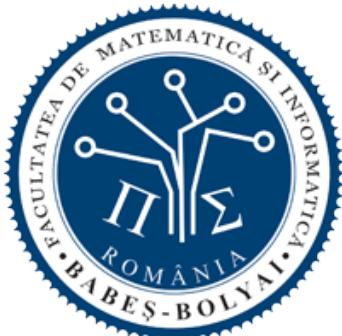


Curs - Probabilități și Statistică 2025/2026
Secția Informatică

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



Teoria Probabilităților

Teoria probabilităților este o disciplină a matematicii care se ocupă de **studiu fenomeneelor aleatoare**.

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

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

Fenomene și procese aleatoare apar, de exemplu, în:

→ pariuri, loto (6 din 49), jocuri de noroc / jocuri online

→ previziuni meteo

→ previziuni economice / financiare, investiții, cumpărături online (predicția comportamentului clienților)

→ sondaje de opinie (analiza unor strategii politice), asigurări (evaluarea riscurilor / pierderilor)



[Sursa: www.financialmarket.ro]

→ în informatică:

▷ sisteme de comunicare, prelucrarea informației, modelarea traficului în rețea, criptografie;

▷ analiza probabilistică a performanței unor algoritmi, fiabilitatea sistemelor, predicții în cazul unor sisteme complexe;

▷ algoritmi de simulare, machine learning, data mining, recunoașterea formelor / a vocii;

▷ generarea de numere aleatoare (pseudo-aleatoare), algoritmi aleatori

▷ 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 dezintegreză sunt complet imprevizibile), sau variațiile de amplitudine din perturbările atmosferice (*atmospheric noise*, folosit de <https://www.random.org/randomness/>), sau comportamentul fotonilor (în mecanica cuantică, când un foton loveste un separator de fascicule -*beam splitter*-, fotonul are sănă de 50% de a fi reflectat și 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 execuție, spațiul de stocare, rezultatul obținut sunt variabile aleatoare (chiar dacă se folosesc aceleași valori input)
- ▷ la anumite tipuri de algoritmi corectitudinea e garantată doar cu o anumită probabilitate.
- Algoritm de tip **Las Vegas** este un algoritm aleator, care returnează la fiecare execuție rezultatul corect (independent de alegerile aleatoare făcute); durata de execuție este o variabilă aleatoare.

Exemplu: Random QuickSort

- Un algoritm aleator pentru care rezultatele obținute sunt corecte *doar* cu o anumită probabilitate se numește algoritm **Monte Carlo**.
 - se examinează probabilitatea cu care rezultatul este corect; probabilitatea de eroare poate fi scăzută semnificativ prin execuții repetitive, independente.

Exemplu:

- ▷ testul Miller-Rabin, care verifică dacă un număr natural este prim sau este număr compus; testul returnează fie răspunsul „numărul este sigur un număr compus” sau răspunsul „numărul este probabil un număr prim”.

Exercițiu: Fie S un vector cu 60 de elemente, din mulțimea $\{0, 1, 2\}$ (ordinea lor este necunoscută; se presupune că sirul conține cel puțin un 0).

→ 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 iterări.

```
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 găsește niciun 0.

Noțiuni introductive:

- **Experiență aleatoare** (experimentul aleator) este acea experiență al cărei rezultat nu poate fi cunoscut decât după încheierea ei.
- **Evenimentul** este rezultatul unei experiențe aleatoare.

Exemple:

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

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

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

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

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


Operații cu evenimente

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

Relații între evenimente

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

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

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

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

asociative

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

și **distributive**

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

satisfac **legile lui De Morgan**

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

Frecvența relativă și frecvența absolută

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

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

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

Definiția clasică a probabilității

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

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

► Prin repetarea de multe ori a unui experiment, în condiții practic identice, frecvența relativă $f_n(A)$ de apariție a evenimentului A este aproximativ egală cu $P(A)$: $f_n(A) \approx P(A)$ pentru valori mari ale lui n .

► Din punct de vedere probabilistic sirul $(f_n(A))_n$ „converge aproape sigur” către $P(A)$ când $n \rightarrow \infty$.

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

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

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

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

Exercițiu: (1) Se alege aleator un număr din mulțimea $\{1, 2, 3, \dots, 99\}$. 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 aplicații web,
- ▷ 6 servere sunt dedicate sarcinilor de învățare automată.

Un nou proces este atribuit aleator unui 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 învățare automată?



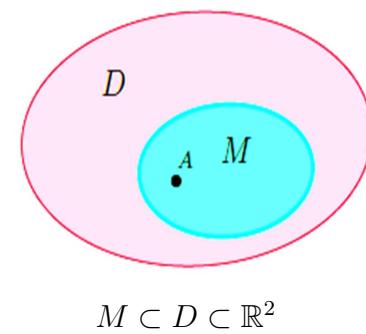
Definiția axiomatică a probabilității

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

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

Alegem aleator un punct $A \in D$ (în acest caz spațiul de selecție este D). Probabilitatea geometrică a evenimentului “ $A \in M$ ” este

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



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

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

→ \mathcal{K} este o mulțime de evenimente și are structura unei σ -algebrelor (vezi Def. 4)

→ P satisface anumite axiome (vezi Def. 5)

Def. 4. O familie \mathcal{K} de evenimente din spațiul de selecție Ω se numește **σ -algebră** dacă sunt satisfăcute condițiile:

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

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

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

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

este o σ -algebră pe mulțimea B . \diamond

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

- (1) $\emptyset, \Omega \in \mathcal{K}$;
(2) $A, B \in \mathcal{K} \implies A \cap B, A \setminus B \in \mathcal{K}$;
(3) $A_n \in \mathcal{K}$, $n \in \mathbb{N}^* \implies \bigcap_{n=1}^{\infty} A_n \in \mathcal{K}$.

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

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

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

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

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

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

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

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

$$P(\{n\}) = \frac{1}{2^{n+1}}, n \in \mathbb{N}$$

$$P(\{n_1, \dots, n_k, \dots\}) = \frac{1}{2^{n_1+1}} + \dots + \frac{1}{2^{n_k+1}} + \dots, \text{ unde } \{n_1, \dots, n_k, \dots\} \subseteq \mathbb{N}.$$

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 spațiu de probabilitate; Def. 5-(3) este îndeplinită, datorită teoremei din analiză, care afirmă că pentru o serie cu termeni pozitivi, schimbarea ordinii termenilor seriei nu schimbă natura seriei și nici suma ei. ♣

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

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

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

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

Exemplu: Dintr-un pachet de 52 de cărți de joc se extrage o carte aleator. Care este probabilitatea p de a extrage a) un as sau o damă de pică? b) o carte cu inimă sau un as?

R.: a) A : s-a extras un as; D : s-a extras damă de pică; A și D sunt două evenimente disjuncte (incompatibile)

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

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

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



Evenimente independente

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

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

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

Exercițiu: Se aruncă un zar de două ori.

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

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

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

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

- (1) A și B sunt independente.
- (2) \bar{A} și B sunt independente.
- (3) A și \bar{B} sunt independente.
- (4) \bar{A} și \bar{B} sunt independente.

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

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

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

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

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

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

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

verde și una este colorată având cele trei culori. Se aruncă tetraedrul și se consideră evenimentele:

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

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

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

Sunt cele 3 evenimente *independente în totalitate*?

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

3) O firmă utilizează două sisteme de securitate independente pentru a detecta activitatea suspectă a rețelei: un firewall care detectează o astfel de activitate cu o probabilitate de 0.7 și un antivirus care o detectează cu o probabilitate de 0.8.

Presupunând că firewall-ul și antivirusul funcționează independent, care este probabilitatea ca:

(a) Ambele sisteme detectează activitatea suspectă?

(b) Cel puțin un sistem detectează activitatea suspectă? ◆

Exemplu istoric - Joc de zaruri (sec. XVII): Un pasionat jucător de zaruri, cavalerul de Méré, susținea în discuțiile sale cu B. Pascal că a arunca un zar de 4 ori pentru a obține cel puțin o dată față șase, este același lucru ca să arunca de 24 ori câte două zaruri pentru a obține cel puțin o dublă de șase. Cu toate acestea, cavalerul de Méré a observat că jucând în modul al doilea (cu două zaruri aruncate de 24 ori), pierdea față de adversarul său, dacă acesta alegea primul mod (aruncarea unui singur zar de 4 ori). Pascal și Fermat au arătat că probabilitatea de câștig la jocul cu un singur zar aruncat de 4 ori este $p_1 \approx 0.5177$, iar probabilitatea $p_2 \approx 0.4914$ la jocul cu două zaruri aruncate de 24 de ori. Deși diferența dintre cele două probabilități este mică, totuși, la un număr mare de partide, jucătorul cu probabilitatea de câștig p_1 câștigă în fața jucătorului cu probabilitatea de câștig p_2 . Practica jocului confirmă astfel justitia raționamentului matematic, contrar credinței lui de Méré.

Estimăm prin simulări Python probabilitățile următoarelor evenimente:

A: se obține cel puțin un 6 în 4 aruncări ale unui zar;

B: se obține cel puțin o pereche (6,6) în 24 de aruncări a două zaruri;

C: se obține cel puțin o pereche (6,6) în 25 de aruncări a două zaruri.

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

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

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

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

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

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

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

```
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 condiționată

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

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

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

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

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

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

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

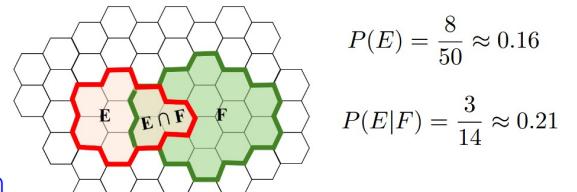
► $P(A|B)$: probabilitatea condiționată a lui A de către B , este **probabilitatea de a se realiza evenimentul A dacă în prealabil s-a realizat evenimentul B** .

► Fie evenimentele $A, B \in \mathcal{K}$ astfel încât $P(A) > 0$ și $P(B) > 0$. Evenimentele A și B sunt **independente** (a se vedea Def. 6), dacă apariția evenimentului A , nu influențează apariția evenimentului B și invers, adică

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

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

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



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

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

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

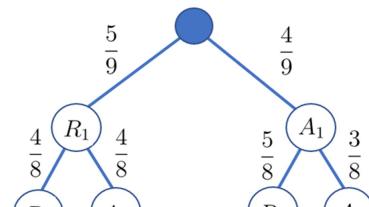
R.: pentru $i \in \{1, 2\}$ fie evenimentele

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

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

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

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

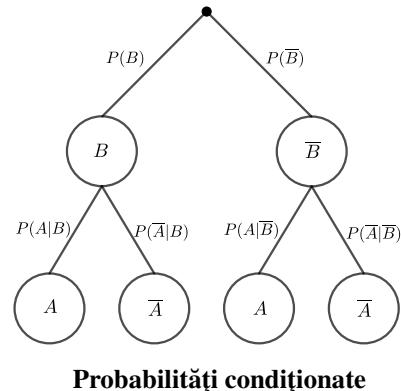


Extragere fără retrunare

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

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

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



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

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

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

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

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

R.: 7 bile albe: 1, 2, 3, 4, 5, 6, 7; 6 bile roșii: 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 roșie;

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 și P.4, scriem succesiv

$$\begin{aligned} p_2 &= P(I_2|R_1) = \frac{P(I_2 \cap R_1)}{P(R_1)} = \frac{P(I_2 \cap R_1 \cap I_1) + P(I_2 \cap R_1 \cap \bar{I}_1)}{P(R_1)} \\ &= \frac{P(I_2|R_1 \cap I_1)P(R_1 \cap I_1) + P(I_2|R_1 \cap \bar{I}_1)P(R_1 \cap \bar{I}_1)}{P(R_1)} = \frac{\frac{6}{12} \cdot \frac{3}{13} + \frac{7}{12} \cdot \frac{3}{13}}{\frac{6}{13}} = \frac{13}{24}. \end{aligned}$$



Exemplu: Ce probabilități calculează programul de mai jos?

② Care sunt valorile teoretice pentru p_1 , p_2 , p_3 , 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 înmulțirii probabilităților)

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

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

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

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

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

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

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

- a) prima bilă să fie verde, iar cea de-a doua albastră?
- b) cele 2 bile să aibă aceeași culoare?
- c) a doua bilă să fie albastră?
- d) prima bilă să fie verde, știind că a doua este albastră?
- e) se mai extrage o a treia bilă; se cere probabilitatea ca prima bilă să fie verde, cea de-a doua albastră și a treia tot albastră.

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

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

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

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

c) folosim formula probabilității totale P.7:

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

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

e) formula de înmulțire a probabilităților P.6:

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

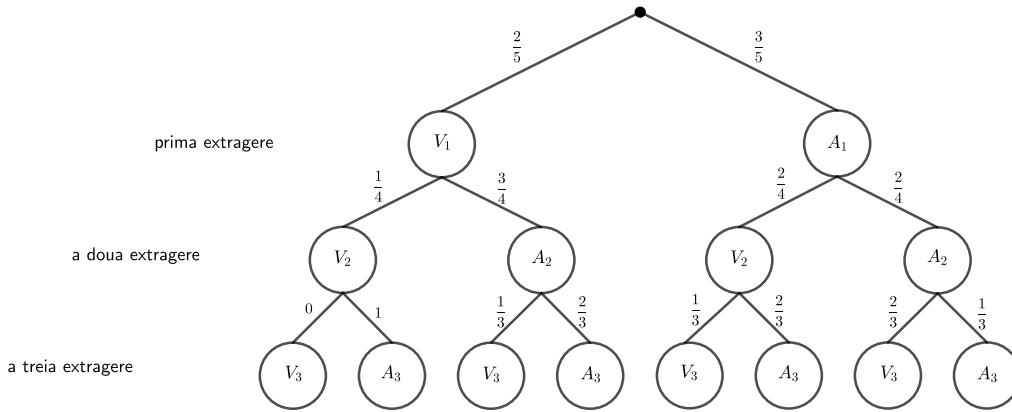


Fig. 3. Extragere fără returnare



Formula lui Bayes

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

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

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

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

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

► Se cunosc:

→ $P(E|H)$ probabilitatea ca să apară datele, știind că ipoteza H este adevarată;

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

acestea reprezintă verosimilitatea (engl. *likelihood*) datelor observate (a informațiilor / evidențelor).

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 conține cuvântul *succes*.

Se cunosc probabilitățile

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

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

Care este probabilitatea ca un email să fie spam, știind că emailul conține 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 partiția $\{S, \bar{S}\}$)

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

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

► Probabilitatea ca un email să fie clasificat spam, știind că emailul conține cuvântul *succes* este 0.636.

- ▷ Dacă un email conține cuvântul *succes*, atunci există aproximativ 63.6% şanse să fie clasificat spam.



P. 7. (Formula lui Bayes)

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

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

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

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

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

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

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

Exemplu (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 predicție / prognoză asupra unui email ales aleator, dacă acesta este spam sau non-spam cu ajutorul filtrului de spam.

Au fost colectate următoarele date statistice:

- ▷ $AP = 400$ (adevărat pozitiv) Numărul de emailuri care sunt de fapt spam și care au fost clasificate ca spam de către filtrul de spam ; $\#(H \cap E)^1$
- ▷ $FP = 210$ (fals pozitiv) Numărul de emailuri care sunt de fapt non-spam și care au fost clasificate ca fiind spam de filtrul de spam ; $\#(\bar{H} \cap E)$
- ▷ $FN = 310$ (fals negativ) Numărul de emailuri care sunt de fapt spam și care au fost clasificate ca non-spam de filtrul de spam ; $\#(H \cap \bar{E})$
- ▷ $AN = 1200$ (adevărat negativ) Numărul de emailuri care sunt de fapt non-spam și care au fost clasificate ca non-spam de filtrul de spam; $\#(\bar{H} \cap \bar{E})$.

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

| | | starea actuală (realitatea) | | predicția |
|--|--|-----------------------------|---------------------------------|-------------|
| | | H (email este spam) | \bar{H} (email este non-spam) | |
| 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) | | | predicția |
|--|--|------------------------------|---------------------------------|-------|-----------|
| | | H : email este spam | \bar{H} : email este non-spam | total | |
| 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 elemente din $H \cap E$.

Pe baza datelor statistice: a) probabilitatea ca un email, despre care se știe 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 știe 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 | <i>machine learning (ML)</i> |
| măsuri de performanță | <i>measuring the performance of a binary classification model</i> |
| valoarea predictivă pozitivă= $\frac{AP}{AP+FP}$ | <i>positive predictive value; precision</i> |
| valoarea predictivă negativă= $\frac{AN}{AN+FN}$ | <i>negative predictive value</i> |
| sensibilitatea= $\frac{AP}{AP+FN}$ | <i>recall; probability of detection; true positive rate</i> |
| specificitatea= $\frac{AN}{AN+FP}$ | <i>true negative rate</i> |
| acuratețea= $\frac{AP+AN}{AP+FP+AN+FN}$ | <i>accuracy</i> |

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

Variable aleatoare

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

Exemplu: 1) La aruncarea a două zaruri, suma numerelor obținute este o variabilă aleatoare

$S : \Omega \rightarrow \{2, 3, \dots, 12\}$, unde Ω conține toate evenimentele elementare ce se pot obține la aruncarea a două zaruri, adică $\Omega = \{(\omega_i^1, \omega_j^2) : i, j = \overline{1, 6}\}$, unde

(ω_i^1, ω_j^2) este evenimentul elementar: la primul zar s-a obținut numărul i și la al doilea zar s-a obținut numărul j , unde $i, j = \overline{1, 6}$.

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

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

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

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

Notație 1. *variabilă / variabile aleatoare* → v.a.

O variabilă aleatoare este:

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

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

V.a. discrete: exemple de *v.a. numerice discrete*: numărul produselor defecte produse de o anumită linie de producție 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* (→ se clasifică în categorii): prognoza meteo: *ploios, senin, înnorat, cețos*; calitatea unor servicii: *nesatisfăcătoare, satisfăcătoare, bune, foarte bune, exceptionale*, etc.

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

Variabile aleatoare numerice - definiție formală

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

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

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

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

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

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

- ▷ O variabilă aleatoare discretă X este caracterizată de distribuția 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 $i \in I$.
- $\mathbb{X} = (X_1, \dots, X_m)$ este un **vector aleator discret** dacă fiecare componentă a sa este o variabilă aleatoare discretă.

Distribuții discrete clasice

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

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

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

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

► Python: `scipy.stats.randint`

```
# Exemplu 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 frecvențele absolute:", count)
print("Valorile:", x, "au frecvențele relative:", count/N)
bar(x, count/N, width=0.8,color="cyan", edgecolor="black")
# deseneaza histograma frecvențelor relative
plt.grid()
plt.xlabel("valorile")
plt.ylabel("frecvențe relative")
plt.title("Unid(6)")
xticks(range(0,b))
show()
```

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

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

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

$X = 0 \Leftrightarrow$ dacă \bar{A} apare; $X = 1 \Leftrightarrow$ dacă A apare

$\Rightarrow X \sim Bernoulli(p)$ cu $p := P(A)$

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



► Python: `scipy.stats.bernoulli`

Distribuția 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: $X \in \{0, 1, \dots, n\}$

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

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

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 promotional către 100 de abonați. Pe baza unor date statistice, se știe că fiecare abonat citează emailul cu probabilitatea 0.25 (independent de ceilalți abonați). 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 variabilă aleatoare; X = numărul de emailuri deschise (din cele n trimise)
- ② Care este probabilitatea ca exact 30 de persoane să citească emailul promotional?
- ② Care este probabilitatea ca mai puțin de 20 de persoane să îl citească?



→ formula binomială $(a + b)^n = \sum_{k=0}^n C_n^k a^k b^{n-k}$ pentru $a = p$ și $b = 1 - p$ se obține

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

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 frecvențele relative:",count/N)
bar(z,count/N, width=0.8,color="yellow", edgecolor="black")
plt.grid()
plt.xlabel("valorile distributiei")
plt.ylabel("frecvențe 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()
```

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

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

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

▷ **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

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

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

Exerciții: 1) O rețea de laborator este compusă din 15 calculatoare. Rețeaua 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 puțin 10; c) exact 10

calculatoare?

2) Sondaj de opinie: O companie PR studiază modul în care oamenii interacționează cu mass-media în prezent. Pentru a înțelege comportamentul publicului, se realizează un sondaj. Fiecare participant la sondaj este întrebat: *Pentru a vă informa, ce tip de mass-media utilizați cel mai frecvent?*

Participanții trebuie să aleagă una dintre următoarele opțiuni:

Televiziune/Rețele sociale/Ziare tipărite/Știri online/Radio.

Pe baza cercetărilor anterioare și a tendințelor de interacțiune cu mass-media, următoarele date statistice sunt disponibile:

Televiziune: 30%

Rețele sociale: 35%

Ziare tipărite: 5%

Știri online: 20%

Radio: 15%

Compania dorește să calculeze probabilitatea de a observa următoarea distribuție într-un eșantion de 100 de persoane:

30 persoane aleg Televiziunea

40 aleg Rețelele sociale

0 aleg Ziarele tipărite

30 aleg Știrile online

0 aleg Radio.

Cu ce formulă se calculează această probabilitate?



Distribuția 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 și n_2 bile negre. Se extrag **fără returnare** n bile.

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

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

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

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

► Python: `scipy.stats.hypergeom`

▷ **Modelul urnei cu r culori și bilă neretrunată:** fie n_i = numărul inițial 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

$$\begin{aligned} P(X_1 = k_1, \dots, X_r = k_r) &= \text{probabilitatea de a obține } k_i \text{ bile având culoarea } c_i, i = \overline{1, r}, \\ &\text{din } n = k_1 + \dots + k_r \text{ extrageri fără returnarea bilei extrase,} \\ &= \frac{C_{n_1}^{k_1} \cdot \dots \cdot C_{n_r}^{k_r}}{C_{n_1+\dots+n_r}^n}. \end{aligned}$$

▷ (X_1, \dots, X_r) este un vector aleator discret și urmează **distribuția hipergeometrică multidimensională**

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

▷ Cazul $r = 2$ corespunde distribuției hipergeometrice.

► Python: `scipy.stats.multivariate_hypergeom`

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

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

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

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

Scriem

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

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

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

2) Loto 6 din 49: Care este probabilitatea de a nimeri exact 4 numere câștigătoare?

R.: Între cele 49 de bile exact $n_1 = 6$ sunt câștigătoare ("bilele albe") și $n_2 = 43$ necâștigătoare ("bilele negre"). Probabilitatea ca din $n = 6$ extrageri fără returnare, exact $k = 4$ numere să fie câștigătoare (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".

Scrieți distribuția de probabilitate a v.a. X .

3) O echipă de marketing se pregătește să testeze o nouă campanie promoțională. Baza lor de date cu clienți conține:

- 390 de clienți atenți la buget
- 310 de clienți fideli mărcii
- 250 de cumpărători impulsivi
- 50 de clienți VIP.

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

- 15 de clienți atenți la buget
- 10 clienți fideli mărcii
- 10 cumpărători impulsivi
- 5 clienți VIP?



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

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

- $A = \text{succes}$ cu $P(A) = p$, $\bar{A} = \text{insucces}$ $P(\bar{A}) = 1 - p$
- se repetă (independent) experimentul până apare prima dată A ("succes")

- v.a. X arată de câte ori apare \bar{A} (numărul de “insuccese”) până la apariția primului A (“succes”) \Rightarrow valori posibile: $X \in \{0, 1, \dots\}$

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

► Python: `scipy.stats.geom`; atenție valorile generate sunt de la 1; adică $P(Y = k) = p(1 - p)^{k-1}$ pentru $k \in \{1, 2, \dots\}$, 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 recepție corectă a mesajului $\Rightarrow X$ are distribuție geometrică.



Exercițiu: 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.

- Determinații distribuția de probabilitate a v.a. X .
- Calculați probabilitatea $P(X < 90)$.

Variabile aleatoare independente

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

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

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

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

V.a. X și Y sunt independente $\iff \forall (i, j) \in I \times J$ evenimentele A_i și B_j sunt independente (a se vedea Def. 6).

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

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

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

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

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

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

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

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

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

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

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

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

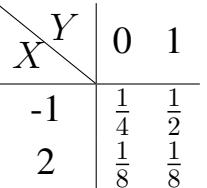
$$p_{ij} := P((X, Y) = (x_i, y_j)) = P(\{X = x_i\} \cap \{Y = y_j\}), \quad p_{ij} > 0 \quad \forall i \in I, j \in J,$$

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

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

| | | y_j | | | |
|-------|----------|----------|----------|----------|----------|
| | | \vdots | \vdots | \vdots | \vdots |
| x_i | \vdots | \dots | p_{ij} | \dots | \vdots |
| | \vdots | \vdots | \vdots | \vdots | \vdots |

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

următorul tabel: 

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

etc.

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

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

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

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

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

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

Operații cu variabile aleatoare (numerice)

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

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

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

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

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

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

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

Exerciții: (1) Fie X, Y v.a. independente, având distribuțiile

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

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

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

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

(3) Într-o echipă de dezvoltare software s-au implementat un set (o colecție) de 100 de teste independente, care verifică funcționalitatea unei noi aplicații. Aceste teste sunt ruleate 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 dorește modelarea probabilității ca un anumit număr de teste să treacă într-o execuție completă a setului de teste.

(a) Care este probabilitatea ca exact 95 de teste să treacă?

(b) Probabilitatea ca cel puțin 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 $100 - X$?



| | | |
|---|----|----|
| 1 | x1 | y1 |
| 2 | x1 | y2 |
| 3 | x2 | y1 |
| 4 | x2 | y1 |
| 5 | x2 | y2 |
| 6 | x2 | y2 |
| 7 | x3 | y1 |
| 8 | x3 | y1 |

Setul de date

| | | |
|---|----|----|
| 3 | x2 | y1 |
| 5 | x2 | y2 |
| 1 | x1 | y1 |
| 3 | x2 | y1 |
| 7 | x3 | y1 |
| 2 | x1 | y2 |
| 7 | x3 | y1 |
| 2 | x1 | y2 |
| 5 | x2 | y2 |
| 4 | x2 | y1 |

Bootstrap 1

| | | |
|---|----|----|
| 1 | x1 | y1 |
| 4 | x2 | y1 |
| 7 | x3 | y1 |
| 2 | x1 | y2 |
| 7 | x3 | y1 |
| 2 | x1 | y2 |
| 5 | x2 | y2 |
| 3 | x2 | y1 |

Bootstrap 2

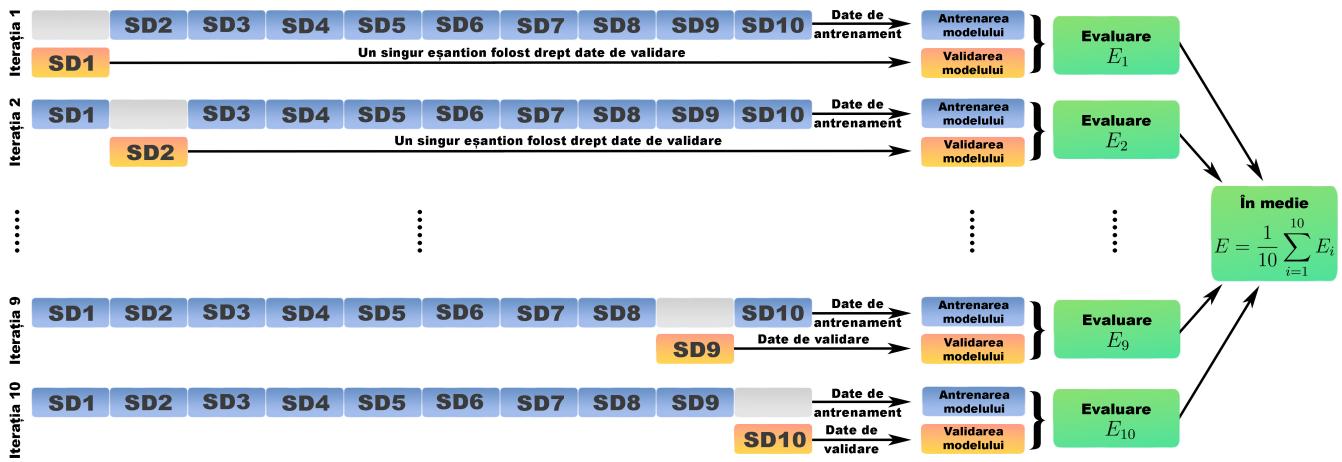
| | | |
|---|----|----|
| 5 | x2 | y2 |
| 2 | x1 | y2 |
| 1 | x1 | y1 |
| 7 | x3 | y1 |
| 6 | x2 | y2 |
| 8 | x3 | y1 |
| 3 | x2 | y1 |
| 1 | x1 | y1 |

Bootstrap 3

Metoda bootstrap

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

Bootstrapping este o procedură statistică care re-eșantinează un singur set de date pentru a crea mai multe eșantioane (folosind simulări). Aceste eșantioane sunt folosite pentru a face inferențe statistice asupra setului inițial de date.



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

► Metoda validării încrucișate (engl. *cross validation*)

Validarea încrucișată este o tehnică de evaluare a unui model de învățare automată și de testare a performanței acestuia. Metoda este folosită pentru compararea și selectarea unui model adecvat în cazul unei probleme specifice de modelare predictivă.

În cazul validării încrucișate (k -fold cross validation), eșantionul original de date este împărțit aleatoriu în k sub-eșantioane de dimensiuni egale. Din cele k sub-eșantioane, un singur sub-eșantion este folosit ca *date de validare* pentru testarea modelului, iar celelalte $k - 1$ sub-eșantioane sunt utilizate ca *date de antrenament*. Procesul de validare încrucișată se repetă apoi de k ori, fiecare dintre cele k sub-eșantioane fiind utilizat exact o dată ca date de validare. Avantajul acestei metode constă în faptul că toate observațiile sunt utilizate atât pentru antrenare, cât și pentru validare, iar fiecare observație este utilizată pentru validare exact o dată. Validarea încrucișată cu $k=10$ (sau $k=5$) este utilizată în mod obișnuit.

Atunci când $k = n$ (numărul de observații), validarea încrucișată este echivalentă cu validarea încrucișată numită în engleză *leave-one-out*.

Clasificarea naivă Bayes

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

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

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

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

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

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

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

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

R.: Dacă se cunoaște tipul zarului ales, atunci aruncările sunt în mod evident independente: $P(A_1 \cap A_2|Z_1) = \frac{1}{36} = P(A_1|Z_1) \cdot P(A_2|Z_1)$.

Din formula probabilității totale P.5 avem:

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

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

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

$$\Rightarrow P(A_1 \cap A_2) \neq P(A_1)P(A_2) \Rightarrow A_1 \text{ și } A_2 \text{ nu sunt independente.} \quad \clubsuit$$

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

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

Exemplu de clasificare naivă Bayes

| | Vreme | Timp | Trafic |
|----|---------------|---------------|------------------|
| 1 | înnorat | noapte | relaxat |
| 2 | zăpadă | seară | aglomerat |
| 3 | senin | noapte | relaxat |
| 4 | ploaie | seară | aglomerat |
| 5 | înnorat | amiază | aglomerat |
| 6 | senin | amiază | aglomerat |
| 7 | senin | dimineață | relaxat |
| 8 | ploaie | noapte | relaxat |
| 9 | înnorat | dimineață | aglomerat |
| 10 | zăpadă | noapte | aglomerat |
| 11 | senin | seară | relaxat |
| 12 | zăpadă | amiază | relaxat |
| 13 | înnorat | seară | aglomerat |
| 14 | ploaie | dimineață | aglomerat |
| 15 | zăpadă | dimineață | aglomerat |
| 16 | ploaie | amiază | ? |

Tabel de date obținute în urma unor observații

mea este ploioasă și este amiază, ce previziune se poate face despre trafic (aglomerat a sau relaxat r)?

Se face următoarea presupunere **naivă**: atributele sunt **condițional independente**, dacă se știe (cunoaște) clasificarea, adică

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

pentru fiecare $v \in \{p, z, s, \hat{i}\}$, $ti \in \{di, am, se, no\}$, $\mathbf{t} \in \{\mathbf{a}, \mathbf{r}\}$. 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 probabilitățile claselor și probabilitățile condiționate ale atributelor, cunoscând clasa.

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

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

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

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

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

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

și

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

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

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

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

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

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

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

★

Valoarea medie a unor variabile aleatoare discrete

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

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

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

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

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

- $E(aX + b) = aE(X) + b$ pentru orice $a, b \in \mathbb{R}$;
- $E(X + Y) = E(X) + E(Y)$;
- Dacă X și Y sunt v.a. independente, atunci $E(X \cdot Y) = E(X)E(Y)$.
- Dacă $g : \mathbb{R} \rightarrow \mathbb{R}$ e o funcție astfel încât $g(X)$ este v.a., atunci

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

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

► Python: `numpy.mean(x) = $\frac{1}{n}(x_0 + \dots + x_{n-1})$` pentru $x = [x_0, \dots, x_{n-1}]$

► Fie $x = [x_0, \dots, x_{n-1}]$ valori aleatoare ale unei v.a. X , atunci

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

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

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

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

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

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

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

Așadar jucătorul câștigă în medie 5 u.m.

```
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))
```

Exercițiu: Variabila aleatoare X descrie de câte ori apare pană de curent în rețea (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 comerț pe internet estimează că fiecare astfel de pană de curent în rețea duce la o pierdere de 200 Ron. Calculați valoarea medie a pierderilor zilnice ale acestei companii (datorate lipsei de curent). Estimați această valoare medie cu ajutorul unor simulări în Python.

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

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

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

Exemplu: Se aruncă patru zaruri. Fie X_i v.a. care indică numărul apărut la al i -lea zar.

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

Def. 18. Funcția de repartiție $F : \mathbb{R} \rightarrow [0, 1]$ a unei variabile aleatoare 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 cumulative distribution function, prescurtat cu cdf.

Exemplu: Funcția de repartiție $F_X : \mathbb{R} \rightarrow [0, 1]$ a v.a. discrete X este

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

Determinați valoarea medie a lui X .



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

- (1) $F(b) - F(a) = P(X \leq b) - P(X \leq a) = P(a < X \leq b) \forall a, b \in \mathbb{R}, a < b.$
- (2) F este monoton crescătoare, adică pentru orice $x_1 < x_2$ rezultă $F(x_1) \leq F(x_2)$.
- (3) F este continuă la dreapta, adică $\lim_{x \searrow x_0} F(x) = F(x_0) \forall x_0 \in \mathbb{R}$.
- (4) $\lim_{x \rightarrow \infty} F(x) = 1$ și $\lim_{x \rightarrow -\infty} F(x) = 0$.

Observație:

- ▷ Orice funcție $F : \mathbb{R} \rightarrow \mathbb{R}$, care are proprietățile (1), (2), (3) din **P.10** este o funcție de repartiție.
- ▷ Funcția de repartiție a unei v.a. descrie complet comportamentul probabilistic al acelei v.a.
- în Python: se calculează $F(x) = P(X \leq x)$ pentru $X \sim \text{Bin}(n, p)$ cu `scipy.stats.binom.cdf(x, n, p)`, iar pentru $X \sim \text{Huge}(n, n_1, n_2)$ cu `scipy.stats.hypergeom.cdf(x, n_1+n_2, n_1, n)`.

```
#Fie o urnă 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 și nenumărabil de valori într-un interval sau reuniune de intervale (v.a. poate lua orice valoare din intervalul considerat);

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

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

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

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

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

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

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

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

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

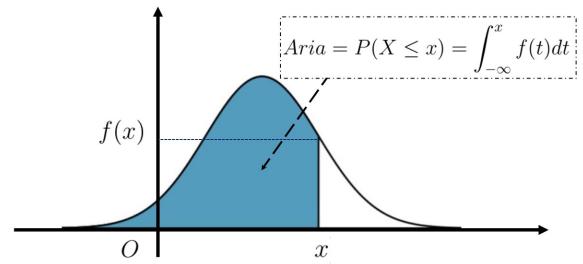
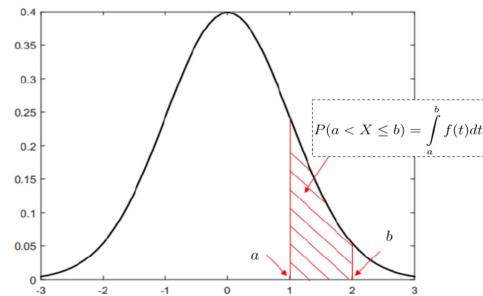
(2) $\int_{-\infty}^{\infty} f(t)dt = 1$;

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

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

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

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



- (6) $P(X \in M) = \int_M f(t)dt$, $M \subseteq \mathbb{R}$;
- (7) F este o funcție monoton crescătoare și continuă pe \mathbb{R} ;
- (8) $\lim_{x \rightarrow \infty} F(x) = 1$ și $\lim_{x \rightarrow -\infty} F(x) = 0$.
- (9) dacă F este derivabilă în punctul x , atunci $F'(x) = f(x)$.

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

(2) Fie f_1 o funcție de densitate pentru v.a. X și fie $f_2 : \mathbb{R} \rightarrow \mathbb{R}$ astfel încât $f_1(t) = f_2(t)$ pentru orice $t \in \mathbb{R} \setminus \mathcal{N}$, unde \mathcal{N} este o mulțime cel mult numărabilă. Atunci f_2 este o funcție de densitate pentru aceeași v.a. X . Este suficient să cunoaștem o funcție de densitate în orice punct din \mathbb{R} exceptând, eventual, o mulțime cel mult numărabilă de puncte. Funcția de densitate asociată unei v.a. nu este unică (unicitate în sensul egalității în toate punctele din \mathbb{R}). Proprietățile integralelor implică

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

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

(3) *Funcția de fiabilitate* (sau funcția de supraviețuire) este folosită în studiul fiabilității unor a sisteme, și a duratei de funcționare până la prima defectiune

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

Exemple de distribuții clasice continue

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

- funcția de densitate este

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

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

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

- funcția 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 numește *distribuția standard normală*.
- Distribuția normală se aplică în: măsurarea erorilor (de ex. termenul eroare în analiza regresională), în statistică (teorema limită centrală, teste statistice) etc.

```
# 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()
```

► **Distribuția exponențială:** $X \sim Exp(\lambda)$, $\lambda > 0$

- funcția de densitate este

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

```
# 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()

```

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

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

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

unde funcția Gamma este

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

```

#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()

```

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

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

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

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

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

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

♣

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

Vector aleator

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

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

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

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

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

F_X și F_Y se numesc **funcții de repartiție marginale**.

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

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

ceea ce este echivalent cu

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

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

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

Sunt X și Y v.a. independente? Determinați f_X , respectiv f_Y . ♣