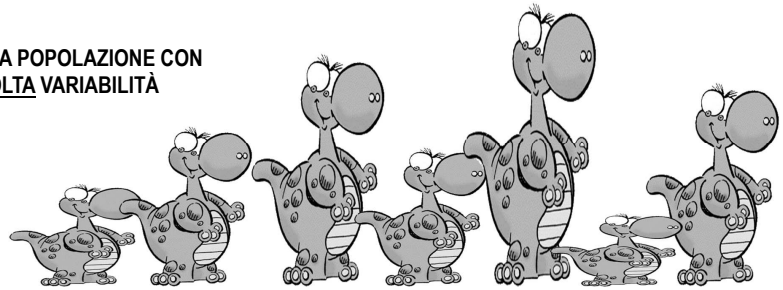
LEZIONI DI STATISTICA MEDICA

Indici di dispersione

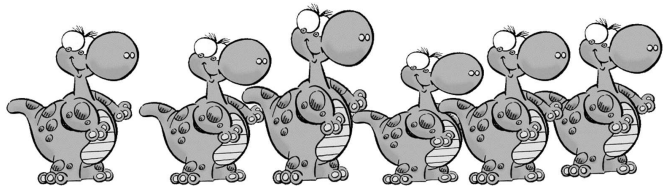
Sezione di Epidemiologia & Statistica Medica
Università degli Studi di Verona

la variabile d'interesse è l'ALTEZZA

UNA POPOLAZIONE CON MOLTA VARIABILITÀ



UNA POPOLAZIONE CON POCA VARIABILITÀ



INDICI DI DISPERSIONE

(measures of dispersion)

1. CAMPO DI VARIAZIONE (range)
2. DISTANZA INTERQUARTILE
3. DEVIANZA
4. VARIANZA
5. DEVIAZIONE STANDARD
6. COEFFICIENTE DI VARIAZIONE

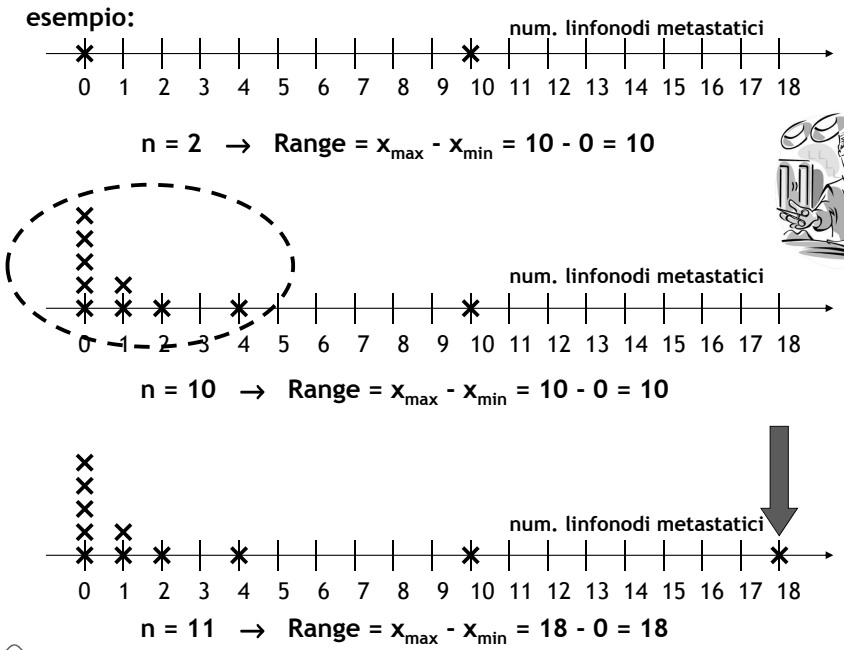
RANGE (CAMPO DI VARIAZIONE)

$$\text{Range} = x_{max} - x_{min}$$

differenza tra il valore massimo e il valore minimo osservati

- ✓ Si basa soltanto sui valori estremi della distribuzione e non tiene conto dei valori intermedi
- ✓ E' molto influenzato da osservazioni anomale (***outliers***)
- ✓ Tende ad aumentare al crescere del numero delle osservazioni

esempio:



$n = 2 \rightarrow$ **Range** $= x_{max} - x_{min} = 10 - 0 = 10$

$n = 10 \rightarrow$ **Range** $= x_{max} - x_{min} = 10 - 0 = 10$

$n = 11 \rightarrow$ **Range** $= x_{max} - x_{min} = 18 - 0 = 18$

DISTANZA INTERQUARTILE

$$IQR = Q_3 - Q_1$$

differenza tra il III° quartile (Q3) ed il I° quartile (Q1)

- ✓ In questo intervallo ricade la metà dei valori osservati, posta esattamente al centro della distribuzione.
- ✓ Non è influenzata da osservazioni anomale o estreme.

esempio: Statura matricole della Facoltà di Medicina (A.A. 95/96)

Range $= x_{max} - x_{min} = 193 - 162 =$ **31 cm**

MASCHI

Statura	Freq.	Cumul.
162	1	1
168	1	2
169	1	3
170	3	6
172	2	8
174	2	10
175	5	15
176	3	18
177	3	21
178	3	24
179	1	25
181	1	26
182	2	28
183	2	30
184	1	31
188	1	32
192	1	33
193	1	34
Totale	34	

Calcolo del I° quartile:
(rango percentilico = 25)

1. rango = (34+1) * 25 / 100
= 35 / 4 ≈ 9

2. I° quartile = **174 cm**

Calcolo del III° quartile:
(rango percentilico = 75)

1. rango = (34+1) * 75 / 100
= 35 * 3 / 4 ≈ 26

2. III° quartile = **181 cm**

IQR = Q3 - Q1 = 181 - 174 = 7 cm

DEVIANZA

Nella popolazione

dimensione della popolazione

$$\sum_{i=1}^{N} (x_i - \mu)^2$$

media nella popolazione (*parametro*)

Nel campione

dimensione del campione

$$\sum_{i=1}^{n} (x_i - \bar{x})^2$$

media nel campione (*statistica*)

- ✓ E' un indice di dispersione definito sulla base del concetto di **scarto rispetto ad un punto centrale della distribuzione**.
- ✓ E' la base delle misure di dispersione per variabili quantitative (da essa discendono la Varianza e la Deviazione Standard).

DEVIANZA

formula per il calcolo

PROPRIETA' DELLA SOMMATORIA:

Se a è una costante:

1) $\sum_{i=1}^{n} a = n \cdot a$

2) $\sum_{i=1}^{n} a x_i = a \sum_{i=1}^{n} x_i$

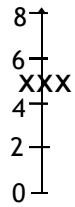
$$\sum (x-\bar{x})^2 = \sum (x^2 - 2\bar{x}x + \bar{x}^2) =$$
$$= \sum x^2 - \sum 2\bar{x}x + \sum \bar{x}^2 =$$
$$= \sum x^2 - 2\bar{x}\sum x + N\bar{x}^2 =$$
$$= \sum x^2 - 2\frac{\sum x}{N}\sum x + N\frac{(\sum x)^2}{N^2} =$$
$$= \sum x^2 - 2\frac{(\sum x)^2}{N} + \frac{(\sum x)^2}{N} =$$

$$\sum (x-\bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{N}$$

- formula usata nella pratica per semplificare il calcolo
- la differenza al 2° membro assume sempre valore positivo **!!!**

esempio:

$\bar{x}$ = 15 / 3 = 5

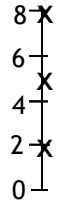


Σx_i = 5 + 5 + 5 = 15

devianza = (5 - 5)² + (5 - 5)² + (5 - 5)² = **0**

Σx_i = 4 + 5 + 6 = 15

devianza = (4 - 5)² + (5 - 5)² + (6 - 5)² = **2**



Σx_i = 2 + 5 + 8 = 15

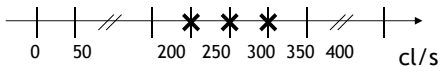
devianza = (2 - 5)² + (5 - 5)² + (8 - 5)² = **18**

esempio:

FEV_1	n_i	p_i	$(x_i-\bar{x})$	$(x_i-\bar{x})^2$
250	1	0.33	-50	2500
300	1	0.33	0	0
350	1	0.33	50	2500
TOT	3	1	0	5000

$\bar{x}$ = (250+300+350) / 3 = 300 cl/s

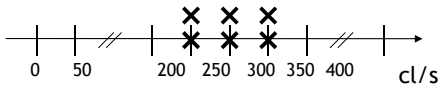
n = 3 → devianza = 5000 cl^2/s^2



FEV_1	n_i	p_i	$(x_i-\bar{x})n_i$	$(x_i-\bar{x})^2n_i$
250	2	0.33	-50*2	2500*2
300	2	0.33	0	0
350	2	0.33	50*2	2500*2
TOT	6	1	0	10000

$\bar{x}$ = (250*2+300*2+350*2) / 6 = = 300 cl/s

n = 6 → devianza = 10000 cl^2/s^2



La devianza raddoppia anche se la variabilità è costante, perché aumenta il numero delle osservazioni!

VARIANZA

- E' una devianza media ossia la devianza rapportata al numero delle osservazioni campionarie (*n*) o della popolazione (*N*).
- E' la media aritmetica dei quadrati degli scarti delle singole osservazioni dalla loro media.

Nella popolazione

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

parametro

Nel campione

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

statistica

Gradi di Libertà

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^{\,2} - (\sum_{i=1}^{n} x_i)^2 / n}{n-1}$$

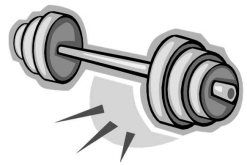
VARIANZA:
formula per il calcolo

I GRADI DI LIBERTÀ rappresentano il numero di osservazioni indipendenti del campione, dal momento che sui dati disponibili è già stata calcolata una statistica (la media campionaria).

La **VARIANZA**:

- Tiene conto di tutte le osservazioni ed è dunque influenzata da eventuali osservazioni anomale (*outliers*).
- Non è direttamente confrontabile con la media o altri indici di posizione in quanto l'unità di misura è elevata al quadrato.

VARIANZA PONDERATA



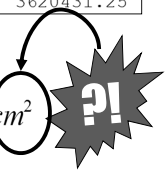
Quando le osservazioni sono raggruppate in una distribuzione di frequenza (*in k classi*):

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{n-1} = \frac{\sum_{i=1}^{k} n_i x_i^{\,2} - \frac{(\sum_{i=1}^{k} n_i x_i)^2}{n}}{n-1}$$

esempio: distribuzione di frequenza della statura delle matricole di Medicina dell'Università di Verona nell'A.A. 95/96

CLASSE	PUNTO CENTRALE (x_i)	FREQUENZA ASSOLUTA	$n_i * x_i$	$n_i * x_i^2$
[150-155)	152.5	1	152.5* 1 = 152.5	$(152.5)^2$* 1 = 23256.25
[155-160)	157.5	8	157.5* 8 = 1260.0	$(157.5)^2$* 8 = 198450.00
[160-165)	162.5	24	162.5*24 = 3900.0	$(162.5)^2$*24 = 633750.00
[165-170)	167.5	34	5695.0	953912.50
[170-175)	172.5	27	4657.5	803418.75
[175-180)	177.5	19	3372.5	598618.75
[180-185)	182.5	9	1642.5	299756.25
[185-190)	187.5	1	187.5	35156.25
[190-195]	192.5	2	385.0	74112.50
TOTALE		125	21252.5	3620431.25

$$s^2 = \frac{\sum_{i=1}^{n} n_i x_i^{\,2} - (\sum_{i=1}^{n} n_i x_i)^2 / n}{n-1} = \frac{362043125 - (212525)^2 / 125}{124} = 57.1 cm^2$$



DEVIAZIONE STANDARD

Nella popolazione

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Nel campione
(d.s. corretta)

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

- Ha sempre valore positivo
- E' una misura della **dispersione della variabile intorno alla media**
- E' una misura di **distanza dalla media**, direttamente confrontabile con le misure di posizione, essendo calcolata con la stessa unità di misura.

esempio: distribuzione di frequenza della statura delle matricole di Medicina dell'Università di Verona nell'A.A. 95/96

CLASSE	PUNTO CENTRALE (x_i)	FREQUENZA ASSOLUTA	$n_i * x_i$	$n_i * x_i^2$
[150-155)	152.5	1	152.5* 1 = 152.5	$(152.5)^2$* 1 = 23256.25
[155-160)	157.5	8	157.5* 8 = 1260.0	$(157.5)^2$* 8 = 198450.00
[160-165)	162.5	24	162.5*24 = 3900.0	$(162.5)^2$*24 = 633750.00
[165-170)	167.5	34	5695.0	953912.50
[170-175)	172.5	27	4657.5	803418.75
[175-180)	177.5	19	3372.5	598618.75
[180-185)	182.5	9	1642.5	299756.25
[185-190)	187.5	1	187.5	35156.25
[190-195]	192.5	2	385.0	74112.50
TOTALE		125	21252.5	3620431.25

$$s = \sqrt{\frac{\sum_{i=1}^{n} n_i x_i^{\,2} - (\sum_{i=1}^{n} n_i x_i)^2 / n}{n-1}} = \sqrt{\frac{3620431.25 - (21252.5)^2 / 125}{124}} = \sqrt{57.1} = 7.6cm$$

ESEMPIO

I dati seguenti si riferiscono al livello di emoglobina (X) in g/100 ml misurato in un campione di 70 donne:

Raggruppate i dati in intervalli di ampiezza 1 g/100 ml.

Determinate la varianza e la deviazione standard della distribuzione (dati raggruppati in intervalli di classe).

9	11,4	12,9
9,3	11,4	13
9,4	11,4	13,1
9,7	11,5	13,1
10,2	11,6	13,2
10,2	11,6	13,3
10,3	11,7	13,3
10,4	11,7	13,4
10,4	11,8	13,4
10,5	11,8	13,5
10,6	11,9	13,5
10,6	11,9	13,6
10,7	12	13,7
10,8	12	13,7
10,8	12,1	14,1
10,9	12,1	14,6
10,9	12,1	14,6
10,9	12,2	14,7
11	12,3	14,9
11	12,5	15
11,1	12,5	
11,1	12,7	
11,2	12,9	
11,2	12,9	
11,3	12,9	

SOLUZIONE

CLASSE	PUNTO CENTRALE (x_i)	FREQUENZA ASSOLUTA (n_i)
[9-10)	9.5	4
[10-11)	10.5	14
[11-12)	11.5	19
[12-13)	12.5	14
[13-14)	13.5	13
[14-15]	14.5	6
TOTALE		70

VARIANZA

$$s^2 = \frac{\sum_{i=1}^{n} x_i^{\,2} n_i - (\sum_{i=1}^{n} x_i n_i)^2 / n}{n-1} = \frac{10235.50 - (841.0)^2 / 70}{69} = 1.91(g/100mL)^2$$

DEVIAZIONE STANDARD

$$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^{\,2} n_i - (\sum_{i=1}^{n} x_i n_i)^2 / n}{n-1}} = \sqrt{1.91} = 1.38g/100mL$$

esercizio

1. range & IQR

2. indici di posizione e dispersione (dati in classi)

ESERCIZIO-I

I dati seguenti si riferiscono al livello di emoglobina (X) in g/100 ml misurato in un campione di 70 donne:

Determinate il range e la distanza interquartile della distribuzione (dati individuali).

9	11,4	12,9
9,3	11,4	13
9,4	11,4	13,1
9,7	11,5	13,1
10,2	11,6	13,2
10,2	11,6	13,3
10,3	11,7	13,3
10,4	11,7	13,4
10,4	11,8	13,4
10,5	11,8	13,5
10,6	11,9	13,5
10,6	11,9	13,6
10,7	12	13,7
10,8	12	13,7
10,8	12,1	14,1
10,9	12,1	14,6
10,9	12,1	14,6
10,9	12,2	14,7
11	12,3	14,9
11	12,5	15
11,1	12,5	
11,1	12,7	
11,2	12,9	
11,2	12,9	
11,3	12,9	

SOLUZIONE-I

valore minimo

9	11,4	12,9
9,3	11,4	13
9,4	11,4	13,1
9,7	11,5	13,1
10,2	11,6	13,2
10,2	11,6	13,3
10,3	11,7	13,3
10,4	11,7	13,4
10,4	11,8	13,4
10,5	11,8	13,5
10,6	11,9	13,5
10,6	11,9	13,6
10,7	12	13,7
10,8	12	13,7
10,8	12,1	14,1
10,9	12,1	14,6
10,9	12,1	14,6
10,9	12,2	14,7
11	12,3	14,9
11	12,5	**15**
11,1	12,5	
11,1	12,7	
11,2	12,9	
11,2	12,9	
11,3	12,9	

valore massimo

$$\text{Range} = x_{max} - x_{min} =$$
$$= 15 - 9 = \mathbf{6\ g/100\ ml}$$

SOLUZIONE-I

Calcolo del I° quartile (*rango percentile = 25*):

1. **rango = (70+1) * 25 / 100 = 71 / 4 ≈ 18**
2. **I° quartile = 10.9 g/100 ml**

Calcolo del III° quartile (*rango percentile = 75*):

1. **rango = (70+1) * 75 / 100 = 71*3 / 4 ≈ 53**
2. **III° quartile = 13.1 g/100 ml**

$$IQR = Q_3 - Q_1 = 13.1 - 10.9 =$$
$$= \mathbf{2.2\ g/100\ ml}$$

9	11,4	12,9
9,3	11,4	13
9,4	11,4	13,1
9,7	11,5	13,1
10,2	11,6	13,2
10,2	11,6	13,3
10,3	11,7	13,3
10,4	11,7	13,4
10,4	11,8	13,4
10,5	11,8	13,5
10,6	11,9	13,5
10,6	11,9	13,6
10,7	12	13,7
10,8	12	13,7
10,8	12,1	14,1
10,9	12,1	14,6
10,9	12,1	14,6
10,9	12,2	14,7
11	12,3	14,9
11	12,5	15
11,1	12,5	
11,1	12,7	
11,2	12,9	
11,2	12,9	
11,3	12,9	

In alcune situazioni il confronto della variabilità all'interno di due gruppi di osservazioni utilizzando la deviazione standard è fuorviante

Due variabili diverse:

In 91 ragazze matricole di Medicina a Verona nell'A.A. 95/96,
la media del **peso** era pari a **55.1 Kg** e la deviazione standard era pari a **5.7 Kg**,
la media della **statura** era pari a **166.1 cm** e la deviazione standard era pari a **6.1 cm.**

E' maggiore la variabilità del peso o la variabilità della statura?

1. Le variabili misurate nei due gruppi sono diverse *(le osservazioni nei due gruppi sono espresse con diverse unità di misura)*

Due gruppi con valori medi molto distanti

Tre neonati pesano rispettivamente **3**, **4** e **5 Kg** (media = **4 Kg**; dev.st. = **1 Kg**).
Tre bambini di 1 anno pesano **10**, **11** e **12 Kg** (media = **11 Kg**; dev.st. = **1 Kg**).

La deviazione standard è uguale nei due gruppi, ma il buon senso suggerisce che la variabilità del peso sia maggiore nei neonati.

1. La variabile misurata è la stessa ma i valori medi delle osservazioni nei due gruppi sono molto distanti *(le osservazioni nei due gruppi sono su diversi ordini di grandezza)*

COEFFICIENTE DI VARIAZIONE PERCENTUALE

CV% = (deviazione standard / media) * 100%

Ci permette di misurare la variabilità **indipendentemente** dalla grandezza e dalla scala di misura delle osservazioni

	Media	Dev. standard	**CV**
Neonati	4 Kg	1 Kg	**25.0 %**
Bambini 1 anno	11 Kg	1 Kg	**9.1 %**

La variabilità del peso è maggiore nei neonati.

	Media	Dev. standard	**CV**
Peso	55.1 Kg	5.7 Kg	**10.3 %**
Statura	166.1 cm	6.1 cm	**3.7 %**

La variabilità del peso è maggiore della variabilità della statura.