Teoria Probabilităţilor

Teoria probabilităţilor este o disciplină a matematicii care se ocupă de studiul fenomenelor aleatoare.

- *aleator* = care depinde de o împrejurare viitoare şi nesigură; supus întâmplării
- provine din latină: *aleatorius*; *alea* (lat.) = zar; joc cu zaruri; joc de noroc; şansă; risc

↪ se măsoară *şansele pentru succes* sau *riscul pentru insucces* al unor evenimente

Fenomene şi procese aleatoare apar în:
→ jocuri de noroc, pariuri, loto (6 din 49)
→ previziuni meteo → previziuni economice / financiare (evaluarea stocurilor)
→ sondaje de opinie, asigurări (evaluarea riscurilor, pierderilor)

→ **în informatică**: criptografie, sisteme de comunicare, prelucrarea informaţiei, modelarea traficului în reţea, analiza probabilistică a unor algoritmi, fiabilitatea sistemelor, algoritmi de simulare, machine learning, data mining, recunoaşterea formelor / a vocii, generarea de numere aleatoare, algoritmi aleatori: de tip Monte-Carlo, de tip Las Vegas etc.
Exemple de întrebări:
→ Cum este concepută memoria cache pentru a maximiza viteza RAM a calculatoarelor?
→ Reţelele de calculatoare: Care este probabilitatea ca un pachet de date să fie recepţionat corect, atunci când canalul de transmisie este perturbat? În medie câte date sunt transmise corect?

Algoritmi aleatori

Def. 1. *Un algoritm pe cursul executării căruia se iau anumite decizii aleatoare este numit* ***algoritm aleator (randomizat)****.*

▷ durata de execuţie, spaţiul de stocare, rezultatul obţinut sunt variabile aleatoare (chiar dacă se folosesc aceleaşi valori input)
▷ la anumite tipuri de algoritmi corectitudinea e garantată doar cu o anumită probabilitate
▷ în mod paradoxal, incertitudinea ne poate oferi mai multă eficienţă
Exemplu: Random QuickSort, în care elementul pivot este selectat aleator
• Algoritm de tip **Las Vegas** este un algoritm aleator, care returnează la fiecare execuţie rezultatul corect (independent de alegerile aleatoare făcute); durata de execuţie este o variabilă aleatoare.
Exemplu: Random QuickSort
• Un algoritm aleatoriu pentru care rezultatele obţinute sunt corecte doar cu o anumită probabilitate se numeşte algoritm **Monte Carlo**.
↪ se examinează probabilitatea cu care rezultatul este corect; probabilitatea de eroare poate fi scăzută semnificativ prin execuţii repetate, independente;
Exemplu:
1) testul Miller-Rabin, care verifică dacă un număr natural este prim sau este număr compus; rezultatul testului returnează fie răspunsul: "numărul este sigur un număr compus" sau răspunsul: "numărul este probabil un număr prim"; testul nu returnează valorile divizorilor numărului compus;
2) problema tăieturii minime într-un graf (algoritmul lui D. Karger: random min-cut)
⟶ De care tip este următorul algoritm (scris în pseudocod)?
Input: Fie A(1),...,A(200) un vector cu 200 de elemente, din care 100 sunt egale cu 0 şi restul egale cu 1 (ordinea lor este necunoscută).
Output: Să se găsească un 0 în vector.

```
algorithm(array A, size n)
begin
repeat
randomly select one element from A
until 0 is found
end
```

Răspuns: Algoritm de tip Las Vegas.
Versiunea Monte Carlo a problemei formulate anterior: se dă k numărul maxim de iteraţii

```
find_MC(array A, size n, k)
begin
    i=0
    repeat
        randomly select one element from A
        i = i + 1
    until i=k or 0 is found
end
```

▷ dacă 0 este găsit, atunci algoritmul se încheie cu rezultatul corect, altfel algoritmul nu găseşte niciun 0; probabilitatea de a găsi pe 0 după k iteraţii este

$$P(\text{“0 este găsit după } k \text{ iteraţii”}) = 1 - (1/2)^k.$$

Noţiuni introductive:

• Experienţa aleatoare este acea experienţă al cărei rezultat nu poate fi cunoscut decât după încheierea ei.
• Evenimentul este rezultatul unui experiment.

Exemple:

▷ Experiment: aruncarea a două zaruri, eveniment: ambele zaruri indică 1
▷ experiment: aruncarea unei monede, eveniment: moneda indică pajură
▷ experiment: extragerea unei cărţi de joc, eveniment: s-a extras as
▷ experiment: extragerea unui număr la loto, eveniment: s-a extras numărul 27

• evenimentul imposibil, notat cu $\emptyset$, este evenimentul care nu se realizează niciodată la efectuarea experienţei aleatoare
• evenimentul sigur este un eveniment care se realizează cu certitudine la fiecare efectuare a experienţei aleatoare
• spaţiul de selecţie, notat cu Ω, este mulţimea tuturor rezultatelor posibile ale experimentului considerat
  ◇ spaţiul de selecţie poate fi finit sau infinit
• dacă A este o submulţime a lui Ω atunci A se numeşte eveniment aleator, iar dacă A are un singur element atunci A este un eveniment elementar.

▷ O *analogie între evenimente şi mulţimi* permite o scriere şi în general o exprimare mai comode ale unor idei şi rezultate legate de conceptul de eveniment aleator.

Exemplu: Experimentul: aruncarea unui zar, spaţiul de selecţie: $\Omega = \{e_1, e_2, e_3, e_4, e_5, e_6\}$,
e_i: s-a obţinut numărul i $(i = 1, \ldots, 6)$
$e_1, e_2, e_3, e_4, e_5, e_6$ sunt evenimente elementare
A: s-a obţinut un număr par $\Rightarrow A = \{e_2, e_4, e_6\}$
$\bar{A}$: s-a obţinut un număr impar $\Rightarrow \bar{A} = \{e_1, e_3, e_5\}$

Operaţii cu evenimente

• dacă $A, B \subseteq \Omega$, atunci evenimentul reuniune $A \cup B$ este un eveniment care se produce dacă cel puţin unul din evenimentele A sau B se produce

• dacă $A, B \subseteq \Omega$, atunci evenimentul intersecţie $A \cap B$ este un eveniment care se produce dacă cele două evenimente A şi B se produc în acelaşi timp

• dacă $A \subseteq \Omega$ atunci evenimentul contrar sau complemetar $\bar{A}$ este un eveniment care se realizează atunci când evenimentul A nu se realizează

• dacă $A, B \subseteq \Omega$ ele sunt evenimente incompatibile dacă sunt disjuncte: $A \cap B = \emptyset$

• dacă $A, B \subseteq \Omega$, atunci evenimentul diferenţă $A \setminus B$ este un eveniment care se produce dacă A are loc şi B nu are loc, adică

$$A \setminus B = A \cap \overline{B}$$

Relaţii între evenimente

• dacă $A, B \subseteq \Omega$, atunci A implică B, dacă producerea evenimentului A conduce la producerea evenimentului B: $A \subseteq B$

• dacă A implică B şi B implică A, atunci evenimentele A şi B sunt egale: $A = B$

Proprietăţi ale operaţiilor între evenimente $A, B, C \subseteq \Omega$

Operaţiile de reuniune şi intersecţie sunt operaţii **comutative**:

$$A \cup B = B \cup A, \quad A \cap B = B \cap A$$

asociative

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

şi **distributive**

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

satisfac legile lui De Morgan

$$\overline{A \cup B} = \bar{A} \cap \bar{B}, \quad \overline{A \cap B} = \bar{A} \cup \bar{B}.$$

Are loc $\bar{\bar{A}} = A$.

Frecvenţa relativă şi frecvenţa absolută

fie A un eveniment asociat unei experienţe, repetăm experienţa de n ori (în aceleaşi condiţii date) şi notăm cu $k_n(A)$ numărul de realizări ale evenimentului A; frecvenţa relativă a evenimentului A este numărul

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

$k_n(A)$ este frecvenţa absolută a evenimentului A.

Definiţia clasică a probabilităţii

într-un experiment în care cazurile posibile sunt finite la număr şi au aceleaşi şanse de a se realiza, **probabilitatea** unui eveniment A este numărul

$$P(A) = \frac{\text{numărul de cazuri lui } A}{\text{numărul total de cazuri posibile}}.$$

$\triangleright$ Prin repetarea de multe ori a unui experiment, în condiţii practic identice, frecvenţa relativă $f_n(A)$ de apariţie a evenimentului A este aproximativ egală cu $P(A)$

$$f_n(A) \approx P(A), \text{ dacă } n \to \infty.$$

Exemplu: Experiment: Se aruncă 4 monede. Evenimentul A: *cele 4 monede indică pajură exact de 3 ori* ; experimentul s-a repetat de $n = 100$ de ori şi evenimentul A a apărut de 22 de ori.

$$f_n(A) =?, \qquad P(A) =?$$

Răspuns: $f_n(A) = \frac{22}{100} = 0.22$

$\Omega = \{(c,c,c,c), (c,p,p,p), \ldots, (p,p,p,c), (p,p,p,p)\}$

$A = \{(c,p,p,p), (p,c,p,p), (p,p,c,p), (p,p,p,c)\}$

$\Rightarrow P(A) = \frac{4}{2^4} = 0.25$

Definiţia axiomatică a probabilităţii

Definiţia clasică a probabilităţii poate fi utilizată numai în cazul în care numărul cazurilor posibile este finit. Dacă numărul evenimentelor elementare este infinit, atunci există evenimente pentru care probabilitatea în sensul clasic nu are nici un înţeles.

O teorie formală a probabilităţii a fost creată în anii '30 ai secolului XX de către matematiciuanul rus **Andrei Nikolaevici Kolmogorov**, care, în anul **1933**, a dezvoltat teoria axiomatică a probabilităţii în lucrarea sa *Conceptele de bază ale Calculului Probabilităţii*.

$\hookrightarrow$ unui eveniment aleator $A \subseteq \Omega$ i se asociază un număr $P(A)$, probabilitatea de apariţie a evenimentului A

$\hookrightarrow P : \mathcal{K} \to \mathbb{R}$ este o funcţie astfel încât: orice eveniment aleator $A \in \mathcal{K}$

$\mathcal{K}$ are structura unei σ-algebre (vezi Def. 2)

P satisface anumite axiome (vezi Def. 3)

Def. 2. *O familie $\mathcal{K}$ de evenimente din spaţiul de selecţie Ω se numeşte **σ-algebră** dacă sunt satisfăcute condiţiile:*

(i) $\mathcal{K} \neq \emptyset$;

(ii) dacă $A \in \mathcal{K}$, *atunci* $\overline{A} \in \mathcal{K}$;

(iii) dacă $A_n \in \mathcal{K}$, $n \in \mathbb{N}$, *atunci* $\bigcup_{n=1}^{\infty} A_n \in \mathcal{K}$.

*Perechea $(\Omega, \mathcal{K})$ se numeşte **spaţiu măsurabil**.*

Exemplu: 1) Dacă $A \subset \Omega$ atunci $\mathcal{K} = \{\emptyset, A, \bar{A}, \Omega\}$ este o σ-algebră.
2) $\mathcal{P}(\Omega)$:= mulţimea tuturor submulţimilor ale lui Ω este o σ-algebră.

P. 1. *Proprietăţi ale unei σ-algebre: Dacă $\mathcal{K}$ este o σ-algebră în Ω, atunci au loc proprietăţile:*

(1) $\emptyset, \Omega \in \mathcal{K}$;

(2) $A, B \in \mathcal{K} \Longrightarrow A \cap B,\ A \setminus B \in \mathcal{K}$;

(3) $A_n \in \mathcal{K}$, $n \in \mathbb{N} \Longrightarrow \bigcap_{n=1}^{\infty} A_n \in \mathcal{K}$.

Def. 3. *Fie $\mathcal{K}$ o σ-algebră în Ω. O funcţie $P : \mathcal{K} \to \mathbb{R}$ se numeşte* ***probabilitate*** *dacă satisface axiomele:*

(i) $P(\Omega) = 1$

(ii) $P(A) \geq 0$ *pentru orice* $A \in \mathcal{K}$;

(iii) pentru orice şir $(A_n)_{n\in\mathbb{N}}$ *de evenimente două câte două disjuncte (adică* $A_i \cap A_j = \emptyset$ *pentru orice* $i \neq j$*) din* $\mathcal{K}$ *are loc*

$$P\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} P(A_n)$$

Tripletul $(\Omega, \mathcal{K}, P)$ *format din spaţiul măsurabil* $(\Omega, \mathcal{K})$ *şi probabilitatea* $P : \mathcal{K} \to \mathbb{R}$ *se numeşte* ***spaţiu de probabilitate***.

P. 2. *Fie* $(\Omega, \mathcal{K}, P)$ *un spaţiu de probabilitate. Au loc proprietăţile:*

(1) $P(\overline{A}) = 1 - P(A)$ *şi* $0 \leq P(A) \leq 1$

(2) $P(\emptyset) = 0$

(3) $P(A \setminus B) = P(A) - P(A \cap B)$

(4) $A \subseteq B \Longrightarrow P(A) \leq P(B)$*, adică* P *este monotonă.*

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

Exerciţiu: Să se arate că pentru $\forall\, A, B, C \in \mathcal{K}$ are loc:

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

Şiruri de evenimente aleatoare

P. 3. *Fie* $(\Omega, \mathcal{K}, P)$ *un spaţiu de probabilitate. Au loc următoarele proprietăţi:*
(1) dacă $(A_n)_n$ *este un şir crescător de evenimente din* $\mathcal{K}$*, adică* $A_n \in \mathcal{K}$ *şi* $A_n \subseteq A_{n+1}$ $\forall n \in \mathbb{N}$*, atunci*

$$\lim_{n\to\infty} P(A_n) = P\Big(\bigcup_{n=1}^{\infty} A_n\Big).$$

(2) dacă $(B_n)_n$ *este un şir descrescător de evenimente din* $\mathcal{K}$*, adică* $B_n \in \mathcal{K}$ *şi* $B_{n+1} \subseteq B_n$ $\forall n \in \mathbb{N}$*, atunci*

$$\lim_{n\to\infty} P(B_n) = P\Big(\bigcap_{n=1}^{\infty} B_n\Big).$$

Exemplu: $\Omega := [0,1]$, $\mathcal{K} := \mathcal{B}([0,1])$ (σ-algebra mulţimilor boreliene din $[0,1]$), P:= măsura Lebesgue (de exemplu: $P([a,b]) = P((a,b)) = P((a,b]) = P([a,b)) = b - a$, unde $0 \leq a < b \leq 1$)

$$A_n = \left[\frac{1}{n}, \frac{1}{2} - \frac{1}{n}\right], n \geq 4 \qquad \Rightarrow \bigcup_{n=4}^{\infty} A_n = ?$$

Spre ce valoare converge şirul $\Big(P(A_n)\Big)_n$?

2. $\Omega = [0,1], \mathcal{K} := \mathcal{B}([0,1])$

$$B_n = \left[\frac{1}{4} - \frac{1}{n}, \frac{3}{4} + \frac{1}{n}\right], n \geq 4 \qquad \Rightarrow \bigcap_{n=4}^{\infty} B_n = ?$$

Spre ce valoare converge şirul $\Big(P(B_n)\Big)_n$?

Probabilitate condiţionată

Def. 4. *Fie $(\Omega, \mathcal{K}, P)$ un spaţiu de probabilitate şi fie $A, B \in \mathcal{K}$.* ***Probabilitatea condiţionată a evenimentului A de evenimentul B*** *este $P(\cdot|B) : \mathcal{K} \to [0,1]$ definit prin*

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

dacă $P(B) > 0$. $P(A|B)$ este probabilitatea apariţiei evenimentului A, ştiind că evenimentul B s-a produs.

P. 4. *Pentru $A, B \in \mathcal{K}, P(A) > 0, P(B) > 0$ au loc:*

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

$$P(A|B) = 1 - P(\bar{A}|B).$$

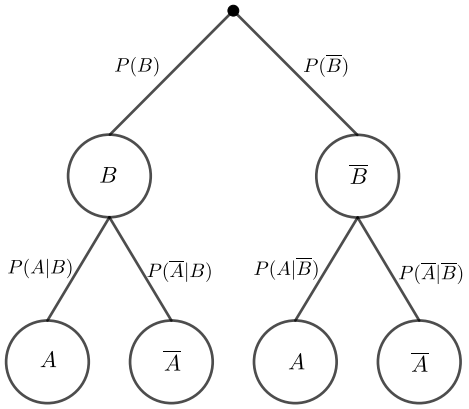


Fig.1. Probabilităţi condiţionate

Def. 5. *O familie $A_1, \ldots A_n$ de evenimente din Ω se numeşte* ***partiţie*** *sau* ***sistem complet de evenimente*** *a lui Ω, dacă $\bigcup_{i=1}^{n} A_i = \Omega$ şi pentru fiecare $i, j \in \{1, \ldots, n\}$, $i \neq j$, evenimentele A_i şi A_j sunt disjuncte, adică $A_i \cap A_j = \emptyset$.*

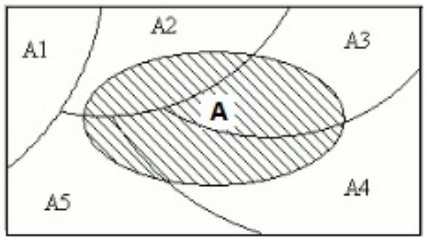


Fig.2. Partiţie $A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5 = \Omega$

Exemplu: Dacă $B \subset \Omega$ atunci $\{B, \bar{B}\}$ formează o partiţie a lui Ω.

P. 5. *(Formula probabilităţii totale) Într-un spaţiu de probabilitate $(\Omega, \mathcal{K}, P)$ considerăm partiţia $A_1, ..., A_n$ a lui Ω cu $P(A_i) > 0$ şi $A_i \in \mathcal{K} \; \forall \; i \in \{1, \ldots, n\}$, şi fie $A \in \mathcal{K}$. Atunci are loc*

$$P(A) = \sum_{i=1}^{n} P(A|A_i)P(A_i).$$

Evenimente independente

Def. 6. *Fie $(\Omega, \mathcal{K}, P)$ un spaţiu de probabilitate. Evenimentele $A, B \in \mathcal{K}$ sunt **evenimente independente** dacă*

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

Observaţie: Evenimentele $A, B \in \mathcal{K}$ sunt **independente**, dacă apariţia evenimentului A, nu influenţează apariţia evenimentului B şi invers, adică

$$P(A|B) = P(A) \text{ şi } P(B|A) = P(B),$$

dacă $P(A) > 0$ şi $P(B) > 0$.

P. 6. *(Formula lui Bayes) Într-un spaţiu de probabilitate $(\Omega, \mathcal{K}, P)$ considerăm partiţia $A_1, \ldots, A_n$ a lui Ω cu $P(A_i) > 0$ şi $A_i \in \mathcal{K}\ \forall\ i \in \{1, ...n\}$, şi fie $A \in \mathcal{K}$ astfel încât $P(A) > 0$. Atunci,*

$$P(A_j|A) = \frac{P(A_j)P(A|A_j)}{\sum_{i=1}^{n} P(A_i)P(A|A_i)} \quad \forall\ j \in \{1, 2, ..., n\}.$$

$\triangleright$ $P(A_i)$ sunt **probabilităţi apriori** pentru A_i, numite şi ipoteze (aserţiuni), A se numeşte **evidenţă** (dovadă, premisă). Cu formula lui Bayes se calculează probabilităţile pentru ipoteze, cunoscând evidenţa: $P(A_i|A)$, acestea se numesc **probabilităţi posterioare**; $P(A|A_i)$ reprezintă verosimilitatea datelor observate.
$\triangleright$ Se pot calcula probabilităţile *cauzelor*, date fiind *efectele*; formula lui Bayes ne ajută să diagnosticăm o anumită situaţie sau să testăm o ipoteză.

P. 7. *(Regula de înmulţire) Fie $(\Omega, \mathcal{K}, P)$ un spaţiu de probabilitate şi fie $A_1, \ldots, A_n \in \mathcal{K}$ astfel încât*

$$P(A_1 \cap \cdots \cap A_{n-1}) > 0.$$

Atunci,

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

Exemplu: Într-o urnă sunt 2 bile verzi şi 3 bile albastre. Se extrag 2 bile succesiv, fără returnare. Care este probabilitatea ca
a) prima bilă să fie verde, iar cea de-a doua albastră?
b) cele 2 bile să aibă aceeaşi culoare?
c) a doua bilă să fie albastră?
d) prima bilă să fie verde, *ştiind* că a doua este albastră?
e) se mai extrage o a treia bilă; se cere probabilitatea ca prima bilă să fie verde, cea de-a doua albastră şi a treia tot albastră.

Răspuns: Notăm pentru $i \in \{1, 2, 3\}$:
A_i: la a i-a extragere s-a obţinut bilă albastră;
V_i: la a i-a extragere s-a obţinut bilă verde;
a) folosim P4: $P(V_1 \cap A_2) = P(A_2|V_1)P(V_1) = \frac{2}{5} \cdot \frac{3}{4}$
b) $P((V_1 \cap V_2) \cup (A_1 \cap A_2)) = P(V_1 \cap V_2) + P(A_1 \cap A_2) = P(V_2|V_1)P(V_1) + P(A_2|A_1)P(A_1) = \frac{1}{4} \cdot \frac{2}{5} + \frac{2}{4} \cdot \frac{3}{5}$
c) folosim formula probabilităţii totale P5:

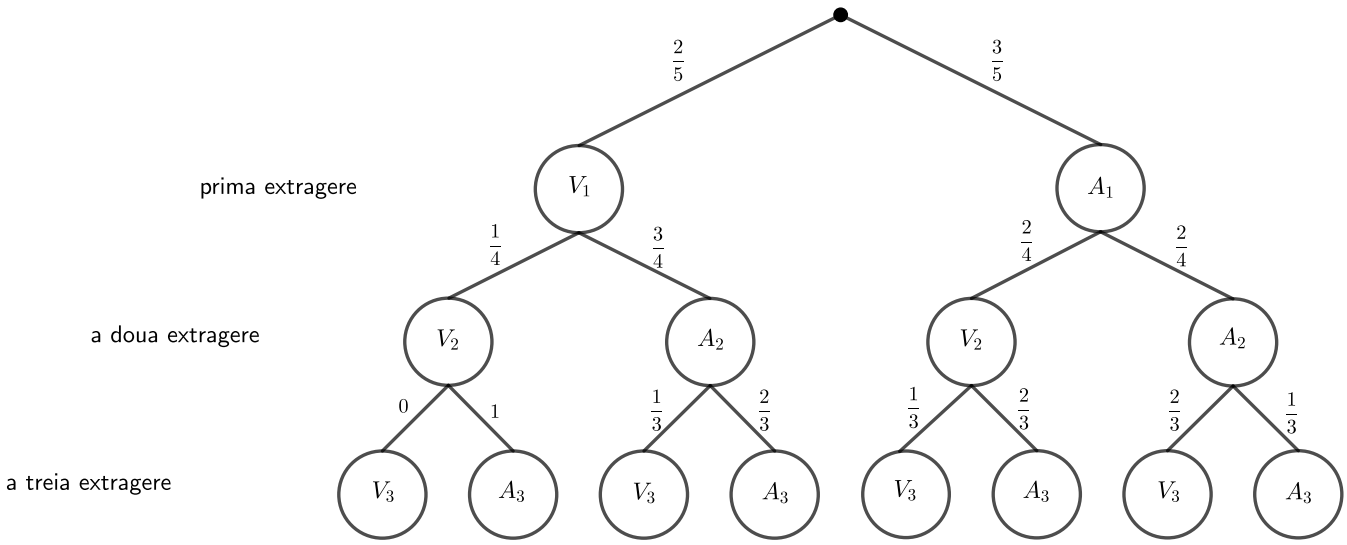


Fig. 3. Extragere fără returnare

$P(A_2) = P(A_2|V_1)P(V_1) + P(A_2|A_1)P(A_1) = \frac{3}{4} \cdot \frac{2}{5} + \frac{2}{4} \cdot \frac{3}{5}$

d) folosim P4: $P(V_1|A_2) = \frac{P(V_1 \cap A_2)}{P(A_2)} = \frac{P(A_2|V_1)P(V_1)}{P(A_2)} = \frac{\frac{3}{4} \cdot \frac{2}{5}}{\frac{3}{4} \cdot \frac{2}{5} + \frac{2}{4} \cdot \frac{3}{5}}$

e) formula de înmulţire a probabilităţilor P7:
$P(V_1 \cap A_2 \cap A_3) = P(V_1) \cdot P(A_2|V_1) \cdot P(A_3|V_1 \cap A_2) = \frac{2}{5} \cdot \frac{3}{4} \cdot \frac{2}{3}$.

Def. 7. *Fie $(\Omega, \mathcal{K}, P)$ un spaţiu de probabilitate. $B_1, \ldots, B_n$ sunt n* ***evenimente independente*** *din $\mathcal{K}$ dacă*

$$P(B_{i_1} \cap \cdots \cap B_{i_m}) = P(B_{i_1}) \cdot \ldots \cdot P(B_{i_m})$$

pentru orice submulţime finită $\{i_1, \ldots, i_m\} \subseteq \{1, 2, ..., n\}$.

P. 8. *Fie $(\Omega, \mathcal{K}, P)$ un spaţiu de probabilitate şi fie $A, B \in \mathcal{K}$. Sunt echivalente afirmaţiile:*

(1) *A şi B sunt independente.*

(2) *$\overline{A}$ şi B sunt independente.*

(3) *A şi $\overline{B}$ sunt independente.*

(4) *$\overline{A}$ şi $\overline{B}$ sunt independente.*

Variable aleatoare

Exemplu: Un jucător aruncă două monede $\Rightarrow \Omega = \{(c,p), (c,c), (p,c), (p,p)\}$ (c=cap; p=pajură)

X indică de câte ori a apărut pajură: $\Rightarrow X : \Omega \to \{0, 1, 2\}$

$\Rightarrow P(X = 0) = P(X = 2) = \frac{1}{4}, P(X = 1) = \frac{1}{2}$,

$\triangleright$ Prescurtare: variabilă aleatoare $\to$ v.a.

O variabilă aleatoare este
discretă, dacă ia un număr finit de valori $(x_1, \ldots, x_n)$ sau un număr infinit numărabil de valori $(x_1, \ldots, x_n, \ldots)$

continuă, dacă valorile sale posibile sunt nenumărabile şi sunt într-un interval (sau reunine de intervale) sau în $\mathbb{R}$

V.a. discrete: exemple de v.a. numerice discrete: suma numerelor obţinute la aruncarea a 4 zaruri, numărul produselor defecte produse de o anumită firmă într-o săptămână; numărul apelurilor telefonice într-un call center în decursul unei ore; numărul de accesări ale unei anumite pagini web în decursul unei anumite zile (de ex. duminica); numărul de caractere transmise eronat într-un mesaj de o anumită lungime; exemple de v.a. categoriale ($\rightarrow$ se clasifică în categorii): prognoza meteo: ploios, senin, înnorat, ceţos; calitatea unor servicii: nesatisfăcătoare, satisfăcătoare, bune, foarte bune, excepţionale ...

v.a. continue sunt v.a. numerice: timpul de funcţionare până la defectare a unei piese electronice, temperatura într-un oraş, viteza înregistrată de radar pentru maşini care parcurg o anumită zonă ...

Variabile aleatoare numerice - definiţie formală

Def. 8. *Fie* $(\Omega, \mathcal{K}, P)$ *spaţiu de probabilitate* $X : \Omega \to \mathbb{R}$ *este o variabilă aleatoare, dacă*

$$\{\omega \in \Omega \,:\, X(\omega) \leq x\} \in \mathcal{K} \;\; \textit{pentru fiecare } x \in \mathbb{R}.$$

Variabile aleatoare discrete $X : \Omega \to \{x_1, x_2, \ldots, x_i, \ldots\}$

Distribuţia de probabilitate a v.a. discrete X

$$X \sim \begin{pmatrix} x_1 & x_2 & \ldots & x_i & \ldots \\ p_1 & p_2 & \ldots & p_i & \ldots \end{pmatrix} = \begin{pmatrix} x_i \\ p_i \end{pmatrix}_{i \in I}$$

$I \subseteq \mathbb{N}$ (mulţime de indici) probabilităţile: $p_i = P(X = x_i) > 0$, $i \in I$, cu $\sum_{i \in I} p_i = 1$.

$\triangleright$ Variabilele aleatoare discrete sunt caracterizate de distribuţiile lor.
$\triangleright$ $\{X = x_i\} = \{\omega \in \Omega : X(\omega) = x_i\}$ este un eveniment din $\mathcal{K}$ pentru fiecare $i \in I$.

Distribuţii discrete clasice

Distribuţia discretă uniformă: $X \sim Unif(n)$

$$X \sim \begin{pmatrix} 1 & 2 & \ldots & n \\ \frac{1}{n} & \frac{1}{n} & \ldots & \frac{1}{n} \end{pmatrix}$$

Exemplu: Se aruncă un zar, fie X v.a. care indică numărul apărut

$$\Rightarrow X \sim \begin{pmatrix} 1 & 2 & \ldots & 6 \\ \frac{1}{6} & \frac{1}{6} & \ldots & \frac{1}{6} \end{pmatrix}$$

Matlab/Octave: *unidrnd*

Distribuţia Bernoulli: $X \sim Bernoulli(p)$, $p \in (0, 1)$

$$X \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$$

Exemplu: în cadrul unui experiment poate să apară evenimentul A (*succes*) sau $\bar{A}$ (*insucces*)

$X = 0 \Leftrightarrow$ dacă $\bar{A}$ apare

$X = 1 \Leftrightarrow$ dacă A apare
$\Rightarrow X \sim Bernoulli(p)$ cu $p := P(A)$

$$X \sim \begin{pmatrix} 0 & 1 \\ 1 - P(A) & P(A) \end{pmatrix}$$

generare în Matlab/Octave:

```
function X=bernoulli(p,n)
nr=rand(1,n);
X=(nr<=p); % vector de date avand distributia Bernoulli(p)
end
%%%%%%%%
function nr=bernoulli(p,n)
nr=floor(rand(1,n)+p);
end
%%%%%%%%
binornd(1,p,1,n)
```

Distribuţia binomială: $X \sim Bino(n, p), n \in \mathbb{N}^*, p \in (0, 1)$
în cadrul unui experiment poate să apară evenimentul A (*succes*) sau $\bar{A}$ (*insucces*)
• $A =$ succes cu $P(A) = p$, $\bar{A} =$ insucces $P(\overline{A}) = 1 - p$
• se repetă experimentul de n ori
• v.a. X= numărul de succese în n repetări independente ale experimentului $\Rightarrow$ valori posibile: $X \in \{0, 1, \ldots, n\}$

$$P(X = k) = C_n^k p^k (1 - p)^{n-k}, \quad k = 0, \ldots, n.$$

Exemplu: Un zar se aruncă de 10 ori, fie X v.a. care indică de câte ori a apărut numărul 6 $\Rightarrow Bino(10, \frac{1}{6})$.
$\rightarrow$ de reamintit: formula binomială

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

pentru $a = p$ şi $b = 1 - p$ se obţine

$$1 = \sum_{k=0}^{n} C_n^p p^k (1 - p)^{n-k}$$

Matlab/Octave: *binornd*

▷ Distribuţia binomială corespunde modelului cu extragerea bilelor dintr-o urnă cu returnarea bilelor după fiecare extragere.
Exemplu: Într-o urnă sunt n_1 bile albe şi n_2 bile negre. Se extrag cu returnare n bile, din care k sunt albe şi $n - k$ negre; fie v.a. X = numărul de bile albe extrase $\Rightarrow X \sim Bino(n, p)$ cu $p = \frac{n_1}{n_1+n_2}$.
Exerciţiu: O reţea de laborator este compusă din 15 calculatoare. Reţeaua a fost atacată de un virus nou, care atacă un calculator cu o probabilitatea 0.4, independent de alte calculatoare. Care este probabilitatea ca virusul a atacat a) cel mult 10 computere; b) cel puţin 10 calculatoare; c) exact 10 calculatoare.

Distribuţia hipergeometrică: $X \sim Hyge(n, n_1, n_2)$
Într-o urnă sunt n_1 bile albe şi n_2 bile negre. Se extrag **fără returnare** n bile.
Fie v.a. X = numărul de bile albe extrase $\Rightarrow$ valori posibile pentru X sunt $\{0, 1, \ldots, n^*\}$ cu

$$n^* = \min(n_1, n) = \begin{cases} n_1 & \text{dacă } n_1 < n \text{ (mai puţine bile albe decât numărul de extrageri)} \\ n & \text{dacă } n_1 \geq n \text{ (mai multe bile albe decât numărul de extrageri)} \end{cases}$$

Fie $n_1, n_2, n \in \mathbb{N}$ cu $n \leq n_2$ şi notăm $n^* = \min(n_1, n)$.

$$\Rightarrow \; P(X = k) = \frac{C_{n_1}^k C_{n_2}^{n-k}}{C_{n_1+n_2}^n}, \quad k = 0, \ldots, n^*.$$

Matlab/Octave: *hygernd*

Exemplu: Loto 6 din 49 $\rightarrow$ Care este probabilitatea de a nimeri exact 4 numere câştigătoare?
Răspuns: Între cele 49 de bile exact $n_1 = 6$ sunt câştigătoare ("bilele albe") şi $n_2 = 43$ necâştigătoare ("bilele negre"). Care este probabilitatea ca din $n = 6$ extrageri fără returnare, exact $k = 4$ numere să fie câştigătoare?

$$\Rightarrow \; P(X = 4) = \frac{C_6^4 C_{43}^2}{C_{49}^6}$$

Distribuţia geometrică $X \sim Geo(p)$
În cadrul unui experiment poate să apară evenimentul A (*succes*) sau $\bar{A}$ (*insucces*)
• $A =$ succes cu $P(A) = p$, $\quad \bar{A} =$ insucces $P(\overline{A}) = 1 - p$
• se repetă (independent) experimentul până apare prima dată A ("succes")
• v.a. X arată de câte ori apare $\bar{A}$ (numărul de "insuccese") *până* la apariţia primului A ("succes") $\Rightarrow$ valori posibile: $X \in \{0, 1, \ldots\}$

$$P(X = k) = p(1 - p)^k \quad \text{pentru } k \in \{0, 1, 2, \ldots\}.$$

Matlab/Octave: *geornd*

Exemplu: X v.a. ce indică numărul de retransmisii printr-un canal cu zgomot (canal cu perturbări) până la prima recepţionare corectă a mesajului; X are distribuţie geometrică.

Variabile aleatoare independente

Def. 9. *Variabilele aleatoare discrete X (care ia valorile $\{x_i, i \in I\}$) şi Y (care ia valorile $\{y_j, j \in J\}$) sunt* ***independente****, dacă şi numai dacă*

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j) \quad \forall\, i \in I, j \in J.$$

Notaţie: $P(X = x_i, Y = y_j) = P(\{X = x_i\} \cap \{Y = y_j\})$.
Observaţie: X şi Y sunt **variabilele aleatoare independente** $\Leftrightarrow$

$$P(X \leq x, Y \leq y) = P(X \leq x) \cdot P(Y \leq y) \quad \forall\, x, y \in \mathbb{R}.$$

Def. 10. $(X_1, \ldots, X_n)$ *este un* ***vector aleator discret*** *dacă fiecare componentă a sa este o variabiă aleatoare discretă.*

$\triangleright$ Vectorii aleatori sunt caracterizaţi de distribuţiile lor! De exemplu, un vector aleator cu 2 componente:

$$(X, Y) \sim \begin{pmatrix} (x_i, y_j) \\ p_{ij} \end{pmatrix}_{(i,j) \in I \times J}$$

unde $I, J \subseteq \mathbb{N}$ sunt mulţimi de indici, $p_{ij} = P((X, Y) = (x_i, y_j)) = P(\{X = x_i\} \cap \{Y = y_j\})$, $p_{ij} > 0 \; \forall$ $i \in I, j \in J$, iar $\sum_{(i,j) \in I \times J} p_{ij} = 1$.

Uneori distribuţia vectorului (X, Y) se dă sub formă tabelară:

$X \backslash Y$	...	y_j	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	...	p_{ij}	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Observaţie: Dacă X şi Y sunt v.a. independente, atunci

$$p_{ij} = P(\{X = x_i\} \cap \{Y = y_j\}) = P(X = x_i)P(Y = y_j) \quad \forall\, i \in I, j \in J. \tag{1}$$

▷ Dacă X şi Y sunt v.a. independente, şi se ştiu distribuţiile lor, atunci distribuţia vectorului aleator (X, Y) se determină pe baza formulei (1).
▷ Dacă se cunoaşte distribuţia vectorului aleator (X, Y) distribuţiile lui X şi Y se determină astfel:

$$P(X = x_i) = \sum_{j \in J} p_{ij} \quad \forall i \in I$$

$$P(Y = y_j) = \sum_{i \in I} p_{ij} \quad \forall j \in J.$$

Operaţii cu variabile aleatoare (numerice)

• Cunoscând distribuţia vectorului (X, Y) cum se determină distribuţia pentru $X + Y$, $X \cdot Y$, $X^2 - 1$, $2Y$?

Exemplu: Fie vectorul aleator discret (X_1, X_2) cu distribuţia dată de următorul tabel:

$X_1 \backslash X_2$	0	1	2
1	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{2}{16}$
2	$\frac{1}{16}$	$\frac{5}{16}$	$\frac{5}{16}$

Determinaţi:
a) distribuţiile variabilelor aleatoare X_1 şi X_2;
b) distribuţiile variabilelor aleatoare $X_1 + X_2$ şi $X_1 \cdot X_2$, $X_1^2 - 1$;
c) dacă variabilele aleatoare X_1 şi X_2 sunt independente sau dependente.
a) $X_1 \sim \begin{pmatrix} 1 & 2 \\ \frac{5}{16} & \frac{11}{16} \end{pmatrix}$ şi $X_2 \sim \begin{pmatrix} 0 & 1 & 2 \\ \frac{3}{16} & \frac{6}{16} & \frac{7}{16} \end{pmatrix}$.
b) $X_1 + X_2 \sim \begin{pmatrix} 1 & 2 & 3 & 4 \\ \frac{2}{16} & \frac{2}{16} & \frac{7}{16} & \frac{5}{16} \end{pmatrix}$ şi $X_1 \cdot X_2 \sim \begin{pmatrix} 0 & 1 & 2 & 4 \\ \frac{3}{16} & \frac{1}{16} & \frac{7}{16} & \frac{5}{16} \end{pmatrix}$, $X_1^2 - 1 \sim \begin{pmatrix} 0 & 3 \\ \frac{5}{16} & \frac{11}{16} \end{pmatrix}$
c) X_1 şi X_2 nu sunt independente, pentru că $\frac{2}{16} = P(X_1 = 1, X_2 = 0) \neq P(X_1 = 1)P(X_2 = 0) = \frac{5}{16} \cdot \frac{3}{16}$.

• Cunoscând distribuţiile variabilelor aleatoare independente (discrete) X şi Y, cum se determină distribuţia pentru $X + Y$, $X \cdot Y$?

Exerciţiu: Fie X,Y v.a. independente, având distribuţiile

$$X \sim \begin{pmatrix} 0 & 1 \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}, \quad Y \sim \begin{pmatrix} -1 & 0 & 1 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

a) Care sunt distribuţiile v.a. $2X + 1$, Y^2, dar distribuţia vectorului aleator (X, Y)?
b) Care sunt distribuţiile v.a. $X + Y$, $X \cdot Y$, $\max(X, Y)$, $\min(X, Y^2)$?

Def. 11. ***Valoarea medie a unei variabilă aleatoare discrete*** *(numerice) X, care ia valorile $\{x_i, i \in I\}$, este*

$$E(X) = \sum_{i \in I} x_i P(X = x_i),$$

dacă $\sum_{i \in I} |x_i| P(X = x_i) < \infty$.

▷ Valoarea medie a unei variabile aleatoare caracterizează *tendinţa centrală* a valorilor acesteia.

P. 9. *Fie X v.a. discretă. Au loc proprietăţile:*
→ $E(aX+b) = aE(X)+b$ *pentru orice* $a, b \in \mathbb{R}$*;*
→ $E(X+Y) = E(X)+E(Y)$;
→ *Dacă X şi Y sunt variabile aleatoare independente, atunci* $E(X \cdot Y) = E(X)E(Y)$.
→ *Dacă* $g : \mathbb{R} \to \mathbb{R}$ *e o funcţie, atunci* $E(g(X)) = \sum_{i \in I} g(x_i)P(X = x_i)$ *(dacă* $\sum_{i \in I} |g(x_i)|P(X = x_i) < \infty$*).*

Matlab/Octave: *mean*

Exemplu: Joc: Se aruncă un zar; dacă apare 6, se câştigă 3 u.m. (unităţi monetare), dacă apare 1 se câştigă 2 u.m., dacă apare 2,3,4,5 se pierde 1 u.m. În medie cât va câştiga sau pierde un jucător după 30 de repetiţii ale jocului?
Răspuns: Fie X v.a. care indică venitul la un joc

$$X \sim \begin{pmatrix} 2 & -1 & 3 \\ \frac{1}{6} & \frac{4}{6} & \frac{1}{6} \end{pmatrix}$$

Pentru $i \in \{1, ..., 30\}$ fie X_i venitul la al i-lea joc; X_i are aceeaşi distribuţie ca X. Venitul mediu al jucătorului după 30 de repetiţii ale jocului este

$$E(X_1 + ... + X_{30}) = E(X_1) + ... + E(X_{30}) = 30 \cdot E(X) = 30 \cdot \frac{1}{6} \cdot (2 - 4 + 3) = 5 \text{ (u.m.).}$$

Aşadar jucătorul *câştigă în medie* 5 u.m.
Problemă:
Input: Fie A(1),...,A(200) un vector cu 200 de elemente, din care 50 sunt egale cu 0, 70 egale cu 1 şi 80 sunt egale cu 2 (ordinea lor este necunoscută).
Output: Să se găsească un 0 în vector, alegând aleator un element din şir şi verificând dacă acesta este 0.
Întrebare: În *medie* câte iteraţii sunt necesare *înainte* să apară primul 0?

```
clear all
A=[zeros(1,50), zeros(1,70)+1,zeros(1,80)+2];
index=randperm(length(A));
A=A(index);
c=0;
i=randsample(length(A),1);
while A(i)~=0
c=c+1;
i=randsample(length(A),1); % i=randi(length(A));
end
fprintf('nr. iteratii: %d \n',c)

clc
clear all
A=[zeros(1,50), zeros(1,70)+1,zeros(1,80)+2];
s=[];
N=100;
for j=1:N
```

```
index=randperm(length(A));
A=A(index);
c=0;
i=randsample(length(A),1)
A(i)
while A(i)~=0
c=c+1;
i=randsample(length(A),1) % i=randi(length(A));
end
s=[s,c];
end
mean(s)
fprintf('nr. mediu de iteratii: %4.3f \n',mean(s))
```

Probabilitatea să apară la orice iteraţie 0 este $p = \frac{50}{200} = 0.25$.
Notăm cu X v.a. care indică numărul de iteraţii necesare *înainte* să apară primul $0 \Rightarrow X \sim Geo(p)$.
Numărul mediu de iteraţii necesare *înainte* să apară primul 0 este $E(X)$. Să se calculeze această valoare medie!

Variabile aleatoare continue

V.a. continuă: ia un număr infinit şi nenumărabil de valori într-un interval sau reuniune de intervale (v.a. poate lua orice valoare din intervalul considerat);
$\triangleright$ v.a. continue pot modela caracteristici fizice precum timp (de ex. timp de instalare, timp de aşteptare), greutate, lungime, poziţie, volum, temperatură (de ex. X e v.a. care indică durata de funcţionare a unui dispozitiv până la prima defectare; X e v.a. care indică temperatura într-un oraş la ora amiezii)
$\triangleright$ ea este caracterizată de o funcţie de densitate

Def. 12. ***Funcţia de densitate*** $f : \mathbb{R} \to \mathbb{R}$ *a unei v.a. continue este funcţia pentru care are loc*

$$P(X \leq x) = \int_{-\infty}^{x} f(t)dt, \ \forall\, x \in \mathbb{R}.$$

P. 10. *Fie f funcţia de densitate a unei v.a. continue X. Au loc proprietăţile:*
(1) $f(t) \geq 0$ *pentru orice* $t \in \mathbb{R}$*;*

(2) $\int_{-\infty}^{\infty} f(t)\, dt = 1;$

(3) $P(a < X \leq b) = \int_{a}^{b} f(t)dt, \ \forall\, a, b \in \mathbb{R}, a < b.$

Observaţie: Orice funcţie $f : \mathbb{R} \to \mathbb{R}$, care are proprietăţile (1), (2) din **P.10** este o funcţie de densitate.

Exemple de distribuţii clasice continue

Distribuţia uniformă pe un interval $[a, b]$**:** $X \sim Unif[a, b]$, $a, b \in \mathbb{R}$ cu $a < b$

• funcţia de densitate este

$$f(t) = \begin{cases} \frac{1}{b-a}, \text{pentru } t \in [a, b] \\ 0, \text{ pentru } t \in \mathbb{R} \setminus [a, b] \end{cases}$$

Matlab/Octave:
când $a = 0, b = 1$ *rand(M,N)* returnează o matrice $M \times N$ de valori aleatoare din [0,1]
unifrnd(a,b,M,N) sau *(b-a)rand(M,N)+a*

Distribuţia normală (Gauss): $X \sim N(m, \sigma^2)$, $m, \sigma \in \mathbb{R}$ cu $\sigma > 0$

• funcţia de densitate este

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t-m)^2}{2\sigma^2}\right), t \in \mathbb{R}.$$

• Pentru $m = 0, \sigma = 1$: $N(0,1)$ se numeşte *distribuţia standard normală.*

• Distribuţia normală se aplică în: măsurarea erorilor (de ex. termenul eroare în analiza regresională), în statistică (teorema limită centrală, teste statistice) etc.

Matlab/Octave:

normrnd(m,σ,M,N)

Friedrich Gauss şi legea normală $N(m, \sigma^2)$ (bancnota de 10 DM)

Distribuţia exponenţială: $X \sim Exp(\lambda)$, $\lambda \in \mathbb{R}$ cu $\lambda > 0$

• funcţia de densitate este

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, & \text{pentru } t > 0 \\ 0, & \text{pentru } t \leq 0 \end{cases}$$

Matlab/Octave:

exprnd($\frac{1}{\lambda}$,M,N)

Def. 13. ***Funcţia de repartiţie*** $F : \mathbb{R} \to [0,1]$ *a unei variabile aleatoare* X *(discrete sau continue) este*

$$F(x) = P(X \leq x) \quad \forall x \in \mathbb{R}.$$

P. 11. *Funcţia de repartiţie* F *a unei variabile aleatoare* X *(discrete sau continue) are următoarele proprietăţi:*

(1) F *este monoton crescătoare, adică pentru orice* $x_1 < x_2$ *rezultă* $F(x_1) \leq F(x_2)$.

(2) $\lim\limits_{x \to \infty} F(x) = 1$ *şi* $\lim\limits_{x \to -\infty} F(x) = 0$.

(3) F *este continuă la dreapta, adică* $\lim\limits_{x \searrow x_0} F(x) = F(x_0)\ \forall\, x_0 \in \mathbb{R}$.

(4) $P(a < X \leq b) = F(b) - F(a)\ \forall a, b \in \mathbb{R}, a < b$.

Observaţie importantă:

▷ Orice funcţie $F : \mathbb{R} \to \mathbb{R}$, care are proprietăţile (1), (2), (3) din **P.11** este o funcţie de repartiţie.

Matlab/Octave
- $X \sim Unif[a,b]$: *unifpdf(x,a,b)* calculează $f(x)$; *unifcdf(x,a,b)* calculează $F(x)$
- $X \sim N(m,\sigma^2)$: *normpdf(x,m,σ)* calculează $f(x)$; *normcdf(x,m,σ)* calculează $F(x)$
- $X \sim Exp(\lambda)$: *exppdf(x,$\frac{1}{\lambda}$)* calculează $f(x)$; *expcdf(x,$\frac{1}{\lambda}$)* calculează $F(x)$

V.a. discretă	V.a. continuă
• caracterizată de distribuţia de probabilitate discretă $X \sim \begin{pmatrix} x_i \\ P(X=x_i) \end{pmatrix}_{i\in I}$	• caracterizată de funcţia de densitate f care verifică $P(X \le x) = \int_{-\infty}^{x} f(t)dt$
• funcţia de repartiţie $F(x) = P(X \le x) \ \ \forall x \in \mathbb{R}$	• funcţia de repartiţie $F(x) = P(X \le x) \ \ \forall x \in \mathbb{R}$
• $F(x) = \sum_{i\in I: x_i \le x} P(X = x_i) \ \forall x \in \mathbb{R}$	• $F(x) = \int_{-\infty}^{x} f(t)dt \quad \forall x \in \mathbb{R}$
• F este funcţie continuă la dreapta	• F este funcţie continuă în orice punct $x \in \mathbb{R}$
• F este discontinuă în punctele $x_i, \ \forall\, i \in I$	
• $\forall\, a < b, a, b \in \mathbb{R}$ $P(a \le X \le b) = \sum_{\substack{i\in I \\ a\le x_i \le b}} P(X = x_i)$	• $\forall\, a < b, a, b \in \mathbb{R}$ $P(a \le X \le b) = P(a < X \le b)$ $= P(a \le X < b) = P(a < X < b) = \int_a^b f(t)\,dt$
• $P(X = a) = 0$ dacă $a \notin \{x_i : i \in I\}$	• $\forall\, a \in \mathbb{R} \Rightarrow P(X = a) = \int_a^a f(t)\,dt = 0$
	• dacă F este derivabilă în punctul $x \Rightarrow F'(x) = f(x)$.

Exemplu: Fie X v.a. care indică timpul de funcţionare neîntreruptă (în ore) până la prima defectare a unui aparat, pentru care $P(X > x) = 2^{-x}, x > 0$ şi $P(X > x) = 1, x \le 0$. Să se determine f_X şi $P(2 < X < 3)$.

Exemplu: V.a. X urmează legea uniformă pe $[a,b]$ cu $0 < a < b$. Să se arate că $P(X > 0) = 1$ şi să se determine funcţia densitate de probabilitate a variabilei aleatoare $Y = \ln\left(\frac{1}{X}\right)$.

Soluţie: Funcţia de densitate a lui X este:

$$f_X(x) = \begin{cases} \frac{1}{b-a}, \text{pentru } x \in [a,b] \\ 0, \text{ pentru } x \in \mathbb{R} \setminus [a,b] \end{cases}.$$

Atunci are loc

$$P(X > 0) = 1 - P(X \le 0) = 1 - \int_{-\infty}^{0} f_X(x)dx = 1.$$

Scriem succesiv

$$F_Y(y) = P(Y < y) = P\left(\ln\left(\frac{1}{X}\right) < y\right) = 1 - F_X(e^{-y}).$$

Derivăm în raport cu y

$$f_Y(y) = f_X(e^{-y})e^{-y}.$$

Folosind definiţia lui f_X, obţinem

$$f_Y(y) = \begin{cases} \dfrac{e^{-y}}{b-a}, & y \in [-\ln b, -\ln a] \\ 0, & \text{altfel}\,. \end{cases}$$

$\triangle$

Generarea de numere pseudo-aleatoare ce urmează o distribuţie discretă dată (metoda inversei)

Se dau $(x_1, \ldots, x_n)$ (valorile) şi $(p_1, \ldots, p_n)$ (probabilităţile lor). Realizaţi un program care generează N numere pseudo-aleatoare, care urmează *distribuţia discretă*

$$X \sim \begin{pmatrix} x_1 & x_2 & \ldots & x_n \\ p_1 & p_2 & \ldots & p_n \end{pmatrix},$$

folosind numere aleatoare uniform distribuite pe [0,1].

Procedeul de generare al numerelor aleatoare $Y(i)$, $i = \overline{1, N}$, este:

• Se citesc valorile $x_1, x_2, \ldots, x_n$ şi probabilităţile corespunzătoare $p_1, p_2, \ldots, p_n$, precum şi numărul N. Fie $p_0 = 0$.

• Se generează N numere aleatoare uniform distribuite pe [0,1]: $U(i)$, $i = \overline{1, N}$.

• Pentru fiecare $i = \overline{1, N}$: $Y(i) = x_k$ dacă şi numai dacă

$$p_0 + p_1 + \cdots + p_{k-1} < U(i) \leq p_0 + p_1 + \cdots + p_k,\ k \in \{1, \ldots, n\}.$$

• Se returnează: $Y(i), i = \overline{1, N}$.

Verificarea procedeului: deoarece U urmează legea uniformă, avem pe baza procedeului de mai sus:
$P(\text{“se generează } x_k\text{”}) = P(p_0 + p_1 + \cdots + p_{k-1} < U \leq p_0 + p_1 + \cdots + p_k) = p_k,\ k = 1, \ldots, n$, deci numerele generate urmează legea de distribuţie discretă dată.

Problemă: Conform statisticilor medicale 46% din oameni au grupa sanguină **0**, 40% au grupa sanguină **A**, 10% au grupa sanguină **B** şi 4% au grupa sanguină **AB**. Simulaţi de $N(= 100, 1000)$ ori stabilirea grupei sanguine a unei persoane alese aleator şi afişaţi frecvenţa de apariţie a fiecărei grupe sanguine. Comparaţi rezultatele obţinute cu cele teoretice.

```
function Y=ivtdiscret(x,p,N) %inverse transform method
Y=zeros(1,N);
q=cumsum(p);
for i=1:N
    U=unifrnd(0,1);
    Y(i)=x(find(U<=q,1));
end
%Aplicatia
clc
clear all
close all
N=1000;%numarul de simulari
x=[0,1,2,3];% valorile variabilei aleatoare
p=[0.46,0.4,0.1,0.04];%probabilitatile
Y=ivtdiscret(x,p,N);%generarea celor N numere
ny=hist(Y,length(unique(Y)));%determinarea frecventei de aparitie
[p' ny'/N]
```

Generarea de numere pseudo-aleatoare ce urmează o distribuţie continuă dată (metoda inversei)

Fie X o variabilă aleatoare ce are funcţia de repartiţie F. Din teorie se ştie că F este continuă şi monoton crescătoare. Presupunem că F *este inversabilă*, adică există F^{-1}: pentru orice $y \in (0,1)$ există un unic $x \in \mathbb{R}$ astfel încât $F(x) = y$, ceea ce este echivalent cu $F^{-1}(y) = x$.

Procedeul de generare a numerelor aleatoare $Y(i)$, $i = \overline{1,N}$, *care au aceeaşi distribuţie ca* X *este*:
• Se citeşte numărul N, se defineşte funcţia F^{-1}.
• Se generează N numere aleatoare uniform distribuite pe [0,1]: $U(i)$, $i = \overline{1,N}$.
• Pentru fiecare $i = \overline{1,N}$: $Y(i) = F^{-1}(U(i))$.
• Se afişează: $Y(i), i = \overline{1,N}$.

Verificarea procedeului: Fie variabila aleatoare $U \sim Unif[0,1]$ şi definim variabila aleatoare $Y = F^{-1}(U)$. Arătăm că Y are aceeaşi funcţie de repartiţie ca X: pentru orice $y \in \mathbb{R}$ are loc

$$F_Y(y) = P(Y \leq y) = P(F^{-1}(U) \leq y) = P(U \leq F(y)) = F(y),$$

deci Y are aceeaşi distribuţie ca X, pentru că $F_Y = F$.
Exemplu: Dacă $X \sim Exp(\lambda)$, atunci

$$F(x) = 1 - e^{-\lambda x},\ x > 0 \Longrightarrow F^{-1}(y) = -\frac{\ln(1-y)}{\lambda},\ y \in (0,1)$$

adică $-\frac{1}{\lambda}\ln(U) \sim Exp(\lambda)$, dacă $U \sim Unif[0,1]$.

Vector aleator continuu

Def. 14. $(X_1, \ldots, X_n)$ *este un* ***vector aleator continuu*** *dacă fiecare componentă a sa este o variabiă aleatoare continuă.*

Def. 15. $f_{(X,Y)} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ *este* ***funcţia de densitate a vectorului aleator continuu*** (X,Y)*, dacă*

$$P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{(X,Y)}(s,t)dsdt \ \ \forall x, y \in \mathbb{R}.$$

Def. 16. $F_{(X,Y)} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ *este* ***funcţia de repartiţie a vectorului aleator*** (X,Y) *(discret sau continuu), dacă*

$$F_{(X,Y)}(x,y) = P(X \leq x, Y \leq y) \quad \forall\, x, y \in \mathbb{R}.$$

Observaţie: Cunoscând distribuţia vectorului aleator continuu (X,Y), cum determinăm distribuţia v.a. continue X respectiv Y?
$\triangleright$ dacă se cunoaşte $f_{(X,Y)}$: f_X, respectiv f_Y, se determină cu:

$$f_X(x) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)dy,\ \forall x \in \mathbb{R}, f_Y(y) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)dx,\ \forall y \in \mathbb{R};$$

$\triangleright$ dacă se cunoaşte $F_{(X,Y)}$: F_X, respectiv F_Y, se determină cu:

$$F_X(x) = \lim_{y \to \infty} F_{(X,Y)}(x,y), \quad F_Y(y) = \lim_{x \to \infty} F_{(X,Y)}(x,y).$$

Def. 17. $X_1, \ldots, X_n$ *sunt* ***n variabilele aleatoare independente*** *(discrete sau continue), dacă*

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = P(X_1 \leq x_1) \cdot \ldots \cdot P(X_n \leq x_n)\ \forall\, x_1, \ldots, x_n \in \mathbb{R}.$$

Observaţie ($n = 2$ in definiţia de mai sus): X şi Y sunt **două variabilele aleatoare independente**, dacă

$$P(X \leq x, Y \leq y) = P(X \leq x) \cdot P(Y \leq y) \quad \forall\, x, y \in \mathbb{R},$$

adică

$$F_{(X,Y)}(x, y) = F_X(x) \cdot F_Y(y) \quad \forall\, x, y \in \mathbb{R}.$$

P. 12. *Variabilele aleatoare continue X (cu funcţia de densitate f_X) şi Y (cu funcţia de densitate f_Y) sunt independente, dacă şi numai dacă*

$$f_X(x) f_Y(y) = f_{(X,Y)}(x, y) \; \forall\, (x, y) \in \mathbb{R}^2,$$

unde $f_{(X,Y)}$ este funcţia de densitate a vectorului aleator (X, Y).

Exemplu: (X_1, X_2) are *distribuţie uniformă* pe $I = [a_1, b_1] \times [a_2, b_2]$, cu $a_1, a_2, b_1, b_2 \in \mathbb{R}$, $a_1 < b_1, a_2 < b_2$ dacă

$$f(x_1, x_2) = \begin{cases} \dfrac{1}{(b_1 - a_1)(b_2 - a_2)} & \text{dacă } (x_1, x_2) \in I \\ 0 & \text{dacă } (x_1, x_2) \notin I. \end{cases}$$

Funcţia de repartiţie are expresia

$$F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(t_1, t_2) dt_1 dt_2 = \left(\frac{x_1 - a_1}{b_1 - a_1} \right)_* \cdot \left(\frac{x_2 - a_2}{b_2 - a_2} \right)_*,$$

unde

$$u_* \;=\; \begin{cases} 0 & \text{dacă } u < 0 \\ u & \text{dacă } 0 \leq u \leq 1 \\ 1 & \text{dacă } 1 < u. \end{cases}$$

P. 13. *Pentru un vector aleator continuu (X, Y) au loc proprietăţile:*

1. $\displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{(X,Y)}(u, v)\, du dv = 1.$
2. *$F_{(X,Y)}$ este funcţie continuă.*
3. *Dacă $F_{(X,Y)}$ este derivabilă parţial în x şi y are loc:*

$$\frac{\partial^2 F_{(X,Y)}(x, y)}{\partial x \partial y} = f_{(X,Y)}(x, y).$$

4. $P((X, Y) \in A) = \underbrace{\int\int}_{A} f_{(X,Y)}(u, v) du dv, \; A \subset \mathbb{R}^2.$

Exemplu: Fie (X, Y) vector aleator continuu, având funcţia de repartiţie

$$F_{(X,Y)}(x,y) = \begin{cases} (1-e^{-x})(1-e^{-2y}) & \text{dacă } x>0 \text{ şi } y>0 \\ 0 & \text{în rest} \end{cases}$$

Sunt X şi Y v.a. independente? Să se calculeze $P(1 \le X \le 2 \le Y \le 3)$.
Soluţie: Se calculează $F_X(x) = 1 - e^{-x}$ pentru $x > 0$ şi $F_X(x) = 0$ pentru $x \le 0$, precum şi $F_Y(y) = 1 - e^{-2y}$ pentru $y > 0$ şi $F_Y(y) = 0$ pentru $y \le 0$. Se verifică

$$F_{(X,Y)}(x,y) = F_X(x) \cdot F_Y(y) \quad \forall\, x, y \in \mathbb{R}.$$

Deci, X şi Y sunt v.a. independente.
Se observă că $X \sim Exp(1)$ şi $Y \sim Exp(2)$.

$$P(1 \le X \le 2 \le Y \le 3) = \int_1^2 \int_2^3 f_X(u) f_Y(v) du dv = (e^{-1} - e^{-2})(e^{-4} - e^{-6}) \approx 0.00368.$$

Operaţii cu v.a. continue

Proprietate: Fie (X, Y) vector aleator continuu cu funcţia de densitate $f_{(X,Y)}$. Atunci $X + Y$ şi $X \cdot Y$ sunt v.a. continue, având funcţiile de densitate:

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_{(X,Y)}(u, z-u) du \quad \forall\, z \in \mathbb{R},$$

$$f_{X \cdot Y}(z) = \int_{-\infty}^{\infty} \frac{1}{|u|} f_{(X,Y)}\Big(u, \frac{z}{u}\Big) du \quad \forall\, z \in \mathbb{R}.$$

Dacă X şi Y sunt v.a. independente, atunci

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(u) f_Y(z-u) du \quad \forall\, z \in \mathbb{R},$$

$$f_{X \cdot Y}(z) = \int_{-\infty}^{\infty} \frac{1}{|u|} f_X(u) f_Y\Big(\frac{z}{u}\Big) du \quad \forall\, z \in \mathbb{R}.$$

Def. 18. ***Valoarea medie a unei v.a. continue*** *X, care are funcţia de densitate f, este*

$$E(X) = \int_{-\infty}^{\infty} t f(t) dt$$

dacă $\int_{-\infty}^{\infty} |t| f(t) dt < \infty$.

▷ Valoarea medie a unei variabile aleatoare caracterizează tendinţa centrală a valorilor acesteia.

P. 14. *Proprietăţie ale valorii medii; fie X, Y v.a. continue:*
$\rightarrow E(aX + b) = aE(X) + b$ *for all* $a, b \in \mathbb{R}$*;*
$\rightarrow E(X + Y) = E(X) + E(Y)$;
$\rightarrow$ *Dacă X şi Y sunt variabile aleatoare independente, atunci* $E(X \cdot Y) = E(X)E(Y)$.
$\rightarrow$ *Dacă* $g : \mathbb{R} \rightarrow \mathbb{R}$ *e o funcţie, astfel încât* $g(X)$ *este o v.a. continuă, atunci*

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx,$$

dacă $\int_{-\infty}^{\infty} |g(x)| f_X(x) dx < \infty$.

Exemplu: Dacă $X \sim N(m, \sigma^2)$, să se arate că $E(X) = m$.
Exemplu: Durata drumului parcurs de un elev dimineaţa de acasă până la şcoala este o v.a. uniform distribuită între 20 şi 26 minute. Dacă elevul porneşte la 7:35 (a.m.) de acasă şi are ore de la 8 (a.m.), care este probabilitatea ca elevul să ajungă la timp la şcoală? *În medie* cât durează drumul elevului până la şcoală?
Răspuns: fie X (v.a.) =durata drumului parcurs până la şcoală (în minute) $\Rightarrow X \sim Unif[20, 26]$

$$P(\text{“elevul ajunge la timp la şcoală”}) = P(X \leq 25) = \frac{25 - 20}{26 - 20} = \frac{5}{6}.$$

$$E(X) = \int_{20}^{26} x \frac{1}{26 - 20} dx = \frac{20 + 26}{2} = 23 \text{ (minute)}.$$

Def. 19. ***Varianţa (dispersia)*** *unei variabile aleatoare X (discrete sau continue) este*

$$V(X) = E\Big((X - E(X))^2\Big),$$

(dacă valoarea medie $E\Big((X - E(X))^2\Big)$ *există). Valoarea* $\sqrt{V(X)}$ *se numeşte* ***deviaţia standard*** *a lui X.*

▷ Varianţa unei variabile aleatoare caracterizează împrăştierea (dispersia) valorilor lui X în jurul valorii medii $E(X)$.

Exemplu: Dacă $X \sim N(m, \sigma^2)$, să se arate că $V(X) = \sigma^2$.

P. 15. *Proprietăţi ale varianţei:*
$\rightarrow V(X) = E(X^2) - E(X)^2$.
$\rightarrow V(aX + b) = a^2 V(X) \; \forall \; a, b \in \mathbb{R}$.
$\rightarrow$ *Dacă X şi Y sunt variabile aleatoare independente, atunci* $V(X + Y) = V(X) + V(Y)$.

Def. 20. ***Covarianţa*** *variabilelor aleatoare*

$$\text{cov}(X, Y) = E\Big((X - E(X))(Y - E(Y))\Big),$$

dacă există $E\Big((X - E(X))(Y - E(Y))\Big)$.
Dacă $\text{cov}(X, Y) = 0$*, atunci X şi Y sunt* ***variabile aleatoare necorelate***.
Coeficientul de corelaţie *(Pearson) al variabilelor aleatoare X şi Y*

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{V(X)V(Y)}},$$

dacă $\text{cov}(X, Y), V(X), V(Y)$ *există şi* $V(X) \neq 0, V(Y) \neq 0$.

▷ Coeficientul de corelaţie măsoară intensitatea relaţiei între variabilele X şi Y, fiind o măsură a dependenţei liniare între variabilele X şi Y.
▷ Dacă v.a. X şi Y sunt *independente*, atunci $\text{cov}(X,Y)=0$, adică X şi Y sunt v.a. *necorelate*.

P. 16. *Fie X, Y v.a. (discrete sau continue), au loc proprietăţile:*

(1) $\text{cov}(X,X)=V(X)$.

(2) $\text{cov}(X,Y)=E(X\cdot Y)-E(X)E(Y)$.

(3) $V(aX+bY)=a^2V(X)+b^2V(Y)+2ab\,\text{cov}(X,Y)\ \forall\ a,b\in\mathbb{R}$.

(4) $-1\le\rho(X,Y)\le 1$.

(5) $|\rho(X,Y)|=1\Rightarrow Y$ *depinde liniar de* X.

Matlab/Octave: *mean, var, std, cov, corrcoef*

Def. 21. $(X_n)_n$ *este* ***şir de v.a. independente****, dacă* $\forall\ \{i_1,\dots,i_k\}\subset\mathbb{N}$ *v.a.* $X_{i_1},\dots,X_{i_k}$ *sunt independente, adică*

$$P(X_{i_1}\le x_{i_1},\dots,X_{i_k}\le x_{i_k})=P(X_{i_1}\le x_{i_1})\cdot\dots\cdot P(X_{i_k}\le x_{i_k})$$

$\forall\ x_{i_1},\dots,x_{i_k}\in\mathbb{R}$.

Exemplu: a) X_n= v.a. care indică numărul apărut la a n-aruncare a unui zar $\Rightarrow (X_n)_n$ şir de v.a. independente
b) Se aruncă o monedă

$$X_n=\begin{cases}0 & :\ \text{la a } n\text{-a aruncare a apărut } \textit{cap},\\ 1 & :\ \text{la a } n\text{-a aruncare a apărut } \textit{pajură}.\end{cases}$$

$\Rightarrow (X_n)_n$ şir de v.a. independente.

Def. 22. *Şirul de v.a.* $(X_n)_n$ ***converge aproape sigur*** *la v.a. X, dacă*

$$P(\omega\in\Omega:\lim_{n\to\infty}X_n(\omega)=X(\omega))=1.$$

Notaţie: $X_n\overset{\text{a.s.}}{\to}X$

► Cu alte cuvinte, convergenţa a.s. $X_n\overset{\text{a.s.}}{\to}X$ impune ca $(X_n(\omega))_n$ să conveargă la $X(\omega)$ pentru fiecare $\omega\in\Omega$, *cu excepţia* unei mulţimi "mici" de probabilitate nulă, adică mulţimea (evenimentul)

$$M=\{\omega\in\Omega:(X_n(\omega))_n\ \textbf{nu converge}\ \text{la}\ X(\omega)\}\ \text{are}\ P(M)=0,\ \text{dacă}\ X_n\overset{a.s.}{\to}X.$$

Exemple:

$$X_n\sim\begin{pmatrix}-\frac{1}{n} & \frac{1}{n}\\ 0.5 & 0.5\end{pmatrix}\Rightarrow\ X_n\xrightarrow{a.s.}???$$

Exemplu:
$\Omega:=[0,1]$ spaţiul de selecţie, fie P probabilitatea pe [0,1] indusă de măsura Lebesgue pe [0,1], adică pentru $\forall\alpha<\beta$ din $[0,1]$ are loc

$$P\Big([\alpha,\beta]\Big)=P\Big([\alpha,\beta)\Big)=P\Big((\alpha,\beta]\Big)=P\Big((\alpha,\beta)\Big):=\beta-\alpha\ \text{(lungimea intervalului)}$$

1) $X_n(\omega) = \omega + \omega^n,\ \omega \in [0,1], n \geq 1 \Rightarrow\ X_n \xrightarrow{a.s.}$???
Răspuns:

$$\lim_{n\to\infty} X_n(\omega) = \begin{cases} \omega & \text{pentru } \omega \in [0,1) \\ 2 & \text{pentru } \omega = 1. \end{cases}$$

Fie $X(\omega) = \omega$ pentru fiecare $\omega \in \Omega$

$$\Rightarrow\ \{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = \omega\} = [0,1)$$

$$\Rightarrow\ P(\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = \omega) = P([0,1)) = 1.$$

$$X_n \overset{a.s.}{\to} X.$$

2) $X_n(\omega) = (-1)^n\omega,\ \omega \in [0,1], n \geq 1 \Rightarrow\ X_n \xrightarrow{a.s.}$???
Răspuns: $(X_n)_n$ nu converge a.s. spre o v.a.; şirul $((X_n(\omega))_n$ este convergent doar în $\omega = 0$.

Frecvenţe relative şi absolute

Fie A un eveniment asociat unei experienţe, repetăm experienţa de n ori (în aceleaşi condiţii date) şi notăm cu k_n numărul de realizări ale evenimentului A; **frecvenţa relativă** a evenimentului A este numărul

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

$k_n(A)$ este **frecvenţa absolută** a evenimentului A.

Experiment: Se aruncă o monedă de n ori; A: se obţine *pajură*

n	frecvenţă absolută $k_n(A)$	frecvenţă relativă $f_n(A)$
100	48	0.48
1000	497	0.497
10000	5005	0.5005

$h_n(A) \xrightarrow{a.s.} \frac{1}{2}$ (a se vedea **P.19**)

Legea tare a numerelor mari (LTNM)

Legea numerelor mari se referă la descrierea rezultatelor unui experiment repetat de foarte multe ori. Conform acestei legi, rezultatul mediu obţinut se apropie tot mai mult de valoarea aşteptată, cu cât experimentul se repetă de mai multe ori. Aceasta se explică prin faptul că abaterile aleatoare se compensează reciproc.
Legea numerelor mari are două formulări: legea slabă a numerelor mari (LSNM) şi legea tare a numerelor mari (LTNM).
△ **Scurt istoric**: Jacob Bernoulli (1655 -1705) a formulat LSNM pentru frecvenţa relativă a unui experiment şi a dat răspunsul la întrebarea "*Putem aproxima empiric probabilităţile*?" (în opera publicată postum, în 1713, *Ars conjectandi*):
▷ Teorema lui Bernoulli afirmă: *"Frecvenţele relative converg în probabilitate la probabilitatea teoretică."*
▷ În cadrul unui experiment poate să apară evenimentul A (*succes*) sau $\bar{A}$ (*insucces*).

• $X_i = 0 \Leftrightarrow$ dacă $\bar{A}$ apare în a i-a repetiţie a experimentului
• $X_i = 1 \Leftrightarrow$ dacă A apare în a i-a repetiţie a experimentului
$\Rightarrow X_i \sim Bernoulli(p)$ cu $p := P(A)$

$$X_i \sim \begin{pmatrix} 0 & 1 \\ 1 - P(A) & P(A) \end{pmatrix}$$

• $X_1, ..., X_n$ sunt v.a. independente
• frecvenţa relativă de apariţie a lui A este

$$f_n(A) = \frac{1}{n}(X_1 + ... + X_n);\ f_n(A) \text{ este o v.a..}$$

• $(X_n)_n$ verifică LSNM, adică

$$\lim_{n\to\infty} P\left(\left|f_n(A) - P(A)\right| > \varepsilon\right) = 0\ \forall\ \varepsilon > 0.$$

△

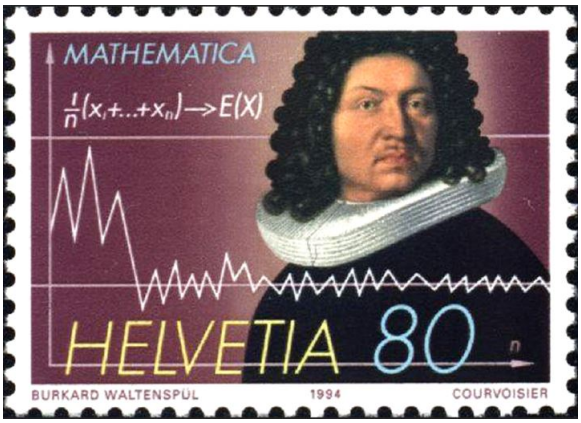


Fig. 5. Jacob Bernoulli (timbru emis în 1994 cu ocazia Congresului Internaţional al Matematicienilor din Elveţia)

Def. 23. *Şirul de v.a.* $(X_n)_n$ *cu* $E|X_n| < \infty\ \forall\ n \in \mathbb{N}$ *verifică* ***legea tare a numerelor mari (LTNM)*** *dacă*

$$P\left(\omega \in \Omega : \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}\Big(X_k(\omega) - E(X_k)\Big) = 0\right) = 1,$$

adică

$$\frac{1}{n}\sum_{k=1}^{n}\Big(X_k - E(X_k)\Big) \xrightarrow{a.s.} 0.$$

P. 17. *Fir* $(X_n)_n$ *şir de v.a. independente cu* $E|X_n| < \infty\ \forall n \in \mathbb{N}$ *şi* $\sum_{n=1}^{\infty} \frac{1}{n^2}V(X_n) < \infty \Rightarrow (X_n)_n$ *verifică* ***LTNM****, adică*

$$\frac{1}{n}\sum_{k=1}^{n}\Big(X_k - E(X_k)\Big) \xrightarrow{a.s.} 0.$$

P. 18. *Fie* $(X_n)_n$ *şir de v.a. independente având aceeaşi distribuţie (notăm* $m = E(X_n)\ \forall\ n \in \mathbb{N}$*)* $\Rightarrow (X_n)_n$ *verifică* ***LTNM****, adică*

$$\frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{a.s.} E(X).$$

În simulări: $\frac{1}{n}(X_1 + \cdots + X_n) \approx E(X),$ dacă n este suficient de mare.
Aplicaţie Matlab/Octave: Simulare LTNM

```
clear all
clf
hold on
n=300;
x=unidrnd(6,1,n);
for i=1:n
    s(i)=sum(x(1:i))/i;
     y(i)=i;
     plot(y(i),s(i),'b.')
     plot(y(i),3.5,'k-')
end
plot(y,s,'r-')
xlabel('Nr. aruncari zar')
ylabel('Media numerelor aparute')
```

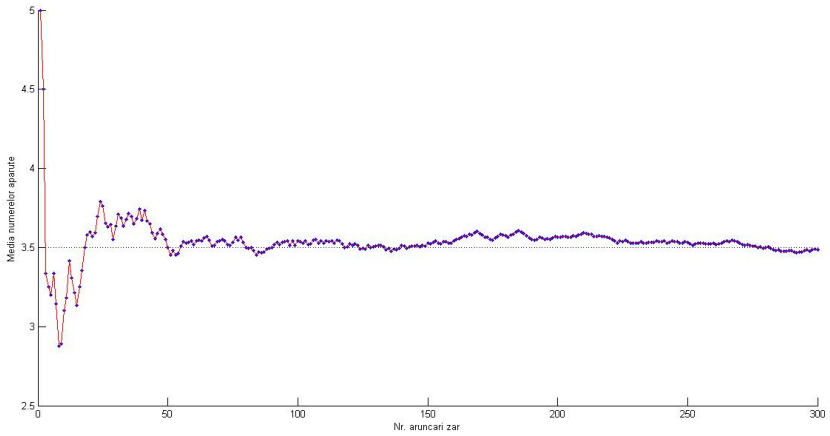


Fig. 4. LTNM

P. 19. *Fie A un eveniment asociat unei experienţe, repetăm experienţa de n ori (în aceleaşi condiţii date). LTNM: cu cât repetăm mai des un experiment ($n \to \infty$), cu atât mai bine aproximează frecvenţa relativă $f_n(A)$ a evenimentului A probabilitatea sa (teoretică) de apariţie $P(A)$:*

$$f_n(A) \xrightarrow{a.s.} P(A), \text{ dacă } n \to \infty.$$

În simulări: $f_n(A) \approx P(A),$ dacă n este suficient de mare.

Metode Monte Carlo pentru integrare numerică

Fie $g : [a, b] \to [0, \infty)$ o funcţie integrabilă dată şi $M > 0$ astfel încât $g(x) \leq M$, oricare ar fi $x \in [a, b]$. Considerăm următorii paşi pentru aproximarea integralei $\int_a^b g(x)\, dx$ folosind valori aleatoare.

Metoda 1:
► fie $(U_n)_n$ şir de v.a. independente cu $U_n \sim Unif[a,b]$ şi fie $X_n = g(U_n)$.
► $(X_n)_n$ satisface LTNM (a se vedea P.17), adică

$$\frac{1}{n}\sum_{k=1}^{n}\Big(X_k - E(X_k)\Big) \xrightarrow{a.s.} 0.$$

► are loc $E(X_k) = \dfrac{1}{b-a}\displaystyle\int_a^b g(t)dt \; \forall k \in \mathbb{N}$

$$\Rightarrow \frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow{a.s.} \frac{1}{b-a}\int_a^b g(t)dt$$

► în simulări:

$$\int_a^b g(t)dt \approx (b-a)\frac{1}{n}(g(U_1) + \cdots + g(U_n)) \text{ , dacă } n \text{ este suficient de mare}$$

Metoda 2:
► fie X_k, Y_k v.a. independente $X_k \sim Unif[a,b]$, $Y_k \sim Unif[0,M]$ $\forall k = \overline{1,n}$.
► fie $N_n = \#\{k \in \{1,...,n\} : Y_k \le g(X_k)\}$
► $\displaystyle\int_a^b g(t)dt =$ aria sub graficul funcţiei g

$$P(\text{ punct aleator sub graficul funcţiei } g) \approx \frac{N_n}{n}$$

$$\frac{\text{Aria sub graficul funcţiei } g}{\text{Aria dreptunghiului } [a,b]\times[0,M]} = \frac{1}{(b-a)M}\int_a^b g(t)dt \quad \text{(probabilitate geometrică)}$$

$$\Rightarrow \int_a^b g(t)dt \approx \frac{N_n}{n}\cdot(b-a)M \text{ , dacă } n \text{ este suficient de mare}$$

► în simulări:

$$\Rightarrow \int_a^b g(t)dt \approx \frac{\#\{k \in \{1,...,n\} : Y_k \le g(X_k)\}}{n}\cdot(b-a)M \text{ , dacă } n \text{ este suficient de mare.}$$

Statistică matematică

► Statistica matematică este o ramură a matematicii aplicate, care se ocupă de colectarea, gruparea, analiza şi interpretarea datelor referitoare la anumite fenomene în scopul obţinerii unor previziuni;
• statistica descriptivă: metode de colectare, organizare, sintetizare, prezentare şi descriere a datelor numerice (sau nenumerice) într-o formă convenabilă
• statistica inferenţială: metode de interpretare a rezultatelor obţinute prin metodele statisticii descriptive, utilizate apoi pentru luarea deciziilor.
► O *colectivitate* sau *populaţie statistică* $\mathcal{C}$ este o mulţime de elemente care au anumite însuşiri comune ce fac obiectul analizei statistice. Numărul elementelor populaţiei se numeşte *volumul populaţiei.*
Exemple de populaţii statistice: mulţimea persoanelor dintr-o anumită ţară, localitate, zonă etc. într-un anumit

an, multimea gospodăriilor din Romania la un moment dat, mulţimea consumatorilor unui produs, mulţimea societăţilor care produc un anumit produs, angajaţii unei societăţi, studenţii unei facultăţi.
► *Eşantionul* $\mathcal{E}$ reprezintă o submulţime a unei populaţii statistice $\mathcal{E} \subset \mathcal{C}$, constituită dupa criterii bine stabilite:
a) să fie aleatoare;
b) toate elementele colectivităţii să aibe aceeaşi şansă de a fi alese în eşantion;
c) eşantionul să fie reprezentativ (structura eşantionului să fie apropiată de structura populaţiei);
d) volumul eşantionului să fie suficient de mare.
► *Unitatea statistică* (indivizii) este elementul, entitatea de sine stătătoare a unei populaţii statistice, care posedă o serie de trăsături caracteristice ce-i conferă apartenenţa la populaţia studiată.
De exemplu: *unitatea statistică simplă*: un salariat, un student, un agent economic, o trăsătură, o părere; *unitatea statistică complexă*: o grupă de studenţi sau o echipă de salariaţi, o familie sau o gospodărie, o categorie de mărfuri.
► *Variabila statistică* sau *caracteristica* reprezintă o însuşire, o proprietate măsurabilă a unei unităţi statistice, întâlnită la toate unităţile care aparţin aceleiaşi colectivităţi şi care prezintă variabilitate de la o unitate statistică la alta. Caracteristica sau variabila statististică corespunde unei variabile aleatoare.
Exemple de caracteristici: vârsta, salariul, preferinţele politice, preţul unui produs, calitatea unor servicii, nivelul de studii.
a) variabile (caracteristici) continue, iau un număr infinit şi nenumărabil de valori într-un interval sau reuniune de intervale (de ex.: greutatea, înălţimea, valoarea glicemiei, temperatura aerului)
b) variabile (caracteristici) discrete, iau număr finit sau infinit dar numărabil de valori discrete (de ex.: numări elevi ai unei şcoli, numărul liceelor existente într-un oraş, valoarea IQ)
▷ acestea (a) si b)) sunt variabile numerice (cantitative)
c) variabile (caracteristici) nominale (de ex.: culoarea ochilor, ramura de activitate, religia)
d) variabile (caracteristici) nominale ordinale (de ex.: starea de sănătate / calitatea unor servicii - precară, mai bună, bună, foarte bună)
e) variabile (caracteristici) dihotomiale (binare) (de ex.: stagiul militar - satisfăcut/nesatisfăcut, starea civilă - căsătorit/necăsătorit, genul - masculin/feminin)
▷ acestea (c),d),e)) sunt variabile calitative
▷ variabilele nominale mai sunt numite variabile categoriale
► *Datele statistice* reprezintă observaţiile rezultate dintr-o cercetare statistică, sau ansamblul valorilor colectate în urma unei cercetări statistice.
De exemplu: un angajat al unei companii are o vechime de 6 ani în muncă. Angajatul reprezintă unitatea statistică, vechimea în muncă este caracteristica (variabila) cercetată, iar 6 este valoarea acestei caracteristici.

O *colectivitate* (populaţie) $\mathcal{C}$ este cercetatată din punctul de vedere al caracteristicii (variabilei statistice) X.

Distribuţia caracteristicii X poate fi

1) complet specificată (de ex.: $X \sim Exp(3)$, $X \sim Bin(10, 0.3)$, $X \sim N(0, 1)$)

2) specificată, dar depinzând de unul sau mai mulţi parametri necunoscuţi
(de ex.: $X \sim Exp(\lambda)$, $X \sim Bin(10, p)$, $X \sim N(m, \sigma^2)$)
3) necunoscută: $X \sim ?$

• în cazurile 2) şi 3) parametrii necunoscuţi sau distribuţia necunoscută
$\hookrightarrow$ se estimează $\rightarrow$ teoria estimaţiei
$\hookrightarrow$ se testează $\rightarrow$ teste statistice

► Fie $\mathcal{E} \subset \mathcal{C}$ un eşantion. Se numesc date de selecţie relative la caracteristica X datele statistice $x_1, \ldots, x_n$ obţinute prin cercetarea indivizilor care fac parte din eşantionul $\mathcal{E}$.
► Datele de selecţie $x_1, \ldots, x_n$ pot fi considerate ca fiind valorile unor variabile aleatoare $X_1, \ldots, X_n$, numite

variabile de selecţie şi care se consideră a fi variabile aleatoare independente şi având aceeaşi distribuţie ca X.
► Fie $x_1, \ldots, x_n$ datele statistice pentru caracteristica cercetată X, notăm cu $X_1, \ldots, X_n$ variabilele de selecţie corespunzătoare. Fie $g : \mathbb{R}^n \to \mathbb{R}$ o funcţie astfel încât $g(X_1, \ldots, X_n)$ este o variabilă aleatoare.
$g(X_1, \ldots, X_n)$ se numeşte funcţie de selecţie sau estimator
$g(X_1, \ldots, X_n)$ se numeşte valoarea funcţiei de selecţie sau valoarea estimatorului.
► $g(X_1, \ldots, X_n)$ este estimator nedeplasat pentru parametrul necunoscut θ, dacă

$$E(g(X_1, \ldots, X_n)) = \theta.$$

► $g(X_1, \ldots, X_n)$ este estimator consistent pentru parametrul necunoscut θ, dacă

$$g(X_1, \ldots, X_n) \overset{a.s.}{\to} \theta.$$

► Fie $g_1(X_1, \ldots, X_n)$ şi $g_2(X_1, \ldots, X_n)$ estimatori nedeplasaţi pentru parametrul necunoscut θ. $g_1(X_1, \ldots, X_n)$ este mai eficient decât $g_2(X_1, \ldots, X_n)$, dacă $V(g_1) < V(g_2)$.

• **Exemple de estimatori (funcţii de selecţie)** sunt: media de selecţie, dispersia de selecţie, funcţia de repartiţie empirică (de selecţie), mediana de selecţie.
▷ Estimatorii (funcţiile de selecţie) se folosesc în statistică pentru estimarea punctuală a unor parametri necunoscuţi, pentru obţinerea unor intervale de încredere pentru parametri necunoscuţi, pentru verificarea unor ipoteze statistice.

Fie $x_1, \ldots, x_n$ datele statistice pentru caracteristica cercetată X, notăm cu $X_1, \ldots, X_n$ variabilele de selecţie corespunzătoare:

► media de selecţie

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

► valoarea mediei de selecţie

$$\bar{x}_n = \frac{1}{n}(x_1 + \cdots + x_n)$$

► varianţa (dispersia) de selecţie

$$\tilde{S}_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2$$

► valoarea varianţei (dispersiei) de selecţie

$$\tilde{s}_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x}_n)^2$$

► abaterea standard de selecţie

$$\tilde{S}_n = \left(\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2\right)^{\frac{1}{2}}$$

► valoarea abaterii standard de selecţie

$$\tilde{s}_n = \left(\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x}_n)^2\right)^{\frac{1}{2}}$$

► funcţia de repartiţie empirică $\hat{F}_n : \mathbb{R} \times \Omega \to \mathbb{R}$

$$\hat{F}_n(x) = \frac{\#\{i \in \{1, ..., n\} : X_i \leq x\}}{n}, x \in \mathbb{R}$$

► valoarea funcţiei de repartiţie empirice

$$\hat{F}_n(x) = \frac{\#\{i \in \{1, ..., n\} : x_i \leq x\}}{n}, x \in \mathbb{R}$$

► fie $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ variabilele de selecţie ordonate crescător; definim mediana de selecţie

$$\hat{X}_m = \begin{cases} X_{(\frac{n+1}{2})}, \text{ dacă } n \text{ este impar,} \\ \frac{1}{2}(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}) \text{ dacă } n \text{ este par.} \end{cases}$$

► fie $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ datele de selecţie ordonate crescător; valoarea medianei de selecţie este

$$\hat{x}_m = \begin{cases} x_{(\frac{n+1}{2})}, \text{ dacă } n \text{ este impar,} \\ \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}) \text{ dacă } n \text{ este par.} \end{cases}$$

► coeficientul de corelaţie Pearson a două caracteristici X şi Y (datele de selecţie corespunzătoare sunt $y_1, ..., y_n$)

$$r_P := \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}.$$

Observaţii:
1) Media de selecţie $\bar{X}_n$ este estimator nedeplasat şi consistent pentru media teoretică $E(X)$ a caracteristicii X (în practică $E(X) \approx \bar{x}_n$).
2) Varianţa (dispersia) de selecţie $\tilde{S}_n^2$ este estimator nedeplasat şi consistent pentru varianţa teoretică $V(X)$ a caracteristicii X (în practică $V(X) \approx \tilde{s}_n^2$).
3) Funcţia de repartiţie de selecţie calculată în $x \in \mathbb{R}$: $\hat{F}_n(x)$ este estimator nedeplasat şi consistent pentru $F_X(x)$ valoarea funcţiei de repartiţie teoretice în x (în practică $F_X(x) \approx \hat{F}_n(x)$).
4) Mediana de selecţie este un estimator nedeplasat pentru mediana teoretică $z_{0.5}$, dacă distribuţia caracteristicii X este simetrică.
5) r_P măsoară gradul de dependenţă liniară între variabilele X şi Y (are loc $-1 \leq r_P \leq 1$), r_P estimează coeficientul de corelaţie $\rho(X, Y)$

$$\begin{cases} r_P = -1 & \text{relaţie liniară negativă perfectă} \\ r_P = 0 & \text{nu există relaţie liniară} \\ r_P = 1 & \text{relaţie liniară pozitivă perfectă.} \end{cases}$$

Metoda momentelor pentru estimarea parametrilor necunoscuţi $\theta = (\theta_1, \ldots, \theta_r)$ pentru distribuţia caracteristicii cercetate X

de exemplu:

$X \sim Exp(\lambda)$ parametrul necunoscut: $\theta = \lambda$
$X \sim N(m, \sigma^2)$ parametri necunoscuţi: $(\theta_1, \theta_2) = (m, \sigma)$
$X \sim Unif[a, b]$ parametri necunoscuţi: $(\theta_1, \theta_2) = (a, b)$

Fie $x_1, \ldots, x_n$ datele statistice pentru caracteristica cercetată X şi fie $X_1, \ldots, X_n$ variabilele de selecţie corespunzătoare.
Se rezolvă sistemul

$$\begin{cases} E(X^k) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^k, \\ k = \{1, ..., r\} \end{cases}$$

cu necunoscutele $\theta_1, \ldots, \theta_r$.
Soluţia sistemului $\hat{\theta}_1, \ldots, \hat{\theta}_r$ este estimatorul pentru parametrii necunoscuţi ai distribuţiei caracteristicii X.

Exemplu: Folosind metoda momentelor, să se estimeze parametrul necunoscut $\theta := a$ pentru $X \sim Unif[0, a]$; se dau datele statistice: 0.1,0.3,0.9,0.49,0.12,0.31,0.98,0.73, 0.13,0.62.

Avem cazul: $r = 1$, calculăm $E(X) = \frac{a}{2}$, $n = 10$, $\bar{x}_n = 0.468$. Se rezolvă

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i \Longrightarrow \frac{a}{2} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Estimatorul pentru parametrul necunoscut a este

$$\hat{a}(X_1, ..., X_n) = \frac{2}{n}\sum_{i=1}^{n} X_i,$$

unde $X_1, \ldots, X_n$ sunt variabilele de selecţie. Valoarea estimatorului este

$$\hat{a}(x_1, ..., x_n) = \frac{2}{n}\sum_{i=1}^{n} x_i = 0.936\,.$$

Parametrul necunoscut a este estimat cu valoarea 0.936.
► Este $\hat{a}(X_1, ..., X_n)$ un estimator nedeplasat pentru parametrul a?

Metoda verosimilităţii maxime pentru estimarea parametrului necunoscut θ al distribuţiei caracteristicii cercetate X

Fie $x_1, \ldots, x_n$ datele statistice pentru caracteristica cercetată X şi fie $X_1, \ldots, X_n$ variabilele de selecţie corespunzătoare. Notăm

$$L(x_1, \ldots, x_n; \theta) = \begin{cases} P(X = x_1) \cdot \ldots \cdot P(X = x_n), \text{ dacă } X \text{ e v.a. discretă} \\ f_X(x_1) \cdot \ldots \cdot f_X(x_n), \text{ dacă } X \text{ e v.a. continuă.} \end{cases}$$

Aceasta este funcţia de verosimilitate pentru parametrul θ şi datele statistice $x_1, \ldots, x_n$.

Metoda verosimilităţii maxime se bazează pe principiul că valoarea cea mai verosimilă (cea mai potrivită) a parametrului necunoscut θ este aceea pentru care funcţia de verosimilitate $L(x_1, \ldots, x_n; \theta)$ ia valoarea maximă:

$$(1) \qquad L(x_1, \ldots, x_n; \hat{\theta}) = \max_{\theta} L(x_1, \ldots, x_n; \theta).$$

Se rezolvă sistemul $\frac{\partial L}{\partial \theta} = 0$ şi se arată că $\frac{\partial^2 L}{\partial \theta^2} < 0$. Deseori este mai practic să se considere varianta transformată
$\frac{\partial lnL}{\partial \theta} = 0$ cu $\frac{\partial^2 lnL}{\partial \theta^2} < 0$. În unele situaţii (1) se rezolvă prin alte metode.

Observaţie: Dacă distribuţia caracteristicii cercetate depinde de k parametri necunoscuţi $(\theta_1, \ldots, \theta_k)$ atunci se rezolvă sistemul

$$\frac{\partial L}{\partial \theta_j} = 0, j = \overline{1,k} \text{ şi se arată că matricea } \left(\frac{\partial^2 L}{\partial \theta_i \partial \theta_j}\right)_{1 \le i \le j \le k} \text{ este negativ definită.}$$

Se poate lucra şi cu varianta transformată:

$$\frac{\partial lnL}{\partial \theta_j} = 0, j = \overline{1,k} \text{ şi se arată că matricea } \left(\frac{\partial^2 lnL}{\partial \theta_i \partial \theta_j}\right)_{1 \le i \le j \le k} \text{ este negativ definită.}$$

O matrice M este negativ definită dacă $y^t M y < 0$ pentru orice $y \in \mathbb{R}^n \setminus \{0_n\}$.

Exemplu: Folosind metoda verosimilităţii maxime să se estimeze parametrul $\theta := p \in (0,1)$ al distribuţiei Bernoulli,
$X \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$, cu datele statistice: 0,1,1,0,0,0,1,0,1,0.

$\Rightarrow n = 10, x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 0 ...; P(X = x) = p^x (1-p)^{1-x}, x \in \{0,1\}$

$$\Rightarrow L(x_1, \ldots, x_n; p) = P(X = x_1) \cdot \ldots \cdot P(X = x_n) = p^{x_1 + \cdots + x_n} (1-p)^{n - (x_1 + \cdots + x_n)}$$

$$\Rightarrow lnL(x_1, \ldots, x_n; p) = (x_1 + \cdots + x_n) ln(p) + (n - (x_1 + \cdots + x_n)) ln(1-p)$$

$$\frac{\partial lnL}{\partial p} = 0 \Rightarrow p = \frac{1}{n}(x_1 + \cdots + x_n).$$

Are loc: $\frac{\partial^2 lnL}{\partial p^2} < 0$.

Estimatorul de verosimilitate maximă pentru parametrul necunoscut p este

$$\hat{p}(X_1, \ldots, X_n) = \frac{1}{n}(X_1 + \cdots + X_n) = \bar{X}_n,$$

unde $X_1, \ldots, X_n$ sunt variabilele de selecţie. Valoarea estimată este

$$\hat{p}(x_1, \ldots, x_n) = \frac{1}{n}(x_1 + \cdots + x_n) = \bar{x}_n = \frac{4}{10} = 0.4\,.$$

► Este $\hat{p}(X_1, ..., X_n)$ un estimator nedeplasat pentru parametrul p?

Intervale de încredere şi teste statistice

Noţiuni de bază

► Fie $\alpha \in (0,1)$ nivelul de semnificaţie (probabilitatea de risc). Cuantila de ordin α pentru distribuţia caracteristicii cercetate X este numărul $z_\alpha \in \mathbb{R}$ pentru care

$$P(X < z_\alpha) \leq \alpha \leq P(X \leq z_\alpha).$$

• dacă $\alpha = 0.5$ atunci $z_{0.5}$ se numeşte mediana
• dacă X este v.a. continuă, atunci: z_α este cuantilă de ordin $\alpha \iff P(X \leq z_\alpha) = \alpha \iff F_X(z_\alpha) = \alpha$

• $\alpha \cdot 100\%$ din valorile lui X sunt mai mici sau egale cu z_α

Distribuţii de probabilitate continue frecvent folosite în statistică

▷ distibuţia normală $N(m, \sigma^2)$
cuantila $z_\alpha = norminv(\alpha)$; funcţia de repartiţie $F_{N(0,1)}(x) = normcdf(x)$
▷ distibuţia Student $St(n)$
cuantila $t_\alpha = tinv(\alpha, n)$; funcţia de repartiţie $F_{St(n)}(x) = tcdf(x, n)$
▷ distibuţia Chi-pătrat $\chi^2(n)$
cuantila $c_\alpha = chi2inv(\alpha, n)$; funcţia de repartiţie $F_{\chi^2(n)}(x) = chi2cdf(x, n)$
De exemplu:
$norminv(0.01) = -2.3263, norminv(1 - 0.01) = 2.3263,$
$tinv(0.05, 10) = -1.8125, tinv(1 - 0.05, 10) = 1.8125,$
$chi2inv(0.05, 10) = 3.9403, chi2inv(1 - 0.05, 10) = 18.307.$

Intervale de încredere

Fie $x_1, \ldots, x_n$ datele statistice pentru caracteristica cercetată X, a cărei distribuţie depinde de parametrul necunoscut θ; notăm cu $X_1, \ldots, X_n$ variabilele de selecţie corespunzătoare. Fie $\alpha \in (0,1)$ *nivelul de semnificaţie*; $1 - \alpha$ se numeşte *nivelul de încredere*.
Se caută doi estimatori $g_1(X_1, \ldots, X_n)$ şi $g_2(X_1, \ldots, X_n)$ astfel încât

$$P(g_1(X_1, \ldots, X_n) < \theta < g_2(X_1, \ldots, X_n)) = 1 - \alpha \;\Leftrightarrow\; P\Big(\theta \notin \Big(g_1(X_1, \ldots, X_n), g_2(X_1, \ldots, X_n)\Big)\Big) = \alpha$$

► $\Big(g_1(X_1, \ldots, X_n), g_2(X_1, \ldots, X_n)\Big)$ se numeşte **interval de încredere bilateral pentru parametrul necunoscut** θ
► $\Big(g_1(x_1, \ldots, x_n), g_2(x_1, \ldots, x_n)\Big)$ este valoarea intervalului de încredere pentru parametrul necunoscut θ
► $g_1(X_1, \ldots, X_n)$ este limita inferioară a intervalului de încredere, valoarea sa este $g_1(x_1, \ldots, x_n)$
► $g_2(X_1, \ldots, X_n)$ este limita superioară a intervalului de încredere, valoarea sa este $g_2(x_1, \ldots, x_n)$
► probabilitatea ca parametrul necunoscut θ să fie în intervalul $\Big(g_1(X_1, \ldots, X_n), g_2(X_1, \ldots, X_n)\Big)$ este $1 - \alpha$ (nivelul de încredere)
► există şi **intervale de încredere unilaterale**: $\Big(-\infty, g_3(X_1, \ldots, X_n)\Big)$, $\Big(g_4(X_1, \ldots, X_n), \infty\Big)$, estimatorii g_3 şi g_4 sunt astfel încât

$$P(\theta < g_3(X_1, \ldots, X_n)) = 1 - \alpha, \text{ respectiv } \quad P(g_4(X_1, \ldots, X_n) < \theta) = 1 - \alpha$$

► $\Big(-\infty, g_3(x_1,\ldots,x_n)\Big)$ $\Big(g_4(x_1,\ldots,x_n),\infty\Big)$ sunt valorile intervalelor de încredere unilaterale pentru parametrul necunoscut θ

► probabilitatea ca parametrul necunoscut θ să fie în intervalul $\Big(-\infty, g_3(X_1,\ldots,X_n)\Big)$ este $1-\alpha$, respectiv probabilitatea ca θ să fie în intervalul $\Big(g_4(X_1,\ldots,X_n),\infty\Big)$ este $1-\alpha$.

Teoreme de bază folosite în Statistică:

P. 20. *(Teorema limită centrală) Fie $(X_n)_n$ un şir de v.a. independente, care au aceeaşi distribuţie. Fie $m = E(X_n)$ şi $\sigma^2 = V(X_n) > 0\ \forall\ n \geq 1$. Are loc*

$$\lim_{n\to\infty} P\left(\frac{\frac{1}{n}(X_1+\cdots+X_n)-m}{\frac{\sigma}{\sqrt{n}}} < b\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{b} e^{-\frac{t^2}{2}}dt = F_{N(0,1)}(b),$$

pentru orice $b \in \mathbb{R}$.

P. 21. *Fie $X_1,\ldots,X_n$ variabile de selecţie pentru $X \sim N(m,\sigma^2)$, atunci:*

(1) pentru media de selecţie are loc $\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$;

(2) pentru media de selecţie şi abaterea standard de selecţie are loc $\dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} \sim St(n-1)$;

(3) pentru varianţa de selecţie are loc $\frac{n-1}{\sigma^2}\tilde{S}_n^2 \sim \chi^2(n-1)$.

În aplicaţii:

• dacă $X \sim N(m,\sigma^2)$ sau $n > 30$ şi X are o distribuţie necunoscută, atunci $\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

• dacă X are distribuţie, care nu este distribuţia normală, se estimează pentru $n > 30$ (suficient de mare)

$$P\left(\bar{X}_n + \frac{a\sigma}{\sqrt{n}} < m < \bar{X}_n + \frac{b\sigma}{\sqrt{n}}\right) \approx F_{N(0,1)}(b) - F_{N(0,1)}(a) = normcdf(b,0,1) - normcdf(a,0,1).$$

• dacă luăm $a = -z_{1-\frac{\alpha}{2}} = z_{\frac{\alpha}{2}}, b = z_{1-\frac{\alpha}{2}}$ (cuantilele distribuţiei $N(0,1)$)

$$P\left(\bar{X}_n - \frac{\sigma}{\sqrt{n}}\cdot z_{1-\frac{\alpha}{2}} < m < \bar{X}_n + \frac{\sigma}{\sqrt{n}}\cdot z_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha$$

▷ valoarea intervalului de încredere bilateral pentru media teoretică $m = E(X)$, când varianţa este cunoscută, este

$$\left(\bar{x}_n - \frac{\sigma}{\sqrt{n}}\cdot z_{1-\frac{\alpha}{2}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}}\cdot z_{1-\frac{\alpha}{2}}\right)$$

• dacă $X \sim N(m,\sigma^2)$ sau $n > 30$ şi X are o distribuţie necunoscută, atunci $\dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} \sim St(n-1)$

• dacă X are distribuţie, care nu este distribuţia normală, se estimează pentru $n > 30$ (suficient de mare)

$$P\left(a < \frac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} < b\right) \approx F_{St(n-1)}(b) - F_{St(n-1)}(a) = tcdf(b,n-1) - tcdf(a,n-1).$$

• dacă luăm $a = -t_{1-\frac{\alpha}{2}} = t_{\frac{\alpha}{2}}, b = t_{1-\frac{\alpha}{2}}$ (cuantile ale distribuţiei $St(n-1)$)

$$P\left(\bar{X}_n - \frac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}} < m < \bar{X}_n + \frac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha$$

▷ valoarea intervalului de încredere bilateral pentru media teoretică $m = E(X)$, când varianţa este necunoscută, este

$$\left(\bar{x}_n - \frac{\tilde{s}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}, \bar{x}_n + \frac{\tilde{s}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\right)$$

• dacă $X \sim N(m, \sigma^2)$

$$P\left(a < \frac{n-1}{\sigma^2}\tilde{S}_n^2 < b\right) \approx F_{\chi^2(n-1)}(b) - F_{\chi^2(n-1)}(a) = chi2cdf(b, n-1) - chi2cdf(a, n-1).$$

• dacă luăm $a = c_{\frac{\alpha}{2}}, b = c_{1-\frac{\alpha}{2}}$ (cuantile ale distribuţiei $\chi^2(n-1)$)

$$P\left(\frac{n-1}{c_{1-\frac{\alpha}{2}}} \cdot \tilde{S}_n^2 < \sigma^2 < \frac{n-1}{c_{\frac{\alpha}{2}}} \cdot \tilde{S}_n^2\right) \approx 1-\alpha$$

▷ valoarea intervalului de încredere bilateral pentru varianţa teoretică σ^2

$$\left(\frac{n-1}{c_{1-\frac{\alpha}{2}}} \cdot \tilde{s}_n^2, \frac{n-1}{c_{\frac{\alpha}{2}}} \cdot \tilde{s}_n^2\right)$$

Exemplu: Media de selecţie a lungimii a 100 de şuruburi este 15.5 cm, iar varianţa de selecţie este 0.09 cm^2. Să se construiască un interval de încredere 99% bilateral pentru media (teoretică) a lungimii şuruburilor.
Soluţie: valoarea intervalului de încredere bilateral pentru media teoretică m, când varianţa este necunoscută, este

$$\left(\bar{x}_n - \frac{\tilde{s}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}, \bar{x}_n + \frac{\tilde{s}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\right)$$

cu $\bar{x}_n = 15.5, \tilde{s}_n = 0.3$ $(\tilde{s}_n^2 = 0.09), t_{1-\frac{\alpha}{2}} = tinv(1-\frac{\alpha}{2}, 99) = 2.6264, \sqrt{n} = 10$

Interval de încredere pentru media $m = E(X)$ caracteristicii cercetate X, când dispersia $\sigma^2 = V(X)$ este cunoscută

► se dau $\alpha \in (0,1)$, σ, datele statistice $x_1, \ldots, x_n$

► dacă $X \sim N(m, \sigma^2)$ sau $n > 30$ şi X are o distribuţie necunoscută, atunci $\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$

► cuantilele legii normale $N(0,1)$:
$z_{1-\frac{\alpha}{2}} = norminv(1-\frac{\alpha}{2})$, $z_{1-\alpha} = norminv(1-\alpha)$, $z_\alpha = norminv(\alpha)$

• *interval de încredere bilateral*: se caută doi estimatori $g_1(X_1, \ldots, X_n)$ şi $g_2(X_1, \ldots, X_n)$ astfel încât pentru media teoretică $m = E(X)$ să avem:

$$P\Big(g_1(X_1, \ldots, X_n) < m < g_2(X_1, \ldots, X_n)\Big) = 1-\alpha$$

▷ din $P\left(\left|\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}\right| < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$ avem : $\left|\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}\right| < z_{1-\frac{\alpha}{2}} \Leftrightarrow \bar{X}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} < m < \bar{X}_n + \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$

• *intervalul de încredere bilateral* pentru $m = E(X)$ (media teoretică) este $\left(\bar{X}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\ ,\ \bar{X}_n + \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\right)$

▷ se calculează valoarea intervalului de încredere $\left(\bar{x}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\ ,\ \bar{x}_n + \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\right)$

• *interval de încredere unilateral*: se caută doi estimatori $g_3(X_1, \ldots, X_n)$, $g_4(X_1, \ldots, X_n)$ astfel încât pentru media teoretică $m = E(X)$ să avem:

$$P(m < g_3(X_1, \ldots, X_n)) = 1 - \alpha, \quad P(g_4(X_1, \ldots, X_n) < m) = 1 - \alpha$$

▷ din $P\left(\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} > z_\alpha\right) = 1 - \alpha$ avem : $\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} > z_\alpha \Leftrightarrow m < \bar{X}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_\alpha$

• $g_3(X_1, \ldots, X_n) = \bar{X}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_\alpha$; valoarea intervalul de încredere este $\left(-\infty\,, \bar{x}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_\alpha\right)$

▷ din $P\left(\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} < z_{1-\alpha}\right) = 1 - \alpha$ avem : $\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} < z_{1-\alpha} \Leftrightarrow m > \bar{X}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}$

• $g_4(X_1, \ldots, X_n) = \bar{X}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}$; valoarea intervalului de încredere este $\left(\bar{x}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}\ ,\ \infty\right)$

Interval de încredere pentru media $m = E(X)$ caracteristicii cercetate X, când dispersia $V(X)$ este necunoscută

► se dau $\alpha \in (0, 1)$, datele statistice $x_1, \ldots, x_n$

► dacă $X \sim N(m, \sigma^2)$ sau $n > 30$ şi X are o distribuţie necunoscută, atunci $\dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} \sim St(n-1)$

► cuantilele legii Student $St(n-1)$:
$t_{1-\frac{\alpha}{2}} = tinv(1-\frac{\alpha}{2}, n-1)$, $t_{1-\alpha} = tinv(1-\alpha, n-1)$, $t_\alpha = tinv(\alpha, n-1)$

• *interval de încredere bilateral*: se caută doi estimatori $g_1(X_1, \ldots, X_n)$ şi $g_2(X_1, \ldots, X_n)$ astfel încât pentru media teoretică $m = E(X)$ să avem:

$$P\Big(g_1(X_1, \ldots, X_n) < m < g_2(X_1, \ldots, X_n)\Big) = 1 - \alpha$$

▷ din $P\left(\left|\dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}}\right| < t_{1-\frac{\alpha}{2}}\right) = 1-\alpha$ avem : $\left|\dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}}\right| < t_{1-\frac{\alpha}{2}} \Leftrightarrow \bar{X}_n - \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}} < m < \bar{X}_n + \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}$

• *intervalul de încredere bilateral* pentru $m = E(X)$ (media teoretică) este $\left(\bar{X}_n - \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\ ,\ \bar{X}_n + \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\right)$

▷ se calculează valoarea intervalului de încredere $\left(\bar{x}_n - \dfrac{\tilde{s}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\ ,\ \bar{x}_n + \dfrac{\tilde{s}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\right)$

• *interval de încredere unilateral*: se caută doi estimatori $g_3(X_1, \ldots, X_n)$, $g_4(X_1, \ldots, X_n)$ astfel încât pentru media teoretică $m = E(X)$ să avem:

$$P(m < g_3(X_1, \ldots, X_n)) = 1 - \alpha, \quad P(g_4(X_1, \ldots, X_n) < m) = 1 - \alpha$$

$\triangleright$ din $P\left(\dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} > t_\alpha\right) = 1-\alpha$ avem : $\dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} > t_\alpha \Leftrightarrow m < \bar{X}_n - \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_\alpha$

• $g_3(X_1,\ldots,X_n) = \bar{X}_n - \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_\alpha$; valoarea intervalul de încredere este $\left(-\infty\,,\, \bar{x}_n - \dfrac{\tilde{s}_n}{\sqrt{n}} \cdot t_\alpha\right)$

$\triangleright$ din $P\left(\dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} < t_{1-\alpha}\right) = 1-\alpha$ avem : $\dfrac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} < t_{1-\alpha} \Leftrightarrow m > \bar{X}_n - \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\alpha}$

• $g_4(X_1,\ldots,X_n) = \bar{X}_n - \dfrac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\alpha}$; valoarea intervalului de încredere este $\left(\bar{X}_n - \dfrac{\tilde{s}_n}{\sqrt{n}} \cdot t_{1-\alpha}\,,\, \infty\right)$

Interval de încredere pentru varianţa (dispersia) $\sigma^2 = V(X)$ caracteristicii cercetate X

► se dau $\alpha \in (0,1)$, datele statistice $x_1,\ldots,x_n$
► dacă $X \sim N(m,\sigma^2)$, atunci $\frac{n-1}{\sigma^2}\tilde{S}_n^2 \sim \chi^2(n-1)$
► cuantilele χ^2 (Chi-pătrat) cu $n-1$ grade de libertate:
$c_{1-\frac{\alpha}{2}} = chi2inv(1-\frac{\alpha}{2}, n-1)$, $c_{\frac{\alpha}{2}} = chi2inv(\frac{\alpha}{2}, n-1)$, $c_{1-\alpha} = chi2inv(1-\alpha, n-1)$, $c_\alpha = chi2inv(\alpha, n-1)$

• *interval de încredere bilateral*: se caută doi estimatori $g_1(X_1,\ldots,X_n)$ şi $g_2(X_1,\ldots,X_n)$ astfel încât pentru varianţa teoretică $m = E(X)$ să avem:

$$P\Big(g_1(X_1,\ldots,X_n) < \sigma^2 < g_2(X_1,\ldots,X_n)\Big) = 1-\alpha$$

$\triangleright$ din $P\Big(c_{\frac{\alpha}{2}} < \frac{n-1}{\sigma^2}\cdot\tilde{S}_n^2 < c_{1-\frac{\alpha}{2}}\Big) = 1-\alpha$ avem : $c_{\frac{\alpha}{2}} < \frac{n-1}{\sigma^2}\cdot\tilde{S}_n^2 < c_{1-\frac{\alpha}{2}} \Leftrightarrow \frac{n-1}{c_{1-\frac{\alpha}{2}}}\cdot\tilde{S}_n^2 < \sigma^2 < \frac{n-1}{c_{\frac{\alpha}{2}}}\cdot\tilde{S}_n^2$

• *intervalul de încredere bilateral* pentru $\sigma^2 = V(X)$ (varianţa teoretică) este $\left(\frac{n-1}{c_{1-\frac{\alpha}{2}}}\cdot\tilde{S}_n^2 < \sigma^2 < \frac{n-1}{c_{\frac{\alpha}{2}}}\cdot\tilde{S}_n^2\right)$

$\triangleright$ se calculează valoarea intervalului de încredere $\left(\frac{n-1}{c_{1-\frac{\alpha}{2}}}\cdot\tilde{s}_n^2 < \sigma^2 < \frac{n-1}{c_{\frac{\alpha}{2}}}\cdot\tilde{s}_n^2\right)$

• *interval de încredere unilateral*: se caută doi estimatori $g_3(X_1,\ldots,X_n)$, $g_4(X_1,\ldots,X_n)$ astfel încât pentru varianţa teoretică $\sigma^2 = V(X)$ să avem:

$$P(\sigma^2 < g_3(X_1,\ldots,X_n)) = 1-\alpha, \quad P(g_4(X_1,\ldots,X_n) < \sigma^2) = 1-\alpha$$

$\triangleright$ din $P\left(\frac{n-1}{\sigma^2}\cdot\tilde{S}_n^2 > c_\alpha\right) = 1-\alpha$ avem : $\frac{n-1}{\sigma^2}\cdot\tilde{S}_n^2 > c_\alpha \Leftrightarrow \sigma^2 < \frac{n-1}{c_\alpha}\cdot\tilde{S}_n^2$

• $g_3(X_1,\ldots,X_n) = \frac{n-1}{c_\alpha}\cdot\tilde{S}_n^2$; valoarea intervalul de încredere este $\left(-\infty\,,\, \frac{n-1}{c_\alpha}\cdot\tilde{s}_n^2\right)$

$\triangleright$ din $P\left(\frac{n-1}{\sigma^2}\cdot\tilde{S}_n^2 < c_{1-\alpha}\right) = 1-\alpha$ avem : $\frac{n-1}{\sigma^2}\cdot\tilde{S}_n^2 < c_{1-\alpha} \Leftrightarrow \sigma^2 > \frac{n-1}{c_{1-\alpha}}\cdot\tilde{S}_n^2$

• $g_4(X_1,\ldots,X_n) = \frac{n-1}{c_{1-\alpha}}\cdot\tilde{S}_n^2$; valoarea intervalului de încredere este $\left(\frac{n-1}{c_{1-\alpha}}\cdot\tilde{s}_n^2\,,\, \infty\right)$

Teste statistice

Fie $x_1,\ldots,x_n$ datele statistice pentru caracteristica cercetată X, notăm cu $X_1,\ldots,X_n$ variabilele de selecţie corespunzătoare.
► Ipoteza statistică este o presupunere relativă la un parametru necunoscut θ
► Metoda de stabilire a veridicităţii unei ipoteze statistice se numeşte test (criteriu de verificare).
► Rezultatul testării se foloseşte apoi pentru luarea unor decizii (cum ar fi: eficienţa unor medicamente, strategii

de marketing, alegerea unui produs etc.).
► Se formulează ipoteza nulă H_0 şi ipoteza alternativă H_1, privind parametrul θ; fie θ_0 o valoare dată

$$\text{I. } H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0$$

$$\text{II. } H_0 : \theta \leq \theta_0 \quad H_1 : \theta > \theta_0$$

$$\text{III. } H_0 : \theta \geq \theta_0 \quad H_1 : \theta < \theta_0$$

Se dă $\alpha \in (0, 1)$ nivelul de semnificaţie (probabilitatea de risc). Formularea unui test revine la construirea unei regiuni critice $U \subset \mathbb{R}^n$ (pentru cazurile I, II, respectiv III) astfel încât

$$P((X_1, \ldots, X_n) \in U | H_0) = \alpha$$

ceea ce este echivalent cu

$$P((X_1, \ldots, X_n) \notin U | H_0) = 1 - \alpha$$

Concluzia testului:
$(x_1, \ldots, x_n) \notin U \Rightarrow$ ipoteza H_0 este admisă
$(x_1, \ldots, x_n) \in U \Rightarrow$ ipoteza H_0 este respinsă, în favoarea ipotezei H_1

► O colectivitate este testată în raport cu caracteristica X.
- test pentru valoarea medie $E(X)$
  - ▷ când varianţa teoretică $V(X)$ este cunoscută: testul lui Gauss (testul Z)
  - ▷ când varianţa teoretică $V(X)$ este necunoscută: Student Test (testul T)
- test pentru abaterea standard teoretică $\sqrt{V(X)}$ sau pentru varianţa teoretică $V(X)$: testul χ^2
- test asupra proporţiei (test Gauss aproximativ)
- test pentru independenţa a două caracteristici

Paşii în efectuarea unui test statistic:
- Care parametru se testează? Care test este potrivit?
- Care este ipoteza nulă H_0 şi care este ipoteza alternativă H_1?
- Care este nivelul de semnificaţie (probabilitatea de risc) α ?
- Calculul valorii estimatorului pe baza datelor statistice
- Concluzia testului

Test pentru media $m = E(X)$ caracteristicii cercetate X, când varianţa $\sigma^2 = V(X)$ este cunoscută (testul Z, testul Gauss)

► se dau $\alpha \in (0, 1)$, m_0, σ

► dacă $X \sim N(m, \sigma^2)$ sau $n > 30$ şi X are o distribuţie necunoscută, atunci $\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$

► folosind datele statistice $x_1, \ldots, x_n$, se calculează $z = \dfrac{\bar{x}_n - m_0}{\frac{\sigma}{\sqrt{n}}}$

► cuantilele legii normale $N(0, 1)$:
$z_{1-\frac{\alpha}{2}} = norminv(1 - \frac{\alpha}{2})$, $z_{1-\alpha} = norminv(1 - \alpha)$, $z_\alpha = norminv(\alpha)$

	I. H_0: $m = m_0$ H_1: $m \neq m_0$	II. H_0: $m \leq m_0$ H_1: $m > m_0$	III. H_0: $m \geq m_0$ H_1: $m < m_0$
Se acceptă H_0 dacă	$\lvert z\rvert < z_{1-\frac{\alpha}{2}}$	$z < z_{1-\alpha}$	$z > z_{\alpha}$
Se respinge H_0 în favoarea lui H_1, dacă	$\lvert z\rvert \geq z_{1-\frac{\alpha}{2}}$	$z \geq z_{1-\alpha}$	$z \leq z_{\alpha}$

► în Octave/Matlab: *ztest*

► *regiunea critică* $U \subset \mathbb{R}^n$ pentru testul mediei, când varianţa este cunoscută:

I. $U = \left\{(u_1, \ldots, u_n) \in \mathbb{R}^n : \left|\frac{\bar{u}_n - m_0}{\frac{\sigma}{\sqrt{n}}}\right| \geq z_{1-\frac{\alpha}{2}}\right\}$, unde $\bar{u}_n = \frac{1}{n}(u_1 + \cdots + u_n)$

II. $U = \left\{(u_1, \ldots, u_n) \in \mathbb{R}^n : \frac{\bar{u}_n - m_0}{\frac{\sigma}{\sqrt{n}}} \geq z_{1-\alpha}\right\}$

III. $U = \left\{(u_1, \ldots, u_n) \in \mathbb{R}^n : \frac{\bar{u}_n - m_0}{\frac{\sigma}{\sqrt{n}}} \leq z_{\alpha}\right\}$

Test pentru media $m = E(X)$ caracteristicii cercetate X, când varianţa $V(X)$ este necunoscută (Testul T, testul Student)

► se dau $\alpha \in (0,1)$, m_0

► dacă $X \sim N(m, \sigma^2)$ sau $n > 30$ şi X are o distribuţie necunoscută, atunci $\frac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} \sim St(n-1)$

► folosind datele statistice $x_1, \ldots, x_n$ se calculează $t = \frac{\bar{x}_n - m_0}{\frac{\tilde{s}_n}{\sqrt{n}}}$

► cuantilele legii Student cu $n-1$ grade de libertate $St(n-1)$:
$t_{1-\frac{\alpha}{2}} = tinv(1-\frac{\alpha}{2}, n-1)$, $t_{1-\alpha} = tinv(1-\alpha, n-1)$, $t_{\alpha} = tinv(\alpha, n-1)$

	I. H_0: $m = m_0$ H_1: $m \neq m_0$	II. H_0: $m \leq m_0$ H_1: $m > m_0$	III. H_0: $m \geq m_0$ H_1: $m < m_0$
Se acceptă H_0 dacă	$\lvert t\rvert < t_{1-\frac{\alpha}{2}}$	$t < t_{1-\alpha}$	$t > t_{\alpha}$
Se respinge H_0 în favoarea lui H_1, dacă	$\lvert t\rvert \geq t_{1-\frac{\alpha}{2}}$	$t \geq t_{1-\alpha}$	$t \leq t_{\alpha}$

► în Octave/Matlab: *ttest*

► regiunea critică $U \subset \mathbb{R}^n$ pentru testul mediei, când varianţa este necunoscută:

I. $U = \left\{(u_1, \ldots, u_n) \in \mathbb{R}^n : \left|\frac{\bar{u}_n - m_0}{\frac{\tilde{\sigma}_n}{\sqrt{n}}}\right| \geq t_{1-\frac{\alpha}{2}}\right\}$, unde $\bar{u}_n = \frac{1}{n}(u_1 + \cdots + u_n)$,

$\tilde{\sigma}_n = \left(\frac{1}{n-1}\sum_{k=1}^{n}(u_k - \bar{u}_n)^2\right)^{\frac{1}{2}}$

II. $U = \left\{(u_1, \ldots, u_n) \in \mathbb{R}^n : \frac{\bar{u}_n - m_0}{\frac{\tilde{\sigma}_n}{\sqrt{n}}} \geq t_{1-\alpha}\right\}$

III. $U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n : \dfrac{\bar{u}_n - m_0}{\frac{\tilde{\sigma}_n}{\sqrt{n}}} \leq t_\alpha \right\}$

Test asupra proporţiei p pentru caracteristica $X \sim Bernoulli(p)$ (testul Gauss aproximativ)

► se dau $\alpha \in (0,1)$, p_0

► dacă $X \sim Bernoulli(p)$ şi $np(1-p) \geq 10$, atunci $\dfrac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$

► folosind datele statistice $x_1, \ldots, x_n$ se calculează $z = \dfrac{\bar{x}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

► cuantilele legii normale $N(0,1)$:
$z_{1-\frac{\alpha}{2}} = norminv(1-\frac{\alpha}{2})$, $z_{1-\alpha} = norminv(1-\alpha)$, $z_\alpha = norminv(\alpha)$

	I. H_0: $p = p_0$ H_1: $p \neq p_0$	II. H_0: $p \leq p_0$ H_1: $p > p_0$	III. H_0: $p \geq p_0$ H_1: $p < p_0$
Se acceptă H_0 dacă	$\lvert z \rvert < z_{1-\frac{\alpha}{2}}$	$z < z_{1-\alpha}$	$z > z_\alpha$
Se respinge H_0 în favoarea lui H_1, dacă	$\lvert z \rvert \geq z_{1-\frac{\alpha}{2}}$	$z \geq z_{1-\alpha}$	$z \leq z_\alpha$

► regiunea critică $U \subset \mathbb{R}^n$

I. $U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n : \left| \dfrac{\bar{u}_n - p_0}{\frac{\sigma}{\sqrt{n}}} \right| \geq z_{1-\frac{\alpha}{2}} \right\}$, unde $\bar{u}_n = \dfrac{1}{n}(u_1 + \cdots + u_n)$

II. $U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n : \dfrac{\bar{u}_n - p_0}{\frac{\sigma}{\sqrt{n}}} \geq z_{1-\alpha} \right\}$

III. $U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n : \dfrac{\bar{u}_n - p_0}{\frac{\sigma}{\sqrt{n}}} \leq z_\alpha \right\}$

Test pentru varianţa $\sigma^2 = V(x)$ / abaterea standard $\sigma = \sqrt{V(x)}$ / a caracteristicii cercetate X

► se dau $\alpha \in (0,1)$, σ_0

► dacă $X \sim N(m, \sigma^2)$, atunci $\frac{n-1}{\sigma^2}\tilde{S}_n^2 \sim \chi^2(n-1)$

► folosind datele statistice $x_1, \ldots, x_n$ se calculează $c = \dfrac{n-1}{\sigma_0^2} \cdot \tilde{s}_n^2$

► cuantilele χ^2 (Chi-pătrat) cu $n-1$ grade de libertate:
$c_{1-\frac{\alpha}{2}} = chi2inv(1-\frac{\alpha}{2}, n-1)$, $c_{\frac{\alpha}{2}} = chi2inv(\frac{\alpha}{2}, n-1)$, $c_{1-\alpha} = chi2inv(1-\alpha, n-1)$, $c_\alpha = chi2inv(\alpha, n-1)$

	I. H_0: $\sigma = \sigma_0$ H_1: $\sigma \neq \sigma_0$	II. H_0: $\sigma \leq \sigma_0$ H_1: $\sigma > \sigma_0$	III. H_0: $\sigma \geq \sigma_0$ H_1: $\sigma < \sigma_0$
Se acceptă H_0, dacă	$c_{\frac{\alpha}{2}} < c < c_{1-\frac{\alpha}{2}}$	$c < c_{1-\alpha}$	$c > c_\alpha$
Se respinge H_0 în favoarea lui H_1, dacă	$c \notin (c_{\frac{\alpha}{2}}, c_{1-\frac{\alpha}{2}})$	$c \geq c_{1-\alpha}$	$c \leq c_\alpha$

► în Octave/Matlab: *vartest*
► regiunea critică $U \subset \mathbb{R}^n$ pentru testul varianţei:

I. $U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n : \frac{1}{\sigma_0^2} \sum_{k=1}^{n} (u_k - \bar{u}_n)^2 \notin \left(c_{\frac{\alpha}{2}}, c_{1-\frac{\alpha}{2}} \right) \right\}$, unde $\bar{u}_n = \frac{1}{n} (u_1 + \cdots + u_n)$

II. $U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n : \frac{1}{\sigma_0^2} \sum_{k=1}^{n} (u_k - \bar{u}_n)^2 \geq c_{1-\alpha} \right\}$

III. $U = \left\{ (u_1, \ldots, u_n) \in \mathbb{R}^n : \frac{1}{\sigma_0^2} \sum_{k=1}^{n} (u_k - \bar{u}_n)^2 \leq c_\alpha \right\}$

Test pentru independenţa a două caracteristici discrete X şi Y

► fie $\alpha \in (0, 1)$ probabilitatea de risc
► fie X v.a., care are valorile posibile $\{a_1, \ldots, a_r\}$ şi Y v.a., care are valorile posibile $\{b_1, \ldots, b_s\}$
► se dau datele statistice (x_i, y_j), $i \in I_X, j \in I_Y$, corespunzătoare caracteristicii (X, Y) (I_X, I_Y sunt mulţimile de indici)
► fie (X_i, Y_j), $i \in I_X, j \in I_Y$, perechile de variabile de selecţie corespunzătoare caracteristicii (X, Y)
► ipoteza nulă şi ipoteza alternativă
H_0 : X şi Y sunt independente
H_1 : X şi Y nu sunt independente

► se consideră estimatorii şi valorile lor corespunzătoare

Estimatorul	Valoarea estimatorului
• $N_{ij} = \#\{(k,l) \in I_X \times I_Y : X_k = a_i \text{ şi } Y_l = b_j\}$	$n_{ij} = \#\{(k,l) \in I_X \times I_Y : x_k = a_i \text{ şi } y_l = b_j\}$
• $N_{i\cdot} := \sum_{j=1}^{s} N_{ij}$	$n_{i\cdot} := \sum_{j=1}^{s} n_{ij}$
• $N_{\cdot j} := \sum_{i=1}^{r} N_{ij}$	$n_{\cdot j} := \sum_{i=1}^{r} n_{ij}$
• $N := \sum_{i=1}^{r} \sum_{j=1}^{s} N_{ij}$	$n := \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}$

► din punct de vedere teoretic are loc

$$\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(N_{ij} - \frac{N_{i\cdot} \cdot N_{\cdot j}}{N} \right)^2}{\frac{N_{i\cdot} \cdot N_{\cdot j}}{N}} \sim \chi^2((r-1)(s-1))$$

► practic, se calculează

$$x = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot} \cdot n_{\cdot j}}{n}}$$

şi se determină cuantila de ordin $1-\alpha$ a distribuţiei $\chi^2((r-1)(s-1))$, adică $c_{1-\alpha} = chi2inv(1-\alpha, (r-1)(s-1))$
► concluzia testului:
dacă $x \leq c_{1-\alpha}$**, atunci se acceptă** H_0
dacă $x > c_{1-\alpha}$**, atunci se respinge** H_0.

Erori în efectuarea testelor statistice

realitatea / decizia	H_0 este adevărată	H_1 este adevărată
se respinge H_0	Eroare de tip I	decizie corectă
se acceptă H_0	decizie corectă	Eroare de tip II

$P(\text{Eroare de tip I}) = P(\text{se respinge } H_0 | H_0 \text{ este adevărată}) = \alpha$

$P(\text{Eroare de tip II}) = P(\text{se acceptă } H_0 | H_1 \text{ este adevărată}) \overset{\text{notaţie}}{=} \beta$

Exemple de probleme:

1. Un profesor a înregistrat pe parcursul mai multor ani rezultatele elevilor săi. Calificativul unui elev este o v.a. cu valoarea între 1 şi 100, având abaterea standard egală cu 12. Actuala clasă are 36 de elevi şi media calificativelor lor este 73.2. Se poate afirma din punct de vedere statistic că media calificativelor din actuala clasă este egală cu 73.5? ($\alpha = 0.05$)
Soluţie: H_0: $m = 73.5$, H_1: $m \neq 73.5$, testul Z (Gauss) pentru medie, când varianţa este cunoscută $\sigma = 12$.

2. Specificaţiile unui anumit medicament indică faptul că fiecare comprimat conţine în medie 2.4 g de substanţă activă. 100 de comprimate alese la întâmplare din producţie sunt analizate şi se constată că ele conţin în medie 2.5 g de substanţă activă cu o deviaţie standard de 0.2 g. Se poate spune că medicamentul respectă specificaţiile (cu $\alpha = 0.01$)?
Soluţie: H_0: $m = 2.4$ cu H_1: $m \neq 2.4$, testul Student.

3. Un manager este suspicios că un utilaj, care umple anumite cutii cu ceai, trebuie înlocuit cu unul mult mai precis. 121 de cutii cu ceai sunt cântărite. S-a obţinut o medie de 196.6 g şi o abatere (deviaţie) standard de 2.09 g pentru acest eşantion.
a) Să se testeze dacă abaterea standard a utilajului este de 2 g.
b) Sunt datele suficiente pentru a concluziona, că utilajul trebuie reglat pentru că nu pune 200 g de ceai într-o cutie? ($\alpha = 0.01$)
Soluţie: a) H_0: $\sigma = 2$ cu H_1: $\sigma \neq 2$, test pentru abaterea standard (adică testul pentru varianţă)
b) H_0: $m = 200$ cu H_1: $m \neq 200$, testul Student.

4. Se dau datele statistice referitoare la preferinţele de vacanţă ale bărbaţilor (B) şi femeilor (F):

pref. / gen	plajă	munte
B	209	280
F	225	248

.

Sunt preferinţele de vacanţă independente de gen (B,F)? (pentru $\alpha = 0.05$)
Soluţie: test pentru independenţă; cele 2 caracteristici sunt
X: genul (valori posibile: *B*,*F*), $r = 2$;
Y: preferinţele de vacanţă (valori posibile: *plajă*, *munte*), $s = 2$.
$\triangleright$ din tabel avem: $n_{11} = 209, n_{12} = 280, n_{21} = 225, n_{22} = 248$
$\Rightarrow n_{1\cdot} = 489, n_{\cdot 1} = 434, n_{2\cdot} = 473, n_{\cdot 2} = 528, n = 962$

$$x = \frac{(209 - \frac{489\cdot 434}{962})^2}{\frac{489\cdot 434}{962}} + \frac{(280 - \frac{489\cdot 528}{962})^2}{\frac{489\cdot 528}{962}} + \frac{(225 - \frac{473\cdot 434}{962})^2}{\frac{473\cdot 434}{962}} + \frac{(248 - \frac{473\cdot 528}{962})^2}{\frac{473\cdot 528}{962}} \approx 2.2622$$

$\triangleright$ are loc $chi2inv(1 - 0.05, 1) = 3.8415 > x$, aşadar se acceptă ipoteza H_0, cele două caracteristici sunt independente, adică preferinţele de vacanţă nu depind de gen!

Reţele Bayes

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

Def. 24. *Evenimentele $A, B \in \mathcal{K}$ sunt **condiţional independente**, cunoscând evenimentul $C \in \mathcal{K}$, dacă şi numai dacă*

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

P. 22. *Au loc echivalenţele:*

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

Demonstraţie: • Pentru prima echivalenţă: "$\Rightarrow$"

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

"$\Leftarrow$"

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

• $P(A \cap B|C) = P(A|C)P(B|C) \Leftrightarrow P(B|A \cap C) = P(B|C)$ se demonstrează analog. ■

Def. 25. *Fie*
$\triangleright$ *X v.a. discretă, care ia valorile $\{x_i : i \in I\}$*
$\triangleright$ *Y v.a. discretă, care ia valorile $\{y_j : j \in J\}$*
$\triangleright$ *Z v.a. discretă, care ia valorile $\{z_k : k \in K\}$*

*V.a. X, Y sunt **condiţional independente** cunoscând v.a. Z, dacă şi numai dacă*

$$P(X = x_i, Y = y_j|Z = z_k) = P(X = x_i|Z = z_k)P(Y = y_j|Z = z_k) \ \ \forall i \in I, j \in J, k \in K.$$

Vom introduce noţiunea de condiţional independenţă pentru mai multe v.a. discrete.

Def. 26. *Pentru $i \in \{1, 2, ..., m\}$ fie X_i v.a. discrete, care ia valori din mulţimea $\mathcal{X}_i$; fie Z v.a. discretă, care ia valori din mulţimea $\mathcal{Z}$.* ***V.a.*** *$X_1, ..., X_m$* ***sunt condiţional independente cunoscând v.a.*** *Z, dacă şi numai dacă*

$$P(X_1 = x_1, ..., X_m = x_m | Z = z) = P(X_1 = x_1 | Z = z)...P(X_m = x_m | Z = z) \ \forall x_i \in \mathcal{X}_i, i = \overline{1, m}, z \in \mathcal{Z}.$$

Def. 27. *Fie $X, Y_1, ..., Y_m, Z_1, ..., Z_n$ v.a. discrete cu valori respectiv în mulţimile $\mathcal{X}$, $\mathcal{Y}_1, ..., \mathcal{Y}_m$, $\mathcal{Z}_1, ..., \mathcal{Z}_n$.* ***V.a.*** *$X$* ***este condiţional independentă de v.a.*** *$Y_1, ..., Y_m$****, cunoscând v.a.*** *$Z_1, ..., Z_n$, dacă şi numai dacă pentru orice $x \in \mathcal{X}, y_i \in \mathcal{Y}_i, i = \overline{1, m}, z_j \in \mathcal{Z}_j, j = \overline{1, n}$ are loc*

$$\begin{aligned} &P(X = x, \bigcap_{i=1}^{m} \{Y_i = y_i\} | \bigcap_{j=1}^{n} \{Z_j = z_j\}) \\ &= P(X = x | \bigcap_{j=1}^{n} \{Z_j = z_j\}) P(\bigcap_{i=1}^{m} \{Y_i = y_i\} | \bigcap_{j=1}^{n} \{Z_j = z_j\}) \,. \end{aligned}$$

P. 23. *Fie X v.a. care este condiţional independentă de v.a. $Y_1, ..., Y_m$, cunoscând v.a. $Z_1, ..., Z_n$. Fie $i_1, ..., i_k \in \{1, ..., m\}$ indici distincţi, atunci X este condiţional independentă de v.a. $Y_{i_1}, ..., Y_{i_k}$, cunoscând $Z_1, ..., Z_n$; pentru orice $x \in \mathcal{X}, y_{i_l} \in \mathcal{Y}_{i_l}, l = \overline{1, k}, z_j \in \mathcal{Z}_j, j = \overline{1, n}$ au loc relaţiile*

$$\begin{aligned} &P(X = x, \bigcap_{l=1}^{k} \{Y_{i_l} = y_{i_l}\} | \bigcap_{j=1}^{n} \{Z_j = z_j\}) \\ &= P(X = x | \bigcap_{j=1}^{n} \{Z_j = z_j\}) P(\bigcap_{l=1}^{k} \{Y_{i_l} = y_{i_l}\} | \bigcap_{j=1}^{n} \{Z_j = z_j\}) \,, \end{aligned}$$

$$P(X = x | \bigcap_{l=1}^{k} \{Y_{i_l} = y_{i_l}\}, \bigcap_{j=1}^{n} \{Z_j = z_j\}) = P(X = x | \bigcap_{j=1}^{n} \{Z_j = z_j\}) \,,$$

şi

$$P(\bigcap_{l=1}^{k} \{Y_{i_l} = y_{i_l}\} | X = x, \bigcap_{j=1}^{n} \{Z_j = z_j\}) = P(\bigcap_{l=1}^{k} \{Y_{i_l} = y_{i_l}\} | \bigcap_{j=1}^{n} \{Z_j = z_j\}).$$

Observaţie: Se mai foloseşte următoarea scriere compactă:
V.a. X **este condiţional independentă de v.a.** $Y_1, ..., Y_m$**, cunoscând v.a.** $Z_1, ..., Z_n$, dacă şi numai dacă

$$P(X, Y_1, ..., Y_m | Z_1, ..., Z_n) = P(X | Z_1, ..., Z_n) P(Y_1, ..., Y_m | Z_1, ..., Z_n).$$

Fie X v.a. care este condiţional independentă de v.a. $Y_1, ..., Y_m$, cunoscând v.a. $Z_1, ..., Z_n$. Properietăţile din **P.23** se pot rescrie în formă compactă astfel: dacă $i_1, ..., i_k \in \{1, ..., m\}$ sunt indici distincţi, atunci

$$\begin{aligned} &P(X, Y_{i_1}, ..., Y_{i_k} | Z_1, ..., Z_n) \\ &= P(X | Z_1, ..., Z_n) P(Y_{i_1}, ..., Y_{i_k} | Z_1, ..., Z_n) \,, \end{aligned}$$

$$P(X | Y_{i_1}, ..., Y_{i_k}, Z_1, ..., Z_n) = P(X | Z_1, ..., Z_n) \,,$$

şi

$$P(Y_{i_1}, ..., Y_{i_k}|X, Z_1, ..., Z_n) = P(Y_{i_1}, ..., Y_{i_k}|Z_1, ..., Z_n).$$

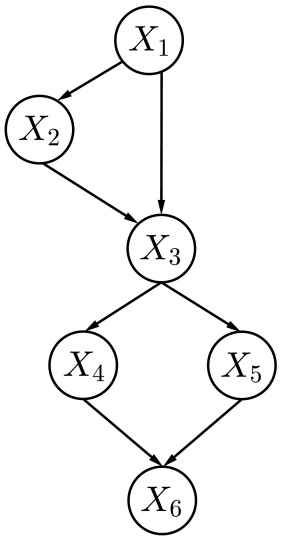
▲

Observaţie: Folosind **P.22** au loc echivalenţele:
V.a. X şi Y condiţional independente, cunoscând Z
$\Leftrightarrow$ $P(X, Y|Z) = P(X|Z)P(Y|Z)$
$\Leftrightarrow$ $P(X|Y, Z) = P(X|Z)$
$\Leftrightarrow$ $P(Y|X, Z) = P(Y|Z)$. ▲

Reţele Bayes

Reţeaua Bayes este un graf orientat aciclic (i.e. nu conţine niciun drum orientat închis).
▷ Nodul Y este *părinte* pentru nodul X, dacă există o muchie orientată de la Y la X. Mulţimea părinţilor lui X se notează cu $p(X)$. Dacă X este nod rădăcină, atunci $p(X) = \emptyset$.
▷ Nodul Y este *descendent* al nodului X, dacă există un drum orientat de la X la Y. Mulţimea descendenţilor lui X se notează cu $de(X)$.
Într-o reţea în care există o structură cauzală, nodurile din $p(X)$ reprezintă cauzele pentru X, iar nodurile din $de(X)$ sunt efectele nodului X.
▷ Nodul Y este *nondescendent* al nodului X, dacă nu este descendent al nodului X. Mulţimea nondescendeţilor lui X se notează cu $nd(X)$.
▷ Fiecare nod $X_1, ..., X_n$ din reţea este identificat cu o variabilă aleatoare şi este definit pe acelaşi spaţiu de probabilitate $(\Omega, \mathcal{K}, P)$; probabilităţile $P(X_j|p(X_j)), j = \overline{1,n}$ sunt date; are loc convenţia $P(X_j|p(X_j)) = P_{X_j}$, dacă X_j este nod rădăcină (P_{X_j} este distribuţia de probabilitate a lui X_j, a se vedea secţiunea cu v.a. discrete).
▷ Properietatea reţelei Bayes: orice nod X şi nondescendenţii săi $nd(X)$ sunt *condiţional independenţi*, dacă se cunosc valorile părinţilor $p(X)$; dacă $p(X) = \emptyset$, atunci X şi $nd(X)$ sunt independenţi.

Exemplul 1: Se dă reţeaua Bayes din figura alăturată, în care $X_1, ..., X_6$ sunt variabile aleatoare binare.
▷ Au loc proprietăţile:
• Mulţimile de noduri corespunzătoare părinţilor, descendenţilor, nondescendenţilor sunt:
$p(X_1) = \emptyset, p(X_2) = \{X_1\}, p(X_3) = \{X_1, X_2\},$
$p(X_4) = p(X_5) = \{X_3\}, p(X_6) = \{X_4, X_5\}$
$de(X_1) = \{X_2, X_3, X_4, X_5, X_6\},$
$de(X_2) = \{X_3, X_4, X_5, X_6\},$
$de(X_3) = \{X_4, X_5, X_6\},$
$de(X_4) = de(X_5) = \{X_6\}, de(X_6) = \emptyset,$
$nd(X_2) = \{X_1\}, nd(X_3) = \{X_1, X_2\},$
$nd(X_4) = \{X_1, X_2, X_3, X_5\}$
$nd(X_5) = \{X_1, X_2, X_3, X_4\},$
$nd(X_6) = \{X_1, X_2, X_3, X_4, X_5\};$
• probabilităţile (asociate nodurilor), care definesc reţeaua Bayes sunt:
$P_{X_1}, P(X_2|X_1), P(X_3|X_1, X_2), P(X_4|X_3), P(X_5|X_3), P(X_6|X_4, X_5);$
• independenţe condiţionate:
▷ X_4 este condiţional independentă de X_1, X_2, X_5, cunoscând X_3



Reţea Bayes

$$\Rightarrow P(X_4|X_1, X_2, X_3, X_5) = P(X_4|X_3),$$

$\triangleright$ X_5 este condiţional independentă de X_1, X_2, X_4, cunoscând X_3

$$\Rightarrow P(X_5|X_1, X_2, X_3, X_4) = P(X_5|X_3),$$

$\triangleright$ X_6 este condiţional independentă de X_1, X_2, X_3, cunoscând X_4, X_5

$$\Rightarrow P(X_6|X_1, X_2, X_3, X_4, X_5) = P(X_6|X_4, X_5);$$

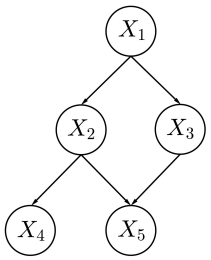
• (exemplu de calcul în reţea Bayes) se ştiu $P(X_1{=}1){=}0.5$, $P(X_2{=}1|X_1{=}1){=}0.6$, $P(X_3{=}1|X_1{=}1, X_2{=}1){=}0.5$, $P(X_4{=}1|X_3{=}1){=}0.4$, $P(X_4{=}1|X_3{=}0){=}0.3$, atunci să se calculeze $P(X_4{=}1, X_2{=}1, X_1{=}1)$:

$$\begin{aligned}
&P(X_4{=}1, X_2{=}1, X_1{=}1)\\
&{=}P(X_4{=}1, X_3{=}1, X_2{=}1, X_1{=}1) + P(X_4{=}1, X_3{=}0, X_2{=}1, X_1{=}1)\\
&{=}P(X_1{=}1)P(X_2{=}1|X_1{=}1)P(X_3{=}1|X_1{=}1, X_2{=}1)P(X_4{=}1|X_3{=}1)\\
&\quad + P(X_1{=}1)P(X_2{=}1|X_1{=}1)P(X_3{=}0|X_1{=}1, X_2{=}1)P(X_4{=}1|X_3{=}0)\\
&{=}0.105.
\end{aligned}$$

♦

Exemplul 2: Se dă reţeaua Bayes din figura alăturată, în care $X_1, ..., X_5$ sunt variabile aleatoare binare. Se ştiu probabilităţile:
$P(X_1 = 0) = 0.4$, $P(X_2 = 0|X_1 = 0) = 0.2$, $P(X_2 = 0|X_1 = 1) = 0.5$

$P(X_3 = 0|X_1 = 0) = 0.3$, $P(X_3 = 0|X_1 = 1) = 0.4$,
$P(X_4 = 0|X_2 = 0) = 0.2$, $P(X_4 = 0|X_2 = 1) = 0.5$,
$P(X_5 = 0|X_2 = 0, X_3 = 0) = 0.5$, $P(X_5 = 0|X_2 = 0, X_3 = 1) = 0.2$,
$P(X_5 = 0|X_2 = 1, X_3 = 0) = 0.7$, $P(X_5 = 0|X_2 = 1, X_3 = 1) = 0.4$.



Reţea Bayes

a) Să se calculeze

$$P(X_3 = 1|X_2 = 1), P(X_1 = 0, X_3 = 1), P\left(\bigcap_{i=1}^{5}\{X_i = 1\}\right).$$

b) Să se scrie distribuţia de probabilitate a variabilei aleatoare X_3.
Rezolvare: a) Se calculează: $P(X_1 = 1) = 1 - P(X_1 = 0) = 0.6$

$$P(X_2 = 1|X_1 = 0) = 1 - P(X_2 = 0|X_1 = 0) = 0.8; P(X_2 = 1|X_1 = 1) = 1 - P(X_2 = 0|X_1 = 1) = 0.5;$$
$$P(X_3 = 1|X_1 = 0) = 1 - P(X_3 = 0|X_1 = 0) = 0.7; P(X_3 = 1|X_1 = 1) = 1 - P(X_3 = 0|X_1 = 1) = 0.6;$$
$$P(X_4 = 1|X_2 = 0) = 1 - P(X_4 = 0|X_2 = 0) = 0.8; P(X_4 = 1|X_2 = 1) = 1 - P(X_4 = 0|X_2 = 1) = 0.5;$$
$$P(X_5 = 1|X_2 = 1, X_3 = 1) = 1 - P(X_5 = 0|X_2 = 1, X_3 = 1) = 0.6.$$

a) Are loc:

$$P(X_3 = 1|X_2 = 1) = \frac{P(X_3 = 1, X_2 = 1)}{P(X_2 = 1)}.$$

Folosind formula probabilităţilor totale şi proprietăţile reţelelor Bayes (X_2 este condiţional independentă de X_3, cunoscând X_1)[1]:

- $P(X_3 = 1, X_2 = 1) = P(X_3 = 1, X_2 = 1|X_1 = 0)P(X_1 = 0) + P(X_3 = 1, X_2 = 1|X_1 = 1)P(X_1 = 1) =$
$= P(X_3 = 1|X_1 = 0)P(X_2 = 1|X_1 = 0)P(X_1 = 0) + P(X_3 = 1|X_1 = 1)P(X_2 = 1|X_1 = 1)P(X_1 = 1)$
- $P(X_2 = 1) = P(X_2 = 1|X_1 = 0)P(X_1 = 0) + P(X_2 = 1|X_1 = 1)P(X_1 = 1)$.

[1]Orice nod X şi nondescendenţii săi $nd(X)$ sunt condiţional independenţi, dacă se dau valorile părinţilor $p(X)$.

Are loc

$$P(X_1 = 0, X_3 = 1) = P(X_3 = 1|X_1 = 0)P(X_1 = 0).$$

Folosind regula de înmulţire şi proprietăţile reţelelor Bayes (X_2 este condiţional independentă de X_3, cunoscând X_1; X_4 este condiţional independentă de X_1, X_3, cunoscând X_2; X_5 este condiţional independentă de X_1, X_4, cunoscând X_2, X_3)

$$\begin{aligned}P(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1, X_5 = 1) &= P(X_1 = 1)P(X_2 = 1|X_1 = 1)P(X_3 = 1|X_1 = 1, X_2 = 1) \cdot \\ &\cdot P(X_4 = 1|X_1 = 1, X_2 = 1, X_3 = 1)P(X_5 = 1|X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1) = \\ &= P(X_1 = 1)P(X_2 = 1|X_1 = 1)P(X_3 = 1|X_1 = 1)P(X_4 = 1|X_2 = 1)P(X_5 = 1|X_2 = 1, X_3 = 1).\end{aligned}$$

b) $P(X_3 = 0) = P(X_3 = 0|X_1 = 0)P(X_1 = 0) + P(X_3 = 0|X_1 = 1)P(X_1 = 1) = 0.12 + 0.24 = 0.36 \Rightarrow P(X_3 = 1) = 0.64$

$$\Rightarrow X_3 \sim \begin{pmatrix} 0 & 1 \\ 0.36 & 0.64 \end{pmatrix}.$$

♦