

**Curs - Probabilități și Statistică 2022/2023**  
**Secția Informatică**

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



# Teoria Probabilităților

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

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

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

Fenomene și procese aleatoare apar, de exemplu, în:

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



[Sursa: [www.financialmarket.ro](http://www.financialmarket.ro)]

→ **în informatică:**

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

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

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

**Exemplu:** Generarea de valori aleatoare (în Octave)

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

```

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

```

### Algoritmi aleatori

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

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

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

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

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

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

Exemplu:

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

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

```

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

```

**Răspuns:** Algoritm de tip Las Vegas.

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

```

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

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

```

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

### Noțiuni introductive:

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

### Exemple:

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

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

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

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

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



### Operații cu evenimente

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

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

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

### Relații între evenimente

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

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

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

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

#### asociative

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

#### și distributive

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

satisfac **legile lui De Morgan**

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

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

### Frecvența relativă și frecvența absolută

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

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

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

### Definiția clasică a probabilității

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

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

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

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

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

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

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

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

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



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

Cu toate acestea, cavalerul de Méré a observat că jucând în modul al doilea (cu două zaruri aruncate de 24 ori), pierdea față de adversarul său, dacă acesta alegea primul mod (aruncarea unui singur zar de 4 ori). Pascal și Fermat au arătat că probabilitatea de câștig la jocul cu un singur zar aruncat de 4 ori este  $p_1 \approx 0.5177$ , iar probabilitatea  $p_2 \approx 0.4914$  la jocul cu două zaruri aruncate de 24 de ori. Deși diferența dintre cele două probabilități este mică, totuși, la un număr mare de partide, jucătorul cu probabilitatea de câștig  $p_1$  câștigă în fața jucătorului cu probabilitatea de câștig  $p_2$ . Practica jocului confirmă astfel justetea raționamentului matematic, contrar credinței lui de Méré.



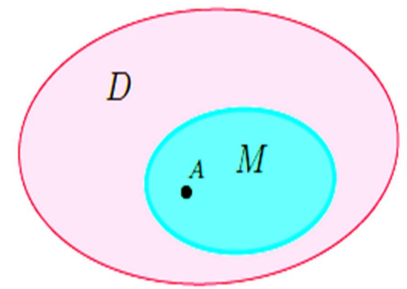
### Definiția axiomatică a probabilității

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

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

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

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



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

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

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

$\hookrightarrow \mathcal{K}$  este o mulțime de evenimente și are structura unei  $\sigma$ -algebre (vezi Def. 4)

$\hookrightarrow P$  satisface anumite axiome (vezi Def. 5)

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

(1)  $\mathcal{K}$  este nevidă;

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

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

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

Fie  $M = \{1, 2, 3, \dots\}$  și fie

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



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

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

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

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

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

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

este o  $\sigma$ -algebră pe mulțimea  $B$ . ◇

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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



## Evenimente independente

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

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

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

(1)  $A$  și  $B$  sunt independente.

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

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

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

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

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

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

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

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

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

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

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

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

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

Sunt cele 3 evenimente *independente în totalitate*?

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

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

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

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

### Exemplu:

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

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

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

```
clear all
clc
M=input('M='); % numar maxim de iteratii; M >= 1
% de ex. M=3
U=[0 0 0 0 0 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3];
% 0,1,2,3 apar fiecare cu probabilitatea 5/20=1/4
S=U(randperm(length(U)))
% S este o permutare aleatoare a vectorului initial U

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

Se calculează probabilitățile (teoretice) ale următoarelor evenimente:

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

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

probabilitatea evenimentului complementar este

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



### Probabilitate condiționată

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

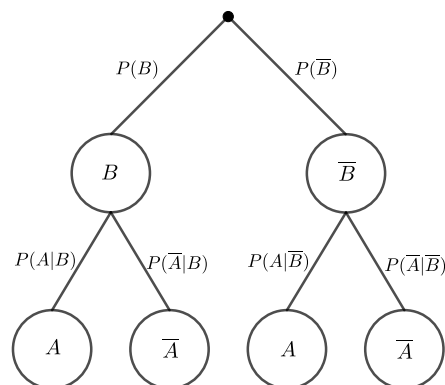
a)  $P(A_2|R_1) = \frac{4}{8}$ . b)  $P(R_1 \cap R_2) = P(R_2|R_1)P(R_1) = \frac{4}{8} \cdot \frac{5}{9}$ .



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

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

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



**Fig.1. Probabilități condiționate**

**Def. 9.** O familie  $\{H_1, \dots, H_n\} \subset \mathcal{K}$  de evenimente din  $\Omega$  se numește **partiție** sau **sistem complet de evenimente** a lui  $\Omega$ , 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  $\Omega$ .



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

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

**Exemplu:** Într-o urnă sunt 7 bile albe, notate cu 1, 2, 3, 4, 5, 6, 7, și 6 bile roșii notate cu 8, 9, 10, 11, 12, 13. Se extrage o bilă. a) Știind că bila extrasă este roșie, care este probabilitatea  $p_1$ , ca numărul de pe bilă să fie divizibil cu 4? b) Știind că prima bilă este roșie, care este probabilitatea

$p_2$ , ca o a doua bilă extrasă să indice un număr impar? (Prima bilă nu s-a returnat în urnă!)

R.: Se consideră evenimentele:

$A_1$ : prima bilă extrasă are înscris un număr divizibil cu 4;

$B_1$ : prima bilă extrasă este roșie;

$C_1$ : prima bilă extrasă are înscris un număr impar;

$C_2$ : a doua bilă extrasă are înscris un număr impar.

a)  $p_1 = P(A_1|B_1) = \frac{2}{6}$ .

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

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



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

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

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

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

►  $p_1$  estimează probabilitatea condiționată ca numărul ales aleator să fie divizibil cu 3, știind că

s-a extras un număr impar;

► p2 estimează probabilitatea condiționată ca numărul ales aleator să provină din mulțimea  $\{6, 7, 8, 9, 10\}$ , știind că s-a extras un număr par;

Algoritmul este de tip Monte-Carlo!

Care sunt valorile teoretice pentru probabilitățile p1, p2 din acest exemplu?



### P. 6. (Formula înmulțirii probabilităților)

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

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

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

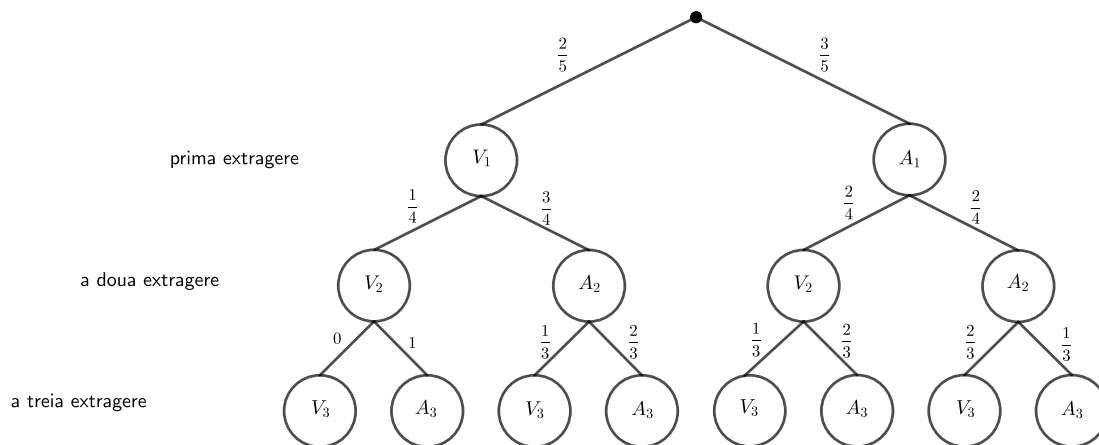
b)  $P((V_1 \cap V_2) \cup (A_1 \cap A_2)) = P(V_1 \cap V_2) + P(A_1 \cap A_2) = P(V_2|V_1)P(V_1) + P(A_2|A_1)P(A_1) = \frac{1}{4} \cdot \frac{2}{5} + \frac{2}{4} \cdot \frac{3}{5}$

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





**Fig. 3. Extragere fără returnare**

e) formula de înmulțire a probabilităților P.6:

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



## Formula lui Bayes

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

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

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

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

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

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

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

Formula lui Bayes este în acest caz

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

## P. 7. (Formula lui Bayes)

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

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

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

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

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

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

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

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

$H$ : o persoană aleasă aleator dintr-o populație are o anumită alergie  $\mathcal{A}$

$E$ : testul clinic returnează pozitiv privind alergia  $\mathcal{A}$

$\bar{E}$ : testul clinic returnează negativ privind alergia  $\mathcal{A}$

▷ din statistici anterioare sunt cunoscute:

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

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

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

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

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

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

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

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

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

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

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

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

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

Matricea de confuzie (engl. confusion matrix)

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

Matricea de confuzie construită cu datele statistice

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

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

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

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

<sup>1</sup> numărul de elemente din  $H \cap E$



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

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

### Variable aleatoare

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

**Exemplu:** 1) La aruncarea a două zaruri, suma numerelor obținute este o variabilă aleatoare  $S : \Omega \rightarrow \{2, 3, \dots, 12\}$ , unde  $\Omega$  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  $(\omega_i^1, \omega_j^2)$  este evenimentul elementar: la primul zar s-a obținut numărul  $i$  și la al doilea zar s-a obținut numărul  $j$ , unde  $i, j = \overline{1, 6}$ .

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

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

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

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

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

O variabilă aleatoare este:

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

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

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

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

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

### Variabile aleatoare numerice - definiție formală

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

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

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

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

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

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

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

### Distribuții discrete clasice

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

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

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

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

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

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

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

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

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



generare în Matlab/Octave:

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

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

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

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

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

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

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

$\rightarrow$  are loc **formula binomială**

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

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

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

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

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

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

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

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

$A$ : cel mult  $m$  erori apar în transmiterea celor  $N$  biți

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



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

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

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



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

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

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

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

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

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

Matlab/Octave: `hygernd( $n_1 + n_2, n_1, n, \dots$ )` generează valori aleatoare;

`hygepdf( $x, n_1 + n_2, n_1, n$ )` calculează  $P(X = x)$ , dacă  $X \sim Hyge(n, n_1, n_2)$ .

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

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

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

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

Scriem

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

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

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

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

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

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





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

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

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

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

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

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



## Variabile aleatoare independente

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

următorul tabel: 

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

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

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

b) Sunt  $X$  și  $Y$  v.a. independente?

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

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

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

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

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

### Observații:

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

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

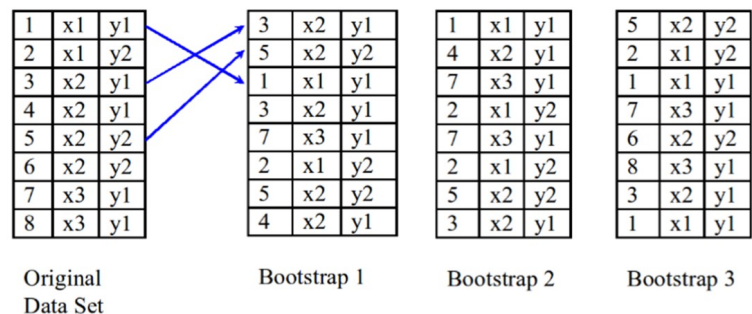
▷  $(X_1, \dots, X_r)$  este un vector aleator discret.

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

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

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

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



din urnă,  $i = \overline{1, r}$ ;

$$p(k_1, \dots, k_r; n) = \begin{array}{l} \text{probabilitatea de a obține } k_i \text{ bile cu culoarea } i, i = \overline{1, r}, \\ \text{din } n = k_1 + \dots + k_r \text{ extrageri fără returnarea bilei extrase,} \\ \text{în care ordinea de extragere a bilelor de diverse culori nu contează} \end{array}$$

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

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

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

### Operații cu variabile aleatoare (numerice)

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

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

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

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

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

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

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

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


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


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

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

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

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

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

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

### Clasificarea naivă Bayes

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

Fie  $(\Omega, \mathcal{K}, P)$  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:

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

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

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

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

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

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

### Exemplu de clasificare naivă Bayes

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

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

Considerăm evenimentul următor, denumit *vector de attribute*:

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

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

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

Tabel de date obținute în urma unor observații

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

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

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

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

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

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

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

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

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

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

și

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

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

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

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

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

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

★

## Valoarea medie a unor variabile aleatoare discrete

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

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

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

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

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

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

→  $E(X + Y) = E(X) + E(Y)$ ;

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

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

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

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

Matlab/Octave: `mean(x)`

pentru  $x = [x(1), \dots, x(n)]$ , se calculează  $\text{mean}(x) = \frac{1}{n}(x(1) + \dots + x(n))$

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

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

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

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

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

Așadar jucătorul câștigă în medie 5 u.m.



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

### Exercițiu:

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

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

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

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

```
end
fprintf('nr. mediu de iteratii: %4.3f \n', mean(s))
```

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

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

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

Numărul mediu de iterații necesare *înainte* să apară primul 0 este  $E(X)$ . Se poate arăta că  $E(X) = \frac{1-p}{p} = \frac{1-0.25}{0.25} = 3$ . ▼

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## Variabile aleatoare continue

V.a. continuă: ia un număr infinit și nenumărabil de valori într-un interval sau reuniune de intervale (v.a. poate lua orice valoare din intervalul considerat);

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

▷ ea este caracterizată de funcția de densitate.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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