

SAMPLE STATISTICS

When dealing with a dataset that is *unreadable* in its raw form because it is too vast, it becomes necessary to introduce numerical quantities capable of capturing its most relevant characteristics. These summary measures of the data are called **sample statistics or indices**.

Notice that a statistic is a numerical quantity whose value is only determined by the data.

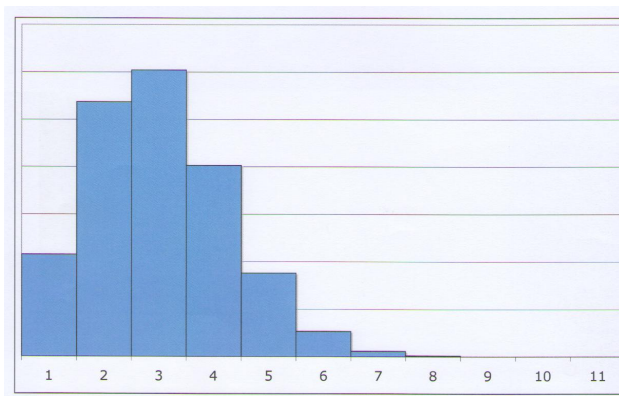
For example, the **mode** (the value with the highest frequency) is a statistic, as are the **mean**, the **median**, the **variance**, and the **standard deviation**.

Gli indici (statistiche) servono per misurare quantitativamente caratteristiche che possono essere osservate qualitativamente sui grafici.

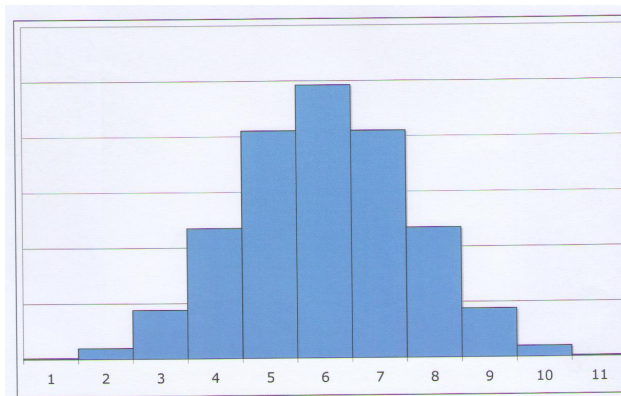
- **indici di posizione** che misurano la tendenza centrale dell'insieme dei dati: (**media, moda, mediana**)
- **indici di dispersione**: che danno una misura di quanto i valori siano lontani dal centro: (**varianza, deviazione standard**)

MODA

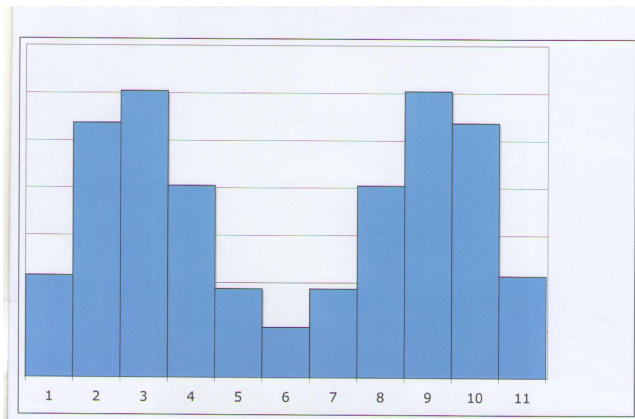
Si chiama **moda** campionaria il valore che si verifica con maggiore frequenza. Osservando il grafico qui sotto si vede subito che la moda è 3, inoltre c'è un solo valore corrispondente al massimo per cui questo grafico si chiama **unimodale**.



Quello qui sotto invece è ancora un grafico unimodale ma a differenza del precedente è **simmetrico**. La simmetria è centrata intorno alla moda che vale 6.



Il grafico qui sotto è invece **bimodale**, ossia ci sono due valori (3 e 9) che hanno la stessa massima frequenza. La moda assume quindi due valori (ci sono di due distinte classi modalì).



MEDIA

Si definisce **media campionaria** di un campione di taglia n , la media aritmetica dei dati

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

Esempio. La rilevazione della temperatura massima (in gradi centigradi) fatta a Roma il 1 agosto per 10 anni ha fornito i seguenti dati:

30.1 28.2, 31.3, 22.6, 27.5, 29.4, 24.5, 27.8, 33.2, 22.8

La media campionaria vale

$$\begin{aligned} \bar{x} &= \frac{1}{10} \left(30.1 + 28.2 + 31.3 + 22.6 + 27.5 + 29.4 \right. \\ &\quad \left. + 24.5 + 27.8 + 33.2 + 22.8 \right) = \frac{277.4}{10} = 27.74 \end{aligned}$$

PROPRIETÀ DELLA MEDIA

Se ciascun valore viene incrementato di una costante c (positiva o negativa) allora anche la media campionaria viene incrementata di c .

In formule: si definisce la variabile y con valori $y_1, \dots, y_n$,

$$y_i = x_i + c, \quad i = 1, \dots, n$$

allora, indicando con $\bar{y}$ la media campionaria della variabile y ,

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

La dimostrazione di questa asserzione è una semplice verifica:

$$\begin{aligned}\bar{y} &= \frac{1}{n} \left[(x_1 + c) + (x_2 + c) + \dots + (x_n + c) \right] = \\ &= \frac{1}{n} (x_1 + x_2 + \dots + x_n + nc) = \bar{x} + c\end{aligned}$$

Esempio: consideriamo la variabile x di taglia $n = 5$ con valori 6, 7, 5, 5, 7. La media campionaria è

$$\bar{x} = \frac{6 + 7 + 5 + 5 + 7}{5} = \frac{30}{5} = 6$$

Sommiamo $c = -5$ ad ogni valore di x e otteniamo la variabile y con valori

$$1 \quad 2 \quad 0 \quad 0 \quad 2$$

La media campionaria della variabile y è

$$\bar{y} = \frac{1 + 2 + 0 + 0 + 2}{5} = 1$$

per cui

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

Se ciascun valore della variabile x viene moltiplicato per una costante a (positiva o negativa), allora anche la media campionaria viene moltiplicata per a .

In formule: si definisce la variabile y con valori $y_1, \dots, y_n$,

$$y_i = a x_i, \quad i = 1, \dots, n$$

allora, indicando con $\bar{y}$ la media campionaria della variabile y ,

$$\bar{y} = a\bar{x}$$

Anche la dimostrazione di questa asserzione è una semplice verifica:

$$\begin{aligned}\bar{y} &= \frac{1}{n} (a x_1 + a x_2 + \dots + a x_n) \\ &= \frac{a}{n} (x_1 + x_2 + \dots + x_n) = a\bar{x}\end{aligned}$$

Questa proprietà è utile "per cambiare unità di misura".

Nel 1994 le entrate medie mensili in Italia per la voce "turismo" sono state 3.192.310 lire. Per conoscerle in euro basta usare $a = 1/1936,27$ (conversione Lire $\rightarrow$ Euro). Quindi le entrate medie mensili in euro per la voce turismo sono:

$$\bar{y} = \frac{3.192.310}{1936,27} = 1.648,69$$

Riassumendo. Se a partire dai valori $x_1, \dots, x_n$ della variabile x si definisce la variabile y con valori

$$y_i = ax_i + c, \quad a \neq 0, c \text{ numeri reali}$$

si ha

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

CALCOLO DELLA MEDIA CON LE FREQUENZE

Esempio. Abbiamo i seguenti dati disposti in una tabella delle frequenze assolute

Valore	Freq. assoluta
4	1
6	4
7	2
TOTALE	7

Quindi l'insieme dei dati originali è composto da 7 valori che disposti in modo crescente sono:

4 6 6 6 6 7 7

La media campionaria è dunque

$$\bar{x} = \frac{4 + 6 + 6 + 6 + 6 + 7 + 7}{7} = \frac{1 \cdot 4 + 4 \cdot 6 + 2 \cdot 7}{7} = 6$$

In generale sia x una variabile numerica di taglia n con k valori distinti ordinati $x_1 < x_2 < \dots < x_k$ con frequenze assolute $n_1, n_2, n_3, \dots, n_k$ ($n_1 + n_2 + \dots + n_k = n$). La media campionaria è data da

$$\bar{x} = \frac{n_1 \cdot x_1 + n_2 \cdot x_2 + \dots + n_k \cdot x_k}{n}$$

Si osservi che

$$\bar{x} = \frac{n_1}{n} x_1 + \frac{n_2}{n} x_2 + \dots + \frac{n_k}{n} x_k$$

e poichè $f_i = \frac{n_i}{n}$ è la frequenza relativa, la media può essere calcolata tramite le frequenze relative usando la formula

$$\bar{x} = f_1 x_1 + f_2 x_2 + \dots + f_k x_k = \sum_{j=1}^k f_j x_j$$

Esercizio 1. Supponiamo di sapere che la metà dei valori di un campione sono uguali a 10, un sesto sono uguali a 20 e un terzo sono uguali a 30. Quanto vale la media campionaria?

$$\bar{x} = \frac{1}{2} 10 + \frac{1}{6} 20 + \frac{1}{3} 30 = 18,33$$

Esercizio 2. Il numero di settimane trascorse per un campione di 7 persone da quando hanno completato il corso di guida a quando hanno ottenuto la patente sono

$$2 \quad 110 \quad 5 \quad 7 \quad 6 \quad 7 \quad 3$$

Calcolare la media campionaria:

$$\bar{x} = \frac{2 + 110 + 5 + 7 + 6 + 7 + 3}{7} = \frac{140}{7} = 20$$

Tutti i valori tranne 1 sono molto minori della media. Un punto debole della media come indicatore del centro di un insieme di dati. è che il suo valore è ampiamente influenzato da un valore estremo.

MEDIANA

Si introduce un'altro indicatore che indichiamo con $m = \text{mediana campionaria}$.

Si dispongono i valori degli n dati in ordine crescente:

- Se n è dispari allora m è il **valore intermedio**.
- Se n è pari allora m è la **media dei due valori intermedi**.

In altri termini, i dati $x_1, \dots, x_n$ sono disposti in ordine crescente

- se n è **dispari** $m = x_{\frac{n+1}{2}}$
- se n è **pari** $m = \frac{1}{2}[x_{\frac{n}{2}} + x_{\frac{n}{2}+1}]$

La definizione assicura che a destra della mediana cadono lo stesso numero di dati che a sinistra.

Esempio 1. Calcoliamo la mediana per i dati dell'esercizio 2

2 110 5 7 6 7 3

dove $n = 7$. In ordine crescente

2 3 5 6 7 7 110

La mediana campionaria è il **quarto** valore, quindi $m = 6$.

Esempio 2. Il campione è di taglia $n = 6$ ed i dati in ordine crescente sono

4 5 7 8 9 50

La mediana campionaria è la media aritmetica dei valori intermedi:

$$m = \frac{7 + 8}{2} = \frac{15}{2} = 7.5$$

La media campionaria di questi dati è

$$\bar{x} = \frac{4 + 5 + 7 + 8 + 9 + 50}{6} = \frac{83}{6} = 13,8$$

Osserva che la mediana campionaria non risente dei valori estremi. Infatti se il dato 50 viene sostituito con 10:

4 5 7 8 9 10

la mediana è la stessa mentre **la media diventa 7,1**

$$\bar{x} = \frac{4 + 5 + 7 + 8 + 9 + 10}{6} = \frac{43}{6} = 7,1$$

E' informativo calcolare entrambe le statistiche per descrivere la tendenza centrale di un insieme di dati. La media campionaria prende in considerazione tutti i valori, invece la mediana considera soltanto 1 o 2 valori centrali dei dati e quindi non dipende dai valori estremi.

ESERCISES ON POSITION INDICES

Esercizio 1. I dati di una variabile numerica di taglia $n = 11$ sono:

Valore	Freq. assoluta
10	3
11	1
16	3
21	1
23	3
TOTALE	11

Determinare la media campionaria e la mediana campionaria.

Poichè $n = 11$ è dispari la mediana è

$$m = x_{\frac{11+1}{2}} = x_6 = 16$$

$$\bar{x} = \frac{1}{11}[3 \times 10 + 11 + 3 \times 16 + 21 + 3 \times 23] = 16,27$$

Per dati approssimativamente simmetrici rispetto alla mediana la media e la mediana sono vicine.

Esercizio 2.

Una compagnia di assicurazioni ha rilevato il numero di incidenti nel periodo 1996-2000 relativo a 25 assicurati

0, 1, 0, 2, 5, 0, 1, 4, 3, 2, 0, 1, 0, 5, 2, 0, 0, 6, 1, 1, 0, 3, 1, 2, 2

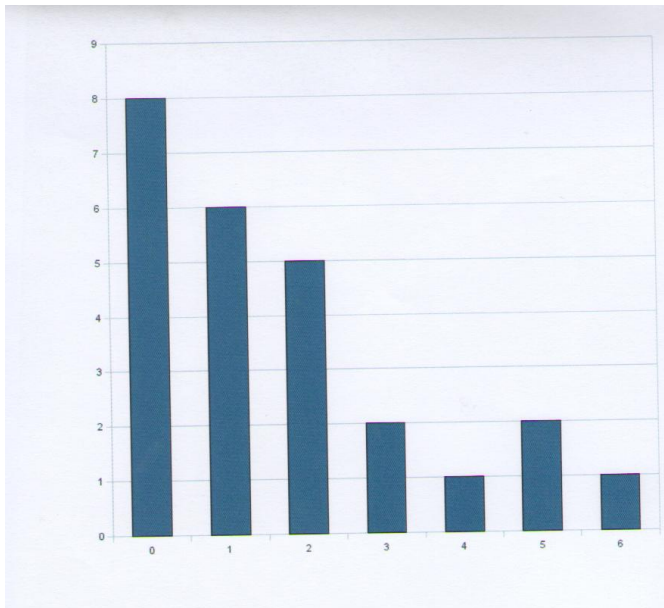
- (1)** Rappresentare i dati in una tabella delle frequenze e con un diagramma a barre.
- (2)** Calcolare la media, la mediana e la moda campionarie.
- (3)** Con quale frequenza non si è dovuto risarcire più di un sinistro?

Ordiniamo i dati

0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 4, 5, 5, 6

Valori	Freq. assoluta	Freq. relativa	Freq. percentuale
0	8	0.32	32
1	6	0.24	24
2	5	0.2	20
3	2	0.08	8
4	1	0.04	4
5	2	0.08	8
6	1	0.04	4
TOTALE	25	1	100

Diagramma a barre



La media campionaria è :

$$\bar{x} = \frac{6 + 2 \cdot 5 + 3 \cdot 2 + 4 + 5 \cdot 2 + 6}{25} = \frac{42}{25} = 1,68$$

La taglia del campione è $n = 25$ quindi la mediana è il valore che corrisponde al tredicesimo valore nella lista ordinata, ossia:

$$m = x_{13} = 1$$

La moda è il valore con la frequenza più alta ed è quindi uguale a 0.

Per calcolare con quale frequenza relativa non si è dovuto risarcire più di un sinistro, osserviamo che il numero degli assicurati che ha subito al più un incidente (cioè 0 oppure 1) è dato dalla somma delle frequenze assolute del valore 0 e del valore 1 $= 8+6=14$. La frequenza relativa cercata è quindi $\frac{14}{25} = 0.56$.

Esercizio 3.

Si è analizzato un campione di 770 incidenti motociclistici verificatisi a Roma nel 1980 e nel 1981. Gli incidenti sono stati classificati secondo la gravità del danno alla testa sofferto dal motociclista.

0= nessun danno, 1= minore, 2= moderato,
3= grave senza pericolo di vita, 4= grave con peric. di vita,
5= critico, sopravvivenza incerta, 6= fatale.

In **331** incidenti il motociclista **indossava il casco**
mentre negli altri **439 non indossava il casco.**

I dati sono stati riassunti in una tabella in cui:

- la prima colonna indica il valore,
- la seconda la frequenza assoluta con casco
- la terza la frequenza assoluta senza casco.

Valore	Freq. ass. con casco	Freq. ass. senza casco
0	248	227
1	58	135
2	11	33
3	3	14
4	2	3
5	8	21
6	1	6
TOTALE	331	439

Determinare la media e la mediana campionaria del danno alla testa sia per i motociclisti con casco che per quelli senza.

Indichiamo con

x la variabile numerica corrispondente ai motociclisti **con casco**

y la variabile numerica corrispondente ai motociclisti **senza casco**.

Valore	Freq. ass. con casco: x	Freq. ass. senza casco: y
0	248	227
1	58	135
2	11	33
3	3	14
4	2	3
5	8	21
6	1	6
TOTALE	331	439

La media campionaria con casco è

$$\bar{x} = \frac{0 \cdot 248 + 1 \cdot 58 + 2 \cdot 11 + 3 \cdot 3 + 4 \cdot 2 + 5 \cdot 8 + 6 \cdot 1}{331} = 0.43$$

La media campionaria senza casco è

$$\bar{y} = \frac{0 \cdot 227 + 1 \cdot 135 + 2 \cdot 33 + 3 \cdot 14 + 4 \cdot 3 + 5 \cdot 21 + 6 \cdot 6}{439} = 0.90$$

I dati indicano che i motociclisti con casco hanno sofferto, **in media**, danni minori di quelli senza casco.

Valore	Freq. ass. con casco	Freq. ass. senza casco
0	248	227
1	58	135
2	11	33
3	3	14
4	2	3
5	8	21
6	1	6
TOTALE	331	439

Le mediane campionarie con e senza casco = 0 entrambe. La mediana campionaria si calcola mettendo i dati in ordine crescente se **n è dispari** (come nel nostro caso) allora **$m = x_{\frac{n+1}{2}}$** **Con il**

casco: $\frac{n+1}{2} = \frac{332}{2} = 166$ **$x_{166} = 0$**

Senza casco: $\frac{n+1}{2} = \frac{440}{2} = 220$ **$x_{220} = 0$.**