Curs - Probabilităţi şi Statistică 2025/2026
Secţia Informatică

Facultatea de Matematică şi Informatică
Universitatea Babeş-Bolyai, Cluj-Napoca
Conf. Dr. Habil. Hannelore Lisei

Teoria Probabilităţilor

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

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

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

[Sursa: www.financialmarket.ro]

Fenomene şi procese aleatoare apar, de exemplu, în:

$\rightarrow$ pariuri, loto (6 din 49), jocuri de noroc / jocuri online
$\rightarrow$ previziuni meteo
$\rightarrow$ previziuni economice / financiare, investiţii, cumpărături online (predicţia comportamentului clienţilor)
$\rightarrow$ sondaje de opinie (analiza unor strategii politice), asigurări (evaluarea riscurilor / pierderilor)

$\rightarrow$ **în informatică**:
$\triangleright$ sisteme de comunicare, prelucrarea informaţiei, modelarea traficului în reţea, criptografie;
$\triangleright$ analiza probabilistică a performanţei unor algoritmi, fiabilitatea sistemelor, predicţii în cazul unor sisteme complexe;
$\triangleright$ algoritmi de simulare, machine learning, data mining, recunoaşterea formelor / a vocii;
$\triangleright$ generarea de numere aleatoare (pseudo-aleatoare), algoritmi aleatori
$\triangleright$ se pot genera numere cu „adevărat aleatoare” (*true random numbers*), folosind ca surse fenomene fizice, ca de exemplu surse radioactive (momentele de timp în care particulele se dezintegrează sunt complet imprevizibile), sau variaţiile de amplitudine din perturbările atmosferice (*atmospheric noise*, folosit de `https://www.random.org/randomness/`), sau comportamentul fotonilor (în mecanica cuantică, când un foton loveşte un separator de fascicule -*beam splitter*-, fotonul are şansa de 50% de a fi reflectat şi 50% de a trece); etc.

Exemplu: Generarea de valori aleatoare (în Python)

```
# Exemplu 1
import numpy as np
n=4
r = np.random.rand(n)
print(n,"valori aleatoare din intervalul (0,1):",r)
N = np.random.randint(-1,6,size=n+3)
print(n+3,"numere intregi aleatoare din intervalul [-1,5]:",N)
L = ["AB", "XY", "EF", "MN","FG"]
print(n,"-extrageri aleatoare cu returnare:",np.random.choice(L,size=n,replace=True))
print(n,"-extrageri aleatoare fara returnare:",np.random.choice(L,size=n,replace=
                                                     False))

# Exemplu 2
import numpy as np
n=30
R = np.random.randint(1,7,size=n)
print(n,"valori aleatoare:\n",R)
x= sum(R==2)
print("Rezultat ......",x)
```

Algoritmi aleatori

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

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

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

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

• Un algoritm aleator pentru care rezultatele obţinute sunt corecte *doar* cu o anumită probabilitate se numeşte algoritm **Monte Carlo**.
↪ se examinează probabilitatea cu care rezultatul este corect; probabilitatea de eroare poate fi scăzută semnificativ prin execuţii repetate, independente.
Exemplu:
▷ testul Miller-Rabin, care verifică dacă un număr natural este prim sau este număr compus; testul returnează fie răspunsul „numărul este sigur un număr compus” sau răspunsul „numărul este probabil un număr prim”.

Exerciţiu: Fie S un vector cu 60 de elemente, din mulţimea $\{0, 1, 2\}$ (ordinea lor este necunoscută; se presupune că şirul conţine cel puţin un 0).
$\longrightarrow$ De care tip este următorul algoritm?

```
import numpy as np
N=60
S = np.random.randint(0,3, size = N)
k=1
i= np.random.randint(low=0, high=N)
while S[i] != 0:
    print("iteratia:",k)
    print("S[",i,"]=",S[i])
    i= np.random.randint(low=0, high=N)
    k=k+1
if S[i]==0:
    print("iteratia:",k)
    print("S[",i,"]=",S[i])
print("S-a gasit aleator un 0.")
```

Răspuns: Algoritm de tip Las Vegas, algoritmul se încheie întotdeauna cu găsirea unui 0.

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

```
import numpy as np
print("a doua versiune")
N=50
S = np.random.randint(3,size=N)
print(S)
#un vector cu N  elemente, din multimea {0,1,2}
M=3  #nr maxim de iteratii  M>1
a=True
for k in range(M) :
    print("iteratia:",k+1)
    i= np.random.randint(low=0, high=N)
    print("S[",i,"]=",S[i])
    if S[i] == 0:
        print("la iteratia",k+1,"s-a gasit aleator un 0.")
        a=False
        break
if a:
    print("In",k+1,"iteratii nu s-a gasit niciun 0.")
```

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

Noţiuni introductive:

• Experienţa aleatoare (experimentul aleator) este acea experienţă al cărei rezultat nu poate fi cunoscut decât după încheierea ei.
• Evenimentul este rezultatul unei experienţe aleatoare.

Exemple:

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

▷ experienţa: extragerea unei cărţi de joc, eveniment: s-a extras un as
▷ experienţa: extragerea unui număr la loto, eveniment: s-a extras numărul 27
• evenimentul imposibil, notat cu $\emptyset$, este evenimentul care nu se realizează niciodată la efectuarea experienţei aleatoare
• evenimentul sigur este un eveniment care se realizează cu certitudine la fiecare efectuare a experienţei aleatoare
• spaţiul de selecţie, notat cu Ω, este mulţimea tuturor rezultatelor posibile ale experienţei considerate

◇ spaţiul de selecţie poate fi finit sau infinit

• dacă A este o submulţime a lui Ω atunci A se numeşte eveniment aleator, iar dacă A are un singur element atunci A este un eveniment elementar.
▷ O *analogie între evenimente şi mulţimi* permite o scriere şi o exprimare mai comode ale unor idei şi rezultate legate de conceptul de eveniment aleator.

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

Operaţii cu evenimente

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

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

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

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

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

• Au loc relaţiile: $A \cup \bar{A} = \Omega, \quad A \cap \bar{A} = \emptyset, \quad \bar{\bar{A}} = A.$

Relaţii între evenimente

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

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

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

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

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

asociative

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

şi **distributive**

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

satisfac **legile lui De Morgan**

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

Frecvenţa relativă şi frecvenţa absolută

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

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

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

Definiţia clasică a probabilităţii

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

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

► Prin repetarea de multe ori a unui experiment, în condiţii practic identice, frecvenţa relativă $f_n(A)$ de apariţie a evenimentului A este aproximativ egală cu $P(A)$: $f_n(A) \approx P(A)$ pentru valori mari ale lui n.
► Din punct de vedere probabilistic şirul $\big(f_n(A)\big)_n$ „converge aproape sigur” către $P(A)$ când $n \to \infty$.

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

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

R.: $f_n(A) = \frac{22}{100} = 0.22$ este frecvenţa relativă a evenimentului A;

$P(A) = \frac{4}{2^4} = 0.25$ probabilitatea (teoretică) a evenimentului A. ♠

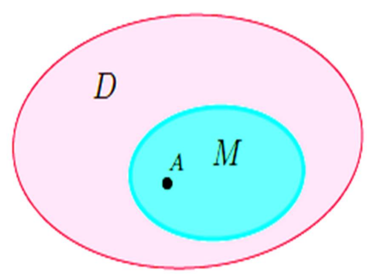
Exerciţiu: (1) Se alege aleator un număr din mulţimea $\{1, 2, 3, ...99\}$. Care este probabilitatea ca acesta să nu fie divizibil nici cu 4, nici cu 6?
(2) Un centru de calcul dispune de 24 de servere:
▷ 10 servere sunt rezervate pentru baze de date,
▷ 8 servere sunt pentru aplicaţii web,
▷ 6 servere sunt dedicate sarcinilor de învăţare automată.
Un nou proces este atribuit aleator unuia dintre servere.
▷ Care este probabilitatea ca procesul să nu ruleze pe un server de baze de date?
▷ Care este probabilitatea ca procesul să ruleze pe un server web sau de învăţare automată? ♢

Definiţia axiomatică a probabilităţii

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

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

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



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

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

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

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

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

(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* $\bigcup_{n=1}^{\infty} A_n \in \mathcal{K}$.

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

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

este o σ-algebră pe mulţimea B. ♢

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

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

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

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

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

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

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

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

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

Tripletul $(\Omega, \mathcal{K}, P)$ *se numeşte* ***spaţiu de probabilitate***.

Exemplu: 1) Cea mai simplă (funcţie de) probabilitate se obţine pentru cazul unui *spaţiu de selecţie finit* Ω: fie $\mathcal{K} = \mathcal{P}(\Omega)$ (mulţimea tuturor submulţimilor lui Ω) şi $P : \mathcal{K} \to \mathbb{R}$ 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 şi corespunde *definiţiei clasice a probabilităţii unui eveniment* (a se vedea Def. 3).
2) Fie $\Omega = \mathbb{N} = \{0, 1, 2, \ldots\}$, $\mathcal{K} = \mathcal{P}(\mathbb{N})$ şi $P : \mathcal{K} \to \mathbb{R}$ definită prin
$P(\{n\}) = \frac{1}{2^{n+1}}$, $n \in \mathbb{N}$
$P(\{n_1, ..., n_k, ...\}) = \frac{1}{2^{n_1+1}} + ... + \frac{1}{2^{n_k+1}} + ...$, unde $\{n_1, ..., n_k, ...\} \subseteq \mathbb{N}$.

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 spaţiu 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 şi nici suma ei. ♣

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

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

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

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

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

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

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

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

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

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

♡

Evenimente independente

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

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

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

Exerciţiu: Se aruncă un zar de două ori.

A: primul număr este 6; B: al doilea număr este 5; C: primul număr este 1.

Sunt A şi B evenimente independente? Sunt A şi B evenimente disjuncte?

Sunt A şi C evenimente independente? Sunt A şi C evenimente disjuncte? ♣

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

(1) A *şi* B *sunt independente.*

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

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

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

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

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

pentru orice submulţime finită $\{i_1, \dots, i_m\} \subseteq \{1, 2, ..., n\}$*, unde* $m \geq 2$*.*

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

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

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

Exemplu: 1) Din Def. 6 şi Def. 7 deducem că, independenţa (în totalitate) implică şi independenţa a două câte două evenimente. Afirmaţia inversă, însă, nu are loc. Drept (contra)exemplu putem lua experimentul aleator ce constă în aruncarea unui tetraedru regulat, ale cărui patru feţe sunt vopsite astfel: una este roşie, una este albastră, una este

verde şi una este colorată având cele trei culori. Se aruncă tetraedrul şi se consideră evenimentele:
R: tetraedrul cade pe o parte ce conţine culoarea roşie;
A: tetraedrul cade pe o parte ce conţine culoarea albastră;
V: tetraedrul cade pe o partea ce conţine culoarea verde.
Sunt cele 3 evenimente *independente în totalitate*?
2) Pentru a verifica dacă n evenimente distincte $B_1, \ldots, B_n$ sunt independente în totalitate câte relaţii trebuie verificate?
3) O firmă utilizează două sisteme de securitate independente pentru a detecta activitatea suspectă a reţelei: un firewall care detectează o astfel de activitate cu o probabilitate de 0.7 şi un antivirus care o detectează cu o probabilitate de 0.8.
Presupunând că firewall-ul i antivirusul funcţionează independent, care este probabilitatea ca:
(a) Ambele sisteme detectează activitatea suspectă?
(b) Cel puţin un sistem detectează activitatea suspectă? ♦

Exemplu istoric - Joc de zaruri (sec. XVII): Un pasionat jucător de zaruri, cavalerul de Méré, susţinea în discuţiile sale cu B. Pascal că a arunca un zar de 4 ori pentru a obţine cel puţin o dată faţa şase, este acelaşi lucru cu a arunca de 24 ori câte două zaruri pentru a obţine cel puţin o dublă de şase. Cu toate acestea, cavalerul de Méré a observat că jucând în modul al doilea (cu două zaruri aruncate de 24 ori), pierdea faţă de adversarul său, dacă acesta alegea primul mod (aruncarea unui singur zar de 4 ori). Pascal şi Fermat au arătat că probabilitatea de câştig la jocul cu un singur zar aruncat de 4 ori este $p_1 \approx 0.5177$, iar probabilitatea $p_2 \approx 0.4914$ la jocul cu două zaruri aruncate de 24 de ori. Deşi diferenţa dintre cele două probabilităţi este mică, totuşi, la un număr mare de partide, jucătorul cu probabilitatea de câştig p_1 câştigă în faţa jucătorului cu probabilitatea de câştig p_2. Practica jocului confirmă astfel justeţea raţionamentului matematic, contrar credinţei lui de Méré.
Estimăm prin simulări Python probabilităţile următoarelor evenimente:
A: se obţine cel puţin un 6 în 4 aruncări ale unui zar;
B: se obţine cel puţin o pereche (6,6) în 24 de aruncări a două zaruri;
C: se obţine cel puţin o pereche (6,6) în 25 de aruncări a două zaruri.

Calculăm probabilităţile teoretice pentru evenimentele A, B, C: $\bar{A}$ este evenimentul că niciun 6 nu apare în 4 aruncări ale unui zar

$$\Longrightarrow P(\bar{A}) = \left(\frac{5}{6}\right)^4 \Longrightarrow P(A) = 1 - \left(\frac{5}{6}\right)^4 \approx 0.5177.$$

$\bar{B}$ este evenimentul că nicio pereche $(6,6)$ nu apare în 24 de aruncări a două zaruri

$$\Longrightarrow P(\bar{B}) = \left(\frac{35}{36}\right)^{24} \Longrightarrow P(B) = 1 - \left(\frac{35}{36}\right)^{24} \approx 0.4914\,.$$

Analog $P(C) = 1 - \left(\frac{35}{36}\right)^{25} \approx 0.5055$. Comparăm probabilităţile teoretice ale celor trei evenimente

$$P(B) < \frac{1}{2} < P(C) < P(A).$$

Concluzie: Evenimentul A are şansele cele mai mari de câştig. $\diamondsuit$

```
import random
import numpy
a=0
N=10000
for _ in range(N):
    x=random.choices([1,2,3,4,5,6],k=4) # alegere aleatoare cu returnare
    a=a+(x.count(6)>0)
print("din simulari P(A) este:",a/N)
b=0
for _ in range(N):
    x1=random.choices([1,2,3,4,5,6],k=24)
    x2=random.choices([1,2,3,4,5,6],k=24)
    s=numpy.add(x1,x2)
    b=b+(sum(s==12)>0)
print("din simulari P(B) este:",b/N)
c=0
for _ in range(N):
    y1=random.choices([1,2,3,4,5,6],k=25)
    y2=random.choices([1,2,3,4,5,6],k=25)
    s=numpy.add(y1,y2)
    c=c+(sum(s==12)>0)
print("din simulari P(C) este:",c/N)
X=[a,b,c]
str="ABC"
z=sorted([a,b,c])
i0= X.index(z[0]) # index din X pt care este probabilitatea cea mai mica
i1= X.index(z[1])
i2= X.index(z[2]) # index din X pt care este probabilitatea cea mai mare
print("P(",str[i0],")< P(",str[i1],")< P(",str[i2],")")
# probabilitatile evenimentelor afisate in ordine crescatoare
```

Probabilitate condiţionată

În anumite situaţii este necesar să cunoaştem probabilitatea unui eveniment particular, care urmează să aibă loc, ştiind deja că alt eveniment a avut loc.

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

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

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

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

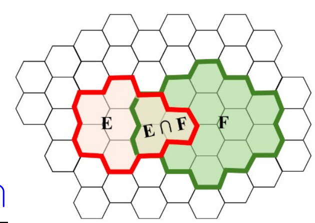
$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

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

➤ $P(A|B)$: probabilitatea condiţionată a lui A de către B, este **probabilitatea de a se realiza evenimentul A dacă în prealabil s-a realizat evenimentul B**.
➤ Fie evenimentele $A, B \in \mathcal{K}$ astfel încât $P(A) > 0$ şi $P(B) > 0$. Evenimentele A şi B sunt **independente** (a se vedea Def. 6), dacă apariţia evenimentului A, nu influenţează apariţia evenimentului B şi invers, adică

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

➤ Într-un experiment în care cazurile posibile sunt finite la număr şi au aceleaşi şanse de a se realiza, atunci
se poate folosi



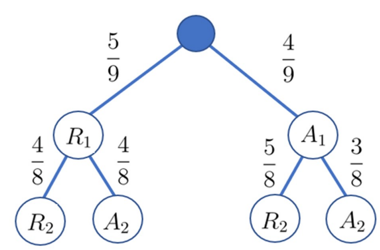
$P(E) = \frac{8}{50} \approx 0.16$

$P(E|F) = \frac{3}{14} \approx 0.21$

$$P(E|F) = \frac{\text{numărul de cazuri favorabile apariţiei lui } E \cap}{\text{numărul de cazuri favorabile pentru apariţia lui } F}$$

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

a) Ştiind că prima bilă este roşie, care este probabilitatea (condiţionată) ca a doua bilă să fie albă?
b) Care este probabilitatea ca ambele bile să fie roşii?
R.: pentru $i \in \{1, 2\}$ fie evenimentele
R_i: la a i-a extragere s-a obţinut o bilă roşie;
$A_i = \bar{R}_i$: la a i-a extragere s-a obţinut o bilă albă;
a) $P(A_2|R_1) = \frac{4}{8}$.
b) $P(R_1 \cap R_2) = P(R_2|R_1)P(R_1) = \frac{4}{8} \cdot \frac{5}{9}$. ♣

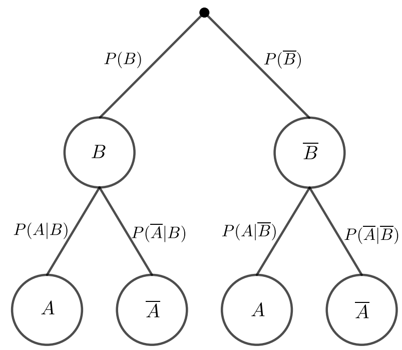


Extragere fără retrunare

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

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

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



Probabilităţi condiţionate

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

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

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

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

Exemplu: Într-o urnă sunt 7 bile albe, notate cu 1, 2, 3, 4, 5, 6, 7, şi 6 bile roşii notate cu 8, 9, 10, 11, 12, 13. Se extrag succesiv fără returnare două bile. **a)** Ştiind că prima bilă extrasă este roşie, care este probabilitatea p_1, ca numărul de pe bilă să fie divizibil cu 4? **b)** Ştiind că prima bilă este roşie, care este probabilitatea p_2, ca o a doua bilă extrasă să indice un număr impar?

R.: 7 bile albe: 1, 2, 3, 4, 5, 6, 7;   6 bile roşii: 8, 9, 10, 11, 12, 13.

Se consideră evenimentele:

D_1: prima bilă extrasă are înscris un număr divizibil cu 4;

R_1: prima bilă extrasă este roşie;

I_1: prima bilă extrasă are înscris un număr impar;

I_2: a doua bilă extrasă are înscris un număr impar.

a) $p_1 = P(D_1|R_1) = \frac{2}{6}$.

b) $p_2 = P(I_2|R_1) =$? Folosim Def.8 şi P.4, scriem succesiv

$$p_2 = P(I_2|R_1) = \frac{P(I_2 \cap R_1)}{P(R_1)} = \frac{P(I_2 \cap R_1 \cap I_1) + P(I_2 \cap R_1 \cap \bar{I}_1)}{P(R_1)}$$

$$= \frac{P(I_2|R_1 \cap I_1)P(R_1 \cap I_1) + P(I_2|R_1 \cap \bar{I}_1)P(R_1 \cap \bar{I}_1)}{P(R_1)} = \frac{\frac{6}{12} \cdot \frac{3}{13} + \frac{7}{12} \cdot \frac{3}{13}}{\frac{6}{13}} = \frac{13}{24}\,.$$

♡

Exemplu: Ce probabilităţi calculează programul de mai jos?

(?) Care sunt valorile teoretice pentru `p1`, `p2`, `p3`, din acest exemplu?

```
import random; import numpy
c1,c2,a1,a2=0,0,0,0
N=10000
A= list(range(1,21))
for _ in range(N):
 i=numpy.random.randint(len(A))
 v=A[i]
 c1=c1+(v%2)
 c2=c2+((v%2)==0)
 a1=a1+(v%2)*((v%3)==0);
 a2=a2+ ((v%2)==0)*(6<=v and v<=10)
p1=a1/c1
p2=a2/c2
p3=c1/N
print(f"p1={p1:.6f}")
print(f"p2={p2:.6f}")
print(f"p3={p3:.6f}")
```

♣

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

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

Observaţie: 1) Formula înmulţirii probabilităţilor 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 $A_1, \ldots, A_n$ sunt *independente în totalitate*, atunci formula înmulţirii probabilităţilor are forma

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2) \ldots P(A_n).$$

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

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

b) cele 2 bile să aibă aceeaşi culoare?

c) a doua bilă să fie albastră?

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

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

R.: Notăm pentru $i \in \{1, 2, 3\}$ evenimentele:
A_i: la a i-a extragere s-a obţinut bilă albastră; V_i: la a i-a extragere s-a obţinut bilă verde;
a) folosim P.4: $P(V_1 \cap A_2) = P(A_2|V_1)P(V_1) = \frac{3}{4} \cdot \frac{2}{5}$

b) $P((V_1 \cap V_2) \cup (A_1 \cap A_2)) = P(V_1 \cap V_2) + P(A_1 \cap A_2) = P(V_2|V_1)P(V_1) + P(A_2|A_1)P(A_1) = \frac{1}{4} \cdot \frac{2}{5} + \frac{2}{4} \cdot \frac{3}{5}$

c) folosim formula probabilităţii totale P.7:
$P(A_2) = P(A_2|V_1)P(V_1) + P(A_2|A_1)P(A_1) = \frac{3}{4} \cdot \frac{2}{5} + \frac{2}{4} \cdot \frac{3}{5}$

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

e) formula de înmulţire a probabilităţilor P.6:
$P(V_1 \cap A_2 \cap A_3) = P(V_1) \cdot P(A_2|V_1) \cdot P(A_3|V_1 \cap A_2) = \frac{2}{5} \cdot \frac{3}{4} \cdot \frac{2}{3}$.

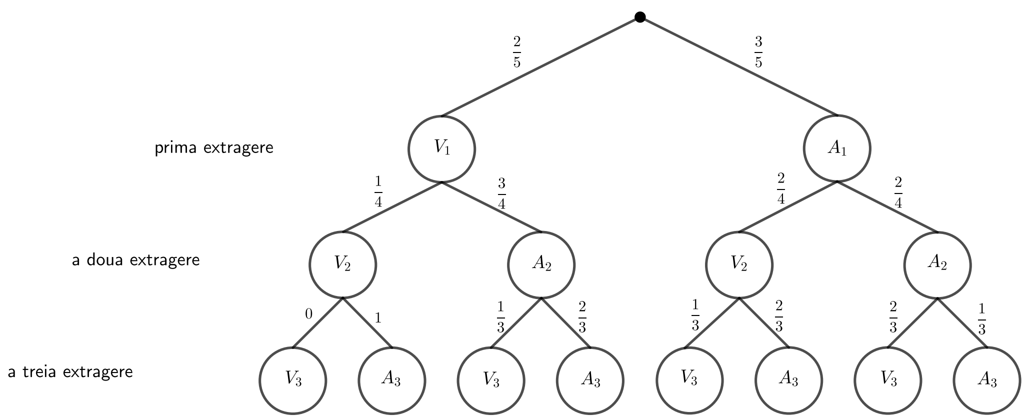


Fig. 3. Extragere fără returnare

♣

Formula lui Bayes

Formula lui Bayes este o metodă de a „corecta” (a revizui, a îmbunătăţi) pe baza unor noi date (informaţii) disponibile o probabilitate determinată apriori.

▶ Se porneşte cu o estimare pentru probabilitatea unei anumite ipoteze H (engl. *hypothesis*).

$\hookrightarrow$ $P(H)$ probabilitatea ca ipoteza H să fie adevărată, numită şi ***probabilitatea apriori***

▶ Dacă avem noi date (date de antrenare, dovezi, informaţii, evidenţe - engl. *evidence*) E, ce privesc ipoteza H, se poate calcula o probabilitate „corectată” pentru ipoteza H, numită probabilitate posterioară (a-posteriori)

$\hookrightarrow$ probabilitatea condiţionată $P(H|E)$ este ***probabilitatea posterioară*** (corectată de cunoaşterea noilor date / informaţii / evidenţe);

▶ Se cunosc:

$\hookrightarrow$ $P(E|H)$ probabilitatea ca să apară datele, ştiind că ipoteza H este adevarată;

$\hookrightarrow$ $P(E|\bar{H})$ probabilitatea ca să apară datele, ştiind că ipoteza H este falsă

acestea reprezintă verosimilitatea (engl. *likelihood*) datelor observate (a informaţiilor / evidenţelor).

Exemplu: Un clasificator de emailuri este conceput pentru a detecta mesajele `spam`. Fiecare email este clasificat într-una dintre cele două categorii:

- S: un email este `spam`
- $\bar{S}$: un email nu este `spam`

C: un email conţine cuvântul *succes*.

Se cunosc probabilităţile

$$P(S) = 0.2, \text{ deci } P(\bar{S}) = 0.8 \text{ (probabilităţile apriori)},$$

$$P(C|S) = 0.7, \; P(C|\bar{S}) = 0.1.$$

Care este probabilitatea ca un email să fie `spam`, ştiind că emailul conţine cuvântul *succes*?

R: Scriem succesiv

$$P(S|C) = \frac{P(S \cap C)}{P(C)} = \frac{P(C|S) \cdot P(S)}{P(C)} = \frac{P(C|S) \cdot P(S)}{P(C|S) \cdot P(S) + P(C|\bar{S}) \cdot P(\bar{S})}.$$

Calculăm (folosim P.5 cu partiţia $\{S, \bar{S}\}$)

$$P(C) = P(C|S) \cdot P(S) + P(C|\bar{S}) \cdot P(\bar{S}) = 0.7 \cdot 0.2 + 0.1 \cdot 0.8) = 0.22$$

$$\Longrightarrow P(S|C) = \frac{0.7 \cdot 0.2}{0.22} = \frac{0.14}{0.22} \approx 0.636 \text{ (probabilitatea posterioară)}.$$

▷ Probabilitatea ca un email să fie clasificat `spam`, ştiind că emailul conţine cuvântul *succes* este 0.636.

▷ Dacă un email conţine cuvântul *succes*, atunci există aproximativ 63.6% şanse să fie clasificat `spam`. ♣

P. 7. (Formula lui Bayes)
Într-un spaţiu de probabilitate $(\Omega, \mathcal{K}, P)$ *considerăm partiţia* $\{H_1, \ldots, H_n\}$ *a lui* Ω *cu* $H_i \in \mathcal{K}$ *şi* $P(H_i) > 0 \; \forall \, i \in \{1, ...n\}$*, şi fie* $E \in \mathcal{K}$ *astfel încât* $P(E) > 0$*. Atunci,*

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

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

Exemplu (problemă de clasificare): **Cât de bun este filtrul de spam?**

- H: un email este `spam` (în realitate)
- $\bar{H}$: un email este `non-spam` (în realitate)

Un filtru de spam trebuie să clasifice emailurile în
`spam` (evenimentul E) sau `non-spam` (evenimentul $\bar{E}$).

Se pune problema de a face o predicţie / prognoză asupra unui email ales aleator, dacă acesta este `spam` sau `non-spam` cu ajutorul filtrului de spam.

Au fost colectate următoarele date statistice:

$\triangleright$ $AP = 400$ (adevărat pozitiv) Numărul de emailuri care sunt de fapt `spam` şi care au fost clasificate ca `spam` de către filtrul de spam ; $\#(H \cap E)$[1]

$\triangleright$ $FP = 210$ (fals pozitiv) Numărul de emailuri care sunt de fapt `non-spam` şi care au fost clasificate ca fiind `spam` de filtrul de spam ; $\#(\bar{H} \cap E)$

$\triangleright$ $FN = 310$ (fals negativ) Numărul de emailuri care sunt de fapt `spam` şi care au fost clasificate ca `non-spam` de filtrul de spam ; $\#(H \cap \bar{E})$

$\triangleright$ $AN = 1200$ (adevărat negativ) Numărul de emailuri care sunt de fapt `non-spam` şi care au fost clasificate ca `non-spam` de filtrul de spam; $\#(\bar{H} \cap \bar{E})$.

Matricea de confuzie este utilizată pentru a vizualiza performanţa unui clasificator (de exemplu, a filtrului de spam).

		starea actuală (realitatea)		
		H (email este `spam`)	$\bar{H}$ (email este `non-spam`)	total
predicţia	E (email este clasificat `spam`)	AP	FP	AP+FP
	$\bar{E}$ (email este clasificat `non-spam`)	FN	AN	FN+AN
	total	AP+FN	FP+AN	AP+FP+FN+AN

Matricea de confuzie (engl. *confusion matrix*)

		starea actuală (realitatea)		
		H: email este `spam`	$\bar{H}$: email este `non-spam`	total
predicţia	E: email este clasificat `spam`	400 (adevărat pozitiv AP)	210 (fals pozitiv FP)	610
	$\bar{E}$: email este clasificat `non-spam`	310 (fals negativ FN)	1200 (adevărat negativ AN)	1510
	total	710	1410	2120

Matricea de confuzie construită cu datele statistice din acest exemplu

[1] $\#(H \cap E)$ = numărul de alemente din $H \cap E$.

Pe baza datelor statistice: a) probabilitatea ca un email, despre care se ştie că fost clasificat `spam`, să fie în realitate `spam`, este

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

b) probabilitatea ca un email, despre care se ştie că fost clasificat `non-spam`, să fie în realitate `non-spam` este

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

♣

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

★ Probabilităţile condiţionate sunt folosite în probleme de clasificare, în teoria deciziilor, în predicţie, în diagnosticare, etc.

Variable aleatoare

$\to$ Variabilele aleatoare apar ca funcţii, ce depind de rezultatul (aleator) al efectuării unui anumit experiment.

Exemplu: 1) La aruncarea a două zaruri, suma numerelor obţinute este o variabilă aleatoare
$S : \Omega \to \{2, 3, ..., 12\}$, unde Ω conţine toate evenimentele elementare ce se pot obţine 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 obţinut numărul i şi la al doilea zar s-a obţinut 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: $\Rightarrow X : \Omega \to \{0, 1, 2\}$

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

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

O variabilă aleatoare este:

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

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

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

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

Variabile aleatoare numerice - definiţie formală

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

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

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

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

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

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

▷ O variabilă aleatoare discretă X este caracterizată de distribuţia de probabilitate!
▷ Notăm $\{X = x_i\} = \{\omega \in \Omega : X(\omega) = x_i\}, i \in I$; acesta este un eveniment din $\mathcal{K}$ pentru fiecare $i \in I$.
$\mathbb{X} = (X_1, \ldots, X_m)$ este un **vector aleator discret** dacă fiecare componentă a sa este o variabiă aleatoare discretă.

Distribuţii discrete clasice

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

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

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

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

➤ Python: `scipy.stats.randint`

```
# Examplu Unid(6)  - Histograma
from scipy.stats import randint
import numpy
import matplotlib.pyplot as plt
from matplotlib.pyplot import bar, show, xticks, grid
N=4000
a=1; b=7
R = randint.rvs(a, b, size = N)
#print ("Valori aleatoare: \n", R)
x, count = numpy.unique(R, return_counts=True)
print("Valorile:",x,"au frecventele absolute:",count)
print("Valorile:",x,"au frecventele relative:",count/N)
bar(x,count/N, width=0.8,color="cyan", edgecolor="black")
# deseneaza histograma frecventelor relative
plt.grid()
plt.xlabel("valorile")
plt.ylabel("frecvente relative")
plt.title("Unid(6)")
xticks(range(0,b))
show()
```

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

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

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

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

♣

➤ Python: `scipy.stats.bernoulli`

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

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

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

Exemple: 1) Un zar se aruncă de 10 ori, fie X v.a. care indică de câte ori a apărut numărul 6 $\Rightarrow X \sim Bino(10, \frac{1}{6})$.
2) O echipă de marketing trimite un email promoţional către 100 de abonaţi. Pe baza unor date statistice, se ştie că fiecare abonat citeaşte emailul cu probabilitatea 0.25 (independent de ceilalţi abonaţi). Definim:
▷ n numărul de încercări; $n = 100$ de emailuri trimise
▷ p probabilitatea de succes; $p = 0.25$ (probabilitatea ca un email să fie citit)
▷ X variabila aleatoare; X = numărul de emailuri deschise (din cele n trimise)
? Care este probabilitatea ca exact 30 de persoane să citească emailul promoţional?
? Care este probabilitatea ca mai puţin de 20 de persoane să îl citească? ♣

$\rightarrow$ formula binomială $(a+b)^n = \sum_{k=0}^{n} C_n^k a^k b^{n-k}$ pentru $a = p$ şi $b = 1-p$ se obţine

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

Exemplu: Un client accesează o dată pe zi o anumită pagină web cu probabilitatea 0.6. Cu ce probabilitate clientul accesează această pagină în total de 3 ori în următoarele 10 zile? R.: $C_{10}^3 0.6^3 0.4^7$

➤ Python: `scipy.stats.binom`

```
# Exemplu distributia binomiala Bino(10,0.6)
import numpy
import matplotlib.pyplot as plt
from matplotlib.pyplot import bar, grid, show, xticks
from scipy.stats import binom
N=1000
n=10; p=0.6 # se genereaza date pentru distributia Bino (n,p)
data = binom.rvs(n, p,size= N)
z, count = numpy.unique(data, return_counts=True)
print("Valorile",z,"au frecventele relative:",count/N)
bar(z,count/N, width=0.8,color="yellow", edgecolor="black")
plt.grid()
plt.xlabel("valorile distributiei")
plt.ylabel("frecvente relative")
plt.title("Bino(10,0.6)")
s=sum(data==3)
print("Clientul acceseaza pag. de 3 ori in urmatoarele 10 zile cu prob.", s/N)
#estimarea probabilitatii teoretice din simulari
xticks(range(0,n+1))
show()
```

$\triangleright$ Distribuţia binomială corespunde **modelului cu extragerea bilelor dintr-o urnă cu bile de două culori şi cu returnarea bilei după fiecare extragere**:

Într-o urnă sunt n_1 bile albe şi n_2 bile negre. Se extrag cu returnare n bile; fie v.a. X_1 = numărul de bile albe extrase; X_2 = numărul de bile negre extrase

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

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

$P(X_1 = k_1, \dots, X_r = k_r) =$ probabilitatea de a obţine k_i bile având culoarea c_i, $i = \overline{1,r}$, din $n = k_1 + \dots + k_r$ extrageri *cu returnarea* bilei extrase

$$= \frac{n!}{k_1! \dots k_r!} \cdot p_1^{k_1} \cdot \dots \cdot p_r^{k_r}$$

▷ $(X_1, ..., X_r)$ este un vector aleator discret şi urmează distribuţia multinomială
▷ distribuţia multinomială modelează experimentele în care se extrag cu returnare un număr specific $n = k_1 + ... + k_r$ de obiecte (elemente) din r categorii, dintr-o mulţime dată de obiecte, care au probabilităţile $p_1 = \frac{n_1}{n_1+...+n_r}, ..., p_r = \frac{n_r}{n_1+...+n_r}$, unde n_i este numărul de obiecte din categoria a i-a ($i = \overline{1, r}$).
▷ cazul $r = 2$ corespunde distribuţiei binomiale (modelul binomial cu bile de două culori într-o urnă)
➤ Python: `scipy.stats.multinomial`

Exerciţii: 1) O reţea de laborator este compusă din 15 calculatoare. Reţeaua a fost atacată de un virus nou, care atacă un calculator cu o probabilitatea 0.4, independent de alte calculatoare. Care este probabilitatea ca virusul a atacat
a) cel mult 10; b) cel puţin 10; c) exact 10
calculatoare?
2) Sondaj de opinie: O companie PR studiază modul în care oamenii interacţionează cu mass-media în prezent. Pentru a înţelege comportamentul publicului, se realizează un sondaj. Fiecare participant la sondaj este întrebat: *Pentru a vă informa, ce tip de mass-media utilizaţi cel mai frecvent?*
Participanţii trebuie să aleagă una dintre următoarele opţiuni:
`Televiziune/Reţele sociale/Ziare tipărite/Ştiri online/Radio`.
Pe baza cercetărilor anterioare şi a tendinţelor de interacţiune cu mass-media, următoarele date statistice sunt disponibile:
`Televiziune`: 30%
`Reţele sociale`: 35%
`Ziare tipărite`: 5%
`Ştiri online`: 20%
`Radio`: 15%
Compania doreşte să calculeze probabilitatea de a observa următoarea distribuţie într-un eşantion de 100 de persoane:
30 persoane aleg `Televiziunea`
40 aleg `Reţelele sociale`
0 aleg `Ziarele tipărite`
30 aleg `Ştirile online`
0 aleg `Radio`.
Cu ce formulă se calculează această probabilitate?

♣

Distribuţia hipergeometrică: $X \sim Hyge(n, n_1, n_2),\ n, n_1, n_2 \in \mathbb{N}^*$
Într-o urnă sunt n_1 bile albe şi n_2 bile negre. Se extrag **fără returnare** n bile.

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

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

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

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

➤ Python: `scipy.stats.hypergeom`

▷ **Modelul urnei cu r culori şi bilă nereturnată:** fie n_i =numărul iniţial de bile având culoarea c_i din urnă, $i = \overline{1, r}$; fie X_i v.a. ce indică numărul de bile de culoarea c_i, $i = \overline{1, r}$, după n extrageri *fără returnarea bilei extrase*, iar ordinea de extragere a bilelor de diverse culori nu contează

$P(X_1 = k_1, \ldots, X_r = k_r) =$ probabilitatea de a obţine k_i bile având culoarea c_i, $i = \overline{1, r}$, din $n = k_1 + \ldots + k_r$ extrageri *fără returnarea* bilei extrase,

$$= \frac{C_{n_1}^{k_1} \cdot \ldots \cdot C_{n_r}^{k_r}}{C_{n_1+\ldots+n_r}^n}.$$

▷ $(X_1, ..., X_r)$ este un vector aleator discret şi urmează distribuţia hipergeometrică multidimensională

▷ distribuţia hipergeometrică multidimensională modelează experimentele în care se extrag fără returnare un număr specific $n = k_1 + \ldots + k_r$ de obiecte (elemente) din r categorii, dintr-o mulţime finită de obiecte $n_1 + \ldots + n_r$, unde n_i este numărul iniţial de obiecte din categoria a i-a ($i = \overline{1, r}$).

▷ Cazul $r = 2$ corespunde distribuţiei hipergeometrice.

➤ Python: `scipy.stats.multivariate_hypergeom`

Exemple: 1) Într-o urnă sunt $n_1 = 2$ bile albe şi $n_2 = 3$ bile negre. Se extrag fără returnare $n = 3$ bile. Fie v.a. X = numărul de bile albe extrase. Vom calcula $P(X = 1)$:

Prima metodă: Pentru $i \in \{1, 2, 3\}$ fie evenimentele

A_i: la a i-a extragere s-a obţinut bilă albă

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

Scriem

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

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

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

2) Loto 6 din 49: Care este probabilitatea de a nimeri exact 4 numere câştigătoare?
R.: Între cele 49 de bile exact $n_1 = 6$ sunt câştigătoare ("bilele albe") şi $n_2 = 43$ necâştigătoare ("bilele negre"). Probabilitatea ca din $n = 6$ extrageri fără returnare, exact $k = 4$ numere să fie câştigătoare (ordinea nu contează) este $p = \frac{C_6^4 C_{43}^2}{C_{49}^6}$.
▷ Fie v.a. X numărul de numere ghicite, jucând cu o singură variantă la "Loto 6 din 49". Scrieţi distribuţia de probabilitate a v.a. X.
3) O echipă de marketing se pregăteşte să testeze o nouă campanie promoţională. Baza lor de date cu clienţi conţine:

390 de clienţi atenţi la buget
310 de clienţi fideli mărcii
250 de cumpărători impulsivi
50 de clienţi VIP.

Din cauza unor constrângeri bugetare, se selectează aleator (şi fără returnare) 40 de clienţi pentru campania de testare. Care este probabilitatea ca eşantionul selectat să includă:

15 de clienţi atenţi la buget
10 clienţi fideli mărcii
10 cumpărători impulsivi
5 clienţi VIP? ♢

Distribuţia geometrică $X \sim Geo(p), p \in (0,1)$
În cadrul unui experiment poate să apară evenimentul A (*succes*) sau $\bar{A}$ (*insucces*)
• $A =$ succes cu $P(A) = p$, $\quad \bar{A} =$ insucces $P(\bar{A}) = 1 - p$
• se repetă (independent) experimentul până apare prima dată A ("succes")

• v.a. X arată de câte ori apare $\bar{A}$ (numărul de "insuccese") *până la* apariţia primului A ("succes") $\Rightarrow$ valori posibile: $X \in \{0, 1, \ldots\}$

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

➤ Python: `scipy.stats.geom`; atenţie valorile generate sunt de la 1; adică $P(Y = k) = p(1-p)^{k-1}$ pentru $k \in \{1, 2, \ldots\}$, iar $X = Y - 1$ cu $X \sim Geo(p)$.

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

♠

Exerciţiu: Considerăm v.a. X ca fiind numărul format astfel: dintr-o cutie cu 9 bile numerotate de la 1 la 9 sunt extrase aleator, succesiv, fară returnare, 2 bile, formând astfel un număr din două cifre, prima cifră fiind numărul primei bile, iar cea de-a doua cifră, fiind numărul celei de-a doua bile extrase.
a) Determinaţii distribuţia de probabilitate a v.a. X.
b) Calculaţi probabilitatea $P(X < 90)$.

Variabile aleatoare independente

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

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

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

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

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

P. 8. *Fie variabilele aleatoare discrete X (care ia valorile $\{x_i, i \in I\}$) şi Y (care ia valorile $\{y_j, j \in J\}$). Sunt echivalente afirmaţiile:*
(1) *X şi Y sunt v.a. sunt independente;*
(2) *$P(X \leq x, Y \leq y) = P(X \leq x) \cdot P(Y \leq y) \quad \forall\, x, y \in \mathbb{R}$.*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$\triangleright$ Dacă X şi Y sunt v.a. independente, şi se ştiu distribuţiile lor, atunci distribuţia vectorului aleator (X, Y) se determină pe baza formulei (1).

$\triangleright$ Dacă se cunoaşte distribuţia vectorului aleator (X, Y) distribuţiile lui X şi Y se determină astfel:

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

Operaţii cu variabile aleatoare (numerice)

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

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

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

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

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

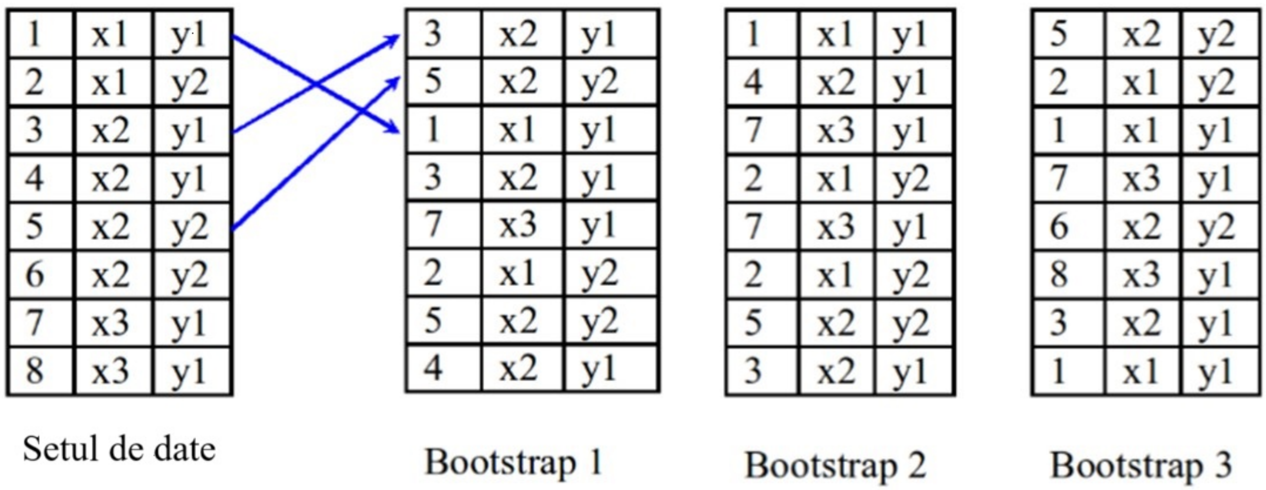
Exerciţii: **(1)** Fie X,Y v.a. independente, având distribuţiile

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

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

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

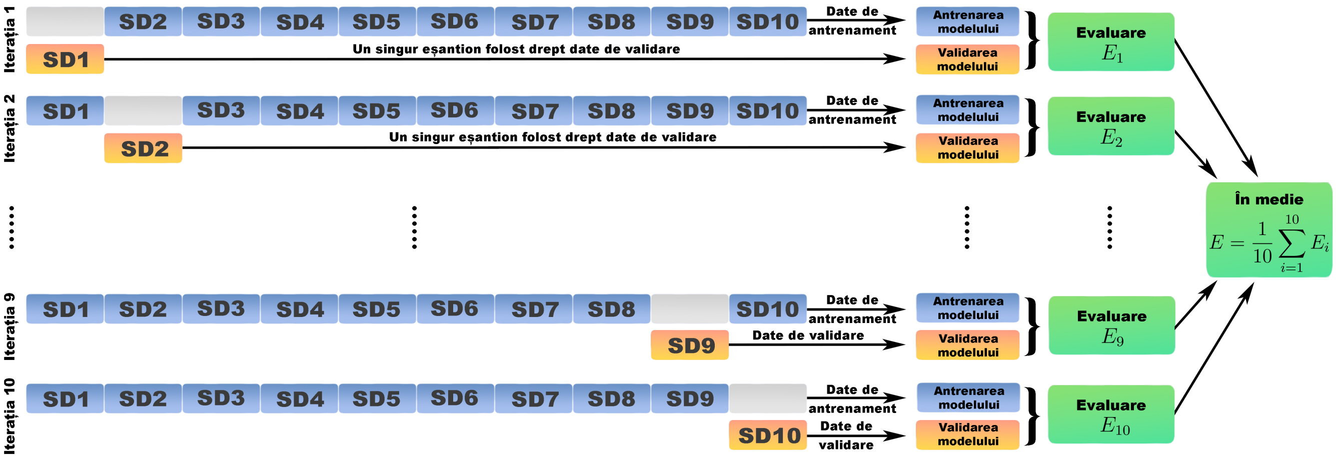
(3) Într-o echipă de dezvoltare software s-au implementat un set (o colecţie) de 100 de teste independente, care verifică funcţionalitatea unei noi aplicaţii. Aceste teste sunt rulate de fiecare dată când se face o modificare în cod. Pe baza datelor statistice existente, s-a constatat că fiecare test are o probabilitate de 95% să treacă (adică să nu detecteze o eroare). Pentru a estima fiabilitatea sistemului, se doreşte modelarea probabilităţii ca un anumit număr de teste să treacă într-o execuţie completă a setului de teste.
(a) Care este probabilitatea ca exact 95 de teste să treacă?
(b) Probabilitatea ca cel puţin 90 de teste să treacă?
(c) Dacă X este numărul de teste care trec, să se calculeze $P(X > 95)$?
(d) Ce descrie variabila aleatoare $100 - X$? ♣



Metoda bootstrap

► Extragerea cu returnare este folosită în **metoda bootstrap** (engl. *bootstrapping*), care este o metodă utilizată pentru a estima proprietăţile statistice dintr-un set de date. Tehnica implică re-eşantionarea (engl. *resampling*), folosind datele dintr-un singur set de date cu n observaţii. Un set de *date bootstrap* este format din n observaţii *alese aleator cu returnare şi independent* din setul de date iniţial.

Bootstrapping este o procedură statistică care re-eaņtionează un singur set de date pentru a crea mai multe eşantioane (folosind simulări). Aceste eşantione sunt folosite pentru a face inferenţe statistice asupra setului iniţial de date.



Validarea încrucişată, $k = 10$ (SD=set de date)

▶ Metoda validării încrucişate (engl. *cross validation*)

Validarea încrucişată este o tehnică de evaluare a unui model de învăţare automată şi de testare a performanţei acestuia. Metoda este folosită pentru compararea şi selectarea unui model adecvat în cazul unei probleme specifice de modelare predictivă.

În cazul validării încrucişate (k-*fold cross validation*), eşantionul original de date este împărţit *aleatoriu* în k sub-eşantioane de dimensiuni egale. Din cele k sub-eşantioane, un singur sub-eşantion este folosit ca *date de validare* pentru testarea modelului, iar celelalte $k - 1$ sub-eşantioane sunt utilizate ca *date de antrenament*. Procesul de validare încrucişată se repetă apoi de k ori, fiecare dintre cele k sub-eşantioane fiind utilizat exact o dată ca date de validare. Avantajul acestei metode constă în faptul că toate observaţiile sunt utilizate atât pentru antrenare, cât şi pentru validare, iar fiecare observaţie este utilizată pentru validare exact o dată. Validarea încrucişată cu k=10 (sau k=5) este utilizată în mod obişnuit.
Atunci când $k = n$ (numărul de observaţii), validarea încrucişată este echivalentă cu validarea încrucişată numită în engleză *leave-one-out*.

Clasificarea naivă Bayes

În învăţarea automată, clasificatorii bayesieni naivi sunt o familie de clasificatori probabilistici simpli, bazaţi pe aplicarea formulei lui Bayes (a se vedea P.5) cu ipoteze "naive" de independenţă condiţionată între atribute (engl. *features*), cunoscând clasificarea. Pentru unele tipuri de modele de probabilitate, clasificatorii bayesieni naivi pot fi antrenaţi foarte eficient. În aplicaţii practice pentru modelele bayesiene naive se foloseşte *metoda probabilităţii maxime*. Noţiunea folosită în acest context este condiţional independenţa între v.a.

Fie $(\Omega, \mathcal{K}, P)$ un spaţiu de probabilitate. De asemenea considerăm că toate probabilităţile condiţionate sunt definite (adică condiţionarea se face în raport cu un eveniment a cărui probabilitate nu este 0).

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

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

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 şi A_2 condiţional independente, cunoscând Z_1? Sunt A_1 şi A_2 independente?
R.: Dacă se cunoaşte 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 probabilităţii totale P.5 avem:
$P(A_1) = P(A_1|Z_1)P(Z_1) + P(A_1|Z_2)P(Z_2) = \frac{1}{6} \cdot \frac{1}{2} + \frac{5}{6} \cdot \frac{1}{2} = \frac{1}{2}$,
$P(A_2) = P(A_2|Z_1)P(Z_1) + P(A_2|Z_2)P(Z_2) = \frac{1}{6} \cdot \frac{1}{2} + \frac{5}{6} \cdot \frac{1}{2} = \frac{1}{2}$,
$P(A_1 \cap A_2) = P(A_1 \cap A_2|Z_1)P(Z_1) + P(A_1 \cap A_2|Z_2)P(Z_2) = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{2} + \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{2} = \frac{13}{36}$.
$\Longrightarrow P(A_1 \cap A_2) \neq P(A_1)P(A_2) \Longrightarrow$ A_1 şi A_2 nu sunt independente. ✾

Def. 15. *Fie X, Y, Z v.a. discrete, care iau valori în mulţimile $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. V.a. X este* ***condiţional independentă*** *de Y, cunoscând (ştiind) 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

	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	dimineaţă	**relaxat**
8	ploaie	noapte	**relaxat**
9	înnorat	dimineaţă	**aglomerat**
10	zăpadă	noapte	**aglomerat**
11	senin	seară	**relaxat**
12	zăpadă	amiază	**relaxat**
13	înnorat	seară	**aglomerat**
14	ploaie	dimineaţă	**aglomerat**
15	zăpadă	dimineaţă	**aglomerat**
16	**ploaie**	**amiază**	**?**

Tabel de date obţinute în urma unor observaţii

Se doreşte *clasificarea* **traficului** $\boldsymbol{T}$ pe un anumit bulevard, în *clasele*: *aglomerat* $\boldsymbol{a}$ sau *relaxat* $\boldsymbol{r}$, în funcţie de următoarele *atribute* cu valorile lor posibile:

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

Considerăm evenimentul următor, denumit *vector de atribute*:

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

Se caută o clasă pentru E, stabilind care din următoarele probabilităţi este mai mare: $P(\boldsymbol{T = a}|E)$ sau $P(\boldsymbol{T = r}|E)$; aceasta este **metoda de probabilitate maximă.** *Ştiind că vremea este ploioasă şi este amiază, ce previziune se poate face despre trafic (aglomerat* $\boldsymbol{a}$ *sau relaxat* $\boldsymbol{r}$*)*?

Se face următoarea presupunere *naivă*: atributele sunt **condiţional independente**, dacă se ştie (cunoaşte) clasificarea, adică

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

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

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

► Folosind datele din tabel, determinăm mai întâi probabilităţile claselor şi probabilităţile condiţionate 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	$\mathbf{T=a}$	$\mathbf{T=r}$	$P(V=...\mid\mathbf{T=a})$	$P(V=...\mid\mathbf{T=r})$
p	2	1	$\frac{2}{9}$	$\frac{1}{6}$
z	3	1	$\frac{3}{9}$	$\frac{1}{6}$
s	1	3	$\frac{1}{9}$	$\frac{3}{6}$
$\hat{\imath}$	3	1	$\frac{3}{9}$	$\frac{1}{6}$

Ti	$\mathbf{T=a}$	$\mathbf{T=r}$	$P(Ti=...\mid\mathbf{T=a})$	$P(Ti=...\mid\mathbf{T=r})$
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 şi a ipotezei de independenţă condiţionată, deducem că:

$$P(\mathbf{T=a}|E)=\frac{P(E|\mathbf{T=a})P(\mathbf{T=a})}{P(E)}=\frac{P(V=p,Ti=am|\mathbf{T=a})P(\mathbf{T=a})}{P(E)}$$

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

şi

$$P(\mathbf{T=r}|E)=\frac{P(E|\mathbf{T=r})P(\mathbf{T=r})}{P(E)}=\frac{P(V=p,Ti=am|\mathbf{T=r})P(\mathbf{T=r})}{P(E)}$$

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

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

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

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

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

şi 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 $\sum\limits_{i \in I} |x_i| P(X = x_i) < \infty$.

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

P. 9. *Fie X şi Y v.a. discrete. Au loc proprietăţile:*
→ $E(aX + b) = aE(X) + b$ *pentru orice* $a, b \in \mathbb{R}$;
→ $E(X + Y) = E(X) + E(Y)$;
→ *Dacă X şi Y sunt v.a. independente, atunci* $E(X \cdot Y) = E(X)E(Y)$.
→ *Dacă $g : \mathbb{R} \to \mathbb{R}$ e o funcţie astfel încât $g(X)$ este v.a., atunci*

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

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

➤ Python: $\texttt{numpy.mean}(x) = \frac{1}{n}(x_0 + ... + x_{n-1})$ pentru $x = [x_0, ..., x_{n-1}]$
➤ Fie $x = [x_0, \ldots, x_{n-1}]$ valori aleatoare ale unei v.a. X, atunci

$$E(X) \approx \texttt{numpy.mean}(x) = \frac{1}{n}(x_0 + \cdots + x_{n-1}) \text{ pentru } n \text{ suficient de mare}$$

```
# Exemplu numpy.mean
import numpy
x = [[1, 3], [5, 9]]
print("media aritmetica (matrice):",numpy.mean(x))
y=[-1,0,-2,0,1,2,2,1,0,1]
print("media aritmetica (vector):",numpy.mean(y))
```

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

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

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

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

Aşadar jucătorul *câştigă în medie* 5 u.m.

```
import numpy as np
s=[]
N=10000
for _ in range(N):
    jocuri = np.random.choice([-1,-1,-1,-1,2,3],size=30,replace=True)
    s.append(sum(jocuri))
print("Castigul mediu (dupa 30 jocuri):",numpy.mean(s))
```

Exerciţiu: Variabila aleatoare X descrie de câte ori apare pana de curent în reţea (pe parcursul unei zile, într-o anumită localitate)

$$P(X = 0) = 0.9, P(X = 1) = 0.08, P(X = 2) = 0.02\,.$$

O companie de comerţ pe internet estimează că fiecare astfel de pană de curent în reţea duce la o pierdere de 200 Ron. Calculaţi valoarea medie a pierderilor zilnice ale acestei companii (datorate lipsei de curent). Estimaţi această valoare medie cu ajutorul unor simulări în Python.

Def. 17. *Fie $X_1, \ldots, X_n$ cu $n \in \mathbb{N}$, $n \geq 2$, variabile aleatoare discrete, care iau valori în mulţimile $\mathcal{X}_1, \ldots, \mathcal{X}_n$. $X_1, \ldots, X_n$ sunt* ***variabile aleatoare independente****, dacă şi numai dacă*

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

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

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

Def. 18. ***Funcţia de repartiţie*** *$F : \mathbb{R} \to [0, 1]$ a unei variabile aleatoare discrete X, care ia valorile $\{x_i, i \in I\}$, este*

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

În lb. engleză denumirea este `cummulative distribution function`, prescurtat cu `cdf`.

Exemplu: Funcţia de repartiţie $F_X : \mathbb{R} \to [0,1]$ a v.a. discrete X este

$$F_X(x) = P(X \leq x) = \begin{cases} 0, & \text{dacă } x < -2 \\ 0.5, & \text{dacă } -2 \leq x < 1 \\ 0.7, & \text{dacă } 1 \leq x < 2 \\ 1, & \text{dacă } 2 \leq x\,. \end{cases}$$

Determinaţi valoarea medie a lui X. ♣

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

Observaţie:
▷ Orice funcţie $F : \mathbb{R} \to \mathbb{R}$, care are proprietăţile (2), (3), (4) din **P.10** este o funcţie de repartiţie.
▷ Funcţia de repartiţie a unei v.a. descrie complet comportamentul probabilistic al acelei v.a.
➤ în Python: se calculează $F(x) = P(X \leq x)$
pentru $X \sim Bino(n, p)$ cu `scipy.stats.binom.cdf(`x, n, p`)`,
iar pentru $X \sim Hyge(n, n_1, n_2)$ cu `scipy.stats.hypergeom.cdf(`$x, n_1{+}n_2, n_1, n$`)`.

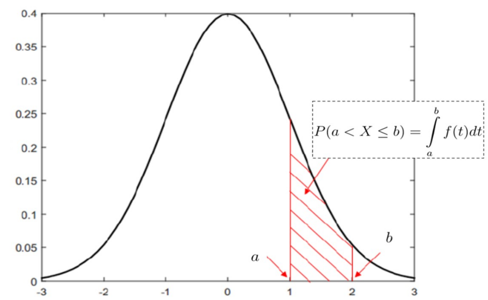
```
#Fie o urna cu 10 bile, din care 5 sunt rosii; X (v.a.)= cate bile rosii au fost
                                        extrase
#in 5 extrageri cu returnare; se reprezinta grafic functia de repartitie a lui X
import scipy.stats
import matplotlib.pyplot as plt
import numpy as np
n=5
p=0.5
x = np.linspace(-2, n+2, 101)
y=scipy.stats.binom.cdf(x,n,p)
plt.plot(x, y, "r.")
for t in range(n+1):
    plt.plot(t, scipy.stats.binom.cdf(t,n,p), "ko")
    plt.plot(t, scipy.stats.binom.cdf(t-(n+4)/100,n,p), 'ko', mfc='none')
plt.xlabel("x")
plt.ylabel("F(x)=P(X <= x)")
plt.title("Functia de repartitie a lui X")
plt.xticks(range(-2,n+3))
plt.grid()
plt.show()
```

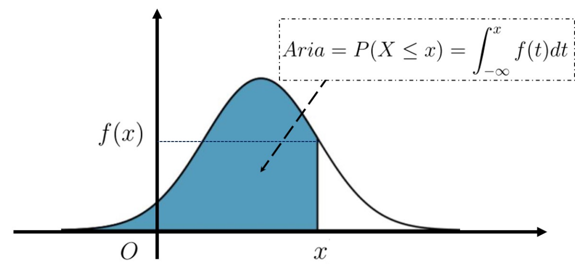
Variabile aleatoare continue

V.a. continuă: ia un număr infinit şi nenumărabil de va-lori într-un interval sau reuniune de intervale (v.a. poate lua orice valoare din intervalul conside-rat);

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

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





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

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

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

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

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

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

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

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

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

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

(5) $P(X \in M) = \int_M f(t)dt,\ M \subseteq \mathbb{R}$ *pentru care integrala există;*

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

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

(7) *F este o funcţie monoton crescătoare şi continuă pe $\mathbb{R}$;*
(8) $\lim\limits_{x\to\infty} F(x) = 1$ *şi* $\lim\limits_{x\to-\infty} F(x) = 0$.
(9) *dacă F este derivabilă în punctul x, atunci $F'(x) = f(x)$.*

Observaţii: (1) Orice funcţie $f : \mathbb{R} \to \mathbb{R}$, care are proprietăţile (1), (2) din **P.11** este o funcţie de densitate.
(2) Fie f_1 o funcţie de densitate pentru v.a. X şi fie $f_2 : \mathbb{R} \to \mathbb{R}$ astfel încât $f_1(t) = f_2(t)$ pentru orice $t \in \mathbb{R} \setminus \mathcal{N}$, unde $\mathcal{N}$ este o mulţime *cel mult numărabilă.* Atunci f_2 este o funcţie de densitate pentru aceeaşi v.a. X. Este suficient să cunoaştem o funcţie de densitate în orice punct din $\mathbb{R}$ exceptând, eventual, o mulţime cel mult numărabilă de puncte. Funcţia de densitate asociată unei v.a. nu este unică (unicitate în sensul egalităţii în toate punctele din $\mathbb{R}$). Proprietăţile integralelor implică

$$\int_{-\infty}^{x} f_1(t)dt = \int_{-\infty}^{x} f_2(t)dt, \ \forall\, x \in \mathbb{R},$$

$$\int\limits_a^b f_1(t)dt = \int\limits_a^b f_2(t)dt, \forall\, a < b, \ a, b \in \mathbb{R}.$$

(3) *Funcţia de fiabilitate* (sau funcţia de supravieţuire) este folosită în studiul fiabilităţii unor a sisteme, şi a duratei de funcţionare până la prima defecţiune

$$R(x) = P(X > x), \ x \in \mathbb{R}$$

Exemple de distribuţii clasice continue

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

• funcţia de densitate este

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

Python:
▷ pentru $a = 0, b = 1$: `random.random()` returnează o valoare aleatoare din $[0, 1]$
▷ `scipy.stats.uniform.rvs`$(a, b - a,$ `size` $= N)$ returnează N valori aleatoare uniform distribuite din $[a, b]$

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

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

• funcţia de densitate este

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}, t \in \mathbb{R}.$$

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

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

```
# functia de densitate pentru distributia normala
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import norm
x = np.linspace(-8, 8, 101)
sigma= [1, 2, 4]   # valori pentru sigma
for t in sigma:
    y = norm.pdf(x, loc=0, scale=t)  # scale= deviatia standard
    plt.plot(x, y, label=f"Norm(0, {t}$^2$)")
plt.xlabel("x")
plt.ylabel("functia de densitate")
plt.title("Distributia normala")
plt.legend()
plt.show()
```

➦**Distribuţia exponenţială:** $X \sim Exp(\lambda)$, $\lambda > 0$

• funcţia de densitate este

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

```
# functia de densitate pentru distributia exponentiala
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import expon
x = np.linspace(0, 4, 100)
L= [1, 2, 3]   # valori pentru Lambda
for t in L:
```

```
    y = expon.pdf(x, scale=1/t)
    plt.plot(x, y, label=f"Exp({t})")
plt.xlabel("x")
plt.ylabel("functia de densitate")
plt.title("Distributia exponentiala")
plt.legend()
plt.show()
```

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

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

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

unde funcţia Gamma este

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

```
#Exemplu functii de densitate
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import t,norm
x = np.linspace(-5, 5, 100)
degrees_of_freedom = [1, 2, 5]
# diferite grade de libertate pt distributia Student T(.)
# functii de densitate
for df in degrees_of_freedom:
    y = t.pdf(x, df)
    plt.plot(x, y, label=f"T({df})")
z0=norm.pdf(x,0,1)
plt.plot(x, z0, label="N(0,1)")
plt.xlabel('x')
plt.ylabel("functia de densitate")
plt.title("distributia Student si distributia normala standard")
plt.legend()
plt.show()
```

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

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

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

Exemplu: Fie $X \sim Exp(0.5)$ v.a. care indică timpul de funcţionare a unei baterii (câte luni funcţionează bateria). Să se calculeze a) $P(2 \leq X \leq 4)$; b) $P(X > 3)$.

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

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

♣

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

Vector aleator

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

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

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

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

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

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

F_X şi F_Y se numesc **funcţii de repartiţie marginale**.

Def. 21. X *şi* Y *sunt* ***variabile aleatoare independente*** *(discrete sau continue), dacă şi numai dacă*

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

ceea ce este echivalent cu

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

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

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

Sunt X şi Y v.a. independente? Determinaţi f_X, respectiv f_Y. ♣

Valoarea medie a unei variabile aleatoare continue

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

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

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

P. 12. *Proprietăţi 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 şi Y sunt variabile aleatoare* ***independente****, atunci* $E(X \cdot Y) = E(X)E(Y)$.
→ *Dacă $g : \mathbb{R} \to \mathbb{R}$ e o funcţie, astfel încât $g(X)$ este o v.a. continuă, atunci*

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

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

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

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

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

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

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

♣

Varianţa unei variabile aleatoare

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

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

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

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

P. 13. *Proprietăţi ale varianţei (pentru v.a. discrete sau continue):*
$\rightarrow$ $V(X) = E(X^2) - E^2(X)$.
$\rightarrow$ $V(aX + b) = a^2 V(X)\ \forall\, a, b \in \mathbb{R}$.
$\rightarrow$ *Dacă X şi Y sunt variabile aleatoare* ***independente****, atunci*

$$V(X + Y) = V(X) + V(Y).$$

► Python: `numpy.mean, numpy.var, numpy.std`
Fie $x = [x_0, \dots, x_{n-1}]$ valori aleatoare ale unei v.a. X

$$E(X) \approx \texttt{numpy.mean}(x) = \frac{1}{n}(x_0 + \cdots + x_{n-1}) \text{ pentru } n \text{ suficient de mare}$$

$$V(X) \approx \texttt{numpy.var}(x) = \frac{1}{n}\sum_{i=0}^{n-1}(x_i - \texttt{numpy.mean}(x))^2 \text{ pentru } n \text{ suficient de mare}$$

$$Std(X) \approx \texttt{numpy.std}(x) = \left(\frac{1}{n}\sum_{i=0}^{n-1}(x_i - \texttt{numpy.mean}(x))^2\right)^{\frac{1}{2}} \text{pentru } n \text{ suficient de mare}.$$

Observaţie: Dacă Z este o v.a. (cu $V(Z) > 0$) atunci v.a. $\tilde{Z} := \frac{Z - E(Z)}{Std(Z)}$ este **versiunea standardizată** a v.a. Z, iar $E(\tilde{Z}) = 0$ şi $V(\tilde{Z}) = 1$.

Proprietăţi (recapitulare)

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

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

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

V.a. $X_1, \dots, X_n$ sunt independente şi folosind P.13, obţinem

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

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

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

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

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

Din P.11-(2) rezultă

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

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

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

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

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

3) Să se arate că: $X \sim N(\mu, \sigma^2) \Longleftrightarrow \frac{X-\mu}{\sigma} \sim N(0,1)$.
R.: Funcţia de densitate pentru distribuţia $N(\mu, \sigma^2)$ este

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

Funcţia de densitate pentru $N(0,1)$ este

$$f_{N(0,1)}(y) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\}, y \in \mathbb{R}.$$

Notăm cu $Y = \frac{X-\mu}{\sigma}$.
"$\Rightarrow$" Pentru orice $y \in \mathbb{R}$ are loc:

$$F_Y(y) = P(Y \le y) = P\left(\frac{X-\mu}{\sigma} \le y\right) = P(X \le \sigma y + \mu) = F_X(\sigma y + \mu).$$

Prin derivare în raport cu y se obţine

$$f_Y(y) = F'_Y(y) = F'_X(\sigma y + \mu) \cdot \sigma = f_X(\sigma y + \mu) \cdot \sigma = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\},$$

care este funcţia de densitate pentru $N(0,1)$. Deci $Y \sim N(0,1)$.
"$\Leftarrow$" Pentru orice $x \in \mathbb{R}$ are loc:

$$F_X(x) = P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = P\left(Y \le \frac{x-\mu}{\sigma}\right) = F_Y\left(\frac{x-\mu}{\sigma}\right).$$

Prin derivare în raport cu x se obţine

$$f_X(x) = F'_X(x) = F'_Y\left(\frac{x-\mu}{\sigma}\right) \cdot \frac{1}{\sigma} = f_Y\left(\frac{x-\mu}{\sigma}\right) \cdot \frac{1}{\sigma} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

care este funcţia de densitate pentru $N(\mu,\sigma^2)$. Deci $X \sim N(\mu,\sigma^2)$. ♣

Def. 24. $f_{(X,Y)} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ *este* ***funcţia de densitate a vectorului aleator continuu*** (X,Y)*, dacă*

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

unde $F_{(X,Y)}$ *este funcţia de repartiţie a vectorului aleator* (X,Y)*.*

Proprietăţile funcţiei de densitate din P.11 sunt generalizate pentru funcţia de densitate a unui vector aleator continuu.

P. 14. *Pentru un* ***vector aleator continuu*** (X,Y) *au loc proprietăţile:*

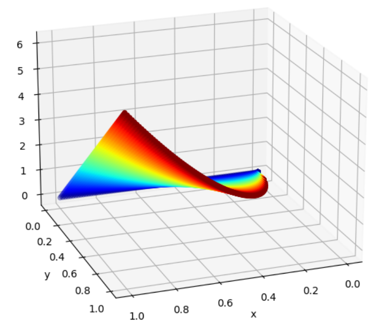
1. $f_{(X,Y)}(s,t) \ge 0 \; \forall \; s,t \in \mathbb{R}$.
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{(X,Y)}(u,v)dudv = 1.$

3. $F_{(X,Y)}$ *este funcţie continuă pe* $\mathbb{R}^2$.

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

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

5. $P((X,Y) \in M) = \underbrace{\int\int}_{M} f_{(X,Y)}(u,v)dudv,\ M \subset \mathbb{R}^2$*, pentru care integrala există.*

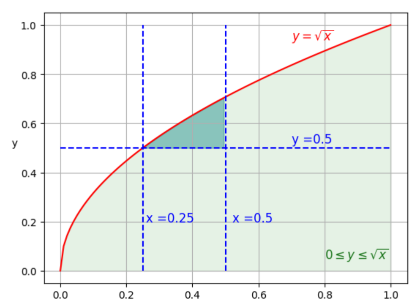


Exemplul 1

Exemplul 1: Fie $f_{(X,Y)} : \mathbb{R}^2 \to [0,\infty)$ densitatea de probabilitate a vectorului aleator (X,Y) definită prin

$$f_{(X,Y)}(x,y) = \begin{cases} 6xy, & \text{dacă } 0 \le x \le 1, 0 \le y \le \sqrt{x} \\ 0, & \text{altfel.} \end{cases}$$

Să se determine $P(X \le 0.5 \le Y)$.



R.: Folosim P. 14 pentru a calcula

$$P(X \le 0.5 \le Y) = \int_{0.25}^{0.5} \left(\int_{0.5}^{\sqrt{x}} 6xy\, dy \right) dx = \tfrac{5}{128}. \clubsuit$$

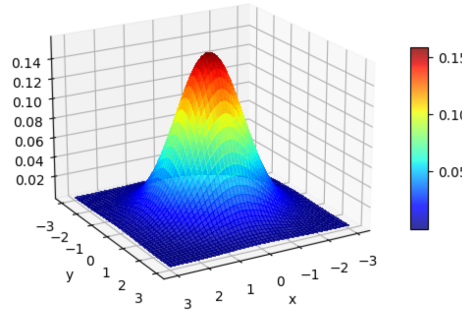
► Dacă se cunoaşte funcţia de densitate $f_{(X,Y)}$ pentru vectorul aleator continuu (X,Y), atunci f_X, respectiv f_Y, se determină cu

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

f_X şi f_Y se numesc **funcţii de densitate marginale**.

Exemplul 2: Distribuţia normală bidimensională standard: (X, Y) are funcţia de densitate

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



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

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

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

$\Longrightarrow X, Y \sim N(0,1)$. ♣

P. 15. *Variabilele aleatoare continue X (cu funcţia de densitate f_X) şi Y (cu funcţia de densitate f_Y) sunt* ***independente****, dacă şi numai dacă*

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

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

Exemplul 3: (X, Y) are *distribuţie uniformă* pe $I = [a_1, b_1] \times [a_2, b_2]$, cu $a_1, a_2, b_1, b_2 \in \mathbb{R}$, $a_1 < b_1, a_2 < b_2$ dacă

$$f_{(X,Y)}(x,y) = \begin{cases} \dfrac{1}{(b_1 - a_1)(b_2 - a_2)}, & \text{dacă } (x,y) \in I \\ 0, & \text{dacă } (x,y) \notin I. \end{cases}$$

X şi Y v.a. independente.
R.: Cu (4) se calculează

$$f_X(x) = \begin{cases} \dfrac{1}{b_1 - a_1} & \text{dacă } x \in [a_1, b_1] \\ 0 & \text{dacă } x \in \mathbb{R} \setminus [a_1, b_1]. \end{cases} \quad \text{şi } f_Y(y) = \begin{cases} \dfrac{1}{b_2 - a_2} & \text{dacă } y \in [a_2, b_2] \\ 0 & \text{dacă } y \in \mathbb{R} \setminus [a_2, b_2]. \end{cases}$$

$\Longrightarrow X \sim Unif[a_1, b_1], Y \sim Unif[a_2, b_2]$ (a se vedea distribuţia uniformă pe un interval, pg. 39). Se observă $f_{(X,Y)} = f_X \cdot f_Y \Longrightarrow X$ şi Y sunt v.a. independente! ✿

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

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

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

Def. 25. $(X_n)_n$ *este* ***şir de v.a. independente****, dacă* $\forall \{i_1,\ldots,i_k\} \subset \mathbb{N}$ *v.a.* $X_{i_1},\ldots,X_{i_k}$ *sunt independente, adică pentru oricare* $x_{i_1},\ldots,x_{i_k} \in \mathbb{R}$ *are loc*

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

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

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

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

Def. 26. *Şirul de v.a.* $(X_n)_n$ ***converge aproape sigur (a.s.)*** *la v.a.* X*, dacă*

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

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

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

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

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

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

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

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

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

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

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

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

Legea tare a numerelor mari (LTNM)

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

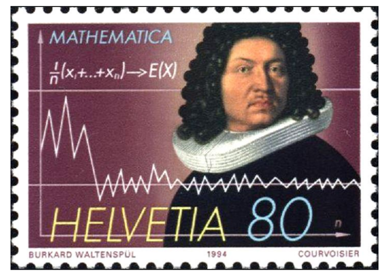


Fig. 5. Jacob Bernoulli (timbru emis în 1994 cu ocazia Congresului Internaţional al Matematicienilor din Elveţia)

Def. 27. *Şirul de v.a.* $(X_n)_n$ *cu* $E|X_n| < \infty\ \forall\, n \in \mathbb{N}$ *verifică* ***legea tare a numerelor mari*** *(**LTNM**) dacă*

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

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

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

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

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

```
#LTNM  X_n - nr aparut la aruncarea n
import matplotlib.pyplot as plt
from scipy.stats import randint
import numpy as np
N=400 #de cate ori aruncam zarul
X=randint.rvs(1,7,size=N)
sume = np.cumsum(X)
# sume cumulative: X[0], X[0]+X[1], ... , X[0]+X[1]+...+X[N-1]
S=[]
for n in range(N):
```

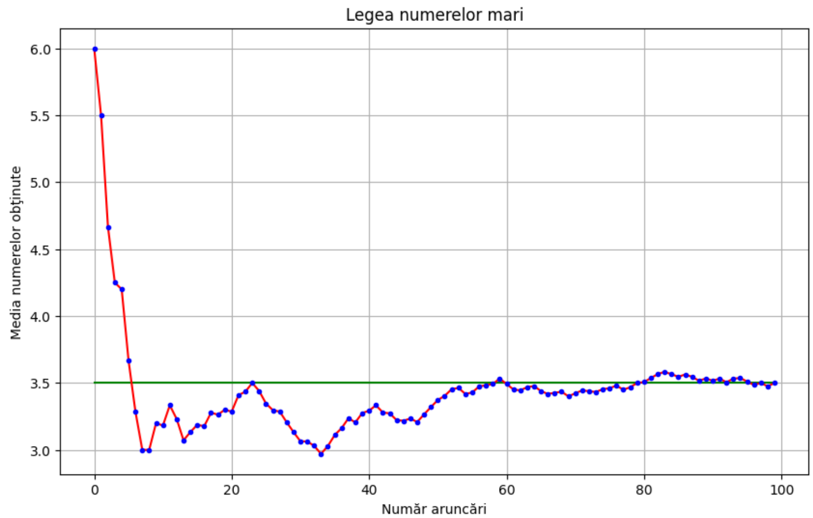


Fig. 4. Simulare LTNM

```
    S.append(sume[n]/(n+1))
    # se aduga in S valoarea (X[0]+X[1]+...+X[n]) /(n+1)
print("Valoare medie estimata prin simulari",np.mean(X))
# egala cu  sume[N-1]/N
v=[1,2,3,4,5,6]
e=np.mean(v)
print("Valoarea medie teoretica ",f"{e:3.2f}")
t=[i for i in range(0,N)]
plt.figure(figsize=(8,6))
plt.plot(t,e*np.ones(N),"g-")
plt.plot(t,S,"r-")
plt.plot(t,S,"b.")
plt.xlabel("Numar de aruncari ale zarului")
plt.ylabel("Valoarea medie a numerelor obtinute ")
plt.title("Legea numerelor mari")
plt.grid()
plt.show()
```

Exemplul 2: Fie $X_1, ..., X_n, ... \sim Unif[-1, 1]$ v.a. independente. Spre ce valoare converge a.s. şirul

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

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

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

▲

Exemplul 3: Fie $(X_n)_n$ şir de v.a. independente, având aceeaşi distribuţie ca v.a. X şi varianţă finită: $E(X_n) = E(X) \in \mathbb{R}$, $V(X_n) = V(X) \in \mathbb{R}$ pentru fiecare $n \in \mathbb{N}^*$. Definim $Y_n = (X_n - E(X))^2\ \forall\ n \in \mathbb{N}^* \Rightarrow (Y_n)_n$ este şir de v.a. independente, având

aceeaşi distribuţie ca v.a. $(X - E(X))^2$ şi $E(Y_n) = E((X - E(X))^2) = V(X)\ \forall\, n \in \mathbb{N}^*$.
P.16 $\Rightarrow (Y_n)_n$ verifică **LTNM**

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

adică

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Statistică matematică

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

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

• statistica inferenţială: metode de interpretare a rezultatelor obţinute prin metodele statisticii descriptive, utilizate apoi pentru luarea deciziilor.

► O *colectivitate* sau *populaţie statistică* $\mathcal{C}$ este o mulţime de elemente care au anumite însuşiri comune ce fac obiectul analizei statistice. Numărul elementelor populaţiei se numeşte *volumul populaţiei.*

Exemple de populaţii statistice: mulţimea persoanelor dintr-o anumită ţară, localitate, zonă etc. într-un anumit an; multimea gospodăriilor din Romania la un moment dat; mulţimea consumatorilor unui anumit produs; mulţimea societăţilor care produc un anumit produs; angajaţii unei societăţi; studenţii unei facultăţi.

► *Eşantionul* $\mathcal{E}$ reprezintă o submulţime a unei populaţii statistice $\mathcal{E} \subset \mathcal{C}$, constituită după criterii bine stabilite:

a) să fie aleatoare;

b) toate elementele colectivităţii să aibă aceeaşi şansă de a fi alese în eşantion;

c) eşantionul să fie reprezentativ (structura eşantionului să fie apropiată de structura populaţiei);

d) volumul eşantionului să fie suficient de mare.

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

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

► *Variabila statistică* sau *caracteristica* reprezintă o însuşire, o proprietate măsurabilă a unei unităţi statistice, întâlnită la toate unităţile care aparţin aceleiaşi colectivităţi şi care prezintă variabilitate de la o unitate statistică la alta. *Caracteristica sau variabila statistică corespunde unei variabile aleatoare.*

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

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

b) variabile (caracteristici) discrete $\to$ iau număr finit sau infinit dar numărabil de valori

discrete (de ex.: numări elevi ai unei şcoli, numărul liceelor existente într-un oraş, valoarea IQ)
$\triangleright$ caracteristicile de la a) şi b) sunt variabile numerice (cantitative)
c) variabile (caracteristici) nominale (de ex.: culoarea ochilor, ramura de activitate, religia)
d) variabile (caracteristici) nominale ordinale (de ex.: starea de sănătate / calitatea unor servicii - precară, mai bună, bună, foarte bună)
e) variabile (caracteristici) dihotomiale (binare) (de ex.: stagiul militar - satisfăcut/nesatisfăcut, starea civilă - căsătorit/necăsătorit)
$\triangleright$ caracteristicile de la c),d),e) sunt variabile calitative
$\triangleright$ variabilele nominale mai sunt numite variabile categoriale
$\blacktriangleright$ *Datele statistice* reprezintă observaţiile rezultate dintr-o cercetare statistică, sau ansamblul valorilor colectate în urma unei cercetări statistice.
De exemplu: un angajat al unei companii are o vechime de 6 ani în muncă. Angajatul reprezintă unitatea statistică, vechimea în muncă este caracteristica (variabila) cercetată, iar 6 este valoarea acestei caracteristici.
O *colectivitate* (populaţie) $\mathcal{C}$ este cercetatată din punctul de vedere al caracteristicii (variabilei statistice) X.

Distribuţia caracteristicii X de poate fi

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

2) specificată, dar depinzând de unul sau mai mulţi parametri necunoscuţi
de ex.: $X \sim Exp(\lambda)$, $X \sim Bino(10, p)$, $X \sim N(m, \sigma^2)$
3) necunoscută: $X \sim ?$
• în cazul 2) parametrii sunt necunoscuţi, iar în cazul 3) distribuţia este necunoscută
$\hookrightarrow$ se estimează, folosind teoria estimaţiei şi intervalele de încredere
$\hookrightarrow$ se testează, folosind testele statistice
$\blacktriangleright$ Fie $\mathcal{E} \subset \mathcal{C}$ un eşantion. Se numesc date de selecţie relative la caracteristica X datele statistice $x_1, \ldots, x_n$ obţinute prin cercetarea indivizilor care fac parte din eşantionul $\mathcal{E}$.
$\blacktriangleright$ Datele de selecţie $x_1, \ldots, x_n$ pot fi considerate ca fiind valorile unor variabile aleatoare $X_1, \ldots, X_n$, numite variabile de selecţie şi care se consideră a fi variabile aleatoare independente şi având aceeaşi distribuţie ca X.
$\blacktriangleright$ Fie $x_1, \ldots, x_n$ datele statistice pentru caracteristica cercetată X, notăm cu $X_1, \ldots, X_n$ variabilele de selecţie corespunzătoare. Fie $g : \mathbb{R}^n \to \mathbb{R}$ o funcţie astfel încât $g(X_1, \ldots, X_n)$ este o variabilă aleatoare.
$g(X_1, \ldots, X_n)$ se numeşte funcţie de selecţie sau estimator
$g(x_1, \ldots, x_n)$ se numeşte valoarea funcţiei de selecţie sau valoarea estimatorului.

Estimarea punctuală este valoarea atribuită unui parametru necunoscut pe baza statisticii construite din eşantion.

Fie $X_1, \ldots, X_n$ variabilele de selecţie corespunzătoare caracteristicii cercetate X, a cărei distribuţie depinde de parametrul necunoscut θ.

Def. 28. *$g(X_1, \ldots, X_n)$ este estimator nedeplasat pentru parametrul necunoscut θ, dacă*

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

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

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

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

▷ Un estimator nedeplasat furnizează estimări corecte ale parametrului *în medie* pe baza mai multor eşantioane.
▷ Estimatorul nu este consistent, dacă acesta nu converge către valoarea reală a parametrului, chiar dacă există un număr mare de date statistice.

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

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

► media de selecţie (empirică)

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

► valoarea mediei de selecţie

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

▷ Media de selecţie $\bar{X}_n$ este un estimator nedeplasat şi consistent pentru media teoretică $E(X)$ a caracteristicii X; se folosesc simulări pentru $E(X) \approx \bar{x}_n$; `numpy.mean`
► varianţa (dispersia) de selecţie (empirică)

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

▶ valoarea varianţei (dispersiei) de selecţie

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

▷ Varianţa de selecţie S_n^2 este un estimator nedeplasat şi consistent pentru varianţa teoretică $V(X)$ a caracteristicii X; se folosesc simulări pentru $V(X) \approx s_n^2$;
`numpy.var(..., ddof=1)`
▶ abaterea standard de selecţie (empirică)

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

▶ valoarea abaterii standard de selecţie

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

▷ Deviaţia standard de selecţie S_n **nu** este un estimator nedeplasat pentru deviaţia standard teoretică $Std(X) = \sqrt{V(X)}$ a caracteristicii X; el este un estimator consistent pentru deviaţia standard teoretică $Std(X)$ a caracteristicii X; în simulări se foloseşte $Std(X) \approx s_n$; `numpy.std(..., ddof=1)`
▶ momentul centrat de selecţie (empiric) de ordinul doi

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

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

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

▷ Momentul centrat de selecţie de ordinul doi M_n *nu este un estimator nedeplasat* pentru varianţa teoretică $V(X)$ a caracteristicii X; el este un estimator consistent pentru varianţa teoretică $V(X)$ a caracteristicii X (a se vedea Exemplul 3, pg. 52); se folosesc simulări pentru $V(X) \approx m_n$; `numpy.var(..., ddof=0)`
▶ funcţia de repartiţie empirică $\mathcal{F}_n : \mathbb{R} \times \Omega \to [0, 1]$

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

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

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

▷ Funcţia de repartiţie de selecţie $\mathcal{F}_n(x, \cdot)$ calculată în $x \in \mathbb{R}$ este un estimator nedeplasat şi consistent pentru $F_X(x)$, care este valoarea funcţiei de repartiţie teoretice calculată în x; în simulări $F_X(x) \approx \mathcal{F}_n(x)$; `scipy.stats.ecdf`

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

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

este un estimator *nedeplasat* şi *consistent* pentru parametrul necunoscut p.
Se dau datele statistice $x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 0, x_5 = 0, x_6 = 1, x_7 = 1, x_8 = 0$. Să se calculeze valoarea estimatorului $\hat{p}$.
(b) Fie $n > 2$. Considerăm estimatorul $\bar{p}(X_1, ..., X_n) = \frac{1}{2}(X_1 + X_n)$. Este estimatorul $\hat{p}$ mai eficient decât estimatorul $\bar{p}$?
R.: (a) $X \sim Bernoulli(p) \Longrightarrow E(X) = p, V(X) = p(1 - p)$;

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

LTNM (a se vedea P.16) implică

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

Deci, $\hat{p}(X_1, ..., X_n)$ este un estimator nedeplasat şi consistent pentru parametrul necunoscut p.
Folosind datele statistice $x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 0, x_5 = 0, x_6 = 1, x_7 = 1, x_8 = 0$, valoarea estimată pentru p este

$$p \approx \hat{p}(x_1, ..., x_8) = \frac{1}{8}(x_1 + ... + x_8) = \bar{x}_8 = \frac{3}{8} = 0.375\,.$$

(b) Observăm $E(\bar{p}) = p; V(\bar{p}) = \frac{p(1-p)}{2} > V(\hat{p}) = \frac{p(1-p)}{n}$. Deci estimatorul $\hat{p}$ este *mai eficient* decât estimatorul $\bar{p}$. ♢
Exemplul 2: (a) Fie $(X_n)_n$ un şir de variabile de selecţie pentru caracteristica X. Varianţa empirică $S_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2$ este un estimator nedeplasat şi consistent pentru

varianţa teoretică $V(X)$ a caracteristicii X.
R.: (a) Notăm $m = E(X), \sigma^2 = V(X)$. Are loc

$$(n-1)S_n^2 = \sum_{k=1}^{n}(X_k - \bar{X}_n)^2 = \sum_{k=1}^{n}(X_k^2 - 2X_k\bar{X}_n + \bar{X}_n^2) = \sum_{k=1}^{n} X_k^2 - n\bar{X}_n^2$$
$$= \sum_{k=1}^{n} X_k^2 - \frac{1}{n}\Big(\sum_{k=1}^{n} X_k\Big)^2 = \frac{n-1}{n}\sum_{k=1}^{n} X_k^2 - \frac{1}{n}\sum_{\substack{k,j=1\\k\neq j}}^{n} X_kX_j.$$

$X_1, \ldots, X_n$ sunt v.a. independente şi au aceeaşi distribuţie ca X:

$$E(X_kX_j) = E(X_k)E(X_j) = m^2 \qquad \forall k, j \in \{1, ..., n\}, k \neq j$$
$$E(X_k^2) = V(X_k) + E^2(X_k) = V(X) + E^2(X) = \sigma^2 + m^2 \qquad \forall k \in \{1, ..., n\}$$

Scriem succesiv

$$E(S_n^2) = \frac{1}{n-1}\left(\frac{n-1}{n} \cdot n(\sigma^2 + m^2) - \frac{1}{n} \cdot n(n-1)m^2\right) = \sigma^2 = V(X)$$

$\Longrightarrow S_n^2$ este estimator nedeplasat pentru $V(X)$.

Mai sus s-a demonstrat că

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

LTNM (a se vedea P.16) implică

$$\frac{1}{n}(X_1^2 + ... + X_n^2) \xrightarrow{a.s.} \sigma^2 + m^2 \implies \frac{1}{n-1}(X_1^2 + ... + X_n^2) \xrightarrow{a.s.} \sigma^2 + m^2$$

şi

$$\bar{X}_n^2 = \frac{1}{n}(X_1 + ... + X_n) \xrightarrow{a.s.} m \implies \frac{n}{n-1}\bar{X}_n^2 \xrightarrow{a.s.} m^2.$$

În concluzie $S_n^2 \xrightarrow{a.s.} \sigma^2$, deci S_n^2 este estimator consistent pentru $\sigma^2 = V(X)$.
(b) Deoarece $M_n = \frac{n-1}{n}S_n^2$, obţinem că $E(M_n) = \frac{n-1}{n}E(S_n^2) = \frac{n-1}{n}V(X)$, deci M_n **nu** este un estimator nedeplasat pentru $V(X)$, dar este un estimator consistent pentru $V(X)$.

▲

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

de exemplu:

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

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

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

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

Exemplu 1: Folosind metoda momentelor, să se estimeze parametrul necunoscut $\theta := a$ pentru $X \sim Unif[0, a]$; se dau datele statistice: 0.1,0.3,0.9,0.49,0.12,0.31,0.98,0.73, 0.13,0.62.

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

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

Valoarea estimatorului este

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

Estimatorul pentru parametrul necunoscut a este

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

Parametrul necunoscut a este estimat cu valoarea 0.936.
► Este $\hat{a}(X_1, ..., X_n)$ un estimator nedeplasat pentru parametrul a?
R.: Da, se arată că $E(\hat{a}(X_1, ..., X_n)) = a$. ♡

Exemplu 2:
Folosind metoda momentelor, să se estimeze parametrii necunoscuţi $\theta_1 := \mu$ şi $\theta_2 = \sigma^2$ 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 $n = 10$, $x_1, ..., x_n$ sunt datele statistice, iar $X_1, ..., X_n$ sunt variabile de selecţie. 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. 46). Se rezolvă

$$\begin{cases} \mu & = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i \\ \sigma^2 + \mu^2 & = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^2 \end{cases} \Longrightarrow \text{ are soluţia } \begin{cases} \hat{\mu}(x_1, ..., x_n) & = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i \\ \hat{\sigma}^2(x_1, ..., x_n) & = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i^2 - \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i\right)^2 \end{cases}$$

Valorile estimatorilor sunt

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

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

Estimatorii sunt

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

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

♣

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

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

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

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

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

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

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

Deseori este mai practic să se considere varianta transformată $\frac{\partial \ln L}{\partial \theta} = 0$ cu $\frac{\partial^2 \ln L}{\partial \theta^2} < 0$. În unele situaţii (1) se rezolvă prin alte metode; de exemplu în cazul în care $\frac{\partial L}{\partial \theta} = 0$ nu are soluţie (echivalent cu $\frac{\partial \ln L}{\partial \theta} = 0$ nu are soluţie). Reamintire: dacă $a, b > 0$, atunci au loc proprietăţile:

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

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

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

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

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

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

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

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

Estimatorul de verosimilitate maximă pentru parametrul necunoscut p este

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

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

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

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

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

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

Se poate lucra şi cu varianta transformată:

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

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

Intervale de încredere şi teste statistice

Noţiuni de bază

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

Def. 29. ***Cuantila de ordin*** *α pentru distribuţia caracteristicii cercetate X este numărul $z_\alpha \in \mathbb{R}$ pentru care*

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

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

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

• $\alpha \cdot 100\%$ din valorile lui X sunt mai mici sau egale cu z_α
De exemplu, pentru $\alpha = 0.5$ şi X v.a.continuă: 50% din valorile aleatoare ale lui X sunt mai mici sau egale cu $z_{0.5}$ (mediana), adică $P(X \le z_{0.5}) = 0.5$.

Exemple: (1) Fie $X \sim Exp(\lambda)$. Să se determine cuantila de ordin α.
R: Funcţia de repartiţie este $F(x) = 1 - e^{-\lambda x}$, pentru $x > 0 \Longrightarrow$ cuantila de ordin α este $z_\alpha = -\frac{\ln(1-\alpha)}{\lambda}$.
(2) Fie $X \sim \begin{pmatrix} 1 & 3 & 5 & 7 \\ 0.2 & 0.35 & 0.35 & 0.1 \end{pmatrix}$ v.a. discretă. Să se determine mediana $z_{0.5}$.

R: $P(X < 3) = 0.2 \le 0.5 \le P(X \le 3) = 0.2 + 0.35 = 0.55 \Longrightarrow z_{0.5} = 3$ este mediana.

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

$\triangleright$ distibuţia normală $N(0,1)$
funcţia de repartiţie $F_{N(0,1)}(x)=$`norm.cdf`$(x,0,1)$;
cuantila $z_\alpha=$ `norm.ppf`$(\alpha,0,1)$, adică $F_{N(0,1)}(z_\alpha)=\alpha$;
$\triangleright$ distibuţia Student $T(n)$
funcţia de repartiţie $F_{T(n)}(x)=$ `t.cdf`(x,n);
cuantila $t_\alpha=$ `t.ppf`(α,n), adică $F_{T(n)}(t_\alpha)=\alpha$;
$\triangleright$ distibuţia Chi-pătrat $\chi^2(n)$
funcţia de repartiţie $F_{\chi^2(n)}(x)=$ `chi2.cdf`(x,n);
cuantila $c_\alpha=$ `chi2.ppf`(α,n), adică $F_{\chi^2(n)}(c_\alpha)=\alpha$;
Exemple: `norm.ppf`$(0.01,0,1)=-2.3263$, `norm.ppf`$(1-0.01,0,1)=2.3263$,
`t.ppf`$(0.05,10)=-1.8125$, `t.ppf`$(1-0.05,10)=1.8125$,
`chi2.ppf`$(0.05,10)=3.9403$, `chi2.ppf`$(1-0.05,10)=18.307$.
`ppf - percent point function (inverse of cdf)`

```
#Exemplu - Cuantile
from scipy.stats import norm, t, chi2
alfa=0.01
z1=norm.ppf(alfa,0,1)
z2=norm.ppf(1-alfa,0,1)
print("Cuantile ale distributiei N(0,1):", "z_alfa=",z1,"z_{1-alfa}=",z2)
n=10
alfa=0.05
t1=t.ppf( alfa,n )
t2=t.ppf(1-alfa,n)
print(f"Cuantile ale distributiei Student T({n}):", "t_alfa=",t1,"t_{1-alfa}=",t2)
c1=chi2.ppf(alfa,n)
c2=chi2.ppf(1-alfa,n)
print(f"Cuantile ale distributiei Chi-Patrat({n}):", "c_alfa=",c1,"c_{1-alfa}=",c2)
```

P. 18. *Pentru orice $\alpha\in(0,1)$ cuantilele distribuţiei normale $N(0,1)$ satisfac relaţia $z_\alpha=-z_{1-\alpha}$, iar pentru cuantilele distribuţiei Student $T(n)$ are loc $t_\alpha=-t_{1-\alpha}$.*

Observaţie: P.18 implică $z_{0.5}=t_{0.5}=0$.

Intervale de încredere

În paragrafele anterioare s-a văzut cum poate fi estimat (punctual) un parametru necunoscut, folosind datele dintr-un eşantion. Se pune problema cât este de bună această estimare a parametrului necunoscut, adică vom calcula o anumită ”marjă de eroare”. Presupunem că studiem media (teoretică) a timpului de aşteptare la un anumit ghişeu al unei bănci. Prin studierea unui eşantion de volum 200 s-a constatat că media de selecţie a timpului de aşteptare este $\bar{x}_{200} = 10$ (minute). Dacă considerăm un alt eşantion probabil obţinem o altă valoare pentru $\bar{x}_{200}$.

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

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

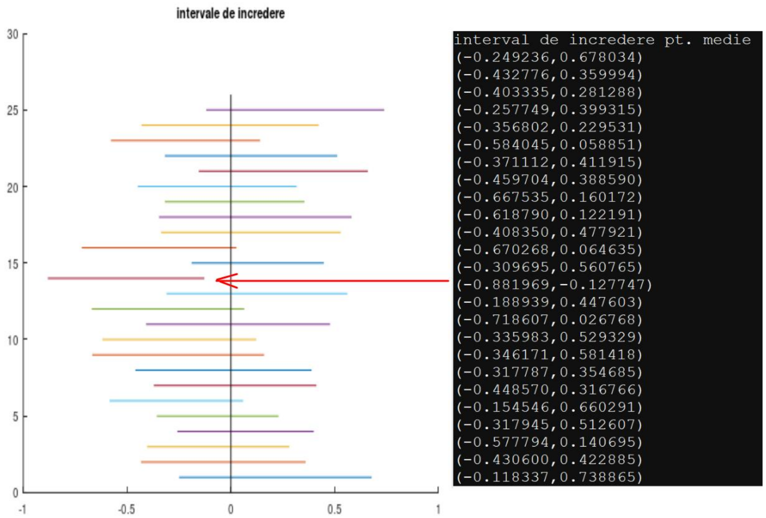
$$P\Big(g_1(X_1, \ldots, X_n) < \theta < g_2(X_1, \ldots, X_n)\Big) = 1 - \alpha$$

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

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

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

➡ **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 extremităţile sale sunt v.a. Interpretarea corectă a lui $1 - \alpha$ este următoarea: dacă, facem un număr foarte mare de selecţii (din mai multe eşantioane) şi calculăm de fiecare dată intervalul de încredere cu nivelul de



În această simulare: din 25 de intervale de încredere, un interval nu conţine *valoarea reală* 0; parametrul necunoscut este θ=media teoretică; datele statistice au fost generate, cu `norm.rvs(0,1)`, iar $1-\alpha$=0.95

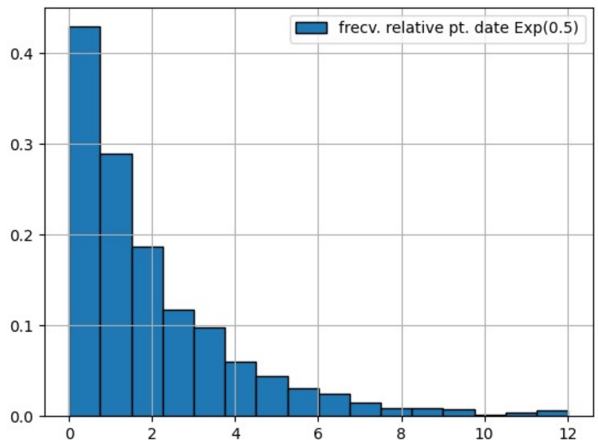
încredere $1-\alpha$, atunci $(1-\alpha)\cdot 100\%$ din aceste intervale conţin valoarea reală a parametrului θ. În exemplul de mai sus un interval de încredere, din cele 25 construite, nu conţine valoarea reală 0.

Recapitulare (notaţii)

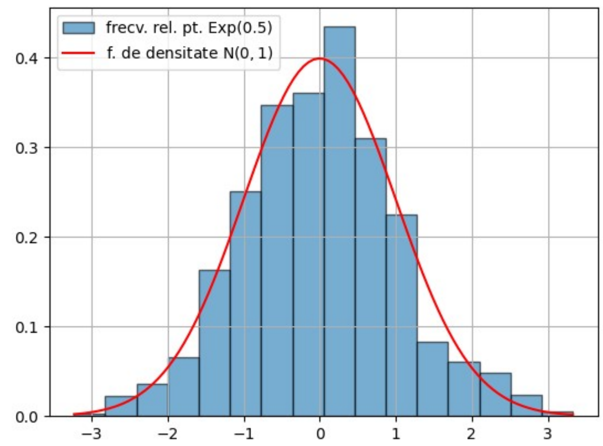
Variabilele de selecţie pentru caracteristica X: $X_1, ..., X_n$ sunt v.a. independente, au aceeaşi distribuţie ca X	datele statistice pentru caracteristica X: $x_1, ..., x_n$ sunt valorile (numerice) ale v.a. $X_1, ..., X_n$
Estimator	**Valoarea estimatorului**
media de selecţie $\bar{X}_n = \frac{1}{n}(X_1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

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

Reamintim: $X \sim N(\mu, \sigma^2)$ are $E(X) = \mu$, $V(X) = \sigma^2$ (calculele pe pg. 46).



(a) Histogramă cu 16 clase construită din 1000 date aleatoare $Exp(0.5)$



(b) Histogramă cu 16 clase construită din 1000 date aleatoare pentru $\frac{\bar{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}}$, iar $X_1, ..., X_n \sim Exp(0.5), n = 800$

Exemplificare TLC pentru $\frac{\bar{X}_n-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$, pentru $X_1, ..., X_n \sim Exp(0.5), n = 800, \mu = \sigma = 2$

P. 20. (Teorema limită centrală - TLC) *Fie $X_1, \ldots, X_n$ variabile de selecţie pentru caracteristica X. Fie $\mu = E(X_k)$ şi $\sigma^2 = V(X_k) > 0 \; \forall \, k \in \{1, ..., n\}$.*

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) \text{ pentru } n \text{ suficient de mare } (n > 30).$$

▷ TLC afirmă: distribuţia versiunii standardizate[2] a mediei de selecţie converge către distribuţia normală standard $N(0,1)$, chiar dacă variabilele de selecţie (caracteristica cercetată X) nu urmează o distribuţie normală
▷ echivalent putem reformula TLC astfel: media de selecţie $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$ este approximativ normal distribuită cu media μ şi varianţa $\frac{\sigma^2}{n}$, adică $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$.

➡ **Consecinţă** (la P. 20): pentru orice $a < b$ are loc
$$P\left(a < \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < b\right) \approx F_{N(0,1)}(b) - F_{N(0,1)}(a) = \texttt{norm.cdf}(b,0,1)\text{-}\texttt{norm.cdf}(a,0,1)$$
pentru n suficient de mare $(n > 30)$.

Exemplu: Dacă $(X_n)_{1 \le n \le 100}$ sunt variabile de selecţie pentru caracteristica $X \sim Bernoulli(0.5)$, să se estimeze $P(0.35 < \bar{X}_{100} < 0.65)$, folosind P.20 (TLC).

[2] Dacă Z este o v.a. (cu $V(Z) > 0$) atunci v.a. $\tilde{Z} := \frac{Z - E(Z)}{Std(Z)}$ este **versiunea standardizată** a v.a. Z, iar $E(\tilde{Z}) = 0$ şi $V(\tilde{Z}) = 1$.

R.: Se calculează $\mu = E(X_n) = E(X) = 0.5$ şi $\sigma = \sqrt{V(X_n)} = \sqrt{V(X)} = 0.5$ şi se scrie

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

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

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

aşadar pentru o caracteristică de tip $Bernoulli(0.5)$, media de selecţie $\bar{X}_{100}$ aparţine cu o probabilitate foarte mare intervalului $(0.35, 0.65)$.
Observaţie: $(0.35, 0.65)$ *nu* este valoarea unui interval de încredere! ♡

Interval de încredere pentru media $\mu = E(X)$ a caracteristicii cercetate X, când varianţa $\sigma^2 = V(X)$ 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 selecţie 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 obţinut 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 selecţie corespunzătoare caracteristicii cercetate X
▶ construim intervale de încredere pentru parametrul *necunoscut* $\mu = E(X)$
▶ dacă $X \sim N(\mu, \sigma^2)$ sau $n > 30$ şi X are o distribuţie necunoscută, atunci P20, respectiv P.19, implică

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

▶ cuantilele legii normale $N(0, 1)$:
$z_{1-\frac{\alpha}{2}} = \texttt{norm.ppf}(1-\frac{\alpha}{2}, 0, 1)$, $z_{1-\alpha} = \texttt{norm.ppf}(1-\alpha, 0, 1)$, $z_\alpha = \texttt{norm.ppf}(\alpha, 0, 1)$

• un *interval de încredere bilateral* pentru $\mu = E(X)$ (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:

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

$$\overset{(5)}{=} F_{N(0,1)}(z_{1-\frac{\alpha}{2}}) - F_{N(0,1)}(-z_{1-\frac{\alpha}{2}}) \overset{P.18}{=} F_{N(0,1)}(z_{1-\frac{\alpha}{2}}) - F_{N(0,1)}(z_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha$$

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

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

Interval de încredere pentru **media** $E(X)$ când varianţa $\sigma^2 = V(X)$ este cunoscută:	Expresia intervalului de încredere, folosind datele statistice
bilateral	$\Big(\bar{x}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\ ,\ \bar{x}_n + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\Big)$
unilateral stâng (oferă limită superioară)	$\Big(- \infty\ , \bar{x}_n - \frac{\sigma}{\sqrt{n}} \cdot z_\alpha\Big)$
unilateral drept (oferă limită inferioară)	$\Big(\bar{x}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}\ ,\ \infty\Big)$

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

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

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

Exerciţiu: Cum se modifică intervalul de încredere bilateral introdus pentru media teoretică, dacă:
(a) creşte deviaţia standard σ?
(b) creşte nivelul de încredere $1 - \alpha$?
(c) creşte volumul n al eşantionului?