program Matlab:

```
load x1000.dat;y=x1000;
n=length(y);
c0=var(y,1);
C(1:n-1)=0;
for i=1:n-1
    temp=0;
    for j=1:n-i
        temp=temp+(y(j)-mean(y))*(y(i+j)-mean(y));
    end
    C(i)=temp/n;
end
```

```
R=C/c0;
h=1:(n-1);
stem(h,R)
hold on
ls(1:(n-1))=1.96/sqrt(n);
li=-ls;
plot(h,ls,'r',h,li,'r')
```

4.2.7. GRAFICUL DE INCERTITUDINI (BOOTSTRAP PLOT)

Scop: determinarea incertitudinii unei statistici.

Din setul de date se extrage un eşantion de volum mai mic sau egal cu volumul datelor iniţiale. Eşantionul se obţine cu returnare, astfel încât fiecare element poate fi prelevat de mai multe ori sau deloc. Acest proces se repetă de mai multe ori, în general cel puţin de 500 ori. Valorile calculate (mediana, mijlocul intervalului de variaţie) se constituie într-o estimare a repartiţiei eşantionului.

Pentru fiecare eşantion se calculează indicatorul dorit, ex. mediana. În urma eşantionărilor repetate se obţin 500 de valori pentru mediană. Pentru determinarea incertitudinii, se ordonează şirul de 500 valori şi se selectează valoarea din poziţia 25, respectiv 475. În aceste condiţii incertitudinea este determinată cu un grad de încredere de 90% (25/500 = 0.05, 475/500 = 0.95, deci valorile determinate reprezintă o estimare a cvantilei). Cele 2 valori reprezintă limita inferioară şi cea superioară pentru incertitudine.

Reprezentare:

- pe axa verticală valoarea calculată pentru fiecare eşantion;
- pe axa orizontală – numărul eşantionului.

În general, acest grafic se asociază şi cu histograma aferentă, pentru a urmări atât localizarea, cât şi împrăştierea statisticii datelor.

În fig. 4.12 se prezintă graficul pentru medie, mediană şi centrul intervalului de variaţie, împreună cu histogramele aferente. Se observă că variabilitatea minimă o are centrul intervalului de variaţie. În aceste condiţii, acest indicator se recomandă pentru estimarea localizării setului de date ce a generat graficul.

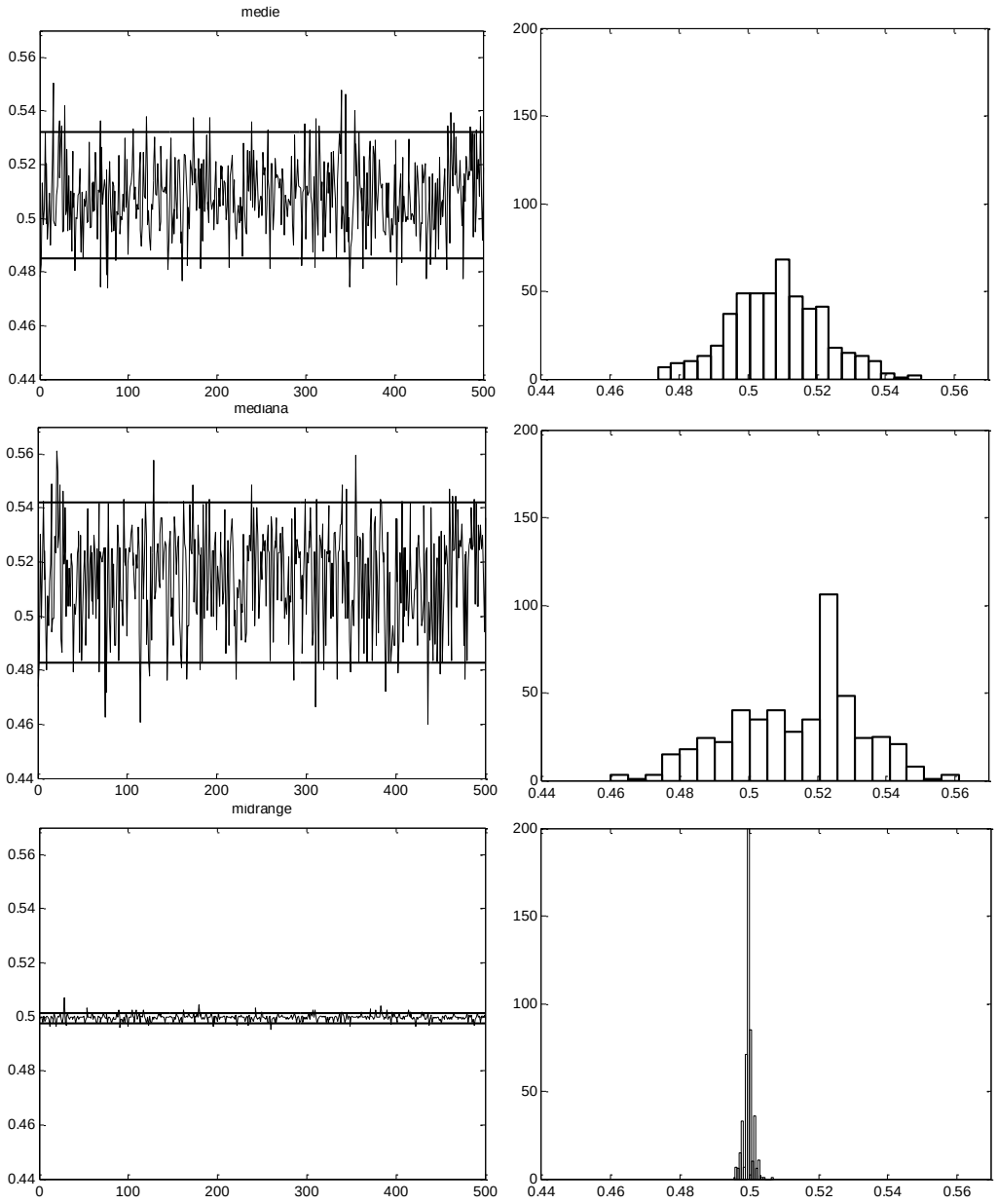


Fig. 4.12. Graficul pentru incertitudini

Construirea acestor grafice se poate face cu următoarea secvenţă de program:

```
load y1_500.dat
y=y1_500;
n=length(y);z(1:n)=0;
```

```
medie(1:500)=0;mediana(1:500)=0;midrange(1:500)=0;
for i=1:500
    indice=randsample(n,n,true);
    for j=1:n
        z(j)=y(indice(j));
    end
    medie(i)=mean(z);
    mediana(i)=median(z);
    midrange(i)=(min(z)+max(z))/2;
end
x=1:500;
medie_ord=sort(medie);
mediana_ord=sort(mediana);
midrange_ord=sort(midrange);
li(1:500)=medie_ord(25);ls(1:500)=medie_ord(475);
figure,plot(x,li,x,ls,x,medie),title('medie'),axis([0 500 .44
.57])
figure,hist(medie,20),axis([0.44 .57 0 200])
li(1:500)=mediana_ord(25);ls(1:500)=mediana_ord(475);
figure,plot(x,li,x,ls,x,mediana),title('mediana'),axis([0 500
.44 .57])
figure,hist(mediana,20),axis([0.44 .57 0 200])
li(1:500)=midrange_ord(25);ls(1:500)=midrange_ord(475);
figure,plot(x,li,x,ls,x,midrange),title('midrange'),axis([0 500
.44 .57])
figure,hist(midrange,20),axis([0.44 .57 0 200])
```

În program s-au impus scalările graficelor pentru a facilita comparaţia între cei trei indicatori.

4.2.8. GRAFICUL CVANTILĂ – CVANTILĂ (Q – Q PLOT)

Scop: Verifică dacă două seturi de date provin din aceeaşi distribuţie de probabilitate. De fapt, se verifică dacă cele 2 seturi de date provin din populaţii având aceeaşi repartiţie. De asemenea, cu ajutorul acestui tip de grafic se poate compara un eşantion cu o distribuţie de probabilitate prescrisă, pentru a verifica ipoteza provenienţei datelor din distribuţia de probabilitate respectivă. În această situaţie se ajunge la grafice de probabilitate, cum este cazul graficului probabilităţii normale, prezentat anterior.

Reprezentare: Această metodă grafică reprezintă cvantila primului set de date în raport cu cvantila setului al doilea, conform metodologiei prezentate în 3.2.2. Fie cele două seturi de date: X: x_1, x_2, ...x_n şi Y: y_1, y_2, ... y_m cu m ≤ n. Se ordonează ambele şiruri crescător: $X' = x'_1, x'_2, \dots x'_n$ şi $Y' = y'_1, y'_2, \dots y'_n$. În cazul

când eşantioanele au volume egale (n = m) se reprezintă puncte având coordonatele egale cu cvantilele corespondente ale celor două seturi de date. (În acest caz coordonatele punctelor sunt chiar elementele şirurilor ordonate $P_i(x_i', y_i')$).
Un avantaj major al acestei metode grafice este faptul că nu necesită eşantioane de volum egal. În situaţia când m < n, se vor reprezenta puncte y_i' în raport cu cvantila (i-0.5)/m a celuilalt şir.

Interpretare: În cazul când datele provin din populaţii cu aceeaşi repartiţie, punctele prezintă doar o abatere uşoară faţă de dreapta de referinţă. Cu cât punctele sunt situate la o distanţă mai mare de linia de referinţă, cu atât este mai evident faptul că seturile provin din repartiţii diferite. Graficul oferă informaţii calitative şi pentru ca rezultatul să aibă credibilitate este necesar ca volumul datelor să fie mare.

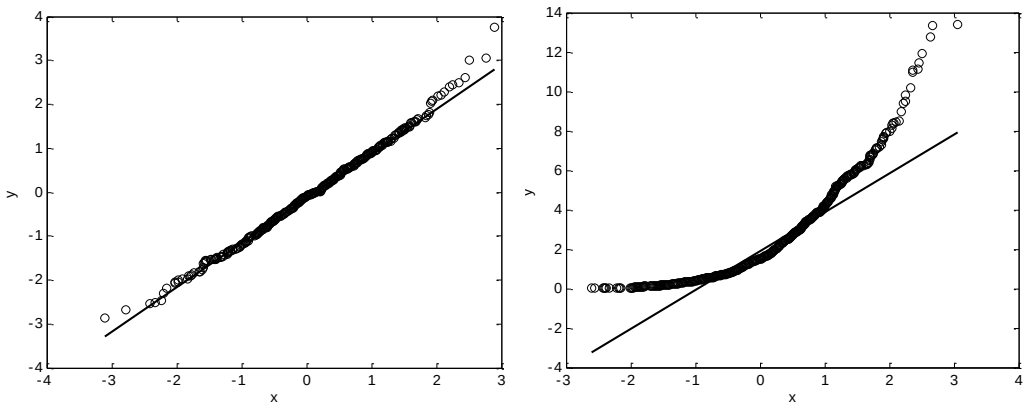


Fig. 4. 13. Grafice cvantila – cvantila

Principalul avantaj al acestei metode constă în faptul că eşantioanele nu trebuie să fie de acelaşi volum.

Dacă datele provin din populaţii ale căror repartiţie diferă doar prin localizare, punctele se vor situa pe o dreaptă aproximativ paralelă cu linia de referinţă, dar decalată în sus sau jos faţă de aceasta.

În fig. 4.13 se prezintă două grafice cvantilă – cvantilă, primul generat pe baza a două eşantioane provenite din repartiţia normală, iar cel de-al doilea pentru un eşantion provenit dintr-o repartiţie normală şi unul din repartiţie exponenţială. Se observă alinierea la o dreaptă în primul caz, ceea ce confirmă apartenenţa la o aceeaşi repartiţie a celor două eşantioane. Pentru a interpreta

mai uşor graficul, la grafic se mai ataşează o dreaptă obţinută pe baza cvantilelor inferioare şi superioare ale celor două şiruri. Dreapta din grafice este determinată de două puncte, un punct are coordonatele egale cu cvantila inferioară a şirului x, respectiv y, iar cel de-al doilea punct are coordonatele egale cu cvantilele superioare ale şirurilor. Dreapta se extinde până la valorile minime, respectiv maxime ale datelor. În Matlab, acest grafic se poate construi direct cu comanda **`qqplot(x,y)`**, unde x şi y sunt vectori ce conţin seriile de date.

Dacă există 2 seturi de date este important să se ştie dacă ipoteza unei repartiţii comune este justificată. Dacă da, localizarea şi variabiliatea pot fi estimate pentru ambele loturi. Dacă loturile diferă, este important să se cunoscă clar diferenţele.

Graficul cvantilă – cvantilă este similar graficului de probabilitate. Pentru un grafic de probabilitate, cvantilele unui set de date se înlocuiesc cu cele ale repartiţiei teoretice. În fig. 4.14 se prezintă graficul repartiţiei uniforme pentru două seturi de date.

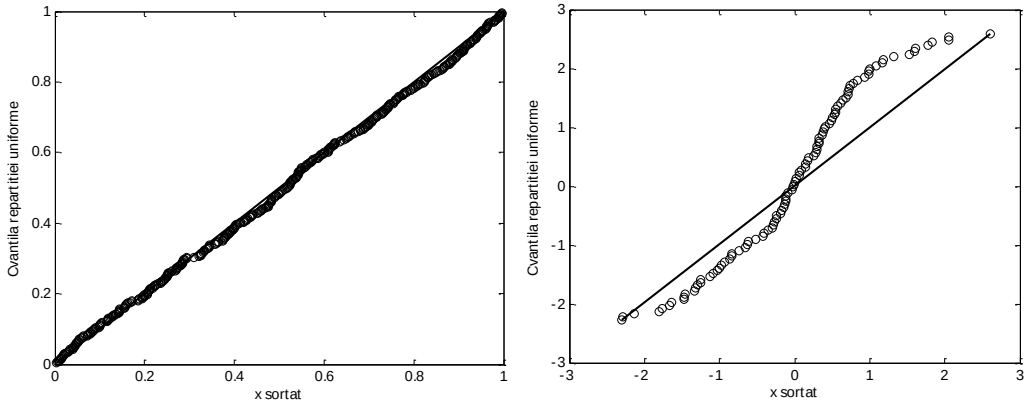


Fig. 4.14. Graficul repartiţiei uniforme

Se observă în primul grafic din fig. 4.14, că punctele se aliniază pe o dreaptă, ceea ce confirmă provenienţa setului de date dintr-o distibuţie uniformă. În cel de-al doilea caz, apar abateri semnificative de la dreaptă, dovadă a faptului că punctele nu provin dintr-o distribuţie de acest tip. Setul de date utilizat în acest caz a fost generat dintr-o distribuţie normală.

Construirea unui astfel de grafic se face cu următoarea secvenţă de program:

```
load y1_500.dat
x=y1_500;
n=length(x);
```

```
prob=((1:n)-.5)/n;% se calculeaza probabilitatile
corespunzatoare setului de date
qp=unifinv(prob,min(x),max(x));% se determina cvantilele
repartitiei teoretice
plot(sort(x),qp,'o')
% Se ataseaza dreapta de referinta
xa=quantile(x,.25);xb=quantile(x,.75);
ya=0.25;yb=0.75;
x_dr=[xa xb];y_dr=[ya yb];
figure,plot(sort(x),qp,'o',x_dr,y_dr)
m=(yb-ya)/(xb-xa);
ymin=ya+m*(min(x)-xa);
ymax=ya+m*(max(x)-xa);
dreapta_x=[min(x) max(x)];dreapta_y=[min(qp) max(qp)];
figure,plot(sort(x),qp,'o',dreapta_x,dreapta_y)
```

Prin înlocuirea funcţiei unifinv, ce defineşte cvantila distribuţiei uniforme, cu funcţia corespunzătoare altei distribuţii, se poate construi orice grafic de probabilitate dorit.

4.3. TEHNICI CANTITATIVE ASOCIATE ANALIZEI EXPLORATORII

Analiza exploratorie a datelor se asociază cu o serie de tehnici cantitative. Analiza cu tehnici grafice oferă doar informaţii calitative, ce trebuie validate cu metode cantitative. Majoritatea tehnicilor cantitative se împart în două mari categorii:

a – estimări cu intervale de încredere;

b – testarea ipotezelor.

De exemplu, cel mai uzual indicator al localizării este media. Media teoretică a populaţiei se obţine ca valoarea medie a variabilei aleatoare ce reprezintă caracteristica investigată a populaţiei. Ea se estimează prin calcularea mediei unui eşantion şi este o estimare punctuală. Intervalul de încredere extinde estimarea punctuală prin includerea unei incertitudini. Prin extragerea diferitelor eşantioane rezultă diferite medii. Se determină un interval pe baza unui nivel de încredere. Acest interval conţine valoarea medie teoretică cu o anumită probabilitate.

Prin testarea ipotezei, în loc să se asocieze un interval de încredere, se respinge o anumită afirmaţie referitoare la parametrul propus pe baza

informaţiei extrase din eşantion.

De exemplu, ipoteza nulă poate fi una din afirmaţiile despre parametru:

- media populaţiei este egală cu 10;
- deviaţia standard este egală cu 5;
- mediile a două populaţii sunt egale;
- deviaţiile standard pentru cinci populaţii sunt egale.

Respingera ipotezei nule înseamnă că este falsă. Acceptarea unei ipoteze nu înseamnă că este adevărată, doar că nu avem dovada să credem acest lucru. De aceea testarea ipotezei se face prin două ipoteze, una ce nu inspiră încredere – ipoteza nulă (H_0), şi o ipoteză ce se crede adevărată – ipoteza alternativă (H_1).

4.3.1. ESTIMĂRI ALE LOCALIZĂRII

O sarcină fundamentală în numeroase analize statistice este estimarea unui parametru de localizare pentru repartiţie, găsirea unei valori tipice sau centrale pentru a descrie cel mai bine datele.

În primul rând trebuie definită o valoare tipică. Pentru şiruri unidimensionale există trei estimări:

- media:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_1^n x_i,$$

este valoarea utilizată cel mai frecvent pentru caracterizarea unei repartiţii;

- mediana – valoarea ce împarte setul de date în două părţi egale:

$$x_{me} = x_{\frac{n+1}{2}}, \quad n = 2k+1$$

$$x_{me} = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right), \quad n = 2k$$

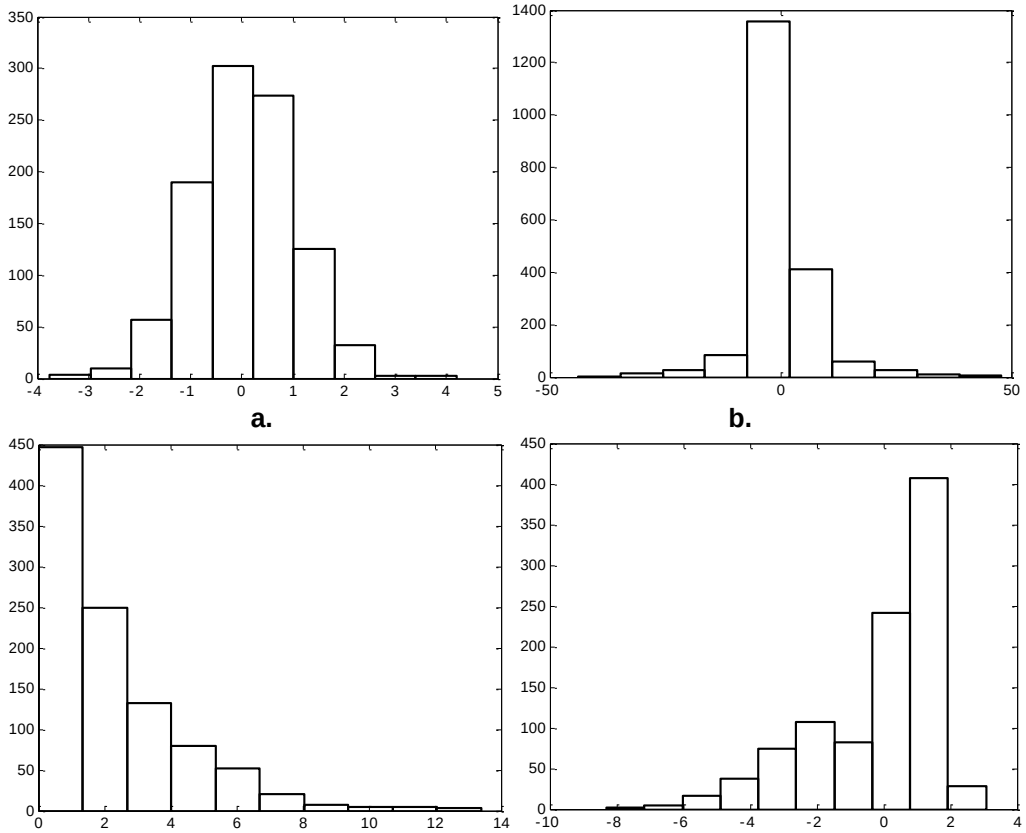
- modulul - valorea ce apare cu cea mai mare frecvenţă din toate valorile şirului de date. Nu este neapărat o valoare unică.

Sunt necesare mai multe posibilităţi de estimare a localizării, deoarece în anumite situaţii se preferă una din ele datorită repartiţiei datelor.

Pe baza repartiţiei normale (având $\mu = 0$, $\sigma = 1$) s-a generat un şir de date a cărui histogramă este prezentată în fig. 4.15.a. Media este 0.0434, mediana

0.0666, iar modulul este -0.0521. Repartiţia normală este simetrică, cu cozi moderate şi un singur punct de maxim al funcţiei densitate de probabilitate. Pentru repartiţii normale: media, mediana şi modulul sunt echivalente. Dacă histograma sau graficul de probabilitate indică faptul că datele pot fi aproximate de o repartiţie normală, este rezonabil să se utilizeze media pentru estimarea localizarii.

Pe baza repartiţia Cauchy a fost generat şirul de date reprezentat în fig.4.15.b; are media 0.2965, mediana 0.0021 şi modulul -43.9168. Repartiţia Cauchy este simetrică, dar are extremităţi cu anvergură mare şi un singur vârf în centru. Această repartiţie are proprietatea interesantă că prin colectarea unui număr mai mare de date, nu se generează o estimare mai precisă a mediei (repartiţia Cauchy nu are medie, [9]. Acest lucru înseamnă că media nu are semnificaţie ca măsură a localizării. Mediana este o măsură mai potrivită. Valorile mari din extremităţi distorsionează media, ceea ce nu se întâmplă cu mediana, ce depinde de indicele elementelor din şir.



c. **d.**

Fig. 4.15. Diferite tipuri de repartiţii

În general, pentru seturi de date cu cozi extinse, mediana este o estimarea recomandată pentru localizare.

Pentru şirul de date provenit dintr-o repartiţie exponenţială (fig. 4.15.c), media este 2.2191, mediana 1.5204 şi modulul 0.0037. Repartiţia exponenţială este asimetrică. Pentru repartiţia asimetrică media este diferită de mediană. Media este deplasată spre extremitatea cu anvergură mai mare. Asimetrie dreapta înseamnă medie mai mare decât mediana şi invers. În fig. 4.15.d este reprezentată o repartiţie cu asimetrie stânga. În acest caz media este -0.1574, mediana 0.6234, iar modulul -8.3048. Se poate observa că se respectă condiţia media < mediana.

La repartiţia asimetrică nu este evident care dintre indicatori caracterizează cel mai bine localizarea repartiţiei. Se recomandă să se indice toţi trei.

Există ma multe alternative pentru medie şi mediană pentru măsurarea localizării. Aceste alternative s-au dezvoltat pentru date ce nu au o repartiţie normală. În cazul repartiţiei normale, media este estimatorul optim. Indicatorii alternativi se adresează repartiţiilor cu coadă extinsă.

Indicatori alternativi**:**

- media intervalului intercvantilic – calculează media între cvantila inferioară şi superioară (25% şi 75%);
- media scurtată – se calculează pentru datele cuprinse între procentilele 5% - 95%;
- media scurtată modificată – valorile până la cvantila 5% se fac egale cu cea de 5% iar cele mai mari de 95% se modifică la valoarea de 95%;
- centrul intervalului de variaţie este egal cu media aritmetică dintre minim şi maxim, (min+max)/2.

În caracterizarea localizării unei serii de date se preferă estimarea cu ajutorul unui interval de încredere, deoarece estimarea punctuală este influenţată de eşantion. Intervalul de încredere oferă o indicaţie a cantităţii de incertitudine din estimarea mediei. Cu cât intervalul este mai îngust, cu atât estimarea este mai precisă. Intervalul de încredere pentru un eşantion din distribuţia normală de medie şi dispersie necunoscută este:

$$\bar{x} \pm t_{1-\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}, \qquad (4.8)$$

unde $t_{1-\frac{\alpha}{2}}(n-1)$ este 1- α/2 cvantila repartiţiei t cu n-1 grade de libertate şi se calculează în Matlab apelând funcţia **`tinv`**. Probabilitatea ca media teoretică să aparţină acestui interval este (1 - α).

Se remarcă faptul că lăţimea intervalului de încredere est controlată de doi factori:

- volumul eşantionului, n. Pe măsură ce n creşte, intervalul scade în raport cu $\sqrt{n}$ şi rezultă că o modalitate de a obţine o estimare mai precisă este creşterea volumului eşantionului;
- deviaţia standard, S. Cu cât S este mai mare, cu atât creşte şi intervalul de încredere.

Pentru a verifica dacă media unei populaţii normal distribuite are o valoare prescrisă μ_0, se testează ipoteza H_0 contra alternativei H_1:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

pe baza datelor dintr-un eşantion din populaţie y_1, y_2, ...y_n folosind funcţia test:

$$T = \frac{\bar{y} - \mu_0}{S / \sqrt{N}}$$

Nivel de semnificaţie: α = 0.05.
Regiune critică: se respinge H_0 dacă $|T| > t_{1-\frac{\alpha}{2}}(n-1)$.

Pentru alţi estimatori: mediana, media intervalului intercvantilic determinarea intervalului de încredere este dificilă din punct de vedere matematic. O alternativă este construirea graficului de incertitudini.

4.3.1. ESTIMĂRI ALE VARIABILITĂŢII

O sarcină fundamentală în multe analize statistice este caracterizarea variabilităţii (împrăştierea) unui set de date. La informaţiile legate de variabilitate a datelor, există două componente de bază:

1 - cât de împrăştiate sunt valorile datelor în apropierea centrului;
2 - cât de împrăştiate sunt datele la extremităţi.

Diferiţi indicatorilor statistici vor oferi informaţii cu ponderi diferite din acest punct de vedere. Alegerea unui estimator al localizării este determinată de componenta ce prezintă cel mai mare interes. Histograma este un instrument

grafic ce evidenţiază ambele componente.

Pentru date unidimensionale, există mai mulţi indicatori statistici pentru variabilitate:

1. Dispersia de selecţie

$$S^2 = \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2 / \left(N - 1\right)$$

Dispersia este, de fapt, o aproximare a mediei pătratelor distanţelor de la puncte la valoarea medie. Prin ridicarea la pătrat se atribuie ponderi mai mari valorilor ce sunt situate mai departe de medie. Deci, ea poate fi afectată mult de comportamentul extremităţilor.

2. Deviaţia standard

$$S = \sqrt{\frac{1}{N-1} \sum \left(Y_i - \bar{Y}\right)^2}$$

Deviaţia standard se exprimă în aceeaşi unitate de măsură ca şi dateleasociate eşantionului. Este afectată de comportamentul extremităţilor.

3. Intervalul de variaţie (IV) sau amplitudinea setului de date egală cu diferenţa dintre valoarea maximă şi valoarea minimă a şirului. Este un estimator ce se bazează doar pe valorile extreme. Nu oferă nici o informaţie în privinţa comportamentului valorilor centrale

4. Media deviaţiei absolute (AAD – average absolute deviation):

$$AAD = \sum_{i=1}^{N} \frac{\left|Y_i - \bar{Y}\right|}{N}. \qquad (4.9)$$

Acest indicator este mai puţin afectat de observaţiile extreme decât dispersia şi deviaţia standard. Este stabil numericşi se foloseşte în investigaţiile bazate pe calculator.

5. Mediana deviaţiei absolute (MAD – median absolute deviation)

$$MAD = median\left(\left|Y_i - \bar{Y}\right|\right), \qquad (4.10)$$

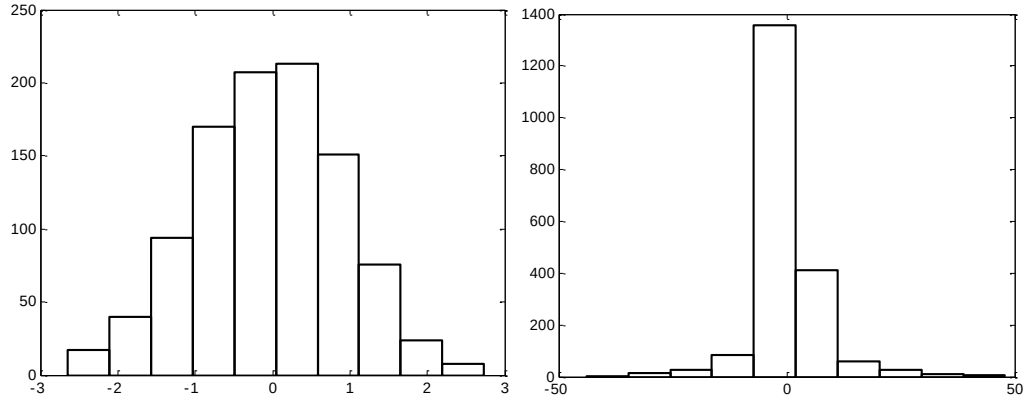
este un indicator ce este şi mai puţin afectat de valorile extreme, deoarece ele afectează mult mai puţin valorile medianei.

6. Lungimea intervalul intercvantilic: $q_{0.75} - q_{0.25}$, este diferenţa dintre cvantila superioară şi inferioară a setului de date şi determină variabilitatea punctelor din zona centrală.

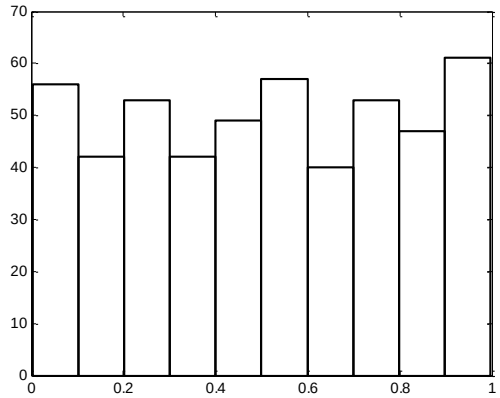
În concluzie, dispersia de selecţie, deviaţia standard, media deviaţiei absolute şi mediana deviaţiei absolute măsoară ambele aspecte ale variabilităţii. Ele diferă prin faptul că AAD şi MAD nu oferă ponderi valorilor extreme. Pe de altă parte, intervalul de variabilitate se bazează doar pe două puncte.

Necesitatea folosirii acestor indicatori se ilustrează în exemplele următoare. În fig. 4.16 sunt prezentate histogramele a trei seturi de date asociate cu indicatori ai variabilităţii: deviaţia standard, intervalul de variaţie sau amplitudinea şi mediana deviaţiei absolute.

Primul set de date (fig. 4.16.a) provine dintr-o repartiţie normală. Repartiţia normală este simetrică, cu extremităţile de anvergură moderată, un singur maxim al densităţii de probabilitate. Există o simetrie faţă de centru, caz în care mediana deviaţiei absolute este mai mică decât deviaţia standard. Dacă histograma sau graficul probabilităţii normale indică faptul că datele sunt bine aproximate de o repartiţie normală este rezonabil să se utilizeze deviaţia standard ca indicator al variabilităţii.



a. σ= 0.9435; IV= 5.3759; MAD= 0.6543 **b. σ= 7.4492; IV= 91.8277; MAD= 1.9493**



c. σ=0.2943; IV=0.9946; MAD= 0.2552

Fig. 4.16. Histograme asociate cu indicatori ai variabilităţii

În fig. 4.16.b se prezintă histograma unui set de date provenind dintr-o repartiţie Cauchy: σ= 7.4492; IV= 91.8277; MAD= 1.9493. Comparând repartiţia normală cu cea Cauchy se observă că repartiţia are un vârf mai pronunţat, scade mai rapid în apropierea centrului şi are extremităţi de anvergură mai mare. Datorită acestor extremităţi, deviaţia standard are tendinţa să crească comparativ cu repartiţia normală. Extremitatea de anvergură mai mare se reflectă în valoarea IV. Repartiţia are o proprietate interesantă – şi anume prin colectarea unui număr mai mare de date nu se obţine o medie sau o deviaţie standard mai precisă. Deci pentru eşantion şi populaţie cei doi indicatori sunt egali. Acest lucru înseamnă că pentru seturi de date provenind din repartiţia Cauchy deviaţia standard nu este o măsură a împrăştierii. Din histogramă se observă că toate datele sunt situate între –5 şi 5. Cu toate acestea, foarte puţine valori extreme determină creşterea mare a lui σ şi IV. MAD este doar puţin diferită de cea a repartiţiei normale. În acest caz, MAD este măsura potrivită pentru variabilitate.

Deşi repartiţia Cauchy este un caz extrem, ea ilustrează importanţa extremităţilor în măsurarea variabilităţii. Valorile extreme pot distorsiona deviaţia standard. Ele, însă, nu afectează MAD. Pentru repartiţii cu extremităţi pronunţate, MAD sau intervalul intercvantilic constituie indicatori mai buni pentru variabilitate.

Cel de-al treilea set de date (fig. 4.16.c) provine dintr-o repartiţie uniformă: σ=0.2943; IV=0.9946; MAD= 0.2552. Repartiţia are extremităţi trunchiate. În acest caz deviaţia standard şi MAD au valori mai apropiate decât în celelalte trei exemple, ce au extremităţi de diferite anverguri.

Dacă histograma şi graficul probabilităţii normale indică faptul că datele se pot aproxima cu o repartiţie normală, are sens utilizarea deviaţiei standard ca indicator al variabilităţii. În cazul când, datele au alt tip de repartiţie, cu extremităţi de anvergură mare, utilizarea altor indicatori: MAD sau IV este mai potrivită. IV se utilizează în unele aplicaţii, cum ar fi controlul calităţii, pentru simplitatea calculelor. În plus, prin compararea IV cu deviaţia standard avem o indicaţie legată de împrăştierea datelor la extremităţi. Acest indicator, IV, trebuie utilizat cu precauţie fiind bazat pe doar două valori.

BIBLIOGRAFIE

1. Davidescu, A., Analiza şi procesarea datelor cu aplicaţii în Matlab, Ed. Politehnica Timişoara, 2003.
2. Fleming, M., Nellis, J., Principles of Applied Statistics, Routledge Co., London, 1994.
3. Hanselman, D., Littlefield, B., The Student Edition of Matlab®. User's Guide, Prentice Hall, 1995.
4. Hanselman, D., Littlefield, B., Mastering Matlab® 5. A comprehensive tutorial and reference, Prentice Hall, 1998.
5. J de Leeuw, WEB Statistics Book, **http://www.stat.ucla.edu/textbook**
6. Martinez, W., Martinez, A., Computational Statistics Handbook with Matlab, Chapman&Hall/CRC, Washington DC, 2002.
7. Montgomery, D., Design and Analysis of Experiments, John Willey&Sons, Singapore, 1991.
8. Montgomery, D., Runger, G., Applied Statistics and Probability for Engineers, John Wiley&Sons, New York, 2006.
9. Petrişor, E., Probabilităţi şi statistică. Aplicaţii în economie şi inginerie, Ed. Politehnica, Timişoara, 2005.
10. Engineering Statistics Handbook : http://www.itl.nist.gov/div898/handbook/eda/eda.htm