Curs - Probabilităţi şi Statistică 2021/2022

Facultatea de Matematică şi Informatică
Universitatea Babeş-Bolyai, Cluj-Napoca

Teoria Probabilităţilor

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

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

$\hookrightarrow$ se măsoară *şansele pentru succes* sau *riscul pentru insucces* al unor evenimente

[Sursa: www.financialmarket.ro]

Fenomene şi procese aleatoare apar, de exemplu, în:

$\to$ jocuri de noroc, pariuri, loto (6 din 49)
$\to$ previziuni meteo
$\to$ previziuni economice / financiare
$\to$ sondaje de opinie, asigurări (evaluarea riscurilor, pierderilor)

$\to$ **în informatică**:
$\triangleright$ sisteme de comunicare, prelucrarea informaţiei, modelarea traficului în reţea;
$\triangleright$ analiza probabilistică a unor algoritmi, fiabilitatea sistemelor;
$\triangleright$ algoritmi de simulare, machine learning, data mining, recunoaşterea formelor sau a vocii;
$\triangleright$ generarea de numere aleatoare, algoritmi aleatori: de tip Monte-Carlo, de tip Las Vegas etc.

Octave online: `https://octave-online.net`

Exemplu: Generarea de valori aleatoare (în Octave/Matlab)

```
a=rand % valoare aleatoare între 0 şi 1
v1=rand(1,10)    %vector cu 10 valori aleatoare între 0 şi 1
a=4; b=10;
v2=a+(b-a)*rand(1,15) %vector cu 15 valori  aleatoare între 4 şi 10
z=randi(6,1,20)
%vector cu 20 de valori aleatoare din mulţimea {1,2,3,4,5,6}
```

Exerciţiu: Generaţi un vector cu 100 de valori aleatoare `0` şi `1`, în care `0` şi `1` au aceleaşi şanse de apariţie.

Răspuns: `floor(2*rand(1,100))` sau `randi(2,1,100)-1`

Algoritmi aleatori

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

▷ durata de execuţie, spaţiul de stocare, rezultatul obţinut sunt variabile aleatoare (chiar dacă se folosesc aceleaşi valori input)

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

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

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

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

• Un algoritm aleatoriu pentru care rezultatele obţinute sunt corecte *doar* cu o anumită probabilitate se numeşte algoritm **Monte Carlo**.
$\hookrightarrow$ se examinează probabilitatea cu care rezultatul este corect; probabilitatea de eroare poate fi scăzută semnificativ prin execuţii repetate, independente;
Exemplu:
▷ testul Miller-Rabin, care verifică dacă un număr natural este prim sau este număr compus; testul returnează fie răspunsul "numărul este sigur un număr compus" sau răspunsul "numărul este probabil un număr prim";
Exerciţiu: Fie S(1),...,S(300) un vector cu 300 de elemente, din mulţimea $\{0, 1, 2\}$ (ordinea lor este necunoscută). $\longrightarrow$ De care tip este următorul algoritm (scris în Octave)?

```
S=randi(3,1,300)-1;
k=0;
do
    k=k+1;
    i=randi(300);
until (S(i) == 0)
i  %  indicele, pentru care S(i)=0
k  % număr iteraţii până se găseşte aleator un 0
```

Răspuns: Algoritm de tip Las Vegas.

Versiunea Monte Carlo a problemei formulate anterior: se dă M numărul maxim de iteraţii.

```
M=3;
S=randi(3,1,300)-1;
k=0;
do
    k=k+1 ;
    i=randi(300);
until ( (S(i) == 0) || (k==M) )
i  % indicele, pentru care S(i)=0  sau pentru care k==M
k
% număr iteraţii până se găseşte
% aleator un 0 sau programul s-a oprit
S(i)
```

▷ dacă 0 este găsit, atunci algoritmul se încheie cu rezultatul corect, altfel algoritmul nu găseşte niciun 0.

Noţiuni introductive:

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

• Evenimentul este rezultatul unui experiment.

Exemple:

▷ Experiment: aruncarea a două zaruri, eveniment: ambele zaruri indică 1

▷ experiment: aruncarea unei monede, eveniment: moneda indică pajură

▷ experiment: extragerea unei cărţi de joc, eveniment: s-a extras as

▷ experiment: extragerea unui număr la loto, eveniment: s-a extras numărul 27

• evenimentul imposibil, notat cu $\emptyset$, este evenimentul care nu se realizează niciodată la efectuarea experienţei aleatoare

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

• spaţiul de selecţie, notat cu Ω, este mulţimea tuturor rezultatelor posibile ale experimentului considerat

    ⋄ spaţiul de selecţie poate fi finit sau infinit

• dacă A este o submulţime a lui Ω atunci A se numeşte eveniment aleator, iar dacă A are un singur element atunci A este un eveniment elementar.

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

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

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

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

Operaţii cu evenimente

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

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

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

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

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

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

Relaţii între evenimente

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

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

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

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

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

asociative

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

şi **distributive**

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

satisfac legile lui De Morgan

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

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

Frecvenţa relativă şi frecvenţa absolută

Def. 2. *Fie A un eveniment asociat unei experienţe, repetăm experienţa de n ori (în aceleaşi condiţii date) şi notăm cu $r_n(A)$ numărul de realizări ale evenimentului A; frecvenţa relativă a evenimentului A este numărul*

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

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

Definiţia clasică a probabilităţii

Def. 3. *Într-un experiment în care cazurile posibile sunt finite la număr şi au aceleaşi şanse de a se realiza,* ***probabilitatea*** *unui eveniment A este numărul*

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

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

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

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

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

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

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

$A = \{(c, p, p, p), (p, c, p, p), (p, p, c, p), (p, p, p, c)\} \Rightarrow P(A) = \frac{4}{2^4} = 0.25$ ♠

Exemplu - Joc de zaruri (sec. XVII): Un pasionat jucător de zaruri, cavalerul de Méré, susţinea în discuţiile sale cu B. Pascal că a arunca un zar de 4 ori pentru a obţine cel puţin o dată faţa şase, este acelaşi lucru cu a arunca de 24 ori câte două zaruri pentru a obţine cel puţin o dublă de şase. Cu toate acestea, cavalerul de Méré a observat că jucând în modul al doilea (cu două zaruri aruncate de 24 ori), pierdea faţă de adversarul său, dacă acesta alegea primul mod (aruncarea unui singur zar de 4 ori). Pascal şi Fermat au arătat că probabilitatea de câştig la jocul cu un singur zar aruncat de 4 ori este $p_1 \approx 0.5177$, iar probabilitatea $p_2 \approx 0.4914$ la jocul cu două zaruri aruncate de 24 de ori. Deşi diferenţa dintre cele două probabilităţi este mică, totuşi, la un număr mare de partide, jucătorul cu probabilitatea de câştig p_1 câştigă în faţa jucătorului cu probabilitatea de câştig p_2. Practica jocului confirmă astfel justeţea raţionamentului matematic, contrar credinţei lui de Méré.

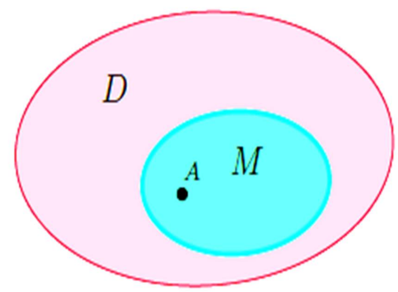
♠

Definiţia axiomatică a probabilităţii

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

Probabilitatea geometrică: Măsura unei mulţimi corespunde lungimii în $\mathbb{R}$, ariei în $\mathbb{R}^2$, volumului în $\mathbb{R}^3$. Fie $M \subset D \subset \mathbb{R}^n$, $n \in \{1, 2, 3\}$, mulţimi cu măsură finită.
Alegem aleator un punct $A \in D$ (în acest caz spaţiul de selecţie este D). Probabilitatea geometrică a evenimentului "$A \in M$" este

$$P(A \in M) := \frac{\text{măsura}(M)}{\text{măsura}(D)}.$$



$M \subset D \subset \mathbb{R}^2$

O teorie formală a probabilităţii 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 probabilităţii în lucrarea sa *Conceptele de bază ale Calculului Probabilităţii*.

$\twoheadrightarrow$ $P : \mathcal{K} \to \mathbb{R}$ este o funcţie astfel încât oricărui eveniment aleator $A \in \mathcal{K}$ i se asociază valoarea $P(A)$, ***probabilitatea de apariţie a evenimentului*** A

$\hookrightarrow$ $\mathcal{K}$ este o mulţime de evenimente şi are structura unei σ-algebre (vezi Def. 4)
$\hookrightarrow$ P satisface anumite axiome (vezi Def. 5)

Def. 4. *O familie $\mathcal{K}$ de evenimente din spaţiul de selecţie Ω se numeşte **σ-algebră** dacă sunt satisfăcute condiţiile:*

(i) $\mathcal{K}$ este nevidă;

(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 numeşte **spaţiu măsurabil**.*

Exemple: 1) Dacă $\emptyset \neq A \subset \Omega$ atunci $\mathcal{K} = \{\emptyset, A, \bar{A}, \Omega\}$ este o σ-algebră.
2) $\mathcal{P}(\Omega)$:= mulţimea tuturor submulţimilor ale lui Ω este o σ-algebră.
3) Dacă $(\Omega, \mathcal{K})$ este un spaţiu măsurabil şi $\emptyset \neq B \subseteq \Omega$, atunci

$$B \cap \mathcal{K} = \{B \cap A : A \in \mathcal{K}\}$$

este o σ-algebră pe mulţimea B, iar $(B, B \cap \mathcal{K})$ este un spaţiu măsurabil.

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

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

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

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

Def. 5. *Fie $\mathcal{K}$ o σ-algebră în Ω. O funcţie $P : \mathcal{K} \to \mathbb{R}$ se numeşte* ***probabilitate*** *dacă satisface axiomele:*

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

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

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

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

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

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

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

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

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

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

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

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

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

b) Pentru $A_1, ..., A_n \in \mathcal{K}$ care e formula similară de calcul pentru $P(A_1 \cup A_2 \cup ... \cup A_n)$?

Exemplu: Dintr-un pachet de 52 de cărţi de joc se extrage o carte aleator. Care este probabilitatea p de a extrage a) un as sau o damă de pică? b) o carte cu inimă sau un as?
R.: a) A: s-a extras un as; D: s-a extras damă de pică; A şi D sunt două evenimente disjuncte (incompatibile)

$$p = P(A\cup D) = P(A) + P(D) = \frac{4+1}{52};$$

b) I: s-a extras o carte cu inimă; I şi A *nu* sunt evenimente incompatibile

$$p = P(I\cup A) = P(I) + P(A) - P(I\cap A) = \frac{13+4-1}{52} = \frac{4}{13}.$$

♡

Evenimente independente

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

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

Observaţie: Fie evenimentele $A, B \in \mathcal{K}$. Evenimentele A şi B sunt **independente**, dacă **apariţia evenimentului A, nu influenţează apariţia evenimentului B şi invers**. Două evenimente se numesc **dependente** dacă probabilitatea realizării unuia dintre ele depinde de faptul că celălalt eveniment s-a produs sau nu.
Exerciţiu: Se aruncă un zar de două ori.
A: primul număr este 6; B: al doilea număr este 5; C: primul număr este 1.
Sunt A şi B evenimente independente?
Sunt A şi C evenimente independente? Sunt A şi C evenimente disjuncte?
Sunt B şi C evenimente dependente? ♣

P. 3. *Fie $(\Omega, \mathcal{K}, P)$ un spaţiu de probabilitate şi fie $A, B \in \mathcal{K}$. Sunt echivalente afirmaţiile:*

(1) *A şi B sunt independente.*

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

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

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

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

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

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

Exemplu: Se dă algoritmul de tip Monte-Carlo

```
M=input('M=') % numar maxim de iteratii; M >= 1
M=3 % maximale Anzahl Iterationen; M >= 1
S=[0 0 0 0 0 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3];
% 0,1,2,3 apar fiecare cu probabilitatea  5/20=1/4
S=S(randperm(length(S))) % permutare aleatoare a lui S
k=0;
do
    k=k+1;
    i=randi(20);
    % se alege o valoare aleatoare S(i)
until ( (S(i) == 0) || (k==M) )
fprintf('k: %d \n',k)
fprintf('S(%d): %d \n',i,S(i))
```

Se calculează probabilităţile următoarelor evenimente (din punct de vedere teoretic):

$$P(\text{“primul 0 este găsit la a } M\text{-a iteraţie”}) = \left(\frac{3}{4}\right)^{M-1} \cdot \frac{1}{4},$$

$$P(\text{“0 nu este găsit în } M \text{ iteraţii”}) = \left(\frac{3}{4}\right)^{M},$$

probabilitatea evenimentului complementar este

$$P(\text{“cel puţin un 0 este găsit în } M \text{ iteraţii”}) = 1 - \left(\frac{3}{4}\right)^{M} \longrightarrow 1, \text{ când } M \to \infty.$$

♣

Exemplu: 1) $A, B, C \in \mathcal{K}$ sunt trei evenimente independente (în totalitate), dacă

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

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

2) Cele 4 feţe ale unui tetraedru regulat sunt vopsite astfel: una este roşie, una este albastră, una este verde şi una este colorată având cele trei culori. Se aruncă tetraedrul şi se consideră evenimentele: R: tetraedrul cade pe o parte ce conţine culoarea roşie; A: tetraedrul cade pe o parte ce conţine culoarea albastră; V: tetraedrul cade pe o partea ce conţine culoarea verde.
Sunt cele 3 evenimente independente în totalitate?

R.: Nu, cele 3 evenimente nu sunt independente în totalitate pentru că $P(R \cap A \cap V) = \frac{1}{4} \neq P(R)P(A)P(V) = \frac{1}{8}$.
3) Pentru a verifica dacă n evenimente distincte $B_1, \ldots, B_n$ sunt independente în totalitate câte relaţii trebuie verificate?
R.: $C_n^2 + C_n^3 + ... + C_n^n = 2^n - C_n^0 - C_n^1 = 2^n - 1 - n$. ♦

Probabilitate condiţionată

Def. 8. *Fie $(\Omega, \mathcal{K}, P)$ un spaţiu de probabilitate şi fie $A, B \in \mathcal{K}$.* ***Probabilitatea condiţionată a evenimentului*** *A* ***de evenimentul*** *B este $P(\cdot|B) : \mathcal{K} \to [0,1]$ definită prin*

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

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

Observaţie: Fie evenimentele $A, B \in \mathcal{K}$ astfel încât $P(A) > 0$ şi $P(B) > 0$. Evenimentele A şi B sunt **independente** (a se vedea Def. 6), dacă apariţia evenimentului A, nu influenţează apariţia evenimentului B şi invers, adică

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

Exemplu: Se extrag succesiv fără returnare două bile dintr-o urnă cu 4 bile albe şi 5 bile roşii.
a) Ştiind că prima bilă este roşie, care este probabilitatea (condiţionată) ca a doua bilă să fie albă?
b) Care este probabilitatea ca ambele bile să fie roşii?
R.: pentru $i \in \{1, 2\}$ fie evenimentele

R_i: la a i-a extragere s-a obţinut o bilă roşie;

$A_i = \bar{R}_i$: la a i-a extragere s-a obţinut o bilă albă;

a) $P(A_2|R_1) = \frac{4}{8}$. **b)** $P(R_1 \cap R_2) = P(R_2|R_1)P(R_1) = \frac{4}{8} \cdot \frac{5}{9}$. ♣

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

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

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

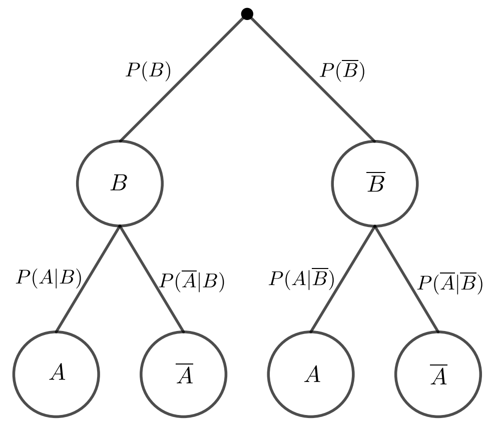


Fig.1. Probabilităţi condiţionate

Def. 9. *O familie $\{H_1, \dots H_n\} \subset \mathcal{K}$ de evenimente din Ω se numeşte* ***partiţie*** *sau* ***sistem complet de evenimente*** *a lui Ω, dacă $\bigcup\limits_{i=1}^{n} H_i = \Omega$ şi pentru fiecare $i, j \in \{1, \dots, n\}$, $i \neq j$, evenimentele H_i şi H_j sunt disjuncte, adică $H_i \cap H_j = \emptyset$.*

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

P. 5. ***(Formula probabilităţii totale)*** *Într-un spaţiu de probabilitate $(\Omega, \mathcal{K}, P)$ considerăm partiţia $\{H_1, ..., H_n\}$ a lui Ω cu $H_i \in \mathcal{K}$ şi $P(H_i) > 0 \; \forall \, i \in \{1, \dots, n\}$, şi fie $A \in \mathcal{K}$. Atunci are loc*

$$P(A) = P(A|H_1)P(H_1) + ... + P(A|H_n)P(H_n).$$

Exemplu: Într-o urnă sunt 7 bile albe, notate cu 1,2,3,4,5,6,7, şi 6 bile roşii notate cu 8,9,10,11,12,13. Se extrage o bilă. **a)** Ştiind că bila extrasă este roşie, care este probabilitatea (condiţionată) p_1, ca numărul înscris să fie divizibil cu 4? **b)** Ştiind că prima bilă este roşie, care este probabilitatea (condiţionată) p_2, ca o a doua bilă extrasă să fie un număr impar? (Prima bilă nu s-a returnat în urnă!)
R.: Se consideră evenimentele:
A_1: prima bilă extrasă are înscris un număr divizibil cu 4;
B_1: prima bilă extrasă este roşie;
C_1: prima bilă extrasă are înscris un număr impar;
C_2: a doua bilă extrasă are înscris un număr impar.
a) $p_1 = P(A_1|B_1) = \frac{2}{6}$.

b) $p_2 = P(C_2|B_1) = ?$ Folosim Def.8 şi P.4, scriem succesiv

$$p_2 = P(C_2|B_1) = \frac{P(C_2 \cap B_1)}{P(B_1)} = \frac{P(C_2 \cap B_1 \cap C_1) + P(C_2 \cap B_1 \cap \bar{C}_1)}{P(B_1)}$$
$$= \frac{P(C_2|B_1 \cap C_1)P(B_1 \cap C_1) + P(C_2|B_1 \cap \bar{C}_1)P(B_1 \cap \bar{C}_1)}{P(B_1)} = \frac{\frac{6}{12} \cdot \frac{3}{13} + \frac{7}{12} \cdot \frac{3}{13}}{\frac{6}{13}} = \frac{13}{24}.$$

♡

Exemplu: Ce probabilităţi calculează programul de mai jos? Ce tip de algoritm aleator este?
▶ `randi(imax,n,m)` generează o `n`×`m` matrice cu valori întregi aleatoare (pseudoaleatoare) între `1` şi `imax`.

```
clear all
ci=0;
cp=0;
c=0;
a=0;
b=0;
N=1000;
A=[1:20];
for i=1:N
  r= randi(length(A));
 v=A(r);
 ci=ci+mod(v,2);
 cp=cp+(mod(v,2)==0);
  c=c+ mod(v,2)*(mod(v,3)==0);
  a=a+ mod(v,2)*(6<=v && v<=10);
  b=b+ (mod(v,2)==0)*(v>=14);
end
p1=c/ci
p2=a/ci
p3=b/cp
```

R.: Se extrage aleator un număr din şirul `A=[1,2,...,20]`.
▶`p1` estimează probabilitatea condiţionată ca numărul ales aleator să fie divizibil cu 3, *ştiind* că s-a extras un număr impar;
▶ `p2` estimează probabilitatea condiţionată ca numărul ales aleator să provină din mulţimea $\{6, 7, 8, 9, 10\}$, *ştiind* că s-a extras un număr impar;

► `p3` estimează probabilitatea condiţionată ca numărul ales aleator să provină din mulţimea $\{14, 15, 16, 17, 18, 19, 20\}$, *ştiind* că s-a extras un număr par.
Algoritmul este de tip Monte-Carlo! ♣

Exerciţiu: Să se calculeze valorile teoretice pentru probabilităţile `p1`, `p2`, `p3` din exemplul anterior! ♠

P. 6. ***(Regula de înmulţire)*** *Fie $(\Omega, \mathcal{K}, P)$ un spaţiu de probabilitate şi fie $A_1, \ldots, A_n \in \mathcal{K}$ astfel încât $P(A_1 \cap \cdots \cap A_{n-1}) > 0$. Atunci,*

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

Exemplu: Într-o urnă sunt 2 bile verzi şi 3 bile albastre. Se extrag 2 bile succesiv, fără returnare. Care este probabilitatea ca
a) prima bilă să fie verde, iar cea de-a doua albastră?
b) cele 2 bile să aibă aceeaşi culoare?
c) a doua bilă să fie albastră?
d) prima bilă să fie verde, *ştiind* că a doua este albastră?
e) se mai extrage o a treia bilă; se cere probabilitatea ca prima bilă să fie verde, cea de-a doua albastră şi a treia tot albastră.
R.: Notăm pentru $i \in \{1, 2, 3\}$ evenimentele:
A_i: la a i-a extragere s-a obţinut bilă albastră; V_i: la a i-a extragere s-a obţinut bilă verde;
a) folosim P.4: $P(V_1 \cap A_2) = P(A_2|V_1)P(V_1) = \frac{3}{4} \cdot \frac{2}{5}$
b) $P((V_1 \cap V_2) \cup (A_1 \cap A_2)) = P(V_1 \cap V_2) + P(A_1 \cap A_2) = P(V_2|V_1)P(V_1) + P(A_2|A_1)P(A_1) = \frac{1}{4} \cdot \frac{2}{5} + \frac{2}{4} \cdot \frac{3}{5}$
c) folosim formula probabilităţii totale P.7:
$P(A_2) = P(A_2|V_1)P(V_1) + P(A_2|A_1)P(A_1) = \frac{3}{4} \cdot \frac{2}{5} + \frac{2}{4} \cdot \frac{3}{5}$
d) folosim P.4: $P(V_1|A_2) = \frac{P(V_1 \cap A_2)}{P(A_2)} = \frac{P(A_2|V_1)P(V_1)}{P(A_2)} = \frac{\frac{3}{4} \cdot \frac{2}{5}}{\frac{3}{4} \cdot \frac{2}{5} + \frac{2}{4} \cdot \frac{3}{5}}$
e) formula de înmulţire a probabilităţilor P.6:
$P(V_1 \cap A_2 \cap A_3) = P(V_1) \cdot P(A_2|V_1) \cdot P(A_3|V_1 \cap A_2) = \frac{2}{5} \cdot \frac{3}{4} \cdot \frac{2}{3}$.

♣

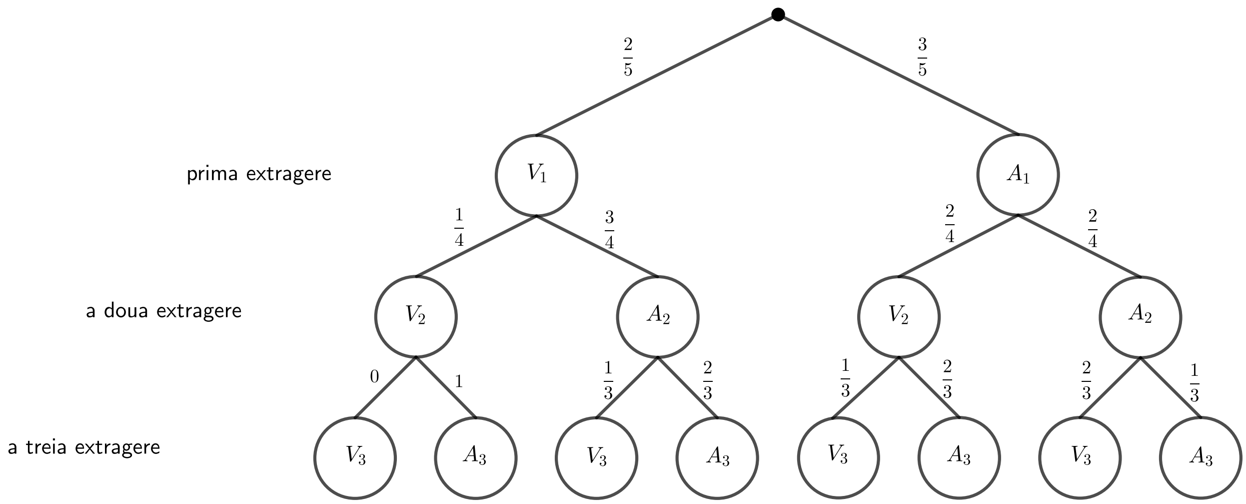


Fig. 3. Extragere fără returnare

Formula lui Bayes

Formula lui Bayes este o metodă de a "corecta" (a revizui, a îmbunătăţi) pe baza unor noi date (informaţii) disponibile o probabilitate determinată apriori. Se porneşte cu o estimare pentru probabilitatea unei anumite ipoteze H (engl. *hypothesis*). Dacă avem noi date (date de antrenare, dovezi, informaţii, evidenţe - engl. *evidence*) E, ce privesc ipoteza H, se poate calcula o probabilitate "corectată" pentru ipoteza H, numită probabilitate posterioară (a-posteriori).
$\hookrightarrow$ $P(H)$ probabilitatea ca ipoteza H să fie adevărată, numită şi *probabilitatea apriori*;
$\hookrightarrow$ probabilitatea condiţionată $P(H|E)$ este *probabilitatea posterioară* (corectată de cunoaşterea noilor date / informaţii);
$\hookrightarrow$ $P(E|H)$ probabilitatea ca să apară datele (informaţiile), ştiind că ipoteza H este adevarată;
$\hookrightarrow$ $P(E|\bar{H})$ probabilitatea ca să apară datele (informaţiile), ştiind că ipoteza H este falsă (ipoteza $\bar{H}$ este adevarată).
Folosind P.5 are loc:

$$P(E) = P(E|H)\cdot P(H) + P(E|\bar{H})\cdot P(\bar{H}) = P(E|H)\cdot P(H) + P(E|\bar{H})\cdot(1 - P(H)).$$

Formula lui Bayes este în acest caz

$$P(H|E) = \frac{P(E|H)\cdot P(H)}{P(E)} = \frac{P(E|H)\cdot P(H)}{P(E|H)\cdot P(H) + P(E|\bar{H})\cdot P(\bar{H})}.$$

P. 7. (Formula lui Bayes)
Într-un spaţiu de probabilitate $(\Omega, \mathcal{K}, P)$ *considerăm partiţia* $\{H_1, \ldots, H_n\}$ *a lui* Ω *cu* $H_i \in \mathcal{K}$

şi $P(H_i) > 0 \; \forall \, i \in \{1, ...n\}$, *şi fie* $E \in \mathcal{K}$ *astfel încât* $P(E) > 0$. *Atunci,*

$$P(H_j|E) = \frac{P(E|H_j)P(H_j)}{P(E)} = \frac{P(E|H_j)P(H_j)}{P(E|H_1)P(H_1) + ... + P(E|H_n)P(H_n)} \quad \forall \, j \in \{1, 2, ..., n\}.$$

▷ pentru $i \in \{1, 2, ..., n\}$ $P(H_i)$ sunt **probabilităţi apriori** pentru H_i, numite şi ipoteze (aserţiuni; engl. *hypothesis*)
▷ E se numeşte **evidenţă** (dovadă, premisă, informaţie; engl. *evidence*);
▷ cu formula lui Bayes se calculează probabilităţile pentru ipoteze, cunoscând evidenţa: $P(H_j|E)$, $j \in \{1, 2, ..., n\}$, care se numesc **probabilităţi posterioare** (ulterioare);
▷ $P(E|H_i)$, $i \in \{1, 2, ..., n\}$, reprezintă verosimilitatea (engl. *likelihood*) datelor observate.
▷ Se pot calcula probabilităţile *cauzelor*, date fiind (cunoscând / ştiind) *efectele*; formula lui Bayes ne ajută să diagnosticăm o anumită situaţie sau să testăm o ipoteză.

Exemplu: Considerăm evenimentele (în teste clinice):
H: o persoană aleasă aleator dintr-o populaţie are o anumită alergie $\mathcal{A}$
E: testul clinic returnează `pozitiv` privind alergia $\mathcal{A}$
$\bar{E}$: testul clinic returnează `negativ` privind alergia $\mathcal{A}$
▷ din statistici anterioare sunt cunoscute:
$p = P(H)$, probabilitatea ca o persoană selectată aleator din populaţie să sufere de alergia $\mathcal{A}$;
sensibilitatea testului $s_1 = P(E|H)$;
specificitatea testului $s_2 = P(\bar{E}|\bar{H})$;
▷ probabilitatea de a obţine răspuns `fals pozitiv` este $P(E|\bar{H}) = 1 - s_2$;
▷ un test clinic bun implică valori apropiate de 1 pentru s_1 şi s_2;
▶ cunoscând p, s_1, s_2 se doreşte a se determina *valoarea predictivă* $P(H|E)$:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)} = \frac{P(E|H) \cdot P(H)}{P(E|H) \cdot P(H) + P(E|\bar{H}) \cdot P(\bar{H})} = \frac{s_1 \cdot p}{s_1 \cdot p + (1 - s_2) \cdot (1 - p)}.$$

♣

Variabile aleatoare

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

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

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

Notaţie 1. *variabilă/variabile aleatoare* $\to$ *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

$(x_1, \ldots, x_n, \ldots)$

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

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

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

Variabile aleatoare numerice - definiţie formală

Def. 10. *Fie* $(\Omega, \mathcal{K}, P)$ *spaţiu de probabilitate.* $X : \Omega \to \mathbb{R}$ *este o variabilă aleatoare, dacă*

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

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

Def. 11. ***Distribuţia de probabilitate*** *a v.a. discrete* X

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

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

▷ O variabilă aleatoare discretă X este caracterizată de **distribuţia de probabilitate** $P[X]$:

(1) $\quad P[X] : \{x_1, x_2, \ldots\} \to [0, 1]$, definită prin $P[X](x) = P(X = x) \; \forall \, x \in \{x_1, x_2, \ldots\}$.

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

Distribuţii discrete clasice

Distribuţia discretă uniformă: $X \sim Unid(n),\ n \in \mathbb{N}^*$

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

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

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

Matlab/Octave: `unidrnd`$(n, ...)$, `randi`$(n, ...)$ generează valori aleatoare; `unidpdf`(x, n) calculează $P(X = x)$, dacă $X \sim Unid(n)$.

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

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

Exemplu: în cadrul unui experiment poate să apară evenimentul A (*succes*) sau $\bar{A}$ (*insucces*)
$X = 0 \Leftrightarrow$ dacă $\bar{A}$ apare; $X = 1 \Leftrightarrow$ dacă A apare
$\Rightarrow X \sim Bernoulli(p)$ cu $p := P(A)$

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

♣

generare în Matlab/Octave:

```
n=1000;
p=0.3;
nr=rand(1,n);
X=(nr<=p) % vector de date avand distributia Bernoulli(p)
%%%%%%%%
Y=floor(rand(1,n)+p)% vector de date avand distributia Bernoulli(p)
%%%%%%%%
```

Distribuţia binomială: $X \sim Bino(n, p), n \in \mathbb{N}^*, p \in (0, 1)$
în cadrul unui experiment poate să apară evenimentul A (*succes*) sau $\bar{A}$ (*insucces*)
• $A =$ succes cu $P(A) = p$, $\quad \bar{A} =$ 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: $X \in \{0, 1, \dots, n\}$

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

$$X \sim Bino(n, p) \iff X \sim \begin{pmatrix} k \\ C_n^k p^k (1-p)^{n-k} \end{pmatrix}_{k \in \{0, \dots, n\}}$$

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

$\rightarrow$ are loc formula binomială

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

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

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

Matlab/Octave: `binornd`$(n, p, ...)$ generează valori aleatoare; `binopdf`(x, n, p) calculează $P(X = x)$, dacă $X \sim Bino(n, p)$.

$\triangleright$ Distribuţia binomială corespunde **modelului cu extragerea bilelor dintr-o urnă cu bile de două culori şi cu returnarea bilei după fiecare extragere**:
Într-o urnă sunt n_1 bile albe şi n_2 bile negre. Se extrag cu returnare n bile; fie v.a. X_1 = numărul de bile albe extrase; X_2 = numărul de bile negre extrase
$\Rightarrow X_1 \sim Bino(n, p_1)$ cu $p_1 = \frac{n_1}{n_1+n_2}$, $X_2 \sim Bino(n, p_2)$ cu $p_2 = \frac{n_2}{n_1+n_2}$.

Exemplu: Fie un canal de comunicare binară care transmite cuvinte codificate de N biţi fiecare. Probabilitatea transmiterii cu succes a unui singur bit este p, iar probabilitatea unei erori este $1 - p$. Presupunem, de asemenea, că un astfel de cod este capabil să corecteze până la m erori (într-un cuvânt), unde $0 \leq m \leq N$. Se ştie că transmiterea biţilor succesivi este independentă, atunci probabilitatea transmiterii cu succes a unui cuvânt este $P(A)$, unde
A: cel mult m erori apar în transmiterea celor N biţi

$$P(A) = \sum_{k=0}^{m} C_N^k p^{N-k} (1-p)^k.$$

♠

Exerciţii: 1) Un client accesează o dată pe zi o anumită pagină web, care oferă produse bio, cu probabilitatea 0.4. Cu ce probabilitate clientul accesează această pagină în total de 3 ori în următoarele 6 zile?
2) O reţea de laborator este compusă din 15 calculatoare. Reţeaua a fost atacată de un virus nou, care atacă un calculator cu o probabilitatea 0.4, independent de alte calculatoare. Care este probabilitatea ca virusul a atacat a) cel mult 10 computere; b) cel puţin 10 calculatoare; c) exact 10 calculatoare? ♣

Distribuţia hipergeometrică: $X \sim Hyge(n, n_1, n_2),\ n, n_1, n_2 \in \mathbb{N}^*$

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

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

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

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

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

Matlab/Octave: `hygernd`($n_1 + n_2, n_1, n, \ldots$) generează valori aleatoare;
`hygepdf`($x, n_1 + n_2, n_1, n$) calculează $P(X = x)$, dacă $X \sim Hyge(n, n_1, n_2)$.

Exemplu: 1) Într-o urnă sunt $n_1 = 2$ bile albe şi $n_2 = 3$ bile negre. Se extrag fără returnare $n = 3$ bile. Fie v.a. X = numărul de bile albe extrase. Vom calcula $P(X = 1)$ cu două metode:
Prima metodă: Pentru $i \in \{1, 2, 3\}$ fie evenimentele
A_i: la a i-a extragere s-a obţinut bilă albă
$N_i = \bar{A}_i$: la a i-a extragere s-a obţinut bilă neagră.
Scriem

$$P(X = 1) = P(A_1 \cap N_2 \cap N_3) + P(A_1 \cap N_2 \cap N_3) + P(A_1 \cap N_2 \cap N_3),$$

$$P(A_1 \cap N_2 \cap N_3) = P(A_1)P(N_2|A_1)P(N_3|A_1 \cap N_2) = \frac{2}{5} \cdot \frac{3}{4} \cdot \frac{2}{3} = \frac{1}{5}$$

$$P(N_1 \cap A_2 \cap N_3) = P(N_1)P(A_2|N_1)P(N_3|N_1 \cap A_2) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} = \frac{1}{5}$$

$$P(N_1 \cap N_2 \cap A_3) = P(N_1)P(N_2|N_1)P(A_3|N_1 \cap N_2) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} = \frac{1}{5}$$

$$\Rightarrow P(X = 1) = \frac{3}{5}.$$

A doua metodă: O bilă albă din două se poate alege în $C_2^1 = 2$ moduri, două bile neagre din trei se pot alege în $C_3^2 = 3$ moduri, trei bile din cinci se pot alege în $C_5^3 = 10$ moduri

$$\Rightarrow P(X = 1) = \frac{C_2^1 \cdot C_3^2}{C_5^3} = \frac{2 \cdot 3}{10} = \frac{3}{5}.$$

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

numere să fie câştigătoare?

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

$\diamondsuit$

Distribuţia geometrică $X \sim Geo(p), p \in (0,1)$
În cadrul unui experiment poate să apară evenimentul A (*succes*) sau $\bar{A}$ (*insucces*)
• A = succes cu $P(A) = p$, $\quad \bar{A}$ = insucces $P(\bar{A}) = 1 - p$
• 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 apariţia primului A ("succes") $\Rightarrow$ valori posibile: $X \in \{0, 1, \ldots\}$

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

Matlab/Octave: `geornd`$(p, \ldots)$ generează valori aleatoare; `geopdf`(x, p) calculează $P(X = x)$, dacă $X \sim Geo(p)$.

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

$\spadesuit$

Variabile aleatoare independente

Def. 12. *Variabilele aleatoare discrete X (care ia valorile $\{x_i, i \in I\}$) şi Y (care ia valorile $\{y_j, j \in J\}$) sunt **independente**, dacă şi numai dacă*

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

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

Observaţie: Fie evenimentele $A_i = \{X = x_i\}, i \in I$, şi $B_j = \{Y = y_j\}, j \in J$.
V.a. X şi Y sunt independente $\Longleftrightarrow \forall\, (i,j) \in I \times J$ evenimentele A_i şi B_j sunt independente (a se vedea Def. 6).

Exemplu: Se aruncă o monedă de 10 ori. Fie X v.a. care indică de câte ori a apărut pajură în primele cinci aruncări ale monedei; fie Y v.a. care indică de câte ori a apărut pajură în ultimele cinci aruncări ale monedei. X şi Y sunt v.a. independente. Care este distribuţia de probabilitate a lui X, respectiv Y?

P. 8. *Fie variabilele aleatoare discrete X (care ia valorile $\{x_i, i \in I\}$) şi Y (care ia valorile $\{y_j, j \in J\}$). Sunt echivalente afirmaţiile:*

(1) X *şi* Y *sunt v.a. sunt independente;*
(2) $P(X=x|Y=y)=P(X=x) \quad \forall x\in\{x_i, i\in I\}, y\in\{y_j, j\in J\}$;
(3) $P(Y=y|X=x)=P(Y=y) \quad \forall x\in\{x_i, i\in I\}, y\in\{y_j, j\in J\}$;
(4) $P(X\le x, Y\le y)=P(X\le x)\cdot P(Y\le y) \quad \forall\, x,y\in\mathbb{R}$.

Def. 13. $\mathbb{X}=(X_1,\ldots,X_m)$ *este un* ***vector aleator discret*** *dacă fiecare componentă a sa este o variabiă aleatoare discretă.*
Fie $K\subseteq\mathbb{N}$ *o mulţime de indici şi fie date* $\mathbb{x}_k := (x_{1,k},...,x_{m,k})\in\mathbb{R}^m, k\in K$.
Dacă $\mathbb{X}:\Omega\to\{\mathbb{x}_k, k\in K\}$ *este un vector aleator discret, atunci*

$$P(\mathbb{X}=\mathbb{x}_k) := P\big(\{\omega\in\Omega : \mathbb{X}(\omega)=\mathbb{x}_k\}\big),\ k\in K,$$

determină ***distribuţia de probabilitate a vectorului aleator discret*** $\mathbb{X}$

$$\mathbb{X}\sim\begin{pmatrix} \mathbb{x}_k \\ P(\mathbb{X}=\mathbb{x}_k)\end{pmatrix}_{k\in K}.$$

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

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

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

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

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

Exemplu: Fie vectorul aleator discret (X,Y) cu distribuţia dată de

următorul tabel:

$X \backslash Y$	0	1
-1	$\frac{1}{4}$	$\frac{1}{2}$
2	$\frac{1}{8}$	$\frac{1}{8}$

$\Longrightarrow P(X=-1,Y=0)=\frac{1}{4}, P(X=-1,Y=1)=\frac{1}{2}$, etc.

a) Să se determine $P(X=-1)$, $P(X\le 3)$, respectiv $P(Y=1)$, $P(Y\le -1)$.
b) Sunt X şi Y v.a. independente?

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

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

$\triangleright$ Dacă X şi Y sunt v.a. independente, şi se ştiu distribuţiile lor, atunci distribuţia vectorului aleator (X, Y) se determină pe baza formulei (2).
$\triangleright$ Dacă se cunoaşte distribuţia vectorului aleator (X, Y) distribuţiile lui X şi Y se determină astfel:

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

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

Exemplu:
$\triangleright$ **Modelul urnei cu r culori cu returnarea bilei după fiecare extragere:** fie p_i probabilitatea de a extrage o bilă cu culoarea i, $i = \overline{1, r}$ dintr-o urnă; fie X_i v.a. ce indică numărul de bile de culoarea i, $i = \overline{1, r}$ după n extrageri *cu returnarea bilei extrase*, iar ordinea de extragere a bilelor de diverse culori nu contează

$$\begin{aligned} P(X_1 = k_1, \dots, X_r = k_r) &= \text{probabilitatea de a obţine } k_i \text{ bile cu culoarea } i,\ i = \overline{1, r}, \\ &\quad\ \text{din } n = k_1 + \dots + k_r \text{ extrageri } \textit{cu returnarea} \text{ bilei extrase} \\ &= \frac{n!}{k_1! \dots k_r!} \cdot p_1^{k_1} \cdot \dots \cdot p_r^{k_r}, \end{aligned}$$

$\triangleright$ $(X_1, ..., X_r)$ este un vector aleator discret.
$\triangleright$ cazul $r = 2$ corespunde distribuţiei binomiale (modelul binomial cu bile de două culori într-o urnă, a se vedea pg. 18): (X_1, X_2) este un vector aleator discret, iar $X_1 + X_2 = n$; X_1 şi X_2 *nu* sunt v.a. independente.

Operaţii cu variabile aleatoare (numerice)

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

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

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

. Determinaţi: a) distribuţiile variabilelor aleatoare X_1 şi X_2;
b) distribuţiile variabilelor aleatoare $X_1 + X_2$ şi $X_1 \cdot X_2$, $X_1^2 - 1$;
c) dacă variabilele aleatoare X_1 şi X_2 sunt independente sau dependente.

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

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

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

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

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

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

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

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

Exerciţiu: Se aruncă două zaruri. a) Să se scrie distribuţia de probabilitate pentru variabila aleatoare, care este suma celor două numere apărute. b) Să se scrie distribuţia de probabilitate pentru variabila aleatoare, care este produsul celor două numere apărute. ♠

Def. 14. ***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 *tendinţa centrală* a valorilor acesteia.

P. 9. *Fie X şi Y v.a. discrete. Au loc proprietăţile:*

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

→ *$E(X + Y) = E(X) + E(Y)$;*

→ *Dacă X şi Y sunt v.a. independente, atunci $E(X \cdot Y) = E(X)E(Y)$.*

→ *Dacă $g : \mathbb{R} \to \mathbb{R}$ e o funcţie astfel încât $g(X)$ este v.a., 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(x)`
pentru $x = [x(1), ..., x(n)]$, se calculează $\texttt{mean}(x) = \frac{1}{n}\big(x(1) + ... + x(n)\big)$

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

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

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

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

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

```
clear all
A=[zeros(1,50), zeros(1,70)+1,zeros(1,80)+2];
index=randperm(length(A));
A=A(index);
c=0;
i=randi(length(A));
while A(i)~=0
c=c+1;
i=randi(length(A));
end
fprintf('nr. iteratii inainte sa apara primul 0: %d \n',c)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
clear all
A=[zeros(1,50), zeros(1,70)+1,zeros(1,80)+2];
s=[];
N=1000;
for j=1:N
index=randperm(length(A));
A=A(index);
```

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

Probabilitatea să apară la orice iteraţie 0 este $p = \frac{50}{200} = 0.25$.
Notăm cu X v.a. care indică numărul de iteraţii necesare *înainte* să apară primul 0
$\Rightarrow X \sim Geo(p)$.
Numărul mediu de iteraţii necesare *înainte* să apară primul 0 este $E(X)$. Se poate arăta că $E(X) = \frac{1-p}{p} = \frac{1-0.25}{0.25} = 3$. ▼

Def. 15. *Fie $X_1, \ldots, X_n$ cu $n \in \mathbb{N}$, $n \geq 2$, variabile aleatoare discrete, care iau valori în mulţimile $\mathcal{X}_1, \ldots, \mathcal{X}_n$. $X_1, \ldots, X_n$ sunt* ***variabile aleatoare independente****, dacă*

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(X_1 = x_1) \cdot \ldots \cdot P(X_n = x_n)$$

pentru fiecare $x_1 \in \mathcal{X}_1, \ldots, x_n \in \mathcal{X}_n$.

Exemplu: Se aruncă patru zaruri. Fie X_i v.a. care indică numărul apărut la al i-lea zar.
a) X_1, X_2, X_3, X_4 sunt v.a. independente;
b) $X_1 + X_2$ şi $X_3 + X_4$ sunt v.a. independente;
c) $X_1 + X_2 + X_3$ şi X_4 sunt v.a. independente.

Def. 16. ***Funcţia de repartiţie*** *$F : \mathbb{R} \to [0, 1]$ a unei variabile aleatoare X discrete, care ia valorile $\{x_i, i \in I\}$, este*

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

Exemplu: Fie v.a. discretă X dată prin:

$$P(X = -1) = 0.5,\ P(X = 1) = 0.3,\ P(X = 4) = 0.2\,.$$

$\Longrightarrow X$ are funcţia de repartiţie $F_X : \mathbb{R} \to [0, 1]$

$$F_X(x) = P(X \leq x) = \begin{cases} 0, & \text{dacă } x < -1 \\ 0.5, & \text{dacă } -1 \leq x < 1 \\ 0.5 + 0.3 = 0.8, & \text{dacă } 1 \leq x < 4 \\ 0.5 + 0.3 + 0.2 = 1, & \text{dacă } 4 \leq x\,. \end{cases}$$

P. 10. *Funcţia de repartiţie F a unei variabile aleatoare discrete X are următoarele proprietăţi:*
(1) $F(b) - F(a) = P(X \leq b) - P(X \leq a) = P(a < X \leq b)\ \forall a, b \in \mathbb{R}, a < b.$
(2) *F este monoton crescătoare, adică pentru orice $x_1 < x_2$ rezultă $F(x_1) \leq F(x_2)$.*
(3) *F este continuă la dreapta, adică $\lim\limits_{x \searrow x_0} F(x) = F(x_0)\ \forall\ x_0 \in \mathbb{R}$.*
(4) $\lim\limits_{x\to\infty} F(x) = 1$ *şi* $\lim\limits_{x\to-\infty} F(x) = 0.$

Matlab/Octave: `binocdf(`x, n, p`)`, `hygecdf(`$x, n_1{+}n_2, n_1, n$`)`, `geocdf(`x, p`)` calculează $F(x) = P(X \leq x)$ pentru $X \sim Bino(n, p)$, $X \sim Hyge(n_1 + n_2, n_1, n)$, respectiv $X \sim Geo(p)$.

```
pkg load statistics
clear all
close all
% X~Bino(n,p)
n=5; % nr. repetari ale experimentului
p=0.4; %probabilitatea de a obtine succes
x=-1:0.001:6;
y=binocdf(x,n,p);
plot(x,y,'r.')
title('FUNCTIA DE REPARTITIE - Distr. binomiala')
```

Variabile aleatoare continue

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

Def. 17. ***Funcţia de densitate*** *a unei v.a. continue X este funcţia $f : \mathbb{R} \to \mathbb{R}$ pentru care are loc*

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

Funcţia $F : \mathbb{R} \to [0, 1]$ definită prin

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

se numeşte ***funcţia de repartiţie*** *a v.a. continue X.*

P. 11. *Fie f funcţia de densitate şi F funcţia de repartiţie a unei v.a. continue X. Au loc proprietăţile:*

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

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

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

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

(5) *pentru* $\forall\, a < b, a, b \in \mathbb{R}$ *au loc*

$$F(b)-F(a) = P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) = \int\limits_a^b f(t)dt;$$

(6) F *este o funcţie monoton crescătoare şi continuă pe* $\mathbb{R}$*;*

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

(8) *dacă* F *este derivabilă în punctul* x*, atunci* $F'(x) = f(x)$*.*

Observaţie: Orice funcţie $f : \mathbb{R} \to \mathbb{R}$, care are proprietăţile (1), (2) din **P.11** este o funcţie de densitate.

Exemple de distribuţii clasice continue

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

• funcţia de densitate este

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

Matlab/Octave:

▷ pentru $a = 0, b = 1$: `rand`(M, N) returnează o matrice $M \times N$ cu valori aleatoare din [0,1]

▷ `unifrnd`(a, b, M, N), respectiv $(b - a)$`rand`$(M, N) + a$ returnează o matrice $M \times N$ cu valori aleatoare din $[a, b]$

▷ pentru $X \sim Unif[a, b]$: `unifpdf`(x, a, b) calculează $f(x)$, iar `unifcdf`(x, a, b) calculează $F(x) = P(X \leq x)$.

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

• funcţia de densitate este

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

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

• Pentru $m = 0, \sigma = 1$: $N(0, 1)$ se numeşte *distribuţia standard normală.*
• Distribuţia normală se aplică în: măsurarea erorilor (de ex. termenul eroare în analiza regresională), în statistică (teorema limită centrală, teste statistice) etc.
Matlab/Octave: `normrnd`(m, σ, M, N) returnează o matrice $M \times N$ cu valori aleatoare;
$\triangleright$ pentru $X \sim N(m, \sigma^2)$: `normpdf`(x, m, σ) calculează $f(x)$, iar `normcdf`(x, m, σ) calculează $F(x) = P(X \leq x)$.

➡Distribuţia exponenţială: $X \sim Exp(\lambda)$, $\lambda > 0$
• funcţia de densitate este

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

Matlab/Octave: `exprnd`$\left(\frac{1}{\lambda}, M, N\right)$ returnează o matrice $M \times N$ cu valori aleatoare;
$\triangleright$ pentru $X \sim Exp(\lambda)$: `exppdf`$\left(x, \frac{1}{\lambda}\right)$ calculează $f(x)$, iar `expcdf`$\left(x, \frac{1}{\lambda}\right)$ calculează $F(x) = P(X \leq x)$.

```
pkg load statistics
clear all
close all
figure
title('Functia de densitate a legii exponentiale')
hold on
L=[1,2,4]; %  lambda parametru
t=[-1:0.01:2];
plot(t, exppdf(t,1/L(1)), 'r*')
plot(t, exppdf(t,1/L(2)), 'b*')
plot(t, exppdf(t,1/L(3)), 'g*')
legend('lambda=1','lambda=2','lambda=4')
```

➡**Distribuţia Student:** $X \sim St(n)$, $n \in \mathbb{N}^*$

• distribuţia Student cu $n \in \mathbb{N}^*$ grade de libertate are funcţia de densitate

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \ t \in \mathbb{R}$$

unde funcţia Gamma este

$$\Gamma(a) = \int_0^\infty v^{a-1}\exp(-v)dv, \ a > 0$$

Matlab/Octave: $\texttt{trnd}(n, M, N)$ returnează o matrice $M \times N$ cu valori aleatoare;
$\triangleright$ pentru $X \sim St(n)$: $\texttt{tpdf}(x, n)$ calculează $f(x)$, iar $\texttt{tcdf}(x, n)$ calculează $F(x) = P(X \leq x)$.

➡**Distribuţia Chi-pătrat:** $X \sim \chi^2(n)$, $n \in \mathbb{N}^*$

• distribuţia χ^2 cu $n \in \mathbb{N}^*$ grade de libertate are funcţia de densitate

$$f(x) = \begin{cases} 0, & \text{dacă } x \leq 0 \\ \dfrac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} \cdot x^{\frac{n}{2}-1} \cdot \exp\left(-\dfrac{t}{2}\right), & \text{dacă } x > 0, \end{cases}$$

Matlab/Octave: $\texttt{chi2rnd}(n, M, N)$ returnează o matrice $M \times N$ cu valori aleatoare;
$\triangleright$ pentru $X \sim \chi^2(n)$: $\texttt{chi2pdf}(x, n)$ calculează $f(x)$, iar $\texttt{chi2cdf}(x, n)$ calculează $F(x) = P(X \leq x)$.

Exemplu: Fie $X \sim Exp(0.5)$ v.a. care indică timpul de funcţionare a unei baterii (câte luni funcţionează bateria). Folosind simulări, să se estimeze a) $P(2 \leq X \leq 4)$; b) $P(X > 3)$ şi să se compare rezultatele obţinute cu rezultatele teoretice.

```
pkg load statistics
N=10000;
X=exprnd(2,1,N);
p=sum((2<=X)&(X<=4))/N
q=sum(X>3)/N
> p=0.23280
> q=0.22060
```

$$P(2 \leq X \leq 4) = \int_2^4 0.5e^{-0.5t}dt = -e^{-0.5t}\Big|_2^4 = e^{-1} - e^{-2} \approx 0.23254$$

$$P(X > 3) = 1 - \int_{-\infty}^3 0.5e^{-0.5t}dt = \int_3^\infty 0.5e^{-0.5t}dt = -e^{-0.5t}\Big|_3^\infty = e^{-1.5} \approx 0.22313 \quad \clubsuit$$

Matlab/Octave:

Distribuţia v.a. discrete X	Generare valori aleatoare	Funcţia de repartiţie $F_X(x) = P(X \leq x)$	Probabilitate $P(X = x)$
$Bino(n, p)$	`binornd`(n, p)	`binocdf`(x, n, p)	`binopdf`(x, n, p)
$Unid(n)$	`unidrnd`(n)	`unidcdf`(x, n)	`unidpdf`(x, n)
$Hyge(n,n_1,n_2)$	`hygernd`(n_1+n_2,n_1,n)	`hygecdf`(x,n_1+n_2,n_1,n)	`hygepdf`(x,n_1+n_2,n_1,n)
$Geo(p)$	`geornd`(p)	`geocdf`(x, p)	`geopdf`(x, p)

Distribuţia v.a. continue X	Generare valori aleatoare	Funcţia de repartiţie $F_X(x) = P(X \leq x)$	Funcţia de densitate $f_X(x)$
$Unif[a, b]$	`unifrnd`(a, b)	`unifcdf`(x, a, b)	`unifpdf`(x, a, b)
$N(m, \sigma^2)$	`normrnd`(m, σ)	`normcdf`(x, m, σ)	`normpdf`(x, m, σ)
$Exp(\lambda)$	`exprnd`$(\frac{1}{\lambda})$	`expcdf`$(x, \frac{1}{\lambda})$	`exppdf`$(x, \frac{1}{\lambda})$

Observaţie: Dacă în cadrul aceluiaşi program Matlab/Octave se generează valori aleatoare (de exemplu cu `rand`, `randi`, `binornd`, `hygernd`, `unidrnd`, `geornd`, `unifrnd`, `normrnd`, `exprnd`, etc.) atunci acestea pot fi considerate ca fiind valorile unor variabile aleatoare independente.

Proprietăţi

V.a. discretă	V.a. continuă
• caracterizată de distribuţia de probabilitate discretă $X \sim \begin{pmatrix} x_i \\ P(X=x_i) \end{pmatrix}_{i\in I}$	• caracterizată de funcţia de densitate f $P(X \leq x) = \int_{-\infty}^{x} f(t)dt$
• $\sum_{i\in I} P(X=x_i) = 1$	• $\int_{-\infty}^{\infty} f(t)dt = 1$
• $P(X \in A) = \sum_{i\in I:\, x_i\in A} P(X=x_i)$	• $P(X \in A) = \int_A f(t)dt$
• funcţia de repartiţie $F(x)=P(X \leq x)\ \forall x \in \mathbb{R}$	• funcţia de repartiţie $F(x)=P(X \leq x)\ \forall x\in\mathbb{R}$
• $F(x) = \sum_{i\in I:\, x_i\leq x} P(X=x_i)\ \forall x \in \mathbb{R}$	• $F(x) = \int_{-\infty}^{x} f(t)dt \quad \forall x \in \mathbb{R}$
• F este funcţie continuă la dreapta	• F este funcţie continuă în orice punct $x \in \mathbb{R}$
• F este discontinuă în punctele $x_i,\ \forall\, i \in I$	
• $\forall\, a < b, a, b \in \mathbb{R}$ $P(a \leq X \leq b) = \sum_{i\in I: a\leq x_i\leq b} P(X=x_i)$	• $\forall\, a < b, a, b \in \mathbb{R}$ $P(a \leq X \leq b) = P(a < X \leq b)$ $= P(a \leq X < b) = P(a < X < b) = \int_a^b f(t)\,dt$
• $P(X=a) = 0$ dacă $a \notin \{x_i : i \in I\}$	• $P(X=a) = \int_a^a f(t)\,dt = 0\ \forall\, a \in \mathbb{R}$
	• dacă F este derivabilă în punctul x $\Rightarrow\ F'(x) = f(x)$.

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

Vector aleator continuu

Def. 18. $(X_1, \dots, X_n)$ *este un* ***vector aleator continuu*** *dacă fiecare componentă a sa este o variabiă aleatoare continuă.*

Def. 19. $f_{(X,Y)} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ *este* ***funcţia de densitate a vectorului aleator continuu*** (X, Y)*, dacă*

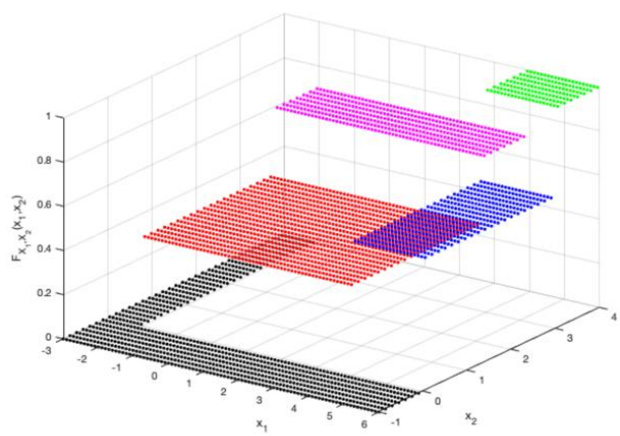
$$P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \left(\int_{-\infty}^{y} f_{(X,Y)}(s,t)dt \right) ds \;\; \forall x, y \in \mathbb{R}.$$

Def. 20. $F_{(X,Y)} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ *este* ***funcţia de repartiţie a vectorului aleator*** (X, Y) *(discret sau continuu), dacă*

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

Exemplu: *Vectorul aleator discret* (X_1, X_2) este dat prin următorul tabel de contingenţă:

X_1 \ X_2	0	3
-2	0.4	0.3
4	0.2	0.1



Funcţia de repartiţie $F_{(X_1,X_2)}$

$\Longrightarrow$ (X_1, X_2) are funcţia de repartiţie $F_{(X_1,X_2)} : \mathbb{R} \times \mathbb{R} \to [0, 1]$

$$F_{(X_1,X_2)}(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2) = \begin{cases} 0, & \text{dacă } x_1 < -2 \text{ sau } x_2 < 0 \\ 0.4, & \text{dacă } -2 \leq x_1 < 4 \text{ şi } 0 \leq x_2 < 3 \\ 0.7, & \text{dacă } -2 \leq x_1 < 4 \text{ şi } 3 \leq x_2 \\ 0.6, & \text{dacă } 4 \leq x_1 \text{ şi } 0 \leq x_2 < 3 \\ 1, & \text{dacă } 4 \leq x_1 \text{ şi } 3 \leq x_2\,. \end{cases}$$

♣

P. 12. *Pentru un* ***vector aleator continuu*** (X, Y) *au loc proprietăţile:*

1. $\displaystyle\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} f_{(X,Y)}(u,v)\,dv\right)du = 1.$
2. $F_{(X,Y)}$ *este funcţie continuă.*
3. *Dacă* $F_{(X,Y)}$ *este derivabilă parţial în* (x, y), *atunci are loc:*
$$\frac{\partial^2 F_{(X,Y)}(x,y)}{\partial x \partial y} = f_{(X,Y)}(x,y).$$
4. $\displaystyle P((X,Y)\in A) = \underbrace{\int\int}_{A} f_{(X,Y)}(u,v)dudv,\ A \subset \mathbb{R}^2.$

► Dacă se cunoaşte funcţia de repartiţie $F_{(X,Y)}$ pentru vectorul aleator (X, Y) (discret sau continuu), atunci F_X, respectiv F_Y, se determină cu

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

Exemplu: Funcţia de repartiţie a vectorului aleator (X, Y) este $F_{(X,Y)} : \mathbb{R}\times\mathbb{R} \to [0, 1]$

$$F_{(X,Y)}(x,y) = \begin{cases} 0, & \text{dacă } x < 0 \text{ sau } y < 1 \\ x(y-1), & \text{dacă } 0 \le x < 1 \text{ şi } 1 \le y < 2 \\ x, & \text{dacă } 0 \le x < 1 \text{ şi } 2 \le y \\ y-1, & \text{dacă } 1 \le x \text{ şi } 1 \le y < 2 \\ 1, & \text{dacă } 1 \le x \text{ şi } 2 \le y\,. \end{cases}$$

Ce distribuţie au X, respectiv Y?
R.: Se determină F_X, F_Y cu (3) şi se calculează $f_X = F'_X, f_Y = F'_Y$; se obţine $X \sim Unif[0,1], Y \sim Unif[1,2]$. ♣

► Dacă se cunoaşte funcţia de densitate $f_{(X,Y)}$ pentru vectorul aleator continuu (X, Y), atunci f_X, respectiv f_Y, se determină cu

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

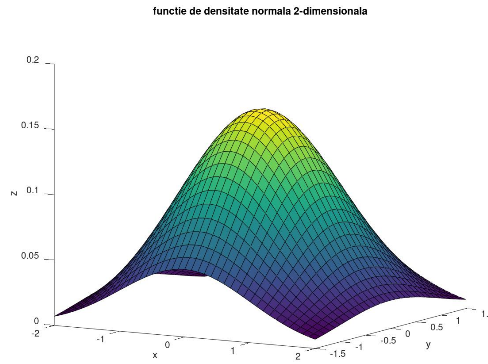
Exemplu pentru o distribuţie normală bidimensională: (X, Y) are funcţia de densitate (graficul acestei funcţii este dat în figura alăturată)

$$f_{(X,Y)}(x, y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}}, \ \ x, y \in \mathbb{R}.$$

$$\Longrightarrow f_X(x) = \int_{-\infty}^{\infty} f_{(X,Y)}(x, y) dy = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \ \forall x \in \mathbb{R},$$

$$\Longrightarrow f_Y(y) = \int_{-\infty}^{\infty} f_{(X,Y)}(x, y) dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}, \ \forall y \in \mathbb{R}.$$

$$\Longrightarrow \ X, Y \sim N(0, 1).$$



$f_{(X,Y)}$ pentru distribuţia normală bidimensională

♣

Def. 21. *$X_1, \ldots, X_n$ sunt* ***n variabilele aleatoare independente*** *(discrete sau continue), dacă*

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

Observaţie ($n = 2$ in definiţia de mai sus): X_1 şi X_2 sunt **două variabilele aleatoare independente** (discrete sau continue), dacă

$$P(X_1 \leq x_1, X_2 \leq x_2) = P(X_1 \leq x_1) \cdot P(X_2 \leq x_2) \quad \forall \ x_1, x_2 \in \mathbb{R},$$

adică

$$F_{(X_1,X_2)}(x_1, x_2) = F_{X_1}(x_1) \cdot F_{X_2}(x_2) \quad \forall \ x_1, x_2 \in \mathbb{R}.$$

P. 13. *Variabilele aleatoare continue X_1 (cu funcţia de densitate f_{X_1}) şi X_2 (cu funcţia de densitate f_{X_2}) sunt independente, dacă şi numai dacă*

$$f_{(X_1,X_2)}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) \ \ \forall \ x_1, x_2 \in \mathbb{R},$$

unde $f_{(X_1,X_2)}$ este funcţia de densitate a vectorului aleator (X_1, X_2).

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

$$f_{(X_1,X_2)}(x_1, x_2) = \begin{cases} \dfrac{1}{(b_1 - a_1)(b_2 - a_2)} & \text{dacă } (x_1, x_2) \in I \\ 0 & \text{dacă } (x_1, x_2) \notin I. \end{cases}$$

Cu (4) se calculează

$$f_{X_1}(x_1) = \begin{cases} \dfrac{1}{b_1 - a_1} & \text{dacă } x_1 \in [a_1, b_1] \\ 0 & \text{dacă } x_1 \in \mathbb{R} \setminus [a_1, b_1]. \end{cases} \quad \text{şi } f_{X_2}(x_2) = \begin{cases} \dfrac{1}{b_2 - a_2} & \text{dacă } x_2 \in [a_2, b_2] \\ 0 & \text{dacă } x_2 \in \mathbb{R} \setminus [a_2, b_2]. \end{cases}$$

$\Longrightarrow X_1 \sim Unif[a_1, b_1], X_2 \sim Unif[a_2, b_2]$;

se observă $f_{(X_1,X_2)} = f_{X_1} \cdot f_{X_2} \Longrightarrow X_1$ şi X_2 sunt v.a. independente! ✿

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

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

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

R.: Se calculează $F_X(x) = 1 - e^{-x}$ pentru $x > 0$ şi $F_X(x) = 0$ pentru $x \leq 0$, precum şi $F_Y(y) = 1 - e^{-2y}$ pentru $y > 0$ şi $F_Y(y) = 0$ 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 şi Y sunt v.a. independente.

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

✿

Def. 22. ***Valoarea medie a unei v.a. continue*** *X, care are funcţia de densitate f, este*

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

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

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

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

$\rightarrow E(X + Y) = E(X) + E(Y)$;

$\rightarrow$ *Dacă X şi Y sunt variabile aleatoare* ***independente****, atunci* $E(X \cdot Y) = E(X)E(Y)$.

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

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

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

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

$$\Longrightarrow \ f_X(t) = \begin{cases} \frac{1}{26-20} = \frac{1}{6}, & \text{dacă } 20 \le t \le 26 \\ 0, & \text{în rest.} \end{cases}$$

$$P(\text{“elevul ajunge la timp la şcoală”}) = P(X \le 25) = \int_{-\infty}^{25} f_X(t)dt = \int_{20}^{25} \frac{1}{6}dt = \frac{25-20}{6} = \frac{5}{6}.$$

$$E(X) = \int_{-\infty}^{\infty} t f_X(t)dt = \int_{20}^{26} t \cdot \frac{1}{6}dt = \frac{1}{6} \cdot \frac{t^2}{2}\Big|_{20}^{26} = 23 \text{ (minute)}.$$

♣

Def. 23. ***Varianţa (dispersia)*** *unei variabile aleatoare X (discrete sau continue) este*

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

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

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

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

Exemple: 1) Fie $X \sim Bino(n, p)$. Să se arate că $E(X) = np$ şi $V(X) = np(1 - p)$.
R.: Pentru $i \in \{1, \dots, n\}$ fie $X_i \sim Bernoulli(p)$ (adică $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$), astfel încât $X_1, \dots, X_n$ sunt v.a. independente. Observăm că $X_1 + \ldots + X_n \sim Bino(n, p)$. Deci, $X_1 + \ldots + X_n$ şi X au aceeaşi distribuţie, aşadar ele au aceeaşi valoare medie şi aceeaşi varianţă

$$E(X) = E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n) = p + ... + p = np.$$

V.a. $X_1, \dots, X_n$ sunt independente şi folosind P.15, obţinem

$$V(X) = V(X_1 + \cdots + X_n) = V(X_1) + \cdots + V(X_n) = np(1 - p) = np(1 - p).$$

2) Dacă $X \sim N(m, \sigma^2)$ să se arate că $E(X) = m,\ V(X) = \sigma^2$.
R.: Funcţia de densitate a lui X este

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

Când $m = 0$ şi $\sigma = 1$ obţinem *funcţia de densitate a distribuţiei normale standard*

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}, x \in \mathbb{R}.$$

Din P.11-(2) rezultă

$$\int_{-\infty}^{\infty} \varphi(t)dt = 1.$$

În calculele de mai jos utilizăm schimbarea de variabilă $t = \dfrac{x-m}{\sigma}$

$$\begin{aligned} E(X) = \int_{-\infty}^{\infty} x f_X(x)dx &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\} dx \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t \exp\left\{-\frac{t^2}{2}\right\} dt + m \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{t^2}{2}\right\} dt \\ &= 0 + m \int_{-\infty}^{\infty} \varphi(t)dt = m\ . \end{aligned}$$

Folosind aceeaşi schimbare de variabilă şi apoi integrare prin părţi, avem

$$\begin{aligned} V(X) &= E[(X-m)^2] = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-m)^2 \exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\} dx \\ &= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 \exp\left\{-\frac{t^2}{2}\right\} dt = \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t \left(-\exp\left\{-\frac{t^2}{2}\right\}\right)' dt \\ &= t\left(-\exp\left\{-\frac{t^2}{2}\right\}\right)\Bigg|_{-\infty}^{\infty} - \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(-\exp\left\{-\frac{t^2}{2}\right\}\right) dt \\ &= 0 - 0 + \sigma^2 \int_{-\infty}^{\infty} \varphi(t)dt = \sigma^2. \end{aligned}$$

3) Vectorul aleator (X, Y) are funcţia de densitate

$$f_{(X,Y)} : \mathbb{R}^2 \to \mathbb{R} \qquad f_{(X,Y)}(x,y) = \begin{cases} x - y, & \text{dacă } 0 \le x \le 1 \text{ şi } -1 \le y \le 0 \\ 0, & \text{altfel .} \end{cases}$$

Să se calculeze $E(X)$ şi $E(X^2)$.
R.:

$$f_X(x) = \int_{-\infty}^{\infty} f_{(X,Y)}(x,y)dy = \begin{cases} \displaystyle\int_{-1}^{0} (x-y)\,dy = x + \frac{1}{2}, & \text{dacă } 0 \le x \le 1 \\ 0, & \text{altfel .} \end{cases}$$

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)dx = \int_0^1 x\left(x+\frac{1}{2}\right)dx = \frac{7}{12}.$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)dx = \int_0^1 x^2\left(x+\frac{1}{2}\right)dx = \frac{5}{12}.$$

♣

▶ Matlab/Octave: `mean, var, std`
Fie $x = [x_1, \ldots, x_n]$ valorile unei v.a. X

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

$mean(x) \approx E(X)$ pentru n suficient de mare

$$var(x,1) = \frac{1}{n}\sum_{i=1}^{n}(x_i - mean(x))^2, \quad var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - mean(x))^2$$

$var(x,1) \approx V(X), var(x) \approx V(X)$ pentru n suficient de mare

$$std(x,1) = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - mean(x))^2\right)^{\frac{1}{2}}, std(x) = \left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - mean(x))^2\right)^{\frac{1}{2}}$$

$std(x,1) \approx Std(X), std(x) \approx Std(X)$ pentru n suficient de mare

Def. 24. $(X_n)_n$ *este* ***şir de v.a. independente****, dacă* $\forall\ \{i_1, \ldots, i_k\} \subset \mathbb{N}$ *v.a.* $X_{i_1}, \ldots, X_{i_k}$ *sunt independente, adică*

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

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

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

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

$\Rightarrow (X_n)_n$ şir de v.a. independente.
c) X_n= v.a. care indică numărul apărut la al n-lea joc de ruletă
$\Rightarrow (X_n)_n$ şir de v.a. independente. ♣

Def. 25. *Şirul de v.a. $(X_n)_n$* ***converge aproape sigur (a.s.)*** *la v.a. X, dacă*

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

Notaţie: $X_n \xrightarrow{\text{a.s.}} X$
► Cu alte cuvinte, convergenţa aproape sigură $X_n \overset{\text{a.s.}}{\to} X$ impune ca $\big(X_n(\omega)\big)_n$ să conveargă la $X(\omega)$ pentru fiecare $\omega \in \Omega$, *cu excepţia* unei mulţimi "mici" de probabilitate nulă;
dacă $X_n \xrightarrow{a.s.} X$ atunci evenimentul

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

Exemple: 1) În spaţiul de probabilitate $(\Omega, \mathcal{K}, P)$ fie $A \in \mathcal{K}$ cu $P(A) = 0.4$ şi $P(\bar{A}) = 0.6$:

$$X_n(\omega) = \begin{cases} 1 + \frac{1}{n}, & \text{pentru } \omega \in A \\ -\frac{1}{n}, & \text{pentru } \omega \in \bar{A}. \end{cases} \Longrightarrow P(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = ???\}) = 1.$$

Definim

$$X(\omega) = \begin{cases} 1, & \text{pentru } \omega \in A \\ 0, & \text{pentru } \omega \in \bar{A}. \end{cases} \Longrightarrow P(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = P(A) + P(\bar{A}) = 1.$$

Aşadar $X_n \xrightarrow{\text{a.s.}} X$.
2) Fie $\Omega := [0, 1]$ spaţiul de selecţie, P probabilitatea pe [0,1] indusă de măsura Lebesgue pe [0,1], adică pentru $\forall \alpha < \beta$ din $[0, 1]$ are loc

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

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

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

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

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

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

Frecvenţe relative şi absolute (a se vedea Def.2): Fie A un eveniment asociat unei experienţe, repetăm experienţa de n ori (în aceleaşi condiţii date) şi notăm cu r_n numărul de realizări ale evenimentului A; **frecvenţa relativă** a evenimentului A este numărul

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

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

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

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

$f_n(A) \xrightarrow{a.s.} \frac{1}{2}$ (a se vedea P.17)

Legea tare a numerelor mari (LTNM)

Legea numerelor mari se referă la descrierea rezultatelor unui experiment repetat de foarte multe ori. Conform acestei legi, rezultatul mediu obţinut se apropie tot mai mult de valoarea aşteptată, cu cât experimentul se repetă de mai multe ori. Aceasta se explică prin faptul că abaterile aleatoare se compensează reciproc.

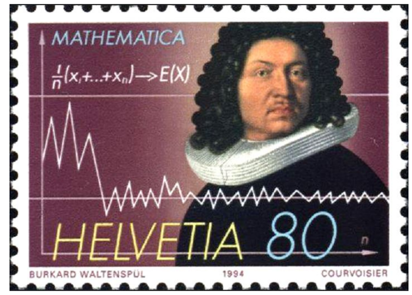


Fig. 5. Jacob Bernoulli (timbru emis în 1994 cu ocazia Congresului Internaţional al Matematicienilor din Elveţia)

Legea numerelor mari are două formulări: **legea slabă a numerelor mari (LSNM)** şi **legea tare a numerelor mari (LTNM)**.

▲ **Scurt istoric**: Jacob Bernoulli (1655 -1705) a formulat LSNM pentru frecvenţa relativă a unui experiment şi a dat răspunsul la întrebarea "*Putem aproxima empiric probabilităţile*?" (în opera publicată postum, în 1713, *Ars conjectandi*); ▷ Teorema lui Bernoulli afirmă: *"Frecvenţele relative converg în probabilitate la probabilitatea teoretică."*

Def. 26. *Şirul de v.a.* $(X_n)_n$ *cu* $E|X_n| < \infty\ \forall\ n \in \mathbb{N}$ *verifică* ***legea tare a numerelor mari (LTNM)*** *dacă*

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

adică

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

P. 16. *Fie* $(X_n)_n$ *şir de v.a. independente având aceeaşi distribuţie şi există* $m = E(X_n)\ \forall\, n \in \mathbb{N}$. $\Rightarrow (X_n)_n$ *verifică* ***LTNM****, adică*

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

În simulări: $\dfrac{1}{n}(X_1 + \cdots + X_n) \approx m,$ dacă n este suficient de mare.

Exemplu 1: Fie $X_1, ..., X_n, ... \sim Unid(6)$ v.a. independente; are loc $E(X_n) = \frac{1+2+3+4+5+6}{6} = 3.5\ \forall\ n \geq 1$. Folosind P.16 rezultă că $(X_n)_n$ verifică **LTNM**, adică $\dfrac{1}{n}(X_1 + ... + X_n) \xrightarrow{a.s.} 3.5$.

Simulare LTNM (Matlab/Octave):

```
pkg load statistics
clear all
close all

n=1000;
x=unidrnd(6,1,n);
for i=1:n
    s(i)=mean(x(1:i)); %media primelor i valori
end
fprintf('valoarea medie din simulari %5.3f\n', mean(x))
  % este egala cu s(n)
vmt=mean([1:6]);   % val medie teoretica in acest exemplu
fprintf('valoarea medie teoretica %5.3f\n',vmt)
figure
hold on
plot([1:n],vmt*ones(1,n),'g-')
plot([1:n],s,'r-')
plot([1:n],s,'b.')
xlabel('Nr. aruncari zar')
ylabel('Media numerelor aparute')
```

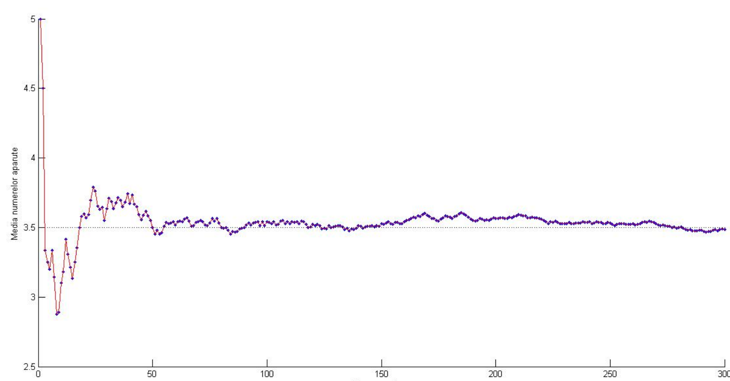


Fig. 4. Simulare LTNM

Exemplu 2: Fie $(X_n)_n$ şir de v.a. independente, având aceeaşi distribuţie ca v.a. X şi varianţă finită: $E(X_n) = E(X) \in \mathbb{R}$, $V(X_n) = V(X) \in \mathbb{R}$ pentru fiecare $n \in \mathbb{N}^*$.
Definim $Y_n = (X_n - E(X))^2 \; \forall \; n \in \mathbb{N}^* \Rightarrow (Y_n)_n$ este şir de v.a. independente, având aceeaşi distribuţie ca v.a. $(X - E(X))^2$ şi $E(Y_n) = E((X - E(X))^2) = V(X) \; \forall \; n \in \mathbb{N}^*$.
P.16 $\Rightarrow (Y_n)_n$ verifică **LTNM**

$$\frac{1}{n}\Big(Y_1 + ... + Y_n\Big) \xrightarrow{a.s.} V(X),$$

adică

$$\frac{1}{n}\Big((X_1 - E(X))^2 + ... + (X_n - E(X))^2\Big) \xrightarrow{a.s.} V(X).$$

Caz particular: Fie $X_1, ..., X_n, ... \sim Unid(6)$ v.a. independente; are loc $E(X_n) = \frac{1+2+3+4+5+6}{6} = \frac{7}{2} = 3.5$, $V(X_n) = E(X_n^2) - E^2(X_n) = \frac{35}{12} \approx 2.91666\ \forall\ n \geq 1$. Folosind P.16 rezultă că $(Y_n)_n = \big((X_n - 3.5)^2\big)_n$ verifică **LTNM**, adică $\dfrac{1}{n}\big((X_1 - 3.5)^2 + ... + (X_n - 3.5)^2\big) \xrightarrow{a.s.} \dfrac{35}{12}$.

```
pkg load statistics
clear all
close all
n=1000;
x=unidrnd(6,1,n);
for i=1:n
    z(i)=var(x(1:i),1); %varianta primelor i valori
end
fprintf('varianta din simulari %5.3f\n', var(x,1))
  % este egala cu z(n)
v=[1:6];
vt= mean(v.^2)-(mean(v))^2;
fprintf('varianta teoretica %5.3f\n', vt)
figure
hold on
plot([1:n],vt*ones(1,n),'g-')
plot([1:n],z,'r-')
plot([1:n],z,'b.')
xlabel('Nr. aruncari zar')
ylabel('Varianta numerelor aparute')
```

Exemplu 3: Fie $X_1, ..., X_n, ... \sim Unif[-1, 1]$ v.a. independente. Spre ce valoare converge a.s. şirul

$$Z_n = \frac{1}{n}(X_1^2 + ... + X_n^2),\ n \in \mathbb{N}^*\ \ ?$$

R.: Aplicăm P.16 pentru şirul de v.a. independente $(X_n^2)_n \Longrightarrow Z_n \xrightarrow{a.s.} E(X_1^2)$. Calculăm

$$E(X_1^2) = \int_{-1}^{1} t^2 \frac{1}{1-(-1)} dt = \frac{1}{2} \cdot \frac{t^3}{3}\Big|_{-1}^{1} = \frac{1}{3}.$$

$\Longrightarrow Z_n \xrightarrow{a.s.} \frac{1}{3}$. ▲

P. 17. *Fie A un eveniment asociat unei experienţe, repetăm experienţa de n ori (în aceleaşi condiţii date şi independent unele de altele). LTNM: cu cât repetăm mai des un experiment ($n \to \infty$), cu atât mai bine aproximează frecvenţa relativă $f_n(A)$ a evenimentului A probabilitatea sa teoretică de apariţie $P(A)$:*

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

În simulări: $f_n(A) \approx P(A),$ dacă n este suficient de mare.
Demonstraţie pentru P.17: Aplicăm P.16 pentru şirul de v.a. independente $(X_n)_n$, unde

$$X_n = \begin{cases} 1, & \text{dacă } A \text{ apare în a } n\text{- a execuţie a experimentului} \\ 0, & \text{dacă } \bar{A} \text{ apare în a } n\text{- a execuţie a experimentului} \end{cases}$$

$$\Longrightarrow X_n \sim \begin{pmatrix} 0 & 1 \\ 1-P(A) & P(A) \end{pmatrix} \Longrightarrow X_n \sim Bernoulli\big(P(A)\big)$$

$$\Longrightarrow E(X_n) = 0 \cdot (1-P(A)) + 1 \cdot P(A) = P(A) \ \ \forall n \in \mathbb{N}^*.$$

$$\text{P.16} \Longrightarrow \frac{1}{n}(X_1 + ... + X_n) \xrightarrow{a.s.} P(A).$$

Dar $\dfrac{1}{n}(X_1 + ... + X_n) = f_n(A)$ (frecvenţa relativă a lui A) $\Longrightarrow f_n(A) \xrightarrow{a.s.} P(A).$ $\square$

Statistică matematică

► Statistica matematică este o ramură a matematicii aplicate, care se ocupă de *colectarea*, *gruparea*, *analiza* şi *interpretarea datelor* referitoare la anumite fenomene în scopul obţinerii unor previziuni;
• statistica descriptivă: metode de colectare, organizare, sintetizare, prezentare şi descriere a datelor numerice (sau nenumerice) într-o formă convenabilă
• statistica inferenţială: metode de interpretare a rezultatelor obţinute prin metodele statisticii descriptive, utilizate apoi pentru luarea deciziilor.
► O *colectivitate* sau *populaţie statistică* $\mathcal{C}$ este o mulţime de elemente care au anumite însuşiri comune ce fac obiectul analizei statistice. Numărul elementelor populaţiei se numeşte *volumul populaţiei.*
Exemple de populaţii statistice: mulţimea persoanelor dintr-o anumită ţară, localitate, zonă etc. într-un anumit an; multimea gospodăriilor din Romania la un moment dat; mulţimea consumatorilor unui anumit produs; mulţimea societăţilor care produc un anumit produs; angajaţii unei societăţi; studenţii unei facultăţi.
► *Eşantionul* $\mathcal{E}$ reprezintă o submulţime a unei populaţii statistice $\mathcal{E} \subset \mathcal{C}$, constituită după criterii bine stabilite:
a) să fie aleatoare;
b) toate elementele colectivităţii să aibe aceeaşi şansă de a fi alese în eşantion;
c) eşantionul să fie reprezentativ (structura eşantionului să fie apropiată de structura populaţiei);
d) volumul eşantionului să fie suficient de mare.
► *Unitatea statistică* (indivizii) este elementul, entitatea de sine stătătoare a unei populaţii statistice, care posedă o serie de trăsături caracteristice ce-i conferă apartenenţa la populaţia studiată. De exemplu: *unitatea statistică simplă*: un salariat, un student, un agent economic, o trăsătură, o părere; *unitatea statistică complexă*: o grupă de studenţi sau o echipă de salariaţi, o familie sau o gospodărie, o categorie de mărfuri.
► *Variabila statistică* sau *caracteristica* reprezintă o însuşire, o proprietate măsurabilă a unei unităţi statistice, întâlnită la toate unităţile care aparţin aceleiaşi colectivităţi şi care prezintă variabilitate de la o unitate statistică la alta. Caracteristica sau variabila statististică corespunde unei variabile aleatoare.
Exemple de caracteristici: vârsta, salariul, preferinţele politice, preţul unui produs, calitatea unor servicii, nivelul de studii.
a) variabile (caracteristici) continue $\rightarrow$ iau un număr infinit şi nenumărabil de valori într-un interval sau reuniune de intervale (de ex.: greutatea, înălţimea, valoarea glicemiei, temperatura aerului)

b) variabile (caracteristici) discrete $\rightarrow$ iau număr finit sau infinit dar numărabil de valori discrete (de ex.: numări elevi ai unei şcoli, numărul liceelor existente într-un oraş, valoarea IQ)
$\triangleright$ caracteristicile de la a) şi 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)
$\triangleright$ caracteristicile de la c),d),e) sunt variabile calitative
$\triangleright$ variabilele nominale mai sunt numite variabile categoriale
$\blacktriangleright$ *Datele statistice* reprezintă observaţiile rezultate dintr-o cercetare statistică, sau ansamblul valorilor colectate în urma unei cercetări statistice.
De exemplu: un angajat al unei companii are o vechime de 6 ani în muncă. Angajatul reprezintă unitatea statistică, vechimea în muncă este caracteristica (variabila) cercetată, iar 6 este valoarea acestei caracteristici.

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

Distribuţia caracteristicii X poate fi

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

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

• în cazul 2) parametrii sunt necunoscuţi, iar în cazul 3) distribuţia este necunoscută
$\hookrightarrow$ se estimează $\rightarrow$ teoria estimaţiei / intervale de încredere
$\hookrightarrow$ se testează $\rightarrow$ teste statistice

$\blacktriangleright$ Fie $\mathcal{E} \subset \mathcal{C}$ un eşantion. Se numesc date de selecţie relative la caracteristica X datele statistice $x_1, \ldots, x_n$ obţinute prin cercetarea indivizilor care fac parte din eşantionul $\mathcal{E}$.
$\blacktriangleright$ Datele de selecţie $x_1, \ldots, x_n$ pot fi considerate ca fiind valorile unor variabile aleatoare $X_1, \ldots, X_n$, numite variabile de selecţie şi care se consideră a fi variabile aleatoare independente şi având aceeaşi distribuţie ca X.
$\blacktriangleright$ Fie $x_1, \ldots, x_n$ datele statistice pentru caracteristica cercetată X, notăm cu $X_1, \ldots, X_n$ variabilele de selecţie corespunzătoare. Fie $g : \mathbb{R}^n \rightarrow \mathbb{R}$ o funcţie astfel încât $g(X_1, \ldots, X_n)$ este o variabilă aleatoare.
$g(X_1, \ldots, X_n)$ se numeşte funcţie de selecţie sau estimator

$g(x_1, \ldots, x_n)$ se numeşte valoarea funcţiei de selecţie sau valoarea estimatorului.

• **Exemple de estimatori (funcţii de selecţie)** sunt: media de selecţie, dispersia de selecţie, momentul centrat de selecţie de ordinul doi, funcţia de repartiţie empirică.

▷ Estimatorii (funcţiile de selecţie) se folosesc în statistică pentru estimarea punctuală a unor parametri necunoscuţi, pentru obţinerea unor intervale de încredere pentru parametri necunoscuţi, pentru verificarea unor ipoteze statistice.

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

▶ media de selecţie (empirică)

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

▶ valoarea mediei de selecţie

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

▶ varianţa (dispersia) de selecţie (empirică)

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

▶ valoarea varianţei (dispersiei) de selecţie

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

▶ abaterea standard de selecţie (empirică)

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

▶ valoarea abaterii standard de selecţie

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

▶ momentul centrat de selecţie (empiric) de ordinul doi

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

▶ valoarea momentului centrat de selecţie (empiric) de ordinul doi

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

▶ funcţia de repartiţie empirică $\mathcal{F}_n : \mathbb{R} \times \Omega \to [0,1]$

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

▶ valoarea (expresia) funcţiei de repartiţie empirice $\mathcal{F}_n : \mathbb{R} \to [0,1]$

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

Def. 27. $g(X_1,\ldots,X_n)$ *este estimator nedeplasat pentru parametrul necunoscut* θ*, dacă*

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

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

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

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

Observaţii:

1) Media de selecţie $\bar{X}_n$ este un estimator nedeplasat şi consistent pentru media teoretică $E(X)$ a caracteristicii X; în simulări $E(X) \approx \bar{x}_n$.
În Octave: `mean(d)`, unde `d` este vectorul datelor statistice.
2) Varianţa de selecţie S_n^2 este un estimator nedeplasat şi consistent pentru varianţa teoretică $V(X)$ a caracteristicii X; în simulări $V(X) \approx s_n^2$.
În Octave: `var(d)`, unde `d` este vectorul datelor statistice.
2*) Momentul centrat de selecţie de ordinul doi M_n *nu* este un estimator nedeplasat pentru varianţa teoretică $V(X)$ a caracteristicii X; el este un estimator consistent pentru varianţa teoretică $V(X)$ a caracteristicii X; în simulări se foloseşte şi $V(X) \approx m_n$.
În Octave: `var(d,1)`, unde `d` este vectorul datelor statistice.
3) Deviaţia standard de selecţie S_n nu este un estimator nedeplasat pentru deviaţia standard teoretică $Std(X) = \sqrt{V(X)}$ a caracteristicii X; el este un estimator consistent pentru deviaţia standard teoretică $Std(X)$ a caracteristicii X; în simulări se foloseşte $Std(X) \approx s_n$.
În Octave: `std(d)`, unde `d` este vectorul datelor statistice.

4) Funcţia de repartiţie de selecţie $\mathcal{F}_n(x,\cdot)$ calculată în $x \in \mathbb{R}$ este un estimator nedeplasat şi consistent pentru $F_X(x)$, care este valoarea funcţiei de repartiţie teoretice calculată în x; în simulări $F_X(x) \approx \mathcal{F}_n(x)$.
În Octave: `empirical_cdf(`x`,d)` $= \mathcal{F}_n(x)$, unde `d` este vectorul datelor statistice şi `length(d)=`n.

```
pkg load statistics  % exemple de estimatori
clear all
close all
d=randsample([4:10],400,1);
% note (la o anumita materie) in clasa a X-a intr-un anumit oras
% extragere cu repetitie (de 400 de ori) din vectorul [4,5,6,7,8,9,10]
% distributia teoretica X: P(X=k)=1/7 pentru k in {4,5,6,7,8,9,10}
note=[4:10];
 m=mean(d)  % valoarea mediei de selectie
 m_teor=mean(note) %media teoretica E(X)
 v=var(d)   % valoarea variantei de selectie
 v1=var(d,1) % valoarea momentului centrat  de selectie de ordinul 2
 v1_teor=var(note,1) %varianta teoretica V(X)
                    % sau altfel:   mean(note.^2)-mean(note)^2
 st=std(d) % valoarea deviatiei standard de selectie
 st1=std(d,1)
 st1_teor=std(note,1) %deviatia standard teoretica Std(X)=sqrt(V(X))
figure(1)
hold on
 x=4:0.01:10;
 y=empirical_cdf(x,d); %valoarea functiei de repartitie de selectie
 plot(x,y,'r*') % graficul functiei de repartitie de selectie
 y_teor=empirical_cdf(x,note); %valoarea functiei de repartitie teoretice
 plot(x,y_teor,'b*')
 legend('F. de repartitie de selectie', 'F. de repartitie teoretica')
 title('FUNCTIA DE REPARTITIE EMPIRICA / TEORETICA')
figure(2)
 h=hist(d,[4:10])
 bar([4:10],h/length(d),'hist');
 title('HISTOGRAMA FRECVENTELOR RELATIVE')
figure(3)
 bar([4:10],h,'hist');
 title('HISTOGRAMA FRECVENTELOR ABSOLUTE')
```

Exemplu: Fie $(X_n)_n$ şirul variabilelor de selecţie pentru caracteristica cercetată $X \sim Bernoulli(p)$, unde $p \in (0,1)$ este parametru necunoscut. Estimatorul

$$\hat{p}(X_1,...,X_n) = \frac{1}{n}(X_1 + ... + X_n) = \bar{X}_n \text{ (media de selecţie)}$$

este un estimator *nedeplasat* şi *consistent* pentru parametrul necunoscut p.
R.: $X \sim Bernoulli(p) \Longrightarrow E(X) = p$;

$$\Longrightarrow E\Big(\hat{p}(X_1, ..., X_n)\Big) = \frac{1}{n}\Big(E(X_1) + ... + E(X_n)\Big) = E(X) = p.$$

LTNM (a se vedea P.16) implică

$$\hat{p}(X_1, ..., X_n) = \frac{1}{n}(X_1 + ... + X_n) \xrightarrow{a.s.} p.$$

Deci, $\hat{p}(X_1, ..., X_n)$ este un estimator nedeplasat şi consistent pentru parametrul necunoscut p.
Dacă $x_1, \ldots, x_n \in \{0, 1\}$ sunt date statistice, atunci valoarea estimată pentru p este

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

$\diamondsuit$

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

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

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

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

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

Exemplu 1: 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.

R.: Fie $X_1, \ldots, X_n$ variabilele de selecţie. Avem cazul: $r = 1$, calculăm $E(X) = \frac{a}{2}$, $n = 10$, $\bar{x}_n = 0.468$. Se rezolvă

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

Estimatorul pentru parametrul necunoscut a este

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

Valoarea estimatorului este

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

Parametrul necunoscut a este estimat cu valoarea 0.936.

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

R.: Da, se arată că $E(\hat{a}(X_1, ..., X_n)) = a$. ♡

Exemplu 2:

Folosind metoda momentelor, să se estimeze parametrii necunoscuţi $\theta_1 := m$ şi $\theta_2 = \sigma^2$ pentru $X \sim N(m, \sigma^2)$; se dau datele statistice:

$$0.831, 0.71, -0.2, -0.04, 2.08, -1.2, 0.448, -0.18, -0.27, -0.55\,.$$

R.: Fie $n = 10$, iar $X_1, ..., X_n$ variabile de selecţie. Avem cazul: $r = 2$, calculăm $E(X) = m$, $E(X^2) = V(X) + E^2(X) = \sigma^2 + m^2$ (a se vedea exemplul de pe pg. 37), $\bar{x}_n = 0.1629$ (calculat în Octave cu `mean(x)`, unde `x` este vectorul datelor statistice), $m_n = 0.7346$ (calculat în Octave cu `var(x,1)`). Se rezolvă

$$\begin{cases} m & = \frac{1}{n}\sum_{i=1}^{n} X_i \\ \sigma^2 + m^2 & = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{cases} \Longrightarrow \text{are soluţia} \begin{cases} \hat{m} & = \frac{1}{n}\sum_{i=1}^{n} X_i \\ \hat{\sigma}^2 & = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \end{cases}$$

Estimatorii sunt

$$\hat{m}(X_1, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n \text{ (media de selecţie)},$$

$$\hat{\sigma}^2(X_1,...,X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = M_n \text{ (momentul centrat de selecţie de ordinul doi)}$$

Valorile estimatorilor sunt

$$\hat{m}(x_1,...,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}_n = 0.1629\,,$$

$$\hat{\sigma}^2(x_1,...,x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2 = m_n = 0.7346\,.$$

♣

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

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

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

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

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

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

Se rezolvă sistemul $\frac{\partial L}{\partial \theta} = 0$ şi se arată că $\frac{\partial^2 L}{\partial \theta^2} < 0$.

Deseori este mai practic să se considere varianta transformată

$\frac{\partial lnL}{\partial \theta} = 0$ cu $\frac{\partial^2 lnL}{\partial \theta^2} < 0$. În unele situaţii (1) se rezolvă prin alte metode.

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

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

Se poate lucra şi cu varianta transformată:

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

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

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

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

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

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

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

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

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

Intervale de încredere şi teste statistice

Noţiuni de bază

► Fie $\alpha \in (0,1)$ nivelul de semnificaţie (probabilitatea de risc).

Def. 28. ***Cuantila de ordin*** *α pentru distribuţia caracteristicii cercetate X este numărul $z_\alpha \in \mathbb{R}$ pentru care*

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

*Dacă $\alpha = 0.5$ atunci $z_{0.5}$ se numeşte **mediană**.*

▶ dacă X este v.a. continuă, atunci: z_α este cuantilă de ordin $\alpha \iff P(X \le z_\alpha) = \alpha \iff F_X(z_\alpha) = \alpha$
▶ dacă F_X este funcţie inversabilă, atunci $z_\alpha = F_X^{-1}(\alpha)$

• $\alpha \cdot 100\%$ din valorile lui X sunt mai mici sau egale cu z_α
• Matlab/Octave: `quantile`

Distribuţii de probabilitate continue frecvent folosite în statistică și cuantilele lor corespunzătoare

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

• Pentru cuantilele distribuţiei normale $N(0,1)$ are loc $z_\alpha = -z_{1-\alpha}$ pentru orice $\alpha \in (0,1)$;
• pentru cuantilele distribuţiei Student $St(n)$ are loc $t_\alpha = -t_{1-\alpha}$ pentru orice $\alpha \in (0,1)$.

Exemplu: Să se arate că: **a)** $X \sim N(0,1) \iff -X \sim N(0,1)$;
b) pentru cuantilele distribuţiei normale $N(0,1)$ are loc $z_\alpha = -z_{1-\alpha}$ pentru orice $\alpha \in (0,1)$;
c) proprietatea analoagă are loc și pentru distribuţia Student $St(n)$, adică $t_\alpha = -t_{1-\alpha}$ pentru orice $\alpha \in (0,1)$.
R.: a) Fie $x \sim N(0,1)$. Scriem pentru orice $u \in \mathbb{R}$

$$F_{-X}(u) = P(-X \le u) = P(X > -u) = 1 - P(X \le -u) = 1 - F_X(-u).$$

Aceasta implică

$$f_{-X}(u) = F'_{-X}(u) = F'_X(-u) = f_X(-u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}},\ \forall u \in \mathbb{R}.$$

Deci $-X \sim N(0,1)$. Folosind rezultatul deja demonstrat şi relaţia $X = -(-X)$, obţinem că $-X \sim N(0,1) \Longrightarrow X \sim N(0,1)$.
b) Fie $X \sim N(0,1)$ şi $z_\alpha, z_{1-\alpha}$ cuantile ale sale. Rezultă că

$$P(X \leq z_\alpha) = \alpha,\ P(X \leq z_{1-\alpha}) = 1-\alpha.$$

Scriem şi folosim faptul că $-X$ şi X urmează distribuţia $N(0,1)$

$$P(X \leq z_\alpha) = \alpha = 1 - P(X \leq z_{1-\alpha}) = P(X > z_{1-\alpha}) = P(-X < -z_{1-\alpha}) = P(X < -z_{1-\alpha})$$
$$= P(X \leq -z_{1-\alpha}).$$

Pentru distribuţia $N(0,1)$ cuantila z_α e unic determinată din relaţia $P(X \leq z_\alpha) = \alpha$ (pentru că F_X e o funcţie inversabilă şi atunci $z_\alpha = F_X^{-1}(\alpha)$), aşadar obţinem că $z_\alpha = -z_{1-\alpha}$.
c) Raţionamentul este analog. Se foloseşte $X \sim St(n) \Longleftrightarrow -X \sim St(n)$. ♣

Intervale de încredere

În paragrafele anterioare s-a văzut cum poate fi estimat un parametru necunoscut, folosind datele dintr-un eşantion. Se pune problema cât este de bună această estimare a parametrului necunoscut, adică vom calcula o anumită ”marjă de eroare”.

Presupunem că studiem media (teoretică) a timpului de aşteptare la un anumit ghişeu al unei bănci. Prin studierea unui eşantion de volum 200 s-a constatat că media de seleţie a timpului de aşteptare este $\bar{x}_{200} = 10$ (minute). Dacă considerăm un alt eşantion probabil obţinem o altă valoare pentru $\bar{x}_{200}$.

Problemă: putem construi un interval (aleator) care să acopere valoarea reală a parametrului necunoscut studiat cu o anumită probabilitate dată (numit nivel de încredere)?
Pe baza datelor din eşantion acest interval aleator va deveni un interval numeric.

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

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

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

► $\Big(g_1(X_1, \ldots, X_n), g_2(X_1, \ldots, X_n)\Big)$ se numeşte **interval de încredere bilateral pentru parametrul necunoscut** θ

► $\Big(g_1(x_1,\dots,x_n), g_2(x_1,\dots,x_n)\Big)$ este **valoarea intervalului de încredere** pentru parametrul necunoscut θ

► $g_1(X_1,\dots,X_n)$ 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 θ să fie în intervalul $\Big(g_1(X_1,\dots,X_n), g_2(X_1,\dots,X_n)\Big)$ este $1-\alpha$ (nivelul de încredere)

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

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

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

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

➡ Nu este corect să afirmăm că probabilitatea ca intervalul numeric construit (din datele statistice) să cuprindă valoarea reală a lui θ este $1-\alpha$. Intervalul de încredere este un interval aleator, deci extremităţile sale sunt v.a. Prin urmare interpretarea corectă a lui $1-\alpha$ este următoarea: dacă, facem un număr foarte mare de selecţii (din mai multe eşantioane) şi calculăm de fiecare dată intervalul de încredere cu nivelul de încredere $1-\alpha$, atunci $(1-\alpha)\cdot 100\%$ din aceste intervale vor conţine valoarea reală pentru θ.

P. 18. (Teorema limită centrală) *Fie $(X_n)_n$ un şir de v.a. independente, care au aceeaşi distribuţie. Fie $m = E(X_n)$ şi $\sigma^2 = V(X_n) > 0\ \forall\, n \geq 1$. Are loc*

$$\lim_{n\to\infty} P\left(\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \leq b\right) = F_{N(0,1)}(b) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{b} e^{-\frac{t^2}{2}}dt,$$

pentru orice $b \in \mathbb{R}$, iar $\bar{X}_n = \frac{1}{n}(X_1+\dots+X_n)$.

➡ $F_{N(0,1)}(b)$=`normcdf`$(b,0,1)$ funcţia de repartiţie a legii normale standard $N(0,1)$

➡ **Consecinţă** (la P. 18): pentru orice $a<b$ are loc

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

Exemplul 1: Dacă $(X_n)_{1\leq n\leq 100}$ sunt variabile de selecţie pentru caracteristica cercetată $X \sim Bernoulli(0.5)$, să se estimeze $P(0.35 < \bar{X}_{100} < 0.65)$!
R.: Se calculează $m = E(X_n) = 0.5$ şi $\sigma = \sqrt{V(X_n)} = 0.5$ şi se scrie

$$P(0.35 < \bar{X}_{100} < 0.65) = P\left(-3 < \frac{\bar{X}_{100} - 0.5}{\frac{0.5}{\sqrt{100}}} < 3\right).$$

Cf. P. 18 şi a consecinţei de mai sus

$$\Longrightarrow P\left(-3 < \frac{\bar{X}_{100} - 0.5}{\frac{0.5}{\sqrt{100}}} < 3\right) \approx \texttt{normcdf}(3,0,1) - \texttt{normcdf}(-3,0,1) = 0.9973$$

$$\Longrightarrow P\Big(\bar{X}_{100} \in (0.35, 0.65)\Big) \approx 0.9973,$$

aşadar pentru o caracteristică de tip $Bernoulli(0.5)$, media de selecţie $\bar{X}_{100}$ aparţine cu o probabilitate foarte mare intervalului $(0.35, 0.65)$. ♡

Exemplul 2: Se ştie că 40% din populaţia unui orăşel susţine un anumit candidat la alegerile viitoare. Dacă $(X_n)_{1\leq n\leq 600}$ sunt variabile de selecţie pentru distribuţia $Bernoulli(0.4)$, adică $\forall\, n \in \{1, ..., 600\}$

$X_n = 1 \Longleftrightarrow$ persoana a n-a votează acest candidat,

$X_n = 0 \Longleftrightarrow$ persoana a n-a nu votează acest candidat,

şi $X_n \sim Bernoulli(0.4)$. Estimaţi $P(\bar{X}_{600} > 0.43)$, calculaţi $E(\bar{X}_{600})$ şi $V(\bar{X}_{600})$.
R.: Dacă $(X_n)_{1\leq n\leq 600}$ sunt variabile de selecţie pentru $Bernoulli(0.4)$, se calculează $m = E(X_n) = 0.4$ şi $\sigma^2 = V(X_n) = 0.24\ \forall\, n \in \mathbb{N}^*$ şi se doreşte estimarea probabilităţii

$$P(\bar{X}_{600} > 0.43) = 1 - P(\bar{X}_{600} \leq 0.43).$$

Cf. P. 18

$$\Longrightarrow P(\bar{X}_{600} \leq 0.43) = P\left(\frac{\bar{X}_{600} - 0.4}{\sqrt{\frac{0.24}{600}}} \leq \frac{0.43 - 0.4}{\sqrt{\frac{0.24}{600}}}\right) = P\left(\frac{\bar{X}_{600} - 0.43}{\sqrt{\frac{0.24}{600}}} \leq 1.5\right)$$

$$\approx F_{N(0,1)}(1.5) = \texttt{normcdf}(1.5, 0, 1) = 0.93319$$

$\Longrightarrow P(\bar{X}_{600} > 0.43) \approx 0.066807.$
$E(\bar{X}_{600}) = \frac{1}{600}\Big(E(X_1) + ...E(X_{600})\Big) = 0.4$ şi
$V(\bar{X}_{600}) = \frac{1}{600^2}\Big(V(X_1) + ... + V(X_{600})\Big) = \frac{1}{600} \cdot 0.24 = 0.0004.$ ♠

Exerciţiu: 100 de zaruri sunt aruncate. Folosind P.18 (Teorema limită centrală), estimaţi probabilitatea ca suma numerelor obţinute să fie între 300 şi 400!

P. 19. *Fie $X_1, \dots, X_n$ variabile de selecţie pentru $X \sim N(m, \sigma^2)$, atunci pentru* ***media de selecţie*** *are loc* $\dfrac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$.

Reamintim: $X \sim N(m, \sigma^2) \Longrightarrow E(X) = m, V(X) = \sigma^2$ (a se vedea calculele de pe pg. 37).

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

▶ se dau $\alpha \in (0,1)$, σ, datele statistice $x_1, \dots, x_n$
▶ construim intervale de încredere pentru parametrul *necunoscut* $m = E(X)$
▶ dacă $X \sim N(m, \sigma^2)$ sau $n > 30$ şi X are o distribuţie necunoscută, atunci P. 18 şi P. 19 implică

$$\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \tag{5}$$

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

• un *interval de încredere bilateral* pentru $m = E(X)$ (media teoretică) când dispersia este cunoscută este

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

deoarece:

$$P\Big(\bar{X}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} < m < \bar{X}_n + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\Big) = P\Big(-z_{1-\frac{\alpha}{2}} < \frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}}\Big)$$
$$= F_{N(0,1)}(z_{1-\frac{\alpha}{2}}) - F_{N(0,1)}(-z_{1-\frac{\alpha}{2}}) \overset{(5)}{=} F_{N(0,1)}(z_{1-\frac{\alpha}{2}}) - F_{N(0,1)}(z_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha$$

• *intervale de încredere unilaterale*: $\Big(-\infty, \bar{X}_n - \dfrac{\sigma}{\sqrt{n}} \cdot z_\alpha\Big), \Big(\bar{X}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}, \infty\Big)$, adică

$$P\Big(m < \bar{X}_n - \frac{\sigma}{\sqrt{n}} \cdot z_\alpha\Big) = 1 - \alpha,\ \ P\Big(\bar{X}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha} < m\Big) = 1 - \alpha\,.$$

Interval de încredere pentru **media teoretică** când dispersia σ^2 este cunoscută:	Expresia intervalului de încredere, folosind datele statistice
bilateral	$\left(\bar{x}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\ ,\ \bar{x}_n + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\right)$
unilateral	$\left(-\infty\, , \bar{x}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{\alpha}\right)$
	$\left(\bar{x}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}\ ,\ \infty\right)$

Exemplu: Un profesor a înregistrat pe parcursul mai multor ani rezultatele elevilor săi la un anumit tip de test. Punctajul unui elev este o v.a. $X \in (0, 100)$, având abaterea standard egală cu 10. Media de selecţie a calificativelor a 144 de elevi este 68. Dacă $\alpha = 0.05$, să se construiască un interval de încredere bilateral pentru valoarea medie $E(X)$ a punctajului obţinut de un elev la un anumit test.
R:

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

unde $n = 144, \sigma = 10, \bar{x}_n = 68, \alpha = 0.05, z_{1-\frac{\alpha}{2}} = \texttt{norminv}(1 - \frac{0.05}{2}, 0, 1) \approx 1.96$. Valoarea intervalului de încredere bilateral este $(66.367, 69.633)$. ♣

P. 20. *Fie $X_1, \ldots, X_n$ variabile de selecţie pentru $X \sim N(m, \sigma^2)$, atunci pentru* ***media de selecţie şi abaterea standard de selecţie*** *are loc* $\dfrac{\bar{X}_n - m}{\frac{S_n}{\sqrt{n}}} \sim St(n-1)$.

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

▶ se dau $\alpha \in (0, 1)$, datele statistice $x_1, \ldots, x_n$
▶ dacă $X \sim N(m, \sigma^2)$ sau $n > 30$ şi X are o distribuţie necunoscută, atunci P.20 implică

$$\frac{\bar{X}_n - m}{\frac{S_n}{\sqrt{n}}} \sim St(n-1)$$

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

• un *interval de încredere bilateral* pentru $m = E(X)$ (media teoretică), când dispersia este necunoscută este: $\Big(\bar{X}_n - \dfrac{S_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\ ,\ \bar{X}_n + \dfrac{S_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\Big)$, adică

$$P\Big(\bar{X}_n - \frac{S_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}} < m < \bar{X}_n + \frac{S_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\Big) = 1 - \alpha$$

• *intervale de încredere unilaterale* $\Big(-\infty, \bar{X}_n - \dfrac{S_n}{\sqrt{n}} \cdot t_\alpha\Big)$, $\Big(\bar{X}_n - \dfrac{S_n}{\sqrt{n}} \cdot t_{1-\alpha}\ ,\ \infty\Big)$, adică

$$P\Big(m < \bar{X}_n - \frac{S_n}{\sqrt{n}} \cdot t_\alpha\Big) = 1 - \alpha, \quad P\Big(\bar{X}_n - \frac{S_n}{\sqrt{n}} \cdot t_{1-\alpha} < m\Big) = 1 - \alpha$$

Interval de încredere pentru **media teoretică** când dispersia este necunoscută	Expresia intervalului de încredere, folosind datele statistice
bilateral	$\Big(\bar{x}_n - \dfrac{s_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\ ,\ \bar{x}_n + \dfrac{s_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\Big)$
unilateral	$\Big(-\infty\ ,\bar{x}_n - \dfrac{s_n}{\sqrt{n}} \cdot t_\alpha\Big)$
	$\Big(\bar{x}_n - \dfrac{s_n}{\sqrt{n}} \cdot t_{1-\alpha}\ ,\ \infty\Big)$

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

$$\Big(\bar{x}_n - \frac{s_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}, \bar{x}_n + \frac{s_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\Big)$$

unde $\bar{x}_n = 15.5, s_n = 0.3\ (s_n^2 = 0.09), \alpha = 0.01, t_{1-\frac{\alpha}{2}} = \texttt{tinv}(0.995, 99) = 2.6264, \sqrt{n} = 10$.
Valoarea intervalului de încredere bilateral este $\big(15.421208, 15.578792\big)$. ♣

P. 21. *Fie $X_1, \ldots, X_n$ variabile de selecţie pentru $X \sim N(m, \sigma^2)$, atunci pentru **varianţa de selecţie** are loc $\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2(n-1)$.*

Exemplu: Timpul necesar unei unităţi CPU pentru a realiza un anumit tip de operaţii are distribuţie normală cu media 20 de secunde şi abaterea standard 3 secunde. Într-un eşantion de 25 de astfel de operaţii, care este probabilitatea ca varianţa de selecţie (a timpului necesar tipului de operaţii

studiate) să depăşească 12 secunde?
R: Vom folosi P.21. Scriem succesiv

$$P(S_{25}^2 > 12) = P\Big(\frac{25-1}{3^2}S_{25}^2 > \frac{25-1}{3^2}\cdot 12\Big) = 1 - P\Big(\frac{24}{9}S_{25}^2 \le 32\Big).$$

Dar $\frac{24}{9}S_{25}^2 \sim \chi^2(25-1)$ (cf. P.21)

$$\Longrightarrow P(S_{25}^2 > 12) = 1 - F_{\chi^2(24)}(32) = 1 - \texttt{chi2cdf}(32, 24) \approx 1 - 0.87301 = 0.12699\,.$$

■

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

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

▶ dacă $X \sim N(m, \sigma^2)$, atunci P.21 implică $\frac{n-1}{\sigma^2}S_n^2 \sim \chi^2(n-1)$

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

• un *interval de încredere bilateral* pentru varianţa teoretică $\sigma^2 = V(X)$ este: $\Big(\frac{n-1}{c_{1-\frac{\alpha}{2}}}\cdot S_n^2, \frac{n-1}{c_{\frac{\alpha}{2}}}\cdot S_n^2\Big)$, adică

$$P\Big(\frac{n-1}{c_{1-\frac{\alpha}{2}}}\cdot S_n^2 < \sigma^2 < \frac{n-1}{c_{\frac{\alpha}{2}}}\cdot S_n^2\Big) = 1-\alpha$$

• *intervale de încredere unilaterale*: $\Big(0, \frac{n-1}{c_\alpha}\cdot S_n^2\Big)$, $\Big(\frac{n-1}{c_{1-\alpha}}\cdot S_n^2\,,\ \infty\Big)$, adică

$$P\Big(\sigma^2 < \frac{n-1}{c_\alpha}\cdot S_n^2\Big) = 1-\alpha, \quad P\Big(\frac{n-1}{c_{1-\alpha}}\cdot S_n^2 < \sigma^2\Big) = 1-\alpha.$$

Interval de încredere pentru **varianţa (dispersia) teoretică $V(X)$**	Expresia intervalului de încredere, folosind datele statistice
bilateral	$\Big(\frac{n-1}{c_{1-\frac{\alpha}{2}}}\cdot s_n^2, \frac{n-1}{c_{\frac{\alpha}{2}}}\cdot s_n^2\Big)$
unilateral	$\Big(0\,, \frac{n-1}{c_\alpha}\cdot s_n^2\Big)$
	$\Big(\frac{n-1}{c_{1-\alpha}}\cdot s_n^2\,,\ \infty\Big)$

Interval de încredere pentru **abaterea standard teoretică** $Std(X)$	Expresia intervalului de încredere, folosind datele statistice
bilateral	$\left(\sqrt{\frac{n-1}{c_{1-\frac{\alpha}{2}}}}\cdot s_n, \sqrt{\frac{n-1}{c_{\frac{\alpha}{2}}}}\cdot s_n\right)$
unilateral	$\left(0\,,\sqrt{\frac{n-1}{c_{\alpha}}}\cdot s_n\right)$
	$\left(\sqrt{\frac{n-1}{c_{1-\alpha}}}\cdot s_n\,,\,\infty\right)$

Exemplu: Durata de funcţionare a unui anumit tip de baterie este 500 de ore. Pe baza unui eşantion s-au testat 64 de baterii şi s-a obţinut media de 525 de ore şi abaterea standard de 25 de ore. Să se construiască un interval de încredere 99%
a) bilateral pentru media (teoretică);
b) unilateral pentru abaterea standard teoretică (care are marginea inferioară 0 şi se cere să se calculeze marginea superioară)
a duratei de funcţionare a acestui tip de baterii (se presupune că durata de funcţionare a acestui tip de baterie urmează distribuţia normală).
R.: a) Valoarea intervalului de încredere bilateral pentru media teoretică, când varianţa este necunoscută, este

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

cu $\sqrt{n}=8, \bar{x}_n=525, s_n=25, \alpha=0.01, t_{1-\frac{\alpha}{2}}=\texttt{tinv}(0.995,63)=2.6561 \Longrightarrow$ valoarea intervalului de încredere bilateral pentru medie este $(516.7, 533.3)$.
b) Expresia intervalului de încredere unilateral pentru abaterea standard (teoretică) este $\left(0,\sqrt{\frac{n-1}{c_{\alpha}}}\cdot s_n\right)$, cu $n=64, s_n=25, \alpha=0.01, c_\alpha=\texttt{chi2inv}(0.01,63)=39.8551 \Longrightarrow$ valoarea intervalului de încredere unilateral pentru abaterea standard este $(0, 31.432)$. ♣

Interval de încredere pentru proporţia necunoscută p, a caracteristicii cercetate $X\sim Bernoulli(p)$

▶ se dau $\alpha\in(0,1)$, datele statistice $x_1,\dots,x_n\in\{0,1\}$
▶ construim intervale de încredere pentru parametrul *necunoscut* $p\in(0,1)$
▶ dacă $X\sim Bernoulli(p)$, atunci P. 18 implică $\dfrac{\bar{X}_n-p}{\sqrt{\frac{p(1-p)}{n}}}\sim N(0,1)$ pentru n suficient de mare
▶ cuantilele legii normale $N(0,1)$:

$z_{1-\frac{\alpha}{2}} = \texttt{norminv}(1-\frac{\alpha}{2}, 0, 1)$, $z_{1-\alpha} = \texttt{norminv}(1-\alpha, 0, 1)$, $z_{\alpha} = \texttt{norminv}(\alpha, 0, 1)$

• *intervalul de încredere bilateral* pentru p:

$$\Big(\bar{X}_n - \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \cdot z_{1-\frac{\alpha}{2}}\ ,\ \bar{X}_n + \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \cdot z_{1-\frac{\alpha}{2}}\Big),$$

adică

$$P\Big(\bar{X}_n - \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \cdot z_{1-\frac{\alpha}{2}} < p < \bar{X}_n + \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \cdot z_{1-\frac{\alpha}{2}}\Big) = 1-\alpha$$

• *intervale de încredere unilaterale*: $\Big(0\, , \bar{X}_n - \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \cdot z_{\alpha}\Big)$,
$\Big(\bar{X}_n - \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \cdot z_{1-\alpha}\, ,\ 1\Big)$, adică

$$P\Big(p < \bar{X}_n - \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \cdot z_{\alpha}\Big) = 1-\alpha,\ \ P\Big(\bar{X}_n - \sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}} \cdot z_{1-\alpha} < p\Big) = 1-\alpha$$

Interval de încredere pentru **proporţia** p	Expresia intervalului de încredere, folosind datele statistice
bilateral	$\Big(\bar{x}_n - \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}} \cdot z_{1-\frac{\alpha}{2}}\ ,\ \bar{x}_n + \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}} \cdot z_{1-\frac{\alpha}{2}}\Big) \cap (0,1)$
unilateral	$\Big(0\, , \bar{x}_n - \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}} \cdot z_{\alpha}\Big) \cap (0,1)$
	$\Big(\bar{x}_n - \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}} \cdot z_{1-\alpha}\, ,\ 1\Big) \cap (0,1)$

Exemplul 1: $p\cdot$ 100% din populaţia unui oraş susţine un anumit candidat la alegerile viitoare, unde $p \in (0,1)$ este parametru necunoscut. S-a ales un eşantion aleatoriu de dimensiunea 2000 şi s-a determinat că 980 de persoane susţin candidatul. Construiţi un interval de încredere bilateral cu nivelul de încredere 95% pentru proporţia p necunoscută.
R.: Intervalul de încredere bilateral este

$$\Big(\bar{x}_n - \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}} \cdot z_{1-\frac{\alpha}{2}}\ ,\ \bar{x}_n + \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}} \cdot z_{1-\frac{\alpha}{2}}\Big) \cap (0,1)\,,$$

unde $n = 2000$, $\alpha = 0.05$, $\bar{x}_n = 980/2000 = 0.49$, $z_{1-\frac{\alpha}{2}} = \texttt{norminv}(1-\frac{0.05}{2}, 0, 1) \approx 1.96$.
Valoarea intervalului de încredere bilateral este $(0.4678, 0.51212)$. ♢