UNIVERSITATEA TEHNICA DE CONSTRUCTII
BUCURESTI

PROBABILITATI SI STATISTICA

Viorel PETREHUS, Sever-Angel POPESCU

BUCURESTI 2005

Cuprins

Cuvânt înainte

Cursul de faţă a fost scris în perioada 1996-1997 de către Viorel Petrehuş (partea I, probabilităţi) şi Sever-Angel Popescu (partea a II-a, statistică) pentru studenţii anului II din Universitatea Tehnică de Construcţii Bucureşti şi a apărut în 1997 multiplicat în atelierele universităţii. El a fost gândit în 14 lecţii, câte una pe săptămână, pe parcursul unui semestru. Fiecare lecţie se încheie cu exerciţii.

Autorii sunt recunoscători tuturor celor care au contribuit cu observaţiile lor la buna organizare a materialului prezentat.

Autorii

Partea I

Probabilităţi

Lecţia 1

Definiţia probabilităţii

1.1 Definiţia clasică

Probabilitatea a fost privită fie dintr-un punct de vedere "psihologic" ca măsurând gradul de siguranţă al observatorului relativ la producerea sau neproducerea unui fenomen, fie "statistic" ca frecvenţa de apariţie a unui fenomen într-un număr mare de experimente independente. Din punct de vedere clasic, definiţia care s-a dovedit cea mai eficientă în calcule a fost aceea care a plecat de la conceptul de egală posibilitate. Acest lucru înseamnă că numărul de posibilităţi într-un experiment este finit şi toate posibilităţile au aceeaşi şansă.

Probabilitatea unui eveniment care constă din mai multe asemenea posibilităţi este raportul dintre numărul cazurilor favorabile şi numărul cazurilor posibile. Utilizarea acestei definiţii presupune că într-un fel sau altul putem număra stările posibile şi pe cele favorabile.

Exemplul 1.1 *Se aruncă un zar de două ori.Să se determine probabilitaţile evenimentelor:*
a) Suma celor doua zaruri este 6.
b) Ambele zaruri au acelasi numãr.

Soluţie. Cazurilor posibile in cele două situaţii sunt (1,1), (1,2), ... (6,6), în număr de 36. Pentru punctul a) numai cazurile (1,5), (2,4), (3,3), (4,2), (5,1) sunt favorabile. Probabilitatea este deci

$$p = \frac{nr.\, cazuri\, favorabile}{nr.\, cazuri\, posibile} = \frac{5}{36}$$

Analog, probabilitatea celui de-al dolea eveniment este p=6/36=1/6.

Exemplul 1.2 *Dintr-un pachet de 36 cărţi se extrag trei la întâmplare. Care este probabilitatea ca cel puţin o carte să fie as?*

Soluţie. Numărul cazurilor posibile este C_{36}^3=7140. Numărul cazurilor favorabile este

$$\overbrace{4 \cdot C_{32}^2}^{un\, as} + \overbrace{C_4^2 \cdot C_{32}^1}^{doi\ \text{aşi}} + \overbrace{C_4^3 \cdot 1}^{trei\ \text{aşi}} = 2180$$

Probabilitatea este deci $p = 2180/7140 \approx 0,30532$..

În exemplul următor vedem cît de dificil este uneori să numărăm cazurile posibile sau cazurile favorabile.

Exemplul 1.3 *O persoană scrie n scrisori către n persoane distincte, le pune în n plicuri şi apoi scrie adresele la întîmplare. Care este probabilitatea ca cel puţin o scrisoare să ajungă la destinatarul potrivit?*

Numărul cazurilor posibile de a scrie adresele este $n!$. Enumerarea cazurilor favorabile este mai dificilă. După o mică dezvoltare a calculului probabilităilor se poate rezolva elegant aceată problemă (exerciţiul 5).

Exemplul 1.4 *Care este probabilitatea ca luînd un punct la întîmplare în pătratul [0,1]*$\times[0,1]$ *el să fie deasupra bisectoarei y=x?*

În acest caz numărul cazurilor posibile şi favorabile este infinit şi definiţia clasică nu se mai aplică.

1.2 Definiţia axiomatică a probabilităţii

În general evenimentele se exprimă prin propoziţii. Propoziţiile obţinute prin operaţiile logicii matematice (sau,şi,non), între propoziţii care exprimă evenimente, exprimă la rândul lor alte evenimente. În cele ce urmează probabilitatea este definită pe o multime Ω de evenimente care odată cu evnimentele A şi B conţine şi evenimentele următoare exprimate prin operaţiile logice sau, şi, non:

$$\begin{array}{lr} A \wedge B & A \text{ si } B \\ A \vee B & A \text{ sau } B \\ \overline{A} & \text{non } A \\ 1 & \text{evenimentul sigur} \\ 0 & \text{evenimentul imposibil} \end{array}$$

De asemenea, presupunem că au loc următoarele relaţii:

a)Comutativitatea

$$A \wedge B = B \wedge A \quad A \vee B = B \vee A$$

b)Asociativitatea

$$A \vee (B \vee C) = (A \vee B) \vee C$$
$$A \wedge (B \wedge C) = (A \wedge B) \wedge C$$

c)Distributivitatea

$$A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C)$$
$$A \vee (B \wedge C) = (A \vee B) \wedge (A \vee C)$$

d) Absorbţia

$$(A \cap B) \cup A = A \quad (A \cup B) \cap A = A$$

e)Legile lui Morgan

$$\overline{A \vee B} = \overline{A} \wedge \overline{B}$$
$$\overline{A \wedge B} = \overline{A} \vee \bar{B}$$

f)Evenimentele 1 şi 0 se caracterizează prin

$$0 \wedge A = 0, \quad 0 \vee A = A$$
$$1 \wedge A = A, \quad 1 \vee A = 1$$

g)Evenimentul non A sau contrarul lui A are proprietăţile:

$$A \wedge \overline{A} = 0 \quad A \vee \overline{A} = 1$$

Remarcăm că aceste relaţii sunt adevărate dacă A, B,.. sunt propoziţii, iar 0 este propoziţia totdeauna falsă şi 1 este propoziţia totdeauna adevărată.

O mulţime de evenimente cu proprietăţile de mai sus se numeşte câmp de evenimente sau algebră de evenimente sau algebră Boole. Relaţiile de mai sus nu sunt independente, deci nu formează un set minimal de axiome pentru algebrele Boole. Pe de altă parte ele implică alte relaţii importante, ca de exemplu $\bar{\bar{A}} = A$. Următoarea teoremă a lui Stone descrie asemenea câmpuri de evenimente prin submulţimi.

Teorema 1.5 *(Stone)Fie M o multime de evenimente cu proprietăţile de mai sus. Atunci există o mulţime X şi o submulţime $\Omega \subset P(X)$ cu proprietăţile :*
1)$\emptyset \in \Omega, X \in \Omega$
2)A,B$\in \Omega \Rightarrow A \cup B \in \Omega$
3)A,B$\in \Omega \Rightarrow A - B \in \Omega$
şi o bijecţie $\sigma : M \to \Omega$ *astfel ca:*
a)$\sigma(A \vee B) = \sigma(A) \cup \sigma(B)$
b)$\sigma(A \wedge B) = \sigma(A) \cap \sigma(B)$
c)$\sigma(\overline{A}) = \overline{\sigma(A)}$
$\sigma(0) = \emptyset \quad \sigma(1) = X$

Teorema arată că în esenţă orice câmp de evenimente poate fi reprezentat prin submulţimi ale aceleiaşi mulţimi X. Operaţiei 'şi' între evenimente îi corespunde intersecţia submulţimilor, operaţiei 'sau' îi corespunde reuniunea, negaţiei îi corespunde complementara, evenimentul sigur corespunde mulţimii totale X, iar evenimentul imposibil corespunde cu mulţimea vidă.

Definiţia 1.6 *O mulţime $\Omega \subset P(X)$ cu proprietăţile 1,2,3 din teorema lui Stone se numeste clan, algebră de evenimente sau cîmp de evenimente. Uneori o vom numi algebră de mulţimi.*

Propoziţia 1.7 *Fie $\Omega \subset P(X)$ un câmp de evenimente.*
a) Dacă $A_1, A_2, ...A_n \in \Omega$ atunci $\bigcup\limits_{i=1.n} A_i \in \Omega$ şi $\bigcap\limits_{i=1..n} A_i \in \Omega$
b) Dacă $A \in \Omega$ atunci $\overline{A} \in \Omega$.

Demonstraţie. b) $\overline{A} = X - A$ şi din proprietaţile 1 şi 3 ale algebrei de mulţimi rezultă $\overline{A} \in \Omega$. a)Dacă A şi B sunt în Ω atunci $A \cap B = \overline{\overline{A} \cup \overline{B}} \in \Omega$ conform cu proprietatea 2 a algebrei de mulţimi şi punctul b). Prin inducţie rezultă acum uşor şi a).

Vedem că în general un număr finit de operaţii de intesecţie, reuniune,diferenţă sau complementară cu mulţimi din Ω are ca rezultat tot o mulţime din Ω. Vom defini acum probabilitatea.

Definiţia 1.8 *Fie $\Omega \subset P(X)$ un câmp de evenimente. Se numeşte probabilitate pe Ω, o funcţie $p{:}\Omega \to R$ cu proprietăţile:*
1) $p(A) \in [0,1]$ pentru oricare $A \in \Omega$
2) $p(A \cup B) = p(A) + p(B)$ dacă $A \cap B = \emptyset$
3) $p(\emptyset) = 0 \quad p(X) = 1$

Tripletul (X, Ω, p) îl vom numi câmp de probabilitate.

Teorema 1.9 *Fie (X, Ω, p) un cîmp de probabilitate. Atunci:*
1) $p(\overline{A}) = 1 - p(A)$
2) Dacă $A_1, A_2, ...A_n$ sunt în Ω şi $A_i \cap A_j = \emptyset$ pentru orice i şi j atunci

$$p(\bigcup_{i=1..n} A_i) = \sum_1^n p(A_i) \tag{1.1}$$

3) $p(A \cup B) = p(A) + p(B) - p(A \cap B)$ A şi B nefiind neapărat disjuncte.
4) $A \subset B$ implică $p(A) \leq p(B)$.

Demonstraţie. 1) $X = \bar{A} \cup A$, $A \cap \bar{A} = \emptyset$, deci $1 = p(X) = p(A) + p(\bar{A})$

2) Pentru n=2 afirmaţia rezultă din definiţia probabilităţii, pe urmă se procedează prin inducţie.

3) Să privim evenimentele ca mulţimi. Atunci $A \cup B = (A \cap \bar{B}) \cup (B \cap \bar{A}) \cup (A \cap B)$, reuniune disjunctă, deci $p(A \cup B) = p(A \cap \bar{B}) + p(B \cap \bar{A}) + p(A \cap B)$. De asemenea $A = (A \cap \bar{B}) \cup (A \cap B)$ deci $p(A) = p(A \cap \bar{B}) + p(A \cap B)$ şi prin urmare $p(A \cap \bar{B}) = p(A) - p(A \cap B)$. Analog $p(B \cap \bar{A}) = p(B) - p(A \cap B)$. Punând aceste valori în expresia precedentă pentru p($A \cup B$) se obţine punctul 3).

4) $A \subset B \Rightarrow B = (B \cap \bar{A}) \cup A$ (disjuncte) $\Rightarrow p(B) = p(B \cap \bar{A}) + p(A) \geq p(A)$.

Observaţia 1.10 *Datorită formulei (1.1) probabilitatea p se mai numeşte finit aditivă.*

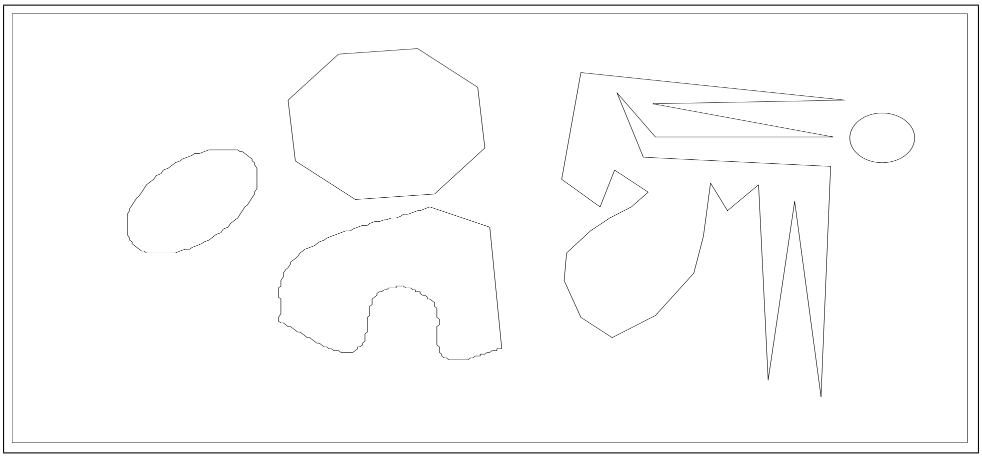
Exemplul 1.11 *Se aruncă două zaruri. Pot apare toate perechile de feţe (i,j) cu* $1 \leq i \leq 6$ *şi* $1 \leq j \leq 6$*, în număr de 36. Fie X mulţimea acestor perechi. Fie* $\Omega = P(X)$ *şi* $p : \Omega \to R$ *definită prin* $p(A) = \frac{numar\ de\ elemente\ din\ A}{numar\ de\ elemente\ din\ X}$*. Funcţia p definită în acest fel este o probabilitate în sensul definiţiei anterioare. Numărul de elemente din X reprezintă numărul cazurilor posibile, iar numărul de elemente din A reprezintă numărul cazurilor favorabile. In acest fel definiţia clasică a probabilităţilor este cuprinsă în definiţia axiomatică de mai sus.*

Acum putem să dăm o soluţie pentru problema pusă în exemplul 4.

Exemplul 1.12 *Fie X=[0,1]×[0,1]. Fie* $\Omega \subset P(X)$ *formată din submulţimile lui X cu frontiera formată dintr-un număr finit de curbe de clasă* C^1 *pe porţiuni (adică fiecare curbă poate avea cel mult un număr finit de puncte unde să nu fie cu derivata continuă). Se verifică uşor că* Ω *este o algebră de mulţimi. Definim acum probabilitatea prin*

$$p(A) = \frac{aria(A)}{aria(X)} \quad pentru\ A \in \Omega$$

Este clar că p astfel definită satisface condiţiile din definiţia axiomatică a probabilităţii. In raport cu această probabilitate evenimentul din exemplul 4 este reprezentat prin A=$\{(x,y) \in X \,|\, y > x\}$ *pentru care* $p(A) = 1/2$*. Asemenea probabilităţi în care X este o submulţime în R,* R^2 *,* R^3*, unde* Ω *este este formată din submulţimi "cu arie" ale lui X, iar* $p(A) = \frac{\int_A f(\xi)d\xi}{\int_X f(\xi)d\xi}$ *se numesc probabilităţi geometrice. Funcţia f în această definiţie se mai numeşte pondere sau densitate şi trebuie să fie positivă.*



Diverse tipuri de figuri pentru care este definită probabilitatea geometrică

1.3 Probabilităţi condiţionate

Fie (X,Ω,p) un câmp de probabilitate, B în Ω, p(B)> 0.

Definiţia 1.13 *Definim $p_B : \Omega \to R$ (sau $p(\cdot\,|B)$) prin*

$$p(A|B) = p_B(A) = \frac{p(A \cap B)}{p(B)} \tag{1.2}$$

şi o numim probabilitatea lui A condiţionată de B.

Teorema 1.14 *În condiţiile definiţiei de mai sus p_B este o probabilitate şi în plus $p(A \cap B) = p(B) \cdot p_B(A)$.*

Demonstraţie. Deoarece $\emptyset \subset A \cap B \subset B$ atunci $0 \leq p_B(A) \leq 1$. De asemenea $p_B(\emptyset) = 0$ şi $p_B(X) = 1$ sunt evidente. Dacă $A_1 \cap A_2 = \emptyset$ atunci şi $(A_1 \cap B) \cap (A_2 \cap B) = \emptyset$, deci

$$\begin{aligned} p_B(A_1 \cup A_2) &= \frac{p((A_1 \cup A_2) \cap B)}{p(B)} = \frac{p((A_1 \cap B) \cup (A_2 \cap B))}{p(B)} \\ &= \frac{p(A_1 \cap B) + p(A_2 \cap B)}{p(B)} = p_B(A_1) + p_B(A_2) \end{aligned}$$

Definiţia 1.15 *Spunem că două evenimente A şi B sunt independente dacă $p(A \cap B) = p(A) \cdot p(B)$, altfel spus dacă $p(A) = p_B(A)$.*

Propoziţia 1.16 *Dacă A şi B sunt independente atunci şi perechile de evenimente ($A,\bar{B}$), ($\bar{A},B$), ($\bar{A},\bar{B}$) sunt independente.*

Demonstraţie. Prin definiţie $p(A \cap B) = p(A)p(B)$. Deoarece $A = (A \cap B) \cup (A \cap \bar{B})$ (reuniune disjunctă), avem

$$p(A) = p(A \cap B) + p(A \cap \bar{B}) = p(A)p(B) + p(A \cap \bar{B})$$

de unde rezultă

$$p(A \cap \bar{B}) = p(A) - p(A)p(B) = p(A)(1 - p(B)) = p(A)p(\bar{B})$$

. Analg se demonstrează independenţa şi în celelalte cazuri.

În mod analog se defineşte independenţa mai multor evenimente:

Definiţia 1.17 *$A_1, A_2, ...A_n$ sunt independente dacă pentru orice $k \leq n$ şi orice evenimente $A_{i_1}, A_{i_2}, ..A_{i_k}$ dintre cele $A_1, ..A_n$ date avem*

$$p(A_{i_1} \cap A_{i_2} \cap ..A_{i_k}) = p(A_{i_1})p(A_{i_2})..p(A_{i_k}). \tag{1.3}$$

Analog cu propoziţia anterioară se poate demonstra şi

Propoziţia 1.18 *Dacă evenimentele $A_1, A_2, ..A_n$ sunt independente atunci şi evenimentele $A_1, \bar{A}_2, ..A_n$ sunt independente (orice combinaţie de A_i sau complementare).*

Demonstraţia este lăsată ca exerciţiu.

Definiţia 1.19 *Spunem că evenimentele $A_1, A_2, ..A_n$ formează o partiţie a lui X dacă*
1) $\bigcup\limits_{i=1..n} A_i = X$
2) $A_i \cap A_j = \emptyset$ pentru orice $i \neq j$.

Avem următoarea

Propoziţia 1.20 *Fie (X, Ω, p) un câmp de probabilitate şi partiţia $A_1, ..A_n$. Atunci pentru orice $B \in \Omega$ avem*

$$p(B) = \sum_{i=1}^{n} p(A_i) \cdot p(B|A_i) \tag{1.4}$$

Formula de mai sus se numeşte formula probabilităţii totale.
Demonstraţie.

$$\begin{gathered} p(B) = p(B \cap X) = p(B \cap (\bigcup_{i=1..n} A_i)) = p(\bigcup_{i=1..n} (B \cap A_i)) = \\ \textstyle\sum_{i=1}^{n} p(B \cap A_i) = \sum_{i=1}^{n} p(A_i)p(B) \end{gathered} \tag{1.5}$$

Q.E.D.
Formula următoare, numită formula lui Bayes, ne dă probabilitatea ca după ce ştim că un eveniment ce poate apare din mai multe cauze s-a realizat, acesta să se fi realizat dintr-o cauză anume. Enunţul precis este:

Propoziţia 1.21 *(Bayes)Fie (X, Ω, p) un câmp de probabilitate şi $A_1, A_2, ..A_n$ o partiţie a lui X. Fie B un eveniment cu probabilitate nenulă. Atunci:*

$$\mathbf{p}(\mathbf{A}_i|B) = \frac{p(A_i)p(B|A_i)}{\sum_{j=1}^{n} p(A_j)p(B|A_j)} \tag{1.6}$$

Demonstraţie.

$$p(A_i|B) = \frac{p(B \cap A_i)}{p(B)} = \frac{p(A_i)p(B|A_i)}{\sum_{j=1}^{n} p(A_j)p(B|A_j)}$$

conform formulei probabilităţii totale.

Exemplul 1.22 *În trei urne cu bile albe şi negre avem compoziţiile: 4+6, 2+8, 4+1. Se pun toate bilele la un loc şi se extrage o bilă la întîmplare. Se constată că este albă. Care este probabilitatea ca ea să fi provenit din urna a treia?*

Soluţie. Fie A_i evenimentul care constă în extragerea unei bile provenind din urna i şi B evenimentul care constă în extragerea unei bile albe. Avem $p(A_1) = \frac{10}{25}$ $p(A_2) = \frac{10}{25}$ $p(A_3) = \frac{5}{25}$ (probabilitatea evenimentului A$_i$ este proporţională cu numărul bilelor provenite din urna i). Mai departe avem $p(B|A_1) = \frac{4}{10}$ $p(B|A_2) = \frac{2}{10}$ $p(B|A_3) = \frac{4}{5}$. Probabilitatea care ni se cere este $p(A_3|B)$ şi conform cu formula lui Bayes

$$p(A_3|B) = \frac{\frac{10}{25}\cdot\frac{4}{5}}{\frac{10}{25}\cdot\frac{4}{10}+\frac{10}{25}\cdot\frac{2}{10}+\frac{5}{25}\cdot\frac{4}{5}} = \frac{2}{5}$$

Observaţia 1.23 *Dacă alegerea unei bile s-ar fi făcut după regulile: i)se alege la întîmplare o urnă; 2)din urna aleasă se extrage la întîmplare o bilă; atunci am avea* $p(A_1) = p(A_2) = p(A_3) = \frac{1}{3}$ *şi tot formula lui Bayes conduce la* $p(A_3|B) = \frac{4}{7}$.

1.4 Rezumat

Probabilitatea are o componentă psihologică în sensul de grad de încredere şi una practică în sensul de frecvenţă de apariţie a unui fenomen într-un număr mare de experienţe independente, dar pentru a putea introduce numere şi a face calcule s-a dovedit utilă o definiţie axiomatică, asemănător cum în geometrie dreapta se introduce axiomatic, independent de mulţimea exemplelor practice. Punctele importante în acest demers sunt:

i) Reprezentarea evenimentelor prin algebre de mulţimi(Stone).

ii) Definiţia probabilităţii ca funcţie pe algebre de evenimente, cu proprietăti asemănătoare cu aria figurilor plane.

In dezvoltarea elementară a calculului cu probabilităţi următoarele trei puncte sunt esenţiale:

iii) Noţiunea de probabilitate condiţionată şi probabilitatea de apariţie simultană a evenimentelor independente.

iv) Formula probabilităţii totale (1.4) .

v) Formula pentru probabilitatea cauzelor (1.6).

FORMULE UTILIZATE FRECVENT :

a) p=$\frac{nr.\,cazuri\,favorabile}{nr.\,cazuri\,posibile}$**.**

b) Definiţia probabilităţii condiţionate $p(A|B) = p_B(A) = \frac{p(A\cap B)}{p(B)}$**(1.2);**

probabilitatea evenimentelor independente $p(A\cap B) = p(A)\cdot p(B)$**(1.3)**

c) Formula probabilităţii totale $p(B) = \sum_{i=1}^{n} p(A_i)\cdot p(B|A_i)$**(1.4) .**

d) Formula lui Bayes p(A$_i|B) = \frac{p(A_i)p(B|A_i)}{\sum_{j=1}^{n} p(A_j)p(B|A_j)}$**(1.6).**

1.5 Exerciţii

1. O urnă conţine jetoane numerotate de la 1 la 8. Se extrag la întîmplare 3 jetoane. Care este probabilitatea ca suma numerelor extrase să fie superioară sau egală cu suma numerelor rămase?

Indicaţie. Probabilitata este raportul dintre numărul cazurilor favorabile şi numărul cazurilor posibile. Sunt posibile $C_8^3 = 56$ cazuri. Suma tuturor numerelor este 36, deci suma celor extrase trebuie să depăşească 18 pentru a fi caz favorabil. Acum se enumeră pur şi simplu toate cazurile favorabile: (8,7,6), (8,7,5), etc.

2. Se aleg la întîmplare a,b,c în intervalul (0,1). Care este probabilitatea ca ecuaţia de gradul doi $ax^2+2bx+c=0$ să aibă rădăcini reale? (se consideră (a,b,c)$\in (0,1)\times(0,1)\times(0,1)$ iar probabilitatea este proporţională cu volumul).

Indicaţie. (a,b,c)$\in (0,1)^3$. Condiţia de rădăcini reale este $4b^2-4ac \geq 0$ sau $1 \geq b \geq \sqrt{ac}$. Fie V=volumul lui $(0,1)^3$ şi V_f =volumul cazurilor favorabile. Avem $V_f = \int\int\int_{V_f} da\cdot dc\cdot db = \int_0^1 da \int_0^1 dc \int_{\sqrt{ac}}^1 db$. Tinând seama că V=1, se obţine imediat probabilitatea $p = V_f/V$.

3. La un teatru sunt 2n persoane la coadă la bilete. n persoane au bancnote de 500 lei iar celelalte n persoane au bancnote de 1000 lei. Un bilet costa 500 lei. Care este probabilitatea ca aşezâindu-se întîmplător la coadă să poată toate persoanele să-şi cumpere bilet, ştiind că initial în casă nu este nici un leu si fiecare persoană primeşte restul de la casă?

Soluţie Ca să numărăm aşezările în care toţi î-şi pot cumpăra bilete, reprezentăm într-un sistem xOy pe cele 2n persoane prin puncte în 1,2,3..2n pe Ox. Fiecărei aşezări îi asociem o funcţie definită pe $\overline{0,2n}$ prin f(0)=0 şi f(i)=f(i-1)+1 dacă persoana i de la rînd are 1000 lei şi f(i)=f(i-1)-1 dacă persoana i are 500 lei. Fiecare traiectorie începe în (0,0) şi se termină în (2n,0), deoarece de cîte ori se urcă se şi coboară. Numărul aranjărilor la coadă este n!n!C_{2n}^n unde C_{2n}^n reprezintă numărul de alegeri ale locurilor de către cei cu 500 lei (sau numărul traiectoriilor) iar n! reprezintă numărul de permutări celor cu 500 lei sau cu 1000 lei intre ei. O traiectorie reprezintă o aşezare nefavorabilă dacă pentru un i avem f(i)=1. În acest caz primul i cu proprietatea de mai sus reprezinta cel cu 1000 lei care nu mai poate primi rest de la casă. Considerînd simetrica acestei traiectorii de la acest i faţă de y=1 obţinem o traiectorie ce se termină în (2n,2) şi are n+1 urcuşuri şi n-1 coborâşuri. Numărul traiectoriilor de acest fel este C_{2n}^{n+1} , deci numărul cazurilor nefavorabile este n!n!C_{2n}^{n+1}. Probabilitatea este deci

$$p = \frac{n!n!C_{2n}^n - n!n!C_{2n}^{n+1}}{n!n!C_{2n}^n} = \frac{1}{n+1}$$

4. Fie (X,Ω,p) un câmp de probabilitate şi $A_1, A_2, ..A_n$ evenimente din Ω. Să se arate *că*

$$p(\underset{i=1..n}{\cup} A_i) = \sum_{i=1}^{n} p(A_i) - \sum_{i\neq j} p(A_i \cap A_j) + \sum_{i\neq j\neq k} p(A_i \cap A_j \cap A_k) + +... + (-1)^{n-1} p(A_1 \cap A_2 \cap .. \cap A_n) \tag{1.7}$$

(formula lui Poincaré).

Indicaţie. Formula a fost demonstrată pentru n=2 în cadrul lecţiei. Presupunem prin inducţie că este adevărată pentru n. Fie $B = \underset{i=1..n}{\cup} A_i$ şi $B_i = A_{n+1} \cap A_i$. Avem

$$\begin{aligned} p(\underset{i=1..n+1}{\cup} A_i) &= p(B \cup A_{n+1}) \\ &= p(B) + p(A_{n+1}) - p(B \cap A_{n+1}) \\ &= (formula \quad 1.7\,) + p(A_{n+1}) - p(\underset{i=1..n}{\cup} B_i) \\ &= (formula \quad 1.7\,) + p(A_{n+1}) - (formula\, 1.7 \;\; pt.\; B_i) \\ &= (formula\, 1.7 \;\; pt.\; n+1) \end{aligned}$$

5. O persoană scrie n scrisori către n persoane distincte, le pune în n plicuri şi apoi scrie adresele la întîmplare. Care este probabilitatea ca cel puţin o scisoare să ajungă la destinatarul potrivit? (problema concordanţelor).

Indicaţie. Fie A_k evenimentul ce constă în faptul că persoana k primeşte scrisoarea potrivită. $p(A_k) = \frac{(n-1)!}{n!} = 1/n$. Există C_n^2 evenimente de tipul $A_i \cap A_j$ şi fiecare are probabilitatea $\frac{(n-2)!}{n!}$ (deoarece două persoane primesc scrisorile potrivite iar celelalta n-2 persoane primesc scrisorile întâmplător în $(n-2)!$ feluri. Analog, există C_n^3 evenimente de tipul $A_i \cap A_j \cap A_k$ şi fiecare are probabilitatea $\frac{(n-3)!}{n!}$, etc. Aplicând formula de mai sus rezultă:

$$\begin{aligned} p &= n \cdot \frac{1}{n} - C_n^2 \cdot \frac{(n-2)!}{n!} + C_n^3 \cdot \frac{(n-3)!}{n!} + ... \\ &= \frac{1}{1!} - \frac{1}{2!} + \frac{1}{3!} + ... \end{aligned}$$

6. Două persoane A şi B s-au înţeles să se întâlnească într-un anumit loc între orele 12 şi 13. Persoana care vine prima aşteaptă 20 de minute şi apoi pleacă. Dacă fiecare persoană vine la întîmplare în acel loc, care este probabilitatea ca persoanele să se întîlnească?

Indicaţie. Prima problemă este cum reprezentăm toate posibilităţile de întâlnire. Fie pe axa x ora de sosire a primei persoane iar pe axa y ora de sosire a celei de a doua persoane.

Toate variantele de sosire se reprezintă deci prin produsul $[12,13]\times[12,13]$. Condiţia de întâlnire este $|x-y|\leq\frac{1}{3}$ ore. Admiţând că probabilitatea este proporţională cu aria, găsim că aria zonei de întîlnire este 5/9 iar aria zonei de sosire este 1. Deci probabilitatea este 5/9.

7. Un magazin se aprovizionează de la trei fabrici A,B.C, cu un anumit produs, în proporţie de 30%,60%,10%. Proporţia de articole defecte dintre cele achiziţionate este de 2%,1%,5%, pentru A, respectiv B şi C. Dacă un cumpărător găseşte că produsul achiziţionat la magazin este cu defecţiuni, care este probabilitatea ca el să fi provenit de la fabrica B?

Indicaţie. Se aplică formula lui Bayes.

8. Intr-un circuit sunt legate în serie rezistenţele R_1, R_2 şi grupul de rezistenţe în paralel: R_3, R_4, R_5. Probabilităţile de defecţiune ale celor cinci elemente independente sunt: 0,1; 0,01; 0,05; 0,04; 0,1*. Care este probabilitatea de întrerupere a circuitului?*

Indicaţie. Fie A_k evenimentul care constă în întreruperea rezistenţei R_k. Avem $p(A_1)=0,1 \quad p(A_2)=0,01 \quad etc.$ Toate aceste evenimente sunt independente şi întreruperea circuitului este: $A_1\cup A_2\cup(A_3\cap A_4\cap A_5)$. Se aplică formula lui Poincaré şi se ţine seama că evenimentele A_k sunt independente.

9. Trei ţintaşi nimeresc o ţintă cu probabilităţile 0,7;0,8;0,9. Fiecare trage câte o lovitură. Care este probabilitatea ca:
a) Toţi trei să nimerească ţinta?
b) Cel puţin unul să nimerească ţinta?
c) Unul singur să nimerească ţinta?

Indicaţie. Fie A_k evenimentul care constă în faptul că trăgătorul k nimereşte. Avem $p(A_1)=0,7; p(A_2)=0,8; p(A_3)=0,9$ iar evenimentele A_1, A_2, A_3 sunt independente. La punctul a) se cere probabilitatea lui $A_1\cap A_2\cap A_3$ care este produsul probabilităţilor $p(A_1)\cdot p(A_2)\cdot p(A_3)$. Analog se consideră şi celelalte cazuri.

10. Fie X={1,2,3,4} şi $\Omega=P(X)$. Se defineşte o probabilitate pe X astfel ca $p(\{1,2,3\})=7/8$, $p(\{2,3,4\})=1/2$, şi $p(\{1,4\})=5/8$. Să se determine $p(\{1\})$ şi $p(\{1,3,4\})$.

11. Intr-o urnă se află bile numerotate 0, 1, 2, ..9. Se extrag 3 bile la întâmplare, fără a pune bila înapoi. Se scrie numărul format din cele trei cifre în ordinea apariţiei. Care este probabilitatea ca numărul să se dividă la 12? Dar dacă extragerea se face cu punerea bilei înapoi?

12. Un grup format din 2n băieţi şi 2n fete este despărţit în două subgrupuri de acelaşi efectiv. Care este probabilitatea ca în fiecare subgrup numărul de băieţi să fie egal cu numărul de fete.

13. Fie $A_1, A_2, ...A_n$ evenimente independente. Care este numărul maxim de evenimente *distincte care se poate obţine din* $A_1, ...A_n$ *prin aplicarea repetată a operaţiilor şi, sau, non?*

14. Pe un segment AB se iau la întâmplare două puncte C şi D, astfel ca $|AC| < |AD|$. *Care este probabilitatea ca să se poată forma un triunghi cu segmentele AC, CD, DB?*

15. O urnă conţine 5 bile albe, 3 bile negre şi 2 bile roşii iar altă urnă conţine 3 bile negre , 2 bile albe şi 5 bile roşii. Se extrage câte o bilă din fiecare urnă. Care este probabilitatea ca să se extragă bile de aceeaşi culoare?

16. Fie evenimentele $A_1, A_2, ..A_n$. *Să se demonstreze că*

$$p(A_1 \cap A_2 \cap ... \cap A_n) \geq p(A_1) + p(A_2) + ... + p(A_n) - (n-1)$$

(inegalitatea lui Boole).

17. Un submarin lansează n torpile asupra unui vas. Dacă fiecare torpilă are probabilitatea $\frac{1}{2}$ *de a lovi vasul, independent de celelalte torpile, care este numărul minim de torpile ce trebuie lansate ca probabilitatea de a fi lovit vasul de cel puţin una să depăşească 0,9?*

Lecţia 2

Variabile aleatoare simple

2.1 Definiţie şi proprietăţi

De fiecare dată când aplicăm teoria probabilităţilor subînţelegem că există o algebră de evenimente realizată ca o mulţime de submulţimi ale unei mulţimi X, şi, o probabilitate definită pe acele evenimente, chiar dacă X şi algebra de evenimente nu sunt exprimate explicit. În general informaţia din X se extrage prin funcţii. În cele ce urmează, X este o mulţime pe care avem definită o algebră de evenimente Ω şi o probabilitate p.

Definiţia 2.1 *O funcţie $f : X \to R$ se numeste variabilă aleatoare (pe scurt v.a.) dacă pentru orice $(a, b) \subset R$ avem $f^{-1}(a, b) \in \Omega$. O funcţie $f : X \to C$ se numeşte variabilă aleatoare dacă $Re(f)$ şi $Im(f)$ sunt variabile aleatoare reale.*

În cele ce urmează variabilele aleatoare considerate iau un număr finit de valori. Asemenea variabile aleatoare le vom numi simple. Fie $v_1, v_2, ..v_n$ valorile distincte ale lui f şi fie $A_i = f^{-1}(v_i)$. În aceste condiţii variabila aleatoare se scrie

$$f(x) = \begin{cases} v_1 \text{ dacă } x \in A_1 \\ v_2 \text{ dacă } x \in A_2 \\ \quad\quad ..... \\ v_n \text{ dacă } x \in A_n \end{cases} \tag{2.1}$$

Este clar că definiţia de mai sus este echivalentă cu a spune că pentru orice valoare v rezultă că $f^{-1}(v) \in \Omega$. Vom mai scrie $A_i = \{f = v_i\}, \quad p_i = p(A_i) = p(f^{-1}(v_i)) = p(f = v_i)$. Este mai puţin important din punctul de vedere al calculului cu probabilităţi, care este mulţimea A_i, cât care sunt probabilităţile $p(A_i)$, $\quad p(A_i \cap A_j)$ etc. Fiecărei variabile aleatoare f îi vom asocia o diagramă (notată tot f):

$$f = \begin{pmatrix} v_1 & v_2 & v_3 & \cdots & v_n \\ p_1 & p_2 & p_3 & \cdots & p_n \end{pmatrix}$$

Valorile $v_1, v_2, ..v_n$ sunt în general distincte iar evenimentele $A_i, i = 1, 2, ..n$ formează o partiţie a lui X , adică A_i şi A_j sunt disjuncte dacă $i \neq j$ şi $\bigcup\limits_{i=1..n} A_i = X$. Este clar că $p_1 + p_2 + .. + p_n = 1$. Putem considera şi diagrame în care unele valori v coincid între ele ceea ce ar corespunde definirii lui f prin (2.1) şi unde am avea de exemplu $v_1 = v_2$ dar neapărat $A_1 \cap A_2 = \emptyset$. În acest caz A_i nu mai este neapărat $f^{-1}(v_i)$ dar $A_1, ..A_n$ formează încă o partiţie, $p_i = p(A_i)$ şi $p_1 + .. + p_n = 1$. De exemplu funcţia f^k ia valorile $(v_i)^k$ pe mulţimile A_i, deci diagrama asociată este:

$$f^k = \begin{pmatrix} v_1^k & v_2^k & \ldots & v_n^k \\ p_1 & p_2 & \ldots & p_n \end{pmatrix}$$

iar dacă $v_1 = -v_2$, k=2 atunci $v_1^2 = v_2^2$ şi avem o diagramă unde unele valori din rândul de sus coincid. Diagramele în care $v_i \neq v_j$ pentru $i \neq j$ le vom numi standard. Dată o variabilă aleatoare simplă, ca mai sus, definim:

Definiţia 2.2 *1)media lui f:* $M(f) = \sum_{i=1}^n v_i p_i = \sum_{i=1}^n v_i \cdot p(f = v_i)$
2)momentul de ordin k: $M_k(f)$ *(sau* m_k*)* $= \sum_{i=1}^n v_i^k p_i = M(f^k)$
3)momentul centrat de ordin k: $\mu_k(f) = \sum_{i=1}^n (v_i - M(f))^k p_k = M\left((f - M(f))^k\right)$
4)dispersia: $D(f) = \sigma^2(f) = \sum_{i=1}^n (v_i - M(f))^2 p_i = M((f - M(f))^2) = \mu_2(f)$
5)funcţia caracteristică $f_c : R \to C$ *sau* $\varphi_f : R \to C$, $f_c(t) = \varphi_f(t) = \sum_{i=1}^n e^{\sqrt{-1} v_i t} p_i$
6)functia de repartiţie $F : R \to R$, $F(t) = p(\{x \in X | f(x) < t\})$.

Observaţia 2.3 *Dacă pentru un i avem* $p_i = 0$ *atunci valoarea corespunzătoare* v_i *o putem exclude din diagramă deoarece nu are nici o contribuţie la caracteristicile numerice ale lui f. Despre două v.a. f şi g, pentru care mulţimea pentru care* $f \neq g$ *are probabilitatea 0, spunem că ele coincid aproape peste tot (pe scurt a.p.t.). Ele au aceleaşi diagrame standard şi aceleaşi caracteristici numerice: medie, dispersie,.... Dacă unei v.a. f i se asociază două diagrame distincte, atunci definiţiile de mai sus pentru medie, dispersie, momente, funcţie caracteristică, dau acelaşi rezultat, indiferent de diagrama folosită. Se poate demonstra acest lucru uşor comparând valorile obţinute dintr-o diagramă cu cele date de diagrama standard (exerciţiu).*

Vom nota în general cu $\{a < f < b\}$ evenimentul $f^{-1}(a, b) = \{x \in X | a < f(x) < b\} \in \Omega$ iar probabilitatea lui cu $p(a < f < b)$. Analog vom nota evenimentele ce se obţin folosind inegalităţi de tipul $\leq, \geq$,etc.

Exemplul 2.4 *Într-o clasă s-au obţinut următoarele note: 5 de către 8 elevi, 7 de către 3 elevi, 8 de către 5 elevi, 10 de către 4 elevi. Funcţia nota ia valorile 5,7,8,10, iar probabilităţile corespunzătoare sunt:* $\frac{8}{20}$ $\frac{3}{20}$ $\frac{5}{20}$ $\frac{4}{20}$. *Diagrama asociată funcţiei este:*

$$nota \begin{pmatrix} 5 & 7 & 8 & 10 \\ \frac{8}{20} & \frac{3}{20} & \frac{5}{20} & \frac{4}{20} \end{pmatrix}$$

media notelor este $M_1 = 5 \cdot \frac{8}{20} + 7 \cdot \frac{3}{20} + 8 \cdot \frac{5}{20} + 10 \cdot \frac{4}{20} = 7,05$
momentul de ordinul doi este $M_2 = 5^2 \cdot \frac{8}{20} + 7^2 \cdot \frac{3}{20} + 8^2 \cdot \frac{5}{20} + 10^2 \cdot \frac{4}{20} = 53,35$
dispersia este $D = (5-7,05)^2 \cdot \frac{8}{20} + (7-7,05)^2 \cdot \frac{3}{20} + (8-7,05)^2 \cdot \frac{5}{20} + (10-7,05)^2 \cdot \frac{4}{20} = 3,6475$
funcţia funcţia caracteristică este $f_c(t) = e^{i5t} \cdot \frac{8}{20} + e^{i7t} \cdot \frac{3}{20} + e^{i8t} \cdot \frac{5}{20} + e^{i10t} \cdot \frac{4}{20}$ *unde* $i = \sqrt{-1}$

Observaţia 2.5 *Media se poate numi centrul valorilor luate de o v.a. iar dispersia este o măsură a împrăştierii valorilor acelei v.a. în jurul mediei. In afară de medie, celelalte caracteristici se utilizează doar pentru v.a. reale.*

Definiţia 2.6 *Două variabile aleatoare reale, f şi g definite pe acelaşi câmp de probabilitate X se numesc independente dacă oricare ar fi intervalele (a,b) şi (c,d) avem*

$$p(f^{-1}(a,b) \cap g^{-1}(c,d)) = p(f^{-1}(a,b)) \cdot p(g^{-1}(c,d))$$

Definiţia independenţei a două v.a. simple se poate da în mai multe feluri aşa cum rezultă din propoziţia următoare:

Propoziţia 2.7 *Fie* (X, Ω, p) *un câmp finit de probabilitate şi f,g două variabile aleatoare simple cu valori reale. Următoarele afirmaţii sunt echivalente:*
1) f şi g sunt independente
2) evenimentele $\{a < f < b\}$ *şi* $\{c < g < d\}$ *sunt independente pentru* $a, b, c, d \in R$
3) pentru orice v,u$\in R$ *evenimentele* $\{f = v\}$ *şi* $\{g = u\}$ *sunt independente*
4) pentru orice submulţimi A,B din R, evenimentele $\{f \in A\}$ *şi* $\{f \in B\}$ *sunt independente*

Demonstraţia este lăsată ca exerciţiu.

Observaţia 2.8 *Sub forma (3) cu* $u, v \in C$ *sau sub forma (4) cu* $A, B \in C$*, din propoziţia anterioară, definiţia independenţei a două v.a. simple se poate enunţa şi pentru v.a. cu valori complexe.*

Propoziţia 2.9 *Dacă* $f, g : X \to C$ *sunt v.a. simple independente iar* $F, G : C \to C$ *atunci* $F \circ f$ *şi* $G \circ g$ *sunt independente.*

Demonstraţie. Avem:

$$\begin{aligned} & p\left((F\circ f=v)\cap(G\circ g=u)\right) \\ = \; & p\left(\left(f\in F^{-1}(v)\right)\cap\left(g\in G^{-1}(u)\right)\right) \\ = \; & p(f\in F^{-1}(v))\cdot p(g\in G^{-1}(u)) \\ = \; & p(F\circ f=v)p(G\circ g=u) \end{aligned}$$

Q.E.D.

De exemplu $(f-a)^2$ şi $(g-b)^2$ sunt independente pentru orice a şi b.

Teorema 2.10 *(proprietăţi ale mediei) Fie (X,Ω,p) un câmp de probabilitate finit şi $f, f_1, f_2, ...$ variabile aleatoare pe X. Atunci:*
1)$f\equiv C$ (constant) $\Rightarrow M(f)=C$
2)$\alpha=const\Rightarrow M(\alpha f)=\alpha M(f)$
3)$M(f_1+f_2)=M(f_1)+M(f_2)$ şi de aici $M(f_1+f_2+..+f_n)=M(f_1)+M(f_2)+..+M(f_n)$, $M(af+bg)=aM(f)+bM(g)$, şi $M(f-M(f))=0$, (a si b fiind constante)
4) Dacă f_1, f_2 sunt independente atunci $M(f_1f_2)=M(f_1)\cdot M(f_2)$.
5) $f\geq 0\Rightarrow M(f)\geq 0$.
6) $f\geq g\Rightarrow M(f)\geq M(g)$.
7) $|M(f)|\leq\max|f|$.

Demonstraţie. 1) şi 2) sunt evidente.

3)Fie $v_1, v_2, ..v_n$ valorile lui f_1 şi $A_i=f_1^{-1}(v_i)$ iar $u_1, ..u_m$ valorile lui f_2 şi $B_j=f_2^{-1}(u_j)$. Variabila aleatoare f_1+f_2 ia valorile v_i+u_j pe evenimentele $\mathrm{A}_i\cap B_j$ cu probabilităţile $p_{ij}=p(A_i\cap B_j)=p(f_1=v_i\cap f_2=u_j)$. Avem acum

$$\begin{gathered} M(f_1+f_2)=\sum_{i=1..n,j=1..m}(v_i+u_j)p_{ij}=\sum_{i=1}^n v_i\sum_{j=1}^m p_{ij}+\sum_{j=1}^m u_j\sum_{i=1}^n p_{ij}= \\ =\sum v_ip(f_1=v_i)+\sum u_jp(f_2=u_j)=M(f_1)+M(f_2) \end{gathered}$$

Am folosit

$$\begin{gathered} \sum_{j=1}^m p_{ij}=\sum_{j=1}^m p(A_i\cap B_j)=\text{(fiind disjuncte)} \\ =p(\bigcup_{j=1..m}(A_i\cap B_j))=p(A_i\cap(\bigcup_{j=1..m}B_j))= \\ p(A_i\cap X)=p(A_i)=p(f=v_i) \end{gathered}$$

şi în mod analog cealaltă sumă este $\sum_{i=1}^n p_{ij}=p(f_2=u_j)$.

4) $M(f_1f_2)=\sum_{i,j}v_iu_jp(f_1=v_i\cap f_2=u_j)=$(din independenţa variabilelor aleatoare) $=\sum_{i,j}v_iu_jp(f_1=v_i)p(f_2=u_j)=\sum_i v_ip(f_1=v_i)\sum_j u_jp(f_2=u_j)=M(f_1)M(f_2)$

5), 6), 7) sunt evidente.

Q.E.D.

Teorema 2.11 *(proprietăţile dispersiei). 1) $f\equiv C$ a.p.t.(constant) $\Leftrightarrow D(f)=0$*
2) $D(af)=a^2D(f)$ a fiind o constantă.
3) Dacă f_1 şi f_2 sunt independente atunci $D(f_1+f_2)=D(f_1)+D(f_2)$

Demonstraţie. 1) $D(f) = \sum (v_i - M(f))^2 p(f = v_i)$ şi suma este zero numai dacă $v_i = M(f)$ pentru orice i cu $p_i \neq 0$, adică $f \equiv M(f)$ a.p.t.

2) Avem

$$D(af) = M((af - M(af))^2) = M(a^2(f - M(f))^2) = a^2 M((f - M(f))^2) = a^2 D(f)$$

3) În formulele de mai jos utilizăm faptul că f_1 şi f_2 independente implică $(f_1 - a)^2$ şi $(f_2 - b)^2$ sunt independente, precum şi proprietăţile mediei:

$$\begin{gathered} D(f_1 + f_2) = M(((f_1 + f_2) - M(f_1 + f_2))^2) = \\ M(((f_1 - M(f_1)) + (f_2 - M(f_2))^2) = \\ M((f_1 - M(f_1))^2 + (f_2 - M(f_2))^2 + 2M((f_1 - M(f_1))(f_2 - M(f_2))) = \\ M((f_1 - M(f_1))^2) + M((f_2 - M(f_2))^2) + 2\underbrace{M(f_1 - M(f_1))}_{=0} \cdot \underbrace{M(f_2 - M(f_2))}_{=0} = \\ D(f_1) + D(f_2) \end{gathered}$$

Q.E.D.

O generalizare a punctului 3) este:

Propoziţia 2.12 *Fie $f_1, f_2, ..f_n$ variabile aleatoare independente două cîte două. Atunci*

$$D(f_1 + f_2 + ... + f_n) = D(f_1) + D(f_2) + .. + D(f_n)$$

Demonstraţia este analoagă celei pentru două funcţii şi e lăsată ca exerciţiu.

Definiţia 2.13 *Dacă f este o variabilă aleatoare, expresia $\frac{f - M(f)}{\sqrt{D(f)}}$ se numeşte deviaţia standard a lui f.*

Se vede imediat că $M(\frac{f-M(f)}{\sqrt{D(f)}}) = 0$ şi $D(\frac{f-M(f)}{\sqrt{D(f)}}) = 1$.

Teorema 2.14 *(proprietăţi ale funcţiei caracteristice) Fie f_c funcţia caracteristică a variabilei aleatoare f. Atunci:*
1) f_c este continuă (chiar uniform continuă) pe R)
2) $f_c(0) = 1, \quad |f_c(t)| \leq 1$ pentru $-\infty < t < \infty$
3) Dacă $g = a \cdot f + b$, cu a şi b constante, atunci $g_c(t) = e^{\sqrt{-1}bt} \cdot f_c(at)$
4) Dacă f_1 şi f_2 sunt independente, atunci $(f_1 + f_2)_c(t) = f_{1c}(t) \cdot f_{2c}(t)$
5) $M_k(f) = \frac{1}{(\sqrt{-1})^k} f_c^{(k)}(0)$
6) Dacă $f_{1c} \equiv f_{2c}$ atunci f_1 şi f_2 au aceleaşi diagrame standard, adică $f_1 = f_2$ a.p.t.

Demonstraţie. 1)f_c este o sumă finită de funcţii continue $f_c(t) = \sum p_k e^{\sqrt{-1}v_k t}$ deci este o funcţie continuă. Având derivata mărginită rezultă că este uniform continuă.

2) $g_c(t) = \sum p_k e^{\sqrt{-1}t(\alpha v_k + \beta)} = e^{\sqrt{-1}t\beta} \sum p_k e^{\sqrt{-1}t\alpha v_k} = e^{\sqrt{-1}\beta t} f_c(\alpha t)$.

3) $f_c(0) = \sum p_k e^{\sqrt{-1}0v_k} = \sum p_k = 1$ şi $|f_c(t)| \leq \sum p_k \left| e^{\sqrt{-1}v_k t} \right| = \sum p_k = 1$

4) Dacă f_1 şi f_2 sunt independente atunci e^{if_1} şi e^{if_2} sunt independente.

Acum observăm că

$$\begin{aligned}(f_1 + f_2)_c(t) &= M(e^{it(f_1+f_2)}) = M(e^{itf_1} \cdot e^{itf_2}) = (\text{fiind indep.}) \\ &= M(e^{itf_1})M(e^{itf_2}) = f_{1c}(t) \cdot f_{2c}(t)\end{aligned}$$

5) $f_c'(t) = i \sum_k p_k v_k e^{itv_k}$ şi deci $f_c'(0) = i \sum_k p_k v_k = iM_1(f)$. Analog prin derivare succesivă se obţine $f^{(n)}(0) = i^n M_n(f)$.

6) Avem, folosind diagrame reduse, $\sum p_k e^{iv_k t} \equiv \sum q_l e^{iu_l t}$, unde v_k sunt diferite între ele şi u_l sunt diferite între ele. Dacă nu se reduc toţi termenii, adică pentru fiecare k să existe l astfel ca $p_k = q_l, v_k = u_l$ atunci după toate reducerile posibile egalitatea devine: $\sum r_s e^{iw_s t} \equiv 0$, $(1 \leq s \leq m)$ unde r_s sunt nenule , iar w_s sunt diferite între ele. Derivând de mai multe ori expresia de mai sus şi punând t=0 în derivate găsim:

$$\begin{cases} r_1 + r_2 + ... + r_m = 0 \\ r_1 w_1 + r_2 w_2 + ... + r_m w_m = 0 \\ \dots\dots\dots\dots\dots \\ r_1 w_1^{m-1} + r_2 w_2^{m-1} + .. + r_m w_m^{m-1} = 0 \end{cases}$$

ceea ce nu se poate decât dacă $r_1 = r_2 = ..r_m = 0$. Contradicţia obţinută arată că reducerea are loc, deci f_1 şi f_2 iau aceleaşi valori, cu aceleaşi probabilităţi.

QED.

Definiţia 2.15 *Spunem că n variabile aleatoare $f_1, ..f_n$ sunt independente dacă pentru orice intervale $I_1, ..I_k$, şi orice alegere $f_{i_1}, ..f_{i_k}$, k≤ n, rezultă că evenimentele $f_{i_1}^{-1}(I_1)$, $f_{i_2}^{-1}(I_2)$, ... , $f_{i_k}^{-1}(I_k)$ sunt independente.*

Exerciţiul 2.16

Dacă variabilele aleatoare $f_1, ..f_n$ sunt independente atunci

$$(f_1 + f_2 + .. + f_n)_c(t) = f_{1c}(t) \cdot f_{2c}(t) \cdot .. \cdot f_{nc}(t)$$

2.2 Spaţiul de probabilitate produs

Fie (X_1,Ω_1, p_1) şi (X_2, Ω_2, p_2) două câmpuri de probabilitate finite. Atunci pe mulţimea $X = X_1 \times X_2$ se poate defini o algebră de mulţimi astfel:

$$\begin{gathered}\Omega = \{A \subset X | A \text{ este reuniune finită de elemente de forma } A_1 \times A_2, \\ \text{cu } A_1 \in \Omega_1 \text{şi } A_2 \in \Omega_2\}\end{gathered}$$

Ţinând seama că Ω_1 şi Ω_2 sunt algebre de evenimente, rezultă că şi Ω este. . Definim $p : \Omega \to R$ prin $p(A_1 \times A_2) = p_1(A_1) \cdot p_2(A_2)$. Dacă A este o reuniune disjunctă de mulţimi de forma $\cup A_{i_i} \times A_{2_i}$ atunci p(A)=$\sum p_1(A_{1_i}) \cdot p_2(A_{2_i})$. Se demonstrează că p este o probabilitate pe Ω iar câmpul (X, Ω, p) se numeşte câmpul produs al câmpurilor (X_1, Ω_1, p_1) şi (X_2, Ω_2, p_2). În mod asemănător se defineşte produsul unui număr finit de câmpuri de probabilitate $(X_1, \Omega_1, p_1), ..$ (X_n, Ω_n, p_n). Spaţiul total este $X = X_1 \times X_2 \times ..X_n$ iar mulţimea de evenimente se defineşte prin $A \in \Omega \Leftrightarrow A$ este o reuniune finită de elemente de forma $A_1 \times A_2$ $\times A_3 \times ... \times A_n$ cu $A_i \in \Omega_i$ pentru orice i. Se arată că Ω este o algebră de evenimente pe care se poate defini o probabilitate care pe evenimentele de forma $A_1 \times A_2 \times .. \times A_n$ are valoarea $p(A_1 \times A_2 \times ..A_n) = p_1(A_1) \cdot p(A_2) \cdot .. \cdot p_n(A_n)$. Este foarte uşor de arătat că evenimentele

$$\begin{array}{c} B_1 = A_1 \times X_2 \times ..X_n \\ B_2 = X_1 \times A_2 \times ..X_n \\ .... \\ B_n = X_1 \times X_2 \times ..A_n \end{array}$$

sunt independente(exerciţiu). De asemenea următoarele variabile aleatoare

$$\begin{array}{c} pr_1(x_1, x_2, ..x_n) = x_1 \\ pr_2(x_1, x_2, ..x_n) = x_n \\ .... \\ pr_n(x_1, x_2, ..x_n) = x_n \end{array}$$

numite proiecţiile canonice ale lui X, sunt independente(exerciţiu). De asemenea dacă $f_k : X_k \to R$ sunt variabile aleatoare atunci variabilele aleatoare $F_k : X \to R$ definite prin $F_k(x_1, x_2, ..x_n) = f_k(x_k)$ sunt independente(exerciţiu).

2.3 Rezumat

Informaţiile dintr-un spaţiu probabilizat se extrag prin funcţii, numite variabile aleatoare. In cazul variabilelor aleatoare ce iau un număr finit de valori (numite şi variabile simple), media este definită într-un mod analog cu media notelor la şcoală. Dispersia e definită ca o măsură a împrăştierii valorilor în jurul mediei. Funcţia caracteristică este o noţiune auxiliară importantă pentru calculul mediei, dispersiei,etc. Ea defineşte unic diagrama variabilei aleatoare.Proprietăţile acestor mărimi sunt listate în cele trei teoreme anterioare. E de remarcat comportarea acestor mărimi la adunarea variabilelor aleatoare independente. Un exemplu important de spaţiu probabilizat şi variabile aleatoare independente este spaţiul produs şi proiecţiile pr_k.

FORMULE UTILIZATE FRECVENT:

a) Definiţiile mediei, momentelor, dispersiei, etc. pentru calculul direct al lor pag(15)

.

b) Proprităţile mediei, dispersiei,funcţiei caracteristice. De remarcat:

i) $M(af+bg) = aM(f) + bM(g)$;
ii) $f \geq 0 \Rightarrow M(f) \geq 0$;
iii) $|M(f)| \leq \max |f|$;
iv) f_1, f_2 independente $\Rightarrow M(f_1 f_2) = M(f_1) M(f_2)$, $D(f_1+f_2) = D(f_1) + D(f_2)$ şi $(f_1+f_2)_c(t) = f_{1c}(t) f_{2c}(t)$;
v) $D(f) = M_2(f) - M^2(f)$
.
vi) $M_k(f) = \frac{1}{(\sqrt{-1})^k} f_c^{(k)}(0)$.

2.4 Exerciţii

1. Fie variabila aleatoare
$$f = \begin{pmatrix} -1 & 2 & 4 & 10 \\ \frac{1}{2} & \frac{1}{10} & \frac{3}{10} & \frac{1}{10} \end{pmatrix}$$
Să se calculeze media, dispersia şi momentul de ordinul 3.

Indicaţie. Se utilizează definiţiile.

2 Fie v.a.
$$f = \begin{pmatrix} -2 & -1 & 2 & 4 & 10 \\ \frac{1}{5} & \frac{1}{10} & \frac{3}{10} & \frac{1}{10} & \alpha \end{pmatrix}$$
Sa se determine α astfel ca f sa fie o v.a. discreta. Să se calculeze media şi dispersia ei.

3. Stiind ca pentru v.a. $f = \begin{pmatrix} -1 & 0 & 1 \\ p_1 & p_2 & p_3 \end{pmatrix}$ avem $M(f) = -\frac{1}{5}$ şi $M(f^2) = \frac{33}{5}$ să se determine p_1, p_2, p_3.

4. Fie o v.a. f cu media 10 . Se cere $b \in R$ astfel ca $M(f+b) = 0$.

5. În N urne se găsesc câte n bile numerotate de la 1 la n, în fiecare. Se extrage câte o bilă din fiecare urnă şi se notează cu f valoarea celui mai mare număr obţinut. Se cere explicitarea probabilităţilor p(f=k), valoarea medie m_n şi limita expresiei m_n/n când n$\rightarrow \infty$.

Indicaţie. $p(f \leq k)$=probabilitatea ca din fiecare urnă să se extragă un număr mai mic sau egal cu k. Numărul cazurilor favorabile este k^N iar numărul celor posibile este n^N, deci probabilitatea este $p(f \leq k) = \frac{k^N}{n^N}$. De aici deducem $p(f=k) = \frac{k^N}{n^N} - \frac{(k-1)^N}{n^N}$. Aplicând definiţia mediei găsim $m_n = n - \sum_{k=1}^{n} \frac{(k-1)^N}{n^N}$ deci $\frac{m_n}{n} \rightarrow 1 - \int_0^1 x^N dx$.

6. O intreprindere recrutează un nou angajat. Se prezintă n candidaţi într-o ordine aleatoare şi primul candidat care trece testul proous de comisie este angajat. Probabilitatea de a trece testul este p pentru fiecare candidat. Fie variabila aleatoaref ce ia valoarea j dacă al j-ulea candidat este admis, şi valoarea 0 dacă nu este nimeni admis.
Să se determine p(f=j), valoarea medie M(f) şi dispersia D(f).

Indicaţie. Ca să fie admis candidatul j trebuie ca toţi candidaţii 1,2,..j-1 să fie respinşi iar j să fie admis. Probabilitatea acestui eveniment este $p(f=j) = q^{j-1}p$, unde $q = 1-p$.

Probabilitatea ca toţi să fie respinşi este q^n. Prin urmare f are diagrama:

$$\begin{pmatrix} 1 & 2 & .. & k & .. & n \\ p & qp & .. & q^{k-1}p & .. & q^n \end{pmatrix}$$

Este foarte simplu acum de a calcula funcţia caracteristică, media şi dispersia.

7. O persoană se deplasează aleator plecând din punctul 0. Cu probabilitatea p face un pas în faţă şi cu probabilitatea q=1-p face un pas înapoi. Valoare unui pas este de 1m iar ritmul este de un pas pe minut. Fie f variabila aleatoare egală cu abscisa la care se ajunge după o oră(un pas în faţă este +1m iar un pas în spate este -1m). Se cer probabilităţile p(f=k), valoarea medie şi dispersia lui f.

Indicaţie. Dacă se fac în faţă k paşi, se ajunge în poziţia k-(60-k)=2k-60. Probabilitatea de a face k paşi în faţă este $C_{60}^k p^k q^{60-k}$(vezi în lecţia 4 despre schema Bernoulli). Deci $p(f = 2k-60) = C_{60}^k p^k q^{60-k}$. Deci funcţia caracteristică este $f_c(t) = \sum_{k=0}^{60} C_{60}^k p^k q^{60-k} e^{i(2k-60)t} = e^{-60ti}\left(pe^{2ti}+q\right)^{60}$. Cu formulele uzuale se deduc acum media, momentele, dispersia.

8. Probabilitatea ca un bec sa reziste la suprasarcina este p=0,90. Se fac 5 incercari de becuri. fie f variabila aleatoare care reprezinta numarul de becuri care au rezistea. Se cere media si dispersia lui f.

9. Se aruncă două zaruri şi se notează cu f suma punctelor obţinute. Se cer media şi dispersia lui f.

Indicaţie. Se poate obţine suma 2 din varianta (1,1), adică un caz din 36. Se poate obţine suma 3 din variantele (1,2) sau (2,1), deci 2 cazuri din 36, etc. De aici rezultă imediat legea lui f etc.

10. O urnă conţine n bile albe şi n bile negre. Se extrag una câte una bilele până se epuizează urna. Fie f variabila aleatoare care pentru o variantă de extragere succesivă a tuturor bilelor ia valoarea k dacă prima bilă albă apare la extragerea k.
a)Să se determine legea lui f
b) Se cere media lui f.

Indicaţie. k ia valori între 1 şi n+1. Ca f să ia valoarea k trebuie să iasă primele k-1 bile negre. Probabilitatea ca să iasă prima bilă neagră este $\frac{n}{2n}$. Condiţionat de aceasta, probabilitatea ca a doua bilă să iasă neagră este $\frac{n-1}{2n-1}$. Deci probabilitatea ca primele două bile să iasă negre este$\frac{n(n-1)}{2n(2n-1)}$. Dacă primele două bile au ieşit negre, probabilitatea ca a treia să iasă neagră este $\frac{n-2}{2n-2}$. Deci probabilitatea ca primele trei bile să iasă negre este $\frac{n(n-1)(n-2)}{2n(2n-1)(2n-2)}$. Analog găsim probabilitatea ca primele k-1 bile să iasă negre $\frac{n(n-1)(n-2)..(n-k+2)}{2n(2n-1)..(2n-k+2)} = \frac{C_n^{k-1}}{C_{2n}^{k-1}}$. Ştiind aceasta, probabilitatea ca la extragerea k să iasă bila albă este $\frac{n}{2n-k+1}$. Deci $p(f=k) = \frac{n}{2n-k+1}\frac{C_n^{k-1}}{C_{2n}^{k-1}}$.

11. Două variabile aleatoare f şi g iau aceleaşi valori $v_1, v_2, ..v_n$. Se mai ştie că $M(f) = M(g)$, $M_2(f) = M_2(g)$,..$M_{n-1}(f) = M_{n-1}(g)$. Să se arate că $p(f = v_k) = p(g = v_k)$ pentru orice k$\in \overline{1..n}$.

Indicaţie. Fie $p_k = p(f = \nu_k)$ şi $p'_k = p(g = \nu_k)$. Avem:

$$M_k(f) = p_1\nu_1^k + p_2\nu_2^k + ...p_n\nu_n^k = M_k(g) = p'_1\nu_1^k + p'_2\nu_2^k + .. + p'_n\nu_n^k$$

pentru k=0,1,2,..n. Prin urmare $p_1, p_2, ..p_n$ şi $p'_1, p'_2, ..p'_n$ satisfac acelaşi sistem nesingular.

12. Să se arate că $D(f + C) = D(f)$, C fiind o constantă.

Indicaţie. Se aplică definiţia dispersiei şi se ţine seama că $M(f + C) = M(f) + C$.

13. Fie X o variabilă aleatoare finită cu media m şi dispersia σ^2. Cum trebuie să fie a şi b (constante) ca variabila aleatoare af+b să aibă media zero şi dispersia unu?

Indicaţie. 0=$M(af+b) = aM(f)+b = am+b$ şi $1 = D(af+b) = D(af) = a^2D(f) = a^2\sigma^2$

14. Două variabile aleatoare independente ξ, η au distribuţiile

$$\xi\begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{8} & \frac{1}{4} & \frac{a}{4} & \frac{a}{2} \end{pmatrix} \quad \eta\begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{2} & b & \frac{1}{8} & \frac{1}{4} \end{pmatrix}$$

a) Sa se determine a şi b

b) Care este distribuţia variabilei $\xi - \eta$?

c) Care este distributia variabilei $2\xi + 3\eta$?

Indicaţie. Pentru ξ avem $\frac{1}{8} + \frac{1}{4} + \frac{a}{4} + \frac{a}{4} = 1$ şi analog pentru η. Pentru $\xi - \eta$ se ţine seama că ξ şi η sunt independente, deci, de exemplu $p(\xi = 1$ şi $\eta = 2) = p(\xi = 1) \cdot p(\eta = 2)$, etc.

15. Doi jucători, A şi B, convin asupra următoarei reguli de joc: la o aruncare cu zarul, dacă apar feţele 1 sau 2 atunci primul jucător dă celui de al doilea 3 lei, iar în caz contrar al doilea jucător dă primului 1 leu. Care este câştigul mediu al fiecărui jucător după n aruncări?

Indicaţie. Probabilitatea de câştig a lui A este 1/3 iar a lui B este de 2/3. Se scriu variabilele aleatoare care reprezinta cştigul pentru A şi B, ca la problema 7.

16. Trei jucători A, B, C pun jos câte o sumă de a, b, respectiv c lei, apoi aruncă o monedă, pe rând, în ordinea A, B, C, A, B, C,... până apare faţa. Primul jucător la care apare faţa ia toţi banii, iar dacă nu a apărut după ce fiecare a aruncat de 3 ori, se anulează jocul.

a) Care sunt câştigurile medii ale fiecărui jucător după 3 jocuri?

b) Cum ar trebui să fie sumele a, b, c ca jocul să fie echitabil?

Lecţia 3

σ Câmpuri de probabilitate

Fie (X, Ω, p) un câmp de probabilitate. Dacă algebra de evenimente verifică în plus axioma:

Oricare ar fi şirul de evenimente $(\mathbf{A}_n)_{n\in N}$, cu $\mathbf{A}_n \in \Omega$ atunci $\cup A_n \in \Omega$,

spunem că Ω este $\sigma-$algebră sau un $\sigma-$câmp de evenimente. Dacă p este o probabilitate pe Ω care în plus verifică relaţia

$$p(\bigcup_{n=1,\infty} A_n) = \sum_{n=1}^{\infty} p(A_n)$$

pentru orice şir A_n de evenimente disjuncte, atunci spunem că p este o σ probabilitate. În aceste condiţii tripletul (X, Ω, p) se numeşte *σ-câmp de probabilitate.*

Teorema 3.1 *Fie (X, Ω, p) un σ -câmp de probabilitate. Atunci:*
1) Dacă $A_n \in \Omega$ pentru orice n, avem şi $\bigcap\limits_{n=1,\infty} A_n \in \Omega$.
2) Dacă $A_1 \supseteq A_2 \supseteq .. \supseteq A_n \supseteq ..$ atunci $p(\bigcap\limits_{n=1,\infty} A_n) = \lim\limits_{n\to\infty} p(A_n)$
4) Dacă $A_1 \subseteq A_2 \subseteq .. \subseteq A_n \subseteq ..$ atunci $p(\bigcup\limits_{n=1,\infty} A_n) = \lim\limits_{n\to\infty} p(A_n)$
3) $p(\bigcup\limits_{n=1,\infty} A_n) \leq \sum_{n=1}^{\infty} p(A_n)$

Demonstraţie. p fiind o probabilitate, toate proprietăţile pentru probabilităţile finit aditive rămân adevărate.

1) $\bigcap\limits_{n=1,\infty} A_n = \overline{\bigcup\limits_{n=1,\infty} \bar{A}_n}$, şi deoarece $\bar{A}_n \in \Omega$, rezultă că reuniunea mulţimilor $\bar{A}_n$ este in Ω, deci şi complementara acestei reuniuni, adică intersecţia mulţimilor A_n este în Ω.

2) Mulţimile $A_k - A_{k+1}$ sunt disjuncte între ele şi sunt disjuncte de $\bigcap\limits_{k=1...\infty} A_k$ iar $A_1 =$

$\left(\bigcap\limits_{k=1..\infty} A_k\right) \cup (A_1 - A_2) \cup (A_2 - A_3) \cup ...$ este o reuniune disjunctă, de unde:

$$p(A_1) = p\left(\bigcap_{k=1..\infty} A_k\right) + \sum_{k=1}^{\infty} p(A_k - A_{k+1})$$

de unde:

$$p\left(\bigcap_{k=1..\infty} A_k\right) = p(A_1) - \lim_{n\to\infty} \sum_{k=1}^{n-1} p(A_k - A_{k+1}) =$$
$$= \lim_{n\to\infty} \left(p(A_1) - \sum_{k=1}^{n-1} p(A_k - A_{k+1})\right) = \lim_{n\to\infty} p(A_n)$$

3) Demonstraţia este asemănătoare cu cea de la punctul 2).

4) $p(A_1 \cup A_2) = p(A_1) + p(A_2) - p(A_1 \cap A_2) \leq p(A_1) + p(A_2)$ şi prin inducţie se demonstrează că $p(A_1 \cup A_2 \cup .. \cup A_n) \leq p(A_1) + p(A_2) + ..p(A_n)$. Acum având $A_1 \subseteq A_1 \cup A_2 \subseteq .. \subseteq (A_1 \cup A_2 .. \cup A_n) \subseteq ..$, rezultă:

$$p\left(\underset{k=1..\infty}{\cup} A_k\right) = \lim_{n\to\infty} p\left(A_1 \cup A_2 \cup .. \cup A_n\right) \leq$$
$$\leq \lim_{n\to\infty} p(A_1) + p(A_2) + ..p(A_n) = \sum_{k=1}^{\infty} p(A_k)$$

QED.

Exemplul 3.2 *Fie $X = [0,1] \times [0,1]$ şi Ω mulţimea submulţimilor lui X cu frontiera formată dintr-un număr finit de segmente. Este clar că Ω este o mulţime închisă la reuniune, intersecţie şi complementară în X, deci este o algebră de mulţimi. O probabilitate pe Ω se poate defini prin formula $p(A) = \frac{aria(A)}{aria(X)} = aria(A)$. Pe de altă parte Ω nu este o σ algebră deorece spre exemplu on cerc plin C nu face parte din Ω deşi este reuniunea unui şir crescător A_n de mulţimi din Ω, cu A_n egal cu poligonul regulat cu 2^n laturi înscris în C. Se demonstrează că există o cea mai mică σ algebră ce conţine pe Ω, fie ea B, iar probabilitatea p (sau aria p) se poate extinde la o σ probabilitate pe B. Probabilitatea (sau aria) astfel definită se numeşte probabilitatea Lebesgue (aria sau măsura Lebesgue).*

In mod asemănător se defineşte măsura Lebesgue pe un segment sau pe un paralelipiped în spaţiu, ca o extindere a noţiunii de lungime a unui segment sau de volum al unui corp mărginit de un număr finit de feţe plane.

3.1 Variabile aleatoare pe σ câmpuri de probabilitate

Definiţia 3.3 *Fie (X, Ω, p) un σ câmp de probabilitate. O funcţie $f : X \to R$ se numeşte variabilă aleatoare dacă pentru orice număr $c \in R$, mulţimea $\{\omega \in \Omega | f(\omega) < c\}$ este un eveniment din Ω. Adesea vom nota mulţimea de mai sus prin $\{f < c\}$ sau $f^{-1}(-\infty, c)$.*

Propoziţia 3.4 *f:X→ R este variabilă aleatoare dacă şi numai dacă este îndeplinită una din condiţiile următoare:*
a)$\forall c \in R \Rightarrow \{f < c\} \in \Omega$
b)$\forall c \in R \Rightarrow \{f \leq c\} \in \Omega$
c) $\forall c \in R \Rightarrow \{f > c\} \in \Omega$
d) $\forall c \in R \Rightarrow \{f \geq c\} \in \Omega$
e) $\forall \alpha, \beta \in R \Rightarrow \{\alpha \leq f < \beta\} \in \Omega$

Demonstraţie.a) este definiţia curentă a variabilei aleatoare. $a) \to b)$ $\{f \leq c\} = \bigcap_n \{f < c + \frac{1}{n}\} \in \Omega$. b)→ $c)$ {f>c}=$X - \{f \leq c\} \in \Omega$ c)→ $d)$ {f≥ $c\} = \bigcap_n \{f > c - \frac{1}{n}\} \in \Omega$ d}→ $e)$ $\{\alpha \leq f < \beta\} = \{f \geq \alpha\} - \{f \geq \beta\} \in \Omega$ e)→ $a)$ {f< $c\} = \bigcup_n \{-n \leq f < c\} \in \Omega$
QED.

Prin urmare o variabilă aleatoare este caracterizată prin faptul că pentru orice interval I⊂ R , închis,deschis,semiînchis,finit sau nu, avem relaţia $f^{-1}(I) \in \Omega$. Spre deosebire de cazul variabilelor aleatoare simple studiate în lecţia trecută, condiţia $f^{-1}(v) \in \Omega$ pentru $v \in R$ nu este suficientă ca f să fie variabilă aleatoare.

Teorema 3.5 *Fie $f : X \to R$ o variabilă aleatoare şi $\alpha \in R$. Atunci următoarele funcţii sunt variabile aleatoare: a)$\alpha + f$; b)$\alpha \cdot f$; c)f^2; d)$\frac{1}{f}$ dacă f$\neq 0$; e) $|f|$.*

Demonstraţie. a) $\{\alpha + f < c\} = \{f < c - \alpha\} \in \Omega$. b) $\{\alpha f < c\} = \{f < \frac{c}{\alpha}\}$ dacă $\alpha > 0$ şi $\{\alpha f < c\} = \{f > c/\alpha\}$ dacă $\alpha < 0$. In ambele cazuri se obţin mulţimi în Ω. c) $\{f^2 < c\} = \{-\sqrt{c} < f < \sqrt{c}\} \in \Omega$. d) $\{\frac{1}{f} < c\} = (\{f > 0\} \cap \{cf > 1\}) \cup (\{f < 0\} \cap \{cf < 1\}) \in \Omega$. e) exerciţiu.
QED.

Teorema 3.6 *Fie $(f_i)_{i \in N}$ un şir de variabile aleatoare. Atunci:*
a)$h(\omega) = \sup\limits_{1 \leq i < \infty} f_i(\omega)$ *este o variabilă aleatoare.*
b) $g(\omega) = \inf\limits_{1 \leq i < \infty} f_i(\omega)$ *este o variabilă aleatoare.*
c) Dacă există $\lim\limits_{i \to \infty} f_i(\omega) = f(\omega)$ *atunci f este o variabilă aleatoare.*

Demonstraţie. a) $\{h > c\} = \bigcup_{i=1}^{\infty} \{f_i > c\} \in \Omega$ b) $\{g < c\} = \bigcup_{i=1}^{\infty} \{f_i < c\} \in \Omega$ c) Fie $\bar{f}_i = \max\limits_{j \geq i} f_j$. Conform cu a) $\bar{f}_i$ este o variabilă aleatoare. $f = \lim\limits_{i \to \infty} f_i = \min\limits_{i} \bar{f}_i$ care este conform cu b) o variabilă aleatoare.
QED.

Teorema 3.7 *Dacă f şi $g$sunt variabile aleatoare atunci $f+g$, fg, $\frac{f}{g}$ sunt variabile aleatoare.*

Demonstraţie. $\{f + g < c\} = \{f < c - g\} = \bigcup_{r \in Q} (\{f < r\} \cap \{c - g > r\}) \in \Omega$, ştiind că toate numerele r∈ Q se pot pune într-un şir. Deci $f+g$ este o variabilă aleatoare. Analog se procedează cu $f-g$. Putem scrie $f\cdot g=\frac{1}{4}\left[(f+g)^2-(f-g)^2\right]$ şi conform punctului anterior fgeste o variabilă aleatoare. Deoarece $\frac{f}{g}=f\cdot\frac{1}{g}$ rezultă că şi $\frac{f}{g}$ este o variabilă aleatoare.

QED.

Din cele de mai sus vedem că dacă a şi b sunt constante atunci $af+bg$ este variabilă aleatoare odată cu f şi g. Deci mulţimea vsriabilelor aleatoare pe o aceeaşi σ algebră formează un spaţiu vectorial. Se poate arăta că dacă f este o variabilă aleatoare şi $F:R\to R$ este o funcţie continuă atunci $F\circ f:X\to R$ este o variabilă aleatoare (demonstraţia este lăsată ca exerciţiu)

Dacă f este o v.a. simplă atunci există o partiţie a spaţiului X=$\bigcup_{i=1}^{n}E_i,\ E_i\cap E_j=\emptyset$ astfel că f este constantă pe E_i. Dacă f şi g sunt variabile aleatoare simple atunci $f\pm g,\ fg,\ f^2,\ f^\alpha,\ |f|$ sunt simple. Orice variabilă aleatoare este limita unui şir de variabile aleatoare simple.

Teorema 3.8 *Fie (X,Ω) o σ algebră şi f o variabilă aleatoare pe X. Atunci există un şir de variabile aleatoare simple $f_n\to f$. Convergenţa este înţeleasă în sensul că pentru orice $x\in X$ rezultă $\lim\limits_{n\to\infty}f_n(x)=f(x)$ în R (convergenţă simplă).*

Demonstraţie. Fie $E_{n,j}=\{\omega\in X|\ \frac{j-1}{2^n}\le f<\frac{j}{2^n}\}\quad -n2^n\le j\le n2^n$. Fie acum f_n definite astfel:

$$f_n(\omega)=\begin{cases}\frac{j-1}{2^n},\ \omega\in E_{nj}\\ n,\ f(\omega)\ge n\\ -n,\ f(\omega)<-n\end{cases}$$

f_n este o variabilă aleatoare simplă şi se vede că $f_n(\omega)\to f(\omega)$. De asemenea se vede că f_n tinde uniform la f dacă f este mărginită. In plus dacă f>0, atunci f_n tinde crescător la f.

QED.

3.2 Media unei variabile aleatoare oarecare

Dacă variabila aleatoare este simplă atunci media ei a fost definită în lecţia anterioară. Fie $f(\omega)=v_i$ dacă $\omega\in E_i$, pentru i=1,2,..n. Atunci media lui f se defineşte prin $M(f)=\sum_i v_i p(E_i)$. Vom mai nota această expresie prin $\int_X f\cdot dp$ sau $\int_X f(\omega)\cdot dp(\omega)$. Dacă variabila aleatoare nu este simplă atunci se aproximează f prin variabile aleatoare simple iar media lui f se aproximează prin media acestora. In continuare vom folosi noţiunea de convergenţă aproape peste tot (pe scurt a.p.t.).

Definiţia 3.9 *Un şir $(f_n)_{n\in N}$ de v.a. se spune că este convergent a.p.t. la variabila aleatoare f dacă şi numai dacă mulţimea $\left\{x\in X|\ \lim\limits_{n\to\infty}f_n(x)\ne f(x)\right\}$ are probabilitatea 0.*

Putem descrie acum felul în care se defineşte media în general:

a) Considerăm un şir de variabile aleatoare simple $(f_n)_{n\in N}$ care converge la f a.p.t.(un asemenea şir există conform teoremei precedente, şi care convege peste tot la f).

b) Stim că $M(f_n) = \int_X f_n dp$ există. Dacă şirul e Cauchy în sensul că $\forall\epsilon > 0, \exists n(\epsilon) \in N, a.\hat{\imath}\forall n, m \in N, n \geq n(\epsilon), m \geq n(\epsilon) \Rightarrow \int_X |f_n - f_m|\, dp < \epsilon$, atunci se arată că exită $\lim\limits_{n\to\infty} \int_X f_n dp$ şi această limită se numeşte media lui f. Ea se notează cu $\int_X f dp$ sau $M(f)$ şi nu depinde de şirul de variabile aleatoare simple care tind Cauchy la f. In cazul particular când f este simplă, se regăsaşte vechea definiţie.

Se demonstrează că proprietăţile mediei demonstrate anterior pentru variabile aleatoare simple se păstrează prin trecere la limită, în general. Avem deci (scriind $\int_X f dp$ în loc de $M(f)$):

1)
$$f = c \Rightarrow \int_X f dp = c$$

2)
$$\int_X (\alpha f + \beta g) dp = \alpha \int_X f dp + \beta \int_X g dp$$

3)
$$\int_X (f_1 + f_2 + .. + f_n) dp = \int_X f_1 dp + \int_X f_2 dp + .. + \int_X f_n dp$$

Spunem că variabilele aleatoare f_1, f_2 sunt independente, la fel ca în cazul variabilelor aleatoare simple, dacă oricare ar fi intervalele I_1, I_2 avem $p(\{f_1 \in I_1\} \cap \{f_2 \in I_2\}) = p(f_1 \in I_1) \cdot p(f_2 \in I_2)$. Cu aceasta putem enunţa următoarea proprietate:

4) Dacă f_1 şi f_2 sunt variabile aleatoare independente atunci:

$$\int_X (f_1 f_2) dp = \left(\int_X f_1 dp\right) \cdot \left(\int_X f_2 dp\right)$$

5)
$$f \geq 0 \Rightarrow \int_X f dp \geq 0$$

6)
$$f \geq g \Rightarrow \int_X f dp \geq \int_X g dp$$

7)
$$\left|\int_X f dp\right| \leq \max |f|$$

Proprietatea următoare este mai specială:

8) Dacă $\lim\limits_{n\to\infty} f_n = f$ a.p.t. şi dacă există o variabilă aleatoare $g \geq 0$ astfel ca $\int_X gdp < \infty$ şi $|f_n| \leq g$ pentru orice n, atunci

$$\lim_{n\to\infty} \int_X f_n dp = \int_X f dp$$

Deci dacă un şir de variabile aleatoare este convergent şi dominat de o v.a. pozitivă cu medie finită, atunci media limitei este egală cu limita mediilor.

O variabilă aleatoare complexă este o funcţie $f : X \to C \quad f = u + iv$ astfel ca u şi v să fie variabile aleatoare reale. Se defineşte media lui f prin $\int_X f dp = \int_X udp + i \int_X vdp$. Se demonstrează că proprietăţile 1)-8) se păstreză.

Momentul de ordin k al unei variabile aleatoare oarecare se defineşte ca pentru variabilele simple

$$M_k(f) = M(f^k) = \int_X f^k dp$$

Momentul centrat de ordinul k se defineşte analog:

$$\mu_k(f) = M\left((f - M(f))^k\right) = \int_X (f - M(f))^k dp$$

Dispersia se defineşte de asemenea ca pentru variabile simple:

$$D(f) = M\left((f - M(f))^2\right) = \int_X (f - M(f))^2 dp = M_2(f) - (M(f))^2$$

Proprietăţile demonstrate pentru dispersie în cazul variabilelor aleatoare simple rămîn adevărate în general, ele deducându-se din proprietăţile mediei, care rămân adevărate în general.

Funcţia caracteristică a unei variabile aleatoare generale se defineşte ca pentru variabile aleatoare simple, cu ajutorul mediei:

$$f_c(t) = M(e^{\sqrt{-1}tf}) = \int_X e^{itf} dp$$

Si în acest caz proprietăţile demonstrate pentru variabile aleatoare simple rămîn adevărate în general.

Unei variabile aleatoare generale nu-i putem ataşa o diagramă ca în cazul variabilelor aleatoare simple, ele luînd în general o infinitate de valori. In cazul particular când valorile se pot enumera $v_1, v_2, ..v_n, ...$ iar evenimentele $E_i = \{f = v_i\}$ au probabilităţile p_i putem asocia lui f diagrama:

$$\begin{pmatrix} v_1 & v_2 & v_3 ... & v_n ... \\ p_1 & p_2 & p_3 ... & p_n ... \end{pmatrix}$$

In acest caz variabilele simple

$$f_n(\omega) = \begin{cases} v_i & \omega \in E_i, i \leq n \\ 0 & \omega \in E_i, i > n \end{cases}$$

tind spre f, şi deci $\int_X f dp = \lim\limits_{n\to\infty} \int_X f_n dp = \lim\limits_{n\to\infty} (v_1p_1 + v_2p_2 + v_3p_3 + ...v_np_n) = \sum_{i=1}^{\infty} v_ip_i = M(f)$. Analog, dispersia este

$$D(f) = \sum_{i=1}^{\infty} (v_i - M(f))^2 p_i$$

Exemplul 3.10 *Fie o variabilă aleatoare ce ia valorile* $0, 1, 2, ...n, ..$ *cu probabilităţile* p_0, p_1, p_2, ... p_n.... *Fie* $G_f(t) = \sum_{i=0}^{\infty} p_it^i$. *Să se arate că:*
a) $M(f) = G'_f(1)$;
b) $M_2(f) = G''_f(1) + G'_f(1)$;
c) $D(f) = G''_f(1) + G'_f(1) - \left(G'_f(1)\right)^2$. *Funcţia* G_f *se numeşte funcţia generatoare a v.a.* f.

Soluţie. a) In acest caz avem $v_i = i$. $G'_f(t) = \sum ip_it^{i-1}$şi deci $G'_f(1) = \sum ip_i = \sum v_ip_i = M(f)$. b) $G''_f(t) = \sum i(i-1)p_it^{i-2}$ deci $G''_f(1) = \sum i^2p_i - \sum ip_i$ adică $G''_f(1) = M_2(f) - M(f)$ de unde rezultă b). c) Avem $D(f) = M_2(f) - (M(f))^2$ ceea ce dă, conform cu a) şi b), exact c).

3.3 Funcţia de repartiţie densitatea de probabilitate

Fie $f : X \to R$ o variaibilă aleatoare definită pe spaţiul probabilizat X. Asociem lui f o funcţie $F : R \to R$ prin formula $F(t) = p(f < t)$. Funcţia Fnu mai apare ca depinzând explicit de X. Ea conţine informaţiile "probabilistice" despre f, independent de natura elementelor din X. Este posibil ca aceeaşi funcţie$F$ să corespundă la variabile aleatoare diferite, definite pe acelaşi spaţiu X sau pe spaţii diferite.

Definiţia 3.11 *Funcţia* $F : R \to R$ *cu* $F(t) = p(f < t)$, f*fiind o variabilă aleatoare, se numeşte funcţia de repartiţie a acestei variabilei aleatoare.*

Definiţia 3.12 *Fie o variabilă aleatoare* f *cu funcţia de repartiţie* F. *O funcţie* $\rho : R \to [0, \infty)$, *integrabilă, cu proprietatea că* $F(t) = \int_{-\infty}^{t} \rho(x)dx$ *se numeşte densitate de repartiţie a varibilei* f.

Teorema 3.13 *Funcţia de repartiţie are următoarele proprietăţi:*
1) F *este monoton crescătoare*

2)$F(-\infty) = \lim_{t\to-\infty} F(t) = 0 \quad F(\infty) = \lim_{t\to\infty} F(t) = 1$
3) F este continuă la stînga.
4) $p(a \le f < b) = F(b) - F(a)$.
5) Reciproc, dacă o funcţie $F : R \to R$ are proprietăţile de mai sus, există un câmp de probabilitate (X, Ω, p) şi o variabilă aleatoare pe X , care are pe F ca funcţie de repartiţie.

Demonstraţie.1) $t_1 < t_2 \Rightarrow \{f < t_1\} \subseteq \{f < t_2\} \Rightarrow p(f < t_1) \le p(f < t_2)$ adică $F(t_1) \le F(t_2)$. 3) Fie $t_n \to t$ monoton crescător. In aceste condiţii mulţimile $\{f < t_n\}$ formează un şir crescător spre $\{f < t\}$ deci $p(f < t) = \lim_{n\to\infty} p(f < t_n)$, adică F este continuă în t la stînga. 2)Şirul de mulţimi $A_n = \{f < t_n\}$ este monoton descrescător cu intersecţia vidă pentru orice şir (t_n) monoton descrescător spre -∞. Din continuitatea probabilităţii, rezultă $F(-\infty) = \lim_{n\to\infty} F(t_n) = \lim_{n\to\infty} p(A_n) = p(\emptyset) = 0$. Analog se demonstrează că $F(\infty) = 1$. 4)Avem $\{f < b\} = \{f < a\} \cup \{a \le f < b\}$, reuniune disjunctă. Luând probabilităţile în ambii membri găsim formula din punctul 4).
5). Se poate lua $X = R$, $\Omega =$ cea mai mică $\sigma-$algebră ce conţine intervalele de forma $[a, b)$, iar probabilitatea este definită prin $p([a, b)) = F(b) - F(a)$. Detaliile nu fac obiectul prezentului curs.

Teorema 3.14 *Densitatea de probabilitate a unei variabile aleatoare are proprietăţile:*
1) $\rho(t) \ge 0$
2) $\int_{-\infty}^{\infty} \rho(t)dt = 1$
3)$p(a \le f < b) = \int_a^b \rho(x)dx$.
4)Reciproc, dacă o funcţie integrabilă $\rho : R \to R$ are proprietăţile 1)-3) atunci există un câmp de probabilitate (X, Ω, p) şi o variabilă aleatoare pe X, care admite pe ρ ca densitate de probabilitate.

Demonstraţie.1) şi 2) sunt evidente iar 3) rezultă din punctul 4) al teoremei precedente. 5) Se poate lua, ca în teorema precedentă, $X = R$, $\Omega =$ cea mai mică $\sigma-$algebră ce conţine intervalele de forma $[a, b)$, iar probabilitatea este definită prin $p([a, b)) = F(b) - F(a)$, unde $F(x) = \int_{-\infty}^{x} \rho(t)\, dt$.

QED.

Funcţia de repartiţie a unei variabile aleatoare există întotdeauna pe când densitatea de repartiţie nu. Dacă există densitate de repartiţie ρ, atunci F este continuă iar în punctele de continuitate ale lui ρ funcţia F este în plus derivabilă şi există relaţia $F'(t) = \rho(t) = \lim_{\Delta t\to 0} \frac{p(t \le f < t+\Delta t)}{\Delta t}$.

Pentru a vedea cum putem calcula efectiv media, dispersia, etc. în general, studiem pe scurt un concept numit integrala Stieltjes.

3.4 Integrala Stieltjes

Fie $F : [a, b] \to R$ o funcţie monoton crescătoare. Vrem să definim $\int_a^b f(x)dF$ pentru o funcţie $f : [a, b] \to R$. Fie

$$\Delta : x_0 < x_1 < ... < x_{n-1} < x_n = b$$

o diviziune a intervalului [a,b] şi fie câte un element $\xi_i \in [x_{i-1}, x_i]$. Definim suma Stieltjes analog cu suma Riemann:

$$S_\Delta = \sum_{i=1}^{n} f(\xi_i)\,(F(x_i) - F(x_{i-1}))$$

Spunem că f este integrabilă Stieltjes în raport cu F dacă pentru orice şir de diviziuni $(\Delta_n)_{n\in N}$ cu norma $|\Delta_n| = \max\limits_i |x_i - x_{i-1}|$ tinzând la zero, şirul de sume Stieltjes are o aceeaşi limită, independent de şirul de diviziuni sau de punctele ξ_i alese. Această limită se notează cu $\int_a^b f(x)dF(x)$ sau $\int_a^b fdF$. Se demonstrează că pentru orice funcţie f continuă integrala Stieltjes există. Următoarele proprietăţi ale integralei Stieltjes seamănă cu cele ale integralei Riemann:

a) Dacă f_1 şi f_2 sunt integrabile Stieltjes atunci $\alpha f_1 + \beta f_2$ este integrabilă şi

$$\int_a^b (\alpha f_1 + \beta f_2)dF = \alpha \int_a^b f_1 dF + \beta \int_a^b f_2 dF.$$

b) Dacă $a < c < b$ atunci $\int_a^b fdF = \int_a^c fdF + \int_c^b fdF$.

c) $f \geq 0$ implică $\int_a^b fdF \geq 0$.

d) $\int_a^b fdF \leq \max\limits_{x\in[a,b]} |f(x)| \cdot (F(b) - F(a))$.

e) Dacă $F = F_1 + F_2$ sumă de două funcţii monotone, atunci $\int_a^b fdF = \int_a^b fdF_1 + \int_a^b fdF_2$.

Demonstraţia acestor proprietăţi nu o facem aici. Să calculăm câteva integrale Stieltjes.

Exemplul 3.15 *Fie $F : [a, b] \to R$, $F(x) = x$. Atunci $F(x_i) - F(x_{i-1}) = x_i - x_{i-1}$ şi integrala Stieltjes devine integrala Riemann.*

Exemplul 3.16 *$F : [a, b] \to R$ este monoton crescătoare, derivabilă, cu derivata $\rho(x) = F'(x)$ integrabilă Riemann. Atunci $F(x_i) - F(x_{i-1}) = \rho(\xi_i^{'})(x_i - x_{i-1})$ (conform teoremei lui Lagrange) şi dacă luăm în suma Stieltjes $\xi_i = \xi_i'$ obţinem suma Riemann pentru funcţia $f\rho$. Deci*

$$\int_a^b fdF = \int_a^b f(x)\rho(x)dx \tag{3.1}$$

Exemplul 3.17 *Fie* $c \in [a,b]$ *şi* $F : [a,b] \to R$ *dată prin*

$$F(x) = \begin{cases} \alpha & x \in [a,c] \\ \beta & x \in (c,b] \end{cases}$$

Atunci în suma Stieltjes numai termenul $f(\xi_i)(x_i - x_{i-1})$ *pentru care* $x_{i-1} \leq c \leq x_i$ *este nenul şi suma este* $(\beta - \alpha)f(\xi_i)$ *care tinde spre* $(\beta - \alpha)f(c)$ *dacă f este continuă în c.Deci*

$$\int_a^b f dF = (\beta - \alpha)f(c) = (F(c+0) - F(c-0)) \cdot f(c) \tag{3.2}$$

Exemplul 3.18 *Fie* $F : [a,b] \to R$ *o funcţie constantă pe porţiuni, cu salturi în punctele* $c_1, c_2, ..c_n$ *unde avem* $F(c_i + 0) - F(c_i - 0) > 0$*. Atunci, la fel ca în exemplul anterior, dacă f este continuă în punctele* c_i *avem*

$$\int_a^b f dF = \sum_{i=1}^{n} f(c_i)\left(F(c_i+0) - F(c_i - 0)\right)$$

Exemplul 3.19 *Fie* F *continuă pe porţiuni, cu discontinuităţi în* $c_1, c_2, ..c_n$ *iar între aceste puncte derivabilă,* $F'(x) = \rho(x)$*, integrabilă. Atunci în* $\int_a^b f dF$ *avem două părţi: una de tipul (3.1) iar alta de tipul (3.2), deci*

$$\int_a^b f dF = \int_a^b f(x)\rho(x)dx + \sum_{i=1}^{n} f(c_i)\left(F(c_i+0) - F(c_i - 0)\right)$$

Se poate defini $\int_{-\infty}^{+\infty} f dF$ prin $\lim\limits_{a \to -\infty, b \to \infty} \int_a^b f dF$. Proprietăţile a)..e) se păstrează la fel ca şi metodele de calcul descrise în exemplele anterioare.

3.5 Media şi funcţia de repartiţie

Fie o variabilă aleatoare f cu valori într-un interval$[a,b)$ şi cu funcţia de repartiţie F. Fie

$$\Delta : x_0 = a < x_1 < .. < x_{n-1} < x_n = b$$

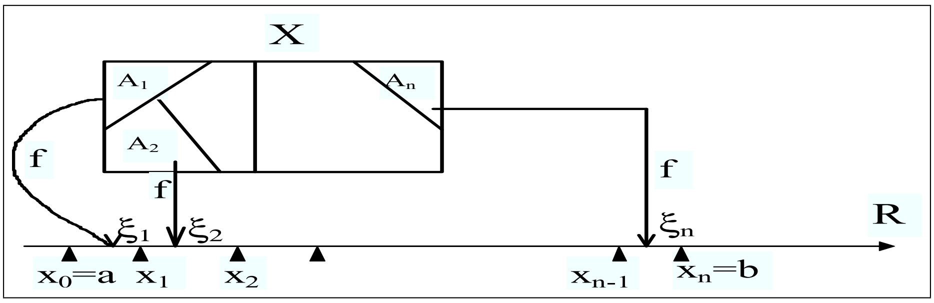
o diviziune a intervalului [a,b]. Luăm cîte un punct $\xi_i \in [x_{i-1}, x_i]$. Fie variabila aleatoare

$$f_\Delta(\omega) = \begin{cases} 0 & dacă \quad f(\omega) \notin [a,b) \\ \xi_i & dacă \quad f(\omega) \in [x_{i-1}, x_i) \end{cases}$$

Altfel spus, dacă A_i=$\{\omega \in X \mid f(\omega) \in [x_{i-1}, x_i)\}$, atunci

$$f_\Delta(\omega) = \begin{cases} \xi_1, \text{ dacă } \omega \in A_1 \\ \xi_2, \text{ dacă } \omega \in A_2 \\ \quad\ldots\ldots\ldots\ldots\ldots \\ \xi_n, \text{ dacă } \omega \in A_n \end{cases}$$

Variabila f_Δ este deci simplă, cu media $M(f_\Delta) = \sum_i \xi_i \cdot p(x_{i-1} \leq f < x_i) = \sum_i \xi_i(F(x_i) - F(x_{i-1}))$. Atunci când norma diviziunii $|\Delta| \to 0$, variabilele f_Δ tind la f (a se vedea figura)



Exemplu de diviziune pentru calculul mediei

deci prin definiţie

$$M(f) = \lim_{|\Delta|\to 0} M(f_\Delta) = \lim_{|\Delta|\to 0} \sum_i \xi_i \left(F(x_i) - F(x_{i-1})\right) = \int_a^b x dF(x) \tag{3.3}$$

Am obţinut astfel formula simplă de mai sus care exprimă media unei variabile aleatoare printr-o integrală Stieltjes. Dacă variabila aleatoare nu este mărginită de a şi b finite, atunci integrala de mai sus trebuie luată între -∞ şi ∞. Variabila aleatoare f^k este aproximată de variabilele aleatoare f_Δ^k care au mediile

$$\sum_i \xi_i^k \left(F(x_i) - F(x_{i-1})\right) \to \int_a^b x^k dF(x)$$

Dispersia lui f este media variabilei $(f - M(f))^2$. Această variabilă aleatoare este aproximată prin variabilele simple $(f_\Delta - M(f))^2$ care au mediile

$$\sum_i \left(\xi_i - M(f)\right)^2 \left(F(x_i) - F(x_{i-1})\right) \to \int_a^b \left(x - M(f)\right)^2 dF(x)$$

Funcţia caracteristică a lui f este media variabilei aleatoare e^{itf} care este aproximată prin variabilele $e^{\sqrt{-1}tf_\Delta}$ cu mediile

$$\sum_i e^{\sqrt{-1}t\xi_i} \left(F(x_i) - F(x_{i-1})\right) \to \int_a^b e^{\sqrt{-1}tx} dF(x)$$

Dacă f nu este mărginită atunci integralele trebuie luate de la -∞ la ∞. Avem deci formulele($\rho = F'$ este densitatea de probabilitate):

$$M(f) = \int_X f dp = \int_{-\infty}^{\infty} x dF(x) = \int_{-\infty}^{\infty} x\rho(x)dx$$

$$M_k(f) = M(f^k) = \int_X f^k dp = \int_{-\infty}^{\infty} x^k dF(x) = \int_{-\infty}^{\infty} x^k \rho(x)dx$$

$$\begin{array}{c} \mu_k(f) = M\left((f - M(f))^k\right) = \int_X (f - M(f))^k \, dp = \\ = \int_{-\infty}^{\infty} (x - M(f))^k \, dF(x) = \int_{-\infty}^{\infty} (x - M(f))^k \, \rho(x)dx \end{array} \qquad (3.4)$$

$$\begin{array}{c} D(f) = M\left((f - M(f))^2\right) = \int_X (f - M(f))^2 \, dp = \\ = \int_{-\infty}^{\infty} (x - M(f))^2 \, dF(x) = \int_{-\infty}^{\infty} (x - M(f))^2 \, \rho(x)dx \end{array}$$

$$f_c(t) = M(e^{\sqrt{-1}tf}) = \int_X e^{\sqrt{-1}tf} dp = \int_{-\infty}^{\infty} e^{\sqrt{-1}tx} dF(x) = \int_{-\infty}^{\infty} e^{\sqrt{-1}tx} \rho(x)dx$$

Prin urmare cunoaşterea funcţiei de repartiţie este suficientă pentru determinarea caracteristicilor numerice ale unei variabile aleatoare. Aşa cum s-a menţionat mai înainte proprietăţile acestor mărimi, demonstrate pentru variabile aleatoare simple, rămîn adevărate în general. Se vede în ultima formulă că funcţia caracteristică este până la un factor transformata Fourier inversă a densităţii de probabilitate, deci, densitatea de probabilitate este determinată de funcţia caracteristică(prin transformata Fourier). Mai precis se demonstrează teorema următoare:

Teorema 3.20 *Fie două variabile aleatoare f şi g, cu funcţiile de repartiţie F, G şi funcţiile caracteristice f_c şi g_c. Dacă funcţiile caracteristice sunt egale, atunci $F = G$.*

Nu demonstrăm această teoremă dar o vom folosi mai departe pentru a pune în evidenţa egalitatea unor funcţii de repartiţie.

Exemplul 3.21 *Fie $X = [0, 1] \times [0, 1]$, Ω = mulţimea figurilor cu arie, $p(A)$ =aria lui A. Fie $f : X \to R$ $f(x, y) = x + y$. Este clar că f este o variabilă aleatoare pentru că $\{f < t\} = \{(x, y) \in X | \, x + y < t\}$ este o figură cu arie. Se cere:*
1) Să se determine un şir mărginit de v.a. simple ce tind la f.
2)Să se determine $\int_X f dp$ utilizând şirul de v.a. de la punctul precedent.
3) Să se determine funcţia de repartiţie şi densitatea de repartiţie(dacă există).
4) Să se determine media şi dispersia folosind funcţia de repartiţie.

Soluţie. 1,2) Putem în mai multe moduri determina şiruri de v.a. simple ce tind la f. De exemplu pentru un n dat fie

$$\begin{array}{rcl} 0 &=& x_0 < x_1 < .. < x_n = 1 \quad x_i = i/n \\ 0 &=& y_0 < y_1 < .. < y_n = 1 \quad y_j = j/n \end{array}$$

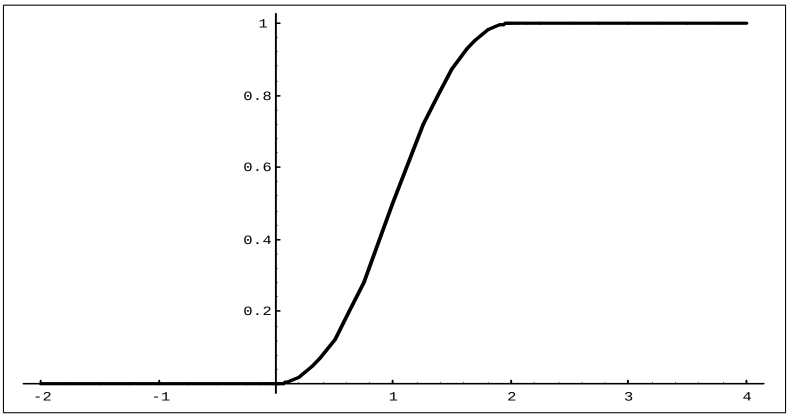
şi fie $f_n(x,y) = (x_{i-1}+x_i)/2 + (y_{j-1}+y_j)/2$ dacă $(x,y) \in [x_{i-1},x_i) \times [y_{j-1},y_j)$, adică valoarea lui $f(x,y)$ în centrul P_{ij} dreptunghiului $[x_{i-1},x_i) \times [y_{j-1},y_j)$ pe care o notăm şi $f(P_{ij})$. Dacă (x,y) nu se află într-un dreptunghi $[x_{i-1},x_i) \times [y_{j-1},y_j)$ atunci definim $f_n(x,y) = 0$. Este clar că $f_n \to f$ a.p.t. şi avem:

$$\begin{aligned}
\int_X f dp &= \lim_{n\to\infty} \int_X f_n dp = \lim_{n\to\infty} \sum_{i,j=1..n} f(P_{ij}) p\left([x_{i-1},x_i) \times [y_{j-1},y_j)\right) = \\
&= \lim_{n\to\infty} \sum_{i,j=1..n} f(P_{ij}) (x_i - x_{i-1}) (y_j - y_{j-1}) = \int\int_X f(x,y) dx dy = \\
&= \int_0^1 \int_0^1 (x+y) dx dy = 1
\end{aligned}$$

3) Funcţia de repartiţie este

$$\begin{aligned}
F(t) &= aria\left(\{(x,y) \in X | x+y < t\}\right) = \\
&= \begin{cases} 0 & t \le 0 \\ \frac{t^2}{2} & 0 \le t \le 1 \\ 1 - \frac{(2-t)^2}{2} & 1 \le t \le 2 \\ 1 & t \ge 2 \end{cases}
\end{aligned}$$

Graficul funcţiei de repartiţie este:

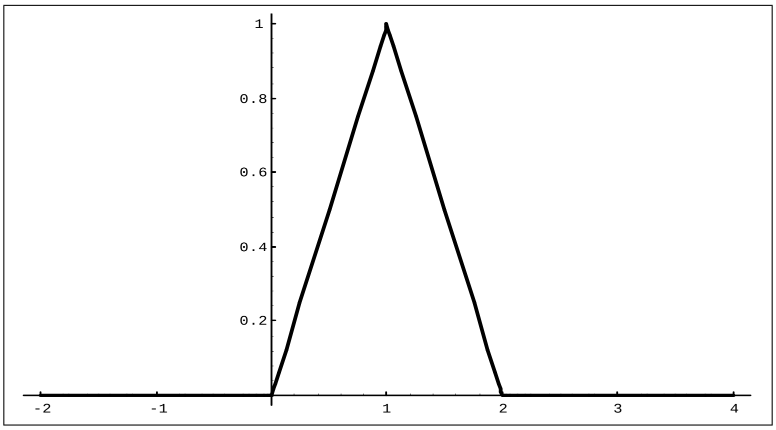


Graficul funcţiei de repartiţie al variabilei aleatoare f

Deoarece F este derivabilă avem densitatea de probabilitate:

$$\rho(t) = \begin{cases} 0 & t \le 0 \\ 2t & 0 \le t \le 1 \\ 2(2-t) & 1 \le t \le 2 \\ 0 & t \ge 2 \end{cases}$$

Graficul densităţii este:



Graficul densităţii de probabilitate a variabilei aleatoare f

4)

$$\begin{aligned} M(f) &= \int_{-\infty}^{\infty} t dF(t) = \int_{-\infty}^{\infty} t\rho(t)dt = \int_{0}^{2} t\rho(t)dt = \\ &= \int_{0}^{1} 2t^2 dt + \int_{1}^{2} t \cdot 2(2-t)dt = 1 \end{aligned}$$

Analog se poate calcula dispersia.

3.6 Rezumat

Pentru a putea lucra şi cu variabile aleatoare cu un număr infinit de valori trbuie lărgit conceptul de algebră de evenimente la cel de σ algebră iar conceptul de probabilitate la cel de σ probabilitate. Concret:

i) O σ algebră de evenimente(mulţimi) este o algebră de evenimente(mulţimi) cu proprietatea că dacă A_n sunt în algebră pentru orice n, atunci $\bigcup\limits_{n=1..\infty} A_n$ este în algebră.

ii) O probabilitate este σ probabilitate dacă pentru orice şir de evenimente (mulţimi) disjuncte A_n avem $p(\bigcup\limits_{n=1..\infty} A_n) = \sum_{n=1}^{\infty} p(A_n)$.

iii) σ probabilitatea are o remarcabilă proprietate de continuitate: dacă A este intersecţia unui şir monoton descrescător sau reuniunea unui şir monoton crescător de evenimente A_n atunci $p(A) = \lim\limits_{n\to\infty} p(A_n)$.

iv) O variabilă aleatoare generală este o funcţie reală care întoarce intervalele în evenimente, deci au sens expresii ca $p(f = a)$ sau $p(a \leq f < b)$ etc. Operaţiile uzuale cu funcţii, inclusiv trecerea la limită, aplicate variabilelor aleatoare duc tot la variabile aleatoare. Orice variabilă aleatoare este limită de variabile aleatoare simple (ce iau un număr finit de valori).

v) Caracteristicile numerice ale unei variabile aleatoare se definesc prin limita caracteristicilor corespunzătoare ale variabilelor aleatoare simple ce tind la variabila generală. Această limită nu există întotdeauna. Proprietăţile acestor caracteristici: medie, dispersie, momente, funcţie caracteristică, demonstrate anterior pentru variabile aleatoare simple se păstrează prin trecere la limită şi pentru variabile aleatoare generale.

vi) Pentru calculul concret al caracteristicilor numerice se introduce funcţia de repartiţie definită prin $F(t) = p(f < t)$, Ffiind funcţia de repartiţie a variabilei aleatoare f. Toate informaţiile numerice legate de f sunt conţinute în funcţia de repartiţie F, independent de spaţiul probabilizat pe care e definită f.

vii) Densitatea de probabilitate ρ, este derivata funcţiei de repartiţie (dacă există), adică $\rho(x) = \lim\limits_{\Delta t \to 0} \frac{p(t \leq f < t + \Delta t)}{\Delta t}$,sau, mai exact se defineşte prin: $F(x) = \int_{-\infty}^{x} \rho(t)dt$. Proprietăţile funcţiei de repartiţie şi ale densităţii de probabilitate sunt studiate mai sus.

viii) Pentru a ajunge la scopul propus, exprimarea caracteristicilor numerice ale unei variabile aleatoare prin intermediul funcţiei de repartiţie, avem nevoie de un concept auxiliar numit integrala Stieltjes

$$\int_a^b f(x)dg(x) = \lim_{|\Delta| \to \infty} \sum f(\xi_i)\left(g(x_i) - g(x_{i-1})\right)$$

, o generalizare a integralei Riemann. Proprietăţile integralei Stieltjes seamănă în mare măsură cu cele ale integralei Riemann.

ix) In final, prin formulele (3.4), media, dispersia, etc. sunt exprimate direct prin funcţia de repartiţie ori prin densitatea de probabilitate, independent de spaţiul pe care e definită variabila aleatoare.

____

FORMULE UTILIZATE FRECVENT:

a) Pentru o probabilitate $\sigma-$aditivă: Dacă $A_1 \supseteq A_2 \supseteq .. \supseteq A_n \supseteq ..$ atunci $p(\bigcap\limits_{n=\overline{1,\infty}} A_n) = \lim\limits_{n\to\infty} p(A_n)$ iar dacă $A_1 \subseteq A_2 \subseteq .. \subseteq A_n \subseteq ..$ atunci $p(\bigcup\limits_{n=\overline{1,\infty}} A_n) = \lim\limits_{n\to\infty} p(A_n)$.

b) $\int_a^b f dF = \int_a^b f(x)\rho(x)dx + \sum_{i=1}^{n} f(c_i)\left(F(c_i+0) - F(c_i-0)\right)$ **unde** F **e funcţia de repartiţie,** ρ **e densitatea de probabilitate iar** c_i **sunt punctele de discontinuitate ale lui** F

c)Formulele (3.4) pentru medie, momente, etc.

d) $p(a \leq f < b) = F(b) - F(a) = \int_a^b \rho(x)\,dx$.

3.7 Exerciţii

1. Fie o variabilă aleatoare :

$$f = \begin{pmatrix} 0 & 1 & 2 & ... & n & ... \\ \alpha p^0 & \alpha p & \alpha p^2 & ... & \alpha p^n & ... \end{pmatrix}$$

a)Să se determine α astfel ca diagrama de mai sus să corespundă unei variabile aleatoare.
b) Să se calculeze media , dispersia şi funcţia caracteristică.

Indicaţie. Trebuie ca $\sum_{k=0}^{\infty} \alpha p^k = 1$. De aici rezultă α. O problemă analoagă este rezolvată în cadrul lecţiei.

2. Fie variabila aleatoare uniformă

$$\begin{pmatrix} 1 & 2 & ... & k & .. & n \\ \frac{1}{n} & \frac{1}{n} & ... & \frac{1}{n} & .. & \frac{1}{n} \end{pmatrix}$$

Să se determine funcţia caracteristică, media şi dispersia.

Indicaţie. Se utilizează definiţiile.

3. O urnă conţine s bile din care s-1 negre şi una albă. Se fac extrageri din urnă cu punerea bilei înapoi pînă se extrage bila albă. Fie variabila aleatoare f care ia valoarea k dacă bila albă apare la extragerea k. Se cere valoarea medie a lui X.

Indicaţie. Probabilitatea de a extrage o bilă albă este p=1/s iar probabilitatea de a extrage o bilă neagră este q=(s-1)/s. Probabilitatea ca bila albă să apară abia la extragerea k este $p_k = \left(\frac{s-1}{s}\right)^{k-1} \cdot \frac{1}{s}$. Mai departe problema seamănă cu 1).

4. Un jucător trage la o ţintă; probabilitatea de a o nimeri este $p \in (0,1)$ iar probabilitatea de a rata lovitura este $q = 1 - p$. Partida se termină când fie numărul lovirilor este cu 2 mai mare ca al ratărilor, caz în care cîştigă partida, fie când numărul ratărilor este cu 2 mai mare ca numărul lovirilor, caz în care pierde partida. Fie A_n evenimentul care constă în a câştiga la tragerea n iar B_n evenimentul care constă în a pierde la tragerea n.
a) $p(A_{2n+1}) =?$ $p(B_{2n+1}) =?$.
b) Care e probabilitatea ca jucătorul sa câştige?
c) Care e probabilitatea ca jucătorul să piardă?
d) Care e probabilitatea ca partida să nu se termine niciodată?
e)Fie X variabila aleatoare care ia valoarea k dacă partida se termină la tragerea k. Se ce probabilităţile p(X=k) şi valoarea medie a lui X.

Indicaţie. Dacă jucătorul câştigă (sau pierde) la mutarea k, atunci nimereşte (sau ratează) cu două lovituri în plus faţă de situaţia contrară. Prin urmare numărul loviturilor este par, deci $p(A_{2n+1}) = p(B_{2n+1}) = 0$. Fie p_{2l} probabilitatea ca partida să fie câştigată la tragerea 2l, fie q_{2l} probabilitatea ca partida să fie pierdută la tragere 2l şi fie r_{2l} probabilitatea de "remiză' la tragerea 2l. Avem $p_{2l} = r_{2l-2} \cdot p^2$ $q_{2l} = r_{2l-2} \cdot q^2$ $r_{2l} = r_{2l-2} \cdot (1 - p^2 - q^2)$.

5. a)Pentru care λ funcţia

$$\rho(x) = \begin{cases} \lambda\,(-x^2+1) & x \in [-1,1] \\ 0 & x \notin [-1,1] \end{cases}$$

este o densitate de probabilitate?
b) Se cere media şi dispersia variabilei cu densitatea ρ.
c) Care este probabilitatea ca o v.a. cu densitatea ρ *să ia valori între 0,2 şi 0,5?*
Indicaţie. λ rezultă din $\int_{-\infty}^{\infty}\rho(x)dx = 1$. In rest se folosesc formulele **(3.4).**

6. a)O variabilă aleatoare f are densitatea

$$\rho(x) = \begin{cases} 0, & x < 0 \\ e^{-x}, & x \geq 0 \end{cases}$$

Care este funcţia de repartiţie şi densitatea de probabilitate a variabilelor f^2, $2f+1$, e^f.

b) Să se arate că dacă o v.a. f *are densitatea* $\rho(x)$ *iar* $g : R \to R$ *este bijecţie derivabilă cu* $g' > 0$, *atunci variabila aleatoare* $g \circ f$ *are densitatea* $\rho\left(g^{-1}(t)\right)\frac{1}{g'(t)}$.

Indicaţie. Fie G funcţia de repartiţie a variabilei f^2. Avem

$$\begin{aligned} G(t) &= p(f^2 < t) = p(-\sqrt{t} < f < \sqrt{t}) = \int_{-\sqrt{t}}^{\sqrt{t}} \rho(x)dx = \\ &= \int_0^{\sqrt{t}} e^{-x}dx = 1 - e^{-\sqrt{t}} \quad \textit{pentru} \quad t \geq 0 \end{aligned}$$

In mod evident G(t)=0 pentru t$\leq$ 0. Densitatea lui f^2 se obţine prin derivarea lui G. Analog se procedează şi cu celelalte expresii de f.

7. Se dă funcţia

$$f(x) = \begin{cases} 0 & x < 0 \\ x - 1 & 0 \leq x \leq 1 \\ x^2 & 1 < x < 2 \\ 5 - x & x \geq 2 \end{cases}$$

Să se calculeze $\int_{-1}^{5}(2+x)df(x)$.

Indicaţie. Se aplică tehnicile de calcul ale integralei Stieltjes din lecţie.

8. Funcţia de repartiţie a unei v.a. continue este

$$F(x) = \begin{cases} a, \ x \leq 0 \\ bx^2, \ x \in (0,1) \\ c, \ x \geq 1 \end{cases}$$

Se cer a,b,c, şi $P\left(1/4 \leq f \leq 1/2)\right)$.

9. Sa admitem că timpul de aşteptare într-o staţie de metrou este o v.a. cu funcţia de repartiţie:

$$F(t) = \begin{cases} 0, \ t \leq 0 \\ t/4, \ 0 < t \leq 2 \\ 1/2, \ 1 < t \leq 4 \\ t/8, \ 4 < t \leq 8 \\ 1, \ t > 8 \end{cases}$$

a) Se cere densitatea de probabilitate.
b) Care este probabilitatea ca un călător să aştepte mai mult de 3 minute?
c) Care este probabilitatea ca un călător să atepte mai mult de 3 minute ştiind că a aşteptat mai mult de 1 minut?
d) Care este probabilitatea ca un călător să atepte mai puţin de 3 minute ştiind că a aşteptat mai mult de 1 minut?

10. Funcţia de repartiţie a unei v.a. X este

$$F_X(t) = \begin{cases} 0,\ t < 0 \\ x^2,\ t \in (0,1] \\ 1,\ t > 1 \end{cases}$$

Se cer $M(X)$, $D(X)$, $\phi_X(t)$, $F_{3X+2}(t)$, $F_{X^2}(t)$.

11. Fie o v.a. cu funcţia de repartiţie F care admite medie şi dispersie finite. Să se arate că

$$\int_{-\infty}^{\infty} (x-a)^2\, dF(x)$$

este minimă când a=M(f).

Indicaţie. Fie $i(a) = \int_{-\infty}^{\infty} (x-a)^2\, dF(x)$. La extrem trebuie să avem $i'(a) = 0$.

12. O variabilă aleatoare f ia valori în intervalul $[a, b]$. *Să se arate că* $M(f) \in [a, b]$ *şi* $D(f) \in [0, (b-a)^2/4]$.

Indicaţie. Se exprimă media şi dispersia prin integrale Stieltjes apoi se majorează corespunzător mărimile de sub integrală. Se poate utiliza problema precedentă.

13. Intr-un sistem de axe rectangulare xOy se dau A(2,0), B(0,1), O(0,0). Se iau la întâmplare $A' \in [O, A]$ *şi* $B' \in [O, B]$. A' *şi* B' *au repartiţii uniforme. Se cer:*
a) Valoarea medie a ariei triunghiului $OA'B'$.
b) Valoarea medie a perimetrului triunghiului $OA'B'$.
c) Valoarea medie a lungimii segmentului $A'B'$.

Indicaţie. Se procedează ca în exemplul 7).

14. Se iau la întâmplare două puncte în intervalul $[0, 1]$. *Care este funcţia de repartiţie a distanţei dintre ele? Care e valoarea medie a acestei distanţe?*

Indicaţie. Se procedează ca la problema 7).

15. O variabilă aleatoare f are densitatea $\rho(x) = \frac{1}{\pi}\frac{1}{1+x^2}$. *Se cere repartiţia variabilei* $g = \min(|f|, 1)$ *precum şi media lui g.*

Indicaţie. Fie G funcţia de repartiţie a lui g. Dacă $t \leq -1$ atunci

$$G(t) = p(g < t) = p(\min(|f|, 1) < t) = p(\emptyset) = 0$$

In mod analog $G(t) = 1$ pentru $t > 1$. Dacă $t \in (0, 1)$ atunci

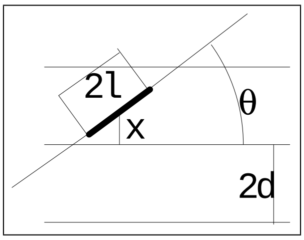
$$\begin{aligned} G(t) &= p(g < t) = p(\min(|f|, 1) < t) = p(|f| < t) \\ &= p(-t < f < t) = \int_{-t}^{t} \frac{1}{\pi} \frac{1}{1 + x^2} dx = \frac{2}{\pi} \arctan(t) \end{aligned}$$

Asemănător se poate calcula G(1) şi apoi media cu formulele**3.4**

16. Doi jucători A şi B, având fiecare un capital de a respectiv b lei, joacă până la ruina unuia din ei. La un joc şansele de câştig sunt p pentru jucătorul A şi q=1-p pentru jucătorul B. La fiecare joc, cel ce pierde plăteşte 1 leu câştigătorului. Care sunt şansele de câştig ale fiecăruia?

17. Problema lui Buffon. Pe un plan orizontal sunt trasate drepte paralele între ele la aceeaşi distanţă $2d$. *Se aruncă pe acest plan un ac de lungime* $2l < 2d$. *Care este probabilitatea ca acul să intersecteze o paralelă oarecare?*

Indicaţie. Starea de intersecţie a acului cu reţeaua de linii depinde de orientarea acului, θ, şi de distanţa x a mijlocului acului faţă de cea mai apropiată linie (vezi figura).



$\theta \in [0, \pi)$ iar $x \in [0, d]$. Pentru intersecţie trebuie ca $x \leq d \cos \theta$. Spaţiul poziţiilor posibile este $X = [0, \pi) \times [0, d]$, iar mulţimea poziţiilor favorabile intersecţiei este $A = \{(\theta, x) \in X \mid x \leq d \cos \theta\}$. In cazul când toate poziţiile sunt egal probabile, probabilitatea unui eveniment este proporţională cu aria mulţimii punctelor prin care se reprezintă acel eveniment.

Lecţia 4

Legi clasice

Am văzut că valorile caracteristice ale unei variabile aleatoare sunt determinate de funcţia de repartiţie (sau de densitatea de probabilitate, dacă există). In cazul variabilelor simple, caracteristicile numerice sunt determinate de diagrama asociată lor. In cele ce urmează sunt studiate câteva legi frecvent întîlnite în aplicaţii. Când discutăm o asemenea lege subînţelegem că există un câmp de probabilitate (X, Ω, p) şi o funcţie $f : X \to R$ care are ca funcţie de repartiţie pe F şi ca densitate pe ρ. Pentru calcule nu este nevoie să avem explicit date X, Ω, p şi f, ci numai pe F sau ρ.

4.1 Repartiţia binomială

Fie variabila aleatoare simplă

$$f = \begin{pmatrix} 0 & 1 & \ldots & k & \ldots & n \\ C_n^o p^0 q^n & C_n^1 p^1 q^{n-1} & & C_n^k p^k q^{n-k} & & C_n^n p^n q^0 \end{pmatrix}$$

unde $p \in [0,1] \quad p+q=1 \quad n \in N$. O asemenea variabilă aleatoare se numeşte binomială. Funcţia caracteristică este

$$f_c(t) = \sum_{k=0}^{n} e^{itk} C_n^k p^k q^{n-k} = \left(pe^{it} + q\right)^n$$

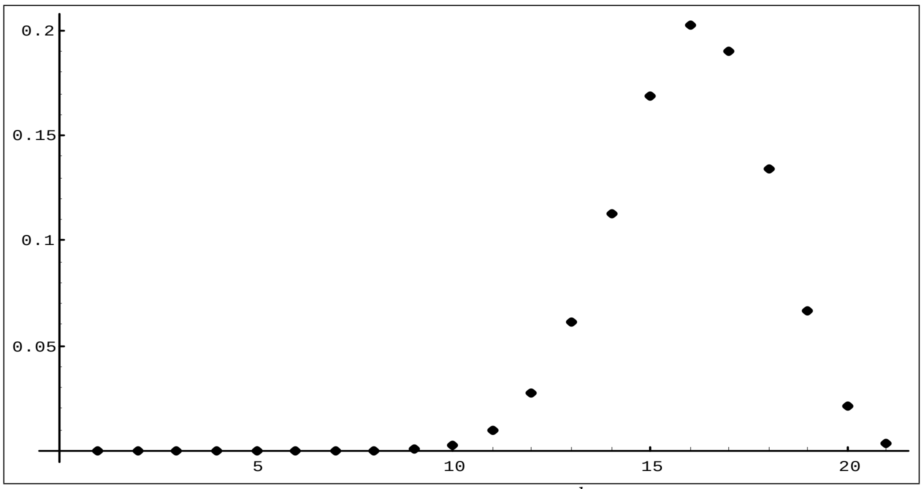
unde $i = \sqrt{-1} \in C$. De aici obţinem:

a) $M(f) = \frac{1}{i} f_c'(0) = \frac{1}{i}\, n \left(pe^{it} + q\right)^{n-1} pie\Big|_{t=0}^{it} = np$

b) $M_2(f) = \frac{1}{i^2} f_c''(0) = n^2p^2 + npq$

c) $D = M_2(f) - \left(M(f)\right)^2 = npq$

Fie p probabilitatea cu care apare A într-o experienţă independentă. Atunci în n experienţe independente apariţia de k ori a evenimentului A se poate face în C_n^k feluri, egal cu numărul de variante în care sunt plasate cele k experienţe unde apare A printre toate cele n experienţe. Probabilitatea fiecărui şir de n experienţe, cu k apariţii ale evenimentului A, este $p^k q^{n-k}$, $q = 1 - p$. Sumând probabilităţile tuturor acestor şiruri favorabile găsim că probabilitatea ca evenimentul A să apară de k ori este $C_n^k p^k (1-p)^{n-k}$. Această distribuţie a fost studiată intensiv de Bernoulli şi se mai numeşte şi repartiţie Bernoulli. Pentru n=20 şi p=3/4 probabilităţile $C_n^k p^k (1-p)^{n-k}$ arată astfel:



Valorile pentru $C_{20}^k \left(\frac{3}{4}\right)^k \frac{1}{2^{20-k}}$

Se vede că probabilităţile care diferă substanţial de 0 apar pentru valori ale lui k în jurul lui np, care în acest caz este 15. In lecţia 5 vom demonstra că acest lucru are loc pentru orice $p \in (0,1)$ şi n suficient de mare.

4.2 Repartiţia Poisson

Spunem că f este o varibilă aleatoare de tip Poisson dacă ia valori întregi pozitive şi $p(f = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ unde $\lambda > 0$. Diagrama unei asemenea variabile aleatoare este

$$f = \begin{pmatrix} 0 & 1 & 2 & \ldots & n & \ldots \\ e^{-\lambda} & e^{-\lambda}\frac{\lambda}{1!} & e^{-\lambda}\frac{\lambda^2}{2!} & \ldots & e^{-\lambda}\frac{\lambda^n}{n!} & \ldots \end{pmatrix}$$

λ se numeşte parametrul variabilei aleatoare. Funcţia caracteristică este

$$f_c(t) = \sum_{k=0}^{\infty} e^{itk} e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\left(e^{it}\lambda\right)^k}{k!} = e^{-\lambda} e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)}$$

Prin urmare:
a) $M(f) = \frac{1}{i} f_c'(0) = \lambda$

b) $M_2(f) = \frac{1}{i^2} f_c''(0) = \lambda^2 + \lambda$
c) $D = M_2 - M^2 = \lambda$
Probabilităţile $e^{-\lambda}\frac{\lambda^k}{k!}$ apar în felul următor:

Propoziţia 4.1 *Fie un şir de variabile aleatoare binomiale*

$$f_n = \begin{pmatrix} 0 & 1 & ... & k & ... & n \\ q^n & C_n^1 pq^{n-1} & ... & C_n^k p^k q^{n-k} & ... & p^n \end{pmatrix}$$

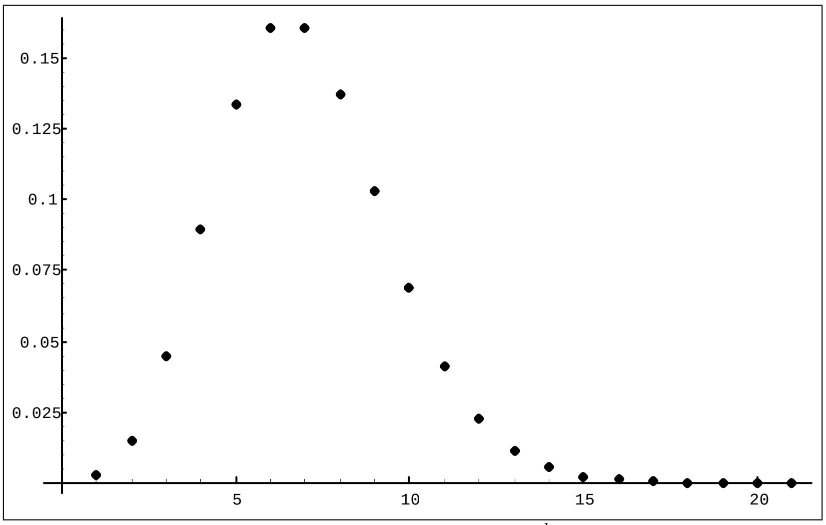
în care $np \to \lambda$ *când* $n \to \infty$*. Atunci* $p(f_n = k) \to e^{-\lambda}\frac{\lambda^k}{k!}$ *când* $n \to \infty$*.*

Demonstraţie. Vom demonstra această afirmaţie pentru $np = \lambda$. Avem

$$p(f_n = k) = C_n^k p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k} =$$
$$= \frac{n(n-1)(n-2)..(n-k+1)}{k!} p^k q^{n-k} = \frac{n^k\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)...\left(1-\frac{k-1}{n}\right)}{k!}\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^{n-k} \to$$
$$\to \frac{\lambda^k}{k!} e^{-\lambda}$$

QED.

Prin urmare $\lambda^k e^{-\lambda}/k!$ reprezintă probabilitatea ca un anumit eveniment A să apară de k ori într-un şir foarte mare de n experienţe independente, probabilitatea de apariţie a evenimentului A într-o singură experienţă find foarte mică, de tipul λ/n. Valoarea medie a numărului de apariţii este $np = \lambda$. Valorile $p(f = k)$ pentru $k = 0, 1, ..10$, la o valoare a parametrului $\lambda = 6$ se reprezintă astfel:



Valorile pentru $e^{-6}\frac{6^k}{k!}$

Se vede că valorile $p(f = k)$ care diferă substanţial de zero se concentrează în jurul valorii k=$\lambda = 6$.

Propoziţia 4.2 *Fie f şi g două variabile Poisson independente, de parametri λ şi μ . Atunci $f+g$ este o variabilă Poisson de parametru $\lambda+\mu$.*

Demonstraţie. Deoarece f şi g sunt independente avem

$$\begin{aligned}(f+g)_c(t) = f_c(t)\cdot g_c(t) = e^{\lambda(e^{it}-1)}\cdot e^{i\mu(e^{it}-1)} = \\ = e^{i(\lambda+\mu)\left(e^{it}-1\right)}\end{aligned}$$

Deoarece funcţia caracteristică determină complet funcţia de repartiţie rezultă că $f+g$ este o variabilă Posson de parametru $\lambda+\mu$.
QED

4.3 Repartiţia uniformă

Spunem că o v.a. are o repartiţie uniformă dacă densitatea ei este:

$$\rho(x) = \begin{cases} \frac{1}{b-a} & x\in[a,b] \\ 0 & x\notin[a,b] \end{cases}$$

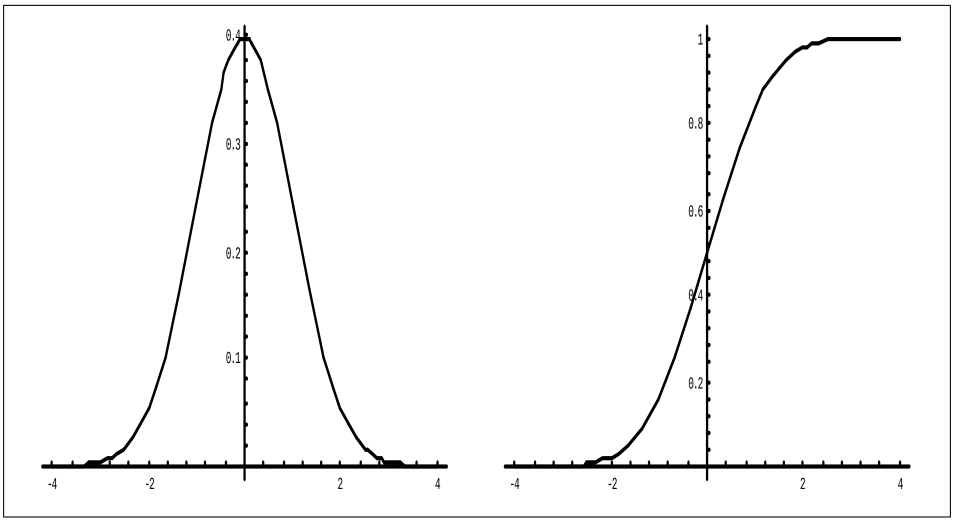
unde $a<b$ sunt reale. Funcţia caracteristică este $\int_a^b e^{itx}\frac{1}{b-a}dx = \frac{e^{ibt}-e^{iat}}{it}$, media este $\frac{b+a}{2}$, etc.

4.4 Repartiţia Normală

Spunem că o variabilă aleatoare f admite o repartiţie normală de parametri m şi $\sigma>0$ dacă densitatea ei de probabilitate este:

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}$$

O variabilă aleatoare cu repartiţia normală, de medie m şi dispersie σ^2 o vom numi de tipul $N(m,\sigma)$. Graficul ρ pentru $m=0$ şi $\sigma=1$ al densităţii de probabilitate şi al funcţiei de repartiţie, arată astfel:



Densitatea de probabilitate şi funcţia de repartiţie a unei variabile normale

Deoarece $\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} dx =$ (prin schimbarea $t = \frac{x-m}{\sigma}$) $= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt = 1$, rezultă că ρ este o densitate de probabilitate.

Vom determina la început funcţia caracteristică.

$$f_c(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-\frac{(x-m)^2}{2\sigma^2}} dx =$$
$$\text{(schimbarea } y = \tfrac{x-m}{\sigma}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{it(m+\sigma y)-\frac{y^2}{2}} dy =$$
$$= e^{itm-\frac{\sigma^2 t^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(y-it\sigma)^2}{2}} dy = e^{itm-\frac{\sigma^2 t^2}{2}}$$

Am folosit

$$\int_{-\infty}^{\infty} e^{-\frac{(y-it\sigma)^2}{2}} dy = \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy = \sqrt{2\pi} \tag{4.1}$$

. Acest lucru se arată astfel:

$$\begin{gathered} 0 = 2\pi i \sum Res\left(e^{-\frac{z^2}{2}}, z_k\right) = \\ = \int_{-R}^{R} e^{-\frac{x^2}{2}} dx + \int_{R}^{R-it\sigma} e^{-\frac{z^2}{2}} dz - \int_{-R}^{-R-it\sigma} e^{-\frac{z^2}{2}} dz + \int_{-R-it\sigma}^{-R} e^{-\frac{z^2}{2}} dz \end{gathered} \tag{4.2}$$

Suma reziduurilor este în dreptunghiul $[-R,\ R,\ R-it\sigma,\ -R-it\sigma]$. Pe latura $[R, R-it\sigma]$ putem scrie $z = R - ist\sigma \quad cu\ 0 \le s \le t\sigma$ deci $z^2 = R^2 - s^2t^2\sigma^2 - i2Rst\sigma$ deci $\int_{R}^{R-it\sigma} e^{-\frac{z^2}{2}} dz = e^{-R^2/2} \int_0^{t\sigma} e^{s^2t^2\sigma^2/2} e^{i2Rst\sigma} (-it\sigma)\, ds \to 0$ când $R \to \infty$. Pe latura $[-R-it\sigma, -R]$ se procedează analog şi se găseşte limita integrale tot zero. Trecând la limită $R \to 0$ în (4.2) găsim prima egalitate din (4.1). A doua egalitate se ştie din anul I.

Cunoscând funcţia caracteristică se pot determina caracteristicile numerice:

a) $M(f) = \frac{1}{i} f_c'(0) = m$

b)$M_2(f) = \frac{1}{i^2} f_c''(0) = \sigma^2 + m^2$

c) $D(f) = M_2 - M^2 = \sigma^2$
d) $\mu_{2k+1} = 0 \quad \mu_{2k} = \frac{(2k)!}{2^k k!}\sigma^{2k}$. (exerciţiu)
In anumite condiţii avem următoarea relaţie între două variabile norale:

Propoziţia 4.3 *a) Fie f_1 şi f_2 două variabile aleatoare normale independente, de parametri m_1, σ_1 respectiv m_2, σ_2. Atunci $f_1 + f_2$ este normală de parametri $m = m_1 + m_2$, $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$.*
b) Dacă $f_1, f_2 ... f_n$ sunt v.a. independente şi normale de tip $N(m, \sigma)$, atunci $\frac{f_1+f_2+...+f_n}{n}$ este v.a. normală de tip $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$.

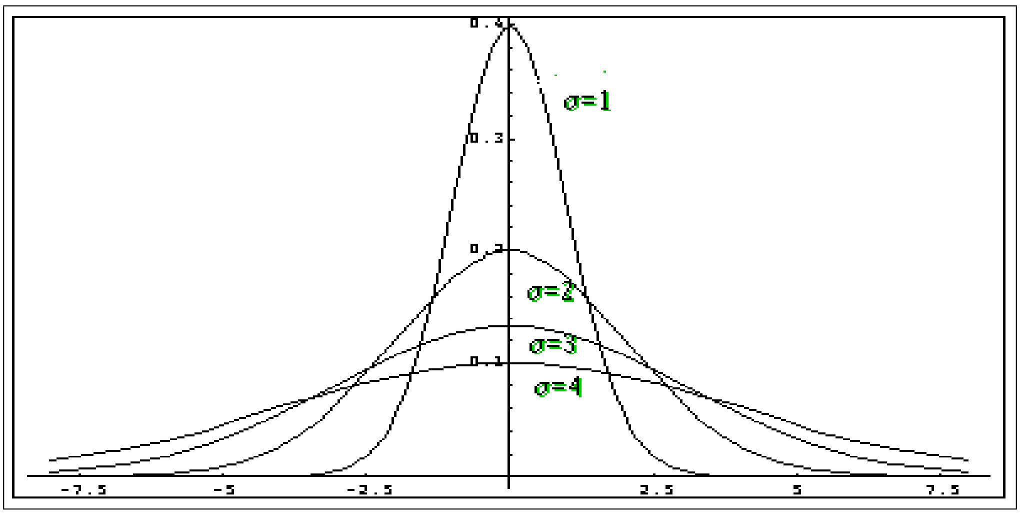
Demonstraţie. a) $f_{1c}(f) = e^{im_1 t - \frac{\sigma_1^1 t^2}{2}} \quad f_{2c}(t) = e^{im_2 - \frac{\sigma_2^2 t^2}{2}}$

$$(f_1 + f_2)_c(t) = f_{1c}(t) f_{2c}(t) = e^{i(m_1+m_2)t - \frac{\left(\sigma_1^2+\sigma_2^2\right)t^2}{2}} = e^{imt - \frac{\sigma^2 t^2}{2}}$$

cu $m = m_1 + m_2 \quad \sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$.
b) Ca la punctul a) găsim $(f_1 + f_2 + ... f_n)_c(t) = e^{i \cdot n \cdot m \cdot t - \frac{n\sigma^2}{2} t^2}$. Din teorema 3, lecţia 2, rezultă

$$\left(\frac{f_1 + f_2 + .. + f_n}{n}\right)_c (t) = (f_1 + f_2 + .. + f_n)_c \left(\frac{t}{n}\right) = e^{imt - \frac{1}{2}\left(\frac{\sigma}{\sqrt{n}}\right)^2 t^2}$$

Cu aceasta b) este demonstrat.
QED.
Pentru $\sigma = 1, 2, 3, 4$ graficele densităţilor apar în continuare. Se vede că pentru σ mare (adică dispersie mare) graficul este mai împrăştiat (aplatizat) în jurul mediei $m = 0$.



Graficele densităţilor unor variabile normale
pentru diverse valori ale lui σ

Legea normală a apărut în legătură cu teoria erorilor de măsurare. Să presupunem că o mărime z este determinată prin măsurători. In mod normal se vor face erori. Fie $p(z,x)dx$ probabilitatea ca să obţinem rezultatul între x ş x+dx, dacă valoarea exactă este z. p(z,x) este deci densitatea de probabilitate a rezultatului măsurătorii. Dacă facem n măsurători independente atunci probabilitatea ca să obţinem rezultatele în intervalele $[x_1 + dx_1]$, $[x_2 + dx_2]$,..$[x_n + dx_n]$ este

$$p(z,x_1)dx_1 p(z,x_2)dx_2 \cdot ..p(z,x_n)dx_n = F(z,x_1,x_2,..x_n)dx_1..dx_n \tag{4.3}$$

Care este expresia cea mai potrivită pentru funcţia $p(z,x)$? In mare Gauss a plecat de la următoarele ipoteze pentru a determina pe $p(z,x)$:[1]

1) $p(z,x) = \phi(z-x)$ (adică erori se fac cu aceeaşi probabilitate în stânga ca şi în dreapta valorii exacte şi probabilitatea de a obţine prin măsurare o anumită valoare x depinde doar de depărtarea faţă de valoarea exactă z).

2) Coeficientul lui $dx_1 dx_2..dx_n$ în n măsurători independente (4.3) este maxim pentru $z = (x_1 + x_2 + .. + x_n)/n$.

3) ϕ este cel puţin de clasă C^2.

Cu aceste ipoteze găsim astfel legea erorilor p(z,x):

$F(z,x_1,...) = \phi(z-x_1)\phi(z-x_2)\cdot ..\cdot \phi(z-x_n) > 0$ este maxim în acelaşi timp cu $\ln(F(z,x_1,x_2,..))$.

Prin urmare:

$$(\ln(F(z,x_1,..))'_x = \frac{\phi'(z-x_1)}{\phi(z-x_1)} + .. + \frac{\phi'(z-x_n)}{\phi(z-x_n)} = 0$$

pentru $z = (x_1 + x_2 + ... + x_n)/n$. Notând $g(x) = \frac{\phi'(z-x)}{\phi(z-x)}$ avem

$$g(x_1) + g(x_2) + ...g(x_n) = 0$$

de fiecare dată când

$$x_1 + x_2 + ... + x_n = nz$$

adică

$$g(x_1) + g(x_2) + .. + g(x_{n-1}) + g(nz - x_1 - x_2 - .. - x_{n-1}) \equiv 0$$

Prin derivare după x_1 obţinem:

$$g'(x_1) - g'(nz - x_1 - .. - x_{n-1}) \equiv 0$$

şi datorită arbitrarităţii valorilor $x_1, ..x_{n-1}$ găsim că

$$g'(x) \equiv a = const$$

[1]Demonstraţia este adaptată după H. Poincaré-Calcul des probabilités, Gauthiers-Villars,Paris,1912

de unde

$$g(x) = ax + b$$

Deoarece $0 = g(x_1) + g(x_2) + .. + g(nz - x_1 - .. - x_{n-1}) = anz + nb$ rezultă $b = -az$, deci:

$$\frac{\phi'(z-x)}{\phi(z-x)} = a(x-z)$$

sau

$$\frac{\phi'(t)}{\phi(t)} = -at$$

de unde

$$\begin{aligned} \ln(\phi(t)) &= -a\frac{t^2}{2} + c \\ \phi(t) &= e^c e^{-a\frac{t^2}{2}} \end{aligned}$$

Deci:

$$p(z,x) = \phi(z-x) = ke^{-a\frac{(z-x)^2}{2}}$$

Deoarece $p(z,x)$ este o densitate de probabilitate trebuie ca

$$\int_{-\infty}^{\infty} p(z,x)dx = 1$$

Aceasta implică $a > 0$ şi $k = \sqrt{\frac{a}{2\pi}}$. Dacă luăm $a = \frac{1}{\sigma^2}$ atunci densitatea p devine:

$$p(z,x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-z)^2}{2\sigma^2}}$$

adică ceea ce am numit legea normală.

Alte motive pentru care legea normală este importantă vom vedea la considerarea teoremelor de tip limită centrală. Fiind frecvent folosită valorile funcţiei de repartiţie pentru $m = 0$, $\sigma = 1$, adică $\int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ au fost tabelate. Mai precis s-a introdus funcţia

$$\Phi(t) = \int_0^t \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

. Cu ajutorul acestei funcţii avem

$$p(a \le f < b) = \int_a^b \rho(x)dx = \Phi(b) - \Phi(a)$$

pentru o v.a. de tipul $N(0,1)$. Φ este o funcţie antisimetrică, $\Phi(a) = -\Phi(-a)$. Valorile funcţiei Φ s-au tabelat pentru diverse valori ale lui t. Dacă f este o v.a. normală de parametri m *şi* σ atunci variabila $Z = \frac{f-m}{\sigma}$ este normală cu media 0 şi dispersia 1 (exerciţiu), deci cu funcţia de repartiţie tabelată.

4.5 Repartiţia exponenţială negativă

Se numeşte astfel o repartiţie cu densitatea

$$\rho(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geq 0 \end{cases}$$

unde $\lambda > 0$.

Plecând de la definiţie se găseşte $f_c(t) = \frac{\lambda}{\lambda - it} \quad M = \frac{1}{\lambda} \quad M_2 = \frac{2}{\lambda^2} \quad D = \frac{1}{\lambda^2}$. Toate aceste calcule sunt lăsate ca exerciţiu.

4.6 Repartiţia Gamma

Se numte astfel o repartiţie cu densitatea

$$\rho_{\alpha,\beta}(x) = \begin{cases} 0 & x \leq 0 \\ \frac{1}{\Gamma(\alpha+1)\beta^{\alpha+1}} x^\alpha e^{-\frac{x}{\beta}} & x > 0 \end{cases}$$

unde $\alpha > -1$ şi $\beta > 0$. Cel mai frecvent se foloseşte cazul $\beta = 1 \quad \alpha + 1 = m$, când avem

$$\rho(x) == \rho_{m+1,0}(x) = \begin{cases} 0 & x \leq 0 \\ \frac{1}{\Gamma(m)} x^{m-1} e^{-x} & x > 0 \end{cases}$$

Reamintim că funcţia gamma se defineşte astfel

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dx$$

pentru $x > 0$. Următoarele proprietăţi se cunosc de la cursul de analiză:

$$\Gamma(x+1) = x\Gamma(x)$$
$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$

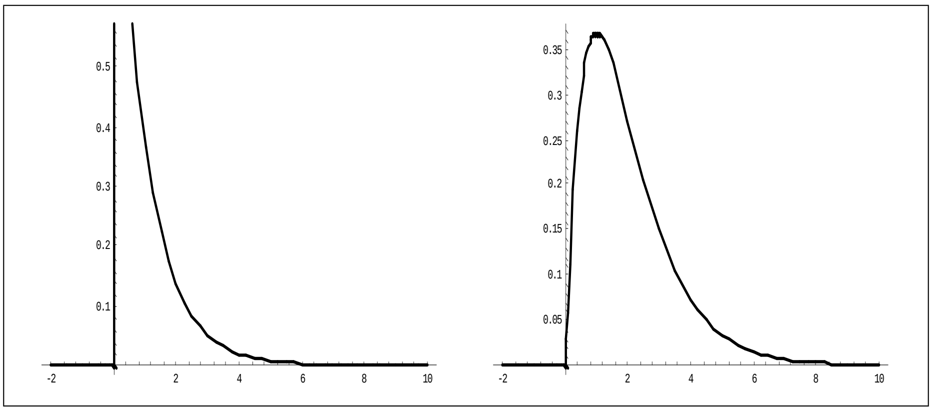
Funcţia caracteristică se calculeazaă astfel:

$$
\begin{aligned}
f_c(t) &= \frac{1}{\Gamma(\alpha+1)\beta^{\alpha+1}}\int_0^\infty e^{itx}x^\alpha e^{-\frac{x}{\beta}}dx \\
&= \frac{1}{\Gamma(\alpha+1)\beta^{\alpha+1}}\int_0^\infty \sum_{n=0}^{\infty}\frac{(itx)^n}{n!}x^\alpha e^{-\frac{x}{\beta}}dx \\
&= \frac{1}{\Gamma(\alpha+1)\beta^{\alpha+1}}\sum_{n=0}^{\infty}\frac{(it)^n}{n!}\int_0^\infty x^{n+\alpha} e^{-\frac{x}{\beta}}dx \\
&= \sum_{n=0}^{\infty}(it)^n\frac{1}{\Gamma(\alpha+1)\beta^{\alpha+1}}\frac{\beta^{n+\alpha+1}}{n!}\int_0^\infty y^{n+\alpha}e^{-y}dy \\
&= \sum_{n=0}^{\infty}(it)^n\frac{1}{\Gamma(\alpha+1)\beta^{\alpha+1}}\frac{\beta^{n+\alpha+1}}{n!}\Gamma(n+\alpha+1) \\
&= \sum_{n=0}^{\infty}(it)^n\frac{\beta^{n+\alpha+1}(n+\alpha)(n+\alpha-1)\cdots(\alpha+1)\Gamma(\alpha+1)}{\Gamma(\alpha+1)\beta^{\alpha+1}n!} \\
&= \sum_{n=0}^{\infty}(-it\beta)^n\cdot\frac{(-\alpha-1)(-\alpha-2)\cdots(\alpha-n)}{n!} = (1-i\beta t)^{-\alpha-1}
\end{aligned}
$$

Am folosit aici dezvoltarea

$$(1+x)^\gamma = 1+\frac{\gamma}{1!}x+\frac{\gamma(\gamma-1)}{2!}x^2+...+\frac{\gamma(\gamma-1)...(\gamma-n+1)}{n!}x^n+..$$

Caracterisicile numerice sunt:
a) $M(f)=\frac{1}{i}f_c'(0)=\beta(\alpha+1)$
b) $M_2(f)=\frac{1}{i^2}f_c''(0)=\beta^2(\alpha+1)(\alpha+2)$
c) $D(f)=M_2(f)-M(f)=\beta^2(\alpha+1)$
In cazul $\beta=1$ $\alpha=m-1$ se obţine $f_c(t)=(1-mt)^{-m}$, media=m, dispersia=m. Tipul de mai sus de variabilă aleatoare îl vom numi $\gamma(\alpha,\beta)$ iar în cazul particular considerat îl vom numi $\gamma(m)$. Pentru $m=0$ şi $m=2$ graficele densităţilor apar în continuare:



Densităţile $\gamma(m)$ pentru $m = 0$ şi $m = 2$.

Propoziţia următoare este simplu de demonstrat:

Propoziţia 4.4 *Dacă f şi g sunt variabile aleatoare independente de tipurile $\gamma(m_1)$ respectiv $\gamma(m_2)$ atunci $f + g$ este de tipul $\gamma(m_1 + m_2)$.*

Demonstraţie. Exerciţiu
QED.

4.7 Repartiţia $\mathfrak{X}^2$(hi pătrat)

Se numete astfel o v.a. cu densitatea :

$$\rho(x) = \begin{cases} 0 & x \leq 0 \\ \frac{1}{2^{\frac{s}{2}}\sigma^s\Gamma(\frac{s}{2})} x^{\frac{s}{2}-1} e^{-\frac{x}{2\sigma^2}} & x > 0 \end{cases}$$

s este un număr natural numit numărul gradelor de libertate, iar $\sigma > 0$. Acest tip de variabilă aleatoare îl vom numi $H(s, \sigma)$; se vede imediat că este acelaşi cu $\gamma(\frac{s}{2} - 1, 2\sigma^2)$. Prin urmare avem:

a) $f_c(t) = (1 - 2\sigma^2 it)^{-\frac{s}{2}}$
b) $M = s\sigma^2$
c) $M_2 = s\,(s + 2)\,\sigma^4$
c) $D = 2s\sigma^4$

Iată cum se poate ajunge la o distribuţie $\mathfrak{X}^2$ placând de la distribuţii normale:

Propoziţia 4.5 *Fie $f_1, f_2, ..f_s$ variabile aleatoare independente de tipu $N(0, \sigma)$. Atunci variabila aleatoare $g = f_1^2 + f_2^2 + .. + f_s^2$ este de tipul $H(s, \sigma)$.*

Demonstraţie. Fie $F_i(t)$ funcţia de repartiţie a variabilei f_i^2. Avem

$$F_i(t) = p(f_i^2 < t) = p(-\sqrt{t} < f_i < \sqrt{t}) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\sqrt{t}}^{\sqrt{t}} e^{-\frac{x^2}{2\sigma^2}}dx = \\ = \frac{2}{\sigma\sqrt{2\pi}}\int_0^{\sqrt{t}} e^{-\frac{x^2}{2\sigma^2}}dx$$

De aici prin derivare determinăm densităţile pentru f_i^2, notate ρ_i. Avem

$$\rho_i(t) = F_i^{'}(t) = \frac{1}{\Gamma(\frac{1}{2})\,(2\sigma^2)^{1/2}} t^{-\frac{1}{2}} e^{-\frac{t}{2\sigma^2}}$$

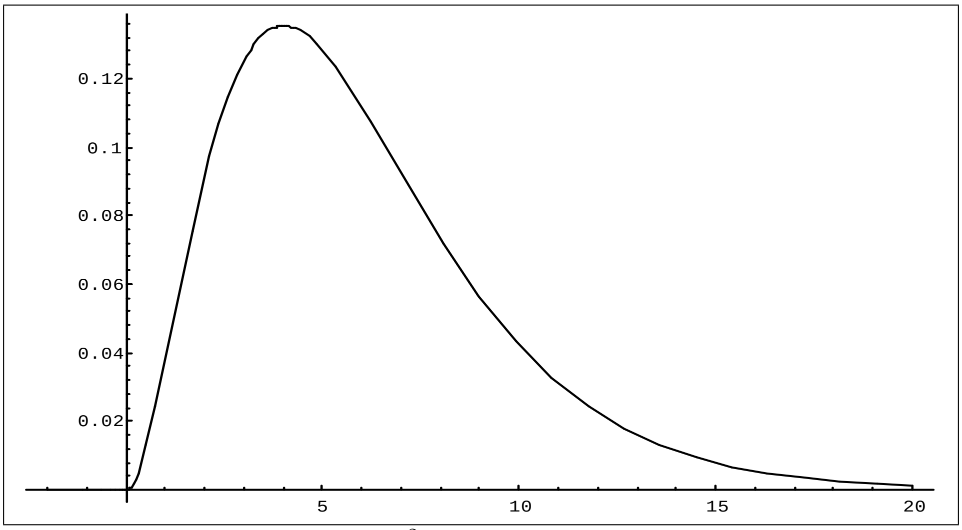
pentru $t > 0$ ceea ce pune în evidenţă faptul că f_i^2 sunt distribuţii de tip $\gamma(-\frac{1}{2}, 2\sigma^2)$, deci $(f_i^2)_c\,(t) = (1 - 2\sigma^2 it)^{-\frac{1}{2}}$. Deoarece f_1^2, $f_2^2, ..f_s^2$ sunt independente rezultă că

$$\left(f_1^2 + f_2^2 + .. + f_s^2\right)_c(t) = \left(f_1^2\right)_c(t) \cdot \left(f_2^2\right)_c(t) \cdots \left(f_s^2\right)_c(t) = \left(1 - 2\sigma^2 ti\right)^{-\frac{s}{2}}$$

care este chiar funcţia caracteristică a unei distribuţii $H(m, \sigma)$.

QED.

Pentru $\sigma = 1$ şi $s = 6$ grade de libertate, graficul densităţii arată astfel:



Densitatea χ^2 pentru $s = 6$ şi $\sigma = 1$.

4.8 Repartiţia Student

O v.a. f, cu densitatea de probabilitate

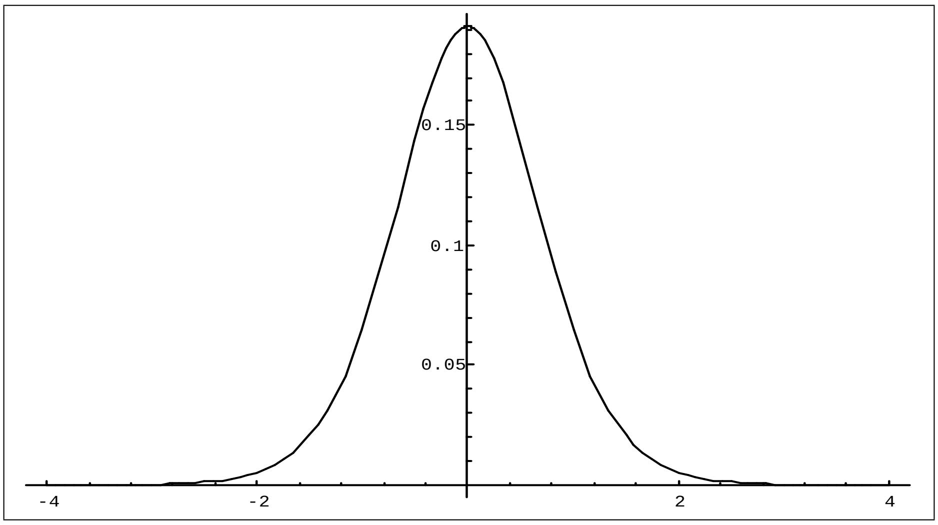
$$\rho(x) = \frac{\Gamma\left(\frac{s+1}{2}\right)}{\sqrt{\pi s}\,\Gamma\left(\frac{s}{2}\right)}\left(1 + \frac{x^2}{s}\right)^{-\frac{s+1}{2}}$$

se numeşte variabilă Student. Tipul acesta de v.a. îl vom nota $S(s)$. Demonstraţiile afirmaţiilor de mai jos în legătură cu distribuţiile Student se vor da în lecţia 6.

a) $M(f) = 0$ şi în general $\mu_{2k+1}(f) = 0$ pentru $0 \leq k < \frac{s}{2}$. Pentru $k \geq \frac{s}{2}$, $\mu_{2k+1}(f)$ nu există.

b) $\mu_{2k}(f) = M_{2k}(f) = \frac{s^r}{\sqrt{\pi}} \frac{\Gamma\left(k+\frac{1}{2}\right)\Gamma\left(\frac{s}{2}-k\right)}{\Gamma\left(\frac{s}{2}\right)} = \frac{s^k 1 \cdot 3 \cdot \ldots \cdot (2k-1)}{(s-2)\cdot(s-4)\cdot\ldots(s-2k)}$.

c) Dacă ξ este o v.a. de tip normal $N(0, \sigma)$ şi η este de tip χ^2cu s grade de libertate, adică de tip $H(s, \sigma)$ atunci v.a.$\frac{\xi}{\sqrt{\frac{\eta}{s}}}$ este de tip Student $S(s)$. Pentru $s = 6$, graficul densităţii este următorul:



Densitatea Student pentru $s = 6$

Această repartiţie este frecvent întâlnită în Statistică şi este tabelată pentru diverse valori ale lui s.

4.9 Rezumat

Variabilele aleatoare anterioare apar în multe modele de probleme reale. De aceea au fost studiate mai în detaliu.

i) Modelul binomial descrie probabilitatea de apariţie de k ori a unui eveniment într-un şir de n experienţe independente, dacă probabilitatea de apariţie a evenimentului într-o experienţă este p. Această probabilitate este $C_n^k p^k (1-p)^{n-k}$.

ii) Modelul Poisson descrie probabilitatea de aparitie de k ori a unui eveniment într-un şir mare de n experienţe independente, dacă probabilitatea de apariţie a evenimentului într-o experienţă este foarte mică, de tipul λ/n. Numărul mediu de apariţii este acelaşi, λ, iar probabilitatea este $e^{-\lambda}\lambda^k/k!$.

iii) Legea normlă, cu densitatea $\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$ este legată de distribuţia erorilor de măsurare. m este valoarea medie a valorii măsurate iar σ^2 este măsura dispersiei valorilor găsite în jurul mediei.

iv) Distribuţia χ^2 cu s grade de libertate are densitatea

$$\rho(x) = \begin{cases} 0 & x \leq 0 \\ \frac{1}{2^{\frac{s}{2}} \sigma^s \Gamma(\frac{s}{2})} x^{\frac{s}{2}-1} e^{-\frac{x}{2\sigma^2}} & x > 0 \end{cases}$$

şi este distribuţia sumei pătratelor a s variabile aleatoare normale independente, de medie 0 şi dispersie σ^2. Această distribuţie apare frecvent în statistică.

Alte legi clasice sunt date în exerciţii sau vor fi studiate mai târziu, pe măsură ce vor fi folosite. E de remarcat utilitatea funcţiei caracteristice în calculul momentelor.

———————————————————————————————

4.10 Exerciţii

1. Să se arate că

$$\mu_{2k} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-m)^{2k} e^{-\frac{(x-m)^2}{2\sigma^2}} dx = (2k-1)\,\sigma^2 \mu_{2k-2}$$

de unde

$$\mu_{2k} = \frac{(2k)!}{2^k k!}$$

Indicaţie. Cu substituţia $t = \frac{x-m}{\sigma}$ se găseşte

$$\mu_{2k} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2k} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2k-1} \left(-e^{-\frac{x^2}{2}}\right)' dx = ...$$

2. Fie o v.a. cu densitatea $\rho(x) = \frac{1}{2\lambda} e^{-\frac{|x-m|}{\lambda}}$ $\dot{\lambda} > 0$ *(distribuţia lui Laplace).*
a)Să se arate că $f_c(t) = \frac{e^{imt}}{1+\lambda^2 t^2}$.
b) Să se determine media şi dispersia.

Indicaţie. Se aplică definiţiile.

3. Să se arate că dacă o v.a. f *este de tipul* $N(m, \sigma^2)$*, atunci* $g = \frac{(f-m)^2}{2\sigma^2}$ *este de tipul* $\gamma(\frac{1}{2}) = \gamma(-\frac{1}{2}, 0)$.

Indicaţie. Se arată mai întâi că h=$\frac{f-m}{\sigma}$ este de tipul $N(0,1)$. Pe urmă după modelul de la distribuţia χ^2 se arată că $g = h^2$ e de tipul cerut.

4. Fie $f_1, f_2, ..f_k$ *variabile aleatoare independente, fiecare având o repartiţie exponenţială negativă.*
a) Să se arate că f=$f_1 + f_2 + .. + f_k$ *are ca funcţie caracteristică* $f_c(t) = \frac{\lambda^k}{(\lambda - it)^k}$.
b) Fie variabila aleatoare g *cu densitatea* $\rho(x) = \frac{\lambda^k}{\Gamma(k)} x^k e^{-\lambda x}$ *dacă* $x \geq 0$ *şi* $\rho(x) = 0$ *dacă* $x < 0$*. Să se arate că* g *este de tipul* $\gamma(\alpha, \beta)$ *pentru un* α *şi* β*. Sa se determine funcţia caracteristică a lui* g.
c) Sa se arate că $f_1 + f_2 + .. + f_k$ *şi* g *au aceleaşi repartiţii.*

Indicaţie. Utilizăm funcţiile caracteristice. Funcţia caracteristică a lui $f_1 + f_2 + .. + f_n$ este produsul funcţiilor caracteristice ale variabilelor $f_1, ..f_n$. Aceste funcţii caracteristice sunt calculate în lecţia prezentă.

5. Să se arate că într-o distribuţie Poisson de parametru λ valoarea k cea mai probabilă satisface inegalitatea $\lambda - 1 \leq k \leq \lambda$.

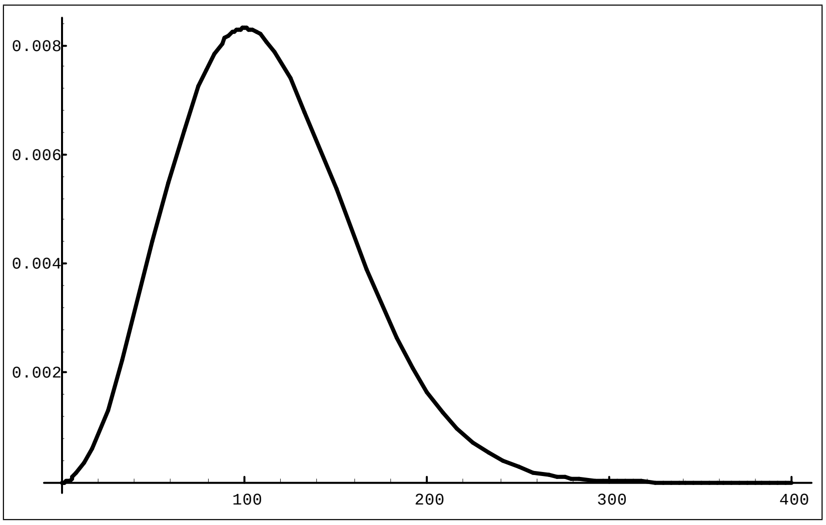
Indicaţie. Se face raportul a două probabilităţi vecine, $p(f = k)$ şi $p(f = k + 1)$.

6. Viteza absolută a unei molecule într-un gaz prfect este o mărime aleatoare. Conform teoriei lui Maxwell densitatea de probabilitate a acestei viteze este:

$$\begin{aligned} \rho(x) &= \frac{4x^2}{c^3\sqrt{\pi}} e^{-\frac{x^2}{c^2}}, \text{ pentru } x > 0 \\ \rho(x) &= 0, \text{ pentru } x \leq 0 \end{aligned}$$

unde c este o constantă. Să se determine viteza medie, energia cinetică medie şi dispersiile lor.

Indicaţie. Pentru c=100, graficul distribuţiei vitezelor este:



Distribuţia vitezelor într-un gaz perfect la c=100

Viteza medie=$\int_{-\infty}^{\infty} x\rho(x)dx$, energia cinetică medie=$\int_{-\infty}^{\infty} \frac{mx^2}{2}\rho(x)dx$, unde m este masa unei molecule. Integralele se fac prin părţi sau se folosesc rezultatele de la legea normală (pb. 1).

7. Distribuţia cu densitatea

$$\rho(x) = \begin{cases} 0 & x \leq 0 \\ \frac{1}{x\beta\sqrt{2\pi}} e^{-\frac{(\ln x - \alpha)^2}{2\beta^2}} & x > 0 \end{cases}$$

se numeşte normal logaritmică. Se cere media şi dispersia unei astfel de distribuţii.

Indicaţie. Dacă ξ este o v.a. cu astfel de lege atunci:

$$\begin{aligned} M(\xi) &= \int_0^\infty x \frac{1}{x\beta\sqrt{2\pi}} e^{-\frac{(\ln x-\alpha)^2}{2\beta^2}} dx \\ &= \frac{1}{\sqrt{2\pi}\beta} \int_{-\infty}^\infty e^{t-\frac{(t-\alpha)^2}{2\beta^2}} dt = e^{\alpha+\frac{\beta^2}{2}}. \end{aligned}$$

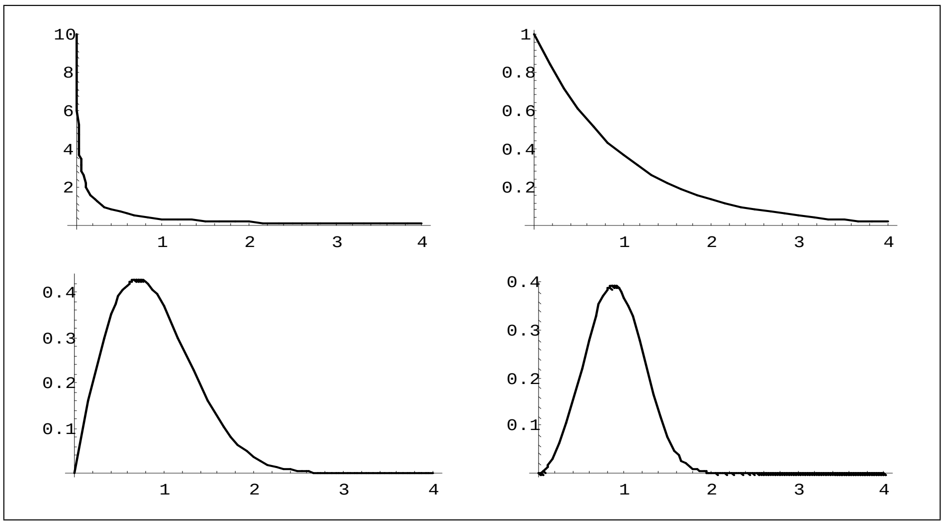
Analog se procedează pentru dispersie, găsindu-se $D(f) = e^{2\alpha+\beta^2}\left(e^{\beta^2}-1\right)$ A.N. Kolmogorov a arătat că aceasta este legea de distribuţie a diametrelor particulelor într-un proces de măcinare.

8. O variabilă aleatoare f cu densitatea

$$\rho(x) = \begin{cases} 0 & x \le 0 \\ c\alpha x^{\alpha-1} e^{-cx^\alpha} & x > 0 \end{cases}$$

se numşte variabilă Weibull.
a) Care media lui f?
b) Care e dispersia lui f?
c) Care este momentul de ordin k al lui f?

Indicaţie. Folosind formulele din lecţia 3 pentru momentul de ordin k şi utilizând schimbarea $x^\alpha = t$ ajungem la o funcţie Γ. Acum media şi dispersia se calculează imediat. In figura următoare sunt reprezentate graficele densităţilor pentru $\alpha = 0,5;\ 1;\ 2;\ 3$ la c=1.



Densitatea Weibull pentru $\alpha = 0,5;\ 1;\ 2;\ 3$.

9. Să se arate că funcţia $\rho(x)$ care maximizează expresia

$$I = -\int_{-\infty}^\infty \rho(x) \ln(\rho(x))\, dx$$

cu condiţiile

$$\int_{-\infty}^{\infty} \rho(x)\, dx = 1 \text{şi} \int_{-\infty}^{\infty} x^2 \rho(x)\, dx = \sigma^2$$

este $\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$. *(I se numeşte entropia densităţii* ρ*. Prin urmare distribuţia normală are entropie maximă, la o dispersie dată).*

Indicaţie. Se ştie din Calculul Variaţional că funcţia $\rho(x)$ care realizează extremul funcţionalei I, cu cele două constrângeri, trebuie să satisfacă ecuaţia $\frac{\partial F}{\partial \rho} - \frac{\partial}{\partial \rho}\left(\frac{\partial F}{\partial \rho'}\right) = 0$, unde $F(x, \rho, \rho') = -\rho \ln \rho + \lambda_1 \rho + \lambda_2 x^2 \rho$.

10. Să se arate că distribuţia pe $[0, \infty)$ *care maximizează entropia şi are media* $m > 0$*, este este exponenţială negativă* $\rho(x) = \frac{1}{m} e^{-\frac{x}{m}}$ *.*

Indicaţie. Trebuie maximizată integrala $I = -\int_0^\infty \rho(x) \ln \rho(x)\, dx$ în condiţiile $\int_0^\infty \rho(x)\, dx = 1$, $\int_0^\infty x\rho(x)\, dx = m$. (vezi problema 9).

11. Să se arate că distribuţia pe intervalul [a,b] care maximizează entropia, este distribuţia uniformă.

12. Intr-o urnă se găsesc a bile albe şi b bile negre. Se extrag n bile din care k sunt albe şi $n - k$ *sunt negre (*$k \leq a$, $n - k \leq b$*). Extragerea se face fără punerea bilei înapoi (schema bilei neîntoarse). Să se arate că probabilitatea acestui eveniment este* $\frac{C_a^k C_b^{n-k}}{C_{a+b}^n}$*. Să se arate că pentru variabila aleatoare*

$$f = \begin{pmatrix} \ldots & k & \ldots \\ \ldots & \frac{C_a^k C_b^{n-k}}{C_{a+b}^n} & \ldots \end{pmatrix}$$

cu $k \in [\max(0, n-b),\ \min(a, n)]$ *media este* $M(f) = n\frac{a}{a+b}$ *şi dispersia este* $D(f) = \frac{a+b-n}{a+b-1} n \frac{a}{a+b} \frac{b}{a+b}$.

Indicaţie. Se utilizează identitatea $\sum_k C_a^k C_b^{n-k} = C_{a+b}^n$ unde însumarea se face între limitele specificate în problemă pentru k (identitatea se poate obţine egalând coeficienţii lui x^n în $(1+x)^a (1+x)^b = (1+x)^{a+b}$). Pentru medie se utilizează $kC_a^k = aC_{a-1}^{k-1}$ iar pentru dispersie se utilizează $k(k-1) C_a^k = a(a-1) C_{a-2}^{k-2}$.

13. Faţă de un adversar la fel de tare ce este mai probabil să se câştige: trei partide din patru sau cinci partide din opt?

14. Pentru o variabilă aleatoare se defineşte asimetria prin $\gamma_1 = \frac{\mu_3}{\sqrt{\mu_2^3}}$ *şi excesul prin* $\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$*. Să se calculeze valorile acestor mărimi pentru variabilele aleatoare din acestă lecţie.*

Lecţia 5

Legi limită

In cele ce urmează vom studia semnificaţia dispersiei, precum şi importanţa repartiţiei normale pentru probabilităţi. Incepem cu inegalitatea lui Cebîşev.

Teorema 5.1 *Fie f o variabilă aleatoare cu medie şi dispersie finite. Atunci pentru $a > 0$ avem:*

$$p\left(|f - M(f)| > a\right) \leq \frac{D(f)}{a^2}$$

sau

$$p\left(|f - M(f)| \leq a\right) \geq 1 - \frac{D(f)}{a^2} \tag{5.1}$$

Demonstraţie.

$$p\left(|f - M(f)| > a\right) = \int_{|x-M(f)|>a} dF(x) \leq \int_{|x-M(f)|>a} \frac{(x-M(f))^2}{a^2} dF(x) \leq$$

$$\leq \frac{1}{a^2} \int_{|x-M(f)|>a} \left(x - M(f)\right)^2 dF(x) \leq \frac{1}{a^2} \int_{-\infty}^{\infty} \left(x - M(f)\right)^2 dF(x) = \frac{D(f)}{a^2}$$

QED.

Putem pune acest lucru şi în alt mod dacă scriem $a = \nu\sigma = \nu\sqrt{D}$. In acest caz avem $\frac{D}{a^2} = \frac{1}{\nu^2}$. Inegalitatea devine

$$\begin{aligned} p\left(|f - M(f)| > \nu\sigma\right) &\leq \frac{1}{\nu^2} \text{ sau} \\ p\left(|f - M\left(f\right)| \leq \nu\sigma\right) &\geq 1 - \frac{1}{v^2} \end{aligned}$$

Aceasta înseamnă că variabila aleatoare ia valori în intervalul $[M(f) - \nu\sigma, M(f) + \nu\sigma]$ cu o probabilitate mai mare ca $1 - \frac{1}{\nu^2}$. In cazul particular al unei variabile normale $f \in N(m, \sigma)$ găsim că f ia valori în intervalul $[m - 3\sigma, m + 3\sigma]$ cu probabilitatea $\geq 1 - \frac{1}{9} = 0,88..$ adică aria de sub graficul densităţii $\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$ cuprins între $x = m - 3\sigma$ şi $x = m + 3\sigma$ este cel puţin $0,88$...Cu cât σ este mai mic, cu atât mai mic este intervalul în care funcţia ia valori cu probabilitate mare (valori în jurul mediei m).

Observaţia 5.2 *Estimarea dată de inegalitatea lui Cebîşev este destul de grosieră în cazuri practice şi se înlocuieşte cu alte estimări mai eficiente. De exemplu*

$$p(-3 < f < 3) = \Phi(3) - \Phi(-3) = 2\Phi(3) = 0,9973..$$

pentru $f \in N(0,1)$ este o estimare mai bună ca cea dată de inegalitatea lui Cebîşev

$$p(-3 < f < 3) \geq 0,88..$$

5.1 Legea numerelor mari

Următoare teoremă este o ilustrare a aplicării inegalităţii lui Cebîşev.

Teorema 5.3 *(Legea numerelor mari)Fie $f_1, f_2, ...f_n, ..$ un şir de v.a. independente două câte două şi cu dispersiile mărginite de aceeaşi constantă: $D(f_n) \leq C$, pentru orice $n \in N$. Atunci pentru orice $\epsilon > 0$ avem:*

$$\lim_{n\to\infty} p\left(\left|\frac{f_1 + f_2 + ... + f_n}{n} - \frac{M(f_1) + ... + M(f_n)}{n}\right| \leq \epsilon\right) = 1 \tag{5.2}$$

Mai precis

$$p\left(\left|\frac{f_1 + f_2 + ... + f_n}{n} - \frac{M(f_1) + ... + M(f_n)}{n}\right| \leq \epsilon\right) \geq 1 - \frac{C}{n\epsilon^2} \tag{5.3}$$

Demonstraţie. Fie $g_n = \frac{f_1+f_2+...+f_n}{n}$. Avem $M(g_n) = \frac{M(f_1)+...+M(f_n)}{n}$ şi $D(g_n) = \frac{1}{n^2}(D(f_1) + ... + D(f_n)) \leq \frac{C}{n^2}$. Inegalitatea lui Cebâşev pentru g_n dă (5.3) iar prin trecere la limită se obţine (5.2).

QED.

Teorema se interpretează astfel: oricare ar fi $\epsilon > 0$, **dacă n este suficient de mare**, atunci funcţia $(f_1 + f_2 + ... + f_n)/n$ diferă cu mai puţin de ϵ de constanta $(M(f_1) + ... + M(f_n))/n$ pe o mulţime cu probabilitate foarte aproape de 1. Dacă $f_1, ...f_n, ..$ au aceeaşi medie m atunci avem:

Corolarul 5.4 *Fie v.a $f_1, f_2, ...f_n, ..$ cu aceeaşi medie m şi dispersiile mărginite de C. Atunci pentru orice $\epsilon > 0$ avem:*

$$\lim_{n\to\infty} p\left(\left|\frac{f_1 + f_2 + ... + f_n}{n} - m\right| \leq \epsilon\right) = 1 \tag{5.4}$$

Un alt caz particular este:

Corolarul 5.5 *(teorema lui Bernoulli) Fie μ_n numărul de apariţii ale unui eveniment A în n experienţe independente. Probabilitatea de realizare a lui A într-o singură experienţă este $\alpha \in [0,1]$. Atunci pentru orice $\epsilon > 0$ avem*

$$\lim_{n\to\infty} p\left(\left|\frac{\mu_n}{n} - \alpha\right| \leq \epsilon\right) = 1 \tag{5.5}$$

Observaţia 5.6 *Cu alte cuvinte pentru un $\epsilon > 0$ oricât de mic, dacă numărul de experienţe n este suficient de mare atunci frecvenţa relativă $\frac{\mu_n}{n}$ este aproape sigur în jurul probabilităţii p de apariţie a evenimentului într-o singură experienţă.*

Demonstrţie. Fie şirul de v.a.

$$f_n(\omega) = \begin{cases} 1 & \text{,dacă la experienţa n s-a realizat A} \\ 0 & \text{, dacă la experinţa n s-a realizat } \bar{\text{A}} \end{cases} \tag{5.6}$$

Aici ω este un şir de experienţe independente. Fiecare v.a. f_n are diagrama

$$f_n \begin{pmatrix} 1 & 0 \\ \alpha & \beta \end{pmatrix}$$

cu$\beta = 1 - \alpha$. Toate variabilele f_n au mediile α şi dispersiile $\alpha\beta$. Ţinând seama că

$$(f_1 + f_2 + .. + f_n) = \mu_n$$

prin aplicarea formulei (5.4) rezultă teorema lui Bernoulli.

O altă demunstraţie se poate da astfel:

i) μ_n are o distribuţie Bernoulli cu $p(\mu_n = k) = C_n^k \alpha^k \beta^{n-k}$.

ii)$M(\mu_n) = n\alpha$ şi $D(\mu_n) = n\alpha\beta$ deci $M(\frac{\mu_n}{n}) = \alpha$ şi $D(\frac{\mu_n}{n}) = \frac{\alpha\beta}{n}$.

iii)Aplicând teorema lui Cebâşev lui μ_n/n găsim

$$\begin{aligned} 1 &\geq p(\left|\frac{\mu_n}{n} - \alpha\right| \leq \epsilon) \geq 1 - \frac{D(\mu_n)}{\epsilon^2} \\ &= 1 - \frac{pq}{n\epsilon^2} \geq 1 - \frac{1}{4n\epsilon^2} \to 1 \end{aligned} \tag{5.7}$$

când $n \to \infty$, pentru că $\alpha\beta = \alpha(1-\alpha) \leq \frac{1}{4}$ atunci când $0 \leq \alpha \leq 1$.

QED.

Observaţia 5.7 *Avem*

$$\sum_{\left|\frac{k}{n} - p\right| \leq \epsilon} C_n^k \alpha^k \beta^{n-k} = p(\left|\frac{\mu_n}{n} - \alpha\right| \leq \epsilon) \geq 1 - \frac{1}{4n\epsilon^2} \tag{5.8}$$

sau

$$\sum_{\left|\frac{k}{n} - \alpha\right| > \epsilon} C_n^k \alpha^k \beta^{n-k} \leq \frac{1}{4n\epsilon^2} \tag{5.9}$$

Datorită frecvenţei schemei binomiale şi dificultăţii de a calcula $C_n^k \alpha^k \beta^{n-k}$ s-au dezvoltat tehnici de calcul aproximativ, mai rapide, pentru aceste mărimi.

5.2 Teoreme limită centrală

Avem nevoie în continuare de următoarea formulă pentru factorial:

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \cdot e^{\frac{\theta_n}{12n}},\ 0 < \theta_n < 1$$

numită formula lui **Stirling.** Demonstraţia acestei formule se poate găsi într-un curs de Analiză matematică.

Teorema 5.8 *(fornula locală de Moivre-Laplace)Fie* $0 < p < 1$ *şi* $x_k = \frac{k-np}{\sqrt{npq}}$, $q = 1-p$. *Atunci*

$$\frac{C_n^k p^k q^{n-k}}{\frac{1}{\sqrt{npq}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x_k^2}{2}}} \to 1 \tag{5.10}$$

când $n \to \infty$, *uniform pentru* $a \le x_k \le b$, *cu* a *şi* b *finite.*

Demonstraţie. Avem:

$$\sqrt{npq}C_n^k p^k q^{n-k} = \sqrt{npq}\sqrt{\frac{n}{2\pi k(n-k)}}\frac{n^n p^k q^{n-k}}{k^k(n-k)^{n-k}} \cdot e^{\theta_{n,k}} \tag{5.11}$$

unde $\theta_{n,k} = \frac{\theta_n}{n} - \frac{\theta_k}{k} - \frac{\theta_{n-k}}{n-k}$.

i) Din $a \le \frac{k-np}{\sqrt{npq}} \le b$ deducem:

$$\begin{aligned} k &\ge np + a\sqrt{npq} = np(1 + a\sqrt{\frac{q}{np}}) \\ n-k &\ge nq - b\sqrt{npq} = nq(1 - b\sqrt{\frac{q}{nq}}) \end{aligned}$$

deci:

$$\begin{aligned} 0 &< |\theta_{n,k}| < \frac{\theta_n}{n} + \frac{\theta_k}{k} + \frac{\theta_{n-k}}{n-k} < \frac{1}{12}(\frac{1}{n} + \frac{1}{k} + \frac{1}{n-k}) < \\ &< \frac{1}{12n}\left(1 + \frac{1}{p + a\sqrt{\frac{pq}{n}}} + \frac{1}{q - \sqrt{\frac{pq}{n}}}\right) \end{aligned}$$

Prin urmare $\theta_{n,k}$ tinde uniform la zero când $n \to \infty$ şi $a \le x_k \le b$, deci $e^{\theta_{n,k}}$ tinde uniform la 1.

ii) Scriind că $k = np + x_k\sqrt{npq}$, găsim:

$$\sqrt{npq}\sqrt{\frac{n}{2\pi k(n-k)}} = \sqrt{\frac{1}{2\pi}}\sqrt{\frac{1}{\left(1 + x_k\sqrt{\frac{q}{np}}\right)\left(1 - x_k\sqrt{\frac{q}{np}}\right)}} \to \frac{1}{\sqrt{2\pi}}$$

uniform pentru $a \leq x_k \leq b$.

iii) Mai avem în (5.11) de aflat limita lui

$$A_{n,k} = \frac{n^n p^k q^{n-k}}{k^k (n-k)^{n-k}}$$

Avem:

$$\begin{aligned}
\ln(A_{n,k}) &= \ln\left(\left(\frac{np}{k}\right)^k \cdot \left(\frac{nq}{n-k}\right)^{n-k}\right) \\
&= -k\ln\left(\frac{k}{np}\right) - (n-k)\ln\left(\frac{n-k}{nq}\right) \\
&= -\left(np + x_k\sqrt{npq}\right)\ln\left(\frac{np + x_k\sqrt{npq}}{np}\right) - \\
&\quad -\left(nq - x_k\sqrt{npq}\right)\ln\left(\frac{nq - x_k\sqrt{npq}}{nq}\right) \\
&= -\left(np + x_k\sqrt{npq}\right)\ln\left(1 + x_k\sqrt{\frac{q}{np}}\right) - \\
&\quad -\left(nq - x_k\sqrt{npq}\right)\ln\left(1 - x_k\sqrt{\frac{p}{nq}}\right)
\end{aligned}$$

Dezvoltând Taylor cei doi logaritmi, găsim:

$$\begin{aligned}
\ln(A_{n,k}) &= -\left(np + x_k\sqrt{npq}\right)\left(x_k\sqrt{\frac{q}{np}} - \frac{1}{2}\frac{x_k^2 q}{np} + O\left(\frac{1}{n^{3/2}}\right)\right) - \\
&\quad -\left(nq - x_k\sqrt{npq}\right)\left(-x_k\sqrt{\frac{p}{nq}} - \frac{1}{2}\frac{x_k^2 p}{nq} + O\left(\frac{1}{n^{3/2}}\right)\right)
\end{aligned}$$

deoarece x_k sunt mărginiţi de a şi b. După ce facem înmulţirile şi reducem termenii rezultă:

$$\ln\left(A_{n,k}\right) = -\frac{x_k^2}{2} + O\left(\frac{1}{\sqrt{n}}\right)$$

Deci

$$A_{n,k} = e^{-\frac{x_k^2}{2}} e^{O\left(1/\sqrt{n}\right)} \rightarrow e^{-\frac{x_k^2}{2}}$$

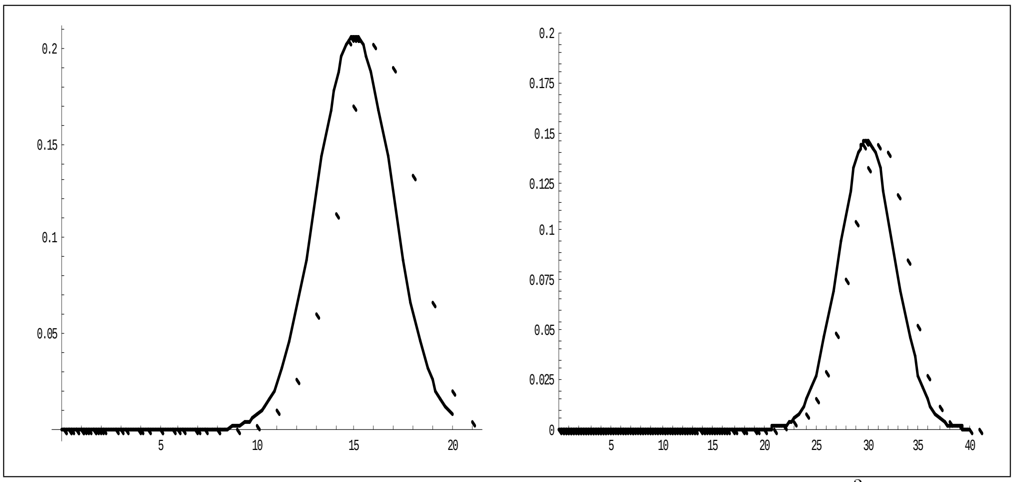
Folosind limitele de la i), ii), iii) găsim:

$$\sqrt{npq}\, C_n^k p^k q^{n-k} \rightarrow \frac{1}{\sqrt{2\pi}} e^{-\frac{x_k^2}{2}}$$

uniform pentru $a \leq x_k \leq b$.

QED.

In figura următoare apar pentru $n = 20$ şi $n = 40$, $p = 3/4$, $q = 1/4$ atât valorile $C_n^k p^k q^{n-k}$ cât şi graficelel funcţiilor $\frac{1}{\sqrt{npq}}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-np)^2}{2npq}}$ care pentru $x = x_k$ au ca valoare $\frac{1}{\sqrt{npq}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x_k^2}{2}}$



Probabilităţile $C_n^k p^k q^{n-k}$ şi graficele funcţiilor $\frac{1}{\sqrt{npq}}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-np)^2}{2npq}}$
pentru p=3/4 şi n=20 respectiv n=40

Teorema de Moivre-Laplace locală se mai poate scrie

$$\sqrt{npq}C_n^k p^k q^{n-k} = \frac{1}{\sqrt{2\pi}}e^{-\frac{x_k^2}{2}}\left(1+\epsilon_{n,k}\right)$$

cu $\epsilon_{n,k} \to 0$ când $n \to \infty$, uniform pentru k astfel încât -$\infty < a \leq x_k \leq b < \infty$.

Forma sub care se utilizează cel mai frecvent acest rezultat este:

Teorema 5.9 *(formula integrală de Moivre-Laplace) Fie* $0 < p < 1, q = 1 - p$. *Atunci:*

$$\lim_{n\to\infty} \sum_{a \leq \frac{k-np}{\sqrt{npq}} \leq b} C_n^k p^k q^{n-k} = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx \tag{5.12}$$

limita fiind uniformă în -$\infty \leq a \leq b \leq \infty$.

Demonstraţie(schiţă) Considerăm cazul a şi b finite. Fie $x_k = \frac{k-np}{\sqrt{npq}}$. Suma din teoremă se scrie:

$$\sum_{a \leq x_k \leq b} \sqrt{npq}C_n^k p^k q^{n-k}\frac{1}{\sqrt{npq}} \tag{5.13}$$

Dar conform formulei locale $\sqrt{npq}C_n^k p^k q^{n-k} = \frac{1}{\sqrt{2\pi}}e^{-\frac{x_k^2}{2}}\left(1+\epsilon_{n,k}\right)$ cu $|\epsilon_{n,k}| \leq \epsilon_n \to 0$ independent de k, dacă $a \leq x_k \leq b$ şi $x_k - x_{k-1} = \frac{1}{\sqrt{npq}}$. Deci

$$\begin{aligned}\sqrt{npq}C_n^k p^k q^{n-k} &= \frac{1}{\sqrt{2\pi}}e^{-\frac{x_k^2}{2}} + \epsilon_{n,k}e^{-\frac{x_k^2}{2}} \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{x_k^2}{2}} + \epsilon'_{n,k}\end{aligned}$$

Deci suma (5.13) se scrie:

$$\sum_{a\le x_k\le b}\frac{1}{\sqrt{2\pi}}e^{-\frac{x_k^2}{2}}(x_k-x_{k-1})+\sum_{a\le x_k\le b}\epsilon_{n,k}e^{-\frac{x_k^2}{2}}(x_k-x_{k-1})$$

Prima sumă de mai sus tinde când $n\to\infty$ către $\int_a^b\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=\Phi(b)-\Phi(a)$ iar a doua e majorată de $\epsilon_n(b-a)$ deci tinde la zero când $n\to\infty$. Prin urmare teorema e demonstrată pentru a şi b finite. O examinare mai atentă a situaţiei arată că teorema este adevărată şi pentru a sau b infinite, iar limita este uniformă în a şi b.

QED.

Care sunt valorile n pentru care aproximarea probabilităţilor $C_n^kp^kq^{n-k}$ prin formulele de Moivre-Laplace este suficient de bună? In general se consideră că:

i) n$\ge$ 30 p$\approx\frac{1}{2}$, atunci formulele de Moivre-Laplace dau o aproximaţie satisfăcătoare.

ii) n$\ge$ 30 , p$\le\frac{1}{10}$ $np=\lambda\le 10$ atunci formulele distribuţiei Poisson dau o aproximaţie suficient de bună: $C_n^kp^kq^{n-k}\approx e^{-\lambda}\frac{\lambda^k}{k!}$.

iii)$n\ge 30$ $p\le\frac{1}{10}$ $np\ge 10$ atunci sunt satifăcătoare formulele de Moivre-Laplace, adică $C_n^kp^kq^{n-k}\simeq\frac{1}{\sqrt{npq}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x_k^2}{2}}$.

Exemplul 5.10 *O firmă de asigurări a distribuit poliţe de asigurare la 10.000 de persoane de aceeaşi vârstă şi grup social contra unei sume de 20.000 lei şi în caz de accident firma îi plăteşte persoanei 2.000.000 lei. Probabilitatea de accident pentru acest grup de persoane este p=6/1000.*
a)Care este probabilitatea ca firma să dea faliment?
b)Care e probabilitatea ca firma să câştige cel puţin 50 milioane lei?

Soluţie. a)Firma încasează 200 milioane lei. Probabilitatea de a da faliment este probabilitatea ca să plătească mai mult de 200 milioane lei, adică numărul celor accidentaţi să depăşească 100. Avem o schemă binomială cu n=10.000, p=0,006, np=60, q=0,994. Se cere suma probabilităţilor $C_n^kp^kq^{n-k}$ pentru k$\ge$ 100. Conform formulei de Moivre-Laplace avem:

$$\begin{aligned}\sum_{k\ge 100}C_n^kp^kq^{n-k} &= \sum_{\frac{100-np}{\sqrt{npq}}\le\frac{k-np}{\sqrt{npq}}\le\frac{n-np}{\sqrt{npq}}}C_n^kp^kq^{n-k}\\ &\approx \int_{\frac{100-60}{\sqrt{10000\cdot 0,006\cdot 0,994}}}^{\frac{10000-60}{\sqrt{10000\cdot 0,006\cdot 0,994}}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx\\ &= \Phi(1287)-\Phi(5,17)\approx 0,5-0,5=0\end{aligned}$$

Probabilitatea de a da faliment este aproape zero.

b) Firma câştigă cel puţin 50 milioane dacă plăteşte sub 150 milioane, deci dacă numărul celor accidentaţi este sub 75. Probabilitatea cerută este deci

$$
\begin{aligned}
p &= \sum_{k\leq 75} C_n^k p^k q^{n-k} = \sum_{\frac{k-np}{\sqrt{npq}}=\frac{0-np}{\sqrt{npq}}}^{\frac{k-np}{\sqrt{npq}}=\frac{75-np}{\sqrt{npq}}} C_n^k p^k q^{n-k} \approx \\
&\approx \int_{x=\frac{0-10000\cdot 0{,}006}{\sqrt{10000\cdot 0{,}006\cdot 0{,}994}}}^{x=\frac{75-10000\cdot 0{,}006}{\sqrt{10000\cdot 0{,}006\cdot 0{,}994}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx = \int_{-5{,}485}^{1{,}371} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx \\
&= \Phi(1,942) - \Phi(-7,769) = \Phi(1,942) + \Phi(7,769) \\
&= 0,474 + 0,5 = 0,974
\end{aligned}
$$

Prin urmare căştigul va fi de peste 50 milioane cu o probabilitate foarte mare.

Alte exemple de aplicare a formulelor de Moivre-Laplace sunt date în exerciţii.

Teoremele de Moivre-Laplace fac parte dintr-o clasă mai largă de teoreme cunoscute sub numele de teoreme limită centrală. Ştim că în anumite condiţii media unui şir de variabile aleatoare independente tinde spre o constantă (legea numerelor mari). Cum tinde această medie de variabile aleatoare spre o constantă? Fie şirul de v.a independente f_n dat de (5.6). Atunci $\mu_n = f_1 + f_2 + ... + f_n$ are o distribuţie Bernoulli, deci $p(\mu_n = k) = C_n^k p^k q^{n-k}$. Pe de o parte ştim din legea numerelor mari că $\frac{f_1+f_2+...+f_n}{n} \to p$. Pe de altă parte

$$
\begin{aligned}
& p\left(a \leq \frac{(f_1 - M(f_1)) + (f_2 - M(f_2)) + ... + (f_n - M(f_n))}{\sqrt{\sum_{k=1}^n D(f_k)}} < b\right) \\
& = p\left(a \leq \frac{n}{\sqrt{\sum_{k=1}^n D(f_k)}}\left(\frac{f_1+..+f_n}{n} - \frac{M(f_1)+..+M(f_n)}{n}\right) < b\right) \\
&= p\left(a \leq \frac{f_1 + f_2 + .. + f_n - np}{\sqrt{\sum_{k=1}^n pq}} < b\right) \\
&= p\left(a \leq \sqrt{\frac{n}{pq}}\left(\frac{f_1 + f_2 + .. + f_n}{n} - p\right) < b\right) \\
&= p\left(a \leq \frac{\mu_n - np}{\sqrt{npq}} < b\right) = \sum_{a\leq\frac{k-np}{\sqrt{npq}}\leq b} p(\mu_n = k) \\
&= \sum_{a\leq\frac{k-np}{\sqrt{npq}}\leq b} C_n^k p^k q^{n-k} \to \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}}\,dx. \text{ (cf. 5.12)}
\end{aligned}
$$

Prin urmare variabila aleatoare $\frac{f_1+f_2+...+f_n}{n} - p$ se comportă, pentru valori mari ale lui n ca o variabilă aleatoare normală, de tip $N(0,1)$ multiplicată cu $\sqrt{\frac{pq}{n}}$. Acest fenomen este mai general.

Anume, relaţia:

$$
\begin{aligned}
& p\left(a \leq \frac{(f_1 - M(f_1)) + (f_2 - M(f_2)) + ... + (f_n - M(f_n))}{\sqrt{\sum_{k=1}^{n} D(f_k)}} < b\right) \\
& = p\left(a \leq \frac{n}{\sqrt{\sum_{k=1}^{n} D(f_k)}}\left(\frac{f_1+..+f_n}{n} - \frac{M(f_1)+..+M(f_n)}{n}\right) < b\right) \\
& \stackrel{n\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx
\end{aligned} \tag{5.14}
$$

are loc dacă:

Varianta I: $(f_k)_{k\in N}$ **este un şir de v.a. pe acelaşi câmp X, independente două câte două, cu aceeaşi funcţie de repartiţie, cu dispersie finită şi nenulă.**

Varianta II: $(f_k)_{k\in N}$ **este un şir de v.a. pe acelaşi câmp X, independente două câte două şi**

$$
\lim_{n\to\infty} \frac{\sqrt[3]{M\left(|f_1 - M(f_1)|^3\right) + ... + M\left(|f_n - M(f_n)|^3\right)}}{\sqrt{\sum_{k=1}^{n} D(f_k)}} = 0
$$

Mai există şi alte variante de condiţii necesare pentru ca relaţia (5.14) să aibă loc. Asemenea teoreme poartă numele de teoreme limită centrală şi pun în evidenţă că în condiţii foarte generale medierea unui număr mare de v.a. independente conduce la o constantă plus o variabilă aleatoare normală. Forma exactă este relaţia (5.14).

5.3 Rezumat

Prima carte de probabilităţi a fost publicată în 1704 de J. Bernoulli unde a fost demonstrată legea numerelor mari sub forma (5.5) şi unde afirmă că s-a gândit 20 de ani la această teoremă.. Ulterior au apărut generalizări ale ei sub diverse forme, ca de exemplu (5.2). O demonstraţie elegantă a acestor teoreme se bazeaza pe inegalitatea lui Cebâşev (5.1).

In esenţă legea numerelor mari spune că în condiţii generale media unui număr mare de v.a. independente diferă puţin de o constantă. Acest lucru este exprimat precis prin formula (5.3).

Teoremele de tip limită centrală aduc o pecizare legii numerelor mari. Anume, diferenţa dintre media unui şir de v.a. independente şi constanta la care convege această medie este aproximativ o v.a. normală, de tip $N(0,1)$ înmulţită cu o constantă care la rândul ei tinde la zero. Forma matematică a acestui enunţ este (5.14).

In aplicaţii apare frecvent schema Bernoulli şi deci calculul probabilităţilor $C_n^k p^k q^{n-k}$. Formulele de Moivre-Laplace (locală 5.10 şi globală 5.12) ne pemit să calculăm aproximativ aceste probabilităţi cu ajutorul densităţii normale $\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ (cazul local) sau primitivei ei, funcţia Φ (cazul global). In unele cazuri $C_n^k p^k q^{n-k}$ se pot aproxima prin probabilitătile legii Poisson. Limitele în care aproximaţiile probabilităţilor $C_n^k p^k q^{n-k}$ prin legea normală sau Poisson sunt considerate satisfăcătoare sunt date în pag. **??**.

Formula lui Stirling este frecvent folosită pentru evaluarea factorialelor. Există şi o generalizare a ei pentru evaluarea funcţiei Gamma.

FORMULE UTILIZATE FRECVENT:

a) Inegalitatea lui Cebâşev: $p\left(|f - M(f)| \leq \epsilon\right) \geq 1 - \frac{D(f)}{\epsilon^2}$**.**

b) Legea numerelor mari: $\lim\limits_{n\to\infty} p\left(\left|\frac{f_1+f_2+\ldots+f_n}{n} - \frac{M(f_1)+\ldots+M(f_n)}{n}\right| \leq \epsilon\right) \geq 1 - \frac{C}{n\epsilon^2}$ **dacă toate dispersiile sunt mărginite de C; varianta Bernoulli:** $\lim\limits_{n\to\infty}\left(\left|\frac{\mu_n}{n} - p\right| \leq \epsilon\right) = 1$ **unde** μ_n **este numărul de apariţii ale unui eveniment A în n experienţe independente iar p este probabilitatea de apariţie a lui A într-un singur experiment.**

c) Formula locală de Moivre-Laplace: $\lim\limits_{n\to\infty} \frac{\sqrt{npq}\,C_n^k p^k q^{n-k}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{x_k^2}{2}}} = 1$**.**

d) Formula globală de Moivre-Laplace: $\lim\limits_{n\to\infty} \sum_{a \leq \frac{k-np}{\sqrt{npq}} \leq b} C_n^k p^k q^{n-k} = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx$**,** **$0<p<1$, q=1-p.**

5.4 Exerciţii

1. Se aruncă o monedă de 10000 ori. Care e probabilitatea ca banul să apară de mai puţin 4550 ori?

Indicaţie. Se procedează ca în exemplul din lecţie.

2. Probabilitatea de a lovi o ţintă este p=1/1000. Care e probabilitatea ca din 5000 de lovituri, ţinta să fie atinsă de cel puţin 2 ori?

Indicaţie. Se utilizează de Moivre-Laplace.

3. Probabilitatea ca un anumit produs să fie defect este p=0,005. Care e probabilitatea ca dintr-un lot de 10000 produse luate la întâmplare să fie mai puţin de 70 defecte?

Indicaţie. Se utilizează formula de Moivre-Laplace.

4. O firmă de asigurări are m salariaţi cu un salariu mediu de s lei pe an. O persoană asigurată cu a lei primeşte o despăgubire de d lei dacă păţeşte ceva. Probabilitatea de a păţi ceva este p. a)Care este probabilitatea ca veniturile companiei să fie în acel an mai mari sau egale cu c_0.b) Care e numărul minim de asiguraţi de la care firma înregistrează un venit brut mai mare ca c_0 cu probabilitatea de cel puţin 0,99?

Indicaţie. Se utilizează formula de Moivre-Laplace.

5. Să se demonstreze că pentru $x > 0$ funcţia $\phi(x) = \int_x^\infty e^{-\frac{t^2}{2}}dt$, satisface inegalităţile:

$$\frac{x}{1+x^2}e^{-\frac{x^2}{2}} \leq \phi(x) \leq \frac{1}{x}e^{-\frac{x^2}{2}}$$

Indicaţie. Pentru prima inegalitate se considetă $f(x) = \frac{x}{1+x^2}e^{-\frac{x^2}{2}} - \Phi(x)$. Avem $f(0) = 0$. Se derivează şi se studiază variaţia lui f. Analog pentru a doua inegalitate.

6. Două vase identice conţin fiecare 10^{22} molecule. Vasele sunt puse în contact şi moleculele se mişcă liber între vase. După omogenizare, care e probabilitatea ca în vasul A să fie cu a zece miliarda parte mai multe molecule ca în B? (admitem că pentru fiecare moleculă probabilitatea de a fi în A este egală cu probabilitatea de a fi în B şi este 1/2).

Indicaţie. Se utilizează formula de Moivre-Laplace şi ex. precedent.

7. Pe scara unui bloc sunt 40 apartamente şi în fiecare apartament este căte un frigider de putere 0,6 kW. In medie un frigider consumă energie electrică 3 ore din 24. Care este probabilitatea ca la un moment dat să fie absorbită de la reţea, de ctre frigidere, o putere mai mare de 4 kW?

Indicaţie. Se utilizează de Moivre-Laplace.

8. a) Să se arate că dacă $\epsilon > 0$ şi $\delta > 0$ sunt date ,iar $n \geq \frac{1}{4\delta\epsilon^2}$ atunci:

$$\sum_{\left|\frac{k}{n}-x\right|>\epsilon} C_n^k x^k (1-x)^{n-k} \leq \delta$$

unde $x \in [0,1]$.

b) Să se arate că dacă $f : [0,1] \to R$ este continuă, deci mărginită $|f| \leq M$ atunci ϵ, δ şi n fiind ca la pct. a) avem:

$$-2M\delta \leq \sum_{\left|\frac{k}{n}-x\right|>\epsilon} C_n^k x^k (1-x)^{n-k} \left(f(\frac{k}{n}) - f(x)\right) \leq 2M\delta$$

c) Să se arate că în condiţiile de la pct. b)

$$-\omega_{x,\epsilon} \leq \left(\sum_{\left|\frac{k}{n}-x\right|\leq\epsilon} C_n^k x^k (1-x)^{n-k} \left(f(\frac{k}{n}) - f(x)\right)\right) \leq \omega_{x,\epsilon}$$

unde $\omega_{x,e}$ este oscilaţia funcţiei f pe intervalul $[x-\epsilon, x+\epsilon]$, adică diferenţa între minimul şi maximul funcţiei pe acest interval.

d) Folosind a), b) şi c) să se arate că şirul de polinoame

$$B_n(x) = \sum_{k=0..n} C_n^k x^k (1-x)^{n-k} f(\frac{k}{n})$$

converge uniform la f pe $[0,1]$ *oricare ar fi funcţia continuă f. (polinoamele lui Bernstein).*

Indicaţie. a) Se utilizează inegalitatea lui Cebâşev pentru schema Bernoulli (vezi obs. 5.9).

b) Se ţine seama de şi de $|f| \le M$. c) Se ţine seama de $\left|f(\frac{k}{n}) - f(x)\right| \le \omega_{x,\epsilon}$ şi $\sum_{k=0..n} C_n^k x^k (1-x)^{n-k} = 1$.

d) Se pune diferenţa $f(x) - B_n(x)$ ca în b) şi c) apoi se ţine seama că f e uniform continuă.

9. Să presupunem că f *este o variabilă aleatoare nenegativă, cu media* m*. Să se arate că este adevărată inegalitatea*

$$p(f \ge c) \le \frac{m}{c}$$

(inegalitatea lui Markov).

10. Fie ν_r *momentul centrat absolut de ordin* r *al unei variabile aleatoare* f *care are media* m*, definit prin* $\nu_r = M(|f-m|^r)$*. Să se arate că*

$$p(|f-m| \ge \varepsilon) \le \frac{\nu_r}{\varepsilon^r}$$

11. Vrem să verificăm dacă un zar este falsificat sau nu, căutând care este probabilitatea r *de apariţie a feţei şase . Pentru aceasta alegem* $\varepsilon = 0,01$*, aruncăm zarul de un număr* n *de ori şi observăm care este numărul* k *de apariţii ale feţei şase.*

Să se determine o valoare cât mai mică pentru n *astfel ca probabilitatea* $p\left(\left|\frac{k}{n} - r\right| \le \varepsilon\right)$ *să fie cel puţin* $0,99$ *utilizând:*

a) inegalitatea lui Cebâşev

b) formula integrală de Moivre-Laplace.

12. In problema acului lui Buffon (vezi lecţia 3, exerciţii), probabilitatea de intersectare a paraleleor este $p = \frac{2l}{\pi a}$ *unde* l *este lungimea acului iar* a *este distanţa dintre paralele. De câte aruncări este nevoie ca probabilitatea ca* $\left|\pi - \frac{2 \cdot l \cdot n}{a \cdot k}\right| < 0,01$ *să fie cel puţin* $0,99$*. (*n *este numărul de aruncări iar* k *este numărul de intersectări ale reţelei de drepte paralele).*

Lecţia 6

Dependenţa între variabilele aleatoare

6.1 Coeficientul de corelaţie

Definiţia 6.1 *Fie $f_1, f_2 : X \to R$ două v.a. cu mediile m_1, m_2 şi dispersiile $D_1 = \sigma_1^2 = M\left((f_1 - m_1)^2\right)$, $D_2 = \sigma_2^2 = M\left((f_2 - m_2)^2\right)$. Se numeşte coeficient de corelaţie al variabilelor aleatoare f_1, f_2 numărul:*

$$c(f_1, f_2) = \frac{M\left((f_1 - m_1) \cdot (f_2 - m_2)\right)}{\sqrt{D_1 \cdot D_2}} \tag{6.1}$$

Mărimea $cov\,(f_1, f_2) = M\left((f_1 - m_1) \cdot (f_2 - m_2)\right)$ se numeşte covarianţa lui f_1şi f_2.

Spunem că două funcţii f_1, şi f_2 coincid aproape peste tot (a.p.t.) dacă $p(f_1 \neq f_2) = 0$. Importanţa acestui coeficient se va vedea în continuare.

Teorema 6.2 *(Inegalitatea Cauchy-Buniakovski). Fie $f_1, f_2 : X \to R$ două v.a . cu momente de ordinul doi finite. Atunci:*

$$M^2\left(f_1 f_2\right) \leq M\left(f_1^2\right) \cdot M\left(f_2^2\right) \tag{6.2}$$

Egalitatea poate avea loc dacă şi numai dacă $f_2 = af_1$sau $f_1 = af_2$ a.p.t. cu a=const.

Demonstraţie. $M\left((f_1 + xf_2)^2\right) \geq 0$ pentru orice $x \in R$, deci $M\left(f_1^2\right) + 2M\left(f_1 f_2\right) x + M\left(f_2^2\right) x^2 \geq 0$ pentru orice x. De aici rezultă că discriminantul funcţiei de gradul doi este negativ: $\Delta = M^2\left(f_1 f_2\right) - M\left(f_1^2\right) M\left(f_2^2\right) \leq 0$ adică tocmai inegalitatea Cauchy-Buniakovski. Dacă $\Delta = 0$ atunci există un x_0 real, soluţie a ecuaţiei de gradul doi

$$M\left((f_1 + x_0 f_2)^2\right) = 0$$

de unde rezulă că $(f_1 + x_0 f_2)^2$ ia valoarea zero pe o mulţime cu probabilitatea 1, adică $f_1 = -x_0 f_2$ a.p.t.

In cazul în care inecuaţia nu ar fi de gradul 2, deci $M(f_2^2) = 0$, rezultă $f_2 = 0$ a.p.t., deci $f_1 f_2 = 0$ a.p.t., deci (6.2) are din nou loc deoarece ambii membri sunt 0. In acest caz $f_2 = 0 \cdot f_1$.

Teorema 6.3 *Fie $f_1, f_2 : X \to R$ două v.a. cu mediile m_1, m_2 şi dispersiile $D_1 = \sigma_1^2 = M\left((f_1 - m_1)^2\right)$, $D_2 = \sigma_2^2 = M\left((f_2 - m_2)^2\right)$ finite. Atunci:*

a) $c(f_1, f_2) \in [-1, 1]$.

b) $c(f_1, f_2) = \pm 1$ este echivalent cu $f_1 = af_2 + b$ sau $f_2 = af_1 + b$ a.p.t. pentru nişte constante a şi b potrivit alese.

c) f_1 şi f_2 independente implică $c(f_1, f_2) = 0$ (dar nu şi invers).

Demonstraţie. a) Se aplică inegalitatea Cauchy-Buniakovski pentru $f_1 - m_1$ şi $f_2 - m_2$.

b) Dacă $c(f_1, f_2) = \pm 1$ atunci în inegalitatea Cauchy-Buniakovski semnul $\leq$ devine $=$ şi teorema precedentă implică $(f_2 - m_2) = a(f_1 - m_1)$ deci $f_2 = af_1 + m_2 - a \cdot m_1$sau $(f_1 - m_1) = a \cdot (f_2 - m_2)$ de unde rezultă $f_1 = af_2 + b$ pentru a şi b potrivit alese.

c) f_1 şi f$_2$ independente implică $(f_1 - m_1)$ şi $(f_2 - m_2)$ independente deci:

$$M((f_1 - m_1)(f_2 - m_2)) = M(f_1 - m_1) M(f_2 - m_2) = 0 \cdot 0 = 0$$

Prin urmare $c(f_1, f_2) = 0$.

QED.

Exemplul 6.4 *Fie două şiruri finite de numere $(x_i)_{1 \leq i \leq n}$ şi $(y_i)_{1 \leq i \leq n}$. Să considerăm o partiţie a unei mulţimi X astfel: $X = A_1 \cup A_2 \cup A_3 \cup .. \cup A_n$ cu toate mulţimile $A_i \neq \emptyset$. Definim pe X o probabilitate prin $p(A_i) = \frac{1}{n}$ şi fie variabilele aleatoare $f_1, f_2 : X \to R$, definite prin $f_1(\omega) = x_i$ dacă $\omega \in A_i$ şi $f_2(\omega) = y_i$ dacă $\omega \in A_i$. Avem :*

$$\begin{aligned}
m_1 &= M(f_1) = \frac{\sum_{i=1}^n x_i}{n}; \quad m_2 = M(f_2) = \frac{\sum_{i=1}^n y_i}{n} \\
D_1 &= D(f_1) = \frac{\sum_{i=1}^n (x_i - m_1)^2}{n}; \quad D_2 = D(f_2) = \frac{\sum_{i=1}^n (y_i - m_2)^2}{n} \\
m_{1,2} &= M((f_1 - m_1)(f_2 - m_2)) = \frac{\sum_{i=1}^n (x_i - m_1)(y_i - m_2)}{n} \\
c(f_1, f_2) &= \frac{m_{1,2}}{\sqrt{D_1 D_2}} = \frac{\sum_{i=1}^n (x_i - m_1)(y_i - m_2)}{\sqrt{\sum_{i=1}^n (x_i - m_1)^2} \sqrt{\sum_{i=1}^n (y_i - m_2)^2}} \qquad (6.3)
\end{aligned}$$

Conform teoremei precedente, egalitatea cu ± 1 a coeficientului 6.3 este echivalentă cu $f_2 = af_1 + b$ sau cu alte cuvinte $y_i = ax_i + b$ pentru toate valorile $1 \leq i \leq n$. Dacă $|c(f_1, f_2)|$ este suficient de aproape de unu atunci punctele (x_i, y_i) sunt aproximativ pe o dreaptă. Dreapta ce aproximează "cel mai bine" acest nor de puncte se numeşte dreapta de regresie a lui y în x şi se determină prin metoda celor mai mici pătrate. Această dreaptă de forma $y = ax + b$ se determină din condiţia ca $\sum (y_i - ax_i - b)^2$ să fie minimă (a se vedea lecţia 8 sau lecţia 14). Dacă se caută dependenţe de forma $y_i = be^{ax_i}$, prin logaritmare se găseşte $\ln(y_i) = ax_i + \ln(b)$. O asemenea dependenţă există, conform celor de mai sus dacă $c(f_1, \ln(f_2)) = \pm 1$. Dacă $|c(f_1, \ln(f_2))|$ este suficient de aproape de unu atunci punctele $(x_i, \ln(y_i))$ sunt aproximativ pe o dreaptă care se determină prin metoda celor mai mici pătrate.Găsim astfel a şi $\ln(b)$ care ne dau dependenţa $\ln(y_i) = ax_i + \ln(b)$, adică $y_i = be^{ax_i}$.

6.2 Variabile aleatoare bidimensionale

Fie (X, Ω, p) un σ câmp de probabilitate şi fie f_1, $f_2 : X \to R$ două variabile aleatoare. Legătura între variabilele aleatoare este pusă în evidenţă prin funcţia lor comună de repartiţie. Să considerăm $f : X \to R^2$, definită prin $f(\omega) = (f_1(\omega), f_2(\omega)) \in R^2$. Funcţia f are calitatea că pentru orice produs de intervale $I_1 \times I_2 \subset R^2$ mulţimea $f^{-1}(I_1 \times I_2) = f_1^{-1}(I_1) \cap f_2^{-1}(I_2) \in \Omega$.

Definiţia 6.5 *Se numeşte v.a. bidimensională pe câmpul de probabilitate (X, Ω, p) o funcţie $f : X \to R^2$ cu proprietatea că pentru orice intervale I_1 şi I_2 din R, rezultă $f^{-1}(I_1 \times I_2) \in \Omega$.*

Vom mai spune uneori variabilă aleatoare dublă în loc de variabilă aleatoare bidimensională. Notând $f_1(\omega)$ şi $f_2(\omega)$ cele două componente ale lui f, vedem că o v.a. bidimensională este asociată unei perechi de v.a. unidimensionale.

Definiţia 6.6 *Se numeşte funcţie de repartiţie bidimensională a lui f, funcţia $F : R^2 \to R$ definită prin:*

$$\begin{aligned} F(x_1, x_2) &= p\left(f^{-1}((-\infty, x_1) \times (-\infty, x_2))\right) \\ &= p((f_1 < x_1) \cap (f_2 < x_2)) \end{aligned}$$

Teorema 6.7 *1)Funcţia de repartiţie bidimensională are proprietăţile:*
a) F este crescătoare în fiecare argument.
b) $F(x_1, -\infty) = 0, F(-\infty, x_2) = 0, \quad F(\infty, \infty) = 1.$
c) F este continuă la stânga în fiecare argument.
d) Pentru orice $a_1 < b_1$ şi $a_2 < b_2$ avem

$$\begin{aligned} & p((a_1 \le f_1 < b_1) \cap (a_2 \le f_2 < b_2)) \\ = \ & F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2) \ge 0 \end{aligned} \tag{6.4}$$

2) Reciproc, orice funcţie $F : R^2 \to R$ cu proprietăţile a)-d) de mai sus este funcţie de repartiţie pentru o v.a. bidimensională potrivit aleasă.

Demonstraţie.a) $x_1 < x_1'$ implică $\{f_1 < x_1 \cap f_2 < x_2\} \subset \{f_1 < x_1' \cap f_2 < x_2\}$. Luând probabilităţile ambilor membri găsim $F(x_1, x_2) \leq F(x_1', x_2)$. Analog se arată monotonia în x_2.

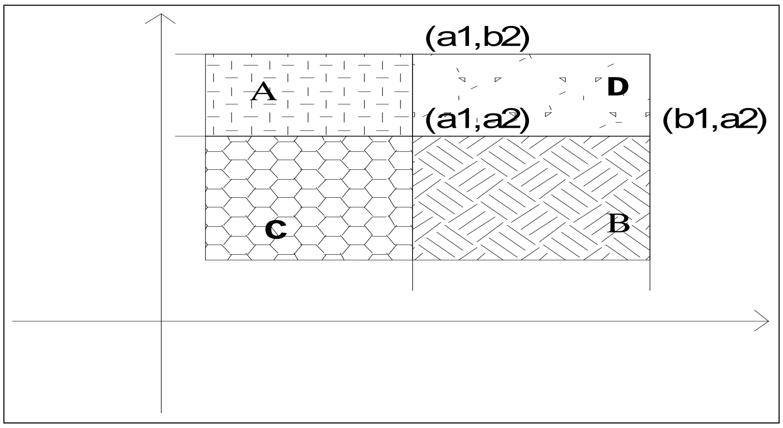
b) Prin definiţie $F(x_1, -\infty) = \lim\limits_{x_2 \to -\infty} F(x_1, x_2)$. Insă pentru orice şir monoton $x_{2,n} \to -\infty$ avem $\{f_1 < x_1 \cap f_2 < x_{2,n+1}\} \subset \{f_1 < x_1 \cap f_2 < x_{2,n}\}$ şi $\bigcap\limits_{n=1,\infty} \{f_1 < x_1 \cap f_2 < x_{2,n}\} = \emptyset$, deci $F(x_1, x_{2,n}) \to 0$, deci $F(x_1, -\infty) = 0$. Analog se arată şi celelalte egalităţi.

c) Continuitatea la stânga se arată ca pentru v.a. unidimensionale (lecţia 3).

d) Fie regiunile din x_1Ox_2 ca în figura următoare:

$A = \{(x_1, x_2) \mid -\infty < x_1 < a_1, a_2 \leq x_2 < b_2\}$, etc.(vezi figura). Fie evenimentele:

$A' = \{\omega \in X \mid (f_1(\omega), f_2(\omega)) \in A\}$ şi analog B', C', D'. Observăm că $D' = (a_1 \leq f_1 < b_1) \cap (a_2 \leq f_2 < b_2)$. Avem $p(A') = F(a_1, b_2) - F(a_1, a_2)$, $p(C') = F(a_1, a_2)$, $p(B') = F(b_1, a_2) - F(a_1, a_2)$.



Scriem acum

$$\begin{aligned} F(b_1, b_2) &= p(A' \cup B' \cup C' \cup D') = p(A') + p(B') + p(C') + p(D') \\ &= F(a_1, b_2) - F(a_1, a_2) + +F(b_1, a_2) - F(a_1, a_2) + F(a_1, a_2) + p(D') \end{aligned}$$

de unde rezultă pentru $p(D')$ formula 6.4.

QED.

Definiţia 6.8 *Se numeşte densitate de probabilitate a v.a. $f : X \to R^2$ o funcţie $\rho : R^2 \to R$ pozitivă, integrabilă, astfel ca $\int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \rho(t_1, t_2) dt_1 dt_2 = F(x_1, x_2)$ pentru $(x_1, x_2) \in R^2$ F fiind funcţia de repartiţie a lui f.*

Vom mai numi ρ densitatea mixtă a variabilelor f_1 şi f_2 unde $f = (f_1, f_2)$.

Teorema 6.9 *1) Densitatea de probabilitate ρ are propriatăţile:*

a) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \rho(t_1, t_2) dt_1 dt_2 = 1$.

b) $p((a_1 \leq f_1 < b_1) \cap (a_2 \leq f_2 < b_2)) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \rho(x_1, x_2) dx_1 dx_2$.

2) Invers, orice funcţie $\rho : R^2 \to R$ *pozitivă şi integrabilă cu proprietea a) de mai sus este densitate de probabilitate pentru o v.a. bidimensională potrivit aleasă.*

Demonstraţie. a), b) sunt evidente din definiţie. Punctul 2) se poate demonstra luând $X = R^2$, probabilitatea unui paralelipiped $[a, b) \times [c, d)$ se defineşte prin $\int_a^b \int_c^d \rho(x_1, x_2) dx_1 dx_2$, iar cele două variabile aleatoare sunt $f_1, f_2 : R^2 \to R$, $f_1(x_1, x_2) = x_1$, $f(x_1, x_2) = x_2$.

QED.

Cunoscând F=funcţia de repartiţie şi ρ =densitatea de probabilitate bidimensională a lui $f = (f_1, f_2)$ putem calcula relativ uşor diverse mărimi legate de f_1şi f_2.

1. Funcţia F_1de repartiţie a lui f_1 este

$$F_1(x_1) = p(f_1 < x_1) = F(x_1, \infty) = \int_{-\infty}^{x_1} dt_1 \int_{-\infty}^{\infty} \rho(t_1, x_2) dx_2 \tag{6.5}$$

iar densitatea este

$$\rho_1(x_1) = \int_{-\infty}^{\infty} \rho(x_1, x_2) dx_2. \tag{6.6}$$

2. Media lui f_1 este

$$m_1 = M(f_1) = \int_{-\infty}^{\infty} x_1 \rho_1(x_1) dx_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 \rho(x_1, x_2) dx_1 dx_2 \tag{6.7}$$

3. Dispersia este

$$D(f_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 - m_1)^2 \rho(x_1, x_2) dx_1 dx_2. \tag{6.8}$$

4. Formule analoage sunt valabile şi pentru f_2. De exemplu

$$m_2 = M(f_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{-\infty} x_2 \rho(x_1, x_2) dx_1 dx_2$$

5. Dacă $g : R^2 \to R$ este o funcţie continuă, atunci $g(f_1, f_2) : X \to R$ dată de $g(f_1, f_2)(\omega) = g(f_1(\omega), f_2(\omega))$ este o v.a. Media acestei v.a. se calculează cu densitatea mixtă prin

$$M(g) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x_1, x_2) \rho(x_1, x_2) dx_1 dx_2 \tag{6.9}$$

6. In particular luînd $g = (f_1 - m_1)(f_2 - m_2)$ apoi $g = (f_1 - m_1)^2$ şi $g = (f_2 - m_2)^2$ găsim pentru coeficientul de corelaţie:

$$\begin{aligned} c(f_1, f_2) &= \frac{M((f_1 - m_1)(f_2 - m_2))}{\sqrt{M((f_1 - m_1)^2)}\sqrt{M((f_2 - m_2)^2)}} \\ &= \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_1 - m_1)(x_2 - m_2) \rho(x_1, x_2) dx_1 dx_2}{\sqrt{D(f_1)}\sqrt{D(f_2)}} \end{aligned} \tag{6.10}$$

7. Dacă f_1 şi f_2 sunt independente atunci

$$F(x_1, x_2) = p(f < x_1 \cap f_2 < x_2) = p(f_1 < x_1) \cdot p(f_2 < x_2) \ = F_1(x_1) \cdot F_2(x_2) \tag{6.11}$$

De aici rezultă

$$\rho(x_1, x_2) = \rho_1(x_1) \cdot \rho_2(x_2). \tag{6.12}$$

Se vede uşor că şi reciproc este adevărat: dacă funcţia de repartiţie ori densitatea de probabilitate se descompune în produs de două funcţii, una depinzând de x_1 şi cealaltă depinzând de x_2 atunci f_1 şi f_2 sunt independente (exerciţiu).

Dintre formulele de mai sus doar (6.9) mai necesită demonstraţie, celelalte fiind evidente.

a) Considerăm cazul când f_1şi f_2 sunt mărginite, deci $f = (f_1, f_2) : X \to [a, b) \times [c, d)$. Divizăm intervalele $[a, b]$ şi $[c, d]$ prin $\Delta_x = (x_i)_{0 \leq i \leq n}$ respectiv $\Delta_y = (y_j)_{0 \leq j \leq m}$şi notăm $\Delta = \Delta_x \times \Delta_y$, $|\Delta| = \max\{|\Delta_x|, |\Delta_y|\}$, $D_{ij} = [x_{i-1}, x_i) \times [y_{j-1}, y_j)$, $A_{ij} = f^{-1}(D_{ij}) \subset X$. Variabila aleatoare

$$g_\Delta(\omega) = \begin{cases} g(x_i, y_j) & \text{dacă } \omega \in A_{ij} \\ 0 & \text{în caz contrar} \end{cases}$$

este simplă, şi tinde la $g(f_1, f_2)$ atunci când $|\Delta_x| \to 0$ şi $|\Delta_y| \to 0$. Deci

$$\begin{aligned} M(g_\Delta) &= \sum g(x_i, y_j)\, p(A_{ij}) \\ &= \sum g(x_i, y_j) \underbrace{\int\int_{D_{ij}} \rho(x, y)\, dxdy}_{p(A_{ij})} \\ &= \sum g(x_i, y_j) \underbrace{\rho\left(\xi_i, \eta_j\right)(x_i - x_{i-1})(y_j - y_{j-1})}_{\textit{din teorema de medie}} \\ &\underset{|\Delta| \to 0}{\longrightarrow} \int\int_{[a,b]\times[c,d]} g(x, y)\, \rho(x, y)\, dxdy \end{aligned}$$

b) Demonstraţia se poate extinde acum şi la cazul când f_1şi f_2 nu sunt mărginite dar integrala (6.9) este convergentă. Argumentele fiind lungi, însă standard, nu le reproducem aici.

Repartiţiile pentru f_1 şi f_2 se mai numesc repartiţii marginale ale lui f..

Exemplul 6.10 *Fie $D = [a, b] \times [c, d]$. O variabilă aleatoare $f = (f_1, f_2)$ cu valori în D care are densitatea:*

$$\rho(x_1, x_2) = \begin{cases} 0 & (x_1, x_2) \notin D \\ \frac{1}{aria(D)} & (x_1, x_2) \in D \end{cases}$$

se numeşte uniformă. Conform celor de la pct. 6) rezultă că f_1 ş f_2 sunt independente.

Exemplul 6.11 *Fie o v.a.* (f_1, f_2)*bidimensională cu densitatea:*

$$\rho(x_1, x_2) = \frac{\sqrt{\det(A)}}{2\pi} e^{-\frac{1}{2}Q(x_1,x_2)}$$

unde $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ *este o matrice pozitiv definită iar*

$$Q(x_1, x_2) = \sum_{i,j=1}^{2} a_{i,j}(x_i - m_i)(x_j - m_j)$$

O astfel de variabilă aleatoare se numeşte normală. Se cere:
a) $\sigma_1^2 = D(f_1)$ *şi* $\sigma_2^2 = D(f_2)$*; b)* $c(f_1, f_2)$*.*

Soluţie. a) Avem nevoie de $M(f_1) = \int\int x_1\rho(x_1, x_2)\,dx_1dx_2$. Se ştie că matricea A este simetrică. Pentru aceste integrale este utilă următoarea schimbare de coordonate:
i) Se determină λ_1 şi λ_2 valorile proprii ale matricei A, din ecuaţia:

$$\begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = 0$$

ii) Se determină vectorii proprii ai matricei A şi se normalizează. Aceşti vectori puşi pe două coloane dau o matrice ortogonală $R = \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix}$, $R^t = R^{-1}$.

iii) Se ştie că $R^tAR = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$. Facem schimbarea $\begin{pmatrix} x_1 - m_1 \\ x_2 - m_2 \end{pmatrix} = R \cdot \begin{pmatrix} x_1' \\ x_2' \end{pmatrix}$. Prin această schimbare Q devine $Q(x_1, x_2) = \lambda_1 x_1'^2 + \lambda_2 x_2'^2$ iar iacobianul $\frac{D(x_1,x_2)}{D\left(x_1',x_2'\right)} = \det(R) = 1$. Avem de asemenea $\det(A) = \lambda_1\lambda_2$. Toate aceste lucruri se studiază la cursul de algebră liniară. Utilizând $\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}}dx = 1$, $\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2\sigma^2}}dx = \sigma^2$, $\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} xe^{-\frac{x^2}{2\sigma^2}}dx = 0$, (a se vedea lecţia 4, legea normală) găsim:
iv)

$$\begin{aligned} M(f_1) &= \frac{\sqrt{\lambda_1\lambda_2}}{2\pi}\int_{-\infty}^{\infty} dx_1' \int_{-\infty}^{\infty} (m_1 + r_{11}x_1' + r_{12}x_2')\, e^{-\frac{\lambda_1}{2}x_1'^2 - \frac{\lambda_2}{2}x_2'^2} dx_2' \\ &= \frac{\sqrt{\lambda_1}}{\sqrt{2\pi}} \cdot \frac{\sqrt{\lambda_2}}{\sqrt{2\pi}} \cdot m_1 \cdot \int_{-\infty}^{\infty} e^{-\frac{\lambda_1}{2}x_1'^2} dx_1' \cdot \int_{-\infty}^{\infty} e^{-\frac{\lambda_2}{2}x_2'^2} dx_2' = m_1 \end{aligned}$$

Analog găsim $M(f_2) = m_2$.
Putem acum calcula dispersiile.

a)

$$\begin{aligned}
\sigma_1^2 &= \int\int (x_1 - m_1)^2 \rho(x_1, x_2)\, dx_1 dx_2 \\
&= \frac{\sqrt{\lambda_1\lambda_2}}{2\pi} \int\int (r_{11}x_1' + r_{12}x_2')^2 e^{\frac{-\lambda_1 x_1'^2 - \lambda_2 x_2'^2}{2}} dx_1' dx_2' \\
&= r_{11}^2 \cdot \frac{1}{\lambda_1} + r_{12}^2 \cdot \frac{1}{\lambda_2}
\end{aligned}$$

In mod analog găsim $\sigma_2^2 = r_{21}^2 \cdot \frac{1}{\lambda_1} + r_{22}^2 \cdot \frac{1}{\lambda_2}$
b) Covarianţa lui f_1şi f_2 este:

$$\begin{aligned}
cov(f_1, f_2) &= \frac{\sqrt{\lambda_1\lambda_2}}{2\pi} \int\int (r_{11}x_1' + r_{12}x_2')(r_{21}x_1' + r_{22}x_2') e^{\frac{-\lambda_1 x_1'^2 - \lambda_2 x_2'^2}{2}} dx_1' dx_2' \\
&= \frac{\sqrt{\lambda_1\lambda_2}}{2\pi} \int\int \left(r_{11}r_{21}x_1'^2 + r_{12}r_{22}x_2'^2\right) e^{\frac{-\lambda_1 x_1'^2 - \lambda_2 x_2'^2}{2}} dx_1' dx_2' \\
&= r_{11}r_{21} \cdot \frac{1}{\lambda_1} + r_{12}r_{22} \cdot \frac{1}{\lambda_2}
\end{aligned}$$

. De aici rezultă $c(f_1, f_2)$ şi punctul b) e terminat. Mai remarcăm că deoarece $R^t A R = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$, rezultă

$$A^{-1} = R \begin{pmatrix} \frac{1}{\lambda_1} & 0 \\ 0 & \frac{1}{\lambda_2} \end{pmatrix} R^t = \begin{pmatrix} \sigma_1^2 & cov(f_1, f_2) \\ cov(f_1, f_2) & \sigma_2^2 \end{pmatrix}$$

6.3 Funcţia de repartiţie condiţionată

In această secţiune (X, Ω, p) este un câmp de probabilitate, $f : X \to R$ o v.a. cu funcţia de repartiţie F,iar α, β sunt numere reale pozitive. Dacă există următoarea limită

$$\begin{aligned}
&\lim_{\alpha,\beta \to 0} \frac{p(B \cap (x - \alpha \le f < x + \beta))}{p(x - \alpha \le f < x + \beta)} \\
&= \lim_{\alpha,\beta \to 0} \frac{p(B \cap (x - \alpha \le f < x + \beta))}{F(x + \beta) - F(x - \alpha)}
\end{aligned}$$

atunci o notăm $p(B|f = x)$. Putem da acum definiţia:

Definiţia 6.12 *Dacă $x \in R$ şi $B \in \Omega$, numim probabilitatea lui B condiţionată de $f = x$ mărimea $p(B \mid f = x)$.*

E posibil ca $p(f = x) = 0$ şi deci să nu putem defini $p(B|f = x) = \frac{p(B\cap(f=x))}{p(f=x)}$ aşa ca în lecţia 1. In asemenea situaţie aplicăm trecerea la limită de mai sus. Ca funcţie de B, $p(\cdot\,|f = x)$ este o probabilitate finit aditivă pe o subalgebră a lui Ω. Dacă nu există pericol de confuzie vom scrie $p(B\,|\,x)$ în loc de $p(B\,|\,f = x)$.

Fie $g : X \to R$ este o v.a. iar $h : X \to R^2$ $\quad h = (f, g)$ o v.a. dublă care are funcţia de repartiţie H. Vom nota cu ρ_f densităţile care se referă la f, cu ρ_g densităţile care se referă la g, cu ρ densitatea variabilei bidimensionale (f, g). Notăm

$$\begin{aligned} G(t\,|\,f = x) &= p(g < t|f = x) \\ &= \lim_{\alpha,\beta\to 0} \frac{p((g < t) \cap (x - \alpha \le f < x + \beta))}{p((x - \alpha \le f < x + \beta))} \\ &= \lim_{\alpha,\beta\to 0} \frac{H(x + \beta, t) - H(x - \alpha, t)}{H(x + \beta, \infty) - H(x - \alpha, \infty)} \end{aligned} \tag{6.13}$$

dacă limita există. Dacă limita există în orice $t \in R$ şi are ca funcţie de t proprietăţile unei funcţii de repartiţie, atunci definim:

Definiţia 6.13 *Se numeşte funcţia de repartiţie a lui g condiţionată de $f = x$ funcţia $G(\cdot\,|f = x) : R \to R$ definită prin* (6.13).

Se vede că dacă H este derivabilă, cu densitatea ρ avem

$$G(t|f = x) = \frac{\frac{\partial H(x,t)}{\partial x}}{\frac{\partial H(x,\infty)}{\partial x}} = \frac{\int_{-\infty}^{t} \rho(x, y)\, dy}{\int_{-\infty}^{\infty} \rho(x, y)\, dy} \tag{6.14}$$

Funcţiile $G(\cdot)$ şi $G(\cdot\,|f = x)$ sunt diferite. Se utilizează frecvent deosebirea funcţiilor prin "semnătura" lor, adică prin tipul de argumente, chiar dacă pentru nume se utilizează acelaşi simbol. Dacă nu există pericol de confuzie vom scrie şi $G(t|x)$ în loc de $G(t|f = x)$.

Definiţia 6.14 *Se numeşte densitate de probabilitate a lui g condiţionată de $f = x$ o funcţie reală $\rho_g(\cdot\,|f = x)$, integrabilă, pozitivă, astfel ca $\int_{-\infty}^{t} \rho_g(y|f = x)\, dy = G(t|f = x)$.*

O asemenea densitate nu există întotdeauna. Dacă există atunci în punctele ei de continuitate avem:

$$\frac{\partial G(t|f = x)}{\partial t} = \rho_g(t|f = x) = \frac{\rho(x, t)}{\int_{-\infty}^{\infty} \rho(x, y)\, dy} \quad \text{(cf. 6.14)}$$

Şi aici vom scrie $\rho_g(t|x)$ sau $\rho(t|x)$în loc de $\rho_g(t|f = x)$ dacă nu există pericol de confuzie.

In studiul probabilităţilor condiţionate din lecţia 1, cele mai importante formule studiate au fost formula probabilităţii totale şi formula lui Bayes. Cum arată aceste formule în contextul prezent?

1. **Formula probabilităţii totale:**

Dacă $p(B|x)$ este o funcţie continuă atunci

$$p(B) = \int_{-\infty}^{\infty} p(B|f = x)\, dF(x) = \int_{-\infty}^{\infty} p(B|f = x)\, \rho_f(x)\, dx$$

în particular

$$G(t) = \int_{-\infty}^{\infty} G(t|f = x)\, dF(x) = \int_{-\infty}^{\infty} G(t|f = x)\, \rho_f(x)\, dx$$

1. **Formula lui Bayes**

$$\rho_f(x|g = y) = \frac{\rho_f(x)\, \rho_g(y|f = x)}{\int_{-\infty}^{\infty} \rho_g(y|f = x)\, \rho_f(x)\, dx}$$

Dacă nu există pericol de confuzie putem scrie:

$$\rho(x|y) = \frac{\rho(x)\, \rho(y|x)}{\int_{-\infty}^{\infty} \rho(y|x)\, \rho(x)\, dx}$$

Nu insistăm asupra demonstraţiilor acestor formule. Ele se pot obţine prin divizarea intervalului unde se găsesc valorile lui x, aplicarea formulei clasice a probabilităţii totale sau a lui Bayes aşa cum au fost studiate în lecţia 1 şi trecerea la limită pentru diviziuni din ce în ce mai fine.

6.4 Distribuţia sumei şi câtului

Fie $f, g : X \to R$ două v.a. şi $h : X \to R^2$, $h = (f, g)$ cu repartiţia H şi densitatea ρ. Care este repartiţia sumei f+g şi a câtului $\frac{f}{g}$ în eventualitatea că $g \neq 0$? In general dacă r este o funcţie continuă $r : R^2 \to R$ atunci repartiţia lui $r(f, g)$ este:

$$F_{r(f,g)}(t) = p(r(f, g) < t) = \int\int_{r(x,y)<t} \rho(x, y)\, dxdy \tag{6.15}$$

Pentru sumă şi produs avem formule mai precise. Alte exemple sunt date la exerciţii.

6.4.1 Distribuţia sumei

In condiţiile de mai sus suma $f + g$ are distribuţia

$$F_{f+g}(t) = p(f + g < t) = \int\int_{x+y<t} \rho(x, y)\, dxdy = \int_{-\infty}^{t} du \int_{-\infty}^{\infty} \rho(v, u - v)\, dv$$

Am utilizat schimbarea $u = x + y$, $v = x$. Dacă f, g sunt independente atunci $\rho(x, y) = \rho_f(x)\rho_g(y)$ şi densitatea sumei devine

$$\rho_{f+g}(t) = F'_{f+g}(t) = \int_{-\infty}^{\infty} \rho_f(v)\rho_g(t-v)\,dv \tag{6.16}$$

ceea ce se numeşte produsul de convoluţie al densităţilor ρ_f şi ρ_g.

6.4.2 Distribuţia câtului

Fie f şi g ca mai înainte şi în plus g$\neq$ 0. In aceste condiţii câtul $\frac{f}{g}$ are distribuţia:

$$\begin{aligned} F_{\frac{f}{g}}(t) &= p\left(\frac{f}{g} < t\right) = \iint\limits_{y>0, x<ty} \rho(x,y)\,dxdy + \iint\limits_{y<0, x>ty} \rho(x,y)\,dxdy \\ &= \int_0^{\infty} dy \int_{-\infty}^{ty} \rho(x,y)\,dx + \int_{-\infty}^{0} dy \int_{ty}^{\infty} \rho(x,y)\,dx \end{aligned}$$

Prin derivare în raport cu t găsim

$$\rho_{\frac{f}{g}}(t) = \int_{-\infty}^{\infty} |y|\,\rho(ty, y)\,dy \tag{6.17}$$

In particular pentru f şi g independente găsim:

$$\rho_{\frac{f}{g}}(t) = \int_{-\infty}^{\infty} |y| \cdot \rho_f(ty) \cdot \rho_g(y) \cdot dy \tag{6.18}$$

6.5 Distribuţia Student

Exemplul 6.15 *Fie f normală de tipul $N(0, \sigma)$ iar g o variabilă χ^2 de tipul $H(s, \sigma)$, independentă de f. Se cer:*

a) Densitatea lui $\sqrt{\frac{g}{s}}$.

b) Densitatea lui $h = \frac{f}{\sqrt{\frac{g}{s}}}$.

c) Momentele lui h.

Soluţie. Densitatea lui f este

$$\rho_f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$$

iar a lui g este

$$\rho_g(y) = \frac{1}{2^{\frac{s}{2}}\sigma^s\Gamma\left(\frac{s}{2}\right)} y^{\frac{s}{2}-1} e^{-\frac{y}{2\sigma^2}} \text{pentru y>0 şi 0 în caz contrar.}$$

a) $F_{\sqrt{\frac{g}{s}}}(y) = p\left(\sqrt{\frac{g}{s}} < y\right) = p\left(g < y^2 s\right) = F_g\left(y^2 s\right)$, pentru $y > 0$ şi 0 pentru $y \leq 0$. Prin urmare densitatea lui $\sqrt{\frac{g}{s}}$ este

$$\begin{aligned}\rho_{\sqrt{\frac{g}{s}}}(y) &= \frac{d}{dy}F_{\sqrt{\frac{g}{s}}}(y) = 2syF_g'\left(y^2 s\right) \\ &= 2sy\rho_g\left(y^2 s\right) = \frac{2sy}{2^{\frac{s}{2}}\sigma^s\Gamma\left(\frac{s}{2}\right)}\left(y^2 s\right)^{\frac{s}{2}-1} e^{-\frac{y^2 s}{2\sigma^2}}\end{aligned}$$

pentru y>0 ş 0 pentru y≤ 0.

b) Conform cu (6.18) avem pentru $h = \frac{f}{\sqrt{g/s}}$ densitatea

$$\begin{aligned}\rho_h(t) &= \int_{-\infty}^{\infty} |y| \cdot \rho_f(ty) \cdot \rho_g(y) \cdot dy \\ &= \int_0^{\infty} y\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{t^2y^2}{2\sigma^2}} \cdot \frac{2sy}{2^{\frac{s}{2}}\sigma^s\Gamma\left(\frac{s}{2}\right)}y^{s-2}s^{\frac{s}{2}-1}e^{-\frac{sy^2}{2\sigma^2}}dy \\ &= \frac{2s^{\frac{s}{2}}}{2^{\frac{s+1}{2}}\sqrt{\pi}\sigma^{s+1}\Gamma\left(\frac{s}{2}\right)}\int_0^{\infty} e^{-\frac{\left(t^2+s\right)y^2}{2\sigma^2}}y^s dy \quad \text{(schimbăm } sy^2 = u\text{)} \\ &= \frac{1}{2^{\frac{s+1}{2}}\sqrt{\pi s}\sigma^{s+1}\Gamma\left(\frac{s}{2}\right)}\int_0^{\infty} e^{-\frac{\left(1+\frac{t^2}{s}\right)}{2\sigma^2}u}u^{\frac{s+1}{2}-1}du \\ &= \frac{1}{2^{\frac{s+1}{2}}\sqrt{\pi s}\sigma^{s+1}\Gamma\left(\frac{s}{2}\right)}\frac{\Gamma\left(\frac{s+1}{2}\right)}{\left(\frac{1+\frac{t^2}{s}}{2\sigma^2}\right)^{\frac{s+1}{2}}} = \frac{\Gamma\left(\frac{s+1}{2}\right)}{\sqrt{\pi s}\Gamma\left(\frac{s}{2}\right)}\left(1+\frac{t^2}{s}\right)^{-\frac{s+1}{2}}\end{aligned}$$

c) Media lui h, $M(h)$, este zero, h având densitatea pară. Pentru momentele lui h folosim definiţia:

$$M_k(h) = \int_{-\infty}^{\infty} x^k\rho_h(x)\,dx = \int_{-\infty}^{\infty}(x - M(h))^k\rho_h(x)\,dx = \mu_k(h)$$

Rezultă momente de ordin impar zero. Inlocuind pe $\rho_h(x)$ cu expresia de mai înainte ş făcând schimbarea $\frac{x^2}{s} = u$ şi apoi $\frac{u}{1+u} = y$ ajungem la

$$\begin{aligned}\mu_{2k} &= \frac{\Gamma\left(\frac{s+1}{2}\right)s^{k+\frac{1}{2}}}{\sqrt{\pi s}\Gamma\left(\frac{s}{2}\right)2}\int_0^1 y^{\left(k+\frac{1}{2}\right)-1}(1-y)^{\left(\frac{s}{2}-k\right)-1}dy \\ &= \frac{\Gamma\left(\frac{s+1}{2}\right)s^{k+\frac{1}{2}}}{\sqrt{\pi s}\Gamma\left(\frac{s}{2}\right)2}\frac{\Gamma\left(k+\frac{1}{2}\right)\Gamma\left(\frac{s}{2}-k\right)}{\Gamma\left(\frac{s+1}{2}\right)} \\ &= \frac{s^k}{\sqrt{\pi}}\frac{\Gamma\left(k+\frac{1}{2}\right)\Gamma\left(\frac{s}{2}-k\right)}{\Gamma\left(\frac{s}{2}\right)} \\ &= \frac{s^k 1\cdot 3\cdot ..\cdot(2k-1)}{(s-2)\cdot(s-4)\cdot ..\cdot(s-2k)}\end{aligned}$$

dacă $2k < s$. Pentru celelalte valori ale lui k momentele nu există.

Definiţia 6.16 *O variabilă aleatoare cu densitatea*

$$\rho(t) = \frac{\Gamma\left(\frac{s+1}{2}\right)}{\sqrt{\pi s}\,\Gamma\left(\frac{s}{2}\right)}\left(1 + \frac{t^2}{s}\right)^{-\frac{s+1}{2}}$$

*se numeşte variabilă **Student.** Tipul acesta de v.a. îl notăm $S(s)$. Graficul acestei densităţi se găseşte la lecţia 4.*

O asemennea variabilă aleatoare este raportul dintre o variabilă normală cu media zero şi dispersia σ^2 pe de o parte şi $\sqrt{\frac{g}{s}}$ unde g este o variabilă de tipul $H(s,\sigma)$ adică χ^2 cu s grade de libertate obţinută ca sumă a s pătrate de variabile normale de tip $N(0,\sigma)$ (vezi lecţia 4), pe de altă parte.

6.6 Distribuţia Snedecor-Fisher

Exemplul 6.17 *Fie f_1 şi f_2 două variabile aleatoare independente de tip $H(n_1,\sigma)$ şi $H(n_2,\sigma)$. Se cere densitatea variabilei $g = \frac{f_1/n_1}{f_2/n_2}$.*

Soluţie. f_1 şi f_2 sunt variabile χ^2. Deoarece ele sunt independente densitatea variabilei $h = (f_1, f_2)$ este

$$\rho(x,y) = \rho_{f_1}(x)\,\rho_{f_2}(y) = \begin{cases} \frac{1}{2^{\frac{n_1+n_2}{2}}\sigma^{n_1+n_2}\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}x^{\frac{n_1}{2}-1}y^{\frac{n_2}{2}-1}e^{-\frac{x+y}{2}}, & x,y>0 \\ 0, & \text{în rest} \end{cases}$$

dacă $x \geq 0$, $y \geq 0$ şi 0 în rest. Aplicând aceeaşi tehnică precum în exemplul precedent rezultă :

$$\rho_g(x) = \begin{cases} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}\frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}x^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}, & x>0 \\ 0, & \text{in rest} \end{cases} \tag{6.19}$$

dacă $x \geq 0$ şi 0 în caz contrar.

Definiţia 6.18 *O variabilă aleatoare cu densitatea (6.19) se numeşte variabilă **Snedecor-Fisher.** Acest tip îl vom nota $S(n_1,n_2)$.*

6.7 Exerciţii

1. Fie şirul de date

$$\begin{array}{rrrrrrr} x = & 1 & 3 & -1 & 4 & 7 & 9 \\ y = & 0 & 4 & -3 & 8 & 12 & 18 \end{array}$$

a)Care este coeficientul de corelaţie între aceste şiruri de date?
b) Dacă este mai mare de 0.95, să se determine prin metoda celor mai mici pătrate dreapta de regresie a lui y în x.

Indicaţie. Coeficientul de corelaţie (6.3) este $0,99$ deci punctele sunt aproximativ pe o dreaptă.

2. Fie şirul de date

$$\begin{array}{rcccccc} x & = & 1 & 1,5 & 2 & 2,5 & 3 \\ y & = & 3,3 & 4,23 & 5,44 & 6,98 & 8,96 \end{array}$$

Să se studieze existenţa unei dependenţe de forma $y = be^{ax}$ între date.

Indicaţie. Valorile lui $\ln(y)$ sunt

$$1,19 \quad 1,44 \quad 1,69 \quad 1,94 \quad 2,19$$

Coeficientul de corelaţie între x ş $\ln(y)$ este aproximativ 1, prin urmare $\ln(y) = ax + d$ deci $y = e^d e^{ax} = be^{ax}$. Prin metoda celor mai mici pătrate găsim $a = 0,5 \quad d = 0,69$ deci $b = 2$.

3. Variabilele aleatoare independente f_1 şi f_2 sunt uniform distribuite în $[0,1]$. Se cer:
a) Distribuţiile pentru $\min(f_1, f_2)$ şi $\max(f_1, f_2)$.
b) Distribuţia sumei $f_1 + f_2$.

Indicaţie. Variabila dublă (f_1, f_2) are densitatea $\rho(x,y) = 1$ în $(0,1) \times (0,1)$ şi 0 în rest. Fie F funcţia de repartiţie a lui $\min(f_1, f_2)$. Avem

$$\begin{aligned} F(x) &= p(f_1 < x \cup f_2 < x) = 1 - p\left(\overline{f_1 < x \cup f_2 < x}\right) \\ &= 1 - p(f_1 \geq x \cap f_2 \geq x) = 1 - p(f_1 \geq x) \cdot p(f_2 \geq x) \\ &= 1 - (1-x)(1-x) \end{aligned}$$

pentru $x \in [0,1]$. In mod analog se calculează repartiţia lui $\mathrm{mmax}(f_1, f_2)$. Pentru sumă se foloseşte formula (6.16)

4. Fie X şi Y două v.a. cu coeficientul de corelaţie r. Care este coeficientul de corelaţie al variabilelor $aX + b$ şi $cY + d$, unde $a, b, c, d \in R$?

5. Numărul de maşini pe şoseaua Bucureşti-Ploieşti, dinspre Bucureşti spre Ploieşti, într-un interval de timp I este o variabilă Poisson cu parametrul λ_1, iar numărul de maşini în *sens contrar, în acelaşi interval de timp, este o variabilă Poisson cu parametrul* λ_2*. Dacă cele două v.a. sunt independente, care este distribuţia numărului total de maşini pe şosea în acelaşi interval de timp?*

6. Un cuplu de v.a. $f = (f_1, f_2)$ *are densitatea dublă*

$$\rho(x,y) = \begin{cases} \frac{4(x+3y)e^{-x-3y}}{5}, & \text{dacă } x \geq 0 \quad y \geq 0 \\ 0 & \text{în caz contrar} \end{cases}$$

i) Să se determine densităţile marginale pentru f_1 *şi* f_2 *.*
ii) Să se calculeze densitatea condiţională a lui X dacă Y=y.
ii) Să se calculeze coeficientul de corelaţie dintre X şi Y.

7. Se trage asupra unei ţinte plane aşezate în $(0,0)$*. Măsurătorile indică o distribuţie normală a absciselor* X *a punctelor de impact de tip* $N(0,2)$ *şi la fel pentru ordonatele* Y*. De asemenea* X *şi* Y *sunt independente. Se cer:*
a) Distribuţia distanţei $\sqrt{X^2+Y^2}$ *de la punctul de impact până la ţintă.*
b) Probabilitatea ca o lovitură specificată să cadă în discul $X^2+Y^2 \leq 4$*.*
c) Probabilitatea ca din 4 lovituri cel puţin una să cadă în discul $X^2+Y^2 \leq 4$*.*
d) Probabilitatea ca din 100 lovituri cel puţin 25 să cadă în discul $X^2+Y^2 \leq 4$*.*

Indicaţie. a) $\rho_1(x) = \frac{1}{\sqrt{2\pi}2}e^{-\frac{x^2}{8}}$ $\rho_2(x) = \frac{1}{\sqrt{2\pi}2}e^{-\frac{y^2}{8}}$ deci $\rho(x,y) = \frac{1}{8\pi}e^{-\frac{x^2+y^2}{8}}$. Fie $F(t)$ funcţia de repartiţie a lui $\sqrt{X^2+Y^2}$. Avem pentru $F(t) = 0$ pentru $t < 0$ iar pentru $t \geq 0$:

$$F(t) = p\left(X^2+Y^2 < t^2\right) = \frac{1}{8\pi}\int\int_{x^2+y^2<t^2} e^{-\frac{x^2+y^2}{8}}dxdy$$

şi se trece la coordonate polare. b) Probabilitatea este $p = F(2)$. c) Probabilitatea căutată este $1-(1-p)^4$. d) Răspunsul este $\sum_{k=25}^{100} C_{100}^k p^k (1-p)^{n-k}$ şi se foloseşte formula integrală de Moivre-Laplace.

8. Perechea de v.a. (f_1, f_2) *admite densitatea dublă* $\rho(x,y) = \frac{1}{2\pi}\frac{1}{(1+x^2+y^2)^{3/2}}$*. Se cere:*
a) Să se arate că f_1 *şi* f_2 *sunt v.a. de tip Cauchy, adică au densitatea* $\rho(x) = \frac{1}{\pi}\frac{1}{1+x^2}$*.*
b) Să se calculeze $p(f_1^2+f_2^2 \leq 4)$*.*

Indicaţie. a) $\rho_1(x) = \int_{-\infty}^{\infty}\rho(x,y)\,dy = ...$ b) Se procedează ca la punctul a) de la problema precedentă.

9. Durata de viaţă a unui bec ce funcţionează continuu este o variabilă aleatoare cu densitatea

$$\rho(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

Intr-un anumit loc trebuie să funcţioneze continuu un bec şi se dispune de un număr de 10 becuri pentru înlocuire în caz că cel în funcţiune se defectează. Se cer:

a) Valoarea lui λ dacă durata medie de viaţă a unui bec este de 20 zile (ziua va fi în problemă unitatea de măsură pentru timp).
b) Să se arate că dacă f_1 şi f_2 sunt independente, cu densitatea ρ atunci f_1+f_2 are densitatea

$$\rho_1(t) = \begin{cases} \lambda^2 t e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

c) Să se arate că dacă $f_1, f_2, ...f_n$ sunt independente şi au densitatea ρ de mai sus, atunci $f_1 + f_2 + ... + f_n$ are densitatea

$$\rho_n(t) = \begin{cases} \frac{\lambda^n t^{n-1}}{(n-1)!} e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

d) Care e probabilitatea ca prin înlocuirea becurilor arse să poată fi menţinută lumina aprinsă peste 200 zile?

Indicaţie. a) $M(f_1) = \frac{1}{\lambda}$ deci $\lambda = \frac{1}{20}$. b,c) Se aplică (5.13) sau se procedează ca în lecţia 4, exerciţiul nr. 4. d) In datele de la c) avem $\lambda = \frac{1}{20}$, $n = 10$, iar probabilitatea cerută este:

$$\begin{aligned} P &= p(f_1 + f_2 + ... + f_{10} \geq 200) = \int_{200}^{\infty} \frac{1}{20^{10} \times 9!} t^9 e^{-\frac{t}{20}} dt \\ &= \frac{10^{10}}{9!} \int_1^{\infty} u^9 e^{-10u} du \simeq 0,457 \end{aligned}$$

10) Două variabile aleatoare independente f_1 şi f_2 au distribuţii normale $N(0, \sigma)$. Să se arate că $f_1^2 + f_2^2$ şi $\frac{f_1}{f_2}$ sunt independente.

Indicaţie. (f_1, f_2) are densitatea $\frac{1}{4\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$. $\frac{f_1}{f_2}$ este definită a.p.t. deoarece $p(f_2 = 0) = 0$. Fie $\phi : R^2 - \{(x, 0) | x \in R\} \to R^2$ definită prin $(x, y) \xrightarrow{\phi} \left(r = x^2 + y^2,\ s = \frac{x}{y}\right)$. Avem $\frac{D(r,s)}{D(x,y)} = -2(s^2 + 1)$. De asemenea $x^2 = \frac{rs^2}{1+s^2}$ $y^2 = \frac{r}{1+s^2}$. Vedem că pentru orice pereche (r, s) există două perechi (x, y) ce îi corespund prin transformarea dată. Pentru $D \subset R^2$ suficient de mic $\phi^{-1}(D) = D_1 \cup D_2$ cu D_1 şi D_2 disjuncte în corespondenţă bijectivă şi diferenţiabilă cu D. Avem

$$\begin{aligned} & p\left(\left(f_1^2 + f_2^2, \frac{f_1}{f_2}\right) \in D\right) \\ &= p\left((f_1, f_2) \in \phi^{-1}(D)\right) \\ &= \frac{1}{4\pi\sigma^2} \int\int_{D_1} e^{-\frac{x^2+y^2}{2\sigma^2}} dxdy + \frac{1}{4\pi\sigma^2} \int\int_{D_2} e^{-\frac{x^2+y^2}{2\sigma^2}} dxdy \\ &= 2 \cdot \frac{1}{4\pi\sigma^2} \int\int_D e^{-\frac{r}{2\sigma^2}} \frac{1}{2(s^2+1)} drds \end{aligned}$$

Prin urmare perechea de v.a. $\left(f_1^2 + f_2^2, \frac{f_1}{f_2}\right)$ are densitatea $\frac{1}{4\pi\sigma^2}e^{-\frac{r}{2\sigma^2}} \cdot \frac{1}{1+s^2}$. Conform (6.12) şi remarcii de după ea rezultă că $f_1^2 + f_2^2, \frac{f_1}{f_2}$ sunt independente.

11. Să se arate că dacă $f : X \to R^2$ este o v.a. bidimensională cu densitatea $\rho_f(x_1, x_2)$ şi $g : R^2 \to R^2$ este o bijecţie de clasă C^1 cu iacobianul nenul atunci densitatea variabilei duble $g \circ f$ este $\rho_{g\circ f}(y_1, y_2) = \rho_f\left(g^{-1}(y_1, y_2)\right)\frac{D(x_1,x_2)}{D(y_1,y_2)}$.

12. Să se calculeze mediile, dispersiile, corelaţia distribuţiilor marginale pentru densitatea dublă

$$\rho(x_1, x_2) = \frac{\sqrt{12}}{2\pi}e^{-2\left(x_1^2 + x_1x_2 + x_2^2\right)}$$

Indicaţie. A se vedea exemplul 3.

13. O variabilă aleatoare bidimensională Z are densitatea dublă

$$\rho(x_1, x_2) = \begin{cases} \alpha(x_1 + x_2 + 2) \text{ dacă } (x_1, x_2) \in [0, 2] \times [1, 3] \\ 0, \text{ in rest} \end{cases}$$

Să se determine:
i) α.
ii) Densităţile distribuţiilor marginale
iii) $p(Z \in \{(x_1, x_2) \mid \quad x_2 > x_1 + 1\})$.
iv) Fie Z_1 şi Z_2 cele două distribuţii marginale. Care este media lui $Z_1 + Z_2$.

Indicaţie. $\int\int_{R^2} \rho(x_1, x_2)\, dx_1dx_2 = 1$. Se utilizează formula 6.9

14. O anumită lungime este împrţită în două şi sunt măsurate cele două porţiuni de către două persoane în mod independent. Eroarea de măsurare este o variabilă aleatoare cu densitatea $\rho(x) = \begin{cases} \frac{3}{4}\left(1 - x^2\right) \text{ pentru } x \in [-1, 1] \\ 0 \text{ în rest} \end{cases}$ pentru fiecare persoană. Care este probabilitatea ca eroarea în măsurarea întregii lungimi să fie mai mare ca 1?

Indicaţie. Cunoaştem densităţile a două variabile aleatoare independente. Deci rezultă imediat densitatea sumei şi de aici răspunsul la întrebare. Altfel, putem folosi formula 6.9

Lecţia 7

Procese aleatoare

In modelarea probabilistică a unor fenomene, presupunem implicit existenţa unei mulţimi X de factori necontrolabili, mulţime care nu poate fi exact definită, ce afectează desfăşurarea fenomenelor, precum şi existenţa unei probabilitǎti pe această mulţime. Rezultatul observaţiei unui fenomen îl reprezentăm printr-un număr sau mai multe numere. Observând de mai multe ori acelaşi fenomen, chiar reproducând condiţiile de desfăşurare cât mai fidel posibil, rezultatele numerice pot ieşi diferite. Intervin aici acei factori necontrolabili care fac să fluctueze rezultatele numerice ale experienţei într-o anumită zonă de valori. De exemplu, măsurând independent cu o ruletă obişnuită de 2m o distanţă destul de lungă, să zicem în jur de 100 m, persoane diferite vor ajunge la rezultate diferite. Nu putem specifica exact ce factori contribuie la apariţia diferenţei între rezultate şi care este contribuţia fiecăruia, dar putem admite existenţa unei probabilităţi pe această mulţime X de factori necontrolabili.

Fie acea mărime numerică pe care o urmărim, egală cu ξ. După ce fixăm valorile factorilor controlabili, ξ devine o funcţie $\xi : X \to R$, pe mulţimea factorilor necontrolabili, adică o variabilă aleatoare. Dacă ξ reprezintă o mărime ce evoluează în timp atunci $\xi : T \times X \to R$, deci variabila aleatoare este $\xi(t, \cdot) : X \to R$, pentru fiecare $t \in T \subset R$. Mai notăm această variabilă aleatoare prin ξ_t.

Definiţia 7.1 O asemenea familie de variabile aleatoare, ce depinde de un parametru $t \in T \subset R$ se numeşte proces aleator sau proces stochastic.

Fie $F(t, x)$ sau funcţia de repartiţie a acestei variabile aleatoare, adică:

$$F(t, x) = p(\{\omega \in X | \xi(t, \omega) < x\})$$

sau pe scurt

$$F(t, x) = p(\xi(t, \cdot) < x) = p(\xi_t < x)$$

Uneori vom mai nota această funcţie $F_t(x)$ pentru a sublinia că e funcţie de repartiţie în x şi depinde de parametrul t. In timp ce ξ nu poate fi specificată exact, neavând o descriere a mulţimii X, F este o funcţie de variabilele reale t, x. Cum funcţia de repartiţie conţine toate informaţiile probabilistice despre o v.a., un proces stochastic va fi descris prin funcţia sa de repartiţie $F(t,x)$. Vom caracteriza procesul stochastic prin $F_t(x)$ sau derivata ei $\rho(t,x) = \rho_t(x) = \frac{\partial F(t,x)}{\partial x}$. Dacă ξ_t ia o valori într-o mulţime discretă $\{v_1, ..v_n, ..\}$ atunci funcţia de repartiţie este complet determinată de $p_k(t) = p(\xi_t = v_k)$ sau prin $F_t(x) = \sum_{v_k<x} p_k(t)$ deci, procesul stochastic va fi descris prin aceste funcţii. Mai jos studiem câteva tipuri de procese stochastice.

7.1 Procese Poisson

Să considerăm numărul particulelor cosmice care pătrund într-un anumit volum V, într-un interval de timp. Acest număr depinde de factori incontrolabili şi apare ca o valoare aleatoare. Presupunem că ξ_t este un proces aleator pentru valori ale timpului $t \in [0,\infty)$ care ia doar valori naturale 0,1,2,...k,... , şi anume pentru $t > 0$, ξ_t ia valoarea k dacă în intervalul de timp $(0,t]$ au pătruns în volumul V, k particule cosmice. Pentru fiecare $t \geq 0$, ξ_t este o variabilă aleatoare. Pentru $t < t + \Delta t$, variabila aleatoare $\xi_{t+\Delta t} - \xi_t$ reprezintă numărul de particule cosmice care au pătruns în volumul V în intervalul de timp $(t, t+\Delta t]$. Variabilele ξ_t nu sunt independente deoarece valoarea lui $\xi_{t+\Delta t}$ depinde de valoarea lui ξ_t. Următoarele ipoteze asupra variabilelor ξ_t apar ca naturale:

a) *Absenţa post efectului,* adică numărul de particole care în intervalul de timp $[a,b)$ intră în volumul V este independent de numărul de particule care au pătruns anterior în acest volum şi de momentele la care au pătruns. Exprimăm acest fapt astfel: dacă $0 \leq a_1 < b_1 \leq a_2 < b_2 \leq ..a_n < b_n$ atunci avem v.a. $\xi_{b_1} - \xi_{a_1}, ...\xi_{b_n} - \xi_{a_n}$ care au ca valoare numărul de particule ce au pătruns în V în intervalele de timp $(a_1, b_1]$, ... $(a_n, b_n]$; ipoteza a) înseamnă că aceste v.a. sunt independente, adică

$$\begin{aligned} & p\left(\xi_{b_1} - \xi_{a_1} = k_1, \xi_{b_2} - \xi_{a_2} = k_2, ..., \xi_{b_n} - \xi_{a_n} = k_n\right) \\ = \; & p\left(\xi_{b_1} - \xi_{a_1} = k_1\right) \cdot p\left(\xi_{b_2} - \xi_{a_2} = k_2\right) \cdot ... \cdot p\left(\xi_{b_n} - \xi_{a_n} = k_n\right) \end{aligned} \tag{7.1}$$

pentru orice $k_1, ...k_n \in N$ şi orice număr n de intervale.

b) *Staţionaritatea*, înseamnă că numărul de particule ce pătrunde în volum într-un interval de timp depinde doar de lungimea intervalului de timp şi nu de plasarea acelui interval pe axa timpului. Exprimăm acest lucru prin:

$$p(\xi_b - \xi_a = k) = p\left(\xi_{a+T} - \xi_{b+T} = k\right) \tag{7.2}$$

oricare ar fi $0 \leq a < b$, $T \geq 0$ şi oricare ar fi $k \in N$.

c) *Ordinaritatea*, adică probabilitatea ca în intervalul $[t, t+\Delta t)$ să pătrundă în volum mai mult de o particulă este zero în raport cu Δt. Mai precis:

$$p\left(\xi_{t+\Delta t} - \xi_t > 1\right) = O(\Delta t)$$

Prin $O(\Delta t)$ se întelege o funcţie reală O de argument Δt, astfel ca $\lim\limits_{\Delta t \to 0} \frac{O(\Delta t)}{\Delta t} = 0$.

Definiţia 7.2 *Un proces aleator ξ_t, $t \in [0,\infty)$ unde ξ_t ia valori numere naturale şi unde sunt satisfăcute condiţiile a), b), c) de mai sus se numeşte proces Poisson.*

Multe alte fenomene din natură apar ca satisfăcând cerinţele a), b), c) : dezintegrarea spontană a nucleelor radioactive, venirea la coadă a maşinilor la o staţie de benzină, numărul de apeluri la o centrală telefonică etc.

Problema pe care vrem să o rezolvăm în continuare este de a determina probabilităţile $p_k(t) = p\left(\xi_{a+t} - \xi_a = k\right)$, adică probabilitatea ca în intr-un interval de lungime t să intre k particule cosmice în volumul V. Conform cu b) aceste probabilităţi nu depind de a.

Teorema 7.3 *Fie un proces Poisson ca mai sus. Atunci există $\lambda > 0$ astfel ca $p_1(t) = \lambda t + 0(t)$ şi*

$$p_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \tag{7.3}$$

Demonstraţie.(Schiţă) Fie $r = p_0(1)$. Impărţim intervalul $[0,1]$ prin

$$0 < \frac{1}{n} < \frac{2}{n} < .. < \frac{n}{n}$$

, pentru un $n \in N$. Din staţionaritate rezultă

$$p_0\left(\frac{1}{n}\right) = p\left(\xi_{1/n} - \xi_0 = 0\right) = p\left(\xi_{2/n} - \xi_{1/n} = 0\right) = ... = p\left(\xi_1 - \xi_{\frac{n-1}{n}} = 0\right)$$

Utilizând acum faptul că nu există post efect, avem:

$$\begin{aligned} r &= p_0(1) = p\left(\xi_{1/n} - \xi_0 = 0,\, \xi_{2/n} - \xi_{1/n} = 0,\, \xi_1 - \xi_{\frac{n-1}{n}} = 0\right) = \\ &= p\left(\xi_{1/n} - \xi_0 = 0\right) \cdot p\left(\xi_{2/n} - \xi_{1/n} = 0\right) \cdot ... \cdot p\left(\xi_1 - \xi_{\frac{n-1}{n}} = 0\right) = \\ &= \left(p\left(\xi_{1/n} - \xi_0 = 0\right)\right)^n = \left(p_0(1/n)\right)^n \end{aligned}$$

De aici rezultă $p_0\left(\frac{1}{n}\right) = r^{\frac{1}{n}}$Evenimentul că în intervalul $(0, \frac{k}{n}]$ nu intră nici o particulă în V, este echivalent cu intersecţia evenimentelor " în intervalul $(\frac{i-1}{n}, \frac{i}{n}]$ nu intră vreo particulă în V", pentru $1 \leq i \leq k$. Cum aceste evenimente sunt independente, rezultă $p_0\left(\frac{k}{n}\right) = \left(p_0\left(\frac{1}{n}\right)\right)^k = r^{\frac{k}{n}}$. Acum, deoarece $p_0(t)$ este crescătoare în t, rezultă uşor că $p_0(t) = r^t$. Cum r este o probabilitate, $r \in (0,1)$, deci $r = e^{-\lambda}$ pentru un $\lambda > 0$, deci $p_0(t) = e^{-\lambda t}$. Am demonstrat deci formula 7.3 pentru k=0.

Mai departe avem

$$p_0(t) + p_1(t) + \underbrace{\sum_{k>1} p_k(t)}_{=0(t) \text{ conform cu c)}} = 1$$

adică $e^{-\lambda t} + p_1(t) + 0(t) = 1$. De aici rezultă

$$p_1(t) = 1 - e^{-\lambda t} + O(t) = \lambda t + O(t)$$

Pentru a determina $p_k(t)$ este uşor de justificat formula

$$\begin{aligned} p_k(t+\Delta t) &= \sum_{i=0}^{k} p_i(\Delta t)\, p_{k-i}(t) \\ &= e^{-\lambda \Delta t} p_k(t) + (\lambda \Delta t + 0(\Delta t)\, p_{k-1}(t) + \underbrace{\sum_{i=2}^{k} p_i(\Delta t)\, p_{k-i}(t)}_{=0(\Delta t)\text{ conform cu c)}} \\ &= p_k(t) - \lambda \Delta t p_k(t) + \lambda \Delta t p_{k-1}(t) + 0(\Delta t) \end{aligned}$$

pentru k$\geq$ 1. Prin trecerea lui $p_k(t)$ în membrul stâng şi divizarea la Δt, apoi trecerea la limită $\Delta t \to 0$, rezultă

$$\frac{dp_k(t)}{dt} = -\lambda p_k(t) + \lambda p_{k-1}(t) \tag{7.4}$$

Dacă ţinem seama de condiţiile $p_0(t) = e^{-\lambda t}$, $p_k(0) = 0$ (din ordinaritate), atunci sistemul 7.4 se rezolvă şi se găseşte $p_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$.

QED.

Probabilităţile $p_k(t) = p(\xi_t = k)$ sunt la fel ca la o v.a. Poisson (vezi lecţia 4), deci media lui ξ_t este $M(\xi_t) = \lambda t$ şi dispersia $D(\xi_t) = \lambda t$.

7.2 Procese Markov discrete

Definiţia 7.4 *Un proces stochastic se numeşte proces Markov discret dacă ξ_t poate lua un număr finit de valori $V = \{v_1, v_2, ...v_n\}$, timpul t variază într-o mulţime discretă de valori $T = \{t_1, t_2, ...t_n, ...\}$ şi*

$$p\left(\xi_{t_k} = v \,|\, \xi_{t_{k-1}} = v_{i_{k-1}}, \xi_{t_{k-2}} = v_{i_{k-2}}, .., \xi_{t_1} = v_{i_1}\right) = p\left(\xi_{t_k} = v \,|\, \xi_{t_{k-1}} = v_{i_{k-1}}\right)$$

pentru orice $v, v_{i_{k-1}}, ...v_{i_1} \in V$.

Cu alte cuvinte probabilitatea ca sistemul (variabila aleatoare ξ) să ajungă în momentul t_k în starea v_{i_k} depinde doar de starea $v_{i_{k-1}}$ de la momentul imediat anterior, t_{k-1}, şi nu depinde de stările la momentele t_{k-2}, $t_{k-3}, ...$ t_1. Fără a pierde din generalitate putem considera mulţimea stărilor $V = \{1, 2, ...n\}$, iar mulţime valorilor temporale $T = \{0, 1, 2, ...n, ...\}$. Vom nota cu $p_{i,j}(k) = p\left(\xi_k = j \,|\, \xi_{k-1} = i\right)$, adică $p_{i,j}(k)$ este probabilitatea ca în cazul când la momentul k-1 sistemul este în starea i, el să ajungă în starea j la momentul următor, k. Aceste mărimi se numesc *probabilităţi de tranziţie*. Deoarece din starea i se poate ajunge la momentul următor doar în stările 1,2,3,...n, trebuie să avem $p_{i,1}(k) + p_{i,2}(k) + ... + p_{i,n}(k) = 1$. Cel mai simplu proces de acest tip este acela în care probabilităţile de tranziţie $p_{i,j}(k)$ nu depind de momentul k. In acest caz avem $p_{i,j}(k) = p_{i,j}$. Matricea p cu elementele $p_{i,j}$ se numeşte matricea de tranziţie. Intr-o astfel de matrice $\sum_{j=1}^{n} p_{i,j} = 1$, $p_{i,j} \geq 0$.

Exemplul 7.5 *Să presupunem că o particulă se poate mişca între două bariere $a < b$ trecând prin puncte intermediare*

$$a = x_1 < x_2 < ... < x_n = b$$

Dacă particula se găseşte în poziţia x_i atunci cu probabilitatea r se deplasează înainte în poziţia x_{i+1} şi cu probabilitatea $s = 1 - p$ se deplasează înapoi în poziţia x_{i-1}. In capetele a şi b particula este respinsă în poziţia imediat vecină. Pentru 4 poziţii posibile, matricea de tranziţie este

$$p = \begin{pmatrix} 0 & 1 & 0 & 0 \\ s & 0 & r & 0 \\ 0 & s & 0 & r \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Fie acum $p_{i,j}^{(k)} = p(\xi_k = j \,|\, \xi_0 = i)$, adică $p_{i,j}^{(k)}$ este probabilitatea ca sistemul să se găsească la momentul k în starea j, dacă la momentul iniţial 0 s-a găsit în starea i. Fie $p^{(k)}$ matricea de tranziţie de la momentul 0 la momentul k, având elementele $p_{i,j}^{(k)}$. Din formula probabilităţii totale găsim $p(\xi_k = j \,|\, \xi_0 = i) = \sum_{l=0}^{n} p\left(\xi_{k-1} = l \,|\, \xi_0 = i\right) \cdot p\left(\xi_k = j \,|\, \xi_{k-1} = l,\, \xi_0 = i\right)$
$= \sum_{l=0}^{n} p\left(\xi_{k-1} = l \,|\, \xi_0 = i\right) \cdot p\left(\xi_k = j \,|\, \xi_{k-1} = l\right)$
altfel spus

$$p_{i,j}^{(k)} = \sum_{l=0}^{n} p_{i,l}^{(k-1)} \cdot p_{l,j}$$

Din această formulă rezultă prin inducţie

$$p^{(k)} = p^k$$

adică $p^{(k)}$ este puterea de ordinul k a matricei de tranziţie.

Exemplul 7.6 *Pentru matricea de tranziţie din exemplul 1 găsim*

$$p^2 = \begin{pmatrix} s & 0 & r & 0 \\ 0 & s + r \cdot s & 0 & r^2 \\ s^2 & 0 & r + r \cdot s & 0 \\ 0 & s & 0 & r \end{pmatrix}$$

Pentru r=3/4 şi s=1/4 matricea de tranziţie devine:

$$p = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1/4 & 0 & 3/4 & 0 \\ 0 & 1/4 & 0 & 3/4 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Pentru valori mari ale lui k găsim:

$$p^{(100)} = \begin{pmatrix} 0,0769231 & 0 & 0,923077 & 0 \\ 0 & 0,307692 & 0 & 0,692308 \\ 0.0769231 & 0 & 0,923077 & 0 \\ 0 & 0,307692 & 0 & 0,692308 \end{pmatrix}$$

Cum arată probabilitătile $p_{i,j}^{(k)}$ pentru valori mari ale lui k? Următoarea teoremă aduce lămuriri în această privinţă.

Teorema 7.7 *Dacă pentru $m \in N$ are loc inegalitatea $p_{i,j}^{(m)} > 0$ pentru orice i, j, atunci există $\lim_{k\to\infty} p^{(k)}$ şi*

$$\lim_{k\to\infty} p_{i,j}^{(k)} = p_j$$

independent de i.

Altfel spus, probabilitătile de tranziţie de la starea i în momentul 0, la starea j în momentul k, se stabilizează pentru $k \to \infty$ la valori independente de starea la momentul 0.

Pentru demonstraţie se poate consulta bibliografia.

Exemplul 7.8 *Fie matricea de tranziţie*

$$p = \begin{pmatrix} 1/8 & 2/8 & 2/8 & 3/8 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/10 & 5/10 & 2/10 & 2/10 \\ 7/20 & 3/20 & 5/20 & 5/20 \end{pmatrix}$$

Utilizând calculatorul găsim

$$p^7 = \begin{pmatrix} 0,214013 & 0,283041 & 0,238095 & 0,264851 \\ 0,21402 & 0,283038 & 0,238095 & 0,264846 \\ 0,214014 & 0,28304 & 0,238095 & 0,26485 \\ 0,214024 & 0,283037 & 0,238095 & 0,264844 \end{pmatrix}$$

$$p^{100} = \begin{pmatrix} 0,214018 & 0,283039 & 0,238095 & 0,264848 \\ 0,214018 & 0,283039 & 0,238095 & 0,264848 \\ 0,214018 & 0,283039 & 0,238095 & 0,264848 \\ 0,214018 & 0,283039 & 0,238095 & 0,264848 \end{pmatrix}$$

Se vede cum de la p^7 se stabilizează primele trei zecimale ale probabilităţilor limită, la p^{100} fiind stabilizate cel puţin 6 zecimale.

7.3 Procese de naştere şi moarte

Să presupunem că într-un proces stochastic ξ_t *ia valori numere naturale* şi că următoarele condiţii sunt îndeplinite:

i) $p\left(\xi_{t+\Delta t} = k+1 | \xi_t = k\right) = \lambda_k \cdot \Delta t + O\left(\Delta t\right),\ \lambda > 0$,(aceasta este probabilitatea unui proces de naştere în intervalul $(t, t+\Delta t)$)

ii) $p\left(\xi_{t+\Delta t} = k-1 | \xi_t = k\right) = \mu_k \cdot \Delta t + O\left(\Delta t\right),\ \mu > 0, \quad k \geq 1$(aceasta este probabilitatea unui proces de moarte în intervalul $(t, t+\Delta t)$.

iii) $p\left(\left|\xi_{t+\Delta t} - \xi_t\right| > 1\right) = O\left(\Delta t\right)$ (aceasta este probabilitatea a mai mult de o naştere sau o moarte în intervalul $(t, t+\Delta t)$)

Definiţia 7.9 *Un proces stochastic cu valori naturale şi care îndeplineşte condiţiile i)-iii) de mai sus se numeşte proces de naştere şi moarte.*

Dacă $\mu_k = 0$ atunci procesul se numeşte proces de naştere iar dacă $\lambda_k = 0$ se numeşte proces de moarte. Deoarece pentru $k \geq 1$ avem

$$\begin{aligned} 1 &= \sum_{n=0}^{\infty} p\left(\xi_{t+\Delta t} = n | \xi_t = k\right) \\ &= p\left(\xi_{t+\Delta t} = k-1 | \xi_t = k\right) + p\left(\xi_{t+\Delta t} = k | \xi_t = k\right) \\ &\quad + p\left(\xi_{t+\Delta t} = k+1 | \xi_t = k\right) + p\left(\left(\left|\xi_{t+\Delta t} - k\right| \geq 2\right) | \xi_t = k\right) \end{aligned}$$

rezultă imediat din i). ii), iii) că

iv) $p\left(\xi_{t+\Delta t} = k | \xi_t = k\right) = 1 - \lambda_k \Delta t - \mu_k \Delta t + O\left(\Delta t\right)$.

Problema care se pune la aceste procese constă în determinarea probabilităţilor $p\left(\xi_t = k\right) = p_k\left(t\right)$ **cunoscând** $p_k\left(0\right) = p\left(\xi_0 = k\right)$**.**

Soluţie. Deoarece evenimentele $\left\{\xi_t = k\right\}_{k \in N}$ sunt disjuncte avem conform formulei probabilităţii totale pentru $k \geq 1$

$$\begin{aligned} p\left(\xi_{t+\Delta t} = k\right) &= p\left(\xi_t = k\right) \cdot p\left(\xi_{t+\Delta t} = k | \xi_t = k\right) && (7.5) \\ &\quad + p\left(\xi_t = k+1\right) \cdot p\left(\xi_{t+\Delta t} = k | \xi_t = k+1\right) && (7.6) \\ &\quad + p\left(\xi_t = k-1\right) \cdot p\left(\xi_{t+\Delta t} = k | \xi_t = k-1\right) && (7.7) \\ &\quad + p\left(\left|\xi_t - k\right| \geq 2\right) \cdot p\left(\xi_{t+\Delta t} = k \middle| \left|\xi_t - k\right| \geq 2\right) && (7.8) \end{aligned}$$

Folosind relaţia iv) în (7.5), relaţia ii) în (7.6), relaţia i) în (7.7), relaţia iii) în (7.8) găsim:

$$\begin{aligned} p_k(t+\Delta t) &= p_k(t)\cdot(1-\lambda_k\Delta t-\mu_k\Delta t+O(\Delta t)) \\ &+p_{k+1}(t)\cdot\left(\mu_{k+1}\Delta t+O(\Delta t)\right) \\ &+p_{k-1}(t)\cdot(\lambda_{k-1}\Delta t+O(\Delta t)) \\ &+O(\Delta t) \end{aligned}$$

Trecând $p_k(t)$ în membrul stâng, divizând prin Δt, apoi trecând la limită $\Delta t \to 0$ găsim

$$p_k'(t) = \lambda_{k-1}p_{k-1}(t) - (\lambda_k+\mu_k)\,p_k(t) + \mu_{k+1}p_{k+1}(t) \tag{7.9}$$

Pentru cazul $k=0$ o analiză similară conduce la:

$$p_0'(t) = -\lambda_0 p_0(t) + \mu_1 p_1(t) \tag{7.10}$$

Rezolvarea acestui sistem infinit de ecuaţii diferenţiale este anevoioasă, însă în practică după un proces de tranziţie haotic urmează o stabilizare, în care $p_k(t)$ sunt constante. Ecuaţiile în acest caz sunt

$$\begin{cases} 0 = -\lambda_0 p_0 + \mu_1 p_1 \\ 0 = \lambda_{k-1}p_{k-1} - (\lambda_k+\mu_k)\,p_k + \mu_{k+1}p_{k+1} \end{cases} \tag{7.11}$$

Din prima ecuaţie rezultă $p_1 = \frac{\lambda_0}{\lambda_1}p_0$. Folosind celelalte ecuaţii găsim succesiv $p_1 = \frac{\lambda_0\lambda_1}{\mu_1\mu_2}$, ş.a.m.d. In general obţinem:

$$p_n = \frac{\lambda_0\lambda_1\ldots\cdot\lambda_{n-1}}{\mu_1\mu_2\ldots\cdot\mu_n}p_0 \tag{7.12}$$

Valoarea lui p_0 se determină din condiţia

$$\sum_{i=0}^{\infty} p_i = 1 \tag{7.13}$$

şi acest lucru este posibil numai dacă $\sum_n \frac{\lambda_0\lambda_1\ldots\cdot\lambda_{n-1}}{\mu_1\mu_2\ldots\cdot\mu_n} < \infty$.

In acest fel poate fi modelat fenomenul de aşteptat la coadă.

7.3.1 Model de aşteptare cu o singură staţie de deservire şi un număr mare de unităţi ce au nevoie de serviciile staţiei

Se poate imagina o astfel de situaţie la o benzinărie cu o singură staţie la care vin aleator maşini pentru alimentare. Maşinile provin dintr-o populaţie mare astfel încât coada nu este influenţată de diminuarea numărului de maşini care necesită alimentare.

Prima problemă care se pune este modelarea caracterului întîmplător al venirilor în staţie şi al plecărilor din staţie. Fie λ numărul mediu de veniri în unitatea de timp. Deci într-un timp Δt vor fi în medie $\lambda\Delta t$ veniri. Dacă numărul de maşini din care se vine la coadă este n iar nevoia de benzină este întîmplătoare şi independentă de la o maşină la alta atunci probabilitatea de a veni la coadă k maşini în intervalul $(t, t + \Delta t)$ este $C_n^k p^k q^{n-k}$ unde p este probabilitatea ca o maşină să aibă nevoie de benzină iar $q = 1 - p$ (vezi legea binomială). Numărul mediu de veniri este np iar pe de altă parte este $\lambda \cdot \Delta t$. Deci $np = \lambda\Delta t \Rightarrow p = \frac{\lambda\Delta t}{n}$. In aceste condiţii ştim lecţia 4 că că $C_n^k p^k q^{n-k} \to e^{-\lambda\Delta t}\frac{(\lambda\Delta t)^k}{k!}$ atunci când $n \to \infty$. Cum n este mare, putem considera că probabilitatea ca la coadă să vină k maşni în intervalul $(t, t+\Delta t)$ este $e^{-\lambda\Delta t}\frac{(\lambda\Delta t)^k}{k!}$. **Numărul de veniri la coadă în intervalul de timp Δt poate fi considerat ca o variabilă aleatoare de tip Poisson:**

$$\begin{pmatrix} 0 & 1 & ... & k & ... \\ e^{-\lambda\Delta t} & e^{-\lambda\Delta t}\frac{\lambda\Delta t}{1} & ... & e^{-\lambda\Delta t}\frac{(\lambda\Delta t)^k}{k!} & ... \end{pmatrix}$$

Prin urmare :

a) Probabilitatea ca în intervalul $(t, t + \Delta t)$ să vină în staţie o maşină este $e^{-\lambda\Delta t}\frac{\lambda\Delta t}{1} = \lambda \cdot \Delta t + O(\Delta t)$.

b) Probabilitatea ca să nu vină în staţie nici o maşină în intervalul $(t, t+\Delta t)$ este $e^{-\lambda\Delta t} = 1 - \lambda \cdot \Delta t + O(\Delta t)$.

c) Probabilitatea ca în intervalul $(t, t + \Delta t)$ să vină în staţie mai mult de o maşină este $\sum_{k\geq 2} e^{-\lambda\Delta t}\frac{(\lambda\Delta t)^k}{k!} = O(\Delta t)$.

In mod analog, fie μ numărul mediu de maşini deservite de o staţie în unitatea de timp(presupunând că are continuu de lucru). Intr-un interval de lungime Δt vor fi în medie deservite $\mu \cdot \Delta t$ maşini. Numărul real al celor deservite poate fi mai mare sau mai mic decît media şi depinde de factori întâmplători, ca necesarul de benzină, întîrzieri produse de şofer, etc. Admitem că:

a') Probabilitatea ca în intervalul $(t, t + \Delta t)$ să plece din staţie o maşină, dacă există vreuna, este $\mu\Delta t + O(\Delta t)$

b') Probabilitatea ca în intervalul $(t, t + \Delta t)$ să nu plece din staţie nici o maşină, dacă există vreuna, este $1 - \mu\Delta t + O(\Delta t)$

c') Probabilitatea ca în intervalul $(t, t + \Delta t)$ să plece din staţie mai mult de o maşină este $O(\Delta t)$.

Ipotezele făcute asupra plecărilor din staţie sunt asemănătoare cu condiţiile a), b),c) îndeplinite de probabilităţile de venire în staţie.

Fie acum familia de variabile aleatoare ξ_t, $t \in T = [0, \infty)$, unde valoarea lui ξ_t este egală cu numărul de maşini în staţie la momentul t. Ipotezele făcute asupra venirilor şi plecărilor, independente unele de altele, se mai scriu:

i)

$$\begin{aligned} p\left(\xi_{t+\Delta t}=k+1|\xi_t=k\right) &= \underbrace{(\lambda\cdot\Delta t+O\left(\Delta t\right))}_{o\,venire}\cdot\underbrace{(1-\mu\Delta t-O\left(\Delta t\right))}_{nici\,o\,plecare}+ \\ +\underbrace{O\left(\Delta t\right)}_{mai\,mult\,de\,o\,venire\,sau\,o\,plecare} &= \lambda\cdot\Delta t+O\left(\Delta t\right) \end{aligned}$$

ii)

$$\begin{aligned} p\left(\xi_{t+\Delta t}=k-1|\xi_t=k\right) &= \underbrace{(\mu\cdot\Delta t+O\left(\Delta t\right))}_{o\,plecare}\cdot\underbrace{(1-\lambda\Delta t-O\left(\Delta t\right))}_{nici\,o\,venire}+ \\ +\underbrace{O\left(\Delta t\right)}_{mai\,mult\,de\,o\,plecare\,sau\,venire} &= \mu\cdot\Delta t+O\left(\Delta t\right) \end{aligned}$$

din condiţia a') de la plecări şi condiţia a) de la veniri (o plecare şi nici o venire).

iii) $p\left(\left|\xi_{t+\Delta t}-\xi_t\right|>1\right)=O\left(\Delta t\right)$ din condiţiile c) de la veniri şi c') de la plecări

Prin urmare avem de a face cu un proces de naştere şi moarte. Mărimile $p_k\left(t\right)=p\left(\xi_t=k\right)$ satisfac deci ecuaţiile (7.9)-(7.10), iar la stabilizare ecuaţiile (7.11). Soluţia la stabilizare este (7.12) unde $\lambda_k=\lambda$ şi $\mu_k=\mu$, adică

$$p_k=\left(\frac{\lambda}{\mu}\right)^k p_0 \tag{7.14}$$

iar din $1=p_0\sum_{k=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^k=p_0\frac{1}{1-\frac{\lambda}{\mu}}$ rezultă

$$p_0=1-\frac{\lambda}{\mu}$$

Stabilizarea se poate face doar cu condiţia $0<\frac{\lambda}{\mu}<1$ astfel încât seria $\sum\left(\frac{\lambda}{\mu}\right)^k$ să fie convergentă.

7.3.2 Model de aşteptare cu o singură staţie iar numărul de unităti care au nevoie de serviciile staţiei este limitat la o valoare dată.

Presupunem acum că populaţia de unde provin maşinile (unităţile) în aşteptare este limitată la un număr de N exemplare. Coada care se formează depinde de cât de des au nevoie maşinile de serviciile staţiei şi cât de prompte sunt aceste servicii. In ce priveşte capacitatea de deservire a staţiei, nu apar elemente care să modifice ipotezele a'), b'), c') anterioare. In ce priveşte venirile în staţie, ele apar în mod normal proporţionale cu numărul maşinilor în activitate (deci care nu sunt aşezate la coadă). Vom face ipoteza că maşinile au nevoie independent de serviciile staţiei, şi dacă în staţie sunt k maşini, atunci probabilitatea ca să mai vină una în următoarele Δt unităţi de timp este $(N-k)\,\lambda\Delta t + 0\,(\Delta t)$ dacă $k < N$ şi zero dacă $k = N$. Astfel familia de variabile aleatoare ξ_t care au ca valoare numărul de maşini în staţie este un proces de naştere şi moarte cu $\lambda_k = (N-k)\,\lambda$,iar $\mu_k = \mu$ pentru $0 \leq k \leq N$. Prin urmare, conform formulelor (7.12) rezultă

$$p_k = \frac{N\,(N-1)\cdot ...\,(N-k+1)\,\lambda^k}{\mu^k} p_0, \quad 0 \leq k \leq N \tag{7.15}$$

Valoarea lui p_0 se determină din condiţia ca $\sum_{k=0}^{n} p_k = 1$.

Un asemenea proces ar putea modela de exemplu coada la reparaţii într-o intreprindere cu N maşini dacă serviciul de reparaţii are doar un mecanic. Un model mai realist este:

7.3.3 Model de aşteptare cu n staţii de deservire şi cu N unităti ce trebuie deservite (1<n<N)

In acest caz avem de a face cu un proces de naştere şi moarte în care $\lambda_k = N - k$ pentru $0 \leq k \leq N$ (probabilităţile de a se veni la staţie pentru alimentare, reparaţii etc. sunt proporţionale cu numărul unităţilor în serviciu), şi $\mu_k = k\mu$ dacă $1 \leq k < n$, $\mu_k = n\mu$ dacă $n \leq k \leq N$ (probabilitatea ca o unitate să părăsească staţia este proporţională cu numărul celor în curs de deservire). Conform formulei (7.12) găsim

$$p_k = \begin{cases} C_N^k \left(\frac{\lambda}{\mu}\right)^k p_0 & \text{pentru } 1 \leq k < n-1 \\ \frac{k!}{n!n^{k-n}} C_N^k \left(\frac{\lambda}{\mu}\right)^k p_0 & \text{pentru } n \leq k \leq N \end{cases} \tag{7.16}$$

Si aici p_0 se determină din condiţia ca $\sum p_k = 1$.

7.4 Procese aleatoare staţionare

Un proces aleator se numeşte staţionar dacă proprietăţile sale statistice sunt invariante la o translaţie a timpului. Fie F_t funcţia de repartiţie a v.a. ξ_t şi fie ρ_i densitatea de probabilitate dacă există. In continuare în această secţiune timpul t aparţine mulţimii $T = (-\infty, \infty)$ sau mulţimii $T = [0, \infty)$, în afară de cazul când se specifică o altă situaţie.

Definiţia 7.10 *Procesul aleator ξ_t se numeşte staţionar de ordinul întâi dacă $F_t = F_{t+\tau}$ pentru orice τ.*

Prin urmare pentru procese staţionare de ordinul unu, funcţia de repartiţie este aceeaşi pentru toate variabilele ξ_t.

Să considerăm acum pentru două valori t_1, t_2 vectorul aleator $\left(\xi_{t_1}, \xi_{t_2}\right)$. Acest vector are o funcţie de repartiţie mixtă (vezi lecţia 6) definită prin $F_{t_1,t_2}(x,y) = p\left(\xi_{t_1} < x, \xi_{t_2} < y\right)$.

Definiţia 7.11 *Spunem că procesul aleator ξ_t este staţionar de ordinul doi dacă funcţia de repartiţie mixtă este invariantă la o translaţie a timpului. Mai precis $F_{t_1,t_2} = F_{t_1+\tau,t_2+\tau}$ pentru orice τ.*

In mod analog se poate defini un proces staţionar de ordinul n, prin funcţia de repartiţie n dimensională a vectorului aleator $\left(\xi_{t_1}, \xi_{t_2}, ...\xi_{t_n}\right)$. Dacă un proces aleator este staţionar de ordin n, atunci el este staţionar de orice ordin $0 \leq k \leq n$.

Pentru un proces staţionar media,dispersia, momentele de diverse ordine ale variabilei ξ_t nu depind de t.Variabilele aleatoare ξ_t nu sunt în general independente. Dacă m este media variabilei ξ_t (independent de t) atunci coeficientul de corelaţie

$$c\left(\xi_{t_1}, \xi_{t_2}\right) = \frac{M\left(\left(\xi_{t_1} - m\right) \cdot \left(\xi_{t_2} - m\right)\right)}{\sqrt{M\left(\left(\xi_{t_1} - m\right)^2\right) \cdot M\left(\left(\xi_{t_2} - m\right)^2\right)}}$$

(vezi lecţia 6) este şi el invariant la o translaţie în timp, adică $c\left(\xi_{t_1}, \xi_{t_2}\right) = c\left(\xi_{t_1+\tau}, \xi_{t_2+\tau}\right)$ pentru orice τ. Insă în multe aplicaţii ale probabilităţilor se utilizează doar media, dispersia pentru variabilele aleatoare individuale şi coeficientul de corelaţie pentru perechile de variabile aleatoare. Aceste mărimi pot fi invariante la o translaţie a timpului fără ca funcţia de repartiţie să fie. De aceea s-au studiat în mod special procesele aleatoare pentru care aceste mărimi sunt invariante la o translaţie a timpului, fără să se ceară şi invarianţa funcţiei de repartiţie.

Definiţia 7.12 *Un proces aleator $(\xi_t)_{t\in T}$ se numeşte staţionar în sens larg dacă*
a) media $M(\xi_t) = m$ este independentă de $t \in T$.
b) dispersia $M(\xi_t) = \sigma^2$ este independentă de $t \in T$.
c) $c\left(\xi_{t_1}, \xi_{t_2}\right) = c\left(\xi_{t_1+\tau}, \xi_{t_1+\tau}\right)$ pentru orice $t_1, t_2, t_1 + \tau, t_2 + \tau \in T$.

Observaţia 7.13 *Condiţiile de mai sus sunt echivalente cu*
a') media $M(\xi_t) = m$ este independentă de $t \in T$.
b') $M\left(\xi_t \xi_{t+\tau}\right)$ este funcţie doar de τ (deci independentă de t)
Demonstraţia echivalenţei celor două seturi de condiţii este lăsată ca exerciţiu.

In continuare vom nota $R(t_1, t_2) = M\left(\xi_{t_1} \xi_{t_2}\right)$ şi o vom numi *funcţia de autocorelaţie* a procesului aleator. Dacă procesul este staţionar în sens larg, această funcţie este invariantă la o translaţie în timp, deci depinde doar de diferenţa $t_2 - t_1$. Vom nota în acest caz funcţia de autocorelaţie prin $R(\tau) = M\left(\xi_t \xi_{t+\tau}\right) = M(\xi_0 \xi_\tau)$. Câteva din proprietăţile funcţiei de autocorelaţie pentru procese staţionare în sens larg, apar în continuare:

1.
$$R(0) = M_2(\xi_t)$$

pentru că $R(0) = M\left(\xi_t \xi_{t+0}\right) = M_2(\xi_t)$.

2.
$$R(-\tau) = R(\tau)$$

pentru că $R(-\tau) = M(\xi_t \xi_{t-\tau}) = M(\xi_{t-\tau}\xi_t) = R(\tau)$.

3.

$$R(0) \geq |R(\tau)|$$

pentru că din $M\left((\xi_t \pm \xi_{t+\tau})^2\right) \geq 0$ rezultă $M(\xi_t^2) + M(\xi_{t+\tau}^2) \pm 2M(\xi_t \xi_{t+\tau}) \geq 0$, sau $2R(0) \pm 2R(\tau) \geq 0$.

4. Pentru orice $n \in N$ şi $\tau_1, \tau_2, ..\tau_n \in R$ forma pătratică

$$\sum_{i,j=1}^{n} R(\tau_i - \tau_j) x_i x_j$$

este pozitiv definită pentru că este egală cu $M\left(\left(\sum_{i=1}^n x_i \xi_{t+\tau_i}\right)^2\right) \geq 0$, oricare ar fi t astfel ca $t + \tau_i \in T$ pentru orice i.

Definiţia 7.14 *Procesul aleator staţionar în sens larg, $(\xi_t)_{t\in T}$, se numeşte continuu dacă $\lim_{t\to 0} M\left((\xi_t - \xi_0)^2\right) = 0$. Acest lucru este echivalent cu $\lim_{t\to t_0} M\left((\xi_t - \xi_{t_0})^2\right) = 0$ pentru orice $t_0 \in T$.*

Se mai spune că procesul aleator este continuu în medie pătratică.

5. Dacă procesul aleator staţionar în sens larg $(\xi_t)_{t\in T}$ este continuu atunci funcţia de autocorelaţie $R(\tau)$ este continuă (exerciţiu).

6. Funcţia de autocorelare a unui proces staţionar în sens larg, continuu, se poate pune sub forma

$$R(\tau) = \int_{-\infty}^{\infty} e^{i\tau\omega} dF(\omega)$$

unde $F(x)$ este mărginită, monoton crescătoare, continuă la stânga. Dacă F admite o derivată, f, ceea ce se întâmplă dacă $|R|$ scade destul de rapid când $|\tau| \to \infty$, atunci $R(\tau) = \int_{-\infty}^{\infty} e^{i\tau\omega} f(\omega)\, d\omega$. Demonstraţia acestui nu face obiectul acestui curs. $F(\omega)$ se numeşte funcţia spectrală a procesului (ξ_t) iar $f(\omega)$ se numeşte densitatea spectrală. Din formulele de inversiune ale transformării Fourier găsim că

$$f(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-i\tau\omega} R(\tau)\, d\tau$$

Mai mult, ţinând seama că $R(\tau)$ este pară se poate arăta că

$$R(\tau) = \int_{-\infty}^{\infty} \cos(\tau\omega)\, dF(\omega)$$

7. Orice proces aleator continuu staţionar în sens larg se poate aproxima prin procese aleatoare standard. Anume, pentru orice $A > 0$ şi orice $\varepsilon > 0$ există un număr finit n de

variabile aleatoare ϕ_1, η_1, ϕ_2, η_2,...ϕ_n, η_n independente două câte două şi există n numere reale λ_1, λ_2, ...λ_n astfel ca pentru orice $t \in [-A, A]$ are loc relaţia

$$M\left(\left(\xi_t - \sum_{k=1}^{n} (\phi_k \cos \lambda_k t + \eta_k \sin \lambda_k t)\right)^2\right) < \varepsilon$$

7.5 Exerciţii

1. Care este probabilitatea ca în cazul unui serviciu cu 5 staţii de deservire a 20 de unităţi şi cu raportul între coeficienţii de venire şi deservire $\rho = \frac{\lambda}{\mu} = \frac{1}{5}$, să nu poată fi la un moment dat satisfăcută o cerere? Dar dacă $\rho = \frac{4}{5}$? (o asemenea situaţie poate fi într-o intreprindere unde o 5 mecanici întreţin 20 de maşini).

Soluţie. Avem un proces de aşteptare cu n=5 staţii şi N=20 unităţi, deci putem aplica formulele (7.16), pentru $k = n$, de unde

$$p = \frac{\sum_{k=5}^{20} \frac{k!}{n! \cdot n^{k-n}} C_{20}^k \rho^k}{\sum_{k=0}^{4} C_{20}^k \rho^k + \sum_{k=5}^{20} \frac{k!}{n! \cdot n^{k-n}} C_{20}^k \rho^k}$$

Cu un computer, pentru $\rho = 1/5$ găsim $p = 0,275$, iar pentru $\rho = 4/5$ găsim $p = 0,9993$.

2. Să presupunem că numărul de staţii este n iar solicitările vin dintr-o populaţie foarte mare, deci numărul de veniri nu este influenţat de numărul de unităţi în aşteptare. Care este probabilitatea ca o solicitare să poată fi satisfăcută imediat? Caz particular $n = 5$, $\rho = \frac{\lambda}{\mu} = \frac{1}{2}$?

Soluţie. In acest caz avem un proces de naştere şi moarte în care $\lambda_k = \lambda$ şi în care $\mu_k = k \cdot \mu$ pentru $0 \leq k \leq n$ şi $\mu_k = n\mu$ pentru $k \geq n$ (probabilitatea de a ieşi o unitate din sistem este proporţională cu numărul unităţilor în curs de deservire). Găsim, conform cu (7.12)

$$p_k = \begin{cases} \frac{\rho^k}{k!} p_0, \text{ pentru } k \leq n \\ \frac{\rho^k}{n! n^{k-n}} p_0, \text{ pentru } k > n \end{cases}$$

Condiţia $\sum p_k = 1$ conduce la

$$p_0 = \frac{1}{\sum_{k=0}^{n} \frac{\rho^k}{k!} + \frac{n^n}{n!} \sum_{k=n+1}^{\infty} \left(\frac{\rho}{n}\right)^k} = \frac{1}{\sum_{k=0}^{n} \frac{\rho^k}{k!} + \frac{\rho^{n+1}}{n!(n-\rho)}}$$

In cazul concret al problemei, o solicitare poate fi satisfăcută dacă numărul de unităţi din sistem este mai mic de n. Probabilitatea căutată este

$$p = \sum_{k=0}^{n-1} p_k = \frac{\sum_{k=0}^{n-1} \frac{\rho^k}{k!}}{\sum_{k=0}^{n} \frac{\rho^k}{k!} + \frac{\rho^{n+1}}{n!(n-\rho)}}$$

Pentru $n = 5$, $\rho = 1/2$ găsim $p = 0,998$. Dacă venirile sunt mai pronunţate, de exemplu $\rho = \frac{\lambda}{\mu} = 3$ (probabilitatea de solicitare a unui serviciu într-un interval scurt Δt este de trei ori mai mare ca probabilitatea de ieşire de la un server în lucru), atunci găsim o probabilitate de satisfacere imediată a cererii mai mică, $p = \frac{263}{343} = 0,7638$.

3. La o centrală telefonică vin apeluri aleatoare, independente. Centrala poate deservi simultan $n = 50$ de cereri, iar apelurile care vin când centrala este ocupată se anulează. Presupunând că raportul $\rho = \frac{\lambda}{\mu}$ dintre probabilitatea $\lambda\Delta t$ de venire a unui apel într-un timp scurt Δt şi probabilitatea $\mu\Delta t$ de eliberare într-un timp scurt Δt a unui circuit ocupat este $\rho = 50$, să se calculeze probabilitatea ca un apel telefonic să fie anulat.

Soluţie. Şi în acest caz avem un proces de naştere şi moarte. Fie p_k probabilitatea ca în centrală să fie k circuite ocupate. Coeficientul λ_k pentru "naştera unui apel" este fix, egal cu λ. Coeficientul μ_k pentru "moartea unui apel" este $\mu_k = k \cdot \mu$, pentru că dacă sunt k circuite ocupate atunci probabilitatea de eliberare a unui circuit creşte de k ori faţă de cazul unui singur circuit ocupat. Conform cu formulele (7.12) avem

$$\begin{aligned} p_k &= \frac{\lambda_0\lambda_1 \ldots \cdot \lambda_{k-1}}{\mu_1\mu_2 \ldots \cdot \mu_k} p_0 = \frac{\lambda^k}{k!\mu^k} p_0 = \frac{\rho^k}{k!} p_0, \text{ pentru } 0 \le k \le n \\ p_k &= 0, \text{ pentru } k > n \end{aligned}$$

Din condiţia $\sum_{k=0}^{n} p_k = 1$ rezultă

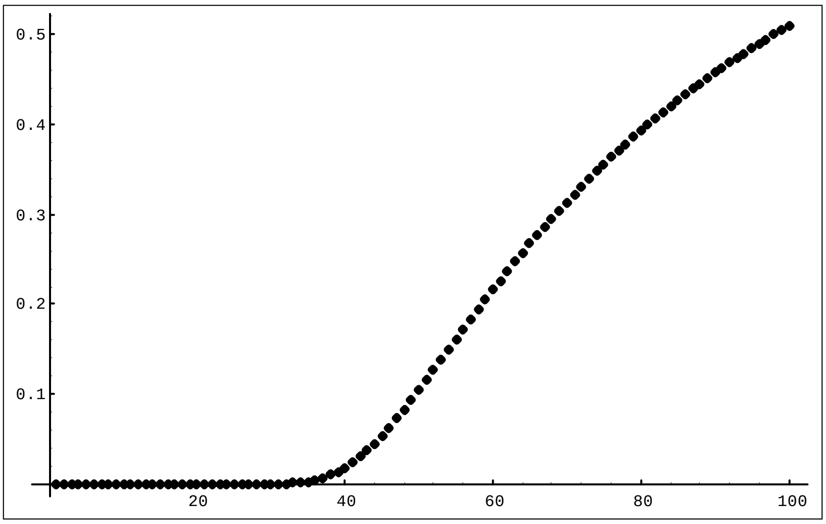
$$p_0 = \frac{1}{\sum_{k=0}^{n} \frac{\rho^k}{k!}}$$

Probabilitatea cerută în problemă este

$$p = p_n = \frac{\frac{\rho^n}{n!}}{\sum_{k=0}^{n} \frac{\rho^k}{k!}}$$

unde trebuie luat $n = 50$, $\rho = 50$. Cu calculatorul se obţine $p = 0.104787$.

Pe măsură ce solicitarea centralei creşte, adică ρ creşte, devine din ce în ce mai mare probabilitatea ca un apel să fie anulat. Mai jos apare graficul acestei probabilităţi în funcţie de gradul de încărcare ρ al centralei.



Creşterea probabilităţii p (verticală) de refuz a unui apel
cu creşterea factorului ρ (orizontală) de încărcare al centralei

4. O particulă se mişcă între doi pereţi doar prin poziţiile $x_1 < x_2 < x_3 < x_4$ astfel: dacă este într-o poziţie interioară atunci cu probabilitatea $\frac{3}{4}$ face un salt în poziţia din faţă şi cu probabilitatea $\frac{1}{4}$ în poziţia imediat din spate; dacă se găseşte în prima sau în ultima poziţie, rămâne definitiv acolo.

Să se descrie comportarea particulei după un număr mare de salturi.

Indicaţie. Matricea probabilităţilor de trecere din o poziţie în alta este:

$$p = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/4 & 0 & 3/4 & 0 \\ 0 & 1/4 & 0 & 3/4 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Cu calculatorul găsim de exemplu

$$p^{100} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0.30769 & 4.4679 \cdot 10^{-37} & 0 & 0,692308 \\ 0,07692 & 0 & 4,4679 \cdot 10^{-37} & 0,923077 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Aspectul matricei p^{100} pune în evidenţă că după un număr mare de salturi particula este captată de un perete sau altul, probabilităţile de a rămâne în poziţii intermediare fiind foarte mici. Care este valoarea exactă a matricei p^{∞}?

5. Fie $(\eta_i)_{i \in Z}$ o familie de variabile aleatoare independente, care au toate repartiţia

$$\begin{pmatrix} 1 & -1 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

Fie $f : R \to R$ *prin* $f(t) = \begin{cases} 1, \ t \in (-\frac{b}{2}, \frac{b}{2}] \\ 0, \text{ în caz contrar} \end{cases}$. *Definim acum procesul aleator* $(\xi_t)_{t \in R}$ *prin* $\xi_t = \sum_{i \in Z} f(t - i \cdot b)\, \eta_i$. *Să se calculeze funcţia de autocorelaţie* $R(t_1, t_2)$. *Este procesul staţionar în sens larg?*

Soluţie. Pe intervalul $(-\frac{b}{2}, \frac{b}{2}]$, $\xi_t = \eta_0$; pe intervalul $(\frac{b}{2}, \frac{3b}{2}]$, $\xi_t = \eta_1$; ... pe intervalul $I_i = (ib - \frac{b}{2}, ib + \frac{b}{2}]$, $\xi_t = \eta_i$. Prin urmare $\xi_{t_1}\xi_{t_2} = \eta_i\eta_j$ dacă $t_1 \in I_i$ şi $t_2 \in I_j$. Deoarece variabilele η_i sunt independente rezultă că $M\left(\xi_{t_1}\xi_{t_2}\right) = M\left(\eta_i\eta_j\right)$ este diferită de zero doar dacă $i = j$, adică $t_1, t_2 \in I_i$, caz în care avem $R(t_1, t_2) = M(\eta_i^2) = 1$. Funcţia de autocorelaţie nu este invariantă la o translaţie în timp deci procesul nu este staţionar în sens larg.

6. Fie procesul aleator $\xi_t = \sum_{k=1}^{n} a_k (\phi_k \cos \lambda_k t + \eta_k \sin \lambda_k t)$ *unde* a_k *şi* $\lambda_k \in R$, *iar* ϕ_k *şi* η_k *sunt variabile aleatoare independente, de medie 0 şi dispersie 1, pentru k=1,2,...n. Să se arate că procesul este staţionar în sens larg şi să se determine autocorelaţia procesului.*

Răspuns. $R(\tau) = \sum a_k^2 \cos \lambda_k \tau$.

Partea II

Statistică

Lecţia 8

Statistica descriptivă

In cele ce urmează vom încerca să explicăm ce este statistica, cum diferă ea de teoria probabilităţilor, ce o leagă de aceasta, care sunt părtile ei componente si cum începe demersul practic într-o problemă de statistică (adică vom spune câteva cuvinte despre statistica descriptivă).

Atunci când omul nu a mai putut *intui* a început să *măsoare.* Măsurătorile şi observaţiile au devenit prima treaptă spre întelegerea legilor *naturii.* Dar, în acest fel, omul nu mai poate să cunoască direct *realitatea*, el poate numai să o *aproximeze* succesiv prin *modele fizice* si apoi prin *modele matematice.* Dar aceste modele nu descriu *exact* Realitatea. Ele o aproximează şi apar asa numitele erori. Unele erori sunt previzibile, altele însă sunt întâmplătoare (aleatoare). Aceste ultime erori (aleatoare) au si ele legile lor de manifestare. Apar deci fenomenele aleatoare descrise prin variabilele aleatoare. Teoria probabilitătilor pleacă de la ipoteza că se cunosc exact aceste variabile aleatoare (prin functiile de probabilitate, prin functiile de repartitie, prin functiile caracteristice, etc.). *Statistica* pleacă de la măsuratorile brute si caută să regăsească modelul probabilistic teoretic exact care se află în spatele acestor măsuratori. Partea "empirică" a statisticii care se ocupă de prelucrarea datelor obtinute prin măsuratori sau observatii se numeste *statistică descriptivă.* Aparatul matematic al teoriei probabilitătilor, pus în functiune pentru a studia si interpreta aceste date, în dorinta de a recupera modelul probabilistic real, care guvernează fenomenul măsurat sau observat, formează *inferenta statistică.* După ce cercetătorul capătă informatii suficient de clare despre fenomenul probabilistic studiat, el va trebui să *actioneze* optim potrivit acestor informatii. Apare deci *teoria deciziei statistice,* care este o ramură importantă a statisticii.

8.1 Statistica unei variabile

Multimea de obiecte studiată se numeste *populatie.* Un obiect separat dintr-o populatie dată se numeste *individ* sau *membru al populatiei.* Trăsătura comună a tuturor membrilor populatiei care ne interesează în studiul nostru se numeste *caracteristică.* Caracteristicile pot fi *cantitative* (înaltime, greutate, notă la examen, abscisa unui punct în plan, etc...) sau *calitative* (culoarea ochilor, sex, loc de nastere, etc...). Oricum statistica lucrează cu numere, caracteristicilor calitative li se atasază *coduri numerice.*

Exemplul 8.1 *Ne interesează statistica ploilor în Bucuresti pe anul 1995, zilnic. Aici populatia este multimea zilelor din anul 1995, un individ al populatiei este o zi anume din acest an, de exemplu 3 ianuarie, iar caracteristica calitativă este faptul că a plouat sau nu în acea zi. Dacă a plouat punem 1 si dacă nu, putem 0. Numerele 1 si 0 reprezintă coduri în statistica respectivă.*

Presupunem în continuare că avem numai caracteristici cantitative ale unor populatii, mai exact avem multimi brute de numere reale, sau tabele de numere reale. Privim aceste numere atasate unei populatii ca fiind valori ale unei variabile aleatoare X. Vom spune pe scur: "*fie populatia X*".

Exemplul 8.2 *O masină produce piese cilindrice fine, cu diametru standard fixat ϕ= 3 cm. Fiecare piesă are o abatere de la acest diametru, măsurată în microni. Aceste abateri formează o "populatie" în sensul de mai sus, mai bine zis valorile unei variabile aleatoare X. Noi nu putem să precizăm de la început ce abatere va avea o piesă luată la întâmplare, dar putem face o selectie de n piese si putem măsura abaterile lor: $x_1, x_2, ..., x_n$. Fiecare x_i reprezintă o valoare a v.a. X care, teoretic vorbind, are o densitate de probabilitate $\rho(x)$ si o functie de repartitie $F(x)$.*

Definiţia 8.3 *O mulţime de n observaţii independente asupra unei caracteristici numerice X a unei populaţii P, care ne dă n valori $x_1, x_2, ...x_n$, se numeşte selecţie de volum n. Şirul de valori $(x_i)_{1\leq i\leq n}$ îl vom numi serie statistică discretă.*

In exemplul de mai sus facem o *selectie de volum n* din multimea pieselor si construim asa numita *functie de repartitie empirică* $F_n^*(x)$.

Definiţia 8.4 *Se numeşte funcţie de repartiţie empirică asociată unei variabile aleatoare X şi unei selecţii $\{x_1, x_2, ..., x_n\}$,funcţia $F_n^* : R \to R$*

$$F_n^*(x) = \frac{nr.\ de\ valori\ x_j < x}{n} = \frac{k_x}{n}$$

Teorema de mai jos pune în evidenţă că funcţiile de repartiţie empirice aproximează oricât de bine funcţia reală de repartiţie.

Teorema 8.5 *Fie P o populatie statistică si X variabila aleatoare atasată ei cu functia de repartitie $F(x)$. Pentru o selectie de volum n: $\{x_1, x_2, ..., x_n\}$ construim ca mai sus functia de repartitie empirică $F_n^*(x)$. Atunci*

$$\Pr ob\{\mid F(x) - F_n^*(x) \mid\geq \epsilon\} \to 0$$

când $n \to \infty$, pentru orice $\epsilon > 0$, fixat. Altfel spus $F_n^*(x) \to F(x)$ în probabilitate.

Demonstratie Să notăm cu p =Prob$\{X < x\} = F(x)$, si cu $F_n^*(x) = \frac{k_x}{n}$ (vezi definitia de mai sus). Notăm cu $\eta_1, ..., \eta_n$ v.a. construite astfel: η_j are valoarea 1 dacă $x_j < x$ si 0 în caz contrar. Variabilele $\eta_1, ...\eta_n$ sunt independente (ca valori ale unor observaţii independente) şi au distribuţia

$$\begin{pmatrix} 1 & 0 \\ \text{p} & \text{1-p} \end{pmatrix}$$

Avem $M(\eta_i) = p,\ D(\eta_i) = p(1-p)$. Este clar că v.a. $Y_n = \frac{\eta_1 + \cdots + \eta_n}{n}$ are media p şi dispersia

$$D(Y_n) = \frac{1}{n^2}(D(\eta_1) + ... + D(\eta_n)) = \frac{p(1-p)}{n}$$

(a se vedea proprietăţile mediei şi dispersiei, Lecţia 2). Aplicăm acum inegalitatea lui Cebîsev lui Y_n si găsim că

$$\begin{aligned} & \text{Prob}\,(\mid F_n^*(x) - F(x) \mid \geq \epsilon) \\ = \quad & \text{Prob}\,(|Y_n - p| \geq \epsilon) \leq \frac{D(Y_n)}{\epsilon^2} = \frac{p^2(1-p)^2}{n\epsilon^2} \end{aligned}$$

Cum partea dreaptă tinde la 0 când $n \to \infty$ rezultă că $F_n^*(x) \to F(x)$ în probabilitate, când $n \to \infty$.

QED.

In urma oricărei selectii de volum n dintr-o populatie de numere se obtine un sir finit de n numere numit *serie statistică* (de volum n). Cum construim o densitate de probabilitate empirică? Pentru a răspunde la această întrebare grupăm termenii unei serii statistice în intervale disjuncte: $I_1, I_2, ..., I_k$, după criterii mai mult sau mai putin subiective. Asociem fiecărui interval I_j mijlocul lui, M_j. Punctului M_j îi asociem *frecventa relativă* a v.a. empirice pe intervalul I_j, adică câtul dintre numărul n_j al acelor x_i care se află în I_j si n (volumul întregii selectii): n_j/n. Este clar că în felul acesta obtinem o v.a. $\widetilde{X}_n \begin{pmatrix} \widetilde{x}_j \\ n_j/n \end{pmatrix}$ $j = 1, 2, ..., k$, unde $\widetilde{x}_j$ este abscisa punctului M_j. Graficul functiei de probabilitate al v.a. $\widetilde{X}_n$ se numeste *histogramă* asociată selectiei $x_1, ..., x_n$ si împărtirii în intervale $I_1, ..., I_k$. Dacă unim printr-o linie poligonală punctele de coordonate $(\widetilde{x}_j, n_j/n)$ obtinem un *poligon al frecventelor* ce aproximează de fapt graficul functiei densitate de probabilitate al v.a. X pe un interval finit $(I_1 \cup I_2 \cup \cdots \cup I_k)$ care contine numerele $x_1, ..., x_n$.

Pentru o selectie dată $\{x_1, ..., x_n\}$ se introduc diferiti *indicatori empirici* care dau anumite informatii despre întreaga populatie.

Definiţia 8.6 *Fie* $\{x_1, x_2, ..x_n\}$ *o selecţie de volum n.*

i) $m^* = \frac{x_1 + x_2 + ..x_n}{n}$ *se numeşte media empirică.*

ii) $m_r^* = \frac{x_1^k + x_2^k + .. + x_n^k}{n}$ *se numeşte momentul empiric de ordinul r.*

iii) $\mu_k^* = \frac{(x_1 - m^*)^k + (x_2 - m^*)^k + ... + (x_n - m^*)^k}{n}$ *se numeşte momentul empiric centrat de ordin k.*

iv) $S^{*2}=D^{*}=\sigma^{*2}=\frac{(x_1-m^*)^2+(x_2-m^*)^2+...+(x_n-m^*)^2}{n}$ *se numeşte dispersia empirică sau varianţa empirică.* σ^* *se numeşte deviaţia standard.*

v) $S'^{*2}=\frac{(x_1-m^*)^2+(x_2-m^*)^2+...+(x_n-m^*)^2}{n-1}$ *se numeşte dispersia empirică modificată.*

vi) Valoarea $\alpha \in R$ *astfel ca numărul de valori* $x_i \leq \alpha$ *este egal cu numărul de valori* $x_i \geq \alpha$*, se numeşte mediană. Dacă există mai multe asemenea valori pentru* α*, atunci ele formează un interval şi mediana este prin definiţie mijlocul acestui interval.*

vii) Valoarea x_i *cu frecvenţa maximă de apariţie se numeşte modul selecţiei. (este posibil să nu fie unic)*

viii) Se numeşte prima cvartilă a selecţiei, cel mai mic x *astfel ca numărul de valori* $x_j \leq x$ *să fie* $\geq \frac{1}{4}n$ *. A treia cvartilă este cea mai mică valoare* x_i *astfel ca numărul de valori* $x_j \leq x_i$ *să fie* $\geq \frac{3}{4}n$*. Analog se defineşte a p-a cuantilă de ordin q ca cea mai mică valoare* x_i *astfel ca numărul de valori* $x_j \leq x_i$ *să fie* $\geq \frac{p}{q}n$*.*

Observaţia 8.7 •*In cazul când datele sunt grupate pe intervale, definiţiile de mai sus se referă la mijloacele intervalelor, fiecare mijloc fiind considerat de atâtea ori câte valori se află în el.*

•*In general dacă o valoare* x_i *se repetă atunci vom nota cu* n_i *numărul de apariţii, şi cu* $f_i=\frac{n_i}{n}$ *frecvenţa relativă. Formulele de mai sus pot fi scrise* $m^*=\frac{\sum n_i x_i}{n}=\sum f_i x_i$*,* $\sigma^{*2}=\sum f_i\left(x_i-m^*\right)^2$*, etc. Insumarea se face acum numai după valorile* x_i *distincte. Seria statistică o vom nota în acest caz* $(x_i,n_i)_{1\leq i\leq p}$*, punând în evidenţă de câte ori apare fiecare valoare.*

• *Media şi mediana descriu "centrul" valorilor de selecţie iar dispersia este o măsură a împrăştierii acestor valori în jurul "centrului". Modul indică în ce zonă sunt cele mai probabile valori. Cuantilele indică în ce zone se află un anumit procent de valori.*

Propoziţia 8.8 *Următoarele formule au loc:*

$$S^{*2}=\frac{n\sum x_i^2\ -\ \left(\sum x_i\right)^2}{n^2} \tag{8.1}$$

$$S^{*\prime\,2}=\frac{n\sum x_i^2\ -\ \left(\sum x_i\right)^2}{n(n-1)} \tag{8.2}$$

$$S^{*\prime\,2}=\frac{\sum x_i^2}{n-1}-\left(\frac{n}{n-1}\right)m^{*2} \tag{8.3}$$

Demonstraţie. Sunt calcule simple lăsate ca exerciţiu.

Observaţia 8.9 *In general în calcule nu utilizăm notaţiile* m^*, D^**, etc. ci m, D, ... Am introdus aici notaţiile* m^*, D^**,.. pentru a le distinge de m=media teoretică, D=dispersia teoretică, etc., care se vor introduce în lecţia următoare.*

Exemplul 8.10 *O firmă este interesată de timpul mediu al convorbirilor telefonice şi de distributia acestor timpi fată de timpul mediu (dispersia) pe durata a 40 convorbiri telefonice consecutive. Timpii s-au rotunjit în minute şi rezultatul sondajului a dat următorii timpi: 4, 6, 4, 4, 7, 2, 3, 1, 2, 1, 1, 4, 9, 8, 11, 12, 3, 2, 1, 1, 3, 9, 4, 5, 7, 7, 9, 10, 10, 1, 2, 2, 3, 11, 12, 10, 1, 1, 3, 4. Să se facă şi o histogramă a frecventelor relative şi un grafic al functiei de repartitie pentru acest sondaj.*

Solutie Facem mai întâi următorul tabel :

timpi de convorbire t_i	*numărul* convorbirilor n_i	*frecv. relativă* $f_i = n_i/n$	*frecv. cumulată* $F_{(t_i)}$
1 min	8	8/40	8/40
2 min	5	5/40	$\frac{8}{40} + \frac{5}{40} = \frac{13}{40}$
3 min	5	5/40	$\frac{8}{40} + \frac{5}{40} + \frac{5}{40} = \frac{18}{40}$
4 min	6	6/40	24/40
5 min	1	1/40	25/40
6 min	1	1/40	26/40
7 min	3	3/40	29/40
8 min	1	1/40	30/40
9 min	3	3/40	33/40
10 min	3	3/40	36/40
11 min	2	2/40	38/40
12 min	2	2/40	$40/40 = 1$

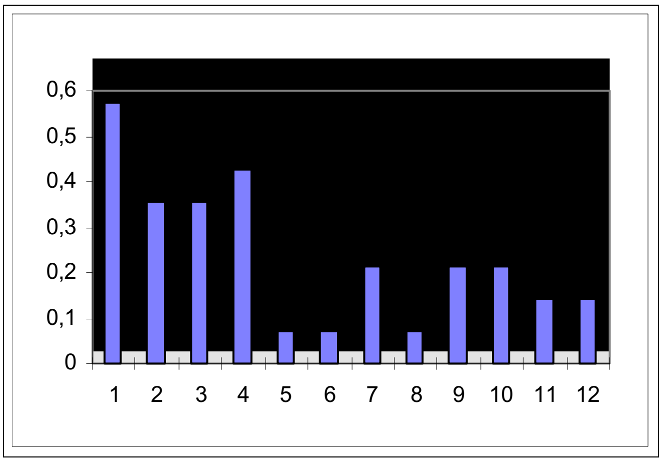
• *media* convorbirilor este

$$\begin{aligned} m^* &= \sum t_i \cdot n_i/n = \frac{1}{n}\sum t_i \cdot f_i \\ &= \frac{1}{40}(1 \cdot 8 + 2 \cdot 5 + 3 \cdot 5 + 4 \cdot 6 + 5 \cdot 1 + +6 \cdot 1 \\ &\quad +7 \cdot 3 + 8 \cdot 1 + 9 \cdot 3 + 10 \cdot 3 + 11 \cdot 2 + 12 \cdot 2) \\ &= 5 \end{aligned}$$

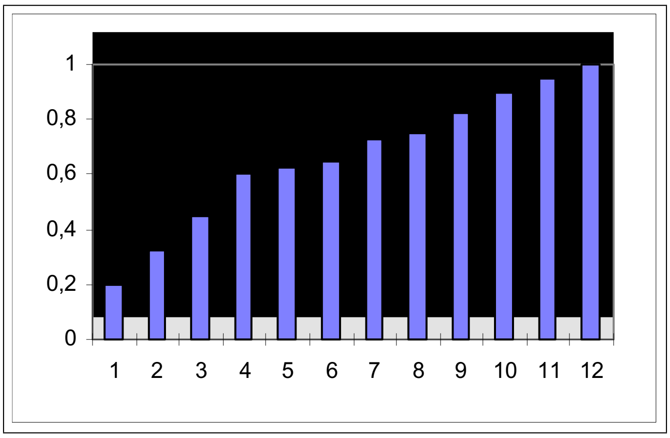
• *dispersia* empirică este

$$\begin{aligned} & \sum (t_i - m^*)^2 \cdot n_i/n \\ = & \left(\frac{1}{40}(1 - m^*)^2 \cdot 8 + (2 - m^*)^2 \cdot 5 + (3 - m^*)^2 \cdot 5 + \cdots + (12 - m^*)^2 \cdot 2\right) \\ = & \ 13,179 \end{aligned}$$

• *mediana* este 4, prima cvartilă este 2, a treia cvartilă este 8,25, modul este 1.
• *histograma frecvenţelor* este



- *histograma frecvenţelor cumulate* este:



Exemplul 8.11 *Se dă o selectie de 150 de numere* $\{x_1, x_2, ..., x_{150}\}$ *cu media de selectie m=102, 42. Aceste numere se grupează în 8 intervale* $[81,5; 87,5)$, $[87,5; 93,5)$, ..., $[123,5; 129,5)$, *de lungime 6 unităti. Ele se repartizează în aceste intervale după cum urmează: în primul interval avem 2 numere* (n_1*=2), în al doilea 23 de numere* (n_2*=23),* f_3*=22,* n_4*=65,* n_5*=20,* n_6*=10,* n_7*=0,* n_8*=8.a) Să se calculeze media selecţiei. b) Să se calculeze dispersia selectiei.*

Solutie a) este lăsat ca exerciţiu; se găseşte $m^* = 102,42$.
b) Se face următorul tabel de calcule:

x_j (*mijlocul int.*)	n_j	$x_j - m^*$	$(x_j - m^*)^2$	$(x_j - m^*)^2 n_j$
$84,5$	2	$-17,92$	$321,1264$	$642,2528$
$90,5$	23	$-11,92$	$142,0864$	$3267,9872$
$96,5$	22	$-5,92$	$35,0464$	$771,0208$
$102,5$	65	$0,08$	$0,0064$	$0,4160$
$108,5$	20	$6,08$	$36,9664$	$739,3280$
$114,5$	10	$12,08$	$145,9264$	$1459,2640$
$120,5$	0	$18,08$	$326,8864$	$0,0006$
$126,5$	8	$24,08$	$579,8464$	$4638,7712$
				$11519,0400$

Găsim $S^{*2} = \frac{\sum(x_j-m)^2 n_j}{n} = \frac{11519,04}{150} = 76,79$. Pentru verificare putem folosi formula $S^{*2} = \frac{\sum x_j^2 n_j}{n} - m^{*2}$ care este mai comodă, dar cere o coloană separată cu calculul lui x_j^2.

8.2 Statistica a două variabile

Să presupunem că avem două caracteristici numerice care se urmăresc, de exemplu înălţimea şi greutatea. Prin testare se găseşte următoarea situaţie: x_i sunt greutăţile, y_j sunt înălţimile observate (grupate pe intervale), iar la întretăierea coloanei i cu linia j se află numărul de cazuri observate, $n_{i,j}$.

$\frac{x_i \to}{y_j \downarrow}$	43	48	53	58	$n_{.,j}$
152	20	8	2	0	30
157	2	18	1	4	25
162	0	1	10	4	15
167	0	1	4	15	20
$n_{i,.}$	22	28	16	23	N=80

Notăm o asemenea serie de observaţii prin $(x_i, y_j, n_{i,j})_{\substack{1\le i\le p\\1\le j\le q}}$. Avem de exemplu la x_2=48 şi y_1=152 un număr de $n_{2,1}$ =8 cazuri înregistrate.

Se definesc următoarele mărimi:

i) $n_{i,.} = \sum_j n_{i,j}$, $n_{.,j} = \sum_i n_{i,j}$, N=$\sum_{i,j} n_{i,j}$. Seria $(x_i, n_{i,.})$ se numeşte seria marginală în x, iar seria $(y_j, n_{.,j})$ se numeşte seria marginală în y. $f_{i,.} = \frac{n_{i,.}}{N}$ şi $f_{.,j} = \frac{n_{.,j}}{N}$ se numesc frecvenţe marginale, iar $f_{i,j} = \frac{n_{i,j}}{N}$ se numeşte frecvenţa dublă.

ii) $m_x^* = \frac{\sum_{i,j} n_{i,j} x_i}{N} = \frac{\sum_i n_{i,.} x_i}{N}$ şi $m_y^* = \frac{\sum_{i,j} n_{i,j} y_j}{N} = \frac{\sum_j n_{.,j}\, y_j}{N}$ se numesc medii marginale.

ii)
$$\sigma_x^{*2} = \frac{\sum_{i,j} n_{i,j}(x_i - m_x^*)^2}{N} = \frac{\sum_i n_{i,.}\,(x_i - m_x^*)^2}{N} = \frac{\sum_i n_{i,.} x_i^2}{N} - m_x^{*2}$$

şi

$$\sigma_y^{*2} = \frac{\sum_{i,j} n_{i,j}(y_j - m_y^*)^2}{N} = \frac{\sum_j n_{.,j}\left(y_j - m_y^*\right)^2}{N} = \frac{\sum_j n_{.,j} y_j^2}{N} - m_y^{*2}$$

se numesc dispersii (varianţe) marginale.

iv) Covarianţa seriei este numărul

$$cov(x,y) = \frac{\sum_{i,j} n_{i,j}\left(x_i - m_x^*\right)\left(y_j - m_y^*\right)}{N} = \frac{\sum_{i,j} n_{i,j} x_i y_j}{N} - m_x^* m_y^*.$$

v) Coeficienrtul de corelaţie liniară al seriei este $\rho_{x,y} = \frac{cov(x,y)}{\sigma_x^* \sigma_y^*}$.

In cazul de mai sus găsim

$$m_x^* = \frac{22 \cdot 43 + 28 \cdot 48 + 16 \cdot 53 + 23 \cdot 58}{80} = 49,438;$$

$$m_y^* = \frac{30 \cdot 152 + 25 \cdot 157 + 15 \cdot 162 + 20 \cdot 167}{80} = 157,313;$$

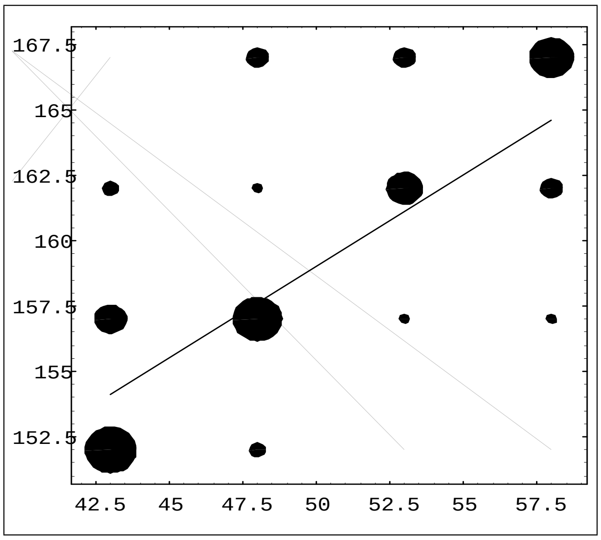
$$\sigma_x^{*2} = \frac{\begin{array}{c} 22 \cdot (43 - 49,438)^2 + 28 \cdot (48 - 49,438)^2 + \\ +16 \cdot (53 - 49,438)^2 + 23 \cdot (58 - 49,438)^2 \end{array}}{80} = 28,246;$$

$$\sigma_y^{*2} \frac{\begin{array}{c} 30 \cdot (152 - 157,313)^2 + 25 \cdot (157 - 157,313)^2 + \\ +15 \cdot (162 - 157,313)^2 + 20 \cdot (167 - 157,313)^2 \end{array}}{80} = 26,465;$$

$$cov(x,y) = \frac{\begin{array}{c} 20 \cdot (43 - 49,438) \cdot (152 - 157,313) + \\ +8 \cdot (48 - 49,38) \cdot (152 - 157,313) + \\ ... + 15 \cdot (58 - 49,438) \cdot (167 - 157,313) \end{array}}{80} = 18,926;$$

$$\rho_{x,y} = \frac{18,926}{\sqrt{28,246 \cdot 26,465}} = 0,692.$$

Reprezentarea grafică a datelor se face prin discuri pline: în punctul (x_i, y_j) se pune un disc cu aria proporţională cu numărul de observaţii care au dat greutatea x_i şi înălţimea y_j. Se obţine histograma:



(dreapta nu face parte din histogramă; vezi în continuare)

Două selectii de acelasi volum n din două populatii diferite, $\{x_1, ..., x_n\}$ şi $\{y_1, ..., y_n\}$ se zic corelate prin functia $y = f(x)$ dacă $y_k = f(x_k)$, pentru $k = 1, 2, ..., n$.

Dacă $f(x) = ax + b$, corelatia se zice *liniară*. Am văzut în lecţia 6 că $\rho_{x,y} = \pm 1$ este echivalent cu faptul că punctele (x_i, y_j) sunt de-a lungul unei drepte. Dacă $\left|\rho_{x,y}\right|$ este apropiat de 1 atunci datele (x_i, y_j) sunt aproximativ pe o dreaptă $y = ax + b$. Reluăm aici, în varianta folosită în aplicaţii acest lucru.

Teorema 8.12 *Fie $(x_i, y_j, n_{i,j})_{\substack{1\le i\le p\\ 1\le j\le q}}$ o selecţie dublă. Atunci:*

a) $-1 \le \rho_{x,y} \le 1$ şi semnul egal apare dacă şi numai dacă punctele (x_i, y_j), pentru $n_{i,j} \ne 0$, sunt coliniare.

b) Ecuaţia dreptei $y = ax+b$, unde coeficienţii a, b sunt determinaţi de condiţia ca expresia

$$\Phi(a,b) = \sum_{(x_i,y_j)=observat} (y_j - ax_i - b)^2 = \sum_{i,j} n_{i,j}\,(y_j - ax_i - b)^2$$

să fie minimă, este:

$$y = \frac{cov(x,y)}{\sigma_x^{*2}}\,(x - m_x^*) + m_y^* \tag{8.4}$$

Această dreaptă se numeşte dreapta de regresie a lui y în x.

Demonstraţie.

a) Deoarece $n_{i,j} \ge 0$, atunci expresia $E = \sum_{i,j} n_{i,j}\left(t(y_j - m_y^*) + (x_i - m_x^*)\right)^2$ este pozitivă pentru orice $t \in R$. Ridicând la pătrat, găsim: $t^2\sigma_y^{*2} + 2cov\,(x,y)\,t + \sigma_x^{*2} \ge 0$ pentru orice $t \in R$. Prin urmare $\Delta = 4cov^2\,(x,y) - 4\sigma_y^{*2}\sigma_x^{*2} \le 0$. Prin împărţire cu $\sigma_x^{*2}\sigma_y^{*2}$ găsim $\rho_{x,y}^2 \le 1$, adică $-1 \le \rho_{x,y} \le 1$. Dacă $\rho = \pm 1$ atunci $\Delta = 0$ deci există t_0 astfel ca $E = 0$, deci fiecare paranteză este egală cu 0, deci pentru orice i, j, pentru care $n_{i,j} \neq 0$ avem $t_0 \left(y_j - m_y^*\right) + x_i - m_x^* = 0$, deci punctele (x_i, y_j) cu $n_{i,j} \neq 0$ sunt coliniare.

b) $\frac{\partial \Phi(a,b)}{\partial a} = 0$, $\frac{\partial \Phi(a,b)}{\partial b} = 0$ formează un sistem liniar în a şi b cu soluţiile $a = \frac{cov(x,y)}{\sigma_x^{*2}}$, $b = m_y^* - am_x^*$.

QED.

Analog se determină dreapta de regresie a lui x în y. Cele două drepte sunt distincte. Ele coincid doar dacă datele (x_i, y_j) sunt coliniare. In cazul de mai sus găsim $y = 0,67x + 124,188$, dreaptă care este reprezentată pe histograma datelor.

Dacă $f(x) = ax^2 + bx + c$, corelatia se zice *parabolică*. Coeficienţii se determină din condiţia ca $\Phi(a, b, c) = \sum_{i,j} n_{i,j} \left(y_j - ax_i^2 - bx_i - c\right)^2$ să fie minimă.

Dacă $f(x) = ae^{bx}$, corelatia se zice *exponentială* şi se reduce prin logaritmare tot la o corelaţie liniară: $\ln(f(x)) = \ln(a) + bx$. Coeficienţii $\alpha = \ln(a)$ şi b se determină din condiţia ca $\Phi(\alpha, b) = \sum_{i,j} n_{i,j} \left(\ln(y_j) - \alpha - bx_i\right)^2$ să fie minimă.

Dacă $f(x) = ax^b$, atunci avem prin logaritmare $\ln(f(x)) = \ln(a) + b\ln(x)$, şi, la fel ca mai sus se deduc coeficienţii $\alpha = \ln(a)$ şi b din condiţia $\Phi(\alpha, b) = \sum_{i,j} n_{i,j} \left(\ln(y_j) - \alpha - b\ln(x_i)\right)^2$ să fie minimă.

In multe situaţii pentru fiecare x_i avem doar o valoare y pe care o notăm y_i, deci valorile (x_i, y_i) sunt pe graficul unei funcţii. Determinarea unei funcţii care ajustează datele respective *prin metoda celor mai mici pătrate* constă în propunerea unui model de funcţie, $f(x, a, b, ..)$, şi determinarea parametrilor $a, b, ..$ din condiţia $\Phi(a, b, ..) = \sum_i \left(y_i - f(x_i, a, b, ..)\right)^2$ să fie minimă.

8.3 Exerciţii

1. S-a făcut un sondaj preelectoral pe un esantion de 100 persoane. Am notat cu A, B, C, D, E candidatii, cu F răspunsul "nedecis" şi cu G răspunsul "nu intentionez să votez". Să se construiască o histogramă cu functia de distributie şi altă histogramă cu functia de repartitie (frecventa cumulată) pentru acest sondaj dacă răspunsurile sunt date în următorul tabel: C, A, A, B, E, F, F, C, C, C, A, B, A, A, A, E, F, A, B, G, D, B, B, C, F, G, G, D, D, D, B, A, B, B, B, F, G, B, C, A, E, C, C, D, G, A, A, E, E, E, C, D, D, E, G, G, A, B, B, A, F, F, G, G, G, G, A, A, A, B, B, C, C, A, A, D, D, E, F, G, A, B, C, C, D, A, E, F, A, B, F, G, A, B, C, D, A, A, B, E.

Solutie Aici trebuie mai întâi să codificăm numeric literele (optiunile electoratului) A, B, C, D, F, G. De exremplu, propunem următoarea codificare:

G$\longleftrightarrow$0
F$\longleftrightarrow$1
A$\longleftrightarrow$4
B$\longleftrightarrow$5
C$\longleftrightarrow$6

D$\longleftrightarrow$7

2. Două grupe de 10 studenti A şi B au obtinut următoarele note la examenul de statistică:

A : 8, 5, 6, 6, 7, 9, 4, 3, 5, 6

B : 9, 6, 7, 8, 6, 10, 5, 4, 6 ,7.

Să se găsească cea mai bună corelatie liniară între cele două selectii. Să se găsească valoarea deviatiei pătratice. Să se facă acelasi lucru pentru o corelatie de tip parabolic şi să se compare deviatiile pătratice.

3. Fie selectia {0, 1, -1, -1, -2, 1, 1, -1, 2, 3, 1, 4, 3, -1, 0, 0, 3, -1, -2, -2} dintr-o populatie anume. Fie X v.a. care guvernează populatia. Să se aproximeze cu ajutorul selectiei numărul $P(0 \leq X \leq 2)$. *Se cere graficul functiei de repartitie pentru această selectie şi o histogramă a frecventelor.*

4. S-a făcut un sondaj asupra pretului (în centi) galonului de benzină premium asupra a 30 statii luate la întâmplare. De aici a rezultat selectia: 65, 58, 64, 68 , 52, 48, 59, 59, 56, 63, 61, 66, 52, 57, 60, 62, 55, 55, 64, 71, 61, 63, 46, 53, 60, 57, 58, 57, 54, 58. Se cere graficul poligonului de frecventă (relativă) dacă: a) grupăm datele în intervale de lungime 3, cu 60 centrul unui asemenea interval; b) grupăm datele în intervale de lungime 5, cu 60 ca centru al unui asemenea interval. Calculati pentru aceste grupări media si dispersia de selectie. Găsiti graficul functiilor de repartitie empirice.

5. La un concurs 12 studenti au obtinut următorele punctaje: 18, 15, 19, 27, 13, 30, 24, 11, 5, 16, 17, 20. Calculati media, mediana, deviatia standard si deviatia absolută medie. Construiti functia empirică de frecventă cumulată (f. de repartitie) si interpretati rezultatele obtinute.

Lecţia 9

Statistici. Estimarea parametrilor

Amintim că o populatie P este o multime de obiecte din care se fac selectii finite (de volum $n < \infty$). Populatia se poate identifica cu multimea tuturor observatiilor potentiale pe care le putem face asupra obiectelor ei. Pentru fiecare obiect al selecţiei se testează valoarea unei caracteristici numerice, X. Admitem că pe P există o probabilitate şi că X este o variabilă aleatoare. Distributia (functia de repartitie) a *v.a.* X se numeste *distributia populatiei după caracteristica X.*

Exemplul 9.1 *Intr-o magazie grâul este amestecat cu neghină. Populaţia P este aici totalitatea boabelor din magazie (câteva sute de milioane). Fie $X : P \to R$,*

$$X(bob) = \begin{cases} 1 \text{ } daca \text{ } e \text{ } grau \\ 0 \text{ } daca \text{ } e \text{ } neghina \end{cases}$$

Probabilitatea p ca un bob să fie de tip A este definită prin:

$$p(A) = \frac{nr.\ de\ boabe\ de\ tipul\ A}{nr.\ de\ boabe\ din\ magazie}, \quad A \subset P$$

Aici p nu se poate determina experimental exact din cauza numărului mare de boabe, dar teoretic p există. Valoarea medie a lui X înmulţită cu 100 este procentul de boabe de grâu din magazie, lucru important.

Exemplul 9.2 *Să presupunem că mai multe persoane, sau aceeaşi persoană în mai multe rânduri, măsoară independent o lungime, de aproximativ 1 km folosind o ruletă de 2 m. Evident că se vor obţine rezultate diferite datorită unei game largi de cauze incontrolabile. Putem în acest caz considera P ca mulţimea tuturor complexelor de cauze necontrolabile care influenţează rezultatul măsurătorii sau putem considera P ca mulţimea tuturor măsurătorilor posibile. Oricum P nu este o mulţime pe care o putem explicita ca în cazul precedent. Admitem însă că pe P există o probabilitate iar o măsurătoare înseamnă o manifestare a unui complex ω de cauze necontrolabile care conduc la un rezultat $X(\omega)$, în cazul nostru X fiind o lungime.*

Prin urmare caracteristica lungime apare ca o funcţie $X : P \to R$. Admitem că X este o variabilă aleatoare, adică (vezi lecţiile 2, 3) $\{\omega \in P | \, X(\omega) < r\} \subset P$ este o mulţime pe care este definită probabilitatea p.

Statistica Matematică se ocupă, printre altele, cu problema determinării repartiţiei unei variabile aleatoare X ca în exemplele de mai sus, prin experimente. In general n experimente conduc la n valori numerice $x_1, ...x_n$. Ce operaţii trebuie făcute cu valorile x_1, $...x_n$ pentru a găsi caracteristici ale lui X şi ce încredere putem avea în rezultatele obţinute?

In continuare prezentăm felul în care putem considera rezultatele x_1, $x_2, .. x_n$ ale lui X în n experienţe independente ca valori a n variabile aleatoare independente X_1, $X_2, ... X_n$. La o primă lectură se poate sări peste această parte, remarcându-se doar concluziile.

Fie P ca mai înainte spaţiul probabilizat al cauzelor incontrolabile, fie $\Omega \subset P(P)$ $\sigma-$algebra submulţimilor lui P pentru care e definită σ probabilitatea p. Notăm P^∞ şirurile de elemente din P. Deci $\omega \in P^\infty$ dacă şi numai dacă $\omega = (\omega_k)_{k\in N}$ şi pentru orice k, $\omega_k \in P$. Următoarele submulţimi ale lui P^∞ :

$$\begin{aligned} A &= A_1 \times A_2 \times ... \times A_n \times P \times P \times ... \\ &= \left\{(\omega_k)_{k\in N} \,|\omega_k \in A_k \text{ pentru } 1 \leq k \leq n\right\} \end{aligned} \tag{9.1}$$

unde $A_k \in \Omega$ pentru orice k, se numesc paralelipipede. Aici n nu este fixat ci poate fi orice număr natural. Fie Ω^∞ submulţimile lui P^∞ care sunt reuniuni finite de paralelipipede. Se arată că aceaste mulţimi formează o algebră. Pe această algebră putem defini o unică probabilitate p' astfel ca pentru paralelipipede să avem:

$$p'(A) = p(A_1) \cdot p(A_2) \cdot ...p(A_n) \tag{9.2}$$

unde p este probabilitatea pe P. Definiţia seamănă cu definiţia volumului unui paralelipiped în funcţie de lungimile laturilor sale. Asemenea probabilităţi nu sunt suficiente pentru nevoile de calcul. E nevoie de o proprietate de continuitate de genul: $B_1 \subset B_2 \subset ...B_k \subset ...$ cu $B = \cup_{k=1,\infty} B_k$ implică $p(B) = \lim\limits_{n\to\infty} p(B_k)$. Construcţia unei astfel de probabilităţi peP^∞ se realizează astfel:

a) Se extinde Ω^∞ la cea mai mică σ algebră (deci algebră de mulţimi închisă şi la reuniuni numărabile) notată $\Omega^{(\infty)}$

b) Probabilitatea p' definită pe Ω^∞ ese extinde unic la o σ probabilitate pe $\Omega^{(\infty)}$ notată $p^{(\infty)}$

Probabilitatea $p^{(\infty)}$ se numeşte probabilitate produs. Detaliile de construcţie nu fac obiectul acestui curs. Putem remarca asemănarea construcţiei probabilităţii produs cu a volumului corpurilor plecând de la lungime. Aşa cum în afară de reuniuni finite de paralelipipede există şi alte corpuri cu volum, tot aşa apar în $\Omega^{(\infty)}$ şi alte mulţimi care au probabilitate, în afară de reuniunile finite de tipul (9.1).

Definiţia 9.3 *Mulţimea P^∞ împreună cu $\Omega^{(\infty)} \subset P(P^\infty)$ şi cu probabilitatea $p^{(\infty)} : \Omega^{(\infty)} \to R$ se numeşte produsul infinit al câmpului de probabilitate (P, Ω, p).*

Observaţia 9.4 *In lecţia 2 am introdus produsul finit al unor câmpuri de probabilitate. Faţă de cazul considerat acolo, aici avem două lucruri în plus:*
a) pentru a avea o σ probabilitate pe produs trebuie extinsă algebra de mulţimi formată din reuniuni finite de mulţimi paralelipipedice la o σ algebră
b) Am luat în consideraţie o infinitate de factori în produs .

Observaţia 9.5 *In principiu nu e nevoie de o cunoaştere detaliată a produsului de câmpuri de probabilitate. Este suficient să ştim că el există şi că probabilitatea unei mulţimi paralelipipedice este produsul probabilităţilor factorilor (formula 9.2).*

Fie acum X o v.a. pe P, $X : P \to R$. In aceste condiţii pe P^∞ avem un şir de v.a. definite prin:

$$X_i : P^\infty \to R, \quad X_i(\omega) = X_i\left((\omega_k)_{k\in N}\right) = X(\omega_i)$$

pentru orice $i \in N$. Prin urmare X_i aplicată unui şir este valoarea lui X pe componenta a i a şirului $\omega = (\omega_k)_{k\in N} \in P^\infty$. Aceste v.a. sunt independente şi la fel distribuite (adică au aceeaşi funcţie de repartiţie, deci aceleaşi caracteristici numerice). Pe produsul finit $P^n = P \times P ... \times P$ avem în mod analog variabilele aleatoare X_i definite prin formula de mai sus dar cu $\omega \in P^n$. Ele sunt independente şi la fel distribuite.

In concluzie, *mai multe măsurători ale unei mărimi apar în statistică astfel:*

a) Urmărim o componentă numerică a unui fenomen, să zicem notată cu X.

b) Acea caracteristică depinde de o seamă de factori dintr-o mulţime P, în general neexplicită.

c) Admitem că pe P există o probabilitate p, iar X:P$\to$ R este o variabilă aleatoare.

d) Prin n experienţe independente găsim pentru X valorile $x_1, x_2, ...x_n$.

e) $x_1, x_2, ...x_n$ apar ca valorile a n variabile aleatoare X_1, X_2, $...X_n$ definite pe spaţiul produs P^n sau pe P^∞. Aceste v.a. sunt independente şi la fel distribuite ca X. Vom numi X_1, X_2, $...X_n$ variabile aleatoare de selecţie asociate lui X. X_i reprezintă rezultatul experienţei i. In cele ce urmează vom considera toate variabilele X_i definite pe aceeaşi mulţime P^∞.

Ne ocupăm în continuare de operaţiile pe care le facem cu rezultatele x_1, x_2, $...x_n$ pentru a obţine caracteristici ale variabilei aleatoare X. Vom folosi doar acele operaţii în care mulţimea P, care nu este explicită, nu intervine efectiv. Probabilitatea pe P^∞ o vom nota uneori cu Prob alteori cu p.

Definiţia 9.6 *Se numeşte statistică un şir $(G_n)_{n\in N}$ de variabile aleatoare $G_n : P^\infty \to R$.*

Toate statisticile utilizate de noi vor fi de forma următoare:
i) se dă un şir de funcţii $g_n : R^n \to R$
ii) avem o v.a. $X : P \to R$
Fie X_1, X_2,...X_n,... variabilele aleatoare asociate. Definim statistica:

$$g_n(X_1, X_2, ...X_n) : P^\infty \to R$$

, astfel

$$g_n(X_1, X_2, ..X_n)(\omega) = g_n(X_1(\omega), X_2(\omega), ...X_n(\omega)) \text{ pentru } \omega \in P^\infty$$

Uneori vom folosi termenul de statistică pentru şirul de variabile aleatoare construite mai sus pe P^∞.

Exemple de statistici frecvent folosite

Fie X o v.a. cu funcţia de repartiţie $F : R \to R$, şi X_1, X_2, $...X_n$ variabilele de selecţie asociate. Vom folosi următoarele notaţii:

a1) m=$\int_{-\infty}^{\infty} xdF(x)$ = media lui X (Lecţia 3)

b1) M$(X_1, X_2, ...X_n) = \bar{X}_{(n)} = \frac{X_1+X_2+...+X_n}{n} : P^\infty \to R$, o v.a. numită *media de selecţie.* Uneori o vom nota cu $\bar{X}_{(n)}$ pentru a pune în evidenţă dependenţa de n, alteori o vom nota simplu $\bar{X}$. Astfel,

c1) m$^* = m^*(x_1, x_2, ...x_n) = \frac{x_1+x_2+...+x_n}{n}$ este valoarea mediei de selecţie pentru rezultatele $x_1, x_2, ...x_n$ obţinute în cele n experienţe, numită şi *media empirică (Lec*ţia 8).

ak) $m_k = \int_{-\infty}^{\infty} x^k dF(x)$ =momentul de ordin k al lui X (Lecţia 3)

bk) M$_k(X_1, X_2, ...X_n) = \frac{X_1^k+X_2^k+...+X_n^k}{n}$, o v.a. numită *momentul de selecţie de ordin k.*

ck) m$_k^* = m_k^*(x_1, x_2, ...x_n) = \frac{x_1^k+x_2^k+...+x_n^k}{n}$ *=momentul empiric de ordin k (Lec*ţia 8)

ak0) $\mu_k = \int_{-\infty}^{\infty} (x-m)^k dF(x)$ = momentul centrat de ordin k al lui X (Lecţia 3)

bk0) M$_{0k}(X_1, X_2, ...X_n) = \frac{\left(X_1-\bar{X}\right)^k+\left(X_2-\bar{X}\right)^k+...+\left(X_n-\bar{X}\right)^k}{n} : P^\infty \to R$, o v.a. numită *momentul centrat de ordin k.*

ck0) $\mu_k^* = \mu_k^*(x_1, x_2...x_n) = \frac{(x_1-m^*)^k+(x_2-m^*)^k+...+(x_n-m^*)^k}{n}$ *=momentul centrat de ordin k, empiric (Lec*ţia 8).

a20) D=$\sigma^2 = \int_{-\infty}^{\infty} (x-m)^2 dF(x)$ = dispersia lui X (Lecţia 3).

b20) D$(X_1, ...X_n) = S^2(X_1, ...X_n) = \frac{\left(X_1-\bar{X}\right)^2+\left(X_2-\bar{X}\right)^2+...+\left(X_n-\bar{X}\right)^2}{n}$:P$^\infty \to R$, *o v.a. numită dispersia de selecţie.*

$$S'^2(X_1, X_2, ...X_n) = \frac{\left(X_1-\bar{X}\right)^2+\left(X_2-\bar{X}\right)^2+...+\left(X_n-\bar{X}\right)^2}{n-1} : P^\infty \to R$$

se numeşte *dispersia de selecţie modificată.*

c20) s^{*2}=D$^* = D^*(x_1, x_2, ...x_n) = \frac{(x_1-m^*)^2+(x_2-m^*)^2+...+(x_n-m^*)^2}{n}$ = *dispersia empirică de selecţie* (Lecţia 8), iar $s'^{*2} = \frac{(x_1-m^*)^2+(x_2-m^*)^2+...+(x_n-m^*)^2}{n-1}$ se numeşte *dispersia empirică modificată.*

E de aşteptat ca v.a. de selecţie de mai sus să aproximeze într-un fel sau altul marimile corespunzătoare ale variabilei X.

Propoziţia 9.7 *Avem relaţia*

$$S^2(X_1, X_2, ..., X_n) = M_2(X_1, X_2, ..., X_n) - \overline{X}^2 \tag{9.3}$$

Demonstratie

$$\begin{aligned} S^2(X_1, X_2, ..., X_n) &= \frac{1}{n}\sum\left(X_i - \overline{X}\right)^2 = \frac{1}{n}\sum\left(X_i^2 - 2\overline{X}X_i + \overline{X}^2\right) \\ &= \frac{1}{n}\sum X_i^2 - \frac{2}{n}\cdot\overline{X}\cdot\sum X_i + \frac{1}{n}\cdot n\overline{X}^2 \\ &= \frac{1}{n}\sum X_i^2 - 2\cdot\overline{X}\cdot\overline{X} + \overline{X}^2 \\ &= M_2(X_1, X_2, ..., X_n) - \overline{X}^2 \end{aligned}$$

QED.
Fie X o v.a. şi $X_1, X_2, ..., X_n...$ variabilele de selecţie asociate. Fie de asemenea A$\in R$.

Definiţia 9.8 *Se numeşte estimator sau funcţie de estimaţie pentru A, o statistică $(g_n)_{n\in N}$ astfel ca pentru orice $\epsilon > 0$ să avem:*

$$\lim_{n\to\infty} Prob(|g_n(X_1, X_2, ...X_n) - A| > \epsilon) = 0$$

Cu alte cuvinte, $\epsilon > 0$ fiind dat, pentru valori mari ale lui n este foarte puţin probabil ca variabila aleatoare $g_n(X_1, X_2, ...X_n)$ să ia valori în afara intervalului $[A-\epsilon, A+\epsilon]$, adică este foarte puţin probabil ca numărul
$g_n(x_1, x_2, ...x_n)$ să fie în afara intervalului $[A-\epsilon, A+\epsilon]$. In aceste condiţii, după un număr de n experienţe, considerăm pe $g_n(x_1, x_2, ...x_n)$ ca o aproximaţie bună pentru A. Este posibil să ne înşelăm, dar probabilitatea de a ne înşela este mică, pentru n mare. Statistica nu ne oferă răspunsuri sigure ci doar aproximaţii în care putem avea un grad mai mic sau mai mare de încredere. Se acceptă acele aproximaţii în care avem un grad mai mare de încredere.

Definiţia 9.9 *O statistică $(g_n(X_1, ...X_n))_{n\in N}$ se numeşte corectă sau deplasată relativ la valoarea A dacă avem:*

1) $\lim\limits_{n\to\infty}$ M$(g_n(X_1, X_2, ..., X_n)) = A$.
2) $\lim\limits_{n\to\infty} D(g_n(X_1, X_2, ..., X_n)) = 0$.
şi se numeşte *absolut corectă sau nedeplasată* dacă în plus M$(g_n(X_1, X_2, ..., X_n)) = A$.

Condiţiile 1) şi 2) din definiţia de mai sus pun în evidenţă situaţii în care o statistică oarecare $(g_n)_{n\in N}$ este un estimator pentru o valoare A. Teorema de mai jos pune în evidenţă importanţa condiţiilor din definiţia anterioară.

Teorema 9.10 *Dacă statistica $(g_n(X_1, X_2, ..., X_n))_{n\in N}$ este corectă relativ la A atunci ea este un estimator al lui A, adică pentru orice $\epsilon > 0$ avem*

$$\lim_{n\to\infty} Prob(|g_n(X_1, X_2, ...X_n) - A| > \epsilon) = 0.$$

Demonstraţie. Conform cu inegalitatea lui Cebâşev (Lecţia 5) pentru un $\epsilon > 0$ avem:

$$\text{Prob}(|g_n(X_1, X_2, ...X_n) - M(g_n(X_1, X_2, ...X_n))|>\epsilon)\leq$$
$$\leq \frac{D(g_n(X_1,X_2,...X_n))}{\epsilon^2}$$

Acum ţinând seama de 1) şi 2) din definiţia corectitudinii rezultă

$$Prob(|g_n(X_1, X_2, ...X_n) - A| > \epsilon) \to 0$$

când $n \to \infty$, deci statistica $(g_n(X_1, X_2, ..., X_n))_{n\in N}$ este un estimator al lui A.

QED.

Arătăm acum că funcţiile de selecţie introduse cu ocazia notaţiilor precedente sunt estimatori pentru valorile corespunzătoare ale variabilei X.

Teorema 9.11 *a) Statistica media de selectie:*

$$g_n(X_1, X_2, ..., X_n) = \bar{X}_{(n)} = (X_1 + X_2 + \cdots + X_n)/n$$

estimează media m =M(X) a v.a. X absolut corect.

b) Statistica

$$h_n(X_1, X_2, ..., X_n) = \frac{1}{n}\sum X_i^r$$

estimează absolut corect momentul de ordin r, m_r, al v.a. X.

c) Statistica

$$S^2(X_1, X_2, ..., X_n) = \frac{1}{n}\sum \left(X_i - \overline{X}_{(n)}\right)^2$$

estimează corect dar nu absolut corect dispersia v.a. X, σ^2 =D(X).

d) $S'^2(X_1, X_2, ..., X_n) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_{(n)}\right)^2$, adică dispersia de selectie modificată, aproximează absolut corect dispersia v.a. X.

Demonstratie Trebuie verificate condiţiile 1 şi 2 din definiţia statisticii corecte.

a) Verificăm 1): M(g_n)= $\frac{1}{n}\sum$M(X$_i$)= $\frac{1}{n}\cdot n\cdot m = m$ =M(X).

Verificăm 2): D(g_n)= $\frac{1}{n^2}\sum$D(X$_i$)= $\frac{D(X)}{n}$, deoarece $X_1, X_2, ..., X_n$ sunt independente. Prin urmare D(g_n)$\to 0$, când $n\to\infty$.

b) Verificăm 1): M(h_n)= $\frac{1}{n}\sum$M(X$_i^r$)= $\frac{1}{n}\cdot n\cdot m_r = m_r$

Verificăm 2): D(h_n)= $\frac{1}{n^2}\sum$D(X$_i^r$)= $\frac{1}{n}$D(X^r)$\to 0$, când $n\to\infty$. Retinem formula:

$$\mathbf{D}(\overline{\mathbf{X}}_{(\mathbf{n})} = \mathbf{D}(\mathbf{X})/\mathbf{n} \qquad (9.4)$$

c) încercăm să verificăm 1):

$$
\begin{aligned}
& M\left(S^2\left(X_1, X_2, ..., X_n\right)\right) \overset{\text{din } 9.3}{=} M\left(\mathrm{M}_2\left(X_1, X_2, ..., X_n\right) - \overline{X}^2_{(n)}\right) \\
= \; & M\left(\mathrm{M}_2\left(X_1, X_2, ..., X_n\right)\right) - \frac{1}{n^2} M\left(\left(\sum X_i\right)^2\right) \\
& = M(X^2) - \frac{1}{n^2} M\left(\left(\sum X_i\right)\left(\sum X_j\right)\right) \\
= \; & M(X^2) - \frac{1}{n^2} M\left(\sum X_i^2\right) - \frac{1}{n^2} M\left(\sum_{i \neq j} X_i X_j\right) \\
& = M(X^2) - \frac{n}{n^2} M\left(X^2\right) - \frac{1}{n^2} \sum_{i \neq j} M\left(X_i X_j\right) \\
& \overset{X_i, X_j independente.}{=} \frac{n-1}{n} M(X^2) - \frac{1}{n^2} \cdot n(n-1) M(X)^2 \\
= \; & \frac{n-1}{n}\left(\mathrm{M}(\mathrm{X}^2) - \mathrm{M}(\mathrm{X})^2\right) = \frac{n-1}{n} D(X) \to D\left(X\right)
\end{aligned}
$$

Prin urmare S^2 *nu* estimează absolut corect dispersia v.a. X. Retinem formula deja găsită:

$$
\mathbf{M}\left(\mathrm{S}^{\mathbf{2}}\left(\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_n}\right)\right) = \frac{\mathbf{n-1}}{\mathbf{n}} \mathbf{D}(\mathbf{X}) \tag{9.5}
$$

E clar că $M\left(S'^2\left(X_1, X_2, ..., X_n\right)\right) = D\left(X\right)$.
Verificăm 2): lăsăm ca exercitiu pentru cititor verificarea formulei

$$
D(S'^{\,2}) = \frac{1}{n}\left(m_4 - \frac{n-3}{n-1}\left[\mathrm{D(X)}\right]^2\right) \tag{9.6}
$$

unde m_4 este momentul de ordin 4 al v.a. X.
Se vede clar de aici că $\mathrm{D(S'^{\,2})} \to 0$, când $n \to \infty$. De asemenea $D\left(S^2\right) = \left(\frac{n-1}{n}\right)^2 D\left(S'^2\right) \to 0$ când $n \to \infty$.
QED.

Observaţia 9.12 *In practică se foloseste $S'^{\,2}$ în locul lui S^2 deoarece dă rezultate mai bune după cum ne arată teorema 2. Totusi formula (9.6) ne spune că pentru n suficient de mare si statistica S^2 poate fi folosită ca estimator al dispersiei v.a. X. Din definitia lui $S'^{\,2}$ si din formula (9.6) găsim formula utilă:*

$$
D(S^{\,2}) = \frac{(n-1)^2}{n^3}\left(m_4 - \frac{n-3}{n-1}\left[\mathrm{D(X)}\right]^2\right) \tag{9.7}
$$

Exerciţiul 9.13 *Fie selectia {0, 1, 1, 0, 1, 1, 2, 0, 0, 2}. Să se estimeze absolut corect dispersia populatiei din care provine această selectie.*

Soluţie. Media este estimată absolut corect de media empirică $m^* = 8/10 = 0,8$. Dispersia este estimată absolut corect de dispersia modificată empirică

$$\begin{aligned} s'^{*2} &= \frac{1}{9}[(0-0,8)^2 + (1-0,8)^2 + (1-0,8)^2 + (0-0,8)^2 + (1-0,8)^2 \\ &\quad +(1-0,8)^2 + +2-0,8)^2 + (0-0,8)^2 + (0-0,8)^2 + (2-0,8)^2] \\ &= 0,56 \end{aligned}$$

Observaţia 9.14 *Deoarece dispersia se mai numeste si varianţă vom folosi si noi uneori varianţa de selectie pentru dispersia de selectie.*

9.1 Principiul verosimilitătii maxime

Presupunem că P este o populatie unde se urmăreşte caracteristica numerică X, care este o variabilă aleatoare cu densitatea de probabilitate $f(x;\theta)$, θ fiind un parametru necunoscut. Cunoastem doar forma matematică a functiei $f(x;\theta)$. De exemplu dacă ştim că X este o v.a. normală cu media θ, necunoscută dar cu dispersia σ^2 cunoscută, atunci $f(x;m) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}}$.

Pentru determinarea lui θ facem o selectie care dă rezultatele $\{x_1,...,x_n\}$şi încercăm pe baza lor să estimăm pe θ. Deoarece v.a. de selectie $X_1,...,X_n$ sunt independente, probabilitatea ca X_1 să ia valoari în intervalul $[x_1, x_1+dx_1)$, X_2 să ia valori în $[x_2, x_2+dx_2)$, ..., X_n să ia valori în $[x_n, x_n+dx_n)$ este dată de $f(x_1;\theta)\cdot\ f(x_2;\theta)\cdots\ f(x_n;\theta)dx_1dx_2...dx_n = L\left(x_1,...,x_n;\theta\right)dx_1dx_2....dx_n$. Această functie L se numeste *functia de verosimilitate* şi va fi folosită pentru estimarea lui θ.

Dacă X ia valori discrete, atunci $f(x,\theta)$ este probabilitatea ca X să ia valoarea x. De exemplu, în cazul distributiei Poisson, $f(x;\theta) = e^{-\theta}\cdot\frac{\theta^x}{x!}$, cu $x \in \mathbf{N}$, reprezintă probabilitatea ca $X = x$, iar θ este parametrul necunoscut (pe care urmează să-l estimăm!). Probabilitatea ca în n selecţii independente să se obţină rezultatele $x_1, x_2, ...x_n$ este $f\left(x_1,\theta\right)\cdot f\left(x_2,\theta\right)...\cdot f\left(x_n,\theta\right) = L\left(x_1,x_2,...x_n;\theta\right)$ care se numeşte şi în acest caz funcţia de verosimilitate.

Functia L este determinată de volumul selectiei n si depinde de θ. Metoda verosimiltătii maxime constă în următorul principiu (axiomă): *valoarea "cea mai verosimilă" (cea mai potrivită în acest sens!) a parametrului θ este aceea pentru care funcţia $L\left(x_1,...,x_n;\theta\right)$ este maximă.* După cum ştim de la Analiza matematică, această cerintă are loc dacă avem:

$$\frac{\partial L\left(x_1,...,x_n;\theta\right)}{\partial\theta} = 0 \tag{9.8}$$

adică θ este un punct crirtic pentru $L\left(x_1,...,x_n;\theta\right)$.

Ecuatia (9.8) în practică se dovedeste dificilă. De aceea cel mai des se foloseste observatia: $L(x_1, ..., x_n; \theta)$ este maximă dacă si numai dacă ln $L(x_1, ..., x_n; \theta)$ este maximă (functia logaritmică este strict crescătoare). Deci (9.8) este echivalentă cu :

$$\frac{\partial \ln\ L(x_1, ..., x_n; \theta)}{\partial \theta} = 0 \tag{9.9}$$

care poartă numele de *ecuatie a verosimilitătii maxime.* Rezolvăm ecuatia (9.9), sau ecuatia (9.8) si găsim $\theta = \Theta_n(x_1, ...x_n)$. Ca estimator (funcţie de estimare) pentru θ luăm variabila aleatoare $\Theta_n(X_1, X_2, ...X_n)$, care, pentru selecţia $\{x_1, x_2, ...x_n\}$ dă rezultatul $\Theta_n(x_1, x_2, ...x_n)$.

Se poate demonstra că în condiţii foarte generale, pentru selecţii mari, statistica $\Theta(X_1, X_2, ...X_n)$ obţinută prin metoda verosimilităţii maxime, are o distribuţie aproximativ normală, cu media egală cu θ=valoarea adevărată a parametrului şi dispersia

$$D(\Theta) = \frac{1}{-n\int_{-\infty}^{\infty}\left(\frac{\partial^2 \ln(f(x,\theta))}{\partial\theta^2}\right) f(x;\theta)\,dx} = \frac{1}{n\int_{-\infty}^{\infty}\left(\frac{\partial \ln\ f(x;\theta)}{\partial\ \theta}\right)^2 f(x;\theta)\,dx}$$

Dacă distribuţia este discretă atunci integralele din formula precedentă devin sume.

Exemplul 9.15 *Presupunem că populatia are distributia Poisson (cazul evenimentelor rare). Functia de probabilitate este $f(k;\lambda) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$, k =0, 1, 2, ... Ne interesează să estimăm parametrul λ prin metoda verosimilitătii maxime. Pentru aceasta facem o selectie $\{x_1, x_2, ..., x_n\} \subset \{0, 1, 2, ...\}$.*

$$L(x_1, ..., x_n; \lambda) = f(x_1; \lambda) \cdot f(x_2; \lambda) \cdot \cdots \cdot f(x_n; \lambda) = e^{-n\lambda} \cdot \frac{\lambda^{\sum x_k}}{x_1! x_2! ... x_n!}$$

$$\ln L(x_1, ..., x_n; \lambda) = -n\lambda + \left(\sum x_k\right) \ln\lambda - \sum \ln(x_k!)$$

$\frac{\partial \ln L}{\partial\lambda} = 0$ ne furnizează $\lambda = \frac{\sum x_k}{n}$, deci un estimator pentru λ este $\Lambda_n(X_1, ..X_n) = \frac{X_1+X_2+...X_n}{n}$ adică media de selectie. Deoarece λ este media lui X (variabilă Poisson), Λ_n este un estimator absolut corect pentru λ.

Este el oare si cel mai *eficient*, în sensul că are dispersia cea mai mică? Este greu de răspuns la această întrebare. Totusi avem un rezultat puternic care face oarecare lumină:

Teorema 9.16 *(Rao-Cramer) Dacă statistica $g_n(X_1, ..., X_n)$ dă un estimator eficient (cu dispersia minimă , în multimea tuturor estimatorilor absolut corecti pentru θ), atunci*

$$D(g_n(X_1, ..., X_n)) = \frac{1}{n\int\limits_{-\infty}^{\infty}\left(\frac{\partial \ln\ f(x;\theta)}{\partial\ \theta}\right)^2 f(x;\theta)dx}$$

sau

$$D\left(g_n\left(X_1, ..., X_n\right)\right) = \frac{1}{n \sum\limits_{x=0}^{\infty} \left(\frac{\partial \ln\ f(x;\theta)}{\partial\ \theta}\right)^2 f(x;\theta)} \tag{9.10}$$

dacă distribuţia este cu valori discrete.

Fără demonstraţie.

Ne întoarcem la exemplul anterior. Stim că $D(\Lambda_n) = D\left(\frac{\sum X_k}{n}\right) = \frac{D(X)}{n} = \frac{\lambda}{n}$. $f(x,\lambda) = \frac{e^{\lambda}\lambda^x}{x!}$, deci $\frac{\partial \ln f(x,\lambda))}{\partial\lambda} = -1 + \frac{x}{\lambda}$. De aici rezultă

$$\begin{aligned}
& \sum_{x=0}^{\infty} \left(\frac{\partial \ln f\left(x,\lambda\right)}{\partial\lambda}\right)^2 f\left(x,\lambda\right) \\
= & \sum_{x=0}^{\infty} \left(1 - 2\frac{x}{\lambda} + \frac{x^2}{\lambda^2}\right) e^{-\lambda}\frac{\lambda^x}{x!} \\
= & \ e^{-\lambda}\left(\underbrace{\sum_{x=0}^{\infty}\frac{\lambda^x}{x!}}_{=e^{\lambda}} - 2\underbrace{\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!}}_{=e^{\lambda}} + \sum_{x=1}^{\infty}\underbrace{\frac{\lambda^{x-2}}{(x-1)!}(x-1+1)}_{=e^{\lambda}+\frac{1}{\lambda}e^{\lambda}}\right) \\
= & \ \frac{1}{\lambda}
\end{aligned}$$

Prin urmare $\frac{1}{n \sum\limits_{x=0}^{\infty} \left(\frac{\partial \ln\ f(x;\lambda)}{\partial\ \lambda}\right)^2 f(x;\lambda)} = \frac{\lambda}{n} = D\left(\Lambda_n\right)$. Rezultă din teorema Rao-Cramer că statistica medie de selectie este si un estimator eficient pentru λ. Putem spune acum că $\lambda = \left(\sum x_k\right)/n$ este o estimatie "foarte bună" în toate sensurile.

Exemplul 9.17 *Să se estimeze parametrul* p *al unei distribuţii Bernoulli*

$$\begin{pmatrix} 1 & 0 \\ p & 1-p \end{pmatrix}$$

prin metoda verosimilităţii maxime.

Soluţie.

$$f\left(x,p\right) = \begin{cases} p \text{ dacă x=1} \\ 1-p \text{ dacă x=0} \end{cases}$$

Funcţia de verosimilitate este $L\left(x_1, x_2, ...x_n\right) = p^{n_1}\left(1-p\right)^{n-n_1}$ unde n_1 este numărul de realizări ale lui 1. $\frac{\partial \ln L}{\partial p} = 0$ devine $\frac{\partial(n_1 \ln p + (n-n_1)\ln(1-p))}{\partial p} = 0$ adică $\frac{n_1}{p} - \frac{n-n_1}{1-p} = 0$ care are soluţia $p = \frac{n_1}{n}$. Valoarea n_1 este valoarea variabilei $X_1 + X_2 + ... + X_n$, unde X_i este variabila de selecţie a cărei valoare este 1 dacă la experienţa i se obţine rezultatul 1 şi are valoarea 0 în caz contrar. Prin urmare statistica ce estimează parametrul p este

$$\mathcal{P} = \frac{X_1 + X_2 + ... + X_n}{n}$$

care este chiar media de selecţie. La fel ca în cazul repartiţiei Poisson se arată că

$$\begin{aligned} D(\mathcal{P}) &= \frac{pq}{n} \\ &= \frac{1}{n\sum_{x=0}^{1}\left(\frac{\partial \ln f(x;p)}{\partial p}\right)^2 f(x;p)} = \frac{1}{n\left(\left(\frac{\partial \ln(1-p)}{\partial p}\right)^2(1-p) + \left(\frac{\partial \ln p}{\partial p}\right)^2 p\right)} \end{aligned}$$

deci estimarea lui p este absolut corectă (exerciţiu).

Dacă avem de estimat mai mulţi parametri $\theta_1, \theta_2, ..\theta_p$, ştiind că densitatea de probabilitate a variabilei aleatoare X este $f(x; \theta_1, ...\theta_p)$, atunci în mod analog cu cazul unui singur parametru, principiul verosimilităţii maxime spune că în urma a n experienţe independente care dau rezultatele $x_1, x_2, ...x_n$, se aleg pentru parametri acele valori care maximizează funcţia de verosimilitate $L(x_1, ..x_n; \theta_1, ..\theta_p) = f(x_1; \theta_1, \theta_2, ...\theta_p) \cdot f(x_2; \theta_1, ..\theta_p) \cdot ..f(x_n; \theta_1, \theta_2, ..\theta_p)$ sau ceea ce este acelaşi lucru acele valori care maximizează $\ln L(x_1, x_2, ..x_n; \theta_1, ...\theta_p)$. Aceasta implică:

$$\begin{cases} \frac{\partial \ln L(x_1,x_2,..x_n;\theta_1,...\theta_p)}{\partial \theta_1} = 0 \\ \frac{\partial \ln L(x_1,x_2,..x_n;\theta_1,...\theta_p)}{\partial \theta_2} = 0 \\ ..... \\ \frac{\partial \ln L(x_1,x_2,..x_n;\theta_1,...\theta_p)}{\partial \theta_p} = 0 \end{cases} \tag{9.11}$$

Exemplul 9.18 *Să presupunem că v.a. X are o distribuţie normală cu $f(x; m; \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$. In acest caz avem $\ln L = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{\sum(x_i-m)^2}{2\sigma^2}$. Sistemul (9.11) devine:*

$$\begin{cases} \frac{1}{\sigma^2}\sum(x_i - m) = 0 \\ -\frac{n}{\sigma} + \frac{\sum(x_i-m)^2}{\sigma^3} = 0 \end{cases}$$

care are ca soluţii $m = \frac{\sum x_i}{n} = m^*(x_1, ..x_n)$ şi $\sigma = \sqrt{\frac{\sum_i(x_i-m^*)^2}{n}}$.

Ştim din această lecţie că M$(X_1, X_2, ...X_n) = \bar{X}_{(n)} = \frac{X_1+X_2+...+X_n}{n}$ şi D$(X_1, ...X_n) = \frac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+...+(X_n-\bar{X})^2}{n}$ sunt estimaţii corecte pentru media şi dispersia unei variabile aleatoare, în cazul nostru pentru m şi σ^2.

9.2 Metoda momentelor (K. Pearson)

Dată selectia $\{x_1, ..., x_n\}$ noi putem calcula momentul de ordin k al selectiei: $m_k^* = \frac{\sum_i x_i^k}{n}$, pentru orice k =0, 1, 2, ... Obtinem astfel estimatori pentru medie, dispersie ,momente de diferite ordine. Funcţia caracteristică $X_c(t)$ are toate derivatele în $t = 0$ date de $X_c^{(k)}(0) = i^k M^k(X)$. Prin urmare în condiţii foarte generale, care asigură că $X_c(t)$ este analitică (se poate dezvolta în serie convergentă de puteri în jurul oricărui punct), rezultă că momentele $M^k(X)$ determină pe $X_c(t)$ care la rândul ei determină repartiţia lui X (vezi Lecţia 3). Această observaţie a fost folosită de K. Pearson pentru a găsi estimatori pentru parametrii unei legi de probabilitate.

Fie $\rho(x; \theta_1, \theta_2, ...\theta_p)$ densitatea de probabilitate a v.a. X, unde parametrii $\theta_1, ...\theta_p$ sunt necunoscuţi. Există relaţiile:$m_k = \int_{-\infty}^{\infty} \rho(x; \theta_1, \theta_2, ...\theta_p) x^k dx$ pentru orice k. Am văzut că m_k este estimat de m_k^*. Egalând valoarea teoretic exactă cu estimarea practică, $m_k = m_k^*$, adică:

$$\left\{ \int_{-\infty}^{\infty} x^k \cdot \rho(x; \theta_1, \theta_2, ...\theta_p) dx = \frac{\sum_{i=1}^{i=n} x_i^k}{n} \right. \tag{9.12}$$

pentru k=1,2,...p, obţinem un sistem care dă prin rezolvare $\theta_k = \Theta_n(x_1, x_2, ...x_n)$, pentru k=1,2,..p.

Ca estimatori pentru θ_k se iau v.a. $\Theta_k(X_1, X_2, ...X_n)$.

Dacă v.a. X este dicretă atunci integrala dim (9.12) devine sumă, la fel ca în cazul metodei verosimilităţii maxime.

Exemplul 9.19 *Fie v.a. cu densitatea $\rho(x; \lambda) = \frac{1}{\Gamma(\lambda)} x^{\lambda-1} e^{-x}$, pentru $\lambda > 0 \quad x > 0$. Aici parametrul este λ. Se cere o metodă de a estima pe λ prin selecţii. Dacă folosim metoda momentelor găsim*

$$\int_0^{\infty} x \cdot \frac{1}{\Gamma(\lambda)} x^{\lambda-1} e^{-x} dx = \frac{x_1 + x_2 + ...x_n}{n}$$

adică $\lambda = \frac{x_1+x_2+...x_n}{n}$.

Exemplul 9.20 *Densitatea de probabilitate a unei v.a. X are forma:*

$$f(x; a; b; c) = \begin{cases} a + 2bx, \ x \in [0, c] \\ 0, \text{ în rest} \end{cases}$$

Rezultatele unei selecţii de volum n=3 dau pentru X valorile $\{x_1, x_2, x_3\}$={-1,0,1} . Să se estimeze papametrii a, b, c prin metoda momentelor.

Mai întâi punem conditia ca $\int\limits_{-\infty}^{\infty} f(x; a; b; c) dx = 1$, de unde găsim relatia:

$$ac + bc^2 = 1 \tag{9.13}$$

Calculăm acum media v.a. X:

$$M(X) = \int_0^c x(a+2bx)dx = \frac{ac^2}{2} + \frac{2bc^3}{3} \tag{9.14}$$

momentul de ordin 2:

$$M_2(X) = \int_0^c x^2(a+2bx)dx = \frac{ac^3}{3} + \frac{2bc^4}{4} \tag{9.15}$$

Momentele de selectie m_1^* =(-1+0+1)/3=0 si $m_2^* = \left((\text{-}1)^2 + 0^2 + 1^2\right)/3$=2/3 vor estima pe $M(X)$ si pe $M_2(X)$. Deci vom obtine sistemul nelinear de ecuatii:

$$\begin{cases} ac + bc^2 = 1 \\ \frac{ac^2}{2} + \frac{2bc^2}{3} = 0 \\ \frac{ac^3}{3} + \frac{2bc^4}{4} = \frac{2}{3} \end{cases} \tag{9.16}$$

In general rezolvarea acestor sisteme (care se obtin folosind metoda momentelor) este foarte complicată. Din ecuatia a doua găsim $c = -\frac{3a}{4b}$. înlocuim expresia lui c în prima si în ultima ecuatie si găsim:

$$\begin{array}{c} -\frac{3a^2}{4b} + \frac{9a^2}{16b} = 1 \\ -\frac{9a^4}{64b^3} + \frac{81a^4}{512b^3} = \frac{2}{3} \end{array} \tag{9.17}$$

Din prima ecuatie a sistemului (9.17) găsim $b = -\frac{3a^2}{16}$. Inlocuim expresia lui b în ecuatia a doua si găsim că $2a^2 = -8$, lucru imposibil. Concluzia este că selectia {-1,0,1} *nu poate fi* pentru variabila X. Sodajul respectiv este eronat sau X nu are densitatea de probabilitate propusă. De altfel, examinând cu atenţie rezultatele sondajului vedem că valoarea -1 în principiu nu ar fi trebuit să se obţină deoarece X are densitatea pozitivă pe [0,c].

9.3 Exerciţii

1. Pentru o selectie de volum n dintr-o distributie exponentială($\rho(t) = \lambda e^{-\lambda t}$, dacă $t \geq 0$ si 0 în rest) cu parametrul λ, să se găsească un estimator pentru λ folosind metoda verosimilitătii maxime. Presupunem că $\rho(t)$ este densitatea de probabilitate a duratei dintre doua sosiri succesive la o staie de benzină. Se cronometrează 11 sosiri şi se găsesc următorii timpi între ele 4, 3, 6, 1, 1, 4, 2, 6, 1, 3. Calculati prin metoda verosimilitătii maxime pe λ .

2. Considerăm o selectie de volum n dintr-o populatie cu distributia gama ($f(t) = \lambda(\lambda t)^{r-1}e^{-\lambda t} \cdot \frac{1}{(r-1)!}$, dacă $t \geq 0$ si 0 în rest). Găsiti un estimator pentru λ prin metoda verosimilitătii maxime si un altul prin metoda momentelor.

3. Viata unui bec electric, măsurată în numărul de ore de functionare continuă până când se arde, se presupune uniform distribuită cu parametrii a si b:

$$f(x) = \begin{cases} 1/(b-a)\ ,\ a \leq x \leq b \\ 0\ ,\ \text{în rest.} \end{cases}$$

Se face o selectie de n becuri si se notează cu $x_1, ..., x_n$ timpii de functionare ai acestora până când se ard. Determinati estimatori pentru a si b prin metoda momentelor.

4. Functia de probabilitate a v.a. X este dată de

$$f(x) = \begin{cases} \frac{2b(c-bx)}{c^2},\ \text{dacă } 0 \leq x \leq \frac{c}{b} \\ 0,\ \text{în rest} \end{cases}$$

. Stim că media M(X)=$\frac{\mathrm{c}}{3\mathrm{b}}$ *si* $\sigma_X^2 = \frac{\mathrm{c}^2}{18\mathrm{b}^2}$.

i) dacă c=3, este oare media de selectie M a unui esantion de volum n un estimator nedeplasat pentru parametrul b?

ii) dacă b=1/3, este M un estimator pentru c? (P(|M-c| $< \epsilon) \to 1$, când $n \to \infty$). *Indicatie: folositi inegalitatea lui Cebîsev sau teoria din această lecţie.*

5. Fie statistica $g(X_1, X_2, ..., X_n) = a_1X_1 + a_2X_2 + \cdots + a_nX_n$, cu $a_1, ..., a_n \in \mathbf{R}$. Cum trebuie să fie numerele $a_1, ..., a_n$, astfel încât g să fie un estimator nedeplasat pentru media m a populatiei? Indicatie: cereti M(g)=m.

6. Dacă $g(X_1, X_2, ..., X_n)$ este un estimator nedeplasat pentru parametrul θ, este adevărat că si g^2 este un estimator nedeplasat pentru θ^2?

7. Greutatea unor utilaje produse de o firmă este distribuită normal cu dispersia cunoscută σ_X^2, dar cu media m necunoscută. Fie statisticile

$G = \frac{(\mathrm{X}_1+\mathrm{X}_2+\mathrm{X}_3+\mathrm{X}_4+...+\mathrm{X}_n)}{n}$ *şi* $H = (\mathrm{X}_1+2\mathrm{X}_2+3\mathrm{X}_3+4\mathrm{X}_4 + ... + n\mathrm{X}_n) \cdot \frac{2}{n(n+1)}$, $n \in N$.

a) Să se arate că G şi H sunt estimatori nedeplasaţi pentru m.

b) Care estimator are dispersia mai mică?

Lecţia 10

Intervale de încredere

Definiţia 10.1 *Fie P o populatie, θ un parametru al ei şi $g = g(X_1, ..., X_n)$, $h = h(X_1, ..., X_n)$ două statistici astfel încât $g(X_1, ..., X_n) \leq h(X_1, ..., X_n)$, adică oricare ar fi selectia $\{x_1, ..., x_n\}$ să avem că $g(x_1, ..., x_n) \leq h(x_1, ..., x_n)$. Spunem că intervalul $[g, h]$ este un interval de încredere pentru parametrul θ , de nivel de încredere α dacă avem relatia:*

$$Prob\{g \leq \theta \leq h\} \geq \alpha \tag{10.1}$$

Numărul ε=1–α se mai numeste *prag de încredere*. De obicei α se exprimă în procente, de exemplu pentru α=0,95 putem scrie α=95%.

Cerinta (10.1) trebuie înteleasă astfel: dacă după un număr mare de selectii $\{x_1, ..., x_n\}$, să zicem N, K dintre ele dau intervale $[g(x_1, x_2..x_n), h(x_1, x_2, ...x_n)]$ cu proprietatea că $\theta \in [g, h]$ (pentru fiecare selectie fixată, intervalul devine interval obisnuit, numeric), atunci K/N$\geq \alpha$. Altfel spus, intervalele $[g, h]$ acoperă pe θ în proportie de cel puţin α %(de exemplu ,dacă $\alpha = 1/5 = 20/100$, $\alpha = 20\%$).

Definiţia 10.2 *Pentru un α, un interval de încredere $[g_\alpha, h_\alpha]$ de lungime minimă, astfel încât $Prob\{g_\alpha \leq \theta \leq h_\alpha\} = \alpha$, se zice interval de încredere eficient, relativ la încrederea α.*

Pentru calculele următoare vom avea nevoie de teorema:

Teorema 10.3 *Fie X o v.a. normală, de tip $N(m, \sigma)$. Fie $X_1, X_2, ...X_n$ variabilele de selecţie asociate cu X. Atunci avem:*

a) Variabila $\bar{X} = \frac{X_1+X_2+...+X_n}{n}$ este de tipul $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$.

b) Variabila $\left(\frac{X_1-m}{\sigma}\right)^2 + \left(\frac{X_2-m}{\sigma}\right)^2 + ... + \left(\frac{X_n-m}{\sigma}\right)^2$ este de tip $H(n)$ adică este o variabilă χ^2 standard cu n grade de libertate.

c) Variabila $\left(\frac{X_1-\bar{X}}{\sigma}\right)^2 + \left(\frac{X_2-\bar{X}}{\sigma}\right)^2 + ... + \left(\frac{X_n-\bar{X}}{\sigma}\right)^2$ este de tip $H(n-1)$ adică este de tip χ^2 cu n-1 grade de libertate şi este independentă faţă de variabila $\bar{X}$.

d) Variabila $\sqrt{(n-1)\,n}\frac{\bar{X}-m}{\sqrt{\left(X_1-\bar{X}\right)^2+\left(X_2-\bar{X}\right)^2+...+\left(X_n-\bar{X}\right)^2}}$ *este de tip Student cu n-1 grade de libertate.*

Demonstraţie. a) Această afirmaţie este demonstrată în lecţia 4, secţiunea "Repartiţia normală".

b) Deoarece $\left(\frac{X_i-m}{\sigma}\right)$ sunt normale de tip $N(0,1)$ şi independente, afirmaţia de la acest punct rezultă din lecţia 4, secţiunea "Distribuţia χ^2".

c) Faptul că $\bar{X}$ şi $\left(\frac{X_1-\bar{X}}{\sigma}\right)^2+\left(\frac{X_2-\bar{X}}{\sigma}\right)^2+...+\left(\frac{X_n-\bar{X}}{\sigma}\right)^2$ sunt independente nu se demonstrează în acest curs. Acum scriem că $X_i-m=\left(X_i-\bar{X}\right)+\left(\bar{X}-m\right)$, de unde $\sum\left(X_i-m\right)^2=\sum\left(X_i-\bar{X}\right)^2+\sum\left(\bar{X}-m\right)^2+2(\bar{X}-m)\sum\left(X_i-\bar{X}\right)$. Dar $\sum\left(X_i-\bar{X}\right)=0$, deoarece $\bar{X}=\frac{1}{n}\sum X_i$. Prin urmare

$$\sum\left(\frac{X_i-m}{\sigma}\right)^2 = \sum\left(\frac{X_i-\bar{X}}{\sigma}\right)^2+\frac{n(\bar{X}-m)^2}{\sigma^2}=$$
$$\sum\left(\frac{X_i-\bar{X}}{\sigma}\right)^2+\left(\frac{\bar{X}-m}{\sigma/\sqrt{n}}\right)^2$$

Membrul stâng este de tip $H(n)$, iar în membrul doi avem o sumă de v.a. independente, dintre care a doua este de tip $H(1)$ fiind pătratul unei v.a. normale, de tip $N(0,1)$ (vezi lecţia 4). Prin urmare am găsit $\chi^2_{(n)}=?+\chi^2_{(1)}$.Comparând această relaţie cu $\chi^2_{(p+q)}=\chi^2_{(p)}+\chi^2_{(q)}$, unde indicii de jos indică numărul de grade de libertate (vezi lecţia 4), găsim că $\sum\left(\frac{X_i-\bar{X}}{\sigma}\right)^2$este de tip $\chi^2_{(n-1)}$.

d) Conform cu lecţia 6, secţiunea "Distribuţia Student", variabila aleatoare

$$\frac{\frac{\bar{X}-m}{\sigma/\sqrt{n}}}{\sqrt{\frac{\sum\left(\frac{X_i-\bar{X}}{\sigma}\right)^2}{n-1}}}=\sqrt{(n-1)\,n}\frac{\bar{X}-m}{\sqrt{\left(X_1-\bar{X}\right)^2+\left(X_2-\bar{X}\right)^2+...+\left(X_n-\bar{X}\right)^2}}$$

fiind de tipul $\frac{f}{\sqrt{\frac{g}{n-1}}}$, cu f de tipul $N(0,1)$ şi g de tipul $H(n-1)=H(n-1,1)$, rezultă că are o distribuţie Student cu $n-1$ grade de libertate.

QED.

10.1 Intervale de încredere pentru medie

Să considerăm o caracteristică numerică X *care are o disribuţie normală* de medie m şi dispersie σ^2. Dacă în urma unei selecţii de volum n s-au obţinut rezultatele $x_1, x_2, ...x_n$ pentru X, atunci, conform celor arătate în lecţia trecută valoarea $\frac{x_1+x_2+..x_n}{n}$ este o estimare bună pentru m iar $\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+...(x_n-\bar{x})^2}{n}$ este o estimare bună pentru σ^2. Ce încredere putem avea în aceste estimări? In continuare vom da un răspuns la această întrebare?

10.1.1 Dispersia este cunoscută

Să considerăm cazul când dispersia σ^2 este cunoscută. Valorile $x_1, x_2, ...x_n$ sunt valorile variabilelor aleatoare de selecţie, independente, $X_1, X_2, ...X_n$, care au aceeaşi distribuţie normală ca X. Deoarece variabila $\bar{X} = \frac{X_1+X_2+...X_n}{n}$ este normală cu media m şi dispersia $\frac{\sigma^2}{n}$ rezultă că variabila $Z = \frac{\bar{X}-m}{\sigma/\sqrt{n}}$ este normală cu media 0 şi dispersia 1. Ca urmare:

$$P\left(-a \leq Z \leq a\right) = \Phi\left(a\right) - \Phi\left(-a\right) = 2\Phi\left(a\right)$$

Dacă înlocuim pe Z cu $\frac{\frac{X_1+X_2+...X_n}{n}-m}{\sigma/\sqrt{n}}$ găsim:

$$P\left(\frac{X_1+X_2+...+X_n}{n} - a\frac{\sigma}{\sqrt{n}} \leq m \leq \frac{X_1+X_2+...+X_n}{n} + a\frac{\sigma}{\sqrt{n}}\right) = 2\Phi\left(a\right)$$

Prin urmare intervalul intervalul $\left[\frac{x_1+x_2+...x_n}{n} - a\frac{\sigma}{\sqrt{n}}, \frac{x_1+x_2+...x_n}{n} + a\frac{\sigma}{\sqrt{n}}\right]$ este un interval de încredere pentru m cu nivelul de încredere $2\Phi\left(a\right)$. Introducând pragul de încredere ε, avem $1-\varepsilon = 2\Phi\left(a\right)$ sau $\Phi\left(a\right) = \frac{1-\varepsilon}{2}$. Am demonstrat deci:

Propoziţia 10.4 *Fie X o variabilă normală de dispersie cunoscută σ^2 şi de medie m necunoscută. Dacă $\varepsilon \in (0,1)$ şi $a \in \mathbb{R}_+$, atunci, la selectii de volum n, o conditie suficientă ca intervalul $\left[\overline{X} - a\frac{\sigma}{\sqrt{n}}, \overline{X} + a\frac{\sigma}{\sqrt{n}}\right]$ să fie interval de încredere de nivel $1-\varepsilon$ pentru media m, este ca a să verifice ecuaţia $\Phi\left(a\right) \geq \frac{1}{2}\left(1-\varepsilon\right)$.*

QED.

Observaţia 10.5 *La acelaşi prag de încredere ε, creşterea volumului n de selecţie conduce la un interval de încredere mai scurt.*

Exemplul 10.6 *Fie P o populatie normală de variantă (dispersie) cunoscută σ^2 şi de medie m necunoscută (de estimat). Considerăm selectii de volum fixat n. Vom găsi un interval de încredere,de nivel de încredere 95% pentru medie, dacă alegem astfel pe a încât $\Phi\left(a\right) \geq \frac{1}{2}\frac{95}{100}$. Din tabelul pentru Φ găsim $a \geq 1,96$. Deci un interval de încredere de nivel 95% va fi de forma:*

$$\left[\overline{X} - 1,96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1,96\frac{\sigma}{\sqrt{n}}\right].$$

Exemplul 10.7 *O firmă produce piese cilindrice de diametru ϕ =10 mm. Abaterile de la acest diametru impus respectă o lege normală de variatie (dispersie) egală cu 0,04 mm (practica a arătat acest lucru). Se face un sondaj pe 100 de piese şi se găseste că media de selectie (empirică) este de 10,01 mm. Să se găsească un interval de estimatie pentru media reală cu nivelul de încredere de 90%.*

Soluţie Aici n=100, σ=0,2, $\overline{X}_{(100)}$=0,01, (1-ε)=0,90, deci ε =0,10. Din tabelul funcţiei Φ găsim $\Phi(a) \geq \frac{90}{2 \cdot 100} = 0,45$ pentru $\alpha \geq 1,65$. Deci, un interval de estimatie pentru media reală este: $\left[10,01 - 1,65 \cdot \frac{0,2}{10}, \quad 10,01 + 1,65 \cdot \frac{0,2}{10}\right] = [9,977, \quad 10,043]$.

Ce informatie obtine de aici producătorul? El este sigur în proportie de 90% că abaterea medie de la diametru real ϕ =10 mm este de cel mult 0,043 mm.

10.1.2 Dispersia este necunoscută

Am văzut până acum că dacă dispersia unei populatii normale este cunoscută putem estima prin intervale de încredere media populatiei cu ajutorul v.a. normale standard $Z = \frac{\bar{X}-m}{\sigma/\sqrt{n}}$, unde $\overline{X}$ este media de selectie, iar m este media reală a populatiei.

Dacă media m nu este cunoscută atunci putem folosi punctul d) al teoremei precedente care spune că variabila

$$\begin{aligned} T &= \sqrt{(n-1)\,n}\frac{\bar{X}-m}{\sqrt{\left(X_1-\bar{X}\right)^2+\left(X_2-\bar{X}\right)^2+...+\left(X_n-\bar{X}\right)^2}} \\ &= \sqrt{n-1}\frac{\bar{X}-m}{S} \end{aligned}$$

are o distribuţie Student cu n-1 grade de libertate. Aici utilizat notaţia (vezi lecţia 9) $S^2 = \frac{\left(X_1-\bar{X}\right)^2+\left(X_2-\bar{X}\right)^2+...+\left(X_n-\bar{X}\right)^2}{n}$. Aşa cum se vede în lecţia 4, densitatea de probabilitate este simetrică faţă de x=0, deci pentru funcţia de repartiţie $F(t)$ avem relaţia $F(-t) = 1 - F(t)$. Această observatie ne ajută să folosim tabelul II pentru găsirea cuantilelor corespunzătoare acestei distributii. Pe coloana din stânga a tabelului avem gradele de libertate $\nu = n - 1$ (n volumul selectiei), pe prima linie orizontală avem valorile functiei F(t) de la 0,60 până la 0,999. Fie de aflat la $\nu = n - 1 = 4$ valoarea lui a astfel ca F(a)=0,40. Avem 1-F(a)=F($-a$)=0,60 şi pentru 0,60 avem cuantila în tabel: $-a$=0,271. Deci $a = -0,271$.

Să punem aceste rezultate în următoarea propozitie:

Propoziţia 10.8 *Fie P o populatie normală cu media m şi dispersia σ^2 necunoscute. Pentru orice n, pentru un prag $\varepsilon \in (0,1)$ şi $a \in \mathbb{R}_+$, o conditie suficientă ca intervalul*

$$\left[\overline{X} - a\frac{S}{\sqrt{n-1}}; \overline{X} + a\frac{S}{\sqrt{n-1}}\right]$$

să fie interval de încredere de nivel $1-\varepsilon$ (sau de prag ε) pentru media m, este ca a să fie cuantilă de ordin $1-\varepsilon/2$ a distributiei Student cu $n-1$ grade de libertate (adică $F(a) = 1-\varepsilon/2$).

Demonstraţie. Relaţia $P\left(\overline{X}-a\frac{S}{\sqrt{n-1}}\leq m\leq\overline{X}+a\frac{S}{\sqrt{n-1}}\right)=1-\varepsilon$ se mai poate scrie

$$\begin{aligned}P\left(-a\leq\sqrt{n-1}\frac{\bar{X}-m}{S}\leq a\right) &= 1-\varepsilon,\\ \textit{sau}\quad P\left(-a\leq T\leq a\right) &= 1-\varepsilon\end{aligned}$$

Dar

$$\begin{aligned}P\left(-a\leq T\leq a\right) &= F(a)-F(-a)=\\ F(a)-1+F(a) &= 2F(a)-1=2\left(1-\varepsilon/2\right)-1=1-\varepsilon\end{aligned}$$

QED.

Exemplul 10.9 *Presupunem că în exemplul precedent nu cunoastem dispersia 0,04 mm şi că o estimăm cu formula* $S^2=\frac{\left(X_1-\bar{X}\right)^2+\left(X_2-\bar{X}\right)^2+...+\left(X_n-\bar{X}\right)^2}{n}$ *găsind-o ca fiind egală cu 0,09 mm. Avem* $1-(\varepsilon/2)=0,95$*,* $n=100$*,* $S=0,3$ *(în cazul nostru). Cuantila corespunzătoare lui 0,95 o găsim din tabelul cu distribuţia Student. La* $\nu=n-1=99$ *nu găsim date dar putem folosi linia lui* $\nu=120$*, deoarece cuantilele vecine diferă putin unele de altele (pentru acelasi prag bineînteles). Aici găsim a=1,1658. Cu acest a găsit, intervalul de încredere va fi:* $\left[10{,}01\text{-}1{,}658\frac{0,3}{\sqrt{99}}\ ;\ 10{,}01+1{,}658\frac{0,3}{\sqrt{99}}\right]$*, adică [9,9798,10,0599]. Să observăm că a=1,65 pentru situatia când am folosit v.a. Z şi a=1,658 pentru situatia când am folosit v.a. T. Acest lucru se explică, deoarece pentru n mare (mai mare ca 40), în cazul nostru 100, cele două cuantile diferă foarte putin.*

10.2 Intervale de încredere pentru dispersie

Dacă media m a variabilei aleatoare X este cunoscută, atunci putem folosi punctul b) al teoremei precedente care spune că variabila $\left(\frac{X_1-m}{\sigma}\right)^2+\left(\frac{X_2-m}{\sigma}\right)^2+...+\left(\frac{X_n-m}{\sigma}\right)^2$ este de tip $H\left(n\right)$ iar dacă media m nu este cunoscută putem folosi punctul c) al teoremei, anume că variabila aleatoare $\left(\frac{X_1-\bar{X}}{\sigma}\right)^2+\left(\frac{X_2-\bar{X}}{\sigma}\right)^2+...+\left(\frac{X_n-\bar{X}}{\sigma}\right)^2$ este de tip $H\left(n-1\right)$, în scopul de a determina intervale de încredere pentru dispersie.

$$\frac{nS^2}{\sigma^2}=\chi^2_{(n-1)} \tag{10.2}$$

adică $\frac{nS^2}{\sigma^2}$ este v.a χ^2 cu $n-1$ grade de libertate. Dacă în loc de S^2 se foloseste estimatorul nedeplasat $S'^{\,2}$ se obtine formula

$$\frac{(n-1)S'^{\,2}}{\sigma^2}=\chi^2_{(n-1)} \tag{10.3}$$

Teorema 10.10 *Fie* $\varepsilon \in (0, 1)$. *Un interval de încredere de nivel* $100(1 - \varepsilon)$ *procente pentru dispersia* σ^2 *a unei populatii normale cu media cunoscută* m, *în cazul selectiilor de volum* n *, este*

$$\left[\frac{(n-1)S'^{\,2}}{b}, \frac{(n-1)S'^{\,2}}{a}\right] \tag{10.4}$$

unde a este cuantila de ordin $\varepsilon/2$ şi b este cuantila de ordin $1-(\varepsilon/2)$ a distributiei χ^2 cu $n-1$ grade de libertate.

Demonstratie Notăm cu $F(t)$ functia de repartitie a v.a. $\chi^2_{(n-1)}$. Avem $F(a) = \varepsilon/2$ şi $F(b) = 1 - (\varepsilon/2)$. Atunci $P\left(a \le \frac{(n-1)S'^{\,2}}{\sigma^2} \le b\right) = F(b) - F(a) = 1 - (\varepsilon/2) - (\varepsilon/2) = 1 - \varepsilon$. Dar, din $a \le \frac{(n-1)S'^{\,2}}{\sigma^2} \le b$ găsim că $\frac{(n-1)S'^{\,2}}{b} \le \sigma^2 \le \frac{(n-1)S'^{\,2}}{a}$. Prin urmare, intervalul din (10.4) acoperă pe σ^2 cu probabilitatea $1 - \varepsilon$.

QED.

Exemplul 10.11 *Media erorilor de măsurare a lungimilor unor baghete metalice este de 3 mm. Presupunem că aceste erori respectă legea normală cu media 3 mm şi dispersia necunoscută. Se face o selectie de volum 4: {-1, 4, 4, 1}. Se cere un interval de estimatie pentru* σ^2 *cu pragul de încredere de 90%.*

Soluţie În cazul nostru aplicăm Teorema 10.9 cu 1-ε =0,90, deci ε =0,10. Căutăm cuantilele pentru $\varepsilon/2$ =0,05 şi 1-($\varepsilon/2$)=0,95, când n-1=3 (grade de libertate). Găsim a=0,351846 şi b=7,81473 în Tabelul III. Calculăm acum $S'^{\,2} = \frac{1}{3}\left((-1-3)^2 + (4-3)^2 + (4-3)^2 + (1-3)^2\right) = \frac{22}{3}$. Intervalul va fi deci $\left[\frac{22}{7{,}81}; \frac{22}{0{,}35}\right]$ = [2,81;62,85]. Se observă că intervalul este destul de mare, deci precizia pentru σ^2 este mică, chiar dacă apare cu probabilitate mare!

10.3 Intervale de încredere pentru cîtul a două dispersii

Fie acum două populatii distincte, normal distribuite. Facem o selectie de volum n_1 din prima populatie şi o selectie de volum n_2 din a doua populatie. Stim din formula (10.3) că

$$\frac{(n_1-1)S_1'^{\,2}}{\sigma_1^2} = \chi^2_{(n_1-1)} \tag{10.5}$$

$$\frac{(n_2-1)S_2'^{\,2}}{\sigma_2^2} = \chi^2_{(n_2-1)}$$

unde $\sigma_1, S_1'^{\,2}$ şi $\sigma_2, S_2'^{\,2}$ sunt dispersiile şi dispersiile de selectie modificate pentru cele două populatii. Notăm cu $\nu_1 = n_1 - 1$ şi cu $\nu_2 = n_2 - 1$. Notăm cu F (de la Fischer) v.a.

$$\frac{S_1'^2/\sigma_1^2}{S_2'^2/\sigma_2^2} = \frac{\left[\chi^2_{(\nu_1)}/\nu_1\right]}{\left[\chi^2_{(\nu_2)}/\nu_2\right]} \tag{10.6}$$

Această v.a. are o densitate de probabilitate ce depinde de doi parametri ν_1 şi ν_2 iar formula ei este complicată din punct de vedere matematic (vezi lecţia 4). Ea apare ca un cât de v.a. χ^2, înmultit cu un număr care depinde de ν_1 şi ν_2, adică ν_2/ν_1. Vom mai nota o asemenea variabilă F_{ν_1,ν_2} pentru a pune în evidenţă cei doi parametri de care depinde. Tabelul IV ne furnizează fractilele acestei distributii numai pentru ordinele 0,95; 0,975 şi 0,99. Pe coloane apar valorile parametrului ν_1 şi pe linii apar valorile parametrului ν_2. De exemplu, pentru $n_1 = 10$, $n_2 = 6$, $\nu_1 = 9$, $\nu_2 = 5$, presupunem că după selectie am obtinut F=7. Ne uităm la cuantila de ordin 0,95 şi găsim valoarea 3,48. Valoarea selectiei, 7, este mai mare decât 3,48, deci cade în partea opusă, adică în partea cu probabilitatea 5%. Prin urmare "inferenta" noastră asupra câtului $\frac{S_1'^2}{S_2'^2}$ *nu* este adevărată cu 95% probabilitate. În practică este utilă relatia:

$$P\left(F_{(\nu_1;\nu_2)} \geq c\right) = P\left(F_{(\nu_2;\nu_1)} \leq \frac{1}{c}\right) \tag{10.7}$$

Aici $F_{(\nu_2;\nu_1)}$ se numeste *inversa* v.a. $F_{(\nu_1;\nu_2)}$.

Din aceste observaţii rezultă imediat:

Teorema 10.12 *Avem relaţia*

$$\begin{aligned} P\left(\frac{aS_2'^2}{S_1'^2} \leq \frac{\sigma_2^2}{\sigma_1^2} \leq \frac{bS_2'^2}{S_1'^2}\right) &= P\left(a \leq \frac{S_1'^2/\sigma_1^2}{S_2'^2/\sigma_2^2} \leq b\right) = \\ F_{\nu_1,\nu_2}(b) - F_{\nu_1,\nu_2}(a) &= 1-\varepsilon \end{aligned}$$

dacă a şi b sunt alese astfel ca $F_{\nu_1,\nu_2}(b) = 1 - \frac{\varepsilon}{2}$ *şi* $F(a) = \frac{\varepsilon}{2}$. *In aceste condiţii un interval de încredere pentru* $\frac{\sigma_1^2}{\sigma_2^2}$, *cu nivelul de încredere* $(1-\varepsilon)$ *este* $\left[\frac{aS_2'^2}{S_1'^2}, \frac{bS_2'^2}{S_1'^2}\right]$.

QED.

10.4 Intervale de încredere în cazul unor selecţii mari

Dacă $f(x,\theta)$ este densitatea de probabilitate a variabilei aleatoare X atunci din $\int_{-\infty}^{\infty} f(x,\theta)dx = 1$ rezultă prin derivare în raport cu θ că $\int_{-\infty}^{\infty} \frac{\partial f}{\partial \theta}(x,\theta)\,d\theta = 0$ sau $\int \frac{\partial \ln(f(x,\theta))}{\partial \theta} f(x,\theta)dx = 0$. Deci variabila aleatoare $\frac{\partial \ln(f(x,\theta))}{\partial \theta}$ are media 0 şi dispersia $\int_{-\infty}^{\infty}\left(\frac{\partial \ln(f(x,\theta))}{\partial \theta}\right)^2 f(x,\theta)dx$. Presupunând că dispersia este finită , rezultă din legea limită centrală (vezi lecţia 5) că pentru n

mare, variabila aleatoare

$$\frac{\frac{\partial \ln(f(X_1,\theta))}{\partial \theta}+\frac{\partial \ln(f(X_1,\theta))}{\partial \theta}+...+\frac{\partial \ln(f(X_1,\theta))}{\partial \theta}}{\sqrt{n\int_{-\infty}^{\infty}\left(\frac{\partial \ln(f(x,\theta))}{\partial \theta}\right)^2 f(x,\theta)dx}}$$

unde X_1, X_2,... X_n sunt variabilele de selecţie asociate cu X, are o distribuţie aproximativ normală cu media 0 şi dispersia 1. Avem deci:

$$\Pr ob\left(a<\frac{\frac{\partial \ln(f(X_1,\theta))}{\partial \theta}+\frac{\partial \ln(f(X_1,\theta))}{\partial \theta}+...+\frac{\partial \ln(f(X_1,\theta))}{\partial \theta}}{\sqrt{n\int_{-\infty}^{\infty}\left(\frac{\partial \ln(f(x,\theta))}{\partial \theta}\right)^2 f(x,\theta)dx}}<b\right)\approx\int_a^b\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx \qquad (10.8)$$

Un interval de încredere pentru θ,cu nivelul de încredere α, se poate obţine pentru n mare astfel:

- Se determină prin n experienţe independente valorile x_1,.. x_n pentru X_1,.. X_n.
- Se determină a şi b astfel ca $\int_a^b \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=\alpha$.
- Din formula 10.8 rezultă că mulţimea valorilor θ care verifică inegalitatea

$$a<\frac{\frac{\partial \ln(f(X_1,\theta))}{\partial \theta}+\frac{\partial \ln(f(X_1,\theta))}{\partial \theta}+...+\frac{\partial \ln(f(X_1,\theta))}{\partial \theta}}{n\cdot\sqrt{\int_{-\infty}^{\infty}\left(\frac{\partial \ln(f(x,\theta))}{\partial \theta}\right)^2 f(x,\theta)dx}}<b$$

este o mulţime de încredere pentru θ cu încrederea α. In unele cazuri această mulţime este un interval.

10.5 Rezumat

1. Fie $\varepsilon>0$, $\varepsilon\in(0,1)$. Vom numi interval de încredere de prag ε (sau de nivel de încredere $1-\varepsilon$) pentru parametrul λ două statistici $\Lambda_1=f_1(X_1,...,X_n)$ şi $\Lambda_2=f_2(X_1,...,X_n)$ astfel încât $P(\Lambda_1\leq\lambda\leq\Lambda_2)\geq 1-\varepsilon$.

Pentru selectii efective $x_1,...,x_n$, vom nota valoarea statisticii Λ_1 cu $\widehat{\lambda}_1$şi a statisticii Λ_2 cu $\widehat{\lambda}_2$. Intervalul numeric $\left[\widehat{\lambda}_1,\widehat{\lambda}_2\right]$ este considerat încă ca interval de încredere de nivel $1-\varepsilon$ (se mai spune de prag ε) pentru parametrul estimat λ.

2. Dacă există două statistici $Y=f(X_1,...,X_n)$ şi $Z=g(X_1,...,X_n)$ astfel încât v.a. $T=\frac{Y-\mu}{Z}$ să fie normală redusă sau Student cu d grade de libertate şi t_ε un număr pozitiv astfel încât P$(|\mathrm{T}|>t_\varepsilon)\leq\varepsilon$, atunci $[Y-t_\varepsilon|\mathrm{Z}|,Y+t_\varepsilon|\mathrm{Z}|]$ este un interval de încredere de prag ε pentru media μ, adică:

$$P\left(Y-t_\varepsilon|\mathrm{Z}|\leq\mu\leq Y+t_\varepsilon|\mathrm{Z}|\right)\geq 1-\varepsilon$$

• Fie acum de estimat σ^2 a v.a. X. Presupunem că am găsit o statistică Y a.î. T=$\frac{Y}{\sigma^2}$ are distributia χ^2 cu d grade de libertate şi fie două numere t'_ε şi t''_ε astfel încât P$(T \geq t'_\varepsilon) \leq \frac{\varepsilon}{2}$ şi P$(T \leq t''_\varepsilon) \leq \frac{\varepsilon}{2}$. Atunci $\left[\frac{Y}{t'_\varepsilon}, \frac{Y}{t''_\varepsilon}\right]$ este un interval de încredere pentru σ^2 de prag ε: P$\left(\frac{Y}{t'_\varepsilon} \leq \sigma^2 \leq \frac{Y}{t''_\varepsilon}\right) \geq 1-\varepsilon$.

Se alege Y a.î. să aibă cât mai multe grade de libertate.

FORMULE UTILIZATE FRECVENT

In formulele de mai jos nivelul de încredere este $1-\varepsilon$, iar rezultatele a n măsurători independente ale unei caracteristici numerice cu distribuţie normală sunt $x_1, x_2, ...x_n$.

1. Un interval de încredere pentru media m a unei variabile aleatoare normale, dacă se cunoaşte dispersia σ^2 este:

$$\left[\frac{x_1+x_2+...x_n}{n} - a\frac{\sigma}{\sqrt{n}}, \frac{x_1+x_2+...x_n}{n} + a\frac{\sigma}{\sqrt{n}}\right]$$

unde a se alege astfel ca $\Phi(a) = 0,5 - \frac{\varepsilon}{2}$.

2. Un interval de încredere pentru media m a unei v.a. normale, dacă nu se cunoaşte dispersia, este:

$$\left[m^* - a\frac{s^*}{\sqrt{n-1}}, m^* + a\frac{s^*}{\sqrt{n-1}}\right]$$

unde $m^* = \frac{x_1+x_2+...x_n}{n}$, $s^* = \sqrt{\frac{(x_1-m^*)^2+(x_2-m^*)^2+...(x_n-m^*)^2}{n}}$, iar a se alege astfel ca $F(a) = 1-\frac{\varepsilon}{2}$, F fiind funcţia de repartiţie a unei variabile Student cu $n-1$ grade de libertate.

3. Un interval de încredere pentru dispersia σ^2 a unei v.a. normale este:

$$\left[\frac{(x_1-m^*)^2+...+(x_n-m^*)^2}{b}, \frac{(x_1-m^*)^2+...+(x_n-m^*)^2}{a}\right]$$

unde $F(a) = \frac{\varepsilon}{2}$, iar $F(b) = 1-\frac{\varepsilon}{2}$, F fiind funcţia de repartiţie a unei variabile χ^2 cu $n-1$ grade de libertate.

4. Un interval de încredere pentru câtul dispersiilor $\frac{\sigma_2^2}{\sigma_1^2}$a două v.a. independente este:

$$\left[a \cdot \frac{\frac{\sum_{i=1}^{n_2}(y_i-m'^*)^2}{n_2-1}}{\frac{\sum_{i=1}^{n_1}(x_i-m^*)^2}{n_1-1}}, b \cdot \frac{\frac{\sum_{i=1}^{n_2}(y_i-m'^*)^2}{n_2-1}}{\frac{\sum_{i=1}^{n_1}(x_i-m^*)^2}{n_1-1}}\right]$$

unde $m^* = \frac{x_1+x_2+...x_{n_1}}{n_1}$, $m'^* = \frac{y_1+y_2+...y_{n_2}}{n_2}$, n_1şi n_2 sunt volumele celor două selecţii, $\nu_1 = n_1-1$, $\nu_2 = n_2-1$, iar a şi b sunt alese astfel ca $1-\frac{\varepsilon}{2} = F_{\nu_1,\nu_2}(b)$, $\frac{\varepsilon}{2} = F_{\nu_1,\nu_2}(a)$

10.6 Exerciţii rezolvate

1. Atunci când se nasc 2 copii simultan (gemeni) probabiliatea ca ei să fie gemeni adevărati este λ. Se presupune că:

a) 2 gemeni adevărati au întotdeauna acelasi sex şi probabilitatea ca ei să fie băieti este $\frac{1}{2}$;

b) 2 gemeni falsi au sexe diferite şi probabilitatea ca unul dintre ei să băiat este $\frac{1}{2}$;

i)În cursul nasterii a 2 gemeni se consideră evenimentele: A= (2 băieti); B= (2 fete); C= (1 băiat şi 1 fată). Calculati în functie de λ probabilitătile p(A), p(B), p(C).

ii) În cursul a 1000 de nasteri se realizează evenimentul C de 328 de ori. Dati pentru λ un interval de încredere de prag ε=0,05.

iii) Observăm acum n nasteri de gemeni şi notăm cu Y_C numărul de realizări ale evenimentului C. Ce lege guvernează v.a. Y_C? Definiti cu ajutorul lui Y_C un esantion nedeplasat Z pentru λ. Calculati varianta lui Z. Dati pentru n mare o conditie independentă de λ şi suficientă pentru a putea defini cu ajutorul lui Z un interval de încredere de prag ε=0,05 a cărui lungime să fie mai mică decât un $a \in \mathbb{R}$, dat. Caz particular $a = \frac{1}{100}$.

Soluţie i) Notăm cu V evenimentul: << gemeni adevărati >> şi cu F: << gemeni falsi >>. Atunci $\mathrm{A}=(\mathrm{A}\cap\mathrm{V})\cup(\mathrm{A}\cap\mathrm{F})$ şi $p(\mathrm{A})=p(\mathrm{V})\cdot p_V(\mathrm{A})+p(\mathrm{F})\cdot p_F(\mathrm{A})=\lambda\cdot\frac{1}{2}+(1-\lambda)\frac{1}{4}=\frac{\lambda+1}{4}$. La fel $p(\mathrm{B})=\frac{\lambda+1}{4}$ şi $p(\mathrm{C})=\frac{1-\lambda}{2}$.

ii) Fie X v.a. care are valoarea 1, dacă se realizează evenimentul C, şi 0 altfel. Este clar că $\mathrm{M(X)}=\mu=\frac{1-\lambda}{2}$. $\overline{\mathrm{X}}=\frac{1}{n}Y_C=\frac{328}{1000}=0,328$. Cum $\mathrm{X}=\mathrm{X}_i$, avem că $\mathrm{X}_i^2=\mathrm{X}_i$, deci $\mathrm{S}^2=\frac{1}{n}\sum_i \mathrm{X}_i^2-(\overline{\mathrm{X}})^2=\overline{\mathrm{X}}\left(1-\overline{\mathrm{X}}\right)$. Cum $\mathrm{T}=\frac{\overline{\mathrm{X}}-\mathrm{M(X)}}{\sqrt{\mathrm{S}^2/n}}$ este practic normală redusă, egalitatea $\mathrm{P}(|\mathrm{T}|\geq 1,96)=0,05$ dă pentru μ intervalul de încredere cerut: $\overline{\mathrm{X}}-1,96\frac{\mathrm{S}}{\sqrt{n}}\leq\mu\leq\overline{\mathrm{X}}+1,96\frac{\mathrm{S}}{\sqrt{n}}$, sau $0,299\leq\frac{1-\lambda}{2}\leq 0,357$. De aici rezultă intervalul de încredere căutat pentru λ: $0,286\leq\lambda\leq 0,402$.

iii) Y_C este binomială cu $p=\frac{1-\lambda}{2}$. Avem deci $\mathrm{M(Y}_C)=\frac{n(1-\lambda)}{2}$ şi $\mathrm{D(Y}_C)=\frac{n(1-\lambda^2)}{4}$. Egalitatea $\lambda=1-\frac{2}{n}\mathrm{M(Y}_C)$ dă pentru λ estimatorul nedeplasat $\mathrm{Z}=1-\frac{2}{n}\mathrm{Y}_C$ de dispersie $\frac{4}{n^2}\mathrm{D(Y}_C)=\frac{1-\lambda^2}{n}$. Când $n\to\infty$, Y_C este practic gaussiană (normală), deci şi Z este la fel. Considerăm deci $\mathrm{T}=\frac{\mathrm{Z}-\lambda}{\sqrt{(1-\lambda^2)/n}}$ care este gaussiană redusă. Egalitatea $\mathrm{P}(|\mathrm{T}|\geq 1,96)=0,05$ ne permite să scriem $\mathrm{P}\left(|\mathrm{Z}-\lambda|\geq 1,96\sqrt{\left(1-\lambda^2\right)/n}\right)=0,05$, şi cum $\sqrt{\left(1-\lambda^2\right)/n}\leq\sqrt{1/n}$ avem $\mathrm{P}(|\mathrm{Z}-\lambda|\geq 1,96/\sqrt{n})<0,05$, de unde $\mathrm{Z}-\frac{1,96}{\sqrt{n}}\leq\lambda\leq\mathrm{Z}+\frac{1,96}{\sqrt{n}}$. Lungimea lui va fi mai mică decât a atunci când $a\geq\frac{2\times 1,96}{\sqrt{n}}$, sau $n\geq\left(\frac{3,92}{a}\right)^2$. Pentru $a=\frac{1}{100}$ găsim $n\geq 153664$.

2. Se măsoară forta de compresiune X (în Kg/cm^3) a cimentului din care sunt confectionati cilindri mici, limită de la care ei se sparg. Pentru n=10 cilindri se observă următoarele presiuni:

19,6	19,9	20,4	19,8	20,5
21,0	18,5	19,7	18,4	19,4

Presupunem că X are o lege gaussiană (normală).

i) Dati un interval de încredere de prag $\varepsilon = 0,1$ pentru M(X).

ii) Dati o estimare nedeplasată $\widehat{\sigma}^2$ pentru varianta σ^2 a v.a. X, găsiti apoi un interval de încredere de prag 0,1 pentru σ^2.

iii) Presupunem că σ^2=0,69. Găsiti pentru M(X) un nou interval de încredere de prag 0,1. Comparati cu rezultatul de la 1).

Solutie i) Calculăm $\overline{\mathrm{X}} = 19,72$ şi $n\mathrm{S}^2 = 6,0960$. T=$\frac{\overline{\mathrm{X}}-\mathrm{M(X)}}{\sqrt{\mathrm{S}^2/(n-1)}}$ este o v.a. Student cu $n-1 = 9$ grade de libertate, avem $\mathrm{P}(|\mathrm{T}| > t_\varepsilon) = 0,1$ pentru $t_\varepsilon = 1,833$. Intervalul de încredere cerut este deci $\overline{\mathrm{X}} - t_\varepsilon \frac{\mathrm{S}}{\sqrt{n-1}} \leq$M(X)$\leq \overline{\mathrm{X}} + t_\varepsilon \frac{\mathrm{S}}{\sqrt{n-1}}$, sau $19,243 \leq$M(X)$\leq 20,197$.

ii) O estimatie nedeplasată a lui σ^2 este $\widehat{\sigma}^2 = \frac{n\mathrm{S}^2}{n-1} = 0,6773$. Pe de altă parte stim că U=$\frac{n\mathrm{S}^2}{\sigma^2}$ are distributia χ^2(Pearson) cu $n-1 = 9$ grade de libertate. Avem deci $\mathrm{P}(\mathrm{U} > t'_\varepsilon) = 0,05$ pentru $t''_\varepsilon = 3,33$. De aici găsim pentru σ^2 intervalul de încredere: $\frac{n\mathrm{S}^2}{t'_\varepsilon} \leq \sigma^2 \leq \frac{n\mathrm{S}^2}{t''_\varepsilon}$, adică $0,36 \leq \sigma^2 \leq 1,83$.

iii) Dacă stim dispersia $\sigma^2 = 0,69$ putem folosi faptul că $\overline{X}$ este gaussiană $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$ şi deci $\mathrm{T}' = \frac{\overline{\mathrm{X}}-\mathrm{M(X)}}{\sqrt{\sigma^2/n}}$ este gaussiană redusă. Egalitatea $\mathrm{P}(|\mathrm{T}'| > t'_\varepsilon) = 0,1$ ne conduce la $t'_\varepsilon = 1,6449$. Prin urmare, găsim un interval de încredere de prag $\frac{1}{10}$ pentru M(X): $\overline{\mathrm{X}} - t'_\varepsilon\sqrt{\sigma^2/n} \leq$M(X)$\leq \overline{\mathrm{X}} + t'_\varepsilon\sqrt{\sigma^2/n}$, sau $19,287 \leq$M(X)$\leq 20,153$. Acest interval este mai mic decât acela găsit la 1) deoarece acum avem dispersia dată.

10.7 Exerciţii

1. Notăm cu X vârsta în ani la care un om devine bunic. Presupunem că X are distributia normală cu varianta 225. 9 persoane luate la înâmplare au declarat că au devenit bunici la: 42, 56, 68, 56, 48, 36, 45, 71 şi 64 ani.

a) Calculati media şi dispersia de selectie.

b) Găsiti un interval de încredere de 80% pentru medie.

c) Găsiti un interval de încredere de 95% pentru medie.

2. În cadrul unui proces de estimare a mediei unei populatii oarecare, un statistician vrea ca probabilitatea ca media de selectie să difere de media adevărată cu mai puţin de 0,2σ să fie mai mare de 0,95.

a) Ce volum de selectie trebuie să folosească?

b) Dar dacă volumul de selectie este 100, care este marja de aproximare (în σ unităti) a mediei reale cu media de selectie, pentru ca să se obtină un prag de încredere de 0,95?

c) Dar dacă se ştie că populaţia este normală care trebuie sa fie volumul de selecţie ca

$$\Pr ob\left(\left|\bar{X} - m\right| \leq 0,2 \cdot \sigma\right) \geq 0,95$$

?

3. Fie distributia student T cu 12 grade de libertate.
a) găsiti fractile pentru 0,10; 0,60 şi 0,95.
b) găsiti media şi dispersia.
c) P(T<-0,695).
d)P(-2,179<T<1,356).

4. Fie distributia T cu 6 grade de libertate.
a) găsiti P(T$\geq$ 1).
b) găsiti fractila pentru 0,20.

5. Fie P o populatie distribuită normal. Se face selectia: -2, -1, 0, 0, 1, 2, 2, 3, 0. Să se găsească un interval de încredere pentru media populatiei cu pragul de încredere de 80%, folosind distributia T.

6. Dintr-un lot de 100 piese, 4 sunt găsite defecte. Presupunem că procesul de productie se comportă ca un proce de tip Bernoulli. Determinati intervale de încredere de 95% şi de 99% pentru probabilitatea ca luând o piesă la întâmplare ea să fie defectă.

7. Se urmăreste indicatorul Dow-Jones (indicator la bursa americană) la stocurile industriale de la o zi la alta. Se presupune că acesta variază după o distributie normală. Se face un sondaj pe parcursul a 81 de zile şi se obtine o medie de selectie de 0,20 şi o dispersie de selectie de 1,50. Găsiti un interval de încredere pentru medie de 90%.

8. Dintr-o populatie noramlă (m=24, σ^2=9) se ia un esantion de volum 8. Notăm cu W suma pătratelor celor 8 valori standardizate $\left(Z_i^2 = \left(\frac{X_i-m}{\sigma}\right)^2\right)$. *Găsiti: a) P(W$\geq$20,09); b) P(2,73<W<5,07); c) P(W>2,18).*

11. Pentru distributia F cu 5 grade de libertate la numărător şi cu 8 grade de libertate la numitor, găsiti : a) fractilele de ordin 0,025 şi 0,99; b) P(F $\leq$3,69); c) P(F $\geq$1,22).

12. Durata de viaţă a unui bec electric este o variabila aleatoare normala. Testul pe 16 becuri a arătat o valoare medie de viaţă de 3000 ore şi o abatere standard $\sigma^* = 20$. *Să se determine intervale de încredere pentru medie şi abatere standard cu pragul de risc* $\varepsilon = 0,1$.

13. Un producător de rulmenţi pretinde că diametrul mediu al rulmenţilor, în mm este de 10 mm cu o dispersie de 10^{-4}. *Admitem că diametrul este o v.a. normală. Pe un lot de* 20 *rulmenţi măsuraţi s-a găsit că diametrul mediu* $9,98$ *mm şi o dispersie empirică* $0,0002$. *Să se determine intervale de încredere pentru diametrul mediu şi dispersie cu încrederea* $0,99$. *Valorile pretinse de fabricant se află în aceste intervale?*

Lecţia 11

Ipoteze statistice. Teste statistice

11.1 Ipoteze şi testarea lor

În continuare vom face ipoteze asupra parametrilor unor populatii, stiind în prealabil clasa de distributii din care fac parte (de exemplu: normală, Bernoulli, Poisson, etc.). Vom folosi rezultatele obtinute în Lectia 10 asupra estimării prin intervale de încredere a unor parametri remarcabili pentru distributii cunoscute (media şi dispersia pentru populatii normale, de exemplu).

O ipoteză statistică este o ipoteză făcută asupra unor însusiri statistice ale unei populatii P.Ea este *simplă*, dacă se referă la întreaga informatie care determină distributia populatiei, de exemplu ipoteza:

H: populatia este normală de medie m=10 şi dispersie σ^2=225, sau

H: populatia este Bernoulli cu p=0,3.

Ipoteza poate fi *compusă* dacă se referă numai la o parte din informatiiile ce ar putea determina distributia populatiei. Iată un exemplu de distributie compusă:

H: populatia este normală de medie 40, sau

H: populatia este Poisson (nu facem nici o ipoteză asupra mediei λ).

In cazul ipotezelor compuse, ceilalti parametrii care impreuna cu cei testati ar duce la determinarea completa a distributiei, se estimeaza dintr-o selectie (sau mai multe) facuta asupra populatiei.

O ipoteză poate fi *exactă*,de exemplu ipoteza H: media populatiei Poisson este λ=3, sau poate fi *inexactă*: H: media populatiei normale este $m \geq 5$.

În aparentă noi lucrâm cu o singură ipoteză H. De fapt lucrăm cu două ipoteze: H=H_0 şi H_1, ipoteza contrară ipotezei H_0. În cele ce urmează vom considera două ipoteze alternative H_0 şi H_1. Nu intotdeauna ipoteza H_1 reprezinta negatia logica obisnuita a ipotezei H_0. De exemplu, H_0: media populatiei este m=30, H_1: media populatiei s-a micsorat, adica este m<30.

Operatia de comparare a două ipoteze statistice în lumina informatiilor furnizate de selectie se numeste *test statistic.* Dacă testul statistic se referă la unul sau mai multi parametri ce apar în legea ce defineste populatia spunem că testul este *parametric.* Dintre cele două ipoteze H_0 şi H_1, una dintre ele, notată cu H_0, ocupă locul central: testăm pe H_0 "împotriva" alternativei H_1. La finalul testării statisticianul, fie acceptă pe H_0, fie că respinge pe H_0 în favoarea ipotezei H_1. Oricum el trebuie să ia "*o decizie*". Fiecare test statistic implică o statistică de selectie, adică o functie continuă de tipul $g(X_1, ..., X_n)$, unde n este volumul selectiei, iar $X_1, ..., X_n$ sunt variabilele de selectie. Anumite valori ale statisticii conduc la acceptarea ipotezei H_0, alte valori ale ei conduc la respingerea acestei ipoteze. Vom vorbi deci de un *domeniu de respingere (de neacceptare). Regula de decizie* este dată de fapt de specificarea domeniului de respingere al ipotezei H_0, deoarece se consideră că domeniul complementar domeniului de respingere este exact domeniul de acceptare pentru ipoteza H_0.

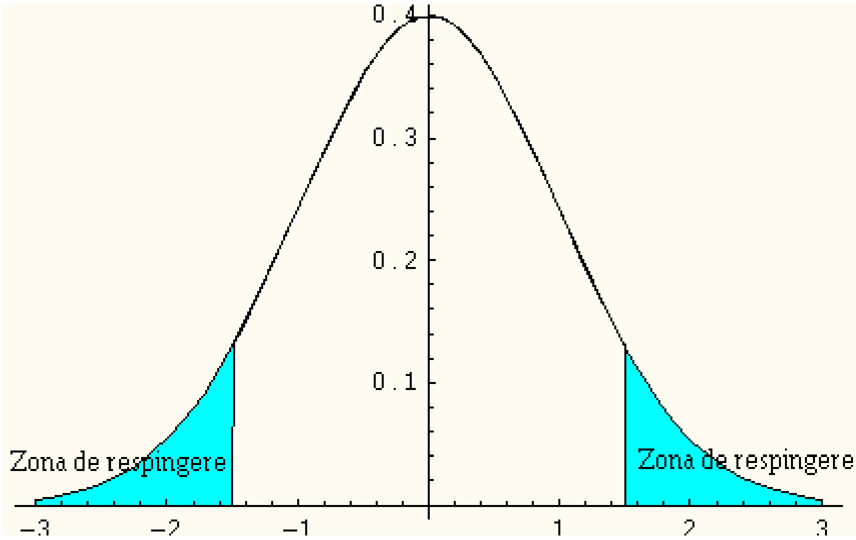
Mai exact, daca ipoteza H_0, numita si *ipoteza nula*, este adevarata atunci, in urma unei selectii concrete, este foarte probabil ca valoarea calculata pentru v.a. $g(X_1, ..., X_n)$ sa se gaseasca intr-un interval "de probabilitate mare". Acest lucru se intampla deoarece statistica $g(X_1, ..., X_n)$ are o repartitie bine determinata de ipoteza H_0 si eventual de unele estimari facute in urma selectiei concrete ce apare in problema. Deci noi trebuie sa stabilim o *zona de acceptare*, adica o submultime A din $\mathbb{R}$ a.i. probabilitatea ca o valoare a v.a. $g(X_1, ..., X_n)$ sa apartina multimii A sa fie destul de mare (de obicei se ia ca fiind $\geq 0,9$). Multimea $\mathbb{R}\backslash\backslash\mathrm{A}$ se zice *zona de respingere* si probabilitatea ca v.a. $g(X_1, ..., X_n)$ sa ia o valoare in $\mathbb{R}\backslash\backslash\mathrm{A}$ este foarte mica ($\leq 0,1$). Numarul $\varepsilon = prob\,(g(X_1, ..., X_n) \in \mathbb{R}\backslash\backslash A)$ se numeste *prag de semnificatie* pentru testul pe care il vom constitui, iar statistica $g(X_1, ..., X_n)$ este o v.a. care depinde de v.a. de selectie $X_1, ..., X_n$ si este legata de ipoteza H_0. De exemplu, daca H_0 se refera la media populatiei, $g(X_1, ..., X_n)$ va fi media de selectie standardizata, adica $Z = \frac{\overline{X}-m}{\sigma/\sqrt{n}}$, unde m este media ce rezulta din ipoteza H_0, σ este deviatia standard (presupusa cunoscuta), n este volumul selectiei, iar $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$, este v.a. media de selectie. *Ipoteza alternativa*, H_1 este ipoteza ce rezulta *natural in urma negarii ipotezei H_0*. De exemplu, daca H_0 este "$m = 30$" si noi stim sigur ca media *nu poate sa creasca* in urma experimentului ce apare in problema, atunci H_1 va fi "$m < 30$". Daca nu stim nimic despre modul in care se schimba media, este natural sa consideram ipoteza alternativa ca fiind "$m \neq 30$", adica "$m < 30$" sau "$m > 30$". Iata deci cum functioneaza in general un *test parametric* referitor la parametrul θ al unei populatii P:

1) construim ipotezele H_0 (*ipoteza nula*) si H_1 (*ipoteza alternativa*) asupra parametrului θ.

2) construim v.a. $g(X_1, ..., X_n)$ care are o distributie (repartitie) cunoscuta daca consideram pe H_0 adevarata.

3) precizam pragul de semnificatie ε pentru v.a. $g(X_1, ..., X_n)$ (ε este mic, $\leq 0,1$).

4) reprezentam grafic (schematic si nu exact) zonele de respingere si respectiv de acceptare, $\rho(g)$ =densitatea de probabilitate a v.a. g.



Zona tipica de respingere a unui test

Aria hasurata este $\varepsilon/2 + \varepsilon/2 = \varepsilon = prob(g$ sa ia valori in zona de respingere).

5) calculam $g_{calc.} = g(x_1, ..., x_n)$ pentru valorile efective ale unei selectii furnizate de problema. Daca $g_{calc} \in \{$zonei de acceptare$\}$ vom spune ca acceptam ipoteza H_0 cu pragul de semnificatie ε. Daca $g_{calc} \in \{$zonei de respingere$\}$ acceptam ipoteza alternativa H_1 cu pragul de semnificatie ε.

11.1.1 Testul Z privind media unei populatii normale cu dispersia cunoscuta σ^2

Vrem sa testam ipoteza:

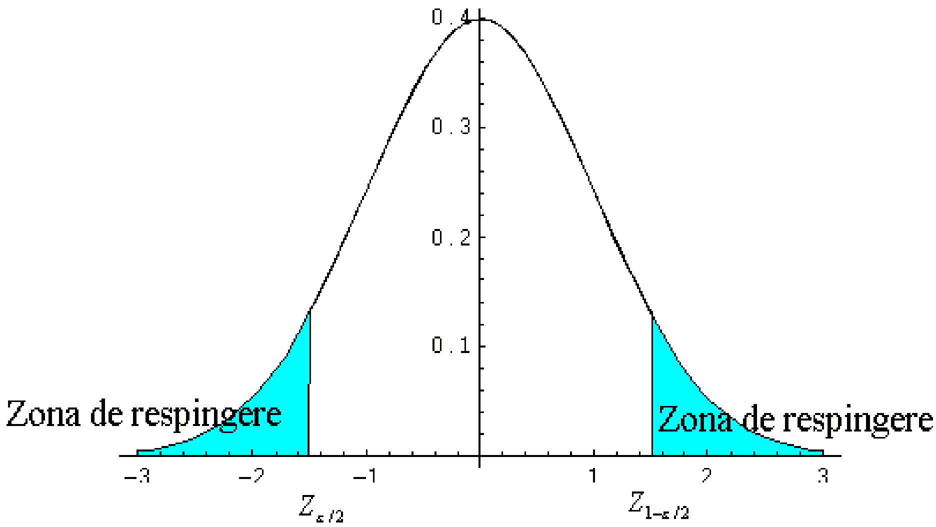
• H_0: $m = m_0$, m_0 specificat (m este media populatiei)

H_1: $m \neq m_0$

• Consideram statistica $Z = \frac{\overline{X} - m_0}{\sigma/\sqrt{n}}$. Stim ca Z este o v.a. normala redusa (are media 0 si dispersia 1), pentru n mare. Aici σ este precizat de problema.

• Fie $\varepsilon \in (0; 0,1]$ pragul de semnificatie ales ($\varepsilon = 0,05; 0,01$, etc.)

•



Zona de respingere pentru un test bilateral

Calculam pe $Z_{\varepsilon/2}$ ca fiind cuantila de ordin $\varepsilon/2$, adica $F(Z_{\varepsilon/2}) = 1 - \varepsilon/2$ (vezi TABELUL I)

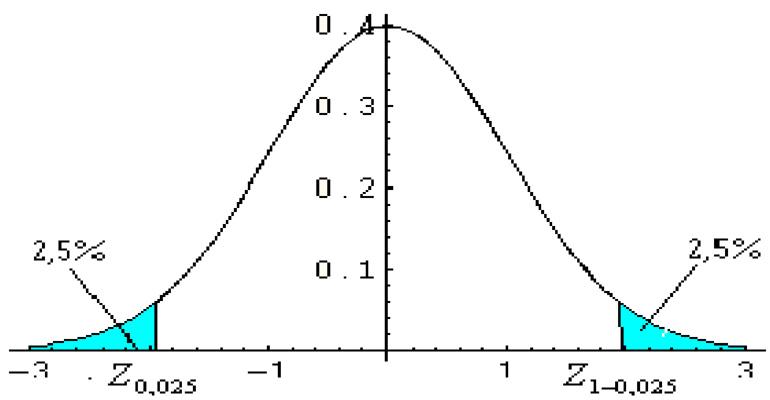
• calculam $Z_{calc} = \frac{\overline{X}-m_0}{\sigma/\sqrt{n}}$ pentru selectia din problema.

Daca $|Z_{calc}| \leq Z_{\alpha/2}$, acceptam ipoteza H_0 cu pragul de semnificatie ε.

Daca $|Z_{calc}| > Z_{\alpha/2}$, respingem ipoteza H_0 (acceptam H_1) cu pragul de semnificatie ε.

Exemplul 11.1 *Testati cu un prag (nivel) de semnificatie de 5% daca o selectie de volum 1, $x_1 = 172$ provine dintr-o populatie normala cu media $m = 150$ si dispersia fixata (cunoscuta) $\sigma^2 = 100$.*

Solutie *H_0: $m = 150$; H_1: $m \neq 150$; $Z = \frac{\overline{X}-m}{\sigma/\sqrt{n}} = \frac{X-m}{\sigma}$, deoarece $n = 1$ si $\overline{X} = X_1 = X$.* $F(Z_{0,025}) = 1 - 0,025 = 0,975$.



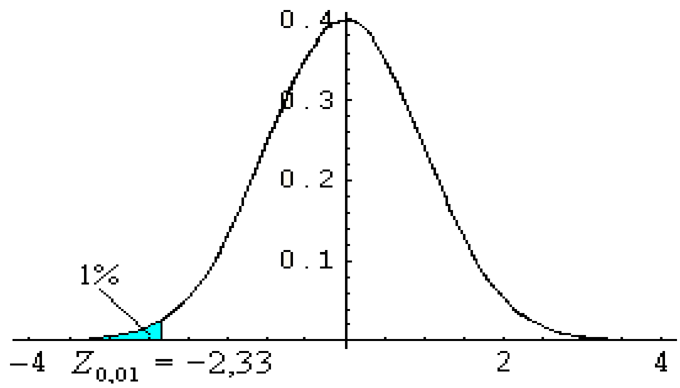
Deci, din TABELUL I gasim ca $Z_{0,025}$ (cuantila de ordin 0,025) este 1,96.

$Z_{calc} = \frac{172-150}{10} = 2,2$. *Deoarece* $|Z_{calc}| > 1,96$ *respingem* H_0 *(acceptam* H_1*) cu pragul de semnificatie de 5%. Adica este putin probabil ca selectia sa provina dintr-o populatie normala cu media* $m = 150$ *si dispersia* $\sigma^2 = 100$.

Aici am folosit un test bilateral (zona de respingere este simetrica fata de origine, adica are "2 cozi") deoarece pentru H_1, m *poate fi tot asa de bine* < 150 *sau* > 150.

Exemplul 11.2 *Testati cu un prag de semnificatie* $\varepsilon = 1\%$ *daca selectia de volum 1,* $x_1 = 54$, *a fost facuta dintr-o populatie normala cu media* $m = 65$ *si dispersia* $\sigma^2 = 30$, *sau daca media este mai mica decat* 65.

Solutie H_0: $m = 65$; H_1: $m < 65$. *Vom avea deci un test unilateral (la stanga, cu o singura "coada").* $Z = \frac{\overline{X}-m}{\sigma/\sqrt{n}} = \frac{X-m}{\sigma}$.



Zona de respingere pentru un test unilateral

Deoarece in tabelele statistice se dau numai valorile functiei de repartitie normala, $F(z)$, *pentru* $z \geq 0$, *va trebui sa folosim proprietatile de simetrie ale densitatii de probabilitate* $\rho(z)$. *Avem ca* $P(Z < Z_{0,01}) = F(Z_{0,01}) = 0,01$. *Deci* $F(-Z_{0,01}) = P(Z < -Z_{0,01}) = 1 - P(Z < Z_{0,01}) = 1 - 0,01 = 0,99$, *deci, din TABELUL I gasim ca* $-Z_{0,01} = 2,33$, *adica* $Z_{0,01} = -2,33$.

Calculam $Z_{calc} = \frac{54-65}{\sqrt{30}} = -2,01$. *Cum* $-2,01 > -2,33$ *vom accepta ipoteza* H_0 *cu pragul de semnificatie de 1%. Cum pragul este mic si* Z_{calc} *este foarte "aproape" de valoarea critica* $-2,33$ *statisticianul are dubii serioase asupra rezultatului si va trebui sa considere si alta selectie si sa foloseasca un test cu semnificatie mai mare, de exemplu de 5% pentru a fi "mai sigur" de concluzia pe care o da.*

Exemplul 11.3 *Din 100 de seminte plantate 83 au germinat. Folositi aproximarea distributiei binomiala cu o distribatie normala pentru a testa pretentia comerciantului ca 90% din* *seminte germineaza. Folositi doua teste: unul cu pragul de semnificatie de 5%, altul cu pragul de 1%.*

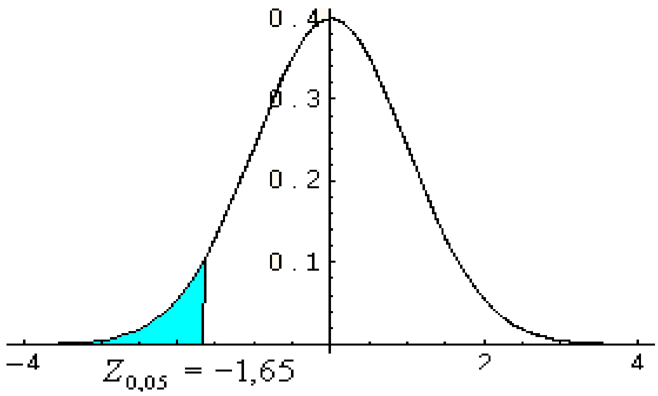
Solutie *Fie* X *v.a. care numara cate seminte au germinat din cele* n. $X \sim Bin(n,p)$, *unde* $n = 100$.

H_0: $p = 0,9$ *(rata de germinare este de 90%).*

H_1: $p < 0,9$ *(rata de germinare este mai mica de 90%).*

Vom avea deci un test unilateral, deoarece este putin probabil ca vanzatorul sa sustina o rata de germinare mai mica decat aceea reala.

Pentru pragul $\varepsilon = 0,05$ *avem:*



$P(Z < Z_{0,05}) = 0,05$; *deci* $P(Z < -Z_{0,05}) = 0,95 = F(-Z_{0,05})$.

Din TABELUL I gasim ca $-Z_{0,05} = 1,65$, *deci* $Z_{0,05} = -1,65$.

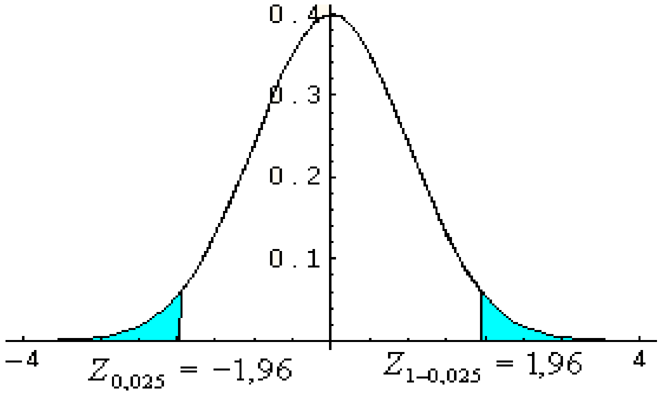
Cum H_0: $X \sim Bin(100; 0,9)$ *avem ca* $X \sim N(np, npq) = N(90,9)$, *deci* $Z_{calc} = \frac{X-np}{\sqrt{npq}} = \frac{83-90}{3} = -2,33$.

Dar $Z_{calc} = -2,33 < -1,65$ *si deci va trebui sa resping* H_0 *cu pragul de 5%. Adica fabricantul de seminte... minte!*

Pentru pragul $\varepsilon = 0,01$, $\phi(-Z_{0,01}) - 0,99$, *deci* $-Z_{0,01} = 2,32$, *sau* $Z_{0,01} = -2,32$. *Cum* $Z_{calc} = -2,33 < -2,32$, *dar aproape insensibil mai mic, testul in acst caz* ***nu*** *poate fi concludent deoarece valoarea calculata* Z_{calc} *este prea aproape de valoarea critica* $-2,32$. *Prin urmare, fabricantul... minte, dar nu minte prea mult! Este nevoie de alte solutii pentru a capata o certitudine mai mare.*

Exemplul 11.4 *O masina produce benzi elastice cu tensiuni de rupere normal distribuite cu media* $m = 45N$ *si* $\sigma = 4,36N$. *Intr-o zi s-a facut o selectie de volum 50 si s-a gasit media selectiei* $\overline{x} = 43,46N$. *Testati cu un prag de semnificatie de 5% daca acest lucru indica sau nu o schimbare a mediei tensiunilor de rupere.*

Solutie *H_0: $m = 45$ (media nu s-a schimbat)*
H_1: $m \neq 45$ (media s-a schimbat)-test bilateral!



$\overline{X} \sim N\left(m, \frac{\sigma^2}{n}\right)$, $m = 45N$; $\sigma = 4,36$; $n = 50$.

$Z_{calc} = \frac{\overline{x}-m}{\sigma/\sqrt{n}} = \frac{43{,}46-45}{43{,}36/\sqrt{50}} = -2,4975 < -1,96$. *Prin urmare respingem ipoteza H_0 cu pragul de semnificatie de 5%.*

Un interval de incredere de nivel 95% pentru medie este $\overline{x} \pm 1,96\frac{\sigma}{\sqrt{n}} = (42,25; 44,67)$. *Vedem ca* $45 \notin (42,25; 44,67)$. *Cea mai mica valoare a lui σ astfel incat 45 sa fie in intervalul de incredere $\overline{x} \pm 1,96\frac{\sigma}{\sqrt{n}}$ este $\sigma = 5,56$ (vezi ecuatia $43,46 + 1,96\frac{\sigma}{\sqrt{50}} = 45$).*

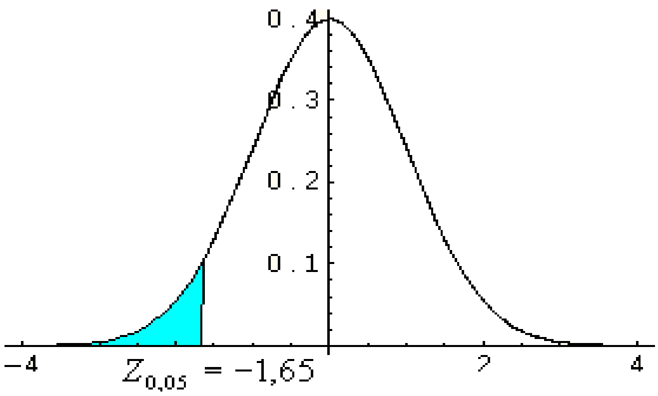
Exemplul 11.5 *Tensiunea de rupere a unor cabluri produse de o fabrica este normal distribuita cu media 6000N si deviatia standard $\sigma = 150N$. Gasiti probabilitatea ca un cablu luat la intamplare se aiba tensiunea de rupere > 6200N.*

S-a modificat procesul de productie si media tensiunilor de rupere se modifica. Se aleg 6 cabluri la intamplare dupa aceasta modificare, se testeaza si se gaseste o medie de rupere $\overline{x} = 5920N$. Testati cu un prag de 5% daca dupa modificare media tensiunilor s-a micsorat. Gasiti o constanta C a.i. noi sa putem spune cu un nivel de incredere de 90% ca media de rupere este mai mare decat C.

Soluţie $X \sim N(6000, 150^2)$;

$$\begin{aligned} P\left(X > 6200\right) &= P\left(\frac{X-m}{\sigma} > \frac{6200-m}{\sigma}\right) = P\left(Z > 1,333\right) = 1 - P\left(Z \leq 1,333\right) \\ &= 1 - F\left(1,333\right) = 1 - 0,90 = 0,1. \end{aligned}$$

$\overline{x} = 5920N$; H_0: $m = 6000N$; H_1: $m < 6000N$;
$\overline{X} \sim N\left(m, \frac{\sigma^2}{n}\right) = N\left(6000, \frac{150^2}{6}\right)$;



$Z_{calc} = \frac{\overline{x}-m}{\sigma/\sqrt{n}} = \frac{5920-6000}{150/\sqrt{6}} = -1,306 > -1,65$, *deci acceptam ipoteza* H_0 *cu pragul de semnificatie 5%.*

• *Trebuie sa gasim C a.i.* $P(C < m < \infty) = 0,9$, *sau* $P(-C > -m) = 0,9$, *sau inca* $P\left(\frac{\overline{X}-C}{\sigma/\sqrt{6}} > \frac{\overline{X}-m}{\sigma/\sqrt{6}} = Z\right) = 0,9$. *Deci* $F\left(\frac{\overline{X}-C}{\sigma/\sqrt{6}} > Z\right) = 0,9$ *si de aici gasim ca* $\frac{\overline{X}-C}{\sigma/\sqrt{6}} = Z_{0,9} = 1,29$. *Prin urmare* $C = \overline{x} - 1,29\frac{150}{\sqrt{6}}$.

Exemplul 11.6 *O distributie normala se crede a avea media 50. Se face o selectie de volum 100 si se gaseste o medie de 52,6 si o deviatie standard de selectie de 14,5. Testati cu nivelul de 5% daca media populatiei a crescut.*

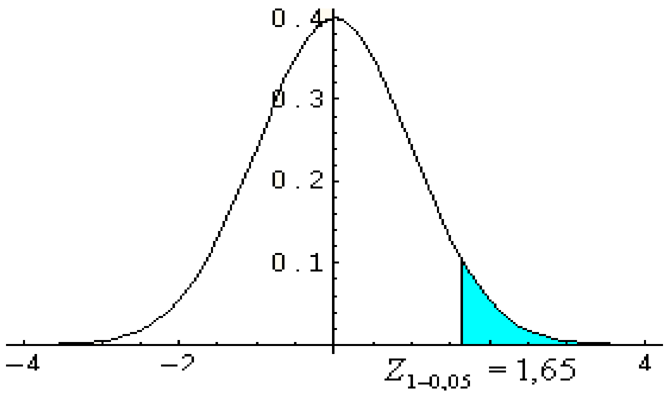
Solutie *Fie m media reala si* σ^2 *dispersia reala a populatiei.* $\overline{x} = 52,6$, *iar* $s = \sqrt{s^2} = 14,5$,

unde $s^2 = \frac{1}{n}\sum(x-\overline{x})^2$.

H_0: $m = 50$ *(nu exista o schimbare a mediei)*

H_1: $m > 50$ *(media populatiei a crescut)*

$\overline{X} \sim N\left(m, \frac{\sigma^2}{n}\right)$. *Estimam pe* σ^2 *cu* $\widehat{\sigma}^2 = \frac{ns^2}{n-1} \approx s^2$, *deoarece* $n = 100$ *este considerat mare* $(100 > 30)$. *Folosim deci statistica* $Z = \frac{\overline{X}-m}{\widehat{\sigma}/\sqrt{n}}$ *si calculam* $Z_{calc} = \frac{52,6-50}{14,5/\sqrt{100}} = 1,793$.



Cum $Z_{calc} = 1,793 > 1,645$, *vom respinge* H_0, *adica acceptam* H_1 *cu pragul de semnificatie de 5%. Deci acceptam ca mesia a crescut cu pragul de semnificatie de 5%.*

11.1.2 Testul T privind media unei populatii normale cu dispersia estimata prin estimatorul nedeplasat S'^2

Ipoteza nula H$_0$: $m = m_0$ se refera la o populatie normala careia nu ii cunoastem dispersia. Aceasta se va estima fie prin estimatorul deplasat $S^2 = \frac{1}{n}\sum(x-\overline{x})^2$, fie prin estimatorul nedeplasat $S^{/\,2} = \frac{1}{n-1}\sum(x-\overline{x})^2$. Daca volumul selectiei n este mare ($n \geq 30$) atunci $S^{/\,2} \approx S^2$ si *putem* folosi ca statistica pentru un test de semnificatie, statistica Z (vezi Exemplul 11.1.6). Daca insa volumul selectiei este mic *nu* mai putem folosi aceasta statistica ci este indicat sa folosim statistica Student cu $n-1$ grade de libertate, $T = \frac{\overline{X}-m_0}{S^{/}/\sqrt{n}}$, unde $S^{/} = \sqrt{S^{/\,2}}$.

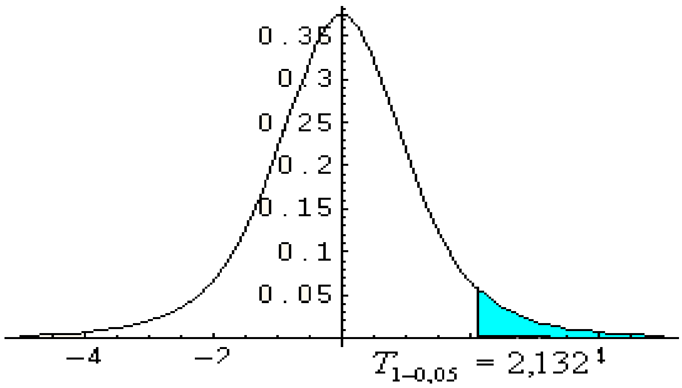
In rest, testul lucreaza exact ca testele de la 11.1.1.

Exemplul 11.7 *Se testeaza rezistenta in ohmi pentru 5 bucati de cablu si se gasesc valorile: 1,51; 1,49; 1,54; 1,52; 1,54. Daca cablul ar fi din argint pur,rezistenta lui ar fi de 1,5 ohmi. Daca argintul nu este pur, rezistenta creste. Testati cu un nivel de semnificatie de 5% faptul ca argintul din cablu nu este pur.*

Solutie H_0: $m = 1,5$ *ohmi.*

H_1: $m > 1,5$ *ohmi (test unilateral)*

Deoarece esantionul selectiei este mic ($n = 5$) *vom folosi statistica student* $T = \frac{\overline{X}-m}{S^{/}/\sqrt{n}} = \frac{\overline{X}-m}{S/\sqrt{n-1}}$, *cu* $n-1=4$ *grade de libertate.*



F($T_{0,05}$) =*P*(*T*< $T_{0,05}$) = 0*,95. Din TABELUL II, pentru* $\nu = 4$ *grade de libertate gasim ca* $T_{0,05} = 2,132$.

$T_{calc} = \frac{\overline{x}-m}{S/\sqrt{n-1}} = \frac{1,52-1,50}{0,019/2} = 2,105$, *deoarece* $\overline{x} = \frac{7,6}{5} = 1,52$, *iar* $S^2 = \frac{1}{n}\sum(x-\overline{x})^2 = \frac{0,0018}{5} = 0,00036$, *sau* $S = 0,019$.

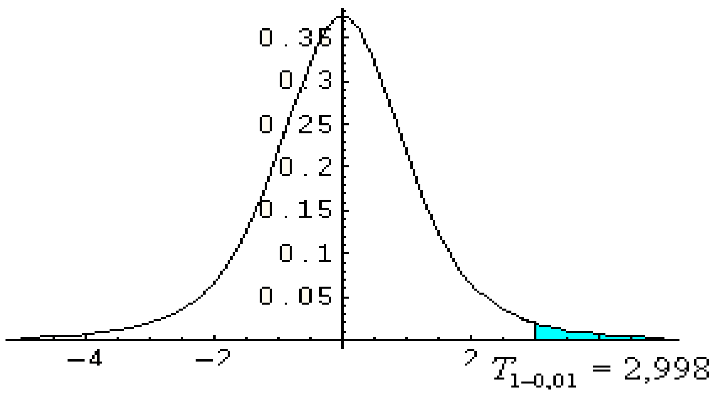
Deoarece $T_{calc} < 2,132$, *valoarea critica a testului, atunci trebuie sa acceptam ipoteza* H_0 *cu nivelul de semnificatie 5%.*

Exemplul 11.8 *Se fac 8 observatii dintr-o populatie normala si gasim* $\overline{x} = 4,65$ *si* $\sum(x-\overline{x})^2 = 0,74$. *Testati cu nivelul de semnificatie de 2% daca media distributiei este 4,3.*

Solutie H_0: $m = 4,3$

H_1: $m \neq 4,3$ *(test bilateral)*

$T = \frac{\overline{X}-m}{S/\sqrt{n-1}}$; $T_{calc} = \frac{4,65-4,3}{\frac{0,74}{8\sqrt{7}}} = 3,05$.



Cum $T_{calc} > 2,998$, resping H_0 cu nivelul de semnificatie de 2%.

11.1.3 Test pentru proportia de succese

Sa notam cu P_s =(numarul de realizari ale evenimentului A din n incercari)$/n$, adica proportia de realizari a evenimentului A, cu X v.a. $\begin{pmatrix} 1 & 0 \\ p & q \end{pmatrix}$, unde $q = 1 - p$, iar p este probabilitatea teoretica pentru a se realiza evenimentul A la o incercare. Daca punem $X_1 = X_2 = \cdots = X_n = X$, este clar ca P_s este chiar $\overline{X} = \frac{X_1+X_2+\cdots+X_n}{n}$. Pentru n mare, stim ca $\overline{X} \sim \mathrm{N}(\mathrm{M}(\mathrm{P}_s), \mathrm{D}(\mathrm{P}_s))$, dar $\mathrm{M}(\mathrm{P}_s) = \mathrm{M}(\overline{X}) = p$, $\mathrm{D}(\mathrm{P}_s) = \mathrm{D}(\overline{X}) = \frac{D(X)}{n} = \frac{p-p^2}{n} = \frac{pq}{n}$. Deci

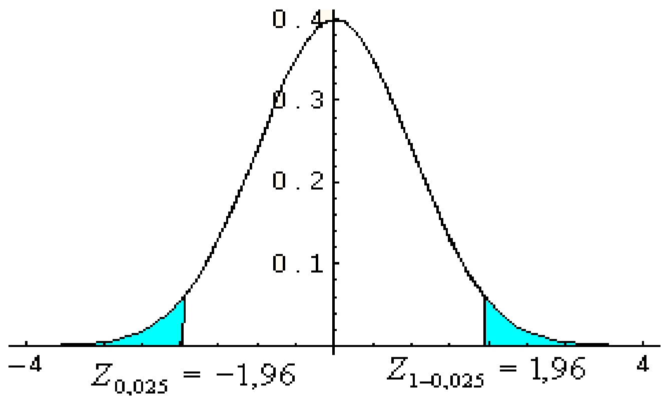
P_s ="proportia de succese"$\sim \mathrm{N}\left(p, \frac{pq}{n}\right)$. Fie $Z = \frac{\mathrm{P}_s - p}{\sqrt{\frac{pq}{n}}} \sim \mathrm{N}(0,1)$. Testul proportiei de succese se realizeaza cu ajutorul statisticii Z dupa cum se va vedea in exemplul urmator.

Exemplul 11.9 *La o universitate americana senatul sustine ca nu se face discriminare sexuala la admitere. Se aleg 500 studenti si se gasesc 267 baieti. Testati cu nivelul de semnificatie 5% daca senatul universitatii spune adevarul sau nu.*

Solutie *H_0: $p = 0,5$=probabilitatea ca un student sa fie baiat;*

H_1: $p \neq 0,5$ (test bilateral)

$Z = \frac{P_s - p}{\sqrt{\frac{pq}{n}}}$; $z_{calc} = \frac{\frac{267}{500} - 0,5}{\sqrt{\frac{0,5 \cdot 0,5}{500}}} = 1,52.$



Cum $z_{calc} = 1,52 < 1,96$, acceptam ipoteza H_0 cu pragul de semnificatie de 5%. Prin urmare este foarte probabil ca senatul sa spuna adevarul.

Exemplul 11.10 ***([Hays], pag. 447)*** *Un producător de frigidere pretinde că temperatura medie în compartimentul de congelare este de 10 grade Fahrenheit (aproximativ -12,3 grade Celsius). Vrem să vedem adevărul acestei afirmatii şi facem ipotezele:* $H_0 : m \leq 10$ *(ipoteza nulă), versus* $H_1 : m = 10$ *(ipoteza alternativă). Facem un sondaj pe un lot de 16 frigidere alese la întâmplare şi măsurăm temperaturile în congelatoarele acestora. Găsim că media selectiei este 10,24 grade, iar dispersia de selectie modificată (nedeplasată) este de* $S'^{\,2}$*=0,36. Presupunem că distributia temperaturilor este normală (deoarece apare un fenomen de repartitie de "erori"). Cum dispersia este calculată tot din selectie folosim v.a. de selectie* T*. Calculăm* $t = \frac{10{,}24-10}{0{,}6/4} = 1,6$*. Vrem sa folosim statistica* T *cu pragul de semnificatie de 5% pentru a vedea daca producatorul are dreptate sau nu. Testul va fi unilateral deoarece zona de respingere este data de* $m > 10$*.* $T_{0,95} = 1,753$ *pentru 15 grade de libertate, dupa cum se poate constata in TABELUL II.* $T_{calc} = 1,6 < 1,753$*. Prin urmare acceptam ipoteza* H_0 *cu pragul de 5%.*

Observaţia 11.11 *a) În exemplul de mai sus regiunea de respingere este:* $m > 10$*, concentrată într-o singură directie, adică la dreapta lui 10. Un astfel de test se zice directional (sau unilateral).*

b) Dacă vrem să testăm o ipoteză despre dispersia unei populatii: $H_0 : \sigma^2 = \sigma_0^2$*, unde* σ_0 *este dată, va trebui să utilizăm testul* χ^2 *(vezi Lectia 12). Aici* $H_1 : \sigma^2 \neq \sigma_0^2$*, şi* α *este dat. Cele două (*α *şi* H_1*) descriu regiunea de respingere. Statistica folosită este* $\chi^2_{(n-1)} = \frac{(n-1)S'^{\,2}}{\sigma_0^2}$*, unde* $(n-1)$ *reprezintă numărul gradelor de libertate. Se urmează apoi aceeasi cale ca şi în cazul mediei (vezi Rezumatul şi exemplele date acolo).*

11.1.4 Testul T pentru compararea a două esantioane

Fie $\{x_1, ..., x_n\}$ şi $\{x'_1, ..., x'_m\}$ două esantioane. H$_0$:<<cele două esantioane provin din aceeasi populatie>>.

Fie $\alpha \in$(0,1) un prag de semnificatie. Notăm cu $T_\nu = \frac{\overline{X}-\overline{X}'}{S'}\sqrt{\frac{nm}{n+m}}$, unde $\overline{X} = \left(\sum x_i\right)/n$, $\overline{X}' = \left(\sum x'_i\right)/m$, $S'^{\,2} = \frac{1}{\nu}\left(\sum\left(x_i - \overline{X}\right)^2 + \sum\left(x'_i - \overline{X}'\right)^2\right)$, iar $\nu = n + m - 2$. R. A. Fisher a arătat că T_ν tinde către o repartitie Student cu ν grade de liberate. Calculăm T_ν. Găsim cuantila de ordin 1–$\alpha/2$ în Tabelul corespunzător lui T pentru ν grade de libertate şi o notăm cu $T_{\alpha/2}$. Testul functionează astfel:

- Dacă $|T_\nu| > T_{\alpha/2}$ atunci vom respinge ipoteza H$_0$ cu pragul de semnificatie α.
- Dacă $|T_\nu| \leq T_{\alpha/2}$ atunci vom accepta ipoteza H$_0$ cu pragul de semnificatie α.

Exemplul 11.12 ***(R. A. Fisher)*** *8 ghivece cu fire de orez au fost supuse la descărcări electrice. Altele 9 au fost ferite de descărcări. Rezultatul recoltei a fost (număr de spice):*

Izolate: 17, 27, 18, 25, 27, 29, 27, 23, 17;

Electrizate: 16, 16, 20, 16, 20, 17 ,15, 21.

Să se testeze ipoteza H_0*:<<electricitatea influentează cresterea orezului>>.*

11.2 Tipuri de erori. Reguli de decizie

Vom incepe cu un exemplu.

Exemplul 11.13 ***([Hays], pag. 404)*** *Un economist are două ipoteze asupra implicatiilor ce derivă din cresterea impozitelor la un anumit moment dat. Prima ipoteză este că după această crestere 80% din populatie va trebui să-şi reducă economiile. A doua ipoteză este că numai 40% din populatie va trebui să facă acest lucru. Cum s-ar putea afla care ipoteză este adevărată?*

Soluţie. S-ar putea ca nici una dintre cele două ipoteze să nu fie adevărată. Totusi, aici, noi vom considera că *sigur* una ditre ele este adevărată. Vom nota:

H_0: p=0,8

H_1: p=0,4, unde p este proportia de consumatori care urmează să-şi reducă economiile datorită cresterii impozitelor. Iată că impozitele au crescut şi economistul nostru vrea să vadă care dintre cele două ipoteze ale sale (fiecare are în spatele ei rationamente şi teorii economice sofisticate) este adevărată. Pentru aceasta face un sondaj pe un esantion de n consumatori. Deoarece fiecare din consumatori spune DA sau NU (şi-a redus sau nu şi-a redus economiile) avem un proces de tip Bernoulli cu n dat şi p dat. Pentru H_0 trebuie să considerăm p=0,8, iar pentru H_1 trebuie să considerăm p=0,4. Presupunem, pentru usurintă, că n=10. Notăm cu r numărul acelor consumatori care au răspuns cu DA (dintre cei n chestionati). Statistica de seletie care poate fi comparată cu p este R/n, unde R este v.a. ce pate lua valorile r: 0,1,...,n. Valorile teoretice pe care le poate lua R/n şi probabilitătile lor le găsim în următorul tabel (am folosit distributia binomială, vezi Tabelul V):

r	r/n	$P(r/n \mid p=0,8)$ aprox. 4 zecimale	$P(r/n \mid p=0,4)$ aprox. 4 zecimale
0	0	0	$0,006$
1	$0,1$	0	$0,040$
2	$0,2$	0	$0,121$
3	$0,3$	$0,001$	$0,215$
4	$0,4$	$0,006$	$0,251$
5	$0,5$	$0,026$	$0,2$
6	$0,6$	$0,088$	$0,111$
7	$0,7$	$0,201$	$0,042$
8	$0,8$	$0,302$	$0,011$
9	$0,9$	$0,268$	$0,002$
10	1	$0,107$	$0,0001$

(11.1)

Să presupunem că statisticianul are datele de selectie ale sondajului, are deci raportul R/n calculat din sondaj. El are nevoie de o *REGULĂ DE DECIZIE* pentru a putea alege H_0 sau H_1. Există foarte multe posibilităti de a construi astfel de reguli de decizie. Unele sunt mai "bune", altele nu sunt asa de "bune". Vom alege acum următoarea regulă: <<Dacă R/n <0,8, alegem H_1; dacă $R/n \geq$0,8, alegem H_0 >>(REGULA 1).

Ce se poate întâmpla după ce statisticianul a folosit această regulă?

El poate gresi sau nu. Să calculăm probabilitătile în toate cele patru cazuri care pot să apară:

$$\begin{array}{cccc} & & \textit{Situatia} & \textit{reală} \\ & & \mathrm{H}_0 & \mathrm{H}_1 \\ \textit{Decizia} & \mathrm{H}_0 & \text{corect} & \text{eroare II} \\ \textit{luată} & \mathrm{H}_1 & \text{eroare I} & \text{corect} \end{array} \tag{11.2}$$

De exemplu, să presupunem că din selectie am obtinut R/n $\geq$0,8 şi totusi în realitate p=0,4. Deci am ales H_0 şi totusi H_1 este adevărată. Apare al doilea tip de eroare (eroare II). Să calculăm probabiltatea acestei erori (folosind tabelul de mai sus):

$$\begin{aligned} & P\,(R/n \geq 0{,}8 \mid p{=}0{,}4) \\ = \; & P\,(R/n{=}0{,}8 \mid p{=}0{,}4) + P\,(R/n{=}0{,}9 \mid p{=}0{,}4) + P\,(R/n{=}1 \mid p{=}0{,}4) \\ = \; & 0,011 + 0,002 + 0,0001 \approx 0,013. \end{aligned}$$

Primul tip de eroare (eroare I) apare dacă alegem H_1 şi totusi H_0 este adevărată. Probabilitatea acesteia este:

$$\begin{aligned} & P\,(R/n < 0{,}8 \mid p{=}0{,}8) \\ = \; & +0 + 0 + 0,001 + 0,006 + 0,026 + 0,088 + 0,201 = 0,322. \end{aligned}$$

Cele două situatii corecte au următoarele probabilităti:
P($R/n \geq$ 0,8 | p=0,8) =0,677,
P($R/n <$ 0,8 | p=0,4) =0,987.
Punem aceste rezultate în următorul tabel:

$$\begin{array}{cccc} & & \textit{Situatia} & \textit{reală} \\ & & \mathrm{H}_0 & \mathrm{H}_1 \\ \textit{Decizia} & \mathrm{H}_0 & 0{,}677 & 0{,}013 \\ \textit{luată} & \mathrm{H}_1 & 0{,}323 & 0{,}987 \end{array} \tag{11.3}$$

Ce spune acest tabel? Dacă după selectie R/n <0,8, este foarte probabil ca p=0,4. Oricum este mai probabil acest lucru decât faptul că R/n $\geq$0,8 şi p=0,8. Este foarte putin probabil "să gresim" cu această regulă de decizie, deoarece 0,323+0,013<0,677+0,987.

Iată o altă regulă de decizie: <<Dacă R/n <0,6, alegem H_1; dacă R/n $\geq$0,6, alegem H_0 >>(REGULA 2). Tabelul corespunzător acestei reguli este următorul:

		Situatia	*reală*
		H_0	H_1
Decizia	H_0	0,966	0,116
luată	H_1	0,034	0,834

(11.4)

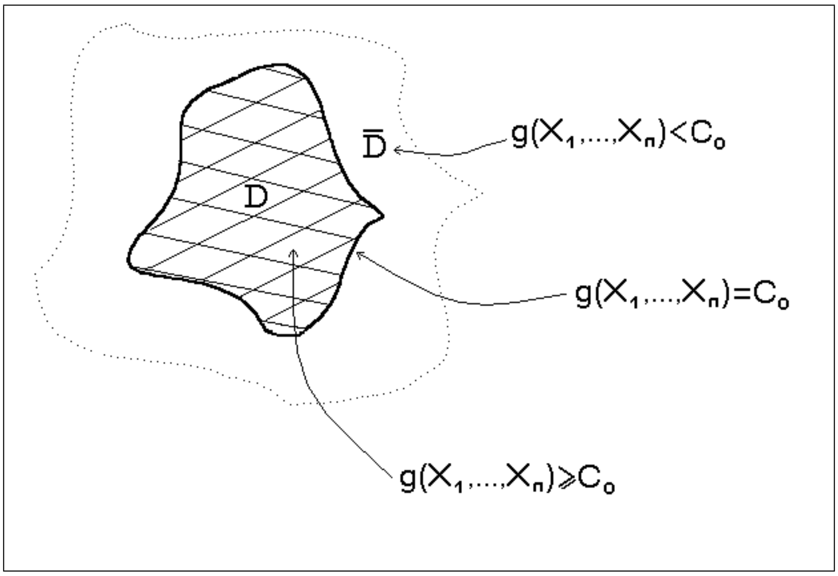
Este clar că dintre cele două reguli de decizie este "mai bună" a doua regulă deoarece probabilitătile de eroare sunt mici.

În prima regulă de decizie valoarea lui R/n=0,8 (care face trecerea de la zona ipotezei H_0 la zona ipotezei H_1) se numeste *valoarea critică a lui* R/n (adică a statisticii R/n).

Pentru a doua regulă de decizie valoarea critică a statisticii R/n este 0,6.

• În general, frontiera dintre domeniul de acceptare D şi domeniul de respingere $\overline{D}$, formează multimea de puncte în care statistica de testare $g\left(X_1, ..., X_n\right)$ ia valoarea critica C_0.

În figura următoare avem reprezentarea grafică a *domeniului de acceptare* D pentru ipoteza H_0, a *domeniului de respingere* $\overline{D}$ pentru ipoteza H_0 şi a multimii punctelor de valoare critică pentru următoarea regulă de decizie: <<Dacă $g\left(X_1, ..., X_n\right)$ $\geq\mathrm{C}_0$, acceptă H_0; dacă $g\left(X_1, ..., X_n\right)$ $<\mathrm{C}_0$, acceptă H_1 >>. Aici C_0 este valoarea critică a testului.



Regiunea de acceptare si regiunea de respingere a unui test

Desigur că statistica $g\left(X_1, ..., X_n\right)$ este aleasă astfel încât să fie o legătură naturală între ea şi ipotezele H_0 şi H_1. Se consideră, de asemenea, că H_0 şi H_1 se exclud reciproc $\left(D \cap \overline{D} = \emptyset\right)$.

Vom presupune în continuare că H_0 este *ipoteza care se testează***.** Statisticianul poate face două tipuri de erori:

• *Eroare de tipul I* dacă respinge H_0, ea fiind adevărată;
• *Eroare de tipul II* dacă acceptă H_0, ea nefiind neadevărată.

Notăm cu $\alpha = P$(eroare de tipul I)$= P$(respinge H_0 $|H_0$ este adevărată);
$\beta = P$(eroare de tipul II)$= P$(acceptă H_0 $|H_0$ este falsă).

Tabelele (11.3) şi (11.4) din Exemplul 11.1 se generalizează la următorul tabel:

		Situatia	*reală*
		H_0	H_1
Decizia	Accept H_0	1-α	β
luată	Resping H_0	α	1-β

(11.5)

Orice regulă de decizie are un cuplu de numere (α, β).

Idealul ar fi ca α şi β să fie 0, sau foarte mici.

Dintre două reguli de decizie cu(α_1, β_1) şi (α_2, β_2) astfel încât $\alpha_1 \leq \alpha_2$, $\beta_1 \leq \beta_2$, vom elimina pe cea de-a doua. Spunem că regula de decizie cu (α_1, β_1) *domină* (este mai tare) regula de decizie cu (α_2, β_2). În cazul Exemplului 11.1 cele două reguli nu se pot compara, ele nu se domină una pe alta.

O regulă de decizie dominată de o altă regulă de decize se zice *inadmisibilă.*

Vom da acum un exemplu de regulă de decizie inadmisibilă: <<Dacă $R/n \in (-\infty; 0,2) \cup (0,8; \infty)$, alegem H_1; dacă $R/n \in [0,2; 0,8]$, alegem H_0 >>(REGULA III).

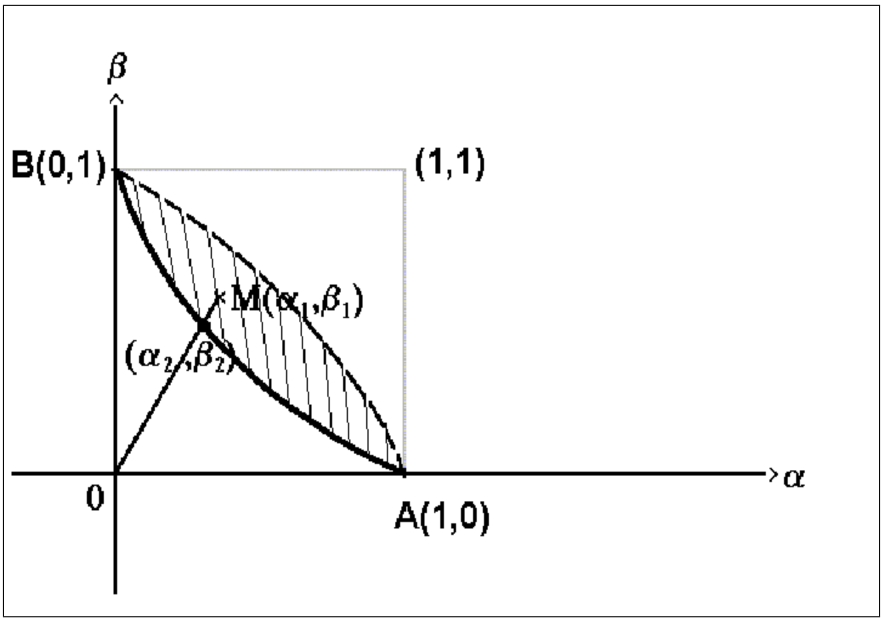
Tabelul corespunzător ei este:

		Situatia	*reală*
		H_0	H_1
Decizia	Accept H_0	0,625	0,952
luată	Resping H_0	0,375	0,048

(11.6)

Aici α_2=0,375, β_2=0,952. Cum α_1=0,323 şi β_1=0,013 erau probabilitatile de eroare pentru prima regulă, rezultă că regula <<Dacă $R/n \in (-\infty; 0,2) \cup (0,8; \infty)$, alegem H_1; dacă $R/n \in [0,2; 0,8]$, alegem H_0 >>(REGULA III) este inadmisibilă deoarece este dominată de prima regulă de decizie. Nu vom lucra niciodată cu reguli de decizie despre care stim că sunt inadmisibile. Dacă se analizează îndeaproape regula <<Dacă $R/n \in (-\infty; 0,2) \cup (0,8; \infty)$, alegem H_1; dacă $R/n \in [0,2; 0,8]$, alegem H_0 >>(REGULA III) se constată că ea contrazice chiar "bunul simt" probabilistic (De ce?).

În general statisticianul este interesat de modul in care variaza probabilitatile de eroare α şi β atunci când el schimba legea de decizie. Evident că pentru orice statistica de testare $g(X_1, ..., X_n)$ şi pentru orice lege de decizie fixată avem un α şi un β bine determinati. Păstrăm statistica de testare fixată şi variem legile de decizie. Se obtine un domeniu *al probabilităţilor de eroare* (un *domeniu de risc)* dacă α se consideră abscisa şi β ordonata punctului (α,β) — vezi figura următoare:



Domeniul probabilităţilor de eroare

Curba groasă $\widehat{AB}$ din figura anterioară este curba corespunzătoare tuturor cuplurilor (α,β) care derivă din decizii "bune" sau admisibile. Oricare alt punct (α_1,β_1) din domeniul de risc care *nu* se află pe curba $\widehat{AB}$ provine dintr-o decizie inadmisibilă, deoarece aceasta este dominată de decizia corespunzătoare punctului (α_2,β_2) de pe curba $\widehat{AB}$ şi aflat la intersectia dintre această curbă şi segmentul OM.

Experienta arată că dacă volumul de selectie n creste, atunci curba $\widehat{AB}$ "se apropie" de origine, adică riscul devine mai mic cel putin pentru deciziile admisibile (deoarece precizia de predictie asupra populatiei creste odată cu n).

Singura problemă care rămâne pentru statistician este aceea legată de alegerea regulii de decizie admisibile. Aici intervine "negocierea" între cazurile "α mare, β mic", sau invers, "α mic, β mare", (α,β)$\in \widehat{AB}$.

Situatia neutră este aceea când α=β. În practică această negociere depinde de factori subiectivi sau obiectivi dar.

• Din punct de vedere istoric ipoteza H_0 se numeste *ipoteza nulă* (nevinovătia prezumtivă în cazul unui acuzat!), iar ipoteza H_1 se zice *ipoteză alternativă*(vinovătia acuzatului). În practică, numărul α (notat uneori şi cu ε) se dă ca fiind 0,05 sau 0,01 (rareori se utilizează altă valoare). α reprezintă probabilitatea de a respinge H_0 cu toate că H_0 este adevărată.

Presupunem că H_0 este adevărată şi alegem asa numita *regiune (domeniu) de respingere* în concordantă cu ipoteza H_1. Să notăm domeniul de respingere cu Res = {*acele valori ale statisticii de testare* $g\left(X_1,...,X_n\right)$ *pentru care* H_1 *este adevărată*}. Găsim multimea Res din ipoteza:

$$P\left(g\left(X_1,...,X_n\right)\in Res \mid \mathrm{H}_0\text{=adevărată}\right) = \alpha(dat)$$

Dacă la o selectie de volum n: $X_1, ..., X_n$, valoarea v.a. $g(X_1, ..., X_n) \in Res$, spunem că *ipoteza H_1 este adevărată cu pragul de semnificatie* α(cu conditia ca α să fie mic, ca mai sus).

Vom da acum câteva exemple de testare a unor ipoteze statistice în lumina celor spuse mai sus.

Exemplul 11.14 ***([Hays], pag. 415)*** *Un muncitor poate să realizeze X piese pe oră. După îndelungate cercetări statistice s-a stabilit că X este o v.a. normală cu media m=138 şi deviatia standard σ=20. Un inginer pretinde că poate aduce o inovatie în procesul de productie astfel încât să ridice media la m=142 piese pe oră, fără a perturba normalitatea v.a. X şi pe σ. Este chemat un statistician să testeze pretentia inginerului.*

Solutie Introducem două ipoteze:

H$_0$: m=138

H$_1$: m=142

Facem o selectie de 100 muncitori care lucrează câte o oră fiecare. Alegem pragul de semnificatie α=0,05=P(respingem H$_0$ |H$_0$ este adevărată). Vrem acum să găsim regiunea de respingere în acest caz concret. Stim că un estimator bun pentru media m este media de selectie $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$, care este normal distribuită cu media m şi deviatia standard (pentru selectia noastră n=100) $\sigma_{\overline{X}} = 20/\sqrt{100} = 2$. Chiar dacă ipoteza de normalitate a v.a. X nu este întrutotul adevărată (sau chiar falsă!), deoarece n=100 este mare, din teorema limită centrală, rezultă că $\overline{X}$ este distribuită normal. Dacă H$_0$ este adevărată, atunci $\overline{X}$ este normal distribuită cu media 138 şi deviatia standard 2, pe când, dacă H$_1$ este adevărată, m=142 şi deviatia standard tot egală cu 2. Vedem de aici că valorile mari ale v.a. $\overline{X}$ favorizează ipoteza H$_1$, iar valorile mici ale v.a. $\overline{X}$ favorizează ipoteza H$_0$.

Regiunea de respingere va fi deci de forma: <<Respinge H$_0$ dacă $\overline{X} \geq C$ >>. Aici C reprezintă valoarea critică a regulii de decizie: <<Dacă $\overline{X} < C$, alegem H$_0$; dacă $\overline{X} \geq C$, alegem H$_1$ >>.

Vrem acum să determinăm valoarea critică C a statisticii $\overline{X}$ dacă stim pe α=0,05. Scriem că α =P(respinge H$_0$ |H$_0$=adevărată)=P($\overline{X} \geq C \mid m$=138)=0,05. Dar P($\overline{X} \geq C \mid m$=138)= P$\left(Z \geq \frac{C-138}{2}\right)$, unde $Z = \frac{\overline{X}-138}{2}$ este v.a. standard normală cu media 0 şi cu deviatia standard 1. Vrem ca P$\left(Z \geq \frac{C-138}{2}\right) = 0,05 = 1 - F\left(\frac{C-138}{2}\right)$, unde $F(z)$ este functia de repartitie (functia de distributie cumulată) a v.a. normale stanard Z. Din Tabelul I găsim că $\frac{C-138}{2}$ este 1,65, adică cuantila de ordin 1-0,05=0,95 a v.a. Z. De aici rezultă că punctul critic C=141,30. Acum putem calcula şi probabilitatea

$$\begin{aligned}
\beta &= P(accept\ H_0 \mid H_1 = adevărată) = P(\overline{X} < 141,30 \mid m = 142) \\
&= P\left(Z < \frac{141,30 - 142}{2}\right) = P\left(Z < -0,35\right) = F(-0,35) = 1 - F(0,35) = 0,36.
\end{aligned}$$

Tabelul corespunzător regulii de decizie de mai sus este:

		Situatia	*reală*
		H_0	H_1
Decizia	Accept H_0	0,95	0,36
luata	Resping H_0	α=0,05	0,64

(11.7)

Dacă din calcule $\overline{X} < 141,30$, atunci alegem H_0, adică rămânem la vechiul procedeu de lucru.

Să comentăm putin rezultatele din tabelul (11.7). α=0,05 , chiar dacă $\overline{X} \geq 141,30$, adică aleg H_1 şi resping H_0, probabilitatea de eroare în cazul când H_0 este adevărată, este foarte mică. Prin urmare, dacă alegem noul procedeu şi-l înlocuiesc cu primul (primul fiind mai bun) riscul este mai mic, aproape zero. Totusi, dacă alegem H_0 ($\overline{X} < 141,30$)— rămânem la vechiul procedeu în timp ce noul procedeu este mai bun, riscul este mai mare: β=0,36.

Statisticianul vrea să micsoreze şi acest risc β fără însă a mări pe α=0,05 (putem micsora pe β dacă reducem valoarea critică la 140, de exemplu; dar, în acest caz creste şi α). Teoretic stim că α şi β se micsorează dacă mărim volumul selectiei.

Vom mări pe n la 400. În acest caz deviatia standard a v.a. $\overline{X}$ devine $\sigma_{\overline{X}} = 20/\sqrt{400} = 1$. Determinăm valoarea critică ca mai sus şi găsim $C = 139,65$. În acest caz

$$\begin{aligned} \beta &= P(\textit{acceptă } H_0 \mid H_1 = \textit{adevărată}) = P(\overline{X} < 139,65 \mid m = 142) \\ &= P\left(Z < \frac{139,65 - 142}{1}\right) = P\left(Z < -2,35\right) = 0,01. \end{aligned}$$

Iată deci că am redus pe β de la 0,36 (când n=100) la 0,01 (când n=400).

Acum statiticianul poate să răspundă cu riscuri foarte mici (α, β mici) dacă este bine să schimbăm procedeul de producere al pieselor (accept H_1 şi resping H_0) sau, dimpotrivă, să lucrăm după acelasi procedeu (accept H_0 şi resping H_1).

În primul caz pot gresi cu 5%, iar în al doilea caz cu 1%. Facem deci un sondaj de 400 muncitori/oră. Calculăm $\overline{X}$ şi vedem dacă $\overline{X} < 139,65$ sau $\overline{X} \geq 139,65$.

Observaţia 11.15 *Pentru valoarea critică C=141,30, regiunea de respingere este [141,30;∞), pe când pentru valoarea critică C=139,65, regiunea de respingere este [139,65;∞).*

11.3 Puterea unui test statistic

Fie θ un parametru al unei populatii şi ipoteza: $H_0 : \theta = \theta_0$

Pentru o regiune de respingere dată şi pentru o valoare particulară a lui θ, de exemplu θ_1, putem calcula P(resping $H_0 \mid \theta = \theta_1$). Dar, dacă $\theta_1 = \theta_0$, această ultimă probabilitate este chiar α. Dacă $\theta_1 \neq \theta_0$ notăm cu $H_1 : \theta = \theta_1$ şi probabilitatea de mai sus devine P(resping H_0 |H_1=adevărată)=1-P(acceptă H_0 |H_1=adevărată)=1-β.

Definiţia 11.16 *Numărul P(resping H_0 | $\theta = \theta_1$), calculat mai sus, se numeste puterea testului ipotezei H_0 contra alternativei H_1.*

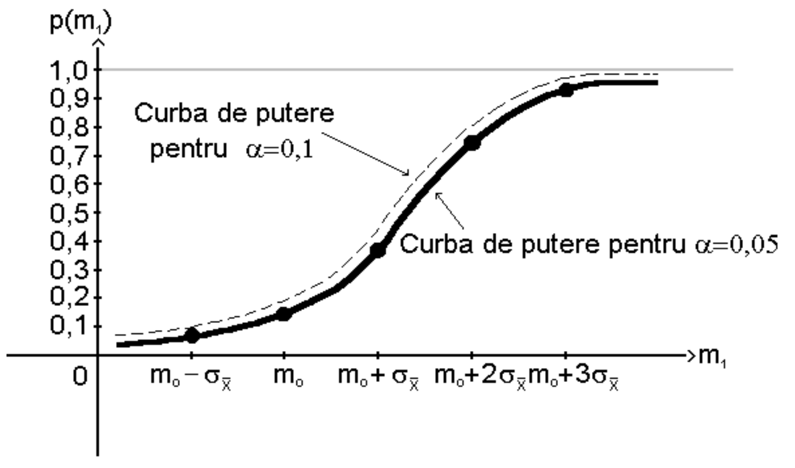
În Exemplul 11.2 puterea testului $\mathrm{H}_0 : m = 138$ contra alternativei $\mathrm{H}_1 : m = 142$ este egală cu 1-β=1-0,36=0,64, dacă regiunea de respingere este [141,30;∞). Un test este cu atât mai puternic cu cât ipoteza $\mathrm{H}_0 : \theta = \theta_0$ este "mai departe" decât valoarea reală a lui θ, $\mathrm{H}_1 : \theta = \theta_1$. Mai mult, 1-$\beta$ mare înseamnă β mic, adică un test este cu atât mai puternic cu cât este mai putin probabil să accept H_0 când ea de fapt nu este adevărată. Să amintim că noi alegem α foarte mic. Acest α determină zona de respingere (regiunea de respingere). Pentru această zonă de respingere determinăm pe β dacă stim pe θ_1. Prin urmare puterea unui test 1-β depinde de θ_1. Ea va fi maximă pentru acel θ_1 pentru care riscul de a accepta pe H_0 când ea este falsă, comparativ cu H_1, este minim (aici H_1 este considerată adevărată!).

Exemplul 11.17 *Să reluăm exemplul anterior într-un cadru mai general. Presupunem că vrem să studiem media unei populatii şi facem ipoteza nulă (initială, sau de bază): $H_0 : m = m_0$, şi ipoteza alternativă: $H_1 : m = m_1$. Presupunem că $m_1 > m_0$ şi că v.a. media de selectie $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$ este normală (pentru n mare $\overline{X}$ poate fi considerată normală) cu deviatia standard $\sigma_{\overline{X}}$ cunoscută. Deoarece valorile mari ale v.a. $\overline{X}$ favorizează H_1 şi valorile mici ale v.a. $\overline{X}$ favorizează H_0, este natural să căutăm zona de respingere de forma $\overline{X} \geq C$, unde C este valoarea critică a regulii de decizie:<<respinge H_0, dacă $\overline{X} \geq C$; acceptă H_0, dacă $\overline{X} < C$ >>. Ca şi în exemplul 11.2 găsim că pentru pentru α=0,05 valoarea critică $C = m_0 + 1,65\sigma_{\overline{X}}$. Acum putem calcula puterea acestui test pentru orice valoare dată lui m_1. De exemplu, dacă $m_1 = m_0 + \sigma_{\overline{X}}$ puterea testului este P(respinge H_0 | $m = m_0 + \sigma_{\overline{X}}$)=P($\overline{X} \geq m_0 + 1,65\sigma_{\overline{X}}$ | $m = m_0 + \sigma_{\overline{X}}$)=P$\left(Z \geq \frac{(m_0+1,65\sigma_{\overline{X}})-(m_0+\sigma_{\overline{X}})}{\sigma_{\overline{X}}}\right)$=P($Z \geq 0,65$)=0,26.*

Dacă $m_1 = m_0 + 3\sigma_{\overline{X}}$, puterea testului creste: P(respinge H_0 | $m = m_0 + 3\sigma_{\overline{X}}$)=P($Z \geq -1,35$)= 0,91. Această ultimă valoare este mare, deci, dacă cumva media $m = m_0 + 3\sigma_{\overline{X}}$, atunci testul de mai sus poate detecta acest lucru cu o probabiltate de 0,91.

• Numim *curbă de putere* a testului de mai sus graficul *functiei de putere* P($\overline{X} \geq m_0 + 1,65\sigma_{\overline{X}}$ | $m = m_1$), ca functie de variabila m_1. Să o notăm cu $p(m_1)$.

Două teste se pot *compara* şi putem spune care dintre ele este mai tare. Mai mult, în anumite conditii putem alege testul "cel mai puternic". Noi însă nu ne ocupăm aici de aceste lucruri. În cadrul rezumatului vom lucra cu functia $p(m_1)$ şi vom construi "cel mai puternic test" în sensul precizat acolo. Ne limităm la analiza functiei de putere din figura următoare:



Curba de putere pentru α=0,05 si α=0,1

Curba de putere depinde de alegerea pragului de semnificatie α. Este natural ca pentru α mai mare puterea testului să crească, după cum se vede în figură. De asemenea, dacă mărim volumul de selectie n, puterea testului creste pentru valori $m_1 > m_0$ şi scade pentru valori $m_1 < m_0$ (de ce?).

• În exemplele 2 şi 3 s-a testat media în cazul când dispersia populatiei era cunoscută. Putem observa că dacă micsorăm dispersia, puterea testului creste. Pe baza teoriei intervalelor de încredere (Lectia 10) se pot construi teste parametrice în care şi dispersia este necunoscută.

11.4 Rezumat

• Partea teoretica

Fie X un model statistic pentru o populatie P, model care depinde de un parametru $\theta \in \Theta$. Fie H o submultime specificată în Θ. Ea va fi numită ipoteză. Fie K=$\Theta \backslash H$. Spunem că:

— H se realizează dacă $\theta \in$H (prin natura lucrurilor)

— H *nu* se realizează dacă $\theta \in$K.

În Lectia 11 am notat H cu H_0 şi K cu H_1.

Să notăm cu Ω spatiul tuturor observatiilor pe care le vom face asupra populatiei P (spatiul de selectie). Pentru fiecare $\theta \in \Theta$ avem o distributie p_θ pentru X. Dacă pentru un sondaj $\omega \in \Omega$, găsim că $\theta \in$H, spunem că acceptăm ipoteza H, dacă $\theta \in$K spunem că respingem ipoteza H. Spatiul de selectie se împarte în două părti distincte: Ω=A$\cup$R, unde A este *zona de acceptare* a ipotezei H, adică A=$\{\omega \in \Omega \mid \theta \in$H$\}$, iar R=$\{\omega \in \Omega \mid \theta \in$K$\}$ este *zona de respingere* a ipotezei H.

Să notăm cu p_θ (R) $= \beta_\theta$ =probabilitatea de a respinge ipoteza H când de fapt θ este bine ales de natură. Se zice *eroare de primul tip* respingerea ipotezei H când ea este adevărată:

$\theta \in$H şi $\omega \in$R. Ea are probabilitatea β_θ.

Se zice *eroare de al doilea tip* acceptarea ipotezei H când ea de fapt nu este adevărată: $\theta \in$H şi $\omega \in$A. Testul are *pragul* ε dacă $\beta_\theta \leq \varepsilon$, $(\forall)$ $\theta \in$H. Functia $\theta \rightsquigarrow \beta_\theta$, $\theta \in$K se zice *puterea testului*.

Pentru a face un test se alege o statistică Y care este "mică" când H se realizează şi este "mai mare" când H nu se realizează. Se alege un număr y. Regiunea de respingere este de forma: Resping H$\Longleftrightarrow$Y$\geq y$. Testul este de prag ε dacă $\theta \in$H implică $p_\theta\left(\mathrm{Y} \geq y\right) \leq \varepsilon$. Dacă nu se dă dinainte pragul ε, se observă valorile y ale statisticii Y şi maximul probabilităţii $p_\theta\left(\mathrm{Y} \geq y\right)$ când $\theta \in$H. Dacă acest maxim este mic se respinge ipoteza H. Dacă acest maxim este mare se acceptă ipoteza H.

Partea practica

1) *Test asupra mediei m a unei populatii normale cu dispersia cunoscuta* σ^2

$Z = \frac{\overline{X}-m}{\sigma/\sqrt{n}} \sim N\left(0,1\right)$; $\mathrm{H}_0 : m = m_0$; $\mathrm{H}_1 : m \neq m_0$ (test bilateral), sau $\mathrm{H}_1 : m > m_0$ (sau $m < m_0$) (test unilateral)

2) *Test asupra mediei m cu dispersia* σ^2 *necunoscuta* $(n \geq 30)$

Se estimeaza σ^2 cu $S^2 = \frac{1}{n}\sum\left(x-\overline{x}\right)^2$ si se foloseste statistica $Z = \frac{\overline{X}-m}{S/\sqrt{n}} \sim N\left(0,1\right)$. H_0 si H_1 apar exact ca la **1)**.

3) *Test asupra mediei m cu dispersia* σ^2 *necunoscuta* $(n \leq 30)$

Se estimeaza σ^2 cu $S^{/\,2} = \frac{1}{n-1}\sum\left(x-\overline{x}\right)^2 = \frac{nS^2}{n-1}$ si se foloseste statistica

$$T = \frac{\overline{X}-m}{S/\sqrt{n-1}}\left(= \frac{\overline{X}-m}{S^{/}/\sqrt{n}}\right) \sim t\left(n-1\right)$$

adică distributia Student cu $n-1$ grade de libertate. H_0 si H_1 apar exact ca la **1)**.

4) *Test pentru proportia de succese*

$P_s \sim N\left(p, \frac{pq}{n}\right)$, $q = 1-p$. $Z = \frac{P_s-p}{\sqrt{\frac{pq}{n}}} \sim N\left(0,1\right)$. H_0 si H_1 apar exact ca la **1)**.

5) *Erori de tipul I si erori de tipul II*

	Situatia reala	*Concluzia noastra*		
(1)	H_0 este adevarata	Acceptam H_0	Decizie corecta	
(2)	H_0 este adevarata	Respingem H_0	Decizie falsa	Eroare I
(3)	H_0 este falsa	Acceptam H_0	Decizie falsa	Eroare II
(4)	H_0 este falsa	Respingem H_0	Decizie corecta	

$\alpha = P$(Eroare de tipul I)$= P$(resping $\mathrm{H}_0 \mid \mathrm{H}_0 =$ adevarata)

$\beta = P$(Eroare de tipul II)$= P$(accept $\mathrm{H}_0 \mid \mathrm{H}_1 =$ adevarata)

11.5 Exerciţii rezolvate (mai dificile)

1.Fie X o v.a. Poisson de medie λ. Considerăm ipoteza H: $\lambda \leq 0,5$. Fie X_1,...,X_{20} un esantion din X şi punem Y=X_1+$X_2 + \cdots +X_{20}$. Definiti cu ajutorul lui Y (care are tot o distributie Poisson) cel mai puternic test posibil de prag 0,05. Pentru aces test determinati:

1) Probabilitatea erorii de primul tip pentru $\lambda = 0,3$.

2) Probabilitatea erorii de al doilea tip pentru $\lambda = 0,8$.

Presupunem că găsim Y=20. Cu ce prag se poate accepta H?

Soluţie. Cum Y=20$\overline{\mathrm{X}}$ este natural să considerăm un test de forma: Respinge H$\Longleftrightarrow$Y$\geq y$. Când H se verifică, cea mai mare probabilitate de a respinge H se obtine pentru $\lambda = 0,5$. În acest caz Y este Poisson de parametru $20 \times 0,5 = 10$. Tabelul arată ca P(Y$\geq y$)$\leq 0,05$ pentru $y \geq 16$. Testul cerut se obtine pentru cel mai mic y posibil (pentru a avea puterea maximă). Deci testul este: Respinge H$\Longleftrightarrow$Y$\geq$ 16. Pentru $\lambda = 0,3$, Y este Poisson de parametru $20 \times 0,3 = 6$ şi probabiliatea erorii de primul tip este P(Y$\geq$ 16)=0,0005. Pentru $\lambda = 0,8$, Y este Poisson de parametru $20 \times 0,8 = 16$ şi probabilitatea erorii de al doilea tip este P(Y$\leq$ 15)=0,4667.

Presupunem că Y=20. Maximul lui P(Y$\geq$ 20), când H este verificată, se obtine pentru $\lambda = 0,5$. În acest caz Y este Poisson de parametru 10. Se găseste în acest caz P(Y$\geq$ 20)=0,0035 şi trebuie să acceptăm ipoteza pentru un prag $\leq 0,0035$.

2.Fie X o v.a. gaussiană de dispersie 1 şi de medie 0 sau 1. Vrem să testăm ipoteza H: M(X)=0 contra alternativei M(X)=1 cu un prag $\varepsilon = 0,05$. Se ia pentru acesta un esantion de volum n.

1) Definiti un test (cel mai puternic posibil).

2) Plecând de la ce valoare a lui n testul obtinut va avea puterea $\geq 0,9$?

Soluţie. Considerăm un test de forma: Respinge H$\Longleftrightarrow$ $\overline{\mathrm{X}} \geq y$. Fie λ =M(X). Variabila aleatoare T=$\sqrt{n}\left(\overline{\mathrm{X}} - \lambda\right)$ este gaussiană redusă. În tabele găsim P(T$\geq t$)=0,05 pentru t=1,6449. P(T$\geq u$)=0,05 pentru u=–1,2816.

1) n este considerat fix. Să găsim y astfel încât testul să fie de prag ε. Dacă H se verifică, $\lambda = 0$ şi T=$\sqrt{n} \cdot \overline{\mathrm{X}}$. Testul va avea pragul ε dacă P($\overline{\mathrm{X}} \geq y$)$\leq \varepsilon$, sau P(T$\geq \sqrt{n}y$)$\leq \varepsilon$, adică $\sqrt{n}y \geq t$. Testul va fi cel mai puternic dacă $y = \frac{t}{\sqrt{n}}$, deci avem testul:

Respinge H$\Longleftrightarrow$ $\overline{\mathrm{X}} \geq \frac{1{,}6449}{\sqrt{n}}$.

2) Presupunem că H *nu* se verifică, adică $\lambda = 1$ şi T=$\sqrt{n}\overline{\mathrm{X}} - \sqrt{n}$. Se comite o eroare de al doilea tip pentru $\overline{\mathrm{X}} \leq y$ sau T$\leq \sqrt{n}y - \sqrt{n} = t - \sqrt{n}$. Puterea testului va fi deci $\geq 0,9$ dacă P(T$\leq t - \sqrt{n}$), sau $t - \sqrt{n} < u$, deci $\sqrt{n} > t - u = 2,9265$. Cel mai mic n pentru care puterea testului este $\geq 0,9$ este $n = 9$.

3. Fie X pretul aceluiasi articol luat la întâmplare din 15 magazine. Găsim tabelul acestor *preturi în \$:*

42,7	42,6	43,0	43,5	42,8
43,1	43,6	42,9	41,6	42,8 .
42,9	43,2	42,6	43,1	43,1

a) Se poate admite ipoteza M(X)=43,0?

b) Se poate admite ipoteza D(X)=0,1? În ambele cazuri se ia pragul $\varepsilon = 0,05$ *şi se consideră legea gaussiană pentru X.*

Soluţie. 1) Avem că $n = 15$, $\overline{\mathrm{X}} = 42,9$, $\mathrm{S}^2 = 0,2$. De aici deducem că $|\mathrm{T}_0| = \frac{|\overline{\mathrm{X}}-43|}{\sqrt{\mathrm{S}^2/(n-1)}} = 0,8366$. O variabilă aleatoare cu $n-1$=14 grade de libertate are o probabilitate de cel putin 0,40 pentru a lua o astfel de valoare. Ipoteza M(X)=43 este perfect acceptabilă.

2) O estimare nedeplasată pentru σ^2 =D(X) este $\frac{n}{n-1}\mathrm{S}^2 = 0,21$. Această valoare este dublul valorii testate 0,1. Vrem să vedem dacă nu cumva ea este prea mare. Pentru aceasta folosim faptul că $\frac{n\mathrm{S}^2}{\sigma^2}$ are o distributie χ^2 (Pearson) cu $n-1$ grade de libertate. Găsim că $\frac{n\mathrm{S}^2}{0,1} = 30$. Dar χ^2 cu 14 grade de libertate are o probabilitate $< 0,01$ pentru a lua o astfel de valoare ridicată. Cu pragul 0,05 va trebui deci să respingem ipoteza $\sigma^2 = 0,1$.

4. Într-un oras A, 300 de locuitori din 1500 interogati declară că nu s-au uitat niciodată la TV. Într-un alt oras B, 320 din 1800 declară acelasi lucru. Ce credeti despre ipoteza H: proportia de locuitori care nu se uită la TV este aceeasi în ambele orase. (pragul 0,05).

Solutie Vom accepta pentru cele două v.a. X şi Y ce reprezintă numărul acelora care nu se uită deloc la TV că au o distributie uniformă. Presupunem că $\mathrm{X}_1,...,\mathrm{X}_{1500}$ şi $\mathrm{Y}_1,...,\mathrm{Y}_{1800}$ sunt independente. Avem că $m = 1500$, $n = 1800$, $\overline{\mathrm{X}} = \frac{300}{1500} = 0,2$, $\overline{\mathrm{Y}} = \frac{320}{1800} = 0,172$; $\mathrm{S}^2_{\mathrm{X}} = \frac{1}{m}\sum \overline{\mathrm{X}}_i^2 - \overline{\mathrm{X}}^2 = \overline{\mathrm{X}}\left(1-\overline{\mathrm{X}}\right)$, $\mathrm{S}^2_{\mathrm{Y}} = \overline{\mathrm{Y}}\left(1-\overline{\mathrm{Y}}\right)$; $|\mathrm{T}_0| = \frac{|\overline{\mathrm{Y}}-\overline{\mathrm{X}}|}{\sqrt{\frac{1}{m}\mathrm{S}^2_{\mathrm{X}}+\frac{1}{n}\mathrm{S}^2_{\mathrm{Y}}}} = 1,61$.

O variabilă gaussiană redusă are o probabilitate mai mare decât 0,1 pentru a lua o valoare atât de mare. Prin urmare, la pragul de 0,05 ipoteza H se acceptă.

5. Pentru a măsura o masă μ *putem folosi două procedee A şi B. Rezultatul măsurătorii lui* μ *prin procedeul A este o v.a. gaussiană X cu media* μ *şi cu dispersia* σ^2_{X}*. Prin procedeul B, avem o variabilă gaussiană Y cu media* μ *şi dispersia* σ^2_{Y}*. S-au făcut 8 măsuri independente pentru* μ *prin procedeul A care au condus, la o dispersie de selectie empirică egală cu 0,24. S-au făcut apoi 12 măsurători independente pentru aceeasi masă* μ *prin procedeul B şi s-a obtinut dispersia de selectie 0,08. Ce credeti despre ipoteza H: cele două procedee A şi B au aceeasi precizie? (prag* $\varepsilon = 0,05$*)*

Soluţie. Testul se referă la ipoteza H: $\sigma^2_{\mathrm{X}} = \sigma^2_{\mathrm{Y}}$, cu $m = 8$, $n = 12$, $\mathrm{S}^2_{\mathrm{X}} = 0,24$, $\mathrm{S}^2_{\mathrm{Y}} = 0,08$. De aici s-ar putea crede că $\sigma^2_{\mathrm{X}} > \sigma^2_{\mathrm{Y}}$. Vrem să vedem dacă nu cumva raportul $\frac{S^2_{\mathrm{X}}}{S^2_{\mathrm{Y}}}$ nu este prea mare (relativ la pragul ε) pentru ca ipoteza H să fie verificată.

Stim că variabilele aleatoare $\frac{mS^2_{\mathrm{X}}}{\sigma^2_{\mathrm{X}}}$ şi $\frac{nS^2_{\mathrm{Y}}}{\sigma^2_{\mathrm{Y}}}$ sunt χ^2 cu $m-1=7$ şi respectiv $n-1=11$ grade de libertate. Dacă cumva ipoteza H ar fi verificată v.a. U=$\frac{mS^2_{\mathrm{X}}/(m-1)}{nS^2_{\mathrm{Y}}/(n-1)}$ ar avea o distributie F (Fisher-Snedecor) cu $\left\{ \begin{array}{c} m-1 \\ n-1 \end{array} \right\} = \left\{ \begin{array}{c} 7 \\ 11 \end{array} \right\}$ grade de libertate. În cazul nostru U$=\frac{8\times 0,24/7}{12\times 0,08/11} = 3,14$.

Se vede din cercetarea directă a tabelului pentru distributia F că probabilitatea ca v.a. acesta să ia o astfel de valoare este ceva mai mică decât 0,05. Vom fi deci obligati să respingem ipoteza că σ_{X}^2 este cu mult mai mare decât σ_{Y}^2. Totusi nu putem avea mare certitudine că σ_{X}^2 $=\sigma_{\mathrm{Y}}^2$.

11.6 Exerciţii propuse

1. Se arunca o moneda de 64 de ori. Testati la un nivel de semnificatie de 5% daca moneda este corecta, sau daca moneda este contrafacuta in favoarea "capului", sau daca

a) apare "capul" de 38 ori;

b) apare "capul" de 42 ori.

R: a) z=1,5, corecta; b) z=2,5, contrafacuta.

2. Un producator de hrana pentru caini pretinde ca 8 din 10 caini prefera hrana produsa de el hranei produse de alti producatori. Se aleg 120 de caini si se gasesc 88 care sa prefere aceasta hrana. Testati cu un nivel de 5% daca pretentia producatorului este corecta.

R: z=-1,826, afirmatia producatorului se respinge.

3. O masina produce componente a caror greutati variaza dupa o lege normala cu greutatea medie de 15,4 g si cu deviatia standard $\sigma = 2,3$ g. Masina s-a uzat cu timpul si se banuieste ca greutatea medie a pieselor produse de ea s-a micsorat. Se face o selectie aleatoare de 81 de piese si se gaseste ca masa medie a lor $\overline{x} = 15$ g. Oare acest lucru ne indreptateste pe noi sa credem cu un nivel de semnificatie de 5% ca greutatea medie a pieselor s-a micsorat? Presupunem ca deviatia standard σ nu s-a schimbat.

R: z=-1,565, Nu.

4. Un producator de castet audio pretinde ca o caseta care dureaza 90 min. de obicei, dureaza de fapt 92 min. in medie cu deviatia standard $\sigma = 1,8$ min. Se selecteaza 36 de benzi si se incearca. Cel care verifica casetele respinge pretentia producatorului la un nivel de 5% si spune ca media de timp al unei casete este mai mica decat 92 min. Ce puteti spune despre valoarea mediei de selectie pe care a cercetat-o statisticianul, valoare care l-a condus la decizia luata?

R: $\overline{x} < 91,51$ min.

5. După un sondaj făcut asupra a 300 de studenţi s-a găsit că 35% fumează. Se poate respinge ipoteza $H_0 : p = 0,4$ în favoarea ipotezei $H_1 : p \neq 0,4$ la un prag de semnificaţie de 0,05?

6. Un esantion de volum 25 a fost extras dintr-o populatie normala, $X \sim N(m,4)$*. Media de selectie* $\overline{x}$ *este 10,72. Fie ipoteza nula* $H_0 : m = 10$ *si ipoteza alternativa: a)* $H_1 : m > 10$*; b)* $H_1 : m \neq 10$*. Gasiti in ambele cazuri nivelul de semnificatie la acre trebuie sa respingem ipoteza nula in favoarea ipotezei alternative.*

R: a) $\varepsilon \geq 3,59\%$*; b)* $\varepsilon \geq 7,18\%$*.*

7. Un producator de becuri sustine ca media de viata a unui bec electric produs de el este de 2000 de ore. Se iau 64 de becuri la intamplare si se testeaza viata lor in ore, x*.*

Se obtine $\sum x = 127808$, $\sum (x - \overline{x})^2 = 9694,6$*. Oare la un nivel de semnificatie de 2% putem spune ca producatorul si-a supraestimat produsul? Presupunem ca durata de viata a unui bec are distributia normala.*

R: $n = 64 > 30$ *este mai mare; deci estimam* σ *cu* $\widehat{\sigma} = S = \sqrt{\sum (x - \overline{x})^2 / n}$*.*

Folosim statistica $Z = \frac{\overline{X} - m}{\widehat{\sigma}/\sqrt{n}}$ *si gasim* $z_{calc} = -1,95$ *(test unilateral). Producatorul* **nu** *si-a supraestimat produsul la nivelul de 2%.*

8. Dintr-o populatie normala X *se extrage un esantion de volum 40 cu* $\sum x = 24$ *si* $\sum x^2 = 596$*. Testati (test bilateral) la nivelul de semnificatie de 5% daca media populatiei este 0 (estimati* σ *cu* S*).*

R: $z = 0,995$*. Se accepta afirmatia.*

9. La un examen se analizeaza punctajul x *obtinut de fiecare candidat in parte. Se aleg 250 de candidati si gasim ca* $\sum x = 11872$ *si* $\sum x^2 = 646193$*. Gasiti un interval de incredere de nivel 90% pentru media* m*. Testati ipoteza* $H_0 : m = 49,5$ *impotriva ipotezei* $H_1 : m < 49,5$ *cu nivelul de semnificatie* $\alpha\%$ *si gasiti* α *a.i.* H_0 *sa fie respinsa.*

R: $(45,6; 49,4)\,; \alpha > 4$*.*

10. Un esantion de volum 8 dintr-o populatie normala are $\overline{x} = 4,65$, $\sum (x - \overline{x})^2 = 0,74$*. Testati la un nivel de semnificatie de 2% daca media distributiei este 4,3.*

R: Folositi statistica $T = \frac{\overline{X} - m}{S/\sqrt{n-1}} \in t(n-1)$*, deoarece esantionul este mic* $(8 < 30)$*;* $t_{calc} = 3,05$ *(test bilateral). Respingem afirmatia.*

11. O masina produce ace de otel de lungume 2 cm. Se face o selectie de 10 ace si se gaseste: 1,98; 1,96; 1,99; 2,00; 2,01; 1,95; 1,97; 1,96; 1,97; 1,99. Presupunand ca lungimile acelor sunt normal distribuite, testati cu 1% daca masina functioneaza bine sau nu.

R: $n = 10 < 30$*; folosim deci statistica T;* $t_{calc} = -3,601$*. Nu functioneaza bine.*

12. Se aleg 8 femei la intamplare si li se masoara nivelul de colesterol: 3,1; 2,8; 1,5; 1,7; 2,4; 1,9; 3,3; 1,6.

Vrem sa testam daca nivelul mediu de colesterol este 3,1.

a) Presupunand ca selectia face parte dintr-o populatie normala aratati de ce testul T este cel mai potrivit.

b) Aplicati testul asupra mediei 3,1 si gasiti cu nivelul de semnificatie de 2% daca media 3,1 este corecta sau nu.

c) Fasiti un interval de incredere de 90% pentru nivelul mediu al colesterolului.

R: b) Resping H_0 *(m=3,1); c) (1,81;2,72).*

13. Teoria prezice realizarea unui eveniment A cu probabilitatea p=0,4. Se experimenteaza realizarea evenimentului A de 400 de ori. Din cele 400 de incercari A se realizeaza de 140 de ori. Testati cu nivelul de semnificatie de 1% daca p este mai mica sau nu decat 0,4.

R: Se modeleaza cu proportia de succese: $P_s \sim N\left(p, \frac{pq}{n}\right)$*;* $Z=\frac{P_s-p}{\sqrt{\frac{pq}{n}}}$; $z_{calc} = -2,04$. *Nu.*

14. Se cerceteaza proportia de studenti care au un calculator personal. Din 200 de studenti, 143 au un calculator personal. Testati cu un nivel de 5% ipoteza H_0*:p=75% (probabilitatea ca un student sa aiba un calculator) contra ipotezei* H_1 *: probabilitatea p este mai mica decat 75%.*

R: $z_{calc} = -1,143$. *Se accepta.*

15. Gasiti probabilitatile erorilor de tipul I si celor de tipul II in testarea urmatoarelor ipoteze. Se stie ca o cutie poate sa contina fie (H_0) *10 jetoane albe si 90 negre, fie* (H_1) *50 jetoane albe si 50 negre. Pentru a testa ipoteza* H_0 *impotriva ipotezei* H_1 *se aleg 4 jetoane din cutie fara a le pune inapoi. Daca toate 4 jetoane sunt negre, accept* H_0. *Altfel, o resping (regula de decizie a testului).*

Indicatie H_0 *: cutia contine 10 jetoane albe si 90 negre.*

H_1 *: cutia contine 50 jetoane albe si 50 negre.*

P(Eroare I)=P(resping $H_0 \mid H_0$ *este adevarata)=P(cel putinun jetoneste alb|in cutie sunt 10 albe si 90 negre)=1-P(toate 4 sunt negre|in cutie sunt 10 albe si 90 negre)=*$1-\frac{90}{100}\cdot\frac{89}{99}\cdot\frac{88}{98}\cdot\frac{87}{97} = 1-0,652 = 0,348$.

P(Eroare II)=$0,059$.

16. Pentru datarea specimenelor arheologice se foloseste faptul ca aceste specimene emit particule radioactive. Numarul de particule emise in n minute are o distributie Poiison cu parametrul $n\lambda$, *unde* λ *este un parametru ce depinde de varsta specimenului.*

Se fac doua ipoteze asupra varstei unui specimen

H_0 *: specimenul este de 7000 de ani* $(\lambda = 0,1)$

H_1 *: specimenul este de 15000 de ani* $(\lambda = 4,0)$

S-a decis sa se contorizeze numarul X al particulelor radioactive emise in n minute de specimen si

Acceptam H_0, *daca* $X \leq 1$ *(respingem* H_1*)*

Acceptam H_1, *daca* $X \geq 2$ *(respingem* H_0*)*

Daca $n = 1$, *se cere: a) P(resping* $H_0 \mid H_0$ *=adevarata) si b) P(resping* $H_1 \mid H_1$ *=adevarata).*

Presupunem acum ca P(resping $H_1 \mid H_1$ *=adevarata)*$\leq 0,001$; *aratati ca numarul minim de minute complete necesare inregistrarii este de 3 minute.*

Pentru acest numar de 3 minute sa se calculeze P(resping $H_0 \mid H_0$ *=adevarata).*

Indicatie $X \sim Po\,(n\lambda)$. *Daca* $n = 1$, $X \sim Po\,(\lambda)$ *si a) P(resping* $H_0 \mid H_0$ *=adevarata)=*$P(X \geq 2 \mid X \sim Po\,(1,0))$. $P(X \geq 2) = 1 - P(X = 0) - P(X = 1) = 1 - e^{-1} - e^{-1} = 1 - 2e^{-1} = 1 - 0,736 = 0,264$. *Deci P(resping* $H_0 \mid H_0 =$ *adevarata)*$= 0,246$.

b) P(resping $H_1 \mid H_1$ *=adevarata)=*$P(X \leq 1 \mid X \sim Po\,(4)) = 5e^{-4} = 0,092$. *Daca P(resping* $H_1 \mid H_1$ *=adevarata)*$\leq 0,001$, *atunci* $P(X \leq 1 \mid X \sim Po\,(4n)) \leq 0,01$. *Daca* $X \sim Po\,(4n)$, *atunci* $P\,(X \leq 1) = e^{-4n}\,(1 + 4n) \leq 0,001$ *si gasim* $n = 3$ *si P(resping* $H_0 \mid H_0$ *=adevarata)=* $P\,(X \geq 2 \mid X \sim Po\,(3 \cdot 1)) = 1 - 4e^{-3} = 0,801$.

17. Se fac doua ipoteze asupra functiei densitate de probabilitate pentru o v.a. X.

$$H_0 : f\,(x) = \begin{cases} \frac{1}{4}\,(x+1)\,, 0 < x < 2 \\ 0,\ altfel \end{cases}$$

$$H_1 : f\,(x) = \begin{cases} \frac{1}{4}x^3, 0 < x < 2 \\ 0,\ altfel \end{cases}$$

Se construieste urmatorul test. Se face o singura observatie asupra v.a. X si daca $X < k$, *cu k dat,* $0 < k < 2$, *atunci* H_0 *se accepta. Altfel* H_1 *se accepta: a) Gasiti k a.i.* $P($*Eroare I*$) = 0,1$. *b) Cu valoarea lui k de la a) gasiti* $P($*Eroare II*$)$

Indicatie *Faceti graficele pentru* $f(x)$ *in fiecare din cazurile* H_0 *si* H_1. *P(accept* $H_1 \mid H_0$ *=adevarata)*$= 0,1$ *implica* $P\big(X \geq k \mid f\,(x) = \frac{1}{4}\,(x+1)\big) = 0,1$ *sau* $\int_k^2 \frac{1}{4}\,(x+1)\,dx = 0,1$, *de unde* $k = 1,86$.

$P($*Eroare II*$) = P\left(X < 1,86 \mid f\,(x) = \frac{1}{4}x^3\right) = \int_0^{1,86} \frac{1}{4}x^3 dx = 0,748$.

18. Dintr-o populatie normala $N(m, 36)$ *se ia un esantion de volum 100. Un cercetator vrea sa testeze ipotezele* $H_0 : m = 65$, $H_1 : m > 65$. *El decise sa foloseasca urmatoarea regula de decizie:*

accept H_0 *daca media de selectie* $\overline{x} \leq 66,5$

resping H_0 *daca* $\overline{x} > 66,5$.

a) Gasiti $P($*Eroare I*$)$;

b) Daca el foloseste alternativa $H_1 : m = 67,9$, *gasiti* $P($*Eroare II*$)$.

c) Ce valoare critica trebuie sa considere el pentru media de selectie daca vrea ca

$$P(\textit{Eroare I}) = P(\textit{Eroare II})$$

Indicatie *a) Sub* H_0 *avem:* $\overline{X} \sim N\left(65, \frac{36}{100}\right)$. *El respinge* H_0 *daca* $\overline{x} > 66,5$, *adica* $P\left(\overline{X} > 66,5\right) = P\,(Z > 2,5) = 0,00621$. *Deci* $P($*Eroare I*$) = 0,00621$.

b) Daca el considera $H_1 : m = 67,9$, *atunci sub* H_1 *avem ca* $\overline{X} \sim N\left(67,9, \frac{36}{100}\right)$ *si*

$$P(\textit{Eroare II}) = P\,(\textit{accept } H_0 \mid H_1 \textit{este adevarata}) = P\left(\overline{X} \leq 66,5 \mid m = 67,9\right)$$

$$P\left(\overline{X} \leq 66,5 \mid m = 67,9\right) = P\left(Z \leq -2,333\right) = 0,00982.$$

c) $P\big(\overline{X} > \overline{x} \mid H_0\text{este adevarata}\big) = P\left(\overline{X} \leq \overline{x} \mid H_1\text{este adevarata}\right)$, *deci* $\overline{x}$ *se afla la mijlocul segmentului* $[65; 67,9]$, *adica* $\overline{x} = 66,45$.

19. Ingredientele care se amesteca pentru a forma betonul au asemenea proportii incat rezistenta medie la rupere sa fie de 2000N. Daca rezistenta medie la rupere cade sub 1800N atunci compozitia trebuie schimbata. Distributia rezistentei de rupere este normal distribuita cu deviatia standard de 200N.

Se iau esantioane pentru a se cerceta ipotezele:

$H_0 : m = 2000N$

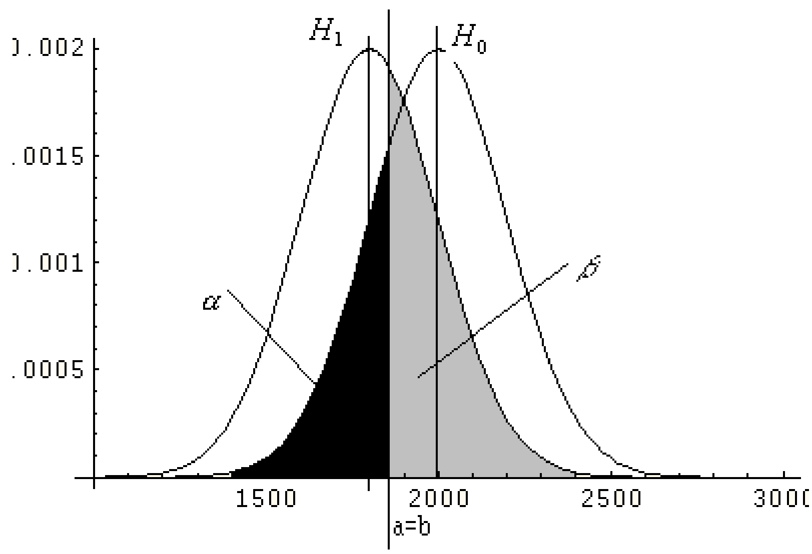
$H_1 : m = 1800N$

Cate esantioane trebuie testate pentru ca sa avem:

$P(\textit{Eroare I}) = \alpha = 0,05$ *si*

$P(\textit{Eroare II}) = \beta = 0,1.$

Indicatie *Sub* H_0 *avem ca* $X \sim N\left(2000, 200^2\right)$, *deci* $\overline{X} \sim N\left(2000, \frac{200^2}{n}\right)$



Daca Z_α *este cuantila ce corespunde lui* $\alpha = 0,05$, *avem ca* $Z_\alpha = -1,645$ *si valoarea corespunzatoare a lui* $\overline{X}$ *este* $a = 2000 - 1,645\left(\frac{200}{\sqrt{n}}\right)$. *Sub* H_1, $\overline{X} \sim N\left(1800, \frac{200^2}{n}\right)$, *deci, pentru* $Z_\beta = 1,282$ $(\beta = 0,1)$ *avem ca* $b = 1800 + 1,282\left(\frac{200}{\sqrt{n}}\right)$. *Din* $a = b$ *gasim* $n = 8,57$, *deci vom lua 9 probe.*

20. Se aleg aleator esantioane de 400 seminte dintr-un anumit sortiment. Probabilitatea ca o seminta sa germineze este egala cu α. *Sa notam cu* X *v.a. ce reprezinta numarul de seminte care au germinat din totalul semintelor dintr-un esantion.*

Folositi o aproximare convenabila a modelului probabilistic cu un model normal pentru a determina:

a) $P(X \leq 340 \mid \alpha = 0,9)$

b) $P(X \geq 340 \mid \alpha = 0,8)$

c) Controlorul de calitate al semintelor stie ca α *este fie* $0,8$, *fie* $0,9$. *Sa presupunem ca, de fapt, din totalul de 400 de semninte dintr-un esantion au germinat numai* x. *Controlorul decide ca valoarea lui* α *este* $0,8$ *daca*

$$Z = P(X \geq x \mid \alpha = 0,8) - P(X \leq x \mid \alpha = 0,9)$$

este pozitiv. Altfel el decide ca $\alpha = 0,9$. *Gasiti care este decizia controlorului pentru fiecare dintre cazurile* $x = 330$, $x = 340$, $x = 350$.

R: a) 0,000577; b)0,00738; c)0,8; 0,8; 0,9.

Lecţia 12

Testul neparametric χ^2

12.1 Principiul testului χ^2

Vom introduce acest test prin analiza atentă a unui exemplu.

Exemplul 12.1 *(fictiv) S-a făcut un sondaj într-un oras din România pe un esantion de 200 de tineri bărbati de 27 de ani, în 1985, asupra situatiei pregătirii lor scolare. Ei au fost împărtiti în 6 categorii după cum urmează:*

1. *cu studii superioare terminate;*
2. *cu studii superioare neterminate dar începute;*
3. *cu liceul de 12 ani terminat;*
4. *cu liceul de 12 ani neterminat dar început;*
5. *cu scoala generală de 8 ani terminată;*
6. *cu scoala generală de 8 ani începută şi neterminată.*

Acelasi sondaj se repetă (în aceleasi conditii) în 1995. Vom nota cu $f_{o,j}, j = 1, 2, ..., 6$ frecventa (absolută) observată în 1995 pentru categoria "j" din cele 6 expuse mai sus. Vom nota cu $f_{s,j}$ frecventa (sperată, la care ne asteptăm şi în 1995!) găsită în 1985. Iată tabelul cu cele două sondaje:

categoria "j"	*Frecventa observată în 1995, $f_{o,j}$*	*Frecventa observată în 1985, $f_{s,j}$*
1	35	36
2	40	34
3	83	64
4	16	26
5	26	34
6	0	6
	– – –	– – –
	200	200

Facem următoarea ipoteză statistică (inferentă statistică) H_0:<<distributia populatiei în 1995 este aceeasi cu distributia populatiei din 1985>>.

Este "*natural*" (de ce?) ca "discrepanta" sau deviatia dintre cele două situatii să o măsurăm prin suma

$$\sum_{j=1}^{6} \frac{(f_{o,j} - f_{s,j})^2}{f_{s,j}} \tag{12.1}$$

Se arată (Teorema lui K. Pearson) că pentru n mare (cel putin 20-25), în cazul nostru n=200, această sumă tinde către valoarea distributiei χ^2 cu 6-1 grade de libertate în ipoteza că H_0 este adevărată: avem aceiasi distributie, adică distributiile celor două sondaje concordă. În general, dacă am repartizat elementele din selectie în J clase, pentru un volum mare ($n \geq 25$) suma (12.1) cu J în loc de 6, este aproximativ egală cu valoarea distributiei χ^2 cu $J-1$ grade de libertate.

Ipoteza H_0 are sanse să fie adevărată dacă valoarea sumei (12.1) este mică. Găsim

$$\chi^2 = \frac{(35-36)^2}{36} + \frac{(40-34)^2}{34} + \cdots + \frac{(0-6)^2}{6} = 18,46$$

Căutăm în Tabelul III pe linia corespunzătoare $\nu = 5$ (grade de libertate) şi găsim că 16,7496≤18.46≤20,515.

Dar

$$\begin{aligned} P(16,7496 &\leq \chi^2 \leq 20,515) \\ &= F(20,515) - F(16,7496) = 0,999 - 0,995 = 0,004 \end{aligned}$$

deci extrem de mică. Prin urmare, este foarte putin probabil ca cele două distributii să concorde. Deci *se respinge ipoteza H_0 cu probabilitatea* 1-0,004=0,986.

Retinem din Exemplul 1 că ori de câte ori vrem să facem o ipoteză asupra probabilitătii ca două distributii să concorde putem folosi "testul χ^2", adică metodologia din 12.1, cu conditia ca numărul observatiilor să fie mare ($n \geq 25$), grupele de divizare ale sondajului să fie disjuncte şi numărul lor k să respecte următoarele reguli:

$$\begin{aligned} 25 &< n \leq 100, k \in (10,15) \\ 100 &< n < 200, k \in (15,18) \\ 200 &\leq n < 400, k \in (18,20) \\ 400 &\leq n < 1000, k \in (25,30) \\ 1000 &\leq n < 2000, k \in (35,40) \end{aligned}$$

Aceste reguli au fost deduse din practică şi nu teoretic.

Prezentăm acum situatia generală în care putem utiliza testul de concordanta χ^2. Facem următoarea ipoteză: H$_0$:<<Presupunem că populatia P are o functie de probabilitate $g(x)$, specificată (de exemplu normală cu $m = 0$ şi $\sigma = 2$)>>.

Vrem să folosim testul χ^2 pentru a "măsura" câtă dreptate avem să presupunem acest lucru pornind de la un sondaj de volum n: $x_1, x_2, ..., x_n$ din populatia P. După regulile de mai sus împărtim datele $x_1, x_2, ..., x_n$ în J grupe disjuncte. Notăm cu ν_j frecventa absolută în grupa "j"(numărul acelor x_i care se află în grupa "j"). Notăm cu p_j probabilitatea teoretică (obtinută cu ajutorul functiei de probabilitate $g(x)$, sau cu ajutorul functiei de repartitie corespunzătoare $G(x)$) ca un element x din populatia P să se afle în grupa j. Atunci frecventa teoretică este np_j, şi suma care măsoară deviatia din formula (12.1) devine:

$$d = \sum_{j=1}^{J} \frac{(\nu_j - np_j)^2}{np_j} \tag{12.2}$$

Calculăm acest număr d. Privim Tabelul III şi încercăm să "estimăm" probabilitatea ca χ^2 să aibă valoarea d. De obicei se fixează *un prag de semnificatie* $\alpha \in (0, 1)$. Se caută cuantila χ^2_α corespunzătoare acestui prag, adică acea valoare a lui χ^2 astfel încât functia de repartitie a lui χ^2 să aibă valoarea α. Cum $\mathrm{P}(\chi^2 \leq \chi^2_\alpha) = \alpha$, din definitia functiei de repartitie, vom face următorul rationament:

• dacă $d \geq \chi^2_\alpha$, *vom accepta ipoteza* H_0 (distributia populatiei este de forma prescrisă) *cu pragul de semnificatie* α.

• dacă $d < \chi^2_\alpha$, *vom respinge ipoteza* H_0 *cu pragul de semnificatie* α.

Acesta este testul de concordantă χ^2. De obicei α se ia mic: α=0,05; 0,01; 0,001.

Exemplul 12.2 *Se fac 500 de măsurători asupra erorilor date de un aparat de măsură de la bordul unui avion. Ele se împart în 8 intervale consecutive. Freventele absolute cu care apar aceste erori pe un interval sunt date în următorul tabel:*

I_j :	$[-4;-3)$	$[-3;-2)$	$[-2;-1)$	$[-1;0)$	$[0;1)$	$[1;2)$	$[2;3)$	$[3;4)$
ν_j :	6	25	72	133	120	88	46	10

Utilizând testul χ^2 *să se verifice cu pragul de semnifictie* $\alpha = 0,95$ *dacă distributia erorilor este normală cu media estimată la* $m = 0,168$ *şi dispersia estimată la* $\sigma^2 = 1448^2$.

Solutie Aici avem un exemplu mai complicat deoarece cei doi parametri au deja estimări date. Vom avea mereu două relatii de legătură: $0,168 = \frac{1}{500}\sum x_i$ şi $1448^2 = \frac{1}{500}\sum (x_i - 0,168)^2$.

Deci "*numărul gradelor de libertate*" va scădea cu doi. Prin urmare χ^2 va avea 8-1-2=5 grade de libertate.

Calculăm probabilitatea teoretică pe fiecare interval $[x_i, x_{i+1})$:

$$p_j = \Phi\left(\frac{x_{j+1} - m}{\sigma}\right) - \Phi\left(\frac{x_j - m}{\sigma}\right)$$

, unde Φ este funcţia lui Laplace. Găsim pentru np_j următoarele valori (corespunzătoare intervalelor precizate deja în tabelul de mai sus): 6,2; 26,2; 71,2; .122,2; 131,8; 90,5; 32,8; 10,5. Calculăm $d = \sum_{j=1}^{8} \frac{(v_j - np_j)^2}{np_j} = 3,94$.

Cuantila $\chi^2_{0,95}$ pentru 5 grade de libertate este 11,0705 (vezi Tabelul III). Deci P($\chi^2 \leq 11,0705$)=0,95,adică P($\chi^2 > 11,0705$)=0,05.

Cum $d = 3,94 < 11,0705$ acceptăm ipoteza de normalitate cu probabilitatea 0,95 (adică cu 95%), sau cu pragul $\alpha = 0,05$.

Vom prezenta acum testul χ^2 (se citeste chi pătrat sau chi doi) dintr-un punct de vedere mai general.

Fie evenimentele $\{A_1, A_2, ..., A_r\}$ cu probabilitatea p_i =P(A_i), $i = \overline{1,r}$, cunoscute mai mult sau mai putin. Putem cere ca $\sum p_i = 1$, adică $\{A_1, A_2, ..., A_r\}$ să fie o descompunere a evenimentului singur. În esentă testul χ^2 îşi propune să testeze o ipoteză oarecare H privind probabilităile $p_1, ..., p_r$. Se fac sondaje de volum mare pentru fiecare A_i, $i = \overline{1,r}$. Se notează cu Y_i numărul acelor probe care sunt favorabile evenimentului A_i.

Cazul I Se dau numerele $p'_1, p'_2, ..., p'_r \geq 0$ cu $\sum p'_i = 1$ şi se consideră ipoteza H: $p_1 = p'_1$, $p_2 = p'_2, ..., p_r = p'_r$ (problemă de concordantă). Dacă H este adevărată statistica (Helmert-Pearson)

$$T = \sum_{1 \leq i \leq r} \frac{(Y_i - np'_i)^2}{np'_i} = \left(\sum_i \frac{Y_i^2}{np'_i} \right) - n \tag{12.3}$$

are practic (adică pentru n mare) distributia χ^2 cu $r-1$ grade de libertate. Pentru a arăta ultima egalitate în (12.3) am folosit egalitătile $\sum Y_i = n$ şi $\sum p_i = 1$. Testul constă în a respinge ipoteza H dacă T ia o valoare semnificativ prea mare relativ la χ^2 (vezi exemplul 12.1). De exemplu dacă se dau v.a. X şi legea Q cunoscută, pentru a testa dacă H: X are distributia Q, descompunem dreapta reală $\mathbb{R}$ în subintervale disjuncte: $\mathbb{R} = A_1 \cup A_2 \cup \cdots \cup A_r$, considerăm evenimentele $(X \in A_i)_{i=\overline{1,r}}$ şi punem p'_i =Q(A_i).

Cazul II Fie acum r functii pozitive $f_1(x_1, ..., x_s), ..., f_r(x_1, ..., x_s)$ de variabile $x_1, ..., x_s$ cu $s < r$, astfel încât $f_1(x_1, ..., x_s) + \cdots + f_r(x_1, ..., x_s) = 1$. Considerăm ipoteaza H: Există numerele $\lambda_1, ..., \lambda_s$ (estimate printr-un procedeu oarecare) astfel încât $p_i = f_i(\lambda_1, ..., \lambda_s)$ pentru $1 \leq i \leq r$. Dacă H este adevărată se pot defini (plecând de la valorile $Y_1, ..., Y_r$) estimatori convenabili pentru $\lambda_1, ..., \lambda_s$. Fie $\widehat{\lambda}_1 = g_i(Y_1, ..., Y_r)$, $1 \leq i \leq s$, un estimator pentru λ_i. De aici găsim estimtori convenabili ai probabilităţilor p_i: $\widehat{p}_i = f_i\left(\widehat{\lambda}_1, ..., \widehat{\lambda}_s\right)$ pentru $1 \leq i \leq r$. Statisica

$$T = \sum_{1 \leq i \leq r} \frac{(Y_i - n\widehat{p}_i)^2}{n\widehat{p}_i} = \left(\sum_i \frac{Y_i^2}{n\widehat{p}_i} \right) - n \tag{12.4}$$

are practic o distributie χ^2 cu $r - s - 1$ grade de libertate (apar încă s legături datorită relatiilor de estimare). Testul constă în a respinge ipoteza H dacă T ia o valoare semnificativ prea mare relativ la χ^2 (vezi Exemplul 2). Se pune acum problema de a găsi estimtori "*buni*" pentru $\lambda_1, ..., \lambda_n$. Iată câteva reguli bazate pe însăsi definitia "*formulei de deviatie*" (Helmert-Pearson) dată în (12.1) şi (12.2) şi pe alte notiuni ce apar în Lectia 9.

Regula I (χ^2 minim) Se iau pentru $\widehat{\lambda}_1, ..., \widehat{\lambda}_s$ acele functii de $\mathrm{Y}_1, ..., \mathrm{Y}_r$ care minimizează expresia (deviaţia):

$$d = \sum_i \frac{(\mathrm{Y}_i - n\widehat{p}_i)^2}{n\widehat{p}_i} = \min \tag{12.5}$$

adică acele $\widehat{\lambda}_1, ..., \widehat{\lambda}_s$ care verifică ecuatiile diferentiale.

$$\sum_{1 \le i \le r} \frac{\mathrm{Y}_i^2}{\widehat{p}_i^2} \cdot \frac{\partial \widehat{p}_i}{\lambda \widehat{\lambda}_j} = 0, 0 \le j \le s \tag{12.6}$$

Se pune conditia grad $d\left(\widehat{\lambda}_1, ..., \widehat{\lambda}_s\right) = 0$şi se foloseste faptul că $\sum\limits_i \frac{\partial \widehat{p}_i}{\partial \widehat{\lambda}_j} = 0$,deoarece $\sum_j \widehat{p}_j = 1$.

Regula II (verosimilitatea maximă) După un rationament asemănător cu acela din Lectia 9 (Principiul verosimilitătii maxime) se deduce că $\widehat{\lambda}_1, ..., \widehat{\lambda}_n$ trebuie să verifice ecuatiile diferentiale:

$$\sum_{1 \le i \le r} \frac{\mathrm{Y}_i}{\widehat{p}_i} \cdot \frac{\partial \widehat{p}_i}{\partial \widehat{\lambda}_j} = 0 \tag{12.7}$$

pentru $1 \le j \le s$.

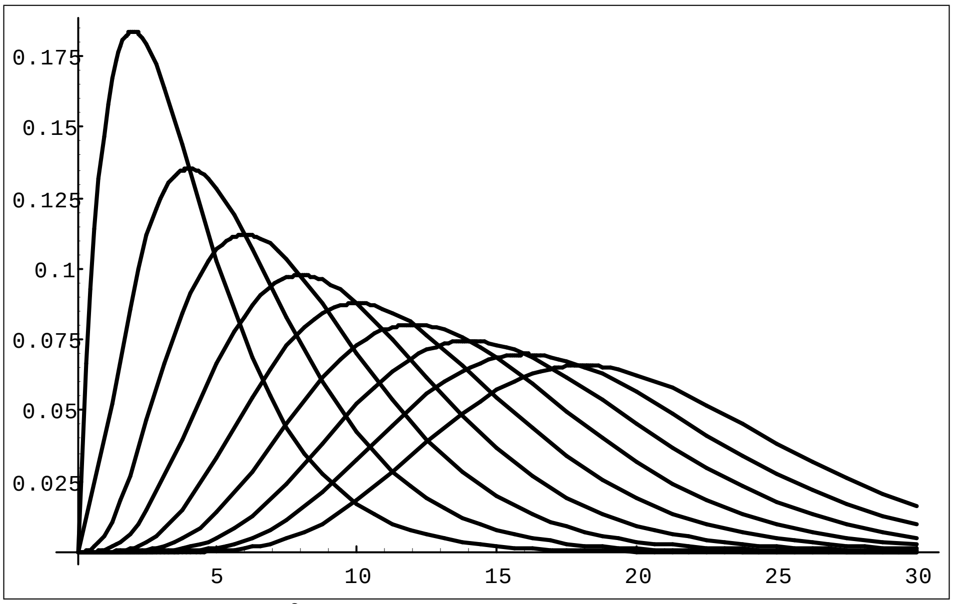
Regula III În locul formulelor (12.6) se pot folosi formulele:

$$\sum_{1 \le i \le r} \frac{\widehat{p}_i}{\mathrm{Y}_i} \cdot \frac{\partial \widehat{p}_i}{\partial \widehat{\lambda}_j} = 0 \tag{12.8}$$

pentru $1 \le j \le s$

(χ^2 minim modificat).

Această distribuţie atât de folosită are pentru diverse valori ale lui n, densităţile ca în graficul următor:



Densităţile χ^2 pentru n=4, 6, 8, 10, 12, 14, 16, 18, 20

Să aplicăm acum cele de mai sus la câteva tipuri de probleme.

12.1.1 Teste asupra formei unei distributii

Fie P=Q$(\lambda_1, ..., \lambda_s)$ o familie de legi pe $\mathbb{R}$ care depinde de s parametrii $\lambda_1, ..., \lambda_s$ şi X o v.a. aleatoare. Pentru a testa ipoteza: H: X *se supune unei anume legi din familia* P, împărtim pe $\mathbb{R}$ în r subintervale: $\mathbb{R}$ =A$_1 \cup \cdots \cup$A$_r$, şi punem $f_i\left(\lambda_1, ..., \lambda_s\right)$ =Q$(\lambda_1, ..., \lambda_s)\left(\mathrm{A}_i\right)$.

Exemplul 12.3 *Fie X o v.a. ce poate lua valorile 0, 1, 2,..., k,.... Testăm ipoteza: H: X se supune unei legi Poisson.*

Familia legilor Poisson depinde de un singur parametru $\lambda > 0$. Ca estimator "*bun*" pentru λ se alege $\widehat{\lambda}$= media de selectie (eventual după grupaje convenabile). Avem $s = 1$ în acest caz.

Exemplul 12.4 *Vrem să testăm pentru o v.a. X continuă următoarea ipoteză H: X se supune unei legi gaussiene. Aici P este familia N(μ, σ^2) care depinde de doi parametri: μ şi σ^2. După cum stim este convenabil să estimăm media μ cu media de selectie şi pe σ^2 cu $S'^{\,2}$. Aici $s = 2$.*

12.1.2 Teste de independentă

Fie X şi Y două v.a. definite pe aceeasi categorie de probe. Pentru a testa ipoteza H: X *şi* Y *sunt independente*, se descompune $\mathbb{R}$ în două partitii: $\mathbb{R}$ =E$_1 \cup \cdots \cup$E$_a$=F$_1 \cup \cdots \cup$F$_b$ şi considerăm cele $r = a \cdot b$ evenimente: X$\in$E$_i$ şi Y$\in$F$_j$, pentru $1 \leq i \leq a$, $1 \leq j \leq b$.

Se face un număr mare de sondaje independente de tipul (X_1,Y_1),... (X_n, Y_n). Notăm cu N_{ij} numărul realizărilor $X \in E_i$ şi $Y \in F_j$. Se introduc statisticile

$$\widetilde{N}_{i\bullet} = \sum_{1\le j\le b} N_{ij} \qquad \widetilde{N}_{\bullet j} = \sum_{1\le i\le a} N_{ij} \tag{12.9}$$

$$T = n\sum_{i,j}\frac{\left(N_{ij} - \frac{1}{n}\widetilde{N}_{i\bullet}\widetilde{N}_{\bullet j}\right)^2}{\widetilde{N}_{i\bullet}\widetilde{N}_{\bullet j}} = n\left[\left(\sum_{i,j}\frac{N_{ij}^2}{\widetilde{N}_{i\bullet}\widetilde{N}_{\bullet j}}\right) - 1\right] \tag{12.10}$$

Dacă H este adevărată, T se supune practic unei legi χ^2 cu $(a-1)(b-1)$ grade de libertate. Testul constă în a respinge ipoteza H dacă T ia o valoare semnificativ prea mare relativ la χ^2. Dacă ipoteza H este verificată numărul $\widehat{p}_{ij} = \frac{1}{n^2}\widetilde{N}_{i\bullet}\widetilde{N}_{\bullet j}$ este un estimator bun pentru probabilitatea p_{ij} =P($X \in E_i$şi $Y \in F_j$).

12.1.3 Teste de omogenitate

Se consideră t variabile aleatoare $X_1, ...,X_t$. Pentru a testa ipoteza H: $X_1, ...,X_t$ *satisfac aceeasi lege*, se face partitia $\mathbb{R} = A_1 \cup A_2 \cup \cdots \cup A_r$, pentru fiecare i se ia un esantion $(X_{i1}, ..., X_{in_i})$ de volum mare, $x_i \in X_i$, $n = n_1 + \cdots + n_i$, v.a. $X_{11}, ...,X_{in_i}$ fiind considerate independente. Pentru $1 \le i \le t$ se notează Y_{ij} numărul de indici k astfel încât $X_{ik} \in A_j$. Se introduc statisticile

$$\widetilde{Y}_{\bullet j} = \sum_{1\le i\le t}\widetilde{Y}_{ij} \qquad T = \sum_{\substack{1\le i\le t\\ 1\le i\le r}}\frac{\left(Y_{ij} - \frac{n_i}{n}\widetilde{Y}_{\bullet j}\right)^2}{\frac{n_i}{n}\widetilde{Y}_{\bullet j}} = n\left[\sum_{i,j}\frac{Y_{ij}^2}{n_i\widetilde{Y}_{\bullet j}} - 1\right]. \tag{12.11}$$

Dacă H este adevărată, T se supune practic unei legi χ^2 cu $(t-1)(r-1)$ grade de libertate. Testul constă în a respinge H atunci când T ia o valoare semnificativ mare relativ la χ^2. Dacă H este adevărată $\widehat{p}_j = \frac{1}{n}\widetilde{Y}_{\bullet j}$ este un estimator bun pentru $P(X_i \in A_j)$ şi nici nu depinde de i. Dacă $r = 2$ (12.10) devine

$$T = \frac{n^2}{\widetilde{Y}_{\bullet 1}\widetilde{Y}_{\bullet 2}}\sum_i\frac{Y_{i1}^2}{n_i} - n\frac{\widetilde{Y}_{\bullet 1}}{\widetilde{Y}_{\bullet 2}} \tag{12.12}$$

Exemplul 12.5 *Teoria lui Mendel asupra eredității ne previne că dacă crestem 2 tipuri de plante va trebui să obtinem produse de tipul A, B, C, D în proportie de 9, 3, 3 şi 1. După experiente se observă că s-au obtinut 154 produse de tipul A, 44 de tip B, 63 de tip C şi 21 de tip D. Ce părere aveti în acest caz de teoria lui Mendel (prag $\varepsilon = 0,05$)?*

Solutie Aici evenimentele A_1, A_2, A_3, A_4 sunt A, B, C şi D. Teoria lui Mendel prevede că $p_1 = \frac{9}{16}$, $p_2 = \frac{3}{16}$, $p_3 = \frac{3}{16}$, $p_4 = \frac{1}{16}$ (deoarece 9+3+3+1=16). Suntem în *Cazul I* al testului χ^2

cu: $p_1' = \frac{9}{16}$, $p_2' = \frac{3}{16}$, $p_3' = \frac{3}{16}$, $p_4' = \frac{1}{16}$. Experientele conduc la , $\mathrm{Y}_1 = 154$, $\mathrm{Y}_2 = 44$, $\mathrm{Y}_3 = 63$, $\mathrm{Y}_4 = 21$, $n = 154 + 44 + 63 + 21 = 282$.

Dacă U este variabila χ^2 cu $r-1 = 3$ grade de libertate, avem că P(U$\geq 7,81$)=0,05. Testul de prag 0,05 se scrie: Respinge H$\Longleftrightarrow \chi^2 \geq 7,81$. În cazul nostru avem np_1', $np_2' = np_3' = 53$, $np_4' = 18$ şi

$$\chi^2 = \frac{(154-159)^2}{159} + \frac{(44-53)^2}{53} + \frac{(63-53)^2}{53} + \frac{(21-18)^2}{18} = 4,06 < 7,81$$

. Se acceptă deci teoria lui Mendel ca fiind adevărată cu 0,95=1-0,05.

Exemplul 12.6 *Fie p probabilitatea pentru ca o piesă din echipamentul fabricat de o anumită masină să aibă defecte. Vrem să testăm ipoteza H: $p = 0,2$. Pentru aceasta luăm 100 de piese şi constatăm că 22 dintre ele au defecte. Care este probabilitatea, dacă ipoteza este adevărată, pentru ca χ^2 să aibă o valoare mare? Cum interpretati acest lucru?*

Solutie Aici A_1=succes şi A_2=esec. Fie $p_1 = p =$P(A_1) şi $p = 1 - p = q =$P(A_2). Vom avea $p_1' = p'$ şi $p_2' = q' = 1 - p'$. Fie Y_1=numărul de succese în cursul a n încercări(=Y) şi Y_2=$n-$Y. Avem

$$T = \frac{(\mathrm{Y} - np')^2}{np'} + \frac{[n - \mathrm{Y} - n(1-p')]^2}{n(1-p')} = \frac{(\mathrm{Y} - np')^2}{np'q'} = Z^2$$

unde Z=$\frac{\mathrm{Y}-np'}{\sqrt{np'q'}}$. Testul χ^2 capătă deci următoarea formă: Respinge H$\Longleftrightarrow$ (T $\geq y$) , sau dacă $|\mathrm{Z}| \geq \sqrt{y}$. Se stie că dacă H este adevărată Z este practic gaussiană redusă (teorema limită centrală).

Aplicatie numerică: $p' = 0,2$, $n = 100$, Y= 22. Valoarea lui Z^2 este $\mathrm{Z}^2 = \frac{(22-20)^2}{100\times0,2\times0,8} = \frac{1}{4}$. Probabilitatea ca variabila gaussiană redusă să ia o valoare absolută $> \sqrt{\frac{1}{4}} = 0,5$ este 0,616. Se acceptă deci această ipoteză H.

Exemplul 12.7 *O v.a. X ia numai valorile 0, 1, 2, 3 ,4. Vrem să testăm dacă această v.a. se spune legii binomiale cu $p = \frac{1}{3}$ şi $n = 4$ (numărul de probe). S-au făcut pentru aceasta 324 încercări independente care au condus la următoarele rezultate, f_i fiind frecventa absolută a valorii i:*

i	*0*	*1*	*2*	*3*	*4*
f_i	*67*	*122*	*94*	*38*	*3*

Ce concluzie trageti?

Solutie Fie A_i evenimentul: X=i; P(A_i)=$\mathrm{C}_4^i \cdot \left(\frac{1}{3}\right)^i \cdot \left(\frac{2}{3}\right)^{4-i}$, pentru $0 \leq i \leq 4$. Avem deci $p_0' = \frac{16}{81}$; $p_1' = \frac{32}{81}$; $p_2' = \frac{24}{81}$; $p_3' = \frac{8}{81}$; $p_4' = \frac{1}{81}$; $\mathrm{Y}_0 = 67$; $\mathrm{Y}_1 = 122$; $\mathrm{Y}_2 = 94$; $\mathrm{Y}_3 = 38$; $\mathrm{Y}_4 = 3$. Deoarece Y_4 este mic se grupează A_3 cu A_4 şi găsim A_0, A_1, A_2 şi B=($\mathrm{A}_3 \cup \mathrm{A}_4$) cu p_0', p_1', p_2' şi $q' = p_3' + p_4'$. Experienta a condus la Y_0, Y_1, Y_2 şi Z=Y_3+Y_4. Se obtine χ^2=1,15. Probabilitatea ca χ^2 cu 4-1=3 grade de libertate să ia o valoare de această mărime este $> 0,1$. Deci ipoteza se acceptă.

Exemplul 12.8 *Gazul de esapament al unui motor contine particule solide. Se consideră ipoteza H: numărul X al acestor particule continut într-un volum mic V de gaz se supune unei legi Poisson. Pentru a testa această ipoteză luăm 400 esantioane de acelasi volum V şi se găsesc 1872 de particule reprezentate după următorul tabel (n_i reprezintă numărul de esantione care au continut i particule):*

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
n_i	0	20	43	53	86	70	54	37	18	10	5	2	2	0	0

Aceste rezultate permit oare să acceptăm ipoteza H?

Solutie $\overline{\mathrm{X}} = \frac{1}{n}\sum in_i = \frac{1872}{400} = 4,68$. Se grupează clasele corespunzătoare lui $i = 0$ şi $i = 1$, cât şi clasele corespunzătoare lui $i \geq 10$. Se obtin în final 10 clase numerotate de la 1 la 10 care au probabilitătile estimate prin $\widehat{p_i}$:

$\widehat{p_i} = e^{-4,68}(1 + 4,68)$, $\widehat{p_i} = e^{-4,68} \times \frac{(4,68)^i}{i\,!}$, pentru $2 \leq i \leq 9$, $\widehat{p_i} = \frac{4,68}{i}\widehat{p}_{i-1}$, pentru $3 \leq i \leq 9$.

şi $\widehat{p}_{10} = 1 - \sum\limits_{1\leq i\leq 9} \widehat{p_i}$. Avem următorul tabel:

i	Y_i	$\widehat{p_i}$	$n\widehat{p_1}$	$\frac{(\mathrm{Y}_i - n\widehat{p_i})^2}{n\widehat{p_i}}$
0 sau 1	20	0,0527	21,1	0,0552
2	43	0,1016	40,6	0,1372
3	53	0,1585	63,4	1,7060
4	86	0,1855	74,2	1,8764
5	70	0,1736	69,4	0,0044
6	54	0,1354	54,1	0,0004
7	37	0,0905	36,2	0,0176
8	18	0,0529	21,1	0,4720
9	10	0,0275	11,0	0,0908
≥ 10	9	0,0218	8,7	0,0152
				4,3772

Deci suma termenilor de pe ultima coloană este chiar $\chi^2 \approx 4,38$. Probabilitatea pentru ca o variabilă χ^2 cu $r - s - 1 = 10 - 1 - 1 = 8$ grade de libertate să fie $\geq 4,38$ este $> 0,1$. Prin urmare H se va accepta.

Exemplul 12.9 *Vrem să examinăm dacă aptitudinile manuale ale unui individ sunt independente de vedere. Pentru aceasta se definesc 2 caractere X şi Y:X ia valorile 1, 2 sau 3 care corespund faptului că individul este mai abil cu mâna stângă, la fel de abil cu ambele mâini, sau mai abil cu mâna dreaptă. Y ia valorile 1, 2 sau 3 după cum individul vede mai bine cu ochiul stâng, cu ambii ochi sau vede mai bine cu ochiul drept. Facem deci ipoteza că X şi Y sunt v.a. independente. În tabelul următor avem rezultatele observatiilor făcute asupra a 413* *persoane. De exemplu, am găsit 20 persoane cu X=2 şi Y=3, etc.*

X \ *Y*	*1*	*2*	*3*
1	*34*	*62*	*28*
2	*27*	*28*	*20*
3	*57*	*105*	*52*

.

Solutie Cu notatiile din *Aplicatia 2* din această lectie avem

$$\widetilde{\mathrm{N}}_{1\bullet} = 34 + 62 + 28 = 124$$

$$\widetilde{\mathrm{N}}_{2\bullet} = 75, \widetilde{\mathrm{N}}_{3\bullet} = 214$$

$$\widetilde{\mathrm{N}}_{\bullet 1} = 34 + 27 + 57 = 118$$

$$\widetilde{\mathrm{N}}_{\bullet 2} = 196, \widetilde{\mathrm{N}}_{\bullet 3} = 52in = 34 + 62 + \cdots + 52 = 413.$$

Pentru numerele $\frac{1}{n}\widetilde{\mathrm{N}}_{i\bullet}\widetilde{\mathrm{N}}_{\bullet j}$ avem tabelul:

i \ j	1	2	3
1	35	59	30
2	21	35	18
3	61	101	52

şi

$\chi^2 = \frac{(34-35)^2}{35} + \cdots + \frac{(52-52)^2}{52} = 3,5$. Cum v.a. χ^2 are $(a-1)(b-1) = 2 \cdot 2 = 4$ grade de libertate, ipoteza de independentă se acceptă (vezi Tabelul ?).

Exemplul 12.10 *Inteligenta unui copil se clasează în 6 nivele: de la A— foarte slabă, până la F— foarte bună. S-au clasat 1725 copii la întâmplare alesi din 8 scoli numerotate de la 1 la 8 şi s-au obtinut rezultate următoare:*

scoala \ nivelul	A	B	C	D	E	F
1	6	18	36	43	39	4
2	14	25	52	87	54	13
3	—	—	1	8	37	—
4	14	19	60	94	73	12
5	33	69	132	187	85	14
6	45	50	69	72	66	15
7	—	6	20	8	9	1
8	18	32	37	36	12	—

Aceste rezultate ne permit oare să acceptăm ipoteza după care repartitia nivelelor de inteligenta este aceeasi în oricare dintre scoli?

Solutie Este vorba despre un test de omogenitate (vezi *Aplicatia 3*). Aici $t = 8$, $r = 6$, numerele Y_{ij} se află în tabel la intersectia liniei i cu coloana j, de exemplu $\mathrm{Y}_{23} = 52$. Avem $n_1 = 6+18+36+43+39+4 = 146$, $n_2 = 243$, $n_3 = 46$, $n_4 = 272$, $n_5 = 520$, $n_6 = 317$, $n_7 = 44$, $n_8 = 135$. La fel $\widetilde{\mathrm{Y}}_{\bullet 1} = 6 + 14 + 14 + 33 + 44 + 18 = 130$, $\widetilde{\mathrm{Y}}_{\bullet 2} = 219$, $\widetilde{\mathrm{Y}}_{\bullet 3} = 407$, $\widetilde{\mathrm{Y}}_{\bullet 4} = 535$, $\widetilde{\mathrm{Y}}_{\bullet 5} = 375$, $\widetilde{\mathrm{Y}}_{\bullet 6} = 59$. Calculăm acum numerele $\frac{\mathrm{Y}_{11}^2}{n_1\widetilde{\mathrm{Y}}_{\bullet 1}} = 0,001897$,..., $\frac{\mathrm{Y}_{85}^2}{n_8\widetilde{\mathrm{Y}}_{\bullet 5}} = 0,002844$ şi suma lor $\sum\limits_{i,j} \frac{\mathrm{Y}_{ij}^2}{n_i\widetilde{\mathrm{Y}}_{\bullet j}} = 1,0784$. De aici găsim că $\chi^2 = 1725 \times (1,0784 - 1) = 135$. Variabila aleatoare χ^2 cu $(t-1)(r-1) = 35$ grade de libertate are o probabilitate mai mică decât 0,001 ca să fie ≥ 135. Deci ipoteza va trebui categoric respinsă.

12.2 Rezumat

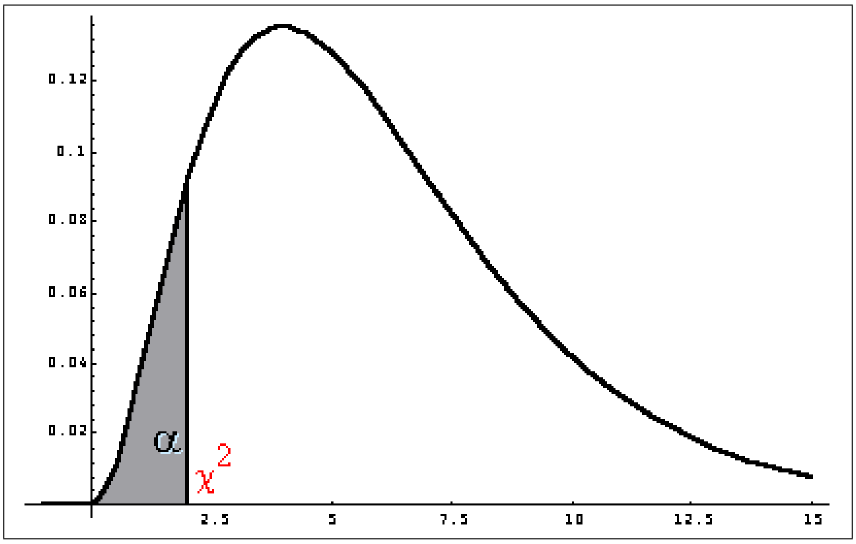
Fie X o v.a. ce guvernează populatia P. Nu cunoastem legea statistică a v.a. X şi vrem să testăm dacă această lege este o lege cunoscută Q (dată). Pentru aceasta considerăm un esantion de volum n ($n \geq 25$): $x_1, x_2, ..., x_n$. Grupăm aceste date în intervale $J_1, J_2, ..., J_r$ astfel încât $J_i \cap J_j = \phi$, pentru $i \neq j$ şi $J_1 \cup J_2 \cup \cdots \cup J_r = \mathbb{R}$ să ia valoarea în intervalul J_i. Este clar că nu cunoastem aceste p_i-uri. Stim doar că $\sum p_i = 1$. Fie p_i' probabilitatea ca v.a. X să ia valoarea în intervalul J_i dacă am presupune că X are legea Q. Aceste numere se pot calcula folosind tabelele distributiei cunoscute Q. Notăm cu Y_i numărul acelor x_j-uri care se află în intervalul J_i. Y_i este de fapt frecventa absolută empirică de selectie pe intervalul J_i. Frecventa absolută teoretică (potrivit legii Q) pe intervalul J_i este egală cu np_i'. Ipoteza pe care o facem este următoarea: H: $p_i = p_i'$, $(\forall)$ $i = \overline{1,r}$.

Construim acum statistica Helmert-Pearson: $\mathrm{T}=\sum\limits_{i=1}^{r} \frac{\left(\mathrm{Y}_i - np_i'\right)^2}{np_i'}$. Ea măsoară "deviatia" legii reale a v.a. X de la legea presupusă Q. Pentru n mare T tinde să aibă distributia χ^2 (Pearson) cu $r-1$ grade de libertate dacă cumva X ar avea într-adevăr legea de distributie Q.

• Calculăm numărul T în cazul nostru şi îl notăm cu χ^*.

• Fie acum $\alpha \in (0,1)$ (de odicei α se ia mic α =0,05; 0,01; 0,001), un număr real dat pe care îl vom numi *prag de semnificatie*.

• Fie χ^2_α cuantila de ordin α pentru v.a. χ^2, adică aceea valoare pentru care P($\chi^2 < \chi^2_\alpha$)=α (vezi figura)



• Testul χ^2 functionează astfel:

— *Dacă valoarea calculată din selectie* $\chi^* > \chi^2_\alpha$ *vom accepta ipoteza* H *cu pragul de semnificatie* α, adică este foarte probabil ca ipoteza să fie adevărată.

— *Dacă valoarea calculată* $\chi^* \leq \chi^2_\alpha$ *vom respinge ipoteza* H *cu pragul de semnificatie* α.

Aceasta este testul χ^2 *simplu (Cazul I)*

• Dacă însă legea Q are s parametri de estimat tot din selectie, atunci va trebui să micsorăm numărul gradelor de libertate cu s, adică să avem $r-s-1$ grade de libertate, deoarece apar $s+1$ legături: $\sum p_i = 1$ şi cele s legături de estimare.

12.3 Exercitii rezolvate

1. Se arunca 4 monede simultan de 160 de ori. Se observa de fiecare data de cate ori a aparut "capul".

x =de cate ori poate sa apara capul, cand aruncam o data cele 4 monede	0	1	2	3	4
f =de cate ori s-a realizat x in cele 160 de aruncari x in cele 160 de aruncari	5	35	67	41	12

Sa se testeze ipoteza H_0 : monedele nu sunt falsificate, cu pragul de 5%.

Solutie Fie X v.a. care ia valorile x = de cate ori apare "capul" cand aruncam o data cele 4 monede (sau de 4 ori aceeasi moneda). Avem ca:

$P(X=x)=C_4^x\cdot\left(\frac{1}{2}\right)^x\cdot\left(\frac{1}{2}\right)^{4-x}=\left(\frac{1}{2}\right)^4\cdot C_4^x.$

Prin urmare frecventa absoluta sperata (daca presupunem H_0 adevarata) de aruncari pentru x este $160\cdot P(X=x)$. Obtinem deci urmatorul tabel:

x	0	1	2	3	4
$P(X=x)$	1/16	4/16	6/16	4/16	1/16
E_i =fr. abs. sperata cand $x=i$ sperata cand $x=i$	10	40	60	40	10
O_i =fr. abs. observata cand $x=i$	5	35	67	41	12
$\frac{(O_i-E_i)^2}{E_i}$	2,5	0,625	0,817	0,025	0,4

Folosim testul χ^2 cu ν =5 clase-1 restrictie ($\sum O_i=160$) =4 grade de libertate.

Avem deci $\chi^2_{calc}=\sum\limits_{i=1}^{5}\frac{(O_i-E_i)^2}{E_i}=4,365<\chi^2_{0,05}(4)=9,49$. Prin urmare acceptam H_0 cu pragul de 95%.

2. La o fabrica de caramizi se aleg 500 de pachete de cate 5 caramizi la intervale regulate de-a lungul unei saptamani si se numara de fiecare data cate caramizi defecte sunt in fiecare pachet. S-au obtinut urmatoarele rezultate:

x =nr.de caramizi defecte intr-un singur pachet	0	1	2	3	4	5
f =nr. de pachete care au avut x caramizi defecte	170	180	120	20	8	2

Testati cu ajutorul testului χ^2 cu nivelul de semnificatie 5% daca numarul caramizilor defecte urmeaza legea binomiala.

Solutie Deoarece in legea binomiala Bin$(n=5;p)$ nu stim pe p, trebuie sa-l estimam din esantion:

$np=\overline{x}=\frac{\sum f\cdot x}{\sum f}=1,044$, deci $p=0,2088$.

Am presupus ca H_0 : "legea este binomiala" este adevarata. Atunci v.a. X = nr. caramizilor defecte intr-un esantion de 5 caramizi va fi binomiala cu $n=5$ si $p=0,2088\equiv$ probabilitatea ca o caramida luata la intamplare sa fie defecta. Avem ca frecventa absoluta sperata de "i" caramizi defecte este $E_i=500\cdot P(X=i)=500\cdot C_5^i\,(0,2088)^i\,(0,7912)^{5-i}$, unde 0,7912=1-0,2088, iar $i=0,1,2,3,4,5$. Cum $E_4=4$ si $E_5=0$, clasele E_3, E_4, si E_5 le cumulam intr-o singura clasa renotata cu $E_3\,(x\geq 3)$. La fel, cumulam pe $O_3=20$, $O_4=8$ si $O_5=2$ intr-o noua clasa $O_3\,(x\geq 3)$. Obtinem deci urmatorul tabel:

x	0	1	2	$x\geq 3$
E_i=fr. abs. sperata cand $x=i$, daca acceptam legea binomiala	155	205	108	32
O_i=fr. abs. observata cand $x=i$, data in datele problemei	170	180	120	30
$\frac{(O_i-E_i)^2}{E_i}$	1,452	3,049	1,333	0,125

Folosim testul χ^2 cu $\nu = 4$ clase-2 restrictii($\sum O_i = 500$ si $\overline{x} = 1,044$ este impus prin estimare)= 2 grade de libertate.

Avem deci $\chi^2_{calc} = 5,959 = \sum\limits_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} < \chi^2_{0,05}(2) = 5,99$. Acceptam deci ipoteza H_0 (repartitia este binomiala) cu pragul de semnificatie 5%. Avem o oarecare neincredere deoarece 5,959 este prea aproape de valoarea critica 5,99. Se indica sa se repete experienta pentru mai multa siguranta.

3. In 100 de meciuri o echipa de fotbal a inscris goluri dupa cum urmeaza:

x =nr. de goluri inscrise intr-un meci	0	1	2	3	4	5	6	7
f =nr. de meciuri in care echipa a inscris x goluri	14	18	29	18	10	7	3	1

Testati cu χ^2 cu 5% daca golurile inscrise se repartizeaza dupa o lege Poisson.

Solutie Deoarece nu cunoastem pe λ in legea lui Poisson $Po(\lambda)$, il vom estima din media de selectie:

$\overline{x} = \frac{\sum f \cdot x}{\sum f} = \frac{230}{100} = 2,3$, deci $\lambda \approx \overline{x} = 2,3$.

$P(X = x) = \frac{e^{-2,3} \cdot (2,3)^x}{x!}$, $x = 0, 1, 2, ...$

Calculam $E_i = 100 \cdot P(X = i) \equiv$ frecventa absoluta sperata (teoretica) a "i" goluri inscrise de echipa in cele 100 de meciuri. Aici $P(X = i)$ este probabilitatea ca echipa sa inscrie "i" goluri intr-un singur meci. Folosim testul χ^2 cu $\nu = 6$ clase - 2 restrictii ($\sum O_i = 100$ si $\overline{x} = 2,3$ este impus)=4 grade de libertate. Teoretic avem 9 clase: E_0, E_1,..., E_7, E_8 $(i \geq 8)$. Cum $E_5 = 5,4$; $E_6 = 2,1$; $E_7 = 0,7$, $E_8 = 0,2$, cumulam clasele E_5, E_6, E_7 si E_8 intr-una singura, desemnata tot cu $E_5 = 8,4$. Vom obtine tabelul:

x	0	1	2	3	4	5
E_i	10,0	23,1	26,5	20,3	11,7	8,4
O_i	14	18	29	18	10	11
$\frac{(O_i - E_i)^2}{E_i}$	1,6	1,126	0,236	0,261	0,247	0,805

Folosim deci testul χ^2 cu 4 grade de libertate. Cum $\chi^2_{calc} = \sum\limits_{i=0}^{5} \frac{(O_i - E_i)^2}{E_i} = 4,275 < \chi^2_{0,05}(4) = 9,49$ vom accepta ipoteza H_0 : "distributia golurilor urmeaza o lege Poisson" cu nivelul de incredere de 95%.

12.4 Exerciţii

1. Se testează greutatea unui grup de 50 de bărbati şi se obtin valorile (în Kg):

66	78	82	75	94	77	69	74	68	60
96	78	89	61	75	95	60	79	83	71
79	62	67	97	78	85	76	65	71	75
86	84	75	81	68	63	62	75	76	77
73	65	88	87	60	62	71	78	85	72

Folositi testul χ^2 *pentru a decide cu pragul de* α*=0,90 dacă populatia este normală cu media 75 şi dispersia 625. Faceti acelasi lucru dar cu media şi dispersia estimate din selectie. Împărtiti esantionul în zece intervale.*

2. Se urmăresc accidentele mortale pe o portiune dintr-un drum national timp de 100 săptămâni. Timp de 45 de săptămâni nu a fost nici un accident mortal. În 29 de săptămâni s-a produs un accident, în 17 săptămâni 2 accidente, iar în 9 săptămâni au avut loc 3 accidente. Folositi testele χ^2 *şi K-S pentru a vedea, cu pragul* α*=0,90, dacă distributia accidentelor urmează modelul Poisson sau nu. Parametrul se estimează din selectie.*

3. Se consideră o prismă care are ca baze două triunghiuri echilaterale B_1 *şi* B_2 *şi ca fete laterale* A_1, A_2 *şi* A_3. *Se aruncă prisma de 500 de ori şi se constată că prisma a căzut de :*
$Y_1 = 111$ *ori pe fata* A_1
$Y_2 = 113$ *ori pe fata* A_2
$Y_3 = 118$ *ori pe fata* A_3
$Z_1 = \ 81$ *ori pe fata* B_1
$Z_2 = \ 77$ *ori pe fata* B_2. *Testati ipoteza după care cele 3 fete laterale şi cele 2 baze au aceeasi probabilitate* $\frac{1}{5}$ *(pragul* $\varepsilon = 0,05$ *şi* $\varepsilon = 0,01$*).*

Indicatie. Avem Cazul I: $r = 5$ *şi* $p'_1 = p'_2 = \cdots = p'_5 = \frac{1}{5}$. *Găsim* χ^2*=15,04. Analizând tabelul vedem că trebuie să respingem ipoteza pentru cele două praguri.*

4. Vrem să testăm ipoteza H după care o anumită v.a. X este gaussiană cu media 1,1 şi dispersia 0,2. S-au făcut 1000 de probe independente care au condus la rezultatele următoare:

0,6	0,7	0,8	0,9	1	1,1	1,2	1,3	1,4	1,5
26	51	107	168	200	193	138	80	29	8

adică X a luat de 26 de ori o valoare mai mică decât 0,6 şi de 51 de ori o valoare cuprinsă în intervalul $[0,6; 0,7)$, *etc. Testati ipoteza H cu pragurile* $\varepsilon = 0,05$, $\varepsilon = 0,01$ *şi* $\varepsilon = 0,001$.

Indicatie Se partitionează $\mathbb{R}$ *după cum indică problema. Se găseste* $\chi^2 = 13,97$. *Folosim tabelul distributiei* χ^2 *cu* $10 - 1 = 9$ *grade de libertate. Ipoteza trebuie respinsă cu pragul 0,05 şi 0,01, dar trebuie să o acceptăm cu pragul 0,001.*

5. Se testează ipoteza după care culoarea ochilor este independentă de culoarea părului. Pentru aceasta se introduc 2 caractere X şi Y:

X ia valorile 1, 2, 3, 4 după cum ochii sunt albastri, gri sau bruni.

Y ia valorile 1, 2, 3, 4 după cum părul este blond, brun, negru sau roscat. Se testează un număr de persoane şi rezultatele le găsim în tabelul următor:

X \ Y	1	2	3	4
1	1768	807	189	47
2	946	1387	746	53
3	1125	438	288	16

.

Indicatie Aici $(a-1)(b-1) = 2 \times 3 = 6$ *şi* $\chi^2 = 1075$. *Chiar cu pragul de* $\varepsilon = 0,001$ *ipoteza nu este acceptabilă.*

6. Un articol de calitătile A, B sau C poate fi fabricat după 2 metode numerotate cu 1 şi 2. Se examinează 100 de articole şi se găseste tabelul:

	A	B	C
1	20	19	11
2	12	31	7

. *Putem accepta oare ipoteza că, calitatea articolului nu depinde de modul său de fabricare (se ia* $\varepsilon = 0,1$ *şi* $\varepsilon = 0,01$*)?*

Indicatie. Este vorba de un test de omogenitate cu $r = 3$ *şi* $t = 2$. *Avem că* $\chi^2 = 5,77$ *pentru o variabilă aleatoare* χ^2 *cu 2 grade de libertate. Cu pragul de 0,1 trebuie să respingem ipoteza, dar cu pragul de 0,01 trebuie să o acceptăm.*

Lecţia 13

Alte teste neparametrice

13.1 Testul de concordantă Kolmogorov-Smirnov

Acest test este din multe puncte de vedere "mai bun" decât testul χ^2. El se aplică bine şi pentru esantioanele mici ($n \leq 25$). Dacă pentru testul χ^2 trebuia să grupăm datele, pentru testul K-S nu este nevie decât să calculăm functia de repartitie empirică asociată selectiei efectuate. El poate fi utilizat bine şi pentru a compara două distributii.

Presupunem că populatia P are ca functie de repartitie teoretică functia $\mathrm{F}_T(x)$. Efectuăm un sondaj de volum n: $\{x_1, ..., x_n\}$ şi notăm cu $\mathrm{F}_S(x)$ functia de repartitie empirică asociată acestui sondaj (vezi Lectia 8). Pentru a măsura deviatia functiei $\mathrm{F}_T(x)$ de la functia $\mathrm{F}_S(x)$ se introduce statistica lui Kolmogorov: $\mathrm{D}=\max\limits_x |\mathrm{F}_S(x)\text{-}\mathrm{F}_T(x)|$. Dacă populatia P are într-adevăr repartitia $\mathrm{F}_T(x)$ atunci se cunoaste distributia v.a. D (vezi Tabelul XIX). Pe prima coloană în acest tabel avem volumul de selectie. Pe prima linie orizontală avem 5 valori ale pragului de semnificatie: α =0,80; 0,85; 0,90; 0,95 şi 0,99.

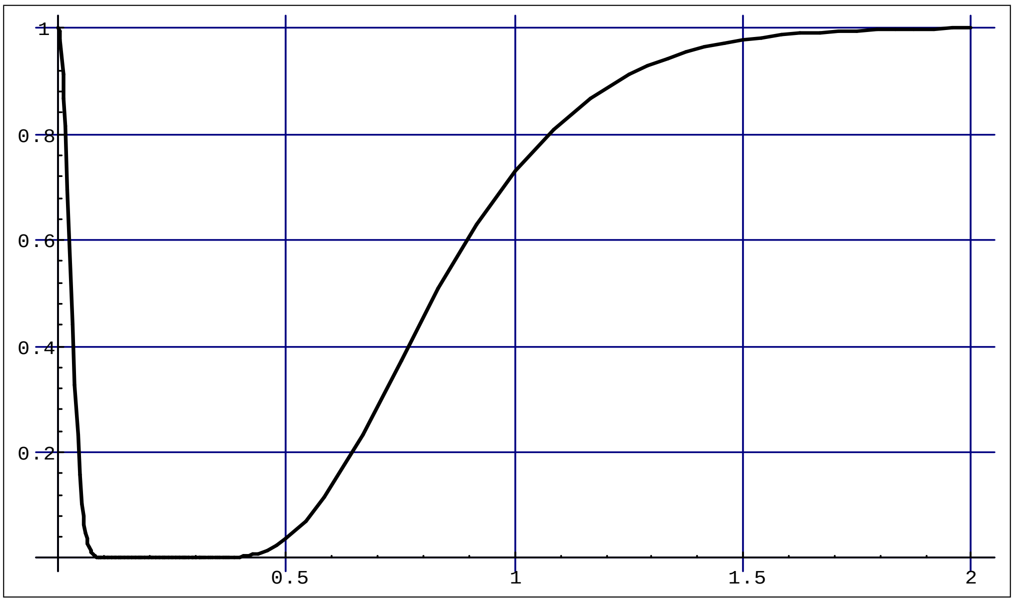
Tabelul XIX este construit pe baza teoremei lui A. N. Kolmogorov:

Teorema 13.1 *Fie X o v.a., $F_T(x)$ functia ei de repartitie considerată continuă şi $F_n(x)$ o functie empirică de repartitie asociată unei selectii de volum n:$\{x_1, ..., x_n\}$, dintr-o populatie cu v.a. X. Atunci avem relatia*

$$\lim_{n\to\infty} P\left(D \leq \frac{\lambda}{\sqrt{n}}\right) = K(\lambda) = \sum_{-\infty}^{\infty} (-1)^k \cdot \exp\left(-2k^2\lambda^2\right) \tag{13.1}$$

pentru orice $\lambda > 0$.

Graficul funcţiei $K(\lambda)$ apare în figura următoare:



Graficul funcţiei $K(\lambda)$

Exemplul 13.2 *Se face sondajul {-2, -1, -1, 0, 0, 1, 2, 2} dintr-o populatie P. Să se testeze cu testul K-S dacă populatia este normală cu media 0 şi dispersia 2 (date şi nu estimate!) cu pragul de semnificatie* α *=0,80.*

Solutie Trebuie să construim functia de repartitie empirică, $F_S(x)$— notată de noi cu $F_n^*(x)$ în Lectia 8.

$$F_S(x) = \begin{cases} 0,\ x \leq -2 \\ \frac{1}{8},\ x \in (-2,-1] \\ \frac{3}{8},\ x \in (-1,0] \\ \frac{5}{8},\ x \in (0,1] \\ \frac{6}{8},\ x \in (1,2] \\ 1,\ x > 2 \end{cases}$$

Deoarece P(X< x)=$F_T(x)$ şi cum v.a. $\frac{X}{\sigma} = \frac{X}{\sqrt{2}}$ este normală redusă (de tipul N(0,1)) avem că P(X< x)=P$\left(\frac{X}{\sigma} < \frac{x}{\sigma}\right)$=$\Phi\left(\frac{x}{\sigma}\right)$. Deci $F_T(x) = \Phi\left(\frac{x}{\sigma}\right)$, unde Φ este functia lui Laplace tabelată în Tabelul I. Prin urmare va trebui să calculăm $\Phi\left(\frac{x}{\sqrt{2}}\right)$ în punctele x =-2, -1, 0, 1, 2. Cum $\Phi(-y) = 1 - \Phi(y)$, rămâne numai (deoarece $\Phi(0) = 0,5$) să calculăm din Tabelul I $\Phi\left(\frac{2}{\sqrt{2}}\right) = \Phi(\sqrt{2}) = \Phi(1,41) = 0,92$ şi $\Phi\left(\frac{1}{\sqrt{2}}\right) = \Phi\left(\frac{\sqrt{2}}{2}\right) = \Phi(0,70) = 0,75$. Prin urmare $\Phi\left(-\frac{2}{\sqrt{2}}\right) = 1 - 0,92 = 0,08$ şi $\Phi\left(-\frac{1}{\sqrt{2}}\right) = 1 - 0,75 = 0,25$. Calculăm acum diferentele dintre $F_S(x)$ şi $F_T(x)$ pe fiecare interval ce apare în definitia functiei $F_S(x)$. Găsim

x	$-\infty$	-2	-1	0	1	2	∞
$F_S(x)$	0	0	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{5}{8}$	$\frac{6}{8}$	1
$F_T(x)$	0	$0,08$	$0,25$	$0,5$	$0,75$	$0,92$	1
$F_S(x) - F_T(x)$	0	$-0,08$	$-0,125$	$-0,125$	$-0,125$	$-0,170$	0

.

Prin urmare D=max $|F_S(x)$-$F_T(x)| = 0,170$. Cuantila pentru $\alpha = 0,80$ şi $n = 8$ (din Tabelul XIX) este $D_\alpha = 0,358$. Cum P(D$< 0,358$)=0,80 rezultă că putem accepta ipoteza că populatia P este normală cu pragul de semnificatie 80% (deoarece 0,170<0,358).

Observaţia 13.3 *Acest test se poate aplica când avem de comparat două distributii:*

$$D = \max|F_{S_1}(x)\text{-}F_{S_2}(x)|$$

etc.,unde cele două functii empirice de repartitie care apar corespund celor două distributii.

Exerciţiul 13.4 *Folositi testul K-S pentru exemplul 12.1*

13.2 Testul lungimilor (secventelor)

Fie X şi Y două variabile aleatoare. Vrem să testăm următoarea ipoteză:

H: X şi Y au aceeasi lege de distributie.

Pentru aceasta considerăm un esantion de volum m al v.a. X:$(X_1, ..., X_m)$ şi un esantion de volum n al v.a. Y:$(Y_1, ..., Y_2)$. Considerăm acum sirul de $m + n$ variabile aleatoare $(X_1, ..., X_m, Y_1, ..., Y_n)$. Presupunem că ele sunt independente. Dacă X şi Y ar avea aceeasi lege de distributie am putea "amesteca" oricum aceste variabile şi rationamentele noastre nu s-ar schimba la trecerea de la $(X_1, ..., X_n)$ la $(Y_1, ..., Y_m)$.

Introducem următoarea v.a. R. Facem câte o selectie de volum m din X: $x_1, ..., x_m$ şi de volum n din Y: $y_1, ..., y_n$ considerăm sirul de numere $x_1, ..., x_m, y_1, ..., y_n$. Asezăm acum aceste numere în ordinea crescătoare şi nu ne interesează decât faptul că ele sunt din X sau din Y. Vom nota o astfel de situatie sub forma: ω=XXYXYYYXXXXYY. Aici m= numărul X-lor, adică m=7, n= numărul Y-lor, adică n=6. În sirul ω cel mai mic număr din sirul $x_1, ..., x_m, y_1, ..., y_n$, este din $(x_1, ..., x_m)$,următorul în ordine crescătoare este tot din acest sir,al treilea în ordine crescătoare este din $(y_1, ..., y_n)$,etc. Cel mai mare număr din ω este din $(y_1, ..., y_n)$. Dacă cumva $x_{i_1} = x_{i_2}$ le asezăm unul după altul în orice ordine, etc.

Vom numi *secventă (lungime)* în sirul ω orice subsir format numai din X sau numai din Y. De exemplu în ω avem următoarele secvente:

$$\underbrace{\text{XX}}_{\text{S1}}\quad \underbrace{\text{Y}}_{\text{S2}}\quad \underbrace{\text{X}}_{\text{S3}}\quad \underbrace{\text{YYY}}_{\text{S4}}\quad \underbrace{\text{XXXX}}_{\text{S5}}\quad \underbrace{\text{YY}}_{\text{S6}}$$

Prima secventa este S1: XX
A doua secventa este S2: Y
A treia secventa este S3: X
$\vdots$ $\vdots$
A sasea secventă este S6: YY.

Pentru selectia noastră v.a. R ia valoarea 6, adică numărul secventelor de X-şi sau de Y-ci care apar după rearanjarea în ordine crescătoare a esantionului $(x_1, ..., x_m, y_1, ..., y_n)$.

Dacă ipoteza H este adevărată, functia de repartitie a v.a. R, F(x)=P(R $\leq x$) se calculează tinând seama de următoarele observatii de natură combinatorică:

Dacă de exemplu $m \leq n$,

$$P(R = 2s) = \frac{2}{C_{m+n}^{m}} \cdot C_{m-1}^{s-1} \cdot C_{m-1}^{s-1}, \ pentru \quad 1 \leq s \leq m;$$

$$P(R = 2s + 1) = \frac{1}{C_{m+n}^{m}} \cdot [C_{m-1}^{s} \cdot C_{m-1}^{s-1} + C_{n-1}^{s} \cdot C_{n-1}^{s-1}], \ pentru \quad 1 \leq s \leq m;$$

$$P(R = 2m + 1) = \frac{1}{C_{m+n}^{m}} \cdot C_{n-1}^{m}, \ dacă \quad m < n;$$

$$P(R = r) = 0$$

în toate celelalte cazuri rămase.

Se demonstrează că statistica v.a. R este asimptotică $(m, n$ mari) gaussiană cu

$$M(R) = 1 + \frac{2mn}{m+n}, D(R) = \frac{2mn\,(2mn - m - n)}{(m+n)^2\,(m+n-1)} \tag{13.2}$$

Să notăm cu

$$T = \frac{\text{R-M(R)}}{\sqrt{\text{D(R)}}} \tag{13.3}$$

Variabila T tinde către o v.a. gaussiană redusă.

Un test de preg ϵ se construieste astfel:

— Fie r cel mai mare întreg x astfel încât F(x)$\leq \epsilon$.

— Testul va fi: Respinge H$\Longleftrightarrow$R$^* \leq r$, unde R* este valoarea efectivă a v.a. R obtinută din sondaje (= numărul secventelor). Pragul este ϵ şi acest test este cel mai puternic pentru acest prag.

Desigur că am putea construi şi alte teste plecând de la formula (13.3) în analogie cu testele de semnificatie sau ca testul χ^2. Pentru m şi n mici se calculează direct F(x) după formulele (13.1), iar pentru m şi n mari se foloseste (13.3).

Exemplul 13.5 *Se cântăresc 10 mere de două calităti A şi B şi se găsesc următoarele rezultate (în grame):*

A:	*192*	*197*	*207*	*182*	*191*
B:	*212*	*201*	*209*	*214*	*203*

Ne permit oare aceste rezultate să respingem ipoteza H: greutatea unui măr urmează aceeasi lege de distributie indiferent de calitatea lui A sau B (cu pragul ϵ=0,05)? Calculati probabilitea erorii de primul tip pentru acest test. Utilizati şi aproximarea gaussiană şi comparati rezultatele.

Solutie Aici avem $m = n = 5$; $C_{m+n}^m = 252$. Să notăm
N(x)= C_{m+n}^m·P(R=x)=225·P(R=x). Avem N(2s)=$2\left(C_4^{s-1}\right)^2$, pentru 1≤s≤5,
N(2s+1)=$2 \cdot C_4^{s-1} \cdot C_4^{s-1}$, pentru 1≤s≤4 şi N($x$)=0 în celelalte cazuri. Deducem de aici legea de distributie pentru v.a. R:

x	2	3	4	5	6	7	8	9	10
$252 \cdot$ P(R=x)	2	8	32	48	72	48	32	8	2

.

Căutăm cel mai mare x cu F(x)≤0,05, sau 252 F(x)≤ 12,60. Cum avem 252 F(x)=$\sum_{i \leq x}$N(i), găsim că 252 F(2)=2; 252 F(3)=10; 252 F(4)=42.

Cel mai mare x este deci 3 şi testul devine: Respinge H⟺R* <3. Probabilitatea erorii de primul tip pentru acest test este P(R≤3)=F(3)=$\frac{10}{252}$=0,0397. Să utilizăm acum aproximarea gaussiană estimând M(R) şi D(R) direct din selectie.

$$M(R) = 1 + \frac{2 \cdot 5 \cdot 5}{5+5} = 6.$$

$$D(R) = \frac{2 \cdot 5 \cdot 5 \cdot 40}{10 \cdot 9} = 2,222.$$

Avem că

$$P(R \leq 3,5) = P\left(\frac{\mathrm{R}-6}{\sqrt{2,222}} \leq \frac{-2,5}{\sqrt{2,222}}\right) = 0,047$$

Să observăm că rezulatele sunt comparabile chiar dacă m şi n şi sunt mici.
În exemplul nostru ω=XXXXYYXYYY, deoarece selectia ordonată este:

182 191 192 197 201 203 207 209 212 214

.

Avem deci 4 secvente. Cum R* = 4 > 3 va trebui să acceptăm ipoteza H cu pragul ϵ=0,05.

13.3 Testul lui Wilcoxon I (cazul observatiilor necuplate)

Definiţia 13.6 *Fie X şi Y două v.a. Variabila aleatoare Y se zice stochastic superioară v.a. X dacă* $(\forall)$ $z \in \mathbf{R}$ *avem că* $P(Y\leq z)\leq P(X\leq z)$*, cu inegalitate strictă cel putin pentru un* $z = z_1$*. Presupunem că avem alternativele:*

— sau X şi Y au aceeasi lege
— sau Y este stochastic superioară lui X.

Fie ipoteza H: X şi Y au aceeasi lege. Ca şi la testul secventelor construim sirul ω=XXYX..., de exemplu ω=XYXXYYXYY. Statistica lui Wilcoxon T este v.a. care are ca valoare T= suma numerelor care arată locurile pe care le ocupă X în sirul ω. Aici avem T=1+3+4+7=15. Testul este de forma: Respinge H$\Longleftrightarrow$T$\leq t$, unde t este valoarea lui T din selectie. Dacă H este verificată, P(T$\leq t$)=$\frac{1}{C^m_{m+n}}$înmultit cu numărul probelor favorabile relatiei T$\leq t$ (pentru m şi n mici se poate calcula direct F(t)). Aici lucrăm cu selectiile $X_1, ..., X_m, Y_1, ..., Y_n$. Dacă m şi n sunt mari, T este aproape gaussiană cu:

$$M(T) = \frac{m\,(m+n+1)}{2} \quad D(T) = \frac{mn\,(m+n+1)}{12} \tag{13.4}$$

Exemplul 13.7 *Se încearcă un tratament mediacl nou asupra unui grup de persoane de aceesi vârstă, bolnave grav cu o maladie de tip cardiac. S-a notat timpul (în ani) după care aceste persoane tratate mai trăiau încă, cât şi timpul după care alte persoane de aceeasi vârstă, bolnave de aceeasi boală, dar netratate, mai trăiau. S-au obtinut următoarele rezultate:*

Tratate:	1,2	6,3	6,5	7,8	11,2	15,6
Netratate:	0,4	3,5	4,8	6,7		

Testati ipoteza H potrivit căreia tratamentul nu prelungeste viata unui bolnav (prag 0,05).

Solutie Aplicăm testul lui Wilcoxon. X este v.a. care exprimă durata de viată a unui bolnav netratat şi Y a unui bolnav tratat. Este clar că v.a. Y este stochastic superioară v.a. X. Ordonăm crescător sirul $\{x_1, ..., x_m, y_1, ..., y_n\}$ şi găsim ω=XYXXYYXYYY. Aici T=1+3+4+7=15. Avem m=4 şi n=6. Să numărăm cazurile favorabile conditiei T$\leq$15. Un caz îl avem mai sus: (1, 3, 4, 7), adică am asezat într-un sir pozitiile lui X în ω. Cazurile favorabile vor fi: (1, 2, 3, 4), (1, 2, 3, 5), (1, 2, 3, 6), (1, 2, 3, 7), (1, 2, 3, 8), (1, 2, 3, 9), (1, 2, 4, 5), (1, 2, 4, 6), (1, 2, 4, 7), (1, 2, 4, 8), (1, 2, 5, 6), (1, 2, 5, 7), (1, 3, 4, 5), (1, 3, 4, 6), (1, 3, 4, 7) şi (1, 3, 5, 6). Sunt în total 16 posibilităti favorabile. Numărul tuturor posibilitătilor este $C^4_{10} = 210$ (= numărul modurilor în care putem aseza patru litere X într-un sir de10 litere X şi Y). Dacă H ar fi adevărată ar trebui să avem P(T$\leq$15)=$\frac{16}{210}$=0,076. Prin urmare, cum 0,076>0,05 trebuie să acceptăm ipoteza H cu pragul 0,05 şi trebuie să o respingem cu pragul 0,1.

13.4 Testul semnelor

Presupunem că avem două două v.a. X şi Y definite pe aceeasi categorie de probe, astfel încât P(X=Y)=0. Vrem să testăm ipoteza: H: P(Y>X)≥P(X>Y) contra ipotezei K: P(Y>X)<P(X>Y).

Pentru aceasta se fac n observatii independente $(X_1,Y_1),\ldots,(X_n,Y_n)$ ale v.a. (X,Y). Se notează cu V numărul acelor cupluri (X_i,Y_i) pentru care $Y_i < X_i$. Testul este de forma:

Respinge H$\Longleftrightarrow$V$\leq v$, unde v este valoarea lui V din sondajul respectiv.

Dacă H este adevărată cea mai mare valoare posibilă pentru P(V$\leq v$) se obtine dacă presupunem că V satisface legea lui Bernoulli $B\left(n, \frac{1}{2}\right)$.

Exemplul 13.8 *O firmă vrea să testeze un nou ingredient adăogat unei creme antisolare. Se fac testări pe 7 voluntari şi pe spatele fiecăruia se aplică cremă antisolară astfel: pe jumătatea superioară se aplică crema veche, iar pe jumătatea inferioară se aplică crema cu ingredientul respectiv. Se expun la soare cei 7 voluntari şi se observă măsura în care pielea lor se înnegreste. Se obtine tabelul următor:*

Voluntarul nr.	1	2	3	4	5	6	7
crema veche	42	51	31	61	44	55	48
crema nouă	38	53	36	52	33	49	36

.

Solutie Notăm cu Y v.a. corespunzătoare înnegririi pielii cu vechea cremă şi cu X v.a. pentru noua cremă, deoarece se presupune că prin adăogarea ingredientului valorile lui X vor fi în general mai mici decât cele ale lui Y. Avem n=7. Aplicăm testul semnelor. Numărul cuplurilor (X,Y) pentru care Y<X este V=2. Dacă H este adevărată V are o distributie $B\left(7, \frac{1}{2}\right)$ şi deci

$$P(V \leq 2) = \left(\frac{1}{2}\right)^7 \left[C_7^0 + C_7^1 + C_7^2\right] = \frac{29}{128} = 0,23 > 0,1$$

Este deci foarte probabil să găsim pentru V valori mai mari decât 2. Prin urmare, cu acest test al semnelor (V reprezintă câte semne "+" avem în diferenta X-Y) ipoteza H trebuie acceptată (cu pragul 0,1).

13.5 Testul lui Wilcoxon II (cazul observatiilor cuplate)

Fie X, Y două v.a. definite pe aceeasi ctegorie de probe. Fie Z=Y-X. Vrem să testăm ipoteza H: Z este o variabilă aleatoare simetrică, adică are aceeasi lege ca şi v.a. -Z, contra ipotezei K: Z este stochastic superioară lui -Z. Să observăm că cele 2 ipoteze *nu* sunt contrare şi deci *nu* acoperă gama de posibilităti.

Pentru aceasta facem n observatii independente (X_1,Y_1),...,(X_n,Y_n) ale v.a. (X,Y). Aranjăm apoi în ordinea crescătoare a modulelor lor diferentele Z=Y−X şi nu retinem decât semnele lor. Găsim de exemplu ω=(− − − + − + + + +). Notăm cu W suma numerelor ce exprimă pozitia semnelor minus. În cazul nostru W=1+2+3+5=11. Testul are forma următoare: Respinge H⟺W≤ w, unde w este valoarea v.a. W obtinută din sondaj.

Dacă H este adevărată, multimea Ω a celor 2^n posibilităti pentru semnul ”−” este înzestrată cu legea uniformă (este unica lege probabilistică pe o multime finită de evenimente, care face ca aceste evenimente să fie egal probabile). Dacă n nu este prea mare legea W se obtine prin numărarea directă a cazurilor favorabile. Dacă n este mare W este aproximativ gaussiană cu:

$$M(W) = \frac{n\,(n+1)}{4} iD(W) = \frac{n\,(n+1)\,(2n+1)}{24} \tag{13.5}$$

Exemplul 13.9 *Reluăm exemplul 13.6 şi vrem să-i aplicăm testul Wilcoxon II (ϵ=0,1).*

• Ordonăm valorile v.a. Y−X în ordinea crescătoare a valorilor lor absolute: −2, 4, −5, 6, 9, 11, 12. Găsim sirul de semne ω=(− + − + + + +). Suma indicilor cu semnul minus este W=1+3=4. Numărul cazurilor posibile este 2^7=128. Aici w=4. Cazurile favorabile (W≤4) sunt: (+ + + + + + +), (− + + + + + +), (+ − + + + + +), (+ + − + + + +), (+ + + − + + +), (− − + + + + +), (− + − + + + +), adică 7. De aici avem că P(W≤4)=$\frac{7}{128}$=0,054.

Cu pragul 0,1 va trebui să respingem ipoteza H.

Observaţia 13.10 *Dacă comparăm rezultatul obisnuit cu acela din Exemplul 13.6 aparent găsim o contradictie. Aceasta se explică deoarece în testul semnelor nu tinem seama decât de numărul semnelor minus şi nu de pozitia lor în sirul ω. În testul lui Wilcoxon se tine seama că aceste semne sunt plasate la început, şi nu oriunde în sirul ω. Acest lucru face ca valoarea V să fie în general mult mai mare decât valoarea W. De aici apare clar că rezultatul testului Wilcoxon II sr trebui să fie ”mai demn” de crezut de firmă decât rezultatul testului semnelor. Comparatia dintre cele două teste se face de obicei de la caz la caz şi se interpretează rezultatul potrivit situatiei particulare studiate. Evident că aici, pentru firmă, este convenabil testul Wilcoxon II şi nu testul semnelor, care nu pare concludent. În plus, în testul Wilcoxon II se presupune ”ceva” în plus de la început (K nu este aplternativa ipotezei H).*

13.6 Exerciţii

1. Se dau două esantioane independente de volum 20 din v.a. X şi Y:

X:	147	193	238	225	252	143	178	209
	259	263	226	179	253	262	181	169
	210	233	248	194				
Y:	240	254	192	157	168	170	207	222
	201	215	217	243	172	183	197	241
	182	163	173	167				

Cu fiecare din pragurile ϵ=0,05; ϵ=0,1, testati ipoteza H: X şi Y au aceeasi lege,
1) cu testul lungimilor;
2) cu testul Wilcoxon I.
Explicati eventualele contradictii.

Indicatie 1) R=15. Prin aproximarea gaussiană pentru ϵ=0,05 găsim

$$P(R \leq 15) \cong 0{,}0388$$

Se respinge deci H.

2. Suma indicilor lui X este T=462. Folosim aproximarea gaussiană pentru ϵ=0,05 şi găsim
$P(T \leq 462) > 0{,}5$, deci ipoteza H se acceptă. Aici trebuie să respingem H deoarece se vede clar că cele două legi "sunt" departe una de alta. Testul Wilcoxon I "nu a mers" deoarece nerealizarea lui H$\Longrightarrow$Y este stochastic superioară lui X, lucru evident neadevărat, din sondaj. Deci trebuie să acceptăm pe H, dar cu rezerve. Probabil că cele două legi nu sunt aceleasi dar se "întrepătrund".

3. Se testează un medicament nou pe un lot de 13 soareci şi se obtin următoarele rezultate relative la o anumită analiză ("mare" înseamnă înrăutătirea stării individului):

soareci netratati:	45	88	16	6	28	122	62	13
soareci tratati:	23	104	2	9	30			

Folositi testul lui Wilcoxon I pentru a testa ipoteza H: medicamentul nu dă rezultate. (prag ϵ=0,05 şi ϵ=0,1).

Indicatie. Se foloseste aproximarea gaussiană şi se găseste că P(T$\leq$30)$\approx$0,255. Ipoteza se acceptă.

4. Douăzeci de stupi cu albine se lasă pe aceeasi perioadă a anului în două zone diferite A (zece stupi) şi B (zece stupi) timp de 20 de ani. Se observă câte kilograme de miere se obtin de la ei în fiecare an în cele 2 zone:

Anul	A	B	Anul	A	B
1	68,3	72,5	11	32,2	31,9
2	60,1	56,0	12	63,3	58,1
3	52,2	55,8	13	54,2	52,7
4	41,7	39,2	14	47,0	46,2
5	32,0	31,4	15	91,9	90,2
6	30,9	35,5	16	56,1	55,4
7	39,3	39,2	17	79,6	75,1
8	42,0	41,1	18	81,2	86,6
9	37,7	43,3	19	78,4	75,3
10	33,5	31,7	20	46,6	43,8

Să se testeze cu testul semnelor, apoi cu testul Wilcoxon II (ϵ=0,05) ipoteza: H: cele două zone melifere A şi B sunt tot atât de productive.

Indicatie. Fie X v.a. ce măsoară greutatea mierii provenită din zona A şi Y v.a. corespunzătoare zonei B.

Testul semnelor Z=Y–X conduce la legea binomială $B\left(20, \frac{1}{2}\right)$ şi deci

$$P(V \leq 4)=0{,}00591 < 0{,}05$$

, lucru ce conduce la respingerea ipotezei H.

Testul Wilcoxon II conduce la W=71. Aproximăm W cu legea gaussiană şi găsim

$$P(W \leq 71)=P\left(\frac{\mathrm{W}-105}{\sqrt{717,5}} \leq -1{,}27\right)=0{,}102 > 0{,}05$$

, deci ipoteza H se acceptă în aces caz. Deoarece avem putine "minusuri" (doar 4) vom prefera testul Wilcoxon II. Considerăm numai o pură întâmplare că avem putine minusuri. În general, când numărul minusurilor în testul semnelor este mic, nu putem să ne bazăm pe acest test.El este "slab" în acest caz.

Lecţia 14

Analiza dispersiei şi analiza regresiei

14.1 Analiza dispersiei

Vom analiza aici cea mai simplă problemă dispersională.

Problemă Se consideră s variabile aleatoare gaussiene X_1,...,X_s de aceeasi dispersie necunoscută σ^2. Se notează cu m_i =M(X_i). Vrem să testăm următoarea ipoteză: H: $m_1 = m_2 = \cdots = m_s$, adică toate v.a. au aceeasi medie.

Presupunem că avem pentru fiecare $i = 1, ..., s$ câte un esantion de volum n_i :X_{i_1},X_{i_2}, ...,$X_{i_{n_i}}$, al v.a. X_i. Presupunem că toate cele $n = n_1+n_2+\cdots+n_s$ v.a. X_{11},...,X_{sn_s} sunt independente. Notăm cu

$$\overline{X}_i = \left(\frac{1}{n_i} X_{i_1} + \cdots + X_{i_{n_i}}\right), \text{pentru i=1,...,s} \tag{14.1}$$

$$\overline{X} = \frac{1}{n} \sum_{\substack{1\le i\le s\\ 1\le j\le n_i}} X_{ij} = \sum_{1\le i\le s} \frac{n_i}{n} \overline{X}_i \tag{14.2}$$

$$Q_A = \sum_{1\le i\le s} n_i \left(\overline{X}_i - \overline{X}\right)^2 = \left(\sum_{1\le i\le s} n_i \overline{X}_i^2\right) - n\overline{X}^2 \tag{14.3}$$

$$Q_R = \sum_{\substack{1\le i\le s\\ 1\le j\le n_i}} \left(X_{ij} - \overline{X}_i\right)^2 = \left(\sum_{\substack{1\le i\le s\\ 1\le j\le n_i}} X_{ij}^2\right) - \sum_{1\le i\le n} n_i \overline{X}_i^2. \tag{14.4}$$

• Se stie că statistica Q_R/σ^2 se spune unei legi Pearson cu $n-s$ grade de libertate şi că variabila aleatoare.

• $U_{ij} = \sqrt{\frac{(n_i+n_j)(n-s)}{n_i n_j}} \frac{\left[\overline{X}_i - \overline{X}_j - (m_i - m_j)\right]}{\sqrt{Q_R}}$ are o distributie Student cu $n-s$ grade de liberate pentru orice i,j ca mai sus.

• De asemenea, statistica $W = \frac{(n-s)Q_A}{(s-1)Q_R}$ se supune unei legi Fisher–Snedecor (distributia F) cu $\left\{\begin{array}{c} s-1 \\ n-s \end{array}\right\}$ grade de libertate, dacă ipoteza H este adevărată .

Această ultimă observatie va constitui esenta testului următor: Respinge H$\Longleftrightarrow$W$\geq w$, unde w este cel mai mare număr astfel încât P(FS$\geq w$)$\leq \epsilon$, unde FS este v.a. Fisher–Snedecor cu $\left\{\begin{array}{c} s-1 \\ n-s \end{array}\right\}$ grade de libertate, iar ϵ este un prag de semnificatie, considerat mic, de exemplu ϵ=0,05; 0,1; etc. Este posibil ca în tabele să găsim cuantilele de ordin 0,95; 0,9; etc. Se trece atunci la probabilitatea evenimentului contrar, etc.

• Dacă datele X_{ij} sunt mari se înlocuiesc acestea cu datele $aX_{ij}+b$, unde $a \neq 0$, $b \in \mathbf{R}$ sunt alese astfel încât numerele $aX_{ij}+b$ să devină mici. Prin această schimbare v.a. U_{ij} şi W *nu* se modifică, deci testul decurge exact ca mai sus pentru noile date.

Exemplul 14.1 *Pe patru soluri diferite A_1, A_2, A_3, A_4 se plantează orz. Se fac selectii de volume diferite din tulpini de orz ajunse la maturitate din cele patru soluri şi se notează lungimea acestora (în cm):*

A_1	A_2	A_3	A_4
380	350	354	376
376	358	360	344
360	356	362	342
368	376	352	372
372	338	366	374
366	342	372	360
374	366	362	
382	350	344	
	344	342	
	364	358	
		351	
		348	
		348	

Se notează cu X_i lungimea aleatoare a unei tulpini de orz de pe terenul A_i. Se presupune că X_i sunt gaussiene cu aceeasi dispersie σ^2. Fie pragul ϵ=0,05.

1) Testati ipoteza H: X_1, X_2, X_3, X_4 au aceeasi medie.

2) Testati ipoteza H: X_2, X_3 şi X_4 au aceeasi medie.

Solutie Deoarece datele sunt mari le centrăm cu ajutorul transformării Z_i=X_i–330. Obtinem un nou tabel:

A_1	A_2	A_3	A_4
50	20	24	46
46	28	30	14
30	26	32	12
38	46	22	42
42	8	36	44
36	12	42	30
44	36	32	
52	20	14	
	14	12	
	34	28	
		21	
		18	
		18	

1) Folosim o analiză dispersională pentru a testa ipoteza H. Aici avem s=4, n_1=8, n_2=10, n_3=13, n_4=6, n=8+10+13+6=37; $n_1\overline{Z}_1$=338; $n_2\overline{Z}_2$=244; $n_3\overline{Z}_3$=188; $n\overline{Z}=n_1\overline{Z}_1+\cdots+n_4\overline{Z}_4$=1099.

Prin urmare $\sum\limits_{1\leq i\leq 4} n_i\overline{Z}_i^2$=34449, $n\overline{Z}^2$=32640, deci Q_A=1809. Cum $\sum\limits_{i,j} Z_{ij}^2$=38229, avem Q_R=3780. De aici W=5,26.

Probabilitatea ca o v.a. Fisher–Snedecor cu $\left\{\begin{matrix}3\\33\end{matrix}\right\}$ grade de libertate să ia o valoare asemănătoare lui W este inferioară lui 0,01. Prin urmare ipoteza H se respinge.

2) În acest caz s'=3, $n'=n_2+n_3+n_4$=29, Q'_A=199; Q'_R=3400 şi W'=0,76. Această valoare a lui W' este prea mică, deci ipoteza se acceptă în acest caz

14.2 Analiza regresiei

Fie X şi Y două variabile aleatoare şi $X_1,...,X_n,Y_1,...,Y_n$ v.a. de selectie de acelasi volum n. Se pune problema de a studia *legătura* (dacă există) dintre cele două v.a. X şi Y numai din analiza unor cupluri de selectie de tipul $\{x_1,...,x_n\},\{y_1,...,y_n\}$. Este posibil ca cele două v.a. să *nu* fie "corelate", adică *coeficientul de corelatie* $\rho_{XY}=\frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{M\left((X-\overline{X})(Y-\overline{Y})\right)}{\sigma_X\sigma_Y}$ să fie zero. Acest lucru nu înseamnă că nu poate exista o relatie functională de forma F(X,Y)=0 între v.a. X şi Y. Această egalitate poate să nu fie "deterministă". Sau poate să fie astfel,dar noi să nu putem descrie matematic această functie de legătură. În Lectia 6 s-a arătat că X şi Y sunt legate între ele printr–o relatie liniară: Y=aX+b, sau X=cX+d, dacă şi numai dacă $\rho_{XY}=\pm1$. De regulă noi facem sondaje în urma cărora estimăm coeficientul de corelatie printr–o formulă empirică de forma:

$$\rho^*_{\mathrm{XY}} = \frac{\sum\limits_{i=1}^{n} (x_i - \bar{\mathrm{x}}) \cdot (y_i - \bar{y})}{n \cdot \sigma^*_x \cdot \sigma^*_y} \tag{14.5}$$

unde $\bar{\mathrm{x}}=m^*_x = \frac{\sum\limits_{i=1}^{n} x_i}{n}$ şi $\bar{\mathrm{y}}=m^*_y = \frac{\sum\limits_{i=1}^{n} y_i}{n}$, $\sigma^*_x = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}$, $\sigma^*_y = \sqrt{\frac{\sum(y_i-\bar{y})^2}{n}}$.

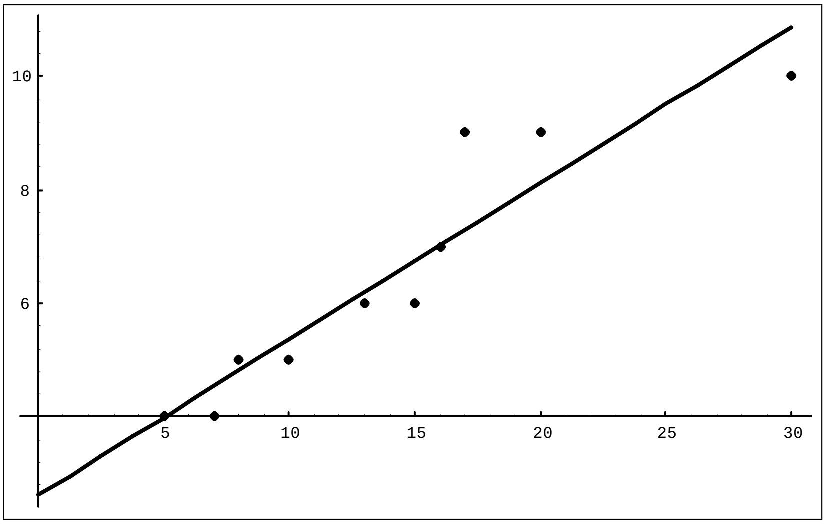
Dacă ρ^*_{XY} se apropie de +1 sau de –1 putem spera ca X şi Y să fie corelate liniar. Dacă nu, se continuă investigatia prin analize mai fine.

Să începem prin a examina următorul exemplu:

Exemplul 14.2 *Ne interesează care este legătura între numărul de ore afectate de un student de inteligentă medie studiului Analizei matemetice (lunar) şi rezultatele obtinute de acesta la examen. În urma unui sondaj efectuat pe 10 studenti s–au obtinut următoarele rezultate:*

Student	*Nr. ore: x*	*Nota: y*
1	*5*	*4*
2	*7*	*4*
3	*8*	*5*
4	*10*	*5*
5	*13*	*6*
6	*15*	*6*
7	*16*	*7*
8	*17*	*9*
9	*20*	*9*
10	*30*	*10*

Punem într-un grafic aceste date:



Dreapta de regresie

Se pune problema dacă aceste puncte sunt "foarte" apropiate de o dreaptă. Mai exact, să notăm cu (x_i, y_i), $i = \overline{1,n}$ valorile obtinute dintr-un sondaj pentru v.a. X şi Y. Există $b_0, b_1 \in \mathbb{R}$ astfel încât diferentele $e_i = y_i - b_0 - b_1 x_i$ să fie "mici"? Care sunt "cei mai buni" b_0 şi b_1 care să facă acest lucru? Sau poate există $b_0, b_1, ..., b_k$ astfel încât diferentele $e_i = y_i - b_0 - b_1 x_i - b_2 x_i^2 - \cdots - b_k x_i^k$ să fie mici pentru orice $i = \overline{1,n}$?

Definiţia 14.3 *Ecuatia*

$$y_i = \beta_0 + \beta_1 x_i + e_i, i = \overline{1,n} \tag{14.6}$$

se numeste *model de regresie simplă (sau liniară)*, iar ecuatia

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + e_i \tag{14.7}$$

unde $i = \overline{1,n}$, se numeste *model de regresie multiplă.*

De exemplu, pentru $k = 2$ se numeste *regresie parabolică*, etc.

Noi ne vom ocupa aici în exclusivitate cu regresia liniară (simplă). Vom interpreta y_i ca fiind valorile unei v.a.. La fel vom interpreta valorile *erorilor* e_i. De asemenea vom interpreta β_0 şi β_1 ca fiind valorile unor v.a. pe care le vom determina prin metoda celor mai mici pătrate.

14.2.1 Metoda celor mai mici pătrate (C. F. Gauss)

Dacă vrem să aproximăm $y_i \approx \beta_0 + \beta_1 x_i$ eroarea comisă este e_i, $i = \overline{1,n}$. Vom pune conditia ca suma pătratelor erorilor e_i să fie minimă:

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \text{minimă}. \tag{14.8}$$

(vezi şi Lecţia 8). Este usor de arătat că S=S(β_0, β_1) are un singur minim pentru $\beta_0 = b_0$, $\beta_1 = b_1$, unde b_0 şi b_1 reprezintă solutia sistemului liniar

$$\frac{\partial \text{S}}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \tag{14.9}$$

$$\frac{\partial \text{S}}{\partial \beta_1} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)\, x_i = 0 \tag{14.10}$$

Notăm cu $\overline{y} = n^{-1}\sum\limits_{i=1}^{n} y_i$ şi cu $\overline{x} = n^{-1}\sum\limits_{i=1}^{n} x_i$.
Atunci, din (14.9) rezultă că :

$$b_0 = \overline{y} - b_1\overline{x} \tag{14.11}$$

Dar b_0 şi b_1 verifică şi (14.10):

$$\sum_{i=1}^{n} y_1 x_i - nb_0 - b_1\sum_{i=1}^{n} x_i^2 = 0 \tag{14.12}$$

(14.11) şi (14.12) ne conduc la expresia lui b_1:

$$b_1 = \left(\sum_{i=1}^{n} y_i x_i - n\overline{xy}\right) \Big/ \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) \tag{14.13}$$

Nu este greu de arătat că b_1 se mai poate scrie sub formă ”centrată”:

$$b_1 = \frac{\sum\limits_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2} = \frac{cov\,(x, y)}{\sigma_x^{*2}} \tag{14.14}$$

Cu acestă expresie a lui b_1 venim în (14.11) şi găsim

$$b_0 = \overline{y} - \overline{x}\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x}) \Big/ \sum_{i=1}^{n} (x_i - \overline{x})^2 = \bar{y} - \bar{x}\frac{cov\,(x, y)}{\sigma_x^{*2}} \tag{14.15}$$

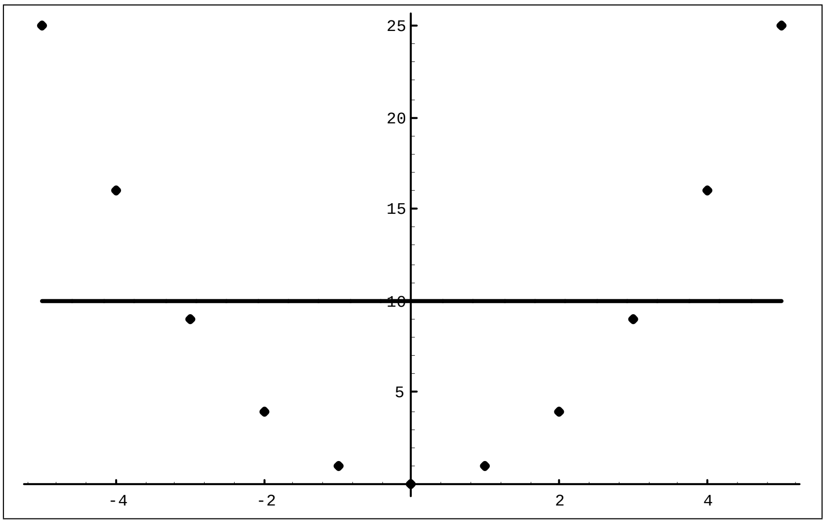
(Vezi şi formulele corespunzătoare din Lecţia 8). În exemplul 14.2 metoda celor mai mici pătrate ne dă $b_0 = 2,621$, $b_1 = 0,274$, $y = b_0 + b_1 x$ fiind dreapta cea mai apropiată de norul de puncte respectiv(vezi figura ”Dreapta de regresie”).

14.2.2 Conditiile Gauss–Markov pentru metoda celor mai mici pătrate

Conditiile Gauss–Markov sunt conditii naturale care se impun v.a. e_i, $i = \overline{1, n}$. Prima conditie cere ca media v.a. e_i să fie zero:

$$M(e_i) = 0, \quad \text{pentru} \quad \text{orice} \quad i = \overline{1, n} \tag{14.16}$$

Să observăm că oricum $\sum_{i=1}^{n} e_i$=0, deci $\sum_{i=1}^{n}$M(e_i)=0, în general.
Figura următoare ne arată un caz în care nu are loc (14.16).

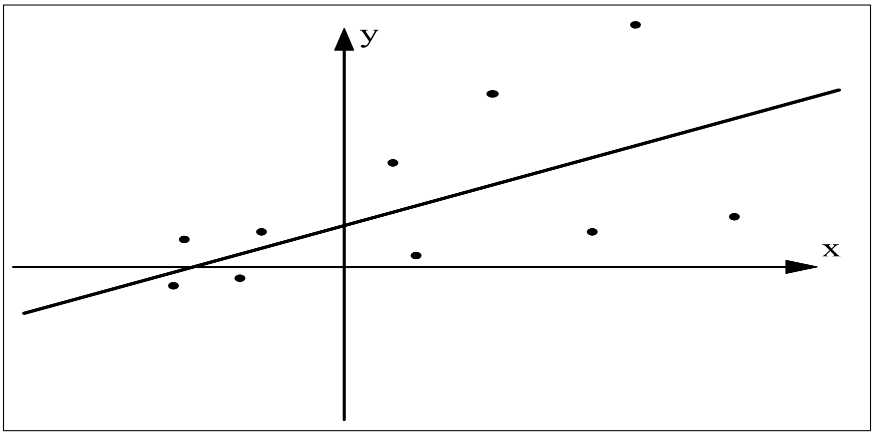


Un caz în care nu este îndeplinită prima condiţie Gauss-Markov

A doua conditie Gauss–Markov cere ca dispersiile v.a. e_i să fie constante, adică

$$D(e_i) = \sigma^2 = \text{constantă, dar necunoscută}$$

Figura următoare ne prezintă o situatie în care *nu* are loc (14.17) deoarece D(e_i) cresc odată cu cresterea i–urilor.



Un caz în care nu este îndeplinită a doua condiţie Gauss Markov

Uneori chiar un singur punct poate să facă condtiile (14.16) sau (14.17) neadevărate.

Dacă observatiile noastre sunt corelate unele cu altele *nu* putem face mai târziu aprecieri pertinente asupra lor. De aceea vom micsora numărul acestor observatii până când acestea vor deveni necorelate.

Ultima conditie Gauss–Markov se referă tocmai la acest lucru. Se cere ca $\mathrm{cov}\,(e_i, e_j) = 0$. Cum M($e_i$)=M($e_j$)=0, rămâne doar conditia:

$$M(e_i e_j) = 0, \text{ pentru orice } i \neq j. \tag{14.17}$$

Definiţia 14.4 *Metoda celor mai mici pătrate se spune că este o metodă bună dacă variabilele aleatoare erori, e_i, i=1, 2,..., n îndeplinesc cele trei conditii Gauss–Markov (14.16), (14.17) şi (14.18).*

14.2.3 Măsura deviatiei la metoda celor mai mici pătrate

Am văzut mai sus că S=$\sum\limits_{i=1}^{n} e_i^2$ măsoară deviatia adevăratelor y_i de la valorile estimate prin metodă , $\widehat{y}_i = b_0 + b_1 x_i$, deoarece $e_i = y_i - \widehat{y}_i$. Cum nu se doreste ca măsura deviatiei să depindă de unitatea de măsură, se lucrează cu altă mărime, oarecum relativă.

Definiţia 14.5 *Fie modelul de regresie liniară $y_i = \beta_0 + \beta_1 x_i + e_i$, $i = 1, 2, ..., n$.*

a) Dacă $\beta_0 \neq 0$ se ia ca măsură a deviatiei expresia: $R^2 = 1 - \sum\limits_{i=1}^{n} e_i^2 \Big/ \sum\limits_{i=1}^{n} (y_i - \overline{y})^2$.

b) Dacă $\beta_0 = 0$ se ia ca măsură a deviatiei expresia: $R^2 = 1 - \sum\limits_{i=1}^{n} e_i^2 \Big/ \sum\limits_{i=1}^{n} y_i^2$.

• **Media şi varianta v.a.** b_0 **şi** b_1

Să interpretăm acum pe y_i şi pe e_i ca variabile aleatoare, pe β_0 şi β_1 ca niste constante (parametri), iar pe b_0 şi b_1 ca variabile aleatoare care iau diferite valori la fiecare selectie în parte.

Teorema 14.6 *Fie modelul de regresie liniară $y_i = \beta_0 + \beta_1 x_i + e_i$, $i = \overline{1,n}$ în care v.a. e_i, $i = \overline{1,n}$ verifică cele 3 condiţii Gauss–Markov. Atunci avem relatiile*

$$M(b_0) = \beta_0 \qquad D(b_0) = \sigma^2 \left[n^{-1} + \frac{\overline{x}^2}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2} \right]$$

$$M(b_1) = \beta_1 \qquad D(b_1) = \sigma^2 \diagup \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{14.18}$$

Demonstratie Deoarece $\sum_{i=1}^{n} (x_i - \overline{x}) = 0$ rezultă că

$$\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x}) = \sum_{i=1}^{n} y_i (x_i - \overline{x}) \tag{14.19}$$

Din (14.20) şi (14.14) rezultă că

$$\begin{aligned} b_1 &= \sum_{i=1}^{n} c_i y_i, \text{ unde} \\ c_i &= (x_i - \overline{x}) \diagup \sum_{i=1}^{n} (x_i - \overline{x}) \end{aligned} \tag{14.20}$$

Cu aceste notatii avem:

$$\begin{aligned} \sum_{i=1}^{n} c_i &= 0 \\ \sum_{i=1}^{n} c_i x_i &= \sum_{i=1}^{n} c_i (x_i - \bar{x}) = 1 \text{ , de unde} \\ \sum_{i=1}^{n} c_i^2 &= \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{aligned}$$

De aici, aplicând operatorul de medie relatiei (14.21) în care $y_i = \beta_0 + \beta_1 x_i + e_i$ găsim:

$$M(b_1) = \sum_{i=1}^{n} c_i M(y_i) = \beta_0 \underbrace{\sum_{i=1}^{n} c_i}_{=0} + \beta_1 \underbrace{\sum_{i=1}^{n} c_i x_i}_{=1} + \sum c_i \underbrace{M(e_i)}_{=0} = \beta_1 \tag{14.21}$$

Calculăm acum D(b_1):

$$D(b_1) = \sum_{i=1}^{n} c_i^2 D(y_i) = \left(\sum_{i=1}^{n} c_i^2 \right) \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{14.22}$$

Acum, deoarece M($\overline{y}$)=$\frac{\sum\limits_{i=1}^{n} M(y_i)}{n}$=$\beta_0 + \beta_1\overline{x}$, rezultă că

$$M(b_0) = M\left(\overline{y} - b_1\overline{x}\right) = \beta_0 + \beta_1\overline{x} - \overline{x}M(b_1) = \beta_0 \tag{14.23}$$

Calculăm acum D(b_0):

$b_0 = n^{-1}\sum\limits_{i=1}^{n} y_i - \overline{x}\sum\limits_{i=1}^{n} c_i y_i = \sum\limits_{i=1}^{n}\left(n^{-1} - \overline{x}c_i\right)y_i$, deci

$$\begin{aligned} D(b_0) &= \sum_{i=1}^{n}\left[n^{-1} - \overline{x}c_i\right]^2 D(y_i) \\ &= \sigma^2\sum_{i=1}^{n}\left[n^{-2} - 2n^{-1}\overline{x}c_i + \overline{x}^2c_i^2\right] \\ &= \sigma^2\left[n^{-1} + \frac{\overline{x}^2}{\sum\limits_{i=1}^{n}\left(x_i - \overline{x}\right)^2}\right] \end{aligned}$$

deoarece $\sum\limits_{i=1}^{n} c_i = 0$.

În cazul în care $\beta_0 = 0$, avem direct din $\frac{\partial \mathrm{S}}{\partial \beta_1} = 0$ că

$$b_1 = \frac{\sum\limits_{i=1}^{n} y_i x_i}{\sum\limits_{i=1}^{n} x_i^2} \qquad e_i = y_i - b_1 x_i \tag{14.24}$$

$\left(\text{în general}\sum\limits_{i=1}^{n} e_i \neq 0\right)$.

Înlocuim în (14.25) pe y_i cu $\beta_1 x_i + e_i$ şi găsim că

$$b_1 = \frac{\beta_1\sum\limits_{i=1}^{n} x_i^2}{\sum\limits_{i=1}^{n} x_i^2} + \frac{\sum\limits_{i=1}^{n} e_i x_i}{\sum\limits_{i=1}^{n} x_i^2} = \beta_1 + \frac{\sum\limits_{i=1}^{n} e_i x_i}{\sum\limits_{i=1}^{n} x_i^2} \tag{14.25}$$

De aici rezultă că

$$D(b_1) = \sigma^2\frac{\sum\limits_{i=1}^{n} x_i^2}{\left(\sum\limits_{i=1}^{n} x_i^2\right)^2} = \frac{\sigma^2}{\sum\limits_{i=1}^{n} x_i^2}$$

şi că M(b_1)=β_1, deoarece M(e_i)=0, pentru orice $i = \overline{1, n}$.□

Corolarul 14.7 *Cu formele de mai sus, rezultă din Teorema 14.6 că b_0 este un estimator nedeplasat al parametrului β_0 şi că b_1 este un estimator nedeplasat al parametrului β_1.*□

Formulele (14.19) arată că D(b_0) şi D(b_1) contin pe σ^2 care este necunoscut. De obicei σ^2 se estimează cu estimatorul nedeplasat (verificarea nu este simplă!)

$$s^2 = (n-1)^{-1}\sum_{i=1}^{n} e_i^2 \tag{14.26}$$

14.2.4 Intervale de încredere şi teste pentru β_0 şi β_1

Presupunem că avem modelul de regresie liniară: $y_i = \beta_0 + \beta_1 x_i + e_i$, $i = \overline{1, n}$, unde (y_i) şi (e_i) sunt v.a., iar β_0 şi β_1 sunt considerati parametri statistici care au fost estimati mai sus prin b_0 şi b_1.

Presupunem că acest model îndeplineşte conditiile Gauss–Markov. De asemenea facem presupunerea că v.a. (e_i) au o distributie normală N(0,σ^2). Atunci rezultă (vezi Lectia 4) că (y_i) au o distributie normală N($\beta_0 + \beta_1 x_i$, σ^2). Cum b_0 şi b_1 sunt combinatii liniare de (y_i)-uri rezultă că şi ele sunt v.a. cu mediile şi dispersiile date în formula (14.19).

Se poate arăta că v.a. $\left(b_j - \beta_j\right) \diagup \sqrt{\mathrm{D}(b_j)}$ este o v.a. Student cu $n-2$ grade de liberate, pentru $j = 0, 1$ (dacă $\beta_0 \neq 0$).

Folosim acum teoria testelor de semnificatie şi găsim că intervalul

$$\left(b_j - \sqrt{\mathrm{D}(b_j)}\mathrm{T}_{n-2,\alpha/2} \qquad b_j + \sqrt{\mathrm{D}(b_j)}\mathrm{T}_{n-2,\alpha/2}\right)$$

este un interval de încredere pentru β_j, j=0,1, de (1–α)×100 procente. Aici T$_{n-2}$, $\alpha/2$ este cuantila de ordin $\alpha/2$ a distributiei Student cu n–2 grade de libertate.

În lumina rezultatelor de mai sus vom studia pe scurt următoarea situatie.

Fie X$_1$, X$_2$,..., X$_n$ variabile aleatoare gaussiene independente cu aceeasi dispersie σ^2 astfel încât să existe două constante a, b cu proprietatea: M(X$_i$)=$a + bt_i$, 1≤ $i \leq n$.

Ca şi mai sus (înlocuim pe Y$_i$ cu M(X$_i$)!) introducem următoarele notatii:

$$\begin{aligned}
\bar{t} &= \frac{1}{n}\sum_{i=1}^{n} t_i; \overline{\mathrm{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i \\
S_t^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(t_i - \bar{t}\right)^2 = \frac{1}{n}\sum_{i} t_i^2 - \bar{t}^2;
\end{aligned}$$

$$S_{\mathrm{X}}^2 = \frac{1}{n}\sum_i \left(\mathrm{X}_i - \overline{\mathrm{X}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i - \overline{\mathrm{X}}^2$$

$$S_{t\mathrm{X}}^2 = \frac{1}{n}\sum_i \left(t_i - \overline{t}\right)\left(\mathrm{X}_i - \overline{\mathrm{X}}\right) = \frac{1}{n}\sum_i t_i X_i - \overline{t}\,\overline{\mathrm{X}}$$

Statisticile (care dau estimatori pentru a şi b): $\widehat{b} = \mathrm{S}_{t\mathrm{X}}^2 \big/ \mathrm{S}_t^2$ şi $\widehat{a} = \overline{\mathrm{X}} - \overline{t}\widehat{b}$ sunt v.a. gaussiene cu $\mathrm{M}(\widehat{b}) = \mathrm{b}$,

$$M(\widehat{a}) = a, \qquad D(\widehat{a}) = \frac{\sigma^2}{n}\left(1 + \frac{\overline{t}^2}{\mathrm{S}_t^2}\right), \qquad D(\widehat{b}) = \frac{\sigma^2}{n\mathrm{S}_t^2}$$

(vezi Teorema 1). Se poate arăta că statistica

$$\rho_{t\mathrm{X}}^2 = \frac{1}{n}\sum_i \left(X_i - \widehat{a} - \widehat{b}t_i\right)^2 = \frac{\mathrm{S}_t^2 \cdot \mathrm{S}_{\mathrm{X}}^2 - \left(\mathrm{S}_{t\mathrm{X}}^2\right)^2}{\mathrm{S}_t^2} = S_{\mathrm{X}}^2 - \widehat{b} S_{t\mathrm{X}}^2$$

este independentă de statisticile $\widehat{a}$ şi $\widehat{b}$. Se poate arăta de asemenea că $\frac{n}{\sigma^2}\rho_{t\mathrm{X}}^2$ se supune unei legi Pearson (χ^2) cu n–2 grade de libertate. Această observatie ne permite să construim un test de semnificatie şi un nou tip de intervale de încredere pentru parametrii a şi b.

Mai mult, se poate arăta că statisticile

$$T = \frac{\sqrt{\mathrm{S}_t^2}\left(\widehat{b} - b\right)}{\sqrt{\frac{1}{n-2}\rho_{t\mathrm{X}}^2}} = \frac{\sqrt{n-2}\,\mathrm{S}_t^2\left(\widehat{b} - b\right)}{\sqrt{\mathrm{S}_t^2 \cdot \mathrm{S}_{\mathrm{X}}^2 - \left(\mathrm{S}_{t\mathrm{X}}^2\right)^2}}$$

şi

$$U = \frac{\sqrt{n(n-2)}\mathrm{S}_t^2(\widehat{a} - a)}{\sqrt{\sum_i t_i^2}\sqrt{\mathrm{S}_t^2 \cdot \mathrm{S}_{\mathrm{X}}^2 - \left(\mathrm{S}_{t\mathrm{X}}^2\right)^2}} = \frac{\sqrt{\mathrm{S}_t^2}(\widehat{a} - a)}{\sqrt{\frac{1}{n-2}\rho_{t\mathrm{X}}^2\left(\mathrm{S}_t^2 + \overline{t}^2\right)}}$$

satisfac legea Student cu n–2 grade de libertate. Si ele pot fi folosite pentru cei doi parametri a şi b.

Exemplul 14.8 *Un biolog studiază cresterea unei specii de plante, pe mai multe exemplare, într–un interval de timp dat. La începutul perioadei planta avea (în mm) înăltimea initială t. La sfârsitul perioadei ea a avut înăltimea X. S-au făcut 10 probe:*

t	57	60	52	49	56	46	51	63	49	57
X	86	93	77	67	81	70	71	91	67	82

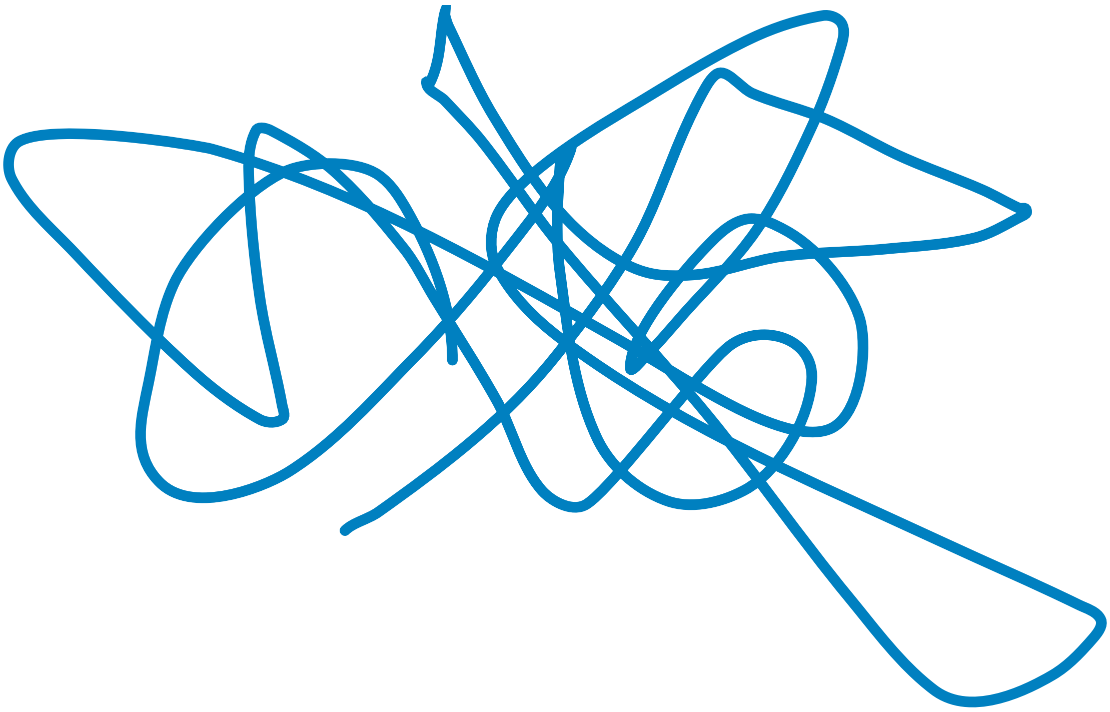
1) Găsiti estimatori punctuali pentru a, b şi σ^2.
2) Estimati înăltimea unei plante la sfârsitul perioadei dacă initial ea a avut 52 mm.
3) Dati pentru b un interval de încredere de 95%.

Solutie 1) Folosim formulele de mai sus şi găsim: n=10; $n\bar{t}$=540; $n\overline{\mathrm{X}}$=785; $\sum_{i=1}^{10} t_i^2$=29426; $\sum_i \mathrm{X}_i^2$=62459; $\sum_i t_i \mathrm{X}_i$=42836; $\bar{t}$=54; $\overline{\mathrm{X}}$=78,5; $n\mathrm{S}_t^2$=266; $n\mathrm{S}_{t\mathrm{X}}^2$=446; $n\mathrm{S}_{\mathrm{X}}^2$=836,5. De aici se găseste pentru b estimarea $\widehat{b}$=1,677 şi pentru a, $\widehat{a}$=–12,06.

Dar $n\rho_{t\mathrm{X}}^2$=88,5. Folosim acum faptul că $\frac{n\rho_{t\mathrm{X}}^2}{\sigma^2}$ se supune unei legi Pearson cu n–2=8 grade de libertate. Pentru σ^2 avem estimarea $\widehat{\sigma}^2$=$\frac{n\rho_{t\mathrm{X}}^2}{8}$=11,06.

2) O plantă cu t=52 la începutul perioadei, va avea la sfârsitul duratei lungimea $\widehat{a}+52\widehat{b}$=75,14.

3) Dacă Y=$\widehat{b}$ şi Z=$\sqrt{\frac{\rho_{t\mathrm{X}}^2}{\mathrm{S}_t^2(n-2)}}$, v.a. T=$\frac{\mathrm{Y}-b}{\mathrm{Z}}$ se supune unei legi Student cu n–2=8 grade de libertate. Avem deci că P(|T| $> t_\varepsilon$)=0,05 (=95%) pentru t_ε=2,306. Se obtine de aici un interval de încredere de prag 95% pentru b: $\widehat{b} - t_\varepsilon\,|\mathrm{Z}| \leq b \leq \widehat{b} + t_\varepsilon\,|\mathrm{Z}|$. Dar Z=$\sqrt{\frac{n\rho_{t\mathrm{X}}^2}{n(n-2)\mathrm{S}_t^2}}$ şi deci $t_\varepsilon\,|\mathrm{Z}|$=0,47. Găsim în final intervalul 1,207$\leq b \leq$2,147.



Bibliografie

1. Alain Cambrouze - Probabilités et Statistique, Press Universitaires de France, 1993
2. G. Ciucu, V. Craiu - Introducere teoria probabilităţilor şi Statistica matematic, Editura Didactică şi Pedagogică, Bucureşti, 1971
3. Harald Cramer - Mathematical Methods of Statistics, Princeton University Press, 1946
4. W. Feller - An introduction to Probability Theory and Its Applications, Vol. I, John Wiley&Sons, Inc. 1960
5. B.V. Gnedenko - The theory of Probability, Mir, Moscow, 1969
6. M. Iosifescu, Gh. Mihoc, R. Theodorescu - Teoria probabilităţilor şi Statistică matematică, Editura Tehnică, Bucureşti, 1966
7. P. Jaquard - Probabilités, Statistique (Culegere de probleme), Masson, Paris, 1972
8. A. Krief, S. Levy - Calcul des Probabilités (Exercices), Hermann, Paris, 1982
9. D. Lungu, D. Chiocel - Metode probabilistice In calculul construcţiilor, Editura Tehnică, 1982
10. Ashis Sen, Muni Srivastava - Regression Analysis, Theory, Methods and Applications, Springer Texts in Statistics, Springer-Verlag, New-York, Inc. 1990
11. H. Ventsel - Théorie des probabilités, Edition Mir, Moscou, 1973
12. R.L. Winkler, W. L. Hays - Statistics, Holt, Reinhart and Winston, N.Y., 1975

TABELUL 0. Funcţia Laplace $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz$

x	0	1	2	3	4	5	6	7	8	9
0.0	0.00000	00399	00798	01197	01595	01994	02392	02790	03188	03586
0.1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535
0.2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409
0.3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173
0.4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793
0.5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240
0.6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490
0.7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524
0.8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327
0.9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891
1.0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214
1.1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298
1.2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147
1.3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774
1.4	41924	42073	42220	42364	42507	42647	42785	42922	43056	43189
1.5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408
1.6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449
1.7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327
1.8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062
1.9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670
2.0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169
2.1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574
2.2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899
2.3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158
2.4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361
2.5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520
2.6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643
2.7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736
2.8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807
2.9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861

3.0	0.49865	3.1	49903	3.2	49931	3.3	49952	3.4	49966
3.5	49977	3.6	49984	3.7	49989	3.8	49993	3.9	49995
4.0	499968								
4.5	499997								
5.0	49999997								

I

TABELE

TABELUL 1 DISTRIBUŢIA NORMALĂ STANDARD

$$F(z)=\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$$

- Exemple: $F(1,02)=0,8461358$, $F(-3,40)=1-F(3,40)=1-0,9996631=0,0003369$.

z	$F(z)$	z	$F(z)$	z	$F(z)$	z	$F(z)$
.00	.5000000	.36	.6405764	.72	.7642375	1.08	.8599289
.01	.5039894	.37	.6443088	.73	.7673049	1.09	.8621434
.02	.5079783	.38	.6480273	.74	.7703500	1.10	.8643339
.03	.5119665	.39	.6517317	.75	.7733726	1.11	.8665005
.04	.5159534	.40	.6554217	.76	.7763727	1.12	.8686431
.05	.5199388	.41	.6590970	.77	.7793501	1.13	.8707619
.06	.5239222	.42	.6627573	.78	.7823046	1.14	.8728568
.07	.5279032	.43	.6664022	.79	.7852361	1.15	.8749281
.08	.5318814	.44	.6700314	.80	.7881446	1.16	.8769756
.09	.5358564	.45	.6736448	.81	.7910299	1.17	.8789995
.10	.5398278	.46	.6772419	.82	.7938919	1.18	.8809999
.11	.5437953	.47	.6808225	.83	.7967306	1.19	.8829768
.12	.5477584	.48	.6843863	.84	.7995458	1.20	.8849303
.13	.5517168	.49	.6879331	.85	.8023375	1.21	.8868606
.14	.5556700	.50	.6914625	.86	.8051055	1.22	.8887676
.15	.5596177	.51	.6949743	.87	.8078498	1.23	.8906514
.16	.5635595	.52	.6984682	.88	.8105703	1.24	.8925123
.17	.5674949	.53	.7019440	.89	.8132671	1.25	.8943502
.18	.5714237	.54	.7054015	.90	.8159399	1.26	.8961653
.19	.5753454	.55	.7088403	.91	.8185887	1.27	.8979577
.20	.5792597	.56	.7122603	.92	.8212136	1.28	.8997274
.21	.5831662	.57	.7156612	.93	.8238145	1.29	.9014747
.22	.5870644	.58	.7190427	.94	.8263912	1.30	.9031995
.23	.5909541	.59	.7224047	.95	.8289439	1.31	.9049021
.24	.5948349	.60	.7257469	.96	.8314724	1.32	.9065825
.25	.5987063	.61	.7290691	.97	.8339768	1.33	.9082409
.26	.6025681	.62	.7323711	.98	.8364569	1.34	.9098773
.27	.6064199	.63	.7356527	.99	.8389129	1.35	.9114920
.28	.6102612	.64	.7389137	1.00	.8413447	1.36	.9130850
.29	.6140919	.65	.7421539	1.01	.8437524	1.37	.9146565
.30	.6179114	.66	.7453731	1.02	.8461358	1.38	.9162067
.31	.6217195	.67	.7485711	1.03	.8484950	1.39	.9177356
.32	.6255158	.68	.7517478	1.04	.8508300	1.40	.9192433
.33	.6293000	.69	.7549029	1.05	.8531409	1.41	.9207302
.34	.6330717	.70	.7580363	1.06	.8554277	1.42	.9221962
.35	.6368307	.71	.7611479	1.07	.8576903	1.43	.9236415

II

z	F(z)	z	F(z)	z	F(z)	z	F(z)
1.44	.9250663	1.77	.9616364	2.10	.9821356	2.43	.9924506
1.45	.9264707	1.78	.9624620	2.11	.9825708	2.44	.9926564
1.46	.9278550	1.79	.9632730	2.12	.9829970	2.45	.9928572
1.47	.9292191	1.80	.9640697	2.13	.9834142	2.46	.9930531
1.48	.9305634	1.81	.9648521	2.14	.9838226	2.47	.9932443
1.49	.9318879	1.82	.9656205	2.15	.9842224	2.48	.9934309
1.50	.9331928	1.83	.9663750	2.16	.9846137	2.49	.9936128
1.51	.9344783	1.84	.9671159	2.17	.9849966	2.50	.9937903
1.52	.9357445	1.85	.9678432	2.18	.9853713	2.51	.9939634
1.53	.9369916	1.86	.9685572	2.19	.9857379	2.52	.9941323
1.54	.9382198	1.87	.9692581	2.20	.9860966	2.53	.9942969
1.55	.9394292	1.88	.9699460	2.21	.9864474	2.54	.9944574
1.56	.9406201	1.89	.9706210	2.22	.9867906	2.55	.9946139
1.57	.9417924	1.90	.9712834	2.23	.9871263	2.56	.9947664
1.58	.9429466	1.91	.9719334	2.24	.9874545	2.57	.9949151
1.59	.9440826	1.92	.9725711	2.25	.9877755	2.58	.9950600
1.60	.9452007	1.93	.9731966	2.26	.9880894	2.59	.9952012
1.61	.9463011	1.94	.9738102	2.27	.9883962	2.60	.9953388
1.62	.9473839	1.95	.9744119	2.28	.9886962	2.70	.9965330
1.63	.9484493	1.96	.9750021	2.29	.9889893	2.80	.9974449
1.64	.9494974	1.97	.9755808	2.30	.9892759	2.90	.9981342
1.65	.9505285	1.98	.9761482	2.31	.9895559	3.00	.9986501
1.66	.9515428	1.99	.9767045	2.32	.9898296	3.20	.9993129
1.67	.9525403	2.00	.9772499	2.33	.9900969	3.40	.9996631
1.68	.9535213	2.01	.9777844	2.34	.9903581	3.60	.9998409
1.69	.9544860	2.02	.9783083	2.35	.9906133	3.80	.9999277
1.70	.9554345	2.03	.9788217	2.36	.9908625	4.00	.9999683
1.71	.9563671	2.04	.9793248	2.37	.9911060	4.50	.9999966
1.72	.9572838	2.05	.9798178	2.38	.9913437	5.00	.9999997
1.73	.9581849	2.06	.9803007	2.39	.9915758	5.50	.9999999
1.74	.9590705	2.07	.9807738	2.40	.9918025		
1.75	.9599408	2.08	.9812372	2.41	.9920237		
1.76	.9607961	2.09	.9816911	2.42	.9922397		

TABELUL II FRACTILELE (CUANTILELE) DISTRIBUȚIEI *T*

- $v = 1, 2, \ldots, 30, 40, \ldots$ reprezintă numărul gradelor de libertate.
- $\alpha = 0{,}60; 0{,}75; 0{,}90; \ldots$ reprezintă ordinul fractilei
- dacă ordinul fractilei este mai mic decât 0,50 ($\alpha = F(t) < 0{,}50$), atunci folosim formula $t = -u$, unde $F(u) = 1-\alpha$, deoarece $F(-t) = 1-F(t)$.
- Exemple: pentru $v = 11$, $\alpha = 0{,}95$, fractila este 1,796; pentru $v = 5$, $\alpha = 0{,}025$, fractila este -2,571.

$F(t)$

v	.60	.75	.90	.95	.975	.99	.995	.999
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	.289	0.816	1.886	2.920	4.303	6.965	9.925	22.326
3	.277	.765	1.638	2.353	3.182	4.541	5.841	10.213
4	.271	.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	.265	.718	1.440	1.943	2.447	3.143	3.707	5.208
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.785
8	.262	.706	1.397	1.860	2.306	2.896	3.355	4.501
9	.261	.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	.260	.697	1.363	1.796	2.201	2.718	3.106	4.025
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.930
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.852
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.686
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.646
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.610
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.527
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.505
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.485
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.435
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.421
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.408
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	.255	.681	1.303	1.684	2.021	2.423	2.704	3.307
60	.254	.679	1.296	1.671	2.000	2.390	2.660	3.232
120	.254	.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	3.090

TABELUL III FRACTILELE (CUANTILELE) DISTRIBUȚIEI χ^2

▪ Exemplu: pentru 14 grade de libertate fractila de ordin 0,025 este 5,62872.

ν	$F(\chi^2)$						
	.005	.01	.025	.05	.10	.25	.50
1	$392704 \cdot 10^{-10}$	$157088 \cdot 10^{-9}$	$982069 \cdot 10^{-9}$	$393214 \cdot 10^{-8}$	0.0157908	0.1015308	0.454937
2	0.0100251	0.0201007	0.0506356	0.102587	0.210720	0.575364	1.38629
3	0.0717212	0.114832	0.215795	0.351846	0.584375	1.212534	2.36597
4	0.206990	0.297110	0.484419	0.710721	1.063623	1.92255	3.35670
5	0.411740	0.554300	0.831211	1.145476	1.61031	2.67460	4.35146
6	0.675727	0.872085	1.237347	1.63539	2.20413	3.45460	5.34812
7	0.989265	1.239043	1.68987	2.16735	2.83311	4.25485	6.34581
8	1.344419	1.646482	2.17973	2.73264	3.48954	5.07064	7.34412
9	1.734926	2.087912	2.70039	3.32511	4.16816	5.89883	8.34283
10	2.15585	2.55821	3.24697	3.94030	4.86518	6.73720	9.34182
11	2.60321	3.05347	3.81575	4.57481	5.57779	7.58412	10.3410
12	3.07382	3.57056	4.40379	5.22603	6.30380	8.43842	11.3403
13	3.56503	4.10691	5.00874	5.89186	7.04150	9.29906	12.3398
14	4.07468	4.66043	5.62872	6.57063	7.78953	10.1653	13.3393
15	4.60094	5.22935	6.26214	7.26094	8.54675	11.0365	14.3389
16	5.14224	5.81221	6.90766	7.96164	9.31223	11.9122	15.3385
17	5.69724	6.40776	7.56418	8.67176	10.0852	12.7919	16.3381
18	6.26481	7.01491	8.23075	9.39046	10.8649	13.6753	17.3379
19	6.84398	7.63273	8.90655	10.1170	11.6509	14.5620	18.3376
20	7.43386	8.26040	9.59083	10.8508	12.4426	15.4518	19.3374
21	8.03366	8.89720	10.28293	11.5913	13.2396	16.3444	20.3372
22	8.64272	9.54249	10.9823	12.3380	14.0415	17.2396	21.3370
23	9.26042	10.19567	11.6885	13.0905	14.8479	18.1373	22.3369
24	9.88623	10.8564	12.4011	13.8484	15.6587	19.0372	23.3367
25	10.5197	11.5240	13.1197	14.6114	16.4734	19.9393	24.3366
26	11.1603	12.1981	13.8439	15.3791	17.2919	20.8434	25.3364
27	11.8076	12.8786	14.5733	16.1513	18.1138	21.7494	26.3363
28	12.4613	13.5648	15.3079	16.9279	18.9392	22.6572	27.3363
29	13.1211	14.2565	16.0471	17.7083	19.7677	23.5666	28.3362
30	13.7867	14.9535	16.7908	18.4926	20.5992	24.4776	29.3360
40	20.7065	22.1643	24.4331	26.5093	29.0505	33.6603	39.3354
50	27.9907	29.7067	32.3574	34.7642	37.6886	42.9421	49.3349
60	35.5346	37.4848	40.4817	43.1879	46.4589	52.2938	59.3347
70	43.2752	45.4418	48.7576	51.7393	55.3290	61.6983	69.3344
80	51.1720	53.5400	57.1532	60.3915	64.2778	71.1445	79.3343
90	59.1963	61.7541	65.6466	69.1260	73.2912	80.6247	89.3342
100	67.3276	70.0648	74.2219	77.9295	82.3581	90.1332	99.3341

V

ν	$F(\chi^2)$ .75	.90	.95	.975	.99	.995	.999
1	1.32330	2.70554	3.84146	5.02389	6.63490	7.87944	10.828
2	2.77259	4.60517	5.99147	7.37776	9.21034	10.5966	13.816
3	4.10835	6.25139	7.81473	9.34840	11.3449	12.8381	16.266
4	5.38527	7.77944	9.48773	11.1433	13.2767	14.8602	18.467
5	6.62568	9.23635	11.0705	12.8325	15.0863	16.7496	20.515
6	7.84080	10.6446	12.5916	14.4494	16.8119	18.5476	22.458
7	9.03715	12.0170	14.0671	16.0128	18.4753	20.2777	24.322
8	10.2188	13.3616	15.5073	17.5346	20.0902	21.9550	26.125
9	11.3887	14.6837	16.9190	19.0228	21.6660	23.5893	27.877
10	12.5489	15.9871	18.3070	20.4831	23.2093	25.1882	29.588
11	13.7007	17.2750	19.6751	21.9200	24.7250	26.7569	31.264
12	14.8454	18.5494	21.0261	23.3367	26.2170	28.2995	32.909
13	15.9839	19.8119	22.3621	24.7356	27.6883	29.8194	34.528
14	17.1170	21.0642	23.6848	26.1190	29.1413	31.3193	36.123
15	18.2451	22.3072	24.9958	27.4884	30.5779	32.8013	37.697
16	19.3688	23.5418	26.2962	28.8454	31.9999	34.2672	39.252
17	20.4887	24.7690	27.5871	30.1910	33.4087	35.7185	40.790
18	21.6049	25.9894	28.8693	31.5264	34.8053	37.1564	42.312
19	22.7178	27.2036	30.1435	32.8523	36.1908	38.5822	43.820
20	23.8277	28.4120	31.4104	34.1696	37.5662	39.9968	45.315
21	24.9348	29.6151	32.6705	35.4789	38.9321	41.4010	46.797
22	26.0393	30.8133	33.9244	36.7807	40.2894	42.7956	48.268
23	27.1413	32.0069	35.1725	38.0757	41.6384	44.1813	49.728
24	28.2412	33.1963	36.4151	39.3641	42.9798	45.5585	51.179
25	29.3389	34.3816	37.6525	40.6465	44.3141	46.9278	52.620
26	30.4345	35.5631	38.8852	41.9232	45.6417	48.2899	54.052
27	31.5284	36.7412	40.1133	43.1944	46.9630	49.6449	55.476
28	32.6205	37.9159	41.3372	44.4607	48.2782	50.9933	56.892
29	33.7109	39.0875	42.5569	45.7222	49.5879	52.3356	58.302
30	34.7998	40.2560	43.7729	46.9792	50.8922	53.6720	59.703
40	45.6160	51.8050	55.7585	59.3417	63.6907	66.7659	73.402
50	56.3336	63.1671	67.5048	71.4202	76.1539	79.4900	86.661
60	66.9814	74.3970	79.0819	83.2976	88.3794	91.9517	99.607
70	77.5766	85.5271	90.5312	95.0231	100.425	104.215	112.317
80	88.1303	96.5782	101.879	106.629	112.329	116.321	124.839
90	98.6499	107.565	113.145	118.136	124.116	128.299	137.208
100	109.141	118.498	124.342	129.561	135.807	140.169	149.449