DESCRIZIONE DEI DATI – PARTE II

Indici di posizione

Numeri riassuntivi che forniscono indicazioni sull'**ordine di grandezza** del fenomeno

Indici di dispersione

Numeri riassuntivi che forniscono informazioni sulla **variabilità** (eterogeneità) del fenomeno

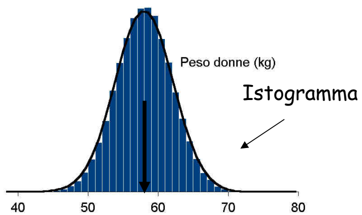
Media Aritmetica (<u>SOLO var. QUANTITATIVE</u>)

Dato un **campione di n unità** su cui è stata rilevata la **variabile X** $\{x_1, x_2, x_3, \ldots\ldots, x_n\}$

la media aritmetica è $\bar{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

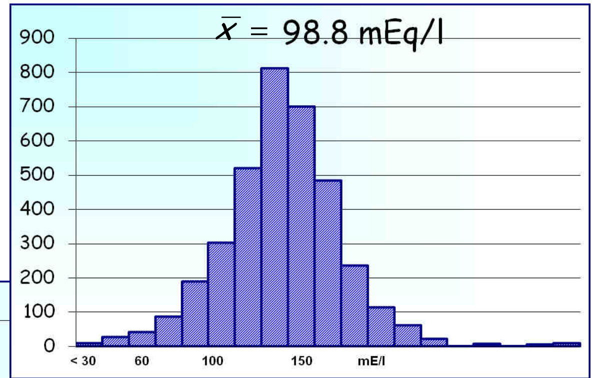
N.B. l'espressione $\sum_{i=1}^{n} x_i$ si legge "sommatoria di x_i con i esteso da 1 a n"

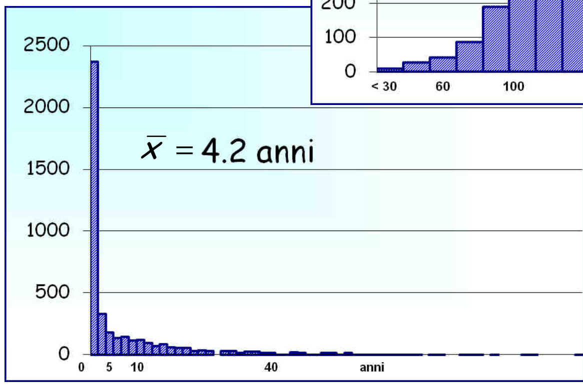
Si presta bene a sintetizzare distribuzioni **simmetriche**



E' meno utile quando la distribuzione è **asimmetrica**.

Concentrazione di cloro nel sudore (simmetrica)

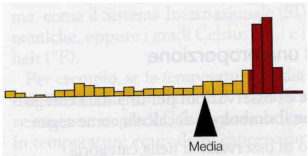


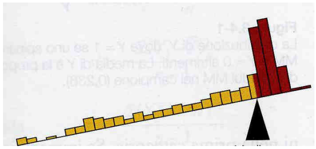


Eta' alla diagnosi nella fibrosi cistica (asimmetria positiva)

Proprietà della media aritmetica (1)

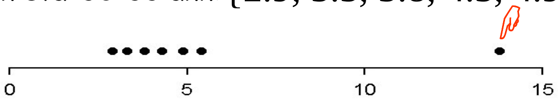
La media aritmetica rappresenta il **baricentro** della distribuzione.





Questa proprietà ha come effetto indesiderato la forte dipendenza dai valori estremi

Durata del travaglio (ore) per il secondo parto naturale in 7 donne in eta' 30-33 anni $\{2.9, 3.3, 3.8, 4.3, 4.9, 5.4, 13.8\}$



$\bar{x}$ = 5.5 ore , ben poco rappresentativa dell'insieme di dati

Esempio

Il numero di infortunati ricoverati in un pronto soccorso in 5 periodi di un'ora è: 28 16 24 31 27.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{28+16+24+31+27}{5} = \frac{126}{5} = 25.2$$

La media di infortuni ricoverati nel pronto soccorso è di 25.2 per ora.

Esempio

Sette soggetti dopo una dieta hanno riscontrato le seguenti perdite di peso (kg) :

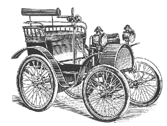
13 3.2 7.4 4.3 8.5 5.9 10.0

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{13+3.2+7.4+4.3+8.5+5.9+10}{7} = \frac{52.3}{7} = 7.5$$

La media della perdita di peso è pari a 7.5 kg.

Media aritmetica: Automobili

N° di automobili per famiglia



x_i	f_i
0	2
1	20
2	25
3	3
Tot.	50

con le frequenze assolute

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{n} = \frac{0 \cdot 2 + ... 3 \cdot 3}{50} = \frac{79}{50} = 1.6$$

con le frequenze relative

$$\bar{x} = \sum_{i=1}^{k} x_i \cdot p_i = 0 \cdot 0.04 + ... 3 \cdot 0.06 = 1.6$$

Media aritmetica: Neonati

Lunghezza supina (cm) in un campione di 60 neonati

Estremi di classe	Valore centrale	Freq. semplici		Freq.cumulate	
		f	p%	F	P%
44.25 ⊣ 45.75	**45.0**	2	*3.3*	2	*3.3*
45.75 ⊣ 47.25	**46.5**	5	*8.3*	7	*11.7*
47.25 ⊣ 48.75	**48.0**	7	*11.7*	14	*23.3*
48.75 ⊣ 50.25	**49.5**	14	*23.3*	28	*46.7*
50.25 ⊣ 51.75	**51.0**	16	*26.7*	44	*73.3*
51.75 ⊣ 53.25	**52.5**	9	*15.0*	53	*88.3*
53.25 ⊣ 54.75	**54.0**	5	*8.3*	58	*96.7*
54.75 ⊣ 56.25	**55.5**	1	*1.7*	59	*98.3*
56.25 ⊣ 57.75	**57.0**	1	*1.7*	60	*100.0*

Per il calcolo della media è necessario considerare come valore rappresentativo di ogni classe il suo valore centrale $_{c}x_i$

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{n} = \frac{45 \cdot 2 + 46.5 \cdot 5 + ... + 57 \cdot 1}{60} = \frac{3022.5}{60} = 50.4$$

$$\bar{x} = \sum_{i=1}^{k} x_i \cdot p_i = ...$$

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot p_i\%}{100} = ...$$

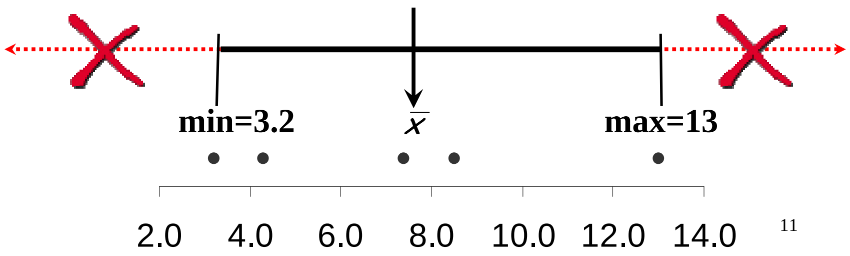
La media calcolata sui dati raggruppati in classi rappresenta una approssimazione di quella determinata a partire dai singoli dati.

Proprietà della media aritmetica (2)

La media aritmetica è sempre compresa tra il più piccolo ed il più grande dei valori osservati

$x_{(1)} \le \quad \le x_{(n)}$

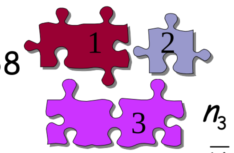
dove $x_{(1)} = \min\{x_1, x_2,..., x_n\}$ $x_{(n)} = \max\{x_1, x_2,..., x_n\}$



Proprietà della media aritmetica (3)

Altezza (m) relativa a 85 ragazzi che frequentano 3 classi diverse

$n_1 = 20$ $\bar{x}_1 = 1.68$



$n_2 = 15$ $\bar{x}_2 = 1.60$

$n_3 = 50$ $\bar{x}_3 = 1.90$

$$\bar{x} = \frac{1.68 \cdot 20 + 1.60 \cdot 15 + 1.90 \cdot 50}{85} = \frac{152.6}{85} = 1.80$$

La media di un insieme di osservazioni organizzate in k gruppi è pari alla media ponderata delle medie parziali $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$ con pesi uguali alla numerosità dei sottogruppi $n_1, n_2, ..., n_k$

$$\bar{x} = \frac{\sum_{i=1}^{k} \bar{x}_i \cdot n_i}{n}$$

Proprietà della media aritmetica (4)

La somma degli scarti delle osservazioni dalla media è pari a

???

Esempio:

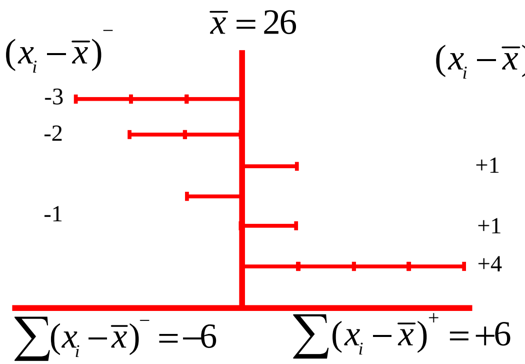
Voti esami sostenuti: $\{23, 24, 27, 25, 27, 30\}$

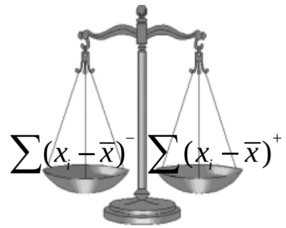
$\overline{x} = 26$

(23-***26***)+(24-***26***)+…

Proprietà della media aritmetica (4)

Voti esami sostenuti: $\{23, 24, 27, 25, 27, 30\}$





Proprietà della media aritmetica (4)

La somma degli scarti delle osservazioni dalla media è nulla

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

Considerazione

La proporzione di soggetti che presentano una caratteristica è la media aritmetica di una variabile che assume valore

- 1 quando la caratteristica è presente
- 0 quando è assente

Es: Presenza di gravi complicazioni dopo un intervento chirurgico con frequenza 24 (12%) su 200 interventi

Se x_i=1 quando la complicazione è presente e x_i=0 quando è assente, si puo' scrivere

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{200} = \frac{24}{200} = 0.12$$

Esempio : Allattamento al seno

Tavola 7 - Donne che hanno partorito nei cinque anni precedenti l'intervista per tempo trascorso dopo il parto prima del tentativo di allattamento al seno e ripartizione geografica- Dati provvisori 2004- 2005 (per 100 donne con le stesse caratteristiche)

Ripartizioni geografiche	Dopo quanto tempo ha attaccato al seno il bambino per la prima volta dopo il parto				
	Subito dopo il parto	Dopo poche ore/entro il primo giorno	Il giorno dopo	Dopo più di due giorni	Totale
Nord-Ovest	55,1	37,0	4,7	3,2	100,0
Nord-Est	59,9	33,1	3,1	3,9	100,0
Centro	43,9	46,2	6,3	3,6	100,0
Sud	38,3	51,3	7,0	3,4	100,0
Isole	46,7	42,0	7,1	4,2	100,0
Italia	**48,4**	**42,4**	**5,6**	**3,6**	100,0

Esempio : Allattamento al seno

Si osservi che gli elementi nell'ultima riga (**Italia**) non sono ottenuti come media della corrispondente colonna.

Ad esempio, il 48.4% di donne allattano subito dopo il parto

	Subito dopo il parto
Nord-Ovest	55,1
Nord-Est	59,9
Centro	43,9
Sud	38,3
Isole	46,7
Italia	**48,4**

è diverso da

(55.1+59.9+43.9+38.3+46.7)/5 =48.7

Esempio : Allattamento al seno

Per calcolare la % di donne che allattano subito dopo il parto in Italia bisogna risalire alle frequenze assolute di donne che allattano subito dopo il parto nelle varie ripartizioni territoriali

Se, ad esempio, il totale di donne analizzate è (2600) rispettivamente : 510 , 490, 510, 580, 510, le frequenze assolute di donne che hanno allattato subito dopo il parto sono

	Subito dopo il parto
Nord-Ovest	0.551*510 = 281
Nord-Est	0.599*490 = 294
Centro	0.439*510 = 224
Sud	0.383*580 = 222
Isole	0.467*510 = 238
Italia	1259

Le donne che allattano subito dopo il parto in Italia sono

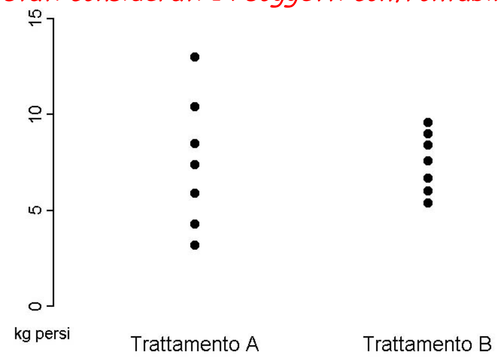
1259/2600 = 0.484 = 48.4%

...la media non dice tutto

E' più efficace una dieta alimentare (A) o un trattamento farmacologico (B) per diminuire di peso?

Per valutare l'entità della perdita di peso (kg) che si verifica dopo trattamento (A o B) sono stati considerati 14 soggetti confrontabili.

Trattamento	
A	**B**
13.0	8.4
3.2	5.4
7.4	7.6
4.3	6.0
8.5	9.6
5.9	6.7
10.0	9.0



Cosa avremmo concluso?

Varianza (SOLO var. QUANTITATIVE)

Si vuole costruire un indicatore che riassume la variabilità del fenomeno

1) usando tutte le osservazioni del campione

2) Misurando il grado di dispersione dei dati rispetto ad un "particolare" valore

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

…ma, poiché $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ è necessario valutare altre possibili misure

La varianza è la media dei quadrati degli scarti delle singole osservazioni (x_i) rispetto alla media campionaria ($\bar{x}$):

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

- può assumere valori strettamente positivi;
- vale 0 in assenza di variabilità; (es. 3.5, 3.5, 3.5, 3.5)
- è tanto più elevata quanto più i dati sono dispersi in un intervallo ampio di valori;
- è fortemente influenzata dall'eventuale presenza di dati estremi per il fatto che si utilizzano i quadrati delle distanze;
- ha per unità di misura il quadrato della scala del fenomeno.

Deviazione Standard

Per avere un indicatore con la stessa unita' di misura della media si prende la radice quadrata della varianza

a partire da un insieme di dati enumerato

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

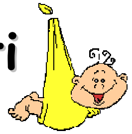
a partire da una tabella di frequenza

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

Trattamento A		Trattamento B	
$_A x_i$	$(_A x_i - \bar{x}_A)^2$	$_B x_i$	$(_B x_i - \bar{x}_B)^2$
13.0	30.57	8.4	0.76
3.2	18.25	5.4	4.54
7.4	0.01	7.6	0.01
4.3	10.06	6.0	2.34
8.5	1.06	9.6	4.28
5.9	2.47	6.7	0.69
10.0	6.39	9.0	2.16
	Somma = 68.79		Somma = 14.78

$s_A = \sqrt{11.47} = 3.39$ $\qquad$ $s_B = \sqrt{2.46} = 1.57$

Deviazione standard: Neonati



Lunghezza supina (cm) in un campione di 60 neonati

Estremi di classe	Valore centrale	Freq. semplici f	Freq. semplici p%	Freq.cumulate F	Freq.cumulate P%
44.25 ⊣ 45.75	**45.0**	2	*3.3*	2	*3.3*
45.75 ⊣ 47.25	**46.5**	5	*8.3*	7	*11.7*
47.25 ⊣ 48.75	**48.0**	7	*11.7*	14	*23.3*
48.75 ⊣ 50.25	**49.5**	14	*23.3*	28	*46.7*
50.25 ⊣ 51.75	**51.0**	16	*26.7*	44	*73.3*
51.75 ⊣ 53.25	**52.5**	9	*15.0*	53	*88.3*
53.25 ⊣ 54.75	**54.0**	5	*8.3*	58	*96.7*
54.75 ⊣ 56.25	**55.5**	1	*1.7*	59	*98.3*
56.25 ⊣ 57.75	**57.0**	1	*1.7*	60	*100.0*

$$s = \sqrt{s^2} = \sqrt{\frac{(45-50.4)^2 \cdot 2 + \ldots + (57-50.4)^2 \cdot 1}{60-1}} = \ldots = 2.5$$

Media e Deviazione Standard

Riassumono 'posizione' e 'variabilita' del fenomeno

Il 68% delle osservazioni sono nell'intervallo $[\bar{x} - s \;,\; \bar{x} + s]$

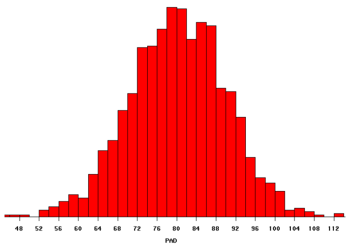
Il 95% delle osservazioni sono nell'intervallo $[\bar{x} - 2 \cdot s \;,\; \bar{x} + 2 \cdot s]$

Il 99% delle osservazioni sono nell'intervallo $[\bar{x} - 3 \cdot s \;,\; \bar{x} + 3 \cdot s]$

- Sono adatti solo a rappresentare distribuzioni simmetriche (con forma approssimativamente normale)
- Permettono di effettuare confronti tra fenomeni nella stessa unita' di misura e con lo stesso ordine di grandezza

Esempio: Pressione diastolica

La PAD è stata misurata in un campione di 1500 uomini tra i 35 e 44 anni. I risultati sono rappresentati con un istogramma delle frequenze relative divise per ampiezza della classe di PAD (classi di 2 mmHg)



Per completare la descrizione del fenomeno potrei calcolare
- un indice di posizione : la media $\bar{x} = 80$
- un indice di dispersione : la deviazione standard $s = 12$

Da un punto di vista descrittivo si potrà dire che :

1) Il 68% delle persone hanno PAD tra 80-12 e 80+12
2) Il 95% delle persone hanno PAD tra 80-2*12 e 80+2*12
3) Il 99% delle persone hanno PAD tra 80-3*12 e 80+3*12

Ai fini predittivi su cio' che mi aspetto nella generica persona in questa classe di età (non necessariamente appartenente allo studio) si potrà dire che :

- c'è una probabilità di 0. 68 di avere PAD tra 68 e 92
- e cosi' via…

Coefficiente di variazione

Abbiamo visto dei metodi empirici per :

- farci un'idea dell'andamento e della distribuzioni di un fenomeno
- confrontare fenomeni nella stessa unita' di misura (e.g. farmaco vs dieta)

Come confrontare la variabilita' di fenomeni espressi nella stessa unita' di misura ma con ordini di grandezza diversi ?

Come confrontare la variabilita' di fenomeni diversi tra loro ?

Coefficiente di variazione

Il coefficiente di variazione (CV) è il rapporto fra la deviazione standard e la media aritmetica:

$$CV = \frac{s}{\bar{x}}$$

È un numero puro che può assumere valori positivi o negativi a seconda del segno della media.

- l'unita' di misura viene eliminata
- la variabilita' viene standardizzata per l'ordine di grandezza del fenomeno

CV : Esempio

Medesimo fenomeno (reddito) in gruppi con ordine di grandezza differente (operai e miliardari)

Reddito (migliaia di €)			
Operai		**Miliardari**	
1	20	**1**	800020
2	60	**2**	800060

$\bar{x}_o = 40 \, mila \; euro$

$\bar{x}_M = 800040 \, mila \; euro$

$s_o = s_M = 28.3 \, mila \; euro$

$$CV_o = \frac{28.3}{40} = 0.71 \qquad CV_M = \frac{28.3}{800040} = 0.00004$$

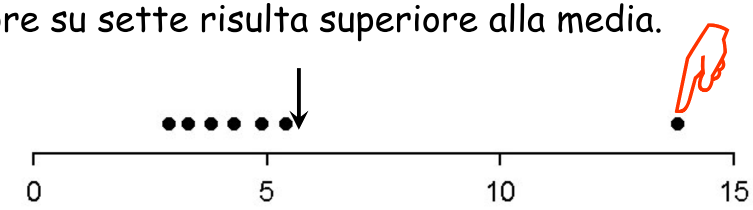
La variabilità del reddito dei miliardari risulta trascurabile rispetto all'ordine di grandezza del fenomeno.

Mediana

Durata del travaglio (ore) per il secondo parto naturale in 7 donne in eta' 30-33 anni

{3.8, 3.3, 2.9, 4.3, 13.8, 5.4, 4.9}

la media di 5.49 ore non è un valore tipico dell'insieme di osservazioni, dato che un solo valore su sette risulta superiore alla media.



La **mediana** è quella modalità tale per cui l'insieme delle osservazioni risulta essere per metà inferiore e per metà superiore ad essa.

Es: Durata del travaglio (ore) $\{3.8, 3.3, 2.9, 4.3, 13.8, 5.4, 4.9\}$

Per calcolare la mediana di una **variabile quantitativa**:

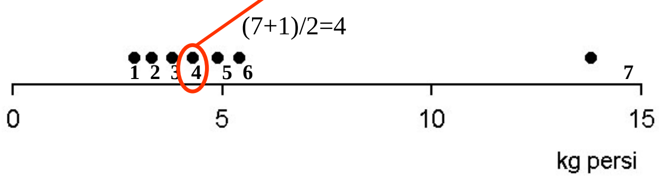
si individua quella modalità che è più grande di circa il 50% delle osservazioni e più piccola del restante 50%:

$$\{\underbrace{2.9, 3.3, 3.8,}_{\text{3 osservazioni}}\ 4.3,\ \underbrace{4.9, 5.4, 13.8}_{\text{3 osservazioni}}\}$$

Mediana da un insieme di dati enumerati

Per **n dispari**, la mediana è quel valore che occupa la posizione $\frac{n+1}{2}$ nell'insieme ordinato:

$$\{2.9, 3.3, 3.8, 4.3, 4.9, 5.4, 13.8\}$$

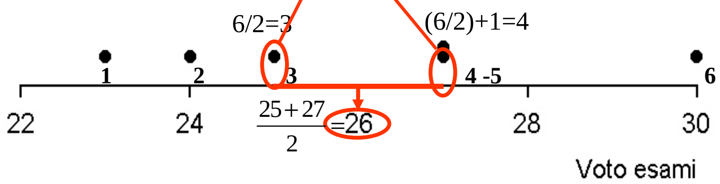


Mediana da un insieme di dati enumerati

Per **n pari**, la mediana è la media tra i valori nelle posizioni $\frac{n}{2}$ e $\left(\frac{n}{2}+1\right)$ nell'insieme ordinato:

Es: Voto esami

$$\{23, 24, 25, 27, 27, 30\}$$



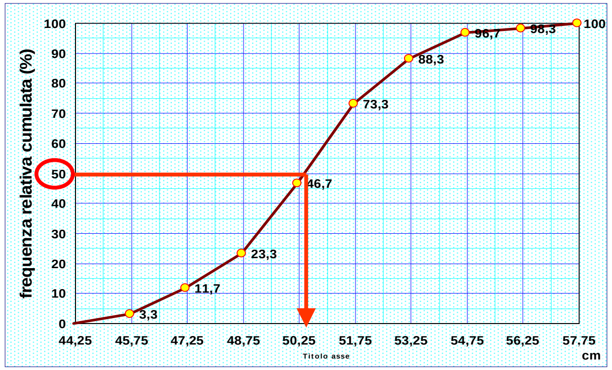
Mediana da una tabella di frequenza

Estremi di classe	Valore centrale	Freq. semplici f	Freq. semplici p%	Freq.cumulate F	Freq.cumulate P%
44.25 ⊣ 45.75	**45.0**	2	*3.3*	2	*3.3*
45.75 ⊣ 47.25	**46.5**	5	*8.3*	7	*11.7*
47.25 ⊣ 48.75	**48.0**	7	*11.7*	14	*23.3*
48.75 ⊣ 50.25	**49.5**	14	*23.3*	28	*46.7*
50.25 ⊣ 51.75	**51.0**	16	*26.7*	44	*73.3*
51.75 ⊣ 53.25	**52.5**	9	*15.0*	53	*88.3*
53.25 ⊣ 54.75	**54.0**	5	*8.3*	58	*96.7*
54.75 ⊣ 56.25	**55.5**	1	*1.7*	59	*98.3*
56.25 ⊣ 57.75	**57.0**	1	*1.7*	60	*100.0*

1) ci si può limitare alla **classe mediana**:

(50.25-51.75) .....oppure....

Mediana da un grafico delle frequenze cumulate

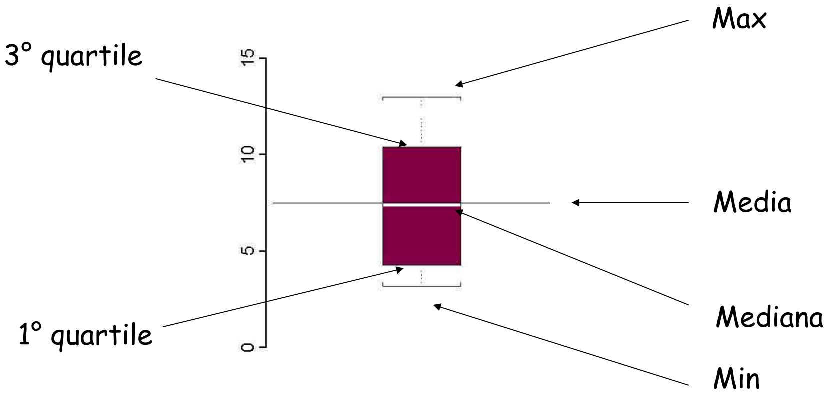


Mediana e Media

La mediana risulta essere meno sensibile della media alla presenza di dati anomali.

	Mediana	$\overline{x}$
{2.9, 3.3, 3.8, 4.3, 4.9, 5.4, 13.8}	4.3	5.49
{2.9, 3.3, 3.8, 4.3, 4.9, 5.4, 130.8}	4.3	**22.2**

Grafico Box-plot - scatola e baffi



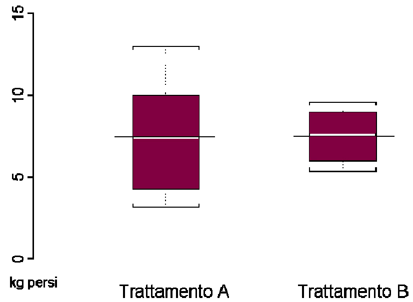
- Il box plot riassume 'posizione' e 'variabilita' del fenomeno
- E' adatto a rappresentare anche distribuzioni asimmetriche

Esempio: Diete

Kg persi con dieta (A) e farmaco (B)

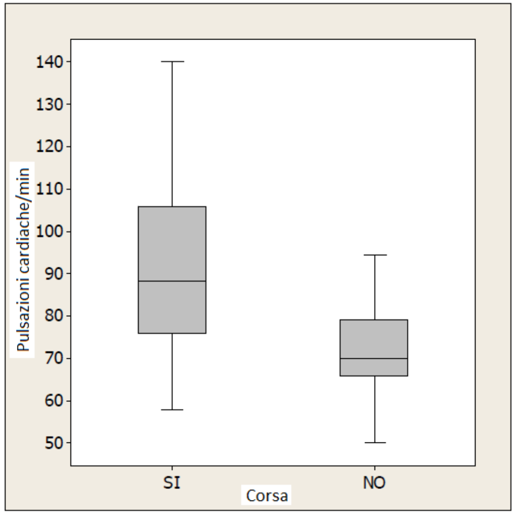
Tratt.	Media	Mediana
A	7.47	7.4
B	7.53	7.6

1) A e B sono sovrapponibili in termini di media e mediana

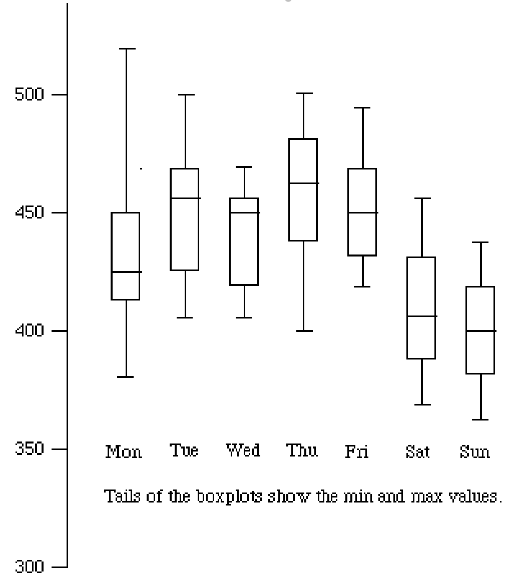


2) I dati relativi ad A sono più dispersi di quelli relativi a B

Pulsazioni cardiache e attività fisica



Nascite in ospedali canadesi



Moda

La **moda** è quella modalità/valore più frequente, che "va piu' di moda"

moda = Lunga

Tipi di gonna portate dalle donne italiane

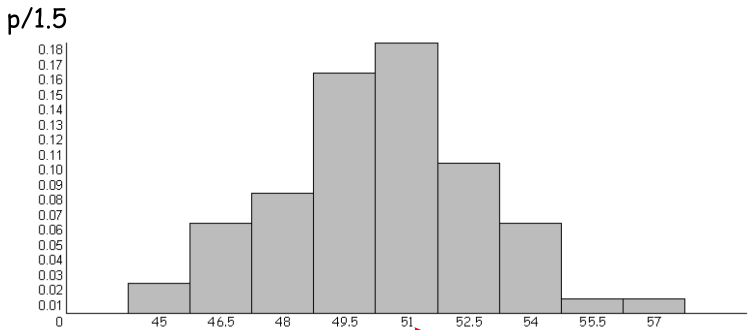
Mini	Corta	Midi	Longuette	Lunga	Tot.
120	57	70	87	230	564

moda = 27

Voto esami {23, 24, 25, 27, 27, 30}

Esempio: lunghezza bambini

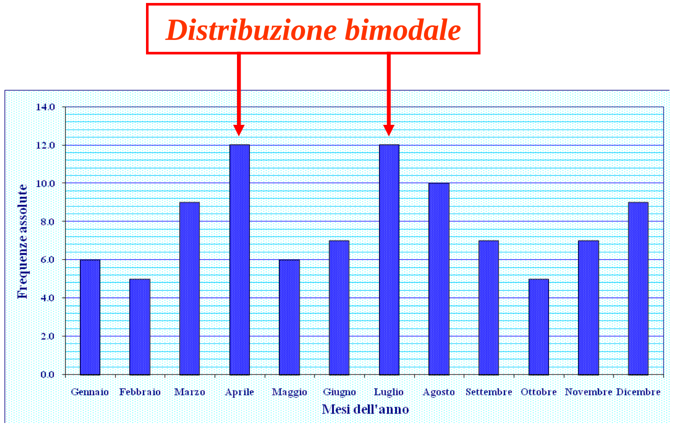
Lunghezza supina (cm) in un campione di 60 neonati.



La moda è la classe 50.25 -| 51.75 con una frequenza relativa di 26.7%

Esempio: mese di nascita

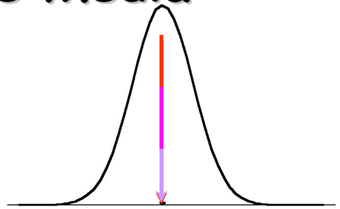
Mese di nascita

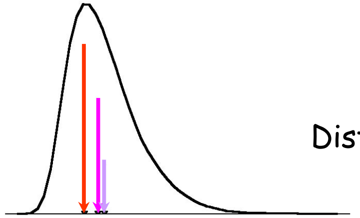


Posizione relativa di Moda, Mediana e Media

Distribuzione simmetrica

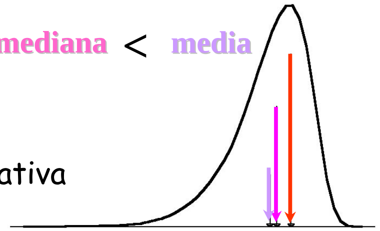
moda = **mediana** = **media**





Distribuzione con asimmetria positiva

moda < **mediana** < **media**

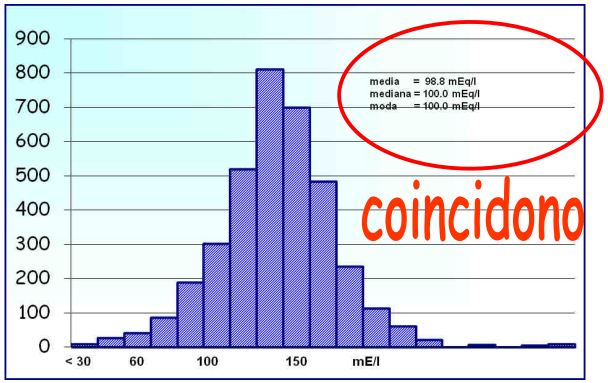


Distribuzione con asimmetria negativa

media < **mediana** < **moda**

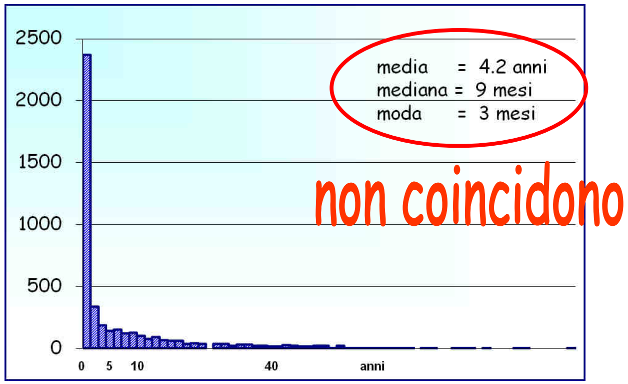
Esempio

Concentrazione di cloro nel sudore (simmetrica)



Esempio

Eta' alla diagnosi nella fibrosi cistica (asimmetria positiva)



INDICI DI POSIZIONE
Tabella riassuntiva

	Moda	Mediana	Media Aritmetica
Var. quantitativa Continua/Discreta	✓	✓	✓
Var. qualitativa Ordinale	✓	✓	
Var. qualitativa Nominale	✓		

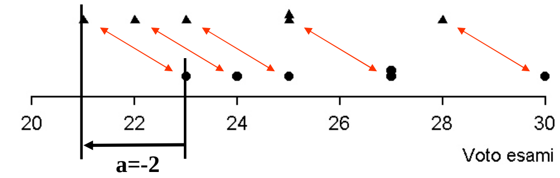
Proprietà algebriche di media e varianza (1)

Spostamento dell'origine $x_i \rightarrow x_i + a$

la media di aumenta di *a*

la deviazione standard non cambia

(ogni valore conserva la medesima distanza dalla media)



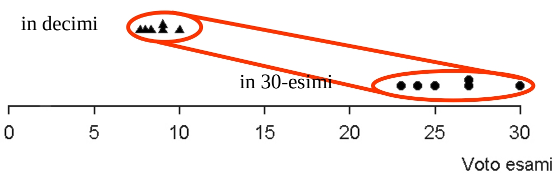
Proprietà algebriche di media e varianza (2)

Cambiamento di scala $x_i \rightarrow x_i \cdot a$

la media viene moltiplicata per *a*

la deviazione standard viene moltiplicata per *a*

(la distribuzione si allunga/restringe proporzionalmente ad *a*)



Esercizio per lo studente

Lunghezza supina (cm) in un campione di 60 neonati

Estremi di classe	Valore centrale	Freq. semplici		Freq.cumulate	
		f	p%	F	P%
44.25 ⊣ 45.75	**45.0**	2	*3.3*	2	*3.3*
45.75 ⊣ 47.25	**46.5**	5	*8.3*	7	*11.7*
47.25 ⊣ 48.75	**48.0**	7	*11.7*	14	*23.3*
48.75 ⊣ 50.25	**49.5**	14	*23.3*	28	*46.7*
50.25 ⊣ 51.75	**51.0**	16	*26.7*	44	*73.3*
51.75 ⊣ 53.25	**52.5**	9	*15.0*	53	*88.3*
53.25 ⊣ 54.75	**54.0**	5	*8.3*	58	*96.7*
54.75 ⊣ 56.25	**55.5**	1	*1.7*	59	*98.3*
56.25 ⊣ 57.75	**57.0**	1	*1.7*	60	*100.0*

- Rappresentare graficamente il fenomeno mediante un boxplot