DIPARTIMENTO DI STORIA, PATRIMONIO CULTURALE, FORMAZIONE E SOCIETÀ

Corso di Laurea Triennale

in

*Scienze del Turismo* (SciTUR)

# Indici di posizione

## Unità 2

a cura di
Andreina Anna D'Arpino

# INDICE

# INDICI DI POSIZIONE

## Valori sintetici

### Introduzione

*Gli indici statistici descrittivi hanno lo scopo di mettere in luce particolari aspetti di una distribuzione statistica e sono ritenuti utili per la soluzione di determinati problemi. Vengono utilizzati come sintesi dell'informazione fornita dalla distribuzione, di cui sono considerati valori rappresentativi (sempre sulla base degli obiettivi di studio).*

Un valore rappresentativo di un'intera distribuzione, per esempio un valore attorno a cui i dati si "addensano", viene denominato ***indice di posizione***. La conoscenza di un indice di posizione non può sostituire, in ogni circostanza, quella dell'intera distribuzione. Poiché distribuzioni, anche molto diverse, possono dare luogo ad uno stesso indice di posizione, è opportuno disporre almeno di un ulteriore indicatore il quale misuri la complessiva "distanza", dall'indice di posizione prescelto, dei valori della distribuzione; esso viene denominato ***indice di dispersione.***

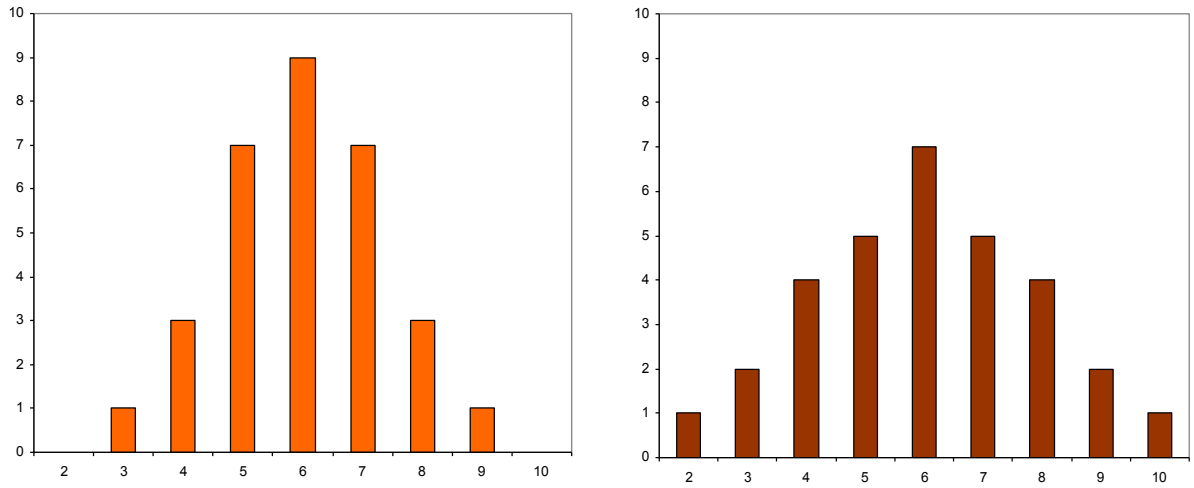



Figura 1

Osserviamo i due diagrammi a barre riportati in alto, entrambe le distribuzioni rappresentate hanno media uguale a 6, ma dispersione diversa: la prima a sinistra risulta meno dispersa rispetto alla seconda.

## Indici di posizione

Gli indici di posizione sono anche detti **medie** e si distinguono in **medie analitiche** e **medie di posizione**.

***Definizione***

*Le medie analitiche si possono applicare soltanto a caratteri quantitativi e sono calcolate mediante operazioni algebriche a partire dalle misure osservate.*
*Le medie di posizione richiedono operazioni quali l'ordinamento ed il confronto dei dati ed esse possono essere applicate sia a caratteri qualitativi ordinati sia a caratteri quantitativi.*
*La moda è l'unico indice che può essere utilizzato anche per caratteri qualitativi sconnessi.*

Sono medie analitiche: la **media aritmetica**, la **media armonica** e la **media geometrica.**
Sono medie di posizione: la **mediana**, i **quartili** e la **moda**.

## Le medie analitiche

### Media aritmetica

La media aritmetica, o semplicemente media, fornisce una misura dell'intensità complessiva del fenomeno ripartita in maniera esatta fra tutte le osservazioni.

**Definizione**

*La media aritmetica (semplice) di n misure:* $x_1, x_2, x_3, \ldots\ldots, x_n$ *è il numero reale M che si ottiene dividendo la loro somma per il numero n dei dati stessi:*

$$M = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

*che si può anche scrivere:*

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

*dove il simbolo* $\sum$ *(detto sommatoria) indica la somma dei termini* $x_i$ *, attribuendo ad i, successivamente, tutti i valori compresi tra 1 ed n.*

## Esempio

**La media aritmetica dei seguenti valori: 7, 13, 21, 40, 100 è data dalla somma di detti termini, divisa per il loro numero, cioè:**

$$M = \frac{7+13+21+40+100}{5} = \frac{181}{5} = 36{,}2$$

Più in generale, se in una distribuzione il valore $x_i$ compare con la frequenza $n_i$ ($i = 1,2,\ldots, k$) dove k rappresenta il numero delle modalità del carattere, in modo che risulti $n_1 + n_2 + n_3 + \ldots + n_k = n$, si può applicare la seguente formula

$$M = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{n_1 + n_2 + \ldots + n_k}$$

o, in forma più compatta:

$$M = \frac{\sum_{i=1}^{k} x_i n_i}{n}$$

**Definizione**

*La media aritmetica calcolata utilizzando le frequenze si chiama media aritmetica ponderata dei k valori $x_1, x_2, x_3, \ldots\ldots, x_k$ di pesi rispettivi $n_1, n_2, n_3, \ldots\ldots, n_k$.*

## Osservazione

*Se i dati sono organizzati in classi, per calcolare la media è necessario cercare il valore centrale [1] di ciascuna classe, operando prima la semisomma dei due estremi e poi procedendo come nel caso della distribuzione di frequenza.*

## Esempio

[1] Il ricorso al valore centrale della classe equivale ad ipotizzare che la frequenza del carattere sia concentrata su tale valore. Ad esempio, se 1.500 individui appartengono alla classe di età 0-14, ai fini del calcolo della media significa che si attribuisce ai 1.500 individui un'età pari a 7 (semisomma delle età 0 e 14).

**Riprendiamo la nostra indagine sui televisori venduti e calcoliamo la media aritmetica della variabile "dimensione del televisore" a partire dalla distribuzione per classi.**

*Procedura per il calcolo della media aritmetica per dati raggruppati in classi.*
*Variabile "dimensione dei televisori venduti nel 2014".*

| Modalità | Frequenza assoluta $n_i$ | Valore centrale $x_i$ | $x_i \cdot n_i$ |
|---|---|---|---|
| 14-26 | 5 | 20 | 100 |
| 27-32 | 2 | 29,5 | 59 |
| 33-37 | 4 | 35 | 140 |
| 38-42 | 6 | 40 | 240 |
| 43-46 | 6 | 44,5 | 267 |
| 47-51 | 3 | 49 | 147 |
| 52-57 | 4 | 54,5 | 218 |
| Totale | 30 | | 1171 |

Ne deriva che: $$M = \frac{\sum_{i=1}^{7} x_i n_i}{\sum_{i=1}^{7} n_i} = \frac{1171}{30} = 39{,}03$$

La dimensione media dei televisori venduti è pari a 39 pollici. Ovviamente se si calcola la media aritmetica a partire dalla distribuzione semplice (cfr. tab. 6-Unità 1) si ottiene, in generale, un valore diverso da quello che si è ottenuto con la distribuzione per classi. Nel nostro caso la dimensione media risulta pari a 39,63 pollici. Si conferma quanto avevamo sottolineato sulla perdita di informazione nel passaggio dal dato semplice alla distribuzione per classi.

## Scarto dalla media

### Definizione

*Dati i valori* $x_1, x_2, x_3, \ldots\ldots, x_n$ *e la loro media aritmetica M, si definiscono scarti dalla media le differenze:*

$$x_1 - M \qquad x_2 - M \qquad x_3 - M \qquad \ldots \qquad x_n - M$$

*Proprietà*

*La media aritmetica, comunque calcolata, gode delle seguenti proprietà:*

- *La media aritmetica è sempre compresa tra gli estremi della distribuzione.*

$$x_1 \leq M \leq x_n$$

- *La media aritmetica è tale che la somma degli scarti da essa è nulla, ossia*

$$\sum_{i=1}^{n}(x_i - M) = 0 \,,$$

*infatti, si ha:*

$$\sum_{i=1}^{n}(x_i - M) = \sum_{i=1}^{n} x_i - nM \qquad \text{dividendo per } \mathbf{n}: \Rightarrow \quad \frac{\sum_{i=1}^{n} x_i}{n} - M = 0.$$

- *Dati due insiemi di misure:* $x_1, x_2, x_3, \ldots\ldots, x_n$ *e* $y_1, y_2, y_3, \ldots\ldots, y_n$ *la media aritmetica delle somme* $x_1 + y_1, x_2 + y_2, x_3 + y_3, \ldots\ldots, x_n + y_n$ *è uguale alla somma delle medie aritmetiche dei due insiemi di misure, come si ricava banalmente.*

$$M_x + M_y = M_{x+y}$$

- *Fra le varie somme che si ottengono addizionando i quadrati degli scarti fra i termini di una successione ed un valore qualsiasi* $\overline{M}$ *è minima quella in cui* $\overline{M}$ *coincide con la media aritmetica M, dei termini della distribuzione.*

$$\sum_{i=1}^{n}(x_i - M)^2 = \text{minimo}$$

## Osservazione

*La proprietà della media aritmetica (la somma degli scarti dalla media è nulla) rende necessario, per misurare poi la dispersione del fenomeno, ricorrere alla proprietà su indicata, la quale garantisce che il risultato ottenuto, utilizzando la media aritmetica, sarà il più piccolo possibile. E' proprio questa proprietà che è alla base della costruzione degli indici di dispersione più diffusi in statistica: varianza e deviazione standard.*

## Media armonica

### Definizione

*Data la distribuzione di n valori* $x_1, x_2, x_3, \ldots\ldots, x_n$ *non nulli, di un carattere quantitativo, si dice media armonica di tali valori, il reciproco della media aritmetica dei reciproci dei valori dati:*

$$M_a = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

che si può anche scrivere:

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

Se vi è una distribuzione di frequenza si ricava la formula per la media armonica ponderata:

$$M_a = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$$

## Osservazione

*La media armonica non gode di alcuna delle proprietà che caratterizzano la media aritmetica.*
*Essa viene utilizzata quando i termini di un fenomeno sono il reciproco di un altro di cui si conoscono già i dati.*
*Un esempio tipico è il potere d'acquisto della moneta che è uguale al reciproco del prezzo della merce; quindi per trovare il potere d'acquisto medio si calcola la media armonica dei prezzi.*

## Esempi

**Un bene ha, su vari mercati, i seguenti prezzi:**

| | | | | |
|---|---|---|---|---|
| 9,5 | 10 | 12 | 13,5 | 15 |

Calcoliamo il potere d'acquisto medio rispetto a 1 euro.
I relativi poteri d'acquisto riferiti a 1 euro sono:

| | | | | |
|---|---|---|---|---|
| 0,105 | 0,100 | 0,083 | 0,074 | 0,067 |

e la loro media aritmetica semplice è:

$$M = \frac{0{,}105 + 0{,}100 + 0{,}083 + 0{,}074 + 0{,}067}{5} = 0{,}086$$

il cui reciproco è 1,16 che ci fornisce il prezzo medio nei 5 mercati.

Applicando direttamente la media armonica:

$$M_a = \frac{5}{\frac{1}{9{,}5} + \frac{1}{10} + \frac{1}{12} + \frac{1}{13{,}5} + \frac{1}{15}} = 1{,}16$$

Un altro caso in cui viene utilizzata **la media armonica** è la velocità media che si calcola come media armonica delle velocità registrate, in quanto il reciproco della velocità è uguale al tempo occorso per un'unità di spazio.

**Dobbiamo determinare la velocità media impiegata per percorrere la distanza di km 500 da tre auto, conoscendo i tempi impiegati da ciascun veicolo, e cioè:**

| | |
|---|---|
| Prima auto | 5h 20m 10' |
| Seconda auto | 5h 00m 5' |
| Terza auto | 4h 58m 30' |

Calcoliamo la media armonica dei tempi impiegati, ottenendo il tempo medio.

$$M_a = \frac{3}{\frac{1}{5.20.10} + \frac{1}{5.00.5} + \frac{1}{4.58.30}} = 5^h 5^m 56'$$

Dalla formula della velocità $v = \frac{s}{t} = \frac{500}{5.5.56} = 98$ km all'ora.

## Media geometrica.

### Definizione

*Dati n valori positivi* $x_1, x_2, x_3, \ldots\ldots, x_n$ *che rappresentano le misure di un carattere quantitativo, si dice media geometrica semplice la radice n-esima aritmetica del loro prodotto:*

$$M_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

In analogia con quanto considerato per la media aritmetica, nel caso in cui le misure siano fornite mediante distribuzione di frequenza in cui il valore $x_i$ compare con la frequenza $n_i$ $(i = 1,2,\ldots,k)$, avremo che:

se $x_1$ è presente $n_1$ volte dovendo eseguire un prodotto si dovrà moltiplicare $x_1$ $n_1$ volte, questo coincide con l'elevare alla $n_1$ il valore $x_1$, questa proprietà vale per tutti i termini.

$$M_g = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}}$$

Essendo $n = n_1 + n_2 + \ldots . n_k$

### Osservazione

*L'impiego della media geometrica dipende dalla natura del problema, essa è più adatta quando si richiede un indice che consenta una* ***equiripartizione*** *del prodotto dei termini. In generale la usiamo nel caso in cui i dati rappresentano un fenomeno che abbia una tendenza ad aumentare o diminuire in progressione geometrica.*

***Proprietà:***

*La media geometrica gode delle seguenti proprietà:*

*La media geometrica dei reciproci è uguale al reciproco della media geometrica:*

$$\sqrt[n]{\frac{1}{x_1 \cdot x_2 \cdot \ldots \cdot x_n}} = \frac{1}{\sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}} = \frac{1}{M_g}$$

## Esempio

**Si sono rilevati i prezzi al consumo delle mele verdi in quattro città. Calcolare il prezzo medio tramite la media geometrica.**

| Città A | Città B | Città C | Città D |
|---|---|---|---|
| 1,20 | 1,20 | 0,90 | 1,40 |

$$M_g = \sqrt[4]{1{,}20^2 \cdot 0{,}90 \cdot 1{,}40} = 1{,}16$$

# Le medie di posizione

Esistono altri indici di posizione, non analitici, che forniscono informazioni diverse sulla distribuzione statistica.

## Moda

Nel modo comune di dire "quest'anno va di moda" o "va di moda" significa indicare qual è la tendenza comune alla maggioranza degli individui.

**Definizione**

*In statistica la moda di una serie di n dati è il valore che presenta la maggior frequenza. Nel caso di dati raggruppati in classi è la classe k che presenta la maggiore frequenza. Può essere calcolata sia per dati qualitativi che quantitativi. Può accadere che nessuna classe o nessun valore sia più frequente di altri, come pure che due classi abbiano frequenza più elevata (comportamento bimodale).*

### Esempio

**Riprendiamo l'indagine sui televisori venduti nel 2014 e calcoliamo la moda della distribuzione "tipo".**
Come si può facilmente notare il valore di massima frequenza appartiene alla modalità **LCD** con 14 televisori venduti. La moda è il tipo **LCD** che ha avuto le vendite maggiori.

| Modalità | Frequenza assoluta $n_i$ |
|---|---|
| CRT (tradizionale) | 6 |
| **LCD** | **14** |
| Plasma | 10 |
| Totale | 30 |

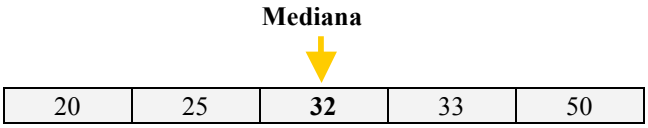
## Mediana

### Definizione

*La mediana, detta anche valore centrale o mediano, di una serie di n dati ordinati è rappresentata dal valore centrale (se n è dispari) o dalla media aritmetica dei due valori centrali (se n è pari).*

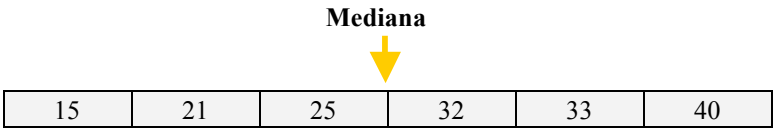
La mediana ha il compito di separare le osservazioni in due parti esattamente uguali (un 50% di valore inferiore e un 50% di valore superiore alla mediana stessa).
Se la distribuzione è semplice basta disporre i termini in ordine crescente o decrescente e individuare:

- Se **n** è **dispari** il valore centrale.




Nel caso della distribuzione in alto, essendo n = 5 il valore mediano è il valore che occupa il 3° posto Me = 32.

- Se **n** è **pari** non esiste un valore centrale.




Nella distribuzione riportata in alto, composta da **6** termini, non abbiamo un termine centrale, bensì due: 25 e 32, la **mediana** si ottiene calcolando la **media aritmetica** dei **due termini**:

$$M_e = \frac{25+32}{2} = 28{,}5$$

### Osservazione

*In generale*

- *se n è dispari la mediana è rappresentata dal termine che occupa il* $\left(\frac{n+1}{2}\right)^o$ *posto;*
- *se n è pari la mediana è rappresentata dalla media aritmetica dei termini che si trovano a* $\left(\frac{n}{2}\right)^o$ e $\left(\frac{n}{2}\right)^o + 1$

Nel caso di **distribuzione di frequenza** è opportuno ricorrere alle frequenze cumulate.

## Esempio

**Calcolare la mediana dei voti all'esame di statistica di 25 studenti riportati nella seguente tabella.**

| $x_i$ | $n_i$ | $N_i$ |
|---|---|---|
| **15** | **3** | **3** |
| **18** | **5** | **8** |
| **20** | **10** | **18** |
| **25** | **3** | **21** |
| **30** | **4** | **25** |
| | **25** | |

Essendo n = 25, la mediana si troverà a $\left(\frac{25+1}{2}\right)=13^{o}$ posto, dall'esame delle frequenze cumulate si può osservare che il tredicesimo studente si trova tra quelli che hanno preso 20 .
Quindi il **valore mediano** è pari a 20.

## Quartili, percentili

In un insieme di n dati ordinati la mediana è stata definita come il valore che separa l'insieme in due parti uguali.
Estendendo tale concetto, possiamo definire i valori che separano l'insieme in 4, 10 o 100 parti uguali: parleremo rispettivamente di ***quartili*** ($Q_1, Q_2$ e $Q_3$), **percentili** ($P_1, P_2, \ldots, P_{99}$).

In generale tali indici possono essere chiamati **quantili**.

### Osservazione

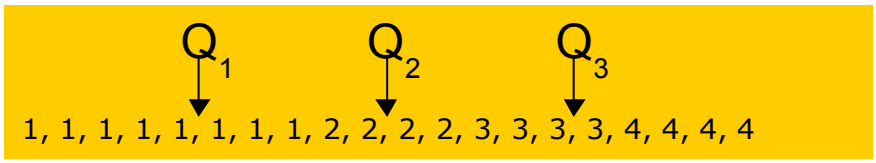
*La mediana è un unico valore, i quartili sono 3, i percentili 99. La mediana coincide con il 2° quartile (e quindi con il 50° percentile).*

### Esempio

**Data la seguente distribuzione di 5 termini:**

1, 1, 2, 3, 4

Per poter determinare i **quartili** occorre portare la distribuzione ad un numero di termini **divisibile** per **4**, moltiplico ciascun termine per 4 ne ottengo una simile di 20 termini:




- Il 1° quartile, essendo n pari, sarà la semisomma tra $\left(\frac{n}{4}\right)^o$ e $\left(\frac{n}{4}\right)^o + 1$, cioè tra il 5° e 6° termine.
- Il 2° quartile, essendo n pari, sarà la semisomma tra: $\left(\frac{n}{2}\right)^o$ e $\left(\frac{n}{2}\right)^o + 1$, cioè tra il 10° e 11° termine.
- Il 3° quartile, essendo n pari, sarà la semisomma tra: $\left(\frac{3}{4}n\right)^o$ e $\left(\frac{3}{4}n\right)^o + 1$, cioè tra il 15° e 16° termine.

Cioè:

$$Q_1 = 1,\ Q_2 = 2 \text{ e } Q_3 = 3$$

# Riepilogo formule

| | |
|---|---|
| Media aritmetica semplice | $M = \frac{\sum_{i=1}^{n} x_i}{n}$ |
| Media aritmetica ponderata | $M = \frac{\sum_{i=1}^{k} x_i n_i}{n}$ |
| Media armonica semplice | $M_a = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$ |
| Media armonica ponderata | $M_a = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$ |
| Media geometrica semplice | $M_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$ |
| Media geometrica ponderata | $M_g = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}}$ |
| Mediana | **Se n è dispari al posto** $\left(\frac{n+1}{2}\right)^o$ <br> **Se n è pari tra** $\left(\frac{n}{2}\right)^o$ **e** $\left(\frac{n}{2}\right)^o + 1$ |
| Quartili | **Posto n è pari e multiplo di 4:** <br> **1° Quartile** <br> $\left(\frac{n}{4}\right)^o$ **e** $\left(\frac{n}{4}\right)^o + 1$ <br> **2° Quartile** <br> $\left(\frac{n}{2}\right)^o$ **e** $\left(\frac{n}{2}\right)^o + 1$ <br> **3° Quartile** <br> $\left(\frac{3}{4}n\right)^o$ **e** $\left(\frac{3}{4}n\right)^o + 1$ |

# Indici di dispersione

*Come già espresso precedentemente, è opportuno completare la descrizione del collettivo, utilizzando indici che permettano di valutare la variabilità delle osservazioni.*

I principali indici di dispersione (o di variabilità) sono: il **campo di variazione**, la **varianza**, la **deviazione standard**, lo **scarto semplice medio** e il **coefficiente di variazione**.
Tali indici sono utilizzati per sintetizzare di quanto la distribuzione statistica sia addensata attorno ad una misura di localizzazione.

## Il campo di variazione

Il **campo di variazione** (**range**) è dato semplicemente dalla differenza tra il valore più grande e quello più piccolo del campione.

$$CV = x_n - x_1$$

### Esempio

**Due aziende (A e B) producono succhi di frutta che confezionano in bottiglie della capacità di 1 litro. Si prendono a caso in esame 5 bottiglie dei succhi A e 5 bottiglie dei succhi B e si rileva, con uno strumento molto preciso, il contenuto di ciascuna bottiglia:**

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | **Media** |
|---|---|---|---|---|---|---|
| *Campione A (l)* | 0.97 | 1.00 | 0.94 | 1.03 | 1.06 | 1.00 |
| *Campione B (l)* | 1.06 | 1.01 | 0.88 | 0.91 | 1.14 | 1.00 |

Come si vede, la media dei due campioni è del tutto identica e vale esattamente 1 litro. Il semplice calcolo del campo di variazione, che vale:

$$\mathbf{CV(A)} = 1.06 - 0.94 = 0.12$$
$$\mathbf{CV(B)} = 1.14 - 0.88 = 0.26$$

permette di dire che, in base ai campioni raccolti, il contenuto effettivo delle bottiglie del campione B presenta una maggiore variabilità di quello di A.

# Lo scarto interquartile

**Definizione**

*Lo scarto interquartile è dato dal valore assoluto della differenza tra il 3° quartile e 1° quartile:*

$$SQ = Q_3 - Q_1$$

*Esso delimita la zona centrale della distribuzione che contiene il 50% delle osservazioni.*
Anche noto come **campo di variazione interquartile** è un'altra misura di variabilità non influenzata dai valori estremi.

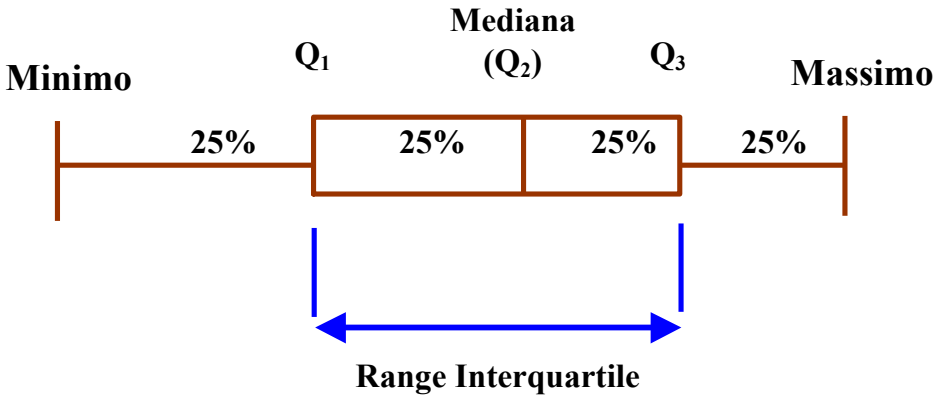



Figura 2

Il grafico sopra riportato, denominato **Box-plot,** rappresenta in modo compatto la distribuzione statistica attraverso alcuni indici sintetici:

- **Indici di posizione** (misurati con la mediana ed i quartili) e rappresentati nel grafico con una linea all'interno del box (mediana) ed i due estremi del box stesso (primo e terzo quartile).

- **Indice di variabilità** (misurato con la differenza interquartile) e rappresentato dalla base del rettangolo (box) .

Sono inoltre, riportati: il **valore minimo** della distribuzione (primo segmento verticale) e il **valore massimo** (ultimo segmento verticale).

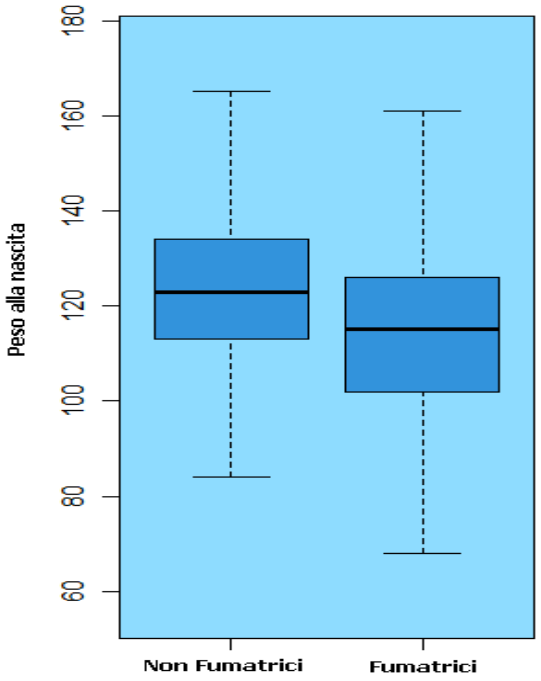
La dimensione dell'altezza (o base se messo in verticale) del rettangolo non rappresenta alcuna informazione, come pure la posizione del Box-Plot, che può essere posto sia verticalmente che orizzontalmente.

Il Box plot risulta molto utile quando si debbano confrontare più distribuzioni.

## Esempio

**Supponiamo di voler confrontare il peso alla nascita (in once) di neonati da madri fumatrici e non fumatrici.**

| | Non fumatrici | Fumatrici |
|---|---|---|
| **Min** | **55** | **58** |
| **$Q_1$** | **113** | **102** |
| **$Q_2$** | **123** | **115** |
| **$Q_3$** | **134** | **126** |
| **Media ar.** | **123,05** | **114,11** |
| **Max** | **176** | **163** |
| **Range** | **121** | **105** |
| **Range interquartile** | **21** | **24** |




Il box-plot è un ottimo strumento per visualizzare le differenze tra i due gruppi. Nella tabella sono riportati i pesi relativi ai neonati provenienti da madri fumatrici e non fumatrici; i neonati di madri fumatrici hanno mediamente un peso alla nascita inferiore rispetto a quello delle madri non fumatrici. Tutti e tre gli indici di posizione ($Q_1$, $Q_2$ e $Q_3$) sono minori nel gruppo dei neonati provenienti da madri fumatrici. La variabilità dei pesi, misurata con lo scarto interquartile risulta maggiore rispetto ai neonati di madri non fumatrici.

## Osservazione

*Anche se il campo di variazione è un indice di variabilità piuttosto elementare, in molti processi produttivi è l'unico indice utilizzato come elemento di controllo del processo stesso. Indipendentemente dalla variabilità che caratterizza le unità prodotte, il processo produttivo sarà ritenuto soddisfacente solo se le misure ricadranno all'interno del range prefissato.*

Se, però, si vuole tenere conto anche dei valori intermedi occorre utilizzare qualche altro strumento. La prima cosa che viene in mente è di "misurare" quanto i singoli valori differiscano dalla media della distribuzione. Supponiamo che la media in questione sia la media aritmetica (ma può essere un altro valor medio qualsiasi). Possiamo calcolare gli scarti dalla media cioè le differenze fra ciascun valore osservato e la media aritmetica.
Poiché la media è compresa fra il valore più piccolo e quello più grande, alcuni scarti sono positivi e altri negativi.

## Esempio

**Calcoliamo gli scarti dalla media aritmetica per i dati delle due aziende produttrici di succhi di frutta.**

Per calcolare gli scarti basta sottrarre ad ogni valore riportato in tabella il valore della media. Si ottiene così:

| Unità | Campione A | Campione B |
|---|---|---|
| 1 | -0,03 | 0,06 |
| 2 | 0 | 0,01 |
| 3 | -0,06 | -0,12 |
| 4 | 0,03 | -0,09 |
| 5 | 0,06 | 0,14 |
| Somma | 0 | 0 |

Osserviamo che il valore assoluto dei singoli scarti risulta maggiore per i gruppi in cui le misure mostrano maggiore variabilità e che, comunque, la somma degli scarti risulta sempre nulla [2].

[2] Per una nota proprietà della media aritmetica.

## La varianza

Quanto detto finora indica che la variabilità e gli scarti sono fra loro legati e che, quindi, si può pensare di misurare la variabilità di un fenomeno statistico considerando e sintetizzando la distribuzione degli scarti. Posto ciò, resta il fatto che tale sintesi non può essere fatta calcolando semplicemente la media degli scarti, dato che questa è nulla in quanto scarti positivi e scarti negativi si vanno a compensare. L'inconveniente può essere superato ricorrendo ad un espediente: anziché considerare la **media degli scarti** consideriamo la **media degli scarti al quadrato** [3]. Questo indice si chiama **varianza**.

**Definizione**

*La varianza è un indice usato per misurare la dispersione o variabilità, cioè l'addensamento maggiore (poca dispersione o variabilità) o minore (molta dispersione) dei dati attorno alla media aritmetica ed è definito come segue:*

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - M)^2}{n}$$

*Proprietà*

- *La varianza è sempre positiva.*

  Il numeratore, infatti, è somma di addendi tutti positivi.

- *La varianza è uguale a zero se (e solo se) la variabilità è nulla.*

  In questo caso, si ha: $x_1 = x_2 = .... = x_n = M$

  e, quindi tutti gli addendi che figurano al numeratore della precedente relazione sono uguali a zero; viceversa, se la varianza è uguale a zero, essendo il numeratore somma di termini tutti non negativi, è necessario che questi siano tutti nulli.

  $$x_1 = x_2 = .... = x_n = M$$

- *La varianza è tanto più elevata quanto più elevata è la variabilità.*

  Se la variabilità è più elevata, i termini al numeratore tenderanno, pertanto, ad essere più grandi.

[3] Elevando al quadrato ciascuno scarto diventa positivo.

## Esempio

**Calcoliamo la varianza dei due campioni di bottiglie di succo di frutta.:**

| Unità | Campione A $(x_i - M_A)^2$ | Campione B $(x_i - M_B)^2$ |
|---|---|---|
| 1 | 0,0009 | 0,0036 |
| 2 | 0 | 0,0001 |
| 3 | 0,0036 | 0,00144 |
| 4 | 0,0009 | 0,0081 |
| 5 | 0,0036 | 0,0196 |
| Somma | **0,009** | **0,0458** |

Dal calcolo otteniamo i seguenti valori delle varianze:

$$\sigma^2(A) = \frac{0{,}009}{5} = 0.0018 \qquad \sigma^2(B) = \frac{0{,}0458}{5} = 0.00916$$

Come si può facilmente osservare tali indici non sono più espressi in litri, avendo elevato ogni scarto al quadrato. Si conferma la maggior variabilità del campione B.

Nel caso di una **distribuzione di frequenza** per calcolare la varianza è necessario moltiplicare ciascuno scarto per la relativa frequenza. Otteniamo così, la seguente formula:

$$\sigma^2 = \frac{\sum_{i=1}^{k} (x_i - M)^2 n_i}{n}$$

## Esempio

**Calcoliamo la varianza per il carattere: voto conseguito all'esame di statistica dei 25 studenti considerati.**

| $x_i$ | $n_i$ | $x_i n_i$ | $x_i - M$ | $(x_i - M)^2$ | $(x_i - M)^2 \cdot n_i$ |
|---|---|---|---|---|---|
| 15 | 3 | 45 | -6,20 | 38,44 | 115,32 |
| 18 | 5 | 90 | -3,20 | 10,24 | 51,2 |
| 20 | 10 | 200 | -1,20 | 1,44 | 14,4 |
| 25 | 3 | 75 | 3,80 | 14,44 | 43,32 |
| 30 | 4 | 120 | 8,80 | 77,44 | 309,76 |
| | 25 | 530 | | | 534 |

Per calcolare la varianza facciamo riferimento alla formula nel caso di distribuzione di frequenza.

Il **valore medio** risulta: $M = \frac{530}{25} = 21{,}2$

Dalla formula : $\sigma^2 = \frac{\sum_{i=1}^{k} (x_i - M)^2 n_i}{n}$ avremo $\sigma^2 = \frac{534}{25} = 21{,}36$

## Lo scarto quadratico medio

Per misurare il grado di variabilità di una distribuzione, è preferibile, il più delle volte, calcolare la **radice quadrata** (positiva) della varianza.

**Definizione**

*Lo scarto quadratico medio, o scostamento quadratico medio o scarto standard si ottiene dal calcolo della radice quadrata della varianza:*

$$\sigma = \sqrt{\sigma^2} = \sqrt{\left( \frac{1}{n} \sum_{i=1}^{n} (x_i - M)^2 \right)} \; .$$

*Utilizzando lo scarto standard ci si riconduce ad un indice di variabilità espresso nella stessa unità di misura della variabile considerata. Come per la varianza, maggiore è la variabilità dei valori di un insieme di dati e maggiore è la deviazione standard, la quale assume valore nullo solo nel caso in cui tutti i valori siano uguali.*

**Osservazione**

*Osserviamo che si potrebbero, in teoria, definire altri indici di dispersione Il motivo della scelta privilegiata della deviazione standard è nella proprietà di minimo della media quadratica, relativa alle variabili scarto, rispetto alla media aritmetica, la quale rende particolarmente significativo ed utile $\sigma$ come indice di dispersione.*

**Esempio**

**Calcoliamo lo scarto quadratico medio o deviazione standard dei due campioni A e B.**

$$\sigma(A) = \sqrt{0{,}0018} = 0.04243 \qquad \sigma^2(B) = \sqrt{0{,}00916} = 0.0957$$

## Gli scostamenti semplici medi

Altre misure di variabilità sono gli scarti semplici medi che si ottengono come **media aritmetica delle differenze**, in **valore assoluto**, tra i valori osservati $x_1, x_2, x_3, \ldots\ldots, x_n$ di una variabile $x$ e un indice di posizione.
A seconda della media scelta si può ottenere uno specifico scarto semplice medio. Per esempio, se come media scegliamo la media aritmetica **M**, si ha **lo scarto semplice medio dalla media aritmetica:**

$$SM = \sum_{i=i}^{n} \frac{|x_i - M|}{n}.$$

**Osservazione**

*Come la deviazione standard, anche questo indice di dispersione è omogeneo e si annulla solo quando tutte le unità presentano la stessa modalità.*

Se invece di considerare le differenze dalla media aritmetica consideriamo quelle dalla **mediana Me** otteniamo lo scarto semplice medio dalla mediana:

$$SMe = \sum_{i=1}^{n} \frac{|x_i - M_e|}{n}.$$

Questo è ancora un indice omogeneo e, inoltre, gode di una proprietà di minimo analoga a quella di cui gode $\sigma$ rispetto alla media.

*Proprietà*

*La somma degli scarti in valore assoluto dalla mediana è un minimo*

$$SM = \sum_{i=1}^{n} |x_i - M| = \mathbf{minim}$$

## Il coefficiente di variazione

Tutti gli indici presentati, non consentono di effettuare confronti essendo legati all'unità di misura attraverso la quale è espresso il fenomeno. Chiaramente la variabilità delle misure non dipende dall'unità di misura utilizzata, così per rendere più facilmente confrontabili le misure della dispersione si costruisce il **coefficiente di variazione.**

**Definizione**

*Il coefficiente di variazione che si indica con CV, è il rapporto tra il valore dello scarto quadratico medio e il valore della media. L' indice ottenuto è un numero puro indipendente dall'unità di misura utilizzata.*

**Esempio**

**In un collettivo in cui sono state rilevate le stature, in cm, e i pesi, in kg, risulta:**

| | Media aritemetica | Deviazione **standard** |
|---|---|---|
| **peso** | M = 67,6 kg | $\sigma$ = 7,6 kg |
| **statura** | M =171,7 cm | $\sigma$ = 7,7 cm |

**Quale delle due distribuzioni è più dispersa? In altre parole, risulta più variabile il peso o la statura?**

Calcoliamo il coefficiente di variazione nei due gruppi:

$$CV = \frac{7,6}{67,6} = 0,112 \qquad CV = \frac{7,7}{171,7} = 0,045$$

Come si può notare, c'è una maggiore variabilità per la variabile **peso** rispetto alla variabile **altezza** (quasi il triplo).

# La forma di una distribuzione

Gli indici di posizione e di variabilità di una distribuzione di frequenza non esauriscono le informazioni contenute nei dati. Un altro aspetto da prendere in considerazione è la **forma**.

**Definizione**

*Quando si parla di forma ci si riferisce, in particolare a due aspetti:*

- *La simmetria o meno di una distribuzione rispetto al centro di gravità o media aritmetica;*
- *Il grado di appiattimento della distribuzione, sempre attorno al suo centro di gravità, rispetto ad una distribuzione particolare che viene chiamata distribuzione normale e che vedremo in dettaglio più avanti.*

Per avere un'idea della simmetria di una distribuzione basta guardare la curva riportata al centro, in una distribuzione di questo tipo media aritmetica, mediana e moda coincidono.

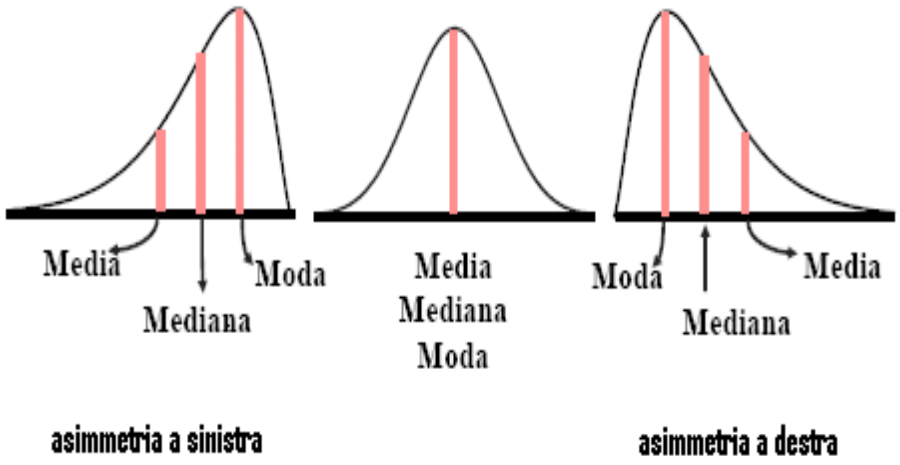



Figura 3

Si ha una forma simile al grafico a destra quando c'è una **asimmetria positiva** o una coda destra, esistono cioè un maggior addensamento dei dati in corrispondenza di valori inferiori alla media.

Nel caso rappresentato dal primo grafico, al contrario, abbiamo una coda a sinistra o un'**asimmetria negativa** dovuta alla presenza di un maggior addensamento dei dati per valori superiori alla media aritmetica.

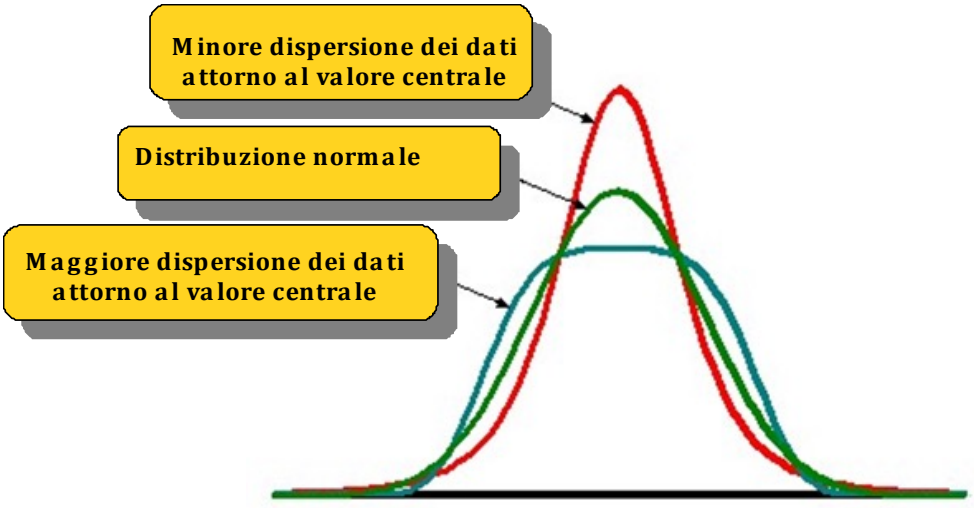



Figura 3

Il maggiore o minore addensamento dei dati attorno al valore centrale fa assumere alla curva una forma più o meno appuntita come si può osservare nella fig. 4.

## Riepilogo formule

| | |
|---|---|
| Campo di variazione | $CV = x_n - x_1$ |
| Varianza per distribuzione semplice | $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - M)^2}{n-1}$ |
| Varianza per distribuzione di frequenza | $\sigma^2 = \frac{\sum_{i=1}^{k}(x_i - M)^2 n_i}{n-1}$ |
| Scostamento semplice medio dalla media aritmetica | $SM = \sum_{i=i}^{n} \frac{\lvert x_i - M \rvert}{n}$ |
| Scostamento semplice medio dalla mediana | $SMe = \sum_{i=1}^{n} \frac{\lvert x_i - M_e \rvert}{n}$ |
| Deviazione standard o scostamento quadratico medio | $\sigma = \sqrt{\sigma^2}$ |
| Differenza interquartile | $SQ = Q_3 - Q_1$ |
| Coefficiente di variazione | $CV = \frac{\sigma}{M}$ |