Probabilități și statistică

CUPRINS

Introducere

Teoria probabilităţilor şi statistica matematică se aplică în majoritatea domeniilor ştiinţei, începând cu ştiinţele exacte şi inginereşti şi finalizând cu ştiinţele socio-economice, în special acolo unde există condiţii de risc şi incertitudine şi unde este necesară adoptarea unor decizii riguros argumentate.

Una dintre construcţiile de bază în fundamentele statisticii şi teoriei probabilităţilor, precum şi în justificarea aplicării acestora în alte domenii, este dată de "legea numerelor mari" teoremă binecunoscută care îi aparţine matematicianului Jakob Bernoulli (1654-1705), fiind apărută în lucrarea postumă "Ars conjectandi" (1713). Printre alţi matematicieni care au rămas celebri în teoria probabilităţilor şi statistică, îi amintim pe: de Moivre, Laplace, Gauss, Bertrand, Poincaré, Cebîşev, Liapunov, Markov, Borel, Kolmogorov, Glivenko. De asemenea, şcoala românească de probabilităţi, fondată de Octav Onicescu, şi reprezentată de nume precum Gheorghe Mihoc şi Marius Iosifescu, a adus contribuţii semnificative în dezvoltarea acestui domeniu.

Cartea de faţă îşi propune să vină în sprijinul studenţilor care au ca disciplină de studiu, în cadrul a diferite specializări, disciplina *Probabilităţi şi statistică,* oferindu-le acestora o gamă largă de aspecte teoretice, însoţite de exemple şi aplicaţii. Ca structură, cartea se fundamentează pe baza a treisprezece capitole, şase dintre acestea fiind dedicate *Teoriei probabilităţilor,* respectiv şapte capitole, *Statisticii matematice.*

În *Capitolul 1,* sunt prezentate concepte de bază ale teoriei probabilităţilor, mai precis, experienţe aleatoare, evenimente, probabilitate, reguli de calcul cu probabilităţi, în timp ce în *Capitolul 2,* sunt abordate noţiuni precum variabile aleatoare, caracteristici numerice ale variabilelor aleatoare, funcţia caracteristică, funcţia generatoare de momente. În *Capitolul 3* sunt prezentate principalele legi de probabilitate ale variabilelor aleatoare discrete şi anume: legea discretă uniformă, legea binomială şi cazul său particular legea Bernoulli, legea binomială cu exponent negativ şi cazul particular legea geometrică, legea hipergeometrică şi legea Poisson (legea evenimentelor rare), iar în *Capitolul 4* sunt prezentate principalele legi de probabilitate ale variabilelor aleatoare continue, şi anume: legea continuă uniformă (rectangulară), legea normală (Gauss-Laplace), legea log-normală, legea gamma, legea beta, legea χ^2 (Helmert-Pearson), legea Student (t) şi cazul său particular legea Cauchy, legea Snedecor şi legea Fisher, legea Weibull şi cazul său particular, legea exponenţială. *Capitolul 5* se construieşte în jurul convergenţei, de diferite tipuri, a variabilelor aleatoare, fiind menţionate convergenţa aproape sigură, convergenţa în probabilitate, convergenţa în repartiţie, precum şi legile numerelor mari şi teorema limită centrală. Partea aferentă teoriei probabilităţilor se încheie cu *Capitolul 6,* dedicat algoritmilor de simulare a variabilelor aleatoare. În *Capitolul 7,* se studiază elemente de

statistică descriptivă şi aspecte privind organizarea datelor, cât şi analiza acestora, punându-se accentul pe modalităţile de reprezentare, dar şi pe găsirea diverselor mărimi caracteristice. Noţiunile sunt însoţite de exemple adecvate şi actuale. În *Capitolul 8* sunt prezentate noţiuni de teoria selecţiei, începând cu descrierea generării unor valori particulare ale variabilelor aleatoare discrete sau continue şi continuând cu legi de probabilitate ale variabilelor de eşantionare. *Capitolul 9* conţine o scurtă introducere în teoria estimaţiei. Se prezintă, cu multe exemple, conceptele de estimator nedeplasat şi estimator de maximă verosimilitate. În *Capitolul 10*, sunt prezentate intervalele de încredere pentru principalii parametri statistici. Astfel, capitolul debutează cu fundamentarea formei generale a unui interval de încredere, ca metodă de estimare statistică, după care sunt prezentate pe rând, intervalul de încredere pentru medie, incluzând cazul când dispersia este necunoscută, respectiv cazul particular al unei proporţii, apoi interval de încredere pentru diferenţa a două medii, respectiv, interval de încredere pentru dispersie şi pentru raportul a două dispersii, toate acestea însoţite de exemple practice. *Capitolul 11* prezintă succint teoria deciziei. Sunt descrise noţiunile de ipoteză statistică, test statistic, tipuri de erori, nivel de semnificaţie, putere a unui test, p-valoare. Exemplele sunt luate din practica testării şirurilor binare în ceea ce priveşte caracterul aleator. În *Capitolul 12*, sunt prezentate tehnici de analiză a regresiei, atât prin intermediul aspectelor teoretice, cât şi prin intermediul unor exemple. În primul paragraf sunt trecute în revistă noţiunile de bază, fiind definite diverse tipuri de modele de regresie, urmând ca în ultimele trei paragrafe spaţiul să fie alocat cu precădere modelului liniar. Astfel, este prezentată metoda celor mai mici pătrate în estimarea parametrilor necunoscuţi ai unui model liniar multiplu, sunt realizate inferenţe asupra estimatorilor unui model liniar în ipotezele clasice Gauss-Markov, capitolul încheindu-se cu aspecte care ţin de previziunea şi analiza rezultatelor unei regresii liniare. Cartea se încheie cu *Capitolul 13* în care sunt abordate aspecte care ţin de statistica bayesiană şi noţiuni de teoria credibilităţii.

Cartea de faţă a fost elaborată în cadrul proiectului POSDRU/56/1.2/S/32768, "Formarea cadrelor didactice universitare şi a studenţilor în domeniul utilizării unor instrumente moderne de predare-învăţare-evaluare pentru disciplinele matematice, în vederea creării de competenţe performante şi practice pentru piaţa muncii", de către un colectiv de autori, cadre didactice universitare, astfel: capitolele 1 şi 2, Lucia Căbulea, capitolele 3 şi 4, Rodica Luca-Tudorache, capitolele 5, 6 şi 13, Gheorghiţă Zbăganu, capitolele 7 şi 8, Ariana Pitea, capitolele 9 şi 11, Ioan Rasa, respectiv capitolele 10 şi 12, Nicoleta Breaz.

Finanţat din Fondul Social European şi implementat de către Ministerul Educaţiei, Cercetării, Tineretului şi Sportului, în colaborare cu The Red Point, Oameni şi Companii, Universitatea din Bucureşti, Universitatea Tehnică de Construcţii din Bucureşti, Universitatea „Politehnica" din Bucureşti, Universitatea din Piteşti, Universitatea Tehnică „Gheorghe Asachi" din Iaşi, Universitatea de Vest din Timişoara, Universitatea „Dunărea de Jos" din Galaţi, Universitatea Tehnică din Cluj-Napoca, Universitatea "1 Decembrie 1918" din

Alba-Iulia, proiectul contribuie în mod direct la realizarea obiectivului general al Programului Operaţional Sectorial de Dezvoltare a Resurselor Umane – POSDRU şi se înscrie în domeniul major de intervenţie 1.2 Calitate în învăţământul superior.

Proiectul are ca obiectiv adaptarea programelor de studii ale disciplinelor matematice la cerinţele pieţei muncii şi crearea de mecanisme şi instrumente de extindere a oportunităţilor de învăţare.

Evaluarea nevoilor educaţionale obiective ale cadrelor didactice şi studenţilor legate de utilizarea matematicii în învăţământul superior, masterate şi doctorate precum şi analizarea eficacităţii şi relevanţei curriculelor actuale la nivel de performanţă şi eficienţă, în vederea dezvoltării de cunoştinţe şi competenţe pentru studenţii care învaţă discipline matematice în universităţi, reprezintă obiective specifice de interes în cadrul proiectului. Dezvoltarea şi armonizarea curriculelor universitare ale disciplinelor matematice, conform exigenţelor de pe piaţa muncii, elaborarea şi implementarea unui program de formare a cadrelor didactice şi a studenţilor interesaţi din universităţile partenere, bazat pe dezvoltarea şi armonizarea de curriculum, crearea unei baze de resurse inovative, moderne şi funcţionale pentru predarea-învăţarea-evaluarea în disciplinele matematice pentru învăţământul universitar sunt obiectivele specifice care au ca răspuns materialul de faţă.

Formarea de competenţe cheie de matematică şi informatică presupune crearea de abilităţi de care fiecare individ are nevoie pentru dezvoltarea personală, incluziune socială şi inserţie pe piaţa muncii. Se poate constata însă că programele disciplinelor de matematică nu au întotdeauna în vedere identificarea şi sprijinirea elevilor şi studenţilor potenţial talentaţi la matematică. Totuşi, studiul matematicii a evoluat în exigenţe până a ajunge să accepte provocarea de a folosi noile tehnologii în procesul de predare - învăţare - evaluare pentru a face matematica mai atractivă.

În acest context, analiza flexibilităţii curriculei, însoţită de analiza metodelor şi instrumentelor folosite pentru identificarea şi motivarea studenţilor talentaţi la matematică ar putea răspunde deopotrivă cerinţelor de masă, cât şi celor de elită.

Viziunea pe termen lung a acestui proiect preconizează determinarea unor schimbări în abordarea fenomenului matematic pe mai multe planuri: informarea unui număr cât mai mare de membri ai societăţii în legătură cu rolul şi locul matematicii în educaţia de bază în instrucţie şi în descoperirile ştiinţifice menite să îmbunătăţească calitatea vieţii, inclusiv popularizarea unor mari descoperiri tehnice, şi nu numai, în care matematica cea mai avansată a jucat un rol hotărâtor. De asemenea, se urmăreşte evidenţierea a noi motivaţii solide pentru învăţarea şi studiul matematicii la nivelele de bază şi la nivel de performanţă; stimularea creativităţii şi formarea la viitorii cercetători matematicieni a unei atitudini deschise faţă de însuşirea aspectelor specifice din alte ştiinţe, în scopul participării cu succes în echipe mixte de cercetare sau a abordării unei cercetări inter şi multidisciplinare; identificarea unor forme de pregătire adecvată de matematică pentru viitorii studenţi ai disciplinelor

matematice, în scopul utilizării la nivel de performanță a aparatului matematic în construirea unei cariere profesionale.

Capitolul 1

Teoria probabilităților

1.1. Formalizarea experiențelor aleatoare

1.1.1. Evenimente

Definiția 1.1.1. *Realizarea practică a unui ansamblu de condiții bine precizat poartă numele de experiență sau probă.*

Definiția 1.1.2. *Prin eveniment vom înțelege orice rezultat al unei experiențe despre care putem spune că s-a realizat sau că nu s-a realizat, după efectuarea experimentului considerat. Evenimentele se pot clasifica în: evenimente sigure; evenimente imposibile, evenimente aleatoare.*

Definiția 1.1.3. *Evenimentul sigur este evenimentul care se produce în mod obligatoriu la efectuarea unei probe şi se notează cu* Ω.

Definiția 1.1.4. *Evenimentul imposibil este evenimentul care în mod obligatoriu nu se produce la efectuarea unei probe şi se notează cu* ϕ.

Definiția 1.1.5. *Evenimentul aleator este evenimentul care poate sau nu să se realizeze la efectuarea unei probe şi se notează prin litere mari A, B, C, ..., sau prin litere mari urmate de indici* A_i, B_i,….

Definiția 1.1.6. *Evenimentul contrar evenimentului A se notează* $\bar{A}$ *şi este evenimentul ce se realizează numai atunci când nu se realizează evenimentul A.*

Definiția 1.1.7. *Un eveniment se numeşte:*

1) *elementar dacă se realizează ca rezultat al unei singure probe; se notează cu* ω.

2) compus dacă acesta apare cu două sau mai multe rezultate ale probei considerate.

Definiția 1.1.8. *Mulțimea tuturor evenimentelor elementare generate de un experiment aleator se numeşte spațiul evenimentelor elementare (spațiul de selecție) şi se notează cu* Ω. *Acesta poate fi finit sau infinit.*

Observația 1.1.9. *O analogie între evenimente şi mulțimi permite o scriere şi în general o exprimare mai comode ale unor idei şi rezultate legate de conceptul de eveniment. Astfel, vom înțelege evenimentul sigur ca mulțime a tuturor evenimentelor elementare, adică:* $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ *şi orice eveniment compus ca o submulțime a lui* Ω. *De asemenea, putem vorbi despre mulțimea tuturor părților lui* Ω *pe care o notăm prin P(*Ω*), astfel că pentru un eveniment compus A putem scrie, în contextul analogiei dintre evenimente şi mulțimi, că* $A \subseteq \Omega$ *sau* $A \in P(\Omega)$.

Exemplul 1.1.10. *Fie un zar, care are cele şase fețe marcate prin puncte de la 1 la 6. Se aruncă zarul pe o suprafață plană netedă. Dacă notăm cu* ω_i = *evenimentul "apariția feței cu i puncte",* $i = \overline{1,6}$, *atunci spațiul evenimentelor elementare ataşat experimentului cu un zar este dat prin* $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$.

Evenimentul sigur Ω *este "apariția feței cu un număr de puncte* ≤ 6*".*

Evenimentul imposibil ϕ *este "apariția feței cu 7 puncte".*

1.2. Relații între evenimente

Definiția 1.2.1. *Spunem că evenimentul A implică evenimentul B şi scriem* $A \subset B$, *dacă realizarea evenimentului A atrage după sine şi realizarea evenimentului B.*

Observația 1.2.2. $A \subset B$ *şi* $B \subset C$ *rezultă* $A \subset C$ *- proprietatea de tranzitivitate a relației de implicare.*

Definiția 1.2.3. *Spunem că evenimentele A şi B sunt echivalente (egale) dacă avem simultan* $A \subset B$ *şi* $B \subset A$.

Definiția 1.2.4. *Prin reunirea evenimentelor A şi B vom înțelege evenimentul notat* $A \cup B$ *care constă în realizarea a cel puțin unuia dintre evenimentele A şi B. Deoarece evenimentele A şi B sunt submulțimi formate cu evenimentele elementare ale spațiului* Ω, *rezultă că reunirea evenimentelor poate fi scrisă astfel:*

$$A \cup B = \{\omega \in \Omega / \omega \in A \text{ sau } \omega \in B\}$$

Observația 1.2.5. *Dacă notăm prin K mulțimea tuturor evenimentelor asociate unui experiment aleator avem:*

1. $\forall A, B \in K \Rightarrow A \cup B = B \cup A$ *(comutativitatea);*
2. $\forall A, B, C \in K \Rightarrow (A \cup B) \cup C = A \cup (B \cup C)$ *(asociativitatea);*

3. Dacă $A, B \in K$ şi $A \subset B \Rightarrow A \cup B = B$ *(evident* $A \cup \Omega = \Omega$, $A \cup \phi = A$, $\Omega \cup \phi = \Omega$ şi $A \cup \overline{A} = \Omega$).

Definiția 1.2.6. *Prin intersecția evenimentelor A şi B vom înțelege evenimentul notat* $A \cap B$ *care constă în realizarea simultană a ambelor evenimente. Intersecția evenimentelor A şi B poate fi scrisă sub forma:*

$$A \cap B = \{\omega \in \Omega / \omega \in A \text{ şi } \omega \in B\}$$

Observația 1.2.7. *Au loc relațiile următoare:*

1. $\forall A, B \in K \Rightarrow A \cap B = B \cap A$ *(comutativitatea)*
2. $\forall A, B, C \in K \Rightarrow (A \cap B) \cap C = A \cap (B \cap C)$ *(asociativitatea)*
3. Dacă $A, B \in K$ şi $A \subset B$ *atunci* $A \cap B = A$ *(evident* $A \cap \Omega = A$, $A \cap \emptyset = \emptyset$, $\Omega \cap \emptyset = \emptyset$ şi $A \cap A = A$).
4. $\forall A \in K \Rightarrow A \cap \overline{A} = \emptyset$

Definiția 1.2.8. *Spunem că evenimentele A şi B sunt incompatibile dacă* $A \cap B = \emptyset$, *adică realizarea lor simultană este imposibilă, şi spunem că sunt compatibile dacă* $A \cap B \neq \emptyset$, *adică este posibilă realizarea lor simultană. Evenimentele A şi B sunt contrare unul altuia dacă* $A \cup B = \Omega$ *şi* $A \cap B = \emptyset$, *adică realizarea unuia constă din nerealizarea celuilalt.*

Definiția 1.2.9. *Se numeşte diferența evenimentelor A şi B, evenimentul notat A-B care se realizează atunci când se realizează evenimentul A şi nu se realizează evenimentul B. Diferența evenimentelor poate fi scrisă sub forma:*

$$A - B = \{\omega \in \Omega / \omega \in A \text{ şi } \omega \notin B\}$$

Observația 1.2.10. *Evident avem* $A - B = A \cap \overline{B}$ şi $E - A = \overline{A}$.

Au loc relațiile lui De Morgan: $\overline{A \cup B} = \overline{A} \cap \overline{B}$ şi $\overline{A \cap B} = \overline{A} \cup \overline{B}$ şi *respectiv generalizările* $\overline{\bigcup_{i \in I} A_i} = \bigcap_{i \in I} \overline{A_i}$; $\overline{\bigcap_{i \in I} A_i} = \bigcup_{i \in I} \overline{A_i}$.

Teorema 1.2.11. *Dacă evenimentele A, B, C, D* $\in K$, *atunci sunt adevărate următoarele afirmații:*

i) $A - B = A - (A \cap B)$
ii) $A - B = (A \cup B) - B$
iii) $A = (A - B) \cup (A \cap B)$
iv) $(A - B) \cap (B - A) = \phi$
v) $A \cup B = A \cup [B - (A \cap B)]$
vi) $A \cap (B - C) = (A \cap B) - (A \cap C)$
vii) $(A - B) \cap (C - D) = (A \cap C) - (B \cup D)$

Definiția 1.2.12. *Evenimentele A şi B sunt dependente dacă realizarea unuia depinde de realizarea celuilalt şi sunt independente dacă realizarea unuia nu depinde de realizarea celuilalt.*

O mulțime de evenimente sunt independente în totalitatea lor dacă sunt independente câte două, câte trei etc.

Pentru evenimentele independente în totalitatea lor vom folosi şi denumirea de evenimente independente.

1.3. Câmp de evenimente

Definiția 1.3.1. *O mulțime nevidă de evenimente* $K \subseteq P(\Omega)$ *se numeşte corp dacă satisface axiomele:*

i) $\forall A \in K \Rightarrow \overline{A} \in K$

ii) $\forall A, B \in K \Rightarrow A \cup B \in K$.

Cuplul (Ω, K) *se numeşte câmp finit de evenimente, în cazul în care K este un corp.*

Observația 1.3.2.

1. *Într-un câmp finit de evenimente* (Ω, K) *sunt adevărate afirmațiile*:

a. $A, B \in K \Rightarrow A - B \in K$

b. *Evident* $\phi \in K$ şi $\Omega \in K$.

c. *Dacă* $A, B \in K$ *atunci* $A \cap B \in K$.

2. *Dacă mulțimea evenimentelor elementare este numărabilă, o mulțime* $K \subseteq P(\Omega)$ *se numeşte corp borelian (sau* σ*-corp, sau* σ*-algebră) pe* Ω, *în condițiile:*

i) $\forall A \in K \Rightarrow \overline{A} \in K$,

ii) *dacă* $I \subseteq N$, $I \neq \phi$ şi $A_i \in K$, $(\forall) i \in I$, *atunci* $\bigcup_{i \in I} A_i \in K$,

iii) $\Omega \in K$.

Perechea (Ω, K) *în care K este un* σ*-corp se numeşte câmp borelian (câmp infinit) de evenimente.*

Definiția 1.3.3. *Într-un câmp finit de evenimente* (Ω, K), *evenimentele* $A_i \in K$, $i = \overline{1, n}$, *formează un sistem complet de evenimente (sau o partiție a câmpului) dacă:*

i) $\bigcup_{i=1}^{n} A_i = \Omega$

ii) $A_i \cap A_j = \phi \quad \forall i \neq j, \; i, j = \overline{1, n}$

Observația 1.3.4. *Evenimentele elementare* ω_i, $\mathrm{i}=\overline{1,\mathrm{n}}$, *corespunzătoare unei probe formează un sistem complet de evenimente care se mai numeşte sistem complet elementar.*

Propoziția 1.3.5. *Dacă* $\Omega=\{\omega_1,\omega_2,...,\omega_n\}$ *atunci câmpul de evenimente corespunzător conține* 2^{n} *evenimente.*

Demonstrație

Pentru un experiment de n rezultate elementare şi prin urmare pentru un eveniment sigur compus din n evenimente elementare, vom avea diverse evenimente compuse din acestea după cum urmează:

– evenimente compuse din câte zero evenimente elementare $=C_n^0$

– evenimente compuse din câte un eveniment elementar $=C_n^1$

– evenimente compuse din câte două evenimente elementare $=C_n^2$

– - - - - - - - - - - - - - - - -

– evenimente compuse din câte k evenimente elementare $=C_n^k$

– evenimente compuse din câte n evenimente elementare $=C_n^n$

şi prin urmare, numărul total de evenimente ale lui K este egal cu

$$C_n^0+C_n^1+...+C_n^k+...+C_n^n=2^n$$

1.4. Câmp de probabilitate

Definiția 1.4.1.(axiomatică a probabilității) *Fie* (Ω,K) *un câmp finit de evenimente. Se numeşte probabilitate pe câmpul considerat o funcție* $\mathrm{P}:\mathrm{K}\to\mathrm{R}$ *care satisface axiomele:*

i) $P(A)\geq 0,\ \forall A\in K$,

ii) $\mathrm{P}(\Omega)=1$,

iii) $\mathrm{P}(\mathrm{A}\cup\mathrm{B})=\mathrm{P}(\mathrm{A})+\mathrm{P}(\mathrm{B}),\ \forall\,\mathrm{A},\mathrm{B}\in\mathrm{K},$ şi $\mathrm{A}\cap\mathrm{B}=\phi$.

Definiția 1.4.2. *Se numeşte câmp finit de probabilitate tripletul* $\{\Omega,,\mathrm{K},\mathrm{P}\}$ *unde cuplul* (Ω,K) *este un câmp finit de probabilitate, iar* $\mathrm{P}:\mathrm{K}\to\mathrm{R}$ *este o probabilitate pe K.*

Observația 1.4.3. *În cazul în care câmpul de evenimente* (Ω,K) *este infinit (K este infinită) probabilitatea P definită pe K satisface axiomele:*

i) $\mathrm{P}(\mathrm{A})\geq 0,\forall\mathrm{A}\in\mathrm{K}$

ii) $\mathrm{P}(\Omega)=1$

iii) $P\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} P(A_i)$ dacă $A_i \cap A_j = \emptyset$ $i \neq j$, $i, j \in I$, $A_i \in K$,

I-o mulțime de indici cel mult numărabilă.

Propoziția 1.4.4. *Au loc relațiile:*

1. $P(\emptyset) = 0$
2. $P(\overline{A}) = 1 - P(A)$
3. $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$ dacă $A_i \cap A_j = \emptyset$ $i \neq j$, $i, j = \overline{1,n}$

Demonstrație

1) Din relațiile $\emptyset \cup \Omega = \Omega$ şi $\emptyset \cap \Omega = \emptyset$ aplicând axioma iii) din definiția probabilității avem $P(\Omega) = P(\emptyset \cup \Omega) = P(\emptyset) + P(\Omega)$ şi rezultă $P(\emptyset) = 0$

2) Din relațiile $A \cup \overline{A} = \Omega$ *şi* $A \cap \overline{A} = \emptyset$ aplicând axioma iii) din definiția probabilității avem $P(A \cup \overline{A}) = P(A) + P(\overline{A})$ adică

$P(\Omega) = P(A) + P(\overline{A})$ şi rezultă $P(\overline{A}) = 1 - P(A)$

3) Demonstrăm prin inducție matematică

Pentru n = 2 $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ relația este adevărată conform axiomei iii) din definiția probabilității.

Presupunem relația adevărată pentru n – 1 evenimente, adică

$P\left(\bigcup_{i=1}^{n-1} A_i\right) = \sum_{i=1}^{n-1} P(A_i)$ şi demonstrăm pentru n evenimente

$$P\left(\bigcup_{i=1}^{n} A_i\right) = P\left[\left(\bigcup_{i=1}^{n-1} A_i\right) \cup A_n\right] = P\left(\bigcup_{i=1}^{n-1} A_i\right) + P(A_n) = \sum_{i=1}^{n-1} P(A_i) + P(A_n) = \sum_{i=1}^{n} P(A_i)$$

dacă s-a folosit ipoteza de inducție şi s-a ținut seama că $\left(\bigcup_{i=1}^{n-1} A_i\right) \cap A_n = \emptyset$.

Definiția 1.4.5. (clasică a probabilității) *Probabilitatea unui eveniment A este egală cu raportul dintre numărul evenimentelor egal probabile favorabile evenimentului A şi numărul total al evenimentelor egal probabile.*

Altă formulare: probabilitatea unui eveniment este raportul între numărul cazurilor favorabile evenimentului şi numărul cazurilor posibile.

Observația 1.4.6.

1) *Conform acestei definiții nu putem stabili probabilitatea unui eveniment ce aparține unui câmp infinit de evenimente.*

2) *Definiția clasică se aplică numai atunci când evenimentele elementare sunt egal posibile.*

Exemplul 1.4.7. *Considerăm experiența de aruncare a unui zar. Evenimentele elementare sunt egal posibile şi avem 6 cazuri posibile. Notăm cu A evenimentul "apariția unei fețe cu număr par de puncte* ≤ 6 *" numărul cazurilor favorabile evenimentului A este 3. Deci* $\mathrm{P(A)} = \frac{3}{6} = \frac{1}{2}$.

Exemplul 1.4.8. *Dintr-o urnă cu 15 bile numerotate de la 1 la 15 se extrage o bilă la întâmplare. Se consideră evenimentele:*

A = obținerea unui număr prim;
B = obținerea unui număr par;
C =obținerea unui număr divizibil prin 3.

Să calculăm probabilitățile acestor evenimente.

Rezolvare

În această experiență aleatoare numărul total al cazurilor posibile este 15.

Pentru A numărul cazurilor favorabile este 6, adică {2, 3, 5, 7, 11, 13}, deci $\mathrm{P(A)} = \frac{6}{15} = \frac{2}{5}$.

Pentru B numărul cazurilor favorabile este 7, adică {2, 4, 6, 8, 10, 12, 14}, deci $\mathrm{P(B)} = \frac{7}{15}$.

Pentru C, numărul cazurilor favorabile este 5, adică { 3, 6, 9, 12, 15}, deci $\mathrm{P(C)} = \frac{5}{15} = \frac{1}{3}$.

1.5. Reguli de calcul cu probabilități

P_1) *Probabilitatea diferenței: Dacă* $\mathrm{A, B} \in \mathrm{K}$ *şi* $\mathrm{A} \subset \mathrm{B}$ *atunci*

P(B-A)=P(B)-P(A)

Demonstrație

Din relațiile B = A ∪ (B - A) şi A ∩ (B - A) = Ø aplicând axioma iii) avem $P(B) = P\left[A \cup (B - A)\right] = P(A) + P(B - A)$

P_2) *Probabilitatea reunirii (formula lui Poincaré):*

Dacă $\mathrm{A, B} \in \mathrm{K}$ *atunci* $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Demonstrație

Din relațiile $A \cup B = A \cup \left[B - (A \cap B)\right]$ şi $A \cap \left[B - (A \cap B)\right] = \text{Ø}$ aplicând axioma iii) avem

$$P(A \cup B) = P(A) + P\left[B - (A \cap B)\right] = P(A) + P(B) - P(A \cap B)$$

dacă s-a folosit P_1.

Generalizare:

Dacă $A_1, A_2, \ldots A_n$ sunt evenimente compatibile atunci

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

P_3) *Probabilităţi condiţionate: Dacă* $\mathrm{P(B)} \neq 0$ *atunci raportul* $\dfrac{\mathrm{P(A \cap B)}}{\mathrm{P(B)}}$ *îl numim probabilitatea lui A condiţionată de B şi notăm* $P_B(A)$ *sau* $\mathrm{P(A|B)}$.

Demonstraţie

Arătăm că $P_B(A)$ satisface axiomele probabilităţii:

i) $P_B(A) \geq 0$ deoarece $P(A \cap B) \geq 0$ şi $P(B) > 0$

ii) $P_B(E) = \dfrac{P(B \cap E)}{P(B)} = \dfrac{P(B)}{P(B)} = 1$

iii) Fie A_1 şi $A_2 \in K$ şi $A_1 \cap A_2 = \emptyset$. Avem

$$P_B(A_1 \cup A_2) = \frac{P\big(B \cap (A_1 \cup A_2)\big)}{P(B)} = \frac{P\big[(B \cap A_1) \cup (B \cap A_2)\big]}{P(B)} =$$

$$= \frac{P(B \cap A_1) + P(B \cap A_2)}{P(B)} = \frac{P(B \cap A_1)}{P(B)} + \frac{P(B \cap A_2)}{P(B)} = P_B(A_1) + P_B(A_2),$$

dacă $(B \cap A_1) \cap (B \cap A_2) = \emptyset$.

Observaţia 1.5.1.

1) *Oricărui câmp de evenimente* (Ω, K) *îi putem ataşa un câmp de probabilitate condiţionat* $\{\Omega, K, P_B\}$.

2) $\mathrm{P(A \cap B) = P(B) \cdot P_B(A)}$ - *formula de calcul a intersecţiei a două evenimente dependente. Are loc o generalizare: dacă* A_1, A_2, $\ldots A_n$ *sunt evenimente dependente atunci*

$$\mathrm{P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1) \cdot P_{A_1}(A_2) \cdot P_{A_1 \cap A_2}(A_3) \ldots P_{\bigcap_{i=1}^{n-1} A_i}(A_n).}$$

3) *Dacă evenimentele A şi B sunt independente atunci* $P_B(A) = P(A)$ *şi* $\mathrm{P(A \cap B) = P(A) \cdot P(B)}$ - *formula de calcul a intersecţiei a două evenimente independente.*

Generalizare:

Dacă A_1, A_2, $\ldots A_n$ sunt evenimente independente atunci

$$\mathrm{P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i).}$$

4) Dacă evenimentele A şi B se condiţionează reciproc şi $P(A) \neq 0, P(B) \neq 0$ atunci $P(A) \cdot P_A(B) = P(B) \cdot P_B(A)$.

P_4) *Probabilitatea reunirii evenimentelor independente. Dacă A_1, A_2, …A_n sunt evenimente independente, atunci:* $P\left(\bigcup_{i=1}^{n} A_i\right) = 1 - \prod_{i=1}^{n} (1 - P(A_i))$

Demonstraţie

Folosind relaţiile lui De Morgan $\overline{\bigcup_{i=1}^{n} A_i} = \bigcap_{i=1}^{n} \overline{A_i}$ şi faptul că A_i sunt evenimente independente implică

$$P\left(\bigcup_{i=1}^{n} A_i\right) = 1 - P\left(\overline{\bigcup_{i=1}^{n} A_i}\right) = 1 - P\left(\bigcap_{i=1}^{n} \overline{A_i}\right) = 1 - \prod_{i=1}^{n} P(\overline{A_i}) = 1 - \prod_{i=1}^{n} (1 - P(A_i))$$

P_5) *Inegalitatea lui Boole: A_1, A_2, …A_n, sunt evenimente dependente atunci* $P\left(\bigcap_{i=1}^{n} Ai\right) \geq \sum_{i=1}^{n} P(A_i) - (n-1) = 1 - \sum_{i=1}^{n} P(\overline{A_i})$

Demonstraţie

Verificăm inegalitatea din enunţ prin inducţie matematică.

Pentru n = 2 avem $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$ dacă $P(A_1 \cup A_2) \leq 1$ şi rezultă $P(A_1 \cap A_2) \geq P(A_1) + P(A_2) - 1$ relaţia este adevărată.

Presupunem inegalitatea adevărată pentru n-1 adică

$P\left(\bigcap_{i=1}^{n-1} A_i\right) \geq \sum_{i=1}^{n-1} P(A_i) - (n-2)$ şi demonstrăm pentru n.

Avem succesiv

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P\left[\left(\bigcap_{i=1}^{n-1} A_i\right) \cap A_n\right] \geq P\left(\bigcap_{i=1}^{n-1} A_i\right) + P(A_n) - 1 \geq$$

$$\geq \sum_{i=1}^{n-1} P(A_i) - (n-2) + P(A_n) - 1 = \sum_{i=1}^{n} P(A_i) - (n-1)$$

dacă s-a ţinut seama de ipoteza de inducţie.

P_6) *Formula probabilităţii totale: Dacă A_1,A_2, …A_n este un sistem complet de evenimente şi $X \in K$ atunci $P(X) = \sum_{i=1}^{n} P(A_i) \cdot P_{A_i}(X)$.*

Demonstrație

Din ipoteza că A_i, $i=\overline{1,n}$ este un sistem complet de evenimente rezultă că $X=(A_1\cap X)\cup(A_2\cap X)\cup\ldots\cup(A_n\cap X)$

Deoarece $A_i\cap A_j=\emptyset,\ i\neq j,\ i,j=\overline{1,n}$ avem că

$(X\cap A_i)\cap(X\cap A_j)=\emptyset,\ i\neq j,\ i,j=\overline{1,n}$

Avem succesiv

$$P(X)=P\left(\bigcup_{i=1}^{n}(A_i\cap X)\right)=\sum_{i=1}^{n}P(A_i\cap X)=\sum_{i=1}^{n}P(A_i)\cdot P_{A_i}(X)$$

P7) *Formula lui Bayes: Dacă A_1, A_2, …A_n este un sistem complet de evenimente al câmpului (Ω, K) şi* $X\in K$ *atunci:*

$$P_X(A_i)=\frac{P(A_i)\cdot P_{A_i}(X)}{\sum_{i=1}^{n}P(A_i)\cdot P_{A_i}(X)},\ i=\overline{1,n}$$

Demonstrație

Deoarece $P(X\cap A_i)=P(X)\cdot P_X(A_i)$ şi $P(X\cap A_i)=P(A_i)\cdot P_{A_i}(X)$ avem $P(X)\cdot P_X(A_i)=P(A_i)\cdot P_{A_i}(X)$, deci $P_X(A_i)=\frac{P(A_i)\cdot P_{A_i}(X)}{P(X)}=\frac{P(A_i)\cdot P_{A_i}(X)}{\sum_{i=1}^{n}P(A_i)\cdot P_{A_i}(X)}$ dacă s-a folosit formula probabilității totale.

Exemplul 1.5.2. *Cele 26 de litere ale alfabetului, scrise fiecare pe un cartonaş, sunt introduse într-o urnă. Se cere probabilitatea ca extrăgând la întâmplare de 5 ori câte un cartonaş şi aşezându-le în ordinea extragerii să obținem cuvântul LUCIA.*

Rezolvare

Notăm prin X evenimentul căutat, deci de a obține prin extrageri succesive cuvântul LUCIA, de asemenea notăm prin A_1 = evenimentul ca la prima extragere să obținem litera L; A_2 = evenimentul ca la a doua extragere să obținem litera U; A_3 = evenimentul ca la a treia extragere să obținem litera C; A_4 = evenimentul ca la a patra extragere să obținem litera I; A_5 = evenimentul ca la a cincea extragere să obținem litera A.

Atunci evenimentul X are loc dacă avem

$X=A_1\cap A_2\cap A_3\cap A_4\cap A_5$.

Rezultă:

$$P(X) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot P(A_4|A_1 \cap A_2 \cap A_3) \cdot$$
$$\cdot P(A_5|A_1 \cap A_2 \cap A_3 \cap A_4) = \frac{1}{26} \cdot \frac{1}{25} \cdot \frac{1}{24} \cdot \frac{1}{23} \cdot \frac{1}{22}.$$

Exemplul 1.5.3. *Dacă probabilitatea ca un automobil să plece în cursă într-o dimineață friguroasă este de 0,6 şi dispunem de două automobile de acest fel, care este probabilitatea ca cel puțin unul din automobile să plece în cursă într-o dimineață friguroasă?*

Rezolvare

Dacă notăm prin A_1 şi A_2 evenimentele ca primul respectiv, al doilea automobil să plece în cursă şi prin X evenimentul căutat, deci ca cel puțin unul dintre automobile să plece în cursă, avem: $X = A_1 \cup A_2$, iar $P(X) = P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$, deoarece evenimentele A_1 şi A_2 sunt compatibile (cele două automobile pot să plece în cursă deodată). Cum $P(A_1) = P(A_2) = 0,6$, iar evenimentele A_1 şi A_2 sunt independente între ele (plecarea unui automobil nu depinde de plecarea sau neplecarea celuilalt), deci $P(A_1 \cap A_2) = P(A_1)P(A_2) = (0,6)^2$. Se obține că $P(X) = 0,6 + 0,6 - (0,6)^2 = 0,84$.

Exemplul 1.5.4. *Trei secții ale unei întreprinderi* S_1, S_2, S_3 *depăşesc planul zilnic de producție cu probabilitățile de respectiv 0,7; 0,8 şi 0,6. Să se calculeze probabilitățile evenimentelor:*

A - cel puțin o secție să depăşească planul de producție.

B - toate secțiile să depăşească planul de producție.

Rezolvare

Fie A_i evenimentul ca secția S_i să depăşească planul de producție.

Avem: $A = A_1 \cup A_2 \cup A_3$, deci

$P(A) = P(A_1 \cup A_2 \cup A_3) = 1 - P(\overline{A}_1 \cap \overline{A}_2 \cap \overline{A}_3) = 1 - P(\overline{A}_1) \cdot P(\overline{A}_2) \cdot P(\overline{A}_3) =$ $1 - (1-0,7)(1-0,8)(1-0,6) = 1 - 0,3 \cdot 0,2 \cdot 0,4 = 0,976$.

$B = A_1 \cap A_2 \cap A_3$ şi ținând seama de independența evenimentelor, avem:

$P(B) = P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2) \cdot P(A_3) = 0,7 \cdot 0,8 \cdot 0,6 = 0,336$.

Exemplul 1.5.5. *O presă este considerată că satisface standardul de fabricație dacă trei caracteristici sunt satisfăcute. Dacă aceste caracteristici A, B şi C sunt satisfăcute cu probabilitățile* $P(A) = \frac{9}{10}$, $P(B) = \frac{7}{11}$ *şi* $P(C) = \frac{11}{12}$, *atunci probabilitatea ca să fie satisfăcute toate trei caracteristicile se poate evalua cu formula lui Boole. Astfel se poate scrie:*

$P(\mathrm{A}\cap\mathrm{B}\cap\mathrm{C})\geq 1-\left[\mathrm{P}(\overline{\mathrm{A}})+\mathrm{P}(\overline{\mathrm{B}})+\mathrm{P}(\overline{\mathrm{C}})\right]$, *adică*

$$P(\mathrm{A}\cap\mathrm{B}\cap\mathrm{C})\geq 1-\left(\frac{1}{10}+\frac{4}{11}+\frac{1}{12}\right)=\frac{229}{660}.$$

Exemplul 1.5.6. *Un sortiment de marfă dintr-o unitate comercială provine de la trei fabrici diferite în proporţii, respectiv* $\frac{1}{3}$ *de la prima fabrică,* $\frac{1}{6}$ *de la a doua fabrică şi restul de la fabrica a treia. Produsele de la cele trei fabrici satisfac standardele de fabricaţie în proporţie de 90%, 95% şi respectiv 92%. Un client ia la întâmplare o bucată din sortimentul de marfă respectiv.*

a) Care este probabilitatea ca produsul să satisfacă standardele de fabricaţie?

b) Care este probabilitatea ca produsul să fie defect şi să provină de la prima fabrică?

Rezolvare

a) Notăm cu $\mathrm{A}_1, \mathrm{A}_2$ şi A_3 evenimentele ca produsul cumpărat să fie de la prima, a doua, respectiv a treia fabrică. Aceste trei evenimente formează un sistem complet de evenimente şi au probabilităţile $\mathrm{P}(\mathrm{A}_1)=\frac{1}{3}, \mathrm{P}(\mathrm{A}_2)=\frac{1}{6}$ şi $P(A_3)=\frac{1}{2}$. Dacă A este evenimentul că produsul cumpărat de client satisface standardele de fabricaţie, atunci $\mathrm{P}(\mathrm{A}|\mathrm{A}_1)=0{,}90$, $\mathrm{P}(\mathrm{A}|\mathrm{A}_2)=0{,}95$ şi $\mathrm{P}(\mathrm{A}|\mathrm{A}_3)=0{,}92$. Folosind formula probabilităţii totale se obţine:

$$P(A)=P(A_1)\cdot P(A|A_1)+P(A_2)\cdot P(A|A_2)+P(A_3)\cdot P(A|A_3)=$$

$$=\frac{1}{9}\cdot 0{,}90+\frac{1}{6}\cdot 0{,}95+\frac{1}{2}0{,}92=\frac{5{,}51}{6}=0{,}918$$

b) Folosind formula lui Bayes, avem:

$$\mathrm{P}(\mathrm{A}_1|\overline{\mathrm{A}})=\frac{\mathrm{P}(\mathrm{A}_1)\mathrm{P}(\overline{\mathrm{A}}|\mathrm{A}_1)}{\mathrm{P}(\mathrm{A}_1)\mathrm{P}(\overline{\mathrm{A}}|\mathrm{A}_1)+\mathrm{P}(\mathrm{A}_2)\mathrm{P}(\overline{\mathrm{A}}|\mathrm{A}_2)+\mathrm{P}(\mathrm{A}_3)\mathrm{P}(\overline{\mathrm{A}}|\mathrm{A}_3)}=$$

$$=\frac{\frac{1}{3}\cdot 0{,}10}{\frac{1}{3}\cdot 0{,}10+\frac{1}{6}\cdot 0{,}05+\frac{1}{2}\cdot 0{,}08}=\frac{0{,}2}{0{,}49}=0{,}408.$$

Exemplul 1.5.7. *Un student solicită o bursă de studii la 3 universităţi. După trimiterea actelor necesare, acesta poate obţine bursă de la universitatea i* (U_i) *sau nu* ($\overline{U_i}$), $1\leq i\leq 3$. *Scrieţi evenimentele ce corespund următoarelor situaţii :*

a) primeşte o bursă;

b) primeşte cel mult o bursă;
c) primeşte cel puţin o bursă;
d) primeşte cel puţin două burse.

Rezolvare

a) Bursa primită poate fi de la prima universitate, caz în care celelalte nu-i acordă bursă, sau de la a doua, caz în care prima şi a treia nu-i acordă bursă, sau de la a treia, caz în care primele două nu-i acordă bursă. Avem astfel evenimentul

$$A = (U_1 \cap \overline{U_2} \cap \overline{U_3}) \cup (\overline{U_1} \cap U_2 \cap \overline{U_3}) \cup (\overline{U_1} \cap \overline{U_2} \cap U_3) .$$

b) Avem două variante : studentul nu primeşte nici o bursă sau studentul primeşte o bursă. Obţinem evenimentul

$$B = (\overline{U_1} \cap \overline{U_2} \cap \overline{U_3}) \cup A .$$

c) Evenimentul poate fi scris ca reuniunea a trei evenimente : studentul primeşte o bursă, două burse, trei burse. Astfel $C = A \cup E \cup F$, unde $E = (U_1 \cap U_2 \cap \overline{U_3}) \cup (\overline{U_1} \cap U_2 \cap U_3) \cup (U_1 \cap \overline{U_2} \cap U_3)$,

iar $F = U_1 \cap U_2 \cap U_3$.

d) Avem $D = E \cup F$. Altfel, evenimentul D este contrar evenimentului B, deci $D = \overline{B} = \overline{(\overline{U_1} \cap \overline{U_2} \cap \overline{U_3}) \cup A}$.

Exemplul 1.5.8. *Într-un grup de studenţi aflaţi în excursie se găsesc 6 fete şi 9 băieţi. Se aleg la întâmplare doi studenţi pentru a cerceta traseul. Care este probabilitatea ca cei doi să fie :*

a) băieţi;
b) fete;
c) un băiat şi o fată;
d) cel puţin un băiat;
e) primul băiat şi a doua fată;
f) de acelaşi sex.

Rezolvare

Notăm cu A_1 şi A_2 evenimentele alegerii unui băiat la prima, respectiv a doua alegere. La primul punct avem de calculat probabilitatea $P(A_1 \cap A_2)$. Întrucât a doua alegere depinde de prima avem :

$$P(A_1 \cap A_2) = P(A_1)P(A_2 \mid A_1) = \frac{9}{15} \cdot \frac{8}{14} = \frac{12}{35} ,$$

deoarece alegând un băiat mai rămân în grup 14 studenţi între care 8 băieţi. Evenimentul de la punctul b) se scrie astfel : $B = \overline{A_1} \cap \overline{A_2}$. Deci

$$P(B) = P(\overline{A_1} \cap \overline{A_2}) = P(\overline{A_1})P(\overline{A_2} \mid \overline{A_1}) = \frac{6}{15} \cdot \frac{5}{14} = \frac{1}{7} .$$

Evenimentul de la punctul c) este $C=(A_1\cap\overline{A_2})\cup(A_2\cap\overline{A_1})$ aşadar $P(C)=P(A_1\cap\overline{A_2})+P(A_2\cap\overline{A_1})$, ($A_1\cap\overline{A_2}, A_2\cap\overline{A_1}$ sunt incompatibile)

Dar $P(A_1\cap\overline{A_2})=P(A_1)P(\overline{A_2}/A_1)=\frac{9}{15}\cdot\frac{6}{14}$,

iar $P(A_2\cap\overline{A_1})=P(\overline{A_1})P(A_2/\overline{A_1})=\frac{6}{15}\cdot\frac{9}{14}$

de unde $P(C)=2\cdot\frac{9}{15}\cdot\frac{6}{14}=\frac{18}{35}$.

Am obținut şi probabilitatea evenimentului de la punctul e) $P(A_1\cap\overline{A_2})$. Evenimentul de la punctul d) se exprimă astfel : $D=A_1\cup A_2$.

El este contrar evenimentului : $B=\overline{A_1}\cap\overline{A_2}$, prin urmare

$P(D)=1-P(B)=1-\frac{1}{7}=\frac{6}{7}$. Evenimentul de la ultimul punct f) este

$F=(A_1\cap A_2)\cup(\overline{A_1}\cap\overline{A_2})$. Cum $(A_1\cap A_2)\cap(\overline{A_1}\cap\overline{A_2})=\Phi$ cele două evenimente sunt incompatibile şi deci

$$P(F)=P(A_1\cap A_2)+P(\overline{A_1}\cap\overline{A_2})=\frac{12}{35}+\frac{1}{7}=\frac{17}{35}.$$

Exemplul 1.5.9. *La un examen de licență participă mai mulți absolvenți, între care numai trei din străinătate. Probabilitatea ca primul student să promoveze este ¾, probabilitatea ca al doilea să promoveze este 4/5, iar pentru al treilea 5/6. Să se determine probabilitățile ca :*

a) toți cei trei studenți să promoveze;

b) cel puțin unul să promoveze examenul.

Rezolvare

Fie A_i evenimentul promovării examenului de către studentul i, i=1,2,3. Evenimentul de la punctul a) este $A=A_1\cap A_2\cap A_3$, iar de la punctul b) este $B=A_1\cup A_2\cup A_3$. Evenimentele A_i sunt independente (rezultatele celor 3 studenți nedepinzând unul de celelalte), deci

$$P(A)=P(A_1)P(A_2)P(A_3)=\frac{3}{4}\cdot\frac{4}{5}\cdot\frac{5}{6}=\frac{1}{2}.$$

Folosind proprietățile probabilității avem :

$$P(B)=P(A_1\cup A_2\cup A_3)=P(A_1\cup A_2)+P(A_3)-P((A_1\cup A_2)\cap A_3)=$$
$$=P(A_1)+P(A_2)-P(A_1\cap A_2)+P(A_3)-P((A_1\cap A_3)\cup(A_2\cap A_3))=$$

$$= P(A_1) + P(A_2) - P(A_1 \cap A_2) + P(A_3) - [P(A_1 \cap A_3) + P(A_2 \cap A_3) -$$
$$- P((A_1 \cap A_3) \cap (A_2 \cap A_3)) = P(A_1) + P(A_2) + P(A_3) -$$
$$- P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3).$$

Ținând seama de independența evenimentelor A_i , i=1,2,3, avem:

$$P(B) = P(A_1) + P(A_2) + P(A_3) - P(A_1)P(A_2) - P(A_1)P(A_3) - P(A_2)P(A_3) +$$
$$+ P(A_1)P(A_2)P(A_3) = \frac{3}{4} + \frac{4}{5} + \frac{5}{6} - \frac{3}{4} \cdot \frac{4}{5} - \frac{3}{4} \cdot \frac{5}{6} - \frac{4}{5} \cdot \frac{5}{6} + \frac{3}{4} \cdot \frac{4}{5} \cdot \frac{5}{6} = \frac{119}{120}.$$

Exemplul 1.5.10. *Din mai multe controale asupra activităților a trei magazine se apreciază că în proporție de 90%, 80%, 70%, cele trei magazine au declarat marfa vândută. La un nou control, comisia de control solicită 50 de documente privind activitatea comercială: 20 de la primul magazin, 15 de la al doilea, 15 de la al treilea. Dintre acestea se alege unul la întâmplare pentru a fi verificat:*

a) Cu ce probabilitate documentul ales este corect (înregistrat)?

b) Constatând că este corect, cu ce probabilitate el aparține primului magazin?

Rezolvare

a) Notăm cu A_1, A_2, A_3 evenimentul ca documentul controlat să provină de la primul, al doilea şi respectiv al treilea magazin. Avem astfel

$$P(A_1) = \frac{20}{50}; P(A_2) = \frac{15}{50}; P(A_3) = \frac{15}{50} \ .$$

Fie A evenimentul ca documentul controlat să fie corect. Atunci A / A_1, A / A_2, A / A_3 reprezintă evenimentul ca documentul controlat să fie corect ştiind că el provine de la primul, al doilea, al treilea magazin. Prin urmare : $P(A/A_1)=0,90$; $P(A/A_2)=0,80$; $P(A/A_3)=0,70$. Cum $\{A_1, A_2, A_3\}$ este un sistem complet de evenimente

$$A_1 \cup A_2 \cup A_3 = E, A_1 \cap A_2 = A_1 \cap A_3 = A_2 \cap A_3 = \Phi$$

aplicând formula probabilității totale avem :

$$P(A) = P(A_1)P(A/A_1) + P(A_2)P(A/A_2) + P(A_3)P(A/A_3) =$$
$$= \frac{20}{50} \cdot 0,90 + \frac{15}{50} \cdot 0,80 + \frac{15}{50} \cdot 0,70 = 0,81.$$

b) Aplicând formula lui Bayes avem :

$$P(A_1 / A) = \frac{P(A_1)P(A/A_1)}{\sum_{i=1}^{3} P(A_i)P(A/A_i)} = \frac{\frac{20}{50} \cdot 0{,}90}{0{,}81} = \frac{0{,}36}{0{,}81} = \frac{4}{9} \ .$$

(A_1/A reprezintă evenimentul ca documentul controlat să provină de la primul magazin ştiind că a fost corect).

Capitolul 2

Variabile aleatoare

Variabila aleatoare este una din noţiunile fundamentale ale teoriei probabilităţilor şi a statisticii matematice. În cadrul unei cercetări experimentale se constată că între valorile numerice măsurate există diferenţe chiar dacă rămân neschimbate condiţiile de desfăşurare ale experimentului.

Dacă ne referim la o singură măsurătoare, variabila aleatoare este acea mărime care ăn cadrul unui experiment poate lua o valoare necunoscută aprioric. Pentru un şir de măsurători, variabila aleatoare este o noţiune care-l caracterizează din două puncte de vedere:

- caracterizare din punct de vedere cantitativ – variabila ne dă informaţii privind valoarea numerică a mărimii măsurate;
- caracterizare din punct de vedere calitativ – variabila aleatoare ne dă informaţii privind frecvenţa de apariţie a unei valori numerice într-un şir.

Dacă valorile numerice ale unui şir de date aparţin mulţimii numerelor întregi sau raţionale atunci se defineşte o variabilă aleatoare discretă, iar în cazul aparteneţei valorilor la mulţimea numerelor reale se defineşte o variabilă aleatoare continuă.

2.1. Variabile aleatoare discrete

În ciuda faptului că după repetarea unui experiment de un număr mare de ori intervine o anumită regularitate în privinţa apariţiei unor rezultate ale acestuia, nu se poate preciza niciodată cu certitudine care anume dintre rezultate va apare într-o anumită probă. Din acest motiv cuvântul sau conceptul „aleator” trebuie înţeles sau gândit în sensul că avem de-a face cu experimente sau fenomene care sunt guvernate de legi statistice (atunci când există un anumit grad de incertitudine privind apariţia unui rezultat sau reapariţia lui) şi nu de legi deterministe (când ştim cu certitudine ce rezultat va apare sau nu). Pentru ca astfel de experimente sau fenomene să fie cunoscute şi prin urmare studiate, sunt importante şi necesare două lucruri şi anume:

1. rezultatele posibile ale experimentului, care pot constitui o mulţime finită, infinită sau numărabilă sau infinită şi nenumărabilă;
2. legea statistică sau probabilităţile cu care este posibilă apariţia rezultatelor experimentului considerat.

În linii mari şi într-un înţeles mai larg, o mărime care ia valori la întâmplare sau aleatoriu dintr-o mulţime oarecare posibilă se numeşte variabilă aleatoare (sau întâmplătoare). Se poate da şi o definiţie riguroasă.

Definiția 2.1.1. *Fie câmpul de probabilitate {Ω, K, P}. Numim variabilă aleatoare de tip discret o aplicație X : Ω → R care verifică condițiile:*

i) are o mulțime cel mult numărabilă de valori;

ii) $\forall x \in R \ (X = x) \in K$

Observația 2.1.2.

1) Dacă K = P(Ω) atunci ii) este automat îndeplinită;

2) O variabilă aleatoare de tip discret este deci o funcție univocă de forma

X : Ω → {x1, x2, ... xn, ...} ⊂ R;

3) Se obişnuieşte ca valorile variabilei să se noteze în ordine crescătoare adică $x_1 < x_2 < x_3 ... < x_n < ....$, *xi ∈ R, i = 1, 2, …*

4) Evenimentele Ai = X-1(xi) = {ω ∈ Ω / X(ω) = xi } ∈ K, oricare ar fi i = 1, 2, 3, …,

X-1 : {x1, x2, … xn, …} → K este inversa funcției X.

Definiția 2.1.3. *Numim distribuția sau repartiția variabilei aleatoare X de tip discret, tabloul de forma* $X : \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I}$ *unde xi,* $i \in I$, *sunt valorile posibile ale variabilei aleatoare X iar pi este probabilitatea cu care variabila considerată X ia valoarea xi , adică pi = P(X = xi),* $i \in I$ *mulțimea I putând fi finită sau cel mult numărabilă.*

Observația 2.1.4.

1) Evenimentele (X = xi), i ∈ I formează un sistem complet de evenimente şi $\sum_{i \in I} p_i = 1$.

2) Variabila aleatoare pentru care mulțimea valorilor este un interval finit sau infinit pe axa numerelor reale este variabilă aleatoare continuă.

3) Forma cea mai generală a unei variabile aleatoare aparținând unei clase de variabile aleatoare de tip discret se numeşte lege de probabilitate discretă.

Definiția 2.1.5. *Spunem că variabilele aleatoare X şi Y care au respectiv distribuțiile* $X\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I}$ *şi* $Y\begin{pmatrix} y_j \\ q_j \end{pmatrix}_{j \in J}$ *sunt independente dacă*

$P(X = xi , Y = yj) = P(X = xi) P(Y = yj), \ \forall (i, j) \in I \times J.$

Definiția 2.1.6. *Fie variabilele aleatoare X, Y care au respectiv distribuțiile* $X\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I}$ *şi* $Y\begin{pmatrix} y_j \\ q_j \end{pmatrix}_{j\in J}$ *atunci variabila aleatoare sumă X+Y, produs* $\mathrm{X}\cdot\mathrm{Y}$ *şi cât* $\frac{X}{Y}$ *(dacă* $y_j \neq 0, \forall j \in J$*) vor avea distribuțiile* $X+Y\begin{pmatrix} x_i + y_j \\ p_{ij} \end{pmatrix}_{(i,j)\in IxJ}$, $\mathrm{X}\cdot\mathrm{Y}\begin{pmatrix} \mathrm{x_i y_j} \\ \mathrm{p_{ij}} \end{pmatrix}_{(\mathrm{i,j})\in \mathrm{IxJ}}$, *respectiv* $\frac{X}{Y}\begin{pmatrix} \frac{x_i}{y_j} \\ p_{ij} \end{pmatrix}_{(i,j)\in IxJ}$ *unde pij = P(X = xi, Y = yj)* $(i,j)\in \mathrm{IxJ}$.

Definiția 2.1.7. *Se numeşte*

a) *produs al variabilei aleatoare X prin constanta reală a, variabila aleatoare notată prin* $aX:\begin{pmatrix} ax_i \\ p_i \end{pmatrix}_{i\in I}$

b) *sumă a variabilei aleatoare X cu constanta reală a, variabila aleatoare notată prin* $a+X:\begin{pmatrix} a+x_i \\ p_i \end{pmatrix}_{i\in I}$

c) *putere a variabilei aleatoare X de exponent k,* $k \in Z$, *variabila aleatoare* $X^k:\begin{pmatrix} x_i^k \\ p_i \end{pmatrix}_{i\in I}$ *cu condiția ca operațiile* x_i^k, $i \in I$, *să aibă sens.*

Observația 2.1.8. *Au loc relațiile* $\sum_{j\in J} p_{ij} = p_i, \forall i \in I$ *şi* $\sum_{i\in I} p_{ij} = q_j, \forall j \in J$.

Dacă variabilele X,Y sunt independente atunci $p_{ij} = p_i q_j, \forall (i,j) \in I \times J$

Definiția 2.1.9. *Fie* $\{\Omega, K, P\}$ *un câmp de probabilitate, iar* $X: \Omega \to R$ *o variabilă aleatoare. Numim funcție de repartiție ataşată variabilei aleatoare X funcția* $F: R \to [0, 1]$, *definită prin* $F(x) = P(X \le x)$, $\forall x \in R$, *adică* $F(x) = \sum_{x_i \le x} p_i, x \in R$.

Dacă nu există pericol de confuzie, funcția de repartiție a variabilei aleatoare X se notează prin F.

Propoziția 2.1.10. (proprietăți ale funcției de repartiție)

1. $\forall \mathrm{a,b} \in \mathrm{R},\ \mathrm{a} < \mathrm{b}$ *avem:*

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

Demonstrație

Avem succesiv

$$P(a \le X < b) = P(X < b, \overline{X < a}) = P[(X < b) - (X < a)] =$$
$$= P(X < b) - P(X < a) = F(b) - P(X = b) - F(a) + P(X = a)$$

dacă s-a ținut seama de relația $(X < a) \subset (X < b)$ şi s-a folosit probabilitatea diferenței.

$$P(a < X < b) = P[(a \le X < b) - (X = a)] = P(a \le X < b) - P(X = a) =$$
$$= F(b) - P(X = b) - F(a) + P(X = a) - P(X = a) = F(b) - P(X = b) - F(a)$$

dacă s-a folosit relația demonstrată anterior.

2. F este nedescrescătoare pe **R**,

adică $\forall\ x_1, x_2 \in R,\ x_1 < x_2 \Rightarrow F(x_1) \le F(x_2)$

Demonstrație

$$0 \le P(x_1 < X \le x_2) = F(x_2) - F(x_1) \Rightarrow F(x_1) \le F(x_2)$$

3. $\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1$

Demonstrație

$$\lim_{x \to -\infty} F(x) = \lim_{x \to -\infty} P(X \le x) = P(\phi) = 0$$
$$\lim_{x \to +\infty} F(x) = \lim_{x \to +\infty} P(X \le x) = P(E) = 1$$

4. $\forall x \in \mathrm{R}, \mathrm{F}(x - a) = \mathrm{F}(x)$ (F este continuă la stânga în fiecare punct $x \in R$)

Exemplul 2.1.11. *Se consideră variabila aleatoare discretă*

$\mathrm{X}: \begin{pmatrix} 1 & 2 & 3 & 4 \\ \mathrm{p}^2 & \frac{7}{4}\mathrm{p} & \frac{1}{3} & \frac{1}{6} \end{pmatrix}$. *Care este probabilitatea ca X să ia o valoare mai mică sau egală cu 3?*

Rezolvare

Pentru ca X să fie o variabilă aleatoare trebuie ca $\mathrm{p} \ge 0$ şi $p^2 + \frac{7}{4} \cdot p + \frac{1}{3} + \frac{1}{6} = 1$. Se obține soluția acceptabilă $p = \frac{1}{4}$. Se calculează

probabilitatea cerută prin intermediul evenimentului contrar şi anume $P(X \le 3) = 1 - P(X = 4) = 1 - \frac{1}{6} = \frac{5}{6}$ sau

$$P(X \le 3) = P(X = 1) + P(X = 2) + P(X = 3) = \frac{1}{16} + \frac{7}{16} + \frac{1}{3} = \frac{5}{6}.$$

Exemplul 2.1.12. *Se dau variabilele aleatoare independente:*

$$X: \begin{pmatrix} -1 & 0 & 1 \\ p + \frac{1}{6} & q + \frac{1}{3} & \frac{1}{3} \end{pmatrix}; \; Y: \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{3} & 2p - q & 12p^2 \end{pmatrix}.$$

a) Să se scrie distribuţia variabilei 2XY.

b) Pentru ce valori ale lui c avem: $P(X + Y = c) > \frac{2}{9}$?

Rezolvare

Pentru ca X şi Y să fie variabile aleatoare se impun condiţiile:

$p + \frac{1}{6} \ge 0; q + \frac{1}{3} \ge 0; 2p - q \ge 0$ şi apoi : $\begin{cases} p + \frac{1}{6} + q + \frac{1}{3} + \frac{1}{3} = 1 \\ \frac{1}{3} + 2p - q + 12p^2 = 1 \end{cases}$, rezultă valorile acceptabile $p = \frac{1}{6}$ şi q = 0. Deci variabilele aleatoare au repartiţiile:

$X: \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}; \; Y: \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$. Avem :

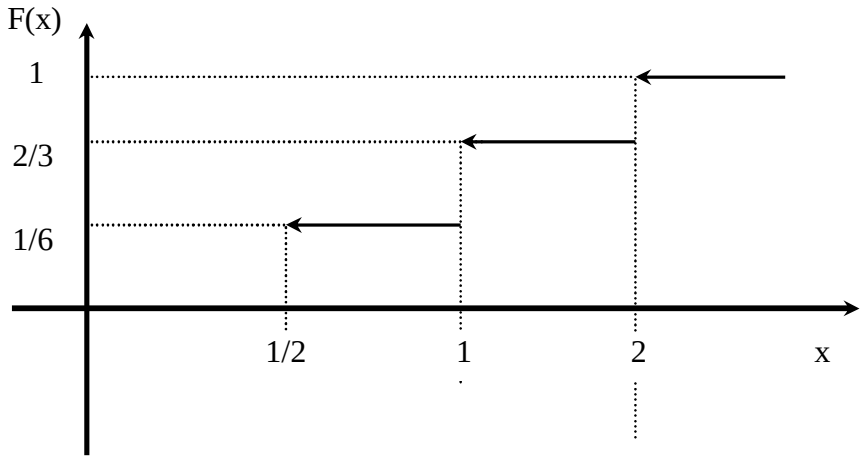
a) $2XY: \begin{pmatrix} -2 & 0 & 2 \\ \frac{2}{9} & \frac{5}{9} & \frac{2}{9} \end{pmatrix}$

b) $X + Y: \begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ \frac{1}{9} & \frac{2}{9} & \frac{3}{9} & \frac{2}{9} & \frac{1}{9} \end{pmatrix}$, deci $P(X + Y = c) > \frac{2}{9}$ corespunde situaţiei $P(X + Y = 0) = \frac{3}{9} > \frac{2}{9}$ adică c = 0.

Exemplul 2.1.13. *Variabila aleatoare X cu distribuţia următoare:*

$X:\begin{pmatrix} \frac{1}{2} & 1 & 2 \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} \end{pmatrix}$, *are funcția de repartiție:* $F(x) = P(X \le x) = \begin{cases} 0, \text{dacă } x \le \frac{1}{2}, \\ \frac{1}{6}, \text{dacă } \frac{1}{2} < x \le 1, \\ \frac{2}{3}, \text{dacă } 1 < x \le 2, \\ 1, \text{dacă } x > 2 \end{cases}$

Graficul funcției de repartiție este:



2.2. Vector aleator bidimensional

Definiția 2.2.1. *Fie câmpul de probabilitate {Ω, K, P}. Spunem că U=(X,Y) este vector aleator bidimensional de tip discret dacă aplicația* $U: \Omega \to R^2$ *verifică condițiile:*

i) are o mulțime cel mult numărabilă de valori;

ii) $\forall (x, y) \in R^2, (X = x, Y = y) \in K$.

Definiția 2.2.2. *Numim distribuția sau repartiția vectorului aleator (X,Y) de tip discret tabloul:*

X \ Y	y_1	y_j .....
x_1	p_{11}	p_{1j} .....
⋮	⋮	⋮
x_i	p_{i1}	p_{ij} .....
⋮	⋮	⋮
....	.....	.....

unde (x_i, y_j) *sunt valorile pe care le ia vectorul aleator (X,Y), iar* $p_{ij} = P(X = x_i, Y = y_j)$.

Evident $\sum_{(i,j)\in I\times J} p_{ij} = 1.$

Definiția 2.2.3. *Numim funcție de repartiție ataşată vectorului aleator bidimensional funcția* $F: R^2 \to [0,1]$, *definită prin:*

$$F(x,y) = P(X \le x,\ Y \le y),\ \forall (x,y) \in R^2.$$

Propoziția 2.2.4.(proprietățile funcției de repartiție a unui vector aleator bidimensional de tip discret)

1. *dacă* $a<b$ *şi* $c<d$, *atunci*
   $P(a < X \le b, c < Y \le d) = F(b,d) - F(a,c).$
2. *F(x,y) este nedescrescătoare în raport cu fiecare argument.*
3. $\lim_{x\to-\infty} F(x,y) = \lim_{y\to-\infty} F(x,y) = \lim_{\substack{x\to-\infty\\ y\to-\infty}} F(x,y) = 0;\ \lim_{\substack{x\to\infty\\ y\to\infty}} F(x,y) = 1.$
4. *F(x,y) este continuă la stânga în raport cu fiecare argument.*

Observația 2.2.5. *Dacă (X,Y) are funcția de repartiție F, iar variabilele X şi Y au funcțiile de repartiție* F_X *şi respectiv* F_Y, *atunci:*

$$F_X(x) = \lim_{y\to\infty} F(x,y) \ \text{şi}\ F_Y(y) = \lim_{x\to\infty} F(x,y).$$

Exemplul 2.2.6. *Se consideră vectorul aleator discret (X,Y) cu repartiția dată în tabelul:*

X \ Y	2	6
1	0,20	0,10
3	0,05	0,15
4	0,45	0,05

a) să se determine repartiția variabilelor X,Y, X+Y;
b) să se stabilească dacă X şi Y sunt independente sau nu;
c) să se calculeze $F\left(\frac{7}{2},5\right).$

Rezolvare

a) Variabila X are repartiția:

X: $\begin{pmatrix} 1 & 3 & 4 \\ p_1 & p_2 & p_3 \end{pmatrix}$, unde $\begin{matrix} p_1 = p_{11} + p_{12} = 0{,}20 + 0{,}10 = 0{,}30 \\ p_2 = p_{21} + p_{22} = 0{,}05 + 0{,}15 = 0{,}20 \\ p_3 = p_{31} + p_{32} = 0{,}45 + 0{,}05 = 0{,}50 \end{matrix}$, adică

X: $\begin{pmatrix} 1 & 3 & 4 \\ 0{,}30 & 0{,}20 & 0{,}50 \end{pmatrix}$.

Analog, variabila Y are repartiția Y: $\begin{pmatrix} 2 & 6 \\ q_1 & q_2 \end{pmatrix}$, unde

$\begin{matrix} q_1 = p_{11} + p_{21} + p_{31} = 0{,}20 + 0{,}05 + 0{,}45 = 0{,}70 \\ q_2 = p_{12} + p_{22} + p_{32} = 0{,}10 + 0{,}15 + 0{,}05 = 0{,}30 \end{matrix}$, adică Y: $\begin{pmatrix} 2 & 6 \\ 0{,}70 & 0{,}30 \end{pmatrix}$.

Avem: X+Y: $\begin{pmatrix} 3 & 4 & 6 & 7 & 8 & 10 \\ 0{,}20 & 0{,}05 & 0{,}45 & 0{,}10 & 0{,}15 & 0{,}05 \end{pmatrix}$.

b) Pentru verificarea independenței variabilelor X,Y, efectuăm un control, de exemplu:

P(X=1) P(Y=2) = $0{,}30 \cdot 0{,}70 = 0{,}21$, iar P[(X=1)$\cap$(Y=2)] = $p_{11} = 0{,}20$. Cum $0{,}21 \neq 0{,}20$, deducem că X şi Y sunt dependente.

c) F($\frac{7}{2}$,5) = $P(X \le \frac{7}{2}, Y \le 5) = P[(X = 1, Y = 2) \cup (X = 3, Y = 2)] =$

=P(X=1,Y=2) +P(X=3,Y=2) = 0,20+0,05 = 0,25.

Definiția 2.2.7. *Fie variabila aleatoare X având funcția de repartiție* F, *vom spune că X este variabilă aleatoare de tip continuu dacă funcția de repartiție se poate reprezenta sub forma:*

$$F(x) = \int_{-\infty}^{x} \rho(t)dt, \quad \forall x \in R.$$

Funcția $\rho : R \to R$ *se numeşte densitate de probabilitate a variabilei aleatoare X.*

Propoziția 2.2.8. *Au loc afirmațiile:*

1) $\forall x \in R, \ \rho(x) \ge 0$.

2) $F'(x) = \rho(x)$ *a.p.t. pe* ***R****.*

3) $P(a \le X < b) = \int_a^b \rho(x)dx$.

4) $\int_{-\infty}^{\infty} \rho(x)dx = 1$.

Observația 2.2.9.

1. Pentru o variabilă de tip continuu P(X=a)= 0, deci $P(a \le X < b) = P(a < X < b) = P(a \le X \le b) = P(a < X < b) = \int_a^b \rho(x)dx$.

2. $\rho(x) = F'(x) = \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{P(x \le X < x+\Delta x)}{\Delta x}$, *deci când* Δx *este mic avem* $P(x \le X < x + \Delta x) \approx \rho(x) \cdot \Delta x$.

Definiția 2.2.10. *Fie vectorul aleator (X,Y) având funcția de repartiție F, spunem că (X,Y) este un vector aleator de tip continuu, dacă funcția de repartiție F se poate pune sub forma:*

$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} \rho(s,t)\ ds\ dt,\ \forall (x,y) \in R^2$, *iar funcția* $\rho : R^2 \to R$ *se numeşte densitate de probabilitate a vectorului aleator (X,Y).*

Observația 2.2.11. *Dacă* ρ *este densitate de probabilitate pentru (X,Y), iar* ρ_X *şi* ρ_Y *densități de probabilitate pentru X, respectiv Y au loc:*

1) $\rho(x,y) \ge 0\ , \forall (x,y) \in R^2$.

2) $\frac{\partial^2 F(x,y)}{\partial x \partial y} = \rho(x,y)$ *a.p.t. pe* R^2.

3) $P((X,Y) \in D) = \iint_D \rho(x,y)\ dx\ dy,\ D \subset R^2$.

4) $\int\int_{R^2} \rho(x,y)\ dx \cdot dy = 1$.

5) $\rho_X(x) = \int_{-\infty}^{\infty} \rho(x,y) dy,\ \forall x \in R$; $\rho_Y(y) = \int_{-\infty}^{\infty} \rho(x,y) dx,\ \forall y \in R$.

Definiția 2.2.12. *Spunem că variabilele aleatoare de tip continuu X şi Y sunt independente dacă F(x,y)* $= F_X(x) \cdot F_Y(y)$, $\forall (x,y) \in R^2$.

Aplicația 2.2.13. *Funcția* $\rho(x,y) = \begin{cases} kxy^2, & (x,y) \in [1,2] \times [1,3] \\ 0, & \text{în rest} \end{cases}$ *este densitate de probabilitate dacă* $\rho(x,y) \ge 0$ *şi* $\iint_{R^2} \rho(x,y) dxdy = 1$ *ceea ce implică ecuația în k,* $k \int_1^2 \int_1^3 xy^2 dxdy = 1$, *verificată pentru* $k = \frac{1}{13}$.

În acest caz funcția de repartiție va fi

$$F(x,y) = \int_1^x \int_1^y uv^2 du dv = \begin{cases} 0 & , \text{dacă } x < 1 \text{ sau } y < 1 \\ \frac{1}{78}(x^2-1)(y^3-1) & , \text{dacă } (x,y) \in [1,2] \times [1,3] \\ \frac{1}{26}(y^3-1) & , \text{dacă } y \in [1,3] \text{ şi } x > 2 \\ \frac{1}{2}(x^2-1) & , \text{dacă } y \in [1,2] \text{ şi } y > 3 \\ 1 & , \text{dacă } x > 2 \text{ şi } y > 3 \end{cases}$$

şi deducem de asemenea că funcţiile de repartiţie marginale sunt, respectiv,

$$F_X(x)=\begin{cases}0 & ,x<1\\ \frac{1}{3}(x^2-1) & ,x\in[1,2]\\ 1 & ,x>2\end{cases}\ ;\ F_Y(y)=\begin{cases}0 & ,y<1\\ \frac{1}{3}(x^2-1) & ,y\in[1,3]\\ 1 & ,y>3\end{cases}$$

2.3. Caracteristici numerice asociate variabilelor aleatoare

Fie $\{\Omega, K, P\}$ un câmp de probabilitate şi $X:\Omega\to R$ o variabilă aleatoare. În afara informaţiilor furnizate de funcţia de repartiţie $F(x)$ sau chiar de repartiţia probabilistă (discretă (p_i) sau continuă $(\rho(x))$ ale unei variabile aleatoare X, de un real folos teoretic şi practic sunt şi informaţiile pe care le conţin anumite caracteristici numerice (valoarea medie, dispersia, abaterea medie pătratică sau diverse alte momente) ale lui X despre această variabilă aleatoare.

Valoarea medie (speranţa matematică)

Definiţia 2.3.1. *Fie {Ω, K, P} un câmp borelian de probabilitate şi variabila aleatoare X : Ω → **R** cu distribuţia* $X\begin{pmatrix}x_i\\ p_i\end{pmatrix}$, $i\in I$. *Se numeşte valoare medie, caracteristica numerică* $E(X)=\sum_{i\in I}x_i\,p_i$.

Observaţia 2.3.2.

1) Dacă I este finită, valoarea medie există.

2) Dacă I este infinit numărabilă, E(X) există când seria care o defineşte este absolut convergentă.

Definiţia 2.3.3. *Fie{Ω, K, P} un câmp borelian de probabilitate şi variabila aleatoare X : Ω → **R** de tip continuu* $\mathrm{X}\begin{pmatrix}\mathrm{x}\\ \rho(\mathrm{x})\end{pmatrix}$, $\mathrm{x}\in\mathrm{R}$. *Se numeşte valoarea medie a variabilei X, caracteristica numerică* $E(X)=\int_{-\infty}^{+\infty}x\,\rho(\mathrm{x})\mathrm{dx}$. *Valoarea medie există atunci când integrala improprie care o defineşte este convergentă.*

Propoziţia 2.3.4.(proprietăţile valorii medii) *Au loc afirmaţiile :*

1) $E(a\,\mathrm{X}+\mathrm{b})=\mathrm{a}\,\mathrm{E}(\mathrm{X})+\mathrm{b},\forall\mathrm{a},\mathrm{b}\in\mathrm{R}$

2) E(X + Y) = E(X) + E(Y)

3) X,Y independente $\Rightarrow$ *E(X Y) = E(X) E(Y)*

Demonstraţie

a) Fie variabilele aleatoare de tip discret X, Y având repartițiile $X\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I}$, $Y\begin{pmatrix} y_j \\ q_j \end{pmatrix}_{j\in J}$

1. Avem $E(aX+b)=\sum_{i\in I}(ax_i+b)p_i=\sum_{i\in I}ax_ip_i+\sum_{i\in I}bp_i=aE(X)+b$

dacă variabila $aX+b$ are repartiția $aX+b\begin{pmatrix} ax_i+b \\ p_i \end{pmatrix}_{i\in I}$ și $\sum_{i\in I}p_i=1.$

2. Variabila X+Y are repartiția $X+Y\begin{pmatrix} x_i+y_j \\ p_{ij} \end{pmatrix}_{(i,j)\in I\times J}$,

$p_{ij}=P(X=x_i,Y=y_j)$

Rezultă

$$E(X+Y)=\sum_{i\in I}\sum_{j\in J}(x_i+y_i)p_{ij}=\sum_{i\in I}\sum_{j\in J}x_ip_{ij}+\sum_{i\in I}\sum_{j\in J}y_jp_{ij}=$$

$$=\sum_{i\in I}x_ip_i+\sum_{j\in J}y_jq_j=E(X)+E(Y)$$

dacă s-au folosit relațiile $\sum_{i\in I}p_{ij}=q_j$ și $\sum_{j\in J}p_{ij}=p_i$

3. Variabila XY are repartiția $XY\begin{pmatrix} x_iy_j \\ p_iq_j \end{pmatrix}_{(i,j)\in I\times J}$ dacă X și Y sunt independente.

Avem $E(XY)=\sum_{i\in I}\sum_{j\in J}x_iy_jp_iq_j=\sum_{i\in I}x_ip_i\sum_{j\in J}y_jq_j=E(X)E(Y)$

b) Presupunem ca X și Y sunt variabile aleatoare de tip continuu.

1. Dacă notăm prin Y = aX + b, a ≠ 0, atunci se obține că

$\rho_Y(x)=\dfrac{\rho_X(\frac{x-b}{a})}{|a|}$, pentru orice x ∈ R.

Avem: $E(aX+b)=E(Y)=\int_{-\infty}^{+\infty}x\rho_Y(x)dx=\dfrac{1}{|a|}\int_{-\infty}^{+\infty}x\rho_Y(\dfrac{x-b}{a})dx$ de unde prin schimbarea de variabilă $u = (x - b)/a$, $dx = adu$, obținem

$$E(aX+b)=\int_{-\infty}^{+\infty}(au+b)\rho_X(u)du=a\int_{-\infty}^{+\infty}u\rho_X(u)du+b\int_{-\infty}^{+\infty}\rho_X(u)du=aE(X)+b$$

2. Dacă notăm prin Z = X + Y, variabila care are densitatea de probabilitate ρ_Z, iar densitatea de probabilitate a vectorului (X, Y) o notăm prin ρ, atunci:

$$E(X+Y)=E(Z)=\int_{-\infty}^{+\infty}s\rho_Z(x)dx=\int_{-\infty}^{+\infty}x(\int_{-\infty}^{+\infty}\rho(u,x-u)du)dx$$

Schimbăm ordinea de integrare, apoi schimbarea de variabilă $x - u = t$, $dx = dt$, şi obţinem

$$E(X+Y) = \int_{-\infty}^{+\infty}(\int_{-\infty}^{+\infty} x\rho(u, x-u)dx)du = \int_{-\infty}^{+\infty}(\int_{-\infty}^{+\infty}(t+u)\rho(u,t)dt)du =$$

$$= \int_{-\infty}^{+\infty} t(\int_{-\infty}^{+\infty}\rho(u,t)du)dt + \int_{-\infty}^{+\infty} u(\int_{-\infty}^{+\infty}\rho(u,t)dt)du = \int_{-\infty}^{+\infty} t\rho_X(t)dt + \int_{-\infty}^{+\infty} u\rho_Z(u)du = E(Y) + E(X)$$

3. Dacă notăm prin V =X Y, care are densitatea de probabilitate ρ_V, iar densitatea de probabilitate a vectorului (X, Y) o notam prin ρ, atunci

$$E(XY) = E(V) = \int_{-\infty}^{+\infty} x\rho_V(x)dx = \int_{-\infty}^{+\infty} x(\int_{-\infty}^{+\infty}\rho(u, \frac{x}{u})\frac{dx}{|u|})dx$$

Schimbăm ordinea de integrare, apoi facem schimbarea de variabilă $x / u = t$, $dx = udt$, şi obţinem:

$$E(XY) = \int_{-\infty}^{+\infty}(\int_{-\infty}^{+\infty}\frac{x}{|u|}\rho(u, \frac{x}{u})dx)du = \int_{-\infty}^{+\infty}(\int_{-\infty}^{+\infty} tu\rho(u,t)dt)du =$$

$$= \int_{-\infty-\infty}^{+\infty+\infty}\int tu\rho_X(u)\rho_Y(t)dtdu = \int_{-\infty}^{+\infty} u\rho_X(u)du \int_{-\infty}^{+\infty} t\rho_Y(t)dt = E(X)E(Y)$$

Dispersia

Definiţia 2.3.5. *Fie {Ω, K, P} un câmp borelian de probabilitate şi variabila aleatoare X : Ω →* ***R.*** *Se numeşte dispersia (varianţa) variabilei aleatoare X, caracteristica numerică* $Var(X) = E\left[(X - E(X))^2\right]$ *iar* $\sigma(X) = \sqrt{Var(X)}$ *se numeşte abatere medie pătratică.*

În mod explicit, dispersia are expresia $Var(X) = \sum_{i \in I}(x_i - E(X))^2 \cdot p_i$, $I \subset N$, dacă X este o variabilă aleatoare discretă sau

$Var(X) = \int_R (x - M(X))^2 \rho(x)dx$, dacă X este o variabilă aleatoare continuă.

Dispersia este un indicator numeric al gradului de împrăştiere (sau de dispersare) a valorilor unei variabile aleatoare în jurul valorii medii a acesteia.

Propoziţia 2.3.6.(proprietăţile dispersiei)

a) $Var(X) = E(X^2) - [E(X)]^2$

b) $Var(aX + b) = a^2 Var(X), \forall a, b \in R$

c) X,Y independente $\Rightarrow Var(X+Y) = Var(X) + Var(Y)$

Demonstraţie

a) $Var(X) = E\left[(X - E(X))^2\right] = E\left[X^2 - 2E(X)X + (E(X))^2\right]$
$= E(X^2) - 2E(X)E(X) + [E(X)]^2 = E(X^2) - [E(X)]^2$

dacă s-a făcut un calcul formal.

b) Folosind proprietățile valorii medii şi definiția dispersiei avem:

$Var(aX + b) = E\left[(aX + b - aE(X) - b)^2\right] = E\left[a^2(X - E(X))^2\right] =$
$= a^2 E\left[(X - E(X))^2\right] = a^2 Var(X)$

c) Dacă X, Y sunt independente avem $E(XY) = E(X)E(Y)$. Calculăm

$Var(X + Y) = E\left[(X + Y - E(X + Y))^2\right] = E\left[((X - E(X)) + (Y - E(Y)))^2\right] =$
$= E\left[(X - E(X))^2 + (Y - E(Y))^2 + 2(X - E(X))(Y - E(Y))\right] =$
$= E\left[(X - E(X))^2\right] + E\left[(Y - E(Y))^2\right] + 2E(X - E(X))E(Y - E(Y)) =$
$= Var(X) + Var(Y)$

dacă s-a ținut seama că $E(X - E(X)) = 0$

Propoziția 2.3.7.(Inegalitatea lui Cebîşev) *Dacă variabila aleatoare X are valoare medie şi dispersie atunci* $\forall \varepsilon > 0$ *are loc inegalitatea* $P(|X - E(X)| < \varepsilon) \geq 1 - \frac{Var(X)}{\varepsilon^2}$ *sau inegalitatea echivalentă cu aceasta* $P(|X - E(X)| \geq \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}$.

Demonstrație

Presupunem că X este o variabilă aleatoare de tip continuu, având densitatea de probabilitate $\rho(x)$. Atunci

$$Var(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 \rho(x)dx \geq \int_D (x - E(X))^2 p(x)dx$$

unde $D = \{x / |x - E(X)|\} \geq \varepsilon$, deoarece $|x - E(X)| \geq \varepsilon$, avem că $(x - E(X))^2 \geq \varepsilon$. Deci, avem

$$\int_D (x - E(X))^2 \rho(x)dx \geq \varepsilon^2 \int_D \rho(x)dx = \varepsilon^2 P(|X - E(X)| \geq \varepsilon)$$

Am obținut că $Var(X) \geq \varepsilon^2 P(|X - E(X)| \geq \varepsilon)$, rezultă

$$P(|X - E(X)| \geq \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}$$

Folosind probabilitatea evenimentului contrar se obține şi cealaltă formă a inegalității: $P(|X - E(X)| < \varepsilon) = 1 - P(|X - E(X)| \geq \varepsilon) \geq 1 - \frac{Var(X)}{\varepsilon^2}$

Aplicația 2.3.8. *Dacă X este o variabilă aleatoare discretă*

$$X:\begin{pmatrix} -2 & -1 & 0 & 1 & 2 & 3 \\ \frac{1}{12} & \frac{2}{12} & \frac{2}{12} & \frac{5}{12} & \frac{1}{12} & \frac{1}{12} \end{pmatrix}$$

atunci deducem că:

$$E(X) = -2\frac{1}{12} - 1\frac{2}{12} + 0\frac{2}{12} + 1\frac{5}{12} + 2\frac{1}{12} + 3\frac{1}{12} = \frac{1}{2}$$

$$E(X^2) = 4\frac{1}{12} + 1\frac{2}{12} + 0\frac{2}{12} + 1\frac{5}{12} + 4\frac{1}{12} + 9\frac{1}{12} = 2$$

$$Var(X) = E(X^2) - [E(X)]^2 = 2 - \frac{1}{4} = \frac{7}{4}$$

$$\sigma(X) = \sqrt{Var(X)} = \frac{\sqrt{7}}{2}$$

Aplicația 2.3.9. *Dacă X este variabilă aleatoare continuă*

$$X:\begin{pmatrix} x \\ \rho(x) \end{pmatrix},\ \rho(x) = \begin{cases} \frac{x}{4}, & x \in [1,3] \\ 0, & \text{în rest} \end{cases}$$

atunci deducem că:

$$E(X) = \int_1^3 x\rho(x)dx = \frac{x^3}{12}\Big|_1^3 = \frac{13}{6}\ ,\ E(X^2) = \int_1^3 x^2\rho(x)dx = \frac{x^4}{16}\Big|_1^3 = 5$$

$$Var(X) = E(X^2) - [E(X)]^2 = 5 - \frac{169}{36} = \frac{11}{36}$$

$$\sigma(X) = \sqrt{Var(X)} = \frac{\sqrt{11}}{6}$$

Momente

Definiția 2.3.10. *Fie {Ω, K, P} un câmp borelian de probabilitate şi variabila aleatoare X : Ω → **R**. Se numeşte moment inițial (obişnuit) de ordin k al variabilei aleatoare X, caracteristica numerică* $m_k = E(X^k)$

Observația 2.3.11. *a) Pentru k=1 avem* $m_1 = E(X)$ *iar pentru k=2,* $Var(X) = m_2 - m_1^2$

b) Dacă X este variabilă de tip discret având repartiția $X:\begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i\in I}$,

$\sum_{i\in I} p_i = 1$ *atunci* $m_k = \sum_{i\in I} x_i^k p_i$

c) Dacă X este variabilă de tip continuu $X:\begin{pmatrix} x \\ \rho(x) \end{pmatrix}_{x\in R}$ *atunci* $m_k = \int_R x^k \rho(x)dx$

Definiția 2.3.12. *Se numeşte moment centrat de ordin k al variabilei aleatoare X, caracteristica numerică* $\mu_k = E\left[(X - E(X))^k\right]$, *adică*

$$\mu_k = \begin{cases} \sum_{i\in I} (x_i - E(X))^k \cdot p_i & , \text{ X discretă} \\ \int_R (x_i - E(X))^k \rho(x)dx & , \text{X continuă} \end{cases}$$

Observația 2.3.13. *Pentru k=1 avem* $\mu_1 = 0$, *iar pentru k=2,* $\mu_2 = Var(X)$

Teorema 2.3.14. *Între momentele centrate şi momentele inițiale există următoarea relație:* $\mu_k = \sum_{i=0}^{k} (-1)^i C_k^i m_{k-i} m_1^i$.

Demonstrație

Avem

$$\mu_k = E\left[(X - E(X))^k\right] = E\left[(X - m_1)^k\right] = E\left(\sum_{i=0}^{k} C_k^i X^{k-i} (-m_1)^i\right) =$$

$$= E\left[\sum_{i=0}^{k} (-1)^i C_k^i X^{k-i} m_1^i\right] = \sum_{i=0}^{k} (-1)^i C_k^i E(X^{k-i}) m_1^i = \sum_{i=0}^{k} (-1)^i C_k^i m_{k-i} m_1^i$$

Observația 2.3.15. *În statistica matematică se utilizează de regulă primele patru momente centrate:* $\mu_1, \mu_2, \mu_3, \mu_4$.

Definiția 2.3.16. *Se numeşte momentul inițial de ordinul (r,s) al vectorului aleator (X,Y) caracteristica numerică* $m_{rs} = E(X^r Y^s)$, *adică*

$$m_{rs} = \begin{cases} \sum_{i\in I} \sum_{j\in J} x_i^r y_j^s p_{ij} & , (X,Y) \text{ discret} \\ \iint_{R^2} x^r y^s \rho(x,y)dxdy & , (\text{X}, \text{Y}) \text{ continuu} \end{cases}$$

Definiția 2.3.17. *Se numeşte moment centrat de ordin (r,s) al vectorului aleator (X,Y), caracteristica numerică*

$\mu_{rs} = E\left[(X - E(X))^r (Y - E(Y)^s\right]$, *adică*

$$\mu_{rs} = \begin{cases} \sum_{i\in I}\sum_{j\in J}(x_i - E(X))^r (y_j - E(Y))^s p_{ij} & ,(X,Y) \text{ discret} \\ \iint_{R^2}(x - E(X))^r (y - E(Y))^s \rho(x,y)dxdy & ,(\mathrm{X},\mathrm{Y}) \text{ continuu} \end{cases}$$

Observația 2.3.18.

$m_{1o} = E(X),\ m_{01} = E(Y),\ \mu_{20} = Var(X),\ \mu_{02} = Var(Y)$

Corelație sau covarianță

Definiția 2.3.19. *Se numeşte corelaţia sau covarianţa variabilelor aleatoare X şi Y, caracteristica numerică*

$C(X,Y) = E[(X - E(X))(Y - E(Y))]$ adică $\mathrm{C(X, Y)} = \mu_{11}$

Observația 2.3.20.

1) $C(X,Y) = E(XY) - E(X)E(Y), \mathrm{C(X, Y)} = m_{11} - m_{10}m_{01}$

Dacă X, Y independente $\Rightarrow \mathrm{C(X, Y)} = 0$ *, dar nu şi reciproc.*

$C(X,X) = Var(X)$

$C\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j C(X_i, Y_j)$, *oricare ar fi variabilele aleatoare* X_i *şi* Y_j *şi oricare ar fi constantele reale* a_i *şi* b_j, $1 \le i \le m, 1 \le j \le n$

$C(X,Y) = C(Y,X)$*, oricare ar fi X şi Y.*

Definiția 2.3.21. *Se numeşte coeficient de corelație relativ la variabilele aleatoare X şi Y caracteristica numerică*

$$\mathrm{r(X, Y)} = \frac{C(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Observația 2.3.22.

1) X, Y independente $\Rightarrow \mathrm{r(X, Y)} = 0$ *reciproc nu este adevărat;*

2) Spunem că X,Y sunt necorelate dacă r(X,Y) =0

Proprietăți:

a) $|\mathrm{r(X, Y)}| \le 1$

b) $r(X,Y) = +1 \Leftrightarrow Y = a\,\mathrm{X} + \mathrm{b}, a > 0$

c) $r(X,Y) = -1 \Leftrightarrow Y = aX + b, a < 0$

Observația 2.3.23. *În practică se mai spune că:*

1) X şi Y sunt pozitiv perfect corelate dacă $r(X,Y) = 1$*;*

2) X şi Y sunt negativ perfect corelate dacă $r(X,Y) = -1$*;*

3) X şi Y sunt puternic pozitiv (sau negativ) corelate dacă

$0{,}75 \le r(X,Y) < 1$ *(sau* $-1 < r(X,Y) \le -0{,}75$*);*

4) X şi Y sunt slab pozitiv (sau negativ) corelate dacă $0 < r(X,Y) < 0{,}25$ *(sau* $-0{,}25 < r(X,Y) < 0$*);*

Marginile valorice decizionale fiind alese convenţional.

Aplicaţia 2.3.24. *Fie (X,Y) un vector aleator discret a cărui repartiţie probabilistă este dată de tabelul de mai jos.*

Calculaţi coeficientul de corelaţie r(X,Y).

X \ Y	-1	0	1	2	p_i
-1	1/6	1/12	1/12	1/24	9/24
0	1/24	1/6	1/12	1/24	8/24
1	1/24	1/24	1/6	1/24	7/24
q_j	6/24	7/24	8/24	3/24	1

Rezolvare

Pe baza formulelor corespunzătoare, deducem imediat:

$$E(X) = -1\frac{9}{24} + 0\frac{8}{24} + 1\frac{7}{24} = -\frac{2}{24} = -\frac{1}{12}$$

$$E(X^2) = 1\frac{9}{24} + 0\frac{8}{24} + 1\frac{7}{24} = \frac{16}{24} = \frac{2}{3}$$

$$Var(X) = E(X^2) - [E(X)]^2 = \frac{2}{3} - \frac{1}{144} = \frac{95}{144}$$

$$E(Y) = -1\frac{6}{24} + 0\frac{7}{24} + 1\frac{8}{24} + 2\frac{3}{24} = \frac{8}{24} = \frac{1}{3}$$

$$E(Y^2) = 1\frac{6}{24} + 0\frac{7}{24} + 1\frac{1}{24} + 4\frac{3}{24} = \frac{26}{24} = \frac{13}{12}$$

$$Var(Y) = E(Y^2) - [E(Y)]^2 = \frac{13}{12} - \frac{1}{9} = \frac{35}{36}$$

$$E(XY) = -1\cdot\left(-1\frac{1}{6} + 0\frac{1}{12} + 1\frac{1}{12} + 2\frac{1}{24}\right) + 0\cdot\left(-1\frac{1}{24} + 0\frac{1}{6} + 1\frac{1}{12} + 2\frac{1}{24}\right) +$$

$$+1\cdot\left(-1\frac{1}{24} + 0\frac{1}{24} + 1\frac{1}{6} + 2\frac{1}{24}\right) = \frac{5}{24}$$

$$C(X,Y) = E(XY) - E(X)E(Y) = \frac{5}{24} + \frac{1}{36} = \frac{17}{72}$$

$$r(X,Y) = \frac{C(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\frac{17}{12}}{\sqrt{\frac{95}{144}\cdot\frac{35}{36}}} \approx 0{,}295$$

Observaţia 2.3.25. *Coeficientul de corelaţie* $r(X,Y)$ *reprezintă prima măsură a corelaţiei sau gradului de dependenţă în sens clasic. Introdusă de către statisticianul englez K. Pearson în anul 1901 ca rod al colaborării acestuia cu antropologul englez F. Galton (care a avut prima idee de măsurare a corelaţiei sub denumirea de variaţie legată), această măsură a gradului de dependenţă a fost criticată încă de la apariţiei ei pentru diverse motive, printre care şi aceea că:*

1) *este dependentă de valorile vectorului aleator* (X,Y) *şi ca urmare nu este aplicabilă pentru cazul variabilelor aleatoare necantitative;*
2) *nu este precisă în cazul independenţei şi al necorelării deoarece dacă* $r(X,Y)=0$ *nu există un răspuns categoric (în sensul independenţei sau necorelării);*
3) *nu poate fi extinsă la mai mult de două variabile aleatoare sau chiar la doi sau mai mulţi vectori aleatori, fapte cerute de practică.*

Dacă la prima obiecţie a dat chiar K. Pearson un răspuns, pentru celelalte două obiecţii nu s-au dat răspunsuri clare decât după apariţia în 1948 a teoriei matematice a informaţiei, rezultate remarcabile în acest sens obţinând şcoala românească de matematică sub conducerea lui Silviu Guiaşu introducând măsurile entropice ale dependenţei dintre variabile aleatoare şi vectori aleatori (în anii 1974-1978) cu o largă aplicabilitate teoretică şi practică.

În ciuda tuturor criticilor ce i s-au adus, coeficientul de corelaţie clasic (sau coeficientul Galton-Pearson) este cel mai frecvent utilizat în practică şi, pentru că este cel mai simplu în utilizare.

Definiţia 2.3.26. *Fiind dat vectorul aleator* $Z=(X_1,X_2,...,X_n)$ $Z:E\to R^n$, *se numeşte valoare medie a acestuia şi se notează cu* $E(Z)$, *dacă există, vectorul n-dimensional ale cărui componente sunt valorile medii ale componentelor lui Z adică:*

$$E(Z)=E(X_1,X_2,...,X_n)=\left(E(X_1),E(X_2),...,E(X_n)\right).$$

Se numeşte matrice de covarianţă (sau de corelaţie) a vectorului Z şi se notează prin $C(Z)$, *dacă există, matricea* $C(Z)=\left(c_{ij}\right)_{\substack{i=\overline{1,n}\\ j=\overline{1,n}}}=\left(C\left(X_i,Y_j\right)\right)_{i,j=\overline{1,n}}$

Observaţia 2.3.27.

a) *Pentru cazul unui vector aleator bidimensional, a nu se face confuzie între media produsului componentelor X şi Y, care este* $E(XY)$ *şi media vectorului* (X,Y) *care este* $E(X,Y)$.
b) *Uneori matricea de corelaţie* $C(Z)$ *se mai notează şi cu* $\Gamma(Z)$.
c) *Desfăşurat matricea de covarianţă* $C(Z)$ *are forma:*

$$C(Z)=\begin{pmatrix} Var(X_1) & C(X_1,X_2) & \dots & C(X_1,X_n) \\ C(X_2,X_1) & Var(X_2) & \dots & C(X_2,X_n) \\ \dots & \dots & \dots & \dots \\ C(X_n,X_1) & C(X_n,X_2) & \dots & Var(X_n) \end{pmatrix}$$

şi ca urmare a proprietăţilor corelaţiei, constatăm că matricea $C(Z)$ *este simetrică.*

d) Pornind de la definiţia coeficientului de corelaţie şi de la matricea de corelaţie, dacă toate componentele lui Z sunt neconstante, atunci putem introduce matricea coeficienţilor de corelaţie $R(Z)$ *a cărei formă dezvoltată este:*

$$R(Z)=\begin{pmatrix} 1 & r(X_1,X_2) & \dots & r(X_1,X_n) \\ r(X_2,X_1) & 1 & \dots & r(X_2,X_n) \\ \dots & \dots & \dots & \dots \\ r(X_n,X_1) & r(X_n,X_2) & \dots & 1 \end{pmatrix}$$

Ambele forme ale matricei de corelaţie a vectorului aleatoriu Z reprezintă de fapt tabele ale măsurării gradului de dependenţă dintre componentele lui Z, considerate două câte două.

Aplicaţia 2.3.28. *Fie variabilele aleatoare:*

$$X_1:\begin{pmatrix} -1 & 1 \\ p_1 & p_2 \end{pmatrix};\ X_2:\begin{pmatrix} 1 & 3 \\ q_1 & q_2 \end{pmatrix};\ X_3:\begin{pmatrix} 2 & 4 \\ r_1 & r_2 \end{pmatrix}$$

a căror repartiţie comună notată (p_{ijk}), $1\le i,j,k\le 2$, *este:*

$$p_{111}=\frac{1}{16};\ p_{112}=\frac{1}{16};\ p_{121}=\frac{1}{32};\ p_{122}=\frac{3}{32};$$
$$p_{211}=\frac{1}{8};\ p_{212}=\frac{1}{4};\ p_{221}=\frac{1}{16};\ p_{222}=\frac{5}{16}.$$

Să se determine repartiţiile bidimensionale şi unidimensionale ale vectorului aleator tridimensional $Z=(X_1,X_2,X_3)$ *şi matricele de corelaţie* $C(Z)$ *şi* $R(Z)$.

Rezolvare

Avem imediat repartiţiile bidimensionale

$$\left.\begin{matrix} p_{11\bullet}=p_{111}+p_{112}=\frac{1}{8} & ; & p_{12\bullet}=p_{121}+p_{122}=\frac{1}{8} \\ p_{21\bullet}=p_{211}+p_{212}=\frac{3}{8} & & p_{22\bullet}=p_{221}+p_{222}=\frac{3}{8} \end{matrix}\right\} \text{ pentru } (X_1,X_2)$$

$$\left.\begin{matrix} p_{1\bullet 1}=p_{111}+p_{121}=\frac{3}{32} & ; & p_{1\bullet 2}=p_{112}+p_{122}=\frac{5}{32} \\ p_{2\bullet 1}=p_{211}+p_{221}=\frac{8}{16} & & p_{2\bullet 2}=p_{212}+p_{222}=\frac{9}{16} \end{matrix}\right\} \text{pentru } (X_1,X_3)$$

$$\left.\begin{array}{ll} p_{\bullet 11} = p_{111} + p_{211} = \dfrac{3}{16} & p_{\bullet 12} = p_{112} + p_{212} = \dfrac{5}{32} \\ p_{\bullet 21} = p_{121} + p_{221} = \dfrac{3}{32} & p_{\bullet 22} = p_{122} + p_{222} = \dfrac{13}{32} \end{array}\right\} \text{pentru } (X_2, X_3)$$

şi ca urmare putem scrie următoarele tabele de repartiţie bidimensionale:

X_2 / X_1	1	3	p_i
-1	1/8	1/8	¼
1	3/8	3/8	¾
q_j	1/2	1/2	1

X_3 / X_1	2	4	p_i
-1	3/32	5/32	1/4
1	3/16	9/16	3/4
r_k	9/32	23/32	1

X_3 / X_2	2	4	q_i
1	3/16	5/16	1/2
3	3/32	13/32	1/2
r_k	9/32	21/32	1

din care se observă şi repartiţiile unidimensionale (repartiţiile variabilelor aleatoare considerate X_1, X_2, X_3). Din aceste tabele deducem prin calcul imediat:

$$E(X_1) = \frac{1}{2};\ E(X_1^2) = 1;\ Var(X_1) = \frac{3}{4}$$

$$E(X_2) = 2;\ E(X_2^2) = 5;\ Var(X_2) = 1$$

$$E(X_3) = \frac{55}{16};\ E(X_3^2) = \frac{101}{8};\ Var(X_3) = \frac{207}{256}$$

$$E(X_1 \cdot X_2) = 1;\ E(X_1)E(X_2) = 1;\ C(X_1, X_2) = 0;\ r(X_1, X_2) = 0$$

$$E(X_1 \cdot X_3) = \frac{29}{16};\ E(X_1)E(X_3) = \frac{55}{32};\ C(X_1, X_3) = \frac{3}{32};\ r(X_1, X_3) = \sqrt{\frac{3}{207}} \approx 0{,}12$$

$$E(X_2 \cdot X_3) = \frac{113}{16};\ E(X_2)E(X_3) = \frac{55}{8};$$

$$C(X_2, X_3) = \frac{3}{16};\ r(X_2, X_3) = \frac{3}{\sqrt{207}} \approx 0{,}21$$

şi ca urmare putem scrie matricele de corelaţie:

$$C(Z) = \begin{pmatrix} 3/4 & 0 & 3/32 \\ 0 & 1 & 3/16 \\ 3/32 & 3/16 & 207/256 \end{pmatrix} \text{ şi } R(Z) = \begin{pmatrix} 1 & 0 & 0{,}12 \\ 0 & 1 & 0{,}21 \\ 0{,}12 & 0{,}21 & 1 \end{pmatrix}$$

constatând că X_1 şi X_2 sunt independente în timp ce între X_3 şi X_1 sau X_3 şi X_2 există o anumită dependenţă chiar dacă nu este puternică.

Alte caracteristici numerice

Definiţia 2.3.29. *Se numeşte mediana unei variabile aleatoare X, caracteristica numerică M_e care verifică relaţia:*

$$P(X \geq M_e) \geq \frac{1}{2} \leq P(X \leq M_e)$$

Observația 2.3.30. *1. Dacă F este funcția de repartiție şi este continuă atunci* M_e *se determină din ecuația* $F(M_e)=\frac{1}{2}$.

2. Dacă $M_e \in (a,b]$ *atunci se ia* $M_e = \frac{a+b}{2}$

Definiția 2.3.31. *Se numeşte valoare modală sau modul a variabilei aleatoare X orice punct de maxim local al distribuției lui X (în cazul discret) respectiv al densității de probabilitate (în cazul continuu).*

Observația 2.3.32. *Dacă există un singur punct de maxim local spunem că legea lui X este unimodală altfel o numim plurimodală.*

Definiția 2.3.33. *Se numeşte asimetria (coeficientul lui Fischer) variabilei aleatoare X caracteristica numerică definită prin* $s = \frac{\mu_3}{\sigma^3}$.

Definiția 2.3.34. *Se numeşte exces al variabilei aleatoare X, caracteristica numerică definită prin* $\mathrm{e} = \frac{\mu_4}{\sigma^4} - 3$.

Observația 2.3.35.

1) *Dacă e<0 atunci graficul distribuției are un aspect turtit şi legea se numeşte platicurtică.*

2) Dacă e>0 atunci graficul distribuției are un aspect ascuțit şi legea va fi numită leptocurtică.

3) Dacă e = 0 atunci repartițiile sunt mezocurtice.

Definiția 2.3.36. *Dacă X este o variabilă aleatoare cu funcția de repartiție* $F(x)$*, se numesc cuartile (în număr de trei) ale lui X (sau ale repartiției lui X) numerele* q_1, q_2 *şi* q_3 *cu proprietățile:*

$$\begin{cases} F(q_1) \le \frac{1}{4} \\ F(q_1+0) \ge \frac{1}{4} \end{cases} \quad \begin{cases} F(q_2) \le \frac{1}{2} \\ F(q_2+0) \ge \frac{1}{2} \end{cases} \quad \begin{cases} F(q_3) \le \frac{3}{4} \\ F(q_3+0) \ge \frac{3}{4} \end{cases}$$

Observăm că $q_2 = M_e$.

Exemplul 2.3.37. *Se consideră variabila aleatoare* $\mathrm{X}: \begin{pmatrix} -1 & 0 & 2 \\ 0{,}2 & 0{,}3 & 0{,}5 \end{pmatrix}$.

Să se calculeze: E(X), E(3X), E(4X-2), Var(X), σ_X.

Rezolvare

$$E(X)=\sum_{i=1}^{3}x_i p_i=-1\cdot 0{,}2+0\cdot 0{,}3+2\cdot 0{,}5=0{,}8$$

$$E(3X)=3E(X)=3\cdot 0{,}8=3{,}4;\quad E(4X-2)=4E(X)-2=4\cdot 0{,}8-2=1{,}2$$

$$Var(X)=E(X^2)-\left[E(X)\right]^2=2{,}2-0{,}64=1{,}56;$$

$$E(X^2)=(-1)^2\cdot 0{,}2+0^2\cdot 0{,}3+2^2\cdot 0{,}5=2{,}2;\quad \sigma_X=\sqrt{Var(X)}=\sqrt{1{,}56}=1{,}24$$

Exemplul 2.3.38. *Să se calculeze valoarea medie şi dispersia variabilei aleatoare care are densitatea de probabilitate*

$$\rho(x)=\begin{cases}1-|1-x|, \text{ dacă } x\in(0,2)\\ 0, \text{ altfel}\end{cases}$$

Rezolvare

Observăm că: $\rho(x)=\begin{cases}x, \text{ dacă } 0<x\le 1\\ 2-x, \text{ dacă } 1<x<2\\ 0, \text{ altfel}\end{cases}$

Ținând seama de definiție avem:

$$E(X)=\int_{-\infty}^{+\infty}x\,\rho(x)dx=\int_0^1 x^2dx+\int_1^2 x(2-x)dx=\frac{x^3}{3}\Big|_0^1+x^2\Big|_1^2-\frac{x^3}{3}\Big|_1^2=1$$

$$E(X^2)=\int_{-\infty}^{+\infty}x^2\rho(x)dx=\int_0^1 x^3dx+\int_1^2 x^2(2-x)dx=\frac{x^4}{4}\Big|_0^1+2\frac{x^3}{3}\Big|_1^2-\frac{x^4}{4}\Big|_1^2=\frac{7}{6}$$

$$Var(X)=E(X^2)-\left[E(X)\right]^2=\frac{7}{6}-1=\frac{1}{6}$$

Exemplul 2.3.39. *Fie vectorul aleator (X,Y) cu densitatea de probabilitate* $\rho(x,y)=\begin{cases}k(x+y+1), x\in[0,1], y\in[0,2]\\ 0, \text{în rest}\end{cases}$ *Se cere:*

a) să se determine constanta k;

b) să se determine densitățile marginale;

c) să se cerceteze dacă X şi Y sunt independente sau nu;

d) să se calculeze coeficientul de corelație între X şi Y.

Rezolvare

a) din condițiile $\rho(x,y)\ge 0\Rightarrow k\ge 0$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\rho(x,y)dxdy=k\int_0^1 dx\int_0^2(x+y+1)dy=1\Rightarrow k=1/5.$$

Deci $\rho(x,y)=\begin{cases}\frac{1}{5}(x+y+1), x\in[0,1], y\in[0,2]\\ 0, \text{în rest}\end{cases}$

b) $\rho_X(x) = \int_{-\infty}^{+\infty} \rho(x,y)dy = \frac{1}{5}\int_0^2 (x+y+1)dy = \frac{2x+4}{5}, x \in [0,1]$

$$\Rightarrow \rho_X(x) = \begin{cases} \frac{2x+4}{5}, x \in [0,1] \\ 0, \text{altfel} \end{cases}$$

$$\rho_y(y) = \int_{-\infty}^{+\infty} \rho(x,y)dx = \frac{1}{5}\int_0^1 (x+y+1)dx = \frac{2y+3}{10}, y \in [0,2]$$

$$\Rightarrow \rho_Y(y) = \begin{cases} \frac{2y+3}{10}, y \in [0,2] \\ 0, \text{altfel} \end{cases}$$

c) X şi Y nu sunt independente deoarece: $\rho(x,y) \neq \rho_X(x) \cdot \rho_Y(y)$

d) $E(X) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x\rho(x,y)dxdy = \int_{-\infty}^{+\infty} x\rho_X(x)dx = \frac{1}{5}\int_0^1 x(2x+4)dx = \frac{8}{15}$

$$E(Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \rho(x,y)dxdy = \int_{-\infty}^{+\infty} y\rho_Y(y)dy = \frac{1}{10}\int_0^2 y(2y+3)dy = \frac{17}{15}$$

$$m_2(X) = E(X^2) = \int_{-\infty}^{+\infty} x^2\rho_X(x)dx = \frac{1}{5}\int_0^1 x^2(2x+4)dx = \frac{11}{30}$$

$$m_2(Y) = E(Y^2) = \int_{-\infty}^{+\infty} y^2\rho_Y(y)dy = \frac{1}{10}\int_0^2 y^2(2y+3)dy = \frac{8}{15}$$

Deci $Var(X) = E(X^2) - [E(X)]^2 = \frac{11}{30} - \frac{64}{225} = \frac{37}{450}$

$$Var(Y) = E(Y^2) - [E(Y)]^2 = \frac{8}{5} - \frac{289}{225} = \frac{71}{225}$$

$$E(X \cdot Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy\ \rho(x,y)dxdy = \frac{1}{5}\int_0^1 dx\int_0^2 xy(x+y+1)dy =$$

$$= \frac{1}{5}\int_0^1 \left(2x^2 + \frac{14}{3}x\right)dx = \frac{9}{15} \Rightarrow C(X,Y) = E(XY) - E(X)E(Y) = .$$

$$= \frac{9}{15} - \frac{8}{15}\cdot\frac{12}{15} = \frac{-1}{225}$$

Se obține: $r(X,Y) = \frac{C(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{-\frac{1}{225}}{\sqrt{\frac{37}{450}\cdot\frac{71}{225}}} = -0,02758$

Exemplul 2.3.40. *Se ştie că, dacă două variabile aleatoare X şi Y sunt independente, atunci coeficientul lor de corelație este nul. Reciproca nu este adevărată. Iată un vector aleator discret (X,Y), în care X şi Y sunt dependente şi totuşi* $r = 0$.

X \ Y	-1	1	2
0	1/16	2/16	3/16
1	2/16	0	2/16
2	1/16	2/16	3/16

Rezolvare

Calculăm repartițiile marginale:

$$X:\begin{pmatrix} 0 & 1 & 2 \\ 6/16 & 4/16 & 6/16 \end{pmatrix};\quad Y:\begin{pmatrix} -1 & 1 & 2 \\ 4/16 & 4/16 & 8/16 \end{pmatrix}$$

Avem: $E(X)=1;\ E(X^2)=\frac{7}{4};\ Var(X)=\frac{3}{4};\ \sigma_X=\frac{\sqrt{3}}{2}$

$$E(Y)=1;\ E(Y^2)=\frac{5}{2};\ Var(Y)=\frac{3}{2};\ \sigma_Y=\frac{\sqrt{10}}{2}$$

$$\mathrm{X}\cdot\mathrm{Y}:\begin{pmatrix} -2 & -1 & 0 & 2 & 4 \\ 1/16 & 2/16 & 6/16 & 4/16 & 3/16 \end{pmatrix},\ \mathrm{E(XY)=1}$$

$$\Rightarrow\ r=\frac{\mathrm{E(X\cdot Y)-E(X)E(Y)}}{\sigma_X\cdot\sigma_Y}=\frac{1-1}{\frac{\sqrt{3}}{2}\cdot\frac{\sqrt{10}}{2}}=0$$

Exemplul 2.3.41. *Fie X o variabilă aleatoare care are densitatea de probabilitate definită prin:* $\rho(\mathrm{x})=\begin{cases}0, \mathrm{x}\notin(0,2)\\ 1/2, \mathrm{x}\in(0,2)\end{cases}$.

a) Să se determine modulul şi mediana

b) Să se calculeze momentul de ordin k, $m_k(x)$..

Rezolvare

a) Conform definiției, M_0 este valoarea pentru care $\rho(\mathrm{x})\to\max$. adică $M_0\in(0,2)$ adică există o infinitate de valori modale situate pe segmentul (0,2). M_e se determină din ecuația $F(M_e)=\frac{1}{2}$.

Cum $F(M_e) = P(X \le M_e) = \int_0^{M_e} \rho(x)dx = \frac{M_e}{2} \Rightarrow M_e = 1.$

b) $m_k(X) = E(X^k) = \int_{-\infty}^{+\infty} x^k \cdot \frac{1}{2} dx = \frac{2k}{k+1}.$

2.4. Funcția caracteristică. Funcția generatoare de momente

Definiția 2.4.1. *Fie câmpul de probabilitate* $\{\Omega, K, P\}$ *şi variabilele aleatoare X şi Y definite pe Ω cu valori reale. Se numeşte variabilă aleatoare complexă* $Z = X + iY$, $i^2 = -1$, *iar valoarea medie a acesteia notată cu* $E(Z)$ *este dată de relația* $E(Z) = E(X) + i\,\mathrm{E}(Y)$ *dacă mediile* $E(X)$ *şi* $E(Y)$ *există.*

Observația 2.4.2. *Dacă X este o variabilă aleatoare expresia* $e^{itX} = \cos tX + i \sin tX$, $t \in R$ *defineşte de asemenea o variabilă aleatoare şi* $\left|e^{itX}\right|^2 = \cos^2 tX + \sin^2 tX = 1$

Definiția 2.4.3. *Fie X o variabilă aleatoare reală. Se numeşte funcția caracteristică a lui X o funcție* $\varphi_X : R \to C$ *dată de relația* $\varphi_X(t) = \varphi(t) = E(e^{itX})$, *care explicit poate fi scrisă sub forma*

$$\varphi_X(t) = \begin{cases} \sum_{k \in K} p_k e^{itx_k} & \text{, este de tip discret} \\ \int_R e^{itx} \rho(x)dx & \text{, este de tip continuu} \end{cases}$$

Propoziția 2.4.4. *Funcția caracteristică are următoarele proprietăți:*

1) $\varphi(0) = 1$ *şi* $|\varphi(t)| \le 1, \forall\, t \in R$

2) Dacă X_j, $j = \overline{1,m}$ *sunt variabile aleatoare independente în totalitate cu funcțiile caracteristice* $\left(\varphi_{X_j}(t) = \varphi_j(t), \mathrm{j} = \overline{1,\mathrm{m}}\right)$, *atunci funcția caracteristică a variabilei aleatoare sumă* $X = X_1 + X_2 + ... + X_m$ *este*

$$\varphi_X(t) = \varphi_1(t)\varphi_2(t)...\varphi_m(t) = \prod_{j=1}^{m} \varphi_j(t)$$

3) Dacă $Y = aX + b$, a *şi* $b \in R$, *atunci* $\varphi_Y(t) = \varphi_X(at)e^{itb}$

4) Dacă X admite momente inițiale de orice ordine atunci funcția caracteristică admite derivate de orice ordin şi are loc relația

$$m_r(X) = E(X^r) = \frac{1}{i^r} \varphi_X^{(r)}(0)$$

Demonstrație

1) $\varphi(0) = E(1) = 1$ şi

$|\varphi(t)| = \left|\sum_{k\in K} p_k e^{itx_k}\right| \le \sum_{k\in K} p_k \left|e^{itx}\right| = \sum_{k\in K} p_k = 1$ dacă X este de tip discret şi

$|\varphi(t)| = \left|\int_R e^{itx}\rho(x)dx\right| \le \int_R \left|e^{itx}\right|\rho(x)dx = \int_R \rho(x)dx = 1$ dacă X este de tip continuu.

2) Având în vedere proprietăţile valorii medii, putem scrie că

$$\varphi_X(t) = E(e^{itX}) = E(e^{itX_1}\cdot e^{itX_2} \dots e^{itX_m}) = \prod_{j=1}^{m} E(e^{itX_j}) = \prod_{j=1}^{m} \varphi_j(t)$$

3) Tot ca urmare a proprietăţilor mediei avem:

$$\varphi_Y(t) = E(e^{itY}) = E(e^{itaX}\cdot e^{itb}) = e^{itb}\varphi_{aX}(t) = e^{itb}\varphi_X(at)$$

4) Observăm că $\varphi_X^{(r)}(t) = E(X^r e^{itx} i^r) = i^r E(X^r e^{itX})$ şi rezultă $\varphi_X^{(r)}(0) = i^r E(X^r) = i^r m_r(X)$ *.q.e.d*

Observaţia 2.4.5. *Folosirea relaţiei de la punctul 4) este recomandabilă doar atunci când calcularea momentelor este mai comodă prin această relaţie decât pornind direct de la definiţia acestora.*

Aplicaţia 2.4.6.

1) Dacă $X:\begin{pmatrix} -1 & 0 & 1 \\ 1/6 & 1/2 & 1/3 \end{pmatrix}$ *atunci*

$$\varphi(t) = \frac{1}{6}e^{-it} + \frac{1}{2} + \frac{1}{3}e^{it} = \frac{2e^{it} + e^{-it} + 3}{6}$$

2) Dacă $X:\begin{pmatrix} x \\ 2x \end{pmatrix}$, $x \in [0,1]$ *atunci* $\varphi(t) = \int_0^1 2xe^{itx}dx = \frac{2-2ix}{e^2}e^{itx}$

3) Dacă $X:\begin{pmatrix} x \\ e^{-x} \end{pmatrix}$, $x \ge 0$ *atunci* $\varphi(t) = \int_0^\infty e^{-x}e^{itx}dx = \frac{1+it}{1+t^2}$

Definiţia 2.4.7. *Fie X o variabilă aleatoare reală definită pe câmpul de probabilitate* $\{\Omega, K, P\}$. *Se numeşte funcţie generatoare de momente, dacă există, funcţia* $G: R \to R$, *dată de relaţia* $G_X(t) = G(t) = E(e^{tX})$ *care explicit poate fi scrisă sub forma*

$$G(t) = \begin{cases} \sum_{k\in K} p_k e^{tx_k} & \text{, X este de tip discret} \\ \int_R e^{tx}\rho(x)dx & \text{, X este de tip continuu} \end{cases}$$

cu condiţia existenţei expresiilor corespunzătoare.

Propoziţia 2.4.8. *Funcţia generatoare de momente are următoarele proprietăţi:*

1) $G(0)=1$

2) Dacă X_j, $1\le j\le m$, *sunt independente în totalitate şi au funcţiile generatoare* $G_j(t)$, $j=\overline{1,m}$, *atunci funcţia generatoare a variabilei aleatoare* $X=X_1+X_2+...+X_m$ *este*

$$G_X(t)=\prod_{j=1}^{m}G_j(t)$$

3) Dacă $Y=aX+b$, *a şi* $b\in R$, *atunci*

$G_Y(t)=G_X(at)\cdot e^{tb}$

4) Dacă X admite momente iniţiale de orice ordin, atunci funcţia generatoare admite derivate de orice ordin în punctul zero şi $G_X^{(r)}(0)=E(X^r)=m_r(X)$, $r=1,2,...$

Aplicaţia 2.4.9.

1) Dacă $X:\begin{pmatrix}-1 & 0 & 1\\ 1/6 & 1/12 & 2/3\end{pmatrix}$ *atunci*

$$G(t)=\frac{1}{6}e^{-t}+\frac{1}{2}+\frac{1}{3}e^{-t}=\frac{2e^{-t}+e^{-t}+3}{6}$$

2) Dacă $X:\begin{pmatrix}x\\ \lambda e^{-\lambda x}\end{pmatrix}$, $x\ge 0$, $\lambda>0$ *atunci*

$$G(t)=\lambda\int_0^{\infty}e^{-\lambda x}e^{tx}dx=\frac{\lambda}{\lambda-t},\ dacă\ t<\lambda$$

iar în caz contrar nu există.

3) Dacă $X:\begin{pmatrix}k\\ C_n^k p^k q^{n-k}\end{pmatrix}_{k=\overline{0,n}}$, $p,q>0$, $p+q=1$, *atunci*

$$G(t)=\sum_{k=0}^{n}C_n^k p^k q^{n-k}e^{tk}=(pe^t+q)^n$$

$$G'(t)=npe^t(pe^t+q)^{n-1};\ G''(t)=npe^t(pe^t+q)^{n-1}+n(n-1)p^2e^{2t}(pe^t+q)^{n-2}$$

$$G'(0)=np=E(X);\ G''(0)=n^2p^2+npq=E(X^2)$$

2.5. Probleme rezolvate

Aplicaţia 2.5.1. *Fie variabilele aleatoare independente :*

$$X:\begin{pmatrix}0 & 1 & 2\\ 1/2 & 1/4 & 1/4\end{pmatrix}\ şi\ Y:\begin{pmatrix}-1 & 1 & 2\\ 1/3 & 1/6 & 1/2\end{pmatrix}.$$

Să se scrie variabilele aleatoare : 2X, Y^2, *X+Y, XY, 2X+3Y, X/Y, max(X,Y),* $\sqrt{X}$.

Rezolvare

Probabilităţile corespunzătoare valorilor lui 2X, Y^2, $\sqrt{X}$ sunt aceleaşi cu cele corespunzătoare lui X şi respectiv Y. Avem:

$$2X:\begin{pmatrix}0 & 2 & 4\\ 1/2 & 1/4 & 1/4\end{pmatrix},\ \sqrt{X}:\begin{pmatrix}0 & 1 & \sqrt{2}\\ 1/2 & 1/4 & 1/4\end{pmatrix},\ Y^2:\begin{pmatrix}1 & 4\\ 1/2 & 1/2\end{pmatrix}.$$

$$P(Y^2=1)=P(Y=-1\vee Y=1)=P(Y=-1)+P(Y=1)=\frac{1}{3}+\frac{1}{6}=\frac{1}{2}.$$

Deoarece X şi Y sunt independente avem că $p_{ij} = p_iq_j, 1 \le i, j \le 3$. De exemplu $p_{12}=P(X=0,Y=1)=P(X=0)P(Y=1)=\frac{1}{2}\cdot\frac{1}{6}=\frac{1}{12}$. Obţinem

$$X+Y:\begin{pmatrix}0-1 & 0+1 & 0+2 & 1-1 & 1+1 & 1+2 & 2-1 & 2+1 & 2+2\\ 1/6 & 1/12 & 1/4 & 1/12 & 1/24 & 1/8 & 1/12 & 1/24 & 1/8\end{pmatrix}$$

adică

$$X+Y:\begin{pmatrix}-1 & 0 & 1 & 2 & 3 & 4\\ 1/6 & 1/12 & 1/6 & 7/24 & 1/6 & 1/8\end{pmatrix}.$$

Analog

$$X\cdot Y:\begin{pmatrix}0\cdot(-1) & 0\cdot 1 & 0\cdot 2 & 1\cdot(-1) & 1\cdot 1 & 1\cdot 2 & 2\cdot(-1) & 2\cdot 1 & 2\cdot 2\\ 1/6 & 1/12 & 1/4 & 1/12 & 1/24 & 1/8 & 1/12 & 1/24 & 1/8\end{pmatrix}$$

de unde

$$X\cdot Y:\begin{pmatrix}-2 & -1 & 0 & 1 & 2 & 4\\ 1/12 & 1/12 & 1/2 & 1/24 & 1/6 & 1/8\end{pmatrix}.$$

Cum 2X şi 3Y au repartiţiile

$$2X:\begin{pmatrix}0 & 2 & 4\\ 1/2 & 1/4 & 1/4\end{pmatrix},\ 3Y:\begin{pmatrix}-3 & 3 & 6\\ 1/3 & 1/6 & 1/2\end{pmatrix},$$

obţinem repartiţia lui 2X + 3Y :

$$2X+3Y:\begin{pmatrix}-3 & -1 & 1 & 3 & 5 & 6 & 7 & 8 & 10\\ 1/6 & 1/12 & 1/12 & 1/12 & 1/24 & 1/4 & 1/24 & 1/8 & 1/8\end{pmatrix}.$$

La fel obţinem:

$$\frac{X}{Y}:\begin{pmatrix}-2 & -1 & 0 & 1/2 & 1 & 2\\ 1/12 & 1/12 & 1/2 & 1/8 & 1/6 & 1/24\end{pmatrix},$$

$$\max(X,Y):\begin{pmatrix}0 & 1 & 2\\ 1/6 & 5/24 & 5/8\end{pmatrix}.$$

Aplicaţia 2.5.2. *Fie X şi Y două variabile aleatoare discrete ale căror repartiţii probabiliste comună (p_{ij}) şi marginale (p_i) şi (q_j) sunt date în tabelul următor :*

X \ Y	-1	0	1	p_i
-1	1/8	1/12	1/6	3/8
1	1/24	1/4	1/3	5/8
q_j	1/6	1/3	1/2	1

a) Să se scrie variabilele aleatoare X şi Y.
b) Să se precizeze dacă X şi Y sunt independente.
c) Să se scrie variabilele $X+Y$, $X\cdot Y$, X^2, Y^2, Y/X.

Rezolvare

a)Din tabelul de repartiție de mai sus deducem că

$$X:\begin{pmatrix}-1 & 1\\ 3/8 & 5/8\end{pmatrix} \text{ şi } Y:\begin{pmatrix}-1 & 0 & 1\\ 1/6 & 1/3 & 1/2\end{pmatrix}.$$

b)Dacă X şi Y ar fi independente atunci

$$p_{11}=P(X=-1,Y=-1)=p_1q_1=P(X=-1)P(Y=-1),$$

ceea ce nu are loc întrucât $\frac{1}{8}\neq\frac{3}{8}\cdot\frac{1}{6}$.

c)Deoarece

$$p_{11}=\frac{1}{8},\ p_{12}=\frac{1}{12},\ p_{13}=\frac{1}{6},\ p_{21}=\frac{1}{24},\ p_{22}=\frac{1}{4},\ p_{23}=\frac{1}{3},$$

obținem

$$X+Y:\begin{pmatrix}-2 & -1 & 0 & 1 & 2\\ 1/8 & 1/12 & 1/6+1/24 & 1/4 & 1/3\end{pmatrix},$$

$$X\cdot Y:\begin{pmatrix}-1 & 0 & 1\\ 1/6+1/24 & 1/12+1/4 & 1/8+1/3\end{pmatrix},$$

$$\frac{X}{Y}:\begin{pmatrix}-1 & 0 & 1\\ 1/24+1/6 & 1/12+1/4 & 1/8+1/3\end{pmatrix}.$$

Repartițiile lui X^2 şi Y^2 rezultă imediat din cele ale lui X şi Y:

$$X^2:\begin{pmatrix}1\\ 1\end{pmatrix},\ Y^2:\begin{pmatrix}0 & 1\\ 1/3 & 2/3\end{pmatrix}.$$

Aplicația 2.5.3. *Fie variabila aleatoare discretă* $X:\begin{pmatrix}1 & 2 & 3 & 3 & 3\\ p & p^2 & p & p^2 & p^2\end{pmatrix}$.

a) Să se determine p.
b) Să se calculeze funcția de repartiție a lui X.
c) Să se calculeze probabilitățile:

$$P(X<1),P(X<3),P(X>4),P(1{,}5<X<3{,}2),P(X\geq 2{,}1),$$

$$P(3{,}1<X\,|\,X>2{,}8),P(1{,}5<X\leq 3\,|\,2<X<4).$$

Rezolvare

a) Trebuie să avem $p+p^2+p+p^2+p^2=1$ şi $p\geq 0$. Rezultă $3p^2+2p=1$ şi $p\geq 0$, adică $p=1/3$.

b) Cum $F(x)=P(X\leq x)$ rezultă că $F(x)=0$ dacă $x\leq 1$,

$F(x)=p=1/3$ dacă $x\in(1,2]$,

$F(x)=P(X=1)+P(X=2)=p+p^2=4/9$ dacă $x\in(2,3]$,

F(x) = P(X=1) + P(X=2) + P(X=3) = p + p^2 + p = 7/9 dacă x $\in$ (3,4],

F(x) = P(X=1) + P(X=2) + P(X=3) + P(X=4) = p + p^2 + p + p^2 = 8/9 dacă x $\in$ (4,5] şi F(x) = 1 dacă x > 5 .

Deci :

$$F(x)=\begin{cases}0, & x\le 1;\\ \frac{1}{3}, & 1<x\le 2;\\ \frac{4}{9}, & 2<x\le 3;\\ \frac{7}{9}, & 3<x\le 4;\\ \frac{8}{9}, & 4<x\le 5;\\ 1, & x>5.\end{cases}$$

c) Avem

P(X<1) = P(Φ) = 0, P(X<3)=P(X=1)+P(X=2)=4/9,

P(X>4)=P(X=5)=1/9, P(1,5<X<3,2)=P(X=2)+P(X=3)=4/9,

P(X≥2,1)=P(X=3)+P(X=4)+P(X=5)=5/9,

$$P(3{,}1<X\,|\,X>2{,}8)=\frac{P(3{,}1<X,\,X>2{,}8)}{P(X>2{,}8)}=\frac{P(X>3{,}1)}{P(X>2{,}8)}=\frac{2p^2}{p+2p^2}=\frac{2}{5}$$

P(1,5<X≤3/2<X<4)=P(X=2 sau X=3 / X=3) = P(X=3/X=3) =1.

Aplicația 2.5.4. *Determinați constanta $a \in R$ pentru ca funcția f dată mai jos să fie densitate de repartiție şi apoi să se determine funcția de repartiție corespunzătoare. Să se calculeze mediana, cuantilele şi valoarea modală a variabilei aleatoare X având densitatea de probabilitate ρ(x):*

$$\rho(x)=\begin{cases}2x, & x\in[0,1/2]\\ a-\frac{2x}{3}, & x\in(1/2,2]\\ 0, & altfel.\end{cases}$$

Rezolvare

Avem $\int_{-\infty}^{+\infty}\rho(x)dx=1$, de unde $\int_0^{1/2}xdx+\int_{1/2}^{2}\left(a-\frac{2x}{3}\right)dx=1$ sau

$x^2\Big|_0^{1/2}+\left(ax-\frac{x^2}{3}\right)\Big|_{1/2}^{2}=1$. Rezultă a = 4/3 şi deci

$$F(x)=\int_{-\infty}^{x}\rho(t)dt=\begin{cases}0, & x<0\\ x^2, & x\in[0,1/2]\\ \frac{1}{4}+\int_{1/2}^{x}\frac{4-2t}{3}dt, & x\in(1/2,2]\\ 1, & x>2\end{cases}=\begin{cases}0, & x<0\\ x^2, & x\in[0,1/2]\\ \frac{4x-x^2-1}{3}, & x\in(1/2,2]\\ 1, & x>2\end{cases}$$

Întrucât F este continuă vom avea F(x) = F(x+0) , $\forall x$, deci

$F(M_e(X)) \le \frac{1}{2}$ şi $F(M_e(X)) \ge \frac{1}{2}$, adică $F(M_e(X)) = \frac{1}{2}$. Aceasta se realizează pentru $\frac{4x - x^2 - 1}{3} = \frac{1}{2}$, de unde $x = 2 - \sqrt{\frac{3}{2}} \in (1/2, 2]$.

Aşadar $M_e(X) = 2 - \sqrt{\frac{3}{2}} = c_2$. Se observă că F(1/2)=1/4 deci $c_1 = \frac{1}{2} c_3$ rezultă din F(x)=3/4, adică $\frac{4x - x^2 - 1}{3} = \frac{3}{4}$, de unde $c_3 = 2 - \frac{\sqrt{3}}{2}$.

Deoarece ρ este crescătoare pe [0,1/2] şi descrescătoare pe (1/2,2], x=1/2 este punct de maxim (singurul), prin urmare $M_o(X) = \frac{1}{2}$.

Aplicaţia 2.5.5. *Să se determine variabilele aleatoare independente*

$$X : \begin{pmatrix} x & x+1 & x+2 & x+3 \\ p & 2p & 3p & 4p \end{pmatrix} \text{ şi } Y : \begin{pmatrix} y & 2y & 3y \\ q & q^2 & q^2 \end{pmatrix},$$

ştiind că E(X)=2 şi E(Y)=7. Să se calculeze apoi E(2X+3Y), Var(X), Var(Y) şi Var(2X+3Y).

Rezolvare

Deoarece X este o variabilă aleatoare trebuie să avem p+2p+3p+4p=1, adică p=1/10. Atunci

E(X) = x·p+(x+1)·2p+(x+2)·3p+(x+3)·4p = 10px+20p = x+2.

Cum E(X) = 2 rezultă că x = 0.

Analog $q+q^2+q^2=1$, adică $2q^2+q-1=0$, de unde q=1/2. Rezultă că

$E(Y)=y \cdot q+2y \cdot q^2+3y \cdot q^2=\frac{7}{4}y$. Cum E(Y) = 7 avem că y=4. Tablourile de repartiţie ale lui X şi Y vor fi

$$X : \begin{pmatrix} 0 & 1 & 2 & 3 \\ \frac{1}{10} & \frac{1}{5} & \frac{3}{10} & \frac{2}{5} \end{pmatrix}, \quad Y : \begin{pmatrix} 4 & 8 & 12 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}.$$

Folosind proprietăţile mediei avem

E(2X+3Y) = E(2X)+E(3Y) = 2E(X)+3E(Y) = 2·2+3·7 = 25.

Pentru calcularea dispersiilor avem nevoie să calculăm mediile lui X^2 şi Y^2. Acestea au tablourile de repartiţie

$$X^2 : \begin{pmatrix} 0 & 1 & 4 & 9 \\ \frac{1}{10} & \frac{1}{5} & \frac{3}{10} & \frac{2}{5} \end{pmatrix}, \quad Y^2 : \begin{pmatrix} 16 & 64 & 144 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}, \text{ astfel că}$$

$$E(X^2) = 0 \cdot \frac{1}{10} + 1 \cdot \frac{1}{5} + 4 \cdot \frac{3}{10} + 9 \cdot \frac{2}{5} = 5 \text{ şi}$$

$E(Y^2) = 16 \cdot \frac{1}{2} + 64 \cdot \frac{1}{4} + 144 \cdot \frac{1}{4} = 60$. Atunci

Var(X) = $E(X^2)-E(X)^2$ = 5-4 = 1 şi Var(Y) = $E(Y^2)-E(Y)^2$ = 60-49 = 11.

Cum X şi Y sunt independente, rezultă că şi 2X şi 3Y sunt independente şi avem

Var(2X+3Y) = Var(2X)+Var(3Y) = 4Var(X)+9Var(Y) = 4·1+9·11 = 103.

Aplicația 2.5.6. *Să se determine variabilele aleatoare X şi Y ale căror repartiții sunt date incomplet în tabelul de mai jos, ştiind că E(X)=17 şi Var(Y)=1. Să se calculeze apoi E(XY) şi Var(X-Y).*

X \ Y	-b	0	b	pi
a	1/5	1/10		
a^2		2/5		3/5
qj			1/5	

Rezolvare

Deoarece $p_1+p_2=1$ rezultă că $p_1 = 1-\frac{3}{5} = \frac{2}{5}$. Mai departe

$p_{11}+p_{12}+p_{13}=p_1$, adică $\frac{1}{5}+\frac{1}{10}+p_{13} = \frac{2}{5}$, deci $p_{13} = \frac{1}{10}$. Cum

$p_{13}+p_{23}=q_3$ rezultă că $p_{23} = \frac{1}{5}-\frac{1}{10} = \frac{1}{10}$. Dar $p_{21}+p_{22}+p_{23}=p_2$, adică $p_{21} = \frac{3}{5}-\frac{2}{5}-\frac{1}{10} = \frac{1}{10}$. Din $p_{11}+p_{21}=q_1$ şi $p_{22}+p_{12}=q_2$, rezultă că $q_1 = \frac{3}{10}$ şi $q_2 = \frac{1}{2}$. Obținem astfel

$X : \begin{pmatrix} a & a^2 \\ \frac{2}{5} & \frac{3}{5} \end{pmatrix}$, $Y : \begin{pmatrix} -b & 0 & b \\ \frac{3}{10} & \frac{1}{2} & \frac{1}{5} \end{pmatrix}$. Astfel $E(X) = a \cdot \frac{2}{5} + a^2 \cdot \frac{3}{5} = 17$, adică $3a^2+2a-85=0$, de unde $a_1=5$, $a_2=-17/3$. Deoarece $E(Y) = -\frac{3b}{10} + 0 \cdot \frac{1}{2} + \frac{b}{5} = -\frac{b}{10}$ şi

$E(Y^2) = (-b)^2 \cdot \frac{3}{10} + 0^2 \cdot \frac{1}{2} + b^2 \cdot \frac{1}{5} = \frac{b^2}{2}$, rezultă că

$Var(Y) = E(Y^2) - E(Y)^2 = \frac{b^2}{2} - \frac{b^2}{100} = \frac{49b^2}{100}$. Din ipoteză Var(Y)=1, astfel că $b^2 = \frac{100}{49}$, adică $b = \frac{10}{7}$. Din tabloul repartiției comune (p_{ij}) avem

$XY:\begin{pmatrix} -a^2b & -ab & 0 & ab & a^2b \\ \frac{1}{10} & \frac{1}{5} & \frac{1}{2} & \frac{1}{10} & \frac{1}{10} \end{pmatrix}$ şi $X-Y:\begin{pmatrix} a-b & a^2-b & a & a^2 & a+b & a^2+b \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{2}{5} & \frac{1}{5} & \frac{1}{10} \end{pmatrix}$

Astfel $E(XY)=-\frac{a^2b}{10}-\frac{ab}{5}+\frac{ab}{10}+\frac{a^2b}{10}=-\frac{ab}{10}=-\frac{a}{7}$.

Dacă a=5, E(XY)=-5/7, iar dacă a=-17/3, E(XY)=17/21. Pentru calcularea dispersiei lui X-Y, avem nevoie de:

$$E(X-Y)=\frac{a-b}{10}+\frac{a^2-b}{10}+\frac{a}{10}+\frac{2a^2}{5}+\frac{a+b}{5}+\frac{a^2+b}{10}=\frac{6a^2+4a+b}{10}$$

$$E[(X-Y)^2]=\frac{(a-b)^2}{10}+\frac{(a^2-b)^2}{10}+\frac{a^2}{10}+\frac{2a^4}{5}+\frac{(a+b)^2}{5}+\frac{(a^2+b)^2}{10}=$$

$$=\frac{6a^4+4a^2+2ab+5b^2}{10}$$

Pentru a=5, b=10/7 avem E(X-Y)=$\frac{120}{7}$ şi $E[(X-Y)^2]=\frac{18985}{49}$,

deci $Var(X-Y)=\frac{18985}{49}-\frac{14400}{49}=\frac{4585}{49}$.

Aplicația 2.5.7. *Să se determine parametrii care apar în repartițiile următoare şi să se calculeze apoi E(X) şi Var(X), X fiind o variabilă aleatoare, având repartiția respectivă:*

a) $X:\begin{pmatrix} n \\ \frac{1}{4}q^n \end{pmatrix}, n\in N, q>0$ *;*

b) $X:\begin{pmatrix} x \\ a(x^2+2x) \end{pmatrix}, x\in[0,1], a>0$ *;*

c) $X:\begin{pmatrix} x \\ \rho(x) \end{pmatrix}, \rho(x)=\begin{cases} a^2x^3, & x\in[0,1] \\ \frac{1}{2}ax, & x\in(1,2] \\ 0, & altfel. \end{cases}$

Rezolvare

a)Din condiția $\sum_{n=0}^{\infty}\frac{1}{4}q^n=1$ rezultă că $\frac{1}{4}\cdot\frac{1}{1-q}=1\Rightarrow 1-q=\frac{1}{4}\Rightarrow q=\frac{3}{4}$

$$E(X)=\sum_{n=0}^{\infty}n\cdot\frac{1}{4}q^n=\frac{1}{4}q\sum_{n=0}^{\infty}n\cdot q^{n-1}=\frac{1}{4}q\sum_{n=0}^{\infty}(q^n)'=\frac{1}{4}q\left(\sum_{n=0}^{\infty}q^n\right)'=$$

Atunci

$$=\frac{1}{4}q\cdot\left(\frac{1}{1-q}\right)'=\frac{1}{4}q\frac{1}{(1-q)^2}=\frac{1}{4}\cdot\frac{3}{4}\cdot\frac{1}{\frac{1}{16}}=3$$

$$E(X^2)=\sum_{n=0}^{\infty}n^2\cdot\frac{1}{4}q^n=\frac{1}{4}q\sum_{n=0}^{\infty}n(q^n)'=\frac{1}{4}q\left(\sum_{n=0}^{\infty}nq^n\right)'=q\left(\frac{1}{4}\sum_{n=0}^{\infty}nq^n\right)'=$$

$$=\frac{1}{4}q\left[\frac{1}{(1-q)^2}\right]'=\frac{1}{4}q\frac{2}{(1-q)^3}=24.$$

Astfel Var(X) = $E(X^2)$-$E(X)^2$ = 24-9 = 15.

b) Din condiția $\int_0^1 a(x^2+2x)dx=1$ rezultă că

$a\left(\frac{x^3}{3}+x^2\right)\Big|_0^1=1\Rightarrow a\cdot\frac{4}{3}=1\Rightarrow a=\frac{3}{4}$. Atunci

$$E(X)=\int_0^1 x\cdot\rho(x)dx=\int_0^1 x\cdot\frac{3}{4}(x^2+2x)dx=\frac{3}{4}\left(\frac{x^4}{4}+\frac{2x^3}{3}\right)\Big|_0^1=\frac{11}{16},$$

$$E(X^2)=\int_0^1 x^2\cdot\rho(x)dx=\int_0^1 x^2\cdot\frac{3}{4}(x^2+2x)dx=\frac{3}{4}\left(\frac{x^5}{5}+\frac{2x^4}{4}\right)\Big|_0^1=\frac{21}{40}.$$

Astfel $Var(X)=E(X^2)-E(X)^2=\frac{21}{40}-\left(\frac{11}{16}\right)^2=\frac{67}{1280}$.

c) Din condiția $\int_0^2\rho(x)dx=1$ rezultă că

$\int_0^1 a^2x^3dx+\int_1^2\frac{ax}{2}dx=1\Rightarrow a^2\cdot\frac{x^4}{4}\Big|_0^1+a\cdot\frac{x^2}{4}\Big|_1^2=1\Rightarrow\frac{a^2}{4}+\frac{3a}{4}=1\Rightarrow\begin{cases}a=1\\a=4\end{cases}$ Cum

$\rho(x)\geq 0$ numai a=1 convine, astfel că :

$$E(X)=\int_0^2 x\cdot\rho(x)dx=\int_0^1 x\cdot x^3dx+\int_1^2 x\cdot\frac{x}{2}dx=\frac{x^5}{5}\Big|_0^1+\frac{x^3}{6}\Big|_1^2=\frac{1}{5}+\frac{7}{6}=\frac{41}{30}$$

$$E(X^2)=\int_0^2 x^2\cdot\rho(x)dx=\int_0^1 x^2\cdot x^3dx+\int_1^2 x^2\cdot\frac{x}{2}dx=\frac{x^6}{6}\Big|_0^1+\frac{x^4}{8}\Big|_1^2=\frac{49}{24},$$

$$Var(X)=E(X^2)-E(X)^2=\frac{49}{24}-\left(\frac{41}{30}\right)^2=\frac{313}{1800}.$$

Aplicația 2.5.8. *Fie vectorul aleator (X,Y) cu X, Y variabile aleatoare independente, a cărui densitate de probabilitate este*

$\rho(x,y)=\frac{a}{(1+x^2)(1+y^2)}$ *. Să se determine:*

a) funcția de repartiție corespunzătoare;

b) $P((X,Y)\in[0,1)\times[0,1))$.

Rezolvare

a) Avem $\iint_{R^2}\rho(x,y)dxdy=1$, deci $\int_{-\infty}^{+\infty}\frac{a}{(1+x^2)(1+y^2)}dxdy=1$. Rezultă că $a\int_{-\infty}^{+\infty}\frac{dx}{1+x^2}\int_{-\infty}^{+\infty}\frac{dy}{1+y^2}=1\Rightarrow a\cdot arctgx\Big|_{-\infty}^{+\infty}\cdot arctgy\Big|_{-\infty}^{+\infty}=1\Rightarrow a=\frac{1}{\pi^2}$.

Atunci $F(x,y)=\frac{1}{\pi^2}\int_{-\infty}^{x}\int_{-\infty}^{y}\frac{dudv}{(1+u^2)(1+v^2)}=\frac{1}{\pi^2}\int_{-\infty}^{x}\frac{du}{1+u^2}\int_{-\infty}^{y}\frac{dv}{1+v^2}=$

$=\left(\frac{1}{\pi}arctgx+\frac{1}{2}\right)\left(\frac{1}{\pi}arctgy+\frac{1}{2}\right)$

b) $P((X,Y)\in D)=\frac{1}{\pi^2}\int_0^1\int_0^1\frac{dudv}{(1+u^2)(1+v^2)}=$

$=F(1,1)-F(1,0)-F(0,1)+F(0,0)=\frac{1}{16}$

Aplicația 2.5.9. *Fie vectorul aleator (X,Y) cu densitatea de probabilitate*

$$\rho(x,y)=\begin{cases}ax^2y, & (x,y)\in[0,1]\times[0,2]\\ 0, & altfel.\end{cases}$$

a) Să se determine constanta a.

b) Să se calculeze funcția de repartiție F(x,y) şi funcțiile marginale $F_X(x)$ şi $F_Y(y)$.

Rezolvare

a) Avem $\int_0^1\int_0^2\rho(x,y)dxdy=1$, adică $a\int_0^1x^2dx\int_0^2ydy=1$, rezultă că $\frac{2a}{3}=1$ de unde $a=\frac{3}{2}$.

b) Avem $F(x,y)=\int_{-\infty}^{x}\int_{-\infty}^{y}\rho(x,y)dxdy$.

Dacă x < 0 sau y < 0, atunci f(x,y)=0 şi deci F(x,y)=0.

Dacă x > 1 şi y > 2, atunci F(x,y)=1.

Dacă (x,y) $\in[0,1]\times[0,2]$ avem

$$F(x,y)=\int_0^x\int_0^y\frac{3}{2}u^2vdudv=\frac{3}{2}\int_0^xu^2du\int_0^yvdv=\frac{x^3y^2}{4}.$$

Dacă x $\in[0,1]$ şi y > 2 avem

$$F(x,y)=\int_0^x\int_0^2\frac{3}{2}u^2vdudv=\frac{3}{2}\int_0^xu^2du\int_0^2vdv=x^3.$$

Dacă x > 1 şi y $\in[0,2]$ obținem

$$F(x,y)=\int_0^1\int_0^y\frac{3}{2}u^2vdudv=\frac{3}{2}\int_0^1u^2du\int_0^yvdv=\frac{y^2}{4}.$$

Astfel $F(x,y)=\begin{cases} 0, & x<0 \text{ sau } y<0 \\ \dfrac{x^3y^2}{4}, & (x,y)\in[0,1]\times[0,2] \\ x^3, & x\in[0,1],\ y>2 \\ \dfrac{y^2}{4}, & x>1,\ y\in[0,2] \\ 1, & x>1,\ y>2 \end{cases}$

Funcțiile de repartiție marginale sunt

$$F_X(x)=F(x,\infty)=\begin{cases} 0, & x<0 \\ x^3, & x\in[0,1] \\ 1, & x>1 \end{cases}$$

$$F_Y(y)=F(\infty,y)=\begin{cases} 0, & y<0 \\ \dfrac{y^2}{4}, & y\in[0,2] \\ 1, & y>2 \end{cases}$$

Aplicația 2.5.10. *Fie vectorul aleator (X,Y) având densitatea de probabilitate*

$$\rho(x,y)=\begin{cases} e^{-(x+y)}, & x\ge 0,\ y\ge 0 \\ 0, & altfel \end{cases}$$

Să se calculeze:

a) P(X<1,Y<1), P(X+Y<1), P(X+Y≥2), P(X≥1/Y≥1), P(X<2Y), P(X=n) ;

b) Funcția de repartiție F(x,y) şi funcțiile de repartiție marginale $F_X(x)$, $F_Y(y)$;

c) Densitățile de repartiție marginale $\rho_X(x)$, $\rho_Y(y)$;

d) Momentele obişnuite de ordin (k,s) ;

e) Corelația variabilelor X şi Y.

Rezolvare

a) $P(X<1,Y<1)=\int_{-\infty}^{1}\int_{-\infty}^{1}\rho(x,y)dxdy=$

$=\int_0^1\int_0^1 e^{-x-y}dxdy=(-e^{-x})\Big|_0^1(-e^{-y})\Big|_0^1=(1-e^{-1})^2$

$P(X+Y<1)=\iint\limits_{x+y<1}\rho(x,y)dxdy=\int_0^1\int_0^{1-x}e^{-x-y}dxdy=\int_0^1 e^{-x}(-e^{-y})\Big|_0^{1-x}dx=$

$=\int_0^1(e^{-x}-e^{-1})dx=(-e^{-x})\Big|_1^0-e^{-1}=1-2e^{-1}$,

$P(X+Y\ge 2)=1-P(X+Y<2)=1-\iint\limits_{x+y<2}\rho(x,y)dxdy=$

$=1-\int_0^2\int_0^{2-x}e^{-x-y}dxdy=1-\int_0^2 e^{-x}(e^{-y})\Big|_0^{2-x}dx=1-\int_0^2(e^{-x}-e^{-2})dx=3e^{-2}$

$$P(X \geq 1/Y \geq 1) = \frac{P(X \geq 1, Y \geq 1)}{P(Y \geq 1)}$$

$$P(X \geq 1, Y \geq 1) = \int_1^{+\infty}\int_1^{+\infty} e^{-x-y}dxdy = \left(\int_1^{+\infty} e^{-x}dx\right)^2 = \left((-e^{-x})\Big|_1^{+\infty}\right)^2 = e^{-2}$$

$$P(Y \geq 1) = \int_0^{+\infty}\int_1^{+\infty} e^{-x-y}dxdy = -e^{-x}\Big|_0^{+\infty}(-e^{-y})\Big|_1^{+\infty} = e^{-1}$$

$$\Rightarrow P(X \geq 1, Y \geq 1) = e^{-1}$$

$$P(X < 2Y) = \iint\limits_{0<x<2y} e^{-x-y}dxdy = \int_0^{+\infty}\int_0^{\frac{x}{2}} e^{-x}e^{-y}dydx = \int_0^{+\infty} e^{-x}(-e^{-y})\Big|_0^{x/2}dx =$$

$$= \int_0^{+\infty}(e^{-x} - e^{\frac{-3x}{2}})dx = \left(-e^{-x} + \frac{2}{3}e^{\frac{-3x}{2}}\right)\Big|_0^{+\infty} = 1 - \frac{2}{3} = \frac{1}{3}, \text{ P(X=Y)=0.}$$

b) Avem $F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} \rho(u,v)dudv$.

Dacă x > 0 sau y < 0 ρ(x,y)=0, deci F(x,y)=0. Dacă x ≥0 şi y ≥0 avem

$F(x,y) = \int_0^x\int_0^y e^{-u-v}dudv = \int_0^x e^{-u}du\int_0^y e^{-v}dv = (1-e^{-x})(1-e^{-y})$.

Funcţiile de repartiţie marginale sunt

$F_X(x) = F(x,\infty) = 1-e^{-x}$ şi $F_Y(y) = F(\infty,y) = 1-e^{-y}$.

c) Densităţile de probabilitate marginale $\rho_X(x)$ şi $\rho_Y(y)$ sunt derivatele funcţiilor de repartiţie marginale:

$$\rho_X(x) = \begin{cases} 0, & x<0 \\ e^{-x}, & x \geq 0 \end{cases}, \quad \rho_Y(y) = \begin{cases} 0, & y<0 \\ e^{-y}, & y \geq 0 \end{cases}.$$

d) $\sigma_{k,s} = M(X^kY^s) = \int_0^{+\infty}\int_0^{+\infty} x^k y^s e^{-x-y}dxdy =$

$= \int_0^{+\infty} x^k e^{-x}dx\int_0^{+\infty} y^s e^{-y}dy = \Gamma(k+1)\Gamma(s+1) = k!s!$,

unde $\Gamma(p) = \int_0^{+\infty} x^{p-1}e^{-x}dx$ este funcţia gama a lui Euler şi are proprietatea că $\Gamma(p+1) = p!$ pentru $p \in N$.

e) Avem

$\text{cov}(X,Y) = E(XY) - E(X)E(Y) = m_{11} - \int_0^{+\infty} xe^{-x}dx\int_0^{+\infty} ye^{-y}dy =$

$= 1 - \Gamma(2)^2 = 1-1 = 0$.

Aplicaţia 2.5.11. *Fie (X,Y) un vector aleator discret a cărui repartiţie probabilistă este dată în tabelul de mai jos. Să se calculeze coeficientul de corelaţie r(X,Y) şi să se scrie ecuaţiile dreptelor de regresie.*

X \ Y	-1	0	1	2	p_i
-1	1/10	1/5	1/10	0	2/5
0	1/20	0	1/10	1/20	1/5
1	1/10	1/10	1/20	3/20	2/5
q_j	1/4	3/10	1/4	1/5	1

Rezolvare

Pe baza formulelor corespunzătoare, deducem imediat:

$$E(X)=-1\cdot\frac{2}{5}+0\cdot\frac{1}{5}+1\cdot\frac{2}{5}=0,\ E(Y)=-1\cdot\frac{1}{4}+0\cdot\frac{3}{10}+1\cdot\frac{1}{4}+2\cdot\frac{1}{5}=\frac{2}{5},$$

$$E(X^2)=(-1)^2\cdot\frac{2}{5}+0^2\cdot\frac{1}{5}+1^2\cdot\frac{2}{5}=\frac{4}{5}\ ,$$

$$E(Y^2)=(-1)^2\cdot\frac{1}{4}+0^2\cdot\frac{3}{10}+1^2\cdot\frac{1}{4}+2^2\cdot\frac{1}{5}=\frac{13}{10}\ ,$$

$$Var(X)=E(X^2)-E(X)^2=\frac{4}{5},$$

$$Var(Y)=E(Y^2)-E(Y)^2=\frac{13}{10}-\frac{4}{25}=\frac{57}{50}$$

$$E(XY)=(-1)\cdot(-1)\cdot\frac{1}{10}+(-1)\cdot0\cdot\frac{1}{5}+(-1)\cdot1\cdot\frac{1}{10}+(-1)\cdot2\cdot0+0\cdot(-1)\cdot\frac{1}{20}+$$

$$+0\cdot0\cdot0+0\cdot1\cdot\frac{1}{10}+0\cdot2\cdot\frac{1}{20}+1\cdot(-1)\cdot\frac{1}{10}+1\cdot0\cdot\frac{1}{10}+1\cdot1\cdot\frac{1}{20}+1\cdot2\cdot\frac{3}{20}=\frac{1}{4}$$

$$\mathrm{cov}(X,Y)=E(XY)-E(X)E(Y)=\frac{1}{4},$$

$$r(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sqrt{Var(X)Var(Y)}}=\frac{0{,}25}{\frac{2}{\sqrt{5}}\cdot\frac{\sqrt{57}}{\sqrt{50}}}\ .$$

Ecuațiile dreptelor de regresie sunt :

$$\frac{x-0}{\frac{4}{5}}=+0{,}25\cdot\frac{y-\frac{2}{5}}{\frac{57}{50}}\ ,\quad \frac{y-\frac{2}{5}}{\frac{57}{50}}=+0{,}25\cdot\frac{x-0}{\frac{4}{5}}\ .$$

Aplicația 2.5.12. *Să se determine funcția caracteristică şi funcția generatoare de momente şi apoi să se calculeze, pornind de la acestea, momentele* m_1 *şi* m_2*, pentru următoarele variabile aleatoare:*

a) $X:\begin{pmatrix}0 & 1 & 2\\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4}\end{pmatrix}$ *;*

b) $X:\begin{pmatrix}x\\ e^{-x}\end{pmatrix}, x\geq 0$.

Rezolvare

a) Avem

$$\varphi_X(t)=e^{0\cdot it}\cdot\frac{1}{2}+e^{1\cdot it}\cdot\frac{1}{4}+e^{2\cdot it}\cdot\frac{1}{4}=\frac{2+e^{it}+e^{2it}}{4}\ \text{şi}$$

$G_X(t) = \frac{1}{2}e^{0\cdot t} + \frac{1}{4}e^{1\cdot t} + \frac{1}{4}e^{2\cdot t} = \frac{2+e^t+e^{2t}}{4}$.

Primele două derivate ale acestor funcții sunt:

$\varphi_X^{'}(t) = \frac{i}{4}(e^{it} + 2e^{2it})$, $\varphi_X^{''}(t) = -\frac{1}{4}(e^{it} + 4e^{2it})$,

$G_X^{'}(t) = \frac{1}{4}(e^t + 2e^{2t})$, $G_X^{''}(t) = \frac{1}{4}(e^t + 4e^{2t})$.

Obținem

$\varphi_X^{'}(0) = \frac{3i}{4}, \varphi_X^{''}(0) = -\frac{5}{4}$, $G_X^{'}(0) = \frac{3}{4}, G_X^{''}(0) = \frac{5}{4}$, de unde

$\sigma_1(X) = E(X) = \frac{\varphi_X^{'}(0)}{i} = G_X^{'}(0) = \frac{3}{4}$ şi

$\sigma_2(X) = E(X^2) = \frac{\varphi_X^{''}(0)}{i^2} = G_X^{''}(0) = \frac{5}{4}$.

b) $\varphi_X(t) = \int_0^{+\infty} e^{itx}\cdot e^{-x}dx = \int_0^{+\infty}(\cos tx + i\sin tx)\cdot e^{-x}dx =$

$= \int_0^{+\infty}\cos tx\cdot e^{-x}dx + i\int_0^{+\infty}\sin tx\cdot e^{-x}dx = A + iB$

$A = \int_0^{+\infty}\cos tx\cdot e^{-x}dx = \int_0^{+\infty}\cos tx\cdot(-e^{-x})'dx =$

$= -\cos tx\cdot e^{-x}\Big|_0^{+\infty} - \int_0^{+\infty} t\sin tx\cdot e^{-x}dx = 1 - tB$,

$B = \int_0^{+\infty}\sin tx\cdot(-e^{-x})dx = -\sin tx\cdot e^{-x}\Big|_0^{+\infty} + \int_0^{+\infty} t\cos tx\cdot e^{-x}dx = tA$.

Obținem A = 1 - t^2A, adică A = $\frac{1}{1+t^2}$ şi B = $\frac{t}{1+t^2}$.

Astfel $\varphi_X(t) = \frac{1+it}{1+t^2}$.

Apoi $G_X(t) = \int_0^{+\infty} e^{tx}e^{-x}dx = \int_0^{+\infty} e^{x(t-1)}dx = \frac{e^{x(t-1)}}{t-1}\Big|_0^{+\infty} = \frac{1}{1-t}, t<1$.

Primele două derivate sunt

$\varphi_X^{'}(t) = \frac{i-2t-it^2}{(1+t^2)^4}$, $\varphi_X^{''}(t) = \frac{6it^3+14t^2-10it-2}{(1+t^2)^5}$,

$G_X^{'}(t) = \frac{1}{(1-t)^2}$, $G_X^{''}(t) = \frac{2}{(1-t)^3}$.

Obținem

$m_1(X) = E(X) = \frac{\varphi_X^{'}(0)}{i} = G_X^{'}(0) = 1$ şi

$m_2(X) = E(X^2) = \frac{\varphi_X^{''}(0)}{i^2} = G_X^{''}(0) = 2$.

2.6. Probleme propuse

Aplicația 2.6.1. *Se consideră vectorul aleator (X,Y) cu densitatea de probabilitate:* $\rho(x,y)=\begin{cases} A\sqrt{xy}, daca \quad x,y\rangle 0, y\langle x\langle 1 \\ 0, \quad altfel \end{cases}$. *Să se determine:*

a) *constanta reală A;*
b) *densitățile de probabilitate* ρ_X, ρ_Y *pentru variabilele aleatoare X, Y;*
c) *probabilitățile* $P(0<X<\frac{1}{2}, 0<Y<\frac{1}{2})$ *şi* $P(X<\frac{1}{2}\,/Y<\frac{1}{2})$.

Aplicația 2.6.2. *La patru unități alimentare din oraş se poate găsi zilnic pâine proaspătă cu probabilitățile* $p_1=0.8$, $p_2=0.9$, $p_3=0.95$ *şi respectiv* $p_4=0.85$. *Fie X numărul unităților alimentare din cele patru la care se găseşte pâine proaspătă într-o zi fixată. Să se determine:*

a) *distribuția variabilei aleatoare X;*
b) *valoarea medie, dispersia, abaterea medie pătratică, mediana şi modul variabilei aleatoare X.*

Aplicația 2.6.3. *Fie (X,Y) coordonatele unui punct luminos ce reprezintă o țintă pe un ecran radar circular şi care urmează legea uniformă pe domeniul* $D = \{(x,y)\in R^2 / x^2+y^2 \le r^2\}$. *Să se determine valoarea medie şi dispersia distanței* $Z = \sqrt{X^2+Y^2}$ *de la centrul ecranului până la punctul luminos.*

Aplicația 2.6.4. *Folosind inegalitatea lui Cebîşev, să se arate că*

$$P(0<X<2(m+1)) \ge \frac{m}{m+1},$$

dacă variabila aleatoare X are densitatea de probabilitate

$$\rho(x)=\begin{cases} \frac{x^m}{m!}e^{-x}, daca \quad x\rangle 0 \\ 0, daca \quad x\le 0 \end{cases}.$$

Aplicația 2.6.5. *Probabilitatea ca o persoană să găsească loc la un hotel este p = 0.8. În decursul unei luni de zile, la hotelul respectiv s-au prezentat 4000 de persoane. Fie X numărul persoanelor care au găsit loc la hotel din totalul de 4000. Să se determine probabilitatea ca:*

a) *numărul persoanelor care au găsit loc la hotel să fie cuprins între 3000 şi 3400;*
b) *numărul persoanelor care au găsit loc la hotel să nu depăşească 3000;*
c) *numărul persoanelor care nu au găsit loc la hotel să fie mai mic decât 500.*

Aplicația 2.6.6. *Fie variabilele aleatoare independente:*

$X:\begin{pmatrix} 0 & 1 & 2 \\ 1/6 & 1/2 & 1/3 \end{pmatrix}$ *şi* $Y:\begin{pmatrix} -1 & 0 & 1 & 2 \\ 1/8 & 1/2 & 1/4 & 1/8 \end{pmatrix}$

Să determine variabilele aleatoare: X+Y; X-Y; X·Y; X^2; Y^2; X^3; Y^3; *2X; 3Y; 2X+3Y; 3Y-2X;* $\sqrt{X}$.

Aplicația 2.6.7. *Fie X şi Y două variabile aleatoare discrete ale căror repartiții probabiliste comune* (p_{ij}) *şi unidimensionale* (p_i) *şi* (q_j) *sunt date în tabelul de mai jos:*

Y / X	-1	0	1	2	p_i
-1	1/12	1/24	1/24	1/48	3/16
1	1/48	1/24	1/48	1/24	1/8
2	1/48	1/3	1/6	1/6	11/16
q_j	1/8	5/12	11/48	11/48	1

a) *Scrieți variabilele aleatoare X şi Y;*
b) *Precizați dacă variabilele aleatore X şi Y sunt independente sau nu şi justificați răspunsul;*
c) *Scrieți variabilele aleatoare: X+Y; X-Y; X·Y;* $\frac{Y}{X}$; X^2; Y^3; *3X-2Y;*

Aplicația 2.6.7. *Fie variabilele aleatoare independente:*

$$X:\begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ \frac{1}{10} & \frac{1}{5} & \frac{2}{5} & \frac{1}{10} & \frac{1}{5} \end{pmatrix} \text{ şi } Y:\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 \\ \frac{1}{12} & \frac{1}{12} & \frac{1}{2} & \frac{1}{6} & \frac{1}{12} & \frac{1}{12} \end{pmatrix}$$

a) *Calculați E(X), E(*X^2*), Var(X), E(Y), E(*Y^2*) şi Var(Y);*
b) *Care dintre următoarele mărimi pot fi calculate şi care nu şi de ce?*
E(2X+3Y); Var(2X+3Y); E(X^2*+Y); Var(*X^2*+Y);*
E(X^2*+*Y^2*); Var(*X^2*+*Y^2*); E(XY).*
c) *Calculați mărimile de la punctul b) pentru care răspunsul este favorabil.*

Aplicația 2.6.8. *Să se determine, în fiecare caz, variabila aleatoare X şi apoi să se calculeze E(X) şi Var(X).*

a) $X:\begin{pmatrix} x \\ pq^x \end{pmatrix}$, $x\in N$, $p > 0$, $q > 0$, b) $X:\begin{pmatrix} x \\ \frac{ax}{x!}b \end{pmatrix}$, $x\in N$, $a > 0$, $b > 0$

c) $X:\begin{pmatrix} x \\ ax \end{pmatrix}$, $x\in[0,1]$, $a\in R$, d) $X:\begin{pmatrix} x \\ a(3x^2+2x) \end{pmatrix}$, $x\in[0,1]$, $a\in R$

e) $X:\begin{pmatrix} x \\ e^{a|x|} \end{pmatrix}$, $a \in R$, $x \in R$, *f)* $X:\begin{pmatrix} x \\ \rho(x) \end{pmatrix}$, $x \in R$,

$$\rho(x)=\begin{cases} ax & , x \in [0,1] \\ \dfrac{a^2x^2}{2} & , x \in [1,2],\ a \in R \\ 0 & , \text{în rest} \end{cases}$$

Aplicația 2.6.9. *Fie variabilele aleatoare discrete:* $X:\begin{pmatrix} 0 & 1 \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$ *şi* $Y:\begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} \end{pmatrix}$. *Dacă* $P(X=0,Y=0)=\lambda$ *şi* $P(X=1,Y=1)=\frac{1}{3}$, *să se determine repartiția comună a vectorului aleatoriu (X,Y) în funcție de* $\lambda \in R$. *Calculați apoi coeficientul de corelație* $r(X,Y)$ *şi precizați dacă există valori ale lui* λ *pentru care X şi Y să fie independente.*

Aplicația 2.6.10. *Fie* (X,Y) *un vector aleatoriu continuu cu densitatea de repartiție* $\rho(X,Y)=\begin{pmatrix} a(xy^2+x^2y) & , (x,y) \in [1,2]\times[2,3] \\ 0 & , \text{în rest} \end{pmatrix}$, $a > 0$

a) Determinați densitatea de repartiție $\rho(X,Y)$ *şi densitățile de repartiție marginale corespunzătoare* $\rho_X(x)$ *şi* $\rho_Y(y)$*;*

b) Calculați coeficientul de corelație $r(X,Y)$.

Aplicația 2.6.11. *Calculați funcția caracteristică şi funcția generatoare de momente pentru fiecare dintre variabilele aleatoare:*

a) $X:\begin{pmatrix} 1 & 2 & 3 \\ \frac{1}{6} & \frac{1}{3} & \frac{1}{2} \end{pmatrix}$, *b)* $X:\begin{pmatrix} -2 & -1 & 0 & 1 & 2 \\ \frac{1}{10} & \frac{1}{5} & \frac{2}{5} & \frac{1}{5} & \frac{1}{10} \end{pmatrix}$,

c) $X:\begin{pmatrix} x \\ 3x^2 \end{pmatrix}$, $x \in [0,1]$, *d)* $X:\begin{pmatrix} x \\ xe^{-x} \end{pmatrix}$, $x \geq 0$

şi apoi verificați dacă momentele obținute pe cale directă coincid cu cele obținute cu ajutorul acestor funcții.

Aplicația 2.6.12. *Verificați dacă funcțiile următoare definesc repartiții ale unor variabile aleatoare discrete şi apoi calculați E(X) şi Var(X) pentru fiecare dintre ele.*

a) $P(k)=\frac{1}{\theta}\frac{(\ln\theta)^k}{k!}, k \in N, \theta > 1$, *b)* $P(k)=\frac{1}{\theta^k \cdot k!}e^{-\frac{1}{\theta}}, k \in N, \theta > 0$,

c) $P(k)=\theta^k(1-\theta)^{1-k}, k \in \{0,1\}, 0 < \theta < 1$

Capitolul 3

Legi clasice de probabilitate (repartiții) ale variabilelor aleatoare discrete

Introducere

Vom prezenta în acest capitol principalele legi de probabilitate ale variabilelor aleatoare discrete, şi anume: legea discretă uniformă, legea binomială şi cazul său particular legea Bernoulli, legea binomială cu exponent negativ şi cazul particular legea geometrică, legea hipergeometrică şi legea Poisson (legea evenimentelor rare).

3.1. Legea discretă uniformă

Definiția 3.1.1. *Variabila aleatoare discretă X urmează* ***legea discretă uniformă*** *dacă are tabloul repartiției*

$$X:\begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}, \quad unde\ p_k = \frac{1}{n},\ k = 1,2,\dots n,\ n \in \mathrm{N}^*. \tag{3.1.1}$$

Vom mai spune că variabila aleatoare X dată de formula (3.1.1) are o repartiție discretă uniformă.

Din tabloul repartiției variabilei aleatoare X se observă că

$$\sum_{k=1}^{n} p_k = \sum_{k=1}^{n} \frac{1}{n} = 1.$$

Teorema 3.1.2. *Dacă variabila aleatoare X are repartiție discretă uniformă cu tabloul repartiției (3.1.1), atunci valoarea medie şi dispersia sa sunt*

$$E(X) = \frac{1}{n}\sum_{k=1}^{n} x_k, \quad Var(X) = \frac{1}{n}\sum_{k=1}^{n} x_k^2 - \frac{1}{n^2}\left(\sum_{k=1}^{n} x_k\right)^2. \tag{3.1.2}$$

Demonstrație

Din formulele de calcul ale mediei şi dispersiei obținem

$$E(X)=\sum_{k=1}^{n}x_k p_k=\sum_{k=1}^{n}x_k\cdot\frac{1}{n}=\frac{1}{n}\sum_{k=1}^{n}x_k,$$

$$Var(X)=E(X^2)-[E(X)]^2=\sum_{k=1}^{n}x_k^2 p_k-\left(\sum_{k=1}^{n}x_k p_k\right)^2$$

$$=\sum_{k=1}^{n}x_k^2\cdot\frac{1}{n}-\left(\sum_{k=1}^{n}x_k\cdot\frac{1}{n}\right)^2=\frac{1}{n}\sum_{k=1}^{n}x_k^2-\frac{1}{n^2}\left(\sum_{k=1}^{n}x_k\right)^2.$$

q.e.d.

Observația 3.1.3. *Deoarece* $Var(X)\geq 0$*, din relația a doua (3.1.2) avem*

$$n\sum_{k=1}^{n}x_k^2\geq\left(\sum_{k=1}^{n}x_k\right)^2,$$

relație utilă în diverse aplicații practice, şi care rezultă direct şi din inegalitatea lui Cauchy-Buniakovski-Schwarz.

Propoziția 3.1.4. *Dacă variabila aleatoare discretă* X *are repartiție uniformă cu tabloul repartiției (3.1.1), atunci funcția sa caracteristică este*

$$\varphi(t)=\frac{1}{n}\sum_{k=1}^{n}e^{itx_k},\quad t\in R. \tag{3.1.3}$$

Demonstrație
Conform formulei de calcul pentru funcția caracteristică, avem

$$\varphi(t)=\sum_{k=1}^{n}p_k e^{itx_k}=\frac{1}{n}\sum_{k=1}^{n}e^{itx_k},\quad \forall t\in R.q.e.d.$$

Propoziția 3.1.5. *Dacă variabila aleatoare discretă* X *are repartiție uniformă şi ia valorile* $x_k=k,\ k=1,2,...,n$*, adică are tabloul repartiției*

$$X:\begin{pmatrix}1 & 2 & \cdots & n\\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n}\end{pmatrix},$$

atunci

$$E(X)=\frac{n+1}{12},\ Var(X)=\frac{n^2-1}{12},$$

$$\varphi(t)=\frac{\sin\frac{nt}{2}}{n\sin\frac{t}{2}}e^{\frac{i(n+1)t}{2}},\ \forall t\in R,\ t\neq 2k\pi,\ k\in Z;\quad \varphi(t)=1,\ t=2k\pi,\ k\in Z.$$

Demonstrație

Din formulele (3.1.2), pentru $x_k=k,\ k=1,2,...,n,$ obținem

$$E(X)=\frac{1}{n}\sum_{k=1}^{n}x_k=\frac{1}{n}\sum_{k=1}^{n}k=\frac{n+1}{2},$$

$$Var(X)=\frac{1}{n}\sum_{k=1}^{n}x_k^2-\frac{1}{n^2}\left(\sum_{k=1}^{n}x_k\right)^2=\frac{1}{n}\sum_{k=1}^{n}k^2-\frac{1}{n^2}\left(\sum_{k=1}^{n}k\right)^2$$

$$=\frac{1}{n}\cdot\frac{n(n+1)(2n+1)}{6}-\frac{1}{n^2}\cdot\frac{n^2(n+1)^2}{4}=\frac{n^2-1}{12}.$$

Din formula (3.1.3), obținem pentru funcția caracteristică

$$\varphi(t)=\frac{1}{n}\sum_{k=1}^{n}e^{itk}=\frac{e^{it}}{n}\cdot\frac{1-e^{\text{int}}}{1-e^{it}}=\frac{e^{it}}{n}\cdot\frac{1-\cos nt-i\sin nt}{1-\cos t-i\sin t}$$

$$=\frac{e^{it}}{n}\cdot\frac{2\sin^2\frac{nt}{2}-2i\sin\frac{nt}{2}\cos\frac{nt}{2}}{2\sin^2\frac{t}{2}-2i\sin\frac{t}{2}\cos\frac{t}{2}}=\frac{e^{it}}{n}\cdot\frac{\sin\frac{nt}{2}\left(\cos\frac{nt}{2}+i\sin\frac{nt}{2}\right)}{\sin\frac{t}{2}\left(\cos\frac{t}{2}+i\sin\frac{t}{2}\right)}$$

$$=\frac{e^{it}\sin\frac{nt}{2}}{n\sin\frac{t}{2}}\left(\cos\frac{(n-1)t}{2}+i\sin\frac{(n-1)t}{2}\right)=\frac{\sin\frac{nt}{2}}{n\sin\frac{t}{2}}e^{\frac{i(n+1)t}{2}},\ \forall t\in R,\ t\neq 2k\pi,\ k\in Z,$$

$\varphi(t)=1,\ t=2k\pi,\ k\in Z.$

q.e.d.

3.2. Legea binomială. Legea Bernoulli

Definiția 3.2.1. *Variabila aleatoare discretă X urmează* ***legea binomială*** *(X are o repartiție binomială) cu parametrii n şi p (*$n\in N,\ \ 0<p<1$*) dacă ia valorile* $0,1,2,...,n$ *cu probabilitățile*

$$P(X=k)=C_n^k p^k q^{n-k},\ k=0,1,2,...,n, \qquad (3.2.1)$$

unde $q=1-p$.

Tabloul repartiției variabilei aleatoare X este

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

Se observă că $\sum_{k=0}^{n} C_n^k p^k q^{n-k} = \sum_{k=0}^{n} C_n^k p^{n-k} q^k = (p+q)^n = 1.$

Exemplul 3.2.2. *Dacă* $A_1, A_2, ..., A_n$ *sunt evenimente independente şi* $P(A_i)=p,\ i=1,2,...,n,$ *iar* X *reprezintă numărul evenimentelor care se realizează în cadrul unei experienţe , atunci* X *are repartiţie binomială cu parametrii n şi p (conform schemei lui Bernoulli).*

Exemplul 3.2.3. *Dacă* A *este un eveniment legat de o anumită experienţă şi probabilitatea ca* A *să se producă când efectuăm o singură dată experienţa este* $P(A)=p$, *atunci variabila aleatoare care are ca valori numărul realizărilor lui* A *când efectuăm de* n *ori experienţa are repartiţie binomială cu parametrii n şi p.*

Teorema 3.2.4. *Dacă variabila aleatoare* X *are repartiţie binomială cu parametrii n şi p, atunci valoarea medie şi dispersia sa sunt*

$$E(X)=np,\quad Var(X)=npq. \qquad (3.2.2)$$

Demonstraţie

Valoare medie a variabilei aleatoare X este

$$E(X)=0\cdot C_n^0 p^0 q^n + 1\cdot C_n^1 pq^{n-1} + 2\cdot C_n^2 p^2 q^{n-2} + \cdots + n\cdot C_n^n p^n q^0 = \sum_{k=0}^{n} kC_n^k p^k q^{n-k}.$$

Pentru a calcula suma de mai sus vom considera polinomul

$$P(x)=(px+q)^n = C_n^0 p^n x^n + C_n^1 p^{n-1} q x^{n-1} + \cdots + C_n^{n-1} pq^{n-1} x + C_n^n q^n$$
$$= \sum_{k=0}^{n} C_n^k p^{n-k} q^k x^{n-k} = \sum_{k=0}^{n} C_n^k p^k q^{n-k} x^k.$$

Derivând polinomul de mai sus obținem

$$P'(x) = np(px+q)^{n-1} = nC_n^0 p^n x^{n-1} + (n-1)C_n^1 p^{n-1} q x^{n-2} + \cdots$$
$$+ C_n^{n-1} p q^{n-1} + 0 \cdot C_n^n q^n = \sum_{k=0}^{n} k C_n^k p^k q^{n-k} x^{k-1}. \tag{3.2.3}$$

Luând $x = 1$ în relația (3.2.3), obținem $\sum_{k=0}^{n} k C_n^k p^k q^{n-k} = np(p+q)^{n-1}$, de unde rezultă că $E(X) = np$.

Pentru a calcula dispersia lui X vom folosi formula

$$Var(X) = E(X^2) - [E(X)]^2.$$

Media variabilei X^2 este $E(X^2) = \sum_{k=0}^{n} k^2 C_n^k p^k q^{n-k}$.

Înmulțim relația (3.2.3) cu *x* şi obținem

$$xP'(x) = npx(px+q)^{n-1} = nC_n^0 p^n x^n + (n-1)C_n^1 p^{n-1} q x^{n-1} + \cdots C_n^{n-1} p q^{n-1} x$$
$$+ 0 \cdot C_n^n q^n = \sum_{k=0}^{n} k C_n^k p^k q^{n-k} x^k.$$

Dacă derivăm relația de mai sus deducem că

$$P'(x) + xP''(x) = np(px+q)^{n-1} + n(n-1)p^2 x(px+q)^{n-2} = \sum_{k=0}^{n} k^2 C_n^k p^k q^{n-k} x^{k-1}.$$

Luând $x = 1$ în relația de mai sus deducem că $E(X^2) = np + n(n-1)p^2$.
Obținem astfel dispersia lui *X*

$$Var(X) = np + n(n-1)p^2 - n^2 p^2 = np - np^2 = npq.$$

q.e.d.

Propoziția 3.2.5. *Dacă λ este modulul (valoarea cea mai probabilă) a unei variabile aleatoare X cu repartiție binomială cu parametrii n şi p, atunci*

$$np - q \le \lambda \le np + p,$$

unde $q = 1 - p$.

Demonstrație

Dacă λ este modulul variabilei X atunci

$$P(X=\lambda-1)\le P(X=\lambda),\quad P(X=\lambda+1)\le P(X=\lambda).$$

Inegalitățile de mai sus ne conduc la sistemul

$$\begin{cases} C_n^{\lambda-1}p^{\lambda-1}q^{n-\lambda+1}\le C_n^{\lambda}p^{\lambda}q^{n-\lambda} \\ C_n^{\lambda+1}p^{\lambda+1}q^{n-\lambda-1}\le C_n^{\lambda}p^{\lambda}q^{n-\lambda} \end{cases} \Rightarrow \begin{cases} \dfrac{q}{n-\lambda+1}\le\dfrac{p}{\lambda} \\ \dfrac{p}{\lambda+1}\le\dfrac{q}{n-\lambda} \end{cases} \Rightarrow \begin{cases} \lambda\le np+p \\ \lambda\ge np-q, \end{cases}$$

de unde rezultă concluzia propoziției. *q.e.d.*

Propoziția 3.2.6. *Dacă variabila aleatoare X are repartiție binomială cu parametrii n şi p, atunci funcția sa caracteristică este*

$$\varphi(t)=(pe^{it}+q)^n,\ t\in R. \tag{3.2.4}$$

Demonstrație

Conform formulei pentru funcția caracteristică avem

$$\varphi(t)=\sum_{k=0}^{n}C_n^k p^k q^{n-k}e^{itk}=\sum_{k=0}^{n}C_n^k(pe^{it})^k q^{n-k}=(pe^{it}+q)^n,\ \forall t\in R.$$

q.e.d.

Teorema 3.2.7. *Dacă variabilele independente X şi Y au repartiții binomiale cu parametrii n şi p, respectiv m şi p, atunci variabila X+Y are repartiție binomială cu parametrii m+n şi p.*

Demonstrație

Deoarece X ia valorile $0,1,\ldots,n$, iar Y ia valorile $0,1,\ldots,m$, rezultă că variabila $X+Y$ va lua valorile $0,1,\ldots,n+m$. Variabila $X+Y$ are valoarea k ($k\in\{0,1,\ldots,n+m\}$) dacă ($X=0$ şi $Y=k$) sau ($X=1$ şi $Y=k-1$) sau ... sau ($X=k$ şi $Y=0$). Atunci vom obține

$$P(X+Y=k)=P\left(\bigcup_{j=0}^{k}\{X=j,Y=k-j\}\right)=\sum_{j=0}^{k}P(X=j,Y=k-j)$$

$$=\sum_{j=0}^{k}P(X=j)P(Y=k-j)=\sum_{j=0}^{k}C_n^j p^j q^{n-j}C_m^{k-j}p^{k-j}q^{m-k+j}$$

$$=p^k q^{m+n-k}\sum_{j=0}^{k}C_n^j C_m^{k-j}=C_{m+n}^k p^k q^{m+n-k}.$$

Am folosit mai sus faptul că evenimentele X şi Y sunt independente, şi de asemenea am utilizat formula $\sum_{j=0}^{k}C_n^j C_m^{k-j}=C_{m+n}^k$, care poate fi dedusă egalând coeficientul lui x^k din dezvoltările $(1+x)^n(1+x)^m$ şi $(1+x)^{n+m}$.

Deci am obținut

$$P(X+Y=k)=C_{m+n}^k p^k q^{m+n-k},\quad \forall k=0,1,\ldots,n+m,$$

adică variabila $X+Y$ are o repartiție binomială cu parametrii $m+n$ şi p.

Concluzia teoremei mai poate fi obținută folosind proprietatea de la funcții caracteristice care spune că funcția caracteristică a sumei a două variabile aleatoare independente cu funcțiile caracteristice $\varphi_1(t)$ şi $\varphi_2(t), t\in R$, are forma $\varphi(t)=\varphi_1(t)\varphi_2(t),\ t\in R.$ Astfel folosind relația (3.2.4) deducem că funcția caracteristică a variabilei $X+Y$ este

$$\varphi(t)=\left(pe^{it}+q\right)^n\left(pe^{it}+q\right)^m=\left(pe^{it}+q\right)^{m+n},\ \forall t\in R.$$

Din expresia de mai sus a funcției c tragem concluzia că variabila aleatoare $X+Y$ are repartiție binomială cu parametrii $m+n$ şi p. *q.e.d.*

Teorema 3.2.8. *(Bernoulli) Un eveniment are probabilitatea de realizare p atunci când facem o singură dată experiența de care este legat. Dacă α_n este numărul de realizări ale evenimentului când repetăm experiența de n ori, atunci*

$$\lim_{n\to\infty}P\left(\left|\frac{\alpha_n}{n}-p\right|\geq\varepsilon\right)=0, \qquad (3.2.5)$$

oricare ar fi $\varepsilon>0$.

Demonstrație

Variabila aleatoare α_n care are ca valori numărul de realizări ale evenimentului din problemă are repartiție binomială cu parametrii n şi p. Conform Teoremei

3.2.2 avem $E(\alpha_n) = np$ şi $Var(\alpha_n) = npq$. Variabila aleatoare $\frac{\alpha_n}{n}$ va avea atunci valoarea medie, dispersia şi abaterea medie pătratică

$$m = E\left(\frac{\alpha_n}{n}\right) = \frac{1}{n}E(\alpha_n) = \frac{np}{n} = p,$$

$$Var\left(\frac{\alpha_n}{n}\right) = \frac{1}{n^2}Var(\alpha_n) = \frac{pq}{n}, \quad \sigma_X = \sqrt{\frac{pq}{n}}.$$

Vom folosi acum Inegalitatea lui Cebâşev pentru variabila $\frac{\alpha_n}{n}$ şi $a = \varepsilon$.

Obținem

$$P\left(\left|\frac{\alpha_n}{n} - p\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2} = \frac{pq}{n\varepsilon^2}.$$

Deoarece $\lim_{n\to\infty} \frac{pq}{n\varepsilon^2} = 0$, din inegalitatea de mai sus rezultă inegalitatea (3.2.5).

q.e.d.

Observația 3.2.9. *O îmbunătățire a inegalității (3.2.5) este dată de teorema lui Borel, care spune că în condițiile Teoremei 3.2.8 are loc relația*

$$P\left(\frac{\alpha_n}{n} \to p\right) = 1.$$

Aplicația 3.2.10. *În cadrul unei experiențe evenimentele independente* $A_1, A_2, \ldots A_n$ *au probabilitățile de realizare* $P(A_k) = p_k$, $k = 1,2,\ldots,n$. *Să se calculeze valoarea medie şi dispersia numărului de evenimente care se realizează atunci când experiența are loc.*

Rezolvare

Să notăm cu X variabila aleatoare care are ca valori numărul de evenimente care se realizează în cadrul experienței. Valorile variabilei X sunt $0,1,2,\ldots,n$. Probabilitatea ca X să ia valoarea k ($k = 0,1,2,\ldots,n$) este, conform schemei lui Poisson (schema binomială generalizată) coeficientul lui x^k din polinomul

$$Q(x) = (p_1 x + q_1)(p_2 x + q_2)\cdots(p_n x + q_n), \qquad (3.2.6)$$

unde $q_i = 1 - p_i$, $i = 1,2,\ldots,n$. Dacă scriem desfăşurat pe $Q(x)$ sub forma

$$Q(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n, \qquad (3.2.7)$$

atunci tabloul repartiției variabilei *X* este

$$X : \begin{pmatrix} 0 & 1 & 2 & \cdots & n \\ a_0 & a_1 & a_2 & \cdots & a_n \end{pmatrix}.$$

Suma tuturor elementelor de pe linia a doua a tabloului de mai sus este *1*. Într-adevăr

$$a_0 + a_1 + \cdots + a_n = Q(1) = (p_1 + q_1)(p_2 + q_2)\cdots(p_n + q_n) = 1.$$

Valoarea medie a variabilei *X* este $E(X) = \sum_{k=0}^{n} k a_k$. Ideea de demonstrație a teoremei este asemănătoare cu cea a Teoremei 3.2.4. Vom deriva polinomul Q, scris sub cele două forme de mai sus (3.2.6) şi (3.2.7). Derivând relația (3.2.7) obținem

$$Q'(x) = a_1 + 2a_2 x + 3a_3 x^2 + \ldots + n a_n x^{n-1}, \qquad (3.2.8)$$

de unde rezultă

$$Q'(1) = a_1 + 2a_2 + 3a_3 + \cdots + n a_n = \sum_{k=1}^{n} k a_k = M(X).$$

Pe de altă parte parte, derivând relația (3.2.6) obținem

$$Q'(x) = p_1 \prod_{k \neq 1} (p_k x + q_k) + p_2 \prod_{k \neq 2} (p_k x + q_k) + \cdots + p_n \prod_{k \neq n} (p_k x + q_k), \qquad (3.2.9)$$

iar pentru $x = 1$ deducem $Q'(1) = p_1 + p_2 + \cdots + p_n$. Rezultă că

$$E(X) = \sum_{k=1}^{n} p_k. \qquad (3.2.10)$$

Pentru a calcula dispersia, vom calcula mai întâi $E(X^2) = \sum_{k=0}^{n} k^2 a_k$. Înmulțim relația (3.2.8) cu *x* şi obținem

$$xQ'(x) = a_1 x + 2a_2 x^2 + 3a_3 x^3 + \cdots + n a_n x^n.$$

Derivând egalitatea de mai sus rezultă

$$Q'(x)+xQ''(x)=a_1+2^2a_2x+3^2a_3x^2+\cdots+n^2a_nx^{n-1}.$$

Pentru $x=1$ deducem din relația de mai sus $Q'(1)+Q''(1)=\sum_{k=1}^{n}k^2a_k$. Deci

$$E(X^2)=Q'(1)+Q''(1)=\sum_{k=1}^{n}p_k+Q''(1). \qquad (3.2.11)$$

Derivăm acum relația (3.2.9) pentru a determina $Q''(x)$; obținem

$$Q''(x)=p_1\left[p_2\prod_{j\neq1,2}(p_jx+q_j)+p_3\prod_{j\neq1,3}(p_jx+q_j)+\cdots+p_n\prod_{j\neq1,n}(p_jx+q_j)\right]$$
$$+\cdots+p_n\left[p_1\prod_{j\neq1,n}(p_jx+q_j)+p_2\prod_{j\neq2,n}(p_jx+q_j)+\cdots+p_{n-1}\prod_{j\neq1,n-1}(p_jx+q_j)\right].$$

Pentru $x=1$ obținem din relația de mai sus

$$Q''(1)=p_1\sum_{k\neq1}p_k+p_2\sum_{k\neq2}p_k+\cdots+p_n\sum_{k\neq n}p_k=p_1[E(X)-p_1]+p_2[E(X)-p_2]$$
$$+\cdots+p_n[E(X)-p_n]=E(X)(p_1+p_2+\cdots+p_n)-(p_1^2+p_2^2+\cdots+p_n^2)$$
$$=[E(X)]^2-\sum_{k=1}^{n}p_k^2.$$

Din relația (3.2.11) şi din relația de mai sus deducem că

$$E(X^2)=\sum_{k=1}^{n}p_k+[E(X)]^2-\sum_{k=1}^{n}p_k^2. \qquad (3.2.12)$$

Folosind acum relațiile (3.2.10) şi (3.2.12) rezultă că dispersia lui *X* este

$$Var(X)=E(X^2)-[E(X)]^2=\sum_{k=1}^{n}p_k+[E(X)]^2-\sum_{k=1}^{n}p_k^2-[E(X)]^2$$
$$=\sum_{k=1}^{n}p_k-\sum_{k=1}^{n}p_k^2=\sum_{k=1}^{n}p_k(1-p_k)=\sum_{k=1}^{n}p_kq_k.$$

O altă metodă mai simplă pentru calculul mediei şi dispersiei variabilei *X* este următoarea: să notăm cu X_k variabila aleatoare care are ca valori pe *1* dacă A_k se

realizează, şi pe *0* dacă A_k nu se realizează, pentru $k = 1, 2, \ldots, n.$ Tablourile de repartiţie pentru variabilele X_k sunt

$$X_k : \begin{pmatrix} 1 & 0 \\ p_k & q_k \end{pmatrix}, \quad k = 1, 2, \ldots, n.$$

Atunci numărul evenimentelor care se realizează este $X = X_1 + \cdots + X_n$, deci media sa va fi

$$E(X) = \sum_{k=1}^{n} E(X_k) = \sum_{k=1}^{n} p_k.$$

Deoarece variabilele aleatoare X_k, $k = 1, 2, \ldots, n$ sunt independente, atunci dispersia varaibilei *X* va fi

$$Var(X) = \sum_{k=1}^{n} Var(X_k) = \sum_{k=1}^{n} \left\{ E(X_k^2) - [E(X_k)]^2 \right\} = \sum_{k=1}^{n} (p_k - p_k^2) = \sum_{k=1}^{n} p_k q_k.$$

q.e.d.

Pentru $n = 1,$ legea binomială este cunoscută şi sub numele de **legea Bernoulli** cu parametrul *p*. Variabila aleatoare *X* care urmează legea Bernoulli cu parametrul *p* admite doar două valori posibile *0* şi *1* cu probabilităţile de realizare *q=1-p* şi *p*, având tabloul repartiţiei

$$X : \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}, \quad q = 1 - p.$$

Valoarea medie şi dispersia variabilei *X* sunt $E(X) = p$ şi $Var(X) = pq$.

O variabilă aleatoare cu repartiţie binomială cu parametrii *n* şi *p* dată de Definiţia 3.2.1 este suma a *n* variabile aleatoare independente cu repartiţii Bernoulli cu acelaşi parametru *p*.

3.3. Legea binomială cu exponent negativ. Legea geometrică

Definiţia 3.3.1. *Variabila aleatoare X urmează* ***legea binomială cu exponent negativ*** *(X are repartiţie binomială cu exponent negativ) cu parametrii m şi p* ($m \in N^*$, $0 < p < 1$) *dacă ia valorile* $m, m+1, m+2, \ldots$ *cu probabilităţile*

$$P(X = k) = C_{k-1}^{m-1} p^m q^{k-m}, \quad k \geq m, \qquad (3.3.1)$$

unde q=1-p.

Tabloul repartiției variabilei aleatoare *X* este

$$X:\ \begin{pmatrix} m & m+1 & m+2 & \cdots \\ C_{m-1}^{m-1}p^m q^0 & C_m^{m-1}p^m q & C_{m+1}^{m-1}p^m q^2 & \cdots \end{pmatrix}.$$

Exemplul 3.3.2. *O experiență se efectuează până la cea de-a m-a realizare a unui eveniment A legat de ea. Dacă probabilitatea acestui eveniment când se face o singură dată experiența este p, atunci numărul X de efectuări ale experienței este variabilă aleatoare care are repartiție binomială cu exponent negativ cu parametrii m şi p. Într-adevăr, evenimentul* $\{X=k\}$ *se scrie ca intersecția a două evenimente: „în primele k-1 efectuări ale experienței evenimentul A se produce de m-1 ori" şi „în a k-a efectuare a experienței se produce A". Probabilitatea primului din aceste două evenimente este* $C_{k-1}^{m-1}p^{m-1}q^{k-m}$*, conform schemei lui Bernoulli, iar probabilitatea celui de-al doilea este p. Deci*

$$P(X=k)=pC_{k=1}^{m-1}p^{m-1}q^{k-m}=C_{k-1}^{m-1}p^m q^{k-m},\quad k=m,\,m+1,\ldots$$

Teorema 3.3.3. *Dacă variabila aleatoare X are repartiție binomială cu exponent negativ cu parametrii m şi p, atunci valoarea medie şi dispersia sa sunt*

$$E(X)=\frac{m}{p},\quad Var(X)=\frac{mq}{p^2}. \tag{3.3.2}$$

Demonstrație

Valoarea medie a variabilei aleatoare *X* este

$$E(X)=mC_{m-1}^{m-1}p^m q^0+(m+1)C_m^{m-1}p^m q+(m+2)C_{m+1}^{m-1}p^m q^2+\cdots$$
$$=\sum_{k=m}^{\infty}kC_{k-1}^{m-1}p^m q^{k-m}.$$

Pentru a calcula suma seriei de mai sus, pornim de la dezvoltarea în serie de puteri a funcției $(1-x)^{-m}$, şi anume

$$(1-x)^{-m}=1+\frac{m}{1!}x+\frac{m(m+1)}{2!}x^2+\cdots,\qquad x\in(-1,1).$$

Pentru $m\in N^*$ seria binomială de mai sus se scrie sub forma

$$(1-x)^{-m} = C_{m-1}^{0} + C_m^1 x + C_{m+1}^2 x^2 + \cdots = \sum_{k=m}^{\infty} C_{k-1}^{k-m} x^{k-m}, \quad x \in (-1,1)$$

sau

$$(1-x)^{-m} = C_{m-1}^{m-1} + C_m^{m-1} x + C_{m+1}^{m-1} x^2 + \cdots = \sum_{k=m}^{\infty} C_{k-1}^{m-1} x^{k-m}, \quad x \in (-1,1). \qquad (3.3.3)$$

Folosind seria de mai sus (cu $x=q$), observăm că suma tuturor probabilităților de pe linia a doua a tabloului repartiției variabilei X este *1*. Într-adevăr

$$\sum_{k=m}^{\infty} C_{k-1}^{m-1} p^m q^{k-m} = p^m \sum_{k=m}^{\infty} C_{k-1}^{m-1} q^{k-m} = p^m (1-q)^{-m} = \left(\frac{1}{p} - \frac{q}{p}\right)^{-m} = 1.$$

Numele repartiției binomiale cu exponent negativ provine din observația că termenii $P(X=k) = C_{k-1}^{m-1} p^m q^{k-m}$, $k \geq m$ sunt termenii generali ai dezvoltării $\left(\frac{1}{p} - \frac{q}{p}\right)^{-m}$.

Relația (3.3.3) se scrie echivalent astfel

$$\frac{x^m}{(1-x)^m} = \sum_{k=m}^{\infty} C_{k-1}^{m-1} x^k \qquad (3.3.4)$$

Derivând relația (3.3.4) obținem

$$\frac{m x^{m-1}}{(1-x)^{m+1}} = \sum_{k=m}^{\infty} k C_{k-1}^{m-1} x^{k-1}, \quad x \in (-1,1). \qquad (3.3.5)$$

Pentru $x=q$ din relația (3.3.5) deducem

$$\frac{m q^{m-1}}{p^{m+1}} = \sum_{k=m}^{\infty} k C_{k-1}^{m-1} q^{k-1}. \qquad (3.3.6)$$

Atunci valoarea medie a variabilei X este, folosind relația (3.3.6)

$$E(X) = \sum_{k=m}^{\infty} k C_{k-1}^{m-1} p^m q^{k-m} = p^m q^{-m+1} \sum_{k=m}^{\infty} k C_{k-1}^{m-1} q^{k-1} = p^m q^{-m+1} \frac{m q^{m-1}}{p^{m+1}} = \frac{m}{p}.$$

Pentru a calcula dispersia variabilei X, vom calcula mai întâi $E(X^2)$, şi anume

$$E(X^2)=\sum_{k=m}^{\infty}k^2C_{k-1}^{m-1}p^m q^{k-m}=p^m q^{1-m}\sum_{k=m}^{\infty}k^2C_{k-1}^{m-1}q^{k-1}.$$

Înmulțim relația (3.3.5) cu x şi obținem

$$\frac{mx^m}{(1-x)^{m+1}}=\sum_{k=m}^{\infty}kC_{k-1}^{m1}x^k,\quad x\in(-1,1).$$

Prin derivarea relației de mai sus rezultă

$$\frac{x^{m-1}m(m+x)}{(1-x)^{m+2}}=\sum_{k=m}^{\infty}k^2C_{k-1}^{m-1}x^{k-1},\quad x\in(-1,1).$$

Pentru $x=q$ din relația obținută deducem

$$\sum_{k=m}^{\infty}k^2C_{k-1}^{m-1}q^{k-1}=\frac{q^{m-1}m(m+q)}{p^{m+2}}.$$

Rezultă atunci că $E(X^2)$ este

$$E(X^2)=p^m q^{1-m}\frac{q^{m-1}m(m+q)}{p^{m+2}}=\frac{m(m+q)}{p^2},$$

iar dispersia varaibilei aleatoare X este

$$Var(X)=E(X^2)-[E(X)]^2=\frac{m^2+mq}{p^2}-\frac{m^2}{p^2}=\frac{mq}{p^2}.$$

q.e.d.

Propoziția 3.3.4. *Dacă variabila aleatoare X are repartiție binomială cu exponent negativ cu parametrii n şi p, atunci funcția sa caracteristică este*

$$\varphi(t)=\left(\frac{pe^{it}}{1-qe^{it}}\right)^m,\quad t\in R. \tag{3.3.7}$$

Demonstrație

Conform formulei pentru funcția caracteristică avem

$$\varphi(t)=\sum_{k=m}^{\infty}C_{k-1}^{m-1}p^m q^{k-m}e^{itk}=p^m e^{itm}\sum_{k=m}^{\infty}C_{k-1}^{m-1}(qe^{it})^{k-m}$$

$$=p^m e^{itm}(1-qe^{it})^{-m}=\left(\frac{pe^{it}}{1-qe^{it}}\right)^m,\ \forall t\in R,$$

conform relației (3.3.3), care este adevărată şi pentru numere complexe $x\in C,\ |x|<1.$ *q.e.d.*

Pentru $m=1$ legea binomială cu exponent negativ cu parametrul p se mai numeşte **legea geometrică** cu parametrul p. Tabloul repartiției unei variabile aleatoare X cu repartiție geometrică cu parametrul p este următorul

$$X:\ \begin{pmatrix} 1 & 2 & 3 & \cdots \\ p^m q^0 & p^m q & p^m q^2 & \cdots \end{pmatrix},$$

unde $q=1-p$. Conform relațiilor (3.3.2) şi (3.3.7) media, dispersia şi funcția caracteristică ale lui X sunt

$$E(X)=\frac{1}{p},\ Var(X)=\frac{q}{p^2},\ \varphi(t)=\frac{pe^{it}}{1-qe^{it}},\ t\in R.$$

3.4. Legea hipergeometrică

Definiția 3.4.1. *Variabila aleatoare X urmează* ***legea hipergeometrică*** *(X are repartiție hipergeometrică) cu parametrii a, b şi n* $(a,b,n\in N^*,\ n\le a+b)$ *dacă poate lua orice valoare întreagă între* $\max(0,n-b)$ *şi* $\min(n,a)$ *şi*

$$P(X=k)=\frac{C_a^k C_b^{n-k}}{C_{a+b}^n},\ \forall k\in[\max(0,n-b),\min(n,a)]. \qquad (3.4.1)$$

Pentru calculul mediei şi dispersiei variabilei aleatoare X cu repartiție hipergeometrică vom presupune fără a restrânge generalitatea problemei că $n\le b$ şi $n\le a,$ deci $\max(0,n-b)=0$ şi $\min(n,a)=n.$ Atunci tabloul repartiției variabilei X este

$$X:\ \begin{pmatrix} 0 & 1 & 2 & \cdots & n \\ \dfrac{C_a^0 C_b^n}{C_{a+b}^n} & \dfrac{C_a^1 C_b^{n-1}}{C_{a+b}^n} & \dfrac{C_a^2 C_b^{n-2}}{C_{a+b}^n} & \cdots & \dfrac{C_a^n C_b^0}{C_{a+b}^n} \end{pmatrix}.$$

Se observă că suma probabilităților de pe linia a doua a tabloului de mai sus este *1*. Într-adevăr avem

$$\sum_{k=0}^{n}\frac{C_a^k C_b^{n-k}}{C_{a+b}^n}=\frac{1}{C_{a+b}^n}\sum_{k=0}^{n}C_a^k C_b^{n-k}=\frac{1}{C_{a+b}^n}C_{a+b}^n=1.$$

Exemplul 3.4.2. Dacă dintr-o urnă care conține *a* bile albe şi *b* bile negre se extrag *n* bile una câte una, fără întoarcerea bilei extrase în urnă (sau se extrag *n* bile simultan), iar *X* este numărul de bile albe extrase, atunci *X* are repartiție hipergeometrică cu parametrii *a, b* şi *n,* conform schemei hipergeometrice (schema bilei neîntoarse).

Teorema 3.4.3. *Dacă variabila aleatoare X are repartiție hipergeometrică cu parametrii a, b şi n, cu* $n\le b$ şi $n\le a$, *atunci valoarea medie şi dispersia sa sunt*

$$E(X)=ap,\quad Var(X)=npq\frac{a+b-n}{a+b-1}, \tag{3.4.2}$$

unde $p=\frac{a}{a+b}$, $q=1-p=\frac{b}{a+b}$.

Demonstrație

Valoarea medie a variabilei aleatoare *X* este

$$E(X)=\sum_{k=0}^{n}k\frac{C_a^k C_b^{n-k}}{C_{a+b}^n}=\frac{a}{C_{a+b}^n}\sum_{k=1}^{n}C_{a-1}^{k-1}C_b^{n-k}=\frac{a}{C_{a+b}^n}C_{a+b-1}^{n-1}=\frac{an}{a+b}=ap.$$

Pentru calculul dispersiei, vom calcula mai întâi media variabilei X^2; avem

$$E(X^2)=\sum_{k=0}^{n}k^2\frac{C_a^k C_b^{n-k}}{C_{a+b}^n}=\sum_{k=1}^{n}k(k-1)\frac{C_a^k C_b^{n-k}}{C_{a+b}^n}+\sum_{k=0}^{n}k\frac{C_a^k C_b^{n-k}}{C_{a+b}^n}$$

$$=\frac{a(a-1)}{C_{a+b}^n}\sum_{k=2}^{n}C_{a-2}^{k-2}C_b^{n-k}+E(X)=\frac{a(a-1)}{C_{a+b}^n}C_{a+b-2}^{n-2}+\frac{an}{a+b}=\frac{an(a-1)(n-1)}{(a+b)(a+b-1)}$$

$$+\frac{an}{a+b}.$$

Deci dispersia lui *X* este

$$Var(X) = E(X^2) - [E(X)]^2 = \frac{an(a-1)(n-1)}{(a+b)(a+b-1)} + \frac{an}{a+b} - \frac{a^2n^2}{(a+b)^2}$$
$$= \frac{abn(a+b-n)}{(a+b)^2(a+b-1)} = npq\frac{a+b-n}{a+b-1}.$$

Folosind modelul urnei din Exemplul 3.4.2, valoarea medie a variabilei *X* din problemă se poate calcula şi în felul următor. Considerăm o urnă cu *a* bile albe şi *b* bile negre din care se extrag una câte una *n* bile (fără întoarcerea bilei în urnă) şi considerăm variabilele aleatoare X_k, pentru $k = 1,2,\ldots,n$, unde variabila X_k ($k = 1,2,\ldots,n$) are ca valori numărul de bile albe obținute la extragerea *k* (*1* dacă obținem bilă albă şi *0* dacă obținem bilă neagră). Pentru X_1, avem

$$P(X_1 = 1) = \frac{a}{a+b}, \quad P(X_1 = 0) = \frac{b}{a+b}.$$

Deci tabloul repartiției variabilei X_1 este X_1: $\begin{pmatrix} 1 & 0 \\ \frac{a}{a+b} & \frac{b}{a+b} \end{pmatrix}$. Pentru variabila X_2 obținem (în urnă au rămas *a+b-1* bile)

$$P(X_2 = 1) = P(X_2 = 1/X_1 = 1)P(X_1 = 1) + P(X_2 = 1/X_1 = 0)P(X_1 = 0)$$
$$= \frac{a-1}{a+b-1}\cdot\frac{a}{a+b} + \frac{a}{a+b-1}\cdot\frac{b}{a+b} = \frac{a(a+b-1)}{(a+b-1)(a+b))} = \frac{a}{a+b},$$
$$P(X_2 = 0) = P(X_2 = 0/X_1 = 1)P(X_1 = 1) + P(X_2 = 0/X_1 = 0)P(X_1 = 0)$$
$$= \frac{b}{a+b-1}\cdot\frac{a}{a+b} + \frac{b-1}{a+b-1}\cdot\frac{b}{a+b} = \frac{b(a+b-1)}{(a+b)(a+b-1)} = \frac{b}{a+b}.$$

Deci tabloul repartiției variabilei X_2 este X_2: $\begin{pmatrix} 1 & 0 \\ \frac{a}{a+b} & \frac{b}{a+b} \end{pmatrix}$. Pentru variabila X_3 avem (în urnă au rămas *a+b-2* bile)

$$P(X_3 = 1) = P(X_3 = 1/X_2 = 1)P(X_2 = 1) + P(X_3 = 1/X_2 = 0)P(X_2 = 0)$$
$$= [P((X_3 = 1/X_2 = 1)/X_1 = 1)P(X_1 = 1) + P((X_3 = 1/X_2 = 1)/X_1 =$$
$$= 0)P(X_1 = 0)]P(X_2 = 1) + [P((X_3 = 1/X_2 = 0)/X_1 = 1)P(X_1 = 1) +$$
$$+ P((X_3 = 1/X_2 = 0)/X_1 = 0)P(X_1 = 0)]P(X_2 = 0)$$

$$=\left(\frac{a-2}{a+b-2}\cdot\frac{a}{a+b}+\frac{a-1}{a+b-2}\cdot\frac{b}{a+b}\right)\frac{a}{a+b}+\left(\frac{a-1}{a+b-2}\cdot\frac{a}{a+b}\right.$$
$$\left.+\frac{a}{a+b-2}\cdot\frac{b}{a+b}\right)\frac{b}{a+b}=\frac{(a-2)a^2+2(a-1)ab+ab^2}{(a+b-2)(a+b)^2}$$
$$=\frac{a(a+b)(a+b-2)}{(a+b-2)(a+b)^2}=\frac{a}{a+b}.$$

Asemănător se arată că $P(X_3=0)=\frac{b}{a+b}(=1-P(X_3=1))$. Deci tabloul repartiției variabilei X_3 este X_3: $\begin{pmatrix}1 & 0\\ \frac{a}{a+b} & \frac{b}{a+b}\end{pmatrix}$.

Se arată în acelaşi mod că toate variabilele X_k, $k=1,2,\ldots,n$ au acelaşi tablou de repartiție X_k: $\begin{pmatrix}1 & 0\\ \frac{a}{a+b} & \frac{b}{a+b}\end{pmatrix}$, $k=1,2,\ldots,n$, (deşi ele sunt variabile dependente).

Cu ajutorul variabilelor X_k, $k=1,2,\ldots,n$, variabila X se scrie $X=\sum_{k=1}^{n}X_k$, deci media variabilei X este

$$E(X)=\sum_{k=1}^{n}E(X_k)=\sum_{k=1}^{n}\frac{a}{a+b}=\frac{na}{a+b}=np.$$

Pentru dispersie nu mai putem scrie că $Var(X)$ este egală cu $\sum_{k=1}^{n}Var(X_k)$, deoarece variabilele X_k, $k=1,2,\ldots,n$ nu sunt independente. *q.e.d.*

3.5. Legea Poisson (legea evenimentelor rare)

Definiția 3.5.1. *Variabila aleatoare X urmează* ***legea Poisson*** *(X are repartiție Poisson) cu parametrul* λ $(\lambda>0)$ *dacă poate lua orice valoare întreagă pozitivă şi*

$$P(X=k)=\frac{\lambda^k}{k!}e^{-\lambda},\quad k=0,1,2,\ldots \tag{3.5.1}$$

Tabloul repartiției variabilei X este

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

Pentru a verifica că suma probabilităților de pe linia a doua a tabloului de mai sus este *1*, vom folosi dezvoltarea în serie de puteri a funcției $f(x)=e^x$, şi anume

$$e^x = 1+\frac{x}{1!}+\frac{x^2}{2!}+\cdots+\frac{x^k}{k!}+\cdots, \quad \forall x\in R. \tag{3.5.2}$$

Folosind relația (3.5.2) pentru $x=\lambda$, avem

$$\sum_{k=0}^{\infty} P(X=k) = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1.$$

Teorema 3.5.2. *Dacă variabila aleatoare X are repartiție Poisson cu parametrul* λ, *atunci valoarea medie şi dispersia sa sunt*

$$E(X)=\lambda, \quad Var(X)=\lambda. \tag{3.5.3}$$

Demonstrație

Valoarea medie a variabilei X este

$$E(X)=\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda}e^{\lambda} = \lambda.$$

Pentru dispersie, calculăm mai întâi media variabilei X^2. Obținem

$$E(X^2)=\sum_{k=0}^{\infty} k^2\frac{\lambda^k}{k!}e^{-\lambda} = \sum_{k=1}^{\infty}(k^2-k)\frac{\lambda^k}{k!}e^{-\lambda} + \sum_{k=1}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} = \sum_{k=2}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda}$$
$$+E(X) = \lambda^2 e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!} + E(X) = \lambda^2 e^{-\lambda}e^{\lambda} + \lambda = \lambda^2+\lambda.$$

Atunci dispersia lui X este

$$Var(X) = E(X^2)-[E(X)]^2 = \lambda^2+\lambda-\lambda^2 = \lambda. \; q.e.d.$$

Propoziția 3.5.3. *Dacă variabila aleatoare X are repartiție Poisson cu parametrul λ, atunci funcția sa caracteristică este*

$$\varphi(t) = e^{\lambda(e^{it}-1)}, \quad t \in R. \tag{3.5.4}$$

Demonstrație

Folosind formula de calcul de la funcția caracteristică, avem

$$\varphi(t) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} e^{itk} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{itk} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^{it})^k}{k!} = e^{-\lambda} e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)}, \ \forall t \in R.$$

q.e.d.

Teorema 3.5.4. *Variabilele aleatoare independente X_1 şi X_2 au repartiții Poisson cu parametrii λ_1 şi respectiv λ_2. Atunci variabila aleatoare $X_1 + X_2$ are repartiție Poisson cu parametrul $\lambda_1 + \lambda_2$.*

Demonstrație

Variabila aleatoare $X_1 + X_2$ are ca valori pe $0,1,2,\ldots$ Fie $k \geq 0$ întreg. Atunci avem

$$P(X_1 + X_2 = k) = P\left(\bigcup_{j=0}^{k} (X_1 = j,\ X_2 = k - j)\right) = \sum_{j=0}^{k} P(X_1 = j,\ X_2 = k - j)$$

$$= \sum_{j=0}^{k} P(X_1 = j) P(X_2 = k - j) = \sum_{j=0}^{k} \frac{\lambda_1^j}{j!} e^{-\lambda_1} \frac{\lambda_2^{k-j}}{(k-j)!} e^{-\lambda_2} = \frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{j=0}^{k} C_k^j \lambda_1^j \lambda_2^{k-j}$$

$$= \frac{(\lambda_1 + \lambda_2)^k}{k!} e^{-(\lambda_1+\lambda_2)}.$$

Deducem astfel că variabila aleatoare $X_1 + X_2$ are repartiție Poisson cu parametrul $\lambda_1 + \lambda_2$. Folosind funcțiile caracteristice ale variabilelor X_1, X_2, exprimate cu ajutorul formulei (3.5.4), şi anume $\varphi_1(t) = e^{\lambda_1(e^{it}-1)}$, $\varphi_2(t) = e^{\lambda_2(e^{it}-1)}$, $t \in R$, deducem că funcția caracteristică a variabilei $X_1 + X_2$ este

$$\varphi(t) = \varphi_1(t)\varphi_2(t) = e^{\lambda_1(e^{it}-1)} \cdot e^{\lambda_2(e^{it}-1)} = e^{(\lambda_1+\lambda_2)(e^{it}-1)}, \quad \forall t \in R.$$

Din expresia de mai sus a funcției caracteristice rezultă că variabila $X_1 + X_2$ are repartiție Poisson cu parametrul $\lambda_1 + \lambda_2$. *q.e.d.*

Legătura dintre repartiția binomială şi repartiția Poisson este dată de următoarea teoremă.

Teorema 3.5.5. *Fie* $k \in N$ *fixat, iar pentru* $n > k$ *considerăm variabilele* X_n *care au repartiţii binomiale cu parametrii n şi* p_n, *astfel încât toate să aibă aceeaşi valoare medie* λ. *Atunci are loc relaţia*

$$\lim_{n\to\infty} P(X_n = k) = \frac{\lambda^k}{k!} e^{-\lambda}. \qquad (3.5.5)$$

Rezolvare

Deoarece variabilele X_n, $n > k$ au aceeaşi valoare medie λ, deducem conform primei relaţii din (3.5.3) că valoarea medie a acestor variabile este $E(X_n) = np_n = \lambda$. Deci $p_n = \lambda / n$. Atunci obţinem

$$\lim_{n\to\infty} P(X_n = k) = \lim_{n\to\infty} C_n^k p_n^k q_n^{n-k} = \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k} \cdot \lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda},$$

adică relaţia (3.5.5). *q.e.d.*

Observaţia 3.5.6. *Relaţia (3.5.5) ne arată că dacă* p_n *este suficient de mic şi n suficient de mare, atunci putem aproxima repartiţia binomială cu parametrii n şi* p_n, *prin repartiţia Poisson de parametru* $\lambda = np_n$. *Din acest motiv repartiţia Poisson se mai numeşte* ***legea evenimentelor rare***. *Dacă* $n \geq 30$ *şi* $np < 5$ *atunci repartiţia Poisson cu parametrul* $\lambda = np$ *este o bună aproximare a repartiţiei binomiale cu parametrii n şi p.*

Aplicaţia 3.5.7. *Să se calculeze momentele iniţiale* $m_3(X)$ *şi* $m_4(X)$, *precum şi momentele centrate* $\mu_3(X)$ *şi* $\mu_4(X)$ *pentru o variabilă aleatoare X cu repartiţie Poisson cu parametrul* λ.

Rezolvare

Din demonstraţia teoremei 3.5.2 ştim că $m_1(X) = E(X) = \lambda$, iar $m_2(X) = E(X^2) = \lambda^2 + \lambda$. Momentul iniţial de ordinul al treilea al variabilei *X* este $m_3(X) = E(X^3) = \sum_{k=0}^{\infty} k^3 \frac{\lambda^k}{k!} e^{-\lambda}$.

Deoarece $k^3 = k(k-1)(k-2) + 3k(k-1) + k$, vom scrie pe $m_3(X)$ astfel

$$m_3(X)=\sum_{k=2}^{\infty}k(k-1)(k-2)\frac{\lambda^k}{k!}e^{-\lambda}+3\sum_{k=1}^{\infty}k(k-1)\frac{\lambda^k}{k!}e^{-\lambda}+\sum_{k=1}^{\infty}k\frac{\lambda^k}{k!}e^{-\lambda}$$
$$=\lambda^3e^{-\lambda}\sum_{k=3}^{\infty}\frac{\lambda^{k-3}}{(k-3)!}+3\lambda^2e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}+\lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}=\lambda^3e^{-\lambda}e^{\lambda}$$
$$+3\lambda^2e^{-\lambda}e^{\lambda}+\lambda e^{-\lambda}e^{\lambda}=\lambda^3+3\lambda^2+\lambda.$$

Apoi momentul centrat de ordinul al treilea este

$$\mu_3(X)=E(X-\lambda)^3=E(X^3-3\lambda X^2+3\lambda^2X-\lambda^3)=E(X^3)-3\lambda E(X^2)$$
$$+3\lambda^2E(X)-\lambda^3=m_3-3\lambda m_2+3\lambda^2m_1-\lambda^3=\lambda^3+3\lambda^2+\lambda-3\lambda(\lambda^2+\lambda)$$
$$+3\lambda^3-\lambda^3=\lambda.$$

Pentru momentul inițial de ordinul al patrulea avem $m_4(X)=E(X^4)=\sum_{k=0}^{\infty}k^4\frac{\lambda^4}{k!}e^{-\lambda}$. Deoarece $k^4=k(k-1)(k-2)(k-3)+6k(k-1)(k-2)+7k(k-1)+k$, momentul $m_4(X)$ se scrie astfel

$$m_4(X)=\sum_{k=3}^{\infty}k(k-1)(k-2)(k-3)\frac{\lambda^k}{k!}e^{-\lambda}+6\sum_{k=2}^{\infty}k(k-1)(k-2)\frac{\lambda^k}{k!}e^{-\lambda}$$
$$+7\sum_{k=1}^{\infty}k(k-1)\frac{\lambda^k}{k!}e^{-\lambda}+\sum_{k=1}^{\infty}k\frac{\lambda^k}{k!}e^{-\lambda}=\lambda^4e^{-\lambda}\sum_{k=4}^{\infty}\frac{\lambda^{k-4}}{(k-4)!}+6\lambda^3e^{-\lambda}\sum_{k=3}^{\infty}\frac{\lambda^{k-3}}{(k-3)!}$$
$$+7\lambda^2e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}+\lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}=\lambda^4e^{-\lambda}e^{\lambda}+6\lambda^3e^{-\lambda}e^{\lambda}+7\lambda^2e^{-\lambda}e^{\lambda}$$
$$+\lambda e^{-\lambda}e^{\lambda}=\lambda^4+6\lambda^3+7\lambda^2+\lambda.$$

Apoi momentul centrat de ordinul al patrulea este

$$\mu_4(X)=E(X-\lambda)^4=E(X^4-4\lambda X^3+6\lambda^2X^2-4\lambda^3X+\lambda^4)=E(X^4)$$
$$-4\lambda E(X^3)+6\lambda^2E(X^2)-4\lambda^3E(X)+\lambda^4=m_4-4\lambda m_3+6\lambda^2m_2-4\lambda^3m_1+\lambda^4$$
$$=\lambda^4+6\lambda^3+7\lambda^2+\lambda-4\lambda(\lambda^3+3\lambda^2+\lambda)+6\lambda^2(\lambda^2+\lambda)-4\lambda^4+\lambda^4=3\lambda^2+\lambda.$$

Capitolul 4

Legi clasice de probabilitate (repartiții) ale variabilelor aleatoare continue

Introducere

Vom prezenta în acest capitol principalele legi de probabilitate ale variabilelor aleatoare continue, şi anume: legea continuă uniformă (rectangulară), legea normală (Gauss-Laplace), legea log-normală, legea gamma, legea beta, legea χ^2 (Helmert-Pearson), legea Student (t) şi cazul său particular legea Cauchy, legea Snedecor şi legea Fisher, legea Weibull şi cazul său particular legea exponențială.

4.1. Legea continuă uniformă (rectangulară)

Definiția 4.1.1. *Variabila aleatoare X urmează* ***legea continuă uniformă (rectangulară)*** *(X are repartiție uniformă) cu parametrii* μ *şi* ω $(\mu \in R, \omega > 0)$ *dacă densitatea sa de probabilitate (repartiție) este funcția*

$$f(x) = \begin{cases} \dfrac{1}{\omega}, & x \in \left[\mu - \dfrac{\omega}{2}, \mu + \dfrac{\omega}{2}\right], \\ 0, & x \in \left(-\infty, \mu - \dfrac{\omega}{2}\right) \cup \left(\mu + \dfrac{\omega}{2}, \infty\right). \end{cases} \tag{4.1.1}$$

Observăm că funcția f este o densitate de probabilitate, deoarece $f(x) \geq 0,\ \forall x \in R,$ şi

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{\mu-\frac{\omega}{2}}^{\mu+\frac{\omega}{2}} \frac{1}{\omega}\,dx = \frac{x}{\omega}\Bigg|_{\mu-\frac{\omega}{2}}^{\mu+\frac{\omega}{2}} = 1.$$

Funcția de repartiție a variabilei X este

$$F(x)=\begin{cases}0, & x\in\left(-\infty,\mu-\dfrac{\omega}{2}\right],\\ \dfrac{1}{\omega}\left[x-\mu+\dfrac{\omega}{2}\right], & x\in\left(\mu-\dfrac{\omega}{2},\mu+\dfrac{\omega}{2}\right],\\ 1, & x\in\left(\mu+\dfrac{\omega}{2},\infty\right).\end{cases} \tag{4.1.2}$$

Graficele funcțiilor *f* şi *F* sunt date în Figura 4.1.1 şi respectiv Figura 4.1.2.

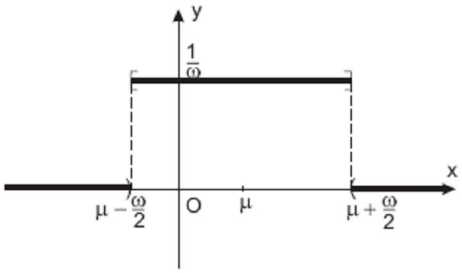


Figura 4.1.1

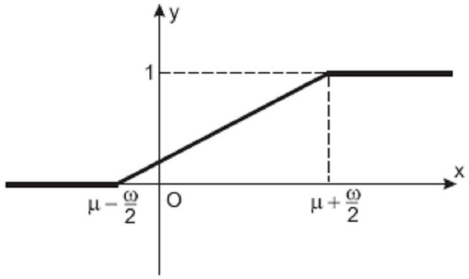


Figura 4.1.2

Teorema 4.1.2. *Dacă variabila aleatoare X are repartiție uniformă cu parametrii* μ şi ω, *atunci valoarea medie şi dispersia sa sunt*

$$E(X)=\mu,\quad Var(X)=\frac{\omega^2}{12}. \tag{4.1.3}$$

Demonstrație

Valoarea medie a variabilei *X* este

$$E(X)=\int_{-\infty}^{\infty}xf(x)\,dx=\int_{\mu-\frac{\omega}{2}}^{\mu+\frac{\omega}{2}}\frac{x}{\omega}dx=\frac{1}{2\omega}x^2\Big|_{\mu-\frac{\omega}{2}}^{\mu+\frac{\omega}{2}}=\mu,$$

iar dispersia sa este

$$Var(X)=\int_{-\infty}^{\infty}(x-\mu)^2\,f(x)\,dx=\int_{\mu-\frac{\omega}{2}}^{\mu+\frac{\omega}{2}}(x-\mu)^2\,\frac{1}{\omega}dx=\frac{1}{3\omega}(x-\mu)^3\Big|_{\mu-\frac{\omega}{2}}^{\mu+\frac{\omega}{2}}=\frac{\omega^2}{12}.$$

q.e.d.

Propoziția 4.1.3. *Funcția caracteristică a unei variabile aleatoare X cu repartiție uniformă cu parametrii* μ *şi* ω *este*

$$\varphi(t)=\begin{cases}\dfrac{i}{t\omega}\left(e^{it\left(\mu-\frac{\omega}{2}\right)}-e^{it\left(\mu+\frac{\omega}{2}\right)}\right), & t\in R^*,\\ 1, & t=0.\end{cases} \qquad (4.1.4)$$

Demonstrație
Pentru $t\neq 0,$ avem

$$\varphi(t)=\int_{-\infty}^{\infty}e^{itx}f(x)\,dx=\frac{1}{\omega}\int_{\mu-\frac{\omega}{2}}^{\mu+\frac{\omega}{2}}e^{itx}\,dx=\frac{1}{it\omega}e^{itx}\Big|_{\mu-\frac{\omega}{2}}^{\mu+\frac{\omega}{2}}=\frac{i}{t\omega}\left(e^{it\left(\mu-\frac{\omega}{2}\right)}-e^{it\left(\mu+\frac{\omega}{2}\right)}\right),$$

iar $\varphi(0)=\int_{-\infty}^{\infty}f(x)\,dx=1.$ *q.e.d.*

Aplicația 4.1.4. *Să se calculeze momentele* $m_2(X)$, $m_3(X)$ *şi* $\mu_3(X)$ *pentru o variabilă aleatoare X cu repartiție uniformă cu parametrii* μ *şi* ω.

Rezolvare
Momentul inițial de ordinul al doilea este

$$m_2(X)=\int_{-\infty}^{\infty}x^2f(x)\,dx=\int_{\mu-\frac{\omega}{2}}^{\mu+\frac{\omega}{2}}\frac{1}{\omega}x^2\,dx=\frac{1}{3\omega}x^3\Big|_{\mu-\frac{\omega}{2}}^{\mu+\frac{\omega}{2}}=\frac{1}{3\omega}\left[\left(\mu+\frac{\omega}{2}\right)^3\right.$$
$$\left.-\left(\mu-\frac{\omega}{2}\right)^3\right]=\frac{1}{3\omega}\left(3\mu^2\omega+\frac{\omega^3}{4}\right)=\mu^2+\frac{\omega^2}{12}.$$

Apoi momentul inițial de ordinul al treilea este

$$m_3(X)=\int_{-\infty}^{\infty}x^3f(x)\,dx=\int_{\mu-\frac{\omega}{2}}^{\mu+\frac{\omega}{2}}x^3\frac{1}{\omega}dx=\frac{1}{4\omega}x^4\Big|_{\mu-\frac{\omega}{2}}^{\mu+\frac{\omega}{2}}=\frac{1}{4\omega}\left[\left(\mu+\frac{\omega}{2}\right)^4\right.$$
$$\left.-\left(\mu-\frac{\omega}{2}\right)^4\right]=\frac{1}{4\omega}(4\mu^3\omega+\mu\omega^3)=\mu^3+\frac{\mu\omega^2}{4},$$

iar momentul centrat de ordinul al treilea este

$$\mu_3(X)=\int_{-\infty}^{\infty}(x-\mu)^3f(x)\,dx=\int_{\mu-\frac{\omega}{2}}^{\mu+\frac{\omega}{2}}(x-\mu)^3\frac{1}{\omega}dx=\frac{1}{4\omega}(x-\mu)^4\Big|_{\mu-\frac{\omega}{2}}^{\mu+\frac{\omega}{2}}=0.$$

Aplicația 4.1.5. *Variabilele aleatoare X şi Y sunt independente. X are repartiție uniformă pe intervalul* [0,1], *iar Y are repartiție uniformă pe intervalul* [0,2]. *Să se calculeze densitatea de probabilitate a variabilei X+Y.*

Rezolvare

Dacă f_1 şi f_2 sunt densitățile de probabilitate ale variabilelor aleatoare *X* şi respectiv *Y*, atunci $\int_{-\infty}^{\infty} f_1(x)\,dx = 1$ şi $\int_{-\infty}^{\infty} f_2(x)\,dx = 1.$ Deducem din aceste relații că parametrii variabilei *X* sunt $\omega_1 = 1$ şi $\mu_1 = \frac{1}{2},$ iar parametrii variabilei *Y* sunt $\omega_2 = 2$ şi $\mu_2 = 1.$ Atunci densitățile de probabilitate vor avea forma

$$f_1(x) = \begin{cases} 1, & x \in [0,1], \\ 0, & x \notin [0,1], \end{cases} \qquad f_2(x) = \begin{cases} \frac{1}{2}, & x \in [0,2], \\ 0, & x \notin [0,2]. \end{cases}$$

Deoarece *X* şi *Y* sunt independente, densitatea de probabilitate a variabilei *X+Y* este dată de formula

$$f(x) = \int_{-\infty}^{\infty} f_1(x-y)\, f_2(y)\, dy.$$

Observăm că $f_1(x-y) \neq 0$ pentru $x - y \in [0,1]$ sau echivalent $y \in [x-1, x]$, iar $f_2(y) \neq 0$ pentru $y \in [0,2]$. Deci $f_1(x-y)\, f_2(y) \neq 0$ pentru $y \in [x-1, x] \bigcap [0,2].$ Avem următoarele patru cazuri:

a) Dacă $x \notin [0,3],$ atunci $[x-1, x] \bigcap [0,2] = \Phi,$ deci $f(x) = 0.$

b) Dacă $x \in [0,1),$ atunci $[x-1, x] \bigcap [0,2] = [0, x]$ şi

$$f(x) = \int_0^x f_1(x-y)\, f_2(y)\, dy = \int_0^x \frac{1}{2}\, dy = \frac{x}{2}.$$

c) Dacă $x \in [1,2),$ atunci $[x-1, x] \bigcap [0,2] = [x-1, x]$ şi

$$f(x) = \int_{x-1}^x f_1(x-y)\, f_2(y)\, dy = \int_{x-1}^x \frac{1}{2}\, dy = \frac{1}{2}.$$

d) Dacă $x \in [2,3],$ atunci $[x-1, x] \bigcap [0,2] = [x-1, 2]$ şi

$$f(x) = \int_{x-1}^2 f_1(x-y)\, f_2(y)\, dy = \int_{x-1}^2 \frac{1}{2}\, dy = \frac{3-x}{2}.$$

Deci densitatea de probabilitate a variabilei $X+Y$ este funcția

$$f(x)=\begin{cases} x/2, & x\in[0,1),\\ 1/2, & x\in[1,2),\\ (3-x)/2, & x\in[2,3],\\ 0, & x\notin[0,3]. \end{cases}$$

4.2. Legea normală (Gauss-Laplace). Legea normală standard (legea normală centrată redusă)

Definiția 4.2.1. *Variabila aleatoare X urmează* ***legea normală (Gauss-Laplace)*** *(X are repartiție normală) cu parametrii* m *şi* σ $(m\in R, \sigma>0)$ *dacă densitatea sa de probabilitate (repartiție) este funcția*

$$f(x;m,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad \forall x\in R. \tag{4.2.1}$$

O variabilă aleatoare cu repartiție normală cu parametrii m şi σ se notează cu $N(m,\sigma^2)$.

Funcția f de mai sus se numește *densitatea de repartiție normală* sau *gaussiană*. Observăm că f este o densitate de probabilitate, deoarece $f(x)>0,\ \forall x\in R$ şi $\int_{-\infty}^{\infty} f(x)\,dx=1$. Într-adevăr, pentru a verifica ultima relație, în integrala de mai sus facem schimbarea de variabilă $\frac{x-m}{\sigma\sqrt{2}}=y$. Rezultă că $dx=\sigma\sqrt{2}\,dy$. Dacă $x\to-\infty$ atunci $y\to-\infty$, iar dacă $x\to\infty$ atunci $y\to\infty$. Obținem astfel

$$\int_{-\infty}^{\infty} f(x)\,dx=\frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-y^2}\,dy=\frac{2}{\sqrt{\pi}}\int_{0}^{\infty} e^{-y^2}\,dy=1.$$

Am folosit mai sus integrala lui Euler-Poisson $\int_0^{\infty} e^{-y^2}\,dy=\sqrt{\pi}/2$.

Graficul funcției f are formă de clopot (vezi Figura 4.2.1). Dreapta de ecuație $x=m$ este axă de simetrie pentru acest grafic, iar pentru $x=m$ se obține valoarea maximă a funcției f, şi anume $\frac{1}{\sigma\sqrt{2\pi}}$. Punctele $x=m-\sigma$ şi $x=m+\sigma$ sunt puncte de inflexiune.

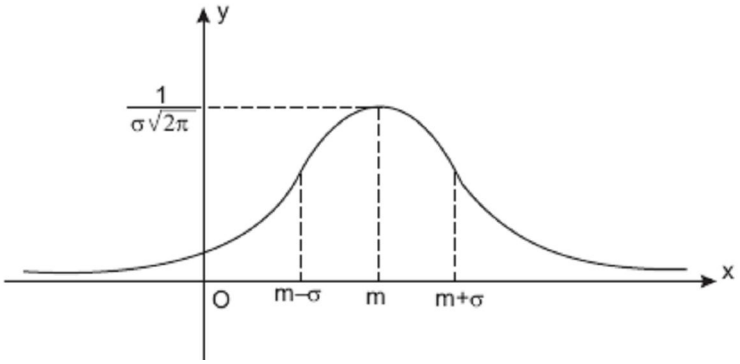


Figura 4.2.1

Pentru $m = 0$ şi $\sigma = 1$ funcția *f* dată de relația (4.2.1) devine

$$f(x;0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in R. \tag{4.2.2}$$

Vom spune despre o variabilă aleatoare X care are ca densitate de probabilitate funcția (4.2.2) că urmează **legea normală standard** sau **legea normală centrată redusă**.

Pentru a determina funcția de repartiție $F(x;m,\sigma)$ a unei variabile aleatoare X cu repartiție normală cu parametrii m şi σ, vom determina mai întâi funcția de repartiție pentru o variabilă aleatoare cu repartiție normală standard, notată cu $F(x;0,1)$ şi numită funcția de repartiție normală standard. Conform relației de legătură dintre *f* şi *F*, avem

$$F(x;0,1) = \int_{-\infty}^{x} f(t;0,1)\,dt = \int_{-\infty}^{0} f(t;0,1)\,dt + \int_{0}^{x} f(t;0,1)\,dt = \frac{1}{2} + \Phi(x), \tag{4.2.3}$$

unde $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2}\,dt$. Funcția Φ de mai sus se numeşte *funcția integrală a lui Laplace*, pentru valorile căreia sunt întocmite tabele.

Dacă variabila aleatoare X urmează legea normală cu parametrii m şi σ, atunci variabila aleatoare $Y = \frac{1}{\sigma}(X - m)$ urmează legea normală cu parametrii 0 şi 1. Într-adevăr, dacă F_1 este funcția de repartiție a variabilei *Y*, atunci

$$F_1(x) = P(Y \le x) = P(X \le m + \sigma x) = F(m + \sigma x; m, \sigma), \tag{4.2.4}$$

iar densitatea de probabilitate a variabilei *Y* este

$$f_1(x) = F_1'(x) = \sigma\, f(m+\sigma x; m, \sigma) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = f(x;0,1), \quad \forall\, x \in R.$$

Deci $F_1(x) = F(x;0,1),\ x \in R,$ adică variabila Y urmează legea normală cu parametrii 0 şi 1. Din ultima relație şi relația (4.2.4) obținem

$$F(x;0,1) = F(m+\sigma x; m, \sigma),\ \ x \in R. \tag{4.2.5}$$

Rezultă astfel că pentru variabila aleatoare X cu repartiție normală cu parametrii m şi σ, funcția de repartiție este

$$F(x;m,\sigma) = F\left(\frac{x-m}{\sigma};0,1\right) = \frac{1}{2} + \Phi\left(\frac{x-m}{\sigma}\right). \tag{4.2.6}$$

Teorema 4.2.2. *Dacă variabila aleatoare X are repartiție normală cu parametrii* m *şi* σ, *atunci valoarea medie şi dispersia sa sunt*

$$E(X) = m, \quad Var(X) = \sigma^2. \tag{4.2.7}$$

Demonstrație

Valoarea medie a variabilei aleatoare X este

$$E(X) = \int_{-\infty}^{\infty} x f(x;m,\sigma)\, dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx.$$

În integrala de mai sus vom face schimbarea de variabilă $\frac{x-m}{\sigma} = y,$ de unde rezultă $dx = \sigma\, dy;$ pentru $x \to -\infty$ rezultă $y \to -\infty,$ iar pentru $x \to \infty$ rezultă $y \to \infty.$ Deci obținem

$$E(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma y + m) e^{-y^2/2} \sigma\, dy = \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y e^{-y^2/2}\, dy$$
$$+ \frac{m}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} dy = m \int_{-\infty}^{\infty} f(y;0,1)\, dy = m.$$

Dispersia variabilei X este

$$Var(X) = \int_{-\infty}^{\infty} (x-m)^2 f(x;m,\sigma)\, dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-m)^2 e^{-\frac{(x-m)^2}{2\sigma^2}}\, dx.$$

Folosind schimbarea de variabilă de mai sus, obținem

$$Var(X)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\sigma^2y^2e^{-y^2/2}\sigma\,dy=\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}y^2e^{-y^2/2}dy$$

$$=\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}y\left(ye^{-y^2/2}\right)dy=-\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}y\left(e^{-y^2/2}\right)'dy=-\frac{\sigma^2}{\sqrt{2\pi}}ye^{-y^2/2}\Big|_{-\infty}^{\infty}$$

$$+\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-y^2/2}\,dy=\sigma^2.$$

Deducem de aici că abaterea medie pătratică a variabilei *X* este $\sigma_X=\sigma$. *q.e.d.*

Propoziția 4.2.3. *Dacă variabila aleatoare X are repartiție normală cu parametrii* m *şi* σ, *atunci funcția sa caracteristică este*

$$\varphi(t)=e^{imt-\frac{\sigma^2t^2}{2}},\quad t\in R. \tag{4.2.8}$$

Demonstrație

Funcția caracteristică a variabilei *X* este

$$\varphi(t)=\int_{-\infty}^{\infty}e^{itx}f(x)\,dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{itx}\cdot e^{-\frac{(x-m)^2}{2\sigma^2}}dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{x^2-2mx+m^2-2\sigma^2itx}{2\sigma^2}}dx$$

$$=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{x^2-2x(m+\sigma^2it)+(m+\sigma^2it)^2+m^2}{2\sigma^2}}\cdot e^{\frac{(m+\sigma^2it)^2}{2\sigma^2}}dx$$

$$=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{m^2+\sigma^4i^2t^2+2mit\sigma^2-m^2}{2\sigma^2}}\int_{-\infty}^{\infty}e^{-\frac{[x-(m+\sigma^2it)]^2}{2\sigma^2}}dx,\quad t\in R.$$

Folosind schimbarea de variabilă $\frac{x-(m+\sigma^2it)}{\sigma\sqrt{2}}=u$, obținem

$$\varphi(t)=\frac{1}{\sqrt{\pi}}e^{imt-\frac{\sigma^2t^2}{2}}\int_{-\infty}^{\infty}e^{-u^2}du=e^{imt-\frac{\sigma^2t^2}{2}},\quad \forall t\in R.\,q.e.d.$$

Propoziția 4.2.4. *Dacă variabila aleatoare X are repartiție normală cu parametrii* m *şi* σ, *iar* $a, b, k\in R,\ k>0$, *atunci*

a) $P(a<X<b)=\Phi\left(\frac{b-m}{\sigma}\right)-\Phi\left(\frac{a-m}{\sigma}\right);$ (4.2.9)

b) $P(|X-m|<k\sigma)=2\Phi(k).$ (4.2.10)

Demonstrație

a) Din relația (4.2.6) şi continuitatea funcției F deducem că

$$P(a < X < b) = F(b;m,\sigma) - F(a;m,\sigma) = \left[\frac{1}{2} + \Phi\left(\frac{b-m}{\sigma}\right)\right] - \left[\frac{1}{2} + \Phi\left(\frac{a-m}{\sigma}\right)\right]$$

$$= \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right).$$

b) Conform relației (4.2.9) obținem

$$P(|X-m| < k\sigma) = P(m-k\sigma < X < m+k\sigma) = \Phi(k) - \Phi(-k) = 2\Phi(k),$$

deoarece funcția Φ este impară $(\Phi(-k) = -\Phi(k))$. *q.e.d.*

Observația 4.2.5. *Dacă luăm* $k = 3$ *în relația (4.2.10) rezultă*

$$P(|X-m| < 3\sigma) = 2\Phi(3) \cong 0{,}9974. \qquad (4.2.11)$$

Relația (4.2.11) ne spune că aproape toate valorile variabilei X sunt situate în intervalul $(m-3\sigma, m+3\sigma)$. *Egalitatea din (4.2.11) este cunoscută sub numele de regula celor şase* σ.

Propoziția 4.2.6. *Dacă variabila aleatoare X are repartiție normală cu parametrii* m *şi* σ, *atunci momentele sale centrate sunt*

$$\mu_{2p-1}(X) = 0, \quad \mu_{2p}(X) = (2p-1)!!\sigma^{2p}, \quad \forall\, p \in N^*, \qquad (4.2.12)$$

unde $(2p-1)!! = 1 \cdot 3 \cdot 5 \cdots (2p-1)$.

Demonstrație

Momentul centrat de ordinul *k* este

$$\mu_k(X) = \int_{-\infty}^{\infty} (x-m)^k f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-m)^k e^{-\frac{(x-m)^2}{2\sigma^2}}\,dx. \qquad (4.2.13)$$

Obținem astfel

$$\mu_0(X) = \int_{-\infty}^{\infty} f(x)\,dx = 1, \quad \mu_1(X) = \int_{-\infty}^{\infty} (x-m) f(x)\,dx = 0, \quad \mu_2(X) = Var(X) = \sigma^2.$$

Pentru $k \geq 2$, în integrala din relația (4.2.13) vom face schimbarea de variabilă $\frac{x-m}{\sigma\sqrt{2}} = y$, de unde deducem că $dx = \sigma\sqrt{2}\,dy$; apoi pentru $x \to -\infty$ rezultă $y \to -\infty$, iar pentru $x \to \infty$ rezultă $y \to \infty$. Obținem atunci

$$\mu_k(X) = \frac{(\sigma\sqrt{2})^k}{\sqrt{\pi}} \int_{-\infty}^{\infty} y^k e^{-y^2}\,dy = -\frac{(\sigma\sqrt{2})^k}{2\sqrt{\pi}} \int_{-\infty}^{\infty} y^{k-1}\left(e^{-y^2}\right)'\,dy$$

$$= -\frac{(\sigma\sqrt{2})^k}{2\sqrt{\pi}} y^{k-1}e^{-y^2}\Big|_{-\infty}^{\infty} + (k-1)\frac{(\sigma\sqrt{2})^k}{2\sqrt{\pi}} \int_{-\infty}^{\infty} y^{k-2}e^{-y^2}\,dy$$

$$= (k-1)\frac{(\sigma\sqrt{2})^2}{2} \cdot \frac{(\sigma\sqrt{2})^{k-2}}{\sqrt{\pi}} \int_{-\infty}^{\infty} y^{k-2}e^{-y^2}\,dy = (k-1)\sigma^2\mu_{k-2}.$$

Deci am dedus relația de recurență $\mu_k(X) = (k-1)\sigma^2\mu_{k-2}(X)$. Deoarece $\mu_0(X) = 1$, $\mu_1(X) = 0$, din relația de mai sus rezultă relațiile (4.2.12). q.e.d.

Propoziția 4.2.7. *Dacă variabiele aleatoare indepedente X şi Y au repartiții normale cu parametrii* m_1 *şi* σ_1, *respectiv* m_2 *şi* σ_2, *atunci variabilele X+Y şi X-Y au repartiții normale cu parametrii* $m = m_1 + m_2$ *şi* $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$, *respectiv* $m = m_1 - m_2$ *şi* $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$.

Demonstrație

Vom calcula mai întâi densitățile de probabilitate pentru variabilele *X+Y* şi *X-Y*, pentru cazul particular $m_1 = m_2 = 0$ şi $\sigma_1 = \sigma_2 = 1$. Dacă notăm cu *f* şi *g* densitățile de probabilitate ale variabilelor *X* şi *Y*, atunci conform relației (4.2.2) avem

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

Densitatea de probabilitate a variabilei *X+Y* este

$$h(x)=\int_{-\infty}^{\infty} f(x-y)g(y)\,dy=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\frac{(x-y)^2}{2}}\cdot e^{-\frac{y^2}{2}}\,dy=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\left(y^2-xy+\frac{x^2}{2}\right)}\,dy$$

$$=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\left(y-\frac{x}{2}\right)^2}\cdot e^{-\frac{x^2}{4}}dy \overset{y-x/2=z}{=} \frac{1}{2\pi}e^{-\frac{x^2}{4}}\int_{-\infty}^{\infty} e^{-z^2}\,dz=\frac{1}{2\sqrt{\pi}}e^{-\frac{x^2}{4}}$$

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

Deducem astfel că $X+Y$ urmează şi ea legea normală cu parametrii $m=0$ şi $\sigma=\sqrt{2}.$

Asemănător pentru densitatea de probabilitate a variabilei aleatoare X-Y obținem

$$k(x)=\int_{-\infty}^{\infty} f(x+y)g(y)\,dy=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\frac{(x+y)^2}{2}}\cdot e^{-\frac{y^2}{2}}\,dy=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\left(y^2+xy+\frac{x^2}{2}\right)}\,dy$$

$$=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\left(y+\frac{x}{2}\right)^2}\cdot e^{-\frac{x^2}{4}}dy \overset{y+x/2=z}{=} \frac{1}{2\pi}e^{-\frac{x^2}{4}}\int_{-\infty}^{\infty} e^{-z^2}\,dz=\frac{1}{2\sqrt{\pi}}e^{-\frac{x^2}{4}}$$

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

Rezultă că X-Y urmează legea normală cu parametrii $m=0$ şi $\sigma=\sqrt{2}.$

În cazul general al variabilelor X şi Y cu repartiții normale cu parametrii m_1 şi σ_1, respectiv m_2 şi σ_2, putem să facem un calcul asemănător celui de sus, sau se poate raționa mai simplu folosind funcțiile caracteristice ale variabilelor. Conform relației (4.2.8) funcțiile caracteristice ale variabilelor X şi Y sunt

$$\varphi_1(t)=e^{im_1t-\frac{\sigma_1^2t^2}{2}},\quad \varphi_2(t)=e^{im_2t-\frac{\sigma_2^2t^2}{2}},\quad \forall t\in R.$$

Atunci funcția caracteristică a variabilei aleatoare $X+Y$ este

$$\varphi(t)=\varphi_1(t)\cdot\varphi_2(t)=e^{i(m_1+m_2)t-\frac{\left(\sigma_1^2+\sigma_2^2\right)t^2}{2}},\quad \forall t\in R.$$

Se observă că φ este tocmai funcția caracteristică corespunzătoare legii normale cu parametrii $m=m_1+m_2$ şi $\sigma=\sqrt{\sigma_1^2+\sigma_2^2}$.

Asemănător funcția caracteristică a variabilei aleatoare X-Y este

$$\tilde{\varphi}(t) = \varphi_1(t) \cdot \varphi_2(-t) = e^{i(m_1 - m_2)t - \frac{(\sigma_1^2 + \sigma_2^2)t^2}{2}}, \quad \forall t \in R.$$

Se observă că $\tilde{\varphi}$ este funcția caracteristică corespunzătoare legii normale cu parametrii $m = m_1 - m_2$ *şi* $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$. *q.e.d.*

Legea normală reprezintă „cazul limită" al multor legi de probabilitate. Astfel menționăm următorul rezultat de legătură între repartiția normală şi repartiția Poisson.

Teorema 4.2.8. *Dacă variabila aleatoare* X_λ $(\lambda > 0)$ *are repartiție Poisson cu parametrul* λ, *atunci funcția de repartiție a variabilei aleatoare* $\frac{X_\lambda - \lambda}{\sqrt{\lambda}}$ *tinde către funcția de repartiție normală standard, pentru* $\lambda \to \infty$.

În ipotezele Teoremei 4.2.8 se mai spune că variabila aleatoare $\frac{X_\lambda - \lambda}{\sqrt{\lambda}}$ este asimptotic normală.

Legătura dintre repartiția binomială şi repartiția normală este dată în următoarea teoremă.

Teorema 4.2.9. (Moivre-Laplace) *Dacă variabila aleatoare* X_n *are repartiție binomială cu parametrii n şi p (p nu depinde de n), iar X are repartiție normală standard, atunci*

$$\lim_{n\to\infty} P\left(a < \frac{X_n - np}{\sqrt{npq}} < b\right) = P(a < X < b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx. \qquad (4.2.14)$$

Teorema 4.2.9 ne spune că pentru valori mari ale lui *n* putem folosi tabelele legii normale pentru studiul variabilelor aleatoare cu repartiții binomiale. Pentru $n \geq 30$, există următoarea regulă practică care ne permite să aproximăm repartiția binomială prin cea normală:

- Dacă $\min(np, nq) \geq 10$, atunci aproximarea este foarte bună.
- Dacă $\min(np, nq) \in (5,10)$, atunci aproximarea este acceptabilă, dacă nu este nevoie de mare precizie.
- Dacă $\min(np, nq) \leq 5$, atunci nu se foloseşte această aproximare.

Teorema lui Moivre-Laplace este un caz particular al aşa numitei *Teorema limită centrală,* care spune că funcția de repartiție a unei sume de variabile aleatoare independente tinde în condiții destul de generale către funcția de

repartiție normală. De fapt, prin Teorema limită centrală se înțelege un grup de teoreme care tratează problema repartiției limită a sumelor de variabile aleatoare (nu întotdeauna independente). Rezultatul cel mai general în cazul variabilelor aleatoare independente este teorema lui Lindeberg-Fellar. În aplicații este foarte util un caz particular al acestei teoreme, prezentat mai jos.

Teorema 4.2.10. (Liapunov) *Fie variabilele aleatoare independente* X_n, $n \geq 1$ *cu mediile şi dispersiile* $m_n = E(X_n)$, $\sigma_n^2 = Var(X_n)$, $n \geq 1$, *care au momente centrate absolute de ordinul al treilea* $\rho_n^3 = E(|X_n - m_n|^3)$, $n \geq 1$.

Dacă $\lim_{n\to\infty} \frac{\rho(n)}{\sigma(n)} = 0$, *unde* $\sigma(n) = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2}$ *şi* $\rho(n) = \sqrt[3]{\rho_1^3 + \rho_2^3 + \cdots + \rho_n^3}$, *atunci* $\lim_{n\to\infty} F_n(x) = F(x;0,1)$, *unde* F_n *este funcția de repartiție a variabilei aleatoare* $X = \frac{1}{\sigma(n)} \sum_{k=1}^{n} (X_k - m_k)$.

Aplicația 4.2.11. *Se aruncă o monedă de* 256 *de ori. Care este probabilitatea ca numărul de apariții ale „stemei" să fie cuprins între* 112 *şi* 144*?*

Rezolvare

Să notăm cu *X* variabila aleatoare care are ca valori numărul de apariții ale „stemei", atunci când se aruncă moneda de 256 de ori. Variabila *X* are repartiția binomială cu parametrii *n*=256 şi *p*=1/2 (probabilitatea ca la o aruncare să apară „stema"). În această aplicație trebuie să calculăm $P(112<X<144)$. Deoarece $E(X)=np=128$, iar $\sigma_X = \sqrt{npq} = 8$, atunci are loc relația

$$P(112 < X < 144) = P\left(-2 < \frac{X-128}{8} < 2\right).$$

Vom folosi Teorema 4.2.9 (Moivre-Laplace) şi vom aproxima repartiția $\frac{X-128}{8}$ cu repartiția normală standard *Y*. Obținem

$$P\left(-2 < \frac{X-128}{8} < 2\right) \cong P(-2 < Y < 2) = \Phi(2) - \Phi(-2) = 2\Phi(2) \cong 0{,}95.$$

Aplicația 4.2.12. *De câte ori trebuie să aruncăm un zar corect astfel încât probabilitatea ca abaterea frecvenței relative a feței* 1 *de la numărul p*=1/6 *să fie cuprinsă între* -0,03 *şi* 0,03*, este* 0,95*.*

Rezolvare

Dacă în *n* aruncări fața 1 apare de α_n ori, vom determina pe *n* astfel încât să aibă loc relația

$$P\left(-0{,}03 < \frac{\alpha_n}{n} - p < 0{,}03\right) = 0{,}95. \qquad (4.2.15)$$

Relația (4.2.15) se mai poate scrie astfel

$$P\left(\left|\frac{\alpha_n - np}{\sqrt{npq}}\right| < \frac{0{,}03\sqrt{n}}{\sqrt{pq}}\right) = 0{,}95, \quad q = 1 - p = \frac{5}{6}. \qquad (4.2.16)$$

Folosind relațiile (4.2.14) şi (4.2.9) (cu $m = 0,\ \sigma = 1$) pentru $b = -a = \frac{0{,}03\sqrt{n}}{\sqrt{pq}}$, din relația (4.2.16) deducem că

$$2\Phi\left(\frac{0{,}03\sqrt{n}}{\sqrt{pq}}\right) = 0{,}95 \Rightarrow \Phi\left(\frac{0{,}18\sqrt{n}}{\sqrt{5}}\right) = 0{,}475$$

$$\Rightarrow F\left(\frac{0{,}18\sqrt{n}}{\sqrt{5}}\right) = \frac{1}{2} + \Phi\left(\frac{0{,}18\sqrt{n}}{\sqrt{5}}\right) = 0{,}975.$$

Folosind un tabel cu valorile funcției Φ, rezultă că $0{,}18\sqrt{n}/\sqrt{5} = 1{,}96$. Obținem $n \cong 592{,}8$. Deci trebuie să aruncăm zarul de 593 de ori pentru a fi verificată relația (4.2.15).

Aplicația 4.2.13. *O maşină produce o piesă circulară. Piesa este bună dacă diametrul său d este cuprins între* 3,99 *cm şi* 4,01 *cm. Care este probabilitatea producerii unei piese defecte de către maşina respectivă, ştiind că d are repartiție normală cu media* 4,002 *cm şi abaterea medie pătratică* 0,005 *cm ?*

Rezolvare

Conform formulei (4.2.9), probabilitatea ca o piesă să fie bună este

$$P(3{,}99 < d < 4{,}01) = \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right) = \Phi\left(\frac{4{,}01 - 4{,}002}{0{,}005}\right)$$

$$-\Phi\left(\frac{3{,}99 - 4{,}002}{0{,}005}\right) = \Phi(1{,}6) - \Phi(-2{,}4) = \Phi(1{,}6) + \Phi(2{,}4) \cong 0{,}937.$$

Rezultă atunci că probabilitatea ca piesa să fie defectă este 1-0,937=0,063.

Aplicaţia 4.2.14. *O maşină produce o piesă circulară. Când maşina este bine reglată, diametrul d al pieselor are repartiţie normală cu media* 10 *cm şi abaterea medie pătratică* 0,08 *cm. Se iau la întâmplare* 4 *piese fabricate de maşină şi se constată că media aritmetică a diametrelor acestor piese este* 10,14 *cm. Se poate afirma că maşina s-a dereglat ?*

Rezolvare

Fie d_1, d_2, d_3, d_4 diametrele celor 4 piese alese la întâmplare. Acestea sunt 4 variabile aleatoare indepedente cu repartiţii normale cu media 10 cm şi abaterea medie pătratică 0,08. Atunci variabila $d = (d_1 + d_2 + d_3 + d_4)/4$ are de asemenea repartiţie normală cu media 10 cm şi abaterea medie pătratică 0,04 cm. Într-adevăr, avem

$$m = E(d) = E\left(\frac{d_1 + d_2 + d_3 + d_4}{4}\right) = \frac{1}{4}\sum_{i=1}^{4} E(d_i) = 10,$$

$$Var(d) = Var\left(\frac{d_1 + d_2 + d_3 + d_4}{4}\right) = \frac{1}{16}\sum_{i=1}^{4} Var(d_i) = 0{,}0016, \quad \sigma_d = 0{,}04.$$

Conform regulei celor şase σ (relaţia (4.2.11)), avem

$$P(|d - m| < 3\sigma) = 2\Phi(3) \cong 0{,}997 \Leftrightarrow P(m - 3\sigma < d < m + 3\sigma) \cong 0{,}997,$$

de unde rezultă că $P(9{,}88 < d < 10{,}12) \cong 0{,}997$. Din această ultimă relaţie deducem că d ia valori în afara intervalului (9,88; 10,12) cu o probabilitate mai mică de 0,003. Deci este aproape sigur că maşina s-a defectat.

4.3. Legea log-normală

Definiţia 4.3.1. *Variabila aleatoare X urmează* ***legea log-normală (logaritmică normală)*** *(X are repartiţie log-normală) cu parametrii* m *şi* σ $(m \in R, \sigma > 0)$ *dacă densitatea sa de probabilitate (repartiţie) este funcţia*

$$f(x; m, \sigma) = \begin{cases} \dfrac{1}{\sigma x\sqrt{2\pi}} e^{-\frac{(\ln x - m)^2}{2\sigma^2}}, & x > 0, \\ 0, & x \le 0. \end{cases} \qquad (4.3.1)$$

Funcţia f din (4.3.1) este o densitate de probabilitate, deoarece $f(x) \ge 0, \ \forall x \in R$ şi $\int_{-\infty}^{\infty} f(x)\,dx = 1$. Într-adevăr pentru ultima relaţie avem

$$\int_{-\infty}^{\infty} f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}}\int_0^{\infty}\frac{1}{x}e^{-\frac{(\ln x-m)^2}{2\sigma^2}}\,dx \overset{\frac{\ln x-m}{\sigma}=t}{=} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\,dt = 1.$$

Funcția de repartiție a variabilei log-normale X cu parametrii m şi σ este $F(x)=0$ pentru $x \le 0$, iar pentru $x>0$ este

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \frac{1}{\sigma\sqrt{2\pi}}\int_0^{x}\frac{1}{t}e^{-\frac{(\ln t-m)^2}{2\sigma^2}}\,dt \overset{\frac{\ln t-m}{\sigma}=u}{=} \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{\ln x-m}{\sigma}} e^{-\frac{u^2}{2}}\,du$$
$$= F\left(\frac{\ln x-m}{\sigma};0,1\right),$$

unde $F(x;0,1)$ este funcția de repartiție normală standard. Deducem din ultima relație că pentru valorile pozitive ale lui X, variabila $Y = \dfrac{\ln X - m}{\sigma}$ urmează legea normală standard.

Teorema 4.3.2. *Dacă variabila aleatoare X are repartiție log-normală cu parametrii m şi σ, atunci momentele inițiale de ordinul k sunt*

$$m_k(X) = e^{mk+\frac{\sigma^2k^2}{2,}}, \quad k \in N^*. \tag{4.3.2}$$

Demonstrație

Conform formulei pentru momentele inițiale de ordinul k, avem

$$m_k(X) = \int_{-\infty}^{\infty} x^k f(x)\,dx = \int_0^{\infty}\frac{1}{\sigma x\sqrt{2\pi}}x^k e^{-\frac{(\ln x-m)^2}{2\sigma^2}}\,dx \overset{\ln x=y}{=}$$
$$= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{ky}\cdot e^{-\frac{(y-m)^2}{2\sigma^2}}\,dy = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{y^2-2my+m^2-2\sigma^2yk}{2\sigma^2}}\,dy$$
$$= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{y^2-2y(m+\sigma^2k)+(m+\sigma^2k)^2-(m+\sigma^2k)^2+m^2}{2\sigma^2}}\,dy$$
$$= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{[y-(m+\sigma^2k)]^2-\sigma^4k^2-2m\sigma^2k}{2\sigma^2}}\,dy = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{\sigma^4k^2+2mk\sigma^2}{2\sigma^2}}\int_{-\infty}^{\infty} e^{-\frac{[y-(m+\sigma^2k)]^2}{2\sigma^2}}\,dy.$$

Notăm $(y-m-\sigma^2k)/(\sigma\sqrt{2}) = u$ şi obținem

$$m_k(X) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{\sigma^2k^2+2mk}{2}}\int_{-\infty}^{\infty} e^{-u^2}\sigma\sqrt{2\pi}\,du = e^{mk+\frac{\sigma^2k^2}{2}}.q.e.d.$$

Din relația (4.3.2) deducem că media şi dispersia variabilei X cu repartiție log-normală cu parametrii m şi σ sunt

$$E(X)=m_1(X)=e^{m+\frac{\sigma^2}{2}},\quad Var(X)=m_2(X)-[m_1(X)]^2=e^{2m+\sigma^2}(e^{\sigma^2}-1). \quad (4.3.3)$$

4.4. Legea gamma

Definiția 4.4.1. *Variabila aleatoare X urmează* ***legea gamma*** *(X are repartiție gamma) cu parametrul* $p\,(p>0)$ *dacă densitatea sa de probabilitate (repartiție) este funcția*

$$f(x)=\begin{cases}\dfrac{x^{p-1}e^{-x}}{\Gamma(p)}, & x>0,\\ 0, & x\le 0,\end{cases} \quad (4.4.1)$$

unde $\Gamma(p)=\int_0^\infty x^{p-1}e^{-x}\,dx,\ p>0$ *este integrala lui Euler de al doilea tip sau funcția gamma a lui Euler.*

Funcția f din (4.4.1) este o densitate de probabilitate, deoarece $f(x)\ge 0,\ \forall x\in R$ şi

$$\int_{-\infty}^{\infty} f(x)\,dx=\frac{1}{\Gamma(p)}\int_0^\infty x^{p-1}e^{-x}\,dx=1.$$

În propoziția următoare vom prezenta câteva proprietăți ale funcției Γ (pentru demonstrațiile lor vezi [21]).

Propoziția 4.4.2. *Integrala (funcția) lui Euler* $\Gamma(p)=\int_0^\infty x^{p-1}e^{-x}\,dx,\ p\in R$ *verifică următoarele proprietăți:*

a) $\Gamma(p)$ *este convergentă pentru p>0 şi divergentă pentru* $p\le 0$;
b) $\Gamma(p+1)=p\Gamma(p),\ \forall p>0$;
c) $\Gamma(n)=(n-1)!,\ \forall n\in N^*$;
d) $\Gamma(1/2)=\sqrt{\pi}$.

Teorema 4.4.3. *Dacă variabila aleatoare X are repartiție gamma cu parametrul p, atunci momentele inițiale de ordinul k sunt*

$$m_k(X) = p(p+1)\cdots(p+k-1),\ k \in N^*. \tag{4.4.2}$$

Demonstrație

Momentul inițial de ordinul *k* este

$$m_k(X) = \int_{-\infty}^{\infty} x^k f(x)\,dx = \frac{1}{\Gamma(p)}\int_0^{\infty} x^{k+p-1}e^{-x}\,dx = \frac{\Gamma(k+p)}{\Gamma(p)}$$

$$= \frac{(k+p-1)\Gamma(k+p-1)}{\Gamma(p)} = \cdots = \frac{(k+p-1)\cdots(p+1)p\Gamma(p)}{\Gamma(p)}$$
$$= (k+p-1)(k+p-2)\cdots(p+1)p.$$

Am folosit mai sus proprietatea b) din Propoziția 4.4.2. q.e.d.

Din relația (4.4.2) deducem că media şi dispersia variabilei *X* cu repartiție gamma cu parametrul *p* sunt

$$E(X) = m_1(X) = p,\quad Var(X) = m_2(X) - [m_1(X)]^2 = p. \tag{4.4.3}$$

Definiția 4.4.4. *Variabila aleatoare X urmează* ***legea gamma generalizată*** *(X are repartiție gamma generalizată) cu parametrii* $p > 0$ *şi* $\lambda > 0$ *dacă densitatea sa de probabilitate (repartiție) este funcția*

$$f(x) = \begin{cases} \dfrac{\lambda^p x^{p-1}e^{-\lambda x}}{\Gamma(p)}, & x > 0, \\ 0, & x \le 0, \end{cases} \tag{4.4.4}$$

Într-un mod asemănător cu demonstrația Teoremei 4.4.3 se arată următoarea teoremă.

Teorema 4.4.5. *Dacă variabila aleatoare X are repartiție gamma generalizată cu parametrii* $p > 0$ *şi* $\lambda > 0$*, atunci momentele inițiale de ordinul k sunt*

$$m_k(X) = \frac{p(p+1)\cdots(p+k-1)}{\lambda^k},\ k \in N^*, \tag{4.4.5}$$

deci media şi dispersia sa sunt

$$E(X) = m_1(X) = \frac{p}{\lambda},\quad Var(X) = m_2(X) - [m_1(X)]^2 = \frac{p}{\lambda^2}. \tag{4.4.6}$$

Propoziția 4.4.6. *Dacă variabila aleatoare X are repartiție gamma generalizată cu parametrii* $p > 0$ *şi* $\lambda > 0$, *atunci funcția sa caracteristică este*

$$\varphi(t) = \left(1 - \frac{it}{\lambda}\right)^{-p}, \quad t \in R. \tag{4.4.7}$$

Demonstrație

Pentru determinarea funcției caracteristice a variabilei *X* vom folosi formula care dă momentul de ordinul *k* al variabilei *X* în funcție de derivatele funcției caracteristice *c* în punctul 0, şi anume $m_k = \varphi^{(k)}(0) \cdot i^{-k}$, $(m_0 = 1)$, precum şi dezvoltarea în serie de puteri a funcției φ care are forma următoare

$$\varphi(t) = \sum_{k=0}^{\infty} \frac{\varphi^{(k)}(0)}{k!} t^k = \sum_{k=0}^{\infty} \frac{m_k i^k}{k!} t^k = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} m_k.$$

Folosind formula (4.4.5) pentru momentele inițiale ale variabilei *X*, obținem pentru funcția caracteristică formula

$$\varphi(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} m_k = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \cdot \frac{p(p+1)\cdots(p+k-1)}{\lambda^k}$$

$$= \sum_{k=0}^{\infty} \frac{p(p+1)\cdots(p+k-1)}{k!} \left(\frac{it}{\lambda}\right)^k = \left(1 - \frac{it}{\lambda}\right)^{-p}, \ \forall t \in R.$$

q.e.d.

Din relația (4.4.7) pentru $\lambda = 1$, deducem că funcția caracteristică a unei variabile *X* cu repartiție gamma cu parametrul *p* este

$$\varphi(t) = (1 - it)^{-p}, \quad t \in R. \tag{4.4.8}$$

Propoziția 4.4.7. *Dacă variabilele aleatoare independente X şi Y au repartiții gamma cu parametrii p şi respectiv q, atunci variabila X+Y are repartiție gamma cu parametrul p+q.*

Demonstrație

Să notăm cu *f* şi *g* densitățile de probabilitate pentru variabilele *X* şi *Y*. Deci

$$f(x)=g(x)=\begin{cases}\dfrac{1}{\Gamma(p)}x^{p-1}e^{-x}, & x>0,\\ 0, & x\le 0.\end{cases}$$

Densitatea de probabilitate h a variabilei $X+Y$ este $h(x)=\int_{-\infty}^{\infty} f(x-y)g(y)\,dy.$

Dacă $x\le 0$ atunci

$$h(x)=\int_{-\infty}^{0} f(x-y)g(y)\,dy+\int_{0}^{\infty} f(x-y)g(y)\,dy=0,$$

(în prima integrală $g(y)=0$, iar în a doua integrală $f(x-y)=0$).
Dacă $x>0,$ atunci $f(x-y)\neq 0$ pentru $x-y\ge 0\Leftrightarrow y\le x,$ iar $g(y)\neq 0$ pentru $y\ge 0.$ Deci $f(x-y)g(y)\neq 0$ dacă $y\in[0,x]$ şi atunci

$$h(x)=\int_0^x f(x-y)g(y)\,dy=\int_0^x \frac{1}{\Gamma(p)}(x-y)^{p-1}e^{-x+y}\,\frac{1}{\Gamma(q)}y^{q-1}e^{-y}\,dy$$

$$=\frac{e^{-x}}{\Gamma(p)\Gamma(q)}\int_0^x (x-y)^{p-1}y^{q-1}\,dy.$$

Pentru calculul ultimei integrale de mai sus facem schimbarea de variabilă $y=tx$, deci $dy=xdt$; pentru $y\to 0$ rezultă $t\to 0,$ iar pentru $y\to x$ rezultă $t\to 1.$ Obținem astfel

$$h(x)=\frac{e^{-x}}{\Gamma(p)\Gamma(q)}\int_0^1 (x-tx)^{p-1}(tx)^{q-1}x\,dt=\frac{e^{-x}x^{p+q-1}}{\Gamma(p)\Gamma(q)}\int_0^1 x^{q-1}(1-x)^{p-1}\,dt$$

$$=e^{-x}x^{p+q-1}\,\frac{B(q,p)}{\Gamma(p)\Gamma(q)}=\frac{e^{-x}x^{p+q-1}}{\Gamma(p+q)}.$$

Rezultă că variabila $X+Y$ are repartiție gamma cu parametrul $p+q$.

Concluzia acestei propoziții o putem obține mai direct folosind funcțiile caracteristice. Fie φ_1 şi φ_2 funcțiile caracteristice ale variabilelor X şi Y, şi anume

$$\varphi_1(t)=(1-it)^{-p},\quad \varphi_2(t)=(1-it)^{-q},\quad t\in R.$$

Atunci funcția caracteristică a variabilei aleatoare $X+Y$ este

$$\varphi(t)=\varphi_1(t)\varphi_2(t)=(1-it)^{-(p+q)},\quad t\in R.$$

Deducem din relația obținută că $X+Y$ are repartiție gamma cu parametrul $p+q$. q.e.d.

Propoziția 4.4.8. *Dacă variabila aleatoare X urmează legea normală cu parametrii m şi* σ, *atunci variabila aleatoare* $Y = \frac{1}{2\sigma^2}(X-m)^2$ *are o repartiție gamma cu parametrul 1/2.*

Demonstrație

Să notăm cu $G(x)$ funcția de repartiție a variabilei Y. Deoarece $Y \geq 0$, atunci pentru $x \leq 0$ rezultă că $F(x) = P(Y \leq x) = 0$. Pentru $x > 0$ avem

$$G(x) = P(Y \leq x) = P\left(\frac{(X-m)^2}{2\sigma^2} \leq x\right) = P\left(-\sqrt{x} \leq \frac{X-m}{\sigma\sqrt{2}} \leq \sqrt{x}\right)$$

$$= P(m-\sigma\sqrt{2x} \leq X \leq m+\sigma\sqrt{2x}) = \Phi\left(\frac{m+\sigma\sqrt{2x}-m}{\sigma}\right) - \Phi\left(\frac{m-\sigma\sqrt{2x}-m}{\sigma}\right)$$

$$= 2\Phi(\sqrt{2x}) = \frac{\sqrt{2}}{\sqrt{\pi}}\int_0^{\sqrt{2x}} e^{-t^2/2}\,dt.$$

Atunci pentru $x > 0$ densitatea de probabilitate g a variabilei aleatoare Y este

$$g(x) = G'(x) = \frac{1}{\sqrt{\pi x}}e^{-x} = \frac{1}{\sqrt{\pi}}x^{-1/2}e^{-x},$$

iar pentru $x \leq 0$, $g(x)=0$.

Deoarece $\sqrt{\pi} = \Gamma(1/2)$ (conform Propoziției 4.4.2, d)), deducem forma funcției g, şi anume

$$g(x) = \begin{cases} \frac{1}{\Gamma(1/2)}x^{-1/2}e^{-x}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

adică variabila Y are repartiție gamma cu parametrul 1/2. q.e.d.

Aplicația 4.4.9. *Să se calculeze momentele centrate* $\mu_3(X)$ *şi* $\mu_4(X)$ *pentru o variabilă aleatoare X cu repartiție gamma cu parametrul p.*

Rezolvare

Momentul centrat de ordinul al treilea este

$$\mu_3(X)=\int_{-\infty}^{\infty}(x-p)^3 f(x)\,dx=\int_{-\infty}^{\infty}x^3 f(x)\,dx-3p\int_{-\infty}^{\infty}x^2 f(x)\,dx+3p^2\int_{-\infty}^{\infty}xf(x)\,dx$$
$$-p^3\int_{-\infty}^{\infty}f(x)\,dx=m_3-3pm_2+3p^2m_1-p_3=p(p+1)(p+2)-3p^2(p+1)$$
$$+3p^3-p^3=2p,$$

iar momentul centrat de ordinul al patrulea este

$$\mu_4(X)=\int_{-\infty}^{\infty}(x-p)^4 f(x)\,dx=\int_{-\infty}^{\infty}x^4 f(x)\,dx-4p\int_{-\infty}^{\infty}x^3 f(x)\,dx+6p^2\int_{-\infty}^{\infty}x^2 f(x)\,dx$$
$$-4p^3\int_{-\infty}^{\infty}xf(x)\,dx+p^4\int_{-\infty}^{\infty}f(x)\,dx=m_4-4pm_3+6p^2m_2-4p^3m_1+p^4$$
$$=p(p+1)(p+2)(p+3)-4p^2(p+1)(p+2)+6p^3(p+1)-4p^4+p^4=3p^2+6p.$$

4.5. Legea beta

Definiția 4.5.1. *Variabila aleatoare X urmează* ***legea beta*** *(X are repartiție beta) cu parametrii p şi* $q\,(p,q>0)$ *dacă densitatea sa de probabilitate (repartiție) este funcția*

$$f(x)=\begin{cases}\dfrac{1}{B(p,q)}x^{p-1}(1-x)^{q-1}, & x\in(0,1),\\ 0, & x\notin(0,1),\end{cases}\qquad(4.5.1)$$

unde $B(p,q)=\int_0^1 x^{p-1}(1-x)^{q-1}\,dx$, $p,q>0$ *este integrala lui Euler de primul tip sau funcția beta a lui Euler.*

Funcția f din relația (4.5.1) este o densitate de probabilitate, deoarece $f(x)\geq 0,\ \forall x\in R$ şi

$$\int_{-\infty}^{\infty}f(x)\,dx=\frac{1}{B(p,q)}\int_0^1 x^{p-1}(1-x)^{q-1}\,dx=1.$$

În propoziția următoare vom prezenta câteva proprietăți ale funcției B (pentru demonstrațiile lor vezi [21]).

Propoziția 4.5.2. *Integrala lui Euler* $B(p,q)=\int_0^1 x^{p-1}(1-x)^{q-1}\,dx$, $p,q\in R$ *verifică următoarele proprietăți*

a) $B(p,q)$ *este convergentă pentru p>0 şi q>0, şi în rest este divergentă;*

b) $B(p,q) = B(q,p),\ \forall\, p,q > 0;$

c) $B(p,q) = \dfrac{p-1}{p+q-1} B(p-1,q),\ \forall\, p > 1,\ q > 0;$

d) $B(p,q) = \dfrac{q-1}{p+q-1} B(p,q-1),\ \forall\, p > 0,\ q > 1;$

e) $B(p,q) = \dfrac{(p-1)(q-1)}{(p+q-1)(p+q-2)} B(p-1,q-1),\ \forall\, p > 1,\ q > 1;$

f) $B(p,n) = B(n,p) = \dfrac{(n-1)!}{p(p+1)\cdots(p+n-1)},\ \forall\, n \in N^*,\ p > 0;$

g) $B(m,n) = \dfrac{(m-1)!(n-1)!}{(m+n-1)!},\ \forall\, m, n \in N^*;$

h) $B(p+1,q) + B(p,q+1) = B(p,q),\ \forall\, p, q > 0;$

i) $qB(p+1,q) = pB(p,q+1),\ \forall\, p, q > 0;$

j) $B(p,q) = \int_0^\infty \dfrac{t^{p-1}}{(1+t)^{p+q}} dt,\ \forall\, p, q > 0;$

k) $B(p,q) = \dfrac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)},\ \forall\, p, q > 0;$

l) $B(p,1-p) = \Gamma(p)\Gamma(1-p) = \dfrac{\pi}{\sin(p\pi)},\ \forall\, p \in (0,1).$

Teorema 4.5.3. *Dacă variabila aleatoare X are repartiție beta cu parametrii p şi* $q\,(p, q > 0)$, *atunci momentele inițiale de ordinul k sunt*

$$m_k(X) = \frac{p(p+1)\cdots(p+k-1)}{(p+q)(p+q+1)\cdots(p+q+k-1)},\ k \in N^*. \qquad (4.5.2)$$

Demonstrație
Momentul inițial de ordinul *k* este

$$m_k(X)=\int_{-\infty}^{\infty}x^k f(x)\,dx=\frac{1}{B(p,q)}\int_0^1 x^{k+p-1}(1-x)^{q-1}\,dx=\frac{B(k+p,q)}{B(p,q)}$$
$$=\frac{k+p-1}{p+q+k-1}\cdot\frac{B(k+p-1,q)}{B(p,q)}=\cdots=\frac{(k+p-1)\cdots(p+1)p}{(p+q+k-1)\cdots(p+q)}\cdot\frac{B(p,q)}{B(p,q)}$$
$$=\frac{p(p+1)\cdots(p+k-1)}{(p+q)(p+q+1)\cdots(p+q+k-1)},$$

conform Propoziției 4.5.2, c). q.e.d.

Din relația (4.5.2) deducem că media şi dispersia variabilei *X* cu repartiție beta cu parametrii *p* şi *q* sunt

$$E(X)=m_1(X)=\frac{p}{p+q},\quad Var(X)=m_2(X)-[m_1(X)]^2=\frac{pq}{(p+q)^2(p+q+1)}. \tag{4.5.3}$$

4.6. Legea χ^2 (Helmert-Pearson)

Definiția 4.6.1. *Variabila aleatoare X urmează* ***legea*** χ^2 ***(Helmert-Pearson)*** *(X are repartiție* χ^2*) cu parametrul n (*$n\in N^*$*) dacă densitatea sa de probabilitate (repartiție) este funcția*

$$f(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}, & x>0,\\ 0, & x\le 0.\end{cases} \tag{4.6.1}$$

Parametrii *n* din Definiția 4.6.1 se mai numesc *grade de libertate*, astfel că o să mai spunem că *X* are repartiție χ^2 cu *n* grade de libertate.

Funcția *f* din relația (4.6.1) este o densitate de probabilitate deoarece $f(x)\ge 0,\ \forall x\in R$ şi

$$\int_{-\infty}^{\infty}f(x)\,dx=\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}\int_0^{\infty}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}\,dx=1.$$

Pentru verificarea relației de mai sus , facem în integrala dată schimbarea de variabilă $x/2=y$, de unde rezultă $dx=2dy$; dacă $x \to 0$ atunci $y \to 0$, iar dacă $x \to \infty$ atunci $y \to \infty$. Obținem astfel

$$\int_{-\infty}^{\infty} f(x)\,dx = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_0^{\infty} (2y)^{\frac{n}{2}-1} e^{-y} \cdot 2\,dy = \frac{2^{\frac{n}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_0^{\infty} y^{\frac{n}{2}-1} e^{-y}\,dy$$

$$= \frac{1}{\Gamma\left(\frac{n}{2}\right)} \Gamma\left(\frac{n}{2}\right) = 1.$$

În Figura 4.6.1 sunt reprezentate grafic funcțiile f pentru diverse valori ale parametrului n.

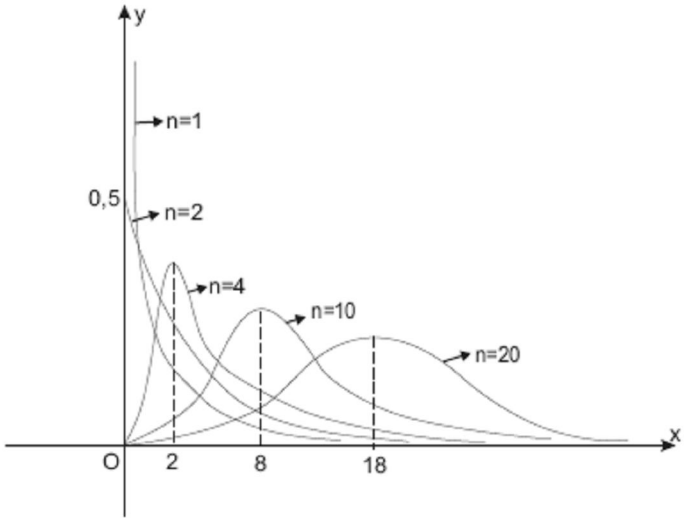


Figura 4.6.1

Teorema 4.6.2. *Dacă variabila aleatoare X are repartiție* χ^2 *cu n grade de libertate, atunci momentele inițiale de ordinul k sunt*

$$m_k(X) = n(n+2)\cdots(n+2k-2), \quad k \in N^*. \tag{4.6.2}$$

Demonstrație

Momentul inițial de ordinul k este

$$m_k(X) = \int_{-\infty}^{\infty} x^k f(x)\,dx = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \int_0^{\infty} x^{\frac{n}{2}+k-1} e^{-\frac{x}{2}}\,dx.$$

Cu aceeaşi schimbare de variabilă de mai sus folosită pentru calculul integralei $\int_{-\infty}^{\infty} f(x)dx$, obținem

$$m_k(X) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}\int_0^{\infty}(2y)^{\frac{n}{2}+k-1}e^{-y}\cdot 2\,dy = \frac{2^{\frac{n}{2}+k}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}\int_0^{\infty} y^{\frac{n}{2}+k-1}e^{-y}\,dy$$

$$= \frac{2^k\Gamma\left(\frac{n}{2}+k\right)}{\Gamma\left(\frac{n}{2}\right)} = 2^k\cdot\frac{n}{2}\left(\frac{n}{2}+1\right)\cdots\left(\frac{n}{2}+k-1\right) = n(n+2)\cdots(n+2k-2).$$

q.e.d.

Din relația (4.6.2) deducem că media şi dispersia variabilei *X* cu repartiție χ^2 cu *n* grade de libertate sunt

$$E(X) = m_1(X) = n, \quad Var(X) = m_2(X) - [m_1(X)]^2 = 2n. \tag{4.6.3}$$

Definiția 4.6.3. *Variabila aleatoare X urmează* ***legea*** χ^2 ***generalizată*** *(X are repartiție* χ^2 *generalizată) cu parametrii n şi* σ ($n \in N^*$, $\sigma > 0$)
dacă densitatea sa de probabilitate (repartiție) este funcția

$$f(x) = \begin{cases} \dfrac{1}{(\sigma\sqrt{2})^n\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}e^{-\frac{x}{2\sigma^2}}, & x > 0, \\ 0, & x \le 0. \end{cases} \tag{4.6.4}$$

Într-un mod asemănător cu demonstrația Teoremei 4.6.2 se arată următoarea teoremă.

Teorema 4.6.4. *Dacă variabila aleatoare X are repartiție* χ^2 *generalizată cu parametrii n şi* σ ($n \in N^*$, $\sigma > 0$), *atunci momentele inițiale de ordinul k sunt*

$$m_k(X) = n(n+2)\cdots(n+2k-2)\sigma^{2k}, \quad k \in N^*, \tag{4.6.5}$$

deci media şi dispersia sa sunt

$$E(X) = m_1(X) = n\sigma^2, \quad Var(X) = m_2(X) - [m_1(X)]^2 = 2n\sigma^4. \tag{4.6.6}$$

Propoziția 4.6.5. *Dacă variabila aleatoare X are repartiție* χ^2 *generalizată cu parametrii n şi* σ ($n \in N^*$, $\sigma > 0$), *atunci funcția sa caracteristică este*

$$\varphi(t) = (1 - 2it\sigma^2)^{-\frac{n}{2}}, \quad t \in R. \tag{4.6.7}$$

Demonstrație
Folosind formula din demonstația Propoziției 4.4.6, obținem

$$\varphi(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} m_k = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \cdot (2\sigma^2)^k \frac{n}{2}\left(\frac{n}{2}+1\right)\cdots\left(\frac{n}{2}+k-1\right)$$

$$= \sum_{k=0}^{\infty} (2it\sigma^2)^k \frac{\frac{n}{2}\left(\frac{n}{2}+1\right)\cdots\left(\frac{n}{2}+k-1\right)}{k!} = (1-2it\sigma^2)^{-\frac{n}{2}}, \ \forall t \in R.$$

q.e.d.

Din relația (4.6.7) pentru $\sigma = 1$, deducem că funcția caracteristică a unei variabile *X* cu rpartiție χ^2 cu *n* grade de libertate este

$$\varphi(t) = (1-2it)^{-\frac{n}{2}}, \quad t \in R. \tag{4.6.8}$$

Legăturile dintre repartiția χ^2 şi repartiția normală sunt prezentate în următoarele două teoreme.

Teorema 4.6.6. *Dacă variabila* X_n *are repartiție* χ^2 *cu n grade de libertate* ($n \geq 1$) *atunci densitatea de probabilitate a variabilei* $\frac{X_n - n}{\sqrt{2n}}$ *tinde pentru* $n \to \infty$ *către densitatea de repartiție normală standard.*

Teorema 4.6.7. *Dacă fiecare dintre variabilele aleatoare independente* $X_1, X_2, \ldots, X_n$ *are repartiție normală standard, atunci variabila aleatoare* $X = X_1^2 + X_2^2 + \cdots + X_n^2$ *are repartiție* χ^2 *cu n grade de libertate.*

Propoziția 4.6.8. *Dacă variabilele aleatoare independente X şi Y au repartiții* χ^2 *cu m, respectiv n grade de libertate, atunci X+Y are repartiție* χ^2 *cu m+n grade de libertate.*

Demonstrație
Să notăm cu *f* şi *g* densitățile de probabilitate ale variabilelor *X* şi *Y*, adică avem

$$f(x)=\begin{cases}\dfrac{1}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)}x^{\frac{m}{2}-1}e^{-\frac{x}{2}}, & x>0,\\ 0, & x\le 0,\end{cases}\qquad g(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}}, & x>0,\\ 0, & x\le 0.\end{cases}$$

Atunci densitatea de probabilitate a variabilei $X+Y$ este

$$h(x)=\int_{-\infty}^{\infty} f(x-y)g(y)\,dy.$$

Folosind un raționament asemănător celui întâlnit în demonstrația Propoziției 4.4.7, deducem că $h(x)=0,\ \forall x\le 0$, iar pentru $x>0$ obținem

$$h(x)=\int_0^x \frac{1}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)}(x-y)^{\frac{m}{2}-1}e^{-\frac{x-y}{2}}\cdot\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}y^{\frac{n}{2}-1}e^{-\frac{y}{2}}\,dy$$

$$=\frac{e^{-\frac{x}{2}}}{2^{\frac{m+n}{2}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\int_0^x (x-y)^{\frac{m}{2}-1}y^{\frac{n}{2}-1}\,dy.$$

Notăm $y=tx$, de unde rezultă $dy=xdt$; dacă $y\to 0$ atunci $t\to 0$, iar dacă $y\to x$ atunci $t\to 1$. Obținem

$$h(x)=\frac{e^{-\frac{x}{2}}}{2^{\frac{m+n}{2}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\int_0^1 (x-tx)^{\frac{m}{2}-1}(tx)^{\frac{n}{2}-1}x\,dt$$

$$=\frac{e^{-\frac{x}{2}}}{2^{\frac{m+n}{2}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}x^{\frac{m+n}{2}-1}\int_0^1 (1-t)^{\frac{m}{2}-1}t^{\frac{n}{2}-1}dt=\frac{e^{-\frac{x}{2}}x^{\frac{m+n}{2}-1}}{2^{\frac{m+n}{2}}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}B\left(\frac{m}{2},\frac{n}{2}\right)$$

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

Deducem că $X+Y$ are repartiție χ^2 cu $m+n$ grade de libertate.

Concluzia acestei propoziții o putem obține mai direct folosind funcțiile caracteristice. Fie φ_1 şi φ_2 funcțiile caracteristice ale variabilelor *X* şi *Y*, şi anume

$$\varphi_1(t) = (1-2it)^{-\frac{m}{2}}, \quad \varphi_2(t) = (1-2it)^{-\frac{n}{2}}, \quad t \in R.$$

Atunci funcția caracteristică a variabilei aleatoare $X+Y$ este

$$\varphi(t) = \varphi_1(t)\varphi_2(t) = (1-2it)^{-\frac{m+n}{2}}, \quad t \in R.$$

Deducem din relația obținută că $X+Y$ are repartiție χ^2 cu $m+n$ grade de libertate. q.e.d.

4.7. Legea Student (t). Legea Cauchy

Definiția 4.7.1. *Variabila aleatoare X urmează* ***legea Student (t)*** *(X are repartiție Student) cu parametrul n ($n \in N^*$) dacă densitatea sa de probabilitate (repartiție) este funcția*

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad x \in R. \qquad (4.7.1)$$

Parametrii *n* din Definiția 4.7.1 se mai numesc *grade de libertate*, astfel că o să mai spunem că *X* are repartiție Student cu *n* grade de libertate.

Funcția *f* din relația (4.7.1) este o densitate de probabilitate deoarece

$$f(x) \geq 0, \ \forall x \in R \text{ şi } \int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Pentru a verifica ultima relație, avem

$$\int_{-\infty}^{\infty} f(x)\,dx = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\int_{-\infty}^{\infty}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}dx = \frac{2\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\int_{0}^{\infty}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}dx.$$

În ultima integrala de mai sus facem schimbarea de variabilă $x^2/n = y,\ x \geq 0,$ de unde rezultă

$$dx = \frac{\sqrt{n}}{2\sqrt{y}}dy.$$

Intervalul $[0,\infty)$ se transformă în intervalul $[0,\infty)$ şi astfel obținem

$$\int_{-\infty}^{\infty} f(x)dx = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n}{2}\right)}\int_0^{\infty}\frac{y^{-1/2}}{(1+y)^{\frac{n}{2}+\frac{1}{2}}}dy = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n}{2}\right)}\cdot B\left(\frac{1}{2},\frac{n}{2}\right)$$

$$= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n}{2}\right)}\cdot\frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} = 1.$$

În relațiile de mai sus am folosit proprietățile j) şi k) ale funcției *B* din Propoziția 4.5.2 şi proprietatea d) a funcției Γ din Propoziția 4.4.2.

Teorema 4.7.2. *Dacă variabila aleatoare X are repartiție Student cu n grade de libertate, atunci momentele inițiale de ordinul k sunt*

$$\begin{aligned} &m_{2k+1}(X) = 0, \quad 2k+1 < n, \\ &m_{2k}(X) = \frac{n^k(2k-1)!!}{(n-2)(n-4)\cdots(n-2k)}, \quad 2k < n,\ k \in N^*. \end{aligned} \tag{4.7.2}$$

Demonstrație

Pentru $2k+1 < n,$ momentele de ordin impar sunt nule. Într-adevăr

$$m_{2k+1}(X) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\int_{-\infty}^{\infty} x^{2k+1}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}dx = 0, \tag{4.7.3}$$

deoarece funcția *f* este impară (integrala de mai sus este convergentă). Dacă $2k+1 \geq n$ integrala din (4.7.3) este divergentă, deci *X* nu are momente inițiale de ordinul $2k+1.$

Pentru momentul de ordin par $m_{2k}(X),$ cu $2k < n$ avem

$$m_{2k}(X)=\frac{2\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\int_0^{\infty}x^{2k}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}dx. \qquad (4.7.4)$$

Pentru calculul ultimei integrale folosim aceeaşi schimbare de variabilă ca cea folosită mai sus pentru verificarea relației $\int_{-\infty}^{\infty}f(x)\,dx=1.$ Notăm $x^2/n=y,\ x\geq 0,$ de unde rezultă $dx=\frac{\sqrt{n}}{2\sqrt{y}}dy.$ Intervalul $[0,\infty)$ se transformă în intervalul $[0,\infty)$ şi astfel obţinem

$$m_{2k}(X)=\frac{2\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\int_0^{\infty}(ny)^k(1+y)^{-\frac{n+1}{2}}\frac{\sqrt{n}}{2\sqrt{y}}dy=\frac{n^k\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n}{2}\right)}\int_0^{\infty}y^{k-\frac{1}{2}}(1+y)^{-\frac{n+1}{2}}dy$$

$$=\frac{n^k\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n}{2}\right)}\int_0^{\infty}y^{\left(k+\frac{1}{2}\right)-1}(1+y)^{-\left[\left(k+\frac{1}{2}\right)+\left(\frac{n}{2}-k\right)\right]}dy.$$

Folosind proprietăţile funcţiei *B* din Propoziţia 4.5.2 deducem din relaţia de mai sus

$$m_{2k}(X)=\frac{n^k\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n}{2}\right)}B\left(k+\frac{1}{2},\frac{n}{2}-k\right)=\frac{n^k\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n}{2}\right)}\cdot\frac{\Gamma\left(k+\frac{1}{2}\right)\Gamma\left(\frac{n}{2}-k\right)}{\Gamma\left(\frac{n+1}{2}\right)}$$

$$=\frac{n^k}{\sqrt{\pi}}\cdot\frac{\Gamma\left(k+\frac{1}{2}\right)\Gamma\left(\frac{n}{2}-k\right)}{\Gamma\left(\frac{n}{2}\right)}=\frac{n^k}{\sqrt{\pi}}\cdot\frac{\left(k-\frac{1}{2}\right)\Gamma\left(k-\frac{1}{2}\right)\Gamma\left(\frac{n}{2}-k\right)}{\left(\frac{n}{2}-1\right)\Gamma\left(\frac{n}{2}-1\right)}$$

$$= \frac{n^k}{\sqrt{\pi}} \cdot \frac{\left(k-\frac{1}{2}\right)\left(k-\frac{3}{2}\right)\Gamma\left(k-\frac{3}{2}\right)\Gamma\left(\frac{n}{2}-k\right)}{\left(\frac{n}{2}-1\right)\left(\frac{n}{2}-2\right)\Gamma\left(\frac{n}{2}-2\right)} = \cdots$$

$$= \frac{n^k}{\sqrt{\pi}} \cdot \frac{\left(k-\frac{1}{2}\right)\left(k-\frac{3}{2}\right)\cdots\frac{1}{2}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}-k\right)}{\left(\frac{n}{2}-1\right)\left(\frac{n}{2}-2\right)\cdots\left(\frac{n}{2}-k\right)\Gamma\left(\frac{n}{2}-k\right)} = \frac{n^k(2k-1)!!}{(n-2)(n-4)\cdots(n-2k)}.$$

Dacă $2k \geq n$ integrala din relația (4.7.4) este divergentă, deci variabila *X* nu are momente inițiale de ordinul *2k.* *q.e.d.*

Din relațiile (4.7.2) deducem că dacă $n>1$ atunci media variabilei *X* este $E(X)=0$, iar dacă $n>2$ atunci dispersia variabilei *X* este $Var(X)=\frac{n}{n-2}$.

Legăturile dintre repartiția Student şi repartiția normală sunt prezentate în următoarele două teoreme.

Teorema 4.7.3. *Dacă* f_n *este densitatea de probabilitate a unei repartiţii Student cu n grade de libertate atunci*

$$\lim_{n\to\infty} f_n(x) = f(x;0,1), \ \forall x \in R,$$

unde $f(x;0,1)$ *este densitatea de repartiţie normală standard.*

Teorema 4.7.4. *Dacă fiecare dintre variabilele aleatoare independente* $X_1, X_2, \ldots, X_{n+1}$ *are repartiţie normală cu parametrii* $m=0$ *şi* σ, *atunci variabila aleatoare*

$$X = \sqrt{n}\frac{X_{n+1}}{\sqrt{X_1^2 + \cdots X_n^2}}$$

are repartiţie Student cu n grade de libertate.

Pentru $n=1$ legea (repartiția) Student se mai numeşte **legea (repartiția) Cauchy.** O variabilă aleatoare *X* cu repartiție Cauchy are densitatea de probabilitate funcția

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad x \in R.$$

Din observațiile făcute după demonstația Teoremei 4.7.2 , deducem că variabila aleatoare *X* cu repartiția Cauchy nu are nici valoare medie şi nici dispersie.

4.8. Legea Snedecor. Legea Fisher

Definiția 4.8.1. *Variabila aleatoare X urmează* ***legea Snedecor*** *(X are repartiție Snedecor) cu parametrii* n_1 *şi* n_2, $(n_1, n_2 \in N^*)$ *dacă densitatea sa de probabilitate (repartiție) este funcția*

$$f(x)=\begin{cases}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}\cdot\frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}x^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}, & x>0,\\ 0, & x\le 0.\end{cases} \qquad (4.8.1)$$

Parametrii n_1 şi n_2 din Definiția 4.8.1 se mai numesc *grade de libertate,* astfel că o să mai spunem că *X* are repartiție Snedecor cu n_1 şi n_2 grade de libertate.

Funcția *f* din relația (4.8.1) este o densitate de probabilitate, deoarece $f(x)\ge 0,\ \forall x\in R$ şi $\int_{-\infty}^{\infty} f(x)\,dx=1.$ Pentru a verifica ultima relație, avem

$$\int_{-\infty}^{\infty} f(x)\,dx=\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}\cdot\frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\int_0^{\infty} x^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}dx.$$

În ultima integrală de mai sus facem schimbarea de variabilă $n_1 x/n_2 = y.$ Astfel obținem

$$\int_{-\infty}^{\infty} f(x)\,dx = \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \int_0^{\infty} \left(\frac{n_2 y}{n_1}\right)^{\frac{n_1}{2}-1} (1+y)^{-\frac{n_1+n_2}{2}} \frac{n_2}{n_1}\,dy$$

$$= \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \int_0^{\infty} y^{\frac{n_1}{2}-1} (1+y)^{-\frac{n_1+n_2}{2}}\,dy = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} B\left(\frac{n_1}{2}, \frac{n_2}{2}\right) = 1.$$

Teorema 4.8.2. *Dacă variabila aleatoare X are repartiție Snedecor cu parametrii* n_1 *şi* n_2, $(n_1, n_2 \in N^*)$, *atunci momentele inițiale de ordinul k sunt*

$$m_k(X) = \frac{n_2^k}{n_1^k} \cdot \frac{n_1(n_1+2)\cdots(n_1+2k-2)}{(n_2-2)(n_2-4)\cdots(n_2-2k)}, \quad 2k < n_2,\ k \in N^*. \qquad (4.8.2)$$

Demonstație

Variabila *X* are momente inițiale de ordinul *k* pentru $2k < n_2$. Avem

$$m_k(X) = \int_{-\infty}^{\infty} x^k f(x)\,dx = \int_0^{\infty} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} x^{k+\frac{n_1}{2}-1} \left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}} dx.$$

Pentru calculul ultimei integrale folosim aceeaşi schimbare de variabilă ca cea folosită mai sus pentru verificarea relației $\int_{-\infty}^{\infty} f(x)\,dx = 1$. Notăm $n_1 x / n_2 = y$ şi obținem

$$m_k(X) = \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \int_0^{\infty} \left(\frac{n_2 y}{n_1}\right)^{k+\frac{n_1}{2}-1} (1+y)^{-\frac{n_1+n_2}{2}} \cdot \frac{n_2}{n_1}\,dy$$

$$= \left(\frac{n_2}{n_1}\right)^k \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \int_0^\infty y^{k+\frac{n_1}{2}-1}(1+y)^{-\frac{n_1+n_2}{2}}\, dy$$

$$= \left(\frac{n_2}{n_1}\right)^k \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} B\left(k+\frac{n_1}{2}, \frac{n_2}{2}-k\right)$$

$$= \left(\frac{n_2}{n_1}\right)^k \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \cdot \frac{\Gamma\left(k+\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}-k\right)}{\Gamma\left(\frac{n_1+n_2}{2}\right)}$$

$$= \left(\frac{n_2}{n_1}\right)^k \frac{\left(k+\frac{n_1}{2}-1\right)\Gamma\left(\frac{n_1}{2}+k-1\right)\Gamma\left(\frac{n_2}{2}-k\right)}{\Gamma\left(\frac{n_1}{2}\right)\left(\frac{n_2}{2}-1\right)\Gamma\left(\frac{n_2}{2}-1\right)} = \cdots$$

$$= \left(\frac{n_2}{n_1}\right)^k \frac{\left(k+\frac{n_1}{2}-1\right)\left(k+\frac{n_1}{2}-2\right)\cdots\frac{n_1}{2}\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}-k\right)}{\Gamma\left(\frac{n_1}{2}\right)\left(\frac{n_2}{2}-1\right)\left(\frac{n_2}{2}-2\right)\cdots\left(\frac{n_2}{2}-k\right)\Gamma\left(\frac{n_2}{2}-k\right)}$$

$$= \left(\frac{n_2}{n_1}\right)^k \frac{n_1(n_1+2)\cdots(n_1+2k-2)}{(n_2-2)(n_2-4)\cdots(n_2-2k)}.$$

q.e.d.

Din relația (4.8.3) deducem că pentru $n_2 > 2$ media variabilei X este $E(X) = \frac{n_2}{n_2-2}$, iar pentru $n_2 > 4$ dispersia variabilei X este

$$Var(X) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-4)(n_2-2)^2}.$$

Legătura dintre repartiția normală şi repartiția Snedecor este dată în următoarea teoremă.

Teorema 4.8.3. *Dacă variabilele aleatoare independente* $X_1, X_2, ..., X_{n_1}, X_{n_1+1}, ...$ $X_{n_1+n_2}$ *urmează legea normală cu parametrii 0 şi* σ *atunci variabila aleatoare*

$$X = \frac{n_2}{n_1} \cdot \frac{X_1^2 + \cdots + X_{n_1}^2}{X_{n_1+1}^2 + \cdots + X_{n_1+n_2}^2}$$

are repartiție Snedecor cu parametrii n_1 *şi* n_2.

Definiția 4.8.4. *Variabila aleatoare X urmează* ***legea Fisher*** *(X are repartiție Fisher) cu parametrii* n_1 *şi* n_2, $(n_1, n_2 \in N^*)$ *dacă densitatea sa de probabilitate (repartiție) este funcția*

$$f(x) = 2\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} e^{n_1 x}\left(1+\frac{n_1}{n_2}e^{2x}\right)^{-\frac{n_1+n_2}{2}}, \quad x \in R. \qquad (4.8.3)$$

Teorema 4.8.5. *Dacă variabila aleatoare X are repartiție Snedecor cu parametrii* n_1 *şi* n_2, *atunci variabila aleatoare* $Y = \frac{1}{2}\ln X$ *are o repartiție Fisher cu parametrii* n_1 *şi* n_2.

Demonstrație

Dacă notăm funcția de repartiție *X* cu *F*, atunci funcția de repartiție *G* a variabilei *Y* este

$$G(x) = P(Y \le x) = P\left(\frac{1}{2}\ln X \le x\right) = P(\ln X \le 2x)$$
$$= P(X \le e^{2x}) = F(e^{2x}), \ \forall x \in R.$$

Deci densitatea de probabilitate a variabilei *Y* este

$$g(x) = G'(x) = 2e^{2x}F'(e^{2x}) = 2e^{2x} f(e^{2x})$$
$$= 2e^{n_1 x}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(1+\frac{n_1}{n_2}e^{2x}\right)^{-\frac{n_1+n_2}{2}}, \quad \forall x \in R. \quad q.e.d.$$

4.9. Legea Weibull. Legea exponențială

Definiția 4.9.1. *Variabila aleatoare X urmează* ***legea Weibull*** *(X are repartițieWeibull) cu parametrii* λ *şi* α $(\lambda > 0, \alpha > 0)$ *dacă densitatea sa de probabilitate (repartiție) este funcția*

$$f(x) = \begin{cases} \lambda\alpha x^{\alpha-1} e^{-\lambda x^{\alpha}}, & x > 0, \\ 0, & x \le 0. \end{cases} \tag{4.9.1}$$

Funcția f din (4.9.1) este o densitate de probabilitate, deoarece $f(x) \ge 0, \forall x \in R$ şi $\int_{-\infty}^{\infty} f(x)\,dx = \lambda\alpha \int_0^{\infty} x^{\alpha-1} e^{-\lambda x^{\alpha}}\,dx = 1$. Într-devăr pentru verificarea ultimei relații facem schimbarea de variabilă $\lambda x^{\alpha} = y$, de unde rezultă $x = (y/\lambda)^{1/\alpha}$ şi $dx = \dfrac{1}{\alpha\lambda^{\frac{1}{\alpha}}} y^{\frac{1}{\alpha}-1} dy$. Atunci

$$\int_{-\infty}^{\infty} f(x)\,dx = \lambda\alpha \int_0^{\infty} \left(\frac{y}{\lambda}\right)^{\frac{\alpha-1}{\alpha}} e^{-y} \cdot \frac{1}{\alpha\lambda^{\frac{1}{\alpha}}} y^{\frac{1}{\alpha}-1} dy = \int_0^{\infty} e^{-y} dy = 1.$$

Teorema 4.9.2. *Dacă variabila aleatoare X are repartiție Weibull cu parametrii* λ *şi* α $(\lambda > 0, \alpha > 0)$, *atunci valoarea medie şi dispersia sa sunt*

$$E(X) = \lambda^{-\frac{1}{\alpha}} \Gamma\left(\frac{1}{\alpha} + 1\right), \quad Var(X) = \lambda^{-\frac{2}{\alpha}} \left[\Gamma\left(\frac{2}{\alpha} + 1\right) - \Gamma^2\left(\frac{1}{\alpha} + 1\right)\right]. \tag{4.9.2}$$

Demonstrație

Folosind schimbarea de variabilă de mai sus, obținem pentru media variabilei X

$$E(X) = \lambda\alpha \int_0^{\infty} x^{\alpha} e^{-\lambda x^{\alpha}}\,dx = \lambda\alpha \int_0^{\infty} \frac{y}{\lambda} e^{-y} \cdot \frac{1}{\alpha\lambda^{\frac{1}{\alpha}}} y^{\frac{1}{\alpha}-1} dy$$

$$= \frac{1}{\lambda^{\frac{1}{\alpha}}} \int_0^{\infty} y^{\frac{1}{\alpha}} e^{-y} dy = \lambda^{-\frac{1}{\alpha}} \Gamma\left(\frac{1}{\alpha} + 1\right).$$

Pentru dispersia lui X calculăm mai întâi media lui X^2. Avem

$$E(X^2) = \lambda\alpha \int_0^{\infty} x^{\alpha+1} e^{-\lambda x^{\alpha}}\,dx = \lambda\alpha \int_0^{\infty} \left(\frac{y}{\lambda}\right)^{\frac{\alpha+1}{\alpha}} e^{-y} \cdot \frac{1}{\alpha\lambda^{\frac{1}{\alpha}}} y^{\frac{1}{\alpha}-1} dy = \lambda^{-\frac{2}{\alpha}} \Gamma\left(\frac{2}{\alpha} + 1\right).$$

Atunci dispersia lui *X* este

$$Var(X) = \lambda^{-\frac{2}{\alpha}}\left[\Gamma\left(\frac{2}{\alpha}+1\right) - \Gamma^2\left(\frac{1}{\alpha}+1\right)\right].$$

q.e.d.

Pentru $\alpha = 1$ legea (repartiția) Weibull se mai numește **legea exponențială (repartiția exponențială)** de parametru λ. Deci densitatea de probabilitate a unei variabile aleatoare *X* cu repartiție exponențială cu parametrul λ este

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0, \\ 0, & x \le 0. \end{cases}$$

Din relația (4.9.2) rezultă că media şi dispersia unei variabile aleatoare *X* cu repartiție exponențială cu parametrul λ sunt

$$E(X) = \frac{1}{\lambda}\Gamma(2) = \frac{1}{\lambda}, \quad Var(X) = \frac{1}{\lambda^2}[\Gamma(3) - \Gamma^2(2)] = \frac{1}{\lambda^2},$$

deoarece $\Gamma(2) = 1,\ \Gamma(3) = 2.$

Observația 4.9.3. *Repartiția exponențială cu parametrul λ este o repartiție gamma generalizată cu parametrii $p = 1$ şi λ.*

Propoziția 4.9.4. *Dacă variabila aleatoare X are repartiție exponențială cu parametrul λ, atunci funcția sa caracteristică este*

$$\varphi(t) = \left(1 - \frac{it}{\lambda}\right)^{-1}, \quad t \in R. \tag{4.9.3}$$

Demonstrație

Din Observația 4.9.3 şi relația (4.4.7) din Propoziția 4.4.6, deducem că funcția caracteristică a variabilei *X* este dată de expresia din (4.9.3).

Funcția caracteristică se poate calcula şi direct folosind formula din definiția sa. Avem

$$\varphi(t) = E\left(e^{itX}\right) = \int_{-\infty}^{\infty} e^{itx} f(x)\,dx = \int_0^{\infty} e^{itx} \cdot \lambda e^{-\lambda x}\,dx = \lambda \int_0^{\infty} e^{(it-\lambda)x}\,dx$$

$$= \lambda \int_0^{\infty} e^{-\lambda x}[\cos(tx) + i\sin(tx)]\,dx = \lambda \underbrace{\int_0^{\infty} e^{-\lambda x}\cos(tx)\,dx}_{I_1} + i\lambda \underbrace{\int_0^{\infty} e^{-\lambda x}\sin(tx)\,dx}_{I_2}.$$

Pentru I_1 obținem

$$I_1 = \int_0^\infty e^{-\lambda x}\cos(tx)\,dx = -\frac{1}{\lambda}\int_0^\infty \left(e^{-\lambda x}\right)'\cos(tx)\,dx = -\frac{1}{\lambda}e^{-\lambda x}\cos(tx)\Big|_0^\infty$$
$$-\frac{t}{\lambda}\int_0^\infty e^{-\lambda x}\sin(tx)\,dx = \frac{1}{\lambda}+\frac{t}{\lambda^2}\int_0^\infty \left(e^{-\lambda x}\right)'\sin(tx)\,dx = \frac{1}{\lambda}+\frac{t}{\lambda^2}\left(e^{-\lambda x}\sin(tx)\Big|_0^\infty\right.$$
$$\left.-t\int_0^\infty e^{-\lambda x}\cos(tx)\,dx\right) = \frac{1}{\lambda}-\frac{t^2}{\lambda^2}\int_0^\infty e^{-\lambda x}\cos(tx)\,dx.$$

Rezultă astfel relația $I_1 = \frac{1}{\lambda}-\frac{t^2}{\lambda^2}I_1$, de unde obținem că $I_1 = \frac{\lambda}{\lambda^2+t^2}$. Pentru I_2 din relațiile de mai sus deducem $I_2 = \frac{t}{\lambda^2+t^2}$. Obținem astfel pentru $\varphi(t)$ expresia $\varphi(t) = \left(1-\frac{it}{\lambda}\right)^{-1}$, $\forall t \in R$. q.e.d.

Aplicația 4.9.5. *Fie variabilele independente X_1 şi X_2 cu repartiții exponențiale cu parametrii λ_1, respectiv λ_2. Să se determine densitatea de probabilitate a variabilei $X_1 + X_2$. Să se generalizeze apoi rezultatul obținut la cazul a n variabile aleatoare independente cu repartiții exponențiale cu acelaşi parametru λ.*

Rezolvare

Să notăm cu f_1 şi f_2 densitățile de probabilitate ale variabilelor X_1 şi X_2, adică

$$f_1(x) = \begin{cases} \lambda_1 e^{-\lambda_1 x}, & x > 0, \\ 0, & x \le 0, \end{cases} \qquad f_2(x) = \begin{cases} \lambda_2 e^{-\lambda_2 x}, & x > 0, \\ 0, & x \le 0. \end{cases}$$

Dacă $\lambda_1 \neq \lambda_2$ atunci densitatea de probabilitate f a variabilei $X_1 + X_2$ este 0 pentru $x \le 0$, iar pentru $x > 0$ avem

$$f(x) = \int_{-\infty}^{\infty} f_1(x-y)\,f_2(y)\,dy = \int_0^x \lambda_1 e^{-\lambda_1(x-y)}\cdot\lambda_2 e^{-\lambda_2 y}\,dy$$
$$= \lambda_1\lambda_2 e^{-\lambda_1 x}\int_0^x e^{(\lambda_1-\lambda_2)y}dy = \frac{\lambda_1\lambda_2}{\lambda_1-\lambda_2}e^{-\lambda_1 x}e^{(\lambda_1-\lambda_2)y}\Big|_0^x = \frac{\lambda_1\lambda_2}{\lambda_1-\lambda_2}\left(e^{-\lambda_2 x}-e^{-\lambda_1 x}\right).$$

Deci funcția f are forma $f(x) = \begin{cases} \frac{\lambda_1\lambda_2}{\lambda_1-\lambda_2}\left(e^{-\lambda_2 x}-e^{-\lambda_1 x}\right), & x > 0, \\ 0, & x \le 0. \end{cases}$

Dacă $\lambda_1 = \lambda_2 = \lambda$ atunci $f(x) = 0, \ \forall x \le 0,$ iar pentru $x > 0$ obținem

$$f(x) = \int_0^x \lambda e^{-\lambda(x-y)} \cdot \lambda e^{-\lambda y}\, dy = \lambda^2 e^{-\lambda x} \int_0^x dy = \lambda^2 x e^{-\lambda x}.$$

Deci densitatea de probabilitate a variabilei $X_1 + X_2$ este

$$f(x) = \begin{cases} \lambda^2 x e^{-\lambda x}, & x > 0, \\ 0, & x \le 0. \end{cases}$$

Pentru trei variabile aleatoare independente X_1, X_2, X_3 cu repartiții exponențiale cu parametrul λ, densitatea de probabilitate a variabilei $X_1 + X_2 + X_3$ este $h(x) = 0, \ \forall x \le 0$ și pentru $x > 0$ avem

$$h(x) = \int_0^x \lambda^2 (x-y) e^{-\lambda(x-y)} \cdot \lambda e^{-\lambda y}\, dy = \lambda^3 e^{-\lambda x} \int_0^x (x-y)\, dy = \frac{\lambda^3}{2} x^2 e^{-\lambda x}.$$

Deci

$$h(x) = \begin{cases} \frac{\lambda^3}{2} x^2 e^{-\lambda x}, & x > 0, \\ 0, & x \le 0. \end{cases}$$

Prin inducție matematică se arată că pentru variabilele aleatoare independente $X_1, X_2, ..., X_n$ cu repartiții exponențiale cu parametrul λ, densitatea de probabilitate a variabilei $X = X_1 + X_2 + \cdots + X_n$ este funcția

$$k(x) = \begin{cases} \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x}, & x > 0, \\ 0, & x \le 0. \end{cases}$$

Rezultă că variabila *X* are o repartiție gamma generalizată cu parametrii *n* și λ.

Capitolul 5

Convergența variabilelor aleatoare

5.1.Convergența aproape sigură şi convergența în probabilitate

Fie (Ω,K,P) un spațiu probabilizat şi $(X_n)_n$ un şir de variabile aleatoare. Fie de asemenea X o altă variabilă aleatoare. Ce înseamnă că X_n tinde la X ?

Desigur, variabilele aleatoare sunt funcții $X_n:\Omega \to \Re$. Pentru funcții se ştie deja ce înseamnă că X_n converge la X : că $X_n(\omega) \to X(\omega)$ pentru orice $\omega \in \Omega$. Mai există şi noțiunea de convergență uniformă: $X_n \xrightarrow{u} X$ dacă $\sup_{\omega\in\Omega} |X_n(\omega) - X(\omega)| \longrightarrow 0$.

În teoria probabilităților aceste două tipuri de convergență nu prea sunt de folos. Situația tipică întîlnită în practica statistică este următoarea: se face un experiment cu rezultatele posibile $r_1,\ldots,r_k$. Rezultatul obținut la al *n*-ulea experiment este X_n.

(De exemplu se aruncă un zar despre care nu ştim dacă este corect sau falsificat; în acest caz rezultatele sunt 1,2,3,4,5,6)

Nu avem cum să prezicem rezultatul X_n, dar putem spera să aproximăm probabilitățile p_i de apariție a rezultatului r_i, dacă facem unele ipoteze acceptabile. De exemplu, dacă acceptăm că toate variabilele aleatoare X_n au aceeaşi repartiție, *F*. Am putea număra procentajele f_i de apariție a rezultatului r_i şi spera că ele nu diferă mult de "adevăratele" probabilități p_i.

De altfel, de aici a şi pornit teoria probabilităților.

În cazul zarului nostru, am putea să îl aruncăm de 6*n* de ori – de exemplu – şi să vedem dacă frecvențele obținute diferă mult de *n*. Dacă da, (deocamdată nu ştim ce înseamnă "prea mult", se va vedea la capitolul privind intervalele de încredere) am putea spera că dacă o să mărim numărul de aruncări ale zarului ne vom apropia din ce în ce mai mult de "adevăratele valori" $p_i = P(X = r_i)$.

Să formalizăm puțin contextual.

Avem un şir de variabile aleatoare $(X_n)_n$ care sunt identic repartizate cu repartiția necunoscută $F = \begin{pmatrix} r_1 & r_2 & \ldots & r_k \\ p_1 & p_2 & \ldots & p_k \end{pmatrix}$. Vrem să găsim numerele

$p_1,...,p_k$. Pentru aceasta calculăm valorile $Y_{n,i} = \frac{\left|\left\{j \le n : X_j = r_i\right\}\right|}{n}$ şi ne gîndim că, poate, $Y_{n,i}$ converge la p_i dacă $n \to \infty$.

Aceasta este problema de bază.

Nu ne putem aştepta ca $Y_{n,i}$ să conveargă la p_i în sensul obişnuit al cuvîntului, adică să aibă loc convergenţa pentru orice *scenariu* $\omega \in \Omega$: de exemplu, s-ar putea ca rezultatul r_i să apară mereu sau nici măcar să nu apară, deşi $p_i > 0$.

Dar poate că $Y_{n,i}$ converge la p_i în majoritatea covîrşitoare a cazurilor?

Într-adevăr, aşa se întîmplă dacă se mai adaugă nişte ipoteze. **Aceasta este legea numerelor mari.**

Mai întîi să lămurim ce înseamnă "***în majoritatea covîrşitoare a cazurilor***": înseamnă **aproape sigur.**

Definiţia 5.1.1. *Spunem că* $(X_n)_n$ *converge la* X ***P-aproape sigur*** *şi scriem* $X_n \xrightarrow{P-a.s.} X$ *(sau* $X_n \to X$ *(mod P)) dacă* $P(\limsup X_n = \liminf X_n = X) = 1$. *Dacă* $P(\limsup X_n = \liminf X_n) = 1$, *spunem că* $(X_n)_n$ *este un şir P- convergent aproape sigur.*

Observaţia 5.1.2. *Dacă probabilitatea P se subînţelege, putem renunţa la litera "P" şi scriem doar* $X_n \xrightarrow{a.s.} X$

Observaţia 5.1.3. *Evident că dacă* X_n *converge la* X *punctual - în sensul obişnuit al cuvîntului - atunci* $X_n \xrightarrow{a.s.} X$. *La fel de evident este că diferenţa între convergenţa punctuală şi cea aproape sigură poate fi foarte mare.*

Observaţia 5.1.4. *Tot evident este că limita aproape sigură a unui şir de variabile aleatoare nu este unică, ci numai* ***unică aproape sigur****. Într-adevăr, dacă* $X_n \xrightarrow{a.s.} X$ *şi* $X = Y$ *(a.s.), atunci putem la fel de bine să spunem că* $X_n \xrightarrow{a.s.} Y$.

Observaţia 5.1.5. *La fel de evident este că* $X_n \xrightarrow{a.s.} X \Leftrightarrow X_n - X \xrightarrow{a.s} 0$, *deoarece*

$P(\limsup X_n = \liminf X_n = X) = P(\limsup(X_n - X) = \liminf(X_n - X) = 0)$.

Exemplu 5.1.6. *Fie* $\Omega = \Re$, $K = B(\Re)$, $X_n(\omega) = \sin(n\omega)$, $P = (\delta_0 + \delta_\pi) = \begin{pmatrix} 0 & \pi \\ .5 & .5 \end{pmatrix}$. *Şirul* X_n *este divergent, singurele puncte* ω *în care el*

converge sunt cele de forma $\omega = k\pi$. *Totuşi,* $X_n \xrightarrow{P-a.s.} 0$, *deoarece* $P(X_n = 0) = 1$. *Sau, dacă schimbăm puţin şi luăm* $X_n(\omega) = \sin^n(\omega)$, *acesta converge la* 0 *cu excepţia punctelor de forma* $\pi/2 + k\pi$. *Dacă luăm, de data aceasta,* $P = (\delta_{\pi/2} + \delta_{5\pi/2}) = \begin{pmatrix} \pi/2 & 5\pi/2 \\ .5 & .5 \end{pmatrix}$, *atunci* $X_n \xrightarrow{P-a.s.} 1$.

Exemplul 5.1.7. *Fie* $\Omega = [0,1)$, $K = B([0,1))$, P = Uniform(0,1) (= măsura Lebesgue pe Ω!) *şi* $(a_n)_{n\geq 1}$ *un şir strict crescător de numere pozitive cu proprietatea că* $\lim a_n = \infty$ *dar* $\lim(a_{n+1} - a_n) = 0$ *şi, în plus,* $a_{n+1} - a_n \leq 1$. *De exemplu putem lua* $a_n = \ln n$ *sau* $a_n = \sqrt{n}$.

Fie $d(x) = x - [x]$ *partea* zecimală a lui x, $A_n = d([a_n, a_{n+1})) := \{d(x) : x \in [a_n, a_{n+1})\}$ şi $X_n = 1_{A_n}$. *Observăm că* $P(A_n) = a_{n+1} - a_n \to 0$. *(De exemplu, dacă* $a_n = \ln n$, *atunci* $A_1 = [0, \ln 2], A_2 = [\ln 2, 1) \cup [0, \ln 3 - 1), A_3 = [\ln 3 - 1, \ln 4 - 1)$, *etc.).* $A_4 = [\ln 4 - 1, 1) \cup [0, \ln 5 - 2)$. *Atunci limsup* $A_n = \Omega$ *şi liminf* $A_n = \varnothing$ *deci* $\limsup X_n = 1_{\limsup A_n} = 1$ *şi* $\liminf X_n = 1_\varnothing = 0$.

(Într-adevăr,

$$\operatorname{limsup} A_n = \bigcap_n \bigcup_k d([a_{n+k}, a_{n+k+1})) = \bigcap_n d\left(\bigcup_k [a_{n+k}, a_{n+k+1})\right) = \bigcap_n d([a_n, \infty)) = \Omega \qquad \textit{iar}$$

$$\liminf A_n = \bigcup_n \bigcap_k d([a_{n+k}, a_{n+k+1})) = \varnothing)$$

Exemplul 5.1.7 este unul extrem, în care limita superioară şi cea inferioară nu coincid nicăieri.

Exemplul 5.1.8. *Fie, pe spaţiu probabilizat de la exemplul anterior,* $X_n = n1_{\left(0, \frac{1}{n}\right)}$. *Atunci* X_n *converge evident la 0, deci converge şi a.s.*

Cum putem decide dacă $X_n \to X$? În cazul cel mai simplu, un răspuns este dat de

Propoziţia 5.1.9. *Fie* (Ω, K, P) *un spaţiu probabilizat şi* $(A_n)_n$ *un şir de evenimente din K. Presupunem că seria* $\sum P(A_n)$ *este convergentă. Atunci*

$1_{A_n} \xrightarrow{a.s.} 0$.

Demonstraţie

Fie $B_n = A_n \cup A_{n+1} \cup \ldots$. Cum P(B_n) $\le \sum_{k\ge 0} P(A_{n+k})$ şi seria $\sum P(A_n)$ este convergentă, $P(B_n) \to 0$. Dar şirul de mulţimi $(B_n)_n$ este monoton descrescător, deci $1_{B_n} \to 1_B$ cu B = $\bigcap_{n=1}^{\infty} B_n$ şi $P(B) = \lim P(B_n) = 0$.

Dar $\limsup 1_{A_n} = 1_{\limsup A_n} = 1_{\bigcap_n \bigcup_k A_{n+k}} = 1_{\bigcap B_n} = 1_B$ de unde $0 \le \lim\inf 1_{A_n} \le \limsup 1_{A_n} = 1_B$. Cum $P(\limsup 1_{A_n} \ne \liminf 1_{A_n}) \le P(B) = 0$, urmează că şirul $(1_{A_n})_n$ converge aproape sigur. Putem lua ca limită a sa, X, orice variabilă aleatoare $X = 1_A$ unde A este o mulţime neglijabilă. Cel mai simplu e să spunem că $1_{A_n} \to 0$. *.q.e.d*

De aici rezultă un criteriu important cu care putem verifica dacă $X_n \xrightarrow{a.s.} X$

Propoziţia 5.1.10. *Fie* $(X_n)_n$ *un şir de variabile aleatoare. Presupunem că pentru orice* $\varepsilon > 0$ *seria* $\sum P(|X_n| > \varepsilon)$ *este convergentă. Atunci* $X_n \xrightarrow{a.s.} 0$.

Demonstraţie

Fie $B = \{\omega \in \Omega : X_n(\omega)$ nu converge la $0\}$. Atunci B = $\bigcup_N B_N$ unde $B_N = \{\omega \in \Omega : |X_n(\omega)| > 1/N$ de o infinitate de ori$\}$. Cum şirul de mulţimi $(B_N)_N$ este crescător, P(B) = lim P(B_N) – aplicăm din nou proprietatea de continuitate monotonă a probabilităţii. Dar seria $\sum P(|X_n| > 1/N)$ este convergentă, deci $P(B_N) = 0 \ \forall$ N de unde $P(B) = 0$ *q.e.d.*

Observaţia 5.1.11. *Condiţia din propoziţia de mai sus nu este decît suficientă, nu şi necesară. De exemlu dacă avem un şir descrescător de mulţimi* $A_1 \supseteq A_2 \supseteq \ldots\ldots$ *cu* $P(A_n) = 1/n$, *atunci* 1_{A_n} *este un şir descrescător, deci are o limită de forma* 1_A *cu* $A = \bigcap_n A_n$. Deci $1_{A_n} \to 1_A$. *Cum* $P(A) = 0$, *putem scrie, conform observaţiei 3, că* $1_{A_n} \xrightarrow{a.s.} 0$.

Observaţia 5.1.12. *Convergenţele obişnuite sunt date de o distanţă. Adică putem scrie că* $X_n \to X \Leftrightarrow \forall\, \varepsilon > 0$ *există* n_ε *astfel ca* $n \ge n_\varepsilon \Rightarrow d(X_n, X) \le \varepsilon$..
Aceste convergenţe au următoarea proprietate: dacă din orice subşir al lui (X_n) *se poate extrage un sub-subşir care este convergent şi limita sa este aceeaşi, să zicem X, atunci* $X_n \to X$. *Înr-adevăr, dacă* $\lim X_n \ne X$, *atunci*

există un $\varepsilon > 0$ şi un subşir $(k_n)_n$ astfel ca $d(X_{k_n}, X) > \varepsilon$, deci din subşirul (X_{k_n})n nu putem extrage nici un sub-subşir care să tindă la ε. Spunem că aceste convergenţe ***sunt topologice****. Convergenţa aproape sigură* ***nu este topologică.*** *Într-adevăr, dacă ne uităm la exemplul extrem 1.1.7, în care şirul nu are absolut nici o limită aproape sigură, vedem că el are proprietatea ciudată că din orice subşir al său se poate extrage un sub-subşir care converge la 0. Motivul este că $P(A_n) \to 0$. Altfel zis, şirul 1_{A_n} converge totuşi la 0, dar în probabilitate.*

Definiţia 5.1.13. *Spunem că şirul de variabile aleatoare (X_n) converge în probabilitate la X (şi scriem $X_n \xrightarrow{P} X$) dacă* $\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0 \forall \varepsilon > 0.$

Din definiţie rezultă imediat că un şir de forma $X_n = 1_{A_n}$ converge în probabilitate la 0 dacă şi numai dacă $P(A_n) \to 0$. Acesta este cazul exemplului 1.1.7, care ne arată cum se prea poate ca un şir divergent aproape sigur să conveargă, totuşi, în probabilitate.

Convergenţa în probabilitate este una mult mai uşor de manipulat, deoarece ea este topologică: provine de la distanţă.

Propoziţia 5.1.14. *Fie X,Y două variabile aleatoare. Definim* $d(X,Y) = E(|X - Y| \wedge 1)$. *Atunci funcţia d este o semidistanţă pe spaţiul L(Ω,K) al tuturor variabilelor aleatoare şi, în plus, avem echivalenţa* $X_n \xrightarrow{P.} X \Leftrightarrow d(X,X_n) \to 0$

Demonstraţie

Cum este evident că $a,b \geq 0 \Rightarrow a\wedge 1 + b\wedge 1 \geq (a+b)\wedge 1$, este clar că d este o semidistanţă: $d(X,Y) + d(Y,Z) = E[(|X - Y| \wedge 1) + (|Y - Z| \wedge 1)] \geq E(|X - Z| \wedge 1)$.

Pe de altă parte, dacă notăm cu Z variabila aleatoare $|X - Y|$, observăm că dacă $\varepsilon \in (0,1)$, putem scrie $E(Z \wedge 1) = E(Z \wedge 1; Z \leq \varepsilon) + E(Z \wedge 1; Z > \varepsilon) \leq \varepsilon P(Z \leq \varepsilon) + P(Z > \varepsilon)$ de unde

$$E(Z \wedge 1) \leq \varepsilon + P(Z > \varepsilon) \qquad (5.1.1)$$

Apoi $E(Z \wedge 1; Z \leq \varepsilon) + E(Z \wedge 1; Z > \varepsilon) \geq E(Z \wedge 1; Z > \varepsilon) \geq E(\varepsilon; Z > \varepsilon)$ de unde

$$E(Z \wedge 1) \geq \varepsilon P(Z > \varepsilon) \qquad (5.1.2)$$

Din (5.1.1) şi (5.1.2) deducem cleştele

$$\varepsilon P(Z > \varepsilon) \leq E(Z \wedge 1) \leq \varepsilon + P(Z > \varepsilon) \qquad (5.1.3)$$

Să presupunem acum că $d(X,X_n) \to 0$. Din prima inegalitate de la (5.1.3) deducem inegalitatea $P(|X - X_n| > \varepsilon) \leq \frac{E(|X - X_n| \wedge 1)}{\varepsilon}$, adică $P(|X - X_n| > \varepsilon) \leq \frac{d(X, X_n)}{\varepsilon} \Rightarrow \lim_n P(|X - X_n| > \varepsilon) = 0$; aşadar $d(X,X_n) \to 0 \Rightarrow X_n \xrightarrow{P} X$

Reciproc, dacă pentru orice $\varepsilon > 0$ avem că $\lim_n P(|X - X_n| > \varepsilon) = 0$, din a doua inegalitate de la (5.1.3) rezultă că $\lim\sup_{n\to\infty} d(X,X_n) \leq \varepsilon$ pentru orice $\varepsilon > 0$. Dar ε este arbitrar, deci $\lim d(X,X_n) = 0$. *q.e.d.*

Un caz care implică imediat convergența în probabilitate este convergența în L^p..

Definiția 5.1.15. *Spunem că şirul de variabile aleatoare* $(X_n)_n$ *converge în* L^p *la* X *(şi notăm acest lucru cu* $X_n \xrightarrow{L^p} X$*) dacă* $\lim_n \|X_n - X\|_p = 0$.[1]

Propoziția 5.1.16. *Dacă* $X_n \xrightarrow{L^p} X$ *pentru un anumit* $p \geq 1$, *atunci* $X_n \xrightarrow{P} X$.

Demonstrație

Pentru $p \in [1,\infty)$ folosim inegalitatea evidentă $P(|Z| > \varepsilon) \leq \frac{\|Z\|_p^p}{\varepsilon^p}$[2] : într-adevăr, dacă $X_n \xrightarrow{L^p} X$, atunci $\lim_n P(|X - X_n| > \varepsilon) \leq \lim_n \frac{\|X - X_n\|_p^p}{\varepsilon^p} = 0$. Iar dacă $p = \infty$, atunci este şi mai evident: $X_n \xrightarrow{L^\infty} X \Rightarrow X_n \xrightarrow{L^p} X \Rightarrow X_n \xrightarrow{P} X$. *q.e.d.*

Observația 5.1.17. *Totuşi, dacă* $X_n \xrightarrow{P} X$, *din orice subşir al său se poate extrage un sub-subşir care converge la* X *aproape sigur. Într-adevăr, aplicăm* Propoziția 5.1.10 *Prin procedeul diagonal, putem alege un subşir* $(k_n)_n$ *în aşa fel încît seria* $\sum P(|X_{k_n}| > \varepsilon)$ *să fie convergentă pentru orice* $\varepsilon > 0$.

Concluzie. Între cele trei tipuri de convergență există următoarele implicatii:

$X_n \xrightarrow{L^\infty} X \Rightarrow X_n \xrightarrow{L^p} X$ ($p < \infty$) $\Rightarrow X_n \xrightarrow{P} X$ şi $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{P} X \Rightarrow X_{k_n} \xrightarrow{a.s.} X$

Aşadar, convergența în probabilitate este cea mai slabă. Este de remarcat că nu există alte implicații. Astfel:

-Şirul de la exemplul 5.1.3 converge peste tot, dar nu în L^1, deci nici în L^p cu $p \geq 1$

-Şirul de la exemplul 5.1.2 converge în probabilitate şi în L^p pentru orice $1 \leq p < \infty$, dar nu a.s.

-Dacă luăm $X_n = \alpha_n 1_{\left(0,\frac{1}{n}\right)}$ pe $\Omega = (0,1)$, P = Uniform(0,1), care converge punctual la 0, atunci pentru orice p>1 putem găsi un şir $(\alpha_n)_n$ astfel ca X_n

---

[1] Amintim că dacă X este o variabilă aleatoare, şi $p \in [1,\infty]$ atunci $\|X\|_p := \left(E|X|^p\right)^{\frac{1}{p}}$ dacă $p < \infty$ iar $\|X\|_\infty := \lim_{p\to\infty} \|X\|_p = \text{ess sup}\,|X|$. Inegalitatea normelor ne spune că funcția $p \mapsto \|X\|_p$ este crescătoare.

[2] $\|Z\|_p^p = E|Z|^p > E\left(|Z|^p; |Z| > \varepsilon\right) > \varepsilon^p P(|Z| > \varepsilon)$

$\xrightarrow{L^{p'}}$0 pentru $p' < p$ *dar* X_n nu converge nicăieri dacă $p' \geq p$. De exemplu $\alpha_n = n^{-\frac{1}{p}}$.

5.2. Legi ale numerelor mari şi aplicaţii

5.2.1 Legea slabă

Revenim la problema de bază enunţată în paragraful precedent: putem spera să găsim repartiţia unei variabile aleatoare „empiric", adică făcînd observaţii asupra ei?

Aşa pusă fiind, problema nu are sens.

Pentru noi niciodată nu putem face mai multe observaţii asupra unei variabile aleatoare, ci numai una. Dacă aruncăm un zar de n ori, noi nu observăm o variabilă aleatoare, ci un şir de variabile aleatoare $X_1,...,X_n$. Putem accepta că aceste variabile aleatoare au aceeaşi repartiţie.

Atunci problema capătă sens: avem un şir de variabile aleatoare $(X_n)_n$ care sunt identic repartizate şi am dori să îi aproximăm repartiţia, pe baza observaţiilor făcute asupra lui.

Nu cumva am putea să îi calculăm, aproximativ, media? În definitiv $\mu = EX_j$ nu se numeşte degeaba „medie"!

Răspunsul este: uneori, da.

Propoziţia 5.2.1. ***Legea slabă a numerelor mari.****(WLLN)*[3]
Fie $(X_n)_n$ *un şir de variabile aleatoare identic repartizate şi necorelate din* L^2 (*adică avînd şi moment de* **ordin** 2). *Fie* $\mu = EX_j$, $\sigma^2 = \mathrm{Var}(X_j)$ *şi fie* $\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}$ *media empirică*[4]. *Atunci* $\overline{X}_n \xrightarrow{P} \mu$.

Ca un caz particular, dacă $X_n \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$, *atunci* $\overline{X}_n = \frac{\left|\left\{j \leq n : X_j = 1\right\}\right|}{n}$ $\xrightarrow{P} p$. *Acum mediile* $\overline{X}_n$ *se notează cu* f_n *şi se numesc* ***frecvenţe relative.***

Demonstraţie

Centrăm variabilele: fie $Y_n = X_n - \mu$. Atunci $\overline{X}_n - \mu = \frac{Y_1 + Y_2 + ... + Y_n}{n}$.

Urmează $\| \overline{X}_n - \mu \|_2^2 = E(\overline{X}_n - \mu)^2 = \frac{1}{n^2} \sum_{1 \leq i, j \leq n} EY_iY_j$. Dar $EY_iY_j = \mathrm{cov}(X_i,X_j) = 0$

[3] abreviere internaţională: Weak Law of Large Numbers
[4] Se mai numeşte şi "media de selecţie"

dacă $i \neq j$. Deci $E(\overline{X}_n - \mu)^2 = \frac{1}{n^2} \sum_{1 \leq i \leq n} EY_i^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$. Concluzie: $\overline{X}_n \xrightarrow{L^2} \mu$ deci, conform cu Propoziția 1.1.16, $\overline{X}_n \xrightarrow{P} \mu$. *q.e.d.*

Rezultatul nu mai rămîne adevărat dacă renunțăm la ipoteza necorelării. De exemplu, dacă X şi Y sunt două variabile aleatoare cu aceeaşi repartiție şi pe baza lor construim şirul $X_n = X$ dacă n este par şi X_n = Y dacă este impar, atunci $\overline{X}_n$ va converge peste tot la $\frac{X+Y}{2}$, care nu numai că nu coincide cu media, dar mai este şi variabilă aleatoare.

Se pot da exemple de şiruri de variabile aleatoare identic repartizate pentru care media empirică nu converge nicăieri. De exemplu se iau doua variabile aleatoare X şi Y si cu ajutorul lor se construieste un sir format doar din X şi Y care nu are limită Cesaro.

Observația 5.2.2. *Legea slabă se mai numeşte* ***Teorema lui Bernoulli.*** *Ea a dat primul răspuns la întrebarea: putem aproxima empiric probabilitățile? Forma sa verbală repetată de sute de ani este:* ***Frecvențele relative converg în probabilitate la adevărata probabilitate.***

Acest rezultat nu este satisfăcător. De aceea teorema lui Bernoulli se numeşte „Legea slabă". Faptul că $\overline{X}_n$ converge în probabilitate la μ nu înseamnă de loc că şirul $\overline{X}_n(\omega)$ converge la μ pentru toate scenariile ω, nici măcar că aşa ceva se întîmplă pentru „marea lor majoritate". E suficient să privim exemplul 5.1.7. E adevărat că el conține un subşir care converge la μ (observația 5.1.17) dar asta ne ajută prea puțin.

Am dori un rezultat „tare" care să ne asigure că $\overline{X}_n$ converge la μ aproape sigur. Dacă eliminăm ipoteza ca natura ne joacă o farsă urîtă, atunci putem fi siguri că dacă tot repetăm un experiment în anumite condiții, ne vom apropia oricît de media cea "adevărată".[5]

Chiar aşa se şi întîmplă.

5.2.2 Legea tare

Prețul este să înlocuim ipoteza ca variabilele X_n sunt necorelate cu ipoteza mult mai tare că ele sunt independente.

[5] Aici e de fapt o problemă de filosofie a satisticii: decretăm că evenimentele de probabilitate 0 nu se întîmplă. E o discuție fascinantă, dar mai complicată.

Fie atunci $X:\Omega \to \Re^\infty$ vectorul cu componentele $X = (X_1,X_2,....)$. Componentele sale X_j sunt proiecțiile $\mathrm{pr_j}(X)$. În loc să scriem $\overline{X_n}$, scriem $\frac{1}{n}(pr_1(X)+...+pr_n(X))$. Avantajul este că acum avem o singură variabilă, anume X. Introducem şi funcția t: $\Re^\infty \to \Re^\infty$ definită prin

$$t(x_1,x_2,.....) = (x_2,x_3,....) \Leftrightarrow pr_j(t(x)) = pr_{j+1}(x)\ \forall\ j \geq 1 \qquad (5.2.1)$$

care se numeşte **shiftul canonic**. Dacă notăm cu $f:\Re^\infty \to \Re$ prima proiecție ($f = pr_1$) atunci putem scrie

$$\overline{X}_n = \frac{1}{n}\left(f + f\circ t + f\circ t^2 +...+ f\circ t^{n-1}\right)(X) \qquad (5.2.2)$$

unde prin t^k înțelegem $t\circ t\circ...\circ t$, compunerea de n ori a lui t cu ea însăşi.

Scrierea are avantajul că ne permite să ne concentrăm asupra unui şir Cesaro în care apar doar două variabile: funcția f şi shiftul t. Cu studiul unor asemenea şiruri se ocupă o disciplină matematică numită **teorie ergodică.**

Ca să putem aplica rezultatele din teoria ergodică, ar trebui verificat că shiftul invariază măsura.

Este vorba despre repartiția $P_1 = P\circ X^{-1}$ a vectorului X. Ea este o probabilitate pe spațiul produs $(\Re^\infty,B^\infty(\Re))$ [6] În general nu ştim să o calculăm, dar dacă facem ipoteza suplimentară că variabilele X_n sunt independente şi identic repartizate, atunci este uşor de văzut că $P_1 = F^\infty$, unde F este repartiția comună a variabilelor aleatoare X_n.[7]

Lema 5.2.3. *Pe spațiul* $(\Re^\infty,B^\infty(\Re))$ *avem că*

(i) $P_1\circ t^{-1} = P_1$

(ii) *Dacă* $A \in B^\infty(\Re)$ *are proprietatea că* $t^{-1}(A) = A$, *atunci* $P_1(A) \in \{0,1\}$

Demonstrație

(i) Trebuie arătat că $P_1(t^{-1}(A)) = P_1(A)\ \forall\ A \in B^\infty(\Re)$. Din motive elementare[8] este suficient să verificăm acest lucru pentru un bloc de lungime n, $A = B_1\times B_2\times...\times B_n\times\Re\times\Re\times...$

---

[6] σ-algebra produs se definește ca fiind cea mai mică σ-algebră generată de blocuri. Un ***bloc de lungime n*** este o mulțime de forma $B_1\times B_2\times...\times B_n\times\Re\times\Re\times.....$ unde mulțimile B_n sunt boreliene. Mulțimea D a blocurilor este în mod evident stabilă la intersecții finite, deci $B^\infty(\Re)$ coincide cu U-sistemul generat de D.

[7] Amintim că dacă Π_j sunt o probabilitate pe un spațiu măsurabil $(E,\mathcal{E})$, atunci $P = \Pi_1\otimes\Pi_2\otimes....$ este probabilitatea pe E^∞ cu proprietatea că $P(B_1\times B_2\times...\times B_n\times E\times E\times.....) = \Pi_1(B_1)\Pi_2(B_2)...\Pi_n(B_n)$. Că o asemenea probabilitate (numită probabilitate produs) există, nu este evident: existența ei este dată de teorema lui Kolmogorov. Dacă $\Pi_1 = \Pi_2 = ...= \Pi$, atunci probabilitatea produs se notează cu Π^∞. Acesta este cazul nostru.

[8] Motivul elementar este că dacă două probabilități definite pe o aceeaşi σ-algebră coincid pe un sistem de generatori închis la intersecții finite al σ-algebrei , atunci ele coincid. În cazul de față acest sistem de generatori este mulțimea D a blocurilor.

Să remarcăm că $t^{-1}(A) = \Re \times A$. Într-adevăr, $x \in t^{-1}(A) \Leftrightarrow t(x) \in A \Leftrightarrow (x_2,x_3,.....) \in B_1 \times B_2 \times ... \times B_n \times \Re \times \Re \times .. \Leftrightarrow x_2 \in B_1,\ x_3 \in B_2,...,x_{n+1} \in B_n \Leftrightarrow x \in \Re \times B_1 \times B_2 \times ... \times B_n \times \Re \times \Re \times .. = \Re \times A$. Deci $P_1(t^{-1}(A)) = F^{\infty}(\Re \times B_1 \times B_2 \times ... \times B_n \times \Re \times \Re \times ..) = F(\Re)F(B_1)...F(B_n) = F(B_1)...F(B_n) = P_1(A)$. Adică t invariază probabilitatea P_1.

(ii) Lucrăm cu indicatori, căci este mai comod. Ipoteza $t^{-1}(A) = A$ devine $1_A \circ t = 1_A$. Dacă punem f în loc de 1_A se pune întrebarea ce putem spune despre o funcție f care are proprietatea că $f = f \circ t$. Aplicăm această funcție vectorului X şi avem că $f(X) = f(t(X)) = f(t(t(X))) = ...$ sau, scris explicit, că $f(X_1,X_2,...) = f(X_2,X_3,...) = f(X_3,X_4,...) = ...$

Aici intervine în forță ipoteza independenței.

Deci variabila aleatoare $Y = f(X)$ este independentă de X_1 (de vreme ce $Y = f(X_2,X_3,...)$!) şi de X_2 (de vreme ce $Y = f(X_3,X_4,...)$!), şi de X_3, adică de toate variabilele aleatoare X_n. Deci $f(X)$ este independentă de X, adică şi de $f(X)$. Înseamnă că Y este independentă de ea însăşi. Dar atunci ea este constantă aproape sigur.

Concluzie: dacă $1_A \circ t = 1_A$, atunci 1_A = constant (mod P_1). Această constantă, firește, nu poate fi decît 0 sau 1: deci $P(A) \in \{0,1\}$. *q.e.d.*

Definiția 5.2.4. *Fie* (Ω,K,P) *un spațiu probabilizat. Şi fie* $t : \Omega \to \Omega$ *măsurabilă. Spunem că t este ergodică față de P dacă* $P_1 \circ t^{-1} = P_1$ *şi* $t^{-1}(A) = A \Rightarrow P(A) \in \{0,1\}$

Acum suntem în contextul firesc al teoriei ergodice. Putem aplica:

Teorema 5.2.5.(Teorema ergodică) *Fie* $(\Omega,K.P)$ *un spațiu probabilizat şi* $f \in L^1(\Omega,K,P)$. *Fie, de asemenea,* $t:\Omega \to \Omega$ *o funcție ergodică. Atunci*

$$\frac{1}{n}\left(f + f \circ t + f \circ t^2 + ... + f \circ t^{n-1}\right) \xrightarrow{a.s.} \mathrm{E}f = \int f dP$$

Demonstrația este departe de a fi evidentă şi nu o vom da aici.[9]

Şi avem ceva mult mai tare decît am sperat, anume

Teorema 5.2.6. *Fie* (Ω,K,P) *un spațiu probabilizat,* (E,E) *un spațiu măsurabil,* $X = (X_n)_n$ *un şir de variabile aleatoare* $X_n: \Omega \to E$ *cu proprietatea că shiftul* $t: E^{\infty} \to E^{\infty}$ *este ergodic față de repartiția sa* $P_1 = P \circ X^{-1}$. *Fie* $f: E^{\infty} \to \Re$ *o funcție măsurabilă cu proprietatea că* $f(X) \in L^1(\Omega,K,P)$. *Atunci* $\frac{1}{n}\left(f + f \circ t + f \circ t^2 + ... + f \circ t^{n-1}\right)(X)$ *converge P – aproape sigur la Ef(X).*

[9] Cititorul interesat poate găsi multe despre această teoremă aici: http://en.wikipedia.org/wiki/Ergodic_theory. O demonstrație frumoasă se poate găsi in cursul lui I. Cuculescu (1974) sau Tudor. Demonstrația originală a autorului (Birkhoff 1931) este aici: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1076138/?page=1

Demonstrație

Nu avem decît să aplicăm teorema ergodică pe spațiul $(E^{\infty},\mathcal{E}^{\infty},P\circ X^{-1})$. *q.e.d.*

Avantajul acestei formulări a legii tari a numerelor mari este că se poate generaliza şi la alte tipuri de şiruri de variabile aleatoare – de exemplu la lanțuri Markov sau la proces staționare.

Ca să o înțelegem mai bine, o să îi scriem cîteva particularizări.

Corolarul 5.2.7. *Fie $(X_n)_n$ un şir de variabile aleatoare i.i.d. cu valori într-un spațiu măsurabil (E,E) şi fie f: $E^{\infty} \to \mathfrak{R}$ o funcție măsurabilă cu proprietatea că $f(X_1,X_2,....) \in L^1$. Atunci*

$$\frac{1}{n}\left(f(X_1,X_2,..)+f(X_2,X_3,..)+...+f(X_n,X_{n+1},..)\right)\xrightarrow{P-a.s.}Ef(X_1,X_2,..) \quad (5.2.3)$$

Şi acest enunț este foarte general. Ca să îl putem aplica, ar trebui să putem calcula membrul drept. Particularizăm la cazuri calculabile. De exemplu dacă variabilele aleatoare sunt reale iar f depinde doar de o mulțime finită de componente:

Corolarul 5.2.8. *Fie $(X_n)_n$ un şir de variabile aleatoare i.i.d. şi fie f:$\mathfrak{R}^k \to \mathfrak{R}$ o funcție măsurabilă cu proprietatea că $f(X_1,X_2,...,X_k) \in L^1$. Atunci*

$$\frac{1}{n}\left(f(X_1,X_2,..,X_k)+f(X_2,X_3,..,X_{k+1})+...+f(X_n,X_{n+1},..,X_{n+k-1})\right)\xrightarrow{P-a.s.}Ef(X_1,X_2,...,X_k) \quad (5.2.4)$$

Avantajul este că acum chiar putem calcula membrul drept, folosind formula de transport.

Dacă $P\circ X_n^{-1} = F$, atunci $Ef(X_1,...,X_k) = \int f dF^k$.

Dacă, de exemplu, F are o densitate ρ față de măsura Lebesgue, atunci $Ef(X_1,...,X_k) = \int f(x_1,x_2,...,x_k)\rho(x_1)\rho(x_2)...\rho(x_k)dx_1dx_2...dx_k$.

O particularizare şi mai mare este cea cu care am început: dacă $f = pr_1$.

Corolarul 5.2.9. Legea tare a numerelor mari (SLLN)[10]

Fie $(X_n)_n$ un şir de variabile aleatoare i.i.d. din L^1. Fie F repartiția lor şi $\mu = EX_1 = \int xdF(x)$. *Atunci* $\overline{X}_n = \frac{X_1+X_2+...+X_n}{n}$ *converge a.s. la* μ..

Exemplul 5.2.10. *Fie X_n variabile aleatoare pozitive i.i.d. şi repartiția F. Atunci media geometrică* $\sqrt[n]{X_1X_2...X_n}$ *converge a.s. la* $e^{E\ln X_1} = e^{\int \ln x dF(x)}$. (*Într-adevăr,*

[10] abreviere internațională: **S**trong **L**aw of **L**arge **N**umbers

logaritmînd expresia avem $\frac{1}{n}(\ln X_1 + \ln X_2 + ... + \ln X_n) \to E\ln X_1$). *Dacă* $P(X_n = 0) > 0$, *atunci limita este 0.*

Exemplul 5.2.11. *Media lor armonică converge la* $1/E\frac{1}{X_1}$. *Într-adevăr,*

$$\frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + ... + \frac{1}{X_n}} \to \frac{1}{E\frac{1}{X_1}}.$$

Exemplul 5.2.12. ***(Procese de reînnoire)***. *Fie* $(\sigma_n)_n$ *un şir de variabile aleatoare i.i.d. strict pozitive a.s. şi* $T_0 = 0$ *iar* $n \geq 1 \Rightarrow T_n = \sigma_1 + ... + \sigma_n$. *Fie* $N(t) = max\{k : T_k < t\}$. $N(t)$ *este un contorcare numără cîte variabile* σ_j *au apărut pînă la momentul t. Atunci* $\frac{N(t)}{t} \xrightarrow{a.s.} \frac{1}{E\sigma_1}$. *Într-adevăr,* $N(t) = n \Leftrightarrow T_n \leq t < T_{n+1}$. *Dacă* $t \to \infty$, *atunci* $n \to \infty$ *şi avem cleştele* $\frac{T_n}{n} \leq \frac{t}{N(t)} < \frac{T_{n+1}}{n}$ *în care termenii din stînga şi din dreapta au aceeaşi limită, anume* $E\sigma_1$.

Posibilitatea statisticii: teorema lui Glivenko

Revenim la problema inițială: putem spera să aproximăm, empiric, funcția de repartiție a unei variabile aleatoare X?

Dacă dispunem de un şir de observații independente asupra ei, răspunsul este, **da**.

Cu condiția să punem problema corect.

"Observații independente asupra lui X" înseamnă de fapt un şir de variabile aleatoare $(X_n)_n$ care sunt **independente, identic repartizate şi avînd aceeaşi repartiție ca *X***.

Exemplu 5.2.13. *Se dă un zar, posibil falsificat şi dorim să estimăm probabilitățile* $p_i = P(X = i)$, $1 \leq i \leq 6$. *X este rezultatul unei aruncări a zarului.*

Exemplu 5.2.14. *Se aruncă la întîmplare două puncte A, B într-un pătrat de latură L =* 1. *Segmentul AB are o lungime aleatoare* $X \in [0, \sqrt{2}]$. *Am dori să îi găsim funcția de repartiție în ipoteza că „la întîmplare" înseamnă că punctele A şi B sunt vectori aleatori independenți repartizați uniform în pătrat.*

În ambele cazuri problema este similară, deşi cu grad de dificultate tehnică diferit.

Fie $(X_n)_n$ un şir de variabile i.i.d. cu funcția de repartiție *F*. Deci $F(x) = P(X_j \leq x)$. Îi ataşăm **funcția de repartiție empirică** şi arătăm că ea converge la F.

Definiția 5.2.15. *Fie* $(X_n)_n$ *un şir de variabile aletoare i.i.d. Şirul de variabile aleatoare* $F_n(x) = \frac{1}{n}\left|\left\{j \le n : X_j \le x\right\}\right|$ *se numeşte* ***funcția de repartiție empirică calculată în x.***

Propoziția 5.2.16. *Pentru orice* $n \ge 1$, $Y_n := nF_n(x)$ *sunt variabile aleatoare repartizate Binomial(n, F(x)). Deci* $EY_n = F(x)$, $\mathrm{Var}(Y_n) = nF(x)(1 - F(x))$. *În plus,* $F_n(x) \xrightarrow{a.s.} F(x)$
În cuvinte: ***Funcția de repartiție empirică converge aproape sigur la adevărata funcție de repartiție.***

Demonstrație

Fie $Z_j = 1_{\{X_j \le x\}}$. Variabilele Z_j sunt i.i.d. repartizate Binomial(1,$F(x)$) iar $F_n(x)$ nu este altcineva decît $\frac{Z_1 + Z_2 + ... + Z_n}{n}$ care, conform SLLN converge a.s. la $EZ_1 = F(x)$. *q.e.d.*

Şi totuşi, **se poate şi mai bine**.

Este adevărat că $F_n(x)$ converge punctual la $F(x)$. Dar funcția de repartiție $F:\Re \to [0,1]$ este definită pe o mulțime nenumărabilă.

Nu cumva mulțimea $\Omega_0 := \{\omega \in \Omega \mid F_n(x)(\omega) \to F(x)$ pentru orice $x \in \Re\}$ poate să fie de probabilitate mică, sau chiar 0? Noi am vrea să estimăm funcția de repartiție F în **toate** punctele, nu numai într-un singur x!

Din fericire, lucrurile nu stau aşa. Nu numai că $P(\Omega_0) = 1$ (ceea ce tranşează problema), dar se poate demonstra chiar mai mult:

Teorema 5.2.17. (**Teorema lui Glivenko)** *Fie* $\Delta_n(\omega) = sup\ |F_n(x)(\omega) - F(x)|$ *distanța uniformă dintre funcția de repartiție empirică după n observații şi adevărata funcție de repartiție. Atunci* $\Delta_n \xrightarrow{a.s.} 0$

Funcția de repartiție empirică converge aproape sigur uniform la adevărata funcție de repartiție.

Nu vom da demonstrația acestui rezultat.[11]

Semnalăm că ea este foarte mult îmbunătățită dacă funcția de repartiție F este continuă. Un rezultat de matematică grea este următorul

Teorema 5.2.18. (Teorema Kolmogorov – Smirnov) *Dacă F este continuă, atunci*

[11] Cititorul interesat poate consulta Teoria probabilităților de Cuculescu (1974) sau Tudor (1980). Sau poate vedea mai multe referințe aici
http://en.wikipedia.org/wiki/Glivenko%E2%80%93Cantelli_theoremhttp://en.wikipedia.org/wiki/Glivenko%E2%80%93Cantelli_theorem

$$\lim_{n\to\infty} P(\sqrt{n}\,\Delta_n \le x) = \frac{\sqrt{2\pi}}{x}\sum_{i\ge 1} e^{-\frac{(2i-1)^2\pi^2}{8x^2}}$$

Metoda Monte Carlo

Metoda se foloseşte într-o serie de probleme unde apar calcule prea grele pentru a fi abordate determinist. Ele sunt de două tipuri: calcul de integrale sau probleme de optimizare.

Calcul de integrale.

Să presupunem că vrem să calculăm o integrală de forma

$$I = \int_C f(x_1,...,x_k)dx_1dx_2...dx_k = \int f 1_C d\lambda^k \quad (5.2.5)$$

unde C este un compact de măsură Lebesgue pozitivă şi f o funcție integrabilă pe acel compact. Ideea este să generăm un şir de vectori aleatori independenți X_n repartizați uniform în compactul C. Atunci $\frac{1}{n}(f(X_1)+...+f(X_n)) \xrightarrow{a.s.} Ef(X_1)$. Dar, conform formulei de transport, E$f(X_1)$ = $\int f dU_C = \frac{1}{\lambda^k(C)}\int_C f(x_1,...,x_k)dx_1dx_2...dx_k$ .Deci algoritmul este

$$\int_C f(x_1,...,x_k)dx_1dx_2...dx_k = \lambda^k(C)\times[\text{a.s. } \lim_{n\to\infty} \frac{1}{n}(f(X_1)+...+f(X_n)) \ ] \quad (5.2.6)$$

A simula un vector aleator repartizat uniform într-un compact nu este un lucru simplu. Uneori însă, este simplu: dacă $C = [a_1,b_1]\times[a_2,b_2]\times...\times[a_k,b_k]$. Atunci $X_n = (X_{n,1},...,X_{n,k})$ este uşor de simulat: componentele sale sunt variabile aleatoare independente repartizate Uniform(ai,bi). Toate mediile de programare (R, C++, Java, R, Matlab, Excel etc) au în dotare cel puțin generatoare de numere pseudoaleatoare, repartizate Uniform(0,1)

Revenind la Exemplul 1.1.11. Următoarea secvența (sau, cum i se mai spune, ***„script***") din mediul de programare „ R":

```
segment1<-function(n)
{xa=runif(n);xb=runif(n);ya=runif(n);yb=runif(n)
d=sqrt((xa-xb)^2+(ya-yb)^2)
d}
```

face *n* simulari (instrucțiunea `xa=runif(n)` produce un vector de lungime n cu componentele variabile aleatoare repartizate Uniform(0,1)

Cu instrucțiunea

```
> d<-segment1(1000000);summary(d)
```

generăm 1000000 de segmente cărora le calculăm lungimea. Apoi ni se furnizează minimul, maximul, prima cuantilă, mediana, media şi cuantila a treia.

Sigur că este vorba de cuantilele de selecție (**se poate arăta, tot pe baza SLLN că şi cuantilele de selecție converg la adevăratele cuantile**). Conform SLLN ne aşteptăm ca aceste variaile aleatoare (căci asta sunt!) să nu oscileze prea tare şi să ne dea informații despre repartiția variabilei aleatoare $X = \|A\text{-}B\|$.

Iată rezultatul a 10 simulări de cîte 1 milion de segmente

```
   Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
   0.0007994   0.3276000   0.5118000   0.5214000   0.7049000
1.3560000
   0.0002333   0.3278000   0.5116000   0.5211000   0.7043000
1.4010000
   0.0003137   0.3284000   0.5117000   0.5214000   0.7043000
1.3670000
   0.0005139   0.3279000   0.5119000   0.5213000   0.7044000
1.3760000
   0.0006177   0.3277000   0.5117000   0.5210000   0.7043000
1.3880000
   0.0008395   0.3284000   0.5119000   0.5212000   0.7041000
1.3860000
   0.0004015   0.3283000   0.5123000   0.5215000   0.7046000
1.3860000
   0.0008071   0.3279000   0.5113000   0.5209000   0.7039000
1.3790000
   0.0006265   0.3282000   0.5113000   0.5213000   0.7045000
1.3820000
   0.0005321   0.3284000   0.5126000   0.5215000   0.7046000
1.3800000
```

Observăm că media reprezintă o remarcabilă stabilitate: primele două zecimale nu se schimbă. La fel şi mediana. Putem avea o idee despre precizia estimării comparînd cu lucruri cunoscute: ştim că maximul esențial al lui X este $\sqrt{2}$ şi minimul este 0. Maximul empiric al lui X pare să fie ≈ 1.38 iar minimul pare a fi ≈ 0.

Calculul exact este aproape imposibil de făcut. Teoretic, avem de calculat următoarele

$$F(x) = P(X \le x) = P((x_A - x_B)^2 + (y_A - y_B)^2 \le x^2) \quad (5.2.7)$$

$$EX = \mathrm{E}\sqrt{(x_A - x_B)^2 + (y_A - y_B)^2} \quad (5.2.8)$$

unde $A(x_A,y_A)$, $B(x_B,y_B)$ sunt punctele aleatoare repartizate uniform în pătratul unitate.

Formal, se dau patru variabile aleatoare independente: x_A, x_B, y_A, y_B şi se cere să se calculeze cantitățile (5.2.7) şi (5.2.8). Prima revine la a calcula măsura Lebesgue 4-dimensională λ^4 a mulțimii $M_x = \{ x_A, x_B, y_A, y_B \in [0,1]: (x_A - x_B)^2 + (y_A - y_B)^2 \le x^2\}$

Iar a doua, pe baza formulei de transport, la a calcula integrala

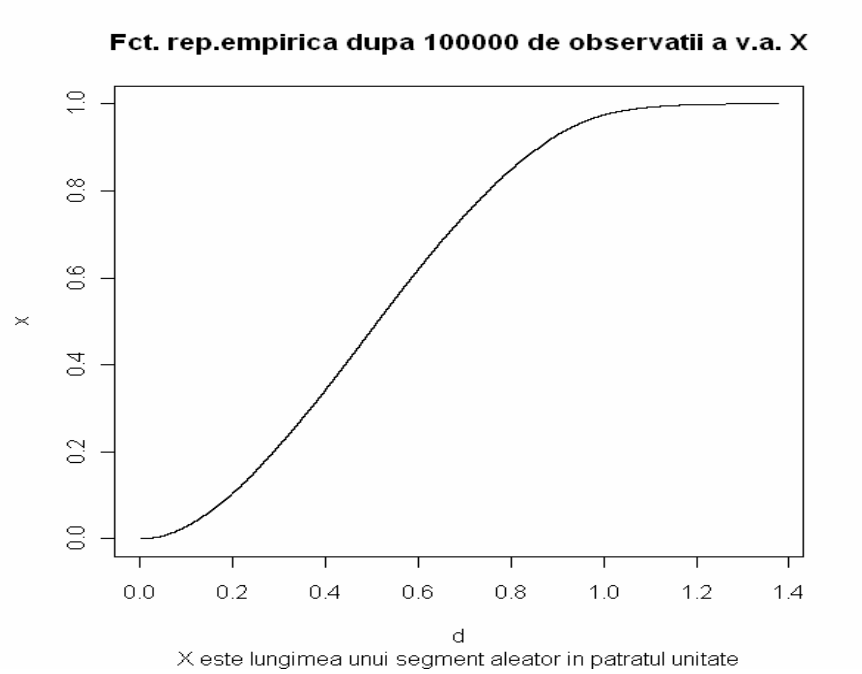
$$EX = \int_0^1\int_0^1\int_0^1\int_0^1 \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}\, dx_A dx_B dy_A dy_B$$

Prima cantitate se poate calcula exact, cu mult efort; pentru al doua, nu există formule prin cuadraturi.

Ca să avem o idee de puterea metodei, o putem compara cu metodele deterministe de calcul ale integralelor multiple. Tot o sumă avem de făcut, după ce luăm cîte o diviziune pe fiecare axă. Dacă diviziunea este echidistantă cu 20 de puncte, vom avea de calculat valoarea funcției de integrat în 204 = 160.000 de puncte. Nu e sigur că eroarea va fi mai mică!

Putem concluziona că X este o variabilă aleatoare cu media ≈0.521 şi mediana≈.512

De curiozitate, prezentăm şi graficul unei funcții de repartiție empirice după 100.000 de probe.



Calcul de maxime-minime.

O problemă fundamentală în matematica aplicată este de a găsi maximele şi minimele unei funcții $f:C \to \Re$ unde C este un domeniu din $\Re^{kk}$. Există multe metode deterministe de a face acest lucru – acesta este domeniul teoriei optimizării. Există două tipuri de agoritmi : determinişti şi probabilişti.

Ideea de bază a algoritmilor probabilişti este de a arunca o ploaie de puncte „la întîmplare" cu mai multă sau mai puțină inteligență .

Propoziția 5.2.19. *Fie $f:C \to \Re$ o funcție mărginită definită pe compactul cu interior nevid $C \subseteq \Re^k$. Fie $(X_n)_n$ un şir de variabile aleatoare repartizate uniform în C şi $Y_n = max(f(X_1),...,f(X_n))$, $Z_n = min(f(X_1),..,f(X_n))$. Atunci $(Y_n)_n$ este un şir crescător de variabile aleatoare, $(Y_n)_n$ este un şir descrescător . Primul*

converge aproape sigur la Esssupf iar al doilea la Essinf f.[12] *Dacă f este continuă, atunci* $Y_n \xrightarrow{a.s.} max(f)$ *şi* $Z_n \xrightarrow{a.s.} min(f)$

Demonstraţie

Fie F funcţia de repartiţie a variabilei aleatoare f(X) unde $X \sim$ Uniform(C). Deci $F(x) = P(f(X) \le x) = P(X \in f^{-1}(-\infty,x]) = \lambda^k(f^{-1}(-\infty,x])/\lambda^k(C)$.

Fie M = ess inf f. Atunci $P(Y_n \le M - \varepsilon) = P(\max(f(X_1),...,f(X_n)) \le M - \varepsilon)$. Dar variabilele $f(X_k)$ sunt independente, deci

$P(\max(f(X_1),...,f(X_n)) \le M - \varepsilon) = P(f(X_1) \le M - \varepsilon, f(X_2) \le M - \varepsilon,... f(X_n)) \le M - \varepsilon)$

$= P(f(X_1) \le M - \varepsilon)P(f(X_2) \le M - \varepsilon,)...P(f(X_n)) \le M - \varepsilon) = F^n(M - \varepsilon)$. Cum M este supremul esenţial, $F(M - \varepsilon) < 1$, deci $F^n(M - \varepsilon) \to 0$. Deci $P(Y_n \le M - \varepsilon) \to 0$. Dar şirul Y_n, fiind crescător, are o limită, Y_∞. Rezultă că $P(Y_\infty \le M - \varepsilon) = 0 \ \forall \ \varepsilon$. Deci $Y_\infty \ge M$. Pe de altă parte, $Y_n \le M$ deoarece toate variabilele aleatoare $f(X_j)$ au această proprietate. Înseamnă că $Y_\infty = M$ P-a.s. $\Leftrightarrow$ $Y_\infty = M$ (λ^k a.p.t.).

Dacă funcţia este continuă, nu mai este nevoie de precauţia ca ea să fie mărginită: sigur că este, deoarece orice funcţie continuă duce compacte în compacte. Mai mult, atunci maximul ei coincide cu supremul esenţial din următorul motiv: fie M = max f. Atunci mulţimea $\{x \in C \mid f(x) > M - \varepsilon\}$ este deschisă în C, deci are interior nevid. Orice mulţime de interior nevid are măsură pozitivă. Adică M are proprietatea care defineşte supremul esenţial.

Demonstrarea afirmaţiilor legate de minim sau infimul esenţial este analogă. *q.e.d.*

Observaţia 5.2.20. *Acesta este cel mai simplu algoritm, deoarece lucrează „în orb", fără memorie. Perfecţionarea lui duce la „algoritmii genetici". Dacă apelăm şi la un minimum de memorie, reţinînd punctele de minim şi maxim găsite, atunci putem avea o idee despre Argmin f şi Argmax f*[13]

Exemplul 5.2.21. *Pentru a vedea cît este de tare algoritmul, să luăm o funcţie căreia îi putem calcula extremele, de exemplu* $f(x,y) = x \wedge y - xy$, $f:[0,1]^2 \to \Re$. *Verificaţi că max f = f(½, ½) = ¼ , minf =*0. Puncte de minim sunt o infinitate – frontiera pătratului unitate, dar există un singur punct de maxim.. Aplicăm metoda Monte Carlo şi să vedem ce rezultă. După 1000 de simulări a rezultat minimul min f = 4.069785e-07 (în loc de 0), maximul max f = 0.2480471 (în loc de 0.25), punctul de maxim z = (0.5315545, 0.5295112) (în loc de (0.5, 0.5)!) şi un punct de minim de coordonate (0.8046973, 2.083834e-06) . Dacă însă facem

[12] Supremul esenţial al unei funcţii definite pe un compact de interior nevid C este un număr M cu proprietatea că $f(x) \le M$ pentru aproape toţi $x \in M$ şi măsura Lebesgue a mulţimii $\{x \in C \mid f(x) > M - \varepsilon\}$ este pozitivă $\forall \ \varepsilon>0$. Similar, infimul esenţial este un număr m cu proprietatea că $f(x) \ge M$ pentru aproape toţi $x \in M$ şi măsura Lebesgue a mulţimii $\{x \in C \mid f(x) < M + \varepsilon\}$ este pozitivă $\forall \ \varepsilon>0$. A nu se confunda cu supremul şi infimul De exemplu, dacă $f = 1_\Delta - 1_{\Delta'}$ cu Δ diagonala pătratului unitate din plan, şi Δ' cealaltă diagonală atunci sup f = 1, inf f = -1, dar ess sup f = ess inf f = 0.

[13] O notaţie pentru punctele în care se găseşte maximul sau minimul.

10000 de simulări, obținem minf =*9.063544e-09, max f = 0.2499187, Argmax(f) = (0.4966409, 0.4967819), Argmin(f) =(0.9999011, 9.165588e-05)* .

De regulă algoritmii Monte Carlo nu se aplică decît in extremis – dacă nu avem altceva mai bun.

Viteza de convergență la WLLM şi SLLM este dată de $\sigma^2 = Var(X_i)$.

5.3. Convergența în repartiție

Spre deosebire de convergențele din capitolul precedent, convergența în repartiție nu se referă ca convergența variabilelor aleatoare, ci la cea a repartițiilor lor. Propoziția „X_n converge la X în repartiție" cu notația $X_n \xrightarrow{D} X$ trebuie înțeleasă în sensul „**repartițiile variabilelor aleatoare X_n converg la repartiția lui X "**.

De obicei, noțiunea de convergență este legată de o topologie: un şir de repartiții F_n are limita F dacă în afara oricărei vecinătăți a lui F există cel mult un număr finit de termeni ai şirului.

Şi tot de obicei, noțiunea de vecinătate este legată de o **distanță**: o vecinătate a unui punct F este o mulțime care conține o bilă de centru F şi rază ε.

Convergența tare

Cele mai folosite distanțe sunt date de norme: $d(F,G) = \|F - G\|$.

Repartițiile variabilelor aleatoare sunt probabilități pe dreaptă. Probabilitățile sunt măsuri finite pe $(\Re, B(\Re))$. Diferența a două măsuri finite este o măsură cu semn.

Aici trebuie amintite unele lucruri elementare: măsurile finite cu semn pe un spațiu măsurabil $(E,\mathcal{E})$ formează spațiu vectorial. O măsură cu semn $\mu: E \to \Re$ se poate întotdeauna scrie sub forma $\mu = \mu_+ - \mu_-$ unde μ_+ şi μ_- reprezintă partea pozitivă şi partea negativă a măsurii. Aceasta este descompunerea Hahn – Jordan.[14]

Mai mult, există o mulțime $H \subseteq E$ [15] cu proprietatea $\mu_+(E) = \mu(H)$ şi $\mu_-(E) = -\mu(E \setminus H)$. Ea are proprietatea că $\mu(A \cap H) \geq 0$, $\mu(A \setminus H) \leq 0$ pentru orice $A \in E$.

Măsura $|\mu| := \mu_+ + \mu_-$ se numeşte **variația** lui μ iar funcția definită prin $\|\mu\| = |\mu|(E) = 2\mu(H) - \mu(E)$ este o normă: $\|\mu\| = 0 \Rightarrow \mu = 0$ şi $\|\mu_1 + \mu_2\| \leq \|\mu_1\| + \|\mu_2\|$ $\forall$ μ_1, μ_2 măsuri cu semn.

---

[14] Vezi orice manual de teoria măsurii sau, de exemplu aici: http://www.math.purdue.edu/~zhang24/SignedMeasure.pdf

[15] Ea se numeşte *mulțimea Hahn* ataşată lui μ. Dacă, de exemplu, $\mu = \rho \cdot \nu$, unde ν este o măsură oarecare, atunci $H = \{\rho > 0\}$

Norma se calculează uşor dacă μ are o densitate faţă de o măsură adevărată, ν: mai precis, dacă $\mu = \rho \cdot \nu$, atunci $\|\mu\| = \int |\rho| d\nu$ [16].

De exemplu dacă F şi G sunt două repartiţii discrete cu acelaşi suport, $F = \sum_j p_j \delta_{x_j}$, $G = \sum_j q_j \delta_{x_j}$, atunci $\nu = \sum_j \delta_{x_j}$, densitatea lui F ar fi $(p_j)_j$, cea a lui G ar fi $(q_j)_j$ iar distanţa între F şi G ar fi $\|F - G\| = \sum_j |p_j - q_j|$.

Definiţia 5.3.1. *Fie $(E,\mathcal{E})$ un spaţiu măsurabil. Fie $(F_n)_n$ şi F probabilităţi pe el. Spunem că F_n converge tare la F dacă $\|F - F_n\| \to 0$. Notăm acest lucru prin „$F_n \xrightarrow{s} F$”.*

Propoziţia 5.3.2. *Fie $(E,\mathcal{E})$ un spaţiu măsurabil.*

(i). *Dacă F şi G sunt două probabilităţi pe E, atunci $d(F, G) \in [0,2]$. Dacă $d(F,G) = 0$, atunci $F = G$ iar dacă $d(F,G) = 2$, atunci există o mulţime A în aşa fel încît $F(A) = 0$ şi $G(A) = 1$. Spunem că F şi G sunt* ***singulare.***

(ii). *Dacă ν este o măsură oarecare şi $F_n = f_n \cdot \nu$, $F = f \cdot \nu$ sunt probabilităţi, atunci $F_n \xrightarrow{s} F \Leftrightarrow f_n \xrightarrow{L^1(E,\mathcal{E},\nu)} f$. O condiţie suficientă ca f_n să conveargă la f în L^1 este ca f_n să convearga la f ν - aproape sigur.*

(iii). *Dacă $F_n \xrightarrow{s} F$, atunci $F_n(A) \to F(A)$ $\forall A \in \mathcal{E}$. Reciproca nu este adevărată. Ca un caz particular, dacă F_n şi F sunt repartiţii de pe dreapta reală, atunci funcţiile lor de repartiţie converg: $F_n((-\infty,x]) \to F((-\infty,x])$.* [17]

Demonstraţie

(i). Fie $\mu = F - G$.Deci $\mu(E) = F(E) - G(E) = 1 - 1 = 0$. Fie H mulţimea Hahn ataşată lui μ. Atunci $\|\mu\| = 2\mu(H) - \mu(E) = 2\mu(H)$. Dacă $\|\mu\| = 2$, atunci $\mu(H) = 1 \Rightarrow F(H) - G(H) = 1$. Dar F şi G sunt probabilităţi, deci $F(H) = 1$ şi $G(H) = 0$.

(ii). $\|F_n - F\| = \int |f - f_n| d\nu$, deci e clară echivalenţa $F_n \xrightarrow{s} F \Leftrightarrow f_n \xrightarrow{L^1(E,\mathcal{E},\nu)} f$. Interesantă este cealaltă afirmaţie, deoarece în general nu este adevărat că dacă f_n converge la f aproape sigur, converge şi în L^1 (vezi exemplul 1.1.8). Dar la noi există condiţia foarte tare ca fn să conveargă la **o densitate de probabilitate**. Folosind egalitatea $|x| = 2x_+ - x$ şi avem

---

[16] Notaţia $\mu = \rho.\nu$ este acceptată de majoritatea matematicienilor pentru a desemna măsura de densitate ρ şi bază ν. Precis, sensul este $(\rho \cdot \nu)(A) = \int \rho 1_A d\nu$ pentru orice $A \in \mathcal{E}$. Alţi autori folosesc în acelaşi scop notaţia $d\mu = \rho d\nu$, care are avantajele ei, deoarece atunci cînd calculăm integrala, densitatea „iese în faţă”: $\int f d(\rho \cdot \nu) = \int f \rho d\nu$.

[17] Se obişnuieşte notaţia $F(x)$ în loc de $F((-\infty,x])$. Dacă suntem atenţi, nu este nici un pericol de confuzie: *dacă A este o mulţime, $F(A)$* înseamnă probabilitatea lui A şi *dacă x e punct, $F(x)$ înseamnă $F((-\infty,x])$.*

$\int |f - f_n| d\nu = 2\int (f - f_n)_+ d\nu + \int (f - f_n) d\nu = 2\int (f - f_n)_+ d\nu$. Şirul $(f - f_n)_+$ tinde ν-a.s. la 0 este dominat de f care este în L^1. Teorema de convergenţă dominată ne spune atunci că putem comuta limita cu integrala: $\lim_n \int |f - f_n| d\nu = \int \lim_n |f - f_n| d\nu = 0$.

(iii). $|F_n(A) - F(A)| \le |F_n - F|(A) \le \| F_n - F \|$. Pentru a doua afirmaţie, vezi exemplul 1.3.7 de mai jos. □

Exemple de aplicare

Exemplul 5.3.3. *Dacă $(p_n)_n$ este un şir de probabilităţi cu proprietatea că $np_n \to \lambda$, cu $\lambda > 0$, atunci repartiţiile Binomial(n,p_n) converg tare la Poisson(λ). Într-adevăr, putem lua $\nu = \sum_{n\ge 0} \delta_n$ măsura cardinal cu suport mulţimea numerelor naturale şi densităţile $f_n(i) = C_n^i p_n^i (1-p_n)^{n-i}$, $f(i) = \frac{\lambda^i}{i!}$. Verificaţi că $f_n \to f$.*

Exemplul 5.3.4. *Fie $F_n = \left(1 - \frac{1}{n}\right)\delta_0 + \frac{1}{n}\delta_n$ Atunci $F_n \xrightarrow{s} \delta_0$.*

Exemplul 5.3.5. *Dacă $a_n \to a$ şi $b_n \to b$, atunci Uniform(a_n,b_n) $\to$ Uniform(a,b).*

Exemplul 5.3.6. *Mai general, familiile de repartiţii obişnuite: (Geometric(p), Negbin(k,p), Gamma(k,λ), N(μ,σ^2) etc) sunt continue în parametru: $\lambda_n \to \lambda \Rightarrow$ Gamma(k,λ_n) $\xrightarrow{s}$ Gamma(k,λ); $\mu_n \to \mu$, $\sigma_n \to \sigma \Rightarrow N(\mu_n,\sigma_n) \xrightarrow{s} N(\mu,\sigma)$. Nu avem decît să verificăm că densiăţile converg.*

Exemplul 5.3.7. *Fie $A_n = \bigcup_{k=0}^{2^{n-1}-1} (\frac{2k}{2^n}, \frac{2k+1}{2^n}]$ şi $F_n = (2 \cdot 1_{A_n}) \cdot \lambda$, unde λ este măsura Lebesgue. Fie f = Uniform(0,1) = $1_{(0,1)} \cdot \lambda$. Atunci $\| F_n - F \| = 1$ (într-adevăr, norma este egală cu $\int |1_{A_n} - 1_{(0,1)}| d\lambda$ iar $(1_{A_n} - 1_{(0,1)})(x) = \begin{cases} 1 & \text{dacă } x \in A_n \\ -1 & \text{dacă } x \in (0,1) \setminus A_n \\ 0 & \text{în rest} \end{cases}$).*

Totuşi, $F_n(A) \to F(A)$ pentru orice mulţime boreliană A din următorul motiv:remarcăm că $F_n(A) \le 2F(A)$ pentru orice $A \in B(\mathfrak{R})$. Fie atunci $C = \{A \in B(\mathfrak{R}): F_n(A) \to F(A)\}$. Familia C este un u-sistem (singurul lucru cu probleme este să arătăm că dacă (A_k) este un şir de mulţimi disjuncte cu proprietatea că $F_n(A_k) \to F(A_k)$, atunci şi $F_n(\bigcup_k A_k) \to F(\bigcup_k A_k) \Leftrightarrow \sum_k F_n(A_k) \to \sum_k F(A_k)$ ceea ce rezultă din faptul că seriile $\sum_k F_n(A_k)$ sunt dominate de $2\sum_k F(A_k)$). Acest u-sistem conţine intervalele A = $(-\infty,x]$. Într-adevăr, pentru $x \le 0$ sau $x \ge 1$ funcţiile de repartiţie

chiar coincide: $F_n(x) = F(x)$. *Dacă* $0 < x < 1$ *se arată imediat prin inducţie că* $F_n(x) = x + \frac{|[2^{n-1}x + \frac{1}{2}] - 2^{n-1}x|}{2^{n-1}}$ *deci* $0 \le F_n - F \le 2^{-n}$.

Deci f*uncţiile de repartiţie converg uniform,* $(F_n - F)(A)$ *converge la* 0 *pentru orice A şi totuşi* F_n ***nu*** *converge tare la F.*

Exemplul 5.3.8. *Dacă F este o repartiţie discretă şi G este o repartiţie continuă, atunci* $\|F - G\| = 2$ *– adică maximul posibil. În* ***consecinţă niciodată nu se poate aproxima în sensul tare o repartiţie continuă cu un şir de repartiţii discrete.***

În multe situaţii se pune problema evaluării momentelor unei variabile aleatoare aproximîndui repartiţia cu alta. Schema mentală este „Dacă $X_n \xrightarrow{D} X$, atunci poate că şi $EX_n \to EX$”

Este oare convergenţa tare suficientă pentru a asigura convergenţa momentelor?

Uneori chiar aşa se întîmplă, dar în general răspunsul este negativ. Dacă luăm, de exemplu, $X_n \sim F_n := \begin{pmatrix} 0 & n \\ 1-\frac{1}{n} & \frac{1}{n} \end{pmatrix}$, $X = 0 \sim \delta_0$, vedem că deşi $F_n \xrightarrow{s} F$, $EX_n = 1$ nu converge la $EX = 0$. Problema convergenţei momentelor este dificilă.[18]

Convergenţa slabă

Convergenţa tare, deşi cea mai naturală, nu răspunde satisfăcător problemelor de statistică. Toate repartiţiile vizibile în statistică sunt repartiţii empirice, deci sunt discrete. Exemplul 5.3.7 ne arată că nu se pot aproxima repartiţiile continui cu repartiţii discrete. Ce puţin nu în sensul tare.

Teorema lui Glivenko ne spune că (uneori, vezi mai sus) funcţiile de repartiţie empirice converg la adevărata funcţie de repartiţie. O idee ar fi să decretăm că

Definiţia 5.3.9.(Definiţie intermediară) *Fie* $(F_n)_n$ *un şir de repartiţii pe dreaptă. Fie F o altă repartiţie. Spunem că* $F_n \to F$ *dacă* $F_n(x) \to F(x)$ *pentru orice* $x \in \Re$.

Dar această definiţie are două neajunsuri. Amîndouă serioase.

În primul rînd, ca să fie ceva care corespunde intuiţiei, ar trebui ca, dacă $x_n \to x$, atunci şi δ_{x_n} să conveargă la δ_x. Şi nu este aşa. De exemplu, dacă $x_n = \frac{1}{n}$, $x_n \to$

[18] Cititorul poate consulta, de exemplu, Ioan Cuculescu, *Teoria Probabilităţilor,* Bucureşti, All, 1998, pg 281-368.

0, dar funcțiile de repartiție sunt $F_n(x) = 1_{[\frac{1}{n},\infty)}(x) \to 1_{(0,\infty)}(x)$; limita nu este o funcție de repartiție fiindcă nu este continuă la dreapta. Am fi vrut să conveargă la funcția de repartiție a lui δ_0, care este $1_{[0,\infty)}$.

În al doilea rînd, ar fi preferabilă o definiție care să se poată extinde şi la alte spații măsurabile, nu numai la dreaptă. Am vrea să dăm un sens, de exemplu, şi noțiunii de convergență dacă avem de a face cu repartiții în plan sau în spațiu.

De aceea s-a ales altă definiție. Ea are sens pe spații mai generale, dar ne mulțumim aici cu spațiile euclidiene, care sunt cel mai bine cunoscute.

Definiția 5.3.10. *Fie $(F_n)_n$, F repartiții pe spațiul euclidian $(\Re^d, B(\Re^d))$. Spunem că F_n converge slab la F dacă $\int f dF_n \to \int f dF$ pentru orice funcție continuă şi mărginită f.*[19]

Notăm acest fapt prin „$F_n \Rightarrow F$”

Dacă X_n ,X sunt vectori aleatori cu repartițiile F_n şi F, în locul notației „$F_n \Rightarrow F$” se foloseşte, prin abuz de limbaj, notația „$X_n \xrightarrow{D} X$” care se citeşte „X_n converge în repartiție la X”. Se admite şi notația „$X_n \xrightarrow{D} F$” , care se citeşte „X_n converge în repartiție la F”.

Proprietățile cele mai importante ale acestei noțiuni sunt sintetizate în următorul rezultat – Teorema Portmanteau.

Propoziția 5.3.11. (Teorema Portmanteau)

Fie $E = \Re^d$, $d \geq 1$, $E = B(\Re^d)$ şi $(F_n)_n$, F probabilități pe (E, E).

Atunci următoarele proprietăți sunt echivalente

(i) $\int f dF_n \to \int f dF$ *pentru orice f continuă şi mărginită*

(ii) $\int f dF_n \to \int f dF$ *pentru orice f uniform continuă şi mărginită*

(iii) $\limsup F_n(C) \leq F(C)$ *pentru orice mulțime închisă din E*

(iv) $\liminf F_n(D) \geq F(D)$ *pentru orice mulțime deschisă din E*

(v) $\lim F_n(A) = F(A)$ *pentru orice mulțime A cu frontieră F- neglijabilă (adică $F(\overline{A} \setminus Int(A)) = 0$!).*

(vi) *(doar pentru $d = 1$):* $F_n(x) \to F(x)$ $\forall$ *x punct de continuitate pentru F*

(vii) *(doar pentru $d = 1$):* $F_n(x) \to F(x)$ $\forall$ $x \in \Gamma$ *unde Γ este o mulțime numărabilă densă din $\Re$.*

Nu vom demonstra această teoremă. Unele implicații sunt simple, altele mai laborioase.[20]

---

[19] *Mulțimea funcțiilor reale continue şi mărginite pe $\Re^d$ se notează cu $C_b(\Re^d)$.*

[20] Ioan Cuculescu, *Teoria Probabilităților* sau http://en.wikipedia.org/wiki/Convergence_of_measures. Sunt sute de cărți care conțin demonstrația.

Observația 5.3.12. *Toate punctele teoremei de mai sus pot scrise în termini de variabile aleatoare. Dacă în loc de F_n şi F punem X_n şi X obținem următoarele caracterizări echivalente ale faptului că $X_n \xrightarrow{D} X$:*

(i) $Ef(X_n) \to Ef(X)$ *pentru orice f continuă şi mărginită*

(ii) $Ef(X_n) \to Ef(X)$ *pentru orice f uniform continuă şi mărginită*

(iii) $limsup\ P(X_n \in C) \le P(X \in C)$ *pentru orice mulțime închisă din E*

(iv) $liminf\ P(X_n \in D) \ge P(X \in D$ *pentru orice mulțime deschisă din E*

(v) $lim\ P(X_n \in A) = P(X \in A)$ *pentru orice mulțime A cu frontieră F-neglijabilă*

Observația 5.3.13. *Nu trebuie să credem că dacă F_n sunt funcții de repartiție şi $F_n \to F$, atunci şi F este funcție de repartiție. Exemplele sunt nenumărate:* $1_{[n,\infty)}$ *sunt funcții de repartiție care converg la 0, la fel şi funcțiile de repartiție pentru Uniform(0,n) (adică* $F_n(x) = \min(\frac{x_+}{n},1)$ *) etc. Este fenomenul cunoscut ca "escape to infinity" sau "se pierde masă spre infinit".*

Observația 5.3.14. *Dacă dorim să avem o familie de probabilități care să fie relative compactă (din orice şir să se poată extrage un subşir Cauchy), atunci trebuie ca ea să fie "tight": pentru orice $\varepsilon > 0$ să existe un compact C cu proprietatea ca $F(C) > 1 - \varepsilon$ pentru toate probabilitățile F din acea familie (Teorema lui Prohorov)*[21].

Remarcăm următoarea consecință imediată a teoremei Portmanteau:

Corolarul 5.3.15. *Fie X_n.X vectori aleatori. Dacă $X_n \xrightarrow{P} X$, atunci $X_n \xrightarrow{D} X$. Convergența în probabilitate implică convergența în repartiție.*

Demonstrație

Dacă $X_n \xrightarrow{P} X$, atunci conține un subşir care converge a.s. la X. Dacă f este o funcție continuă, atunci $f(X_n)$ converge aproape sigur la $f(X)$. Dacă f este şi mărginită, teorema de convergență dominată arată că $Ef(X_n) \to Ef(X)$.

Ce avem de făcut dacă dorim să verificăm o conjectură de tipul "$F_n \Rightarrow F$"? Nici una din rețetele din Teorema Portmanteau nu pare să funcționeze. Nici măcar în cazul unidimensional: este uşor de zis "verifică dacă $F_n(x)$ converge la $F(x)$ pe o mulțime densă", dar e mai greu de făcut.

Există un instrument care ajută în multe cazuri, anume **funcția caracteristică**.

[21] De exemplu http://en.wikipedia.org/wiki/Prokhorov%27s_theorem

Definiția 5.3.16. *Fie F o repartiție pe* $(\Re^d, B(\Re^d))$. *Funcția* $\varphi_F: \Re^d \to \mathbb{C}$ *definită prin*

$$\varphi_F(t) = \int e^{it'x} dF(x) \quad (5.3.1)$$

se numeşte funcția caracteristică a lui F. (Un punct x din $\Re^d$ *este un vector coloană; t' este transpusul lui t deci t'x este produsul scalar dintre t şi x : t'x =* $t_1x_1 + t_2x_2 + ...+ t_dx_d$*). Dacă F este repartiția unui vector aleator d-dimensional, X, atunci scriem* φ_X *în loc de* φ_F *şi, pe baza formulei de transport, avem*

$$\varphi_X(t) = Ee^{it'X} \quad (5.3.2)$$

Funcția caracteristică are proprietatea de a fi multiplicativă: dacă X şi Y sunt vectori independenți, atunci $\varphi_{X+Y} = \varphi_X\varphi_Y$. Dacă este de clasă C^∞, atunci X are toate momentele finite şi ele se pot calcula prin derivări. Ea are însă o proprietate suplimentară pe care analogul său real (funcția generatoare de momente $m_X(t) = Ee^{t'X}$) nu o are: aceea că domeniul său de definiție este acelaşi indiferent de repartiția F şi că ea caracterizează astfel, repartiția. Mai precis, avem

Propoziția 5.3.17.
(i) *Fie F şi G două repartiții pe* $(\Re^d, B(\Re^d))$ *cu proprietatea că* $\varphi_F = \varphi_G$. *Atunci F = G (Teorema de unicitate).*
(ii) *Presupunem că* $(F_n)_n$ *este un şir de repartiții cu proprietatea că şirul* φ_{F_n} *este convergent şi limita sa,* φ, *este continuă în 0. Atunci există o repartiție F cu proprietatea că* $\varphi_F = \varphi$ *şi* $F_n \Rightarrow F$. *Sau, în termeni de variabile aleatoare: dacă* $(X_n)_n$ *sunt vectori aleatori d-dimensionali independenți şi* $\varphi_{X_n} \to \varphi$, φ *continuă în* 0, *atunci există o repartiție F ca* $X_n \xrightarrow{D} F$.

Nu vom demonstra nici această teoremă fundamentală. [22]

Cu ajutorul ei însă putem arăta

Propoziția 5.3.18.
(i). *Dacă* $F_n \Rightarrow F$ *şi* $G_n \Rightarrow G$, *atunci* $F_n \otimes G_n \to F \otimes G$ *(sau, acelaşi lucru în termeni de variabile aleatoare dacă* $X_n \xrightarrow{D} X$, $Y_n \xrightarrow{D} Y$,X_n *independent de* Y_n, *atunci* $(X_n, Y_n) \xrightarrow{D} (X, Y)$)
(ii). *Dacă* $F_n \Rightarrow F$ şi $G_n \Rightarrow G$, a*tunci* $F_n * G_n \to F * G$ (*în termeni de variabile aleatoare : dacă* $X_n \xrightarrow{D} X$, $Y_n \xrightarrow{D} Y$,X_n *independent de* Y_n, *atunci* $X_n + Y_n \xrightarrow{D} X + Y$)

[22] Ioan Cuculescu, *Teoria Probabilităților* sau Lukacs, E. (1970). *Characteristic functions*. London: Griffin.

Demonstrația se reduce la verificarea faptului banal că $\varphi_{F\otimes G} = \varphi_F\varphi_G$ şi $\varphi_{F*G} = \varphi_F\varphi_G$. *q.e.d.*

5.4. Teorema limită centrală

Am văzut că în varianta ei cea mai simplu de înțeles, Legea numerelor mari spune că dacă X_n sunt variabile aleatoare i.i.d., atunci media empirică $\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}$ converge aproape sigur la adevărata medie $\mu = EX$. Pentru a studia mai amănunțit viteza de convergență, ar trebui să calculăm mărimea $p(\varepsilon) = P(|\overline{X}_n - \mu| > \varepsilon)$; am dori să ştim cîte observații ne-ar trebui pentru ca această probabilitate să fie mică. Dacă scriem $\overline{X}_n = \frac{S_n}{n}$, probabilitatea în cauză s-ar scrie sub forma $p(\varepsilon) = P(|S_n - n\mu| > n\varepsilon) = P(S_n \in (-\infty, n(\mu-\varepsilon)) \cup (n(\mu+\varepsilon),\infty))$

Dacă notăm cu F_n repartiția sumelor S_n, atunci probabiliitatea respectivă s-ar putea scrie $p(\varepsilon) = F(n(\mu - \varepsilon)) + 1 - F(n(\mu+\varepsilon))$

Ce s-ar putea spune despre această cantitate?

Se ştie că repartiția sumei unor variabile aleatoare independente este convoluția erepartițiilor termenilor. Întrebarea este: cum se comportă convoluțiile de multe repartiții?

Să studiem un exemplu în care se pot face calcule. Să zicem că $X_n \sim U(0,1)$. Atunci vectorul $X:=(X_1,...,X_n)$ este repartizat uniform în cubul $[0,1]^n$. Deci $P(S_n \le x) = P(\mathbf{X} \in A_x)$ unde

$$A_x = \{\mathbf{x} \in [0,1]^n : x_1 + ... + x_n \le x\} \qquad (5.4.1)$$

Dacă $x \le 1$, este uşor de făcut calculul: $P(\mathbf{X} \in A_x)$ este volumul simplexului $S_n(x) = \{\mathbf{x \ge 0}: x_1 + ... + x_n \le x\}$ care, din rațiuni de simetrie este 1/n! din volumul cubului, adică $F_n(x) = \frac{x^n}{n!}$. Dacă însă $x > 1$, atunci trebuie scăzute din el volumele celor n simplexe de latură $x - 1$ care apar (făceți un desen în cazul n = 3!). Deci $A_x = S_n(x) \setminus \bigcup_{j=1}^{n} A_j(x)$ cu $A_j(x) = \{\mathbf{x} \in S_n(x): x_i > 1\}$. Intersecția unei familii finite de asemenea mulțimi e de aceeaşi formă. Aplicînd principiul includerii şi excluderii, găsim

$$F_n(x) = \frac{1}{n!}\left(x^n - C_n^1(x-1)^n + C_n^2(x-2)^n - .....\right) \qquad (5.4.2)$$

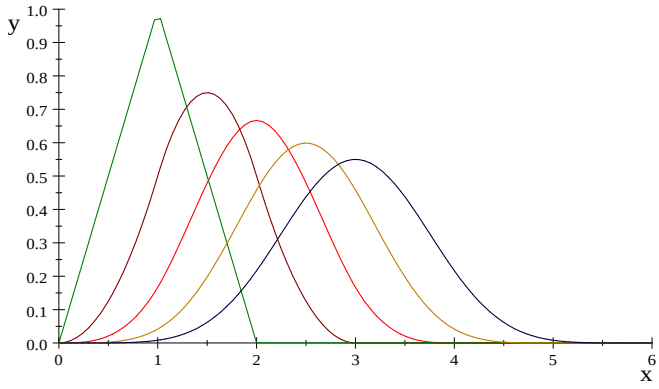
Unde suma se face atîta vreme cît $x - k \ge 0$. Ca să nu avem probleme de sumare, scriem

$$F_n(x) = \frac{1}{n!}\sum_{k=0}^{\infty} C_n^k (-1)^k (x-k)_+^n \qquad (5.4.3)$$

unde x+ este partea pozitivă a lui x. Derivînd (sumele sunt, totuşi, finite) găsim densităţile

$$f_n(x) = \frac{1}{(n-1)!}\sum_{k=0}^{\infty} C_n^k (-1)^k (x-k)_+^{n-1} \qquad (5.4.4)$$

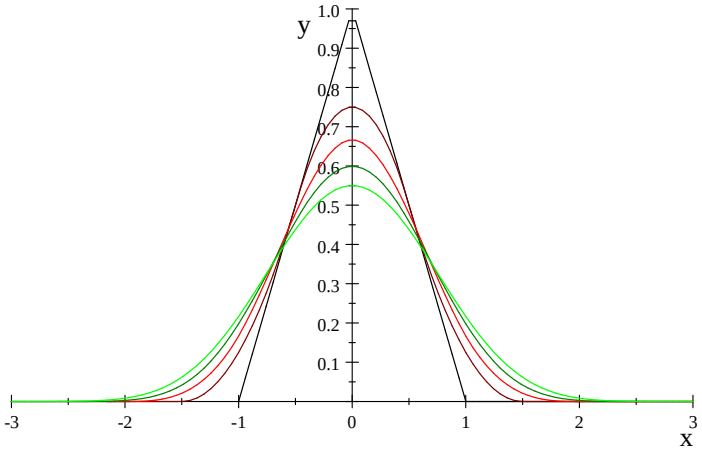
În figura de mai jos am făcut graficele densităţilor f_j cu $2 \le j \le 6$.
Se observă cum ele capătă o formă specifică, de clopot.



Ca sa putem compara mai bine densităţile, ar trebui să facem ca aceste densităţi să aibă aceeaşi axă de simetrie. Adică să centrăm variabilele aleatoare X_n, scăzînd din ele media. Astfel obţinem sumele centrate $S_{n,c} = S_n - n\mu$ unde $\mu = EX_n = ½$. Funcţiile de repatiţie centrate, notate ad-hoc cu $F_{n,c}$ se calculează imediat după foemula evidentă

$$F_{n,c}(x) = F_n(x + n\mu)\ ,\ f_{n,c}(x) = f_n(x + n\mu) \qquad (5.4.5)$$

Obţinem cinci grafice care se pot compara mai bine, fiindcă au aceeaşi axă de simetrie.



Totuşi, dispersiile tind la infinit, aşa că şi densităţile centrate vor tinde la 0. Nu putem să le comparăm bine. Ca să facem să aibă toate aceeaşi dispersie, împărţim la abaterea medie pătratică a lui S_n, care este $\sigma\sqrt{n}$, unde $\sigma^2 = \mathrm{Var}(X_n) = \frac{1}{12}$. Obţinem sumele centrate şi normate

$$s_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \tag{5.4.6}$$

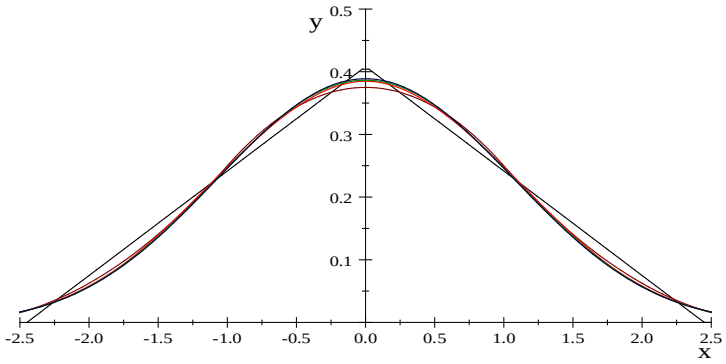
care au funcţiile de repartiţie notate cu

$$\Phi_n(x) = P(s_n \le x) = F_n(n\mu + \sigma\sqrt{n}\, x) \tag{5.4.7}$$

şi densităţile

$$\gamma_n(x) = \Phi_n'(x) = \sigma\sqrt{n}\, f_n(n\mu + \sigma\sqrt{n}\, x) \tag{5.4.8}$$

Mai jos am făcut graficele celor cinci densităţi centrate şi normate. Se observă cum se stabilizează.



Cine este limita?

Să luăm un caz particular, în care chiar putem face calcule: să zicem că X_n ~ Exp(1) sunt toate repartizate exponențial standard. Verificați imediat prin inducție că

$$f_{n+1}(x) = \frac{x^n}{n!}e^{-x}1_{(0,\infty)}(x) \qquad (5.4.9)$$

Acum $\mu = \sigma = 1$, deci conform cu (1.4.8) avem $\gamma_{n+1}(x)=\sqrt{n+1}\, f_n(n+1+\sqrt{n+1}\, x)$ adică

$$\gamma_{n+1}(x) = \sqrt{n+1}\frac{\left(n+1+\sqrt{n+1}x\right)^n}{n!}e^{-\left(n+1+\sqrt{n+1}x\right)}1_{(0,\infty)}(n+1+\sqrt{n+1}x) \qquad (5.4.10)$$

Dacă n este mare, $n+1 + x\sqrt{n+1}$ devine pozitiv, deci putem renunța la indicator.Aplicăm formula lui Stirling, $n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$ şi avem $\gamma_{n+1}(x) \approx$

$$\sqrt{\frac{n+1}{2\pi n}}\left(\frac{n+1+\sqrt{n+1}x}{n}\right)^n e^{n-\left(n+1+\sqrt{n+1}x\right)} = \sqrt{\frac{n+1}{2\pi n}}\left(\frac{n+1+\sqrt{n+1}x}{n+1}\right)^n\left(\frac{n+1}{n}\right)^n e^{-\left(1+\sqrt{n+1}x\right)}.$$

Trecînd la limită avem

$$\lim_{n\to\infty}\gamma_n(x) = \lim\left(e^{-1}\sqrt{\frac{n+1}{2\pi n}}\left(\frac{n+1}{n}\right)^n\right)\lim\left(\frac{n+1+\sqrt{n+1}x}{n+1}\right)^n e^{-\sqrt{n+1}x} = \sqrt{\frac{1}{2\pi}}\lim\left(\left(1+\frac{x}{\sqrt{n+1}}\right)^n e^{-\sqrt{n+1}x}\right)$$

$$=\sqrt{\frac{1}{2\pi}}\mathrm{L}$$

Logaritmăm: $\ln L = \lim\left(n\ln\left(1+\frac{x}{\sqrt{n+1}}\right) - x\sqrt{n+1}\right)$. Dezvoltăm logaritmul în serie $(\ln(1+t) = t - t^2/2 + t^3/3 - t^4/4 + ...)$ şi avem

$$\ln L = \lim\left(\frac{n}{n+1}(n+1)\left(\frac{x}{\sqrt{n+1}} - \frac{x^2}{2(n+1)} + \frac{x^3}{3(n+1)\sqrt{n+1}} - \frac{x^4}{4(n+1)^2} + ... \right) - x\sqrt{n+1}\right)$$

$$= \lim\left(\left(x\sqrt{n+1} - \frac{x^2}{2} + \frac{x^3}{3\sqrt{n+1}} - \frac{x^4}{4(n+1)} + ... \right) - x\sqrt{n+1}\right) = - x^2/2.$$

Concluzie
Şirul densităților centrate şi normate converge, în cazul în care X_n sunt repartizate Exp(1) la funcția $\gamma(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Aceasta chiar este o densitate, este densitate repartiției noemale standard N(0,1)! (Atenție: limita unui şir de densități nu este obligatoriu o densitate, după cum ne putem convinge cu densitățile $f_n = 1_{[0,n]}(x)/n$ care converg la 0!)

Am verificat pe un caz particular

Teorema 5.4.1(Teorema limită centrală locală) *Dacă densitatea comună a variabilelor aleatoare i.i.d. din* L^2 , $(X_n)_n$, *este mărginită, atunci densitățile* γ_n *ale sumelor centrate şi normate* $s_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ *converge la densitatea repartiției normale standard :* $\gamma_n(x) \to \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. *În consecință, conform Propoziției 5.3.2.(ii)*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{s} N(0,1) \qquad (5.4.11)$$

Afirmația este valabilă într-un context şi mai general, anume dacă putem demonstra că există un n începînd de la care γn este o mărginită. Demonstrația depăşeşte cu mult cadrul acestui manual. Cine chiar este interesat o poate găsi de exemplu, în Y. Ptohorov, Y. Rozanov, Probability Theory, Springer 1969, pp 190-194. Alte demonstrații, mai recente se pot găsi pe Internet, cu Google.

Dar dacă variabilele aleatoare X_n sunt discrete, atunci problema locală nu are sens. Se poate demonstra un rezultat mai slab.

Teorema 5.4.2.(Teorema Limită Centrală (TLC))
Fie $(X_n)_n$ *un şir de variabile aleatoare i.i.d. din* L^2. *Fie* $\mu = EX_1$ şi $\sigma^2 = \mathrm{Var}(X_1)$. *Atunci*

$$\frac{X_1 + X_2 + ... + X_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{D} N(0,1)$$

Scrisă explicit, afirmaţia este că

$$lim_{n\to\infty} P\left(\frac{X_1 + ... + X_n - n\mu}{\sigma\sqrt{n}} \le x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

Observaţia 5.4.3. *Funcţia de repartiţie a repartiţiei* N(0,1) *se notează cu* Φ *şi este tabelată de peste 100 de ani. Acum nu se mai folosec tabelele, deoarece toate softurile matematice o calculează. Deci*

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

Demonstraţie
Dacă acceptăm teorema de convergenţă a funcţiilor caracteristice, demonstraţia este simplă. Fie $Y_n = X_n - \mu$ variabilele centrate şi $\varphi(t) = Ee^{itY}$ funcţia lor caracteristică. Ştim de la proprietăţile funcţiilor caracteristice că dacă Y_n sunt din L^2, atunci φ este derivabilă de două ori. Ne interesează că φ e derivabilă de două ori în 0: deci putem scrie $\varphi(t) = \varphi(0) + t\varphi'(0) + \frac{t^2}{2}\varphi''(0) + o(t)\cdot t^2$ unde o(t) este o funcţie continuă şi $o(0) = 0$. Dar $\varphi(0) = 1$, $\varphi'(0) = EY_n = 0$ şi $\varphi''(0) = -EY_n^2 = -\sigma^2$. Deci $\varphi(t) = 1 - \frac{t^2\sigma^2}{2} + o(t)$.
Calculăm funcţia caracteristică a lui s_n, notată cu φ_n:

$$\varphi_n(t) = E\, e^{\frac{it}{\sigma\sqrt{n}}(Y_1+...+Y_n)} = \varphi^n\left(\frac{t}{\sigma\sqrt{n}}\right) = \left(1 - \frac{t^2\sigma^2}{2n\sigma} + o\left(\frac{t}{\sigma\sqrt{n}}\right)\frac{t^2}{n\sigma^2}\right)^n.$$

Atunci

$$\lim_n \varphi_n(t) = \exp[\lim\left(n\left(-\frac{t^2}{2n} + \frac{t^2}{n\sigma^2}o\left(\frac{t}{\sigma\sqrt{n}}\right)\right)\right)] = \exp\left(-\frac{t^2}{2} + \frac{t^2}{\sigma^2}\lim o\left(\frac{t}{\sigma\sqrt{n}}\right)\right) = e^{-\frac{t^2}{2}}.$$

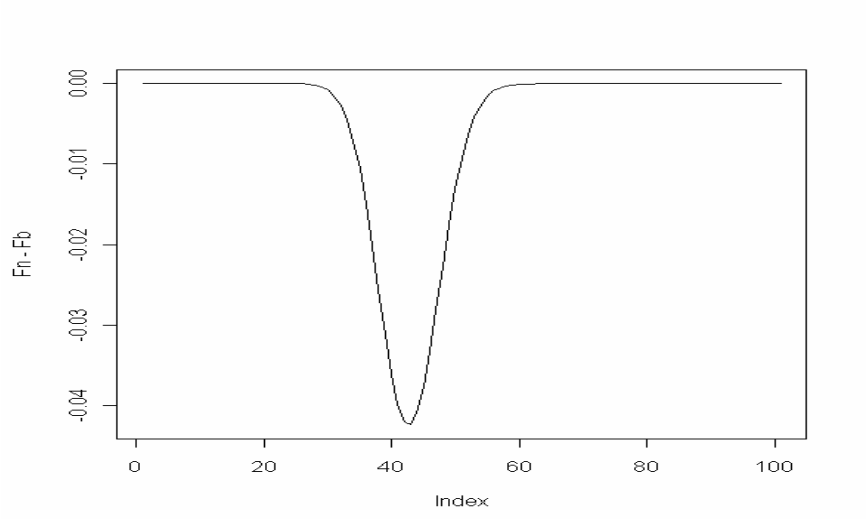
Dar funcția $\varphi(t) = e^{-\frac{t^2}{2}}$ este funcția caracteristică a repartiției normale standard. Teorema este demonstrată. *q.e.d.*

Exemplu de aplicare:

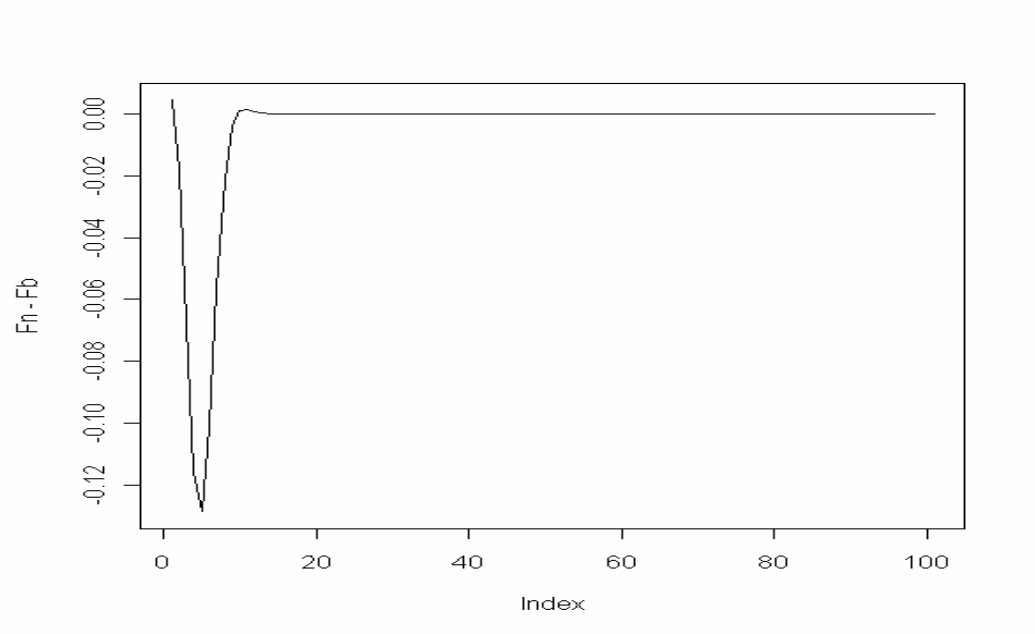
Exemplul 5.4.4. *Repartiția binomială. Dacă* $X_n \sim \begin{pmatrix} 0 & 1 \\ q & p \end{pmatrix}$, *atunci* $S_n \sim$ *Binomial(n,p). Dacă n este destul de mare putem aproxima* $P(S_n \le a)$ *cu TLC astfel: Avem* $\mu = p$, $\sigma = \sqrt{pq}$ *. Deci*

$$P(S_n \le a) = P\left(\frac{S_n - np}{\sqrt{npq}} \le \frac{a - np}{\sqrt{npq}}\right) \approx \Phi\left(\frac{a - np}{\sqrt{npq}}\right).$$

Se ştie că dacă $npq \ge 20$, *aproximarea este foarte bună. Prezentăm un grafic cu diferențele dintre funcția de repartiție a repartiției Binomial(100,0.42) şi cea a repartiției* $N(np, \sqrt{npq}^2) = N(42, \sqrt{24.36}^2)$ *calculate pentru* $x \in [0,100]$



Pe axa Ox sunt valorile lui x iar pe axa Oy valorile diferenței dintre cele două funcții de repartiție. Maximul este maimic de 0.05. Daca însă luam un caz extrem, cu p =0.042, situația se schimba



Diferența dintre probabilități poate depăşi 0.12, ceea ce este imens. Explicația este că acum μ*=4.2 şi probabilitatea ca X ~ Binomial(100,0.042) să ia valorile 4 sau 5 este mare : 0.3653754. Ca să fie aplicabilă aproximarea normală trebuie ca toate probabilitățile P(X = k) să fie mici. Dacă produsul np este mic, deşi n este mare, atunci este preferabilă aproximarea cu repartiția Poisson(np).*

Capitolul 6

Simularea variabilelor aleatoare

În dotarea calculatoarelor există de mult generatoare de numere aleatoare, care simulează destul de bine un şir de variabile aleatoare repartizate Uniform(0,1). A simula o variabilă aleatoare înseamnă ca, pe baza acestui generator de numere aleatoare să se construiască variabile aleatoare avînd o repartiție dată.

Formal, problema s-ar pune aşa: se dă o variabilă aleatoare U~Uniform(0,1) şi o funcție de repartiție, F. Să se construiască o funcție f astfel ca funcția de repartiție a variabilei aleatoare $X = f(U)$ să fie F.

Sau, mai general, se dau k variabile aleatoare $(U_j)_{1\le j\le k}$ i.i.d., repartizate Uniform(0,1) şi se cere să se construiască o funcție $f:\Re^k \to \Re$ ca în aşa fel încît variabila aleatoare $X = f(U_1,...,U_k)$ să aibă funcția de repartiție F.

Există mai multe medii de programare care fac acest lucru în cazul repartițiilor clasice. De exemplu, în „R", mediu de programare gratuit, care se poate descărca de pe internet există posibilitatea simulării (şi nu numai) cel puțin a următoarelor repartiții

Repartiția	Numele ei în R	Parametri
beta	beta	α,β
binomială	binom	k,p
Cauchy	cauchy	m,a
χ^2	chisQ	n
exponentială	exp	λ
F	f	m,n
gamma	gamma	ν,λ
geometrică	geom	p
hipergeometrică	hyper	a, n, k[1]
log-normală	lnorm	μ,σ
logistică	logis	μ,σ [2]
negativ binomială	nbinom	k,p
normală	norm	μ,σ^3
Poisson	pois	λ

[1] Extragem k bile dintr-o urnă cu a bile albe şi n bile negre; X este numărul de bile albe.

[2] Repartiția logistică are funcția de repartiție $\frac{1}{1+e^{-\frac{x-\mu}{\sigma}}}$. Este mai rar folosită..

[3] În codificarea "R", x=rnorm(1,μ,σ) produce o variabila aleatoare $X \sim N(\mu,\sigma^2)$. Parametrul al doilea reprezintă abaterea medie pătratică şi nu varianța. Uneori se face confuzie.

Student	t	n
uniformă	unif	a,b
Weibull	weibull	k, λ
Wilcoxon	wilcox	m, n

Pentru a genera un vector cu n componente i.i.d. cu aceste repartiţii se pune în faţa numelui lor din "R" litera r. 4 Apoi se pune numărul de variabile aleatoare dorite şi parametrii repartiţiei.

De exemplu:

x=rnorm(100,10,4);x : x este un vector cu 100 de componente repartizate $N(10,4^2)$

x=rbinom(10,10,.4);x

[1] 3 5 1 5 4 6 4 5 6 3 : x este un vector cu 10 componente repartizate Binomial(10,.4)

x=rnbinom(6,10,.4);x

[1] 19 8 9 25 8 19 : x este un vector cu 10 componente repartizate Negbin(10,.4)

x=rhyper(10,4,6,6);x

[1] 3 2 3 3 2 2 2 2 1 3 : x este un vector cu 10 componente ~Hypergeometric(4,6,6)

Întrebarea este cum generăm noi variabile aleatoare cu o repartiţie care nu face parte din cele simulate de mediile de programare?

6.1. Simularea repartiţiilor pe dreaptă

Algoritmul general: metoda cuantilei

Cum putem folosi o variabilă U ~ Uniform(0,1) pentru a simula o variabilă aleatoare X cu o repartiţie dată?

Să presupunem pentru început că funcţia de repartiţie a lui X este bijectivă. Mai precis, presupunem că există un interval $I \subseteq \Re$ astfel ca $F: I \to [0,1]$ să fie bijectivă. Cum este şi crescătoare, fireşte că ar trebui ca F să fie şi continuă – dacă ar fi discontinuă într-un punct a, atunci imaginea sa, Im(F), nu ar conţine intervalul ($F(a\text{-}0)$, $F(a)$).

[4] Dacă vrem să le calculăm funcţia de repartiţie punem *p*, pentru densitate punem *d* iar pentru cuantile (vezi mai jos) punem *Q*. De exemplu
x=pnorm(1,0,1);y=dnorm(1,0,1);z=Qnorm(0.01,0,1) va produce numerele x = 0.8413447 (căci $\Phi(1)$ = 0.8413447), $y = 0.2419707 = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ cu $x = 1$ şi $z = -2.326348 = \Phi^{-1}(0.02)$

Observația fundamentală este că variabila aleatoare $X = F^{-1}(U)$ are exact funcția de repartiție F. Acesta este **algoritmul inversei funcției de repartiție**.
Într-adevăr, dacă $x \in [0,1]$, atunci , ținînd seama de ipoteza că U ~ Uniform (0,1) $\Leftrightarrow P(U \le x) = x \; \forall \; x \in [0,1]$, avem $P(X \le x) = P(F^{-1}(U) \le x) = P(F(F^{-1}(U)) \le F(x))$ deci

$$P(X \le x) = P(U \le F(x)) = F(x). \qquad (6.1.1)$$

În general, funcțiile de repartiție nu sunt bijective. Dacă, de exemplu, $X \sim$ Uniform($\{1,2,...,n\}$), atunci $F(x) = \frac{\lceil x_+ \rceil}{n} \wedge 1$ ia doar valorile k/n cu $0 \le k \le 1$. Este o situație tipică pentru repartițiile discrete. Dar, chiar dacă repartiția este continuă, e posibil ca F să nu fie injectivă. De exemplu, dacă $X \sim$ Uniform($[0,1] \cup [2,3]$), atunci F_X este constantă pe intervalul $[1,2]$(verificați: $F(x) = (x_+ \wedge 1 + (x\text{-}2)_+ \wedge 1)/2$).

Ce se mai poate salva în acest caz din algoritmul anterior?

Ideea este să înlocuim inversa cu **cuantila.**

Definiția 6.1.1. *Fie F o funcție de repartiție. O cuantilă a sa este orice funcție reală $Q = Q_F$ definită pe* (0,1) *cu proprietatea că* $F(Q(u)\text{-}0) \le u \le F(Q(u)) \; \forall \; u$.

Dacă F este inversabilă, există o unică cuantilă, anume inversa lui F : $Q = F^{-1}$

Dacă nu, pot exista o infinitate. De exemplu, dacă X ~ Binomial (1, ½) = Uniform($\{0,1\}$), atunci $F(x) = ½$ pentru $x \in [0,1)$. Verificați că orice funcție de forma $Q(u) = 1_{(½,1)}(u) + \alpha 1_{\{½\}}$ cu $\alpha \in (0,1)$ este o cuantilă.

Propoziția 6.1.2. *Fie* $F:\Re \to [0,1]$ *o funcție de repartiție și* $Q:(0,1) \to \Re$ *o cuantilă a sa. Fie U o variabilă aleatoare repartizată Uniform(0,1). Atunci* $X = Q(U)$ *este o variabilă aleatoare avînd funcția de repartiție F.*

Demonstrație

Observăm că orice cuantilă este o funcție crescătoare. Într-adevăr, dacă $u<v$, atunci $Q(u) \le Q(v)$ deoarece în caz contrar, $Q(u) > Q(v) \Rightarrow F(Q(u) - 0) \ge F(Q(v)) \Rightarrow u \ge F(Q(u)\text{-}0) \ge F(Q(v)) \ge v \Rightarrow u \ge v$.

Arătăm că $\{U < F(x)\} \subseteq \{ Q(U) \le x\} \subseteq \{ U \le F(x) \}$și va fi suficient, pentru că atunci rezultă că $P(U < F(x)) \le P(Q(U) \le x) \le P(U \le F(x))$ de unde, cum $U \sim$ Uniform(0,1), deducem că $P(Q(U) \le x) = F(x)$.

Presupunem $Q(U) \le x$. Atunci $F(Q(U)) \le F(x)$. Dar, din definiția cuantilei, $U \le F(Q(U))$, deci $U \le F(x)$. Am demonstrat că

$$\{ Q(U) \le x \} \subseteq \{ U \le F(x)\} \qquad (6.1.2)$$

Să presupunem acum că $Q(U) > x$. Atunci $F(Q(U) - 0) \ge F(x)$. Dar, tot din definiția cuantilei. $U \ge F(Q(U) - 0)$, deci $U \ge F(x)$.Așadar $\{ Q(U) > x \} \subseteq \{ U \ge F(x)\}$

Trecînd la complementară,avem că

$$\{ Q(U) \le x \} \supseteq \{ U < F(x)\} \qquad (6.1.3)$$

Din (6.1.2) și (6.1.3) deducem că $\{ U < F(x)\} \subseteq \{ Q(U) \le x \} \subseteq \{ U \le F(x)\}$. *q.e.d.*.

Două sunt cazurile extreme care interesează în calcule: **cuantilă inferioară şi cuantila superioară.**

Cuantila inferioară, notată cu Q -, se defineşte prin relația

$$Q^{-}(u) = \sup\{F < u\} = \inf\{F \geq u\} \qquad (6.1.4)$$

iar cea superioară, notată cu Q^{+} se defineşte prin

$$Q^{+}(u) = \sup\{F \leq u\} = \inf\{F > u\} \qquad (6.1.5)$$

Propoziția 6.1.3. *Funcțiile definite prin relațiile* (6.1.4) şi (6.1.5) *sunt cuantile. Orice altă cuantilă Q este cuprinsă între ele:* $Q^{-} \leq Q \leq Q^{+}$.

Demonstrație

Din definiția supremului avem următoarele lucruri evidente

$F(Q^{-}(u) - \varepsilon) < u$, $F(Q^{-}(u) + \varepsilon) \geq u \ \forall \varepsilon > 0$

$F(Q^{+}(u) - \varepsilon) \leq u$, $F(Q^{+}(u) + \varepsilon) > u \ \forall \varepsilon > 0$

Trecînd la limită cu $\varepsilon \downarrow 0$, deducem că $F(Q^{-}(u) - 0) \leq u$, $F(Q^{-}(u) + 0) \geq u$ şi $F(Q^{+}(u) - 0) \leq u$, $F(Q^{+}(u) + 0) \geq u$. Dar F este continuă la dreapta, deci ambele funcții verifică definiția cuantilei.*q.e.d.*

Exemplul 6.1.4. *Dacă* $X \sim Uniform(\{0,1\})$ *atunci* $F(x) = ½$ *pentru* $x \in [0,1)$. *Cuantilele inferioară şi superioară sunt* $Q^{-}(u) = 1_{(½,1)}(u)$, $Q^{+}(u) = 1_{[½,1)}(u) = [2u]$ *deci* $X = [2U]$ *este o variabilă aleatoare cu repartiția cerută.*

Exemplul 6.1.5. *Dacă* $X \sim Uniform(\{1,2,...,n\})$, *atunci* $Q^{+}(u) = 1 + [nu] \Rightarrow X = 1 + [nU]$

Exemplul 6.1.6. *Dacă* $X \sim Exponential(1)$, *atunci* $F(x) = 1 - e^{-x}$ *este chiar bijectivă de la* $(0,\infty)$ *la* $(0,1)$, *deci* $Q^{+} = Q^{-} = F^{-1} \Rightarrow Q(x) = -\ln(1-u)$. *Putem pune* $X = -\ln(1-U)$. *Dar U şi* $1 - U$ *au aceeaşi repartiție,* $Uniform(0,1)$, *deci putem la fel de bine să punem* $X = -\ln U$.

Exemplul 6.1.7. *Dacă* $X \sim Negbin(1,p)$, *atunci* $F(x) = 1 - q^{[x]+1}$, *unde* $x \geq 0$ *şi* $q = 1 - p$. *O cuantilă a sa este* $Q(u) = \left[\frac{\ln u}{\ln(1-p)}\right]$. *Deci* $X = \left[\frac{\ln u}{\ln(1-p)}\right]$ *este o variabilă aleatoare cu repartiția cerută.*

Exemplul 6.1.8. *În general, dacă repartiția F este discretă,*

$F = \begin{pmatrix} a_1 & a_2 & a_3 & \dots & \dots \\ p_1 & p_2 & p_3 & \dots & \dots \end{pmatrix}$ *cu* $a_1 < a_2 < a_3 < ...$, *atunci cuantilele ei sunt*

$Q^{+}(u) = a_1 1_{(0,s_1)}(u) + a_2 1_{[s_1,s_2)}(u) + a_3 1_{[s_2,s_3)}(u) + ...$ *şi*

$Q^{-1}(u) = a_1 1_{(0,s_1]}(u) + a_2 1_{(s_1,s_2]}(u) + a_3 1_{(s_2,s_3]}(u) + ...$

unde $s_1 = p_1$, $s_2 = p_1 + p_2$, $s_3 = p_1 + p_2 + p_3$ ...

Deci variabila aleatoare $X = \sum_{k\geq 1} s_k 1_{[s_{k-1}, s_k)}(U)$ *are exact repartiția* F .

Iată un script în „R” *care* calculează cuantila unei repartiții discrete $\begin{pmatrix} a_1 & a_2 & a_3 & \dots & a_k \\ p_1 & p_2 & p_3 & \dots & p_k \end{pmatrix}$. Aici „a” este vectorul care contine valorile variabilei aleatoare X; „p” este vectorul care cuprinde probabilitățile ca să se ia aceste valori. Ambii vectori au lungimea k. Vectorul „o” este permutarea care trebuie făcută pentru ca numerele a_i să fie puse în ordine crescătoare iar „F” este repartiția propriu-zisă; acum e scrisă canonic. Vectorul „s” cuprinde sumele $s_k = p_1 + p_2 + ... + p_3$. Interesant este cît de simplu este de găsit locul lui u: instrucțiunea which($s >= u$) produce mulțimea $J(u) = \{i: s_i \geq u\}$ căreia i se ia primul element $i = \min(J(u)) = Q^+(u)$.

```
cuantila<-function(u,a,p)
{o=order(a)  #sortez pe a, pentru ca se poate sa
nu fie in ordine
F=rbind(a[o],p[o]); a=F[1,]; p=F[2,]
#calculez sumele partiale
s=p; for (i in 1:length(a)){s[i]=sum(p[1:i])}
#caut locul lui u
v=which(s>=u);i=min(v); q=a[i]
q}
```

Apelarea funcției se face cu instrucțiunea `q=cuantila(u,a,p)`
Un exemplu concret: să se simuleze n =1000 variabile aleatoare avînd repartiția $F = \frac{1}{10}\begin{pmatrix} -6 & 0 & 1 & 2 & 3 \\ 2 & 1 & 2 & 1 & 4 \end{pmatrix}$

```
n=1000;a=c(-
6,0,1,2,3);p=c(2,1,2,1,4)/10;u=runif(n,0,1)
x=u;for (i in 1:n) {x[i]=cuantila(u[i],a,p)}
```

Pentru a verifica dacă e bine, folosim instrucțiunea „table(x)” care arată repartiția empirică a unui vector x : sub fiecare valoare diferită pe care o ia vectorul se scrie numărul de apariții a acelei valori (frecvența absolută). În cazul nostru avem:

```
table(x)
x
 -6   0   1   2   3
195  99 190  98 418
```

Valorile frecvențelor absolute corespund celor teoretice, care ar fi trebuit să fie 200,100, 200,100,400. Deci ese plauzibil. Există metode de verificare a acurateței modelulului dat de o repartiție : de exemplu metoda „qqplot”.[5]

Pentru valori relativ mici ale lungimii k a vectorilor a şi p algoritmul este satisfăcător şi rapid. Poate fi folosit pînă la $k \approx 1000$.

Dacă, însă, vectorii au lungime prea mare încep să apară erori de maşină şi, pe de altă parte, viteza lui scade. Este şi normal. De aceea, dacă se poate, ar fi bine să fie folosiți algoritmi rapizi bazați pe diverse le repartiții care se pot obține din repartiția uniformă.

Problema este de a calcula cuantila dacă repartiția F nu este neapărat discretă şi cu suportul format dintr-o mulțime cu număr mic de elemente.

Chiar dacă avem funcția de repartiție F dată printr-o formulă analitică (de exemplu $F(x) = pF_1 + qF_2$ cu $F_1 = \text{Exp}(1)$, $F_2 = \text{Gamma}(2,1) \Rightarrow F(x) = 1 - (1 + qx)e^{-x}$) nu avem formule pentru a calcula inversa $F^{-1}(u)$ (în cazul de mai sus ar trebui rezolvată ecuația $1 - (1 + qx)e^{-x} = u$, $0<u<1$, care este o ecuație transcendentă). Pentru a ieşi din dilemă ar trebui să facem o discretizare a lui F. Înlocuim funcția de repartiție F cu repartiția $F_d = \begin{pmatrix} a_1 & a_2 & \dots & a_k & \alpha \\ F(a_1) & F(a_2)-F(a_1) & \dots & F(a_k)-F(a_{k-1}) & 1-F(a_k) \end{pmatrix}$ unde α este un număr mare şi aplicăm algoritmul cuantilei pentru această repartiție. Eventual un algoritm modificat în care înlocuim funcția F_d cu o linie poligonală care unește punctele de coordonate $(a_j, F(a_j))_j$. Sigur că pierdem din precizie.

Algoritmi speciali bazați pe proprietăți ale repartițiilor

Două sunt operațiile mai importante care se fac cu repartițiile: mixtura şi convoluția.

Definiția 6.1.9. *Fie $(F_j)_{1\le j\le n}$ o mulțime de repartiții pe dreaptă şi fie $(p_j)_{1\le j\le n}$ numere pozitive şi de sumă 1. Atunci repartiția $F = p_1F_1 + ...+ p_nF_n$ se numeşte mixtură de F. Dacă $X_j \sim F_j$ sunt variabile aleatoare independente, atunci repartiția sumei lor $S=X_1 + ...+ X_n$ este $F_1*F_2*...*F_n$.*

De exemplu, dacă $F(x) = pF_1 + qF_2$ cu $F_1 = \text{Exp}(1)$, $F_2 = \text{Gamma}(2,1)$, atunci F este o mixtură de exponențială cu Gamma.

Observația este că dacă ştim să simulăm variabilele X_i, atunci simulăm mai uşor variabilele X (cu repartiția F) şi S (cu repartiția $F_1*F_2*...*F_n$), fără a trece prin metoda pseudoinversei. Pentru S nu avem ce comenta, dar nu este absolut evident cum construim pe X.

[5] Chiar în “R” există instrucțiunea „qqplot”.

Propoziția 6.1.10. *Fie* $X_j \sim F_j$ *cu* $1 \le j \le n$ *şi fie* J *o variabilă aleatoare independentă de* $(X_j)_j$ *cu repartiția* $J \sim \begin{pmatrix} 1 & 2 & \dots & \dots & n \\ p_1 & p_2 & \dots & \dots & p_n \end{pmatrix}$. *Atunci variabila* $X := X_J$ *are repartiția* $F = \sum_{j=1}^{n} p_j F_j$.

Demonstrație

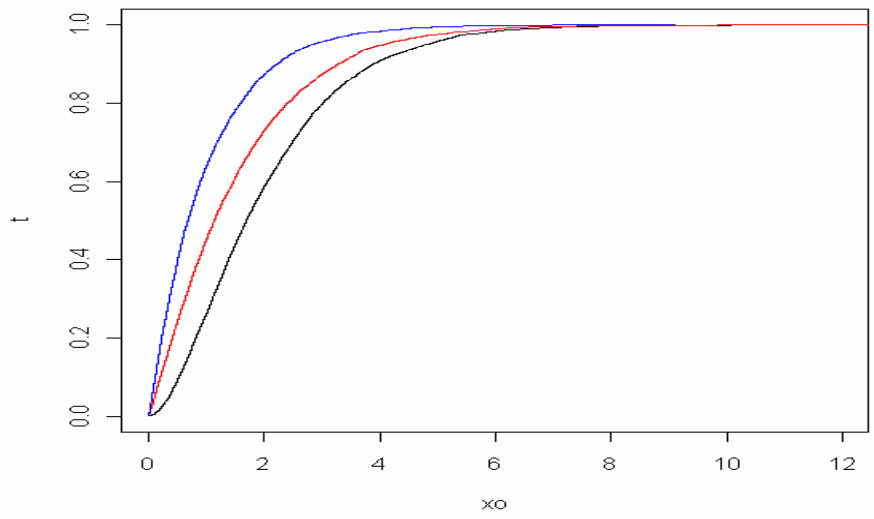
$P(X \le x) = P(X_J \le x) = \sum_{j=1}^{n} P(X_J \le x; J = j) = \sum_{j=1}^{n} P(X_j \le x) P(J = j) = \sum_{j=1}^{n} F_j(x) p_j$ *q.e.d.*

Deci, revenind la exemplul nostru cu F = pGamma(1,1) + qGamma(2,1) algoritmul exact este: simulăm variabila $J \sim \begin{pmatrix} 1 & 2 \\ p & q \end{pmatrix}$. Dacă J = 1, simulăm X ~ Exp(1) = Gamma(1,1) iar dacă J=2 simulăm X = Gamma(2,1). După cum se vede în scriptul următor, care foloseşte funcția „cuantila" din paragraful anterior

```
mixtura<-function(n,p) # simulează n variabile aleatoare ~ pExp(1)+qgamma(2,1
{a=c(1,2);pr=c(p,1-p)
x1<-rexp(n,1)          # simulează n variabile aleatoare ~ Exp(1)
;x2<-rgamma(n,2,1)     # simulează n variabile aleatoare ~ Gamma(2,1)
z<-rbind(x1,x2)        # se formează cu ele o matrice z de tip 2×n
x=x1                   # se inițializează x
for (i in 1:n)
{u=runif(1,0,1)        # se generează o variabilă aleatoare U ~ Uniform(0,1)
j=cuantila(u,a,pr)     # se generează variabila aleatoare J
x[i]=z[j,i]            # x = z(J,.)
}
x                      # se simulează n variabile aleatoare ~ Exp(1)
}
```

Ne putem convinge că este aşa dacă încercăm mai multe variante de p. Pentru p = 0 avem variabile repartizate Gamma(2,1); pentru p = ½ este o mixtură cu ponderi egale iar pentru p = 1 avem doar variabile repartizate Exp(1). Scriptul următor face cîte 5000 de simulări de fiecare tip şi apoi face greficul celor trei funcții de repartiție empirice, care ar trebui să semene cu cele adevărate

```
t=1:5000;t=t/5000
xo=mixtura(5000,0);xo=sort(xo)
xm=mixtura(5000,0.5);xm=sort(xm)
xu=mixtura(5000,1);xu=sort(xu)
plot(xo,t,type="l");lines(xm,t,col="red");
lines(xu,t,col="blue")
```



Exemplul 6.1.11. *Să presupunem că nu avem acces la un software performant, dar vrem să simulăm o variabilă aleatoare X ~ Gamma(n,λ) cu n număr întreg. Cum facem?*

Soluție

Variabilele $X_j = (-\ lnU_j)/\lambda$ *sunt repartizate Exp(λ). Dar Gamma*(n,λ) = Exp$(\lambda)^{*n}$ *exponențiala convolutată cu ea însăşi de n ori. Deci soluția este* $X = -\frac{\sum_{j=1}^{n} \ln U_j}{\lambda}$

Dar dacă vrem să simulăm $X \sim N(\mu,\sigma^2)$ şi nu avem decît un software de bază? Fireşte, este suficient să simulăm Y ~ N(0,1) şi apoi să punem X = μ + σY. Dar cum simulăm o normală standard? Algoritmul cuantilei nu ne dă decît o aproximare, pentru că nu putem calcula exact nici măcar Φ, funcția caracteristică a repartiției N(0,1), cu atît mai puțin să ăi calculăm şi cuantila. Putem face o aproximare bună, este adevărat, dar asta cere timp.

Propoziția 6.1.12. ***Metoda Box – Muller.*** *Fie U,V ~ Uniform(0,1) două variabile aleatoare independente. Atunci variabilele* $X = \sin(2\pi U)\sqrt{-2\ln V}$, $Y = \cos(2\pi U)\sqrt{-2\ln V}$ *sunt două variabile aleatoare independente repartizate* N(0,1). *Deci* $(X,Y) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$

Demonstrație

Avem de arătat că dacă $f:\Re^2 \to \Re$ este o funcție continuă şi mărginită, atunci $Ef(X,Y) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-\frac{x^2+y^2}{2}}dxdy$

Din formula de transport

$$Ef(X,Y) = \int_0^1\int_0^1 f(\sqrt{-2\ln u}\cos(2\pi v), \sqrt{-2\ln u}\sin(2\pi v))dudv \qquad (6.1.6)$$

Facem schimbarea de variabilă

$$x = \sqrt{-2\ln u}\cos(2\pi v), y = \sqrt{-2\ln u}\sin(2\pi v) \qquad (6.1.7)$$

Rezultă $x^2 + y^2 = -2\ln u$, de unde

$$u = e^{-\frac{x^2+y^2}{2}} \qquad (6.1.8)$$

Iacobianul este $\frac{D(x.y)}{D(u,v)} = \begin{vmatrix} -\frac{\sin(2\pi v)}{u\sqrt{-2\ln u}} & -2\pi\sqrt{-2\ln u}\cos(2\pi v) \\ -\frac{\cos(2\pi v)}{u\sqrt{-2\ln u}} & 2\pi\sqrt{-2\ln u}\sin(2\pi v) \end{vmatrix} = \frac{2\pi}{u}$

Imaginea mulțimii $[0,1]\times[0,1]$ prin această transformare este $\Re^2 \setminus \{0\}$.Cum punctul $\{0\}$este o mulțime neglijabilă față de măsura Lebesgue în plan, deducem din relația (6.1.8) că

$dxdy = \frac{2\pi}{u}dudv \Rightarrow dudv = \frac{u}{2\pi}dxdy = \frac{e^{-\frac{x^2+y^2}{2}}}{2\pi}dxdy$ deci relația (6.1.6) devine

$$\int_0^1\int_0^1 f\left(\sqrt{-2\ln u}\cos(2\pi v), \sqrt{-2\ln u}\sin(2\pi v)\right)dudv = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)e^{-\frac{x^2+y^2}{2}}dxdy \qquad (6.1.9)$$

exact ceea ce trebuia verificat.*q.e.d.*

Prin metoda Box – Muller se generează variabile aleatoare normale folosind doar două variabile uniform repartizate. Este evident o mare simplificare.

Există şi o metodă exactă de a simula repartiția Poisson(λ), fără a calcula cuantilele. Uneori ea este mai rapidă.

Propoziția 6.1.13. *Fie* $(U_n)_n \sim$ *Uniform*(0,1) *un şir de variabile aleatoare i.i.d. Fie*

$$N = \min\{n \geq 1 : U_1U_2...U_n < e^{-\lambda}\} - 1 \qquad (6.1.10)$$

Atunci $N \sim$ *Poison(*λ*).*

Demonstrație

Fie $\sigma_n = -\frac{\ln U_n}{\lambda}$ şi $T_n = \sigma_1 + ...+ \sigma_n$. Atunci σ_n sunt i.i.d. repartizate Exponential(λ). Logaritmînd expresia din (6.1.10) avem $N = \min\{n : T_n > 1\}$. Deci $P(N = 0) = P(T_1>1) = e^{-\lambda}$ iar dacă $n \geq 1$, atunci $P(N = n) = P(T_n \leq 1, T_{n+1} > 1\} = P(T_n \leq 1) - P(T_{n+1} \leq 1) = P(T_{n+1} > 1) - P(T_n > 1)$. Dar $T_n \sim$ Gamma(n,1), deci

$$P(T_n > t) = \left(1+\frac{\lambda t}{1!}+\frac{\lambda^2 t^2}{2!}+....+\frac{\lambda^{n-1}t^{n-1}}{(n-1)!}\right)e^{-\lambda t} \qquad (6.1.11)$$

Rezultă că $P(T_{n+1} > t) - P(T_n > t) = \frac{\lambda^n t^n}{n!}e^{-\lambda t}$. Am demonstrat chiar mai mult: că dacă punem $N(t) = \min\{n : T_n > t\}$, atunci $N(t) \sim$ Poisson(λt). [6]*q.e.d.*

Deci dacă dorim să simulăm, de exemplu, N ~ Poisson(1), înmulţim un şir de variabile aleatoare uniform repartizate pînă cînd produsul lor devine mai mic decît 1/e. Dacă pentru acest lucru a fost nevoie de n variabile aleatoare, atunci declarăm că N = n -1.

În statistica Bayesiană apare frecvent repartiţia Beta(m,n). Dacă m şi n sunt numere întregi, şi aceste repartiţii se pot simula fără a se calcula cuantilele.

Propoziţia 6.1.14. *Fie* $(U_j)_{1 \leq j \leq n} \sim$ *Uniform(0,1) independente. Sortăm aceste variabile aleatoare sub forma (* $U_{(1)} < U_{(2)} < ...< U_{(n)}$*). Atunci* $U_{(k)} \sim$ Beta($k, n+1-k$).

Demonstraţie

Fie $A_{n,k}$ = { exact k dintre variabilele aleatoare U_j sunt mai mici decît x } Evident $P(A_{n,k}) = C_n^k x^k (1-x)^{n-k}$. Să observăm că evenimentul ($U_{(k)} \leq x$) se poate scrie sub forma $A_{n,k} \cup A_{n,k+1} \cup ...\cup A_{n,n}$. Cum mulţimile $(A_{n,j})_{0 \leq j \leq n}$ sunt disjuncte, găsim că funcţia de repartiţie a lui $(U)_{(k)}$ este $P(U_{(k)} \leq x) = C_n^k x^k (1-x)^{n-k} + C_n^{k+1} x^{k+1}(1-x)^{n-k+1} +....+ C_n^n x^n (1-x)^{n-n}$, care este exact funcţia de repartiţie a unei variabile aleatoare Beta($k,n+1-k$). Densitatea sa este

$$f_k(x) = kC_n^k x^{k-1}(1-x)^{n-k} \qquad (6.1.12)$$

Exemplul 6.1.15. *Să simuleze o variabilă aleatoare X ~ Beta(10,2). Acum n = 10 + 2 – 1 = 11. Simulăm 11 variabile uniforme şi le sortăm; o luăm pe cea de a zecea. Dacă avem la dispoziţie un program de sort, nu e nici o problemă.*

De exemplu, în „R" secvenţa ar fi:

```
u = runif(11);u=sort(u);x=u[10]
```

[6] Procesul stochastic $(N(t))_{t\geq 0}$ se numeşte **procesul Poisson de intensitate λ.** Se poate demonstra că el este **cu creşteri independente**, adică dacă $t_1 < t_2 < ...< t_n$, atunci variabilele aleatoare ($N(t_1)$, $N(t_2) - N(t_1)$,....., $N(t_n) - N(t_{n-1})$) sunt independente.

6.2. Algoritmul general: teorema de descompunere

Ideea de bază este următoarea: orice repartiţie d-dimensională se poate scrie ca produsul dintre o probabilitate pe dreaptă şi mai multe probabilităţi de trecere.

Pentru a înţelege, să studiem următorul exemplu simplu

Un punct aleator bidimensional $Z = X + iY$ (îl scriem ca număr complex, este mai simplu)

$$Z\sim \frac{1}{8}\begin{pmatrix} 0 & 1 & 2 & 1+i & 2i & -1+i & i & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Atunci $X \sim \frac{1}{8}\begin{pmatrix} -1 & 0 & 1 & 2 \\ 2 & 3 & 2 & 1 \end{pmatrix}$. Dacă X = -1, atunci Y poate lua două valori, 0 şi 1. repartiţia sa condiţionată de faptul că X = -1 se o scriem (abuz de notaţie, dar sugestiv)

$(Y \mid X = -1) \sim \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$. Analog, avem $(Y \mid X = 0) \sim \frac{1}{3}\begin{pmatrix} 0 & 1 & 2 \\ 1 & 1 & 1 \end{pmatrix}$, $(Y \mid X = 1) \sim \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$ şi, în sfîrşit, $(Y \mid X = 2) \sim \delta_0$.

Pentru a-l genera, simulăm mai întîi prima componentă, cu algoritmul cuantilei. Apoi avem patru variante: dacă X = -1, sau X = 1 simulăm pe Y $\sim \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$, dacă X = 0 simulăm $Y \sim \frac{1}{3}\begin{pmatrix} 0 & 1 & 2 \\ 1 & 1 & 1 \end{pmatrix}$ iar dacă X = 2, punem Y = 0.

Ideea este că P(X = i, Y = j) = P(X = i)P(Y = j | X = i). Pentru fiecare valoare a lui X avem o altă repartiţie pentru Y, **repartiţia condiţionată**. Generăm pe X şi, apoi, depinzînd de valoarea lui X simulăm pe Y cu repartiţia condiţionată (Y | X = x).

Formal, acest lucru se poate generaliza astfel:

Algoritmul general. *Fie $Z = (X,Y) \in E \times F$ un vector aleator. Să presupunem că repartiţia lui X este Π şi că există o familie de repartiţii pe F, $(Q_x)_{x \in E}$ cu proprietatea că $P(Y \in B \mid X) = Q_X(B)$. Atunci repartiţia vectorului Z este $\Pi \otimes Q$.*

Familia de probabilităţi $(Q_x)_{x \in E}$ se numeşte **repartiţia lui Y condiţionată de X.** Un mod intuitiv de a o nota este $F_{Y \mid X}$. Un mod şi mai intuitiv este să scriem $Q_x = F_{Y \mid X = x}$. Aceasta este notaţia din statistică. Avem de lămurit ce înseamnă probabilitatea condiţionată $P(Y \in B \mid X)$. Dacă X este discretă, nu e nici o problemă: $P(Y \in B \mid X) = \sum_{x \in A} P(Y \in B \mid X = x) 1_{(X=x)}$ unde $A = \{x \in E \mid P(X = x) > 0\}$. Dar dacă X este continuă, avem o problemă, deoarece P(Y∈B | X = x) nu are sens.

Definiția 6.2.1. *Probabilitate condiționată, medie condiționată. Fie* (Ω,K,P) *un spațiu probabilizat,* $Y: \Omega \to F$, $X: \Omega \to E$ *două variabile aleatoare cu valori în spații măsurabile*(E,E) *și* (F,F). *Fie* $B \in F$. *Atunci definim* $P(Y \in B \mid X) = E(1_B(Y) \mid X)$.

Dacă $U: \Omega \to \Re$ *este o variabilă aleatoare, cu moment de ordin 1, atunci definim*

$$E(U \mid X) = \varphi(X) \Leftrightarrow E(U\psi(X)) = E(\varphi(X)\psi(X))$$

$\forall\ \psi: E \to \Re$ *măsurabilă și mărginită.*

Se demonstrează că media condiționată există și că, în caz că U *are și moment de ordin 2, ea are următoarea proprietate de optim*

$$E(U - \varphi(X))^2 \le E(U - h(X))^2\ \forall\ h:\Re \to \Re \text{ măsurabilă ca } E[h(X)]^2 < \infty.$$

Media condiționată are trei proprietăți importante, care se folosesc în practică:

- Dacă X și Y sunt independente, atunci $E[h(Y) \mid X] = Eh(Y)$ (la independență nu contează condiționarea). Cum o variabilă aleatoare constantă este independentă de orice altă variabilă aleatoare, deducem că dacă X este constantă, $E[h(Y) \mid X] = Eh(Y)$.
- $E[h(X) \mid X] = h(X)$ și, mai general, $E[h(X)Y \mid X] = h(X)E[Y \mid X]$ (funcțiile X-măsurabile se comportă precum constantele). Aici h este o funcție măsurabilă și mărginită.
- $E[E[Z \mid X,Y] \mid X] = E[Z \mid X]$ (proprietatea de iterativitate).

  În particular, dacă X = constant (mod P) avem $E[E[Z \mid Y]] = E[Z]$.

Deci, a spune că repartiția lui $(Y \mid X)$ este Q înseamnă a spune că

$$E[h(Y) \mid X] = \int h(y)dQ_X(y) \text{ pentru orice funcție măsurabilă } h: E \to \Re \qquad (6.2.1)$$

Un alt mod de a scrie același lucru (de multe ori mai comod) este

$$E[h(Y) \mid X] = \int h(y)Q(X, dy) \qquad (6.2.\ 2)$$

Relația (2.1.1) se poate prelungi la funcții de două variabile

$$E[h(X,Y) \mid X] = \int h(X, y)Q(X, dy) \qquad (6.2.3)$$

Într-adevăr, formula (6.2.3) este imediată dacă $h(x,y) = f(x)g(y)$, apoi se prelungește la indicatori de mulțimi de forma $h = 1_{A\times B}$; familia mulțimilor C pentru care formula este adevărată este un u – sistem, deci această familie conține σ-algebra $E \otimes F$ etc .

Existența repartițiilor condiționate este dată de:

Teorema 6.2.2 ***Teorema de dezintegrare.*** *Fie* P *o probabilitatea pe* $E\times F$ *unde* $E = \Re^m$ și $F = \Re^n$. *Fie* $\Pi(A) = P(A\times F)$. *Atunci există o probabilitate de trecere de la* E *la* F, *notată cu* Q *în așa fel încît* $P = \Pi\otimes Q$

Nu vom demonstra această teoremă. Se găsește în manualele de teoria probabilităților, de exemplu G. Ciucu, C. Tudor, *Teoria probabilităților*, Editura

Academiei, Bucureşti 1981 sau I. Cuculescu, *Teoria Probabilităţilor,* Editura ALL, Bucureşti 1998. Sau se poate căuta pe internet, Dissintegration Theorem.

Important este să lămurim ce spune teorema şi să înţelegem cum se aplică.

Definiţia 6.2.3. *Fie (E, $\mathcal{E}$) şi (F, $\mathcal{F}$) spaţii măsurabile. O funcţie $Q:E\times \mathcal{F}\to[0,1]$ se numeşte probabilitate de trecere de la E la F dacă*

(i) *aplicaţia $x\mapsto Q(x, B)$ este măsurabilă pentru orice $B\in F$*

(ii) *aplicaţia $B\mapsto Q(x, B)$ este o probabilitate pe $\mathcal{F}$ pentru orice $x\in E$*

Putem gîndi o probabilitate de trecere şi altfel: ca o colecţie de probabilităţi pe *F,* $(Q_x)_{x\in E}$. Condiţia (i) este una tehnică, pentru a se putea face calcule.

Exemplul 6.2.4. *Familiile clasice de repartiţii de pe dreaptă pot fi gîndite ca fiind probabilităţi de trecere: Q_λ = Exponential(λ) este o probabilitate de trecere de la* $(0,\infty)$ *la* $(0,\infty)$, $Q_{(n,p)}$ *= Binomial(n,p) este o probabilitate de trecere de la* $\{1,2,3,...\}\times[0,1]$ *la $\mathfrak{R}$, etc..*

Produsul dintre o probabilitate pe E şi una de trecere de la E la F se defineşte astfel:

Definiţia 6.2.5. *Fie (E,E) şi (F,F) două spaţii măsurabile, Π o probabilitate pe E şi Q o probabilitate de trecere de la E la F. Atunci $\Pi\otimes Q$ este o probabilitate pe* $(E\times F, \mathcal{E}\otimes\mathcal{F})$ *definită prin*

$$\Pi\otimes Q(C) = \int Q_x(C(x,.))d\Pi(x) \qquad (6.2.4)$$

unde $C(x,.) = \{y\in F \mid (x,y)\in C\}$. Dacă observăm că $1_{C(x,.)(y)} = 1_C(x,y)$, atunci putem scrie formula (2.1.4) ca

$$\Pi\otimes Q(C) = \int\int 1_C(x,y.)dQ_x(y)d\Pi(x). \qquad (6.2.5)$$

Astfel obţinem o formulă de integrare faţă de probabilitatea produs.

Definirea în acest mod a produsului este motivată de următorul rezultat care justifică algoritmul general

Propoziţia 6.2.6. *Fie Z = (X,Y) un vector aleator cu valori în spaţiul măsurabil $(E\times F, \mathcal{E}\otimes\mathcal{F})$. Dacă Π este repartiţia lui X şi Q este repartiţia lui Y condiţionată de X, atunci repartiţia lui Z este $\Pi\otimes Q$.*

Demonstraţie

Fie $f:E\times F\to\mathfrak{R}$ o funcţie măsurabilă. Avem $\int f(x,y)d\Pi\otimes Q(x,y) = \int(\int f(x,y)dQ_x(y))d\Pi(x) = \int h(x)d\Pi(x)$ unde $h(x) = \int f(x,y)dQ_x(y)$.

Cum Π este repartiția lui X, formula de transport spune că $\int h(x)d\Pi(x) = Eh(X)$. Dar $h(X) = \int f(X,y)dQ_X(y) = E[f(X,Y) \mid X]$ (conform cu (2.1.3), deci $Eh(X) = E[E[f(X,Y) \mid X]] = Ef(X,Y)$ (proprietatea de iterativitate). Aşadar $Ef(X,Y) = \int f(x,y)d\Pi \otimes Q(x,y)$. Cum egalitatea este valabilă pentru orice f măsurabilă şi mărginită, rezultă că repartiția lui Z = (X,Y) este $\Pi \otimes Q$.

Scrierea statistică a acestei propoziții este

$$F_{(X,Y)} = F_X \otimes F_{(Y|X)} \qquad (6.2.6)$$

Principiul este

„înmulțim repartiția lui *X* cu repartiția lui *Y* condiționată de X”.

Avantajul este că formula se poate generaliza

$$F(X,Y,Z) = F_X \otimes F_{(Y|X)} \otimes F_{(Z|XY)} \qquad (6.2.7)$$

În general am avea

$$F_{(X_1,X_2,...,X_n)} = F_{X_1} \otimes F_{(X_3|X_1,X_2)} \otimes ... \otimes F_{(X_n|X_{1,...,}X_{n-1})} \qquad (6.2.8)$$

Există două situații în care aplicarea algoritmului nu pune probleme:

Cazul discret

Să se simuleze un vector aleator $Z \sim \begin{pmatrix} z_1 & z_2 & z_3 & ... & z_n \\ p_1 & p_2 & p_3 & ... & p_n \end{pmatrix}$ unde $z_j = (z_{j,1},...,z_{j,d})$ sunt vectori d – dimensionali. Atunci formula (6.2.8) revine la

$$P(X_1=x_1,X_2=x_2,...,X_d=x_d) = P(X_1=x_1)P(X_2=x_2 \mid X_1=x_1)...P(X_d=x_d \mid X_1=x_1,...,X_{d-1}=x_{d-1}) \qquad (6.2.9)$$

Cazul absolut continuu. Acum presupunem că vectorul d - dimensional Z are densitatea f_Z. Atunci se verifică imediat că şi repartițiile $F_{(X_k|X_1,...,X_{k-1})}$ au densități, notate $f_{(X_k|X_1,...,X_{k-1})}$ care se calculează astfel: mai întîi calculăm densitățile celor d vectori de dimensiune $k \le d$. Este imediat că $f_{(X_1,...,X_k)}(x_1,...,x_k) = \int_{\Re^{d-k}} f_Z(x_1,...,x_k,x_{k+1},...,x_d)dx_{k+1}dx_{k+2}...dx_d$. Apoi punem

- $f_{X_1}(x_1) = \int_{\Re^{d-1}} f(x_1,x_2,...,x_d)dx_2dx_3...dx_d$
- $f_{X_2|X_1}(x_2) = \dfrac{f_{(X_1,X_2)}(X_1,x_2)}{f_{X_1}(X_1)}$
- $f_{X_3|X_1,X_2}(x_3) = \dfrac{f_{(X_1,X_2,X_3)}(X_1,X_2,x_3)}{f_{(X_1,X_2)}(X_1,X_2)}$
- .....

Dacă folosim notațiile statistice formulele devin mai uşor de înțeles

$$f_{X_1}(x_1)=\int_{\Re^{d-1}} f(x_1,x_2,...,x_d)dx_2dx_3...dx_d \ , \ f_{X_2|X_1=x_1}(x_2)=\frac{f_{(X_1,X_2)}(x_1,x_2)}{f_{X_1}(x_1)}$$

$$f_{X_3|X_1=x_1,X_2=x_2}(x_3)=\frac{f_{(X_1,X_2,X_3)}(x_1,x_2,x_3)}{f_{(X_1,X_2)}(x_1,x_2)},... \qquad (6.2.10)$$

Exemplul 6.2.7. *Să se simuleze un vector X repartizat Multinomial (100; 0.1,0.2,0.7).*

Soluție

Repartiția multinomială Multinomial (n ; $p_1,p_2,p_3,...,p_k$) este o repartiție k – dimensională discretă definită prin densitățile discrete

$$p(i_1,...,i_k) = \frac{n!}{i_1!i_2!...i_k!} p_1^{i_1} p_2^{i_2}...p_k^{i_k}$$

Aici $(i_1,...,i_k) \in \{0,1,...,n\}^k$. În cazul acesta, discret, putem să renunțăm la teorie şi să simulăm vectorul nostru Z ca pe orice variabilă aleatoare discretă: codificăm cumva vectorul $(i_1,...,i_k)$ – de exemplu îl gîndim că ar fi un număr scris în baza $(n+1)$ – şi apoi aplicăm metoda cuantilei. În cazul particular de mai sus, am avea de a face cu o variabilă aleatoare cu $(100+1)^3$ componente. Se poate, dar **nu vă sfătuiesc**.

Dar, dacă ținem seama de interpretarea probabilistică a acestei repartiții (se extrag cu revenire n bile dintr-o urnă cu bile de k culori diferite, urnă în care p_j este proporția bilelor de culoare „j" !) se verifică imediat că

$$X_1 \sim \text{Binomial}(n,p_1),\ (X_2 \mid X_1) \sim \text{Binomial}(n - X_1, \frac{p_2}{p_2+...+p_k}),$$

$$(X_3 \mid X_1, X_2) \sim \text{Binomial}(n - X_1 - X_2, \frac{p_3}{p_3+...+p_k}),\ etc$$

În cazul nostru concret X1 ~ Binomial(100, 0.1); $(X_2 \mid X_1) \sim$ Binomial(100 – X_1, 0.2/0.9) iar $X_3 = 100 - X_1 - X_2$.

Secvența de instrucțiuni care simulează în „R" N = 10 asemenea vectori este

```
n=100;p1=.1;p2=.2/(1-p1);N=10
x1=rbinom(N,n,p1);x2=rbinom(N,n-x1,p2);x3=n-x1-
x2;x=cbind(x1,x2,x3);x
      x1 x2 x3
 [1,]  8 23 69
 [2,]  5 23 72
 [3,]  8 20 72
 [4,] 14 24 62
 [5,] 13 23 64
 [6,]  8 16 76
 [7,]  4 32 64
 [8,]  8 25 67
 [9,] 10 21 69
[10,] 10 21 69
```

Exemplul 6.2.8. *Să se simuleze un vector X ~ Uniform(C) unde C este sfera tridimensională unitară de rază 1:* $C = \{(x,y,z) \in \Re^3 \mid x^2 + y^2 + z^3 \le 1\}$

Soluție

Amintim că un vector X se numeşte repartizat uniform într-un compact din $\Re^k$ de volum nenul dacă $P(X \in B) = \lambda^k(B \cap C) / \lambda^k(C)$. O definiție alternativă este că densitatea sa $f_X = \alpha 1_C$ unde $\alpha = 1/ \lambda^k(C)$. Aici λ^k este măsura Lebesgue k – dimensională.

În cazul nostru $f_X(x,y,z) = \frac{3}{4\pi} 1_C(x,y,z)$ - deoarece volumul sferei este $4\pi \cdot 1^3/ 3$.

Folosim formulele *(6.2.10.)* Mai întîi, $f_{X_1}(x) = \frac{3}{4\pi} \int 1_C(x,y,z) dydz$. Integrala aceasta reprezintă aria secțiunii prin sferă făcută prin punctul (x,0,0); secțiunea are formă de cerc cu raza r = $\sqrt{1-x^2}$, deci aria este $\pi(1- x^2)$, deci $f_{X_1}(x) = \frac{3(1-x^2)}{4}$. *Apoi,* $f_{(X_1,X_2)}(x,y) = \frac{3}{4\pi} \int 1_C(x,y,z) dz$; integrala este lungimea secțiunii făcute în sferă (x,y,0) , care este intervalul [$-\sqrt{1-x^2-y^2}$ $\sqrt{1-x^2-y^2}$], deci avem $f_{(X_1,X_2)}(x,y) = \frac{3\sqrt{1-x^2-y^2}}{2\pi} 1_{\{x^2+y^2 \le 1\}}$. Deci avem în concluzie

$$f_{X_1}(x) = \frac{3(1-x^2)}{4}$$

$$f_{X_2|X_1=x}(y) = \frac{2}{\pi} \frac{\sqrt{1-x^2-y^2}}{1-x^2} 1_{\left[-\sqrt{1-x^2}, \sqrt{1-x^2}\right]}(y)$$

$$f_{X_3|X_1=x, X_2=y}(z) = \frac{2(1-x^2)}{3\sqrt{1-x^2-y^2}} 1_{\left[-\sqrt{1-x^2-y^2}, \sqrt{1-x^2-y^2}\right]}(z)$$

Principial ar trebui să simulăm pe X_1 după prima densitatea. Cu valoarea lui X_1 = x astfel obținută simulăm pe X_2 după a doua. Găsim $X_2 = y$ şi, cu x şi y astfel găsiți, simulăm pe X_3.

Se poate face, dar necesită timp şi la calculul cuantilelor se pierde mult din precizie. Poate e loc şi de mai bine.*q.e.d.*

Exemplul 6.2.9. *Să se simuleze un vector* $X \sim N\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}4 & 3\\3 & 9\end{pmatrix}\right)$

Soluție

În general, dacă $X \sim N(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix},\begin{pmatrix}\sigma_1^2 & r\sigma_1\sigma_2\\ r\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix})$, atunci se demonstrează uşor, folosind funcția caracteristică sau ce generatoare de momente, că

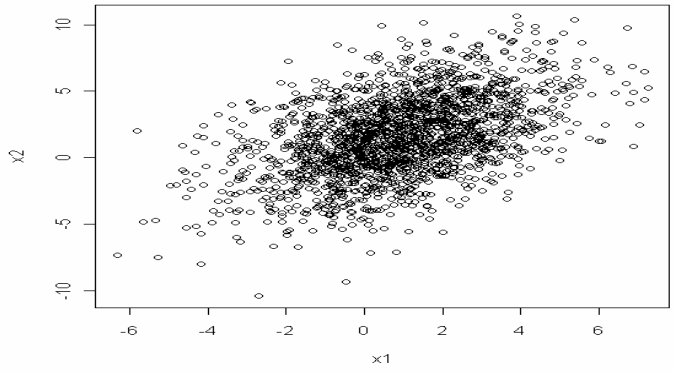
$X_1 \sim N(\mu_1,\sigma_1^2)$ şi $(X_2 \mid X_1) \sim N(\mu_2 + r(X_1-\mu_1)\frac{\sigma_2}{\sigma_1},\left(\sigma_2\sqrt{1-r^2}\right)^2)$

În cazul nostru $\mu_1 = \mu_2 = 1$, $\sigma_1 = 2$, $\sigma_2 = 3$, $r = 0.5$. În „R” este foarte simplu:

```
mu1=1;mu2=1;sig1=2;sig2=3;r=.5;N=10
x1=rnorm(N,mu1,sig1);mu2yx=mu2+r*x1*sig2/sig1
x2=rnorm(N,mu2yx,sig2*sqrt(1-r^2));x2
cbind(x1,x2)
```

```
           x1         x2
 [1,]  1.4348022 -0.4392765
 [2,]  2.9756048  1.3576312
 [3,]  1.4325336  5.1186160
 [4,]  3.6110443  0.2845141
 [5,] -1.6309399 -0.8586991
 [6,] -1.1127450 -3.2264889
 [7,] -1.7916329 -3.2797160
 [8,]  0.2004633  5.0036828
 [9,]  3.4808525  5.0187472
[10,] -0.2603470  2.8354261
```

În figura alăturată vedem rezultatul a 5000 de simulări.



Pentru dimensiuni mai mari, algoritmul general nu mai este satisfăcător nici pentru simularea vectorilor aleatori normali, deoarece formulele de calcul pentru $(X_j \mid X_1,...,X_{j-1})$ devin tot mai complicate.

6.3. Algoritmi speciali: repartiții uniforme

Problema 6.3.1. *Fie $C \subseteq \Re^k$ o mulțime compactă de volum $\lambda^k(C) > 0$. Să se simuleze un vector $X \sim Uniform(C)$.*

Am văzut că algoritmul general pune mari dificultăți chiar în cazurile simple. De exemplu, chiar şi pentru o mulțime simplă, cum ar fi simplexul $C = \Delta_3 = \{(x,y,z) \in [0,1]^3 : x + y + z \le 1\}$,unde densitățile sunt uşor de calculat am avea probleme: $f = 6 1_C$, $f_{X_1}(x) = 3(1-x)^2 1_{(0,1)}(x)$, $f_{X_1,X_2}(x,y) = 6(1-x-y)1_{\{x+y<1\}} \Rightarrow f_{X_2|X_1=x}(y) = \frac{2(1-x-y)}{(1-x)^2} 1_{[1-x,1]}(y)$. La prima densitate e uşor de calculat cuantila, la a doua e greu.

Algoritmul acceptare / respingere. Ideea este să generăm vectori aleatori unifom repartizați într-o mulțime mai mare, unde este comod de făcut aceasta, şi să reținem doar acei vectori care sunt C. Cel mai comod este să includem C într-un hiperparalelipiped $[a_1,b_1]\times...\times[a_k,b_k]$

Formal, rezultatul este următorul:

Propoziția 6.3.2 *Fie $C \subseteq A \subseteq \Re^k$ două compacte de măsură pozitivă şi fie $(X_n)_n$ un şir de vectori aleatori i.i.d. repartizați Uniform(C). Fie $N = \inf\{n : X_n \in C\}$ şi $Z = X_N$.*

Atunci $Z \sim Uniform(C)$.

Demonstrație

Fie $p = P(X_n \in C)$. Atunci $P(N = n) = p(1-p)^{n-1}$, deci probabilitatea ca nu nimerim niciodată în C este 0. Fie $B \subseteq C$ o mulțime boreliană. Atunci

$$P(Z \in B) = P(X_N \in B) = \sum_{n\ge1} P(X_N \in B, N = n) =$$

$$\sum_{n\ge1} P(X_j \notin C \forall j < n, X_n \in B) = \frac{P(X_n \in B)}{P(X_n) \in C)} \sum_{n\ge1} P(X_j \notin C \forall j < n, X_n \in C) =$$

$$\frac{P(X_n \in B)}{P(X_n) \in C)} \sum_{n\ge1} p(1-p)^{n-1} = \frac{P(X_n \in B)}{P(X_n) \in C)} = \frac{\lambda^k(B)/\lambda^k(A)}{\lambda^k(C)/\lambda^k(A)} = \frac{\lambda^k(B)}{\lambda^k(C)}$$, exact ce voiam.

Viteza algoritmului depinde, evident, de raportul dintre volumul lui C şi volumul lui A. În cazul $C = \Delta_3$, putem lua $A = [0,1]^3$ – mai bine nici nu se poate. Cum raportul dintre volume este 1/6, ne aşteptăm ca în jur de o şesime din punctele generate Uniform(A) să fie şi în C. Chiar aşa şi este: după 120 de simulări, am reuşit să nimerim de 23 de ori în A.

```
N=120;x1=runif(N);x2=runif(N);x3=runif(N);s=x1+x2+x3
v=which(s<1);nf=length(v)
xb1=1:nf;xb2=1:nf;xb3=1:nf
for              (i              in              1:nf)
{xb1[i]=x1[v[i]];xb2[i]=x2[v[i]];xb3[i]=x3[v[i]]}
x=cbind(xb1,xb2,xb3);x
```

```
               xb1         xb2        xb3
 [1,] 0.261076623 0.001725541 0.59342688
 [2,] 0.329976494 0.618442961 0.02125208
 [3,] 0.070146402 0.406062445 0.41304084
 [4,] 0.551778257 0.067009293 0.28671116
 [5,] 0.229001845 0.150103106 0.29779559
 [6,] 0.145313528 0.550778716 0.19904163
 [7,] 0.019184817 0.133237032 0.68866607
 [8,] 0.180765220 0.394726553 0.23718938
 [9,] 0.619285529 0.085581634 0.19468893
[10,] 0.012022830 0.403672086 0.19543799
[11,] 0.420006088 0.054411151 0.36632268
[12,] 0.042866380 0.024161682 0.77753557
[13,] 0.394432793 0.135697418 0.09784523
[14,] 0.342567456 0.325553098 0.11622695
[15,] 0.439633231 0.135284625 0.04868877
[16,] 0.002924559 0.660988999 0.21707588
[17,] 0.442618400 0.198682676 0.14334936
[18,] 0.055533753 0.397702978 0.21110361
[19,] 0.102202073 0.703533423 0.04253352
[20,] 0.101106154 0.433682927 0.38915822
[21,] 0.458446015 0.205739238 0.10033640
[22,] 0.067286825 0.133509017 0.29028897
[23,] 0.229641613 0.620405543 0.06375227
```

Dacă însă k, dimensiunea simplexului, în loc să fie 3 era 7, atunci volumul său era 1/7! = 1/5040. Din 120 de simulări era foarte probabil ca nici una să nu fie în mulțimea dorită.

Avem şi o veste bună: dacă C este un simplex, atunci există algoritmi rapizi care generează vectori repartizați uniform acolo. El se bazează pe următoarea descoperire

Propoziția 6.3.3. *Simularea rapidă a vectorilor* ***repartizați uniform într-un simplex****. Fie n+1 variabile aleatoare i.i.d.* $(X_j)_{1 \le j \le n+1}$ *repartizate Exponential(1). Fie S suma lor şi* $Y_j = \frac{X_j}{S}$. *Atunci vectorul* $Y = (Y_j)_{1 \le j \le n}$ *este repartizat Uniform*(Δ_n) *unde* $\Delta_n = \{\mathbf{x} \in [0,1]^n : x_1 + x_2 + ... + x_n \le 1\}$.

Demonstrație

Fie $f : \Re^n \to \Re$ o funcție măsurabilă şi mărginită. Avem de calculat Ef(Y) în speranța că rezultatul va fi $n! \int_{\Delta_n} f d\lambda^n$. Din formula de transport, avem

$$Ef(Y) = \int f\left(\frac{x_1}{s}, \frac{x_2}{s}, ..., \frac{x_n}{s}\right) e^{-s} 1_{[0,\infty)^{n+1}}(x_1, ..., x_n, x_{n+1}) dx_1 dx_2 .... dx_{n+1}$$. Aici s este suma $x_1 + ... + x_{n+1}$

Facem schimbarea de variabile

$y_1 = x_1/s$, $y_2 = x_2/s$,..., $y_n = x_n/s$, $y_{n+1} = s$. Jacobianul ei este $\frac{D(x_1,...,x_{n+1})}{D(y_1,...y_{n+1})} = s^n$,

imaginea mulțimii $[0,\infty)^n \times [0,\infty)$ prin ea este $\Delta_n \times [0,\infty)$ deci integrala devine

$$Ef(Y) = \int f(y_1, y_2, ..., y_n) s^n e^{-s} 1_{\Delta_n}(y_1, ..., y_n) 1_{[0,\infty)}(s) dy_1 dy_2 .... dy_n ds =$$

$$\int f(y_1, y_2, ..., y_n) 1_{\Delta_n}(y_1, ..., y_n) dy_1 dy_2 .... dy_n \int_0^\infty s^n e^{-s} ds$$, adică exact ce doream – căci a doua integrală este n!. *q.e.d.*

În general o mulțime se numeşte simplex n dimensional dacă este anvelopa convexă a unei mulțimi de n+1 puncte şi, dacă, în plus, are interiorul nevid. De exemplu, pentru n = 2, orice triunghi ABC este un simplex, cu condiția ca nu cumva cele trei puncte să fie coliniare. În spațiu, orice tetraedru ABCD este simplex dacă nu cumva cele patru puncte sunt coplanare.
Frumusețea este că simplexul S = S(a) cu vîrfurile $a_1a_2 \ldots a_na_{n+1}$ se poate descrie întotdeauna sub forma

$$S = \{a_1X_1 + a_2X_2 + \ldots + a_nX_n + a_{n+1}(1 - X_1 - X_2 - \ldots - X_n) : (X_1, X_2, \ldots, X_n \in \Delta_n\} \quad (6.3.1)$$

Şi de aici se vede ce avem de făcut pentru a simula un vector uniform repartizat acolo: simulăm vectori aleatori repartizați unifom în Δ_n.

Exemplul 6.3.4. *Să se simuleze un vector aleator repartizat U(C) unde C este triunghiul ABC*

Soluție. Fie a,b,c afixele celor trei puncte. Atunci

$$X = \frac{aX_1 + bX_2 + cX_3}{X_1 + X_2 + X_3}$$ *unde X_j sunt i.i.d şi ~ Exponential(1).*

Mai există un caz în care suntem norocoşi: dacă X ~ N(μ,C). Atunci folosim următorul truc:

Propoziția 6.3.5. *Fie C o matrice pozitiv definită şi A o matrice simetrică cu proprietatea că* $A^2 = C$. *Fie* $Y \sim N(0,\mathbf{I}_k)$ *(adică un vector normal standard – cu toate componentele i.i.d ~ N(0,1)). Atunci* $X = AY + \mu$ *este un vector repartizat* $N(\mu,C)$.

Demonstrație
Evident. X este normal , $EX = \mu$ şi cov(X) = cov(AY + μ) = Acov(Y)A' = AA' = A^2. *q.e.d.*

Ca să construim construim pe A scriem matricea C la forma diagonala C = ODO' unde O este matricea vectorilor proprii, care sunt ortogonali, D este matricea diagonală care cu valorile proprii, înlocuim D cu $\sqrt{D}$ - adică cu matricea care pe diagonală are radăcinile pătrate ale valorilor proprii şi punem A = O$\sqrt{D}$ O'.

Dacă avem un software care e în stare să calculeze vectorii proprii, nu e nici o problemă. Iată, de exemplu, o funcție în „R" exact acest lucru

```
     #simularea unei repartitii normale k-dimensionale
     normal<-function(mu,cov)
     {k=length(mu);jor=eigen(cov);valp=jor[[1]]      #
valp sunt valorile proprii
     vecp=jor[[2]]  # vecp are vectorii proprii
     kk=k^2;diag=1:kk;diag=diag-diag;dim(diag)=c(k,k)
     for (i in 1:k) {diag[i,i]=sqrt(valp[i])}
     a=vecp%*%diag%*%t(vecp)
     u=rnorm(k,0,1);x=a%*%u+mu
     x}
```

Funcția se apelează prin comanda

```
x=normal(mu,c)
```

Pentru a o apela este nevoie de vectorul mu = μ al mediilor, si de matricea de covarianță C. Iată un exemplu:simulăm 10 vectori aleatori N(μ,*C*) cu

$\mu=(1,2,3)$ şi $C = \begin{pmatrix} 4 & 1 & -1 \\ 1 & 9 & 2 \\ -1 & 2 & 16 \end{pmatrix}$.

```
xx= 1:30; dim(xx) = c(10,3=
for (i in 1:10) {xx[i,]=normal(mu,c)}; xx
```

```
            [,1]       [,2]        [,3]
 [1,]  1.3356960  2.7749180 15.06867494
 [2,]  2.3000149  1.3544057 -1.01883329
 [3,] -1.3323644  0.8461128 10.02604282
 [4,]  0.9394982  0.8378104  7.92561663
 [5,]  2.0869182  0.7712231 -7.01464171
 [6,]  0.2167368  1.6627487  0.05478921
 [7,]  0.7415051  3.6121115  3.01237874
 [8,]  0.5663550  1.6125076 -1.61308129
 [9,]  0.6628614 -1.1853937 12.85524488
[10,]  5.2891190  0.4922314  0.82281214
```

Capitolul 7

Statistică descriptivă

Introducere

Statistica descriptivă este ramura statisticii ce se ocupă cu prezentarea, organizarea şi interpretarea unei colectii de date. Descrierea acestor informaţii se poate face grafic (prin liste, grafice liniare, de distribuţie etc.), sau prin indicatori statistici (medie, mediană, abatere etc.).

7.1. Prezentarea datelor statistice

Analiza statistică a unui fenomen începe cu statistica formală (culegerea datelor asupra fenomenului respectiv şi înregistrarea datelor). Datele sunt apoi analizate şi interpretate, cu ajutorul statisticii matematice.

Definiţia 7.1.1. *Prin populaţie statistică (populaţie) se înţelege orice mulţime care formează obiectul unei analize statistice. Elementele unei populaţii statistice se numesc unităţi statistice sau indivizi.*

Caracteristica este trăsătura comună unităţilor unei populaţii statistice. Valoarea numerică a caracteristicii se numeşte variabilă aleatoare. De exemplu, dacă ne referim la repartiţia componenţilor unei echipe de fotbal, după înălţime, constatăm că mulţimea sportivilor formează populaţia statistică, fiecare fotbalist este o unitate statistică şi înalţimea este caracteristica studiată.

Matematic, o populaţie statistică este o partiţie a unei mulţimi E, $E=\{A_1, \dots, A_n\}$, submulţimile $A_1, \dots, A_n$ fiind clase. Unităţile statistice care compun o clasă A_i sunt alese pe baza unei relaţii de echivalenţă, care reprezintă caracteristica populaţiei.

Caracteristicile pot fi calitative sau cantitative. Caracteristicile cantitative pot fi măsurate folosind numere reale. Integrarea datelor cantitative în text are anumite avantaje, dar tabelele statistice permit realizrea unor comparaţii.

În tabelul 7.1.1, avem informaţiile privind durata medie a vieţii în România, in perioada 1998- 2007 (conform Institutului Naţional de Statistică), prezentate sub forma de tabel, evidenţiindu-se, astfel, aspectele importante ale datelor. Obesrvăm astfel că aceasta valoare creşte, începând cu anul 2003, după o scădere nesemnificativă.

Anul	Durata medie de viata
1998	69,24
1999	69,74
2000	70,53
2001	71,19
2002	71,18
2003	71,01
2004	71,32
2005	71.76
2006	72,22
2007	72,61

Tabelul 7.1.1

Reprezentarea grafică realizată pentru studierea schimbărilor sau pentru compararea variabilelor statistice se numeşte grafic. Există mai multe astfel de reprezentări.

Reprezentarea cu batoane foloseşte batoane verticale sau orizontale, a căror lungime simbolizează valorile variabilei statistice. Batoanele verticale se folosesc, de obicei, pentru caracteristici care variază în timp. Între batoanele consecutive se lasă, de regulă, un spaţiu de jumătate de unitate.

Figura 7.1.1 este reprezentarea cu batoane pentru datele din tabelul 7.1.1.

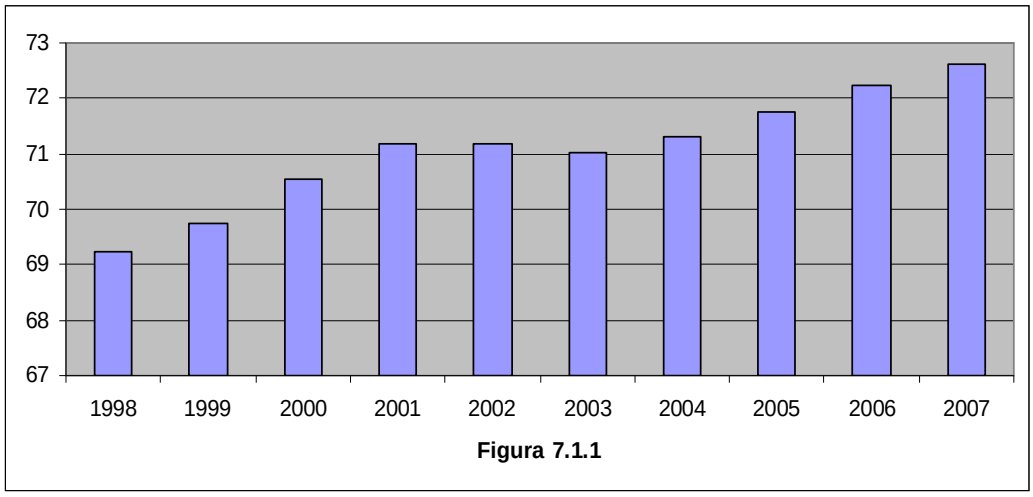


Figura 7.1.1

Reprezentarea cu batoane orizontale prezintă variante adaptate, de exemplu reprezentarea pe componente. Diagrama cu batoane grupate furnizează o metodă de prezentare a părților componente ale unui întreg, fără realizarea unei comparații cu întregul.

Dacă ne referim la datele din Tabelul 7.1.2, privind structura populatiei pe medii (urban, rural), data furnizate de

Tabelul 7.1.2

Anul	Urban(%)	Rural(%)
1960	32,1	67,9
1970	36,9	63,1
1980	45,8	54,2
1990	54,3	45,7
2000	54,6	45,4
2001	54,6	45,4
2002	53,3	46,7
2003	53,4	46,6
2004	54,9	45,1
2005	54,9	45,1
2006	55,2	44,8
2007	55,1	44,9
2008	55,0	45,0

Institutul Național de Statistică, se obține reprezentarea cu batoane orizontale pe componente, din figura 7.1.2.

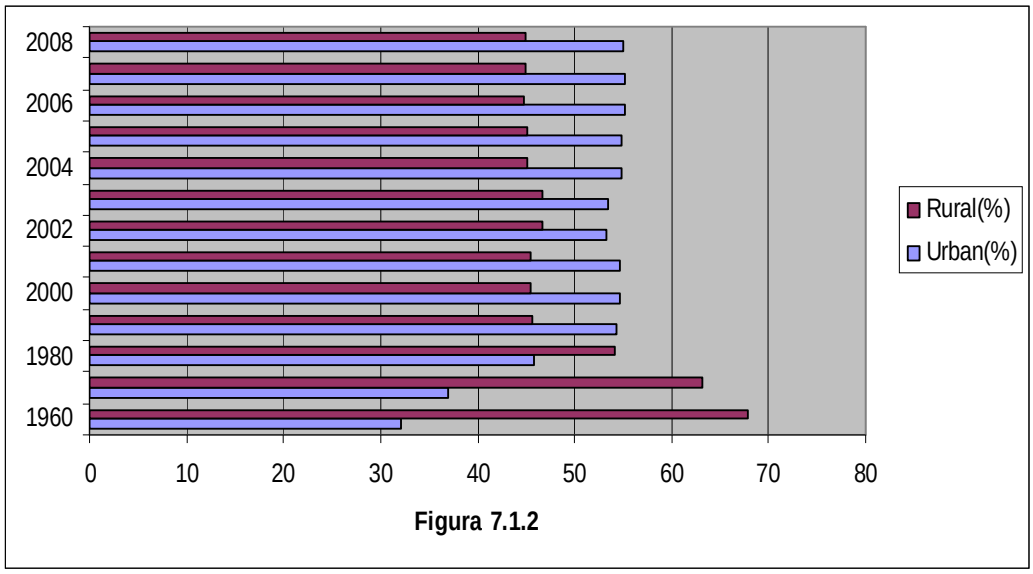


Figura 7.1.2

Graficul liniar pe porțiuni este format din segmente de dreaptă ce se obțin prin unirea perechilor de valori corespunzătoare ale unei perechi de variabile diferite.

În tabelul 7.1.3, sunt prezentate datele furnizate de Institutul Național de Statistică privind totalul numărului de imigranți, în perioada 2003-2008.

Anul	Total imigranți
2003	3267
2004	2987
2005	3704
2006	7714
2007	9575
2008	10030

Tabelul 7.1.3

Pentru acest tabel, am realizat graficul liniar pe porțiuni din figura 7.1.3.

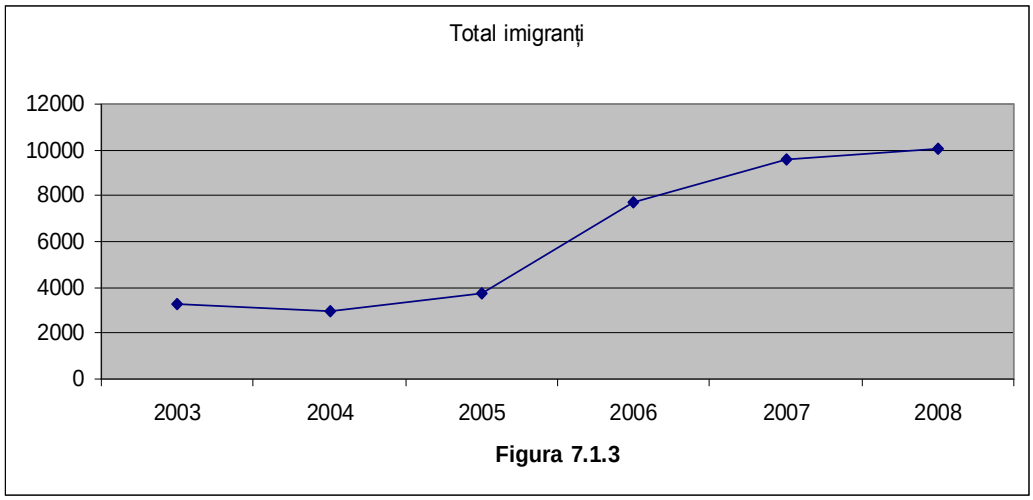


Figura 7.1.3

Diagrama circulară arată descompunerea unui întreg în părțile sale componente. Ele se exprimă ca procente din total şi sunt reprezentate prin segmente de cerc, unghiurile la centru având măsuri egale cu procentul corespunzător din 360^0**.**

Figura 7.1.4 arată structura cheltuielilor din domeniul cercetare-dezvoltare, din punctul de vedere al surselor de finanțare, în România, în 2001.

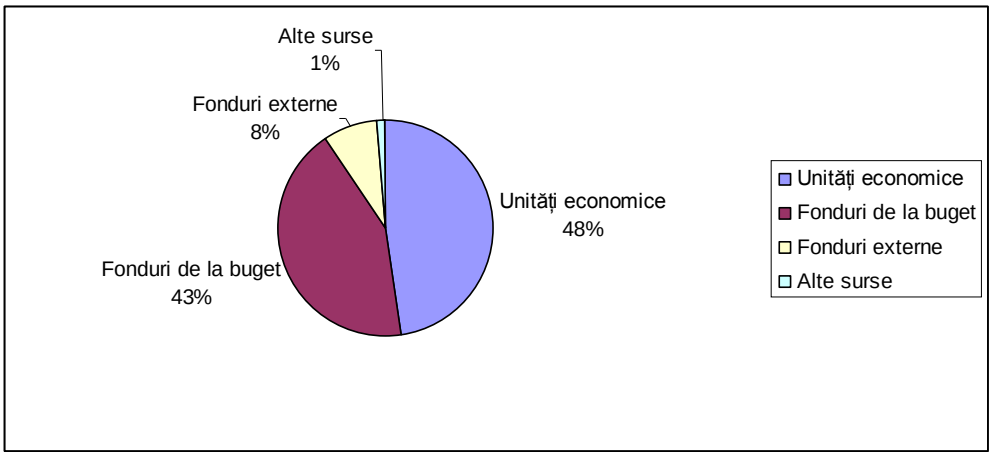


Figura 7.1.4

În continuare, ne vom referi la distribuţii şi reprezentarea lor prin diagrame şi tabele.

Definiţia 7.1.2. *O variabilă statistică se numeşte discretă dacă ea nu poate lua decât valori izolate în intervalul său de variaţie. Ea se numeşte continuă dacă poate lua toate valorile posibile în intervalul său de variaţie.*

Ca exemplu de variabilă discretă, ne putem referi la numărul capitolelor unei cărţi, numărul articolelor produse de o fabrică etc.. Pentru cazul continuu, putem da ca exemplu înălţimea unei persoane, ora sosirii unui tren etc..

Ne referim în continuare la cazul variabilei continue.

Să considerăm un eşantion de 40 de angajaţi al căror salariu brut exprimat în mii lei, la începutul lunii ianuarie, conduce la datele din tabelul 7.1.4.

0,831	0,904	0,896	0,961	0,981
0,956	1,705	1,591	1,156	1,221
1,587	0,991	1,981	1,459	1,861
0,82	1,141	1,452	1,344	1,42
1,805	1,052	1,731	1,75	0,976
1,091	1,201	1,895	0,972	1,071
1,605	0,989	1,858	1,081	1,492
1,594	1,354	1,946	1,671	1,057

Tabelul 7.1.4

O descriere precisă a seriei statistice obţinute se realizează prin construirea unui tabel al frecvenţelor, în care observaţiile sunt clasificate în raport cu numărul unităţilor statistice care se află între anumite limite. Tabelul 7.1.5 prezintă frecvenţele pentru datele anterioare, privind salariile. Astfel, marginile claselor de valori (0,8; 0.95 ...) din tabelul 7.1.5 sunt limitele sau marginile clasei de valori. Media aritmetică a limitelor unei clase se numeşte

mijlocul sau valoarea centrală a clasei. Diferența dintre cea mai mare şi cea mai mică margine se numeşte domeniu sau amplitudine. Frecvența absolută este dată

Limitele clasei	Mijlocul clasei	Frecvența absolută	Frecvența relativă(%)	Frecvența cumulată absolută	Frecvența cumulată relativă(%)
[0,8;0,95)	0,875	4	10	4	10
[0,95;1,1)	1,025	12	30	16	40
[1,1;1,25)	1,175	5	12,5	21	52,5
[1,25;1,4)	1,325	2	5	23	57,5
[1,4;1,55)	1,475	5	12,5	28	70
[1,55;1,7)	1,625	5	12,5	33	82,5
[1,7;1,85)	1,775	4	10	37	92,5
[1,85;2)	1,925	3	7,5	40	100

Tabelul 7.1.5

de numărul unităților statistice aflate între limitele unei clase, iar cea relativă este raportul dintre frecvența absolută şi numărul total al unităților statistice. În cazul în care nu este precizat, prin frecvență se înțelege frecvență relativă. Mulțimea frecvențelor (absolute sau relative), împreună cu clasele lor formează frecvența distribuției. Frecvența cumulată a unei clase este suma frecvențelor până la clasa respectivă, clasele fiind ordonate crescător.

În general, este indicată utilizarea a 10-20 clase de valori.

Histograma este o reprezentare cu batoane, fără spațiu între acestea. Ea prezintă marginile claselor pe axa orizontală şi frecvențele pe cea verticală.

Histograma pentru datele din tabelul 7.1.4 este prezentată in figura 7.1.5.

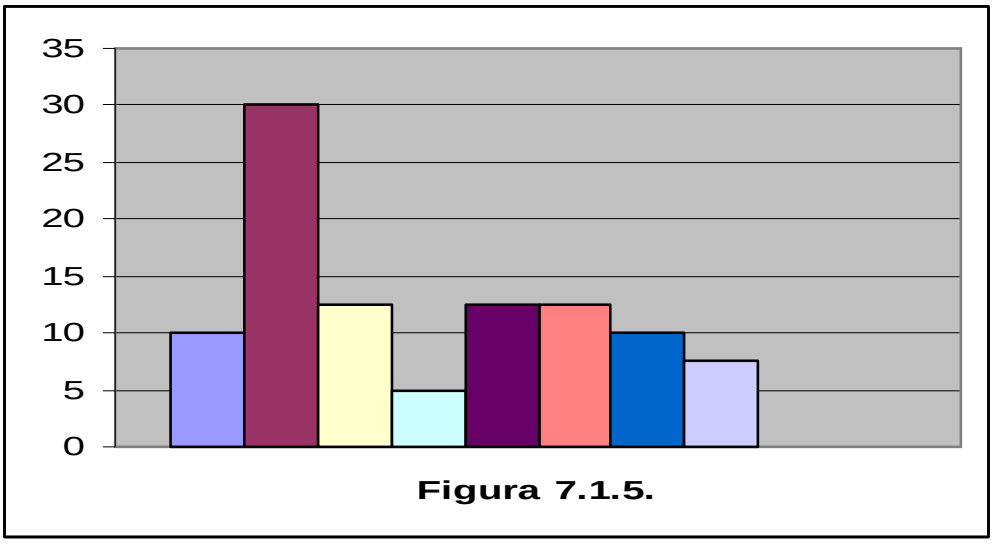


Figura 7.1.5.

Poligonul frecvențelor este un grafic liniar pe porțiuni, mijloacele claselor fiind reprezentate pe axa orizontală şi frecvențele pe cea verticală. Fiecare mijloc are o frecvență, marcată printr-un punct. Punctele consecutive se unesc prin segmente de dreaptă, rezultând poligonul frecvențelor.

Figura 7.1.6 prezintă poligonul frecvențelor pentru datele din tabelul 7.1.4 (pe axa absciselor s-au folosit rotunjiri, pentru a da doar două zecimale).

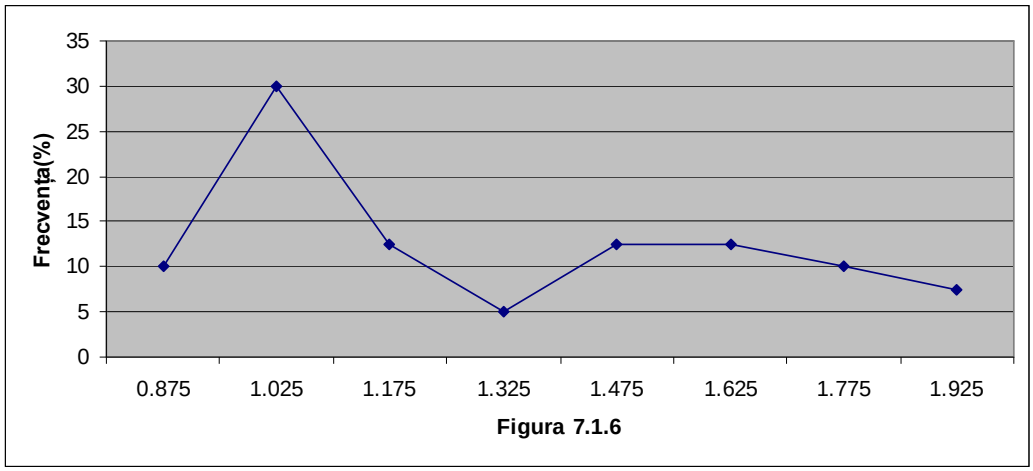


Figura 7.1.6

Poligonul frecvențelor cumulate este un grafic liniar pe porțiuni, care se realizează similar cu poligonul frecvențelor, singura schimbare fiind aceea ca în locul frecvențelor apar frecvențele cumulate.

7.2. Caracteristici numerice

Ne vom referi acum la descrierea informațiilor folosind indicatori statistici. În acest sens, există două mari categorii: măsuri ale tendinței centrale (media, mediana, moda etc.) şi măsuri ale variației sau împrăştierii (amplitudinea, abaterea etc.).

În continuare, prezentăm principalii indicatori ai tendinței centrale.

Într-o distribuție (ne referim la variabilă continuă), clasa cu cea mai mare valoare a frecvenței este clasa modală, iar mijlocul acesteia este moda variabilei.

În tabelul 7.1.5, clasa modală este [0,95;1,1), iar moda este 1,025.

Să ne referim acum la o mulțime de date de selecție (variabilă discretă), moda este valoarea cu frecvența maximă.

Să considerăm o grupă formată din 20 de studenți care susțin un test la matematică, obținându-se rezultatele din tabelul 8.2.1. Aici, moda este 8.

Definiția 7.2.1. *Pentru cazul discret, mediana unei mulțimi* $x_1, x_2, \ldots, x_m$ *(datele de selecție sunt ordonate crescător) este valoarea de mijloc,* $x_{(m+1)/2}$*, dacă m este impar, şi media celor două valori de mijloc,* $\frac{1}{2}(x_{m/2} + x_{m/2+1})$*, dacă m este par.*

De exemplu, media mulțimii 5,6,8,9,12 este 8, iar pentru 15, 18, 20, 14, 28, 30 se obține mediana $\frac{1}{2}(20+24)=22$.

Nota	Frecvența absolută	Frecvența relativă(%)
3	1	5
4	1	5
5	2	10
6	2	10
7	5	25
8	6	30
9	2	10
10	1	5

Tabelul 7.2.1

Definiția 7.2.2. *Pentru o variabilă continuă, clasa cu frecvența cea mai mică ce are proprietatea că frecvența cumulată asociată este mai mare decât* $\frac{m}{2}$, *m fiind numărul total de clase, se numeşte clasa medianei.*

Notând cu *m* numărul total de clase, m_d mediana distribuției, f_i frecvența pentru clasa $[x_{i-1},x_i)$, F_i frecvența cumulată pentru clasa $[x_{i-1},x_i)$ şi $[x_{j-1},x_j)$ clasa modală. Efectuând o interpolare, obținem următoarea

Definiția 7.2.3. *Într-o distribuție, valoarea medianei este dată de relația*

$$m_d=x_i+h_i\frac{0,5-F_{i-1}}{f_i},$$

unde $h_i=x_i-x_{i-1}$, $F_{i-1}<0,5$ *;i* $F_i>0,5$.

Pentru datele din tabelul 8.1.5, clasa mediană este [1,1;1,25), iar mediana este $m_d=1,1+1,5\frac{0,5-0,4}{0,125}=1,122$.

Definiția 7.2.4. *Media(de selecție) a unei mulțimi* $x_1, x_2, ..., x_m$ *se defineşte prin*

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i .$$

De exemplu, dacă lista de prețuri, în lei, pentru centrale termice, este următoarea: 9900, 10300, 11200, 12500, 7600, 17500, costul mediu al unei centrale este

$$\bar{x} = \frac{1}{6}(9\,900 + 10\,300 + 11\,200 + 12\,500 + 7\,600 + 17\,500) = 11500 .$$

Dacă $x_1, x_2, ..., x_k$ sunt valorile distincte ale lui X, iar n_i este frecvența lui x_i, formula se rescrie

$$\bar{x} = \frac{\sum_{i=i}^{k} n_i x_i}{\sum_{i=i}^{k} n_i} .$$

Notând $f_i = \frac{n_i}{m}$, rezultă $\bar{x} = \sum_{i=1}^{k} f_i x_i$, ("media ponderată").

Definiția 7.2.5. *Considerăm un tabel al frecvențelor cu k clase. Dacă* $x_1^*, x_2^*, ..., x_m^*$ *sunt mijloacele claselor,* $n_1, n_2, ..., n_k$ *frecvențele lor absolute şi* $f_1, f_2, ..., f_k$ *frecvențele lor relative, atunci media distribuției este*

$$\bar{x} = \frac{\sum_{i=i}^{k} n_i x^*_i}{\sum_{i=i}^{k} n_i},$$

mai precis

$$\bar{x} = \sum_{i=1}^{k} f_i x_i^* .$$

Pentru datele din tabelul 7.1.5, media este

$$\bar{x} = \frac{4 \cdot 0{,}875 + 12 \cdot 1{,}025 + 5 \cdot 1{,}175 + 2 \cdot 1{,}375 + 5 \cdot 1{,}475 + 5 \cdot 1{,}625 + 4 \cdot 1{,}775 + 3 \cdot 1.925}{4 + 12 + 5 + 2 + 5 + 5 + 4 + 3}$$

$= 1{,}3175.$

Se observă că media nu dă o imagine completă a datelor de selecție sau a distribuției. De exemplu, mulțimile {2, 2, 2, 5, 8, 8, 8}, {3, 3, 5, 5, 5, 7, 7}, {4, 4, 4, 5, 6, 6, 6} au aceeaşi medie, dar au structuri diferite. Acesta este motivul

pentru care sunt introduse măsuri ale variației, care să arate gradul de împrăştiere a datelor în jurul mediei.

Definiția 7.2.6. *Pentru o variabilă discretă, diferența dintre cea mai mare şi cea mai mică valoare a selecției se numeşte amplitudine.*

Pntru o variabilă continuă, amplitudinea este diferența dintre limita superioară a clasei cu cele mai mari margini şi limita inferioară a clasei cu cele mai mici margini.

Definiția 7.2.7. *Fie* $x_1, x_2, ..., x_m$ *date de selecție având media* $\bar{x}$ *. Abaterea medie se defineşte prin relația*

$$a.m. = \frac{1}{m}\sum_{i=1}^{m} | x_i - \bar{x} |.$$

Să considerăm următoarele date de selecție: 12, 15, 13, 20, 13. Media lor este $\bar{x} = \frac{1}{5}(12+15+13+20+13) = 14{,}6$, în timp ce abaterea medie are valoarea

$$a.m. = \frac{1}{5}(|12-14{,}6| + |15-14{,}6| + |13-14{,}6| + |20-14{,}6| + |13-14{,}6|)$$

$= 2{,}32.$

Altfel spus, valorile de selecție diferă în medie cu 2,32 față de media 14,6.

Fie $x_1, x_2, ..., x_k$ valorile distincte ale lui *X*, având media $\bar{x}$, iar n_i frecvența lui x_i . Atunci

$$a.m. = \frac{\sum_{i=i}^{k} n_i \, | x_i - \bar{x} |}{\sum_{i=i}^{k} n_i}.$$

Notând cu $f_i = \frac{n_i}{m}$ frecvența relativă, rezultă $\bar{x} = \sum_{i=1}^{k} f_i \, | x_i - \bar{x} |$.

Pentru datele din tabelul 7.2.1, abaterea medie este $a.m. = 1{,}3$.

Definiția 7.2.8. *Fie o variabilă continuă cu un tabel al frecvențelor cu k clase. Dacă* $x_1^*, x_2^*, ..., x_m^*$ *sunt mijloacele claselor,* $n_1, n_2, ..., n_k$ *frecvențele lor absolute şi* $f_1, f_2, ..., f_k$ *frecvențele lor relative, atunci abaterea medie este*

$$a.m. = \frac{\sum_{i=i}^{k} n_i \mid x^*_i - \bar{x} \mid}{\sum_{i=i}^{k} n_i},$$

adică

$$a.m. = \sum_{i=1}^{k} f_i \mid x_i^* - \bar{x} \mid .$$

Pentru tabelul 7.1.5, calculăm abaterea medie şi obținem valoarea $\frac{119,85}{40} = 2,99625$.

Definiția 7.2.9. *Fie $x_1, x_2, ..., x_m$ date de selecție cu media $\bar{x}$. Dispersia se defineşte astfel*

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2 .$$

$\sigma = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2}$ *este abaterea de selecție (empirică) standard.*

Fie $x_1, x_2, ..., x_k$ valorile distincte ale lui *X*, cu media $\bar{x}$, iar n_i frecvența lui x_i. Formula pentru calculul dispersiei devine

$$\sigma^2 = \frac{\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^2}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^2}{m}.$$

Dacă frecvența relativă este $f_i = \frac{n_i}{m}$, rezultă $\sigma^2 = \sum_{i=1}^{k} f_i\left(x_i - \bar{x}\right)^2$.

Dispersia corespunzătoare datelor din tabelul 7.2.1 este

$$\sigma^2 = \frac{1}{20}\Big[1\cdot(3-7)^2 + 1\cdot(4-7)^2 + 2\cdot(5-7)^2 + 2\cdot(6-7)^2 + 5\cdot(7-7)^2 +$$

$$6\cdot(8-7)^2 + 2\cdot(9-7)^2 + 1\cdot(10-7)^2\Big] = 2,8 ,$$

în timp ce abaterea standard este $\sigma = 1,673$.

Definiția 7.2.10. *Fie o variabilă continuă cu un tabel al frecvențelor cu k clase. Dacă* $x_1^*, x_2^*, \ldots, x_k^*$ *sunt mijloacele claselor,* $n_1, n_2, \ldots, n_k$ *frecvențele lor absolute şi* $f_1, f_2, \ldots, f_k$ *frecvențele lor relative, atunci media distribuției este*

$$\sigma^2 = \frac{\sum_{i=i}^{k} n_i \left(x^*_i - \bar{x}\right)^2}{\sum_{i=i}^{k} n_i},$$

mai precis

$$\sigma = \sum_{i=1}^{k} f_i \left(x_i^* - \bar{x}\right)^2 \quad .$$

Pentru datele din tabelul 7.1.5, dispersia este $\sigma^2 = \frac{445{,}2750}{40} = 1{,}114$, iar abaterea este $\sigma = 1{,}055$.

7.3. Corelație. Regresie

Legătura dintre două sau mai multe variabile poartă numele de corelație. Conexiunea aceasta se poate prezenta sub mai multe forme, cea mai simplă fiind relația $y = f(x)$, unde *f* este o funcție de variabila *x*, egalitate ce arată că lui *x* îi corespunde o valoare bine determinată a lui *y*.

Mai jos sunt prezentate notele la algebră şi geometrie obținute de zece studenți.

Nr. crt.	Nota algebră	Nota geometrie
1	5	5
2	6	5
3	6	6
4	6	7
5	7	7
6	7	8
7	8	8
8	9	9
9	10	9
10	10	10

Tabelul 7.3.1

Vom reorganiza datele de mai sus sub forma unui tabel cu două intrări, astfel: notele la algebră sunt reprezentate pe axa absciselor, iar cele la geometrie pe cea a ordonatelor.

	5	6	7	8	9	10	m_Y
10							9,5
9							9
8							8
7							7,5
6							6
5							5
m_X	5,5	6	6,5	7,5	9,5	10	

Tabelul 7.3.2

Vom indica existența unui student ce ia note corespunzătoare unui pătrat printr-un punct situat în interiorul acestuia. De exemplu, studentul cu notele 7 şi 8 se va regăsi în pătratul (7,8). Acest tablou se numeşte tablou de corelație.

Observăm că, în general, creşterea notelor la algebră este însoțită de creţterea notelor la geometrie. Astfel, între aceste variabile există o corelație pozitivă. Punctele din interiorul tabloului de corelație se grupează în jurul unei diagonale a pătratului, deci se poate afirma că avem o corelație liniară.

În figurile 7.3.1 şi 7.3.2, am reprezentat grafic pe m_X în raport cu x ți pe m_Y în raport cu y.

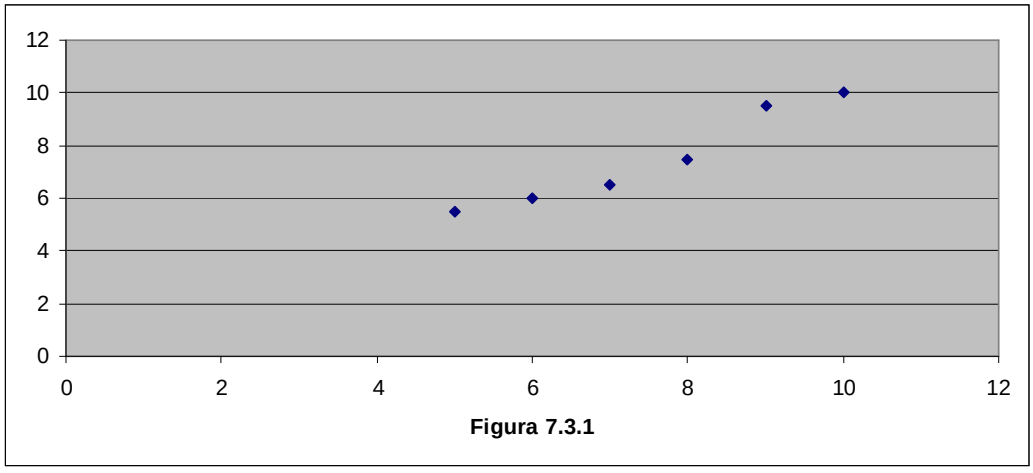


Figura 7.3.1

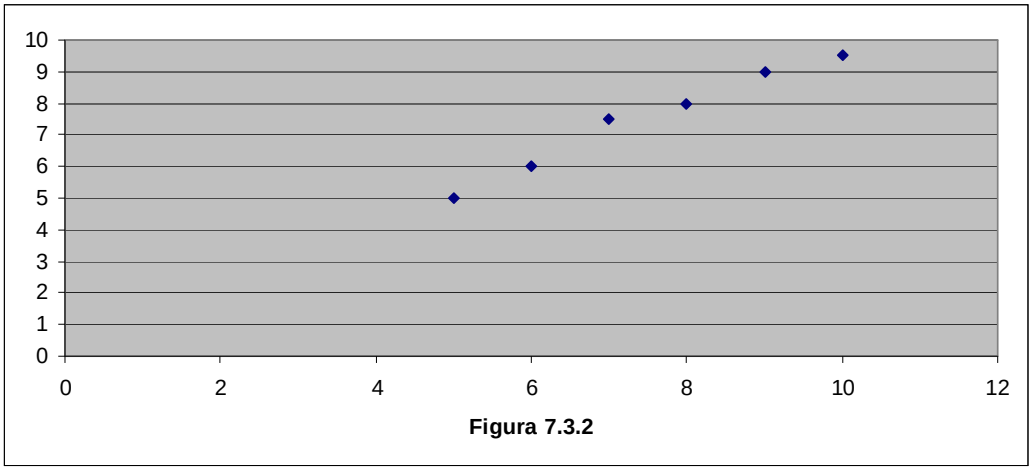


Figura 7.3.2

m_X şi m_Y se numesc funcţii de regresie. În figurile de mai sus, graficele lor se grupează în jurul primei bisectoare. În figura 7.3.1, avem regresia lui *y* asupra lui *x*, iar în figura 7.3.2, regresia lui *x* asupra lui *y*.

Pentru fixarea ideilor, să considerăm *V(X,Y)* un vector în plan, *X* şi *Y* fiind variabile aleatoare. Statistic, suntem interesaţi de situaţia în care vectorul *V* ia un număr finit de valori, $v_{ij} = v_{ij}(a_j, b_k)$, $j = \overline{1,m}$, $k = \overline{1,n}$, vecorul având componentele a_j, b_k. Notând $P(V = v_{jk}) = p_{jk}$, avem $p_{jk} \geq 0$ şi $\sum_{j=1}^{m}\sum_{k=1}^{n} p_{jk} = 1$. Tabelul 7.3.3 se numeşte tabel de corelaţie.

	b_1	b_2		b_n	Y
a_1					p_1
a_2					p_2
...	...	...	...	...	...
...	...	...	...	...	...
a_m	p_{m1}	p_{m2}		p_{mn}	p_m
X	p_1	p_2		p_n	

Tabelul 7.3.3

Au loc următoarele relaţii:

$$\sum_{k=1}^{n} p_{jk} = p_j, \ \ j = \overline{1,m}; \qquad \sum_{j=1}^{m} p_{jk} = p_k, \ \ k = \overline{1,n};$$

$$\sum_{k=1}^{n} p_k = 1, \ \sum_{j=1}^{m} p_j = 1.$$

p_j, respectiv p_k, se numesc probabilități marginale şi reprezintă probabilitatea ca $X = a_j$, respectiv $Y = b_k$, penrtu orice valoare a celeilalte variabile.

Similar cu valorile medii pentru variabilele unidimensionale se definesc cele pentru variabilele bidimensionale. Dacă notăm

$$m_{rs} = \sum_{j=1}^{m}\sum_{k=1}^{n} p_{jk} a_j^r b_k^s ,$$

obținem, ca valori particulare:

$$m_{10} = \sum_{j=1}^{m}\sum_{k=1}^{n} p_{jk} a_j = \sum_{j=1}^{m} a_j$$

$$m_{01} = \sum_{j=1}^{m}\sum_{k=1}^{n} p_{jk} b_k ,$$

care se notează, de obicei, prin m_1 şi respectiv m_2, ele fiind coordonatele centrului de greutate.

Momentele centrate se definesc prin:

$$\mu_{rs} = E\left[(X - m_1)^r (Y - m_2)^s\right] = \sum_{j=1}^{m}\sum_{k=1}^{n} (a_j - m_1)^r (b_k - m_2)^s .$$

Momentele centrate de ordinul al doilea au denumiri speciale în teoria probabilităților. Astfel, μ_{20} se notează cu σ_1^2 şi reprezintă dispersia lui *X*, μ_{02}, notat σ_2^2, este dispersia lui *Y*. μ_{11} este covarianța variabilelor *X* şi *Y*.

Dacă notăm prin $E(X \mid Y = b_j) = \dfrac{\sum_{i=1}^{m} p_{ij} a_i}{p_j}$ valoarea medie a lui *X*, condiționată de $Y = b_j$, obținem modul de comportare al variabilei X atunci când *Y* ia valoarea b_j. Având în vedere toate valorile (a_i, b_j), $i = \overline{1,m}$, iar b_j rămâne fix, diversele medii condiționate referitoare la a_i caracterizează modul de variație al variabilei X pentru $Y = b_j$ fixat.

Se poate caracteriza această variație pintr-o relație $X = f(Y)$, astfel ca $m_2\left(X - f(b_j) \mid Y = b_j\right)$ să fie minimă, m_2 fiind momentul de ordinul al doilea.

În general, se încearcă o exprimare de forma $Y = \alpha x + \beta$, astfel ca $E(Y - \alpha X - \chi)^2$ să fie minimă. Având în vedere că

$$E(Y - \alpha X - \chi)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij}\left(b_j - \alpha a_i - \beta\right)^2,$$

şi derivând în raport cu α, β se obține dreapta de regresie $y - m_2 = \frac{\rho\sigma_2}{\sigma_1}(x - m_1)$. Am notat prin $\rho = \frac{E(X - m_1)(Y - m_2)}{\sqrt{E(X - m_1)^2 (Y - m_2)^2}}$ coeficientul de corelație.

Capitolul 8

Teoria selecției

Introducere

Teoria selecției s-a dezvoltat datorită necesităților practice. În multe situații, apare necesitatea de a obține informații relevante despre mulțimi cu număr mare de elemente, neexistând posibilitatea reală de a studia fiecare element în parte. În aceste cazuri, se poate examina o selecție (eşantion) din mulțime, în ideea ce informația obținută este utilă pentru întreaga populație studiată. Numeroase aplicații au condus la axioma potrivit căreia un eşantion dă informații utile despre întreaga mulțime şi că, pe măsură ce selecția creşte ca volum, datele obținute sunt din ce în ce mai fidele.

8. 1. Generarea valorilor particulare ale unei variabile aleatoare

Simularea unei variabile aleatoare este utilizată atât în statistică (procedee de eşantionare), cât şi în modelarea stochastică. Această simulare se poate realiza cu ajutorul unor obiecte (zar, ruletă), sau algoritmizarea generării unor valori numerice pe calculator. Calculatoarele din ce în ce mai performante au determinat, în principal, orientarea spre cea de a doua metodă.

În ceea ce priveşte clasificarea, există generatoare de variabile aleatoare uniform repartizate şi neuniform repartizate, ultimele construindu-se, în general, pe baza celor din prima categorie.

Variabile aleatoare discrete uniform repartizate. Fie X o variabilă aleatoare discretă $P(X=i)=\frac{1}{n}, \quad i=\overline{1,n}$. Simularea acestei variabile aleatoare utilizează o funcție $g: I^k \to I$, unde I este mulțimea numerelor întregi reprezentate în calculator. Pornind de la valorile inițiale $x_1, x_2, ..., x_n$ şi folosind relația de recurență $x_n = g(x_{n-k},...x_{n-1}), \ n > k$, se generează o secvență de numere. Şirul (x_n) este periodic, deoarece I este finită. Acest generator este suficient de bun dacă perioada lui este mare în raport cu numărul de valori generate şi valorile generate nu sunt secvențial corelate, adică secvențe de p valori succesiv generate ocupă spațiul de dimensiune p. Aceste două condiții pot fi îndeplinite printr-o alegere adecvată a funcției g. Metodele congruențiale sunt cele mai utilizate. Ele se bazează pe o relație de recurență de forma

$$x_n = f(x_{n-k},...,x_{n-1}) \bmod m$$

unde $f : I^k \to I$.

Este comod să alegem funcția liniară $f(x_{n-k},...,x_{n-1}) = a_1 x_{n-1} + ... + a_k x_{n-k} + c$, $a_1, ...a_k, c$ fiind valori întregi.

Metoda generează valori cuprinse între 0 şi m-1, motiv pentru care *m* se alege cât mai mare. Pentru k=0, avem generatori de ordinul întâi. În acest caz, a=16807, c=0, m= $2^{31} - 1$ sau a=24298, c=99991, m=199017 sunt câteva seturi de valori adecvate.

Variabile aleatoare continue. Simularea unei variabile aleatoare continue uniforme pe [0,1] se realizează prin generarea de valori întregi uniform repartizate pe mulțimea {0, 1, ..., m-1} şi prin împărțirea lor prin m-1. Pentru o variabilă aleatoare uniform repartizată pe (0,1) se generează valori din mulțimea {1, ..., m-1}, care se împart la m. O variabilă aleatoare *X*, uniform repartizată pe (a,b) se generează cu ajutorul unei variabile aleatoare *Y*, uniform repartizată pe (0,1) după care se efectuează schimbarea de variabilă $X = (b-a)Y + a$.

În cele ce urmează, ne vom referi la metoda inversării funcției de repartiție.

Fie *X* o variabilă aleatoare având drept funcție de repartiție $F : \mathbf{R} \to [0,1]$**,** atunci inversa ei se defineşte ca $F^{-1}(y) = \{x \in \boldsymbol{R} \mid F(x) \geq u\}$, $(\forall)$ $u \in [0,1]$, rezultând astfel o posibilitate de algoritmizare în vederea simulării variabilei. Pentru diverse repartiții concrete, se obțin diverşi algoritmi.

Repartiții discrete. Fie *X* o variabilă aleatoare discretă, *X*: $\begin{pmatrix} x_i \\ p_i \end{pmatrix}$, $i = \overline{1,n}$,

$\sum_{i=1}^{n} p_i$, $p_i > 0$. Această variabilă are funcția de repartiție

$$F(x) = \begin{cases} F_1 = 0, x \leq x_i; \\ \dots \\ F_i = \sum_{k=1}^{n-1} p_k, x_{i-1} < x \leq x_i; \\ 1, x > x_n. \end{cases}$$

Atunci inversa funcției de repartiție se determină astfel:

$$F^{-1}(u) = x_i,$$

ori de câte ori $F(x_{i-1}) < u < F(x_i)$, $i = \overline{1,n}$, unde $x_0 = -\infty$, $F(x_0) = 0$. Algoritmul de simulare constă în generarea unei valori *u,* uniform repartizate în (0,1), şi în determinarea indicelui *i* astfel încât $F_{i-1} < u \leq F_i$.

Exemplul. 8.1.1. *Să considerăm o variabilă aleatoare binomială X, cu funcția de repartiție*

$$F(x)=\begin{cases}0, x\le 0;\\ \sum_{i=0}^{k} C_n^i p^i (1-p)^{n-1}, k-1<x\le k\\ 1, x>n.\end{cases}$$

X poate fi simulată prin construirea tabelului asociat funcției de repartiție şi aplicarea metodei clasice de căutare. O altă posibilitate este de a simula extragerile cu revenire şi de a număra de câte ori se produce evenimentul de probabilitate p. În acest sens, se consideră N, cel mai mic număr natural pentru care $n_1 = Np$ şi $n_2 = Np$ *sunt naturale şi se construieşte tabelul corespunzător, cu valorile* $t_1 = t_2 = ... = t_n = 1,\ t_{n_1+1} = t_{n_1+2} = ...t_{n_1+n_2} = 0.$

Exemplul 8.1.2. *Fie acum o variabilă aleatoare geometrică X, având funcția de repartiție*

$$F(x)=\begin{cases}0, x\le 0;\\ \sum_{i=0}^{k} pq^k, k-1<x\le k,\\ 1, x>n\end{cases}$$

unde $p+q=1,\ p\in(0,1)$.
X poate fi simulată prin aplicarea modelului clasic de căutare sau prin simularea variabilei aleatoare cu ajutorul monedei trucate în aşa fel încât probabilitatea apariției valorii sa fie p, iar cea a stemei sa fie q. Astfel, variabila X arată la a câta încercare se obține stema prima dată.

Repartițiile continue. În situația în care se cunoaşte expresia inversei funcției de repartiție, se poate utiliza forma generală a algoritmului. În caz contrar, se utilizează algoritmul specific repartițiilor discrete. În prealabil se determină intervalul $(F_{i-1}, F_i]$ ce conține pe *u*, calculându-se

$$x^* = x_{i-1} + (x_i - x_{i-1})\frac{u - F_{i-1}}{F_i - F_{i-1}}.$$

Facem precizarea că aici x^* reprezintă valoarea obținută, şi nu x_i.

Exemplul 8.1.3. *Fie X o variabilă aleatoare exponențială, cu funcția de repartiție* $F(x)=1-e^{-\lambda x},\ x,\lambda>0$. *Inversa lui F este* $F^{-1}(u)=-\frac{1}{\lambda}\ln(1-u)$, *iar u este uniform repartizată în (0,1). Astfel, şi* $1-u$ *este uniform repartizată* în *(0,1), rezultând astfel imediat algoritmul de simulare.*

Exemplul 8.1.4 *Fie X o variabilă aleatoare Weibull, cu funcția de repartiție* $F(x)=1-e^{ax^b}$, $a,b>0$. *Algoritmul de simulare se dudcce uşor, pe baza faptului că inversa funcției F este* $F^{-1}(u)=\left(-\frac{1}{a}\ln(1-u)\right)^{\frac{1}{b}}$.

Exemplul 8.1.5. *Ne referim acum la simularea variabilei aleatoare normal repartizate, mai precis la metoda bazată pe teorema limită centrală. Conform acesteia, dacă* $X_1, X_2, ...X_n \in N(\mu,\sigma)$, *atunci şirul* $Z_n=\frac{\sqrt{n}}{\sigma}\left(\frac{X_1+...X_n}{n}-\mu\right)$ *este convergent în repartiție către* $Z \in N(0,1)$. *Să considerăm* $\mu=\frac{1}{2}$ şi $\sigma^2=\frac{1}{12}$. *Atunci obținem* $Z_n=\frac{X_1+...+X_n-\frac{n}{2}}{\frac{\sqrt{n}}{\sqrt{12}}}$.

Pentru $n=12$, *algoritmul de simulare devine simplu, întrucât* $Z_n=X_1+...+X_{12}-6$.

Exemplul 8.1.6. *Generarea variabilelor* χ^2 *şi Student se bazează pe legătura lor cu repartiția normală standard.*

8. 2. Variabile de eşantionare

Fenomenele din natură, studiile sociologice au consacrat metoda sondajului. Problema care se pune este de a determina caracteristica unei populații formată din *N* indivizi, prin prisma rezultatelor x_1, x_2, ... , x_n obținute prin *n* experiențe indepenente. x_i se numesc valori de eşantionare, iar populația este distribuția teoretică. Dacă populația este infinită, sau poate fi considerată infinită, singura metodă de cercetare pentru determinarea caracteristicilor distribuției teoretice este metoda selecției (sondajele de opinie). Acelaşi lucru se realizează şi pentru un număr finit de indivizi (controlul antidoping realizat pe un număr mic de componenți ai unei echipe).

Definiția 8.2.1. *O subcolectivitate a unei colectivități cercetate se numeşte selecție sau sondaj. Numărul elementelor selecției este volumul selecției.*

Definiția 8.2.2. *Valorile obținute pentru indivizii care intră în selecție privind caracteristica X de numesc date de selecție relative la caracteristica X.*

Pentru o selecție de volum n vom nota datele de selecție cu x_1, x_2, ..., x_n.

Definiția 8.2.3. *Datele de selecție x_1, x_2, ..., x_n sunt valorile unor variabile aleatoare X_1, X_2, ..., X_n , care se vor numi variabile de selecție.*

Pe parcursul întregului capitol vom avea în vedere notațiile menționate în definiția anterioară.

Definiția 8.2.4. *O funcție de variabile de selecție a cărei valoare devine cunoscută când variabilele de selecție sunt înlocuite prin valorile de selecție se numeşte statistică.*

Sondajul este operația de colectare a elementelor unui eşantion din populația statistică examinată. Există sondaje cu revenire (bernoulliene), când elementul extras din populația considerată este reintrodus în colectiv înainte de efectuarea unei noi extrageri, sau sondaje fără revenire.

Colectarea elementelor din eşantion conduce la realizarea mai multor tipuri de sondaje. Sondajul pur aleator se obține cand unitățile statistice au aceeaşi probabilitate de a fi alese din eşantion (sunt echiprobabile). Dacă se prestabileşte un principiu, se efectuează un sondaj dirijat. În cazul în care populația examinată este împărțită în grupuri (straturi) în raport cu o caracteristică prestabilită, avem un sondaj mixt. Acestea pot fi de mai multe feluri: sondaje stratificate simple fără revenire, sondaje stratificate (tipice) în două faze (se aleg mai întâi *r* straturi din cele deja existente, iar după aceea se fac extrageri aleatoare din fiecare strat). Dacă din fiecare start tipic se extrage un număr de unități ales astfel încât raportul dintre volumul eşantionului de strat şi volumul stratului să coincidă cu raportul dintre volumul eşantionului general şi volumul total al populației, se realizează un sondaj stratificat proporțional. Un sondaj în care volumul eşantionului nu este fixat inițial şi prelucrarea continuă până când un anumit eveniment se realizează se numeşte sondaj secvențial.

Cercetarea şi perfecționarea metodelor de analiză a datelor experimentale privind un anumit fenomen depind de volumul eşantionului ales. Datele pot fi ordonate după anumite criterii, spre exemplu: momentul din timp şi locul în care s-a produs fenomenul frecvența apariției acestuia.

Definiție 8.2.5. *Frecvența absolută n_i reprezintă numărul de apariții ale unui rezultat în cele n experimente efectuate asupra eşantionului, în timp ce frecvența relativă f_i este raportul dintre frecvența absolută şi volumul eşantionului.*

Există trei moduri în care pot fi organizate rezultatele $x_1, x_2, ..., x_n$ ale măsurătorilor. Primul dintre acestea este $x_1 < x_2 < ... < x_n$, unde $n_i = 1$, $f_i = \frac{1}{n}$, $i = \overline{1,n}$, n fiind volumul eşantionului, n_i frecvențele absolute ale apariției

valorilor x_i corespunzătoare, iar f_i sunt frecvențele relative, cel de al doilea ar fi

$x_1 < x_2 < ... < x_n$, $n_i \neq 1$, $f_i = \frac{n_i}{n}$, $i = \overline{1,k}$.

Să presupunem acum că măsurătorile pot fi grupate în *k* intervale de valori, de lungime egală, fiecărui interval corespunzându-i un reprezentant $\hat{x}_i$, $i = \overline{1,k}$. În această situație, frecvențele absolute asociate fiecărui interval sunt egale cu numărul de valori ale caracteristicii măsurate în intervalul respectiv. Apare astfel cel de al treilea tip de serie statistică, şi anume

$$\hat{x}_1 < \hat{x}_2 < ... < \hat{x}_k, \; n_i \neq 1, \; f_i = \frac{n_i}{n} \; , \; i = \overline{1,k}.$$

Deosebirea față de cel de al doilea tip constă în faptul că în serie apar reprezentanții intervalelor.

Fie *X* caracteristica examinată. Ea poate fi caracterizată prin

$$X : \begin{pmatrix} x_i \\ f_i \end{pmatrix}, \sum_{i=1}^{n} f_i = 1,$$

sau prin funcția empirică de repartiție, notată F_n^*. Pentru seriile din primele două categorii, aceasta are forma

$$F_n^*(x) = \begin{cases} 0, x < x_1, \\ \sum_{j=1}^{i-1} f_j, x_{i-1} \le x < x_i, i = \overline{1,k+1.} \end{cases}$$

Pentru ultimul tip, funcția empirică de repartiție este dată de

$$F_n^*(x) = \begin{cases} 0, x < l_0, \\ \sum_{j=1}^{i-1} f_j + \frac{x - l_{i-1}}{d} f_i, l_{i-1} \le x < l_i, i = \overline{1,k+1,} \end{cases}$$

k fiind numărul intervalelor (l_{i-1}, l_i), iar *d* lungimea acestor intervale.

Este uşor de remarcat faptul că funcția empirică de repartiție este analogul funcției de repartiție a unei variabile aleatoare discrete finite.

Datorită diversității mărimilor obținute, şi aici sunt necesare analiza şi organizarea datelor. Tendințele de grupare şi împrăştiere în jurul unei tendințe maxime se măsoară cu indicatori care se definesc similar cu cei definiți in capitolul 8, de aceea nu îi vom reaminti.

Definiția 8.2.6. *Media de selecție (momentul de ordin 1) este definită prin relația*

$$\overline{X_n} = \frac{1}{n} \sum_{j=1}^{n} X_j .$$

Dacă x_j reprezintă valoarea observată a variabilei X_j, valoarea numerică a acestei statistici este

$$\overline{x_n} = \frac{1}{n}\sum_{j=1}^{n} x_j .$$

Definiția 8.2.7. *Momentul centrat de selecție de ordinul r este dat de relația*

$$\mu_r = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X})^r ,$$

notație pe care o vom folosi şi pentru valoarea momentului de selecție de ordin r,

$$\mu_r = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^r .$$

Rezultă de aici valoarea dispersiei de selecție,

$$s^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2 .$$

În paragraful următor vom demonstra proprietăți privind media de selecție şi dispersia.

Exemplul 8.2.8. *Să considerăm un eşantion de 20 de clienți ai unui magazin alimentar. Ne propunem să studiem frecvența X cu care clienții fac apel la serviciile magazinului de-a lungul unei săptămâni şi să cercetăm cheltuielile lunare Y în zeci de lei ale clienților pentru achiziționarea de produse din magazinul respectiv. Datele de selecție sunt următoarele (în ordinea în care au fost obținute):*

X : 2,1,1,4,3,2,5,6,1,2,3,2,3,4,6,2,4,3,2,1;

Y : 90,9,101,88,85,77,102,100,86,97,76,121,113,110,96,9,2,108,112,109,103.

Datele de selecție pentru caracteristica X au $n = 6$ *valori distincte, rezultând, astfel, pentru aceasta, distribuția empirică*

$$X : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 6 & 4 & 3 & 1 & 2 \end{pmatrix}.$$

În cazul lui Y, vom face o grupare a datelor de selecție corespunzătoare, în intervalele $[70,80)$, $[80,90)$, *..., rezultând, în acest mod, următoarea distribuție empirică* $Y : \begin{pmatrix} 75 & 85 & 95 & 105 & 115 & 125 \\ 2 & 4 & 4 & 6 & 3 & 1 \end{pmatrix}$.

Mediile de selecție ale celor două caracteristici sunt:

$$\overline{X} = \frac{1}{20}(4\cdot 1 + 6\cdot 2 + 4\cdot 3 + 3\cdot 4 + 1\cdot 5 + 2\cdot 6) = 2{,}85$$

$$\overline{Y} = \frac{1}{20}(2\cdot 75 + 4\cdot 85 + 4\cdot 95 + 6\cdot 105 + 3\cdot 115 + 1\cdot 125) = 98{,}5.$$

Pentru momentele centrate de selecție de ordinul al doilea, obținem:

$$\mu_2(X)=\frac{1}{20}\sum_{j=1}^{20}(x_j-\overline{X})^2=\frac{1}{20}\Big(4\cdot(1-2{,}85)^2+6\cdot(2-2{,}85)^2+4\cdot(3-2{,}85)^2+$$
$$3\cdot(4-2{,}85)^2+1\cdot(5-2{,}85)^2+2\cdot(6-2{,}85)^2\ \Big)=2{,}3275.$$

$$\mu_2(Y)=\frac{1}{20}\sum_{j=1}^{20}\left(y_k-\overline{Y}\right)^2=\frac{1}{20}\Big(2\cdot(75-98{,}5)^2+4\cdot(85-98{,}5)^2+4\cdot(95-98{,}5)^2$$
$$6\cdot(105-98{,}5)^2+3\cdot(115-98{,}5)^2+1\cdot(125-98{,}5)^2\Big)=182{,}5.$$

Funcțiile de repartiție de selecție ale celor două caracteristici sunt:

$$F^*_{20,X}(x)=\begin{cases}0, & x\le 1;\\ \frac{1}{5}, & 1<x\le 2;\\ \frac{1}{2}, & 2<x\le 3;\\ \frac{7}{10}, & 3<x\le 4;\\ \frac{17}{20}, & 4<x\le 5;\\ \frac{9}{10}, & 5<x\le 6;\\ 1, & x>6\end{cases},$$

respectiv

$$F^*_{20,Y}(y)=\begin{cases}0, & y\le 1;\\ \frac{1}{10}, & 75<y\le 85;\\ \frac{3}{10}, & 85<y\le 95;\\ \frac{1}{2}, & 95<y\le 105;\\ \frac{4}{5}, & 105<y\le 115;\\ \frac{19}{20}, & 115<y\le 125;\\ 1, & y>6.\end{cases}$$

Repartiții statistice bidimensionale. Ne îndreptăm acum atenția asupra populațiilor statistice care au două caracteristici (cantitative sau calitative). Să considerăm X şi Y caracteristici cantitative ale unei populații, pentru care s-au determinat valorile $x_1, x_2, ..., x_r$, respectiv $y_1, y_2, ..., y_s$. Fie n_{ij} frecvențele absolute ale cazurilor pentru care $X = x_i$ şi $Y = y_j$, $i = \overline{1,r}$, $j = \overline{1,s}$. n fiind momentul selecției, în mod evident relația $\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = n$. Frecvențele relative sunt $\frac{f_{ij}}{n} = 1$, $\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = n$, şi sunt trecute într-un tabel de corelație, asemănător unei matrice cu r linii şi s coloane.

Definiția 8.2.9. *Momentul de selecție de ordinul k în raport cu X şi Y este dat de relațiile:*

$$m_{k0} = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} x_i^k = \sum_{i=1}^{r} f_{i0} x_i^k = \overline{x},$$

$$m_{0k} = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} y_i^k = \sum_{i=1}^{r} f_{0j} x_i^k = \overline{y},$$

unde $f_{i0} = \sum_{j=1}^{s} f_{ij}$ şi $f_{0i} = \sum_{i=1}^{r} f_{ij}$.

Definiția 8.2.10. *Momentul de selecție de ordin h în raport cu X de ordin k în raport cu Y se defineşte prin*

$$m_{hk} = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} x_i^h y_j^k .$$

Momentele centrate μ_{hk} *se definesc în mod asemănător.*

Definiția 8.2.11. *Momentul centrat mixt de ordinul al doilea,*

$$\mu_{11} = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij}(x_i - \bar{x})(y_j - \bar{y}),$$

reprezintă covarianța de selecție, în timp ce coeficientul de corelație este $\rho = \frac{\mu_{11}}{s_1 s_2}$, $\bar{s}_1^2, \bar{s}_2^2$ *fiind dispersiile de selecție ale celor două variabile individuale. Coeficientul de corelație reprezintă, de fapt, o măsură a dependenței celor două variabile X şi Y.*

8.3. Legi de probabilitate ale variabilelor de eşantionare

Principiul fundamental al statisticii matematice afirmă că frecvența experimentală de apariție a unui eveniment converge către frecvența teoretică, datorat lui Bernoulli.

Teorema 8.3.1. (Bernoulli) *Fie* a_n *numărul de apariții ale unui eveniment A în n experimente independente şi p probabilitatea de realizare acestui eveniment în fiecare experiență. Dacă* $f_n = \frac{a_n}{n}$ *este frecvența relativă de apariție a acestui evenimentului, atunci şirul* (f_n) *converge în probabilitate către p.*

Demonstrație

Deoarece $a_n = nf_n$, a_n este o variabilă binomială, aşadar $E(a_n) = np$ şi $Var(a_n) = np(1-p)$. Au loc următoarele relații, având în vedere inegalitatea lui Cebîşev:

$$P\left(\left| f_{n1-} - p\right| < \varepsilon\right) = P\left(\left|a_n - np\right| < n\varepsilon\right) =$$

$$P\left(\left|a_n - M(a_n) < n\varepsilon\right|\right) > 1 - \frac{D(a_n)}{n^2\varepsilon^2} =$$

$$= 1 - \frac{p(1-p)}{n\varepsilon^2}.$$

De aici rezultă că $\lim_{n\to\infty} P\left(\left| f_n - p\right| < \varepsilon\right) = 1$, *q.e.d.*

Această teoremă permite doar evaluarea directă a probabilității *p* de producere a unui eveniment. Când ne referim la o variabilă aleatoare, pentru obținerea de informații globale trebuie să facem apel la teorema lui Glivenko.

Teorema 8.3.2. *Fie F funcția de repartiție a statisticii X, şi* F_n^* *funcția de repartiție de selecție corespunzătoare unei selecții bernoulliene de volum n, atunci*

$$P\left(\lim_{n\to\infty}\max_{x\to\infty}\left|F_n^*(x)-F(x)\right|=0\right)=1.$$

Pentru demonstrație, îndrumăm cititorul spre lucrarea [11].

Teoremele de convergență arată condițiile în care repartiția statistică (empirică) tinde către cea teoretică. Aceasta din urmă nu este cunoscută, de cele mai multe ori, ea putând fi apreciată doar cu ajutorul momentelor de diferite ordine ale variabilei considerate *X*. În această situație, apare însă problema de a studia în ce măsură diversele momentele de selecție converg către momentele teoretice. Trebuie să precizăm că valorile variabilei *X* rezultate din măsurători sunt, de asemenea, variabile aleatoare, numite variabile de selecție. Ele depind de eşantionul ales, aşadar momentele de selecție devin, la rândul lor, variabile aleatoare.

Dacă *X* este o variabilă aleatoare examinată printr-o selecție de volum *n*, obținută printr-un sondaj pur aleator, care are momentele $E(X^k)$ şi dispersia $Var(X)$, atunci variabilele de selecție $X_1, X_2, ..., X_n$ sunt independente, au aceeaşi repartiție ca şi variabila inițială *X*, aceleaşi momente şi aceeaşi dispersie.

Teorema 8.3.3. *Dacă repartiția teoretică a unei variabile este normală, de medie* μ *şi dispersie* σ^2, *atunci distribuția mediei de selecție obținută prin sondaj pur aleator este de asemenea normală.*

Demonstrație

Variabilele de selecție sunt independente şi normal repartizate. Folosind notațiile anunțate la începutul paragrafului, media de selecție, $\overline{X}=\sum_{i=1}^{k} f_i X_i$, este o combinație liniară de variabile repartizate normal, aşadar are tot repartiție normală, parametrii fiind însă următorii:

$$M(\overline{X})=M\left(\sum_{i=1}^{k} f_i X_i\right)=\sum_{i=1}^{k} M(X_i)=\mu\sum_{i=1}^{k} f_i=\mu;$$

$$Var(\overline{X}) = Var\left(\sum_{i=1}^{k} f_i X_i\right) = \sigma^2 \sum_{i=1}^{k} f_i^2 \text{ , q.e.d.}$$

Observația 8.3.4. *În cazul unei serii statistice din primul tip descris anterior, dispersia mediei are valoarea* $\frac{\sigma^2}{n}$.

Teorema 8.3.5. *Fie* $X_{j1}, X_{j2}, ..., X_{jn_j}$, $j = \overline{1,k}$, *selecții independente din populații normale* $N(\mu_j, \sigma_j^2)$ şi $\overline{X_j}$, $j = \overline{1,k}$, *mediile de selecție. Atunci variabila* $Y = \sum_{j=1}^{k} a_j \overline{X}_j$ *este de asemenea normală, de parametri* $\sum_{j=1}^{k} a_j \mu_j$, *respectiv* $\sum_{j=1}^{k} a_j^2 \frac{\sigma_j^2}{n_j}$.

Demonstrație

Aplicând teorema precedentă, rezultă că $\overline{X_j} \in N\left(\mu_j, \frac{\sigma_j^2}{n_j}\right)$. Variabila *Y*, fiind combinație liniară de variabile normal repartizate, este tot normal repartizată, calcule asemănătoare celor din teorema anterioară conducând la determinarea parametrilor săi. *q.e.d.*

Teorema 8.3.6. *Dacă* $X \in N(0, \sigma^2)$ *şi* $X_1, X_2, ..., X_n$ *reprezintă variabile de selecție obținute prin sondaj pur aleator, atunci* $Y = \sum_{i=1}^{n} X_i^2 \in \chi^2(n, \sigma)$.

Demonstrație

Considerăm $Y_i = X_i^2$ şi $X_i \in N(0, \sigma^2), i = \overline{1,n}$. Au loc relațiile:

$$F_{Y_i}(x) = P(Y_i < x) = P(X_i^2 < x) = P(-\sqrt{x} < X_i < \sqrt{x}), > 0.$$

Deoarece $F_{Y_i}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\sqrt{x}}^{\sqrt{x}} e^{-u^2/2\sigma^2} du$ şi $f_{Y_i}(x) = \frac{1}{\sigma\sqrt{2\pi x}} e^{-x^2/2\sigma^2}$, rezultă că $Y_i \in \chi^2(n, \sigma)$. Funcția caracteristică a lui Y arată că ea este variabilă din $\chi^2(n, \sigma)$, *q.e.d.*

Teorema 8.3.7. *Fie* $X=(X_1,X_2,...,X_n)^t$ *o selecție obținută prin sondaj pur aleator dintr-o populație caracterizată de o repartiție normală redusă și* $A=(a_{ij})_{1\le i,\,j\le n}$ *o martice ortonormală. Atunci variabilele* $V_j=\sum_{k=1}^{n}a_{jk}X_k$, $j=\overline{1,n}$, *sunt independente și normal repartizate, de parametri 0 și 1.*

Demonstrație

Dacă $V=(V_1,V_2,...,V_n)^t$, deoarece $V=AX$ și matricea A este ortonormală, au loc relațiile:

$$\sum_{j=1}^{n}V_j^2=V^tV=X^tA^tAX=X^tX=\sum_{j=1}^{n}V_j^2.$$

Funcția caracteristică a vectorului V devine, succesiv:

$$\varphi_V(t_1,t_2,...,t_n)=E\left(e^{i\sum_{k=1}^{n}t_kv_k}\right)=\frac{1}{\sqrt{2\pi}}\int_{R^n}\prod_{k=1}^{n}e^{it_kx_k-\frac{1}{2}x_k^2}dx_1...dx_n=$$

$$=\frac{1}{\sqrt{2\pi}}\prod_{k=1}^{n}\int_{-\infty}^{\infty}e^{it_kx_k-\frac{1}{2}x_k^2}dx_k=\prod_{k=1}^{n}e^{t_k^2/2}=\prod_{k=1}^{n}\varphi_{X_k}(t_k)=\prod_{k=1}^{n}\varphi_{X_k}(t).$$

Variabilei V_j i se asociază funcția caracteristică următoare

$$\varphi_{V_j}(t)=E\left(e^{it\sum_{k=1}^{n}a_{jk}x_k}\right)=$$

$$\prod_{k=1}^{n}\varphi_{X_k}(ta_{jk})=\prod_{k=1}^{n}e^{\frac{t^2a_{jk}^2}{2}}=e^{\frac{t^2}{2}\sum_{k=1}^{n}a_{jk}^2}=e^{\frac{t^2}{2}}=\varphi_{X_k}(t).$$

Acest lucru arată că V_j sunt normal redus repartizate. Având în vedere că funcția caracteristică a vectorului V este chiar produsul funcțiilor caracteristice ale variabilelor componente, rezultă că V_j sunt variabile independente, *q.e.d.*

Teorema 8.3.8. *Dacă* $X=(X_1,X_2,...,X_n)^t$ *este o selecție obținută prin sondaj pur aleator dintr-o populație caracterizată de o distribuție normală redusă, atunci variabilele:*

$$U=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i=\sqrt{n}\,\overline{X},$$

$$V=\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\sum_{i=1}^{n}X_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}X_i\right)^2$$

sunt independente. În plus, $U\in N(0,1)$ şi $V\in\chi^2(n-1,1)$.

Demonstraţie

U este o combinaţie liniară de variabile independente identic repartizate, aşadar *U* este normal repartizată, cu parametrii:

$$E(U)=0,$$

$$Var(U)=Var\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}nVar(X_i)=1.$$

Să considerăm acum o matrice ortonormală *A*, în care $a_{1k}=\frac{1}{n},\forall k,$ şi $V=AX=(V_1,V_2,...V_n)$. Având în vedere teorema anterioară, $V_j\in N(0,1)$. Egalităţile

$$\sum_{j=2}^{n}V_j^2=\sum_{i=1}^{n}X_i^2-\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_i\right)^2=\sum_{i=1}^{n}X_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}X_i\right)^2=V$$

conduc la concluzia că $V\in\chi^2(n-1,1)$, *q.e.d.*

Teorema 8.3.9. *Fie* $X=(X_1,X_2,...,X_n)$ *o selecţie obţinută prin sondaj pur aleator, dintr-o populaţie normal distribuită, cu parametri* μ *şi* σ^2. *Atunci:*

$$U=\frac{\overline{X}-\mu}{\sigma}\sqrt{n}\in N(0,1),$$

$$V=\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2\in\chi^2(n-1,1).$$

Demonstraţie

Notând $Y_i=\frac{X_i-\mu}{\sigma},i=\overline{1,n},$ rezultă că $Y_i\in N(0,1)$. Concluzia rezultă uşor, verificând condiţiile din teorema anterioară prin intermediul variabilelor ajutătoare abia introduse, *q.e.d.*

Capitolul 9

Teoria estimației

9.1. Estimatori nedeplasați

Să considerăm o variabilă aleatoare X a cărei lege de probabilitate conține un parametru θ. Fie $X_1, ..., X_n$ variabile aleatoare independente care au aceeaşi distribuție ca şi X.

Alegem o anumită funcție $t(X_1, ..., X_n)$ pe care o vom utiliza ca estimator al lui θ; cu alte cuvinte, dacă dispunem de valorile $x_1, ..., x_n$ obținute experimental, numărul $t(x_1, ..., x_n)$ va fi considerat ca estimator al parametrului θ.

Definiția 9.1.1. *$t(X_1, ..., X_n)$ se numeşte estimator nedeplasat al parametrului θ dacă media lui $t(X_1, ..., X_n)$ este egală cu θ pentru orice valoare posibilă a lui θ.*

Exemplul 9.1.2. *Fie μ media lui X. Atunci*

$$t(X_1, ..., X_n) = \frac{1}{n}(X_1 + ... + X_n)$$

este un estimator nedeplasat al lui μ, fiindcă în mod clar media variabilei aleatoare

$$\frac{1}{n}(X_1 + ... + X_n) \text{ este egală cu } \frac{1}{n}(\mu + ... + \mu) = \mu .$$

Exemplul 9.1.3. *Fie μ media lui X şi σ^2 dispersia lui X. Pentru un n fixat, notăm*

$$\overline{X} = \frac{1}{n}(X_1 + ... + X_n) .$$

Am văzut în exemplul anterior că $\overline{X}$ este un estimator nedeplasat al mediei μ. În calitate de estimator al dispersiei lui X putem alege pe

$$S^2 = \frac{1}{n}(X_1^2 + ... + X_n^2) - \overline{X}^2 .$$

Un calcul algebric elementar ne convinge că media variabilei aleatoare S^2 este egală cu $\frac{n-1}{n}\sigma^2$. Aceasta arată că S^2 nu este estimator nedeplasat; deducem însă imediat că media lui

$$\frac{n}{n-1} S^2$$

este egală cu σ^2, deci $\frac{n}{n-1} S^2$ este estimator nedeplasat pentru dispersia σ^2. Cu alte cuvinte, dacă dispunem de valorile experimentale $x_1, ..., x_n$ vom estima dispersia σ^2 prin

$$\frac{n}{n-1}\left(\frac{x_1^2 + ... + x_n^2}{n} - \left(\frac{x_1 + ... + x_n}{n} \right)^2 \right) .$$

9.2. Estimatori de maximă verosimilitate

Fie $f(x; \theta)$ densitatea de probabilitate a unei variabile aleatoare *X*, în care θ este parametrul care urmează să fie estimat. Să presupunem că avem la dispoziție valorile experimentale $(x_1, ..., x_n)$ ale lui *X*, obținute în urma unei selecții de volum *n*.

Definiția 9.2.1. *Funcția*

$$L(x_1, ..., x_n; \theta) = f(x_1; \theta) ... f(x_n; \theta)$$

se numeşte funcția de verosimilitate.

Definiția 9.2.2. *Un estimator de maximă verosimilitate este un estimator care maximizează pe L ca funcție de θ.*

Exemplul 9.2.3. *Să considerăm densitatea exponențială*

$$f(x; \theta) = \begin{cases} \theta e^{-\theta x}, & x > 0; \\ 0, & x \le 0. \end{cases}$$

Funcția de verosimilitate este

$$L = \theta e^{-\theta x_1} ... \theta e^{-\theta x_n} = \theta^n e^{-\theta (x_1 + ... + x_n)}.$$

Derivata lui L ca funcție de θ este egală cu

$$\theta^{n-1} e^{-\theta (x_1 + ... + x_n)} (n - (x_1 + ... + x_n)\theta) .$$

Deducem imediat că L îşi atinge maximumul pentru

$$\theta = \frac{n}{x_1 + ... + x_n} .$$

Aşadar, estimatorul de maximă verosimilitate al parametrului θ din legea exponențială este

$$\frac{n}{X_1 + ... + X_n} .$$

Exemplul 9.2.4. *Să considerăm variabila X cu densitatea uniformă*

$$f\ (x;\ \theta)\ =\ \begin{cases} \dfrac{1}{\theta}\ , & 0 \le x \le \theta\ ; \\ \\ 0 \qquad , & \text{in rest}\ . \end{cases}$$

Ținând seama de natura acestei variabile, valorile experimentale $(x_1, ..., x_n)$ *satisfac condiția* $0 \le x_i \le \theta\ ,\ i = 1, ..., n$.

Funcția de verosimilitate este

$$L\ (x_1\ ,\ ...,\ x_n\ ;\ \theta)\ =\ \frac{1}{\theta^n}\ .$$

Valoarea ei este cu atât mai mare, cu cât θ *este mai mic. Datorită condițiilor impuse, cea mai mică valoare posibilă a lui* θ *este* $\max\{x_1, ..., x_n\}$.

Prin urmare, estimatorul de maximă verosimilitate este $\max\{X_1, ..., X_n\}$.

Exemplul 9.2.5. *Să considerăm variabila aleatoare X cu densitatea*

$$f\ (x;\ \theta)\ =\ \begin{cases} (1+\theta)\ x^{\theta}\ , & 0 \le x \le 1\ ; \\ \\ 0 \qquad , & \text{in rest}\ . \end{cases}$$

Funcția de verosimilitate este

$$L\ (x_1\ ,\ ...,\ x_n\ ;\ \theta)\ =\ (1+\theta)^n\ (x_1\ ...\ x_n)^{\theta}\ .$$

Derivata lui L ca funcție de θ *este egală cu*

$$(1+\theta)^{n-1}(x_1\ ...\ x_n)^{\theta}\ (n + (1+\theta)\ln(x_1\ ...\ x_n))\ .$$

Funcția L își atinge maximumul pentru

$$\theta\ =\ -\ \frac{n}{\ln x_1 + ... + \ln x_n}\ -\ 1\ ,$$

deci estimatorul de maximă verosimilitate este în acest caz

$$-\ \frac{n}{\ln X_1 + ... + \ln X_n}\ -\ 1\ .$$

Exemplul 9.2.6. *Fie X o variabilă normală cu densitatea*

$$f\ (x;\ \theta)\ =\ \frac{1}{\sqrt{2\pi}}\ e^{-(x-\theta)^2/2}\ .$$

Funcția de verosimilitate este

$$L = \frac{1}{\left(\sqrt{2\pi}\right)^n} e^{-\left[(x_1-\theta)^2+\ldots+(x_n-\theta)^2\right]/2} .$$

Se constată imediat că ea îşi atinge maximumul pentru

$$\theta = \frac{x_1+\ldots+x_n}{n} .$$

Deci, estimatorul de maximă verosimilitate este

$$\frac{X_1+\ldots+X_n}{n} .$$

Să observăm că θ coincide cu media variabilei X, deci, estimatorul de maximă verosimilitate este tocmai estimatorul nedeplasat despre care a fost vorba în Exemplul 9.1.2.

Exemplul 9.2.7. *Să considerăm o variabilă normală X cu densitatea*

$$f(x;\theta) = \frac{1}{\theta\sqrt{2\pi}} e^{-x^2/2\theta^2} .$$

Funcţia de verosimilitate este acum

$$L = \frac{1}{\left(\sqrt{2\pi}\right)^n} \frac{1}{\theta^n} e^{-\left(x_1^2+\ldots+x_n^2\right)/2\theta^2} .$$

Ea îşi atinge maximumul pentru

$$\theta = \left(\frac{1}{n}\left(x_1^2+\ldots+x_n^2\right)\right)^{1/2} ,$$

deci, estimatorul de maximă verosimilitate este în acest caz

$$\left(\frac{1}{n}\left(X_1^2+\ldots+X_n^2\right)\right)^{1/2} .$$

Capitolul 10

Estimarea prin interval de încredere

Introducere

Estimația punctuală a unui parametru θ, necunoscut la nivelul unei populații, deşi constituie o informație în legătură cu acesta, nu poate fi utilizată fără a avea o imagine şi asupra mărimii probabilistice a erorii de estimare. Apare astfel necesitatea estimării unui parametru prin aşa numitul, interval de încredere. În capitolul de față vom prezenta forma generală a intervalului de încredere, expresia intervalului de încredere pentru medie inclusiv cazul particular al unei proporții, interval de încredere pentru diferența a două medii, interval de încredere pentru dispersie şi respectiv interval de încredere pentru raportul a două dispersii. Pe tot parcursul capitolului va fi prezentat doar cazul când eşantionul se formează pe baza unei selecții simple aleatoare formată prin extrageri independente.

10.1. Forma generală a intervalului de încredere

Fie o populație statistică A, variabila statistică X studiată prin intermediul unei selecții simple de volum n, formată prin extrageri independente, $(X_1, X_2, ..., X_n)$ şi un parametru necunoscut θ asociat variabilei X, pentru care se obține pe baza selecției estimatorul $\overline{\theta}(X_1, X_2, ..., X_n)$, având densitatea de probabilitate $f(\overline{\theta})$.

Definiția 10.1.1. *Se numeşte* ***interval de încredere pentru parametrul necunoscut θ, cu nivelul de semnificație*** $\alpha \in (0,1)$, *intervalul*

$$(h_1(\overline{\theta}(X_1, ..., X_n)),\ h_2(\overline{\theta}(X_1, ..., X_n))) \qquad (10.1.1)$$

susceptibil de a conține valoarea lui θ *cu o probabilitate* $(1-\alpha)$, *unde* $h_1(\overline{\theta}) = h_1(\overline{\theta}(X_1, ..., X_n))$ *şi* $h_2(\overline{\theta}) = h_2(\overline{\theta}(X_1, ..., X_n))$ *rezultă din legea de probabilitate dată prin* $f(\overline{\theta})$.

După observarea statistică a eşantionului (selecției), cele două limite ale intervalului mai sus menționat pot fi determinate numeric. Intervalul care încadrează parametrul θ are proprietatea că acoperă valoarea θ în $100(1-\alpha)$ din cazuri. Cu cât nivelul de semnificație α este mai mic cu atât θ are şanse mai mari de a se afla în acel interval (de regulă $\alpha \le 0{,}05$). Probabilitatea $1-\alpha$ este probabilitatea de garantare a intervalului, valorile acceptate fiind de regulă peste 0,95.

Observația 10.1.2. *Obținerea celor două limite ale intervalului se realizează pornind de la faptul că pentru* $\bar{\theta}(X_1, X_2, ..., X_n)$ *(sau pentru o altă variabilă aleatoare de lege de probabilitate cunoscută a cărei expresie îl conține pe θ) se pot determina pe baza legii de probabilitate cunoscute, două valori particulare* $a_1(\theta)$ *şi* $a_2(\theta)$ *astfel încât să aibă loc relația:*

$$P(a_1(\theta) < \bar{\theta} < a_2(\theta)) = \int_{a_1(\theta)}^{a_2(\theta)} f(\bar{\theta}) d\bar{\theta} = 1-\alpha\,, \qquad (10.1.2)$$

formulă ce se poate scrie, prin artificii matematice, astfel:

$$P(h_1(\bar{\theta}(X_1,....,X_n)) < \theta < h_2(\bar{\theta}(X_1,...X_n))) = \int_{a_1(\theta)}^{a_2(\theta)} f(\bar{\theta}) d\bar{\theta} = 1-\alpha\,. \quad (10.1.3)$$

.

Observația 10.1.3. *Dacă estimatorul* $\bar{\theta}$ *este un estimator centrat (nedeplasat), deci* $E(\bar{\theta}) = \theta$ *intervalul de încredere este simetric în raport cu* $\bar{\theta}$*, vom avea* $h_1(\bar{\theta}) = \bar{\theta} - \Delta\bar{\theta}$ *şi* $h_2(\bar{\theta}) = \bar{\theta} + \Delta\bar{\theta}$ *unde* $\Delta\bar{\theta}$ *reprezintă eroarea limită de estimare sub formă absolută. Astfel intervalul de încredere se poate scrie:*

$$P(\bar{\theta} - \Delta\bar{\theta} < \theta < \bar{\theta} + \Delta\bar{\theta}) = \int_{a_1(\theta)}^{a_2(\theta)} f(\bar{\theta}) d\bar{\theta} = 1-\alpha \qquad (10.1.4)$$

unde $\Delta\bar{\theta}$ *este de obicei proporțională cu dispersia estimatorului, prin urmare se scrie de forma* $\Delta\bar{\theta} = z \cdot \sigma_{\bar{\theta}}$.

Probabilitatea $1-\alpha$ şi eroarea limită constituie împreună o măsură a preciziei estimării parametrului θ prin $\bar{\theta}(X_1,...,X_n)$. De regulă acestea se fixează apriori, astfel:

$$R_{\bar{\theta}} = \frac{\Delta\bar{\theta}}{\theta} \cdot 100 \le 5\% \quad \text{şi} \quad \alpha \le 5\%\,. \qquad (10.1.5)$$

Pe baza acestor parametri ai peciziei va rezulta dimensiunea eşantionului prin care se asigură precizia dorită, folosind în acest scop formula $\Delta\overline{\theta} = z \cdot \sigma_{\overline{\theta}}$ în care se va vedea că dispersia estimatorului $\sigma_{\overline{\theta}}$ depinde şi de volumul *n* al selecţiei, iar *z* depinde de nivelul de semnificaţie stabilit şi de legea de probabilitate implicată.

10.2. Interval de încredere pentru medie

Vom considera în cele ce urmează, cazul când parametrul necunoscut θ pe care dorim să-l estimăm prin interval de încredere este media unei variabile aleatoare.

Propoziţia 10.2.1. *Intervalul de încredere pentru media unei variabile X ce urmează legea normală* $N(\mu,\sigma^2)$ *cu* $\mu \in IR$ *necunoscut şi* $\sigma^2 > 0$ *cunoscut este de forma* $(\overline{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}})$, cu

$$P(\overline{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1-\alpha \qquad (10.2.1)$$

unde $\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$ *este media de selecţie corespunzătoare unui eşantion obţinut prin extrageri aleatoare independente,* $\alpha \in (0,1)$ *este nivelul de semnificaţie iar* $z_{1-\frac{\alpha}{2}}$ *este cuantila de ordin* $1-\frac{\alpha}{2}$ *a repartiţiei normale standard dată prin funcţia de repartiţie (funcţia integrală Laplace - Gauss),* $\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$.

Demonstraţie

Întrucât $X \in N(\mu,\sigma^2)$, iar media de selecţie este de forma $\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$, extragerile fiind independente, urmează că $\overline{X} \in N\left(\mu, \frac{\sigma^2}{n}\right)$ şi mai departe, avem că variabila

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

urmează legea normală standard (centrată şi redusă), $N(0,1)$. În consecință, fiind dat un prag de semnificație $\alpha \le 0{,}05$, vom găsi două numere z_1 şi z_2 astfel încât:

$$P(z_1 < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_2) = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{t^2}{2}} dt = \phi(z_2) - \phi(z_1) = 1 - \alpha$$

unde ϕ este funcția de repartiție a legii normale standard, Laplace-Gauss, $\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$, ale cărei valori sunt tabelate. Mai departe, prin artificii de calcul folosind probabilitatea unor evenimente echivalente, avem că:

$$P(\overline{X} - z_2 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} - z_1 \cdot \frac{\sigma}{\sqrt{n}}) = \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{t^2}{2}} dt = \phi(z_2) - \phi(z_1) = 1 - \alpha .$$

Pentru α fixat, se pot determina o infinitate de numere z_1 şi z_2, astfel încât $\phi(z_2) - \phi(z_1) = 1 - \alpha$, însă pentru o precizie cât mai bună a intervalului de încredere, ne va interesa un interval de lungime minimă, ori acesta se obține pentru cazul când $z_1 = -z_2$, prin urmare, avem

$$P(\overline{X} - z_2 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_2 \cdot \frac{\sigma}{\sqrt{n}}) = \frac{1}{\sqrt{2\pi}} \int_{-z_2}^{z_2} e^{-\frac{t^2}{2}} dt = \phi(z_2) - \phi(-z_2) = 1 - \alpha .$$

În acest caz z_2 se determină din relația $\phi(z_2) - \phi(-z_2) = 1 - \alpha$ cu $\phi(-z_2) = 1 - \phi(z_2)$, adică $\phi(z_2) = 1 - \frac{\alpha}{2}$, ceea ce înseamnă că $z_2 = z_{1-\frac{\alpha}{2}}$ este cunatila de ordin $1 - \frac{\alpha}{2}$, a repartiției normale standard.

Prin urmare, intervalul este dat prin formula

$$P(\overline{X} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha \ . \ q.e.d.$$

Observația 10.2.2. *Valorile funcției de repartiție Laplace - Gauss,* $\phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$ *sunt tabelate, adică pentru probabilitatea auxiliară* $1-\frac{\alpha}{2}$, *vom găsi în tabel valoarea cuantilei* $z_2=z_{1-\frac{\alpha}{2}}$ *însă există şi tabele, care prezintă valorile funcției* $\tilde{\phi}(x)=\frac{1}{\sqrt{2\pi}}\int_{0}^{x}e^{-\frac{t^2}{2}}dt$, $x>0$, *caz în care* z_2 *este corespunzătoare valorii* $\frac{1-\alpha}{2}$.

Observația 10.2.3. *Folosind notația:* $\Delta\overline{X}=z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$, *se poate scrie intervalul de încredere ca un caz particular al formulei (10.1.4), (*$\overline{X}$ *fiind estimator nedeplasat) şi anume:*

$$P(\overline{X}-\Delta\overline{X}<\mu<\overline{X}+\Delta\overline{X})=1-\alpha\,.$$

Determinarea efectivă a celor două limite presupune cunoaşterea expresiei matematice a estimatorului $\overline{X}$ *şi a erorii medii pătratice de estimare* $\sigma_{\overline{X}}$, *fiecare variind de la un tip de sondaj la altul, aici fiind prezentat doar cazul unui sondaj aleator simplu cu extrageri independente. Cu cât lungimea* $\Delta\overline{X}$ *a intervalul de încredere, respectiv nivelul de semnificație* α *sunt mai mici, cu atât estimația parametrului necunoscut este mai bună.*

Observația 10.2.4. *Pentru selecții de volum mare, intervalul de încredere pentru medie, precizat în Propoziția 10.2.1 este valabilă şi pentru cazul în care variabila X urmează o lege oarecare de probabilitate, datorită Teoremei limită centrală.*

Exemplul 10.2.5. *Să presupunem că dispunem de valorile unei variabile* $X\in N(\mu,2^2)$, *obținute printr-o selecție simplă cu extrageri independente, de volum 25 şi de medie* $\overline{x}=55$, *obiectivul fiind acela de a estima media la nivelul populației,* μ, *necunoscută, printr-un interval de încredere de 95%.*

Soluție

Întrucât $\alpha=0{,}05$, obținem din *Anexa 1*, $z_{1-\frac{\alpha}{2}}=1{,}96$ şi mai departe, pentru 95% din cazuri,

$$55-1{,}96\cdot\frac{2}{\sqrt{25}}<\mu<55+1{,}96\cdot\frac{2}{\sqrt{25}}\,.$$

Prin urmare, în 95% din cazuri, intervalul (54,216 - 55,784) va acoperi valoarea necunoscută a parametrului μ.

Propoziția 10.2.6. *Intervalul de încredere de tip* $1-\alpha$, *pentru media unei variabile X ce urmează legea normală* $N(\mu,\sigma^2)$ *cu* $\mu \in IR$ *necunoscut şi* $\sigma^2 > 0$ *necunoscut este de forma* $(\overline{X} - t_{n-1,\,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \overline{X} + t_{n-1,\,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}})$, cu

$$P(\overline{X} - t_{n-1,\,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} < \mu < \overline{X} + t_{n-1,\,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}) = 1-\alpha \qquad (10.2.2)$$

unde $\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$ *este media de selecție corespunzătoare unui eşantion obținut prin extrageri aleatoare independente,* $s^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + ... + (X_n - \overline{X})^2}{n-1}$ *este estimația absolut corectă a lui* σ^2, $\alpha \in (0,1)$ *este nivelul de semnificație iar* $t_{n-1,1-\frac{\alpha}{2}}$ *este cuantila de ordin* $1-\frac{\alpha}{2}$ *a repartiției Student cu n-1 grade de libertate.*

Demonstrație

Conform legilor de probabilitate ale variabilelor de eşantionare, atunci când $x \in N(\mu,\sigma^2)$, variabila $T = \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}}$ urmează legea Student cu *n-1* grade de libertate, întrucât se obține ca raport între două variabile independente, o variabilă ce urmează legea normală standard şi radicalul unei alte variabile de tip χ^2 raportată la numărul gradelor de libertate. Procedând ca în cazul când σ^2 este cunoscut se obține intervalul de încredere corespunzător. *q.e.d.*

Exemplul 10.2.7. *Să considerăm estimarea printr-un interval de încredere de tip 98%, a notei medii de repartiție normală, pornind de la următoarele date: 5, 10, 7, 6, 9, 8.*

Soluție

Întrucât $\alpha = 0{,}02$, $n = 6$, obținem din *Anexa 2*, $t_{n-1,1-\frac{\alpha}{2}} = 3{,}365$. Pe baza selecției avem $\overline{x} = 7{,}5$, $s = 1{,}87$ şi mai departe, în 98% din cazuri,

$$7{,}5 - 3{,}365 \cdot \frac{1{,}87}{\sqrt{6}} < \mu < 7{,}5 + 3{,}365 \cdot \frac{1{,}87}{\sqrt{6}}.$$

Prin urmare, în 98% din cazuri, intervalul (4,93 - 10) va acoperi valoarea necunoscută a parametrului μ.

Observația 10.2.8. *Pentru selecții de volum mare, diferența între valorile cuantilelor repartiției Student şi cele ale repartiției normale standard este neglijabilă. De asemenea este neglijabilă şi diferența dintre estimatorul absolut corect* $s^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + ... + (X_n - \overline{X})^2}{n-1}$ *şi estimatorul* $\mu_2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + ... + (X_n - \overline{X})^2}{n}$, *prin urmare, se poate utiliza formula (10.2.1) cu* σ^2 *înlocuit de* μ_2.

Propoziția 10.2.9. *Pentru selecții de volum mare, intervalul de încredere de tip* $1-\alpha$ *, pentru media unei variabile X ce urmează legea Bernoulli de parametru necunsocut p este de forma*

$$(\hat{p} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}) \qquad (10.2.3)$$

unde $\hat{p} = \frac{X_1 + X_2 + ... + X_n}{n}$ *este media de selecție corespunzătoare unui eşantion obținut prin extrageri aleatoare independente,* $\alpha \in (0,1)$ *este nivelul de semnificație iar* $z_{1-\frac{\alpha}{2}}$ *este cuantila de ordin* $1-\frac{\alpha}{2}$ *a repartiției normale standard.*

Demonstrație

Media unei variabile aleatoare ce urmează legea Bernoulli de parametru necunoscut *p*, are pentru selecții de volum mare, conform *Observației 10.2.4,* o repartiție aproximativ normală cu media egală cu *p* şi abaterea medie pătratică aproximativ egală cu $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, prin urmare aplicând *Propoziția 10.2.1* se obține intervalul dorit. *q.e.d.*

O astfel de medie exprimă de fapt proporția *p* de indivizi dintr-o populație care au o anumită caracteristică şi este tratată ca media unei variabile *X* cu valorile 0 şi 1, prin valoarea 1 notând valoarea variabilei X pentru indivizii care au această caracteristică.

Exemplul 10.2.10. *La un sondaj electoral, 40 din 100 de persoane chestionate s-au pronunțat în favoarea unui candidat, scopul fiind determinarea unui interval de încredere de tip 95% pentru procentul de alegători favorabil acelui candidat.*

Soluție

În 95% din cazuri intervalul va fi $\frac{40}{100}-1{,}96\sqrt{\frac{0{,}4\cdot 0{,}6}{100}}<p<\frac{40}{100}+1{,}96\sqrt{\frac{0{,}4\cdot 0{,}6}{100}}$, adică (0,304- 0,496), prin urmare în 95% din cazuri, procentul favorabil candidatului va fi aproximativ între 30% şi 50%.

10.3. Interval de încredere pentru diferența a două medii

Vom considera în cele ce urmează problema determinării intervalelor de încredere pentru diferența a două medii, utilă în cazul în care dorim să avem o informație privind diferența de comportament a unei variabile, de la o populație la alta.

Propoziția 10.3.1. *Fie două populații studiate în raport cu variabila X, pentru prima populație variabila fiind de tipul* $N\left(\mu_1,\sigma_1^{\,2}\right)$ *iar pentru a doua, de tipul* $N\left(\mu_2,\sigma_2^{\,2}\right)$*, cu* μ_1,μ_2 *necunoscute, respectiv* $\sigma_1^{\,2},\sigma_2^{\,2}$ *cunoscute. Fie de asemenea două selecții bazate pe extrageri independente de volume* n_1 şi respectiv n_2*, din cele două populații, cu mediile de selecție* $\overline{X}_1,\overline{X}_2$ *. Pentru un nivel de semnificație,* $\alpha\in(0,1)$*, intervalul de încredere pentru diferența celor două medii este de forma*

$$\overline{X}_1-\overline{X}_2-z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{\sigma_1^{\,2}}{n_1}+\frac{\sigma_2^{\,2}}{n_2}}<\mu_1-\mu_2<\overline{X}_1-\overline{X}_2+z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{\sigma_1^{\,2}}{n_1}+\frac{\sigma_2^{\,2}}{n_2}} \quad (10.3.1)$$

unde $z_{1-\frac{\alpha}{2}}$ *este cuantila de ordin* $1-\frac{\alpha}{2}$ *a repartiției normale standard.*

Demonstrație

Variabila aleatoare $Z=\frac{\left(\overline{X}_1-\overline{X}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{\sigma_1^{\,2}}{n_1}+\frac{\sigma_2^{\,2}}{n_2}}}$ urmează legea normală standard. *q.e.d.*

Exemplul 10.3.2. *Vom determina intervalul de încredere de tip 95% pentru diferența mediilor unei variabile în două populații diferite în care se presupune*

că repartițiile variabilei sunt normale, de medii necunoscute μ_1 *respectiv* μ_2*şi dispersii cunoscute* $\sigma_1^{\,2} = 2^2, \sigma_2^{\,2} = 3^2$*, pornind de la selecțiile de volum* $n_1 = 7$ *şi respectiv* $n_2 = 8$*, de medie,* $\bar{x}_1 = 22$ *şi respectiv,* $\bar{x}_2 = 20$*.*

Soluție

În 95% din cazuri intervalul va fi

$$\left((22-20) - 1{,}96 \cdot \sqrt{\frac{4}{7} + \frac{9}{8}} < \mu_1 - \mu_2 < (22-20) + 1{,}96 \cdot \sqrt{\frac{4}{7} + \frac{9}{8}}\right).$$

Propoziția 10.3.3. *Fie două populații studiate în raport cu variabila X, pentru prima populație variabila fiind de tipul* $N\left(\mu_1, \sigma_1^{\,2}\right)$ *iar pentru a doua, de tipul* $N\left(\mu_2, \sigma_2^{\,2}\right)$*, cu* μ_1, μ_2 *necunoscute, respectiv* $\sigma_1^{\,2}, \sigma_2^{\,2}$ *necunoscute dar egale. Fie de asemenea două selecții bazate pe extrageri independente de volume* n_1 *şi respectiv* n_2*, din cele două populații cu mediile de selecție* $\bar{X}_1, \bar{X}_2$ *şi estimatorii absolut corecți ai dispersiilor,* $s_1^{\,2}, s_2^{\,2}$*. Pentru un nivel de semnificație,* $\alpha \in (0,1)$*, intervalul de încredere pentru diferența celor două medii este de forma*

$$\bar{X}_1 - \bar{X}_2 - t_{\gamma, 1-\frac{\alpha}{2}} \cdot S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < \bar{X}_1 - \bar{X}_2 + t_{\gamma, 1-\frac{\alpha}{2}} \cdot S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (10.3.2)$$

unde $t_{\gamma, 1-\frac{\alpha}{2}}$ *este cuantila de ordin* $1 - \frac{\alpha}{2}$ *a repartiției Student cu* $\lambda = n_1 + n_2 - 2$ *grade de libertate iar* $S^2 = \frac{(n_1 - 1)s_1^{\,2} + (n_2 - 1)s_2^{\,2}}{n_1 + n_2 - 2}$.

Demonstrație

Variabila aleatoare $T = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)}{S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ urmează legea Student cu $n_1 + n_2 - 2$ grade de libertate. *q.e.d.*

Exemplul 10.3.4. *În vederea comparării comportamentului unei variabile în două populații normale, având aceeaşi dispersie, se pune problema determinării unui interval de încredere de tip 95% pentru diferența mediilor variabilei în cele*

două populații, pornind de la două seturi de date de volum 10 şi 12 cu mediile $\bar{x}_1 = 2{,}5$ *şi* $\bar{x}_2 = 2$, *respectiv* $s_1^{\,2} = 0{,}2$ şi $s_2^{\,2} = 0{,}22$.

Soluție

În 95% din cazuri intervalul va fi

$$\left((2{,}5-2) - t_{\gamma,1-\frac{\alpha}{2}} \cdot S \cdot \sqrt{\frac{1}{10}+\frac{1}{12}} < \mu_1 - \mu_2 < (2{,}5-2) - t_{\gamma,1-\frac{\alpha}{2}} \cdot S \cdot \sqrt{\frac{1}{10}+\frac{1}{12}} \right),$$

cu $t_{\gamma,1-\frac{\alpha}{2}} = t_{20,0.975} = 2{,}086$ *şi* $S^2 = \dfrac{(10-1)\cdot 0{,}2 + (12-1)\cdot 0{,}22}{10+12-2}$.

Propoziția 10.3.5. *Fie două populații studiate în raport cu variabila X, pentru prima populație variabila fiind de tipul* $N\left(\mu_1, \sigma_1^{\,2}\right)$ *iar pentru a doua, de tipul* $N\left(\mu_2, \sigma_2^{\,2}\right)$, *cu* μ_1, μ_2 *necunoscute, respectiv* $\sigma_1^{\,2}, \sigma_2^{\,2}$ *necunoscute şi diferite. Fie de asemenea două selecții bazate pe extrageri independente de volume* n_1 *şi respectiv* n_2, *din cele două populații cu mediile de selecție* $\bar{X}_1, \bar{X}_2$ *şi estimatorii absolut corecți ai dispersiilor,* $s_1^{\,2}, s_2^{\,2}$. *Pentru un nivel de semnificație,* $\alpha \in (0,1)$, *intervalul de încredere pentru diferența celor două medii este de forma*

$$\bar{X}_1 - \bar{X}_2 - t_{\gamma,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^{\,2}}{n_1}+\frac{s_2^{\,2}}{n_2}} < \mu_1 - \mu_2 < \bar{X}_1 - \bar{X}_2 + t_{\gamma,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^{\,2}}{n_1}+\frac{s_2^{\,2}}{n_2}} \qquad (10.3.3)$$

unde $t_{\gamma,1-\frac{\alpha}{2}}$ *este cuantila de ordin* $1-\dfrac{\alpha}{2}$ *a repartiției Student cu* γ *grade de libertate unde* $\dfrac{1}{\gamma} = \dfrac{c^2}{n_1-1} + \dfrac{(1-c)^2}{n_2-1}$ *iar* $c = \dfrac{\frac{s_1^{\,2}}{n_1-1}}{\frac{s_1^{\,2}}{n_1-1}+\frac{s_2^{\,2}}{n_2-1}}$.

Demonstrație

Variabila aleatoare $T = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^{\,2}}{n_1}+\frac{s_2^{\,2}}{n_2}}}$ urmează legea Student cu γ grade de libertate. *q.e.d.*

Exemplul 10.3.6. *În vederea comparării comportamentului unei variabile în două populații normale, având dispersii diferite, se pune problema determinării unui interval de încredere de tip 95% pentru diferența mediilor variabilei în cele două populații, pornind de la două seturi de date de volum 10 şi 12 cu mediile* $\bar{x}_1 = 2{,}5$ *şi* $\bar{x}_2 = 2$, *respectiv* $s_1^{\,2} = 0{,}2$ şi $s_2^{\,2} = 0{,}22$.

Soluție

În 95% din cazuri intervalul va fi

$$\left((2{,}5-2) - t_{\gamma,1-\frac{\alpha}{2}} \cdot \sqrt{\frac{0{,}2}{10} + \frac{0{,}22}{12}} < \mu_1 - \mu_2 < (2{,}5-2) - t_{\gamma,1-\frac{\alpha}{2}} \cdot S \cdot \sqrt{\frac{0{,}2}{10} + \frac{0{,}22}{12}} \right),$$

cu $t_{\gamma,1-\frac{\alpha}{2}}$ determinat din *Anexa 2*, pentru $1-\frac{\alpha}{2} = 0{,}975$ şi γ determinat din relația $\frac{1}{\gamma} = \frac{c^2}{9} + \frac{(1-c)^2}{11}$ cu $c = \frac{\frac{0{,}2}{9}}{\frac{0{,}2}{9} + \frac{0{,}22}{11}}$.

Observația 10.3.7. *Dacă volumul selecțiilor este mare, se poate considera că variabilele aleatoare utilizate în Propoziția 10.3.3 şi Propoziția 10.3.5 au repartiția normală, folosindu-se astfel formula (10.3.1) în care* $\sigma_1^{\,2}, \sigma_2^{\,2}$ *se estimează prin momentele centrate de ordin 2, corespunzătoare celor două selecții. De asemenea, pentru selecții mari, formula (10.3.1) se poate utiliza şi pentru populații în care repartiția variabilei nu este normală.*

10.4. Interval de încredere pentru dispersie şi raportul a două dispersii

Un alt parametru care se poate estima prin interval de încredere este dispersia (varianța) unei variabile. De asemenea, după cum s-a putut observa în paragraful anterior, pentru a obține estimații privind diferența a două medii este util să avem informații despre raportul dintre dispersiile corespunzătoare, atunci când ele nu se cunosc, aspecte ce vor fi prezentate în acest paragraf.

Propoziția 10.4.1. *Intervalul de încredere de tip* $1-\alpha$, *pentru dispersia unei variabile X ce urmează legea normală* $N(\mu, \sigma^2)$ *cu* $\mu \in IR$ *necunoscut şi* $\sigma^2 > 0$ *necunoscut este de forma*

$$\frac{n \cdot s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} < \sigma^2 < \frac{n \cdot s^2}{\chi^2_{n-1,\frac{\alpha}{2}}} \qquad (10.4.1)$$

unde

$$s^2 = \frac{\left(X_1 - \overline{X}\right)^2 + \left(X_2 - \overline{X}\right)^2 + ... + \left(X_n - \overline{X}\right)^2}{n-1}$$

este estimația absolut corectă a lui σ^2 *pentru o selecție de volum n, bazată pe extrageri independente,* $\alpha \in (0,1)$ *este nivelul de semnificație iar* $\chi^2_{n-1,1-\frac{\alpha}{2}}$ *și* $\chi^2_{n-1,\frac{\alpha}{2}}$ *sunt cuantilele de ordin* $1-\frac{\alpha}{2}$ *și* $\frac{\alpha}{2}$ *ale repartiției* χ^2 *cu n-1 grade de libertate.*

Demonstrație

Întrucât, conform legilor de probabilitate ale variabilelor de eșantionare, cum o sumă de pătrate de *n* variabile aleatoare independente, de medie 0 și repartiție normală are o repartiție de tip χ^2 cu *n-1* grade de libertate, avem că variabila $\mathrm{X}^2 = \frac{(n-1)\cdot s^2}{\sigma^2}$ urmează legea χ^2 cu *n-1* grade de libertate, prin urmare pentru un nivel de semnificație α, găsim două numere χ_1^2 și χ_2^2, astfel încât $P\left(\chi_1^2 < \mathrm{X}^2 < \chi_2^2\right) = 1-\alpha$. Dacă alegem $\chi_1^2 = \chi^2_{n-1,\frac{\alpha}{2}}$ și $\chi_2^2 = \chi^2_{n-1,1-\frac{\alpha}{2}}$ cuantilele de ordin $1-\frac{\alpha}{2}$ și $\frac{\alpha}{2}$ ale repartiției χ^2 cu n-1 grade de libertate, obținem $\chi^2_{n-1,\frac{\alpha}{2}} < \frac{(n-1)\cdot s^2}{\sigma^2} < \chi^2_{n-1,1-\frac{\alpha}{2}}$, de unde prin artificii de calcul, rezultă intervalul dorit. *q.e.d.*

Exemplul 10.4.2. *În vederea estimării variabilității unui instrument de măsurare, se pune problema determinării unui interval de încredere de tip 98% pentru* σ^2 *pornind de la 5 măsurători independente presupuse ca fiind extrase dintr-o repartiție normală, pentru care s-a calculat* $s^2 = 2$.

Soluție

În 98% din cazuri intervalul va fi $\frac{5\cdot 2}{\chi^2_{4,0.99}} < \sigma^2 < \frac{5\cdot 2}{\chi^2_{4,0.01}}$, unde $\chi^2_{4,0.99} = 13{,}28$ și $\chi^2_{4,0.01} = 0{,}297$ rezultă din *Anexa 3*, cu valorile tabelate ale legii χ^2.

Propoziția 10.4.3. *Fie două populații studiate în raport cu variabila X, pentru prima populație variabila fiind de tipul* $N\left(\mu_1, {\sigma_1}^2\right)$ *iar pentru a doua, de tipul* $N\left(\mu_2, {\sigma_2}^2\right)$, *cu* μ_1, μ_2 *necunoscute, respectiv* ${\sigma_1}^2, {\sigma_2}^2$ *necunoscute. Fie de asemenea două selecții bazate pe extrageri independente de volume* n_1 *şi respectiv* n_2, *din cele două populații cu mediile de selecție* $\overline{X}_1, \overline{X}_2$ *şi estimatorii absolut corecți ai dispersiilor,* ${s_1}^2, {s_2}^2$. *Pentru un nivel de semnificație,* $\alpha \in (0,1)$, *intervalul de încredere pentru raportul celor două dispersii va fi*

$$\frac{1}{f_{n_1-1,n_2-1,1-\frac{\alpha}{2}}} \cdot \frac{{s_1}^2}{{s_2}^2} < \frac{{\sigma_1}^2}{{\sigma_2}^2} < \frac{1}{f_{n_1-1,n_2-1,\frac{\alpha}{2}}} \cdot \frac{{s_1}^2}{{s_2}^2} \qquad (10.4.2)$$

unde $f_{n_1-1,n_2-1,1-\frac{\alpha}{2}}$ *şi* $f_{n_1-1,n_2-1,\frac{\alpha}{2}}$ *sunt cuantilele de ordin* $1-\frac{\alpha}{2}$ *şi* $\frac{\alpha}{2}$ *ale repartiției Fisher-Snedecor cu* n_1-1 *şi* n_2-1 *grade de libertate.*

Demonstrație

Cum raportul a două variabile de tip χ^2 este o variabilă de tip Fisher, avem că variabila $F = \frac{\frac{{s_1}^2}{{\sigma_1}^2}}{\frac{{s_2}^2}{{\sigma_2}^2}}$ urmează legea Fisher cu n_1-1 şi n_2-1 grade de libertate şi raționaând ca în propoziția anterioară, se obține intervalul dorit. *q.e.d.*

Exemplul 10.4.4. *În vederea comparării comportamentului unei variabile în două populații normale, se pune mai întâi problema determinării unui interval de încredere de tip 95% pentru raportul dispersiilor în cele două populații, pornind de la două seturi de date fiecare de volum 10 şi 12 cu* ${s_1}^2 = 0{,}2$ şi ${s_2}^2 = 0{,}22$.

Soluție

În 95% din cazuri intervalul este $\frac{1}{f_{9,11,0.975}} \cdot \frac{0,2}{0,22} < \frac{\sigma_1^{\,2}}{\sigma_2^{\,2}} < \frac{1}{f_{9,11,0.025}} \cdot \frac{0,2}{0,22}$,

cu $f_{9,11,0.975} = 3,59$ determinat din *Anexa 4*, cu valorile tabelate ale legii Fisher

și $f_{9,11,0.025} = \frac{1}{f_{9,11,0.975}} = 0,278$.

Capitolul 11

Teoria deciziei

11.1. Decizii „empirice”

În viața de zi cu zi suntem nevoiți să luăm decizii – importante sau mai puțin importante, la fiecare pas. Când traversăm strada printr-un loc nepermis, culegem în mod empiric câteva informații (distanța până la cel mai apropiat vehicul, viteza cu care se apropie de noi, etc.), apoi luăm decizia de a traversa. De fiecare dată ne asumăm riscul ca decizia să fie greşită, şi uneori, decizia chiar este greşită.

11.2. Decizii statistice

Firma de automobile F trebuie să cumpere o mare cantitate de anvelope şi are de ales între mărcile **A1** şi **A2**. Înainte de a lua o decizie, firma testează **n** anvelope de tip **A1** şi **n** anvelope de tip **A2**. Având la dispoziție măsurătorile care rezultă în urma testărilor, firma decide că anvelopele **A1** sunt superioare. Această decizie se bazează pe testarea unui număr **n** de anvelope, care – prin forța lucrurilor – este mic în raport cu numărul mare de anvelope care vor fi cumpărate. Prin urmare, există riscul ca decizia să fie greşită.

Poate fi „măsurat” acest risc?

Triumful teoriei statistice a deciziilor constă tocmai în posibilitatea de a măsura gradul de risc în termenii unor probabilități obiective. Prelucrând statistic rezultatele numerice ale măsurătorilor, statisticienii firmei **F** pot comunica managerului că probabilitatea ca decizia să fie greşită este mai mică decât, să zicem, 0.05. Tinând seama de toți factorii implicați, managerul poate decide că riscul este acceptabil. Dacă probabilitatea ca decizia să fie greşită este prea mare, managerul poate cere teste suplimentare, sau poate lua în considerare parametrii suplimentari.

Când decidem să traversăm strada printr-un loc nepermis, acționăm pe baza unor probabilități subiective, şi deci gradul de risc este apreciat în mod subiectiv.

Când avem la dispoziție rezultatele numerice ale unor experimente, teoria statistică a deciziilor ne permite să măsuram gradul de risc asociat unei decizii în termenii unor probabilități obiective.

În cele ce urmează, vom prezenta anumite metode de a calcula astfel de probabilități obiective şi – implicit – de a aprecia gradul de risc asociat unei decizii. Metode de acest tip vor fi aplicate la testarea şirurilor binare cu scopul de a decide dacă sunt aleatoare.

11.3. Ipoteze statistice

Într-o accepțiune largă, prin ipoteză statistică înțelegem o ipoteză asupra unui fenomen aleator. Putem formula, de exemplu, ipoteza asupra naturii distribuției unei variabile aleatoare: normală, binomială, Poisson etc. Sau, dacă natura distribuției este precizată, putem formula ipoteze asupra valorilor numerice ale parametrilor care intervin în structura legii respective de probabilitate.

Ipoteza care urmează să fie testată se notează cu $\mathbf{H_0}$. Este necesar să formulăm şi ipoteză alternativă, notată cu $\mathbf{H_1}$.

Dacă, de exemplu, $\mathbf{H_0}$ este ipoteza că un anumit parametru **p**, are o valoare numerică $\mathbf{p_0}$, atunci $\mathbf{H_1}$ poate fi ipoteza că **p** are o alta valoare numerică $\mathbf{p_1}$. Alt exemplu ar putea fi ipoteza $\mathbf{H_1: p \neq p_1;}$ cu alte cuvinte, $\mathbf{H_1}$ poate fi pur şi simplu ipoteza că $\mathbf{H_0}$ este falsă.

Un test statistic are menirea de a recomanda acceptarea ipotezei $\mathbf{H_0}$ (şi deci respingerea lui $\mathbf{H_1}$) sau respingerea lui $\mathbf{H_0}$ (și deci acceptarea lui $\mathbf{H_1}$).

11.4. Teste statistice

Un test statistic se bazează pe un experiment în urma căruia, sub ipoteza $\mathbf{H_0}$, putem deduce valoarea numerică a unei statistici **X.** În spațiul valorilor pe care le poate lua **X** vom izola o submulțime numită *zona critică* a testului. Dacă valoarea numerică a lui **X** furnizată de experiment aparține zonei critice, vom decide că respingem ipoteza $\mathbf{H_0}$ şi acceptăm ipoteza $\mathbf{H_1}$; în caz contrar, acceptăm $\mathbf{H_0}$ şi respinge $\mathbf{H_1}$.

Întrucât rezultatul experimentului este influențat de factori aleatori, decizia noastră nu este infailibilă: ea poate fi eronată.

11.5. Tipuri de erori

Să presupunem că $\mathbf{H_0}$ este adevărată, dar că – în ciuda acestui fapt – datorită factorilor aleatori, valoarea lui **X** obținută în urma experimentului aparține zonei critice. Noi vom decide să respingem ipoteza $\mathbf{H_0}$, dar această decizie va fi evident greşită. În acest caz se spune ca am comis o *eroare de tip I.*

Cealaltă eroare posibilă este de a accepta ipoteza $\mathbf{H_0}$ când în realitate ea este falsă; aceasta se întâmplă când, deşi $\mathbf{H_0}$ este falsă, valoarea lui **X** în urma experimentului se plasează în afara zonei critice. În acest caz avem de a face cu o *eroare de tip II.*

11.6. Nivel de semnificație

Probabilitatea de a comite o eroare de tip I se notează cu $\boldsymbol{\alpha}$ şi se numeşte *nivel de semnificație* al testului. Dacă, de exemplu, $\boldsymbol{\alpha=0.05}$ şi aplicăm testul de 1000 de ori, în aproximativ 50 de cazuri vom respinge în mod eronat ipoteza $\mathbf{H_0}$.

Probabilitatea de a comite o eroare de tip II se notează cu $\boldsymbol{\beta}$.

Desigur că dorim să proiectăm teste pentru care probabilitățile de eroare $\boldsymbol{\alpha}$ şi $\boldsymbol{\beta}$ să fie mici. În anumite situații, se caută minimizarea sumei $\boldsymbol{\alpha + \beta}$. În alte cazuri se fixează nivelul de semnificație $\boldsymbol{\alpha}$ şi se caută testul pentru care $\boldsymbol{\beta}$ să fie minimă. În cazuri complexe – printre care şi testarea şirurilor binare – calculul lui $\boldsymbol{\beta}$ este dificil sau chiar imposibil, aşa că se fixează doar nivelul de semnificație $\boldsymbol{\alpha}$.

Aceste considerații sunt exemplificate în următorul exemplu.

Considerăm o anumită distribuție de probabilitate a cărei formă o cunoaştem, dar în structura căreia intra un parametru necunoscut $\boldsymbol{\theta}$.

Fie $\mathbf{H_0}$: $\boldsymbol{\theta = \theta_0}$, şi fie **K** zona critică a testului. Avem

$$\alpha = P\left(x \in K \mid H_0 \text{ este adevarata}\right) = P\left(x \in K \mid \theta = \theta_0\right)$$

Întrucât distribuția de probabilitate este complet specificată, probabilitatea de mai sus poate fi calculată.

i. Fie $\mathbf{H_1}$: $\boldsymbol{\theta = \theta_1}$.

Atunci

$$\beta = P\left(x \notin K \mid H_0 \text{ este falsa}\right) = P\left(x \notin K \mid H_1 \text{ este adevarata}\right) = P\left(x \notin K \mid \theta = \theta_1\right)$$

Din nou distribuția de probabilitate este complet specificată, deci probabilitatea $\boldsymbol{\beta}$ poate fi şi ea calculată.

ii. Fie $\mathbf{H_1}$: $\boldsymbol{\theta \neq \theta_0}$. Acum

$$\beta = P\left(x \notin K \mid \theta \neq \theta_0\right) .$$

De data aceasta, parametrul $\boldsymbol{\theta}$ nu mai este specificat, deci calculul lui $\boldsymbol{\beta}$ este dificil sau chiar imposibil.

11.7. Un exemplu

Notăm cu **X** timpul dintre două semnalizări succesive ale unui contor Geiger. Din studiul dezintegrării radioactive se ştie că variabila aleatoare **X** are densitatea exponențială**:**

$$f(t) = \begin{cases} \theta e^{-\theta t}, t > 0 \\ 0 \ , \text{ in rest} . \end{cases}$$

unde **θ** este un parametru care depinde de natura materialului radioactiv.

Un fizician doreşte să testeze valoarea lui **θ** pentru un anumit material radioactiv.

Din anumite considerente teoretice sau experimentale, el ştie că **θ** poate lua fie valoarea 1, fie valoarea 2; intuiția fizicianului favorizează valoarea 2.

Aşadar se testează ipoteza:

$$H_0 : \theta = 2$$

în prezența ipotezei alternative:

$$H_1 : \theta = 1 .$$

Pentru simplitatea exprimării, vom presupune că se face o singură observație asupra variabilei **X**, cu alte cuvinte, se măsoară lungimea unui singur interval de timp dintre două scintilații consecutive ale contorului Geiger; desigur că în practică testarea se va baza pe mai multe astfel de observații.

Notăm cu *x* valoarea măsurată a lui **X**. Alegem drept zonă critică a testului intervalul (1, +∞). Aceasta înseamnă că:

Dacă x > 1, vom respinge ipoteza $\mathbf{H_0}$ şi vom accepta ipoteza $\mathbf{H_1}$;

Dacă 0 < x ≤ 1, vom accepta ipoteza $\mathbf{H_0}$ şi vom respinge ipoteza $\mathbf{H_1}$.

Să calculăm nivelul de semnificație al testului. Avem

$$\alpha = P(x > 1 \mid H_0 \text{ adevarata}) = P(x > 1 \mid \theta = 2) .$$

Pentru **θ**=2, densitatea de probabilitate este

$$f(t) = 2e^{-2t}, t > 0$$

şi deci

$$\alpha = \int_1^{\infty} f(t)dt = 2\int_1^{\infty} e^{-2t}dt = 0.13 .$$

Aşadar, probabilitatea de a comite o eroare de tipul I (i.e., de a respinge în mod eronat ipoteza $\mathbf{H_0}$) este egală cu 0.13.

Să calculăm acum probabilitatea de a comite o eroare de tipul II (adică de a accepta eronat ipoteza $\mathbf{H_0}$, ceea ce este totuna cu a respinge eronat ipoteza $\mathbf{H_1}$).

$$\alpha = P(0 < x \le 1 \mid H_1 \text{ adevarata}) = P(0 < x \le 1 \mid \theta = 1) .$$

Pentru **θ**=1, densitatea de probabilitate este

$$f(t) = e^{-t}, t > 0 \quad ,$$

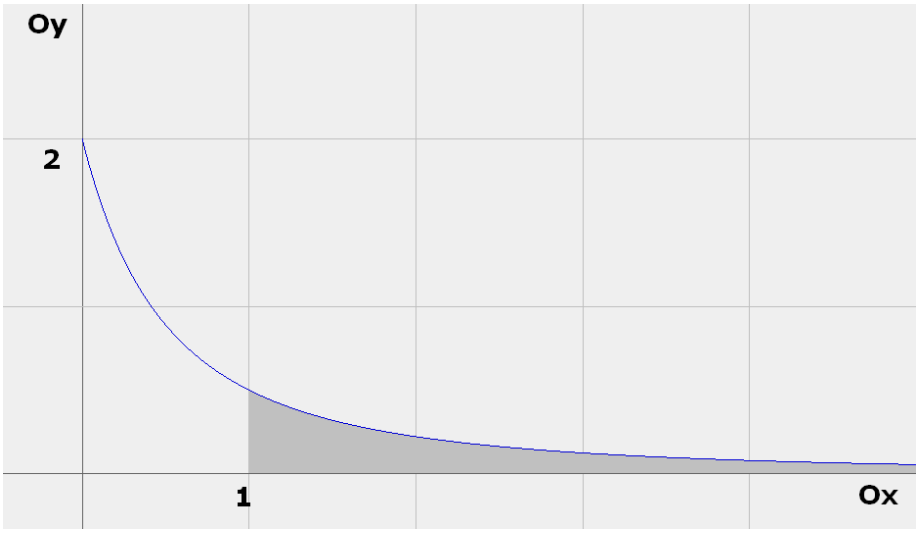
şi atunci avem

$$\beta = \int_0^1 e^{-t} dt = 0.63 .$$

Desigur, probabilitatea unei erori de tipul II este mare, dar aceasta se explică prin faptul că testul nostru se bazează pe o singura observație.

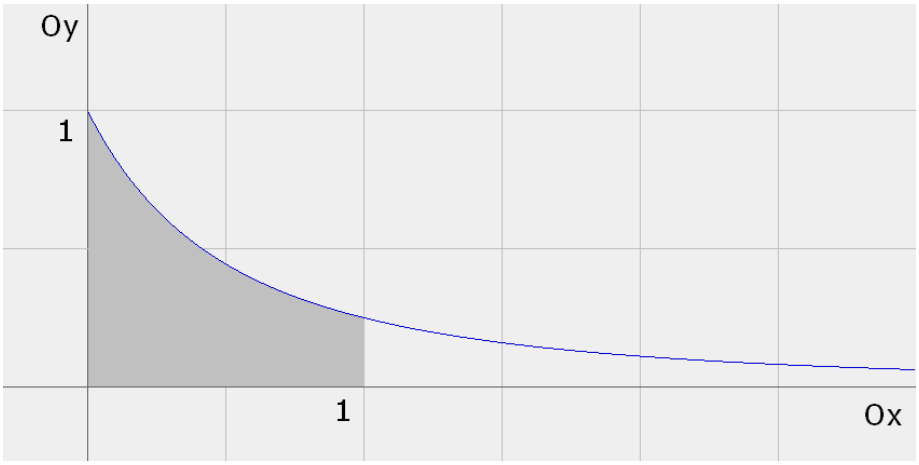
Este utilă o interpretare geometrică a faptelor de mai sus.

Pentru **θ**=2, graficul densității de probabilitate este schițat în figura de mai jos.



Zona critică este intervalul (1, +∞) de pe axa Ox. Probabilitatea ca observația **x** să aparțină zonei critice este egală cu aria suprafeței delimitate de grafic şi axa $0x$ deasupra zonei critice. Nivelul de semnificație **α** este egal cu această arie.

Pentru **θ**=1 avem graficul de mai jos:



Zona necritică este intervalul (0,1], iar probabilitatea **β** este egală cu aria delimitată de grafic şi axa $0x$ deasupra zonei necritice.

Să construim acum alt test, cu acelaşi nivel de semnificație, dar în care regiunea critică să fie un interval de forma (0,a). Aceasta înseamnă să determinăm numărul **a**>0 astfel încât

$$P\left(0 < x < a \,|\, \theta = 2\right) = 0.13 \ .$$

Condiția de mai sus se transcrie în forma echivalentă:

$$2\int_0^a e^{-2t}\,dt = 0.13$$

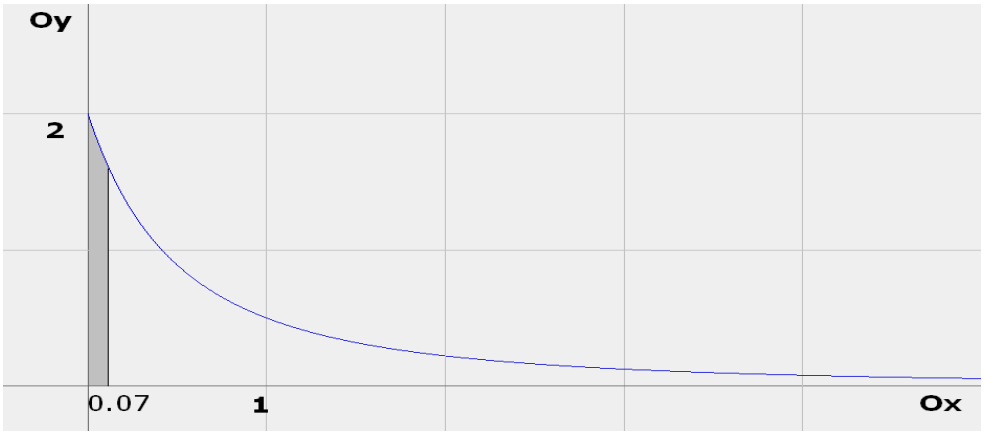
de unde se poate deduce **a**=0.07.

În aceste condiții, probabilitatea unei erori de tipul II va fi

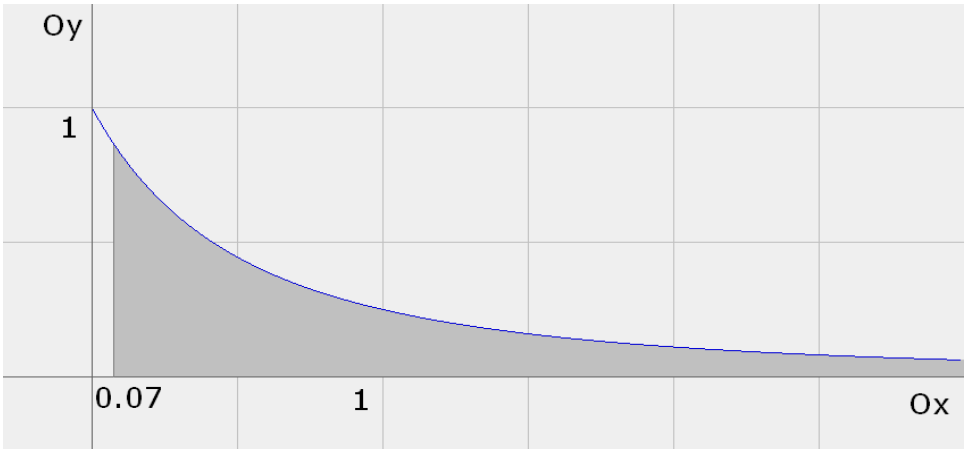
$$\beta = P\left(x > a \,|\, \theta = 1\right) = \int_{0.07}^{\infty} e^{-t} dt = 0.93 \ .$$

Constatăm acum că probabilitatea **β** este mai mare decât în situația precedentă, prin urmare, testul anterior, bazat pe zona critică $(1, +\infty)$, este superior.

Interpretarea geometrică în cazul al doilea se deduce din următoarele grafice.



Nivelul de semnificație **α** este egal cu aria suprafeței situate deasupra zonei critice $(0,0.07)$. Graficul corespunde valorii **θ**=2.



Probabilitatea **β** este egală cu aria suprafeței situate deasupra zonei necritice $(0.07, +\infty)$. Graficul este trasat pentru valoarea **θ**=1.

11.8. Relația dintre probabilitățile α şi β

În fiecare situație concretă este important să ştim care dintre cele două erori posibile ar produce cele mai mari prejudicii, şi să minimalizăm probabilitatea de a comite acea eroare. Intuitiv este clar că dacă se micşorează valoarea lui **α**, va creşte valoarea lui **β**, şi invers. Acest fapt poate fi ilustrat geometric în condițiile exemplului din secțiunea anterioară.

Să reluăm ipotezele:

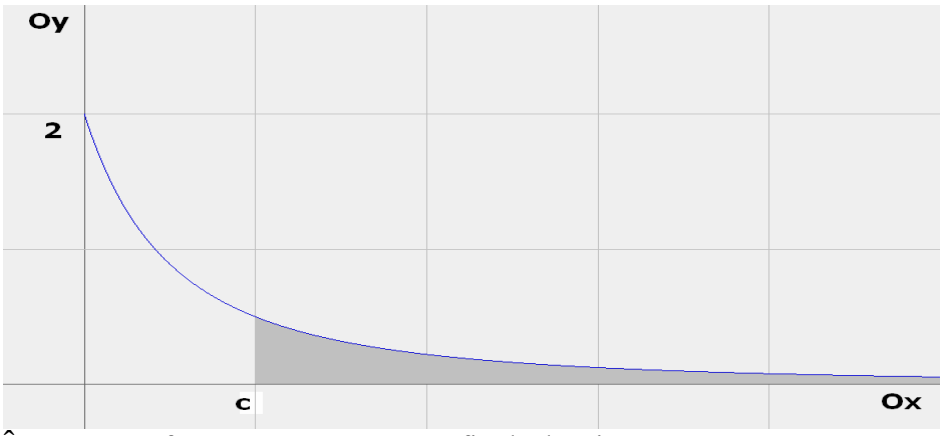
$$H_0 : \theta = 2 \ ;$$

$$H_1 : \theta = 1 \ .$$

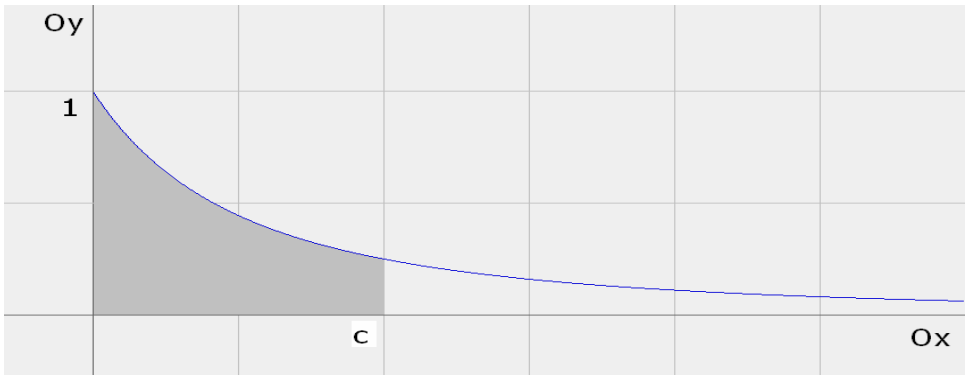
referitoare la parametrul **O** din densitatea de probabilitate

$$f(t) = \begin{cases} \theta \, e^{-\theta t}, \ t > 0 \\ 0 \ , \ t \le 0 \, . \end{cases}$$

Alegem drept zona critică a testului un interval de forma $(c, +\infty)$, unde $c > 0$.



În această figură este trasat graficul densității pentru **θ=2.** Nivelul de semnificație **α** coincide cu aria de sub grafic, la dreapta lui **c**.



Pentru **θ**=1, graficul densității de probabilitate este trasat în figura de mai sus; probabilitatea **β** coincide cu aria suprafeței de sub grafic, la stânga lui **c**.

Este clar că alegerea lui **c** determină valorile lui **α** şi **β**. A micşora pe α înseamnă a muta punctul c spre dreapta; evident că aşa îl vom mări pe **β**. Invers, a-l micşora pe **β** înseamnă a muta punctul **c** spre stânga, ceea ce îl măreşte pe **α**.

O practică des întâlnită în situații concrete este aceea de a fixa un anumit nivel de semnificație (de obicei **α**= 0.05, sau **α**= 0.01, **α**= 0.001); apoi, dintre testele cu acest nivel **α** se caută unul pentru care **β** să fie cât mai mic cu putință. Intuitiv este clar că dacă numărul de observații pe care se bazează testul creşte, atunci **β** scade; însă un număr sporit de observații poate angaja costuri suplimentare considerabile.

11.9. Puterea unui test

Să considerăm din nou densitatea de probabilitate

$$f(t) = \begin{cases} \theta\, e^{-\theta t}, & t > 0 \\ 0\ , & t \le 0\,. \end{cases}$$

Vom testa ipoteza

$$H_0 : \theta = 2 \ ;$$

în prezenta ipotezei alternative

$$H_1 : \theta < 2 \ .$$

care va înlocui ipoteza alternativă **θ=1** considerată anterior. Considerăm drept zonă critică intervalul $(1, +\infty)$, ceea ce înseamnă ca nivelul de semnificație este **α**=0.13.

De data aceasta, ipoteza alternativă $\mathbf{H_1}$ nu mai specifică o valoare cunoscută a lui **θ,** așa încât nu mai putem indica o valoare numerică a probabilității **β**; putem însă determina expresia lui **β** ca funcție de variabila $\theta \in (0,2)$.

Într-adevăr $\beta(\theta)$ este probabilitatea ca x să aparțină zonei necritice când valoarea parametrului este **θ**

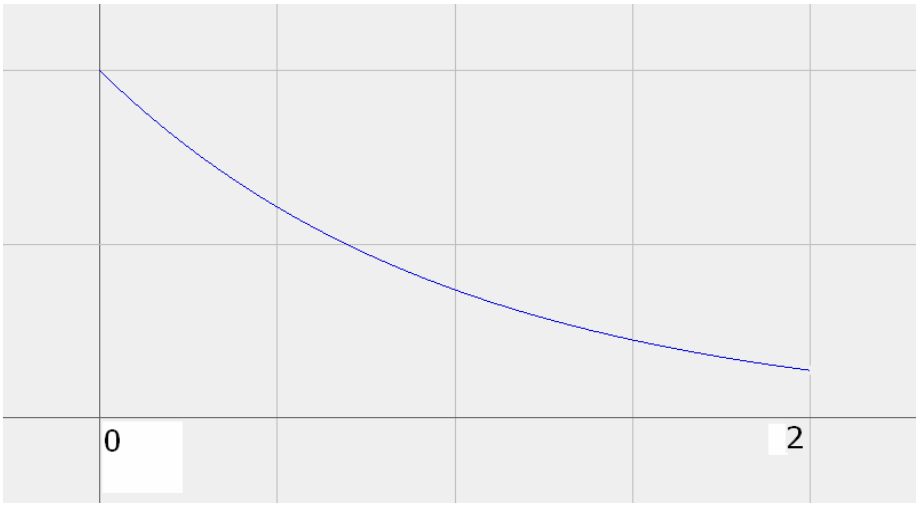
$$\beta(\theta) = \int_0^1 \theta e^{-\theta t} dt = -e^{-\theta t} \Big|_0^1 = 1 - e^{-\theta} \ .$$

Așadar, probabilitatea ca noi să acceptăm în mod eronat ca valoare a parametrului numărul 2, când adevărata valoare este **θ,** va fi egală cu $1 - e^{-\theta}$.

În general, funcția $p(\theta) = 1 - \beta(\theta)$ se numește *funcția de putere* a testului, sau *puterea testului*. Această funcție descrie probabilitatea ca testul să respingă ipoteza $\mathbf{H_0}$ atunci când ea este eronată.

În cazul exemplului de mai sus avem:

$P(\theta) = 1 - e^{-\theta}$; graficul acestei funcții este schițat mai jos.



Pentru fiecare valoare **θ** a parametrului, funcția de putere **P(θ)** descrie probabilitatea ca x să aparțină zonei critice; în particular, **P(2) = 0.13**.

Dacă **θ** = 1, valoarea eronată 2 este respinsă cu probabilitatea 0.37, iar dacă **θ** =0.5, aceeași valoare eronată 2 este respinsă cu probabilitate 0.61.

11.10. Încă un exemplu

Ni se dă o monedă şi ni se cere să decidem dacă este sau nu trucată. Avem de testat ipoteza

H_0 : moneda nu este trucată

în prezența ipotezei alternative

H_1 : moneda este trucată

Astfel formulate, ipotezele **H_0** şi **H_1** sunt de natură calitativă; le putem transcrie sub o formă cantitativă, numerică, dacă notăm cu **p** probabilitatea de a obține fața A la o aruncare a monedei (şi deci **q** = **1 - p** va fi probabilitatea de a obține fața B).

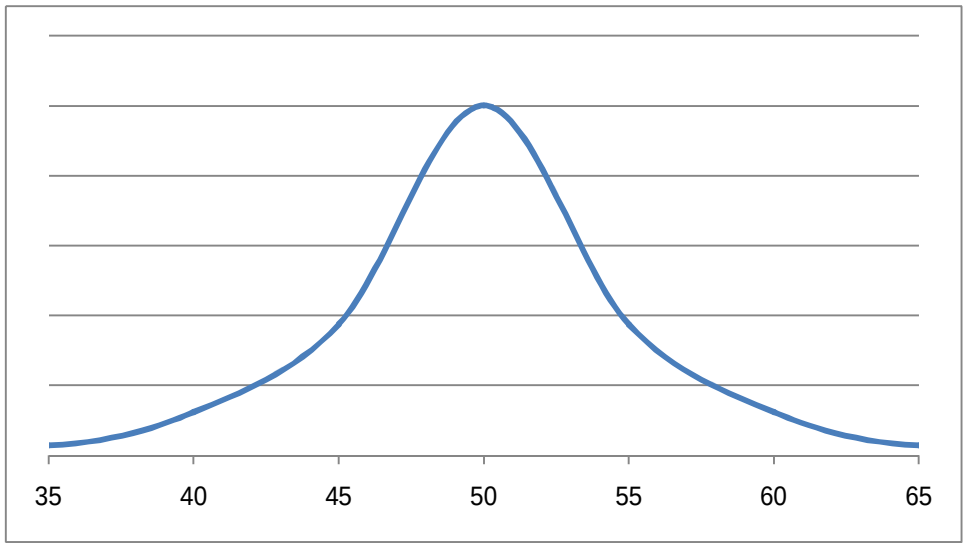
Atunci putem scrie:

$$H_0 : p = \frac{1}{2},$$

$$H_1 : p \neq \frac{1}{2}$$

Testul va consta din a arunca moneda de 100 de ori. Să notăm prin **X** numărul de apariții ale feței **A** în cele 100 de aruncări. Sub ipoteza **H_0**, **X** este o variabilă aleatoare repartizată binomial cu parametrii $n = 100$ şi $p = \frac{1}{2}$. Media lui **X** va fi **np** = 50, iar dispersia **npq** = 25.

Aproximând distribuția binomială printr-o distribuție normală cu media 50 şi dispersia 25, densitatea lui **X** va fi aproximativ cea schițată in figura următoare.



Vom determina zona critică în aşa fel încât nivelul de semnificație să fie 0.05. Folosind tabele pentru legea normală deducem

$$P(40 \leq x \leq 60) = 0.95$$

Înseamnă că zona necritică va fi intervalul [40,60], iar cea critică exteriorul acestui interval.

Aşadar testul nostru poate fi rezumat astfel:

1. Dacă fața **A** apare de un număr de ori mai mic decât 40 sau mai mare decât 60, moneda este trucată. Probabilitatea ca această concluzie să fie greşită este mai mică decât 0.05.
2. Dacă fața **A** apare de un număr de ori cuprins între 40 şi 60, nu există suspiciuni (la nivelul de semnificație 0.05) că moneda ar fi trucată.

Să presupunem acum că în realitate **p** = 0.7, dar noi nu ştim asta. Atunci distribuția reală a lui **X** va fi aproximativ normală cu media 70 şi dispersia 21. Atunci probabilitatea **β** pentru testul nostru va fi

$$\beta = P(40 \leq x \leq 60 \mid p = 0.7) = 0.02$$

Folosind funcția de putere, avem

$$P(0.7) = 1 - \beta(0.7) = 0.98$$

Alte valori ale funcției de putere, calculate în mod similar, sunt

$$P(0.2) = P(0.8) = 1$$
$$P(0.3) = P(0.7) = 0.98$$
$$P(0.4) = P(0.6) = 0.5$$

Cu alte cuvinte, dacă **p**=0.3 testul nostru va respinge ipoteza greşită **p**=0.5 cu probabilitatea 0.98.

11.11. Testarea şirurilor binare

Avem în vedere şiruri finite (numite şi secvențe) formate cu simbolurile 0 şi 1. Un astfel de şir aleator poate fi interpretat ca rezultat al aruncărilor unei monede netrucate având fețele notate cu 0 şi 1. Aruncările sunt independente unele de altele şi rezultatele aruncărilor până la un anumit moment nu influențează în nici un fel rezultatele aruncărilor viitoare.

Acest experiment ideal este neconvenabil pentru scopuri practice. În practică şirurile binare sunt produse de generatoare, şi ele urmează să fie testate din punct de vedere al caracterului aleator.

11.12. Testarea statistică a şirurilor binare

Să considerăm un şir binar care urmează să fie testat. Vom formula ipoteza

H_0 : şirul dat este aleator

şi ipoteza alternativă

H_1 : şirul dat nu este aleator.

Alegem un nivel de semnificație **α**, adică probabilitatea de a comite o eroare de tip I. O valoare mare a lui **α** indică un risc mare ca testul să respingă ipoteza $\mathbf{H_0}$ când ea este în realitate adevărată; cu alte cuvinte, un risc mare ca testul să declare drept nealeatoare şiruri care au fost produse în mod aleator.

O eroare de tip II înseamnă în cazul de față să acceptăm drept aleator un şir produs de un generator imperfect. Probabilitatea **β** de a comite o astfel de eroare depinde de natura imperfecțiunii generatorului, şi este dificil de estimat în practică. În mod curent se consideră că o valoare prea mică a lui **α** măreşte riscul unei erori de tip II, cu alte cuvinte măreşte riscul de a accepta drept aleatoare şiruri produse de un generator imperfect.

Este deci important să alegem nivelul de semnificație **α** adecvat problemei concrete pe care o avem de rezolvat. În practică se foloseşte un nivel de semnificație **α** cuprins între 0.001 şi 0.05; se alege de multe ori **α**= 0.01.

Fiecare test se bazează pe o statistică **X** a cărei valoare numerică se calculează pornind de la şirul considerat. De obicei se aleg statistici care pot fi calculate în mod eficient şi care urmează o lege normală sau χ^2.

Valoarea **x** a statisticii **X** pentru şirul dat se compară cu valoarea aşteptată de la un şir aleator.

a. Să presupunem că statistica **X** cu care lucrăm este distribuită **N(0,1)**, şi că ia fie valori foarte mici, fie valori foarte mari pentru şirurile nealeatoare. Folosind tabele pentru legea normală fixăm un prag $\mathbf{x_\alpha}$ astfel încât

$$P(X \geqslant x_\alpha) = P(X \leqslant -x_\alpha) = \frac{\alpha}{2}$$

Zona critică a testului va fi $(-\infty, -x_\alpha) \cup (x_\alpha, +\infty)$. Dacă valoarea **x** a lui **X**, calculată pentru şirul considerat, aparține zonei critice, şirul este considerat nealeator; cu alte cuvinte, ipoteza $\mathbf{H_0}$ este respinsă la nivelul de semnificație **α**.

Dacă $x \in [-x_\alpha, x_\alpha]$, com accepta ipoteza $\mathbf{H_0}$; nu sunt suspiciuni (la nivel de semnificație **α**) că şirul ar fi produs de un generator imperfect.

De exemplu, dacă **α**=0.05, atunci $\mathbf{x_\alpha}$ = 1.96; probabilitatea ca un □ir aleator să fie respins ca nealeator este de 0.05.

b. Să presupunem acum că statistica **X** este distribuită χ^2 cu **γ** grade de libertate, şi că ia valori foarte mari pentru şirurile nealeatoare. Pragul $\mathbf{x_\alpha}$ se determină (folosind tabele pentru legea χ^2) din condiția

$$P(X \geqslant x_\alpha) = \alpha.$$

Zona critică a testului va fi $(x_\alpha, +\infty)$. Dacă valoarea **x** a lui **X**, calculată pentru şirul testat, aparține zonei critice, ipoteza $\mathbf{H_0}$ va fi respinsă la nivel de semnificație **α**, adică şirul va fi considerat nealeator. Dacă $x \in [0, x_\alpha]$ acceptăm ipoteza $\mathbf{H_0}$; nu sunt suspiciuni că şirul ar fi nealeator.

De exemplu, dacă **γ**=5 şi **α**=0.025, atunci $\mathbf{x_\alpha}$=12.83; probabilitatea ca un şir aleator să fie declarat, în mod eronat, ca nealeator este de 0.025.

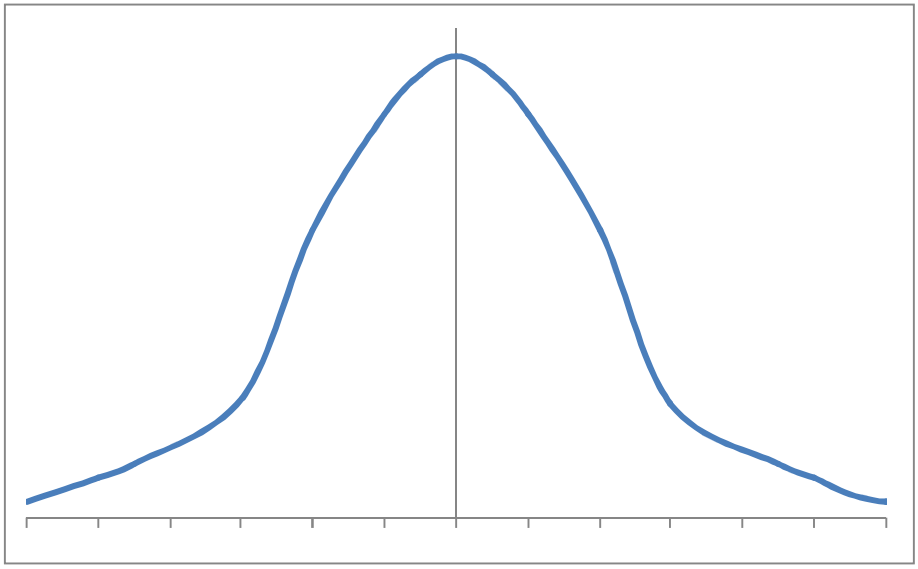
11.13. Noțiunea de P-valoare

Uneori nu se lucrează cu pragul $\mathbf{x}_{\alpha}$, ci se preferă folosirea aşa-numitei *P-valori*. Vom prezenta această noțiune în cadrul exemplelor (a) şi (b) de mai sus.

a. Fie **x** valoarea numerică a lui **X** pentru şirul testat. Probabilitatea

$$P(|X| > |x|)$$

se numeşte *P-valoare.*



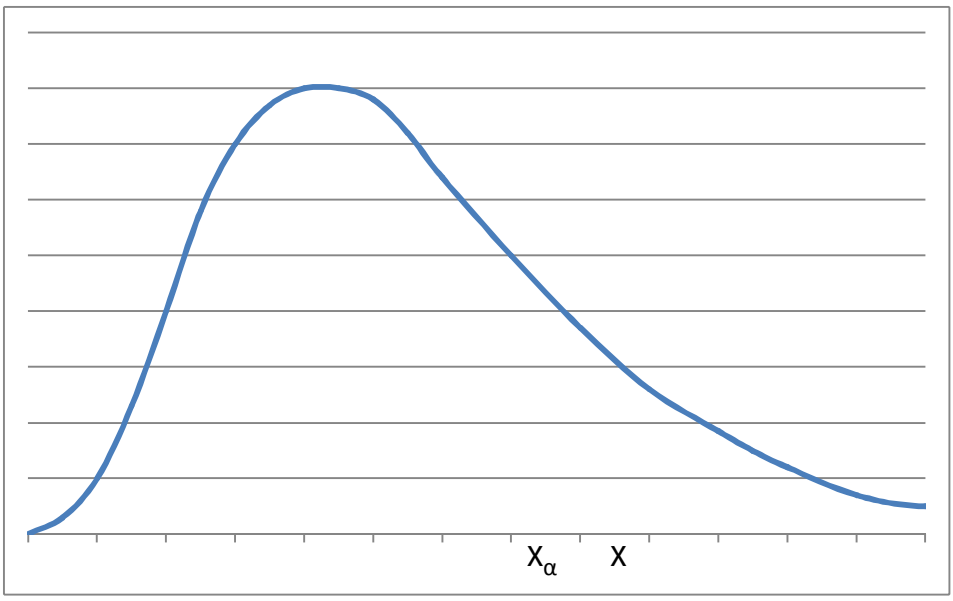
Examinând figura de mai sus deducem că urmatoarele afirmații sunt echivalente.

1) **X** aparține zonei critice
2) $|x| > x_{\alpha}$
3) $P(|X| > |x|) < P(|X| > x_{\alpha})$
4) $P(|X| > |x|) < \alpha$

Din echivalența condițiilor (1) şi (4) deducem că ipoteza $\mathbf{H_0}$ va fi respinsă dacă şi numai dacă P-valoarea este mai mică decât $\boldsymbol{\alpha}$.

Aşadar şirul testat va fi respins ca nealeator (la nivelul de semnificaşie $\boldsymbol{\alpha}$) dacă şi numai dacă P-valoarea calculată pentru el este mai mică decât $\boldsymbol{\alpha}$.

b. În condițiile acestui exemplu, descrise în secțiunea precedentă, fie x valoarea numerică a lui **X** pentru şirul testat. Probabilitatea $P\left(|X| > |x|\right)$ se numeşte *P-valoare.*



Examinând această figură conchidem că următoarele afirmații sunt echivalente:

1) **x** aparține zonei critice
2) **$x > x_\alpha$**
3) $P(X > x) < P(X > x_\alpha)$
4) $P(X > x) < \alpha$.

Echivalența condițiilor (1) şi (4) arată că ipoteza $\mathbf{H_0}$ va fi respinsă la nivelul de semnificație **α** dacă şi numai dacă P-valoarea este mai mică decât **α**. şirul testat va fi respins ca nealeator dacă şi numai dacă P-valoarea calculată pentru el este mai mică decât **α**.

Iată, în rezumat, două concluzii.

1. Dacă testul a fost proiectat pentru nivelul de semnificație **α**=0.001, din 1000 de şiruri produse aleator, el va respinge, în medie, unul singur ca fiind nealeator. Dacă pentru un şir dat P-valoarea este mai mică decât 0.001, şirul va fi declarat nealeator, iar probabilitatea ca această decizie să fie greşită este 0.001.
2. Din 1000 de şiruri aleatoare, un test cu nivelul de semnificație α=0.01 va respinge, în media, 10 şiruri ca fiind nealeatoare. Un şir cu P-valoarea mai mică decât 0.01 va fi declarat nealeator, iar nivelul nostru de încredere în corectitudinea acestei decizii este 99%.

11.14. Un exemplu: statistică repartizată normal

Am observat deja mai sus că aruncând o monedă netrucată generăm un şir binar aleator. Prin urmare, având de testat un şir dat, ne putem imagina că el a apărut ca rezultat al aruncărilor unei monede, şi rămâne să testăm dacă moneda este netrucată.

O astfel de testare a fost descrisă anterior, la nivelul de semnificație **α**=0.05; tot acolo am văzut ce fel de considerații pot fi făcute în legătură cu probabilitatea **β** (de a accepta ca aleator un şir care de fapt nu este aleator) şi în legătură cu puterea testului.

În cazul specific al şirurilor binare, metoda respectivă de testare poate fi descrisă astfel. Considerăm un şir binar de lungime **n**. Formulăm ipoteza

H_0 : şirul este aleatoriu

Statistica $\mathbf{S_n}$ va indica numărul de apariții ale cifrei 1. Sub ipoteza $\mathbf{H_0}$, $\mathbf{S_n}$ este repartizată binomial, cu parametrii **n** şi $\frac{1}{2}$, adică

$$P(S_n = k \mid p = \frac{1}{2}) = \binom{n}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{n-k} = \binom{n}{k}2^{-n}, \qquad k = 0,1,2,\ldots,n.$$

Notăm $X = \frac{S_n - \frac{n}{2}}{\sqrt{n/4}}$.

Atunci **X** este repartizată aproximativ N(0,1), adică

$$P(X > z) = P(X < -z) = 1 - \Phi(z),$$

unde

$$\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{t^2}{2}}\,dt, \quad z>0.$$

Fixând nivelul de semnificație **α**, vom determina pragul $\mathbf{x_\alpha}$ din relația

$$1 - \Phi(x_\alpha) = \frac{\alpha}{2}$$

De exemplu, $\mathbf{x_{0.05}} = 1.96$, iar $\mathbf{x_{0.01}} = 2.58$.

Pentru un şir dat, să notăm cu **x** valoarea numerică a lui **X**:

$$x = \frac{S_n - \frac{n}{2}}{\sqrt{n/4}}.$$

Atunci P-valoarea asociată şirului va fi

$$P(|X| > |x|) = 2(1 - \Phi(|x|))$$

Cu scop ilustrativ, să considerăm exemplul şirului binar

1011010101

Să fixăm nivelul de semnificație $\boldsymbol{\alpha} = 0.01$, ceea ce determină pragul $\mathbf{x_\alpha} = 2.58$.

Avem $\mathbf{n}=10$, $\mathbf{S_n}=6$, deci

$$x = \frac{6-5}{\sqrt{10/4}} = \frac{2}{\sqrt{10}} = \frac{\sqrt{10}}{5} = 0.63$$

Întrucât $0 < \mathbf{x} < \mathbf{x_\alpha}$, acceptăm ipoteza că şirul este aleator.
De altfel putem calcula şi *P-valoarea* pentru acest şir, ea este

$$2(1 - \Phi(0.63)) = 0.52 > 0.01$$

Să considerăm acum şirul binar

1111011111

Avem $\mathbf{n}=10$, $\mathbf{S_n}=9$, deci

$$x = \frac{9-5}{\sqrt{10/4}} = \frac{8}{\sqrt{10}} = \frac{4\sqrt{10}}{5} = 2.52$$

P-valoarea calculată pentru acest şir este

$$2\left(1-\Phi(2.52)\right)=0.012$$

şirul trece (aproape la limită) testul cu nivelul de semnificație 0.01, dar este respins de testul cu nivelul de semnificație 0.05.

11.15. Alt exemplu: statistică repartizată χ^2

Fie **M** şi **N** numere naturale fixate. Considerăm un şir binar de lungime **n** = **MN**, pe care îl împărțim in **N** blocuri consecutive de lungime **M**. Notăm cu $\mathbf{M_i}$ numărul de cifre 1 în blocul *i*, $i \in \{1,2,\dots,N\}$.

Sub ipoteza

$\mathbf{H_0}$: şirul este aleator,

$\mathbf{M_i}$ este o variabilă aleatoare binomială cu parametrii **M** şi ½ ; altfel spus,

$$P(M_i=k\,|H_0)=\binom{M}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{M-k}=\binom{M}{k}2^{-M}$$

Media lui $\mathbf{M_i}$ este $\frac{M}{2}$, iar dispersia $\frac{M}{4}$.

În aceste condiții variabila aleatoare

$$X=\sum_{i=1}^{N}\left(\frac{M_i-\frac{M}{2}}{\sqrt{M/4}}\right)^2$$

este repartizată aproximativ cu **N** grade de libertate; altfel spus, densitatea ei de probabilitate este funcția

$$f(t)=\frac{1}{2^{N/2}\Gamma(N/2)}t^{N/2-1}e^{-t/2},t>0$$

Este comod să notăm cu $\pi_i=\frac{M_i}{M}$ frecvența relativă a cifrei 1 in blocul *i*. Atunci putem scrie

$$X=4M\sum_{i=1}^{N}(\pi_i-\frac{1}{2})^2$$

Această variabilă aleatoare χ^2cu **N** grade de libertate poate fi folosită la testarea şirului, aşa cum se arată în exemplul de mai sus.

Cu scop ilustrativ, să considerăm şirul

011001101

Alegem **M**=3, **N**=3, şi considerăm blocurile

011, 001, 101

Cu notațiile anterioare,

$$\pi_1 = \frac{2}{3}, \pi_2 = \frac{1}{3}, \pi_3 = \frac{2}{3}$$

Acum putem calcula valoarea **x** a variabilei **X**:

$$x = 12\left(\left(\frac{2}{3} - \frac{1}{2}\right)^2 + \left(\frac{1}{3} - \frac{1}{2}\right)^2 + \left(\frac{2}{3} - \frac{1}{2}\right)^2\right) = 1.$$

Lucrăm cu 3 grade de libertate; din tabele găsim că *P-valoarea* şirului este 0.80, deci şirul este considerat aleator la nivelele de semnificație 0.01 şi 0.05. Aceeaşi concluzie se obține comparând valoarea **x**=1 cu pragurile $\mathbf{x_{0.01}}$ = 11.341 şi $\mathbf{x_{0.05}}$ = 7.815.

Capitolul 12

Analiza regresiei

Introducere

Analiza de regresie îşi are originile în nenumăratele probleme practice, care apar atunci când dorim să înţelegem şi să cuantificăm aspectul cauză-efect, în studiul a două sau mai multe fenomene, de natură diversă. Principala noţiune cu care operează acest capitol este noţiunea de model de regresie. Vom vedea în cele ce urmează câteva generalităţi ale problemei regresiei, prezentându-se principalele tipuri de modele de regresie, apoi se va fundamenta modelul liniar multiplu, incluzând particularităţile modelului liniar simplu, estimarea coeficienţilor modelului prin metoda celor mai mici pătrate, inferenţa asupra modelului în ipotezele Gauss-Markov, precum şi aspecte privind previziunea pe baza modelului de regresie.

12.1. Modele de regresie

Să considerăm, spre exemplu, că fiecare element al unei populaţii statistice posedă o caracteristică numerică, *X* şi o alta *Y*. Pentru a vedea cum afectează valorile lui *X*, realizările variabilei *Y*, este necesară studierea posibilei corelaţii existente între cele două variabile. Un exemplu clasic este acela care studiază înălţimea unei persoane, în funcţie de cea a tatălui.

În cazul legăturilor statistice, care conţin ca şi caz particular, aferent dependenţei totale, legătura funcţională, unei singure valori, *x*, a variabilei *X*, i se asociază o repartiţie de valori a variabilei *Y*, de medie $f(x)$, $x \in D$, D fiind mulţimea valorilor variabilei *X*.

Definiţia 12.1.1. *Dacă pentru fiecare valoare,* $x \in D$*, a lui X , Y este o variabilă aleatoare cu distribuţia de probabilitate depinzând de x, vom numi* ***funcţie de regresie a lui Y pe X****, funcţia* $f(x)$*, definită cu ajutorul valorii medii condiţionate,*

$$f(x) = E(Y|x), x \in D. \qquad (12.1.1)$$

Ţinând cont de caracteristicile valorii medii condiţionate, o legătură directă între Y şi X va fi dată atunci, de ***modelul de regresie simplă***

$$Y = f(X) + \varepsilon, \tag{12.1.2}$$

unde ε *satisface condiţiile* $E(\varepsilon) = 0$ şi $Var(\varepsilon)$ - *minimă* şi *are semnificaţia de* ***eroare de specificare****, eroare datorată faptului că variabila/variabilele luate în considerare nu sunt singure suficiente, pentru a explica în totalitate, fenomenul cuantificat de Y.*

Vom considera în cele ce urmează, aşa cum se întâmplă şi în practică, mai multe variabile cauză (variabile exogene, predictori, regresori), $X_1, X_2, ..., X_p$, pentru variabila efect, *Y* (variabilă endogenă).

Definiţia 12.1.2. ***Modelul de regresie multiplă*** *este modelul de forma*

$$Y = f(X_1, X_2, ..., X_p) + \varepsilon, \tag{12.1.3}$$

unde ε *reprezintă o variabilă aleatoare, pentru care* $E(\varepsilon) = 0$ *şi* $V(\varepsilon)$ *mică.*

Deşi $X_1, X_2, ..., X_p$ sunt considerate variabile deterministe, *Y* împrumută de la ε, caracterul aleator, termenul eroare, ε, fiind cel care transformă modelul matematic, strict funcţional, în unul statistic.

Odată specificat un model de regresie, este necesar să determinăm, sau măcar să estimăm, funcţia de regresie *f*, pe baza unor date de selecţie. Astfel, spre exemplu, pentru un model de regresie simplă, se porneşte de la datele $(x_i, y_i), i = \overline{1,n}$, care reprezintă de fapt, o selecţie de volum *n*, pentru variabilele *X* şi *Y* şi se obţine modelul observaţional $y_i = f(x_i) + \varepsilon_i, i = \overline{1,n}$. E bine de ştiut faptul că, ε_i poate cuprinde pe lângă erorile de specificare şi erori de observare, în cazul în care valorile variabilei *Y* şi eventual, ale variabilei *X*, ne sunt puse la dispoziţie, în urma unor măsurători, posibil afectate de mici erori. Desigur, se presupune că în modelul de regresie nu intră erorile sistematice, ci doar cele aleatoare. Din punct de vedere al punctului de plecare, în procesul de estimare a funcţiei de regresie *f*, deosebim regresia parametrică şi regresia neparametrică, cea din urmă nefiind obiectivul acestui capitol. Dacă în determinarea funcţiei de regresie se pleacă de la ideea (desigur pe cât posibil fundamentată), că funcţia f are o anumită formă atunci vorbim de regresia parametrică.

Definiţia 12.1.3. *Modelul de regresie în care funcţia este de forma*

$$f(X_1, X_2, ..., X_p) = f(X_1, X_2, ..., X_p; \alpha_1, \alpha_2, ..., \alpha_q),\ \alpha_i \in IR,\ i = \overline{1,q}, \tag{12.1.4}$$

se numeşte ***model de regresie parametrică****.*

Plecând de la definiție, se observă că un model de regresie parametrică presupune cunoscută forma funcției de regresie, *f* (mai bine zis a estimatorului căutat), excepție făcând un număr finit de parametri necunoscuți, adică $f(\cdot)=f(\cdot,\alpha)$, cu $\alpha=(\alpha_1,...,\alpha_q)'\in B\subseteq IR^q$. Evident, dacă $f(\cdot,\alpha)$ este cunoscută, estimarea lui *f*, într-un model de regresie parametrică, revine la estimarea lui α. Folosind metode de estimare adecvate bazate pe minimizarea erorii din model, cum ar fi ***criteriul celor mai mici pătrate,*** e posibil să se estimeze, din date, vectorul α şi implicit *f*. Reprezentarea grafică a estimatorului funcției *f*, obținut prin astfel de metode, va fi o curbă care ajustează, cel mai bine datele, din mulțimea de curbe permise, prin specificarea modelului.

Modelele de regresie parametrică pot depinde, într-o manieră liniară sau neliniară, de parametri.

Definiția 12.1.4. *Vom spune că avem o* ***regresie liniară,*** *dacă funcția f este liniară în variabilele* $X_1,X_2,...,X_p$ *adică asupra funcției de regresie facem presupunerea că are forma*

$$f(X_1,X_2,...,X_p;\alpha_1,\alpha_2,...,\alpha_p)=\sum_{k=1}^{p}\alpha_k X_k\ . \qquad (12.1.5)$$

Orice altă formă a funcției f *presupune* ***regresie neliniară.***

Forma liniară a funcției de regresie şi metoda celor mai mici pătrate utilizată în scopul estimării parametrilor sunt cele mai des întâlnite, în analiza regresională. Modelele liniare sunt cele mai simple şi mai utilizate modele, multe dintre modelele neliniare şi chiar neparametrice, făcând apel la caracteristicile acestora. Tehnicile de estimare punctuală şi inferențială, utilizate în determinarea modelului liniar, țin de un domeniu important în analiza regresională şi anume, regresia liniară.

Există însă o grupă de modele neliniare, care pot fi tratate tot prin intermediul tehnicilor regresiei liniare şi anume, modelele neliniare liniarizabile, o parte dintre acestea fiind şi modelele liniare în parametri.

Definiția 12.1.5. *Se numeşte* ***model de regresie liniarizabil în parametri,*** *modelul în care funcția de regresie este de forma*

$$f(X_1,X_2,...,X_p;\alpha_1,\alpha_2,...,\alpha_q)=\sum_{k=1}^{q}\alpha_k\varphi_k(X_1,X_2,...,X_p), \quad (12.1.6)$$

adică este presupusă liniară, în raport cu parametrii $\alpha_1,...,\alpha_p$.

Astfel de modele pot fi liniarizate prin substituțiile,

$$\varphi_k\left(X_1, X_2, ..., X_p\right) = Z_k,\ k = \overline{1,q},$$

un exemplu fiind ***modelul de regresie polinomială***, în care

$$f\left(X;\alpha_0,\alpha_1,...,\alpha_q\right) = \alpha_0 + \alpha_1 X + ... + \alpha_q X^q \qquad (12.1.7)$$

şi din care pentru *q=1* derivă ***modelul de regresie liniară simplă***, dat prin

$$f\left(X;\alpha_0,\alpha_1\right) = \alpha_0 + \alpha_1 X\ . \qquad (12.1.8)$$

Un astfel de model este şi ***modelul hiperbolic***, cu funcţia de regresie $f\left(X;\alpha_1,\alpha_2\right) = \alpha_1 + \frac{\alpha_2}{X}$, liniarizabil prin substituţia $\frac{1}{X} = Z$. Tot din grupa modelelor neliniare, dar liniarizabile, fac parte şi modelele care se reduc la modelul liniar, în urma mai multor operaţii: logaritmare, substituţie, etc. Un exemplu este ***modelul exponenţial***, dat de funcţia $f\left(X;\alpha_1,\alpha_2\right) = \alpha_1 \cdot {\alpha_2}^X$, care se liniarizează prin logaritmarea $\log f\left(X;\alpha_1,\alpha_2\right) = \log\alpha_1 + X \cdot \log\alpha_2$ şi substituţiile $F\left(X,A,B\right) = \log f\left(X;\alpha_1,\alpha_2\right)$ şi $A = \log\alpha_1, B = \log\alpha_2$, obţinându-se modelul liniar dat prin $F(X,A,B) = A + X \cdot B$. Există însă şi alte modele neliniare, care nu pot fi liniarizate, cum ar fi de exemplu, modelul $Y = \alpha_0 + \alpha_1 X^{\alpha_2} + \varepsilon$. Este deja bine cunoscută formularea, că modelele neliniare, liniarizabile prin substituţie, adică acelea în care se presupune că funcţia de regresie este liniară în parametri, ţin de regresia liniară, deoarece studiul lor se face cu tehnicile acesteia. Modelele neliniare în parametri, dar liniarizabile în urma unor operaţii, cum ar fi logaritmarea, pot fi tratate, atât cu tehnici ale regresiei liniare, cât şi cu cele ale regresiei neliniare. Aplicarea regresiei liniare pe modelul transformat are avantajul că estimatorii se bucură de proprietăţi mai bune şi dezavantajul că modelul obţinut este doar o aproximare a celui iniţial (a se vedea cazul modelului exponenţial, în care erorile intră aditiv în modelul iniţial). Modelele neliniarizabile ţin exclusiv de regresia neliniară, regresie în care tehnicile nu mai pot fi fundamentate, pe avantajele obţinute din liniaritate.

Pentru modelele neliniare, sistemul care derivă din criteriul celor mai mici pătrate fiind neliniar, se întâmpină de cele mai multe ori, dificultăţi de rezolvare, motiv pentru care, atunci când modelul este liniarizabil, se preferă mai întâi liniarizarea lui, care va duce la un sistem liniar de ecuaţii normale şi nu aplicarea directă a criteriului, deşi în acest fel, se obţine doar o aproximare rezonabilă a modelului iniţial. În cazul modelelor care nu pot fi liniarizate, se aplică tehnici ale regresiei neliniare, bazate în special pe metode iterative, cum ar fi metoda iterativă Gauss - Newton.

12.2. Modelul liniar. Estimarea parametrilor modelului prin metoda celor mai mici pătrate

În acest paragraf, ne ocupăm de aspecte privind ajustarea modelului liniar. Sunt amintite forma teoretică, forma observațională/matriceală şi forma ajustată a modelului liniar, precum şi condiția care garantează existența soluției ajustării de cele mai mici pătrate.

Definiția 12.2.1. *Se numeşte* ***model regresional liniar multiplu****, între variabila Y şi variabilele* $X_1, X_2, ..., X_p$*, modelul*

$$Y = \sum_{k=1}^{p} \alpha_k X_k + \varepsilon . \qquad (12.2.1)$$

Problema regresiei liniare constă în studiul comportării variabilei *Y*, în raport cu factorii $X_1, X_2, ..., X_p$, în ipoteza (12.2.1). Acest studiu revine la evaluarea parametrilor (coeficienților) de regresie, $\alpha_1, \alpha_2, ..., \alpha_p$ şi a termenului aleator, ε. Estimarea coeficienților de regresie se face pe baza unei selecții de volum *n*. Pentru datele de selecție (şi atunci când este cazul pentru variabilele de selecție), vom folosi următoarele notații:

$$y' = (y_1, y_2, ..., y_n) \in IR^n, x = (x_1, ..., x_p) = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}, \; n > p . \qquad (12.2.2)$$

În cazul în care analistul are control asupra alegerii variabilelor, $X_1, X_2, ..., X_p$, matricea *x* se numeşte matrice de design. Pentru parametrii $\alpha_k, k = \overline{1, p}$, şi pentru erorile $\varepsilon_i, i = \overline{1, n}$, corespunzătoare datelor de selecție, vom folosi de asemenea, notațiile matriceale :

$$\alpha' = (\alpha_1, \alpha_2, ..., \alpha_p) \in IR^p \;,\; \varepsilon' = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n) \in IR^n . \qquad (12.2.3)$$

Pentru datele de selecție corespunzătoare lui „i”, modelul (12.2.1) devine modelul observațional (cu datele observate),

$$y_i = \sum_{k=1}^{p} \alpha_k x_{ik} + \varepsilon_i , \qquad (12.2.4)$$

ceea ce, pentru $i=\overline{1,n}$, duce la forma matriceală a modelului liniar observațional,

$$y = x\alpha + \varepsilon\,, \tag{12.2.5}$$

în care y, x, α şi ε sunt cele din notațiile (12.2.2) şi (12.2.3). În vederea estimării lui α, se ajustează modelul (12.2.5), printr-o condiție de minim asupra erorii, care aşa cum am subliniat încă din paragraful introductiv al acestui capitol, este de dorit să fie mică. Ne vom opri aici, doar asupra ***ajustării prin criteriul celor mai mici pătrate,*** care constă în minimizarea expresiei

$$\varepsilon'\varepsilon = \sum_{i=1}^{n} \varepsilon_i^2\ . \tag{12.2.6}$$

Definiția 12.2.2. *Se numeşte* ***model liniar, ajustat prin criteriul celor mai mici pătrate****, modelul*

$$y = xa + e\,, \tag{12.2.7}$$

unde $a' = (a_1, a_2, ..., a_p) \in IR^p$ *realizează minimul expresiei (12.2.6), iar* $e'e$*, cu* $e' = (e_1, e_2, ..., e_n) \in IR^n$ */este valoarea minimă obținută.*

Sistemul de ecuații, la care revine condiția de minim, este

$$x'xa = x'y \tag{12.2.8}$$

şi se numeşte ***sistemul de ecuații normale (Gauss),*** ataşat modelului (12.2.7). Notațiile a şi e desemnează estimatori punctuali ai lui α şi ε, atunci când $y_i, i=\overline{1,n}$, $x_{ik}, i=\overline{1,n}, k=\overline{1,p}$, desemnează variabilele de selecție şi estimații punctuale (valori nenule), atunci când prin $y_i, i=\overline{1,n}$, $x_{ik}, i=\overline{1,n}, k=\overline{1,p}$, înțelegem date de selecție. Obținerea estimatorilor de cele mai mici pătrate a, pentru coeficienții de regresie necunoscuți α, depinde aşadar, de existența inversei matricei $x'x$, care revine la condiția $rang(x'x) = p$. Se cunoaşte următorul rezultat, care dă condiții de existență (unică) a estimatorilor de cele mai mici pătrate (a se vedea de exemplu [30]).

Teorema 12.2.3. *Dacă* $rang(x) = p$*, atunci soluția ajustării prin criteriul celor mai mici pătrate este dată de formula*

$$a = (x'x)^{-1} x'y\,. \tag{12.2.9}$$

Condiția $rang(x) = p$ revine la independența liniară a vectorilor $x_1, x_2, ..., x_p$. Amintim în continuare, câteva noțiuni utilizate şi în teoria modelelor de regresie oarecare.

Definiția 12.2.4. *Se numeşte* ***valoare ajustată*** *a lui y (în modelul liniar), valoarea* $\hat{y}' = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n) \in IR^n$, *definită de* $\hat{y} = xa$. *Se numeşte* ***matrice de influență*** *a modelului, matricea H care transformă valoarea y, în valoarea ajustată* $\hat{y}$, *adică,* $\hat{y} = Hy$, matricea de influență în modelul liniar fiind de forma $H = x(x'x)^{-1}x'$. *Se numeşte* ***reziduu,*** *valoarea* $e = y - \hat{y} = (I - H)y$.

Un caz particular al modelului liniar, care face mai uşor trecerea la modelul liniar simplu (cu o singură variabilă exogenă), este modelul liniar cu termen constant.

Definiția 12.2.5. *Se numeşte* ***model liniar cu termen constant****, un model liniar în care una dintre variabile este înlocuită de constanta 1.*

Modelul observațional, scris pe baza datelor de selecție în forma matriceală, va arăta atunci astfel,

$$y = x_0\alpha_0 + u\alpha_p + \varepsilon, \tag{12.2.10}$$

cu notațiile $\alpha_p \in IR$, $x_0 = (x_1, ..., x_{p-1}) \in M_{n,p-1}$, $\alpha_0' = (\alpha_1, \alpha_2, ..., \alpha_{p-1}) \in IR^{p-1}$, $u' = (1, ..., 1) \in IR^n$. Dacă notăm $x = (x_0 : u)$, $\alpha = (\alpha_0', \alpha_p)$, modelul poate fi scris în forma modelului liniar oarecare, $y = x\alpha + \varepsilon$. O teoremă similară cu *Teorema 12.2.3* are loc şi în cazul modelului liniar cu termen constant.

Teorema 12.2.6. *Fie matricea de centrare,* $P = I - \frac{1}{n}uu'$. *Notăm cu* $\tilde{z}$, *vectorul centrat corespunzător lui* $z \in IR^n$, *adică* $\tilde{z} = Pz = (z_1 - \bar{z}, z_2 - \bar{z}, ..., z_n - \bar{z})'$, *cu* $\bar{z}$, *media de selecție. Dacă* $rang(x) = p$ *şi* $x = (x_0 : u)$, *ajustarea prin metoda celor mai mici pătrate are soluția unică dată de*

$$a_0 = (a_1, a_2, ..., a_{p-1})' = (\tilde{x}_0'\tilde{x}_0)^{-1}\tilde{x}_0'\tilde{y},$$
$$a_p = \bar{y} - \sum_{k=1}^{p-1} a_k \overline{x_k}, \tag{12.2.11}$$

unde $\tilde{x}_0$ *este matricea obținută din matricea* x_0, *prin centrarea vectorilor de pe coloană, iar* $\bar{y}$ *şi* $\overline{x_k}$, *notează mediile de selecție corespunzătoare valorilor* y_i *respectiv,* $x_{ik}, i = \overline{1,n}$.

Observația 12.2.7. *Dacă în Definiția 12.2.5, se consideră* $p = 2$, *obținem* ***modelul de regresie liniară simplă,***

$$Y = \alpha X + \beta + \varepsilon, \tag{12.2.12}$$

care cu ajutorul datelor de selecție se scrie,

$$y_i = \alpha x_i + \beta + \varepsilon_i, \quad i = \overline{1,n}, \; y_i, x_i, \varepsilon_i, \alpha, \beta \in IR.$$

În acest caz, relațiile (12.2.11) dau reprezentarea estimatorilor *a* şi *b*, ai coeficienților α şi β şi anume,

$$a = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{\operatorname{cov}(x,y)}{S_x^2},$$
$$b = \bar{y} - a\bar{x}, \tag{12.2.13}$$

unde $\operatorname{cov}(x,y)$ reprezintă covarianța între *x* şi *y*, iar S_x^2, varianța lui *x*.

Exemplul 12.2.8. *Teoria economică privind gestiunea portofoliului susține că rentabilitatea unei acțiuni este influențată de modificarea indicelui general al bursei, adică de evoluția pieței în general (modelul de piață Sharp-Markowitz, ([17])). Pornind de la această idee, se pune problema determinării estimatorilor de cele mai mici pătrate pentru coeficienții unui model liniar simplu care să descrie corelația dintre rata rentabilității unor acțiuni, Y şi rata rentabilității pieței, X. Se va considera un eşantion de 15 zile, cu valorile:*

Y(%): -2,8; 0,2; 1,6; 3,9; 0,2; 2,4; 4,4; 18,6; 1,5; -17,9; 0,5; 0; 0,8; 0,1; -0,8
X(%): -0,5; 1,5; 3,2; 5,8; 3,3; 7,4; 5,8; 16,0; -2,0; -13,1; 0,7; 0,9; 2,9; 1,8; 1,1.

Soluție

Pe baza eşantionului, obținem $\bar{y} = 0{,}85$, $\bar{x} = 2{,}32$, $\operatorname{cov}(x,y) = 41$, $S_x^2 = 36$, prin urmare, conform formulelor de calcul pentru a şi b, avem

$a = \frac{41}{36} = 1{,}13$ şi $b = 0{,}85 - \frac{41}{36} \cdot 2{,}32 = -1{,}76$ şi modelul ajustat este $Y = -1{,}76 + 1{,}13 \cdot X + \varepsilon$.

12.3. Modelul liniar clasic Gauss - Markov. Inferențe asupra estimatorilor unui model liniar

În acest paragraful tratăm aspectul probabilist al regresiei, cercetând calitățile estimatorilor de cele mai mici pătrate, calități obținute sub anumite ipoteze de natură probabilistă, făcute asupra erorii. Sunt amintite ipotezele clasice, calitățile estimatorilor în aceste ipoteze, câteva statistici utile în inferența asupra coeficienților, precum şi intervalele de încredere privind coeficienții şi de asemenea, câteva din testele cunoscute, referitoare la coeficienți.

În continuare, pe tot parcursul acestui capitol, se păstrează notațiile referitoare la forma teoretică, forma matriceală şi forma ajustată a modelului, precum şi condiția $rang(x) = p$.Vom începe prin a preciza ***ipotezele clasice*** în care se lucrează într-un model liniar, ipoteze care, deşi nu neapărat de neînlocuit, duc la bune proprietăți ale estimatorilor. Aceste ipoteze se referă la distribuția erorilor şi anume,

$$E(\varepsilon) = \theta,\ \theta = (0,0,...,0)' \in IR^n, \tag{12.3.1}$$

$$Var(\varepsilon) = E(\varepsilon \cdot \varepsilon') = \sigma^2 I, \tag{12.3.2}$$

$$\varepsilon \in N \tag{12.3.3}$$

sau altfel spus,

$$\varepsilon \in N(\theta, \sigma^2 I). \tag{12.3.4}$$

Condiția (12.3.2) se poate scrie şi sub forma relațiilor

$$Var(\varepsilon_i) = \sigma^2, \forall i = \overline{1,n}, \tag{12.3.5}$$

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j, i, j = \overline{1,n}. \tag{12.3.6}$$

Condițiile puse asupra erorii se transferă asupra variabilei aleatoare, adică avem $y \in N(x\alpha, \sigma^2 I)$.

Definiția 12.3.1. *Ipotezele (12.3.1) şi (12.3.2) sunt cunoscute şi sub denumirea de* ***ipoteze Gauss-Markov****, modelul liniar sub aceste ipoteze, numindu-se* ***modelul liniar Gauss-Markov****. Pentru eroarea care satisface aceste ipoteze se foloseşte şi denumirea de* ***zgomot alb****. Datorită ipotezei (12.3.5), modelul se numeşte* ***homoscedastic****, iar datorită ipotezei (12.3.6),* ***model cu erori necorelate****. Dacă la ipotezele Gauss-Markov, se adaugă şi ipoteza normalității (12.3.3), atunci modelul liniar este cunoscut şi sub denumirea de* ***model liniar clasic****, cele trei ipoteze fiind apelate ca* ***ipotezele liniarității modelului****, deşi acestea nu țin neapărat de un model liniar. Un* ***model clasic*** *(liniar sau nu) este de fapt, un* ***model cu erori normale*** *((12.3.3)),* ***independente*** *((12.3.6)) şi* ***identic distribuite*** *((12.3.1) şi (12.3.5)) sau prescurtat* ***i.i.d.***

Înainte de a aminti principalele rezultate, cu privire la calitatea estimatorilor de cele mai mici pătrate, a şi e, pentru α şi ε, obținute sub ipotezele (12.3.4), vom sublinia faptul că y, ε, a şi e sunt vectori aleatori, în timp ce α este un vector determinist şi de asemenea, $x_1, x_2, ..., x_p$, din matricea variabilelor de selecție, $x = (x_1, x_2, ..., x_p)$, sunt vectori determinişti.

Teorema 12.3.2. *În ipotezele Gauss-Markov, au loc următoarele afirmații:*
i)Estimatorul de cele mai mici pătrate, a, al lui α *, este nedeplasat,* $E(a) = \alpha$ *şi are varianța* $Var(a) = \sigma^2 (x'x)^{-1}$*. De asemenea,* $Var(\hat{y}) = \sigma^2 H$ *, unde H este matricea de influență a modelului.*
ii)Estimatorul a, al lui α *, este liniar în observațiile lui y.*
iii)Estimatorul a al lui α *este optimal, adică oricare alt estimator* $\tilde{a}$ *, nedeplasat pentru* α *şi liniar în observațiile lui y, are varianța mai mare decât varianța lui a, în sensul că,* $Var(a_k) \le Var(\tilde{a}_k)$*,* $k = \overline{1, p}$*.*

Teorema 12.3.3. *În ipotezele Gauss-Markov, estimatorii*

$$s^2 = \frac{1}{n-p} \sum_{i=1}^{n} e_i^2 = \frac{1}{n-p} e'e \ , \qquad S = s^2 (x'x)^{-1} \tag{12.3.7}$$

sunt estimatori nedeplasați pentru σ^2*, respectiv* $Var(a)$*.*

În plus, au loc şi următoarele proprietăți:

Propoziția 12.3.4. *În ipotezele Gauss-Markov, avem:*
i)Estimatorul e al lui ε *este de varianță* $Var(e) = \sigma^2 Q$*, unde* $Q = I - H$*, H matricea de influență ataşată modelului.*

ii)Dacă $c \in M_{q,p}$, *atunci ca este estimator optimal pentru* $c\alpha$, *în clasa estimatorilor nedeplasaţi, care sunt transformări liniare ale lui y.*

iii)Estimatorii a şi e sunt necorelaţi, adică $\operatorname{cov}(a,e) = \theta$, $\theta = (0,0,...,0)' \in IR^n$.

Ipoteza normalităţii erorilor aduce noi proprietăţi ale estimatorilor.

Propoziţia 12.3.5. *Într-un model liniar clasic (cu erori normale şi i.i.d.-(12.3.4)), estimatorul obţinut prin metoda celor mai mici pătrate este un estimator eficient pentru* α.

Atunci când se cunoaşte legea de probabilitate a variabilei *y* (prin intermediul legii erorilor), putem vorbi şi despre estimatori de verosimilitate maximă, respectiv de regresie de verosimilitate maximă. Într-un model liniar, supus ipotezelor clasice (12.3.4), estimatorul lui α, de cele mai mici pătrate, coincide cu estimatorul lui α, de verosimilitate maximă, în timp ce estimatorul de verosimilitate maximă, $S^2_{y/x}$, pentru σ^2 este nedeplasat, unde :

$$S^2_{y/x} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n} e'e = \frac{n-p}{n} s^2. \qquad (12.3.8)$$

De asemenea, în ipoteza normalităţii, ataşată ipotezelor Gauss-Markov, se pot stabili şi următoarele proprietăţi, care furnizează statistici ce se vor dovedi utile în inferenţa estimatorilor.

Propoziţia 12.3.6. *Dacă* $\varepsilon \in N(\theta, \sigma^2 I)$, *atunci avem*

i)Estimatorii a şi s^2 *sunt independenţi şi* $a \in N\left(\alpha, \sigma^2 (x'x)^{-1}\right)$.

ii) $\aleph^2 = (n-p)\dfrac{s^2}{\sigma^2} = \dfrac{1}{\sigma^2}\sum_{i=1}^{n} e_i^2 \in \aleph^2(n-p)$.

iii)Fie $U = \dfrac{\delta' a - \delta'\alpha}{\sigma\sqrt{\delta'(x'x)^{-1}\delta}}$, $\delta' = (\delta_1,...,\delta_p) \in IR^p$. *Oricare ar fi* $\delta \in IR^p$, U *şi* s^2 *sunt independente şi* $U \in N(0,1)$. *Mai mult,* $T = \dfrac{\delta' a - \delta'\alpha}{s\sqrt{\delta'(x'x)^{-1}\delta}} \in T(n-p), \forall \delta \in IR^p$.

Aceste proprietăţi se bazează pe următorul rezultat din teoria probabilităţilor ([30]):

Lema 12.3.7. *Fie vectorul* $\varepsilon \in IR^n$, *care urmează legea normală, de medie* θ *şi matrice de varianţă,* $\sigma^2 I$; *fie de asemenea, matricele* $Q \in M_{n,n}(IR)$, $Q' = Q$,

$Q^2 = Q$, $r = rang(Q)$, $L \in M_{n,n}(IR)$, $LQ = \theta$. *Atunci,* $\frac{1}{\sigma^2}\varepsilon' Q\varepsilon \in \aleph^2(r)$, $L\varepsilon \in N$ *şi vectorul aleator* $L\varepsilon$, *împreună cu variabila aleatoare* $\frac{1}{\sigma^2}\varepsilon' Q\varepsilon$, *sunt independenţi.*

Statisticile amintite în *Propoziţia 12.3.6,* pot servi la realizarea inferenţei asupra estimatorilor de cele mai mici pătrate. Începând de aici, în tot restul paragrafului, vom presupune adevărate ipotezele modelului liniar clasic. Ne vom referi, pentru început, la inferenţa prin intervale de încredere.

Interval de încredere pentru coeficienţi

Dacă în statistica *T* din *Propoziţia 12.3.6.iii,* se particularizează $\delta' = \delta_k' = (0,0,...,1,...,0) \in IR^p$, se obţine statistica Student, cu $n - p$ grade de libertate,

$$T_k = \frac{a_k - \alpha_k}{s_k}, \tag{12.3.9}$$

unde s_k^2 este produsul dintre s^2 şi elementul al *k*-lea, diagonal, al matricei $(x'x)^{-1}$. Pe baza acesteia, se poate construi un interval de încredere pentru coeficienţi, de forma

$$P\left(a_k - s_k \cdot t_{n-p,1-\frac{\varphi}{2}} < \alpha_k < a_k + s_k \cdot t_{n-p,1-\frac{\varphi}{2}}\right) = 1 - \varphi, \tag{12.3.10}$$

unde $t_{n-p,1-\frac{\varphi}{2}}$ este cuantila de ordin $1-\frac{\varphi}{2}$, a unei variabile Student, cu $n - p$ grade de libertate, care rezultă pentru un nivel de semnificaţie φ, fixat, din relaţia

$$P\left(|T_k| < t_{n-p,1-\frac{\varphi}{2}}\right) = 1 - \varphi, \tag{12.3.11}$$

Desigur, în cazul în care σ este cunoscut, nu mai este nevoie de operaţia de studentizare şi atunci, se poate folosi statistica *U,* din *Propoziţia 12.3.6 iii.* Pot fi de asemenea elaborate regiuni de încredere, pentru $x\alpha \in IR^n$ şi $\alpha \in IR^p$, regiuni pe care nu le vom aminti aici. Atunci când σ este necunoscut, pe lângă estimaţia punctuală *s,* se poate da şi un interval de încredere pentur varianţă.

Interval de încredere pentru varianță

Pe baza statisticii $\aleph^2$, cu $n-p$ grade de libertate, din *Propoziția 12.3.6.ii*, se poate construi un interval de încredere pentru σ^2, de forma

$$P\left((n-p)\frac{s^2}{\aleph^2_{n-p,1-\frac{\varphi}{2}}} < \sigma^2 < (n-p)\frac{s^2}{\aleph^2_{n-p,\frac{\varphi}{2}}}\right) = 1-\varphi, \qquad (12.3.12)$$

unde $\aleph^2_{n-p,1-\frac{\varphi}{2}}$ şi $\aleph^2_{n-p,\frac{\varphi}{2}}$ sunt cuantilele de ordin $1-\frac{\varphi}{2}$, respectiv $\frac{\varphi}{2}$, ale unei variabile $\aleph^2$, cu $n-p$ grade de libertate, determinate astfel încât pentru un nivel de semnificație φ, fixat, să aibă loc relația

$$P\left(\aleph^2_{n-p,\frac{\varphi}{2}} < \aleph^2 < \aleph^2_{n-p,1-\frac{\varphi}{2}}\right) = 1-\varphi. \qquad (12.3.13)$$

În ultima parte a acestui paragraf, vom prezenta un al doilea aspect al inferenței asupra estimatorilor şi anume, testările de ipoteze asupra coeficienților.

Testul T pentru coeficienții unui model liniar

Ipotezele acestui test sunt ipoteza nulă, $H_0 : \alpha_k = \alpha_k^{(0)}$ şi alternativa ei, $H_1 : \alpha_k \neq \alpha_k^{(0)}$, ceilalți coeficienți fiind în afara ipotezelor.

Testul se fundamentează pe o statistică de tip Student, cu $n-p$ grade de libertate, T_k, precizată în formula (12.3.9). Pentru nivelul de semnificație φ, rezultă cuantila $t_{n-p,1-\frac{\varphi}{2}}$, de ordin $1-\frac{\varphi}{2}$, a unei variabile Student, cu $n-p$ grade de libertate aşa încât,

$$P\left(|T_k| < t_{n-p,1-\frac{\varphi}{2}} \middle| H_0\right) = 1-\varphi.$$

Se calculează valoarea $t_k = \frac{a_k - \alpha_k^{(0)}}{s_k}$, a statisticii T_k, pe baza datelor de selecție şi se respinge ipoteza nulă, dacă $|t_k| > t_{n-p;1-\frac{\varphi}{2}}$.

Un caz particular al acestui test este cel al semnificației coeficientului α_k, test bazat pe ipotezele $H_0 : \alpha_k = 0$ şi $H_1 : \alpha_k \neq 0$.

Următoarele două teste clasice, pe care le vom prezenta, se bazează pe o statistică de tip Fisher-Snedecor, furnizată de următorul rezultat din teoria probabilităților ([19]).

Lema 12.3.8. *Fie vectorul aleator* $\varepsilon' = (\varepsilon_1,...,\varepsilon_n)$*, ce urmează legea normală, cu* $\varepsilon \in N(\theta, \sigma^2 I)$ *şi fie matricele* $x \in M_{n,p}(IR)$, $A \in M_{p,q}(IR)$, $x_0 = xA \in M_{n,q}(IR)$. *Dacă se notează* $Q = I - x(x'x)^{-1}x'$ *şi* $Q_0 = I - x_0(x_0'x_0)^{-1}x_0'$, *atunci statistica*

$$F = \frac{\varepsilon' Q_0 \varepsilon - \varepsilon' Q \varepsilon}{rang(Q_0) - rang(Q)} \Big/ \frac{\varepsilon' Q \varepsilon}{rang(Q)} \tag{12.3.14}$$

urmează legea de probabilitate Fisher-Snedecor, cu $rang(Q_0) - rang(Q)$ *şi* $rang(Q)$ *grade de libertate.*

Vom nota pe tot parcursul acestui paragraf,

$$S_0^2 = \varepsilon' Q_0 \varepsilon \quad \text{şi} \quad S_1^2 = \varepsilon' Q \varepsilon . \tag{12.3.15}$$

Se observă că, dacă matricea *X* din lemă este matricea datelor de selecție dintr-un model liniar, atunci avem $Q = I - H$, unde *H* este matricea de influență. Mai mult, vom avea $S_0^2 = \|e_0\|^2 = e_0'e_0$ şi $S_1^2 = \|e\|^2 = e'e$. În cazul formulării unei ipoteze H_0, asupra coeficienților unui model liniar, ipoteză în cadrul căreia matricea datelor de selecție devine de forma X_0 din lemă, notațiile S_0^2 şi S_1^2 reprezintă suma pătratelor reziduurilor, în modelul redus (obținut în ipoteza H_0), respectiv suma pătratelor reziduurilor, în modelul complet (obținut în ipoteza H_1).

Testul F al egalității între q coeficienți

Se testează ipoteza nulă, $H_0 : \alpha_1 = ... = \alpha_q, q \leq p$, cu alternativa, H_1 :"H_0 – falsă". Pentru un nivel fixat φ, se determină valoarea $f = f_{q-1;n-p;1-\varphi}$, a unei statistici Fisher-Snedecor, cu $q-1$ şi $n-p$ grade de libertate, astfel încât,

$$P\left(F < f \middle| H_0\right) = 1 - \varphi .$$

Ipoteza nulă se va respinge, dacă $f_c > f$, unde $f_c = \frac{n-p}{q-1} \cdot \frac{S_0^2 - S_1^2}{S_1^2}$ este o valoare calculată a statisticii *F* din *Lema 12.3.8*, pe baza datelor de selecție.

Testul F al semnificației a *q* coeficienți

Se testează ipoteza nulă, $H_0 : \alpha_1 = ... = \alpha_q = 0,\ q \le p$ cu alternativa, H_1 :"H_0-falsă". Pentru un nivel fixat φ, se determină o valoare $f = f_{q;n-p;1-\varphi}$, a unei statistici Fisher-Snedecor, cu q şi $n - p$ grade de libertate, astfel încât,

$$P\left(F < f \middle| H_0\right) = 1 - \varphi .$$

Ipoteza nulă se va respinge, dacă $f_c > f$, unde $f_c = \frac{n-p}{q} \cdot \frac{S_0^2 - S_1^2}{S_1^2}$ este o valoare a statisticii *F* din *Lema 12.3.8*, calculată pe baza datelor de selecție.

Exemplul 12.3.9. *Reluând datele din Exemplul 12.2.8, ne propunem să determinăm intervalele de încredere de tip 95% pentru coeficienții modelului liniar simplu.*

Soluție

Aplicând formula (12.3.10), intervalele de încredere de tip 95% pentru coeficienții modelului estimați în Exemplul 12.2.8, vor fi (0.9156, 1.336) pentru *a* şi (-3.086, -0.4451) pentru *b*.

12.4. Previziunea şi analiza rezultatelor unei regresii liniare

Ne propunem, în acest paragraf, să amintim principalele aspecte care țin de previziunea şi analiza rezultatelor unei regresii liniare, deşi, marea parte a aspectelor şi statisticilor considerate aici sunt valabile şi în cazul altor modele. ([5]). Elaborarea unui model de regresie are ca scop, pe lângă determinarea unui mecanism, care să copieze comportamentul dependenței studiate şi acela de a putea previziona, adică de a obține o estimație cât mai bună, pentru o valoare y_0, corespunzătoare unor noi date pentru variabilele $x_1, x_2, ..., x_p$. O astfel de previziune primeşte credit, atunci când specificarea modelului de regresie din care se obține, este corectă. Astfel, înainte de a realiza previziuni asupra variabilei endogene *y*, este necesar să ne asigurăm că ipotezele făcute asupra modelului, în special asupra erorii, sunt valide. Vom discuta pe rând, în acest

paragraf, ipotezele respective, vom aminti tehnicile de verificare a lor, precum şi posibilităţile de corectare a situaţiilor în care ipotezele nu sunt valide. De asemenea, vom aminti câteva statistici care se folosesc pentru a analiza calitatea ajustării datelor, prin model, precum şi intervalul de încredere pentru previziune.

Controlul ipotezelor liniarităţii modelului

În literatură, sub denumirea de ipoteze ale liniarităţii modelului, se întâlnesc de fapt, ipotezele clasice, definite în paragraful anterior (erori normale de medie zero, independente şi identic distribuite), la care se adaugă ipoteza necorelaţiei între erori şi variabilele exogene, $X_1, X_2, ..., X_p$ şi desigur, absenţa corelaţiei între variabilele exogene (absenţa multicoliniarităţii). Verificarea ultimei ipoteze ţine de tehnici ale corelaţiei, care fac obiectul acestui capitol. Nevalidarea acestei ipoteze duce la erori mari în model de aceea, pentru a nu compromite din start modelul, se încearcă satisfacerea acestei condiţii. Ţinând cont de aceste două aspecte, nu vom mai relua aici această ipoteză. E bine de specificat că, deşi au denumirea de ipoteze ale liniarităţii, sunt de fapt presupuneri care nu au legătură cu caracterul liniar al modelului, de aceea le putem întâlni şi la alte modele, unde vor fi analizate în mod asemănător.

Vom analiza în continuare ipotezele rămase, în cadrul unui model liniar. Deoarece aceste ipoteze se referă la erorile din model, pentru verificarea lor se determină modelul, se calculează reziduurile $e_i, i = \overline{1,n}$, precizate în *Definiţia 12.2.4* şi se analizează aceste reziduuri, presupunând că ele constituie de fapt, estimatori pentru erorile $\varepsilon_i, i = \overline{1,n}$.

Ipoteza normalităţii

Ipoteza $\varepsilon \in N$ este necesară pentru obţinerea unor estimatori eficienţi ai coeficienţilor şi de asemenea, pentru obţinerea unor estimatori ce urmează legea normală. Verificarea normalităţii erorilor se poate face prin teste de concordanţă, fie prin intermediul testului lui Massey (vezi [15]), atunci când *n* este mic, fie prin testul Kolmogorov-Smirnow (vezi [23]), atunci când *n* este mare. Ambele teste se bazează pe compararea frecvenţelor cumulate empirice, cu frecvenţele teoretice, corespunzătoare legii normale. Ipoteza normalităţii, deşi necesară, nu este crucială atunci când volumul eşantionului este mare. ***Teorema centrală limită*** (vezi [23]) ne asigură că, atunci când $n \to \infty$, deşi ε nu urmează legea normală, avem că estimatorul de cele mai mici pătrate *a*, al lui α, converge la legea normală. Mai precis, se introduce factorul de normalizare $\sqrt{n}$ şi se obţine că, atunci când *n* este mare, $\sqrt{n}(a-\alpha)$ converge asimptotic, către o variabilă normală, $N(0, n \cdot Var(a))$, de unde $a \underset{n\to\infty}{\to} N(\alpha; Var(a))$. Pentru eşantioane mici, dacă se respinge ipoteza normalităţii erorilor, se reprezintă reziduurile şi se calculează coeficienţii de asimetrie şi boltire ai lui Fisher, relativ la e_i, pentru a

aprecia mărimea deviației de la normalitate. Se vor elimina acele observații pentru care reziduurile sunt foarte mari, aşa încât reziduurile rămase să se apropie mai mult de valori normale şi se va reface modelul doar cu noile observații.

Ipoteza zgomotului alb

Presupunerea $E(\varepsilon)=0$ este necesară în obținerea de estimatori *a*, nedeplasați, adică $E(a)=\alpha$. Această ipoteză arată de asemenea, faptul că erorile din model nu sunt erori sistematice. Verificarea acestei ipoteze se poate realiza prin analiza grafică a reziduurilor sau prin analiza (grafică sau numerică) a intervalelor de încredere, pentru media erorii. Astfel, se reprezintă grafic reziduurile, fiind necesar ca valorile acestora să oscileze în jurul dreptei de ecuație, $y=0$. Pentru prezentarea intervalelor de încredere pentru media erorii, vom aminti mai întâi, noțiunea de reziduu studentizat. Conform *Propoziției 12.3.4.i*, reziduurile sunt în general corelate şi variațele lor depind de locația punctelor. În vederea obținerii unei statistici Student, utilă în testele de ipoteze, precum şi în intervalele de încredere referitoare la erori, este necesară studentizarea reziduurilor, proces care constă în acest caz, într-o scalare care implică obținerea aceleiaşi variațe pentru reziduuri. În [29], se defineşte reziduul studentizat sub forma:

Definiția 12.4.1. *Se numeşte* ***reziduu studentizat****, expresia*

$$t_i=\frac{e_i}{s\sqrt{1-h_i}}\ , \tag{12.4.1}$$

care înainte de eşantionare este o variabilă aleatoare ce urmează legea Student, $T(n-p)$*. Elementul* h_i *este elementul diagonal al matricei de influență H, iar s este eroarea standard a estimației, definită în formula (12.3.7).*

În literatura de specialitate, se întâlnesc de asemenea şi alte forme ale reziduurilor studentizate. Astfel, în pachetul de programe Matlab, utilizat şi în statistică, se foloseşte următorul reziduu studentizat,

$$t_i=\frac{e_i}{s_{(-i)}\sqrt{1-h_i}}\in T(n-p-1), \tag{12.4.2}$$

unde

$$s_{(-i)}^2 = \frac{\|e\|^2}{n-p-1} - \frac{e_i^2}{(n-p-1)(1-h_i)} \tag{12.4.3}$$

este estimatorul varianței erorilor σ^2, obținut prin omiterea datei a *i*-a, $(n-1)-p$, fiind astfel, numărul gradelor de libertate din model.

Acelaşi program foloseşte următorul interval de încredere pentru media erorii, precizat prin limitele sale,

$$e_i \pm t_{n-p-1;1-\frac{\varphi}{2}} s_{(-i)} \sqrt{1-h_i}, i = \overline{1,n}, \tag{12.4.4}$$

unde $t_{n-p-1;1-\frac{\varphi}{2}}$ este cuantila de ordin $1-\frac{\varphi}{2}$, a variabilei Student, cu $n-p-1$ grade de libertate, iar $s_{(-i)}$ este cel din (12.4.3). Aceste intervale pot fi reprezentate printr-un grafic cu bare, împreună cu reziduurile e_i. Intervalele de încredere care nu-l includ pe 0 sunt echivalente cu respingerea ipotezei că $E(\varepsilon) = 0$, respingere făcută pentru pragul de semnificație φ. În cazul respingerii ipotezei de zgomot alb asupra erorii, se elimină acest neajuns, prin scoaterea din model a acelor observații pentru care intervalul de încredere (12.4.4) nu-l conține pe 0.

Ipoteza necorelației erorilor cu variabilele exogene

Ipoteza inexistenței unei corelații între eroare şi câte o variabilă exogenă $X_k, k = \overline{1,p}$, asigură obținerea de estimatori optimali, prin metoda celor mai mici pătrate. Dacă există corelație între acestea, atunci estimatorul *a* va fi doar asimptotic nedeplasat. S-a constatat că deplasarea asimptotică obținută pentru *a* este mică, atunci când coeficientul de corelație liniară simplă, între ε şi X_k, este mic ([15]). Amintim aici, definiția coeficientului de corelație liniară simplă.

Definiția 12.4.2. *Numim **coeficient de corelație liniară simplă** (Bravais-Pearson), între două variabile aleatoare, X şi Y, parametrul*

$$r(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{E(XY) - E(X) \cdot E(Y)}{\sigma_X \cdot \sigma_Y}. \tag{12.4.5}$$

Coeficientul de corelație liniară simplă Pearson, de selecție, are valori cuprinse între –1 şi 1 şi se va calcula cu formula

$$\hat{r}(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \qquad (12.4.6)$$

Dacă $|r|$ se apropie de 1, variabilele sunt liniar corelate. În cazul în care *r* se apropie de 0, nu sunt liniar corelate, iar dacă *X,Y* sunt normale şi $r=0$, variabilele sunt independente.

De asemenea, se cunoaşte că statistica

$$T=\frac{\hat{r}^2}{1-\hat{r}^2}\cdot(n-2) \qquad (12.4.7)$$

urmează legea Student, cu $n-2$ grade de libertate. Aşadar, corelația între variabila reziduală şi variabila exogenă X_k se poate testa printr-un test de corelație, bazat pe ipoteza $H_0: r(X_k,\varepsilon)=0$, ipoteză care presupune necorelația. În cazul în care valoarea calculată a statisticii (12.4.7) depăşeşte cuantila de ordin φ, a unei variabile Student, cu *n-2* grade de libertate, φ fiind pragul de semnificație al testului, se respinge ipoteza H_0. Aceleaşi informații se pot obține, analizând graficul în care reziduurile sunt reprezentate în funcție de valorile observate ale variabilei X_k. Graficul respectiv nu trebuie să dea impresia vreunei tendințe, existența acesteia ducând la concluzia că există corelație. Atunci când corelația este mare, deplasarea asimptotică a estimatorului va fi de asemenea, mare şi e de preferat încercarea unui alt model.

Ipoteza homoscedasticității modelului

În cazul în care condiția de homoscedasticitate (erori identic distribuite), $V(\varepsilon_i)\neq\sigma^2$, nu este îndeplinită, modelul se numeşte ***heteroscedastic***. Într-un astfel de caz, rezultă că intensitatea influenței variabilelor exogene asupra celei endogene diferă de la o observație la alta. Condiția $V(\varepsilon_i)=\sigma_i^2, i=\overline{1,n}$, sau $V(\varepsilon)=diag(\sigma_1^2,\sigma_2^2,...,\sigma_n^2)$ nu afectează nedeplasarea estimatorului *a* de cele mai mici pătrate, însă influențează varianța acestuia. Într-un model heteroscedastic avem, $V(a)\to 0$, numai atunci când lipseşte corelația între σ_i^2 şi $x_{ki}, i=\overline{1,p}$, adică atunci când ordinul de mărime al variabilelor, pentru diverse observații, este acelaşi. Această condiție este însă, foarte rar satisfăcută în practică. Astfel, în general dacă modelul este heteroscedastic, eroarea $V(a)$ a estimatorului creşte, prin urmare, creşte şi eroarea medie de estimare.

Pentru testarea homoscedasticității, se cunosc mai multe metode. Una dintre ele apelează la un test de comparare a mai multor dispersii, cum ar fi testul lui Bartlet (vezi [15]), bazat pe legea $\aleph^2$. Ipoteza H_0 de egalitate a dispersiilor se va respinge, dacă valoarea calculată a statisticii aferente testului depăşeşte cuantila corespunzătoare unui anumit prag de semnificaţie şi legii $\aleph^2$. De asemenea, se poate face un test de ipoteză bazat pe reziduuri. Pentru *n* mare (după [29]), reziduul studentizat t_i trebuie să fie cuprins între –2 şi 2. Se compară t_i calculat cu valoarea critică a distribuției, precizate în formula (12.4.2). Cazul când reziduul t_i este mare generează îndoieli asupra faptului că reziduul are aceeaşi varianță ca şi celelalte, ceea ce duce la nesiguranța ipotezei $V\left(\varepsilon_i^2\right)=\sigma^2, i=\overline{1,n}$.

Pentru corectarea heteroscedasticității se utilizează în general, rescalarea modelului. Spre exemplu, pentru un model de forma

$$y_i = a_1 x_{1i} + a_2 x_{2i} + \varepsilon_i, i=\overline{1,n},$$

dacă

$$V\left(\varepsilon_i\right)=\sigma_i^2 = \lambda x_{1i}^2 x_{2i}^2,$$

se scalează modelul astfel,

$$\frac{y_i}{x_{1i}x_{2i}} = a_1 \frac{1}{x_{2i}} + a_2 \frac{1}{x_{1i}} + \frac{\varepsilon_i}{x_{1i}x_{2i}}$$

şi atunci, $V\left(\frac{\varepsilon_i}{x_{1i}x_{2i}}\right)=\lambda, \forall i=\overline{1,n}$.

Ipoteza independenței erorilor

Această ipoteză revine la

$$V\left(\varepsilon\right)=\sigma_\varepsilon^2 I$$

sau

$$\operatorname{cov}\left(\varepsilon_i,\varepsilon_j\right)=0, i\neq j, i, j=\overline{1,n}.$$

Autocorelația erorilor presupune $\text{cov}(\varepsilon_i, \varepsilon_j) \neq 0, i \neq j$. Ipoteza de independență este necesară în obținerea estimatorilor de varianță minimă (optimali), prin metoda celor mai mici pătrate, dar nu şi în obținerea estimatorilor nedeplasați.

Autocorelația erorilor se regăseşte în principal, în cazul modelelor dinamice (serii de timp). Într-un astfel de model, din cauza proastei specificări, influența erorii unei perioade asupra alteia devine plauzibilă. După [15], autocorelația erorilor poate apărea în modelele statistice, doar dacă rezultatele observării au fost aranjate în prealabil, crescător sau descrescător, în raport cu variabila endogenă *Y*. Pentru validarea ipotezei de independență a erorilor, se poate folosi, atât metoda bazată pe analiza graficului reziduurilor, cât şi un test de corelare. Din punct de vedere grafic, se cercetează comportarea empirică a reziduurilor $e_1,...,e_n$, care nu trebuie să dea impresia unei tendințe. Altfel spus, reziduurile nu trebuie să aibă pentru mai multe observații consecutive aceeaşi comportare, spre exemplu, nu trebuie să fie numai pozitive sau numai negative. Pentru necorelație, ele trebuie să fie împrăştiate aleatoriu în jurul axei absciselor.

Din punct de vedere cantitativ, se poate folosi testul Durbin-Watson (vezi [29]), care verifică ipotezele $H_0 : \varepsilon_i$ necorelate şi

$$H_1 : \varepsilon_i = \delta\varepsilon_{i-1} + u_i,\ \delta > 0 \text{ (proces autoregresiv de ordin I).}$$

Statistica aferentă testului este

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \in [0,4]. \qquad (12.4.8)$$

Pentru a admite ipoteza necorelării, trebuie ca valoarea lui *d* să fie în jurul lui 2. În cazul în care se depistează o autocorelație, se corectează calitatea estimatorilor, folosind metoda celor mai mici pătrate generalizate ([30]).

Odată validate sau corectate ipotezele liniarității modelului, avem asigurate pentru estimatorii de cele mai mici pătrate, calitățile discutate în paragraful 12.3. Un alt aspect în analiza rezultatelor regresiei ține de calitatea ajustării.

Calitatea ajustării

Pentru a analiza calitatea ajustării se pot folosi, atât metode numerice, cât şi metode grafice. Metodele grafice (folosite în special în cazul unui singur predictor) se referă, în mare parte, la analiza reziduurilor, analiză care aşa cum s-a văzut în prima parte a acestui paragraf, se poate realiza şi numeric. Metodele numerice utilizate în analiza calității ajustării se bazează pe interpretarea câtorva statistici.

Analiza reziduurilor

Pe lângă verificarea ipotezelor liniarității, studiul reziduurilor poate fi folosit şi în cadrul altor modele de regresie, în vederea reperării observațiilor aberante(outlieri). În prima parte a paragrafului, s-a văzut deja instrumentarul folosit în analiza reziduurilor.

Astfel, dintre mai multe ajustări (în regresii cu un singur predictor, liniare sau nu), se va prefera aceea în care reziduurile sunt aleator împrăştiate în jurul lui zero, fără a comporta o tendință.

În acelaşi sens, se poate face şi comparația curbelor ajustate, în raport cu norul statistic. Pentru eliminarea observațiilor aberante, se poate folosi, de exemplu, testul fundamentat pe reziduurile studentizate şi pe statistica (12.4.1) sau (12.4.2). Un reziduu prea mare poate indica o valoare aberantă, dar o valoare aberantă nu are neapărat reziduul mare. Un alt criteriu care să desemneze valorile aberante ar fi cel bazat pe intervalul de încredere (12.4.4). Reziduul pentru care intervalul nu conține valoarea zero indică o valoare aberantă.

Influența observațiilor

Am văzut cum poate fi eliminată o observație aberantă. Să vedem acum care ar fi influența unei astfel de observații, în cazul în care ea ar rămâne în model.

Definiția 12.4.3. *Dacă notăm cu* $\hat{y}_{(-i)}$*, valoarea ajustată(prezisă) pentru* y_i *(definită în formula (2.1.10), care se obține atunci când este omisă observația a i-a, vom numi* ***reziduu prezis****, valoarea* $y_i - \hat{y}_{(-i)}$.

Conform cu [29], se poate arăta că

$$y_i - \hat{y}_{(-i)} > \frac{e_i}{1-h_i},$$

aşadar observațiile pentru care h_i e mare şi/sau reziduurile sunt mari (observații aberante), duc la previziuni suspecte, căci $y_i - \hat{y}_{(-i)}$ va fi mult diferit de zero. Ca o măsură a puterii de prezicere a modelului, se cunoaşte statistica

$$PRESS = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_i} \right)^2 . \qquad (12.4.9)$$

O valoare mică a statisticii indică o putere mare de prezicere. Astfel, pentru o bună prezicere, trebuie eliminată influența observațiilor aberante (e_i mare) sau cele care au h_i mare. Influența observațiilor asupra estimatorilor este dată de distanța lui Cook,

$$D_i = \frac{\left(a - a_{(-i)}\right)' (x'x)\left(a - a_{(-i)}\right)}{ps^2} = \frac{1}{p} e_i^2 \frac{h_i}{1-h_i} = \frac{\left\|\hat{y} - \hat{y}_{(-i)}\right\|^2}{ps^2}. \quad (12.4.10)$$

Aici, $\hat{y}_{(-i)} = xa_{(-i)}$, $a_{(-i)}$ este estimatorul obținut prin eliminarea observației *i*, iar celelalte elemente au aceeaşi semnificație ca în paragraful 12.3. O observație se consideră a avea o influență anormală, dacă $D_i > 1$.

Analiza estimațiilor obținute

Estimatorul *a* este determinat cu mai multă acuratețe, atunci când intervalele de încredere (12.3.10), pentru coeficienți, sunt mai restrânse.

În continuare, vom prezenta câteva statistici care pot fi folosite pentru evaluarea şi de asemenea, compararea mai multor modele.

Coeficientul de corelație multiplă

Pentru a putea face legătura între acest coeficient şi coeficientul de corelație liniară multiplă, vom aminti mai întâi definiția acestuia.

Definiția 12.4.4. *Se numeşte* ***coeficient de corelație liniară multiplă****, parametrul*

$$r\left(X_1, X_2, ..., X_p, Y\right) = \sup_{a_i} r\left(\sum_{i=1}^{p} a_i X_i, Y\right), \quad (12.4.11)$$

unde r este coeficientul de corelație liniară simplă, definit în formula (12.4.5).

Coeficientul de corelație liniară multiplă, de selecție (eşantionare), se va nota cu $\hat{r}\left(x_1, x_2, ..., x_p, y\right)$ şi este dat de

$$\hat{r}\left(x_1, x_2, ..., x_p, y\right) = \sup_{a_i} \hat{r}\left(\sum_{i=1}^{p} a_i x_i, y\right). \quad (12.4.12)$$

În cazul particular când $p = 1$, se obține valoarea absolută a coeficientului de corelație liniară simplă. În [30], se demonstrează că, dacă $\hat{y}$ este valoarea ajustată a lui *y* (precizată în *Definiția 12.2.4.*), dintr-un model de regresie liniară cu termen constant, atunci

$$\hat{r}\left(x_1, x_2, ..., x_p, y\right) = \hat{r}\left(\hat{y}, y\right). \quad (12.4.13)$$

Altfel spus, metoda celor mai mici pătrate determină acea combinație liniară între variabilele exogene, pentru care corelația cu variabila endogenă este maximală. În plus, se poate demonstra că

$$\hat{r}\left(x_1, x_2, ..., x_p, y\right) = \frac{\left\|\hat{y} - \overline{y}\right\|}{\left\|y - \overline{y}\right\|}, \tag{12.4.14}$$

sau într-o altă formulare

$$\hat{r}^2\left(x_1, x_2, ..., x_p, y\right) = \frac{\left\|\hat{y} - \overline{y}\right\|^2}{\left\|y - \overline{y}\right\|^2} = 1 - \frac{\left\|y - \hat{y}\right\|^2}{\left\|y - \overline{y}\right\|^2}. \tag{12.4.15}$$

Se poate afirma atunci, că coeficientul de corelație liniară multiplă reprezintă proporția în care variația lui *Y* este explicată prin regresia liniară pe $X_1, X_2, ..., X_p$, cu termen constant. Are loc relația $0 \le \hat{r}\left(x_1, x_2, ..., x_p, y\right)$, $\hat{r}^2\left(x_1, x_2, ..., x_p, y\right) \le 1$, o valoare apropiată de unu fiind o posibilă acoperire a faptului, că modelul liniar este potrivit pentru a explica variabila *Y*. Pentru a realiza inferența asupra estimatorului $\hat{r}$, se folosește statistica de tip Fisher Snedecor, cu $p-1$ și $n-p$ grade de libertate,

$$F = \frac{\hat{r}^2}{1-\hat{r}^2} \cdot \frac{n-p}{p-1} \tag{12.4.16}$$

și se emite ipoteza nulă, $r = 0$, care este echivalentă de fapt, cu ipoteza $\alpha_k = 0, k = \overline{1, p}$, α_0 oarecare, ipoteze ce infirmă regresia liniară. Ipoteza nulă se va respinge, dacă valoarea calculată a statisticii (12.4.16) depășește cuantila corespunzătoare legii Fisher-Snedecor și pragului de semnificație φ. Totuși, chiar dacă $\hat{r} \cong 1$ și valoarea reală *r* este semnificativă $\left(H_1 : r \neq 0\right)$, acest lucru nu indică neapărat o corelație reală. Spre exemplu, în cazul regresiei simple, aceste situații cu privire la $\hat{r}$ și *r* pot apărea, ca urmare a unei corelații paralele pe care, atât variabila *y*, cât și variabila *x*, o poate avea cu o a treia variabilă. Pentru a elimina influența acestei de-a treia variabile, se poate calcula un coeficient de corelație parțială, cu formula

$$\hat{r}_{yx.z} = \frac{\hat{r}_{yx} - \hat{r}_{yz} \cdot \hat{r}_{xz}}{\sqrt{1-\hat{r}_{yz}^2}\sqrt{1-\hat{r}_{xz}^2}},$$

formulă ce poate fi extinsă și în cazul regresiei multiple.

Testul bazat pe r^2, prezentat mai sus, este unul de semnificație al coeficienților de regresie în ansamblul lor şi în acelaşi timp(prin $\hat{r}$), al semnificației regresiei liniare cu termen constant, test folosit în analiza rezultatelor regresiei. Spre exemplu, în Matlab, se întoarce ca şi informație de regresie, valoarea calculată a statisticii *F* şi *p* –valoarea asociată acestei statistici, adică probabilitatea critică,

$$c = 1 - F_{p-1,n-p}\left(F_{calculata}\right). \tag{12.4.17}$$

Dacă $c \le \varphi$, φ prag de semnificație, se respinge ipoteza H_0.

Deoarece $\hat{r}$, dat prin (12.4.13), s-a definit doar pentru $\hat{y}$ rezultat din modelul liniar cu termen constant, nu vom putea vorbi de aceste elemente, în cazul unui model de regresie liniară oarecare. Pentru modelul de regresie liniară fără termen constant, se pot folosi testele F de semnificație asupra unui subansamblu al coeficienților, precum şi testul T asupra fiecărui coeficient în parte, teste prezentate în paragraful 12.3. Un test pentru ansamblul coeficienților se poate face şi pentru modelul de regresie liniară fără termen constant, folosind statistica de tip Fisher-Snedecor, cu *p* şi $n - p$ grade de libertate (vezi [29]):

,

$$F = \frac{1}{ps^2}\left(a - \alpha\right)'\left(x'x\right)\left(a - \alpha\right).$$

Totuşi, atât pentru modelul liniar fără termen constant, cât şi pentru un model de regresie oarecare, se poate defini o caracteristică asemănătoare cu *r*, caracteristică pe care o vom numi coeficient de corelație multiplă(oarecare).

Definiția 12.4.5. *Numim* ***coeficient de corelație (determinație) multiplă*** *(oarecare), parametrul definit prin*

$$R^2_{(y,\hat{y})} = 1 - \frac{S_R^2}{S_T^2}, \tag{12.4.18}$$

unde

$$S_T^2 = \sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2, \text{ sumă de pătrate totală,}$$

$$S_R^2 = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2, \text{ sumă de pătrate reziduală (indusă de reziduuri),} \tag{12.4.19}$$

iar $\hat{y}$ *este valoarea ajustată (prezisă) a lui y, prin modelul considerat.*

Aşadar, coeficientul de corelație multiplă se calculează cu formula

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{S_T^2 - S_R^2}{S_T^2} = \frac{\|y - \bar{y}\|^2 - \|y - \hat{y}\|^2}{\|y - \bar{y}\|^2}. \qquad (12.4.20)$$

Mărimea R^2 reprezintă o măsură a calității ajustării datelor, prin modelul de regresie oarecare, în urma căreia rezultă valoarea ajustată(prezisă), $\hat{y}$. Această mărime se poate folosi pentru a compara mai multe ajustări diferite şi poate fi calculată, pentru orice model de regresie, chiar şi pentru cele neparametrice. Totuşi, în [30], Stapleton atrage atenția, că R^2 bazat pe scări diferite nu sunt comparabile. Această observație se referă la cazul când pentru ajustare e necesară o transformare. De exemplu, $R^2(\hat{y}, y)$ se poate compara cu $R^2(\hat{y}_z, y)$, dar nu cu $R(z, \hat{z})$, dacă $z = g(y)$, $\hat{z}$, valoarea ajustată a lui *z*, din modelul liniar al lui *z* pe *x*, $\hat{y}$, valoarea ajustată a lui *y*, din regresia lui *y* pe *x*, $\hat{y}_z = g^{-1}(\hat{z})$. O valoare mare a lui R^2 va indica o mai bună apropiere a modelului față de date. Acest lucru se poate vizualiza şi grafic, pentru modelul cu un singur regresor. Din formula (12.4.20), se observă că R^2 poate lua şi valori negative, valoarea maximă fiind 1.

Acest parametru este cunoscut în literatura de specialitate, doar sub denumirea de *R*-pătrat. Am preferat să-l numim aici coeficient de corelație multiplă (oarecare), deoarece atunci când $\hat{y}$ provine dintr-un model de regresie liniară cu termen constant, R^2 coincide cu coeficientul de corelație liniară multiplă, $\hat{r}^2$, din formula (12.4.15). Într-adevăr, pentru modelul liniar cu termen constant are loc regula de adunare a varianțelor (ANOVA),

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad (12.4.21)$$

sau altfel,

$$\|y - \bar{y}\|^2 = \|y - \hat{y}\|^2 + \|\hat{y} - \bar{y}\|^2 \qquad (12.4.22)$$

sau încă,

$$S_T^2 - S_R^2 = \|\hat{y} - \bar{y}\|^2 \text{ (sumă indusă de model).} \qquad (12.4.23)$$

În general, această regulă nu are loc, de aceea prima parte a formulei (12.4.15) nu este valabilă pentru R^2 şi deci, $R^2 \neq \hat{r}^2$, diferență confirmată şi de faptul că în general, R^2 poate lua valori negative (mai precis pentru modele fără termen constant), în timp ce $\hat{r}^2$ nu poate. Totuşi R^2 şi $\hat{r}^2$ au aceeaşi semnificație şi anume, precizează proporția, $\left(\frac{S_T^2 - S_R^2}{S_T^2}\right)$, în care un model explică bine datele, numai că $\hat{r}^2$ se referă strict la modelul liniar cu termen constant. Spre exemplu, o valoare de 0,97 a lui R^2 ne arată, că modelul din care rezultă $\hat{y}$ acoperă variabilitatea observațiilor, în proporție de 97%.

O mărime asemănătoare cu R^2 este şi raportul de determinație(corelație), $\frac{V(Y(X))}{V(X)} > 0$, care coincide cu R^2 şi cu $\hat{r}^2$ (a se vedea [29]), pentru regresia liniară simplă. În general, raportul de corelație arată proporția în care factorul *X* explică variabila Y, neprecizând forma corelației, în timp ce coeficientul de corelație arată proporția în care factorul *X* (unidimensional) explică variabila *Y*, prin intermediul unui anumit model de regresie.

Pe lângă R^2, tot ca şi măsură a calității ajustării, se foloseşte o variantă ajustată a sa. Necesitatea acestei ajustări rezidă în faptul că R^2 nu este un criteriu absolut al calității ajustării, în cazul când se compară două modele cu un număr diferit de parametrii. Astfel, dacă *p* creşte în regresia liniară, spre exemplu, implică faptul că în model s-a inclus o nouă variabilă, ceea ce face ca S_R^2 să scadă şi implicit, *R* să crească artificial.

Definiția 12.4.6. *Se numeşte coeficient de corelație ajustat, parametrul*

$$\overline{R}^2 = 1 - \frac{n-1}{n-p}\left(1 - R^2\right). \qquad (12.4.24)$$

Astfel, factorul $n - p$ corectează modificarea artificială a lui R^2, odată cu modificarea lui *p*. Ca şi în cazul lui R^2, o valoare apropiată de 1 a lui $\overline{R}^2$ va arăta o bună ajustare, $\overline{R}^2$ fiind un criteriu bun pentru modele cu un număr diferit de parametri.

Alte statistici care măsoară numeric calitatea ajustării (pe lângă R^2 şi $\overline{R}^2$) sunt S_R^2 din formula (12.4.19) şi *s* din formula (12.3.7). Statistica S_R^2 este cunoscută şi sub denumirea de sumă de pătrate datorată erorilor şi măsoară abaterea totală a variabilei *Y*, de la model. O valoare apropiată de zero indică o bună ajustare. Statistica *s*, numită şi eroare standard a regresiei, este rădăcina pătrată a erorii medii pătratice, întâlnită sub notația MSE sau RMS (media pătratelor erorilor). O valoare a lui *s* apropiată de 0 indică o bună ajustare. Toate aceste statistici pot fi definite şi pentru alte modele decât cele liniare.

Odată stabilit şi validat modelul de regresie, ne interesează problema previziunii.

Intervale de încredere pentru previziune

Pe lângă intervalele de încredere pentru previziune, pentru o nouă observaţie, vom prezenta aici şi intervale de încredere pentru medie (pentru funcţia de regresie *f*), pentru a le putea compara. De asemenea, în literatura de specialitate, se iau în calcul, atât intervale simultane (pentru *x* oarecare), cât şi intervale nesimultane (pentru o singură valoare specificată, x_0). Vom considera aici doar intervale nesimultane.

Cazul regresiei liniare (fără termen constant)

Presupunem că modelul $y = x\alpha + \varepsilon$, $\varepsilon \in N(\theta, \sigma^2 I)$, a fost validat şi că ne interesează estimaţii asupra unei noi valori, y_0, a variabilei *Y*, date fiind valorile $x_{01}, x_{02}, ..., x_{0p}$, ale variabilelor $X_1, X_2, ..., X_p$, adică ne interesează să previzionăm o valoare a lui *y*(necunoscută). Avem

$$y_0 = \alpha_1 x_{01} + ... + \alpha_p x_{0p} + \varepsilon_0 = x_0'\alpha + \varepsilon_0 \qquad (12.4.25)$$

sau folosind modelul ajustat prin cele mai mici pătrate,

$$y_0 = a_1 x_{01} + ... + a_p x_{0p} + e_0 = x_0' a + e_0 .$$

Apar aşadar, două tipuri de erori în estimarea previziunii. Una se datorează noii erori ε_0, cu care valoarea y_0 intră în model, iar alta se datorează erorii cu care s-a estimat α. Sigur, în cazul în care dorim să estimăm media $E(Y|X = x_0) = f(x_0) = x_0'\alpha$, atunci eroarea este mai mică şi se reduce la eroarea indusă de α. Folosim notaţia

$$\hat{y}_0 = x_0' a , \qquad (12.4.26)$$

pentru a desemna previziunea punctuală şi dăm în continuare, intervalul de încredere pentru

$$y_0 = x_0'\alpha + \varepsilon_0 \text{(previziune)},$$

respectiv pentru

$$f(x_0) = E(Y|X = x_0) = x_0'\alpha \text{ (medie)}.$$

Interval de încredere pentru previziune:

$$P\left(\hat{y}_0 - st_{n-p,1-\frac{\varphi}{2}}\sqrt{1+x_0'(x'x)^{-1}x_0} \le y_0 \le \hat{y}_0 + st_{n-p,1-\frac{\varphi}{2}}\sqrt{1+x_0'(x'x)^{-1}x_0}\right) = 1-\varphi\,. \tag{12.4.27}$$

Interval de încredere pentru medie:

$$P\left(\hat{y}_0 - st_{n-p,1-\frac{\varphi}{2}}\sqrt{x_0'(x'x)^{-1}x_0} \le f(x_0) \le \hat{y}_0 + st_{n-p,1-\frac{\varphi}{2}}\sqrt{x_0'(x'x)^{-1}x_0}\right) = 1-\varphi\,. \tag{12.4.28}$$

În ambele intervale, s, x, α şi t au semnificaţia precizată deja în acest capitol. Se observă că intervalul pentru previziune este mai larg, datorită erorii suplimentare care intervine prin ε_0. În cazul modelului liniar cu termen constant, intervalele sunt aceleaşi, diferă doar forma lui x_0 şi x, care vor conţine şi un 1, respectiv o coloană de 1 (a se vedea [29], [30]). În cazul regresiei liniare simple, când funcţia ajustată şi intervalele pot fi vizualizate grafic, avem următoarea situaţie:

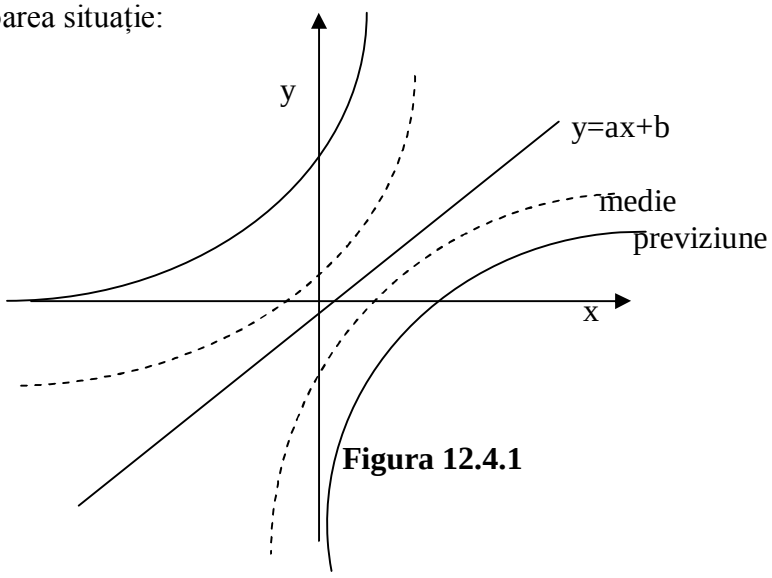


Figura 12.4.1

Avem aşadar şi vizualizarea faptului că intervalul de previziune este mai larg, curbele corespunzătoare încadrând, pentru acelaşi φ, curbele de la intervalul pentru medie, care la rândul lor, încadrează dreapta de regresie. Dacă pe grafic ar fi vizualizate şi datele, s-ar vedea că intervalele de încredere pentru previziuni includ punctele. Ca orice interval de încredere şi intervalele de previziune pot fi instrumente în evaluarea calităţii regresiei, acurateţea previziunii fiind cu atât mai bună, cu cât intervalul este mai mic. Intervalele

simultane sunt asemănătoare celor nesimultane, diferența fiind că se utilizează distribuția Fisher-Snedecor, în loc de distribuția Student.

Exemplul 12.4.7. *Reluând datele din Exemplul 12.2.8, ne propunem să estimăm punctual şi prin interval de încredere de tip 95% rata rentabilității acțiunilor respective corespunzătoare unei rate a pieței de 2%.*

Soluție

Calculând previziunea punctuală obținută pe baza modelului din *Exemplul 12.2.8*, avem $\bar{y}(2) = -1{,}76 + 1{,}13 \cdot 2 \cong 0{,}5$, în timp ce o evaluare a formulei (12.4.27) duce la intervalul de încredere de tip 95% pentru previziune, (-4,42; 5,39).

Capitolul 13

Statistică Bayesiană şi noţiuni de teoria credibilităţii

13.1. Statistică Bayesiană

Un asigurator are un contract care se derulează pe mai mulţi ani. Sumele de bani plătite clienţilor au fost $X_1,X_2,\ldots,X_t$. Ţinînd seama de această experienţă, care ar fi cel mai bun principiu de calcul al primei de asigurare pentru anul t+1?

Pentru asigurator, clientul este un risc, X.

Problema ar fi : cum am putea modifica prima de asigurare h(X) în aşa fel ca să ţin seama de experienţa anilor trecuţi?

Nefiind precis formulată, aceasta NU este o problemă matematică.

O problemă oarecum mai abordabilă ar suna astfel: în teoria asigurărilor, media EX se numeşte **premiul brut** sau **prima brută**. La acesta se mai adaugă diverse sume care ar trebui să ţină seama de profit, de cheltuieli de regie precum şi de alte lucruri ce nu ţin de obiectul studiului nostru.

Problema de bază este găsirea repartiţiei „adevărate" a lui X.

În statistica parametrică clasică se presupune că **repartiţia** F_X aparţine **unei clase de repartiţii** depinzînd de un parametru necunoscut $\theta \in E$. De exemplu, putem crede că X ~ Poisson(θ) sau X ~ Binomial(n,p), sau X ~ Exp(a). În primul caz am avea o familie depinzînd de un singur parametru, iar în al doilea de una depinzînd de doi parametri (căci θ = (n,p) !) . Pentru a merge mai departe ar trebui găsit adevăratul θ. Dacă suntem dispuşi să credem că θ nu se modifică la rîndul lui în timp, am putea privi atunci experienţa acumulată $\underline{X}=(X_1,\ldots,X_t)$ ca o selecţie de volum t dintr-o populaţie F_0. Caz în care s-ar pune problema **estimării** lui θ.

Dacă θ ar fi unidimensional, am putea încerca să găsim intervale de încredere pentru θ, cu un anume risc asumat α. Acesta este punctul de vedere al statisticii parametrice. Ea se foloseşte dacă dispunem de multe date.

În acest capitol vom folosi însă o altă abordare, şi anume cea Bayesiană. Parametrul $\theta \in E$ se numeşte **factor de risc**. Ideea de bază în abordarea bayesiană este că , necunoscînd valoarea adevărată a lui θ, **e ca şi cum factorul de risc θ ar fi la rândul lui o variabilă aleatoare**. O vom nota cu Θ. Formal, avem un spaţiu probabilizat pe care avem definit un vector de observaţii $\underline{X}=(X_1,\ldots,X_t)$ şi o variabilă aleatoare Θ cu valori în E

.

Desigur că apar unele probleme tehnice: va trebui ca spațiul parametrilor E să fie organizat ca un spațiu măsurabil $(E,\mathcal{E})$. De obicei aceasta nu este o problemă, în cazurile clasice. De exemplu, în cazul repartiției Poisson sau exponențiale E = [0,∞); la binomială este N×[0,1] care se organizează în mod natural ca spații măsurabile cu σ-algebra mulțimilor boreliene $\mathcal{B}([0,\infty))$ în primele două cazuri sau cu $\mathcal{P}(\mathrm{N})\otimes\mathcal{B}([0,1])$ în al doilea. La repartiția normală $N(\mu,\sigma^2)$, $E = \Re \times (0,\infty)$, etc.

Putem avea o idee despre repartiția factorului de risc, o credință. Aceasta se numeşte **repartiția apriori** a factorului de risc Θ. În limbaj matematic, repartiția apriori este repartiția inițială a lui Θ. În cele ce urmează ea va fi notată cu *U*. Deci $U(B) = P(\Theta \in B)$.

Dacă parametrul Θ ia valoarea θ , atunci repartiția selecției noastre $\underline{\boldsymbol{X}}$ ar trebui să fie Q(θ). Acesta se numeşte **modelul**. Deci pentru fiecare θ ∈ E , Q(θ) este o repartiție pe $\Re^t$. Aceasta este în definitiv **repartiția condiționată** a lui $\underline{\boldsymbol{X}}$ ştiind că Θ = θ.

Pe scurt o abordare Bayesiană înseamnă o repartiție apriori a parametrului şi un model. Ideea este să obținem din experiență o ajustare a repartiției apriori, numită repartiția aposteriori a parametrului Θ.

Modelul este de fapt o probabilitate de trecere de la mulțimea *E* a parametrilor la mulțimea $\Re^t$ a rezultatelor. Atunci ştim de la capitolul privind simularea variabilelor aleatoare că repartiția vectorului $(\Theta,\underline{X})$ este $U\otimes Q$ iar repartiția lui $\underline{X}$ este *UQ*. Amintim că cele două operații se definesc astfel

$$U\otimes Q(C) = \int Q(\theta, C_\theta)\,dU(\theta) \quad \forall\, C \subseteq E \times \Re^t \text{ măsurabilă} \qquad (13.1.1)$$

$$UQ(B) \overset{Def}{=} U\otimes Q(E\times B) = \int Q(\theta, B)\,dU(\theta) \qquad (13.1.2)$$

Exemplul 13.1.1. *Să presupunem că* $\Theta \sim U([0,1])$ *iar* $\underline{X} \sim Binomial(1,\Theta)^t$. *Mai explicit, modelul este următorul: dacă adevărata valoare a lui* Θ *ar fi* $\Theta = p \in [0,1]$, *atunci* $P(X_1=\varepsilon_1,\ldots, X_t = \varepsilon_t)$ *ar trebui să fie* $p^{\nu(1,\varepsilon)}(1-p)^{\nu(0,\varepsilon)}$ *unde* $\nu(1,\varepsilon) = |\{1\le j \le t \mid \varepsilon_j = 1\}|$ *iar* $\nu(0,\varepsilon) = |\{1\le j \le t \mid \varepsilon_j = 0\}| = t - N(1,\varepsilon)$. *Cum afectează variabila aleatoare N(1) repartiția inițială a lui Θ?*

Exemplul 13.1.2. *O generalizare.* $\Theta \sim U(0,1)$ *iar* $\underline{X} \sim Binomial(\mathrm{n},\Theta)^t$ *unde n este presupus cunoscut. Atunci* $P(X_1=x_1,\ldots,X_t = x_t) = C_n^{x_1}\ldots C_n^{x_t}\theta^{x_1}(1-\theta)^{n-x_1}\ldots\theta^{x_t}(1-\theta)^{n-x_t} = C_n^{x_1}C_n^{x_2}\ldots C_n^{x_t}\theta^S(1-\theta)^{nt-S}$ *unde* $S = S(\underline{x}) = x_1 + \ldots + x_t$.

În Exemplul 13.1.1, repartiția lui $\underline{X}$ este $U(0,1)Q$ unde $Q(\theta) = Binomial(1,\theta)^t$. Deci din (13.1.2) avem

$$P(\underline{X} = \varepsilon) = \int_0^1 \theta^S(1-\theta)^{t-S}d\theta = \beta(S+1, t-S+1) \qquad (13.1.3)$$

unde $\varepsilon \in \mathbf{Z}_2^t$ iar S = ν(1,ε)este definit în Exemplul 13.1.1. Funcția β este funcția β a lui Euler , $\beta(m+1,n+1) = \frac{\Gamma(m+1)\Gamma(n+1)}{\Gamma((m+1)+(n+1))} = \frac{m!n!}{(m+n+1)!}$. Rezultă că $P(\underline{X} = \varepsilon) = \frac{S!(t-S))!}{(t+1)!}$.

În Exemplul 13.1.2 diferența este că $\underline{X}$ este $U(0,1)Q$ unde $Q(\theta) = Binomial(n,\theta)^t$ deci

$$P(\underline{X} = \underline{x}) = \int_0^1 C_n^{x_1} C_n^{x_2} .. C_n^{x_t} \theta^S (1-\theta)^{nt-S} d\theta = C_n^{x_1} C_n^{x_2} .. C_n^{x_t} \beta(S+1, nt-S+1) \quad (13.1.4)$$

Rezumînd cele de mai sus putem concluziona : într-un model Bayesian U este marginala pe spațiul parametrilor a repartiției $U \otimes Q$ a vectorului $(\Theta, \underline{X})$ format din parametrul aleatoriu Θ şi selecția $\underline{X}$, (adică o probabilitate pe $(E \times \Re^t)$) modelul Q(θ) este repartiția selecției condiționată de valoarea pe care o ia parametrul,(o probabilitate de trecere de la E la $\Re^t$) iar a doua marginală, cea de pe $\Re t$ este UQ – repartiția selecției $\underline{X}$. Dacă, aşa cum se întîmplă în aplicații, Θ este la rîndul lui un spațiu standard Borel, teorema de dezintegrare o putem aplica şi marginalei a doua. Deci există o altă probabilitate de trecere U_1 , de data aceasta de la $\Re^t$ la E astfel încît $U = U_1(UQ)$. Din punct de vedere probabilistic Q^* reprezintă repartiția lui Θ condiționată de eşantionul $\underline{X}$. În jargonul Bayesian, U_1 este **repartiția aposteriori** a parametrului Θ după observația $\underline{\boldsymbol{X}}$.

Sensul ar fi că $U_1(\underline{x},A) = P(\Theta \in A \mid \underline{X} = \underline{x})$.

În anumite condiții, suficient de largi pentru statistică, există formule de calcul a repartiției aposteriori U_1.

Să presupunem că repartiția parametrului Θ, deci U , admite o densitate față de o măsură σ-finită τ. De asemenea, presupunem că şi repartiția observației $\underline{X}$ condiționată de Θ = θ este absolut continuă față o altă măsură σ-finită ν, adică admite o densitate q față de ν . Atunci se poate calcula repartiția aposteriori a lui Θ.

Propoziția 13.1.3. *Presupunem că E este o mulțime boreliană dintr-un spațiu euclidian. În plus, presupunem că*

(i). $U = u \cdot \tau$, *unde* τ *este o măsură* σ*-finită pe E;*

(ii). $Q(\theta) = q(\theta) \cdot \nu$ *unde* ν *este o măsură* σ *- finită pe* $\Re^t$

Atunci

$$P \circ (\Theta, \underline{X})^{-1} = f_{\Theta,\underline{X}} \cdot (\tau \otimes \nu) \text{ cu } f_{\Theta,\underline{X}}(\theta,x) = q(\theta,x)u(\theta) \quad (13.1.5)$$

$$P \circ \underline{X}^{-1} = f_{\underline{X}} \cdot \nu \text{ unde } f_{\underline{X}}(\underline{x}) = \int q(\theta,\underline{x})u(\theta)d\tau(\theta) = \int f_{\Theta,\underline{X}}(\theta,\underline{x})\, d\tau(\theta) \quad (13.1.6)$$

$$U_1(\underline{x}) = q^*(\underline{x}) \cdot \tau \text{ unde } q^*(\underline{x},\theta) = \frac{q(\theta,\underline{x})u(\theta)}{\int q(\sigma,\underline{x})u(\sigma)d\tau(\sigma)} \quad (13.1.7)$$

Notații standard. O notație mai sugestivă pentru q(θ) este cea folosită în statistică: $f_{\underline{X} \mid \Theta = \theta}$. Deci modelul este că, dacă parametrul Θ ar lua valoarea θ,

atunci densitatea lui $\underline{X}$ față de ν ar trebui să fie $f_{\underline{X}\mid \Theta=\theta}$. Atunci densitatea comună a vectorului $(\Theta, \underline{X})$ față de $\tau\otimes\nu$ este $f_{\Theta,\underline{X}}$ iar densitatea parametrului Θ în ipoteza că că $\underline{X} = \underline{x}$ este $f_{\Theta\mid \underline{X}=\underline{x}}$. Reformulînd cu aceste notații standard avem

densitatea lui $(\Theta,\underline{X})$ este $f_{\Theta,\underline{X}}(\theta, \underline{x}) = f_{\underline{X}\mid \Theta=\theta}(\underline{x})f_{\Theta}(\theta)$ (13.1.8)

densitatea lui $\underline{X}$ este $f_{\underline{X}}(\underline{x}) = \int f_{\Theta,\underline{X}}(\underline{x},\theta)d\tau(\theta)$ (13.1.9)

densitatea aposteriori a lui Θ dată de rezultatul $\underline{x}$ este

$$f_{\Theta\mid \underline{X}=\underline{x}}(\theta) = \frac{f_{\Theta,\underline{X}}(\theta,\underline{x})}{f_{\underline{X}}(\underline{x})} \qquad (13.1.10)$$

Demontrația se poate găsi, de exemplu, în Gheorghiță Zbăganu, *Metode matematice în teoria şi actuariat, Bucureşti,* Editura Universității 2004, pp 236.

Continuare la exemplele 13.1.1 şi 13.1.2.

Cu notațiile standardizate de mai sus avem

La exemplul 13.1.1: $E = [0,1]$, $\tau = U(0,1)$, $u(\theta) = 1$, $\nu = \mathrm{Card}(\mathbf{Z}_2^t)$ este măsura cardinal pe $\mathbf{Z}_2^t$, densitatea modelului este $f_{\underline{X}\mid \Theta=\theta}(\underline{x}) = \theta^S(1-\theta)^{t-S}$ cu $S = \left|\{1\le j\le t \mid x_j = 1\}\right| = x_1 +\ldots+ x_t$.

Atunci

- $f_{\Theta,\underline{X}}(\theta,\underline{x}) = \theta^S(1-\theta)^{t-S}1_{(0,1)}(\theta)$
- $f_{\underline{X}}(\underline{x}) = \int q(\theta,\underline{x})u(\theta)d\tau(\theta) = \beta(S+1,t-S+1) = \frac{S!(t-S)!}{(t+1)!}$
- $f_{\Theta\mid \underline{X}=\underline{x}}(\theta) = \frac{(t+1)!}{S!(t-S)!}\theta^S(1-\theta)^{t-S}$

Recunoaştem aici că repartiția aposteriori a parametrului Θ este o repartiție Beta(S+1,t-S+1) .

Interpretarea: în urma unei experiment în care au apărut M de „1" şi N de „0" şi în care apriori *nu aveam nici o idee preconceputa asupra lui p* credința noastră asupra parametrului θ ar trebui să fie dată de densitatea aposteriori $u_1(\theta) = \beta_{M+1,N+1}(\theta)$.

La exemplul 13.1.2: E, τ, u sunt aceiaşi, dar $\nu = \mathrm{Card}(\mathbf{Z}_n^t)$;

- $f_{\underline{X}\mid \Theta=\theta}(\underline{x}) = C(\underline{x})\,\theta^S(1-\theta)^{t-S}$ cu $C(\underline{x}) = C_n^{x_1}C_n^{x_2}\ldots C_n^{x_t}$
- $f_{\Theta,\underline{X}}(\theta,\underline{x}) = C(\underline{x})\,\theta^S(1-\theta)^{t-S}1_{(0,1)}(\theta)$
- $f_{\underline{X}}(x) = \frac{C_n^{x_1}C_n^{x_2}\ldots C_n^{x_t}}{(nt+1)C_{nt}^S}$ (vezi (1.1.10)
- $f_{\Theta\mid \underline{X}=x}(\theta) = \frac{1}{\beta(S+1,nt-S+1)}\theta^S(1-\theta)^{nt-S}$ deci repartiția aposteriori a parametrului Θ este Beta(S+1, nt-S+1)

Observație 13.1.4. *Dar dacă aveam o idee preconceută? De exemplu, dacă am fi crezut că p = α?* Atunci statistica Bayesiană nu ne-ar fi de nici un folos. Să presupunem că noi avem o credință apriori că Θ, parametrul nostru

are repartiția $\Theta \sim \begin{pmatrix} \theta_1 & \theta_2 & \dots & \dots & \theta_n \\ p_1 & p_2 & \dots & \dots & \theta_n \end{pmatrix}$. Atunci Q ar fi devenit o matrice stocastică cu n linii şi 2^t coloane: $Q(\theta_j,\underline{x}) = \theta_j^{M(\underline{x})}(1-\theta_j)^{t-M(\underline{x})}$. Cu notațiile din Propoziția 13.1.3 am avea $E = \{\theta_1,\dots,\theta_n\} \subset [0,1]$, $\tau = Card(E)$, $u(\theta_j) = p_j$, $\nu = \text{Card}(\mathbf{Z}_2^t)$, $f_{\Theta,\underline{X}}(\theta,\underline{x}) = \theta_j^{M(\underline{x})}(1-\theta_j)^{t-M(\underline{x})})p_j$ iar

$$f_{\Theta|\underline{X}=\underline{x}}(\theta_j) = \frac{\theta_j^{M(\underline{x})}(1-\theta_j)^{t-M(\underline{x})}\,p_j}{\sum_{i=1}^{n}\theta_i^{M(\underline{x})}(1-\theta_i)^{t-M(\underline{x})}\,p_i} \qquad (13.1.11)$$

În cazul particular în care n = 1 (deci credem orbeşte că $\Theta = \theta_1$) atunci suma de la numitorul din (13.1.11) coincide cu numărătorul, deci $f_{\Theta|\underline{X}=\underline{x}}(\theta_1) = 1$. Ceea ce înseamnă că indiferent ce ne spune experiența, vom continua a crede că $\Theta=\theta_1$!

O explicație este că ***experiența niciodata nu creează noi posibilități explicative,*** cel mult poate anula unele din ele – sau să le facă mai neverosimile.

Exemplu 13.1.5. *Un asigurator are în perspectivă un contract format din riscuri repartizate binomial. El ştie că $X_r \sim Binomial(N, \pi)$, dar nu ştie nici pe N , nici pe π. De exemplu X_r pot fi piesele rebutate dintr-un lot de N piese. Mai ştie că în decursul derulării contractului aceşti parametri nu se schimbă. Pentru a avea o idee ce primă de asigurare să ceară, are la dispoziție un istoric al numărului de rebuturi $X_1,\dots,X_n$. Experiența anterioară îl face să creadă că că N şi π sunt independente şi că $N \sim \sum_{n\geq 1}\alpha_n\varepsilon_n$ iar $\pi \sim \rho\cdot\lambda$ unde $\rho:[0,1] \to [0,\infty)$ este o densitate . Dacă nu are nici o idee despre p- lucru destul de neverosimil - va lua $\rho = 1_{(0,1)}$*

Deci

$E = \mathbf{N}\times[0,1]$, $\Theta = (N,\pi)$,

$\tau = \text{Card}_N \otimes \lambda$,

$\theta = (n,p)$, $f_\Theta(n,p) = \alpha_n\rho(p)$, (*repartiția apriori)*

$f_{\underline{X}|\Theta=(n,p)}(\underline{x}) = \alpha_n C(\underline{x},\theta)p^S(1-p)^{tn-S}$ cu $C(\underline{x},\theta) = C(\underline{x},n) = C_n^{x_1}C_n^{x_2}\dots C_n^{x_t}$

(acesta este modelul propriu zis!)

$f_{\Theta,\underline{X}}(\theta,\underline{x}) = \alpha_n\rho(p)C(\underline{x},\theta)p^S(1-p)^{tn-S}$

Repartiția lui $\underline{X}$ este o mixtură de binomiale. Putem scrie

$$f_{\Theta,\underline{X}}(\theta,\underline{x}) = \rho(p)\alpha_n C(\underline{x},\theta)p^{tM(\underline{x})}(1-p)^{t(n-M(\underline{x}))} \qquad (13.1.12)$$

unde $M(\underline{x})$ este media *aritmetică a primelor t observații, $tM(\underline{x}) = x_1 + \dots + x_t$. Fie $x^* = max(x_1,\dots,x_t)$. Observînd că $n < x^* \Rightarrow C(\underline{x},n) = 0$ şi înlocuind, obținem din că*

$$f_{\underline{X}}(\underline{x}) = \int q(\theta,\underline{x})u(\theta)d\tau(\theta) = \sum_{n=x^*}^{\infty}\alpha_n C(\underline{x},n)\int_0^1 \rho(p)p^{tM(\underline{x})}(1-p)^{t(N-M(\underline{x}))}\,dp \qquad (13.1.13)$$

(*dacă $\pi \sim U(0,1)$ atunci* $f_{\underline{X}}(\underline{x}) = \sum_{n=x^*}^{\infty}\alpha_n \frac{C_n^{x_1}C_n^{x_2}\dots C_n^{x_t}}{(nt+1)C_{nt}^{S}}$) *iar din* (13.1.13)

$$f_{\Theta|\underline{X}=\underline{x}}(n,p) = \frac{\rho(p)\alpha_n C(x,n)p^{tM(\underline{x})}(1-p)^{t(n-M(x))}}{\sum_{k=x^*}^{\infty}\alpha_k C(\underline{x},k)\int p^{tM(\underline{x})}(1-p)^{t(k-M(x))}\rho(p)\mathrm{d}p} \qquad (13.1.14)$$

(*dacă* $\pi \sim U(0,1)$ *atunci* $f_{\Theta|\underline{X}=\underline{x}}(n,p) = \frac{\alpha_n C(\underline{x},n)p^S(1-p)^{tn-S}}{\sum_{k=x^*}^{\infty}\alpha_k \frac{C_k^{x_1}C_k^{x_2}\ldots C_k^{x_t}}{(kt+1)C_{nt}^S}}$

Observăm ceva de bun simț, pentru care nu avem nevoie de multă știință de carte: dacă $n < x^*$, *atunci din (13.1.17),* $f_{\Theta|\underline{X}=\underline{x}}(n,p) = 0$. *Nu o să considerăm posibil ca N să ia valori mai mici decât x**!

Definiția 13.1.6. *Dacă* $\varphi: E \to \Re$ *este o funcție măsurabilă, variabila aleatoare* $\mathrm{E}(\varphi(\Theta)\,|\,\underline{X})$ *se numeşte în limbaj bayesian* ***estimatorul Bayesian cu cele mai mici pătrate*** *al lui* $\varphi(\Theta)$.

Propoziția 13.1.3 are drept corolar o formulă de calcul pentru $\mathrm{E}(\varphi(\Theta)\,|\,\underline{X})$, evidentă datorită formulei de transport:

Propoziția 13.1.7. *Avem*

$$\mathrm{E}(\varphi(\Theta)\,|\,\underline{X}=\underline{x}) = \frac{\int \varphi(\theta) f_{\underline{X}}|_{\Theta=\theta}(\underline{x})u(\theta)\mathrm{d}\tau(\theta)}{\int f_{\underline{X}}|_{\Theta=\sigma}(\underline{x})u(\sigma)\mathrm{d}\tau(\sigma)} \qquad (13.1.15)$$

Să presupunem că variabilele aleatoare X_r sunt toate identic repartizate pentru fiecare valoare posibilă a *factorului de risc* θ . Fie

$$\mu(\Theta) = \mathrm{E}(X_r\,|\,\Theta) \qquad (13.1.16)$$

Aceasta este cea mai bună aproximare pe care o putem face pentru Xr în sensul celor mai mici pătrate. În cele ce urmează nu vom modifica notația: $\mu(\Theta)$ va avea mereu aceeaşi semnificație.

Sensul precis este că dintre toate funcțiile $\psi(\Theta)$ cu care am dori să aproximăm pe X_r în spațiul L^2, *cea pentru care distanța este minimă este* $\mu(\Theta)$.

Ceea ce ne interesează în actuariat este mărimea

$$\mathrm{E}(\mu(\Theta)\,|\,\underline{X}) \ \textit{notată cu}\ g(\underline{X})\,. \qquad (13.1.17)$$

Este mai puțin evident că, în anumite ipoteze, $g(\underline{X})$ este şi cea mai bună aproximare pe care o putem face asupra premiului brut de asigurare viitor (adică

pentru X_{t+1}) ținînd seama de modelul nostru bayesian şi de experiența acumulată, $\underline{X}$.

Definiția 13.1.8. *Două variabile aleatoare X şi Y se numesc* ***condiționat independente*** *față de* Θ *dacă* $P(X \in A, Y \in B \mid \Theta) = P(X \in A)P(Y \in B \mid \Theta)$ *pentru orice A şi B mulțimi boreliene.*

Exemplul 13.1.9. *În exemplele 13.1.1 şi 13.1.5 am presupus tacit că observațiile* $(X_i)_{1 \le i \le t}$ *sunt condiționat independente. Altfel nu puteam să spunem că* $P(X_1 = x_1,...,X_t = x_t) = P(X_1 = x_1)...P(X_t = x_t)$. *Dacă X şi Y sunt condiționat independente, nu rezultă că sunt independente. Într-adevăr, să spunem că X,Y,* Θ *sunt discrete. Atunci*
$P(X = i , Y = j) = \mathrm{E}(P(X = i , Y = j \mid \Theta) = E(P(X = i \mid \Theta)P(Y = j \mid \Theta))$
$\neq \mathrm{E}[P(X = i \mid \Theta)]\mathrm{E}[P(Y = j \mid \Theta)]$. *De exemplu, să zicem că* $(X,Y \mid \Theta) \sim \mathrm{U}(\{\Theta,\Theta+1\})$, $\Theta \sim \mathrm{U}(\{0,1\})$. *Atunci* $P(X = 0, Y = 0) = [P(X = 0, Y = 0 \mid \Theta = 0) + P(X = 0, Y = 0 \mid \Theta = 1)]/2 = [\ ½\ + 0]/2 = 1/8$ *iar* $P(Y = 0) = P(X = 0) = [P(X = 0 \mid \Theta = 0) + \mathrm{P}(X = 0 \mid \Theta = 1)]/2 = (½ + 0)/2 = ¼$. *Produsul este 1/16 şi nu 1/8.*

Propoziția 13.1.10. *Fie X şi Y două variabile aleatoare* ***condiționat independente de*** Θ. *Atunci avem*

$$\mathrm{E}(f(Y) \mid \Theta,X) = \mathrm{E}(f(Y) \mid \Theta) \tag{13.1.18}$$

În consecință

$$\mathrm{E}(f(Y) \mid X) = \mathrm{E}(\mathrm{E}(f(Y) \mid \Theta) \mid X) \tag{13.1.19}$$

Corolarul 13.1.11. *Presupunem că observațiile* X_r *sunt condiționat independente fiind dat* Θ. *Atunci* $\mathrm{E}(X_{t+1} \mid X_1,\ldots,X_t) = \mathrm{E}(\mu(\Theta) \mid X_1,\ldots,X_t) = g(\underline{X})$.

Demonstrație
Este de fapt relația (13.1.19) unde în loc de Y avem X_{t+1} iar în loc de X avem vectorul $\underline{X} = (X_r)_{1 \le r \le t}$. □

Principial, $g(\underline{X})$ se poate calcula, dacă ştim repartiția lui Θ condiționată de $\underline{X}$. O ipoteză destul de optimistă.

În Exemplul 13.1.2 – deci şi în exemplul 13.1.1 – cunoaştem această repartiție: este $\beta_{S+1,\ nt+1-S}$. Cum X_r sunt binomiale condiționat de Θ (adică $P(X_r = j \mid \Theta) = \mathit{Binomial}(n,\Theta)(\{j\})$ – rezultă că $\mu(\Theta) = n\Theta$. Atunci $g(\underline{X}) = \mathrm{E}(n\Theta \mid \underline{X}) = n\mathrm{E}(\Theta \mid \underline{X})$. Dar media unei variabile aleatoare $Y \sim \beta_{m,n}$ este $\frac{m}{m+n}$ de unde obținem estimatorul bayesian pentru X_{t+1} (premiul brut) ca fiind

$$g(\underline{X}) = \frac{n(S+1)}{nt+2} \qquad (13.1.20)$$

Să notăm cu $M = S/t$ media aritmetică a observațiilor (se ştie că M este un estimator nedeplasat şi eficient pentru EX_r în ipoteza că variabilele sunt i.i.d, ceea ce nu este cazul!). Vom nota de asemenea în mod consecvent

$$m = EX_r = E(E(X_r \mid \Theta)) = E(n\Theta) = n\Theta/2 \qquad (13.1.21)$$

(căci am acceptat că $\Theta \sim U(0,1)$). Cu aceste pregătiri putem scrie (1.1.20) sub forma

$$g(\underline{X}) = zM + (1\text{-}z)m \qquad (13.1.22)$$

unde $z = \frac{nt}{nt+2}$. *q.e.d.*

Observația 13.1.12. *Relația (13.1.21) este foarte atractivă: este simplă şi admite o interpretare intuitivă: cel mai bine este să prezicem viitorul sub forma unei mixturi între ideile noastre anterioare (= m) şi experiență (= M). Coeficientul z ne arată ponderea experienței. Dacă* $t \to \infty$, $z \to 1$, *adică e mai bine să ne bazăm pe experiență. Dacă t este mic, atunci este bine de luat în calcul şi modelul nostru teoretic.*

Se pune întrebarea : *nu cumva mereu* $g(\underline{X})$ este cuprins între m şi M ?

Vom da un exemplu că nu este aşa.

Exemplul 13.1.13. *Să presupunem că* $\Theta \sim U(0,1)$ *şi că variabilele aleatoare* X_r *sunt repartizate* $U(\theta,\theta+1)$ *Presupunem de asemenea că ele sunt condiționat independente dacă se ştie* θ. *Deci*

$E = [0,1]$, $\tau = \lambda$, $u(\theta) = 1_{(0,1)}(\theta)$ *(repartiția apriori)*

$Q(\theta) = U(\theta,\theta+1)^t$ *(acesta este modelul propriu zis!)*

$f_{\Theta,\underline{X}}(\theta,\underline{x}) = 1_{(\theta,\theta+1)}(x_1)\, 1_{(\theta,\theta+1)}(x_2)\ldots\, 1_{(\theta,\theta+1)}(x_t)\, 1_{(0,1)}(\theta) = 1_{A(\underline{x})}(\theta)$ *unde*

$A(\underline{x}) = (x_1\text{-}1,x_1) \cap (x_2\text{-}1,x_2) \cap \ldots \cap (x_t\text{-}1,x_t) \cap (0,1) = (x^*\text{-}1,x_*) \cap (0,1)$ *unde*

$x_* = x_1 \wedge x_2 \wedge \ldots \wedge x_t$, $x^* = x_1 \vee x_2 \vee \ldots \vee x_t$

$f_{\underline{X}}(\underline{x}) = \lambda(A(\underline{x})) = ((x_* \wedge 1) - (x^*\text{-}1)_+)_+$

$f_{\Theta \mid \underline{X}=\underline{x}}(\theta) = 1_{A(\underline{x})}/f_{\underline{X}}(\underline{x})$, *deci repartiția lui* Θ *condiționată de* $\underline{X}$ *este* $U(A(\underline{x}))$

Apoi $\mu(\Theta) = E(X_r \mid \Theta) = \Theta + ½$ *(media unei uniforme pe (a,b) este mijlocul intervalului (a+b)/2 ; în cazul nostru* $a = \Theta$ *şi* $b = \Theta+1$*!) deci* $g(\underline{X}) = E(\Theta \mid \underline{X}) + ½ = ((x_* \wedge 1) + (x^*\text{-}1)_+)/2 + ½$ *(căci şi repartiția lui* Θ *condiționată de X este tot uniformă!) . În concluzie*

$$g(\underline{X}) = \begin{cases} \dfrac{x_* + 1}{2} & daca \quad x^* \le 1 \\ \dfrac{x^* + x_*}{2} & daca \quad x_* < 1 < x^* \\ \dfrac{x^* + 1}{2} & daca \quad x_* \ge 1 \end{cases} . \qquad (13.1.23)$$

Pe de altă parte $m = EX_1 = E\mu(\Theta) = ½ + ½ = 1$. *Se poate ca* $g(\underline{X})$ *să nu fie între m şi M: de exemplu, dacă t = 3,* $\underline{x}$ *= (1.1; 1.2; 1.9) atunci M = (1.1+ 1.2 + 1.9)/3 = 1.4 <* $g(\underline{x})$ *= (1 + 1.9)/2 = 1.45.*

13.2. Modelul de credibilitate Bűhlmann

Fie $\underline{X}$ o selecție de volum t . Variabila aleatoare X_r , $1 \le r \le t$ reprezintă suma pe care asiguratorul a plătit-o în anul r. Bănuim că repartiția acestei variabile depinde de un factor de risc, Θ asupra căruia avem o credință – adică o repartiție apriori $U = u \cdot \tau$

Definiția 13.2.1. *Vom numi contract un vector* $(\Theta,\underline{X})$ *unde* $P\circ\Theta^{-1} = U$ şi $\underline{X} = (X_1,\ldots,X_t)$ *reprezintă variabile aleatoare din* L^2 *interpretate fiind ca o istorie a plăților făcute de asigurator la momentele de timp 1,2,…,t.*

Fie $\mu_r(\Theta) = E(X_r \mid \Theta)$

Ceea ce ne interesează este să dăm o predicție asupra plății viitoare X_{t+1} . Istoria plăților până în prezent (momentul t) este $\underline{X}$.

Ca de obicei, dacă nu facem unele ipoteze suplimentare, nu vom putea spune nimic în acest sens.

Vom face ipoteza că pentru fiecare valoare a factorului de risc Θ , variabilele aleatoare $(Xr)_{1 \le r \le t}$ sunt independente şi identic repartizate. Atunci şi variabilele $\mu_r(\Theta)$ vor coincide. Le vom nota cu $\mu(\Theta)$.

Ştim (Corolar 1.1.11) că $E(X_{t+1} \mid X_1,\ldots,X_t) = E(\mu(\Theta) \mid X_1,\ldots,X_t) = g(\underline{X})$

Aceasta este **ipoteza independenței condiționate**.

Scrisă precis, cu notațiile din paragraful anterior, ea devie

$$Q(\theta) = F(\theta)^{\Theta_t} \qquad (13.2.1)$$

adică

$$P(X_1 \in B_1,\ldots,X_t \in B_t \mid \Theta) = P(X_1 \in B_1 \mid \Theta)\ldots P(X_t \in B_t \mid \Theta)$$
$$= F(\Theta,B_1)\ldots F(\Theta,B_t) \;\forall\; B_r \in B(\Re),\ 1 \le r \le t \qquad (13.2.2)$$

Am văzut că semnificația lui $g(\underline{X})$ (**estimatorul Bayesian exact**) este următoarea : dacă notăm cu L mulțimea funcțiilor h: $\Re^t \to \Re$ care sunt măsurabile şi au proprietatea că $h(\underline{X}) \in L^2(\Omega,K,P)$ atunci

$$\| X_{t+1} - g(\underline{X}) \|_2 = \min \{ \| X_{t+1} - h(\underline{X}) \|_2 \mid h \in L \} \quad (13.2.3)$$

adică

$$E(X_{t+1} - g(\underline{X}))^2 = \min \{ E(X_{t+1} - h(\underline{X}))^2 \mid h \in L \} \quad (13.2.4)$$

Cu alte cuvinte g minimizează distanța pătratică între X_{t+1} şi $h(\underline{X})$.

Problema este că în cele mai multe modele realiste g este necalculabil.

Buhlmann a avut ideea să facă un compromis: să caute funcția *h* afină care să minimizeze membrul drept din (13.2.4). Cu alte cuvinte să caute h de forma $h(\underline{x}) = c_0 + <\underline{c}, \underline{x}>$ astfel ca $E(X_{t+1} - h(\underline{X}))^2$ să fie minim.

Să considerăm funcția $\varphi: \Re \times \Re^t \to \Re$ dată de

$$h(c_0, \underline{c}) = E(X_{t+1} - c_0 - c_1X_1 - c_2X_2 - \ldots - c_tX_t)^2 \quad (13.2.5)$$

Problema de optimizat devine:

Găsiți c_0, $\underline{c}$ ca $h(c_0,\underline{c})$ = minim (13.2.6)

De data aceasta problema este simplă. Este vorba de a găsi minimul unei forme pătratice convexe. Fiind strict convexă, are optim unic.

Derivăm *h* după c_0 şi punem condiția ca derivata să se anuleze. Găsim

$$-2E(X_{t+1} - c_0 - c_1X_1 - c_2X_2 - \ldots - c_tX_t) = 0 \quad (13.2.7)$$

(am derivat sub integrală, deoarece putem aplica criteriul lui Lebesgue de dominare: variabilele noastre sunt în L^2. Rezultă

$$c_0 = E(X_{t+1} - c_1X_1 - c_2X_2 - \ldots - c_tX_t). \quad (13.2.8)$$

Dar variabilele aleatoare X_r , fiind condiționat identic repartizate, sunt şi identic repartizate. Media lor se va nota, ca şi în primul capitol, cu $m = EX_r = E\mu(\Theta)$. Înlocuind în (13.2.7) EX_r cu *m* rezultă

$$c_0 = m(1- c_1- c_2 - \ldots - c_t) \quad (13.2.9)$$

Înlocuind în (13.2.5) găsim că avem de optimizat funcția

$$\varphi(\underline{c}) = E(X_{t+1} - m - c_1(X_1- m) - c_2(X_2 -m) - \ldots - c_t(X_t- m))^2 \quad (13.2.10)$$

Să notăm cu Y_r variabilele aleatoare centrate $Y_r = X_r - m$. Atunci funcția (convexă!) de optimizat devine

$$\varphi(\underline{c}) = E(Y_{t+1} - c_1 Y_1 - c_2Y_2 - \ldots - c_t Y_t)^2 \quad (13.2.11)$$

Gradientul ei este

$$\text{Grad } \varphi (\underline{c}) = (-2E(Y_r(Y_{t+1} - c_1 Y_1 - c_2 Y_2 - \ \ldots - c_t Y_t)))_{1 \le r \le t}. \qquad (13.2.12)$$

Ecuația Grad φ ($\underline{c}$) = 0 devine

$$\sum_{j=1}^{t} c_j \ E(Y_r Y_j) = E(Y_r Y_{t+1}) \qquad (1.2.13)$$

Pe de altă parte, dacă j ≠ r variabilele aleatoare Y_j şi Y_r sunt condiționat independente deci

$$E(Y_r Y_j) = E(E(Y_r Y_j \mid \Theta)) = E(E(Y_r \mid \Theta)E(Y_j \mid \Theta)) \qquad (13.2.14)$$

Dar $E(Y_j \mid \Theta) = E(X_j - m \mid \Theta) = E(X_j \mid \Theta) - m = \mu(\Theta) - m = \mu(\Theta) - E\mu(\Theta)$ de unde

$$r \neq j \Rightarrow E(Y_j Y_r) = Var\ \mu(\Theta) \qquad (13.2.15)$$

Vom nota $Var\ \mu(\Theta)$ cu a.

Dacă însă $r = j$ atunci $E(Y_r Y_j) = E(Y_r^2) = E(E(Y_r^2 \mid \Theta)) = E(E(X_r - m)^2 \mid \Theta)) = E(E[(X_r - E(X_r \mid \Theta) + (E(X_r \mid \Theta) - m)]^2 \mid \Theta)) = E(E[(X_r - \mu(\Theta)) + (\mu(\Theta) - m)]^2 \mid \Theta)) = E(E[(X_r - \mu(\Theta)]^2 \mid \Theta)) + 2E([X_r - \mu(\Theta)][\mu(\Theta) - m] \mid \Theta) + E(\mu(\Theta) - m)]^2) = E(Var(X_r \mid \Theta)) + 2E([X_r - \mu(\Theta)] \mid \Theta)[\mu(\Theta) - m]) + Var\ \mu(\Theta) = s^2 + 0 + a = a + s^2$.

Am notat $E(Var(X_r \mid \Theta))$ cu s^2. Cum variabilele aleatoare X_r sunt identic repartizate, notația este corectă . Altfel ar fi trebuit să punem s_r în loc de s. În concluzie

$$E(Y_j Y_r) = a + \delta_{j,r} s^2 \quad \forall\, j,r \in \{1,\ldots,t\} \qquad (13.2.15)$$

Înlocuind în (13.2.12) găsim sistemul

$$\sum_{j=1}^{t} c_j \ (a + \delta_{j,r} s^2) = a \qquad (13.2.16)$$

care se rezolvă foarte simplu. Adunînd toate ecuațiile rezultă $(ta + s^2)(c_1 + \ldots + c_t) = ta$ de unde suma coeficienților $S = c_1 + \ldots + c_t = \dfrac{ta}{ta + s^2}$. Cum sistemul se mai scrie $c_r s^2 + aS = a$, urmează că

$$c_1 = c_2 = \ \ldots = c_t = \frac{a}{ta + s^2}. \qquad (13.2.17)$$

Concluzia finală este

Teorema 13.2.2. *Dacă variabilelele aleatoare* $(X_r)_{r \ge 1}$ *sunt din* L^2 *şi i.i.d. condiționat de* Θ *, atunci estimatorul liniar optim* $h(\underline{X})$ *are forma* $h(\underline{X}) = (1-z)m + zM$ *unde*

$$z = \frac{at}{ta + s^2},\ a = Var\ \mu(\Theta),\ s^2 = \mathrm{E}(Var(X_r \mid \Theta)) \qquad (13.2.18)$$

Definiția 13.2.3 *Numărul z se numeşte* ***coeficientul de credibilitate al lui Buhlmann.***

El nu este o statistică , deoarece depinde de trei parametri neobservabili: a = Var $\mu(\Theta)$; $m = \mathrm{E}\mu(\Theta)$ şi $s^2 = \mathrm{E}(Var(X_r \mid \Theta))$.

Uneori se poate întîmpla să coincidă ca $h(\underline{X})$ să coincidă cu $g(\underline{X})$ – adică estimatorul liniar optim să fie chiar estimatorul bayesian optim.

Pe scurt: dacă avem observațiile $\underline{X} = (X_1,\ldots,X_t)$ şi modelul Bayesian

-$P(X_1 \in B_1,\ldots,X_t \in B_t \mid \Theta) = Q(\Theta,B_1)\ldots Q(\Theta,B_t)\ \forall\ B_r \in B(\Re),\ 1 \le r \le t$

- $Q(\theta)$ are densitatea $f_{\underline{X}\mid\Theta=\theta}\cdot \nu$, probabilitatea apriori cu densitatea $U = u\cdot\tau$

şi notăm $\mathrm{E}(X_r \mid \Theta)$ cu $\mu(\Theta)$, atunci

$$g(\underline{X}) = \mathrm{E}(\mu(\Theta) \mid \underline{X}) = \frac{\int \mu(\theta) f_{\underline{X}\mid\Theta=\theta}(\underline{X}) u(\theta) d\tau(\theta)}{\int f_{\underline{X}\mid\Theta=\sigma}(\underline{X}) u(\sigma) d\tau(\sigma)} = \mathrm{E}(X_{t+1} \mid \underline{X}) \qquad (13.2.19)$$

iar estimatorul Buhlman este $h(\underline{X}) = Mz + (1-z)m$ cu

$$m = \mathrm{E}X_r,\ M = (X_1 + \ldots + X_t)/t,\ z = \frac{at}{at + s^2},\ a = \mathrm{Var}\ \mu(\Theta),\ s^2 = \mathrm{E}(Var(X_1 \mid \Theta)) \qquad (13.2.20)$$

Tot demersul de pînă acum ar fi inutil dacă nu ar exista cazuri întîlnite în statistică care estimatorul Buhlman $h(\underline{X})$ ar coincide cu $g(\underline{X})$. Dăm acum o generalizare a exemplului 1, unde cele două chiar coincid.

Definiția 13.2.4. *Densitatea* $f_{\underline{X}\mid\Theta=\theta}$ *se numeşte familie exponențială dacă este de forma*

$$f_{X\mid\Theta=\theta}(x) = p(x)e^{-\theta x}/q(\theta) \qquad (13.2.21)$$

unde se subînțelege că spațiul parametrilor $E = [0,\infty)$*. Se presupune că funcția* $q(\theta)$ *este derivabilă.*

În acest caz densitatea vectorului $\underline{X}$ este

$$f_{\underline{X}\mid\Theta=\theta}(\underline{x}) = \frac{p(x_1)p(x_2)\ldots p(x_t)}{q^t(\theta)} e^{-\theta S} \text{ unde } S = x_1 + \ldots + x_t \qquad (13.2.22)$$

Să presupunem că densitatea apriori u a variabilei aleatoare Θ este de forma

$$u(\theta) = \frac{q(\theta)^{-\alpha} e^{-\beta\theta}}{C(\alpha,\beta)} \text{ unde } \alpha,\beta \in [0,\infty) \text{ iar } C(\alpha,\beta) \text{ este o constantă de normare.} \qquad (13.2.23)$$

În aceste condiții densitatea aposteriori este

$$f_{\Theta|\underline{X}=\underline{x}}(\theta)= \frac{f_{\underline{X}|\Theta=\theta}(\underline{x})u(\theta)}{\int f_{\underline{X}|\Theta=y}(\underline{x})u(y)d\tau(y)} = A(\alpha,\beta,\underline{x})e^{-\theta(\alpha+S)} / q^{t+\beta}(\theta) \qquad (13.2.24)$$

adică este de același tip ca și densitatea apriori. Spunem că familia aceasta de densități este o **familie conjugată**..

Propoziția 13.2.5. *Dacă modelul bayesian este familie exponențială, densitatea apriori este de forma* (1.2.22) *şi* $u(0) = u(\infty) = 0$ *atunci g şi h , definiți prin* (13.2.18) *şi* (13.2.19) *coincid.*

Un caz particular este dacă modelul este Poisson: repartiția Poisson este de forma (13.2.21). Aici măsura ν este măsura cardinal pe mulțimea numerelor naturale.

Se punem acum problema de a estima pe baza datelor de observație $\underline{X}$ cei trei parametri m, a şi s^2. Acum este o problemă de statistică obișnuită: căutăm trei estimatori nedeplasați pentru aceste cantități. Ideea lui Buhlman a fost de a se apela la mai multe contracte independente de același tip.

Un răspuns este următorul:

Propoziția 13.2.6. *Dacă vectorii* $\underline{X}_j = (X_{j,r})_{1\le r\le t}$, *sunt independenți şi acceptăm la fiecare din ei același model Q, atunci* M_0, $\hat{s}^2$ *şi* $\hat{a}$ *definiți mai jos sunt estimatori consistenți pentru m,* s^2 *şi a*

$$M_0 = \frac{1}{k}\sum_{j=1}^{k} M_j \qquad (13.2.25)$$

$$\hat{s}^2 = \frac{1}{k}\sum_{j=1}^{k}\hat{s}_j^2 \,,\ \hat{s}_j^2 = \frac{\sum_{r=1}^{t}(X_{j,r}-M_j)^2}{t-1} = \frac{\sum_{r=1}^{t}X_{j,r}^2 - tM_j^2}{t-1} \qquad (13.2.26)$$

$$\hat{a} = \frac{\sum_{j=1}^{k}(M_j - M_0)^2}{k-1} - \frac{\hat{s}^2}{t} \qquad (13.2.27)$$

Prin M_j *am notat mediile de selecție ale vectorilor* $\underline{X}_j$.

Mai mult, $\hat{s}^2$ *şi* $\hat{a}$ *sunt estimatori nedeplasați pentru* s^2. *Varianțele sunt*

$$Var(\hat{s}^2) = \frac{1}{k}Var(\hat{s}_1^2)\,,\ Var(M_j) = a + \frac{s^2}{t}\,,\ Var(M_0) = \frac{Var(M_j)}{k} \qquad (13.2.28)$$

Apare o evidentă deosebire față de cazul i.i.d., cînd aceste variante tind la 0 o dată cu creşterea numărului de observații, *t* .De asemenea vedem că nu contează aşa de mult t (= istoricul) cît contează k – numărul de contracte independente.

Corolarul 13.2.7. *Variabila aleatoare* $\hat{z} = \dfrac{\hat{a}t}{\hat{a}t + \hat{s}^2}$ *este un estimator pentru coeficientul de credibilitate z care este consistent în k : adică* $\mathrm{k} \to \infty \Rightarrow \hat{z} \to \mathrm{z}$.

Observația 13.2.8. *În general estimatorul* $\hat{z}$ *nu este nedeplasat, căci nu avem motive să credem că o formulă de tipul* $E\dfrac{X}{X+Y} = \dfrac{EX}{EX+EY}$ *ar putea fi adevărată, chiar în ipoteze restrictive. Dacă X şi Y sunt independente, de exemplu, atunci* $E\dfrac{X}{X+Y}$ *se poate calcula, este diferit de* $\dfrac{EX}{EX+EY}$.

Ca amuzament, dacă *X şi Y* sunt i.i.d. atunci egalitatea este adevărată*!* Ambele valori coincid, în mod evident, cu ½ !

Bibliografie

1. Bărbosu D., Zelina I., *Calculul probabilităţilor*, Editura CUB PRESS 22, Baia Mare, 1998
2. Beganu G. (coordonator), *Teoria probabilităţilor şi statistică matematică, Culegere de probleme*, Editura Meteor Press, Bucureşti 2004
3. Blaga P., *Calculul probabilităţilor şi statistică matematică. Curs şi culegere de probleme*, Universitatea „Babeş-Bolyai" Cluj – Napoca, 1994
4. Blaga P., *Statistică matematică. Ediţia a II-a*, Universitatea „Babeş-Bolyai", Cluj-Napoca, 2001
5. Breaz N., *Modele de regresie bazate pe funcţii spline*, Presa Universitară Clujeană, 2007
6. Breaz N., Jaradat M., *Statistică descriptivă, teorie şi aplicaţii*, Ed. Risoprint, Cluj-Napoca, 2009
7. Căbulea L., *Aproximare în probabilităţi şi statistică*, Editura Aeternitas, Alba Iulia, 2003
8. Căbulea L., Aldea M., *Elemente de teoria probabilităţilor şi statistică matematică*, Editura Didactica, Alba Iulia, 2004
9. Ciucu G., *Elemente de teoria probabilităţilor şi statistică matematică*, Editura Didactică şi Pedagogică, Bucureşti, 1963
10. Ciucu G., Craiu V., *Introducere în teoria probabilităţilor şi statistică matematică*, Editura Didactică şi Pedagogică, Bucureşti, 1971
11. Ciucu G., Craiu V., *Inferenţa statistică,* Editura Didactică şi Pedagogică, Bucureşti, 1974
12. Ciucu G., Craiu V., Săcuiu I., *Culegere de probleme de teoria probabilităţilor*, Editura Tehnică, 1967
13. Ciucu G., Sâmboan G., *Teoria probabilităţilor şi statistică matematică*, Editura Tehnică, Bucureşti, 1962
14. Cojocaru N., Clocotici V., Dobra D., *Metode statistice aplicate în industria textilă*, Editura Tehnică, Bucureşti, 1986
15. Florea I., *Econometrie*, Editura Universităţii din Oradea, 2003
16. Florea I., Parpucea I., Buiga A., *Statistică descriptivă, teorie si aplicaţii*, Editura Continental, Alba Iulia, 1998
17. Florea I., Parpucea I., Buiga A., Lazăr D., *Statistică inferenţială*, Presa Universitară Clujeană, 2000
18. Hoel P.G., *Introduction to Mathematical Statistics*, John Wiley, New York,1971
19. Lehman E.L., *Testing Statistical Hypotheses*, Second edition, Springer, New York-Berlin, 1997

20. Luca-Tudorache R., *Probleme de teoria probabilităţilor*, Editura Tehnopress, Iaşi, 2006.
21. Luca-Tudorache R., *Probleme de analiză matematică. Calcul integral*, Casa de Editură Venus, Iaşi, 2007
22. Mihoc Gh., Micu N., *Teoria probabilităţilor şi statistică matematică*, Editura Didactică şi Pedagogică, Bucureşti, 1980
23. Mihoc I., Fătu C.I., *Calculul probabilităţilor şi statistică matematică*, Casa de editură-Transilvania Press, Cluj – Napoca 2003
24. Moon T.K., Stirling W.C., *Mathematical Methods and Algorithms for Signal Processing*, Prentice Hall, 2000
25. Nicolescu L. J., Stoka M.I., *Matematică pentru ingineri*, vol. II, Editura Tehnică, Bucureşti, 1971
26. Pitea A., Postolache M., *Basic Concepts of Probability and Statistics*, Editura Fair Partners, 2007
27. Reischer C., Sâmboan A., *Culegere de probleme de teoria probabilităţilor şi statistică matematică*, Editura Didactică şi Pedagogică, Bucureşti, 1972
28. Reischer C., Sâmboan G., Teodorescu T., *Teoria probabilităţilor*, Editura Didactică şi Pedagogică, Bucureşti, 1967
29. Saporta G., *Probabilités, analyse des données et statistique*, Editions Technip, Paris, 1990
30. Stapleton J.H., *Linear Statistical Models*, John Wiley & Sons, New York-Chichester-Brisbane, 1995
31. Stark H., Woods J.W., *Probability, Random Processes and Estimation Theory for Engineers*, Prentice Hall, 1986
32. Şabac I. Gh. et all., *Matematici speciale*, vol. II, Editura Didactică şi Pedagogică, Bucureşti, 1983
33. Şabac I. Gh., *Matematici speciale*, vol. II, Editura Didactică şi Pedagogică, Bucureşti, 1965
34. Trandafir R., *Introducere în teoria probabilităţilor*, Editura Albatros, 1979
35. Trâmbiţaş R., *Metode statistice*, Presa Universitară Clujeană, Universitatea „Babeş-Bolyai", Cluj-Napoca, 2000
36. Zbăganu Gh., *Teoria măsurii şi a probabilităţilor*, Editura Universităţii, Bucureşti, 1998
37. Zbăganu Gh., *Metode matematice în teoria riscului şi actuariat*, Editura Universităţii, Bucureşti, 2004

ANEXE

Anexa 1

Valorile funcției Laplace - Gauss

u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7290	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9779	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Anexa 2

Cuantilele repartiției Student

grade\prob.	0.75	0.90	0.95	0.975	0.99	0.995
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.695	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.474	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.319	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797
25	0.684	1.316	1.708	2.060	2.485	2.787
26	0.684	1.315	1.706	2.056	2.479	2.779
27	0.684	1.314	1.703	2.052	2.473	2.771
28	0.683	1.313	1.701	2.048	2.467	2.763
29	0.683	1.311	1.699	2.045	2.462	2.756
30	0.683	1.310	1.697	2.042	2.457	2.750
35	0.681	1.306	1.690	2.030	2.438	2.724
40	0.681	1.303	1.684	2.021	2.423	2.704
80	0.679	1.291	1.671	2.000	2.390	2.660
120	0.677	1.289	1.658	1.980	2.358	2.617
n>120	0.674	1.282	1.645	1.960	2.326	2.576

Anexa 3

Cuantilele repartiției χ^2

grade\prob.	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99
1	0.000	0.000	0.001	0.004	0.016	2.71	3.84	5.02	6.63
2	0.010	0.020	0.051	0.103	0.211	4.60	5.99	7.38	9.21
3	0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34
4	0.207	0.297	0.484	0.711	1.06	7.78	9.48	11.1	13.28
5	0.412	0.554	0.831	1.15	1.61	9.24	11.07	12.8	15.09
6	0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.4	16.81
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.0	18.47
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.5	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.0	21.66
10	2.16	2.56	3.25	3.94	4.87	16.99	18.31	20.5	23.21
11	2.60	3.05	3.82	4.57	5.58	17.27	19.67	21.9	24.72
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.3	26.22
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.7	27.69
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.1	29.14
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.6	30.58
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.8	32.00
17	5.70	6.41	7.56	8.67	10.08	24.77	27.59	30.2	33.41
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.3	34.80
19	6.84	7.63	8.91	10.1	11.65	27.20	30.14	32.9	36.19
20	7.43	8.26	9.59	10.9	12.44	28.41	31.41	34.2	37.57
21	8.03	8.90	10.3	11.6	13.24	29.61	32.67	35.5	38.93
22	8.64	9.54	11.0	12.3	14.04	30.81	33.92	36.8	40.29
23	9.26	10.2	11.7	13.1	14.85	32.01	35.17	38.1	41.64
24	9.89	10.9	12.4	13.8	15.66	33.20	36.41	39.4	42.98
25	10.5	11.5	13.1	14.6	16.47	34.38	37.65	40.6	44.31
26	11.2	12.2	13.8	15.4	17.29	35.56	38.88	41.9	45.64
27	11.8	12.9	14.6	16.2	18.11	36.74	40.11	43.2	46.96
28	12.5	13.6	15.3	16.9	18.94	37.92	41.34	44.5	48.28
29	13.1	14.3	16.0	17.7	19.77	39.09	42.56	45.7	49.59
30	13.8	15.0	16.8	18.5	20.60	40.26	43.77	47.0	50.89
35	17.2	18.5	20.6	22.5	24.8	46.1	49.8	53.2	57.3
40	20.7	22.2	24.4	26.5	29.1	51.8	55.8	59.3	63.7
60	35.5	37.5	40.5	43.2	46.5	74.4	79.1	83.3	88.4

Anexa 4

Cuantilele repartiției Fisher - Snedecor

Gr2\Gr1	Prob.	1	2	3	4	5	6	7	8
	0.95	161.4	199.5	216	225	230	234	237	239
1	0.975	648	800	864	900	922	937	948	957
	0.99	4052	4999	5403	5625	5764	5859	5930	5981
	0.95	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37
2	0.975	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4
	0.99	98.49	99.00	99.17	99.25	99.30	99.33	99.35	99.36
	0.95	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85
3	0.975	17.4	16.0	15.1	15.4	14.9	14.7	14.6	14.5
	0.99	34.12	30.84	29.46	28.71	28.24	27.91	27.7	27.49
	0.95	17.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04
4	0.975	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98
	0.99	21.20	18.00	16.69	15.98	15.52	15.21	15.0	14.80
	0.95	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82
5	0.975	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76
	0.99	16.26	13.27	12.06	11.39	10.97	10.67	10.5	8.10
	0.95	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15
6	0.975	8.07	7.26	6.60	6.23	5.99	5.82	5.70	5.60
	0.99	12.25	10.91	9.78	9.15	8.75	8.47	8.26	8.10
	0.95	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73
7	0.975	8.07	9.54	5.89	5.52	5.29	5.12	4.99	4.90
	0.99	12.25	9.55	8.45	7.85	7.45	7.19	6.99	6.84
	0.95	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44
8	0.975	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43
	0.99	11.26	8.65	7.59	1.01	6.63	6.37	6.18	6.03
	0.95	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23
9	0.975	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10
	0.99	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47
	0.95	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07
10	0.975	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85
	0.99	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06
	0.95	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95
11	0.975	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66
	0.99	9.65	7.20	6.22	5.67	5.32	5.07	4.89	4.74
	0.95	4.75	3.88	3.49	3.26	3.11	3.00	2.91	2.85
12	0.975	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51
	0.99	9.33	6.93	5.95	5.41	5.06	4.82	4.54	4.50
	0.95	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77
13	0.975	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39
	0.99	9.07	6.70	5.74	5.221	4.86	4.62	4.44	4.30
	0.95	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70
14	0.975	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29
	0.99	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14

15	0.95	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64
	0.975	6.20	4.76	4.15	3.80	3.58	3.41	3.29	3.20
	0.99	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00
16	0.95	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59
	0.975	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12
	0.99	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89
17	0.95	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55
	0.975	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06
	0.99	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79
18	0.95	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51
	0.975	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01
	0.99	8.29	7.01	5.09	4.58	4.25	4.01	3.84	3.71
19	0.95	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48
	0.975	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96
	0.99	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63
20	0.95	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45
	0.975	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91
	0.99	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56
21	0.95	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42
	0.975	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87
	0.99	8.02	5.78	4.87	4.37	4.04	3.84	3.64	3.51
22	0.95	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40
	0.975	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84
	0.99	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45
23	0.95	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37
	0.975	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81
	0.99	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41
24	0.95	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36
	0.975	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78
	0.99	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36

Anexa 4 (continuare)

Gr2\Gr1	Prob.	9	10	11	12	13	14	15	16
2	0.95	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4
	0.975	39.4	39.4	39.4	39.4	39.4	39.4	39.4	39.4
	0.99	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.4
3	0.95	8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.69
	0.975	14.5	14.4	14.4	14.3	14.3	14.3	14.3	14.2
	0.99	27.3	27.1	27.1	27.1	27.0	26.9	26.9	26.8
4	0.95	6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.84
	0.975	8.90	8.84	8.79	8.75	8.72	8.69	8.66	8.64
	0.99	14.7	14.5	14.4	14.4	14.3	14.2	14.2	14.2

5	0.95	4.77	4.74	4.70	4.68	4.66	4.64	4.62	4.60
	0.975	6.68	6.62	6.57	6.52	6.49	6.46	6.43	6.41
	0.99	10.2	10.1	9.96	9.89	9.82	9.77	9.72	9.68
6	0.95	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.92
	0.975	5.52	5.46	5.41	5.37	5.33	5.30	5.27	5.25
	0.99	7.98	7.89	7.79	7.72	7.66	7.60	7.56	7.52
7	0.95	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.49
	0.975	4.82	4.76	4.71	4.67	4.63	4.60	4.57	4.54
	0.99	6.72	6.62	6.54	6.47	6.41	6.36	6.31	6.27
8	0.95	3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.20
	0.975	4.36	4.30	4.24	4.20	4.16	4.13	4.10	4.08
	0.99	5.91	5.81	5.73	5.67	5.61	5.56	5.52	5.48
9	0.95	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.99
	0.975	4.03	3.96	3.91	3.87	3.83	3.80	3.77	3.74
	0.99	5.35	5.26	5.18	5.11	5.05	5.00	4.96	4.90
10	0.95	3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.83
	0.975	3.78	3.72	3.66	3.62	3.58	3.55	3.52	3.50
	0.99	4.94	4.85	4.77	4.71	4.65	4.60	4.56	4.52
11	0.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.70
	0.975	3.59	3.53	3.47	3.43	3.39	3.36	3.33	3.30
	0.99	4.63	4.54	4.46	4.40	4.34	4.29	4.25	4.21
12	0.95	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.60
	0.975	3.44	3.37	3.32	3.28	3.24	3.21	3.18	3.15
	0.99	4.39	4.30	4.22	4.16	4.10	4.05	4.01	3.97
13	0.95	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.51
	0.975	3.31	3.25	3.20	3.15	3.12	3.08	3.05	3.03
	0.99	4.19	4.10	4.02	3.96	3.91	3.86	3.82	3.78
14	0.95	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.44
	0.975	3.21	3.15	3.09	3.05	3.01	2.98	2.95	2.92
	0.99	4.03	3.94	3.86	3.80	3.75	3.70	3.66	3.62
15	0.95	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.38
	0.975	3.12	3.06	3.01	2.96	2.92	2.89	2.86	2.84
	0.99	3.89	3.80	3.73	3.67	3.61	3.56	3.52	3.49
16	0.95	2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.33
	0.975	3.05	2.99	2.93	2.89	2.85	2.82	2.79	2.76
	0.99	3.78	3.69	3.62	3.55	3.50	3.45	3.41	3.37
17	0.95	2.49	2.45	2.41	2.38	2.35	2.33	2.29	2.27
	0.975	2.98	2.92	2.87	2.82	2.79	2.75	2.72	2.70
	0.99	3.68	3.59	3.52	3.46	3.40	3.35	3.31	3.27
18	0.95	2.46	2.41	2.37	2.34	2.31	2.29	2.27	2.25
	0.975	2.93	2.87	2.81	2.77	2.73	2.70	2.67	2.64
	0.99	3.60	3.51	3.43	3.37	3.32	3.27	3.23	3.19
19	0.95	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.21
	0.975	2.88	2.82	2.76	2.72	2.68	2.65	2.62	2.59
	0.99	3.52	3.43	3.36	3.30	3.24	3.19	3.15	3.12
20	0.95	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.19
	0.975	2.84	2.77	2.72	2.68	2.64	2.60	2.57	2.55
	0.99	3.46	3.37	3.29	3.23	3.18	3.13	3.09	3.05

21	0.95	2.37	2.32	2.28	2.25	2.22	2.20	2.18	2.16
	0.975	2.80	2.73	2.68	2.64	2.60	2.56	2.53	2.51
	0.99	3.40	3.31	3.24	3.17	3.12	3.07	3.03	2.99
22	0.95	2.34	2.30	2.26	2.23	2.20	2.17	2.15	2.13
	0.975	2.76	2.70	2.65	2.60	2.56	2.53	2.50	2.47
	0.99	3.35	3.26	3.18	3.12	3.07	.302	2.98	2.94
23	0.95	2.32	2.27	2.23	2.20	2.18	2.15	2.13	2.11
	0.975	2.73	2.67	2.62	2.57	2.53	2.50	2.47	2.44
	0.99	3.30	3.21	3.14	3.07	3.02	2.97	2.93	2.89
24	0.95	2.30	2.25	2.21	2.18	2.15	2.13	2.11	2.09
	0.975	2.70	2.64	2.69	2.54	2.50	2.47	2.44	2.41
	0.99	3.26	3.17	3.09	3.03	2.98	2.93	2.89	2.85