Probabilităţi şi statistică

Conf. dr. Cristian Niculescu,
Facultatea de Matematică şi Informatică,
Universitatea din Bucureşti

October 5, 2015

Part I

Probabilităţi

1 Câmp de probabilitate. Operaţii cu evenimente şi formule de calcul pentru probabilităţile acestora. Evenimente independente. Probabilitatea condiţionată. Formula lui Bayes

1.1 Câmp de probabilitate

În teoria probabilităţilor considerăm un experiment cu un rezultat dependent de şansă, care e numit *experiment aleator.* Se presupune că toate rezultatele posibile ale unui experiment aleator sunt cunoscute şi ele sunt elemente ale unei mulţimi fundamentale denumită ca *spaţiul probelor.* Fiecare rezultat posibil este numit *probă* şi un *eveniment* este o submulţime a spaţiului probelor.

Notaţii. Fie Ω mulţime.
$\mathcal{P}(\Omega) := \{A | A \subseteq \Omega\}$.
Fie $A \subseteq \Omega$.
$\overline{A} = \mathcal{C}A := \{a \in \Omega | a \notin A\}$.

Definiţia 1.1. Fie Ω mulţime. $\mathcal{K} \subseteq \mathcal{P}(\Omega)$ se numeşte *corp borelian* sau *σ-algebră* pe Ω dacă şi numai dacă
1) $\mathcal{K} \neq \emptyset$
2) $A \in \mathcal{K} \implies \overline{A} \in \mathcal{K}$
3) $A_1, A_2, ..., A_n, ... \in \mathcal{K} \implies \bigcup_n A_n \in \mathcal{K}$.

$(\Omega, \mathcal{K})$ se numeşte *spaţiu măsurabil* când $\mathcal{K}$ este corp borelian pe Ω.

Proprietăţi. Dacă $(\Omega, \mathcal{K})$ este *spaţiu măsurabil* atunci:

a) $\Omega \in \mathcal{K}$

b) $\emptyset \in \mathcal{K}$

c) $A_1, A_2, ..., A_n \in \mathcal{K} \implies \bigcup_{i=1}^{n} A_i \in \mathcal{K}$.

d) I cel mult numărabilă (i.e. finită sau numărabilă), $A_i \in \mathcal{K}, \forall i \in I \implies \bigcap_{i \in I} A_i \in \mathcal{K}$

e) $A, B \in \mathcal{K} \implies A \backslash B \in \mathcal{K}$.

Demonstraţie. a) $\mathcal{K} \neq \emptyset \implies \exists A \in \mathcal{K} \implies \overline{A} \in \mathcal{K} \implies \Omega = A \cup \overline{A} \cup A \cup A \cup ... \in \mathcal{K}$.

b) $\emptyset = \overline{\Omega} \in \mathcal{K}$.

c) $\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} A_i \cup \emptyset \cup \emptyset \cup ... \in \mathcal{K}$.

d) $\bigcap_{i \in I} A_i = \overline{\bigcup_{i \in I} \overline{A_i}} \in \mathcal{K}$.

e) $A \backslash B = A \cap \overline{B} \in \mathcal{K}$.□

Considerăm un corp borelian pe $\mathcal{K}$ pe un spaţiu Ω de elemente $a, b, c, ...$ cu $\{a\}, \{b\}, \{c\}, ... \in \mathcal{K}$ şi cu submulţimile $A, B, C, ... \in \mathcal{K}$. Unele dintre corespondenţele dintre teoria mulţimilor şi teoria probabilităţilor sunt date în următorul tabel:

Teoria mulţimilor	Teoria probabilităţilor
Spaţiu, Ω	Spaţiul probelor, eveniment sigur
Mulţimea vidă, $\emptyset$	Eveniment imposibil
Elemente $a, b, ...$	Probe $a, b, ...$ (sau evenimente simple)
Mulţimi $A, B, ...$	Evenimente $A, B, ...$
A	Evenimentul A apare
$\overline{A}$	Evenimentul A nu apare
$A \cup B$	Cel puţin unul dintre A şi B apare
$A \cap B$	Ambele A şi B apar
$A \subseteq B$	A este un subeveniment al lui B (i.e. apariţia lui A implică apariţia lui B)
$A \cap B = \emptyset$	A şi B sunt mutual exclusive (i.e. ele nu pot apărea simultan)

$\emptyset$ este considerată un eveniment imposibil deoarece niciun rezultat posibil nu este element al ei. Prin "apariţia unui eveniment" înţelegem că rezultatul observat este un element al acelei mulţimi. Spunem că mai multe evenimente sunt mutual exclusive dacă mulţimile corespunzătoare sunt disjuncte două câte două.

Exemplul 1.1. Considerăm un experiment de calculare a numărului de maşini care virează la stânga la o intersecţie dintr-un grup de 100 de maşini. Rezultatele posibile (numerele posibile de maşini care virează la stânga) sunt $0, 1, 2, ..., 100$. Atunci, spaţiul probelor este $\Omega = \{0, 1, 2, ..., 100\}$ şi $\mathcal{K} = \mathcal{P}(\Omega)$. $A = \{0, 1, 2, ..., 50\}$ este evenimentul "cel mult 50 de maşini virează la stânga". $B = \{40, 41, ..., 60\}$ este evenimentul "între 40 şi 60 (inclusiv) de maşini virează la stânga". $A \cup B$ este evenimentul "cel mult 60 de maşini virează la stânga". $A \cap B$ este evenimentul "între 40 şi 50 (inclusiv) de maşini virează la stânga".

Fie $C = \{81, 82, ..., 100\}$. Evenimentele A şi C sunt mutual exclusive.

Definiţia 1.2. Fie $(\Omega, \mathcal{K})$ spaţiu măsurabil. Funcţia $P : \mathcal{K} \to \mathbb{R}$ se numeşte *probabilitate* pe $(\Omega, \mathcal{K})$ dacă şi numai dacă are următoarele proprietăţi (numite axiomele probabilităţii):

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

Axioma 2: $P(\Omega) = 1$ (normată).

Axioma 3: pentru orice colecţie numărabilă de evenimente mutual exclusive (mulţimi disjuncte două câte două) $A_1, A_2, ... \in \mathcal{K}$,

$P\left(\bigcup_j A_j\right) = \sum_j P(A_j)$ (numărabil aditivă).

$(\Omega, \mathcal{K}, P)$ se numeşte *câmp de probabilitate* dacă şi numai dacă P este probabilitate pe spaţiul măsurabil $(\Omega, \mathcal{K})$.

1.2 Operaţii cu evenimente şi formule de calcul pentru probabilităţile acestora

Proprietăţi. Dacă $(\Omega, \mathcal{K}, P)$ este câmp de probabilitate, atunci:

1) $P(\emptyset) = 0$.

2) pentru orice colecţie finită de evenimente mutual exclusive (mulţimi disjuncte două câte două) $A_1, A_2, ..., A_n \in \mathcal{K}$,

$P\left(\bigcup_{j=1}^{n} A_j\right) = \sum_{j=1}^{n} P(A_j)$ (P este aditivă).

3) $A \subseteq C; A, C \in \mathcal{K} \implies P(A) \leq P(C)$.

4) $A, B \in \mathcal{K} \implies$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

5) (Formula lui Poincare) $A_1, A_2, ..., A_n \in \mathcal{K} \implies$

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

În particular, $A, B, C \in \mathcal{K} \implies$

$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$.

6) $A \in \mathcal{K} \implies P\left(\overline{A}\right) = 1 - P(A)$.

Demonstraţie. 1) $P(\Omega) = P(\Omega \cup \emptyset \cup \emptyset \cup ...) \overset{A3}{=} P(\Omega) + P(\emptyset) + P(\emptyset) + ... \overset{A2}{\implies} 1 = 1 + P(\emptyset) + P(\emptyset) + ... \implies P(\emptyset) + P(\emptyset) + ... = 0 \implies P(\emptyset) = 0.$

2) $P\left(\bigcup_{j=1}^{n} A_j\right) = P\left(\bigcup_{j=1}^{n} A_j \cup \emptyset \cup \emptyset \cup ...\right) \overset{A3}{=} \sum_{j=1}^{n} P(A_j) + P(\emptyset) + P(\emptyset) +$

$... \overset{1)}{=} \sum_{j=1}^{n} P(A_j) + 0 + 0 + ... = \sum_{j=1}^{n} P(A_j)$.

3) $P(C) = P(A \cup (C \backslash A)) \overset{2)}{=} P(A) + \underbrace{P(C \backslash A)}_{\geq 0} \Longrightarrow P(C) \geq P(A)$.

4) $P(A \cup B) = P(A \cup (B \backslash A)) \overset{2)}{=} P(A) + P(B \backslash A) \overset{2)}{=} P(A) + P(B) - P(A \cap B)$.

5) Inducţie.

Pentru $n = 1$ e evident.

Presupunem adevărat pentru n.

Fie $A_1, A_2, ..., A_n, A_{n+1} \in \mathcal{K}$.

$$P\left(\bigcup_{j=1}^{n+1} A_j\right) = P\left(\bigcup_{j=1}^{n} A_j \cup A_{n+1}\right) \overset{4)}{=} P\left(\bigcup_{j=1}^{n} A_j\right) + P(A_{n+1}) - P\left(\left(\bigcup_{j=1}^{n} A_j\right) \cap A_{n+1}\right) \overset{\text{ip. ind.}}{=}$$

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

$$P(A_{n+1}) - P\left(\bigcup_{j=1}^{n} (A_j \cap A_{n+1})\right) \overset{\text{ip. ind.}}{=}$$

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

$$[\sum_{j=1}^{n} P(A_j \cap A_{n+1}) - \sum_{1 \leq i < j \leq n} P((A_i \cap A_{n+1}) \cap (A_j \cap A_{n+1})) +$$

$$\sum_{1 \leq i < j < k \leq n} P((A_i \cap A_{n+1}) \cap (A_j \cap A_{n+1}) \cap (A_k \cap A_{n+1})) - ... + (-1)^{n-1} P\left(\bigcap_{j=1}^{n} (A_j \cap A_{n+1})\right)] =$$

$$\sum_{j=1}^{n+1} P(A_j) - \sum_{1 \leq i < j \leq n} P(A_i \cap A_j) - \sum_{j=1}^{n} P(A_j \cap A_{n+1}) + \sum_{1 \leq i < j < k \leq n} P(A_i \cap A_j \cap A_k) +$$

$$\sum_{1 \leq i < j \leq n} P(A_i \cap A_j \cap A_{n+1}) - ... + (-1)^n P\left(\bigcap_{j=1}^{n+1} A_j\right) =$$

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

6) $1 = P(\Omega) = P\left(A \cup \overline{A}\right) \overset{2)}{=} P(A) + P\left(\overline{A}\right)$. $\square$

Exemplul 1.2. În exemplul 1.1, presupunem că probabilităţile $P(A), P(B)$ şi $P(C)$ sunt cunoscute. Vrem să calculăm $P(A \cup B)$ (probabilitatea ca cel mult 60 de maşini să vireze la stânga) şi $P(A \cup C)$ (probabilitatea ca cel mult 50 de maşini sau între 80 şi 100 de maşini să vireze la stânga). Deoarece A şi C sunt mutual exclusive avem

$P(A \cup C) = P(A) + P(C)$.

Dar

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Informaţia dată este insuficientă pentru a determina $P(A \cup B)$ şi avem nevoie de informaţia adiţională $P(A \cap B)$, care este probabilitatea ca între 40 şi 50 de maşini să vireze la stânga.

1.3 Evenimente independente

Definiţia 1.3. Fie $(\Omega, \mathcal{K}, P)$ câmp de probabilitate. Două evenimente $A, B \in \mathcal{K}$ se numesc *independente* dacă şi numai dacă

$$P(A \cap B) = P(A) P(B).$$

Observaţie. Dacă A şi B sunt evenimente independente, atunci:

$P\left(A \cap \overline{B}\right) = P(A)P\left(\overline{B}\right)$,

$P\left(\overline{A} \cap B\right) = P\left(\overline{A}\right) P(B)$,

$P\left(\overline{A} \cap \overline{B}\right) = P\left(\overline{A}\right) P\left(\overline{B}\right)$.

Demonstraţie. $P(A) = P\left((A \cap B) \cup \left(A \cap \overline{B}\right)\right) = P(A \cap B) + P\left(A \cap \overline{B}\right) \Longrightarrow$ $P\left(A \cap \overline{B}\right) = P(A) - P(A \cap B) = P(A) - P(A) P(B) = P(A)(1 - P(B)) =$ $P(A)P\left(\overline{B}\right)$.

Analog pentru celelalte relaţii.□

Exemplul 1.3. În lansarea unui satelit, probabilitatea unui insucces este q. Care este probabilitatea ca două lansări succesive să eşueze?

Presupunând că lansările satelitului sunt evenimente independente, răspunsul este q^2.

Definiţia 1.4. Fie $(\Omega, \mathcal{K}, P)$ câmp de probabilitate. Evenimentele $A_1, A_2, ..., A_n \in \mathcal{K}$ sunt *independente* dacă şi numai dacă $\forall m = 2, 3, ..., n; k_1, k_2, ..., k_m \in \mathbb{N}$ a. î. $1 \leq k_1 < k_2 < ... < k_m \leq n$,

$$P(A_{k_1} \cap A_{k_2} \cap ... \cap A_{k_m}) = P(A_{k_1}) P(A_{k_2}) ... P(A_{k_m}).$$

În particular, A_1, A_2, A_3 sunt independente dacă şi numai dacă

$P(A_j \cap A_k) = P(A_j) P(A_k), \forall j < k; j, k = 1, 2, 3,$

şi

$P(A_1 \cap A_2 \cap A_3) = P(A_1) P(A_2) P(A_3)$.

Observaţii. 1) Numărul de egalităţi din definiţia independenţei a n evenimente este $2^n - n - 1$.

2) Independenţa două câte două nu conduce în general la independenţă.

Contraexemplu. Fie 3 evenimente A_1, A_2, A_3 definite de

$A_1 = B_1 \cup B_2, A_2 = B_1 \cup B_3, A_3 = B_2 \cup B_3,$

unde B_1, B_2 şi B_3 sunt mutual exclusive, fiecare având probabilitatea $\frac{1}{4}$.

$P(A_1) = P(B_1 \cup B_2) = P(B_1) + P(B_2) = \frac{1}{2}$.

Analog $P(A_2) = P(A_3) = \frac{1}{2}$.

$P(A_1 \cap A_2) = P((B_1 \cup B_2) \cap (B_1 \cup B_3)) = P(B_1 \cup (B_2 \cap B_3)) = P(B_1 \cup \emptyset) =$ $P(B_1) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A_1) P(A_2)$.

Analog $P(A_1 \cap A_3) = P(A_1) P(A_3), P(A_2 \cap A_3) = P(A_2) P(A_3)$.

$P(A_1 \cap A_2 \cap A_3) = P((B_1 \cup B_2) \cap (B_1 \cup B_3) \cap (B_2 \cup B_3)) = P((B_1 \cup (B_2 \cap B_3)) \cap (B_2 \cup B_3)) = P(B_1 \cap (B_2 \cup B_3)) = P((B_1 \cap B_2) \cup (B_1 \cap B_3)) = P(\emptyset \cup \emptyset) = P(\emptyset) = 0$.

$P(A_1) P(A_2) P(A_3) = \frac{1}{8}$.

Deci $P(A_1 \cap A_2 \cap A_3) \neq P(A_1) P(A_2) P(A_3)$.

Evenimentele A_1, A_2, A_3 sunt independente două câte două, dar nu sunt independente.

3) Dacă evenimentele $A_1, A_2, ..., A_n$ sunt independente, atunci înlocuind oricare din A_{k_j} cu complementara $\overline{A_{k_j}}$ în ambii membri din relaţiile din definiţia independenţei, relaţiile obţinute rămân valabile.

Exemplul 1.4. Un sistem compus din 5 componente merge exact atunci când fiecare componentă e bună. Fie $S_i, i = 1, ..., 5$, evenimentul "componenta i e bună" şi presupunem $P(S_i) = p_i$. Care e probabilitatea q ca sistemul să nu meargă?

Presupunând că cele 5 componente merg într-o manieră independentă, fie p probabilitatea de succes.

$$q = 1 - p = 1 - P\left(\bigcap_{i=1}^{5} S_i\right) = 1 - \prod_{i=1}^{5} P(S_i) = 1 - \prod_{i=1}^{5} p_i.$$

1.4 Probabilitatea condiţionată

Definiţia 1.5. Fie $(\Omega, \mathcal{K}, P)$ câmp de probabilitate şi $A, B \in \mathcal{K}$ a. î. $P(B) \neq 0$. *Probabilitatea condiţionată* de B a lui A este dată de

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

Observaţie. În ipotezele definiţiei 1.5, A şi B sunt independente $\iff$ $P(A|B) = P(A)$.

Demonstraţie. A şi B sunt independente $\iff P(A \cap B) = P(A) P(B) \iff \frac{P(A \cap B)}{P(B)} = P(A) \iff P(A|B) = P(A)$.□

Propoziţie. Fie $(\Omega, \mathcal{K}, P)$ câmp de probabilitate şi $B \in \mathcal{K}$ a. î. $P(B) \neq 0$. Atunci funcţia $P_B : \mathcal{K} \to \mathbb{R}, P_B(A) = P(A|B)$ este probabilitate pe$(\Omega, \mathcal{K})$.

Demonstraţie. Verificăm cele 3 axiome ale probabilităţii:

1) $P_B(A) = \frac{P(A \cap B)}{P(B)} \geq 0, \forall A \in \mathcal{K}$, deoarece $P(A \cap B) \geq 0$ şi $P(B) > 0$.

2) $P_B(\Omega) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$.

3) $A_1, A_2, ... \in \mathcal{K}$ colecţie numărabilă de evenimente mutual exclusive $\implies$ $A_1 \cap B, A_2 \cap B, ... \in \mathcal{K}$ mutual exclusive $\implies P_B(A_1 \cup A_2 \cup ...) = \frac{P((A_1 \cup A_2 \cup ...) \cap B)}{P(B)} = \frac{P((A_1 \cap B) \cup (A_2 \cap B) \cup ...)}{P(B)} = \frac{P(A_1 \cap B) + P((A_2 \cap B)) + ...}{P(B)} = \frac{P(A_1 \cap B)}{P(B)} + \frac{P(A_2 \cap B)}{P(B)} + ... = P_B(A_1) + P_B(A_2) + ...$□

Exemplul 1.5. Reconsiderăm exemplul 1.4 presupunând $p_1 > 0$. Care este probabilitatea condiţionată ca primele două componente să fie bune dat fiind că:

a) prima componentă este bună;

b) cel puţin una dintre cele două este bună?

Evenimentul $S_1 \cap S_2$ înseamnă că ambele componente sunt bune, iar $S_1 \cup S_2$ că cel puţin una e bună. Datorită independenţei lui S_1 şi S_2, avem:

a) $P(S_1 \cap S_2 | S_1) = \frac{P(S_1 \cap S_2 \cap S_1)}{P(S_1)} = \frac{P(S_1 \cap S_2)}{P(S_1)} = \frac{P(S_1)P(S_2)}{P(S_1)} = P(S_2) = p_2$.

b) $P(S_1 \cap S_2 | S_1 \cup S_2) = \frac{P(S_1 \cap S_2 \cap (S_1 \cup S_2))}{P(S_1 \cup S_2)} = \frac{P(S_1 \cap S_2)}{P(S_1 \cup S_2)} = \frac{P(S_1)P(S_2)}{P(S_1)+P(S_2)-P(S_1 \cap S_2)} = \frac{p_1 p_2}{p_1+p_2-p_1 p_2}$.

Exemplul 1.6. Determinaţi probabilitatea de a trage, fără înlocuire, 2 aşi succesiv dintr-un pachet de cărţi de joc fără jokeri.

Fie A_1 evenimentul "prima carte trasă este un as" şi similar A_2. Se cere $P(A_1 \cap A_2)$.

$P(A_1) = \frac{4}{52}$ (sunt 4 aşi în cele 52 de cărţi din pachet).

$P(A_2|A_1) = \frac{3}{51}$ (dacă prima carte trasă este un as, au rămas 51 de cărţi dintre care 3 sunt aşi).

$P(A_2|A_1) = \frac{P(A_1 \cap A_2)}{P(A_1)} \implies P(A_1 \cap A_2) = P(A_1) \cdot P(A_2|A_1) = \frac{4}{52} \cdot \frac{3}{51} = \frac{1}{13} \cdot \frac{1}{17} = \frac{1}{221}$.

Propoziţie. Fie $(\Omega, \mathcal{K}, P)$ câmp de probabilitate şi $A_1, A_2, ..., A_n \in \mathcal{K}$ a. î. $P(A_1 \cap A_2 \cap ... \cap A_{n-1}) > 0$. Atunci

$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1) P(A_2|A_1) P(A_3|A_1 \cap A_2) ... P(A_n|A_1 \cap A_2 \cap ... \cap A_{n-1})$.

Demonstraţie. $A_1 \supseteq A_1 \cap A_2 \supseteq ... \supseteq A_1 \cap A_2 \cap ... \cap A_{n-1} \implies P(A_1) \geq P(A_1 \cap A_2) \geq ... \geq P(A_1 \cap A_2 \cap ... \cap A_{n-1}) > 0 \implies$ probabilităţile condiţionate din membrul drept au sens.

$P(A_1) P(A_2|A_1) P(A_3|A_1 \cap A_2) ... P(A_n|A_1 \cap A_2 \cap ... \cap A_{n-1}) = P(A_1) \cdot \frac{P(A_1 \cap A_2)}{P(A_1)} \cdot \frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)} \cdot ... \cdot \frac{P(A_1 \cap A_2 \cap ... \cap A_n)}{P(A_1 \cap A_2 \cap ... \cap A_{n-1})} = P(A_1 \cap A_2 \cap ... \cap A_n)$. □

Definiţia 1.6. Fie $B_i \subseteq \Omega, \forall i \in I$. $(B_i)_{i \in I}$ se numeşte *partiţie* a lui Ω dacă şi numai dacă $(B_i)_{i \in I}$ sunt disjuncte două câte două şi $\bigcup_{i \in I} B_i = \Omega$.

Teorema probabilităţii totale. Fie $(\Omega, \mathcal{K}, P)$ câmp de probabilitate şi $(B_i)_{i \in I} \subset \mathcal{K}$ partiţie cel mult numărabilă a lui Ω a. î. $P(B_i) > 0, \forall i \in I$. Atunci, $\forall A \in \mathcal{K}$,

$$P(A) = \sum_{i \in I} P(A|B_i) P(B_i).$$

Demonstraţie. $(B_i)_{i \in I}$ mutual exclusive, $A \cap B_i \subseteq B_i, \forall i \in I \implies (A \cap B_i)_{i \in I}$ mutual exclusive $\implies \sum_{i \in I} P(A|B_i) P(B_i) = \sum_{i \in I} \frac{P(A \cap B_i)}{P(B_i)} \cdot P(B_i) = \sum_{i \in I} P(A \cap B_i) = P\left(\bigcup_{i \in I} (A \cap B_i)\right) = P\left(A \cap \left(\bigcup_{i \in I} B_i\right)\right) = P(A \cap \Omega) = P(A)$. □

Exemplul 1.7. Să se determine probabilitatea ca un nivel critic al curgerii să fie atins în timpul furtunilor într-un sistem de canalizare pe baza măsurătorilor meteorologice şi hidrologice.

Fie $B_i, i = 1, 2, 3$ diferitele nivele (mic, mediu şi mare) de precipitaţii cauzate de o furtună şi $A_j, j = 1, 2$ nivelele critic, respectiv necritic al curgerii.

Probabilităţile $P(B_i)$ pot fi estimate din înregistrările meteorologice, iar $P(A_j|B_i)$ din analiza curgerii. Presupunem că:

$P(B_1) = 0,5; P(B_2) = 0,3; P(B_3) = 0,2;$

$P(A_1|B_1) = 0; P(A_1|B_2) = 0,2; P(A_1|B_3) = 0,6;$

$P(A_2|B_1) = 1; P(A_2|B_2) = 0,8; P(A_2|B_3) = 0,4.$

Deoarece B_1, B_2, B_3 constituie o partiţie, din teorema probabilităţii totale avem:

$P(A_1) = P(A_1|B_1)P(B_1)+P(A_1|B_2)P(B_2)+P(A_1|B_3)P(B_3) = 0\cdot 0,5+ 0,2\cdot 0,3+0,6\cdot 0,2 = 0,18.$

1.5 Formula lui Bayes

Thomas Bayes a fost un filozof englez.

Teorema lui Bayes. Fie $(\Omega, \mathcal{K}, P)$ câmp de probabilitate şi $A, B \in \mathcal{K}$ a. î. $P(A) \neq 0$ şi $P(B) \neq 0$. Atunci:

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

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

Formula lui Bayes. Fie $(\Omega, \mathcal{K}, P)$ câmp de probabilitate şi $(B_i)_{i\in I} \subset \mathcal{K}$ partiţie cel mult numărabilă a lui Ω a. î. $P(B_i) > 0, \forall i \in I$. Atunci, $\forall A \in \mathcal{K}$ a. î. $P(A) \neq 0, \forall i \in I$,

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j\in I}[P(A|B_j)P(B_j)]}.$$

Demonstraţie. $\frac{P(A|B_i)P(B_i)}{\sum_{j\in I}[P(A|B_j)P(B_j)]} \overset{\text{TPT}}{=} \frac{P(A|B_i)P(B_i)}{P(A)} \overset{\text{TB}}{=} P(B_i|A)$.□

Exemplul 1.8. În exemplul 1.7, să se determine $P(B_2|A_2)$, probabilitatea ca, dat fiind că s-a atins un nivel necritic al curgerii, el să fi fost datorat unei furtuni de nivel mediu. Din formula lui Bayes rezultă

$$P(B_2|A_2) = \frac{P(A_2|B_2)P(B_2)}{\sum_{j=1}^{3}[P(A_2|B_j)P(B_j)]} = \frac{0,8\cdot 0,3}{1\cdot 0,5+0,8\cdot 0,3+0,4\cdot 0,2} = \frac{0,24}{0,5+0,24+0,08} = \frac{0,24}{0,82} = \frac{12}{41} \cong 0,293.$$

Exemplul 1.9. Un canal de comunicare binar simplu transmite mesaje folosind doar 2 semnale, să spunem 0 şi 1. Presupunem că, pentru un canal binar dat, 40%din timp e transmis un 1; probabilitatea ca un 0 transmis să fie corect recepţionat este 0,9 şi probabilitatea ca un 1 transmis să fie corect recepţionat este 0,95. Determinaţi:

a) probabilitatea ca un 1 să fie primit;

b) dat fiind că un 1 este primit, probabilitatea ca un 1 să fi fost transmis.

Fie

A = "1 este transmis"

$\overline{A}$ = "0 este transmis"

B = "1 este primit"

$\overline{B}$ = "0 este primit".

Din ipoteze

$P(A) = 0,4; P\left(\overline{A}\right) = 0,6;$

$P(B|A) = 0,95; P\left(\overline{B}|A\right) = 0,05;$

$P\left(\overline{B}|\overline{A}\right) = 0,9; P\left(B|\overline{A}\right) = 0,1.$

a) Deoarece A şi $\overline{A}$ formează o partiţie, din teorema probabilităţii totale rezultă că

$P(B) = P(B|A)\,P(A) + P\left(B|\overline{A}\right)P\left(\overline{A}\right) = 0,95 \cdot 0,4 + 0,1 \cdot 0,6 = 0,38 + 0,06 = 0,44.$

b) Din teorema lui Bayes,

$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0,95 \cdot 0,4}{0,44} = \frac{0,38}{0,44} = \frac{19}{22} \cong 0,864.$

2 Variabile aleatoare. Variabile aleatoare discrete şi variabile aleatoare continue, cu densitate de repartiţie. Funcţie de repartiţie. Momentele unei variabile aleatoare

2.1 Variabile aleatoare

Considerăm un experiment aleator ale cărui rezultate sunt elemente ale spaţiului probelor Ω din câmpul de probabilitate $(\Omega, \mathcal{K}, P)$. Pentru a construi un model pentru o variabilă aleatoare, presupunem că e posibil să asociem un număr real $X(\omega)$ pentru fiecare rezultat ω, urmând un anumit set de reguli.

Definiţia 2.1. Funcţia X se numeşte *variabilă aleatoare* dacă şi numai dacă
a) $X : \Omega \to \mathbb{R}$, unde $(\Omega, \mathcal{K}, P)$ este câmp de probabilitate şi
b) $\forall x \in \mathbb{R}, \{\omega \in \Omega | X(\omega) \leq x\} \in \mathcal{K}$.

Condiţia b) din definiţie e aşa-numita "condiţie de măsurabilitate". Ea ne asigură că are sens să considerăm probabilitatea evenimentului $\{\omega \in \Omega | X(\omega) \leq x\}$, notat mai simplu $X \leq x$ pentru orice $x \in \mathbb{R}$, sau, mai general, probabilitatea oricărei combinaţii finite sau numărabile de astfel de evenimente.

În continuare, dacă nu e specificat altfel, variabilele aleatoare sunt considerate pe un câmp de probabilitate $(\Omega, \mathcal{K}, P)$.

2.2 Variabile aleatoare discrete şi variabile aleatoare continue, cu densitate de repartiţie

Definiţia 2.2. O variabilă aleatoare se numeşte *discretă* dacă şi numai dacă ia numai valori izolate. Mulţimea valorilor unei variabile aleatoare discrete este cel mult numărabilă.

Definiţia 2.3. O variabilă aleatoare se numeşte *continuă* dacă valorile ei umplu un interval.

Definiţia 2.4. Fie X, variabilă aleatoare continuă. O funcţie $f_X : \mathbb{R} \to \mathbb{R}$ a. î. $f_X(x) \geq 0, \forall x \in \mathbb{R}$ şi $P(X \leq x) = \int_{-\infty}^{x} f_X(u)\, du, \forall x \in \mathbb{R}$ se numeşte *funcţie densitate de repartiţie* sau *funcţie densitate de probabilitate* sau simplu *densitate* a lui X.

2.3 Funcţie de repartiţie

Definiţia 2.5. Fie X variabilă aleatoare. Funcţia $F_X : \mathbb{R} \to \mathbb{R}$,

$$F_X(x) = P(X \leq x),$$

se numeşte *funcţia de repartiţie de probabilitate* sau simplu *funcţia de repartiţie* a lui X.

Indicele X identifică variabila aleatoare. Acest indice e uneori omis când nu e pericol de confuzie.

Proprietăţi ale funcţiei de repartiţie. 1) Există şi are valori între 0 şi 1.

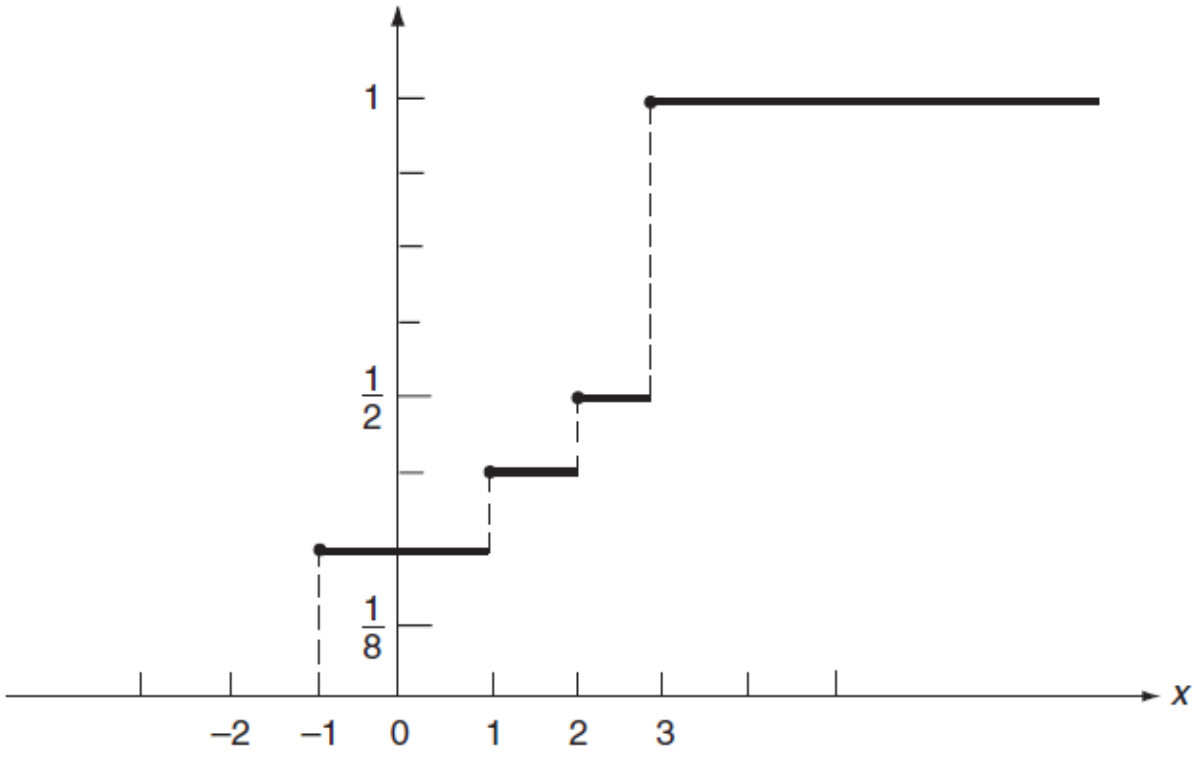
2) E continuă la dreapta şi crescătoare. Mai mult, avem:
$F_X(-\infty) := \lim_{x\to-\infty} F_X(x) = 0$ şi $F_X(\infty) := \lim_{x\to\infty} F_X(x) = 1$.

3) Dacă $a, b \in \mathbb{R}$ a. î. $a < b$, atunci
$P(a < X \leq b) := P(\{\omega \in \Omega | a < X(\omega) \leq b\}) = F_X(b) - F_X(a)$.
Această relaţie rezultă din
$P(X \leq b) = P(X \leq a) + P(a < X \leq b)$.

Exemplul 2.1. Fie X o variabilă aleatoare discretă cu valorile $-1, 1, 2, 3$ luate cu probabilităţile $\frac{1}{4}, \frac{1}{8}, \frac{1}{8}$, respectiv $\frac{1}{2}$. Pe scurt notăm $X \sim \begin{pmatrix} -1 & 1 & 2 & 3 \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & \frac{1}{2} \end{pmatrix}$.
Avem

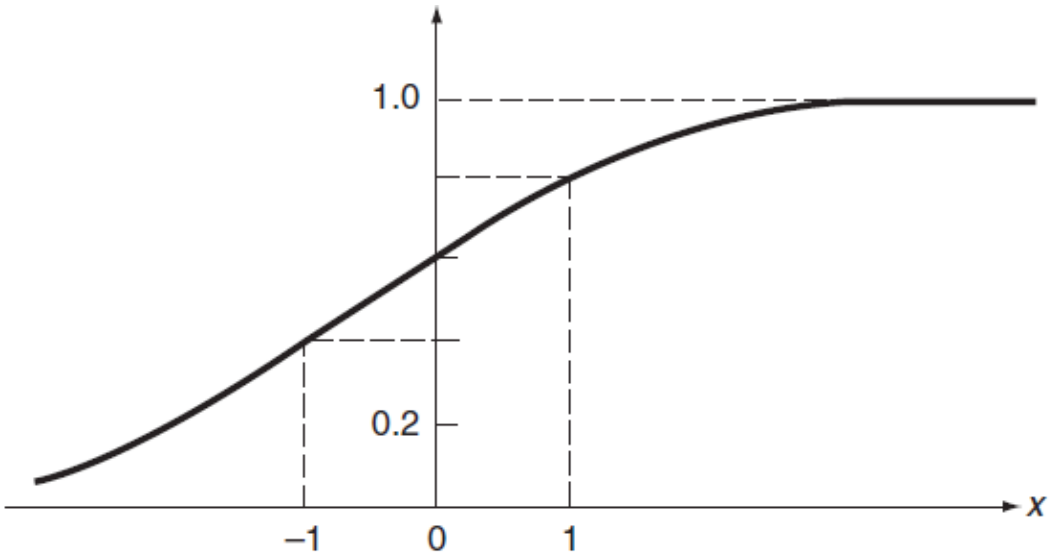
$$F_X(x) = \begin{cases} 0, \text{ pentru } x < -1; \\ \frac{1}{4}, \text{ pentru } -1 \leq x < 1; \\ \frac{3}{8}, \text{ pentru } 1 \leq x < 2; \\ \frac{1}{2}, \text{ pentru } 2 \leq x < 3; \\ 1, \text{ pentru } x \geq 3. \end{cases}$$

Graficul lui F_X este dat mai jos.



Este tipic pentru funcţia de repartiţie a unei variabile aleatoare discrete să crească de la 0 la 1 "în trepte".

Exemplul 2.2. O funcţie de repartiţie tipică pentru o variabilă aleatoare continuă este reprezentată grafic mai jos.



Ea nu are salturi sau discontinuităţi ca în cazul unei variabile aleatoare discrete. Probabilitatea ca X să aibă o valoare într-un anumit interval este dată de proprietatea 3) a funcţiei de repartiţie. Din grafic

$P(-1 < X \leq 1) = F_X(1) - F_X(-1) = 0,8 - 0,4 = 0,4.$

Avem $P(X = a) = 0, \forall a \in \mathbb{R}$.

Observaţie. Se poate defini funcţia de repartiţie şi ca $F_X : \mathbb{R} \to \mathbb{R}, F_X(x) = P(X < x)$. În acest caz proprietăţile 1) şi 2) rămân valabile cu excepţia faptului că funcţia de repartiţie este continuă la stânga şi nu la dreapta, iar proprietatea 3) devine

3') Dacă $a, b \in \mathbb{R}$ a. î. $a < b$, atunci

$P(a \leq X < b) = F_X(b) - F_X(a)$.

Proprietăţi ale densităţii. 1) $f_X(x) = F_X'(x), \forall x$ în care F_X este derivabilă.

2)

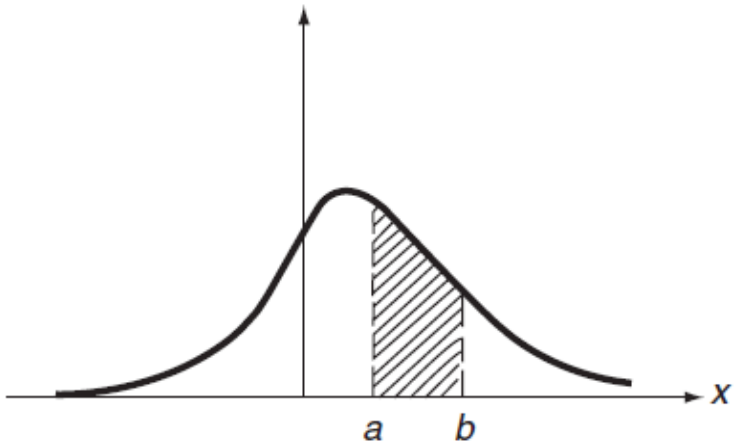
$$F_X(x) = \int_{-\infty}^{x} f_X(u)\, du, \forall x \in \mathbb{R}.$$

3) $\int_{-\infty}^{\infty} f_X(x)\, dx = 1.$

4) Dacă $a, b \in \mathbb{R}$ a. î. $a < b$, atunci

$P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)\, dx.$

Exemplul 2.3. Un exemplu de densitate e reprezentată grafic mai jos.



După cum indică proprietăţile 3) şi 4), aria totală de sub curbă este 1 şi suprafaţa haşurată de la a la b e egală cu $P(a < X \leq b)$.

Observaţie. Cunoaşterea densităţii sau a funcţiei de repartiţie caracterizează complet o variabilă aleatoare continuă.

Exemplul 2.4. Fie $a > 0$. O variabilă aleatoare X a cărei densitate este

$$f_X(x) = \begin{cases} ae^{-ax}, \text{ pentru } x > 0; \\ 0, \text{ altfel,} \end{cases}$$

se numeşte *repartizată exponenţial* (de parametru a). Avem $f_X(x) \geq 0, \forall x \in \mathbb{R}$ şi

$$\int_{-\infty}^{\infty} f_X(x)\,dx = \int_{-\infty}^{0} 0dx + \int_{0}^{\infty} ae^{-ax}dx = 0 - e^{-ax}|_0^{\infty} = 1,$$

deci f_X verifică proprietatea 3).

Dacă $x < 0$, atunci

$$\int_{-\infty}^{x} f_X(u)\,du = \int_{-\infty}^{x} 0du = 0.$$

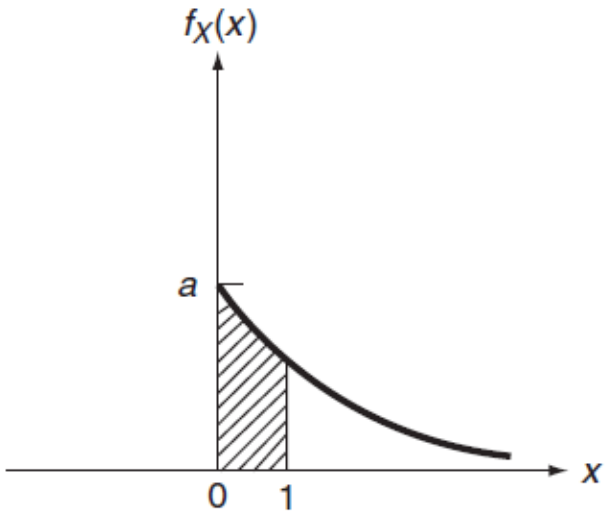
Dacă $x \geq 0$, atunci

$$\int_{-\infty}^{x} f_X(u)\,du = \int_{-\infty}^{0} 0du + \int_{0}^{x} ae^{-au}du = 0 - e^{-au}|_0^x = 1 - e^{-ax}.$$

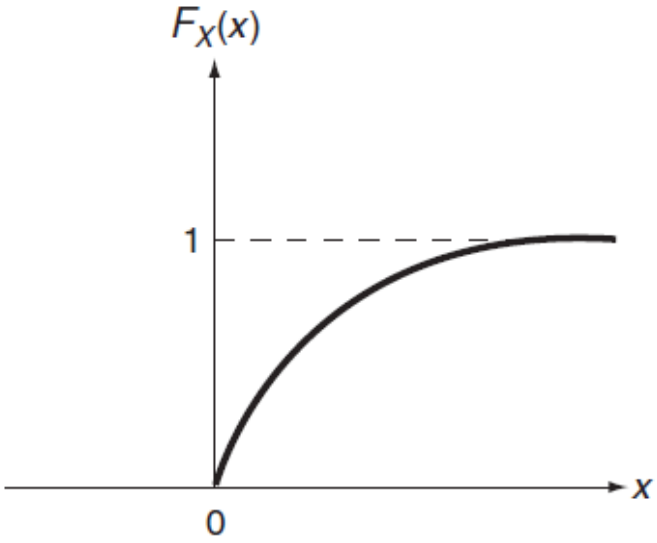
Deci, din proprietatea 2) avem

$$F_X(x) = \begin{cases} 0, \text{ pentru } x < 0; \\ 1 - e^{-ax}, \text{ pentru } x \geq 0. \end{cases}$$

Graficul densităţii e dat în figura (a), iar al funcţiei de repartiţie în figura (b) de mai jos.



(a)



(b)

Calculăm unele probabilităţi folosind f_X.

$P(0 < X \leq 1)$ e egală cu aria de sub graficul lui f_X de la $x = 0$ la $x = 1$, după cum se arată în figura (a). Avem

$$P(0 < X \leq 1) = \int_0^1 f_X(x)\,dx = -e^{-ax}|_0^1 = 1 - e^{-a}.$$

$P(X>3)$ e obţinută calculând aria de sub graficul lui f_X la dreapta lui $x=3$, deci

$P(X>3)=\int_3^{\infty} f_X(x)\,dx=-e^{-ax}|_3^{\infty}=e^{-3a}$.

Aceleaşi probabilităţi pot fi obţinute din F_X astfel:

$P(0<X\leq 1)=F_X(1)-F_X(0)=1-e^{-a}-0=1-e^{-a}$,

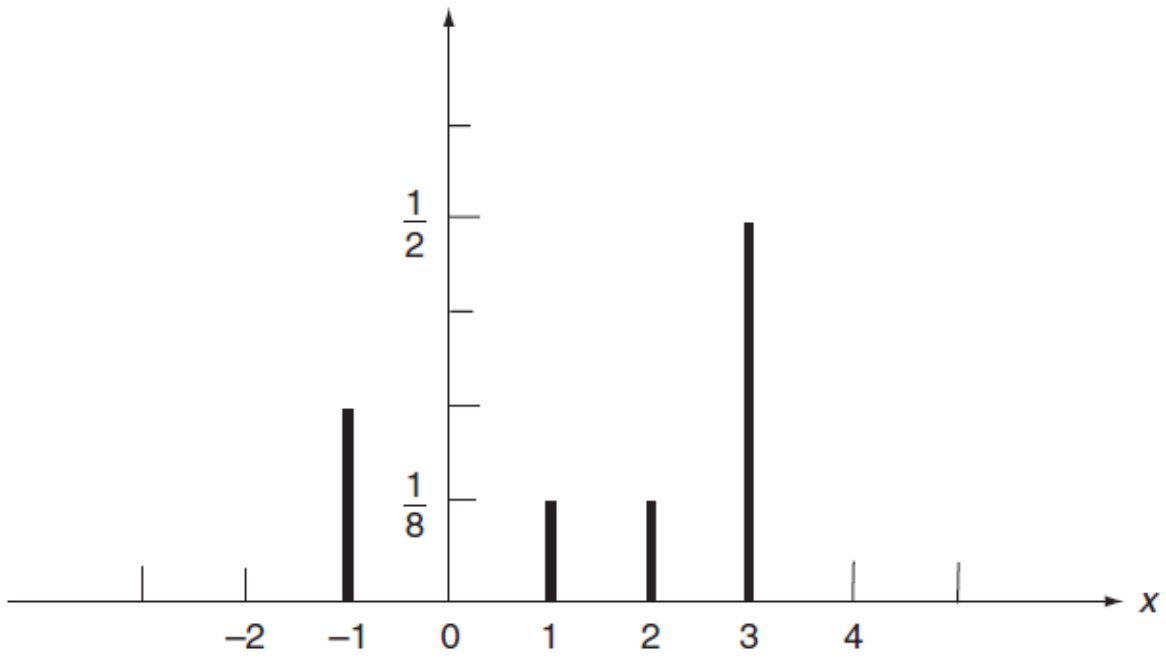
$P(X>3)=F_X(\infty)-F_X(3)=1-\left(1-e^{-3a}\right)=e^{-3a}$.

Mai observăm că $P(0<X\leq 1)=P(0\leq X\leq 1)$ pentru variabile aleatoare continue, deoarece $P(X=0)=0$.

Definiţia 2.6. Fie X variabilă aleatoare discretă. Funcţia $p_X:\mathbb{R}\to\mathbb{R}, p_X(x)=P(X=x):=P(\{\omega\in\Omega|X(\omega)=x\})$ se numeşte *funcţia masă de probabilitate* a lui X, sau, pe scurt, masa lui X.

Din nou indicele X e folosit pentru a identifica variabila aleatoare asociată.

Exemplul 2.5. Funcţia masă de probabilitate a variabilei aleatoare $X\sim\begin{pmatrix}-1 & 1 & 2 & 3\\ \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & \frac{1}{2}\end{pmatrix}$ din exemplul 2.1 e reprezentată mai jos.



Observaţii. 1) Dacă X e variabilă aleatoare discretă cu mulţimea cel mult numărabilă de valori $\{x_1,x_2,...\}$ luate cu probabilităţi nenule, atunci:

$0<p_X(x_i)\leq 1,\forall i$;

$\sum_i p_X(x_i)=1$;

$p_X(x)=0,\forall x\notin\{x_1,x_2,...\}$.

2) Ca şi F_X, specificarea lui p_X caracterizează complet variabila aleatoare discretă X. Mai mult, presupunând $x_1 < x_2 < ...$, relaţiile dintre F_X şi p_X sunt

$$\begin{aligned} p_X(x_1) &= F_X(x_1), \\ p_X(x_i) &= F_X(x_i) - F_X(x_{i-1}), \forall i > 1, \\ F_X(x) &= \sum_{i|x_i \leq x} p_X(x_i), \forall x \in \mathbb{R}. \end{aligned}$$

3) Specificarea lui p_X se face de obicei dând numai valorile pozitive, în restul punctelor subînţelegându-se că e 0.

2.4 Momentele unei variabile aleatoare

Fie X variabilă aleatoare discretă cu valorile $x_1, x_2, ...$ şi funcţia masă de probabilitate p_X sau continuă cu densitatea f_X.

Definiţia 2.7. Numărul real

$$E(X) := \begin{cases} \sum\limits_i x_i p_X(x_i), \text{ pentru } X \text{ discretă;} \\ \int_{-\infty}^{\infty} x f_X(x)\, dx, \text{ pentru } X \text{ continuă,} \end{cases}$$

dacă există, se numeşte *media* lui X şi se mai notează m_X sau simplu m.

Definiţia 2.8. Fie $n \in \mathbb{N}^*$. Numărul real

$$\alpha_n := E(X^n) = \begin{cases} \sum\limits_i x_i^n p_X(x_i), \text{ pentru } X \text{ discretă;} \\ \int_{-\infty}^{\infty} x^n f_X(x)\, dx, \text{ pentru } X \text{ continuă,} \end{cases}$$

dacă există, se numeşte *momentul de ordinul n* al lui X.

Observaţie. Media este momentul de ordinul 1.

Exemplul 2.6. Fie $X \sim \begin{pmatrix} -1 & 1 & 2 & 3 \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & \frac{1}{2} \end{pmatrix}$ din exemplul 2.1.

$E(X) = (-1) \cdot \frac{1}{4} + 1 \cdot \frac{1}{8} + 2 \cdot \frac{1}{8} + 3 \cdot \frac{1}{2} = -\frac{1}{4} + \frac{1}{8} + \frac{1}{4} + \frac{3}{2} = \frac{1}{8} + \frac{3}{2} = \frac{13}{8}$.

Exemplul 2.7. Timpul de aşteptare X (în minute) al unui client la un automat de bilete are densitatea

$f_X(x) = \begin{cases} 2e^{-2x}, \text{ pentru } x > 0; \\ 0, \text{ altfel.} \end{cases}$

Determinaţi timpul mediu de aşteptare.

Integrând prin părţi avem

$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_{-\infty}^{0} 0 dx + \int_0^{\infty} x \cdot 2e^{-2x} dx = 0 - \int_0^{\infty} x \left(e^{-2x}\right)' dx =$

$-xe^{-2x}|_0^{\infty} + \int_0^{\infty} e^{-2x} dx = 0 - \frac{1}{2} e^{-2x}|_0^{\infty} = \frac{1}{2}$ minut.

Proprietăţi ale mediei. Dacă $c \in \mathbb{R}$ este o constantă şi X şi Y sunt variabile aleatoare pe acelaşi câmp de probabilitate $(\Omega, \mathcal{K}, P)$, atunci:

$E(c) = c$,
$E(cX) = cE(X)$,
$E(X+Y) = E(X) + E(Y)$,
$E(X) \leq E(Y)$, dacă $X \leq Y$ (i.e. $X(\omega) \leq Y(\omega), \forall \omega \in \Omega$).

Definiţia 2.9. Fie X variabilă aleatoare. Se numeşte *mediană* a lui X o valoare x_0 a lui X a. î. $P(X \leq x_0) = \frac{1}{2}$ sau, dacă o astfel de valoare nu există, valoarea x_0 a lui X a. î. $P(X < x_0) < \frac{1}{2}$ şi $P(X \leq x_0) > \frac{1}{2}$.

Media lui X poate să nu existe, dar există cel puţin o mediană.

În comparaţie cu media, mediana e uneori preferată ca măsură a tendinţei centrale când repartiţia e asimetrică, în particular când sunt un număr mic de valori extreme în repartiţie. De exemplu, vorbim de mediana veniturilor ca o bună măsură a tendinţei centrale a venitului personal pentru o populaţie. Aceasta e o măsură mai bună decât media, deoarece mediana nu e aşa sensibilă la un număr mic de venituri extrem de mari sau venituri extrem de mici ca media.

Exemplul 2.8. Fie T timpul dintre emisiile de particule la un atom radioactiv. Este stabilit că T e o variabilă aleatoare cu repartiţie exponenţială, adică

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t}, \text{ pentru } t > 0; \\ 0, \text{ altfel,} \end{cases}$$

unde λ e o constantă pozitivă. Variabila aleatoare T se numeşte timpul de viaţă al atomului şi o măsură medie a acestui timp de viaţă este timpul de înjumătăţire, definit ca mediana lui T. Astfel, timpul de înjumătăţire τ e găsit din

$$P(T \leq \tau) = \frac{1}{2} \iff \int_{-\infty}^{\tau} f_T(t)\, dt = \frac{1}{2} \iff 1 - e^{-\lambda\tau} = \frac{1}{2} \iff e^{-\lambda\tau} = \frac{1}{2} \iff \tau = \frac{\ln 2}{\lambda}.$$

Observăm că viaţa medie $E(T)$ este

$$E(T) = \int_{-\infty}^{\infty} t f_T(t)\, dt = \frac{1}{\lambda} \text{ (se calculează analog ca la exemplul 2.7).}$$

Definiţia 2.10. Fie X variabilă aleatoare. Se numeşte *modul* sau *modă* a lui X

a) o valoare x_i luată de X a. î. $p_X(x_i) > p_X(x_{i+1})$ şi $p_X(x_i) > p_X(x_{i-1})$, dacă X e discretă cu valorile $x_1 < x_2 < ...$;

b) un punct de maxim local al lui f_X, dacă X e continuă.

Un modul este astfel o valoare a lui X corespunzătoare unui vârf în funcţia masă de probabilitate sau în densitate.

Termenul *distribuţie unimodală* se referă la o funcţie de repartiţie a unei variabile aleatoare care are un modul unic.

Media, mediana şi modulul coincid atunci când o repartiţie unimodală este simetrică.

Definiţia 2.11. Fie $n \in \mathbb{N}^*$ şi X variabilă aleatoare de medie m. *Momentul centrat de ordinul* n al lui X este

$$\mu_n = E\left((X-m)^n\right) = \begin{cases} \sum_i (x_i - m)^n p_X(x_i), \text{ pentru } X \text{ discretă;} \\ \int_{-\infty}^{\infty} (x-m)^n f_X(x)\,dx, \text{ pentru } X \text{ continuă.} \end{cases}$$

Definiţia 2.12. Fie X variabilă aleatoare. *Varianţa* sau *dispersia* lui X este momentul centrat de ordinul 2 al lui X, μ_2. Se notează cu σ_X^2 sau simplu σ^2 sau $var(X)$.

Valori mari ale lui σ_X^2 implică o întindere mare a valorilor lui X în jurul mediei. Reciproc, valori mici ale lui σ_X^2 implică o concentrare a valorilor lui X în jurul mediei. În cazul extrem când $\sigma_X^2 = 0$, $X = m$ cu probabilitatea 1 (întreaga masă a distribuţiei e concentrată în medie).

Propoziţie. Relaţia dintre dispersia şi momentele lui X este $\sigma^2 = \alpha_2 - m^2$.

Demonstraţie. $\sigma^2 = E\left((X-m)^2\right) = E\left(X^2 - 2mX + m^2\right) = E\left(X^2\right) - 2mE(X) + m^2 = \alpha_2 - 2m^2 + m^2 = \alpha_2 - m^2$.

Alte proprietăţi ale dispersiei. $var(X) \geq 0$,
$var(X+c) = var(X), \forall c \in \mathbb{R}$,
$var(cX) = c^2 var(X), \forall c \in \mathbb{R}$.

Fie X variabilă aleatoare de medie m. Se numeşte *deviaţie standard* a lui X $\sigma_X = \sqrt{E\left((X-m)^2\right)}$.

Un avantaj al folosirii lui σ_X în locul lui σ_X^2 este că σ_X are aceeaşi unitate de măsură ca media. De aceea poate fi comparată cu media pe aceeaşi scală pentru a obţine o măsură a gradului de împrăştiere.

Un număr adimensional (fără unitate de măsură) care caracterizează împrăştierea relativ la medie şi care facilitează compararea variabilelor aleatoare de unităţi diferite este *coeficientul de variaţie* definit de

$$v_X = \frac{\sigma_X}{m_X}.$$

Exemplul 2.9. Fie $X \sim \begin{pmatrix} -1 & 1 & 2 & 3 \\ \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & \frac{1}{2} \end{pmatrix}$ din exemplul 2.1. Să determinăm σ_X^2.

În exemplul 2.6 am văzut că $m_X = \frac{13}{8}$. Avem
$E\left(X^2\right) = (-1)^2 \cdot \frac{1}{4} + 1^2 \cdot \frac{1}{8} + 2^2 \cdot \frac{1}{8} + 3^2 \cdot \frac{1}{2} = \frac{1}{4} + \frac{1}{8} + \frac{1}{2} + \frac{9}{2} = \frac{3}{8} + 5 = \frac{43}{8}$.
$\sigma_X^2 = E\left(X^2\right) - m_X^2 = \frac{43}{8} - \frac{169}{64} = \frac{344-169}{64} = \frac{175}{64}$.

Exemplul 2.10. Determinăm dispersia lui X cu $f_X(x) = \begin{cases} 2e^{-2x}, \text{ pentru } x \geq 0; \\ 0, \text{ altfel.} \end{cases}$.

În exemplul 2.7 am văzut că $m_X = \frac{1}{2}$. Avem, integrând prin părţi
$E\left(X^2\right) = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx = \int_{-\infty}^{0} 0dx + \int_0^{\infty} x^2 \cdot 2e^{-2x}dx = 0 - \int_0^{\infty} x^2 \left(e^{-2x}\right)' dx =$
$-x^2 e^{-2x}|_0^{\infty} + \int_0^{\infty} 2xe^{-2x}dx = 0 + \frac{1}{2} = \frac{1}{2}$, ultima integrală fiind calculată la

exemplul 2.7.

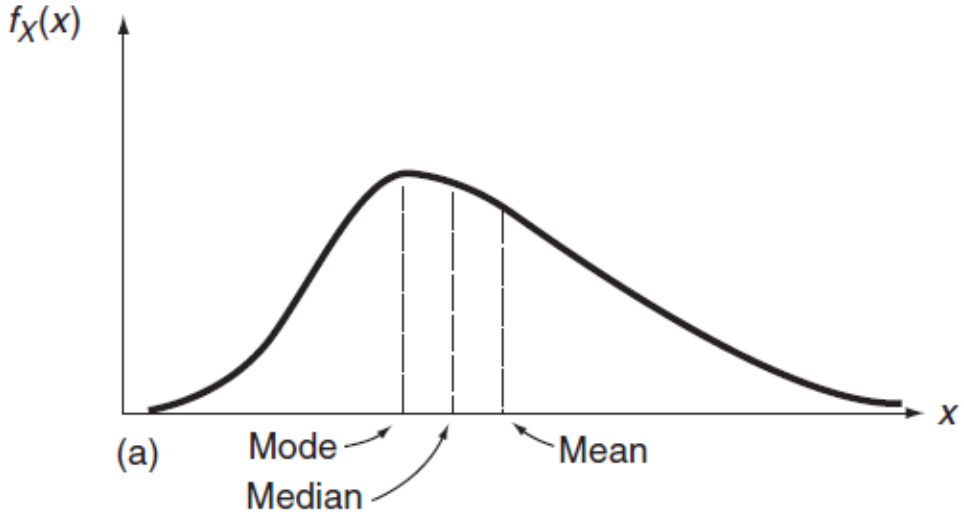
Deci

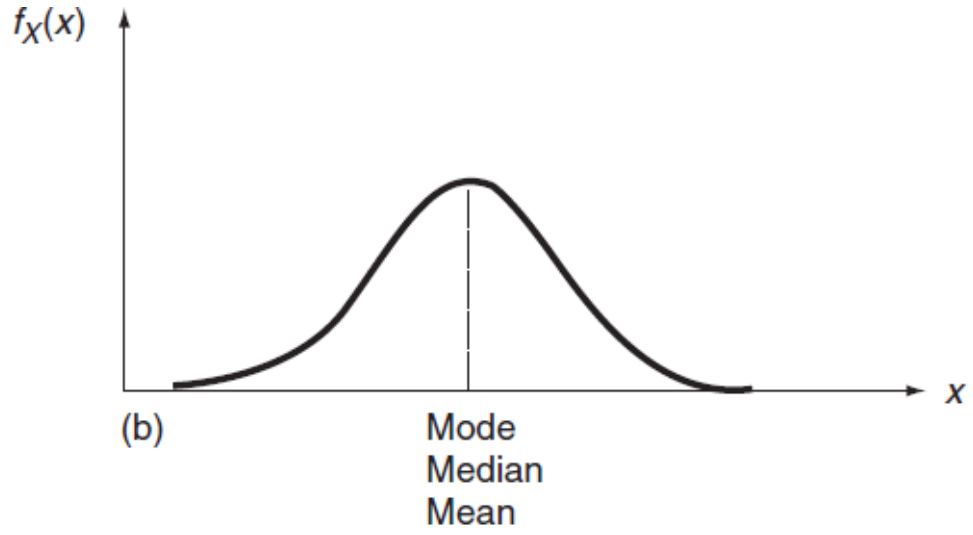
$\sigma_X^2 = E\left(X^2\right) - m_X^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$.

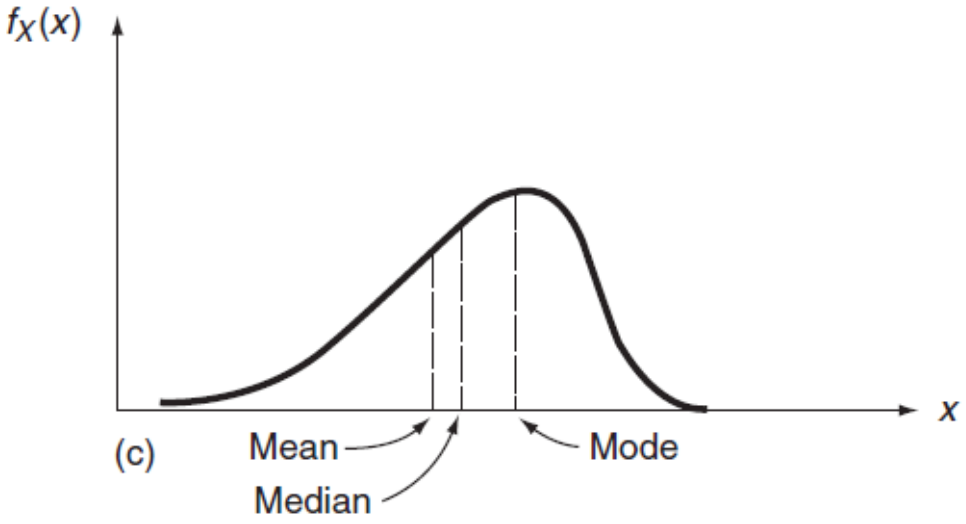
Coeficientul de asimetrie definit de

$\gamma_1 = \frac{\mu_3}{\sigma^3}$

dă o măsură a simetriei unei distribuţii. Este pozitiv când o distribuţie unimodală are o coadă dominantă la dreapta (adică modulul este la stânga mediei) şi negativ în caz contrar. Este 0 când o distribuţie e simetrică în jurul mediei. De fapt, o distribuţie simetrică în jurul mediei are toate momentele centrate de ordin impar 0. În figurile (a), (b) şi (c) sunt reprezentate densităţi cu $\gamma_1 > 0, \gamma_1 = 0$, respectiv $\gamma_1 < 0$.







Gradul de aplatizare a distribuţiei lângă vârfuri poate fi măsurat de *coeficientul de exces* definit de

$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$.

Un $\gamma_2 > 0$ implică un vârf ascuţit în vecinătatea modulului unei distribuţii unimodale, iar $\gamma_2 < 0$ implică, de regulă, un vârf turtit.

2.4.1 Inegalitatea lui Cebîşev

Teorema 2.1. (Inegalitatea lui Cebîşev) Fie X variabilă aleatoare cu media m_X şi deviaţia standard $\sigma_X \neq 0$. Atunci

$$P\left(|X - m_X| \geq k\sigma_X\right) \leq \frac{1}{k^2}, \tag{2.1}$$

pentru orice $k > 0$.

Demonstraţie. Presupunem X continuă. Din definiţie avem

$\sigma_X^2 = \int_{-\infty}^{\infty} (x - m_X)^2 f_X(x)\, dx \geq \int_{|x-m_X| \geq k\sigma_X} (x - m_X)^2 f_X(x)\, dx \geq$

$k^2\sigma_X^2 \int_{|x-m_X| \geq k\sigma_X} f_X(x)\, dx = k^2\sigma_X^2 P\left(|X - m_X| \geq k\sigma_X\right).$

Rezultă relaţia (2.1). Demonstraţia este similară când X este discretă.□

3 Independenţa variabilelor aleatoare. Densitate de repartiţie condiţionată şi formula lui Bayes pentru densităţi de repartiţie. Covarianţă şi corelaţie

3.1 Independenţa variabilelor aleatoare

Toate variabilele aleatoare sunt considerate pe acelaşi câmp de probabilitate $(\Omega, \mathcal{K}, P)$, dacă nu se specifică altfel.

Definiţia 3.1. a) *Funcţia de repartiţie comună* a variabilelor aleatoare X şi Y este definită de

$$F_{XY}(x,y) = P(X \leq x \cap Y \leq y).$$

b) *Funcţia de repartiţie comună* a variabilelor aleatoare $X_1, X_2, ..., X_n$ este definită de

$$F_{X_1X_2...X_n}(x_1, x_2, ..., x_n) = P(X_1 \leq x_1 \cap X_2 \leq x_2 \cap ... \cap X_n \leq x_n).$$

Proprietăţi 3.1. a) $F_{XY}(x,y) \geq 0, \forall x, y \in \mathbb{R}$.

b) F_{XY} e crescătoare în x şi y.

c) F_{XY} e continuă la dreapta în raport cu x şi y.

d) $F_{XY}(-\infty, -\infty) = F_{XY}(-\infty, y) = F_{XY}(x, -\infty) = 0, \forall x, y \in \mathbb{R}$.

e) $F_{XY}(\infty, \infty) = 1$.

f) $F_{XY}(x, \infty) = F_X(x), \forall x \in \mathbb{R}$.

g) $F_{XY}(\infty, y) = F_Y(y), \forall y \in \mathbb{R}$.

h) $\forall x_1, x_2, y_1, y_2 \in \mathbb{R}$ a. î. $x_1 < x_2$ şi $y_1 < y_2$,

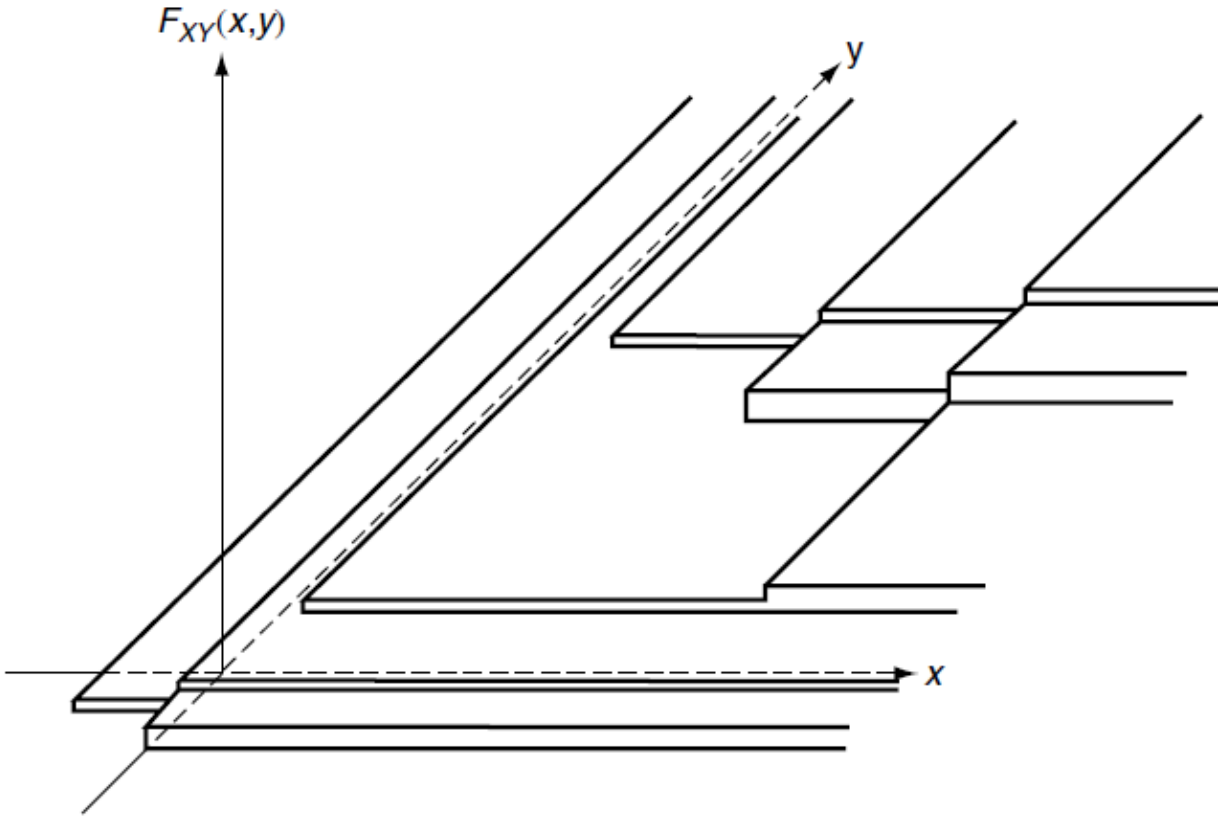
$P(x_1 < X \leq x_2 \cap y_1 < Y \leq y_2) = F_{XY}(x_2, y_2) - F_{XY}(x_1, y_2) - F_{XY}(x_2, y_1) + F_{XY}(x_1, y_1)$.

Demonstrăm de exemplu f):

$F_{XY}(x, \infty) = P(X \leq x \cap Y \leq \infty) = P(X \leq x \cap \Omega) = P(X \leq x) = F_X(x), \forall x \in \mathbb{R}$.

Proprietăţi similare se pot deduce pentru $F_{X_1X_2...X_n}$.

Forma generală a lui F_{XY} poate fi vizualizată din proprietăţile d)-g). În cazul când X şi Y sunt discrete F_{XY} seamănă cu un colţ al unor trepte neregulate, ca în figura de mai jos.



Creşte de la 0 la înălţimea de 1 în direcţia dinspre cadranul 3 spre cadranul 1. Când X şi Y sunt continue F_{XY} este o suprafaţă netedă cu aceleaşi trăsături.

Proprietăţile f) şi g) arată că funcţiile de repartiţie ale variabilelor aleatoare individuale, numite *funcţii de repartiţie marginale*, pot fi calculate din funcţia de repartiţie comună a lor. Reciproca nu este în general adevărată. O situaţie importantă când reciproca este adevărată este când X şi Y sunt independente.

Definiţia 3.2. a) Variabilele aleatoare X şi Y sunt *independente* dacă şi numai dacă $P(X \leq x \cap Y \leq y) = P(X \leq x) P(Y \leq y), \forall x, y \in \mathbb{R}$.

b) Variabilele aleatoare $X_1, X_2, ..., X_n$ sunt *independente* dacă şi numai dacă $P(X_1 \leq x_1 \cap X_2 \leq x_2 \cap ... \cap X_n \leq x_n) = P(X_1 \leq x_1) P(X_2 \leq x_2) ... P(X_n \leq x_n), \forall x_1, x_2, ..., x_n \in \mathbb{R}$.

c) Fie I mulţime infinită. Variabilele aleatoare $(X_i)_{i \in I}$ sunt independente $\iff$ $(X_i)_{i \in J}$ sunt independente, $\forall J \subset I$ finită.

Observaţii. a) X şi Y sunt independente $\iff$ $F_{XY}(x,y) = F_X(x) F_Y(y), \forall x, y \in \mathbb{R}$.

b) $X_1, X_2, ..., X_n$ sunt independente $\iff$ $F_{X_1 X_2 ... X_n}(x_1, x_2, ..., x_n) = F_{X_1}(x_1) F_{X_2}(x_2) ... F_{X_n}(x_n), \forall x_1, x_2, ..., x_n \in \mathbb{R}$.

c) X şi Y sunt independente $\Longrightarrow$

$P(x_1 < X \leq x_2 \cap y_1 < Y \leq y_2) = P(x_1 < X \leq x_2) P(y_1 < Y \leq y_2), \forall x_1, x_2, y_1, y_2 \in \mathbb{R}$ a. î. $x_1 < x_2$ şi $y_1 < y_2$.

Demonstraţie. a), b) Evident.

c) $P(x_1 < X \leq x_2 \cap y_1 < Y \leq y_2) =$
$F_{XY}(x_2, y_2) - F_{XY}(x_1, y_2) - F_{XY}(x_2, y_1) + F_{XY}(x_1, y_1) =$
$F_X(x_2) F_Y(y_2) - F_X(x_1) F_Y(y_2) - F_X(x_2) F_Y(y_1) + F_X(x_1) F_Y(y_1) =$
$(F_X(x_2) - F_X(x_1))(F_Y(y_2) - F_Y(y_1)) = P(x_1 < X \leq x_2) P(y_1 < Y \leq y_2)$.

În general:

$$X_1, X_2, ..., X_n \text{ sunt independente } \iff P\left(\bigcap_{i=1}^{n} X_i \in A_i\right) = \prod_{i=1}^{n} P(X_i \in A_i), \forall A_1, ..., A_n$$

intervale sau mulţimi cu un singur element din $\mathbb{R}$.

Aici $X_i \in A_i = \{\omega \in \Omega | X_i(\omega) \in A_i\}$.

Definiţia 3.3. a) *Funcţia masă de probabilitate comună* a variabilelor aleatoare discrete X şi Y este definită de

$$p_{XY}(x, y) = P(X = x \cap Y = y), \forall x, y \in \mathbb{R}.$$

b) Fie n variabile aleatoare discrete $X_1, X_2, ..., X_n$. *Funcţia masă de probabilitate comună* a lor este definită de

$$p_{X_1 X_2 ... X_n}(x_1, x_2, ...x_n) = P(X_1 = x_1 \cap X_2 = x_2 \cap ... \cap X_n = x_n), \forall x_1, x_2, ..., x_n \in \mathbb{R}.$$

Proprietăţi 3.2. Fie X şi Y variabile aleatoare discrete care iau o mulţime cel mult numărabilă de perechi de valori $(x_i, y_j), i, j = 1, 2, ...$ cu probabilităţi nenule.

a) $p_{XY}(x, y) = 0$ peste tot, exceptând punctele $(x_i, y_j), i, j = 1, 2, ...$ unde ia valori egale cu probabilitatea comună $P(X = x_i \cap Y = y_j)$.

b) $0 < p_{XY}(x_i, y_j) \leq 1$;

c) $\sum_i \sum_j p_{XY}(x_i, y_j) = 1$;

d) $\sum_i p_{XY}(x_i, y) = p_Y(y)$;

e) $\sum_j p_{XY}(x, y_j) = p_X(x)$;

f) $F_{XY}(x, y) = \sum_{i|x_i \leq x} \sum_{j|y_j \leq y} p_{XY}(x_i, y_j), \forall x, y \in \mathbb{R}$.

Acum $p_X(x)$ şi $p_Y(y)$ sunt numite *funcţii masă de probabilitate marginale.*

Proprietăţi similare pot fi scrise pentru $p_{X_1 X_2 ... X_n}$.

Definiţia 3.4. a) *Funcţia densitate de probabilitate comună* a două variabile aleatoare continue X şi Y este o funcţie $f_{XY} : \mathbb{R}^2 \to \mathbb{R}$ astfel încât

$$f_{XY}(x, y) \geq 0, \forall x, y \in \mathbb{R}$$

şi

$$P(X \leq x \cap Y \leq y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}(u, v) \, du dv, \forall x, y \in \mathbb{R}.$$

b) Fie *vectorul aleator* $\mathbf{X}$ cu componente variabilele aleatoare continue $X_1, X_2, ..., X_n$ care au funcţia de repartiţie comună

$F_{\mathbf{X}}(\mathbf{x}) = P(X_1 \leq x_1 \cap X_2 \leq x_2 \cap ... \cap X_n \leq x_n)$,

unde $\mathbf{x}$ este vectorul cu componentele $x_1, x_2, ..., x_n$.

Funcţia densitate comună corespunzătoare este o funcţie $f_{\mathbf{X}} : \mathbb{R}^n \to \mathbb{R}$ astfel încât

$$f_{\mathbf{X}}(\mathbf{x}) \geq 0, \forall \mathbf{x} \in \mathbb{R}^n$$

şi

$$P(X_1 \leq x_1 \cap ... \cap X_n \leq x_n) = \int_{-\infty}^{x_n} ... \int_{-\infty}^{x_1} f_{\mathbf{X}}(u_1, ..., u_n)\, du_1 ... du_n,$$

$\forall \mathbf{x} \in \mathbb{R}^n$.

Proprietăţi 3.3. a) $f_{XY}(x, y) = \frac{\partial^2 F_{XY}}{\partial x \partial y}(x, y), \forall x, y \in \mathbb{R}$ pentru care derivata parţială există.

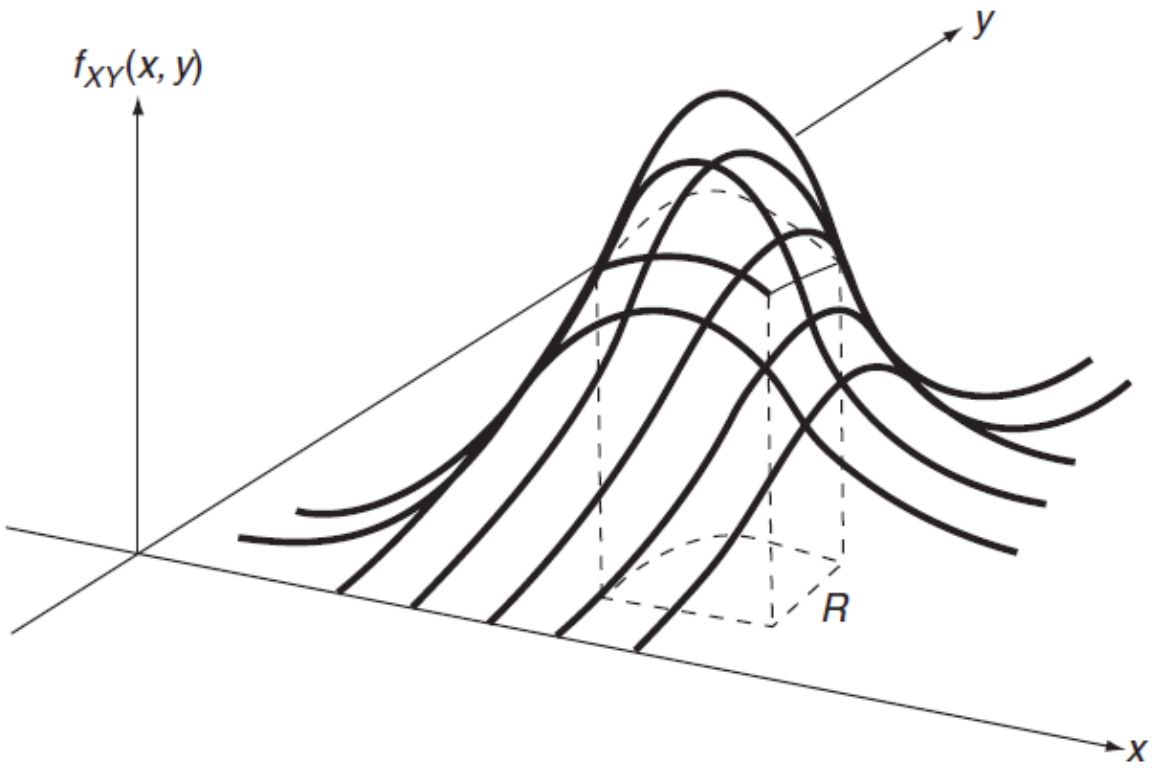
b) Din definiţie,

$$F_{XY}(x, y) = P(X \leq x \cap Y \leq y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}(u, v)\, du dv.$$

c) Dacă $x_1 < x_2$ şi $y_1 < y_2$, atunci

$$P(x_1 < X \leq x_2 \cap y_1 < Y \leq y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f_{XY}(x, y)\, dx dy.$$

d) f_{XY} defineşte o suprafaţă deasupra planului (x, y). După cum indică proprietatea 3.3 c), probabilitatea ca variabilele aleatoare X şi Y să se afle într-o anumită suprafaţă R este egală cu volumul de sub suprafaţa f_{XY} mărginit de acea regiune, ca în figura de mai jos.



e) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\,dxdy = 1.$

Această proprietate rezultă din proprietatea b) punând $x \to \infty$ şi $y \to \infty$ şi arată că volumul total de sub suprafaţa f_{XY} este 1.

f) $\int_{-\infty}^{\infty} f_{XY}(x,y)\,dy = f_X(x).$

Aceasta rezultă din

$F_X(x) = F_{XY}(x,\infty) = \int_{-\infty}^{\infty}\int_{-\infty}^{x} f_{XY}(u,y)\,dudy,$

derivând în raport cu x.

g) $\int_{-\infty}^{\infty} f_{XY}(x,y)\,dx = f_Y(y).$

h) $f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^n F_{\mathbf{X}}}{\partial x_1 \partial x_2 \ldots \partial x_n}(\mathbf{x}), \forall \mathbf{x} \in \mathbb{R}^n$ pentru care derivata parţială există.

Densităţile f_X şi f_Y din proprietăţile f) şi g) se numesc *densităţi marginale* ale lui X, respectiv Y.

3.2 Densitate de repartiţie condiţionată şi formula lui Bayes pentru densităţi de repartiţie

Fie X şi Y variabile aleatoare continue şi $y \in \mathbb{R}$ cu
$f_Y(y) \neq 0$. *Funcţia densitate de repartiţie condiţionată* (pe scurt *densitate condiţionată*) a lui X dat fiind

$$Y = y,$$

notată cu $f_{XY}(x|y)$, este definită de

$$f_{XY}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}. \tag{3.1}$$

În general

$$P(X \in A|Y = y) = \int_A f_{XY}(x|y)\,dx.$$

Când X şi Y sunt independente avem

$$f_{XY}(x|y) = f_X(x)$$

şi

$$f_{XY}(x,y) = f_X(x) f_Y(y),$$

adică densitatea comună e egală cu produsul densităţilor marginale când X şi Y sunt independente.

Exemplul 3.2.1. Fie variabilele aleatoare continue avand densitatea comună

$$f_{XY}(x,y) = \begin{cases} k(x+y), \text{ dacă } (x,y) \in (0,3)^2, \\ 0, \text{ altfel,} \end{cases}$$

unde k este o constantă ce trebuie determinată. Din proprietatea 3.1.3 e),
$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\,dxdy = k\int_0^3\int_0^3 (x+y)\,dxdy =$
$k\int_0^1 \left(\frac{x^2}{2} + xy\right)|_{x=0}^{x=3} dy = k\int_0^3 \left(\frac{9}{2} + 3y\right) dy = k\left(\frac{9}{2}y + 3\frac{y^2}{2}\right)|_0^3 =$
$k\left(\frac{27}{2} + \frac{27}{2}\right) = 27k,$
deci

$$k = \frac{1}{27}.$$

Densitatea marginală a lui X este
$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy =$

$$\begin{cases} \frac{1}{27}\int_0^3 (x+y)\,dy, \text{ dacă } x \in (0,3) \\ 0, \text{ altfel} \end{cases} =$$

$$\begin{cases} \frac{1}{27}\left(xy + \frac{y^2}{2}\right)|_{y=0}^{y=3}, \text{ dacă } x \in (0,3) \\ 0, \text{ altfel} \end{cases} = \begin{cases} \frac{2x+3}{18}, \text{ dacă } x \in (0,3) \\ 0, \text{ altfel} \end{cases}.$$

Datorită simetriei în x şi y a densităţii comune, densitatea marginală a lui Y este

$$f_Y(y) = \begin{cases} \frac{2y+3}{18}, \text{ dacă } y \in (0,3) \\ 0, \text{ altfel.} \end{cases}$$

Deoarece

$$f_{XY}(1,1) = \frac{2}{27} \neq \frac{25}{324} = f_X(1) f_Y(1),$$

X şi Y nu sunt independente.

Densitatea condiţionată a lui X dat fiind

$$Y = y$$

este

$$f_{XY}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{2}{3} \cdot \frac{x+y}{2y+3}, \text{ dacă } (x,y) \in (0,3)^2.$$

Extinderile pentru mai multe variabile aleatoare sunt imediate.

Plecând de la

$$P(A \cap B \cap C) = P(A|B \cap C) P(B|C) P(C)$$

pentru trei evenimente A, B şi C, avem în cazul a trei variabile aleatoare continue X, Y şi Z,

$$f_{XYZ}(x,y,z) = f_{XYZ}(x|y,z) f_{YZ}(y|z) f_Z(z).$$

Pentru cazul a n variabile aleatoare continue $X_1, X_2, ..., X_n$, componente ale vectorului aleator $\mathbf{X}$, putem scrie

$f_{\mathbf{X}}(\mathbf{x}) =$

$f_{X_1 X_2 ... X_n}(x_1|x_2, ..., x_n) f_{X_2 ... X_n}(x_2|x_3, ..., x_n) ... f_{X_{n-1} X_n}(x_{n-1}|x_n)$

$f_{X_n}(x_n)$.

Dacă $X_1, X_2, ..., X_n$ sunt independente, obţinem

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1}(x_1) f_{X_2}(x_2) ... f_{X_n}(x_n).$$

Formula lui Bayes pentru densităţi de repartiţie. Dacă X şi Y sunt variabile aleatoare continue, atunci

$$f_{XY}(x|y) = \frac{f_{YX}(y|x) f_X(x)}{f_Y(y)} = \frac{f_{YX}(y|x) f_X(x)}{\int_{-\infty}^{\infty} f_{YX}(y|\xi) f_X(\xi) d\xi},$$

dacă

$$f_Y(y) \neq 0.$$

3.3 Covarianţă şi corelaţie

Fie X şi Y variabile aleatoare discrete care iau o mulţime cel mult numărabilă de perechi de valori $(x_i, y_j), i, j = 1, 2, ...$ cu probabilităţi nenule sau variabile aleatoare continue.

Definiţia 3.5. Fie $n, m \in \mathbb{N}$.

a) *Momentele comune* α_{nm} ale variabilelor aleatoare X şi Y sunt date de, dacă există,

$$\alpha_{nm} = E(X^n Y^m) = \begin{cases} \sum\limits_i \sum\limits_j x_i^m y_j^n p_{XY}(x_i, y_j), \text{ dacă } X \text{ şi } Y \text{ sunt discrete,} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^n y^m f_{XY}(x, y)\, dxdy, \text{ dacă } X \text{ şi } Y \text{ sunt continue.} \end{cases}$$

b) Similar, *momentele centrate comune* ale lui X şi Y, când există, sunt date de

$\mu_{nm} = E((X - m_X)^n (Y - m_Y)^m)$.

Observaţii. Cu notaţiile folosite aici, mediile lui X şi Y sunt α_{10}, respectiv, α_{01}. De exemplu, folosind definiţia 3.5 a) pentru X şi Y continue, obţinem

$$\alpha_{10} = E(X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{XY}(x, y)\, dxdy = \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{XY}(x, y)\, dydx = \int_{-\infty}^{\infty} x f_X(x)\, dx,$$

unde f_X este densitatea marginală a lui X. Astfel vedem că acest rezultat e identic cu cel din cazul unei singure variabile aleatoare.

Această observaţie este adevărată şi pentru dispersiile individuale. Ele sunt μ_{20}, respectiv μ_{02}, şi pot fi găsite din definiţia 3.5 b) cu înlocuiri corespunzătoare pentru n şi m. Ca şi în cazul unei singure variabile aleatoare avem

$\mu_{20} = \alpha_{20} - \alpha_{10}^2$ sau $\sigma_X^2 = \alpha_{20} - m_X^2$,

respectiv

$\mu_{02} = \alpha_{02} - \alpha_{01}^2$ sau $\sigma_Y^2 = \alpha_{02} - m_Y^2$.

Definiţia 3.6. Se numeşte *covarianţă* a lui X şi Y

$\mu_{11} = cov(X, Y) = E((X - m_X)(Y - m_Y))$.

Covarianţa e o mărime a interdependenţei lui X şi Y.

Proprietatea 3.4. Covarianţa e legată de α_{nm} prin

$\mu_{11} = \alpha_{11} - \alpha_{10}\alpha_{01} = \alpha_{11} - m_X m_Y$.

Demonstraţie. $\mu_{11} = E((X - m_X)(Y - m_Y)) = E(XY - m_Y X - m_X Y + m_X m_Y) = E(XY) - m_Y E(X) - m_X E(Y) + m_X m_Y = \alpha_{11} - \alpha_{10}\alpha_{01} - \alpha_{10}\alpha_{01} + \alpha_{10}\alpha_{01} = \alpha_{11} - \alpha_{10}\alpha_{01}$.□

Definiţia 3.7. *Coeficientul de corelaţie* al lui X şi Y este

$$\rho = \rho(X, Y) = \frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}} = \frac{\mu_{11}}{\sigma_X \sigma_Y}.$$

Proprietatea 3.5. $|\rho| \leq 1$.

Demonstraţie. $[t(X - m_X) + Y - m_Y]^2 \geq 0, \forall t \in \mathbb{R} \implies E\left([t(X - m_X) + Y - m_Y]^2\right) = \mu_{20}t^2 + 2\mu_{11}t + \mu_{02} \geq 0, \forall t \in \mathbb{R} \implies \Delta = 4\mu_{11}^2 - 4\mu_{20}\mu_{02} \leq 0 \implies \mu_{11}^2 \leq \mu_{20}\mu_{02} \implies |\rho| \leq 1$.□

Coeficientul de corelaţie este fără dimensiune. El este şi independent de origine, adică $\forall a_1, a_2, b_1, b_2 \in \mathbb{R}$ cu $a_1, a_2 > 0$ se poate demonstra că

$\rho(a_1X + b_1, a_2Y + b_2) = \rho(X, Y)$.

Proprietatea 3.6. Dacă X şi Y sunt independente, atunci

$\mu_{11} = 0$ şi $\rho = 0$.

Demonstraţie. Fie X şi Y continue.

$\alpha_{11} = E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{XY}(x,y)\,dxdy \overset{\text{indep.}}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)\,f_Y(y)\,dxdy =$
$\int_{-\infty}^{\infty} xf_X(x)\,dx \int_{-\infty}^{\infty} yf_Y(y)\,dy = m_X m_Y \implies \mu_{11} = \alpha_{11} - m_X m_Y = 0 \implies$
$\rho = 0$.

Similar se poate demonstra dacă X şi Y sunt discrete.□

Dacă X şi Y sunt independente, atunci

(3.2) $E(g(X)\,h(Y)) = E(g(X))\,E(h(Y))$,

dacă mediile există.

Când coeficientul de corelaţie al două variabile aleatoare se anulează, spunem că ele sunt *necorelate.*

Observaţii. 1) X şi Y sunt necorelate $\iff E(XY) = E(X)\,E(Y)$. (Rezultă din definiţii şi proprietatea 3.4.)

2) X, Y independente $\implies$ X, Y necorelate. (Rezultă din definiţie şi proprietatea 3.6.)

3) Reciproca nu e adevărată.

Exemplul 3.1. Fie $X \sim \begin{pmatrix} -2 & -1 & 1 & 2 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$ şi $Y = X^2$.

Avem $Y \sim \begin{pmatrix} 1 & 4 \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$,

$$p_{XY}(x,y) = \begin{cases} \frac{1}{4}, \text{ pentru } (x,y) = (-2,4); \\ \frac{1}{4}, \text{ pentru } (x,y) = (-1,1); \\ \frac{1}{4}, \text{ pentru } (x,y) = (1,1); \\ \frac{1}{4}, \text{ pentru } (x,y) = (2,4), \end{cases}$$

$m_X = (-2)\cdot\frac{1}{4} + (-1)\cdot\frac{1}{4} + 1\cdot\frac{1}{4} + 2\cdot\frac{1}{4} = 0$,

$m_Y = 1\cdot\frac{1}{2} + 4\cdot\frac{1}{2} = \frac{5}{2}$,

$\alpha_{11} = (-2)\cdot 4\cdot\frac{1}{4} + (-1)\cdot 1\cdot\frac{1}{4} + 1\cdot 1\cdot\frac{1}{4} + 2\cdot 4\cdot\frac{1}{4} = 0$.

Deci

$\mu_{11} = \alpha_{11} - m_X m_Y = 0 \implies \rho = \frac{\mu_{11}}{\sigma_X\sigma_Y} = 0 \implies X$ şi Y sunt necorelate.

Pe de altă parte,

$P(X \le -2 \cap Y \le 1) = F_{XY}(-2,1) = 0$,

iar

$P(X \le -2)\,P(Y \le 1) = F_X(-2)\cdot F_Y(1) = \frac{1}{4}\cdot\frac{1}{2} = \frac{1}{8}$,

deci

$P(X \le -2 \cap Y \le 1) \ne P(X \le -2)\,P(Y \le 1) \implies X$ şi Y nu sunt independente.

Coeficientul de corelaţie măsoară interdependenţa liniară a variabilelor aleatoare, adică acurateţea cu care o variabilă aleatoare poate fi aproximată printr-o funcţie liniară de cealaltă. Pentru a vedea asta, considerăm problema

aproximării unei variabile aleatoare X printr-o funcţie liniară de o a doua variabilă aleatoare Y, $aY + b$, unde a şi b sunt alese a. î. eroarea medie pătratică e definită de

(3.3) $e = E\left([X - (aY + b)]^2\right)$

este minimă. Avem

$e = E\left(X^2 + a^2Y^2 + b^2 - 2aXY - 2bX + 2abY\right) = E\left(X^2\right) + a^2E\left(Y^2\right) + b^2 - 2aE\left(XY\right) - 2bm_X + 2abm_Y$,

$\frac{\partial e}{\partial a} = 2aE\left(Y^2\right) - 2E\left(XY\right) + 2bm_Y$,

$\frac{\partial e}{\partial b} = 2b - 2m_X + 2am_Y$.

Rezolvând sistemul

$\begin{cases} \frac{\partial e}{\partial a} = 0, \\ \frac{\partial e}{\partial b} = 0, \end{cases}$

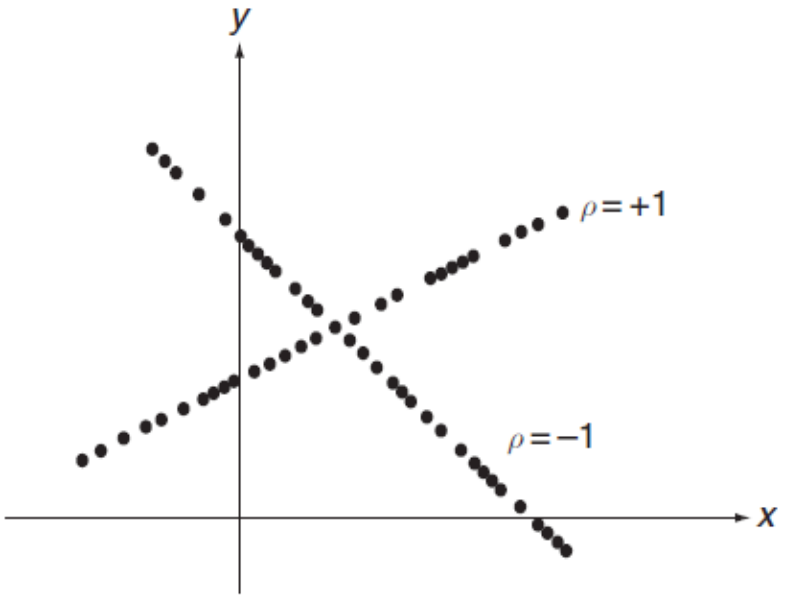
obţinem că minimul e atins când

$a = \frac{\sigma_X \rho}{\sigma_Y}$

şi

$b = m_X - am_Y$.

Înlocuind aceste valori în relaţia (3.3) obţinem eroarea medie pătratică minimă $\sigma_X^2\left(1 - \rho^2\right)$. Vedem că o potrivire exactă în sensul mediei pătratice e atinsă când $|\rho| = 1$ şi aproximarea liniară este cea mai rea când $\rho = 0$. Când $\rho = 1$, X şi Y se numesc *pozitiv perfect corelate*, în sensul că valorile pe care le iau sunt pe o dreaptă cu pantă pozitivă; ele sunt *negativ perfect corelate* când $\rho = -1$ şi valorile lor se află pe o dreaptă cu pantă negativă. Aceste două cazuri extreme sunt ilustrate în figura de mai jos.



Valoarea lui $|\rho|$ descreşte când împrăştierea valorilor în jurul dreptelor creşte.

În demonstraţia faptului că $|\rho| \leq 1$, am obţinut

$\mu_{11}^2 \leq \mu_{20}\mu_{02}$.

Folosind un procedeu similar, putem arăta de asemenea că

$$E^2(XY) \leq E\left(X^2\right) E\left(Y^2\right).$$

Ultimele două relaţii sunt *inegalităţile lui Schwarz*.

Definiţia 3.8. Fie $\mathbf{X}$ un vector coloană aleator cu componentele $X_1, X_2, ..., X_n$ şi $\mathbf{m_X}$ vectorul coloană având componente mediile lui $X_1, X_2, ..., X_n$. *Matricea de covarianţă* este

$$\mathbf{\Lambda} = E\left((\mathbf{X} - \mathbf{m_X})(\mathbf{X} - \mathbf{m_X})^T\right) = \begin{pmatrix} var(X_1) & cov(X_1, X_2) & ... & cov(X_1, X_n) \\ cov(X_2, X_1) & var(X_2) & ... & cov(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_n, X_1) & cov(X_n, X_2) & ... & var(X_n) \end{pmatrix}.$$

$\mathbf{\Lambda}$ este o matrice $n \times n$ cu având pe diagonală varianţe şi în afara diagonalei covarianţe. Deoarece $cov(X_i, X_j) = cov(X_j, X_i)$, matricea de covarianţă este simetrică.

Următorul rezultat este o generalizare a relaţiei (3.2).

Teoremă. Dacă $X_1, X_2, ..., X_n$ sunt independente, atunci

$E(g_1(X_1) g_2(X_2) ... g_n(X_n)) = E(g_1(X_1)) E(g_2(X_2)) ... E(g_n(X_n))$,

unde $g_j(X_j)$ este o funcţie arbitrară de X_j. Se presupune că toate mediile care sunt scrise există.

Propoziţie. Fie $X_1, X_2, ..., X_n$ variabile aleatoare şi $Y = \sum_{i=1}^{n} X_i$. Atunci

$$\sigma_Y^2 = \sum_{i=1}^{n} \sigma_{X_i}^2 + 2 \sum_{1 \leq i < j \leq n} cov(X_i, X_j).$$

În particular, dacă $X_1, X_2, ..., X_n$ sunt independente, atunci

$\sigma_Y^2 = \sum_{i=1}^{n} \sigma_{X_i}^2$.

Lema 3.1. Dacă $\mathbf{X}$ e un vector aleator şi $\mathbf{Y} = A\mathbf{X}$, atunci

$$cov(\mathbf{Y}) = A cov(\mathbf{X}) A^T.$$

4 Principalele repartiţii discrete şi proprietăţile lor

4.1 Repartiţia binomială

O succesiune de probe e făcută astfel încât

a) pentru fiecare probă sunt doar două rezultate posibile, să spunem succes şi eşec;

b) probabilităţile apariţiei acestor rezultate rămân aceleaşi pe durata probelor;

c) probele sunt independente.

Probele făcute în aceste condiţii se numesc *probe Bernoulli.*

Notăm evenimentul "succes" cu S şi evenimentul "eşec" cu F. Fie $P(S) = p$ şi $P(F) = q$, unde $p + q = 1$. Rezultate posibile din efectuarea unei succesiuni de probe Bernoulli pot fi reprezentate simbolic prin

$S \cap S \cap F \cap F \cap S \cap F \cap S \cap S \cap S \cap ... \cap F \cap F$

$F \cap S \cap F \cap S \cap S \cap F \cap F \cap F \cap S \cap ... \cap S \cap F$

$\vdots$

şi, datorită independenţei, probabilităţile acestor rezultate posibile sunt uşor de calculat. De exemplu,

$P(S \cap S \cap F \cap F \cap S \cap F \cap ...F \cap F) = P(S) P(S) P(F) P(F) P(S) P(F) ...P(F) P(F) =$ *ppqqpq...qq*.

Repartiţia unei variabile aleatoare X reprezentând numărul de succese dintr-o succesiune de n probe Bernoulli, indiferent de ordinea în care apar este frecvent de interes considerabil. E clar că X e o variabilă aleatoare discretă care ia valorile $0, 1, 2, ..., n$. Pentru a determina funcţia masă de probabilitate, considerăm $p_X(k)$, probabilitatea de a avea exact k succese în n probe. Acest eveniment poate apărea în la fel de multe moduri precum numărul de submulţimi de k elemente ale unei mulţimi de n elemente. De aici, numărul de moduri în care k succese se pot întâmpla în n probe este

$C_n^k = \frac{n!}{k!(n-k)!}$

şi probabilitatea asociată cu fiecare mod este $p^k q^{n-k}$. Din acest motiv avem

$$p_X(k) = C_n^k p^k q^{n-k}, k = 0, 1, 2, ..., n. \tag{4.1}$$

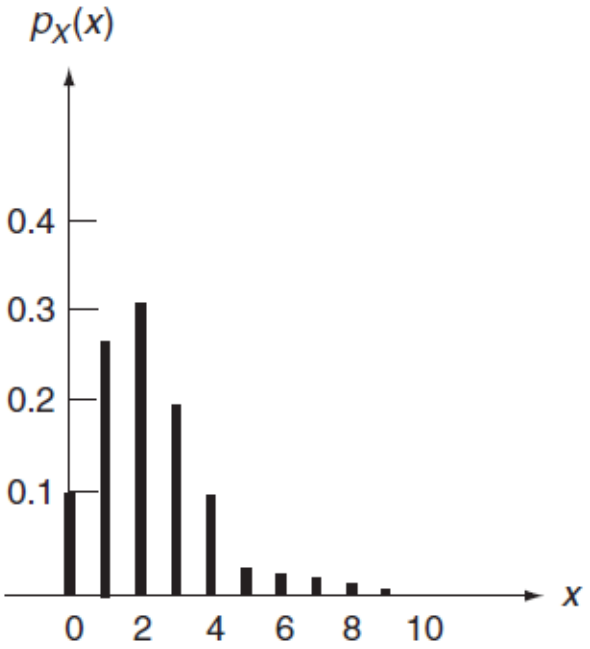
Datorită similarităţii cu termenii binomului lui Newton

$(a+b)^n = \sum_{k=0}^{n} C_n^k a^k b^{n-k}$,

repartiţia definită de relaţia (4.1) se numeşte *repartiţie binomială.* Ea are doi parametri, anume n şi p. O variabilă aleatoare X având repartiţie binomială se notează $X \sim B(n, p)$.

Forma unei repartiţii binomiale e determinată de valorile celor doi parametri ai ei, n şi p. În general, n se dă ca o parte a problemei şi p trebuie estimat din observaţii.

O reprezentare a funcţiei masă de probabilitate $p_X(k)$ pentru $n = 10$ şi $p = 0,2$ este în figura de mai jos.



Vârful repartiţiei se va muta spre dreapta când p creşte, atingând o repartiţie simetrică când $p = 0,5$. Considerăm raportul

$\frac{p_X(k)}{p_X(k-1)} = \frac{C_n^k p^k q^{n-k}}{C_n^{k-1} p^{k-1} q^{n-k+1}} = \frac{\frac{n!}{k!(n-k)!}p}{\frac{n!}{(k-1)!(n-k+1)!}q} = \frac{(n-k+1)p}{kq} = 1 + \frac{(n-k+1)p-kq}{kq} = 1 + \frac{(n+1)p-k(p+q)}{kq} = 1 + \frac{(n+1)p-k}{kq}$.

Observăm că $p_X(k) > p_X(k-1) \iff k < (n+1)p$ şi $p_X(k) < p_X(k-1) \iff k > (n+1)p$. Deci, dacă definim $k^* \in \mathbb{Z}$ prin

$(n+1)p - 1 < k^* \leq (n+1)p$,

valoarea lui $p_X(k)$ creşte de la $k = 0$ şi îşi atinge valoarea maximă când $k = k^*$, apoi descreşte. Dacă $(n+1)p \in \mathbb{Z}$, valoarea maximă se atinge în ambele $p_X(k^* - 1)$ şi $p_X(k^*)$. k^* este astfel un modul al acestei repartiţii şi se numeşte adesea "cel mai probabil număr de succese".

$p_X(k)$ este tabelată ca o funcţie de n şi p. Tabelul A.1 din figurile de mai jos dă valorile ei pentru $n = 2, 3, ..., 10$ şi $p = 0,01; 0,05; ...; 0,5$.

Table A.1 Binomial mass function: a table of

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

for $n = 2$ to 10, $p = 0.01$ to 0.50

n	k	p												
		0.01	0.05	0.10	0.15	0.20	0.25	0.30	$\frac{1}{3}$	0.35	0.40	0.45	0.49	0.50
2	0	0.9801	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4444	0.4225	0.3600	0.3025	0.2601	0.2500
	1	0.0198	0.0950	0.1800	0.2550	0.3200	0.3750	0.4200	0.4444	0.4550	0.4800	0.4950	0.4998	0.5000
	2	0.0001	0.0025	0.0100	0.0225	0.0400	0.0625	0.0900	0.1111	0.1225	0.1600	0.2025	0.2401	0.2500
3	0	0.9703	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2963	0.2746	0.2160	0.1664	0.1327	0.1250
	1	0.0294	0.1354	0.2430	0.3251	0.3840	0.4219	0.4410	0.4444	0.4436	0.4320	0.4084	0.3823	0.3750
	2	0.0003	0.0071	0.0270	0.0574	0.0960	0.1406	0.1890	0.2222	0.2389	0.2880	0.3341	0.3674	0.3750
	3	0.0000	0.0001	0.0010	0.0034	0.0080	0.0156	0.0270	0.0370	0.0429	0.0640	0.0911	0.1176	0.1250
4	0	0.9606	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.1975	0.1785	0.1296	0.0915	0.0677	0.0625
	1	0.0388	0.1715	0.2916	0.3685	0.4096	0.4219	0.4116	0.3951	0.3845	0.3456	0.2995	0.2600	0.2500
	2	0.0006	0.0135	0.0486	0.0975	0.1536	0.2109	0.2646	0.2963	0.3105	0.3456	0.3675	0.3747	0.3750
	3	0.0000	0.0005	0.0036	0.0115	0.0256	0.0469	0.0756	0.0988	0.1115	0.1536	0.2005	0.2400	0.2500
	4	0.0000	0.0000	0.0001	0.0005	0.0016	0.0039	0.0081	0.0123	0.0150	0.0256	0.0410	0.0576	0.0625
5	0	0.9510	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1317	0.1160	0.0778	0.0503	0.0345	0.0312
	1	0.0480	0.2036	0.3280	0.3915	0.4096	0.3955	0.3602	0.3292	0.3124	0.2592	0.2059	0.1657	0.1562
	2	0.0010	0.0214	0.0729	0.1382	0.2048	0.2637	0.3087	0.3292	0.3364	0.3456	0.3369	0.3185	0.3125
	3	0.0000	0.0011	0.0081	0.0244	0.0512	0.0879	0.1323	0.1646	0.1811	0.2304	0.2757	0.3060	0.3125
	4	0.0000	0.0000	0.0004	0.0022	0.0064	0.0146	0.0284	0.0412	0.0488	0.0768	0.1128	0.1470	0.1562
	5	0.0000	0.0000	0.0000	0.0001	0.0003	0.0010	0.0024	0.0041	0.0053	0.0102	0.0185	0.0283	0.0312
6	0	0.9415	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0878	0.0754	0.0467	0.0277	0.0176	0.0156
	1	0.0571	0.2321	0.3543	0.3993	0.3932	0.3560	0.3025	0.2634	0.2437	0.1866	0.1359	0.1014	0.0938
	2	0.0014	0.0305	0.0984	0.1762	0.2458	0.2966	0.3241	0.3292	0.3280	0.3110	0.2780	0.2437	0.2344
	3	0.0000	0.0021	0.0146	0.0415	0.0819	0.1318	0.1852	0.2195	0.2355	0.2765	0.3032	0.3121	0.3125

Table A.1 Continued

n	k	p												
		0.01	0.05	0.10	0.15	0.20	0.25	0.30	$\frac{1}{3}$	0.35	0.40	0.45	0.49	0.50
	4	0.0000	0.0001	0.0012	0.0055	0.0154	0.0330	0.0595	0.0823	0.0951	0.1382	0.1861	0.2249	0.2344
	5	0.0000	0.0000	0.0001	0.0004	0.0015	0.0044	0.0102	0.0165	0.0205	0.0369	0.0609	0.0864	0.0938
	6	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0014	0.0018	0.0041	0.0083	0.0139	0.0156
7	0	0.9321	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0585	0.0490	0.0280	0.0152	0.0090	0.0078
	1	0.0659	0.2573	0.3720	0.3960	0.3670	0.3115	0.2471	0.2048	0.1848	0.1306	0.0872	0.0603	0.0547
	2	0.0020	0.0406	0.1240	0.2097	0.2753	0.3115	0.3177	0.3073	0.2985	0.2613	0.2140	0.1740	0.1641
	3	0.0000	0.0036	0.0230	0.0617	0.1147	0.1730	0.2269	0.2561	0.2679	0.2903	0.2918	0.2786	0.2734
	4	0.0000	0.0002	0.0026	0.0109	0.0287	0.0577	0.0972	0.1280	0.1442	0.1935	0.2388	0.2676	0.2734
	5	0.0000	0.0000	0.0002	0.0012	0.0043	0.0115	0.0250	0.0384	0.0466	0.0774	0.1172	0.1543	0.1641
	6	0.0000	0.0000	0.0000	0.0001	0.0004	0.0013	0.0036	0.0064	0.0084	0.0172	0.0320	0.0494	0.0547
	7	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0005	0.0006	0.0016	0.0037	0.0068	0.0078
8	0	0.9227	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0390	0.0319	0.0168	0.0084	0.0046	0.0039
	1	0.0746	0.2793	0.3826	0.3847	0.3355	0.2670	0.1977	0.1561	0.1373	0.0896	0.0548	0.0352	0.0312
	2	0.0026	0.0515	0.1488	0.2376	0.2936	0.3115	0.2965	0.2731	0.2587	0.2090	0.1569	0.1183	0.1094
	3	0.0001	0.0054	0.0331	0.0839	0.1468	0.2076	0.2541	0.2731	0.2786	0.2787	0.2568	0.2273	0.2188
	4	0.0000	0.0004	0.0046	0.0185	0.0459	0.0865	0.1361	0.1707	0.1875	0.2322	0.2627	0.2730	0.2734
	5	0.0000	0.0000	0.0004	0.0026	0.0092	0.0231	0.0467	0.0683	0.0808	0.1239	0.1719	0.2098	0.2188
	6	0.0000	0.0000	0.0000	0.0002	0.0011	0.0038	0.0100	0.0171	0.0217	0.0413	0.0703	0.1008	0.1094
	7	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0012	0.0024	0.0033	0.0079	0.0164	0.0277	0.0312
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0002	0.0007	0.0017	0.0033	0.0039
9	0	0.9135	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0260	0.0207	0.0101	0.0046	0.0023	0.0020
	1	0.0830	0.2985	0.3874	0.3679	0.3020	0.2253	0.1556	0.1171	0.1004	0.0605	0.0339	0.0202	0.0176
	2	0.0034	0.0629	0.1722	0.2597	0.3020	0.3003	0.2668	0.2341	0.2162	0.1612	0.1110	0.0776	0.0703
	3	0.0001	0.0077	0.0446	0.1069	0.1762	0.2336	0.2668	0.2731	0.2716	0.2508	0.2119	0.1739	0.1641
	4	0.0000	0.0006	0.0074	0.0283	0.0661	0.1168	0.1715	0.2048	0.2194	0.2508	0.2600	0.2506	0.2461
	5	0.0000	0.0000	0.0008	0.0050	0.0165	0.0389	0.0735	0.1024	0.1181	0.1672	0.2128	0.2408	0.2461
	6	0.0000	0.0000	0.0001	0.0006	0.0028	0.0087	0.0210	0.0341	0.0424	0.0743	0.1160	0.1542	0.1641
	7	0.0000	0.0000	0.0000	0.0000	0.0003	0.0012	0.0039	0.0073	0.0098	0.0212	0.0407	0.0635	0.0703
	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0009	0.0013	0.0035	0.0083	0.0153	0.0176
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0003	0.0008	0.0016	0.0020
10	0	0.9044	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0173	0.0135	0.0060	0.0025	0.0012	0.0010
	1	0.0914	0.3151	0.3874	0.3474	0.2684	0.1877	0.1211	0.0867	0.0725	0.0403	0.0207	0.0114	0.0098
	2	0.0042	0.0746	0.1937	0.2759	0.3020	0.2816	0.2335	0.1951	0.1757	0.1209	0.0736	0.0495	0.0439
	3	0.0001	0.0105	0.0574	0.1298	0.2013	0.2503	0.2668	0.2601	0.2522	0.2150	0.1665	0.1267	0.1172
	4	0.0000	0.0010	0.0112	0.0401	0.0881	0.1460	0.2001	0.2276	0.2377	0.2508	0.2384	0.2130	0.2051
	5	0.0000	0.0001	0.0015	0.0085	0.0264	0.0584	0.1029	0.1366	0.1536	0.2007	0.2340	0.2456	0.2461
	6	0.0000	0.0000	0.0001	0.0012	0.0055	0.0162	0.0368	0.0569	0.0689	0.1115	0.1596	0.1966	0.2051
	7	0.0000	0.0000	0.0000	0.0001	0.0008	0.0031	0.0090	0.0163	0.0212	0.0425	0.0746	0.1080	0.1172
	8	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0014	0.0030	0.0043	0.0106	0.0229	0.0389	0.0439
	9	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0005	0.0016	0.0042	0.0083	0.0098
	10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0003	0.0008	0.0010

Calculul lui $p_X(k)$ devine greu când n devine mare; se foloseşte aproximarea Poisson a repartiţiei binomiale.

Funcţia de repartiţie $F_X(x)$ pentru o repartiţie binomială este dată de

$$F_X(x) = \sum_{k=0}^{[x]} C_n^k p^k q^{n-k},$$

unde $[x]$ este partea întreagă a lui x.

Proprietăţi 4.1. Fie $X \sim B(n,p)$. Atunci

a) $m_X = np$;

b) $\sigma_X^2 = npq$.

Demonstraţie. a) $m_X = \sum_{k=0}^{n} k p_X(k) = \sum_{k=1}^{n} k C_n^k p^k q^{n-k} = \sum_{k=1}^{n} k \cdot \frac{n!}{k!(n-k)!} p^k q^{n-k} =$

$\sum_{k=1}^{n} n \cdot \frac{(n-1)!}{(k-1)!(n-1-(k-1))!} p^k q^{n-k} = np \sum_{k=1}^{n} C_{n-1}^{k-1} p^{k-1} q^{n-1-(k-1)} \overset{k-1=i}{=} np \sum_{i=0}^{n-1} C_{n-1}^{i} p^i q^{n-1-i} =$

$np(p+q)^{n-1} = np.$

b) $E(X^2) = \sum_{k=0}^{n} k^2 p_X(k) = \sum_{k=1}^{n} k^2 C_n^k p^k q^{n-k} = np \sum_{k=1}^{n} k C_{n-1}^{k-1} p^{k-1} q^{n-1-(k-1)} =$

$np \left[\sum_{k=1}^{n} (k-1) C_{n-1}^{k-1} p^{k-1} q^{n-1-(k-1)} + \sum_{k=1}^{n} C_{n-1}^{k-1} p^{k-1} q^{n-1-(k-1)} \right] \overset{k-1=i,\ \text{a)}}{=}$

$np\left(\sum_{i=0}^{n-1} iC_{n-1}^{i}p^{i}q^{n-1-i}+1\right)=np\left(m_Y+1\right)\overset{\text{a)}}{=}np\left[\left(n-1\right)p+1\right],$

unde $Y\sim B\left(n-1,p\right).$

$\sigma_X^2=E\left(X^2\right)-m_X^2=np\left[\left(n-1\right)p+1\right]-n^2p^2=n^2p^2-np^2+np-n^2p^2=np\left(1-p\right)=npq.\square$

Faptul că $m_X=np$ sugerează că parametrul p poate fi estimat pe baza valorii medii a datelor observate.

O altă formulare care duce la repartiţia binomială este definirea variabilei aleatoare $X_j, j=1,2,...,n$, reprezentând rezultatul celei de-a j-a probe Bernoulli. Dacă punem

$X_j=\begin{cases}0, \text{ dacă proba } j \text{ este un eşec,}\\ 1, \text{ dacă proba } j \text{ este un succes,}\end{cases}$

atunci

$X=X_1+X_2+...+X_n$

dă numărul de succese în n probe. Din definiţie, $X_1,X_2,...,X_n$ sunt variabile aleatoare independente.

Deoarece

$E\left(X_j\right)=0\cdot q+1\cdot p=p, j=1,2,...,n,$

rezultă că

$E\left(X\right)=E\left(X_1+X_2+...+X_n\right)=E\left(X_1\right)+E\left(X_2\right)+...+E\left(X_n\right)=\underbrace{p+p+...+p}_{n}=np,$

ceea ce coincide cu proprietatea 4.1 a).

Analog, deoarece

$E\left(X_j^2\right)=0^2\cdot q+1^2\cdot p=p, j=1,2,...,n,$

$\sigma_{X_j}^2=E\left(X_j^2\right)-\left[E\left(X_j\right)\right]^2=p-p^2=p\left(1-p\right)=pq, j=1,2,...,n,$

din independenţa lui $X_1,X_2,...,X_n$ rezultă că

$\sigma_X^2=\sum_{j=1}^{n}\sigma_{X_j}^2=\underbrace{pq+pq+...+pq}_{n}=npq,$

ceea ce coincide cu proprietatea 4.1 b).

Teorema 4.1. Fie $X_1\sim B\left(n_1,p\right)$ şi $X_2\sim B\left(n_2,p\right)$ variabile aleatoare independente şi $Y=X_1+X_2$. Atunci $Y\sim B\left(n_1+n_2,p\right).$

4.2 Repartiţia hipergeometrică

Fie Z variabila aleatoare care dă numărul de bile negre care sunt extrase când un eşantion de m bile este extras (fără revenire) dintr-un lot de n bile având n_1 bile negre şi n_2 bile albe ($n_1+n_2=n$). Funcţia masă de probabilitate a variabilei aleatoare Z este

$p_Z\left(k\right)=\frac{C_{n_1}^{k}\cdot C_{n-n_1}^{m-k}}{C_n^m}, k=0,1,...,\min\left(n_1,m\right).$

Spunem că Z are *repartiţie hipergeometrică.*

Se poate arăta că

$m_Z=\frac{mn_1}{n},$

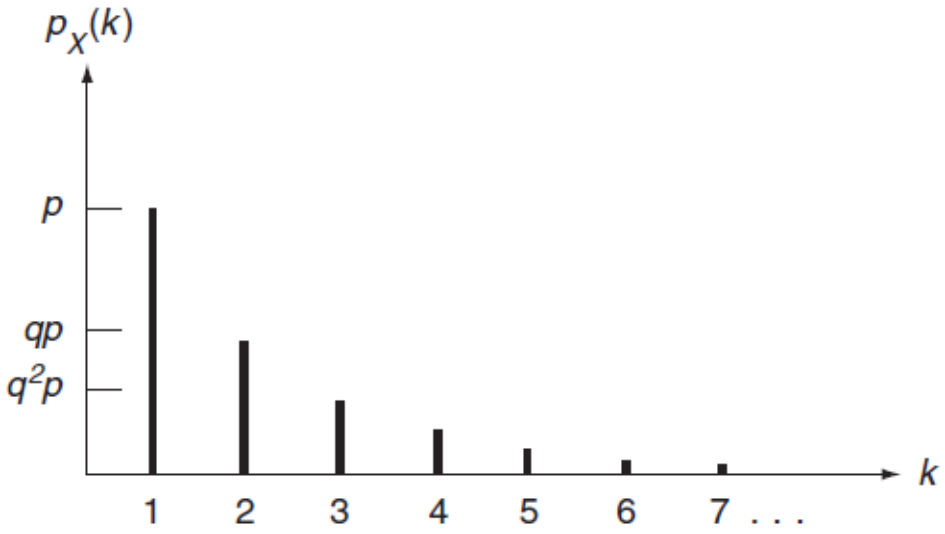
$\sigma_Z^2=\frac{mn_1(n-n_1)(n-m)}{n^2(n-1)}.$

4.3 Repartiţia geometrică

Fie X numărul de probe Bernoulli până la (şi incluzând) prima apariţie a succesului. X e variabilă aleatoare discretă având ca valori toate numerele naturale. Funcţia ei masă de probabilitate este

$$p_X(k) = P\left(\underbrace{F \cap F \cap ... \cap F}_{k-1} \cap S\right) = \underbrace{P(F)\,P(F)\,...P(F)}_{k-1} P(S) = q^{k-1}p, k = 1, 2, ...$$

Această repartiţie este cunoscută ca *repartiţia geometrică* cu parametrul p, unde numele provine de la similaritatea cu termenii progresiei geometrice. O reprezentare a lui $p_X(k)$ e dată mai jos.



Funcţia de repartiţie corespunzătoare este

$$F_X(x) = \sum_{k=1}^{[x]} p_X(k) = \sum_{k=1}^{[x]} q^{k-1}p = p\sum_{k=1}^{[x]} q^{k-1} = (1-q)\sum_{k=1}^{[x]} q^{k-1} = 1 - q^{[x]},$$

dacă $x \geq 1$

şi

$F_X(x) = 0$, dacă $x < 1$.

Media şi dispersia lui X se calculează astfel:

$$E(X) = \sum_{k=1}^{\infty} kq^{k-1}p = p\sum_{k=1}^{\infty} \frac{d}{dq}\left(q^k\right) = p\frac{d}{dq}\sum_{k=1}^{\infty} q^k = p\frac{d}{dq}\left(q\lim_{n\to\infty}\sum_{k=1}^{n} q^{k-1}\right) = p\frac{d}{dq}\left(q\lim_{n\to\infty}\frac{1-q^{n+1}}{1-q}\right) = p\frac{d}{dq}\left(\frac{q}{1-q}\right) = p\cdot\frac{1-q+q}{(1-q)^2} = \frac{1}{p}.$$

$$E\left(X^2\right) = \sum_{k=1}^{\infty} k^2q^{k-1}p = p\left[\sum_{k=1}^{\infty} kq^{k-1} + q\sum_{k=2}^{\infty} k(k-1)q^{k-2}\right] = \frac{1}{p} + pq\sum_{k=2}^{\infty}\frac{d^2}{dq^2}\left(q^k\right) =$$

$\frac{1}{p}+pq\left(\frac{d^2}{dq^2}\sum_{k=2}^{\infty}q^k\right)=\frac{1}{p}+pq\frac{d^2}{dq^2}\left(\frac{q}{1-q}-q\right)=\frac{1}{p}+pq\frac{d^2}{dq^2}\left(\frac{q^2}{1-q}\right)=\frac{1}{p}+pq\frac{d}{dq}\left[\frac{2q(1-q)+q^2}{(1-q)^2}\right]=$
$\frac{1}{p}+pq\frac{d}{dq}\left(\frac{2q-q^2}{(1-q)^2}\right)=\frac{1}{p}+pq\frac{(2-2q)(1-q)^2+\left(2q-q^2\right)2(1-q)}{(1-q)^4}=\frac{1}{p}+q\frac{2\left((1-q)^2+2q-q^2\right)}{(1-q)^2}=$
$\frac{1}{p}+\frac{2q}{(1-q)^2}$.

$\sigma_X^2=E\left(X^2\right)-\left[E\left(X\right)\right]^2=\frac{1}{p}+\frac{2q}{p^2}-\frac{1}{p^2}=\frac{p+2q-1}{p^2}=\frac{q}{p^2}=\frac{1-p}{p^2}$.

4.4 Repartiţia binomială negativă

O generalizare a repartiţiei geometrice este repartiţia variabilei aleatoare X reprezentând numărul de probe Bernoulli necesare pentru apariţia celui de-al r-lea succes, unde $r\in\mathbb{N}^*$ este dat.

Pentru a determina $p_X(k)$ în acest caz, fie A evenimentul ca primele $k-1$ probe să dea exact $r-1$ succese, indiferent de ordinea lor, şi B evenimentul ca un succes să apară la proba k. Atunci, datorită independenţei,

$p_X(k)=P(A\cap B)=P(A)P(B)$.

Avem

$P(B)=p$

şi

$P(A)=p_Z(r-1)=C_{k-1}^{r-1}p^{r-1}q^{k-r}$,

unde $Z\sim B(k-1,p)$.

Obţinem

$$p_X(k)=C_{k-1}^{r-1}p^rq^{k-r},k=r,r+1,.... \tag{4.2}$$

Repartiţia definită de relaţia (4.2) se numeşte repartiţie *binomială negativă* sau *Pascal* şi are parametrii r şi p. Este adesea notată prin $NB(r,p)$. Observăm că ea se reduce la repartiţia geometrică când $r=1$.

O variantă a acestei repartiţii se obţine punând $Y=X-r$. Variabila aleatoare Y este numărul de probe Bernoulli *dincolo* de r necesare pentru realizarea celui de-al r-lea succes, sau poate fi interpretată ca numărul de eşecuri înainte de cel de-al r-lea succes.

Funcţia masă de probabilitate a lui Y, $p_Y(m)$, e obţinută din relaţia (4.2) înlocuind k prin $m+r$:

$p_Y(m)=C_{m+r-1}^{r-1}p^rq^m=C_{m+r-1}^{m}p^rq^m,m=0,1,2,....$

Variabila aleatoare Y are p;roprietatea convenabilă că valorile ei încep de la 0 şi nu de la r ca valorile lui X.

Reamintind o definiţie mai generală a coeficientului binomial

$C_a^j=\frac{a(a-1)...(a-j+1)}{j!},\forall a\in\mathbb{R},j\in\mathbb{N}^*$,

avem

$C_{m+r-1}^m=\frac{(m+r-1)!}{m!(r-1)!}=\frac{(m+r-1)(m+r-2)...r}{m!}=(-1)^m\frac{(-r)(-r-1)...(-r-m+1)}{m!}=$
$(-1)^mC_{-r}^m$.

De aici,

$$p_Y(m)=C_{-r}^mp^r(-q)^m,m=0,1,2,...,$$

acesta fiind motivul pentru numele "repartiţie binomială negativă".

Media şi dispersia variabilei aleatoare X pot fi determinate observând că X poate fi reprezentată prin

$X = X_1 + X_2 + ... + X_r$,

unde X_j este numărul de probe dintre cel de-al $(j-1)$-lea şi (inclusiv) cel de-al j-lea succes. Aceste variabile aleatoare sunt independente, fiecare având repartiţie geometrică în care media este $\frac{1}{p}$ şi dispersia $\frac{1-p}{p^2}$. De aceea media sumei este suma mediilor şi, din independenţă, dispersia sumei este suma dispersiilor, adică:

$m_X = \frac{r}{p}, \sigma_X^2 = \frac{r(1-p)}{p^2}$.

Deoarece $Y = X - r$, avem:

$m_Y = \frac{r}{p} - r, \sigma_Y^2 = \frac{r(1-p)}{p^2}$.

4.5 Repartiţia multinomială

O generalizare a probelor Bernoulli este să relaxăm cererea să fie doar două rezultate posibile pentru fiecare probă. Fie r rezultate posibile pentru fiecare probă, notate cu $E_1, E_2, ..., E_r$ şi fie $P(E_i) = p_i, i = 1, 2, ..., r$ cu $p_1 + p_2 + ... + p_r = 1$.

Dacă variabila aleatoare $X_i, i = 1, 2, ..., r$, reprezintă numărul de E_i într-o succesiune de n probe, funcţia masă de probabilitate comună a lui $X_1, X_2, ..., X_r$ este dată de

$$p_{X_1 X_2 ... X_r}(k_1, k_2, ..., k_r) = \frac{n!}{k_1! k_2! ... k_r!} p_1^{k_1} p_2^{k_2} ... p_r^{k_r}, \tag{4.3}$$

unde $k_j = 0, 1, 2, ...; j = 1, 2, ..., r$, şi $k_1 + k_2 + ... + k_r = n$.

Demonstraţia formulei 4.3. Considerăm evenimentul de a avea exact k_1 rezultate E_1, k_2 rezultate $E_2, ..., k_r$ rezultate E_r în n probe. Acest eveniment poate apărea în la fel de multe moduri precum numărul de moduri de a plasa k_1 litere E_1, k_2 litere $E_2, ..., k_r$ litere E_r în n cutii a. î. fiecare cutie să aibă exact o literă. De aici, numărul de moduri căutat este produsul dintre numărul de moduri de a plasa k_1 litere E_1 în n cutii, numărul de moduri de a plasa k_2 litere în cele $n - k_1$ cutii rămase neocupate, ş.a.m.d., adică

$C_n^{k_1} C_{n-k_1}^{k_2} ... C_{n-k_1-k_2-...-k_{r-1}}^{k_r} = \frac{n!}{k_1!(n-k_1)!} \cdot \frac{(n-k_1)!}{k_2!(n-k_1-k_2)!} \cdot ... \cdot \frac{(n-k_1-k_2-...-k_{r-1})!}{k_r!0!} = \frac{n!}{k_1! k_2! ... k_r!}$

şi probabilitatea asociată cu fiecare mod, datorită independenţei, este $p_1^{k_1} p_2^{k_2} ... p_r^{k_r}$.□

Relaţia (4.3) defineşte funcţia masă de probabilitate comună a *repartiţiei multinomiale*, numită aşa deoarece are forma temenului general din expansiunea multinomială a lui $(p_1 + p_2 + ... + p_r)^n$. Repartiţia multinomială se reduce la repartiţia binomială când $r = 2$, şi cu $p_1 = p, p_2 = q, k_1 = k$ şi $k_2 = n - k$.

Deoarece $X_i \sim B(n, p_i)$, avem

$m_{X_i} = np_i, \sigma_{X_i}^2 = np_1(1 - p_i)$,

şi se poate arăta că avem

$cov(X_i, X_j) = -np_i p_j, \qquad i, j = 1, 2, ..., r, \qquad i \neq j.$

4.6 Repartiţia Poisson

Această repartiţie este folosită în modelele matematice pentru a descrie, într-un interval de timp specific, evenimente ca emisia de particule α dintr-o substanţă radioactivă, sosirile de pasageri la un aeroport, distribuţia particulelor de praf într-un anumit spaţiu, sosirile de maşini la o intersecţie, şi alte fenomene similare.

Pentru a fixa ideile, considerăm problema sosirii pasagerilor la o staţie de autobuz într-un interval de timp specificat. Notăm $X(0,t)$ numărul de sosiri din intervalul de timp $[0,t)$; $X(0,t)$ e o variabilă aleatoare discretă luând valori posibile $0,1,2,...$, iar repartiţia ei depinde de t. Funcţia ei masă de probabilitate se scrie

$$p_k(0,t) = P[X(0,t) = k], \qquad k = 0,1,2,..., \tag{4.4}$$

pentru a arăta dependenţa ei explicită de t.

Facem următoarele ipoteze de bază:

1) Dacă $t_1 < t_2 < ... < t_n$, atunci variabilele aleatoare $X(t_1,t_2), X(t_2,t_3), ..., X(t_{n-1},t_n)$ sunt independente, adică, numerele de pasageri care sosesc în intervale de timp care nu se suprapun sunt independente unul de celălalt.

2) Pentru Δt suficient de mic,

$$p_1(t, t+\Delta t) = \nu \Delta t + o(\Delta t), \tag{4.5}$$

unde $o(\Delta t)$ este o funcţie a. î.

$$\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0. \tag{4.6}$$

Această ipoteză spune că, pentru Δt suficient de mic, probabilitatea de a avea exact o sosire este proporţională cu lungimea Δt. Parametrul ν din relaţia (4.5) este numit *densitatea medie* sau *rata medie a sosirilor*.

3) Pentru Δt suficient de mic,

$$\sum_{k=2}^{\infty} p_k(t, t+\Delta t) = o(\Delta t). \tag{4.7}$$

Această condiţie implică faptul că probabilitatea de a avea două sau mai multe sosiri într-un interval suficient de mic este neglijabilă.

Din relaţiile (4.5) şi (4.7) rezultă

$$p_0(t, t+\Delta t) = 1 - \sum_{k=1}^{\infty} p_k(t, t+\Delta t) = 1 - \nu \Delta t + o(\Delta t). \tag{4.8}$$

Determinăm $p_0(0,t)$. Pentru a nu avea nicio sosire în intervalul $[0, t+\Delta t)$, trebuie să nu avem nicio sosire în ambele subintervale $[0,t)$ şi $[t, t+\Delta t)$. Datorită independenţei sosirilor în intervale nesuprapuse avem

$$p_0(0,t+\Delta t) = p_0(0,t)\,p_0(t,t+\Delta t) = p_0(0,t)\left[1-\nu\Delta t + o(\Delta t)\right]. \qquad (4.9)$$

Rearanjând relaţia (4.9) şi împărţind ambii membri prin Δt obţinem $\frac{p_0(0,t+\Delta t)-p_0(0,t)}{\Delta t} = -p_0(0,t)\left[\nu - \frac{o(\Delta t)}{\Delta t}\right]$.
Punând $\Delta t \to 0$, obţinem ecuaţia diferenţială

$$\frac{dp_0(0,t)}{dt} = -\nu p_0(0,t).$$

Soluţia ei satisfăcând condiţia iniţială $p_0(0,0) = 1$ este

$$p_0(0,t) = e^{-\nu t}. \qquad (4.10)$$

Determinarea lui $p_1(0,t)$ este similară. O sosire în intervalul $[0,t+\Delta t)$ poate fi îndeplinită doar având 0 sosiri în subintervalul $[0,t)$ şi o sosire în $[t,t+\Delta t)$, sau o sosire în $[0,t)$ şi nicio sosire în $[t,t+\Delta t)$. De aici avem

$$p_1(0,t+\Delta t) = p_0(0,t)\,p_1(t,t+\Delta t) + p_1(0,t)\,p_0(t,t+\Delta t). \qquad (4.11)$$

Înlocuind relaţiile (4.5), (4.8) şi (4.10) în relaţia (4.11) obţinem
$p_1(0,t+\Delta t) = e^{-\nu t}(\nu\Delta t + o(\Delta t)) + p_1(0,t)(1-\nu\Delta t + o(\Delta t)) \implies \frac{p_1(0,t+\Delta t)-p_1(0,t)}{\Delta t} = e^{-\nu t}\left(\nu + \frac{o(\Delta t)}{\Delta t}\right) + p_1(0,t)\left(-\nu + \frac{o(\Delta t)}{\Delta t}\right)$.
Punând $\Delta t \to 0$, obţinem

$$\frac{dp_1(0,t)}{dt} = -\nu p_1(0,t) + \nu e^{-\nu t}, \qquad p_1(0,0) = 0, \qquad (4.12)$$

ceea ce conduce la

$$p_1(0,t) = \nu t e^{-\nu t}. \qquad (4.13)$$

Continuând în acest fel, găsim pentru termenul general

$$p_k(0,t) = \frac{(\nu t)^k e^{-\nu t}}{k!}, \qquad k = 0,1,2,.... \qquad (4.14)$$

Relaţia (4.14) dă funcţia masă de probabilitate a lui $X(0,t)$, numărul de sosiri în intervalul de timp $[0,t)$ cu condiţiile de mai sus şi defineşte o *repartiţie Poisson* cu parametrii ν şi t. Parametrii ν şi t pot fi înlocuiţi de un singur parametru $\lambda = \nu t$ şi astfel putem scrie

$$p_k(0,t) = \frac{\lambda^k e^{-\lambda}}{k!}, \qquad k = 0,1,2,.... \qquad (4.15)$$

Media lui $X(0,t)$ este

$$E\left(X\left(0,t\right)\right)=\sum_{k=0}^{\infty}kp_k\left(0,t\right)=e^{-\lambda}\sum_{k=0}^{\infty}\frac{k\lambda^k}{k!}=\lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}=\lambda e^{-\lambda}e^{\lambda}=\lambda. \tag{4.16}$$

Calculăm şi dispersia:

$$E\left(X^2\left(0,t\right)\right)=\sum_{k=0}^{\infty}k^2p_k\left(0,t\right)=e^{-\lambda}\sum_{k=0}^{\infty}\frac{k^2\lambda^k}{k!}=e^{-\lambda}\sum_{k=1}^{\infty}\frac{k\lambda^k}{(k-1)!}=e^{-\lambda}\left[\sum_{k=1}^{\infty}\frac{(k-1)\lambda^k}{(k-1)!}+\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!}\right]=$$

$$e^{-\lambda}\left[\lambda^2\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}+\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}\right]=e^{-\lambda}\left[\lambda^2e^{\lambda}+\lambda e^{\lambda}\right]=\lambda^2+\lambda.$$

$$\sigma^2_{X(0,t)}=E\left(X^2\left(0,t\right)\right)-\left[E\left(X\left(0,t\right)\right)\right]^2=\lambda. \tag{4.17}$$

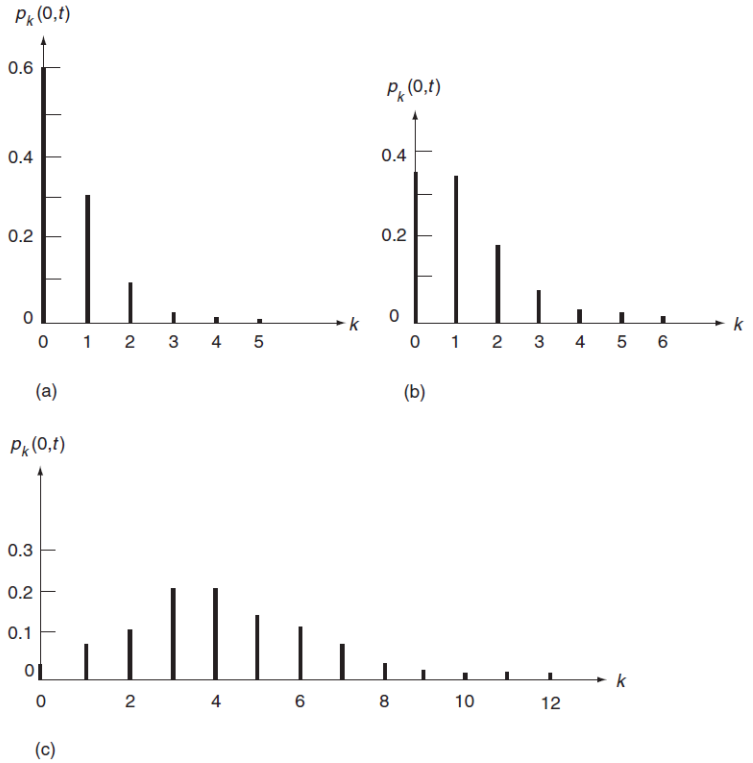
Din relaţia (4.16) se vede că parametrul ν este egal cu media numărului de sosiri pe unitatea de timp; numele "rata medie a sosirilor" pentru ν este astfel justificat. În determinarea valorii acestui parametru într-o problemă dată, el poate fi estimat din observaţii prin $\frac{m}{n}$, unde m este numărul observat de sosiri în n unităţi de timp. Similar, deoarece $\lambda=\nu t$, λ reprezintă numărul mediu de sosiri în intervalul de timp $[0,t)$.

Din relaţiile (4.16) şi (4.17) se vede că media şi dispersia cresc când rata medie creşte. În figura alăturată e reprezentată funcţia masă de probabilitate pentru repartiţia Poisson pentru

a) $\lambda=0,5$,

b) $\lambda=1$,

c) $\lambda=4$.



În general, dacă examinăm raportul $\frac{p_k(0,t)}{p_{k-1}(0,t)}$, cum am făcut la repartiţia binomială, se arată că $p_k(0,t)$ creşte şi apoi descreşte când k creşte, atingându-şi maximul când $k = [\lambda]$.

Tabelul A.2 din figurile de mai jos dă funcţia masă de probabilitate a repartiţiei Poisson pentru anumite valori ale lui λ dintre $0,1$ şi 10. Pentru $\lambda = 10$, avem $p_{23}(0,t) = 0,0002$ şi $p_{24}(0,t) = 0,0001$. În loc de "λt" în tabelul A.2 se va citi "λ".

Table A.2 Poisson mass function: a table of

$$p_k(0,t) = \frac{(\lambda t)^k e^{-\lambda t}}{k!},$$

for $k = 0$ to 24, $\lambda t = 0.1$ to 10

λt	k												
	0	1	2	3	4	5	6	7	8	9	10	11	12
0.1	0.9048	0.0905	0.0045	0.0002	0.0000								
0.2	0.8187	0.1637	0.0164	0.0011	0.0001	0.0000							
0.3	0.7408	0.2222	0.0333	0.0033	0.0002	0.0000							
0.4	0.6703	0.2681	0.0536	0.0072	0.0007	0.0001	0.0000						
0.5	0.6065	0.3033	0.0758	0.0126	0.0016	0.0002	0.0000						
0.6	0.5488	0.3293	0.0988	0.0198	0.0030	0.0004	0.0000						
0.7	0.4966	0.3476	0.1217	0.0284	0.0050	0.0007	0.0001	0.0000					
0.8	0.4493	0.3595	0.1438	0.0383	0.0077	0.0012	0.0002	0.0000					
0.9	0.4066	0.3659	0.1647	0.0494	0.0111	0.0020	0.0003	0.0000					
1.0	0.3679	0.3679	0.1839	0.0613	0.0153	0.0031	0.0005	0.0001	0.0000				
1.1	0.3329	0.3662	0.2014	0.0738	0.0203	0.0045	0.0008	0.0001	0.0000				
1.2	0.3012	0.3614	0.2169	0.0867	0.0260	0.0062	0.0012	0.0002	0.0000				
1.3	0.2725	0.3543	0.2303	0.0998	0.0324	0.0084	0.0018	0.0003	0.0001	0.0000			
1.4	0.2466	0.3452	0.2417	0.1128	0.0395	0.0111	0.0026	0.0005	0.0001	0.0000			
1.5	0.2231	0.3347	0.2510	0.1255	0.0471	0.0141	0.0035	0.0008	0.0001	0.0000			
1.6	0.2019	0.3230	0.2584	0.1378	0.0551	0.0176	0.0047	0.0011	0.0002	0.0000			
1.7	0.1827	0.3106	0.2640	0.1496	0.0636	0.0216	0.0061	0.0015	0.0003	0.0001	0.0000		
1.8	0.1653	0.2975	0.2678	0.1607	0.0723	0.0260	0.0078	0.0020	0.0005	0.0001	0.0000		
1.9	0.1496	0.2842	0.2700	0.1710	0.0812	0.0309	0.0098	0.0027	0.0006	0.0001	0.0000		
2.0	0.1353	0.2707	0.2707	0.1804	0.0902	0.0361	0.0120	0.0034	0.0009	0.0002	0.0000		
2.2	0.1108	0.2438	0.2681	0.1966	0.1082	0.0476	0.0174	0.0055	0.0015	0.0004	0.0001	0.0000	
2.4	0.0907	0.2177	0.2613	0.2090	0.1254	0.0602	0.0241	0.0083	0.0025	0.0007	0.0002	0.0000	
2.6	0.0743	0.1931	0.2510	0.2176	0.1414	0.0735	0.0319	0.0118	0.0038	0.0011	0.0003	0.0001	0.0000
2.8	0.0608	0.1703	0.2384	0.2225	0.1557	0.0872	0.0407	0.0163	0.0057	0.0018	0.0005	0.0001	0.0000
3.0	0.0498	0.1494	0.2240	0.2240	0.1680	0.1008	0.0504	0.0216	0.0081	0.0027	0.0008	0.0002	0.0001
3.2	0.0408	0.1304	0.2087	0.2226	0.1781	0.1140	0.0608	0.0278	0.0111	0.0040	0.0013	0.0004	0.0001
3.4	0.0334	0.1135	0.1929	0.2186	0.1858	0.1264	0.0716	0.0348	0.0148	0.0056	0.0019	0.0006	0.0002
3.6	0.0273	0.0984	0.1771	0.2125	0.1912	0.1377	0.0826	0.0425	0.0191	0.0076	0.0028	0.0009	0.0003
3.8	0.0224	0.0850	0.1615	0.2046	0.1944	0.1477	0.0936	0.0508	0.0241	0.0102	0.0039	0.0013	0.0004
4.0	0.0183	0.0733	0.1465	0.1954	0.1954	0.1563	0.1042	0.0595	0.0298	0.0132	0.0053	0.0019	0.0006
5.0	0.0067	0.0337	0.0842	0.1404	0.1755	0.1755	0.1462	0.1044	0.0653	0.0363	0.0181	0.0082	0.0034
6.0	0.0025	0.0149	0.0446	0.0892	0.1339	0.1606	0.1606	0.1377	0.1033	0.0688	0.0413	0.0225	0.0113
7.0	0.0009	0.0064	0.0223	0.0521	0.0912	0.1277	0.1490	0.1490	0.1304	0.1014	0.0710	0.0452	0.0264
8.0	0.0003	0.0027	0.0107	0.0286	0.0573	0.0916	0.1221	0.1396	0.1396	0.1241	0.0993	0.0722	0.0481
9.0	0.0001	0.0011	0.0050	0.0150	0.0337	0.0607	0.0911	0.1171	0.1318	0.1318	0.1186	0.0970	0.0728
10.0	0.0000	0.0005	0.0023	0.0076	0.0189	0.0378	0.0631	0.0901	0.1126	0.1251	0.1251	0.1137	0.0948

Table A.2 Continued

λt	k										
	13	14	15	16	17	18	19	20	21	22	2
5.0	0.0013	0.0005	0.0002								
6.0	0.0052	0.0022	0.0009	0.0003	0.0001						
7.0	0.0142	0.0071	0.0033	0.0014	0.0006	0.0002	0.0001				
8.0	0.0296	0.0169	0.0090	0.0045	0.0021	0.0009	0.0004	0.0002	0.0001		
9.0	0.0504	0.0324	0.0194	0.0109	0.0058	0.0029	0.0014	0.0006	0.0003	0.0001	
10.0	0.0729	0.0521	0.0347	0.0217	0.0128	0.0071	0.0037	0.0019	0.0009	0.0004	0

Teorema 4.2. Dacă X şi Y sunt variabile aleatoare independente având repartiţii Poisson cu parametrii λ_1, respectiv λ_2, atunci variabila aleatoare $Y = X_1 + X_2$ are repartiţie Poisson cu parametrul $\lambda_1 + \lambda_2$.

Propoziţie. Dacă o variabilă aleatoare X este repartizată Poisson cu parametrul λ, atunci o variabilă aleatoare Y, care este obţinută din X selectând doar cu probabilitatea p fiecare din itemii număraţi de X, este de asemenea repartizată Poisson cu parametrul $p\lambda$.

Demonstraţie. Dat fiind că $X = r$, repartiţia lui Y este binomială cu parametrii r şi p, deci:

$P(Y = k|X = r) = C_r^k p^k (1-p)^{r-k}, \qquad k = 0, 1, 2, ..., r.$

Din teorema probabilităţii totale avem

$$P(Y = k) = \sum_{r=k}^{\infty} P(Y = k|X = r) P(X = r) = \sum_{r=k}^{\infty} C_r^k \frac{p^k (1-p)^{r-k} \lambda^r e^{-\lambda}}{r!} \overset{r=n+k}{=}$$

$$\sum_{n=0}^{\infty} C_{n+k}^k \frac{p^k (1-p)^n \lambda^{n+k} e^{-\lambda}}{(n+k)!} = \frac{(p\lambda)^k e^{-\lambda}}{k!} \sum_{n=0}^{\infty} \frac{[(1-p)\lambda]^n}{n!} = \frac{(p\lambda)^k e^{-\lambda} e^{(1-p)\lambda}}{k!} = \frac{(p\lambda)^k e^{-p\lambda}}{k!}, \qquad k = 0, 1, 2, ...\square$$

Această propoziţie poate fi folosită, de exemplu, pentru situaţii în care Y e numărul de urmaşi ai unei insecte când X e numărul de ouă depuse, sau Y e numărul de uragane dezastruoase când X e numărul total de uragane care apar într-un an dat, sau Y e numărul pasagerilor care nu pot urca la bordul unui zbor dat din cauza suprarezervărilor, când X este numărul de sosiri de pasageri.

4.6.1 Repartiţii spaţiale

Repartiţia Poisson a fost dată pe baza sosirilor în timp, dar acelaşi argument se aplică la repartiţia punctelor în spaţiu. Considerăm repartiţia defectelor într-un material. Numărul de defecte într-un anumit volum are o repartiţie Poisson dacă ipotezele 1-3 sunt valide, cu intervalele de timp înlocuite de volume,

şi este rezonabil să presupunem că probabilitatea de a găsi k defecte în orice regiune depinde numai de volum şi nu de forma regiunii.

Alte situaţii fizice în care repartiţia Poisson e folosită includ numărul de bacterii pe o placă Petri, repartiţia fertilizatoarelor împrăştiate cu avionul pe un câmp şi repartiţia poluanţilor industriali într-o regiune dată.

4.6.2 Aproximarea Poisson a repartiţiei binomiale

Fie $X \sim B(n,p)$.

$p_X(k) = C_n^k p^k (1-p)^{n-k}, \qquad k = 0,1,2,...,n.$

Considerăm cazul când $n \to \infty$ şi $p \to 0$ a. î. $np = \lambda$ rămâne fixat. Observăm că λ este media lui X, care e presupusă a rămâne constantă. Atunci

$p_X(k) = C_n^k \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k}, \qquad k = 0,1,2,...,n.$

Deoarece $n \to \infty$, factorialele $n!$ şi $(n-k)!$ care apar în $C_n^k = \frac{n!}{k!(n-k)!}$ pot fi aproximate prin formula lui Stirling

$n! \cong (2\pi)^{\frac{1}{2}} e^{-n} n^{n+\frac{1}{2}}$.

Obţinem

$$p_X(k) \cong \frac{(2\pi)^{\frac{1}{2}} e^{-n} n^{n+\frac{1}{2}}}{k!(2\pi)^{\frac{1}{2}} e^{-n+k}(n-k)^{n-k+\frac{1}{2}}} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!e^k} \left(\frac{n}{n-k}\right)^{n-k+\frac{1}{2}} \left(1-\frac{\lambda}{n}\right)^{n-k}.$$

Deoarece

$\lim\limits_{n\to\infty} \left(1+\frac{c}{n}\right)^n = e^c,$

avem

$$\lim_{n\to\infty} \left(\frac{n}{n-k}\right)^{n-k+\frac{1}{2}} = \frac{1}{\lim\limits_{n\to\infty}\left(1-\frac{k}{n}\right)^{n-k+\frac{1}{2}}} = \frac{1}{\left[\lim\limits_{n\to\infty}\left(1-\frac{k}{n}\right)^n\right]^{\lim\limits_{n\to\infty}\frac{n-k+\frac{1}{2}}{n}}} = \frac{1}{e^{-k}},$$

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{n-k} = \left[\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n\right]^{\lim\limits_{n\to\infty}\frac{n-k}{n}} = e^{-\lambda}.$$

Obţinem

$p_X(k) \cong \frac{\lambda^k e^{-\lambda}}{k!}, \qquad k = 0,1,....$

Această aproximare Poisson a repartiţiei binomiale uşurează calculele şi se foloseşte în practică atunci când $n > 10$ şi $p < 0,1$. Repartiţia Poisson îşi găseşte aplicabilitate în acest caz în probleme în care probabilitatea apariţiei unui eveniment este mică. De aceea, repartiţia Poisson se mai numeşte adesea *repartiţia evenimentelor rare.*

5 Principalele repartiţii continue, cu densităţi de repartiţie şi proprietăţile lor

5.1 Repartiţia uniformă

O variabilă aleatoare continuă X are *repartiţie uniformă* pe un interval de la a la b $(b > a)$ dacă este egal probabil să ia orice valoare din acest interval. Densitatea lui X este constantă pe intervalul $[a, b]$ şi are forma

$$f_X(x) = \begin{cases} c, & \text{dacă } x \in [a, b], \\ 0, & \text{altfel.} \end{cases}$$

Deoarece

$$\int_{-\infty}^{\infty} f_X(x)\,dx = \int_{-\infty}^{a} 0dx + \int_{a}^{b} cdx + \int_{b}^{\infty} 0dx = 0 + cx|_a^b + 0 = c(b-a),$$

din condiţia

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

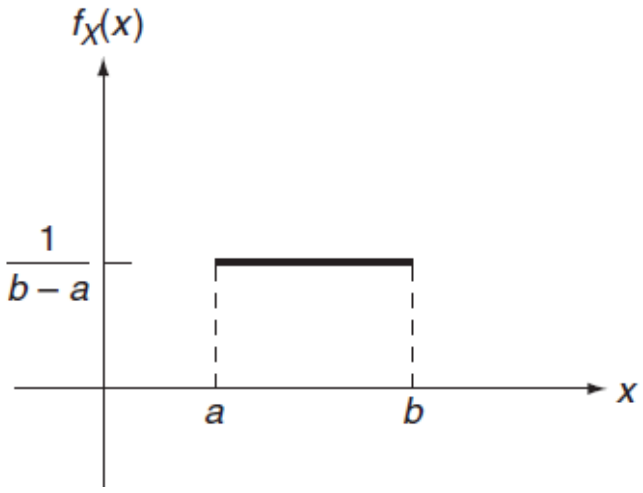
obţinem

$c = \frac{1}{b-a}$.

Deci

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{dacă } x \in [a, b], \\ 0, & \text{altfel.} \end{cases} \tag{5.1}$$

După cum vedem din figura de mai jos, este constantă pe $[a, b]$ şi înălţimea trebuie să fie $\frac{1}{b-a}$ pentru ca aria de sub densitate să fie 1.



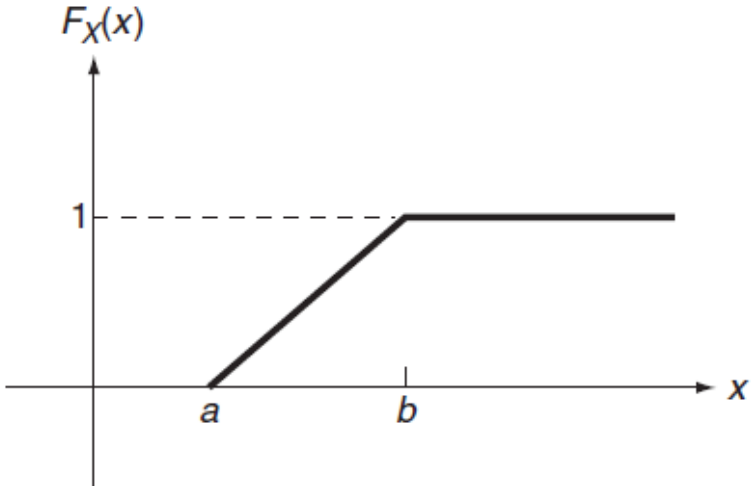
Funcţia de repartiţie a lui X este

$$F_X(x)=\int_{-\infty}^{x} f_X(u)\,du=\begin{cases}\int_{-\infty}^{x} 0du=0, \text{ dacă } x<a;\\ \int_{-\infty}^{a} 0du+\int_{a}^{x}\frac{1}{b-a}du=0+\frac{u}{b-a}|_a^x=\frac{x-a}{b-a}, \text{ dacă } x\in[a,b],\\ \int_{-\infty}^{a} 0du+\int_{a}^{b}\frac{1}{b-a}du+\int_{b}^{x}0du=0+1+0=1, \text{ dacă } x>b.\end{cases}$$

Deci

$$F_X(x)=\begin{cases}0, \text{ dacă } x<a;\\ \frac{x-a}{b-a}, \text{ dacă } x\in[a,b],\\ 1, \text{ dacă } x>b,\end{cases} \tag{5.2}$$

ceea ce este prezentat grafic în figura următoare.



Media şi dispersia lui X sunt

$m_X=\int_{-\infty}^{\infty} xf_X(x)\,dx=\int_{-\infty}^{a}0dx+\int_{a}^{b}\frac{x}{b-a}dx+\int_{b}^{\infty}0dx=0+\frac{x^2}{2(b-a)}|_a^b+0=\frac{b^2-a^2}{2(b-a)}=\frac{a+b}{2}$.

$\sigma_X^2=\int_{-\infty}^{\infty}(x-m_X)^2 f_X(x)\,dx=\int_{-\infty}^{a}0dx+\frac{1}{b-a}\int_{a}^{b}(x-m_X)^2dx+\int_{b}^{\infty}0dx=0+\frac{1}{b-a}\cdot\frac{(x-m_X)^3}{3}|_a^b+0=\frac{(b-m_X)^3-(a-m_X)^3}{3(b-a)}=\frac{(b-a)\left[(b-m_X)^2+(b-m_X)(a-m_X)+(a-m_X)^2\right]}{3(b-a)}=\frac{1}{3}\left[\left(\frac{b-a}{2}\right)^2-\left(\frac{b-a}{2}\right)^2+\left(\frac{b-a}{2}\right)^2\right]=\frac{(b-a)^2}{12}$.

Repartiţia uniformă e una dintre cele mai simple repartiţii şi e folosită de obicei în situaţii unde nu este niciun motiv de a da probabilităţi inegale la valori posibile luate de variabila aleatoare pe un interval dat. De exemplu, timpul de sosire a unui zbor poate fi considerat uniform repartizat pe un anumit interval de timp, sau repartiţia distanţei de la locul încărcăturilor vii pe un pod la un suport terminal poate fi reprezentată adecvat printr-o repartiţie uniformă

pe întinderea podului. Adesea se ataşează o repartiţie uniformă unei anumite variabile aleatoare din cauza unei lipse de informaţie, dincolo de cunoaşterea intervalului de valori.

5.1.1 Repartiţia uniformă bivariată

Fie variabila aleatoare X repartizată uniform pe un interval $[a_1, b_1]$ şi variabila aleatoare Y repartizată uniform pe un interval $[a_2, b_2]$. Mai mult, presupunem că sunt independente. Atunci densitatea comună a lui X şi Y este

$$f_{XY}(x,y) = f_X(x) f_Y(y) = \begin{cases} \frac{1}{(b_1-a_1)(b_2-a_2)}, \text{ pentru } x \in [a_1, b_1] \text{ şi } y \in [a_2, b_2], \\ 0, \text{ altfel.} \end{cases} \tag{5.3}$$

Ia forma unei suprafeţe plate mărginită de $[a_1, b_1]$ de-a lungul axei Ox şi $[a_2, b_2]$ de-a lungul axei Oy.

5.2 Repartiţia Gaussiană sau normală

Cea mai importantă repartiţie în teorie ca şi în aplicaţii este repartiţia *Gaussiană* sau *normală*. O variabilă aleatoare X este *Gaussiană* sau *normală* dacă are densitatea de forma

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right], \qquad -\infty < x < \infty, \tag{5.4}$$

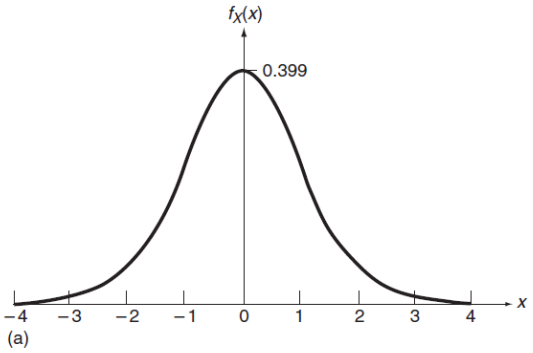
unde m şi σ sunt doi parametri, cu $\sigma > 0$. Alegerea acestor simboluri particulare ca parametri va deveni clară imediat.

Funcţia de repartiţie corespunzătoare este

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{(u-m)^2}{2\sigma^2}\right] du, \qquad -\infty < x < \infty. \tag{5.5}$$

Ea nu poate fi exprimată analitic, dar poate fi evaluată numeric pentru orice x.

Densitatea şi funcţia de repartiţie din relaţiile (5.4) şi (5.5) sunt reprezentate grafic în figurile (a), respectiv (b) următoare, pentru $m = 0$ şi $\sigma = 1$.



Graficul densităţii f_X în acest caz particular este o curbă în formă de clopot, simetrică în jurul originii.

Calculăm media şi dispersia lui X.

$$E(X)=\int_{-\infty}^{\infty} x f_X(x)\,dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} x\exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]dx \overset{\frac{x-m}{\sigma}=u}{=}$$

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(\sigma u+m)\exp\left(-\frac{u^2}{2}\right)\cdot\sigma\,du=\frac{1}{\sqrt{2\pi}}\left[\sigma\int_{-\infty}^{\infty}\underbrace{u\exp\left(-\frac{u^2}{2}\right)}_{\text{impară}}du+m\int_{-\infty}^{\infty}\exp\left(-\frac{u^2}{2}\right)du\right]=$$

$$\frac{1}{\sqrt{2\pi}}\left(\sigma\cdot 0+m\sqrt{2\pi}\right)=m.$$

$$var(X)=\int_{-\infty}^{\infty}[x-E(X)]^2 f_X(x)\,dx=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}(x-m)^2\exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]dx \overset{\frac{x-m}{\sigma}=u}{=}$$

$$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}\sigma^2u^2\exp\left(-\frac{u^2}{2}\right)\cdot\sigma\,du=\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}u^2\exp\left(-\frac{u^2}{2}\right)du=-\frac{\sigma^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}u\left(\exp\left(-\frac{u^2}{2}\right)\right)'du=$$

$$-\frac{\sigma^2}{\sqrt{2\pi}}\left[u\exp\left(-\frac{u^2}{2}\right)|_{-\infty}^{\infty}-\int_{-\infty}^{\infty}\exp\left(-\frac{u^2}{2}\right)du\right]=-\frac{\sigma^2}{\sqrt{2\pi}}\left(0-\sqrt{2\pi}\right)=\sigma^2.$$

Vedem astfel că cei doi parametri m şi σ din repartiţie sunt media şi respectiv deviaţia standard a lui X. Această observaţie justifică alegerea noastră a acestor simboluri speciale pentru ei şi de asemenea pune în evidenţă o proprietate importantă a repartiţiei normale - cunoaşterea mediei şi dispersiei ei

caracterizează complet repartiţia normală. Deoarece ne vom referi frecvent la repartiţia normală, o notăm cu $N\left(m,\sigma^2\right)$. De exemplu, $X \sim N\left(0,9\right)$ înseamnă că X are densitatea dată de relaţia (5.4) cu $m = 0$ şi $\sigma = 3$.

Se poate arăta că momentele centrate ale lui $X \sim N\left(m,\sigma^2\right)$ sunt

$$\mu_n = \begin{cases} 0, \text{ dacă } n \text{ e impar,} \\ 1 \cdot 3 \cdot ... \cdot (n-1)\,\sigma^n, \text{ dacă } n \text{ e par.} \end{cases} \tag{5.6}$$

Coeficientul de exces $\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$ este 0 pentru o repartiţie normală. De aceea ea este folosită ca repartiţie de referinţă pentru γ_2.

5.2.1 Teorema limită centrală

Marea importanţă practică a repartiţiei normale provine din teorema limită centrală.

Teorema 5.1 (Teorema limită centrală) (Lindberg 1922). Fie $\{X_n\}$ un şir de variabile aleatoare independente şi identic repartizate cu mediile m şi dispersiile σ^2. Fie

$$Y = \sum_{j=1}^{n} X_j,$$

şi fie variabila aleatoare normalizată Z definită ca

$$Z = \frac{Y - nm}{\sigma\sqrt{n}}.$$

Atunci funcţia de repartiţie a lui Z, $F_Z\left(z\right)$, converge la $N\left(0,1\right)$ când $n \to \infty$ pentru orice z fixat.

Acest rezultat poate fi extins în câteva direcţii, incluzând cazurile în care Y este o sumă de variabile aleatoare dependente şi neidentic repartizate.

Teorema limită centrală descrie o clasă foarte generală de fenomene aleatoare pentru care repartiţiile pot fi aproximate cu repartiţia normală. Când proprietatea de a fi aleator a unui fenomen fizic este cumularea a multor efecte aleatoare aditive mici, el tinde la o repartiţie normală, indiferent de repartiţiile efectelor individuale. De exemplu, consumul de combustibil la toate automobilele unei anumite mărci, presupus fabricate prin procese identice, diferă de la un automobil la altul. Această proprietate de a fi aleator provine de la o largă varietate de surse, incluzând, printre alte lucruri: inexactităţile inerente în procesele de fabricare, neuniformităţile în materialele folosite, diferenţele în greutate şi alte specificaţii, diferenţe în calitatea combustibilului şi diferenţe în comportamentul şoferilor. Dacă se acceptă faptul că fiecare dintre aceste diferenţe contribuie la proprietatea de a fi aleator a consumului de combustibil, teorema limită centrală ne spune că el tinde la o repartiţie normală. Din acelaşi motiv, variaţiile de temperatură dintr-o cameră, erorile de citire asociate cu un instrument, erorile de ţintire ale unei anumite arme, ş.a.m.d. pot fi aproximate rezonabil prin repartiţii normale.

5.2.2 Tabele de probabilităţi

Datorită importanţei sale, suntem adesea puşi în situaţia să evaluăm probabilităţile asociate cu o variabilă aleatoare normală $X \sim N\left(m, \sigma^2\right)$, ca

$$P\left(a < X \leq b\right) = \frac{1}{\sigma\sqrt{2\pi}} \int_a^b \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right] dx. \tag{5.7}$$

Integrala de mai sus nu poate fi calculată analitic şi este în general calculată numeric. Pentru comoditate sunt date tabele care ne permit să calculăm probabilităţi ca cea din relaţia (5.7).

Tabelarea funcţiei de repartiţie pentru repartiţia normală cu $m = 0$ şi $\sigma = 1$ este dată în tabelul A.3 din figura următoare.

Table A.3 Standardized normal distribution function: a table of

$$F_U(u) = \frac{1}{(2\pi)^{1/2}} \int_{-\infty}^{u} \mathrm{e}^{-x^2/2}\,\mathrm{d}x,$$

for $u = 0.0$ to 3.69

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	0.5733
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	0.7291	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	0.9482	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	0.9778	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	0.8874	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
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

O variabilă aleatoare cu repartiţia $N\left(0,1\right)$ se numeşte variabilă aleatoare normală *standardizată* şi o vom nota cu U. Tabelul alăturat dă $F_U\left(u\right)$ doar pentru punctele din partea dreaptă a repartiţiei (adică pentru $u \geq 0$). Valorile corespunzătoare pentru $u < 0$ sunt obţinute din proprietatea de simetrie a repartiţiei normale standardizate, din relaţia

$$F_U\left(-u\right) = 1 - F_U\left(u\right). \tag{5.8}$$

Mai întâi, tabelul împreună cu relaţia (5.8) pot fi folosite pentru a determina $P(a < U \leq b)$ pentru orice a şi b. De exemplu,

$P(-1,5 < U \leq 2,5) = F_U(2,5) - F_U(-1,5) \overset{(5.8)}{=} F_U(2,5) - [1 - F_U(1,5)] \overset{\text{tabel}}{=} 0,9938 - 1 + 0,9332 = 0,927.$

Mai important, tabelul şi relaţia (5.8) sunt de asemenea suficiente pentru a determina probabilităţile asociate cu variabilele aleatoare normale cu medii si dispersii arbitrare.

Teorema 5.2. Fie $X \sim N\left(m, \sigma^2\right)$. Atunci $\frac{X-m}{\sigma} \sim N(0,1)$, adică

$$U = \frac{X - m}{\sigma}. \tag{5.9}$$

Teorema 5.2 implică faptul că, dacă $X \sim N\left(m, \sigma^2\right)$, atunci

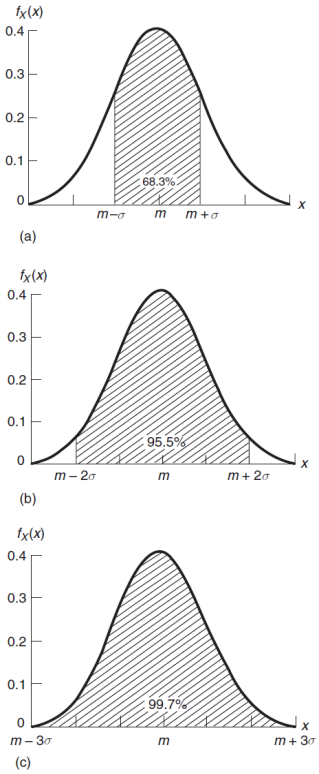
$$P(a < X \leq b) = P(a < \sigma U + m \leq b) = P\left(\frac{a-m}{\sigma} < U \leq \frac{b-m}{\sigma}\right). \tag{5.10}$$

Valoarea din membrul drept poate fi găsită acum din tabel cu ajutorul relaţiei (5.8), dacă este necesar.

Exemplul 5.1. Să calculăm $P(m - k\sigma < X \leq m + k\sigma)$, unde $X \sim N\left(m, \sigma^2\right)$.

$$P(m - k\sigma < X \leq m + k\sigma) \overset{(5.10)}{=} P(-k < U \leq k) = F_U(k) - F_U(-k) \overset{(5.8)}{=} 2F_U(k) - 1. \tag{5.11}$$

Observăm că rezultatul din exemplul 5.1 este independent de m şi σ şi este funcţie doar de k. Astfel, probabilitatea ca X să ia valori între k deviaţii standard în jurul mediei sale depinde numai de k şi este dată de ecuaţia (5.11). Se vede din tabel că aproximativ $68,3\%, 95,5\%$ şi $99,7\%$ din aria de sub o densitate normală se află în zonele $m \pm \sigma, m \pm 2\sigma$, respectiv $m \pm 3\sigma$, după cum se vede din figurile (a)-(c) următoare.



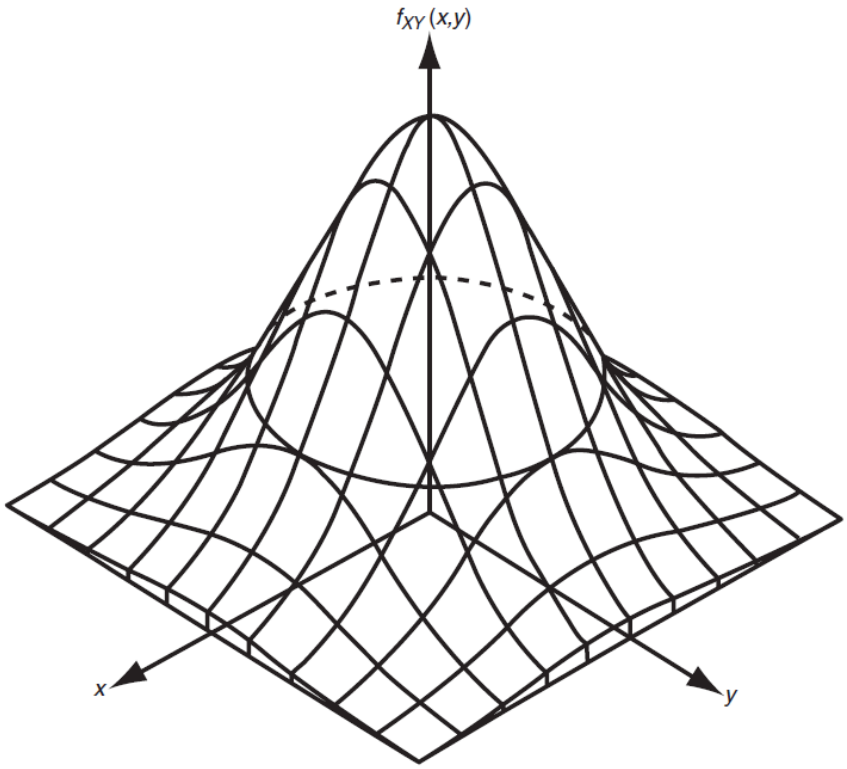
De exemplu, şansele sunt de aproximativ $99,7\%$ ca o mostră selectată aleator dintr-o repartiţie normală să fie în regiunea $m \pm 3\sigma$ (figura (c)).

5.2.3 Repartiţia normală multivariată

Două variabile aleatoare X şi Y se numesc *comun normale* dacă densitatea comună a lor are forma

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-m_X}{\sigma_X}\right)^2 - 2\rho\frac{(x-m_X)(y-m_Y)}{\sigma_X\sigma_Y} + \left(\frac{y-m_Y}{\sigma_Y}\right)^2\right]\right\}, \tag{5.12}$$

unde $(-\infty,-\infty) < (x,y) < (\infty,\infty)$. Relaţia (5.12) descrie *repartiţia normală bivariată*. Sunt cinci parametri asociaţi cu ea: $m_X, m_Y, \sigma_X\,(>0), \sigma_Y\,(>0)$, şi $\rho\ (|\rho| \leq 1)$. O reprezentare grafică tipică a acestei densităţi, pentru $m_X = m_Y = 0$ şi $\sigma_X = \sigma_Y$ este în figura următoare.



Densitatea marginală a variabilei aleatoare X este

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy = \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left[-\frac{(x-m_X)^2}{2\sigma_X^2}\right], \qquad -\infty < x < \infty.$$

Astfel, $X \sim N\left(m_X, \sigma_X^2\right)$. Similar, $Y \sim N\left(m_Y, \sigma_Y^2\right)$ şi $\rho = \frac{\mu_{XY}}{\sigma_X \sigma_Y}$ este coeficientul de corelaţie al lui X şi Y. Vedem astfel că cei cinci parametri conţinuţi în densitatea bivariată $f_{XY}(x,y)$ reprezintă cinci momente importante asociate cu variabilele aleatoare. De asemenea observăm că repartiţia normală bivariată este complet caracterizată de momentele comune de ordinul 1 şi 2 ale lui X şi Y.

Teorema 5.3. Corelaţie zero implică independenţă când variabilele aleatoare sunt comun normale.

Demonstraţie. Punând $\rho = 0$ în relaţia (5.12), obţinem

$$f_{XY}(x,y) = \left\{\frac{1}{\sigma_X\sqrt{2\pi}} \exp\left[-\frac{(x-m_X)^2}{2\sigma_X^2}\right]\right\}\left\{\frac{1}{\sigma_Y\sqrt{2\pi}} \exp\left[-\frac{(x-m_Y)^2}{2\sigma_Y^2}\right]\right\} = f_X(x)\,f_Y(y),$$

adică X şi Y sunt independente.□

Această proprietate nu este valabilă în general.

Avem *repartiţie normală multivariată* când cazul a două variabile aleatoare e extins la cel implicând n variabile aleatoare.

Fie n variabile aleatoare, $X_1, X_2, ..., X_n$. Ele se numesc *comun normale* dacă densitatea lor comună e de forma

$$f_{X_1X_2...X_n}(x_1, x_2, ..., x_n) = f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} |\Lambda|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \Lambda^{-1}(\mathbf{x}-\mathbf{m})\right], \qquad -\infty < \mathbf{x} < \infty,$$

unde $\mathbf{m}^T = (m_1, m_2, ..., m_n) = (E(X_1), E(X_2), ..., E(X_n))$, $\Lambda = \left(\mu_{ij}\right)_{i,j=\overline{1,n}}$ este matricea de covarianţă a lui $\mathbf{X}$, cu

$$\mu_{ij} = E\left((X_i - m_i)(X_j - m_j)\right),$$

iar $|\Lambda|$ este determinantul lui Λ. Din nou vedem că o repartiţie normală comună este complet specificată de momentele comune de ordinul 1 şi 2, aceste momente determinând de asemenea momentele comune de ordine mai mari ca 2. Se poate arăta că, în cazul când variabilele aleatoare $X_1, X_2, ..., X_n$ au mediile 0, toate momentele de ordin impar ale lor sunt 0, şi, pentru n par

$$E(X_1X_2...X_n) = \sum_{m_1,...,m_n} E(X_{m_1}X_{m_2}) E(X_{m_2}X_{m_3}) ... E\left(X_{m_{n-1}}X_{m_n}\right).$$

Suma de mai sus e luată după toate combinaţiile posibile de $\frac{n}{2}$ perechi de n variabile aleatoare şi are $1 \cdot 3 \cdot 5 \cdot ... \cdot (n-3) \cdot (n-1)$ termeni.

5.2.4 Sume de variabile aleatoare normale

Teorema 5.4. Fie $X_1, X_2, ..., X_n$ n variabile aleatoare repartizate comun normal (nu necesar independente). Atunci variabila aleatoare

$$Y = c_1X_1 + c_2X_2 + ... + c_nX_n$$

este repartizată normal, unde $c_1, c_2, ..., c_n$ sunt constante.

Teorema 5.5. Fie $X_1, X_2, ..., X_n$ n variabile aleatoare repartizate normal (nu necesar independente). Atunci variabilele aleatoare $Y_1, Y_2, ..., Y_m$, unde

$$Y_j = \sum_{k=1}^{n} c_{jk}X_k, \qquad j = 1, 2, ..., m,$$

sunt repartizate comun normal.

5.3 Repartiţia lognormală

Definiţia 5.1. Fie

$$X \sim N\left(m_X, \sigma_X^2\right). \tag{5.13}$$

Variabila aleatoare

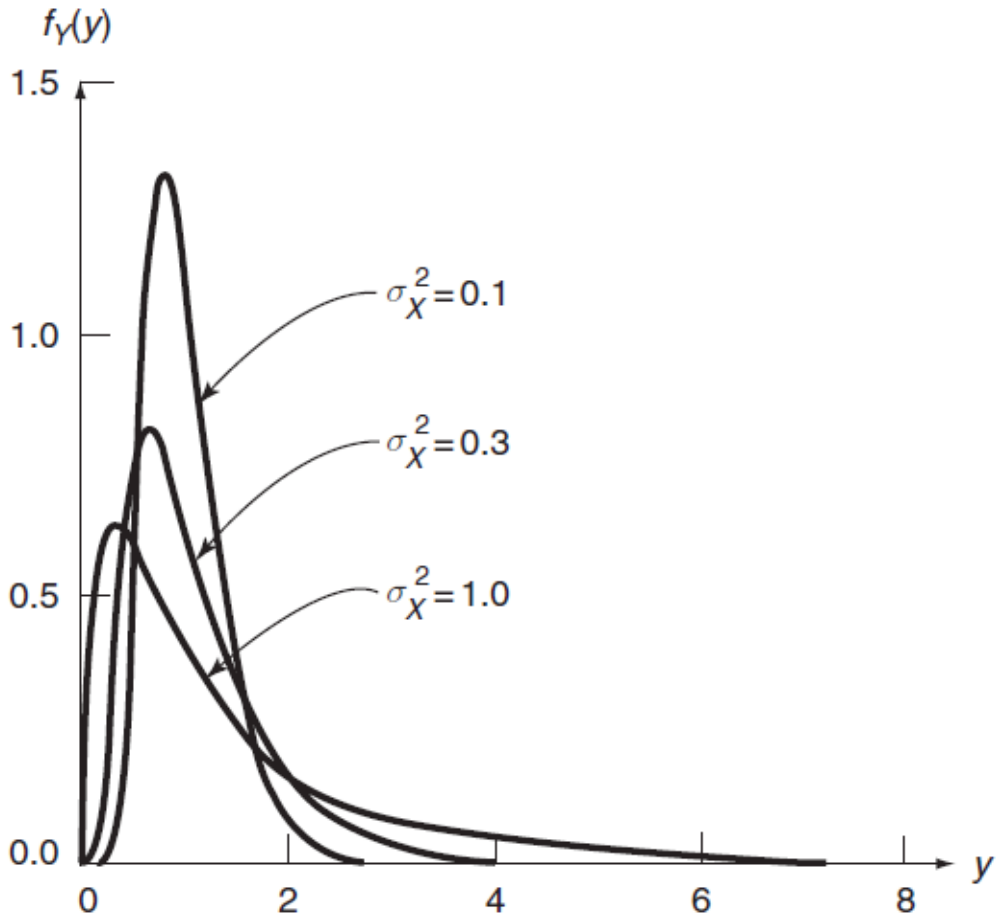
$$Y = e^X \tag{5.14}$$

se spune că are repartiţie *lognormală.*

Densitatea lui Y este

$$f_Y(y) = \begin{cases} \frac{1}{y\sigma_X\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_X^2}\left(\ln y - m_X\right)^2\right], \text{ pentru } y > 0; \\ 0, \text{ altfel.} \end{cases} \tag{5.15}$$

Relaţia (5.15) arată că Y are o repartiţie unilaterală (adică ia valori numai în zona pozitivă a lui y). Această proprietate o face atractivă pentru cantităţi fizice care sunt restricţionate să aibă doar valori pozitive. În plus, $f_Y(y)$ ia multe forme diferite pentru valori diferite ale lui m_X şi σ_X ($\sigma_X > 0$). După cum se vede din figura următoare, care dă graficele lui f_Y pentru $m_X = 0$ şi 3 valori ale lui σ_X^2, densitatea lui Y este asimetrică spre dreapta, această caracteristică devenind mai pronunţată când σ_X creşte.



Parametrii m_X şi σ_X care apar în densitatea lui Y sunt media şi deviaţia standard a lui X, sau $\ln Y$, dar nu ale lui Y. Pentru a obţine o pereche mai naturală de parametri pentru $f_Y(y)$, observăm că, dacă medianele lui X şi Y sunt notate cu θ_X, respectiv θ_Y, definiţia medianei unei variabile aleatoare dă

$\frac{1}{2} = P(Y \leq \theta_Y) = P(X \leq \ln \theta_Y) = P(X \leq \theta_X)$,

de unde

$$\ln \theta_Y = \theta_X.$$

Deoarece, datorită simetriei repartiţiei normale,

$\theta_X = m_X$,

putem scrie

$$m_X = \ln \theta_Y.$$

Scriind $\sigma_X = \sigma_{\ln Y}$, densitatea lui Y poate fi scrisă

$$f_Y(y) = \begin{cases} \frac{1}{y\sigma_{\ln Y}\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2_{\ln Y}} \ln^2\left(\frac{y}{\theta_Y}\right)\right], \text{ pentru } y > 0; \\ 0, \text{ altfel.} \end{cases}$$

În termeni de θ_Y şi $\sigma_{\ln Y}$, media şi dispersia lui Y sunt

$$m_Y = \theta_Y \exp\left(\frac{\sigma^2_{\ln Y}}{2}\right),$$
$$\sigma^2_Y = m^2_Y \left[\exp\left(\sigma^2_{\ln Y}\right) - 1\right].$$

5.3.1 Tabele de probabilităţi

Datorită legăturii dintre repartiţia normală şi repartiţia lognormală prin relaţia (5.14), calculele de probabilităţi implicând o variabilă aleatoare repartizată lognormal pot fi făcute cu ajutorul tabelului de probabilităţi pentru variabile aleatoare normale.

Considerăm funcţia de repartiţie a lui Y. Avem

$$F_Y(y) = P(Y \leq y) = P(X \leq \ln y) = F_X(\ln y), \qquad y > 0.$$

Deoarece media lui X este $\ln \theta_Y$ şi dispersia lui X este $\sigma^2_{\ln Y}$, avem

$$F_Y(y) = F_U\left(\frac{\ln y - \ln \theta_Y}{\sigma_{\ln Y}}\right) = F_U\left[\frac{1}{\sigma_{\ln Y}} \ln\left(\frac{y}{\theta_Y}\right)\right], \qquad y > 0. \tag{5.16}$$

Deoarece F_U este tabelată, relaţia (5.16) poate fi folosită pentru calcule de probabilităţi asociate cu Y, cu ajutorul tabelului probabilităţii normale.

5.4 Repartiţia gamma şi repartiţii în legătură cu aceasta

Repartiţia gamma este unilaterală şi densitatea asociată cu ea este

$$f_X(x) = \begin{cases} \frac{\lambda^\eta}{\Gamma(\eta)} x^{\eta-1} e^{-\lambda x}, \text{ pentru } x > 0; \\ 0, \text{ altfel,} \end{cases} \tag{5.17}$$

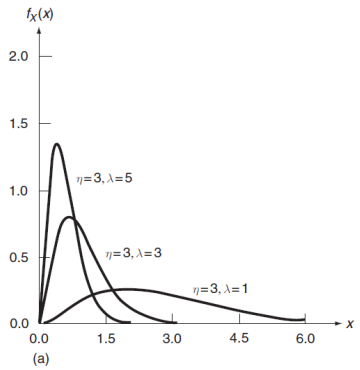
unde $\Gamma(\eta)$ este funcţia gamma:

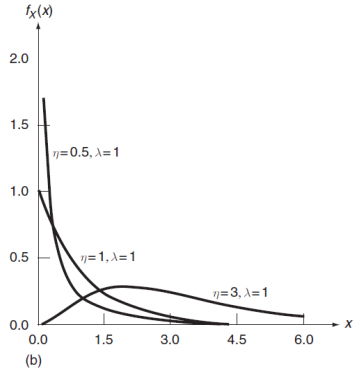
$$\Gamma(\eta) = \int_0^\infty u^{\eta-1} e^{-u} du,$$

care este tabelată, şi

$$\Gamma(n) = (n-1)!, \forall n \in \mathbb{N}^*.$$

Parametrii distribuţiei gamma sunt η şi λ; ambii sunt pozitivi. Deoarece repartiţia gamma este unilaterală, cantităţile fizice care pot lua doar valori pozitive sunt frecvent modelate de ea, servind ca model util datorită versatilităţii ei în sensul că o varietate largă de forme ale densităţii gamma poate fi obţinută variind valorile lui η şi λ, după cum arată figurile (a) şi (b) următoare.





Observăm din aceste figuri că η determină forma repartiţiei în timp ce λ este un parametru de scară pentru repartiţie. În general, densitatea gamma este unimodală, cu vârful în $x = 0$ pentru $\eta \leq 1$ şi în $x = \frac{\eta-1}{\lambda}$ pentru $\eta > 1$.

Se poate arăta că repartiţia gamma este un model pentru timpul cerut pentru un total de exact η sosiri Poisson. Datorită largii aplicabilităţi a sosirilor Poisson, repartiţia gamma are de asemenea numeroase aplicaţii.

Funcţia de repartiţie a variabilei aleatoare X având repartiţia gamma este

$$F_X(x) = \begin{cases} \int_0^x f_X(u)\,du = \frac{\lambda^\eta}{\Gamma(\eta)} \int_0^x u^{\eta-1} e^{-\lambda u} du = \frac{\Gamma(\eta,\lambda x)}{\Gamma(\eta)}, \text{ pentru } x > 0; \\ 0, \text{ altfel.} \end{cases} \tag{5.18}$$

$\Gamma(\eta, u)$ este funcţia gamma incompletă,

$$\Gamma(\eta, u) = \int_0^u x^{\eta-1} e^{-x} dx,$$

care este de asemenea tabelată.

Media şi dispersia variabilei aleatoare gamma repartizate X sunt

$$m_X = \frac{\eta}{\lambda}, \qquad \sigma_X^2 = \frac{\eta}{\lambda^2}. \tag{5.19}$$

5.4.1 Repartiţia exponenţială

Când $\eta = 1$, densitatea gamma dată de relaţia (5.17) se reduce la forma exponenţială

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, \text{ pentru } x > 0; \\ 0, \text{ altfel,} \end{cases} \tag{5.20}$$

unde $\lambda > 0$ este parametrul repartiţiei. Funcţia de repartiţie, media şi dispersia ei se obţin din relaţiile (5.18) şi (5.19) punând $\eta = 1$:

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, \text{ pentru } x \geq 0; \\ 0, \text{ altfel,} \end{cases} \tag{5.21}$$

$$m_X = \frac{1}{\lambda}, \qquad \sigma_X^2 = \frac{1}{\lambda^2}. \tag{5.22}$$

Timpul între sosiri Fie variabila aleatoare $X(0,t)$, numărul de sosiri în intervalul de timp $[0,t)$ şi presupunem că este repartizată Poisson. Fie T variabila aleatoare care dă intervalul de timp dintre 2 sosiri succesive. Funcţia ei de repartiţie este, din definiţie

$$F_T(t) = P(T \leq t) = \begin{cases} 1 - P(T > t), \text{ pentru } t \geq 0; \\ 0, \text{ altfel.} \end{cases}$$

Evenimentul $T > t$ e echivalent cu evenimentul că nu e nicio sosire în intervalul de timp $[0,t)$, sau $X(0,t) = 0$. Deoarece, din relaţia (4.10), $P(X(0,t) = 0) = e^{-\nu t}$, avem

$$F_T(t) = \begin{cases} 1 - e^{-\nu t}, \text{ pentru } t \geq 0; \\ 0, \text{ altfel.} \end{cases}$$

Comparând această expresie cu relaţia (5.21), putem stabili că timpul între 2 sosiri Poisson succesive are o repartiţie exponenţială; parametrul ν din repartiţia lui T este rata medie a sosirilor Poisson.

Deoarece timpurile între sosiri Poisson sunt independente, timpul cerut pentru un total de n sosiri Poisson este o sumă de n variabile aleatoare independente şi repartizate exponenţial. Fie $T_j, j = 1, 2, ..., n$, timpul între sosirile $j-1$ şi j. Timpul cerut pentru un total de n sosiri, notat cu X_n este

$$X_n = T_1 + T_2 + ... + T_n,$$

unde $T_j, j = 1, 2, ..., n$, sunt independente şi repartizate exponenţial cu acelaşi parametru ν. Se poate arăta că X_n este gamma repartizată cu $\eta = n$ şi $\lambda = \nu$. Astfel, repartiţia gamma este potrivită pentru a descrie timpul cerut pentru un total de η sosiri Poisson.

Fiabilitatea şi legea de defectare exponenţială În studiile de fiabilitate, timpul până la defectarea unui component fizic sau un sistem este repartizat exponenţial, dacă unitatea se defectează imediat ce un singur eveniment, ca defectarea unui component, apare, presupunând că astfel de fenomene se întâmplă independent.

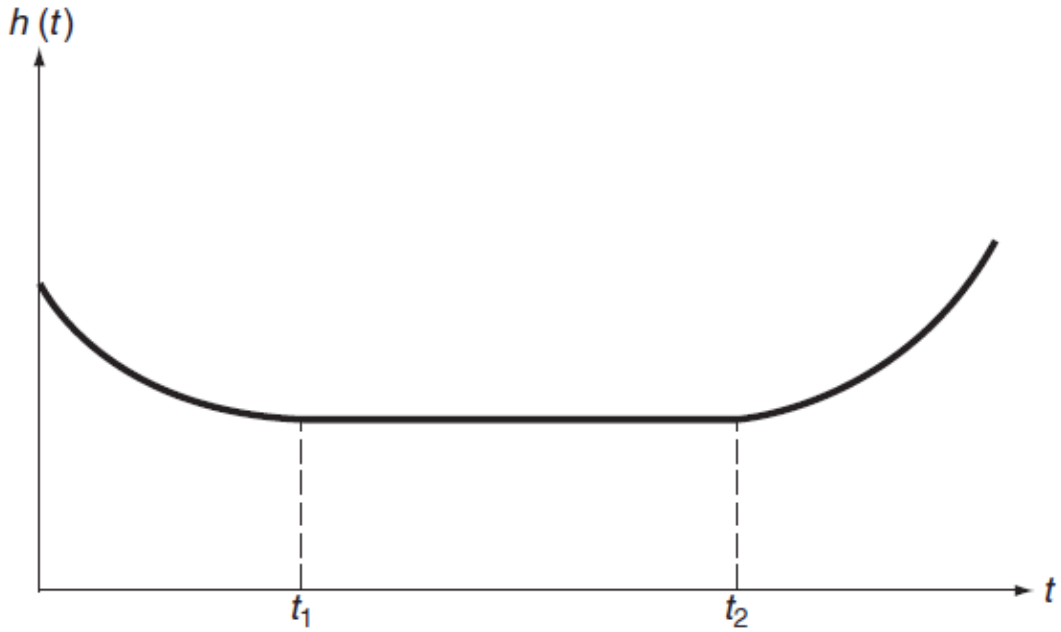
Fie variabila aleatoare T, timpul până la defectarea unui component sau sistem. Funcţia care dă probabilitatea de defecţiune în timpul unei creşteri mici de timp, presupunând că nicio defecţiune nu a apărut înainte de acel timp e notată cu $h(t)$ şi se numeşte *funcţia hazard* sau *rata defectării* şi este definită de

$$h(t)\,dt = P\left(t < T \leq t + dt | T \geq t\right),$$

ceea ce dă

$$h(t) = \frac{f_T(t)}{1 - F_T(t)}. \tag{5.23}$$

În studiile de fiabilitate, o funcţie hazard potrivită pentru multe fenomene ia aşa numita "formă de cadă", arătată în figura următoare.



Porţiunea iniţială a curbei reprezintă "mortalitatea infantilă", atribuibilă defectelor componente şi imperfecţiunilor de fabricare. Porţiunea relativ constantă a curbei $h(t)$ reprezintă perioada "în uz", în care defecţiunea este întâmplătoare. Defecţiunea din uzură de lângă sfârşitul vieţii componentului este arătată ca porţiunea crescătoare a curbei $h(t)$. Fiabilitatea sistemului poate fi

optimizată prin testarea iniţială a defectelor, înainte punerea în funcţiune, până la timpul t_1, pentru a evita defectarea prematură, şi prin înlocuirea parţială la timpul t_2 pentru a evita uzura.

Arătăm că legea de defectare exponenţială este potrivită în timpul perioadei "în uz" a vieţii normale a unui sistem. Înlocuind

$f_T(t) = \lambda e^{-\lambda t}$

şi

$F_T(t) = 1 - e^{-\lambda t}$

în relaţia (5.23) avem

$$h(t) = \lambda.$$

Observăm că parametrul λ din repartiţia exponenţială joacă rolul unei rate de defectare constante.

Am văzut că repartiţia gamma este potrivită pentru a descrie timpul pentru un total de η sosiri Poisson. În contextul legilor de defectare, repartiţia gamma poate fi gândită ca o generalizare a legii de defectare exponenţială pentru sisteme ce se defectează imediat de η evenimente eşuează, presupunând că evenimentele au loc în concordanţă cu legea Poisson. Astfel, repartiţia gamma este potrivită ca model al timpului până la defectare pentru sisteme care au o unitate care funcţionează şi $\eta - 1$ unităţi în standby; aceste unităţi intră în funcţionare pe rând, pe măsură ce se defectează celelalte şi fiecare are o repartiţie exponenţială a timpului până la defectare.

5.4.2 Repartiţia chi-pătrat

Alt caz special important al repartiţiei gamma este repartiţia chi-pătrat (χ^2), obţinută punând $\lambda = \frac{1}{2}$ şi $\eta = \frac{n}{2}$ în relaţia (5.17), unde $n \in \mathbb{N}^*$. Repartiţia χ^2 conţine astfel un parametru n şi are densitatea de forma

$$f_X(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, \text{ pentru } x > 0; \\ 0, \text{ altfel.} \end{cases} \tag{5.24}$$

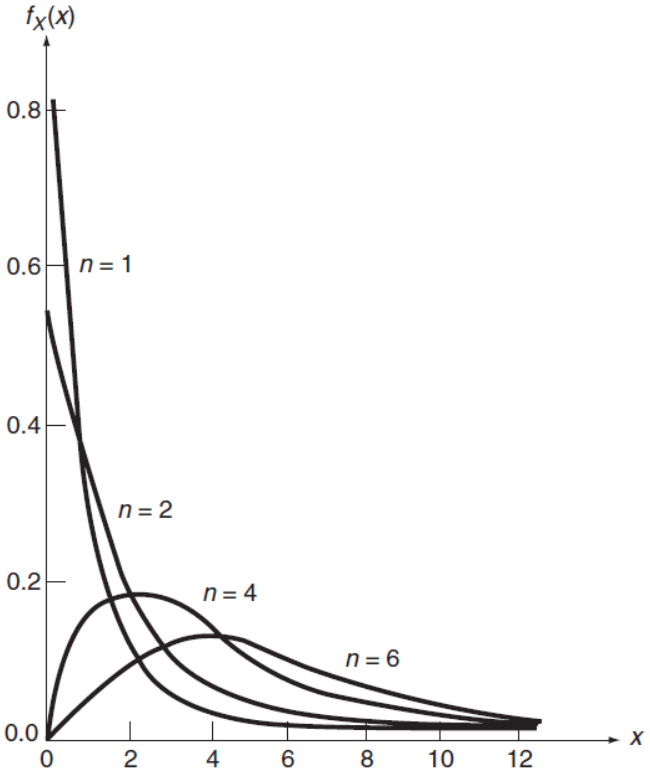
Parametrul n este numărul de *grade de libertate*. Utilitatea acestei repartiţii provine din faptul că o sumă de pătrate de n variabile aleatoare normale standardizate are o repartiţie χ^2 cu n grade de libertate; adică, dacă $U_1, U_2, ..., U_n$ sunt independente şi repartizate $N(0,1)$, atunci suma

$$X = U_1^2 + U_2^2 + ... + U_n^2 \tag{5.25}$$

are o repartiţie χ^2 cu n grade de libertate.

Datorită acestei relaţii, repartiţia χ^2 este una dintre principalele unelte în inferenţa statistică şi testarea ipotezelor.

Densitatea din relaţia (5.24) este reprezentată în figura următoare pentru câteva valori ale lui n.



Se observă că, când n creşte, forma lui $f_X(x)$ devine mai simetrică. Din relaţia (5.25), deoarece X poate fi exprimată ca o sumă de variabile aleatoare identic repartizate, ne asteptăm ca repartiţia χ^2 să tindă la o repartiţie normală când $n \to \infty$ pe baza teoremei limită centrală.

Media şi dispersia unei variabile aleatoare X având o repartiţie χ^2 cu n grade de libertate se obţin din relaţia (5.19):

$$m_X = n, \qquad \sigma_X^2 = 2n.$$

5.5 Repartiţia beta şi repartiţii în legătură cu aceasta

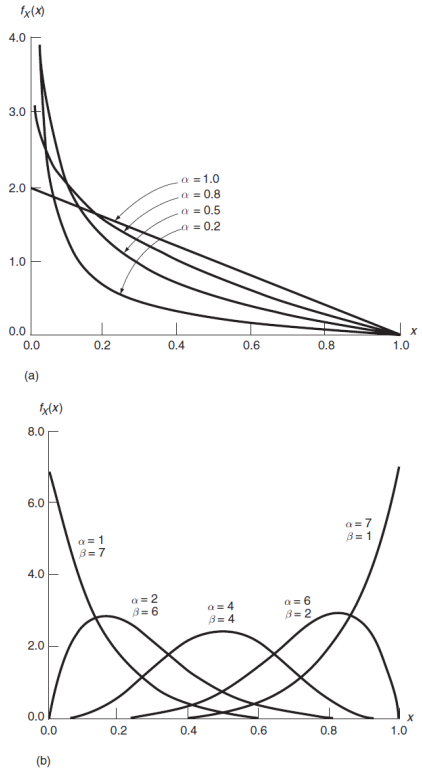
Repartiţia beta este caracterizată de densitatea

$$f_X(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \text{ pentru } 0 < x < 1, \\ 0, \text{ altfel,} \end{cases} \tag{5.26}$$

unde parametrii α şi β sunt pozitivi. Numele repartiţiei vine de la funcţia beta definită prin

$$\mathbf{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Ambii parametri α şi β dau forma repartiţiei; diferite combinaţii ale valorilor lor permit densităţii să ia o largă varietate de forme. Când $\alpha, \beta > 1$, repartiţia este unimodală, cu vârful în $x = \frac{\alpha-1}{\alpha+\beta-2}$. Devine în formă de U când $\alpha, \beta < 1$; este în formă de J când $\alpha \geq 1$ şi $\beta < 1$; ia forma unui J întors când $\alpha < 1$ şi $\beta \geq 1$. În sfârşit, ca un caz special, repartiţia uniformă pe intervalul $(0, 1)$ rezultă când $\alpha = \beta = 1$. Unele din aceste forme posibile sunt în figurile (a), pentru $\beta = 2$, şi (b) următoare.



Media şi dispersia unei variabile aleatoare beta repartizate X sunt

$$\begin{aligned} m_X &= \frac{\alpha}{\alpha+\beta}, \\ \sigma_X^2 &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{aligned} \tag{5.27}$$

Datorită versatilităţii ei ca o repartiţie pe un interval finit, repartiţia beta este folosită pentru a reprezenta un mare număr de cantităţi fizice pentru care valorile sunt restricţionate la un interval identificabil. Câteva din ariile de aplicare sunt limitele de toleranţă, controlul calităţii şi fiabilitatea.

Presupunem că un fenomen aleator Y poate fi observat independent de n ori şi după ce aceste n observaţii independente sunt ordonate crescător, fie $y_1 \leq y_2 \leq ... \leq y_n$ valorile lor. Dacă variabila aleatoare X este folosită pentru a nota proporţia din Y care ia valori între y_r şi y_{n-s+1}, se poate arăta că X are o repartiţie beta cu $\alpha = n - r - s + 1$ şi $\beta = r + s$, adică

$$f_X(x) = \begin{cases} \frac{\Gamma(n+1)}{\Gamma(n-r-s+1)\Gamma(r+s)} x^{n-r-s}(1-x)^{r+s-1}, \text{ pentru } 0 < x < 1, \\ 0, \text{ altfel.} \end{cases}$$

5.5.1 Tabele de probabilităţi

Funcţia de repartiţie asociată repartiţiei beta este

$$F_X(x) = \begin{cases} 0, \text{ pentru } x \leq 0, \\ \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^x u^{\alpha-1}(1-u)^{\beta-1} du, \text{ pentru } 0 < x < 1, \\ 1, \text{ pentru } x \geq 1. \end{cases} \tag{5.28}$$

Are de asemenea forma unei funcţii beta incomplete pentru care valorile pentru valori date ale lui α şi β pot fi găsite din tabele matematice. Funcţia beta incompletă e notată de obicei cu $\mathbf{I}_x(\alpha, \beta)$. Notând $F_X(x)$ cu parametrii α şi β cu $F(x; \alpha, \beta)$, dacă $\alpha \geq \beta$, atunci

$$F(x; \alpha, \beta) = \mathbf{I}_x(\alpha, \beta).$$

Dacă $\alpha < \beta$, atunci

$$F(x; \alpha, \beta) = 1 - \mathbf{I}_{(1-x)}(\beta, \alpha).$$

O altă metodă de evaluare a lui $F_X(x)$ din relaţia (5.28) este de a observa similaritatea în formă dintre $f_X(x)$ şi $p_Y(k)$ a unei variabile aleatoare binomiale Y, pentru cazul când $\alpha, \beta \in \mathbb{N}^*$. Din relaţia (4.1),

$$p_Y(k) = \frac{n!}{k!(n-k)!} p^k q^{n-k}, \qquad k = 0, 1, 2, ..., n. \tag{5.29}$$

Din relaţia (5.26), pentru cazul când $\alpha, \beta \in \mathbb{N}^*$, obţinem

$$f_X(x) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!} x^{\alpha-1}(1-x)^{\beta-1}, \qquad \alpha, \beta = 1, 2, ..., \qquad 0 < x < 1,$$

şi stabilim relaţia

$$f_X(x) = (\alpha+\beta-1) p_Y(k), \qquad \alpha, \beta = 1, 2, ..., \qquad 0 < x < 1,$$

unde $p_Y(k)$ este evaluată în $k = \alpha - 1$, cu $n = \alpha + \beta - 2$ şi $p = x$. De exemplu, valoarea $f_X(0,5)$ cu $\alpha = 2$ şi $\beta = 1$ este egală cu $2p_Y(1)$ cu $n = 1$ şi $p = 0,5$; aici $p_Y(1) = 0,5$ din relaţia (5.29), deci $f_X(0,5) = 1$.

Similar avem

$$F_X(x) = 1 - F_Y(k), \qquad \alpha, \beta = 1, 2, ..., \qquad 0 < x < 1,$$

cu $k = \alpha - 1$, $n = \alpha + \beta - 2$ şi $p = x$. Funcţia de repartiţie F_Y a unei variabile aleatoare binomiale Y este de asemenea tabelată şi poate fi folosită pentru a avantaja aici evaluarea lui $F_X(x)$ asociată cu repartiţia beta.

5.5.2 Repartiţia beta generalizată

Repartiţia beta poate fi generalizată de la una restricţionată la intervalul $(0,1)$ la una acoperind un interval arbitrar (a,b). Fie Y o variabilă aleatoare beta generalizată. Avem

$$Y = (b-a)\,X + a, \tag{5.30}$$

unde X este beta repartizată în conformitate cu relaţia (5.26). Deoarece relaţia (5.30) este o transformare monotonă de la X la Y, avem

$$f_Y(y) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{(b-a)^{\alpha+\beta-1}\Gamma(\alpha)\Gamma(\beta)}\,(y-a)^{\alpha-1}\,(b-y)^{\beta-1}, \text{ pentru } a<y<b, \\ 0, \text{ altfel.} \end{cases} \tag{5.31}$$

5.6 Repartiţii de valori extreme

Cineva preocupat de siguranţa unei structuri este adesea interesat de încărcătura *maximă* şi stresul *maxim* în membrii structurii. În studiile de fiabilitate, repartiţia vieţii unui sistem având n componente în serie (când sistemul se defectează dacă *o* componentă se defectează) este o funcţie de minimul timpilor până la defectarea componentelor, în timp ce pentru un sistem cu aranjament paralel (unde sistemul se defectează când *toate* componentele se defectează) e determinată de repartiţia maximului timpilor până la defectarea componentelor. Aceste exemple punctează preocuparea cu repartiţiile maximului sau minimului valorilor unui număr de variabile aleatoare.

Fie X_j, $j=1,2,...,n$, variabile aleatoare independente şi identic repartizate cu funcţia de repartiţie $F_X(x)$ şi densitatea $f_X(x)$ sau masa $p_X(x)$ şi

$$\begin{aligned} Y_n &= \max(X_1, X_2, ..., X_n), \\ Z_n &= \min(X_1, X_2, ..., X_n). \end{aligned}$$

Funcţia de repartiţie a lui Y_n este

$F_{Y_n}(y) = P(Y_n \le y) = P(X_1 \le y \cap X_2 \le y \cap ... \cap X_n \le y) \overset{\text{indep.}}{=} P(X_1 \le y)\,P(X_2 \le y)\,...P(X_n \le y) = F_{X_1}(y)\,F_{X_2}(y)\,...F_{X_n}(y) = [F_X(y)]^n$.

Deci

$$F_{Y_n}(y) = [F_X(y)]^n$$

Dacă X_j sunt continue, densitatea lui Y_n este

$$f_{Y_n}(y) = \frac{dF_{Y_n}(y)}{dy} = n\,[F_X(y)]^{n-1}\,f_X(y).$$

Funcţia de repartiţie a lui Z_n este

$F_{Z_n}(z) = P(Z_n \le z) = P(X_1 \le z \cup X_2 \le z \cup ... \cup X_n \le z) = 1 - P(X_1 > z \cap X_2 > z \cap ... \cap X_n > z) \overset{\text{indep.}}{=} 1 - P(X_1 > z)\,P(X_2 > z)\,...P(X_n > z) = 1 - [1 - F_{X_1}(z)]\,[1 - F_{X_2}(z)]\,...\,[1 - F_{X_n}(z)] = 1 - [1 - F_X(z)]^n$.

Deci

$$F_{Z_n}(z) = 1 - [1 - F_X(z)]^n .$$

Dacă X_j sunt continue, densitatea lui Z_n este

$$f_{Z_n}(z) = n [1 - F_X(z)]^{n-1} f_X(z) .$$

5.6.1 Repartiţii asimptotice de tip I ale valorilor extreme

Repartiţia asimptotică de tip I a valorilor maxime este repartiţia limitei lui Y_n (când $n \to \infty$) dintr-o repartiţie iniţială $F_X(x)$ a cărei coadă dreaptă este nemărginită şi este de tip exponenţial. Pentru acest caz, putem exprima $F_X(x)$ în forma

$$F_X(x) = 1 - \exp[-g(x)], \tag{5.32}$$

unde $g(x)$ este o funcţie strict crescătoare de x. În această categorie intră, printre altele, repartiţiile normală, lognormală şi gamma.

Fie

$$\lim_{n\to\infty} Y_n = Y.$$

Teorema 5.6. Fie variabilele aleatoare $X_1, X_2, ..., X_n$ independente şi identic repartizate cu aceeaşi funcţie de repartiţie $F_X(x)$. Dacă $F_X(x)$ e de forma dată de relaţia (5.32), avem

$$F_Y(y) = \exp\{-\exp[-\alpha(y-u)]\}, \qquad -\infty < y < \infty,$$

unde $\alpha > 0$ şi u sunt 2 parametri ai repartiţiei.

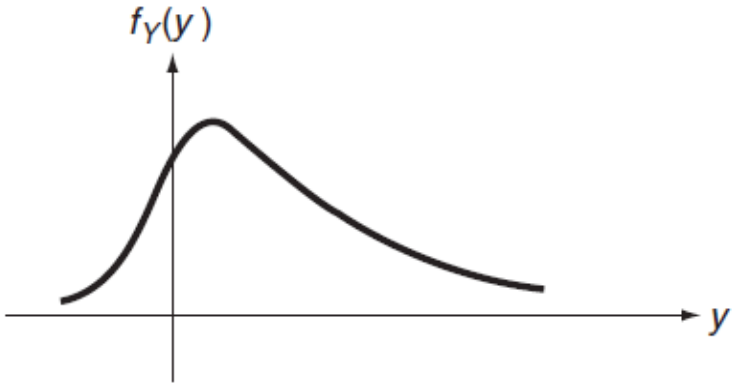
u este modulul repartiţiei, adică valoarea lui y în care $f_Y(y)$ este maximă. Media lui Y este

$$m_Y = u + \frac{\gamma}{\alpha},$$

unde $\gamma \simeq 0,577$ este constanta lui Euler. Dispersia lui Y este

$$\sigma_Y^2 = \frac{\pi^2}{6\alpha^2}.$$

u este parametru de locaţie şi α este parametru de scală al repartiţiei. Coeficientul de asimetrie este în acest caz $\gamma_1 \simeq 1,1396$ şi e independent de α şi u. Acest rezultat arată că repartiţia valorii maxime de tip I are o formă fixă cu o coadă dominantă spre dreapta. O formă tipică pentru $f_Y(y)$ este aratată în figura următoare.



Repartiţia asimptotică de tip I pentru valori minime este repartiţia limitei lui Z_n când $n \to \infty$ dintr-o repartiţie iniţială $F_X(x)$ a cărei coadă stângă este nemărginită şi este de tip exponenţial când descreşte la 0 spre stânga. Un exemplu de $F_X(x)$ care aparţine acestei clase este repartiţia normală.

Dacă punem

$$\lim_{n\to\infty} Z_n = Z,$$

funcţia de repartiţie a lui Z are forma

$$F_Z(z) = 1 - \exp\{-\exp[\alpha(z-u)]\}, \qquad -\infty < z < \infty,$$

unde $\alpha > 0$ şi u sunt din nou cei 2 parametri ai repartiţiei.

Repartiţiile asimptotice de tip I pentru valori maxime şi minime sunt imagini în oglindă una celeilalte. Modulul lui Z este u şi media, dispersia şi coeficientul lui de asimetrie sunt

$$\begin{aligned} m_Y &= u - \tfrac{\gamma}{\alpha}, \\ \sigma_Y^2 &= \tfrac{\pi^2}{6\alpha^2}, \\ \gamma_1 &\simeq -1,1396. \end{aligned}$$

Datorită largii aplicabilităţi, repartiţia valorii maxime de tip I este uneori numită simplu *repartiţia valorii extreme.*

5.6.2 Repartiţii asimptotice de tip II ale valorilor extreme

Repartiţia asimptotică de tip II a valorilor maxime apare ca repartiţia limitei lui Y_n când $n \to \infty$ dintr-o repartiţie iniţială de tip Pareto, adică, $F_X(x)$ e limitată la stânga în 0 şi coada ei dreaptă este nemărginită şi se apropie de 1 conform

$$F_X(x) = 1 - ax^{-k}, \qquad a, k, x > 0. \tag{5.33}$$

Pentru această clasă

$$F_Y(y) = \exp\left[-\left(\frac{y}{v}\right)^{-k}\right], \qquad v, k, y > 0.$$

Cu $F_X(x)$ dat în relaţia (5.33), fiecare X_j are momente doar până la ordinul r, unde r este cel mai mare întreg mai mic decât k. Când $k > 1$, media lui Y este

$$m_Y = v\Gamma\left(1 - \frac{1}{k}\right),$$

şi, când $k > 2$, dispersia are forma

$$\sigma_Y^2 = v^2\left[\Gamma\left(1 - \frac{2}{k}\right) - \Gamma^2\left(1 - \frac{1}{k}\right)\right].$$

Fie Y_I şi Y_{II} variabile aleatoare având repartiţii asimptotice de tip I, respectiv II ale valorilor maxime. Atunci ele sunt legate prin

$$F_{Y_{II}}(y) = F_{Y_I}(\ln y), \qquad y > 0,$$

iar parametrii α şi u din $F_{Y_I}(y)$ sunt legaţi de parametrii k şi v din $F_{Y_{II}}(y)$ prin

$$u = \ln v \text{ şi } \alpha = k.$$

Dacă Y_I şi Y_{II} sunt continue, densităţile lor sunt în relaţia

$$f_{Y_{II}}(y) = \frac{1}{y} f_{Y_I}(\ln y), \qquad y > 0.$$

Repartiţia asimptotică de tip II a valorilor minime apare în condiţii analoage. Cu $F_X(x)$ limitată la dreapta în 0 şi apropiindu-se de 0 la stânga într-o manieră analoagă relaţiei (5.33), avem

$$F_Z(z) = 1 - \exp\left[-\left|\frac{z}{v}\right|^{-k}\right], \qquad v, k > 0, \qquad z < 0.$$

Nu e aşa utilă ca omoloagele de tip I şi III, deoarece în practică repartiţiile iniţiale cerute nu sunt frecvent întâlnite.

5.6.3 Repartiţii asimptotice de tip III ale valorilor extreme

Deoarece repartiţia asimptotică de tip III a valorii maxime este de interes practic limitat, discutăm doar repartiţia valorilor minime.

Repartiţia asimptotică de tip III a valorii minime este repartiţia limitei lui Z_n când $n \to \infty$ pentru o repartiţie iniţială în care coada stângă creşte de la 0 lângă $x = \varepsilon$ în maniera

$$F_X(x) = c(x - \varepsilon)^k, \qquad c, k > 0, \quad x \geq \varepsilon.$$

Această clasă de repartiţii este mărginită la stânga în $x = \varepsilon$. Repartiţia gamma este o astfel de repartiţie cu $\varepsilon = 0$.

Se poate arăta că funcţia de repartiţie asimptotică de tip III a valorii minime este

$$F_Z(z) = 1 - \exp\left[-\left(\frac{z-\varepsilon}{w-\varepsilon}\right)^k\right], \qquad k > 0, \quad w, z > \varepsilon, \tag{5.34}$$

şi, dacă Z e continuă,

$$f_Z(z) = \frac{k}{w-\varepsilon}\left(\frac{z-\varepsilon}{w-\varepsilon}\right)^{k-1} \exp\left[-\left(\frac{z-\varepsilon}{w-\varepsilon}\right)^k\right], \qquad k > 0, \quad w, z > \varepsilon. \tag{5.35}$$

Media şi dispersia lui Z sunt

$$\begin{aligned} m_Z &= \varepsilon + (w-\varepsilon)\,\Gamma\left(1+\tfrac{1}{k}\right), \\ \sigma_Z^2 &= (w-\varepsilon)^2\left[\Gamma\left(1+\tfrac{2}{k}\right) - \Gamma^2\left(1+\tfrac{1}{k}\right)\right]. \end{aligned}$$

Am văzut că repartiţia exponenţială e folosită ca o lege a defectării în studii de fiabilitate, corespunzând unei funcţii hazard constante. Repartiţia dată de relaţiile (5.34) şi (5.35) este frecvent folosită ca un model generalizat al timpului până la defectare pentru cazurile în care funcţia hazard variază cu timpul. Se poate arăta că funcţia hazard

$$h(t) = \frac{k}{w}\left(\frac{t}{w}\right)^{k-1}, \quad t > 0,$$

poate avea o largă varietate de forme, şi densitatea asociată cu ea pentru T, timpul până la defectare, este dată de

$$f_T(t) = \frac{k}{w}\left(\frac{t}{w}\right)^{k-1} \exp\left[-\left(\frac{t}{w}\right)^k\right], \quad w, k, t > 0. \tag{5.36}$$

Aceasta este repartiţia Weibull, numită după Weibull care a descoperit-o în 1939. Relaţia (5.36) este un caz special al relaţiei (5.35), cu $\varepsilon = 0$.

Fie Z_I şi Z_{III} variabile aleatoare având repartiţii asimptotice de tipul I, respectiv III, ale valorilor minime. Atunci

$$F_{Z_{III}}(z) = F_{Z_I}\left[\ln(z-\varepsilon)\right], \quad z > \varepsilon,$$

cu $u = \ln(w-\varepsilon)$ şi $\alpha = k$. Dacă sunt continue,

$$f_{Z_{III}}(z) = \frac{1}{z-\varepsilon} f_{Z_I}\left[\ln(z-\varepsilon)\right], \quad z > \varepsilon.$$

Repartiţiile asimptotice ale valorilor maxime şi minime din aceeaşi repartiţie iniţială pot să nu fie de acelaşi tip. De exemplu, pentru o repartiţie iniţială gamma, repartiţia ei asimptotică a valorii maxime este de tipul I, în timp ce

a valorii minime este de tipul III. Un sistem având n componente în serie cu repartiţii de viaţă gamma independente pentru componentele lui are o repartiţie a timpului până la defectare aparţinând repartiţiei asimptotice de tip III a valorii minime, când n devine mare. Modelul corespunzător pentru un sistem având n componente în paralel este repartiţia asimptotică de tip I a valorii maxime.

6 Funcţii (transformări) de variabile aleatoare

În ştiinţă şi inginerie, multe fenomene sunt bazate pe relaţii funcţionale în care una sau mai multe variabile dependente sunt exprimate în termeni de una sau mai multe variabile independente. De exemplu, distanţa parcursă într-un interval de timp este o funcţie de viteză, ş.a.m.d.

6.1 Funcţii de o variabilă aleatoare

Fie

$$Y = g(X), \tag{6.1}$$

unde g este o funcţie continuă. Dată fiind repartiţia lui X în termeni de funcţia ei de repartiţie, funcţia masă de probabilitate sau densitate, suntem interesaţi de repartiţia corespunzătoare pentru Y şi proprietăţile momentelor lui Y.

6.1.1 Repartiţia

Dacă X este variabilă aleatoare, atunci Y, fiind o funcţie de X definită de relaţia (6.1), este de asemenea o variabilă aleatoare. Fie R_X, *mulţimea valorilor* lui X, şi R_Y mulţimea valorilor lui Y.

Pentru fiecare rezultat $X = x$, rezultă din relaţia (6.1) că $Y = y = g(x)$. Relaţia (6.1) defineşte şi o funcţie de la R_X la R_Y. Probabilităţile asociate cu fiecare punct (în cazul unei variabile aleatoare discrete X) sau fiecare regiune (în cazul unei variabile aleatoare continue X) din R_X sunt transportate în punctul sau regiunea corespunzătoare din R_Y. Repartiţia lui Y este determinată prin completarea acestui proces de transfer pentru fiecare punct (dacă X e discretă) sau fiecare regiune de tipul $Y \leq y$ (dacă X e continuă) din R_Y. Este posibil ca g să ducă mai multe puncte din R_X într-un singur punct din R_Y. Determinarea repartiţiei lui Y depinde astfel de forma lui g din relaţia (6.1).

Variabile aleatoare discrete Fie X variabilă aleatoare discretă cu funcţia masă de probabilitate p_X. Funcţia masă de probabilitate a lui Y este

$$p_Y(y) = \sum_{x \in g^{-1}(y)} p_X(x), \forall y \in R_Y,$$

unde $g^{-1}(y) = \{x \in R_X | g(x) = y\}$.

Exemplul 6.1. Fie $X \sim \begin{pmatrix} -1 & 0 & 1 & 2 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{pmatrix}$. Determinaţi repartiţia lui $Y = 2X^2 + 1$.

Valorile lui Y sunt $g(-1) = 2(-1)^2 + 1 = 3, g(0) = 1, g(1) = 3, g(2) = 9$. Avem $g^{-1}(1) = \{0\}, g^{-1}(3) = \{-1, 1\}, g^{-1}(9) = \{2\}$, deci $p_Y(1) = p_X(0) =$

$\frac{1}{4}, p_Y(3) = p_X(-1) + p_X(1) = \frac{1}{2} + \frac{1}{8} = \frac{5}{8}, p_Y(9) = p_X(2) = \frac{1}{8}$. De aici, $Y \sim \begin{pmatrix} 1 & 3 & 9 \\ \frac{1}{4} & \frac{5}{8} & \frac{1}{8} \end{pmatrix}$.

Dacă $g : R_X \to R_Y$ este injectivă, valorile posibile luate de X sunt $x_1, x_2, ...$, valorile luate de Y sunt $y_1 = g(x_1), y_2 = g(x_2), ...$ şi

$$p_X(x_i) = p_i, \quad i = 1, 2, ...,$$

atunci

$$p_Y(y_i) = p_Y(g(x_i)) = p_i, \quad i = 1, 2, ...$$

Exemplul 6.2. Fie $X \sim \begin{pmatrix} -1 & 0 & 1 & 2 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{pmatrix}$. Determinaţi repartiţia lui $Y = 2X + 1$.

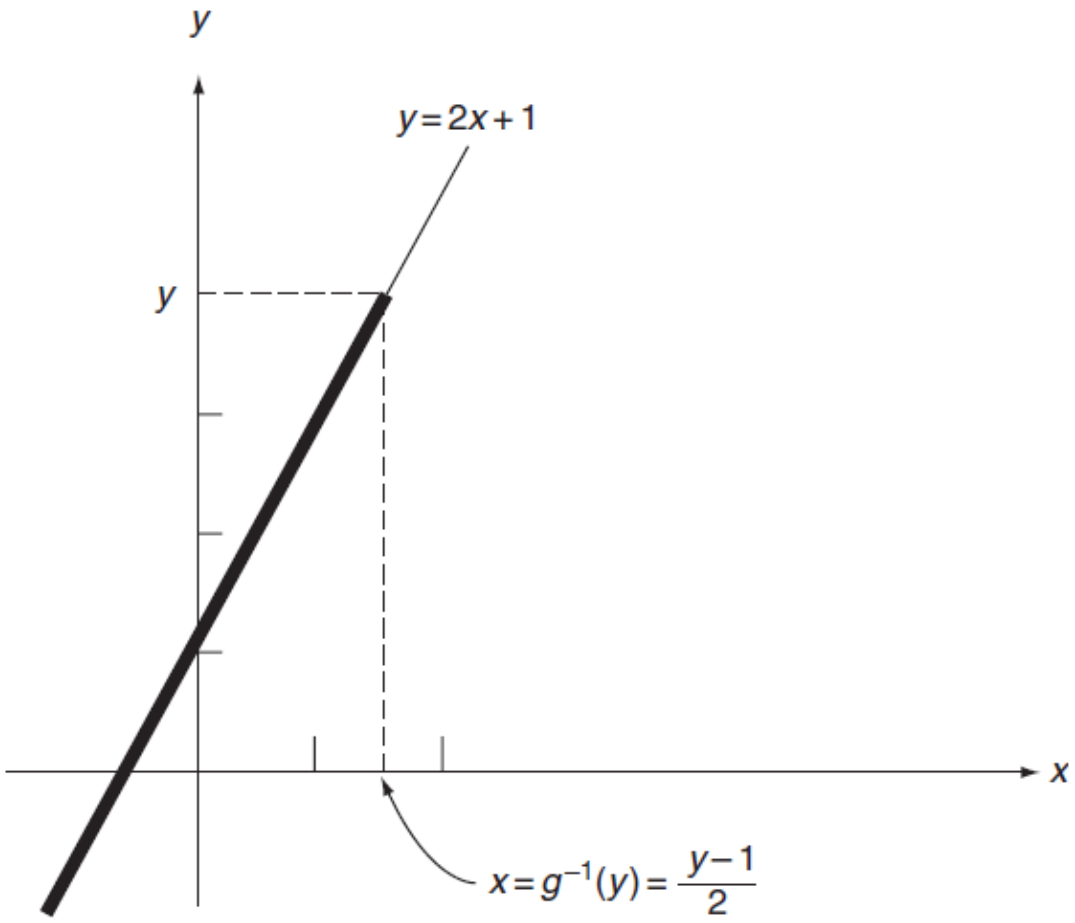
Valorile lui Y sunt $g(-1) = 2(-1) + 1 = -1, g(0) = 1, g(1) = 3, g(2) = 5$. Funcţia g este injectivă, deci $Y \sim \begin{pmatrix} -1 & 1 & 3 & 5 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{pmatrix}$.

Variabile aleatoare continue Considerăm cazul când X este o variabilă aleatoare continuă cu funcţie de repartiţie $F_X(x)$ sau densitate $f_X(x)$ cunoscute.

Începem considerând relaţia

$$Y = g(X) = 2X + 1. \tag{6.2}$$

Transformarea $y = g(x)$ este prezentată grafic în figura următoare.



Funcţia de repartiţie a lui Y este definită de

$$F_Y(y) = P(Y \leq y).$$

Regiunea definită de $Y \leq y$ din R_Y acoperă porţiunea îngroşată a curbei de transformare, după cum e arătat în figura alăturată, care, în R_X corespunde regiunii $g(X) \leq y$, sau $X \leq g^{-1}(y)$, unde

$$g^{-1}(y) = \frac{y-1}{2}$$

este funcţia inversă a lui g, sau soluţia x în funcţie de y a ecuaţiei (6.2). De aici,

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P\left(X \leq g^{-1}(y)\right) = F_X\left(g^{-1}(y)\right). \tag{6.3}$$

(6.3) este relaţia dintre funcţiile de repartiţie ale lui Y şi X.

Relaţia dintre densităţile lui Y şi X se obţine derivând în raport cu y în (6.3):

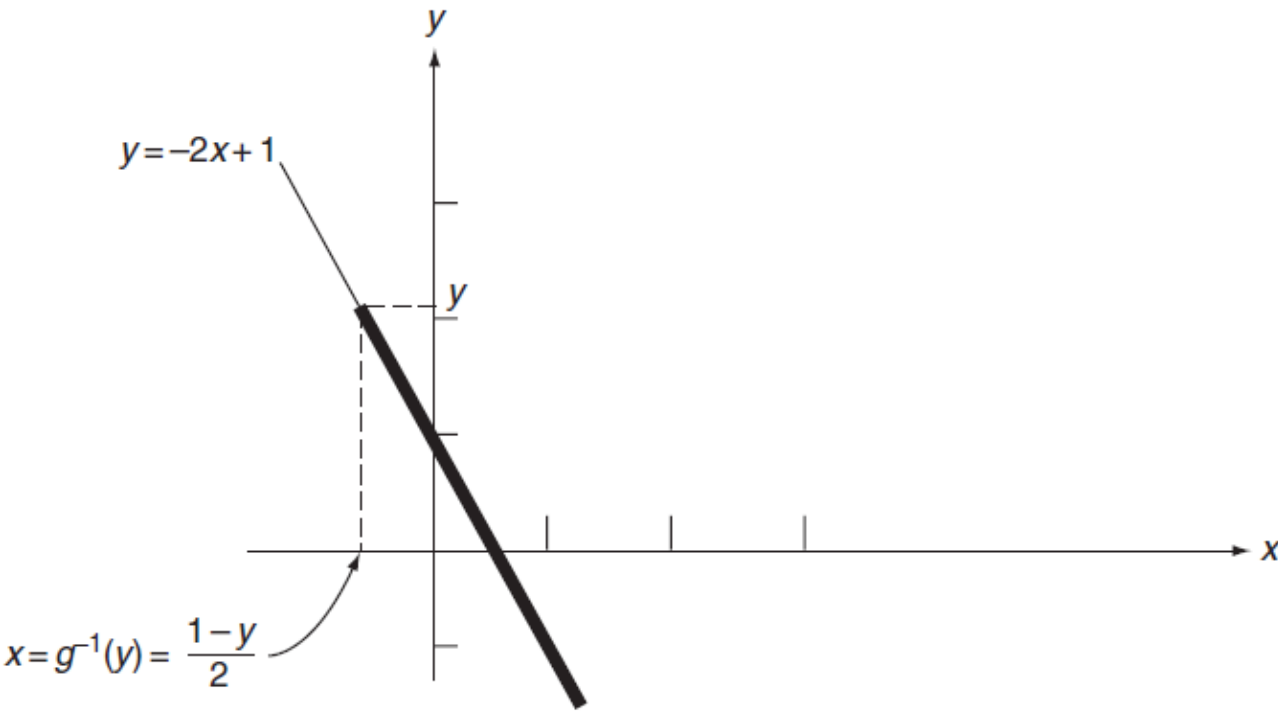
$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{d}{dy}\left[F_X\left(g^{-1}(y)\right)\right] = f_X\left(g^{-1}(y)\right)\frac{dg^{-1}(y)}{dy}. \tag{6.4}$$

Relaţiile (6.3) şi (6.4) au loc nu doar pentru transformarea particulară dată prin relaţia (6.2) dar şi pentru toate funcţiile continue g care sunt strict crescătoare.

Considerăm acum situaţia în care transformarea e dată de

$$Y = g(X) = -2X + 1. \tag{6.5}$$

Plecând din nou cu $F_Y(y) = P(Y \leq y)$ şi raţionând ca mai sus, regiunea $Y \leq y$ din R_Y corespunde regiunii $X \geq g^{-1}(y)$, după cum arată figura următoare.



Deoarece X este continuă, avem

$P\left(X \geq g^{-1}(y)\right) = P\left(X > g^{-1}(y) \cup X = g^{-1}(y)\right) = P\left(X > g^{-1}(y)\right) + P\left(X = g^{-1}(y)\right) = P\left(X > g^{-1}(y)\right) + 0 = P\left(X > g^{-1}(y)\right)$.

De aici, avem în acest caz

$$F_Y(y) = P(Y \leq y) = P\left(X > g^{-1}(y)\right) = 1 - P\left(X \leq g^{-1}(y)\right) = 1 - F_X\left(g^{-1}(y)\right). \tag{6.6}$$

În comparaţie cu relaţia (6.3), relaţia (6.6) dă o legătură diferită între funcţiile de repartiţie ale lui X şi Y, datorită unui g diferit.

Derivând în relaţia (6.6) în raport cu y obţinem relaţia dintre densităţile lui X şi Y:

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{d}{dy}\left[1 - F_X\left(g^{-1}(y)\right)\right] = -f_X\left(g^{-1}(y)\right)\frac{dg^{-1}(y)}{dy}. \tag{6.7}$$

Din nou observăm că relaţiile (6.6) şi (6.7) au loc pentru toate funcţiile continue g care sunt strict descrescătoare.

Deoarece derivata $\frac{dg^{-1}(y)}{dy}$ este totdeauna pozitivă în relaţia (6.4) (pentru că g e strict crescătoare) şi este totdeauna negativă în relaţia (6.7) (pentru că g e strict descrescătoare), rezultatele exprimate de aceste relaţii pot fi combinate în următoarea teoremă:

Teorema 6.1. Fie X o variabilă aleatoare continuă şi $Y = g(X)$, unde g este continuă şi strict monotonă. Atunci

$$f_Y(y) = f_X\left(g^{-1}(y)\right)\left|\frac{dg^{-1}(y)}{dy}\right|. \tag{6.8}$$

Exemplul 6.3. Densitatea lui X este dată de

$$f_X(x) = \frac{a}{\pi\left(x^2 + a^2\right)}, \quad -\infty < x < \infty,$$

unde $a > 0$ (repartiţia Cauchy). Determinaţi densitatea lui $Y = 2X + 1$.

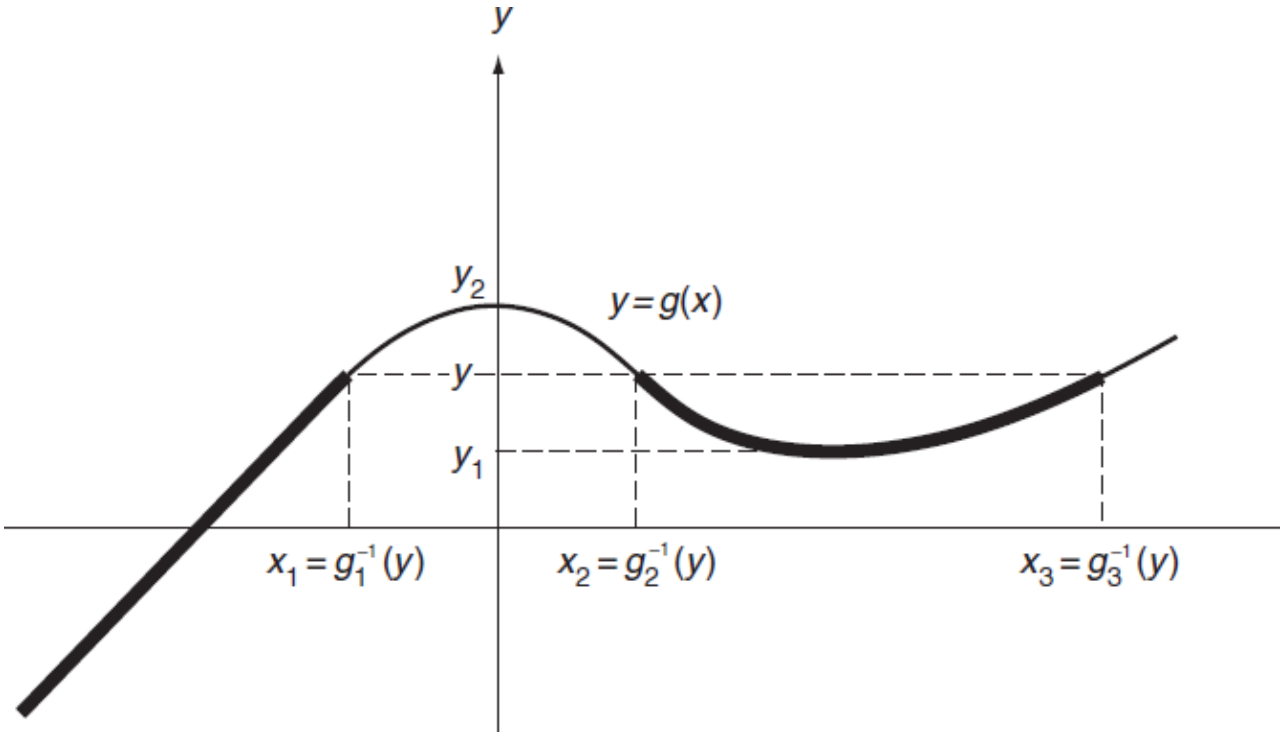
Deoarece $g(x) = 2x + 1$ este continuă şi strict monotonă, aplicăm relaţia (6.8).

$y = 2x + 1 \iff x = \frac{y-1}{2} \implies g^{-1}(y) = \frac{y-1}{2} \implies \frac{dg^{-1}(y)}{dy} = \frac{1}{2}$.

Din relaţia (6.8),

$f_Y(y) = f_X\left(\frac{y-1}{2}\right)\left|\frac{1}{2}\right| = \frac{a}{\pi\left[\left(\frac{y-1}{2}\right)^2 + a^2\right]} \cdot \frac{1}{2} = \frac{2a}{\pi\left[(y-1)^2 + 4a^2\right]}, \quad -\infty < y < \infty.$

Considerăm acum cazul când g nu este strict monotonă. În figura următoare, pentru $y < y_1$ şi $y > y_2$, ecuaţia $g(x) = y$ are soluţie unică şi relaţia (6.8) poate fi folosită pentru a determina densitatea lui Y în aceste intervale.



Pentru $y_1 \leq y \leq y_2$, trebuie să plecăm de la început şi considerăm $F_Y(y) = P(Y \leq y)$. Regiunea definită de $Y \leq y$ în R_Y acoperă porţiunile îngroşate ale curbei $y = g(x)$, după cum este arătat în figură. Astfel,

$$\begin{aligned} &F_Y(y) = P(Y \leq y) = P\left(X \leq g_1^{-1}(y)\right) + P\left(g_2^{-1}(y) < X \leq g_3^{-1}(y)\right) = \\ &P\left(X \leq g_1^{-1}(y)\right) + P\left(X \leq g_3^{-1}(y)\right) - P\left(X \leq g_2^{-1}(y)\right) = \\ &F_X\left(g_1^{-1}(y)\right) + F_X\left(g_3^{-1}(y)\right) - F_X\left(g_2^{-1}(y)\right), \quad y_1 < y < y_2, \end{aligned} \tag{6.9}$$

unde $x_1 = g_1^{-1}(y), x_2 = g_2^{-1}(y), x_3 = g_3^{-1}(y)$ sunt rădăcinile ecuaţiei $g(x) = y$ în termeni de y.

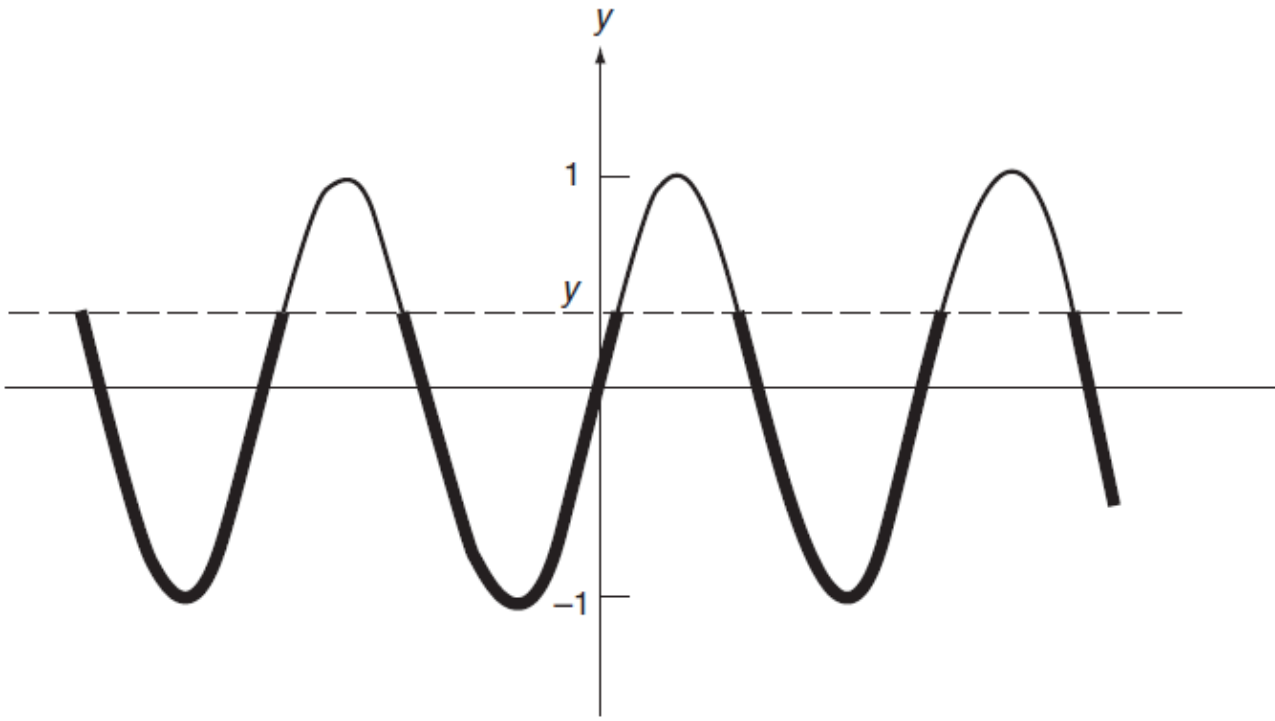
Ca mai sus, relaţia între densităţile lui X şi Y e găsită derivând în relaţia (6.9) în raport cu y:

$$f_Y(y) = f_X\left(g_1^{-1}(y)\right)\frac{dg_1^{-1}(y)}{dy} + f_X\left(g_3^{-1}(y)\right)\frac{dg_3^{-1}(y)}{dy} - f_X\left(g_2^{-1}(y)\right)\frac{dg_2^{-1}(y)}{dy}, \quad y_1 < y < y_2. \tag{6.10}$$

Deoarece derivata $\frac{dg_2^{-1}(y)}{dy}$ e negativă în timp ce celelalte sunt pozitive, relaţia (6.10) ia forma

$$f_Y(y) = \sum_{j=1}^{3} f_X\left(g_j^{-1}(y)\right)\left|\frac{dg_j^{-1}(y)}{dy}\right|, \quad y_1 < y < y_2. \tag{6.11}$$

Figura următoare reprezintă transformarea $y = \sin x$; această ecuaţie are o infinitate numărabilă de rădăcini, $x_1 = g_1^{-1}(y), x_2 = g_2^{-1}(y), ... \ \forall y \in [-1, 1]$.



Urmând procedura de mai sus se poate stabili pentru $F_Y(y)$ o relaţie similară cu (6.9) (dar cu o infinitate de termeni) şi, precum în relaţia (6.11), densitatea lui Y are acum forma

$$f_Y(y) = \sum_{j=1}^{\infty} f_X\left(g_j^{-1}(y)\right) \left| \frac{dg_j^{-1}(y)}{dy} \right|, \quad -1 < y < 1. \tag{6.12}$$

Se observă că $f_Y(y) = 0$ pentru $y \notin [-1, 1]$, deoarece $F_Y(y) = 0$, $\forall y \leq -1$ şi $F_Y(y) = 1, \forall y \geq 1$.

Relaţiile (6.11) şi (6.12) conduc la următorul rezultat:

Teorema 6.2. Fie X o variabilă aleatoare continuă şi $Y = g(X)$, unde g este continuă, şi $g(x) = y$ are un număr cel mult numărabil de rădăcini $x_1 = g_1^{-1}(y), x_2 = g_2^{-1}(y_2), ...$ Atunci:

$$f_Y(y) = \sum_{j=1}^{r} f_X\left(g_j^{-1}(y)\right) \left| \frac{dg_j^{-1}(y)}{dy} \right|, \tag{6.13}$$

unde r este numărul de rădăcini ale ecuaţiei $g(x) = y$.

Relaţia (6.8) este conţinută în această teoremă ca un caz special ($r = 1$).

Exemplul 6.4. Determinaţi densitatea lui $Y = X^2$, unde $X \sim N(0,1)$.

Avem

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad -\infty < x < \infty.$$

$Y \geq 0 \implies F_Y(y) = P(Y \leq y) = P(\emptyset) = 0, \forall y < 0 \implies f_Y(y) = F_Y'(y) = 0, \forall y < 0$. Pentru $y > 0$, rădăcinile lui $x^2 = y$ sunt

$$x_{1,2} = g_{1,2}^{-1}(y) = \pm\sqrt{y}.$$

De aici, folosind relaţia (6.13),

$f_Y(y) = \sum_{j=1}^{2} f_X\left(g_j^{-1}(y)\right) \left|\frac{dg_j^{-1}(y)}{dy}\right| = f_X\left(\sqrt{y}\right)\left|\frac{1}{2\sqrt{y}}\right| + f_X\left(-\sqrt{y}\right)\left|-\frac{1}{2\sqrt{y}}\right| = \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}}$,

deci

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi y}} e^{-\frac{y}{2}}, \text{ pentru } y > 0, \\ 0, \text{ altfel.} \end{cases}$$

Am obţinut repartiţia χ^2 cu un grad de libertate (vezi relaţia (5.24)), deoarece

$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty u^{-\frac{1}{2}} e^{-u} du \overset{u=\frac{x^2}{2}}{=} \sqrt{2}\int_0^\infty x^{-1} e^{-\frac{x^2}{2}} x dx = \sqrt{2}\int_0^\infty e^{-\frac{x^2}{2}} dx = \sqrt{2}\cdot\frac{1}{2}\int_{-\infty}^\infty e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2}}\sqrt{2\pi} = \sqrt{\pi}$.

6.1.2 Momente

Fie $Y = g(X)$ şi presupunem că toate momentele dorite ale lui Y există. Momentul de ordinul n al lui Y este

$$\begin{aligned} &E(Y^n) = E(g^n(X)) = \sum_i g^n(x_i) p_X(x_i), \text{ dacă } X \text{ e discretă,} \\ &E(Y^n) = E(g^n(X)) = \int_{-\infty}^{\infty} g^n(x) f_X(x) dx, \text{ dacă } X \text{ e continuă.} \end{aligned} \tag{6.14}$$

Exemplul 6.5. Fie $X \sim \begin{pmatrix} -1 & 0 & 1 & 2 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{pmatrix}$. Determinaţi media şi dispersia lui $Y = 2X + 1$.

Folosim prima relaţie (6.14).

$E(Y) = E(2X+1) = \sum_i (2x_i + 1) p_X(x_i) = (-1)\cdot\frac{1}{2} + 1\cdot\frac{1}{4} + 3\cdot\frac{1}{8} + 5\cdot\frac{1}{8} = \frac{3}{4}$.

$E\left(Y^2\right) = E\left((2X+1)^2\right) = \sum_i (2x_i+1)^2 p_X(x_i) = 1\cdot\frac{1}{2} + 1\cdot\frac{1}{4} + 9\cdot\frac{1}{8} + 25\cdot\frac{1}{8} = 5$.

$\sigma_Y^2 = E\left(Y^2\right) - E^2(Y) = 5 - \left(\frac{3}{4}\right)^2 = \frac{71}{16}$.

6.2 Funcţii de două sau mai multe variabile aleatoare

Considerăm transformarea

$$Y = g(X_1, X_2, ..., X_n), \tag{6.15}$$

unde $X_1, X_2, ..., X_n$ sunt variabile aleatoare a căror repartiţie comună se cunoaşte. Vrem să determinăm repartiţia lui Y.

Fie $\mathbf{X}$ vectorul aleator n-dimensional cu componentele $X_1, X_2, ..., X_n$ şi $\mathbf{x}$ vectorul n-dimensional cu componentele $x_1, x_2, ..., x_n$.

Dacă $X_1, X_2, ..., X_n$ sunt discrete cu masa comună $p_{\mathbf{X}}(\mathbf{x})$, atunci masa lui Y este

$$p_Y(y) = \sum_{\mathbf{x} \in g^{-1}(y)} p_{\mathbf{X}}(\mathbf{x}), \forall y \in R_Y,$$

unde $g^{-1}(y) = \{\mathbf{x} \in R_{\mathbf{X}} | g(\mathbf{x}) = y\}$, iar $R_{\mathbf{X}}$ este mulţimea valorilor lui $\mathbf{X}$.

Presupunem acum că $X_1, X_2, ..., X_n$ sunt continue şi cunoaştem densitatea comună $f_{\mathbf{X}}(\mathbf{x})$ sau funcţia de repartiţie comună $F_{\mathbf{X}}(\mathbf{x})$. Avem

$$F_Y(y) = P(Y \leq y) = P(g(\mathbf{X}) \leq y) = F_{\mathbf{X}}(\mathbf{x} : g(\mathbf{x}) \leq y). \tag{6.16}$$

Ultima expresie din relaţia (6.16) reprezintă funcţia de repartiţie comună a lui $\mathbf{X}$ pentru care argumentul $\mathbf{x}$ satisface $g(\mathbf{x}) \leq y$. În termeni de $f_{\mathbf{X}}(\mathbf{x})$ este dată de

$$F_{\mathbf{X}}(\mathbf{x} : g(\mathbf{x}) \leq y) = \int_{(R^n : g(\mathbf{x}) \leq y)} \cdots \int f_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x}, \tag{6.17}$$

unde $R^n = \{\mathbf{x} \in \mathbb{R}^n | g(\mathbf{x}) \leq y\}$. $F_Y(y)$ poate fi determinată evaluând integrala din relaţia (6.17).

6.2.1 Sume de variabile aleatoare

Considerăm suma

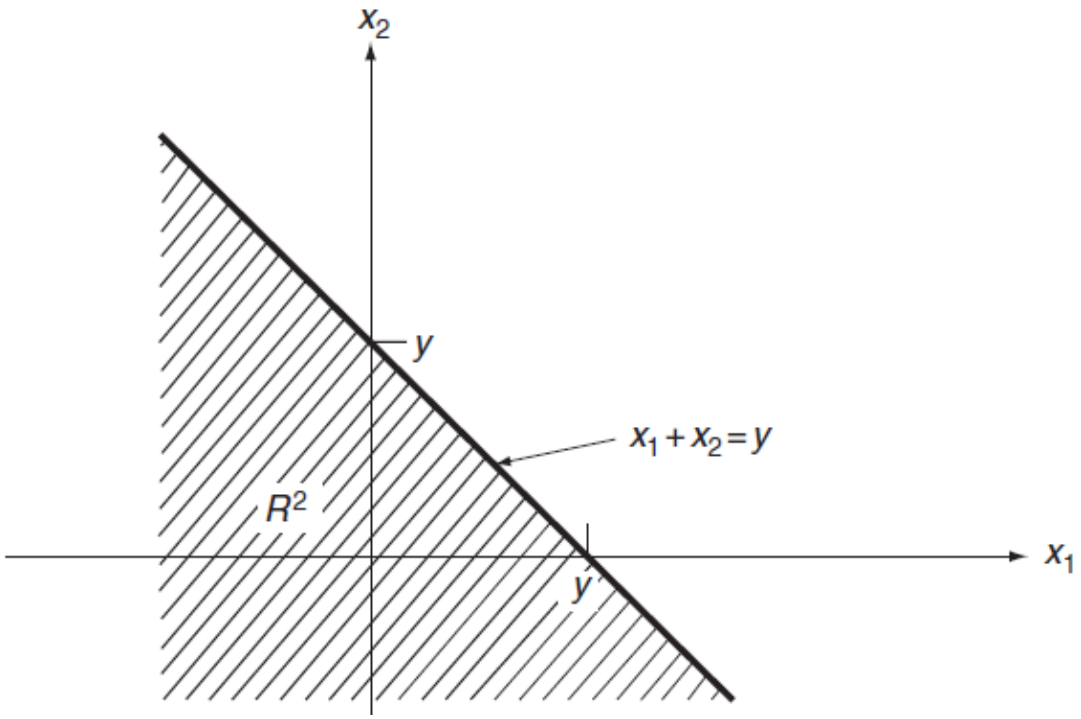
$$Y = g(X_1, X_2, ..., X_n) = X_1 + X_2 + ... + X_n.$$

E suficient să determinăm $f_Y(y)$ pentru $n = 2$. Rezulatatul pentru acest caz poate fi apoi aplicat succesiv pentru a da repartiţia sumei oricărui număr de variabile aleatoare. Pentru $Y = X_1 + X_2$, relaţiile (6.16) şi (6.17) dau

$$F_Y(y) = \iint_{(R^2 : x_1 + x_2 \leq y)} f_{X_1 X_2}(x_1, x_2) \, dx_1 dx_2,$$

şi, după cum se observă din figura următoare, în care este reprezentată $R^2 = \{\mathbf{x} \in \mathbb{R}^2 | x_1 + x_2 \leq y\}$, avem

$$F_Y(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{y-x_2} f_{X_1X_2}(x_1, x_2)\, dx_1 dx_2. \tag{6.18}$$



Derivând în relaţia (6.18) în raport cu y, obţinem

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1X_2}(y - x_2, x_2)\, dx_2. \tag{6.19}$$

Când X_1 şi X_2 sunt independente, relaţia (6.19) devine

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1}(y - x_2)\, f_{X_2}(x_2)\, dx_2. \tag{6.20}$$

Integrala din relaţia (6.20) se numeşte *convoluţia* funcţiilor $f_{X_1}(x_1)$ şi $f_{X_2}(x_2)$.

Teorema 6.3. Fie X_1 şi X_2 variabile aleatoare continue independente şi $Y = X_1 + X_2$. Atunci densitatea lui Y este convoluţia densităţilor lui X_1 şi X_2, adică

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1}(y - x_2)\, f_{X_2}(x_2)\, dx_2 = \int_{-\infty}^{\infty} f_{X_2}(y - x_1)\, f_{X_1}(x_1)\, dx_1. \tag{6.21}$$

6.3 m funcţii de n variabile aleatoare

Fie

$$Y_j = g_j(X_1, ..., X_n), \quad j = 1, 2, ..., m, \quad m \leq n. \tag{6.22}$$

Vrem să obţinem repartiţia comună a variabilelor aleatoare $Y_j, \quad j = 1, 2, ..., m$, ştiind repartiţia comună a variabilelor aleatoare $X_k, \quad k = 1, 2, ..., n$.

Dacă $X_1, X_2, ..., X_n$ sunt discrete cu masa comună $p_{\mathbf{X}}(\mathbf{x})$, atunci masa comună a variabilelor aleatoare $Y_1, ..., Y_m$ este

$$p_{\mathbf{Y}}(\mathbf{y}) = \sum_{\mathbf{x} \in \mathbf{g}^{-1}(\mathbf{y})} p_{\mathbf{X}}(\mathbf{x}), \forall \mathbf{y} \in R_{\mathbf{Y}},$$

unde $\mathbf{g}$ are componentele $g_1, g_2, ..., g_m$, $\mathbf{g}^{-1}(\mathbf{y}) = \{\mathbf{x} \in R_{\mathbf{X}} | \mathbf{g}(\mathbf{x}) = \mathbf{y}\}$, iar $R_{\mathbf{X}}$ şi $R_{\mathbf{Y}}$ sunt mulţimile valorilor lui $\mathbf{X}$, respectiv $\mathbf{Y}$.

Presupunem acum că $X_1, X_2, ..., X_n$ sunt continue.

Considerăm mai întâi cazul $m = n$.

Fie $\mathbf{X}$ şi $\mathbf{Y}$ vectori aleatori n-dimensionali cu componentele $X_1, ..., X_n$, respectiv $Y_1, ..., Y_n$. Relaţia (6.22) se scrie vectorial

$$\mathbf{Y} = \mathbf{g}(\mathbf{X}), \tag{6.23}$$

unde $\mathbf{g}(\mathbf{X})$ are componentele $g_1(\mathbf{X}), g_2(\mathbf{X}), ..., g_n(\mathbf{X})$. Considerăm întâi cazul în care funcţiile g_j din $\mathbf{g}$ sunt continue în raport cu fiecare argument, au derivate parţiale continue şi sunt bijective. Atunci rezultă că funcţiile inverse g_j^{-1} din $\mathbf{g}^{-1}$ definită de

$$\mathbf{X} = \mathbf{g}^{-1}(\mathbf{Y}),$$

există şi sunt unice. Au de asemenea derivate parţiale continue.

Pentru a determina $f_{\mathbf{Y}}(\mathbf{y})$ în termeni de $f_{\mathbf{X}}(\mathbf{x})$, observăm că, dacă o regiune închisă $R_{\mathbf{X}}^n$ din mulţimea valorilor lui $\mathbf{X}$ este dusă într-o regiune închisă $R_{\mathbf{Y}}^n$ din mulţimea valorilor lui $\mathbf{Y}$ prin transformerea $\mathbf{g}$, conservarea probabilităţii dă

$$\int_{R_{\mathbf{Y}}^n} \cdots \int f_{\mathbf{Y}}(\mathbf{y}) \, d\mathbf{y} = \int_{R_{\mathbf{X}}^n} \cdots \int f_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x}, \tag{6.24}$$

unde integralele reprezintă n integrale în raport cu componentele lui $\mathbf{x}$, respectiv $\mathbf{y}$. Urmând regula uzuală de schimbare de variabile în integrale multiple, putem scrie

$$\int_{R_{\mathbf{X}}^n} \cdots \int f_{\mathbf{X}}(\mathbf{x}) \, d\mathbf{x} = \int_{R_{\mathbf{Y}}^n} \cdots \int f_{\mathbf{X}}\left(\mathbf{g}^{-1}(\mathbf{y})\right) |J| \, d\mathbf{y}, \tag{6.25}$$

unde J este Jacobianul transformării, definit ca determinantul

$$J=\begin{vmatrix} \frac{\partial g_1^{-1}}{\partial y_1} & \frac{\partial g_1^{-1}}{\partial y_2} & \cdots & \frac{\partial g_1^{-1}}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial g_n^{-1}}{\partial y_1} & \frac{\partial g_n^{-1}}{\partial y_2} & \cdots & \frac{\partial g_n^{-1}}{\partial y_n} \end{vmatrix}. \tag{6.26}$$

Relaţiile (6.24) şi (6.25) conduc la următorul rezultat.

Teorema 6.4. Pentru transformarea dată de relaţia (6.23), unde $\mathbf{X}$ este un vector aleator continuu şi $\mathbf{g}$ este continuă cu derivate parţiale continue şi bijectivă, densitatea comună a lui $\mathbf{Y}$ este

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}\left(\mathbf{g}^{-1}(\mathbf{y})\right)|J|, \tag{6.27}$$

unde J este definit de relaţia (6.26).

Exemplul 6.6. Fie X_1 şi X_2 variabile aleatoare independente identic repartizate $N(0,1)$. Să se determine densitatea comună a lui $Y_1 = X_1 + X_2$ şi $Y_2 = X_1 - X_2$.

Deoarece sistemul $\begin{cases} x_1 + x_2 = y_1 \\ x_1 - x_2 = y_2 \end{cases}$ are soluţia unică

$$x_1 = g_1^{-1}(y) = \frac{y_1+y_2}{2}, \quad x_2 = g_2^{-1}(y) = \frac{y_1-y_2}{2},$$

$\mathbf{g}$ este bijectivă şi se poate aplica teorema 6.4. Jacobianul este

$$J=\begin{vmatrix} \frac{\partial g_1^{-1}}{\partial y_1} & \frac{\partial g_1^{-1}}{\partial y_2} \\ \frac{\partial g_2^{-1}}{\partial y_1} & \frac{\partial g_2^{-1}}{\partial y_2} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}.$$

Datorită independenţei, relaţia (6.27) conduce la

$f_{Y_1Y_2}(y_1,y_2) = f_{X_1}\left(g_1^{-1}(y)\right) f_{X_2}\left(g_2^{-1}(y)\right)|J| = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{\left(\frac{y_1+y_2}{2}\right)^2}{2}\right] \cdot \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{\left(\frac{y_1-y_2}{2}\right)^2}{2}\right] \cdot \left|-\frac{1}{2}\right| = \frac{1}{4\pi} \exp\left[-\frac{(y_1+y_2)^2}{8}\right] \exp\left[-\frac{(y_1-y_2)^2}{8}\right] = \frac{1}{4\pi} \exp\left(-\frac{y_1^2+y_2^2}{4}\right), \quad y_1, y_2 \in \mathbb{R}.$

Observăm că rezultatul poate fi scris sub forma

$$f_{Y_1Y_2}(y_1,y_2) = f_{Y_1}(y_1) f_{Y_2}(y_2),$$

unde

$$f_{Y_1}(y_1) = \frac{1}{\sqrt{4\pi}} \exp\left(-\frac{y_1^2}{4}\right), \quad y_1 \in \mathbb{R},$$

$$f_{Y_2}(y_2) = \frac{1}{\sqrt{4\pi}} \exp\left(-\frac{y_2^2}{4}\right), \quad y_2 \in \mathbb{R},$$

implicând că, deşi Y_1 şi Y_2 sunt ambele funcţii de X_1 şi X_2, ele sunt independente şi identic repartizate $N(0,2)$.

Relaţia (6.27) este o extindere a relaţiei (6.8), care este pentru cazul special $n = 1$. Similar, o extindere este de asemenea posibilă pentru relaţia (6.13) pentru cazul $n = 1$ când ecuaţia are mai mult de o rădăcină.

Teorema 6.5. În transformarea dată de relaţia (6.23), presupunem că ecuaţia $\mathbf{g}(\mathbf{x}) = \mathbf{y}$ are un număr cel mult numărabil de rădăcini $\mathbf{x}_1 = \mathbf{g}_1^{-1}(\mathbf{y}), \mathbf{x}_2 = \mathbf{g}_2^{-1}(\mathbf{y}), ...$ Atunci

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{j=1}^{r} f_{\mathbf{X}}\left(\mathbf{g}_j^{-1}(\mathbf{y})\right) |J_j|, \tag{6.28}$$

unde r este numărul de rădăcini ale ecuaţiei $\mathbf{g}(\mathbf{x}) = \mathbf{y}$ şi

$$J_j = \begin{vmatrix} \frac{\partial g_{j1}^{-1}}{\partial y_1} & \frac{\partial g_{j1}^{-1}}{\partial y_2} & \cdots & \frac{\partial g_{j1}^{-1}}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial g_{jn}^{-1}}{\partial y_1} & \frac{\partial g_{jn}^{-1}}{\partial y_2} & \cdots & \frac{\partial g_{jn}^{-1}}{\partial y_n} \end{vmatrix}.$$

Aici $g_{j1}^{-1}, g_{j2}^{-1}, ..., g_{jn}^{-1}$ sunt componentele lui $\mathbf{g}_j^{-1}$.

Rezultatele prezentate mai sus pot fi de asemenea aplicate în cazul când dimensiunea lui $\mathbf{Y}$ este mai mică decât cea a lui $\mathbf{X}$. Considerăm transformarea (6.22) în care $m < n$. Pentru a folosi formulele de mai sus, întâi completăm vectorul aleator m-dimensional $\mathbf{Y}$ cu un alt vector aleator $(n-m)$-dimensional $\mathbf{Z}$. Vectorul $\mathbf{Z}$ poate fi construit ca o funcţie simplă de $\mathbf{X}$, în forma

$$\mathbf{Z} = \mathbf{h}(\mathbf{X}), \tag{6.29}$$

unde $\mathbf{h}$ este continuă şi are derivate parţiale continue. Combinând relaţiile (6.22) şi (6.29), avem o transformare de la n variabile aleatoare la n variabile aleatoare, şi densitatea comună a lui $\mathbf{Y}$ şi $\mathbf{Z}$ poate fi obţinută cu relaţia (6.27) sau relaţia (6.28). Densitatea comună a lui $\mathbf{Y}$ singur se găseşte apoi prin integrare în raport cu componentele lui $\mathbf{Z}$.

Part II
Statistică

7 Statistică descriptivă pentru date unudimensionale (reprezentări grafice, funcţie de repartiţie de selecţie, momente de selecţie)

7.1 Histograme şi diagrame de frecvenţe

Fiind dat un set de observaţii independente $x_1, x_2, ..., x_n$, ale unei variabile aleatoare X, un prim pas util este organizarea şi prezentarea lor adecvată astfel încât ele să poată fi uşor inerpretate şi evaluate. Când există un număr mare de date observate, o *histogramă* este o reprezentare grafică excelentă a datelor, facilitând

a) o evaluare a adecvării modelului presupus,

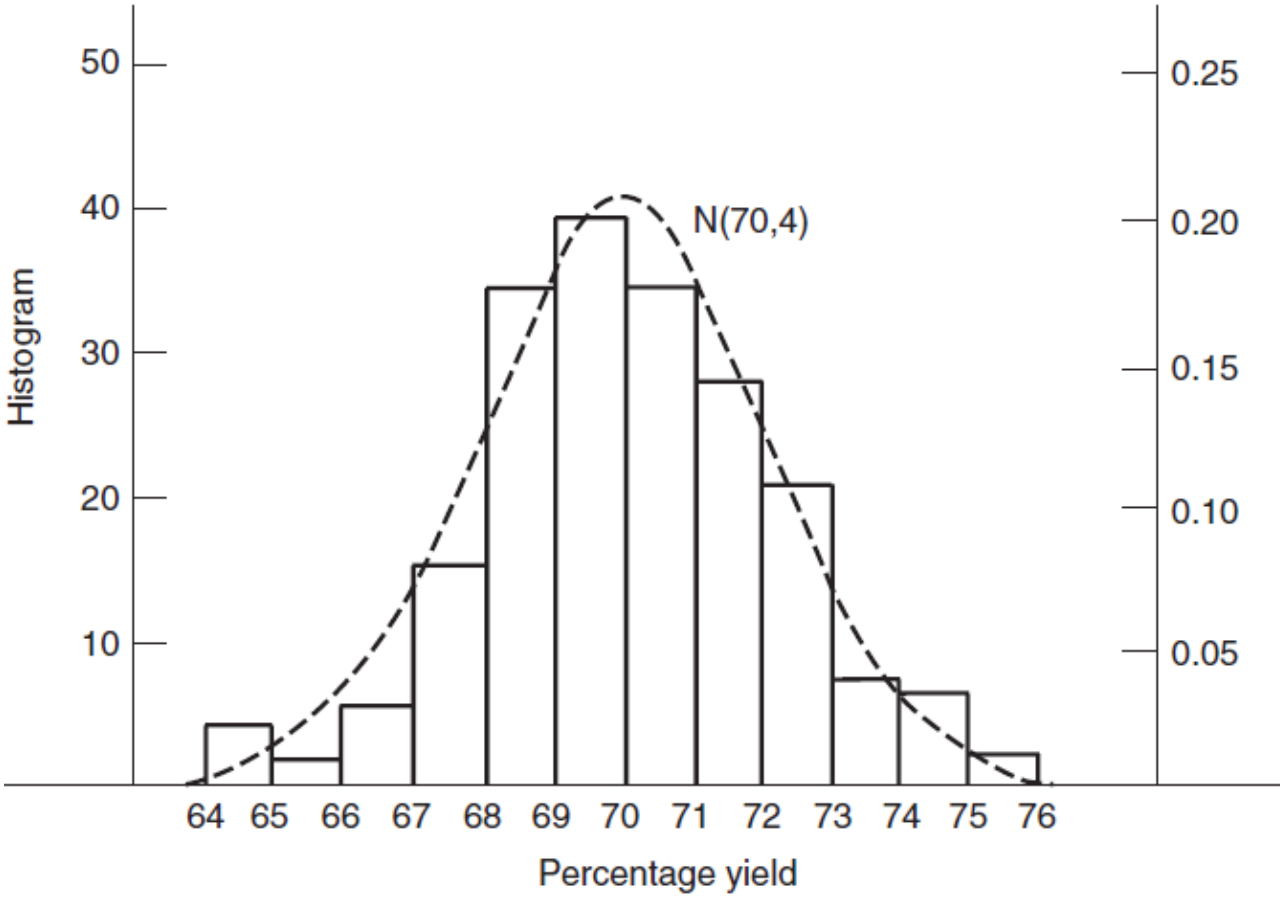
b) estimarea percentilelor repartiţiei,

c) estimarea parametrilor repartiţiei.

Considerăm, de exemplu, un proces chimic care este producerea de loturi (batches) dintr-un material; 200 de valori observate ale randamentului procentual (percentage yield), X, reprezentând un eşantion relativ mare sunt date în tabelul următor.

Batch no.	Yield (%)	Batch no.	Yield (%)	Batch no.	Yield (%)	Batch no.	Yield (%)	Batch no.	Yield (%)
1	68.4	41	68.7	81	68.5	121	73.3	161	70.5
2	69.1	42	69.1	82	71.4	122	75.8	162	68.8
3	71.0	43	69.3	83	68.9	123	70.4	163	72.9
4	69.3	44	69.4	84	67.6	124	69.0	164	69.0
5	72.9	45	71.1	85	72.2	125	72.2	165	68.1
6	72.5	46	69.4	86	69.0	126	69.8	166	67.7
7	71.1	47	75.6	87	69.4	127	68.3	167	67.1
8	68.6	48	70.1	88	73.0	128	68.4	168	68.1
9	70.6	49	69.0	89	71.9	129	70.0	169	71.7
10	70.9	50	71.8	90	70.7	130	70.9	170	69.0
11	68.7	51	70.1	91	67.0	131	72.6	171	72.0
12	69.5	52	64.7	92	71.1	132	70.1	172	71.5
13	72.6	53	68.2	93	71.8	133	68.9	173	74.9
14	70.5	54	71.3	94	67.3	134	64.6	174	78.7
15	68.5	55	71.6	95	71.9	135	72.5	175	69.0
16	71.0	56	70.1	96	70.3	136	73.5	176	70.8
17	74.4	57	71.8	97	70.0	137	68.6	177	70.0
18	68.8	58	72.5	98	70.3	138	68.6	178	70.3
19	72.4	59	71.1	99	72.9	139	64.7	179	67.5
20	69.2	60	67.1	100	68.5	140	65.9	180	71.7
21	69.5	61	70.6	101	69.8	141	69.3	181	74.0
22	69.8	62	68.0	102	67.9	142	70.3	182	67.6
23	70.3	63	69.1	103	69.8	143	70.7	183	71.1
24	69.0	64	71.7	104	66.5	144	65.7	184	64.6
25	66.4	65	72.2	105	67.5	145	71.1	185	74.0
26	72.3	66	69.7	106	71.0	146	70.4	186	67.9
27	74.4	67	68.3	107	72.8	147	69.2	187	68.5
28	69.2	68	68.7	108	68.1	148	73.7	188	73.4
29	71.0	69	73.1	109	73.6	149	68.5	189	70.4
30	66.5	70	69.0	110	68.0	150	68.5	190	70.7
31	69.2	71	69.8	111	69.6	151	70.7	191	71.6
32	69.0	72	69.6	112	70.6	152	72.3	192	66.9
33	69.4	73	70.2	113	70.0	153	71.4	193	72.6
34	71.5	74	68.4	114	68.5	154	69.2	194	72.2
35	68.0	75	68.7	115	68.0	155	73.9	195	69.1
36	68.2	76	72.0	116	70.0	156	70.2	196	71.3
37	71.1	77	71.9	117	69.2	157	69.6	197	67.9
38	72.0	78	74.1	118	70.3	158	71.6	198	66.1
39	68.3	79	69.3	119	67.2	159	69.7	199	70.8
40	70.6	80	69.0	120	70.7	160	71.2	200	69.5

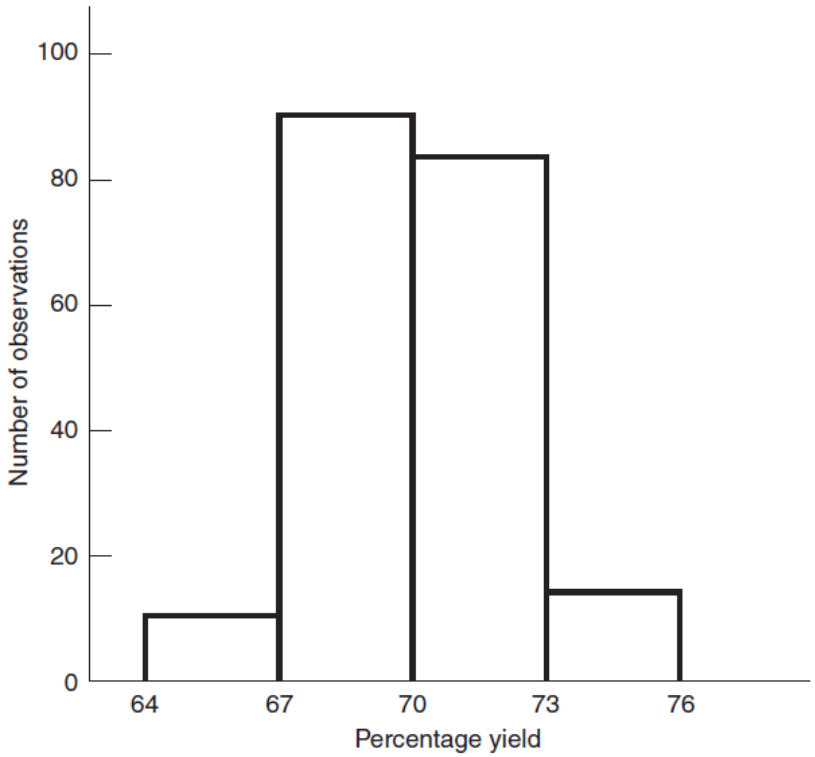
Valorile eşantionului variază de la 64 la 76. Împărţind acest interval în 12 intervale egale şi reprezentând numărul total de randamente observate în fiecare interval ca înălţimea dreptunghiului de deasupra intervalului rezultă histograma după cum e arătat în figura următoare.



O *diagramă de frecvenţe* este obţinută dacă ordonata histogramei este împărţită la numărul total de observaţii, 200 în acest caz, şi la lăţimea intervalului Δ (care se întâmplă să fie 1 în acest exemplu). În dreapta figurii este reprezentată ordonata diagramei de frecvenţe. Vedem că histograma sau diagrama frecvenţelor dă o impresie imediată asupra valorilor, frecvenţei relative şi împrăştierii asociate datelor.

Diagrama frecvenţelor din figură are proprietăţile unei densităţi (suma ariilor dreptunghiurilor este 1). De aici pot fi estimate probabilităţile asociate cu diverse evenimente. De exemplu, probabilitatea unui lot având randament mai mic decât 68% poate fi calculată din diagrama frecvenţelor adunând ariile de la stânga lui 68% şi obţinând $0,13$ $(0,02 + 0,01 + 0.025 + 0.075)$. Similar, probabilitatea unui lot având un randament mai mare de 72% este $0,18$ $(0,105 + 0,035 + 0,03 + 0,01)$. Acestea sunt probabilităţi calculate pe baza datelor observate. Un set diferit de date obţinute din acelaşi proces chimic ar conduce în general la o diagramă de frecvenţe diferită şi deci la valori diferite pentru aceste probabilităţi. În consecinţă, ele sunt estimări ale probabilităţilor $P(X < 68)$ şi $P(X > 72)$ asociate cu variabila aleatoare X.

Pentru acest exemplu, alegerea a 12 intervale este convenabilă datorită intervalului de valori luate de observaţii şi faptului că rezoluţia rezultată este adecvată pentru calculele de probabilităţi de mai sus. În figura următoare este construită o histogramă folosind 4 intervale în loc de 12 pentru acelaşi exemplu.



Se observă că proiectează o impresie vizuală mult diferită şi cu o mai mică acurateţe a comportamentului datelor. Astfel, e important să fie ales numărul de intervale în concordanţă cu informaţia care se vrea extrasă din modelul matematic. Ca un ghid practic, Sturges a sugerat în 1926 ca o valoare aproximativă pentru numărul de intervale k să fie dată de

$$k \simeq 1 + 3,3 \lg n,$$

unde n este numărul de date.

Din punctul de vedere al modelării, este rezonabil să alegem o repartiţie normală ca model probabilistic pentru randamentul procentual X observând că variaţiile aleatoare ale ei sunt rezultanta a numeroase surse aleatoare independente în procesul chimic de fabricare. Dacă aceasta este sau nu o alegere rezonabilă poate fi evaluat intr-un mod subiectiv folosind diagrama frecvenţelor din figura alăturată. Densitatea repartiţiei normale de medie 70 şi dispersie 4 este trasată punctat pe diagrama frecvenţelor, arătând o potrivire rezonabilă. Bazându-ne pe această repartiţie normală, putem calcula probabilităţile de mai

sus, dând o altă evaluare a adecvării modelului. De exemplu, cu ajutorul tabelului valorilor repartiţiei $N(0,1)$,

$P(X<68)=F_U\left(\frac{68-70}{2}\right)=F_U(-1)=1-F_U(1)=1-0,8413=0,1587$,

care se compară cu $0,13$ obţinută folosind diagrama frecvenţelor.

În cele de mai sus, alegerile lui 70 şi 4 ca estimări pentru media, respectiv dispersia lui X, sunt făcute observând că media repartiţiei ar trebui să fie aproape de media aritmetică a selecţiei, adică

$$m_X \simeq \frac{1}{n}\sum_{j=1}^{n} x_j,$$

şi dispersia poate fi aproximată prin

$$\sigma_X^2 \simeq \frac{1}{n}\sum_{j=1}^{n}(x_j-m_X)^2.$$

În cazul unei variabile aleatoare discrete, histograma şi diagrama de frecvenţe obţinute din datele observate iau forma unei reprezentări în batoane, spre deosebire de dreptunghiurile conectate din cazul continuu. Considerăm, de exemplu, repartiţia numărului de accidente pe şofer de-a lungul unei perioade de timp de 6 ani în California. Datele din tabelul următor sunt înregistrările accidentelor pe 6 ani ale 7842 de şoferi californieni.

Number of accidents	Number of drivers
0	5147
1	1859
2	595
3	167
4	54
5	14
> 5	6
	Total = 7842

Bazată pe acest set de observaţii, histograma are forma dată în figura următoare.

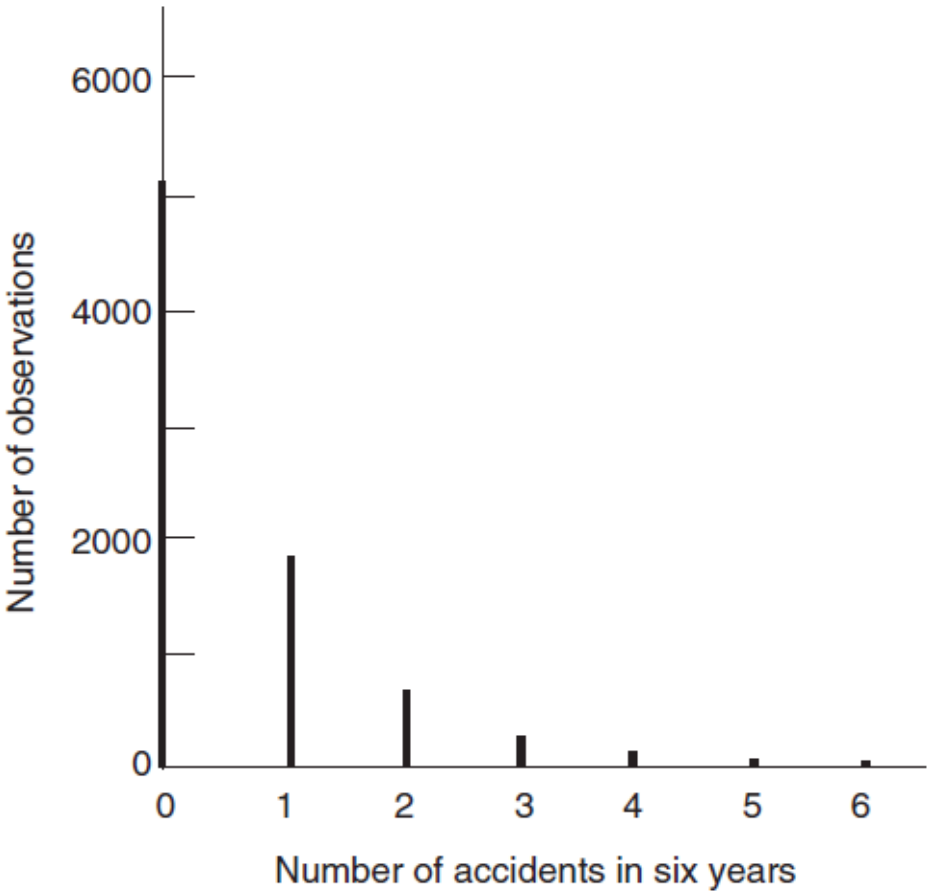


Diagrama frecvenţelor este obţinută în acest caz prin împărţirea ordonatei histogramei la numărul total de observaţii, care este 7842.

7.2 Selecţii şi statistici

În cele mai multe cazuri este imposibil sau nerealist să se observe întreaga populaţie. Unele populaţii au membri care încă nu există (de exemplu, loturile viitoare din producţia de ciment), sau populaţia este prea mare (de exemplu, toate femeile însărcinate). De aceea cercetătorii măsoară doar o parte a populaţiei ţintă, numită selecţie sau eşantion. Dacă facem deducţii despre populaţie folosind informaţia obţinută dintr-o selecţie, atunci este important ca selecţia să fie reprezentativă pentru populaţie, ceea ce poate fi atins dacă toate selecţiile posibile sunt egal probabil să fie alese.

Uneori scopul statisticii este utilizarea selecţiei pentru a estima sau a face unele afirmaţii despre un parametru al populaţiei. Un *parametru* este o măsură descriptivă pentru o populaţie sau pentru o repartiţie a variabilelor aleatoare.

De exemplu, parametrii populaţiei care pot fi de interes includ media, deviaţia standard, cuantile, proporţii, coeficienţi de corelaţie, etc.

Presupunem că un model probabilistic, reprezentat de densitatea $f(x)$, a fost ales pentru un fenomen fizic sau natural pentru care parametrii $\theta_1, \theta_2, ...$ trebuie estimaţi din datele observate independent $x_1, x_2, ..., x_n$. Considerăm pentru moment un singur parametru θ pentru simplitate şi scriem $f(x;\theta)$ pentru o densitate specificată unde θ este parametrul necunoscut care trebuie estimat. Problema estimării parametrului este atunci una a determinării unei funcţii corespunzătoare de $x_1, x_2, ..., x_n$, să spunem $h(x_1, x_2, ..., x_n)$, care dă "cea mai bună" estimare a lui θ.

Fiind dat un set de date independente $x_1, x_2, ..., x_n$, fie

$$\widehat{\theta} = h(x_1, x_2, ..., x_n)$$

o estimare a parametrului θ. Dacă experimentul care a dat setul de date s-ar repeta, am obţine valori diferite pentru $x_1, x_2, ..., x_n$. Funcţia $h(x_1, x_2, ..., x_n)$ aplicată la noul set de date ar da o valoare diferită pentru $\widehat{\theta}$. Vedem astfel că estimarea $\widehat{\theta}$ este ea însăşi o variabilă aleatoare avand o densitate, care depinde atât de forma funcţiei h cât şi de densitatea variabilei aleatoare X. Reprezentarea corespunzătoare a lui $\widehat{\theta}$ e astfel

$$\widehat{\Theta} = h(X_1, X_2, ..., X_n), \tag{7.1}$$

unde $X_1, X_2, ..., X_n$ sunt variabile aleatoare reprezentând o *selecţie* din variabila aleatoare X, care este numită în acest context *populaţie*.

Presupunem că selecţia $X_1, X_2, ..., X_n$ are următoarele proprietăţi:

Proprietatea 1: $X_1, X_2, ..., X_n$ sunt independente.

Proprietatea 2: $f_{X_j}(x) = f_X(x)$, $\forall x, j = 1, 2, ..., n$.

Variabilele aleatoare $X_1, X_2, ..., X_n$ satisfăcând aceste condiţii se numesc o *selecţie aleatore* de mărime n. Cuvântul "aleatore" din această definiţie este de obicei omis pentru scurtime. Daca X este o variabilă aleatoare discretă cu masa $p_X(x)$, atunci $p_{X_j}(x) = p_X(x)$, $\forall j$.

Un set specific de valori observate $(x_1, x_2, ..., x_n)$ este un set de *valori ale selecţiei*.

Datorită proprietăţilor 1 şi 2, densitatea comună este dată de

$$f_{X_1 X_2 ... X_n}(x_1, x_2, ..., x_n) = f_X(x_1) f_X(x_2) ... f_X(x_n).$$

O *statistică* este o funcţie de o selecţie dată $X_1, X_2, ..., X_n$ care nu depinde de parametrul necunoscut. Funcţia $h(X_1, X_2, ..., X_n)$ din (7.1) este o statistică pentru care valoarea poate fi determinată odată ce valorile selecţiei au fost observate. O statistică, fiind o funcţie de variabile aleatoare, este o variabilă aleatoare. Adesea o statistică este folosită pentru următoarele scopuri

- ca o estimare punctuală pentru un parametru al populaţiei,

- pentru a obţine o estimare a intervalului de încredere pentru un parametru, sau,

- ca un test statistic în testarea ipotezelor.

7.2.1 Media de selecţie

Statistica

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

e numită *media de selecţie* a populaţiei X. Fie media şi dispersia populaţiei

$$\begin{array}{l} E(X) = m, \\ var(X) = \sigma^2. \end{array} \tag{7.2}$$

Media mediei de selecţie $\overline{X}$ este

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

Dispersia mediei de selecţie $\overline{X}$ este

$$var(\overline{X}) = E\left(\left(\overline{X} - m\right)^2\right) = E\left(\left[\tfrac{1}{n}\sum_{i=1}^{n}\left(X_i - m\right)\right]^2\right) =$$

$$\tfrac{1}{n^2} E\left(\left[\sum_{i=1}^{n}\left(X_i - m\right)\right]^2\right) =$$

$$\tfrac{1}{n^2} E\left(\sum_{i=1}^{n}\left(X_i - m\right)^2 + 2\sum_{1\le i<j\le n}\left(X_i - m\right)\left(X_j - m\right)\right) =$$

$$\tfrac{1}{n^2}\left[\sum_{i=1}^{n} E\left(\left(X_i - m\right)^2\right) + 2\sum_{1\le i<j\le n} E\left(\left(X_i - m\right)\left(X_j - m\right)\right)\right] =$$

$$\tfrac{1}{n^2}\left[\sum_{i=1}^{n}\sigma^2 + 2\sum_{1\le i<j\le n} E\left(\left(X_i - m\right)\left(X_j - m\right)\right)\right] \overset{\text{independenţă}}{=}$$

$$\tfrac{1}{n^2}\left[\underbrace{\sum_{i=1}^{n}\sigma^2}_{=n\sigma^2} + 2\sum_{1\le i<j\le n}\underbrace{E\left(\left(X_i - m\right)\right)}_{=0}\underbrace{E\left(\left(X_j - m\right)\right)}_{=0}\right] = \tfrac{1}{n^2}\left(n\sigma^2\right) = \tfrac{\sigma^2}{n},$$

deci

$$var(\overline{X}) = \frac{\sigma^2}{n},$$

care este invers proporţională cu mărimea selecţiei n. Când n creşte, dispersia lui $\overline{X}$ descreşte şi repartiţia lui $\overline{X}$ devine ascuţită cu vârful în $E(\overline{X}) = m$. De aici, este clar intuitiv că statistica $\overline{X}$ dă o bună procedură de estimare a mediei populaţiei m.

Deoarece $\overline{X}$ este o sumă de variabile aleatoare independente, pe baza teoremei limită centrală, media de selecţie $\overline{X}$ tinde la o repartiţie normală când

$n \to \infty$. Mai precis, variabila aleatoare $\left(\overline{X} - m\right)\left(\frac{\sigma}{\sqrt{n}}\right)^{-1}$ tinde la $N(0,1)$ când $n \to \infty$.

7.2.2 Dispersia de selecţie

Statistica

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \tag{7.3}$$

se numeşte *dispersia de selecţie* a populaţiei X.

(7.3)$\Rightarrow$

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[(X_i - m) - \left(\overline{X} - m\right)\right]^2 =$

$\frac{1}{n-1}\sum_{i=1}^{n}\left[(X_i - m) - \frac{1}{n}\sum_{j=1}^{n}(X_j - m)\right]^2 =$

$\frac{1}{n-1}\sum_{i=1}^{n}\left\{(X_i - m)^2 - \frac{2}{n}(X_i - m)\sum_{j=1}^{n}(X_j - m) + \frac{1}{n^2}\left[\sum_{j=1}^{n}(X_j - m)\right]^2\right\} =$

$\frac{1}{n-1}\left\{\sum_{i=1}^{n}(X_i - m)^2 - \frac{2}{n}\left[\sum_{i=1}^{n}(X_i - m)\right]^2 + \frac{1}{n}\left[\sum_{i=1}^{n}(X_i - m)\right]^2\right\} =$

$\frac{1}{n-1}\left\{\sum_{i=1}^{n}(X_i - m)^2 - \frac{1}{n}\left[\sum_{i=1}^{n}(X_i - m)\right]^2\right\} =$

$\frac{1}{n-1}\left\{\sum_{i=1}^{n}(X_i - m)^2 - \frac{1}{n}\left[\sum_{i=1}^{n}(X_i - m)^2 + 2\sum_{1\le i<j\le n}(X_i - m)(X_j - m)\right]\right\} =$

$\frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2 - \frac{2}{n(n-1)}\sum_{1\le i<j\le n}(X_i - m)(X_j - m).$

De aici, media lui S^2 este

$E(S^2) =$

$\frac{1}{n}\sum_{i=1}^{n}E\left((X_i - m)^2\right) - \frac{2}{n(n-1)}\sum_{1\le i<j\le n}E\left((X_i - m)(X_j - m)\right) \overset{\text{independenţă}}{=}$

$\frac{1}{n}\sum_{i=1}^{n}\underbrace{E\left((X_i - m)^2\right)}_{=\sigma^2} - \frac{2}{n(n-1)}\sum_{1\le i<j\le n}\underbrace{E(X_i - m)}_{=0}E(X_j - m) = \sigma^2.$

Deci

$$E(S^2) = \sigma^2,$$

unde m şi σ^2 sunt definite de (7.2). Motivul folosirii lui $\frac{1}{n-1}$ în loc de $\frac{1}{n}$ în (7.3) este de a face media lui S^2 egală cu σ^2. Aceasta este o proprietate de dorit pentru S^2 dacă aceasta este folosită la estimarea lui σ^2, adevărata dispersie a lui X.

Dispersia lui S^2 este aflată din

$$var(S^2) = E\left(\left(S^2 - \sigma^2\right)^2\right).$$

Se obţine

$$var(S^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right),$$

unde μ_4 este momentul centrat de ordinul 4 al lui X, adică

$$\mu_4 = E\left((X-m)^4\right).$$

Teorema 7.1. Dacă S^2 este dispersia de selecţie de mărime n a unei populaţii $X \sim N\left(m, \sigma^2\right)$, atunci $\frac{(n-1)S^2}{\sigma^2}$ are o repartiţie χ^2 cu $(n-1)$ grade de libertate.

7.2.3 Momente de selecţie

Momentul de selecţie de ordinul k este

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Observăm că media de selecţie se obţine pentru $k = 1$.

Se poate arăta că

$$\begin{aligned} &E\left(M_k\right) = \alpha_k, \\ &var\left(M_k\right) = \tfrac{1}{n}\left(\alpha_{2k} - \alpha_k^2\right), \end{aligned}$$

unde α_k este momentul de ordinul k al lui X.

Momentul centrat de selecţie de ordinul k este

$$M_k^c = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^k.$$

7.2.4 Statistici de ordine şi funcţia de repartiţie de selecţie

O selecţie $X_1, X_2, ..., X_n$ poate fi pusă în ordine crescătoare. Astfel, pentru această selecţie, statisticile de ordine sunt definite ca

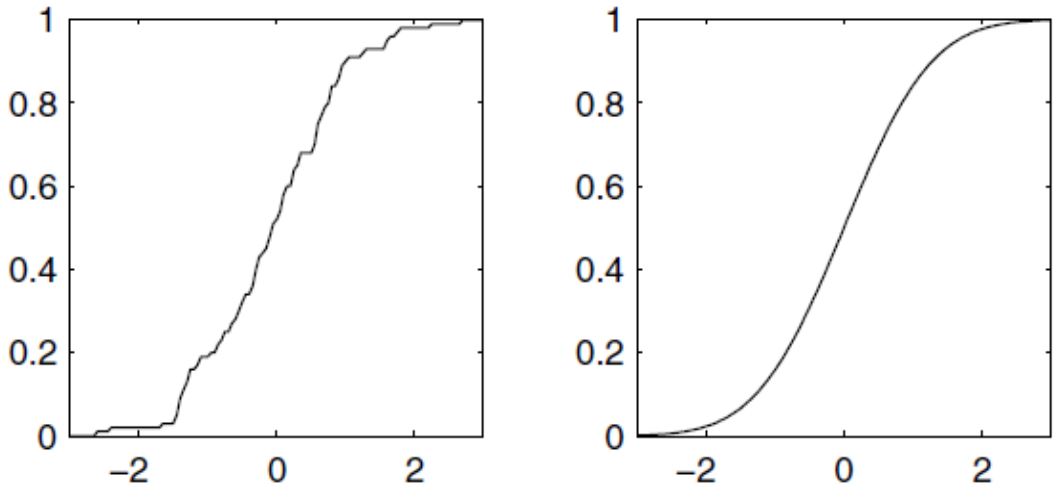
$$X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)},$$

cu $X_{(i)}$ fiind a i-a statistică de ordine.

Funcţia de repartiţie de selecţie $\widehat{F}_n(x)$ este definită ca numărul de date mai mic sau egal cu x împărţit la mărimea selecţiei n. Ea poate fi exprimată în termeni de statistici de ordine astfel:

$$\widehat{F}_n(x) = \begin{cases} 0, \text{ dacă } x < X_{(1)}, \\ \frac{j}{n}, \text{ dacă } X_{(j)} \leq x < X_{(j+1)}, \\ 1, \text{ dacă } x \geq X_{(n)}. \end{cases}$$

Funcţia de repartiţie de selecţie estimează neparametric funcţia de repartiţie a populaţiei X. În figura următoare sunt reprezentate o funcţie de repartiţie de selecţie şi funcţia de repartiţie pentru normala standard.



8 Estimarea parametrilor unor repartiţii prin metoda verosimilităţii maxime

8.1 Criterii de calitate pentru estimări

Statistica

$$\widehat{\Theta} = h\left(X_1, X_2, ..., X_n\right)$$

se numeşte un *estimator* pentru θ.

Odată ce am observat valorile selecţiei $x_1, x_2, ..., x_n$, estimatorul observat,

$$\widehat{\theta} = h\left(x_1, x_2, ..., x_n\right),$$

are o valoare numerică şi va fi numit o *estimare* a parametrului θ.

8.1.1 Nedeplasare

Definiţie. Un estimator $\widehat{\Theta}$ se numeşte estimator *nedeplasat* pentru θ dacă

$$E\left(\widehat{\Theta}\right) = \theta,$$

şi *deplasat* în caz contrar.

Deplasarea estimatorului $\widehat{\Theta}$ este definită ca

$$bias\left(\widehat{\Theta}\right) = E\left(\widehat{\Theta}\right) - \theta.$$

Exemple. 1) Media de selecţie $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ este estimator nedeplasat pentru media populaţiei m deoarece $E\left(\overline{X}\right) = m$.

2) Dispersia de selecţie $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ este estimator nedeplasat pentru dispersia populaţiei σ^2 deoarece $E\left(S^2\right) = \sigma^2$.

3) Estimatorul $S^{2*} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ este estimator deplasat pentru σ^2 deoarece $E\left(S^{2*}\right) = \frac{n-1}{n}\sigma^2$. De aceea pentru definirea dispersiei de selecţie se preferă coeficientul $\frac{1}{n-1}$ in locul alegerii mai naturale $\frac{1}{n}$. Deplasarea lui S^{2*} este

$$bias\left(S^{2*}\right) = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

8.1.2 Dispersie minimă

Definiţie. Fie $\widehat{\Theta}$ un estimator nedeplasat pentru θ. El este un estimator *nedeplasat de dispersie minimă* dacă pentru toţi ceilalţi estimatori nedeplasaţi Θ^* pentru θ din aceeaşi selecţie avem

$$var\left(\widehat{\Theta}\right) \leq var\left(\Theta^*\right).$$

Fiind daţi 2 estimatori nedeplasaţi pentru un parametru dat, cel cu dispersie mai mică este preferat deoarece dispersia mai mică implică faptul că valorile observate ale estimatorului tind sa fie mai aproape de media lui, valoarea adevarată a parametrului.

Exemplu. Am văzut că $\overline{X}$ obţinut dintr-o selecţie de mărime n este un estimator nedeplasat pentru media populaţiei m. Creşte calitatea lui $\overline{X}$ când n creşte?

$E\left(\overline{X}\right) = m$ şi este independentă de volumul eşantionului. Deci $\overline{X}$ rămâne nedeplasat când n creşte. Pe de altă parte

$$var\left(\overline{X}\right) = \frac{\sigma^2}{n},$$

şi descreşte când n creşte. Pe baza criteriului dispersiei minime, calitatea lui $\overline{X}$ ca estimator pentru m creşte când n creşte.□

Teorema 8.1 (inegalitatea Rao-Cramer). Fie $X_1, X_2, ..., X_n$ o selecţie de mărime n din o populaţie X cu densitatea $f(x;\theta)$, unde θ este parametrul necunoscut, şi fie $\widehat{\Theta} = h(X_1, X_2, ..., X_n)$ un estimator nedeplasat pentru θ. Atunci, dispersia lui $\widehat{\Theta}$ satisface inegalitatea

$$var\left(\widehat{\Theta}\right) \geq \left(nE\left(\left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right]^2\right)\right)^{-1}, \qquad \text{(R.C.)}$$

dacă media şi derivata parţială scrise există. Presupunem că se poate deriva in raport cu θ sub integrală. Un rezultat analog se obţine dacă X este discretă înlocuind $f(X;\theta)$ cu $p(X;\theta)$, unde $p(x;\theta)$ este masa populaţiei X.

Demonstraţie. $\theta \overset{\text{nedeplasare}}{=} E\left(\widehat{\Theta}\right) = E\left(h\left(X_1, X_2, ..., X_n\right)\right) \overset{X_1,...,X_n \text{ independente}}{=}$
$\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} h(x_1, ..., x_n) f(x_1;\theta) ... f(x_n;\theta)\, dx_1 ... dx_n \overset{\frac{\partial}{\partial\theta}}{\Longrightarrow}$

$$1 = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} h(x_1, ..., x_n) \left[\sum_{j=1}^{n} \frac{1}{f(x_j;\theta)} \frac{\partial f(x_j;\theta)}{\partial\theta}\right] f(x_1;\theta) ... f(x_n;\theta)\, dx_1 ... dx_n =$$
$$\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} h(x_1, ..., x_n) \left[\sum_{j=1}^{n} \frac{\partial \ln f(x_j;\theta)}{\partial\theta}\right] f(x_1;\theta) ... f(x_n;\theta)\, dx_1 ... dx_n. \tag{8.1}$$

$f(x;\theta)$ densitate $\Rightarrow 1 = \int_{-\infty}^{\infty} f(x_i;\theta)\, dx_i,\ i = 1, 2, ..., n \overset{\frac{\partial}{\partial\theta}}{\Longrightarrow}$
$0 = \int_{-\infty}^{\infty} \frac{\partial f(x_i;\theta)}{\partial\theta} dx_i = \int_{-\infty}^{\infty} \frac{\partial \ln f(x_i;\theta)}{\partial\theta} f(x_i;\theta)\, dx_i,\ i = 1, 2, ..., n.$

Fie $Y=\sum_{j=1}^{n}\frac{\partial \ln f(X_j;\theta)}{\partial\theta}\Rightarrow$

$$E(Y)=\sum_{j=1}^{n}E\left(\frac{\partial \ln f(X_j;\theta)}{\partial\theta}\right)=\sum_{j=1}^{n}\underbrace{\int_{-\infty}^{\infty}\frac{\partial \ln f(x_j;\theta)}{\partial\theta}f(x_j;\theta)\,dx_j}_{=0}=0.$$

$X_1,...,X_n$ independente $\Rightarrow \frac{\partial \ln f(X_1;\theta)}{\partial\theta},...,\frac{\partial \ln f(X_n;\theta)}{\partial\theta}$ independente $\Rightarrow$

$$\sigma_Y^2=\sum_{j=1}^{n}var\left(\frac{\partial \ln f(X_j;\theta)}{\partial\theta}\right)=$$

$$\sum_{j=1}^{n}E\left(\left[\frac{\partial \ln f(X_j;\theta)}{\partial\theta}-\underbrace{E\left(\frac{\partial \ln f(X_j;\theta)}{\partial\theta}\right)}_{=0}\right]^2\right)=\sum_{j=1}^{n}E\left(\left[\frac{\partial \ln f(X_j;\theta)}{\partial\theta}\right]^2\right)$$

$$=\sum_{j=1}^{n}E\left(\left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right]^2\right)=nE\left(\left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right]^2\right).$$

$$1\overset{(8.1)}{=}E(\widehat{\Theta}Y)=E\left(\widehat{\Theta}\right)\underbrace{E(Y)}_{=0}+\rho_{\widehat{\Theta}Y}\sigma_{\widehat{\Theta}}\sigma_Y=\rho_{\widehat{\Theta}Y}\sigma_{\widehat{\Theta}}\sigma_Y\Rightarrow$$

$$\frac{1}{\sigma_{\widehat{\Theta}}^2\sigma_Y^2}=\rho_{\widehat{\Theta}Y}^2\leq 1\Rightarrow$$

$$\sigma_{\widehat{\Theta}}^2\geq\frac{1}{\sigma_Y^2}=\left(nE\left(\left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right]^2\right)\right)^{-1}.\square$$

Definiţie. Marginea inferioară a dispersiei oricărui estimator nedeplasat dată de inegalitatea (RC) este în general o funcţie de θ şi o numim *marginea inferioară Rao-Cramer*. Notaţie

$$\text{MIRC}=\left(nE\left(\left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right]^2\right)\right)^{-1}.$$

Observaţie. $\text{MIRC}=-\left(nE\left(\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right)\right)^{-1}$.

Demonstraţie. $f(x;\theta)$ densitate $\Rightarrow 1=\int_{-\infty}^{\infty}f(x;\theta)\,dx\overset{\frac{\partial}{\partial\theta}}{\Longrightarrow}$

$$0=\int_{-\infty}^{\infty}\frac{\partial f(x;\theta)}{\partial\theta}dx=\int_{-\infty}^{\infty}\frac{\partial \ln f(x;\theta)}{\partial\theta}f(x;\theta)\,dx\overset{\frac{\partial}{\partial\theta}}{\Longrightarrow}$$

$$0=\int_{-\infty}^{\infty}\left[\frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2}f(x;\theta)+\frac{\partial \ln f(x;\theta)}{\partial\theta}\cdot\underbrace{\frac{\partial f(x;\theta)}{\partial\theta}}_{=\frac{\partial \ln f(x;\theta)}{\partial\theta}f(x;\theta)}\right]dx=$$

$$\int_{-\infty}^{\infty}\left\{\frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2}f(x;\theta)+\left[\frac{\partial \ln f(x;\theta)}{\partial\theta}\right]^2f(x;\theta)\right\}dx\Longrightarrow$$

$$\int_{-\infty}^{\infty}\left[\frac{\partial \ln f(x;\theta)}{\partial\theta}\right]^2f(x;\theta)\,dx=-\int_{-\infty}^{\infty}\frac{\partial^2 \ln f(x;\theta)}{\partial\theta^2}f(x;\theta)\,dx\Longrightarrow$$

$$E\left(\left[\frac{\partial \ln f(X;\theta)}{\partial\theta}\right]^2\right)=-E\left(\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right)\Longrightarrow$$

$\text{MIRC} = \left(nE\left(\left[\frac{\partial \ln f(X;\theta)}{\partial \theta}\right]^2\right)\right)^{-1} = -\left(nE\left(\frac{\partial^2 \ln f(X;\theta)}{\partial \theta^2}\right)\right)^{-1}.\square$

Definiţie. Fie $\widehat{\Theta}$ estimator nedeplasat pentru parametrul θ. *Eficienţa* lui $\widehat{\Theta}$ este $e\left(\widehat{\Theta}\right) = \frac{\text{MIRC}}{var\left(\widehat{\Theta}\right)}$.

Observaţie. Eficienţa oricărui estimator nedeplasat este ≤ 1.

Definiţie. Un estimator nedeplasat cu eficienţa 1 se numeşte estimator *eficient*.

Observatie. Din inegalitatea Rao-Cramer rezultă că orice estimator eficient este estimator nedeplasat de dispersie minimă.

Reciproc este adevărat doar dacă există un estimator eficient.

Exemple. 1) Fie $X \sim N\left(m, \sigma^2\right)$. Pe baza unei selecţii fixate de mărime n, este $\overline{X}$ estimator eficient pentru m?

Am văzut că $\overline{X}$ este estimator nedeplasat pentru m. Calculăm MIRC pentru dispersia oricărui estimator nedeplasat pentru m.

$f\left(X;m\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(X-m)^2}{2\sigma^2}\right] \Longrightarrow \ln f\left(X;m\right) = \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{(X-m)^2}{2\sigma^2} \Longrightarrow$
$\frac{\partial \ln f(X;m)}{\partial m} = \frac{X-m}{\sigma^2} \Rightarrow$
$E\left(\left[\frac{\partial \ln f(X;m)}{\partial m}\right]^2\right) = \frac{1}{\sigma^4}E\left((X-m)^2\right) = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2} \Longrightarrow$
$\text{MIRC} = \left(nE\left(\left[\frac{\partial \ln f(X;\theta)}{\partial \theta}\right]^2\right)\right)^{-1} = \frac{\sigma^2}{n}.$

Dar $var\left(\overline{X}\right) = \frac{\sigma^2}{n} \Longrightarrow var\left(\overline{X}\right) = \text{MIRC} \Longrightarrow \overline{X}$ estimator eficient pentru m.

2) Fie $X \sim N\left(0, \sigma^2\right)$, unde σ^2 este un parametru necunoscut ce trebuie estimat dintr-o selecţie de mărime $n > 1$.

a) Să se determine MIRC pentru dispersia oricărui estimator nedeplasat pentru σ^2.

b) Este dispersia de selecţie S^2 un estimator eficient pentru σ^2?

R. a) Notăm $\theta = \sigma^2 \Longrightarrow f\left(X;\theta\right) = \frac{1}{(2\pi\theta)^{\frac{1}{2}}}\exp\left(-\frac{X^2}{2\theta}\right) \Longrightarrow$
$\ln f\left(X;\theta\right) = -\frac{X^2}{2\theta} - \frac{1}{2}\ln 2\pi\theta \Longrightarrow$
$\frac{\partial \ln f(X;\theta)}{\partial \theta} = \frac{X^2}{2\theta^2} - \frac{1}{2\theta} \Longrightarrow \frac{\partial^2 \ln f(X;\theta)}{\partial \theta^2} = -\frac{X^2}{\theta^3} + \frac{1}{2\theta^2} \Longrightarrow$
$E\left(\frac{\partial^2 \ln f(X;\theta)}{\partial \theta^2}\right) = -\frac{1}{\theta^3}\underbrace{E\left(X^2\right)}_{=\sigma^2} + \frac{1}{2\theta^2} = -\frac{\theta}{\theta^3} + \frac{1}{2\theta^2} = -\frac{1}{2\theta^2} \Longrightarrow$
$\text{MIRC} = -\left(nE\left(\frac{\partial^2 \ln f(X;\theta)}{\partial \theta^2}\right)\right)^{-1} = \frac{2\theta^2}{n}.$

b) Am văzut că S^2 este un estimator nedeplasat pentru θ şi

$$\left.\begin{array}{l} var\left(S^2\right) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right) \\ X \sim N\left(0,\sigma^2\right) \Longrightarrow \mu_4 = 3\sigma^4 \end{array}\right\} \Longrightarrow$$

$var\left(S^2\right) = \frac{1}{n}\left(3\sigma^4 - \frac{n-3}{n-1}\sigma^4\right) = \frac{2\sigma^4}{n-1} = \frac{2\theta^2}{n-1} \Longrightarrow$
$e\left(S^2\right) = \frac{\text{MIRC}}{var(S^2)} = \frac{n-1}{n}.$

Observăm că S^2 nu este un estimator eficient pentru θ în acest caz, dar este *asimptotic eficient* în sensul că $\lim\limits_{n\to\infty}\frac{\text{MIRC}}{var(S^2)}=1$.

8.1.3 Consistenţă

Definiţie. Un estimator $\widehat{\Theta}$ bazat pe o selecţie de mărime n este estimator *consistent* pentru θ dacă

$\lim\limits_{n\to\infty}P\left(\left|\widehat{\Theta}-\theta\right|\geq\varepsilon\right)=0,\forall\varepsilon>0.$

Teorema 8.2. Fie $\widehat{\Theta}$ un estimator bazat pe o selecţie de mărime n. Dacă

$$\lim_{n\to\infty}E\left(\widehat{\Theta}\right)=\theta \text{ şi } \lim_{n\to\infty}var\left(\widehat{\Theta}\right)=0,$$

atunci $\widehat{\Theta}$ este un estimator consistent pentru θ.

Demonstraţie. Fie $\varepsilon>0$. Din inegalitatea lui Cebîşev (vezi teorema 2.1), obţinem

$$P\left(\left|\widehat{\Theta}-E\left(\widehat{\Theta}\right)\right|\geq\frac{\varepsilon}{2}\right)\leq\frac{4var\left(\widehat{\Theta}\right)}{\varepsilon^2}.$$

Deoarece

$$\left|\widehat{\Theta}-\theta\right|=\left|\widehat{\Theta}-E\left(\widehat{\Theta}\right)+E\left(\widehat{\Theta}\right)-\theta\right|\leq\left|\widehat{\Theta}-E\left(\widehat{\Theta}\right)\right|+\left|E\left(\widehat{\Theta}\right)-\theta\right|,$$

şi, din $\lim\limits_{n\to\infty}E\left(\widehat{\Theta}\right)=\theta$ rezultă că $\exists n\left(\varepsilon\right)\in\mathbb{N}$ a.î.

$$\left|E\left(\widehat{\Theta}\right)-\theta\right|\leq\frac{\varepsilon}{2},\forall n\geq n\left(\varepsilon\right),$$

avem pentru $n\geq n\left(\varepsilon\right)$

$P\left(\left|\widehat{\Theta}-\theta\right|\geq\varepsilon\right)\leq P\left(\left|\widehat{\Theta}-E\left(\widehat{\Theta}\right)\right|+\left|E\left(\widehat{\Theta}\right)-\theta\right|\geq\varepsilon\right)=P\left(\left|\widehat{\Theta}-E\left(\widehat{\Theta}\right)\right|\geq\varepsilon-\left|E\left(\widehat{\Theta}\right)-\theta\right|\right)\leq P\left(\left|\widehat{\Theta}-E\left(\widehat{\Theta}\right)\right|\geq\frac{\varepsilon}{2}\right).$

Obţinem, pentru $n\geq n\left(\varepsilon\right)$

$$0\leq P\left(\left|\widehat{\Theta}-\theta\right|\geq\varepsilon\right)\leq\frac{4var\left(\widehat{\Theta}\right)}{\varepsilon^2},$$

de unde, trecând la limită rezultă $\lim\limits_{n\to\infty}P\left(\left|\widehat{\Theta}-\theta\right|\geq\varepsilon\right)=0.\square$

8.1.4 Suficienţă

Fie $X_1,X_2,...,X_n$ o selecţie a populaţiei X a cărei repartiţie depinde de parametrul necunoscut θ. Dacă $Y=h\left(X_1,X_2,...,X_n\right)$ este o statistică a. î., pentru orice altă statistică

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

repartiţia condiţionată a lui Z, dat fiind că $Y = y$, nu depinde de θ, atunci Y e numită *statistică suficientă* pentru θ. Dacă în plus $E(Y) = \theta$, atunci Y se numeşte *estimator suficient* pentru θ.

Teorema 8.3. (criteriul de factorizare Fisher-Neymann). Fie

$$Y = h(X_1, X_2, ..., X_n)$$

o statistică bazată pe o selecţie de mărime n a populaţiei continue X. Atunci Y este o statistică suficientă pentru θ dacă şi numai dacă densitatea comună a lui $X_1, X_2, ..., X_n$ poate fi factorizată sub forma

$$\prod_{j=1}^{n} f_X(x_j; \theta) = g_1(h(x_1, ..., x_n), \theta) g_2(x_1, ..., x_n).$$

Dacă X este discretă, condiţia este

$$\prod_{j=1}^{n} p_X(x_j; \theta) = g_1(h(x_1, ..., x_n), \theta) g_2(x_1, ..., x_n).$$

Suficienţa a fost demonstrată de Fisher în 1922, iar necesitatea de Neymann în 1935.

8.2 Metoda verosimilităţii maxime

Introdusă de Fischer în 1922 a devenit cea mai importantă metodă generală de estimare din punct de vedere teoretic.

Fie $f(x; \theta)$ densitatea populaţiei X, unde θ este unicul parametru care trebuie estimat dintr-un set de valori de selecţie $x_1, x_2, ..., x_n$.

Definiţie. *Funcţia de verosimilitate* a unui set de n valori de selecţie din populaţie este

$$L(x_1, x_2, ..., x_n; \theta) = f(x_1; \theta) f(x_2; \theta) ... f(x_n; \theta).$$

În cazul în care X este discretă cu masa $p(x; \theta)$, avem

$$L(x_1, x_2, ..., x_n; \theta) = p(x_1; \theta) p(x_2; \theta) ... p(x_n; \theta).$$

Când valorile de selecţie sunt date, funcţia de verosimilitate L devine o funcţie de o singură variabilă θ. Procedura de estimare pentru θ bazată pe metoda verosimilităţii maxime constă în alegerea, ca o estimare pentru θ, a valorii particulare care maximizează L. Maximul lui $L(\theta)$ apare in cele mai multe cazuri la valoarea lui θ unde $\frac{dL(\theta)}{d\theta} = 0$. De aici, într-un mare număr de cazuri, *estimarea de verosimilitate maximă* (EVM) $\widehat{\theta}$ a lui θ bazată pe valorile de selecţie $x_1, x_2, ..., x_n$ poate fi determinată din

$$\frac{dL\left(x_1, x_2, ..., x_n; \widehat{\theta}\right)}{d\widehat{\theta}} = 0.$$

După cum am văzut mai sus, funcţia L este în forma unui produs de funcţii de θ. Deoarece L este întotdeauna nenegativă şi îşi atinge maximul pentru aceeaşi valoare a lui $\widehat{\theta}$ ca $\ln L$, iar $\ln L$ este în formă de sumă, este în general mai uşor să obţinem EVM $\widehat{\theta}$ rezolvând

$$\frac{d \ln L\left(x_1, x_2, ..., x_n; \widehat{\theta}\right)}{d\widehat{\theta}} = 0,$$

numită *ecuaţia de verosimilitate*.

Soluţia dorită este una unde rădăcina $\widehat{\theta}$ este o funcţie de $x_1, x_2, ..., x_n$, dacă astfel de rădăcină există. Când există mai multe rădăcini ale ecuaţiei de verosimilitate, EVM este rădăcina corespunzătoare maximului global al lui L sau $\ln L$.

În cazul a m parametri, funcţia de verosimilitate devine

$$L\left(x_1, x_2, ..., x_n; \theta_1, ..., \theta_m\right),$$

şi EVM pentru $\theta_1, ..., \theta_m$ sunt obţinuţi rezolvând sistemul de ecuaţii de verosimilitate

$$\frac{\partial \ln L}{\partial \widehat{\theta}_j} = 0, j = 1, 2, ..., m.$$

Revenind la cazul unui singur parametru, să reprezentăm soluţia ecuaţiei de verosimilitate prin

$\widehat{\theta} = h\left(x_1, x_2, ..., x_n\right)$.

Atunci *estimatorul de verosimilitate maximă* (EVM) $\widehat{\Theta}$ pentru θ este

$\widehat{\Theta} = h\left(X_1, X_2, ..., X_n\right)$.

Proprietatea 1: *consistenţa şi eficienţa asimptotică.* Fie $\widehat{\Theta}$ EVM pentru θ din densitatea $f(x; \theta)$ pe baza unei selecţii de mărime n. Atunci EVM $\widehat{\Theta}$ este consistent şi asimptotic eficient.

Un rezultat analog se obţine când populaţia X este discretă. Mai mult, repartiţia lui $\widehat{\Theta}$ tinde la o repartiţie normală când $n \to \infty$.

Proprietatea 2: *proprietatea de invarianţă.* Dacă $\widehat{\Theta}$ este EVM pentru θ, atunci EVM pentru $g(\theta)$ este $g\left(\widehat{\Theta}\right)$, unde g este presupusă bijectivă şi derivabilă în raport cu θ.

De exemplu, dacă $\widehat{\sum}$ este EVM pentru deviaţia standard σ în o repartiţie, atunci EVM pentru dispersia σ^2 este $\widehat{\sum}^2$.

Exemplu. Pentru repartiţia $N(m, \sigma^2)$, să se determine EVM pentru $\theta_1 = m$ şi $\theta_2 = \sigma^2$.

R. Avem

$f\left(x; m, \sigma^2\right) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]$, $x \in \mathbb{R}$.

Funcţia de verosimilitate este

$L\left(x_1, ..., x_n; m, \sigma^2\right) = \prod_{j=1}^{n} \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} \exp\left[-\frac{(x_j-m)^2}{2\sigma^2}\right] =$

$$\left[\frac{1}{(2\pi)^{\frac{1}{2}}\sigma}\right]^n \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-m)^2\right] \Longrightarrow$$

$$\ln L = -\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-m)^2 - \frac{1}{2}n\ln\sigma^2 - \frac{1}{2}n\ln 2\pi.$$

Fie $\theta_1 = m$, $\theta_2 = \sigma^2 \Longrightarrow$

$$\ln L = -\frac{1}{2\theta_2}\sum_{j=1}^{n}(x_j-\theta_1)^2 - \frac{1}{2}n\ln\theta_2 - \frac{1}{2}n\ln 2\pi,$$

$$\frac{\partial \ln L}{\partial \theta_1} = \frac{1}{\theta_2}\sum_{j=1}^{n}(x_j-\theta_1),$$

$$\frac{\partial \ln L}{\partial \theta_2} = \frac{1}{2\theta_2^2}\sum_{j=1}^{n}(x_j-\theta_1)^2 - \frac{n}{2\theta_2}.$$

Sistemul de verosimilitate este

$$\begin{cases} \frac{\partial \ln L}{\partial \widehat{\theta}_1} = 0 \\ \frac{\partial \ln L}{\partial \widehat{\theta}_2} = 0 \end{cases} \Longleftrightarrow \begin{cases} \frac{1}{\widehat{\theta}_2}\sum_{j=1}^{n}\left(x_j-\widehat{\theta}_1\right) = 0 \Longrightarrow \widehat{\theta}_1 = \frac{1}{n}\sum_{j=1}^{n}x_j, \\ \frac{1}{2\widehat{\theta}_2^2}\sum_{j=1}^{n}\left(x_j-\widehat{\theta}_1\right)^2 - \frac{n}{2\widehat{\theta}_2} = 0 \Longrightarrow \widehat{\theta}_2 = \frac{1}{n}\sum_{j=1}^{n}\left(x_j-\widehat{\theta}_1\right)^2. \end{cases}$$

EVM pentru m si σ^2 sunt

$$\widehat{\Theta}_1 = \frac{1}{n}\sum_{j=1}^{n}X_j = \overline{X},$$

$$\widehat{\Theta}_2 = \frac{1}{n}\sum_{j=1}^{n}\left(X_j-\overline{X}\right)^2 = \frac{n-1}{n}S^2.$$

$\overline{X}$ estimator eficient pentru $m \Longrightarrow \widehat{\Theta}_1$ estimator eficient pentru $m \Longrightarrow \widehat{\Theta}_1$ asimptotic eficient pentru m.

$$\left.\begin{array}{l} E\left(\widehat{\Theta}_1\right) = E\left(\overline{X}\right) = m \underset{n\to\infty}{\to} m \\ var\left(\widehat{\Theta}_1\right) = var\left(\overline{X}\right) = \frac{\sigma^2}{n} \underset{n\to\infty}{\to} 0 \end{array}\right\} \overset{\text{teorema 8.2}}{\Longrightarrow} \widehat{\Theta}_1 \text{ estimator consistent pentru } m.$$

$E\left(\widehat{\Theta}_2\right) = \frac{n-1}{n}E\left(S^2\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2 \Longrightarrow \widehat{\Theta}_2$ este estimator deplasat pentru σ^2.

$$\left.\begin{array}{l} E\left(\widehat{\Theta}_2\right) \underset{n\to\infty}{\to} \sigma^2 \\ var\left(\widehat{\Theta}_2\right) = \left(\frac{n-1}{n}\right)^2 var\left(S^2\right) = \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} = \frac{2\sigma^4(n-1)}{n^2} \underset{n\to\infty}{\to} 0 \end{array}\right\} \overset{\text{teorema 8.2}}{\Longrightarrow}$$

$\widehat{\Theta}_2$ estimator consistent pentru σ^2.

$$\left.\begin{array}{l} E\left(\widehat{\Theta}_2\right) \underset{n\to\infty}{\to} \sigma^2 \\ \lim\limits_{n\to\infty} \frac{\text{MIRC}}{var\left(\widehat{\Theta}_2\right)} = \lim\limits_{n\to\infty} \frac{\frac{2\sigma^4}{n}}{\frac{2\sigma^4(n-1)}{n^2}} = \lim\limits_{n\to\infty} \frac{n}{n-1} = 1 \end{array}\right\} \Longrightarrow \widehat{\Theta}_2 \text{ estimator asimptotic eficient pentru } \sigma^2.$$

Exemplu. Presupunem că populaţia X are o repartiţie uniformă pe intervalul $(0,\theta)$. Vrem să determinăm EVM pentru θ.

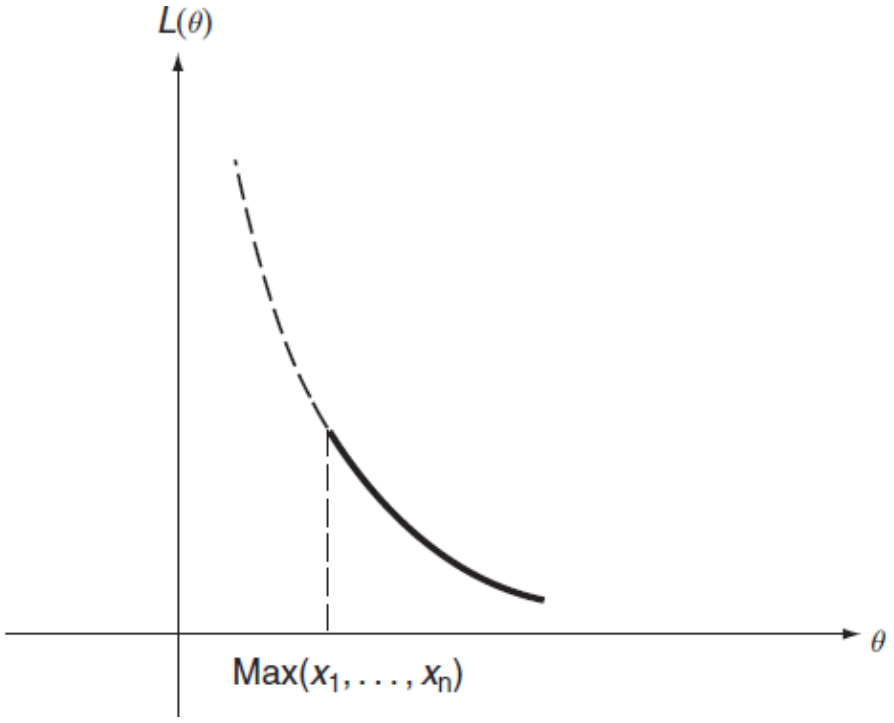
Avem

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}, \text{ pentru } 0 \leq x \leq \theta \\ 0, \text{ altfel.} \end{cases} \tag{8.2}$$

Funcţia de verosimilitate devine

$$L(x_1, x_2, ..., x_n; \theta) = \left(\frac{1}{\theta}\right)^n, \quad 0 \leq x_i \leq \theta, \forall i. \tag{8.3}$$

Graficul lui L este dat în figura următoare.



Observăm că toate valorile selecţiei x_i trebuie să fie mai mici sau egale cu θ, implicând că doar porţiunea curbei de la dreapta lui $\max(x_1, x_2, ..., x_n)$ este aplicabilă. De aici, maximul lui L apare în $\theta = \max(x_1, x_2, ..., x_n)$, sau EVM pentru θ este

$$\widehat{\theta} = \max(x_1, x_2, ..., x_n), \tag{8.4}$$

şi estimatorul de verosimilitate maximă pentru θ este

$$\widehat{\Theta} = \max(X_1, X_2, ..., X_n) = X_{(n)}. \tag{8.5}$$

Observăm că nu am obţinut relaţia (8.4) rezolvând ecuaţia de verosimilitate. Ecuaţia de verosimilitate nu se aplică în acest caz deoarece maximul lui L apare la frontieră şi derivata nu este 0 acolo.

Studiem unele proprietăţi ale lui $\widehat{\Theta}$ dat de relaţia (8.5). Densitatea lui $\widehat{\Theta}$ este (vezi subsecţiunea 5.6)

$$f_{\widehat{\Theta}}(x) = nF_X^{n-1}(x) f_X(x).$$

Cu $f_X(x)$ din relaţia (8.2) şi (din relaţia (5.2), pentru $a = 0$ şi $b = \theta$)

$$F_X(x) = \begin{cases} 0, \text{ pentru } x < 0; \\ \frac{x}{\theta}, \text{ pentru } 0 \leq x \leq \theta; \\ 1, \text{ pentru } x > \theta, \end{cases}$$

avem

$$f_{\widehat{\Theta}}(x) = \begin{cases} \frac{nx^{n-1}}{\theta^n}, \text{ pentru } 0 \leq x \leq \theta; \\ 0, \text{ altfel.} \end{cases}$$

Media şi dispersia lui $\widehat{\Theta}$ sunt

$E\left(\widehat{\Theta}\right) = \int_{-\infty}^{\infty} x f_{\widehat{\Theta}}(x)\, dx = \int_0^{\theta} x \cdot \frac{nx^{n-1}}{\theta^n} dx = \frac{n}{\theta^n} \int_0^{\theta} x^n dx = \frac{n}{\theta^n} \cdot \frac{x^{n+1}}{n+1}|_0^{\theta} = \frac{n}{n+1}\theta,$

$var\left(\widehat{\Theta}\right) = \int_{-\infty}^{\infty} \left(x - E\left(\widehat{\Theta}\right)\right)^2 f_{\widehat{\Theta}}(x)\, dx = \int_0^{\theta} \left(x - \frac{n}{n+1}\theta\right)^2 \frac{nx^{n-1}}{\theta^n} dx =$
$\frac{n}{\theta^n} \int_0^{\theta} \left[x^2 - 2\frac{n}{n+1}\theta x + \left(\frac{n}{n+1}\theta\right)^2\right] x^{n-1} dx = \frac{n}{\theta^n} \left[\frac{x^{n+2}}{n+2} - 2\frac{n}{n+1}\theta \cdot \frac{x^{n+1}}{n+1} + \left(\frac{n}{n+1}\theta\right)^2 \frac{x^n}{n}\right]|_0^{\theta} =$
$n\left[\frac{\theta^2}{n+2} - \frac{2n\theta^2}{(n+1)^2} + \frac{n\theta^2}{(n+1)^2}\right] = n\theta^2\left[\frac{1}{n+2} - \frac{n}{(n+1)^2}\right] = \left[\frac{n}{(n+1)^2(n+2)}\right]\theta^2.$

$E\left(\widehat{\Theta}\right) = \frac{n}{n+1}\theta \neq \theta \implies \widehat{\Theta}$ este deplasat.

$$\left.\begin{array}{l} E\left(\widehat{\Theta}\right) \underset{n\to\infty}{\to} \theta \\ var\left(\widehat{\Theta}\right) = \left[\frac{n}{(n+1)^2(n+2)}\right]\theta^2 \underset{n\to\infty}{\to} 0 \end{array}\right\} \overset{\text{teorema 8.2}}{\Longrightarrow} \widehat{\Theta} \text{ estimator consistent.}$$

9 Estimarea parametrilor repartiţiei normale $N(m, \sigma^2)$: estimatori punctuali, intervale de estimare (de încredere)

9.1 Estimatori punctuali

În secţiunea precedentă am considerat estimarea parametrilor repartiţiei normale prin metoda verosimilităţii maxime.

9.1.1 Metoda momentelor

Metoda momentelor a fost propusă de Pearson în 1894.

Considerăm o densitate $f(x; \theta_1, \theta_2, ..., \theta_m)$ pentru care parametrii $\theta_j, j = 1, 2, ..., m$ sunt de estimat pe baza unei selecţii $X_1, X_2, ..., X_n$ a lui X. Momentele teoretice ale populaţiei X sunt

$$\alpha_i = \int_{-\infty}^{\infty} x^i f(x; \theta_1, ..., \theta_m)\, dx, \quad i = 1, 2, ...$$

Ele sunt, în general, funcţii de parametrii necunoscuţi, adică

$$\alpha_i = \alpha_i(\theta_1, \theta_2, ..., \theta_m).$$

Momentele de selecţie sunt

$$M_i = \frac{1}{n} \sum_{j=1}^{n} X_j^i, \quad i = 1, 2, ...$$

Metoda momentelor sugerează că, pentru a determina estimatorii $\widehat{\Theta}_1, ..., \widehat{\Theta}_m$ din selecţie, egalăm un număr suficient de momente de selecţie cu momentele corespunzătoare ale populaţiei. Procedura determinării lui $\widehat{\Theta}_1, ..., \widehat{\Theta}_m$ constă în următorii paşi:

- Pasul 1: fie

$$\alpha_i\left(\widehat{\Theta}_1, ..., \widehat{\Theta}_m\right) = M_i, \quad i = 1, 2, ..., m. \tag{9.1}$$

  Acestea sunt m ecuaţii de moment cu m necunoscute $\widehat{\Theta}_1, ..., \widehat{\Theta}_m$.

- Pasul 2: rezolvă acest sistem de ecuaţii, determinând $\widehat{\Theta}_1, ..., \widehat{\Theta}_m$, numiţi *estimatorii de moment* pentru $\theta_1, ..., \theta_m$.

Nu este necesar să considerăm m ecuaţii de moment *consecutive* ca în relaţia (9.1); orice m ecuaţii de moment convenabile care conduc la $\widehat{\Theta}_1, ..., \widehat{\Theta}_m$ sunt suficiente. Ecuaţiile momentelor de ordin mai mic sunt totuşi preferate deoarece cer o mai mică manipulare a datelor observate.

Exemplul 9.1. Alegem repartiţia normală ca model pentru randamentul procentual discutat în capitolul 7, adică

$$f\left(x; m, \sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right], \quad x \in \mathbb{R}.$$

Estimăm parametrii $\theta_1 = m$ şi $\theta_2 = \sigma^2$ pe baza celor 200 de valori ale selecţiei date în tabel.

Urmând metoda momentelor, avem nevoie de 2 ecuaţii de moment, şi cele mai convenabile sunt

$$\alpha_1 = M_1 = \overline{X},$$

şi

$$\alpha_2 = M_2.$$

Avem

$$\alpha_1 = \theta_1.$$

De aici, prima dintre aceste ecuaţii de moment dă

$$\widehat{\Theta}_1 = \overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j. \tag{9.2}$$

Proprietăţile acestui estimator au fost discutate în capitolul precedent. El este nedeplasat şi are dispersie minimă între toţi estimatorii nedeplasaţi pentru m.

Ecuaţia momentului de ordinul 2 dă

$$\widehat{\Theta}_1^2 + \widehat{\Theta}_2 = M_2 = \frac{1}{n}\sum_{j=1}^{n} X_j^2,$$

sau

$$\widehat{\Theta}_2 = M_2 - M_1^2 = \tfrac{1}{n}\sum_{j=1}^{n} X_j^2 - \overline{X}^2.$$

Pe de altă parte

$$\tfrac{1}{n}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2 = \tfrac{1}{n}\sum_{j=1}^{n}\left(X_j^2 - 2X_j\overline{X} + \overline{X}^2\right) = \tfrac{1}{n}\left(\sum_{j=1}^{n} X_j^2 - 2\overline{X}\sum_{j=1}^{n} X_j + n\overline{X}^2\right) =$$

$$\tfrac{1}{n}\left(\sum_{j=1}^{n} X_j^2 - 2\overline{X}n\overline{X} + n\overline{X}^2\right) = \tfrac{1}{n}\sum_{j=1}^{n} X_j^2 - \overline{X}^2.$$

Deci

$$\widehat{\Theta}_2 = \frac{1}{n}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2. \tag{9.3}$$

Acesta, după cum am arătat, este un estimator deplasat pentru σ^2.

Estimările $\widehat{\theta}_1$ şi $\widehat{\theta}_2$ ale lui $\theta_1 = m$ şi $\theta_2 = \sigma^2$ pe baza valorilor selecţiei date de tabel sunt, urmând relaţiile (9.2) şi (9.3)

$$\begin{aligned} \widehat{\theta}_1 &= \frac{1}{200}\sum_{j=1}^{200} x_j \simeq 70, \\ \widehat{\theta}_2 &= \frac{1}{200}\sum_{j=1}^{200}\left(x_j - \widehat{\theta}_1\right)^2 \simeq 4, \end{aligned}$$

unde $x_j, \quad j = 1, 2, ..., 200$ sunt valorile de selecţie din tabelul din capitolul 7.

9.2 Intervale de încredere

Exemplul 9.2. 5 valori de selecţie $3; 2; 1, 5; 0, 5; 2, 1$ sunt observate dintr-o repartiţie normală având o medie necunoscută m şi o dispersie cunoscută $\sigma^2 = 9$.

EVM al lui m este media de selecţie $\overline{X}$ şi deci,

$$\widehat{m} = \frac{1}{5}(3 + 2 + 1, 5 + 0, 5 + 2, 1) = 1, 82. \tag{9.4}$$

Vrem să determinăm margini ale unui interval astfel încât, cu un nivel de încredere specificat, media adevarată m să fie în acest interval. $\overline{X}$, fiind sumă de variabile aleatoare normale, este normală cu

$E\left(\overline{X}\right) = m$,

$var\left(\overline{X}\right) = \frac{\sigma^2}{n} = \frac{9}{5}$.

Variabila aleatoare standardizată

$$U = \frac{\overline{X} - E\left(\overline{X}\right)}{\sqrt{var\left(\overline{X}\right)}} = \frac{\sqrt{5}\left(\overline{X} - m\right)}{3} \tag{9.5}$$

este $N(0, 1)$ si are densitatea

$$f_U(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \tag{9.6}$$

Presupunem că cerem ca $P(-u_1 < U < u_1) = 0, 95$. Determinăm u_1.

$0, 95 = P(-u_1 < U < u_1) = \int_{-u_1}^{u_1} f_U(x)\,dx = \int_{-\infty}^{u_1} f_U(x)\,dx - \int_{-\infty}^{-u_1} f_U(x)\,dx =$

$\Phi(u_1) - \Phi(-u_1) =$

$\Phi(u_1) - (1 - \Phi(u_1)) = 2\Phi(u_1) - 1 \Rightarrow$

$\Phi(u_1) = \frac{0,95+1}{2} = 0, 975.$

Din tabelul valorilor funcţiei Φ rezultă $u_1 = 1, 96$ şi

$$P(-1, 96 < U < 1, 96) = \int_{-1,96}^{1,96} f_U(x)\,dx = 0, 95 \tag{9.7}$$

$\Rightarrow P\left(-1,96 < \frac{\sqrt{5}\left(\overline{X}-m\right)}{3} < 1,96\right) = 0,95 \Rightarrow P\left(-5,88 < \sqrt{5}\left(\overline{X}-m\right) < 5,88\right) = 0,95 \Rightarrow$

$P\left(-2,63 < \overline{X} - m < 2,63\right) = 0,95 \Rightarrow P\left(-2,63 < m - \overline{X} < 2,63\right) = 0,95 \Rightarrow$

$$P\left(\overline{X} - 2,63 < m < \overline{X} + 2,63\right) = 0,95. \tag{9.8}$$

Intervalul de 95% încredere pentru m este $\left(\overline{X} - 2,63; \overline{X} + 2,63\right)$. Folosind relaţia (9.4), intervalul de 95% încredere pentru m estimat pe baza observaţiilor este $(\widehat{m} - 2,63; \widehat{m} + 2,63) = (1,82 - 2,63; 1,82 + 2,63) = (-0,81; 4,45)$, adică

$$P\left(-0,81 < m < 4,45\right) = 0,95. \tag{9.9}$$

Probabilitatea ca *intervalul aleator* $\left(\overline{X} - 2,63; \overline{X} + 2,63\right)$ să conţină adevărata medie m a repartiţiei este $0,95$, iar relaţia (9.9) dă intervalul observat bazat pe valorile date ale selecţiei.

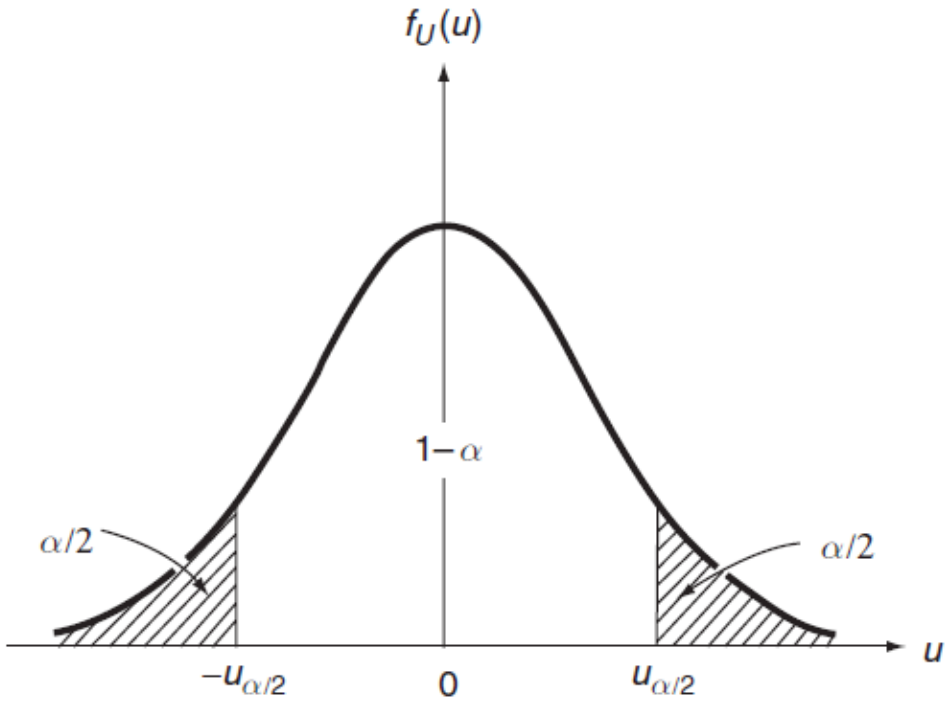
Definiţie. Presupunem că o selecţie $X_1, X_2, ..., X_n$ este extrasă dintr-o populaţie având densitatea $f(x;\theta)$, θ fiind parametrul de estimat. Mai departe presupunem că $L_1(X_1, ..., X_n)$ şi $L_2(X_1, ..., X_n)$ sunt două statistici astfel încât $L_1 < L_2$ cu probabilitatea 1. Intervalul (L_1, L_2) este numit un *interval de încredere* $[100(1-\alpha)]\,\%$ pentru θ dacă L_1 şi L_2 pot fi alese astfel încât

$$P\left(L_1 < \theta < L_2\right) = 1 - \alpha. \tag{9.10}$$

Limitele L_1 şi L_2 sunt numite, respectiv, *limitele inferioară* şi *superioară de încredere* pentru θ, şi $1-\alpha$ este numit *coeficientul de încredere*. Valoarea lui $1-\alpha$ este luată în general $0,9; 0,95; 0,99; 0,995$ sau $0,999$.

Interval de încredere pentru m în $N\left(m, \sigma^2\right)$ cu σ^2 cunoscut Intervalul de încredere dat de relaţia (9.8) estimează media unei populaţii normale cu dispersie cunoscută. În termeni generali, procedura arată că întâi determinăm un interval simetric în U pentru a obţine un coeficient de încredere de $1-\alpha$. Notând cu $u_{\alpha/2}$ valoarea lui U de la care aria de sub $f_U(u)$ este $\alpha/2$, adică $P\left(U > u_{\alpha/2}\right) = \alpha/2$ (vezi figura următoare), avem

$$P\left(-u_{\alpha/2} < U < u_{\alpha/2}\right) = 1 - \alpha. \tag{9.11}$$



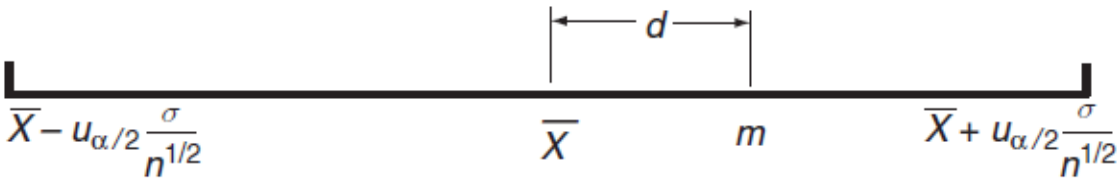
De aici, folosind transformarea dată de relaţia (9.5), avem rezultatul general

$$P\left(\overline{X} - \frac{\sigma u_{\alpha/2}}{\sqrt{n}} < m < \overline{X} + \frac{\sigma u_{\alpha/2}}{\sqrt{n}}\right) = 1 - \alpha. \tag{9.12}$$

Acest rezultat poate fi de asemenea folosit pentru a estima medii ale populaţiilor nenormale cu dispersii cunoscute dacă mărimea selecţiei este suficient de mare pentru a justifica folosirea teoremei limită centrală.

În acest caz, poziţia intervalului este o funcţie de $\overline{X}$ şi de aceea este o funcţie de selecţie. În contrast, mărimea intervalului, este o funcţie doar de mărimea selecţiei n, fiind invers proporţională cu $\sqrt{n}$.

Intervalul de încredere $[100\,(1-\alpha)]\,\%$ pentru m dat de relaţia (9.12) dă de asemenea o estimare a acurateţei estimatorului punctual $\overline{X}$ pentru m. După cum vedem din figura următoare, adevărata medie m se află în intervalul indicat cu $[100\,(1-\alpha)]\,\%$ încredere.



Deoarece $\overline{X}$ este centrul intervalului, distanţa dintre $\overline{X}$ şi m poate fi cel mult egală cu jumătate din mărimea intervalului.

Teorema 9.1. Cu $[100\,(1-\alpha)]\,\%$ încredere, eroarea folosirii estimatorului $\overline{X}$ pentru m este mai mică decât

$$\frac{\sigma u_{\alpha/2}}{\sqrt{n}}.$$

Exemplul 9.3. Fie populaţia X repartizată normal cu dispersia cunoscută σ^2. Determinaţi mărimea minimă n a selecţiei de care e nevoie, astfel încât eroarea estimării mediei m prin $\overline{X}$ să fie mai mică decât o cantitate specificată ε cu $[100\,(1-\alpha)]\,\%$ încredere.

Folosind teorema 9.1, mărimea minimă a selecţiei trebuie să satisfacă

$$\varepsilon = \frac{\sigma u_{\alpha/2}}{\sqrt{n}}.$$

De aici, soluţia este

$$n = \left(\frac{\sigma u_{\alpha/2}}{\varepsilon}\right)^2. \tag{9.13}$$

Interval de încredere pentru m în $N\left(m, \sigma^2\right)$ cu σ^2 necunoscut Diferenţa dintre această problemă şi precedenta este că, deoarece σ nu este cunoscut, nu mai putem folosi

$$U = \left(\overline{X} - m\right)\left(\frac{\sigma}{\sqrt{n}}\right)^{-1}$$

ca variabila aleatoare pentru calculul limitelor de încredere privind media m. Folosim dispersia de selecţie S^2 ca un estimator nedeplasat pentru σ^2 şi considerăm variabila aleatoare

$$Y = \left(\overline{X} - m\right)\left(\frac{S}{\sqrt{n}}\right)^{-1}. \tag{9.14}$$

Variabila aleatoare Y este acum o funcţie de variabilele aleatoare $\overline{X}$ şi S. Determinăm repartiţia lui Y.

Teorema 9.2. (Repartiţia t a lui Student) Considerăm o variabilă aleatoare T definită prin

$$T = U\left(\frac{V}{n}\right)^{-\frac{1}{2}}. \tag{9.15}$$

Dacă $U \sim N(0,1)$, V este repartizată χ^2 cu n grade de libertate, iar U şi V sunt independente, atunci densitatea lui T are forma

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}, \quad t \in \mathbb{R}. \tag{9.16}$$

Această repartiţie este cunoscută ca *repartiţia t a lui Student* cu n grade de libertate; este numită după W. S. Gosset, care a folosit pseudonimul "Student" în publicaţiile sale de cercetare.

După cum se vede din relaţia (9.16), repartiţia t este simetrică în jurul originii. Când n creşte, ea tinde la o repartiţie normală standard.

Întorcându-ne la variabila aleatoare Y definită de relaţia (9.14), fie

$$U = \left(\overline{X} - m\right)\left(\frac{\sigma}{\sqrt{n}}\right)^{-1}$$

şi

$$V = \frac{(n-1)S^2}{\sigma^2}.$$

Atunci

$$Y = U\left(\frac{V}{n-1}\right)^{-\frac{1}{2}}, \tag{9.17}$$

unde $U \sim N(0,1)$, iar V are repartiţia χ^2 cu $(n-1)$ grade de libertate (vezi teorema 7.1). Mai departe, se poate arăta că $\overline{X}$ şi S^2 sunt independente, deci şi U şi V sunt independente. Din teorema 9.2, variabila aleatoare Y are astfel o repartiţie t cu $(n-1)$ grade de libertate.

Variabila aleatoare Y poate fi folosită pentru a stabili intervale de încredere pentru media m. Valoarea lui Y depinde de media necunoscută m, dar repartiţia lui Y nu depinde de m.

Repartiţia t este tabelată în tabelul A.4 următor.

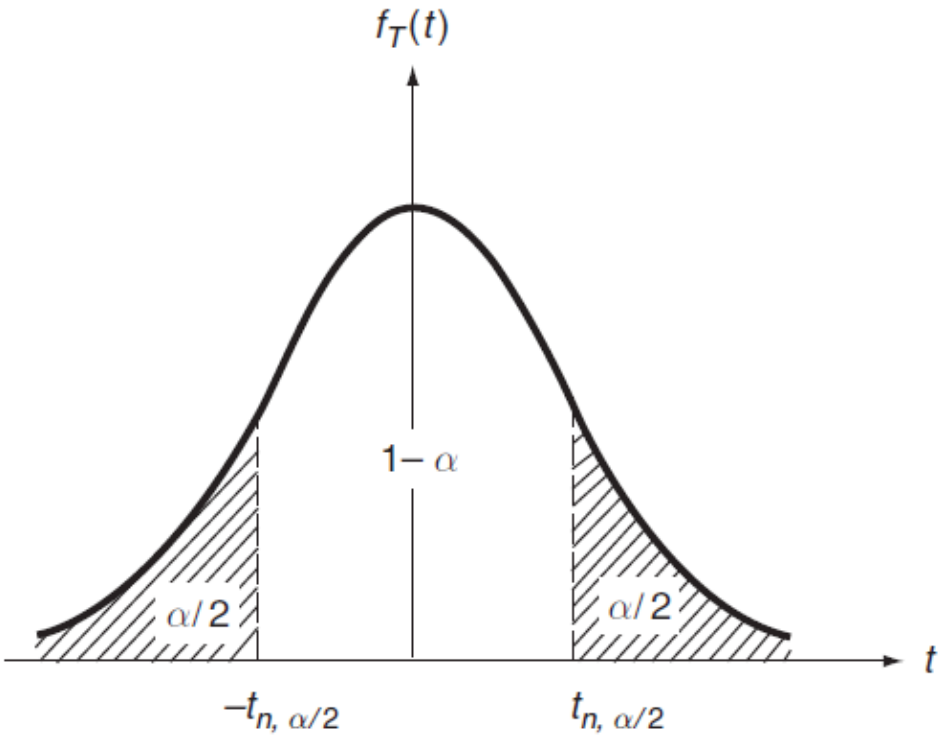
Table A.4 Student's distribution with n degrees of Freedom: a table of $t_{n,\alpha}$ in $P(T > t_{n,\alpha}) = \alpha$, for $\alpha = 0.005$ to 0.10, $n = 1, 2, \ldots$

n	α				
	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.979
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
∞	1.282	1.645	1.960	2.326	2.576

Fie $t_{n,\alpha/2}$ valoarea astfel încât

$$P\left(T > t_{n,\alpha/2}\right) = \frac{\alpha}{2},$$

cu n reprezentând numărul de grade de libertate (vezi figura următoare).



Avem

$$P\left(-t_{n-1,\alpha/2} < Y < t_{n-1,\alpha/2}\right) = 1-\alpha. \tag{9.18}$$

Înlocuind Y din relaţia (9.14) în relaţia (9.18), un interval de încredere $[100\,(1-\alpha)]\,\%$ pentru media m este dat de

$$P\left(\overline{X} - \frac{t_{n-1,\alpha/2}S}{\sqrt{n}} < m < \overline{X} + \frac{t_{n-1,\alpha/2}S}{\sqrt{n}}\right) = 1-\alpha. \tag{9.19}$$

Deoarece $\overline{X}$ şi S sunt funcţii de selecţie, atât poziţia cât şi mărimea intervalului dat mai sus variază de la selecţie la selecţie.

Exemplul 9.4. Presupunem că nivelul anual al căderilor de zăpadă în zona Bufallo este repartizat normal. Folosind înregistrările căderilor de zăpadă din 1969-1970 până în 1978-1979 din tabelul următor, să se determine un interval de încredere 95% pentru media m.

Year	Snowfall
1969–1970	120.5
1970–1971	97.0
1971–1972	109.9
1972–1973	78.8
1973–1974	88.7
1974–1975	95.6
1975–1976	82.5
1976–1977	199.4
1977–1978	154.3
1978–1979	97.3

În acest exemplu $\alpha = 0.05$, $n = 10$, media de selecţie observată este

$$\overline{x} = \frac{1}{10}\left(120.5 + 97 + ... + 97.3\right) = 112.4,$$

şi dispersia de selecţie observată este

$$s^2 = \frac{1}{9}\left[\left(120.5 - 112.4\right)^2 + \left(97 - 112.4\right)^2 + ... + \left(97.3 - 112.4\right)^2\right] = 1414.3.$$

Din tabelul A.4 găsim $t_{9,0.025} = 2.262$. Înlocuind aceste valori în relaţia (9.19), obţinem

$$P\left(85.5 < m < 139.3\right) = 0.95.$$

Interval de încredere pentru σ^2 în $N\left(m, \sigma^2\right)$ Un estimator punctual nedeplasat pentru dispersia σ^2 a populaţiei este S^2. Pentru construirea intervalelor de încredere pentru σ^2, folosim variabila aleatoare

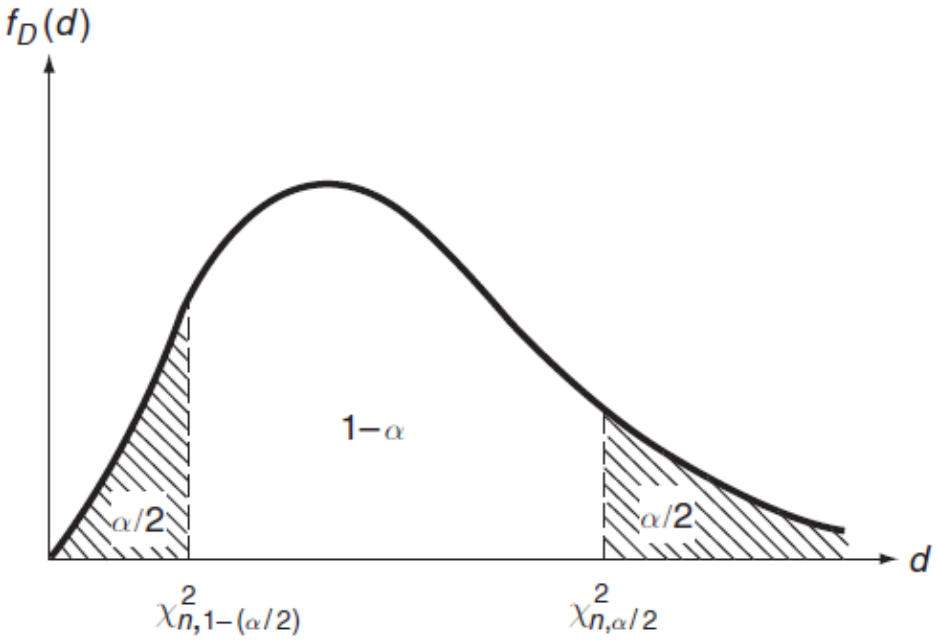
$$D = \frac{\left(n-1\right)S^2}{\sigma^2}, \tag{9.20}$$

care are repartiţie chi-pătrat cu $(n-1)$ grade de libertate (teorema 7.1). Notând $\chi^2_{n,\alpha/2}$ valoarea astfel încât $P\left(D > \chi^2_{n,\alpha/2}\right) = \alpha/2$ cu n grade de libertate, putem scrie (vezi figura următoare)

$$P\left(\chi^2_{n-1,1-(\alpha/2)} < D < \chi^2_{n-1,\alpha/2}\right) = 1 - \alpha, \tag{9.21}$$

care dă, înlocuind D din relaţia (9.20),

$$P\left(\frac{\left(n-1\right)S^2}{\chi^2_{n-1,\alpha/2}} < \sigma^2 < \frac{\left(n-1\right)S^2}{\chi^2_{n-1,1-(\alpha/2)}}\right) = 1 - \alpha. \tag{9.22}$$



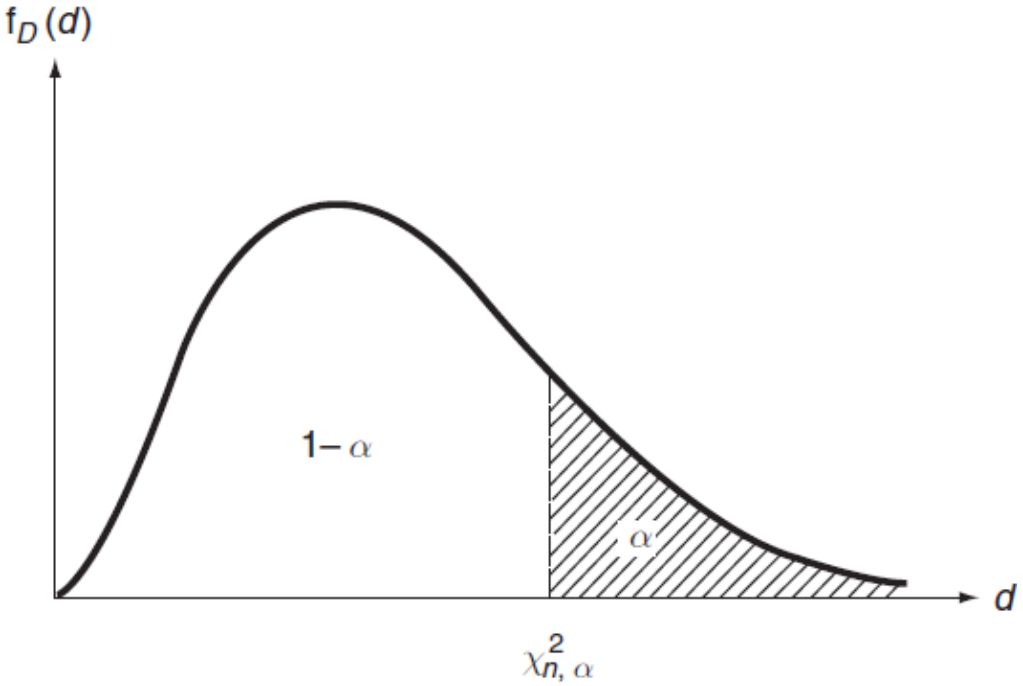
Tabelul A.5 dă valori ale lui $\chi^2_{n,\alpha}$ pentru diferite valori ale lui n şi α.

Table A.5 Chi-squared distribution with n degrees of freedom: a table of $\chi^2_{n,\alpha}$ in $P(\mathrm{D} > \chi^2_{n,\alpha}) = \alpha$, for $\alpha = 0.005$ to 0.995, $n = 1$ to 30

n	α							
	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.0^4393	0.0^3157	0.0^3982	0.0^2393	3.841	5.024	6.635	7.879
2	0.0100	0.0201	0.0506	0.103	5.991	7.378	9.210	10.597
3	0.717	0.115	0.216	0.352	7.815	9.348	11.346	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.832	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.628	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.558
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672

Relaţia (9.22) este folosită pentru construirea intervalelor de încredere pentru σ^2 cu ambele margini finite pentru o populaţie normală. Un interval de încredere pentru σ^2 cu o singură margine finită este dat de (vezi figura următoare)

$$P\left(\sigma^2 > \frac{(n-1)\,S^2}{\chi^2_{n-1,\alpha}}\right) = 1 - \alpha. \tag{9.23}$$



Exemplul 9.5. Determinăm intervalele de încredere 95% pentru σ^2 pentru datele din exemplul 9.4.

Am văzut în exemplul 9.4 că dispersia de selecţie observată este

$$s^2 = 1414.3.$$

Din tabelul A.5 obţinem

$$\chi^2_{9,0.975} = 2.7, \quad \chi^2_{9,0.025} = 19.023, \quad \chi^2_{9,0.05} = 16.919.$$

Relaţiile (9.22) şi (9.23) cu $n = 10$ şi $\alpha = 0.05$ conduc la

$$P\left(669.12 < \sigma^2 < 4714.33\right) = 0.95$$

şi

$$P\left(\sigma^2 > 752.3\right) = 0.95.$$

10 Modele liniare. Estimarea parametrilor prin metoda celor mai mici pătrate. Elemente de analiză discriminantă liniară

10.1 Regresie liniară

Presupunem că variabila aleatoare Y este o funcţie de o variabilă independentă şi relaţia lor este liniară, adică

$$E(Y) = \alpha + \beta x. \tag{10.1}$$

Constantele α şi β sunt necunoscute şi trebuie estimate dintr-o selecţie de valori ale lui Y cu valorile asociate lor ale lui x. Într-un singur experiment, x va avea o anumită valoare x_i şi media lui Y va lua valoarea

$$E(Y_i) = \alpha + \beta x_i. \tag{10.2}$$

Definim variabila aleatoare E prin

$$E = Y - (\alpha + \beta x). \tag{10.3}$$

Modelul de regresie liniară simplă este

$$Y = \alpha + \beta x + E, \tag{10.4}$$

unde E are media 0 şi dispersia σ^2, care coincide cu dispersia lui Y. Valoarea lui σ^2 este necunoscută în general, dar este presupusă a fi constantă şi nu o funcţie de x.

Parametrii necunoscuţi α şi β sunt numiţi *coeficienţi de regresie*, iar variabila aleatoare E reprezintă deviaţia lui Y în jurul mediei.

10.1.1 Estimarea prin metoda celor mai mici pătrate

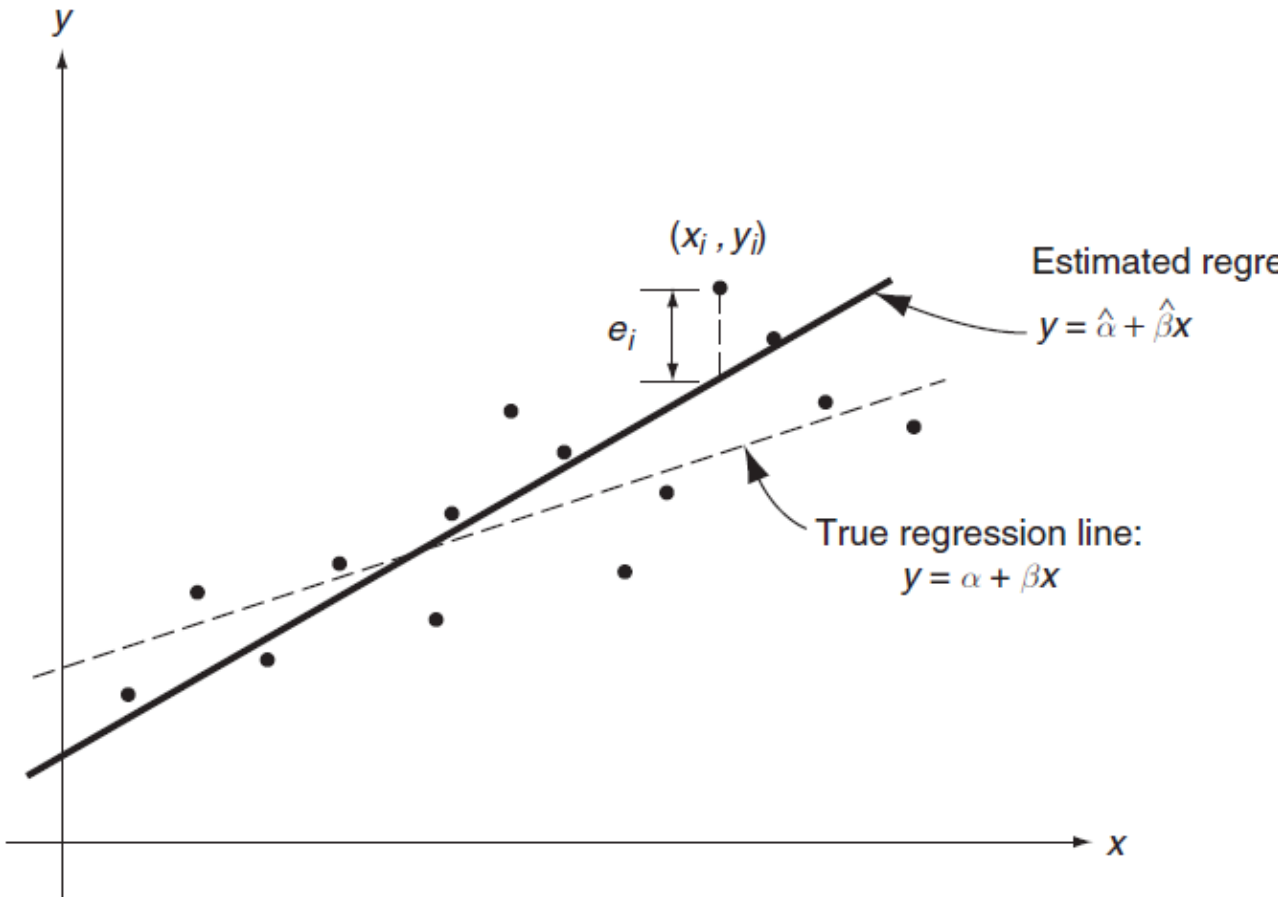
Estimările $\widehat{\alpha}$ şi $\widehat{\beta}$ prin metoda celor mai mici pătrate pentru parametrii de regresie α şi β sunt alese astfel încât suma pătratelor diferenţelor dintre valorile de selecţie observate y_i şi $\widehat{\alpha}+\widehat{\beta}x_i$, valoarea medie estimată a lui Y, este minimă. Fie

$$e_i = y_i - \left(\widehat{\alpha} + \widehat{\beta}x_i\right), i = \overline{1,n}. \tag{10.5}$$

Estimările $\widehat{\alpha}$ şi $\widehat{\beta}$ sunt găsite minimizând

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - \left(\widehat{\alpha} + \widehat{\beta}x_i\right)\right]^2. \tag{10.6}$$

Perechile valorilor de selecţie sunt $(x_1, y_1), ..., (x_n, y_n)$, iar e_i, $i = 1, 2, ..., n$ se numesc *reziduuri*. Figura următoare dă o prezentare grafică a acestei proceduri.



Vedem că reziduurile sunt distanţele verticale dintre y_i, valorile observate ale lui Y, şi estimarea $\widehat{\alpha}+\widehat{\beta}x$ a adevăratei drepte de regresie $\alpha+\beta x$.

Teorema 10.1. Fie modelul de regresie liniară simplă dat de relaţia (10.4). Fie $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$ valorile de selecţie observate ale lui Y cu valorile asociate ale lui x. Presupunem că cel puţin 2 dintre $x_1,...,x_n$ sunt distincte. Atunci estimările prin metoda celor mai mici pătrate ale lui α şi β sunt

$$\widehat{\alpha}=\overline{y}-\widehat{\beta}\overline{x}, \tag{10.7}$$

$$\widehat{\beta}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}, \tag{10.8}$$

unde

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

şi

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Demonstraţie. Fie

$Q\left(\widehat{\alpha},\widehat{\beta}\right) = \sum_{i=1}^{n}\left[y_i - \left(\widehat{\alpha}+\widehat{\beta}x_i\right)\right]^2 \Longrightarrow$

$\frac{\partial Q}{\partial \widehat{\alpha}} = \sum_{i=1}^{n} -2\left[y_i - \left(\widehat{\alpha}+\widehat{\beta}x_i\right)\right],$

$\frac{\partial Q}{\partial \widehat{\beta}} = \sum_{i=1}^{n} -2x_i\left[y_i - \left(\widehat{\alpha}+\widehat{\beta}x_i\right)\right],$

$\frac{\partial^2 Q}{\partial \widehat{\alpha}^2} = \sum_{i=1}^{n} -2(-1) = 2n > 0,$

$\frac{\partial^2 Q}{\partial \widehat{\alpha}\partial \widehat{\beta}} = \sum_{i=1}^{n} -2x_i(-1) = 2\sum_{i=1}^{n} x_i = 2n\overline{x},$

$\frac{\partial^2 Q}{\partial \widehat{\beta}^2} = \sum_{i=1}^{n} -2x_i(-x_i) = 2\sum_{i=1}^{n} x_i^2,$

$D = \begin{vmatrix} \frac{\partial^2 Q}{\partial \widehat{\alpha}^2} & \frac{\partial^2 Q}{\partial \widehat{\alpha}\partial \widehat{\beta}} \\ \frac{\partial^2 Q}{\partial \widehat{\alpha}\partial \widehat{\beta}} & \frac{\partial^2 Q}{\partial \widehat{\beta}^2} \end{vmatrix} = 4n\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = 4n\left(\sum_{i=1}^{n} x_i^2 - 2n\overline{x}^2 + n\overline{x}^2\right) =$

$4n\left(\sum_{i=1}^{n} x_i^2 - 2n\overline{x}\cdot\frac{1}{n}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\overline{x}^2\right) = 4n\sum_{i=1}^{n}\left(x_i^2 - 2\overline{x}x_i + \overline{x}^2\right) =$

$4n\sum_{i=1}^{n}(x_i-\overline{x})^2 > 0 \overset{\frac{\partial^2 Q}{\partial \widehat{\alpha}^2}>0}{\Longrightarrow}$

$HQ = \begin{pmatrix} \frac{\partial^2 Q}{\partial \widehat{\alpha}^2} & \frac{\partial^2 Q}{\partial \widehat{\alpha}\partial \widehat{\beta}} \\ \frac{\partial^2 Q}{\partial \widehat{\alpha}\partial \widehat{\beta}} & \frac{\partial^2 Q}{\partial \widehat{\beta}^2} \end{pmatrix}$ pozitiv definită $\Longrightarrow Q$ (strict) convexă $\Longrightarrow \widehat{\alpha}$ şi $\widehat{\beta}$ care minimizează pe Q sunt soluţia sistemului

$$\begin{cases} \frac{\partial Q}{\partial \widehat{\alpha}} = 0 \\ \frac{\partial Q}{\partial \widehat{\beta}} = 0 \end{cases} \Longleftrightarrow \begin{cases} n\widehat{\alpha} + \widehat{\beta}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \widehat{\alpha}\sum_{i=1}^{n} x_i + \widehat{\beta}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \Longleftrightarrow$$

$$n\widehat{\alpha} + n\overline{x}\widehat{\beta} = n\overline{y} \tag{10.9}$$

$$n\overline{x}\widehat{\alpha} + \widehat{\beta}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \tag{10.10}$$

Ecuaţiile din ultimul sistem se numesc ecuaţii normale. Înmulţind (10.9) cu $\overline{x}$ şi scăzând-o din (10.10) $\Rightarrow$

$$\widehat{\beta}=\frac{\sum_{i=1}^{n}x_iy_i-n\overline{xy}}{\sum_{i=1}^{n}x_i^2-n\overline{x}^2}\overset{\text{vezi calculul lui } D}{=}\frac{\sum_{i=1}^{n}x_iy_i-\overline{x}\sum_{i=1}^{n}y_i-\overline{y}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\overline{xy}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=$$

$$\frac{\sum_{i=1}^{n}[y_i(x_i-\overline{x})-\overline{y}(x_i-\overline{x})]}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}.$$

Din (10.9) $\Rightarrow$

$\widehat{\alpha}=\overline{y}-\overline{x}\widehat{\beta}.\square$

Restabilim rezultalele de mai sus folosind o notaţie matriceală care uşurează calculele şi permite generalizări când considerăm modele mai generale de regresie.

În termeni de valori de selecţie observate $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$, avem sistemul de ecuaţii de regresie observate

$$y_i=\alpha+\beta x_i+e_i,\quad i=1,2,...,n. \tag{10.11}$$

Fie

$$C=\begin{pmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_n\end{pmatrix},\quad \mathbf{y}=\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix},\quad \mathbf{e}=\begin{pmatrix}e_1\\ e_2\\ \vdots\\ e_n\end{pmatrix},$$

şi

$$\boldsymbol{\theta}=\begin{pmatrix}\alpha\\ \beta\end{pmatrix}.$$

Ecuaţiile (10.11) se scriu echivalent

$$\mathbf{y}=C\boldsymbol{\theta}+\mathbf{e}. \tag{10.12}$$

Suma pătratelor reziduurilor dată de relaţia (10.6) este acum

$$Q(\boldsymbol{\theta})=\mathbf{e}^T\mathbf{e}=(\mathbf{y}-C\boldsymbol{\theta})^T(\mathbf{y}-C\boldsymbol{\theta}). \tag{10.13}$$

Estimarea prin metoda celor mai mici pătrate a lui $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}$, este găsită minimizând Q. Deoarece

$$\nabla Q(\boldsymbol{\theta})=-2C^T(\mathbf{y}-C\boldsymbol{\theta})$$

şi

$$HQ(\boldsymbol{\theta})=2C^TC$$

este pozitiv definită, soluţia $\widehat{\boldsymbol{\theta}}$ este obţinută din

$$\nabla Q\left(\widehat{\boldsymbol{\theta}}\right)=0,$$

deci din ecuaţia normală

$$C^T\left(\mathbf{y}-C\widehat{\boldsymbol{\theta}}\right)=0, \tag{10.14}$$

sau

$$C^TC\widehat{\boldsymbol{\theta}}=C^T\mathbf{y},$$

care dă

$$\widehat{\boldsymbol{\theta}}=\left(C^TC\right)^{-1}C^T\mathbf{y}. \tag{10.15}$$

În cele de mai sus, matricea $HQ\left(\boldsymbol{\theta}\right)$ este pozitiv definită şi inversa matricei C^TC există dacă sunt cel puţin 2 valori distincte x_i în selecţie.

Verificăm că relaţia (10.15) este identică cu relaţiile (10.7) şi (10.8) observând că

$$C^TC=\begin{pmatrix}1 & 1 & ... & 1\\ x_1 & x_2 & ... & x_n\end{pmatrix}\begin{pmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_n\end{pmatrix}=\begin{pmatrix}n & n\overline{x}\\ n\overline{x} & \sum\limits_{i=1}^{n}x_i^2\end{pmatrix},$$

$$\left(C^TC\right)^{-1}=\frac{1}{\det C^TC}\left(C^TC\right)^*=\frac{1}{n\left(\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^2\right)}\begin{pmatrix}\sum\limits_{i=1}^{n}x_i^2 & -n\overline{x}\\ -n\overline{x} & n\end{pmatrix},$$

$$C^T\mathbf{y}=\begin{pmatrix}1 & 1 & ... & 1\\ x_1 & x_2 & ... & x_n\end{pmatrix}\begin{pmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{pmatrix}=\begin{pmatrix}n\overline{y}\\ \sum\limits_{i=1}^{n}x_iy_i\end{pmatrix}$$

şi

$$\widehat{\boldsymbol{\theta}}=\left(C^TC\right)^{-1}C^T\mathbf{y}=\frac{1}{n\left(\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^2\right)}\begin{pmatrix}\sum\limits_{i=1}^{n}x_i^2 & -n\overline{x}\\ -n\overline{x} & n\end{pmatrix}\begin{pmatrix}n\overline{y}\\ \sum\limits_{i=1}^{n}x_iy_i\end{pmatrix}=\begin{pmatrix}\dfrac{\overline{y}\sum\limits_{i=1}^{n}x_i^2-\overline{x}\sum\limits_{i=1}^{n}x_iy_i}{\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^2}\\ \dfrac{\sum\limits_{i=1}^{n}x_iy_i-n\overline{xy}}{\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^2}\end{pmatrix}=$$

$$\begin{pmatrix}\overline{y}-\dfrac{\sum\limits_{i=1}^{n}x_iy_i-n\overline{xy}}{\sum\limits_{i=1}^{n}x_i^2-n\overline{x}^2}\overline{x}\\ \dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\end{pmatrix}=\begin{pmatrix}\overline{y}-\widehat{\beta}\overline{x}\\ \dfrac{\sum\limits_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\end{pmatrix}.$$

Exemplul 10.1. Este de aşteptat ca randamentul procentual mediu, Y, dintr-un proces chimic să fie legat liniar de temperatura procesului, x, în $^{\circ}C$. Determinaţi prin metoda celor mai mici pătrate dreapta de regresie pentru $E(Y)$ pe baza a 10 observaţii date în tabelul

i	1	2	3	4	5	6	7	8	9	10
$x\,(^{\circ}C)$	45	50	55	60	65	70	75	80	85	90
y	43	45	48	51	55	57	59	63	66	68

R. Pentru a folosi relaţiile (10.7) şi (10.8) avem nevoie de următoarele cantităţi

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}\left(45 + 50 + ... + 90\right) = \frac{5 \cdot 135}{10} = 67,5,$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{10}\left(43 + 45 + ... + 68\right) = 55,5,$$

$$\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 = 2062,5,$$

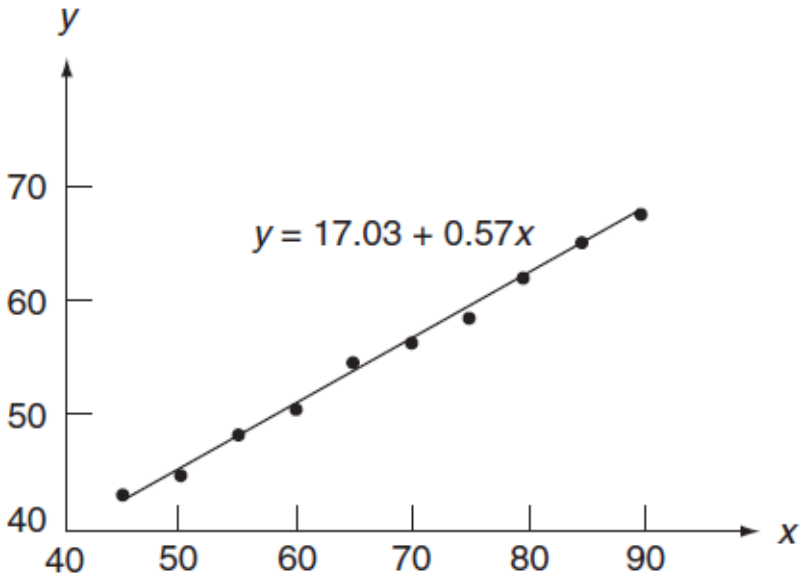
$$\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) = 1182,5.$$

Înlocuirea acestor valori în relaţiile (10.7) şi (10.8) dă

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} = \frac{1182,5}{2062,5} \simeq 0,57,$$

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} \simeq 55,5 - 0,57 \cdot 67,5 \simeq 17,03.$$

Deci dreapta de regresie estimată are ecuaţia $y = 17,03 + 0,57x$ şi este arătată împreună cu valorile selecţiei observate în figura următoare.



Relaţiile de regresie sunt valabile doar pentru intervalul valorilor x reprezentate de date. Astfel, dreapta de regresie estimată în exemplul 10.1 e valabilă doar pentru temperaturi între $45°C$ şi $90°C$. Extrapolarea rezultatului în afara acestui interval poate conduce la greşeli şi nu este valabilă în general.

Analiza regresiei liniare ca cea făcută în exemplul 10.1 este bazată pe presupunerea că adevărata relaţie între $E(Y)$ şi x este liniară. Dacă relaţia este neliniară sau inexistentă, regresia liniară produce rezultate fără sens, chiar dacă o linie dreaptă pare să dea o bună potrivire a datelor.

10.1.2 Proprietăţile estimatorilor prin metoda celor mai mici pătrate

Fie $\widehat{A}$ şi $\widehat{B}$, estimatorii prin metoda celor mai mici pătrate pentru α, respectiv β şi fie

$$\widehat{\mathbf{\Theta}} = \begin{pmatrix} \widehat{A} \\ \widehat{B} \end{pmatrix}. \tag{10.16}$$

Din relaţia (10.15) avem

$$\widehat{\mathbf{\Theta}} = \left(C^T C\right)^{-1} C^T \mathbf{Y}, \tag{10.17}$$

unde

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \tag{10.18}$$

şi $Y_j, j = 1, 2, ..., n$, sunt independente şi identic repartizate în concordanţă cu relaţia (10.4). Astfel, dacă scriem

$$\mathbf{Y} = C\boldsymbol{\theta} + \mathbf{E}, \tag{10.19}$$

atunci $\mathbf{E}$ este un vector aleator cu media 0 şi matricea de covarianţă $\mathbf{\Lambda} = \boldsymbol{\sigma}^2\mathbf{I}$, $\mathbf{I}$ fiind matricea unitate de ordinul n.

Din relaţiile (10.17) şi (10.19), avem

$$E\left(\widehat{\mathbf{\Theta}}\right) = \left(C^TC\right)^{-1}C^TE\left(\mathbf{Y}\right) = \left(C^TC\right)^{-1}C^T\left[C\boldsymbol{\theta}+E\left(\mathbf{E}\right)\right] = \left(C^TC\right)^{-1}C^TC\boldsymbol{\theta} = \boldsymbol{\theta} \tag{10.20}$$

De aici, estimatorii $\widehat{A}$ şi $\widehat{B}$ pentru α, respectiv β, sunt nedeplasaţi.

Matricea de covarianţă asociată cu $\widehat{\mathbf{\Theta}}$ este dată de, după cum rezultă din relaţia (10.17) şi lema 3.1, deoarece $\left(C^TC\right)^{-1}$ este simetrică,

$$cov\left(\widehat{\mathbf{\Theta}}\right) = E\left(\left(\widehat{\mathbf{\Theta}}-\boldsymbol{\theta}\right)\left(\widehat{\mathbf{\Theta}}-\boldsymbol{\theta}\right)^T\right) = \left(C^TC\right)^{-1}C^Tcov\left(\mathbf{Y}\right)C\left(C^TC\right)^{-1}.$$

Dar $cov\left(\mathbf{Y}\right) = \boldsymbol{\sigma}^2\mathbf{I}$; astfel, avem

$$cov\left(\widehat{\mathbf{\Theta}}\right) = \boldsymbol{\sigma}^2\left(C^TC\right)^{-1}C^TC\left(C^TC\right)^{-1} = \boldsymbol{\sigma}^2\left(C^TC\right)^{-1}. \tag{10.21}$$

Elementele de pe diagonala matricei din relaţia (10.21) dau dispersiile lui $\widehat{A}$ şi $\widehat{B}$. În termeni de elementele lui C, putem scrie

$$var\left(\widehat{A}\right) = \frac{\boldsymbol{\sigma}^2\sum_{i=1}^{n}x_i^2}{n\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}, \tag{10.22}$$

$$var\left(\widehat{B}\right) = \frac{\boldsymbol{\sigma}^2}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2} \tag{10.23}$$

Se vede că aceste dispersii descresc când mărimea selecţiei n creşte. Astfel, rezultă din capitolul 8 că aceşti estimatori sunt consistenţi. Pentru un n fixat, dispersia lui $\widehat{B}$ poate fi redusă selectând x_i-urile astfel încât numitorul relaţiei (10.23) este maximizat; asta poate fi făcut împrăştiind x_i-urile cât mai departe posibil. În exemplul 10.1, presupunând că suntem liberi să alegem valorile lui x_i, calitatea lui $\widehat{\beta}$ este îmbunătăţită dacă jumătate din citirile x sunt luate la o extremă a intervalului temperaturii şi cealaltă jumătate la cealaltă extremă. Strategia de selecţie pentru minimizarea $var\left(\widehat{A}\right)$ pentru un n fixat este de a face $\overline{x}$ pe cât posibil cât mai aproape de 0.

Sunt dispersiile date de relaţiile (10.22) şi (10.23) dispersii minime între dispersiile estimatorilor nedeplasaţi pentru α şi β? Un răspuns la această întrebare poate fi aflat comparând rezultatele date de relaţiile (10.22) şi (10.23) cu marginile inferioare Rao-Cramer definite în secţiunea 8.1.2. Pentru a evalua aceste margini, repartiţia lui Y trebuie să fie cunoscută. Totuşi, fără a cunoaşte repartiţia lui Y, se poate arăta că tehnica celor mai mici pătrate conduce la

estimatori liniari nedeplasaţi de dispersie minimă, adică, dintre toţi estimatorii nedeplasaţi care sunt *liniari* în $\mathbf{Y}$, estimatorii prin metoda celor mai mici pătrate au dispersie minimă.

Teorema 10.2. Fie variabila aleatoare Y definită în relaţia (10.4). Fiind dată o selecţie $(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)$ a lui Y cu valorile x asociate, estimatorii prin metoda celor mai mici pătrate $\widehat{A}$ şi $\widehat{B}$ daţi de relaţia (10.17) sunt *estimatori liniari nedeplasaţi de dispersie minimă* pentru α, respectiv β.

Demonstraţie. Considerăm un estimator liniar nedeplasat de forma

$$\mathbf{\Theta}^* = \left[\left(C^T C\right)^{-1} C^T + G\right] \mathbf{Y}. \tag{10.24}$$

Vrem să arătăm că dacă $\mathbf{\Theta}^*$ are dispersie minimă, atunci $G = 0$.

Datorită relaţiei (10.19), cerinţa de nedeplasare conduce la

$$GC = 0. \tag{10.25}$$

Considerăm acum matricea de covarianţă

$$cov\left(\mathbf{\Theta}^*\right) = E\left(\left(\mathbf{\Theta}^* - \boldsymbol{\theta}\right)\left(\mathbf{\Theta}^* - \boldsymbol{\theta}\right)^T\right). \tag{10.26}$$

Folosind relaţiile (10.19), (10.24), (10.25) şi lema 3.1, avem

$$cov\left(\mathbf{\Theta}^*\right) = \sigma^2\left[\left(C^T C\right)^{-1} + GG^T\right].$$

Pentru a minimiza dispersiile asociate cu componentele lui $\mathbf{\Theta}^*$, trebuie să minimizăm fiecare element diagonal al lui GG^T. Deoarece elementul diagonal ii al lui GG^T este dat de

$$\left(GG^T\right)_{ii} = \sum_{j=1}^{n} g_{ij}^2,$$

unde g_{ij} este elementul ij al lui G, trebuie să avem

$$g_{ij} = 0, \forall i, j$$

şi obţinem

$$G = 0. \tag{10.27}$$

□

Teorema de mai sus este un caz special al *teoremei Gauss-Markov*.

Altă comparaţie interesantă este cea dintre estimatorii prin metoda celor mai mici pătrate pentru α şi β şi estimatorii lor de verosimilitate maximă cu o repartiţie cunoscută pentru variabila aleatoare Y. Se poate arăta că estimatorii de verosimilitate maximă pentru α şi β sunt identici cu cei prin metoda celor mai mici pătrate sub ipoteza suplimentară că Y este repartizată normal.

11 Regresie liniară simplă. Estimarea parametrilor prin metoda celor mai mici pătrate. Intervale de încredere pentru parametrii de regresie

11.1 Estimator nedeplasat pentru σ^2

Metoda celor mai mici pătrate nu conduce la un estimator pentru dispersia σ^2 a lui Y, care este în general de asemenea o cantitate necunoscută în modelele de regresie liniară. O alegere intuitivă pentru un estimator pentru σ^2 este

$$\widehat{\Sigma^2} = k \sum_{i=1}^{n} \left[Y_i - \left(\widehat{A} + \widehat{B} x_i \right) \right]^2, \tag{11.1}$$

unde coeficientul k este ales astfel încât $\widehat{\Sigma^2}$ este nedeplasat. Pentru a calcula media lui $\widehat{\Sigma^2}$, observăm că (vezi relaţia (10.7))

$$Y_i - \widehat{A} - \widehat{B} x_i = Y_i - \left(\overline{Y} - \widehat{B} \overline{x} \right) - \widehat{B} x_i = \left(Y_i - \overline{Y} \right) - \widehat{B} \left(x_i - \overline{x} \right). \tag{11.2}$$

De aici rezultă

$$\sum_{i=1}^{n} \left(Y_i - \widehat{A} - \widehat{B} x_i \right)^2 = \sum_{i=1}^{n} \left(Y_i - \overline{Y} \right)^2 - \widehat{B}^2 \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2, \tag{11.3}$$

deoarece (vezi relaţia (10.8))

$$\sum_{i=1}^{n} \left(x_i - \overline{x} \right) \left(Y_i - \overline{Y} \right) = \widehat{B} \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2. \tag{11.4}$$

Se poate arăta că

$$E\left(\widehat{\Sigma^2} \right) = kE \left(\sum_{i=1}^{n} \left(Y_i - \overline{Y} \right)^2 - \widehat{B}^2 \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2 \right) = k \left(n - 2 \right) \sigma^2.$$

De aici, $\widehat{\Sigma^2}$ este nedeplasat pentru $k = \frac{1}{n-2}$, fiind dat de

$$\widehat{\Sigma^2} = \frac{1}{n-2} \sum_{i=1}^{n} \left[Y_i - \left(\widehat{A} + \widehat{B} x_i \right) \right]^2, \tag{11.5}$$

sau, datorită relaţiei (11.3),

$$\widehat{\Sigma^2} = \frac{1}{n-2} \left[\sum_{i=1}^{n} \left(Y_i - \overline{Y} \right)^2 - \widehat{B}^2 \sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2 \right]. \tag{11.6}$$

Exemplul 11.1. Pentru datele din exemplul 10.1, să se determine o estimare nedeplasată pentru σ^2.

Am obţinut în exemplul 10.1

$$\sum_{i=1}^{n}(x_i-\overline{x})^2=2062,5,$$

$$\widehat{\beta}=0,57.$$

În plus, obţinem

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=680,5.$$

Relaţia (11.6) dă

$$\widehat{\sigma^2}=\frac{1}{8}\left[680,5-0,57^2\cdot 2062,5\right]=1,3.$$

Exemplul 11.2. Se face un experiment pe elasticitatea ţesutului plămânului ca o funcţie de proprietăţile de extindere ale plămânului şi măsurătorile din tabelul următor sunt cele ale modulului lui Young al ţesutului (Y), în g/cm^2, la diferite valori ale extinderii plămânului în termeni de tensiune (x), în g/cm^2.

x	2	2.5	3	5	7	9	10	12	15	16	17	18	19	20
y	9.1	19.2	18.0	31.3	40.9	32.0	54.3	49.1	73.0	91.0	79.0	68.0	110.5	130.8

Presupunând că $E(Y)$ este legată liniar de x şi că $\sigma_Y^2=\sigma^2$ (o constantă), determinaţi estimările prin metoda celor mai mici pătrate pentru coeficienţii de regresie şi o estimare nedeplasată pentru σ^2.

În acest caz avem $n=14$. Cantităţile care ne interesează sunt

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i=\frac{1}{14}\left(2+2,5+...+20\right)=11,11,$$

$$\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{1}{14}\left(9,1+19,2+...+130,8\right)=57,59,$$

$$\sum_{i=1}^{n}(x_i-\overline{x})^2=546,09,$$

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=17179,54,$$

$$\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = 2862,12.$$

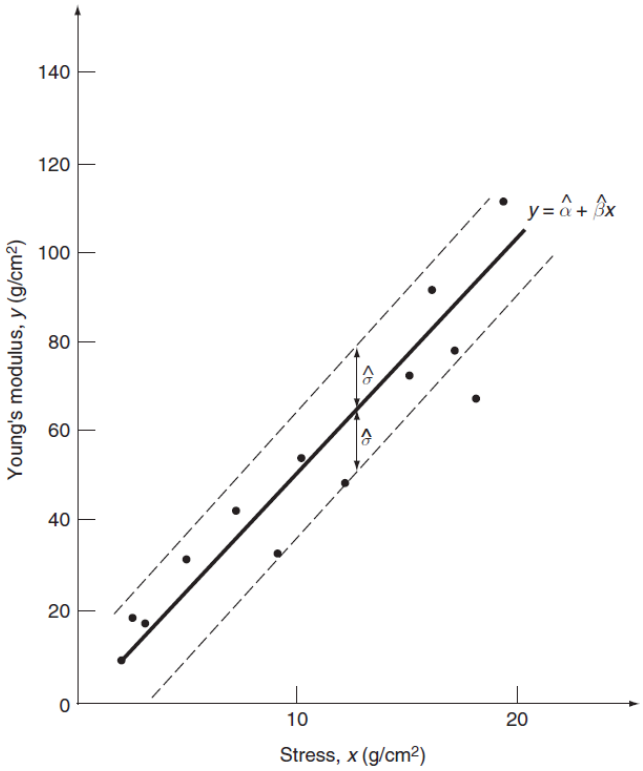
Înlocuirea acestor valori în relaţiile (10.8), (10.7) şi (11.6) dă

$$\widehat{\beta} = \frac{2862,12}{546,09} = 5,24,$$

$$\widehat{\alpha} = 57,59 - 5,24 \cdot 11,11 = -0,63$$

$$\widehat{\sigma^2} = \frac{1}{12}\left[17179,54 - 5,24^2 \cdot 546,09\right] = 182,1.$$

Dreapta de regresie estimată împreună cu datele sunt arătate în figura următoare.



Deviaţia standard estimată este $\widehat{\sigma} = \sqrt{182,1} = 13,49 g/cm^2$ şi σ-banda este de asemenea arătă în figură.

11.2 Intervale de încredere pentru coeficienţii de regresie

Presupunem că $Y \sim N\left(\alpha + \beta x, \sigma^2\right)$. Deoarece estimatorii $\widehat{A}, \widehat{B}$ şi $\widehat{A} + \widehat{B}x$ sunt funcţii liniare de selecţia Y, ei sunt de asemenea variabile aleatoare normale.

Când mărimea selecţiei n este mare, $\widehat{A}, \widehat{B}$ şi $\widehat{A}+\widehat{B}x$ au o repartiţie normală pe baza teoremei limită centrală, indiferent de repartiţia lui Y.

Urmăm calea din secţiunea 9.2 pentru a stabili limite de încredere. Se pot verifica următoarele rezultate:

i) Fie $\widehat{\Sigma^2}$ estimatorul nedeplasat pentru σ^2 definit de relaţia (11.6) şi fie

$$D = \frac{(n-2)\,\widehat{\Sigma^2}}{\sigma^2}. \tag{11.7}$$

Rezultă din secţiunea 9.2 că D este o variabilă aleatoare repartizată χ^2 cu $n-2$ grade de libertate.

ii) Considerăm variabilele aleatoare

$$\frac{\widehat{A}-\alpha}{\sqrt{\dfrac{\widehat{\Sigma^2}\displaystyle\sum_{i=1}^{n} x_i^2}{n\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})^2}}} \tag{11.8}$$

şi

$$\frac{\widehat{B}-\beta}{\sqrt{\dfrac{\widehat{\Sigma^2}}{\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})^2}}}, \tag{11.9}$$

unde, după cum se observă din relaţiile (10.20), (10.22) şi (10.23), α şi β sunt mediile lui $\widehat{A}$, respectiv $\widehat{B}$, iar numitorii sunt deviaţiile standard ale lui $\widehat{A}$, respectiv $\widehat{B}$, cu σ^2 estimat de $\widehat{\Sigma^2}$. Din secţiunea 9.2 rezultă că fiecare dintre aceste variabile aleatoare are o repartiţie t cu $n-2$ grade de libertate.

iii) Estimatorul $\widehat{E(Y)}$ pentru media lui Y este repartizat normal cu media $\alpha+\beta x$ şi dispersia

$$var\left(\widehat{E(Y)}\right) = var\left(\widehat{A}+\widehat{B}x\right) = var\left(\widehat{A}\right)+x^2 var\left(\widehat{B}\right)+2x cov\left(\widehat{A},\widehat{B}\right) = \sigma^2 \frac{\frac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^2 + x^2 - 2x\overline{x}}{\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})^2} =$$

$$\sigma^2\left[\frac{1}{n}+\frac{(x-\overline{x})^2}{\displaystyle\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]. \tag{11.10}$$

De aici, urmând argumentarea din secţiunea 9.2, variabila aleatoare

$$\frac{\widehat{E(Y)} - (\alpha + \beta x)}{\sqrt{\widehat{\Sigma^2}\left[\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\right]}} \tag{11.11}$$

este de asemenea t-repartizată cu $n-2$ grade de libertate.

Pe baza rezultatelor prezentate mai sus putem stabili limite de încredere pentru toţi parametrii de interes. Rezultatele de mai jos sunt o consecinţă directă a celor din secţiunea 9.2.

1) Un interval de $[100\,(1-\gamma)]\,\%$ încredere pentru α este determinat de (vezi relaţia (9.19))

$$L_{1,2} = \widehat{A} \mp t_{n-2,\gamma/2}\sqrt{\frac{\widehat{\Sigma^2}\sum\limits_{i=1}^{n}x_i^2}{n\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}}. \tag{11.12}$$

2) Un interval de $[100\,(1-\gamma)]\,\%$ încredere pentru β este determinat de (vezi relaţia (9.19))

$$L_{1,2} = \widehat{B} \mp t_{n-2,\gamma/2}\sqrt{\frac{\widehat{\Sigma^2}}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}}. \tag{11.13}$$

3) Un interval de $[100\,(1-\gamma)]\,\%$ încredere pentru $E(Y) = \alpha + \beta x$ este determinat de (vezi relaţia (9.19))

$$L_{1,2} = \widehat{E(Y)} \mp t_{n-2,\gamma/2}\sqrt{\widehat{\Sigma^2}\left[\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum\limits_{i=1}^{n}(x_i-\overline{x})^2}\right]} \tag{11.14}$$

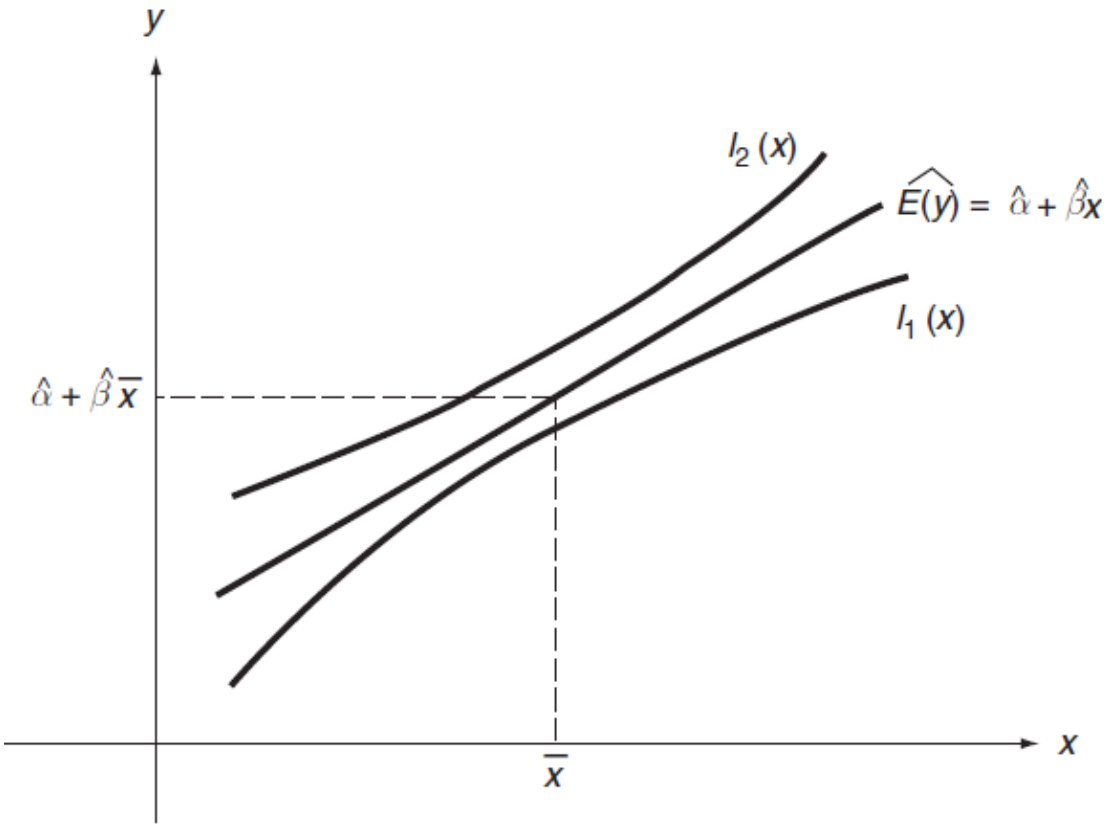
4) Un interval de $[100\,(1-\gamma)]\,\%$ încredere pentru σ^2 cu ambele margini finite este determinat de (vezi relaţia (9.22))

$$\begin{aligned} L_1 &= \frac{(n-2)\widehat{\Sigma^2}}{\chi^2_{n-2,\gamma/2}}, \\ L_2 &= \frac{(n-2)\widehat{\Sigma^2}}{\chi^2_{n-2,1-\gamma/2}}. \end{aligned} \tag{11.15}$$

Un interval de $[100\,(1-\gamma)]\,\%$ încredere pentru σ^2 cu o margine finită este determinat de (vezi relaţia (9.23))

$$L_1 = \frac{(n-2)\,\widehat{\Sigma^2}}{\chi^2_{n-1,\gamma}}. \tag{11.16}$$

În fiecare caz, atât poziţia cât şi mărimea intervalului variază de la selecţie la selecţie. În plus, intervalul de încredere pentru $\alpha + \beta x$ este o funcţie de x. Dacă se reprezintă valorile observate ale lui L_1 şi L_2, ele formează o *bandă de încredere* în jurul dreptei de regresie estimate, după cum este arătat în figura următoare.



Relaţia (11.14) arată că cea mai îngustă lăţime a bandei apare în $x = \overline{x}$. Banda devine mai largă când x de mişcă din $\overline{x}$ în orice direcţie.

Exemplul 11.3. În exemplul 11.2, presupunând că Y este repartizată normal, să se determine o bandă de 95% încredere pentru $\alpha + \beta x$.

Relaţia (11.14) dă limitele de încredere dorite, cu $n = 14, \gamma = 0,05$ şi

$$\widehat{E(Y)} = \widehat{\alpha} + \widehat{\beta}x = -0,63 + 5,24x,$$

$$t_{n-2,\gamma/2} = t_{12,0.025} = 2,179, \text{ din tabelul A.4,}$$

$$\overline{x} = 11,11,$$

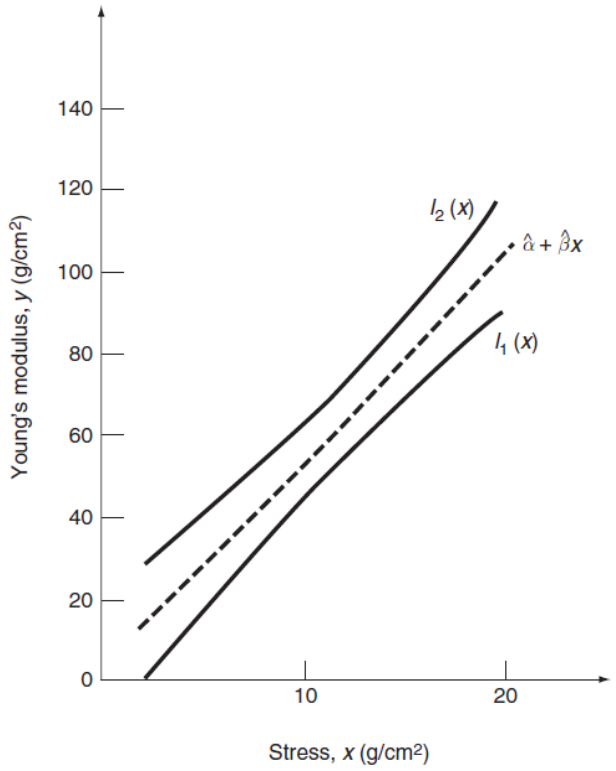
$$\sum_{i=1}^{n} (x_i - \overline{x})^2 = 546,09,$$

$$\widehat{\sigma^2} = 182,1.$$

Limitele de încredere observate sunt date de

$$l_{1,2} = -0,63 + 5,24x \mp 2,179\sqrt{182,1\left[\frac{1}{14} + \frac{(x-11,11)^2}{546,09}\right]}.$$

Rezultatul este arătat grafic în figura următoare.



12 Regresie liniară multiplă. Alte modele de regresie

12.1 Regresie liniară multiplă

În regresia liniară multiplă, modelul ia forma

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m. \tag{12.1}$$

Din nou, presupunem că dispersia lui Y este σ^2 şi este independentă de $x_1, x_2, ..., x_m$. Ca la regresia liniară simplă suntem interesaţi să estimăm cei $m+1$ coeficienţi de regresie $\beta_0, \beta_1, ..., \beta_m$ pe baza unei selecţii a valorilor lui Y cu valorile lor asociate ale lui $(x_1, x_2, ..., x_m)$. O selecţie de mărime n în acest caz ia forma $(x_{11}, x_{21}, ..., x_{m1}, Y_1), (x_{12}, x_{22}, ..., x_{m2}, Y_2), ..., (x_{1n}, x_{2n}, ..., x_{mn}, Y_n)$. Pentru fiecare set de valori $x_{ki}, k = 1, 2, ..., m$, Y_i este o observaţie independentă din populaţia Y definită prin

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + E. \tag{12.2}$$

Ca mai înainte, E este eroarea aleatoare, cu media 0 şi dispersia σ^2.

12.1.1 Estimarea prin metoda celor mai mici pătrate

Fiind date seturile de valori de selecţie observate $(x_{1i}, x_{2i}, ..., x_{mi}, y_i), i = 1, 2, ..., n$, sistemul de ecuaţii de regresie observate este

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_m x_{mi} + e_i, \quad i = 1, 2, ..., n. \tag{12.3}$$

Dacă notăm

$$C = \begin{pmatrix} 1 & x_{11} & x_{21} & ... & x_{m1} \\ 1 & x_{12} & x_{22} & ... & x_{m2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & ... & x_{mn} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

şi

$$\boldsymbol{\theta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix},$$

relaţia (12.3) poate fi scrisă

$$\mathbf{y} = C\boldsymbol{\theta} + \mathbf{e}. \tag{12.4}$$

Comparând relaţia (12.4) cu relaţia (10.12) din regresia liniară simplă, ecuaţiile de regresie observate în ambele cazuri sunt identice cu excepţia faptului că C este acum o matrice $n \times (m+1)$ şi $\boldsymbol{\theta}$ este un vector $(m+1)$-dimensional. Păstrând această diferenţă de dimensiune în minte, rezultatele obţinute în cazul

regresiei liniare simple bazate pe relaţia (10.12) sunt din nou valabile în cazul regresiei liniare multiple. Astfel, fără altă demonstraţie, avem pentru soluţia estimării prin metoda celor mai mici pătrate $\widehat{\boldsymbol{\theta}}$ a lui $\boldsymbol{\theta}$ (vezi relaţia (10.15))

$$\widehat{\boldsymbol{\theta}} = \left(C^T C\right)^{-1} C^T \mathbf{y}. \tag{12.5}$$

Existenţa matricei $\left(C^T C\right)^{-1}$ cere ca să existe cel puţin $(m+1)$ seturi distincte de valori $(x_{1i}, x_{2i}, ..., x_{mi})$ în selecţie. $C^T C$ este o matrice $(m+1) \times (m+1)$ simetrică.

Exemplul 12.1. Consumul lunar mediu de putere electrică (Y) într-o anumită fabrică este considerat a fi liniar dependent de temperatura mediului ambiant (x_1, în $^\circ F$) şi numărul de zile lucrătoare din lună (x_2). Considerăm datele lunare pe un an din tabelul următor.

x_1	20	26	41	55	60	67	75	79	70	55
x_2	23	21	24	25	24	26	25	25	24	25
y	210	206	260	244	271	285	270	265	234	241
			45	33						
			25	23						
			258	230						

Determinaţi estimările prin metoda celor mai mici pătrate ai coeficienţilor de regresie liniară asociaţi.

În acest caz, C este o matrice 12×3 şi

$$C^T C = \begin{pmatrix} 12 & 626 & 290 \\ 626 & 36776 & 15336 \\ 290 & 15336 & 7028 \end{pmatrix},$$

$$C^T \mathbf{y} = \begin{pmatrix} 2974 \\ 159011 \\ 72166 \end{pmatrix}.$$

Astfel avem

$$\widehat{\boldsymbol{\theta}} = \left(C^T C\right)^{-1} C^T \mathbf{y} = \begin{pmatrix} -33,84 \\ 0,39 \\ 10,80 \end{pmatrix},$$

sau

$$\widehat{\beta}_0 = -33,84, \quad \widehat{\beta}_1 = 0,39, \quad \widehat{\beta}_2 = 10,8.$$

Ecuaţia de regresie estimată bazată pe date este astfel

$$\widehat{E\left(y\right)} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 = -33,84 + 0,39x_1 + 10,8x_2.$$

Deoarece relaţia (12.4) coincide cu cea din cazul regresiei liniare simple, multe din rezultatele obţinute acolo privind proprietăţile estimatorilor prin metoda celor mai mici pătrate pot fi duplicate aici ţinând cont doar de noile definiţii ale matricei C şi vectorului $\boldsymbol{\theta}$.

Scriem estimatorul $\widehat{\boldsymbol{\Theta}}$ pentru $\boldsymbol{\theta}$ în forma

$$\widehat{\boldsymbol{\Theta}} = \left(C^T C\right)^{-1} C^T \mathbf{Y}. \tag{12.6}$$

Observăm că

$$E\left(\widehat{\boldsymbol{\Theta}}\right) = \left(C^T C\right)^{-1} C^T E\left(\mathbf{Y}\right) = \boldsymbol{\theta}. \tag{12.7}$$

De aici, estimatorul prin metoda celor mai mici pătrate $\widehat{\boldsymbol{\Theta}}$ este din nou nedeplasat. De asemenea, din relaţia (10.21) rezultă că matricea de covarianţă pentru $\widehat{\boldsymbol{\Theta}}$ este

$$cov\left(\widehat{\boldsymbol{\Theta}}\right) = \sigma^2 \left(C^T C\right)^{-1}. \tag{12.8}$$

Intervale de încredere pentru parametrii de regresie în acest caz pot fi de asemenea stabilite urmând proceduri similare celor din cazul regresiei liniare simple.

12.2 Alte modele de regresie

În ştiinţă şi inginerie e adesea necesar să se considere modele de regresie care sunt neliniare în variabilele independente. Exemple de aceste clase de modele sunt

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + E, \tag{12.9}$$

$$Y = \beta_0 \exp\left(\beta_1 x + E\right), \tag{12.10}$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + E, \tag{12.11}$$

$$Y = \beta_0 \beta_1^x + E. \tag{12.12}$$

Modele polinomiale ca relaţiile (12.9) sau (12.11) sunt încă modele de regresie liniară, deoarece sunt liniare în parametrii necunoscuţi $\beta_0, \beta_1, \beta_2, ...$ De aceea, parametrii pot fi estimaţi folosind tehnici de regresie liniară multiplă. Într-adevăr, punând $x_1 = x$ şi $x_2 = x^2$ în relaţia (12.9), aceasta ia forma unui model de regresie liniară multiplă cu 2 variabile independente şi poate fi analizat ca atare. O echivalenţă similară poate fi stabilită între relaţia (12.11) şi un model de regresie liniară multiplă cu 5 variabile independente.

Considerăm modelul exponenţial dat de relaţia (12.10). Logaritmând ambii membri, avem

$$\ln Y = \ln \beta_0 + \beta_1 x + E. \tag{12.13}$$

În termeni de variabila aleatoare $\ln Y$, relaţia (12.13) reprezintă o ecuaţie de regresie liniară cu coeficienţii de regresie $\ln \beta_0$ şi β_1. Tehnici de regresie liniară se aplică din nou în acest caz. Relaţia (12.12) nu poate fi pusă convenabil într-o formă de regresie liniară.

Exemplul 12.2. În medie, rata creşterii populaţiei (Y) asociată cu un oraş dat variază cu x, numărul de ani după 1970. Presupunând că

$$E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2,$$

calculaţi estimările prin metoda celor mai mici pătrate pentru β_0, β_1 şi β_2 pe baza datelor din tabelul următor.

x	0	1	2	3	4	5
y(%)	1.03	1.32	1.57	1.75	1.83	2.33

Fie $x_1 = x$, $x_2 = x^2$ şi

$$\boldsymbol{\theta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Estimarea prin metoda celor mai mici pătrate pentru $\boldsymbol{\theta}$ este dată de relaţia (12.5) cu

$$C = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{pmatrix}$$

şi

$$y = \begin{pmatrix} 1,03 \\ 1,32 \\ 1,57 \\ 1,75 \\ 1,83 \\ 2,33 \end{pmatrix}.$$

Astfel,

$$\widehat{\boldsymbol{\theta}} = \left(C^T C\right)^{-1} C^T y = \begin{pmatrix} 6 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{pmatrix}^{-1} \begin{pmatrix} 9,83 \\ 28,68 \\ 110,88 \end{pmatrix} = \begin{pmatrix} 1,07 \\ 0,2 \\ 0,01 \end{pmatrix},$$

sau

$$\widehat{\beta}_0 = 1,07, \quad \widehat{\beta}_1 = 0,2 \text{ şi } \widehat{\beta}_2 = 0,01.$$

Observăm în acest exemplu că, deoarece $x_2 = x_1^2$, în matricea C elementele de pe coloana a treia sunt pătratele elementelor corespunzătoare de pe coloana a doua. Pentru modele de regresie polinomială de ordin superior, situaţii de acest tip pot conduce la dificultăţi în inversarea matricei $C^T C$.

13 Estimatori neparametrici ai unei densităţi de repartiţie

13.1 Preliminarii

Fie $X_1, X_2, ..., X_n$ o selecţie independentă de mărime n dintr-o populaţie X cu o presupusă densitate $f(x; \boldsymbol{\theta})$ sau masă $p(x; \boldsymbol{\theta})$, unde $\boldsymbol{\theta}$ poate fi specificat sau nu. Notăm cu *ipoteza* H ipoteza că selecţia reprezintă n valori ale unei variabile aleatoare cu densitatea $f(x; \boldsymbol{\theta})$ sau masa $p(x; \boldsymbol{\theta})$. Această ipoteză este numită *ipoteză simplă* când repartiţia este complet specificată, adică valorile parametrului sunt specificate împreună cu forma funcţională a densităţii sau masei; altfel este o *ipoteză compusă*. Pentru a construi un criteriu pentru testarea ipotezelor este necesar să fie stabilită o ipoteză alternativă împotriva căreia ipoteza H poate fi testată. Un exemplu de ipoteză alternativă este altă repartiţie presupusă sau, ca alt exemplu, ipoteza H poate fi testată împotriva ipotezei alternative că ipoteza H nu este adevărată. Vom considera în continuare ultima alegere.

13.1.1 Erori de tipul I şi de tipul II

Ca şi în estimarea parametrilor, erorile sau riscurile sunt inerente în decizia dacă o ipoteză H ar trebui să fie acceptată sau respinsă pe baza informaţiei dată de selecţie.

Definiţia 13.1. În testarea ipotezei H, o eroare de tipul I e comisă când H este respinsă când de fapt H este adevărată; o eroare de tipul II este comisă când H este acceptată când de fapt H este falsă.

În cele ce urmează, metodele de testare a ipotezelor sunt discutate doar pe baza erorilor de tipul I.

13.2 Testul chi-pătrat al bunătăţii de potrivire

Problema este testarea ipotezei H care specifică repartiţia unei populaţii X comparată cu alternativa că repartiţia lui X nu este de tipul specificat pe baza unei selecţii de mărime n din populaţia X. Testul chi-pătrat (χ^2) al bunătăţii de potrivire a fost elaborat pentru acest scop de Pearson în 1900.

13.2.1 Cazul parametrilor cunoscuţi

Presupunem mai întâi că repartiţia presupusă este complet specificată, fără parametri necunoscuţi. Pentru a testa ipoteza H este folosită o statistică $h(X_1, X_2, ..., X_n)$ a selecţiei, care dă o măsură a deviaţiei repartiţiei observate construită din selecţie de la repartiţia presupusă.

În testul χ^2, statistica utilizată este legată de diferenţa dintre diagrama frecvenţelor construită din selecţie şi o diagramă corespunzătoare construită din repartiţia presupusă. Se împarte spaţiul valorilor lui X în k intervale disjuncte

2 câte 2 $A_1, A_2, ..., A_k$ şi fie N_i numărul X_j-urilor din $A_i, i = 1, 2, ..., k$. Atunci probabilităţile *observate* $P(A_i)$ sunt date de

$$P(A_i) \text{ observată } = \frac{N_i}{n}, \quad i = 1, 2, ..., k. \tag{13.1}$$

Probabilităţile *teoretice* $P(A_i)$ pot fi obţinute din repartiţia presupusă a populaţiei. Le notăm cu

$$P(A_i) \text{ teoretică } = p_i, \quad i = 1, 2, ..., k. \tag{13.2}$$

O alegere logică a statisticii dând o măsură a deviaţiei este

$$\sum_{i=1}^{k} c_i \left(\frac{N_i}{n} - p_i \right)^2, \tag{13.3}$$

care este o măsură naturală a deviaţiei de tipul celor mai mici pătrate. Pearson (1900) a arătat că, dacă luăm coeficientul $c_i = \frac{n}{p_i}$, statistica definită prin expresia (13.3) are proprietăţi convenabile. De aici, alegem ca măsura deviaţiei

$$D = \sum_{i=1}^{k} \frac{n}{p_i} \left(\frac{N_i}{n} - p_i \right)^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{N_i^2}{np_i} - n. \tag{13.4}$$

D este o statistică deoarece este o funcţie de N_i, care sunt, la rândul lor, funcţii de selecţia $X_1, X_2, ..., X_n$. Repartiţia statisticii D este dată în teorema următoare, datorată lui Pearson (1900).

Teorema 13.1. Presupunând că ipoteza H este adevărată, repartiţia lui D definită de relaţia (13.4) tinde la o repartiţie chi-pătrat cu $(k-1)$ grade de libertate când $n \to \infty$. Densitatea ei este dată de (vezi relaţia (5.24))

$$f_D(d) = \begin{cases} \frac{d^{\frac{k-3}{2}} e^{-\frac{d}{2}}}{2^{\frac{k-1}{2}} \Gamma\left(\frac{k-1}{2}\right)}, \text{ pentru } d \geq 0; \\ 0, \text{ altfel.} \end{cases} \tag{13.5}$$

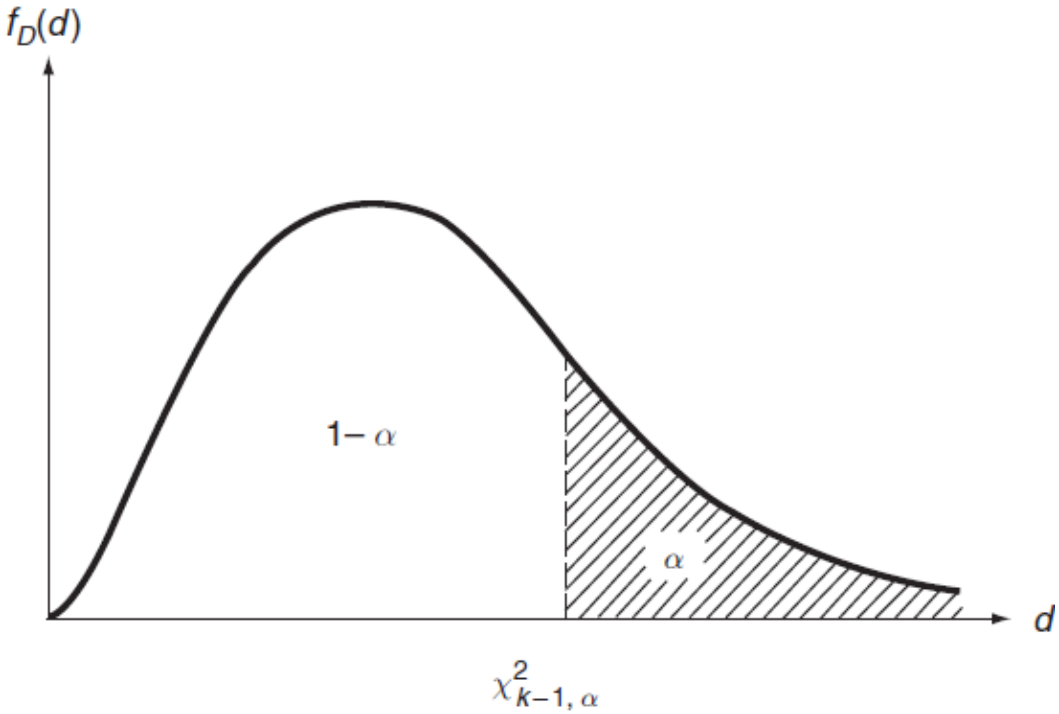
Această repartiţie este independentă de repartiţia presupusă.□

Cu ajutorul teoremei 13.1 poate fi construit un test al ipotezei H considerată mai sus pe baza atribuirii unei probabilităţi erorii de tipul I. Presupunem că vrem să obţinem probabilitatea α pentru eroarea de tipul I. Testul χ^2 sugerează că ipoteza H este respinsă ori de câte ori

$$d = \sum_{i=1}^{k} \frac{n_i^2}{np_i} - n > \chi^2_{k-1,\alpha}, \tag{13.6}$$

şi este acceptată altfel, unde d este valoarea de selecţie a lui D bazată pe valorile de selecţie $x_i, i = 1, ..., n$ şi $\chi^2_{k-1,\alpha}$ este valoarea astfel încât (vezi figura următoare)

$$P\left(D > \chi^2_{k-1,\alpha}\right) = \alpha. \tag{13.7}$$



Deoarece D are o repartiţie chi-pătrat cu $(k-1)$ grade de libertate pentru n mare, o valoare aproximativă pentru $\chi^2_{k-1,\alpha}$ poate fi găsită din tabelul A.5 din cursul 9 pentru repartiţia χ^2 când α este specificat.

Probabilitatea α a erorii de tipul I se numeşte *nivel de semnificaţie* în acest context. După cum vedem din figura de mai sus, reprezintă aria de sub f_D de la dreapta lui $\chi^2_{k-1,\alpha}$. Punând $\alpha = 0,05$, de exemplu, criteriul dat de relaţia (13.6) implică faptul că respingem ipoteza H ori de câte ori măsura deviaţiei d calculată dintr-un set dat de valori de selecţie pică în regiunea de 5%. Cu alte cuvinte, ne aşteptăm să respingem H în aproximativ 5% din timpul când de fapt H este adevărată. Ce nivel de semnificaţie trebuie adoptat într-o situaţie dată depinde de cazul particular implicat. În practică, valori obişnuite pentru α sunt $0,001$, $0,01$ şi $0,05$; o valoare a lui α între 1% şi 5% este *aproape semnificativă*; o valoare între $0,1\%$ şi 1% este *semnificativă*, iar o valoare sub $0,1\%$ este *înalt semnificativă*.

Dăm acum procedura pas cu pas pentru efectuarea testului χ^2 când repartiţia unei populaţii X este complet specificată.

- Pasul 1: împarte spaţiul de valori ale lui X în k intervale convenabile numeric şi disjuncte 2 câte 2 $A_i, i = 1, 2, ..., k$. Fie n_i, numărul de valori ale selecţiei din A_i. De regulă, dacă $n_i < 5$, combină intervalul A_i cu A_{i-1} sau A_{i+1}.

- Pasul 2: calculează probabilităţile teoretice $P(A_i) = p_i, i = 1, 2, ..., k$, cu ajutorul repartiţiei presupuse.
- Pasul 3: construieşte d dat de relaţia (13.6).
- Pasul 4: alege o valoare a lui α şi determină din tabelul A.5 pentru repartiţia χ^2 cu $(k-1)$ grade de libertate valoarea lui $\chi^2_{k-1,\alpha}$.
- Pasul 5: respinge ipoteza H dacă $d > \chi^2_{k-1,\alpha}$. Altfel, acceptă H.

Exemplul 13.1. 300 de becuri sunt testate pentru timpul de ardere t (în ore) şi rezultatul este arătat în tabelul următor.

Burning time, t	Number
$t < 100$	121
$100 \leq t < 200$	78
$200 \leq t < 300$	43
$300 \leq t$	58
	$n = 300$

Presupunem că timpul aleator de ardere este repartizat exponenţial cu timpul mediu de ardere $\frac{1}{\lambda} = 200$ de ore, adică $\lambda = 0,005$ pe oră şi

$$f_T(t) = 0,005e^{-0,005t}, \quad t \geq 0. \tag{13.8}$$

Testaţi această ipoteză folosind testul χ^2 la nivelul de semnificaţie de 5%.

Paşii necesari în efectuarea testului χ^2 sunt indicaţi în tabelul următor.

Interval, A_i	n_i	p_i	np_i	n_i^2/np_i
$t < 100$	121	0.39	117	125.1
$100 \leq t < 200$	78	0.24	72	84.5
$200 \leq t < 300$	43	0.15	45	41.1
$300 \leq t$	58	0.22	66	51.0
	300	1.00	300	301.7

Prima coloană dă intervalele A_i, care sunt alese în acest caz să fie intervalele pentru t date în tabelul anterior. Probabilităţile teoretice $P(A_i) = p_i$ din coloana a treia sunt calculate folosind relaţia (13.8). De exemplu,

$$p_1 = P(A_1) = \int_0^{100} 0,005e^{-0,005t}dt = -e^{-0,005t}|_0^{100} = 1 - e^{-0,5} = 0,39;$$

$$p_2 = P(A_2) = \int_{100}^{200} 0,005e^{-0,005t}dt = -e^{-0,005t}|_{100}^{200} = e^{-0,5} - e^{-1} = 0,24.$$

Numerele teoretice de apariţii prezise de model sunt date în coloana a patra a tabelului, care, când este comparată cu coloana a doua, dă o măsură a bunătăţii de potrivire a modelului cu datele. Coloana 5 ($\frac{n_i^2}{np_i}$) este inclusă pentru a facilita calcularea lui d. Astfel, din relaţia (13.6) avem

$$d = \sum_{i=1}^{k} \frac{n_i^2}{np_i} - n = 301,7 - 300 = 1,7.$$

$k = 4$. Din tabelul A.5 pentru repartiţia χ^2 cu 3 grade de libertate (vezi cursul 9) găsim

$$\chi^2_{3,0.05} = 7,815.$$

Deoarece $d < \chi^2_{3,0.05}$, acceptăm la un nivel de semnificaţie de 5% ipoteza că datele observate reprezintă o selecţie dintr-o repartiţie exponenţială cu $\lambda = 0,005$.

Exemplul 13.2. Înregistrarea numărului de accidente din 6 ani ale 7842 de şoferi californieni este dată în tabelul următor.

Number of accidents	Number of drivers
0	5147
1	1859
2	595
3	167
4	54
5	14
> 5	6
	Total = 7842

Pe baza acestor valori de selecţie, testaţi ipoteza că X, numărul de accidente din 6 ani pe şofer, este repartizat Poisson cu rata medie $\nu = 0,08$ pe an la nivelul de semnificaţie de 1%.

Deoarece X este discretă, alegem intervalele A_i ca în prima coloană a tabelului următor.

Interval, A_i	n_i	p_i	np_i	n_i^2/np_i
$x \leq 0$	5147	0.6188	4853	5459
$0 < x \leq 1$	1859	0.2970	2329	1484
$1 < x \leq 2$	595	0.0713	559	633
$2 < x \leq 3$	167	0.0114	89	313
$3 < x \leq 4$	54	0.0013	10	292
$4 < x \leq 5$	14	0.0001	1	196
$5 < x$	6	0.0001	1	36
	7842	1.0	7842	8413

Intervalul $x > 5$ ar fi fost combinat cu $4 < x \leq 5$ dacă numărul n_7 ar fi fost mai mic decât 5.

Repartiţia presupusă pentru X este

$$p_X(x) = \frac{(\nu t)^x e^{-\nu t}}{x!} = \frac{(0,48)^x e^{-0,48}}{x!}, \quad x = 0, 1, 2, ... \tag{13.9}$$

Astfel avem

$$p_i = P(A_i) = \frac{(0,48)^{i-1} e^{-0,48}}{(i-1)!}, \quad i = 1, 2, ..., 6.$$

$$p_7 = P(A_7) = 1 - \sum_{i=1}^{6} p_i.$$

Aceste valori sunt indicate în coloana a treia a tabelului.

Coloana 5 a tabelului dă

$$d = \sum_{i=1}^{k} \frac{n_i^2}{np_i} - n = 8413 - 7842 = 571.$$

Cu $k = 7$, valoarea lui $\chi^2_{k-1,\alpha} = \chi^2_{6,0.01}$ este găsită din tabelul A.5

$$\chi^2_{6,0.01} = 16,812.$$

Deoarece $d > \chi^2_{6,0.01}$, ipoteza este respinsă la nivelul de semnificaţie 1%.

13.2.2 Cazul parametrilor estimaţi

Considerăm acum situaţia în care parametrii din repartiţia presupusă trebuie de asemenea estimaţi din date. O procedură pentru un test de bunătatea potrivirii

în acest caz este întâi estimarea parametrilor şi apoi efectuarea testului χ^2 pentru parametri cunoscuţi. Apare o complicaţie în aceea că probabilităţile teoretice p_i definite de relaţia (13.2) sunt, fiind funcţii de parametrii repartiţiei, funcţii de selecţie. Statistica D ia acum forma

$$D = \sum_{i=1}^{k} \frac{n}{\widehat{P}_i} \left(\frac{N_i}{n} - \widehat{P}_i \right)^2 = \sum_{i=1}^{k} \frac{N_i^2}{n\widehat{P}_i} - n, \tag{13.10}$$

unde $\widehat{P}_i$ este un estimator pentru p_i şi astfel o statistică. D este acum o funcţie mult mai complicată de $X_1, X_2, ..., X_n$. Care este noua repartiţie a lui D?

Problema determinării repartiţiei limită a lui D în această situaţie a fost mai întâi considerată de Fischer (1922, 1924), care a arătat că, când $n \to \infty$, repartiţia lui D trebuie modificată, şi modificarea depinde de metoda de estimare a parametrilor folosită. Din fericire, pentru o clasă de metode importante de estimare, ca matoda verosimilităţii maxime, modificarea cerută este simplă, şi anume, statistica D tinde la o repartiţie chi pătrat când $n \to \infty$, dar acum cu $(k - r - 1)$ grade de libertate, unde r este numărul de parametri de estimat din repartiţia presupusă. Cu alte cuvinte, este doar necesar de a reduce numărul de grade de libertate în repartiţia limită definită de relaţia (13.5) cu unu pentru *fiecare* parametru estimat din selecţie.

Dăm acum procedura pas cu pas pentru cazul când r parametri din repartiţie trebuie estimaţi din date.

- Pasul 1: împarte spaţiul de valori ale lui X în k intervale convenabile numeric şi disjuncte 2 câte 2 $A_i, i = 1, 2, ..., k$. Fie n_i, numărul de valori ale selecţiei din A_i. De regulă, dacă $n_i < 5$, combină intervalul A_i cu A_{i-1} sau A_{i+1}.

- Pasul 2: estimează cei r parametri prin metoda verosimilităţii maxime din date.

- Pasul 3: calculează probabilităţile teoretice $P(A_i) = p_i, i = 1, 2, ..., k$, cu ajutorul repartiţiei presupuse cu valorile parametrilor estimate.

- Pasul 4: construieşte d dat de relaţia (13.6).

- Pasul 5: alege o valoare a lui α şi determină din tabelul A.5 pentru repartiţia χ^2 cu $(k - r - 1)$ grade de libertate valoarea lui $\chi^2_{k-r-1,\alpha}$. Se presupune că $k - r - 1 > 0$.

- Pasul 6: respinge ipoteza H dacă $d > \chi^2_{k-r-1,\alpha}$. Altfel, acceptă H.

Exemplul 13.3. Sosirile vehiculelor într-un anumit punct au fost înregistrate. Numerele de vehicule sosite în interval de un minut au fost luate pentru 106 minute şi sunt date în tabelul următor.

Vehicles per minute (No.)	Number of occurrences
0	0
1	0
2	1
3	3
4	5
5	7
6	13
7	12
8	8
9	9
10	13
11	10
12	5
13	6
14	4
15	5
16	4
17	0
18	1
	$n = 106$

Pe baza acestor observaţii, determinaţi dacă o repartiţie Poisson este potrivită pentru X, numărul de sosiri pe minut, la nivelul de semnificaţie de 5%.

Repartiţia presupusă este

$$p_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, ..., \tag{13.11}$$

unde parametrul λ trebuie să fie estimat din date. Astfel, $r = 1$.

Întâi determinăm intervale corespunzătoare A_i astfel încât $n_i \geq 5, \forall i$; acestea sunt arătate în prima coloană a tabelului următor.

Interval, A_i	n_i	p_i	np_i	n_i^2/np_i
$0 \leq x < 5$	9	0.052	5.51	14.70
$5 \leq x < 6$	7	0.058	6.15	7.97
$6 \leq x < 7$	13	0.088	9.33	18.11
$7 \leq x < 8$	12	0.115	12.19	11.81
$8 \leq x < 9$	8	0.131	13.89	4.61
$9 \leq x < 10$	9	0.132	13.99	5.79
$10 \leq x < 11$	13	0.120	12.72	13.29
$11 \leq x < 12$	10	0.099	10.49	9.53
$12 \leq x < 13$	5	0.075	7.95	3.14
$13 \leq x < 14$	6	0.054	5.72	6.29
$14 \leq x$	14	0.076	8.06	24.32
	106	1.0	106	119.56

De aici $k = 11$.

Estimarea de verosimilitate maximă pentru λ este dată de

$$\widehat{\lambda} = \overline{x} = \frac{1}{n} \sum_{j=1}^{n} x_j = 9,09.$$

Înlocuirea acestei valori pentru parametrul λ în relaţia (13.11) ne permite să calculăm probabilităţile $P(A_i) = p_i$. De exemplu,

$$p_1 = \sum_{j=0}^{4} p_X(j) = 0,052,$$

$$p_2 = p_X(5) = 0,058.$$

Aceste probabilităţi teoretice sunt date în a treia coloană a tabelului.

Din coloana 5 a tabelului obţinem

$$d = \sum_{i=1}^{k} \frac{n_i^2}{np_i} - n = 119,56 - 106 = 13,56.$$

Tabelul A.5 cu $\alpha = 0,05$ şi $k - r - 1 = 9$ grade de libertate dă

$$\chi^2_{9,0.05} = 16,919.$$

Deoarece $d < \chi^2_{9,0.05}$, repartiţia presupusă cu $\lambda = 9,09$ este acceptată la nivelul de semnificaţie de 5%.

Exemplul 13.4. Pe baza datelor căderilor de zăpadă dintre 1909 şi 1979 din tabelul următor, testaţi ipoteza: căderea anuală de zăpadă poate fi modelată printr-o repartiţie normală la nivelul de semnificaţie de 5%.

Year	Snowfall	Year	Snowfall	Year	Snowfall
1909–1910	126.4	1939–1940	77.8	1969–1970	120.5
1910–1911	82.4	1940–1941	79.3	1970–1971	97.0
1911–1912	78.1	1941–1942	89.6	1971–1972	109.9
1912–1913	51.1	1942–1943	85.5	1972–1973	78.8
1913–1914	90.9	1943–1944	58.0	1973–1974	88.7
1914–1915	76.2	1944–1945	120.7	1974–1975	95.6
1915–1916	104.5	1945–1946	110.5	1975–1976	82.5
1916–1917	87.4	1946–1947	65.4	1976–1977	199.4
1917–1918	110.5	1947–1948	39.9	1977–1978	154.3
1918–1919	25.0	1948–1949	40.1	1978–1979	97.3
1919–1920	69.3	1949–1950	88.7	1979–1980	68.4
1920–1921	53.5	1950–1951	71.4	1980–1981	60.9
1921–1922	39.8	1951–1952	83.0	1981–1982	112.4
1922–1923	63.6	1952–1953	55.9	1982–1983	52.4
1923–1924	46.7	1953–1954	89.9	1983–1984	132.5
1924–1925	72.9	1954–1955	84.6	1984–1985	107.2
1925–1926	74.6	1955–1956	105.2	1985–1986	114.7
1926–1927	83.6	1956–1957	113.7	1986–1987	67.5
1927–1928	80.7	1957–1958	124.7	1987–1988	56.4
1928–1929	60.3	1958–1959	114.5	1988–1989	67.4
1929–1930	79.0	1959–1960	115.6	1989–1990	93.7
1930–1931	74.4	1960–1961	102.4	1990–1991	57.5
1931–1932	49.6	1961–1962	101.4	1991–1992	92.8
1932–1933	54.7	1962–1963	89.8	1992–1993	93.2
1933–1934	71.8	1963–1964	71.5	1993–1994	112.7
1934–1935	49.1	1964–1965	70.9	1994–1995	74.6
1935–1936	103.9	1965–1966	98.3	1995–1996	141.4
1936–1937	51.6	1966–1967	55.5	1996–1997	97.6
1937–1938	82.4	1967–1968	66.1	1997–1998	75.6
1938–1939	83.6	1968–1969	78.4	1998–1999	100.5

Repartiţia asumată pentru X, căderea anuală de zăpadă, este $N\left(m,\sigma^2\right)$ unde m şi σ^2 trebuie să fie esimate din date. Doarece estimatorii de verosimilitate maximă pentru m şi σ^2 sunt $\widehat{M}=\overline{X}$, respectiv $\widehat{\Sigma^2}=\frac{n-1}{n}S^2$, avem

$$\widehat{m}=\overline{x}=\frac{1}{70}\sum_{j=1}^{70}x_j=83,6,$$

$$\widehat{\sigma^2}=\frac{69}{70}s^2=\frac{1}{70}\sum_{j=1}^{70}\left(x_j-83,6\right)^2=777,4.$$

Cu intervalele A_i definite după cum se arată în prima coloană a tabelului următor, probabilităţile teoretice $P\left(A_i\right)$ pot fi calculate cu ajutorul tabelului A.3.

Interval, A_i	n_i	p_i	np_i	n_i^2/np_i
$x \leq 56$	13	0.161	11.27	15.00
$56 < x \leq 72$	10	0.178	12.46	8.03
$72 < x \leq 88$	20	0.224	15.68	25.51
$88 < x \leq 104$	13	0.205	14.35	11.78
$104 < x \leq 120$	8	0.136	9.52	6.72
$120 < x$	6	0.096	6.72	5.36
	70	1.0	70	72.40

De exemplu, primele două dintre aceste probabilităţi sunt
$P(A_1) = P(X \leq 56) = P\left(U \leq \frac{56-83,6}{\sqrt{777,4}}\right) = F_U(-0,99) = 1 - F_U(0,99) = 1 - 0,8389 \simeq 0,161$;
$P(A_2) = P(56 < X \leq 72) = P(-0,99 < U \leq -0,416) = [1 - F_U(0,416)] - [1 - F_U(0,99)] = 0,339 - 0,161 = 0,178$.

Informaţia dată mai sus ne permite să construim restul tabelului. De aici, avem

$$d = \sum_{i=1}^{k} \frac{n_i^2}{np_i} - n = 72,4 - 70 = 2,4.$$

Numărul de grade de libertate în acest caz este $k-r-1 = 6-2-1 = 3$. Tabelul A.5 dă astfel

$$\chi^2_{3,0.05} = 7,815.$$

Deoarece $d < \chi^2_{3,0.05}$ repartiţia normală $N(83.6, 777.4)$ este acceptabilă la nivelul de semnificaţie de 5%.

Statistica D din testul χ^2 este repartizată χ^2 doar când $n \to \infty$. Astfel, testul χ^2 este un test de *selecţie mare*. Ca regulă, $n > 50$ este considerat satisfăcător pentru a îndeplini cerinţa de selecţie mare.

13.3 Testul Kolmogorov-Smirnov

Aşa-numitul *test de bunătatea potrivirii Kolmogorov-Smirnov*, prescurtat testul K-S, este bazat pe o statistică măsurând deviaţia *histogramei cumulative* observate de la funcţia de repartiţie cumulativă presupusă.

Fiind dat un set de valori de selecţie $x_1, x_2, ..., x_n$ observate dintr–o populaţie X, o histogramă cumulativă poate fi construită prin

(a) aranjarea valorilor de selecţie în ordine crescătoare, notate cu $x_{(1)}, x_{(2)}, ..., x_{(n)}$,

(b) determinarea funcţiei de repartiţie observată a lui X în $x_{(1)}, x_{(2)}, ...$, notate prin $F^0\left(x_{(1)}\right), F^0\left(x_{(2)}\right), ...$, din relaţiile $F^0\left(x_{(i)}\right) = \frac{i}{n}$ şi

(c) conectarea valorilor lui $F^0\left(x_{(i)}\right)$ prin segmente de dreaptă.

Statistica test folosită în acest caz este

$$D_2 = \max_{i=1}^{n} \left\{ \left| F^0\left(X_{(i)}\right) - F_X\left(X_{(i)}\right) \right| \right\} = \max_{i=1}^{n} \left\{ \left| \frac{i}{n} - F_X\left(X_{(i)}\right) \right| \right\}, \quad (13.12)$$

unde $X_{(i)}$ este a i-a statistică de ordine a selecţiei. Statistica D_2 măsoară astfel maximul valorilor absolute ale celor n diferenţe dintre funcţia de repartiţie observată şi funcţia de repartiţie presupusă evaluată pentru selecţia observată. În cazul când trebuie estimaţi parametri din repartiţia presupusă, valorile $F_X\left(X_{(i)}\right)$ sunt obţinute folosind estimările.

În timp ce repartiţia lui D_2 este dificil de obţinut analitic, funcţia ei de repartiţie poate fi calculată numeric şi tabelată. Se poate arăta că repartiţia lui D_2 este independentă de repartiţia presupusă şi este o funcţie doar de n, mărimea selecţiei.

Efectuarea testului K-S este asemănătoare cu cea a testului χ^2. La un nivel de semnificaţie specificat α, regula este să se respingă ipoteza H dacă $d_2 > c_{n,\alpha}$; altfel, se acceptă H. Aici, d_2 este valoarea de selecţie a lui D_2, şi valoarea $c_{n,\alpha}$ este definită prin

$$P\left(D_2 > c_{n,\alpha}\right) = \alpha. \quad (13.13)$$

Valorile $c_{n,\alpha}$ pentru $\alpha = 0.01, 0.05$ şi 0.1 sunt date în tabelul A.6 următor ca funcţii de n.

Table A.6 D_2 distribution with sample size n: a table of $c_{n,\alpha}$ in $P(D_2 > c_{n,\alpha}) = \alpha$, for $\alpha = 0.01$ to 0.10, $n = 5, 10, \ldots$

n	α		
	0.10	0.05	0.01
5	0.51	0.56	0.67
10	0.37	0.41	0.49
15	0.30	0.34	0.40
20	0.26	0.29	0.35
25	0.24	0.26	0.32
30	0.22	0.24	0.29
40	0.19	0.21	0.25
Large n	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Sunt şi diferenţe importante între acest test şi testul χ^2. În timp ce testul χ^2 este un test de selecţie mare, testul K-S este valabil pentru toate valorile lui n. Mai departe, testul K-S foloseşte valorile selecţiei în forma lor nealterată şi neagregată, în timp ce prelucrarea datelor este necesară în efectarea testului χ^2. Ca dezavantaje, testul K-S este valabil doar pentru repartiţii continue. De asemenea, valorile lui $c_{n,\alpha}$ din tabelul A.6 sunt bazate pe o repartiţie presupusă specificată complet. Când valorile parametrilor trebuie estimate, nu este disponibilă o metodă riguroasă de ajustare. În aceste cazuri poate fi specificat doar că valorile lui $c_{n,\alpha}$ trebuie reduse puţin.

Procedura pas cu pas pentru efectuarea testului K-S este:

- Pasul 1: rearanjează valorile selecţiei $x_1, x_2, ..., x_n$ în ordine crescătoare şi notează-le $x_{(1)}, x_{(2)}, ..., x_{(n)}$.
- Pasul 2: determină funcţia de repartiţie observată $F^0(x)$ în fiecare $x_{(i)}$ folosind $F^0\left(x_{(i)}\right) = \frac{i}{n}$.
- Pasul 3: determină funcţia de repartiţie teoretică $F_X(x)$ în fiecare $x_{(i)}$ folosind repartiţia presupusă. Parametrii repartiţiei sunt estimaţi din date dacă este necesar.
- Pasul 4: formează diferenţele $\left|F^0\left(x_{(i)}\right) - F_X\left(x_{(i)}\right)\right|$ pentru $i = 1, 2, ..., n$.
- Pasul 5: calculează

$$d_2 = \max_{i=1}^{n} \left\{\left|F^0\left(x_{(i)}\right) - F_X\left(x_{(i)}\right)\right|\right\}.$$

  Determinarea acestei valori maxime cere enumerarea a n cantităţi. Această muncă poate fi redusă puţin reprezentând grafic $F^0(x)$ şi $F_X(x)$ ca funcţii de x şi notând locaţia maximului prin inspecţie.
- Pasul 6: alege valoarea lui α şi determină din tabelul A.6 valoarea lui $c_{n,\alpha}$.
- Pasul 7: respinge ipoteza H dacă $d_2 > c_{n,\alpha}$. Altfel, acceptă H.

Exemplul 13.5. Se fac 10 măsurători ale rezistenţei la întindere a unui material. Ele sunt $30.1, 30.5, 28.7, 31.6, 32.5, 29, 27.4, 29.1, 33.5$ şi 31. Pe baza acestui set de date, testaţi ipoteza că rezistenţa la întindere are o repartiţie normală la nivelul de semnificaţie de 5%.

Reordonarea datelor dă $x_{(1)} = 27.4, x_{(2)} = 28.7, ..., x_{(10)} = 33.5$. Avem

$$F^0(27.4) = 0.1, F^0(28.7) = 0.2, ..., F^0(33.5) = 1.$$

Cu privire la funcţia de repartiţie teoretică, sunt mai întâi obţinute estimările mediei şi dispersiei

$$\widehat{m} = \overline{x} = \frac{1}{10}\sum_{j=1}^{10} x_j = 30.3.$$

$$\widehat{\sigma^2} = \left(\frac{n-1}{n}\right) s^2 = \frac{1}{10}\sum_{j=1}^{10} (x_j - 30.3)^2 = 3.14.$$

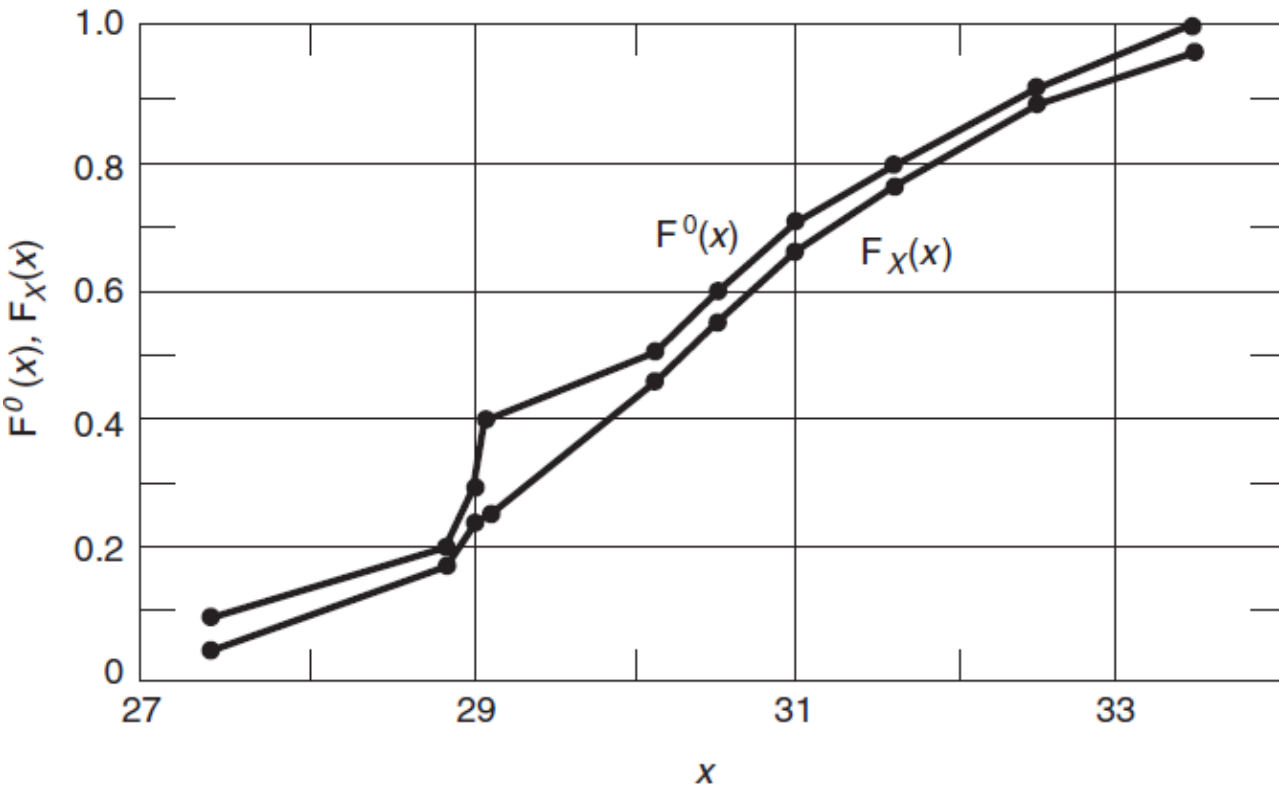
Valorile lui $F_X\left(x_{(i)}\right)$ pot fi găsite acum pe baza repartiţiei $N(30.3, 3.14)$ pentru X. De exemplu, cu ajutorul tabelului A.3 pentru variabila aleatoare normală standardizată U, avem

$$F_X(27.4) = F_U\left(\frac{27.4-30.3}{\sqrt{3.14}}\right) = F_U(-1.64) = 1 - F_U(1.64) = 1 - 0.9495 = 0.0505,$$

$$F_X(28.7) = F_U\left(\frac{28.7-30.3}{\sqrt{3.14}}\right) = F_U(-0.9) = 1 - F_U(0.9) = 1 - 0.8159 = 0.1841,$$

şi aşa mai departe.

Pentru a determina d_2 este constructiv să reprezentăm grafic $F^0(x)$ şi $F_X(x)$ ca funcţii de x, ca în figura următoare.



Se vede din figură că maximul diferenţelor dintre $F^0(x)$ şi $F_X(x)$ apare în $x = x_{(4)} = 29.1$. De aici,

$$d_2 = \left|F^0\left(x_{(4)}\right) - F_X\left(x_{(4)}\right)\right| = 0.4 - 0.2483 = 0.1517.$$

Cu $\alpha = 0.05$ şi $n = 10$, tabelul A.6 dă

$$c_{10,0.05} = 0.41.$$

Deoarece $d_2 < c_{10,0.05}$, acceptăm repartiţia normală $N(30.3, 3.14)$ la nivelul de semnificaţie de 5%.

Deoarece valorile parametrilor au fost de asemenea estimate din date, este mai adecvat să comparăm d_2 cu o valoare puţin mai mică decât 0.41. Datorită faptului că valoarea lui d_2 este mult sub 0.41, concluzia de mai sus este sigură.

14 Introducere în metoda MCMC

În multe aplicaţii ale modelării statistice, analistul de date vrea să folosească un model mai complex pentru un set de date, dar este forţat să recurgă la un model suprasimplificat pentru a folosi tehnicile disponibile. Metoda MCMC (prescurtare de la Markov chain Monte Carlo) este bazată pe simulare şi permite statisticianului sau inginerului să examineze datele folosind modele statistice realiste.

14.1 Inferenţă bayesiană

Incertitudinea despre valorile parametrilor necunoscuţi se poate reprezenta prin repartiţii de probabilitate şi se poate proceda ca şi cum parametrii ar fi cantităţi aleatoare. Dacă D reprezintă datele care sunt observate şi $\boldsymbol{\theta}$ parametrii modelului, atunci, pentru a face orice inferenţă, trebuie să ştim repartiţia de probabilitate comună $P(D, \boldsymbol{\theta})$ peste toate cantităţile aleatoare. Permitem lui $\boldsymbol{\theta}$ să fie multidimensional. Repartiţia comună poate fi scrisă

$$P(D, \boldsymbol{\theta}) = P(\boldsymbol{\theta}) P(D|\boldsymbol{\theta}),$$

unde $P(\boldsymbol{\theta})$ este numită *a priori* şi $P(D|\boldsymbol{\theta})$ este numită *verosimilitate*. O dată ce observăm datele D, putem utiliza teorema lui Bayes pentru a obţine *repartiţia a posteriori* după cum urmează

$$P(\boldsymbol{\theta}|D) = \frac{P(\boldsymbol{\theta}) P(D|\boldsymbol{\theta})}{\int P(\boldsymbol{\theta}) P(D|\boldsymbol{\theta}) \, d\boldsymbol{\theta}}. \tag{14.1}$$

Relaţia (14.1) dă repartiţia lui $\boldsymbol{\theta}$ condiţionată de datele observate D. Deoarece numitorul relaţiei (14.1) nu este o funcţie de $\boldsymbol{\theta}$ (pentru că integrăm în raport cu $\boldsymbol{\theta}$), putem scrie proporţionalitatea

$$P(\boldsymbol{\theta}|D) \propto P(\boldsymbol{\theta}) P(D|\boldsymbol{\theta}) = P(\boldsymbol{\theta}) L(\boldsymbol{\theta}; D).$$

Înţelegerea şi folosirea repartiţiei a posteriori este inima inferenţei bayesiene, unde este de interes a face deducţii folosind diferite trăsături caracteristice ale repartiţiei a posteriori (de exemplu momente, quantile, etc.). Aceste cantităţi pot fi scrise ca medii a posteriori ale funcţiilor de parametrii modelului după cum urmează

$$E(f(\boldsymbol{\theta})|D) = \frac{\int f(\boldsymbol{\theta}) P(\boldsymbol{\theta}) P(D|\boldsymbol{\theta}) \, d\boldsymbol{\theta}}{\int P(\boldsymbol{\theta}) P(D|\boldsymbol{\theta}) \, d\boldsymbol{\theta}}. \tag{14.2}$$

Numitorul din relaţiile (14.1) şi (14.2) este o constantă de proporţionalitate. Uneori, acesta poate fi foarte dificil, dacă nu imposibil, de obţinut. Aceasta este adevărat în special când problema este înalt dimensională, deoarece trebuie integrat în raport cu o mulţime de parametri. Integrarea analitică în aceste expresii a fost o sursă de dificultate în aplicaţiile inferenţei bayesiene, şi adesea au fost folosite modele mai simple pentru a face posibilă analiza. Integrarea Monte Carlo folosind MCMC este un răspuns la această problemă.

Schimbăm notaţia pentru a o face mai generală. Fie $\mathbf{X}$ un vector de d variabile aleatoare, cu repartiţia notată $\pi(\mathbf{x})$. Scopul este de a obţine media

$$E(f(\mathbf{X})) = \frac{\int f(\mathbf{x})\,\pi(\mathbf{x})\,d\mathbf{x}}{\int \pi(\mathbf{x})\,d\mathbf{x}}. \tag{14.3}$$

Cu metoda MCMC trebuie să ştim doar repartiţia lui $\mathbf{X}$ până la constanta de normalizare. Aceasta înseamnă că numitorul din relaţia (14.3) poate fi necunoscut. În cele ce urmează presupunem că $\mathbf{X}$ ia valori în $\mathbb{R}^d$. Metoda poate fi aplicată la variabile aleatoare discrete cu schimbări corespunzătoare.

14.2 Integrarea Monte Carlo

Cele mai multe metode din inferenţa statistică care folosesc simularea pot fi reduse la problema aflării integralelor. Aceasta este o parte fundamentală a metodologiei MCMC.

Integrarea Monte Carlo estimează integrala $E(f(\mathbf{X}))$ din relaţia (14.3) obţinând selecţia $\mathbf{X}_t, t = 1, ..., n$ din repartiţia $\pi(\mathbf{x})$ şi calculând

$$E(f(\mathbf{X})) \approx \frac{1}{n}\sum_{t=1}^{n} f(\mathbf{X}_t). \tag{14.4}$$

Când $\mathbf{X}_t$ sunt independente, aproximarea poate fi făcută oricât de bună crescând pe n. În metoda MCMC, selecţia nu este independentă în cele mai multe cazuri. Aceasta nu limitează utilizarea ei în aflarea integralelor folosind aproximari ca cea din relaţia (14.4). Totuşi, trebuie avută grijă din cauza dependenţei când se determină dispersia estimării din relaţia (14.4).

Exemplul 14.1. Pentru o repartiţie exponenţială cu $\lambda = 1$ aflăm $E\left(\sqrt{X}\right)$ folosind relaţia (14.4). Generăm 500 de variabile aleatoare din repartiţia dată, luăm radical din fiecare şi apoi aflăm media acestor valori. S-a obţinut o estimare de $0,889$. Valoarea obţinută folosind integrarea numerică este $0,886$, destul de aproape de cea obţinută prin metoda Monte Carlo.

$\mathbf{X}_t$ nu trebuie să fie independente atât timp cât sunt generate din *"întregul"* domeniu al lui $\pi(\mathbf{x})$ în proporţiile corecte. Aceast proces de generare poate fi făcut construind un lanţ Markov care are $\pi(\mathbf{x})$ ca repartiţie staţionară.

14.3 Lanţuri Markov

Un lanţ Markov este un şir de variabile aleatoare astfel încât următoarea valoare sau stare a şirului depinde doar de precedenta. Astfel, generăm un şir de variabile aleatoare $\mathbf{X}_0, \mathbf{X}_1, ...$ astfel încât următoarea stare $\mathbf{X}_{t+1}$ cu $t \geq 0$ este repartizată conform cu $P(\mathbf{X}_{t+1}|\mathbf{X}_t)$, care este numit *nucleul de tranziţie*. O realizare a acestui şir este numită de asemenea un lanţ Markov. Presupunem că nucleul de tranziţie nu depinde de t, făcând lanţul omogen în raport cu timpul.

O chestiune care trebuie abordată este cât de sensibil este lanţul faţă de starea sa de start $\mathbf{X}_0$. În anumite condiţii, lanţul va uita starea sa iniţială şi

va converge la o repartiţie staţionară, care este notată cu ψ. Când t creşte, $\mathbf{X}_t$ devin dependente de ψ.

Să presupunem că $\mathbf{X}_t, t = m+1, ..., n$, au repartiţia staţionară ψ. Putem îndepărta primele m iterate şi folosi pe cele $n-m$ rămase împreună cu relaţia (14.4) pentru a obţine o estimare a mediei după cum urmează

$$E\left(f\left(\mathbf{X}\right)\right) \approx \frac{1}{n-m} \sum_{t=m+1}^{n} f\left(\mathbf{X}_t\right). \tag{14.5}$$

Geyer a sugerat în 1992 că m poate fi între 1% şi 2% din n, unde n este suficient de mare pentru a obţine precizie adecvată în estimarea dată de relaţia (14.5).

Cât de mare ar trebui să fie n pentru a obţine precizia cerută în estimare? Calculul dispersiei estimării date de relaţia (14.5) este dificil, deoarece $\mathbf{X}_t$ nu sunt independente. Un mod de a determina n prin simulare este de a genera câteva lanţuri Markov în paralel, fiecare cu o valoare de start diferită. Estimările din relaţia (14.5) sunt comparate şi dacă diferenţa dintre ele este prea mare, atunci lungimea lanţurilor trebuie crescută.

14.4 Analiza rezultatelor

Un analist poate fi interesat în calculul mediilor, dispersiilor marginale, etc. pentru componentele lui $\mathbf{X}$. Dacă $X_{t,j}$ reprezintă componenta j a lui $\mathbf{X}_t$ la pasul t al lanţului, atunci, folosind relaţia (14.5), putem obţine mediile şi dispersiile marginale din

$$\overline{X}_{,j} = \frac{1}{n-m} \sum_{t=m+1}^{n} X_{t,j},$$

şi

$$S^2_{,j} = \frac{1}{n-m-1} \sum_{t=m+1}^{n} \left(X_{t,j} - \overline{X}_{,j}\right)^2.$$

Se pot construi lanţuri Markov cu o anumită repartiţie staţionară de care suntem interesaţi $\pi\left(\mathbf{x}\right)$, numită adesea *repartiţie ţintă*.

BIBLIOGRAFIE

1. Dumitrescu M., Bătătorescu A., Applied statistics using the R System, Editura Universitatii Bucuresti, 2006.

2. Martinez W.L., Martinez A.R., Computational Statistics Handbook with MATLAB, Chapman & Hall, 2002.

3. Saporta G., Probabilité, analyse des données et statistique, Edition Technip, Paris, 1990.

4. Soong, T.T., Fundamentals of probability and statistics for engineers, Wiley, 2004.

5. Tudor C., Teoria Probabilităţilor, Editura Universităţii Bucureşti, 2004.