# STATISTICA DESCRITTIVA
*PER IL CORSO DI LAUREA IN BIOLOGIA*

FRANCESCA PRINARI

Contents



## 1. **Scopi della statistica**

La Statistica é un settore della Matematica Applicata di supporto a varie discipline. Lo studio di un fenomeno (per esempio di tipo biologico) conduce spesso all'analisi di informazioni espresse sotto forma di dati.
La Statistica fornisce concetti e strumenti per

- **organizzare** e **riassumere** in modo significatico i dati raccolti per evidenziare gli aspetti rilevanti ivi contenuti e descrivere quindi le caratteristiche del campione o dell'intera popolazione. Questo é il compito della **Statistica descrittiva**;
- **fare delle previsioni** o dire qualcosa (ossia **inferire**) sull'intera popolazione. Questo é il compito della **Statistica inferenziale**.

**Esempio 1.1.** Le elezioni richiedono l'operazione di contare voti, calcolare percentuali, descrivere la composizione dell'organo eletto. Questi compiti sono di competenza della Statistica descrittiva.
Elaborare i risultati di un sondaggio per capire come si orienterà il voto dell'intera popolazione é di competenza della Statistica inferenziale.

1.1. **Popolazione.** L'indagine statistica viene compiuta su una **popolazione**: essa può essere **reale** (per esempio un insieme di oggetti o di persone) o può essere **virtuale**, per esempio l'insieme di tutte le possibili repliche di un esperimento. Chiameremo **unitá statistica** ogni singolo elemento che costituisce la popolazione. Un sottoinsieme della popolazione si dice **campione** ed é scelto mediante una selezione detta **campionamento**. Quando si esaminano tutte le unitá si parla invece di **censimento**. Attenzione ai campionamenti!!!

1.2. **Tipi di variabili.** Ogni indagine statistica ha un "oggetto di studio" che é una caratteristica della popolazione.

**Definizione 1.2.** *Una* **variabile** *é una caratteristica delle unitá statistiche.*

Esempi di variabili sono il colore degli occhi, l'altezza, il grado di istruzione, il voto di maturità, il gruppo sanguigno...
Una variabile può assumere una molteplicitá di valori detti **modalitá** ed in base alle modalitá una variabile si dice

- **qualitativa** se le modalitá sono espresse "in forma verbale": per esempio
  - **livello di istruzione**: diploma di scuola media, diploma di scuola media superiore, laurea triennale, laurea magistrale, master, dottorato...;
  - **gruppo sanguigno**: A,B,0,AB;
  - **colore degli occhi**: celeste, verde, grigio, castano, nero...;
- **quantitativa** se le modalitá sono espresse **in forma numerica** :altezza, peso, lunghezza...

Una variabile **qualitativa** si dice

- **ordinale** se esiste un ordinamento naturale delle modalitá (e.g.: livello di istruzione);
- **sconnessa** in caso contrario (e.g.: gruppo sanguigno, colore degli occhi).

Una variabile **quantitativa** si dice

- **discreta** se le possibili modalitá sono un insieme numerico finito o al piú numerabile (e.g.:voto di laurea, voto di maturitá...);
- **continua** se le possibili modalitá sono un insieme numerico continuo (e.g.: l'altezza, il peso).

Allo scopo di studiare una fissata variabile, si raccolgono i dati ad essa relativi tramite *sperimentazione* o tramite *osservazione*. Tali dati saranno di tipo *qualitativo* o *quantitativo* a seconda della variabile.

**Definizione 1.3.** *Chiameremo* **variabile statistica** *ogni singola rilevazione della variabile di interesse su un campione.*

Quindi la stessa variabile rilevata su campioni diversi dà luogo, in genere, a variabili statistiche differenti. Per esempio scegliamo come variabile l'altezza dei ragazzi di terza media; in una classe potrebbero esserci molti ripetenti oppure molti anticipatari; i dati raccolti differirebbero molto.

Indicheremo con $Y = (y_1, y_2, \cdots, y_N)$ la generica variabile statistica discreta (dove $N$ é la dimensione del campione).

**Esempio 1.4.** A 20 studenti iscritti al primo anno di università si chiede il voto riportato alla maturità. La variabile é il *Voto di maturità* e supponiamo che i valori rilevati sul campione di studenti siano

$$(76, 90, 70, 74, 80, 84, 78, 80, 70, 94, 76, 82, 74, 80, 70, 76, 76, 94, 100, 82).$$

Questa stringa é la nostra variabile statistica $Y$. $N = 20$ é la dimensione del campione. La variabile di interesse può assumere potenzialmente tutti i valori da 60 a 100 mentre la nostra variabile statistica $Y$ (ossia la variabile "voto" testata sul nostro campione) assume i valori distinti

$$\{z_1, z_2, \cdots, z_{10}\} = \{70, 74, 76, 78, 80, 82, 84, 90, 94, 100\}.$$

## 2. **Frequenze assolute e relative**

Quando il campione é molto numeroso (ossia $N$ é grande), é complicato trascrivere i dati raccolti. Allora, invece di rappresentare la variabile $Y$ con l'intera stringa $(y_1, y_2, \cdots, y_N)$,

- si prende una sola volta ogni singola modalitá;
- si conta "quante volte compare" ogni singola modalità.

**Definizione 2.1.** *Sia $Y$ una variabile statistica. Si dice* **frequenza assoluta** *di una modalità il numero di volte che tale modalitá viene osservata.*

Quindi se indichiamo le modalitá tutte tra loro distinte riscontrate sul campione

$$\{z_1, z_2, \cdots, z_k\}$$

con $k \leq N$, usando le frequenze assolute associate a tali modalitá, si costruiscono tabelle del tipo

| **Modalità** | $z_1$ | $\cdots$ | $z_j$ | $\cdots$ | $z_k$ |
|---|---|---|---|---|---|
| **Frequenza ass.** | $f_1$ | $\cdots$ | $f_j$ | $\cdots$ | $f_k$ |

**Osservazione 2.2.** In una indagine in cui esaminiamo il tipo di gruppo sanguigno di 200 persone, abbiamo che $N = 200$. Essendo pero' i gruppi sanguigni possibili solo 4, nella nostra indagine le modalitá distinte riscontrate sul campione possono essere al piu' 4. Quindi

- se sul campione di $N = 200$ persone troviamo distribuiti tutti e 4 i gruppi sanguigni allora
$$\{z_1, z_2, z_3, z_k\} = \{A, B, AB, 0\};$$
- in teoria, in un campione potremmo anche trovare presente un solo gruppo sanguigno, per esempio tutti hanno il gruppo 0. In tal caso $k = 1$ e abbiamo solo
$$\{z_1\} = \{0\}.$$
- Se invece sul nostro campione troviamo presenti solo i gruppi 0 e $A$ allora $k = 2$ e
$$\{z_1, z_2\} = \{A, 0\}.$$
Cosi' via.

**Esempio 2.3.** Considerando la variabile statistica dell' 1.4 abbiamo

| **Voto** | 70 | 74 | 76 | 78 | 80 | 82 | 84 | 90 | 94 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Freq. ass** | 3 | 2 | 4 | 1 | 3 | 2 | 1 | 1 | 2 | 1 |

Se non sapessimo a la dimensione $N$ del campione, da questa tabella ricaveremmo

$$N = 3 + 2 + 4 + 1 + 3 + 2 + 1 + 1 + 2 + 1 = 20$$

ossia **la somma di tutte le frequenze assolute é uguale alla dimensione $N$ del campione**.

## 3. **Frequenza relativa e percentuale**

**Definizione 3.1.** *Sia $Y$ una variabile statistica. Se $f$ é la frequenza assoluta di una modalitá ed $N$ é la dimensione del campione, si dice* **frequenza relativa** *il rapporto*

$$p = \frac{f}{N}.$$

*La* **frequenza percentuale** *é la frequenza relativa moltiplicata per* 100%

Ovviamente la somma di tutte le frequenze relative é uguale ad 1 mentre la somma delle frequenze percentuali é 100%.

**Esempio 3.2.** La variabile statistica dell' Esempio 1.4 aveva frequenze assolute

| Voto | 70 | 74 | 76 | 78 | 80 | 82 | 84 | 90 | 94 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| Freq.ass | 3 | 2 | 4 | 1 | 3 | 2 | 1 | 1 | 2 | 1 |

Allora dividendo le frequenze assolute per

$$N = 3 + 2 + 4 + 3 + 2 + 1 + 1 + 2 + 1 = 20$$

si ottiene la tabella delle frequenze relative:

| Voto | 70 | 74 | 76 | 78 | 80 | 82 | 84 | 90 | 94 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| Freq.rel | 0.15 | 0.1 | 0.2 | 0.05 | 0.15 | 0.1 | 0.05 | 0.05 | 0.1 | 0.05 |

**Esempio 3.3.** Data la tabella

| Gruppo sanguigno | $A$ | $B$ | $AB$ | 0 |
|---|---|---|---|---|
| Frequenza ass. | 2 | 3 | 1 | 4 |

si dica lunghezza del campione e tabella di distribuzione delle frequenze relative.

Il campione ha dimensione $N = 2 + 3 + 1 + 4 = 10$. Quindi la tabella di distribuzione delle frequenze relative é

| Gruppo sanguigno | $A$ | $B$ | $AB$ | 0 |
|---|---|---|---|---|
| Frequenza ass. | 0.2 | 0.3 | 0.1 | 0.4 |

## 4. **Classi**

**Come visto negli esempi sopra, una distribuzione di frequenza** si può rappresentare mediante una tabella del tipo

| Modalità | $z_1$ | $\cdots$ | $z_j$ | $\cdots$ | $z_k$ |
|---|---|---|---|---|---|
| Frequenza ass. | $f_1$ | $\cdots$ | $f_j$ | $\cdots$ | $f_k$ |

oppure

| Modalità | $z_1$ | $\cdots$ | $z_j$ | $\cdots$ | $z_k$ |
|---|---|---|---|---|---|
| Frequenza rel. | $p_1$ | $\cdots$ | $p_j$ | $\cdots$ | $p_k$ |

In alcuni casi (per esempio con variabili continue) é conveniente "**raggruppare**" le modalitá in "classi contigue e disgiunte" e contare le unità che appartengono a ciascuna classe.

**Esempio 4.1.** Da un'indagine condotta su 200 laureati in Biologia si riscontrano i seguenti voti conseguiti all'esame di Matematica:

| Voto | $[18, 20)$ | $[20, 22)$ | $[22, 24)$ | $[24, 26)$ | $[26, 28)$ | $[28, 30]$ |
|---|---|---|---|---|---|---|
| Freq. ass. | 42 | 50 | 60 | 32 | 12 | 4 |

Dividendo tutte le frequenze assolute per il numero $N = 200$ si ottiene la tabella:

| Voto | $[18, 20)$ | $[20, 22)$ | $[22, 24)$ | $[24, 26)$ | $[26, 28)$ | $[28, 30]$ |
|---|---|---|---|---|---|---|
| Freq. rel. | 0.21 | 0.25 | 0.3 | 0.16 | 0.06 | 0.02 |

4.1. **Distribuzioni di frequenze usando le classi.**

- Si sceglie il **numero** di classi. Si tenga conto che se tale numero é troppo grande, le frequenze relative alle classi possono diventare troppo piccole; se é troppo piccolo, si perde troppa informazione. La **regola di Sturges** suggerisce per $N$ osservazioni un numero di classi pari a $1 + \frac{\log N}{\log 2}$;
- si scelgono **gli estremi delle classi** e si scrivono intervalli del tipo $[z_{j-1}, z_j)$;
- **si contano il numero di volte** in cui la variabile statistica assume valori compresi tra $z_{j-1}$ (incluso) e $z_j$ (escluso) per trovare la tabella di frequenza assolute di questo tipo:

| Classi | $[z_1, z_2)$ | $\cdots$ | $[z_{j-1}, z_j)$ | $\cdots$ | $[z_k, z_{k+1}]$ |
|---|---|---|---|---|---|
| Frequenza ass. | $f_1$ | $\cdots$ | $f_{j-1}$ | $\cdots$ | $f_k$ |

Per trovare la tabella di frequenze relative, occorre dividere le frequenze assolute per la dimensione $N$ del campione che é data da

$$N = f_1 + f_2 + \cdots + f_k = \sum_{i=1}^{k} f_i.$$

Quindi

$$p_j = \frac{f_j}{N} = \frac{f_j}{\sum_{i=1}^{k} f_i}.$$

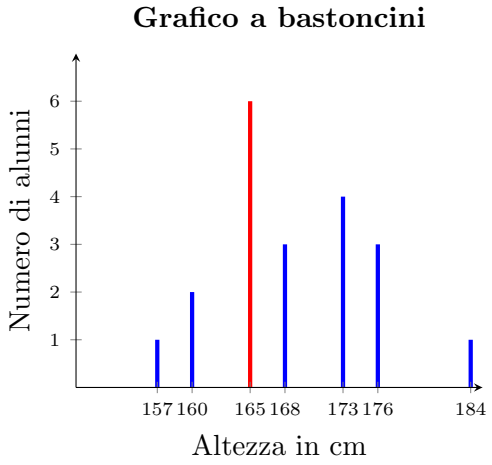


FIGURE 1. **Altezza degli alunni di una classe di terza media**

## 5. Rappresentazione dei dati

Un primo approccio all'analisi di un insieme di dati é costituito dai metodi grafici. Per rappresentare le distribuzione di frequenza (di variabili quantitative o qualitative), si possono utilizzare vari tipi di grafici: a bastoncini, a poligonali, a barre, areogrammi, istogrammi.

5.1. **Grafici a bastoncini.** : sull'asse delle $x$ si posizionano le varie modalità e in corrispondenza di esse si tracciano dei bastoncini di altezza proporzionale o pari alla frequenza assoluta o relativa delle modalitá; così si mettono in evidenza le modalità che hanno la frequenza (assoluta o relativa) più grande.

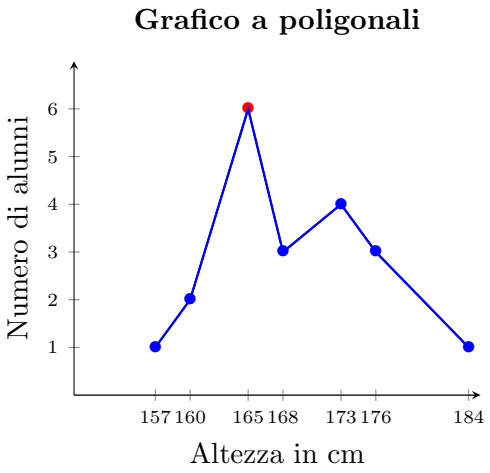


FIGURE 2. **Altezza degli alunni di una classe di prima media**

5.2. **Grafici a poligonali.** Quando si tracciano delle **poligonali** che uniscono gli estremi superiori dei bastoncini, si ottiene un grafico a poligonali:

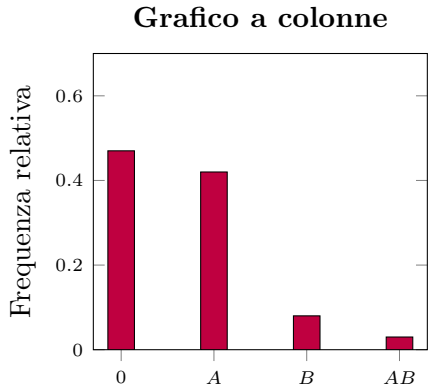


FIGURE 3. **Diffusione dei gruppi sanguigni in Italia**

5.3. **Grafici a barre o a colonne.** Quando alle linee viene dato spessore fino a diventare dei rettangoli, si ottengono i **grafici a barre**: sull'asse delle $x$ si posizionano le varie modalità e in corrispondenza ad esse si tracciano dei rettangoli di altezza proporzionale o pari alla frequenza assoluta o relativa delle modalitá:

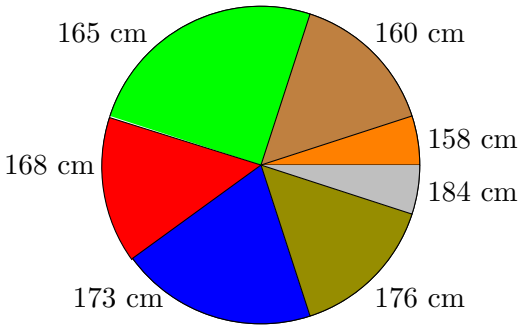


FIGURE 4. **Altezza alunni di III media**

5.4. **Areogramma.** Quando si tratta di visualizzare le diverse parti in cui un tutto é stato suddiviso, si ricorre spesso a un tipo di rappresentazione detto **areogramma** (o torta!).

| Altezza alunni | 158 | 160 | 165 | 168 | 173 | 176 | 184 |
|---|---|---|---|---|---|---|---|
| Freq. ass. | 1 | 2 | 6 | 3 | 4 | 3 | 1 |

Per convenzione le ampiezze dei settori circolari devono essere proporzionali alle grandezze delle corrispondenti parti.

La rappresentazione risulta efficace per mettere in evidenza i mutui rapporti tra le parti.

Supponiamo di avere delle frequenze percentuali $p_1, p_2, \cdots, p_n$ con $p_1+p_2+\cdots+p_n = 100\%$. Allora l'areogramma che rappresenta questi dati é un cerchio suddiviso in $n$ settori circolari di ampiezza rispettivamente $p_1/100 \cdot 360^o, p_2/100 \cdot 360^o, \cdots, p_n/100 \cdot 360^o$ in modo da riempire l'intero cerchio.

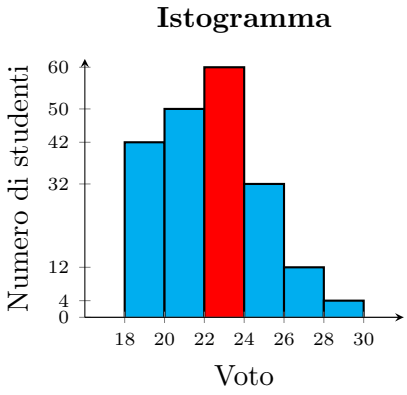


FIGURE 5. **Voto all'esame di Matematica**

5.5. **Istogrammi.** Quando i dati sono raggruppati per classi contigue, allora il grafico a barre delle frequenze si dice **istogramma**.

| Voto | $[18, 20)$ | $[20, 22)$ | $[22, 24)$ | $[24, 26)$ | $[26, 28)$ | $[28, 30]$ |
|---|---|---|---|---|---|---|
| Freq. ass. | 42 | 50 | 60 | 32 | 12 | 4 |

Quando i dati sono raggruppati per classi contigue, allora il grafico a barre delle frequenze si dice **istogramma**. Le altezze sono scelte in modo che le aree dei rettangoli costruiti siano uguali o proporzionali alle frequenze.

Si osservi che se si considera il rettangolo $R_j$ di base $[z_j, z_{j+1}]$ e altezza uguale a $f_j$, allora la sua area é data da

$$\text{base} \cdot \text{altezza} = (z_{j+1} - z_j) \cdot f_j.$$

Quindi se le classi $[z_j, z_{j+1})$ hanno la stessa ampiezza $(z_{j+1} - z_j) = b$, le aree dei rettangoli sono uguali a $b \cdot f_j$ ossia sono proporzionali alle frequenze assolute.

Se invece le classi $[z_j, z_{j+1})$ hanno ampiezze diverse, allo scopo di fare in modo che le aree rimangano proporzionali alle frequenze, si scelgono le altezze dei rettangoli uguali (o proporzionali) a $f_j/(z_j - z_{j-1})$. In tal modo le aree dei rettangoli sono uguali (o proporzionali) a

$$\frac{f_j}{(z_j - z_{j-1}}) \cdot (z_j - z_{j-1}) f_j = f_j.$$

Stesso discorso se al posto delle frequenze assolute si considerano le frequenze relative.

A partire dall'istogramma si può costruire il **poligono di frequenza**: basta tracciare la spezzata che congiunge i punti medi dei lati superiori dei rettangoli dell'istogramma.

## 6. Indici

Allo scopo di riassumere i dati, si introducono degli indici sintetici che

- descrivono intorno a quale valore si concentrano i dati (**indici di posizione o di centralità**); Indici di posizione sono la **moda, media**, la n **media** e la **mediana** che ora introdurremo;
- come si disperdono i dati (**indici di dispersione**): introdurremo la **varianza** e lo **deviazione standard**.

6.1. **La moda.** Questo tipo di rappresentazioni permette subito di individuare la modalitá piú "riscontrata" sul campione, la cosiddetta **moda**.

**Definizione 6.1.** *Sia $Y$ una variabile qualitativa o quantitativa. La* **moda** *é la modalità con frequenza (relativa o assoluta) più alta, ossia la modalità rilevata maggiormente sulla popolazione. Se le modalità sono raggruppate in classi della stessa ampiezza si chiama* **classe modale** *la classe con il numero maggiore di elementi.*

**Esempio 6.2.** Considerando la variabile statistica

- dell'Esempio 2.3 la moda é 76;
- dell'Esempio 4.1 la classe modale é la classe $[18, 20)$.

6.2. **Media.** Supponiamo di avere raccolto un numero finito di dati $y_1, y_2, \cdots, y_N$. Vogliamo trovare un numero $\bar{y}$ che li riassuma. Vediamo se esiste un $\bar{y}$ tale che renda nulla la somma degli errori ossia tale che

$$\sum_{i=1}^{N} (\bar{y} - y_i) = 0.$$

Otteniamo così

$$N\bar{y} - \sum_{i=1}^{N} y_i = 0$$

da cui

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}.$$

**Definizione 6.3.** *Sia $Y = (y_1, y_2, \cdots, y_N)$ una variabile statistica discreta. Si chiama* **media** *o* **media aritmetica** *o* **valor medio** *di $Y$ la quantità*

$$\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{y_1 + y_2 + \cdots y_N}{N}.$$

**Esempio 6.4.** Supponiamo di misurare la lunghezza dell'assone della cellula neurale di un ratto. Ripetiamo 10 volte l'operazione di misurare e otteniamo i seguenti risultati espressi in micrometri: 70.78, 74.22, 74.03, 71.71, 70.97, 73.47, 69.28, 69.62, 72.31, 72.76.

La media delle lunghezze dell'assone della cellula neurale é

$$\frac{70.78 + 74.22 + 74.03 + 71.71 + 70.97 + 73.47 + 69.28 + 69.62 + 72.31 + 72.76}{10} = 71.195.$$

6.2.1. **Calcolo della media usando le frequenze.**

- **Se é nota la distribuzione delle freq. assolute di** $Y$

| **Modalità** | $z_1$ | $\cdots$ | $z_j$ | $\cdots$ | $z_k$ |
|---|---|---|---|---|---|
| **Frequenza ass.** | $f_1$ | $\cdots$ | $f_j$ | $\cdots$ | $f_k$ |

allora la dimensione del campione é

$$N = \sum_{j=1}^{k} f_j$$

e quindi

$$\bar{Y} = \frac{y_1 + y_2 + \cdots y_N}{N} = \frac{\sum_{j=1}^{k} f_j z_j}{\sum_{i=1}^{k} f_j}.$$

**Esempio 6.5.** Da un'indagine condotta su un gruppo di studenti in Biologia si riscontra il seguente grado di soddisfazione (espresso con voto da 0 a 5) sul corso di Matematica:

| **Voto** | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **Frequenza ass.** | 2 | 8 | 40 | 24 | 10 | 16 |

Allora

$$N = 2 + 8 + 40 + 24 + 10 + 16 = 100$$

e il grado medio di soddisfazione é:

$$\frac{0 \cdot 2 + 1 \cdot 8 + 2 \cdot 40 + 3 \cdot 24 + 4 \cdot 10 + 5 \cdot 16}{100} = 2.8.$$

- **Se é nota la distribuzione delle freq. relative della** $Y$

| **Modalità** | $z_1$ | $\cdots$ | $z_j$ | $\cdots$ | $z_k$ |
|---|---|---|---|---|---|
| **Freq. rel.** | $p_1$ | $\cdots$ | $p_j$ | $\cdots$ | $p_k$ |

se $N$ la dimensione del campione (che in generale potrebbe essere non nota), si ha che

$$p_i = \frac{f_i}{N} \Longrightarrow f_i = N p_i$$

e quindi

$$\bar{Y} = \frac{\sum_{j=1}^{k} f_j z_j}{N} = \frac{\sum_{j=1}^{k} N p_j z_j}{N} = N \frac{\sum_{j=1}^{k} p_j z_j}{N}$$

ossia

$$\bar{Y} = \sum_{j=1}^{k} p_j z_j.$$

**Esempio 6.6.** Se consideriamo la tabella relativa ad 1.4

| **Voto** | 70 | 74 | 76 | 78 | 80 | 82 | 84 | 90 | 94 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|
| **Freq.rel** | 0.15 | 0.1 | 0.2 | 0.05 | 0.15 | 0.1 | 0.05 | 0.05 | 0.1 | 0.05 |

per calcolare la media

$$\bar{Y} = 70 \cdot 0.15 + 74 \cdot 0.1 + 76 \cdot 0.2 + 78 \cdot 0.05 + 80 \cdot 0.15$$
$$+82 \cdot 0.1 + 84 \cdot 0.05 + 90 \cdot 0.05 + 94 \cdot 0.1 + 100 \cdot 0.5 = 79.94$$

- **Formule nel caso di classi.** Se le modalitá sono state raggruppate in classi e si ha una delle tabelle

| **Classi** | $[z_1, z_2)$ | $\cdots$ | $[z_j, z_{j+1})$ | $\cdots$ | $[z_k, z_{k+1})$ |
|---|---|---|---|---|---|
| **Frequenza ass.** | $f_1$ | $\cdots$ | $f_j$ | $\cdots$ | $f_k$ |

| **Classi** | $[z_1, z_2)$ | $\cdots$ | $[z_j, z_{j+1})$ | $\cdots$ | $[z_k, z_{k+1})$ |
|---|---|---|---|---|---|
| **Frequenza rel.** | $p_1$ | $\cdots$ | $p_j$ | $\cdots$ | $p_k$ |

per definire $\bar{Y}$ si sostituisce ad ogni classe il valore medio

$$\bar{z}_j = \frac{z_j + z_{j+1}}{2}.$$

Si applicano quindi le formule già viste alla tabelle

| **Classi** | $\bar{z}_1$ | $\cdots$ | $\bar{z}_j$ | $\cdots$ | $\bar{z}_k$ |
|---|---|---|---|---|---|
| **Frequenza ass.** | $f_1$ | $\cdots$ | $f_j$ | $\cdots$ | $f_k$ |
| **Frequenza rel.** | $p_1$ | $\cdots$ | $p_j$ | $\cdots$ | $p_k$ |

ottenendo

$$\bar{Y} = \frac{\sum_{i=1}^{k} f_j \bar{z}_j}{\sum_{j=1}^{k} f_j}$$

e

$$\bar{Y} = \sum_{i=1}^{k} p_j \bar{z}_j.$$

**Esempio 6.7.** Con riferimento all'indagine su un gruppo di laureati in Biologia sul voto $Y$ conseguito all'esame di Matematica si aveva la tabella

| **Voto** | $[18, 20)$ | $[20, 22)$ | $[22, 24)$ | $[24, 26)$ | $[26, 28)$ | $[28, 30]$ |
|---|---|---|---|---|---|---|
| **Freq. rel.** | 0.21 | 0.25 | 0.3 | 0.16 | 0.06 | 0.02 |

Per calcolare il voto medio $\bar{Y}$ conseguito all'esame di Matematica, sostituiamo ad ogni classe il suo valore medio:

| **Voto** | 19 | 21 | 23 | 24 | 27 | 29 |
|---|---|---|---|---|---|---|
| **Freq. rel.** | 0.21 | 0.25 | 0.3 | 0.16 | 0.06 | 0.02 |

Quindi, usando le frequenze relative, si ottiene che:

$$\bar{Y} = 0.21 \cdot 19 + 0.25 \cdot 21 + 0.3 \cdot 23 + 0.16 \cdot 25 + 0.06 \cdot 27 + 0.02 \cdot 29 =$$

$$= 3.99 + 5.25 + 6.9 + 4 + 1.62 + 0.58 = 22.34.$$

6.2.2. **Proprietá della media:**

(1) Se $Y$ é una variabile statistica discreta e $\{z_1, z_2, \cdots, z_k\}$ sono le sue modalitá con $z_1 < z_2 < \cdots < z_k$ allora

$$z_1 \leq \bar{Y} \leq z_k$$

ossia la media é compresa tra il più piccolo e il più grande dei valori osservati. Basta infatti osservare che

$$z_1 = (\sum_{j=1}^{k} p_j)z_1 = \sum_{j=1}^{k} p_j z_1 \leq \sum_{j=1}^{k} p_j z_j = \bar{Y} \leq \sum_{j=1}^{k} p_j z_k = z_k.$$

(2) Sia $Y$ una variabile statistica e sia $W = Y - \bar{Y}$ la variabile "scarto di $Y$ dalla sua media". Allora $\overline{W} = 0$. Infatti

$$\overline{W} = \sum_{i=1}^{k} p_i(y_i - \bar{Y}) = \sum_{i=1}^{k} p_i y_i - (\sum_{i=1}^{k} p_i)\bar{Y} = \bar{Y} - \bar{Y} = 0$$

(3) Se $X = (x_1, x_2, \cdots, x_n)$ e $Y = (z_1, z_2, \cdots, z_k)$ sono due variabile statistiche discrete e

$$Z = (x_1, x_2, \cdots, x_n, z_1, z_2, \cdots, z_k)$$

é quella ottenuta riunendo i due campioni allora

$$\bar{Z} = \frac{n\bar{X} + k\bar{Y}}{n + k}.$$

6.3. **La mediana.** La media risente della presenza di osservazioni anomale. Un solo dato sensibilmente diverso dagli altri può spostare la media in modo significativo. Quando i dati non sono equamente distribuiti attorno alla media, può essere utile riassumerli con un altro valore: la mediana.

Essa si può calcolare per una variabile qualitativa ordinale o quantitativa $Y$ e si indica con $y_{0.5}$.

**Definizione 6.8.** *Sia $Y = (y_1, y_2, \cdots, y_N)$ una variabile statistica con*

$$y_1 \leq y_2 \leq \cdots \leq y_N.$$

*La* **mediana** *corrisponde a quel valore $y_{0.5}$ tale che esattamente* **metà dei dati siano minori o uguale a** $y_{0.5}$ *e* **metà maggiori o uguale a** $y_{0.5}$.

Quindi $y_{0.5}$ é il dato che si trova "a metà " (non é la media!!).
Per calcolarla, vanno messi in ordine crescente gli $N$ dati raccolti sul campione e

- se $N$ é dispari si prende il dato centrale (corrispondente al posto $\frac{N+1}{2}$)
- se $N$ é pari si prende la media aritmetica dei dati centrali ossia si fa la media dei dati che occupano il posto $\frac{N}{2}$ e $\frac{N}{2} + 1$.

La mediana non é influenzata dai valori estremi anomali (anche se dipende dal numero di dati raccolti) ed é poco sensibile ad errori dei dati. Nel gergo statistico si dice che é un indice di posizione robusto.

**Esempio 6.9.** Supponiamo di aver raccolto $N = 9$ (dispari) dati per lunghezza dell'assone, mettiamoli in ordine crescente

$$\underbrace{69.28\,, 69.62,\ 70.78,\ 70.97,}_{\text{quattro dati}}\ \underbrace{71.71}_{\substack{\text{quinto}\\ \text{posto}}}\ \underbrace{,72.31,\ 72.76,\ 73.47\,, 74.03}_{\text{quattro dati}}$$

Quindi la mediana é data dal valore 71.71.

Se facciamo un'altra misurazione e otteniamo il dato 76, $N$ diventa 10 (pari). Quindi

$$\underbrace{69.28\,, 69.62,\ 70.78,\ 70.97,}_{\text{quattro dati}}\ \underbrace{71.71}_{\substack{\text{quinto}\\ \text{posto}}}\ \underbrace{,72.31}_{\substack{\text{sesto}\\ \text{posto}}}\ \underbrace{72.76,\ 73.47,\ 74.03,\ 76}_{\text{quattro dati}}.$$

Quindi la mediana é data dalla media

$$\frac{71.71 + 72.31}{2} = 72,01$$

e anche se l'ultimo dato fosse molto grande , la mediana non subirebbe grosse variazioni.

6.3.1. **Calcolo della mediana nel caso delle classi.**

- **Se sono note le freq. assolute con $N$ dispari.**

  **Esempio 6.10.** Da un'indagine condotta su $N = 201$ laureati in Ingegneria si riscontrano i seguenti voti conseguiti all'esame di Matematica:

| Voto | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F.ass. | 30 | 25 | 25 | 20 | 24 | 18 | 10 | 18 | 12 | 8 | 5 | 3 | 3 |

  Per calcolare la mediana, dobbiamo ordinare tutti i dati (ripetuti tante volte quanto compaiono sul campione) in modo crescente e prendere il $101^{mo}$ dato, in quanto esso lascia alla sua destra 100 dati e altrettanti alla sua sinistra.

  Il 18 compare 30 volte, il 19 compare 25 volte, il 20 compare 25 volte e il 21 compare 20 volte. Poiché

$$30 + 25 + 25 + 20 = 100 \text{ e } 30 + 25 + 25 + 20 + 24 = 124$$

  il $101^{mo}$ dato "cade" nella modalitá 22.

  Quindi la mediana é 22.

- **Se sono note le freq. assolute con $N$ pari.**

  **Esempio 6.11.** Da un'indagine condotta su $N = 200$ laureati in Ingegneria si riscontrano i seguenti voti conseguiti all'esame di Matematica:

| Voto | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F. ass. | 30 | 25 | 25 | 22 | 21 | 19 | 10 | 18 | 12 | 8 | 5 | 3 | 2 |

  Per calcolare la mediana, dobbiamo prendere il $100^{mo}$ e il $101^{mo}$ dato e farne la media.

  Poiché

$$30 + 25 + 25 = 80 \text{ e } 30 + 25 + 25 + 22 = 102$$

  il dato $100^{mo}$ e il $101^{mo}$ cadono nella modalitá 21.

Quindi la mediana é

$$\frac{21+21}{2} = 21$$

.

Se la tabella delle frequenze fosse invece

| Voto | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F.ass. | 30 | 25 | 25 | 20 | 24 | 18 | 10 | 18 | 12 | 8 | 5 | 3 | 2 |

poiché

$$30 + 25 + 25 + 20 = 100 \text{ e } 30 + 25 + 25 + 20 + 24 = 124$$

il dato $100^{mo}$ ha modalitá 21 mentre il dato $101^{mo}$ ha modalitá 22. Quindi la mediana é $\frac{21+22}{2} = 21.5$.

6.4. **Riunione del campione.** Supponiamo di avere un campione con $n = 2$ studenti che hanno preso in media 24 all'esame di Matematica, mentre in un campione di $k = 10$ studenti tutti hanno preso 30.

Qual é la media del voto all'esame di Matematica sul campione di $n + k = 2 + 10 = 12$ studenti? La formula corretta non é

$$\frac{24+30}{2} = 27$$

ma

$$\frac{2 \cdot 24 + 10 \cdot 30}{12} = 29.$$

6.5. **Media ponderata.** In alcuni casi, invece della media aritmetica (quella usata finora) si può decidere di calcolare un valor medio tra i valori $y_1, \cdots, y_N$ che tenga conto che ogni elemento $y_i$ ha un'"importanza" diversa, un "peso" diverso $p_i$.

In tal caso la media si chiama **ponderata** si calcola come

$$\frac{\sum_{i=1}^{N} p_i \cdot y_i}{\sum_{i=1}^{N} p_i}.$$

Per esempio la media degli esami all'università é una media ponderata, ossia calcolata dando ad ogni esame un peso che é dato dal numero di crediti dell'esame.

I crediti di un esame non rappresentano solo quante ore sono dedicate alla frequenza dell'esame ma sono quindi anche il "peso" nel calcolo del voto di presentazione alla tesi di laurea.

Per esempio Matematica = 6CFU. La media (per il calcolo del voto di presentazione alla laurea) é data da

$$\frac{\sum_{i=1}^{N}(\text{CFU dell'i-mo esame}) \cdot (\text{voto dell'i-mo esame})}{\text{numero totale dei crediti degli esami}}$$

dove

$$N = \text{numero di esami}$$

## 7. La varianza

Se $Y = (y_1, y_2, \cdots, y_N)$ é una variabile statistica discreta con media $\bar{Y}$ allora i valori

$$(y_i - \bar{Y})^2$$

sono detti **scarti quadratici**. Sono gli "errori" rispetto alle media.

**Definizione 7.1.** *Si chiama* **varianza** *di $Y$ la media degli scarti quadratici $(y_i - \bar{Y})^2$.*

La varianza di $Y$ viene indicata con $\sigma_Y^2$ o più semplicemente con $\sigma^2$.
Dalla definizione si ha

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \underbrace{(y_i - \bar{Y})^2}_{\text{scarti quadratici}} \quad . \tag{7.1}$$

Tanto più piccola é la varianza di $Y$ tanto più i dati sono concentrati intorno al suo valor medio $\bar{Y}$.

**Definizione 7.2.** *Si definisce* **deviazione standard** *o* **scarto quadratico medio** *di $Y$ la radice quadrata della sua varianza.*

Pertanto la deviazione standard viene indicata con $\sigma_Y$.

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{Y})^2}{N}}.$$

Nell'ambito della statistica inferenziale, a volte si rimpiazza il denominatore $N$ con $N-1$ ottenendo:

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \bar{Y})^2}{N-1}}.$$

Noi useremo sempre la formula con $N$ a denominatore.

**Esempio 7.3.** In un esperimento vengono pesate 15 cavie ottenendo i seguenti pesi in grammi:

$$28, 32, 37, 29, 31, 30, 26, 32, 27, 29, 30, 32, 28, 31, 31.$$

Per usare la formula (7.1) (senza usare le frequenze) dobbiamo calcolare tutti gli scarti quadratici:

$$(28 - 30.2)^2, (32 - 30,2)^2, (37 - 30.2)^2, \cdots (31 - 30.2)^2, (31 - 30.2)^2$$

ossia

$$4.84, 3.24, 46.24, \cdots, 0.64, 0.64$$

da cui segue che

$$\sigma^2 = \frac{4.84 + 3.24 + 46.24 \cdots + 0.64 + 0.64}{15} g^2 = 6.56 g^2.$$

Quindi la deviazione standard é $\sigma = 2.56g$.

**Osservazione 7.4.** Se usiamo la tabella con le frequenze assolute

| Modalità | $z_1$ | $\cdots$ | $z_j$ | $\cdots$ | $z_k$ |
|---|---|---|---|---|---|
| Frequenza ass. | $f_1$ | $\cdots$ | $f_j$ | $\cdots$ | $f_k$ |

dopo aver calcolato la media $\bar{Y}$, possiamo costruire la tabella

| Scarti quadratici | $(z_1-\bar{Y})^2$ | $\cdots$ | $(z_j-\bar{Y})^2$ | $\cdots$ | $(z_k-\bar{Y})^2$ |
|---|---|---|---|---|---|
| Frequenza ass. | $f_1$ | $\cdots$ | $f_j$ | $\cdots$ | $f_k$ |

e quindi calcolare la varianza con la formula

$$\sigma^2 = \frac{\sum_{i=1}^k f_i(z_i-\bar{Y})^2}{N}.$$

**Esempio 7.5.** Nell'esempio precedente 7.3

| **Peso** | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 37 |
|---|---|---|---|---|---|---|---|---|
| **Scarti quadratici** | 17.64 | 10.24 | 4.84 | $\cdots$ | $\cdots$ | $\cdots$ | 3.24 | 46.24 |
| **F.ass.** | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 1 |

e quindi

$$\sigma^2 = \frac{1\cdot 17.64 + 10.24\cdot 1 + 4.84\cdot 2 + \cdots + 3.24\cdot 3. + 46.24\cdot 1}{15} = 6.56.$$

7.1. **Formula "furba".** Sviluppando i quadrati in (7.1) si trova:

$$\frac{1}{N}\sum_{i=1}^N (y_i-\bar{Y})^2 = \frac{1}{N}\sum_{i=1}^N (y_i^2 - 2y_i\bar{Y} + \bar{Y}^2)$$

$$= \frac{1}{N}(\sum_{i=1}^N y_i^2 - 2\sum_{i=1}^N y_i\bar{Y} + \sum_{i=1}^N \bar{Y}^2) = \frac{1}{N}(\sum_{i=1}^N y_i^2 - 2N\bar{Y}\cdot\bar{Y} + N\bar{Y}^2)$$

$$\sigma^2 = \overline{Y^2} - \overline{Y}^2 (\geq 0). \tag{7.2}$$

Quindi la varianza é la **media dei quadrati meno il quadrato della media**.

Quindi, allo scopo di usare la formula (7.2), si devono calcolare

- la media di $Y$ per poi farne il quadrato: otteniamo quindi $\overline{Y}^2$
- scrivere la tabella di $Y^2$ calcolando i quadrati delle singole modalità
- calcolare la media di $Y^2$, ossia $\overline{Y^2}$

**Esempio 7.6.** Nell'esempio precedente 7.3 allo scopo di usare la formula (7.2)

- si devono calcolare tutti i quadrati delle modalitá

| **Peso** | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 37 |
|---|---|---|---|---|---|---|---|---|
| **Peso$^2$** | 676 | 729 | 784 | 841 | 900 | 961 | 1024 | 1369 |
| **F.ass.** | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 1 |

- calcolare la media dei pesi al quadrato

$$\overline{Y^2} = \frac{676\cdot 1 + 729\cdot 1 + 784\cdot 2 + \cdots 961\cdot 3 + 1024\cdot 3 + 1369\cdot 1}{15} = 918.6g^2$$

Quindi

$$\sigma^2 = \overline{Y^2} - (\bar{Y})^2 = 918.06 - (30.2)^2 = 918.06 - 912.04 = 6.56g^2$$

**Osservazione 7.7.** (1) Se usiamo la tabella con le frequenze relative

| Modalità | $z_1$ | $\cdots$ | $z_j$ | $\cdots$ | $z_k$ |
|---|---|---|---|---|---|
| Frequenza ass. | $p_1$ | $\cdots$ | $p_j$ | $\cdots$ | $p_k$ |

si ottiene

$$\sigma^2 = \sum_{i=1}^{k} p_i (z_i - \bar{Y})^2.$$

(2) Se infine $Y$ é una variabile statistica continua e si dispone della tabella di frequenza con le modalità raggruppate in classi $[z_i, z_{i+1})$ per definire $\sigma^2$ é necessario calcolare il punto medio $\bar{z}_i = \frac{z_i + z_{i+1}}{2}$ di ogni classe. Quindi

$$\sigma^2 = \sum_{i=1}^{k} p_i (\bar{z}_i - \bar{Y})^2$$

dove $p_i$ é la frequenza associata alla classe $[z_i, z_{i+1})$.

7.2. **Proprietá della varianza.**

- per definizione $\sigma^2 \geq 0$;
- $\sigma^2(Y + b) = \sigma^2(Y)$ per ogni $b \in \mathbb{R}$, ossia la varianza é invariante per traslazioni;
- $\sigma^2(aY + b) = a^2\sigma^2(Y)$ per ogni $a \in \mathbb{R}$.

**Definizione 7.8.** *Una variabile $Y$ tale che $\bar{Y} = 0$ si dice* **centrata**. *Una variabile $Y$ tale che $\bar{Y} = 0$ e $\sigma^2 = 1$ si dice* **standardizzata**.

Come conseguenza delle proprietà precedenti

- data la variabile standardizzata $Y$ allora la variabile $Z = aY + b$ é tale che $\bar{Z} = b$ e $\sigma(Z) = a$
- data la variabile $Y$ allora $Z = \frac{Y - \bar{Y}}{\sigma_Y}$ é una variabile standardizzata.

(Francesca Prinari) Dipartimento di Matematica e Informatica, Universitá di Ferrara
*Email address*, Francesca Prinari: francesca.prinari@unife.it