CAPITOLUL 1 CALCULUL PROBABILITĂŢILOR

1.1. EVENIMENTE. PROBABILITĂŢI

1.1.1. Evenimente. Tipuri de evenimente. Relaţii între evenimente. Operaţii cu evenimente.

Conceptele fundamentale ale teoriei probabilităţilor sunt cele de *eveniment* şi *probabilitate.*

În teoria probabilităţilor se consideră experimentele ale căror rezultate sunt supuse întâmplării numite şi *experimente aleatoare.*

Deşi rezultatul experimentului nu este cunoscut dinainte, admitem totuşi că mulţimea rezultatelor sale posibile (evenimentelor) ale este cunoscută.

Definiţia 1. Mulţimea tuturor rezultatelor posibile ale unui experiment se numeşte *mulţime fundamentală* sau *spaţiu fundamental* sau încă *spaţiul evenimentelor elementare* şi se notează cu cu Ω. Spaţiul fundamental poate să fie finit sau infinit (numărabil sau nu). Numim *eveniment* orice submulţime a lui Ω. O submulţime formată dintr-un singur element din Ω este un *eveniment elementar*. Acesta este evenimentul ce apare ca rezultat al unei singure probe.

O submulţime a lui Ω formată din cel puţin două evenimente elementare este un *eveniment compus*.

Vom nota evenimentele prin litere majuscule *A,B,C,…,*însoţite uneori de indici: $A_1, A_2, \ldots$

Observaţia 1. Dacă mulţimea fundamentală este finită şi are *n* elemente (zicem în acest caz că Ω are *cardinalul* egal cu *n* şi scriem *card* Ω =*n,* atunci se ştie din teoria mulţimilor că numărul submulţimilor lui Ω (deci numărul tuturor evenimentelor) este egal cu $2^{card\Omega}$. Deci, dacă un experiment are *n* rezultate posibile, atunci legat de acest experiment există 2^n evenimente aleatoare elementare.

Exemplul 1. ***(a).*** La controlul de recepție a mărfurilor, un experiment aleator constă în cercetarea unui lot de marfă, dacă corespunde sau nu din punct de vedere al calității. Proba constă în cercetarea calității unei unități (unui articol) din marfa respectivă. Legat de acest experiment se produc 2 evenimente elementare: evenimentul A („articolul este corespunzător") şi contrarul său $\overline{A}$ („articolul nu este corespunzător").

(b). Considerăm experimentul care constă în aruncarea unui zar pe un plan orizontal. Proba constă în aruncarea zarului şi observarea feței pe care cade. Evenimentele elementere sunt asociate fețelor 1, 2, 3, 4, 5 sau 6. Atunci spațiul fundamental este format din mulțimea acestor rezultate posibile *{1,2,3,4,5,6}*. În acest caz spațiul fundamental este *finit*. Evenimentele posibile (submulțimile lui Ω) sunt în număr de 2^6=64. De exemplu: A = {2, 4, 6} (obținerea unei fețe pare), B = {1,3,5} (obținerea unei fețe impare), C={3} (eveniment elementar).

(c). Fie experimentul care constă în aruncarea succesivă a unui zar până ce se obține fața 3. Spațiul fundamental în acest caz este alcătuit din numărul aruncărilor necesare, care variază de la 1 la infinit : Ω = {1, 2, 3,...,*n*,...}. Spațiul fundamental Ω este *infinit,* însă elementele sale fiind ordonate într-un şir, Ω este un exemplu de spațiu fundamental *numărabil.*

(d). Fie experimentul care constă în măsurarea temperaturii corporale. Proba constă în efectuarea unei măsurarări a temperaturii. Un eveniment elementar constă în rezultatul citirii termometrului. Spațiul fundamental Ω este alcătuit din toate valorile posibile ale temperaturii corporale, astfel putem considera că în Ω intră toate valorile din intervalul [35, 41], sau că Ω = [35,41]. In acest caz, spațiul fundamental este o mulțime *infinită* şi *nenumărabilă*.

În mulțimea evenimentelor distingem unele evenimente remarcabile.

Definiția 2 *(Tipuri de evenimente*). Fie un experiment a cărui mulțime fundamentală este Ω.

- Evenimentul care se realizează cu certitudine în urma efectuării experimentului se numeşte *evenimentul sigur*. El se realizează dacă şi numai dacă se produce cel puțin un eveniment elementar. Ca submulțime a mulțimii fundamentale, evenimentul sigur este însuşi Ω .
- Evenimentul care nu se realizează niciodată în urma efectuării experimentului se numeşte *evenimentul imposibil*. Ca submulțime a mulțimii fundamentale evenimentul imposibil este submulțimea lui Ω care nu are niciun element, adică mulțimea vidă, notată cu $\varnothing$.

- Eveniment care se realizează ori de câte ori nu se realizează un anumit eveniment A se numeşte *contrarul (complementarul)* evenimentului A şi se notează prin $\overline{A}$ sau $\mathbf{C}A = \Omega - A$.

Exemplul 2. Extragerea unei bile albe dintr-o urnă care conţine numai bile albe este un eveniment sigur iar extragerea unei bile negre dintr-o astfel de urnă este un eveniment imposibil.

În cazul extragerii dintr-o urnă ce conţine bile albe şi bile negre, dacă notăm cu A evenimentul extragerii unei bile albe, atunci evenimentul contrar $\overline{A}$, este cel al extragerii unei bile negre. Se observă că nerealizarea lui A este echivalentă cu realizarea lui $\overline{A}$, iar nerealizarea lui $\overline{A}$ este echivalentă cu realizarea lui A.

Evenimentele contrare au următoarele proprietăţi:

Propoziţia 1. Fie Ω mulţimea fundamentală a evenimentelor elementare legate de un anume experiment şi fie A un eveniment oarecare. Atunci avem:

(a) $\overline{\overline{A}} = A$ (contrarul contrarului unui eveniment este evenimentul însuşi).

(b). $\overline{\Omega} = \varnothing$ (contrarul evenimentului sigur este evenimentul imposibil).

(c). $\overline{\varnothing} = \Omega$ (contrarul evenimentului imposibil este evenimentul sigur).

Definiţia 3. Două evenimente A şi B care se pot realiza simultan, adică dacă există probe care realizează atât pe A cât şi pe B se numesc *evenimente compatibile*. În caz contrar, dacă două evenimente A şi B nu se pot realiza simultan, se spune că ele sunt *incompatibile.*

Orice două evenimente elementare distincte sunt incompatibile.

De asemenea, două evenimente contrare unul altuia sunt incompatibile.

Exemplul 3. Considerăm la aruncarea unui zar evenimentele:

$A = \{1,2,3\}$ şi $B = \{2,3,5\}$. Dacă la aruncarea zarului vom obţine ca rezultat apariţia feţei 2, înseamnă că s-au realizat ambele evenimente. Acelaşi lucru se întâmplă dacă obţinem faţa 3. Deci evenimentele A şi B sunt compatibile.

Dacă $C = \{4,5\}$ atunci evenimentele A şi C sunt incompatibile, pe când evenimentele B şi C sunt compatibile.

Definiţia 4. Vom spune că evenimentul A *implică* evenimentul B sau că evenimentul B *este implicat* de evenimentul A, dacă B se realizează ori de câte ori se realizează A, sau dacă realizarea evenimentului A atrage realizarea lui B. Se notează în acest caz $A \subset B$.

Dacă *A* şi *B* sunt două evenimente astfel încât *A* implică *B* şi *B* implică *A,* vom scrie *A=B* şi vom spune că evenimentele *A* şi *B* sunt *echivalente.*

Observaţia 2. Se observă că $A \subset B$ revine la faptul că orice probă care realizează evenimentul *A*, realizează şi evenimentul *B*, adică la incluziunea mulțimii de probe care realizează evenimentul *A* în mulțimea de probe care realizează evenimentul *B*, ceea ce justifică notația $A \subset B$ şi faptul că:

- *evenimentul imposibil implică orice eveniment,* adică $\varnothing \subset A$, oricare ar fi evenimentul *A;*
- *orice eveniment A implică evenimentul sigur* □*, adică A*□□ oricare ar fi evenimentul *A.*

Un eveniment elementar este implicat numai de el însuşi şi de evenimentul imposibil.

Definiţia 5. Fie *A* şi *B* două evenimente legate de o aceeaşi experiență. Evenimentul care constă fie în producerea lui *A* fie a lui *B*, adică atunci când se realizează cel puțin unul din cele două evenimente, se numeşte *reuniunea* sau *disjuncția* celor două evenimente, se notează $A \cup B$ şi se mai citeşte "*A sau B".*

Mulțimea probelor care realizează evenimentul "*A sau B"* este reuniunea dintre mulțimea probelor care realizează pe *A* şi mulțimea probelor care realizează pe *B.* De aceea apare justificată folosirea notației $A \cup B$ pentru evenimentul "*A sau B"*

Definiţia 6. Evenimentul a cărui realizare constă în realizarea atât a evenimentului *A,* cât şi a evenimentului *B*, deci când se realizează ambele evenimente *A* şi *B*, se numeşte *intersecția* sau *conjuncția* celor două evenimente, se notează $A \cap B$ şi se mai citeşte "*A şi B".*

Mulțimea probelor care realizează evenimentul "*A şi B"* este intersecția dintre mulțimea probelor care realizează pe *A* şi mulțimea probelor care realizează pe *B*. De aceea apare justificată folosirea notației $A \cap B$ pentru "*A şi B".*

Dacă evenimentele *A* şi *B* sunt incompatibile, atunci mulțimea probelor care realizează pe *A* nu conține nici o probă care realizează pe *B*. Putem scrie $A \cap B = \varnothing$ (ca operație cu mulțimi).

Astfel, dacă evenimentele *A* şi *B* sunt compatibile scriem $A \cap B \neq \varnothing$ (ca operație cu evenimente).

Definițiile reuniunii şi intersecției se păstrează pentru orice număr finit de evenimente $A_1, A_2, \ldots, A_n$. Astfel, se definesc:

Evenimentul care constă în realizarea cel puțin a unuia din evenimentele $A_1, A_2, \ldots, A_n$ este *reuniunea generalizată* a acestor evenimente şi se notează $A_1 \cup A_2 \cup \ldots \cup A_n$ sau $\bigcup_{i=1}^{n} A_i$.

Evenimentul care constă în realizarea simultană a evenimentelor

$A_1, A_2, \ldots, A_n$ este <u>intersecția generalizată</u> a acestor evenimente şi se notează $A_1 \cap A_2 \cap \ldots \cap A_n$ sau $\bigcap_{i=1}^{n} A_i$.

<u>Definiția 7</u>. Numim <u>*diferența*</u> evenimentelor A şi B, evenimentul care se realizează atunci când se realizează A dar nu se realizează B şi îl notăm A-B.

Diferența evenimentelor şi A se notează □- A= **C**A el reprezentând <u>*contrarul*</u> sau <u>*complementul evenimentului A*</u>.

Operațiile de reuniune, intersecție şi diferență ale evenimentelor

satisfac proprietățile operațiilor cunoscute din algebra mulțimilor.

<u>Exemplul 4</u>. Să considerăm experiența ce constă în aruncarea unui zar. Fie evenimetul A={1,2,3}, B={2,3,4}. Atunci, avem: $A \cup B$ ={1,2,3,4}, $A \cap B$ ={2,3}, A-B={1}, B-A={4}, **C**A={4,5,6}, **C**B={1,5,6}.

<u>Propoziția 2</u>. *(Proprietățile operațiilor cu evenimente).*

Sunt adevărate relațiile de mai jos în care A, B, C sunt evenimente, □ este evenimentul sigur iar ∅ este evenimentul imposibil.

(1). $A \cup A = A$; $A \cap A = A$ *(idempotența).*

(2). $A \cup B = B \cup A$; $A \cap B = B \cap A$ *(comutativitatea).*

(3). $A \cup (B \cup C) = (A \cup B) \cup C$; $A \cap (B \cap C) = (A \cap B) \cap C$ *(asociativitatea).*

(4). $A \cup \varnothing = A$; $A \cap \varnothing = \varnothing$;(5). $A \cup \Omega = \Omega$; $A \cap \Omega = A$.

(6). $\mathbf{C}(A \cup B) = \mathbf{C}A \cap \mathbf{C}B$; $\mathbf{C}(A \cap B) = \mathbf{C}A \cup \mathbf{C}B$. *(relațiile lui De Morgan).*

(7). $A \cup \overline{A} = \Omega$; $A \cap \overline{A} = \varnothing$.

(8). $A \subset A \cup B$; $A \cap B \subset A$.
(9). $A \subset B \Rightarrow A \cup B = B$; $A \subset B \Rightarrow A \cap B = A$.
(10). $A \subset B \Rightarrow A \cup C \subset B \cup C$; $A \subset B \Rightarrow A \cap C \subset B \cap C$.

<u>Definiția 8</u>. Fie Ω mulțimea fundamentală a unei experiențe şi $\mathcal{P}(\Omega)$ mulțimea părților sale şi fie $\mathcal{K} \subset \mathcal{P}(\Omega)$ o mulțime nevidă de evenimente care satisface condițiile:

a). Dacă $A \in \mathcal{K}$ atunci **C**$A \in \mathcal{K}$;

b). Dacă $A, B \in \mathcal{K}$, atunci $A \cup B \in \mathcal{K}$;

Atunci perechea (□, $\mathcal{K}$) se numeşte <u>*câmp finit de evenimente*</u>.

Proprietatea *b)* poate fi extinsă la un număr finit de evenimente:

b') Dacă $A_1, A_2, \ldots, A_n \in \mathcal{K}$ atunci $A_1 \cup A_2 \cup \ldots \cup A_n \in \mathcal{K}$

(Atributul „finit" se referă la faptul că proprietatea *b)* este valabilă pentru un număr finit de evenimente)

Din proprietățile operațiilor cu evenimente deducem:

Propoziția 3. Dacă $(\Omega, \mathcal{K})$ este un câmp finit de evenimente atunci :

a). $\varnothing \in \mathcal{K}$, $\Omega \in \mathcal{K}$;

b). Dacă $A, B \in \mathcal{K}$, atunci $A \cup B \in \mathcal{K}$, $A - B \in \mathcal{K}$, $B - A \in \mathcal{K}$

*(*spunem în acest caz că orice corp de evenimente $\mathcal{K}$ este *închis* față de operațiile de reuniune, intersecție, diferență şi complementară).

1.1.2. Definiţiile probabilităţii.

Pentru măsurarea şanselor de realizare a unui eveniment aleator s-a introdus noțiunea de *probabilitate.* Sunt cunoscute trei definiții ale noțiunii de probabilitate: *Definiția clasică, Definiția statistică şi Definiția axiomatică*.

1º.Definiția clasică a probabilității se bazează pe procedeul de numărare a cazurilor favorabile producerii unui eveniment dintre evenimentele egal posibile (fiecare eveniment are aceeaşi şansă de a se realiza).

Definiția 9. Fie un experiment cu n evenimente elementare *egal posibile* şi fie A un eveniment oarecare (ataşat experimentului). Mulțimea probelor care realizează evenimentul A se va numi *mulțimea cazurilor favorabile lui A.* Presupunem că evenimentul A are m cazuri favorabile, $m \le n$. Se numeşte *probabilitatea evenimentului A*, numărul $P(A) = \frac{m}{n}$, adică *raportul* dintre numărul m al *cazurilor favorabile* realizării lui A şi numărul n al *cazurilor egal posibile .*

Definiția clasică a probabilității se poate folosi numai pentru

experiențele cu evenimente elementare egal posibile (evenimente care considerate împreună au aceeaşi şansă de a se realiza).

Un alt incovenient al defniției apare în cazul în care numărul cazurilor posibile este infnit deoarece în această situație probabilitatea, calculată după definiția clasică, este foarte mică sau egală cu zero.

Exemplul 5. O urnă conține *5* bile albe şi *7* bile negre. Din urnă se extrage la întâmplare o bilă. Să se calculeze probabilitatea evenimentului care constă în extragerea unei bile albe.

Soluție: Fie *A* evenimentul care constă în extragerea unei bile albe.

Pentru a calcula probabilitatea acestui eveniment observăm că avem un număr de 12 cazuri egal posibile (numărul tuturor bilelor) şi un număr de 5 cazuri favorabile lui *A*. Atunci probabilitatea acestui eveniment va fi $P(A)=\frac{5}{12}$.

Exemplul 6. Fie experiența care constă în aruncarea a două zaruri Să se calculeze probabilitatea evenimentului care constă în obținerea sumei punctelor de pe cele două zaruri egală cu 7.

Soluție: Pentru a calcula probabilitatea acestui eveniment observăm că avem un număr de 36 cazuri egal posibile şi anume toate perechile (i,j) cu $i, j = 1,2,3,4,5,6$ şi un număr de 6 cazuri favorabile lui *A* şi anume cazurile (1,6), (2,5), (3,4), (4,3), (5,2) şi (6,1). Atunci probabilitatea acestui eveniment va fi $P(A)=\frac{6}{36}=\frac{1}{6}$.

2º. Definiția statistică a probabilității exprimă probabilitatea cu ajutorul frecvenței de apariție a unui eveniment într-un număr mare de experimente realizate în aceleaşi condiții.

Noțiunea de *frecvență* este noțiune fundamentală în Statistică. Prin frecvența de apariție a unui eveniment *A* cînd se efectuează un număr de experimente realizate în aceleaşi condiții se înțelege raportul dintre numărul $n(A)$ al probelor în care evenimentul *A* s-a produs şi numărul total *n* de probe efectuate:

$$f_n(A)=\frac{n(A)}{n}.$$

Din observații statistice efectuate într-un număr mare de situații a rezultat că că dacă un experiment se repetă de un număr mare de ori, se produce o stabilitate a frecvenței jurul probabilității de apariție a evenimentului considerat. În acest mod s-a impus definiția statistică a probabilității:

Definiția 10. Dacă se efectuează de *n* ori un experiment în aceleaşi condiții iar un eveniment *A* apare de $n(A)$ ori, atunci când $n\to\infty$

$$P(A)=\lim_{n\to\infty} f_n(A).$$

Această legătura între frecvența de apariție a unui eveniment într-un număr mare de experiențe şi probabilitate constituie baza aplicațiilor calculului probabilităților în statistica matematică.

3º. Definiția axiomatică a probabilității introduce probabilitatea ca o funcție definită pe un câmp finit de evenimente cu valori în mulțimea numerelor reale.

În definiția clasică a probabilității ca raport dintre numărul cazurilor favorabile producerii unui eveniment şi numărul cazurilor posibile s-a folosit drept măsură a mulțimilor de evenimente numărul elementelor acestora. În cazul în care mulțimea asociată evenimentelor este finită, numărul *n* al elementelor ce o compun, numit *cardinalul mulțimii*, defineşte aspectul cantitativ al mulțimii, adică *măsura*

mulțimii respective.

Pentru mulțimile infinite (numărabile sau continue) noțiunea de

număr cardinal nu mai poate servi ca măsură.

Se poate generaliza prezentarea modelului de calcul al probabilității *P(A)* a unui eveniment pentru care mulțimile *A* asociate

evenimentelor sunt continue, *măsurile* lor putând fi lungimi, arii, volume, durate de timp, greutate, etc. În acest caz, notând cu *m(Ω) măsura* mulțimii asociate evenimentului sigur şi *m(A)* măsura mulțimii asociate evenimentului *A*, *probabilitatea evenimentului A* se defineşte prin expresia $P(A)=\frac{m(A)}{m(\Omega)}$, adică, *probabilitatea realizării unui eveniment* $A\in\mathcal{K}$, *este dată de raportul dintre măsura mulțimii asociate evenimentului considerat A şi măsura mulțimii asociate evenimentului sigur.*

Măsura posibilității (probabilitatea) producerii unui eveniment a cărui mulțime asociată poate fi finită sau infinită este dată de o funcție reală care este definită axiomatic astfel:

Definiția 11. Fie $(\Omega, \mathcal{K})$ un câmp finit de evenimente. Se numeşte *probabilitate* definită pe câmpul de evenimente (Ω,K) o funcție numerică pozitivă $P : \mathcal{K}\to\boldsymbol{R}_+$, care asociază fiecărui eveniment $A\in\mathcal{K}$ un număr real notat *P(A)* şi care satisface la următoarele axiome:

(1). $P(A)\geq 0,\ \forall\ A\in\mathcal{K}$;

(2). $P(\Omega)=1$;

(3). $P(A\cup B)=P(A)+P(B),\ \forall\ A,B\in\mathcal{K}$ cu $A\cap B=\emptyset$

Un câmp finit de evenimente peste care s-a definit o probabilitate se numeşte *câmp* (sau *spațiu) de probabilitate* şi se notează $(\Omega, \mathcal{K}, P)$.

Axioma (3) se poate extinde prin recurență la un număr finit de evenimente incompatibile două câte două:

(3'). $P\left(\bigcup_{i=1}^{n}A_i\right)=\sum_{i=1}^{n}P(A_i)\ ,\forall\ A_i\in\mathcal{K},\ A_i\cap A_j=\emptyset,\ i\neq j;\ i,j=1,2,\ldots,n$.

Din definiția probabilității rezultă următoarele proprietăți:

Propoziția 4. *(Proprietățile probabilităților).* Fie (Ω,K,P) un spațiu de probabilitate. Atunci au loc proprietățile :

[*P*1]. $\forall\ A\in\mathcal{K}$, atunci $0 \le P(A) \le 1$;

[*P*2]. $P(\varnothing) = 0$;

[*P*3]. $\forall\ A\in\mathcal{K}$ atunci $P(\Omega - A) = 1 - P(A)$;

[*P*4]. $\forall\ A, B\in\mathcal{K}$, $A \subseteq B$, atunci $P(A) \le P(B)$.

[*P*5]. $\forall\ A, B \in \mathcal{K}$, atunci $P(A \cup B)=P(A)+P(B)-P(A\cap B)$.

1.1.3. Probabilităţi condiţionate.

Deseori calculul probabilității apariției unui eveniment *A* este condiționat de realizarea anterioară a unui eveniment *B* cu o anumită probabilitate. De exemplu, dacă într-o urnă sunt *a* bile albe şi *b* bile negre şi extragem două bile una după alta (fără a introduce prima bilă extrasă înapoi în urnă) şi ne interesează probabilitatea evenimentului ca a doua bilă să fie neagră în condiția că prima bilă a fost albă, atunci probabilitatea acestui ultim eveniment este o *probabilitate condiționată.*

Definiţia 12. Fie $(\Omega, \mathcal{K}, P)$ un câmp de probabilitate şi fie *A* şi *B*

două evenimente ale sale astfel încât *B* să nu fie evenimentul imposibil Ø (adică să avem $P(B)>0$). Numărul notat prin $P(A\,|\,B)$ şi definit prin

$$P(A\,|\,B) = \frac{P(A \cap B)}{P(B)}, \qquad (1)$$

se numeşte *probabilitatea evenimentului A condiționat de evenimentul B.*

Dacă $P(A)>0$, se poate defini analog *probabilitatea evenimentului B condiționat de evenimentul A* , prin

$$P(B\,|\,A) = \frac{P(A \cap B)}{P(A)}, \qquad (2)$$

Probabilitatea condiționată $P(A\,|\,B)$ este probabilitatea evenimentului *A* presupunând că evenimentul *B* s-a realizat. La fel se interpretează probabilitatea condiționată $P(B\,|\,A)$.

Din formulele (1) şi (2) obținem imediat:

Propoziţia 5. *(Proprietăţile probabilităţilor condiţionate).*

Fie $(\Omega, \mathcal{K}, P)$ un câmp de probabilitate şi fie A şi B două evenimente ale sale astfel încât $P(A)>0$ şi $P(B)>0$. Atunci are loc egalitatea

$$P(A \cap B) = P(B) \cdot P(A \mid B) = P(A) \cdot P(B \mid A), \qquad (3)$$

numită *regula de înmulţire a probabilităţilor).*

Formula (3) se mai enunţă şi astfel: „ *probabilitatea producerii simultane a evenimentelor A şi B este egală cu produsul dintre probabilitatea unuia din evenimente şi probabilitatea celuilalt eveniment, dacă primul s-a realizat*".

Exemplul 7. Să considerăm experienţa aruncării unui zar şi să notăm cu A evenimentul $A = \{1,2,3\}$ şi cu B evenimentul $B = \{2,3,4\}$. Să se calculeze probabilitatea evenimentului A în ipoteza că B s-a realizat, adică $P(A \mid B)$.

Soluţie: Pentru a calcula probabilitatea evenimentului A în ipoteza că B s-a realizat, aplicăm formula (1) şi avem: $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$. Numărul cazurilor posibile este 6 iar numărul cazurilor favorabile lui B este 3. Apoi $A \cap B = \{2,3\}$ şi acest eveniment are 2 cazuri favorabile. Astfel avem $P(A) = P(B) = \dfrac{3}{6} = \dfrac{1}{2}$ şi $P(A \cap B) = \dfrac{2}{6} = \dfrac{1}{3}$. Deci, $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{1/3}{1/2} = \dfrac{2}{3}$.

Exemplul 8. Într-o grupă de studenţi sunt 10 fete şi 15 băieţi. În ora de seminar se scot la tablă simultan 2 studenţi pentru prezentarea rezolvării proprii a unei probleme. Care este probabilitatea ca ambii studenţi să fie:

a). băieti; *b).* fete; *c).* primul baiat şi al doilea fată?

Soluţie: Fie S_1 evenimentul care constă în ieşirea la tablă a primului student şi S_2 evenimentul care constă în ieşirea la tablă a celui de-al doilea student.

a). Dacă $S_1 = \{$ băiat$\}$ şi $S_2 = \{$ băiat$\}$ atunci avem de calculat :

$$P(S_1 \cap S_2) = P(S_1) \cdot P(S_2 \mid S_1) = \frac{15}{25} \cdot \frac{14}{24} = \frac{7}{20}.$$

b). Dacă $S_1 = \{$ fată$\}$ şi $S_2 = \{$ fată$\}$ atunci avem de calculat :

$$P(S_1 \cap S_2) = P(S_1) \cdot P(S_2 \mid S_1) = \frac{10}{25} \cdot \frac{9}{24} = \frac{3}{20}.$$

c). Dacă $S_1 = \{$ fată$\}$ şi $S_2 = \{$ băiat$\}$ atunci avem de calculat :

$$P(S_1 \cap S_2) = P(S_1) \cdot P(S_2 \mid S_1) = \frac{10}{25} \cdot \frac{15}{24} = \frac{1}{4}.$$

Regula de înmulțire a probabilităților poate fi generalizată la un număr de n evenimente astfel:

Propoziția 6. Fie (Ω, K, P) un câmp de probabilitate şi fie $A_1,\ldots, A_n$ evenimente din K pentru care $P(A_1 \cap A_2 \cap \ldots \cap A_n) \neq 0$, atunci avem:

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2) \cdot \\ \cdot \ldots \cdot P(A_n \mid A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \qquad (4)$$

Într-adevăr, folosind definiția probabilității condiționate, avem*:*

$$P(A_1) = P(A_1),\ P(A_2 \mid_{A_1}) = \frac{P(A_1 \cap A_2)}{P(A_1)}, P(A_3 \mid_{A_1 \cap A_2}) = \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)}, \ldots$$

$$\ldots,\ P(A_n \mid_{A_1 \cap A_2 \cap \ldots A_{n-1}}) = \frac{P(A_1 \cap A_2 \cap \cdots \cap A_n)}{P(A_1 \cap A_2 \cap \cdots \cap A_{n-1})}.$$

Relația (4) rezultă imediat prin înmulțirea membru cu membru a acestor egalități.

Exemplul 9. La jocul Loto "6 din 49" dintr-o urnă care conține bile uniforme numerotate de la 1 la 49, se extrag 6 bile, fără a pune bila extrasă înapoi în urnă. Care este probabilitatea ca varianta (1,2,3,4,5,6) să aibă toate numerele câştigătoare.

Soluție: Fie A_i evenimentul ca la extragerea *i-a* numărul extras să

fie *i.* Avem : $P(A_1) = \frac{1}{49},\ P(A_2 \mid A_1) = \frac{1}{48},\ P(A_3 \mid A_1 \cap A_2) = \frac{1}{47},$

$P(A_4 \mid A_1 \cap A_2 \cap A_3) = \frac{1}{46},\ P(A_5 \mid A_1 \cap A_2 \cap A_3 \cap A_4) = \frac{1}{45},$

$P(A_6 \mid A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) = \frac{1}{44}.$

Atunci conform regulii de înmulțire a probabilităților, formula (4) avem:

$$P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5 \cap A_6) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2) \cdot \\ \cdot P(A_4 \mid A_1 \cap A_2 \cap A_3) \cdot P(A_5 \mid A_1 \cap A_2 \cap A_3 \cap A_4) \cdot \\ \cdot P(A_6 \mid A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) = \\ = \frac{1}{49} \cdot \frac{1}{48} \cdot \frac{1}{47} \cdot \frac{1}{46} \cdot \frac{1}{45} \cdot \frac{1}{44} = \frac{1}{10.068.347.520}$$

Dacă nu se ținea seama de ordinea de apariție a numerelor, avem 6 ! variante formate cu numerele (1,2,3,4,5,6), astfel că probabilitatea ca o astfel de variantă să fie câştigătoare este

$$\frac{1}{44}\cdot\frac{1}{45}\cdot\frac{1}{46}\cdot\frac{1}{47}\cdot\frac{1}{48}\cdot\frac{1}{49}\cdot 6! = \frac{720}{10.068.347.520} \approx 0{,}00000007 \ .$$

Definiţia 13. Fie $(\Omega, \mathcal{K}, P)$ unui câmp de probabilitate. Două evenimente A şi B ale sale se numesc *P - independente* dacă

$$P(A \cap B) = P(A) \cdot P(B) \ , \qquad (5)$$

adică probabilitatea unuia dintre evenimente nu depinde de faptul că celălalt eveniment s-a produs sau nu.

Din regula de înmulțire a probabilităților, (formula (3)), deducem că evenimentele *A* şi *B* sunt *P*-independente dacă şi numai dacă $P(A|B)=P(A)$ şi $P(B|A)=P(B)$.

Observaţia 3. Două evenimente contrare *A* şi $\overline{A}$ nu sunt în general *P*-independente. Într-adevăr, fie $p=P(A)$ şi $P(\overline{A}) = 1-p$. Cum $A \cap \overline{A} = \varnothing$ avem $p(A \cap \overline{A}) = 0$. Pentru ca *A* şi $\overline{A}$ să fie independentre trebuie ca $p(1-p)=0$. Deci o condiție necesară şi suficientă ca evenimentele contrare *A* şi $\overline{A}$ să fie independente, este ca $P(A)=0$ sau $P(A)=1$, adică *A* să fie evenimentul sigur Ω au evenimentul imposibil $\varnothing$.

Independența şi *incompatibilitatea* nu trebuie confundate. Relația de incompatibilitate este definită fără referință la o probabilitate prin rela-ția $A \cap B = \varnothing$, pe când independența depinde de o probabilitate *P*.

În probleme avem de a face de cele mai multe ori cu evenimente a căror independență poate fi inutilă. Este cazul evenimentelor diferite, considerate împreună şi care nu-şi influențează unul altuia rezultatele. De exemplu, aruncarea a două zaruri constă în aruncarea primului zar şi aruncarea celui de-al doilea zar. Şansele apariției feței 1 la primul zar nu influențează şi nu sunt influențate în nici un fel de apariția unei anumite fețe la cel de-al doilea zar.

De asemenea, evenimente independente apar atunci când acestea sunt legate de experimente diferite.

Extindem definiția independenței pentru un număr finit de evenimente.

Definiţia 14. Fie $(\Omega, \mathcal{K}, P)$ un câmp de probabilitate. Evenimentele $A_1, A_2, \ldots, A_n$ sunt *P - independente* dacă şi numai dacă

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_n) \, , \qquad (6)$$

Dacă evenimentele $A_1, A_2, \ldots, A_n$ sunt *P* - independente atunci şi

evenimentele $B_k \in \{A_k, \overline{A}_k\}$, $k = 1,2,...,n$ sunt P - independente.

1.2. FORMULE PENTRU CALCULUL UNOR PROBABILITAŢI

1.2.1. Formule de calcul pentru intersecţii şi reuniuni de evenimente.

1º. Probabilitatea unei intersecții de evenimente.

Reamintim din tema precedentă formula care generalizează *Regula de înmulțire a probabilităților* care dă formula de calcul pentru intersecția unui număr finit de evenimente compatibile*:*

Propoziţia 7. Fie $(\Omega, \mathcal{K}, P)$ un câmp de probabilitate şi fie $A_1,..., A_n$ evenimente din $\mathcal{K}$ pentru care $P(A_1 \cap A_2 \cap ... \cap A_n) \neq 0$, atunci avem:

$$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2) \cdot \\ \cdot ... \cdot P(A_n \mid A_1 \cap A_2 \cap ... \cap A_{n-1}) \tag{7}$$

De asemenea, dacă evenimentele $A_1, A_2, \ldots, A_n$ sunt P - *independente* atunci formula (7) devine:

$$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1) \cdot P(A_2) \cdot ... \cdot P(A_n), \tag{8}$$

2º. Probabilitatea unei reuniuni de evenimente incompatibile.

Propoziţia 8. Fie $(\Omega, \mathcal{K}, P)$ un câmp de probabilitate.

(a). Dacă A şi B sunt evenimente din $\mathcal{K}$, şi incompatibile $(A \cap B = \varnothing)$ atunci :

$$P(A \cup B) = P(A) + P(B) \tag{9}$$

(b). Dacă $A_1, A_2, \ldots, A_n$ sunt evenimente din $\mathcal{K}$ şi incompatibile două câte două $(A_i \cap A_j = \varnothing,\ i \neq j)$ atunci

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) \tag{10}$$

Într-adevăr, formula (9) este chiar axioma (3) din *Definiția axiomatică* a probabilităților. Demonstrarea lui *(b)* se face prin inducție, folosind formula (9).

3º. Probabilitatea unei reuniuni de evenimente compatibile.

Din proprietățile probabilităților rezultă următoarea formulă pentru calculul evenimentuilui reuniune:

Propoziţia 9. Fie $(\Omega, \mathcal{K}, P)$ un câmp de probabilitate.

(a). Dacă A şi B sunt două evenimente oarecare din $\mathcal{K}$, atunci

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{11}$$

(b). Dacă $A_1, A_2, \ldots, A_n$ sunt evenimente oarecare din $\mathcal{K}$, atunci

$$P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} P(A_k) - \sum_{j<k} P(A_j \cap A_k) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \ldots$$

$$+ \ldots + (-1)^{n-1} \cdot P\left(\bigcap_{k=1}^{n} A_k\right), \quad \text{(Formula lui Poincaré)} \tag{12}$$

Exemplul 10. O urnă conține 3 bile albe şi 7 bile negre, iar alta conține 7 bile albe şi 3 bile negre. Din fiecare urnă se extrage câte o bilă. Care este probabilitatea să obținem cel puțin o bilă albă?

Soluție: Fie A evenimentul extragerii unei bile albe din prima urnă şi B evenimentul extragerii unei bile albe din a doua urnă. Avem de calculat probabilitatea evenomentului $A \cup B$.

Avem 10 cazuri posibile pentru fiecare experiență, 3 cazuri favorabile lui A şi 7 cazuri favorabile lui B. Calculând probabilitățile acestor evenimente avem $P(A) = \frac{3}{10}, P(B) = \frac{7}{10}$; apoi, deoarece A şi B sunt independente avem $P(A \cap B) = P(A) \cdot P(B) = \frac{3}{10} \cdot \frac{7}{10} = \frac{21}{100}$. Astfel obținem: $P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{3}{10} + \frac{7}{10} - \frac{21}{100} = \frac{79}{100}$.

4º. Probabilitatea unei reuniuni de evenimente compatibile şi independente.

Propoziţia 10. Fie $(\Omega, \mathcal{K}, P)$ un câmp de probabilitate. Dacă evenimentele $A_1, A_2, \ldots, A_n$ sunt P–independente atunci au loc egalitățile:

$$(a).\ P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} P(A_k) - \sum_{j<k} P(A_j) \cdot P(A_k) +$$

$$+\sum_{i<j<k} P(A_i)\cdot P(A_j)\cdot P(A_k)+\ldots+(-1)^{n-1}\cdot\prod_{k=1}^{n}P(A_k) \qquad (13)$$

(b). $P\left(\bigcup_{k=1}^{n}A_k\right)=1-\prod_{k=1}^{n}(1-P(A_k))$ □

(14) *Demonstraţie:* Pentru *(a)* se ţine seama de formula lui Poincaré . Pentru (*b*), ţinând seama de proprietăţilor operaţiilor cu evenimente şi proprietăţile probabilităţilor avem succesiv :

$$P\left(\bigcup_{k=1}^{n}A_k\right)=1-P\left(\overline{\bigcup_{k=1}^{n}A_k}\right)=1-P\left(\bigcap_{k=1}^{n}\overline{A}_k\right)=1-\prod_{k=1}^{n}P\left(\overline{A}_k\right)=1-\prod_{k=1}^{n}(1-P(A_k))$$ □

Exemplul 11. Se aruncă o monedă de 3 ori. Care este probabilitatea să apară stema cel puţin odată ?. Să se rezolve problema folosind cele două formule.

Soluţie. Dacă *A* este evenimentul apariţiei stemei la prima aruncare, *B* este evenimentul apariţiei stemei la a doua aruncare şi *C* este evenimentul apariţiei stemei la a treia aruncare, atunci avem de calculat $P(A\cup B\cup C)$.Ţinând seama că evenimentele sunt independente, avem:

$$P(A)=P(B)=P(C)=\frac{1}{2},\ P(A\cap B)=P(A\cap C)=P(B\cap C)=\frac{1}{4},\ P(A\cap B\cap C)=\frac{1}{8}.$$

Folosind formula (12) obţinem

$$P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)-P(B\cap C)+$$
$$+P(A\cap B\cap C)=\frac{1}{2}+\frac{1}{2}+\frac{1}{2}-\frac{1}{4}-\frac{1}{4}-\frac{1}{4}+\frac{1}{8}=\frac{7}{8}.$$

Folosim formula (14), avem:

$$P(A\cup B\cup C)=1-\left[1-P(A)\right]\cdot\left[1-P(B)\right]\cdot\left[1-P(C)\right]=1-\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}=1-\frac{1}{8}=\frac{7}{8}.$$

1.2.2. Sistem complet de evenimente. Formula probabilităţii totale. Formula lui Bayes.

Definiţia 15. Fie $(\Omega, \mathcal{K})$ un câmp finit de evenimente. Se numeşte *sistem complet de evenimente* o mulţime de evenimente $E=\{A_1, A_2, \ldots, A_n\}\subset \mathcal{K}$ astfel încât:

(a). Evenimentele sunt incompatibile două câte două (adică $A_i \cap A_j = \varnothing$, pentru $i \neq j$) ;

(b). $A_1 \cup A_2 \cup \cdots A_n = \Omega$.

Propoziţia 11. Fie *(Ω, K, P)* un câmp de probabilitate, fie $E=\{A_1, A_2,\ldots,A_n\}$ un sistem complet de evenimente şi *A* este un evenimement oarecare care se realizează cu unul din aceste evenimente. Atunci are loc formula:

$$P(A) = P(A_1) \cdot P(A \mid A_1) + P(A_2) \cdot P(A \mid A_2) + \cdots + P(A_n) \cdot P(A \mid A_n), \qquad (15)$$

numită *formula probabilităţii totale*.

Demonstraţie : *Într-adevăr* , putem scrie

$$A = A \cap \Omega = A \cap (A_1 \cup A_2 \cup \ldots \cup A_n) = (A_1 \cap A) \cup (A_2 \cap A) \cup \cdots \cup (A_n \cap A)$$

de unde, ţinând seama de faptul că evenimentele $(A_i \cap A), (i = 1,2,\ldots,n)$, sunt incompatibile două cîte două, obţinem:

$$P(A) = P(A_1 \cap A) + P(A_2 \cap A) + \cdots + P(A_n \cap A).$$

Apoi cum pentru fiecare *i= 1, 2, …, n,* folosind formula de calcul

pentru o intersecţie de evenimente avem $P(A_i \cap A) = P(A_i) \cdot P(A \mid A_i)$, de unde deducem

$$P(A) = P(A_1) \cdot P(A \mid A_1) + P(A_2) \cdot P\ (A \mid A_2) + \cdots + P(A_n) \cdot P\ (A \mid A_n).$$

Exemplul 12. Două firme fabrică produse de acelaşi fel pe care le desfac pe o anumită piaţă. Prima firmă produce 40% din necesarul pieţii iar din produsele fabricate 85% corespund normelor de fabricaţie. A doua firmă produce restul de 60% din necesarul pieţii, iar din produsele fabricate 90% corespund normelor de fabricaţie. Se cere probabilitatea ca un produs achiziţionat de pe piaţă să corespundă normelor de fabricaţie.

Soluţie. Să notăm cu *A* evenimentul care constă în faptul că produsul achiziţionat este corespunzător. Notăm cu A_1 evenimentul ca produsul achiziţionat să provină de la prima firmă şi cu A_2 evenimentul ca produsul să provină de la a doua firmă. Atunci avem $A = (A \cap A_1) \cup (A \cap A_2)$, şi aplicând formula probabilităţii totale rezultă:

$$P(A) = P(A_1) \cdot P(A \mid A_1) + P(A_2) \cdot P(A \mid A_2).$$

Dar $P(A_1) = \frac{40}{100}$, $P(A_2) = \frac{60}{100}$, $P(A \mid A_1) = \frac{85}{100}$, $P(A \mid A_2) = \frac{90}{100}$, deci

$$P(A) = \frac{40}{100} \cdot \frac{85}{100} + \frac{60}{100} \cdot \frac{90}{100} = \frac{88}{100} = 0{,}88 .$$

Observaţia 4. Fie A şi B două evenimente oarecare şi $\overline{B}$ contrarul lui B. Evenimentele $\{B, \overline{B}\}$ formează un sistem complet de evenimente. Atunci formula (15) se scrie :

$$P(A) = P(A \mid B) \cdot P(B) + P(A \mid \overline{B}) \cdot P(\overline{B}) \tag{16}$$

sau

$$P(A) = P(A \mid B) \cdot P(B) + P(A \mid \overline{B}) \cdot [1 - P(B)]. \tag{17}$$

Să aplicăm formulele probabilităţilor condiţionate şi cele de mai sus la exemplele următoare:

Exemplul 13. O societate de Asigurări estimează că oamenii se împart în două categorii : aceia care nu prezintă risc mare pentru accidente şi care reprezintă 30% din populaţie şi ceilalţi care prezintă risc moderat. Statisticile pe care le deţine îi arată că primii se accidentează într-un an cu o probabilitate de 0,4, în timp ce cei din a doua categorie, cu o probabilitate de 0,2.

a). Care este probabilitatea ca un nou asigurat să fie victima unui accident în anul care urmează după încheierea poliţei de asigurare?

b). Care este probabilitatea ca noul asigurat să facă parte din categoria de risc major ?

Soluţie : *a).* Notăm cu A evenimentul ca semnatarul poliţei de asigurare să aibă în anul următor un accident şi cu B evenimentul ca acesta să fie din categoria celor cu risc major. Conform formulei (17) avem :

$$P(A) = P(A \mid B) \cdot P(B) + P(A \mid \overline{B}) \cdot [1 - P(B)] = \frac{30}{100} \cdot (0{,}4) + \frac{70}{100} \cdot (0{,}2) = 0{,}26$$

b). Probabilitatea ca noul asigurat să facă parte din categoria de

risc major este $P(B \mid A)$. Aplicând succesiv formulele (1) şi (3)

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(B) \cdot P(B \mid A)}{P(A)} = \frac{(0{,}3) \cdot (0{,}4)}{0{,}26} = \frac{6}{13} .$$

Exemplul 14. Un laborator de analize medicale asigură cu o fiabilitate de 95% detectarea unei anumite boli cînd pacientul o are efectiv. Totuşi, testul poate indica şi rezultate false "pozitive" pentru 1% din persoanele sănătoase la care se aplică (adică 1 persoană din 100 sănătoase poate fi declarată bolnavă). Dacă 0,5 % din populaţie are

efectiv boala respectivă, care este probabilitatea ca o persoană supusă testului să fie cu adevărat bolnavă dacă testul a indicat astfel.

(Probleme asemănătoare : *testul antidoping* la sportivi, *teste de viruşi ş.a.)*

Soluţie: Fie B evenimentul ca persoana supusă testului să fie cu adevărat bolnavă şi A evenimentul ca rezultatul testului să fie pozitiv (adică, să indice boala). Probabilitatea căutată este $P(B \mid A)$. Aplicând succesiv formulele (1), (3) şi (16) avem :

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \mid B) \cdot P(B)}{P(A \mid B) \cdot P(B) + P(A \mid \overline{B}) \cdot P(\overline{B})} =$$

$$= \frac{(0{,}95) \cdot (0{,}005)}{(0{,}95) \cdot (0{,}005) + (0{,}01) \cdot (0{,}995)} = \frac{95}{294} \approx 0{,}323 .$$

Astfel, numai 32% dintre persoanele ale căror rezultate la test sunt "pozitive" sunt cu adevărat bolnave. Acest rezultat este surprinzător deoarece ne-am fi aşteptat la o valoare mult mai ridicată, motiv pentru care un astfel de exemplu este grăitor pentru interpretarea rezultatelor unor teste de acest fel.

Altă rezolvare a problemei: Deoarece 0,5% din populaţie are în mod real boala respectivă, din 200 de persoane testate în medie 1 o va avea. Testul descoperă aceasta cu o probabilitate de 0,95. În medie deci, din 200 de persoane testate, vor fi detectate corect 0,95 cazuri. Pe de altă parte, printre cele 199 persoane sănătoase testul indică bolnave $199 \cdot (0{,}01)$ dintre acestea. Rezumând, la 0,95 cazuri de maladie corect detectată se adaugă în medie 1,99 cazuri fals bolnave (persoane sănătoase în realitate). Făcând proporţia rezultatelor corecte când testul este pozitiv obţinem :

$$P(B \mid A) = \frac{0{,}95}{0{,}95 + 199 \cdot (0{,}01)} = \frac{95}{294} \approx 0{,}323 .$$

Exemplul 15. Un poliţist criminalist, însărcinat cu o anchetă asupra unei crime, este la un moment dat convins cu 60% de culpabilitatea unui anumit suspect. El descoperă o nouă probă care îi permite să afirme că suspectul respectiv este criminalul, acesta având un anumit atribut (era stângaci, sau chel, sau avea părul negru). Or, 20% din populaţie avea acel atribut. Cu ce probabilitate va reaprecia poliţistul culpabilitatea suspectului dacă el găseşte că acesta are atributul respectiv ?

Soluţie : Notăm cu C evenimentul ca suspectul să fie culpabil şi prin A evenimentu ca el să aibă acelaşi atribut ca şi criminalul. Avem de calculat $P(C \mid A)$. Aplicând succesiv formulele (1), (3) şi (16) avem :

$$P(C \mid A) = \frac{P(C \cap A)}{P(A)} = \frac{P(A \mid C) \cdot P(C)}{P(A \mid C) \cdot P(C) + P(A \mid \overline{C}) \cdot P(\overline{C})} =$$

$$= \frac{1 \cdot (0,6)}{1 \cdot (0,6) + (0,2) \cdot (0,4)} = 0,882$$

Deci, convingerea că suspectul este adevăratul criminal a crescut de la 60% la peste 88%.

Fie $E=\{A_1, A_2, \ldots, A_n\}$ un sistem complet de evenimente şi fie A este un evenimemnt oarecare. Un astfel de eveniment poate apărea ca efect al celor n evenimente $A_1, A_2, \ldots, A_n$ (adică A poate apărea odată cu unul şi numai cu unul dintre evenimentele A_i, $i=1,2,\ldots,n$.

Cunoscându-se probabilităţile $P(A_i)$, $i=1,2,\ldots,n$, (care se mai numesc şi *probabilităţi a priori* ale evenimentelor A_i) şi probabilităţile de apariţie a evenimentului A ca efect al evenimentului A_i, deci $P(A \mid A_i)$, $i=1,2,\ldots,n$, se cere a se calcula probabilităţile evenimentelor A_i în situaţia că A s-a produs, adică probabilităţile $P(A_i \mid A)$, numite şi

probabilităţi a posteriori ale evenimentelor A_i .

Propoziţia 12. Fie $(\Omega, \mathcal{K}, P)$ un câmp de probabilitate, fie $E=\{A_1, A_2,\ldots,A_n\}$ un sistem complet de evenimente şi A este un evenimement oarecare care se realizează cu unul din aceste evenimente. Atunci probabilităţile *a posteriori* se calculează cu formula

$$P(A_i \mid A) = \frac{P(A_i) \cdot P(A \mid A_i)}{\sum_{i=1}^{n} P(A_i) \cdot P(A \mid A_i)}, \quad (\textit{formula lui Bayes}) \qquad (18)$$

Demonstraţie : Ţinând seama de regula de înmulţire a probabilităţilor , putem scrie

$$P(A \cap A_i) = P(A) \cdot P(A_i \mid A) = P(A_i) \cdot P(A \mid A_i),$$

de unde deducem: $P(A_i \mid A) = \dfrac{P(A_i) \cdot P(A \mid A_i)}{P(A)}$

Înlocuind în această egalitate $P(A)$ dată de formula (6) a probabilităţii totale, deducem formula (18).

Observaţia 5. Evident că, dacă evenimentele A_i sunt egal posibile, atunci formula lui Bayes se scrie

$$P(A_i \mid A) = \frac{P(A \mid A_i)}{\sum_{i=1}^{n} P(A \mid A_i)}, \qquad (19)$$

Observaţia 6. Dacă tratăm evenimentele A_i ca *ipoteze* posibile într-o anumită problemă, formula lui *Bayes* joacă un rol util în a ne arăta că opiniile *a priori* asupra acestor ipoteze [şi anume $P(A_i)$] trebuie modificate în lumina rezultatului experienței.

Exemplul 16. Două maşini automate care fabrică acelaşi tip de piese şi au aceeaşi productivitate, realizează piese rebut cu probabilitatea $p_1 = 0{,}05$ şi respectiv $p_2 = 0{,}02$. Pentru efectuarea unui control de calitate, din producția celor două maşini se extrage la întâmplare o piesă.

a). Care este probabilitatea ca piesa extrasă să fie rebut? Dar ca ea să fie corespunzătoare?

b). Ştiind că piesa extrasă este un rebut, să se afle probabilitatea ca ea să provină de la prima maşină.

*c).*Ştiind că piesa extrasă este corespunzătoare, care este probabilitatea ca ea să provină de la a doua maşină.

Soluții: *a).* Fie A_i evenimentul ca piesa extrasă să provină de la maşina *"i", i=1,2.* Fie A evenimentul ca piesa extrasă să fie rebut. Evenimentele A_1 şi A_2 formează un sistem complet de evenimente. Probabilitatea lui A se calculează cu formula probabilității totale, unde: $P(A_1) = P(A_2) = 0{,}5$, $P(A \mid A_1) = 0{,}05, P(A \mid A_2) = 0{,}02$.

$$P(A) = P(A_1) \cdot P(A \mid A_1) + P(A_2) \cdot P(A \mid A_2) = 0{,}5 \cdot 0{,}05 + 0{,}5 \cdot 0{,}02 = 0{,}035$$

Fie B evenimentul ca piesa extrasă să fie corespunzătoare, $B = \overline{A}$. Probabilitatea lui B se calculează cu formula:

$$P(B) = P(\overline{A}) = 1 - P(A) = 1 - 0{,}035 = 0{,}965$$

b). Probabilitatea $P(A_1 \mid A)$ se poate calcula cu formula lui Bayes şi se obține:

$$P(A_1 \mid A) = \frac{P(A_1) \cdot P(A \mid A_1)}{P(A)} = \frac{0{,}5 \cdot 0{,}05}{0{,}035} = 0{,}714 .$$

c). Probabilitatea $P(A_2 \mid B)$ se poate calcula cu formula lui Bayes şi se obține:

$$P(A_2 \mid B) = \frac{P(A_2) \cdot P(B \mid A_2)}{P(B)} = \frac{0{,}5 \cdot 0{,}98}{0{,}965} = 0{,}507 .$$

1.2.3. Scheme probabilistice.

În această secțiune punem în evidență acum unele scheme de calcul al probabilităților modelate cu urne, la care se reduc multe probleme de calcul întâlnite în practică şi în teorie.

1º. Schema lui Bernoulli (schema bilei întoarse).

Propoziția 13. Se consideră o urnă în care sunt a bile albe şi b bile negre. Din această urnă se fac n extrageri succesive, punându-se de fiecare dată bila înapoi în urnă (*urna lui Bernoulli*).

Probabilitatea $P(n,k)$ ca din cele n extrase k bile să fie albe şi n-k bile negre este dată de formula:

$$P(n,k) = C_n^k \frac{a^k \cdot b^{n-k}}{(a+b)^n} \tag{20}$$

Demonstrație : Dacă în cadrul unei extrageri notăm cu A evenimentul extragerii unei bile albe de probabilitate şi cu $\overline{A}$ evenimentul contrar (extragerea unei bile negre) atunci probabilitățile

lor sunt: $p = P(A) = \frac{a}{a+b}$ şi $q = P(\overline{A}) = \frac{b}{a+b} = 1 - p$.

Datorită condițiilor de realizare a experimentului (introducerea bilei extrase înapoi în urnă), fapt pentru care această schemă se mai numeşte şi *schema bilei întoarse,* probabilitățile p şi q rămân constante tot timpul experienței. Fie B evenimentul ce constă în realizarea succesiunii

$$\underbrace{A, A, \ldots, A}_{de\ k\ ori} \text{ şi } \underbrace{\overline{A}, \overline{A}, \ldots, \overline{A}}_{de\ (n-k)\ ori}.$$

Notând cu $\overline{A}_i$ evenimentul $\overline{A}$ realizat la experiența de rangul „*i*” *(i=1,2,…,k-1)* şi cu A_i evenimentul A realizat la experiența de rangul „*j*” *(i=k,k+1,,…,n)* rezultă că evenimentul B cerut de problemă este:

$$B = (A_1 \cap A_2 \cap \ldots \cap A_{k-1}) \cap (\overline{A}_k \cap \overline{A}_{k+1} \cap \ldots \cap \overline{A}_n).$$

Probabilitatea acestui eveniment este este:

$$P\left[\left(\underbrace{A \cap A \cap \ldots \cap A}_{de\ k\ ori}\right) \cap \left(\underbrace{\overline{A} \cap \overline{A} \cap \ldots \cap \overline{A}}_{de\ (n-k)\ ori}\right)\right] = p^k q^{n-k}.$$

Numărul succesiunilor în care apare *A* de *k* ori şi $\overline{A}$ de *n-k* ori este de C_n^k. Probabilitatea *P(n;k)* este dată de probabilitatea acestor succesiuni distincte. Cum aceste succesiuni sunt incompatibile şi

echiprobabile, avem: $P(n,k) = C_n^k p^k q^{n-k} = C_n^k \frac{a^k b^{n-k}}{(a+b)^n}$.

Observaţia 7. „*Urna lui Bernoulli*" modelează un experiment care se repetă în aceleaşi condiţii de *n* ori şi în care poate să apară fie un eveniment *A* cu aceeaşi probabilitate $p = P(A)$ fie contrarul său $\overline{A}$ cu probabilitatea $q = P(\overline{A}) = 1 - p$. Probabilitatea *P(n;k)*, ca în cele *n* repetări

ale experimentului, evenimentul *A* să apară de *k* ori, este

$$P(n,k) = C_n^k p^k q^{n-k}, \qquad (21)$$

Aceeaşi problemă rezultă şi din deducerea probabilităţii de apariţie de *k* ori a unui eveniment *A* dacă se efectuează *n* experienţe independente şi dacă în fiecare experienţă probabilitatea de apariţie a evenimentului A este constantă şi este egală cu *p*.

Deoarece probabilitatea *P(n,k)* este coeficientul lui x^k din dezvoltarea binomului $(px+q)^k$, această schemă se mai numeşte *schema binomială* sau că probabilitatea respectivă reprezintă o *lege binomială.*

Exemplul 17. O urnă conţine *3* bile albe şi *4* bile negre. Din această urnă se fac *3* extrageri succesive, punându-se de fiecare dată bila extrasă înapoi. Care este probabilitatea de a obţine *2* bile albe şi 1 bilă neagră?

Soluţie. Suntem în cadrul schemei lui Bernoulli ($p = \frac{3}{7}$, $q = 1 - p = \frac{4}{7}$, *n=3, k=2*). Atunci avem:

$$P(3,2) = C_3^2 \left(\frac{3}{7}\right)^2 \left(\frac{4}{7}\right)^1 = \frac{108}{343} = 0{,}315.$$

Schema lui Bernoulli (binomială) poate fi generalizată la o urnă care coţine bile de mai multe culori astfel:

Propoziţia 14. *(Schema multinomială).* Fie o urnă cu bile de s culori. Fie p_k probabilitatea extragerii unei bile de culoare ($k=1,...,s$). Atunci probabilitatea ca în cele n bile extrase să obţinem n_k bile de culoarea k. ($k=1,...,s$) este dată de formula :

$$P(n;n_1,n_2,\ldots,n_s)=\frac{n!}{n_1!\cdot n_2!\cdot\ldots\cdot n_s!}\cdot p_1^{n_1}\cdot p_2^{n_2}\cdot\ldots\cdot p_s^{n_s} \qquad (22)$$

Exemplul 18. Un magazin primeşte în cursul unei săptămâni 100 de televizoare provenite de la fabricile *A, B, C.* Probabilitatea ca televizoarele să provină de la fabrica *A* este de 0,6; de la fabrica *B* este de 0,2; de la fabrica *C* este de 0,2. Care este probabilitatea ca din cele 100 de televizoare primite, 60 să fi fost realizate la fabrica A, 30 la fabrica B, iar restul la C?

Soluţie : Aplicând schema multinomială avem :

$$P(100;60,30,10)=\frac{100!}{60_1!\cdot 30!\cdot 10!}\cdot(0,6)^{60}\cdot(0,2)^{30}\cdot(0,2)^{10}.$$

2o. Schema lui Pascal (legea geometrică).

Propoziţia 15. În urma efectuării unui experiment poate apărea evenimentul *A* cu probabilitatea p, sau contrariul său cu probabilitatea $q=1-p$. Se repetă experimentul de n ori, în condiţii identice.

Probabilitatea $P(n;k)$ ca în cele n repetări ale experimentului, evenimentul *A* să apară la cel de ordinul k este

$$P(n;k)=p\cdot q^{k-1} \qquad (23)$$

Demonstraţie : Într-adevăr, dacă *B* este evenimentul care constă în realizarea lui *A* la experienţa de ordinul k, atunci este necesar ca la toate cele (k-1) repetări anterioare să fi avut loc evenimentul contrar $\overline{A}$. Notând cu $\overline{A}_i$ evenimentul $\overline{A}$ realizat la experienţa de rangul i rezultă că evenimentul *B* cerut de problemă este:

$$B=\underbrace{(\overline{A}_1\cap\overline{A}_2\cap\ldots\cap\overline{A}_{k-1})}_{\text{de } (k-1) \text{ ori}}\cap A .$$

Evenimentele fiind independente şi $P(A)=p, P(\overline{A}_i)=q$,$i=1,2,\ldots,k-1$, avem:

$$P(B)=P\left[\left(\overline{A}_1\cap\overline{A}_2\cap\ldots\cap\overline{A}_{k-1}\right)\cap A\right]=P(\overline{A}_1)\cdot P(\overline{A}_2)\cdot\ldots\cdot P(\overline{A}_{k-1})\cdot P(A)=q^{k-1}\cdot p$$

Observaţia 8. Pentru diferite valori ale lui k=1,2,..., numerele $p\cdot q^{k-1}$

constituie o progresie geometrică de prim termen p şi de raţie q, de

aceea legea respectivă se mai numeşte şi *legea geometrică.*

Exemplul 19. Considerăm evenimentele: *B,* naşterea unui băiat, cu probabilitatea p=0,49; *F,* naşterea unei fete, cu probabilitatea q=0,51. Care este probabilitatea ca într-o familie de 6 copii un băiat să se nască la a şasea naştere după ce la primele naşteri au fost fete.

Soluţie. Notând cu *A* este evenimentul ca băiatul să apară la a şasea naştere, suntem în cazul legii geometrice cu $k=6$. Avem deci, $P(A)=0,49\cdot 0,51^6=0,0169$.

3o. Schema lui Poisson.

Propoziţia 16. Se fac n experimente independente. În urma experimentului de rang k poate apărea evenimentul *A* cu probabilitatea p_k sau contrarul său $\overline{A}$ cu probabilitatea $q_k=1-p_k$ ori, $k=1,2,\ldots,n$.

Probabilitatea $P(n,m)$ ca în cele n experienţe, evenimentul *A* să apară de m ori, este egală cu coeficientul a_m al lui x^m din dezvoltarea polinomului

$$(p_1x+q_1)(p_2x+q_2)\cdot\ldots\cdot(p_nx+q_n)=a_nx^n+\ldots+a_mx^m+\ldots+a_1x+a_0. \qquad (24)$$

Observaţia 9. Schema lui Poisson se poate realiza printr-un şir de n urne, U_1, U_2, ..., U_n care conţin bile de două culori (a_k bile albe şi b_k bile negre în urna U_k, $k=1,2,\ldots,n$). Se extrage pe rând câte o bilă din fiecare urnă. În urma efectuării extragerii unei bile din urna U_k evenimentul *A* constă în apariţia bilei albe cu probabilitatea p_k iar evenimentul $\overline{A}$ constă în extragerea unei bile negre cu probabilitatea $q_k=1-p_k$.

Observaţia 10. Schema lui Bernoulli este un caz particular al

schemei lui Poisson, când $p_1=p_2=\ldots=p_n$.

Exemplul 20. În trei loturi de produse 3%, 4% şi respectiv 5% sunt defecte. La un control de calitate se extrage la întâmplare câte un produs din fiecare lot. Să se afle probabilitatea ca două dintre produsele alese să fie defecte.

Soluţie. Suntem evident în cadrul schemei lui Poisson, cu datele:

$n=3, m=2, p_1=0,03, q_1=0,97, p_2=0,04, q_2=0,96, p_3=0,05, q_3=0,95$. Probabilitatea căutată este coeficientul lui x^2 din polinomul $(0,03x+0,97)(0,04x+0,96)(0,05x+0,95)$.

4o. Schema bilei neîntoarse.

Propoziţia 17. O urnă conţine a bile albe şi b bile negre. Din această urnă se extrag n bile fără a pune bila extrasă înapoi în urnă. Atunci, probabilitatea ca din cele n bile extrase α să fie albe şi $\beta=n-\alpha$ să fie negre este:

$$P(a,b;\alpha,\beta)=\frac{C_a^{\alpha}\cdot C_b^{\beta}}{C_{a+b}^{\alpha+\beta}} \quad , \tag{25}$$

Într-adevăr, numărul tuturor grupelor de n bile luate din cele $a+b$ bile este C_{a+b}^{n}, unde $n=\alpha+\beta$. Pentru a determina numărul cazurilor favorabile asociem fiecare grupă de α bile albe (în total C_a^{α}) cu fiecare grupă ce conţine β bile negre (în total C_b^{β} grupe). Deci numărul total al cazurilor favorabile este $C_a^{\alpha}\cdot C_b^{\beta}$. Folosind definiţia clasică a probabilităţii se obţine relaţia : $P(a,b;\alpha,\beta)=\frac{C_a^{\alpha}\cdot C_b^{\beta}}{C_{a+b}^{\alpha+\beta}}$.

Observaţia 11. Probleme care se modelează şi se rezolvă cu schema bilei întoarse apar în controlul de calitate la recepţia produselor care ies de de panda de fabricaţie. Astfel dacă avem un lot de N produse printre care se găsesc D produse defecte, se extrag la întâmplare n produse şi se cere probabilitatea ca printre cele n produse să se găsească d produse defecte. Notăm cu $P(N-D,D;n-d,d)$ probabilitatea cerută, atunci $a=N-D,\alpha=n-d,b=N,\beta=d$ şi deci conform propoziţiei precedente avem:

$$P(N-D,D;n-d,d)=\frac{C_{N-D}^{n-d}\cdot C_D^{d}}{C_N^{n}} \tag{26}$$

Exemplul 21. De pe banda de montaj a unei fabrici de televizoare au ieşit un număr de 30 televizoare, dintre care 6 cu defecţiuni. Care este probabilitatea ca la un control de calitate luând prin sondaj 10 televizoare, să se găsească 2 televizoare defecte ?

Soluţie : Conform formulei (26) avem: $P(30,6;24,4)=\frac{C_{24}^{8}\cdot C_6^{2}}{C_{30}^{10}}$.

Schema bilei neîntoarse poate fi generalizată la o urnă care conţine bile de mai multe culori, astfel:

Propoziţia 18. O urnă conţine bile de s culori : a_i bile de culoarea c_i , (i=1,2,,…,s). Atunci, probabilitatea de a obţine n_1 bile de culoarea c_1, n_2 bile de culoarea c_2, etc., când facem $n=n_1+n_2+\ldots+n_m$ extrageri, fără a pune bilele extrase înapoi în urnă, este egală cu

$$P(a_1,a_2,\ldots,a_s;n_1,n_2,\ldots,n_s)=\frac{C_{a_1}^{n_1}\cdot C_{a_2}^{n_2}\cdot\ldots\cdot C_{a_s}^{n_s}}{C_{a_1+a_2+\ldots+a_s}^{n_1+n_2+\ldots+n_s}}, \qquad (27)$$

1.3. VARIABILE ALEATOARE

1.3.1. Distribuţia şi funcţia de repartiţie a unei variabile aleatoare discrete.

Noţiunea de variabilă aleatoare este strâns legată de conceptele fundamentale ale teoriei probabilităţilor şi anume cele de eveniment şi probabilitate.

Din punct de vedere intuitiv o variabilă aleatoare este o mărime X ale cărei valori numerice depind de rezultatul unui experiment aleator, mulţimea acestora fiind bine definită este numită *mulţimea valorilor posibile*. Valorile lui X sunt asociate evenimentelor experimentului şi ele pot fi cunoscute numai după efectuarea acestuia.

Exemplul 1. Să considerăm experimentul care constă în aruncarea a două zaruri. Vom nota cu X suma punctelor de pe cele două zaruri obţinute la o aruncare. Este clar că mulţimea valorilor posibile ale lui X este $\{2,3,\ldots,12\}$. O valoare a lui X depinde de rezultatul experimentului şi se cunoaşte numai după efectuarea sa.

O definiţie riguroasă din punct de vedere matematic a noţiunii de variabilei aleatoare este dată în următoarea definiţie.

Definiţia 1. Fie $\{\Omega, \mathcal{K}, P\}$ un spaţiu de probabilitate şi $\mathcal{E}=\{A_1,A_2,\ldots,A_m\}$ un sistem complet de evenimente ale lui Ω. Se numeşte *variabilă aleatoare* o funcţie $X:\mathcal{E}\rightarrow \boldsymbol{R}$, care ia valorea reală x_i odată cu apariţia evenimentului A_i.

Scriem $X(A_i)=x_i$ sau $A_i=(X=x_i)$. Ultima notaţie semnifică faptul că A_i este evenimentul ca variabila aleatoare X să ia valoarea x_i.

Notăm cu $(X=x)$ mulţimea elementelor lui Ω a căror imagine prin aplicaţia X este numărul real x, adică: $(X=x)=\{a\in\Omega : X(\{a\})=x\}$.

Notăm cu $(X<x)$ mulţimea elementelor lui Ω a căror imagine prin aplicaţia X este un număr real strict mai mic decât x adică:

$$(X<x) = \{ a \in \Omega : X(\{a\}) < x \}.$$

Utilizăm de asemenea notația $(X \le x)$ a cărei interpretare este:

$$(X \le x) = (X=x) \cup (X<x).$$

Mulțimile $(X<x)$, $(X=x)$ şi $(X \le x)$ sunt desigur evenimente.

Observaţia 1. Fie Ω spațiul evenimentelor elementare pe care-l considerăm finit: $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$. Putem defini variabila aleatoare X pe Ω. Considerăm sistemul de evenimente elementare $\mathcal{E}=\{E_1, E_2, \ldots, E_n\}$ unde $E_i = \{\omega_i\}$, $i=1,\ldots,n$. Acesta este evident un sistem complet de evenimente. Notăm cu $X(\Omega)$ mulțimea imaginilor prin X a elementelor lui Ω, adică: $X(\Omega)=\{x_1, x_2, \ldots, x_i, x_{i+1}, \ldots, x_n\}$, unde $x_i = X(E_i) = X(\{\omega_i\})$. Este clar că valorile x_i sunt distincte două câte două, căci evenimentele lui $\mathcal{E}$ sunt două câte două incompatibile. Reciproc, dacă variabila aleatoare X ia una şi numai una din valorile x_i, pentru $i=1,2,\ldots, n$, atunci evenimentele $E_i=(X=x_i)$ sunt incompatibile două câte două şi reuniunea lor este evenimentul sigur. Astfel, mulțimea acestor evenimente $E=\{E_1, E_2, \ldots, E_n\}$, constituie un sistem complet de evenimente (ele nu sunt neapărat evenimente elemenare).

Definiţia 2. Fie X o variabilă aleatoare definită pe spațiul fundamental Ω şi $X(\Omega) \subset \boldsymbol{R}$ mulțimea valorilor sale.

Variabilă aleatoare X se numeşte de *tip discret,* dacă mulțimea $X(\Omega)) \subset \boldsymbol{R}$ formează o mulțime cel mult numărabilă de numere reale (finită sau infinită, valorile sale formând un şir $x_1, x_2, \ldots, x_n, \ldots$).

Variabilă aleatoare X se numeşte de *tip continuu* dacă valorile sale formează un interval umplu un interval $X(\Omega)=(a,b) \subset \boldsymbol{R}$.

În acest paragraf ne ocupăm numai de variabilele aleatoare discrete.

Definiţia 3. Fie X o variabilă aleatoare discretă definită pe spațiul fundamental Ω. Se numeşte *legea de probabilitate* a variabilei aleatoare X sau *distribuția variabilei aleatoare X,* o funcție

$$f : X(\Omega) \to [0,1], \quad f(x)=P(X=x), \quad \forall x \in X(\Omega). \tag{1}$$

Observaţia 2. Legea de probabilitate f a lui X poate fi definită pe întreaga axă reală deoarece:

$x \notin X(\Omega) \Rightarrow (X=x)=\varnothing \Rightarrow P(X=x)=P(\varnothing)=0$.

Astfel putem scrie :

$$f : \boldsymbol{R} \to [0,1], \quad f(x) = \begin{cases} P(X=x), x \in X(\Omega) \\ 0, x \notin X(\Omega) \end{cases} \tag{1'}$$

Definiţia 4. Fie Ω spațiul evenimentelor elementare şi X o variabilă aleatoare discretă definită pe Ω, unde

$X(\Omega)=\{x_1, x_2, \ldots, x_i, x_{i+1}, \ldots, x_n\}$, cu $x_i<x_{i+1}$.

Notând cu $f(x_i)=P(X=x_i)=p_i$, $i=1,2,\ldots,n$, distribuția unei variabile aleatoare discrete X mai poate fi notată astfel:

$$X : \begin{pmatrix} x_i \\ f(x_i) \end{pmatrix}, \ (i=1,2,\ldots,n), \quad \text{sau} \quad X : \begin{pmatrix} x_1 \ldots x_i \ldots x_n \\ p_1 \ldots p_i \ldots p_n \end{pmatrix} \tag{2}$$

numit *tabloul de distribuție* sau *repartiția* variabilei aleatoare discrete *X*.

Într-un astfel de tablou sunt enumerate valorile posibile şi

probabilitățile corespunzătoare.

Propoziția 1. Legea de probabilitate a unei variabile aleatoare discrete *X* are următoarele proprietăți:

(1). $f(x_i) \geq 0, i = 1,2,..., n$; (2). $\sum_{i=1}^{n} f(x_i) = 1$.

Demonstrație. Mai întâi este clar că $f(x_i) \geq 0$, căci această valoare este o probabilitate. Apoi, deoarece $\mathcal{E}=\{E_1,E_2,\ldots,E_n\}$ este un sistem complet de evenimente, unde $E_i=(X=x_i)$ pentru $i=1,\ldots,n$ şi $\bigcup_{i=1}^{n} E_i = \Omega$,

avem: $\sum_{i=1}^{n} f(x_i) = \sum_{i=1}^{n} P(X = x_i) = P\left(\bigcup_{i=1}^{n} E_i\right) = P(\Omega) = 1$.

Observația 3. Ținând seama de faptul că $\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} f(x_i) = 1$, un tablou de distribuție al unei variabile aleatoare discrete se va nota:

$X : \begin{pmatrix} x_i \\ f(x_i) \end{pmatrix}$, $(i = 1,2,...,n)$, $\sum_{i=1}^{n} f(x_i) = 1$ sau $X : \begin{pmatrix} x_1 \ldots x_i \ldots x_n \\ p_1 \ldots p_i \ldots p_n \end{pmatrix}$, $\sum_{i=1}^{n} p_i = 1$.

Dacă spațiul fundamental ? este infinit dar numărabil şi *X* este o variabilă aleatoare definită pe Ω atunci mulțimea valorilor sale $X(\Omega)$ formează un şir $x_1, x_2, \ldots, x_n, \ldots$. Sistemul complet de evenimente este infinit numărabil, adică $E=\{E_1,E_2,\ldots,E_n,\ldots\}$ unde $E_n=(X=x_n)$, $n \in \mathbf{N}^*$ şi $\bigcup_{n=1}^{\infty} E_n = \Omega$. Dacă $p_n = P(E_n) = P(X=x_n), \forall\, n \in \mathbf{N}^*$, atunci

$$\sum_{n=1}^{\infty} p_i = \sum_{n=1}^{\infty} P(X = x_n) = P\left(\bigcup_{n=1}^{\infty} E_i\right) = P(\Omega) = 1,$$

iar tabloul de distribuție al variabilei aleatoare *X* se scrie:

$$X : \begin{pmatrix} x_1 & x_2 & \ldots & x_n & \ldots \\ p_1 & p_2 & \ldots & p_n & \ldots \end{pmatrix} \text{cu } \sum_{n=1}^{\infty} p_n = 1. \qquad (2')$$

Distribuția unei variabile aleatoare discrete este de forma

$X : \begin{pmatrix} x_j \\ f(x_j) \end{pmatrix}$, $f(x_j) = P(X = x_j)$, $j \in J$, $\sum_{j \in J} f(x_j) = 1$, J cel mult numărabilă.

Exemplul 2. Mergând pe un traseu un automobilist întâlneşte patru intersecții semaforizate. La fiecare semafor culoarea roşie durează 60 secunde, cea galbenă 5 secunde iar cea verde 25 secunde. Cele 4 semafoare nu sunt sincronizate şi

presupunem că apariția unei culori la un semafor întâlnit nu depinde de culorile întâlnite la semafoarele anterioare.

Să se scrie tabloul de distribuție al variabilei aleatoare X care reprezintă numărul de semafoare roşii întâlnite de automobilist şi să se reprezinte poligonul de repartiție al acesteia. Să se calculeze media şi dispersia lui X.

Soluție: Fie X variabila aleatoare care reprezintă numărul de semafoare roşii întâlnite de automobilist şi fie $A_k = (X = k)$ evenimentul care reprezintă numărul de semafiare roşii întâlnite în drumul său de către automobilist $(k = 0,1,2,3,4)$. Întrucât la fiecare semafor este aceeaşi situație privind numărul de secunde cât durează fiecare dintre cele trei culori, înseamnă că evenimentul A constând din întâlnirea culorii roşii se poate repeta în aceleați condiții, cu

probabilitatea $p = P(R) = \frac{60}{90} = \frac{2}{3}$, de k ori $(k = 0,1,2,3,4)$. Astfel ne aflăm în cadrul schemei lui Bernoulli (schema bilei întoarse) pentru care avem: $P(X = k) = C_4^k \left(\frac{2}{3}\right)^k \left(\frac{1}{3}\right)^{4-k}$, $k = 0,1,2,3,4$. Atunci:

Pentru $k = 0$, $P(X = 0) = C_4^0 \left(\frac{2}{3}\right)^0 \left(\frac{1}{3}\right)^4 = \frac{1}{81}$;

Pentru $k = 1$, $P(X = 1) = C_4^1 \left(\frac{2}{3}\right)^1 \left(\frac{1}{3}\right)^3 = \frac{8}{81}$

Pentru $k = 2$, $P(X = 2) = C_4^2 \left(\frac{2}{3}\right)^2 \left(\frac{1}{3}\right)^2 = \frac{24}{81}$

Pentru $k = 3$, $P(X = 3) = C_4^3 \left(\frac{2}{3}\right)^3 \left(\frac{1}{3}\right)^1 = \frac{32}{81}$

Pentru $k = 4$, $P(X = 4) = C_4^4 \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^0 = \frac{16}{81}$.

Prin urmare tabloul de distribuție al variabilei aleatoare X este:

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ \frac{1}{81} & \frac{8}{81} & \frac{24}{81} & \frac{32}{81} & \frac{16}{81} \end{pmatrix}.$$

Observaţia 4. Distribuția unei variabile aleatoare discrete se poate reprezenta grafic în plan prin *poligonul de repartiție (distribuție),* care se obține unind printr-o linie poligonală punctele de coordonate $M_i(x_i, p_i)$, $i=1,2,\ldots,n$; în general pe cele două axe se iau unități de măsură diferite.

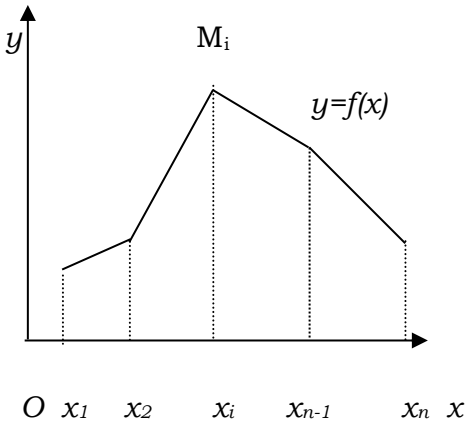


Fig.1. Poligonul de distribuție

Definiţia 5. Fie X o variabilă aleatoare discretă. Se numeşte *funcție de repartiție* a variabilei aleatoare X, funcția

$$F : \boldsymbol{R} \to [0,1], \quad F(x) = P(X < x), \ \forall x \in \boldsymbol{R} \tag{3}$$

Funcția de repartiție a unei variabile aleatoare are următoarele proprietăți:

Propoziţia 2. Fie F funcția de repartiție a unei variabile aleatoare X. Atunci avem:

(1) $0 \le F(x) \le 1, \ \forall \ x \in \boldsymbol{R};$

(2). $F(x_1) \le F(x_2)$, dacă $x_1 < x_2$, (funcția F este nedescrescătoare);

(3). Dacă variabila aleatoare X ia valori în intervalul $[a,b]$, atunci $F(a)=0$ şi $F(b)=1$. În plus, $P(a<X<b) = F(b) - F(a)$.

În particular, dacă argumentul variabilei aleatoare X ia valori pe toată mulțimea $\boldsymbol{R}$ avem :

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0 \text{ şi } F(+\infty) = \lim_{x \to +\infty} F(x) = 1.$$

Demonstrație. (1). Inegalitățile $0 \le F(x) \le 1$, $\forall x \in \mathbf{R}$, sunt adevărate deoarece $F(x)$ este o probabilitate.

(2). Dacă $x_1 < x_2$ atunci putem scrie $F(x_2) - F(x_1) = P(x_1 < X < x_2) \ge 0$.

Prin urmare $F(x_2) \ge F(x_1)$.

(3). Dacă a şi b sunt cea mai mică, respectiv cea mai mare valoare pe care o poate lua argumentul variabilei X, atunci $F(a)=P(X<a)=0$, deoarece evenimentul $(X<a)$ este imposibil şi $F(b)=F(X< b)=1$, deoarece evenimentul $(X< b)$ este sigur. Apoi, pentru $a,b \in \mathbf{R}$, $a<b$, avem

$$(a \le X < b) = (X < b) - (X < a); \quad (X < a) \subset (X < b)$$

şi deci

$$P(a \le X < b) = P(X < b) - P(X < a) = F(b) - F(a).$$

Propoziția 3. Fie X o variabilă aleatoare discretă având distribuția

$$X : \begin{pmatrix} x_1 \dots x_i \dots x_n \\ p_1 \dots p_i \dots p_n \end{pmatrix}, \ \sum_{i=1}^{n} p_i = 1.$$

Atunci funcția sa de repartiție se calculează cu formula:

$$F(x) = \sum_{x_i < x} p_i, \quad \forall x \in \mathbf{R} \qquad (5)$$

Într-adevăr, pentru un punct $x \in \mathbf{R}$, evenimentul $(X < x)$ este reuniunea evenimentelor $(X=x_i)$, până la cel mai mare argument $x_i < x$, adică avem: $(X < x) = \bigcup_{x_i < x} (X = x_i)$. Evenimentele $(X=x_i)$ fiind incompatibile, aplicând operatorul de probabilitate asupra relației precedente, obținem: $F(x) = P(X < x) = \sum_{x_i < x} P(X = x_i) = \sum_{x_i < x} p_i, \quad \forall x \in \mathbf{R}$

adică, *funcția de repartiție $F(x)$ a variabilei aleatoare X este dată de suma probabilităților p_i pentru valorile x_i inferioare lui x.*

Din modul de calcul, funcția de repartiție mai poartă numele şi de *funcție cumulativă* a variabilei aleatoare X.

Exemplul 3. Să se determine funcția de repartiție a variabilei X care are distribuția $X : \begin{pmatrix} 0 & 1 & 2 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}$.

Soluție. Dacă $x \le 0$, $F(x)=P(X< x)=0$;

Dacă $x \le 0$ atunci $F(x)=P(X<x)=P(X=0)=\frac{1}{4}$;

Dacă $1<x \le 2$ atunci avem:

$$F(x)=P(X<x)=P[(X=0) \cup (X=1)] = \frac{1}{4}+\frac{1}{4}=\frac{1}{2}.$$

Dacă $x>2$ atunci $F(x)=P(X<x)=P[(X=0) \cup (X=1) \cup (X=2)]=\frac{1}{4}+\frac{1}{4}+\frac{1}{2}=1$.

Astfel, am obținut: $F(x)=\begin{cases} 0, & x \le 0 \\ 1/4, & 0<x \le 1 \\ 1/2, & 1<x \le 2 \\ 1, & x>2 \end{cases}$.

1.3.2. Operaţii cu variabile aleatoare discrete.

Definiţia 6. Două variabile aleatoare discrete *X* şi *Y* definite pe acelaşi spaţiu de probabilitate *{Ω, K, P}* şi având distribuţiile

$$X:\begin{pmatrix} x_1 & \cdots & x_n \\ p_1 & \cdots & p_n \end{pmatrix} \text{ si } Y:\begin{pmatrix} y_1 & \cdots & y_m \\ q_1 & \cdots & q_m \end{pmatrix},$$

se numesc *P-independente* dacă $\forall i=1,\ldots,n, \forall j=1,\ldots,m$ avem:

$$P\left[(X=x_i)\cap(Y=y_j)\right]=P(X=x_i)\cdot P(Y=y_j)=p_i\cdot q_j \ .$$

Observaţia 5. În toate cazurile când se fac operații cu variabile aleatoare se va presupune că variabilele aleatoare sunt definite pe aceeaşi mulțime de evenimente elementare.

O constantă *a* poate fi interpretată ca o variabilă aleatoare şi anume, ca variabila aleatoare care ia valoarea *a* pentru orice eveniment elementar, iar tabloul său de distribuție este $a:\begin{pmatrix} a \\ 1 \end{pmatrix}$.

În cele ce urmează vom presupune că variabilele aleatoare *X* şi *Y* sunt independente şi au tablourile de distribuție din *Definiția 7:*

Definiţia 7. Se numeşte *suma variabilelor aleatoare X* şi *Y,* variabila aleatoare notată *X+Y,* care ia valoarea x_i+y_j dacă *X* ia valoarea x_i iar *Y* ia valoarea y_j. Această variabilă are distribuția:

$$X+Y:\begin{pmatrix} x_1+y_1 & x_1+y_2 & \cdots & x_i+y_j & \cdots & x_n+y_m \\ r_{11} & r_{12} & \cdots & r_{ij} & \cdots & r_{nm} \end{pmatrix}, \tag{6}$$

unde $r_{ij}, i=1,...,n, j=1,..,m$ sunt probabilităţile:

$$r_{ij}=P(X+Y=x_i+y_j)=P[(X=x_i)\cap(Y=y_j)]=P(X=x_i)\cdot P(Y=y_j)=p_i\cdot q_j .$$

Suma dintre o constantă a şi *variabila aleatoare* X este variabila care ia valoarea $a=x_i$ când X ia valoarea x_i şi deci distribuţia ei este

$$a+X:\begin{pmatrix} a+x_1 & \cdots & a+x_n \\ p_1 & \cdots & p_n \end{pmatrix}. \tag{7}$$

Definiţia 8. Se numeşte *produsul variabilelor aleatoare* X şi Y, variabila aleatoare notată $Z=X\cdot Y$, care ia valoarea $x_i\cdot y_j$ dacă X ia valoarea x_i şi Y ia valoarea y_j. Această variabilă are distribuţia:

$$X\cdot Y:\begin{pmatrix} x_1\cdot y_1 & x_1\cdot y_2 & \cdots & x_i\cdot y_j & \cdots & x_n\cdot y_m \\ r_{11} & r_{12} & \cdots & r_{ij} & \cdots & r_{nm} \end{pmatrix}, \tag{8}$$

unde $r_{ij}, i=1,...,n, j=1,..,m$ sunt probabilităţile

$$r_{ij}=P(X\cdot Y=x_i\cdot y_j)=P[(X=x_i)\cap(Y=y_j)]=P(X=x_i)\cdot P(Y=y_j)=p_iq_j$$

Este clar ca pentru ambele operaţii de adumare şi înmulţire a variabilelor aleatoare avem : $\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij}=\sum_{i=1}^{n} p_i \sum_{j=1}^{m} q_j=1$.

Produsul dintre o constantă a şi *variabila aleatoare* X este variabila, notată aX, care ia valoarea ax_i când X ia valoarea x_i, pentru fiecare $i=1,\ldots,n$ şi deci distribuţia ei este

$$a\cdot X:\begin{pmatrix} ax_1 & \cdots & ax_n \\ p_1 & \cdots & p_n \end{pmatrix} \tag{9}$$

Operaţiie de sumă şi de produs de extind la orice număr finit de variabile aleatoare.

Definiţia 9. Se numeşte *puterea de ordin* k a variabilei aleatoare X, variabila aleatoare notată X^k care ia valoarea x_i^k dacă X ia valoarea x_i . Această variabilă are tabloul de distribuţie:

$$X^k : \begin{pmatrix} x_1^k & \cdots & x_n^k \\ p_1 & \cdots & p_n \end{pmatrix}. \qquad (10)$$

Observaţia 6. În tabloul de distribuţie al al unei puteri, de exemplu al puterii X^2 a unei variabile aleatoare, nu apar decât puteri de forma x_i^2 ale valorilor variabilei nu şi produse de forma $x_i \cdot x_j$ când $x_i \neq x_j$, deoarece probabilitatea ca v.a. X^2 să ia această valoare este nulă.

Într-adevăr, dacă $x_i \neq x_j$ atunci evenimentele $(X = x_i)$ si $(X = x_j)$ sunt disjuncte şi prin urmare avem :

$$P\left(X^2 = x_i \cdot x_j\right) = P\left[(X = x_i) \cap (X = x_j)\right] = P(\varnothing) = 0 .$$

Exemplul 4. Două variabile aleatoare independente au distribuţiile

$$X : \begin{pmatrix} 2 & 3 & 5 \\ 0{,}2 & 0{,}3 & 0{,}5 \end{pmatrix} \quad \text{şi} \quad Y : \begin{pmatrix} 1 & 4 & 6 \\ 0{,}6 & 0{,}2 & 0{,}2 \end{pmatrix}.$$

Să se scrie distribuţiile variabilelor $X+Y$ şi $X \cdot Y$, X^2, Y^2.

Soluţie.: Valorile pe care le ia variabila $X+Y$ sunt

2+1=3, 2+4=6, 2+6=8, 3+1=4, 3+4=7, 3+6=9, 5+1=6, 5+4=9, 5+6=11, care ordonate crescător sunt: 3,4,6,7,8,9,11.

Vom calcula pe rând probabilităţile de obţinere a fiecărei valori. Avem

$$P(X+Y=3) = P\left[(X=2) \cap (Y=1)\right] = P(X=2) \cdot P(y=1) = 0{,}2 \cdot 0{,}6 = 0{,}12;$$

$P(X+Y=4) = 0{,}18$ $P(X+Y=6) = P\left[((X=2) \cap (Y=4) \cup ((X=5) \cap (Y=1))\right] =$

$= P(X=2)P(Y=4) + P(X=5)P(Y=1) = 0{,}2 \cdot 0{,}2 + 0{,}5 \cdot 0{,}6 = 0{,}34$, $P(X+Y=7) = 0{,}06;$

$P(X+Y=8) = 0{,}34$ $P(X+Y=9) = 0{,}16;$ $P(X+Y=11) = 0{,}10$

Astfel, tabloul de distribuţie al variabilei $X+Y$ este

$$X+Y : \begin{pmatrix} 3 & 4 & 6 & 7 & 8 & 9 & 11 \\ 0{,}12 & 0{,}18 & 0{,}34 & 0{,}06 & 0{,}04 & 0{,}16 & 0{,}10 \end{pmatrix}$$

Valorile pe care le ia variabila aleatoare $X \cdot Y$ sunt:

$2\cdot1=2$, $2\cdot4=8$, $2\cdot6=12$, $3\cdot1=3$, $3\cdot4=12$, $3\cdot6=16$, $5\cdot1=5$, $5\cdot4=20$, $5\cdot6=30$, care ordonate crescător sunt: 2, 8, 12, 3, 12, 18, 5, 20, 30.

Probabilităţile corespunzătoare se calculează ca mai sus obţinem:

$$X\cdot Y:\begin{pmatrix} 2 & 3 & 5 & 8 & 12 & 18 & 20 & 30 \\ 0{,}12 & 0{,}18 & 0{,}30 & 0{,}04 & 0{,}10 & 0{,}06 & 0{,}10 & 0{,}10 \end{pmatrix}.$$

Pătratele variabilelor aleatoare X şi Y sunt conform definiţiei:

$$X^2:\begin{pmatrix} 4 & 9 & 25 \\ 0{,}2 & 0{,}3 & 0{,}5 \end{pmatrix},\ Y^2:\begin{pmatrix} 1 & 16 & 36 \\ 0{,}6 & 0{,}2 & 0{,}2 \end{pmatrix}.$$

1.3.3. Caracteristici numerice ale unei variabile aleatoare discrete.

1º. Media unei variabile aleatoare discrete.

Definiţia 10. Dacă X este o variabilă aleatoare discretă având distribuţia $X:\begin{pmatrix} x_1 & \cdots & x_n \\ p_1 & \cdots & p_n \end{pmatrix}$ cu $\sum_{i=1}^{n} p_i = 1$

Se numeşte *valoare medie* sau *media* variabilei aleatoare discrete X, expresia

$$m=M(X)=\sum_{i=1}^{n} p_i x_i\ , \qquad (11)$$

Observaţia 7. Dacă în cazul variabilei aleatoare discrete aceasta are un număr infinit dar numărabil de valori ($n\to\infty$) atunci media are sens dacă seria numerică $\sum_{i=1}^{\infty} p_i x_i$ este convergentă.

Observaţia 8. Media variabilei unei aleatoare X este o medie ponderată cu ponderile $p_1,\ldots,p_n$ şi are următoarea interpretare: ea este valoarea în jurul căreia se grupează valorile variabilei X.

În literatura de specialitate media unei variabilei aleatoare se mai numeşte *speranţa* variabilei aleatoare respective.

Exemplul 5. Să se calculeze media variabilei aleatoare care are distribuţia: $X:\begin{pmatrix} 1 & 2 & 3 \\ 0{,}2 & 0{,}4 & 0{,}4 \end{pmatrix}$

Soluție. a). Conform definiției, avem:

$$m=M(X)=\sum_{i=1}^{n} p_i x_i = 1\cdot 0{,}2+2\cdot 0{,}4+3\cdot 0{,}4=2{,}2\,.$$

Propoziţia 4. (*Proprietăți ale mediei*).

Fie X, Y variabile aleatoare (discreta sau continue) şi a o variabilă aleatoare constantă. Atunci avem:

1. $M(a)=a$;
2. $M(a+X)=a+M(X)$;
3. $M(a\cdot X)=a\cdot M(X)$;
4. $M(X+Y)=M(X)+M(Y)$;
5. $M(X\cdot Y)=M(X)\cdot M(Y)$, dacă X şi Y sunt independente.
6. Dacă $\lambda \le X \le \mu$ atunci $\lambda \le M(X) \le \mu$.

Demonstraţie. Fie a variabila constantă şi X,Y variabile aleatoare discrete având distribuțiile

$$a:\begin{pmatrix} a \\ 1 \end{pmatrix}, X:\begin{pmatrix} x_1 & \cdots & x_n \\ p_1 & \cdots & p_n \end{pmatrix}, Y:\begin{pmatrix} y_1 & \cdots & y_m \\ q_1 & \cdots & q_m \end{pmatrix}, \sum_{i=1}^{n} p_i = \sum_{j=1}^{m} q_i = 1$$

Conform definițiilor operațiilor cu variabile aleatoare avem:

$$a+X:\begin{pmatrix} a+x_1 & \cdots & a+x_n \\ p_1 & \cdots & p_n \end{pmatrix},\ aX:\begin{pmatrix} ax_1 & \cdots & ax_n \\ p_1 & \cdots & p_n \end{pmatrix},$$

$$X+Y:\begin{pmatrix} x_1+y_1 & \cdots & x_i+y_j & \cdots & x_n+y_m \\ r_{11} & \cdots & r_{ij} & \cdots & r_{nm} \end{pmatrix},$$

$$X\cdot Y:\begin{pmatrix} x_1\cdot y_1 & \cdots & x_i\cdot y_j & \cdots & x_n\cdot y_m \\ r_{11} & \cdots & r_{ij} & \cdots & r_{nm} \end{pmatrix},$$

Ținând seama de aceste definiții şi de definiția mediei avem:

1. $M(a)=a\cdot 1=a$;

2. $M(a+X)=\sum_{i=1}^{n}(a+x_i)p_i = a\sum_{i=1}^{n} p_i + \sum_{i=1}^{n} x_i p_i = a+M(X)$;

3. $M(aX)=\sum_{i=1}^{n}(a\cdot x_i)p_i = a\cdot\sum_{i=1}^{n} x_i p_i = a\cdot M(X)$.

4. $M(X+Y)=\sum_{i=1}^{n}\sum_{j=1}^{m}(x_i+y_j)r_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i p_i q_j + \sum_{i=1}^{n}\sum_{j=1}^{m} y_j p_i q_j =$

$$= \sum_{i=1}^{n} x_i p_i \left(\sum_{j=1}^{m} q_j \right) + \sum_{j=1}^{m} y_j q_j \left(\sum_{i=1}^{n} p_i \right) = \sum_{i=1}^{n} x_i p_i + \sum_{j=1}^{m} y_j q_j = M(X) + M(Y)$$

$$5.\ M(X \cdot Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_i y_j r_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{m} x_i y_j p_i q_j = \sum_{i=1}^{n} x_i y_j \sum_{j=1}^{m} p_i q_j = M(X) \cdot M(Y)$$

6. Dacă $\lambda \le x_i \le \mu$, pentru $i = 1,\ldots,n$, atunci înmulțind aceste inegalități cu p_i, însumând şi ținând seama de $\sum_{i=1}^{n} p_i = 1$, obținem succesiv: $\lambda \cdot p_i \le x_i \cdot p_i \le \mu \cdot p_i$, pentru $i = 1,\ldots,n$,

$$\sum_{i=1}^{n} \lambda \cdot p_i \le \sum_{i=1}^{n} x_i \cdot p_i \le \sum_{i=1}^{n} \mu \cdot p_i \Rightarrow \lambda \cdot \sum_{i=1}^{n} p_i \le \sum_{i=1}^{n} x_i \cdot p_i \le \mu \cdot \sum_{i=1}^{n} p_i \Rightarrow$$

$$\Rightarrow \lambda \le M(X) \le \mu \ .$$

2⁰. <u>Dispersia unei variabile aleatoare discrete</u>.

<u>Definiţia 11.</u> Fie *X* o variabilă aleatoare discretă sau continuă având media *m=M(X).* Variabila aleatoare *X-m* se numeşte *abaterea lui* X *de la valoarea medie m.*

<u>Observaţia 9.</u> Din proprietățile mediei, deducem :

$M(X - m) = M(X) - M(m) = m - m = 0$.

Deci pentru orice variabilă aleatoare media abaterilor sale individuale de la valoarea medie este totdeauna egală cu zero. Aceasta înseamnă că nu putem utiliza media acesteia pentru a determine împrăştierea acestor valori faţi de medie.

Din acest motiv ca o măsură a împrăştierii valorilor unei variabile

aleatoare *X* față de media sa vom considera o altă valoare

carecteristică, şi anume *dispersia***.**

<u>Definiţia 12.</u> Se numeşte *dispersia* sau încă *varianța* variabilei aleatoare *X* numărul notat *D(X)=σ²*, reprezentat de media pătratului abaterii lui *X* de valoarea medie *m*, adică

$$D(X)=\sigma^2 = M_2(X-m)=M[(X-m)^2] \qquad (12)$$

Dacă X este o variabilă discretă de distribuție $X:\begin{pmatrix} x_1 & \cdots & x_n \\ p_1 & \cdots & p_n \end{pmatrix}$, atunci *dispersia* lui X este dată de :

$$D(X)=\sum_{i=1}^{n}(x_i-m)^2 p_i \qquad (12')$$

Propoziţia 5. Pentru o variabilă aleatoare discretă X dispersia se poate calcula cu formula:

$$D(X)=M(X^2)-[M(X)]^2 \qquad (13)$$

Demonstraţie. Folosind definiția şi proprietățile mediei, avem :

$D(X)=M[(X-m)^2]=M(X^2-2mX+m^2)=M(X^2)-2mM(X)+M(m^2)=$

$=M\left(X^2\right)-2m^2+m^2=M(X^2)-\left(M(X)\right)^2$.

Exemplul 6. Pentru variabila aleatoare de la *Exemplul 1*, să se calculeze dispersia.

Soluţie. Tabloul de distribuție al variabilei aleatoare X este:

$$X:\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ \frac{1}{81} & \frac{8}{81} & \frac{24}{81} & \frac{32}{81} & \frac{16}{81} \end{pmatrix}.$$

Pentru acesta avem:

$$M(X)=0\cdot\frac{1}{81}+1\cdot\frac{8}{81}+2\cdot\frac{24}{81}+3\cdot\frac{32}{81}+4\cdot\frac{16}{81}=\frac{216}{81}=\frac{8}{3}\approx 2{,}67$$

$$M(X^2)=0\cdot\frac{1}{81}+1\cdot\frac{8}{81}+4\cdot\frac{24}{81}+9\cdot\frac{32}{81}+16\cdot\frac{16}{81}=\frac{648}{81}=8.$$

$$D(X)=M(X^2)-[M(X)]^2=8-\frac{64}{9}=\frac{8}{9}\approx 0{,}89\,.$$

Propoziţia 6. *(Proprietățile dispersiei).*

Fie X, Y variabile aleatoare (discreta sau continue) şi a o variabilă aleatoare constantă. Atunci avem:

1. $D(a)=0$;
2. $D(a\cdot X)=a^2D(X)$;

3. $D(X\pm Y)= D(X)+D(Y)$, dacă variabilele *X şi Y* sunt independente.
4. $D(a+X)=D(X)$;

Demonstraţie. 1. $D(a)=M[(a-M(a))^2]=M[(a-a)^2]= M(0)=0$.

2. $D(aX)=M\{ [aX-M(aX)]^2\}=M\{[a(X-M(X)]^2\}=a^2M[X-M(X)]=a^2D(X)$.

3. $D(X\pm Y)=M[(X\pm Y)^2]-[M(X\pm Y)]^2=M(X^2\pm 2XY+Y^2)-[M(X) \pm M(Y)]^2=$

$=M(X^2) \pm 2M(X)M(Y)+M(Y^2)-[M(X)]^2\pm 2M(X)M(Y)-[M(Y)]^2=D(X)+D(Y)$.

4. $D(a+X)=D(a)+D(X)=D(X)$.

Exemplul 7. Se consideră variabilele aleatoare independente

$$X:\begin{pmatrix} 1 & 2 & 4 \\ 0,7 & 0,1 & 0,2 \end{pmatrix} \text{ şi } Y:\begin{pmatrix} 1 & 4 & 6 & 7 \\ 0,2 & 0,4 & 0,1 & 0,3 \end{pmatrix}$$

Să se calculeze: *(a). $M(2X+4Y)$; (b) $D(2X+4Y)$.*

Soluţie: ***a).*** Folosind proprietăţil mediei putem scrie:

$$M(2X+4Y)=M(2X)+M(4Y)=2M(X)+4M(Y)$$

$M(X)=1\cdot 0,7+2\cdot 0,1+4\cdot 0,2=1,7$, $M(Y)=1\cdot 0,2+4\cdot 0,4+6\cdot 0,1+7\cdot 0,3=4,5$.

Prin urmare $M(2X+4Y)=2\cdot 1,7+4\cdot 4,5=12,4$

b). Folosind proprietăţile mediei putem scrie:

$$D(2X+4Y)=D(2X)+D(4Y)=4D(X)+16D(Y)$$

$M(X^2)=1\cdot 0,7+4\cdot 0,1+16\cdot 0,2=4,3$,

$D(X)=M(X^2)-[M(X)]^2=4,3-(1,7)^2=1,41$.

$M(Y^2)=1\cdot 0,2+16\cdot 0,4+36\cdot 0,1+49\cdot 0,3=24,9$

$D(Y)=M(Y^2)-[M(Y)]^2=24,9-(4,5)^2=4,65$.

Deci $D(2X+4Y)=4\cdot 1,41+16\cdot 4,65=5,64+84,40=90,04$.

Propoziţia 7. Fie $X_1, X_2,\ldots,X_n$, n variabile aleatoare astfel

încât $M(X_i)=m$ şi $D(X_i)=\sigma^2$, $\forall\ i=1,2,\ldots,n$. Dacă $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$

este media aritmedică a variabilelor aleatoare $X_1,X_2,\ldots,X_n$, atunci avem: $M(\overline{X})=m$; $D(\overline{X})=\frac{\sigma^2}{n}$.

Demonstraţie: Folosind proprietăţile mediei şi dispersiei avem:

1). $M(\overline{X})=M\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}M(X_i)=\frac{1}{n}\sum_{i=1}^{n}m=\frac{1}{n}\cdot mn=m$.

2). $D(\overline{X})=D\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\sum_{i=1}^{n}D(X_i)=\frac{1}{n^2}\sum_{i=1}^{n}\sigma^2=\frac{1}{n^2}\cdot n\sigma^2=\frac{\sigma^2}{n}$.

Observaţia 10. Dispersia unei variabile aleatoare *X* măsoară gradul de împrăştiere a valorilor variabilei aleatoare faţă de valoarea medie. Dacă dispersia este mică valorile variabilei aleatoare *X* sunt grupate într-un interval mic în jurul valorii medii.

În aplicaţii este mai comod să se folosească ca măsură a împrăştierii valorilor variabilei aleatoare *X* în jurul valorii medii, numărul definit mai jos.

Definiţia 13. Fie X o variabilă aleatoare de dispersie *D(X).* Numărul $\sigma=\sqrt{D(X)}$ se numeşte *abatere medie pătratică* sau *abatere standard,* ea fiind media centrată de ordinul doi şi se calculează cu

formula :

$$\sigma=\sqrt{\sum_{i=1}^{n}(x_i-m)^2 p_i} \qquad (14)$$

Observaţia 11. Abaterea standard a unei variabile aleatoare are ca dimensiune, dimensiunea variabilei respective şi caracterizează cel mai bine împrăştierea variabilei.

Definiţia 14. Fie variabila aleatoare *X* cu *M*(*X*)=*m* şi $D(X)=\sigma^2\neq 0$.

Variabila aleatoare $Z=\frac{X-m}{\sigma}$ se numeşte *variabila normată* a variabilei

aleatoare *X* (sau *redusa* variabilei aleatoare *X*).

Propoziţia 8. *(Media şi disperesia variabilei normate).*

(a). $M(Z)=0$. *(b).* $D(Z)=1$.

Demonstrație. $(a).\ M(Z)=M\left(\dfrac{X-m}{\sigma}\right)=M\left(\dfrac{X}{\sigma}-\dfrac{m}{\sigma}\right)=\dfrac{1}{\sigma}M(X)-\dfrac{m}{\sigma}=0.$

$(b).\ D(Z)=D\left(\dfrac{X-m}{\sigma}\right)=\dfrac{1}{\sigma^2}D(X-m)=\dfrac{1}{\sigma^2}\left[D(X)-D(m)\right]=\dfrac{1}{\sigma^2}\cdot\sigma^2=1.$

3º. Momentele unei variabile aleatoare discrete.

Definiţia 15. Se numeşte *moment de ordin k* , (unde $k\in\mathbf{N}^*$) al variabilei aleatoare *X*, media variabilei aleatoare X^k , adică numărul notat:

$$m_k = M_k(X)=M(X^k)\,. \tag{15}$$

Dacă variabila aleatoare *X* este o variabilă discretă şi are distribuția $X:\begin{pmatrix}x_1 & \cdots & x_n\\ p_1 & \cdots & p_n\end{pmatrix}$, atunci *momentul de ordin k* al lui *X* este:

$$m_k=M_k(X)=\sum_{i=1}^{n}p_i\,x_i^k \tag{15'}$$

Definiţia 16. Se numeşte *valoare medie de ordin k* (unde $k\in\mathbf{N}^*$) a variabilei aleatoare *X*, rădăcina de ordinul *k* a momentului de ordinul *k*, adică numărul notat:

$$\mu_k=(m_k)^{\frac{1}{k}}=\left(M(X^k)\right)^{\frac{1}{k}}. \tag{16}$$

Dacă variabila aleatoare *X* este o variabilă discretă de distribuție

$X:\begin{pmatrix}x_1 & \cdots & x_n\\ p_1 & \cdots & p_n\end{pmatrix}$, atunci *valoarea medie de ordin k* al lui *X* este

$$\mu_k=\left(\sum_{i=1}^{n}p_i^k\,x_i\right)^{\frac{1}{k}} \tag{16'}$$

Cazuri particulare:

1º. Pentru *k*=1, momentul de ordinul *1* şi *valoarea medie de ordinul 1* ale variabilei aleatoare *X* coincid cu media sa: $m_1=\mu_1=m$.

2º. Pentru *m=2, valoare medie de ordinul 2* al variabilei aleatoare *X*

se mai numeşte *valoarea medie pătratică:* $\mu_2 = \sqrt{\sum_{i=1}^{n} p_i x_i^2}$

Definiţia 17. Fie *X* o variabilă aleatare de medie *m.* Se numeşte *moment centrat de ordin k* al variabilei aleatoare *X*, momentul de ordinul *k* al variabilei aleatoare *X-m* (numită şi abaterea lui *X* de la valoarea medie *m=M(X)* şi se notează cu $M_k(X-m)$.

Dacă variabila aleatoare *X* are distribuția $X:\begin{pmatrix} x_1 & \cdots & x_n \\ p_1 & \cdots & p_n \end{pmatrix}$, atunci *momentul centrat de ordin k* al lui *X* este

$$M_k(X-m) = \sum_{i=1}^{n} (x_i - m)^k p_i \,. \tag{17}$$

3.3.4. Legi de repartiţie discrete.

Am văzut că dacă X este o variabilă aleatoare definită pe spațiul fundamental Ω atunci am definit *legea de probabilitate* a variabilei aleatoare *X* sau *distribuția variabilei aleatoare X* ca fiind funcția

$$f : X(\Omega) \to [0,1], \quad f(x) = P(X = x), \quad \forall x \in X(\Omega) \tag{18}$$

Dacă *X* este o variabilă aleatoare discretă iar $X(\Omega) = \{x_1, x_2, \ldots, x_n\}$ este mulțimea valorilor sale şi dacă notăm cu $p_i = f(x_i) = P(X = x_i), \ \forall i = 1,2,\ldots,n$

atunci distribuția variabilei aleatoare discrete *X* mai poate fi notată astfel:

$$X:\begin{pmatrix} x_i \\ f(x_i) \end{pmatrix}, \ (i = 1,2,\ldots,n), \quad \text{sau } X:\begin{pmatrix} x_1 \ldots x_i \ldots x_n \\ p_1 \ldots p_i \ldots p_n \end{pmatrix} \tag{19}$$

numit *tabloul de distribuție* sau *repartiția* variabilei aleatoare discrete *X*.

Într-un astfel de tablou sunt enumerate valorile posibile şi probabilitățile corespunzătoare.

Observaţia 12. Legea de probabilitate a unei variabile aleatoare discrete X are următoarele proprietăți:

(1). $f(x_i) \geq 0, i = 1,2,..., n$; (2). $\sum_{i=1}^{n} f(x_i) = 1$.

De asemenea am definit *funcția de repartiție* a variabilei aleatoare

X, ca fiind funcția

$$F : \boldsymbol{R} \to [0,1], \quad F(x) = P(X < x), \ \forall x \in \boldsymbol{R} \tag{20}$$

Funcția de repartiție se calculează cu formula:

$$F(x) = \sum_{x_i < x} p_i , \quad \forall x \in \boldsymbol{R} \tag{21}$$

În această secțiune evidențiem principalele repartiții de tip discret ce intervin în aplicații concrete sau în abordarea diverselor aspecte teoretice din teoria probabilităților şi statistica matematică şi anume repartiția binomială şi repartiția Poisson.

1o*. <u>Repartiţia binomială</u> *(corespunzătoare schemei lui Bernoulli).

Reamintim că o experiență aleatoare este de tip binomial dacă la fiecare realizare a sa conduce la două evenimente complementare.

<u>Definiţia 18.</u> Spunem că o variabilă aleatoare *X* are <u>*repartiția binomială*</u> de parametri *n* şi *p*, $(0 < p < 1)$, dacă legea sa de probabilitate este: $f(k) = P(X = k) = C_n^k p^k q^{n-k}$, $k \in \{0,1,\ldots,n\}$, $p, q \in (0,1)$, $p + q = 1$ şi are distribuția*:*

$$X : \begin{pmatrix} 0 & 1 & 2 & \ldots\ldots & k & \ldots\ldots & n \\ q^n & C_n^1 pq^{n-1} & C_n^2 pq^{n-2} & \ldots\ldots & C_n^k p^k q^{n-k} & \ldots\ldots & p^n \end{pmatrix} \tag{22}$$

Mai spunem în acest caz că *X* este repartizată *B(n,p).*

Această variabilă aleatoare este ataşată schemei lui Bernoulli şi ea reprezintă numărul de apariții ale unui eveniment *A* într-un experiență care se efectuează de *n* ori şi în urma căruia se produce evenimentul *A* cu probabilitatea *p* sau contrarul său $\overline{A}$ cu probabilitatea *q*=1-*p.*

Pentru *p=q* repartiția este simetrică. Cu cât diferența între *p* şi *q* este mai mare cu atât asimetria se accentuează.

În numeroase cercetări biologice, agricole, silvice ş. a. Aplicarea legii binomiale prezintă importanță în domeniul variației alternative (prezența sau absența unei însuşiri).

<u>Observaţia 13.</u> Pentru aplicații se foloseşte <u>formula de recurență</u>:

$$f(k+1) = f(k) \cdot \frac{n-k}{k+1} \cdot \frac{p}{q} \tag{23}$$

Cu formula de bază se calculează *f(0)* pentru *k=0,* iar cu formula de recurență se calculează probabilitatea pentru celelalte valori întregi ale lui *k.*

Observaţia 14. Funcţia f este o lege de probabilitate întrucât satisface proprietăţile:

(1). $f(k) \geq 0,\ \forall k \in \{0,1,\ldots,n\}$;

(2). $\sum_{k=0}^{n} f(k) = \sum_{k=0}^{n} C_n^k p^k q^{n-k} = (p+q)^n = 1$.

Din definiţia funcţiei de repartiţie deducem imediat:

Propoziţia 9. Funcţia de repartiţie a variabilei aleatoare care urmează legea binomială este:

$$F(x) = \begin{cases} 0 & , x \leq 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum_{j=0}^{k} C_n^j p^j q^{n-j} & , k < x \leq k+1,\ k = 0,1,\ldots,n-1 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ 1 & , x > 1 \end{cases}$$

Aceasta este o funcţie în trepte (în scară) cu salturi în punctele *0,1,2,…,n.*

Valorile calacteristice ale unei astfel de variabile se calculează folosind rezultate din Algebră privind calculul de sume şi combinatorică.

Propoziţia 10. Fie X o variabilă aleatoare care are distribuţia binomială. Atunci valorile sale caracteristice sunt:

(a). Media: $M(X)=np$.

(b). Momentul de ordinul 2: $m_2=M_2(X)=np(np+q)$;

(c). Dispersia: $D(X)=npq$

Observaţia 15. Să presupunem că sunt date pentru o variabilă aleatoare X, valorile înregistrate, $\{x_1, x_2, \ldots, x_n\}$ şi frecvenţele relative ale acestora $\{f_1, f_2, \ldots, f_n\}$ .Dacă experimentul aleator ce a generat variabila aleatoare X permite aplicarea legii binomiale, se pune problema ajustării frecvenţelor înregistrate (empiric) prin probabilităţile unei legi binomiale corespunzătoare. Pentru identificarea repartiţiei binomiale care ajustează seria frecvenţelor relative empirice, trebuie determinaţi parametrii n şi p. Cum iniţial se cunoaşte volumul eşantionului şi media variabilei X, pe baza formulei mediei repartiţiei binomiale $M(X) = np$, se calculează probabilitatea p după relaţia

$$p = \frac{M(X)}{n}.$$

2º. <u>Repartiţia Poisson</u> (*legea evenimentelor rare).*

<u>Definiţia 19</u>. Spunem că o variabilă aleatoare X are *repartiția Poisson* de parametru $\lambda > 0$ dacă funcţia sa de probabilitate este de forma

$$f(k) = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k \in \mathbf{N}, \ \lambda > 0 \tag{24}$$

(unde e =2,7182 este numărul lui Neper-baza logaritmului natural)

şi are distribuția:

$$X : \begin{pmatrix} 0 & 1 & 2 & \dots & n & \dots \\ \frac{e^{-\lambda}}{0!} & \frac{\lambda e^{-\lambda}}{1!} & \frac{\lambda^2 e^{-\lambda}}{2!} & \dots & \frac{\lambda^n e^{-\lambda}}{n!} & \dots \end{pmatrix} \tag{25}$$

<u>Observaţia 16</u>. Această repartiție este o repartiție asemănătoare cu cea binomială, fiind un caz particular al acesteia. Ea se întâlneşte atunci când probabilitatea p a evenimentului este foarte mică (de unde şi denumirea de *„<u>legea evenimentelor rare</u>”*) şi când numărul n al experienţelor este foarte mare.

Cu alte cuvinte repartiția Poisson este un caz limită al repartiției binomiale pentru n este un numar natural şi $p>0$, unde produsul $np > 0$ să rămână constant.

Repartiția este de tip discret, asimetrică, ea apropiindu-se de cea binomială pe măsură ce parametrul creşte.

Pentru $k=0$, $f(k) = e^{-\lambda}$ iar pentru $k=1,2,3,...$ se foloseşte relația de recurență:

$$f(k+1) = f(k) \cdot \frac{\lambda}{k+1} \tag{26}$$

<u>Observaţia 17.</u> Funcția f este o lege de probabilitate întrucât satisface proprietățile:

(1). $f(k) \geq 0, \ \forall k \in \{0,1,\dots,n\}$;

(2). $\sum_{k=0}^{\infty} f(k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \cdot e^{-\lambda} = e^{-\lambda} \cdot e^{\lambda} = 1$, unde am ținut seama de dezvoltarea în serie de puteri a lui e^{λ} : $e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$.

Din definiția funcției de repartiție deducem imediat:

Propoziţia 11. Fie X o variabilă aleatoare care urmează o repartiție Poisson de parametru $\lambda > 0$. Atunci funcția sa de repartiție este

$$F(x)=\begin{cases}0 & ,x\le 0\\ e^{-\lambda} & ,0<x\le 1\\ \dots\dots\dots\dots\dots & \\ \sum\limits_{j=0}^{k} e^{-\lambda}\dfrac{\lambda^{j}}{j!} & ,k<x\le k+1,\\ \dots\dots\dots\dots\dots & \end{cases} \tag{27}$$

Valorile calacteristice ale unei astfel de variabile se calculează folosind rezultate din Analiza Matematică privind seriile de puteri.

Propoziţia 12. Fie X o variabilă aleatoare care urmează o repartiție Poisson de parametru $\lambda > 0$. Atunci valorile sale caracteristice sunt:

(a). Media: $m=M(X)$

(b). Momentul de ordinul 2: $m_2=M_2(X)$

(c). Dispersia: $D(X)=\lambda$. Abaterea standard este $\sigma=\sqrt{D(X)}=\sqrt{\lambda}$.

Remarcăm că în repartiția Poisson media este egală cu dispersia .

Observaţia 18. Repariția Poisson intervine în aplicații în studiul evenimentelor rare, drept urmare mai este cunoscută şi sub numele de *legea evenimentelor rare.* Enunțăm mai jos câteva exemple de variabile aleatoare care se supun legii de probabilitata a lui Poisson:

- Numărul de persoane care depăşesc vârsta de 100 de ani într-o comunitate umană;
- Numărul de locuri de muncă devenite vacante într-o anumită societate timp de un an;
- Numărul de unități din acelaşi produs vândute de un magazin timp de o zi;
- Numărul de clienți ce intră într-o bancă într-o zi;
- Numărul de particule α emise de un material radioactiv într-un interval de timp dat.
- numărul de autovehicule ce trec printr-o intersecție într-un interval de timp considerat.

3.3.5. Legi de repartiţie continue.

Pentru definirea distribuției unei variabilei aleatoare continue X, în care $X(\Omega)=[a,b]$ unde vom considera intervalul $(x,x+dx)$ a cărui măsură (lungime) dx este diferită de zero. Notăm cu $dP=P(x<X<x+dx)$, adică probabilitatea ca variabila aleatoare să ia o

valoare din intervalul $(x,x+dx)$. Legea de probabilitate a variabilei aleatoare continue X se defineşte astfel:

Definiţia 20. Fie X o variabilă aleatoare continuă definită pe spațiul fundamental Ω şi fie $X(\Omega)=[a,b]$ unde $-\infty \le a < b \le +\infty$. Atunci, cu notațiile de mai sus, se defineşte *legea de probabilitate* sau
densitatea de probabilitate a variabilei aleatoare X ca fiind funcția

$$\varphi : [a,b] \to \mathbf{R}, \text{ definită prin } \varphi(x) = \frac{dP}{dx}, \ \forall x \in [a,b] \quad , \qquad (28)$$

Distribuția variabilei aleatoare continue X se va nota

$$X : \begin{pmatrix} x \\ \varphi(x) \end{pmatrix}, x \in [a,b] \qquad (28')$$

Ca şi în cazul variabilei aleatoare discrete, densitatea de probabilitate poate fi definită pe întreaga mulțime $\mathbf{R}$ a numerelor reale, considerându-se nulă în afara intervalului $[a,b]$.

Astfel putem scrie : $\varphi : \mathbf{R} \to [0,1]$, $\varphi(x) = \begin{cases} \frac{dP}{dx}, x \in [a,b] \\ 0, x \notin [a,b] \end{cases}$ (28")

Propoziţia 13. Legea de probabilitate a unei variabile aleatoare şi continuă X are următoarele proprietăți:

(1) $\varphi(x) \ge 0, \ \forall x \in [a,b]$; (2). $\int_a^b \varphi(x)dx = 1$.

Reprezentarea grafică a distribuției variabilei aleatoare.

Pentru variabila aleatoare de tip continuu densitatea de probabilitate $\varphi(x), x \in [a,b]$ reprezentată grafic, este o curbă continuă numită *curbă de distribuție.*

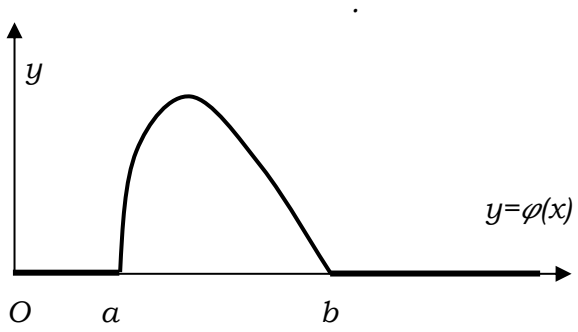


Fig. 2. Curba de distribuție

Funcția de repartiție a unei variabile continue X se defineşte în mod asemănător prin $F : \mathbf{R} \to [0,1]$, $F(x) = P(X < x), \ \forall x \in \mathbf{R}$

iar calculul său se face cu ajutorul integralelor.

Propoziţia 14. Fie X o variabilă aleatoare continuă care are

distribuția $X:\begin{pmatrix} x \\ \varphi(x) \end{pmatrix}, x \in [a,b]$. Atunci funcția sa de repartiție se calculează cu formula:

$$F(x)=\begin{cases} 0, & x \le a \\ \int_a^x \varphi(t)dt, & a < x \le b \\ 1, & x > b \end{cases}, \ \forall x \in \mathbf{R} \qquad (29)$$

Caracteristicile numerice ale unei variabile aleatoare de tip continuu se definesc deasemenea cu ajutorul integralelor astfel :

Definiţia 21. Fie X este o variabilă aleatoare continuă având distribuția

$$X:\begin{pmatrix} x \\ \varphi(x) \end{pmatrix}, x \in [a,b]$$

unde densitatea de probabilitate $\varphi(x)$ este o funcție integrabilă pe intervalul $[a,b]$. Atunci:

(a). Se numeşte *valoarea sa medie* a lui X numărul

$$m=M(X)=\int_a^b x\varphi(x)dx. \qquad (30)$$

(b). Se numeşte *dispersia* sau încă *varianța* variabilei aleatoare X numărul notat $D(X)=\sigma^2$, reprezentat de media pătratului abaterii lui X de valoarea medie sa medie m,

$$D(X)=\int_a^b (x-m)^2\varphi(x)dx. \qquad (31)$$

Numărul $\sigma=\sqrt{D(X)}$ se numeşte *abatere medie pătratică* sau *abatere standard.*

(c). Se numeşte *moment de ordin k* al variabilei aleatoare continue X, media variabilei aleatoare X^k, adică numărul notat:

$$m_k=M_k(X)=\int_a^b x^k\varphi(x)dx. \qquad (32)$$

Dacă în cazul variabilei aleatoare continue extremitățile intervalului de definiție al densității de probabilitate sunt infinite, adică $a \to -\infty$ sau $b \to +\infty$, atunci valoarea medie are sens dacă integrala improprie respectivă este convergentă.

1°. Repartiţia normală.

Repartiția normală este una din cele mai importante repartiții ale teoriei probabilităților şi a fost descoperită de matematicianul german Gauss.

Am văzut că în cazul repartiției binomiale legea de repartiție este $f(k) = C_n^k p^k (1-p)^{n-k}$, reprezentând probabilitatea de apariție de k ori a unui eveniment A dacă se efectuează n experienţe *independente* şi dacă în fiecare experiență probabilitatea de apariție a evenimentului A este constantă şi este egală cu p.

Dacă numărul n creşte ($n \to \infty$), iar probabilitatea de apariție a evenimentului considerat rămâne constantă între 0 şi 1, se ajunge la repartiția normală, care spre deosebire de repartiția binomială este o repartiție continuă. Astfel, repartiția normală se obține din repartiția binomială prin trecere la limită când $n \to \infty$.

Definiţia 22. Spunem că o variabilă aleatoare continuă X urmează o *lege de repartiție normală* $N(m,\sigma)$ cu parametrii m şi σ ($\sigma > 0$) dacă densitatea sa de probabilitate este dată de :

$$\varphi(x;m,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad \forall\, x \in \mathbf{R}. \tag{33}$$

Vom vedea mai târziu că parametrii m şi σ reprezintă media şi respectiv abaterea standard a unei variabile aleatoare repartizată norma.

Observaţia 19. *P*e baza rezultatelor din Analiza Matematică se demonstrează că funcția φ satisface proprietăți asemănătoare proprietăților (1) şi (2) ale unei legi de probabilitate din *Propoziția 13:*

(1). $\varphi(x) \geq 0$, $\forall\, x \in \mathbf{R}$; *(2).* $\int_{-\infty}^{+\infty} \varphi(x;m,\sigma)dx = 1$.

Graficul funcției $y=\varphi(x;m,\sigma)$ depinde de parametrii m şi σ şi are forma unui clopot, numit "*clopotul lui Gauss*".

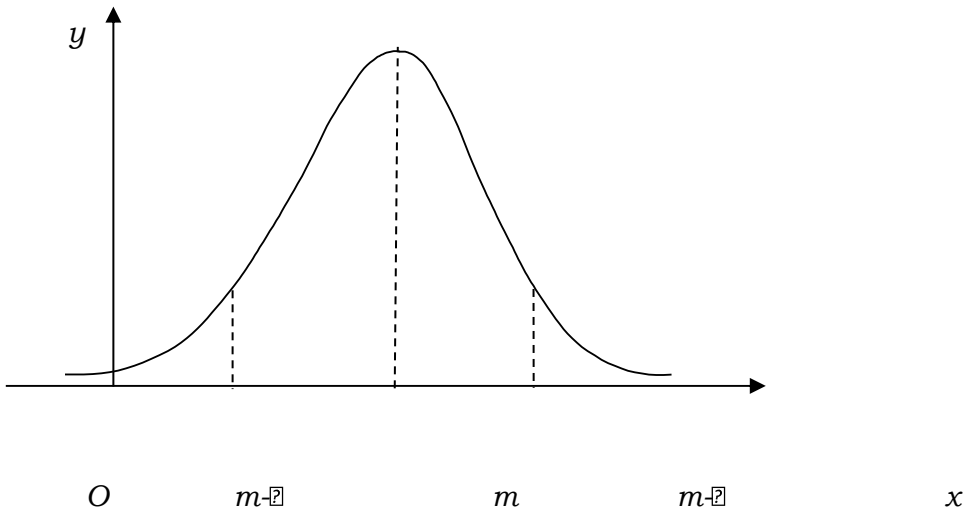


Fig. 3 . Curba de distribuție a repartiției normale.

Graficul este simetric față de dreapta $x=m$.

În punctul $x=m$ graficul admite un punct de maxim de valoare $f_{max} = f(m;m,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}$.

Acest grafic este cu atât mai "turtit" cu cât parametrul este mai mare, el apărând în numitorul ordonatei punctului de maxim.

Punctele $x=m-$ şi $x=m+$ sunt puncte de inflexiune.

Dreapta $y=0$ este asimptotă orizontală la grafic. Curba se apropie repede de axa Ox . În raport cu o abatere $|x-m| < 3\sigma$, diferența față de Ox este de ordinul a 0,003 unități. Astfel, repartiția normală poate fi considerată definită într-un interval închis şi finit.

Legea de repartiție normală se mai numeşte şi *legea de repartiție gaussiană.*

Pentru σ = 1; 2; 3; 4 şi $m=0$ graficele distribuției normale sunt de forma:

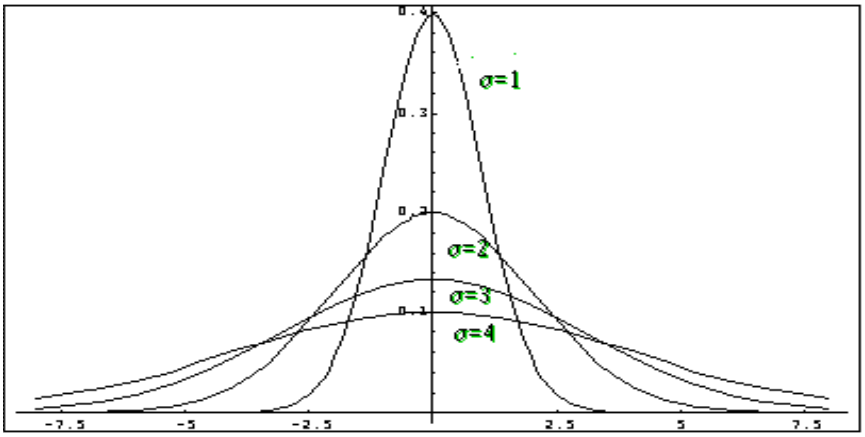


Fig. 4. Graficele distribuției normale pentru σ = 1; 2; 3; 4 şi m=0

Dacă în legea normală generală se face schimbarea de variabilă $u = \frac{x-m}{\sigma}$ se obține funcția $f(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$.

Aceasta funcție corespunde unei densități de probabilitate de parametri *m*=0 şi □□□□

Definiția 23. Pentru *m*=0 şi □=1, legea de repartiție normală se numeşte *legea normală normată* sau *repartiția normală redusă,* notată *N*(0,1).

În acest caz legea de probabilitate este

$$\varphi(x;0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad \forall\ x \in \mathbf{R}. \tag{34}$$

Folosind rezultate din Analiza Matematică se obține expresia funcției de repartiție a legii normale reduse.

Propoziţia 15. Funcția de repartiție a legii normale este :

$$F(x,m,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{(t-m)^2}{2\sigma^2}}dt \tag{35}$$

iar cea a legii normale reduse este

$$F(x,0,1)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt \tag{35'}$$

Observaţia 20. Funcția $F(x,0,1)$ este prin definiție probabilitatea ca variabila redusă să ia valori mai mici ca x şi ea reprezintă aria suprafaței de sub curba normală redusă cuprinsă între -∞ şi valoarea x.

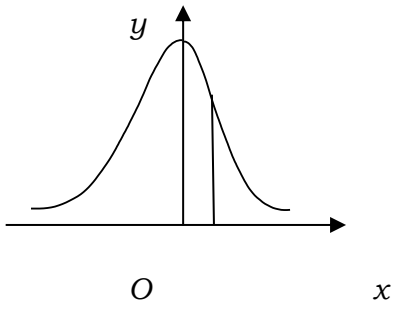


Fig. 5. Probabilitatea $P(X<x)=F(x,0,1)$.

Se arată că funcția de repartiție a legii normale reduse se exprimă prin :

$$F(x;0,1)=\begin{cases}\frac{1}{2}-\Phi(x), & x<0\\ \frac{1}{2}, & x=0\\ \frac{1}{2}+\Phi(x), & x>0\end{cases}, \tag{36}$$

unde $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt$ este *funcția integrală a lui Laplace* (valorile sale fiind date cu precizie în tabele, vezi *ANEXA, Tabela 1*).

Ea reprezintă probabilitatea ca variabila normală redusă să ia valori între 0 şi valoarea *x*: $\Phi(x)=P(0\le X\le x)$.

Funcția de repartiție a unei variabile aleatoare ce urmează legea normală generală este dată de:

$$F(x,m,\sigma)=\frac{1}{2}+\Phi\left(\frac{x-m}{\sigma}\right). \qquad (36')$$

Frecvent se utilizează dublul funcției lui Laplace:

$$P=2\cdot\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-x}^{x} e^{-\frac{t^2}{2}}dt=P(-x\le X\le x)$$

care reprezintă aria suprafeței de sub curba normală redusă cuprinsă între - *x* şi *x* şi poartă numele de *probabilitate de acoperire.*

Probabilitatea complementară $1-P=1-2\Phi(x)$ este probabilitatea ca valorile variabilei să ia valori în afara intervalului [-*x*, *x*] şi ea poartă numele de *probabilitate de risc* sau *nivel (prag) de semnificație.*

Aplicații.

Cu ajutorul rezultatelor de mai sus putem rezolva următoarele tipuri de probleme inverse una alteia privind o variabilă aleatoare care urmează legea normală :

a). Determinarea probabilității ca o variabilă aleatoare repartizată normal, să ia valori într-un interval dat (a,b).

Din proprietățile funcției de repartiție deducem următoarele probabilități:

$$P(a<X<b)=P(X<b)-P(X<a)=F(b)-F(a)=\Phi\left(\frac{b-m}{\sigma}\right)-\Phi\left(\frac{a-m}{\sigma}\right) \qquad (37)$$

Atunci, pentru □□>□0, inegalitatea |*X*-*m*|<□□ este echivalenta cu dubla inegalitate *m*-□□< *X* < *m*+□ , de unde deducem că probabilitatea ca variabila aleatoare *X* care urmează o lege de repartiție normală, să ia valori în intervalul (*m*-□□□*m*+□)□este□

$$P(|X-m|<\varepsilon)=P(m-\varepsilon<X<m+\varepsilon)=\Phi\left(\frac{\varepsilon}{\sigma}\right)-\Phi\left(-\frac{\varepsilon}{\sigma}\right)=2\Phi\left(\frac{\varepsilon}{\sigma}\right),$$

unde am folosit faptul că funcția este o funcție impară.

Pentru $k=\frac{\varepsilon}{\sigma}$ obținem următoarea evaluare a acestei probabilități: $P(|X-m|<k\sigma)=2\cdot\Phi(k)$, (valorile acestor probabilități sunt date în tabele, vezi *ANEXE, Tabela 2*). De exemplu, pentru *k=1,2,3,4* găsim în tabele (1)=0,3413, (2)=0,4772, (3)=0,49865, (4)=0,49997 şi prin urmare avem:

$$P(|X-m|<\sigma)=2\cdot\Phi(1)=0{,}6823\approx 0{,}68;$$

$$P(|X-m|<2\sigma)=2\cdot\Phi(2)=0{,}9544\approx 0{,}95;$$

$$P(|X-m|<3\sigma)=2\cdot\Phi(3)=0{,}9973\approx 0{,}997;$$
$$P(|X-m|<4\sigma)=2\cdot\Phi(4)=0{,}99994\approx 0{,}959.$$

Exemplul 8. Timpul de păstrare în condiții proprii consumului a unui produs alimentar este o variabilă aleatoare *X* distribuită normal de parametrii *m=M(X)=45* zile şi de abatere medie pătratică $\sigma=5$ zile. Să se calculeze probabilitatea ca un produs să se altereze înainte de 50 de zile.

Soluție: Avem:

$$P(X\le 50)=F(50)=\frac{1}{2}+\Phi\left(\frac{50-m}{\sigma}\right)=\frac{1}{2}+\Phi\left(\frac{50-45}{5}\right)=\frac{1}{2}+\Phi(1)=$$
$$=0{,}5+0{,}34=0{,}84.$$

Prin urmare cu o probabilitate de 84 % produsul expiră înainte de 50 de zile.

Exemplul 9. Să presupunem că se recepționează un lot de produse alimentare şi să considerăm că valorile unei caracteristici a sa (cost, greutate, etc.), sunt repartizate după o lege normală de parametrii *m*=200 şi dispersie $\sigma^2=64$. Luând la întâmplare 100 din aceste produse, care este probabilitatea de a se abate cu mai mult de 8 unități de la valoarea medie de 200?

Soluție: Să notăm cu *X* variabila aleatoare care ia aceste valori şi se supune legii normale *N(200,8).* Trebuie să determinăm probabilitatea $P(m-\sigma<X<m+\sigma)=P(200-8<X<200+8)=2\Phi(1)$.

Din tabelul cu valorile funcției lui Laplace Φ obținem $\Phi(1) = 0{,}3413$ şi deci probabilitatea căutată este p = 0,6826, $q = 1 - p$ = 0,3174, adică 31,74% de produse se abat cu mai mult de 8 unități de la valoarea

medie de 200 unități.

b). Determinarea unui interval (a,b) în care ia valori o variabilă aleatoare repartizată normal cu o probabilitate dată.

Pentru P=?? dat, din egalitatea $P(|X - m| < k\sigma) = 2 \cdot \Phi(k)$, deducem

$\Phi(k) = \frac{\alpha}{2}$. Cu ajutorul tabelelor funcției ?(*k*) deducem valoarea lui k şi

atunci intervalul căutat este (*a,b*)=(*m-k*?? *m+k*??

Exemplul 10. O variabilă aleatoare X are distribuția normală cu parametrii m=30 şi ?=10.

(a). Să se determine probabilitatea ca variabila aleatoare să ia valori în intervalul (10,0).

(b). Să se determine un interval în care se găsesc valorile variabilei cu o probabilitate de 0,75.

Soluție. (a). Conform celor spuse mai sus avem:

$$P(10 < X < 50) = \Phi\left(\frac{50-30}{10}\right) - \Phi\left(\frac{10-3}{10}\right). = \Phi(2) - \Phi(-2) =$$

$$= 2\Phi(2) \approx 2 \cdot 0{,}4772 = 0{,}9544$$

(b). Din egalitatea $P(|X - 30| < 10k) = 2 \cdot \Phi(k) = 0{,}75$, găsim ?(*k*)=0,375, de unde deducem $k \approx 1{,}5$. Atunci intervalul în care se găsesc valorile lui X cu probabiliitatea 0,75 este (30-10*k*, 30+10*k*)=(15,45).

Observaţia 21. Dacă X este o variabilă aleatoare care urmează legea binomială *B(n,p)* de parametrii n *şi* p $(0 < p < 1)$ şi dacă n, np şi nq sunt mari atunci putem aprecia că variabila X urmează şi legea normală *N(m,σ)* de parametrii m=np şi $\sigma = \sqrt{npq}$. Prin urmare variabila redusă $Z = \frac{X - m}{\sigma}$ urmează legea normală normată *N(0,1).*

Exemplul 11. Se aruncă un zar de 300 de ori. Notăm cu X variabila aleatoare reprezentând numărul de apariții ale feței cu 5 puncte. Să se calculeze probabilitatea ca în cele 300 de aruncări fața cu 5 puncte să apară de cel puțin 55 de ori.

Soluție: Notăm cu X variabila aleatoare reprezentând numărul de apariții ale feței cu 5 puncte. Legea căreia i se supune X este legea binomială de parametrii $n=300$ şi $p=\frac{1}{6}$. Avem: $m=M(X)=np=50$,

$\sigma^2 = D(X) = npq = 41{,}67$, $\sigma = \sqrt{D(X)} = \sqrt{41{,}67} \approx 6{,}45$.

Cum n=300 este suficient de mare legea binomială poate fi aproximată cu legea normală $N(m,\sigma)$. Atunci putem scrie:

$$P(X \geq 55) = 1 - P(X < 55) = 1 - F(55) = 1 - \left[\frac{1}{2} + \Phi\left(\frac{55-m}{\sigma}\right)\right] =$$

$$= \frac{1}{2} - \Phi\left(\frac{55-50}{6{,}45}\right) = 0{,}5 - \Phi(0{,}77) \approx 0{,}5 - 0{,}28 = 0{,}22 .$$

Prin urmare, cu o probabilitate de 22 % fața cu 5 puncte apare de cel puțin 55 de ori.

Folosind rezultate din Analiza Matematică se determină valorile caracteristice ale legii normale.

Propoziția 16. Valorile sale caracteristice ale unei variabile aleatoare X care urmează legea normală sunt:

(a). Media lui X este: $M(X)=m$.

(b). Momentul de ordinul 2 este: $M_2(X)= m^2+$□

(c)□□□Dispersia lui X este : $D(X)=$□^{2}.

Din rezultatele de mai sus deducem că parametrii m şi □ ai legii normale sunt respectiv, valoarea medie şi abaterea medie pătratică ale distribuției respective.

2°. Repartiția "*t*" (*Student*).

Definiţa 24. Spunem că o variabilă aleatoare *X* urmează o lege de repartiţie *"t"* sau lege de repartiţie *Student* cu *n* grade de libertate dacă densitatea sa de probabilitate este:

$$\varphi(x;n)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\cdot\Gamma\left(\frac{n}{2}\right)}\cdot\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}},\ x\in\boldsymbol{R}\ .$$

(38)

Funcţia $\varphi(x,n)$ este o densitate de probabilitate, adică satisface condiţiile:

$$\varphi(x,n)\geq 0,\ \forall\ x\in\boldsymbol{R}\ \text{şi}\ \int_{-\infty}^{+\infty}\varphi(x;n)dx=1\ .$$

Din definiţia funcţiei de repartiţie deducem imediat:

Funcţia de repartiţie a unei variabile aleatoare ce are o distribuţie "*t*" este:

$$F(t)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\cdot\Gamma\left(\frac{n}{2}\right)}\cdot\int_{-\infty}^{t}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}dx\ ,$$

(39)

şi ea are proprietatea :

$$F(t)=1-F(-t).$$

(40)

În practica statisticii matematice pentru distribuţia *Student* au fost întocmite tabele pentru probabilitatea

$$P=P(-t<X<t)=F(t)-F(-t)=2F(t)-1\ .$$

pentru diferite grade de libertate (vezi ANEXE, tabelul 5). Aceste tabele permit aflarea lui *t,* când se cunoaşte probabilitatea P, determinând mai întâi $F(t)=(1+P)/2$ şi apoi extrăgând valoarea lui $t=t(P,k)$ corespunzătoare, sau, aflarea lui P , când se cunoaşte *t*, extrăgând din tabel *F(t)* şi apoi calculând P cu formula $P=2F(t)-1$.

Observaţia 22. Dacă numărul gradelor de libertate *n* tinde la infinit atunci se arată că $\varphi(t;n)$ tinde către distribuţia normală normată $\varphi(t;0,1)$. Curba lui $\varphi(t;n)$ este asemănătoare cu curba normală. Ea este cu atât mai teşită cu cât numărul gradelor

de libertate n este mic. Cu cât numărul gradelor de libertate creşte, curba lui $\varphi(t;n)$ se apropie de curba lui $\varphi(t;0,1)$. Aceasta are loc când n >30.

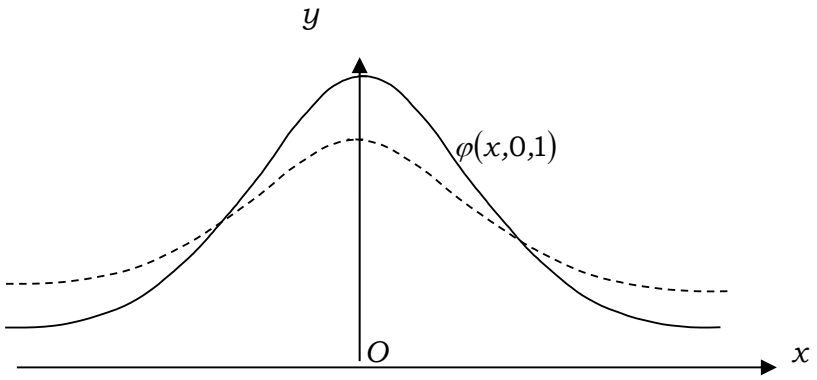


Fig.6 Graficele distribuției normale normate (curba $y = \varphi(x,0,1)$) şi distribuției Student (curba $y = \varphi(x,n)$).

Observaţia 23. O proprietate importantă a legii de probabilitate a distribuției "*t*" este aceea că ea nu depinde de parametrul φ şi prin aceasta are o mare aplicabilitate în Statistică, îndeosebi în determinarea intervalelor de încredere pentru media teoretică m a unei variabile statistice cât şi în verificarea ipotezelor statistice.

Observaţia 24. Se demonstrează că dacă X, X_1, X_2, …, X_n sunt variabile aleatoare independente cu repartiții normate $N(0,1)$, atunci variabila aleatoare $T = \dfrac{X}{\sqrt{\dfrac{1}{n}\sum_{j=1}^{n} X_j^2}}$ urmează o lege de repartiție *Student*

cu n grade de libertate.

Propoziţia 17. Valorile caracteristice ale unei variabile aleatoare X

care urmează legea de repartiție Student cu n grade de libertate sunt:

a). Media lui X este: $M(X)=0$.

b). Momentul de ordinul 2 este: $M_2(X)=\dfrac{n}{n-2}$.

Dispersia lui X este : $D(X) = \dfrac{n}{n-2}$.

d). Abaterea standard este: $\sqrt{D(X)} = \sqrt{\dfrac{n}{n-2}}$.

1.3.6.Inegalitatea lui Cebîşev.

În numeroase aplicații practice ale teoriei probabilităților *inegalitatea lui Cebîşev* oferă informații asupra distribuției unei variabile aleatoare *X*, aceasta dând o margine inferioară pentru probabilitatea evenimentului ca valorile variabilei să se grupeze într-un interval centrat în jurul valorii medii *m=M(X),* adică pentru $P(|X - M(X)| < \varepsilon)$, unde ε>0, este un număr arbitrar. Aceste informații sunt cu atât mai precise, deci probabilitatea respectivă se apropie de valoarea 1, cu cât dispersia variabilei *D(X)* este mai mică.

În cele ce urmează admitem următorul rezultat foarte important:

Teorema 1. Dacă *X* o variabilă aleatoare de medie *m=M(X)* , de dispersie $\sigma^2=D(X)$ şi ε este un număr pozitiv oarecare, atunci are loc inegalitatea

$$P(|X - m| < \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2} \quad , \textit{(Inegalitatea lui Cebîşev)} \tag{41}$$

Observaţia 25. Dacă trecem la evenimentul contrar, inegalilatea lui Cebîşev se scrie :

$$P(|X - m| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}. \tag{42}$$

Observaţia 26. Punând $\frac{\varepsilon}{\sigma} = k$, inegalitatea lui Cebîşev se scrie:

$$P(|X - M(X)| < k\sigma) \geq 1 - \frac{1}{k^2} \tag{43}$$

Se observă că probabilitatea $P(|X - m| < k\sigma)$ creşte odată cu creşterea lui $1 - \frac{1}{k^2}$, care se întâmplă când *k* devine mare.

De exemplu pentru *k*=3 avem $P(m - 3\sigma < X < m + 3\sigma) \geq \frac{8}{9}$, iar pentru *k=6* avem $P(m - 6\sigma < X < m + 6\sigma) \geq \frac{35}{36}$, adică o probabilitate destul de apropiată de 1, ceea ce înseamnă că o variabilă aleatoare ia aproape sigur valorile cuprinse în intervalul $(m - 6\sigma, m + 6\sigma)$. Observăm că abaterile $|X - m|$ mai mari ca 3σ au probabilități foarte mici, deci şansele acestor evenimente de a se produce sunt extrem de reduse.

Inegalitatea lui Cebîşev poate fi aplicată diverselor variabile aletoare cărora le cunoaştem legile de distribuţie şi deci, media şi dispersia.

1º. Cazul variabilei aleatoare care are distribuţia binomială .

Ţinând seama de valorile caracteristice ale mediei şi dispersiei unei

variabilei aleatoare X care urmează legea binomială şi anume

$m = M(X) = np$ şi $\sigma^2 = D(X) = np(1-p)$ obţinem:

Propoziţia 18. Dacă X este o variabilă aleatoare care reprezintă numărul total de apariţii ale unui eveniment A în n probe independente ale unui acelaşi experiment şi p este probabilitatea de apariţie a evenimentului A la oricare probă a experimentului atunci inegalitatea lui Cebîşev aplicată acestei variabile se scrie:

$$P(|X - np| < \varepsilon) \geq 1 - \frac{np(1-p)}{\varepsilon^2}. \qquad (44)$$

Consecinţă. Dacă $Z = \frac{X}{n}$ este variabila aleatoare care reprezintă *frecvenţa relativă* de apariţie a unui eveniment A în n probe ale unui acelaşi experiment şi p este probabilitatea de apariţie a evenimentului A la oricare probă a experimentului atunci inegalitatea lui Cebîşev aplicată acestei variabile se scrie:

$$P\left(\left|\frac{X}{n} - p\right| < \varepsilon\right) \geq 1 - \frac{p(1-p)}{n\varepsilon^2}. \qquad (44')$$

Într-adevăr, ţinând seama de raţionamentul propoziţiei anterioare, notând cu $Z = \frac{X}{n}$ variabila aleatoare care reprezintă frecvenţa relativă de apariţia a evenimentului A în cele n probe independente, pentru această variabilă aleatoare avem: $M(Z) = M\left(\frac{X}{n}\right) = \frac{1}{n}M(X) = \frac{1}{n}np = p$;

$$D(Z) = D\left(\frac{X}{n}\right) = \frac{1}{n^2}D(X) = \frac{1}{n^2}npq = \frac{pq}{n} = \frac{p(1-p)}{n}.$$

Inegalitatea lui Cebâşev pentru variabila $Z = \frac{X}{n}$ se scrie:

$$P\left(\left|\frac{X}{n}-p\right|<\varepsilon\right)\geq 1-\frac{p(1-p)}{n\varepsilon^2}.$$

Exemplul 12. Se aruncă un zar de 1000 de ori. Fie X variabila aleatoare care reprezintă numărul de apariții ale feței cu 3 puncte, $Z=\frac{X}{10^3}$ variabila aleatoare care reprezintă *frecvența relativă* de apariție a feței cu 3 puncte iar $p=\frac{1}{6}$ probabilitatea de producere a acestui eveniment. Utilizând inegalitatea lui Cebâşev să se găsească limita inferioară a probabilității ca frecvența relativă să nu difere de probabilitate cu mai mult de 0,01.

Soluție. Avem de determinat limita inferioară a probabilității

$P\left(\left|\frac{X}{10^3}-\frac{1}{6}\right|<0{,}01\right)$. Aplicăm formula (41). Obținem

$P\left(\left|\frac{X}{n}-p\right|<\varepsilon\right)\geq 1-\frac{p(1-p)}{n\varepsilon^2}$, unde $n=1000, \varepsilon=0{,}01, p=\frac{1}{6}, q=1-\frac{1}{6}=\frac{5}{6}$.

Astfel avem: $P\left(\left|\frac{X}{n}-p\right|<\varepsilon\right)\geq 1-\frac{5}{36\cdot 10^3\cdot 10^{-2}}=1-\frac{5}{360}=\frac{355}{360}\approx 0{,}986$

Exemplul 13. Într-o cercetare ştiințifică se efectuează n experimente, urmărindu-se apariția unei anumite caracteristici. Să se determine numărul minim de experimente astfel încât, cu o probabilitate de cel puțin 0,95, frecvența relativă de apariție să difere în valoare absolută de probabilitatea p cu mai puțin de 10^{-3}.

Soluție. Aplicând inegalitatea lui Cebâşev pentru $\varepsilon=10^{-3}$ se obține:

$P\left(\left|\frac{X}{n}-p\right|<\varepsilon\right)\geq 1-\frac{pq}{n\varepsilon^2}=1-\frac{p(1-p)}{n\cdot 10^{-6}}$. Atunci din inegalitatea $1-\frac{p(1-p)}{n\cdot 10^{-6}}\geq 0{,}95$ deducem $n\geq\frac{p(1-p)\cdot 10^6}{0{,}05}=2p(1-p)\cdot 10^7$.

2º. Cazul variabilei aleatoare care urmează legea normală.

Ținând seama de faptul că pentru legea normală media şi dispersia sunt exact parametrii m şi σ ai acestei legi avem:

Propoziția 19. Dacă X este o variabilă aleatoare continue care urmează *legea normală* de parametrii m şi σ, atunci inegalitatea lui Cebîşev aplicată acesteia se scrie :

$$P\left(|X-m|<\varepsilon\right)\geq 1-\frac{\sigma^2}{\varepsilon^2} \qquad (45)$$

Într-adevăr, am văzut că variabila aleatoare care are distribuția normală de parametrii m şi σ are media $M(X)=m$ şi dispersia $D(X)=\sigma^2$. Atunci inegalitatea lui Cebîşev ia forma: $P\left(|X-m|<\varepsilon\right)\geq 1-\frac{\sigma^2}{\varepsilon^2}$ sau $P\left(|X-m|<k\sigma\right)\geq 1-\frac{1}{k^2}$, unde $k=\frac{\varepsilon}{\sigma}$.

Observaţia 27. Inegalitatea lui Cebîşev dă o margine inferioară pentru probabilitatea $P\left(|X-m|<\varepsilon\right)$ şi se interpretează astfel: cu o probabilitate cel puțin egală cu $1-\frac{\varepsilon^2}{\sigma^2}$ respectiv $1-\frac{1}{k^2}$, variabila aleatoare X ia valori în intervalul ($m-$ $m+$) respectiv ($m-k$ $m+k$).

Mai mult, se poate determina probabilitatea minimă ca variabila aleatoare să ia valori într-un interval dat (a,b), găsind valoarea maximă a lui k din condițiile ($m-k$ $m+k$) (a,b).

Această inegalitate oferă şi o soluție pentru problema inversă: se poate determina un interval minim în care se află valorile unei variabile aleatoare X cu o probabilitate $P=$dată. Se determină k minim din condiția $1-\frac{1}{k^2}\geq\alpha$ şi atunci intervalul minim căutat va fi ($m-k$ $m+k$.

Exemplul 14. Fie X o variabilă aleatoare care reprezintă lungimea unor piese prelucrate de o maşină şi care are media $M(X)=m=50$ şi dispersia $D(X)=\sigma^2=0,1$. Utilizând inegalitatea lui Cebîşev, să se determine probabilitatea ca lungimea X a pieselor să fie cuprinsă între 49,5 cm şi 50,5 cm.

Soluție: Avem de evaluat probabilitatea $P(49,5<X<50,5)$. Deoarece $m=50$, condiția $49,5<X<50,5$ se mai poate scrie $-0,5<X-50<0,5$ sau $|X-50|<0,5$ rezultă $\varepsilon=0,5$. Aplicând

inegalitatea lui Cebâşev obținem: $P\left(|X-50|<0,5\right)\geq 1-\frac{0,1}{0,25}=1-0,4=0,6$.

EXERCIŢII ŞI PROBLEME SUPLIMENTARE.

1. Un student are de pregătit pentru un număr de întrebări (subiecte teoretice) din trei capitole ale cursului. La examen trage un bilet care conține 3 întrebări, câte una din fiecare capitol. Să se scrie evenimentele:

A_0 evenimentul care constă în faptul că studentul nu ştie să răspundă la niciuna dintre cele trei întrebări de pe bilet;

A_1 evenimentul care constă în faptul că studentul ştie să răspundă numai la una dintre cele trei întrebări de pe bilet şi nu ştie să răspundă la celelalte două;

A_2 evenimentul care constă în faptul că studentul ştie să răspundă la două dintre cele trei întrebări de pe bilet şi nu ştie să răspundă la cea de-a treia;

A_3 evenimentul care constă în faptul că studentul ştie să răspundă la toate cele trei întrebări de pe bilet.

B_1 evenimentul care constă în faptul că studentul ştie să răspundă la cel puțin o întrebare de pe bilet;

B_2 evenimentul care constă în faptul că studentul ştie să răspundă la cel puțin două întrebări de pe bilet.

C_1 evenimentul care constă în faptul că studentul ştie să răspundă la cel mult o întrebare de pe bilet;

C_2 evenimentul care constă în faptul că studentul ştie să răspundă la cel mul două întrebări de pe bilet.

2. Mergând pe un traseu un automobilist întâlneşte patru intersecții semaforizate. La fiecare semafor culoarea roşie durează 60 secunde, cea galbenă 5 secunde iar cea verde 25 secunde. Cele 4 semafoare nu sunt sincronizate şi presupunem că apariția unei culori la un semafor întâlnit nu depinde de culorile întâlnite la semafoarele anterioare.

(a). Notând cu R_i , G_i şi V_i evenimentele ca la semaforul „*i*” (i=1,2,3,4) automobilistul să întâlnească respectiv culoarea roşie, cea galbenă sau cea verde, să se calculeze probabilitățile acestor evenimente.

(b). Să se reprezinte evenimentul ca automobilistul să întâlnească pe rând culorile roşu, galben,verde, verde.

(c). Notând cu A_k $(k = 0,1,2,3,4)$ evenimentul ca în drumul său automobilistul să întâlnească k semafoare verzi, să se reprezinte aceste evenimente cu ajutorul evenimentelor R_i , G_i şi V_i (i=1,2,3,4).

3. Se consideră 3 urne U_1, U_2 şi U_3. În fiecare urnă se află cîte 90 bile de trei culori (5 galbene, 25 verzi şi 60 roşii). Din fiecare urnă se extrage cîte o bilă. Notăm respectiv cu G_i, V_i şi R_i (i=1,2,3) evenimentul ca la extragerea din urna „U_i" să apară bila de culoarea galbenă, verde, respectiv roşie.

a). Să se calculeze probabilităţile evenimentelor G_i, V_i şi R_i .

b). Notând cu A_k $(k = 0,1,2,3)$ evenimentul care constă în obţinerea de k bile roşii în cele trei extrageri, să se reprezinte aceste evenimente cu ajutorul evenimentelor R_i , G_i şi V_i (i=1,2,3,4) şi să se calculeze probabilităţile acestor evenimente.

4. O urnă conţine 3 bile albe şi 7 bile negre, iar alta conţine 6 bile albe şi 3 bile negre. Din fiecare urnă se extrage câte o bilă.

a). Care este probabilitatea să obţinem cel puţin o bilă albă?

b). Care este probabilitatea ca cele două bile să fie negre?

c). Care este probabilitatea ca o bilă să fie albă şi alta neagră?

5. O urnă conţine 4 bile albe şi 6 bile negre. Se cere probabilitatea ca extrăgând de 3 ori câte o bilă, fără a pune bila extrasă înapoi după fiecare extragere, să obţinem la prima extragere o bilă albă iar la următoarele extrageri să obţinem câte o bilă neagră.

6. O urnă U_1 conţine 2 bile albe şi o bilă neagră iar o altă urnă U_2 conţine o bilă albă şi 5 bile negre. Se extrage o bilă din urna U_1 şi se introduce în urna U_2, apoi se extrage o bilă din urna U_2. Ştiind că bila extrasă din urna U_2 este albă care este probabilitatea ca bila transferată să fi fost neagră.

7. O urnă U_1 conţine 3 bile albe şi 3 bile negre iar o altă urnă U_2 conţine 2 bile albe şi 4 bile negre. Din aceste urne s-a extras o bilă albă. Care este probabilitatea ca ea să provină din prima urnă?

8. Un lot de piese conţine 5% piese rebut. Controlul de calitate stabileşte ca regulă de acceptare a lotului condiţa ca la 5 verificări consecutive să nu fie nici o piesă rebut. Care este probabilitatea de acceptare a lotului?

9. În trei urne U_1, U_2, U_3 sunt câte 12 bile (albe şi negre), după cum urmează: U_1 (6 albe, 6 negre), U_2 (8 albe, 4 negre) şi U_3 (10 albe, 2 negre). Din fiecare urnă se extrage câte o bilă.

a). Pentru fiecare i=1,2,3 notăm cu A_i evenimentul ca bila extrasă din urna U_i să fie albă şi cu $\overline{A_i}$ evenimentul contrar al acestuia. Să se calculeze probabilităţile : $p_i=P(A_i)$, $q_i=P(\overline{A_i})$, i=1,2,3.

b). Notând cu S_k $(k = 0,1,2,3)$ evenimentul care constă în obținerea de k bile albe în cele trei extrageri, să se reprezinte aceste evenimente cu ajutorul evenimentelor A_i şi $\overline{A_i}$ (i=1,2,3) şi să se calculeze probabilitățile acestor evenimente.

10. O urnă conține 3 bile albe şi 7 bile negre, iar alta conține 7 bile albe şi 3 bile negre. Din fiecare urnă se extrage câte o bilă. Care este probabilitatea să obținem cel puțin o bilă albă?

11. De-a lungul unei şosele sunt trei bariere de cale ferată păzite. Probabilitatea ca un automobil care circulă pe şosea să găsească oricare din bariere deschisă este p=0,8.

Să se scrie variabila aleatoare X care reprezintă numărul de bariere deschise pe care le poate întâlni automobilul.

12. Un student are de pregătit pentru examenul de MATEMATICĂ 30 de întrebări (subiecte teoretice), din care: 12 din capitolul I, 10 din capitolul al II-lea şi 8 din capitolul al III-lea. El, însă, pregătreşte doar 15 subiecte şi anume: 6 din capitolul I, 4 din capitolul al II-lea şi 5 din capitolul al III-lea. La examen trage un bilet care conține 3 întrebări, câte una din fiecare capitol. Să se scrie variabila aleatoare care reprezintă numărul de întrebări la care poate răspunde studentul.

13. Un student este supus unui test grilă cu 25 de întrebări cu răspunsuri multiple. El răspunde la fiecare dintre întrebări în următoarele situații : ştie răspunsul cu probabilitatea *p=0,5* sau îl ghiceşte cu probabilitatea *1-p=0,5***.** Admitem că studentul care ghiceşte răspunde corect la una din cele 25 întrebări posibile. Care este probabilitatea condiționată ca studentul să fi ghicit răspunsul la una din întrebări, dacă a răspuns corect la aceasta.

14. Într-un raft sunt cămăşi de acelaşi fel de talia I şi a-II-a în proporție de 49% (I) şi 51% (II), identic ambalate. Care este probabilitatea ca un cumpărător care doreşte o cămaşă de talia II să o găsească numai la a 6-a încercare?

15. Patru baschetbalişti aruncă mingea la coş. Primul aruncă cu o probabilitatea de 4/5, al doilea cu probabilitatea de 5/6, al treilea cu probabilitatea de 6/7 iar al patrulea cu probabilitatea de 7/8. Dacă fiecare execută câte o aruncare, care este probabilitatea ca trei să marcheze iar al altul să rateze?

16. Un magazin primeşte în cursul unei săptămâni 100 bucăți dintr-o

anumită marfă provenită de la fabricile A, B, C. Probabilitatea ca marfa să

provină de la fabrica A este de 0,6; de la fabrica B este de 0,2; de la fabrica C este de 0,2. Care este probabilitatea ca din cele 100 bucăți primite, 60 să fi fost realizate la fabrica A, 30 la fabrica B, iar restul la C?

17. Un lot de piese conține 5% piese rebut.Controlul de calitate stabileşte ca regulă de acceptare a lotuluicondiția ca la 5 verificări consecutive să nu fie nici o piesă rebut.Care este probabilitatea de acceptare a lotului?

18. O firmă se aprovizionează de la 4 furnizori. Din datele statistice deținute se estimează că doi dintre furnizori onorează contractele cu probabilitatea 0,8, iar ceilalți doi cu probabilitatea 0,9. Se cer probabilitățile următoarelor evenimente:

(a) toți furnizorii onorează contractul;

(b) doar doi furnizori onorează contractul;

(c) nici un furnizor nu onorează contractul;

(d) cel puțin un furnizor onorează contractul.

19. Se experimentează trei tipuri de aparate. Probabilitățile ca prototipurile să corespundă normelor standard sunt respectiv p_1=0,9, p_2=0,8, p_3=0,85. Să se calculeze probabilitatea ca un număr de k , $(k=0,1,2,3)$ prototipuri să corespundă normelor standard.

21. Să considerăm experiența aleatoare a aruncării a două zaruri. Să se determine tabloul de distribuție al variabilei aleatoare X care reprezintă suma punctelor care apar pe cele două zaruri

22. O urnă conține 5 bile albe şi 3 bile negre. Se efectuează 3 extregeri succesive. Notăm cu X variabila aleatoare care ia ca valori numărul de bile albe ce se pot obține în urma celor trei extrageri, când bila se pune înapoi în urnă după fiecare extragere respectiv cu Y variabila aleatoare care ia ca valori numărul de bile albe ce se pot obține în urma celor trei extrageri, când bila extrasă nu se pune în urnă după fiecare extragere.

a). Să se scrie tablourile de distribuție ale variabilelor aleatoare X şi Y.

b). Să se reprezinte poligoanele de repartiție ale celor două variabile aleatoare şi să se calculeze mediile şi dispersiile lor.

23. De-a lungul unei şosele sunt trei bariere de cale ferată păzite. Probabilitatea ca o maşină care circulă pe şosea să găsească oricare din bariere deschisă este p=0,8.

Să se scrie tabloul de distribuție al variabilei aleatoare X care reprezintă numărul de bariere trecute de maşină până la întâlnirea primei bariere închise, să se reprezinte poligonul de repartiție şi să se calculeze media şi dispersia variabilei aleatoare X.

24. În trei urne U_1, U_2, U_3 sunt câte 6 bile (albe şi negre), după cum urmează: U_1 (3 albe, 3 negre), U_2 (4 albe, 2 negre) şi U_3 (5 albe, 1 neagră). Din fiecare urnă se extrage câte o bilă. Pentru fiecare i=1,2,3 notăm cu A_i evenimentul ca bila extrasă din urna U_i să fie albă şi cu $\overline{A_i}$ evenimentul contrar al acestuia.

a). Să se calculeze probabilităţile evenimentelor A_i şi $\overline{A_i}$: $p_i=P(A_i)$, $q_i=P(\overline{A_i})$, i=1,2,3.

b). Notând cu X variabila aleatoare care reprezintă numărul de bile albe ce se pot obţine în urma celor trei extrageri, să se scrie tabloul de distribuţie al acestei variabile aleatoare, să se reprezinte poligonul de repartiţie şi să se calculeze media şi dispersia sa.

25. Într-un atelier trei maşini lucrează acelaşi fel de piese. Prima dă 10% rebuturi, a doua 20 % şi a treia 30 %. Se ia la întâmplare câte o piesă de la fiecare maşină. Fie X variabila aleatoare ce reprezintă numărul de piese bune din cele trei luate la întâmplare. Să se scrie tabloul de distribuţie al variabilei aleatoare X şi să se reprezinte poligonul de distribuţie. Să se calculeze media şi dispersia lui X.

26. Două variabile aleatoare discrete independente au distribuţiile

$$X:\begin{pmatrix}2 & 3 & 5\\ 0{,}2 & 0{,}3 & 0{,}5\end{pmatrix} \quad \text{şi} \quad Y:\begin{pmatrix}1 & 4 & 6\\ 0{,}6 & 0{,}2 & 0{,}2\end{pmatrix}.$$

Să se scrie distribuţiile variabilelor $X+Y$ şi $X\cdot Y$, X^2, Y^2.

27. Fie tabloul : $\begin{pmatrix}1 & 2 & 3 & 4\\ 0{,}2 & 0{,}1 & \alpha & 0{,}3\end{pmatrix}$. Pentru ce valoare a lui α acest tablou reprezintă distribuţia unei variabile aleatoare?

28. Să se determine funcţia de repartiţie, media şi dispersia variabilei X care are distribuţia $X:\begin{pmatrix}0 & 1 & 2 & 3 & 4\\ 0{,}10 & 0{,}15 & 0{,}20 & 0{,}25 & 0{,}40\end{pmatrix}$.

29. Se consideră variabilele aleatoare independente. $X:\begin{pmatrix}1 & 2 & 3 & 4\\ 0{,}1 & 0{,}4 & 0{,}3 & 0{,}2\end{pmatrix}$,

$Y:\begin{pmatrix}0 & 1 & 2 & 3 & 4\\ 0{,}10 & 0{,}15 & 0{,}20 & 0{,}25 & 0{,}40\end{pmatrix}$.

Să se calculeze: *(a). M(5X+2Y). (b) D(5X+2Y).*

30. Durata de viață a becurilor electrice de un anumit tip este o variabilă aleatoare distribuită normal de parametrii $m=M(X)=457$ ore şi de dispersie $\sigma^2 = D(X) = 400$ ore. Să se calculeze probabilitatea ca un bec să se defecteze înainte de 490 ore.

31. Se aruncă un zar de 10^5 ori. Fie X variabila aleatoare care reprezintă numărul de apariții ale feței cu 5 puncte, Z v.a. care reprezintă *frecvența relativă* de apariție a feței cu 5 puncte iar $p = \frac{1}{6}$ probabilitatea de producere a acestui eveniment. Utilizând inegalitatea lui Cebâşev să se găsească limita inferioară a probabilității ca frecvența relativă să nu difere de probabilitate cu mai mult de 0,01.

32. Într-o cercetare ştiințifică se efectuează n experimente, urmărindu-se apariția unei anumite caracteristici. La fiecare probă caracteristica urmărită apare cu probabilitatea p=0,2. Să se determine numărul minim de experimente astfel încât, cu o probabilitate de cel puțin 0,92, frecvența relativă de apariție să difere în valoare absolută de probabilitatea p cu mai puțin de 10^{-3}.

33. O variabilă aleatoare X are distribuție binomială obținută prin efectuarea a n probe independente. Fie $Y=X/n$.

(a). Să se calculeze media şi dispersia variabilei X, pentru n=100, p=0,4.

(b). Pentru p=0,4 şi n=100 să se găsească o margine inferioară pentru probabilitatea $P(20 < X < 60)$.

(c). Pentru p=0,4 să se determine n astfel încât $P(|Y-0,4|<0,2) \geq 0,997$.

(d). Pentru p=0,4 şi n=100 să se determine $\varepsilon > 0$: $P(|Y-0,4|<\varepsilon) \geq 76$.

(e). Pentru n=100, să se determine p astfel încât $P(|Y - p| < 0,1) \geq$

34. O variabilă aleatoare cu distribuția normală X are valorile parametrilor m=10 şi dispersia σ=2.

(a). Să se calculeze probabilitatea ca variabile aleatoare să ia valori mai mici ca 5.

(b). Să se calculeze probabilitatea ca variabila aleatoare să ia valori între 5 şi 15.

(c). Să se determine un interval în care se găsesc valorile variabilei aleatoare cu o probabilitate de 0,75

(d). Să se determine o margine inferioară pentru probabilitatea ca variabila aleatoare să ia valori în intervalul (5,15).

35. O variabilă aleatoare cu distribuție normală are valorile parametrilor m=30 şi σ=10.

(a). Să se determine o margine inferioară pentru probabilitatea ca variabila aleatoare să ia valori în intervalul (10,50).

(b). Să se determine un interval în probabilitatea ca valorile variabilei să se găsească în acest interval să fie cel puțin 0,75.

CAPITOLUL 2 STATISTICĂ MATEMATICĂ

2.1. NOŢIUNI DE BAZĂ ALE STATISTICII MATEMATICE.

2.1.1. Populaţie statistică. Caracteristici.

Statistica matematica este principala aplicație a teoriei probabilităților. În esență, metodele statisticii constau în deducerea unor concluzii referitoare la colectivități mari pe baza cunoasterii unei părți restrânse restrânse a acesteia si extrapolarii rezultatelor.

Conceptele fundamentale ale statisticii matematice sunt cele de *populație statistică* şi *caracteristică.*

Prin *populație statistică* (sau *colectivitate statistică*) se înțelege totalitatea elementelor de aceeasi natură, ce sunt supuse studiului statistic, au o serie de trăsături comune şi sunt generate de acelasi complex de cauze.

Elementele unei populații statistice se numesc *indivizi* sau *unități statistice.* Numărul indivizilor unei populații statistice poartă numele de *volumul populației.*

Caracteristica statistică reprezintă trăsătura comună tuturor indivizilor unei populații statistice. Ea poate fi *cantitativă* dacă se poate exprima printr-un număr relativ la o unitate de măsură şi *calitativă* în caz contrar.

O caracteristică a unei populații statistice, care variază de la o unitate la alta, poartă numele de *variabilă statistică* sau *variabilă aleatoare.*

Variabilele statistice pot fi *discrete,* dacă mulțimea valorilor sale este finită sau cel mult numărabilă şi *continue*, dacă mulțimea valorilor sale umplu un interval.

Caracterizările numerice, cantitative obținute despre unitățile populației statistice cercetate mai poartă numele de *date statistice* iar conținutul specific (semnificația, mesajul) al lor poartă numele de

informație statistică.

Fie P o populație statistică şi X o caracteristică cantitativă a sa. O

astfel de caracteristică se supune unei anumite legi de rapartiție teoretice care nu se cunoaşte. De aceea este necesar studiul mulțimii

valorilor pa care le ia variabila aceasta.

Pentru un individ $i \in P$, notăm cu *X(i)* valoarea caracteristicii *X* atribuită individului "*i*" şi cu *X(P) mulțimea tuturor valorilor caracteristicii X* pentru toți indivizii populației *P*, adică $X(P) = \{X(i) : i \in P\}$.

Dacă volumul *n* al populației este finit studiul mulțimii $X(P)$ a valorilor pe care le ia caracteristica *X* se poate face prin observarea totală, adică prin enumerarea tuturor valorilor luate de *X* pentru toți indivizii populației statistice, caz în care avem:

$$X(P) = \{x_i = X(i) : i = 1,2,\ldots,n\} = \{x_1, x_2, \ldots, x_n\}.$$

Deseori este necesar ca datele statistice să nu fie tratate individual ci grupate pe anumite intervale.

Astfel dacă gruparea se va folosi ca metodă de sistematizare a datelor pentru calcularea indicatorilor derivați si aplicarea analizei statistice, este indicat să se folosească un număr optim de grupe (nu mai mic decât cinci).

Definiția 1. Să considerăm în mulțimea *X(P)* valorile notate x_{min}=min*X(P)*, x_{ma}=max*X(P)*. Numărul

$$A = x_{max} - x_{min} \qquad (1)$$

se numeşte *amplitudinea variației* .

Numărul *r* de intervale în care se împarte şirul datelor statistice se consideră a fi partea întreagă a numărului $1 + 3{,}322 \cdot \lg n$, unde *n* reprezintă volumul populației, adică $r = [1 + 3{,}322 \cdot \lg n]$.

Lungimea intervalelor de grupare (*h*) se calculează cu formula :

$$h = \frac{x_{max} - x_{min}}{r} \qquad \text{(Formula lui Sturges)} \qquad (2)$$

Dacă am stabilit numărul de intervale *r* în care am împărțit mulțimea *X(P)* şi am calculat mărimea intervalelor de grupare, intervalele de grupare se stabilesc pornind de la o valoare x_0 care poate fi x_{min} sau o valoare mai mică şi terminînd cu o valoare x_r care poate fi x_{max} sau o valoare mai mare, obținând şirul de intervale

$[x_0, x_1), \ldots, [x_{i-1}, x_i), \ldots, [x_{r-1}, x_r]$, cu $x_i = x_0 + (i-1)h, \forall i = 1,2,\ldots,r$.

Definiția 2. Mulțimea $X(P) \cap [x_{i-1}, x_i)$, $\forall i = 1,\ldots,r$ se numeşte *interval de grupare* sau *clasă de valori.*

Se numeşte *valoarea centrală a unei clase* sau *centrul clasei* $X(P)\in[x_i,x_{i+1}]$, numărul $x_i^{(c)}$ care este media aritmetică a extremităților acestei clase: $x_i^{(c)}=\frac{x_i+x_{i+1}}{2}$.

Distanța (h) între două limite de clasă sau între două centre de clasă se mai numeşte *interval de clasă* sau *mărimea clasei.*

Limitele intervalelor de grupare se stabilesc fără a fi suprapuse, astfel încât fiecare unitate să fie încadrată într-o singură clasă.

La intervalele cu variație continuă, limita superioară a fiecărui interval se repetă ca limită inferioară a intervalului următor. La intervalele cu variație discretă, limita inferioară este deplasată cu o unitate de măsură față de limita superioară a intervalului precedent.

Exemplul 1. Un sef de serviciu studiază munca a 30 de angajați în legătură cu timpul de muncă pierdut (min) într-o lună:

20	26	26	30	35	35	37	37	37	37
39	41	45	45	45	48	48	48	50	50
54	55	57	57	60	60	65	65	69	70

Să se facă gruparea acestor date statistice pe intervale de variație egale şi să se calculeze frecvențele absolute corespunzătoare.

Soluție : Deteremінăm mai întâi mărimea intervalului de grupare (h) cu formula (2). Pentru aceasta avem :

$r=[1+3{,}322\cdot\lg n]=[1+3{,}322\cdot\lg 30]=5$ şi prin urmare

$$h=\frac{A}{r}=\frac{X_{\max}-X_{\min}}{r}=\frac{70-20}{5}=\frac{50}{5}=10$$

Atunci obținem următoarele intervale declasă :

Grupe de salariați după timp (minute) (intervalul de clasă) $[x_i x_{i+1}]$	Numărul de salariați (frecvența absolută) n_i
[20,30)	3
[30,40)	8
[40,50)	7
[50,60)	6

[60,70]	6
Total	30

Limita inferioară este inclusă în interval (excepție face valoarea 70, care, fiind una singură nu influențează puternic distribuția şi astfel se va include în ultimul interval [60-70).

2.1.2. Repartiţii empirice şi de selecţie. Frecvențe.

Fie P o populație statistică, X o caracteristică cantitativă a sa (o variabilă*)* şi fie $X(P)=\{x_1,x_2,\ldots,x_n\}$ mulțimea valorilor sale, care nu sunt neapărat distincte două câte două. Pentru indivizi diferiți $i\neq j$ putem avea aceeaşi valoare a caracteristicii $x_i=x_j$. De aceea vom reține doar valorile caracteristicii distincte două câte două pe care le renumerotăm obținând mulțimea $\{x_1,x_2,\ldots,x_k\}$ iar pentru fiecare $i=1,2,\ldots,k$ notăm cu n_i numărul de repetiții a valorii x_i. Ca urmare putem întocmi tabloul:

$$X:\begin{pmatrix} x_1\ x_2\ldots x_i\ \ldots x_k \\ n_1\ n_2\ldots n_i\ \ldots n_k \end{pmatrix},\quad \sum_{i=1}^{k} n_i = n. \qquad (3).$$

Definiţia 3. Tabloul (1) se numeşte *repartiția empirică, serie statistică* sau *distribuție de frecvențe* a variabilei *X.*

Numărul n_i care reprezintă numărul indivizilor pentru care valoarea caracteristicii este egală cu x_i se numeşte *frecvența absolută* a valorii x_i.

Numărul $f_i=\dfrac{n_i}{n}$ se numeşte *frecvența relativă* a valorii x_i şi se exprimă prin fracție zecimală f_i=0,... sau în procente $(f_i)_{\%}=f_i\cdot 100$.

Este clar că $0\le f_i\le 1$ şi $\sum_{i=1}^{k} f_i = 1$. Atunci rapartiția empirică a

variabilei *X* se mai scrie: $X:\begin{pmatrix} x_1\ \ x_2\ldots x_i\ \ldots x_k \\ f_1\ \ f_2\ldots f_i\ \ldots f_k \end{pmatrix},\ \sum_{i=1}^{k} f_i = 1.$ (3')

Dacă *P* este o populație de volum mare eventual infinit, numărabil

sau nu, atunci nu se mai poate defini repartiția empirică în baza unei

observări totale. Din acest motiv se face o observare parțială, selectivă, a populației. Fie n numărul indivizilor aleşi la întâmplare din populația supusă observării. Dacă caracteristica X ia pentru indivizii selectați valorile $x_1,x_2,\ldots,x_i,\ldots,x_k$ diferite două câte două, cu frevențele absolute, respectiv, $n_1,n_2,\ldots,n_i,\ldots,n_k$ atunci se poate întocmi repartiția empirică

$$X^*: \begin{pmatrix} x_1 & x_2 \dots x_i & \dots x_k \\ n_1 & n_2 \dots n_i & \dots n_k \end{pmatrix}, \qquad \sum_{i=1}^{k} n_i = n. \tag{4}$$

Definiţia 4. Tabloul (4) se numeşte *repartiţia de selecţie* a variabilei X (s-a notat X^* pentru a deosebi repartiţia de selecţie de repartiţia empirică a variabilei X).

Dacă în locul frecvenţelor absolute se consideră frecvenţele relative, atunci repartiţia de selecţie se scrie

$$X^*: \begin{pmatrix} x_1 & x_2 \dots x_i & \dots x_k \\ f_1 & f_2 \dots f_i & \dots f_k \end{pmatrix}, \qquad \sum_{i=1}^{k} f_i = 1. \; . \tag{4'}$$

Tablourile (3') sau (4') ale repartiţiei empirice respectiv ale repartiţiei de selecţie prin forma lor amintesc de tabloul de distribuţie al unei variabile aleatoare discrete unde locul probabilităţilor este luat de frecvenţele relative.

Prin urmare este justificat a considera o variabilă statistică X a unei populaţii statistice ca fiind o variabilă aleatoare căreia îi putem aplica rezultatele obţinute pentru variabilele aleatoare.

Astfel, pentru o variabilă aleatoare X am definit *funcţia de repartiţie* ca fiind funcţia $F : \boldsymbol{R} \to [0,1]$, ce ataşează fiecărei valori x a variabilei X probabilitatea ca valorile variabilei să fie cel mult egale cu x adică : $F(x) = P(X < x)$.

În cazul variabilei aleatoare discrete funcţia de repartiţie se calculează cu formula $F(x) = \sum_{x_i < x} p_i$, unde $p_i = P(X = x_i)$.

Definiţia 5. Fie X o variabilă statistică discretă având repartiţia empirică:

$$X: \begin{pmatrix} x_1 & x_2 \dots x_i & \dots x_k \\ f_1 & f_2 \dots f_i & \dots f_k \end{pmatrix}, \qquad \sum_{i=1}^{k} f_i = 1$$

Se numeşte *funcţie de repartiţie* a repartiţiei empirice funcţia

$F_n : \boldsymbol{R} \to [0,1]$ definită prin : $F_n(x) = \sum_{x_i < x} f_i$. (5)

În mod asemănător definim *funcţia de repartiţie* a repartiţiei de selecţie.

Indicele n al funcţiei de repartiţie se foloseşte atât pentru a desemna volumul al populaţiei sau al selecţiei cât şi pentru a deosebi funcţia de repartiţie empirică $F_n(x)$ de funcţia de repartiţie teoreticăpentru care folosim notaţia $F(x)$.

Definiţia 6. Pentru un număr real a, numărul indivizilor din mulţimea $\{i \in P : X(i) \le a\}$ se numeşte *frecvenţă absolută cumulată ascendent (crescătoare)* a valorii a şi se notează $n_a \uparrow$.

Pentru un număr real a, numărul indivizilor din mulțimea $\{\{i \in P : X(i) \geq a\}$ se numeşte *frecvență absolută cumulată descendent (descrescătoare)* a valorii a şi se notează $n_a \downarrow$.

În mod analog, schimbând cuvântul "absolut" cu "relativ" se definesc *frecvențele relative cumulate crescătoare şi descrescătoare* notate $f_a \uparrow$ şi respectiv $f_a \downarrow$.

Ca şi pentru unitățile statistice, dacă datele statistice sunt împărțite pe clase de valori, se pot defini noțiunile de frecvență absolută şi frecvență relativă a unei clase de valori.

Definiția 7. Se numeşte *frecvența absolută* a clasei de valori $X(P) \cap [x_{i-1}, x_i]$, $\forall i = 1,\ldots,r$,ca fiind numărul unităților statistice din această clasă adică numărul elementelor mulțimii $\{i \in P : x_{i-1} \leq X(i) \leq x_i\}$.

Raportul dintre frecvența absolută a unei clase de valori şi volumul total n al populației se numeşte *frecvență relativă a clasei.*

În mod asemănător se definesc *frecvențele absolute şi relative cumulate ascendent sau descendent* pentru clase de valori.

2.1.3. Prezentarea datelor statistice. Grafice statistice.

Datele statistice se pot prezenta sub diferite forme:

- *tabele statistice*
- *serii statistice*
- *grafice statistice.*

Tabelele statistice constituie o modalitate de prezentare a datelor statistice şi sunt formate dintr-o rețea de linii, orizontale şi verticale în care sunt încadrate datele statistice.

Tabelele statistice se folosesc atât pentru prezentarea rezultatelor cercetării, ele permițând o bună vizualizare a acestora, cât mai ales, pentru efectuarea diverselor calcule în procesul de prelucrare a datelor. Ele sunt variate şi se folosesc în etapa culegerii datelor, în cursul prelucrării sau al analizei statistice şi pot conține una sau mai multe valori caracteristice.

Tabelele folosite în etapa culegerii datelor, în care se înscriu datele observate sau măsurate în ordinea obținerii lor se numesc *tabele originale.* Tabelele în care datele se ordonează după mărime şi în dreptul fiecăreia se trece numărul de repetiții

de aceleşi valori (frecvența absolută a valorii respective) se numesc *tabele primare de distribuție.*

Seria statistică, prezentată în *Definiția 3*, este tot o formă de prezentare a datelor statistice ce constă într-un tablou matriceal cu două linii, pe prima linie fiind înscrise datele şi pe linia a doua indicatori corespunzători ai datelor individuale (frecvențe absolute sau relative). Ele se folosesc numai în cazul studierii populațiilor cu o singură caracteristică.

Graficele statistice sunt imagini plane sau spațiale, cu caracter convențional şi care prin diferite mijloace plastice de prezentare scot în evidență ceea ce este caracteristic şi esențial pentru obiectul cercetării

în evoluția fenomenelor.

În cazul seriilor statistice cu o caracteristică cantitativă, se întâlnesc în mod curent următoarele reprezentări grafice:

1º <u>Reprezentarea în bare (batoane)</u> se utilizează în cazul seriilor statistice în care caracteristica ia un număr mic de valori şi valorile nu sunt grupate în clase.

Aceasta constă în a considera un sistem de axe rectangulare şi o unitate de măsură potrivită pe fiecare din axe. Pe axa absciselor se trec punctele corespunzătoare valorilor caracteristicii, iar din aceste puncte se ridică segmente verticale de lungime egală cu frecvența absolută sau relativă corespunzătoare.

<u>Exemplul 2.</u> Rezultatele obținute la un examen de 100 de studenți ai aceluiaş an de studiu reprezentate prin note la da 1 la 10 sunt următoarele: 3 note de 1, 4 note de 2, 6 note de 3, 9 note de 4, 12 note de 5, 14 note de 6, 22 note de 7, 18 note de 8, 7 note de 9, 5 note de 10.

a). Să se întocmească tabelul primar de distribuție ce se va completa cu frecvențele relative şi frecvențele absolute cumulate crescator şi descrescător.

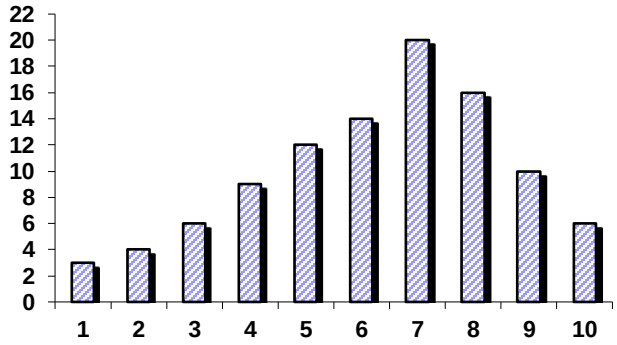
b). Să se reprezinte în bare seria statistică şi să se reprezinte poligoanele frecvențelor absolute cumulate crescator şi descrescător.

<u>Soluție:</u> ***a).*** Tabelul primar de distribuție completat cu frecvențele relative şi frecvențele absolute cumulate crescator şi descrescător este:

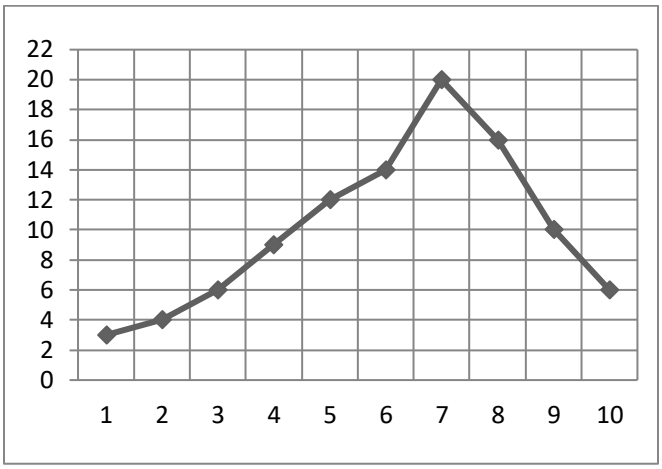
Nota „*i*”	Frecv. abs. n_i	Frecv. relativă f_i	Frecv. abs. cumulată crescător $n_i \uparrow$	Frecv. abs. cumulată descrescător $n_i \downarrow$
1	3	0,03	3	100
2	4	0,04	7	97
3	6	0,06	13	93
4	9	0,09	22	87
5	12	0,12	34	78

6	14	0,14	48	66
7	20	0,22	68	52
8	16	0,18	84	32
9	10	0,10	94	16
10	6	0,06	100	6
	n=100			

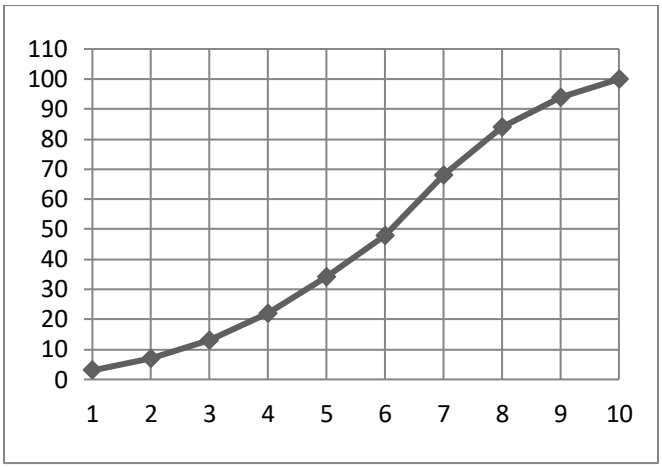
*b).*Reprezentarea în bare este :



Poligonul frecvențelor absolute este:



Poligonul frecvențelor absolute cumulate crescător este:



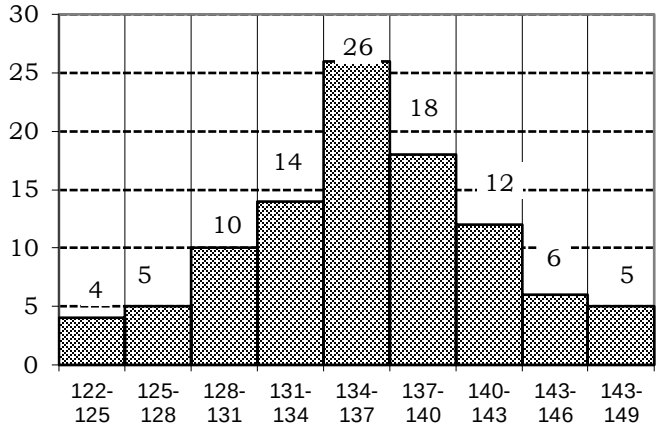
2º Histograma se utilizează atunci când valorile caracteristicii sunt grupate în clase şi are următoarea construcție: pe axa absciselor se iau segmente de lungimi egale între ele, de regulă egale cu amplitudinea claselor şi pe acestea, considerate ca baze, se ridică dreptunghiuri de înălțimi proporționale cu frecvențele absolute sau frecvențele relative ale claselor.

Deseori histogramele se reprezintă plastic sub diferite forme geometrice de spațiale (cilindrice, piramidale, etc).

Exemplul 3. Fie o caracteristică ale cărei valori sunt grupate în clase. Frecvența absolută pe grupe de valori este dată în tabelul:

Clasa de valori $[x_i, x_{i+1}]$	Frecvența absolută a clasei (n_i)	Frecvența absolută cumulată crescător
[122,125]	4	4
[125,128]	5	9
[128,131]	10	19
[131,134]	14	33
[134,137]	26	59
[137,140]	18	77
[140,143]	12	81
[143,146]	6	95
[146,149]	5	100

Soluție: Histograma acestei serii statistice este:



3o. <u>Poligonul frecvențelor</u> se utilizează atât în cazul seriilor cu

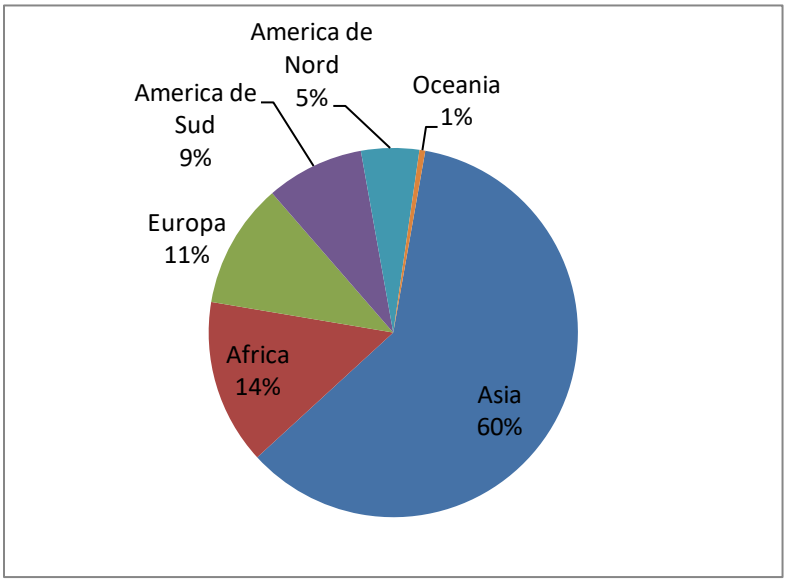
repartiții de frecvențe cât şi în cazul grupării datelor şi se construieşte astfel : În cazul seriilor cu repartiții de frecvență se unesc succesiv punctele din plan de coordonate (x_i, n_i) , unde n_i sunt frecvențele absolute ale valorilor individuale x_i. În cazul seriilor statistice cu datele grupate în clase de valori se unesc succesiv punctele din plan $\left(x_i^{(c)}, n_i\right)$, unde n_i sunt frecvențele absolute ale claselor de valori $[x_i, x_{i+1}]$ iar $x_i^{(c)}$ este valoarea centrală a clasei.

De asemenea se poate reprezenta şi *poligonul frecvențelor cumulate* utilizând în locul frecvențelor absolute (sau relative) ale valorilor individuale sau ale claselor, frecvențele cumulate.

4o. <u>Reprezentarea în sectoare circulare</u>. Pentru a obține rapid o viziune globală de o relativă importanță a diferitelor clase ale statisticii, se folosesc *graficele cu sectoare circulare* (*tip "plăcintă„)* interpretarea lor fiind uşurată dacă clasele reprezentare sunt haşurate sau colorate diferit. Aceste grafice se întocmesc astfel: pe un cerc se consideră sectoare circulare ale căror unghiuri la centru (arce de pe cerc) sunt proprționale cu frecvențele absolute sau relative ale claselor.

<u>Exemplul 4.</u> În anul 2007 populația lumii, repartizată pe continente, era de:

Continentul	Populația (mil. locuitori)	Frecvența relativă (%)
Asia	4.030	60,5
Africa	965	14,0
Europa	731	11,3
America de Sud	572	8,6
America de Nord	339	5,1
Oceania	34	0,5



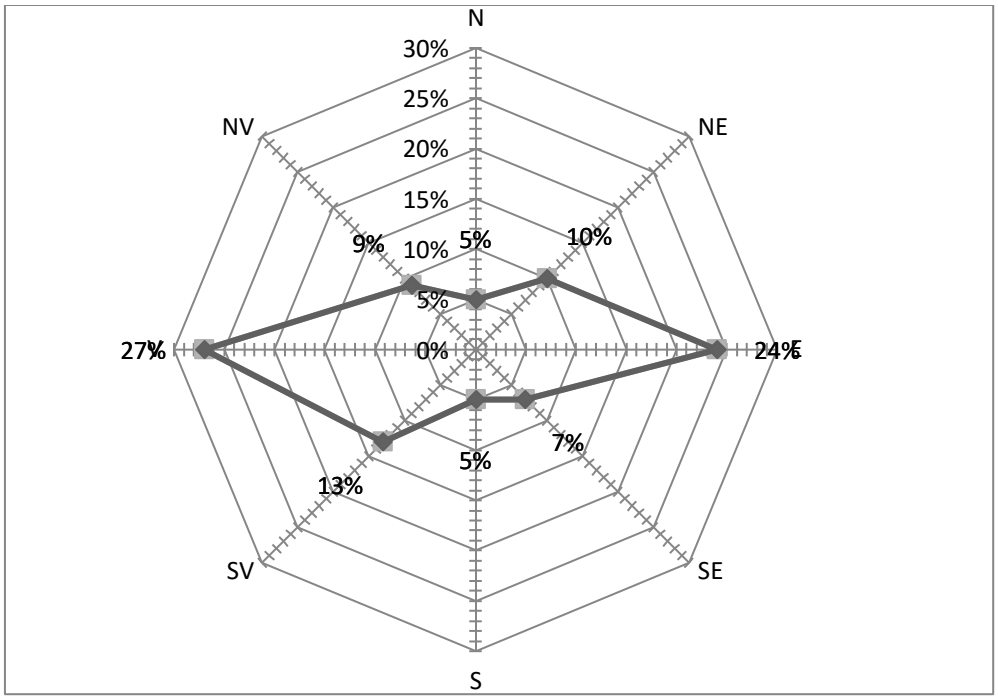
Reprezentarea tip "plăcintă" a seriei statistice de la Exemplul 3.

5º Reprezentarea polară. Când caracterul statistic prezintă o anumită periodicitate acesta se pune în evidență printr-o *reprezentare polară.* Aceasta se construieşte astfel: pe semidrepte de aceeaşi origine şi care împart planul într-un număr de sectoare egale (în funcție de caracterul seriei statistice), se consideră segmente, începând din origine, proporționale cu frecvențele absolute ale claselor şi se unesc

extremitățile acestor segmente.

Exemplul 5. Frecvența medie a vântului pe direcțiile principale şi secundare ale punctelor cardinale înregistrate la Stația meteorologică Craiova în perioada 1950-2000 este dată în tabelul următor

Direcția	N	NE	E	SE	S	SV	V	NV
Frecvența(%)	5	10	24	7	5	13	28	9



Frecvența direcției vântului la Craiova (reprezentarea polară de la exemplul 5)

2.2. INDICATORI STATISTICI

2.2.1. Indicatorii tendinţei centrale. Valori medii şi indicatori de poziţie.

O altă etapă a prelucrării datelor statistice, după întocmirea tabelului primar de distribuție şi reprezentarea grafică, constă în determinarea anumitor mărimi numerice a distribuțiilor de frecvențe numite *indicatori statistici.* Aceştia reprezintă expresii numerice ale caracteristicilor populațiilor şi eşantioanelor şi sunt*: indicatorii tendinței centrale (valorile medii şi indicatorii de poziție), indicatorii variației* şi *momentele.*

1º. Valorile medii.

Valorile medii sunt indicatori care caracterizeaza o populație sau un esantion din punctul de vedere al unei caracteristici studiate.

Definiţia 1. Fiind dată repartiția statistică a unei caracteristici

$$X:\begin{pmatrix} x_1\ x_2 \dots x_i\ \dots x_k \\ n_1\ n_2 \dots n_i \dots n_k \end{pmatrix}, \qquad \sum_{i=1}^{k} n_i = n\ ,$$

se numeşte *media aritmetică (ponderată) a variabilei X* sau *media de selecție* numărul:

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \dots + x_k n_k}{n_1 + n_2 + \dots + n_k} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i \text{ sau } \bar{x} = \sum_{i=1}^{k} x_i f_i \qquad (1)$$

În cazul când valorile variabilei sunt grupate în clase, calculul mediei se face cu ajutorul valorilor centrale ale claselor, şi anume:

$$\bar{x} = \sum_{i=1}^{k} \frac{x_i + x_{i+1}}{2} \cdot f_i \qquad (1')$$

Evident că, în cazul când valorile şirului de observaţii au frecvențele n_i=1, i=1.2,...,n, atunci k=n şi media se calculează cu formula:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1'')$$

şi se numeşte *media aritmetică simplă.*

În practică pentru media aritmetică se pot folosi unele artificii de calcul care sistematizează sau scurtează calculele.

Propoziţia 1. Fie seria statistică

$X:\begin{pmatrix} x_1\ x_2 \dots x_i\ \dots x_k \\ n_1\ n_2 \dots n_i \dots n_k \end{pmatrix}$, $\sum_{i=1}^{k} n_i = n$ şi fie a un număr nenul.

Atunci media seriei statistice se poate calcula cu formulele:

(a). Formula de calcul a mediei cu "zeroului fals":

$$\bar{x} = a + \frac{1}{n}\sum_{i=1}^{k}(x_i - a)\cdot n_i \qquad (2)$$

(b). Formula simplificată pentru seriile cu datele grupate:

$$\overline{x} = a + \frac{h}{n}\sum_{i=1}^{k}\left(\frac{x_i - a}{h}\right)\cdot n_i \,, \qquad (2')$$

unde h este mărimea intervalului de grupare

De regulă se ia pentru a valoarea cu frecvența cea mai mare.

Calculul mediei cu metoda "zeroului fals" este indicat a se face când frecvențele maxime sunt în centrul seriei.

Exemplul 1. Să se calculeze valorile centrale şi media variabielei X a cărei repartiție este dată de tabelul de mai jos (primele două coloane).

Soluție: Calculând valorile centrale ale claselor completăm tabelul cu valorile centrale $x_i^c = \frac{x_i + x_{i+1}}{2}$, cu frecvențele relative f_i şi termenii care intervin în formulele de calcul obținem:

Intervalul $[x_i, x_{i+1}]$	Frecv. abs. (n_i)	Frecv. relat. f_i	Valorile centrale x_i^c	$x_i^c - a$	$f_i(x_i - a)$	$x_i f_i$
122-125	4	0,04	123,5	-12	-0,48	4,940
125-128	5	0,05	126,5	-9	-0,45	6,325
128-131	10	0,1	129,5	-6	-0,60	12,950
131-134	14	0,14	132,5	-3	-0,42	18,550
134-137	26	0,26	135,5	0	0	35,230
137-140	18	0,18	138,5	3	0,54	24,930
140-143	12	0,12	141,5	6	0,72	16,980
143-146	6	0,06	144,5	9	0,54	8,670
146-149	5	0,05	147,5	12	0,60	7,375

Total	100				+0,45	135,95

Calculând media după definiție, însumând termenii ultilei coloane, obținem $\bar{x} = 135{,}95$.

Utilizând metoda zeroului fals şi alegând a=135,5 şi însumând termenii penultimei coloane obținem:

$$\bar{x} = a + \sum_{i=1}^{k}(x_i - a)\cdot f_i = 135{,}5 + 0{,}45 = 135{,}95 .$$

Definiţia 2. Fiind dată repartiția statistică a unei caracteristici

$$X : \begin{pmatrix} x_1\ x_2 \dots x_i\ \dots x_k \\ n_1\ n_2 \dots n_i \dots n_k \end{pmatrix}, \qquad \sum_{i=1}^{k} n_i = n\ ,$$

se numeşte *media pătratică a variabilei X,* rădăcinia pătrată din media aritmetică a pătratelor termenilor seriei, ca medie simplă sau ponderată:

$$\overline{x_p} = \sqrt{\frac{1}{n}\sum_{i=1}^{k} x_i^2 \cdot n_i}\ , \tag{3}$$

Pentru seriile simple, când n_i=1, i=1.2,…,n, atunci k=n iar media pătratică este :

$$\overline{x_p} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}\ , \tag{3'}$$

Observaţia 1. Definiția şi relațiile de calcul ale mediei pătratice conduc la câteva observații importante:

*(a)***.** Media pătratică se foloseste când dăm o importanță mare termenilor mari ai seriei sau în cazul în care termenii seriei au valori pozitive si negative.

(b). În mod frecvent media pătratică se utilizează pentru a caracteriza tendința centrală din ansamblul abaterilor valorilor individuale de la valoarea lor medie.

Între media aritmetică şi media pătratică avem relația: $\bar{x} \le \overline{x_p}$.

2o. Indicatori de poziție.

O mărime care dă informații asupra poziției valorilor principale ale repartiției se numeşte *indicator de poziție.*

Aceşti indicatori sunt: *modul, mediana şi cuantilele.*

Definiţia 3. Prin *mod* sau *valoarea modală* a unei distribuții statistice se înțelege valoarea caracteristicii căreia îi corespunde frecvența absolută (sau în mod egal, frecvența relativă) cea mai mare.

Dacă în tabloul de distribuție al seriei statistice valorile caracteristicii X sunt aşezate în ordine crescătoare, adică

$$X:\begin{pmatrix} x_1 \dots x_i \dots x_n \\ p_1 \dots p_i \dots p_n \end{pmatrix},\ x_1 < \dots < x_i < \dots < x_n,$$

atunci valoarea cea mai probabilă *Mo* este valoarea x_i care corespunde probabilității $p_i = P(X = x_i)$ care satisface dubla inegalitate

$$p_{i-1} \le p_i \ge p_{i+1}.$$

Dacă datele seriei statistice sunt grupate în clase de valori atunci

clasa căreia îi corespunde cea mai mare frecvență se numeşte *clasă*

modală.Dacă *X(P)* ∈[x_i, x_{i+1}] este o clasă modală atunci intervalul [x_i,x_{i+1}] se numeşte *interval modal*. Intervalul modal este acela pentru care frecvența cumulată este mai mare sau egală cu totalul frecvențelor împărțit la doi.

Observaţia 2. Există serii statistice fără mod (unde nici o valoare nu are frecvența maximă, toate valorile având aceeaşi frecvență), cu două sau mai multe valori modale: bimodale, trimodale, etc.

Calculul valorii modale M_0 se face în funcție de modul de prezentare a datelor .

- *Cazul seriei de distribuție de frecvență*. În acest caz modul M_0 este valoarea caracteristicii cu frecvența cea mai mare.
- *Cazul seriei statistice cu datele grupate pe intervale*. În acest caz modul M_0 se determină folosind următorul rezultat :

Propoziţia 2. Calculul valorii modale ale unei distribuții unimodale se bazează pe clasa modală şi pe clasele vecine acesteia şi se face cu ajutorul formulei:

$$M_O = X_0 + h \cdot \frac{\Delta_1}{\Delta_1 + \Delta_2}, \qquad (4)$$

unde: X_0 = limita inferioară a intervalului modal

$\Delta_1 = n_0 - n_{-1}$=diferența dintre frecvența intervalului modal (n_0)şi frecvența intervalului precedent (n_{-1}) ;

$\Delta_2 = n_0 - n_{+1}$= diferența dintre frecvența intervalului modal

(n_0) şi frecvența intervalului următor (n_{+1});

h = mărimea intervalului.

Cu valorile frecvențelor, formula (4) se mai scrie :

$$M_O = X_0 + h \cdot \frac{n_o - n_{-1}}{2n_o - n_{-1} - n_{+1}}, \qquad (4')$$

Definiţia 4. Se numeşte *mediana (valoarea centrală)* a unui şir de observații ordonate crescător sau descrescător, acea valoare din şirul respectiv care împarte şirul în două părți egale ca număr.

Propoziţia 3. Dacă $X : \begin{pmatrix} x_1 & x_2 \dots x_i & \dots x_k \\ f_1 & f_2 \dots f_i & \dots f_k \end{pmatrix}$, $\sum_{i=1}^{k} f_i = 1$ este

rapartiția empirică a unei caracteristici X şi dacă $F_n(x) = \sum_{x_i < x} f_i$ este

funcția de repartiție a acesteia atunci mediana este soluția ecuației:

$$F_n(Me) = \frac{1}{2} \text{ sau } \sum_{x_i < x} f_i = \frac{1}{2}. \qquad (5)$$

Calculul medianei se face în funcție de modul de prezentare a datelor:

- *Cazul seriei simple (formate din valori individuale)*

În acest caz mediana se găseşte la mijlocul şirului, având de o parte şi de alta un număr egal de observații.

La şirurile formate dintr-un număr impar de valori, $2k+1$, mediana este valoarea de rangul $k+1$.

La şirurile formate dintr-un număr par de valori individuale $2k$,

mediana se găseşte între cele două valori din mijlocul şirului şi anume

între valoarea de rang k şi valoarea de rang $k+1$ fiind considerată ca semisuma acestora.

- *Cazul seriei de distribuție de frecvență.*

În acest caz valoarea M_e va fi acea valoare a caracteristicii corespunzătoare primei frecvențe cumulate ascendent ce depăşeste valoarea lui $S_0 = \frac{\sum n_i + 1}{2}$.

- *Cazul seriei statistice cu datele grupate pe intervale* :

Propoziția 4. Dacă notăm cu M_e mediana unei distribuții statistice cu date grupate pe clase de valori atunci:

$$M_e = X_0 + h \cdot \frac{S_0 - \sum n_{pM_e}}{n_{M_e}}, \qquad (6)$$

unde: $S_0 = \frac{\sum n_i + 1}{2}$ dă intervalul median (locul lui M_e);

X_O – limita inferioară a intervalului median ;

h – mărimea intervalului ;

$\sum n_{pM_e}$ =suma frecvențelor cumulate precedente intervalului median;

n_{M_e} = frecvența absolută a intervalului median.

Exemplul 2. Să se calculeze modul şi mediana seriei statistice de mai jos, ce reprezintă distribuția loturilor de produse fabricate după numărul rebuturilor:

Nr. rebuturi	0	1	2	3	4	5
Nr. loturi	10	20	40	15	10	5

Soluție: Întocmim tabelul statistic cu frecvențele absolute şi cumulate crescător:

Nr. Rebuturi din lot	Nr. Loturi (n_i)	Frecvență cumulată $(n_i \uparrow)$
0	10	10
1	20	30
2	40	70
3	15	85
4	10	95
5	5	100
Total	100	

Este clar că numărul de rebuturi care se găsesc în cele mai multe loturi este 2 deci valoarea modală este $M_O=2$.

Valoarea M_e a medianei va fi acea valoare a caracteristicii corespunzătoare primei frecvențe cumulate crescător ce depăşeste valoarea lui $S_0 = \frac{\sum n_i + 1}{2}$. În cazul de mai sus avem : $S_0 = \frac{\sum n_i + 1}{2} = \frac{100 + 1}{2} = 50{,}5$. Prima frecvență mai mare ca S_0 este 70, deci M_e =2.

Exemplul 3. Să se calculeze valoarea modală şi media distribuției de frecvențe reprezentând gruparea după vechime a muncitorilor dintr-o secție :

Grupe	0-5	5-10	10-15	15-20	20-25	25-30	30-35
Nr. muncitori	5	7	10	12	18	15	7

Soluție : Întocmim tabelul primar :

Gruparea muncitorilor după vechime	Număr Muncitori (n_i)	Frecvențe Cumulate ($n_i \uparrow$)
0-5	5	5
5-10	7	12
10-15	10	22
15-20	12	34
20-25	**18**	**52**
25-30	15	67
30-35	7	74
Total	74	

Observăm că intervalul modal este intervalul $[20,25]$. Pentru acesta avem următoarele valori: X_0=20, $\Delta_1 = n_0 - n_{-1}$=18-12=6, $\Delta_2 = n_0 - n_{+1} = 18 - 15 = 3$, h=5.

Astfel, avem : $M_O = X_0 + h \cdot \frac{\Delta_1}{\Delta_1 + \Delta_2} = 20 + 5 \cdot \frac{6}{6+3} = 23{,}33$.

Folosind tabelul statistic cu frecvențele absolute şi cumulate crescător avem : $S_0 = \frac{\sum n_i + 1}{2} = \frac{74+1}{2} = 37{,}5$. Valoarea lui $S_0 = 37{,}5$ sea flă în intervalul $[20,25]$, care este intervalul median. Apoi avem: X_0=20 , h =5, $\sum n_{pM_e}$ = 34, n_{M_e} = 18 .

Deci $M_e = 20 + 5 \cdot \frac{37{,}5 - 34}{18} = 20{,}97$.

<u>Observaţia 3.</u> Mediana se caracterizează prin faptul că nu-şi schimbă valoarea numerică atunci când valorile care se găsesc deasupra sau dedesuptul ei se măresc sau se micşorează. Ea nu este influențată de valorile extreme ale şirului de observații spre deosebire de media aritmetică, care este sensibil influențată. Mediana depinde de mărimea termenului central, respectiv al celor doi termeni centrali şi ea este indicată la stabilirea mediei unei serii statistice cu clase deschise la capete.

O altă caracteristică a medianei este aceea că suma abaterilor observațiilor de la valoarea centrală este un minimum adică mai mică

decât suma abaterilor față de oricare altă valoare.

Mediana M_e şi valoarea modală M_0 se exprimă în aceeaşi unitate de măsură ca şi variabila studiată.

Mediana şi modulul, ca indicatori, nu sunt influențați de termenii seriei, deci nici de valorile aberante, în timp ce media aritmetică sintetizează influența tuturor termenilor.

<u>Relația dintre media aritmetică, valoarea modală şi mediană</u>.

Localizarea în cadrul seriei a valorii mediei aritmetice, a valorii modale şi a medianei conduce la informații despre forma de distribuire a unităților colectivității după caracteristica urmărită. Astfel:

- dacă există egalitatea $\bar{x} = M_0 = M_e$, atunci distribuția frecvențelor este simetrică;
- în cazul unei distribuții unimodale uşor asimetrice, frecvențele sunt uşor deplasate într-o parte sau alta, între cei trei indicatori ai tendinței centrale există următoarea relație, fără să se verifice cu regularitate: $\bar{x} - M_O = 3(\bar{x} - M_e)$.

Sunt cazuri când unul din cei trei indicatori ai tendinței centrale are o semnificație mai puternică.

O altă categorie de indicatori care descriu anumite poziții localizate în mod particular în cadrul seriilor statistice o reprezintă *cuantilele.*

Definiția 5. Fie X o variabilă aleatoare X şi $q \in \mathbf{N}$, $q \geq 2$. Se numesc *q-cuantile de ordinul k* numerele finite $(C_k)_{1 \leq k \leq q-1}$, astfel încât pentru orice $i = 1,2,\ldots,q-1$ avem:

$$(i).\ P(X \geq C_k) \geq \frac{q-k}{q}\ ;\ (ii).\ P(X \leq C_k) \geq \frac{k}{q}. \qquad (7)$$

Altfel spus, *cuantila de ordinul q* (unde $q \in \mathbf{N}$, $q \geq 2$ reprezintă numărul de părți în care a fost împărțită distribuția) şi este acea valore a variabilei aleatoare care marchează trecerea de la o parte la alta.

Propoziția 5. Dacă $X : \begin{pmatrix} x_1 & x_2 \ldots x_i & \ldots x_k \\ f_1 & f_2 \ldots f_i \ldots f_k \end{pmatrix}$, $\sum_{i=1}^{k} f_i = 1$ este rapartiția empirică a unei caracteristici X şi dacă $F_n(x) = \sum_{x_i < x} f_i$ este funcția de repartiție a acesteia atunci pentru orice $k = 1,2,\ldots,q-1$, *q-cuantilele de ordin k* sunt soluții ale ecuațiilor: $F_n(C_k) = \frac{k}{q}$ sau $\sum_{x_i < x} f_i = \frac{q}{k}$. (8) Unele q-cuantile au nume speciale. Dintre acestea cele mai importante sunt:

- *Mediana* este 2-cuantila (*q=2*), este una singură şi se notează cu *Me.* Mediana împarte seria în două părți egale delimitând cîte 50% din observații.
- *Cuartilele* sunt *4*-quantilele (*q=4*), sunt în număr de 3 şi se notează $(Q_k)_{k=1,2,3}$. Ele împart seria în 4 părți egale delimitând cîte 25% din observații. Cuartila de ordin k are proprietatea că un număr de k % dintre valorile secvenței sunt mai mici decât ea şi (4–k)% dintre valori sunt mai mari decât ea.
- *Decilele* sunt *10*-cuantilele (*q=10,* sunt în număr de 9 şi se notează $(D_k)_{k=1,2,\ldots,9}$. Ele împart seria în 10 părți egale delimitând cîte 10 % din observații. Decila de ordinul k are proprietatea că un număr de k % dintre valorile secvenței sunt mai mici decât ea şi (10–k)% dintre valori sunt mai mari decât ea.
- *Centilele (*sau *percentilele)* sunt *100*-cuantilele (*q=100),* sunt în număr de 99 şi se notează $(P_k)_{k=1,2,\ldots,99}$. Ele împart seria în 100 de părți egale delimitând cîte 1% din

observații. Percentila de ordinul k are proprietatea că un număr de k% dintre valorile secvenței sunt mai mici decât ea şi (100 – k)% dintre valori sunt mai mari decât ea.

Observaţia 4. Cuartilele au şi nume speciale şi concid cu anumite percentile. Astfel cuartila Q_1 se mai numeşte şi *cuartila inferioară* şi ea coincide cu percentila P_{25}. Cuartila Q_2 se numeşte *curtila de mijloc* şi ea coincide cu percentila P_{50}. Se observă faptul că cuartila Q_2 este tocmai mediana. Cuartila Q_3 se numeşte *cuartila superioară* şi ea coincide cu percentila P_{75}.

Cuartilele se folosesc pentru a analiza dispersia valorilor secvenței calculându-se cu ajutorul lor aşa-numitul interval *intercuartilic.* El este calculat ca diferența dintre percentila 75 şi percentila 25 . În cazul unei repartiții normale a datelor acest interval trebuie să fie aproximativ 1,35 din abaterea standard a datelor.

Cuartilele folosesc de asemenea la ajustarea termenilor unei serii de date statistice (adică înlocuirea termenilor reali cu termeni teoretici) sau la înlăturarea valorilor aberante (ca mărime în raport cu celelalte) întâlnite în procesul de culegere şi înregistrare a datelor statistice care pot să denatureze indicatorii de localizare centrală. Astfel într-una dintre modalități primele 2,5 % dintre valorile ordonate se înlocuiesc cu percentila $P_{2,5}$ iar ultimele 2,5 % dintre valorile se înlocuiesc cu percentila $P_{97,5}$.

Pentru ***calculul efectiv al q-cuantilelor*** în cazul unei serii statistice se procedează astfel:

q-cuantila de ordin k este valoarea $C_k = x_{j_k}$ din şirul de date ordonat crescător de indice $j_k = \left[\frac{k(n+1)}{q}\right]$, unde

- n reprezintă numărul total de date.
- k este ordinul cuantilei dorite.
- q reprezintă numărul grupurilor în care se împarte setul de date de către cuantilele luate în considerare.

Exemplul 4. Să se determine cuartilele pentru şirul de date: 6, 47, 49, 15, 42, 41, 7, 39, 43, 40, 36.

Soluție. Ordonând valorile în ordine crescătoare obținem:

6, 7, 15, 36, 39, 40, 41, 42, 43, 47, 49. Pentru calculul cuartilelor

$(Q_k)_{k=1,2,3}$. Determinăm indecşii $j_k = \left[\frac{k(n+1)}{q}\right]$, unde *n=11,q=4,*

k=1,2,3. Avem : $j_1 = [3] = 3$, $j_2 = [6] = 6$, $j_3 = [9] = 9$. Atunci cele 3

cuartile sunt : $Q_1 = x_3 = 15$ $Q_2 = x_6 = 40$ $Q_3 = x_9 = 43$

2.2.2. Indicatorii variaţiei. Momente.

Indicatorii tendinței centrale nu dau nici o explicație asupra împrăstierii, respectiv a modului în care termenii seriei se abat între ei sau de la medie.

Aceşti indicatori care dau o caracterizare precisă a unei serii statistice prin care se poate cunoaşte *variația* valorilor individuale (cum se grupează aceste valori în jurul valorii medii, dacă sunt apropiate sau îndepărtate de această valoare), se numesc *indicatorii variației.* Ei sunt de două feluri: *indicatorii simpli ai variației* şi *indicatorii sintetici ai variației.*

1º. Indicatorii simpli ai variaţiei.

Aceşti indicatori se caracterizează prin faptul că se calculează în cifre absolute sau relative, prin compararea valorilor individuale extreme, sau prin compararea fiecărei valori individuale cu valoarea lor medie.

Definiţia 6. Fie dată repartiția statistică a unei caracteristici

$$X : \begin{pmatrix} x_1\ x_2 \dots x_i\ \dots x_k \\ n_1\ n_2 \dots n_i \dots n_k \end{pmatrix}, \qquad \sum_{i=1}^{k} n_i = n$$

Numim *abatere individuală (ecart)* a valorii x_i, o valoare e_i care arată cu câte unități de măsură, sau de câte ori valoarea individuală a caracteristicii este mai mare sau mai mică decât mărimea unui indicator al tendinței centrale(de exemplu media $\bar{x}$)

Abaterile individuale se calculează în cifre absolute sau relative:

Abaterile individuale absolute (e_i):

$$e_i = x_i - \bar{x}, \quad \text{pentru } i = 1,2,\dots,k \qquad (9)$$

Abaterile individuale relative ($e_{i\%}$)

$$e_i\% = \frac{e_i}{\overline{x}} \cdot 100 \text{ , pentru } i = 1,2,\ldots,k \qquad (9')$$

Indicatorii simpli ai variației permit o caracterizare parțială si aproximativă a variației deoarece se calculează pe baza relației între doi

termeni ai seriei, sau între fiecare termen si media lor.

2º. Indicatorii sintetici ai variației.

Spre deosebire de indicatorii simpli, indicatorii sintetici sintetizează într-o singură expresie numerică variația valorilor individuale față de tentința centrală a caracteristicilor urmărite, într-o populație statistică.

Principalii indicatori sintetici cu care se caractrerizează împrăştierea (variația) termenilor seriei față de tendința lor centrală sunt: *abaterea medie adsolută, dispersia. abaterea medie pătratică (sau abaterea standard) şi coeficientul de variație.*

Definiţia 7. Se numeşte *abatere medie absolută (deviație mediei)* a repartiției

$$X : \begin{pmatrix} x_1\ x_2 \ldots x_i\ \ldots x_k \\ n_1\ n_2 \ldots n_i \ldots n_k \end{pmatrix}, \qquad \sum_{i=1}^{k} n_i = n ,$$

numărul notat $\overline{e_x}$ reprezentând media valorilor absolute ale abaterilor individuale ale termenilor seriei față de valoarea medie $\overline{x}$, adică :

$$\overline{e_x} = \frac{1}{n} \sum_{i=1}^{k} |x_i - \overline{x}| \cdot n_i \quad \text{sau} \quad \overline{e_x} = \sum_{i=1}^{k} f_i |x_i - \overline{x}| \qquad (10)$$

Evident că, în cazul când valorile şirului de observații au frecvențele n_i=1, i=1.2,...,k, atunci abaterea mediei se calculează cu formula:

$$\overline{e_x} = \frac{1}{n} \sum_{i=1}^{k} |x_i - \overline{x}| \qquad (10')$$

Pentru seriile de distribuție cu datele grupate pe intervale se consideră drept x_i centrele de interval.

Definiţia 8. Se numeşte *dispersie* a unei distribuții statistice

$$X : \begin{pmatrix} x_1\ x_2 \ldots x_i\ \ldots x_k \\ n_1\ n_2 \ldots n_i \ldots n_k \end{pmatrix}, \quad \sum_{i=1}^{k} n_i = n ,$$

numărul notat $\sigma_{\bar{x}}^2$ reprezentând media pătratelor abaterilor individuale ale termenilor seriei față de valoarea medie $\bar{x}$, adică:

$$\sigma_{\bar{x}}^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i \qquad (11)$$

Evident că, în cazul când valorile şirului de observaţii au frecvențele n_i=1, i=1.2,...,k, atunci k=n şi dispersia se calculează cu formula:

$$\sigma_{\bar{x}}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (11')$$

Observaţia 5. Dispersia este indicele de variație cel mai sigur şi ea dă cele mai bune indicații privind dispersia (împrăştierea) valorilor individuale în jurul valorii medii, ea măsurând gradul de împrăştiere a valorilor selecției în jurul mediei de selecție (centrul de grupare). Cu cât este mai mică dispersia, cu atât valorile seriei statistice se grupează mai mult în jurul valorii medii. O dispersie mare arată că elementele eşantionului au o împrăştiere mare.

Definiţia 9. Se numeşte *abatere medie pătratică* sau *abatere standard* numărul dat de rădăcina pătrată a dispersiei:

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i}\ , \qquad (12)$$

respectiv, în cazul când n_i=1, i=1.2,...,k=n:

$$\sigma_{\bar{x}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\ , \qquad (12')$$

Propoziţia 6. Fie seria statistică de distribuție $X:\begin{pmatrix} x_1\ x_2 \dots x_i\ \dots x_k \\ n_1\ n_2 \dots n_i \dots n_k \end{pmatrix}$, $\sum_{i=1}^{k} n_i = n$,şi $\bar{x} = \frac{1}{n}\cdot\sum_{i=1}^{k} x_i \cdot n_i$, $\overline{x^2} = \frac{1}{n}\cdot\sum_{i=1}^{k} x_i^2 \cdot n_i$.

Atunci dispersia şi abaterea medie pătratică se pot calcula cu formulele:

$$\sigma_x^2 = \overline{x^2} - \bar{x}^2 \qquad (13)$$

$$\sigma_x = \sqrt{\overline{x^2} - \bar{x}^2}\ , \qquad (13')$$

Demonstrație. Într-adevăr, avem:

$$\sigma_{\bar{x}}^2 = \sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i = \sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot f_i = \sum_{i=1}^{k} x_i^2 f_i - 2\bar{x}\sum_{i=1}^{k} x_i f_i + \bar{x}^2 \sum_{i=1}^{k} f_i =$$

$$= \overline{x^2} - 2\bar{x}\cdot\bar{x} + \bar{x}^2 = \overline{x^2} - \bar{x}^2 .$$

Trecând, ca în calculul mediei la alegerea unui punct a potrivit, calculul dispersiei se simplifică prin formulele:

$$\sigma_{\bar{x}}^2 = \left(\sum_{i=1}^{k}(x_i - a)^2 \cdot f_i\right) - \left(\sum_{i=1}^{k} x_i f_i - a\right)^2 = [\overline{(x-a)^2} - (\bar{x} - a)^2] .$$

Observaţia 6. Pentru seriile statistice grupate în clase se procedează ca şi în cazul mediei aritmetice; x_i reprezintă atunci valoarea centrală a clasei $\left(x_i^{(c)}\right)$.

Observaţia 7. În cazul în care se utilizează eşantioane de volum redus aceşti indicatori şi poartă numele de *dispersia modificată* (notată $s_{\bar{x}}^2$) respectiv *abaterea medie pătratică modificată* (notată $s_{\bar{x}}$) şi se determină prin relațiile următoare:

- Pentru dispersie:

$$s_{\bar{x}}^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i \qquad (11'')$$

- Pentru abaterea medie pătratică:

$$s_{\bar{x}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i}\ , \qquad (12'')$$

sau, cu formulele simplificate:

$$s_{\bar{x}}^2 = \frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right) \qquad (13')$$

$$s_{\bar{x}} = \sqrt{\frac{n}{n-1}\left(\overline{x^2} - \bar{x}^2\right)} \qquad (13'')$$

Observaţia 8. Comparând abaterea medie absolută cu abaterea medie pătratică, calculate pentru aceeaşi serie, se constată că:

$$\overline{e_x} \le \sigma_x \text{ sau } \overline{e_x} \approx \frac{4}{5}\sigma_x .$$

Observaţia 9. Deseori în analiza statistică se apelează la *valorile individuale standardizate*. Valorile (datele) numerice standardizate sunt valori inițiale (înregistrate) transformate cu ajutorul medie şi abaterii lor medii pătratice. Deci, prin operația de standardizare fiecare valoare x_i, $(i = 1,..., n)$ se substituie prin $x_i^{(s)}, (i = 1,\ldots,n)$ unde:

$$x_i^{(s)} = \frac{x_i - \bar{x}}{s},\ (i = 1,\ldots,n) \qquad (14)$$

Avantajele principale ale utilizării valorilor standardizate se rezumă la următoarele:

-Elimină unitatea de măsură a variabilei studiate;

-Media lor aritmetică este egală cu zero ($\overline{x^{(s)}} = 0$);

-Dispersia lor este constantă şi egală cu unu $\sigma^2_{\bar{x}^{(s)}} = 1$.

Definiţia 10. Se numeşte *coeficient de omogenitate (coeficient de variaţie)* numărul definit prin raportul, exprimat în procente, între abaterea standard $\sigma_{\bar{x}}$ şi media aritmetică $\bar{x}$ a unei distribuţii

statistice: $(CV_x)_{\%} = \frac{\sigma_{\bar{x}}}{\bar{x}} \cdot 100$

(15)

Observaţia 10. Coeficientul de variaţie este o măsură a dispersiei relative care descrie abaterea medie pătratică ca procent din media aritmetică valorile sale fiind situate în intervalul [0,100]. Cu cât valorile sale sunt mai apropiate de zero, cu atât seria este mai omogenă (media este mai reprezentativă); cu cât valorile sale sunt mai apropiate de 100 cu atât ansamblul valorilor individuale observate este mai eterogen (împrăştierea este mai mare, iar media calculată este mai puţin reprezentativă). Din punct de vedere practic pragul de trecere de la starea de omogenitate la cea de eterogenitate este nivelul de 35%

pentru coeficientul de variaţie. Astfel,

- Dacă $(CV_x)_{\%} \le 35\%$, populaţia este *omogenă*;
- dacă $(CV_x)_{\%} > 35\%$, populaţia este *eterogenă*.

3o. Momente

Definiţia 11. Fiind dată distribuţia statistică

$$X: \begin{pmatrix} x_1\ x_2 \ldots x_i\ \ldots x_k \\ n_1\ n_2 \ldots n_i \ldots n_k \end{pmatrix}, \qquad \sum_{i=1}^{k} n_i = n\ ,$$

se definesc *momentele* distribuţiei, ca *medii aritmetice al abaterilor de la un anumit punct ales ca origine, ridicate la diferite puteri.*

În funcţie de punctul ales ca origine distingem:

(a). Momentele inițiale, notate cu M_r , sunt momentele pentru care punctul de origine al abaterilor este 0. Formulele pentru calculul momentelor inițiale (intervalul de clasă fiind h=1) :

Momentul inițial de ordinul r:

$$M_r = \frac{1}{n}\sum_{i=1}^{k} x_i^r n_i = \sum_{i=1}^{k} x_i^r f_i \qquad (16)$$

În particular,

Momentul de ordinul 1 :

$$M_1 = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i f_i \text{ , (media aritmetică)} \qquad (16')$$

Momentul de ordinul 2:

$$M_2 = \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i = \sum_{i=1}^{k} x_i^2 f_i \text{ .} \qquad (16'')$$

(b). Momentele centrate, notate cu m_r, sunt momentele pentru care punctul de origine al abaterilor este media $\bar{x}$ a distribuției. Formulele pentru calculul momentelor inițiale :

Momentul centrat de ordinul r:

$$m_r = \frac{1}{n}\cdot \sum_{i=1}^{k} (x_i - \bar{x})^r n_i = \sum_{i=1}^{k} (x_i - \bar{x})^r f_i \text{ ,} \qquad (17)$$

În particular, momentele centrate de ordinul 1 şi 2 sunt*:*

$$m_1 = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x}) n_i = \sum_{i=1}^{k} (x_i - \bar{x}) f_i \text{ ,} \qquad (17')$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{k} (x_i - \bar{x})^2 n_i = \sum_{i=1}^{k} (x_i - \bar{x})^2 f_i = \sigma^2 \text{ .} \qquad (17'')$$

Momentele servesc de asemenea la calculul asimetriei şi excesului.

3. Indicii asimetriei. Indicii excesului.

1º . Asimetria.

În urma prelucrării informațiilor se obțin serii de repartiție de frecvență empirice, ce se pot compara cu repartiții teoretice, a căror formă de repartiție este cunoscută. Cea mai frecventă serie de repartiție, către care tind seriile empirice, este distribuția normală, ale cărei frecvențe se distribuie simetric, de o parte şi de alta a frecvenței maxime, plasată în centrul seriei.

O distribuție este *simetrică* dacă observațiile înregistrate sunt egal dispersate de o parte şi alta a valorii lor centrale. Într-o distribuție simetrică cele trei valori cu care se exprimă tendința centrală, valoarea modală (M_O), mediana (M_e) şi media ($\bar{x}$), se confundă. Graficul acestei distribuții are formă de clopot (Clopotul lui Gauss), în raport cu ordonata maximă.

O distribuție *asimetrică (sau oblică)* se caracterizează prin faptul că frecvențele valorilor caracteristicii urmărite sunt deplasate mai mult sau mai puțin, într-o parte şi alta față de tendința centrală (exprimată prin: M_O, M_e sau $\bar{x}$.

Indicele de asimetrie (de oblicitate) ne arata în ce masura media se îndeparteaza de mediana, si implicit, în ce masura curba de istributie normala a datelor se departeaza de mijloc, deplasându-se spre stânga sau spre dreapta. Sunt considerate distributii relativ normale cazurile în care acesti indicatori nu depasesc ±1,96.

Amploarea asimetriei statistice unimodale se caracterizează sintetic cu ajutorul unor coeficienți adimensionali: *Coeficientul de asimetrie a lui Pearson* şi *Coeficientul de asimetrie a lui Johannsen.*

Definiţia 12. Fie X o serie statistică de distribuție

$$X:\begin{pmatrix} x_1\ x_2 \dots x_i\ \dots x_k \\ n_1\ n_2 \dots n_i \dots n_k \end{pmatrix}, \qquad \sum_{i=1}^{k} n_i = n\ ,$$

de medie $\bar{x}$, abatere standard *s* şi valoare modală M_O.

(a). Se numeşte *asimetrie absolută* numărul

$$A_s = \bar{x} - M_O \tag{18}$$

Se numeşte *coeficient (indice) de asimetrie al lui Pearson* numărul notat C_{as}^P care este raportul între asimetria absolută şi abaterea medie pătratică, adică:

$$C_{as}^P = \frac{\bar{x} - M_O}{s} \tag{19}$$

(b). Se numeşte *coeficient (indice) de asimetrie al lui Johannsen* numărul notat C_{as}^J definit de raportul dintre momentul centrat de ordinul 3 şi cubul abaterii standard:

$$C_{as}^J = \frac{m_3}{s^3}\ , \tag{19'}$$

Observaţia 11. Coeficientul de asimetrie al lui Johannsen, deşi are o expresie mai greoaie, este coeficientul cel mai frecvent utilizat în studierea simetriei unei distribuții.

Oricare ar fi expresia coeficientului de asimetrie, notat acum C_{as}, el are o valoare abstractă, arătând mărimea si felul asimetriei, iar valorile lui sunt cuprinse în intervalul (-1, 1).

- Dacă C_{as} = 0, deci $\bar{x} = M_O$, seria este simetrică;

- Dacă $C_{as} \to 0$, seria prezintă o asimetrie mică;
- Dacă $C_{as} \to (\pm 1)$, seria prezintă o asimetrie pronunțată;
- Dacă $C_{as} \in (0,1)$,deci $\bar{x} > M_O$,asimetria este pozitivă(spre stânga)
- Dacă $C_{as} \in (-1,0)$, $\bar{x} < M_O$ asimetria este negativă(spre dreapta)

Observaţia 12. Asimetria distribuțiilor unităților într-o populație după caracteristica urmărită poate fi vizibilă pe reprezentările grafice (histograma, poligonul frecvențelor efective) empirice comparate cu

alura clopotului lui Gauss.

Graficele de mai jos ilustrează cele două cazuri de asimetrii :

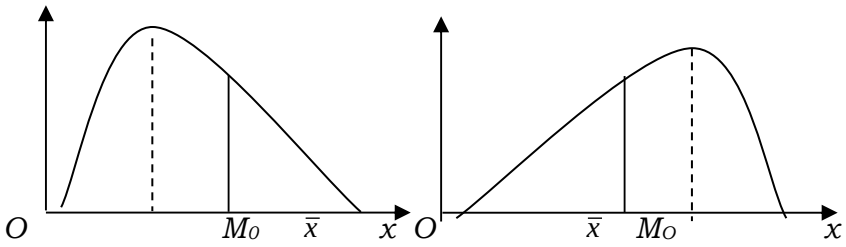


Fig. 1 *a). Asimetrie stângă* *b).Asimetrie dreaptă*

Exemplul 5. Să se calculeze coeficientul de asimetrie (Cas) al caracteristicii reprezentând gruparea după vechime a angajaților de la o firmă şi să se interpreteze rezultatul:

Gruparea sal.după vechime	5-10	10-15	15-20	20-25	25-30	Total
Număr salariați	5	7	15	10	8	45

Soluție. Întocmim tabelul :

Gruparea salariaților după vechime	Nr.sal. n_i	x_i	$\frac{x_i - a}{h}$	$\left(\frac{x_i - a}{h}\right) \cdot n_i$	$\left(\frac{x_i - a}{h}\right)^2 \cdot n_i$
5-10	6	7,5	-2	-12	24
10-15	9	12,5	-1	-9	9
15-20	17	17,5	0	0	0
20-25	10	22,5	1	10	10
25-30	8	27,5	2	16	32
TOTAL	50			5	75

Pentru medie şi dispersie folosim formulele simplificate luând *h=5* şi *a=17,5*:

$$\bar{x} = a + \frac{h}{n}\sum_{i=1}^{k}\left(\frac{x_i - a}{h}\right) \cdot n_i = 17{,}5 + \frac{5}{50} \cdot 5 = 18 .$$

$$\sigma^2_{\bar{x}} = \frac{h^2}{n}\sum_{i=1}^{k}\left(\frac{x_i - a}{h}\right)^2 \cdot n_i - (\bar{x} - a)^2 = \frac{25}{50} \cdot 75 - 0{,}5^2 = 37{,}25 ,$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{37{,}25} = 6{,}10 .$$

Pentru calculul valorii modale folosim formula: $M_O = X_0 + h \cdot \frac{\Delta_1}{\Delta_1 + \Delta_2}$,

unde:

X_0 =15, limita inferioară a intervalului modal

$\Delta_1 = n_0 - n_{-1}$=17-9=8; $\Delta_2 = n_0 - n_{+1}$= 17-10=7; h = 5.

Astfel obținem : $M_O = X_0 + h \cdot \frac{\Delta_1}{\Delta_1 + \Delta_2} = 15 + 5 \cdot \frac{8}{8+7} = 15 + 2{,}67 = 17{,}67.$

Atunci, coeficientul de asimetrie este :

$$C_{as} = \frac{\bar{x} - M_O}{\sigma} = \frac{18 - 17{,}67}{6{,}10} = 0{,}0541 \in (0,1).$$

*S*eria este usor asimetrică pozitivă.

2⁰. Boltirea .

O altă caracteristică a formei curbei de distribuție este *boltirea* sau *aplatizarea.* Există adesea curbe care se abat în ceea ce priveşte boltirea de la curba distribuției normale ("clopotul lui Gauss"), unele fiind mai ascuțite iar altele mai turtite. Această caracteristică se numeşte *exces* al curbei.

Definiţia 13. Fie X o serie statistică de distribuție

$$X:\begin{pmatrix} x_1\ x_2 \dots x_i\ \dots x_k \\ n_1\ n_2 \dots n_i \dots n_k \end{pmatrix}, \qquad \sum_{i=1}^{k} n_i = n,$$

a cărei medie este $\bar{x}$, abatere standard σ şi are momentul centrat de ordinul 4, m_4.

Se numeşte *exces* (sau *coeficient de aplatizare al lui Fischer)* numărul

$$E = \frac{m_4}{\sigma^4} - 3 \tag{20}$$

Observaţia 13. Calculul şi interpretarea coeficienților de aplatizare

prezentați trebuie completat cu analiza graficului distribuției

empirice comparativ cu cel al distribuției normale.

Pentru repartiția normală excesul este E=0. Dacă E>0 atunci curba de distribuție prezintă un exces pozitiv (este mai ascuțită) iar dacă E<0, curba prezintă un exces negativ (este mai turtită) aşa cum se vede din

graficele de mai jos:

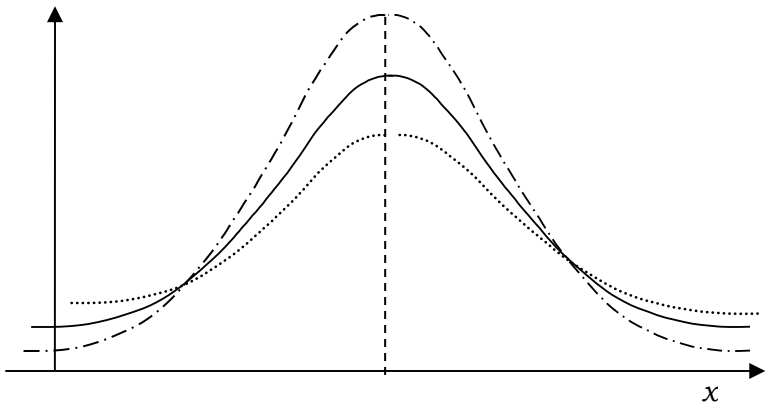


Fig. 2 *Distribuții cu exces pozitiv* ($E>0$) şi cu exces negativ ($E<0$).

Analiza asimetriei şi aplatizării are sens numai în cazul distribuțiilor empirice unidimensionale care prezintă o singură valoare modală.

2.3. CORELAŢIE ŞI REGRESIE

2.3.1. Serii statistice cu două dimensiuni.

În cercetările agricole şi biologice, ca de altfel în toate domeniile de activitate, există o interdependență între fenomene. Apariția şi evoluția

unui fenomen este în strânsă legătură cu o serie de alte fenomene sau factori care intervin în determinarea sau favorizarea acestuia.

În general, deosebim două tipuri de legături: *legături funcționale*

de tip determinist şi *legături statistice* sau *stohastice* (*întâmplătoare).*

Legăturile funcționale exprimă legătura de la cauză la efect între fenomene. Asemenea legături sunt studiate în cadrul ştiințelor exacte, unde având de-a face cu fenomene simple, legătura de la cauză la efect se evidențiază mai uşor şi se exprimă sub formă de lege. În cazul unei legături funcționale unei valori determinate a unei variabile independente X (argument) îi corespunde strict o valoare a variabilei dependente Y (funcție).

Legăturile statistice sunt mai puțin perfecte, se evidențiază mai

greu, exprimând legătura de dependență care există între fenomene.

În cazul corelației statistice fi ecărei valori numerice a variabilei X corespund nu una ci mai multe valori a variabilei Y.Dependența de acest tip are caracter întâmplător şi se numeşte *dependență stochastică.*

Legăturile statistice se pot clasifica în funcție de mai multe criterii:

1. După tipul variabilelor luate în considerație legăturile pot fi

clasificate în*:*

- *corelații statistice* – când legătura se stabileşte între variabile calitative;
- *asocieri statistice* - când în legătură intră cel puțin o variabilă calitativă, nenumerică.

2. După sensul legăturilor dintre variabile, putem avea:

- *legături directe* - pe măsură ce creşte variabila factorială creşte şi cea rezultativă.
- *legături inverse* - pe măsură ce creşte variabila factorială descreşte cea rezultativă.

3. După forma ecuației menită să descrie relația dintre variabile

(adică modelul matematic propriu dependenței studiate) putem avea

- *legături liniare*-dacă dependența variabilei efect *Y* față variabila cauzală *X* este de tip liniar, exprimată printr-o funcție de tipul *y=ax+b.*

- *legături neliniare,* care pot fi exprimate cu ajutorul funcțiilor neliniare (de tip parabolic, hiperbolic, exponențial, etc.)

Studiul dependenței stochastice dintre variabilele aleatoare constă în două aspecte: a*naliza de corelație* şi *analiza de regresie.*

- *Analiza de corelație* ne arată gradul în care o variabilă este dependentă de altă variabilă şi dă măsura dependenței dintre mărimile variabilelor considerate, caracterizată prin *coeficientul de corelație* sau prin *raportul de corelație.* Analiza corelației este specifică variabilelor cantitative, numerice.
- *Analiza de regresie* ne arată cum una dintre variabile este dependentă de altă variabilă, permite previzionarea sa şi constă în determinarea *funcției de regresie* între variabila factorială şi variabila dependentă.

2.3.2. Analiza corelaţiilor. Coeficient de corelaţie

Prin *corelație simplă* se înțelege legătura reciprocă dintre două variabile *X* şi *Y* ale unei populații.

Corelațiile dintre variabile prezintă mare importanță, deoarece cunoscând variația unei însuşiri putem trage concluzii asupra însuşirii sau însuşirilor de care aceasta este legată, fără a recurge la determinări directe.

Corelația poate fi *pozitivă*, atunci când valorile celor două variabilecresc sau descresc în acelaşi timp, sau *negativă,* atunci când valorile unei variabile cresc, cele ale celeilalte variabile descresc.

Metodele cele mai simple de constatare a unei corelații sunt *me-toda grafică* sau *graficul de corelație (corelograma)* şi *tabela de corelație.*

Diagrama de împrăştiere sau *corelograma* indică, în sistemul de coordonate rectangulare, fiecare unitate statistică (fiecare caz individual) printr-un punct

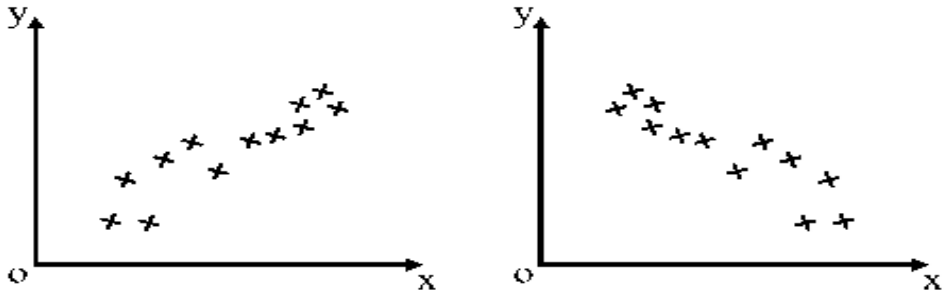


Fig. 1. Corelograme indicând corelatii pozitive (a) sau negative (b)

Forma de distribuire a punctelor pe grafic are aspectul unui "*nor de puncte*" şi ne arată dacă între cele două variabile există o relație. Dacă punctele respective se concentrează în jurul unei anumite curbe (curbă care poate fi liniară sau de altă formă) atunci există posibilitatea stabilirii existenței, a direcției, a formei si intensității legăturilor dintre cele două variabile.

Metoda grafica este utilizată cu bune rezultate pentru alegerea funcției analitice care se studiază (în cazul regresiei şi corelației).

- *Tabelul de corelație* se utilizează în cazul grupării combinate după două variabile numerice.

Fie X şi Y două caracteristici cantitative ale unei populații statistice pentru care se determină valorile distincte $x_1, x_2, \ldots, x_r$ ale lui X respectiv $y_1, y_2, \ldots, y_m$ ale lui Y. Notăm cu n_{ij} *frecvențele absolute* ale cazurilor pentru care $X = x_i$ $(i \in \{1,2,\ldots,r\})$ şi $Y = y_j$ $(j \in \{1,2,\ldots,m\})$.

Dacă n reprezintă numărul unităților din populația statistică la care s-au observat variabilele X şi Y atunci $n = \sum_{i=1}^{r} \sum_{j=1}^{m} n_{ij}$.

Frecvențele relative se definesc prin rapoartele: $f_{ij} = \frac{n_{ij}}{n}$.

Astfel, *tabelul de corelație* numit şi *tabel de contigență* este un tabel

cu dublă intrare, frecvențele absolute sau relative pot fi cuprinse în el similar unei matrice de dimensiune (r, m)

	Valorile caracteristicii dependente Y						Total
	y_1	y_2	$\cdots$	y_j	$\cdots$	y_m	

Valorile caracteristicii de grupare X								
	x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1m}	$n_{1\bullet}$
	x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2m}	$n_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{im}	$n_{i\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	x_r	n_{r1}	n_{r2}	$\cdots$	n_{rj}	$\cdots$	n_{rm}	$n_{r\bullet}$
Total		$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet r}$	$n=\sum_{i=1}^{r} n_{i\bullet}=\sum_{j=1}^{m} n_{\bullet j}$

În cadrul tabelului de corelație întâlnim pe lângă frecvențe absolute ale evenimentelor compuse*:* $\{n_{ij}\}_{i=\overline{1,r},\, j=\overline{1,m}}$ şi *frecvențele marginale absolute*. Acestea sunt :

Frecvențe marginale absolute ale lui X: $n_{i\bullet}=\sum_{j=1}^{m} n_{ij}$, $i=\overline{1,r}$; aceste frecvențe exprimă numărul de unități din populație la care pentru X s-au înregistrat valoarea x_i , indiferent de valoarea înregistrată de variabilea Y.

Frecvențele marginale relative ale lui X sunt definite ca rapoartele

$$f_{i\bullet}=\frac{n_{i\bullet}}{n}=\sum_{j=1}^{m} f_{ij}\ .$$

Frecvențe marginale absolute ale lui Y: $n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$, $j = \overline{1,m}$; aceste frecvențe exprimă numărul de unități din populație la care pentru Y s-au înregistrat valoarea y_j, indiferent de valoarea înregistrată de variabilea *X*.

Frecvențele marginale relative ale lui Y sunt definite ca rapoartele

$$f_{\bullet j} = \frac{n_{\bullet j}}{n} = \sum_{i=1}^{r} f_{ij} \ .$$

În urma grupării combinate ale cărei rezultate se prezintă în tabelul de corelație se obțin:

- *r* distribuții de frecvențe formate după *Y*;

- *m* distribuții de frecvențe formate după *X*;

- o distribuție marginală formată după X , având tabloul

$$X_{marg} : \begin{pmatrix} x_1 & x_2 & \dots & x_r \\ n_{1\bullet} & n_{2\bullet} & \dots & n_{r\bullet} \end{pmatrix};$$

- o distribuție marginală formată după *Y*, având tabloul

$$Y_{marg} : \begin{pmatrix} y_1 & y_2 & \cdots & y_m \\ n_{\bullet 1} & n_{\bullet 2} & \dots & n_{\bullet m} \end{pmatrix}:$$

- o distribuție bidimensională de frecvențe formată simultan după *X* şi *Y*.

Frecvențele din interiorul tabelului permit, la fel ca şi în cazul diagramei de împrăştiere, identificarea existenței, sensului şi chiar a formei dependenței statistice.

Modul de aşezare a frecventelor în jurul diagonalei ne dă posibilitatea să apreciem intensitatea legăturii: concentrarea intensă a frecventelor în jurul diagonalelor indică existența unei legaturi strânse între caracteristici. În alte cazuri, frecventele se grupează pe diverse curbe. Dacă frecvențele se repartizează pe întregul tabel fără nici o regularitate, atunci ori nu există legătura, ori aceasta este foarte slabă.

Direcția legăturii este dată de poziția diagonalei în jurul căreia se grupează frecvențele: când diagonala leagă unghiul din partea stângă-sus al tabelului cu unghiul din partea dreaptă-jos legătura este *directa*, iar când uneşte unghiul partea stângă-jos cu unghiul din partea dreaptă-sus, între cele două caracteristici există o legătură *inversă*.

Exemplul 1. Tabelul de mai jos este un tabel de corelație sunt trecute datele privind frecvențele absolute ale valorilor a două caracteristici *X* şi *Y* reprezentând respectiv diametrul tulpinii unei plante şi procentul de fibre în funcție de diametru.

Să se calculeze frecvențele marginale şi să se stabilească tipul de corelație între cele două caracteristici.

Soluție: Frecvențele marginale sunt calculate pe ultima coloană şi ultima linie.

		Valorile y_j ale caracteristicii Y							
		2	3	4	5	6	7	8	$n_{i\bullet}$
Valorile x_i ale caracteristicii *X*	26	2	3	3	2	0	0	0	**10**
	24	4	5	13	7	4	0	0	**33**
	22	3	6	18	25	10	2	0	**64**
	20	0	1	8	17	18	3	0	**47**
	18	0	0	1	9	8	8	2	**28**
	16	0	0	0	2	3	4	6	**15**
	14	0	0	0	0	0	1	2	**3**
$n_{\bullet j}$		**9**	**15**	**43**	**62**	**43**	**18**	**10**	$n = \sum_{i=1}^{r} n_{i\bullet} = \sum_{j=1}^{m} n_{\bullet j}$ **=200**

Frecvențe marginale absolute ale lui X au fost calculate cu

formulele: $n_{i\bullet} = \sum_{j=1}^{m} n_{ij}$, $i = \overline{1,r}$ şi au fost trecute pe ultima coloană.

Frecvențe marginale absolute ale lui Y au fost calculate cu

formulele: $n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$, $j = \overline{1,m}$ şi au fost trecute pe ultima linie.

Se observă că între cele două caracteristici există o corelație negativă.

Observaţia 1. Dacă variabilele studiate sunt de tip alternativ (dihotomic), atunci celor două variante de răspuns ale fiecăreia (afirmativ şi negativ) li se vor acorda, convențional, valorile numerice 1 şi, respectiv, 0.

În acest caz se foloseşte următorul tabel de corelație:

Variante alternative ale caracteristicii *X*	Variante alternative ale caracteristicii *Y*		TOTAL
	DA $y_1=1$	NU $y_2=0$	
DA ($x_1=1$)	n_{11}	n_{12}	$n_{1\bullet}$
NU ($x_2=0$)	n_{21}	n_{22}	$n_{2\bullet}$
TOTAL	$n_{\bullet 1}$	$n_{\bullet 2}$	

Exemplul 2. Se consideră tablelul de corelație al unei distribuții bidimensionale *(X,Y)* următor:

y / x	1	0
1	8	2
0	5	1

Să se calculeze frecvențele absolute marginale şi să se scrie toate distribuțiile bidimensionale având aceleați distribuții marginale cu distribuția dată.

Soluție: Calculând frecvențele marginale cu formulele cunoscute găsim:

y / x	1	0	$n_{i\bullet}$
1	8	2	**10**
0	5	1	**6**
$n_{\bullet j}$	**13**	**3**	$n=\sum_{i=1}^{r} n_{i\bullet}=\sum_{j=1}^{m} n_{\bullet j}$ **=16**

Pentru a găsi toate distribuțiile bidimensionale care au aceleaşi

frecvențe marginale cu distribuția dată trebuie să găsim frecvențele absolute n_{ij}, $i \in \{1,2\}$, $j \in \{1,2\}$ din tabelul de corelație:

y / x	1	0	$n_{i\bullet}$
1	n_{11}	n_{12}	**10**
0	n_{21}	n_{22}	**6**
$n_{\bullet j}$	**13**	**3**	$n=\sum_{i=1}^{r} n_{i\bullet}=\sum_{j=1}^{m} n_{\bullet j}$ **=16**

Condițiile pe care trebuie să le satisfacă frecvențele absolute n_{ij}, $i \in \{1,2\}$, $j \in \{1,2\}$ sunt $n_{ij} \in N$ şi sistemul:

$$\begin{cases} n_{11}+n_{12}=10 \\ n_{21}+n_{22}=6 \\ n_{11}+n_{21}=13 \\ n_{12}+n_{22}=3 \end{cases}$$

Rezolvând sistemul în mulțimea numerelor naturate obținem distribuțiile:

8	2
5	1

7	3
0	6

9	1
2	4

10	0
3	3

Observaţia 2. O mare parte din distribuțiile cu două variabile întâlnite în cercetările experimentale pun în evidență variabile a căror fluctuație întâmplătoare afectează valorile ambelor variabile şi este de aşteptat ca populația tuturor cazurilor

corespunzătoare să prezinte o *distribuție bidimensională normală* în sensul că fiecare din variabile în parte este de tipul distribuției normale.

Analiza legăturii între variabile se reduce la estimarea unui parametru al distribuției bidimensionale respective printr-o mărime *(r*)calculată pe baza datelor probei de sondaj de care dispunem, numit *coeficient de corelație.*

Definiția 1. Fie X şi Y două variabile având respectiv mediile $M(X)=m_x$ şi $M(Y)=m_y$ şi dispersiile $D(X)=\sigma_x^2$ şi $D(Y)=\sigma_y^2$. Se numeşte *moment de corelație* sau *covarianța* variabilelor X şi Y numărul:

$$cov(X,Y)=M\left[(X-m_x)\cdot(Y-m_y)\right]. \qquad (1)$$

Covarianța variabilelor X şi Y se mai notează σ_{XY}.

Propoziția 1. Covarianța variabilelor aleatoare X şi Y se poate calcula cu formula:

$$cov(X,Y)=M(X\cdot Y)-M(X)\cdot M(Y). \qquad (1')$$

Într-adevăr, efectuând produsul în (1) avem:

$$cov(X,Y)=M\left[(X-m_x)\cdot(Y-m_y)\right]=M\left(X\cdot Y-X\cdot m_y-m_x\cdot Y+m_x\cdot m_y\right)=$$

$$=M(X\cdot Y)-M(X)\cdot m_y-m_x\cdot M(Y)+m_x\cdot m_y=M(X\cdot Y)-M(X)\cdot M(Y).$$

Observația 3. Dacă variabilele aleatoare X şi Y sunt independente, atunci conform proprietăților mediei unei variabile aleratoare rezultă că $M(X\cdot Y)-M(X)\cdot M(Y)=0$ ceea ce conduce la $cov(X,Y)=0$.

Deci două variabile aleatoare independente X şi Y au covarianța nulă. Reciproc nu este adevărat.

Covarianța este un indicator la corelației, exprimând graqdul de împrăştiere a celor două variabile față de mediile respective şi prin aceasta intensitatea legăturii dintre variabile indicând şi sensul acesteia.

Numărul care exprimă măsura corelației este *coeficientul de corelație.*

Definiția 2. Se numeşte *coeficient de corelație* al variabilelor X şi Y numărul ce reprezintă valoarea medie a produsului abaterilor normate:

$$\rho(X,Y)=M\left(\frac{X-m_x}{\sigma_x}\cdot\frac{Y-m_y}{\sigma_y}\right)$$

(2)

Coeficientul de corelație se mai notează pur şi simplu cu ρ.

Observaţia 4. Din proprietățile mediei unei variabile aleatoare deducem că pentru coeficientul de corelație putem folosi una din formele:

$$\rho = \frac{1}{\sigma_x \sigma_y} \cdot M[(X - m_x)(Y - m_y)] = \frac{M(XY) - M(X)M(Y)}{\sigma_x \sigma_y} \qquad (2')$$

sau

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{cov(X,Y)}{\sigma_x \cdot \sigma_y} \qquad (2'')$$

Admitem următoarele rezultate privind coeficientul de corelație:

Propoziţia 2. *(Proprietățile coeficientului de corelație)*

(1). Dacă X şi Y sunt independente atunci $\rho(X,Y) = 0$.

(2). $|\rho(X,Y)| \le 1$, oricare ar fi variabilele aleatoare X şi Y.

(3). Între X şi Y există o dependență liniară dacă şi numai dacă $|\rho(X,Y)| = 1$.

Observaţia 5. Reciproca afirmației (1) din propoziția de mai sus nu este adevărată. Astfel, dacă $\rho(X,Y) = 0$ variabilele aleatoare nu sunt în mod necesar independente, dar dependența lor nu este liniară, ea putând fi de altă natură. În acest caz spunem că cele două variabile sunt necorelate.

În cazul dependenței liniare între două variabile aleatoare X şi Y care au distribuții normale se foloseşte următorul rezultat pe care îl dăm fără demonstrație:

Propoziţia 3. Fie X şi Y două variabile care au distribuții normale între care există o dependență liniară şi fie variabila $Z=Y-(aX+b)$.

Dacă $M(Z^2)$ are valoare minimă atunci: $M(z)=0$ şi $\frac{\sigma_z^2}{\sigma_y^2} = 1 - \rho^2$.

Observaţia 6. Propoziția precedentă arată că dacă funcția liniară $h(x)=ax+b$ este cea mai bună aproximație în medie pătratică a lui y, atunci valoarea medie a abaterii lui y de la funcția liniară $h(x)$ este egală cu zero iar raportul dintre dispersia ei şi dispersia lui y se măsoară prin coeficientul de corelație după relația $\frac{\sigma_z^2}{\sigma_y^2} = 1 - \rho^2$.

Coeficientul de corelație caracterizează măsura *dependenței liniare*

între variabilele *X* şi *Y*. Cu cât |□| se apropie mai mult de 1, cu atât este mai strânsă dependența liniară a variabilelor *X* şi *Y*. Egalitatea |□|=1 înseamnă existența unei dependențe *funcționale liniare* între variabilele *x* şi *y*.

Inversând problema, din formula precedentă putem scrie :

$$\rho^2 = 1 - \frac{\sigma_z^2}{\sigma_y^2} \qquad (4)$$

formulă care poate fi folosită pentru determinarea coeficientului de

corelație $\rho = \rho_{y,x}$ atunci când regresia variabilei *Y* în raport cu variabila *X* este liniară şi variabilele *X* şi *Y* au distribuții normale.

Această valoare a lui □ este numită *coeficientul de corelație în regresia liniară* şi numai în condițiile citate este acelaşi cu coeficientul de corelație dat de (1").

Estimarea covarianței şi a coeficientului de corelație.

Parametrii unei legi de repartiție reprezintă "adevăratele valori" ale indicatorilor legii, adică acele valori care s-ar obține dacă s-ar lucra cu întreaga populație. Indicii selecției (eşantionului) reprezintă valorile "apropiate" de "adevăratele" valori. Ei dau indicații asupra populațiilor statistice, permițînd să se tragă concluzii asupra parametrilor şi se

numesc *estimatori*□

Pentru determinarea estimațiilor parametrilor cu care se studiază dependența variabilelor (coeficientul de corelație şi covarianța) se efectuează un anumit număr *n* de observații (măsurători) asupra variabilelor *X* şi *Y*. Rezultatul celei de-a *i*-a experiențe ne dă o pereche (x_i, y_i), $i=1,2,\ldots,n$. Rezultatul după cele *n* experiențe conduce la şirurile numerice:

$X : x_1, x_2, \ldots, x_i, \ldots, x_n, \quad Y : y_1, y_2, \ldots, y_i, \ldots, y_n.$

Cu aceste valori se determină estimațiile punctuale ale valorilor medii m_x şi m_y, ale abaterilor standard □$_x$ şi □$_y$, ale covarianței □$_{xy}$ precum şi coeficientul de corelație □□

Se arată că:

Propoziția 4. *(Estimațiile mediilor şi dispersiilor)*

(a). Estimațiile mediilor teoretice m_x şi m_y sunt *mediile de selecție* $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ şi $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$;

(b). Estimațiile dispersiilor σ_x^2 şi σ_y^2 sunt *dispersiile de selecție modificate*

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \text{şi} \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \ . \tag{5}$$

<u>Observația 7</u>. Numitorul formulelor pentru dispersie şi pentru abaterea standard, adică *n*-1 (numărul observațiilor micşorat cu o unitate), poartă denumirea de *grade de libertate.*

Faptul că suma pătratelor abaterilor se împarte la *n*-1 şi nu la *n* (pentru a fi vorba de o medie) are următoarea explicație: fie o populație statistică de dispersie σ^2 din care să extragem un eşantion de mărimea n ; dispersia acestui eşantion este s^2 care nu va coincide cu σ^2. Extrăgând din populația statistică un număr foarte mare de eşantioane de mărimea *n*, atunci media $\overline{s^2}$ a dispersiilor s^2 ale acestor eşantioane coincide exact cu σ^2 numai atunci când dispersiile eşantioanelor se calculează împărțind suma pătratelor abaterilor la *n*-1, şi nu la *n*, deaoarece în urma calculării mediei aritmetice (pe care se bazează calculul dispersiei), una din valorile individuale ale eşantionului depinde de celelalte *n*-1 ce pot fi variabile prin însăşi egalitatea care dă media.

Pentru distribuțiile teoretice în care toate valorile individuale ale variabilei sunt libere, dispersia σ^2 şi abaterea standard σ se calculează cu formule asemănătoare doar că sumele abaterilor şi ale pătratelor abaterilor se împart la n şi nu la n-1.

Dacă notăm cu : $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ şi $\overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2$, mediile de selecție

ale pătratelor variabilelor *X* respectiv *Y*, atunci, după efectuarea unor calcule simple în expresiile dispersiilor de selecție modificate s_x^2 şi

s_y^2, găsim:

$$s_x^2 = \frac{n}{n-1}\left[\overline{(x^2)} - (\bar{x})^2\right], \quad s_y^2 = \frac{n}{n-1}\left[\overline{(x^2)} - (\bar{y})^2\right]. \tag{5'}$$

Se arată de asemenea că:

Propoziţia 5. ***(****Estimaţiile covarianţei şi coeficientului de corelaţie)*

(a). Estimatorul covarianţei teoretice σ_{xy} este cantitatea

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \tag{6}$$

(b). Estimatorul coeficientului de corelaţie $\rho = \rho(x,y)$ este numărul

$$r = \frac{s_{xy}}{s_x \cdot s_y} \tag{7}$$

Observaţia 8. Notând cu $\overline{x \cdot y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ media de selecţie a

produsului *XY,* după efectuarea unor calcule simple în expresia covarianţei empirice obţinem:

$$s_{xy} = \frac{n}{n-1}\left(\overline{x \cdot y} - \overline{x} \cdot \overline{y}\right). \tag{6'}$$

Cu aceste estimaţii se construieşte numărul

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}\,. \tag{7}$$

Înlocuind expresiile estimaţiilor s_x, s_y şi s_{xy}, după efectuarea simplificărilor, găsim:

$$r_{xy} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\sqrt{\left(\overline{x^2} - \overline{x}^2\right)\left(\overline{y^2} - \overline{y}^2\right)}} \tag{7'}$$

Definiţia 3. ***(a).*** Numărul $s_{xy} = \frac{n}{n-1}\left(\overline{x \cdot y} - \overline{x} \cdot \overline{y}\right)$ se numeşte *covarianţa de selecţie* sau *covarianţa empirică* a variabilelor *X* şi *Y*.

(b). Numărul $r_{xy} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\sqrt{\left(\overline{x^2} - \overline{x}^2\right)\left(\overline{y^2} - \overline{y}^2\right)}}$ se numeşte *coeficientul empiric de corelaţie* al variabilelor *X* şi *Y*.

Observaţia 9. Din cele prezentate anterior reţinem că folosirea coeficientului de corelaţie este recomandabilă îndeosebi atunci când legătura dintre variabile nu se

abate mult de la liniaritate, iar populația studiată este de tipul distribuțiilor normale bidimensionale. Astfel, în cazul când datele studiate aparţin unei distribuții bidimensionale normale şi relația dintre variabile este liniară coeficientul de corelație are un înțeles statistic bine definit. În caz contrar coeficientul de corelație îşi pierde înțelesul său statistic, iar examinarea semnificației sale statistice devine lipsită de sens.

Exemplul 3. Măsurarea înălțimii X (în cm) şi a greutății Y (în kg) pentru 70 de persoane a condus la distribuția următoare :

Y / X	48-56	56-64	64-72	72-80
160-165	16	8	1	0
165-170	1	10	4	1
170-175	0	4	8	2
175-180	0	1	5	9

a). Considerând pentru fiecare clasă a fiecărei variabile valoarea centrală a clasei să se scrie distribuția corespunzătoare şi pornind de la aceasta să se facă schimbările de variabile (folosind metoda « zeroului fals ») $T = \frac{X - 167,5}{5}$, $Z = \frac{Y - 60}{8}$, să se scrie tabelul de corelație al noii distribuții bidimensionale (T,Y) calculând frecvențele marginale.

*b).*Să se calculeze pentru fiecare variabilă mediile şi dispersiile.

c). Să se calculeze covarianța variabilelor X şi Y precum şi coeficientul de corelație.

Soluție. Mai întâi scriem tabelul de corelație considerând pentru fiecare variabilă valoarea centrală a clasei :

Y / X	52	60	68	76
162,5	16	8	1	0
167,5	1	10	4	1
172,5	0	4	8	2
177,5	0	1	5	9

Considerând « zerourile false » $x_0 = 167{,}5$ şi $y_0 = 60$ şi făcând schimbările de variabile $T = \frac{X - 167{,}5}{5}$, $Z = \frac{Y - 60}{8}$ (numitorii sunt amplitudinile claselor) obținem noul tabel de corelație :

Z / T	-1	0	1	2	$n_{i\bullet}$	$n_{i\bullet} \cdot t_i$	$n_{i\bullet} \cdot t_i^2$
-1	16	8	1	0	**25**	-25	25
0	1	10	4	1	**16**	0	0
1	0	4	8	2	**14**	14	14
2	0	1	5	9	**15**	30	60
$n_{\bullet j}$	**17**	**23**	**18**	**12**	**70**	19	99
$n_{\bullet j} \cdot z_i$	-17	0	18	24	25		
$n_{\bullet j} \cdot z_i^2$	17	0	18	48	83		

Calculăm mediile şi dispersiile variabilelor *T* şi *Z:*

$M(T) = \frac{19}{70} \approx 0{,}27$, $M(Z) = \frac{25}{70} \approx 0{,}36$

$M\left(T^2\right) = \frac{99}{70} \approx 1{,}4114$, $[M(T)]^2 \approx 0{,}0729$, $M\left(Z^2\right) = \frac{83}{70} \approx 1{,}1857$,

$M(TZ) = \frac{1}{70} \sum_{i=1}^{4} \sum_{j=1}^{4} n_{ij} t_i z_j = \frac{1}{70}[16 - 1 + 8 + 4 + 10 + 36] = \frac{73}{70} \approx 1{,}04;$

$[M(Z)]^2 \approx 0{,}1296$,

$D(T) = M(T^2) - [M(T)]^2 \approx 1{,}4114 - 0{,}0729 \approx 1{,}34$.

$D(Z) = M(Z^2) - [M(Z)]^2 \approx 1{,}1857 - 0{,}1296 \approx 1{,}06$.

Folosind proprietățile mediei şi dispersiei şi anume:

$$M(aX + b) = a \cdot M(X) + b,\ D(aX + b) = a^2 \cdot D(X),$$

şi ținând seama că $X = 5T + 167{,}5$, $Y = 8Z + 60$, găsim:

$M(X) = 5M(T) + 167{,}5 \approx 168{,}86$, $M(Y) = 8M(Z) + 60 \approx 62{,}86,$

$D(X) = 25D(T) \approx 33{,}52$, $\sigma_X \approx 5{,}79$, $D(Y) = 64D(Z) \approx 67{,}72$, $\sigma_Y \approx 8{,}23$.

$cov(T,Z)) = M(T \cdot Z) - M(T) \cdot M(Z) \approx 1{,}04 - 0{,}27 \cdot 0{,}36 \approx 0{,}95$.

Ținând seama de proprietatea :

$$cov(aT + b, cZ + d) = ac \cdot cov(T, Z),$$

deducem:

$$cov(X, Y) = cov(5T + 167{,}5{,}8Z + 60) = 5 \cdot 8 \cdot cov(T, Z) \approx 37{,}84 .$$

Coeficientul de corelație este:

$$r_{xy} = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y} \approx \frac{37{,}84}{5{,}79 \cdot 8{,}23} \approx 0{,}79$$

2.3.3. Analiza regresiilor. Regresia liniară

Pentru a răspunde la întrebarea ce fel de corelație există între variabile recurgem la *analiza regresiilor* care constă în determinarea *funcției de regresie* între cele două variabile *X* şi *Y.*

Aşa cum am văzut, felul legăturii între variabilele *X* şi *Y* se poate

obține dacă se observă o anumită concentrare a punctelor din corelogramă în jurul unei anumite curbe în plan, curbă care se numeşte *curba de regresie* şi este reprezentarea geometrică a funcției de regresie.

Coeficientul de corelație ne dă indicații asupra sensului şi intensității legăturii de dependență dintre fenomene, fără a putea preciza, sub aspect cantitativ, cu cât creşte sau scade un fenomen când cel cu care se corelează creşte sau scade cu o anumită cantitate.

Regresia, noțiune strâns legată de noțiunea de corelație, completează corelația şi prin intermediul coeficientului de regresie, stabileşte cu cât creşte sau descreşte sub aspect cantitativ, un fenomen, când cel cu care se corelează creşte sau descreşte cu o unitate de măsură.

Regresia poate fi *simplă şi multiplă*, *liniară şi neliniară.* Ca şi corelația, regresia poate fi *directă*, când fenomenele evoluează în acelaşi sens sau indirectă, când fenomenul evoluează în sens opus. Problema care se pune este deducerea unei funcții teoretice pentru legătura respectivă plecând de la distribuția empirică cunoscută. Această problemă poartă numele de *ajustarea distribuției empirice.* Ea

constă în determinarea unor reprezentări analitice a dependenței funcționale căutate, adică de a alege o formulă care să descrie rezultatele experimentului. Vom considera cazul când punctele corespunzătoare unei serii statistice sunt dispuse aproximativ după o dreaptă. În acest caz legătura cea mai simplă este cea liniară în care unei creşteri a lui x îi corespunde o creştere sau o scădere proporțională a lui y, $y=\square x+\square\square\square\square$numită *ecuația dreptei de regresie.*

Parametrul $\square$ reprezintă ordonata (înălțimea) intersecției dreptei de regresie cu axa y-lor, adică valoarea y corespunzătoare la x=0 iar $\square$ reprezintă panta (înclinarea) față de axa abscisei a dreptei de regresie, adică modificarea lui y atunci când x creşte cu o unitate.

În cazul corelațiilor valorile individuale nu se găsesc niciodată pe o dreaptă, ci sunt mai mult sau mai puțin împrăştiate. Putem totuşi obține în acest caz o linie dreaptă față de care abaterile valorilor individuale să fie minime. Aceasta se realizează atunci când dreapta trece prin valorile medii m_x şi m_y ale variabilelor, adică $m_y=\square\square m_x+\square\square\square$de unde deducem $\square\square m_y-\square\square m_x$. Înlocuind pe $\square\square\square$ în ecuația dreptei de regresie găsim: $y - m_y=\square(x-m_x)$.

Definiția 4. Parametrul $\square$ se numeşte *coeficientul de regresie teoretică* şi este definit prin expresia:

$$\alpha = \rho \cdot \frac{\sigma_y}{\sigma_x} \tag{8}$$

iar *ecuația dreptei de regresie* (a lui y asupra lui x) se scrie:

$$y - m_y = \rho \cdot \frac{\sigma_y}{\sigma_x}(x - m_x) \tag{9}$$

Estimarea parametrilor dreptei de regresie.

Pornind de la această definiție, dacă variabilele X şi Y sunt date prin şirurile numerice:

$$X: x_1, x_2, \ldots, x_i, \ldots, x_n, \quad Y: y_1, y_2, \ldots, y_i, \ldots, y_n.$$

prin estimarea parametrilor teoretici care intervin în expresia coeficientului de regresie teoretic, se obține estimarea acestuia:

Propoziția 6. Estimatorul coeficientului de regresie teoretică $\square\square$ este mărimea

$$a = a_{y,x} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} \tag{8'}$$

iar estimația ecuației dreptei de regresie teoretice este:

$$y-\bar{y}=\frac{\overline{x\cdot y}-\bar{x}\cdot\bar{y}}{\overline{x^2}-\bar{x}^2}\cdot(x-\bar{x}) \qquad (9')$$

Într-adevăr, să considerăm că variabilele X şi Y sunt date prin şirurile numerice: X : $x_1, x_2, \ldots, x_i, \ldots, x_n$, Y : $y_1, y_2, \ldots, y_i, \ldots, y_n$.

Cu aceste valori se determină estimațiile punctuale $\bar{x}$ şi $\bar{y}$, ale valorilor medii m_x şi m_y, s_x şi s_y, ale abaterilor standard σ_x, σ_y, respectiv r şi a coeficientului de corelație ρ. Substituind aceste estimații în expresia lui $a=\rho\cdot\frac{\sigma_y}{\sigma_x}$, obținem $a=r\cdot\frac{s_y}{s_x}$.

Conform rezultatelor din secțiunea precedentă, formulele (5), (5'), (6) şi (7), substituind expresiile corespunzătoare în $a=r\cdot\frac{s_y}{s_x}$, după efectuarea calculelor, deducem că *coeficientul empiric de corelație* (a lui y față de x) se calculează cu formula:

$$a=a_{y,x}=\frac{\overline{x\cdot y}-\bar{x}\cdot\bar{y}}{\overline{x^2}-\bar{x}^2}.$$

Evident că estimația parametrului m_y-m_x este $b=\bar{y}-a\bar{x}$ şi atunci <u>ecuația dreptei de regresie empirică</u> $y=ax+b$ se mai scrie

$$y-\bar{y}=a(x-\bar{x}) \quad \text{sau} \qquad y-\bar{y}=\frac{\overline{x\cdot y}-\bar{x}\cdot\bar{y}}{\overline{x^2}-\bar{x}^2}\cdot(x-\bar{x}).$$

<u>Exemplul 4.</u> Să se determine dreapta de regresie pentru variabilele X şi Y de la *Exemplul 3.*

<u>Soluție.</u> <u>Ecuația dreptei</u> de regresie a lui Y asupra lui X este :

$$y-M(Y)=a[x-M(X)].$$

Pentru datele respective am obținut: *M(X)=168,86, M(Y)=62,86, D(X)=33,52, COV(X,Y)=37,84.* $a=\frac{cov(X,Y)}{D(X)}\approx 1{,}13$ .Substituind *M(X)*, *M(Y)* şi coeficientul a în ecuația dreptei de regresie obținem : $y=1{,}13x-127{,}77$.

<u>Exemplul 5.</u> În urma efectuării a 8 măsurători asupra două caracteristici X şi Y ale unei populații, s-au găsit valorile date în tabelul de mai jos:

Proba	1	2	3	4	5	6	7	8
X: x_i	26,9	26,3	23,6	24,8	29,1	19,6	17,9	19,5
Y: y_i	54,0	52,2	55,5	57,1	54,3	63,2	70,1	70,2

Calculati covarianța şi coeficientul de corelație al variabilelor *X* şi *Y*.

Soluţie. Aşezăm datele în tabelul de mai jos în care trecem în coloane corespunzătoare:

x_i *(cm)*	y_i *(%)*	x_i^2	y_i^2	$x_i y_i$
26,9	54,0	723,61	2916,00	1452,60
26,3	52,2	691,69	2724,84	1372,86
23,6	55,5	556,96	3080,25	1309,80
24,8	57,1	615,04	3260,41	1416,08
29,1	54,3	846,81	2948,48	1580,13
19,6	63,2	384,16	3994,24	1238,72
17,9	70,1	320,41	4914,01	1254,79
19,5	70,2	380,25	4928,04	1368,90
$\sum x_i =$ 187,7	$\sum y_i =$ 476,6	$\sum x_i^2 =$ 4518,93	$\sum y_i^2 =$ 28766,28	$\sum x_i y_i =$ 10993,88
$\overline{x} =$ 23,46	$\overline{y} =$ 59,57	$\overline{x^2} =$ 564,86	$\overline{y^2} =$ 3595,78	$\overline{x \cdot y} =$ 1374,23

Efectuând calculele necesare obținem:

$\overline{x} \cdot \overline{y} = 139751{,}22$, $\overline{x}^2 = 550{,}3716$, $\overline{y}^2 = 3548{,}5849$;

$$s_x^2 = \frac{n}{n-1}\left[\overline{(x^2)} - (\overline{x})^2\right] = \frac{8}{7}(564{,}86 - 550{,}37) = 16{,}56;$$

$$s_y^2 = \frac{n}{n-1}\left[\overline{(y^2)} - (\overline{y})^2\right] = \frac{8}{7}(3595{,}78 - 3548{,}58) = 53{,}94;$$

$$s_{xy} = \frac{n}{n-1}\left(\overline{x \cdot y} - \bar{x} \cdot \bar{y}\right) = \frac{8}{7}(1374{,}23 - 1397{,}51) = -26{,}61;\ r = \frac{s_{xy}}{s_x \cdot s_y} = -0{,}89.$$

Observaţia 10. În secțiunea precedentă am remarcat că coeficientul empiric de corelație r este un estimator al coeficientului de corelație ρ şi acesta are un înțeles statistic bine definit atunci când variabilele X şi Y au distribuții normale şi regresia uneia față de alta este liniară.

Dar liniaritatea regresiei nu constituie un criteriu pentru normalitatea distribuțiilor variabilelor X şi Y , caz în care pentru estimarea parametrilor α şi β ai regresiei nu mai putem folosi estimările parametrilor σ_x, σ_y şi ρ care intervin în definiția coeficientului teoretic de regresie liniară β şi implicit α. În această situație suntem nevoiți să determinăm *dreapta de regresie empirică* care se poate prefigura, prin alte metode decât prin aplicarea formulelor obținute în *Propoziția 6.* de mai sus.

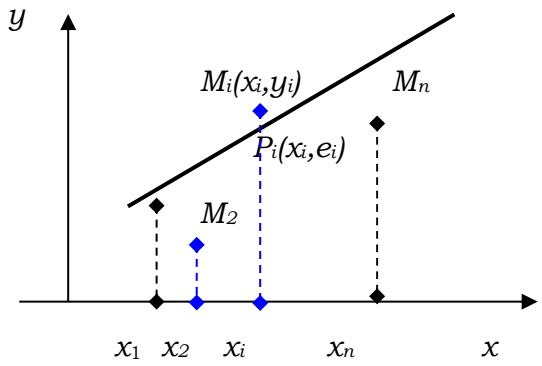
Metoda celor mai mici pătrate .

Una dintre metodele prin care se determină estimatorii parametrilor dreptei de regresie *metoda celor mai mici pătrate.*

Fie variabilele X şi Y sunt date prin şirurile numerice:

$$X: x_1, x_2, \ldots, x_i, \ldots, x_n, \quad Y: y_1, y_2, \ldots, y_i, \ldots, y_n.$$

Presupunem că variabilele X şi Y nu au distribuții normale dar punctele $M_i(x_i,y_i)$, i=1, 2, . . ., n, par a se concentra în jurul unei drepte *(d)* a cărei ecuații vrem să o determinăm.Fie dreapta (*d)* de ecuație $y=ax+b$. În general punctele $M_i(x_i,y_i)$ nu se află pe această dreaptă



Să punem $e_i=ax_i+b$, pentru i=1, 2, …,n, unde a şi b sunt parametrii

reali ai dreptei *(d).* Fie $P_i(x_i,e_i)$ punctul de abscisă x_i de pe dreapta (*d*). Atunci y_i-e_i reprezintă abaterea punctului M_i de la punctul P_i de aceeaşi abscisă de pe dreapta *(d).*

Metoda celor mai mici pătrate constă în determinarea parametrilor *a* şi *b* astfel ca suma pătratelor abaterilor $S(a,b)=\sum_{i=1}^{n}(y_i-e_i)^2$ să fie minimă.

Valorile parametrilor *a* şi *b* folosind principiul metodei celor mai mici pătrate se pot obține prin utilizarea de raționamente de matematici elementare astfel:

Efectuând calculele obținem:

$$S(a,b)=\sum_{i=1}^{n}[y_i-(ax_i+b)]^2=\sum_{i=1}^{n}y_i^2-2\sum_{i=1}^{n}y_i(ax_i+b)+\sum_{i=1}^{n}(ax_i+b)^2=$$

$$=\sum_{i=1}^{n}y_i^2-2a\sum_{i=1}^{n}x_iy_i-2b\sum_{i=1}^{n}y_i+a^2\sum_{i=1}^{n}x_i^2+2ab\sum_{i=1}^{n}x_i+nb^2\ .$$

Astfel *S* se poate scrie ca un polinom de gradul II în *a* şi *b*:

$$S(a,b)=\underbrace{\left(\sum_{i=1}^{n}x_i^2\right)}_{A}\cdot a^2+2\underbrace{\left(\sum_{i=1}^{n}x_i\right)}_{B}ab+\underbrace{n}_{C}b^2+2\underbrace{\left(-\sum_{i=1}^{n}x_iy_i\right)}_{D}a+2\underbrace{\left(-\sum_{i=1}^{n}y_i\right)}_{E}b+\underbrace{\sum_{i=1}^{m}y_i^2}_{F}$$

$=Aa^2+2Bab+Cb^2+2Da+2Eb+F.$

Folosind rezultate din Analiza Matematică se demonstrează că funcția de două variabile $S(a,b)$ are o valoare minimă dacă derivatele sale parțiale în raport cu *a* şi *b* sunt nule :

$$\begin{cases}\dfrac{\partial S}{\partial a}=0\\[2mm]\dfrac{\partial S}{\partial b}=0\end{cases}\Leftrightarrow\begin{cases}Aa+Bb+D=0\\Ba+Cb+E=0\end{cases}$$

Rezolvând acest sistem obținem : $a=\dfrac{BE-CD}{AC-B^2}$, $b=\dfrac{BD-AE}{AC-B^2}$.

Substituind expresiile lui *A,B,C,D,E* găsim:

$$a=\frac{n\sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}=\frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\overline{x^2}-(\bar{x})^2}\quad ,$$

$$b=\frac{BD-AE}{AC-B^2}=\frac{B(CD-BE)-E(AC-B^2)}{C(AC-B^2)}=-\frac{B}{C}\cdot a-\frac{E}{C}=-a\cdot\bar{x}+\bar{y}.$$

Din aceste relații se vede $\bar{y}=a\bar{x}+b$, deci $(\bar{x},\bar{y})$ se găseşte pe dreapta (*d).*

Exemplul 6. Fie seria statistică dată de tabelul :

x_i	1	2	3	4	5	6
y_i	2,10	2,41	2,75	3,03	3,24	3,51

Să se determine dreapta de regresie a lui Y asupra lui X şi cu ajutorul acesteia să se facă prognoza lui *Y* când *X* ia valoarea 10.

Soluţie. Efectuând calculele obținem pentru *n*=6:

$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=3{,}5$; $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=2{,}84$; $\overline{x^2}=\frac{1}{n}\sum_{i=1}^{n}x_i^2=15{,}16$;

$\overline{xy}=\frac{1}{n}\sum_{i=1}^{n}x_iy_i=10{,}74.$ $\qquad a=\frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\overline{x^2}-(\bar{x})^2}=\frac{10{,}74-9{,}94}{15{,}16-12{,}25}=0{,}28$

$b=\bar{y}-a\cdot\bar{x}=2{,}84-0{,}98=1{,}86$

Astfel, ecuația dreaptei de regresie empirică este: y=0,28x+1,86.

Atunci, pentru $x=10$, avem $y=4{,}66$.

2.4. ESTIMAŢII ŞI SEMNIFICAŢIA PARAMETRILOR UNEI REPARTIŢII.

În teoria probabilităților ipoteza de la care se pleacă este cunoaşterea variabilelor aleatoare prin functiile de probabilitate sau prin functiile de repartitie.

Statistica pleacă de la măsuratorile efectuate asupra unei caracteristici si caută să găsească modelul probabilistic teoretic exact căruia i se supune caracteristica respectivă.

Teoria estimatiei urmareste evaluarea parametrilor unei repartiții în general cunoscute. Valorile numerice obtinute se numesc *estimatii* sau *estimatori.*

Estimațiile sunt de două feluri: *estimatii punctuale* şi *estimații prin intervale de încredere.*

2.4.1. Estimaţii punctuale. Metoda verosimilităţii maxime.

Fie o populație statistică de caracteristică X care are funcția (densitatea) de probabilitate $f(x,a)$ depinzând de un singur parametru ce trebuie estimat. Extrăgând succesiv n elemente din această populație se obțin eşantionul de valori $\{x_1,x_2,\ldots, x_n\}$.

Funcția de probabilitate $f(x,a)$ corespunzător valorilor x_i, ia valorile

$$f(x_i,a)=P(X=x_i),\ i=1,2,\ldots,n.$$

Deoarece variabila de selecție presupune realizat evenimentul

$\bigcap_{i=1}^{n}(X=x_i)$, atunci probabilitatea realizării simultane a evenimentelor independente $(X=x_i)$, adică a evenimentului $\bigcap_{i=1}^{n}(X=x_i)$, este

$$P\left(\bigcap_{i=1}^{n}(X=x_i)\right)=\prod_{i=1}^{n}P(X=x_i)=\prod_{i=1}^{n}f(x_i,\lambda).$$

Definiţia 1. Funcția definită prin probabilitatea ca variabila aleatoare X să ia exact valoarile ea x_i, $i=1,2,\ldots,n$ se numeşte *funcția de verosimilitate* a lui X şi se notează:

$$L(\bar{x},\lambda)=\prod_{i=1}^{n}f(x_i,\lambda),\ \text{unde}\ \bar{x}=(x_1,x_2,\ldots,x_n). \qquad (1)$$

Deoarece funcția logaritm natural este o funcție crescătoare, rezultă că funcțiile $L(\bar{x},\lambda)$ şi $\ln L(x,\lambda)$ îşi ating simultan maximul. Or, acesta poate fi atins în punctele în care derivata este nulă, adică

$$\frac{\partial \ln L(\bar{x},\lambda)}{\partial \lambda} = 0 \quad \text{sau} \quad \sum_{i=1}^{n} \frac{\partial \ln f(x_i,\lambda)}{\partial \lambda} = 0 .$$

(2)

Definiţia 2. Ecuaţia $\sum_{i=1}^{n} \frac{\partial \ln f(x_i,\lambda)}{\partial \lambda} = 0$ se numeşte *ecuaţia de verosimilitate.*

Orice soluţie □* a ecuaţiei de verosimilitate se numeşte *estimatorul de maximă verosimilitate* al parametrului □□

Metoda de calcul a estimatorului cu ajutorul ecuaţiei de verosimilitate este numită *metoda verosimilităţii maxime.*

Prin metoda verosimilităţii maxime parametrului □ i se atribuie o valoare numerică, adică un punct □* de pe dreapta reală. O astfel de estimaţie se mai numeşte *estimaţie punctuală.*

2.4.2. Estimaţii prin intervale de încredere.

Fie o populaţie statistică de caracteristică *X* care are funcţia (densitatea) de probabilitate *f(x,*□□ depinzând de un singur parametru □ ce trebuie estimat. Extrăgând succesiv *n* elemente din această populaţie se obţin eşantionul de valori {x_1,x_2,…, x_n}.

O altă procedură de estimare a unui parametru a unei distribuţii teoretice *f(x,)* a unei caracteristici *X* a unei populaţii statistice este ce a *intervalelor de încredere* Aceasta constă în determinarea, pe baza unui eşantion de valori {x_1,x_2,…, x_n} ale lui X, a unui interval $(\lambda_1^*, \lambda_2^*)$, unde $\lambda_1^* = \lambda_1^*(x_1, x_2, ..., x_n), \lambda_2^* = \lambda_2^*(x_1, x_2, ..., x_n)$, astfel încât pentru o probabilitate dată □□, □□□□□□□□1, să avem: $P\left(\lambda_1^* < \lambda < \lambda_2^*\right) = P$.

Relaţia $P\left(\lambda_1^* < \lambda < \lambda_2^*\right) = P$ are următoarea semnificaţie:

Cu probabilitatea □, intervalul $(\lambda_1^*, \lambda_2^*)$ acoperă adevărata valoare a parametrului a.

Definiţia 3. O estimaţie de tipul $P\left(\lambda_1^* < \lambda < \lambda_2^*\right) = P$ se numeşte *estimaţie de încredere.*

Probabilitatea P nu depinde de parametrul λ şi se numeşte *probabilitate de încredere (siguranţă)* sau *nivel de încredere (siguranţă).*

Probabilitatea $\alpha = 1 - P$ poartă numele de *probabilitate (coeficient) de risc, prag de semnificaţie.* sau încă *probabilitatre de transgresiune*.

Intervalul $(\lambda_1^*, \lambda_2^*)$ poartă numele de *interval de încredere* pentru parametrul λ. Diferenţa $\lambda_2^* - \lambda_1^*$ se numeşte *lungimea intervalului de încredere.*

Elementul aleator este intervalul de încredere $(\lambda_1^*, \lambda_2^*)$ şi nu

parametrul . Acest interval depinde de datele de selecţie şi variază de

la o selecţie la alta. Cu cât acest interval este mai mic şi probabilitatea $P = 1 - \alpha$ este mai apropiată de 1, cu atât avem o indicaţie mai precisă asupra parametrului .

Practic pentru α se iau valorile 0,05 ; 0,01 ; 0,001 ; etc.

Mulţimea valorilor de selecţie $(x_1, x_2, \ldots, x_n)$ pentru care

$$\lambda_1^*(x_1, x_2, \ldots, x_n) < \lambda < \lambda_2^*(x_1, x_2, \ldots, x_n), \qquad (3)$$

se numeşte *regiunea de acceptare* pentru parametrul .

***Aplicaţie :* <u>Determinarea intervalelor de încredere pentru media teoretică a legii normale.</u>**

Intervalele de încredere pentru media teoretică m a caracteristicii X a unei populaţii, care urmează o lege de repartiţie normală ( ) se pot determina în două cazuri:

<u>Cazul 1</u>. *<u>Se cunoaşte abaterea standard σ a caracteristicii X.</u>*

Se consideră caracteristica X care urmează legea normală $N(m, \sigma)$ cu $m \in \mathbf{R}$ necunoscut şi $\sigma > 0$ cunoscut. Vom determina un interval de încredere pentru m cu o probabilitate de încredere 1-α dată şi cunoscând datele de selecţie $x_1, x_2, \ldots, x_n$.

Se arată că:

Propoziţia 1. Intervalul de încredere pentru media teoretică m a unei caracteristici care urmează legea normală a cărei abatere standard σ este cunoscută, este:

$$\bar{x} - \frac{\sigma}{\sqrt{n}} \cdot t < m < \bar{x} + \frac{\sigma}{\sqrt{n}} \cdot t \qquad (4)$$

unde:

$\bar{x} = \frac{1}{n}\sum_{=1}^{n} x_i$ este media de selecţie,

σ este abaterea standard (dată, cunoscută),

n este volumul eşantionului,

t se determină din tabela funcţiei *Laplace* $\Phi(t)$ din condiţia $2\Phi(t) = 1 - \alpha$, pentru probabilitatea de încredere $\varepsilon = 1 - \alpha$ precizată.

Exemplul 1. Pentru estimarea unui interval de încredere pentru temperatura medie în luna martie la Craiova, au fost înregistrate şi calculate temperaturile medii lunare în lunile martie de-a lungul a n=50 de ani (1950-1999) de la Staţia Meteorologică Craiova. Pentru această selecţie s-a obţinut media $\bar{x} = 4{,}46^{o}C$.

Cunoscând că variabila X (reprezentând media lunară a temperaturii medii în luna martie la Craiova) are distribuţia normală $N(m,\sigma)$ a cărei abatere standard a variabilei este σ=2,8 să se determine intervalul de încredere pentru media m cu probabilitatea de risc $\alpha = 0{,}05$.

Soluţie. Volumul selecţiei fiind suficient de mare (n=50>30), media de selecţie a variabilei X (reprezentând media lunară a temperaturii medii în luna martie la Craiova) are distribuţia normală *N(0,1).*

Datele problemei cunoscute şi calculate sunt:

$\bar{x} = 4{,}46^{o}C$, σ=2,8, *n=50.*

Apoi pentru $\alpha = 0{,}05$ găsim $2\Phi(t) = 1 - \alpha$ *=0,95*, $\Phi(t) = 0{,}475$, iar din tabela funcţie Laplace găsim *t=1,96.*

Astfel obţinem:

$$m_1 = \bar{x} - \frac{\sigma}{\sqrt{n}} \cdot t = 4{,}46 - \frac{2{,}8}{\sqrt{50}} \cdot 1{,}96 \approx 3{,}70\,,\; m_2 = \bar{x} + \frac{\sigma}{\sqrt{n}} \cdot t = 4{,}46 + \frac{2{,}8}{\sqrt{50}} \cdot 1{,}96 \approx 5{,}21.$$

Deci intervalul de încredere pentru m este $3{,}70 < m < 5{,}21$.

În concluzie în luna martie la Craiova, cu o probabilitate de 0,95 temperatura medie poate avea valori între $3{,}70^0$ C şi $5{,}21^0$ C.

Cazul 2. *Nu se cunoaşte abaterea standard σ a caracteristicii X.*

În cazul în care nu se cunoaşte abaterea standard σ, ea va fi estimată cu ajutorul datelor de selecție.

Ca estimator al dispersiei σ^2 vom lua dispersia de selecție modificată $\tilde{s}^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \overline{X})^2$.

În acest caz se arată că :

Se arată că:

Propoziția 2. Intervalul de încredere pentru media teoretică m a unei caracteristici care urmează legea normală a cărei abatere standard nu este cunoscută, este:

$$\left(\overline{x} - \frac{\tilde{s}}{\sqrt{n}} \cdot t, \overline{x} + \frac{\tilde{s}}{\sqrt{n}} \cdot t\right) \qquad (5)$$

unde:

$\overline{x} = \frac{1}{n}\sum_{=1}^{n} x_i$ este media de selecție,

$\tilde{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k}(x_i - \overline{x})^2}$ este estimatorul abaterii standard σ,

n este volumul eşantionului,

t se determină din tabela repartiției *Student* cu $k=n-1$grade de libertate din condiția $F(t) = 1 - \frac{\alpha}{2}$, corespunzătoare probabilității de risc α precizată (sau $F(t) = \frac{1+P}{2}$, unde P este nivelul de siguranță)

Exemplul 2. Fie X o variabilă aleatoare distribuită normal pentru care s-a realizat următorul eşantion:

i	1	2	3	4	5
x_i	39,75	40,25	39,5	39,50	40,25
i	6	7	8	9	10
xi	40,50	40,00	39,75	40,00	40,00

Se cere să se determine un interval de încredere pentru media m cu coeficientul de risc ⍰⍰=0,05.

Soluţie. Folosind valorile x_i , i=1,…,10, găsim :

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{399{,}50}{10} = 39{,}95\,;\ \tilde{s}^2 = \frac{1}{9}\sum_{i=1}^{10}(x_i - \overline{x})^2 = \frac{1}{9}\cdot 1{,}17 = 0{,}13\,;\ \tilde{s} = 0{,}36\,.$$

Corespunzător la k=n-1=9 grade de libertate, pentru ⍰=0,05,

P=1-⍰⍰⍰0,95, în tabelul repartiţiei *Student* se găseşte t=2,262.

Atunci aplicând formulele obţinem:

$$m_1 = \overline{x} - \frac{\tilde{s}}{\sqrt{n-1}}\cdot t = 39{,}95 - \frac{0{,}36}{\sqrt{9}}\cdot 2{,}262 = 39{,}89,$$
$$m_2 = \overline{x} + \frac{\tilde{s}}{\sqrt{n}}\cdot t = 39{,}95 + \frac{0{,}36}{\sqrt{9}}\cdot 2{,}262 = 40{,}22\ .$$

Deci intervalul de încredere este $(39{,}68;40{,}22)$.

Prin urmare, cu o probabilitate de 0,95 media m se găseşte în intervalul $(39{,}68;40{,}22)$.

2.4.3. Semnificaţia parametrilor unei repartiţii. Ipoteza nulă.

Rezultatele prelucrării datelor de selecţie relativ la o caracteristică a unei populaţii obţinute pe un eşantion din populaţia respectivă au un caracter relativ în sensul că ele nu coincid neapărat cu rezultatele ce s-ar obţine prin prelucrarea întregii populaţii. La fiecare extragere a unei probe dintr-o populaţie, se obţin alte valori pentru medie, varianţă, frecvenţe relative, coeficient de corelaţie, etc., valori care se abat mai mult sau mai puţin faţă de valorile adevărate ale parametrilor populaţiei.

Teoria probabilităţilor oferă însă proceduri pentru evaluarea rezultatelor prelucrărilor datelor selective, permiţînd o estimare, în termeni de probabilitate, a

marjei maxime de eroare ce se poate comite prin utilizarea mărimilor din eşantion în locul celor care caracterizează populaţia.

Ne putem astfel întreba dacă este sau nu semnificativă diferenţa dintre dintre un indice stabil (obţinut pe cale teoretică sau din cercetări anterioare) şi indicele rezultat din cercetarea unei probe sau dacă este sau nu semnificativă o diferenţă dintre indicii privind două sau mai multe probe.

Pentru a putea răspunde la problema semnificaţiei se formulează iniţial o ipoteză care în urma analizei va fi acceptată sau respinsă.

Frecvent se foloseşte *ipoteza nulă* (H_0) care constă în presupunerea că abaterea indicilor estimaţi faţă de parametri populaţiei este zero (nulă) :

Definiţia 4. Fie o repartiţie unidimensională caracterizată de o densitate de probabilitate $f(x,\lambda)$ dependentă de parametrul necunoscut λ. Ipoteza conform căreia λ are valoarea λ_0 , se notează

$$[H_0 : \lambda = \lambda_0]$$

şi poartă numele de *ipoteza nulă.*

Verificarea unei astfel de ipoteze statistice înseamnă supunerea acesteia unor probe, numite *teste de semnificaţie*, operaţii în urma cărora ipoteza se respinge sau se acceptă.

Respingînd ipoteza nulă, se acceptă semnificaţia abaterii respective. Acceptînd-o rezultă că nu există nici un temei pentru a accepta semnificaţia diferenţei.

În luarea deciziei privind respingerea sau acceptarea ipotezei nule sunt posibile următoarele situţii:

Realitatea (*necunoscută*) / Decizia	H_0 *FALSĂ*	H_0 *ADEVĂRATĂ*
RESPINGE H_0	*CORECTĂ*	*ERONATĂ*
ACCEPTĂ H_0	*ERONATĂ*	*CORECTĂ*

Datorită caracterului întâmplător al selecţiei, la verificarea ipotezei nule se vede că există întotdeauna riscul de a lua o decizie eronată. Sunt posibile două tipuri de erori în verificarea ipotezei nule:

Erori de genul I, prin care pe baza rezultatelor prelucrării datelor caracteristice eşantionului, *se respinge ipoteza nulă cînd ea este de fapt adevărată* la nivelul populaţiei ;

Erori de genul II, când, pa baza aceluiaşi eşantion, *se acceptă ipoteza nulă, ea fiind de fapt falsă la nivelul populației.*

Astfel se pot pune în evidentă următoarele probabilități:

- Probabilitatea de respingere a ipotezei nule deşi ea este adevărată (numită *probababilitatea comiterii erorii de ordin unu* sau *riscul de genul I*), notată cu α şi poartă numele de *probabilitate de transgresiune* sau *nivel de semnificație.* Pentru a reduce erorile de genul I trebuie respinse numai ipotezele care se realizează cu o probabilitate mai mică de 5%. În unele situații se resping ipotezele care se realizează cu o probabilitate mai mică de 1% sau 0,1%.
- Probabilitatea de acceptare a ipotezei nule deşi ea este falsă (numită *probabilitatea comiterii erorii de ordin doi* sau *riscul de genul II*), notată cu β. În practică se alege de obicei β=0,10 sau β=0,05.

  Probabilitățile celor două tipuri de erori sunt legate prin relația [illegible]1.

Orice decizie s-ar lua față de ipoteza nulă, totdeauna avem în față un risc, acesta fiind de aspecte contrare. Astfel, dacă ne propunem să micşorăm riscul de a respinge o ipoteză adevărată se micşorează probabilitatea [illegible] dar în acelaşi timp se măreşte probabilitatea [illegible], deci se măreşte riscul de a acepta o ipoteză falsă.

Alegând un test, prin mărirea volumului selecției putem micşora oricât de mult probabilitatea comiterii unei erori, dar nu totdeauna a ambelor. Impunând α (de regulă 0,01 sau 0,05), β rezultă ca o consecința şi invers. Nu se poate afirma care din aceste probabilități trebuie să fie mai mică, neexistând o regulă în această privința.

De exemplu, dacă dorim să verificăm un produs alimentar, produs ce urmează a fi livrat de un furnizor către un beneficiar, din punctul de vedere a unui anumit parametru (cum ar fi, de exemplu, compoziția unui unui anumit ingredient care peste un anumit grad de concentrare devine vătămător) comiterea unei erori de ordinul doi este mai gravă decât comiterea uneia de ordinul unu.

În controlul statistic al calității produselor riscul de genul I (α) mai poartă numele şi de *riscul furnizorului,* care are tot interesul ca probabilitatea de respingere a unui lot bun de produse să fie cât mai mică. Riscul de genul II (β) se mai numeşte şi *riscul beneficiarului*, care este de asemenea interesat ca probabilitatea acceptării unui lot necorespunzător să fie cât mai mică. O organizare rațională a controlului de recepție constă în alegerea de comun acord, de către beneficiar şi de către furnizor a unor riscuri cât mai mici.

Aplicație : **Teste de semnificație al mediei experimentale ale unei distribuții.**

Fie X o variabilă aleatoare definită pe o anumită populație, presupusă că urmează legea normală de parametri m şi □, a cărei de medie teoretică $m=M(X)$ este estimată prin media $\bar{x}$ a unui eşantion $x_1,x_2,\ldots,x_n$ de volum n extras din populația respectivă.

Ca şi în cazul determinării intervalelor de încredere pentru m, studiat în secțiunea precedentă, cercetarea semnificației lui $\bar{x}$ se face testând ipoteza nulă [H_0: $m=m_0$] şi se face considerând două cazuri după cum se cunoaşte, sau nu, abaterea standard □.

O asemenea decizie are intotdeauna la bază calculul intervalului de încredere ce corespunde unui prag de semnificație ales, adică ipoteza nulă se acceptă dacă parametrul aparține intervalului de încredere şi se respinge în caz contrar.

Cazul 1. ***<u>Se cunoaşte</u>*** *<u>abaterea standard □ a caracteristicii X.</u>*

Presupunem cunoscută abaterea standard □ a acestei variabile şi ne propunem să determinăm un interval de încredere pentru media teoretică m.

Testarea ipotezei [H_0 : $m=m_0$] constă în acceptarea sau respingerea intervalului de încredere pentru m cu probabilitatea de încredere $P=1-\alpha$ precizată (unde α este probabilitatea de transgresiune), adică ipoteza H_0 se acceptă dacă m□ $(\bar{x}-\frac{\sigma}{\sqrt{n}}\cdot t;\ \bar{x}+\frac{\sigma}{\sqrt{n}}\cdot t)$ şi se respinge în

caz contrar. Numărul $s_{\bar{x}}=\frac{\sigma}{\sqrt{n}}$ mai poartă numele de <u>*eroarea standard a mediei.*</u>

<u>Practic</u>:

- Cu ajutorul eşantionului $x_1,x_2,\ldots,x_n$ de volum n extras din
- populație se calculează media de selecție $\bar{x}=\frac{1}{n}\sum_{=1}^{n}x_i$.
- Cu valorile date sau calculate n,m_0,σ si $\bar{x}$ se determină valoarea $T_{calculat}=\frac{|\bar{x}-m_0|}{\sigma/\sqrt{n}}$.
- Pentru probabilitatea de încredere P se determină valoarea $t=T_{tabelar}$ din tabela funcției *Laplace Φ(t)* din condiția $2\Phi(t)=P$

pentru probabilitatea de încredere $P=1-\alpha$ precizată.

Se compată $T_{calculat}$ şi $T_{tabelar}$ trăgându-se concluziile:

- dacă $T_{calculat} < T_{tabelar}$, se admite ipoteza nulă H_0;
- dacă $T_{calculat} > T_{tabelar}$, se respinge ipotea nulă H_0.

Observaţia 1. Acest test se aplică atunci când abaterea standard teoretică σ este cunoscută dinainte iar valoarea lui $\bar{x}$ provine dintr-un eşantion de volum mare $(n > 30)$ extras dintr-o populaţie normal distribuită.

Exemplul 3. S-a stabilit experimental că nivelul colesterolului în organismul unui adult este o variabilă aleatoare normală cu dispersia $\sigma^2 = 48{,}4$. O selecţie aleatoare de n = 41 adulţi a dat un nivel mediu observat al colesterolului $\bar{x} = 213$.

Să se testeze ipoteza [H_0 : m = 200] la un nivel de semnificaţie $\alpha = 0{,}05$.

Soluţie. Datele problemei conduc la $n=41$, $\bar{x}=213$, $\sigma = \sqrt{48{,}4} = 6{,}96$ Volumul selecţiei fiind suficient de mare ($n=41>30$), putem aplica acest test şi avem: $T_{calculat} = \dfrac{|\bar{x} - m_0|}{\sigma/\sqrt{n}} = \dfrac{|213-200|}{6{,}96/\sqrt{41}} = \dfrac{13}{6{,}96} \cdot 6{,}4 \approx 11{,}95$.

Pentru P= 0,95 din egalitatea $2\Phi(t)=P$ găsim $\Phi(t)$=0,475 iar din tabel se deduce $t = T_{tabelar} = 1{,}96$.

Cum $T_{calculat} = 11{,}95 > 1{,}96 = T_{tabelar}$ rezultă că se respinge ipoteza [H_0: m=4,40] cu prag de siguranţă P=0,95.

Cazul 2. *Nu se cunoaşte abaterea standard σ a caracteristicii X.*

În cazul în care nu se cunoaşte abaterea standard σ, ea va fi estimată cu ajutorul datelor de selecţie.

Vom lua ca estimator σ abaterea standard modificată $\tilde{s}$.

Testarea ipotezei [H_0 : m=m_0] constă în acceptarea sau respingerea intervalului de încredere pentru m cu probabilitatea de încredere $P = 1-\alpha$ precizată (unde α este probabilitatea de transgresiune), adică ipoteza H_0 se acceptă dacă $m \in \left(\bar{x} - \dfrac{\tilde{s}}{\sqrt{n}} \cdot t, \bar{x} + \dfrac{\tilde{s}}{\sqrt{n}} \cdot t\right)$ şi se respinge în caz contrar.

Practic:

- Cu ajutorul eşantionului $x_1, x_2, \ldots, x_n$ de volum n extras din populație se calculează media de selecție $\bar{x} = \frac{1}{n}\sum_{=1}^{n} x_i$.
- Cu valorile date sau calculate $n, m_0, \tilde{s}$ si $\bar{x}$ se determină valoarea $T_{calculat} = \frac{|\bar{x} - m_0|}{\tilde{s}/\sqrt{n}}$.
- Pentru probabilitatea de încredere *P*, din tabela distribuției Student cu *k=n-1* grade de libertate se determină valoarea $T_{tabelar}$ ca fiind valoarea lui *t* pentru care $F(t) = \frac{1+P}{2}$.

Se compată $T_{calculat}$ şi $T_{tabelar}$ trăgându-se concluziile:

- dacă $T_{calculat} < T_{tabelar}$, se admite ipoteza nulă H_0;
- dacă $T_{calculat} > T_{tabelar}$, se respinge ipotea nulă H_0.

Observaţia 2. Acest test, numit *Testul "t",* se aplică atunci când abaterea standard teoretică σ nu este cunoscută dinainte iar valoarea lui $\bar{x}$ provine dintr-un eşantion de volum mic.

Exemplul 4. Fie *X* caracteristica unei populații statistice pentru care s-a realizat următorul eşantion:

i	1	2	3	4	5
x_i	39,75	40,25	39,5	39,50	40,25
i	6	7	8	9	10
xi	40,50	40,00	39,75	40,00	40,00

Pentru acest eşantion s-au calculat media de selecție şi abaterea medie pătratică de selecție $\bar{x}$ =39,95 şi $\tilde{s} = 0{,}36$.

Să se verifice ipoteza [H_0: *m*=40] pentru un prag de siguranță de *P*=0,95.

Soluție. Avem de verificat ipoteza [H_0: *m*=40]. Valorile calculate sunt : $\bar{x}$ =39,95şi $\tilde{s} = 0{,}36$, $T_{calculat} = \frac{|\bar{x} - m_0|}{\tilde{s}/\sqrt{n}} = \frac{|39{,}95 - 40|}{0{,}36/\sqrt{10}} = 0{,}4394$

Corespunzător la n-1=9 grade de libertate, pentru P=0,95, în tabelul repartiției *Student* se găseşte $T_{tabelar} = 2{,}262$.

Cum $T_{calculat} = 0{,}4394 < T_{tabelar} = 2{,}145$, rezultă că se admite ipoteza [H_0: m=40] cu un prag de siguranță de 95%.

2.5. ANALIZA ERORILOR DE MĂSURARE

2.5.1. Erori de calcul şi de măsurare.

Valorile numerice rezultate ca urmare a măsurării unor mărimi fizice în cadrul unui experiment pot fi afectate de erori.

Definiţia 1. Fie x valoarea reală a unei mărimi şi fie a o aproximație a sa (obținută prin calcul sau ca rezultat al unei măsurări).

Diferența

$$e_x = x - a \tag{1}$$

se numeşte *eroare absolută* de calcul sau de măsurare.

Raportul

$$\varepsilon_x = \frac{e_x}{x} \qquad \text{sau} \qquad \varepsilon_x = \frac{e_x}{e_x + a} \tag{2}$$

se numeşte *eroare relativă.*

Eroarea relativă este mai semnificativă în cazul în care eroarea absolută este în valoare absolută foarte mare sau foarte mică.

În funcție de cauzele producerii erorilor acestea se pot clasifica în:

1⁰. Erori grosolane. Aceste erori apar ca urmare a incorectitudinii efectuării măsurătorilor (neutilizarea corectă a instrumentelor sau a principiilor de folosire a acestora) sau ca rezultat al neatenției operatorului care efectuează măsurătoarea (înregistrând valori pe care le confundă sau altele decât cele observate).

Rezultatele ce conțin erori grosolane constau în abateri foarte mari, diferă esențial ca valoare de rezultatele celorlalte măsurători şi au probabilitate mică de apariție. De aceea aceste erori trebuie eliminate încă în procesul de măsurare. În caz contrar, dacă acest lucru nu a fost făcut, se foloseşte *Criteriul de excludere a erorilor grosolane* care va fi expus mai târziu.

2⁰. Erori sistematice. Sunt acele erori care nu variaza la repetarea masurarii în aceleasi conditii sau variaza în mod determinabil odata cu modificarea conditiilor de masurare. Ele se datoreaza unor cauze bine determinate, se produc întotdeauna în acelasi sens, au valoare constanta în marime si semn sau variază dupa o lege bine determinata si pot fi eliminate prin aplicarea unor corectii.

Erorile sistematice pot fi la rândul lor:

a). *Erori sistematice obiective*:

- *Erori de aparat (instrumentale)*-datorate unor caracteristici constructive ale aparatelor, incorectei etalonari, uzurii;
- *Erori de metodă*- aparute ca urmare a principiilor pe care se bazeaza metoda de masurare, a introducerii unor simplificari sau utilizarii unor relatii empirice;
- *Erori produse de factori externi (erori de influență)*- deosebit de greu de evaluat prin calcule, deoarece nu întotdeauna pot fi cunoscute cauzele si legile de variatie în timp a conditiilor de mediu (temperatura, presiunea, umiditatea, câmpuri magnetice, radiatii etc.). Pentru eliminarea lor se impune asigurarea conditiilor de mediu cerute de producator pentru instalatia de masurat.

b) *Erori sistematice subiective* (de operator), provenind din modul subiectiv în care operatorul apreciaza anumite efecte (coincidente de repere la citirea rezultatelor, intensitati luminoase etc.) si care tin de gradul sau de oboseala, de starea sa psihica sau de anumite deficiente ale organelor de perceptie.

3⁰. Erori aleatoare (întâmplatoare) sunt erorile de măsurate care au rămas după eliminarea tuturor erorilor grosolane şi sistematice apărute. Ele apar apar din cauza unei mulțimi de factori a căror influență individuală este neglijabilă, din care cauză nu există posibilitatea depistării şi înlăturării acestor influențe

Erorile aleatoare sunt inevitabile, nu sunt controlabile şi nu pot fi înlăturate din rezultatele individuale ale măsurătorilor.

Studiul influenței erorilor aleatoare se bazează pe cunoaşterea legilor lor de repartiție.

Se pune problema estimării adevăratei valori a unei mărimi măsurate pe baza rezultatelor mai multor măsurători. După măsurarea repetată a valorii x se obține un şir de valori, fiecare dintre ele conținând o anumită eroare necunoscută. Pe baza acestor măsurători se doreşte calcularea valorii aproximative a lui x cu o eroare cât mai mică posibil.

2.5.2. Repartiția erorilor aleatoare de măsurare.

Erorile aleatoare de măsurare sunt caracterizate de o lege de repartiție bine determinată care poate fi stabilită repetând de un număr mare de ori, în condiții identice, măsurarea unei anumite mărimi şi considerând numărul m de rezultate ale măsurărilor ce cad într-un anumit interval. Raportul $\frac{m}{n}$ dintre acest număr şi numărul n al tuturor măsurătorilor efectuate (numit *frecvența relativă de a cădea în intervalul considerat)* tinde către o anumită constatntă când numărul tuturor măsurătorilor n este suficient de mare. Acest lucru permite aplicarea teoriei probabilităților la studiul erorilor aleatoare de măsurare.

În modelul probabilistic teoretic erorile aleatoare $e_x = x - a$ se consideră ca variabile aleatoare Z ce pot lua orice valoare reală, iar fiecărui interval (z_1, z_2) îi corespunde un număr bine determinat care este probabilitatea ca variabila aleatoare să ia valori în acest interval, notată $P(z_1 < Z < z_2)$ şi care este chiar constanta care aproximează frecvența relativă ca erorile aleatoare să cadă în intervalul considerat:

$$\frac{m}{n} \approx P(z_1 < Z < z_2).$$

Cel mai adesea se consideră ca lege repartiție a erorilor aleatoare de măsurare repartiția normală a cărei densitate de repartiție este:

$$\varphi(x;\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}},$$

(3)

unde parametrul σ $(\sigma > 0)$ caracterizează precizia măsurătorilor.

Cum erorile de măsurare pot fi pozitive sau negative este de interes cazul în care acestea se află într-un interval simetric $(-z,z)$, $z>0$, caz în care probabilitatea din formula (3) se scrie:

$$P(-z<Z<z)=P(|Z|<z)=2\cdot\Phi\left(\frac{z}{\sigma}\right),$$

unde Φ este funcția integrală a lui Laplace ale cărei valori sunt date în tabele (vezi ANEXE, Tabelul 1).

Notând $\frac{z}{\sigma}=t>0$, deducem:

$$P(|Z|<t\sigma)=2\Phi(t). \qquad (4)$$

Trecând la probabilitatea evenimentului contrar, rezultă că probabilitatea ca eroarea aleatoare să depăşească limitele $\pm t\sigma$, $t>0$, este :

$$P(|Z|>t\sigma)=1-2\Phi(t) \qquad (5)$$

Pentru a uşura calculele, valorile probabilității $1-2\Phi(t)$ sunt date în tabela

(vezi ANEXE, Tabelul 2) pentru valorile lui $t\geq 2{,}5$.

Pentru $t=3$ avem $P(|Z|>3\sigma)=1-2\Phi(3)=0{,}0027$, deci evenimentul ca eroarea aleatoare Z să iasă în afara intervalului de limite $\pm 3\sigma$ poate fi considerat ca un eveniment practic imposibil.

Cu atât mai mult, pentru valori mari ale lui t probabilitatea dată de (5) este foarte mică, de exemplu:

$$P(|Z|>4\sigma)=1-2\Phi(4)=6\cdot10^{-5},\quad P(|Z|>5\sigma)=1-2\Phi(5)=6\cdot10^{-7}.$$

Pe baza acestor considerații suntem îndreptățiți să acceptăm următorul rezultat:

Regula de trei sigma *(pentru erorile aleatoare): Erorile aleatoare de măsurare sunt mărginite în valoare absolută de* 3σ.

Observaţia 1. Dacă erorile aleatoare $z=x-a$ urmează legea normală atunci rezultatele măsurătorilor $x=a+z$ au densitatea de probabilitate:

$$\varphi(x;a,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}} \qquad (6)$$

care este legea normală generală de parametrii a şi σ.

Legea de repartiție normală a erorilor reflectă *proprietatea de simetrie a erorilor* (erorile aleatoare de semne diferite se întrâlnesc la fel de des) şi *proprietatea de concentrare a erorilor* (erorile aleatoare de măsurare mici în valoare absolută apar mai frecvent decât cele mari).

Observaţia 2. Aşa cum am dovedit, parametrul σ reprezintă abaterea medie pătratică a legii normale iar pătratul său σ^2, dispersia acestei legi.

Pentru variabila aleatoare a erorilor aleatoare de măsurare care urmează legea normală (3), aceste valori se numesc *indicatori ai preciziei de măsurare* (parametrul σ poartă numele de *eroarea medie pătratică* sau eroarea standard iar pătratul său σ^2, *dispersia erorilor*).

În afara acestora se mai utilizează şi alți indicatori ai preciziei de măsurare, cum ar fi:

Eroarea probabilă:

$$\rho = 0{,}6745 \cdot \sigma , \quad \Phi(\rho) = 0{,}25 . \tag{7}$$

Eroarea medie absolută:

$$\vartheta = \int_{-\infty}^{\infty} |x| \cdot \varphi(x;\sigma) dx = \frac{2\sigma}{\sqrt{2\pi}} = 0{,}7979 \cdot \sigma . \tag{8}$$

Măsura preciziei:

$$h = \frac{1}{\sigma\sqrt{2}} = 0{,}7071 \cdot \sigma . \tag{9}$$

2.5.3. Eliminarea erorilor apărute neaşteptat.

Am văzut în prima secțiune că în cazul apariției în cadrul procesului de măsurare a unor erori grosolane acestea trebuiesc verificate eliminate pe cât posibil încă în cadrul acestui proces.

Dacă o astfel de verificare nu a fost făcută la timpul potrivit atunci problema eliminării unei valori "*apărute în mod neaşteptat*" se rezolvă pe baza comparării acestei

valori cu celelalte rezultate ale măsurătorilor. Presupunând că toate măsurătorile se fac cu acelaşi grad de precizie şi independente una de alta, formulăm criterii de eliminarea a unei erori grosolane după cum se cunoaşte sau nu eroarea medie pătratică σ a măsurătorilor.

1º. *Metodă de eliminare pentru σ cunoscut.*

Să notăm valoarea "apărută neaşteptat" prin x_* şi celelalte valori acceptate ale măsurătorilor cu $x_1, x_2, \ldots, x_n$. Fie $\bar{x}$ media aritmetică a acestor valori. Pentru raportul

$$t = \frac{|x_* - \bar{x}|}{\sigma\sqrt{(n+1)/n}} \qquad (10)$$

calculăm probabilitățile $1-2\Phi(t)$ din Tabelul 2 (ANEXE).

Aceasta ne va da probabilitatea ca raportul considerat să ia întâmplător o valoare mai mare sau egală cu *t*, condiționat de faptul că valoarea x_* nu reprezintă o valoare cu eroare grosolană (adică eroarea rezultatului este întâmplătoare).

Dacă probabilitatea calculată este foarte mică, atunci valoarea "apărută neaşteptat" se consideră a fi cu o eroare grosolană şi ea va fi exclusă din prelucrarea ulterioară a rezultatelor măsurătorilor.

Probabilitatea respectivă trebuie luată nici prea mică deoarece ar putea scăpa erori grosolane şi nici prea mare deoarece am putea exclude şi rezultate cu erori aleatoare, necesare pentru o prelucrare corectă a rezultatelor măsurătorilor.

De obicei se consideră următoarele nivele de excludere:

- Nivelul 5 % (se exclud erorile a căror probabilitate de apariție este sub 0,05);
- Nivelul 1 % (se exclud erorile a căror probabilitate de apariție este sub 0,01).

Pentru un nivel ales α (mic) al probabilității apariției "valorii neaşteptate", se consideră că valoarea x_* conține o eroare grosolană, dacă probabilitatea corespunzătoare raportului t dat de (10) satisface inegalitatea $1-2\Phi(t) < \alpha$. Spunem în aces caz că valoarea x_* conține o eroare grosolană cu nivelul de încredere P=1-□□

Valoarea *t=t(*P*)*, pentru care $1-2\Phi(t) = \alpha$, deci 2□*(t)*=P,se numeşte *valoare critică* a raportului (11) cu siguranța P. Astfel, dacă □□0,01 (nivel 1%) atunci P=0,99 iar valoarea critică *t=t(*P*)*=2,576 (Vezi ANEXE, Tabelul 6) şi de îndată ce raportul (11) depăşeşte această valoare critică, vom putea elimina valoarea "apărut neaşteptat" x_* cu siguranța 0,99.

Exemplul 1. *1.* Fie o serie de $n+1=41$ rezultate ale unor măsurători independente, efectuate cu eroarea medie pătratică $\sigma = 0{,}133$. În aceste măsurători s-a descoperit o valoare "apărută neaşteptat" $x_* = 6{,}866$, iar media aritmetică a celorlalte 40 de măsurători este $\bar{x} = 6{,}500$. Să se decidă dacă valoarea "apărută neaşteptat" conţine o eroare grosolană şi deci poate fi exclusă din prelucrările ulterioare.

Soluţie. Suntem în cazul de eliminare când se cunoaşte precizia de măsurare exprimată prin $\sigma = 0{,}133$. Pentru valorile menţionate calculăm raportul:

$$t = \frac{|x_* - \bar{x}|}{\sigma\sqrt{(n+1)/n}} = \frac{0{,}366}{0{,}133\sqrt{41/40}} \approx 2{,}72 .$$

Din Tabelul 2 din ANEXE, valorii t=2,72 îi corespunde probabilitatea $1 - 2\Phi(t) = 0{,}0066 < 0{,}007$. Prin urmare cu o siguranţă $P > 0{,}993$ se poate considera că valoarea $x_* = 6{,}866$ conţine o eroare grosolană şi se va exclude din prelucrarea ulterioară a rezultatelor măsurătorilor.

2⁰. Metodă de eliminare pentru σ necunoscut.

Dacă eroarea medie pătratică σ a măsurătorilor nu se cunoaşte dinainte, atunci aceasta se estimează pe baza rezultatelor măsurătorilor luând ca estimator al lui ?? abaterea standard modificată $\tilde{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

În acest caz se consideră raportul

$$t = \frac{|x_* - \bar{x}|}{\tilde{s}} \qquad (11)$$

care se compară cu valorile critice $t_n(P)$ din Tabelul 6 (ANEXE), pentru n dat şi

Dacă pentru un număr dat de observaţii n raportul (11) se află între două valori critice cu siguranţele P_1 şi P_2 $(P_1 < P_2)$, atunci cu o siguranţă a concluziei mai mare decât P_1 se poate considera că valoarea "apărută neaşteptat" conţine o eroare grosolană şi o vom elimina din prelucrarea ulterioară.

În caz contrar, dacă siguranţa concluziei se dovedeşte insuficientă, aceasta nu dovedeşte absenţa unei erori grosolane, ci numai faptul că nu avem suficiente motive ca să excludem valoarea "apărută neaşteptat".

Exemplul 2. Considerăm acceptate rezultatele a n măsurări independente de egală precizie, pentru care media aritmetică este $\overline{x} = 6{,}5$ iar abaterea standard modificată $\tilde{s} = 0{,}133$ şi fie cea de a $(n+1)$-a măsurare care conduce la valoare "apărută neaşteptat" $x_* = 6{,}866$. Să se decidă dacă valoarea "apărută neaşteptat" conţine o eroare grosolană şi deci poate fi exclusă, pentru următoarele cazuri:

a). n=40 şi nivelul de siguranţă P=0,99.

b). n=6 şi nivelul de siguranţă P=0,95.

Soluţie. a). Calculăm $t = \dfrac{|x_* - \overline{x}|}{\tilde{s}} = \dfrac{0{,}366}{0{,}133} \approx 2{,}75$.

Fie $n=40$ numărul de rezultate acceptate şi un nivel de siguranţă de P=0,99. Atunci pentru $n=40$ şi nivelul de siguranţă P=0,99 din *Tabelul 6* găsim valoarea critică $t_n(P)=2{,}742$.

Cum $t_n(P)=2{,}742 < t \approx 2{,}75$ rezultă că valoarea $x_* = 6{,}866$ se poate elimina cu o siguranţă a concluziei mai mare decât 0,99.

b). Presupunem acum $n=6$ şi nivelul de siguranţă P=0,95. În acest caz din *Tabelul 6* găsim $t_n(P) = 2{,}78$.

Cum $t_n(P) = 2{,}78 > t \approx 2{,}75$ rezultă că valoarea $x_* = 6{,}866$ nu se exclude cu o siguranţă mai mică decât P=0,95.

2.5.4. Estimări prin intervale de încredere pentru măsurări de precizii egale.

Presupunem că erorile aleatoare de măsurare se supun legii normale şi ne propunem a calcula estimaţiile adevăratei valori a a unei mărimi măsurate prin intervale de încredere simetrice de forma

$$\overline{x} - \varepsilon < a < \overline{x} + \varepsilon \ ,$$

sau

$$|a - \overline{x}| < \varepsilon \, , \qquad (12)$$

unde $\overline{x}$ este media aritmetică a măsurătorilor.

Mărimea ε se determină fixându-se nivelul de încredere (siguranţa estimaţiei) P la una din valorile 0,95 sau 0,99.

Estimarea cu ajutorul intervalelor de încredere se face în două situații după cum se cunoaşte sau nu eroarea medie pătratică σ (su o altă caracteristică a preciziei măsurărilor).

1⁰. Estimarea cu ajutorul intervalului de încredere în cazul cînd ***se cunoaşte precizia măsurărilor.***

Dacă se cunoaşte eroarea medie pătratică σ (su o altă caracteristică a preciziei măsurărilor) atunci intervalul de încredere (12) este de forma

$$|a - \bar{x}| < t(P)\frac{\sigma}{\sqrt{n}}, \tag{13}$$

unde n este numărul de măsurări iar valoarea $t=t(P,n-1)$ se determină fixându-se nivelul de încredere P cu ajutorul relației $2\Phi(t)=P$ (vezi ANEXE, Tabelul 2). Se obține astfel

$$\varepsilon = t(P)\frac{\sigma}{\sqrt{n}}. \tag{14}$$

Exemplul 3. Considerăm zece măsurări ale unei aceleiaşi mărimi fizice date în tabelul de mai jos :

x_i	35,6	35,9	36,1	36,2	36,6	Total
n_i	1	3	3	2	1	10

Presupunem că se cunoaşte precizia măsurărilor (abaterea standard σ=0,28). Se cere să se estimeze prin intervale de încredere adevărata valoare a mărimii măsurate a cu o siguranță P=0,99.

Soluție. Determinăm mai întâi media $\bar{x}$ folosind metoda zeroului fals. Astfel, completăm tabloul de mai jos, luând $x_0 = 36{,}1$ şi folosim metoda zeroului fals:

x_i	n_i	f_i	x_i-x_0	$f_i(x_i-x_0)$	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})^2 f_i$
35,6	1	0,1	-0,5	-0,05	-0,46	0,2116	0,02116
35,9	3	0,3	-0,2	-0,06	-0,16	0,0256	0,00768
36,1	3	0,3	0	0	0,04	0,0016	0,00048
36,2	2	0,2	0,1	0,02	0,14	0,0196	0,00392
36,6	1	0,1	0,5	0,05	0,54	0,2913	0,02913
Total	10			-0,04			0,06237

$$\bar{x} = x_0 + \sum_{i=1}^{k}(x_i - x_0)\cdot f_i = 36{,}1 - 0{,}04 = 36{,}06.$$

Dispersia empirică modificată este

$\tilde{s}^2 = \frac{n}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot f_i = \frac{10}{9} \cdot 0{,}06237 = 0{,}0693$, de unde pentru abaterea standard modificată găsim $\tilde{s} \approx 0{,}26$.

Pentru nivelul de încredere P=□□(*t*)=□,99, adică 1-P=□,01 găsim în Tabelul 2 (ANEXE) *t*(P)=2,576. Prin urmare cu siguranța de 0,99 aavem: $|a - \bar{x}| = |a - 36{,}06| < 2{,}576 \cdot \frac{0{,}28}{\sqrt{10}} = 0{,}23$.

În consecință intervalul de încredere care acoperă a cu încrederea 0,99 este: $(35{,}83;36{,}29)$.

*2⁰. <u>Estimarea cu ajutorul intervalului de încredere în cazul cînd **nu se cunoaşte** precizia măsurărilor.</u>*

Dacă eroarea medie pătratică σ a măsurătorilor nu se cunoaşte dinainte, atunci aceasta se estimează pe baza rezultatelor măsurătorilor luând ca estimator al lui □□ abaterea standard modificată

$$\tilde{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\ .$$

În acest caz intervalul de încredere (13) este de forma

$$|a - \bar{x}| < t(\mathrm{P}, k)\frac{\tilde{s}}{\sqrt{k}}, \qquad (15)$$

unde factorul *t*(P,*k*) depinde nu numai de nivelul de încredere P ci şi de număril *k*=*n-1* al gradelor de libertate (*n* fiind numărul măsurărilor). Valorile lui *t*(P,*k*) sunt date în Tabelul 5 (ANEXE), alcătuit cu ajutorul repartiției Student, adică funcția de repartiție a raportului $(\bar{x} - a)\sqrt{n} / \tilde{s}$; valoarea *t*(P,*k*) se determină astfel încât $P\left(\left|\frac{\bar{x} - a}{\tilde{s}\sqrt{n}} < t\right|\right) = \mathrm{P}$

<u>Exemplul 4.</u> Fie cele 10 măsurări date în problema precedentă, unde precizia măsurărilor nu este cunoscută (valoarea abaterii standard σ este necunoscută). Se cere să se estimeze prin intervale de încredere adevărata valoare a mărimii măsurate a cu o siguranță P=0,99.

<u>*Soluție*</u>. Pe baza rezultatelor măsurărilor am calculat în exemplul anterior media $\bar{x} = 36{,}06$. Cum eroarea medie pătratică σ a măsurătorilor nu se cunoaşte dinainte, atunci aceasta se estimează pe baza rezultatelor măsurătorilor luând ca estimator al lui său abaterea standard empirică modificată pentru care găsim $\tilde{s} \approx 0{,}26$.

Pentru nivelul de încredere P=0,99 şi numărul măsurărilor *n*=10, *k*=*n*-1=9,

determinăm din Tabelul 5 (ANEXE) $t(P,k)= t(0,99;,9)=3,250$. Prin urmare cu siguranța de 0,99 se poate considera

$$|a-\bar{x}| = |a-36,06| < 3,250 \cdot \frac{0,26}{\sqrt{9}} = 0,28 .$$

În consecință intervalul de încredere care acoperă a cu încrederea 0,99 este: $(35,78;36,34)$.

Observaţia 3. Dacă precizia măsurărilor nu este cunoscută intervalul de încredere dat de (16) nu se poate înlocui cu intervalul dat de (14) prin simpla substituire a abaterii standard σ cu abaterea standard empirică modificată $\tilde{s}$ (ceea ce ar reveni la înlocuirea factorului $t(P)$ cu factorul $t(P,k)$), deoarece în acest caz intervalul de încredere este semnificativ mai mare decât cel obținut când se cunoaşte precizia măsurărilor. Valoarea factorului $t(P,k)$ descreşte când numărul gradelor de libertate k creşte indefinit $(k \to \infty)$şi tinde către valoarea factorului $t(P)$. Într-adevăr în acest caz, am văzut că dacă numărul gradelor de libertate n tinde la infinit atunci distribuția Student $\varphi(x;n)$ tinde către distribuția normală normată $\varphi(t;0,1)$.

Observaţia 4. În practica prelucrării datelor obținute din măsurători în formulele (13) şi (15) care dau intervalele de încredere pentru cazurile în care **se cunoaşte** respectiv **nu se cunoaşte** precizia de măsurare (dar care se presupune că este aceeaşi pentru toate măsurătorile), se utilizează frecvent $t(P)=3$ respectiv $t(P,k)=3$ astfel că intervalele de încredere se scriu:

$$|a-\bar{x}| < 3\sigma / \sqrt{n} , \qquad (13')$$

(dacă precizia σ este cunoscută), respectiv

$$|a-\bar{x}| < 3\tilde{s} / \sqrt{n} , \qquad (15')$$

(dacă precizia σ este necunoscută, ea fiind înlocuită cu estimația $\tilde{s}$).

Prima dintre aceste estimații (13') are siguranța $2\Phi(3) = 0,9973 \approx 1-0,003$, independent de numărul măsurărilor.

Siguranța celei de-a doua estimații (15') depinde esențial de numărul n al măsurărilor. Dependența siguranței P de numărul n al măsurărilor pentru estimația (15') este dată în abelul de mai jos:

n	P	n	P
5	0,960	16	0,991
6	0,970	18	0,992
7	0,976	20	0,993

8	0,980	25	0,994
9	0,983	30	0,995
10	0,985	50	0,996
12	0,988	150	0,997
14	0,990	∞	0,9973

EXERCIŢII ŞI PROBLEME SUPLIMENTARE

1. Măsurătorile efectuate prin sondaj aleator asupra înălţimii a 50 de spice dintr-un lot de orz indică următoarele valori (în cm.) date în tabelul de mai jos:

Nr. crt	Înălţime	Nr. crt	Înălţime	Nr. crt	Înălţime	Nr. crt	Înălţime	Nr. crt	Înălţime
1	50,7	11	50,1	21	50.0	31	49,8	41	49,9
2	51,0	12	50,0	22	50,0	32	50,5	42	50,2
3	51,0	13	50,1	23	49,9	33	49,6	43	49,8
4	49,6	14	50,0	24	50,2	34	50,4	44	49,9
5	49,8	15	49,9	25	50,0	35	50,2	45	50,1
6	49,2	16	50,3	26	49,7	36	50,6	46	50,0
7	50,0	17	50,0	27	50,3	37	49,6	47	49,9
8	49,8	18	50,2	28	49,2	38	49,3	48	49,8
9	49,8	19	49,4	29	50,0	39	49,5	49	50,1
10	49,9	20	49,8	30	50,1	40	50,0	50	50,2

a). Să se facă gruparea datelor şi să se determine frecvenţele absolute, relative şi cumulate. Să se facă reprezentarea în bare.

b). Să se determine clase de valori de lungime 1 şi să se determine frecvenţele absolute şi relative ale intervalelor.

c). Să se reprezinte histograma, poligonul frecvenţelor şi poligonul frecvenţelor cumulate.

2. Să se reprezinte în batoane seria statistică dată de:

$$X = \begin{pmatrix} 3,7 & 3,8 & 3,9 & 4,0 & 4,1 & 4,2 & 4,3 & 4,4 & 4,5 & 4,6 & 4,7 & 4,8 \\ 1 & 3 & 7 & 10 & 15 & 22 & 16 & 13 & 6 & 4 & 2 & 1 \end{pmatrix}.$$

3. Distanțele(în km) parcurse cu 1 litru de carburant în cursul a 100 de probe realizate de un acelați vehicul (grupate în clase de amplitudine 0,2 km) sunt date în tabelul de mai jos:

Distanța (*d)* în km	Nr. de probe	Distanța (*d)* în km	Nr. de probe
$8,5< d \leq 8,7$	3	$9,5< d \leq 9,7$	20
$8,7< d \leq 8,9$	6	$9,7< d \leq 9,9$	16
$8,9< d \leq 9,1$	10	$9,9< d \leq 10,1$	9
$9,1< d \leq 9,3$	13	$10,1< d \leq 10,3$	4
$9,3< d \leq 9,5$	17	$10,3< d \leq 10,5$	2

a) Să se completeze tabelul

b)

Clase	Frecv. absolută a clasei *(n_i)*	Frecv. relativă a clasei*(f_i)*	Fecv.abs. cumulată crescător	Val.centra clasei $x_i^{(c)}$
(….;…]	……..	…….	…….	….

b). Să se reprezinte histograma şi poligonul frecvențelor.

4 Temperaturile medii înregistrate la Craiova în lunile *mai* ale anilor 1930-1979 sunt date în tabelul de mai jos:

Anul	0	1	2	3	4	5	6	7	8	9
1930…	8,1	4,0	-0,9	3,2	8,2	6,7	8,8	5,6	7,8	4,1

1940...	3,5	6,3	+0,4	4,3	3,8	6,4	6,4	8,2	5,9	0,3
1950...	5,5	6,9	-1,9	5,1	2,1	3,6	0,0	6,2	2,9	6,0
1960...	4,6	8,0	2,3	2,9	3,2	3,7	6,1	6,6	5,5	-0,1
1970...	5,2	3,6	5,5	3,0	4,9	7,7	3,1	7,2	5,8	6,3

a). Să se facă gruparea în clase, de mărime 2^oC cu convenția ca extremitatea dreaptă a fiecărei clase să nu aparțină clasei (ex. [-2,0;0), [0;2,0), [2,0;4,0), ...);

b). Să se completeze tabela obținută la punctul *a)* cu frecvențele absolute, cu frecvențele relative şi cu valoarea centrală a clasei;

c). Să se reprezinte histograma grupării în clase.

5. Cantitățile lunare de precipitații căzute la Craiova în lunile aprilie ale anilor 1930-1979 sunt date (în litri/m.p.) în tabelul următor

	1930...	1940...	1950...	1960...	1970...
0	55,5	92,0	24,8	39,4	64,4
1	19,6	36,5	40,0	49,4	42,5
2	17,8	33,7	40,8	75,6	16,4
3	7,8	26,9	23,5	33,7	42,6
4	89,0	42,3	52,2	62,6	74,0
5	32,7	35,4	94,3	57,9	43,8
6	22,6	16,3	31,6	65,8	47,1
7	45,3	22,8	65,3	49,5	50,2
8	57,1	37,6	51,4	8,7	31,6
9	28,1	3,9	19,3	31,9	42,7

a). Să se facă gruparea în clase, de mărime 10 litri/mp.

b). Să se completeze tabela obținută la punctul ***a)*** cu frecvențele absolute, cu frecvențele relative şi cu valoarea centrală a clasei;

c). Să se reprezinte histograma grupării în clase.

6. Vârsta indivizilor dintr-un grup de 30 de persoane este dată în tabelul original de mai jos:

20	26	26	30	35	35	37	37	37	37
39	41	45	45	45	48	48	48	50	50
54	55	57	57	60	60	65	65	69	70
21	22	24	32	32	43	40	41	40	42
42	45	52	53	52	54	59	61	62	66

Să se facă gruparea acestor date statistice pe 5 intervale de variație egale şi să se calculeze frecvențele absolute corespunzătoare şi să se reprezinte histograma şi poligonul frecvențelor.

7. Reprezentați circular seria statistică cu valorile date în tabelul:

Clasa	A	B	C	D	E
Frecvența absolută	48	32	24	16	12

8. Statistica naşterilor înregistrate lunar într-un anumit oraş de-a lungul a doi ani consecutivi s-a prezentat astfel:

Luna	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
Număr naşteri	6	8	9	13	18	15	20	24	19	12	11	7

Să se facă reprezentarea polară a seriei statistice.

9. Pentru seriile statistice din exercițiile (1)- (6), de la exercițiile şi problemele suplimentare din paragraful precedent să se calculeze media, valoarea modală, mediana, dispersia şi abaterea medie pătratică.

10. Distanțele(în km) parcurse cu l litru de carburant în cursul a 100 de

probe realizate de un acelați vehicul (grupate în clase de amplitudine 0,2 km) sunt date în tabelul de mai jos:

Distanța (*d)* în km	Nr. de probe	Distanța (*d)* în km	Nr. de probe
$8,5< d \leq 8,7$	3	$9,5< d \leq 9,7$	20

$8,7< d \leq 8,9$	6	$9,7< d \leq 9,9$	16
$8,9< d \leq 9,1$	10	$9,9< d \leq 10,1$	9
$9,1< d \leq 9,3$	13	$10,1< d \leq 10,3$	4
$9,3< d \leq 9,5$	17	$10,3< d \leq 10,5$	2

Să se completeze tabelul

Clase	Frecv. absolută (n_i)	Frecv. relativă (f_i)	Fecv.abs. cum. cresc.	Val.centra clasei (x_i)	$(x_i)^2$
(….;…]	……..	…….	…….	….	…..

şi să se determine media, clasa modală, mediana şi dispersia seriei statistice.

11. În şirul de valori de mai jos sunt prezentate valorile concentratiilor de colesterol în sânge, masurate în mg/dl, pentru un esantion de 50 de pacienti :

250 200 240 210 180 160 210 170 240 140 160 220 150 260 150 180 170 140 180 190 145 220 150 170 210 220 210 230 140 220 230 180 250 230 230 240 170 260 240 220 190 160 180 250 180 160 190 220 260 200

(a). Calculati media, mediana si valoarea modala.

(b). Calculati amplitudinea variatiei, dispersia si abaterea standard.

(c). Calculati frecventele absolute si relative ale valorilor variabilei aleatoare "concentratia colesterolului" si realizati histogramele frecventelor.

(d). *D*eterminati cuartilele Q_1, Q_2 si Q_3 care impart setul de date în patru sectiuni cu numar egal de valori: fiecare parte contine 25% din numarul total de date. Determinati intervalul intercuartilic.

12. Se consideră un eşantion de 20 de clienți, care intră într-un magazin alimentar, pentru a cerceta frecvența *X* cu care clienții fac apel la serviciile magazinului de-a lungul unei săptămâni şi respectiv pentru cercetarea cheltuielilor lunare Y în mii lei ale clienților, pentru procurarea de bunuri alimentare.

S-au obținut următoarele date de selecție pentru *X* şi respectiv *Y*:

X : 1, 2, 1, 4, 3, 2, 5, 6, 1, 2, 3, 2, 3, 4, 6, 2, 4, 3, 1, 2;

Y : 89, 90, 101, 88, 85, 77, 102, 100, 86, 97, 76, 121, 113, 110, 96, 92, 108, 112, 103, 109.

Să se calculeze:

a). Distribuțiile empirice de selecție pentru fiecare din caracteristici.

b). Mediile de selecție, momentele centrate de selecție de ordinul al doilea şi dispersiile de selecție pentru caracteristicile *X* şi *Y*.

13. La recepționarea unei mărfi ambalate în lăzi, care trebuie să aibă greutatea netto de câte 50 kg, s-a efectuat un control prin sondaj, cântărindu-se la întâmplare 15 lăzi, găsindu-se următoarele greutăți:

Nr. Crt	x_i *(kg)*	Nr. Crt	x_i *(kg)*	Nr. Crt	x_i *(kg)*
	49,75	6	50,50	11	19,25
2	50,25	7	50,00	12	1925
3	49,50	8	49,75	13	19,50
4	49,50	9	50,00	14	20,00
5	50,25	10	50,00	15	19,50

Să se calculeze media , dispersia şi abaterea standard.

14. Să se calculeze indicatorii statistici ai seriei statistice de mai jos reprezentând punctajele obținute de un număr de 82 intervievați la un test de aptitudini:

Punctajul obținut	Număr persoane
40 –50	8
50 – 60	14
60 - 70	18
70 – 80	23
80 – 90	12
90 –100	7
Total	82

15. Măsurarea înălțimii X (în cm) şi a greutății Y (în kg) pentru 70 de persoane a condus la distribuția următoare :

Y / X	48-56	56-64	64-72	72-80
160-165	16	8	1	0
165-170	1	10	4	1
170-175	0	4	8	2
175-180	0	1	5	9

a). Considerând pentru fiecare clasă a fiecărei variabile valoarea centrală a clasei să se scrie distribuția corespunzătoare şi pornind de la aceasta să se facă schimbările de variabile (folosind metoda « zeroului fals ») $T = \frac{X - 167{,}5}{5}$, $Z = \frac{Y - 60}{8}$, să se scrie tabelul de corelație al noii distribuții bidimensionale (T,Y) calculând frecvențele marginale.

*b).*Să se calculeze pentru fiecare variabilă mediile şi dispersiile.

c). Să se calculeze covarianța variabilelor X şi Y precum şi coeficientul de corelație.

16. În urma efectuării a 10 măsurători asupra două caracteristici X şi Y ale unei populații, s-au găsit valorile date în tabelul de mai jos:

Proba	**1**	**2**	**3**	**4**	**5**
X: x_i	46,3	46,7	43,6	44,8	47,1
Y: y_i	54,0	52,2	55,5	57,1	54,3
Proba	**6**	**7**	**8**	**9**	**10**
X: x_i	39,6	37,9	39,5	40,8	42,4
Y: y_i	63,2	70,1	70,2	71,8	72,4

Să se reprezinte corelograma şi să se determine covarianța şi coeficientul de corelație al variabilelor X şi Y.

17. Fie caracteristicile *X* şi *Y* reprezentând în procente suprafaţa comercială de expunere a mărfurilor spre vânzare faţă de suprafaţa construită şi respectiv volumul valoric al vânzărilor, raportat la metru pătrat suprafaţă de prezentare a mărfurilor pe lună, în mii lei acestea fiind cunoscute prin următoarele date de selecţie:

X	10	12	15	17	26
Y	40	45	42	53	60

Se cere:

a). Să re reprezinte punctele corelograma seriei şi să se determine coeficientul de corelaţie al variabilelor *X* şi *Y*;

b). Să se determine dreapta de regresie a lui *Y* faţă de *X* şi să se facă prognoza volumului valoric al vânzărilor *Y* când *X* ia valoarea 30.

18. Măsurându-se greutatea şi înălţimea la 10 copii în vârstă de 12-14 ani s-au obţinut datele din tabelul următor:

Nr. crt	1	2	3	4	5	6	7	8	9	10
Greutatea (Kg)	32	33	34	36	40	41	45	47	49	50
Înălţimea (cm)	132	135	139	144	142	147	154	153	156	160

Se cere:

a). Să re reprezinte corelograma seriei.

b). Să se determine covarianţa şi coeficientul de corelaţie al lui *X* şi *Y.*

c). Să se determine dreapta de regresie a lui *Y* asupra lui *X* şi să se determine prognoza asupra lui *Y* când *X=30.*

19. Corespondenta dintre valoarea hemoglobinei glicozilată (HbA1C) si media glicemiilor pe ultimele 3 luni este urmatoarea, dupa **ghidurile** Asociatiei Americane de Diabet sunt date în seria statistică bidimensională de mai jos:

H =hemoglobina glicozilată (%)	6	7	8	9	10	11	12
G=Media glicemiilor (mg/dl)	135	170	205	240	275	310	345

a). Să re reprezinte corelogram seriei. Şi să se calculeze determine covarianţa şi coeficientul de corelaţie al variabilelor H şi G.

b). Să se determine dreapta de regresie a lui G asupra lui H şi cu ajutorul dreptei de regresie să se facă prognoza asupra mediei glicemiei pe ultimele trei dacă unui pacient i se determină hemoglobina glicozilată de 7,1%.

20. Se consideră caracteristica X ce reprezintă greutatea în grame a unor pachete încărcate automat de un dispozitiv. Pentru verificarea normalității lui X, se consideră o selecție de volum n = 100, datele de selecție fiind următoarele:

Greutatea	47–48	48–49	49–50	50–51	51–52	52 – 53
Frecvența	12	18	22	21	19	8

Se cere:

a). Aplicarea testului χ^2, cu nivelul de semnificație α = 0,05, pentru verificarea normalității lui *X*;

b). Aplicarea testului lui Kolmogorov, cu nivelul de semnificație α = 0,05, pentru verificarea normalității lui *X*.

21. Nivelul de calciu în sângele unui adult este în medie 9,5 mgr/decilitru şi $\sigma = 0{,}4$. O clinică măsoară nivelul calciului la 160 de pacienți tineri şi găseşte $\bar{x} = 9{,}3$. Verificați ipoteza $H_0 : m = 9{,}5$ față de $H_1 : m \neq 9{,}5$.

22. La o anumită stație meteorologică media multianuală a cantităților de precipitații înregistrate a fost de 440 mm. În ultimii 10 ani precipitațiile au fost mai reduse, media fiind de 400 mm. Pentru aceeaşi perioadă s-a calculat abaterea standard $\tilde{s} = 20\,\text{mm}$. Cât de semnificativă este diferența dintre cele două medii?

23. Se iau eşantioane din apa rezultată din răcirea la o centrală nucleară. Se consideră că dacă temperatura apei evacuate nu depăşeşte 60^0C nu constituie o primejdie pentru mediul înconjurător.

Se aleg 70 eşantioane de apă şi se măsoară temperatura fiecărui asemenea eşantion. Se obțin rezultatele:

Temperatura în ^{0}C	52	54	58	61	64	65
Frecvența	14	21	18	10	5	2

Să se verifice ipoteza $H_0 : m = 60^0 C$ față de $H_1 : m \neq 60^0 C$ la nivelul de semnificație $\alpha = 0{,}01$.

24. Fie o serie de *n+1*=41 rezultate ale unor măsurători independente, efectuate cu eroarea medie pătratică $\sigma = 0{,}32$. În aceste măsurători s-a descoperit o valoare

“apărută neaşteptat” $x_* = 2{,}65$, iar media aritmetică a celorlalte 40 de măsurători este $\bar{x} = 2{,}52$. Să se decidă dacă valoarea “apărută neaşteptat” conține o eroare grosolană şi deci poate fi exclusă din prelucrările ulterioare.

25. Considerăm rezultatele a n măsurări independente de egală precizie, pentru care media aritmetică este $\bar{x} = 2{,}52$ iar abaterea standard modificată $\tilde{s} = 0{,}32$ şi fie cea de a ($n+1$)-a măsurare care conduce la valoare “apărută neaşteptat” $x_* = 2{,}65$. Să se decidă dacă valoarea “apărută neaşteptat” conține o eroare grosolană şi deci poate fi exclusă din prelucrările ulterioare , pentru următoarele cazuri:

a). n=40 şi nivelul de siguranță P=0,99.

b). n=6 şi nivelul de siguranță P=0,95.

26. Considerăm 20 măsurări ale unei aceleiaşi mărimi fizice date în tabelul de mai jos :

x_i	12,6	12,8	13,2	13,6	14,0	14,4	14,6
n_i	1	3	4	5	4	2	1

Presupunem că se cunoaşte precizia măsurărilor (abaterea standard σ=0,57). Se cere să se estimeze prin intervale de încredere adevărata valoare a mărimii măsurate a cu o siguranță P=0,99.

27. Fie cele 20 măsurări date în problema precedentă, unde precizia măsurărilor nu este cunoscută (valoarea abaterii standard σ este necunoscută). Se cere să se estimeze prin intervale de încredere adevărata valoare a mărimii măsurate a cu o siguranță P=0,99

ANEXE

Integrale improprii. Integrale improprii remarcabile.

Integralele improprii sunt integrale ale funcțiilor definite pe intervale necompacte de una din formele : $[a,b)$, $(a,b]$, (a,b), $[a,\infty)$,

$(-\infty, a]$, (a,∞), $(-\infty, a)$.

În cele ce urmează vom considera funcții definite pe intervale necompacte de forma $[a,b)$.

O funcție f:$[a,b)$ *cu valori in* $\boldsymbol{R}$, care este Riemann-integrabilă pe orice subinterval compact din $[a,b)$ se mai numeşte *local integrabilă* pe $[a,b)$.

Pentru o astfel de funcție limita $\lim\limits_{\substack{\alpha\to b\\ \alpha<b}} \int_a^{\alpha} f(t)dt \overset{not.}{=} \int_a^b f(x)dx$, se numeşte *integrala improprie* a lui f pe intervalul necompact $[a,b)$.

Dacă această limită există şi este finită, spunem că integrala improprie este *convergentă* sau că funcția f este *integrabilă în sens generalizat* pe $[a,b)$.

Fie $f: [a,b)\times J$ o funcție reală de două variabile x şi t (x parcurgând intervalul $[a,b)$ şi t parcurgând intervalul $J\subset\boldsymbol{R}$). Dacă funcția f este integrabilă (în sens impropriu) în raport cu variabila x pe $[a,b)$ adică există $\lim\limits_{\substack{\alpha\to b\\ \alpha<b}} \int_a^{\alpha} f(x,t)dx \overset{not.}{=} \int_a^b f(x,t)dx$ pentru orice $t\in J$, atunci funcția $F: J\to\boldsymbol{R}$, $F(t)=\int_a^b f(x,t)dx$, se numeşte

integrală improprie depinzând de paramatrul t.

Mai spunem în acest caz că integrala improprie depinzând de parametru este simplu convergentă.

În condiții suplimentare privind convergența integralei improprii şi propriatăți de derivabilitate sau integrabilitate ale funcțiilor au loc formulele :

(1). $F'(t)=\int_a^b \frac{\partial f}{dt}(x,t)dx$, adică $\boxed{\left(\int_a^b f(x,t(dx\right)'_t = \int_a^b \frac{\partial f}{dt}(x,t)dx}$

numită *formula de derivare sub semnul integrală în raport cu parametrul a integralei cu parametru.*

(2). Dacă J=[c,d] atunci $\int_c^d F(t)dt = \int_a^b \left(\int_c^d f(x,t)dt \right) dx$ adică:

$$\boxed{\int_c^d \left(\int_a^b f(x,t)dx \right) dt = \int_a^b \left(\int_c^d f(x,t)dt \right) dx}$$

numită *formula de schimbare a uodinii de integrare.*

1°. *Integrala lui Laplace-Poisson :*

$$\boxed{I = \int_0^{+\infty} e^{-x^2} dx = \frac{\sqrt{\pi}}{2}}.$$

Alte forme utile :

a). $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-x^2/2} dx = \frac{1}{2}$; *b).* $\frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-x^2/2} dx = \frac{1}{2}$.

2°. *Funcţia « GAMMA » a lui Euler.*

Se numeşte funcția GAMMA (sau funcția lui Euler de speța a doua) integrala improprie cu parametru:

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

Proprităţile funcţiei Gamma*:*

(a). $\Gamma(p+1)=p\,\Gamma(p)$, $\forall\ p>0$.

(b). $\Gamma(1)=0!$, $\Gamma(2)=1!$, $\Gamma(3)=2!$, $\Gamma(4)=3!\ldots$, $\Gamma(n+1)=n!$, $\forall\ n\in N$;

(c). $\Gamma\left(\frac{1}{2}\right) = 2 \cdot I = \sqrt{\pi}$..

Tabelul 1*. Valorile funcţiei integrale a lui Laplace-Poisson*

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt, \; \Phi(-x) = -\Phi(x)$$

x	Sutimi de x									
	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0040	0080	0120	0160	0199	0239	0279	0319	0359
0,1	3098	0438	0478	0517	0557	0596	0636	0675	0714	0753
0,2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0,3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0,4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0,5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0,6	2257	2291	2324	2357	2389	2422	2454	2486	2517	2549
0,7	2580	2611	2642	2673	2703	2734	2764	2794	2823	2852
0,8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0,9	3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1,0	3413	3437	3461	3485	3508	3531	3554	3577	3599	3621
1,1	3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1,2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1,3	4030	4049	4066	4082	4099	4115	4131	4147	4162	4177
1,4	4192	4207	4222	4263	4251	4265	4279	4292	4036	4319
1,5	4332	4345	4357	4370	4382	4394	4409	4418	4429	4441
1,6	4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1,7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1,8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1,9	4713	4719	4726	4732	4738	4744	4750	4756	4761	4767
2,0	4772	4778	4783	4788	4793	4798	4803	4808	4812	4817
2,1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2,2	4861	4864	4868	4871	4875	4878	4881	4884	4887	4890
2,3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2,4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2,5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2,6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2,7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2,8	4974	4975	4976	4977	4977	4978	4979	4979	4980	4981
2,9	4981	4982	4982	4983	4984	4984	4985	4985	4986	4986

***Tabelul 2**. Valori legate de funcţia $\Phi(t)$*

- Probabilitatea $P = P(|X - m| < k\sigma) = 2 \cdot \Phi(k)$;
- Funcția $k=k(P)$, inversa funcției $P=2\cdot\Phi(k)$.

k	$\Phi(k)$	$1-2\cdot\Phi(k)$	$\alpha=1-P$	$k=k(P)$	$P=1-\alpha$
2,5	0,49379	0,01242	0,05	1,960	0,95
2,6	0,49534	0,00932	0,04	2,054	0,96
2,7	0,49563	0,00693	0,03	2,170	0,97
2,8	0,49744	0,00511	0,02	2,326	0,98
2,9	0,49813	0,00373	0,01	2,576	0,99
3,0	0,49865	0,00270	0,009	2,612	0,991
3,1	0,49903	0,00194	0,008	2,652	0,992
3,2	0,49931	0,00137	0,007	2,697	0,993
3,3	0,49952	0,00097	0,006	2,748	0,994
3,4	0,49966	0,00067	0,005	2,807	0,995
3,5	0,499767	0,000465	0,004	2,878	0,996
3,6	0,499841	0,000318	0,003	2,968	0,997
3,7	0,499892	0,000216	0,002	3,090	0,998
3,8	0,499927	0,000145	0,001	3,291	0,999
3,9	0,499952	0,000096	0,0009	3,320	0,9991
4,0	0,499968	0,000063	0,0008	3,353	0,9992
4,1	0,499979	0,000041	0,0007	3,390	0,9993
4,2	0,499987	0,000027	0,0006	3,432	0,9994
4,3	0,499991	0,000017	0,0005	3,481	0,9995
4,4	0,499995	0,000011	0,0004	3,540	0,9996
4,5	0,4999966	0,0000068	0,0003	3,615	0,9997
4,6	0,4999979	00000041	0,0002	3,720	0,9998
4,7	0,4999987	0,0000025	0,0001	3,891	0,9999
4,8	0,4999992	0,0000016	0,00001	4,417	0,99999
4,9	0,4999995	0,0000009	0,000001	4,892	0,999999
5,0	0,4999997	0,0000006	0,0000001	5,327	0,9999999

Tabelul 3. Valorile $\chi^2_{tabelar}$ (*Testul* χ^2) pentru pragul de siguranţă α şi numărul gradelor de libertate ν.

ν	$\alpha=0,05$	$\alpha=0,01$
1	3,84	6,63
2	5,99	9,21

3	7,81	11,3
4	9,49	13,3
5	11,1	15,1
6	12,6	16,8
7	14,1	18,5
8	15,5	20,1
9	16,9	21,7
10	18,3	23,2
11	19,7	24,7
12	21,0	26,2
13	22,4	27,7
14	23,7	29,1
15	25,0	30,6
16	26,3	32,0
17	27,6	33,4
18	28,9	34,8
19	30,1	36,2
20	31,4	37.6
21	32,7	38,9
22	33,9	40,3
23	35,2	41,6
24	36,4	43,0
25	37,7	44,3
26	38,9	45,3
27	40,1	47,0
28	41,3	48,3
29	42,3	59,6
30	43,8	50,9

Tabelul 4. Valorile *t=t(P,k*) corespunzătoare nivelului de încredere P şi numărului *k=n-1* grade de libertate (*Repartiţia Student*)

P / *k*	0,95	0,99
4	2,776	4,604
5	2,571	4,032
6	2,447	3,707
7	2,365	3,499
8	2,306	3,355

9	2,262	3,250
10	2,228	3,169
11	2,201	3,106
12	2,179	3,055
13	2,160	3,012
14	2,145	2,997
15	2,131	2,947
16	2,120	2,921
18	2,103	2,878
20	2,086	2,845
25	2,060	2,787
30	2,042	2,750
35	2,030	2,724
40	2,021	2,704
45	2,014	2,689
50	2,008	2,677
60	2,000	2,660
70	1,995	2,648
80	1,990	2,639
90	1,987	2,632
100	1,984	2,626
∞	1,960	2,576

Tabelul 5. Valorile critice $t_n(P)$ comparate cu raportul $t = \frac{|x_* - \bar{x}|}{\tilde{s}}$ pentru înlăturarea valorilor „excepţionale” (*n* este numărul rezultatelor acceptate, iar *P* nivelul de încredere).

P / *n*	0,95	0,99
5	3,04	5,04
6	2,78	4,36
7	2,62	3,96
8	2,51	3,71
9	2,43	3,54
10	2,37	3,41
11	2,33	3,31
12	2,29	3,23
13	2,26	3,17
14	2,24	3,12

15	2,22	3,08
16	2,20	3,04
17	2,18	3,01
18	2,17	2,98
20	2,145	2,932
25	2,105	2,852
30	2,079	2,802
35	2,061	2,768
40	2,048	2,742
45	2,038	2,722
50	2,030	2,707
60	2,018	2,683
70	2,008	2,667
80	2,003	2,655
90	1,998	2,646
100	1,994	2,639
∞	1,960	2,576

Pentru valori n>100 valorile critice t_n(P) se pot calcula cu o precizie de pînă la 10^{-3} cu relaţia: $t_n(P) = t_\infty(P) + \frac{t_{100}(P) - t_\infty(P)}{n} \cdot 100$

B I B L I O G R A F I E

[1]. *BĂLAN, V.-Matematici Superioare*, Editura UNIVERSITARIA, Craiova, 2006.

[2]. *BĂLAN, V., BURADA, D.C. -Capitole de Matematici Superioare*, Editura ARVES, Craiova, 2010.

[3].*BĂLAN, V., BURADA, D.C. - Matematică şi Statistică*, Editura ARVES, Craiova, 2011.

[4].*BĂLAN, V., BURADA, D.C. – Teme de Matematici Superioare*, Editura UNIVERSITARIA, Craiova, 2013.

[5].*CEAPOIU, N.- Metode statistice aplicate în experimente agricole şi biologice*, Editura AGRO-SILVICĂ, Bucureşti, 1968**.**

[6].*CENUŞA GH. – Teoria Probabilităţilor- www.ase.ro, biblioteca digitală, cursuri în format digital.*

[7].*CENUŞA GH.* şi col – *Matematici pentru economişti- www.ase.ro, biblioteca digitală, cursuri în format digital.*

[8]. *CRAIU, V- Verificarea ipotezelor statistice,* EDP, Bucureşti, 1972.

[9].*CHIRIŢĂ, S.- Probleme de matematici superioare*, Editura Didactică şi Pedagogică, Bucureşti, 1989.

[10].*FAZLOLLAH, R.***-** *Spaţii Liniare.* Editura Didactică şi Pedagogică, Bucureşti, 1973.

[11].*HARTIA, S. – Programarea liniară în conducerea fermei agricole*. Editura CERES, Bucureşti, 1975.

[12].*IONESCU, H., DINESCU, C, SĂVULESCU, B.- Probleme ale Cercetării Operaţionale*. E.D.P. Bucureşti 1972.

[13].*MIHĂILĂ, N., POPESCU, O. – Matematici speciale aplicate în economie.* E.D.P., Bucureşti 1978.

[14].*MIHOC, GH., MICU, N.-Introducere în teoria probabilităţilor,* Editura Tehnică, Bucureşti, 1970.

[15].*RAFFIN, C.-Statistiques et Probabilités,* Collection *FLASH U, Armand Colin Editeur, Paris, 1993*

[16].*REISCHER, C., SÂMBOAN, A. –Culegere de probleme de teoria probabilităţilor şi statistică matematică*- E.D.P., Bucureşti, 1972.

[17].*RUMŞINSKI,L.Z.- Prelucrarea matematică a datelor experimentale*. Editura Tehnică, Bucureşti, 1974.

[18].*VLADIMIRESCU, I.- Statistică Matematică* Editura UNIVERSITARIA, *Craiova 1998.*