MISURE DI TENDENZA CENTRALE

Psicometria 1 - Lezione 2
Lucidi presentati a lezione

AA 2000/2001 dott. Corrado Caudek

Supponiamo di disporre di un insieme di misure e di cercare un solo valore che, meglio di ciascun altro, sia in grado di “catturare” le caratteristiche della distribuzione nel suo complesso, ovvero rappresenti il valore più tipico della distribuzione.

In altre parole, ci proponiamo di trovare il valore che, meglio di qualunque altro, ci consenta di indovinare il valore di una delle unità di osservazione che fosse tratta a caso dall’insieme delle nostre osservazioni.

Moda, Mediana, Media aritmetica, Media geometrica

Media armonica

MODA

La moda di una distribuzione di frequenze è il punto centrale della classe di misure più frequente.

Distribuzione zeromodale: nessun valore ha una frequenza più elevata degli altri.

Distribuzione unimodale: c'è un solo valore con una frequenza più elevata degli altri. Es. [2, 4, 1, 3, 7, 3, 5, 3]

Distribuzione bimodale: ci sono due valori con una frequenza più elevata degli altri. Es. [7, 4, 7, 3, 7, 3, 5, 3]

Negli istogrammi, la moda coincide con il punto centrale della base del rettangolo con altezza maggiore.

Nelle curve di frequenza, la moda coincide con il valore corrispondente ad un massimo della curva (Fig. 1.17, p. 41).

Affinché una distribuzione sia bimodale è sufficiente che vi siano due massimi. Non è necessario che entrambi abbiano lo stesso valore.

<u>Dato che la moda dipende soltanto dalla frequenza delle osservazioni, è l'unica misura di tendenza centrale per dati in scala nominale.</u>

LIMITI DELLA MODA

Un campione può avere più di una moda.

La moda è molto sensibile alla grandezza e al numero degli intervalli di classe. La moda può cambiare in maniera considerevole cambiando gli intervalli delle classi.

La moda di un campione non fornisce una buona stima della moda della popolazione da cui quel campione è stato tratto.

MEDIANA

La mediana è il valore che occupa la posizione centrale quando le osservazioni di un campione sono **ordinate** in base al loro valore.

Quando ***n* è dispari**, la mediana corrisponde al punteggio dell'individuo numero ***(n + 1)/2.***

Quando n è pari, la mediana corrisponde al valore intermedio tra il punteggio dell'individuo numero ***n /2*** e il punteggio dell'individuo ***(n / 2) + 1***.

6, 6.7, 3.8, 7, 5.8

I valori ordinati sono

3.8, 5.8, 6, 6.7, 7

6, 6.7, 3.8, 7, 5.8, 9.975

I valori ordinati sono:

3.8, 5.8, 6, 6.7, 7, 9.975

2 middle values

$$\text{Median} = \frac{6 + 6.7}{2} = 6.35$$

MEDIA ARITMETICA

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \frac{X_i}{n}$$

Proprietà 1. Se un insieme di osservazioni è costituito da due sottoinsiemi disgiunti di grandezza n_1 e n_2, e medie $\bar{x}_1$ e $\bar{x}_2$ allora la media dell'insieme totale sarà uguale a:

$$\bar{X} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2}$$

Proprietà 2. La somma degli scarti di tutti i punteggi di un campione dalla media è uguale a zero.

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right) = 0$$

$$= \sum X_i - \sum \overline{X} = \sum X_i - n\overline{X} = n\overline{X} - n\overline{X} = 0$$

DIFFERENZE TRA LA MEDIA E LA MEDIANA

La media risente dei cambiamenti effettuati agli estremi di una distribuzione, mentre la mediana è insensibile a questi cambiamenti.

La media è più stabile della mediana, ovvero varia di meno al passare da un campione ad un altro.

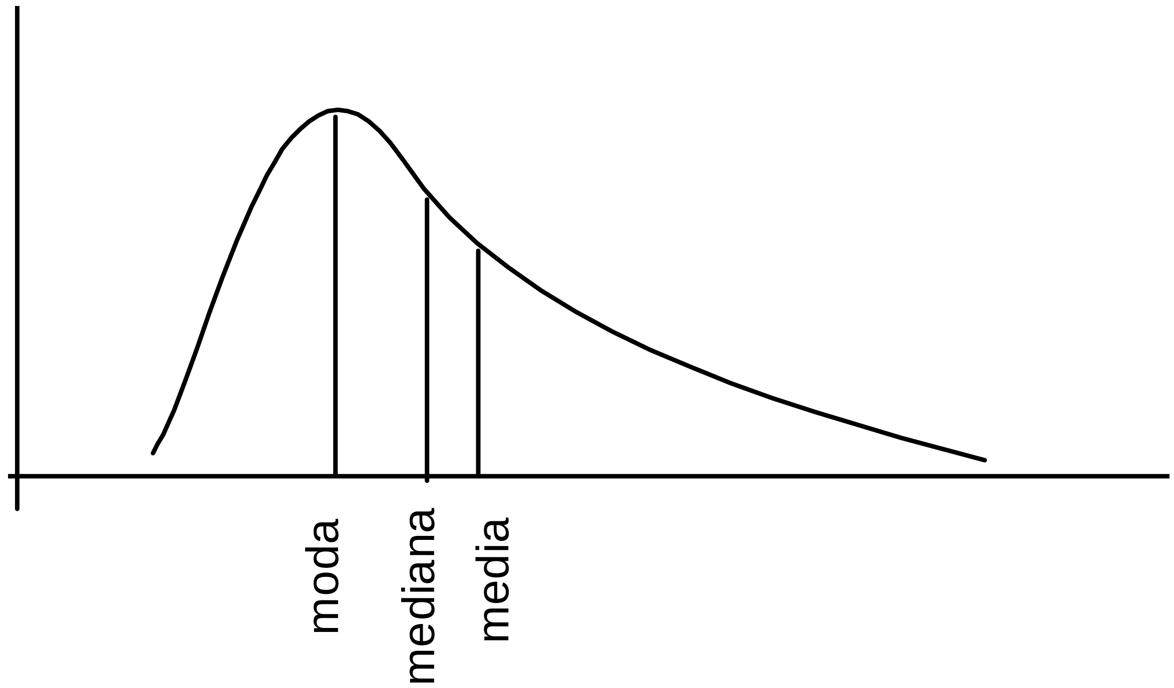
ESERCIZIO

Definite un insieme di numeri casuali di grandezza 3 e un secondo insieme di numeri casuali di grandezza 2. Dimostrate empiricamente la prima proprietà della media.

Definite un insieme di 5 numeri casuali, e dimostrate empiricamente la seconda proprietà della media.

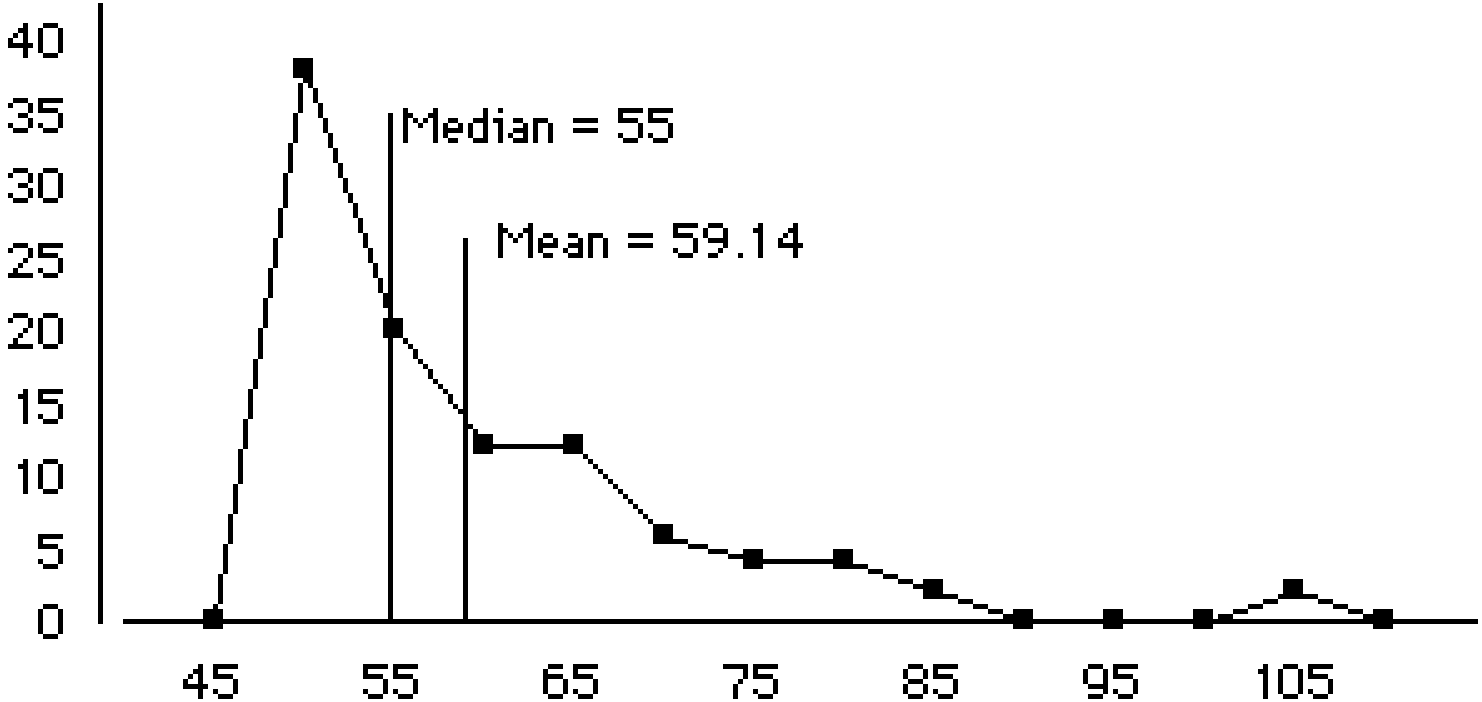
RELAZIONE TRA MISURE DI TENDENZA CENTRALE E CURVE DI FREQUENZA

ASIMMETRIA POSITIVA



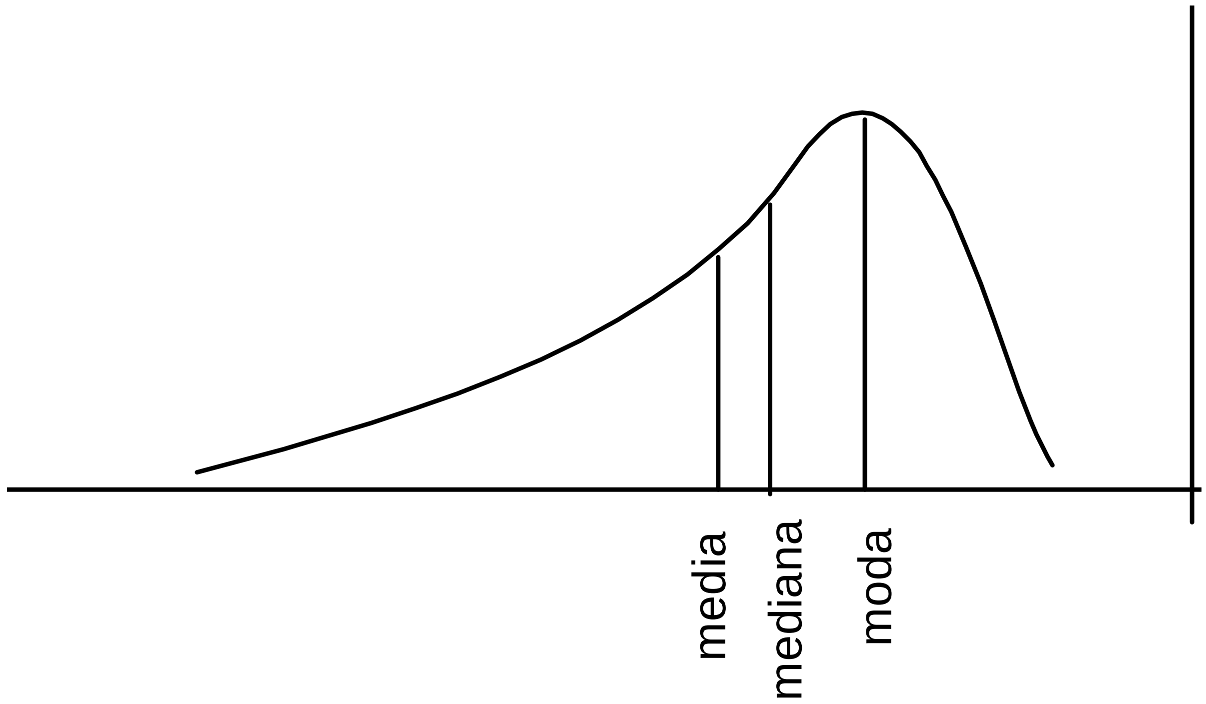
Nel caso di una distribuzione con asimmetria positiva

media > mediana > moda



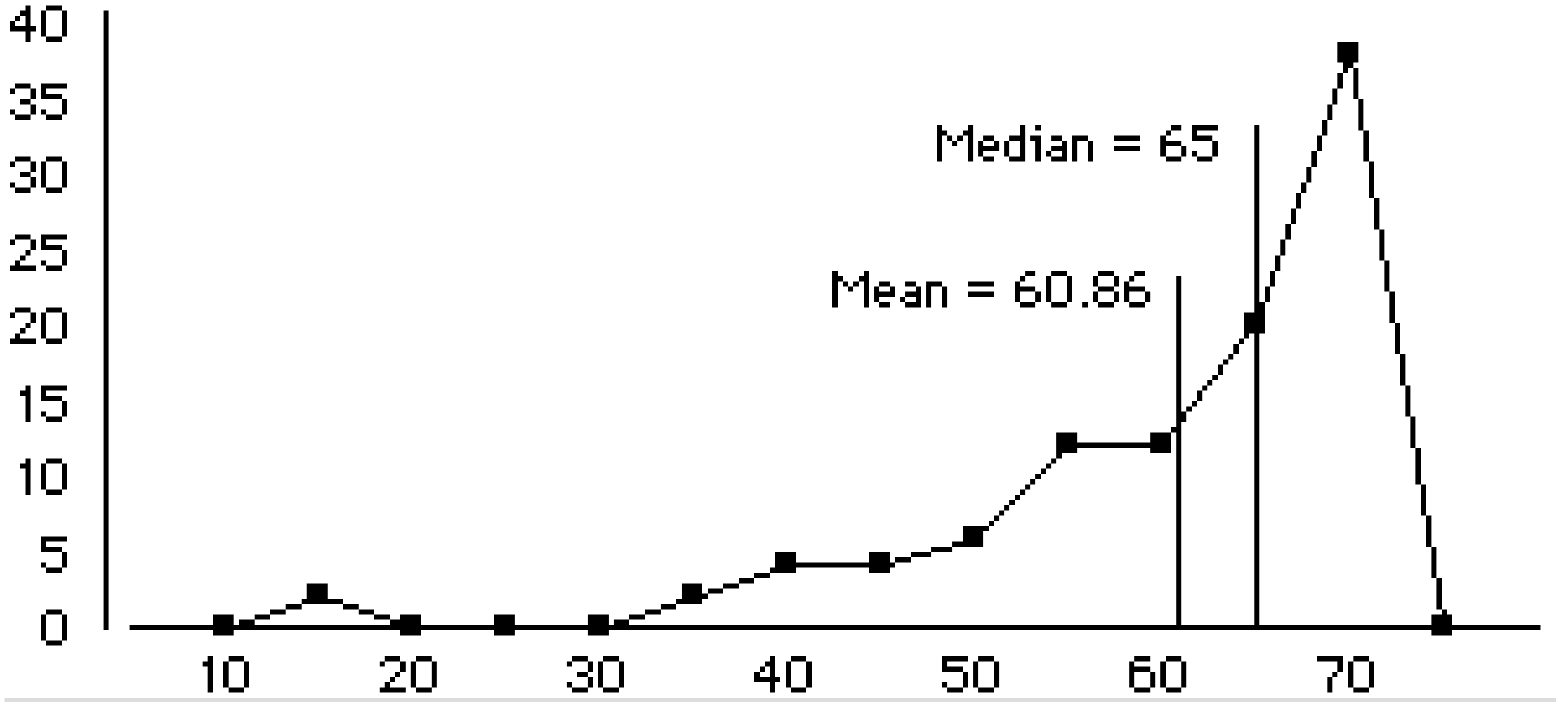
40
35
30
25
20
15
10
5
0
Median = 55
Mean = 59.14
45
55
65
75
85
95
105

ASIMMETRIA NEGATIVA

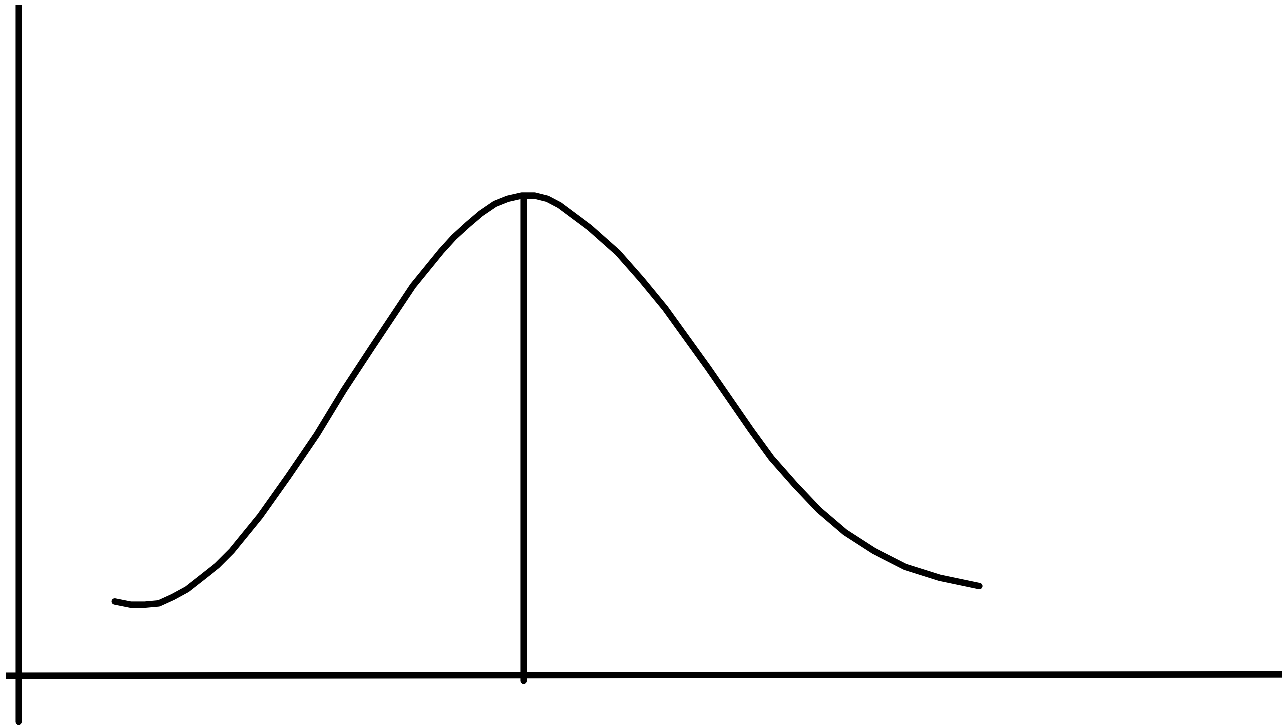


Nel caso di una distribuzione con asimmetria negatuva

media < mediana < moda


40
35
30
25
20
15
10
5
0
10
20
30
40
50
60
70
Median = 65
Mean = 60.86

SIMMETRIA



moda = mediana = media

ESERCIZIO

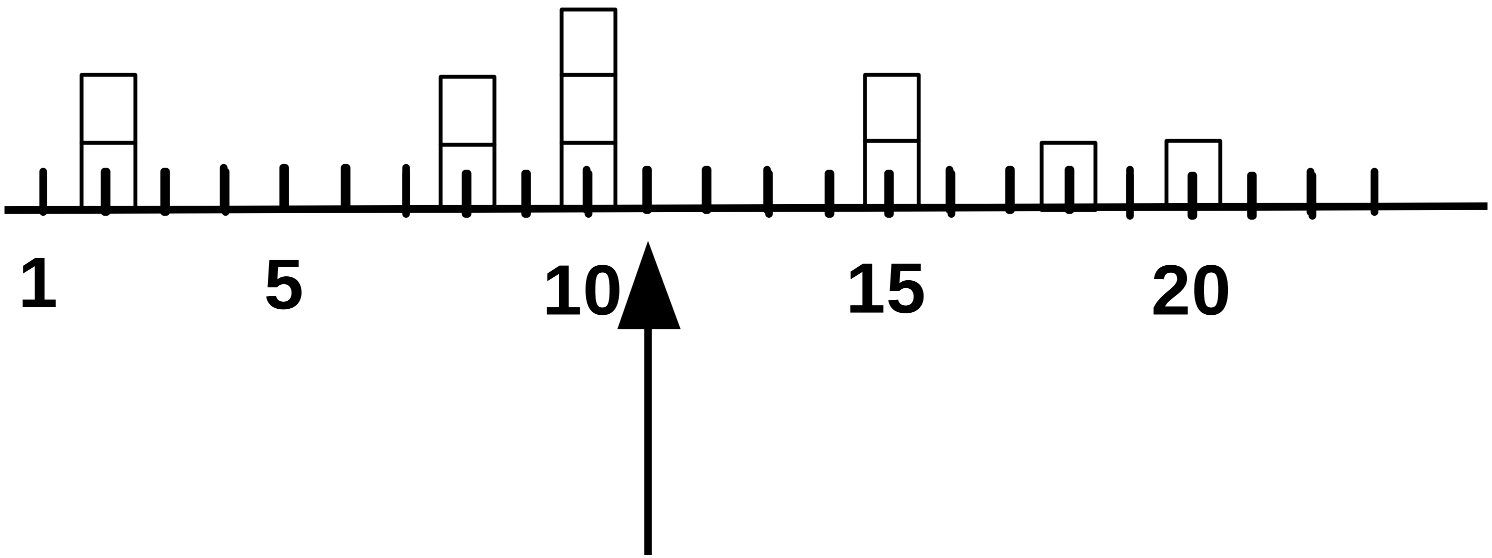
Generate un insieme di numeri di grandezza uguale a 6, tali da costituire una distribuzione con asimmetria positiva, con asimmetria negativa, oppure una distribuzione simmetrica.

Calcolate la *moda*, la *media* e la *mediana* nei tre casi.

La media di una distribuzione può essere accostata ai concetti fisici di *punto di equilibrio* o *centro di gravità*.

2, 2, 8, 8, 10, 10, 10, 15, 15, 18, 20

$$\overline{X} = 11$$



Dato che la somma delle deviazioni dalla media è uguale a zero, se usiamo la media per indovinare il valore di un'osservazione tratta a caso dal campione, allora <u>la media degli errori segnati delle nostre stime sarà uguale a zero</u>.

MEDIA GEOMETRICA

$$G = \sqrt[n]{\prod_{i=1}^{n} X_i}$$

$$\log G = \frac{\sum_{i=1}^{n} \log X_i}{n}$$

La media geometrica è più sensibile della media aritmetica alle variazioni di intensità più piccole, e meno sensibile alle variazioni di intensità più elevate (es. p. 51).

…. ovviamente, può essere usata solo per valori x_i positivi.

INDICI DI POSIZIONE

INDICI DI POSIZIONE

Quantili

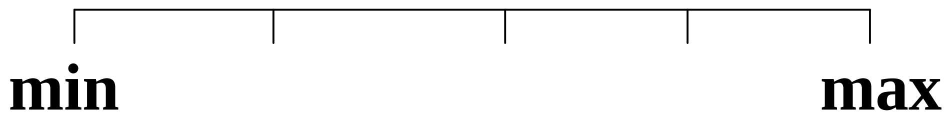
Ranghi percentili

Punti z

Quantili

I quantili si riferiscono ad una suddivisione in parti uguali dei dati ordinati.

Una suddivisione dei dati ordinati in 4 classi aventi frequenze uguali produce i *quartili*.



Il ***primo quartile*** è l'unità di osservazione che ha la proprietà di avere sotto di sé un quarto dei dati della distribuzione.

Il ***secondo quartile*** è uguale alla mediana.

Una suddivisione dei <u>dati ordinati</u> in 10 classi aventi uguale frequenza produce i *decili*.

Il primo **decile** è l’unità di osservazione che ha la proprietà di avere sotto di sé un decimo dei dati della distribuzione.

Una suddivisione in 100 classi di frequenza uguale dei <u>dati ordinati</u> produce i *centili*.

Il primo **centile** è l’unità di osservazione che ha la proprietà di avere sotto di sé un centesimo dei dati della distribuzione.

Per calcolare i quantili si usa lo stesso procedimento usato per il calcolo della mediana.

Per calcolare il 24 percentile, per esempio, si ordinano i dati in senso crescente, e si determina se esiste un valore che abbia sotto di sé il 24% dei dati della distribuzione.

Se tale valore non esiste, una procedura di interpolazione lineare viene usata per trovare il valore esatto del percentile cercato.

Ranghi percentili

Si definisce rango *R* di una data unità di osservazione il numero che indica la posizione che quell'osservazione occupa nell'insieme ordinato a cui appartiene.

Es. *R* = 30 significa che l'osservazione in questione occupa il trentesimo posto nella serie ordinata delle osservazioni che costituiscono il campione.

I **ranghi percentili** si calcolano come:

$$R_p(x) = \frac{k}{n}100$$

Dove *k* è il numero di osservazioni con valore minore o uguale a *X*.

Punti *z* o punti standard

$$z_i = \frac{X_i - \bar{X}}{s}$$

I punti *z* ci dicono quante deviazioni stardard separano l'osservazione considerata dalla media.

La distribuzione dei punteggi standardizzati ha sempre <u>media uguale a zero</u> e <u>deviazione standard uguale a 1</u>.

La trasformazione in punti *z* non cambia la forma della distribuzione.

Dimostrazione di $\bar{z} = 0$

$$\bar{z} = \sum \frac{z_i}{n} = \sum \frac{X_i - \bar{X}}{ns}$$

$$\bar{z} = \frac{1}{ns} \sum (X_i - \bar{X}) = 0$$

Statistiche descrittive

	N	Media	Varianza	Asimmetria		Curtosi	
	Statistica	Statistica	Statistica	Statistica	Errore std	Statistica	Errore std
Punteggio	48	10,00	26,766	,038	,343	-,961	,674
Punteggio z	48	-6,9E-16	1,000	,038	,343	-,961	,674
Validi (listwise)	48						

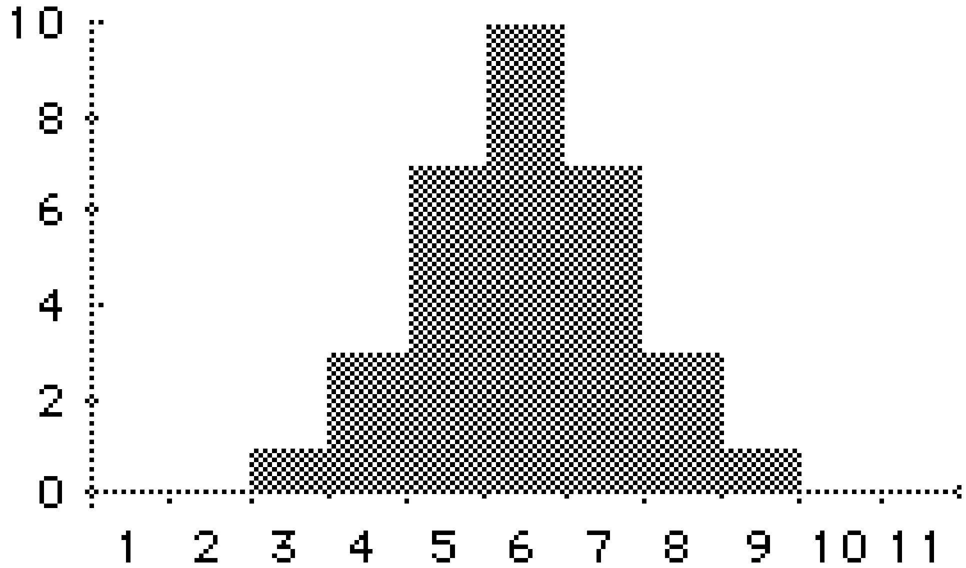
MISURE DI DISPERSIONE

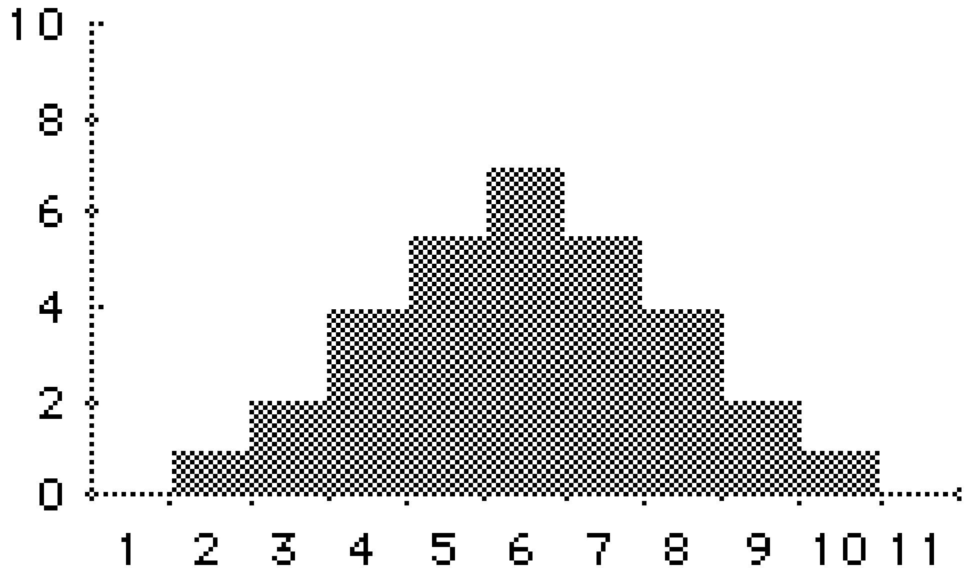
Le misure di dispersione esprimono la tendenza delle singole osservazioni di una distribuzione di allontanarsi dalla tendenza centrale, ovvero la “variabilità” dei dati.

La dispersione esprime la “bontà” o la “povertà” della tendenza centrale quale descrittore di una distribuzione.

[7, 8, 10, 12, 13]

[1, 2, 10, 18, 19]


10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 11


10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 11

Come si può quantificare la variabilità di una distribuzione?

GAMMA (CAMPO DI VARIABILITA')

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

[5, 2, 7, 11, 3, 6, 2]

Gamma = 11 - 2 = 9

DIFFERENZA INTERQUARTILICA

$$DQ = Q_3 - Q_1$$

SEMIDIFFERENZA INTERQUARTILICA

$$SQ = \frac{Q_3 - Q_1}{2}$$

VARIANZA E DEVIAZIONE STANDARD

Lo scarto tra ciascuna osservazione di una distribuzione e la media è dato da:

$$d_i = X_i - \overline{X}$$

Si potrebbe pensare che l'indice più semplice per descrivere la variabilità dei dati sia la media degli ***scarti dalla media***.

Problema:

$$\sum d_i = \sum \left(X_i - \overline{X} \right) = 0$$

Soluzione: elevare gli scarti al quadrato

Varianza: media degli scarti dalla media elevati al quadrato.

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}{n}$$

Esempio di calcolo della varianza

[2, 3, 6, 9, 15]

$$\overline{X} = \frac{2+3+6+9+15}{5} = 7$$

$$S^2 = \frac{(2-7)^2 + (3-7)^2 + (6-7)^2 + (9-7)^2 + (15-7)^2}{5}$$

$$S^2 = \frac{(-5)^2 + (-4)^2 + (-1)^2 + (2)^2 + (8)^2}{5}$$

$$S^2 = \frac{25+16+1+4+64}{5} = 22$$

Deviazione standard: radice quadrata della varianza.

$$S = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}{n}}$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}{n}}$$

$$S = \sqrt{\frac{25 + 16 + 1 + 4 + 64}{5}} = \sqrt{22} = 4.69$$

Formula alternativa per la varianza

$$S^2 = \sum \frac{\left(X_i - \overline{X}\right)^2}{n} = \sum \frac{X_i^2 - 2X_i\overline{X} + \overline{X}^2}{n}$$

$$S^2 = \sum \frac{X_i^2}{n} - 2\overline{X}\sum \frac{X_i}{n} + \sum \frac{\overline{X}^2}{n}$$

$$S^2 = \sum \frac{X_i^2}{n} - 2\overline{X}\sum \frac{X_i}{n} + n\frac{\overline{X}^2}{n}$$

$$S^2 = \sum \frac{X_i^2}{n} - 2\overline{X}\overline{X} + \overline{X}^2 = \sum \frac{X_i^2}{n} - \overline{X}^2$$

$$S^2 = \sum \frac{\left(X_i - \overline{X}\right)^2}{n} = \sum \frac{X_i^2}{n} - \overline{X}^2$$

ESERCIZIO

Per i seguenti dati, calcolate la media, la varianza e la deviazione standard.

Per il calcolo della varianza usate entrambe le formule presentate in precedenza.

2 5 8 7 3

In precedenza abbiamo definito un punto *z* come

$$z_i = \frac{X_i - \overline{X}}{s}$$

e abbiamo dimostrato che la <u>media</u> di una distribuzione standardizzata è uguale a 0.

$$\overline{z} = \sum \frac{z_i}{n} = \sum \frac{X_i - \overline{X}}{ns} = \frac{1}{ns} \sum (X_i - \overline{X}) = \mathbf{0}$$

Dimostriamo adesso che la varianza di una distribuzione standardizzata è uguale a 1.

$$S_z^2 = \sum \frac{\left(z_i - \bar{z}\right)^2}{n} = \sum \frac{z_i^2}{n}$$

$$S_z^2 = \sum \frac{\left(X_i - \bar{X}\right)^2}{nS^2} = \frac{S^2}{S^2} = 1$$

$$S_z = 1$$

Esempio numerico di calcolo della media e della varianza di una distribuzione standardizzata

[2, 3, 6, 9, 15]

$S = sqrt(22) = 4.6904$

$z_1 = (2-7)/S = -1.0660$

$z_2 = (3-7)/S = -0.8528$

$z_3 = (6-7)/S = -0.2132$

$z_4 = (9-7)/S = 0.4264$

$z_5 = (15-7)/S = 1.7056$

Media =
$(z_1+ z_2+ z_3+ z_4+ z_5+)/5 = 0$

Varianza =
$(z_1\text{^}2+ z_2\text{^}2+ z_3\text{^}2+ z_4\text{^}2+ z_5\text{^}2)/5 = 1$

Perché la varianza è calcolata nei termini degli scarti delle singole osservazioni dalla ***media***, anziché da altre misure di tendenza centrale?

La varianza assume il minore valore possibile quando viene calcolata a partire dalla media.

Il metodo che ha la proprietà di minimizzare la somma delle deviazioni al quadrato è detto ***principio dei minimi quadrati***.

$$S_C{}^2 = \sum \frac{\left(X_i - C\right)^2}{n}$$

$$S_C{}^2 = \sum \frac{\left(\left(X_i - \overline{X}\right) + \left(\overline{X} - C\right)\right)^2}{n}$$

$$S_C{}^2 = \sum \frac{\left(X_i - \overline{X}\right)^2 + 2\left(\overline{X} - C\right)\left(X_i - \overline{X}\right) + \left(\overline{X} - C\right)^2}{n}$$

$$S_C{}^2 = \sum \frac{\left(X_i - \overline{X}\right)^2}{n} + 2\left(\overline{X} - C\right)\sum \frac{\left(X_i - \overline{X}\right)}{n} + \sum \frac{\left(\overline{X} - C\right)^2}{n}$$

$$S_C{}^2 = \sum \frac{\left(X_i - \overline{X}\right)^2}{n} + 2\left(\overline{X} - C\right)\sum \frac{\left(X_i - \overline{X}\right)}{n} + \sum \frac{\left(\overline{X} - C\right)^2}{n}$$

$$S_C{}^2 = \sum \frac{\left(X_i - \overline{X}\right)^2}{n} + 0 + \left(\overline{X} - C\right)^2$$

$$S_C{}^2 = S^2 + \left(\overline{X} - C\right)^2$$

Dato che $\left(\overline{X} - C\right)^2$ è elevato al quadrato, deve essere maggiore o uguale a zero. Di conseguenza, $S_C{}^2$ sarà maggiore o uguale a S^2.

COEFFICIENTE DI VARIAZIONE

Il *coefficiente di variazione* è definito come il rapporto tra la deviazione standard e la media:

$$V = \frac{S}{\overline{\overline{X}}}$$

L'indice di *varianza relativa* è uguale al quadrato del coefficiente di variazione:

$$V_{rel}^2 = \frac{S^2}{\overline{\overline{X}}^2}$$

MISURE DI FORMA

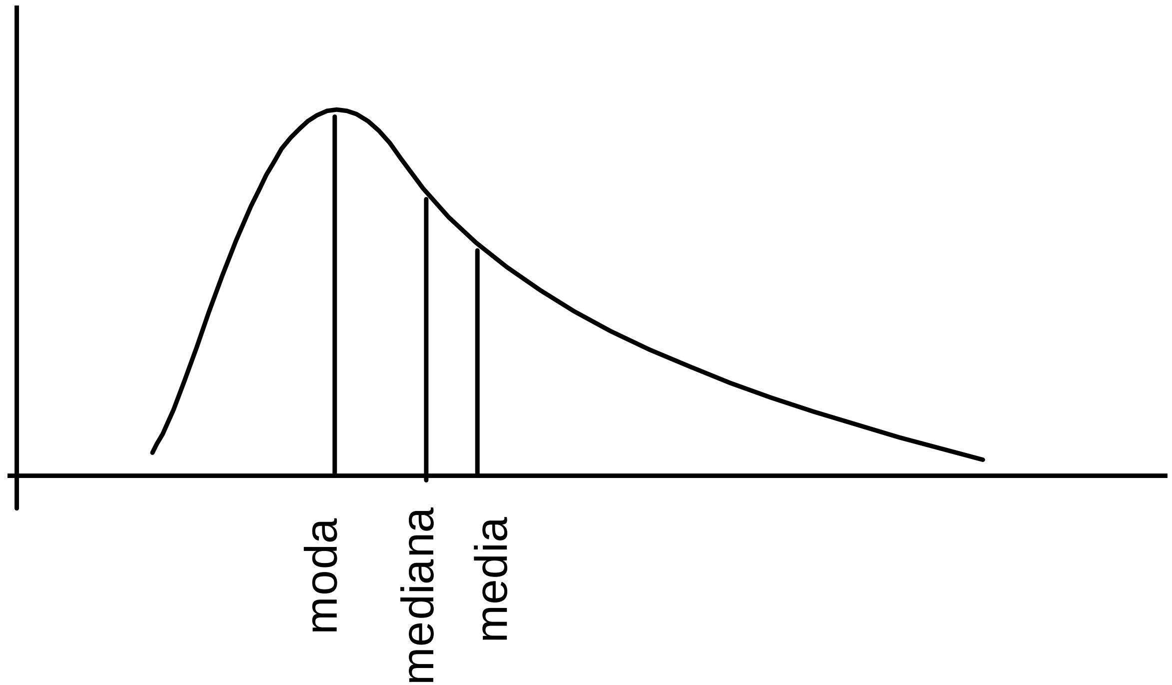
Le distribuzioni di frequenza unimodali possono essere simmetriche, asimmetriche positive o asimmetriche negative, con diversi gradi di curtosi (distribuzioni mesocurtiche, platicurtiche, leptocurtiche).

Come si può quantificare il grado di asimmetria e di curtosi di una distribuzione?

ASIMMETRIA

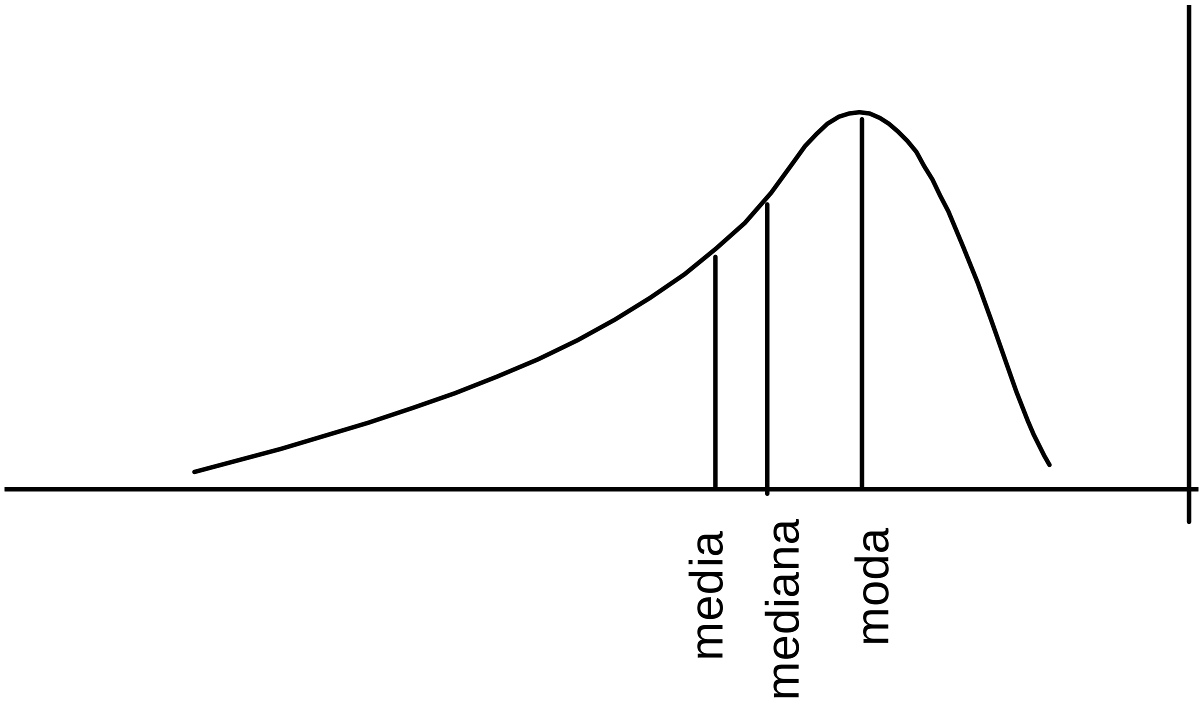
Asimmetria positiva: moda ≤ mediana ≤ media

Media - mediana > 0



Asimmetria negativa: media ≤ mediana ≤ moda

Media - mediana < 0



MISURE ASSOLUTE DI ASIMMETRIA

(1) La differenza tra la media e la mediana, e tra la media e la moda.

Queste differenze saranno negative nel caso di asimmetria negativa, nulle nel caso di simmetria, e positive nel caso di asimmetria positiva.

(2) La differenza tra la mediana e i due quartili:

$$\left(Q_3 - Mdn\right) - \left(Mdn - Q_1\right)$$

Questa differenza sarà negativa nel caso di asimmetria negativa, nulla nel caso di simmetria, e positiva nel caso di asimmetria positiva.

(3) Il momento centrale di terzo ordine.

Si definiscono come **momenti centrali di ordine *r***, i valori m_r:

$$m_r = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^r}{n}$$

Il momento di secondo ordine è la varianza.

Il momento di terzo ordine è negativo nel caso di asimmetria negativa, nullo nel caso di simmetria, e positivo nel caso di asimmetria positiva.

MISURE RELATIVE DI ASIMMETRIA

Le differenze (1) e (2) rapportate alla deviazione standard.

L'indice g_1 di Fisher:

$$g_1 = \frac{m_3}{\sqrt{m_2^3}} = \frac{m_3}{S^3}$$

L'indice g_1 è negativo nel caso di asimmetria negativa, nullo nel caso di simmetria, e positivo nel caso di asimmetria positiva.

Il grado di asimmetria è **trascurabile** se

$$0 < |g_1| < \frac{1}{2}$$

Il grado di asimmetria è **moderato** se

$$\frac{1}{2} < |g_1| < 1$$

Il grado di asimmetria è **elevato** se

$$|g_1| > 1$$

Una distribuzione simmetrica ha:

(1) media uguale a zero ?
(2) varianza uguale a zero ?
(3) asimmetria uguale a zero ?

In una distribuzione simmetrica, quali sono uguali?

(1) la media e la mediana
(2) la media e la moda
(3) la mediana e la moda
(4) tutte le precedenti

CURTOSI

La forma di una distribuzione, oltre che dai momenti, dipende dal grado di addensamento dei valori attorno alla media.

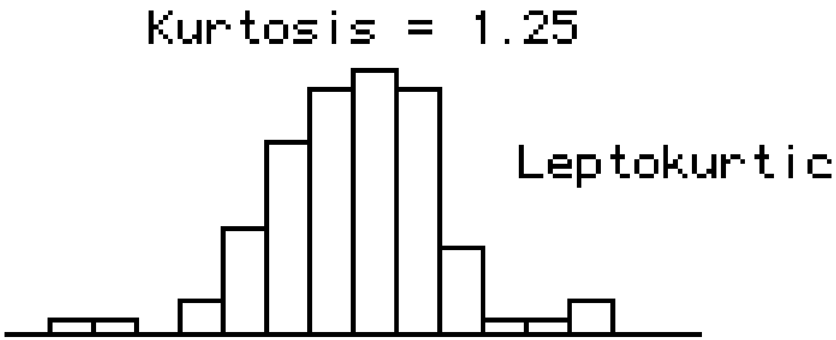
La misura di questo addensamento si chiama **curtosi**.

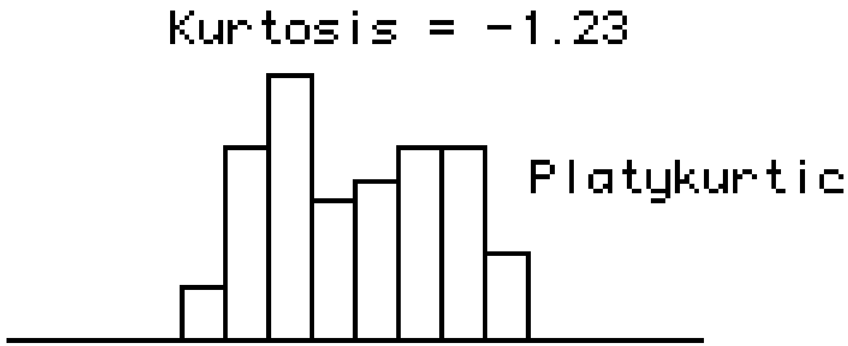
Il grado di curtosi si misura solitamente mediante l'indice g_2 di Fisher:

$$g_2 = \frac{m_4}{m_2^2} - 3$$

L'indice g_2 è uguale a zero nel caso della curva normale, o gaussiana, è maggiore di zero per curve più accentrate della normale (*leptocurtiche*), è minore di zero per curve meno accentrate della normale (*platicurtiche*).

Le seguenti distribuzioni hanno uguale varianza, stesso grado di asimmetria, ma diverso grado di curtosi.





MISURE DI TENDENZA CENTRALE

RELAZIONE TRA MISURE DI TENDENZA CENTRALE E CURVE DI FREQUENZA

INDICI DI POSIZIONE

MISURE DI DISPERSIONE

MISURE DI FORMA