Teoria Probabilităţilor

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

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

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

Fenomene şi procese aleatoare apar, de exemplu, în:
→ jocuri de noroc, pariuri, loto (6 din 49)
→ previziuni meteo
→ previziuni economice / financiare (evaluarea stocurilor)
→ sondaje de opinie, asigurări (evaluarea riscurilor, pierderilor)
→ **în informatică**: sisteme de comunicare, prelucrarea informaţiei, modelarea traficului în reţea, analiza probabilistică a unor algoritmi, fiabilitatea sistemelor, algoritmi de simulare, machine learning, data mining, recunoaşterea formelor / a vocii, generarea de numere aleatoare, algoritmi aleatori: de tip Monte-Carlo, de tip Las Vegas etc.
Exemple de întrebări:
→ Cum este concepută memoria cache pentru a maximiza viteza RAM a calculatoarelor?
→ Reţelele de calculatoare: Care este probabilitatea ca un pachet de date să fie recepţionat corect, atunci când canalul de transmisie este perturbat? În medie câte date sunt transmise corect?
→ Structuri de date: Se caută un tabel hash eficient cu cea mai rapidă căutare (lookup)

Algoritmi aleatori

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

▷ durata de execuţie, spaţiul de stocare, rezultatul obţinut sunt variabile aleatoare (chiar dacă se folosesc aceleaşi valori input)
▷ la anumite tipuri de algoritmi corectitudinea e garantată doar cu o anumită probabilitate
▷ în mod paradoxal, incertitudinea ne poate oferi mai multă eficienţă
Exemplu: Random QuickSort, în care elementul pivot este selectat aleator
• Algoritm de tip **Las Vegas** este un algoritm aleator, care returnează la fiecare execuţie rezultatul corect (independent de alegerile aleatoare făcute); durata de execuţie este o variabilă aleatoare.
Exemplu: Random QuickSort
• Un algoritm aleatoriu pentru care rezultatele obţinute sunt corecte *doar* cu o anumită probabilitate se numeşte algoritm **Monte Carlo**.

$\hookrightarrow$ se examinează probabilitatea cu care rezultatul este corect; probabilitatea de eroare poate fi scăzută semnificativ prin execuţii repetate, independente;
Exemplu:
1) testul Miller-Rabin, care verifică dacă un număr natural este prim sau este număr compus; rezultatul testului returnează fie răspunsul: "numărul este sigur un număr compus" sau răspunsul: "numărul este probabil un număr prim" (acest test nu returnează valorile divizorilor numărului compus);
2) problema tăieturii minime într-un graf (algoritmul lui D. Karger: random min-cut)
$\longrightarrow$ De care tip este următorul algoritm (scris în pseudocod)?
Input: Fie A(1),...,A(200) un vector cu 200 de elemente, din care 100 sunt egale cu 0 şi restul egale cu 1 (ordinea lor este necunoscută).
Output: Să se găsească un 0 în vector.

```
algorithm(array A, size n)
begin
repeat
randomly select one element from A
until 0 is found
end
```

Răspuns: Algoritm de tip Las Vegas.
Versiunea Monte Carlo a problemei formulate anterior: se dă k numărul maxim de iteraţii

```
find_MC(array A, size n, k)
begin
    i=0
    repeat
        randomly select one element from A
        i = i + 1
    until i=k or 0 is found
end
```

$\triangleright$ dacă 0 este găsit, atunci algoritmul se încheie cu rezultatul corect, altfel algoritmul nu găseşte niciun 0; probabilitatea de a găsi pe 0 după k iteraţii este

$$P(\text{"0 este găsit după } k \text{ iteraţii"}) = 1 - (1/2)^k; \quad 1 - (1/2)^k \longrightarrow 1, \text{ când } k \to \infty.$$

Noţiuni introductive:
• Experienţa aleatoare este acea experienţă al cărei rezultat nu poate fi cunoscut decât după încheierea ei.
• Evenimentul este rezultatul unui experiment.
Exemple:

▷ Experiment: aruncarea a două zaruri, eveniment: ambele zaruri indică 1
▷ experiment: aruncarea unei monede, eveniment: moneda indică pajură
▷ experiment: extragerea unei cărţi de joc, eveniment: s-a extras as
▷ experiment: extragerea unui număr la loto, eveniment: s-a extras numărul 27
• evenimentul imposibil, notat cu $\emptyset$, este evenimentul care nu se realizează niciodată la efectuarea experienţei aleatoare
• evenimentul sigur este un eveniment care se realizează cu certitudine la fiecare efectuare a experienţei aleatoare
• spaţiul de selecţie, notat cu Ω, este mulţimea tuturor rezultatelor posibile ale experimentului considerat
◇ spaţiul de selecţie poate fi finit sau infinit
• dacă A este o submulţime a lui Ω atunci A se numeşte eveniment aleator, iar dacă A are un singur element atunci A este un eveniment elementar.
▷ O *analogie între evenimente şi mulţimi* permite o scriere şi în general o exprimare mai comode ale unor idei şi rezultate legate de conceptul de eveniment aleator.
Exemplu: Experimentul: aruncarea unui zar, spaţiul de selecţie: $\Omega = \{e_1, e_2, e_3, e_4, e_5, e_6\}$, e_i: s-a obţinut numărul i $(i = 1, \ldots, 6)$; $e_1, e_2, e_3, e_4, e_5, e_6$ sunt evenimente elementare
A: s-a obţinut un număr par $\Rightarrow A = \{e_2, e_4, e_6\}$
$\bar{A}$: s-a obţinut un număr impar $\Rightarrow \bar{A} = \{e_1, e_3, e_5\}$ ♣

Operaţii cu evenimente

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

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

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

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

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

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

Relaţii între evenimente

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

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

Proprietăţi ale operaţiilor între evenimente $A, B, C \subseteq \Omega$
Operaţiile de reuniune şi intersecţie sunt operaţii **comutative**:

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

asociative

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

şi **distributive**

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

satisfac legile lui De Morgan

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

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

Frecvenţa relativă şi frecvenţa absolută

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

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

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

Definiţia clasică a probabilităţii

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

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

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

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

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

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

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

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

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

Definiţia axiomatică a probabilităţii

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

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

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

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

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

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

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

(i) $\mathcal{K}$ *este nevidă;*

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

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

Perechea $(\Omega, \mathcal{K})$ se numeşte ***spaţiu măsurabil****.*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

♡

Probabilitate condiţionată

Def. 6. *Fie* $(\Omega, \mathcal{K}, P)$ *un spaţiu de probabilitate şi fie* $A, B \in \mathcal{K}$. ***Probabilitatea condiţionată a evenimentului*** A ***de evenimentul*** B *este* $P(\cdot|B) : \mathcal{K} \to [0,1]$ *definită prin*

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

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

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

P. 3. *Pentru* $A, B \in \mathcal{K}, P(A) > 0, P(B) > 0$ *au loc:*

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

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

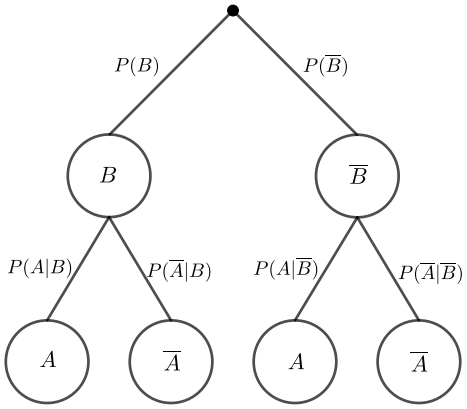
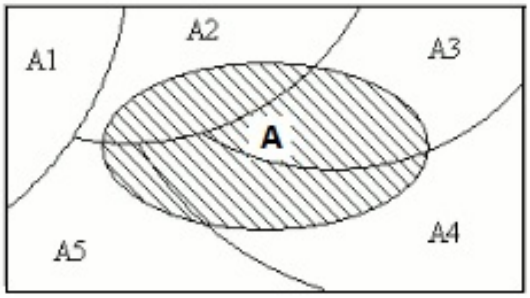


Fig.1. Probabilităţi condiţionate

Def. 7. *O familie* $A_1, \ldots A_n$ *de evenimente din* Ω *se numeşte* ***partiţie*** *sau* ***sistem complet de evenimente*** *a lui* Ω*, dacă* $\bigcup_{i=1}^{n} A_i = \Omega$ *şi pentru fiecare* $i, j \in \{1, \ldots, n\}$, $i \neq j$, *evenimentele* A_i *şi* A_j *sunt disjuncte, adică* $A_i \cap A_j = \emptyset$.



Partiţie $A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5 = \Omega$

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

P. 4. ***(Formula probabilităţii totale)*** *Într-un spaţiu de probabilitate* $(\Omega, \mathcal{K}, P)$ *considerăm partiţia* $H_1, ..., H_n$ *a lui* Ω *cu* $P(H_i) > 0$ *şi* $H_i \in \mathcal{K}\ \forall\ i \in \{1, \dots, n\}$*, şi fie* $A \in \mathcal{K}$*. Atunci are loc*

$$P(A) = \sum_{i=1}^{n} P(A|H_i)P(H_i).$$

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

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

♡

Formula lui Bayes

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

$$P(D) = P(D|I) \cdot P(I) + P(D|\bar{I}) \cdot P(\bar{I}) = P(D|I) \cdot P(I) + P(D|\bar{I}) \cdot (1 - P(I)).$$

Formula lui Bayes este în acest caz

$$P(I|D) = \frac{P(D|I) \cdot P(I)}{P(D)} = \frac{P(D|I) \cdot P(I)}{P(D|I) \cdot P(I) + P(D|\bar{I}) \cdot P(\bar{I})}.$$

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

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

♣

P. 5. (Formula lui Bayes) *Într-un spaţiu de probabilitate* $(\Omega, \mathcal{K}, P)$ *considerăm partiţia* $H_1, \ldots, H_n$ *a lui* Ω *cu* $P(H_i) > 0$ *şi* $H_i \in \mathcal{K}\ \forall\ i \in \{1, ...n\}$*, şi fie* $E \in \mathcal{K}$ *astfel încât* $P(E) > 0$*. Atunci,*

$$P(H_j|E) = \frac{P(H_j)P(E|H_j)}{P(E)} = \frac{P(H_j)P(E|H_j)}{\sum_{i=1}^{n} P(H_i)P(E|H_i)} \quad \forall\ j \in \{1, 2, ..., n\}.$$

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

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

```
clear all
ci=0;
cp=0;
c=0;
a=0;
b=0;
N=1000;
for i=1:N
```

```
 A=[randi(5,1,5),5+randi(8,1,5),13+randi(7,1,10)];
 r= randi(length(A));
 v=A(r);
 ci=ci+mod(v,2);
 cp=cp+(mod(v,2)==0);
  c=c+ mod(v,2)*(mod(v,3)==0);
  a=a+ mod(v,2)*(6<=r && r<=10);
  b=b+ (mod(v,2)==0)*(r>=10);
end
p1=c/ci
p2=a/ci
p3=b/cp
```

R.: Se dă un şir A format din 20 de elemente, în care
• 25% provin din generarea aleatoare şi cu aceeaşi probabilitate (care e 1/5) a unui număr din $\{1, 2, 3, 4, 5\}$
• 25% provin din generarea aleatoare şi cu aceeaşi probabilitate (care e 1/8) a unui număr din $\{6, 7, 8, 9, 10, 11, 12, 13\}$
• 50% provin din generarea aleatoare şi cu aceeaşi probabilitate (care e 1/7) a unui număr din $\{14, 15, 16, 17, 18, 19, 20\}$.
Se extrage aleator un număr din şir.
►`p1` estimează probabilitatea condiţionată ca numărul ales aleator să fie divizibil cu 3, *ştiind* că s-a extras un număr impar;
► `p2` estimează probabilitatea condiţionată ca numărul ales aleator să provină din mulţimea $\{6, 7, 8, 9, 10, 11, 12, 13\}$, *ştiind* că s-a extras un număr impar;
► `p3` estimează probabilitatea condiţionată ca numărul ales aleator să provină din mulţimea $\{14, 15, 16, 17, 18, 19, 20\}$, *ştiind* că s-a extras un număr par.
Algoritmul este de tip Monte-Carlo! ♣

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

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

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

Exemplu: Într-o urnă sunt 2 bile verzi şi 3 bile albastre. Se extrag 2 bile succesiv, fără returnare. Care este probabilitatea ca
a) prima bilă să fie verde, iar cea de-a doua albastră?
b) cele 2 bile să aibă aceeaşi culoare?
c) a doua bilă să fie albastră?

d) prima bilă să fie verde, *ştiind* că a doua este albastră?
e) se mai extrage o a treia bilă; se cere probabilitatea ca prima bilă să fie verde, cea de-a doua albastră şi a treia tot albastră.
R.: Notăm pentru $i \in \{1,2,3\}$ evenimentele:
A_i: la a i-a extragere s-a obţinut bilă albastră; V_i: la a i-a extragere s-a obţinut bilă verde;
a) folosim P.3: $P(V_1 \cap A_2) = P(A_2|V_1)P(V_1) = \frac{3}{4} \cdot \frac{2}{5}$

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

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

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

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

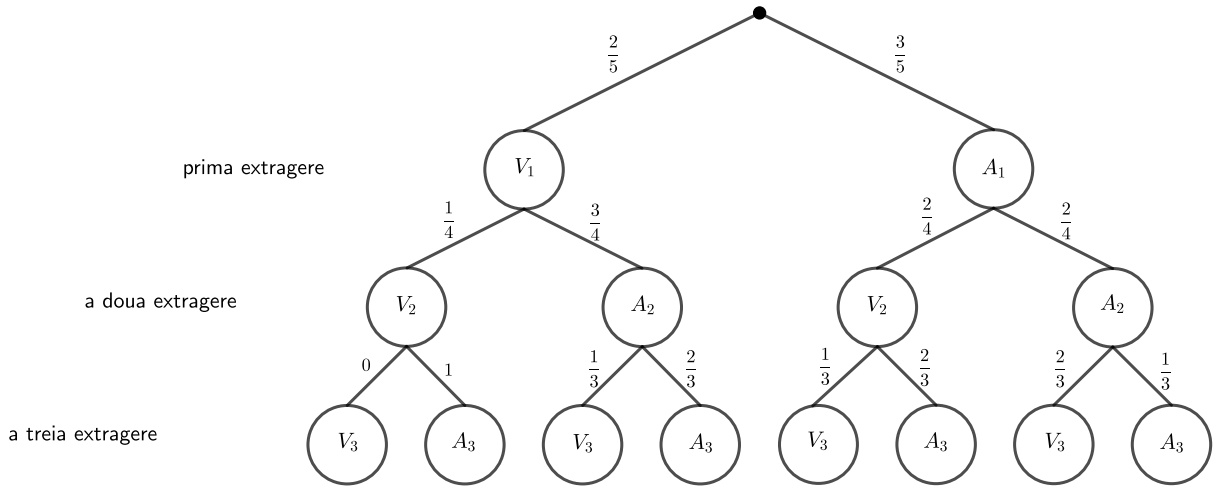


Fig. 3. Extragere fără returnare

♣

Evenimente independente

Def. 8. *Fie* $(\Omega, \mathcal{K}, P)$ *un spaţiu de probabilitate. Evenimentele* $A, B \in \mathcal{K}$ *sunt* ***evenimente independente*** *dacă*

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

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

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

Două evenimente se numesc **dependente** dacă probabilitatea realizării unuia dintre ele depinde de faptul că celălalt eveniment s-a produs sau nu.
Exerciţiu: 1) Se extrag succesiv fără returnare două cărţi de joc dintr-un pachet de cărţi;
A: prima carte extrasă este as;
B: prima carte extrasă este damă
C: a doua carte extrasă este 1;
Sunt A şi B evenimente independente?
Sunt A şi C evenimente independente?
Sunt B şi C evenimente dependente?
2) Se aruncă un zar de două ori.
A: primul număr este 6;
B: al doilea număr este 5
C: primul număr este 1;
Sunt A şi B evenimente independente?
Sunt A şi C evenimente independente?
Sunt B şi C evenimente dependente? ♣

P. 7. *Fie $(\Omega, \mathcal{K}, P)$ un spaţiu de probabilitate şi fie $A, B \in \mathcal{K}$. Sunt echivalente afirmaţiile:*

(1) *A şi B sunt independente.*

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

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

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

Def. 9. *Fie $(\Omega, \mathcal{K}, P)$ un spaţiu de probabilitate. $B_1, \ldots, B_n$ sunt n* ***evenimente independente*** *(în totalitate) din $\mathcal{K}$ dacă*

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

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

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

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

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

2) Cele 4 feţe ale unui tetraedru regulat sunt vopsite astfel: una este roşie, una este albastră, una este verde şi una este colorată având cele trei culori. Se aruncă tetraedrul şi se consideră evenimentele: R: tetraedrul cade pe partea roşie; A: tetraedrul cade pe partea albastră; V: tetraedrul cade pe partea verde. Se poate afirma că $P(R \cap A \cap V) = P(R)P(A)P(V)$? Sunt

cele 3 evenimente independente în totalitate?
3) Pentru a verifica dacă n evenimente distincte $B_1, \dots, B_n$ sunt independente (în totalitate) câte relaţii trebuie verificate? R.: $C_n^2 + C_n^3 + ... + C_n^n = 2^n - n - 1$ ♦

Variable aleatoare

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

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

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

Notaţie 1. *variabilă/variabile aleatoare → 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 şi sunt într-un interval (sau reunine de intervale) sau în $\mathbb{R}$

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

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

Variabile aleatoare numerice - definiţie formală

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

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

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

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

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

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

▷ Variabilele aleatoare discrete sunt caracterizate de distribuţia de probabilitate.
▷ Notăm $\{X = x_i\} = \{\omega \in \Omega : X(\omega) = x_i\}$; $\{X = x_i\}$ este un eveniment din $\mathcal{K}$ pentru fiecare $i \in I$.

Distribuţii discrete clasice

Distribuţia discretă uniformă: $X \sim Unif(n)$

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

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`$(n,...)$

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

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

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

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

♣

generare în Matlab/Octave:

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

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

• se repetă experimentul de n ori
• v.a. X= numărul de succese în n repetări independente ale experimentului $\Rightarrow$ valori posibile: $X \in \{0, 1, \ldots, n\}$

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

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

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

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

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

Matlab/Octave: `binornd`$(n, p, ...)$

▷ Distribuţia binomială corespunde modelului cu extragerea bilelor dintr-o urnă cu returnarea bilelor după fiecare extragere.
Exemplu: Într-o urnă sunt n_1 bile albe şi n_2 bile negre. Se extrag cu returnare n bile; fie v.a. X = numărul de bile albe extrase $\Rightarrow X \sim Bino(n, p)$ cu $p = \frac{n_1}{n_1+n_2}$.
2) Fie un canal de comunicare binară care transmite cuvinte codificate de N biţi fiecare. Probabilitatea transmiterii cu succes a unui singur bit este p, iar probabilitatea unei erori este $1 - p$. Presupunem, de asemenea, că un astfel de cod este capabil să corecteze până la m erori, unde $0 \leq m \leq N$. Se ştie că transmiterea biţilor succesivi este independentă, atunci probabilitatea transmiterii cu succes a cuvântului este $p = P(A)$, unde
A: cel mult m erori apar în transmiterea celor N biţi

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

♠

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

Distribuţia hipergeometrică: $X \sim Hyge(n, n_1, n_2)$
Într-o urnă sunt n_1 bile albe şi n_2 bile negre. Se extrag **fără returnare** n bile.
Fie v.a. X = numărul de bile albe extrase $\Rightarrow$ valori posibile pentru X sunt $\{0, 1, \ldots, n^*\}$ cu

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

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

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

Matlab/Octave: `hygernd`$(n_1 + n_2, n_1, n, ...)$

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

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

♢

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

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

Matlab/Octave: `geornd`$(p, ...)$

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

Variabile aleatoare independente

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

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

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

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

Def. 13. *$\mathbb{X} = (X_1, \ldots, X_m)$ este un **vector aleator discret** dacă fiecare componentă a sa este o variabiă aleatoare discretă.*
Dacă $\mathbb{X}$ este un vector aleator discret care ia valori în mulţimea $\mathbb{X}(\Omega) = \{\mathbb{x}_k : k \in K\} \subset \mathbb{R}^m$, unde $K \subseteq \mathbb{N}$ este o mulţime de indici, atunci

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

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

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

▷ Vectorii aleatori sunt caracterizaţi de distribuţiile lor! De exemplu, un vector aleator cu 2 componente:

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

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

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

X \ Y	...	y_j	...
⋮	⋮	⋮	⋮
x_i	...	p_{ij}	...
⋮	⋮	⋮	⋮

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

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

▷ Dacă X şi Y sunt v.a. independente, şi se ştiu distribuţiile lor, atunci distribuţia vectorului aleator (X, Y) se determină pe baza formulei (1).
▷ Dacă se cunoaşte distribuţia vectorului aleator (X, Y) distribuţiile lui X şi Y se determină astfel:

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

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

Operaţii cu variabile aleatoare (numerice)

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

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

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

Determinaţi:

a) distribuţiile variabilelor aleatoare X_1 şi X_2;

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

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

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

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

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

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

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

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

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

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

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

Def. 14. ***Valoarea medie a unei variabilă aleatoare discrete*** *(numerice) X, care ia valorile $\{x_i, i \in I\}$, este*

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

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

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

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

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

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

Matlab/Octave: `mean(x)`
pentru $x = [x(1), ..., x(n)]$, se calculează $\texttt{mean}(x) = \dfrac{1}{n}(x(1) + ... + x(n))$

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

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

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

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

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

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

```
c=c+1;
i=randi(length(A));
end
fprintf('nr. iteratii: %d \n',c)

clc
clear all
A=[zeros(1,50), zeros(1,70)+1,zeros(1,80)+2];
s=[];
N=100;
for j=1:N
index=randperm(length(A));
A=A(index);
c=0;
i=randi(length(A));
A(i)
while A(i)~=0
c=c+1;
i=randi(length(A));
end
s=[s,c];
end
mean(s)
fprintf('nr. mediu de iteratii: %4.3f \n',mean(s))
```

Probabilitatea să apară la orice iteraţie 0 este $p = \frac{50}{200} = 0.25$.
Notăm cu X v.a. care indică numărul de iteraţii necesare *înainte* să apară primul $0 \Rightarrow X \sim Geo(p)$.
Numărul mediu de iteraţii necesare *înainte* să apară primul 0 este $E(X)$. Să se calculeze această valoare medie! ▼

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

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

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

Notaţie 3. Fie $\mathbb{U}$ un vector aleator discret care ia valori în $\mathcal{U} \subset \mathbb{R}^m$ şi fie $\mathbb{V}$ un vector aleator discret care ia valori în $\mathcal{V} \subset \mathbb{R}^n$, $m, n \in \mathbb{N}^*$. Notăm $P[\mathbb{U}] : \mathcal{U} \to [0, 1]$,

$$P[\mathbb{U}](\mathbb{u}) = P(\mathbb{U} = \mathbb{u}) \; \forall \, \mathbb{u} \in \mathcal{U}$$

$\hookrightarrow P[\mathbb{U}]$ este **distribuţia de probabilitate** a vectorului aleator, dacă $m > 1$ (a se vedea Def. 13), sau a v.a., dacă $m = 1$ (a se vedea Def. 11).

► **Observaţie: (1)** Def. 15 se transcrie mai compact astfel:
$X_1, \ldots, X_n$ sunt **variabile aleatoare independente**

$$\Longleftrightarrow P[X_1, \ldots, X_n] = P[X_1] \cdot \ldots \cdot P[X_n].$$

(2) Dacă U_1 and U_2 sunt 2 v.a. discrete atunci:
$\triangleright$ U_1 au U_2 aceeaşi distribuţie $\Longleftrightarrow P[U_1] = P[U_2]$.
$\triangleright$ U_1 şi U_2 sunt v.a. independente $\Longleftrightarrow P[U_1, U_2] = P[U_1]P[U_2] \Longleftrightarrow P[U_1|U_2] = P[U_1] \Longleftrightarrow P[U_2|U_1] = P[U_2]$.

P. 10. *Dacă $X_1, \ldots, X_n$ sunt variabile aleatoare independente, atunci pentru orice indici diferiţi $i_1, ..., i_k \subset \{1, ..., n\}$ $X_{i_1}, \ldots, X_{i_k}$ sunt variabile aleatoare independente, adică*

$$P[X_{i_1}, \ldots, X_{i_k}] = P[X_{i_1}] \cdot \ldots \cdot P[X_{i_k}].$$

Clasificarea naivă Bayes

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

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

Def. 16. *Evenimentele $A, B \in \mathcal{K}$ sunt **condiţional independente**, cunoscând evenimentul $C \in \mathcal{K}$, dacă şi numai dacă*

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

P. 11. *Au loc echivalenţele:*

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

Demonstraţie: • Pentru prima echivalenţă: "⇒"

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

"$\Leftarrow$"

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

• $P(A\cap B|C)=P(A|C)P(B|C)\Leftrightarrow P(B|A\cap C)=P(B|C)$ se demonstrează analog. ■

Exemplu: Fie Z o v.a. care indică rezultatul aruncării unui zar. Considerăm evenimentele: $A=(Z\in\{1,2\})$, $B=(Z\in\{2,4,6\})$ şi $C=(Z\in\{1,4\})$. Să se arate că:
a) A şi B sunt independente;
b) A şi B nu sunt condiţional independente, cunoscând evenimentul C.
R.: a) $P(A\cap B)=P(Z=2)=\frac{1}{6}=\frac{1}{3}\cdot\frac{1}{2}=P(Z\in\{1,2\})P(Z\in\{2,4,6\})=P(A)P(B)\implies A$ şi B sunt independente.
b)$P(A\cap B|C)=P(Z=2|Z\in\{1,4\})=0$, $P(A|C)=P(Z=1|Z\in\{1,4\})=\frac{1}{2}=P(Z=4|Z\in\{1,4\})=P(B|C)\Rightarrow A$ şi B nu sunt condiţional independente, cunoscând C. ▲

Exerciţiu: Într-o cutie sunt 2 zaruri. La primul zar 3 apare cu probabilitatea $\frac{1}{6}$, iar la celălalt zar (care e măsluit) 3 apare cu probabilitatea $\frac{5}{6}$. Se alege aleator un zar, care este apoi aruncat de 2 ori. Considerăm evenimentele
A_i: "zarul ales indică 3 la aruncarea i", $i\in\{1,2\}$
Z_j:"se alege zarul j", $j\in\{1,2\}$.
Sunt A_1 şi A_2 condiţional independente, cunoscând Z_1? Sunt A_1 şi A_2 independente?
R.: Dacă se cunoaşte tipul zarului ales, atunci aruncările sunt în mod evident independente: $P(A_1\cap A_2|Z_1)=\frac{1}{36}=P(A_1|Z_1)\cdot P(A_2|Z_1)$. Din formula probabilităţii totale P.5 avem: $P(A_2)=P(A_1)=P(A_1|Z_1)P(Z_1)+P(A_1|Z_2)P(Z_2)=\frac{1}{6}\cdot\frac{1}{2}+\frac{5}{6}\cdot\frac{1}{2}=\frac{1}{2}$ şi $P(A_1\cap A_2)=P(A_1\cap A_2|Z_1)P(Z_1)+P(A_1\cap A_2|Z_2)P(Z_2)=\frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{2}+\frac{5}{6}\cdot\frac{5}{6}\cdot\frac{1}{2}=\frac{13}{36}$. Deci $P(A_2|A_1)=\frac{P(A_1\cap A_2)}{P(A_1)}=\frac{13}{18}$. A_1 şi A_2 nu sunt independente, pentru că $P(A_2|A_1)\neq P(A_2)$. ✾

Def. 17. *Fie X,Y,Z v.a. discrete care iau valori în mulţimile $\mathcal{X},\mathcal{Y},\mathcal{Z}$. V.a. X este* ***condiţional independentă*** *de Y, cunoscând (ştiind) v.a. Z, dacă pentru fiecare $x\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}$,are loc*

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

Notaţie 4. Notăm cu $P[\mathbb{U}|\mathbb{V}]:\mathcal{U}\times\mathcal{V}\to[0,1]$ distribuţia de probabilitate condiţionată

$$P[\mathbb{U}|\mathbb{V}](\mathbb{u},\mathbb{v})=P(\mathbb{U}=\mathbb{u}|\mathbb{V}=\mathbb{v})\;\forall\,\mathbb{u}\in\mathcal{U},\mathbb{v}\in\mathcal{V}.$$

► **Observaţie:** Def. 17 se transcrie astfel:

$$P[X,Y|Z]=P[X|Z]P[Y|Z]\,. \tag{2}$$

Folosind **P.11**, rezultă:

P. 12. *V.a. X şi Y sunt condiţional independente, cunoscând $Z \Leftrightarrow$* $P[X,Y|Z] = P[X|Z]P[Y|Z] \Leftrightarrow P[X|Y,Z] = P[X|Z] \Leftrightarrow P[Y|X,Z] = P[Y|Z]$.

Vom introduce noţiunea de condiţional independenţă pentru mai multe v.a. discrete.

Def. 18. *Fie $X, Y_1, \ldots, Y_m, Z_1, \ldots, Z_n$ v.a. discrete. V.a. X este* ***condiţional independentă*** *de $Y_1, \ldots, Y_m$, ştiind (cunoscând) v.a. $Z_1, \ldots, Z_n$, dacă are loc*

$$P[X, Y_1, \ldots, Y_m | Z_1, \ldots, Z_n] = P[X|Z_1, \ldots, Z_n] P[Y_1, \ldots, Y_m | Z_1, \ldots, Z_n].$$

P. 13. *Fie X o v.a. discretă condiţional independentă de v.a. discrete $Y_1, \ldots, Y_m$, cunoscând v.a. discrete $Z_1, \ldots, Z_n$. Dacă $i_1, \ldots, i_k \in \{1, \ldots, m\}$ sunt indici diferiţi, atunci are loc*

$$P[X|Y_{i_1}, \ldots, Y_{i_k}, Z_1, \ldots, Z_n] = P[X|Z_1, \ldots, Z_n],$$
$$P[Y_{i_1}, \ldots, Y_{i_k}|X, Z_1, \ldots, Z_n] = P[Y_{i_1}, \ldots, Y_{i_k}|Z_1, \ldots, Z_n].$$

Exemplu de clasificare naivă Bayes

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

• **vreme** V: ploaie p, zăpadă z, senin s, înnorat *î* (dar nu plouă şi nu ninge) ;

• **timp** Ti: dimineaţă di, amiază am, seară se, noapte no.

Considerăm următorul *tabel de date* obţinute în urma unor observaţii pe bulevard:

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

Considerăm evenimentul următor, denumit *vector de atribute*:

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

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

Se face următoarea presupunere *naivă*: atributele sunt **condiţional independente**, dacă se dă clasificarea. De asemenea, presupunemm că toate probabilităţile condiţionate sunt definite (adică condiţionarea este în raport cu un eveniment a cărui probabilitate nu este 0).

Presupunerea *naivă* în acest exemplu este: V şi Ti sunt condiţinal independente, cunoscând (ştiind) $\mathbf{T}$ (a se vedea Def. 17, respectiv relaţia (2)), adică

$$P[V,Ti|\mathbf{T}] = P[V|\mathbf{T}]P[Ti|\mathbf{T}]. \tag{3}$$

Această notaţie se transcrie prin

$$P(V=v, Ti=ti|\mathbf{T}=\mathbf{t}) = P(V=v|\mathbf{T}=\mathbf{t})P(Ti=ti|\mathbf{T}=\mathbf{t}), \tag{4}$$

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:

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

altfel spus: *probabilitatea să plouă, ştiind că traficul este aglomerat, nu depinde de timp (adică este independentă de timp).*

Din (3) şi P. 12 se deduce că$P[V|Ti,\mathbf{T}] = P[V|\mathbf{T}]$ şi $P[Ti|V,\mathbf{T}] = P[Ti|\mathbf{T}]$, adică

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

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

► Folosind datele din tabel, determinăm mai întâi probabilităţile claselor şi probabilităţile condiţionate ale atributelor, cunoscând clasa.

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

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

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

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

şi

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

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

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

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

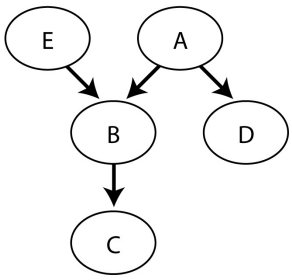
Această concluzie se poate completa la datele din tabelul de date iniţial!

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

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

şi deducem $P(E)=P(V=p,Ti=am)=\frac{11}{270}$. ✸

Reţele Bayes



Reţeaua Bayes este un graf orientat aciclic (i.e. nu conţine niciun drum orientat închis).

▷ Nodul Y este **părinte** pentru nodul X, dacă există o muchie orientată de la Y la X. Mulţimea părinţilor lui X se notează cu $p(X)$. Dacă X este nod rădăcină, atunci $p(X)=\emptyset$. De exemplu: $p(B)=\{E,A\}, p(D)=\{A\}, p(C)=\{B\}, p(E)=p(A)=\emptyset$.

▷ Nodul Y este **descendent** al nodului X, dacă există un drum orientat de la X la Y. Mulţimea descendenţilor lui X se notează cu $de(X)$. De exemplu: $d(E)=\{B,C\}, d(A)=\{B,C,D\}$, $d(B)=\{C\}, d(D)=\emptyset$.

Într-o reţea în care există o structură cauzală, nodurile din $p(X)$ reprezintă *cauzele* pentru X, iar nodurile din $de(X)$ sunt *efectele* nodului X.

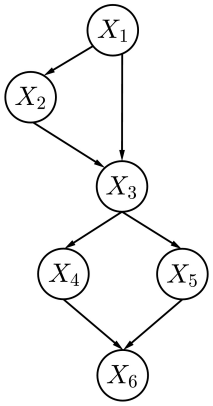
▷ Nodul Y este **nondescendent** al nodului X, dacă nu este descendent al nodului X. Mulţimea nondescendeţilor lui X se notează cu $nd(X)$. De exemplu: $nd(E)=\{A,D\}, nd(A)=\{E\}, nd(B)=\{E,A,D\}, nd(D)=\{E,A,B,C\}, nd(C)=\{E,A,B,D\}$.

▷ Fiecare nod $X_1,...,X_n$ din reţea este identificat cu o variabilă aleatoare şi este definit pe

acelaşi spaţiu de probabilitate $(\Omega, \mathcal{K}, P)$; probabilităţile $P[X_j|p(X_j)], j = \overline{1,n}$ sunt date; are loc convenţia $P[X_j|p(X_j)] = P[X_j]$, dacă X_j este nod rădăcină ($P[X_j]$ este distribuţia de probabilitate a lui X_j, a se vedea Notaţia 3, iar $P[X_j|p(X_j)]$ este distribuţia de probabilitate condiţionată, a se vedea Notaţia 4).

► Properietatea reţelei Bayes: orice nod X şi nondescendenţii săi $nd(X)$ sunt *condiţional independenţi*, dacă se cunosc valorile părinţilor $p(X)$; dacă $p(X) = \emptyset$, atunci X şi $nd(X)$ sunt independenţi.

Exemplul 1: Se dă reţeaua Bayes din figura alăturată, în care $X_1, ..., X_6$ sunt variabile aleatoare binare.



Reţea Bayes

▷ Au loc proprietăţile:

• Mulţimile de noduri corespunzătoare părinţilor, descendenţilor, nondescendenţilor sunt:

$p(X_1) = \emptyset$, $p(X_2) = \{X_1\}$, $p(X_3) = \{X_1, X_2\}$,
$p(X_4) = p(X_5) = \{X_3\}$, $p(X_6) = \{X_4, X_5\}$
$de(X_1) = \{X_2, X_3, X_4, X_5, X_6\}$,
$de(X_2) = \{X_3, X_4, X_5, X_6\}$,
$de(X_3) = \{X_4, X_5, X_6\}$,
$de(X_4) = de(X_5) = \{X_6\}$, $de(X_6) = \emptyset$,
$nd(X_2) = \{X_1\}$, $nd(X_3) = \{X_1, X_2\}$,
$nd(X_4) = \{X_1, X_2, X_3, X_5\}$
$nd(X_5) = \{X_1, X_2, X_3, X_4\}$,
$nd(X_6) = \{X_1, X_2, X_3, X_4, X_5\}$;

• probabilităţile (asociate nodurilor), care definesc reţeaua Bayes sunt:
$P[X_1], P[X_2|X_1], P[X_3|X_1, X_2], P[X_4|X_3], P[X_5|X_3], P[X_6|X_4, X_5]$;

• independenţe condiţionate:

▷ X_4 este condiţional independentă de X_1, X_2, X_5, cunoscând X_3

$$\Rightarrow P[X_4|X_1, X_2, X_3, X_5] = P[X_4|X_3],$$

▷ X_5 este condiţional independentă de X_1, X_2, X_4, cunoscând X_3

$$\Rightarrow P[X_5|X_1, X_2, X_3, X_4] = P[X_5|X_3],$$

▷ X_6 este condiţional independentă de X_1, X_2, X_3, cunoscând X_4, X_5

$$\Rightarrow P[X_6|X_1, X_2, X_3, X_4, X_5] = P[X_6|X_4, X_5];$$

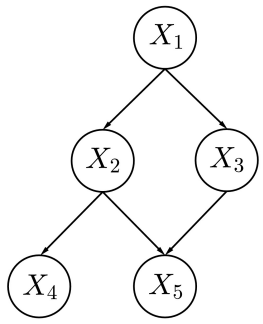
• (exemplu de calcul în reţea Bayes) se ştiu $P(X_1{=}1){=}0.5$, $P(X_2{=}1|X_1{=}1){=}0.6$, $P(X_3{=}1|X_1{=}1, X_2{=}1){=}0.5$, $P(X_4{=}1|X_3{=}1){=}0.4$, $P(X_4{=}1|X_3{=}0){=}0.3$, atunci să se cal-

culeze $P(X_4{=}1, X_2{=}1, X_1{=}1)$:

$$\begin{aligned}
&P(X_4{=}1, X_2{=}1, X_1{=}1)\\
&{=}P(X_4{=}1, X_3{=}1, X_2{=}1, X_1{=}1) + P(X_4{=}1, X_3{=}0, X_2{=}1, X_1{=}1)\\
&{=}P(X_1{=}1)P(X_2{=}1|X_1{=}1)P(X_3{=}1|X_1{=}1, X_2{=}1)P(X_4{=}1|X_3{=}1)\\
&\quad + P(X_1{=}1)P(X_2{=}1|X_1{=}1)P(X_3{=}0|X_1{=}1, X_2{=}1)P(X_4{=}1|X_3{=}0)\\
&{=}0.105.
\end{aligned}$$

♦

Exemplul 2 Se dă reţeaua Bayes din figura alăturată, în care $X_1, ..., X_5$ sunt variabile aleatoare binare. Se ştiu probabilităţile:
$P(X_1 = 0) = 0.4$, $P(X_2 = 0|X_1 = 0) = 0.2$,
$P(X_2 = 0|X_1 = 1) = 0.5$, $P(X_3 = 0|X_1 = 0) = 0.3$,
$P(X_3 = 0|X_1 = 1) = 0.4$,
$P(X_4 = 0|X_2 = 0) = 0.2$, $P(X_4 = 0|X_2 = 1) = 0.5$,
$P(X_5 = 0|X_2 = 0, X_3 = 0) = 0.5$,
$P(X_5 = 0|X_2 = 0, X_3 = 1) = 0.2$,
$P(X_5 = 0|X_2 = 1, X_3 = 0) = 0.7$,
$P(X_5 = 0|X_2 = 1, X_3 = 1) = 0.4$.



Reţea Bayes

a) Să se calculeze

$$P(X_3 = 1|X_2 = 1), P(X_1 = 0, X_3 = 1), P\left(\bigcap_{i=1}^{5}\{X_i = 1\}\right).$$

b) Să se scrie distribuţia de probabilitate a variabilei aleatoare X_3.
Rezolvare: Se calculează: $P(X_1 = 1) = 1 - P(X_1 = 0) = 0.6$

$$P(X_2 = 1|X_1 = 0) = 1 - P(X_2 = 0|X_1 = 0) = 0.8;$$

$$P(X_2 = 1|X_1 = 1) = 1 - P(X_2 = 0|X_1 = 1) = 0.5;$$

$$P(X_3 = 1|X_1 = 0) = 1 - P(X_3 = 0|X_1 = 0) = 0.7;$$

$$P(X_3 = 1|X_1 = 1) = 1 - P(X_3 = 0|X_1 = 1) = 0.6;$$

$$P(X_4 = 1|X_2 = 0) = 1 - P(X_4 = 0|X_2 = 0) = 0.8;$$

$$P(X_4 = 1|X_2 = 1) = 1 - P(X_4 = 0|X_2 = 1) = 0.5;$$

$$P(X_5 = 1|X_2 = 1, X_3 = 1) = 1 - P(X_5 = 0|X_2 = 1, X_3 = 1) = 0.6.$$

a) Are loc:

$$P(X_3 = 1|X_2 = 1) = \frac{P(X_3 = 1, X_2 = 1)}{P(X_2 = 1)}.$$

Folosind formula probabilităţilor totale şi proprietăţile reţelelor Bayes (X_2 este condiţional independentă de X_3, cunoscând X_1)[1]:

- $P(X_3 = 1, X_2 = 1) = P(X_3 = 1, X_2 = 1|X_1 = 0)P(X_1 = 0)$
  $+ P(X_3 = 1, X_2 = 1|X_1 = 1)P(X_1 = 1) =$
  $= P(X_3 = 1|X_1 = 0)P(X_2 = 1|X_1 = 0)P(X_1 = 0)$
  $+ P(X_3 = 1|X_1 = 1)P(X_2 = 1|X_1 = 1)P(X_1 = 1)$
- $P(X_2 = 1) = P(X_2 = 1|X_1 = 0)P(X_1 = 0) + P(X_2 = 1|X_1 = 1)P(X_1 = 1).$

Are loc

$$P(X_1 = 0, X_3 = 1) = P(X_3 = 1|X_1 = 0)P(X_1 = 0).$$

Folosind regula de înmulţire şi proprietăţile reţelelor Bayes (X_2 este condiţional independentă de X_3, cunoscând X_1; X_4 este condiţional independentă de X_1, X_3, cunoscând X_2; X_5 este condiţional independentă de X_1, X_4, cunoscând X_2, X_3)

$$\begin{aligned}
&P(X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1, X_5 = 1)\\
&= P(X_1 = 1)P(X_2 = 1|X_1 = 1)P(X_3 = 1|X_1 = 1, X_2 = 1)\cdot\\
&\quad \cdot P(X_4 = 1|X_1 = 1, X_2 = 1, X_3 = 1)P(X_5 = 1|X_1 = 1, X_2 = 1, X_3 = 1, X_4 = 1) =\\
&= P(X_1{=}1)P(X_2{=}1|X_1{=}1)P(X_3 = 1|X_1 = 1)P(X_4 = 1|X_2 = 1)P(X_5 = 1|X_2 = 1, X_3 = 1).
\end{aligned}$$

b) $P(X_3 = 0) = P(X_3 = 0|X_1 = 0)P(X_1 = 0) + P(X_3 = 0|X_1 = 1)P(X_1 = 1)$
$= 0.12 + 0.24 = 0.36 \Rightarrow P(X_3 = 1) = 0.64$

$$\Rightarrow\ X_3 \sim \begin{pmatrix} 0 & 1 \\ 0.36 & 0.64 \end{pmatrix}.$$

♦

Variabile aleatoare continue

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

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

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

[1]Orice nod X şi nondescendenţii săi $nd(X)$ sunt condiţional independenţi, dacă se dau valorile părinţilor $p(X)$.

P. 14. *Fie f funcţia de densitate a unei v.a. continue X. Au loc proprietăţile:*
(1) $f(t) \geq 0$ *pentru orice* $t \in \mathbb{R}$*;*
(2) $\int_{-\infty}^{\infty} f(t)\, dt = 1$;
(3) $P(a < X \leq b) = \int_a^b f(t)dt \ \forall\, a, b \in \mathbb{R}, a < b$;
(4) $P(X = a) = 0 \ \ \forall\, a \in \mathbb{R}$;
(5) *pentru* $\forall\, a < b, a, b \in \mathbb{R}$ *au loc*

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

Observaţie: Orice funcţie $f : \mathbb{R} \to \mathbb{R}$, care are proprietăţile (1), (2) din **P.14** este o funcţie de densitate.

Exemple de distribuţii clasice continue

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

• funcţia de densitate este

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

Matlab/Octave:
▷ când $a = 0, b = 1$ `rand(M,N)` returnează o matrice $M \times N$ cu valori aleatoare din [0,1]
▷ `unifrnd(a,b,M,N)`, respectiv `(b-a)rand(M,N)+a` returnează o matrice $M \times N$ cu valori aleatoare din $[a, b]$

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

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

• funcţia de densitate este

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

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

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

Matlab/Octave: `normrnd`$(m, \sigma, ...)$

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

• funcţia de densitate este

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

Matlab/Octave: `exprnd`$\left(\frac{1}{\lambda}, ...\right)$

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

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

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

unde

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

este funcţia Gamma.

Matlab/Octave: `trnd`$(n, ...)$

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

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

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

Matlab/Octave: `chi2rnd`$(n, ...)$

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

```
X=exprnd(2,1,10000);
p=mean((2<=X)&(X<=4))
q=mean(X>3)
> p =  0.23280
> q= 0.22060
```

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

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

♣

Def. 20. ***Funcţia de repartiţie*** $F : \mathbb{R} \to [0,1]$ *a unei variabile aleatoare* X *(discrete sau continue) este*

$$F(x) = P(X \leq x) \quad \forall x \in \mathbb{R}.$$

P. 15. *Funcţia de repartiţie* F *a unei variabile aleatoare* X *(discrete sau continue) are următoarele proprietăţi:*
(1) F *este monoton crescătoare, adică pentru orice* $x_1 < x_2$ *rezultă* $F(x_1) \leq F(x_2)$.
(2) $\lim\limits_{x\to\infty} F(x) = 1$ *şi* $\lim\limits_{x\to-\infty} F(x) = 0$.
(3) F *este continuă la dreapta, adică* $\lim\limits_{x\searrow x_0} F(x) = F(x_0)\ \forall\, x_0 \in \mathbb{R}$.
(4) $P(a < X \leq b) = F(b) - F(a)\ \forall a, b \in \mathbb{R}, a < b$.

Observaţie importantă:
▷ Orice funcţie $F : \mathbb{R} \to \mathbb{R}$, care are proprietăţile (1), (2), (3) din **P.15** este o funcţie de repartiţie.
Matlab/Octave:

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

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

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

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

Exemplu: V.a. X urmează legea uniformă pe $[a, b]$ cu $0 < a < b$. Să se arate că $P(X > 0) = 1$ şi să se determine funcţia densitate de probabilitate a variabilei aleatoare $Y = \ln\left(\frac{1}{X}\right)$.

Soluţie: Funcţia de densitate a lui X este:

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

Atunci are loc

$$P(X > 0) = 1 - P(X \le 0) = 1 - \int_{-\infty}^{0} f_X(x)dx = 1.$$

Scriem succesiv

$$F_Y(y) = P(Y < y) = P\left(\ln\left(\frac{1}{X}\right) < y\right) = 1 - F_X(e^{-y}).$$

Derivăm în raport cu y

$$f_Y(y) = f_X(e^{-y})e^{-y}.$$

Folosind definiţia lui f_X, obţinem

$$f_Y(y) = \begin{cases} \dfrac{e^{-y}}{b-a}, & y \in [-\ln b, -\ln a] \\ 0, & \text{altfel}\,. \end{cases}$$

△

Generarea de numere pseudo-aleatoare ce urmează o distribuţie discretă dată (metoda inversei)

Se dau $(x_1, \ldots, x_n)$ (valorile) şi $(p_1, \ldots, p_n)$ (probabilităţile lor). Realizaţi un program care generează N numere pseudo-aleatoare, care urmează *distribuţia discretă*

$$X \sim \begin{pmatrix} x_1 & x_2 & \ldots & x_n \\ p_1 & p_2 & \ldots & p_n \end{pmatrix},$$

folosind numere aleatoare uniform distribuite pe [0,1].

Procedeul de generare al numerelor aleatoare $Y(i)$, $i = \overline{1, N}$, *este*:

• Se citesc valorile $x_1, x_2, \ldots, x_n$ şi probabilităţile corespunzătoare $p_1, p_2, \ldots, p_n$, precum şi numărul N. Fie $p_0 = 0$.

• Se generează N numere aleatoare uniform distribuite pe [0,1]: $U(i)$, $i = \overline{1, N}$.

• Pentru fiecare $i = \overline{1, N}$: $Y(i) = x_k$ dacă şi numai dacă

$$p_0 + p_1 + \cdots + p_{k-1} < U(i) \leq p_0 + p_1 + \cdots + p_k,\ k \in \{1, \ldots, n\}.$$

• Se returnează: $Y(i), i = \overline{1, N}$.

Verificarea procedeului: deoarece U urmează legea uniformă, avem pe baza procedeului de mai sus:
$P(\text{“se generează } x_k\text{”}) = P(p_0+p_1+\cdots+p_{k-1} < U \leq p_0+p_1+\cdots+p_k) = p_k,\ k = 1, \ldots, n,$
deci numerele generate urmează legea de distribuţie discretă dată.

Problemă: Conform statisticilor medicale 46% din oameni au grupa sanguină $\mathbf{0}$, 40% au grupa sanguină $\mathbf{A}$, 10% au grupa sanguină $\mathbf{B}$ şi 4% au grupa sanguină $\mathbf{AB}$. Simulaţi de $N(= 100, 1000)$ ori stabilirea grupei sanguine a unei persoane alese aleator şi afişaţi frecvenţa de apariţie a fiecărei grupe sanguine. Comparaţi rezultatele obţinute cu cele teoretice.

```
function Y=ivtdiscret(x,p,N) %inverse transform method
Y=zeros(1,N);
q=cumsum(p);
for i=1:N
    U=unifrnd(0,1);
    Y(i)=x(find(U<=q,1));
end
%Aplicatia
clc
clear all
close all
N=1000;%numarul de simulari
x=[0,1,2,3];% valorile variabilei aleatoare
p=[0.46,0.4,0.1,0.04];%probabilitatile
Y=ivtdiscret(x,p,N);%generarea celor N numere
ny=hist(Y,length(unique(Y)));%determinarea frecventei de aparitie
[p' ny'/N]
```

Vector aleator continuu

Def. 21. $(X_1, \ldots, X_n)$ *este un* ***vector aleator continuu*** *dacă fiecare componentă a sa este o variabiă aleatoare continuă.*

Def. 22. $f_{(X,Y)} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ *este* ***funcţia de densitate a vectorului aleator continuu*** (X, Y)*, dacă*

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

Def. 23. $F_{(X,Y)} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ *este* ***funcţia de repartiţie a vectorului aleator*** (X, Y) *(discret sau continuu), dacă*

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

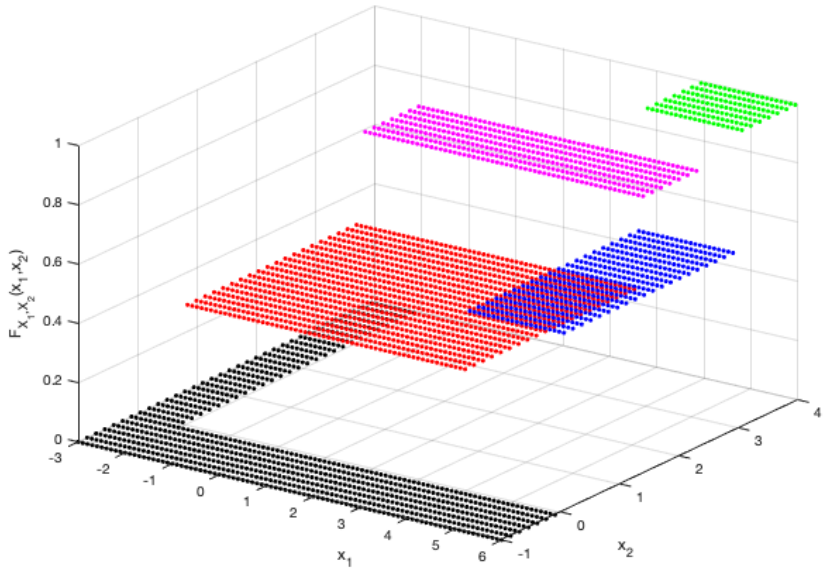
Exemplu: Vectorul aleator (X_1, X_2) este dat prin următorul tabel de contingenţă:

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

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

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

```
%Matlab
clear all; clf; hold on; grid on;
xlim([-3 6]); ylim([-1 4]); zlim([0 1]);
view(35,30)
x=-3:0.1:6; y=-1:0.15:4;
for i=1:length(x)
    for j=1:length(y)
    if x(i)<-2 || y(j)<0
        f(i,j)=0;
```



Funcţia de repartiţie $F_{(X_1,X_2)}$

```
      plot3(x(i),y(j),f(i,j),'k.')
    end
    if -2<=x(i) && x(i)<4 && 0<=y(j)&& y(j)<3
        f(i,j)=0.4;
        plot3(x(i),y(j),f(i,j),'r.')
    end
    if -2<=x(i) && x(i)<4 && 3<=y(j)
        f(i,j)=0.7;
        plot3(x(i),y(j),f(i,j),'m.')
    end
    if 4<=x(i) && 0<=y(j)&& y(j)<3
        f(i,j)=0.6;
        plot3(x(i),y(j),f(i,j),'b.')
    end
    if 4<=x(i) &&  3<=y(j)
        f(i,j)=1;
        plot3(x(i),y(j),f(i,j),'g.')
    end
    end
end
xlabel('x_1'); ylabel('x_2'); zlabel('F_{X_1,X_2}(x_1,x_2)');
```

✸

Observaţie: Cunoscând distribuţia vectorului aleator continuu (X, Y), cum determinăm distribuţia

v.a. continue X respectiv Y?
$\triangleright$ dacă se cunoaşte $f_{(X,Y)}$: f_X, respectiv f_Y, se determină cu

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

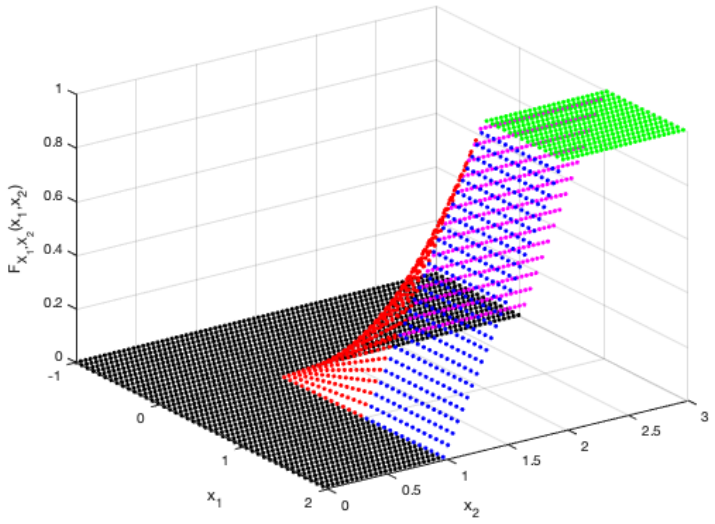
$\triangleright$ dacă se cunoaşte $F_{(X,Y)}$: F_X, respectiv F_Y, se determină cu

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

Exemplu: Funcţia de repartiţie a vectorului aleator (X_1, X_2) este $F_{(X_1,X_2)} : \mathbb{R} \times \mathbb{R} \to [0,1]$

$$F_{(X_1,X_2)}(x_1,x_2) = \begin{cases} 0, & \text{dacă } x_1 < 0 \text{ sau } x_2 < 1 \\ x_1(x_2-1), & \text{dacă } 0 \le x_1 < 1 \text{ şi } 1 \le x_2 < 2 \\ x_1, & \text{dacă } 0 \le x_1 < 1 \text{ şi } 2 \le x_2 \\ x_2 - 1, & \text{dacă } 1 \le x_1 \text{ şi } 1 \le x_2 < 2 \\ 1, & \text{dacă } 1 \le x_1 \text{ şi } 2 \le x_2\,. \end{cases}$$

Ce distribuţie au X_1, respectiv X_2?
R.: Se determină F_{X_1}, F_{X_2} şi $f_{X_1}, f_{X_2} \Longrightarrow X_1 \sim Unif[0,1], X_2 \sim Unif[1,2]$.



Funcţia de repartiţie $F_{(X_1,X_2)}$

Def. 24. *$X_1, \ldots, X_n$ sunt **n variabilele aleatoare independente** (discrete sau continue), dacă*

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

Observaţie ($n = 2$ in definiţia de mai sus): X şi Y sunt **două variabilele aleatoare independente**, dacă

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

adică

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

P. 16. *Variabilele aleatoare continue X (cu funcţia de densitate f_X) şi Y (cu funcţia de densitate f_Y) sunt independente, dacă şi numai dacă*

$$f_X(x)f_Y(y) = f_{(X,Y)}(x,y) \; \forall \; (x,y) \in \mathbb{R}^2,$$

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

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

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

Funcţia de repartiţie are expresia

$$F(x_1,x_2) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f(t_1,t_2)dt_1dt_2 = \left(\frac{x_1-a_1}{b_1-a_1}\right)_* \cdot \left(\frac{x_2-a_2}{b_2-a_2}\right)_*,$$

unde

$$u_* = \begin{cases} 0 & \text{dacă } u < 0 \\ u & \text{dacă } 0 \le u \le 1 \\ 1 & \text{dacă } 1 < u. \end{cases}$$

✿

P. 17. *Pentru un vector aleator continuu (X,Y) au loc proprietăţile:*

1. $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{(X,Y)}(u,v)\,dudv = 1.$
2. *$F_{(X,Y)}$ este funcţie continuă.*
3. *Dacă $F_{(X,Y)}$ este derivabilă parţial în (x,y), atunci are loc:*

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

4. $P((X,Y) \in A) = \underbrace{\int\int}_{A} f_{(X,Y)}(u,v)dudv, \; A \subset \mathbb{R}^2$ *(măsurabilă).*

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

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

Sunt X şi Y v.a. independente? Să se calculeze $P(1 \leq X \leq 2 \leq Y \leq 3)$.
R.: Se calculează $F_X(x) = 1 - e^{-x}$ pentru $x > 0$ şi $F_X(x) = 0$ pentru $x \leq 0$, precum şi $F_Y(y) = 1 - e^{-2y}$ pentru $y > 0$ şi $F_Y(y) = 0$ pentru $y \leq 0$. Se verifică

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

Deci, X şi Y sunt v.a. independente.

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

♡

Operaţii cu v.a. continue

Proprietate: Fie (X, Y) vector aleator continuu cu funcţia de densitate $f_{(X,Y)}$. Atunci $X + Y$ şi $X \cdot Y$ sunt v.a. continue, având funcţiile de densitate:

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_{(X,Y)}(u, z - u) du \quad \forall\, z \in \mathbb{R},$$

$$f_{X \cdot Y}(z) = \int_{-\infty}^{\infty} \frac{1}{|u|} f_{(X,Y)}\Big(u, \frac{z}{u}\Big) du \quad \forall\, z \in \mathbb{R}.$$

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

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(u) f_Y(z - u) du \quad \forall\, z \in \mathbb{R},$$

$$f_{X \cdot Y}(z) = \int_{-\infty}^{\infty} \frac{1}{|u|} f_X(u) f_Y\Big(\frac{z}{u}\Big) du \quad \forall\, z \in \mathbb{R}.$$

Def. 25. ***Valoarea medie a unei v.a. continue*** *X, care are funcţia de densitate f, este*

$$E(X) = \int_{-\infty}^{\infty} t f(t) dt$$

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

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

P. 18. *Proprietăţi ale valorii medii; fie X, Y v.a. continue:*
$\rightarrow$ $E(aX + b) = aE(X) + b$ *for all* $a, b \in \mathbb{R}$*;*
$\rightarrow$ $E(X + Y) = E(X) + E(Y)$;
$\rightarrow$ *Dacă X şi Y sunt variabile aleatoare independente, atunci* $E(X \cdot Y) = E(X)E(Y)$.
$\rightarrow$ *Dacă* $g : \mathbb{R} \rightarrow \mathbb{R}$ *e o funcţie, astfel încât* $g(X)$ *este o v.a. continuă, atunci*

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

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

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

$$P(\text{“elevul ajunge la timp la şcoală”}) = P(X \leq 25) = \frac{25 - 20}{26 - 20} = \frac{5}{6}.$$

$$E(X) = \int_{20}^{26} x \frac{1}{26 - 20} dx = \frac{20 + 26}{2} = 23 \text{ (minute)}.$$

♣

Def. 26. ***Varianţa (dispersia)*** *unei variabile aleatoare X (discrete sau continue) este*

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

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

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

P. 19. *Proprietăţi ale varianţei:*
$\rightarrow$ $V(X) = E(X^2) - E(X)^2$.
$\rightarrow$ $V(aX + b) = a^2 V(X)$ $\forall\, a, b \in \mathbb{R}$.
$\rightarrow$ *Dacă X şi Y sunt variabile aleatoare independente, atunci* $V(X + Y) = V(X) + V(Y)$.

Exemplu: 1) $X \sim Bino(n,p) \Longrightarrow E(X) = np, V(X) = np(1-p)$.
2) Dacă $X \sim N(m, \sigma^2) \Longrightarrow E(X) = m,\ V(X) = \sigma^2$. ♣

► Matlab/Octave: *mean, var, std*

Def. 27. $(X_n)_n$ *este* ***şir de v.a. independente****, dacă* $\forall\ \{i_1, \ldots, i_k\} \subset \mathbb{N}$ *v.a.* $X_{i_1}, \ldots, X_{i_k}$ *sunt independente, adică*

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

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

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

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

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

Def. 28. *Şirul de v.a.* $(X_n)_n$ ***converge aproape sigur*** *la v.a.* X*, dacă*

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

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

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

Exemplu:

$$X_n \sim \begin{pmatrix} -\frac{1}{n} & \frac{1}{n} \\ 0.5 & 0.5 \end{pmatrix} \Rightarrow\ X_n \xrightarrow{a.s.} 0.$$

Exemplu:
$\Omega := [0,1]$ spaţiul de selecţie, fie P probabilitatea pe [0,1] indusă de măsura Lebesgue pe [0,1], adică pentru $\forall \alpha < \beta$ din $[0,1]$ are loc

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

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

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

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

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

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

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

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

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

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

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

$f_n(A) \xrightarrow{a.s.} \frac{1}{2}$ (a se vedea **P.21**, LTNM)

Legea tare a numerelor mari (LTNM)

Legea numerelor mari se referă la descrierea rezultatelor unui experiment repetat de foarte multe ori. Conform acestei legi, rezultatul mediu obţinut se apropie tot mai mult de valoarea aşteptată, cu cât experimentul se repetă de mai multe ori. Aceasta se explică prin faptul că abaterile aleatoare se compensează reciproc.
Legea numerelor mari are două formulări: **legea slabă a numerelor mari (LSNM)** şi **legea tare a numerelor mari (LTNM)**.
△ **Scurt istoric**: Jacob Bernoulli (1655 -1705) a formulat LSNM pentru frecvenţa relativă a unui experiment şi a dat răspunsul la întrebarea "*Putem aproxima empiric probabilităţile*?" (în opera publicată postum, în 1713, *Ars conjectandi*):
▷ Teorema lui Bernoulli afirmă: *"Frecvenţele relative converg în probabilitate la probabilitatea teoretică."*
▷ În cadrul unui experiment poate să apară evenimentul A (*succes*) sau $\bar{A}$ (*insucces*).

• $X_i = 0 \Leftrightarrow$ dacă $\bar{A}$ apare în a i-a repetiţie a experimentului

• $X_i = 1 \Leftrightarrow$ dacă A apare în a i-a repetiţie a experimentului
$\Rightarrow X_i \sim Bernoulli(p)$ cu $p := P(A)$

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

• $X_1, ..., X_n$ sunt v.a. independente
• frecvenţa relativă de apariţie a lui A este

$$f_n(A) = \frac{1}{n}(X_1 + ... + X_n); \quad f_n(A) \text{ este o v.a.}$$

• $(X_n)_n$ verifică LSNM, adică

$$\lim_{n\to\infty} P\left(\left|f_n(A) - P(A)\right| > \varepsilon\right) = 0 \; \forall \, \varepsilon > 0.$$

△

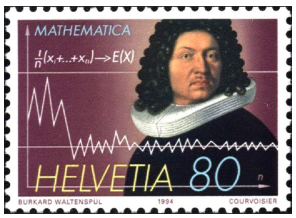


Fig. 5. Jacob Bernoulli (timbru emis în 1994 cu ocazia Congresului Internaţional al Matematicienilor din Elveţia)

Def. 29. *Şirul de v.a.* $(X_n)_n$ *cu* $E|X_n| < \infty \; \forall \; n \in \mathbb{N}$ *verifică* ***legea tare a numerelor mari (LTNM)*** *dacă*

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

adică

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

P. 20. *Fie* $(X_n)_n$ *şir de v.a. independente având aceeaşi distribuţie (şi există* $m = E(X_n) \; \forall \; n \in \mathbb{N}$*)* $\Rightarrow (X_n)_n$ *verifică* ***LTNM****, adică*

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

În simulări: $\frac{1}{n}(X_1 + \cdots + X_n) \approx m,$ dacă n este suficient de mare.
Aplicaţie Matlab/Octave: Simulare LTNM

```
clear all
clf
hold on
n=300;
x=unidrnd(6,1,n);
for i=1:n
    s(i)=sum(x(1:i))/i;
     y(i)=i;
     plot(y(i),s(i),'b.')
     plot(y(i),3.5,'k-')
end
plot(y,s,'r-')
xlabel('Nr. aruncari zar')
ylabel('Media numerelor aparute')
```

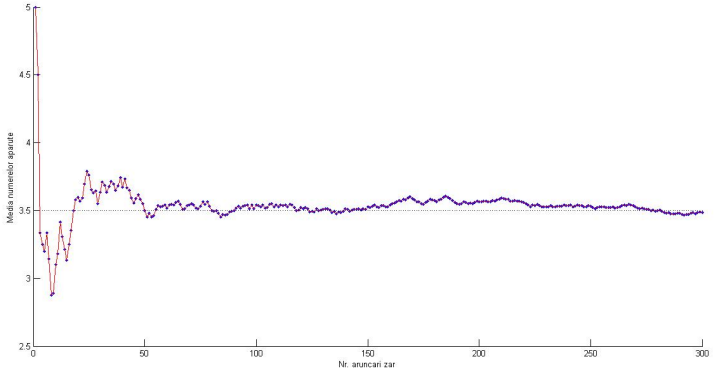


Fig. 4. Simulare LTNM

P. 21. *Fie A un eveniment asociat unei experienţe, repetăm experienţa de n ori (în aceleaşi condiţii date). LTNM: cu cât repetăm mai des un experiment ($n \to \infty$), cu atât mai bine aproximează frecvenţa relativă $f_n(A)$ a evenimentului A probabilitatea sa (teoretică) de apariţie $P(A)$:*

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

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

Statistică matematică

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

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

• statistica inferenţială: metode de interpretare a rezultatelor obţinute prin metodele statisticii

descriptive, utilizate apoi pentru luarea deciziilor.
► O *colectivitate* sau *populaţie statistică* $\mathcal{C}$ este o mulţime de elemente care au anumite însuşiri comune ce fac obiectul analizei statistice. Numărul elementelor populaţiei se numeşte *volumul populaţiei.*
Exemple de populaţii statistice: mulţimea persoanelor dintr-o anumită ţară, localitate, zonă etc. într-un anumit an; multimea gospodăriilor din Romania la un moment dat; mulţimea consumatorilor unui anumit produs; mulţimea societăţilor care produc un anumit produs; angajaţii unei societăţi; studenţii unei facultăţi.
► *Eşantionul* $\mathcal{E}$ reprezintă o submulţime a unei populaţii statistice $\mathcal{E} \subset \mathcal{C}$, constituită după criterii bine stabilite:
a) să fie aleatoare;
b) toate elementele colectivităţii să aibe aceeaşi şansă de a fi alese în eşantion;
c) eşantionul să fie reprezentativ (structura eşantionului să fie apropiată de structura populaţiei);
d) volumul eşantionului să fie suficient de mare.
► *Unitatea statistică* (indivizii) este elementul, entitatea de sine stătătoare a unei populaţii statistice, care posedă o serie de trăsături caracteristice ce-i conferă apartenenţa la populaţia studiată.
De exemplu: *unitatea statistică simplă*: un salariat, un student, un agent economic, o trăsătură, o părere; *unitatea statistică complexă*: o grupă de studenţi sau o echipă de salariaţi, o familie sau o gospodărie, o categorie de mărfuri.
► *Variabila statistică* sau *caracteristica* reprezintă o însuşire, o proprietate măsurabilă a unei unităţi statistice, întâlnită la toate unităţile care aparţin aceleiaşi colectivităţi şi care prezintă variabilitate de la o unitate statistică la alta. Caracteristica sau variabila statististică corespunde unei variabile aleatoare.
Exemple de caracteristici: vârsta, salariul, preferinţele politice, preţul unui produs, calitatea unor servicii, nivelul de studii.
a) variabile (caracteristici) continue $\rightarrow$ iau un număr infinit şi nenumărabil de valori într-un interval sau reuniune de intervale (de ex.: greutatea, înălţimea, valoarea glicemiei, temperatura aerului)
b) variabile (caracteristici) discrete $\rightarrow$ iau număr finit sau infinit dar numărabil de valori discrete (de ex.: numări elevi ai unei şcoli, numărul liceelor existente într-un oraş, valoarea IQ)
$\triangleright$ caracteristicile de la a) şi b) sunt variabile numerice (cantitative)
c) variabile (caracteristici) nominale (de ex.: culoarea ochilor, ramura de activitate, religia)
d) variabile (caracteristici) nominale ordinale (de ex.: starea de sănătate / calitatea unor servicii - precară, mai bună, bună, foarte bună)
e) variabile (caracteristici) dihotomiale (binare) (de ex.: stagiul militar - satisfăcut/nesatisfăcut, starea civilă - căsătorit/necăsătorit)
$\triangleright$ caracteristicile de la c),d),e) sunt variabile calitative
$\triangleright$ variabilele nominale mai sunt numite variabile categoriale

▶ *Datele statistice* reprezintă observaţiile rezultate dintr-o cercetare statistică, sau ansamblul valorilor colectate în urma unei cercetări statistice.
De exemplu: un angajat al unei companii are o vechime de 6 ani în muncă. Angajatul reprezintă unitatea statistică, vechimea în muncă este caracteristica (variabila) cercetată, iar 6 este valoarea acestei caracteristici.

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

Distribuţia caracteristicii X poate fi

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

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

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

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

• **Exemple de estimatori (funcţii de selecţie)** sunt: media de selecţie, dispersia de selecţie, funcţia de repartiţie empirică (de selecţie) .
▷ Estimatorii (funcţiile de selecţie) se folosesc în statistică pentru estimarea punctuală a unor parametri necunoscuţi, pentru obţinerea unor intervale de încredere pentru parametri necunoscuţi, pentru verificarea unor ipoteze statistice.

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

▶ media de selecţie

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

▶ valoarea mediei de selecţie

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

▶ varianţa (dispersia) de selecţie

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

▶ valoarea varianţei (dispersiei) de selecţie

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

▶ abaterea standard de selecţie

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

▶ valoarea abaterii standard de selecţie

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

▶ funcţia de repartiţie empirică $\hat{F}_n : \mathbb{R} \times \Omega \to \mathbb{R}$

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

▶ valoarea funcţiei de repartiţie empirice

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

Def. 30. *$g(X_1, \ldots, X_n)$ este estimator nedeplasat pentru parametrul necunoscut θ, dacă*

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

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

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

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

Observaţii:
1) Media de selecţie $\bar{X}_n$ este estimator nedeplasat şi consistent pentru media teoretică $E(X)$ a caracteristicii X; în practică $E(X) \approx \bar{x}_n$.
2) Varianţa (dispersia) de selecţie $\tilde{S}_n^2$ este estimator nedeplasat şi consistent pentru varianţa teoretică $V(X)$ a caracteristicii X; în practică $V(X) \approx \tilde{s}_n^2$.
3) Funcţia de repartiţie de selecţie calculată în $x \in \mathbb{R}$: $\hat{F}_n(x)$ este estimator nedeplasat şi consistent pentru $F_X(x)$, care este valoarea funcţiei de repartiţie teoretice calculată în x; în practică $F_X(x) \approx \hat{F}_n(x)$.

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

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

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

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

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

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

Avem cazul: $r = 1$, calculăm $E(X) = \frac{a}{2}$, $n = 10$, $\bar{x}_n = 0.468$. Se rezolvă

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

Estimatorul pentru parametrul necunoscut a este

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

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

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

Parametrul necunoscut a este estimat cu valoarea 0.936.

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