# A Statistica descrittiva

## A.1 Indagine statistica

Il termine statistica significa *scienza dello stato*. Questo termine venne usato per la prima volta nel XVI secolo per indicare lo studio dei dati utili al governo degli stati prevalentemente relativi a fenomeni di carattere demografico (nascite, morti, ecc.). Negli anni, la statistica si è estesa ai campi più disparati: fisica, psicologia, ricerca di mercato, indici di gradimento, sondaggi, meteorologia, ... È nata essenzialmente con lo scopo di descrivere i fenomeni (statistica descrittiva), successivamente è divenuta uno strumento utile anche per fare previsioni (statistica inferenziale). A grandi linee si può definire come la scienza che si occupa della raccolta e dell'analisi dei dati relativi ad un certo gruppo di persone, animali o oggetti al fine di descrivere in maniera sintetica un fenomeno che li riguarda e fare eventualmente previsioni sul suo andamento futuro.

Ad esempio, la statistica cerca di rispondere a domande del tipo:

- quanta acqua sarà necessaria in Italia fra 3 anni?
- quanta corrente elettrica sarà necessaria per il fabbisogno nazionale fra 5 anni?
- quale sarà il tasso di disoccupazione nazionale fra 1 anno?

**Definizione A.1.** L'insieme di elementi oggetto dell'indagine statistica è detta *popolazione* o universo, mentre ciascun elemento della popolazione è detto *unità statistica*.

Sono esempi di popolazione statistica gli abitanti di una città in un certo anno, i prezzi di un determinato bene, le temperature massime registrate in una giornata in un particolare luogo, i ciclomotori circolanti in Italia, gli alunni di una scuola.

**Definizione A.2.** Per ogni unità statistica si possono studiare una o più caratteristiche ed ognuna di tali caratteristiche costituisce un *carattere* della popolazione oggetto di indagine. I caratteri possono essere di tipo *qualitativo* o *quantitativo*. Si definisce *modalità* del carattere indagato ciascuno dei diversi modi in cui esso può presentarsi.

Sono esempi di carattere qualitativo il colore degli occhi, il colore dei capelli, il tipo di scuola frequentato, il gradimento di un certo programma televisivo. Le modalità di un carattere qualitativo sono espresse mediante nomi o aggettivi. I caratteri qualitativi sono a loro volta suddivisi in *ordinabili*, cioè può essere definita una relazione di ordine tra essi (per ogni coppia di elementi si può stabilire quale dei due è il primo e quale il secondo – es. il tipo di scuola frequentato è ordinabile a partire dalla scuola dell'infanzia fino alla laurea, il gradimento di un programma televisivo è ordinabile a partire dalla completa mancanza di gradimento fino al gradimento massimo) e *non ordinabili* o *sconnessi* (es. colore degli occhi, colore dei capelli).

Sono invece caratteri quantitativi l'età, l'altezza, il numero di auto prodotte da una fabbrica, ..., ovvero le modalità di un carattere quantitativo sono espresse mediante numeri. I caratteri quantitativi possono essere di tipo *discreto*, quando assumono solo valori puntuali, oppure di tipo *continuo*, quando possono assumere tutti gli infiniti valori compresi in un determinato intervallo. Sono esempi di caratteri quantitativi discreti il numero di figli in una famiglia, i pezzi prodotti in una catena di montaggio; sono esempi di caratteri quantitativi continui l'altezza di una persona, il peso di una persona, la lunghezza di un fiume.

L'indagine statistica può riguardare l'intera popolazione (in tal caso si parla di *censimento*) oppure solo una sua parte (in tal caso si parla di *indagine a campione*). Supponiamo di voler effettuare un'indagine relativa alle persone che fumano in Italia. Il fenomeno collettivo in esame è il fumo, la popolazione di riferimento è costituita dalla popolazione italiana in età adulta, l'unità statistica è rappresentata da ogni cittadino oggetto dell'indagine, i caratteri oggetto dell'indagine possono essere "fumatore/non fumatore", "numero di sigarette fumate", che cosa si fuma (es. pipa, sigaro, sigaretta). Data l'elevata numerosità della popolazione di riferimento la tipologia di indagine preferibile è quella a campione.

A sua volta, l'indagine a campione può essere effettuata su un *campione casuale*, quando si scelgono a caso i campioni all'interno della popolazione o su un *campione stratificato*, quando si suddivide la popolazione in *classi* o *strati* senza specifici criteri e per ogni strato si prende a caso un campione.

✍ *Esercizio proposto:* A.1

## A.2 Fasi di un'indagine statistica

**Definizione A.3.** Dato un carattere oggetto di rilevazione, si definisce *frequenza* il numero delle unità statistiche su cui una sua modalità si presenta.

Affinché un'indagine statistica sia rigorosa (e quindi garantisca un'elevata affidabilità) è necessario che sia strutturata secondo le seguenti fasi:

a) **Studio del problema e impostazione dell'indagine statistica.** Si individua in maniera precisa lo scopo della ricerca, il fenomeno sul quale indagare, la popolazione statistica di riferimento, le singole unità statistiche ed il carattere, o caratteri, oggetto di indagine.
b) **Rilevazione dei dati statistici.** La rilevazione non è altro che la raccolta dei dati statistici riguardanti ogni elemento della popolazione e relativi al fenomeno che si vuole analizzare. La rilevazione può avvenire secondo diverse modalità:

   *rilevazione diretta o globale:* viene eseguita direttamente su tutte le unità statistiche che formano la popolazione;

   *rilevazione indiretta o parziale:* eseguita solo su una parte della popolazione. Si deve scegliere in tal caso un sottoinsieme della popolazione, detto *campione*, che deve essere rappresentativo della popolazione di riferimento, ovvero deve essere il più possibile eterogeneo rispetto alle caratteristiche della popolazione e contenere al suo interno un numero non troppo ristretto di unità.

c) **Spoglio delle schede e tabulazione.** Contemporaneamente o successivamente al rilevamento, i dati raccolti vengono ordinati, suddivisi in classi omogenee e riassunti tramite tabelle dette *tabelle statistiche*.

d) **Rappresentazione dei dati statistici.** La rappresentazione può avvenire attraverso diversi tipi di grafico:

*diagramma cartesiano:* rappresentazione nel piano cartesiano dei valori della variabile sull'asse orizzontale e delle relative frequenze sull'asse verticale;

*ideogramma:* si rappresenta un certo numero di dati con un simbolo;

*diagramma a barre o a colonne:* grafico composto da segmenti o barre (orizzontali o verticali) proporzionali alle frequenze;

*areogramma:* grafico a forma di cerchio composto da settori circolari con aree direttamente proporzionali alle frequenze;

*istogramma:* grafico composto da rettangoli aventi area proporzionale alla frequenza.

e) **Elaborazione dei dati.** Con specifici algoritmi di calcolo, vengono elaborati i dati tabulati al fine di costruire opportuni indici di sintesi.

f) **Interpretazione dei risultati.** Attraverso i grafici e gli indici è possibile descrivere le caratteristiche peculiari del fenomeno analizzato.

Analizziamo in dettaglio le singole fasi che seguono la raccolta dei dati.

### A.2.1 Spoglio delle schede e tabulazione

Dopo aver raccolto i dati per ciascuna modalità del carattere o per ciascuna classe individuata si deve determinare:

- ➡ la *frequenza assoluta*, cioè il numero di volte con cui si presenta una modalità del carattere indagato;
- ➡ la *frequenza relativa*, cioè il rapporto tra la frequenza assoluta e il numero totale dei casi presi in esame;
- ➡ la *frequenza percentuale*, cioè la frequenza relativa moltiplicata per 100.

Si compila poi una tabella di frequenza che sintetizza la raccolta dei dati, come nell'esempio seguente.

---

**Esempio A.1.** La tabella seguente fornisce la distribuzione di frequenze assolute degli alunni di una classe rispetto al carattere sesso.

| Sesso | Femmine | Maschi | Totale |
|---|---|---|---|
| Numero di alunni | 15 | 12 | 27 |

Per costruirla, si è operata la classificazione della popolazione degli alunni della classe rispetto ad un determinato carattere (il sesso), sono state individuate le modalità con cui questo si è manifestato (femmina, maschio) ed è stato effettuato il conteggio delle unità in corrispondenza di ciascuna modalità (frequenza assoluta). Dalle frequenze assolute si ricavano le frequenze relative: 15 alunni su 27 sono femmine: la frazione è di 15/27 femmine sul totale degli alunni. Quindi Dall'operazione 15 diviso 27 otteniamo 0,56 (approssimando a due cifre decimali) che è la frequenza relativa. La frazione può essere espressa in forma percentuale: 0,56 equivale a dire 56 su 100 ed è consuetudine scriverlo in forma percentuale 56%. Tale valore è la frequenza percentuale.

Ripetendo lo stesso procedimento per i maschi si ottiene la seguente tabella delle frequenze:

| Sesso | Frequenza assoluta | Frequenza relativa | Frequenza percentuale |
|---|---|---|---|
| Femmine | 15 | 0,56 | 56% |
| Maschi | 12 | 0,44 | 44% |

Si può concludere che la classe è formata per il 56% da femmine e per il 44% da maschi.

**Esempio A.2.** Supponiamo che i voti elencati di seguito siano quelli riportati in matematica a fine trimestre dagli alunni della tua classe: 5, 4, 6, 8, 8, 7, 7, 6, 5, 5, 6, 7.

Per poter effettuare una lettura più agevole si costruisce una tabella in cui vengono riportati sulla prima colonna i singoli valori rilevati (le modalità del carattere) in ordine crescente, nella seconda la frequenza assoluta, cioè quante volte compare quel determinato voto, nella terza la frequenza relativa e nella quarta quella percentuale (che si ottiene moltiplicando per 100 la frequenza relativa):

| Voto riportato | Frequenza assoluta | Frequenza relativa | Frequenza percentuale |
|---|---|---|---|
| 4 | 1 | $1/12 = 0{,}083$ | 8,30% |
| 5 | 3 | $3/12 = 0{,}25$ | 25,00% |
| 6 | 3 | $3/12 = 0{,}25$ | 25,00% |
| 7 | 3 | $3/12 = 0{,}25$ | 25,00% |
| 8 | 2 | $2/12 = 0{,}167$ | 16,70% |
| Totale | 12 | $12/12 = 1$ | 100% |

**Esempio A.3.** Misurando l'altezza di un gruppo di cani di razza pastore italiano si sono ottenute le seguenti misure in cm:

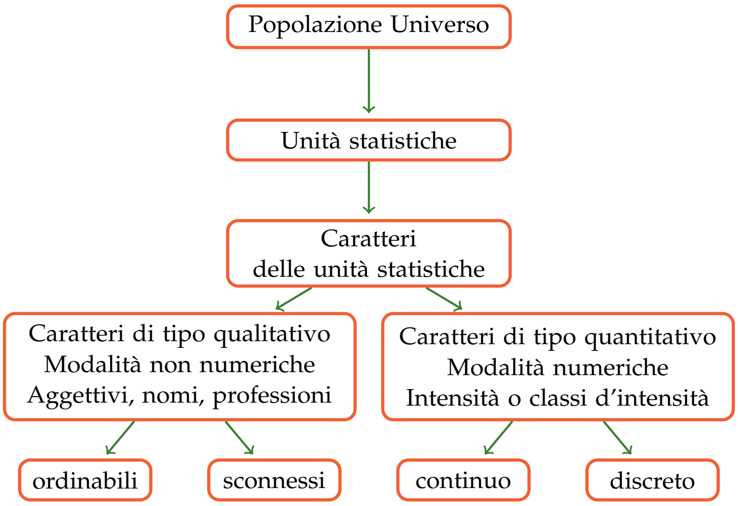
| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 57,1 | 60,8 | 60,7 | 56,2 | 59,5 | 62,4 | 56,1 | 61,2 | 54,5 | 64,5 | 57,5 | 58,3 | 55,2 |
| 58,7 | 57,2 | 56,1 | 58,9 | 57,7 | 53,2 | 59,2 | 58,9 | 54,5 | 55,3 | 62,1 | 59,0 | 58,3 |
| 61,3 | 60,1 | 56,4 | 60,2 | 61,7 | 57,3 | 58,3 | 59,5 | 62,6 | 59,4 | 58,3 | 59,4 | 59,4 |
| 59,3 | 57,6 | 60,0 | 60,7 | 56,7 | 61,1 | 59,8 | 55,3 | 63,9 | 58,0 | 55,2 | 54,9 | 53,8 |

Il carattere indagato nella popolazione cani pastore italiano è di tipo quantitativo continuo; con questo tipo di dati è praticamente impossibile calcolare le frequenze se le altezze non si raggruppano in classi.

Vediamo come procedere: osservando i dati ottenuti si nota che il valore minore è 53,8 mentre il valore maggiore è 64,7. Possiamo allora suddividere i dati in gruppi partendo da 53,0 cm fino a 65,0 cm, formando classi di ampiezza 1cm e ottenendo la seguente tabella:

| Classe (cm) | Frequenza assoluta | Frequenza percent. | Classe (cm) | Frequenza assoluta | Frequenza percent. |
|---|---|---|---|---|---|
| 53,0 - 53,9 | 2 | 3,85% | 59,0 - 59,9 | 9 | 17,31% |
| 54,0 - 54,9 | 3 | 5,77% | 60,0 - 60,9 | 6 | 11,54% |
| 55,0 - 55,9 | 4 | 7,69% | 61,0 - 61,9 | 4 | 7,69% |
| 56,0 - 56,9 | 5 | 9,61% | 62,0 - 62,9 | 3 | 5,77% |
| 57,0 - 57,9 | 6 | 11,54% | 63,0 - 63,9 | 1 | 1,92% |
| 58,0 - 58,9 | 8 | 15,38% | 64,0 - 64,9 | 1 | 1,92% |

Riassumendo




*Esercizi proposti:* A.2, A.3, A.5, A.6, A.7, A.8, A.9

### A.2.2 Rappresentazione grafica

La rappresentazione grafica dei dati statistici facilita notevolmente lo studio delle caratteristiche del fenomeno che si sta esaminando; infatti dopo aver impostato l'indagine, raccolto, classificato ed elaborato i dati nelle tabelle, i dati non sempre si presentano in una forma di facile lettura ed il loro significato e la loro interpretazione rimane poco chiara. Attraverso la rappresentazione grafica, i risultati dell'indagine emergono immediatamente, in maniera diretta e sintetica.

La rappresentazione grafica può avvenire utilizzando diversi tipi di grafico a seconda delle caratteristiche da analizzare.

#### Diagramma cartesiano

La rappresentazione grafica attraverso un diagramma cartesiano dà, in modo immediato, informazioni sull'andamento globale del fenomeno e viene utilizzata prevalentemente per la rappresentazione di serie storiche (per esempio, per rappresentare il numero di auto prodotte per anno da una fabbrica) oppure quando si hanno due caratteri quantitativi e si vuol analizzare il tipo di legame esistente fra di essi.

---

**Esempio A.4.** Consideriamo la tabella statistica relativa alla domanda "quante ore al giorno passi al computer?", posta ad un campione di 50 ragazzi dai 16 ai 24 anni.

Rappresentiamo la tabella attraverso un diagramma cartesiano costruito tracciando due rette perpendicolari, gli assi, quello verticale orientato verso l'alto e quello orizzontale orientato verso destra. Riportiamo sull'asse orizzontale il numero di ore e sull'asse verticale il numero di ragazzi e determiniamo i punti aventi come coordinate (numero ore; numero ragazzi).

Il punto A avrà come coordinate $(0;4)$, il punto B avrà come coordinate $(1;6)$ e così via. Uniamo poi i punti con segmenti e otteniamo il diagramma cartesiano (grafico A.1). Precisamente $A(0;4)$, $B(1;6)$, $C(2;12)$, $D(3;16)$, $E(4;8)$, $F(5;4)$, $G(6;2)$.

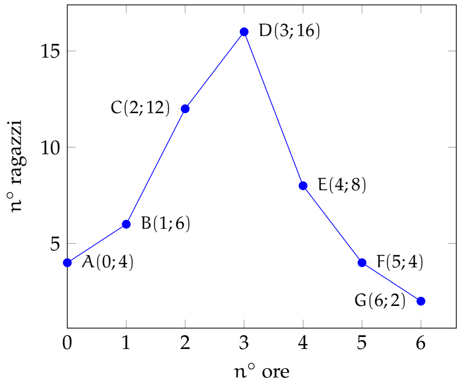



Grafico A.1: Esempio A.4

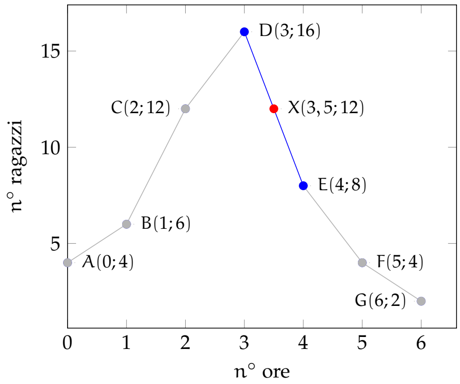



Grafico A.2: Esempio A.4

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Numero di ore | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| Numero di ragazzi | 4 | 6 | 12 | 16 | 8 | 4 | 2 |

Dal grafico A.2 si può notare immediatamente che la maggior parte dei ragazzi trascorre dalle 2 alle 3 ore al computer dato che il picco più alto si ha proprio nei punti C e D. Si può anche notare che, ad esempio, il punto X di coordinate $(3{,}5; 12)$, appartenente al segmento di congiunzione tra i punti D ed E, non ha significato reale, dato che le sue coordinate non sono riportate nella tabella statistica del fenomeno da studiare.

---

### Ideogramma

Nella rappresentazione grafica attraverso *ideogramma* si rappresenta un certo numero di dati con un simbolo che si assume come *unità grafica*; il simbolo deve richiamare l'oggetto dell'indagine e dare quindi una visione immediata del fenomeno. Ad esempio si può utilizzare un uomo stilizzato per rappresentare un dato riguardante il numero di persone che vivono in un determinato territorio, una macchina per la produzione annua di automobili in una fabbrica, e così via. Tale tipo di rappresentazione è spesso usata in campo pubblicitario perché caratterizzata da un evidente impatto visivo.

---

**Esempio A.5.** Un istituto scolastico ha visto aumentare i suoi iscritti, dall'anno scolastico 2003-2004 all'anno 2008-2009 secondo quanto riportato nella seguente tabella:

| Anno scolastico | 2003-04 | 2004-05 | 2005-06 | 2006-07 | 2007-08 | 2008-09 |
|---|---|---|---|---|---|---|
| Iscritti | 150 | 200 | 200 | 325 | 375 | 450 |

Possiamo rappresentare mediante ideogramma i dati contenuti nella tabella statistica. Consideriamo una faccina stilizzata come unità grafica assegnandole il valore di 50 ragazzi iscritti.

 = 50 iscritti

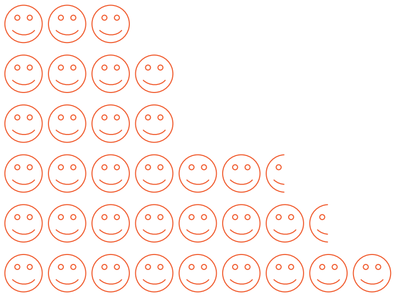
Il numero degli iscritti di ogni anno scolastico sarà rappresentato da tante unità grafiche quanti sono i gruppi di 50 iscritti. Per avere il grafico relativo all'anno 2003-2004 si devono usare tre faccine, in quanto $150 : 50 = 3$.

a.s. 2003-2004 = 

Se la divisione del numero degli iscritti per 50 dà resto, esso si dovrà rappresentare disegnando solo una parte dell'unità grafica, corrispondente alla frazione tra resto e 50. Ad esempio nell' a.s. 2006-2007 ci sono stati 325 iscritti; $325 : 50 = 6$ col resto di 25, quindi 325 sarà uguale a 6 unità grafiche e $\frac{25}{50} = \frac{1}{2}$ unità grafica, cioè mezza faccina, ovvero $325 : 50 = 6{,}5$ cioè 6 faccine e mezzo.

a.s. 2006-2007 = 

Il grafico completo sarà:



| | | |
|---|---|---|
| a.s. 2003-2004 = | | 3 |
| a.s. 2004-2005 = | | 4 |
| a.s. 2005-2006 = | | 4 |
| a.s. 2006-2007 = | | 6 e 1/2 |
| a.s. 2007-2008 = | | 7 e 1/2 |
| a.s. 2008-2009 = | | 9 |

---

#### Diagramma a barre o a colonne

Questo tipo di rappresentazione, detta anche diagramma a nastri o a bastoni, viene usata quando si vuole fornire un'idea delle frequenze delle diverse modalità di un fenomeno. In genere si usa per caratteri qualitativi o quantitativi discreti. Per poter valutare il significato statistico della lunghezza delle barre (o delle colonne) è necessario scegliere opportunamente una scala di riferimento: la larghezza della barra (o della colonna) è arbitraria ma uguale per tutte le barre (o colonne) e la sua lunghezza è proporzionale alla caratteristica che si deve rappresentare. Le barre (o le colonne) possono inoltre essere suddivise in parti di colori diversi per indicare le singole componenti o i singoli fenomeni che si vogliono analizzare.

La differenza fra la rappresentazione a barre e quella a colonne consiste soltanto nell'orientamento del grafico: nel diagramma a barre si indicano le modalità del carattere sull'asse verticale e le frequenze sull'asse orizzontale, mentre in quello a colonne le modalità del carattere sono riportate sull'asse orizzontale e le frequenze su quello verticale.

Di seguito vengono riportate le due tipologie di grafico accompagnate dalla tabella di riferimento:

| Materia | Italiano | Storia | Geografia | Matem. | Scienze | Ed. Fisica | Totale |
|---|---|---|---|---|---|---|---|
| Maschi | 5 | 4 | 4 | 2 | 6 | 5 | 26 |
| Femmine | 3 | 7 | 2 | 3 | 4 | 5 | 24 |

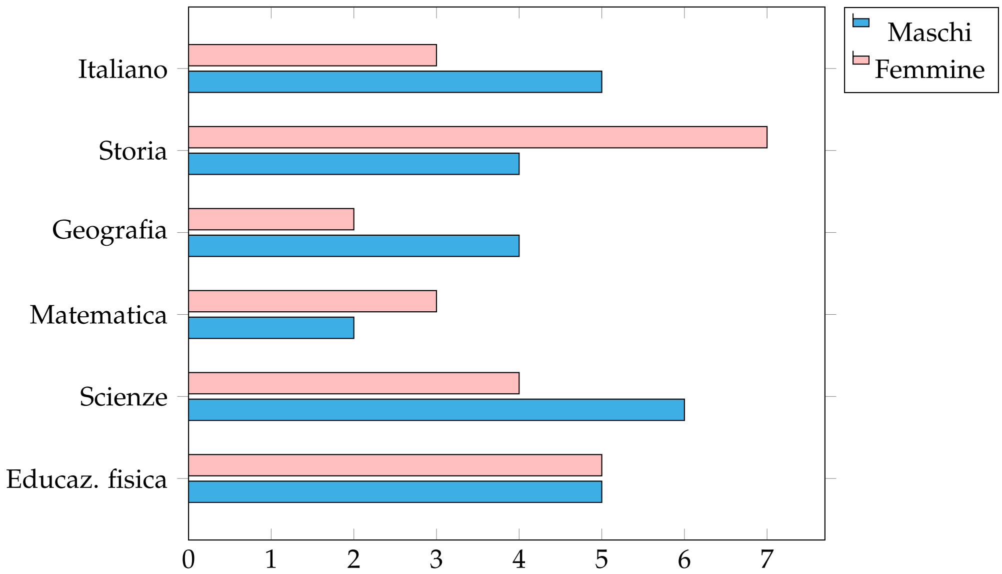



Figura A.1: Diagramma a barre

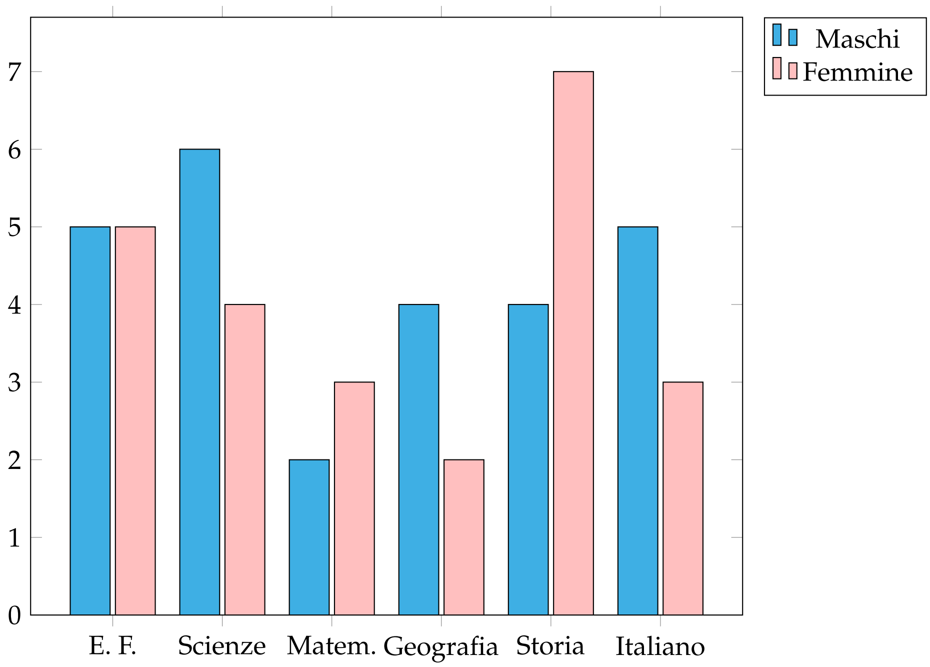



Figura A.2: Diagramma a colonne

#### Areogramma

Questo tipo di rappresentazione, detta anche grafico a torta, viene utilizzato quando si vogliono evidenziare le parti che compongono un fenomeno, per esempio per indicare come si dividono gli alunni di una classe in maschi e femmine, o per rappresentare in che modo le varie voci di spesa incidono sul bilancio familiare. Il grafico si ottiene dividendo un cerchio in settori circolari con aree direttamente proporzionali alle frequenze che rappresentano. Per disegnare l'areogramma, si disegna una circonferenza di diametro arbitrario e si fa corrispondere l'angolo al centro di 360°, con il 100% di frequenza percentuale; per ottenere l'angolo corrispondente ad una certa frequenza percentuale $f_x$ si risolve la proporzione $360° : X° = 100 : f_x \Rightarrow X° = 3{,}6 \cdot f_x$. Si suddivide così la circonferenza negli angoli ottenuti e si evidenziano in maniera differente tra loro i settori circolari ottenuti.

---

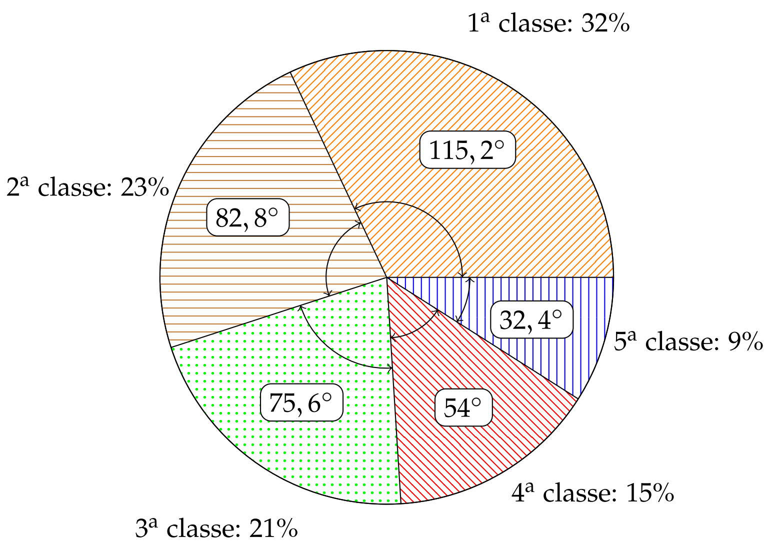
**Esempio A.6.** Consideriamo la seguente tabella statistica che indica gli studenti di un dato istituto scolastico divisi per classe frequentata, in un dato anno.

| Classe | 1ª | 2ª | 3ª | 4ª | 5ª | Totale |
|---|---|---|---|---|---|---|
| Studenti | 320 | 230 | 212 | 152 | 96 | 1 010 |

Nella tabella sono indicate le frequenze assolute; calcoliamo ora le frequenze percentuali degli studenti. Per la 1ª classe si ha: $\frac{320}{1\,010} = 0{,}32$ arrotondato alla seconda cifra decimale, che equivale al 32% e così via per le classi successive.

| Classe | 1ª | 2ª | 3ª | 4ª | 5ª | Totale |
|---|---|---|---|---|---|---|
| Frequenze percentuali | 32% | 23% | 21% | 15% | 9% | 100% |

Rappresentiamo graficamente mediante areogramma i dati contenuti nella tabella precedente.




Per ottenere l'angolo relativo alla frequenza percentuale della 1ª classe si fa: $3{,}6 \cdot 32 = 115{,}2°$ e per la 2ª classe: $3{,}6 \cdot 23 = 82{,}2°$ e così via per le altre classi.

Dal grafico si può notare immediatamente che la classe più frequentata è la prima.

---

### Istogramma

Si utilizza la rappresentazione grafica attraverso istogramma quando il carattere analizzato è di tipo quantitativo ed i dati sono raggruppati in classi.

Prima di tutto si distribuiscono i dati in classi o gruppi e si determina il numero di unità appartenenti a ciascuna classe; questo numero è detto *frequenza della classe*. Riportando tali dati in una tabella si ottiene la distribuzione delle frequenze. Poiché le classi potrebbero avere ampiezze diverse si calcola la *densità di frequenza*, definita come il rapporto fra la frequenza della classe e la relativa ampiezza.

Per disegnare un istogramma si tracciano due assi; sull'asse verticale, orientato verso l'alto, si fissa un segmento unitario e vi si riportano le densità di frequenza. L'asse orizzontale, orientato verso destra, è invece suddiviso in tanti segmenti la cui ampiezza è pari a quella delle singole classi. Il grafico consiste in un insieme di rettangoli aventi per base ogni classe e altezza la densità di frequenza corrispondente. In tal modo l'area di ogni rettangolo rappresenta la frequenza corrispondente a ciascuna classe.

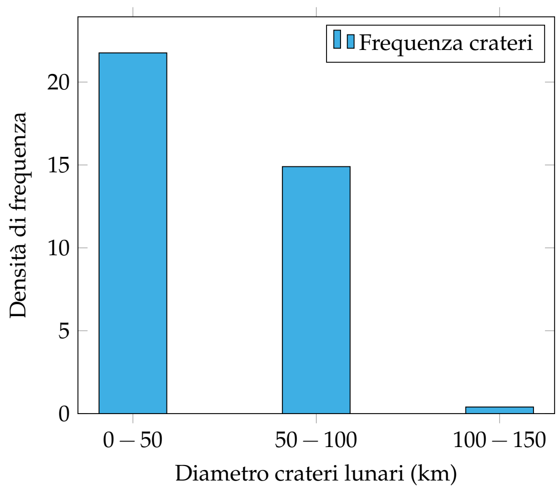
---

**Esempio A.7.** Costruiamo un istogramma a partire dalla distribuzione di frequenza riportata nella seguente tabella:

| Diametro crateri lunari (km) | Numero di crateri |
|---|---|
| $0 - 50$ | 1 088 |
| $50 - 100$ | 745 |
| $100 - 150$ | 20 |

Innanzi tutto per ogni classe dobbiamo determinare la densità di frequenza, che si ottiene dividendo la frequenza assoluta per l'ampiezza della classe:

| Diametro crateri lunari (km) | Densità di freq. |
|---|---|
| $0 - 50$ | $1\,088/50 = 21{,}76$ |
| $50 - 100$ | $745/50 = 14{,}9$ |
| $100 - 150$ | $20/50 = 0{,}4$ |



**Esempio A.8.** Consideriamo la seguente tabella statistica che riporta i giorni di pioggia di ogni mese, in un dato anno e in una data città.

| Mesi | Giorni di pioggia | Mesi | Giorni di pioggia |
|---|---|---|---|
| Gennaio | 15 | Luglio | 1 |
| Febbraio | 10 | Agosto | 3 |
| Marzo | 14 | Settembre | 3 |
| Aprile | 8 | Ottobre | 5 |
| Maggio | 5 | Novembre | 9 |
| Giugno | 2 | Dicembre | 11 |

Dividiamo i mesi dell'anno in classi, raggruppandoli in stagioni. Ad esempio, Luglio, Agosto e Settembre appartengono alla classe dell'Estate e la frequenza di questa classe è data dalla somma delle frequenze di ogni mese, cioè $1+3+3=7$. Si prosegue in questo modo per ogni classe ottenendo così la distribuzione delle frequenze riportata nella tabella.

| Stagioni | Estate | Autunno | Inverno | Primavera |
|---|---|---|---|---|
| Giorni di pioggia | 7 | 25 | 39 | 15 |

Costruisci ora l'istogramma corrispondente alla tabella precedente riportando sull'asse orizzontale le classi (stagioni) e su quello verticale le densità di frequenze.

---

✍ *Esercizi proposti:* A.10, A.11, A.12, A.13, A.14, A.15, A.16, A.17, A.18, A.19, A.20

## A.3 Indici di posizione

Nel caso in cui il carattere considerato nell'indagine sia di tipo quantitativo, l'andamento dei dati raccolti può essere sinteticamente descritto per mezzo di opportuni indici. Gli *indici di posizione* vengono utilizzati per dare un'indicazione sulla distribuzione delle frequenze per mezzo di un solo numero. A seconda del carattere oggetto dell'indagine statistica possono essere utilizzati indici differenti.

### A.3.1 Moda

**Definizione A.4.** La *moda* è la modalità del carattere indagato che si presenta più frequentemente.

In una successione di $n$ modalità $x_1, x_2, \ldots, x_n$ con le relative frequenze $f_1, f_2, \ldots, f_n$, la moda è la modalità che ha la frequenza maggiore. Questo valore può essere calcolato per qualunque tipo di carattere, sia qualitativo che quantitativo. Se il carattere è quantitativo continuo con dati raggruppati in classi non è possibile determinare con esattezza la moda, ci si limita ad individuare la *classe modale* definita come la classe cui è associata la massima densità di frequenza.

---

**Esempio A.9.** Nella tabella seguente sono riportati i numeri degli studenti, divisi per classe, della sezione A di un dato istituto, in un dato anno. Si può osservare che la 1ª classe presenta la frequenza massima di 320 studenti, quindi la moda è la classe prima.

| Classe | 1° | 2° | 3° | 4° | 5° | Totale |
|---|---|---|---|---|---|---|
| Studenti | 320 | 230 | 212 | 152 | 96 | 1 010 |

**Esempio A.10.** La tabella raccoglie i dati relativi alla domanda "quante ore alla settimana pratichi sport?", posta ad un campione di 50 ragazzi dai 18 ai 25 anni. Si può osservare che 12 e 18 ore presentano la frequenza massima 14, quindi si hanno due mode 12 ore e 18 ore. In questo caso la distribuzione è bimodale.

| Numero di ore | 0 | 4 | 8 | 12 | 16 | 18 | 22 | Totale |
|---|---|---|---|---|---|---|---|---|
| Numero di ragazzi | 4 | 1 | 3 | 14 | 8 | 14 | 6 | 50 |

**Esempio A.11.** La tabella seguente è relativa alla distribuzione delle classi di altezza di un gruppo di studenti.

| Altezza (cm) | 160-165 | 165-170 | 170-175 | 175-185 | 185-200 | Totale |
|---|---|---|---|---|---|---|
| Numero di studenti | 5 | 8 | 15 | 10 | 2 | 40 |

Poiché le classi hanno ampiezza diversa è necessario calcolare la densità di frequenza.

| Altezza | 160-165 | 165-170 | 170-175 | 175-185 | 185-200 |
|---|---|---|---|---|---|
| Densità di frequenza | 1 | 1,6 | 3 | 1 | 0,13 |

La massima densità di frequenza si ha in corrispondenza della classe 170-175, essa rappresenta quindi la classe modale.

### A.3.2 Media aritmetica

**Definizione A.5.** La *media aritmetica* (semplice) è il valore ottenuto sommando tutti i dati e dividendo tale somma per il numero dei dati.

Se abbiamo $n$ dati $x_1, x_2, \ldots, x_n$, la media aritmetica semplice $M$ è data da:

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

**Esempio A.12.** Riprendiamo in esame la tabella relativa agli studenti, divisi per classe frequentata di un dato istituto scolastico, in un dato anno e calcoliamone la media aritmetica semplice.

| Classe | 1ª | 2ª | 3ª | 4ª | 5ª | Totale |
|---|---|---|---|---|---|---|
| Studenti | 320 | 230 | 212 | 152 | 96 | 1 010 |

Per calcolare la media aritmetica semplice degli studenti, sommiamo tutti gli studenti delle cinque classi e dividiamo tale somma per il numero delle classi:

$$M = \frac{320 + 230 + 212 + 152 + 96}{5} = \frac{1\,010}{5} = 202.$$

Possiamo dire che si hanno mediamente 202 studenti per ogni classe.

**Definizione A.6.** Si definisce *scarto dalla media* (aritmetica) la differenza tra i valori osservati e la media.

Se $x_1, x_2, \ldots, x_n$ sono i valori osservati e $M$ la loro media aritmetica, gli scarti sono $s_1 = x_1 - M$, $s_2 = x_2 - M, \ldots, s_n = x_n - M$.

**Esempio A.13.** Calcoliamo gli scarti dalla media per la distribuzione "studenti per tipologia di classe frequentata", la cui media è 1 010/5 = 202.

| Classe | 1ª | 2ª | 3ª | 4ª | 5ª | Totale |
|---|---|---|---|---|---|---|
| Studenti | 320 | 230 | 212 | 152 | 96 | 1 010 |
| Scarto | 118 | 28 | 10 | −50 | −106 | 0 |

Si può osservare che vi solo valori superiori alla media e altri inferiori, tanto che lo scarto è rappresentato in alcuni casi da un numero positivo, in altri da un numero negativo. Si può verificare che la somma degli scarti dalla media è nulla, cioè gli scarti positivi compensano sempre quelli negativi.

**Definizione A.7.** La *media aritmetica ponderata* è il valore ottenuto moltiplicando ciascuna modalità del carattere dato con la propria frequenza, sommando tutti i prodotti fra loro e dividendo poi per la somma delle frequenze (che equivale al numero totale $n$ delle unità statistiche considerate).

La media aritmetica ponderata si usa nel caso in cui le unità statistiche sono molte ed è già stata fatta la tabella delle frequenze. Avendo quindi le modalità del carattere $m_1, m_2, \ldots, m_k$ e le relative frequenze $f_1, f_2, \ldots, f_k$, la media aritmetica ponderata $M$ è data da:

$$M = \frac{m_1 \cdot f_1 + m_2 \cdot f_2 + \cdots + m_k \cdot f_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k} m_i \cdot f_i}{\sum_{i=1}^{k} f_i} = \frac{1}{n} \sum_{i=1}^{k} m_i \cdot f_i.$$

**Esempio A.14.** Riprendiamo la tabella dell'esempio precedente relativa alla domanda "quante ore al giorno passi al computer?", posta ad un campione di 52 ragazzi dai 16 ai 24 anni. Calcoliamo la media aritmetica ponderata.

| Numero di ore | 0 | 1 | 2 | 3 | 4 | 5 | 6 | Totale |
|---|---|---|---|---|---|---|---|---|
| Numero di ragazzi | 4 | 6 | 12 | 16 | 8 | 4 | 2 | 52 |

Considerando le 7 modalità del carattere "Numero di ore" riportate nella tabella, si ha:

$$M = \frac{0 \cdot 4 + 1 \cdot 6 + 2 \cdot 12 + 3 \cdot 16 + 4 \cdot 8 + 5 \cdot 4 + 6 \cdot 2}{4 + 6 + 12 + 16 + 8 + 4 + 2} = \frac{142}{52} = 2{,}73.$$

Possiamo dire che, in media, ciascun ragazzo passa circa 3 ore al giorno al computer.

Il valore della media aritmetica semplice effettuata sulle singole unità statistiche coincide con quella ponderata effettuata sul raggruppamento dei dati per modalità del carattere considerato (tabella delle frequenze).

### A.3.3 Mediana

**Definizione A.8.** La *mediana* di una successione di dati disposti in ordine crescente è il valore equidistante dagli estremi, cioè è

- il dato che occupa la posizione centrale, se il numero dei dati è dispari;
- è la media aritmetica dei dati della coppia centrale, se il numero dei dati è pari.

Poiché per calcolare la mediana i dati devono essere ordinati, è bene sottolineare che tale indice non può essere calcolato se il carattere in esame è di tipo qualitativo non ordinabile.

---

**Esempio A.15.** Supponiamo di avere 7 dati disposti in ordine crescente: 5, 8, 10, 14, 18, 20, 25. Allora la mediana è il valore centrale, quello che occupa la quarta posizione, cioè il 14.

**Esempio A.16.** Supponiamo di avere 8 dati disposti in ordine crescente: 1, 5, 8, 10, 14, 18, 20, 25. La mediana è la media aritmetica dei dati che occupano la 4ª e la 5ª posizione, cioè $\frac{10+14}{2} = 12$.

**Esempio A.17.** Supponiamo di avere la distribuzione di frequenza riportata nella tabella. Il numero di osservazioni è pari, quindi la mediana è il valore della variabile che corrisponde alla media dei due valori centrali, rispettivamente quelli che nella serie ordinata occupano il 13° e il 14° posto.

È necessario in questo caso determinare le *frequenze cumulate*. Esse si ottengono sommando le frequenze che hanno un valore della variabile minore o uguale alla modalità considerata. La frequenza cumulata relativa al voto 3 rimane 2, quella relativa al voto 4 si ottiene sommando la frequenza del 3 e la frequenza del 4, cioè $2 + 2 = 4$, la frequenza cumulata relativa al voto 5 si ottiene dalla somma della frequenza del 3, del 4 e del 5 e così via. Il 14° posto corrisponde al voto 6, mentre il 15° posto è il voto 7. La mediana è quindi 6,5.

| Voto | Frequenza | Frequenza cumulata |
|---|---|---|
| 3 | 2 | 2 |
| 4 | 4 | 4+2=6 |
| 5 | 3 | 3+4+2=9 |
| 6 | 5 | 5+3+4+2=14 |
| 7 | 7 | 7+5+3+4+2=21 |
| 8 | 2 | 2+7+5+3+4+2=23 |
| 9 | 2 | 2+2+7+5+3+4+2=25 |
| 10 | 1 | 1+2+2+7+5+3+4+2=26 |
| Totale | 26 | |

---

✍ *Esercizi proposti:* A.21, A.22, A.23, A.24, A.25, A.26, A.27, A.28, A.29, A.30, A.31, A.32

## A.4 Indici di variabilità

Gli *indici di variabilità* vengono calcolati per analizzare in che modo i termini di una distribuzione si concentrano intorno ad un valore medio.

**Definizione A.9.** Il *campo di variazione* è la differenza fra il valore massimo ed il valore minimo assunti dalla variabile: CVar=$x_{max} - x_{min}$.

Tale indice dà un'informazione molto grossolana perché tiene conto solo del primo e dell'ultimo termine della distribuzione e non tiene conto di tutti i valori intermedi. Si considerino, ad esempio, le seguenti distribuzioni di stature non ordinate:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Gruppo A (statura in cm) | 150 | 155 | 155 | 160 | 165 | 180 | 175 |
| Gruppo B (statura in cm) | 150 | 160 | 175 | 170 | 170 | 170 | 180 |

Entrambe le distribuzioni hanno lo stesso valore massimo e lo stesso valore minimo e quindi lo stesso campo di variazione, ma mentre nella prima i valori sono concentrati verso il valore minimo nella seconda si concentrano intorno al valore massimo.

L'indice non dà quindi alcuna indicazione su quest'ultima informazione. Né può essere utilizzato come indice di variabilità la media degli scarti fra le singole osservazioni e la loro media aritmetica perché tale valore è sempre uguale a zero.

### A.4.1 Scarto medio assoluto

**Definizione A.10.** Si definisce *scarto medio assoluto* la media aritmetica dei valori assoluti degli scarti; esso indica quanto i valori rilevati si disperdono intorno al valore medio della distribuzione:

$$s = \frac{|s_1| + |s_2| + \cdots + |s_n|}{n} = \frac{1}{n}\sum_{i=1}^{n} |x_i - M|.$$

Facendo riferimento alla distribuzione

| Classe | 1ª | 2ª | 3ª | 4ª | 5ª | Totale |
|---|---|---|---|---|---|---|
| Studenti | 320 | 230 | 212 | 152 | 96 | 1 010 |

si ha che lo scarto medio assoluto è 62,4. Si può allora affermare che in ogni tipologia di classe si hanno in media $202 \pm 62{,}4$ iscritti.

### A.4.2 Varianza e scarto quadratico medio

L'indice di variabilità più utilizzato è la varianza o lo scarto quadratico medio.

**Definizione A.11.** La *varianza* è la media dei quadrati degli scarti fra le singole osservazioni e la loro media aritmetica:

$$\mathrm{Var} = \frac{\left[(x_1 - M)^2 + (x_2 - M)^2 + \cdots + (x_n - M)^2\right]}{n} = \frac{1}{n}\sum_{i=1}^{n} (x_i - M)^2.$$

Lo *scarto quadratico medio* è la radice quadrata della varianza: $\sigma = \sqrt{\mathrm{Var}}$.

Se i dati si presentano sotto forma di distribuzione di frequenza, la media deve essere ponderata con le singole frequenze, cioè:

$$\mathrm{Var} = \frac{\left[(m_1 - M)^2 \cdot f_1 + (m_2 - M)^2 \cdot f_2 + \cdots + (m_k - M)^2 \cdot f_k\right]}{f_1 + f_2 + \ldots + f_k} = \frac{\sum_{i=1}^{k}(m_i - M)^2 \cdot f_i}{\sum_{i=1}^{k} f_i} =$$
$$= \frac{1}{n}\sum_{i=1}^{k}(m_i - M)^2 \cdot f_i.$$

La varianza assume valore zero quando tutti i valori coincidono con la media ed è tanto più grande quanto più i singoli valori si discostano dalla media. Poiché tale indice è influenzato sia dal valore della media che dall'unità di misura utilizzato, spesso si utilizza un indice detto coefficiente di variazione.

### A.4.3 Coefficiente di variazione

**Definizione A.12.** Il *coefficiente di variazione* è il rapporto fra lo scarto quadratico medio (radice quadrata della varianza) e la media aritmetica:

$$\mathrm{CV} = \frac{\sigma}{M} = \frac{\sqrt{\mathrm{Var}}}{M}.$$

Tale indice risulta di particolare utilità per confrontare distribuzioni diverse.

---

**Esempio A.18.** È dato l'elenco delle stature, in cm, dei ragazzi di una classe: 165, 182, 159, 173, 160, 175, 185, 190, 175, 180, 159, 185, 176, 170, 175, 160, 175, 182, 159, 185.

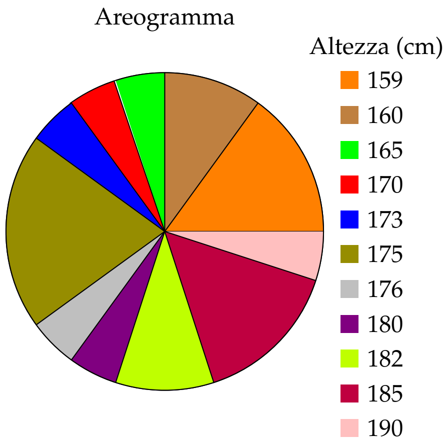
a ) Ordina i dati in una tabella delle frequenze;
b ) rappresenta i dati graficamente;
c ) calcola la media, la mediana e la moda;
d ) calcola la varianza e il coefficiente di variazione.

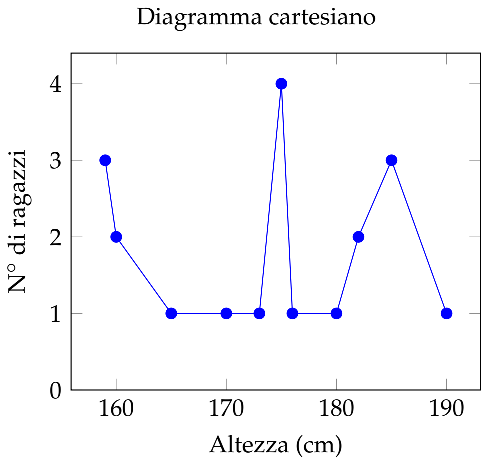
#### Tabella delle frequenze

| Dati | Frequenze assolute | Frequenze relative | Frequenze percentuali |
|---|---|---|---|
| 159 | 3 | 0,15 | 15% |
| 160 | 2 | 0,10 | 10% |
| 165 | 1 | 0,05 | 5% |
| 170 | 1 | 0,05 | 5% |
| 173 | 1 | 0,05 | 5% |
| 175 | 4 | 0,20 | 20% |
| 176 | 1 | 0,05 | 5% |
| 180 | 1 | 0,05 | 5% |
| 182 | 2 | 0,10 | 10% |
| 185 | 3 | 0,15 | 15% |
| 190 | 1 | 0,05 | 5% |
| Totale | 20 | 1 | 100% |

- La somma delle frequenze assolute è pari al numero totale degli studenti;
- la somma delle frequenze relative è 1 (a meno di approssimazioni nel calcolo delle frequenze relative delle singole modalità del carattere);
- la somma delle frequenze percentuali è 100 (a meno di approssimazioni nel calcolo delle frequenze percentuali delle singole modalità del carattere).

### Grafici




### Calcolo della media, mediana e moda

Calcoliamo la media aritmetica:

$$\text{Media} = \frac{1}{20} \cdot (165 + 182 + 159 + 173 + 160 + 175 + 185 + 190 + 175 + 180 + 159 + 185 + \\ + 176 + 170 + 175 + 160 + 175 + 182 + 159 + 185) = 173{,}5.$$

Per determinare la mediana si devono ordinare in modo crescente i dati: 159, 159, 159, 160, 160, 165, 170, 173, 175, 175, 175, 175, 176, 180, 182, 182, 185, 185, 185, 190. Essendo i dati in numero pari si calcola la media dei due dati centrali: Mediana $= (175 + 175)/2 = 175$. Se i dati sono molti è possibile individuare qual è o quali sono i dati centrali utilizzando la tabella delle frequenze opportunamente costruita, cioè con i dati scritti in ordine crescente.

La moda è la modalità del carattere altezza che è più ricorrente, cioè quello con la frequenza più alta: Moda $= 175$.

---

✍ *Esercizi proposti:* A.33, A.34, A.35, A.36, A.37, A.38