

Curs - Probabilități și Statistică 2022/2023
Secția Informatică

Facultatea de Matematică și Informatică
Universitatea Babeș-Bolyai, Cluj-Napoca
Dr. Habil. Hannelore Lisei



Teoria Probabilităților

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

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

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

Fenomene și procese aleatoare apar, de exemplu, în:

- pariuri, loto (6 din 49), jocuri de noroc / jocuri online
- previziuni meteo
- previziuni economice / financiare, investiții, cumpărături online (predicția comportamentului clienților)
- sondaje de opinie (analiza unor strategii politice), asigurări (evaluarea riscurilor / pierderilor)



[Sursa: www.financialmarket.ro]

→ **în informatică:**

- ▷ sisteme de comunicare, prelucrarea informației, modelarea traficului în rețea, criptografie;
- ▷ analiza probabilistică a performanței unor algoritmi, fiabilitatea sistemelor, predicții în cazul unor sisteme complexe;
- ▷ algoritmi de simulare, machine learning, data mining, recunoașterea formelor / a vocii;
- ▷ generarea de numere aleatoare (pseudo-aleatoare cu ajutorul calculatorului), algoritmi aleatori
- ▷ <https://www.random.org/randomness/>

se pot genera numere cu ”adevarat aleatoare” (*true random numbers*), folosind ca sursă un fenomen fizic, ca de exemplu o sursă radioactivă (momentele de timp în care particulele se dezintegrează sunt complet impredictibile - de exemplu *HotBits service* din Elveția), sau variațiile de amplitudine din perturbările atmosferice (atmospheric noise, folosit de Random.org), sau zgomotul de fond dintr-un birou etc.

Octave online: <https://octave-online.net>

Exemplu: Generarea de valori aleatoare (în Octave)

```
clear all %șterge datele folosite anterior
clc % șterge informațiile anterioare afișate pe terminal
a=rand
% o valoare aleatoare în intervalul (0,1)
```

```

v1=rand(1,10)
% un vector cu 10 valori aleatoare în intervalul (0,1)
a=4; b=10;
v2=a+(b-a)*rand(1,15)
%un vector cu 15 valori aleatoare în intervalul (4,10)
A=randi(5,2,4)
% o matrice 2 x 4 de valori aleatoare din mulțimea {1,2,3,4,5},
% fiecare cu șansa de apariție 1/5
v3=randi(2,1,10)-1
% un vector cu 10 valori aleatoare 0 și 1,
% fiecare cu șansa de apariție 1/2

```

Algoritmi aleatori

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

- ▷ durata de execuție, spațiul de stocare, rezultatul obținut sunt variabile aleatoare (chiar dacă se folosesc aceleași valori input)
- ▷ la anumite tipuri de algoritmi corectitudinea e garantată doar cu o anumită probabilitate
- ▷ în mod paradoxal, uneori incertitudinea ne poate oferi mai multă eficiență

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

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

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

↔ se examinează probabilitatea cu care rezultatul este corect; probabilitatea de eroare poate fi scăzută semnificativ prin execuții repetate, independente;

Exemplu:

- ▷ 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”;

Exercițiu: Fie $S(1), \dots, S(300)$ un vector cu 300 de elemente, din mulțimea $\{0, 1, 2\}$ (ordinea lor este necunoscută; se presupune că șirul conține cel puțin un 0). → De care tip este următorul algoritm (scris în Octave)?

```

clear all
clc
disp('prima versiune')
S=randi(3,1,300)-1;
%un vector cu 300 de elemente, din multimea {0,1,2}
k=0;
do
    k=k+1;
    i=randi(300);
until (S(i) == 0)
    % i indicele, pentru care S(i)=0
    % k = număr iterații până se găsește aleator un 0
    fprintf('la a %d-a iteratie s-a gasit aleator 0 in S \n',k)

```

Răspuns: Algoritm de tip Las Vegas.

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

```

clear all
disp('a doua versiune')
M=4;
% număr maxim de iterații
S=randi(3,1,300)-1;
%un vector cu 300 de elemente, din multimea {0,1,2}
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 iterații până se găsește
% aleator un 0 sau programul s-a oprit

if S(i)==0
    fprintf('la a %d-a iteratie s-a gasit aleator 0 in S \n',k)
else
    fprintf('in %d iteratii nu s-a gasit aleator 0 in S \n',k)
endif

```

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

Noțiuni introductive:

- **Experiența aleatoare** este acea experiență al cărei rezultat nu poate fi cunoscut decât după încheierea ei.
- **Evenimentul** este rezultatul unui experiment.

Exemple:

- ▷ Experiment: aruncarea a două zaruri, eveniment: ambele zaruri indică 1
- ▷ experiment: aruncarea unei monede, eveniment: moneda indică pajură
- ▷ experiment: extragerea unei cărți de joc, eveniment: s-a extras as
- ▷ experiment: extragerea unui număr la loto, eveniment: s-a extras numărul 27
- **evenimentul imposibil**, notat cu \emptyset , este evenimentul care nu se realizează niciodată la efectuarea experienței aleatoare
- **evenimentul sigur** este un eveniment care se realizează cu certitudine la fiecare efectuare a experienței aleatoare
- **spațiul de selecție**, notat cu Ω , este mulțimea tuturor rezultatelor posibile ale experimentului considerat
 - ◇ spațiul de selecție poate fi finit sau infinit
- dacă A este o submulțime a lui Ω atunci A se numește **eveniment aleator**, iar dacă A are un singur element atunci A este un **eveniment elementar**.
- ▷ *O analogie între evenimente și mulțimi permite o scriere și o exprimare mai comode ale unor idei și rezultate legate de conceptul de eveniment aleator.*

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

e_i : s-a obținut numărul i ($i = 1, \dots, 6$) ; $e_1, e_2, e_3, e_4, e_5, e_6$ sunt evenimente elementare

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

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



Operații cu evenimente

- dacă $A, B \subseteq \Omega$, atunci **evenimentul reuniune** $A \cup B$ este un eveniment care se produce dacă cel puțin unul din evenimentele A sau B se produce
- dacă $A, B \subseteq \Omega$, atunci **evenimentul intersecție** $A \cap B$ este un eveniment care se produce dacă cele două evenimente A și B se produc în același timp
- dacă $A \subseteq \Omega$ atunci **evenimentul contrar** sau **complementar** \bar{A} este un eveniment care se realizează atunci când evenimentul A nu se realizează
- $A, B \subseteq \Omega$ sunt **evenimente incompatibile (disjuncte)**, dacă $A \cap B = \emptyset$

- dacă $A, B \subseteq \Omega$, atunci **evenimentul diferență** $A \setminus B$ este un eveniment care se produce dacă A are loc și B nu are loc, adică

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

Relații între evenimente

- dacă $A, B \subseteq \Omega$, atunci A **implică** B , dacă producerea evenimentului A conduce la producerea evenimentului B : $A \subseteq B$
- dacă A implică B și B implică A , atunci evenimentele A și B sunt **egale**: $A = B$

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

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

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

asociative

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

și distributive

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

satisfac **legile lui De Morgan**

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

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

Frecvența relativă și frecvența absolută

Def. 2. Fie A un eveniment asociat unei experiențe, repetăm experiența de n ori (în aceleași condiții date) și notăm cu $r_n(A)$ numărul de realizări ale evenimentului A ; **frecvența relativă** a evenimentului A este numărul

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

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

Definiția clasică a probabilității

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

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

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

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

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

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

Răspuns: $f_n(A) = \frac{22}{100} = 0.22$ frecvența relativă a evenimentului A

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

$$A = \{(c, p, p, p), (p, c, p, p), (p, p, c, p), (p, p, p, c)\}$$
$$\Rightarrow P(A) = \frac{4}{2^4} = 0.25 \text{ probabilitatea evenimentului } A.$$



Exemplu istoric - Joc de zaruri (sec. XVII): Un pasionat jucător de zaruri, cavalerul de Méré, susținea în discuțiile sale cu B. Pascal că a arunca un zar de 4 ori pentru a obține cel puțin o dată fața șase, este același lucru cu a arunca de 24 ori câte două zaruri pentru a obține cel puțin o dublă de șase.

Cu toate acestea, cavalerul de Méré a observat că jucând în modul al doilea (cu două zaruri aruncate de 24 ori), pierdea față de adversarul său, dacă acesta alegea primul mod (aruncarea unui singur zar de 4 ori). Pascal și Fermat au arătat că probabilitatea de câștig 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. Deși diferența dintre cele două probabilități este mică, totuși, la un număr mare de partide, jucătorul cu probabilitatea de câștig p_1 câștigă în fața jucătorului cu probabilitatea de câștig p_2 . Practica jocului confirmă astfel justetea raționamentului matematic, contrar credinței lui de Méré.



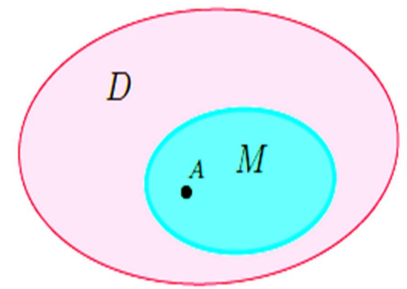
Definiția axiomatică a probabilității

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

Probabilitatea geometrică: Măsura unei mulțimi corespunde lungimii în \mathbb{R} , ariei în \mathbb{R}^2 , volumului în \mathbb{R}^3 . Fie $M \subset D \subset \mathbb{R}^n$, $n \in \{1, 2, 3\}$, mulțimi cu măsură finită.

Alegem aleator un punct $A \in D$ (în acest caz spațiul de selecție 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 probabilității a fost creată în anii '30 ai secolului XX de către matematicianul rus **Andrei Nikolaevici Kolmogorov**, care, în anul **1933**, a dezvoltat teoria axiomatică a probabilității în lucrarea sa *Conceptele de bază ale Calculului Probabilității*.

$\Rightarrow P : \mathcal{K} \rightarrow \mathbb{R}$ este o funcție astfel încât oricărui eveniment aleator $A \in \mathcal{K}$ i se asociază valoarea $P(A)$, **probabilitatea de apariție a evenimentului A**

$\hookrightarrow \mathcal{K}$ este o mulțime de evenimente și 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 spațiul de selecție Ω se numește **σ -algebră** dacă sunt satisfăcute condițiile:

(1) \mathcal{K} este nevidă;

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

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

Observație: În context probabilistic, o familie \mathcal{A} de evenimente din Ω este o *algebră*, dacă au loc (1) și (2) din Def. 4, iar (3) este valabilă pentru un număr finit de evenimente (\mathcal{A} este închisă în raport cu reuniuni finite).

Fie $M = \{1, 2, 3, \dots\}$ și fie

$$\mathcal{A} := \{A \subseteq M : A \text{ sau } \bar{A} = M \setminus A \text{ este finită}\}.$$

\mathcal{A} verifică (1) și (2) din Def. 4, dar \mathcal{A} îndeplinește (3) din Def. 4 doar pentru o reuniune finită de mulțimi din \mathcal{A} ; \mathcal{A} este o algebră, dar *nu* este o σ -algebră, pentru că există un număr infinit numărabil de evenimente $A_n, n \in \{1, 2, 3, \dots\}$, din \mathcal{A} a căror reuniune nu este în \mathcal{A} . De exemplu, fie $A_n = \{2n\}$ pentru $n \in \{1, 2, 3, \dots\}$. Observăm că $A_n \in \mathcal{A}$ pentru fiecare $n \in \{1, 2, 3, \dots\}$, dar

$$\bigcup_{n=1}^{\infty} A_n = \{2, 4, 6, \dots\} \notin \mathcal{A}.$$

Exemple: 1) Dacă $\emptyset \neq A \subset \Omega$ atunci $\mathcal{K} = \{\emptyset, A, \bar{A}, \Omega\}$ este o σ -algebră.

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

3) Dacă \mathcal{K} este o σ -algebră pe Ω și $\emptyset \neq B \subseteq \Omega$, atunci

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

este o σ -algebră pe mulțimea B . ◇

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

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

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

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

Def. 5. Fie \mathcal{K} o σ -algebră pe Ω . O funcție $P : \mathcal{K} \rightarrow \mathbb{R}$ se numește **probabilitate** dacă satisface axiomele:

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

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

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

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

Tripletul (Ω, \mathcal{K}, P) se numește **spațiu de probabilitate**.

Exemplu: 1) Cea mai simplă (funcție de) probabilitate se obține pentru cazul unui *spațiu de selecție finit* Ω : fie $\mathcal{K} = \mathcal{P}(\Omega)$ (mulțimea tuturor submulțimilor lui Ω) și $P : \mathcal{K} \rightarrow \mathbb{R}$ definită astfel

$$P(A) = \frac{\#A}{\#\Omega}, \text{ unde } \#A \text{ reprezintă numărul elementelor lui } A \in \mathcal{P}(\Omega).$$

P astfel definită verifică Def. 5 și corespunde *definiției clasice a probabilității unui eveniment* (a se vedea Def. 3).

2) Fie $\Omega = \mathbb{N} = \{0, 1, 2, \dots\}$, $\mathcal{K} = \mathcal{P}(\mathbb{N})$ și $P : \mathcal{K} \rightarrow \mathbb{R}$ definită prin $P(\{n\}) = \frac{1}{2^{n+1}}$, $n \in \mathbb{N}$.

Are loc $P(\mathbb{N}) = \sum_{n=0}^{\infty} \frac{1}{2^{n+1}} = 1$, iar axiomele din Def. 5 sunt îndeplinite. $(\mathbb{N}, \mathcal{P}(\mathbb{N}), P)$ este un spațiu de probabilitate; Def. 5-(3) este îndeplinită, datorită teoremei din analiză, care afirmă că pentru o serie cu termeni pozitivi, schimbarea ordinii termenilor seriei nu schimbă natura seriei și nici suma ei. ♣

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

$$(1) P(\bar{A}) = 1 - P(A) \text{ și } 0 \leq P(A) \leq 1;$$

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

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

$$(4) A \subseteq B \implies P(A) \leq P(B), \text{ adică } P \text{ este monotonă};$$

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

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

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

Exemplu: Dintr-un pachet de 52 de cărți 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 și 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 și 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 (Ω, \mathcal{K}, P) un spațiu de probabilitate. Evenimentele $A, B \in \mathcal{K}$ sunt **evenimente independente**, dacă

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

Observație: Fie evenimentele $A, B \in \mathcal{K}$. Evenimentele A și B sunt **independente**, dacă apariția evenimentului A , nu influențează apariția evenimentului B și invers. Două evenimente se numesc **dependente** dacă probabilitatea realizării unuia dintre ele depinde de faptul că celălalt eveniment s-a produs sau nu.

Exercițiu: 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 și B evenimente independente? Sunt A și B evenimente disjuncte?

Sunt A și C evenimente independente? Sunt A și C evenimente disjuncte?



P. 3. Fie (Ω, \mathcal{K}, P) un spațiu de probabilitate și fie $A, B \in \mathcal{K}$. Sunt echivalente afirmațiile:

(1) A și B sunt independente.

(2) \bar{A} și B sunt independente.

(3) A și \bar{B} sunt independente.

(4) \bar{A} și \bar{B} sunt independente.

Def. 7. Fie (Ω, \mathcal{K}, P) un spațiu de probabilitate. B_1, \dots, B_n sunt n **evenimente independente (în totalitate)** din \mathcal{K} dacă

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

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

Observație; Din Def. 7 avem $A, B, C \in \mathcal{K}$ sunt trei evenimente independente (în totalitate), dacă

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

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

Exemplu: 1) Din Def. 6 și Def. 7 deducem că, independența (în totalitate) implică și independența a două câte două evenimente. Afirmația inversă, însă, nu are loc. Drept (contra)exemplu putem lua experimentul aleator ce constă în aruncarea unui tetraedru regulat, ale cărui patru fețe sunt vopsite astfel: una este roșie, una este albastră, una este verde și una este colorată având cele trei culori. Se aruncă tetraedrul și se consideră evenimentele:

R : tetraedrul cade pe o parte ce conține culoarea roșie;

A : tetraedrul cade pe o parte ce conține culoarea albastră;

V : tetraedrul cade pe o parte ce conține culoarea verde.

Sunt cele 3 evenimente *independente în totalitate*?

R.: Nu, cele 3 evenimente nu sunt independente în totalitate pentru că $P(R \cap A \cap V) = \frac{1}{4} \neq P(R)P(A)P(V) = \frac{1}{8}$. Însă cele 3 evenimente verifică:

$$P(R \cap A) = P(R)P(A) = \frac{1}{4}; P(V \cap A) = P(V)P(A) = \frac{1}{4}; P(R \cap V) = P(R)P(V) = \frac{1}{4}.$$

2) Pentru a verifica dacă n evenimente distincte B_1, \dots, B_n sunt independente în totalitate câte relații trebuie verificate?

R.: $C_n^2 + C_n^3 + \dots + C_n^n = 2^n - C_n^0 - C_n^1 = 2^n - 1 - n$. ◆

Exemplu:

Se dă algoritmul de tip Monte-Carlo (Exemplu de căutare aleatoare a unui element într-un vector):

Fie S o permutare aleatoare a vectorului $[0 \ 0 \ 0 \ 0 \ 1 \ 1 \ 1 \ 1 \ 1 \ 2 \ 2 \ 2 \ 2 \ 3 \ 3 \ 3 \ 3]$.

Se caută aleator un 2 în S în cel mult M iterații .

```
clear all
clc
M=input('M='); % numar maxim de iteratii; M >= 1
% de ex. M=3
U=[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=U(randperm(length(U)))
% S este o permutare aleatoare a vectorului initial U

% vom cauta aleator un 2 in vectorul S
% iteratiile se repeta pana se gaseste aleator un 2 din S
% sau se atinge numarul maxim M de iteratii
k=0;
do
    k=k+1; % se numara iteratiile
    i=randi(20); % se alege o valoare aleatoare din S -> S(i)
    fprintf('S(%d)=%d \n',i,S(i))
until ( (S(i) == 2) | (k==M) )
if S(i)==2
    fprintf('la a %d-a iteratie s-a gasit aleator 2 in S \n',k)
else
    fprintf('in %d iteratii nu s-a gasit aleator 2 in S \n',k)
endif
```

Se calculează probabilitățile (teoretice) ale următoarelor evenimente:

$$P(\text{"primul 2 este găsit aleator la a } M\text{-a iterație"}) = \left(\frac{3}{4}\right)^{M-1} \cdot \frac{1}{4},$$

$$P(\text{"2 nu este găsit în } M \text{ iterații"}) = \left(\frac{3}{4}\right)^M,$$

probabilitatea evenimentului complementar este

$$P(\text{"cel puțin un 2 este găsit în } M \text{ iterații"}) = 1 - \left(\frac{3}{4}\right)^M \longrightarrow 1, \text{ când } M \rightarrow \infty.$$



Probabilitate condiționată

În anumite situații este necesar să cunoaștem probabilitatea unui eveniment particular, care urmează să aibă loc, știind deja că alt eveniment a avut loc.

▷ Experiment: Se aruncă simultan două zaruri. Notăm cu S suma numerelor rezultate din aruncarea celor două zaruri.

a) $P(S = 11) = ?$

b) Dacă se știe că S este un număr prim, care este probabilitatea ca $S = 11$?

Def. 8. Fie (Ω, \mathcal{K}, P) un spațiu de probabilitate și fie $A, B \in \mathcal{K}$. **Probabilitatea condiționată a evenimentului A de către evenimentul B** este $P(\cdot|B) : \mathcal{K} \rightarrow [0, 1]$ definită prin

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

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

Observație: 1) $P(A|B)$: probabilitatea condiționată a lui A de către B , este **probabilitatea de a se realiza evenimentul A dacă în prealabil s-a realizat evenimentul B** .

2) Într-un experiment în care cazurile posibile sunt finite la număr și au aceleași șanse de a se realiza, atunci se poate folosi

$$P(A|B) = \frac{\text{numărul de cazuri favorabile apariției lui } A \cap B}{\text{numărul total de cazuri posibile pentru apariția lui } B}.$$

3) Fie evenimentele $A, B \in \mathcal{K}$ astfel încât $P(A) > 0$ și $P(B) > 0$. Evenimentele A și B

sunt **independente** (a se vedea Def. 6), dacă apariția evenimentului A , nu influențează apariția evenimentului B și invers, adică

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

Exemplu: Se extrag succesiv fără returnare două bile dintr-o urnă cu 4 bile albe și 5 bile roșii.

a) Știind că prima bilă este roșie, care este probabilitatea (condiționată) ca a doua bilă să fie albă?

b) Care este probabilitatea ca ambele bile să fie roșii?

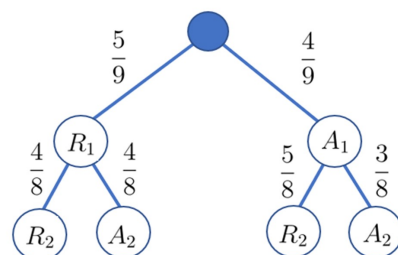
R.: pentru $i \in \{1, 2\}$ fie evenimentele

R_i : la a i -a extragere s-a obținut o bilă roșie;

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

a) $P(A_2|R_1) = \frac{4}{8}$.

b) $P(R_1 \cap R_2) = P(R_2|R_1)P(R_1) = \frac{4}{8} \cdot \frac{5}{9} \cdot \clubsuit$

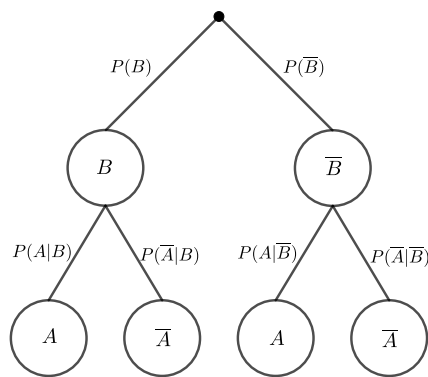


Extragere fără reținere

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

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

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



Probabilități condiționate

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

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

P. 5. (Formula probabilității totale) Într-un spațiu de probabilitate (Ω, \mathcal{K}, P) considerăm partiția $\{H_1, \dots, H_n\}$ a lui Ω cu $H_i \in \mathcal{K}$ și $P(H_i) > 0 \forall i \in \{1, \dots, n\}$, și fie $A \in \mathcal{K}$. Atunci are loc

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

Exemplu: Într-o urnă sunt 7 bile albe, notate cu 1, 2, 3, 4, 5, 6, 7, și 6 bile roșii notate cu 8, 9, 10, 11, 12, 13. Se extrage o bilă. a) Știind că bila extrasă este roșie, care este probabilitatea p_1 , ca

numărul de pe bilă să fie divizibil cu 4? **b)** Știind că prima bilă este roșie, care este probabilitatea p_2 , ca o a doua bilă extrasă să indice 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 roșie;

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 și P.4, scriem succesiv

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



Exemplu: Ce probabilități calculează programul de mai jos? Ce tip de algoritm aleator este?

► `randi(imax,n,m)` generează o $n \times m$ matrice cu valori întregi aleatoare (pseudoaleatoare) între 1 și imax.


```
clear all
ci=0;
cp=0;
c=0;
a=0;
b=0;
N=1000;
A=[1:20];
for i=1:N
    v=A(randi(length(A)));
    ci=ci+mod(v,2);
    cp=cp+(mod(v,2)==0);
    a1=a1+ mod(v,2)*(mod(v,3)==0);
    a2=a2+ (mod(v,2)==0)*(6<=v && v<=10);
end
p1=a1/ci
p2=a2/cp
```

R.: Se extrage aleator un număr din șirul $A = [1, 2, \dots, 20]$.

► p_1 estimează probabilitatea condiționată ca numărul ales aleator să fie divizibil cu 3, *știind* că s-a extras un număr impar;

► p_2 estimează probabilitatea condiționată ca numărul ales aleator să provină din mulțimea $\{6, 7, 8, 9, 10\}$, *știind* că s-a extras un număr par;

Algoritmul este de tip Monte-Carlo!

Care sunt valorile teoretice pentru probabilitățile p_1, p_2 din acest exemplu? 

P. 6. (Formula înmulțirii probabilităților)

Fie (Ω, \mathcal{K}, P) un spațiu de probabilitate și fie $A_1, \dots, A_n \in \mathcal{K}$ astfel încât $P(A_1 \cap \dots \cap A_{n-1}) > 0$. Atunci,

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

Observație: 1) Formula înmulțirii probabilităților a două evenimente ($n = 2$) este

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

2) În cazul, în care evenimentele aleatoare A_1, \dots, A_n sunt *independente în totalitate*, atunci formula înmulțirii probabilităților are forma

$$P(A_1 \cap \dots \cap A_n) = P(A_1)P(A_2) \dots P(A_n).$$

Exemplu: Într-o urnă sunt 2 bile verzi și 3 bile albastre. Se extrag 2 bile succesiv, fără returnare. Care este probabilitatea ca

a) prima bilă să fie verde, iar cea de-a doua albastră?

b) cele 2 bile să aibă aceeași culoare?

c) a doua bilă să fie albastră?

d) prima bilă să fie verde, *știind* că a doua este albastră?

e) se mai extrage o a treia bilă; se cere probabilitatea ca prima bilă să fie verde, cea de-a doua albastră și a treia tot albastră.

R.: Notăm pentru $i \in \{1, 2, 3\}$ evenimentele:

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

a) folosim 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 probabilității 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 înmulțire a probabilităților P.6:

$$P(V_1 \cap A_2 \cap A_3) = P(V_1) \cdot P(A_2|V_1) \cdot P(A_3|V_1 \cap A_2) = \frac{2}{5} \cdot \frac{3}{4} \cdot \frac{2}{3}.$$

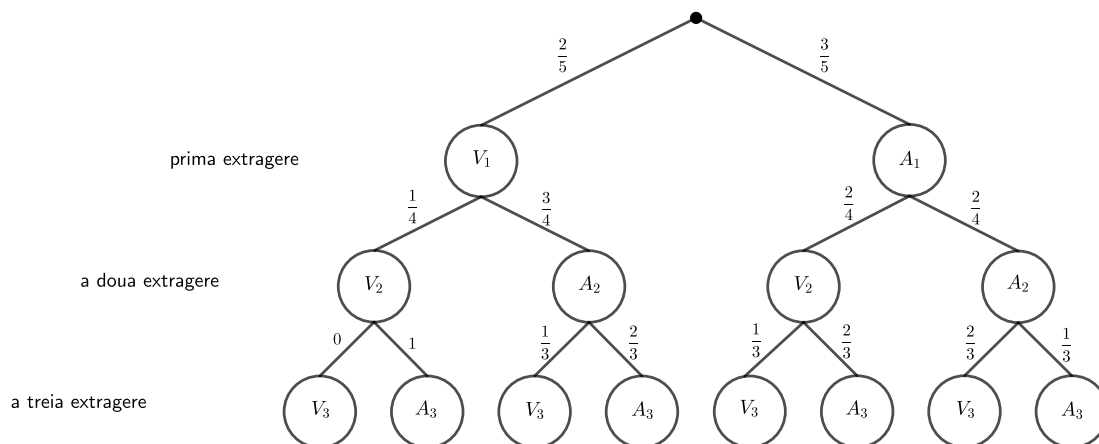


Fig. 3. Extragere fără returnare



Formula lui Bayes

Formula lui Bayes este o metodă de a "corecta" (a revizui, a îmbunătăți) pe baza unor noi date (informații) disponibile o probabilitate determinată apriori. Se pornește cu o estimare pentru probabilitatea unei anumite ipoteze H (engl. *hypothesis*). Dacă avem noi date (date de antrenare, dovezi, informații, evidențe - engl. *evidence*) E , ce privesc ipoteza H , se poate calcula o probabilitate "corectată" pentru ipoteza H , numită probabilitate posterioară (a-posteriori).

↪ $P(H)$ probabilitatea ca ipoteza H să fie adevărată, numită și *probabilitatea apriori*;

↪ probabilitatea condiționată $P(H|E)$ este *probabilitatea posterioară* (corectată de cunoașterea noilor date / informații / evidențe);

↪ $P(E|H)$ probabilitatea ca să apară datele (informațiile), știind că ipoteza H este adevărată;

↪ $P(E|\bar{H})$ probabilitatea ca să apară datele (informațiile), știind că ipoteza H este falsă (ipoteza \bar{H} este adevărată).

Folosind P.5 (cu partiția $\{H, \bar{H}\}$) 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

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

P. 7. (Formula lui Bayes)

Într-un spațiu de probabilitate (Ω, \mathcal{K}, P) considerăm partiția $\{H_1, \dots, H_n\}$ a lui Ω cu $H_i \in \mathcal{K}$ și $P(H_i) > 0 \forall i \in \{1, \dots, n\}$, și fie $E \in \mathcal{K}$ astfel încât $P(E) > 0$. Atunci,

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

▷ pentru $i \in \{1, 2, \dots, n\}$ $P(H_i)$ sunt **probabilități apriori** pentru H_i , numite și ipoteze (asertiuni; engl. *hypothesis*)

▷ E se numește **evidență** (dovadă, premisă, informație; engl. *evidence*);

▷ cu formula lui Bayes se calculează probabilitățile pentru ipoteze, cunoscând evidența: $P(H_j|E)$, $j \in \{1, 2, \dots, n\}$, care se numesc **probabilități posterioare** (ulterioare);

▷ $P(E|H_i)$, $i \in \{1, 2, \dots, n\}$, reprezintă verosimilitatea (engl. *likelihood*) datelor observate.

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

Exemplu: Considerăm evenimentele (în teste clinice, programe de screening):

H : o persoană aleasă aleator dintr-o populație 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 populație să sufere de alergia \mathcal{A} ;
- *sensibilitatea testului* $s_1 = P(E|H)$ probabilitatea ca testul să fie pozitiv, știind că (în timp ce) alergia este prezentă [probabilitatea ca prezența alergiei \mathcal{A} să fi fost corect identificată de test];

- *specificitatea testului* $s_2 = P(\bar{E}|\bar{H})$ probabilitatea ca testul să fie negativ, știind că (în timp ce) alergia nu este prezentă [probabilitatea ca absența alergiei \mathcal{A} să fi fost corect identificată de test];

▷ probabilitatea de a obține răspuns fals pozitiv este $P(E|\bar{H}) = 1 - s_2$ testul este pozitiv, dar persoana (se știe că) nu are alergia \mathcal{A} ;

▷ probabilitatea de a obține răspuns fals negativ este $P(\bar{E}|H) = 1 - s_1$ testul este negativ, dar persoana (se știe că) are alergia \mathcal{A} ;

▷ un test clinic predictiv bun implică valori apropiate de 1 pentru s_1 și s_2 ;

► cunoscând p, s_1, s_2 se dorește a se determina *valoarea predictivă* $P(H|E)$ [este probabilitatea ca o persoană, care are un test pozitiv, să fie corect diagnosticată cu alergia \mathcal{A}]:

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

Date statistice: 2120 persoane au fost testate în cadrul unui program de screening, privind alerggia \mathcal{A} . S-au obținut următoarele informații:

▷ AP=400 este numărul persoanelor adevărat pozitive din setul de testare, adică numărul persoanelor care au alerggia \mathcal{A} și au test pozitiv; $\#(H \cap E)$ ¹

▷ FP=210 este numărul persoanelor fals pozitive din setul de testare adică numărul persoanelor care nu au alerggia \mathcal{A} și au test pozitiv; $\#(\bar{H} \cap E)$

▷ FN=310 este numărul persoanelor fals negative din setul de testare adică numărul persoanelor care au alerggia \mathcal{A} și au test negativ; $\#(H \cap \bar{E})$

▷ AN=1200 este numărul persoanelor adevărat negative din setul de testare, adică numărul persoanelor care nu au alerggia \mathcal{A} și au test negativ; $\#(\bar{H} \cap \bar{E})$.

		starea actuală		
		(+)	(-)	total
predicția	(+)	AP	FP	AP+FP
	(-)	FN	AN	FN+AN
	total	AP+FN	FP+AN	AP+FP+FN+AN

Matricea de confuzie (engl. confusion matrix)

		starea actuală (realitatea)		
		H : are alerggia \mathcal{A} (+)	\bar{H} : nu are alerggia \mathcal{A} (-)	total
predicția	E : test pozitiv \mathcal{A} (+)	400 (adevărat pozitiv AP)	210 (fals pozitiv FP)	610
	\bar{E} : test negativ \mathcal{A} (-)	310 (fals negativ FN)	1200 (adevărat negativ AN)	1510
	total	710	1410	2120

Matricea de confuzie construită cu datele statistice

Pe baza datelor statistice: a) probabilitatea ca o persoană, despre care se știe că are test pozitiv, are în realitate alerggia \mathcal{A} , este

$$P(H|E) = \frac{400}{610} \approx 0.65 \text{ (valoarea predictivă pozitivă);}$$

b) probabilitatea ca o persoană, despre care se știe că are test negativ, nu are în realitate alerggia \mathcal{A} , este

$$P(\bar{H}|\bar{E}) = \frac{1200}{1510} \approx 0.79 \text{ (valoarea predictivă negativă).}$$

¹ numărul de elemente din $H \cap E$



diagnosticare	<i>machine learning (ML)</i>
măsurile de performanță	<i>measuring the performance of a binary classification model</i>
valoarea predictivă pozitivă = $\frac{AP}{AP+FP}$	<i>positive predictive value; precision</i>
valoarea predictivă negativă = $\frac{AN}{AN+FN}$	<i>negative predictive value</i>
sensibilitatea = $\frac{AP}{AP+FN}$	<i>recall; probability of detection; true positive rate</i>
specificitatea = $\frac{AN}{AN+FP}$	<i>true negative rate</i>
acuratețea = $\frac{AP+AN}{AP+FP+AN+FN}$	<i>accuracy</i>

★ Probabilitățile condiționate sunt folosite în probleme de clasificare, în teoria deciziilor, în predicție, în diagnosticare, etc.

Variable aleatoare

→ Variabilele aleatoare apar ca funcții, ce depind de rezultatul (aleator) al efectuării unui anumit experiment.

Exemplu: 1) La aruncarea a două zaruri, suma numerelor obținute este o variabilă aleatoare $S : \Omega \rightarrow \{2, 3, \dots, 12\}$, unde Ω conține toate evenimentele elementare ce se pot obține la aruncarea a două zaruri, adică $\Omega = \{(\omega_i^1, \omega_j^2) : i, j = \overline{1, 6}\}$, unde (ω_i^1, ω_j^2) este evenimentul elementar: la primul zar s-a obținut numărul i și la al doilea zar s-a obținut numărul j , unde $i, j = \overline{1, 6}$.

Astfel, $P(S = 5) = \frac{4}{36}$, $P(S = 6) = \frac{5}{36}$, etc.

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

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

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

Notăție 1. *variabilă/variabile aleatoare* \rightarrow *v.a.*

O variabilă aleatoare este:

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

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

V.a. discrete: exemple de v.a. numerice discrete: suma numerelor obținute la aruncarea a 4 zaruri, numărul produselor defecte produse de o anumită firmă într-o săptămână; numărul

apelurilor telefonice într-un call center în decursul unei ore; numărul de accesări ale unei anumite pagini web în decursul unei anumite zile (de ex. duminică); numărul de caractere transmise eronat într-un mesaj de o anumită lungime; exemple de v.a. categoriale (\rightarrow se clasifică în categorii): prognoza meteo: *plouos, senin, înnorat, cețos*; calitatea unor servicii: *nesatisfăcătoare, satisfăcătoare, bune, foarte bune, excepționale*, etc.

V.a. continue sunt v.a. numerice: timpul de funcționare până la defectare a unei piese electronice, temperatura într-un oraș, viteza înregistrată de radar pentru mașini care parcurg o anumită zonă, cantitatea de apă de ploaie (într-o anumită perioadă), duritatea unui anumit material, etc.

Variabile aleatoare numerice - definiție formală

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

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

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

Def. 11. Distribuția de probabilitate a v.a. discrete X

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

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

\triangleright O variabilă aleatoare discretă X este caracterizată de distribuția de probabilitate! \triangleright Notăm $\{X = x_i\} = \{\omega \in \Omega : X(\omega) = x_i\}$, $i \in I$; acesta este un eveniment din \mathcal{K} pentru fiecare $i \in I$.

Distribuții discrete clasice

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

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

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

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

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

Distribuția Bernoulli: $X \sim \text{Bernoulli}(p), p \in (0, 1)$

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

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

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



generare în Matlab/Octave:

```
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)  
%%%%%%%%
```

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

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

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

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

$$X \sim \text{Bino}(n, p) \iff X \sim \left(C_n^k p^k (1 - p)^{n-k} \right)_{k \in \{0, \dots, n\}}$$

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

\rightarrow are loc **formula binomială**

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

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

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

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

▷ Distribuția binomială corespunde **modelului cu extragerea bilelor dintr-o urnă cu bile de două culori și cu returnarea bilei după fiecare extragere**:

Într-o urnă sunt n_1 bile albe și 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

$\Rightarrow X_1 \sim \text{Bino}(n, p_1)$ cu $p_1 = \frac{n_1}{n_1+n_2}$, $X_2 \sim \text{Bino}(n, p_2)$ cu $p_2 = \frac{n_2}{n_1+n_2}$.

▷ **Exemplu:** Fie un canal de comunicare binară care transmite cuvinte codificate de N biți 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 știe că transmiterea biților 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 biți

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



Exerciții: 1) Un client accesează o dată pe zi o anumită pagină web (care oferă anumite produse) cu probabilitatea 0.4. Cu ce probabilitate clientul accesează această pagină în total de 3 ori în următoarele 6 zile?

2) O rețea de laborator este compusă din 15 calculatoare. Rețeaua a fost atacată de un virus nou, care atacă un calculator cu o probabilitate 0.4, independent de alte calculatoare. Care este probabilitatea ca virusul a atacat

- a) cel mult 10,
- b) cel puțin 10,
- c) exact 10 calculatoare?



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

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

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

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

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

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

Matlab/Octave: `hygernd($n_1 + n_2, n_1, n, \dots$)` 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 și $n_2 = 3$ 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 obținut bilă albă

$N_i = \bar{A}_i$: la a i -a extragere s-a obținut bilă neagră.

Scriem

$$\begin{aligned} P(X = 1) &= P(A_1 \cap N_2 \cap N_3) + P(N_1 \cap A_2 \cap N_3) + P(N_1 \cap N_2 \cap A_3), \\ P(A_1 \cap N_2 \cap N_3) &= P(A_1)P(N_2|A_1)P(N_3|A_1 \cap N_2) = \frac{2}{5} \cdot \frac{3}{4} \cdot \frac{2}{3} = \frac{1}{5} \\ P(N_1 \cap A_2 \cap N_3) &= P(N_1)P(A_2|N_1)P(N_3|N_1 \cap A_2) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} = \frac{1}{5} \\ P(N_1 \cap N_2 \cap A_3) &= P(N_1)P(N_2|N_1)P(A_3|N_1 \cap N_2) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} = \frac{1}{5} \\ \Rightarrow P(X = 1) &= \frac{3}{5}. \end{aligned}$$

A doua metodă: O bilă albă din două se poate alege în $C_2^1 = 2$ moduri, două bile negre 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 câștigătoare?

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

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



Distribuția geometrică $X \sim \text{Geo}(p), p \in (0, 1)$

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

- $A = \text{succes}$ cu $P(A) = p$, $\bar{A} = \text{insucces}$ $P(\bar{A}) = 1 - p$
- se repetă (independent) experimentul până apare prima dată A (“succes”)
- v.a. X arată de câte ori apare \bar{A} (numărul de “insuccese”) *până* la apariția primului A (“succes”) \Rightarrow valori posibile: $X \in \{0, 1, \dots\}$

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

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

Exemplu: X v.a. ce indică numărul de retransmisii printr-un canal cu perturbări (aleatoare) până la (înainte de) prima recepție corectă a mesajului $\Rightarrow X$ are distribuție geometrică.



Variabile aleatoare independente

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

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

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

Observație: Fie evenimentele $A_i = \{X = x_i\}, i \in I$, și $B_j = \{Y = y_j\}, j \in J$.

V.a. X și Y sunt independente $\iff \forall (i, j) \in I \times J$ evenimentele A_i și 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 și Y sunt v.a. independente. Care este distribuția de probabilitate a lui X , respectiv Y ?

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

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

Def. 13. $\mathbb{X} = (X_1, \dots, X_m)$ este un **vector aleator discret** dacă fiecare componentă a sa este o variabilă aleatoare discretă.

Fie $K \subseteq \mathbb{N}$ o mulțime de indici și fie date $\mathbb{x}_k := (x_{1,k}, \dots, x_{m,k}) \in \mathbb{R}^m, k \in K$.

Dacă $\mathbb{X} : \Omega \rightarrow \{\mathbb{x}_k, k \in K\}$ este un vector aleator discret, atunci

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

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

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

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

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

unde $I, J \subseteq \mathbb{N}$ sunt mulțimi de indici,

$p_{ij} := P((X, Y) = (x_i, y_j)) = P(\{X = x_i\} \cap \{Y = y_j\}), p_{ij} > 0 \forall i \in I, j \in J$,

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

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

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

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

următorul tabel:

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

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

a) Să se determine $P(X = -1)$, $P(X \leq 3)$, respectiv $P(Y = 1)$, $P(Y \leq -1)$.

b) Sunt X și Y v.a. independente?

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

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

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

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

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

Observații:

▷ **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

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

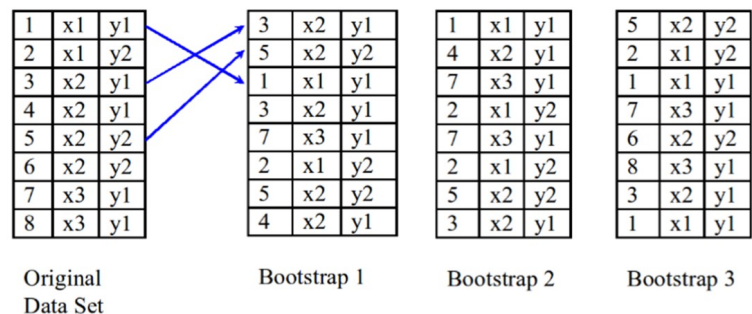
▷ (X_1, \dots, X_r) este un vector aleator discret.

▷ cazul $r = 2$ corespunde distribuției binomiale (modelul binomial cu bile de două culori într-o urnă, a se vedea pg. 22): (X_1, X_2) este un vector aleator discret, iar $X_1 + X_2 = n$; X_1 și X_2 *nu* sunt v.a. independente.

► Extragerea cu returnare (engl. *sampling with replacement*) este folosită în **metoda bootstrap**, care este o metodă utilizată pentru a estima proprietățile statistice dintr-un set de date (de exemplu, date statistice). Tehnica implică reșantionarea (engl. *resampling*) unui număr mare de seturi de date dintr-un singur set de date. Pornind de la

un set de date cu n observații, un set de date bootstrap este un set format din n observații *alese aleator cu returnare* (și independente) din setul de date inițial.

▷ **Modelul urnei cu r culori și bilă nereturnată:** fie n_i = numărul inițial de bile cu culoarea i



din urnă, $i = \overline{1, r}$;

$$\begin{aligned}
 p(k_1, \dots, k_r; n) &= \text{probabilitatea de a obține } k_i \text{ bile cu culoarea } i, i = \overline{1, r}, \\
 &\quad \text{din } n = k_1 + \dots + k_r \text{ extrageri fără returnarea bilei extrase,} \\
 &\quad \text{în care ordinea de extragere a bilelor de diverse culori nu contează} \\
 &= \frac{C_{n_1}^{k_1} \cdot \dots \cdot C_{n_r}^{k_r}}{C_{n_1 + \dots + n_r}^n}.
 \end{aligned}$$

▷ Cazul $r = 2$ corespunde **distribuției hipergeometrice**.

► Extragerea fără returnare (engl. *sampling without replacement*) este folosită în **metoda validării încrucișate** (engl. *k-fold cross validation*): În cazul validării încrucișate (*k-fold cross validation*), eșantionul original de date este împărțit aleatoriu în k sub-eșantioane de dimensiuni egale. Din cele k sub-eșantioane, un singur sub-eșantion este folosit ca date de validare pentru testarea modelului, iar celelalte $k - 1$ sub-eșantioane sunt utilizate ca date de antrenament. Procesul de validare încrucișată se repetă apoi de k ori, fiecare dintre cele k sub-eșantioane fiind utilizat exact o dată ca date de validare. Avantajul acestei metode constă în faptul că toate observațiile sunt utilizate atât pentru antrenare, cât și pentru validare, iar fiecare observație este utilizată pentru validare exact o dată. Validarea încrucișată cu $k=10$ (sau $k=5$) este utilizată în mod obișnuit. Atunci când $k = n$ (numărul de observații), validarea încrucișată este echivalentă cu validarea încrucișată numită în engleză *leave-one-out*.

Operații cu variabile aleatoare (numerice)

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

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

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

. Determinați: a) distribuțiile variabilelor aleatoare X_1 și X_2 ;

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

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

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

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


c) X_1 și 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 distribuțiile variabilelor aleatoare independente (discrete) X și Y , cum se determină distribuția pentru $X + Y$, $X \cdot Y$?

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

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

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

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

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

Clasificarea naivă Bayes

În învățarea automată, clasificatorii bayesieni naivi sunt o familie de clasificatori probabilistici simpli, bazați pe aplicarea formulei lui Bayes (a se vedea P.5) cu ipoteze “naive” de independență condiționată între atribute (engl. *features*), cunoscând clasificarea. Pentru unele tipuri de modele de probabilitate, clasificatorii bayesieni naivi pot fi antrenați foarte eficient. În aplicații practice pentru modelele bayesiene naive se folosește *metoda probabilității maxime*. Noțiunea folosită în acest context este condițional independența între v.a.

Fie (Ω, \mathcal{K}, P) un spațiu de probabilitate. De asemenea considerăm că toate probabilitățile condiționate sunt definite (adică condiționarea se face în raport cu un eveniment a cărui probabilitate nu este 0).

Def. 14. Evenimentele $A, B \in \mathcal{K}$ sunt **condițional independente**, cunoscând evenimentul $C \in \mathcal{K}$, dacă și numai dacă

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

Exemplu: Într-o cutie sunt 2 zaruri. La primul zar 3 apare cu probabilitatea $\frac{1}{6}$, iar la celălalt zar (care e măsluit) 3 apare cu probabilitatea $\frac{5}{6}$. Se alege aleator un zar, care este apoi aruncat de 2 ori. Considerăm evenimentele

A_i : “zarul ales indică 3 la aruncarea i ”, $i \in \{1, 2\}$

Z_j : “se alege zarul j ”, $j \in \{1, 2\}$.

Sunt A_1 și A_2 condițional independente, cunoscând Z_1 ? Sunt A_1 și A_2 independente?

R.: Dacă se cunoaște tipul zarului ales, atunci aruncările sunt în mod evident independente:

$$P(A_1 \cap A_2 | Z_1) = \frac{1}{36} = P(A_1 | Z_1) \cdot P(A_2 | Z_1).$$

Din formula probabilității totale P.5 avem:

$$P(A_1) = P(A_1 | Z_1)P(Z_1) + P(A_1 | Z_2)P(Z_2) = \frac{1}{6} \cdot \frac{1}{2} + \frac{5}{6} \cdot \frac{1}{2} = \frac{1}{2},$$

$$P(A_2) = P(A_2|Z_1)P(Z_1) + P(A_2|Z_2)P(Z_2) = \frac{1}{6} \cdot \frac{1}{2} + \frac{5}{6} \cdot \frac{1}{2} = \frac{1}{2},$$

$$P(A_1 \cap A_2) = P(A_1 \cap A_2|Z_1)P(Z_1) + P(A_1 \cap A_2|Z_2)P(Z_2) = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{2} + \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{2} = \frac{13}{36}.$$

Deci $P(A_2|A_1) = \frac{P(A_1 \cap A_2)}{P(A_1)} = \frac{13}{18} \Rightarrow P(A_2|A_1) \neq P(A_2) \Rightarrow A_1$ și A_2 nu sunt independente. ✱

Def. 15. Fie X, Y, Z v.a. discrete, care iau valori în mulțimile $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. V.a. X este **condițional independentă** de Y , cunoscând (știind) v.a. Z , dacă pentru fiecare $x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}$, are loc

$$P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z).$$

Exemplu de clasificare naivă Bayes

Se dorește *clasificarea traficului* T pe un anumit bulevard, în *clasele*: aglomerat a sau relaxat r , în funcție de următoarele *atribute* cu valorile lor posibile:

- **vreme** V : ploaie p , zăpadă z , senin s , înnorat \hat{i} (dar nu plouă și nu ninge) ;
- **timp** Ti : dimineață di , amiază am , seară se , noapte no .

Considerăm evenimentul următor, denumit *vector de attribute*:

$$E = (V = p) \cap (Ti = am).$$

Se caută o clasă pentru E , stabilind care din următoarele probabilități este mai mare: $P(T = a|E)$ sau $P(T = r|E)$; aceasta este **metoda de probabilitate maximă**. Știind că vremea este ploioasă și este amiază, ce *previziune* se poate face despre trafic?

	Vreme	Timp	Trafic
1	înnorat	noapte	relaxat
2	zăpadă	seară	aglomerat
3	senin	noapte	relaxat
4	ploaie	seară	aglomerat
5	înnorat	amiază	aglomerat
6	senin	amiază	aglomerat
7	senin	dimineață	relaxat
8	ploaie	noapte	relaxat
9	înnorat	dimineață	aglomerat
10	zăpadă	noapte	aglomerat
11	senin	seară	relaxat
12	zăpadă	amiază	relaxat
13	înnorat	seară	aglomerat
14	ploaie	dimineață	aglomerat
15	zăpadă	dimineață	aglomerat

Tabel de date obținute în urma unor observații

Se face următoarea **presupunere naivă**: *atributele sunt condițional independente*, dacă se dă *clasificarea*, adică

$$(2) \quad P(V = v, Ti = ti|\mathbf{T} = \mathbf{t}) = P(V = v|\mathbf{T} = \mathbf{t})P(Ti = ti|\mathbf{T} = \mathbf{t}),$$

pentru fiecare $v \in \{p, z, s, \hat{i}\}$, $ti \in \{di, am, se, no\}$, $\mathbf{t} \in \{a, r\}$. De exemplu, avem:

$$P(V = p, Ti = di|\mathbf{T} = a) = P(V = p|\mathbf{T} = a)P(Ti = di|\mathbf{T} = a).$$

► Folosind datele din tabel, determinăm mai întâi probabilitățile claselor și probabilitățile condiționate ale atributelor, cunoscând clasa.

$\mathbf{T} = \mathbf{a}$	$\mathbf{T} = \mathbf{r}$	$P(\mathbf{T} = \mathbf{a})$	$P(\mathbf{T} = \mathbf{r})$
9	6	$\frac{9}{15}$	$\frac{6}{15}$

V	$\mathbf{T} = \mathbf{a}$	$\mathbf{T} = \mathbf{r}$	$P(V = \dots \mathbf{T} = \mathbf{a})$	$P(V = \dots \mathbf{T} = \mathbf{r})$
p	2	1	$\frac{2}{9}$	$\frac{1}{6}$
z	3	1	$\frac{3}{9}$	$\frac{1}{6}$
s	1	3	$\frac{1}{9}$	$\frac{3}{6}$
\hat{i}	3	1	$\frac{3}{9}$	$\frac{1}{6}$

Ti	$\mathbf{T} = \mathbf{a}$	$\mathbf{T} = \mathbf{r}$	$P(Ti = \dots \mathbf{T} = \mathbf{a})$	$P(Ti = \dots \mathbf{T} = \mathbf{r})$
di	3	1	$\frac{3}{9}$	$\frac{1}{6}$
am	2	1	$\frac{2}{9}$	$\frac{1}{6}$
se	3	1	$\frac{3}{9}$	$\frac{1}{6}$
no	1	3	$\frac{1}{9}$	$\frac{3}{6}$

► Pe baza formulei lui Bayes P. 5 și a ipotezei de independență condiționată, deducem că:

$$\begin{aligned}
 P(\mathbf{T} = \mathbf{a} | E) &= \frac{P(E | \mathbf{T} = \mathbf{a}) P(\mathbf{T} = \mathbf{a})}{P(E)} = \frac{P(V = p, Ti = am | \mathbf{T} = \mathbf{a}) P(\mathbf{T} = \mathbf{a})}{P(E)} \\
 &= \frac{P(V = p | \mathbf{T} = \mathbf{a}) P(Ti = am | \mathbf{T} = \mathbf{a}) P(\mathbf{T} = \mathbf{a})}{P(E)} = \frac{\frac{2}{9} \cdot \frac{2}{9} \cdot \frac{9}{15}}{P(E)} = \frac{1}{P(E)} \cdot \frac{4}{135}
 \end{aligned}$$

și

$$\begin{aligned}
 P(\mathbf{T} = \mathbf{r} | E) &= \frac{P(E | \mathbf{T} = \mathbf{r}) P(\mathbf{T} = \mathbf{r})}{P(E)} = \frac{P(V = p, Ti = am | \mathbf{T} = \mathbf{r}) P(\mathbf{T} = \mathbf{r})}{P(E)} \\
 &= \frac{P(V = p | \mathbf{T} = \mathbf{r}) P(Ti = am | \mathbf{T} = \mathbf{r}) P(\mathbf{T} = \mathbf{r})}{P(E)} = \frac{\frac{1}{6} \cdot \frac{1}{6} \cdot \frac{6}{15}}{P(E)} = \frac{1}{P(E)} \cdot \frac{1}{90}.
 \end{aligned}$$

Deoarece $P(\mathbf{T} = \mathbf{a} | E) > P(\mathbf{T} = \mathbf{r} | E)$, asociem vectorului de atribute

$$E = (V = p) \cap (Ti = am) \text{ clasa } \mathbf{T} = \mathbf{a}.$$

► În plus, putem determina $P(E) = P(V = p, Ti = am)$ astfel: Scriem

$$1 = P(\mathbf{T} = \mathbf{a} | E) + P(\mathbf{T} = \mathbf{r} | E) = \frac{1}{P(E)} \left(\frac{4}{135} + \frac{1}{90} \right)$$

și deducem $P(E) = P(V = p, Ti = am) = \frac{11}{270} \approx 0.04$.

★

Valoarea medie a unor variabile aleatoare discrete

Def. 16. Valoarea medie a unei variabile aleatoare discrete (numerice) X , care ia valorile $\{x_i, i \in I\}$, este

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

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

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

P. 9. Fie X și Y v.a. discrete. Au loc proprietățile:

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

→ $E(X + Y) = E(X) + E(Y)$;

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

→ Dacă $g : \mathbb{R} \rightarrow \mathbb{R}$ e o funcție astfel încât $g(X)$ este 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 $x = [x(1), \dots, x(n)]$, se calculează $\text{mean}(x) = \frac{1}{n}(x(1) + \dots + x(n))$

Exemplu: Joc: Se aruncă un zar; dacă apare 6, se câștigă 3 u.m. (unități monetare), dacă apare 1 se câștigă 2 u.m., dacă apare 2,3,4,5 se pierde 1 u.m. În medie cât va câștiga sau pierde un jucător după 30 de repetiții ale jocului?

Răspuns: Fie X v.a. care indică venitul la un joc

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

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

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

Așadar jucătorul câștigă în medie 5 u.m.


```
%simulare - Exemplul anterior
pkg load statistics
clear all
clc
v=0;
N=2000;
for i=1:N
s=sum(randsample([-1,-1,-1,-1,2,3],30,1)); %venitul dupa 30 de jocuri
% randsample([-1,-1,-1,-1,2,3],30,1) -> 30 de extrageri cu returnare
v=v+s;
endfor
fprintf('venitul mediu (dupa 30 de jocuri): %4.3f u.m. \n',v/N)
```

Exercițiu:

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

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

Întrebare: În medie câte iterații sunt necesare înainte să apară primul 0?

```
clear all
A=[zeros(1,50), zeros(1,70)+1,zeros(1,80)+2];
index=randperm(length(A));
A=A(index);
c=0;
i=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))
```

R.: Probabilitatea să apară la orice iterație 0 este $p = \frac{50}{200} = 0.25$.

Notăm cu X v.a. care indică numărul de iterații necesare *înainte* să apară primul 0

$\Rightarrow X \sim \text{Geo}(p)$.

Numărul mediu de iterații 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. 17. Fie X_1, \dots, X_n cu $n \in \mathbb{N}$, $n \geq 2$, variabile aleatoare discrete, care iau valori în mulțimile $\mathcal{X}_1, \dots, \mathcal{X}_n$. X_1, \dots, X_n sunt **variabile aleatoare independente**, dacă

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

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

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) $X_1 + X_2$ și $X_3 + X_4$ sunt v.a. independente;
- c) $X_1 + X_2 + X_3$ și X_4 sunt v.a. independente.

Def. 18. Funcția de repartiție $F : \mathbb{R} \rightarrow [0, 1]$ 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:

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

$\Rightarrow X$ are funcția de repartiție $F_X : \mathbb{R} \rightarrow [0, 1]$

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

P. 10. Funcția de repartiție F a unei variabile aleatoare discrete X are următoarele proprietăți:

- (1) $F(b) - F(a) = P(X \leq b) - P(X \leq a) = P(a < X \leq b) \quad \forall a, b \in \mathbb{R}, a < b$.
- (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 $\lim_{x \searrow x_0} F(x) = F(x_0) \quad \forall x_0 \in \mathbb{R}$.
- (4) $\lim_{x \rightarrow \infty} F(x) = 1$ și $\lim_{x \rightarrow -\infty} F(x) = 0$.

Matlab/Octave: `binocdf(x,n,p)`, `hygecdf(x,n1+n2,n1,n)`, `geocdf(x,p)` calculează $F(x) = P(X \leq x)$ pentru $X \sim \text{Bino}(n,p)$, $X \sim \text{Hyge}(n_1 + n_2, n_1, n)$, respectiv $X \sim \text{Geo}(p)$.

```
pkg load statistics
clear all
close all
% X urmeaza distributia 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 și nenumărabil de valori într-un interval sau reuniune de intervale (v.a. poate lua orice valoare din intervalul considerat);

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

▷ v.a. continuă este caracterizată de funcția de densitate.

Def. 19. *Funcția de densitate a unei v.a. continue X este funcția $f : \mathbb{R} \rightarrow \mathbb{R}$ pentru care are loc*

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

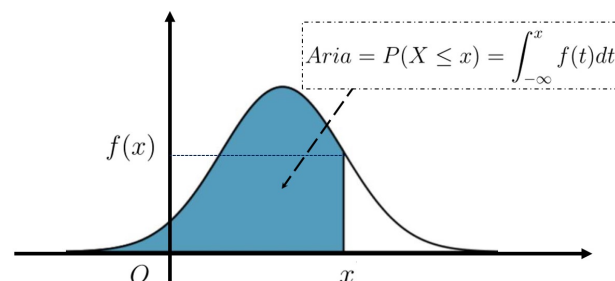
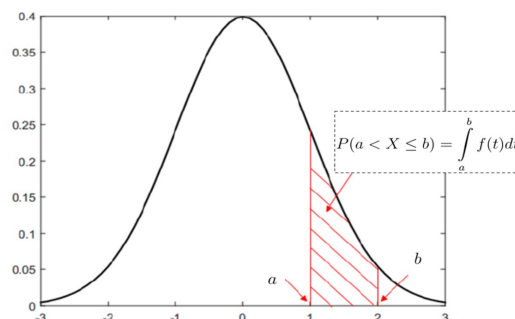
Funcția $F : \mathbb{R} \rightarrow [0, 1]$ definită prin

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

se numește **funcția de repartiție** a v.a. continue X .

P. 11. Fie f funcția de densitate și F funcția de repartiție a unei v.a. continue X . Au loc proprietățile:

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



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

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

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

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

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

$$(6) F \text{ este o funcție monoton crescătoare și continuă pe } \mathbb{R};$$

$$(7) \lim_{x \rightarrow \infty} F(x) = 1 \quad \text{și} \quad \lim_{x \rightarrow -\infty} F(x) = 0.$$

$$(8) \text{ dacă } F \text{ este derivabilă în punctul } x, \text{ atunci } F'(x) = f(x).$$

Observație: Orice funcție $f : \mathbb{R} \rightarrow \mathbb{R}$, care are proprietățile (1), (2) din **P.11** este o funcție de densitate.

Exemple de distribuții clasice continue

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

• funcția de densitate este

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

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)` returnează o matrice $M \times N$ cu valori aleatoare din $[a, b]$

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

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

• funcția de densitate este

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

• Pentru $\mu = 0, \sigma = 1$: $N(0, 1)$ se numește *distribuția standard normală*.

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



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

Matlab/Octave: `normrnd(μ, σ, M, N)` returnează o matrice $M \times N$ cu valori aleatoare;

▷ pentru $X \sim N(\mu, \sigma^2)$: `normpdf(x, μ, σ)` calculează $f(x)$, iar `normcdf(x, μ, σ)` calculează $F(x) = P(X \leq x)$.

➡ **Distribuția exponențială:** $X \sim \text{Exp}(\lambda), \lambda > 0$

- funcția 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($\frac{1}{\lambda}, M, N$)` returnează o matrice $M \times N$ cu valori aleatoare;

▷ pentru $X \sim \text{Exp}(\lambda)$: `exppdf($x, \frac{1}{\lambda}$)` calculează $f(x)$, iar `expcdf($x, \frac{1}{\lambda}$)` calculează $F(x) = P(X \leq x)$.

```
pkg load statistics
clear all
close all
figure
title('Funcția 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')
```

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

- distribuția Student cu $n \in \mathbb{N}^*$ grade de libertate are funcția 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 funcția Gamma este

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

Matlab/Octave: `trnd(n, M, N)` returnează o matrice $M \times N$ cu valori aleatoare;

▷ pentru $X \sim St(n)$: `tpdf(x, n)` calculează $f(x)$, iar `tcdf(x, n)` calculează $F(x) = P(X \leq x)$.

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

- distribuția χ^2 cu $n \in \mathbb{N}^*$ grade de libertate are funcția de densitate

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

Matlab/Octave: `chi2rnd(n, M, N)` returnează o matrice $M \times N$ cu valori aleatoare;

▷ pentru $X \sim \chi^2(n)$: `chi2pdf(x, n)` calculează $f(x)$, iar `chi2cdf(x, n)` calculează $F(x) = P(X \leq x)$.

Exemplu: Fie $X \sim Exp(0.5)$ v.a. care indică timpul de funcționare a unei baterii (câte luni funcționează bateria). Folosind simulări, să se estimeze a) $P(2 \leq X \leq 4)$; b) $P(X > 3)$ și să se compare rezultatele obținute 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
```

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

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



Matlab/Octave:

Distribuția v.a. discrete X	Generare valori aleatoare	Funcția de repartiție $F_X(x) = P(X \leq x)$	Probabilitate $P(X = x)$
$Bino(n, p)$	<code>binornd(n, p)</code>	<code>binocdf(x, n, p)</code>	<code>binopdf(x, n, p)</code>
$Unid(n)$	<code>unidrnd(n)</code>	<code>unidcdf(x, n)</code>	<code>unidpdf(x, n)</code>
$Hyge(n, n_1, n_2)$	<code>hygernd(n_1+n_2, n_1, n)</code>	<code>hygecdf(x, n_1+n_2, n_1, n)</code>	<code>hygepdf(x, n_1+n_2, n_1, n)</code>
$Geo(p)$	<code>geornd(p)</code>	<code>geocdf(x, p)</code>	<code>geopdf(x, p)</code>

Distribuția v.a. continue X	Generare valori aleatoare	Funcția de repartiție $F_X(x) = P(X \leq x)$	Funcția de densitate $f_X(x)$
$Unif[a, b]$	<code>unifrnd(a, b)</code>	<code>unifcdf(x, a, b)</code>	<code>unifpdf(x, a, b)</code>
$N(\mu, \sigma^2)$	<code>normrnd(μ, σ)</code>	<code>normcdf(x, μ, σ)</code>	<code>normpdf(x, μ, σ)</code>
$Exp(\lambda)$	<code>exprnd($\frac{1}{\lambda}$)</code>	<code>expcdf($x, \frac{1}{\lambda}$)</code>	<code>exppdf($x, \frac{1}{\lambda}$)</code>

Observație: Dacă în cadrul aceluiași 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.

Proprietăți

V.a. discretă	V.a. continuă
<ul style="list-style-type: none"> • caracterizată de distribuția de probabilitate discretă $X \sim \left(\begin{matrix} x_i \\ P(X = x_i) \end{matrix} \right)_{i \in I}$ <ul style="list-style-type: none"> • $\sum_{i \in I} P(X = x_i) = 1$ • $P(X \in A) = \sum_{i \in I: x_i \in A} P(X = x_i)$ • funcția de repartiție $F(x) = P(X \leq x) \forall x \in \mathbb{R}$ • $F(x) = \sum_{i \in I: x_i \leq x} P(X = x_i) \forall x \in \mathbb{R}$ • F este funcție continuă la dreapta • F este discontinuă în punctele $x_i, \forall i \in I$ • $\forall a < b, a, b \in \mathbb{R}$ $P(a \leq X \leq b) = \sum_{i \in I: a \leq x_i \leq b} P(X = x_i)$ • $P(X = a) = 0$ dacă $a \notin \{x_i : i \in I\}$ 	<ul style="list-style-type: none"> • caracterizată de funcția de densitate f $P(X \leq x) = \int_{-\infty}^x f(t) dt$ <ul style="list-style-type: none"> • $\int_{-\infty}^{\infty} f(t) dt = 1$ • $P(X \in A) = \int_A f(t) dt$ • funcția de repartiție $F(x) = P(X \leq x) \forall x \in \mathbb{R}$ • $F(x) = \int_{-\infty}^x f(t) dt \quad \forall x \in \mathbb{R}$ • F este funcție continuă în orice punct $x \in \mathbb{R}$ • $\forall a < b, a, b \in \mathbb{R}$ $P(a \leq X \leq b) = P(a < X < b)$ $= P(a \leq X < b) = P(a < X < b) = \int_a^b f(t) dt$ • $P(X = a) = \int_a^a f(t) dt = 0 \quad \forall a \in \mathbb{R}$ • dacă F este derivabilă în punctul x $\Rightarrow F'(x) = f(x)$.

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

Vector aleator continuu

Def. 20. (X_1, \dots, X_n) este un **vector aleator continuu** dacă fiecare componentă a sa este o variabilă aleatoare continuă.

Def. 21. $f_{(X,Y)} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$ este **funcția de densitate a vectorului aleator continuu** (X, Y) , dacă

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

Def. 22. $F_{(X,Y)} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$ este **funcția de repartiție a vectorului aleator** (X, Y) (discret sau continuu), dacă

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

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

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

Calculăm

$$F_{(X_1, X_2)}(3, 5) = P(X_1 \leq 3, X_2 \leq 5) = P(X_1 = -2, X_2 \in \{0, 3\}) = 0.7;$$

$$F_{(X_1, X_2)}(5, 2) = P(X_1 \leq 5, X_2 \leq 2) = P(X_1 \in \{-2, 4\}, X_2 = 0) = 0.6.$$



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

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

2. $F_{(X,Y)}$ este funcție continuă.

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

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

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

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

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

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

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

Ce distribuție au X , respectiv Y ?

R.: Se determină F_X, F_Y cu (3) și se calculează $f_X = F'_X, f_Y = F'_Y$; se obține $X \sim Unif[0, 1], Y \sim Unif[1, 2]$ (a se vedea expresia funcției de densitate a distribuției uniforme pe un interval pe pg. 35). ♣

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

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

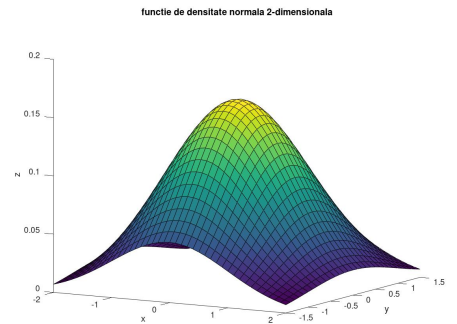
Exemplu pentru o distribuție normală bidimensională: (X, Y) are funcția de densitate (graficul acestei funcții este dat în figura alăturată)

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

$$\stackrel{(4)}{\Rightarrow} f_X(x) = \int_{-\infty}^{\infty} f_{(X,Y)}(x, y) dy = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad \forall x \in \mathbb{R},$$

$$\stackrel{(4)}{\Rightarrow} f_Y(y) = \int_{-\infty}^{\infty} f_{(X,Y)}(x, y) dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}, \quad \forall y \in \mathbb{R}.$$

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



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



```

%graficul unei functii de densitate normala bidimensionala
clear all
close all
figure(1)
hold on
f=@(x,y) (1/(2*pi))*exp(-(x.^2+y.^2)/2);
% functie de densitate normala 2-dimensională
x=-2:0.1:2;
y=-1.5:0.1:1.5;
view(30,10)
[xx,yy]=meshgrid(x,y);
ff=f(xx,yy);
surf(x,y,ff)
title('functie de densitate normala 2-dimensională')
xlabel('x')
ylabel('y')
zlabel('z')
figure(2)
view(30,10)
hold on
title('functie de densitate normala 2-dimensională / animatie')
xlabel('x')
ylabel('y')
zlabel('z')
for i=1:length(x)
for j=1:length(y)
plot3(x(i),y(j),f(x(i),y(j)),'r*')
pause(0.000001)
end
end
end

```

Variabilele aleatoare independente (discrete sau continue)

Def. 23. X_1, \dots, X_n sunt ***n* variabilele aleatoare independente** (discrete sau continue), dacă

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

Observație ($n = 2$ în definiția de mai sus): X_1 și 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 funcția de densitate f_{X_1}) și X_2 (cu funcția de densitate f_{X_2}) sunt independente, dacă și numai dacă

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


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

Exemplul 1: (X_1, X_2) are distribuție uniformă pe $I = [a_1, b_1] \times [a_2, b_2]$, cu $a_1, a_2, b_1, b_2 \in \mathbb{R}$, $a_1 < b_1, a_2 < b_2$ dacă

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

Cu (4) se calculează

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

$\implies X_1 \sim Unif[a_1, b_1], X_2 \sim Unif[a_2, b_2]$; se observă $f_{(X_1, X_2)} = f_{X_1} \cdot f_{X_2} \implies X_1$ și X_2 sunt v.a. independente! 

Exemplul 2: Fie (X, Y) vector aleator continuu, având funcția de repartiție

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

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

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

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

Deci, X și Y sunt v.a. independente.

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



Valoarea medie a unei variabile aleatoare continue

Def. 24. Valoarea medie a unei v.a. continue X , care are funcția de densitate f , este

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

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

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

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

→ $E(X + Y) = E(X) + E(Y)$;

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

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

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

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

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

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

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

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

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



Varianța unei variabile aleatoare

Def. 25. Varianța (dispersia) unei variabile aleatoare X (discrete sau continue) este

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

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

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

P. 15. *Proprietăți ale varianței (pentru v.a. discrete sau continue):*

$$\rightarrow V(X) = E(X^2) - E^2(X).$$

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

$$\rightarrow \text{Dacă } X \text{ și } Y \text{ sunt variabile aleatoare } \textbf{independente}, \text{ atunci } V(X + Y) = V(X) + V(Y).$$

Exemple: 1) Fie $X \sim \text{Bino}(n, p)$. Să se arate că $E(X) = np$ și $V(X) = np(1 - p)$.

R.: Pentru $i \in \{1, \dots, n\}$ fie $X_i \sim \text{Bernoulli}(p)$ (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 $X_1 + \dots + X_n \sim \text{Bino}(n, p)$. Deci, $X_1 + \dots + X_n$ și X au aceeași distribuție, așadar ele au aceeași valoare medie și aceeași varianță

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

V.a. X_1, \dots, X_n sunt independente și folosind P.15, obținem

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

2) Dacă $X \sim N(\mu, \sigma^2)$ să se arate că $E(X) = \mu$, $V(X) = \sigma^2$.

R.: Funcția de densitate a lui X este

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

Când $\mu = 0$ și $\sigma = 1$ obținem funcția de densitate a distribuției 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 $t = \frac{x - \mu}{\sigma}$

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

Folosind aceeași schimbare de variabilă și apoi integrare prin părți, avem

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

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

$$f_{(X,Y)} : \mathbb{R}^2 \rightarrow \mathbb{R} \quad f_{(X,Y)}(x, y) = \begin{cases} x - y, & \text{dacă } 0 \leq x \leq 1 \text{ și } -1 \leq y \leq 0 \\ 0, & \text{altfel.} \end{cases}$$

Să se calculeze $E(X)$ și $E(X^2)$.

R.:

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

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

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



► Matlab/Octave: `mean`, `var`, `std`

Fie $x = [x_1, \dots, x_n]$ valorile unei v.a. X

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

$mean(x) \approx E(X)$ 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. 26. $(X_n)_n$ este **șir de v.a. independente**, dacă $\forall \{i_1, \dots, i_k\} \subset \mathbb{N}$ v.a. X_{i_1}, \dots, X_{i_k} sunt independente, adică

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

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

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

b) Se aruncă o monedă

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

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

c) $X_n =$ v.a. care indică numărul apărut la al n -lea joc de ruletă

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



Def. 27. Șirul de v.a. $(X_n)_n$ **converge aproape sigur (a.s.)** la v.a. X , dacă

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

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

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

dacă $X_n \xrightarrow{\text{a.s.}} X$ atunci evenimentul

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

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

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

Definim

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

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

2) Fie $\Omega := [0, 1]$ spațiul de selecție, P probabilitatea pe $[0, 1]$ (care este numită măsura Lebesgue pe $[0, 1]$), adică pentru $\forall \alpha < \beta$ din $[0, 1]$ are loc

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

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

R.:

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

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

$$\begin{aligned} &\Rightarrow \{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = \omega\} = (0, 1) \\ &\Rightarrow P(\{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = \omega\}) = P((0, 1)) = 1. \\ &X_n \xrightarrow{\text{a.s.}} X. \end{aligned}$$

2b) $X_n(\omega) = (-1)^n \omega$, $\omega \in [0, 1]$, $n \geq 1$; converge $(X_n)_n$ a.s.?

R.: $(X_n)_n$ nu converge a.s. spre o v.a.; șirul $(X_n(\omega))_n$ este convergent doar în $\omega = 0$, iar $P(\{0\}) = 0$. ▲

Frecvențe relative și absolute (a se vedea Def.2): Fie A un eveniment asociat unei experiențe, repetăm experiența de n ori (în aceleași condiții date) și notăm cu r_n numărul de realizări ale evenimentului A ; **frecvența relativă** a evenimentului A este numărul

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

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

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

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

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

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 obținut se apropie tot mai mult de valoarea așteptată, cu cât experimentul se repetă de mai multe ori. Aceasta se explică prin faptul că abaterile aleatoare se compensează reciproc.

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

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



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

Def. 28. Șirul de v.a. $(X_n)_n$ cu $E|X_n| < \infty \forall n \in \mathbb{N}$ verifică **legea tare a numerelor mari (LTNM)** dacă

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

adică

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

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

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

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

Exemplu 1: Fie $X_1, \dots, X_n, \dots \sim Unid(6)$ 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 $\frac{1}{n} (X_1 + \dots + X_n) \xrightarrow{a.s.} 3.5$.
Simulare LTNM (Matlab/Octave):

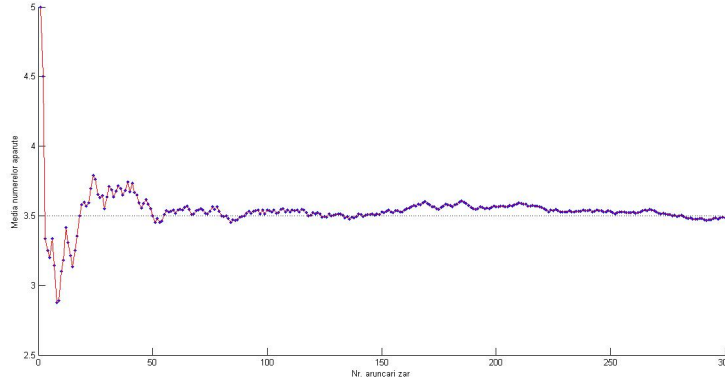


Fig. 4. Simulare LTNM

P. 17. Fie A un eveniment asociat unei experiențe, repetăm experiența de n ori (în aceleași condiții date și independent unele de altele). **LTNM:** cu cât repetăm mai des un experiment ($n \rightarrow \infty$), cu atât mai bine aproximează frecvența relativă $f_n(A)$ a evenimentului A probabilitatea sa teoretică de apariție $P(A)$:

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

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

Demonstrație pentru P.17: Aplicăm P.16 pentru șirul de v.a. independente $(X_n)_n$, unde

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

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

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

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

$$\text{Dar } \frac{1}{n}(X_1 + \dots + X_n) = f_n(A) \text{ (frecvența relativă a lui } A) \Rightarrow f_n(A) \xrightarrow{a.s.} P(A). \quad \square$$

Exemplu 2: Fie $(X_n)_n$ șir de v.a. independente, având aceeași distribuție ca v.a. X și varianță finită: $E(X_n) = E(X) \in \mathbb{R}$, $V(X_n) = V(X) \in \mathbb{R}$ pentru fiecare $n \in \mathbb{N}^*$.

Definim $Y_n = (X_n - E(X))^2 \quad \forall n \in \mathbb{N}^* \Rightarrow (Y_n)_n$ este șir de v.a. independente, având aceeași distribuție ca v.a. $(X - E(X))^2$ și $E(Y_n) = E((X - E(X))^2) = V(X) \quad \forall n \in \mathbb{N}^*$.

P.16 $\Rightarrow (Y_n)_n$ verifică **LTNM**

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

adică

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

Caz particular: Fie $X_1, \dots, X_n, \dots \sim Unid(6)$ v.a. independente; are loc $E(X_n) = \frac{1+2+3+4+5+6}{6} = 3.5$, $V(X_n) = E(X_n^2) - E^2(X_n) = \frac{35}{12} \approx 2.916 \forall n \geq 1$. Folosind P.16 rezultă că $(Y_n)_n = ((X_n - 3.5)^2)_n$ verifică **LTNM**, adică $\frac{1}{n} \left((X_1 - 3.5)^2 + \dots + (X_n - 3.5)^2 \right) \xrightarrow{a.s.} \frac{35}{12}$.

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

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

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

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

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

▲

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 nenumere) 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; mulţimea 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 colectivității să aibe aceeași șansă de a fi alese în eșantion;
- c) eșantionul să fie reprezentativ (structura eșantionului să fie apropiată de structura populației);
- d) volumul eșantionului să fie suficient de mare.

► *Unitatea statistică* (indivizii) este elementul, entitatea de sine stătătoare a unei populații statistice, care posedă o serie de trăsături caracteristice ce-i conferă apartenența la populația studiată. De exemplu: *unitatea statistică simplă*: un salariat, un student, un agent economic, o trăsătură, o părere; *unitatea statistică complexă*: o grupă de studenți sau o echipă de salariați, o familie sau o gospodărie, o categorie de mărfuri.

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

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

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

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

▷ caracteristicile de la a) și b) sunt variabile numerice (cantitative)

c) variabile (caracteristici) nominale (de ex.: culoarea ochilor, ramura de activitate, religia)

d) variabile (caracteristici) nominale ordinale (de ex.: starea de sănătate / calitatea unor servicii - precară, mai bună, bună, foarte bună)

e) variabile (caracteristici) dihotomiale (binare) (de ex.: stagiul militar - satisfăcut/nesatisfăcut, starea civilă - căsătorit/necăsătorit)

▷ caracteristicile de la c),d),e) sunt variabile calitative

▷ variabilele nominale mai sunt numite variabile categoriale

► *Datele statistice* reprezintă observațiile rezultate dintr-o cercetare statistică, sau ansamblul valorilor colectate în urma unei cercetări statistice.

De exemplu: un angajat al unei companii are o vechime de 6 ani în muncă. Angajatul reprezintă unitatea statistică, vechimea în muncă este caracteristica (variabila) cercetată, iar 6 este valoarea acestei caracteristici.

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

Distribuția caracteristicii X poate fi

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

2) specificată, dar depinzând de unul sau mai mulți parametri necunoscuți
(de ex.: $X \sim \text{Exp}(\lambda)$, $X \sim \text{Bin}(10, p)$, $X \sim N(m, \sigma^2)$)

3) necunoscută: $X \sim ?$

- în cazul 2) parametrii sunt necunoscuți, iar în cazul 3) distribuția este necunoscută

↪ se estimează → teoria estimăției / intervale de încredere

↪ se testează → teste statistice

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

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

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

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

$g(x_1, \dots, x_n)$ se numește valoarea funcției de selecție sau **valoarea estimatorului**.

• **Exemple de estimatori (funcții de selecție)** sunt: media de selecție, dispersia de selecție, momentul centrat de selecție de ordinul doi, funcția de repartiție empirică.

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

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

► **media de selecție (empirică)**

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

► **valoarea mediei de selecție**

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

► **varianța (dispersia) de selecție (empirică)**

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

► valoarea varianței (dispersiei) de selecție

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

► abaterea standard de selecție (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 selecție

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

► momentul centrat de selecție (empiric) de ordinul doi

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

► valoarea momentului centrat de selecție (empiric) de ordinul doi

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

► funcția de repartiție empirică $\mathcal{F}_n : \mathbb{R} \times \Omega \rightarrow [0, 1]$

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

► valoarea (expresia) funcției de repartiție empirice $\mathcal{F}_n : \mathbb{R} \rightarrow [0, 1]$

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

Def. 29. $g(X_1, \dots, X_n)$ este *estimator nedeplasat* pentru parametrul necunoscut θ , dacă

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

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

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

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

Observații:

1) Media de selecție \bar{X}_n este un estimator **nedeplasat și consistent pentru media teoretică** $E(X)$ a caracteristicii X ; în simulări $E(X) \approx \bar{x}_n$.

În Octave: `mean(d)`, unde $d=[x_1, \dots, x_n]$ este vectorul datelor statistice.

2) Varianța de selecție S_n^2 este un **estimator nedeplasat și consistent pentru varianța 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 selecție de ordinul doi M_n nu este un estimator nedeplasat pentru varianța teoretică $V(X)$ a caracteristicii X ; el este un **estimator consistent pentru varianța teoretică** $V(X)$ a caracteristicii X ; în simulări se folosește și $V(X) \approx m_n$.

În Octave: `var(d, 1)`, unde d este vectorul datelor statistice.

3) Deviația standard de selecție S_n nu este un estimator nedeplasat pentru deviația standard teoretică $Std(X) = \sqrt{V(X)}$ a caracteristicii X ; el este un **estimator consistent pentru deviația standard teoretică** $Std(X)$ a caracteristicii X ; în simulări se folosește $Std(X) \approx s_n$.

În Octave: `std(d)`, unde d este vectorul datelor statistice.

4) Funcția de repartiție de selecție $\mathcal{F}_n(x, \cdot)$ calculată în $x \in \mathbb{R}$ este un **estimator nedeplasat și consistent pentru** $F_X(x)$, care este valoarea funcției de repartiție 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 și `length(d)=n`.

```
% exemple de estimatori
pkg load statistics
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_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 $(X_n)_n$ șirul variabilelor de selecție pentru caracteristica cercetată $X \sim \text{Bernoulli}(p)$, unde $p \in (0, 1)$ este parametru necunoscut. Estimatorul

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

este un estimator *nedeplasat* și *consistent* pentru parametrul necunoscut p .

R.: $X \sim \text{Bernoulli}(p) \implies E(X) = p$;

$$\implies E\left(\hat{p}(X_1, \dots, X_n)\right) = \frac{1}{n}\left(E(X_1) + \dots + E(X_n)\right) = E(X) = p.$$

LTNM (a se vedea P.16) implică

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

Deci, $\hat{p}(X_1, \dots, X_n)$ este un estimator nedeplasat și consistent pentru parametrul necunoscut p .

Dacă $x_1, \dots, x_n \in \{0, 1\}$ sunt date statistice, atunci valoarea estimată pentru p este

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

◇

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

de exemplu:

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

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

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

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

Se rezolvă sistemul

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

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

Soluția sistemului $\hat{\theta}_1, \dots, \hat{\theta}_r$ sunt valorile estimate pentru parametrii necunoscuți $\theta_1, \dots, \theta_r$ ai distribuției 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, \dots, X_n variabilele de selecție. Avem cazul: $r = 1$, calculăm $E(X) = \frac{a}{2}$, $n = 10$, $\bar{x}_n = 0.468$. Se rezolvă

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

Valoarea estimatorului este

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

Estimatorul pentru parametrul necunoscut a este

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

Parametrul necunoscut a este estimat cu valoarea 0.936.

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

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



Exemplu 2:

Folosind metoda momentelor, să se estimeze parametrii necunoscuți $\theta_1 := \mu$ și $\theta_2 = \sigma^2$ pentru $X \sim N(\mu, \sigma^2)$; 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 $n = 10$, x_1, \dots, x_n sunt datele statistice, iar X_1, \dots, X_n sunt variabile de selecție. Avem cazul: $r = 2$, calculăm $E(X) = \mu$, $E(X^2) = V(X) + E^2(X) = \sigma^2 + \mu^2$ (a se vedea exemplul de pe p. 45), $\bar{x}_n = 0.1629$ (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

$$\begin{cases} \mu = \frac{1}{n} \sum_{i=1}^n x_i \\ \sigma^2 + \mu^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 \end{cases} \Rightarrow \text{are soluția} \begin{cases} \hat{\mu}(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i \\ \hat{\sigma}^2(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i \right)^2 \end{cases}$$

Valorile estimatorilor sunt

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

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

Estimatorii sunt

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

$$\hat{\sigma}^2(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2 = M_n$$

(momentul centrat de selecție de ordinul doi, a se vedea pe p.54).



Metoda verosimilității maxime pentru estimarea parametrului necunoscut θ al distribuției caracteristicii cercetate X

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

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

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

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

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

$\hat{\theta}$ este *punct de maxim global* pentru funcția de verosimilitate. Se rezolvă sistemul $\frac{\partial L}{\partial \theta} = 0$ și se arată că $\frac{\partial^2 L}{\partial \theta^2} < 0$.

Deseori este mai practic să se considere varianta transformată

$\frac{\partial \ln L}{\partial \theta} = 0$ cu $\frac{\partial^2 \ln L}{\partial \theta^2} < 0$. În unele situații (1) se rezolvă prin alte metode; de exemplu în cazul în care $\frac{\partial L}{\partial \theta} = 0$ nu are soluție (echivalent cu $\frac{\partial \ln L}{\partial \theta} = 0$ nu are soluție).

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

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

Se poate lucra și cu varianta transformată:

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

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

Reamintire: dacă $a, b > 0$, atunci au loc proprietățile:

$$\ln(a \cdot b) = \ln a + \ln b, \ln(a^b) = b \cdot \ln a, \ln\left(\frac{a}{b}\right) = \ln a - \ln b.$$

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

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

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

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

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

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

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

Estimatorul de verosimilitate maximă pentru parametrul necunoscut p este

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

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

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

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

Intervale de încredere și teste statistice

Noțiuni de bază

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

Def. 30. Cuantila de ordin α pentru distribuția caracteristicii cercetate X este numărul $z_\alpha \in \mathbb{R}$ pentru care

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

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

► dacă X este v.a. continuă, atunci: z_α este cuantilă de ordin $\alpha \iff P(X \leq z_\alpha) = \alpha \iff F_X(z_\alpha) = \alpha$

► dacă F_X este funcție inversabilă, atunci $z_\alpha = F_X^{-1}(\alpha)$

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

De exemplu, pentru $\alpha = 0.5$ și X v.a. continuă: 50% din valorile aleatoare ale lui X sunt mai mici sau egale cu $z_{0.5}$ (mediana), adică $P(X \leq z_{0.5}) = 0.5$.

• Matlab/Octave: `quantile` (pentru determinarea cuantilei dintr-un set de date)

```
clear all
pkg load statistics
N=1000;
x = normrnd(0,1,1,N);
alfa=[0.25 0.50 0.95];
z = quantile(x,alfa)
% rezultatul
% z = [-0.723985    0.052583    1.592022 ]
```

Exemplu: Fie $X \sim \begin{pmatrix} 1 & 3 & 5 & 7 \\ 0.2 & 0.35 & 0.35 & 0.1 \end{pmatrix}$ v.a. discretă

$\implies P(X < 3) = 0.2 \leq 0.5 \leq P(X \leq 3) = 0.2 + 0.35 = 0.55 \implies z_{0.5} = 3$ este mediana.

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

▷ **distribuția normală** $N(0, 1)$

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

cuantila $z_\alpha = \text{norminv}(\alpha, 0, 1)$ ▷ **distribuția Student** $St(n)$

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

cuantila $t_\alpha = \text{tinv}(\alpha, n)$ ▷ **distribuția Chi-pătrat** $\chi^2(n)$

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

cuantila $c_\alpha = \text{chi2inv}(\alpha, n)$ Exemple: $\text{norminv}(0.01, 0, 1) = -2.3263, \text{norminv}(1 - 0.01, 0, 1) = 2.3263,$

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

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

Proprietăți:

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

Exerciții (cu soluție): Să se arate că: **a)** $X \sim N(0, 1) \iff -X \sim N(0, 1)$;

b) pentru cuantilele distribuției normale $N(0, 1)$ are loc $z_\alpha = -z_{1-\alpha}$ pentru orice $\alpha \in (0, 1)$;

c) pentru cuantilele distribuției Student $St(n)$ are loc $t_\alpha = -t_{1-\alpha}$ pentru orice $\alpha \in (0, 1)$.

R.: a) Fie $x \sim N(0, 1)$. Scriem pentru orice $u \in \mathbb{R}$

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

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 și relația $X = -(-X)$, obținem că $-X \sim N(0, 1) \implies X \sim N(0, 1)$.

b) Fie $X \sim N(0, 1)$ și $z_\alpha, z_{1-\alpha}$ cuantile ale sale. Rezultă că

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

Scriem și folosim faptul că $-X$ și X urmează distribuția $N(0, 1)$

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

Pentru distribuția $N(0, 1)$ cuantila z_α e unic determinată din relația $P(X \leq z_\alpha) = \alpha$ (pentru că F_X e o funcție inversabilă și atunci $z_\alpha = F_X^{-1}(\alpha)$), așadar obținem că $z_\alpha = -z_{1-\alpha}$.

c) Raționamentul este analog. Se folosește $X \sim St(n) \iff -X \sim St(n)$. ♣

Intervale de încredere

În paragrafele anterioare s-a văzut cum poate fi estimat un parametru necunoscut, folosind datele dintr-un eșantion. 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 așteptare la un anumit ghișeu al unei bănci. Prin studierea unui eșantion de volum 200 s-a constatat că media de selecție a timpului de așteptare este $\bar{x}_{200} = 10$ (minute). Dacă considerăm un alt eșantion probabil obținem 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 eșantion acest interval aleator va deveni un interval numeric.

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

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

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

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

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

► $\left(g_1(x_1, \dots, x_n), g_2(x_1, \dots, x_n)\right)$ 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 $(g_1(X_1, \dots, X_n), g_2(X_1, \dots, X_n))$ este $1 - \alpha$ (nivelul de încredere)
- există și **intervale de încredere unilaterale**: $(-\infty, g_3(X_1, \dots, X_n))$, $(g_4(X_1, \dots, X_n), \infty)$, estimatorii g_3 și g_4 sunt astfel încât

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

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

% reprezentare intervale de incredere pentru media teoretica (figura urmatoare)

pkg load statistics

clear all

clc

close all

figure

hold on

a=0.05; %alfa

n=36;

t1=tinv(1-a/2,n-1);

disp('interval de incredere pt. media teoretica')

for i=1:25

x=normrnd(0,1,1,n);

%vectorul de date pt care se calculeaza intervalul de incredere

%poate fi si de ex. %x=unifrnd(-1,1,1,n);

xn=mean(x);

sn=std(x);

fprintf(' (%f,%f)\n', xn-sqrt(1/n)*sn*t1, xn+sqrt(1/n)*sn*t1)

% interval de incredere pentru medie cand varianta este necunoascuta

plot([xn-sqrt(1/n)*sn*t1, xn+sqrt(1/n)*sn*t1],[i,i])

endfor

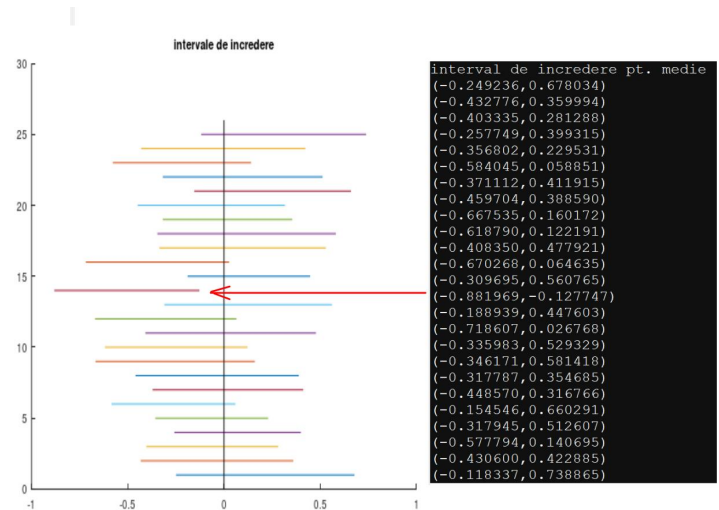
zx=zeros(1,27);

zy=[0:26];

plot(zx,zy,'k-')

title('intervale de incredere')

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



în această simulare: din 25 de intervale de încredere, un interval nu conține *valoarea reală* 0; parametrul necunoscut este $\theta = \text{media teoretică}$; datele statistice au fost generate, de fapt, cu `normrnd(0, 1)`, $1 - \alpha = 0.95$

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

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

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

➡ $F_{N(0,1)}(b) = \text{normcdf}(b, 0, 1)$ funcția de repartiție a legii normale standard $N(0, 1)$

➡ **Consecință** (la P. 18):

➤ $\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$ pentru $n > 30$ (n suficient de mare) și atunci

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

%Teorema limita centrala - exemplificare pentru distributia uniforma

pkg load statistics

close all

clc

N=1000;

x2=[];

m=mean([1:10]);

```

%media teoretica pentru distributia uniforma Unid(10_
sigma= std([1:10]);
%abaterea standard teoretica pentru distributia uniforma Unid(10)
for i=1:N
x1 = randi(10,N,1); % distributia uniforma
x2 = [x2, sqrt(N)*(mean(x1)-m)/sigma];
endfor
n1 = hist(x1,10);
figure
bar([1:10],n1/N) %histograma datelor unei distributii uniforme
figure
hold on
M=floor(sqrt(N)); % M=numar clase
[count,out]=hist(x2,M);
% frecventele absolute ale claselor in "count" si mijloacele claselor in "out"
b=out(2)-out(1); % lungimea unei clase
bar(out,count/(N*b),'hist');
hold on
v=min(x2):0.02:max(x2);
f=normpdf(v,0,1);
plot(v,f,'r*')
title('Teorema limita centrala')

```

Exemplul 1:

Dacă $(X_n)_{1 \leq n \leq 100}$ sunt variabile de selecție pentru caracteristica $X \sim \text{Bernoulli}(0.5)$, să se estimeze $P(0.35 < \bar{X}_{100} < 0.65)$, folosind P.18 (Teorema limită centrală).

R.: Se calculează $m = E(X_n) = E(X) = 0.5$ și $\sigma = \sqrt{V(X_n)} = \sqrt{V(X)} = 0.5$ și 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 și a consecinței de mai sus

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

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

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

Exemplul 2: Se știe că 40% din populația unui orașel susține un anumit candidat la alegerile viitoare. Dacă $(X_n)_{1 \leq n \leq 600}$ sunt variabile de selecție pentru distribuția $\text{Bernoulli}(0.4)$, adică $\forall n \in \{1, \dots, 600\}$

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

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

deci $X_n \sim \text{Bernoulli}(0.4)$. Folosind P.18 (Teorema limită centrală) estimați $P(\bar{X}_{600} > 0.43)$.
Calculați $E(\bar{X}_{600})$ și $V(\bar{X}_{600})$.

R.: Dacă $(X_n)_{1 \leq n \leq 600}$ sunt variabile de selecție pentru $\text{Bernoulli}(0.4)$, se calculează $m = E(X_n) = 0.4$ și $\sigma^2 = V(X_n) = 0.24 \forall n \in \mathbb{N}^*$ și se dorește estimarea probabilității

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

$$\begin{aligned} \text{Cf. P.18} \implies P(\bar{X}_{600} \leq 0.43) &= P\left(\frac{\bar{X}_{600} - 0.4}{\sqrt{\frac{0.24}{600}}} \leq \frac{0.43 - 0.4}{\sqrt{\frac{0.24}{600}}}\right) = P\left(\frac{\bar{X}_{600} - 0.43}{\sqrt{\frac{0.24}{600}}} \leq 1.5\right) \\ &\approx F_{N(0,1)}(1.5) = \text{normcdf}(1.5, 0, 1) = 0.93319 \end{aligned}$$

$$\implies P(\bar{X}_{600} > 0.43) \approx 0.066807.$$

$$E(\bar{X}_{600}) = \frac{1}{600} \left(E(X_1) + \dots + E(X_{600}) \right) = 0.4 \text{ și}$$

$$V(\bar{X}_{600}) = \frac{1}{600^2} \left(V(X_1) + \dots + V(X_{600}) \right) = \frac{1}{600} \cdot 0.24 = 0.0004. \quad \spadesuit$$

Exercițiu: 100 de zaruri sunt aruncate. Folosind P.18 (Teorema limită centrală), estimați probabilitatea ca suma numerelor obținute să fie între 300 și 400!

Recapitulare (notații)

Variabilele de selecție pentru caracteristica X X_1, \dots, X_n sunt v.a. independente, au aceeași distribuție ca X	datele statistice pentru caracteristica X x_1, \dots, x_n sunt valorile (numerice) ale v.a. X_1, \dots, X_n
Estimator	Valoarea estimatorului
media de selecție $\bar{X}_n = \frac{1}{n} (X_1 + \dots + X_n)$	valoarea mediei de selecție $\bar{x}_n = \frac{1}{n} (x_1 + \dots + x_n)$
varianța (dispersia) de selecție $S_n^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X}_n)^2$	valoarea varianței (dispersiei) de selecție $s_n^2 = \frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x}_n)^2$
abaterea standard de selecție $S_n = \left(\frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X}_n)^2 \right)^{\frac{1}{2}}$	valoarea abaterii standard de selecție $s_n = \left(\frac{1}{n-1} \sum_{k=1}^n (x_k - \bar{x}_n)^2 \right)^{\frac{1}{2}}$

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

Reamintim: $X \sim N(\mu, \sigma^2) \implies E(X) = \mu, V(X) = \sigma^2$ (a se vedea calculele de pe pg. 45).

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

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 selecție a calificativelor a 144 de elevi este 68. Dacă $\alpha = 0.05$, să se construiască un interval de încredere bilateral pentru valoarea medie (teoretică) $E(X)$ a punctajului obținut de un elev la test.

- se dau $\alpha \in (0, 1)$, σ , datele statistice x_1, \dots, x_n
- fie X_1, \dots, X_n variabilele de selecție corespunzătoare caracteristicii cercetate X
- construim intervale de încredere pentru parametrul necunoscut $m = E(X)$
- dacă $X \sim N(m, \sigma^2)$ sau $n > 30$ și X are o distribuție necunoscută, atunci P18, respectiv P.19,

implică

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

► cuantilele legii normale $N(0, 1)$:

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

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

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

deoarece:

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

$$\stackrel{(5)}{=} F_{N(0,1)}(z_{1-\frac{\alpha}{2}}) - F_{N(0,1)}(-z_{1-\frac{\alpha}{2}}) = F_{N(0,1)}(z_{1-\frac{\alpha}{2}}) - F_{N(0,1)}(z_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha$$

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

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

Interval de încredere pentru media teoretică $E(X)$ când dispersia $\sigma^2 = V(X)$ este cunoscută:

bilateral

unilateral

Expresia intervalului de încredere, folosind datele statistice

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

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

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

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 selecție a calificativelor a 144 de elevi este 68. Dacă $\alpha = 0.05$, să se construiască un interval de încredere bilateral pentru valoarea medie (teoretică) $E(X)$ a punctajului obținut de un elev la 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 $n = 144, \sigma = 10, \bar{x}_n = 68, \alpha = 0.05, z_{1-\frac{\alpha}{2}} = \text{norminv}(1 - \frac{0.05}{2}, 0, 1) \approx 1.96$. Pe baza datelor statistice valoarea intervalului de încredere bilateral este (66.367, 69.633). ♣

P. 20. Fie X_1, \dots, X_n variabile de selecție pentru $X \sim N(\mu, \sigma^2)$, atunci pentru **media de selecție** și **abaterea standard de selecție** are loc $\frac{\bar{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} \sim St(n-1)$, unde

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

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

Exemplu: Media de selecție a lungimii a 100 de șuruburi este 15.5 cm, iar varianța de selecție este 0.09 cm². Să se construiască un interval de încredere 99% bilateral pentru media (teoretică) a lungimii șuruburilor.

- se dau $\alpha \in (0, 1)$, datele statistice x_1, \dots, x_n
- fie X_1, \dots, X_n variabilele de selecție corespunzătoare caracteristicii cercetate X
- construim intervale de încredere pentru parametrul *necunoscut* $m = E(X)$
- dacă $X \sim N(m, \sigma^2)$ sau $n > 30$ și X are o distribuție necunoscută, atunci P.20 implică

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

- cuantilele legii Student $St(n-1)$:

$$t_{1-\frac{\alpha}{2}} = \text{tinv}(1 - \frac{\alpha}{2}, n-1), t_{1-\alpha} = \text{tinv}(1 - \alpha, n-1), t_{\alpha} = \text{tinv}(\alpha, n-1)$$

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

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

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

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

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

Exemplu: Media de selecție a lungimii a 100 de șuruburi este 15.5 cm, iar varianța de selecție este 0.09 cm². Să se construiască un interval de încredere 99% bilateral pentru media (teoretică) a lungimii șuruburilor.

R.: valoarea intervalului de încredere bilateral pentru media teoretică m , când varianța este necunoscută, este

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

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

P. 21. Fie X_1, \dots, X_n variabile de selecție pentru $X \sim N(\mu, \sigma^2)$, atunci pentru **varianța de selecție** are loc $\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2(n-1)$, unde $S_n^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X}_n)^2$.

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

Exemplu: Media de selecție a lungimii a 100 de șuruburi produse de o anumita firmă este 15.5 cm, iar varianța de selecție este 0.09 cm². Să se construiască un interval de încredere 99% bilateral pentru varianța (teoretică) a lungimii șuruburilor. Dacă varianța este prea mare (adică peste 0.099 cm²), aparatul, care produce șuruburile, trebuie reglat. Se presupune că lungimea unui șurub (produs de această firmă) are o distribuție normală.

- se dau $\alpha \in (0, 1)$, datele statistice x_1, \dots, x_n
- fie X_1, \dots, X_n variabilele de selecție corespunzătoare caracteristicii cercetate X
- construim intervale de încredere pentru parametrul *necunoscut* $\sigma^2 = V(X)$

► 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 distribuției $\chi^2(n-1)$ (Chi-pătrat cu $n-1$ grade de libertate):

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

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

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

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

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

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

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	$\left(0, \sqrt{\frac{n-1}{c_{\alpha}}} \cdot S_n \right)$ $\left(\sqrt{\frac{n-1}{c_{1-\alpha}}} \cdot S_n, \infty \right)$

Exemplul 1: Media de selecție a lungimii a 100 de șuruburi produse de o anumita firmă este 15.5 cm, iar varianța de selecție este 0.09 cm². Să se construiască un interval de încredere 99% bilateral pentru varianța (teoretică) a lungimii șuruburilor. Dacă varianța este prea mare (adică peste 0.099 cm²), aparatul, care produce șuruburile, trebuie reglat. Se presupune că lungimea

unui şurub (produs de această firmă) are o distribuţie normală.

R.: valoarea intervalului de încredere bilateral pentru varianţa teoretică este

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

unde $\bar{x}_n = 15.5$, $s_n^2 = 0.09$, $\alpha = 0.01$, $c_{1-\frac{\alpha}{2}} = \text{chi2inv}(0.995, 99) = 138.99$,
 $c_{\frac{\alpha}{2}} = \text{chi2inv}(0.005, 99) = 66.510$. Valoarea intervalului de încredere bilateral este
(0.064107, 0.133965). Acest interval conţine şi valori peste 0.099, deci aparatul, care produce
şuruburile, trebuie reglat! ♣

Exemplul 2: 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 ne-
cunoscută, 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 $\sqrt{n} = 8$, $\bar{x}_n = 525$, $s_n = 25$, $\alpha = 0.01$, $t_{1-\frac{\alpha}{2}} = \text{tinv}(0.995, 63) = 2.6561 \implies$ 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 $n = 64$, $s_n = 25$, $\alpha = 0.01$, $c_\alpha = \text{chi2inv}(0.01, 63) = 39.8551 \implies$
valoarea intervalului de încredere unilateral pentru abaterea standard este (0, 31.432). ♣

Teste statistice

Fie x_1, \dots, x_n datele statistice pentru caracteristica cercetată X , notăm cu X_1, \dots, X_n variabilele
de selecţie corespunzătoare.

- Ipoteza statistică este o presupunere relativă la un parametru necunoscut θ
- Metoda de stabilire a veridicităţii unei ipoteze statistice se numeşte test (criteriu de verificare).
- Rezultatul testării se foloseşte apoi pentru luarea unor decizii (cum ar fi: eficienţa unor medica-
mente, strategii de marketing, alegerea unui produs etc.).
- Se formulează ipoteza nulă H_0 şi ipoteza alternativă H_1 , privind parametrul θ ; fie θ_0 o valoare

dată

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

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

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

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

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

ceea ce este echivalent cu

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

Concluzia testului:

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

$(x_1, \dots, x_n) \in U \Rightarrow$ ipoteza H_0 este respinsă, în favoarea ipotezei H_1

► O colectivitate este testată în raport cu caracteristica X .

- test pentru valoarea medie teoretică $E(X)$

- ▷ când varianța teoretică $V(X)$ este cunoscută: testul lui Gauss (testul Z)

- ▷ când varianța teoretică $V(X)$ este necunoscută: testul Student (testul T)

- test pentru abaterea standard teoretică $\sqrt{V(X)}$ sau pentru varianța teoretică $V(X)$:
testul χ^2

Pașii pentru efectuarea unui test statistic:

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

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

► se dau $\alpha \in (0, 1)$, m_0 , σ , datele statistice x_1, \dots, x_n

► dacă $X \sim N(m, \sigma^2)$ sau $n > 30$ și X are o distribuție necunoscută, atunci P.18 și P.19 implică

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

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

► cuantilele legii normale $N(0, 1)$:

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

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

► în Octave/Matlab: `ztest`

```
x=normrnd(0,1,1,1000);
[a1,~,a2]=ztest(x,0,1,'tail','both','alpha',0.01) % cazul I
[b1,~,b2]=ztest(x,0,1,'tail','left','alpha',0.01) % cazul II
[c1,~,c2]=ztest(x,0,1,'tail','right','alpha',0.01) % cazul III
```

Observație: 1) Testele statistice și intervalele de încredere: Se observă că

I. $|z| < z_{1-\frac{\alpha}{2}} \iff \bar{x}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} < m_0 < \bar{x}_n + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$, adică m_0 (valoarea testată) aparține intervalului de încredere bilateral (se vedea tabelul de pe pg. 68) \iff se acceptă H_0

II. $z > z_{\alpha} \iff m_0 < \bar{x}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{\alpha}$, adică m_0 (valoarea testată) aparține intervalului de încredere unilateral (se vedea tabelul de pe pg. 68) \iff se acceptă H_0

III. $z < z_{1-\alpha} \iff \bar{x}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha} < m_0$, adică m_0 (valoarea testată) aparține intervalului de încredere unilateral (se vedea tabelul de pe pg. 68) \iff se acceptă H_0

2) *regiunea critică* $U \subset \mathbb{R}^n$ pentru testul mediei, când varianța este cunoscută, are următoarele expresii:

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

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

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

Exemplu: Un profesor a înregistrat pe parcursul mai multor ani rezultatele elevilor săi. Calificativul unui elev este o v.a. cu valoarea între 1 și 100, având abaterea standard egală cu 12. Actuala clasă are 36 de elevi și media calificativelor lor este 73.2. Se poate afirma din punct de vedere statistic că media calificativelor din actuala clasă este egală cu 73.5? ($\alpha = 0.05$)

R.: Se efectuează testul:

$H_0: m = 73.5$, $H_1: m \neq 73.5$, testul Z (Gauss) pentru medie, când varianța este cunoscută $\sigma^2 = 12^2$ (din textul problemei $\sigma = 12$).

Se calculează

$$z = \frac{\bar{x}_n - m_0}{\frac{\sigma}{\sqrt{n}}} = \frac{73.2 - 73.5}{\frac{12}{\sqrt{36}}} = -0.15 \implies |z| < z_{1-\frac{\alpha}{2}} = \text{norminv}(1 - \frac{\alpha}{2}, 0, 1) = 1.96$$

\implies se acceptă H_0 , adică se poate afirma pe baza datelor statistice, că media calificativelor din actuala clasă este egală cu 73.5. ♠

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

► se dau $\alpha \in (0, 1)$, m_0 , datele statistice x_1, \dots, x_n

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

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

► cuantilele legii Student cu $n-1$ grade de libertate $St(n-1)$:

$$t_{1-\frac{\alpha}{2}} = \text{tinv}(1 - \frac{\alpha}{2}, n-1), t_{1-\alpha} = \text{tinv}(1 - \alpha, n-1), t_\alpha = \text{tinv}(\alpha, n-1)$$

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

► în Octave/Matlab: `ttest`

```
x=normrnd(0,1,1,1000);
```

```
[a1,~,a2]=ttest(x,0,'tail','both','alpha',0.01) % cazul I
```

```
[b1,~,b2]=ttest(x,0,'tail','left','alpha',0.01) % cazul II
```

```
[c1,~,c2]=ttest(x,0,'tail','right','alpha',0.01) % cazul III
```

Observație: Se observă că

I. $|t| < t_{1-\frac{\alpha}{2}} \iff \bar{x}_n - \frac{s_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}} < m_0 < \bar{x}_n + \frac{s_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}$, adică m_0 (valoarea testată) aparține intervalului de încredere bilateral (se vedea tabelul de pe pg. 70) \iff se acceptă H_0

II. $t > t_\alpha \iff m_0 < \bar{x}_n - \frac{s_n}{\sqrt{n}} \cdot t_\alpha$, adică m_0 (valoarea testată) aparține intervalului de încredere unilateral (se vedea tabelul de pe pg. 70) \iff se acceptă H_0

III. $t < t_{1-\alpha} \iff \bar{x}_n - \frac{s_n}{\sqrt{n}} \cdot t_{1-\alpha} < m_0$, adică m_0 (valoarea testată) aparține intervalului de încredere unilateral (se vedea tabelul de pe pg. 70) \iff se acceptă H_0

Exemplu: Specificațiile unui anumit medicament indică faptul că fiecare comprimat conține în medie 2.4 g de substanță activă. 100 de comprimate alese la întâmplare din producție sunt analizate și se constată că ele conțin în medie 2.5 g de substanță activă cu o deviație standard de 0.2 g. Se poate spune că medicamentul respectă specificațiile (cu $\alpha = 0.01$)?

R.: $H_0: m = 2.4$ cu $H_1: m \neq 2.4$, testul Student.



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

► se dau $\alpha \in (0, 1)$, σ_0 , datele statistice x_1, \dots, x_n

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

► folosind datele statistice x_1, \dots, x_n se calculează $c = \frac{n-1}{\sigma_0^2} \cdot s_n^2$

► cuantilele χ^2 (Chi-pătrat) cu $n-1$ grade de libertate:

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

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

► în Matlab: `vartest`

```
x=normrnd(0,1,1,1000);
```

```
[a1,~,a2]=vartest(x,1,'tail','both','alpha',0.01) % cazul I
```

```
[b1,~,b2]=vartest(x,1,'tail','left','alpha',0.01) % cazul II
```

```
[c1,~,c2]=vartest(x,1,'tail','right','alpha',0.01) % cazul III
```

Observație: Se observă că

I. $c_{\frac{\alpha}{2}} < c < c_{1-\frac{\alpha}{2}} \iff \sqrt{\frac{n-1}{c_{1-\frac{\alpha}{2}}}} \cdot s_n < \sigma_0 < \sqrt{\frac{n-1}{c_{\frac{\alpha}{2}}}} \cdot s_n$, adică σ_0 (valoarea testată) aparține

intervalului de încredere bilateral (se vedea tabelul de pe pg. 71) \iff se acceptă H_0

II. $c > c_\alpha \iff \sigma_0 < \sqrt{\frac{n-1}{c_\alpha}} \cdot s_n$, adică σ_0 (valoarea testată) aparține intervalului de încredere unilateral (se vedea tabelul de pe pg. 71) \iff se acceptă H_0

III. $c < c_{1-\alpha} \iff \sqrt{\frac{n-1}{c_{1-\alpha}}} \cdot s_n < \sigma_0$, adică σ_0 (valoarea testată) aparține intervalului de încredere unilateral (se vedea tabelul de pe pg. 71) \iff se acceptă H_0

Exemplu: Un manager este suspicios că un utilaj, care umple anumite cutii cu ceai, trebuie înlocuit cu unul mult mai precis. 121 de cutii cu ceai sunt cântărite. S-a obținut o medie de 196.6 g și o abatere standard de 2.09 g pentru acest eșantion.

a) Pe baza datelor statistice se poate afirma că abaterea standard a utilajului este de 2 g?

b) Sunt datele suficiente pentru a concluziona, că utilajul trebuie reglat pentru că nu pune (în medie) 200 g de ceai într-o cutie? ($\alpha = 0.01$)

Să se folosească metoda intervalelor de încredere pentru a obține răspunsurile pentru aceste teste statistice.

R.: $n = 121$, $\bar{x}_n = 196.6$, $s_n = 2.09$, $\sigma_0 = 2$, $m_0 = 200$, $\alpha = 0.01$; vom folosi metoda intervalelor de încredere:

a) $H_0: \sigma = 2$ cu $H_1: \sigma \neq 2$, test pentru abaterea standard

$c_{1-\frac{\alpha}{2}} = \text{chi2inv}(1 - \frac{\alpha}{2}, n - 1)$, $c_{\frac{\alpha}{2}} = \text{chi2inv}(\frac{\alpha}{2}, n - 1)$; calculăm valoarea intervalului de încredere pentru abaterea standard: $\left(\sqrt{\frac{n-1}{c_{1-\frac{\alpha}{2}}}} \cdot s_n, \sqrt{\frac{n-1}{c_{\frac{\alpha}{2}}}} \cdot s_n \right) = (1.764015, 2.464349)$; cum $\sigma_0 = 2$ aparține acestui interval numeric, se acceptă H_0 : se poate afirma că abaterea standard a utilajului este de 2 g.

b) $H_0: m = 200$ cu $H_1: m \neq 200$, testul Student

$t_{1-\frac{\alpha}{2}} = \text{tinv}(1 - \frac{\alpha}{2}, n - 1)$; calculăm valoarea intervalului de încredere pentru medie (când varianța este necunoscută): $\left(\bar{x}_n - \frac{s_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}, \bar{x}_n + \frac{s_n}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}} \right) = (196.109828, 197.090172)$; cum $m_0 = 200$ nu aparține acestui interval numeric se respinge H_0 în favoarea lui H_1 . Utilajul trebuie reglat pentru că nu pune (în medie) 200 g de ceai într-o cutie! ♣

Test pentru independența a două caracteristici discrete X și Y

- fie $\alpha \in (0, 1)$ probabilitatea de risc
- fie X v.a., care are valorile posibile $\{a_1, \dots, a_r\}$ și Y v.a., care are valorile posibile $\{b_1, \dots, b_s\}$
- se dau datele statistice (x_i, y_j) , $i \in I_X, j \in I_Y$, corespunzătoare caracteristicii (X, Y) (I_X, I_Y sunt mulțimile de indici)
- fie (X_i, Y_j) , $i \in I_X, j \in I_Y$, perechile de variabile de selecție corespunzătoare caracteristicii (X, Y)

► ipoteza nulă și ipoteza alternativă

$$H_0 : X \text{ și } Y \text{ sunt independente} \quad H_1 : X \text{ și } Y \text{ nu sunt independente}$$

► se consideră estimatorii și valorile lor corespunzătoare

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

► din punct de vedere teoretic are loc

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

► practic, se calculează

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

și se determină cuantila de ordin $1 - \alpha$ a distribuției $\chi^2((r-1)(s-1))$, adică

$$c_{1-\alpha} = \text{chi2inv}(1 - \alpha, (r-1)(s-1))$$

► concluzia testului:

dacă $x < c_{1-\alpha}$, atunci se acceptă H_0

dacă $x \geq c_{1-\alpha}$, atunci se respinge H_0 în favoarea lui H_1 .

Exemplu: Se dau datele statistice referitoare la preferințele de vacanță ale bărbaților (B) și femeilor (F):

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

. Sunt preferințele de vacanță independente de gen (B,F)? (pentru $\alpha =$

0.05)

R.: test pentru independență; cele 2 caracteristici sunt

X: genul (valori posibile: B,F), $r = 2$;

Y: preferințele de vacanță (valori posibile: *plajă, munte*), $s = 2$.

▷ din tabel avem: $n_{11} = 209, n_{12} = 280, n_{21} = 225, n_{22} = 248$

⇒ $n_{1.} = 489, n_{.1} = 434, n_{2.} = 473, n_{.2} = 528, n = 962$

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

▷ are loc $\chi^2_{inv}(1 - 0.05, 1) = 3.8415 > x$, așadar se acceptă ipoteza H_0 , cele două caracteristici sunt independente, adică preferințele de vacanță nu depind de gen!



Erori în efectuarea testelor statistice

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

adică H_0 este respinsă deși este adevărată

de exemplu: se trage concluzia că un tratament este inefficient pe baza unor interpretări greșite (deși în realitate tratamentul este eficient)

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

adică H_0 nu este respinsă deși este falsă

de exemplu: nu este respins un tratament inefficient (deși în realitate tratamentul este inefficient);

Puterea unui test $= 1 - \beta = 1 - \text{probabilitatea apariției unei erori de tip II}$.

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

Analogie cu procedurile penale (realitatea: acuzatul este vinovat/nevinovat; se ia decizia: acuzatul este vinovat/nevinovat)

decizia \ acuzatul	vinovat	nevinovat
acuzatul este nevinovat	Eroare de tip I	decizie corectă
acuzatul este vinovat	decizie corectă	Eroare de tip II