

Curs - Probabilități și Statistică 2024/2025
Secția Informatică

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



Teoria Probabilităților

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

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

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

Fenomene și procese aleatoare apar, de exemplu, în:

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



[Sursa: www.financialmarket.ro]

→ **în informatică:**

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

▷ <https://www.random.org/randomness/>

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

Exemplu: Generarea de valori aleatoare (în Python)

```
import random
r = random.random()
print("Valoare aleatoare din intervalul (0,1):",r)
N = random.randint(-1,5)
print("Valoare aleatoare din intervalul [-1,5]:",N)
Lista = ["AB", "XY", "EF", "MN","FG"]
print("O alegere aleatoare din Lista: ", random.choice(Lista))
print("Alegere aleatoare de 6-ori din Lista: ", random.choices(Lista, k=6))
# cu returnare
print("Alegere aleatoare de 4-ori din Lista (fara returnare): ", random.sample(Lista,4))
# fara returnare
```

```
import numpy as np
N=30
R = np.random.randint(1,7,size=N)
#un vector cu N de elemente, din multimea {1,2,...6}
print("Numere obtinute la aruncarea de",N,"-ori a zarului:\n",R)
total= sum(R==6)
print("De cate ori s-a obtinut 6 in ",N," aruncari ale zarului:",total)
```

Algoritmi aleatori

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

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

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

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

Exemplu: Random QuickSort

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

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

Exemplu:

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

Exercițiu: Fie S un vector cu 300 de elemente, din mulțimea $\{0, 1, 2\}$ (ordinea lor este necunoscută; se presupune că șirul conține cel puțin un 0).

→ De care tip este următorul algoritm (scris în Python)?

```
import numpy as np
print("Prima versiune")
N=300
S = np.random.randint(0,3, size = N) #se genereaza date
#print(S)
#un vector cu N elemente aleatoare din multimea {0,1,2}
k=1
i= np.random.randint(low=0, high=N)
while S[i] != 0:
    print("iteratia:",k)
    print("S[" ,i, "]= ",S[i])
    i= np.random.randint(low=0, high=N)
    k=k+1
# k = numarul iteratiilor pana se gaseste aleator un 0
if S[i]==0:
    print("iteratia:",k)
    print("S[" ,i, "]= ",S[i]) # i indicele, pentru care S[i]=0
print("La iteratia",k,"s-a gasit aleator un 0.")
```

Răspuns: Algoritm de tip Las Vegas.

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

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

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

Noțiuni introductive:

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

- **Evenimentul** este rezultatul unui experiment.

Exemple:

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

▷ experiment: extragerea unei cărți de joc, eveniment: s-a extras un as

▷ experiment: extragerea unui număr la loto, eveniment: s-a extras numărul 27

- **evenimentul imposibil**, notat cu \emptyset , este evenimentul care nu se realizează niciodată la efectuarea experienței aleatoare

- **evenimentul sigur** este un eveniment care se realizează cu certitudine la fiecare efectuare a experienței aleatoare

- **spațiul de selecție**, notat cu Ω , este mulțimea tuturor rezultatelor posibile ale experimentului considerat

◇ spațiul de selecție poate fi finit sau infinit

- dacă A este o submulțime a lui Ω atunci A se numește **eveniment aleator**, iar dacă A are un singur element atunci A este un **eveniment elementar**.

▷ *O analogie între evenimente și mulțimi* permite o scriere și o exprimare mai comode ale unor idei și rezultate legate de conceptul de eveniment aleator.

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

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

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

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



Operații cu evenimente

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

- dacă $A, B \subseteq \Omega$, atunci **evenimentul intersecție** $A \cap B$ este un eveniment care se produce dacă cele două evenimente A și B se produc în același timp

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

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

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

- *Au loc relațiile:* $A \cup \bar{A} = \Omega$, $A \cap \bar{A} = \emptyset$, $\bar{\bar{A}} = A$.

Relații între evenimente

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

- dacă A implică B și B implică A , atunci evenimentele A și B sunt **egale**: $A = B$

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

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

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

asociative

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

și **distributive**

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

satisfac **legile lui De Morgan**

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

Frecvența relativă și frecvența absolută

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

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

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

Definiția clasică a probabilității

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

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

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

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

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

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

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

$P(A) = \frac{4}{24} = 0.25$ probabilitatea (teoretică) a evenimentului A .



Exercițiu: Se alege aleator un număr din mulțimea $\{1, 2, 3, \dots, 99\}$. Care este probabilitatea ca acesta să nu fie divizibil nici cu 4, nici cu 6?



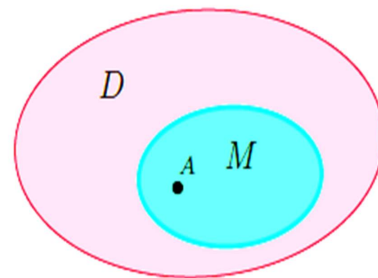
Definiția axiomatică a probabilității

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

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

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

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



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

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

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

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

↪ P satisface anumite axiome (vezi Def. 5)

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

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

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

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

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

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

este o σ -algebră pe mulțimea B . ◇

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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



Evenimente independente

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

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

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

Exercițiu: Se aruncă un zar de două ori.

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

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

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



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

(1) A și B sunt independente.

(2) \bar{A} și B sunt independente.

(3) A și \bar{B} sunt independente.

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

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

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

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

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

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

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

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

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

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

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

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

Sunt cele 3 evenimente *independente în totalitate*?

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

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

Estimăm prin simulări Python probabilitățile următoarelor evenimente:

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

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

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

```
import random
import numpy
a=0
N=10000
for _ in range(N):
    x=random.choices([1,2,3,4,5,6],k=4) # alegere aleatoare cu returnare
    a+= (x.count(6)>0)
print("din simulari P(A) este:",a/N)
b=0
for _ in range(N):
    x1=random.choices([1,2,3,4,5,6],k=24)
    x2=random.choices([1,2,3,4,5,6],k=24)
    s=numpy.add(x1,x2)
    b+= (sum(s==12)>0)
print("din simulari P(B) este:",b/N)
c=0
for _ in range(N):
    y1=random.choices([1,2,3,4,5,6],k=25)
    y2=random.choices([1,2,3,4,5,6],k=25)
    s=numpy.add(y1,y2)
    c+= (sum(s==12)>0)
print("din simulari P(C) este:",c/N)
X=[a,b,c]
```

```

str="ABC"
z=sorted([a,b,c])
i0= X.index(z[0]) # index din X pt care este probabilitatea cea mai mica
i1= X.index(z[1])
i2= X.index(z[2]) # index din X pt care este probabilitatea cea mai mare
print("P(",str[i0],")<P(",str[i1],")<P(",str[i2],")")
# probabilitatile evenimentelor afisate in ordine crescatoare

```

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

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

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

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

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

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

Concluzie: Evenimentul A are șansele cele mai mari de câștig. ◇

Probabilitate condiționată

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

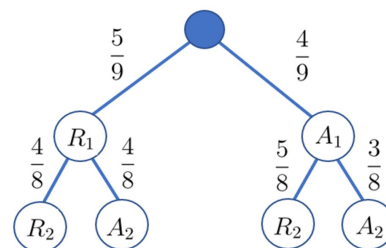
R.: pentru $i \in \{1, 2\}$ fie evenimentele

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

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

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

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

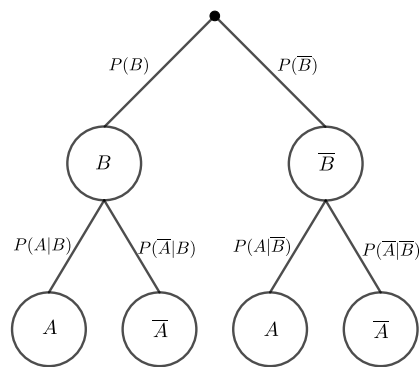


Extragere fără reținere

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

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

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



Probabilități condiționate

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

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



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

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

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

R.: Se consideră evenimentele:

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

B_1 : prima bilă extrasă este roșie;

C_1 : prima bilă extrasă are înscris un număr impar;

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

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

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

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



Exemplu: Ce probabilități calculează programul de mai jos?

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

❓ Care sunt valorile teoretice pentru p_1 , p_2 , p_3 , din acest exemplu?



P. 6. (Formula înmulțirii probabilităților)

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

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

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

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

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

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

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

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

b) cele 2 bile să aibă aceeași culoare?

c) a doua bilă să fie albastră?

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

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

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

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

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

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

c) folosim formula probabilității totale P.7:

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

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

e) formula de înmulțire a probabilităților P.6:

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



Formula lui Bayes

Formula lui Bayes este o metodă de a "corecta" (a revizui, a îmbunătăți) pe baza unor noi date (informații) disponibile o probabilitate determinată apriori. Se pornește cu o estimare pentru probabilitatea unei anumite ipoteze H (engl. *hypothesis*). Dacă avem noi date (date

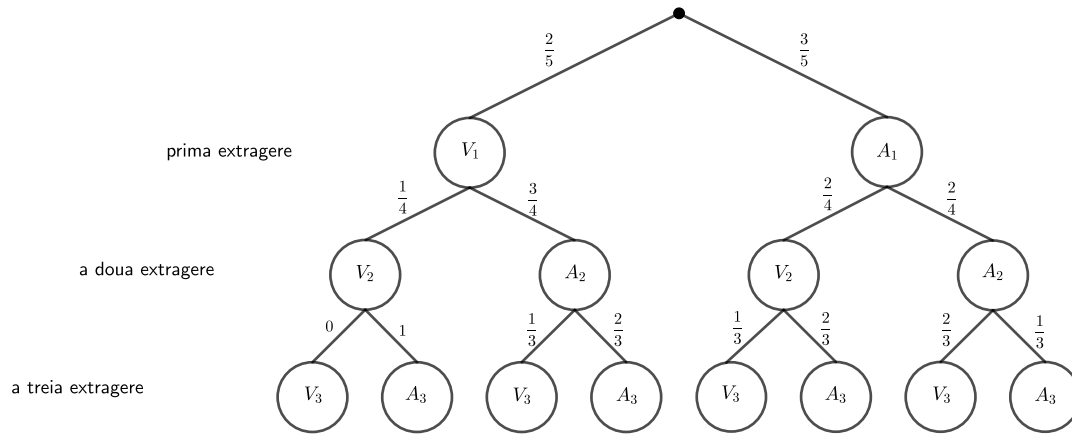


Fig. 3. Extragere fără returnare

de antrenare, dovezi, informații, evidențe - engl. *evidence*) E , ce privesc ipoteza H , se poate calcula o probabilitate "corectată" pentru ipoteza H , numită probabilitate posterioară (a-posteriori).

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

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

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

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

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

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

Formula lui Bayes este în acest caz

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

P. 7. (Formula lui Bayes)

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

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

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

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

- ▷ cu formula lui Bayes se calculează probabilitățile pentru ipoteze, cunoscând evidența: $P(H_j|E)$, $j \in \{1, 2, \dots, n\}$, care se numesc **probabilități posterioare** (ulterioare);
- ▷ $P(E|H_i)$, $i \in \{1, 2, \dots, n\}$, reprezintă verosimilitatea (engl. *likelihood*) datelor observate.
- ▷ Se pot calcula probabilitățile *cauzelor*, date fiind (cunoscând / știind) *efectele*; formula lui Bayes ne ajută să diagnosticăm o anumită situație sau să testăm o ipoteză.

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

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

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

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

▷ din statistici anterioare sunt cunoscute:

- $p = P(H)$, probabilitatea ca o persoană selectată aleator din populație să sufere de alergia \mathcal{A} ;

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

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

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

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

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

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

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

Exemplu cu date statistice: 2120 persoane au fost testate în cadrul unui program de screening, privind alergiile \mathcal{A} .

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

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

Există următoarele informații:

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

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

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

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

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

Matricea de confuzie (engl. *confusion matrix*)

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

Matricea de confuzie construită cu datele statistice din acest exemplu

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

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

¹numărul de elemente din $H \cap E$

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

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



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

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

Variable aleatoare

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

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

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

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

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

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

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

O variabilă aleatoare este:

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

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

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

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

Variabile aleatoare numerice - definiție formală

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

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

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

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

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

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

▷ O variabilă aleatoare discretă X este caracterizată de distribuția de probabilitate!

▷ Notăm $\{X = x_i\} = \{\omega \in \Omega : X(\omega) = x_i\}, i \in I$; acesta este un eveniment din \mathcal{K} pentru fiecare $i \in I$.

Distribuții discrete clasice

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

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

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

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

► Python: `scipy.stats.randint`

```
# Exemplu Unid(6) - Histograma
from scipy.stats import randint
import numpy
import matplotlib.pyplot as plt
from matplotlib.pyplot import bar, show, hist, grid, legend, xticks
N=400
a=1; b=7
R = randint.rvs(a, b, size = N)
print ("Valori aleatoare: \n", R)
x, count = numpy.unique(R, return_counts=True)
print ("Valorile:", x, "au frecvențele:", count)
bar(x, count, width=0.8, color="cyan", edgecolor="black") # deseneaza histograma
plt.grid()
plt.xlabel("valorile")
plt.ylabel("frecvente absolute")
plt.title("Unid(6)")
xticks(range(0,b))
show()
```

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

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

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

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



► Python: `scipy.stats.bernoulli`

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

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

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

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

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

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

→ are loc **formula binomială**

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

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

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

Exemplu: Un client accesează o dată pe zi o anumită pagină web cu probabilitatea 0.6. Cu ce probabilitate clientul accesează această pagină în total de 3 ori în următoarele 10 zile?

R.: $C_{10}^3 0.6^3 0.4^7$

► Python: `scipy.stats.binom`

```
# Exemplu distributia binomiala Bino(10, 0.6)
import numpy
import matplotlib.pyplot as plt
from matplotlib.pyplot import bar, show, hist, grid, legend, xticks
from scipy.stats import binom
```

```

N=1000
n=10; p=0.6 # se genereaza date pentru distributia Bino (n,p)
data = binom.rvs(n, p,size= N)
z, count = numpy.unique(data, return_counts=True)
print("Valorile",z,"au frecventele absolute:",count)
bar(z,count, width=0.8,color="yellow", edgecolor="black")
plt.grid()
plt.xlabel("valorile distributiei")
plt.ylabel("frecvente absolute")
plt.title("Bino(10,0.6)")
s=sum(data==3)
print("Din simulari: clientul acceseaza pag. de 3 ori in urmatoarele 10 zile cu
      probabilitatea", s/N)
xticks(range(0,n+1))

```

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

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

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

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

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



Exerciții: O rețea de laborator este compusă din 15 calculatoare. Rețeaua a fost atacată de un virus nou, care atacă un calculator cu o probabilitatea 0.4, independent de alte calculatoare. Care este probabilitatea ca virusul a atacat
a) cel mult 10; b) cel puțin 10; c) exact 10 calculatoare?



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

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

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

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

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

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

► Python: `scipy.stats.hypergeom`

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

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

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

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

Scriem

$$P(X = 1) = P(A_1 \cap N_2 \cap N_3) + P(N_1 \cap A_2 \cap N_3) + P(N_1 \cap N_2 \cap A_3),$$

$$P(A_1 \cap N_2 \cap N_3) = P(A_1)P(N_2|A_1)P(N_3|A_1 \cap N_2) = \frac{2}{5} \cdot \frac{3}{4} \cdot \frac{2}{3} = \frac{1}{5}$$

$$P(N_1 \cap A_2 \cap N_3) = P(N_1)P(A_2|N_1)P(N_3|N_1 \cap A_2) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} = \frac{1}{5}$$

$$P(N_1 \cap N_2 \cap A_3) = P(N_1)P(N_2|N_1)P(A_3|N_1 \cap N_2) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} = \frac{1}{5}$$

$$\Rightarrow P(X = 1) = \frac{3}{5}.$$

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

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

2) Loto 6 din 49

• Care este probabilitatea de a nimeri exact 4 numere câștigătoare?

R.: Între cele 49 de bile exact $n_1 = 6$ sunt câștigătoare (“bilele albe”) și $n_2 = 43$ necâștigătoare (“bilele negre”). Probabilitatea ca din $n = 6$ extrageri fără returnare, exact $k = 4$ numere să fie câștigătoare (ordinea nu contează) este $p = \frac{C_6^4 C_{43}^2}{C_{49}^6}$.

• Fie v.a. X numărul de numere ghicite, jucând cu o singură variantă la ”Loto 6 din 49”. Scrieți distribuția de probabilitate a v.a. X .

◇

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

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

• A = 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* apariția primului A (“succes”) \Rightarrow valori posibile: $X \in \{0, 1, \dots\}$

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

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

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



Exercițiu: Considerăm v.a. X ca fiind numărul format astfel: dintr-o cutie cu 9 bile numerotate de la 1 la 9 sunt extrase aleator, succesiv, fără returnare, 2 bile, formând astfel un număr din două cifre, prima cifră fiind numărul primei bile, iar cea de-a doua cifră, fiind numărul celei de-a doua bile extrase.

a) Determinații distribuția de probabilitate a v.a. X .

b) Calculați probabilitatea $P(X < 90)$.

Variable aleatoare independente

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

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

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

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

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

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

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

(1) X și Y sunt v.a. sunt independente;

(2) $P(X = x|Y = y) = P(X = x) \quad \forall x \in \{x_i, i \in I\}, y \in \{y_j, j \in J\}$;

(3) $P(Y = y|X = x) = P(Y = y) \quad \forall x \in \{x_i, i \in I\}, y \in \{y_j, j \in J\}$;

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

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

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

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

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

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

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

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

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

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

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

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

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

$\begin{matrix} Y \\ X \end{matrix}$	\dots	y_j	\dots
\vdots	\vdots	\vdots	\vdots
x_i	\dots	p_{ij}	\dots
\vdots	\vdots	\vdots	\vdots

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

următorul tabel:

$\begin{matrix} Y \\ X \end{matrix}$	0	1
-1	$\frac{1}{4}$	$\frac{1}{2}$
2	$\frac{1}{8}$	$\frac{1}{8}$

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

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

b) Sunt X și Y v.a. independente?

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

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

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

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

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

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

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

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

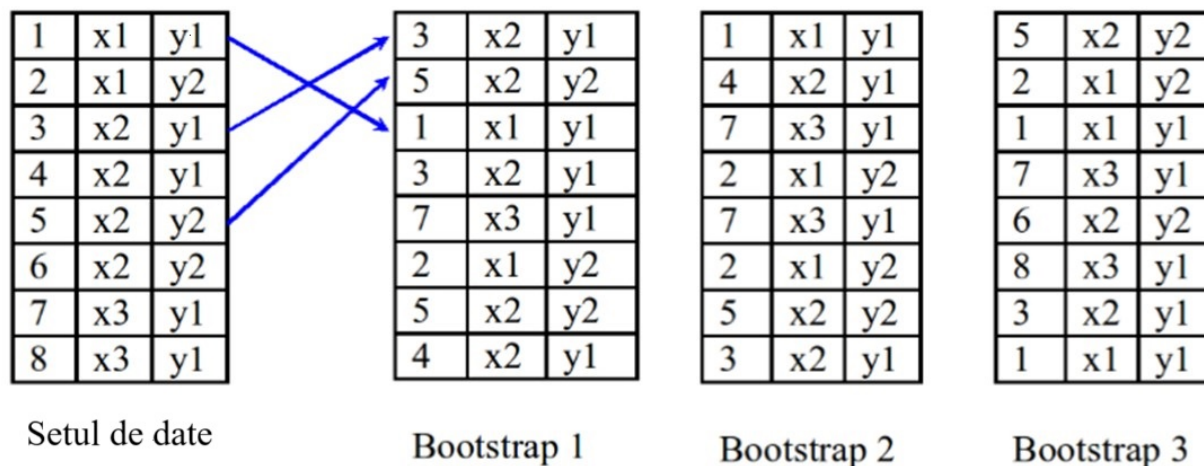
▷ cazul $r = 2$ corespunde distribuției binomiale (modelul binomial cu bile de două culori într-o urnă, a se vedea pg. 22): (X_1, X_2) este un vector aleator discret, iar $X_1 + X_2 = n$; X_1 și X_2 *nu* sunt v.a. independente.

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

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

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

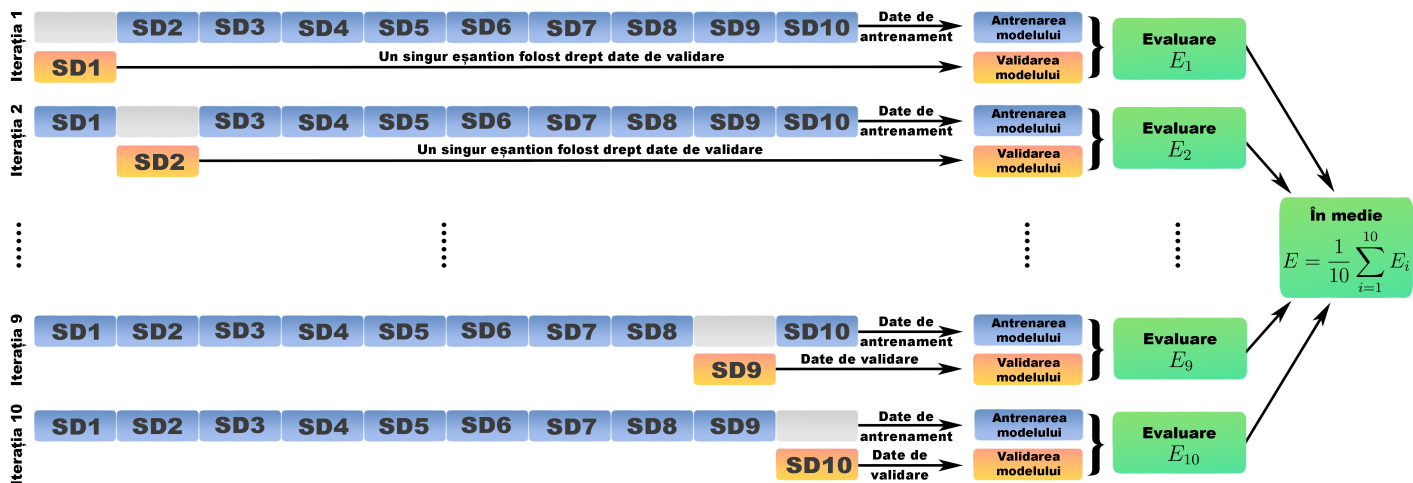
▷ Cazul $r = 2$ corespunde **distribuției hipergeometrice**.



Metoda bootstrap

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

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



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

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

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

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

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

Operații cu variabile aleatoare (numerice)

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

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

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

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

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


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


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

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

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

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

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

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

Clasificarea naivă Bayes

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

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

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

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

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

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

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

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

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

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

Din formula probabilității totale P.5 avem:

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

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

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

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

✱

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

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

Exemplu de clasificare naivă Bayes

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

Tabel de date obținute în urma unor observații

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

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

Considerăm evenimentul următor, denumit *vector de attribute*:

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

Se caută o clasă pentru *E*, stabilind care din următoarele probabilități este mai mare: $P(T = a|E)$ sau $P(T = r|E)$; aceasta este **metoda de probabilitate**

maximă. Știind că vremea este ploioasă și este amiază, ce *previziune* se poate face despre trafic (*aglomerat a* sau *relaxat r*)?

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

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

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

$$P(V = p, Ti = di | T = a) = P(V = p | T = a)P(Ti = di | T = a).$$

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

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

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

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

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

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

și

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

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

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

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

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

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

★

Valoarea medie a unor variabile aleatoare discrete

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

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

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

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

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

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

→ $E(X + Y) = E(X) + E(Y)$;

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

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

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

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

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

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

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

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

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

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

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

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

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

Așadar jucătorul câștigă în medie 5 u.m.

```
import numpy
import random
s=[]
N=1000
for _ in range(N):
    jocuri = random.choices([-1,-1,-1,-1,2,3],k=30)
    s.append(sum(jocuri))
print("Castig mediu (dupa 30 jocuri):",numpy.mean(s))
```

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

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

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

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

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

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

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

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

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

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

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

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

Determinați valoarea medie a lui X .



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

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

Observație:

▷ Orice funcție $F : \mathbb{R} \rightarrow \mathbb{R}$, care are proprietățile (1), (2), (3) din **P.10** este o funcție de repartiție.

► Python:

`scipy.stats.binom.cdf(x,n,p)`, `scipy.stats.hypergeom.cdf(x,M,n,N)` calculează $F(x) = P(X \leq x)$ pentru $X \sim \text{Bino}(n, p)$, respectiv $X \sim \text{Hyge}(M, n, N)$.

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

Variabile aleatoare continuă

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

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

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

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

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

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

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

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

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

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

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

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

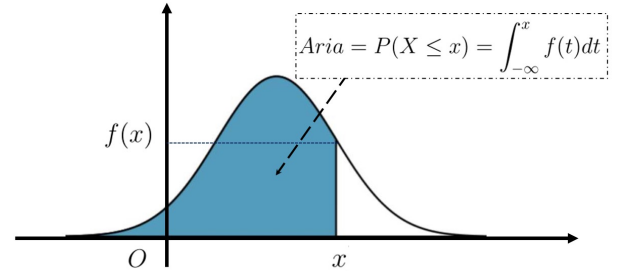
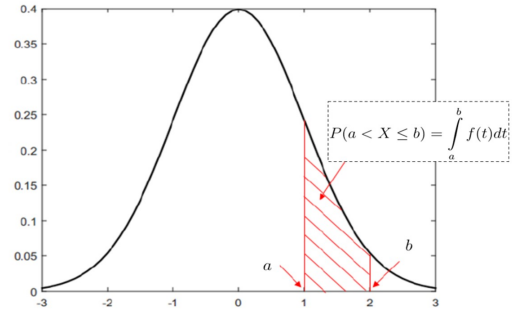
(4) $P(X = a) = 0 \forall a \in \mathbb{R}$;

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

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

(6) $P(X \in M) = \int_M f(t)dt, M \subseteq \mathbb{R}$;

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



$$(8) \lim_{x \rightarrow \infty} F(x) = 1 \quad \text{și} \quad \lim_{x \rightarrow -\infty} F(x) = 0.$$

(9) dacă F este derivabilă în punctul x , atunci $F'(x) = f(x)$.

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

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

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

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

Exemple de distribuții clasice continue

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

• funcția de densitate este

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

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



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

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

- funcția de densitate este

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

- Pentru $\mu = 0, \sigma = 1$: $N(0, 1)$ se numește *distribuția standard normală*.
- Distribuția normală se aplică în: măsurarea erorilor (de ex. termenul eroare în analiza regresională), în statistică (teorema limită centrală, teste statistice) etc.

```
# functia de densitate pentru distributia normala
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import norm
x = np.linspace(-6, 6, 101)
sigma= [1, 1.5, 2] # valori pentru sigma
for t in sigma:
    y = norm.pdf(x, loc=0, scale=t) # scale= deviatia standard
    plt.plot(x, y, label=f"Norm(0, {t**2})")
plt.xlabel("x")
plt.ylabel("functia de densitate")
plt.title("Distributia normala")
plt.legend()
plt.show()
```

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

- funcția de densitate este

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

```
# functia de densitate pentru distributia exponentiala
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import expon
x = np.linspace(0, 4, 100)
L= [1, 2, 3] # valori pentru Lambda
for t in L:
    y = expon.pdf(x, scale=1/t)
    plt.plot(x, y, label=f"Exp({t})")
plt.xlabel("x")
plt.ylabel("functia de densitate")
plt.title("Distributia exponentiala")
plt.legend()
plt.show()
```

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

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

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

unde funcția Gamma este

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

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

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

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

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

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

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

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

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

Vector aleator

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

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

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

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

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

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

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

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

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

ceea ce este echivalent cu

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

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

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

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



Valoarea medie a unei variabile aleatoare continue

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

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

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

P. 12. Proprietăți ale valorii medii: fie X, Y v.a. continue:

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

→ $E(X + Y) = E(X) + E(Y)$.

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

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

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

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

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

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

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

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

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



Varianța unei variabile aleatoare

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

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

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

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

P. 13. *Proprietăți ale varianței (pentru v.a. discrete sau continue):*

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

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

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

► Python: `numpy.mean`, `numpy.var`, `numpy.std`

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

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

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

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

Proprietăți (recapitulare)

V.a. discretă

- caracterizată de distribuția de probabilitate discretă

$$X \sim \left(P(X = x_i) \right)_{i \in I}$$

- $\sum_{i \in I} P(X = x_i) = 1$
- $P(X \in A) = \sum_{i \in I: x_i \in A} P(X = x_i)$
- funcția de repartiție $F(x) = P(X \leq x) \quad \forall x \in \mathbb{R}$
- $F(x) = \sum_{i \in I: x_i \leq x} P(X = x_i) \quad \forall x \in \mathbb{R}$
- F este funcție continuă la dreapta
- F este discontinuă în punctele $x_i, \forall i \in I$
- $\forall a < b, a, b \in \mathbb{R}$

$$P(a \leq X \leq b) = \sum_{i \in I: a \leq x_i \leq b} P(X = x_i)$$

- $P(X = a) = 0$ dacă $a \notin \{x_i : i \in I\}$

- valoarea medie $E(X) = \sum_{i \in I} x_i P(X = x_i)$

V.a. continuă

- caracterizată de funcția de densitate f

$$P(X \leq x) = \int_{-\infty}^x f(t) dt$$

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

$$\bullet P(X \in A) = \int_A f(t) dt$$

- funcția de repartiție $F(x) = P(X \leq x) \quad \forall x \in \mathbb{R}$

$$\bullet F(x) = \int_{-\infty}^x f(t) dt \quad \forall x \in \mathbb{R}$$

- F este funcție continuă în orice punct $x \in \mathbb{R}$

- $\forall a < b, a, b \in \mathbb{R}$

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

$$\bullet P(X = a) = \int_a^a f(t) dt = 0 \quad \forall a \in \mathbb{R}$$

- dacă F este derivabilă în punctul x
 $\Rightarrow F'(x) = f(x)$

$$\bullet \text{valoarea medie } E(X) = \int_{-\infty}^{\infty} t f(t) dt.$$

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

R.: Pentru $i \in \{1, \dots, n\}$ fie $X_i \sim \text{Bernoulli}(p)$ (adică $P(X_i = 1) = p, P(X_i = 0) = 1 - p$), astfel încât X_1, \dots, X_n sunt v.a. independente. Observăm că $X_1 + \dots + X_n \sim \text{Bino}(n, p)$. Deci, $X_1 + \dots + X_n$ și X au aceeași distribuție, așadar ele au aceeași valoare medie și aceeași varianță

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

V.a. X_1, \dots, X_n sunt independente și folosind P.13, obținem

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

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

R.: Funcția de densitate a lui X este

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

Când $\mu = 0$ și $\sigma = 1$ obținem funcția de densitate a distribuției normale standard

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

Din P.11-(2) rezultă

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

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

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

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

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

3) Să se arate că: $X \sim N(\mu, \sigma^2) \iff \frac{X-\mu}{\sigma} \sim N(0, 1)$.

R.: Funcția de densitate pentru distribuția $N(\mu, \sigma^2)$ este

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

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

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

Notăm cu $Y = \frac{X-\mu}{\sigma}$.

” \Rightarrow ” Pentru orice $y \in \mathbb{R}$ are loc:

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

Prin derivare în raport cu y se obține

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

care este funcția de densitate pentru $N(0, 1)$. Deci $Y \sim N(0, 1)$.

” \Leftarrow ” Pentru orice $x \in \mathbb{R}$ are loc:

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

Prin derivare în raport cu x se obține

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

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

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

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

unde $F_{(X,Y)}$ este funcția de repartiție a vectorului aleator (X, Y) .

Proprietățile funcției de densitate din P.11 sunt generalizate pentru funcția de densitate a unui vector aleator continuu.

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

1. $f_{(X,Y)}(s, t) \geq 0 \quad \forall s, t \in \mathbb{R}.$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{(X,Y)}(u, v) du dv = 1.$

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

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

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

$$5. P((X, Y) \in M) = \underbrace{\int \int_M}_{M} f_{(X,Y)}(u, v) du dv, \quad M \subset \mathbb{R}^2.$$

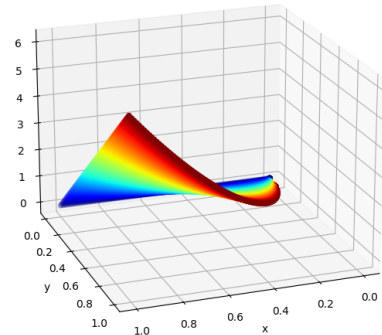
Exemplul 1: Fie $f_{(X,Y)} : \mathbb{R}^2 \rightarrow [0, \infty)$ densitatea de probabilitate a vectorului aleator (X, Y) definită prin

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

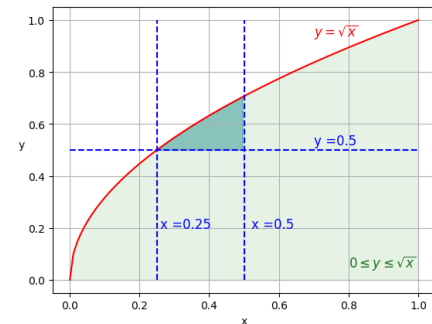
Să se determine $P(X \leq 0.5 \leq Y)$.

R.: Folosim P. 14 pentru a calcula

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



Exemplul 1



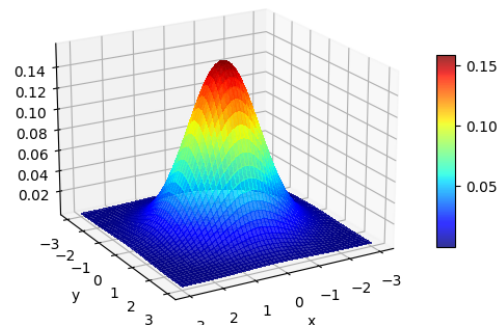
► Dacă se cunoaște funcția de densitate $f_{(X,Y)}$ pentru vectorul aleator continuu (X, Y) , atunci f_X , respectiv f_Y , se determină cu

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

f_X și f_Y se numesc **funcții de densitate marginale**.

Exemplul 2: Distribuția normală bidimensională standard: (X, Y) are funcția de densitate

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



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

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

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



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

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

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


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

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

X și Y v.a. independente.


R.: Cu (4) se calculează

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

$\implies X \sim Unif[a_1, b_1], Y \sim Unif[a_2, b_2]$ (a se vedea distribuția uniformă pe un interval, pg. 37). Se observă $f_{(X,Y)} = f_X \cdot f_Y \implies X$ și Y sunt v.a. independente! 

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

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

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

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

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

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

b) Se aruncă o monedă

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

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

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

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

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

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

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

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

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

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

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

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

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

R.:

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

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

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

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

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

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

R.: $(X_n)_n$ nu converge a.s. spre o v.a.; șirul $(X_n(\omega))_n$ este convergent doar pentru $\omega \in \{0, 1\}$, iar $P(\{0, 1\}) = 0$. ▲

Legea tare a numerelor mari (LTNM)

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



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

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

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

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

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

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

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

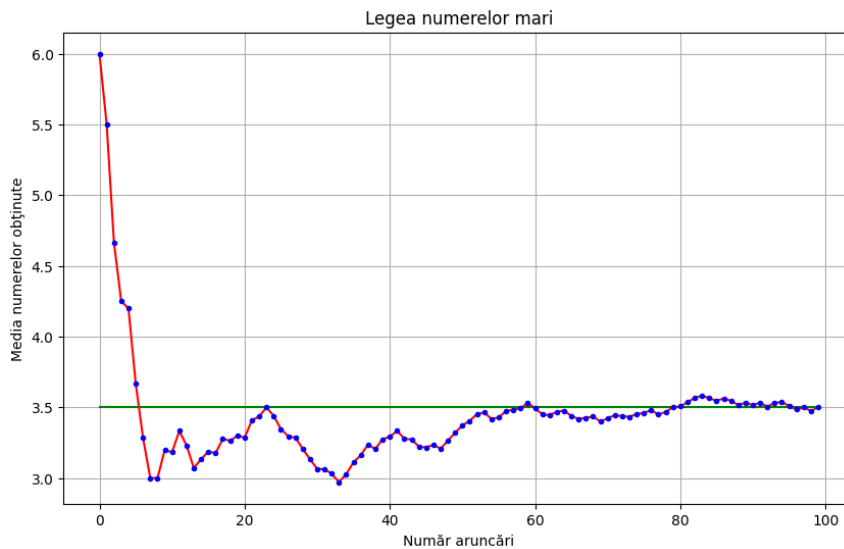


Fig. 4. Simulare LTNM

```

#LTNM X_n - nr aparut la aruncarea n
import matplotlib.pyplot as plt
from scipy.stats import randint
import numpy as np
N=400 #de cate ori aruncam zarul
X=randint.rvs(1,7,size=N)
sume = np.cumsum(X)
# sume cumulative: X[0], X[0]+X[1], ... , X[0]+X[1]+...+X[N-1]
S=[]
for n in range(N):
    S.append(sume[n]/(n+1))
    # se aduga in S valoarea (X[0]+X[1]+...+X[n]) /(n+1)
print("Valoare medie estimata prin simulari",np.mean(X))
# egala cu sume[N-1]/N
v=[1,2,3,4,5,6]
e=np.mean(v)
print("Valoarea medie teoretica ",f"{e:3.2f}")
t=[i for i in range(0,N)]
plt.figure(figsize=(8,6))
plt.plot(t,e*np.ones(N),"g-")
plt.plot(t,S,"r-")
plt.plot(t,S,"b.")
plt.xlabel("Numar de aruncari ale zarului")
plt.ylabel("Valoarea medie a numerelor obtinute ")
plt.title("Legea numerelor mari")
plt.grid()
plt.show()

```

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

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

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

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

▲

Exemplul 3: Fie $(X_n)_n$ șir de v.a. independente, având aceeași distribuție ca v.a. X și varianță finită: $E(X_n) = E(X) \in \mathbb{R}$, $V(X_n) = V(X) \in \mathbb{R}$ pentru fiecare $n \in \mathbb{N}^*$.

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

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

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

adică

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Statistică matematică

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

- statistica descriptivă: metode de colectare, organizare, sintetizare, prezentare și descriere a datelor numerice (sau nenumerice) într-o formă convenabilă
- statistica inferențială: metode de interpretare a rezultatelor obținute prin metodele statisticii descriptive, utilizate apoi pentru luarea deciziilor.

► O *colectivitate* sau *populație statistică* \mathcal{C} este o mulțime de elemente care au anumite însușiri comune ce fac obiectul analizei statistice. Numărul elementelor populației se numește *volumul populației*.

Exemple de populații statistice: mulțimea persoanelor dintr-o anumită țară, localitate, zonă etc. într-un anumit an; mulțimea gospodăriilor din România la un moment dat; mulțimea consumatorilor unui anumit produs; mulțimea societăților care produc un anumit produs; angajații unei societăți; studenții unei facultăți.

► *Eșantionul* \mathcal{E} reprezintă o submulțime a unei populații statistice $\mathcal{E} \subset \mathcal{C}$, constituită după criterii bine stabilite:

- a) să fie aleatoare;
- b) toate elementele colectivității să aibă aceeași șansă de a fi alese în eşantion;
- c) eşantionul să fie reprezentativ (structura eşantionului să fie apropiată de structura populației);
- d) volumul eşantionului să fie suficient de mare.

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

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

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

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

- a) variabile (caracteristici) continue \rightarrow iau un număr infinit și nenumărabil de valori într-un interval sau reuniune de intervale (de ex.: greutatea, înălțimea, valoarea glicemiei, temperatura aerului)
- b) variabile (caracteristici) discrete \rightarrow iau număr finit sau infinit dar numărabil de valori dis-

crete (de ex.: numărul elevi ai unei școli, numărul liceelor existente într-un oraș, valoarea IQ)
 ▷ caracteristicile de la a) și b) sunt variabile numerice (cantitative)
 c) variabile (caracteristici) nominale (de ex.: culoarea ochilor, ramura de activitate, religia)
 d) variabile (caracteristici) nominale ordinale (de ex.: starea de sănătate / calitatea unor servicii - precară, mai bună, bună, foarte bună)
 e) variabile (caracteristici) dihotomiale (binare) (de ex.: stagiul militar - satisfăcut/nesatisfăcut, starea civilă - căsătorit/necăsătorit)

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

▷ variabilele nominale mai sunt numite variabile categoriale

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

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

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

Distribuția caracteristicii X de poate fi

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

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

3) necunoscută: $X \sim ?$

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

↪ se estimează, folosind teoria estimației și intervalele de încredere

↪ se testează, folosind testele statistice

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

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

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

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

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

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

Fie X_1, \dots, X_n variabilele de selecție corespunzătoare caracteristicii cercetate X , a cărei distribuție depinde de parametrul necunoscut θ .

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

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

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

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

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

▷ Un estimator nedeplasat furnizează estimări corecte ale parametrului *în medie* pe baza mai multor eșantioane.

▷ Estimatorul nu este consistent, dacă acesta nu converge către valoarea reală a parametrului, chiar dacă există un număr mare de date statistice.

• Exemple de estimatori (funcții de selecție)

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

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

► media de selecție (empirică)

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

► valoarea mediei de selecție

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

▷ Media de selecție \bar{X}_n este un *estimator nedeplasat și consistent pentru media teoretică* $E(X)$ a caracteristicii X ; se folosesc simulări pentru $E(X) \approx \bar{x}_n$, `numpy.mean`

► varianța (dispersia) de selecție (empirică)

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

► valoarea varianței (dispersiei) de selecție

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

▷ Variația de selecție S_n^2 este un **estimator nedeplasat și consistent pentru variația teoretică** $V(X)$ a caracteristicii X ; se folosesc simulări pentru $V(X) \approx s_n^2$,
`numpy.var(..., ddof=1).`

► **abaterea standard de selecție (empirică)**

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

► valoarea abaterii standard de selecție

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

▷ Deviația standard de selecție S_n **nu** este un estimator nedeplasat pentru deviația standard teoretică $Std(X) = \sqrt{V(X)}$ a caracteristicii X ; el este un **estimator consistent pentru deviația standard teoretică** $Std(X)$ a caracteristicii X ; în simulări se folosește $Std(X) \approx s_n$,
`numpy.std(..., ddof=1).`

► **momentul centrat de selecție (empiric) de ordinul doi**

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

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

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

▷ Momentul centrat de selecție de ordinul doi M_n *nu este un estimator nedeplasat* pentru variația teoretică $V(X)$ a caracteristicii X ; el este un **estimator consistent pentru variația teoretică** $V(X)$ a caracteristicii X ; se folosesc simulări pentru $V(X) \approx m_n$,
`numpy.var(..., ddof=0).`

Exemplul 1: Fie $(X_n)_n$ șirul variabilelor de selecție pentru caracteristica cercetată $X \sim \text{Bernoulli}(p)$, unde $p \in (0, 1)$ este parametru necunoscut.

(a) Estimatorul

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

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

Se dau datele statistice $x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 0, x_5 = 0, x_6 = 1, x_7 = 1, x_8 = 0$. Să se calculeze valoarea estimatorului \hat{p} .

(b) Fie $n > 2$. Considerăm estimatorul $\bar{p}(X_1, \dots, X_n) = \frac{1}{2}(X_1 + X_n)$. Este estimatorul \hat{p} mai eficient decât estimatorul \bar{p} ?

R.: (a) $X \sim \text{Bernoulli}(p) \implies E(X) = p, V(X) = p(1 - p)$;

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

LTNM (a se vedea P.16) implică

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

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

Folosind datele statistice $x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 0, x_5 = 0, x_6 = 1, x_7 = 1, x_8 = 0$, valoarea estimată pentru p este

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

(b) Observăm $E(\bar{p}) = p; V(\bar{p}) = \frac{p(1-p)}{2} > V(\hat{p}) = \frac{p(1-p)}{n}$. Deci estimatorul \hat{p} este *mai eficient* decât estimatorul \bar{p} . \diamond

Exemplul 2: (a) Varianța empirică $S_n^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X}_n)^2$ este un estimator *nedeplasat* și *consistent* pentru varianța teoretică $\sigma^2 = V(X)$ a caracteristicii X .

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

$$\begin{aligned} (n-1)S_n^2 &= \sum_{k=1}^n (X_k - \bar{X}_n)^2 = \sum_{k=1}^n (X_k^2 - 2X_k\bar{X}_n + \bar{X}_n^2) = \sum_{k=1}^n X_k^2 - n\bar{X}_n^2 \\ &= \sum_{k=1}^n X_k^2 - \frac{1}{n} \left(\sum_{k=1}^n X_k \right)^2 = \frac{n-1}{n} \sum_{k=1}^n X_k^2 - \frac{1}{n} \sum_{\substack{k,j=1 \\ k \neq j}}^n X_k X_j. \end{aligned}$$

X_1, \dots, X_n sunt v.a. independente și au aceeași distribuție ca X :

$$\begin{aligned} E(X_k X_j) &= E(X_k)E(X_j) = m^2 \quad \forall k, j \in \{1, \dots, n\}, k \neq j \\ E(X_k^2) &= V(X_k) + E^2(X_k) = V(X) + E^2(X) = \sigma^2 + m^2 \quad \forall k \in \{1, \dots, n\} \end{aligned}$$

Scriem succesiv

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

$$\implies S_n^2 \text{ este estimator nedeplasat pentru } V(X).$$

Mai sus s-a demonstrat că

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

LTNM (a se vedea P.16) implică

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

și

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

În concluzie $S_n^2 \xrightarrow{a.s.} \sigma^2$, deci S_n^2 este estimator consistent pentru $\sigma^2 = V(X)$.

(b) Deoarece $M_n = \frac{n-1}{n} S_n^2$, obținem că $E(M_n) = \frac{n-1}{n} E(S_n^2) = \frac{n-1}{n} V(X)$, deci M_n **nu** este un estimator nedeplasat pentru $V(X)$, dar este un estimator consistent pentru $V(X)$. \diamond

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

de exemplu:

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

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

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

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

Se rezolvă sistemul

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

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

Soluția sistemului $\hat{\theta}_1, \dots, \hat{\theta}_r$ sunt valorile estimate pentru parametrii necunoscuți $\theta_1, \dots, \theta_r$ ai distribuției caracteristicii X .

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

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

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

Valoarea estimatorului este

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

Estimatorul pentru parametrul necunoscut a este

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

Parametrul necunoscut a este estimat cu valoarea 0.936.

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

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



Exemplu 2:

Folosind metoda momentelor, să se estimeze parametrii necunoscuți $\theta_1 := \mu$ și $\theta_2 = \sigma^2$ pentru $X \sim N(\mu, \sigma^2)$; se dau datele statistice:

$$0.831, 0.71, -0.2, -0.04, 2.08, -1.2, 0.448, -0.18, -0.27, -0.55.$$

R.: Fie $n = 10$, x_1, \dots, x_n sunt datele statistice, iar X_1, \dots, X_n sunt variabile de selecție. Avem cazul: $r = 2$, calculăm $E(X) = \mu$, $E(X^2) = V(X) + E^2(X) = \sigma^2 + \mu^2$ (a se vedea exemplul de pe p. 44). Se rezolvă

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

Valorile estimatorilor sunt

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

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

Estimatorii sunt

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

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



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

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

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

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

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

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

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

Deseori este mai practic să se considere varianta transformată

$\frac{\partial \ln L}{\partial \theta} = 0$ cu $\frac{\partial^2 \ln L}{\partial \theta^2} < 0$. În unele situații (1) se rezolvă prin alte metode; de exemplu în cazul în care $\frac{\partial L}{\partial \theta} = 0$ nu are soluție (echivalent cu $\frac{\partial \ln L}{\partial \theta} = 0$ nu are soluție). **Reamintire:** dacă $a, b > 0$, atunci au loc proprietățile:

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

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

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

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

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

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

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

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

Estimatorul de verosimilitate maximă pentru parametrul necunoscut p este

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

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

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

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

Observație:

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

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

Se poate lucra și cu varianta transformată:

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

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

Intervale de încredere și teste statistice

Noțiuni de bază

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

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

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

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

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

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

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

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

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

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

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

▷ distribuția normală $N(0, 1)$

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

cuantila $z_\alpha = \text{norm.ppf}(\alpha, 0, 1)$, adică $F_{N(0,1)}(z_\alpha) = \alpha$;

▷ distribuția Student $T(n)$

funcția de repartiție $F_{T(n)}(x) = \text{t.cdf}(x, n)$;

cuantila $t_\alpha = \text{t.ppf}(\alpha, n)$, adică $F_{T(n)}(t_\alpha) = \alpha$;

▷ distribuția Chi-pătrat $\chi^2(n)$

funcția de repartiție $F_{\chi^2(n)}(x) = \text{chi2.cdf}(x, n)$;

cuantila $c_\alpha = \text{chi2.ppf}(\alpha, n)$, adică $F_{\chi^2(n)}(c_\alpha) = \alpha$;

Exemple: $\text{norm.ppf}(0.01, 0, 1) = -2.3263$, $\text{norm.ppf}(1 - 0.01, 0, 1) = 2.3263$,

$\text{t.ppf}(0.05, 10) = -1.8125$, $\text{t.ppf}(1 - 0.05, 10) = 1.8125$,

$\text{chi2.ppf}(0.05, 10) = 3.9403$, $\text{chi2.ppf}(1 - 0.05, 10) = 18.307$.

ppf - percent point function (inverse of cdf)

```
#Exemplu - Cuantile
from scipy.stats import norm, t, chi2
```

```

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

```

P. 18.

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

Intervale de încredere

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

Problemă: Putem construi un interval (aleator) care să acopere valoarea reală a parametrului necunoscut studiat cu o anumită probabilitate dată (numită nivel de încredere)?

Pe baza datelor din eșantion acest interval aleator va deveni un interval numeric.

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

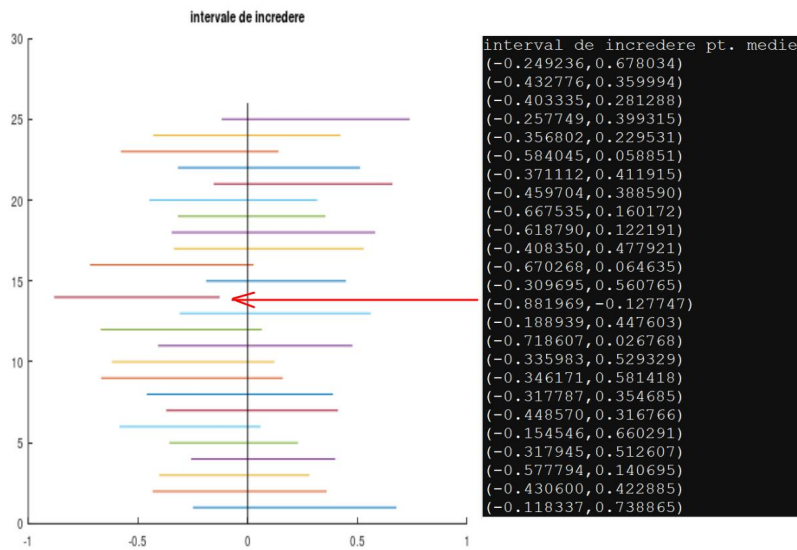
Se caută doi estimatori $g_1(X_1, \dots, X_n)$ și $g_2(X_1, \dots, X_n)$ astfel încât

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

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

► $\left(g_1(x_1, \dots, x_n), g_2(x_1, \dots, x_n)\right)$ este **valoarea intervalului de încredere** pentru parametrul necunoscut θ

► $g_1(X_1, \dots, X_n)$ este limita inferioară a intervalului de încredere, valoarea sa este $g_1(x_1, \dots, x_n)$



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

- $g_2(X_1, \dots, X_n)$ este limita superioară a intervalului de încredere, valoarea sa este $g_2(x_1, \dots, x_n)$
- probabilitatea ca parametrul necunoscut θ să fie în intervalul $(g_1(X_1, \dots, X_n), g_2(X_1, \dots, X_n))$ este $1 - \alpha$ (nivelul de încredere)
- există și **intervale de încredere unilaterale**: $(-\infty, g_3(X_1, \dots, X_n))$ **interval de încredere unilateral stâng**, $(g_4(X_1, \dots, X_n), \infty)$ **interval de încredere unilateral drept**, la care estimatorii g_3 și g_4 sunt construiți astfel încât

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

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

➡ **Nu** este corect să afirmăm că ”probabilitatea ca intervalul numeric construit din datele statistice să cuprindă valoarea reală a parametrului necunoscut θ este $1 - \alpha$ ”. Intervalul de încredere este un *interval aleator*, deci extremitățile sale sunt v.a. Prin urmare interpretarea corectă a lui $1 - \alpha$ este următoarea: dacă, facem un număr foarte mare de selecții (din mai multe eșantioane) și calculăm de fiecare dată intervalul de încredere cu nivelul de încredere $1 - \alpha$, atunci $(1 - \alpha) \cdot 100\%$ din aceste intervale conțin valoarea reală a parametrului θ . În exemplul de mai sus un interval de încredere, din cele 25 construite, nu conține valoarea reală 0.

Recapitulare (notații)

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

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

Reamintim: $X \sim N(\mu, \sigma^2) \implies E(X) = \mu, V(X) = \sigma^2$ (a se vedea calculele de pe pg. 44).

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

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

▷ TLC afirmă: distribuția versiunii standardizate² a mediei de selecție converge către distribuția normală standard $N(0, 1)$, chiar dacă variabilele de selecție (caracteristica cercetată X) nu urmează o distribuție normală

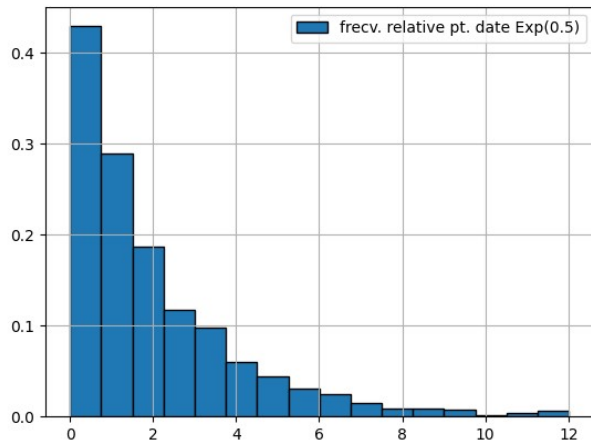
▷ echivalent putem reformula TLC astfel: media de selecție $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$ este aproximativ normal distribuită cu media μ și varianța $\frac{\sigma^2}{n}$, adică $\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n})$ (a se vedea Exemplul 3 de pg. 44).

➡ **Consecință** (la P. 20): pentru orice $a < b$ are loc

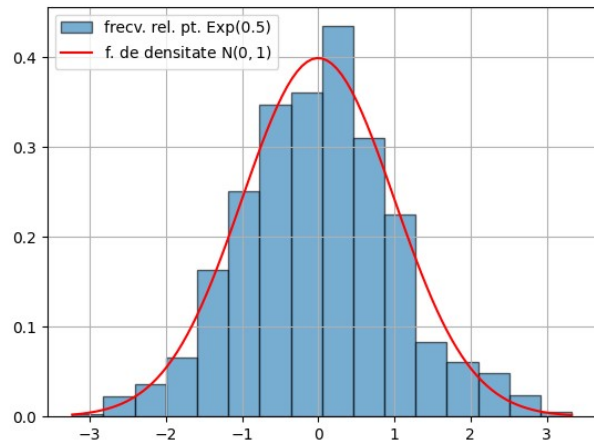
$$P\left(a < \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < b\right) \approx F_{N(0,1)}(b) - F_{N(0,1)}(a) = \text{norm.cdf}(b, 0, 1) - \text{norm.cdf}(a, 0, 1)$$

pentru n suficient de mare ($n > 30$).

²Dacă Z este o v.a. (cu $V(Z) > 0$) atunci v.a. $\tilde{Z} = \frac{Z - E(Z)}{\text{Std}(Z)}$ este **versiunea standardizată** a v.a. Z ; astfel $E(\tilde{Z}) = 0$ și $V(\tilde{Z}) = 1$.



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



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

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

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

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

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

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

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

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

Observație: $(0.35, 0.65)$ *nu* este valoarea unui interval de încredere!



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

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

- se dau $\alpha \in (0, 1)$, σ , datele statistice x_1, \dots, x_n
- fie X_1, \dots, X_n variabilele de selecție corespunzătoare caracteristicii cercetate X
- construim intervale de încredere pentru parametrul *necunoscut* $\mu = E(X)$
- dacă $X \sim N(\mu, \sigma^2)$ sau $n > 30$ și X are o distribuție necunoscută, atunci P20, respectiv P.19, implică

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

- cuantilele legii normale $N(0, 1)$:

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

- un *interval de încredere bilateral* pentru $\mu = E(X)$ (media teoretică) când dispersia este cunoscută este

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

deoarece:

$$\begin{aligned} P\left(\bar{X}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} < \mu < \bar{X}_n + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\right) &= P\left(-z_{1-\frac{\alpha}{2}} < \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}}\right) \\ (5) \quad &\stackrel{P.18}{=} F_{N(0,1)}(z_{1-\frac{\alpha}{2}}) - F_{N(0,1)}(-z_{1-\frac{\alpha}{2}}) \stackrel{P.18}{=} F_{N(0,1)}(z_{1-\frac{\alpha}{2}}) - F_{N(0,1)}(z_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha \end{aligned}$$

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

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

Interval de încredere pentru media $E(X)$ când varianța $\sigma^2 = V(X)$ este cunoscută:	Expresia intervalului de încredere, folosind datele statistice
bilateral	$\left(\bar{x}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} \right)$
unilateral	$\left(-\infty, \bar{x}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{\alpha} \right)$ $\left(\bar{x}_n - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}, \infty \right)$

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

R:

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

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

Exercițiu: Cum se modifică intervalul de încredere bilateral introdus pentru media teoretică, dacă:

- (a) crește deviația standard σ ?
- (b) crește nivelul de încredere $1 - \alpha$?
- (c) crește volumul n al eșantionului?

P. 21. Fie X_1, \dots, X_n variabile de selecție pentru caracteristica X (cu $X \sim N(\mu, \sigma^2)$ sau $n > 30$), atunci pentru **media de selecție și abaterea standard de selecție** are loc

$$\frac{\bar{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} \sim T(n-1).$$

Interval de încredere pentru media $\mu = E(X)$ a caracteristicii cercetate X , când varianța $V(X)$ este necunoscută

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

- se dau $\alpha \in (0, 1)$, datele statistice x_1, \dots, x_n
- fie X_1, \dots, X_n variabilele de selecție corespunzătoare caracteristicii cercetate X
- construim intervale de încredere pentru parametrul *necunoscut* $\mu = E(X)$
- dacă $X \sim N(\mu, \sigma^2)$ sau $n > 30$ și X are o distribuție necunoscută, atunci P.21 implică

$$(6) \quad \frac{\bar{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} \sim T(n-1)$$

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

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

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

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

$$\stackrel{(6)}{=} F_{T(n-1)}(t_{1-\frac{\alpha}{2}}) - F_{T(n-1)}(-t_{1-\frac{\alpha}{2}}) \stackrel{P.18}{=} F_{T(n-1)}(t_{1-\frac{\alpha}{2}}) - F_{T(n-1)}(t_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha$$

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

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

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

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

R.: valoarea intervalului de încredere bilateral pentru media teoretică μ , când varianța este

necunoscută, este

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

unde $\bar{x}_n = 15.5, s_n = 0.3$ ($s_n^2 = 0.09$), $\alpha = 0.01, t_{1-\frac{\alpha}{2}} = t.ppf(0.995, 99) = 2.6264$, $\sqrt{n} = 10$. Valoarea intervalului de încredere bilateral este $(15.421208, 15.578792)$. ♣

P. 22. Fie X_1, \dots, X_n variabile de selecție pentru $X \sim N(\mu, \sigma^2)$, atunci pentru **varianța de selecție** are loc $\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2(n-1)$, unde $S_n^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X}_n)^2$.

Interval de încredere pentru varianța (dispersia) teoretică $\sigma^2 = V(X)$ a caracteristicii cercetate X

Exemplu: Media de selecție a lungimii a 100 de șuruburi produse de o anumita firmă este 15.5 cm, iar varianța de selecție este 0.09 cm². Să se construiască un interval de încredere 99% bilateral pentru varianța (teoretică) a lungimii șuruburilor. Dacă varianța este prea mare (adică peste 0.099 cm²), aparatul, care produce șuruburile, trebuie reglat. Se presupune că lungimea unui șurub (produs de această firmă) are o distribuție normală.

- se dau $\alpha \in (0, 1)$, datele statistice x_1, \dots, x_n
- fie X_1, \dots, X_n variabilele de selecție corespunzătoare caracteristicii cercetate X
- construim intervale de încredere pentru parametrul *necunoscut* $\sigma^2 = V(X)$
- dacă $X \sim N(\mu, \sigma^2)$, atunci P.22 implică $\frac{n-1}{\sigma^2} S_n^2 \sim \chi^2(n-1)$
- cuantilele distribuției $\chi^2(n-1)$ (Chi-pătrat cu $n-1$ grade de libertate):
 $c_{1-\frac{\alpha}{2}} = \text{chi2.ppf}(1 - \frac{\alpha}{2}, n-1)$, $c_{\frac{\alpha}{2}} = \text{chi2.ppf}(\frac{\alpha}{2}, n-1)$,
 $c_{1-\alpha} = \text{chi2.ppf}(1 - \alpha, n-1)$, $c_{\alpha} = \text{chi2.ppf}(\alpha, n-1)$

- un *interval de încredere bilateral* pentru $\sigma^2 = V(X)$ (varianța teoretică) este:
 $\left(\frac{n-1}{c_{1-\frac{\alpha}{2}}} \cdot S_n^2, \frac{n-1}{c_{\frac{\alpha}{2}}} \cdot S_n^2 \right)$, adică

$$P\left(\frac{n-1}{c_{1-\frac{\alpha}{2}}} \cdot S_n^2 < \sigma^2 < \frac{n-1}{c_{\frac{\alpha}{2}}} \cdot S_n^2 \right) = 1 - \alpha$$

- *intervale de încredere unilaterale*: $\left(0, \frac{n-1}{c_{\alpha}} \cdot S_n^2 \right)$, $\left(\frac{n-1}{c_{1-\alpha}} \cdot S_n^2, \infty \right)$, adică

$$P\left(\sigma^2 < \frac{n-1}{c_{\alpha}} \cdot S_n^2 \right) = 1 - \alpha, \quad P\left(\frac{n-1}{c_{1-\alpha}} \cdot S_n^2 < \sigma^2 \right) = 1 - \alpha.$$

Interval de încredere pentru varianța (dispersia) $V(X)$	Expresia intervalului de încredere, folosind datele statistice
bilateral	$\left(\frac{n-1}{c_{1-\frac{\alpha}{2}}} \cdot s_n^2, \frac{n-1}{c_{\frac{\alpha}{2}}} \cdot s_n^2 \right)$
unilateral	$\left(0, \frac{n-1}{c_\alpha} \cdot s_n^2 \right)$ $\left(\frac{n-1}{c_{1-\alpha}} \cdot s_n^2, \infty \right)$

Interval de încredere pentru abaterea standard $Std(X)$	Expresia intervalului de încredere, folosind datele statistice
bilateral	$\left(\sqrt{\frac{n-1}{c_{1-\frac{\alpha}{2}}}} \cdot s_n, \sqrt{\frac{n-1}{c_{\frac{\alpha}{2}}}} \cdot s_n \right)$
unilateral	$\left(0, \sqrt{\frac{n-1}{c_\alpha}} \cdot s_n \right)$ $\left(\sqrt{\frac{n-1}{c_{1-\alpha}}} \cdot s_n, \infty \right)$

Exemplul 1: Media de selecție a lungimii a 100 de șuruburi produse de o anumita firmă este 15.5 cm, iar varianța de selecție este 0.09 cm². Să se construiască un interval de încredere 99% bilateral pentru varianța lungimii șuruburilor. Dacă varianța este prea mare (adică peste 0.099 cm²), aparatul, care produce șuruburile, trebuie reglat. Se presupune că lungimea unui șurub (produs de această firmă) are o distribuție normală.

R.: valoarea intervalului de încredere bilateral pentru varianța teoretică este

$$\left(\frac{n-1}{c_{1-\frac{\alpha}{2}}} \cdot s_n^2, \frac{n-1}{c_{\frac{\alpha}{2}}} \cdot s_n^2 \right)$$

unde $\bar{x}_n = 15.5$, $s_n^2 = 0.09$, $\alpha = 0.01$, $c_{1-\frac{\alpha}{2}} = \text{chi2.ppf}(0.995, 99) = 138.99$,
 $c_{\frac{\alpha}{2}} = \text{chi2.ppf}(0.005, 99) = 66.510$. Valoarea intervalului de încredere bilateral este
(0.064107, 0.133965). Acest interval conține și valori peste 0.099, deci aparatul, care produce
șuruburile, trebuie reglat! ♣

Exemplul 2: Durata de funcționare a unui anumit tip de baterie este 500 de ore. Pe baza unui
eșantion s-au testat 64 de baterii și s-a obținut media de 525 de ore și abaterea standard de 25
de ore. Să se construiască un interval de încredere 99%

a) bilateral pentru media (teoretică);

b) unilateral pentru abaterea standard teoretică (care are marginea inferioară 0 și se cere să se
calculeze marginea superioară)

a duratei de funcționare a acestui tip de baterii (se presupune că durata de funcționare a acestui tip de baterie urmează distribuția normală).

Interval de încredere pentru proporția necunoscută p , a caracteristicii cercetate $X \sim \text{Bernoulli}(p)$

Exemplu: $p \cdot 100\%$ din populația unui oraș susține un anumit candidat la alegerile viitoare, unde $p \in (0, 1)$ este parametru necunoscut. S-a ales un eșantion aleatoriu de dimensiunea 2000 și s-a determinat că 980 de persoane susțin candidatul. Construiți un interval de încredere bilateral cu nivelul de încredere 95% pentru proporția p necunoscută.

- se dau $\alpha \in (0, 1)$, datele statistice $x_1, \dots, x_n \in \{0, 1\}$
- fie X_1, \dots, X_n variabilele de selecție corespunzătoare caracteristicii cercetate X
- construim intervale de încredere pentru parametrul necunoscut $p \in (0, 1)$
- dacă $X \sim \text{Bernoulli}(p)$, atunci P. 20 implică $\frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$ pentru n suficient de mare

► cuantilele legii normale $N(0, 1)$:

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

• *interval de încredere bilateral:* se caută doi estimatori $g_1(X_1, \dots, X_n)$ și $g_2(X_1, \dots, X_n)$ astfel încât pentru proporția p să avem:

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

▷ din

$$\begin{aligned} P\left(-z_{1-\frac{\alpha}{2}} < \underbrace{\frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}}}_{\sim N(0, 1)} < z_{1-\frac{\alpha}{2}}\right) \\ = F_{N(0,1)}(z_{1-\frac{\alpha}{2}}) - F_{N(0,1)}(-z_{1-\frac{\alpha}{2}}) \stackrel{P.18}{=} F_{N(0,1)}(z_{1-\frac{\alpha}{2}}) - F_{N(0,1)}(z_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha \end{aligned}$$

avem:

$$\left| \frac{\bar{X}_n - p}{\sqrt{\frac{p(1-p)}{n}}} \right| < z_{1-\frac{\alpha}{2}} \Leftrightarrow \bar{X}_n - \sqrt{\frac{p(1-p)}{n}} \cdot z_{1-\frac{\alpha}{2}} < p < \bar{X}_n + \sqrt{\frac{p(1-p)}{n}} \cdot z_{1-\frac{\alpha}{2}}$$

▷ se observă că la acest interval capetele depind de parametrul necunoscut p ; având în vedere că \bar{X}_n este un estimator nedeplasat și consistent pentru parametrul necunoscut p în cazul

distribuției $Bernoulli(p)$ (a se vedea Exemplul de pe pagina 56), se înlocuiește p din capetele intervalului cu

$$p \approx \hat{p}(X_1, \dots, X_n) = \bar{X}_n$$

și se obține

- *intervalul de încredere bilateral pentru p :*

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

- *intervale de încredere unilaterale:* $\left(0, \bar{X}_n - \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} \cdot z_{\alpha} \right),$
 $\left(\bar{X}_n - \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} \cdot z_{1-\alpha}, 1 \right)$

Interval de încredere pentru proporția p	Expresia intervalului de încredere, folosind datele statistice
bilateral	$\left(\bar{x}_n - \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}} \cdot z_{1-\frac{\alpha}{2}}, \bar{x}_n + \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}} \cdot z_{1-\frac{\alpha}{2}} \right) \cap (0, 1)$
unilateral	$\left(0, \bar{x}_n - \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}} \cdot z_{\alpha} \right) \cap (0, 1)$ $\left(\bar{x}_n - \sqrt{\frac{\bar{x}_n(1-\bar{x}_n)}{n}} \cdot z_{1-\alpha}, 1 \right) \cap (0, 1)$

Exemplul 1: p · 100% din populația unui oraș susține un anumit candidat la alegerile viitoare, unde $p \in (0, 1)$ este parametru necunoscut. S-a ales un eșantion aleatoriu de dimensiunea 2000 și s-a determinat că 980 de persoane susțin candidatul. Construiți un interval de încredere bilateral cu nivelul de încredere 95% pentru proporția p necunoscută.

R.: Intervalul de încredere bilateral este

$$\left(\bar{x}_n - \sqrt{\frac{\bar{x}_n(1 - \bar{x}_n)}{n}} \cdot z_{1-\frac{\alpha}{2}}, \bar{x}_n + \sqrt{\frac{\bar{x}_n(1 - \bar{x}_n)}{n}} \cdot z_{1-\frac{\alpha}{2}} \right) \cap (0, 1),$$

unde $n = 2000, \alpha = 0.05, \bar{x}_n = 980/2000 = 0.49, z_{1-\frac{\alpha}{2}} = \text{norm.ppf}(1 - \frac{0.05}{2}, 0, 1) \approx 1.96$.
 Valoarea intervalului de încredere bilateral este $(0.4678, 0.51212)$. \diamond

```

# Exemple pentru calcularea unor intervale de incredere
import numpy as np
from scipy.stats import norm, t, chi2, bernoulli
n=400
mu=0
sigma=1
data = norm.rvs(mu, sigma,n) # se genereaza date
m=np.mean(data) #media de selectie
s=np.std(data,ddof=1) #deviatia standard de selectie
v=np.var(data,ddof=1) #varianta de selectie
a=0.01 # alfa=0.01 -> 99% interval de incredere
mu1=m-norm.ppf(1-0.01/2,0,1)*1/np.sqrt(n)
mu2=m+norm.ppf(1-0.01/2,0,1)*1/np.sqrt(n)
print(f"Interval de incredere pentru medie (varianta={sigma**2}):\n ({mu1:.5f},{mu2:.5f})"
      )
m1=m-t.ppf(1-0.01/2,n-1)*s/np.sqrt(n)
m2=m+t.ppf(1-0.01/2,n-1)*s/np.sqrt(n)
print(f"Interval de incredere pentru medie (varianta necunoscuta):\n ({m1:.5f},{m2:.5f})")
v1=v*(n-1)/chi2.ppf(1-a/2,n-1)
v2=v*(n-1)/chi2.ppf(a/2,n-1)
print(f"Interval de incredere pentru varianta:\n ({v1:.5f},{v2:.5f})")
d1=s*np.sqrt((n-1)/chi2.ppf(1-a/2,n-1))
d2=s*np.sqrt((n-1)/chi2.ppf(a/2,n-1))
print(f"Interval de incredere pentru deviatia standard:\n ({d1:.5f},{d2:.5f})")
d=bernoulli.rvs(p=0.4,size=n) # se genereaza date
md=np.mean(d)
p1=md - norm.ppf(1-0.01/2,0,1)* np.sqrt(md*(1-md)/n)
p2=md + norm.ppf(1-0.01/2,0,1)* np.sqrt(md*(1-md)/n)
print(f"Interval de incredere pentru proportie:\n ({p1:.5f},{p2:.5f})")

```