

# Teoria Probabilităților

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

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

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

Fenomene și 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 informației, modelarea traficului în rețea, analiza probabilistică a unor algoritmi, fiabilitatea sistemelor, algoritmi de simulare, machine learning, data mining, recunoașterea 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?

→ Rețelele de calculatoare: Care este probabilitatea ca un pachet de date să fie recepționat 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 execuție, spațiul de stocare, rezultatul obținut sunt variabile aleatoare (chiar dacă se folosesc aceleași valori input)

▷ la anumite tipuri de algoritmi corectitudinea e garantată doar cu o anumită probabilitate

▷ în mod paradoxal, incertitudinea ne poate oferi mai multă eficiență

Exemplu: Random QuickSort, în care elementul pivot este selectat aleator

• Algoritm de tip **Las Vegas** este un algoritm aleator, care returnează la fiecare execuție rezultatul corect (independent de alegerile aleatoare făcute); durata de execuție este o variabilă aleatoare.

Exemplu: Random QuickSort

• Un algoritm aleatoriu pentru care rezultatele obținute sunt corecte doar cu o anumită probabilitate se numește algoritm **Monte Carlo**.

↪ se examinează probabilitatea cu care rezultatul este corect; probabilitatea de eroare poate fi scăzută semnificativ prin execuții 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  $A(1), \dots, A(200)$  un vector cu 200 de elemente, din care 100 sunt egale cu 0 și 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 iterații

```

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 găsește niciun 0; probabilitatea de a găsi pe 0 după  $k$  iterații este

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

### Noțiuni introductive:

- **Experiența aleatoare** este acea experiență 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 cărți 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 experienței aleatoare

• **evenimentul sigur** este un eveniment care se realizează cu certitudine la fiecare efectuare a experienței aleatoare

• **spațiul de selecție**, notat cu  $\Omega$ , este mulțimea tuturor rezultatelor posibile ale experimentului considerat

◊ spațiul de selecție poate fi finit sau infinit

• dacă  $A$  este o submulțime a lui  $\Omega$  atunci  $A$  se numește **eveniment aleator**, iar dacă  $A$  are un singur element atunci  $A$  este un **eveniment elementar**.

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

**Exemplu:** Experimentul: aruncarea unui zar, spațiul de selecție:  $\Omega = \{e_1, e_2, e_3, e_4, e_5, e_6\}$ ,

$e_i$ : s-a obținut numărul  $i$  ( $i = 1, \dots, 6$ )

$e_1, e_2, e_3, e_4, e_5, e_6$  sunt evenimente elementare

$A$ : s-a obținut un număr par  $\Rightarrow A = \{e_2, e_4, e_6\}$

$\bar{A}$ : s-a obținut un număr impar  $\Rightarrow \bar{A} = \{e_1, e_3, e_5\}$

### Operații cu evenimente

- dacă  $A, B \subseteq \Omega$ , atunci **evenimentul reuniune**  $A \cup B$  este un eveniment care se produce dacă cel puțin unul din evenimentele  $A$  sau  $B$  se produce
- dacă  $A, B \subseteq \Omega$ , atunci **evenimentul intersecție**  $A \cap B$  este un eveniment care se produce dacă cele două evenimente  $A$  și  $B$  se produc în același timp
- dacă  $A \subseteq \Omega$  atunci **evenimentul contrar** sau **complementar**  $\bar{A}$  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 diferență**  $A \setminus B$  este un eveniment care se produce dacă  $A$  are loc și  $B$  nu are loc, adică

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

### Relații 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  $B$  și  $B$  implică  $A$ , atunci evenimentele  $A$  și  $B$  sunt **egale**:  $A = B$

### Proprietăți ale operațiilor între evenimente $A, B, C \subseteq \Omega$

Operațiile de reuniune și intersecție sunt operații **comutative**:

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

**asociative**

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

și **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  $\bar{\bar{A}} = A$ .

### Frecvența relativă și frecvența absolută

fie  $A$  un eveniment asociat unei experiențe, repetăm experiența de  $n$  ori (în aceleași condiții date) și notăm cu  $k_n(A)$  numărul de realizări ale evenimentului  $A$ ; **frecvența relativă** a evenimentului  $A$  este numărul

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

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

### Definiția clasică a probabilității

într-un experiment în care cazurile posibile sunt finite la număr și au aceleași șanse de a se realiza, **probabilitatea** unui eveniment  $A$  este numărul

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

▷ Prin repetarea de multe ori a unui experiment, în condiții practic identice, frecvența relativă  $f_n(A)$  de apariție a evenimentului  $A$  este aproximativ egală cu  $P(A)$

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

**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 și evenimentul  $A$  a apărut de 22 de ori.

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

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

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

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

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

### Definiția axiomatică a probabilității

Definiția clasică a probabilității 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 înțeles.

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

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

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

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

$P$  satisface anumite axiome (vezi Def. 3)

**Def. 2.** O familie  $\mathcal{K}$  de evenimente din spațiul de selecție  $\Omega$  se numește  **$\sigma$ -algebră** dacă sunt satisfăcute condițiile:

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

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

(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 numește **spațiu măsurabil**.

**Exemplu:** 1) Dacă  $A \subset \Omega$  atunci  $\mathcal{K} = \{\emptyset, A, \bar{A}, \Omega\}$  este o  $\sigma$ -algebră.

2)  $\mathcal{P}(\Omega)$ := mulțimea tuturor submulțimilor ale lui  $\Omega$  este o  $\sigma$ -algebră.

**P. 1.** Proprietăți ale unei  $\sigma$ -algebre: Dacă  $\mathcal{K}$  este o  $\sigma$ -algebră în  $\Omega$ , atunci au loc proprietățile:

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

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

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

**Def. 3.** Fie  $\mathcal{K}$  o  $\sigma$ -algebră în  $\Omega$ . O funcție  $P : \mathcal{K} \rightarrow \mathbb{R}$  se numește **probabilitate** dacă satisface axiomele:

- (i)  $P(\Omega) = 1$
- (ii)  $P(A) \geq 0$  pentru orice  $A \in \mathcal{K}$ ;
- (iii) pentru orice șir  $(A_n)_{n \in \mathbb{N}}$  de evenimente două câte două disjuncte (adică  $A_i \cap A_j = \emptyset$  pentru orice  $i \neq j$ ) din  $\mathcal{K}$  are loc

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

Tripletul  $(\Omega, \mathcal{K}, P)$  format din spațiul măsurabil  $(\Omega, \mathcal{K})$  și probabilitatea  $P : \mathcal{K} \rightarrow \mathbb{R}$  se numește **spațiu de probabilitate**.

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

- (1)  $P(\bar{A}) = 1 - P(A)$  și  $0 \leq P(A) \leq 1$
- (2)  $P(\emptyset) = 0$
- (3)  $P(A \setminus B) = P(A) - P(A \cap B)$
- (4)  $A \subseteq B \implies P(A) \leq P(B)$ , adică  $P$  este monotonă.
- (5)  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

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

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

### Șiruri de evenimente aleatoare

**P. 3.** Fie  $(\Omega, \mathcal{K}, P)$  un spațiu de probabilitate. Au loc următoarele proprietăți:

- (1) dacă  $(A_n)_n$  este un șir crescător de evenimente din  $\mathcal{K}$ , adică  $A_n \in \mathcal{K}$  și  $A_n \subseteq A_{n+1} \forall n \in \mathbb{N}$ , atunci

$$\lim_{n \rightarrow \infty} P(A_n) = P\left(\bigcup_{n=1}^{\infty} A_n\right).$$

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

$$\lim_{n \rightarrow \infty} P(B_n) = P\left(\bigcap_{n=1}^{\infty} B_n\right).$$

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

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

Spre ce valoare converge șirul  $\left(P(A_n)\right)_n$  ?

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

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

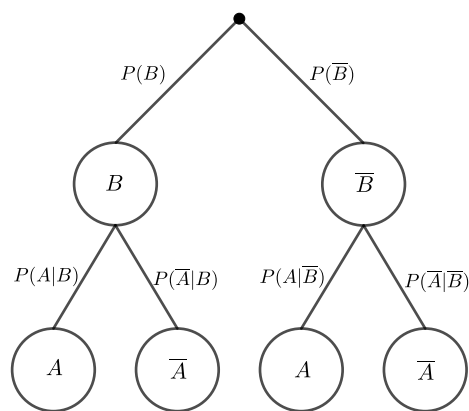
Spre ce valoare converge șirul  $\left(P(B_n)\right)_n$  ?

### Probabilitate condiționată

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

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

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



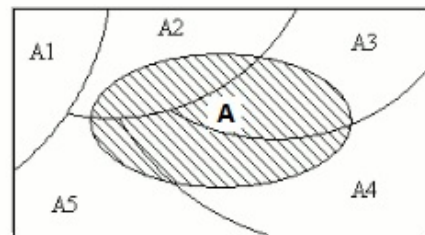
**P. 4.** Pentru  $A, B \in \mathcal{K}, P(A) > 0, P(B) > 0$  au loc:

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

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

**Fig.1. Probabilități condiționate**

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



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

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

**P. 5.** (Formula probabilității totale) Într-un spațiu de probabilitate  $(\Omega, \mathcal{K}, P)$  considerăm partiția  $A_1, \dots, A_n$  a lui  $\Omega$  cu  $P(A_i) > 0$  și  $A_i \in \mathcal{K} \forall i \in \{1, \dots, n\}$ , și 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 spațiu de probabilitate. Evenimentele  $A, B \in \mathcal{K}$  sunt **evenimente independente** dacă

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

**Observație:** Evenimentele  $A, B \in \mathcal{K}$  sunt **independente**, dacă apariția evenimentului  $A$ , nu influențează apariția evenimentului  $B$  și invers, adică

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

dacă  $P(A) > 0$  și  $P(B) > 0$ .

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

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

▷  $P(A_i)$  sunt **probabilități apriori** pentru  $A_i$ , numite și ipoteze (asertiuni),  $A$  se numește **evidență** (dovadă, premisă). Cu formula lui Bayes se calculează probabilitățile pentru ipoteze, cunoscând evidența:  $P(A_i|A)$ , acestea se numesc **probabilități posterioare**;  $P(A|A_i)$  reprezintă verosimilitatea datelor observate.

▷ Se pot calcula probabilitățile *cauzelor*, date fiind *efectele*; formula lui Bayes ne ajută să diagnosticăm o anumită situație sau să testăm o ipoteză.

**P. 7.** (Regula de înmulțire) Fie  $(\Omega, \mathcal{K}, P)$  un spațiu de probabilitate și fie  $A_1, \dots, A_n \in \mathcal{K}$  astfel încât

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

Atunci,

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

**Exemplu:** Într-o urnă sunt 2 bile verzi și 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 aceeași culoare?
- c) a doua bilă să fie albastră?
- d) prima bilă să fie verde, *știind* 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 și a treia tot albastră.

Răspuns: Notăm pentru  $i \in \{1, 2, 3\}$  :

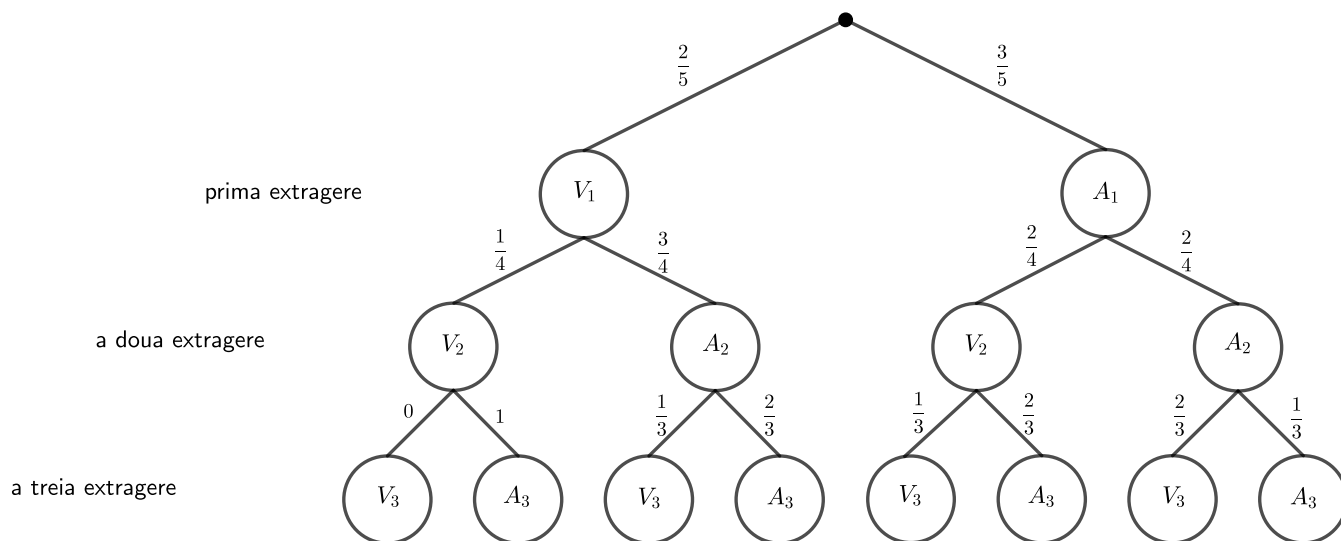
$A_i$ : la a  $i$ -a extragere s-a obținut bilă albastră;

$V_i$ : la a  $i$ -a extragere s-a obținut 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 probabilității 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 înmulțire a probabilităților 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 spațiu de probabilitate.  $B_1, \dots, B_n$  sunt  $n$  **evenimente independente** din  $\mathcal{K}$  dacă

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

pentru orice submulțime finită  $\{i_1, \dots, i_m\} \subseteq \{1, 2, \dots, n\}$ .

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

- (1)  $A$  și  $B$  sunt independente.
- (2)  $\overline{A}$  și  $B$  sunt independente.
- (3)  $A$  și  $\overline{B}$  sunt independente.
- (4)  $\overline{A}$  și  $\overline{B}$  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:  $\Rightarrow X : \Omega \rightarrow \{0, 1, 2\}$

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

▷ Prescurtare: **variabilă aleatoare**  $\rightarrow$  **v.a.**

O variabilă aleatoare este

**discretă**, dacă ia un număr finit de valori  $(x_1, \dots, x_n)$  sau un număr infinit numărabil de valori  $(x_1, \dots, x_n, \dots)$



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

**V.a. discrete:** exemple de v.a. numerice discrete: suma numerelor obținute 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. duminică); numărul de caractere transmise eronat într-un mesaj de o anumită lungime; exemple de v.a. categorice ( $\rightarrow$  se clasifică în categorii): prognoza meteo: ploios, senin, înnorat, cețos; calitatea unor servicii: nesatisfăcătoare, satisfăcătoare, bune, foarte bune, excepționale ...

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

### Variabile aleatoare numerice - definiție formală

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

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

### Variabile aleatoare discrete $X : \Omega \rightarrow \{x_1, x_2, \dots, x_i, \dots\}$

Distribuția de probabilitate a v.a. discrete  $X$

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

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

▷ Variabilele aleatoare discrete sunt caracterizate de distribuțiile lor.

▷  $\{X = x_i\} = \{\omega \in \Omega : X(\omega) = x_i\}$  este un eveniment din  $\mathcal{K}$  pentru fiecare  $i \in I$ .

### Distribuții discrete clasice

**Distribuția discretă uniformă:**  $X \sim Unif(n)$

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

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

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

Matlab/Octave: `unidrnd`

**Distribuția 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  
 $\Rightarrow X \sim \text{Bernoulli}(p)$  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)
```

**Distribuția binomială:**  $X \sim \text{Bino}(n, p), n \in \mathbb{N}^*, p \in (0, 1)$

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

- $A =$  succes cu  $P(A) = p$ ,  $\bar{A} =$  insucces  $P(\bar{A}) = 1 - p$
- se repetă experimentul de  $n$  ori
- v.a.  $X =$  numărul de succese în  $n$  repetări independente ale experimentului  $\Rightarrow$  valori posibile:  $X \in \{0, 1, \dots, n\}$

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

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

$\rightarrow$  de reamintit: formula binomială

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

pentru  $a = p$  și  $b = 1 - p$  se obține

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

Matlab/Octave: *binornd*

▷ Distribuția binomială corespunde modelului cu extragerea bilelor dintr-o urnă cu returnarea bilelor după fiecare extragere.

**Exemplu:** Într-o urnă sunt  $n_1$  bile albe și  $n_2$  bile negre. Se extrag cu returnare  $n$  bile, din care  $k$  sunt albe și  $n - k$  negre; fie v.a.  $X =$  numărul de bile albe extrase  $\Rightarrow X \sim \text{Bino}(n, p)$  cu  $p = \frac{n_1}{n_1 + n_2}$ .

**Exercițiu:** O rețea de laborator este compusă din 15 calculatoare. Rețeaua a fost atacată de un virus nou, care atacă un calculator cu o probabilitate 0.4, independent de alte calculatoare. Care este probabilitatea ca virusul a atacat a) cel mult 10 computere; b) cel puțin 10 calculatoare; c) exact 10 calculatoare.

**Distribuția hipergeometrică:**  $X \sim \text{Hyge}(n, n_1, n_2)$

Într-o urnă sunt  $n_1$  bile albe și  $n_2$  bile negre. Se extrag **fără returnare**  $n$  bile.

Fie v.a.  $X =$  numărul de bile albe extrase  $\Rightarrow$  valori posibile pentru  $X$  sunt  $\{0, 1, \dots, n^*\}$  cu

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

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

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

Matlab/Octave: `hygernd`

Exemplu: Loto 6 din 49  $\rightarrow$  Care este probabilitatea de a nimeri exact 4 numere câștigătoare?

Răspuns: Între cele 49 de bile exact  $n_1 = 6$  sunt câștigătoare (“bilele albe”) și  $n_2 = 43$  necâștigătoare (“bilele negre”). Care este probabilitatea ca din  $n = 6$  extrageri fără returnare, exact  $k = 4$  numere să fie câștigătoare?

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

**Distribuția geometrică**  $X \sim Geo(p)$

În cadrul unui experiment poate să apară evenimentul  $A$  (succes) sau  $\bar{A}$  (insucces)

- $A = \text{succes}$  cu  $P(A) = p$ ,  $\bar{A} = \text{insucces}$   $P(\bar{A}) = 1 - p$
- 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 apariția primului  $A$  (“succes”)  $\Rightarrow$  valori posibile:  $X \in \{0, 1, \dots\}$

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

Matlab/Octave: `geornd`

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

### Variabile aleatoare independente

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

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

Notăție:  $P(X = x_i, Y = y_j) = P(\{X = x_i\} \cap \{Y = y_j\})$ .

**Observație:**  $X$  și  $Y$  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.**  $(X_1, \dots, X_n)$  este un **vector aleator discret** dacă fiecare componentă a sa este o variabilă aleatoare discretă.

▷ Vectorii aleatori sunt caracterizați de distribuțiile lor! De exemplu, un vector aleator cu 2 componente:

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

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

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

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

Observație: Dacă  $X$  și  $Y$  sunt v.a. independente, atunci

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

▷ Dacă  $X$  și  $Y$  sunt v.a. independente, și se știu distribuțiile lor, atunci distribuția vectorului aleator  $(X, Y)$  se determină pe baza formulei (1).

▷ Dacă se cunoaște distribuția vectorului aleator  $(X, Y)$  distribuțiile lui  $X$  și  $Y$  se determină astfel:

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

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

### Operații cu variabile aleatoare (numerice)

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

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

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

Determinați:

a) distribuțiile variabilelor aleatoare  $X_1$  și  $X_2$ ;

b) distribuțiile variabilelor aleatoare  $X_1 + X_2$  și  $X_1 \cdot X_2, X_1^2 - 1$ ;

c) dacă variabilele aleatoare  $X_1$  și  $X_2$  sunt independente sau dependente.

$$a) X_1 \sim \begin{pmatrix} \frac{1}{16} & \frac{2}{16} \\ \frac{5}{16} & \frac{11}{16} \end{pmatrix} \text{ și } X_2 \sim \begin{pmatrix} 0 & 1 & 2 \\ \frac{3}{16} & \frac{6}{16} & \frac{7}{16} \end{pmatrix}.$$

$$b) X_1 + X_2 \sim \begin{pmatrix} 1 & 2 & 3 & 4 \\ \frac{2}{16} & \frac{2}{16} & \frac{7}{16} & \frac{5}{16} \end{pmatrix} \text{ și } X_1 \cdot X_2 \sim \begin{pmatrix} 0 & 1 & 2 & 4 \\ \frac{3}{16} & \frac{1}{16} & \frac{7}{16} & \frac{5}{16} \end{pmatrix}, X_1^2 - 1 \sim \begin{pmatrix} 0 & 3 \\ \frac{5}{16} & \frac{11}{16} \end{pmatrix}$$

$$c) X_1 \text{ și } X_2 \text{ nu sunt independente, pentru că } \frac{2}{16} = P(X_1 = 1, X_2 = 0) \neq P(X_1 = 1)P(X_2 = 0) = \frac{5}{16} \cdot \frac{3}{16}.$$

• Cunoscând distribuțiile variabilelor aleatoare independente (discrete)  $X$  și  $Y$ , cum se determină distribuția pentru  $X + Y, X \cdot Y$ ?

**Exercițiu:** Fie  $X, Y$  v.a. independente, având distribuțiile

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

a) Care sunt distribuțiile v.a.  $2X + 1, Y^2$ , dar distribuția vectorului aleator  $(X, Y)$ ?

b) Care sunt distribuțiile v.a.  $X + Y, X \cdot Y, \max(X, Y), \min(X, Y^2)$ ?

**Def. 11.** *Valoarea medie a unei variabile 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 *tendința centrală* a valorilor acesteia.

**P. 9.** Fie  $X$  v.a. discretă. Au loc proprietățile:

→  $E(aX + b) = aE(X) + b$  pentru orice  $a, b \in \mathbb{R}$ ;

→  $E(X + Y) = E(X) + E(Y)$ ;

→ Dacă  $X$  și  $Y$  sunt variabile aleatoare independente, atunci  $E(X \cdot Y) = E(X)E(Y)$ .

→ Dacă  $g : \mathbb{R} \rightarrow \mathbb{R}$  e o funcție, 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 câștigă 3 u.m. (unități monetare), dacă apare 1 se câștigă 2 u.m., dacă apare 2,3,4,5 se pierde 1 u.m. În medie cât va câștiga sau pierde un jucător după 30 de repetiții 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  $i \in \{1, \dots, 30\}$  fie  $X_i$  venitul la al  $i$ -lea joc;  $X_i$  are aceeași distribuție ca  $X$ . Venitul mediu al jucătorului după 30 de repetiții ale jocului este

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

Așadar jucătorul câștigă în medie 5 u.m.

**Problemă:**

Input: Fie  $A(1), \dots, A(200)$  un vector cu 200 de elemente, din care 50 sunt egale cu 0, 70 egale cu 1 și 80 sunt egale cu 2 (ordinea lor este necunoscută).

Output: Să se găsească un 0 în vector, alegând aleator un element din șir și verificând dacă acesta este 0.

**Întrebare:** În medie câte iterații 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 iterație 0 este  $p = \frac{50}{200} = 0.25$ .

Notăm cu  $X$  v.a. care indică numărul de iterații necesare *înainte* să apară primul 0  $\Rightarrow X \sim Geo(p)$ .

Numărul mediu de iterații 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 și nenumărabil de valori într-un interval sau reuniune de intervale (v.a. poate lua orice valoare din intervalul considerat);

▷ v.a. continue pot modela caracteristici fizice precum timp (de ex. timp de instalare, timp de așteptare), greutate, lungime, poziție, volum, temperatură (de ex.  $X$  e v.a. care indică durata de funcționare a unui dispozitiv până la prima defectare;  $X$  e v.a. care indică temperatura într-un oraș la ora amiezii)

▷ ea este caracterizată de o funcție de densitate

**Def. 12.** *Funcția de densitate*  $f : \mathbb{R} \rightarrow \mathbb{R}$  a unei v.a. continue este funcția pentru care are loc

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

**P. 10.** Fie  $f$  funcția de densitate a unei v.a. continue  $X$ . Au loc proprietățile:

(1)  $f(t) \geq 0$  pentru orice  $t \in \mathbb{R}$ ;

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

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

**Observație:** Orice funcție  $f : \mathbb{R} \rightarrow \mathbb{R}$ , care are proprietățile (1), (2) din **P.10** este o funcție de densitate.

### Exemple de distribuții clasice continue

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

• funcția 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  $(b-a)rand(M,N)+a$

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

- funcția 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 numește *distribuția standard normală*.
- Distribuția 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, sigma, M, N)`



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

**Distribuția exponențială:**  $X \sim \text{Exp}(\lambda)$ ,  $\lambda \in \mathbb{R}$  cu  $\lambda > 0$

- funcția 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(1/lambda, M, N)`

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

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

**P. 11. Funcția de repartiție  $F$  a unei variabile aleatoare  $X$  (discrete sau continue) are următoarele proprietăți:**

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

(2)  $\lim_{x \rightarrow \infty} F(x) = 1$  și  $\lim_{x \rightarrow -\infty} F(x) = 0$ .

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

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

**Observație importantă:**

▷ Orice funcție  $F : \mathbb{R} \rightarrow \mathbb{R}$ , care are proprietățile (1), (2), (3) din **P.11** este o funcție de repartiție.

Matlab/Octave

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

V.a. discretă

- caracterizată de distribuția de probabilitate discretă

$$X \sim \left( \begin{matrix} x_i \\ P(X = x_i) \end{matrix} \right)_{i \in I}$$

- funcția de repartiție  $F(x) = P(X \leq x) \quad \forall x \in \mathbb{R}$

$$F(x) = \sum_{i \in I: x_i \leq x} P(X = x_i) \quad \forall x \in \mathbb{R}$$

- $F$  este funcție continuă la dreapta

- $F$  este discontinuă în punctele  $x_i, \forall i \in I$

- $\forall a < b, a, b \in \mathbb{R}$

$$P(a \leq X \leq b) = \sum_{\substack{i \in I \\ a \leq x_i \leq b}} P(X = x_i)$$

- $P(X = a) = 0$  dacă  $a \notin \{x_i : i \in I\}$

V.a. continuă

- caracterizată de funcția de densitate  $f$  care verifică

$$P(X \leq x) = \int_{-\infty}^x f(t) dt$$

- funcția de repartiție  $F(x) = P(X \leq x) \quad \forall x \in \mathbb{R}$

$$F(x) = \int_{-\infty}^x f(t) dt \quad \forall x \in \mathbb{R}$$

- $F$  este funcție continuă în orice punct  $x \in \mathbb{R}$

- $\forall a < b, a, b \in \mathbb{R}$

$$P(a \leq X \leq b) = P(a < X < b)$$

$$= P(a \leq X < b) = P(a < X < b) = \int_a^b f(t) dt$$

$$\bullet \forall a \in \mathbb{R} \Rightarrow P(X = a) = \int_a^a f(t) dt = 0$$

- dacă  $F$  este derivabilă în punctul  $x \Rightarrow F'(x) = f(x)$ .

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

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

Se dau  $(x_1, \dots, x_n)$  (valorile) și  $(p_1, \dots, p_n)$  (probabilitățile lor). Realizați un program care generează  $N$  numere pseudo-aleatoare, care urmează distribuția discretă

$$X \sim \left( \begin{matrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{matrix} \right),$$

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

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

- Se citesc valorile  $x_1, x_2, \dots, x_n$  și probabilitățile corespunzătoare  $p_1, p_2, \dots, p_n$ , precum și 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 și numai dacă

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

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

Verificarea procedeului: deoarece  $U$  urmează legea uniformă, avem pe baza procedeului de mai sus:

$P(\text{"se generează } x_k") = P(p_0 + p_1 + \dots + p_{k-1} < U \leq p_0 + p_1 + \dots + p_k) = p_k, \quad k = 1, \dots, n$ , deci numerele generate urmează legea de distribuție discretă dată.



**Problemă:** Conform statisticilor medicale 46% din oameni au grupa sanguină **O**, 40% au grupa sanguină **A**, 10% au grupa sanguină **B** și 4% au grupa sanguină **AB**. Simulați de  $N(= 100, 1000)$  ori stabilirea grupei sanguine a unei persoane alese aleator și afișați frecvența de apariție a fiecărei grupe sanguine. Comparați rezultatele obținute 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 distribuție continuă dată (metoda inversei)

Fie  $X$  o variabilă aleatoare ce are funcția de repartiție  $F$ . Din teorie se știe că  $F$  este continuă și 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 aceeași distribuție ca  $X$  este:*

- Se citește numărul  $N$ , se definește funcția  $F^{-1}$ .
- 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 afișează:  $Y(i)$ ,  $i = \overline{1, N}$ .

*Verificarea procedeului:* Fie variabila aleatoare  $U \sim Unif[0, 1]$  și definim variabila aleatoare  $Y = F^{-1}(U)$ . Arătăm că  $Y$  are aceeași funcție de repartiție 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 aceeași distribuție ca  $X$ , pentru că  $F_Y = F$ .

**Exemplu:** Dacă  $X \sim Exp(\lambda)$ , atunci

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

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

### Vector aleator continuu

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

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

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

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

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

**Observație:** Cunoscând distribuția vectorului aleator continuu  $(X, Y)$ , cum determinăm distribuția v.a. continue  $X$  respectiv  $Y$ ?

▷ dacă se cunoaște  $f_{(X,Y)}$ :  $f_X$ , respectiv  $f_Y$ , se determină cu:

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

▷ dacă se cunoaște  $F_{(X,Y)}$ :  $F_X$ , respectiv  $F_Y$ , se determină cu:

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

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

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

**Observație** ( $n = 2$  în definiția de mai sus):  $X$  și  $Y$  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 funcția de densitate  $f_X$ ) și  $Y$  (cu funcția de densitate  $f_Y$ ) sunt independente, dacă și numai dacă

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

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

**Exemplu:**  $(X_1, X_2)$  are distribuție 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

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

Funcția de repartiție 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 proprietățile:

1.  $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{(X,Y)}(u, v) du dv = 1.$

2.  $F_{(X,Y)}$  este funcție continuă.

3. Dacă  $F_{(X,Y)}$  este derivabilă parțial în  $x$  și  $y$  are loc:

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

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

**Exemplu:** Fie  $(X, Y)$  vector aleator continuu, având funcția de repartiție

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

Sunt  $X$  și  $Y$  v.a. independente? Să se calculeze  $P(1 \leq X \leq 2 \leq Y \leq 3)$ .

Soluție: Se calculează  $F_X(x) = 1 - e^{-x}$  pentru  $x > 0$  și  $F_X(x) = 0$  pentru  $x \leq 0$ , precum și  $F_Y(y) = 1 - e^{-2y}$  pentru  $y > 0$  și  $F_Y(y) = 0$  pentru  $y \leq 0$ . Se verifică

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

Deci,  $X$  și  $Y$  sunt v.a. independente.

Se observă că  $X \sim \text{Exp}(1)$  și  $Y \sim \text{Exp}(2)$ .

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

### Operații cu v.a. continue

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

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

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

Dacă  $X$  și  $Y$  sunt v.a. independente, atunci

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

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

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

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

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

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

**P. 14.** Proprietăți ale valorii medii; fie  $X, Y$  v.a. continue:

→  $E(aX + b) = aE(X) + b$  for all  $a, b \in \mathbb{R}$ ;

→  $E(X + Y) = E(X) + E(Y)$ ;

→ Dacă  $X$  și  $Y$  sunt variabile aleatoare independente, atunci  $E(X \cdot Y) = E(X)E(Y)$ .

→ Dacă  $g : \mathbb{R} \rightarrow \mathbb{R}$  e o funcție, astfel încât  $g(X)$  este o v.a. continuă, atunci

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

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

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

**Exemplu:** Durata drumului parcurs de un elev dimineața de acasă până la școală este o v.a. uniform distribuită între 20 și 26 minute. Dacă elevul pornește la 7:35 (a.m.) de acasă și are ore de la 8 (a.m.), care este probabilitatea ca elevul să ajungă la timp la școală? În medie cât durează drumul elevului până la școală?

Răspuns: fie  $X$  (v.a.) = durata drumului parcurs până la școală (în minute)  $\Rightarrow X \sim Unif[20, 26]$

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

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

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

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

(dacă valoarea medie  $E((X - E(X))^2)$  există). Valoarea  $\sigma = \sqrt{V(X)}$  se numește **deviația standard** a lui  $X$ .

▷ Varianța unei variabile aleatoare caracterizează împrăștierea (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.** Proprietăți ale varianței:

→  $V(X) = E(X^2) - E(X)^2$ .

→  $V(aX + b) = a^2V(X) \forall a, b \in \mathbb{R}$ .

→ Dacă  $X$  și  $Y$  sunt variabile aleatoare independente, atunci  $V(X + Y) = V(X) + V(Y)$ .

**Def. 20. Covarianța** variabilelor aleatoare

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

dacă există  $E\left((X - E(X))(Y - E(Y))\right)$ .

Dacă  $\text{cov}(X, Y) = 0$ , atunci  $X$  și  $Y$  sunt **variabile aleatoare necorelate**.

**Coeficientul de corelație** (Pearson) al variabilelor aleatoare  $X$  și  $Y$

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

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

▷ Coeficientul de corelație măsoară intensitatea relației între variabilele  $X$  și  $Y$ , fiind o măsură a dependenței liniare între variabilele  $X$  și  $Y$ .

▷ Dacă v.a.  $X$  și  $Y$  sunt *independente*, atunci  $\text{cov}(X, Y) = 0$ , adică  $X$  și  $Y$  sunt v.a. *necorelate*.

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

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

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

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

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

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

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

**Def. 21.**  $(X_n)_n$  este **șir 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

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

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

**Exemplu:** a)  $X_n$  = v.a. care indică numărul apărut la a  $n$ -aruncare a unui zar  $\Rightarrow (X_n)_n$  șir de v.a. independente

b) Se aruncă o monedă

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

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

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

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

**Notăție:**  $X_n \xrightarrow{\text{a.s.}} X$

► Cu alte cuvinte, convergența a.s.  $X_n \xrightarrow{\text{a.s.}} X$  impune ca  $(X_n(\omega))_n$  să convergă la  $X(\omega)$  pentru fiecare  $\omega \in \Omega$ , cu excepția unei mulțimi “mici” de probabilitate nulă, adică mulțimea (evenimentul)

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

**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:**

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

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

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

Răspuns:

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

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

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

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

$$X_n \xrightarrow{a.s.} X.$$

$$2) X_n(\omega) = (-1)^n \omega, \omega \in [0, 1], n \geq 1 \Rightarrow X_n \xrightarrow{a.s.} ???$$

Răspuns:  $(X_n)_n$  nu converge a.s. spre o v.a.; șirul  $((X_n(\omega))_n$  este convergent doar în  $\omega = 0$ .

### Frecvențe relative și absolute

Fie  $A$  un eveniment asociat unei experiențe, repetăm experiența de  $n$  ori (în aceleași condiții date) și notăm cu  $k_n$  numărul de realizări ale evenimentului  $A$ ; **frecvența relativă** a evenimentului  $A$  este numărul

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

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

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

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

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

### 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 obținut se apropie tot mai mult de valoarea așteptată, 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) și legea tare a numerelor mari (LTNM).

△ **Scurt istoric:** Jacob Bernoulli (1655 -1705) a formulat LSNM pentru frecvența relativă a unui experiment și a dat răspunsul la întrebarea “*Putem aproxima empiric probabilitățile?*” (în opera publicată postum, în 1713, *Ars conjectandi*):

▷ Teorema lui Bernoulli afirmă: “*Frecvențele 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 repetiție a experimentului

•  $X_i = 1 \Leftrightarrow$  dacă  $A$  apare în a  $i$ -a repetiție a experimentului

$\Rightarrow X_i \sim \text{Bernoulli}(p)$  cu  $p := P(A)$

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

•  $X_1, \dots, X_n$  sunt v.a. independente

• frecvența relativă de apariție a lui  $A$  este

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

•  $(X_n)_n$  verifică LSNM, adică

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

△



**Fig. 5.** Jacob Bernoulli (timbru emis în 1994 cu ocazia Congresului Internațional al Matematicienilor din Elveția)

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

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

adică

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

**P. 17.** Fie  $(X_n)_n$  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

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

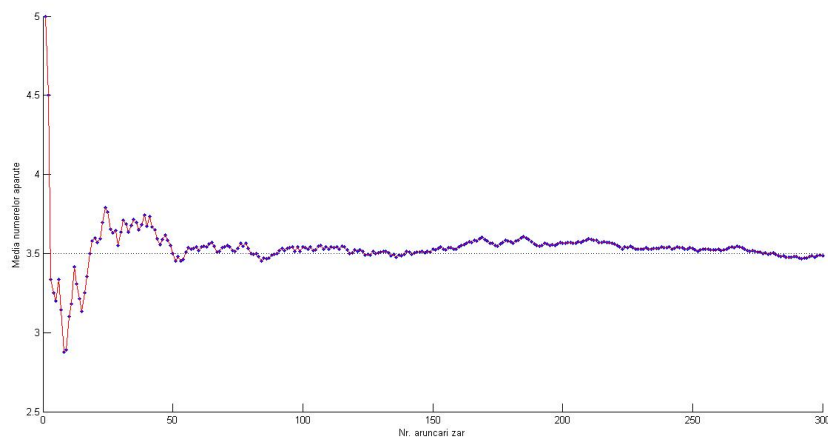
**P. 18.** Fie  $(X_n)_n$  şir de v.a. independente având aceeaşi distribuţie (notăm  $m = E(X_n) \forall n \in \mathbb{N}$ )  $\Rightarrow (X_n)_n$  verifică **LTNM**, adică

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

**În simulări:**  $\frac{1}{n}(X_1 + \dots + X_n) \approx E(X)$ , 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 experiențe, repetăm experiența de  $n$  ori (în aceleași condiții date). *LTNM*: cu cât repetăm mai des un experiment ( $n \rightarrow \infty$ ), cu atât mai bine aproximează frecvența relativă  $f_n(A)$  a evenimentului  $A$  probabilitatea sa (teoretică) de apariție  $P(A)$ :

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

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

### Metode Monte Carlo pentru integrare numerică

Fie  $g : [a, b] \rightarrow [0, \infty)$  o funcție integrabilă dată și  $M > 0$  astfel încât  $g(x) \leq M$ , oricare ar fi  $x \in [a, b]$ .

Considerăm următorii pași pentru aproximarea integralei  $\int_a^b g(x) dx$  folosind valori aleatoare.

#### Metoda 1:

- fie  $(U_n)_n$  șir de v.a. independente cu  $U_n \sim Unif[a, b]$  și fie  $X_n = g(U_n)$ .
- $(X_n)_n$  satisface LTNM (a se vedea P.17), adică

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

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

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

- în simulări:

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

#### Metoda 2:

- fie  $X_k, Y_k$  v.a. independente  $X_k \sim Unif[a, b], Y_k \sim Unif[0, M] \quad \forall k = \overline{1, n}$ .
- fie  $N_n = \#\{k \in \{1, \dots, n\} : Y_k \leq g(X_k)\}$
- $\int_a^b g(t) dt = \text{aria sub graficul funcției } g$

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

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

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

- în simulări:

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

# Statistică matematică

► Statistica matematică este o ramură a matematicii aplicate, care se ocupă de colectarea, gruparea, analiza și interpretarea datelor referitoare la anumite fenomene în scopul obținerii unor previziuni;

- statistica descriptivă: metode de colectare, organizare, sintetizare, prezentare și descriere a datelor numerice (sau nenumere) într-o formă convenabilă

- statistica inferențială: metode de interpretare a rezultatelor obținute prin metodele statisticii descriptive, utilizate apoi pentru luarea deciziilor.

► O *colectivitate* sau *populație statistică*  $\mathcal{C}$  este o mulțime de elemente care au anumite însușiri comune ce fac obiectul analizei statistice. Numărul elementelor populației se numește *volumul populației*.

Exemple de populații statistice: mulțimea persoanelor dintr-o anumită țară, localitate, zonă etc. într-un anumit an, mulțimea gospodăriilor din România la un moment dat, mulțimea consumatorilor unui produs, mulțimea societăților care produc un anumit produs, angajații unei societăți, studenții unei facultăți.

► *Eșantionul*  $\mathcal{E}$  reprezintă o submulțime a unei populații statistice  $\mathcal{E} \subset \mathcal{C}$ , constituită după criterii bine stabilite:

a) să fie aleatoare;

b) toate elementele colectivității să aibe aceeași șansă de a fi alese în eșantion;

c) eșantionul să fie reprezentativ (structura eșantionului să fie apropiată de structura populației);

d) volumul eșantionului să fie suficient de mare.

► *Unitatea statistică* (indivizii) este elementul, entitatea de sine stătătoare a unei populații statistice, care posedă o serie de trăsături caracteristice ce-i conferă apartenența la populația studiată.

De exemplu: *unitatea statistică simplă*: un salariat, un student, un agent economic, o trăsătură, o părere; *unitatea statistică complexă*: o grupă de studenți sau o echipă de salariați, o familie sau o gospodărie, o categorie de mărfuri.

► *Variabila statistică* sau *caracteristica* reprezintă o însușire, o proprietate măsurabilă a unei unități statistice, întâlnită la toate unitățile care aparțin aceleiași colectivități și care prezintă variabilitate de la o unitate statistică la alta. Caracteristica sau variabila statistică corespunde unei variabile aleatoare.

Exemple de caracteristici: vârsta, salariul, preferințele politice, prețul unui produs, calitatea unor servicii, nivelul de studii.

a) variabile (caracteristici) continue, iau un număr infinit și nenumărabil de valori într-un interval sau reuniune de intervale (de ex.: greutatea, înălțimea, valoarea glicemiei, temperatura aerului)

b) variabile (caracteristici) discrete, iau număr finit sau infinit dar numărabil de valori discrete (de ex.: numări elevi ai unei școli, numărul liceelor existente într-un oraș, valoarea IQ)

▷ acestea (a) și 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 observațiile 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* (populație)  $\mathcal{C}$  este cercetată din punctul de vedere al caracteristicii (variabilei statistice)  $X$ .

Distribuția caracteristicii  $X$  poate fi

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

2) specificată, dar depinzând de unul sau mai mulți parametri necunoscuți

(de ex.:  $X \sim \text{Exp}(\lambda)$ ,  $X \sim \text{Bin}(10, p)$ ,  $X \sim N(m, \sigma^2)$ )

3) necunoscută:  $X \sim ?$

- în cazurile 2) și 3) parametrii necunoscuți sau distribuția necunoscută

↪ se estimează → teoria estimăției

↪ se testează → teste statistice

► Fie  $\mathcal{E} \subset \mathcal{C}$  un eșantion. Se numesc **date de selecție** relative la caracteristica  $X$  datele statistice  $x_1, \dots, x_n$  obținute prin cercetarea indivizilor care fac parte din eșantionul  $\mathcal{E}$ .

► Datele de selecție  $x_1, \dots, x_n$  pot fi considerate ca fiind valorile unor variabile aleatoare  $X_1, \dots, X_n$ , numite **variabile de selecție** și care se consideră a fi variabile aleatoare independente și având aceeași distribuție ca  $X$ .

► Fie  $x_1, \dots, x_n$  datele statistice pentru caracteristica cercetată  $X$ , notăm cu  $X_1, \dots, X_n$  variabilele de selecție corespunzătoare. Fie  $g: \mathbb{R}^n \rightarrow \mathbb{R}$  o funcție astfel încât  $g(X_1, \dots, X_n)$  este o variabilă aleatoare.

$g(X_1, \dots, X_n)$  se numește **funcție de selecție** sau **estimator**

$g(X_1, \dots, X_n)$  se numește valoarea funcției de selecție sau **valoarea estimatorului**.

►  $g(X_1, \dots, X_n)$  este **estimator nedeplasat** pentru parametrul necunoscut  $\theta$ , dacă

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

►  $g(X_1, \dots, X_n)$  este **estimator consistent** pentru parametrul necunoscut  $\theta$ , dacă

$$g(X_1, \dots, X_n) \xrightarrow{a.s.} \theta.$$

► Fie  $g_1(X_1, \dots, X_n)$  și  $g_2(X_1, \dots, X_n)$  estimatori nedeplasați pentru parametrul necunoscut  $\theta$ .  $g_1(X_1, \dots, X_n)$  este **mai eficient** decât  $g_2(X_1, \dots, X_n)$ , dacă  $V(g_1) < V(g_2)$ .

• **Exemple de estimatori (funcții de selecție)** sunt: media de selecție, dispersia de selecție, funcția de repartiție empirică (de selecție), mediana de selecție.

▷ Estimatorii (funcțiile de selecție) se folosesc în statistică pentru estimarea punctuală a unor parametri necunoscuți, pentru obținerea unor intervale de încredere pentru parametri necunoscuți, pentru verificarea unor ipoteze statistice.

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

► **media de selecție**

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

► valoarea mediei de selecție

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

► **varianța (dispersia) de selecție**

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

► valoarea varianței (dispersiei) de selecție

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

► abaterea standard de selecție

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

► valoarea abaterii standard de selecție

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

► funcția de repartiție empirică  $\hat{F}_n : \mathbb{R} \times \Omega \rightarrow \mathbb{R}$

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

► valoarea funcției de repartiție empirice

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

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

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

► fie  $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$  datele de selecție ordonate crescător; valoarea **mediane de selecție** 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 corelație Pearson** a două caracteristici  $X$  și  $Y$  (datele de selecție corespunzătoare sunt  $y_1, \dots, y_n$ )

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

### Observații:

- 1) Media de selecție  $\bar{X}_n$  este estimator nedeplasat și consistent pentru media teoretică  $E(X)$  a caracteristicii  $X$  (în practică  $E(X) \approx \bar{x}_n$ ).
- 2) Varianța (dispersia) de selecție  $\tilde{S}_n^2$  este estimator nedeplasat și consistent pentru varianța teoretică  $V(X)$  a caracteristicii  $X$  (în practică  $V(X) \approx \tilde{s}_n^2$ ).
- 3) Funcția de repartiție de selecție calculată în  $x \in \mathbb{R}$ :  $\hat{F}_n(x)$  este estimator nedeplasat și consistent pentru  $F_X(x)$  valoarea funcției de repartiție teoretice în  $x$  (în practică  $F_X(x) \approx \hat{F}_n(x)$ ).
- 4) Mediana de selecție este un estimator nedeplasat pentru mediana teoretică  $z_{0.5}$ , dacă distribuția caracteristicii  $X$  este simetrică.
- 5)  $r_P$  măsoară gradul de dependență liniară între variabilele  $X$  și  $Y$  (are loc  $-1 \leq r_P \leq 1$ ),  $r_P$  estimează coeficientul de corelație  $\rho(X, Y)$

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

## Metoda momentelor pentru estimarea parametrilor necunoscuți $\theta = (\theta_1, \dots, \theta_r)$ pentru distribuția caracteristicii cercetate $X$

de exemplu:

$X \sim Exp(\lambda)$  parametrul necunoscut:  $\theta = \lambda$

$X \sim N(m, \sigma^2)$  parametri necunoscuți:  $(\theta_1, \theta_2) = (m, \sigma)$

$X \sim Unif[a, b]$  parametri necunoscuți:  $(\theta_1, \theta_2) = (a, b)$

Fie  $x_1, \dots, x_n$  datele statistice pentru caracteristica cercetată  $X$  și fie  $X_1, \dots, X_n$  variabilele de selecție corespunzătoare.

Se rezolvă sistemul

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

cu necunoscutele  $\theta_1, \dots, \theta_r$ .

Soluția sistemului  $\hat{\theta}_1, \dots, \hat{\theta}_r$  este estimatorul pentru parametrii necunoscuți ai distribuției 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

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

Estimatorul pentru parametrul necunoscut  $a$  este

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

unde  $X_1, \dots, X_n$  sunt variabilele de selecție. Valoarea estimatorului este

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

Parametrul necunoscut  $a$  este estimat cu valoarea 0.936.

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

## Metoda verosimilității maxime pentru estimarea parametrului necunoscut $\theta$ al distribuției caracteristicii cercetate $X$

Fie  $x_1, \dots, x_n$  datele statistice pentru caracteristica cercetată  $X$  și fie  $X_1, \dots, X_n$  variabilele de selecție corespunzătoare. Notăm

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

Aceasta este funcția de verosimilitate pentru parametrul  $\theta$  și datele statistice  $x_1, \dots, x_n$ .

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

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

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

**Observație:** Dacă distribuția caracteristicii cercetate depinde de  $k$  parametri necunoscuți  $(\theta_1, \dots, \theta_k)$  atunci se rezolvă sistemul

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

Se poate lucra și cu varianta transformată:

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

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 verosimilității maxime să se estimeze parametrul  $\theta := p \in (0, 1)$  al distribuției Bernoulli,

$$X \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}, \text{ cu datele statistice: } 0, 1, 1, 0, 0, 0, 1, 0, 1, 0.$$

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

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

$$\Rightarrow \ln L(x_1, \dots, x_n; p) = (x_1 + \dots + x_n) \ln(p) + (n - (x_1 + \dots + x_n)) \ln(1-p)$$

$$\frac{\partial \ln L}{\partial p} = 0 \Rightarrow p = \frac{1}{n} (x_1 + \dots + x_n).$$

$$\text{Are loc: } \frac{\partial^2 \ln L}{\partial p^2} < 0.$$

**Estimatorul de verosimilitate maximă** pentru parametrul necunoscut  $p$  este

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

unde  $X_1, \dots, X_n$  sunt variabilele de selecție. **Valoarea estimată** este

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

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

# Intervale de încredere și teste statistice

## Noțiuni de bază

► Fie  $\alpha \in (0, 1)$  nivelul de semnificație (probabilitatea de risc). **Cuantila de ordin  $\alpha$**  pentru distribuția 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 numește **mediana**
- dacă  $X$  este v.a. continuă, atunci:  $z_\alpha$  este cuantila 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  $z_\alpha$

## Distribuții de probabilitate continue frecvent folosite în statistică

▷ distribuția normală  $N(m, \sigma^2)$

cuantila  $z_\alpha = \text{norminv}(\alpha)$ ; funcția de repartiție  $F_{N(0,1)}(x) = \text{normcdf}(x)$

▷ distribuția Student  $St(n)$

cuantila  $t_\alpha = \text{tinv}(\alpha, n)$ ; funcția de repartiție  $F_{St(n)}(x) = \text{tcdf}(x, n)$

▷ distribuția Chi-pătrat  $\chi^2(n)$

cuantila  $c_\alpha = \text{chi2inv}(\alpha, n)$ ; funcția de repartiție  $F_{\chi^2(n)}(x) = \text{chi2cdf}(x, n)$

De exemplu:

$$\text{norminv}(0.01) = -2.3263, \text{norminv}(1 - 0.01) = 2.3263,$$

$$\text{tinv}(0.05, 10) = -1.8125, \text{tinv}(1 - 0.05, 10) = 1.8125,$$

$$\text{chi2inv}(0.05, 10) = 3.9403, \text{chi2inv}(1 - 0.05, 10) = 18.307.$$

## Intervale de încredere

Fie  $x_1, \dots, x_n$  datele statistice pentru caracteristica cercetată  $X$ , a cărei distribuție depinde de parametrul necunoscut  $\theta$ ; notăm cu  $X_1, \dots, X_n$  variabilele de selecție corespunzătoare. Fie  $\alpha \in (0, 1)$  *nivelul de semnificație*;  $1 - \alpha$  se numește *nivelul de încredere*.

Se caută doi estimatori  $g_1(X_1, \dots, X_n)$  și  $g_2(X_1, \dots, X_n)$  astfel încât

$$P(g_1(X_1, \dots, X_n) < \theta < g_2(X_1, \dots, X_n)) = 1 - \alpha \Leftrightarrow P\left(\theta \notin \left(g_1(X_1, \dots, X_n), g_2(X_1, \dots, X_n)\right)\right) = \alpha$$

►  $\left(g_1(X_1, \dots, X_n), g_2(X_1, \dots, X_n)\right)$  se numește **interval de încredere bilateral pentru parametrul necunoscut  $\theta$**

►  $\left(g_1(x_1, \dots, x_n), g_2(x_1, \dots, x_n)\right)$  este valoarea intervalului de încredere pentru parametrul necunoscut  $\theta$

►  $g_1(X_1, \dots, X_n)$  este limita inferioară a intervalului de încredere, valoarea sa este  $g_1(x_1, \dots, x_n)$

►  $g_2(X_1, \dots, X_n)$  este limita superioară a intervalului de încredere, valoarea sa este  $g_2(x_1, \dots, x_n)$

► probabilitatea ca parametrul necunoscut  $\theta$  să fie în intervalul  $\left(g_1(X_1, \dots, X_n), g_2(X_1, \dots, X_n)\right)$  este  $1 - \alpha$  (nivelul de încredere)

► există și **intervale de încredere unilaterale**:  $\left(-\infty, g_3(X_1, \dots, X_n)\right)$ ,  $\left(g_4(X_1, \dots, X_n), \infty\right)$ , estimatorii  $g_3$  și  $g_4$  sunt astfel încât

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

- $(-\infty, g_3(x_1, \dots, x_n))$   $(g_4(x_1, \dots, x_n), \infty)$  sunt valorile intervalelor de încredere unilaterale pentru parametrul necunoscut  $\theta$
- probabilitatea ca parametrul necunoscut  $\theta$  să fie în intervalul  $(-\infty, g_3(X_1, \dots, X_n))$  este  $1 - \alpha$ , respectiv probabilitatea ca  $\theta$  să fie în intervalul  $(g_4(X_1, \dots, X_n), \infty)$  este  $1 - \alpha$ .

### Teoreme de bază folosite în Statistică:

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

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

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

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

- (1) pentru media de selecție are loc  $\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$ ;
- (2) pentru media de selecție și abaterea standard de selecție are loc  $\frac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} \sim St(n-1)$ ;
- (3) pentru varianța de selecție are loc  $\frac{n-1}{\sigma^2} \tilde{S}_n^2 \sim \chi^2(n-1)$ .

### În aplicații:

- dacă  $X \sim N(m, \sigma^2)$  sau  $n > 30$  și  $X$  are o distribuție necunoscută, atunci  $\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$
- dacă  $X$  are distribuție, care nu este distribuția normală, se estimează pentru  $n > 30$  (suficient de mare)

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

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

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

▷ valoarea intervalului de încredere bilateral pentru media teoretică  $m = E(X)$ , când varianța 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  $n > 30$  și  $X$  are o distribuție necunoscută, atunci  $\frac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} \sim St(n-1)$
- dacă  $X$  are distribuție, care nu este distribuția normală, se estimează pentru  $n > 30$  (suficient de mare)

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



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

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

▷ valoarea intervalului de încredere bilateral pentru media teoretică  $m = E(X)$ , când varianța 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  $X \sim N(m, \sigma^2)$

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

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

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

▷ valoarea intervalului de încredere bilateral pentru varianța teoretică  $\sigma^2$

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

**Exemplu:** Media de selecție a lungimii a 100 de șuruburi este 15.5 cm, iar varianța de selecție este 0.09 cm<sup>2</sup>. Să se construiască un interval de încredere 99% bilateral pentru media (teoretică) a lungimii șuruburilor.

Soluție: valoarea intervalului de încredere bilateral pentru media teoretică  $m$ , când varianța 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  $\bar{x}_n = 15.5, \tilde{s}_n = 0.03 (\tilde{s}_n^2 = 0.09), t_{1-\frac{\alpha}{2}} = \text{tinv}(1 - \frac{\alpha}{2}, 99) = 2.6264, \sqrt{n} = 10$

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

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

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

► cuantilele legii normale  $N(0, 1)$ :

$$z_{1-\frac{\alpha}{2}} = \text{norminv}(1 - \frac{\alpha}{2}), z_{1-\alpha} = \text{norminv}(1 - \alpha), z_{\alpha} = \text{norminv}(\alpha)$$

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

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

- ▷ din  $P\left(\left|\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}\right| < z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$  avem :  $\left|\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}}\right| < z_{1-\frac{\alpha}{2}} \Leftrightarrow \bar{X}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} < m < \bar{X}_n + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$
- *intervalul de încredere bilateral* pentru  $m = E(X)$  (media teoretică) este  $\left(\bar{X}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\right)$
- ▷ se calculează valoarea intervalului de încredere  $\left(\bar{x}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\right)$

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

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

- ▷ din  $P\left(\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} > z_\alpha\right) = 1 - \alpha$  avem :  $\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} > z_\alpha \Leftrightarrow m < \bar{X}_n - \frac{\sigma}{\sqrt{n}} \cdot z_\alpha$
- $g_3(X_1, \dots, X_n) = \bar{X}_n - \frac{\sigma}{\sqrt{n}} \cdot z_\alpha$ ; valoarea intervalului de încredere este  $\left(-\infty, \bar{x}_n - \frac{\sigma}{\sqrt{n}} \cdot z_\alpha\right)$
- ▷ din  $P\left(\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} < z_{1-\alpha}\right) = 1 - \alpha$  avem :  $\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} < z_{1-\alpha} \Leftrightarrow m > \bar{X}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}$
- $g_4(X_1, \dots, X_n) = \bar{X}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}$ ; valoarea intervalului de încredere este  $\left(\bar{x}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}, \infty\right)$

### 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  $x_1, \dots, x_n$
- ▶ dacă  $X \sim N(m, \sigma^2)$  sau  $n > 30$  și  $X$  are o distribuție necunoscută, atunci  $\frac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}} \sim St(n-1)$
- ▶ cuantilele legii Student  $St(n-1)$ :  
 $t_{1-\frac{\alpha}{2}} = \text{tinu}(1 - \frac{\alpha}{2}, 0, 1)$ ,  $t_{1-\alpha} = \text{tinu}(1 - \alpha, 0, 1)$ ,  $t_\alpha = \text{tinu}(\alpha, 0, 1)$
- *interval de încredere bilateral*: se caută doi estimatori  $g_1(X_1, \dots, X_n)$  și  $g_2(X_1, \dots, X_n)$  astfel încât pentru media teoretică  $m = E(X)$  să avem:

$$P\left(g_1(X_1, \dots, X_n) < m < g_2(X_1, \dots, X_n)\right) = 1 - \alpha$$

- ▷ din  $P\left(\left|\frac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}}\right| < t_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$  avem :  $\left|\frac{\bar{X}_n - m}{\frac{\tilde{S}_n}{\sqrt{n}}}\right| < t_{1-\frac{\alpha}{2}} \Leftrightarrow \bar{X}_n - \frac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}} < m < \bar{X}_n + \frac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}$
- *intervalul de încredere bilateral* pentru  $m = E(X)$  (media teoretică) este  $\left(\bar{X}_n - \frac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}, \bar{X}_n + \frac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\right)$
- ▷ se calculează valoarea intervalului de încredere  $\left(\bar{x}_n - \frac{\tilde{s}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}, \bar{x}_n + \frac{\tilde{s}_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\right)$
- *interval de încredere unilateral*: se caută doi estimatori  $g_3(X_1, \dots, X_n)$ ,  $g_4(X_1, \dots, X_n)$  astfel încât pentru media teoretică  $m = E(X)$  să avem:

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

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

$$\bullet g_3(X_1, \dots, X_n) = \bar{X}_n - \frac{\tilde{S}_n}{\sqrt{n}} \cdot t_\alpha; \text{ valoarea intervalului de \u00e2ncredere este } \left(-\infty, \bar{x}_n - \frac{\tilde{s}_n}{\sqrt{n}} \cdot t_\alpha\right)$$

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

$$\bullet g_4(X_1, \dots, X_n) = \bar{X}_n - \frac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\alpha}; \text{ valoarea intervalului de \u00e2ncredere este } \left(\bar{X}_n - \frac{\tilde{S}_n}{\sqrt{n}} \cdot t_{1-\alpha}, \infty\right)$$

## Interval de \u00e2ncredere pentru varian\u021ba (dispersia) $\sigma^2 = V(X)$ caracteristicii cercetate $X$

\u25b6 se dau  $\alpha \in (0, 1)$ , datele statistice  $x_1, \dots, x_n$

\u25b6 dac\u0103  $X \sim N(m, \sigma^2)$ , atunci  $\frac{n-1}{\sigma^2} \tilde{S}_n^2 \sim \chi^2(n-1)$

\u25b6 cuantilele  $\chi^2$  (Chi-p\u0103trat) cu  $n-1$  grade de libertate:

$$c_{1-\frac{\alpha}{2}} = \text{chi2inv}(1 - \frac{\alpha}{2}, n-1), c_{\frac{\alpha}{2}} = \text{chi2inv}(\frac{\alpha}{2}, n-1), c_{1-\alpha} = \text{chi2inv}(1 - \alpha, n-1), c_\alpha = \text{chi2inv}(\alpha, n-1)$$

\bullet *interval de \u00e2ncredere bilateral*: se caut\u0103 doi estimatori  $g_1(X_1, \dots, X_n)$  \u015fi  $g_2(X_1, \dots, X_n)$  astfel \u00e2nc\u00e2t pentru varian\u021ba teoretic\u0103  $m = E(X)$  s\u0103 avem:

$$P\left(g_1(X_1, \dots, X_n) < \sigma^2 < g_2(X_1, \dots, X_n)\right) = 1 - \alpha$$

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

$$\bullet \text{intervalul de \u00e2ncredere bilateral pentru } \sigma^2 = V(X) \text{ (varian\u021ba teoretic\u0103) este } \left(\frac{n-1}{c_{1-\frac{\alpha}{2}}} \cdot \tilde{S}_n^2 < \sigma^2 < \frac{n-1}{c_{\frac{\alpha}{2}}} \cdot \tilde{S}_n^2\right)$$

$$\triangleright \text{se calculeaz\u0103 valoarea intervalului de \u00e2ncredere } \left(\frac{n-1}{c_{1-\frac{\alpha}{2}}} \cdot \tilde{s}_n^2 < \sigma^2 < \frac{n-1}{c_{\frac{\alpha}{2}}} \cdot \tilde{s}_n^2\right)$$

\bullet *interval de \u00e2ncredere unilateral*: se caut\u0103 doi estimatori  $g_3(X_1, \dots, X_n)$ ,  $g_4(X_1, \dots, X_n)$  astfel \u00e2nc\u00e2t pentru varian\u021ba teoretic\u0103  $\sigma^2 = V(X)$  s\u0103 avem:

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

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

$$\bullet g_3(X_1, \dots, X_n) = \frac{n-1}{c_\alpha} \cdot \tilde{S}_n^2; \text{ valoarea intervalului de \u00e2ncredere este } \left(-\infty, \frac{n-1}{c_\alpha} \cdot \tilde{s}_n^2\right)$$

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

$$\bullet g_4(X_1, \dots, X_n) = \frac{n-1}{c_{1-\alpha}} \cdot \tilde{S}_n^2; \text{ valoarea intervalului de \u00e2ncredere este } \left(\frac{n-1}{c_{1-\alpha}} \cdot \tilde{s}_n^2, \infty\right)$$

## Teste statistice

Fie  $x_1, \dots, x_n$  datele statistice pentru caracteristica cercetat\u0103  $X$ , not\u0103m cu  $X_1, \dots, X_n$  variabilele de selec\u021bie corespunz\u0103toare.

\u25b6 Ipoteza statistic\u0103 este o presupunere relativ\u0103 la un parametru necunoscut  $\theta$

\u25b6 Metoda de stabilire a veridicit\u0103\u021bii unei ipoteze statistice se nume\u015fte test (criteriu de verificare).

\u25b6 Rezultatul test\u0103rii se folose\u015fte apoi pentru luarea unor decizii (cum ar fi: eficien\u021ba unor medicamente, strategii

de marketing, alegerea unui produs etc.).

► Se formulează ipoteza nulă  $H_0$  și ipoteza alternativă  $H_1$ , privind parametrul  $\theta$ ; fie  $\theta_0$  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 semnificație (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

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

ceea ce este echivalent cu

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

Concluzia testului:

$(x_1, \dots, x_n) \notin U \Rightarrow$  ipoteza  $H_0$  este admisă

$(x_1, \dots, x_n) \in U \Rightarrow$  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 varianța teoretică  $V(X)$  este cunoscută: testul lui Gauss (testul Z)
  - ▷ când varianța teoretică  $V(X)$  este necunoscută: Student Test (testul T)
- test pentru abaterea standard teoretică  $\sqrt{V(X)}$  sau pentru varianța teoretică  $V(X)$ : testul  $\chi^2$
- test asupra proporției (test Gauss aproximativ)
- test pentru independența a două caracteristici

Pașii în efectuarea unui test statistic:

- Care parametru se testează? Care test este potrivit?
- Care este ipoteza nulă  $H_0$  și care este ipoteza alternativă  $H_1$ ?
- Care este nivelul de semnificație (probabilitatea de risc)  $\alpha$ ?
- Calculul valorii estimatorului pe baza datelor statistice
- Concluzia testului

**Test pentru media  $m = E(X)$  caracteristicii cercetate  $X$ , când varianța  $\sigma^2 = V(X)$  este cunoscută**

► se dau  $\alpha \in (0, 1)$ ,  $m_0$ ,  $\sigma$

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

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

► cuantilele legii normale  $N(0, 1)$ :

$$z_{1-\frac{\alpha}{2}} = \text{norminv}(1 - \frac{\alpha}{2}), z_{1-\alpha} = \text{norminv}(1 - \alpha), z_\alpha = \text{norminv}(\alpha)$$

	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	$ z  < z_{1-\frac{\alpha}{2}}$	$z < z_{1-\alpha}$	$z > z_\alpha$
Se respinge $H_0$ în favoarea lui $H_1$ , dacă	$ z  \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 varianța este cunoscută:

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

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

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