Statistica Descrittiva

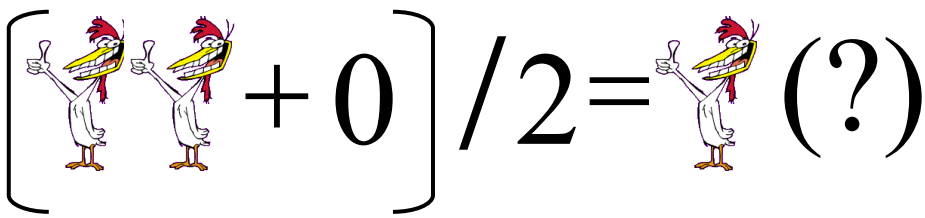
Misure di Posizione

Misure di Dispersione

Sezione di Epidemiologia e Statistica Medica
Università degli Studi di Verona

Il "dilemma" di TRILUSSA

"Me spiego: da li conti che se fanno
seconno le statistiche d'adesso
risurta che te tocca un pollo all'anno:
e, se nun entra ne le spese tue,
t'entra ne la statistica lo stesso
perché c'è un antro che ne magna due"



La Disciplina Statistica

La Statistica, attraverso misure di sintesi (indici o parametri), non ci dice solo quanti "polli mangia" in media una popolazione, ma anche se esistono differenze "alimentari" tra gli individui

INDICI di POSIZIONE

INDICI di DISPERSIONE

Misure della Variabilità del fenomeno oggetto di studio nel collettivo di riferimento

La Sintesi Statistica

Una serie di dati numerici è compiutamente descritta da tre proprietà principali:

- La **tendenza centrale** o **posizione**
- La **dispersione** o **variabilità**
- La **forma**

Queste misure descrittive sintetiche, riassuntive dei dati tabellari, sono chiamate:

- **statistiche**, quando sono calcolate su un campione di dati (si esprimono con lettere dell'alfabeto latino)
- **parametri**, quando descrivono la popolazione od universo dei dati (si esprimono con lettere dell'alfabeto greco

Indici di Posizione

(measures of location or *central tendency)*

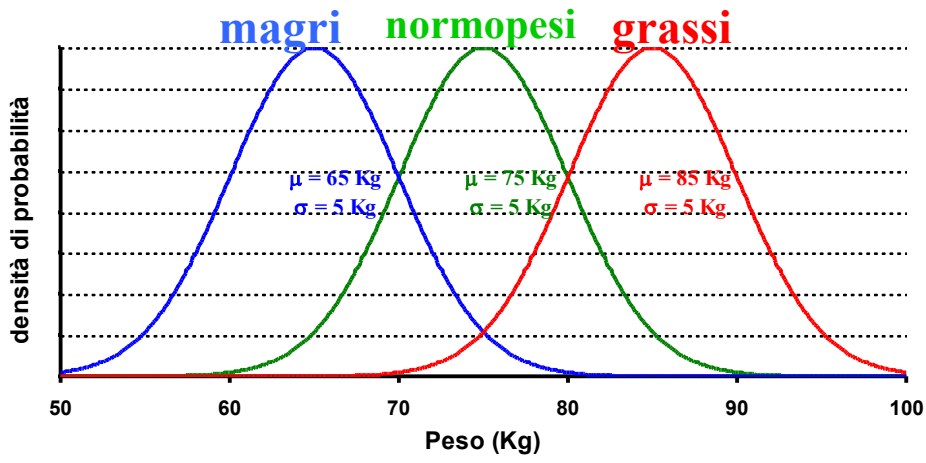
- MEDIA
- MODA
- MEDIANA

Indici di Dispersione

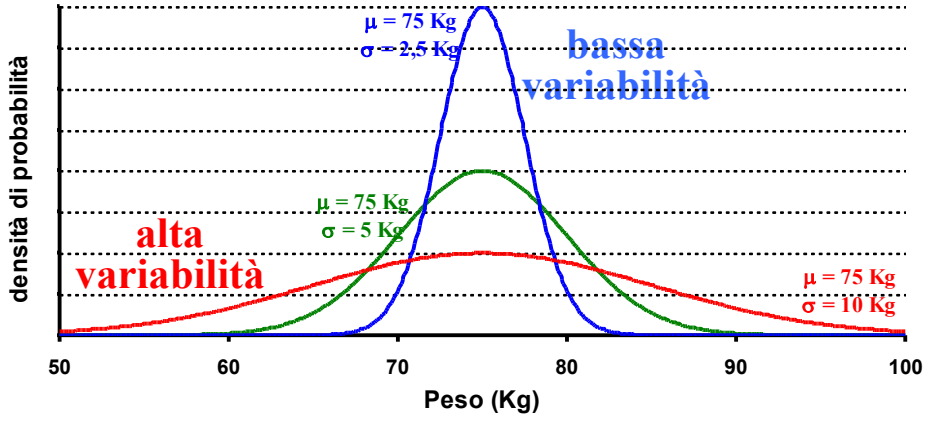
(measures of dispersion)

- CAMPO di VARIAZIONE *(Range)*
- DISTANZA INTERQUARTILE *(Interquartile range)*
- DEVIANZA ⟶ VARIANZA ⟶ DEVIAZIONE STANDARD
- COEFFICIENTE di VARIAZIONE

Queste 3 distribuzioni differiscono per la media (misura di posizione)



Queste 3 distribuzioni differiscono per la deviazione standard (misura di dispersione)



ESEMPLIFICAZIONE

Quali sono le principali MISURE di POSIZIONE nella seguente serie numerica?

x_i	3	15	11	4	5	8	6	4	4
Rango assoluto	*1*	*3*	*3*	*3*	*5*	*6*	*7*	*8*	*9*
Serie ordinata ($x_{(i)}$)	**3**				**5**	**6**	**8**	**11**	**15**

MODA, valore più frequente

MEDIANA, valore centrale in una serie ordinata

MEDIA ($\sum_i x_i / n$) = 60/9 = 6,67

La maggior parte delle variabili biologiche (peso, statura, glicemia) hanno una distribuzione normale, in cui media, mediana e moda coincidono. Alcune variabili (tempo di reazione, tempo di sopravvivenza, numero di linfonodi metastatici, concentrazione serica di IgE) hanno una distribuzione asimmetrica, in cui media e mediana non coincidono.

Esempio:
Negli anni Novanta in un reparto ospedaliero lavoravano 7 medici: 2 specializzandi in formazione, 2 assistenti, 2 aiuti e 1 primario. Il loro reddito era rispettivamente pari a **2, 2, 3, 3, 4, 4** e **25** milioni di lire al mese. Qual è la misura di posizione più adatta a descrivere quest'insieme numerico?

media = $\Sigma x/n = 43/7 =$ **6,14 milioni al mese**
mediana = valore della IV osservazione nella serie ordinata = **3 milioni al mese**

La misura di posizione che descrive meglio il reddito di questi medici è la mediana e non la media.

Esercizio sul calcolo della mediana

Età in anni: 39 25 18 14 69 81 42

1) Ordino i dati in modo crescente
14 18 25 39 42 69 81

2) Calcolo il rango della mediana
n=7 (dispari) rango = (n+1)/2 = (7+1)/2 = 8/2

3) Trovo il valore della quarta osservazione
14 18 25 39 42 69 81

MEDIANA = 39 anni

Esercizio sul calcolo della mediana

Età in anni: 81 72 16 42 38 8

1) Ordino i dati in modo crescente

8 16 72 81

2) Calcolo il rango della mediana

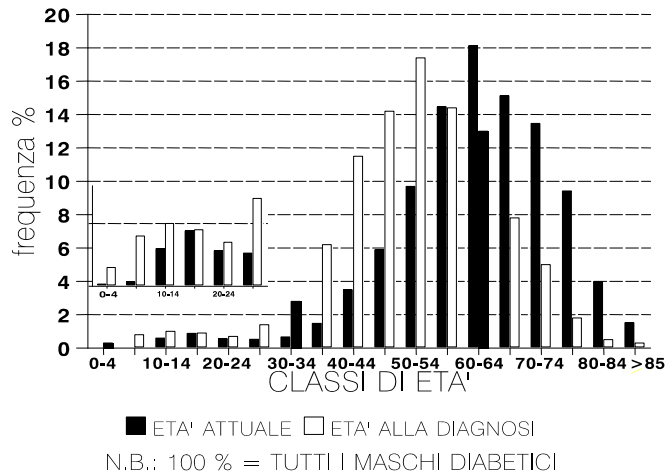
n=6 (pari) rango = (n+1) / 2 = 7/2 = 3.5

3) Faccio la media tra la terza e la quarta osservazione

8 16 38 42 72 81

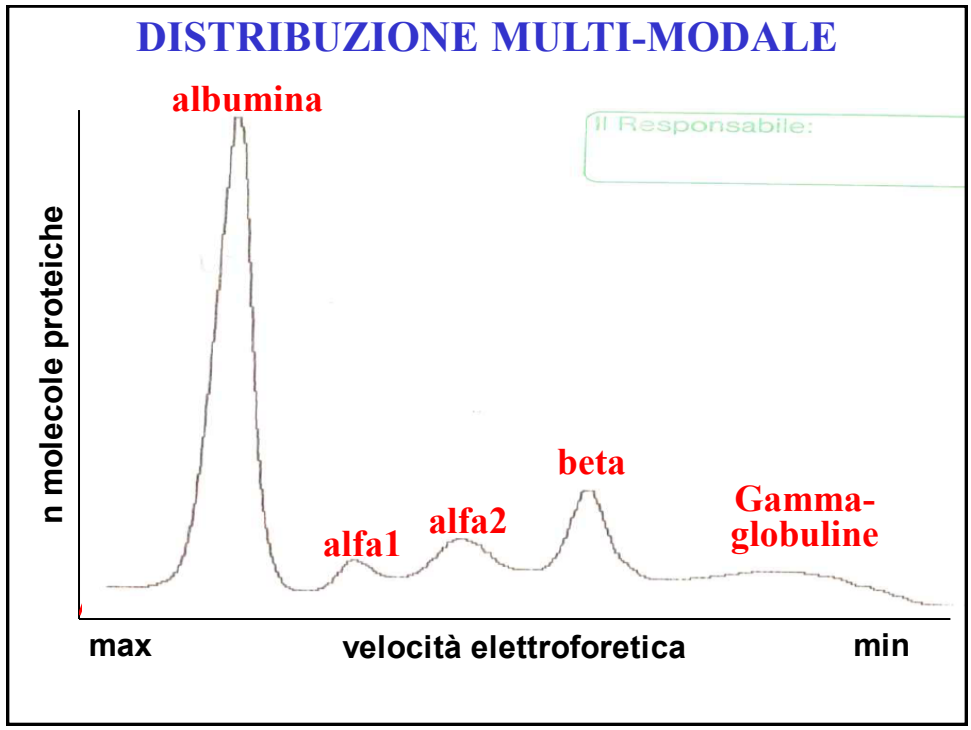
MEDIANA = (38+42)/2 = 40 anni

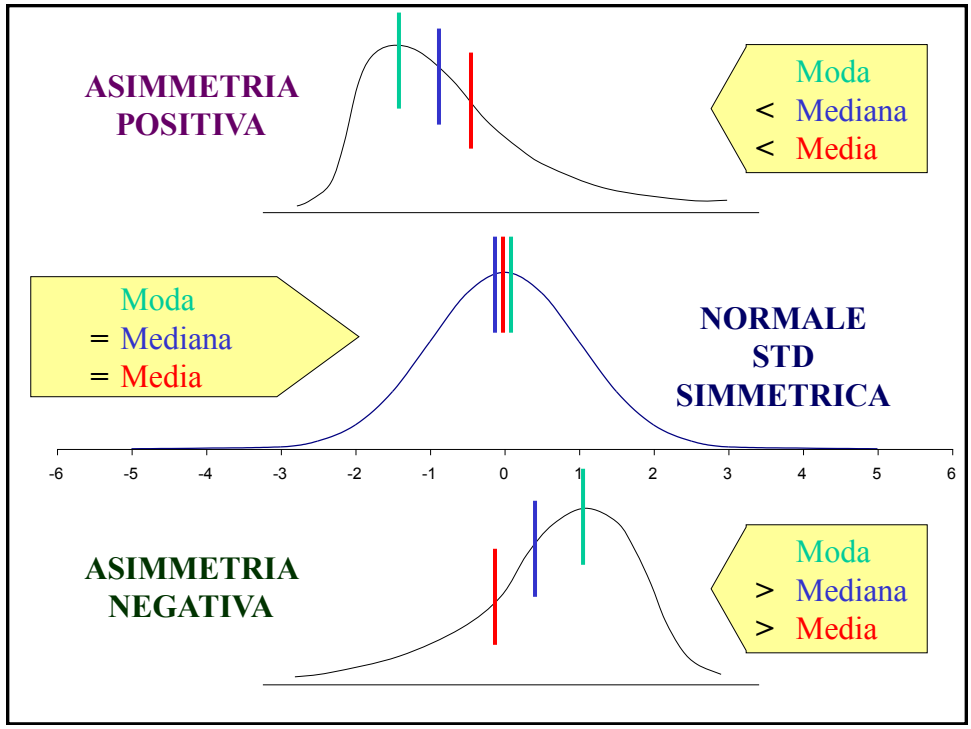
Esempio di distribuzione bimodale (con due mode)
MASCHI DIABETICI a VERONA al 31.12.1986



N.B.: 100 % = TUTTI I MASCHI DIABETICI

Muggeo M, Verlato G, …, de Marco R (1995) The Verona Diabetes Study: a population-based survey on known diabetes mellitus prevalence and 5-year all-cause mortality. *Diabetologia*, 38: 318-325


DISTRIBUZIONE MULTI-MODALE
albumina
alfa1
alfa2
beta
Gamma-globuline
n molecole proteiche
max
velocità elettroforetica
min


ASIMMETRIA POSITIVA
Moda < Mediana < Media
Moda = Mediana = Media
NORMALE STD SIMMETRICA
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
ASIMMETRIA NEGATIVA
Moda > Mediana > Media

Media	Mediana	Moda
La misura di posizione più usata	La misura migliore con distribuzioni asimmetriche (tempo di reazione, tempo di sopravvivenza)	La misura migliore quando un valore ha una frequenza relativa elevata (numero di dita della mano destra)
Facile da trattare matematicamente		
Utilizza tutta l'informazione disponibile sulle unità statistiche ($\Sigma x/n$)		
E' facile calcolare un valore ponderato: $\bar{x} = (\bar{x}_1 n_1 + \bar{x}_2 n_2) / (n_1+n_2)$		
Proprietà dell'equilibrio delle distanze: $\Sigma_i(x_i - \bar{x}) = 0$	Proprietà del minimo delle distanze: $\Sigma\|x - me\| = min$	
Proprietà del minimo degli scarti quadratici: $\Sigma_i(x_i - \bar{x})^2 = min$		

MEDIA PONDERATA

Campione 1 (fantini)
$n_1=30$ $\bar{x}_1=50$ kg

Campione 2 (lottatori di Sumo)
$n_2=10$ $\bar{x}_2=150$ kg

Media complessiva

~~$\bar{x} = (50+150)/2 = 100$ kg~~

Calcolo errato: la media globale deve essere più vicina alla media del campione più numeroso

$$\bar{x} = (n_1 \bar{x}_1 + n_2 \bar{x}_2) / (n_1 + n_2) = (30*50 + 10*150) / (30+10) =$$
$$(1500 + 1500) / 40 = 3000/40 = 75 \text{ kg}$$

	Polli mese	Valore di riferimento	Scarto	Scarto ^{2}
Los Angeles	1	6 media	-5	25
	6		0	0
	11		5	25
Totale	18		0	50

Sostituisco la media con un altro numero

Los Angeles	1	5	-4	16
	6		1	1
	11		6	36
Totale	18		3	53

Los Angeles	1	8	-7	49
	6		-2	4
	11		3	9
Totale	18		-6	62

I PROPRIETA' DELLA MEDIA ARITMETICA:
la somma degli scarti dei singoli valori dalla media aritmetica è ZERO.
II PROPRIETA' DELLA MEDIA ARITMETICA:
la somma del quadrato degli scarti dei singoli valori dalla media aritmetica è la MINIMA POSSIBILE.

Serie aritmetica:

Numero = numero precedente + k

- 3 4 5 6 7 8
- 5 7 9 11

Serie geometrica:

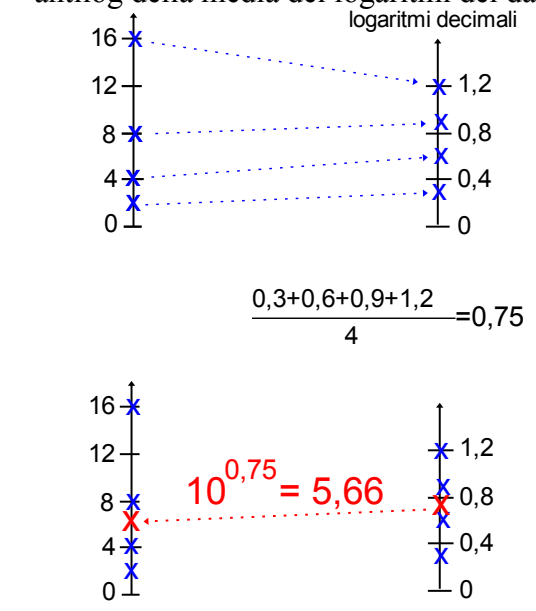
Numero = numero precedente * k

- 4 8 16 32 64 128
- 6 12 24 48 96
- 1/2 1/4 1/8 1/16 1/32

La misura di posizione più indicata in una serie geometrica è la **media geometrica**

Media geometrica

= antilog della media dei logaritmi dei dati



MEDIA da una TABELLA di FREQUENZA

Tempi di degenza (in giorni) per un intervento di emorroidi in un determinato ospedale

Giorni di degenza	Numero di pazienti	Giorni totali
1	9	1*9 = 9
2	15	2*15 = 30
3	12	3*12 = 36
4	9	4* 9 = 36
5	5	5*5 = 25
TOTALE	50	136

MEDIA = $\sum$ nx / $\sum$n = 136/50 = 2,72 giorni

MODA e MEDIANA in una distribuzione di frequenza

moda= 2 giorni

Giorni di degenza	Numero di pazienti	Frequenza cumulativa ass.
1	9	9
2	**15**	24
3	12	**36**
4	9	45
5	5	50
TOTALE	50	

RANGO ASSOLUTO della MEDIANA = (50+1)/2 = 25,5

1 1 1 1 1 1 1 1 1 2
2 2 2 2 2 2 2 2 2 2
2 2 2 2 **3** **3** 3 3 3 3
3 3 3 3 3 3 4 4 4 4
4 4 4 4 4 5 5 5 5 5

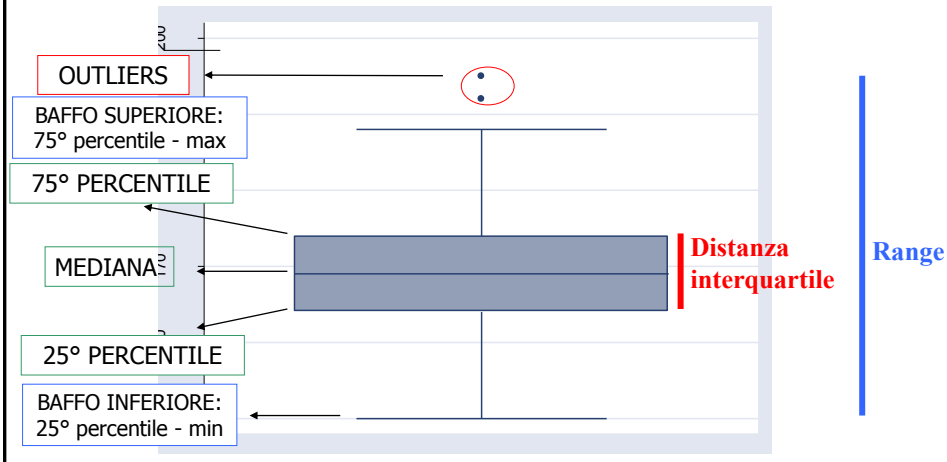
MEDIANA = (3 + 3) / 2 = 3 giorni

Misure di variabilità

Nome italiano	Nome inglese
Campo di variazione	Range
Distanza interquartile	Interquartile range
Devianza (somma di scarti quadratici)	Sum of squares (SSq)
Varianza	Mean Square (MSq)
Deviazione standard	Standard deviation
Coefficiente di variazione	Variation coefficient

RAPPRESENTAZIONE GRAFICA MEDIANTE

BOX-WHISKERS PLOT (GRAFICO SCATOLA E BAFFI)



Range (campo di variazione)

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

(differenza tra il valore massimo e il valore minimo)

Svantaggi

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

Range interquartile o distanza interquartile

$$IQR = Q_3 - Q_1$$

differenza tra il terzo quartile (*75° percentile*)
e il 1° quartile (*25° percentile*)

Osservazioni

- In questo intervallo ricade la metà dei valori, posta esattamente al centro della distribuzione
- Non è molto influenzata da osservazioni anomale o estreme (**statistica robusta**)
- E' adatta a esprimere la variabilità di distribuzioni asimmetriche

DESCRIPTION OF A SERIES OF GASTRIC CANCER PATIENTS

In the series of 921 patients, the total number of dissected lymph nodes was 23,288, with an average of 25.3 ± 16.3 (mean±SD) dissected nodes per case (median 21, range 1-108). The mean number of metastatic nodes was 4.3 ± 7.5 (median 1, range 0-74) in the overall series and 8.3 ± 8.7 (median 5, range 1-74) in pN+ patients.

Bibliografia

De Manzoni G, Verlato G, Roviello F, Morgagni P, Di Leo A, Saragoni L, Marrelli D, Kurihara H, Pasini F, for the Italian Research Group for Gastric Cancer (2002) The new TNM classification of lymph node metastasis minimizes stage migration problems in gastric cancer patients. Brit J Cancer , 87: 171-174

Table 3. Allergy parameters in subjects without self-reported allergic rhinitis and in subjects with perennial, seasonal and perennial+seasonal rhinitis. Absolute frequencies with percentage in brackets are reported for all variables but total IgE, which is expressed as median (interquartile range).

	No rhinitis	**Subjects with self-reported allergic rhinitis**			
	(n=745)	Perennial (n=19)	Seasonal (n=50)	Perennial + seasonal (n=87)	P value
Parental allergy	120/736 (16)	5/19 (26)	21/48 (44)	30/87 (34)	<0.001
Pos. specific IgE					
D.pteronyssinus	56/623 (9)	6/15 (40)	7/43 (16)	19/70 (27)	<0.001
Cat	17/623 (3)	2/15 (13)	4/43 (9)	12/70 (17)	---
Timothy grass	57/623 (9)	3/15 (20)	26/43 (60.5)	39/70 (56)	<0.001
Cl.herbarum	3/623 (0.5)	1/15 (7)	1/43 (2)	3/70 (4)	---
Pariet. judaica	29/623 (5)	1/15 (7)	16/43 (37)	32/70 (46)	<0.001
Total IgE	**36.1 (13.2-101)**	**110.5 (11.6-217.5)**	**87 (38-214.5)**	**106 (50.5-240)**	**<0.001**

Significance of differences was evaluated by chi-squared test for categorical variables and by one-way ANOVA for total IgE after logarithmic transformation. Significance was not evaluated by chi-squared test (---) when cells with expected value<5 exceeded 25%. NS = not significant

Olivieri M, Verlato G, Corsico A, Lo Cascio V, Bugiani M, Marinoni A, de Marco R, for the Italian ECRHS group (2002) Prevalence and features of allergic rhinitis in Italy. Allergy, 57:600-606

Nel primo esempio viene utilizzata come misura di dispersione il **range** per descrivere una casistica nella sua globalità.

Nel secondo esempio viene utilizzata come misura di dispersione la **distanza interquartile**. In questo modo è possibile **confrontare** i livelli di IgE totali fra 4 gruppi di **numerosità molto diversa**: n varia da 19 nel gruppo con rinite allergica perenne a 745 nel gruppo senza rinite.

	Polli/mese	Media	Scarto	Scarto^2	
	5		-1	+1	
Oslo	6	6	0	0	
	7		+1	+1	
Totale	18		0		**2 ← devianza**
Los	1		-5	+25	
Angeles	6	6	0	0	
	11		+5	+25	
Totale	18		0		**50 ← devianza**

Devianza = $\Sigma(x - \bar{x})^2$

(o somma di scarti quadratici)

Somma dei quadrati degli scarti dei singoli valori dalla media

5
6 } Devianza = 2
7

La devianza raddoppia anche se la variabilità rimane costante

5
6
7 } Devianza = 4
5
6
7

Bisogna tener conto della **numerosità**!

Inventiamo la

Varianza = devianza / n

Però, con un campione di 1 soggetto che mangia 6 polli/mese…

Media	Devianza	Varianza non-corretta	Varianza corretta
6	0	0/1 = 0	0/0 = ?

Se noi dividiamo per **n-1** (**1-1=0**) anziché per **n** (1) la varianza è indeterminata, e questo dato rispecchia molto meglio la realtà.

Quindi nell'esempio iniziale **n-1 = 3-1 = 2**

	Media	**Devianza**	**Varianza corretta**
Oslo	**6 polli/mese**	**2 polli^2/mese^2**	**1 polli^2/mese^2**
L.A.	**6 polli/mese**	**50 polli^2/mese^2**	**25 polli^2/mese^2**

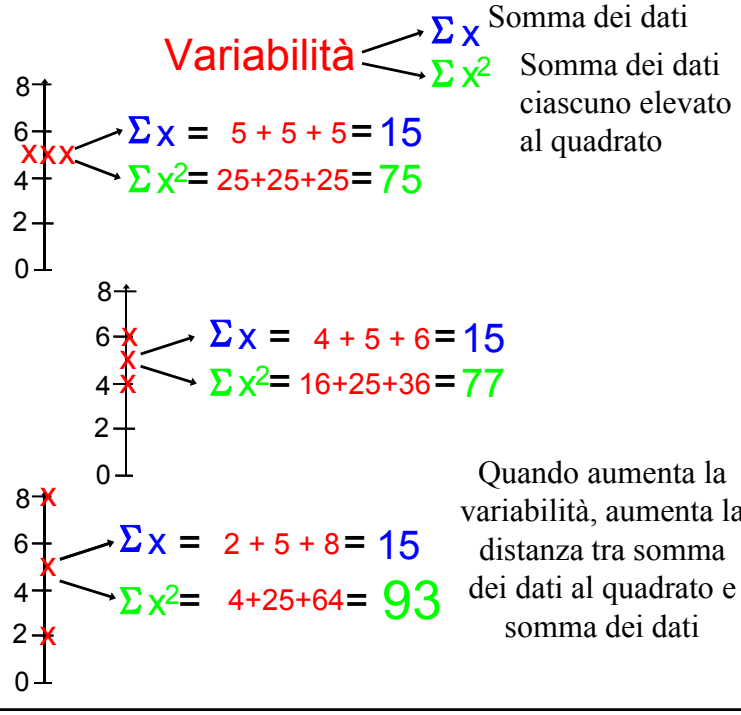
Però, polli^2/mese^2 è una misura un po' difficile!

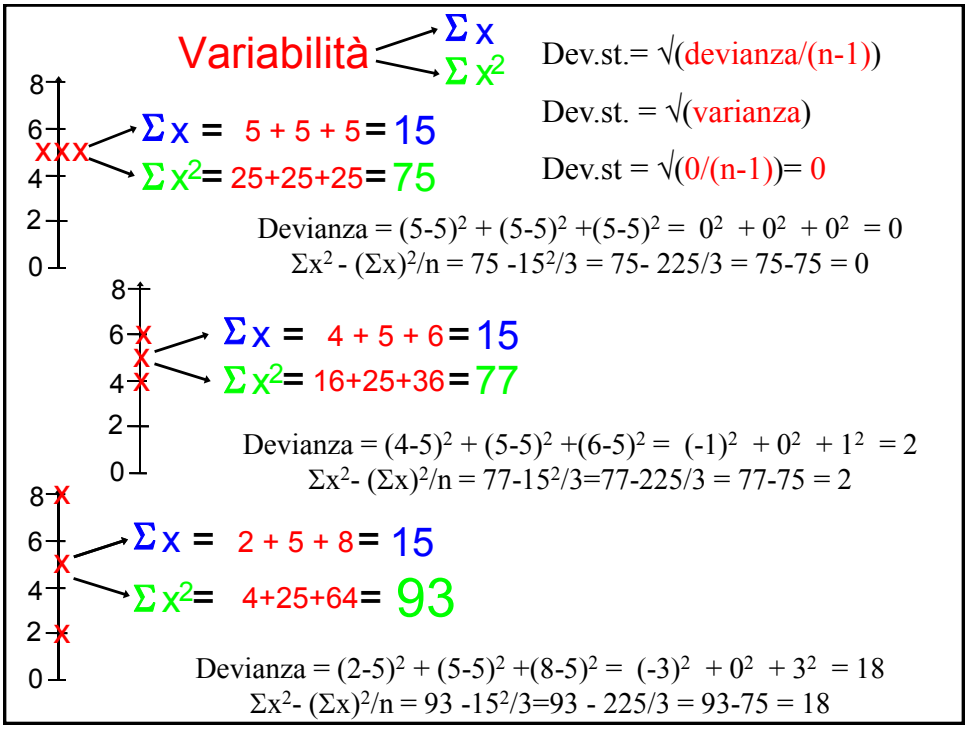
Inventiamo la deviazione standard!

deviazione standard = √ varianza

	Media	Varianza corretta	Deviazione standard
Oslo	**6 polli/mese**	**1 polli^2/mese^2**	**1 pollo/mese**
L.A.	**6 polli/mese**	**25 polli^2/mese^2**	**5 polli/mese**

Oslo: 6 ± 1 polli/mese (media ± DS)
L.A.: 6 ± 5 polli/mese (media ± DS)





Devianza o Somma dei Quadrati (SQ)
(Sum of Squares - SSq)

- Si tratta di un <u>indice di dispersione</u> con riferimento a un <u>centro</u>
- E' la <u>base</u> delle misure di dispersione dei dati, utilizzate in tutta la <u>statistica parametrica</u>.
- Da essa discendono la **Varianza** e la **Deviazione Standard** o **scarto quadratico medio** (sqm)

Formula Euristica **Formula empirica**

$$\sum_{k=1}^{N}(x_k - \bar{x})^2 \quad \Rightarrow \quad \sum_{k=1}^{N}(x_k)^2 - \frac{\left(\sum_{k=1}^{N} x_k\right)^2}{N}$$

A) Varianza o Quadrato Medio (QM)
(Mean Square - MSq)

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

Nella popolazione | **Nel campione (*varianza corretta!*)**

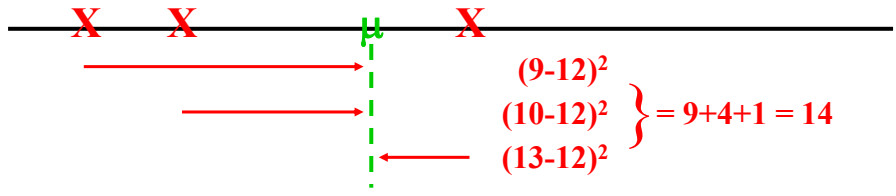
$$\sigma^2 = \frac{\sum_{k=1}^{N}(x_k - \mu)^2}{N} \qquad S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Sigma quadrato

Numerosità Osservazioni

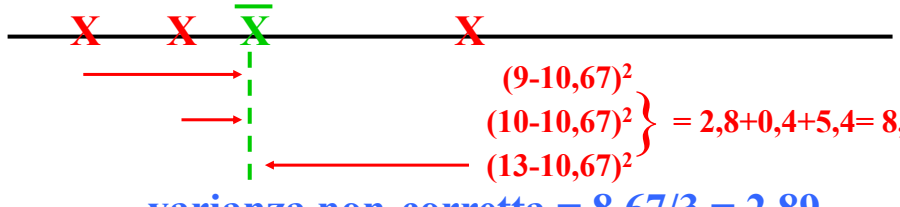
Gradi di Libertà (**gdl**)

Campione: 9, 10, 13 $\mu = 12$



$(9-12)^2$
$(10-12)^2$ } = 9+4+1 = 14
$(13-12)^2$

varianza vera = 14/3 = 4,67



$(9-10{,}67)^2$
$(10-10{,}67)^2$ } = 2,8+0,4+5,4= 8,7
$(13-10{,}67)^2$

varianza non-corretta = 8,67/3 = 2,89

varianza corretta = 8,67/2 = 4,33

B) Varianza

Osservazioni

- E' adatta per distribuzioni simmetriche
- 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 le unità di misura sono elevate al quadrato.
- Ha una notevole importanza nella **teoria statistica.**
- I **gradi di libertà** (*degrees of freedom - df*) rappresentano il numero di osservazioni indipendenti del campione (n -1), dal momento che sui dati disponibili è già stata calcolata una statistica (***x*** *medio*)

A) Deviazione Standard (DS) o (Scarto Quadratico Medio) *(Standard Deviation - SD)*

- Radice quadrata della **Varianza**

Nel campione

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

B) Deviazione Standard

Osservazioni

- E' una misura di distanza dalla media e quindi ha sempre un valore positivo. E' una misura della **dispersione** della variabile casuale intorno alla media
- E' direttamente confrontabile con le misure di posizione, essendo calcolata con la stessa unità di misura
- E' di gran lunga più utilizzata della **varianza** (*che ha un forte valore teorico*) nelle pubblicazioni scientifiche per la sua "praticità d'uso" e immediata confrontabilità con la media

ESERCIZIO

	x_i	x_i^2		$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	3	9		3-6= -3	9
	5	25		5-6= -1	1
	6	36	$\bar{x}$ = 30/5 =6	6-6= 0	0
	7	49		7-6= +1	1
	9	81		9-6= +3	9
totale	30	200		0	20

Devianza = $\Sigma(x - \bar{x})^2 = 20$

oppure

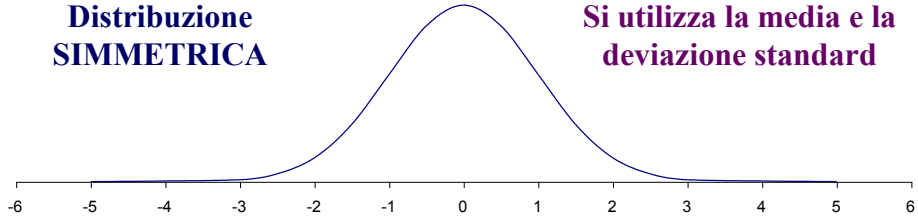
Devianza = $\Sigma x^2 - (\Sigma x)^2/n = 200 - 30^2/5 =$
$= 200 - 900/5 = 200 - 180 = 20$

Varianza = devianza/(n-1) = 20/(5-1) =
20/4 = 5

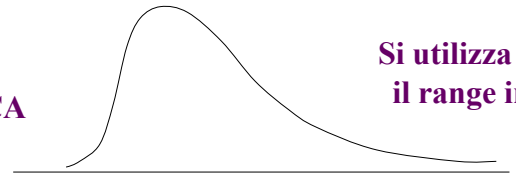
Deviazione standard = $\sqrt{5} = 2{,}24$
$6 \pm 2{,}24$ (media ± DS)

Distribuzione SIMMETRICA

Si utilizza la media e la deviazione standard



Distribuzione ASIMMETRICA



Si utilizza la mediana e il range interquartile

Coefficiente di variazione (CV) - 1

Due gruppi con valori medi molto distanti

Tre neonati pesano rispettivamente **3, 4 e 5 Kg** (media ± DS: 4 ± **1 Kg**).
Tre bambini di 1 anno pesano **10, 11 e 12 Kg** (media ± DS: 11 ± **1 Kg**).
La deviazione standard è uguale nei due gruppi, ma il buon senso suggerisce che la variabilità del peso sia maggiore nei neonati.

Due variabili diverse

In 91 ragazze matricole di Medicina a Verona nell'a.a. 95/96,
il peso era pari a 55,1 ± **5,7 Kg** (media ± DS) con un range di **45-70 Kg,**
la statura era 166,1 ± **6,1 cm** (media±DS) con un range di **150-182 cm.**
E' maggiore la variabilità del peso o la variabilità della statura?

Coefficiente di variazione (CV) - 2

Per rispondere a queste domande è necessario calcolare il **coefficiente di variazione: CV = (deviazione standard / media) * 100**. La deviazione standard viene cioè espressa in percentuale della media.

	Media	Dev. standard	CV
Neonati	4 Kg	1 Kg	**25 %**
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.

Misure di Forma

Misure di Simmetria

1) Coefficiente interquartilico di asimmetria = (Q3-Q2) - (Q2-Q1)
dove Q3, Q2, Q1 =75esimo, 50esimo e 25esimo percentile
Ad esempio, nelle matricole di Medicina di Verona nell'a.a. 95/96 il coefficiente interquartilico di asimmetria vale:

(174,5-169)-(169-164) = 5,5-5 = 0,5 cm

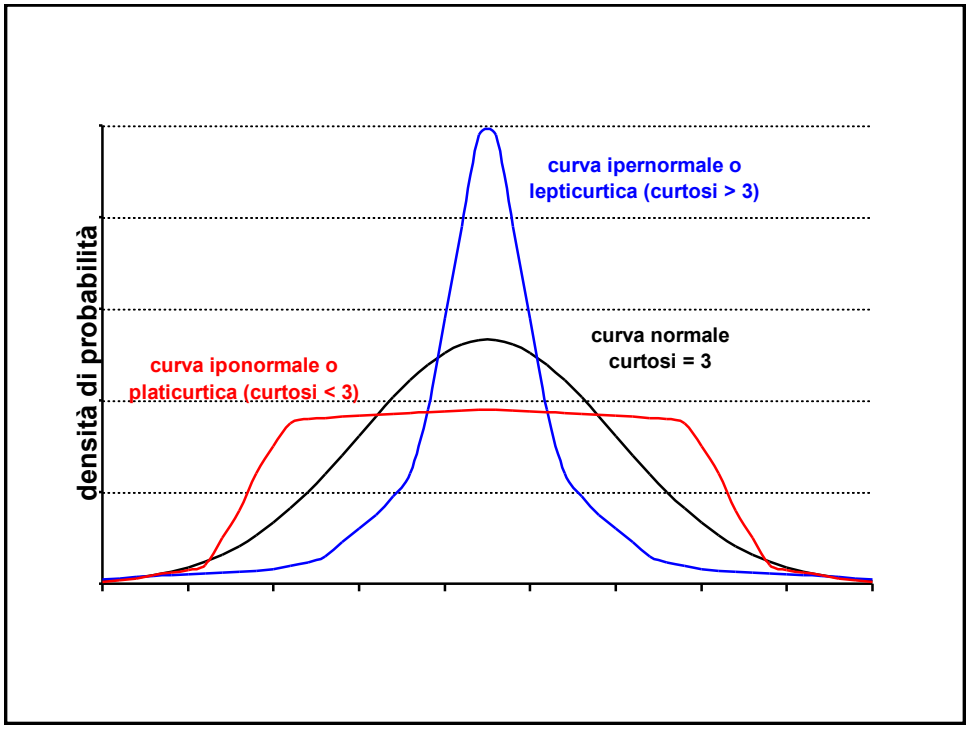
Il coefficiente rileva una lieve asimmetria positiva.

2) Indice di simmetria (skewness) di Pearson = (media - moda) / dev.st.

Misure di Appiattimento (o Curtosi)

1) Indice di Curtosi = misura della concentrazione della distribuzione attorno alla sua media. Indica se la distribuzione è appiattita o presenta un picco in corrispondenza della media.

Indice di Curtosi = $[\Sigma(x-\bar{x})^4/n] / [\Sigma(x-\bar{x})^2/n]^2$


densità di probabilità
curva ipernormale o
lepticurtica (curtosi > 3)
curva normale
curtosi = 3
curva iponormale o
platicurtica (curtosi < 3)