Probabilitate și statistică în Data Science

Partea II. Probabilitatea

Ce ne așteaptă?

1. Ce este probabilitatea
2. Permutații
3. Combinații
4. Intersecții, uniuni, complemente
5. Probabilitatea condiționată
6. Reguli de adunare și înmulțire
7. Teoreme lui Bayes

Notații

$P(A)$ – ***probabilitatea evenimentului A***

$n!$ – ***factorialul numarului n***

${}_nP_r$ – ***numarul permutatiilor din n luate cate r fara repetitie***

${}_nC_r$ – ***numarul combinatiilor din n luate cate r fara repetitie***

${}_{n+r-1}C_r$ – ***numarul combinatiilor din n luate cate r cu repetitie***

$A \cap B$ – ***intersectia evenimentelor A cu evenimentele B***

$A \cup B$ – ***uniunea evenimentelor A cu evenimentele B***

$\bar{A}$ – ***complementul evenimentului A***

$P(A \mid B)$ – ***probabilitatea conditionată a evenimentului A cand a avut loc evenimentul B***

1. Ce este probabilitatea

Noțiuni de probabilitate

- Probabilitatea – o valoarea între 0 și 1 a faptului că un eveniment va avea loc
- De exemplu faptul că va apărea coroana la aruncarea unei monede este 0,5
- Matematic se va scrie: $P(E_{coroana})=0,5$
- Actul aruncării monedei se numește proces
- La realizarea procesului de aruncare a monedei de 5 ori, probabilitatea apariției coroanei rămâne 0,5
- Fiecare proces este independent unul de altul

Experimentele și spațiul fundamental

- **Procesul de aruncare a monedei mai poate fi considerat un experiment**
- **Fiecare rezultat care se exclude reciproc se numește eveniment elementar**
- **Spațiu fundamental reprezintă totalitatea evenimentelor elementare**
- **Exemplu:**

Aruncarea unui cub reprezintă un experiment

Evenimentele elementare sunt:

$E_1=1 \quad E_2=2 \quad E_3=3 \quad E_4=4 \quad E_5=5 \quad E_6=6$

Spațiul fundamental este:

$S = \{E_1, E_2, E_3, E_4, E_5, E_6\}$

Determinarea probabilității

- **Probabilitatea se determină ca numărul evenimentelor elementare favorabile raportat la numărul de elemente ale spațiului fundamental**

- **Exemplu:**

Care este probabilitate apariție feței 6 la aruncare cubului?

Evenimentul elementar favorabil este unui singur:

$$E_6=6$$

Numărul elementelor spațiului fundamental este 6:

$$S = \{E_1 , E_2 , E_3 , E_4, E_5, E_6\}$$

Probabilitatea de apariție a evenimentului:

$$P(E_6) = 1/6$$

2. Permutații

Numărul total a permutațiilor unui set n

- **Permutația unui set de obiecte – aranjarea tuturor obiectelor setului într-o anumită ordine**
- **Permutații posibile a literelor a, b și c**

abc acb bac bca cab cba

- **Numărul total al permutațiilor unui set din n obiecte:**

$$n!$$

- **Exemplu:**

Numărul total al permutațiilor literelor a, b și c?

$$n = 3$$

$$n! = 3! = 3 \times 2 \times 1 = 6$$

Numărul permutațiilor lui n luate câte r fără repetiție

- **În permutație poate fi luat și un subset din r elemente**
- **Numărul totale al permutațiilor unui set din n elemente luate câte r fără repetiție se determina cu formula:**

$$ {}_nP_r = \frac{n!}{(n-r)!} $$

- **Exemplu:**

Câte cuvinte din 4 caractere fără repetiție pot fi formate folosind caractere alfa-numerice?

Numărul n se determina ca suma 26 litere + 10 cifre = 36

Numărul r reprezinte numărul de caractere la formarea unui cuvânt = 4

Numărul permutațiilor se va determina:

$$ {}_{36}P_4 = \frac{36!}{(36-4)!} = \frac{36 \times 35 \times 34 \times 33 \times \cancel{32 \times 31 \dots}}{\cancel{32 \times 31 \dots}} = \mathbf{1,413,720} $$

Numărul permutațiilor lui n luate câte r cu repetiție

- **Numărul totale al permutațiilor unui set din n elemente luate câte r cu repetiție va reprezenta:**

$$n^r$$

- **Exemplu**

Câte numere a câte 4 cifre pot fi formate?

Numărul n reprezintă numărul total de cifre utilizate = 10 (0, 1, 2, 3, 4, 5, 6, 7, 8, 9)

Numărul r reprezinte numărul de cifre într-un număr format = 4

Numărul permutațiilor se va determina:

$$n^r = 10^4 = 10{,}000$$

3. Combinații

Numărul combinațiilor lui n luate câte r fără repetiție

- **Combinația unui set de obiecte – aranjarea neordonată a obiectelor unui set**

- **Numărul totale al combinațiilor unui set din n elemente luate câte r fără repetiție se determina cu formula:**

$$ {}_nC_r = \frac{n!}{r!\,(n-r)!} $$

Combinații vs permutații

- **Câte permutații și câte combinații de 3 litere pot fi compuse din literele ABCDE**

1. Permutații

$$n = 5$$

$$r = 3$$

$${}_5P_3 = \frac{5!}{(5-3)!} = 5 \times 4 \times 3 = \mathbf{60}$$

ABC	ACB	BAC	BCA	CAB	CBA
ABD	ADB	BAD	BDA	DAB	DBA
ABE	AEB	BAE	BEA	EAB	EBA
ACD	ADC	CAD	CDA	DAC	DCA
ACE	AEC	CAE	CEA	EAC	ECA
ADE	AED	DAE	DEA	EAD	EDA
BCD	BDC	CBD	CDB	DBC	DCB
BCE	BEC	CBE	CEB	EBC	ECB
BDE	BED	DBE	DEB	EBD	EDB
CDE	CED	DCE	DEC	ECD	EDC

2. Combinații

$${}_nC_r = \frac{n!}{r!\,(n-r)!} = \frac{5!}{3! \cdot 2!} = \mathbf{10}$$

ABC	ACB	BAC	BCA	CAB	CBA
ABD	ADB	BAD	BDA	DAB	DBA
ABE	AEB	BAE	BEA	EAB	EBA
ACD	ADC	CAD	CDA	DAC	DCA
ACE	AEC	CAE	CEA	EAC	ECA
ADE	AED	DAE	DEA	EAD	EDA
BCD	BDC	CBD	CDB	DBC	DCB
BCE	BEC	CBE	CEB	EBC	ECB
BDE	BED	DBE	DEB	EBD	EDB
CDE	CED	DCE	DEC	ECD	EDC

Numărul combinațiilor lui n luate câte r cu repetiție

- **Numărul totale al combinațiilor unui set din n elemente luate câte r cu repetiție se determină conform formulei:**

$$_{n+r-1}C_r = \frac{(n+r-1)!}{r!\,(n-1)!}$$

- **Exemplu**

Pentru un produs conține 4 ingrediente din 10 disponibile. Câte combinații de produs pot fi realizate cu repetiția ingredientelor?

$n = 10$

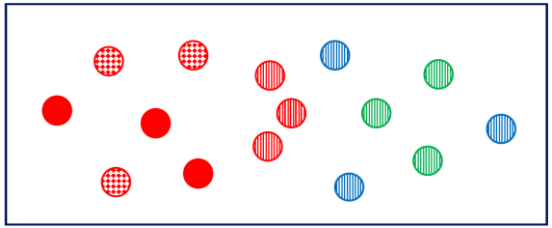
$r = 4$

$$_{n+r-1}C_r = \frac{(n+r-1)!}{r!\,(n-1)!} = \frac{13!}{4!\,(9)!} = 715$$

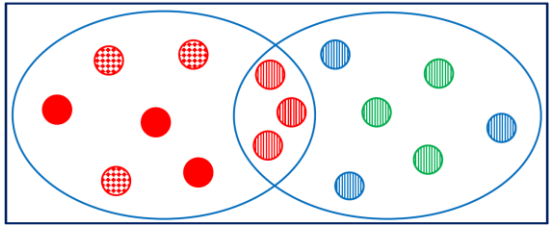
4. Intersecții, uniuni, complemente

Noțiunea intersecției

- Intersecția – descrie spațiul fundamental comun pentru 2 evenimente
- 9 bile sunt de culoare roșie și 9 sunt hașurate cu linii



- Intersecția setului de bile roșii cu setul de bile hașurate cu linii reprezintă cele 3 bile roșii hașurate cu linii



Probabilitatea intersecției

- **Dacă A reprezintă setul bilelor roșii și B setul bilelor hașurate cu linii atunci intersecția A și B se va nota:**

$$A \cap B$$

- **Ordinea intersecție nu are importanță**

$$A \cap B = B \cap A$$

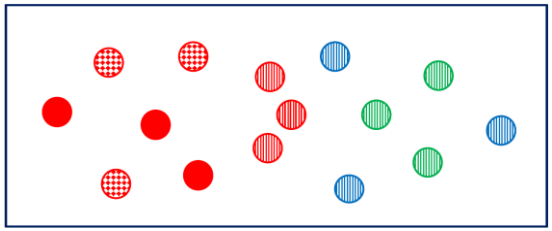
- **Probabilitatea intersecție se consideră ca raportul dintre numărul evenimentelor intersecții și numărul total al evenimentelor:**

- **În cazul bilelor:**

$$P(A \cap B) = \frac{3}{15} = \mathbf{0.2}$$

Uniunea

- **Uniunea – descrie spațiul fundamental total pentru 2 evenimente**
- **Uniunea setului de bile roșii cu setul de bile hașurate cu linii reprezintă toate cele 15 bile**



- **Dacă A reprezintă setul bilelor roșii și B setul bilelor hașurate cu linii atunci uniunea A sau B se va nota:**

$$A \cup B$$

- **Ordinea uniunii nu are importanță**

$$A \cup B = B \cup A$$

Probabilitatea uniunii

- **Probabilitatea uniunii setului A sau B se determină cu relația:**

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

unde: P(A) - probabilitatea setului A
P(B) - probabilitatea setului B
P(A∩B) - probabilitatea intersecției seturilor A și B

- **În cazul bilelor:**

$$P(A \cup B) = \frac{9}{15} + \frac{9}{15} - \frac{3}{15} = \frac{15}{15} = \mathbf{1.0}$$

Complementul

- **Complementul evenimentului – reprezintă orice ce este în afara evenimentului**
- **Complementul evenimentului A este evenimentul nu A și se notează:**

$$\overline{A}$$

- **Probabilitatea evenimentului nu A se determină:**

$$P(\overline{A}) = 1 - P(A)$$

- **În cazul bilelor:**

$$P(\overline{A}) = 1 - P(A) = \frac{15}{15} - \frac{9}{15} = \frac{6}{15} = \mathbf{0.4}$$

5. Probabilitatea condiționată

Evenimente independente

- **O serie de evenimente se consideră independente atunci când rezultatele unui eveniment nu influențează rezultatele altuia**
- **Apariția coroanei la aruncarea monedei nu depinde de rezultatul aruncării precedente**
- **Probabilitatea apariției a două evenimente independente reprezintă produsul probabilității fiecăruia dintre ele**

$$P(E_1E_2) = P(E_1) \times P(E_2)$$

- **Exemplu**

Probabilitatea apariție coroanei la aruncarea 2 două ori a monedei:

$$P(C_1C_2) = P(C_1) \times P(C_2) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

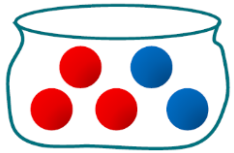
Aruncare 1	Aruncare 2
coroana	coroana
coroana	număr
număr	coroana
număr	număr

Evenimente dependente

- **Un eveniment se consideră dependente atunci când rezultatul acestuia depinde de rezultatul evenimentului anterior**
- **Un eveniment dependent este extragerea unei bile dintr-un bol cu bile colorate fără reîntoarcerea bilei în bol**
- **Exemplu**

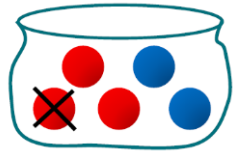
Probabilitatea selectării unei bilei roșii la prima extragere:

$$P(R_1) = \frac{3}{5}$$



Probabilitatea selectării unei bilei roșii la a doua extragere dacă la prima extragere s-a selectat o bilă roșie:

$$P(R_2|R_1) = \frac{2}{4}$$



Probabilitatea condiționată

- **Probabilitatea condiționată - este probabilitatea de apariție a evenimentului A atunci când a avut loc evenimentul B**

- **Probabilitatea condiționată se notează:**

$$P(A|B)$$

- **Probabilitatea condiționată se determină cu relația:**

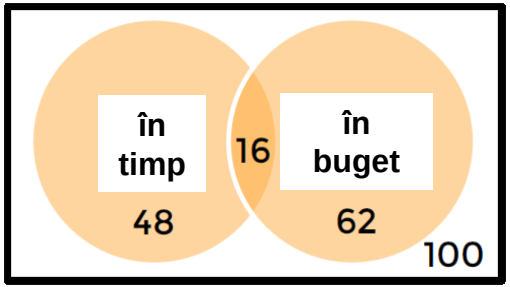
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Exemplu de probabilitate condiționată

- **Într-o companie, la fiecare 100 de proiecte, 48 sunt realizate în timp, 62 se încadrează în bugetul planificat și 16 se încadrează și în timp și în buget. Care este probabilitatea că un proiect realizat în timp se va încadra și în buget?**

A – proiectele realizate în buget

B – proiectele realizate în timp



$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{16}{48} = \mathbf{0.33}$$

6. Reguli de adunare și înmulțire

Regula de adunare

- **Regula de adunare se utilizează la determinarea probabilității unei uniuni**

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

- **Dacă evenimentele sunt reciproc exclusive (apariția unui eveniment exclude apariție celuilalt) atunci regula de adunare devine:**

$$P(A \cup B) = P(A) + P(B) \cancel{- P(A \cap B)}$$

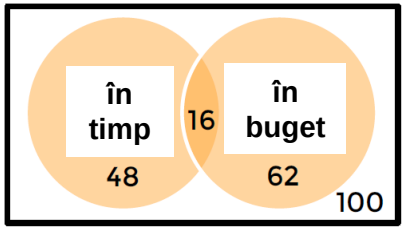
Exemple regula de adunare

- **Într-o companie, la fiecare 100 de proiecte, 48 sunt realizate în timp, 62 se încadrează în bugetul planificat și 16 se încadrează și în timp și în buget. Care este probabilitatea că un proiect se va realiza în timp sau în buget?**

  A – proiectele realizate în buget
  B – proiectele realizate în timp

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

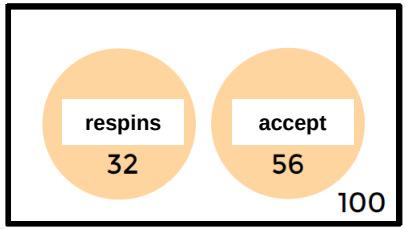
$$= \frac{48}{100} + \frac{62}{100} - \frac{16}{100} = \mathbf{0.94}$$



- **Într-o companie, la fiecare 100 de proiecte, 32 sunt respinse, 56 acceptate și restul nefinisate. Care este probabilitatea că un proiect va fi finisat?**

  A – proiectele respinse
  B – proiectele acceptate

$$P(A \cup B) = P(A) + P(B) = \frac{32}{100} + \frac{56}{100} = \mathbf{0.88}$$



Regula de înmulțire

- **Regula de înmulțire se utilizează la determinarea probabilității intersecției evenimentelor dependente**

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

- **Limita de sus a intervalului de valori admisibile se determină prin adăugarea unei valori egală cu 1,5 IQR în dreapta valorii cuartilei 3**

- **Valorile datelor care nu se încadrează în intervalul de valori admisibile se consideră valori aberante (outliers)**

Exemplu regula de înmulțire

- **Care va fi probabilitatea extragerii celor 4 ași din setul standard de 52 de cărți**

A, B, C, D – evenimentele extragerii celor 4 ași

$$P(A \cap B \cap C \cap D) = P(A) \cdot P(B|A) \cdot P(C|AB) \cdot P(D|ABC)$$

$$= \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times \frac{1}{49} = \frac{24}{6{,}497{,}400} = \frac{1}{270{,}725}$$

7. Teorema lui Bayes

Formala teoremei lui Bayes

- **Relația de calcul a probabilității condiționate este:**

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad dacă \quad P(B) > 0$$

- **Relația de calcul a probabilității intersecției evenimentelor dependente este:**

$$P(A \cap B) = P(A) \cdot P(B|A) \quad dacă \quad P(A) > 0$$

- **Substituind a doua relație în prima rezultă formula teoremei lui Bayes**

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \quad dacă \quad P(A), P(B) > 0$$

Exemplu Teorema lui Bayes (1)

- **O companie a descoperit că 0,2% din produsele sale sunt defecte și a procurat un dispozitiv de diagnosticare a produselor defecte care are precizia de 99%. Care este probabilitatea că un produs diagnosticat ca fiind defect este cu adevărat defect?**

$P(A)$ – *probabilitatea unui produs de a fi defect*

$P(B)$ – *probabilitatea unui produs de a fi diagnosticat defect*

$P(A \mid B)$ – *probabilitatea unui produs de a fi defect dacă a fost diagnosticat defect*

$P(B \mid A)$ – *probabilitatea unui produs de a fi diagnosticat defect dacă e defect*

$P(A) = 0{,}002\ (0{,}2\%)$

$P(B \mid A) = 0{,}99\ (99\%)$

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

Exemplu Teorema lui Bayes (2)

- **Probabilitatea P(B) a unui produs de a fi diagnosticat defect este egală cu suma probabilității produselor defecte diagnosticate ca fiind defecte ($P_{true\text{-}pozitiv}$) și probabilității produselor ne-defecte diagnosticate ca fiind defecte ($P_{false\text{-}pozitiv}$)**

$$P_{true-pozitiv} = P(B \mid A) \times P(A) = 0{,}99 \times 0{,}002 = 0{,}00198$$

$$P_{false-pozitiv} = P(\overline{B \mid A}) \times P(\bar{A}) = (1 - 0{,}99) \times (1 - 0{,}002) = 0{,}00998$$

$$P(B) = P_{true-pozitiv} + P_{false-pozitiv} = 0{,}00198 + 0{,}00998 = 0{,}01196$$

- **Probabilitatea că un produs diagnosticat ca fiind defect este cu adevărat defect:**

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)} = \frac{0{,}99 \times 0{,}002}{0{,}01196} = 0{,}165\ (16{,}5\%)$$