

Statistici Descriptive (1)

Plan de curs

STATISTICI DESCRIPTIVE

Statistici de tendință centrală

Statistici de localizare

Statistici de dispersie (curs 05)

Statistici de asimetrie și aplatizare/boltire (curs 05)

Raport, proporie & Rată (curs 05)

Generalități: Ce sunt statisticile?

- « There are three kinds of lies: lies, damned lies, and statistics. »
(Mark Twain / Benjamin Disraeli)
- **Statistica**: domeniu al matematicii care are ca scop colectarea, descrierea și analiza datelor pentru a extrage concluzii (inferențe).
- **O statistică**: un număr/valoare calculat dintr-un set de date
- **Statistici**: colecție (ansamblu) de numere/valori legate de un set de date

Noțiuni statistice de bază: *populația țintă, accesibilă*

- **Populația**= ansamblu de obiecte/subiecți pe care dorim să o studiem

Exemple:

- ⊗ toți pacienții spitalizați COVID-19 pozitivi
- ⊗ Mulțimea implanturilor dentare

Populația *țintă*: populația pe care dorim să o studiem, populația căreia dorim să extrapolăm rezultatele la sfârșitul studiului pe un eșantion

Populația *accesibilă*: populația la care avem acces: cabinet medical/stomatologic, spital, școală etc.

Populația țintă



Populația accesibilă

Pacienți spitalizați COVID-19 pozitivi



Pacienți COVID-19 pozitivi
urmărire într-un anumit spital

Variable statistice versus Date

Variabila: caracteristica (demografică, clinică etc.) studiată pe diferiți indivizi (aparținând unui eșantion)

- ✓ Variabile calitative (dihotomiale, ordinale, nominale) et cantitative (discrete, continue)
- ✓ Variables independente si variabile dependente

Data: „valoarea” (poate un număr sau un tip/categorie „da”/„nu”) a variabilei
una sau mai multe variabile pot fi măsurate pe un individ ->

Serie statistica de date = valorile variabilelor măsurate în timpul unui studiu

Notiuni de baza: Tipul unei serii statistice

Număr de variabile	Serie statistică
1	uni variată
2	bi variată
3	tri variată
>3	multi variată

Număr de variabile cantitative	Serie statistică
1	uni dimensională
2	bi dimensională
3	tri dimensională
>3	multi dimensională

Tipuri de serii statistice: exemple

- unidimensională: Varsta (ani)

	x ₁	x ₂	x ₃	x ₄	x ₅	x ₆	x ₇
Varsta (ani)	20	45	70	67	60	35	55

- bidimensională: Varsta (ani) și Durata de la debutul infecției COVID-19 și spitalizare(zile)

	x ₁	x ₂	x ₃	x ₄	x ₅	x ₆	x ₇
Varsta (ani)	20	45	70	67	60	35	55
Durata de la debutul infecției COVID-19 și spitalizare(zile)	10	6	4	12	2	10	5

- univariată: Embolie pulmonară (da=1;nu=0)

	x ₁	x ₂	x ₃	x ₄	x ₅	x ₆	x ₇
Embolie pulmonară	0	1	0	1	1	1	0

Statistici descriptive

Măsuri de centralitate <ul style="list-style-type: none">♦ Media♦ Mediana♦ Modulul♦ Valoarea centrală	Măsuri de dispersie / împrăștiere <ul style="list-style-type: none">♦ Amplitudine♦ Variația♦ Deviația standard♦ Coeficientul de variație♦ Eroarea standard
Măsuri de simetriei <ul style="list-style-type: none">♦ Asimetria♦ Boltirea	Măsuri de localizare <ul style="list-style-type: none">♦ Cvartile (decile; percentile)

Statistici descriptive de tendință centrală

Media aritmetică

Mediana

Modulul

Valoarea centrală

Media ponderată

Media geometrică

!!! Statistica/statisticile observate pe un eșantion \approx parametrul/parametrii pe populație

Eșantion versus Populație: Statistică vs. Parametri

Populație → Parametru

Media aritmetică calculată pe o populație

$$\mu = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \dots + x_N}{N}$$

x_i = valoarea variabilei cantitative pentru al i-lea pacient
 N = talie populație

Eșantion → Statistică

Media aritmetică calculată pe un eșantion

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

x_i = valoarea variabilei cantitative pentru al i-lea pacient
 n = talie eșantion

Media aritmetică (Ma)

Avantaje

- ✓ Orice valoare din serie este luată în considerare în calculul mediei
- ✓ suma abaterii de la medie este zero:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

Dezavantaje

- ✗ poate fi influențată de valori extreme
- ✗ modificarea unei singure valori în seria de date va influența (schimba) media
- ✗ are sens doar pentru o variabilă CANTITATIVĂ

Media aritmetică (Ma)

- Media aritmetică: $= (5+6+7+8+8+9+9+10+10)/9 = 8$
- **Excel:** =AVERAGE(A2:I2)

	A	B	C	D	E	F	G	H	I
1	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉
2	5	6	7	8	8	9	9	10	10

- Este parametrul cel mai preferat ca măsură de centralitate atât ca și parametru de descriere a datelor cât și ca estimator
- **Dar**, are semnificație **DOAR DACĂ** variabila de interes este *cantitativă*.

Media aritmetică (Ma)

	A	B	C	D	E	F	G	H	I
1	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉
2	5	6	7	8	8	9	9	10	10

- **Proprietăți:**

1. Orice valoare a seriei este luată în considerare în calculul mediei.

Media aritmetică = $(5+6+7+8+8+9+9+10+10)/9 = 8$

2. **Valorile extreme** pot influența media aritmetică distrugându-i reprezentativitatea.

	A	B	C	D	E	F	G	H	I	J	K
1	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉		Media aritmetică
2	5	6	7	8	8	9	9	10	100		18

3. Media aritmetică se situează printre valorile seriei de date.

Minimum = 5

Maximum = 10

Media aritmetică = 8

	A	B	C	D	E	F	G	H	I
1	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉
2	5	6	7	8	8	9	9	10	10

Media aritmetică (Ma)

4. Suma diferențelor dintre valorile individuale din serie și medie este zero:

$$(5-8) + (6-8) + (7-8) + (8-8) + (8-8) + (9-8) + (9-8) + (10-8) + (10-8) = -3 -2 -1 + 0 + 0 + 1 + 1 + 2 + 2 = -6+6 = 0$$

$$(X_1 - m) + (X_2 - m) + \dots + (X_n - m) = 0$$

	A	B	C	D	E	F	G	H	I
1	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉
2	5	6	7	8	8	9	9	10	10

5. Schimbarea originii scalei de măsură a variabilei X din care provine seria de date are influență asupra mediei. Fie $X'' = X + C$ (unde C este o constantă)

	A	B	C	D	E	F	G	H	I	J	K
1	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉		Media aritmetică
2	5	6	7	8	8	9	9	10	100		18
3											
4	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉		Media aritmetică
5	95	96	97	98	98	99	99	100	190		108

Mediana (Me): definiție și calcul

- **Mediana:** valoarea care împarte seria în 2 grupuri egale de același efectiv.
- **Cum să găsiți/determinați mediana?**
 - sortați datele ascendent
 - uitați-vă la talia eșantionului (« n »)

$$Me = \begin{cases} x_{\frac{n+1}{2}}, & \text{daca } n \text{ este impar} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{daca } n \text{ este par} \end{cases}$$

Mediana (Me):

Avantaje

- ✓ variabile cantitative și calitative ordinale
- ✓ nu este foarte sensibilă la valori extreme, spre deosebire de medie

Dezavantaje

- ✗ nu se pretează la operații algebrice
- ✗ nu se aplică variabilelor calitative nominale

Exemplu de calcul: media și mediana

	continuă	continuă ~ discretă	dihotomială	ordinală
Pacient	Inălțime (cm)	Varsta (ani)	Sex	Categorii de IMC ^(a)
1	170	50	F	Greutate normală
2	160	45	F	Supraponderabilitate
3	187	38	H	Obezitate clasa 1
4	172	25	H	Greutate normală
5	157	65	F	Supraponderabilitate
6	175	56	H	Greutate normală

Variabila: Varsta

Media aritmetică (Ma) =?

$$Ma = (50 + 45 + 38 + 25 + 65 + 56) / 6 = 46,5 \text{ ani}$$

Mediana (Me)=?

Seria statistică ordonată: 25, 38, **45, 50**, 56, 65

Talie eșantion: număr par = 6 =>

$$Me = (45 + 50) / 2 = 47,5$$

Pare (2, 4, ...)

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

MODULUL: definiție și calcul

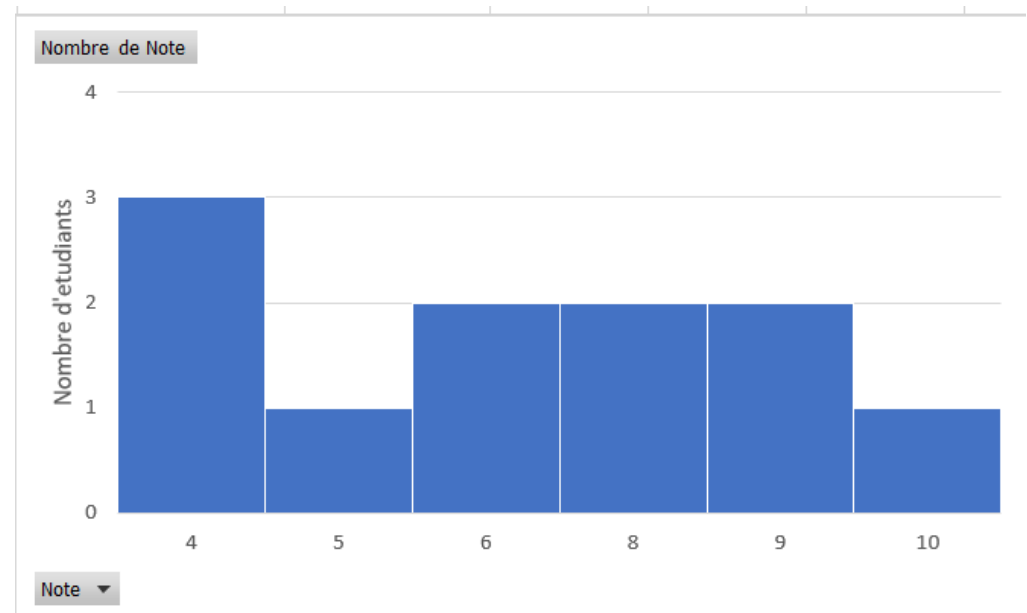
- Definiție: valoarea seriei statistice cu cea mai mare frecvență (valoarea seriei care apare cel mai des)
- nu există o formulă matematică care să o calculeze
- o serie de date statistice poate avea mai multe valori modale.

MODULUL: definiție și calcul

Notele obținute la examenul practic Biostat de către un eșantion de 11 studenți:

4, 9, 5, 8, 6, 4, 9, 10, 8, 6, 5, 4

- Modulul: 4 → serie unimodală

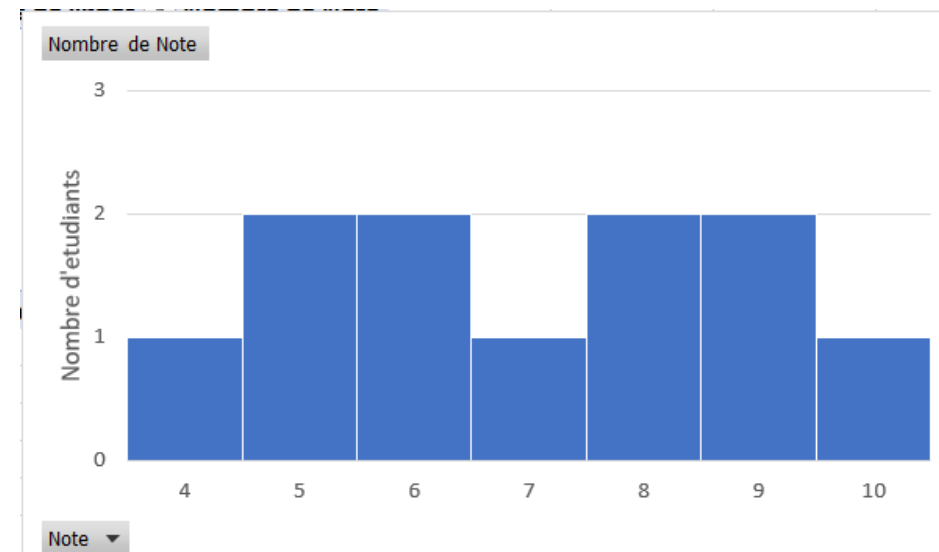


MODULUL: exemplu

Notele obținute la examenul practic Biostat de către un eșantion de 11 studenți:

4, 9, 5, 8, 6, 7, 9, 10, 8, 6, 5

- Modulul: 5, 6, 8, 9
- → serie **multimodală (plurimodală)**



MODUL (Mo): proprietăți

Avantaje

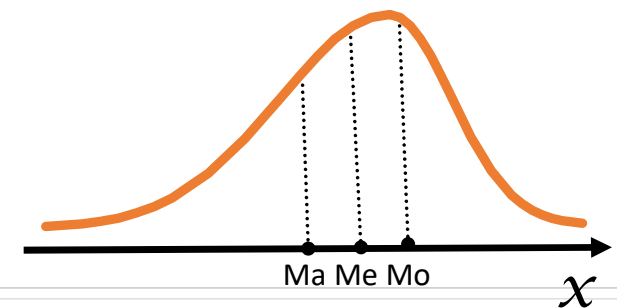
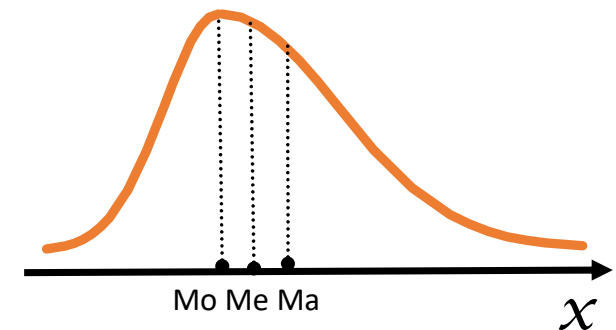
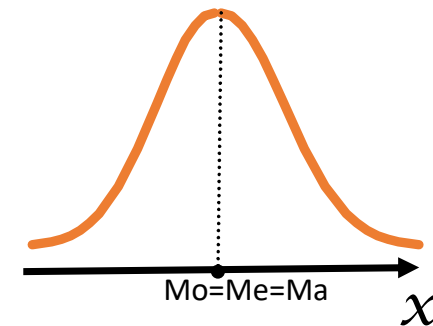
- ✓ sensibilitate scăzută la valorile extreme ale seriei.
- ✓ poate fi un indicator al unei serii de date eterogene
- ✓ dacă datele sunt eterogene (serie bimodală), este mai bine să aveți două valori modale decât o mediană

Dezavantaje

- ✗ nu este potrivit pentru calcule
(Transformați scara de măsurare a seriei de date statistice: $X'' = C \cdot X$, $C = \text{constantă}$)

Poziții relative ale mediei, medianei și modului

- Distribuție **simetrică**:
Modulul \approx Media \approx Mediana
- Distribuție **asimetrică la dreapta**:
Modulul $<$ Mediana $<$ Media
- Distribuție **asimetrică la stânga**:
Media $<$ Mediana $<$ Modulul



Alte măsuri de tendință centrală- Media aritmetică ponderată

- Alte măsuri de tendință centrală

- Media ponderată
$$m_x = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

- ex. Calculul mediei finale la facultate în funcție de numărul de credite

- Media geometrică
$$\sqrt[n]{X_1 X_2 \dots X_n}$$

- Valoarea centrală
$$\frac{X_{\max} + X_{\min}}{2}$$

Alte măsuri de tendință centrală

- **Media ponderată: exemplu**
- Fiecare valoare X_i este înmulțită cu o pondere W_i pozitivă, care indică importanța valorii respective în raport cu celelalte valori:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Credite	6	5	2	4	4	5	2	2			
2	Nr. Crt.	Nume, prenume (initiale)	Anatomie și Histologie	Fiziologie	Educație pentru sănătate	Farmacologie	Nursing	Comportament & Psihologie & Sociologie	Comunicare	Opțional	Media aritmetica	Media ponderata	Media generală
3	1	BR	7	7	7	8	8	5	7	8	7.13	7.00	7.06
4	2	BA	8	8	8	9	9	7	8	6	7.88	7.97	7.92
5	3	BS	9	9	7	7	10	5	7	9	7.88	7.93	7.90
6	4	BR	10	6	8	8	8	6	8	9	7.88	7.80	7.84
7	5	BA	8	7	9	9	9	9	8	8	8.38	8.33	8.35
8	6	CD	9	8	6	9	6	7	8	9	7.75	7.83	7.79
9	7	CE	5	5	7	10	7	10	9	9	7.75	7.43	7.59
10	8	CR	6	10	5	10	5	7	6	6	6.88	7.17	7.02
11	9	CS	5	5	6	5	7	4	8	9	6.13	5.63	5.88
12	10	CT	6	10	10	6	9	9	6	10	8.25	8.10	8.18

- Dacă ponderile W_i sunt alese egale și pozitive atunci se obține media aritmetică obișnuită

MESURI DE LOCALIZARE: cuantile și percentile

- **Cuantilele** (Q_1, Q_2, \dots, Q_{q-1}): **valori remarcabile** care împart seriile de date ordonate în q subseturi (grupuri) consecutive egale.
- **Cuantila de ordinul α ($0 < \alpha < 1$)** este valoarea x_α cu proprietatea că α (%) din date sunt mai mici decât x_α
- **cel mai frecvent utilizate cuantile: cuartilele, cvintilele, decilele, centilele**
- **Percentila de ordinul p :**
 - împarte seria în două submulțimi (astfel încât $p\%$ din valori sunt mai mici decât ea și $(100-p)\%$ din valori sunt mai mari)
 - este cuantila de ordinul $\alpha = p/100$

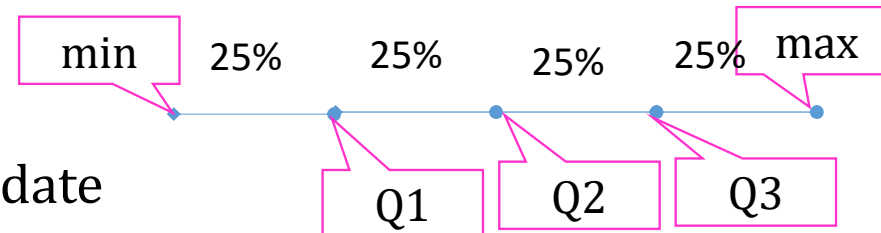
MESURI DE LOCALIZARE: cuantile și percentile

• **Cuartilele (Q_1, Q_2, Q_3):** împart seria de date în patru grupuri având aceeași proporție de date.

- Prima cuartilă (Q_1):
 - 25% din valori $\leq Q_1$, 75% din valori sunt $\geq Q_1$
- A doua cuartilă (Q_2) este mediana seriei (50%)
 - 50% % din valori $\leq Q_2$ / mediana, 50% % din valori $\geq Q_2$ / mediana
- A treia cuartilă (Q_3):
 - 75% din valori sunt $\leq Q_3$, 25% sunt $\geq Q_3$

• **Cvintilele (V_1, V_2, V_3, V_4):**

- Prima cvintilă (V_1) separă ultimele 20% din date
- A doua cvintilă (V_2) separă 40% din date
- A treia cvintilă separă (V_3) 60% din date
- A patra cvintilă (V_4) separă 80% din date



MESURI DE LOCALIZARE: decile și centile

- **Decilele(D_1, \dots, D_9):**

- Prima decilă separă 10% din date
- A doua decilă separă 20% din date
- ...
- A 9-a decilă separă 90% din date

- **Centilele(C_1, \dots, C_{99}): permit împărțirea seriei de date în 100 de subseturi egale)**

- Prima centilă (C_1) separă 1% din date
-
- A 99 centilă (C_{99}) separă 99% din date

STATISTICI DESCRIPTIVE DE DISPERSIE: intervalul intercuartilic

- **Intervalul intercuartilic IQR: [Q1, Q3]**

- **este intervalul dintre prima cuartilă (Q1) și a treia cuartilă (Q3)**

- În articolele științifice, acesta este raportat după mediana

- Ex: Cholesterol mg/dl (mediana, IQR): 189 [162; 211]

- **Distanța intercuartile: EQR: Q3-Q1**

- Uneori, unii cercetători arată doar diferența dintre quartila 3 și quartila 1 (distanța intercuartile $EQR = Q3 - Q1$) în loc de cele două valori ale cuartilelor

- Ex: Colesterol mg/dl (mediana, EQR): 189 (49)

- Cea mai bună reprezentare/raportare este dacă arătăm quartilele (Q1, Q3), nu doar EQR

STATISTICI DESCRIPTIVE DE DISPERSIE: intervalul intercuartilic

- Variabila: Nota la examenul practic

X ₁	9
X ₂	6
X ₃	4
X ₄	9
X ₅	4
X ₆	8
X ₇	8
X ₈	9
X ₉	7
X ₁₀	4
X ₁₁	10
X ₁₂	10

Ordonare

	A	B
1	X ₁₀	4
2	X ₃	4
3	X ₅	4
4	X ₂	6
5	X ₉	7
6	X ₆	8
7	X ₇	8
8	X ₁	9
9	X ₄	9
10	X ₈	9
11	X ₁₁	10
12	X ₁₂	10

Me=8

$$EQR = Q_3 - Q_1 = 9 - 5.5 = 3.5$$

$$A = 10 - 4 = 6$$

$$Me = [X_{12/2} + X_{(12/2+1)}] / 2 = (X_6 + X_7) / 2 = (8 + 8) / 2 = 8$$

Formule Excel:

(Mediana) Me:

$$=MEDIAN(B1:B12)$$

(Intervalul dintre cuartila 3 și 1) IQR:

$$=QUARTILE(B1:B12,3) - QUARTILE(B1:B12,1)$$

(Amplitudinea) A:

$$=MAX(B1:B12) - MIN(B1:B12)$$

STATISTICI DESCRIPTIVE DE DISPERSIE: intervalul intercuartilic

	A	B
1	X ₁₀	4
2	X ₃	4
3	X ₅	4
4	X ₂	6
5	X ₉	7
6	X ₆	8
7	X ₇	8
8	X ₁	9
9	X ₄	9
10	X ₈	9
11	X ₁₁	10
12	X ₁₂	10

Me=8

$$Q_3 - Q_1 = 9 - 5.5 = 3.5$$

$$A = 10 - 4 = 6$$

Me: ½ din studenți au avut nota la examenul practic < 8 și ½ au avut nota > 8

Q1: 25% din studenți au note ≤ 5.5

Q3: 75% din studenți au note ≤ 9

IQR: 50% din studenți au note care nu diferă una față de alta cu mai mult de 3.5 puncte

A: Diferența dintre nota maximă și nota minimă a fost de 6 puncte

- Variabila de studiu: Nota la examenul practic



MULȚUMESC PENTRU ATENȚIE!