Probabilitate și statistică în Data Science

Partea I. Datele

Ce ne așteaptă?

1. Ce sunt datele
2. Măsurarea datelor
3. Măsurarea tendinței centrale
4. Măsurarea dispersiei
5. Determinarea cuartilelor
6. Date bivariate

Notații

x_i – ***valoarea datelor i din setul de date***

n – ***numarul de date ale eșantionului***

$\bar{x}$ – ***valoarea medie a esantionului***

N – ***numarul de date ale populatiei***

μ – ***valoarea medie a polulatiei***

S – ***abaterea standard a esantionului***

σ – ***abaterea standard a populatiei***

S^2 – ***dispersia esantionului***

σ^2 – ***dispersia populatiei***

$cov(X, Y)$ – ***covarianta seturilor X si Y***

$\rho_{X,Y}$ – ***coeficientul de corelatie Person a seturilor X si Y***

1. Ce sunt datele

Noțiuni de date

- **Datele – o colecție de observații referitoare la careva obiecte, fenomene sau persoane**
- **Datele pot fi continue (numerice) – prețul unui produs**
- **Datele pot fi categoriale – culoarea unui produs**

Importanța datelor

- **Datele permit observarea corelației dintre obiecte, procese, fenomene**

  Persoanele care fac sport se adresează mai rar la medic

- **Datele permit predicția comportamentelor în viitor**

  În funcție de tipul de produselor analizate pe site se presupune că va cumpăra un asemenea produs

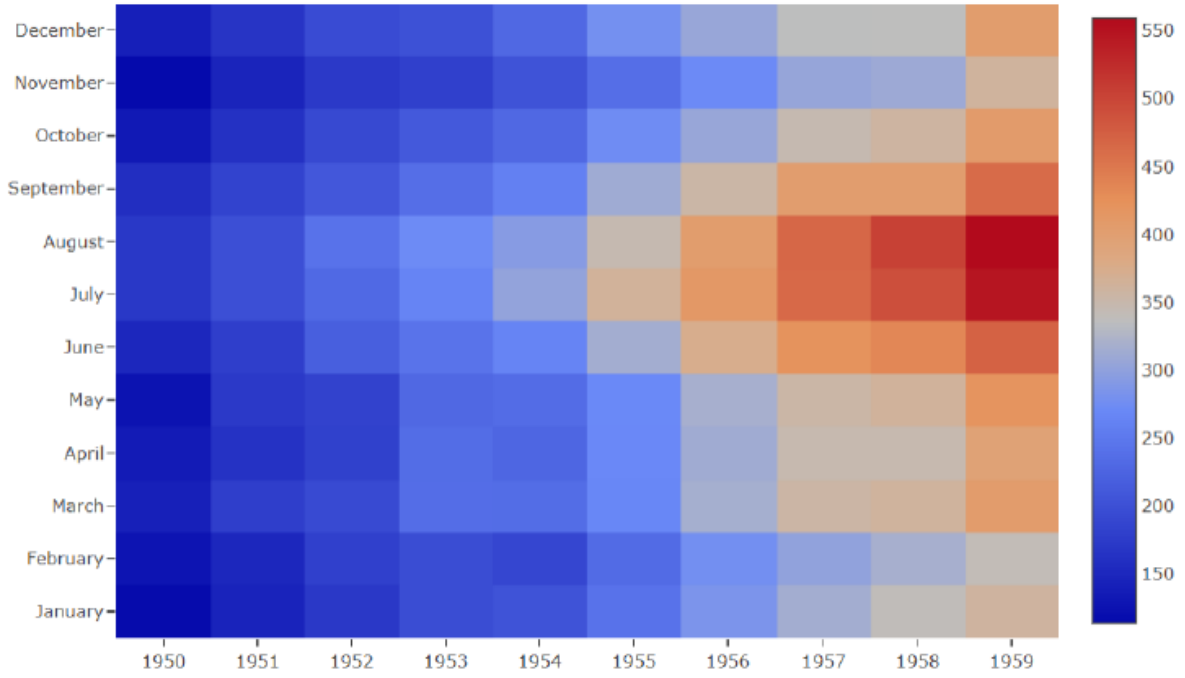
Vizualizarea datelor

- **Tabele**

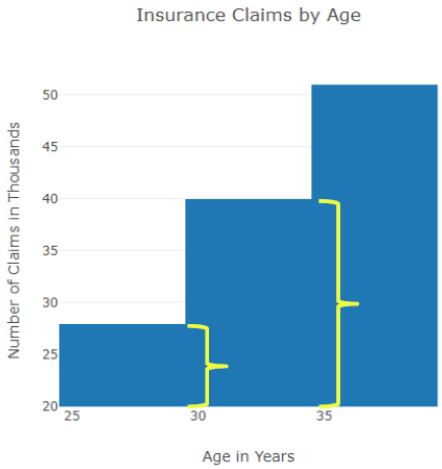
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	year	month	passengers		year	month	passengers		year	month	passengers		year	month	passengers
2	1950	January	115		1952	July	230		1955	January	242		1957	July	465
3	1950	February	126		1952	August	242		1955	February	233		1957	August	467
4	1950	March	141		1952	September	209		1955	March	267		1957	September	404
5	1950	April	135		1952	October	191		1955	April	269		1957	October	347
6	1950	May	125		1952	November	172		1955	May	270		1957	November	305
7	1950	June	149		1952	December	194		1955	June	315		1957	December	336
8	1950	July	170		1953	January	196		1955	July	364		1958	January	340
9	1950	August	170		1953	February	196		1955	August	347		1958	February	318
10	1950	September	158		1953	March	236		1955	September	312		1958	March	362
11	1950	October	133		1953	April	235		1955	October	274		1958	April	348
12	1950	November	114		1953	May	229		1955	November	237		1958	May	363
13	1950	December	140		1953	June	243		1955	December	278		1958	June	435
14	1951	January	145		1953	July	264		1956	January	284		1958	July	491
15	1951	February	150		1953	August	272		1956	February	277		1958	August	505
16	1951	March	178		1953	September	237		1956	March	317		1958	September	404
17	1951	April	163		1953	October	211		1956	April	313		1958	October	359
18	1951	May	172		1953	November	180		1956	May	318		1958	November	310

Vizualizarea datelor

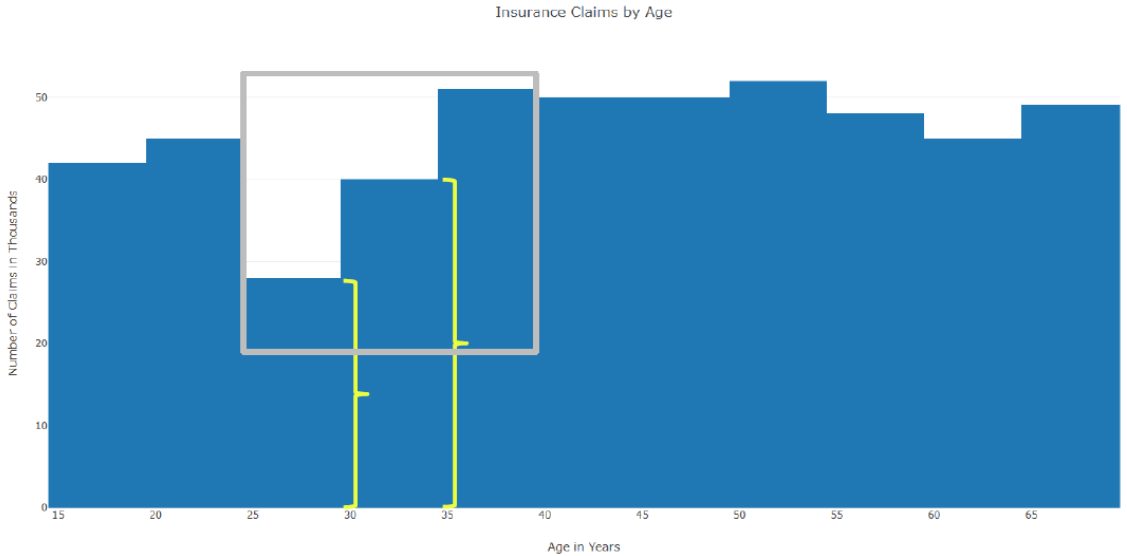
- **Grafice**



Manipularea cu vizualizarea grafică



Grafic manipulator



Grafic real

2. Măsurarea datelor

Nivelul nominal de măsurarea

- **Permite determinarea unei categorii date dintr-o listă predefinită**
- **Nu realizează sortare rezultatelor**
- **Exemple**

Determinarea clasei unui animal: mamifer, reptilă , pește, amfibien, pasăre

Determinarea culorii unui obiect: roșu, verde, albastru, galben, alb, negru

Nivelul ordinal de măsurarea

- **Permite determinarea unei categorii date**
- **Rezultatele pot fi sortate**
- **Lipsește scara de sortare**
- **Exemplu**

Cât de des faceți sport: des, câteodată, rareori, niciodată

Nivelul interval de măsurarea

- **Permite determinarea unei valori numerice**
- **Datele se citesc conform unei scări**
- **Lipsește punctul de referință “zero”**
- **Exemplu**

  Măsurarea temperaturii

Nivelul rată de măsurarea

- **Permite determinarea unei valori numerice**
- **Datele se determină în raport cu o valoarea de referință “zero”**
- **Exemplu**

Determinarea masei unui obiect

Determinarea vârstei unei persoane

Determinarea salarului unui angajat

Populație vs eșantion

- **Populația – fiecare membru a unui grup țintă**
- **Exemplu**

  Toți studenții unei universități

- **Eșantionul – un subset al populației pentru care este posibilă realizarea măsurărilor**
- **Exemplu**

  Un grup de 1000 de studenți ai universității selectat aleator

3. Măsurarea tendinței centrale

Tendința centrală vs dispersie

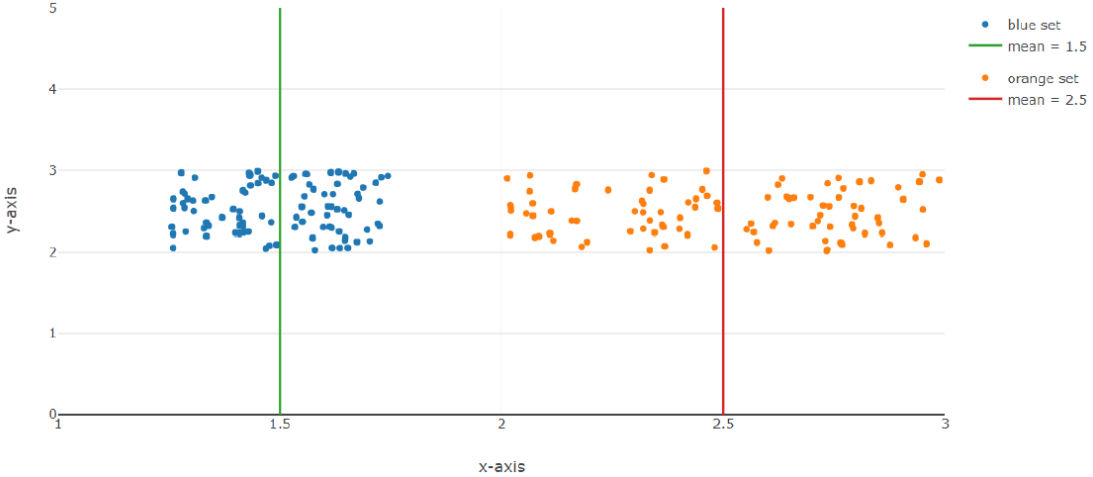
- Tendința centrală – “care este valoarea medie”
- Dispersia – “cât de departe de valoarea medie se află o anumită valoare individuală”
- Tendința centrală descrie “locația” datelor dar nu și “forma” lor
- Valori ale tendinței centrale:
  - Valoarea medie – media aritmetică a datelor
  - Valoarea mediană – valoarea aflată la mijlocul datelor
  - Valoarea modul (mod) – valoarea cea mai des întâlnită

Valoarea medie

- **Valoarea medie a eșantionului de date $\bar{x}$:**

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

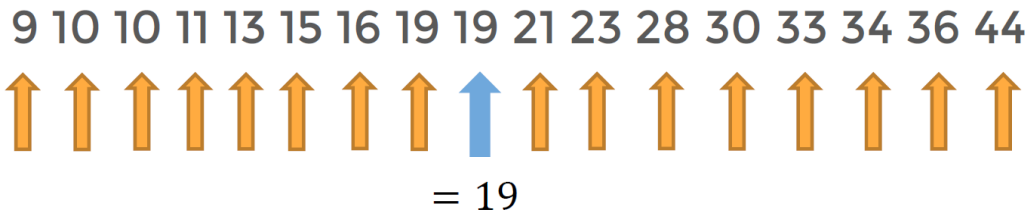
- **Valoarea medie specifică "locația" datelor nu și "împrăștierea" lor**



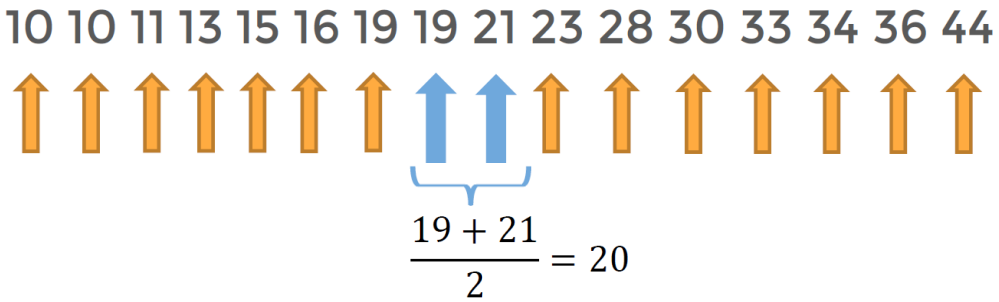
- **Valoarea medie a populație se notează cu litera μ**

Valoarea mediană

- **Valoarea mediană – valoarea situată la mijlocul setului de date după sortarea acestuia**
- **Valoarea mediană pentru un set cu număr impar de date**



- **Valoarea mediană pentru un set cu număr par de date**



Valoarea medie vs valoarea mediană

- Valoare medie este influențată de valorile aberante (outliers)

$$\{2,3,2,3,2,12\}$$

- Valoare medie a setului de mai sus este 4
- Valoare mediană a setului de mai sus este 2,5
- Valoare mediană este mai apropiată de valorile majorității datelor setului

Valoarea modul (mode)

- **Valoarea modul – valoarea cu cea mai mare frecvență de apariție în setul de date**

10 10 11 13 15 16 16 16 21 23 28 30 33 34 36 44

$= 16$

4. Măsurarea dispersiei

Gama de valori (Range)

- **Dispersia permite aprecierea modului de "împăștiere" a datelor**
- **Gama de valori reprezintă diferența dintre valoarea maximă și valoarea minimă a setului de date**

9 10 11 13 15 16 19 19 21 23 28 30 33 34 36 39

$$Range = max - min$$

$$= 39 - 9$$

$$= 30$$

Dispersia (varianța)

- **Dispersia (varianța, pătratul abaterii standard) – se determină ca suma distanțelor pătratice dintre puncte și valoarea medie raportată la numărul de date**
- **Dispersia populației se va determina considerându-se numărul total de date**

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$$

- **Dispersia eșantionului se va determina considerându-se corecția Bessel adică numărul total de date minus 1**

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

- **Dispersia se măsoară în unități pătratice unităților datelor**

Abaterea standard

- **Abaterea standard se determină ca rădăcina pătrată a dispersiei**
- **Abaterea standard a populației**

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

- **Abaterea standard a eșantionului**

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

- **Abaterea standard se măsoară în aceleași unități ca și datele**

5. Determinarea cuartilelor

Cuartilele 1, 2 și 3

- **Cuartilele – o modalitate de descriere a datelor luându-se în considera fiecare punct fără ca acesta să fie agregat**
- **Cuartila 2 va coincide cu valoarea mediană, iar cuartilele 1 și 3 vor fi valorile mediane ale subsetului de date considerate până și respectiv după cuartila 2**

cuartila 1

cuartila 2 sau mediana

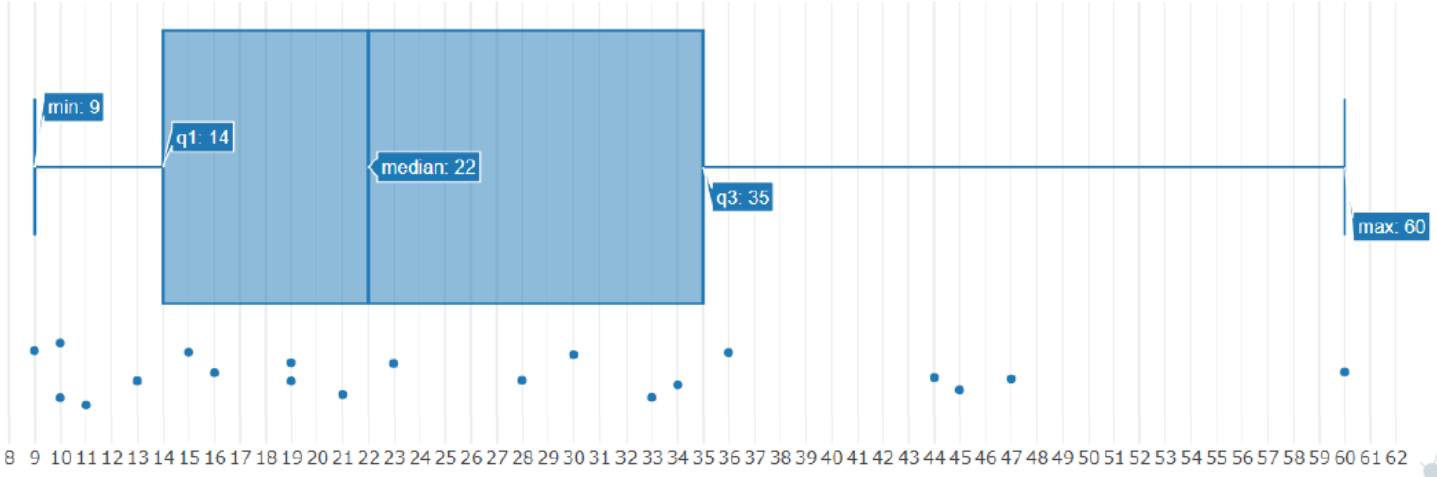
cuartila 3

cuartila 1 = 14
cuartila 2 = 22
cuartila 3 = 35

Gama inter-cuartile (InterQuartile Range - IQR)

- IQR – gama de valori dintre cuartila 1 și 3 în interiorul căreia sunt plasatee 50% dintre setul de date
- Pentru vizualizarea IQR se utilizează grafice de tip boxplot

9 10 10 11 13	15 16 19 19 21	23 28 30 33 34	36 44 45 47 60



Limitele și valorile aberante (outliers)

- IQR se utilizează pentru a detecția valorilor aberante
- Limita de jos a intervalului de valori admisibile se determină prin adăugarea unei valori egală cu 1,5 IQR în stânga valorii cuartilei 1
- 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)

6. Date bivariate

Corelația datelor

- Corelația datelor - determină gradul în care două sau mai multe variabile se mișcă în tandem
- Corelația datelor nu trebuie fi însoțită de cauzalitate
- Corelația poate avea valori in gama -1...1
- Corelația dintre variabile poate fi
  - Corelație pozitivă – ambele seturi de date se mișcă în aceeași direcți și corelația pozitivă puternică ia valoarea 1
  - Corelație negativă sau inversă – seturile de date se mișcă în sensuri opuse și corelația negativă puternică ia valoarea -1

Covarianța datelor

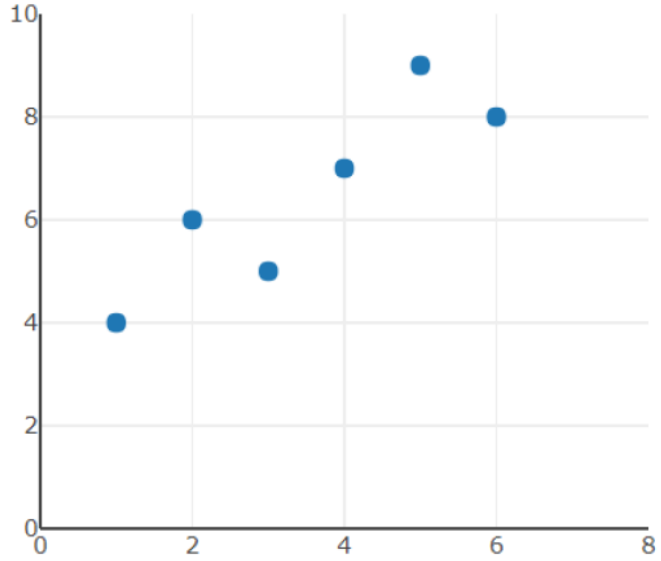
- Covarianța datelor - determină gradul modificarea a valorilor unei variabile la modificarea valorilor altei variabile
- Covarianța necesită ca datele să aibă aceeași scară
- Pentru asigurarea aceleași scări în cadrul determinării covarianței se utilizează procedura de normalizare
- Covarianța poate avea valori în gama -∞...+ ∞
- Covarianța populației se determină cu relația

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

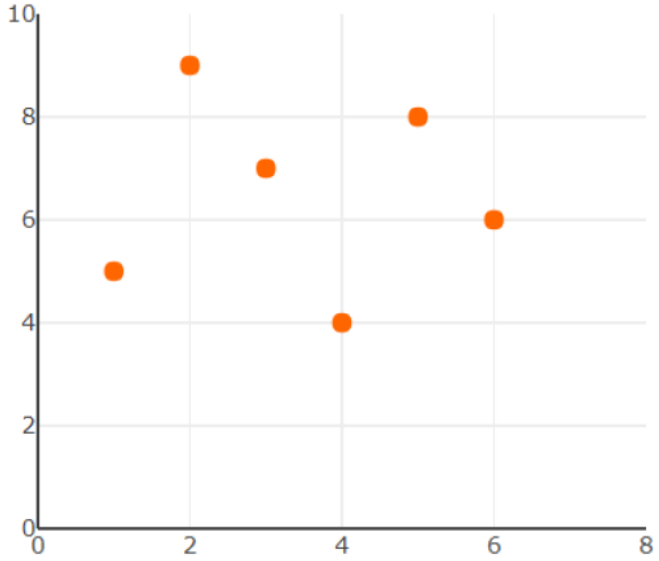
Exemplu de calcularea a covarianței (1)

- Se consideră 2 cazuri a două seturi de date x și y

x	y
1	4
2	6
3	5
4	7
5	9
6	8



x	y
1	5
2	9
3	7
4	4
5	8
6	6



Exemplu de calcularea a covarianței (2)

- **Se calculează valorile medii**

x	y
1	4
2	6
3	5
4	7
5	9
6	8

$$\bar{x} = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\bar{y} = \frac{4+6+5+7+9+8}{6} = 6.5$$

x	y
1	5
2	9
3	7
4	4
5	8
6	6

$$\bar{x} = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\bar{y} = \frac{5+9+7+4+8+6}{6} = 6.5$$

Exemplu de calcularea a covarianței (3)

- Se calculează $(x-\bar{x})$, $(y-\bar{y})$ și $(x-\bar{x})(y-\bar{y})$

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
1	4	-2.5	-2.5	6.25
2	6	-1.5	-0.5	0.75
3	5	-0.5	-1.5	0.75
4	7	0.5	0.5	0.25
5	9	1.5	2.5	3.75
6	8	2.5	1.5	3.75

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
1	5	-2.5	-1.5	3.75
2	9	-1.5	2.5	-3.75
3	7	-0.5	0.5	-0.25
4	4	0.5	-2.5	-1.25
5	8	1.5	1.5	2.25
6	6	2.5	-0.5	-1.25

Exemplu de calcularea a covarianței (4)

- **Se calculează sumele**

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$
1	4	-2.5	-2.5	6.25
2	6	-1.5	-0.5	0.75
3	5	-0.5	-1.5	0.75
4	7	0.5	0.5	0.25
5	9	1.5	2.5	3.75
6	8	2.5	1.5	3.75
			Σ	**15.5**

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$
1	5	-2.5	-1.5	3.75
2	9	-1.5	2.5	-3.75
3	7	-0.5	0.5	-0.25
4	4	0.5	-2.5	-1.25
5	8	1.5	1.5	2.25
6	6	2.5	-0.5	-1.25
			Σ	**-0.5**

Exemplu de calcularea a covarianței (5)

- **Se calculează covarianța**

x	y
1	4
2	6
3	5
4	7
5	9
6	8

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{15.5}{6} = \mathbf{2.583}$$

x	y
1	5
2	9
3	7
4	4
5	8
6	6

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{-0.5}{6} = \mathbf{-0.083}$$

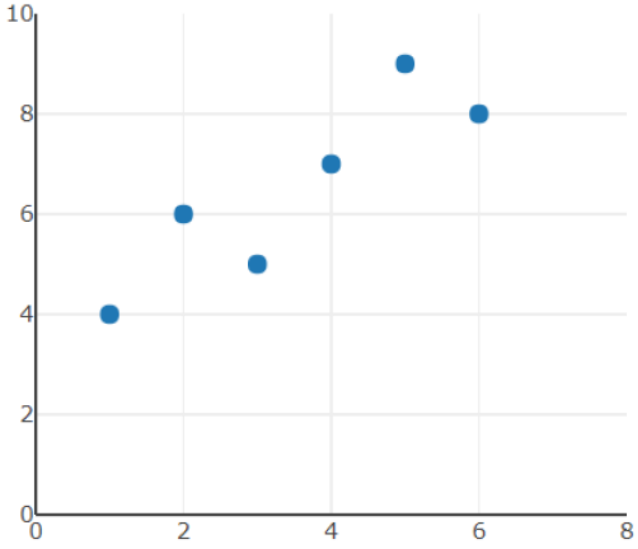
Σ	15.5

Σ	-0.5

Exemplu de calcularea a covarianței (6)

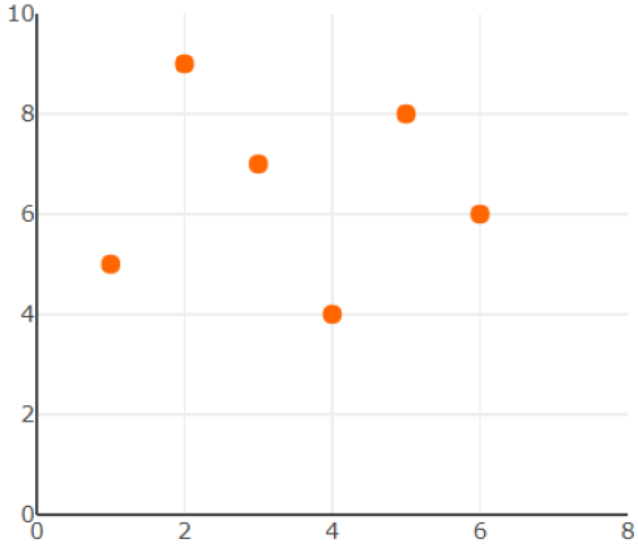
- Se compară covariațele celor 2 cazuri

x	y
1	4
2	6
3	5
4	7
5	9
6	8



cov(x,y) = 2.583

x	y
1	5
2	9
3	7
4	4
5	8
6	6



cov(x,y) = -0.083

Coeficientul de corelație Pearson

- Permite determinarea gradului de variație a unei variabile în funcție de o altă variabilă fără a fi necesară procedura de normalizare
- Coeficientul de corelație Pearson se determină conform formulei

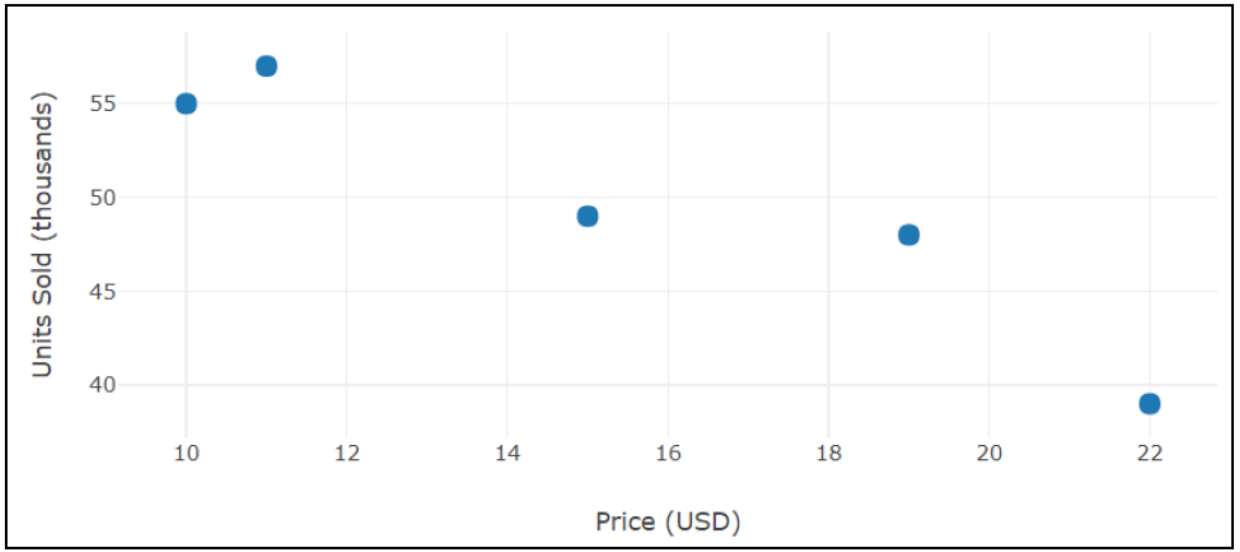
$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{\sum(x-\bar{x})^2}{n}}\sqrt{\frac{\sum(y-\bar{y})^2}{n}}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

- Coeficientul de corelație Pearson poate lua valori în gama -1...1:
  - 1 – corelație pozitivă totală liniară
  - 0 – lipsa corelației liniare
  - -1 – corelație negativă totală liniară

Exemplu de calcularea a coeficientului Pearson (1)

- **Fie datele de vânzarea a unui produs**

Price (USD)	Units Sold (thousands)
10	55
11	57
15	49
19	48
22	39



Exemplu de calcularea a coeficientului Pearson (2)

- **Se calculează valorile medii**

Price (USD)	Units Sold (thousands)
10	55
11	57
15	49
19	48
22	39

$$\bar{x} = \frac{10 + 11 + 15 + 19 + 22}{5} = 15.4$$

$$\bar{y} = \frac{55 + 57 + 49 + 48 + 39}{5} = 49.6$$

Exemplu de calcularea a coeficientului Pearson (3)

- **Se calculează** $(x-\bar{x}), (y-\bar{y}), (x-\bar{x})(y-\bar{y}), (x-\bar{x})^2, (y-\bar{y})^2$ **și sumele**

Price (USD)	Units Sold (thousands)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
10	55	-5.4	5.4	-29.16	29.16	29.16
11	57	-4.4	7.4	-32.56	19.36	54.76
15	49	-0.4	-0.6	0.24	0.16	0.36
19	48	3.6	-1.6	-5.76	12.96	2.56
22	39	6.6	-10.6	-69.96	43.56	112.36
			Σ	**-137.2**	**105.2**	**199.2**

Exemplu de calcularea a coeficientului Pearson (4)

- **Se calculează** coeficientul Pearson conform relației

$$\rho_{X,Y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} = \frac{-137.2}{\sqrt{105.2}\sqrt{199.2}}$$

$$= \frac{-137.2}{10.26 \times 14.11} = \frac{-137.2}{144.8} = \mathbf{-0.948}$$

Σ	-137.2	105.2	199.2