Riassumere i dati

Indici di posizione e di dispersione. Indici di forma

Riassumere I dati

Obiettivo: definire quantità utili per **sintetizzare** un insieme di dati.

Queste misure riassuntive si chiamano **statistiche.**

Definizione

Le quantità numeriche calcolate a partire da un insieme di dati si chiamano *statistiche*

Indici riassuntivi di una distribuzione

- Indici di posizione: media, mediana, moda; percentili.
- Indici di variabilità o dispersione: intervallo di variazione, varianza (o deviazione standard)
- Indici di forma: curtosi, indice di simmetria

Indici di posizione centrale.

Media Campionaria: è la media aritmetica dei valori dei dati

Consideriamo un campione di n dati i cui valori sono $x_1, x_2, \ldots, x_n$

media campionaria $\bar{x}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Media campionaria

Media Campionaria: è la media aritmetica dei valori dei dati

Esempio
Lunghezza (cm) di 6 piantine di basilico
9.3 7.8 6.2 7.0 8.3 9.9

media campionaria

$\bar{x}$=(9.3+7.8+ 6.2+8.3+9.9)/6 cm = 8.1 cm

Proprietà 1

Se ciascun valore è in crementato di una costante c, anche la media viene incrementata di c

Consideriamo i dati $x_1, x_2, \ldots, x_n$

Aggiungiamo a ciascun valore la quantità c.

Otteniamo l nuovi valori

$y_1, y_2, \ldots, y_n,$ dove $y_i = x_i + c$, per i=1,…,n

Allora $\bar{y} = \bar{x} + c$

Esempio

5 nuclei familiari composti da padre, madre e un numero n di figli.

n figli per nucleo familiare

1, 2 ,2, 0, 1

Numero medio di figli $\bar{x}$ = (1+2+2+0+1)/5=1,2

Numero medio di componenti del nucleo familiare

$\bar{y}$ = $\bar{x}$ + 2 = 1,2 + 2 = 2,2

Esempio

Altezze (m) di un campione di 10 studenti

1,64, 1,73, 1,69, 1,82, 1,84, 1,76, 1,70 1,81, 1,77, 1,67

Sottraiamo a ciascun valore 1,70 m e otteniamo i dati

-0,06, 0,03, -0,01 0,12, 0,14 0,06, 0, 0,11, 0,07, -0,03

$\bar{y}$ = 0,04 m $\bar{x}$ = 1,70 + $\bar{y}$ =1,74 m

Proprietà 2

Se ciascun valore è moltiplicato per c, anche la media risulta moltiplicata per c

Consideriamo i dati $x_1, x_2, \ldots, x_n$
Moltiplichiamo ciascun valore per la quantità c.
Otteniamo i nuovi valori

$y_1, y_2, \ldots, y_n,$ dove $y_i = c\, x_i,$ per i=1,…,n

Allora $\bar{y} = c\bar{x}$

Proprietà 2

Esempio. Altezze del campione di 10 studenti espresse in cm (valori precedenti moltiplicati per 100)

164, 173, 169, 182, 184, 176, 170 181, 177, 167

$\bar{y}$ = 100 $\bar{x}$ = 174 cm

Esempio

Numero di cuccioli in 10 cucciolate

3, 4, 4, 3, 5, 5, 4, 3, 6, 5

Media $\bar{x}$= (3+4+4+3+5+5+4+3+6+5)/10 = 4,2

=(3×3 + 3×4 + 3×5 + 6)/10

$$=\frac{F_3\ 3+\ F_4\ 4+\ F_5\ 5+\ F_6\ 6}{10}$$

F_i = frequenza assoluta del valore i

Media campionaria

Insieme di n dati con k valori distinti $x_1, x_2, \ldots, x_k$ (k ≤ n)

con frequenze $F_1, F_2, \ldots, F_k$ (il valore x_i compare F_i volte)

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} F_i \, x_i = \sum_{i=1}^{k} f_i \, x_i \qquad (f_i \text{ freq. rel. di } x_i)$$

(ricordiamo che n= $\sum_{i=1}^{k} F_i$)

Esempio. Numero di orchidee in un CCS di 50 quadrati

n° orchidee	frequenza	frequenza relativa	
0	21	0.42	
1	15	0.30	
2	6	0.12	
3	3	0.06	
4	2	0.04	
5	1	0.02	
8	1	0.02	
12	1	0.02	
totali	50	1.00	

- $\bar{x} = 0 \times 0{,}42 + 1 \times 0{,}30 + 2 \times 0{,}12 + 3 \times 0{,}06 + 4 \times 0{,}04 + 5 \times 0{,}02 + 8 \times 0{,}02 + 12 \times 0{,}02 = 1{,}38$

Scarti

Le differenze tra ciascun valore dei dati e la media campionaria di chiamano scarti

$$s_i = x_i - \bar{x}$$

Esempio. Numero di formiche del legno catturate in 7 trappole

Valori: 25 4 12 9 15 8 20 media=13,3

Scarti: 11,7 -9,3 -1,3 -4,3 1,7 -5,3 6,7

osservazione: la somma degli scarti è uguale a zero

Media e valori estremi

Numero di formiche del legno catturate in 7 trappole

25 4 12 9 15 8 202

$\bar{x}$ = 39,3

La media campionaria è più grande di 6 delle 7 osservazioni ed è molto più piccolo dell'ultima.

La media utilizza il valore effettivo di ogni osservazione, perciò potrà essere distorta da un singolo valore eccezionale (non è robusta).

La mediana

La mediana è il valore intermedio di un insieme di osservazioni che sono state ordinate in ordine crescente

Esempio. Numero di settimane trascorse, per un campione di 7 persone, dalla fine del corso di guida al conseguimento della patente

2, 110, 5, 7, 6, 7, 3

Ordino in modo crescente i dati

2, 3, 5, 6, 7, 7, 110 med= 6

Osservazione: metà delle osservazioni sono più piccole e metà più grandi della mediana

Esempio

Ordiniamo i dati sulle formiche:

4 8 9 12 15 25 202 med=12

La mediana è più robusta della media, ossia non è influenzata da singoli valori estremi.

Nell'esempio la mediana vale 12 qualunque sia il valore della settima osservazione (202, 2002 o 20002).

La mediana

Esempio 1

1 4 7 9 10 12 14 med = 9

Esempio 2

11 13 15 16 19 21 22 25 med = (16 +19) / 2

Se le osservazioni sono in numero dispari la mediana è l'elemento che occupa il posto centrale.

Se le osservazioni sono in numero pari la mediana è la semisomma dei due elementi di posto centrale

Ricetta per calcolare la mediana

- Ordino i valori in modo crescente
- Se il numero di valori (N) è dispari, la mediana è il valore che occupa la posizione (N+1)/2
- Se il numero di dati valori (N) è pari, la mediana è la semisomma tra il valore che occupa la posizione N/2 e il valore successivo (posizione (N/2) + 1)

Media vs mediana

Se i dati sono approssimativamente simmetrici rispetto ai valori centrali, media e mediana sono vicine.

Esempio. Consideriamo i dati

4, 6, 8, 8, 9, 12, 15, 17, 19, 20, 22

med= 12 $\bar{x}$ =12,73

Modifichiamo i dati: 4, 6, 8, 8, 9, 12, 15, 17, 19, 20, 50

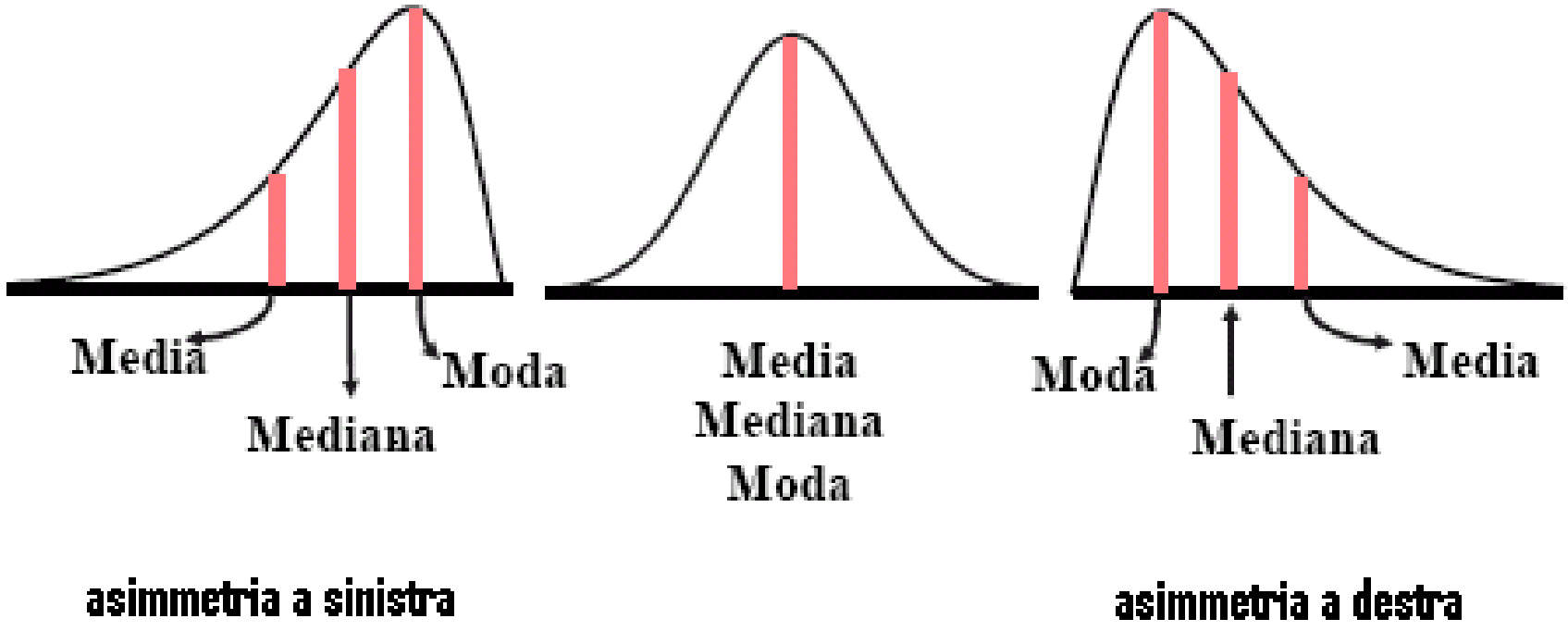
med=12 $\bar{x}$ = 15,27

Moda

La moda campionaria (quando esiste) è il valore che ha frequenza massima.

Se i dati sono divisi in classi, la classe modale (quando esiste) è la classe che ha frequenza massima.

Relazioni tra media mediana e moda



Le misure di posizione centrale

- Moda: è sempre calcolabile, ma è poco potente dal punto di vista informativo.
- Mediana: è calcolabile soltanto per caratteri almeno ordinabili e trascura l'informazione relativa alla grandezza quantitativa dei dati. Ha però il vantaggio di non essere influenzata dai dati estremi.
- Media: è calcolabile soltanto per caratteri quantitativi, è la più informativa, ma è influenzata dai dati estremi.

Moda, mediana, media hanno le stesse unità di misura delle osservazioni individuali.