3. CARACTERISTICI STATISTICE ALE UNEI SERII DE DATE

3.1. INTRODUCERE

Statistica matematică, mai precis metodele furnizate de aceasta s-au implementat puternic în metodologia de lucru a diferite domenii. Apelul la metodele specifice statisticii matematice s-a făcut în principal din două motive:

- existenţa variabilităţii naturale a fenomenelor, proceselor, caracteristicilor etc. aflate sub observaţie;
- necesitatea luării unor decizii asupra acestor fenomene, procese, caracteristici etc.

Statistica matematică, sintetizând o informaţie, de cele mai multe ori parţială, asupra procesului investigat, poate furniza, cu riscuri controlate de operator, baza metodologică pentru adoptarea anumitor decizii, chiar în condiţii specificate de incertitudine. În acest scop, statistica şi-a dezvoltat direcţiile principale: principii, modele şi metode.

La baza statisticii există două concepte de bază: populaţia şi eşantionul.

Populaţia statistică este obiectul de studiu şi poate fi reprezentat de mulţimea produselor ce pot rezulta dintr-un proces tehnologic, de mulţimea valorilor pe care le poate lua o caracteristică de calitate a unui produs etc. Standardele definesc populaţia statistică drept o mulţime de obiecte sau fenomene, calitativ omogene.

Eşantionul reprezintă acea parte a populaţiei asupra căreia experimentatorul aplică metode statistice propriu-zise, pentru a obţine concluzii pe care le extrapolează asupra întregii populaţii.

Această operaţie de extrapolare se numeşte inferenţă statistică. Inferenţa statistică este acea ramură a metodelor ştiinţifice de investigare a unei populaţii, care cu margini specificate de incertitudine exprimată în termeni probabilişti, face trecerea de la observaţii la concluzii privind populaţia, [20].

Inferenţa are două limite:

- informaţia pe care se bazează decizia este de natură aleatoare, fiind constituită din observaţii supuse întâmplării;
- se recunoaşte explicit nesiguranţa concluziilor.

În cazurile practice, populaţiile luate în studiu nu se pot investiga integral, ci doar prin intermediul unuia sau mai multor eşantioane, ceea ce implică apariţia unor riscuri în luarea deciziei finale asupra procesului investigat. Pentru a putea aplica anumite metode cantitative de extragere a informaţiei dorite, este necesar ca populaţia statistică analizată să poată fi reprezentată matematic într-un mod convenabil, adică trebuie să se poată găsi un model statistic al populaţiei studiate, care să încorporeze cât mai realist caracteristicile esenţiale ale populaţiei şi care să nu fie prea complicat la manevrarea analitică. Soluţia acestei probleme o constituie conceptul de variabilă aleatoare – a cărei valoare se atribuie în funcţie de diferite circumstanţe sau evenimente ce se produc într-un experiment. Ex.: timpul de bună funcţionare a unui produs, duritatea unor organe de maşini, rezistenţa la rupere a unor epruvete, o anumită cotă reprezentativă a unei piese etc. Această variabilă aleatoare caracterizează de fapt populaţia respectivă. Comportarea variabilei aleatoare este descrisă, din punct de vedere matematic, de funcţia de repartiţie asociată. Funcţia de repartiţie are o expresie specifică depinzând de tipul variabilei aleatoare: discretă (numărul de impulsuri/unitatea de timp) sau continuă (cota unei piese).

Repartiţia de probabilitate a unei variabile aleatoare discrete se exprimă, de regulă, sub forma unui tablou, în care prima linie conţine toate valorile posibile, iar a doua linie conţine probabilităţile cu care ia aceste valori:

$$X = \begin{bmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{bmatrix}, \tag{3.1}$$

cu condiţia $p_i \in [0,1]$ şi $\sum_{i=1}^{n} p_i = 1$, p_i fiind probabilitatea ca X să ia valoarea x_i. Se adoptă notaţia: $p_i = P(X = x_i)$. De obicei variabilele aleatoare discrete se asociază experimentelor ce constau din numărare.

Spre deosebire de variabilele discrete, cele continue pot lua orice valori într-un anumit interval. În consecinţă, nu se vor mai asocia probabilităţi punctuale ca în cazul anterior, ci se va defini o funcţie pozitivă, f(x), numită densitate de probabilitate, astfel încât aria domeniului cuprins între graficul ei şi axa 0x este egală cu 1. Probabilitatea ca variabila aleatoare să ia valori într-un interval (x;

x+h) este $\int_x^{x+h} f(x)dx$ este probabilitatea ca variabila considerată să ia valori în intervalul (x; x+dx) al dreptei reale (fig. 3.1).

În timpul unui proces de măsurare pot interveni trei tipuri de erori:

- erori aberante;
- erori sistematice;
- erori aleatoare.

Erorile aberante apar ca urmare a încălcării principiilor generale de măsurare sau ca rezultat al neatenţiei experimentatorului. Rezultatele afectate de erori aberante diferă esenţial de restul valorilor şi în consecinţă se pot elimina.

Erorile sistematice sunt determinate de diferiţi factori, cum ar fi: reglarea necorespunzătoare a mijlocului de măsurare, variaţia condiţiilor de mediu (temperatură, presiune, umiditate etc.). Erorile sistematice pot fi depistate şi înlăturate pe baza unor calcule ce pleacă de la principiile fizice ce stau la baza măsurării respective.

Erorile aleatoare apar datorită unei mulţimi de factori, a căror influenţă individuală este neglijabilă şi nu există posibilitatea înlăturării acestor influenţe. Studiul influenţelor erorilor aleatoare se bazează pe cunoaşterea legilor de repartiţie a acestor erori.

Dacă o măsurare se repetă de n ori, iar rezultatele se împart în clase (intervale) de lăţime Δx, se poate calcula frecvenţa relativă cu care rezultatele apar în fiecare clasă:

$$f_i = \frac{n_i}{n}, \tag{3.2}$$

unde n_i este numărul de rezultate aflat în clasa i.

Reprezentarea grafică a frecvenţei relative este o histogramă (Fig. 3.1), care pentru $n \to \infty$ şi $\Delta x \to 0$ devine o funcţie continuă, numită funcţie densitate de repartiţie sau densitate de probabilitate, deoarece prin integrarea ei pe un interval se obţine probabilitatea cu care variabila ia valori în acest interval .

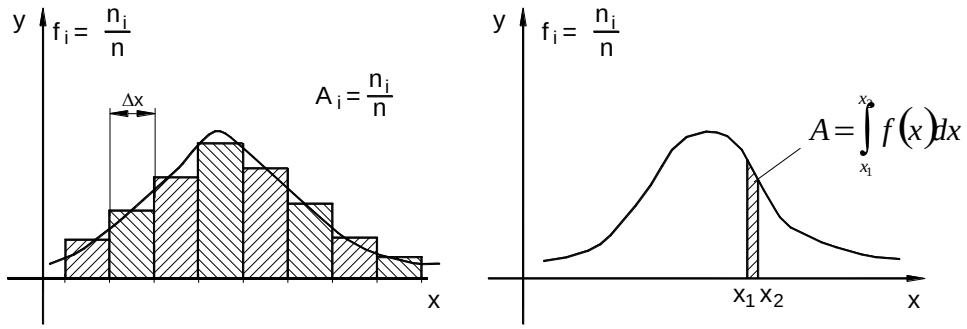


Fig. 3.1 Histograma şi densitatea de repartiţie

În mod natural se impune condiţia:

$$\int_{-\infty}^{\infty} f(x)dx = 1, \tag{3.3}$$

atunci când x ia valori pe toată dreapta reală, adică probabilitatea ca variabila să ia valori reale este 1. Relativ la o variabilă aleatoare prezintă interes probabilităţile unor evenimente de tipul $(X \in I)$ - se citeşte evenimentul X ia valori în intervalul I. Pentru a calcula aceste probabilităţi, notate $P(X \in I)$ este suficient să cunoaştem funcţia de reparţiţie a variabilei aleatoare. Funcţia care asociază oricărui număr real a, probabilitatea ca x să ia valori mai mici sau egale cu a se numeşte funcţie de repartiţie a lui X:

$$F(a) = P(x \le a), \tag{3.4}$$

Legătura dintre densitate şi repartiţie este dată de relaţia:

$$F'(x) = f(x). \tag{3.5}$$

Funcţia de repartiţie are următoarele proprietăţi, [17, 18, 20]:

- valorile funcţiei de repartiţie F(x) aparţin intervalului [0, 1], conform (3.3);
- pentru două numere reale x_1 şi x_2, cu $x_1 < x_2$, funcţia de repartiţie este nedescrescătoare, adică $F(x_1) \le F(x_2)$;
- Dacă densitatea este o funcţie continuă atunci:

$$P(x_1 \le x < x_2) = P(x_1 < x < x_2) = P(x_1 < x \le x_2) = P(x_1 \le x \le x_2). \tag{3.6}$$

Dacă variabila aleatoare este discretă funcţia ei de repartiţie are expresia:

$$F(x)=\begin{cases} 0 & x \le x_1 \\ p_1 & x_1 \le x \le x_2 \\ p_1+p_2 & x_2 < x \le x_3 \\ \dots & \dots \\ p_1+p_2+\dots+p_{n-1} & x_{n-1} < x \le x_n \\ 1 & x \ge x_n \end{cases}. \tag{3.7}$$

Pe baza proprietăţilor funcţiei de repartiţie se deduc relaţiile:

$$\begin{aligned} &P(x_1 \le x \le x_2) = F(x_2) - F(x_1), \\ &P(x \ge x_1) = 1 - F(x_1), \end{aligned} \tag{3.8}$$

relaţii utile în practica industrială:

- probabilitatea ca o variabilă să ia valori între două limite date,
- probabilitatea ca o variabilă să fie mai mare decât o valoare dată sau într-o interpretare şi mai particulară, când variabila x reprezintă numărul de ore de bună funcţionare a unui dispozitiv, aparat atunci $P(x \ge x_1)$ este probabilitatea ca acel dispozitiv, aparat să funcţioneze minimum x_1 ore.

În figura 3.2 sunt prezentate funcţia densitate de probabilitate şi funcţia de repartiţie, fiind marcate ariile ce reprezintă probabilitatea ca:

- variabila ia valori mai mici decât a;
- variabila ia valori mai mari decât b;
- variabila ia valori cuprinse în intervalul [a, b].

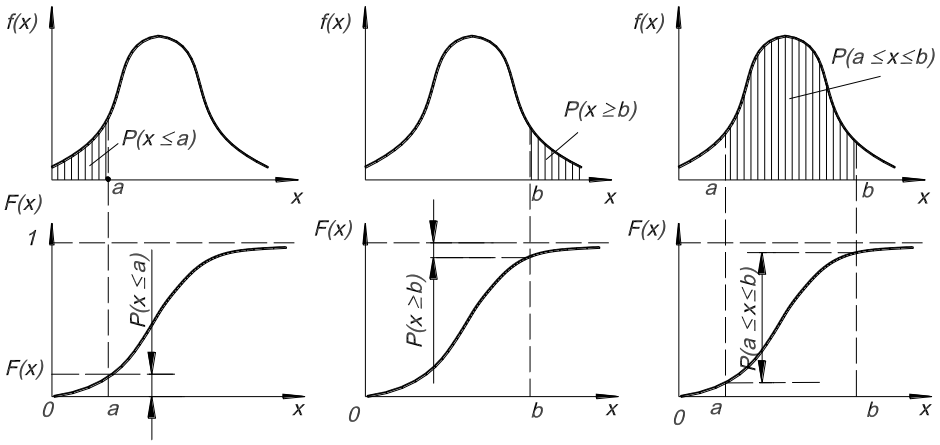


Fig. 3.2 Funcţia densitate de repartiţie şi funcţia de probabilitate

Probabilităţile reprezintă un concept esenţial în înţelegerea analizei statistice, motiv pentru care trebuie bine înţelese. Principalele aspecte ce vor fi subliniate se referă la: atribuirea probabilităţilor şi reguli ale acestora.

Probabilitatea este o măsură numerică ce cuantifică şansa unui eveniment de a se produce într-un experiment. Se măsoară pe o scală de la 0 la 1, 0 corespunzând evenimentului imposibil, iar 1 evenimentului sigur, adică evenimentul ce s-ar produce ori de câte ori s-ar repeta experimentul în aceleaşi condiţii. În plus, suma probabilităţilor tuturor evenimentelor posibile trebuie să fie egală cu 1 ($\sum P_i = 1$), în cazul când experimentul are un număr finit de evenimente ce se exclud reciproc.

Există două modalităţi de atribuire a probabilităţilor unor evenimente, în funcţie de situaţie:

- metoda frecvenţelor relative;
- metoda subiectivă.

Metoda frecvenţelor relative se bazează pe raţionamente, deoarece cu ajutorul logicii se pot determina probabilităţile de apariţie ale unor evenimente. Într-un experiment cu un număr finit de realizări mutual exclusive şi echiprobabile, probabilitatea unui eveniment este numărul cazurilor favorabile/numărul cazurilor posibile. De ex.: când se dă cu banul, există două situaţii posibile, "cap" sau "stemă", evenimente ce sunt echiprobabile (au aceeaşi şansă de apariţie), de aceea $P_{cap} = P_{stemă} = ½ = 0.5$; sau se bazează pe efectuarea unor observaţii (evidenţe empirice), situaţie când:

$$P = \frac{n_i}{N}, \tag{3.9}$$

unde n_i este numărul de evenimente favorabile, iar N numărul total de evenimente. Ex.: dacă se înregistrează numărul de ore de funcţionare a unui bec şi se constată că din 1000 de becuri testate, 100 au funcţionat mai puţin de 500 ore, atunci probabilitatea ca un bec ales aleator să aibă o durată de viaţă mai mică de 500 h este:

$$P = \frac{100}{1000} = 0{,}1.$$

Această metodă nu poate fi aplicată în orice situaţie, iar pentru evenimente viitoare se utilizează metoda subiectivă, conform căreia se atribuie probabilităţi pe baza propriilor considerente. Evident, în acest caz diferite persoane vor atribui probabilităţi diferite unui acelaşi eveniment (probabilitatea de câştigare a unui concurs).

Există numeroase situaţii când trebuie determinată probabilitatea unor evenimente legate. De exemplu, pentru două evenimente, A şi B, ne interesează dacă se vor produce ambele sau dacă va avea loc unul dintre ele. Pentru a putea răspunde la acest tip de întrebare trebuie introduse două reguli fundamentale ale probabilităţilor:

- regula de adunare a probabilităţilor;
- regula de înmulţire a probabilităţilor.

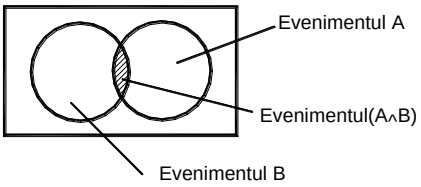
Regula de adunare se aplică în situaţia când există două evenimente şi se doreşte determinarea probabilităţii de apariţie a cel puţin unuia din cele două evenimente. Există două variante ale adunării, în funcţie de tipul evenimentelor; dacă ele se exclud reciproc sau nu. Două evenimente se exclud reciproc, dacă nu se pot produce simultan (cap sau stemă pentru o singură monedă).

Dacă evenimentele se exclud reciproc, probabilitatea de apariţie a cel puţin unui eveniment din cele două este:

$$P(A \cup B) = P(A) + P(B) \tag{3.10}$$

Dacă evenimentele nu se exclud reciproc, atunci “A sau B” înseamnă se produce A, se produce B sau se produc ambele. În acest caz, din suma anterioară trebuie scăzută probabilitatea de producere simultană, pentru a evita însumarea ei dublă (vezi fig. 3. 3):

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{3.11}$$



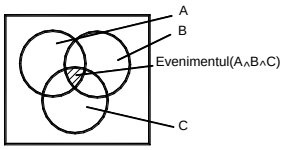


Fig. 3.3 Evenimentul $P(A\cup B)$ **Fig. 3.4 Evenimentul** $P(A\cap B\cap C)$

Această regulă poate fi generalizată pentru un număr de mai multe evenimente. De exemplu, pentru trei evenimente, ce se exclud reciproc, se obţine:

$$P(A\cup B\cup C)=P(A)+P(B)+P(C), \tag{3.12}$$

iar pentru evenimente ce nu se exclud reciproc:

$$\begin{aligned}P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(A\cap B)-P(B\cap C)-\\ -P(A\cap C)+P(A\cap B\cap C)\end{aligned} \tag{3.13}$$

Relaţia între evenimentele implicate în (3.13) este ilustrată în fig. 3.4, unde se observă că zona haşurată este adunată de trei ori (în P(A), P(B) şi P(C)), apoi scăzută de trei ori (în P(A∩B), P(B∩C) şi P(A∩C)), deci trebuie inclusă din nou la sfârşit.

Regula de înmulţire se foloseşte la găsirea probabilităţii de producere simultană a două sau mai multe evenimente (P(A∩B), P(A∩B∩C) etc.). Răspunsul la astfel de probleme depinde de tipul evenimentelor, dacă sunt independente sau nu.

Două evenimente sunt dependente când apariţia unui eveniment influenţează probabilitatea de apariţie a celuilalt. Regula de înmulţire se aplică diferit, în funcţie de tipul evenimentelor:

- pentru evenimente dependente:

$$P(A\cap B)=P(A)\times P(B|A), \tag{3.14}$$

unde P(B| A) reprezintă probabilitatea de producere a evenimentului B, ştiind că A s-a produs:

$$P(B|A)=\frac{P(A\cap B)}{P(A)}. \tag{3.15}$$

- pentru evenimente independente:

$$P(A \cap B) = P(A) \times P(B). \tag{3.16}$$

Expresia P(B| A) reprezintă o probabilitate condiţionată, adică probabilitatea ca B să fie afectat de apariţia lui A. Evident, în cazul a două evenimente independente P(B| A) = P(B). O extindere a acestei idei în contextul mai multor evenimente secvenţiale este cunoscută sub numele de regula lui Bayes. În cazul a două evenimente formula are forma (3.15). Pentru n evenimente mutual exclusive B_1, B_2, …B_n, când A a apărut:

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + ... + P(B_n)P(A|B_n)} \tag{3.17}$$

3.2. STATISTICA DESCRIPTIVĂ

Statistica descriptivă se ocupă, în principal, cu două probleme:

- prezentarea datelor sub formă tabelară şi vizualizarea lor sau a unor caracteristici prin tehnici grafice;
- utilizarea unor indicatori numerici pentru caracterizarea datelor.

Pentru a putea organiza un volum mare de date şi a le prezenta sub formă grafică sau tabelară este necesar să se detecteze eventualele tendinţe de distribuire a lor pe axa reală prin intermediul unor mijloace de analiză. Cele mai utilizate metode grafice de reprezentare sunt: tabele de distribuţie de frecvenţe, pie-chart-uri, histograme, poligoane de frecvenţă, curbe de frecvenţă cumulată etc

3.2.1. ORGANIZAREA DATELOR

Pentru a putea reprezenta tabelele de frecvenţă este necesar ca datele să fie grupate în câteva clase sau intervale semnificative (nici prea multe, nici prea puţine) şi să se contorizeze numărul de valori din fiecare clasă, iar apoi să se calculeze frecvenţele relative, adică numărul datelor din fiecare clasă raportat la numărul total. În general, se recomandă ca numărul claselor să fie între 5 – 15 (în funcţie de mărimea eşantionului avut la dispoziţie) şi cel mai adesea mărimea claselor este egală. De asemenea, trebuie ca fiecare valoare existentă în şir să fie introdusă o singură dată (se stabileşte o convenţie pentru

limitele intervalelor).

În tabelul 3.1 se prezintă un set de date experimentale organizat cu ajutorul tabelului de frecvenţă.

Tabelul 2.1. Tabel de frecvenţă

Date	10.824	10.827	10.775	10.826	10.771	10.802	10.816	10.816	10.829
	10.814	10.815	10.801	10.813	10.821	10.812	10.860	10.850	10.865
	10.810	10.854	10.847	10.844	10.796	10.858	10.797	10.843	10.839
Date ordonate	10.771	10.775	10.796	10.797	10.801	10.802	10.810	10.812	10.813
	10.814	10.815	10.816	10.816	10.821	10.824	10.826	10.827	10.829
	10.839	10.843	10.844	10.847	10.850	10.854	10.858	10.860	10.865

Clase	Frecvenţa absolută	Frecvenţa relativă [%]
10.770 – 10.7867	2	7.40
10.7867 - 10.8023	4	14.81
10.8023 - 10.8180	7	25.92
10.8180 - 10.8337	5	18.51
10.8337 - 10.8493	4	14.81
10.8493 - 10.8650	5	18.51

După construirea tabelului de frecvenţă, în general, pasul următor îl constituie reprezentarea sub o anumită formă grafică.

Un pie-chart este o reprezentare grafică sub forma unui cerc, unde frecvenţele relative sunt utilizate pentru divizarea cercului în sectoare de cerc corespunzătoare fiecărei categorii de variabile. Acest tip de reprezentare se pretează în cazul când datele sunt grupate pe diverse categorii. Ex.: Totalul studenţilor din Facultatea de Mecanică clasificat în studenţi integralişti, studenţi ce au acumulat înter 50 – 60 credite, 40 – 50 credite, mai puţin de 40 credite.

Cel mai utilizat mod de reprezentare îl constituie histogramele. În fig. 3.5 se prezintă histograma aferentă datelor din tabelul 3.1. În cazul când intervalele nu sunt echidistante trebuie ajustată înălţimea dreptunghiurilor astfel încât aria fiecărui dreptunghi să fie proporţională cu frecvenţa corespunzătoare.

Funcţiile din Matlab asociate cu histograma sunt:

- **`N=hist(x)`** – returnează un vector linie, ce conţine numărul de date din vectorul x, aflate în 10 intervale echidistante;

- **`N=hist(x,m)`** – returnează un vector linie, ce conţine numărul de date din vectorul x, aflate în m intervale echidistante;
- **`N=histc(x,limite)`** – contorizează valorile lui x aflate între elementele vectorului limite, ce trebuie sa fie un vector monoton crescător. N va avea dimensiunea cu o unitate mai mică decât vectorul limite, iar N(k) va număra valorile x(i) dacă limite(k)<=x(i)<limite(k+1). Ultimul interval va contabiliza şi valorile egale cu limita superioară;
- **`hist(x)`** – trasează histograma utilizând 10 intervale sau clase;
- **`hist(x,m)`** - trasează histograma utilizând m clase.

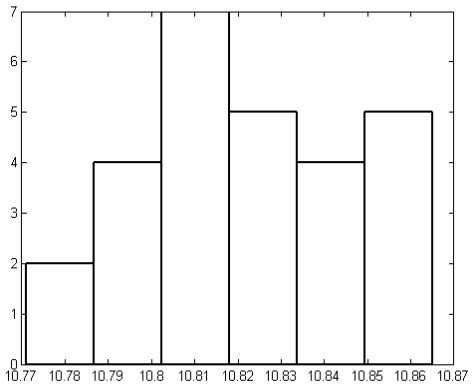


Fig. 3.5. Histograma datelor

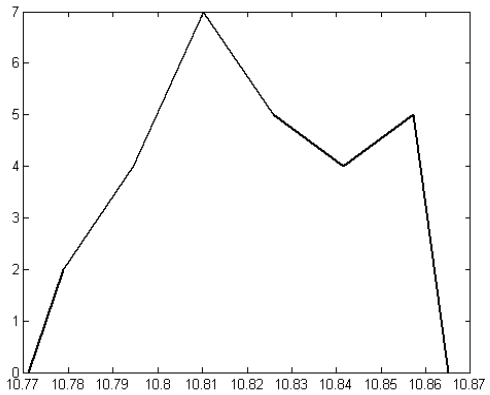


Fig. 3.6. Poligonul de frecvenţe

Poligoanele de frecvenţă se utilizează mai ales în cazul când se doreşte compararea a două distribuţii în aceeaşi reprezentare grafică. Ele se obţin prin reprezentarea frecvenţelor în dreptul mijlocului fiecărei clase şi unirea acestor puncte prin linii drepte. În plus, la extremităţi, primul, respectiv ultimul punct corespunde valorii minime, respectiv maxime din şirul de date. În fig. 3.6. este prezentat poligonul de frecvenţe al datelor cuprinse în tabelul 3.1.

O altă variantă de reprezentare este curba frecvenţelor cumulate, ce permite aflarea răspunsului la întrebări de tipul: câte observaţii sunt egale sau mai mici decât limita superioară a fiecărei clase, respectiv câte sunt mai mari sau egale cu limita inferioară a fiecărei clase. După cum indică şi numele, curba frecvenţelor cumulate se obţine prin reprezentarea frecvenţelor cumulate ale claselor. Cumularea se poate face de la valoarea minimă spre maximă sau invers. În primul caz se obţine o curbă ascendentă, ce oferă informaţii privind

numărul de observaţii mai mic cel mult egal cu limita superioară a fiecărei clase, iar în cazul al doilea, se obţine o curbă descendentă, pe baza căreia se pote determina numărul de observaţii mai mare cel mult egal cu limita inferioară a fiecărui interval. În fig. 3.7 se prezintă curba frecvenţelor cumulate corespunzătoare datelor prezentate în tabelul 3.1.

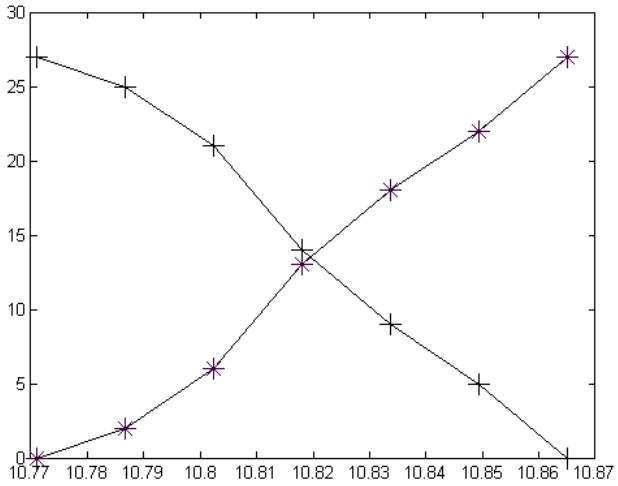


Fig. 3.7. Curba frecvenţelor însumate

3.2.2. INDICATORI NUMERICI DE CARACTERIZARE A DATELOR

Există două caracteristici esenţiale ce se investighează în cazul seturilor de date:

- centrarea sau localizarea;
- concentrarea sau împrăştierea.

Fie o serie de date X: x_1, x_2, … x_n obţinută în urma unui proces de măsurare. Tendinţa centrală a seriei poate fi caracterizată prin: medie, mediană, modul.

Cel mai utilizat indicator de centrare este media. Media unei serii de date este media aritmetică a valorilor sale:

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right) = \frac{1}{n}\sum_{i=1}^{n} x_i \ . \tag{3.18}$$

Media reprezintă centrul de greutate al seriei de date. În Matlab, media seriei

de date, stocate în vectorul x, se calculează cu funcţia **`mean(x)`**.

Mediana este valoarea centrală a unui set de date ordonate crescător. Ea se calculează cu formula:

$$x_{me} = x_{\frac{n+1}{2}}, \quad n = 2k+1$$
$$x_{me} = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right), \quad n = 2k \qquad (3.19)$$

Prima formulă se aplică pentru un număr impar de elemente ale seriei de date, iar a doua când volumul seriei este uj număr par. Mediana caracterizează mai bine valoarea centrală unui serii de date în situaţia când setul de date este asimetric, respectiv apare o concentrare de date la una din extremităţile seriei. În Matlab mediana seriei de date stocate în vectorul x se calculează cu funcţia **`median(x)`**.

Modulul este observaţia cu cea mai mare frecvenţă de apariţie. Există seturi de date cu un singur modul (unimodale) sau cu mai multe (multimodale). Modulul se calculează cu formula:

$$x_{mo} = \bar{x} + 3\left(x_{me} - \bar{x}\right). \qquad (3.20)$$

În cazul unor seturi de date simetrice, media, mediana şi modulul coincid. La repartiţiile asimetrice se poate întâlni cazul $\bar{x} < x_{me} < x_{mo}$ sau $x_{mo} < x_{me} < \bar{x}$ (fig. 3. 8).

Pentru caracterizarea variabilităţii datelor se utilizează o serie de indicatori de concentrare sau împrăştiere. Printre cei mai uzuali se numără: intervalul de variaţie, deviaţia standard, varianţa, coeficientul de variaţie etc.

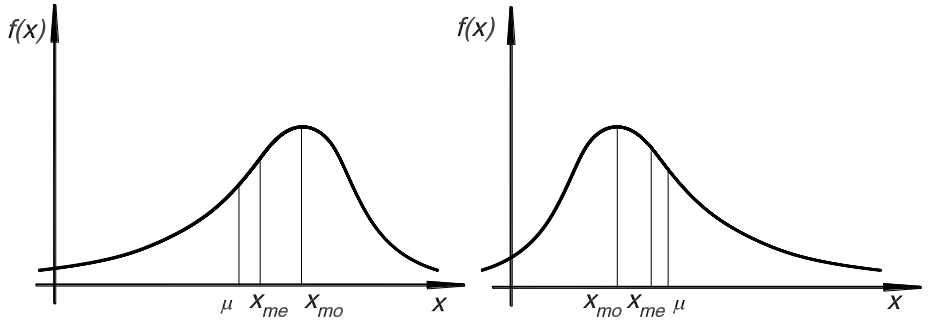


Fig. 3.8 Repartiţii asimetrice

Intervalul de variaţie este diferenţa dintre valoarea maximă şi minimă a seriei

de date. Deoarece acesta se bazează doar pe două valori aparţinătoare seriei de date, în multe situaţii se apelează la p-cvantile.

P-cvantila ($p \in (0,1)$) seriei de date este un număr real, q_p, cu proprietatea că 100p% valori ale seriei sunt mai mici sau egale decât q_p, iar (1 – p)100% valori sunt mai mari sau egale decât q_p. Numărul q_p poate să aparţină seriei sau nu.

Pentru determinarea p-cvantilei unei serii de date se procedează în felul următor [14]:

- se ordonează crescător setul de date X': x_1', x_2', … x_n';
- se calculează partea întreagă a numărului n*p + 0.5, care se notează i;
- p-cvantila este:

$$q(p) = x_i' + \left(x_{i+1}' - x_{i-1}' \right)\{np + 0.5\}, \tag{3.21}$$

unde {np + 0.5} este partea fracţionară a numărului np + 0.5.

În cazul când p = 0.5, $q_{0.5}$ este chiar mediana seriei de date, deoarece i = [n·0.5+0.5] = [(n+1)/2], care pentru n egal cu un număr impar este (n+1)/2, deci 0.5-cvantila este elementul şirului ordonat $x_{(n+1)/2}$, iar în cazul când n este un număr par, 0.5-cvantila devine $\frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n+1}{2}} \right)$, deci se obţin chiar expresiile medianei.

În cazul unor seturi de date de volum mai mare se utilizează procentilele. Procentila 100p, $p \in (0,1)$, este valoarea reală x, cu proprietatea că cel mult 100p% din valoriile seriei de date sunt mai mici decât x şi cel mult 100(1 – p)% sunt mai mari. Dacă p(n+1) este o valoare întreagă, procentila 100p este valoarea de rang p(n+1) din seria ordonată, iar în cazul când p(n+1) nu este număr întreg, procentila seriei este valoarea reală obţinută prin interpolare din valoriile poziţiilor adiacente [14].

În Matlab, există funcţia **`prctile(x,p)`** ce calculează procentila serie de date x, iar $p \in (0,1)$. Pentru seria de date prezentată în tabelul 2.1, procentila 0.3 este 10.8126, iar procentila 0.5, adică mediana, este 10.8210.

Printre cei mai uzuali indicatori de concentrare sau împrăştiere a valorilor seriei de date în jurul mediei se numără dispersia sau varianţa unei serii de date. Având o serie de date x: x_1, x_2, ... x_n cu media $\bar{x}$, dispersia este:

$$S^2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 . \quad (3.22)$$

Pentru o serie statistică:

$$Y = \begin{pmatrix} x_1 & x_2 & \dots & x_k \\ n_1 & n_2 & \dots & n_k \end{pmatrix} \quad \sum_{j=1}^{k} n_j = N \quad (3.23)$$

n_i indică de câte ori apare x_i în seria de date, $\bar{x}$ este media seriei, dispersia este:

$$S^2(Y) = \frac{1}{N-1}\sum_{j=1}^{k}(x_j - \bar{x})^2 n_j . \quad (3.24)$$

În Matlab, varianţa unei serii de date se poate calcula cu funcţiile **`var(x)`**, ce calculează dispersia seriei de date cu formula (3.22), respectiv **`var(x,1)`** ce aplică formula:

$$S^2(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 . \quad (3.25)$$

Având o serie de date cu dispersia $S^2(x)$, valoarea $S = \sqrt{S^2(x)}$ se numeşte abatere standard sau abatere medie pătratică a seriei x. În Matlab, abaterea standard se calculează cu funcţiile **`std(x)`**, respectiv **`std(x,1)`.**

Un indicator al concentrării relative este coeficientul de variaţie. Având o serie de date x, cu media $\bar{x}$ şi abaterea standard S, acesta are expresia:

$$CV(x) = \frac{100S}{\bar{x}} . \quad (3.26)$$

Cu cât coeficientul de variaţie este mai apropiat de 0, seria este mai omogenă şi media $\bar{x}$ mai reprezentativă. Dacă valoarea coeficientului de variaţie tinde spre 100, împrăştierea seriei este mare şi media este mai puţin reprezentativă.

Ultimii indicatori menţionaţi sunt coeficientul de asimetrie (γ_1) şi coeficientul de exces (γ_2):

$$\gamma_1 = \frac{m_3}{S^3}, \quad \gamma_2 = \frac{m_4}{S^4}, \tag{3.27}$$

unde m_3 şi m_4 sunt momentele centrate de ordinul 3, respectiv 4. Momentul centrat de ordinul k se calculează cu formula:

$$m_k = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^k . \tag{3.28}$$

Asimetria şi excesul ajută la identificarea formei de repartiţie a datelor experimentale. Etalonul la această comparaţie este repartiţia normală, pentru care $\gamma_1 = 0$ şi $\gamma_2 = 3$. Excesul mai poartă denumirea de "kurtosis". În funcţie de valoarea coeficientului de exces (pozitiv, negativ sau nul) se pot trage concluzii asupra alurii curbei densităţii de repartiţie (Fig. 3.9).

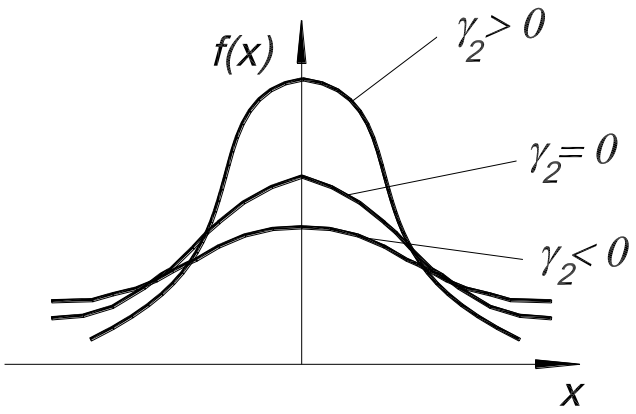


Fig. 3.9 Variaţia formei de repartiţie faţă de coeficientul de exces

Dacă $\gamma_2 = 0$, repartiţia se numeşte mezokurtică, având o formă apropiată de repartiţia normală. Dacă $\gamma_2 > 0$, repartiţia este leptokurtică, fiind mai ascuţită decât cea normală, iar dacă $\gamma_2 < 0$, repartiţia este platokurtică, fiind mai aplatizată.

Coeficientul de asimetrie caracterizează simetria repartiţiei. Repartiţiile simetrice au $\gamma_1 = 0$. În cazul $\gamma_1 > 0$, repartiţia este pozitiv asimetrică, iar în cazul $\gamma_1 < 0$, repartiţia este negativ asimetrică.

În Matlab, momentul centrat de ordinul k al seriei de date stocate în vectorul x

se calculează cu funcţia **`moment(x,k)`**, coeficientul de asimetrie se determină cu funcţia **`skewness(x)`**, iar coeficientul de exces cu **`kurtosis(x)`**.

3.2.3. METODE DE AFIŞARE SAU REPREZENTARE GRAFICĂ ÎN ANALIZA PRIMARĂ A DATELOR

O metodă de afişare şi analiză primară a datelor este diagrama tulpină – frunze (stem and leaf). Ea se pretează la serii mici de date ce au acelaşi ordin de mărime. A fost introdusă de Tukey în 1977 ca o metodă de afişare a datelor într-o listă structurată. În cazul unui set de date ce conţine valori cu minim două cifre se poate construi o astfelde diagramă. Se separă fiecare valoare a şirului de date în două părţi : tulpina şi frunza. Tulpina reprezintă prima cifră a datelor, iar frunza cea de-a doua, respectiv ultima. În cazul valorii 34, 3 este tulpina, iar 4 este frunza; în cazul valorii 126, tulpina este 12, iar frunza este 6.

Fie seria de date x: 54, 25, 43, 3, 28, 39, 78, 32, 54, 93, 27, 33, 22, 78, 75, 83, 62, 76, 77, 67, 77, 80, 4, 26, 18, 10, 34, 30, 36, 43, 78, 41, 24, 91, 90, 63, 87, 55, 60, 49, 37, 39, 51, 66, 10, 76, 34, 47, 51, 34.

Se constată ca fiecare valoare este de ordinul zecilor sau al unităţilor. În acest caz, tulpina este cifra zecilor, iar cifra unităţilor reprezintă frunza. Se ordonează crescător seria. Diagrama tulpină – frunze este constituită din trei coloane. Pe coloana a doua se trec în ordine crescătoare tulpinile, iar în ultima se înregistrează în ordine crescătoare toate frunzele aparţinătoare tulpinei respective (în acest caz cifrele unităţilor). În prima coloană se indică frecvenţele cumulate de la prima clasă până la clasa medianei, respectiv de la ultima clasă până la clasa medianei. Frecvenţa clasei medianei se introduce în paranteză. Diagrama setului de date se prezintă în fig. 3. 10.

Frecvenţe cumulate	Tulpina	Frunze
2	0	34
6	1	0058
12	2	245678
22	3	0234446799
(5)	4	13379
23	5	1445
19	6	02367
14	7	56677888
6	8	037
3	9	013

Fig. 3.10 Diagrama tulpină – frunze

Acest tip de afişare se aseamănă cu histograma, dar pe lângă distribuţia datelor sunt indicate şi valorile respectivei serii. De asemenea, în urma unei astfel de afişări se poate decide modul de divizare a setului de date în clase, în sensul stabilirii mărimii şi numărului acestora.

O metodă de reprezentare grafică este box-plot-ul. El se construieşte pe baza a 5 valori asociate seriei de date: minimul, cvantila inferioară q(0.25), mediana, cvantila superioară q(0.75) şi maximul. Aceste valori pot fi reprezentate pe o axă verticală sau orizontală. Ambele sunt ilustrate în fig. 3.11.

Se desenează un dreptunghi având o latură egală cu diferenţa dintre cvantila superioară şi cea inferioară. În acest dreptunghi se mai trasează o dreaptă paralelă cu latura menţionată, în dreptul valorii medianei. Lungimea laturii dreptunghiului, q(0.75) – q(0.25), este cunoscută sub numele de interval intercvantilic (IQR). Acest interval se multiplică cu un coeficient, de regulă 1.5 şi se determină valoarea q(0.25) - 1.5·IQR şi q(0.75) + 1.5·IQR. Se trasează două segmente ce unesc mijloacele bazelor dreptunghiului cu aceste valori. Aceste segmente se numesc mustăţi (whiskers). datele situate în afara intervalului [q(0.25) – 1.5·IQR; q(0.75) + 1.5·IQR] sunt considerate aberante. Alegerea coeficientului 1.5 nu este standardizată, ea poate fi impusă de utilizator în funcţie de natura datelor şi a experienţei precedente din analiza datelor de un anumit tip.

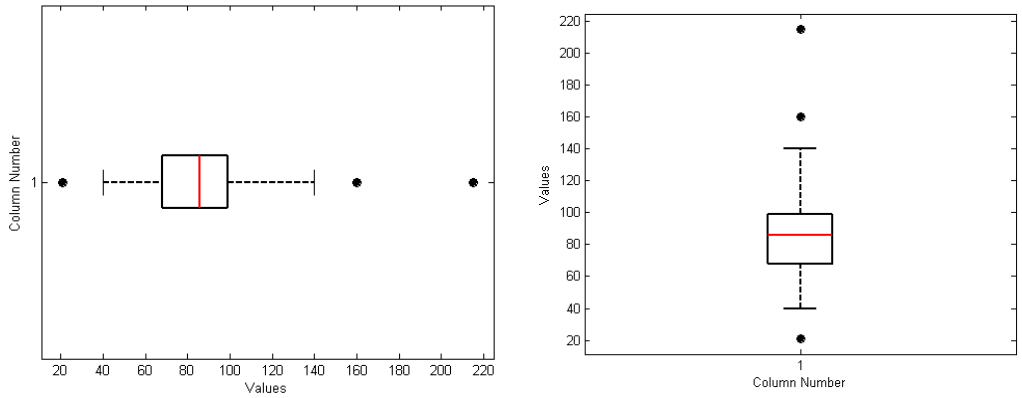


Fig. 3.11 Box-plot

În Matlab, un box-plot se poate trasa cu comanda **`boxplot(x,cod-box,simbol,vertic,whisker)`**, unde x este setul de date, ce poate fi un vector sau o matrice. În cazul când x este o matrice se va trasa câte un box-plot pentru fiecare coloană. **`Cod-box`** este o variabilă ce pote lua valoarea 0

sau 1, dacă este 0 se va desena un dreptunghi, dacă este 1 se va desena un dreptunghi din care se decupează un V. **`simbol`** este caracterul cu care se marchează valoriile din exteriorul intervalului [q(0.25) – 1.5·IQR; q(0.75) + 1.5·IQR] şi este implicit '+'. **`vertic`** este o variabilă ce poate lua valoarea 0 (trasarea se face pe orizontală) sau 1 (trasarea se face pe verticală). **`whisker`** defineşte intervalul, este valoarea cu care se înmulţeşte IQR şi este implicit 1.5. Dacă **`whisker = 0`** se vor marca toate valoriile aflate în afara intervalului [q(0.25), q(0.75)].

În Matlab, intervalul intercvantilic al unei serii de date se poate calcula cu funcţia **`iqr(x)`**.

Acest tip de reprezentare permite indicarea prezumtivelor valori aberante dintr-o serie de date.

3.3. TIPURI DE REPARTIŢII

Determinarea probabilităţilor asociate unor evenimente aleatoare poate fi mult simplificată dacă se construieşte un model matematic ce descrie cu acurateţe situaţiile asociate cu anumite evenimente de interes. Un astfel de model utilizat la determinarea probabilităţilor de producere a unor evenimente este o distribuţie de probabilitate.

Unui experiment aleator în care se determină (prin măsurare sau observare) anumite caracteristici ale unei populaţii, ce pot fi cuantificate numeric, i se asociază una sau mai multe variabile, numite variabile aleatoare. O variabilă aleatoare ce poate lua valori într-o mulţime numărabilă se numeşte variabilă discretă, iar în situaţia când poate lua orice valoare într-un interval real se numeşte variabilă aleatoare continuă.

Comportarea variabilei aleatoare este descrisă din punct de vedere matematic de distribuţia de probabilitate asociată. Distribuţia de probabilitate a unei variabile aleatoare discrete se exprimă sub forma unei matrice:

$$x = \begin{bmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{bmatrix}, \tag{3.29}$$

pe prima linie fiind înregistrate valorile variabilelor, iar pe a doua probabilităţile de realizare, cu condiţiile: $p_i \in [0,1], \sum_{i=1}^{n} p_i = 1.$

Pornind de la evenimente elementare ($X = x_i$) ale unui experiment se pot

determina probabilităţi ale evenimentelor de tipul: $(X \le a), (X \ge a), (a \le X \le b), (a < X < b)$ etc. Aceste probabilităţi se obţin prin însumarea probabilităţilor evenimentelor elementare corespunzătoare, deoarece acestea se exclud reciproc.

Exemplu: Fie x variabila aleatoare ce indică numărul de restanţe înregistrate la sfârşitul unui semestru de studenţii unui an de studiu:

$$x = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 0.05 & 0.3 & 0.25 & 0.15 & 0.1 & 0.05 & 0.05 & 0.05 \end{bmatrix}.$$

a) Să se determine probabilitatea ca un student să aibă maximum 3 restanţe.

b) Dacă un student are minim 3 restanţe, care este probabilitatea ca el să aibă 4 restanţe.

Evenimentul a cărui probabilitate se cere în cazul a) este:

$$(X \le 3) = (X = 0) \cup (X = 1) \cup (X = 2) \cup (X = 3).$$

Deci:

$P(X \le 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 0.05 + 0.3 + 0.25 + 0.15$ = 0.75.

În cazul b) se notează A = (X = 4), evenimentul ca studentul să aibă 4 restanţe şi B = (X ≥ 3), evenimentul ca studentul să aibă minim 3 restanţe. În problemă se cere probabilitatea evenimentului A condiţionată de B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

P(B) = P(X=3) + P(X=4) + P(X=5) + P(X=6) + P(X=7) = 0.15 + 0.1 + 0.05 + 0.05 + 0.05 = 0.4

P(A ∩ B)=P(A)=0.1

P(A|B)=0.1/0.4=0.25

În cazul variabilelor aleatoare discrete funcţia de repartiţie definită prin F(X) = P(X ≤ x) – probabilitatea ca variabila aleatoare X să ia valori mai mici sau egale cu x - este de forma:

$$F(X)=\begin{cases} 0 & X \le x_1 \\ p_1 & x_1 < X \le x_2 \\ p_1 + p_2 & x_2 < X \le x_3 \\ \dots & \dots \\ p_1 + p_2 + \dots + p_{n-1} & x_{n-1} < X \le x_n \\ 1 & X \ge x_n \end{cases}. \quad (3.30)$$

După cum se extrag informaţii dintr-o serie de date prin asocierea unor indicatori numerici, în acelaşi mod se definesc caracteristici pentru variabile aleatoare.

Valoarea medie a unei variabile aleatoare discretă, X, de valori x_i şi probabilităţile $p_i = P(X=x_i)$, i=1, n este numărul notat M(X):

$$M(X)=\sum_{i=1}^{n} x_i p_i \quad (3.31)$$

Fie o variabilă aleatoare discretă X, ce are media m. Dispersia sau varianţa variabilei este:

$$\sigma^2(X)=M\left((X-m)^2\right), \quad (3.32)$$

iar abaterea standard $\sqrt{\sigma^2(X)}$ şi se notează $\sigma(X)$.

Calcularea acestor indicatori este exemplificată în exemplul următor [14]. Fie distribuţia vârstei populaţiei României în 1990 şi 2000. Se raportează vârsta pe categorii, în intervale, reprezentate de mijloacele acestora. Pentru fiecare interval se indică procentul populaţiei ce are vârsta în acel interval:

Interval	Mijloc interval	1990	2000
sub 5 ani	3	7.6	6.4
5-13	9	12.8	11.6
14-17	16	5.3	5.2
18-24	21	12.8	9.0
25-34	30	17.3	12.5
35-44	40	15.1	12.2
45-64	55	18.6	22.5
65-84	75	9.3	16.0
≥85	92	1.2	4.6

Interpretând procentele ca probabilităţi, distribuţia vârstei în 1990 este:

$$X = \begin{bmatrix} 3 & 9 & 16 & 21 & 30 & 40 & 55 & 75 & 92 \\ 0.076 & 0.128 & 0.053 & 0.108 & 0.173 & 0.151 & 0.186 & 0.093 & 0.012 \end{bmatrix}.$$

Vârsta medie a populaţiei României în 1990 era:

M(X)=3(0.076)+9(0.128)+...+92(0.012)=34.455, iar abaterea standard:

$$\sigma(X) = \sqrt{(3-34.455)^2 + ... + (92-34.455)^2} = 21.87.$$

Considerând distribuţia de probabilitate a vârstei populaţiei în 2000, se obţine media M(Y) = 41.2 şi abaterea standard σ(Y) = 25.4. Comparând valorile corespunzătoare se observă că media vâstei creşte, dar şi variabilitatea creşte.

3.3.1. DISTRIBUŢIA BINOMIALĂ

Distribuţia binomială caracterizează variabile discrete, ce se asociază unui proces Bernoulli. Procesul Bernoulli constă din n încercări ale unui experiment, ce are rezultate ce se exclud reciproc (mutual exclusive): succes sau eşec. Încercările sunt independente, adică rezultatul uneia nu influenţează rezultatul celeilalte, iar pentru fiecare încercare probabilitatea succesului este aceeaşi, p.

Într-un experiment Bernoulli prezintă interes numărul de succese şi eşecuri din n încercări. Dacă se notează X variabila aleatoare discretă ce reprezintă numărul de reuşite din n încercări, atunci X va putea lua valorile discrete 0, 1, ... n. Probabilităţile asociate fiecărei valori i vor avea o distribuţie de frecvenţe de tip binomial.

Probabilitatea de a înregistra k succese din n încercări, cu probabilitatea de succes la o încercare p şi de eşec 1 – p este [5,14]:

$$P(X = k) = C_n^k p^k (1-p)^{n-k}, \tag{3.33}$$

unde $C_n^k = \frac{n!}{k!(n-k)!}$, reprezintă combinări de n luate câte k.

Exemple de experimente ce pot fi modelate cu o repartiţie binomială:

- Un medicament are probabilitatea de 90% de a vindeca o anumită maladie. Medicamentul se administrează la 100 de pacienţi, iar în final aceştia sunt vindecaţi sau nu. Dacă X este numărul de pacienţi

vindecaţi, X este o variabilă aleatoare binomială cu parametrii (100; 0.9); adică n = 100, p = 0.9.

- Institutul Naţional de Statistică estimează că există şansa ca 20% din adulţi să fie simpatizanţi al unui anumit partid. 50 de adulţi sunt selectaţi aleator. Dacă X reprezintă numărul adulţilor simpatizanţi ai respectivului partid, X va fi o variabilă aleatoare binomială cu parametrii (50; 0.20).
- Un producător de plăci de calculator are în medie 5% produse defecte. Pentru a monitoriza procesul de producţie se extrage un eşantion de 75 de elemente. Dacă acest eşantion prezintă mai mult de 5 unităţi de produs defecte, procesul de producţie se opreşte. Numărul de produse defecte se poate modela cu o variabilă aleatoare binomială cu parametrii (75; 0.05)

Aplicaţie: 10 muncitori lucrează la un atelier şi 6 din ei au categoria a 5-a. Care este probabilitatea ca într-un grup format din 3 muncitori să fie toţi de categoria a 5-a.

$$P(X=3)=C_{10}^{3}\left(\frac{6}{10}\right)^{3}\left(\frac{4}{10}\right)^{10-3}=120\cdot 0.126\cdot 0.00164=0.425$$

Există o probabilitate de 42.5% ca toţi cei trei muncitori să fie de categoria a cincea.

În numeroase situaţii practice, trebuie calculată probabilitatea de succes care să fie mai mare cel mult egală cu o valoare dată, respectiv mai mică sau egală decât o valoare dată: $P(x \geq k)$, $P(x \leq k)$. În astfel de situaţii se însumează probabilităţile corespunzătoare.

Pentru exemplul anterior să se determine probabilitatea ca nu mai mult de doi muncitori selectaţi să fie de categoria a cincea.

$$P(X\leq 2)=P(X=0)+P(X=1)+P(X=2)=C_{10}^{0}\left(\frac{6}{10}\right)^{0}\left(\frac{4}{10}\right)^{10}+C_{10}^{1}\left(\frac{6}{10}\right)^{1}\left(\frac{4}{10}\right)^{19}+$$

$$+C_{10}^{2}\left(\frac{6}{10}\right)^{2}\left(\frac{4}{10}\right)^{8}=0.0001+0.0016+0.0106=0.0123.$$

În cazul distribuţiei binomiale media şi varianţa sunt:

$$M(x) = np, \tag{3.34}$$

$$\sigma^2(x) = np(1-p).$$

În Matlab există funcţia **`binocdf(a,n,p)`** ce returnează funcţia de repartiţie F(a) a unei variabile aleatoare binomiale de parametri cunoscuţi n şi p. Pentru exemplul anterior se putea calcula probabilitatea ca nu mai mult de doi muncitori să fie de categoria a cincea, P(x≤2), cu **`binocdf(2,10,.6)`**, ce returnează valoarea 0.0123.

Funcţia densitate de probabilitate pentru repartiţia binomială este definită de **`binopdf(x,n,p)`** care returnează valoarea acestei funcţii în punctul x. Rezultatul anterior putea fi obţinut şi prin însumarea funcţiilor densitate de probabilitate pentru valoriile: 0, 1 şi 2;

```
prob=binocdf(0:2,10,0.6)
```

returnează aceeaşi valoare 0.0123.

3.3.2. DISTRIBUŢIA POISSON

Este o distribuţie ce caracterizează variabile discrete. Ea tratează problema evenimentelor aleatoare ce au loc în unitatea de timp sau spaţiu. Ea poate înlocui distribuţia binomială în cazul când numărul de evenimente total este foarte mare, iar şansa de realizare favorabilă foarte mică. Principala utilizare a distribuţiei Poisson este la probleme ce tratează evenimente rare, ce apar într-un interval de timp specificat sau într-o regiune din spaţiu. Ex.: numărul de clienţi dintr-un magazin/oră, numărul de maşini ce intră într-o parcare/zi, numărul de scurgeri de-a lungul unei conducte de petrol, numărul de defecte/unitatea de suprafaţă la o tablă etc.

Distribuţia Poisson este aplicabilă în anumite condiţii:

- evenimentele aleatoare au loc în unitatea de timp sau spaţiu;
- numărul de evenimente favorabile trebuie să fie teoretic infinit;
- producerea unui eveniment este independentă de producerea evenimentelor anterioare sau posterioare.

Dacă aceste condiţii sunt îndeplinite, variabila aleatoare ce contorizează numărul de apariţii a evenimentului într-un interval de timp este o variabilă aleatoare de tip Poisson. În cazul când evenimentele se produc astfel încât în medie apar λ evenimente într-o perioadă de timp sau spaţiu, atunci

probabilitatea ca evenimentul să se producă de k ori este:

$$P(X=k)=\frac{e^{-\lambda}\lambda^{k}}{k!}. \tag{3.35}$$

Pentru calcularea probabilităţii ca evenimentele să se producă de cel mult k ori, se procedează la fel ca în cazul distribuţiei binomiale:

$$P(X\le k)=P(X=0)+P(X=1)+\ldots+P(X=k)$$

Dacă X este o variabilă aleatoare de tip Poisson, cu parametrul λ, atunci media sa este:

$$M(X)=\lambda, \tag{3.36}$$

iar dispersia:

$$\sigma^{2}(X)=\lambda. \tag{3.37}$$

Aplicaţii:

1. Într-un studiu la o spălătorie de maşini s-a constatat că numărul mediu de maşini ce sosesc luni dimineaţa între ora 8 şi 9 este 5. Care este probabilitatea ca într-o anumită zi de luni să sosească exact 5 maşini? Dar care este probabilitatea să vină mai puţin de 3 maşini?

$$P(X=5)=\frac{e^{-5}5^{5}}{5!}=0.1755.$$

În primul caz probabilitatea este 17.55%.

$$P(X<3)=P(X=0)+P(X=1)+P(X=2)=0.0067+0.0337+0.0842=0.1246$$

În cel de-al doilea caz, probabilitatea este de 12.46%.

2. La editarea unei cărţi se utilizează corectarea automată oferită de soft-ul de procesare a textului, în plus la editare se face o nouă corectură. Cu toate acestea un anumit număr de greşeli de editare rămân. Se consideră că numărul de erori tipografice per pagină are o repartiţie Poisson cu parametrul λ = 0.25. Se doreşte calcularea probabilităţii ca pe o pagină să apară cel puţin 2 erori.

$$P(X\ge 2)=1-P(X=0)-P(X=1)=1-e^{-.25}-.25e^{-.25}=0.0265$$

Funcţia de repartiţie Poisson se poate apela în Matlab cu comanda **poisscdf**. Sintaxa este **poisscdf(x,lambda)** - returnează probabilitatea ca variabila să fie mai mică decât o valoare x. Deci în cazul problemei anterioare, rezultatul se poate obţine cu:

```
probabilitate=1-poisscdf(1,0.25)
```

3. Într-o intersecţie s-a constatat că apar în medie 2 accidente pe săptămână (λ = 2). Să se determine care este probabilitatea ca în următoarele 2 săptămâni să aibă loc 3 accidente.

În cazul distribuţiei Poisson numărul de evenimente ce au loc într-un interval depind doar de lungimea intervalului şi sunt independente de punctul de început. Numărul de evenimente ce apar în mai multe intervale de timp, k, este egal cu kλ. În aceste condiţii se poate determina probabilitatea de apariţie a 3 accidente în 2 săptămâni ca fiind:

```
probabilitate=poisscdf(3,2*2)
```

3.3.3. DISTRIBUŢIA UNIFORMĂ

Variabilele aleatoare discrete se asociază unor experimente ce constau în contorizarea anumitor rezultate. Variabilele continue se asociază, în general, unor experimente ce constau din măsurare, deci au rezultate numere reale, valorile variabilei putând fi oricare într-un interval mărginit sau nu. Ex.: cota unei piese, înălţimea sau greutatea populaţiei, rata şomajului etc.

După cum s-a arătat în paragraful 3.1, dacă se cunoaşte funcţia de repartiţie a variabilei aleatoare continue, X, se poate calcula probabilitatea de apariţie a oricărui eveniment: $X < a$, $a < X < b$, $X > b$, $a \le X \le b$ etc.

Fie X o variabilă aleatoare continuă, ce are densitatea de probabilitate f. valoarea medie se notează M(X) şi se calculează:

$$M(X) = \int_{-\infty}^{\infty} x f(x) dx . \tag{3.38}$$

Dispersia sau varianţa unei variabile cu media m este:

$$\sigma^2(X) = M\left((X-m)^2\right) = \int_{-\infty}^{\infty} (x-m)^2 f(x) dx , \tag{3.39}$$

iar abaterea standard este:

$$\sigma(X)=\sqrt{\sigma^2(x)}. \tag{3.40}$$

O variabilă aleatoare distribuită uniform pe intervalul (a, b) are funcţia densitate de probabilitate de forma:

$$f(X;a,b)=\frac{1}{b-a};\ \ a<X<b \tag{3.41}$$

Parametrii pentru repartiţia uniformă sunt capetele intervalului: a şi b. Media şi varianţa unei variabile repartizată uniform sunt:

$$M(X)=\frac{a+b}{2}\ \ , \tag{3.42}$$

$$\sigma^2(X)=\frac{(b-a)^2}{12} \tag{3.43}$$

Funcţia de probabilitate a unei variabile distribuită uniform este:

$$F(X)=\begin{cases}0; & X\le a\\ \frac{x-a}{b-a}; & a<X<b\\ 1; & X\ge b.\end{cases} \tag{3.44}$$

În Matlab există definite funcţia densitate de probabilitate, f, (**`unifpdf`**) şi funcţia de repartiţie, F, (**`unifcdf`**) pentru repartiţia uniformă.

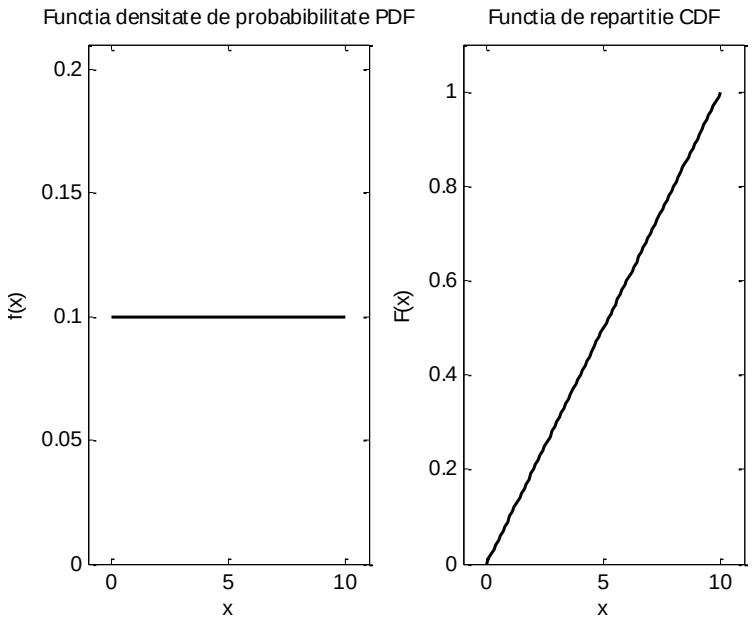


Fig. 3.12 Funcţia densitate de probabilitate şi funcţia de probabilitate pentru o variabilă repartizată uniform în intervalul [0, 10]

În fig. 3.12 se prezintă graficul funcţiei densitate de probabilitate şi funcţiei de probabilitate pentru o variabilă distribuită uniform în intervalul [0,10]. Reprezentarea grafică a fost obţinută cu următoarea secvenţă de program:

```
x=-1:.1:11;%se divizează intervalul [-1,11] cu pasul h=0.1
pdf=unifpdf(x,0,10);
cdf=unifcdf(x,0,10);
subplot(1,2,1),plot(x,pdf),title('Functia densitate de
probabibilitate PDF')
xlabel('x'),ylabel('f(x)'),axis([-1 11 0 .21])
subplot(1,2,2),plot(x,cdf),title('Functia de repartitie CDF')
xlabel('x'),ylabel('F(x)'),axis([-1 11 0 1.1])
```

3.3.4. DISTRIBUŢIA NORMALĂ

Distribuţia normală este o distribuţie fundamentală în statistică, fiind adecvată în modelarea a numeroase fenomene din natură. De asemenea, ea stă la baza inferenţei statistice. Funcţia de densitate de probabilitate are expresia:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\cdot\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \tag{3.45}$$

unde μ este media, iar σ abaterea standard. Graficul acestei funcţii este un clopot. Cu cât σ este mai mic, cu atât repartiţia este mai concentrată (clopotul este mai ascuţit). În figura 3.13 se prezintă câteva exemple de funcţii densitate de probabilitate pentru repartiţii normale. Se observă că pe măsură ce σ creşte înălţimea funcţiei scade, dar creşte anvergura acesteia.

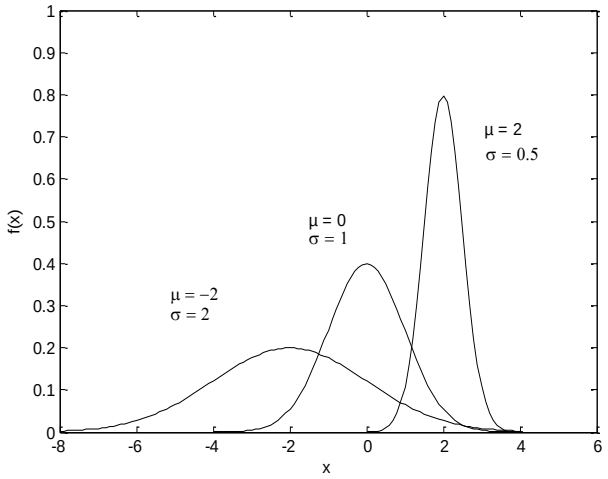


Fig. 3.13 Exemple de funcţii densitate de probabilitate pentru distribuţii normale

Localizarea şi forma repartiţiei depind de parametrii μ şi σ, în sensul că μ determină localizarea repartiţiei pe axă, iar σ forma curbei (fig. 3. 14).

Densitatea de probabilitate este continuă, are formă de clopot şi tinde asimptotic spre 0 pentru $x \to \pm\infty$. Notaţia $X \sim N(\mu, \sigma^2)$ se utilizează pentru a indica că o variabilă aleatoare X este distribuită normal cu media μ şi varianţa σ^2.

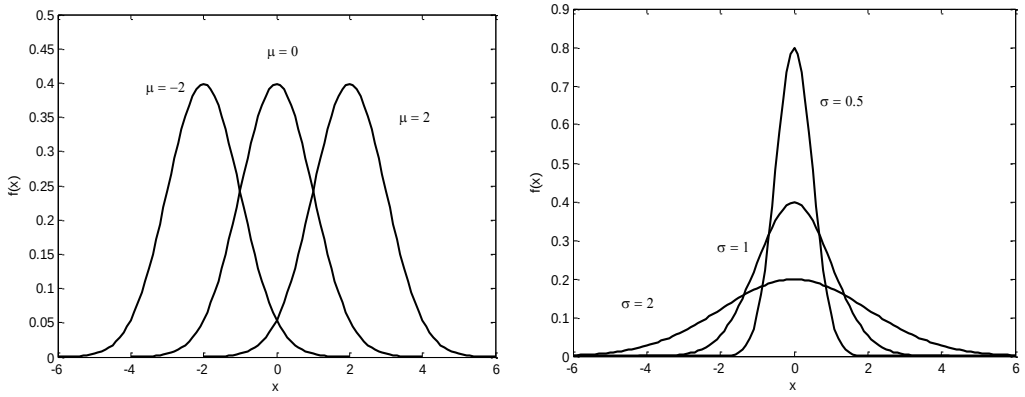


Fig. 3.14. Influenţa parametrilor asupra repartiţiei normale

Pentru orice distribuţie normală, indiferent de valoarea parametrilor μ şi σ, proporţia de observaţii ce aparţin unui interval centrat în μ este aceeaşi (fig. 3.15):

- 68.26% din valorile lui x $\in$ [μ - σ; μ + σ];
- 95.44% din valorile lui x $\in$ [μ - 2σ; μ + 2σ];
- 99.73% din valorile lui x $\in$ [μ - 3σ; μ + 3σ].

Orice variabilă aleatoare normală poate fi transformată în variabilă aleatoare normală normată prin schimbarea de variabilă: z = (x - μ)/σ, deci variabila aleatoare z are o distribuţie normală cu media egală cu 0 şi dispersia 1. În acest mod se poate calcula aria închisă sub curbă pentru diferite valori ale lui x, arie ce este egală cu probabilitatea ca variabila să aparţină intervalului respectiv (delimitat de limitele de integrare). Aceste valori sunt incluse în tabele statistice (Anexa 1). De remarcat că în tabel sunt înregistrate doar ariile pentru valori pozitive. În cazul când se calculează probabilitatea ca z $\in$ [-a, a], valoarea rezultată din tabel trebuie multiplicată cu 2.

Funcţia de repartiţie a unei variabile repartizate normal se defineşte cu relaţia:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} exp\left(-\frac{y^2}{2}\right) dy \qquad (3.46)$$

Funcţia de probabilitate se poate calcula utilizând funcţia eroare, notată cu erf.

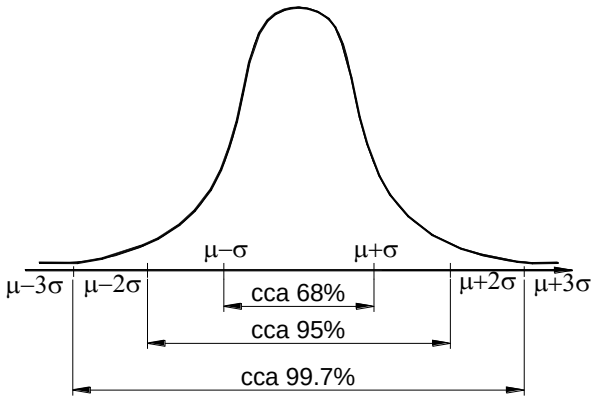


Fig. 3.14. Repartiţia normală

În Matlab sunt definite funcţiile:

- normpdf ce returnează valoriile funcţiei densitate de probabilitate cu sintaxa:
  `y = normpdf(x,miu,sigma)`

unde miu este valoarea mediei, iar sigma este deviaţia standard; implicit miu este 0, iar sigma 1;

- normcdf, ce returnează valoarea funcţiei de repartiţie. Sintaxa este **`p=normpdf(x,miu,sigma)`**
- normspec este o funcţie ce permite calcularea probabilităţii ca o variabilă repartizată normal cu media μ şi deviaţia standard σ să aibă valori între anumite limite. Funcţia reprezintă grafic densitatea de probabilitate restricţionată la intervalul definit de cele două limite. Limitele se indică sub forma unui vector. În cazul când nu există limită inferioară, primul element al vectorului se introduce -∞, iar în cazul inexistenţei limitei superioare, al doilea element al vectorului va fi ∞. Acest lucru se realizează în Matlab cu –Inf, respectiv Inf. Un exemplu se prezintă în continuare. Valorile implicite pentru μ şi σ sunt 0 şi 1.
- norminv este o funcţie ce returnează inversa funcţiei de repartiţie în punctul P. Sintaxa este **`norminv(P,miu,sigma)`**, unde valorile implicite sunt 0 (pentru miu) şi 1 (pentru sigma).

Pentru cazul unei variabile distribuită normal, având media 4 şi deviaţia standard egală cu 1.5, probabilitatea ca variabila să aparţină intervalului [2, 7] se poate calcula în felul următor:

```
% se seteaza limitele intervalului in vectorul limite
limite=[2,7];
miu=4;sigma=1.5;
prob=normspec(limite,miu,sigma)
```

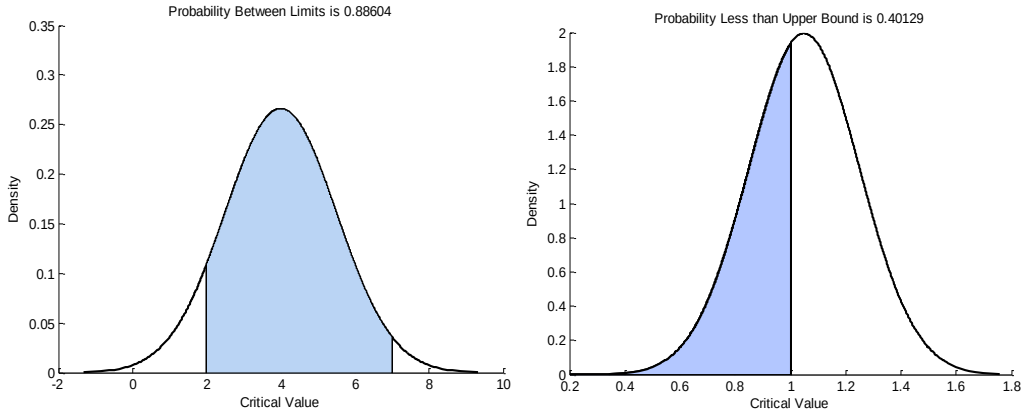


Fig. 3.15 Exemple pentru funcţia normspec

Într-o firmă producătoare de alimente, produsele sunt ambalate în cutii de 1 kg. În realitate masa cutiilor este normal distribuită, cu medie 1.05 kg şi abatere standard de 0.2. Să se determine probabilitatea să existe cutii sub greutatea

specificată.

În acest caz trebuie determinată probabilitatea ca variabila masă să ia valori în intervalul (–∞, 10).

```
limite=(-Inf, 1);
miu=1.05;sigma=.2;
prob=normspec(limite,miu,sigma);
```

3.3.5. DISTRIBUŢIA T

O altă distribuţie cu numeroase aplicaţii în statistică este distribuţia t sau Student. În cazul unei populaţii distribuită normal, dacă se extrag eşantioane şi se calculează media acestora, variabila t calculată cu relaţia:

$$t = \frac{\bar{x} - \mu}{S / \sqrt{n}}, \tag{3.48}$$

are repartiţia t.

Distribuţia t este simetrică, are media 0 şi o varianţă mai mare decât 1 [5]. De fapt, varianţa distribuţiei creşte pe măsură ce n, volumul eşantionului scade. Distribuţia admite ca parametru numărul gradelor de libertate, $\nu = n - 1$. Pe măsură ce numărul gradelor de libertate creşte. distribuţia se apropie de cea normală.

La fel ca în cazul repartiţiei normale standard, există tabele pentru distribuţia t ce permit extragerea valorilor t_α corespunzătoare unei anumite probabilităţi α. Fiecare rând al tabelului corespunde unui anumit număr de grade de libertate. Probabilităţile sunt calculate pentru o singură ramură a distribuţiei, $P(t \geq t_\alpha)$. t_α este α-cvantila repartiţiei Student, adică numărul cu proprietatea că $P(t \leq t_\alpha) = \alpha$.

Dacă se compară din tabele, distribuţia normală standard şi distribuţia t, se constată că pentru valori ale mai mari ale lui n cele două distribuţii sunt aproape identice. Pe baza acestui considerent, în cazurile practice, când $n \geq 30$ se înlocuieşte distribuţia t cu cea normală standard, iar distribuţia t se utilizează mai ales în cazurile când $n < 30$. Mai trebuie menţionat faptul că distribuţia t necesită ca populaţia din care s-a extras eşantionul să nu difere mult de repartiţia normală. În cazul eşantioanelor mici extrase din populaţii asimetrice, nici distribuţia z şi nici t nu pot fi utilizate.

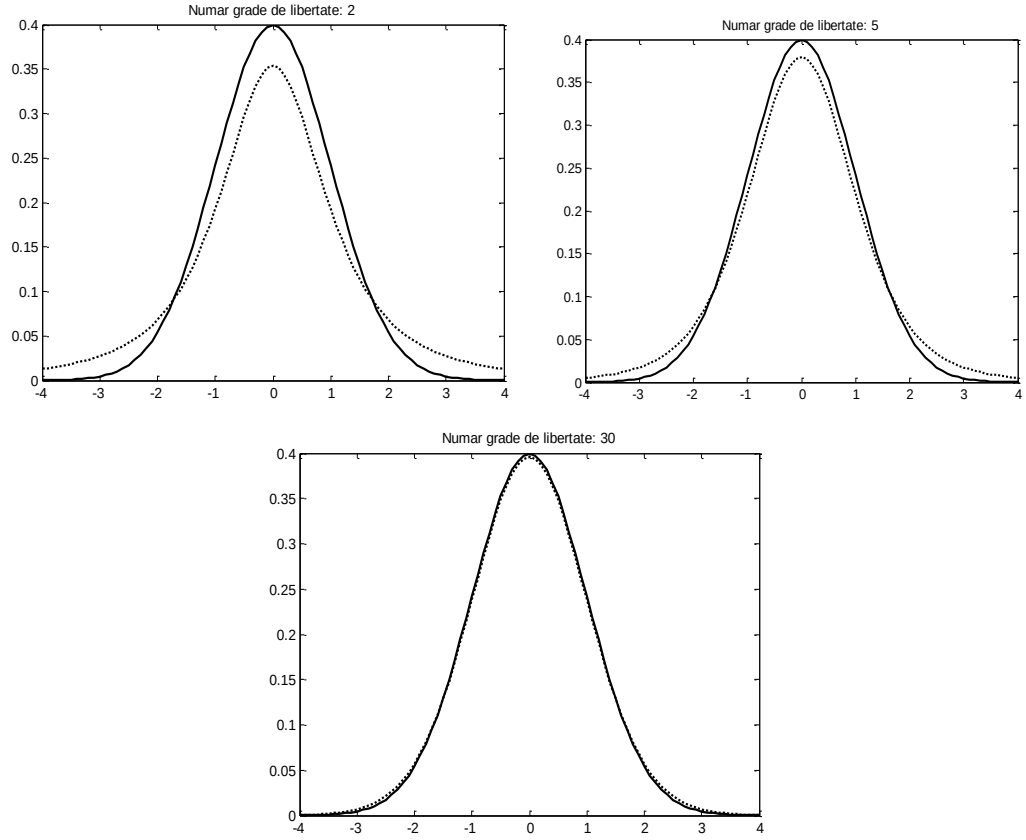


Fig. 3.16 Apropierea dintre repartiţia t şi repartiţia normală pe măsură ce numărul gradelor de libertate creşte

În concluzie se poate afirma că distribuţia t se utilizează în următoarele condiţii:

- deviaţia standard a eşantionului, S, este utilizată pentru estimarea lui σ;
- volumul eşantionului este mic, n < 30;
- populaţia este distribuită aproximativ normal.

În Matlab sunt definite funcţiile tpdf, respectiv tcdf pentru determinarea densităţii de probabilitate şi a probabilităţii unei variabile repartizată conform repartiţiei t. Sintaxa este: **`y=tpdf(x,v)`**, respectiv **`P=tcdf(x,v)`**, unde v este numărul gradelor de libertate. Este definită şi funcţia tinv(P,v), inversa funcţiei de repartiţie, ce returnează cvantila repartiţiei t cu v grade de libertate în punctul P.

3.3.6. DISTRIBUŢIA EXPONENŢIALĂ

O variabilă aleatoare ce are densitatea de probabilitate:

$$f(x;\lambda) = \lambda e^{-\lambda x};\ x \geq 0;\ \lambda > 0. \tag{3.49}$$

Se numeşte variabilă aleatoare exponenţial distribuită de parametru λ. O variabilă aleatoare exponenţial distribuită modelează durata de viaţă a unui produs sau dispozitiv sau lingimea intervalelor de timp între producerea a două evenimente consecutive contorizate de o variabilă aleatoare Poisson. Ex.: intervalelor de timp între sosirea a doi clienţi la o bancă, timpul dintre apelurile telefonice, timpul până la căderea unei piese etc.

Media şi varianţa unei populaţii repartizată exponenţial sunt:

$$M(x) = \frac{1}{\lambda} \tag{3.50}$$

$$\sigma^2(x) = \frac{1}{\lambda^2}. \tag{3.51}$$

Funcţia de probabilitate este de forma:

$$F(x) = \begin{cases} 0; & x < 0 \\ 1 - e^{-\lambda x}; & x \geq 0 \end{cases}. \tag{3.52}$$

Repartiţia exponenţială prezintă o proprietate interesantă şi anume intervalul de timp scurs până la apariţia unui eveniment nu depinde de timpul anterior [10]. Acest lucru se poate scrie:

$$P(X > s + t | X > s) = P(X > t) \tag{3.53}$$

Cu alte cuvinte, acest lucru înseamnă că probabilitatea ca o piesă sau dispozitiv să fie în stare de funcţionare după de s + t unităţi de timp, dacă ea a funcţionat bine deja de timpul s, este egală cu probabilitatea ca piesa să funcţioneze timpul t.

Când această distribuţie se utilizează la reprezentarea intervalelor de timp între evenimente, parametrul λ reprezintă numărul mediu de evenimente produs în unitatea de timp, adică frecvenţa acestora.

Exemplu: Intervalul de timp între sosirea maşinilor într-o intersecţie este în medie de 12 secunde. Se cere probabilitatea ca intervalul de timp să fie cel mult egal cu 10 secunde. În acest caz numărul mediu al maşinilor este egal cu 1/12, iar probabilitatea cerută este egală cu:

$$P(x \leq 10) = 1 - e^{-(1/12)10} = 0.5654$$

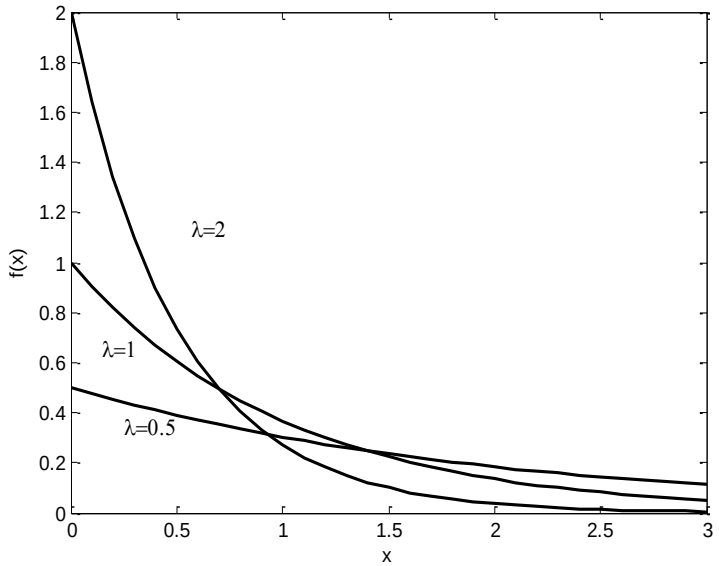


Fig. 3.17 Repartiţia exponenţială pentru diferite valori ale parametrului λ

În Matlab funcţia de repartiţie pentru distribuţia exponenţială este **`expcdf(x,1/lambda)`** cu care se poate calcula probabilitatea ca X să ia valori mai mici sau egale cu x. Trebuie menţionat faptul că în Matlab funcţia de probabilitate pentru repartiţia exponenţială se defineşte cu relaţia:

$$f(x;\mu) = \frac{1}{\mu} e^{-\frac{x}{\mu}};\ x \leq 0;\ \mu > 0. \tag{3.54}$$

În acest context rezultatul aplicaţiei anterioare se poate obţine prin simpla apelare a funcţiei **`expcdf(10,12)`**, ce va returna 0.5654. De asemenea este definită funcţia **`exppdf(x, 1/lambda)`**,ce implementează calculul densităţii de probabilitate, respectiv **`expinv(P,1/lambda)`** inversa funcţiei de repartiţie.

3.3.7. DISTRIBUŢIA GAMMA

Funcţia densitate de probabilitate a repartiţiei gamma are expresia:

$$f(x;\lambda,t) = \frac{\lambda e^{-\lambda x}(\lambda x)^{t-1}}{\Gamma(t)};\ x \geq 0, \tag{3.55}$$

unde t este parametru de formă, iar λ este parametru de localizare. Funcţia gamma este definită de:

$$\Gamma(t) = \int_0^{\infty} e^{-y} y^{t-1} dy, \tag{3.56}$$

relaţie care în cazul valorilor întregi ale parametrului t devine:

$$\Gamma(t) = (t-1)! \tag{3.57}$$

De remarcat că pentru t = 1 distribuţia gamma devine exponenţială. Pentru valori pozitive ale lui t, repartiţia gamma se poate utilize la modelarea intervalului de timp scurs până la producerea a t evenimente, dacă numărul evenimentelor rare are distribuţie Poisson.

Media şi varianţa distribuţiei gamma sunt date de relaţiile:

$$M(x) = \frac{t}{\lambda} \tag{3.58}$$

$$\sigma^2(x) = \frac{t}{\lambda^2}. \tag{3.59}$$

Funcţia de repartiţie a acestei distribuţii are expresia [10]:

$$F(x;\lambda,t) = \begin{cases} 0\,; & x \le 0 \\ \frac{1}{\Gamma(t)} \int_0^{\lambda x} y^{t-1} e^{-y} dy; & x > 0 \end{cases}. \tag{3.60}$$

Relaţia (3.60) se poate evalua în Matlab cu funcţia **`gammacdf(lambda*x,t)`**.

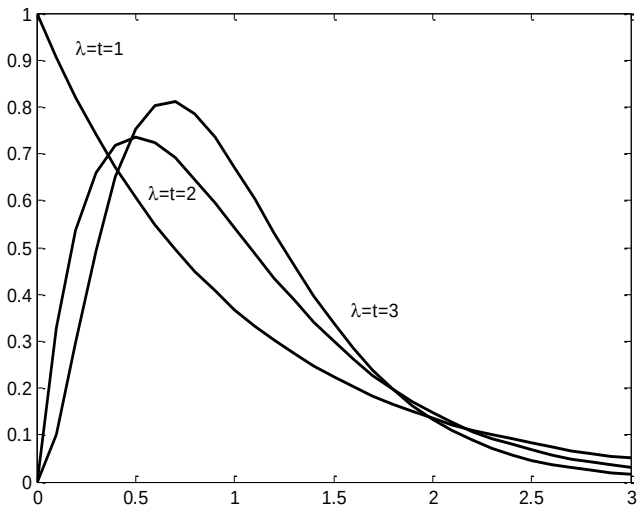


Fig. 3.18 Exemple de funcţii densitate de probabilitate gamma

Alura densităţii de probabilitate gamma pentru diverse valori ale parametrilor se poate observa în fig. 3.18.

Graficul prezentat în figură se obţine cu secvenţa de program:

```
x=0:.1:3;
y1=gampdf(x,1,1);
y2=gampdf(x,2,1/2);
y3=gampdf(x,3,1/3);
plot(x,y1,x,y2,x,y3)
```

La fel ca în cazurile anterioare, în Matlab există definite funcţiile gampdf, respectiv gamcdf, ce pot fi utilizate pentru calcularea densităţii de probabilitate şi funcţiei de repartiţie a distribuţiei gamma.

3.3.8. DISTRIBUŢIA χ^2

Distribuţia χ^2 este un caz particular al distribuţiei gamma, având λ = 0.5 şi t = ν/2, cu ν o valoare întreagă pozitivă, ce reprezintă numărul gradelor de libertate. Repartiţia χ^2 are aplicaţii în testele de concordanţă, cu ajutorul cărora se verifică ipoteza modelării eşantionului cu o anumită repartiţie.

Funcţia densitate de probabilitate pentru o variabilă aleatoare distribuită conform repartiţiei χ^2, cu ν grade de libertate este;

$$f(x,\nu) = \frac{1}{\Gamma(\nu/2)}\left(\frac{1}{2}\right)^{\nu/2} x^{\nu/2-1}e^{-\frac{x}{2}}; \qquad x \geq 0. \tag{3.61}$$

Media şi varianţa distribuţiei se pot obţine pe baza repartiţiei gamma, având valorile:

$$M(x) = \nu \tag{3.62}$$

$$\sigma^2(x) = 2\nu. \tag{3.63}$$

3.4 INFERENŢA STATISTICĂ

Investigările statistice constau în studiul unor caracteristici ale unei populaţii. În acest scop, din populaţie (ce poate fi finită sau infinită) se extrag eşantioane de volum finit. În urma investigării eşantionului se obţin informaţii ce se extrapolează la întreaga populaţie.

Inferenţa statistică dezvoltă metode de investigare a unei populaţii prin sondaj, adică metode şi tehnici de analiză a datelor obţinute prin eşantioane. Având în

vedere că populaţia este caracterizată de câţiva indicatori de interes, statistica formulează inferenţe, adică predicţii, asupra parametrilor populaţiei cu diferite nivele de încredere.

În practică inferenţa statistică se ocupă cu două mari categorii de probleme:

- estimarea valorilor parametrilor unei populaţii (media, dispersia);
- testarea ipotezelor referitoare la parametrii populaţiei.

Spre deosebire de statistica descriptivă, în inferenţa statistică se asociază un model probabilist caracteristicii investigate, în sensul că valorile numerice asociate caracteristicii investigate au o anumită distribuţie de probabilitate, ce este definită de o densitate de probabilitate, f. Această densitate de probabilitate depinde de un parametru necunoscut, $\theta \in R^n$, în cazurile cele mai frecvente n fiind 1 sau 2.

Pe baza tehnicilor dezvoltate de inferenţa statistică se estimează parametrul θ prin:

- o valoare punctuală;
- un interval de valori ce va conţine valoarea estimată cu o probabilitate prescrisă;
- o modalitate de testare a ipotezelor $\theta \le \theta_0, \theta \ge \theta_0$; unde θ_0 este o valoare dată.

3.4.1. EŞANTIONAREA

Dintr-o populaţie de volum N se pot extrage eşantioane de volum n, n<N. Numărul de astfel de eşantioane este mare şi valoarea unei anumite statistici, de ex. media, calculată pentru fiecare eşantion va diferi de la un eşantion la altul. Frecvenţa de distribuţie a tuturor acestor eşantioane dă informaţii primare despre distribuţia unui eşantion.

Conceptul de eşantionare al unei populaţii poate fi ilustrat prin următorul exemplu. Fie o populaţie în care variabila poate lua orice valoare întreagă între 0 şi 4 (incidenţa apariţiei a 4 defecte la un produs). Fie populaţia formată din 5 obiecte, fiecare având 0, 1, 2, 3 sau 4 defecte. Se extrag eşantioane formate din două exemplare. Eşantionarea se poate face cu sau fără returnare. La eşantionarea cu returnare, după extragerea unui element, acesta se introduce din nou în cadrul populaţiei, putând fi ulterior extras.

În tabelul 2.2 se prezintă toate posibilităţile de eşantioane ce apar. Pentru

cazul eşantionării cu înlocuire se obţine: $\mu_{\bar{x}} = 2.0,\ \sigma^2_{\bar{x}} = 1.0$, iar în cazul eşantionării fără înlocuire: $\mu_{\bar{x}} = 2.0,\ \sigma^2_{\bar{x}} = 0.75$. Deoarece s-au luat în considerare toate eşantioanele posibile se poate utiliza notaţia cu indicatorii populaţiei (μ şi σ) şi nu ai eşantionului ($\bar{x}$ şi S).

Se remarcă faptul că cele două medii sunt egale ($\mu_{\bar{x}} = 2.0$), dar varianţa diferă ($\sigma^2_{\bar{x}} = 1.0$, respectiv $\sigma^2_{\bar{x}} = 0.75$).

Tabelul 2.2. Eşantionarea şi prelucrarea datelor în cazul unei populaţii formate din 5 elemente

<table>
<tr><td colspan="5">Eşantionarea cu returnare</td><td colspan="5">Eşantionarea fără returnare</td></tr>
<tr><td>0, 0</td><td>1, 0</td><td>2, 0</td><td>3, 0</td><td>4, 0</td><td rowspan="5">0, 1
0, 2
0, 3
0, 4</td><td rowspan="5">1, 2
1, 3
1, 4</td><td rowspan="5">2, 3
2, 4</td><td rowspan="5">3, 4</td><td rowspan="5"></td></tr>
<tr><td>0, 1</td><td>1, 1</td><td>2, 1</td><td>3, 1</td><td>4, 1</td></tr>
<tr><td>0, 2</td><td>1, 2</td><td>2, 2</td><td>3, 2</td><td>4, 2</td></tr>
<tr><td>0, 3</td><td>1, 3</td><td>2, 3</td><td>3, 3</td><td>4, 3</td></tr>
<tr><td>0, 4</td><td>1, 4</td><td>2, 4</td><td>3, 4</td><td>4, 4</td></tr>
<tr><td colspan="10">Mediile fiecărui eşantion</td></tr>
<tr><td>0.0</td><td>0.5</td><td>1.0</td><td>1.5</td><td>2.0</td><td rowspan="5">0.5
1.0
1.5
2.0</td><td rowspan="5">1.5
2.0
2.5</td><td rowspan="5">2.5
3.0</td><td rowspan="5">3.5</td><td rowspan="5"></td></tr>
<tr><td>0.5</td><td>1.0</td><td>1.5</td><td>2.0</td><td>2.5</td></tr>
<tr><td>1.0</td><td>1.5</td><td>2.0</td><td>2.5</td><td>3.0</td></tr>
<tr><td>1.5</td><td>2.0</td><td>2.5</td><td>3.0</td><td>3.5</td></tr>
<tr><td>2.0</td><td>2.5</td><td>3.0</td><td>3.5</td><td>4.0</td></tr>
<tr><td colspan="10">Distribuţia mediilor</td></tr>
<tr><td>$\bar{x}$</td><td colspan="3">Nr. eşantioane f</td><td>Probabiltatea de apariţie f/∑f</td><td>$\bar{x}$</td><td colspan="2">Nr. eşantioane f</td><td colspan="2">Probabiltatea de apariţie f/∑f</td></tr>
<tr><td>0.0</td><td colspan="3">1</td><td>1/25</td><td>0.5</td><td colspan="2">1</td><td colspan="2">1/10</td></tr>
<tr><td>0.5</td><td colspan="3">2</td><td>2/25</td><td>1.0</td><td colspan="2">1</td><td colspan="2">1/10</td></tr>
<tr><td>1.0</td><td colspan="3">3</td><td>3/25</td><td>1.5</td><td colspan="2">2</td><td colspan="2">2/10</td></tr>
<tr><td>1.5</td><td colspan="3">4</td><td>4/25</td><td>2.0</td><td colspan="2">2</td><td colspan="2">2/10</td></tr>
<tr><td>2.0</td><td colspan="3">5</td><td>5/25</td><td>2.5</td><td colspan="2">2</td><td colspan="2">2/10</td></tr>
<tr><td>2.5</td><td colspan="3">4</td><td>4/25</td><td>3.0</td><td colspan="2">1</td><td colspan="2">1/10</td></tr>
<tr><td>3.0</td><td colspan="3">3</td><td>3/25</td><td>3.5</td><td colspan="2">1</td><td colspan="2">1/10</td></tr>
<tr><td>3.5</td><td colspan="3">2</td><td>2/25</td><td></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>4.0</td><td colspan="3">1</td><td>1/25</td><td></td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>Total</td><td colspan="3">25</td><td></td><td>10</td><td colspan="2"></td><td colspan="2"></td></tr>
</table>

Următorul pas este găsirea relaţiilor dintre indicatorii eşantionului şi cei ai populaţiei. În exemplul prezentat, media populaţiei este:

$$\mu = (0+1+2+3+4)/5 = 2.0,$$

valoare ce este egală cu media determinată în cele două cazuri de eşantionare.

În tabelul 2.3 se prezintă rezultatele obţinute în cazul extragerii de eşantioane de 3, respectiv 4 elemente.

Pentru eşantioanele de volum n = 3 se obţine $\mu_{\bar{x}} = 2.0, \quad \sigma^2_{\bar{x}} = 0.66$, iar pentru cele cu n = 4, $\mu_{\bar{x}} = 2.0, \quad \sigma^2_{\bar{x}} = 0.5$. Se remarcă faptul că media în toate situaţiile este egală cu media populaţiei. De asemenea, se observă că mediile eşantioanelor se distribuie simetric în jurul valorii medii ideale, μ, şi pe măsură ce volumul eşantionului creşte frecvenţa de apariţie a valoriilor apropiate de media ideală creşte. În cazul dispersiei şi abaterii standard apar diferenţe între parametrii populaţiei şi cei ai eşantioanelor. Pe măsură ce volumul eşantionului creşte, dispersia şi abaterea standard a mediei scade. Datorită variaţiei mediei eşantioanelor în jurul mediei populaţiei datorită eşantionării, o eroare de eşantionare apare ori de câte ori o media unui singur eşantion, $\bar{x}$ este utilizată pentru estimarea mediei populaţiei, μ. Din acest motiv termenul de eroare standard se foloseşte uneori în locul abaterii standard, când se referă la distribuţia mediei ($\sigma_{\bar{x}}$).

Tabelul 2.3. Distribuţia mediilor la eşantionarea cu înlocuire

Volumul eşantionului n = 3			Volumul eşantionului n = 4		
$\bar{x}$	Nr. eşantioane f	Probabilitate de apariţie	$\bar{x}$	Nr. eşantioane f	Probabilitate de apariţie
0.00	1	0.008	0.00	1	0.0016
0.33	3	0.024	0.25	4	0.0064
0.67	6	0.048	0.50	10	0.0160
1.00	10	0.080	0.75	20	0.0320
1.33	15	0.120	1.00	35	0.0560
1.67	18	0.144	1.25	52	0.0832
2.00	19	0.152	1.50	68	0.1088
2.33	18	0.144	1.75	80	0.1280
2.67	15	0.120	2.00	85	0.1360
3.00	10	0.080	2.25	80	0.1280
3.33	6	0.048	2.50	68	0.1088
3.67	3	0.024	2.75	52	0.0832
4.00	1	0.008	3.00	35	0.0560
			3.25	20	0.0320
			3.50	10	0.0160
			3.75	4	0.064
			4.00	1	0.0016
Total	125	1.00	Total	625	1.0000

În cazul dispersiei, se observă că pe măsură ce volumul eşantionului creşte varianţa scade. Varianţa populaţiei este:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N} = 2.00$$

La eşantionarea cu returnare, relaţia dintre $\sigma_{\bar{x}}^2$ şi σ^2 este: $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$.

La eşantionarea cu returnare, practic populaţia are un volum infinit, deoarece după fiecare extragere, elementele se introduc din nou, volumul populaţiei rămânând acelaşi. La eşantionarea fără returnare, volumul populaţiei scade pe măsură ce se extrag elemente şi, ca uramare, dispersia distribuţiei scade. Din acest motiv această variantă de extragere a eşantioanelor este mult mai eficientă în colectarea informaţiei. Factorul de scădere a dispersiei depinde de mărimea populaţiei şi de volumul eşantionului, factorul de corecţie datorat numărului finit al populaţiei fiind (N – n)/(N – 1). În acest caz, relaţia dintre $\sigma_{\bar{x}}^2$ şi σ^2 este:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right). \tag{3.64}$$

În concluzie, se poate afirma că în cazul eşantionării cu returnare $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$, iar în cazul eşantionării fără returnare $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$.

În cazul când volumul populaţiei este foarte mare şi volumul eşantionului este mic, valoarea factorului de corecţie poate fi ignorată. De obicei, în situaţia când $n \leq 0.05N$, factorul de corecţie poate fi ignorat şi se poate considera [5]:

$$\begin{aligned} \mu_{\bar{x}} &= \mu \\ \sigma_{\bar{x}} &= \frac{\sigma}{\sqrt{n}} \end{aligned}. \tag{3.65}$$

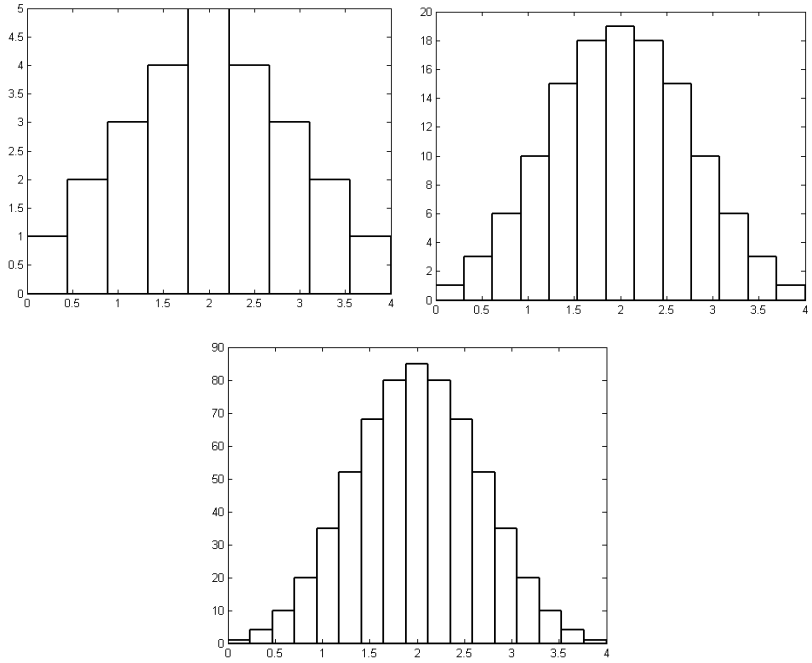


Fig. 3. 19. Histogramele pentru eşantionarea cu 2, 3 şi 4 elemente

În fig. 3.19 se prezintă histogramele pentru cele trei serii de date asociate mediilor $\bar{x}$ ale eşantioanelor cu n = 2, n = 3, respectiv n = 4 elemente. Se observă că pe măsură ce n creşte, histograma aproximează din ce în ce mai bine clopotul se defineşte repartiţia normală. Această tendinţă este adevărată pentru toate eşantioanele, indiferent de distribuţia populaţiei. S-a demonstrat că pe măsură ce volumul eşantionului creşte, distribuţia tinde spre una normală, chiar în situaţia când populaţia are o altfel de distribuţie. Această proprietate este enunţată în teorema limită centrală.

În consecinţă, proprietăţile repartiţiei normale pot fi aplicate pentru determinarea erorii ce apare datorită eşantionării. În practică, teorema limită centrală se poate aplica la eşantioane de volum n ≥ 30. În cazul când populaţia are o distribuţie normală, media eşantionului, indiferent de volumul său, va avea o repartiţie normală.

3.4.2. ESTIMĂRI

În cadrul acestui paragraf se tratează problema estimării indicatorilor statistici ai populaţiei pe baza eşantionării şi determinarea gradului de încredere a

acestei estimări. Estimarea mediei se poate face printr-o estimare punctuală sau cu un interval de încredere.

În cazul estimării punctuale se consideră abaterea standard σ a populaţiei cunoscută. În estimarea punctuală, se utilizează datele dintr-un eşantion pentru a calcula valoarea indicatorului statistic medie, cu alte cuvinte media experimentală, $\bar{x}$, este un estimator punctual pentru μ.

Pentru ca estimarea să fie bună, este necesar să îndeplinească trei condiţii [5]:

- să fie nedeplasată, o estimare este nedeplasată dacă media eşantionului (în acest caz $\mu_{\bar{x}}$) este egală cu parametrul estimat (μ);
- este eficientă, adică deviaţia standard să fie cât mai mică posibil;
- este consistentă, condiţie ce este îndeplinită dacă valoarea estimată tinde spre valoarea adevărată pe măsură ce volumul eşantionului creşte.

Media unui eşantion îndeplineşte aceste condiţii, deci poate fi considerată o estimare punctuală. Trebuie reţinut faptul că o singură estimare a mediei diferă de media populaţiei, datorită erorii de eşantionare. Din acest motiv, în multe situaţii se preferă estimarea printr-un interval de încredere.

Estimarea intervalului de încredere pentru media μ a distribuţiei normale se bazează pe un eşantion de volum n şi medie $\bar{x}$. Intervalul de încrdere, de nivel de încredere de 95% pentru media μ a populaţiei este:

$$\mu = \bar{x} \pm 1.96\sigma_{\bar{x}} = \bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}. \qquad (3.66)$$

Se pot utiliza diferite nivele de încredere în estimarea intervalului valorii medii. În cazul general, nivelul de încredere se notează 100(1 - α)%, unde α reprezintă suma celor două arii din extremele repartiţiei normale (vezi fig. 3.20). Pentru o probabilitate de 95%, α = 0.05. În cazul general, media cu incertitudinea asociată este:

$$\mu = \bar{x} \pm z_{1-\frac{\alpha}{2}}\sigma_{\bar{x}} = \bar{x} \pm z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \qquad (3.67)$$

unde 100(1 - α) este nivelul de încredere, iar $Z_{1-\alpha/2}$ este valoarea variabilei Z ce exclude la fiecare extremitate a distribuţiei o arie de α/2, adică α/2-cvantila distribuţiei normale standard.

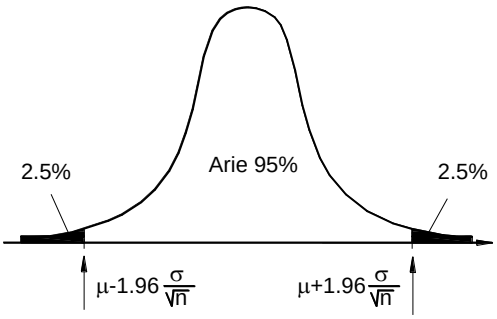


Fig. 3.20. Stabilirea intervalului de încredere

Pentru o probabilitate de 90%, $Z_{\alpha/2}$ = 1.65 şi pentru P = 99%, Z = 2.58.

În mod uzual, σ este o mărime necunoscută, singurele informaţii certe sunt cele legate de eşantion. Se poate demonstra, [5, 10], că S este o estimare nedeplasată a abaterii standard. În cazul eşantioanelor de volum mare, estimarea este foarte bună, însă în cazul eşantioanelor de volum mic, S poate fi subestimat. Din acest motiv la eşantioanele de volum mic, trebuie lărgit intervalul de încredere. Acest lucru se realizează prin utilizarea repartiţiei t.

Similar cu construirea intervalului de încredere pe baza repartiţiei normale standard $\mu = \bar{x} \pm z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$, în cazul utilizării distribuţiei t intervalul este:

$$\mu = \bar{x} \pm t(1-\alpha/2, n-1) \cdot S_{\bar{x}} = \bar{x} \pm t(1-\alpha/2, n-1) \cdot \frac{S}{\sqrt{n}}, \qquad (3.68)$$

unde 100(1-α) este nivelul de încredere.

Cvantilele repartiţiei normale şi t se evaluează cu funcţiile Matlab **`norminv(P,miu,sigma)`**, respectiv **`tinv(P,v)`.**

Ex. 1: Masa unor persoane ce utilizează un ascensor este distribuită normal. Un eşantion furnizează următoarele valori: 71, 85, 68, 72, 58, 76, 74, 80. Să se estimeze cu o probabilitate de 95% valoarea medie.

$$\mu = \bar{x} \pm t(1-\alpha/2, n-1) \cdot \frac{S}{\sqrt{n}}$$

$$t(1-\alpha/2, n-1) = t(0.05/2, 8-1) = 2.365$$

$$\bar{x} = 73,\ \text{S=8.09},\ \Rightarrow S_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{2.365}{\sqrt{8}} = 2.86$$

$$\mu = 73 \pm (2.365 \cdot 2.86) = 73 \pm 6.76$$

$$\mu \in [66.24; 79.76].$$

Valoarea cvatilei repartiţiei t se obţine în Matlab cu **`tinv([0.025 0.975], 7)`**.

Ex. 2: Din 40 de copii ce merg la acceaşi grădiniţă, au fost selectaţi 8 şi s-a măsurat înălţimea lor în cm.: 110, 112, 85, 117, 100, 98, 104, 90. Ştiind că înălţimea este o variabilă aleatoare normal distribuită să se estimeze media înălţimii copiilor.

Se observă că volumul eşantionului este mai mare decât 5% din populaţie (8/40 = 0.20 > 0.05), deci va trebui aplicat un factor de corecţie datorită populaţiei finite.

$$\mu = \bar{x} \pm t(\alpha/2, n-1) \cdot \frac{S}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

$$t(\alpha/2, n-1) = t(0.05/2, 8-1) = 2.365$$

$$\bar{x} = 102,\ \text{S} = 10.994 \Rightarrow S_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{10.994}{\sqrt{8}} = 3.89$$

$$\mu = 102 \pm (2.365 \cdot 3.89)\sqrt{\frac{40-8}{40-1}} = 102 \pm 8.33$$

$\mu \in [93.67;\ 110.33]$.

3.4.3 TESTAREA IPOTEZELOR

Informaţiile legate de o anumită populaţie statistică se obţin prin selectarea unui eşantion, cu ajutorul căruia se estimează parametrii populaţiei respective. Pe baza acestor parametri se poate forma o imagine asupra caracteristicilor analizate. De exemplu, dintr-un lot de piese se extrage un eşantion de n elemente şi se efectuează măsurări ale lungimii l, lungime presupusă a fi o

variabilă repartizată normal cu media μ şi dispersia σ. Pentru eşantionul extras s-a determinat prin calcul media experimentală $\bar{I}$. Se pune problema dacă $\mu = \bar{I}$, problemă ce constituie o ipoteză statistică, adică o presupunere (supoziţie) asupra populaţiei statistice luate în studiu. Chiar dacă ipoteza statistică este avansată pe baza unui eşantion, concluzia referitoare la valoarea parametrului sau la natura repartiţiei se referă la întreaga populaţie.

În mod firesc, o presupunere asupra unor elemente ale repartiţiei ($\mu = \bar{I}$) poate admite o alternativă ($\mu < \bar{I}$). Ipoteza iniţială se numeşte ipoteză nulă şi se notează H_o. Ipoteza alternativă sau concurentă se notează H_1.

Ipotezele statistice nu sunt echivalente cu ipotezele ştiinţifice. În cazul unei ipoteze ştiinţifice este suficient un singur exemplu contrar pentru a o infirma, dar o ipoteză statistică poate fi adevărată, chiar dacă într-o anumită situaţie a fost respinsă ca fiind falsă.

Ipotezele statistice sunt însoţite de două tipuri de erori:

- o eroare apare atunci când se respinge ipoteza iniţială, H_0, în situaţia în care ea este adevărată. Această eroare este probabilitatea de a respinge ipoteza H_0, când în realitate ea este adevărată. Eroarea se notează, de obicei, cu α:
  α = P(respinge $H_0 \mid H_0$ adevărată). (3.69)
- o eroare constă în acceptarea ipotezei H_0, când în realitate ea este falsă. Această eroare se notează cu β:
  β = P(acceptă $H_0 \mid H_0$ falsă). (3.70)

Aceste erori se mai numesc erori de tip I şi II. Este evident că există interesul menţinerii valorilor α şi β la niveluri cât mai mici.

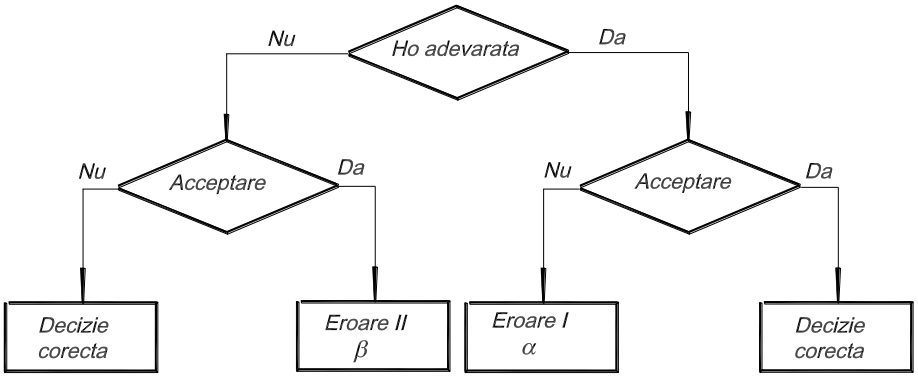


Fig. 3.21 Erori ce apar la verificarea ipotezelor statistice

Metoda efectivă de luare a deciziei asupra acceptării sau respingerii ipotezei H_o se face cu ajutorul unui test sau criteriu statistic, care este un procedeu furnizat de statistica matematică.

Testul impune luarea uneia dintre cele două decizii – acceptare sau respingere – a ipotezei nule cu un anumit risc. Acest lucru este prezentat în Figura 3.21. Se poate observa că valoarea 1 - β este probabilitatea de respingere a ipotezei nule când aceasta este falsă. Valoarea 1 - β poate fi considerată puterea testului respectiv.

Din punct de vedere practic, pentru a verifica ipoteza H_o, cu alternativa H_1, sunt necesare, [8]:

- o statistică (o funcţie de măsurările experimentale) a cărei repartiţie să fie cunoscută sau cel puţin să fie exprimabilă analitic;
- o valoare considerată “critică” cu care să se compare valoarea calculată a statisticii;
- o regulă de decizie prin care să se accepte sau să se respingă H_o (când se respinge H_o, automat se acceptă H_1);
- o valoare a riscului ales α, ce se mai numeşte nivel de semnificaţie al testului.

a. Media unei singure populaţii

În continuarea paragrafului, se tratează problema testării mediei unei populaţii, în sensul că se presupune cunoscută media populaţiei şi se stabileşte dacă un eşantion este extras dintr-o populaţie cu aceeaşi medie.

Ideea de bază în cazul testării ipotezelor este că se începe cu o presupunere asupra unui parametru, cum ar fi media, după care se utilizează eşantionul pentru testarea ipotezelor. Presupunerea admisă, ipoteza nulă se notează H_0. De ex.: Se emite ipoteza că μ este μ_0, adică H_0: $\mu = \mu_0$. În afara ipotezei nule se definesc ipotezele alternative, H_1. Ipoteza alternativă poate fi una din cele trei situaţii posibile, H_1: $\mu \neq \mu_0$; H_1: $\mu > \mu_0$; H_1: $\mu < \mu_0$. Procedura de testare constă în compararea statisticii experimentale, în acest caz $\bar{x}$ cu parametrul corespunzător ipotezei nule, μ_0. Dacă există diferenţe între cele două valori, ipoteza nulă se respinge. De obicei apar diferenţe, fie datorită întâmplării, fie datorită faptului că H_0 este falsă. În acceptarea şi respingerea unei ipoteze statistice intervin cele două tipuri de erori, α şi β, deci niciodată nu se poate lua o decizie cu siguranţă absolută.

În cazul testării ipotezei $\mu = \mu_0$ cu o probabilitate de 95%, există un risc de 5%. În fig. 3.22 se poate observa că apar zone de acceptare şi zone de respingere. În cazul ipotezei alternative H_1: $\mu \neq \mu_0$, zona de respingere, având o arie egală cu α, se împarte în două regiuni distincte, fiecare având o arie egală cu $\alpha/2$ (fig. 2.16 a). Acest tip de test este cu două extremităţi. În cazul celorlalte ipoteze alternative $\mu > \mu_0$ (fig. 2.16 b) şi $\mu < \mu_0$ (fig. 2.16 c) apare o singură zonă de respingere, testele fiind cu o singură extremă.

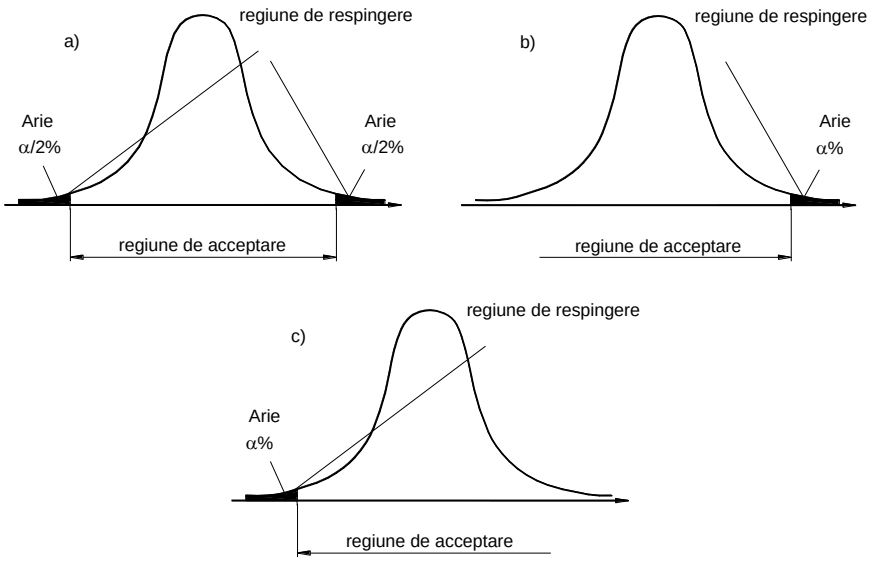


Fig. 3.22 Zone de respingere şi acceptare în testarea ipotezelor

În testarea unei ipoteze există totdeauna 5 paşi ce trebuie urmăriţi:

Pasul 1: Formularea ipotezei nule şi a celei alternative. Ipoteza nulă specifică valoarea parametrului ce urmează a fi testat ($\mu = \mu_0$). Forma ipotezei alternative poate diferi. În cazul testelor cu două extremităţi este importantă doar existenţa unei diferenţe între μ şi μ_0, în timp ce în cazul ipotezelor cu o singură extremitate este important şi sensul acestei diferenţe. Ipotezele se pot formula:

$$\begin{matrix} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{matrix} \quad \text{sau} \quad \begin{matrix} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{matrix} \quad \text{sau} \quad \begin{matrix} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{matrix} .$$

Pasul 2: Specificarea nivelului de semnificaţie α. Cel mai uzual este să se lucreze cu nivelul de semnificaţie de 5% sau 1%. Având stabilit nivelul de semnificaţie, punctul de demarcare a regiunii de acceptare cu cea de respingere reprezintă valoarea critică a testului. Pentru testele cu două extermităţi există două valori critice. Valorile se numesc critice deoarece sunt valorile cu care se compară parametrii experimentali.

Tabelul 2.4. Valori critice în cazul testării mediei

Tip test	Ipoteze		$\bar{x}$	Z	t
	H_0	H_1			
2 extremităţi	$\mu=\mu_0$	$\mu \neq \mu_0$	$\mu_0 \pm Z_{\alpha/2}\sigma_{\bar{x}}$	$\pm Z_{\alpha/2}$	$\pm t(\alpha/2,\nu)$
o extremitate	$\mu=\mu_0$	$\mu > \mu_0$	$\mu_0 + Z_{\alpha}\sigma_{\bar{x}}$	$+ Z_{\alpha}$	$+ t(\alpha,\nu)$
o extremitate	$\mu=\mu_0$	$\mu < \mu_0$	$\mu_0 - Z_{\alpha}\sigma_{\bar{x}}$	$- Z_{\alpha}$	$- t(\alpha,\nu)$

Pasul 3: Selectarea statisticii şi a valorii critice. Există două posibilităţi de construire a unui test:

- calcularea valorii critice a parametrului experimental ($\bar{x}$);
- să se lucreze cu variabila normală standard Z sau variabila t.

Valorile critice în fiecare caz de testare a mediei sunt prezentate în tabelul 2.4.

Pasul 4: Determinarea statisticii testului. Dacă testul este condus pe baza valorii mediei, evident că valoarea statisticii va fi chiar media experimentală. Dacă testul se bazează pe Z sau t, valoarea statisticii va fi o transformare a valorii experimentale în forma standard:

$$Z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} \quad \text{sau } t = \frac{\bar{x} - \mu_0}{S_{\bar{x}}}.$$

Pasul 5: regula de decizie. Odată stabilită valoarea critică se poate concluziona dacă ipoteza nulă se acceptă sau respinge. Este esenţial de remarcat că respingerea ipotezei H_0, implică acceptarea ipotezei H_1.

În cazul testelor cu două extreme:

- se acceptă H_0, dacă $\bar{x}$ aparţine intervalului de acceptare, determinat de cele două valori critice. $\bar{x} \in \left[\mu_0 - Z_{\alpha/2} \cdot \sigma_{\bar{x}}; \mu_0 + Z_{\alpha/2} \cdot \sigma_{\bar{x}}\right]$ şi se respinge H_0 dacă acest lucru nu se întâmplă;
- dacă se lucrează cu Z sau t: se acceptă H_0 dacă valoarea Z, respectiv t cade în interiorul domeniului de acceptare, altfel se respinge H_0.

În cazul testelor cu o extremă:

- cu statistica $\bar{x}$, se acceptă H_0 dacă $\bar{x}$ este mai mare sau mai mic decât valoarea critică, în funcţie de ipoteza alternativă, şi se respinge H_0 dacă nu se îndeplineşte condiţia;
- se procedează similar în cazul variabilei Z sau t.

Trebuie reţinut că la aplicarea unui test statistic nu se obţine niciodată o certitudine. Totdeauna apar erori α şi β. Eroarea de tip α se controlează prin specificarea lui α, nivelul de semnificaţie al testului, specificând riscul asumat de a respinge o ipoteză nulă. Eroarea β nu poate fi controlată simultan cu α, deşi ele sunt interdependente. Singura posibilitate de a reduce ambele erori este să se mărească volumul eşantionului. Cele două riscuri pot fi urmărite numai prin intermediul caracteristicii operative [4].

Exemple

1. În urma unui studiu statistic asupra producţiei unei maşini-unelte s-a constatat că se produc în medie 3.9 repere/h (repere de acelaşi tip), cu o deviaţie standard de 0.6. Urmărind un eşantion de 100 de repere identice pe o altă MU, se constată că producţia are o medie de 4.2 repere/h. Să se testeze ipoteza că producţia de 4.2 repere diferă de cea de 3.9. Se admite un nivel de semnificaţie de 0.05.

Deoarece se cunoaşte valoarea lui σ, se aplică distribuţia normală, cu variabila Z.

Date iniţiale: $\mu = 3.9$; $\sigma = 0.6$; $n = 100$

Ipoteze: H_0: $\mu_0 = 3.9$; H_1: $\mu_0 \neq \mu$

Nivel de semnificaţie $\alpha = 0.05 \Rightarrow Z_{\alpha/2} = Z_{0.025} = \pm 1.96$ (vezi Anexa 1)

Deviaţia standard a mediei: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.6}{\sqrt{100}} = 0.06$

Valori critice: pentru $\bar{x}$: $\mu_0 \pm Z_{\alpha/2}\sigma_{\bar{x}}$ =3.9 ± 1.96·0,06 ⇒3.78 şi 4.02

pentru Z: $\pm Z_{\alpha/2} \Rightarrow$ -1.96 şi 1.96

Testul: valoarea $\bar{x}$ = 4.2 cade în afara intervalului determinat de valorile critice (4.02 < 4.2). Dacă se utilizează variabila Z, statistica este:

$$Z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{4.2 - 3.9}{0.06} = 5,$$

care se găseşte în afara intervalului determinat de valorile critice.

Concluzia: Indiferent de testul aplicat se poate afirma că se respinge ipoteza nulă cu o probabilitate de 95%, deci media producţiei în cel de-al doilea caz diferă de prima. Nu se poate afirma că media eşantionului este mai mare, deoarece s-a utilizat un test cu două extreme.

2. Pe baza ezemplului 1, să se testeze dacă media eşantionului este mai mare decât media populaţiei cu o probabilitate de 95%.

Ipoteze: H_0: $\mu_0 = 3.9$; H_1: $\mu_0 > 3.9$

Nivel de semnificaţie $\alpha = 0.05 \Rightarrow Z_{\alpha} = Z_{0.05} = 1.645$

Deviaţia standard a mediei: $\sigma_{\bar{x}} = 0.06$

Valori critice: pentru $\bar{x}$: $\mu_0 + Z_{\alpha}\sigma_{\bar{x}}$ =3.9 + 1.645·0.06 = 4.0

pentru Z: $+Z_{\alpha} = 1.645$

Testul: valoarea $\bar{x}$ = 4.2 > valoare critică. Dacă se utilizează variabila Z:

$$Z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{4.2 - 3.9}{0.06} = 5 \text{ > valoare critică}$$

Concluzia: În ambele teste se respinge ipoteza nulă şi cu o probabilitate de 95% se acceptă ipoteza alternativă (productivitatea celei de-a doua MU este

mai mare).

3. Un producător de baterii afirmă că bateriile au o durată medie de viaţă de 55 h. Se testează un eşantion de 40 de baterii şi se constată că au o durată de viaţă de 50 h, iar abaterea standard a mediei este 11.734. Este adevărată afirmaţia producătorului cu un nivel de semnificaţie de 1%?

Date iniţiale: $\mu_0 = 55$; $\bar{x} = 50$; S = 11.734; n=40

Ipoteze: H_0: $\mu_0 = 55$; H_1: $\mu_0 \neq 55$

Nivel de semnificaţie $\alpha = 0.01$

Deviaţia standard a mediei: $S_{\bar{x}} = \frac{S}{\sqrt{n}} = \frac{11.734}{\sqrt{40}} = 1.86$

Valori critice: Z: $\pm Z_{\alpha/2} = \pm 2.58$

Testul : $Z = \frac{\bar{x} - \mu_0}{S_{\bar{x}}} = \frac{50 - 55}{1.86} = -2.69$. Statistica calculată > valoarea critică, deci cade în domeniul de respingere.

Concluzie: Cu o probabilitate de 99% se poate susţine că afirmaţia producătorului nu este adevărată.

În caz că se doreşte să se specifice dacă durata medie de viaţă este mai mică decât cea afirmată de producător, trebuie aplicat un test cu o singură extremă.

Ipoteze: H_0: $\mu_0 = 55$; H_1: $\mu_0 < 55$

Nivel de semnificaţie $\alpha = 0.01$

Deviaţia standard a mediei: $S_{\bar{x}} = 1.86$

Valori critice: Z: $-Z_{\alpha} = -2.33$

Testul : $Z = \frac{\bar{x} - \mu_0}{S_{\bar{x}}} = \frac{50 - 55}{1.86} = -2.69$. Statistica cade în interiorul domeniului de respingere, deci se acceptă ipoteza alternativă cu o probabilitate de 99%.

b. Media a două populaţii

În numeroase situaţii practice există două eşantioane şi se pune problema stabilirii dacă între cele două eşantioane apare o diferenţă statistică

semnificativă. În această situaţie ipoteza nulă este egalitatea dintre mediile ideale ale celor două populaţii:

H_0: $\mu_1 = \mu_2$ (sau $\mu_1 - \mu_2 = 0$), μ_1, μ_2 necunoscute

cu alternativele:

H_1: $\mu_1 \neq \mu_2$ (sau $\mu_1 - \mu_2 \neq 0$)

H_1: $\mu_1 > \mu_2$ (sau $\mu_1 - \mu_2 > 0$)

H_1: $\mu_1 < \mu_2$ (sau $\mu_1 - \mu_2 \neq 0$)

În continuare se tratează separat următoarele cazuri:
b1. eşantioane de volum mare;
b2. eşantioane de volum mic;

b1. Eşantioane independente de volum mare

În acest caz volumul celor două eşantioane trebuie să fie mai mare decât 30, n_1, $n_2 \geq 30$. Datorită acestui fapt, testul Z este mai potrivit, în baza teoremei limită centrale. Testul presupune parcurgerea aceloraşi paşi prezentaţi anterior, cu adăugarea unui singur factor, calcularea statisticii Z. Problema care se pune este stabilirea mărimii diferenţei dintre cele două medii. Pentru a putea răspunde la această întrebare, trebuie obţinute informaţii în privinţa distribuţiei $\bar{x}_1 - \bar{x}_2$, pentru toate valorile posibile $\bar{x}_1$ şi $\bar{x}_2$. Având în vedere că volumul eşantioanelor este mai mare decât 30, distribuţia $\bar{x}_1 - \bar{x}_2$ poate fi aproximată cu distribuţia normală. În plus media distribuţiei ($\mu_{\bar{x}_1 - \bar{x}_2}$) este 0, când eşantioanele sunt extrase în condiţiile ipotezei nule ($\mu_1 = \mu_2$). Se poate demonstra, [5], că abaterea standard este:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\ , \tag{3.71}$$

σ_1^2, σ_2^2 fiind varianţele celor două populaţii.

Dacă σ_1^2, σ_2^2 sunt necunoscute, trebuie utilizate valorile experimentale:

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\ . \tag{3.72}$$

Pasul 1: H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 \neq 0$ sau H_1: $\mu_1 - \mu_2 > 0$ sau H_1: $\mu_1 - \mu_2 < 0$.

Pasul 2: Stabilirea nivelului de semnificaţie α.

Pasul 3: Calcularea statisticii Z corespunzătoare diferenţei $\mu_1 - \mu_2$:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_{\bar{x}_1 - \bar{x}_2}}. \qquad (3.73)$$

În ipoteza $\mu_1 - \mu_2 = 0$, Z devine:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{S_{\bar{x}_1 - \bar{x}_2}}. \qquad (3.74)$$

Pasul 4: Compararea statisticii calculate cu valoarea critică, stabilită pe baza nivelului de semnificaţie.

Pasul 5: Acceptarea sau respingerea ipotezei nule.

Ex.: În vederea achiziţionării unor rulmenţi, un beneficiar testează doi producători. Testarea s-a făcut în condiţii identice. Datele obţinute sunt centralizate In tabelul următor:

	volum	medie [h funcţionare]	S
lot 1	$n_1 = 35$	9800	975
lot 2	$n_2 = 40$	10200	120

Să se testeze dacă: a. cele două loturi diferă; b. lotul 1 este inferior lotului 2.

a. Este necesar un test cu două extreme

Ipoteze: H_0: $\mu_1 - \mu_2 = 0$; H_1: $\mu_1 - \mu_2 \neq 0$.

Nivel de semnificaţie $\alpha = 0.05$

Deviaţia standard a mediei: $S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = \sqrt{\frac{975^2}{35} + \frac{120^2}{40}} = 28.01$

Valori critice: $\pm Z_{\alpha/2} = \pm 1.96$

Test: $Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_{\bar{x}_1 - \bar{x}_2}} = \frac{9800 - 10200}{28.01} = -14.28$

Concluzie: $|Z| > |Z_{\alpha/2}|$, valoarea calculată cade în zona de respingere, deci H_0 se respinge cu o probabilitate de 95%. Atenţie: Nu se obţin informaţii în privinţa

superiorităţii unui anumit producător.

b. Se aplică un test cu o singură extremă.

Ipoteze: H_0: $\mu_1 - \mu_2 = 0$; H_1: $\mu_1 - \mu_2 < 0$.

Nivel de semnificaţie $\alpha = 0.05$

Deviaţia standard a mediei: $S_{\bar{x}_1 - \bar{x}_2} = 28.01$

Valori critice: $- Z_{\alpha/2} = - 1.645$

Test: $Z = -14.28$

Concluzie: $Z < Z_{\alpha}$, deci H_0 se respinge şi se acceptă durata de viaţă mai ridicată a celui de-al doilea lot.

b2. Eşantioane de volum mic

În situaţia când σ_1 şi σ_2 nu sunt cunoscute, iar n_1, $n_2 < 30$ se recurge la distribuţia t. În condiţiile ipotezei nule, H_0: $\mu_1 - \mu_2 = 0$, testul statistic t se calculează cu formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}}, \tag{3.75}$$

unde $S_{\bar{x}_1 - \bar{x}_2}$ se calculează cu relaţia (2.45). Este important de remarcat că statistica t doar aproximează distribuţia t. Numărul gradelor de libertate este dat de expresia:

$$\nu = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2 \Big/ (n_1 - 1) + \left(\frac{S_2^2}{n_2}\right)^2 \Big/ (n_2 - 1)}. \tag{3.76}$$

Numărul gradelor de libertate se rotunjeşte în jos, la cel mai apropiat întreg.

În cazul particular, când se poate considera că varianţele celor două populaţii sunt egale ($\sigma_1^2 = \sigma_2^2$), se poate obţine un test mai puternic. Acest lucru poate fi considerat adevărat în cazul proceselor de lungă durată, la care din determinări anterioare s-a constatat că varianţele sunt egale. În acest caz, având două eşantioane de volum n_1, n_2, cu varianţele S_1^2, S_2^2, se calculează

varianţa compusă:

$$S_p = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}, \tag{3.77}$$

Iar abaterea standard este:

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}. \tag{3.78}$$

Statistica t, în cazul ipotezei nule (H_0: $\mu_1 - \mu_2 = 0$) este:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}}, \tag{3.79}$$

iar numărul gradelor de libertate este $\nu = n_1 + n_2 - 2$.

Exemple

1. În vederea stabilirii nivelului de pregătire a studenţilor din anul I, de la două direcţii de specializare, s-a testat un număr de studenţi pe baza unor chestionare. Ambele eşantioane au fost formate de 10 studenţi, iar rezultatele centralizate sunt prezentate în continuare (valoarea înregistrată fiind numărul de puncte realizate).

	volum	medie	S
profil 1	$n_1 = 10$	60.5	8.2
profil 2	$n_2 = 10$	62.5	10.2

Să se stabilească dacă există diferenţe semnificative între cele două eşantioane.

Datorită faptului că studenţii provin din specializări diferite nu sunt motive să se presupună că varianţele cele două populaţii ar fi egale. Se va aplica testul t, varianţele fiind necunoscute şi volumul eşantioanelor mic.

Ipoteze: H_0: $\mu_1 - \mu_2 = 0$; H_1: $\mu_1 - \mu_2 \neq 0$.

Nivel de semnificaţie $\alpha = 0.05$

Deviaţia standard a mediei: $S_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} = \sqrt{\frac{8.2^2}{10}+\frac{10.2^2}{10}} = 4.14$

Valori critice: $\pm t_{\alpha/2} = \pm 2.11$ cu $\nu = 17$

$$\nu = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2\Big/(n_1-1)+\left(\frac{S_2^2}{n_2}\right)^2\Big/(n_2-1)} = 17.2 \Rightarrow \nu = 17$$

Test: $t = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{S_{\bar{x}_1-\bar{x}_2}} = \frac{60.5-62.5}{4.14} = -0.48$

Concluzie: Deoarece $|t| < |t_{\alpha/2}|$ nu se poate respinge ipoteza nulă, deci cu o probabilitate de 95% între cele două specializări nu există diferenţe statistice semnificative.

2. Un echipament identic este utilizat în două societăţi la fabricarea aceloraşi componente. Pe o perioadă de o săptămână se înregistrează numărul defectelor în fiecare societate, obţinându-se următoarele rezultate sumarizate:

	volum	medie defecte	S
societate 1	$n_1 = 7$	8.14	3.24
societate 2	$n_2 = 7$	9.71	3.73

Să se stabilească dacă există diferenţe semnificative între cele două societăţi, respectiv dacă cea de-a doua societate are mai multe defecte decât prima.

Având în vedere că procesul de fabricaţie este identic se poate presupune că varianţele (necunoscute) sunt egale.

Ipoteze: H_0: $\mu_1 - \mu_2 = 0$; H_1: $\mu_1 - \mu_2 \neq 0$.

Nivel de semnificaţie $\alpha = 0.05$

Deviaţia standard a mediei: $S_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2}}$ cu

$$S_p = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{(7-1)3.24^2 + (7-1)3.73^2}{7+7-2} = 12.205$$

$$S_{\bar{x}_1-\bar{x}_2} = 1.867$$

Valori critice: $\pm t_{\alpha/2} = \pm 2.179$ cu $\nu = n_1 + n_2 - 2 = 12$

Test: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_{\bar{x}_1-\bar{x}_2}} = \frac{8.14 - 9.71}{1.867} = -0.84$

Concluzie: $|t| < |t_{\alpha/2}|$ (0.84 < 2.179), deci nu se poate respinge ipoteza nulă; cu o probabilitate de 95% se poate afirma că nu există diferenţe semnificative între cele două eşantioane.

Ipoteze: H_0: $\mu_1 - \mu_2 = 0$; H_1: $\mu_1 - \mu_2 < 0$.

Nivel de semnificaţie $\alpha = 0.05$

Deviaţia standard a mediei: $S_{\bar{x}_1-\bar{x}_2} = 1.867$

Valori critice: $\pm t_{\alpha} =$ -1.782 cu $\nu = n_1 + n_2 - 2 = 12$

Test: $t = -0.84$

Concluzie: $t > -t_{\alpha}$, deci nu se poate respinge H_0, cu o probabilitate de 95% cele două procese de fabricaţie sunt la fel.

BIBLIOGRAFIE

1. Baron, T., Metode statistice pentru analiza şi controlul calităţii producţiei, EDP, Bucureşti, 1979.
2. Cicală, E.F., Metode de prelucrare statistică a datelor experimentale, Ed. Politehnica Timişoara, 1999.
3. Constantinescu, I., ş.a, Prelucrarea datelor experimentale cu calculatoarele numerice, ET, Bucureşti, 1980.
4. Davidescu, A., Metrologie generală, Ed. Politehnica Timişoara, 2001.
5. Fleming, M., Nellis, J., Principles of Applied Statistics, Routledge Co., London, 1994.
6. Hanselman, D., Littlefield, B., The Student Edition of Matlab®. User's Guide, Prentice Hall, 1995.

7. Hanselman, D., Littlefield, B., Mastering Matlab® 5. A comprehensive tutorial and reference, Prentice Hall, 1998.
8. Iliescu, D.V., Vodă, V., Statistică şi toleranţe, ET, Bucureşti, 1977.
9. J de Leeuw, WEB Statistics Book, **http://www.stat.ucla.edu/textbook**
10. Martinez, W., Martinez, A., Computational Statistics Handbook with Matlab, Chapman&Hall/CRC, Washington DC, 2002.
11. Montgomery, D., Design and Analysis of Experiments, John Willey&Sons, Singapore, 1991.
12. Montgomery, D., Runger, G., Applied Statistics and Probability for Engineers, John Wiley&Sons, New York, 2006.
13. Nichici, A., ş.a., Prelucrarea datelor experimentale, Lito UPT, Timişoara, 1990.
14. Petrişor, E., Probabilităţi şi statistică. Aplicaţii în economie şi inginerie, Ed. Politehnica, Timişoara, 2005.
15. Ross, S., Initiation aux probabilités, Presses polytechniques romandes, Lausanne, 1987.
16. Ruegg, A., Probabilités et statistique, Presses polytechniques romandes, Lausanne, 1985.
17. Tiron, M., Teoria erorilor de măsurare şi metoda celor mai mici pătrate, ET, Bucureşti, 1972.
18. Tiron, M., Prelucrarea statistică şi informaţională a datelor de măsurare, ET, Bucureşti, 1976.
19. Tovissi, L., Vodă, V., Metode statistice, Ed. Ştiinţifică şi Enciclopedică, Bucureşti, 1982.
20. Wheeler, D., Chambers, D, Understanding Statistical Process Control, SPC Press, Tenessee, 1986.