STATISTICA DESCRITTIVA

Le misure di tendenza centrale

OBIETTIVO

Individuare un indice che rappresenti significativamente un insieme di dati statistici.

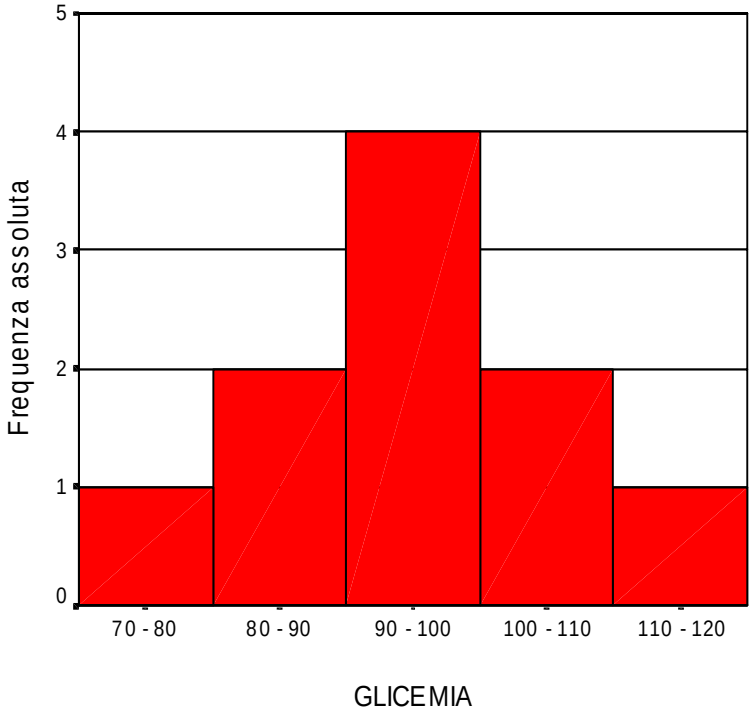
Esempio Nella tabella seguente sono riportati i valori del tasso glicemico rilevati su 10 pazienti:

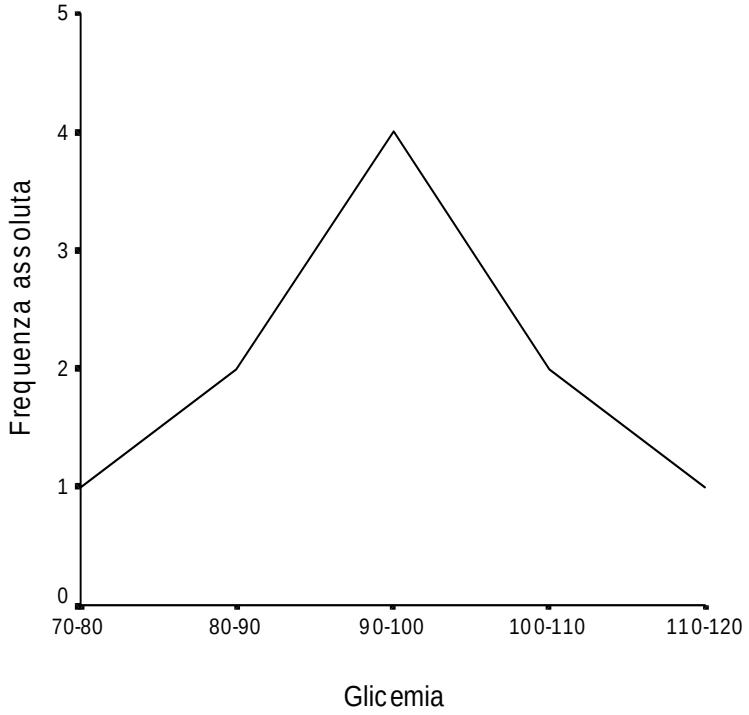
Paziente	Glicemia (mg/100cc)
1	$x_1=103$
2	$x_2=97$
3	$x_3=90$
4	$x_4=119$
5	$x_5=107$
6	$x_6=71$
7	$x_7=94$
8	$x_8=81$
9	$x_9=92$
10	$x_{10}=96$
Totale	**950**

Calcolo delle frequenze di ogni classe: assolute e relative percentuali

Classi di valori di glicemia	Frequenza assoluta	Frequenza relativa
70 —| 80	1	1 / 10 ·100% = 10 %
80 —| 90	2	2 / 10 ·100% = 20 %
90 —| 100	4	4 / 10 ·100% = 40 %
100 —| 110	2	2 / 10 ·100% = 20 %
110 —| 120	1	1 / 10 ·100% = 10 %
Totale	**10**	**100 %**

Costruzione dell'istogramma e del poligono di frequenza





LE MISURE DI POSIZIONE

- ✓ media aritmetica;
- ✓ mediana;
- ✓ moda;
- ✓ media armonica;
- ✓ media geometrica.

LA MEDIA ARITMETICA

DEFINIZIONE: La **media aritmetica** è quel valore che avrebbero tutte le osservazioni se non ci fosse la variabilità (casuale o sistematica).

Più precisamente, è quel valore che sostituito a ciascun degli **n** dati ne fa rimanere costante la somma.

dato un insieme di n elementi $\{x_1, x_2, \ldots x_n\}$

Si dice **media aritmetica semplice** di n numeri il numero che si ottiene dividendo la loro somma per n.

$$\bar{x} = \frac{x_1 + x_2 + \ldots x_n}{n}$$

Formalmente possiamo esprimere la media aritmetica semplice attraverso la seguente formula:

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

Nell’Esempio in esame si ha:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{950}{10} = 95\, mg\ /100\, cc$$

Esempio Riportiamo i tempi di sopravvivenza (mesi) di 19 pazienti con cancro dell'addome

Mesi di sopravvivenza (x_i)	Frequenza (f_i)	$x_i \cdot f_i$
8,5	2	17
9,2	4	36,8
7,3	8	58,4
6,8	2	13,6
10,1	3	30,3
Totale	19	156,1

Si dice **media aritmetica pesata** di n numeri:

$$\frac{x_1 \cdot p_1 + x_2 \cdot p_2 + \ldots + x_m \cdot p_m}{p_1 + p_2 + \ldots + p_m}$$

Dove i pesi p_j sono le frequenze assolute di ogni modalità

Nell’esempio precedente la media aritmetica (ponderata) è data da:

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i} = \frac{156\ ,1}{19} = 8,2$$

Proprietà della media aritmetica:

- ✓ compresa tra il minimo dei dati e il massimo dei dati;
- ✓ $\sum_i (x_i - \overline{x}) = 0$ la somma degli scarti dalla media è zero;
- ✓ $\sum_i (x_i - z)^2$ assume valore minimo per z = media aritmetica;
- ✓ la media dei valori: $k \cdot x_i$ è pari a la media aritmetica $\cdot$ k (dove k è un numero reale qualsiasi)
- ✓ la media dei valori: $x_i \pm h$ è pari a: media aritmetica$\pm$h (dove h è un numero reale qualsiasi).

Lunghezza(cm) in un campione di 66 neonati

55.9	51.3	53.0	50.5	54.9	53.4	53.7	50.0	53.8	52.5	55.6
47.9	54.3	56.0	51.8	54.1	55.6	57.6	53.3	51.1	54.3	52.3
55.3	52.4	56.3	53.7	54.4	54.5	52.5	52.7	51.4	55.5	52.7
57.4	51.7	50.8	49.4	52.0	53.7	54.8	53.5	49.5	50.4	56.4
48.5	53.1	49.5	53.2	53.1	52.6	54.3	54.9	53.7	55.2	51.7
51.4	51.0	52.6	52.8	59.3	56.4	51.5	58.9	52.3	54.6	53.8

la media aritmetica dei 66 valori di lunghezza è:

=(55.9+51.3+53.0+50.5+54.9+53.4+…+53.8)/66

= 3517.500/66

= 53.295

MEDIA per dati raggruppati in classi

x_i	f_i	%	$X_i f_i$
48.0	2	3.03	96.00
49.5	3	4.55	148.50
51.0	12	18.18	612.00
52.5	15	22.73	787.50
54.0	14	21.21	756.00
55.5	10	15.15	555.00
57.0	5	7.58	285.00
58.5	4	6.06	234.00
60.0	1	1.52	60.00
Somma	66	100	3534.00

Nell'*esempio* del campione di 66 misure di lunghezza dei neonati:

$$\bar{x} = \frac{48.0 \times 2 + 49.5 \times 3 + \ \ldots\ 60.0 \times 1}{2 + 3 + \ \ldots\ 1} = \frac{3534.0}{66} = 53.545$$

La **media aritmetica** è la misura di posizione più usata ma. A volte, altre misure come la **mediana** e la **moda** si dimostrano utili.

Si consideri un campione di valori di VES (*velocità di eritrosedimentazione*, mm/ora) misurati in 7 pazienti

{8, 5, 7, 6, 35, 5, 4}

In questo caso, la media che è = 10 mm/ora non è un valore tipico della distribuzione: soltanto un valore su 7 è superiore alla media!

Limite della media aritmetica:

è notevolmente influenzata dai valori estremi della distribuzione.

Esempio Età alla morte di 5 soggetti

$x_1 = 34$ anni; $x_2 = 70$ anni; $x_3 = 74$ anni;
$x_4 = 64$ anni; $x_5 = 68$ anni.

La media aritmetica è pari a:

$$\bar{x} = (34 + 70 + 74 + 64 + 68) / 5 = 62\, anni$$

LA MEDIANA

DEFINIZIONE: La **mediana (Me)** è quell'osservazione che bipartisce la distribuzione in modo tale da lasciare al "di sotto" lo stesso numero di termini che lascia al "di sopra".

L'idea che e alla base della **mediana** e di cercare un numero che sia più grande di un 50% delle osservazioni e più piccolo del restante 50%.

Ritornando all'Esempio della Glicemia, per il calcolo della mediana è necessario disporre i dati in ordine crescente:

71, 81, 90, 92, 94, 96, 97, 103, 107, 119

$$Me = (94+96)/2 = 95 \text{ mg/100 cc}$$

Il fatto che mediana e media aritmetica in questo caso coincidano non è casuale in quanto la distribuzione è simmetrica.

Ma, in generale, ciò non avviene.

Vantaggio nell’uso della mediana: non è influenzata dalle osservazioni aberranti o estreme.

Le fasi operative per il calcolo della mediana sono le seguenti:

1) ordinamento crescente dei dati;

2) se il numero di dati n è dispari, la mediana corrisponde al dato che occupa la (n+1)/2 esima posizione

3) se il numero di dati n è pari, la mediana è data dalla media aritmetica dei due dati che occupano la posizione n/2 e quella n/2+1.

In presenza di una distribuzione di frequenze è necessario considerare le frequenze cumulate

Voti ordinati (x_i)	Frequenze (f_i)	Freq. Cum. (F_i)	Freq.Cum. (F_i%)
18	2 (10.5)	2	10.5
20	4 (21.0)	2+4 = 6	31.5
22	8 (42.1)	4+8 = 14	73.6
24	2 (10.5)	14+2 = 16	84.1
27	2 (10.5)	16+2 = 18	94.6
30	1 (5.4)	18+1 = 19	100
Totale	**19**	**19**	

Voti ordinati	Frequenze	Freq.Cum. F_i	Freq.Cum. F_i%
18	2 (10.5)	2	10.5
20	4 (21.0)	6	31.5
22	8 (42.1)	14	73.6
	2 (10.5)	16	84.1
	2 (10.5)	18	94.6
	1 (5.4)	19	100
	2 (10.5)	**19**	

La Mediana

I QUANTILI

✓Generalizzano la mediana.

✓L'idea alla base di un **quantile-p** dove $p \in$ [0; 1] e di cercare un numero che sia più grande *p*% dei dati osservati e più piccolo del restante (1-p%) dei dati

I quantili con *p* uguale a 0;25, 0;50 e 0;75 vengono chiamati rispettivamente il primo, il secondo e il terzo **quartile**.
Dividono la popolazione in quattro parti uguali.
Si osservi che il 2 quartile coincide con la mediana.

I quantili con *p* = 0,01; … ; 0,99 si chiamano **percentili**.

LA MODA

DEFINIZIONE: La **Moda (Mo)** è l'osservazione che si verifica con maggiore frequenza in una data distribuzione.

Si possono avere anche più valori modali.

quale misura di posizione usare?

A quale misura di tendenza centrale ci riferiamo?

- Il proprietario di una ditta afferma "Lo stipendio mensile nella nostra ditta è 2.700 euro"
- Il sindacato dei lavoratori dice che “lo stipendio medio è di 1.700 euro”.
- L'agente delle tasse dice che “lo stipendio medio è stato di 2.200 euro”.

Queste risposte diverse sono state ottenute tutte dai dati della seguente tabella.

Stipendio mensile	N° di lavoratori
1.300	2
1.700	22
2.200	19
2.600	3
6.500	2
9.400	1
23.000	1

Media aritmetica	= lire 2.700
Mediana	= lire 2.200
Moda	= lire 1.700

interpretazione delle misure di posizione

- La **media aritmetica** indica che, se il denaro fosse distribuito in modo che ciascuno ricevesse la stessa somma, ciascun dipendente avrebbe avuto 2.700 euro

- La **moda** ci dice che la paga mensile più comune è di 1.700.euro

- La moda si considera spesso come il valore tipico dell'insieme di dati poiché è quello che si presenta più spesso. **Non tiene però conto degli altri valori** e spesso in un insieme di dati vi è **più di un valore** che corrisponde alla definizione di moda.

- La **mediana** indica che circa metà degli addetti percepiscono meno di 2.200.euro, e metà di più.

- La mediana **non è influenzata dai valori estremi** eventualmente presenti ma solo dal fatto che essi siano sotto o sopra il centro dell'insieme dei dati.

Relazione tra media, mediana e moda

In una distribuzione perfettamente simmetrica, la media, la mediana e la moda hanno lo stesso valore. In una distribuzione asimmetrica, la media si posiziona nella direzione dell'asimmetria. Nelle distribuzioni di dati biologici, l'asimmetria è quasi sempre verso destra (asimmetria positiva, verso i valori più elevati), e quindi la media è > della mediana o della moda

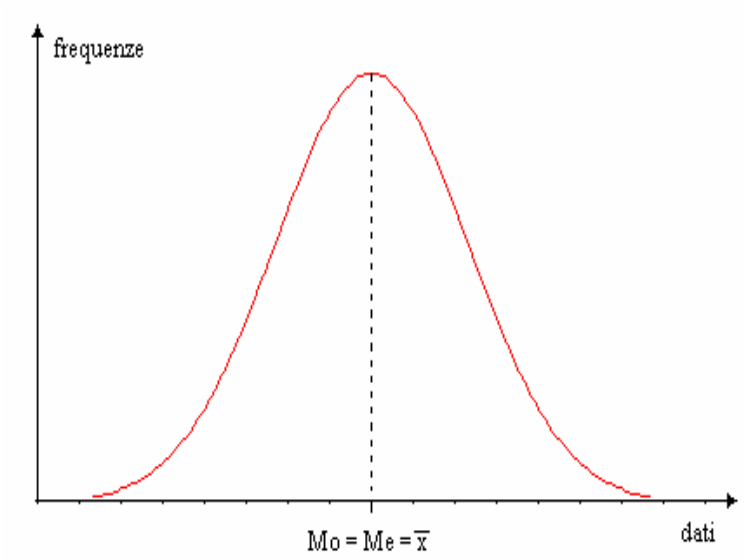


Fig. I - Curva Simmetrica

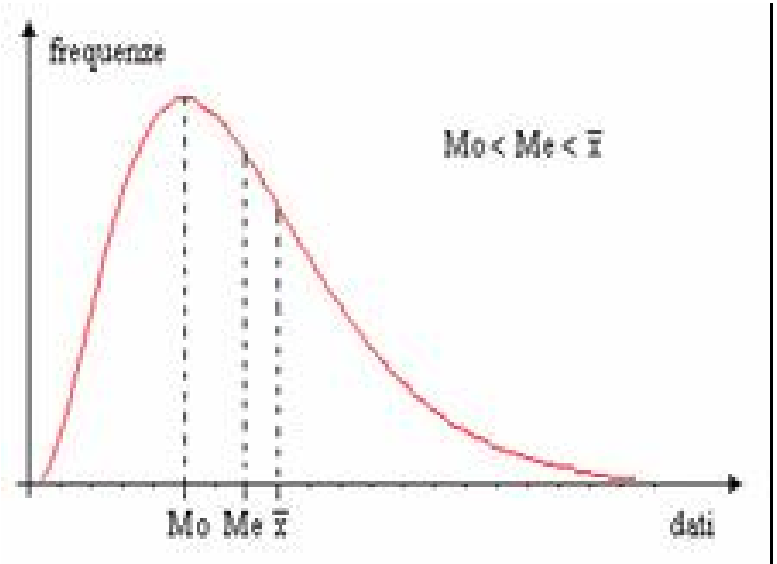


Fig. II - Curva asimmetrica a destra

STATISTICA DESCRITTIVA

Le misure di variabilità

In assenza di variabilità in una popolazione la statistica non sarebbe necessaria: un singolo *elemento* o unità campionaria sarebbe sufficiente a determinare tutto ciò che occorre sapere su una popolazione. Ne consegue, perciò, che nel presentare informazioni su un campione non è sufficiente fornire semplicemente una misura della *media ma servono informazioni sulla variabilità*.

Esempio Si considerino inizialmente, le seguenti due distribuzioni di valori riferiti all'età di 10 individui:

Soggetti	I gruppo	II gruppo
1 2 3 4 5	20aa 30aa 40aa 50aa 60aa	10aa 25aa 40aa 55aa 70aa
Tot	200aa	200aa
Media Aritmetica	200aa/5=40aa	200aa/5=40aa

LE MISURE DI VARIABILITÀ

- ✓ Campo di variazione (range);
- ✓ Devianza;
- ✓ Varianza;
- ✓ Deviazione Standard;
- ✓ Coefficiente di variazione (variabilità relativa).

IL CAMPO DI VARIAZIONE O RANGE

DEFINIZIONE: Il **Campo di variazione o Range** corrisponde alla differenza fra la modalità più piccola e la modalità più grande della distribuzione

$$R = X_{max} - X_{min}$$

Limiti del campo di variazione:

- ✓ è troppo influenzato dai valori estremi;

- ✓ tiene conto dei due soli valori estremi, trascurando tutti gli altri.

Occorre allora un indice di dispersione che consideri tutti i dati (e non solo quelli estremi), confrontando questi con il loro valor medio.

1ª idea →

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

2ª idea →

$$\sum_{i=1}^{n} |x_i - \overline{x}|$$

3ª idea →

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

LA DEVIANZA

DEFINIZIONE: La somma dei quadrati degli scarti dalla media aritmetica

$$\sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2 f_i$$

Esempio 9. Valori del tasso glicemico in 10 soggetti

x_i (glicemia mg/100cc)	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
103	+8	64
97	+2	4
90	-5	25
119	+24	576
107	+12	144
71	-24	576
96	+1	1
$\bar{x} = 95$	94	1596

La quantità 1596 esprime la *Devianza della distribuzione* (Dev).

LA VARIANZA

DEFINIZIONE: La somma dei quadrati degli scarti dalla media aritmetica divisi per la numerosità

$$\sum_{i=1}^{n} \left(x_i - \overline{x} \right)^2 f_i / N$$

LA DEVIAZIONE STANDARD

DEFINIZIONE: La radice quadrata della varianza

$$\sqrt{\frac{\sum_{i=1}^{N} (x_i - \overline{x})^2 \, f_i}{n - 1}}$$

Calcolare la deviazione standard (DV) delle seguenti 10 osservazioni (mm):

81 79 82 83 80 78 80 87 82 82

1. Si calcoli la media, $\bar{x}$:

$$\bar{x} = \frac{\sum x}{n} = \frac{814}{10} = 81.40$$

2. Si calcolino gli scarti dalla media sottraendo da ciascun valore la media; si elevi al quadrato tale quantità (il quadrato elide il segno -):

(81-81.4)2= 0.16 (78-81.4)2= 11.56

(79-81.4)2= 5.76 (80-81.4)2= 1.96

(82-81.4)2= 0.36 (87-81.4)2= 31.36

(83-81.4)2= 2.56 (82-81.4)2= 0.36

(80-81.4)2= 1.96 (82-81.4)2= 0.36

3. Si sommino tali quantità: la somma è pari a 56.4. La somma $\Sigma (x - \overline{x})^2$ è detta somma dei quadrati degli scarti o, più semplicemente, somma dei quadrati.

4. Si divida tale quantità per il numero di osservazioni meno 1:

$$\frac{\text{sommadei quadrati}}{(n-1)} = \frac{\sum\left(x-\bar{x}\right)^2}{n-1} = \frac{56.4}{9} = 6.27$$

5. La deviazione standard è la radice quadrata di tale valore:

$$DS = \sqrt{6.27} = 2.50\,\text{mm}$$

Quindi la deviazione standard del campione di 10 unità estratto dalla popolazione è pari a 2.50 mm.

SCARTO INTERQUARTILE

Scarto interquartile = (3° quartile)-(1° quartile)

E' molto più *resistente* della varianza in presenza di poche osservazioni estreme. Per questo motivo e usato soprattutto nelle situazioni in cui si sospetta la possibile presenza di osservazioni anomale.

IL COEFFICIENTE DI VARIAZIONE

$$C.V. = \frac{\text{(deviazione standard)}}{\text{(media aritmetica)}}$$

La variabilità guarda alle differenze tra le unità sperimentali. E' pero evidente che il significato pratico delle differenze può dipendere dal livello del fenomeno considerato.
Può quindi essere interessante disporre di una qualche misura di variabilità *aggiustata* in qualche maniera per tenere conto del livello del fenomeno.

Esempio

Data la media e la deviazione standard di campioni di (a) neonati, (b) bambini di tre anni e (c) bambini di 10 anni, dobbiamo chiederci se la **variabilità relativa** si modifica con l'età.

(a) Neonati $\overline{x}$ = 3,1 Kg; DS = 0,23 Kg

CV = 0,23/3,1× 100 = 7,4%

(b) Bambini di 3 anni $\overline{x}$ = 16,0 Kg; DS = 4,5 Kg

CV = 4,5/16,0 × 100 = 28,1 %

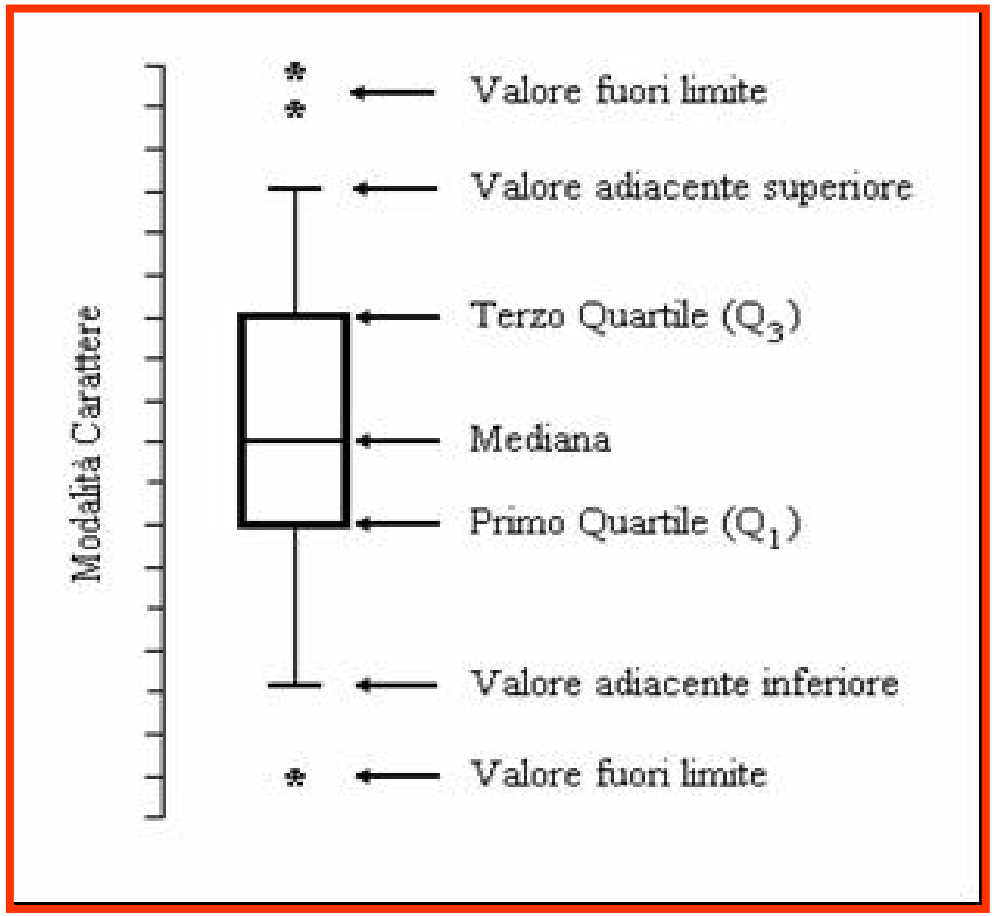
(b) Bambini di 10 anni $\overline{x}$ = 35,0 Kg; DS = 13,8 Kg

CV = 13,8/35,0 × 100 = 39,4 %

Osservando i tre valori del **CV**, si può notare che la **variabilità relativa** aumenta con l'età.

BOX-PLOT

Il nome deriva dall'inglese (*box and whiskers plot* spesso, anche in italiano, abbreviato in *boxplot*).



INDICI DI SIMMETRIA

Distribuzione *simmetrica*:

le osservazioni equidistanti dalla mediana (coincidente in questo caso col massimo centrale) presentano la stessa frequenza relativa

Un esempio importante è fornito dalla curva di ***distribuzione normale***

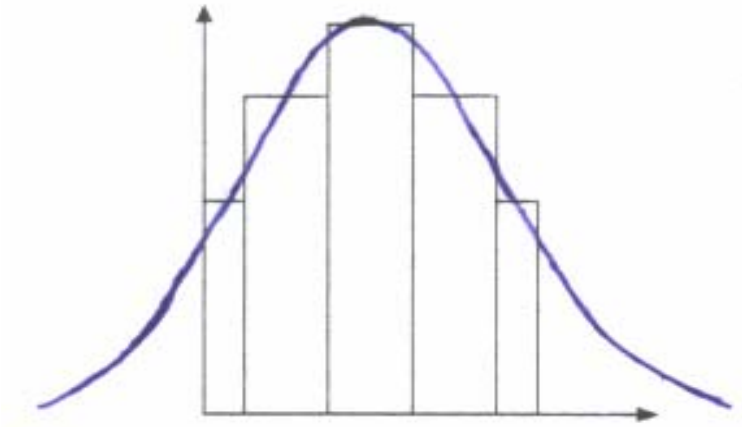


Fig. I - Media = Mediana

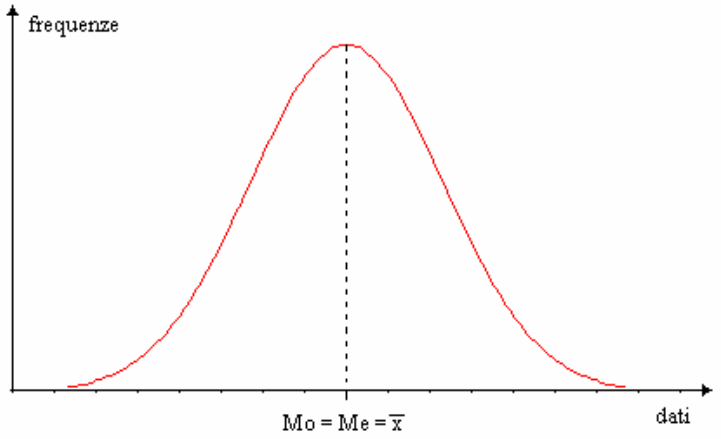


Fig. I - Curva Simmetrica

Distribuzione ***asimmetrica positiva***:

la curva di frequenza ha una *coda* più lunga a destra del massimo centrale.

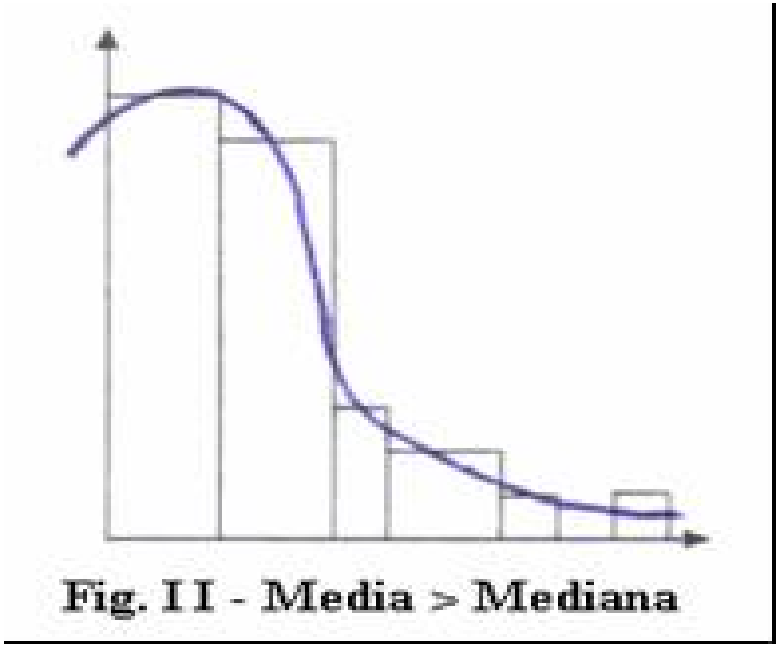


Fig. I I - Media > Mediana

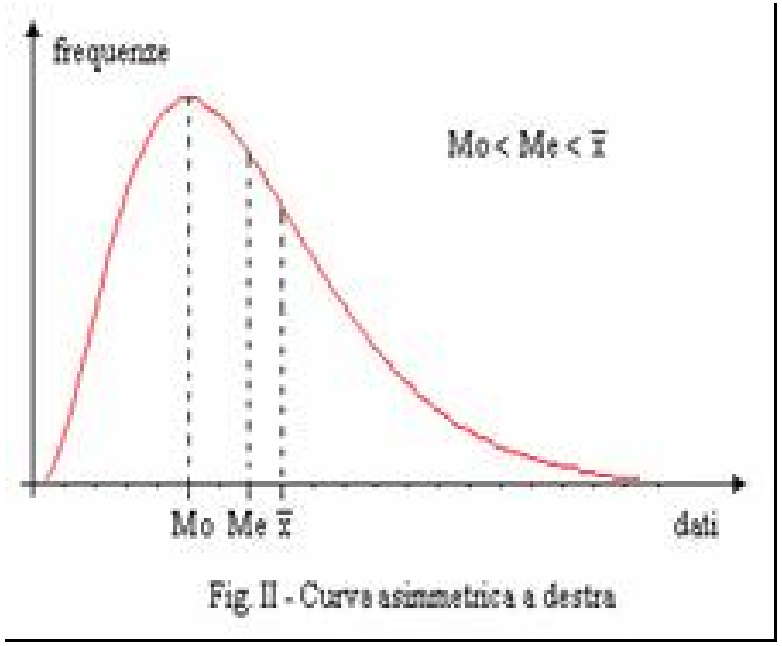


Fig. II - Curva asimmetrica a destra

Distribuzione ***asimmetrica negativa***:

la curva di frequenza ha una *coda* più lunga a sinistra del massimo centrale

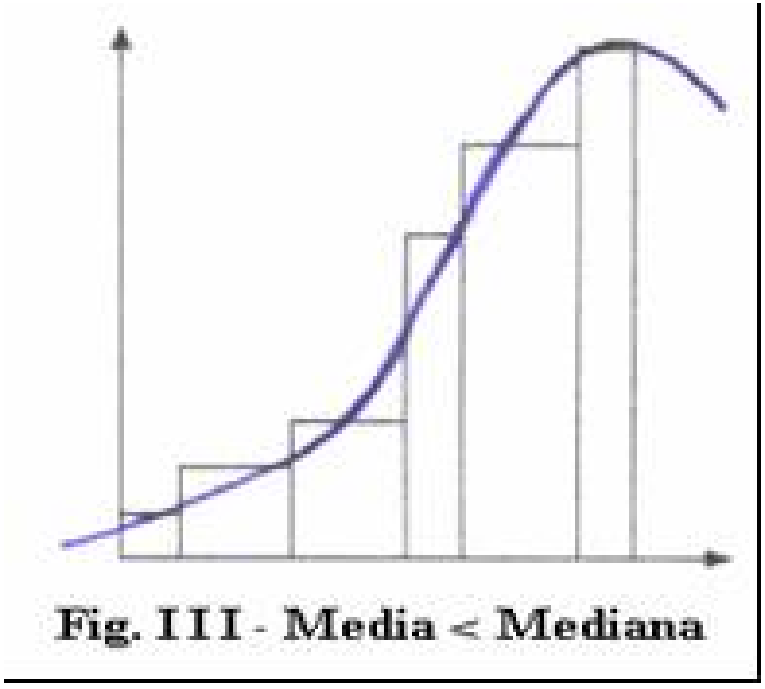


Fig. III - Media < Mediana

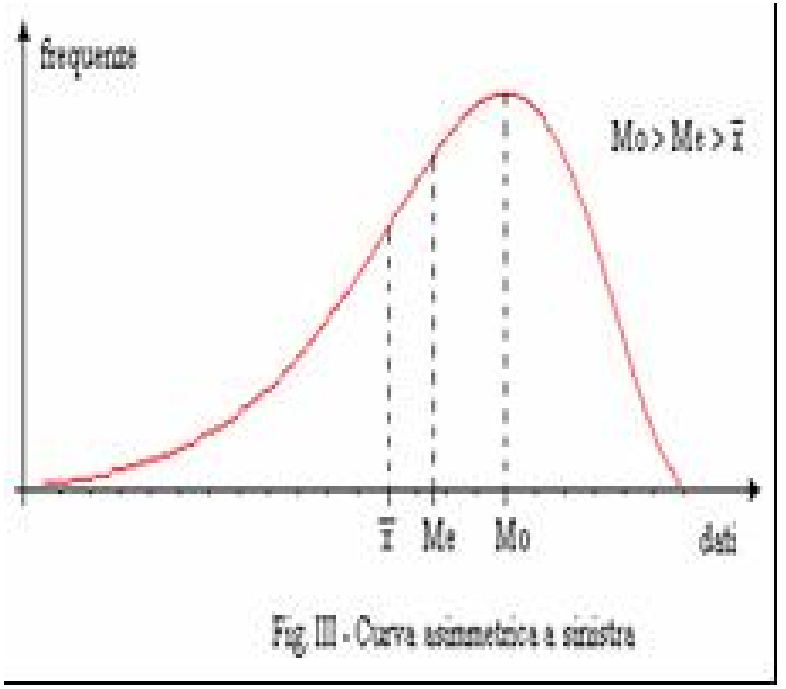


Fig. III - Curva asimmetrica a sinistra