CAPITOLO

16

Statistica descrittiva

Mille numeri in fila non dicono niente; una tabella, un grafico e due indici ben scelti raccontano una storia. Il mestiere della statistica è non farle dire bugie.

OBIETTIVI

1. Riconoscere popolazione, campione, unità statistiche, caratteri e modalità di un'indagine.
2. Costruire tabelle di frequenze assolute, relative, percentuali e cumulate, anche per classi di valori.
3. Disegnare e leggere ortogrammi, areogrammi, istogrammi e spezzate, e smascherare i grafici ingannevoli.
4. Calcolare e interpretare media aritmetica e ponderata, mediana e moda, scegliendo l'indice adatto.
5. Misurare la variabilità dei dati con campo di variazione, scarto semplice medio, varianza e deviazione standard.
6. Usare il foglio di calcolo come strumento per elaborare i dati.

INDICE DEL CAPITOLO

16.1 L'indagine statistica 2
16.2 Dati e frequenze 4
16.3 Rappresentazioni grafiche 6
16.4 Indici di posizione centrale 12
16.5 Indici di variabilità 16

Matematica 1 · Parte VI · Dati e statistica · versione 1.0
Capitolo 16: Statistica descrittiva · pagine 415–440 del volume completo · estratto per lo studio

Il registro elettronico ti dice la media dei voti; l'app del bus promette che la corsa dura «in media» 20 minuti; un sondaggio annuncia che «il 37% dei ragazzi» preferisce una certa cosa. Da dove vengono questi numeri? E ci si può fidare?

La *statistica* è la disciplina che raccoglie, organizza, rappresenta e interpreta i dati. In questo capitolo ci occupiamo della sua parte *descrittiva*, che riassume i dati osservati con tabelle, grafici e indici, senza pretendere di prevedere ciò che non si è osservato; il calcolo delle probabilità, che rende possibili le previsioni, lo incontrerai il prossimo anno. Useremo frazioni e percentuali (capitoli 3 e 4), il piano cartesiano (capitolo 6) e il calcolo letterale (capitolo 7). Un'avvertenza importante: **tutti i dati di questo capitolo sono inventati**, a scopo didattico, e riguardano classi, scuole e persone immaginarie.

16.1 L'indagine statistica

16.1.1 Popolazione, unità statistiche, campione

Supponiamo che il consiglio di istituto di un liceo immaginario voglia sapere come gli studenti arrivano a scuola, per chiedere al Comune più corse di autobus. L'insieme di tutti gli studenti del liceo è la **popolazione** dell'indagine: l'insieme di tutti gli elementi su cui si vuole indagare. Ciascuno studente è una **unità statistica**, cioè un singolo elemento della popolazione sul quale si rileva l'informazione. La popolazione non è necessariamente fatta di persone: possono essere le famiglie di un comune, le lampadine prodotte in un giorno da una fabbrica, i giorni di un anno in cui si misura la temperatura.

Se si intervistano *tutte* le unità della popolazione si fa una *rilevazione totale*, o censimento. Spesso però è impossibile o troppo costoso: per controllare la durata delle lampadine bisognerebbe accenderle tutte finché non si bruciano! Allora si sceglie un **campione**, cioè un sottoinsieme della popolazione, e si fa una *rilevazione campionaria*. Perché le conclusioni valgano anche per la popolazione, il campione deve essere *rappresentativo*: se il consiglio di istituto intervistasse soltanto gli studenti che aspettano alla fermata dell'autobus, otterrebbe un risultato distorto. In questo capitolo non ci occupiamo di come si costruisce un buon campione (serve il calcolo delle probabilità): descriviamo i dati che abbiamo.

16.1.2 Caratteri e modalità

L'aspetto che si osserva su ogni unità statistica si chiama **carattere**; i diversi valori che il carattere può assumere si chiamano **modalità**. Nell'indagine sul tragitto casa-scuola il carattere è «mezzo di trasporto usato» e le modalità sono «a piedi», «bicicletta», «autobus», «auto», «treno». Se invece si rileva il «tempo impiegato», le modalità sono numeri: 12 minuti, 25 minuti…

I caratteri si classificano così.

- Un **carattere qualitativo** ha modalità espresse da parole. È *sconnesso* (o nominale) se le modalità non hanno un ordine naturale (mezzo di trasporto, colore degli occhi, sport preferito); è *ordinabile* se le modalità hanno un ordine (titolo di studio; giudizio «insufficiente, sufficiente, buono, ottimo»; taglia S, M, L, XL).
- Un **carattere quantitativo** ha modalità numeriche, sulle quali ha senso fare calcoli. È *discreto* se assume valori isolati, di solito ottenuti contando (numero di fratelli, voti interi in decimi, libri letti in un anno); è *continuo* se può assumere tutti i valori di un intervallo, e di solito si ottiene misurando (altezza, massa, tempo, temperatura).

Tabella 16.1 Esempi di caratteri e del loro tipo.

Unità statistica	Carattere	Tipo	Esempi di modalità
studente	mezzo di trasporto	qualitativo sconnesso	a piedi, autobus, treno
studente	giudizio in condotta	qualitativo ordinabile	sufficiente, buono, ottimo
famiglia	numero di figli	quantitativo discreto	0, 1, 2, 3
studente	tempo casa-scuola	quantitativo continuo	7,5 min; 23 min
giorno	temperatura massima	quantitativo continuo	18,4 °C; 21,0 °C

Attenzione: un carattere non è quantitativo solo perché le modalità sono scritte con cifre. Il CAP o il numero di maglia di un calciatore sono etichette, e farne la media non ha senso.

16.1.3 Le fasi di un'indagine

Un'indagine statistica seria procede per fasi.

1. **Obiettivo.** Si stabilisce che cosa si vuole sapere e perché.
2. **Progettazione.** Si definiscono popolazione, unità, caratteri e modalità; si sceglie tra censimento e campione.
3. **Rilevazione.** Si raccolgono i dati: questionari, interviste, misure, archivi.
4. **Spoglio e organizzazione.** Si contano le risposte e si ordinano in tabelle di frequenze.
5. **Rappresentazione.** Si costruiscono grafici adatti al tipo di carattere.
6. **Elaborazione.** Si calcolano gli indici (medie, mediana, moda, indici di variabilità).
7. **Interpretazione.** Si traggono le conclusioni, dichiarando i limiti dell'indagine.

Quando non si raccolgono i dati di persona si usano *fonti* affidabili: in Italia la statistica ufficiale è prodotta soprattutto dall'**ISTAT** (Istituto nazionale di statistica), in Europa da **Eurostat**. Un dato senza fonte, senza data e senza la definizione di ciò che si è misurato va preso con molta cautela.

16.2 Dati e frequenze

16.2.1 Frequenze assolute, relative e percentuali

Torniamo all'indagine sul tragitto. In una classe prima immaginaria di 24 studenti si raccolgono le risposte alla domanda «Con quale mezzo vieni a scuola di solito?». Contando quante volte compare ciascuna modalità si ottiene la sua **frequenza assoluta** f. La somma delle frequenze assolute è il numero n delle unità osservate.

La frequenza assoluta da sola non permette confronti: 9 studenti su 24 sono tanti, 9 su 900 sono pochi. Per questo si usa la **frequenza relativa**, il rapporto tra la frequenza assoluta e il numero totale dei dati, e la **frequenza percentuale**, che è la frequenza relativa moltiplicata per 100:

$$f_r = \frac{f}{n} \qquad \text{frequenza percentuale} = f_r \cdot 100$$

Tabella 16.2 Mezzo di trasporto usato dagli studenti di una classe immaginaria ($n = 24$). Le percentuali sono arrotondate ai decimi.

Modalità	f	f_r	%
a piedi	5	5/24 ≈ 0,208	20,8
bicicletta	3	3/24 = 0,125	12,5
autobus	9	9/24 = 0,375	37,5
auto	4	4/24 ≈ 0,167	16,7
treno	3	3/24 = 0,125	12,5
totale	**24**	**1**	**100**

Le frequenze relative sono frazioni con lo stesso denominatore n; la loro somma è $\frac{n}{n} = 1$, e la somma delle percentuali è 100. Quando le percentuali sono arrotondate la somma può risultare 99,9 o 100,1: non è un errore di calcolo, ma un effetto degli arrotondamenti, che conviene segnalare in una nota alla tabella.

16.2.2 Frequenze cumulate

Se le modalità si possono ordinare (carattere qualitativo ordinabile o quantitativo) ha senso chiedersi: «quanti dati sono *minori o uguali* a una certa modalità?». La risposta è la **frequenza cumulata**: la somma delle frequenze di quella modalità e di tutte quelle che la precedono. L'ultima frequenza cumulata è sempre n (o il 100%).

Tabella 16.3 Voti (in decimi) dell'ultima verifica di matematica in una classe immaginaria di 24 studenti.

Voto	f	%	f cumulata	% cumulata
4	1	4,2	1	4,2
5	3	12,5	4	16,7
6	8	33,3	12	50,0
7	6	25,0	18	75,0
8	4	16,7	22	91,7
9	2	8,3	24	100,0

Leggendo la colonna delle cumulate si risponde subito a domande come: «quanti studenti hanno preso meno di 6?» (4, cioè la cumulata del 5) oppure «quanti hanno preso almeno 7?» (24 − 12 = 12, cioè il 50%). Per il mezzo di trasporto, invece, le cumulate non hanno senso: «minore o uguale a bicicletta» non vuol dire niente.

16.2.3 Classi di valori

Quando un carattere è continuo, o discreto con moltissime modalità diverse, una tabella con una riga per ogni valore sarebbe lunghissima e quasi tutta fatta di frequenze 1. Si raggruppano allora i dati in **classi**: intervalli consecutivi che non si sovrappongono e coprono tutti i dati. Per evitare ambiguità ai bordi, in questo libro una classe include l'estremo inferiore ed esclude quello superiore: la classe $[10; 20[$ contiene i tempi t con $10 \le t < 20$, e un tempo di esattamente 20 minuti va nella classe successiva.

Di ogni classe $[a; b[$ si considerano gli *estremi* a e b, l'**ampiezza** $b - a$ e il **valore centrale** $\dfrac{a+b}{2}$, che rappresenta la classe nei calcoli. Le classi possono avere tutte la stessa ampiezza oppure ampiezze diverse, quando i dati sono molto addensati in una zona e radi in un'altra. Di solito si usano da 5 a 10 classi: con troppo poche si perde l'informazione, con troppe si torna alla lista dei dati. Raggruppando si perde qualcosa: dalla tabella per classi non si ricava più il valore esatto dei singoli dati.

ESEMPIO SVOLTO

Problema. Si rileva il tempo impiegato per arrivare a scuola da 40 studenti di un liceo immaginario e si ottiene la tabella per classi qui sotto. Completala con ampiezza, valore centrale e frequenza percentuale di ogni classe.

Tabella 16.4 Tempo casa-scuola di 40 studenti (dati inventati).

Classe (minuti)	f	Ampiezza	Valore centrale	%
[0; 10[	6	10	5	15
[10; 20[	12	10	15	30
[20; 30[	10	10	25	25
[30; 45[	9	15	37,5	22,5
[45; 75[	3	30	60	7,5
totale	**40**			**100**

Soluzione. L'ampiezza è la differenza tra gli estremi: per esempio $45 - 30 = 15$. Il valore centrale è la media degli estremi: $\frac{30 + 45}{2} = 37{,}5$. La percentuale è $\frac{f}{40} \cdot 100$: per la classe [30; 45[si ha $\frac{9}{40} = 0{,}225$, cioè 22,5%. Le classi hanno ampiezze diverse (10, 15, 30 minuti): i tempi lunghi sono pochi e sarebbe inutile spezzarli in tante classi quasi vuote. Lo stesso dato ci servirà per l'istogramma (paragrafo 16.3.3).

16.3 Rappresentazioni grafiche

Un buon grafico mostra a colpo d'occhio ciò che in una tabella bisogna cercare; ma ogni tipo di grafico è adatto solo a certi dati.

16.3.1 L'ortogramma

L'**ortogramma**, o diagramma a barre, rappresenta ogni modalità con un rettangolo (una barra) di base fissa e altezza proporzionale alla frequenza. Le barre sono separate da spazi uguali, per ricordare che le modalità sono distinte. È adatto a caratteri qualitativi e quantitativi discreti. Sull'asse verticale si possono riportare le frequenze assolute, relative o percentuali: la forma del grafico non cambia, cambia solo la scala. Le barre possono anche essere orizzontali, comodo quando i nomi delle modalità sono lunghi.

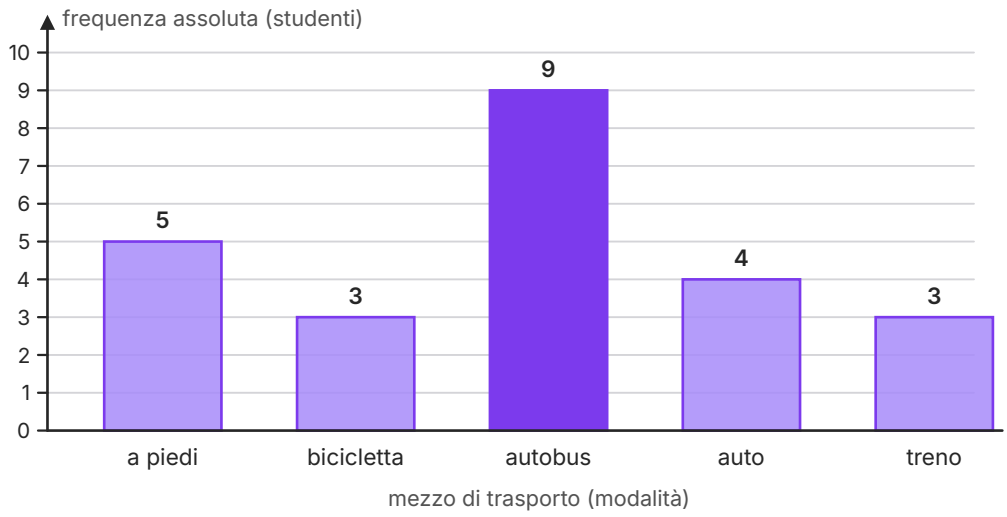


Figura 16.1 Ortogramma dei dati della tabella 16.2. L'altezza di ogni barra è la frequenza assoluta della modalità; la barra più alta, quella dell'autobus, indica la moda (paragrafo 16.4.4). L'asse verticale parte da 0.

Nella figura 16.1 si vede subito che l'autobus è il mezzo più usato e che bicicletta e treno sono alla pari. Se le modalità non hanno un ordine naturale, le barre si possono disporre anche in ordine di frequenza decrescente.

Una variante è il **pittogramma**, in cui le frequenze sono rappresentate da file di figurine (un omino, un libro, un'automobile), ognuna delle quali vale una quantità fissata, dichiarata in una legenda: per esempio «ogni omino = 2 studenti». Tutte le figurine devono avere la stessa dimensione: una frequenza maggiore si rappresenta con *più* figurine, non con una figurina *più grande* (paragrafo 16.3.5).

16.3.2 L'areogramma

L'**areogramma**, o diagramma a torta, rappresenta il totale con un cerchio e ogni modalità con un settore circolare di area proporzionale alla frequenza. Poiché l'area di un settore è proporzionale al suo angolo al centro, basta che l'angolo sia proporzionale alla frequenza. Il cerchio intero (360°) corrisponde a tutti i dati, cioè alla frequenza relativa 1; dunque l'angolo α del settore si ottiene con la proporzione $\alpha : 360° = f : n$, cioè

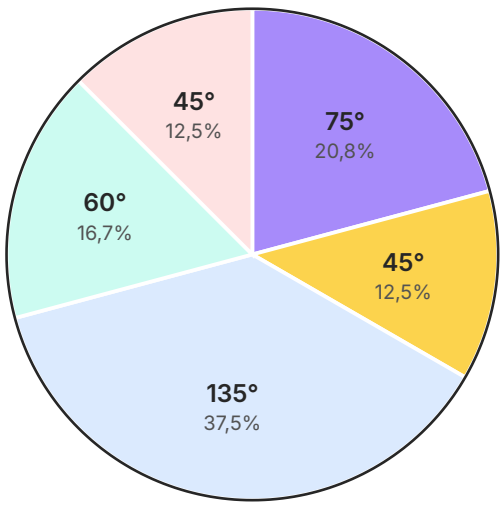
$$\alpha = f_r \cdot 360°$$

L'areogramma è adatto quando interessa il rapporto tra ciascuna parte e il tutto, come nella composizione di un bilancio o nelle preferenze di voto; funziona bene con poche modalità (al massimo 6 o 7), mentre con molte fette sottili diventa illeggibile.

ESEMPIO SVOLTO

Problema. Costruisci l'areogramma dei dati della tabella 16.2.

Soluzione. Calcoliamo gli angoli: a piedi $\frac{5}{24} \cdot 360° = 75°$; bicicletta $\frac{3}{24} \cdot 360° = 45°$; autobus $\frac{9}{24} \cdot 360° = 135°$; auto $\frac{4}{24} \cdot 360° = 60°$; treno $\frac{3}{24} \cdot 360° = 45°$. Controllo: $75° + 45° + 135° + 60° + 45° = 360°$. Con il goniometro si tracciano i settori uno dopo l'altro, partendo per esempio dal raggio verticale e procedendo in senso orario (figura 16.2). Un trucco per il calcolo mentale: qui ogni studente vale $360° : 24 = 15°$.



modalità	f	angolo
a piedi	5	75°
bicicletta	3	45°
autobus	9	135°
auto	4	60°
treno	3	45°
totale	**24**	**360°**

angolo = $f_r \cdot 360°$

Figura 16.2 Areogramma dei dati della tabella 16.2. Ogni settore ha un angolo al centro proporzionale alla frequenza; in ogni settore sono indicati l'angolo e la percentuale. I settori sono disegnati in senso orario a partire dal raggio verticale.

ERRORE FREQUENTE

Arrotondando gli angoli all'unità la somma può non tornare 360°. Con le frequenze 3, 4, 6, 9 su $n = 22$ si ottengono 49,09°, 65,45°, 98,18°, 147,27°; arrotondati: 49°, 65°, 98°, 147°, che danno 359°. Il grado mancante va assegnato al settore con la parte decimale più vicina a 1 tra quelli arrotondati per difetto (65,45° diventa 66°): ora la somma è 360°. Non si «aggiusta» a caso la fetta più grande. Controlla sempre la somma prima di disegnare.

16.3.3 L'istogramma

Per un carattere quantitativo raggruppato in classi si usa l'**istogramma**: sull'asse orizzontale si riportano le classi, in scala, una accanto all'altra senza spazi (perché le classi sono intervalli contigui); su ciascuna classe si costruisce un rettangolo la cui *area* è proporzionale alla frequenza della classe.

Se tutte le classi hanno la stessa ampiezza, l'area è proporzionale all'altezza e si può usare come altezza la frequenza. Se le ampiezze sono diverse, invece, l'altezza deve essere la **densità di frequenza**, cioè la frequenza divisa per l'ampiezza della classe:

$$d = \frac{f}{b - a} \qquad \text{così l'area del rettangolo è } (b - a) \cdot d = f$$

La densità dice quanti dati cadono, in media, in ogni unità di ampiezza: nell'esempio del tempo casa-scuola si misura in «studenti al minuto».

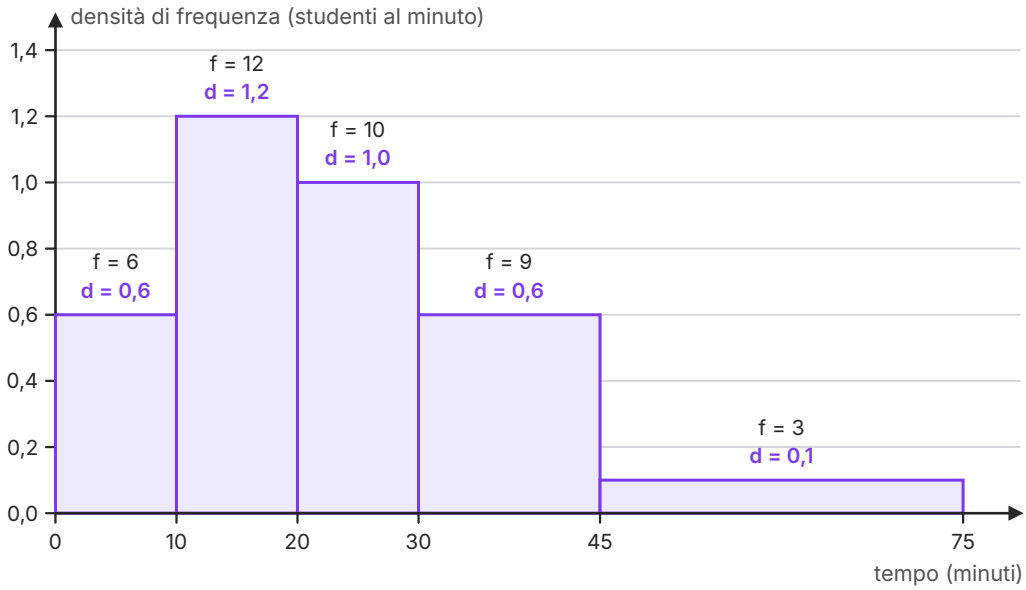


Figura 16.3 Istogramma del tempo casa-scuola (tabella 16.4). Le classi hanno ampiezze diverse, perciò l'altezza di ogni rettangolo è la densità di frequenza $d = f$: ampiezza, e l'area è la frequenza f.

ESEMPIO SVOLTO

Problema. Disegna l'istogramma dei dati della tabella 16.4.

Soluzione. Le classi hanno ampiezze diverse, quindi calcoliamo le densità: $\frac{6}{10} = 0{,}6$; $\frac{12}{10} = 1{,}2$; $\frac{10}{10} = 1{,}0$; $\frac{9}{15} = 0{,}6$; $\frac{3}{30} = 0{,}1$ studenti al minuto. Sull'asse orizzontale segniamo 0, 10, 20, 30, 45, 75 in scala; sopra ciascuna classe costruiamo un rettangolo alto quanto la densità (figura 16.3). Verifica delle aree: per esempio $15 \cdot 0{,}6 = 9$ e $30 \cdot 0{,}1 = 3$, proprio le frequenze.

Osserva la classe [30; 45[: ha 9 studenti, quasi quanti la classe [20; 30[(10 studenti), ma il suo rettangolo è molto più basso, perché quei 9 studenti sono «sparsi» su 15 minuti invece che su 10. Se avessimo usato le frequenze come altezze, il rettangolo largo della classe [45; 75[, con area $30 \cdot 3 = 90$, sarebbe sembrato una volta e mezza quello della classe [0; 10[, con area $10 \cdot 6 = 60$, pur contenendo la metà degli studenti; e quasi grande quanto quello della classe [10; 20[(area $10 \cdot 12 = 120$), che ha quattro volte i suoi studenti.

16.3.4 Il diagramma cartesiano e la spezzata

Quando un carattere viene rilevato in istanti successivi (giorni, mesi, anni) si ottiene una *serie temporale*. Si rappresenta con un *diagramma cartesiano*: sull'asse orizzontale il tempo, su quello verticale il valore del carattere, come nei grafici del capitolo 6. Ogni rilevazione è un punto; unendo i punti consecutivi con segmenti si ottiene una *spezzata*, che mette in evidenza l'andamento: crescite, cali, stagionalità.

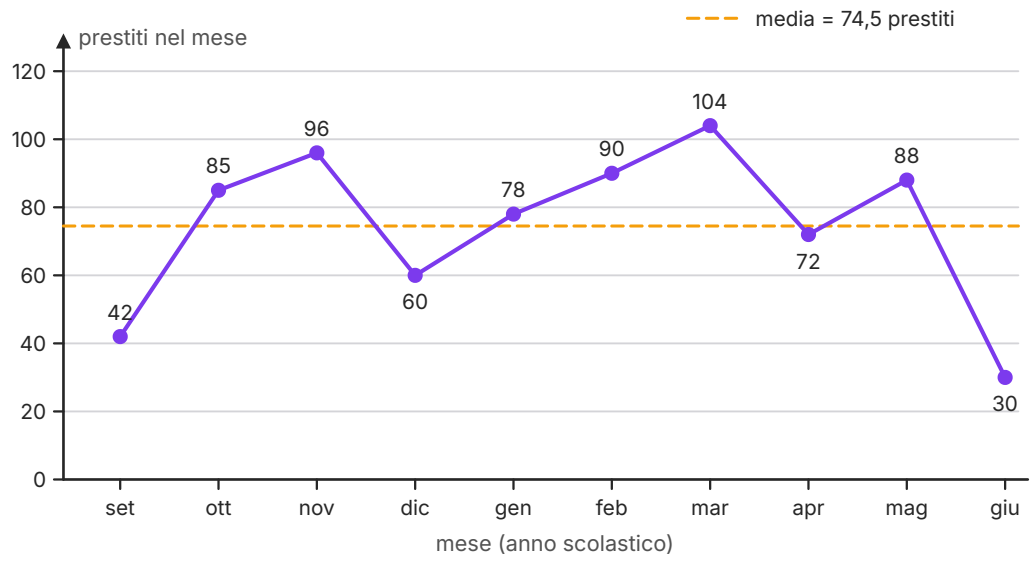


Figura 16.4 Prestiti mensili della biblioteca di una scuola immaginaria durante un anno scolastico (dati inventati). La linea tratteggiata indica la media mensile, 74,5 prestiti: il totale è 745.

Nella figura 16.4 si nota il calo di dicembre e quello di giugno, mesi con molti giorni di vacanza, e il massimo di marzo (104 prestiti). Attenzione: i segmenti guidano solo l'occhio; un valore «intermedio» tra due mesi non è stato misurato.

16.3.5 Grafici ingannevoli

Un grafico può dire il falso anche se tutti i numeri scritti sopra sono veri. Ecco i trucchi più comuni.

- **Asse verticale che non parte da zero.** In un ortogramma la lunghezza delle barre deve essere proporzionale alle frequenze. Se l'asse parte da un valore diverso da 0, le differenze vengono ingigantite (figura 16.5).
- **Figure che crescono in due dimensioni.** Se per rappresentare un valore doppio si raddoppiano sia la base sia l'altezza di una figurina, l'area diventa quadrupla, e l'occhio percepisce l'area. Con figure solide (cubi, sacchi di monete in prospettiva) l'effetto è ancora peggiore: raddoppiando lo spigolo il volume diventa 8 volte più grande.
- **Torte in prospettiva («3D»).** Le fette in primo piano sembrano più grandi di quelle sullo sfondo, anche se hanno lo stesso angolo.
- **Scale diverse o non dichiarate** in due grafici affiancati, e classi di ampiezza diversa disegnate con le frequenze come altezze (paragrafo 16.3.3).

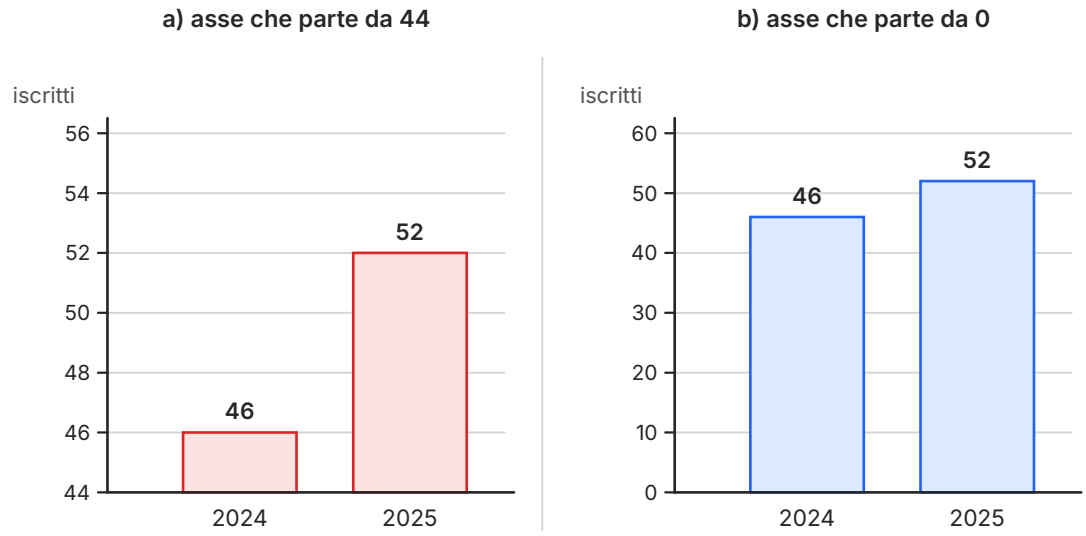


Figura 16.5 Gli iscritti a un corso pomeridiano di una scuola immaginaria passano da 46 a 52. a) Con l'asse che parte da 44 le barre sono alte 2 e 8 unità: la seconda sembra quattro volte la prima. b) Con l'asse che parte da 0 si vede l'aumento reale, del 13% circa.

Nella figura 16.5a la barra del 2025 è quattro volte più alta di quella del 2024, perché $(52 - 44) : (46 - 44) = 8 : 2 = 4$. Ma l'aumento vero è $\frac{52 - 46}{46} = \frac{6}{46} \approx 0{,}13$, cioè circa il 13%. In una spezzata un asse che non parte da zero può essere legittimo, se la scala è ben visibile; in un ortogramma, dove si confrontano le lunghezze delle barre, no.

APPROFONDIMENTO

I grafici statistici che oggi diamo per scontati sono un'invenzione di poco più di due secoli fa. L'ingegnere ed economista scozzese William Playfair pubblicò nel 1786 *The Commercial and Political Atlas*, con serie temporali del commercio inglese disegnate come linee e il primo diagramma a barre; nel 1801, nello *Statistical Breviary*, introdusse il diagramma a torta. Nell'Ottocento l'infermiera britannica Florence Nightingale usò diagrammi a settori circolari (detti «a rosa») per mostrare che, nella guerra di Crimea, molti più soldati morivano di malattie evitabili che di ferite.

16.4 Indici di posizione centrale

Un grafico mostra la forma dei dati; spesso però serve un solo numero che li rappresenti, un «valore tipico». Gli *indici di posizione centrale* sono numeri che riassumono dove si collocano i

dati. Ne studiamo tre: media, mediana e moda. Nelle formule chiameremo x_1, x_2, ..., x_n gli n dati.

16.4.1 La media aritmetica

La **media aritmetica** di n dati numerici è la loro somma divisa per n. Si indica con $\bar{x}$ (si legge «x segnato» o «x medio»):

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

La media è il valore che ogni dato avrebbe se il totale fosse distribuito in parti uguali: se 6 amici spendono in tutto 78 € in pizzeria, la media è 13 €, la quota di ciascuno «alla romana». Si calcola solo per caratteri quantitativi, è sempre compresa tra il dato minimo e il dato massimo, e si esprime nella stessa unità di misura dei dati.

Si chiama *scarto* di un dato dalla media la differenza $x_i - \bar{x}$: è positivo per i dati maggiori della media, negativo per quelli minori. La media ha una proprietà fondamentale.

IN SINTESI

Proprietà della media. La somma degli scarti dei dati dalla loro media aritmetica è zero.

Ipotesi: $\bar{x}$ è la media di $x_1, \ldots, x_n$. **Tesi:** $(x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}) = 0$.

Dimostrazione. Nella somma degli scarti il termine $\bar{x}$ compare n volte con il segno meno; raccogliendo, la somma diventa $(x_1 + x_2 + \ldots + x_n) - n\bar{x}$. Per la definizione di media $x_1 + x_2 + \ldots + x_n = n\bar{x}$, quindi la somma è $n\bar{x} - n\bar{x} = 0$. c.v.d.

In termini fisici: se i dati fossero pesi uguali appoggiati su un'asta graduata senza peso, la media sarebbe il punto in cui mettere il fulcro perché l'asta resti in equilibrio (figura 16.6). Gli scarti negativi «pesano» quanto quelli positivi.

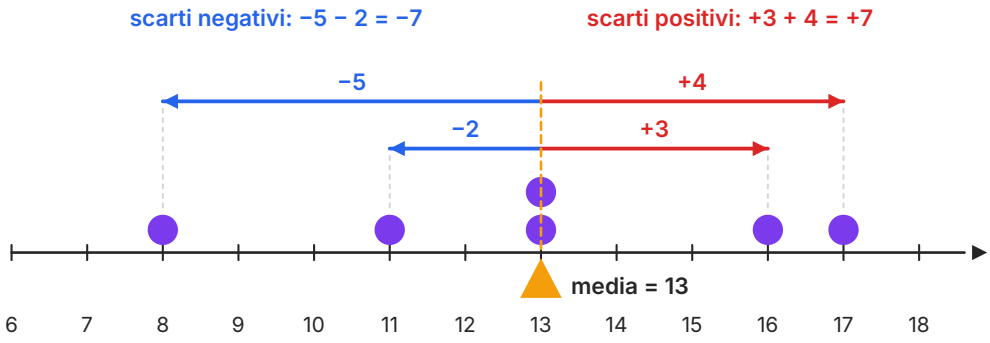


Figura 16.6 I punti segnati da una giocatrice di basket in 6 partite (dati inventati): 8, 11, 13, 13, 16, 17. La media è 78 : 6 = 13. Gli scarti dalla media sono −5, −2, 0, 0, +3, +4 e hanno somma zero: la media è il punto di equilibrio dei dati.

16.4.2 La media ponderata

A volte i dati non hanno tutti la stessa importanza. Si chiama **media ponderata** dei valori x_1, ..., x_k con *pesi* p_1, ..., p_k (numeri positivi) il numero

$$\overline{x} = \frac{x_1\,p_1 + x_2\,p_2 + \ldots + x_k\,p_k}{p_1 + p_2 + \ldots + p_k}$$

Se tutti i pesi sono uguali, la media ponderata coincide con la media aritmetica. I pesi possono essere espressi anche in percentuale: se una scuola immaginaria stabilisce che la verifica scritta conta il 50%, l'interrogazione il 30% e la relazione di laboratorio il 20%, uno studente con voti 7, 6 e 9 ha media $7 \cdot 0{,}5 + 6 \cdot 0{,}3 + 9 \cdot 0{,}2 = 3{,}5 + 1{,}8 + 1{,}8 = 7{,}1$ (i pesi in forma decimale hanno somma 1, quindi non serve dividere).

Il caso più importante è quello di una *tabella di frequenze*: ogni modalità va contata tante volte quanto è la sua frequenza, quindi le frequenze fanno da pesi. Per dati raggruppati in classi si usano come valori i valori centrali; il risultato è solo una *stima* della media vera, perché si suppone che i dati di ogni classe siano tutti uguali al valore centrale.

ESEMPIO SVOLTO

Problema. Calcola la media dei voti della tabella 16.3 e stima la media del tempo casa-scuola della tabella 16.4.

Soluzione. Voti: è una media ponderata con pesi uguali alle frequenze.

$$\bar{x} = \frac{4 \cdot 1 + 5 \cdot 3 + 6 \cdot 8 + 7 \cdot 6 + 8 \cdot 4 + 9 \cdot 2}{24} = \frac{4 + 15 + 48 + 42 + 32 + 18}{24} = \frac{159}{24} = 6{,}625$$

Tempo casa-scuola: usiamo i valori centrali 5; 15; 25; 37,5; 60 con le frequenze 6, 12, 10, 9, 3:

$$\bar{x} \approx \frac{5 \cdot 6 + 15 \cdot 12 + 25 \cdot 10 + 37{,}5 \cdot 9 + 60 \cdot 3}{40} = \frac{977{,}5}{40} \approx 24{,}4 \text{ minuti}$$

16.4.3 La mediana

La **mediana** è il valore che occupa la posizione centrale quando i dati sono disposti in ordine crescente: almeno metà dei dati è minore o uguale a essa e almeno metà è maggiore o uguale. Si trova così:

- si **ordinano** i dati (passaggio indispensabile);
- se n è **dispari**, la mediana è il dato di posto $\frac{n+1}{2}$;
- se n è **pari**, i dati centrali sono due, quelli di posto $\frac{n}{2}$ e $\frac{n}{2}+1$, e la mediana è la loro media.

Esempi. I dati 7, 3, 9, 4, 6, ordinati, sono 3, 4, 6, 7, 9: con $n = 5$ la mediana è il terzo dato, 6. I dati 2, 10, 5, 8, 3, 7, ordinati, sono 2, 3, 5, 7, 8, 10: con $n = 6$ i dati centrali sono il terzo e il quarto, e la mediana è $\frac{5+7}{2} = 6$, che non è uno dei dati.

Con una tabella di frequenze si usano le cumulate. Nella tabella 16.3 i voti sono $n = 24$; i dati centrali sono il 12° e il 13°. La cumulata del 6 è 12: il 12° voto è 6, mentre il 13° è già un 7. La mediana è quindi $\frac{6+7}{2} = 6{,}5$. La mediana si può calcolare anche per un carattere qualitativo ordinabile (per esempio la mediana dei giudizi di una classe può essere «buono»), purché, nel caso pari, i due dati centrali coincidano.

16.4.4 La moda

La **moda** è la modalità con la frequenza più alta. È l'unico dei tre indici che si può usare per qualsiasi carattere, anche qualitativo sconnesso: la moda della tabella 16.2 è «autobus», quella della tabella 16.3 è il voto 6. Una distribuzione può avere due mode (si dice *bimodale*) o più; se

tutte le modalità hanno la stessa frequenza, la moda non è significativa. Per dati raggruppati in classi si parla di *classe modale*: quella con la densità di frequenza più alta, cioè con il rettangolo più alto nell'istogramma. Nella tabella 16.4 è [10; 20[.

DA RICORDARE

Per i voti della tabella 16.3 i tre indici sono diversi: media 6,625, mediana 6,5, moda 6. Non c'è niente di strano: misurano cose diverse. La **media** usa il valore di tutti i dati; la **mediana** usa solo il loro ordine; la **moda** usa solo le frequenze. Per un carattere qualitativo sconnesso esiste solo la moda; per uno ordinabile, moda e mediana; per uno quantitativo, tutti e tre.

16.4.5 Valori anomali: quale indice scegliere?

Un **valore anomalo** è un dato molto lontano da tutti gli altri: può essere un errore di trascrizione (un 80 scritto al posto di 8) oppure un dato vero ma eccezionale. Consideriamo gli stipendi mensili netti dei 7 dipendenti di una piccola ditta immaginaria, titolare compresa:

1500 €, 1600 €, 1600 €, 1700 €, 1800 €, 1900 €, 8000 €

La media è 18100 : 7 ≈ 2586 €; la mediana è il 4° dato, 1700 €; la moda è 1600 €. Sei dipendenti su sette guadagnano meno della media! Il singolo stipendio di 8000 € «trascina» la media verso l'alto, mentre la mediana non se ne accorge: se al posto di 8000 € ci fossero 80 000 €, la mediana resterebbe 1700 €. Si dice che la mediana è un indice *robusto*, cioè poco sensibile ai valori anomali. Togliendo lo stipendio della titolare, la media scende a 10100 : 6 ≈ 1683 € e la mediana diventa 1650 €: ora i due indici sono vicini.

Ecco qualche criterio per scegliere.

- La **media** è l'indice più usato per dati quantitativi senza valori anomali, e l'unico che permette di ricostruire il totale ($n \cdot \overline{x}$).
- La **mediana** è preferibile quando ci sono valori anomali o la distribuzione è molto asimmetrica, come per redditi, prezzi delle case, tempi di attesa.
- La **moda** serve per i caratteri qualitativi e quando interessa la scelta «più frequente» (la taglia da ordinare in maggior quantità, il mezzo più usato).

16.5 Indici di variabilità

16.5.1 Perché la media non basta

Due linee di autobus immaginarie collegano lo stesso quartiere al liceo. Per 10 giorni si registra la durata del viaggio, in minuti:

linea A: 18, 19, 19, 20, 20, 20, 21, 21, 21, 21 linea B: 14, 16, 16, 17, 20, 21, 22, 22, 25, 27

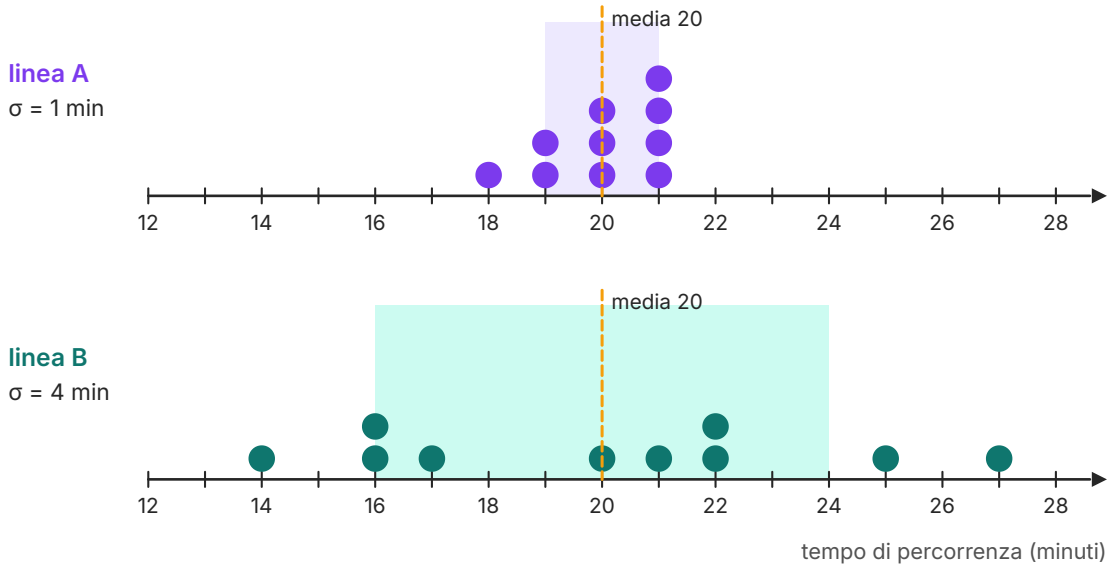


Figura 16.7 Durata del viaggio in 10 giorni con due linee di autobus immaginarie: ogni pallino è un giorno. Le medie sono uguali (20 minuti), ma i dati della linea B sono molto più dispersi. La fascia colorata va da media − σ a media + σ: σ = 1 minuto per la linea A, σ = 4 minuti per la linea B (paragrafo 16.5.4).

Entrambe le serie hanno somma 200 e quindi media 20 minuti. Eppure le due linee sono molto diverse: con la A si arriva sempre tra i 18 e i 21 minuti, con la B si può impiegare da 14 a 27 minuti (figura 16.7). Se devi essere puntuale alla prima ora, la linea A è più affidabile. La media non dice nulla di quanto i dati sono *dispersi*: servono gli *indici di variabilità*, che misurano quanto i dati si allontanano tra loro o dalla media.

16.5.2 Il campo di variazione

L'indice più semplice è il **campo di variazione**, la differenza tra il dato massimo e il dato minimo: $x_{\text{max}} - x_{\text{min}}$. Per la linea A è 21 − 18 = 3 minuti, per la linea B 27 − 14 = 13 minuti. Si calcola in un attimo, ma dipende solo dai due dati estremi: un solo valore anomalo lo fa esplodere (per gli stipendi del paragrafo precedente è 8000 − 1500 = 6500 €).

16.5.3 Lo scarto semplice medio

Per tener conto di tutti i dati viene spontaneo fare la media degli scarti dalla media. Ma sappiamo già che la loro somma è sempre zero: gli scarti positivi e quelli negativi si compensano. Allora si prendono gli scarti in valore assoluto, cioè le distanze dei dati dalla media. Lo **scarto semplice medio** è la media dei valori assoluti degli scarti:

$$S = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots + |x_n - \bar{x}|}{n}$$

Per la giocatrice della figura 16.6 gli scarti sono −5, −2, 0, 0, 3, 4; i loro valori assoluti hanno somma 5 + 2 + 0 + 0 + 3 + 4 = 14, quindi $S = \frac{14}{6} = \frac{7}{3} \approx 2{,}33$ punti: in media, una partita si scosta di poco più di 2 punti dalla media.

16.5.4 Varianza e deviazione standard

Un altro modo per eliminare i segni è elevare gli scarti al quadrato. La **varianza** è la media dei quadrati degli scarti dalla media:

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}$$

La varianza ha un difetto: si misura nel quadrato dell'unità dei dati (se i dati sono in minuti, la varianza è in «minuti al quadrato»). Per tornare all'unità originale se ne prende la radice quadrata, cioè il numero positivo il cui quadrato è la varianza. Si ottiene la **deviazione standard**, detta anche *scarto quadratico medio*:

$$\sigma = \sqrt{\sigma^2}$$

In questo libro σ è la deviazione standard di **tutti** i dati considerati, trattati come l'intera popolazione: per questo al denominatore della varianza c'è n, il numero dei dati. Quando la varianza non è un quadrato perfetto, la radice quadrata si calcola con la calcolatrice e si arrotonda: per esempio, se $\sigma^2 = 5$, allora $\sigma \approx 2{,}24$, perché $2{,}24^2 = 5{,}0176$. Numeri come la radice quadrata di 5, che come quella di 2 (capitolo 3) non sono razionali, li studierai il prossimo anno.

ESEMPIO SVOLTO

Problema. Calcola campo di variazione, scarto semplice medio, varianza e deviazione standard dei punti segnati dalla giocatrice della figura 16.6: 8, 11, 13, 13, 16, 17.

Soluzione. La media è $\overline{x} = 78 : 6 = 13$. Il campo di variazione è $17 - 8 = 9$ punti. Organizziamo i calcoli in una tabella.

Tabella 16.5 Calcolo degli indici di variabilità.

x	8	11	13	13	16	17	somma **78**
scarto $x - 13$	−5	−2	0	0	+3	+4	somma **0**
|scarto|	5	2	0	0	3	4	somma **14**
scarto^2	25	4	0	0	9	16	somma **54**

La somma degli scarti è 0, come deve essere (è un utile controllo). Scarto semplice medio: $S = \frac{14}{6} \approx 2{,}33$ punti. Varianza: $\sigma^2 = \frac{54}{6} = 9$. Deviazione standard: $\sigma = 3$ punti, perché $3^2 = 9$.

Per le due linee di autobus si trova, con lo stesso procedimento, $\sigma = 1$ minuto per la linea A e $\sigma = 4$ minuti per la linea B (esercizio guidato: la somma dei quadrati degli scarti è 10 per A e 160 per B). La deviazione standard dà un'idea della distanza «tipica» dei dati dalla media: nella figura 16.7 la fascia da $\overline{x} - \sigma$ a $\overline{x} + \sigma$ è stretta per la linea A e larga per la linea B. Rispetto allo scarto semplice medio, la deviazione standard pesa di più gli scarti grandi, ed è l'indice di variabilità più usato nelle scienze. Una sua proprietà utile: se si aggiunge uno stesso numero a tutti i dati, la media aumenta di quel numero ma gli scarti non cambiano, quindi σ resta la stessa.

APPROFONDIMENTO

Perché n e non $n - 1$? Sulle calcolatrici scientifiche e nei fogli di calcolo troverai due deviazioni standard: una con n al denominatore della varianza, come in questo libro, e una con $n - 1$. La prima descrive i dati che si hanno, considerati come un'intera popolazione. La seconda si usa quando i dati sono un campione e si vuole *stimare* la variabilità della popolazione: dividendo per n la stima risulterebbe in media un po' troppo piccola. Con pochi dati la differenza si sente: per i 6 dati dell'esempio svolto la varianza diventa $54 : 5 = 10{,}8$ invece di 9.

16.5.5 Il coefficiente di variazione

Come confrontare la variabilità di grandezze diverse, per esempio l'altezza e la massa degli studenti di una scuola? Le deviazioni standard si misurano in unità diverse (cm e kg) e non si possono confrontare. Si usa allora il **coefficiente di variazione**, il rapporto tra deviazione standard

e media, spesso espresso in percentuale: CV = $\frac{\sigma}{\bar{x}}$ (si usa per dati positivi). Se in un gruppo immaginario di ragazzi l'altezza ha media 168 cm e σ = 8,4 cm, e la massa ha media 60 kg e σ = 9 kg, si ha CV = 0,05 = 5% per l'altezza e CV = 0,15 = 15% per la massa: in proporzione, la massa varia tre volte più dell'altezza.

IN SINTESI

Posizione. Media $\bar{x} = \frac{x_1 + \ldots + x_n}{n}$; media ponderata $\frac{x_1 p_1 + \ldots + x_k p_k}{p_1 + \ldots + p_k}$; mediana: valore centrale dei dati ordinati (media dei due centrali se n è pari); moda: modalità più frequente.

Variabilità. Campo $x_{\max} - x_{\min}$; scarto semplice medio S = media dei valori assoluti $|x_i - \bar{x}|$; varianza σ^2 = media degli $(x_i - \bar{x})^2$ (con n al denominatore); deviazione standard $\sigma = \sqrt{\sigma^2}$; coefficiente di variazione $\frac{\sigma}{\bar{x}}$.

COLLEGAMENTI

Il foglio di calcolo. Con un foglio di calcolo (in versione italiana) gli indici si ottengono con poche funzioni. Se i 24 voti della tabella 16.3 sono scritti nelle celle da B2 a B25:

- `=MEDIA(B2:B25)` restituisce 6,625; `=MEDIANA(B2:B25)` restituisce 6,5; `=MODA(B2:B25)` restituisce 6;
- `=MAX(B2:B25)-MIN(B2:B25)` dà il campo di variazione, 5;
- `=DEV.ST.POP(B2:B25)` dà la deviazione standard con n al denominatore, circa 1,25 (attenzione: `DEV.ST` usa invece $n - 1$); `=VAR.POP(B2:B25)` dà la varianza;
- `=FREQUENZA(B2:B25;D2:D6)`, dove D2:D6 contiene i limiti superiori delle classi, conta quanti dati cadono in ciascuna classe (attenzione: questa funzione conta i dati *minori o uguali* al limite, cioè usa classi chiuse a destra).

Nella versione italiana gli argomenti si separano con il punto e virgola. Il foglio di calcolo è uno strumento: fa i conti senza errori, ma non sceglie al posto tuo quale indice o quale grafico ha senso.

SINTESI

- Un'indagine statistica riguarda una popolazione di unità statistiche; se si osserva solo un sottoinsieme si parla di campione, che deve essere rappresentativo.
- Il carattere è ciò che si osserva, le modalità sono i suoi valori. I caratteri possono essere qualitativi (sconnessi od ordinabili) o quantitativi (discreti o continui).
- Frequenza assoluta f; relativa $f_r = f : n$; percentuale $f_r \cdot 100$; cumulata, solo per modalità ordinabili.
- I dati continui si raggruppano in classi $[a; b[$ di ampiezza $b - a$, rappresentate dal valore centrale $(a + b) : 2$.
- Ortogramma (barre separate) per caratteri qualitativi e discreti; areogramma, con angoli $\alpha = f_r \cdot 360°$, per le parti di un tutto; spezzata per le serie temporali.
- Nell'istogramma i rettangoli sono contigui e la loro area è proporzionale alla frequenza: con classi di ampiezza diversa l'altezza è la densità di frequenza f : ampiezza.
- Un grafico è ingannevole se l'asse delle barre non parte da zero, se le figure crescono in area o in volume, se le scale non sono dichiarate.
- Media aritmetica: somma dei dati diviso n; la somma degli scarti dalla media è zero. Media ponderata: ogni valore pesa secondo il suo peso (per esempio la frequenza).
- Mediana: il valore centrale dei dati ordinati (con n pari, la media dei due centrali). Moda: la modalità più frequente. La mediana è robusta rispetto ai valori anomali, la media no.
- Indici di variabilità: campo di variazione, scarto semplice medio, varianza σ^2 (media dei quadrati degli scarti, con n al denominatore) e deviazione standard σ, sua radice quadrata.
- Il coefficiente di variazione $\sigma : \overline{x}$ confronta la variabilità di grandezze con unità o ordini di grandezza diversi.

ESERCIZI

Comprendere

1. ● Il carattere «colore degli occhi», rilevato sugli studenti di una classe, è:
   - **A.** quantitativo discreto
   - **B.** qualitativo ordinabile
   - **C.** qualitativo sconnesso
   - **D.** quantitativo continuo

2. ● Il carattere «numero di fratelli e sorelle» è:

   A. quantitativo discreto
   B. quantitativo continuo
   C. qualitativo ordinabile
   D. qualitativo sconnesso

3. ● In un areogramma, una modalità con frequenza relativa 0,15 è rappresentata da un settore di:

   A. 15°
   B. 36°
   C. 45°
   D. 54°

4. ● La mediana dei dati 3, 8, 5, 10, 7, 2 è:

   A. 5,5
   B. 6
   C. 7,5
   D. 5

5. ● Quale di questi indici cambia di meno se a un insieme di dati si aggiunge un valore anomalo molto grande?

   A. la media aritmetica
   B. il campo di variazione
   C. la media ponderata con pesi uguali
   D. la mediana

6. ●● In un istogramma le classi hanno ampiezze diverse. L'altezza di ciascun rettangolo deve essere:

   A. la densità di frequenza della classe
   B. la frequenza assoluta della classe
   C. il valore centrale della classe
   D. la frequenza cumulata della classe

7. ●● Se si aggiunge 5 a ciascuno dei dati, la deviazione standard:

   A. aumenta di 5
   B. aumenta di 25
   C. non cambia
   D. viene moltiplicata per 5

8. ● Vero o falso? Correggi le affermazioni false. a) Il campione è un sottoinsieme della popolazione. b) La somma delle frequenze relative è 100. c) Le frequenze cumulate hanno senso anche per un carattere qualitativo sconnesso. d) La media aritmetica di un insieme di dati è sempre uno dei dati. e) Una distribuzione può avere più di una moda. f) Nell'istogramma i rettangoli sono separati da spazi.

9. ● Completa. a) Il singolo elemento della popolazione su cui si rileva un carattere si chiama _____. b) I diversi valori che un carattere può assumere si chiamano _____. c) Il rapporto tra frequenza assoluta e numero totale dei dati è la _____. d) La differenza tra il dato massimo e il dato minimo si chiama _____. e) La radice quadrata della varianza si chiama _____.

Calcolare

10. ● Il numero di libri letti in un anno da 25 studenti (dati inventati) è: 1, 0, 2, 5, 4, 4, 4, 0, 1, 3, 1, 3, 3, 2, 1, 2, 4, 0, 2, 1, 2, 2, 5, 3, 2. a) Costruisci la tabella delle frequenze assolute, relative, percentuali e cumulate. b) Quanti studenti hanno letto al massimo 2 libri? Che percentuale rappresentano? c) Disegna l'ortogramma.

11. ● Le altezze in centimetri di 20 studenti sono: 168, 169, 158, 166, 161, 181, 175, 177, 165, 174, 163, 183, 172, 171, 170, 179, 175, 186, 164, 152. Raggruppale nelle classi [150; 160[, [160; 170[, [170; 180[, [180; 190[e scrivi per ciascuna frequenza, frequenza percentuale e valore centrale.

12. ●● In una classe di 26 studenti lo sport preferito è: calcio 9, pallavolo 8, nuoto 6, basket 3. Calcola gli angoli dell'areogramma arrotondati al grado, controlla che la somma sia 360° e, se necessario, correggi gli arrotondamenti.

13. ●● L'età dei visitatori di una mostra in una mattina (dati inventati) è distribuita così: [0; 10[anni 12 persone; [10; 20[30; [20; 40[36; [40; 60[28; [60; 90[24. a) Calcola la densità di frequenza di ogni classe. b) Quale classe ha il rettangolo più alto nell'istogramma? È anche la classe con la frequenza più alta? c) Disegna l'istogramma.

14. ● Calcola media, mediana e moda dei dati: a) 12, 15, 9, 15, 11, 18, 14; b) 21, 25, 19, 30, 25, 22, 28, 26.

15. ●● In un condominio immaginario di 40 famiglie il numero di figli è: 0 figli 6 famiglie, 1 figlio 14, 2 figli 13, 3 figli 5, 4 figli 2. Calcola media, mediana e moda.

16. ● Il voto finale di un laboratorio è la media ponderata di tre voti: relazioni (peso 50%), voto 6,5; prova pratica (peso 30%), voto 7; quaderno (peso 20%), voto 8. Calcola il voto finale.

17. ●● a) I dati 6, 7, 5, 8, x hanno media 7: trova x. b) Quattro numeri hanno media 12; aggiungendo un quinto numero la media diventa 13. Qual è il quinto numero?

18. ●● Per i dati 5, 8, 9, 12, 12, 14 calcola campo di variazione, scarto semplice medio, varianza e deviazione standard.

19. ●● a) Calcola varianza e deviazione standard dei dati 2, 4, 6, 8 (approssima σ ai centesimi con la calcolatrice). b) Quanto valgono varianza e deviazione standard dei dati 10, 10, 10, 10? Spiega il risultato.

20. ● In un gruppo di ragazzi l'altezza ha media 165 cm e deviazione standard 6,6 cm; la massa ha media 55 kg e deviazione standard 8,8 kg. Calcola i due coefficienti di variazione e stabilisci quale grandezza è relativamente più variabile.

21. ●● Usando le classi e i valori centrali dell'esercizio 11, stima l'altezza media dei 20 studenti. Poi calcola la media esatta dai dati originali e confronta i due risultati.

Dimostrare e argomentare

22. ●● Dimostra che la somma degli scarti di tre numeri a, b, c dalla loro media è zero, svolgendo il calcolo letterale con $\overline{x} = \dfrac{a+b+c}{3}$.

23. ●● Si aggiunge uno stesso numero k a tutti i dati $x_1, \ldots, x_n$. Dimostra che la nuova media è $\overline{x} + k$ e che gli scarti dalla media non cambiano. Che cosa si deduce per la deviazione standard?

24. ● Trova l'errore. Marco deve calcolare la mediana dei dati 7, 3, 9, 4, 6 e scrive: «I dati sono 5, quindi la mediana è il terzo, cioè 9». Spiega l'errore e correggilo.

25. ●● Spiega perché in un istogramma con classi di ampiezza diversa non si può usare la frequenza come altezza dei rettangoli. Fai riferimento ai dati della tabella 16.4.

26. ●●● La classe A, di 20 studenti, ha la media di 7 in una verifica; la classe B, di 25 studenti, ha la media di 6. Giulia dice: «La media di tutti i 45 studenti è (7 + 6) : 2 = 6,5». Ha ragione? Calcola la media corretta e spiega in quale caso il ragionamento di Giulia sarebbe giusto.

Problemi

27. ●● Un volantino mostra gli iscritti a un corso pomeridiano (dati inventati): 46 nel 2024 e 52 nel 2025, con un ortogramma il cui asse verticale parte da 44 (figura 16.5a). a) Quante volte la seconda barra è più alta della prima? b) Di quanto per cento sono aumentati davvero gli iscritti? c) Come andrebbe corretto il grafico?

28. ●● Nel primo periodo Luca ha preso in matematica tre voti, 5, 6,5 e 5,5, e gli manca una verifica, con lo stesso peso delle altre. Quale voto deve prendere, almeno, per avere una media di almeno 6? Risolvi con una disequazione.

29. ●● In laboratorio di fisica si misura sei volte il periodo di oscillazione di un pendolo, ottenendo (in secondi): 1,42; 1,45; 1,40; 1,44; 1,43; 1,46. a) Calcola il valor medio, arrotondato ai centesimi. b) Calcola la semidispersione massima, cioè metà del campo di variazione, che in fisica si usa come stima dell'incertezza. c) Scrivi la misura nella forma «valor medio ± incertezza».

30. ●● Gli stipendi mensili dei 7 dipendenti di una ditta immaginaria sono 1500, 1600, 1600, 1700, 1800, 1900, 8000 euro. a) Calcola media, mediana e moda. b) La ditta pubblicizza: «Da noi si guadagnano in media più di 2500 euro al mese». È falso? È corretto? c) Quale indice useresti per descrivere lo stipendio tipico, e perché?

31. ●● Con riferimento ai tempi delle due linee di autobus del paragrafo 16.5.1, verifica che la deviazione standard è 1 minuto per la linea A e 4 minuti per la linea B. Se hai un appuntamento importante e parti 25 minuti prima, quale linea ti conviene? Motiva la risposta usando gli indici.

Con il computer

32. ●● Scrivi i 20 dati dell'esercizio 34 (un dato per studente) nelle celle da A1 ad A20 di un foglio di calcolo. a) Con =CONTA.SE(A1:A20;0), =CONTA.SE(A1:A20;1), … costruisci la tabella delle frequenze assolute e disegna l'ortogramma. b) Controlla con =MEDIA, =MEDIANA e =MODA i risultati dell'esercizio 32 e calcola =DEV.ST.POP(A1:A20) (approssima ai centesimi). c) Guardando il grafico, quali di queste affermazioni sono giustificate dai dati? «Più della metà degli studenti della classe ha fatto meno di 3 assenze»; «Nella scuola la media delle assenze è 2,8»; «Uno studente su dieci ha fatto più di 5 assenze».

33. ●●● **Progetto: un'indagine nella tua classe.** Con i compagni scegli un carattere qualitativo (per esempio il mezzo con cui si arriva a scuola) e uno quantitativo (per esempio i minuti impiegati, o le ore di sonno della notte precedente) e raccogli i dati di tutta la classe in un foglio di calcolo. a) Classifica i due caratteri e costruisci le tabelle di frequenza (per il carattere quantitativo continuo usa le classi). b) Scegli e disegna il grafico più adatto a ciascun carattere. c) Calcola media, mediana, moda e deviazione standard del carattere quantitativo. d) Scrivi due conclusioni che i dati permettono di trarre e una che *non* permettono, spiegando perché.

Verso la verifica

34. ●● Le assenze nel primo periodo dei 20 studenti di una classe immaginaria sono: 0 assenze 3 studenti; 1 assenza 4; 2 assenze 5; 3 assenze 3; 4 assenze 2; 5 assenze 1; 8 assenze 1; 12 assenze 1. a) Costruisci la tabella con frequenze relative, percentuali e cumulate. b) Calcola media, mediana, moda e campo di variazione. c) Ricalcola media, mediana, moda e campo di variazione escludendo lo studente con 12 assenze (approssima ai centesimi). Quali indici cambiano e quali no? Perché?

35. ●●● Trova cinque numeri interi che abbiano media 8, mediana 9, moda 10 (unica) e campo di variazione 6. Spiega perché la soluzione è unica.

36. ●●● **Sfida.** Cinque numeri interi positivi hanno media 10, mediana 8 e una sola moda, uguale a 7. Qual è il massimo campo di variazione possibile? Scrivi i cinque numeri che lo realizzano e giustifica ogni passaggio.

PER APPROFONDIRE

- [HUF] Huff, *Mentire con le statistiche*: un classico divertente e ancora attuale sui modi in cui medie scelte ad arte, campioni distorti e grafici truccati possono ingannare il lettore.
- [POL] Pólya, *Come risolvere i problemi di matematica*: il metodo per affrontare un problema (comprendere, progettare, eseguire, verificare) si applica bene anche alle indagini statistiche.
- [MIUR] *Indicazioni nazionali per i licei*: collocano la statistica descrittiva nel tema «Dati e previsioni» del primo biennio, insieme all'uso consapevole degli strumenti di calcolo.

Glossario del capitolo

Areogramma
Diagramma a torta: ogni modalità è un settore circolare di angolo $\alpha = f_r \cdot 360°$.

Campione
Sottoinsieme della popolazione su cui si effettua la rilevazione; deve essere rappresentativo della popolazione.

Campo di variazione
Differenza tra il dato massimo e il dato minimo.

Carattere
Aspetto che si osserva sulle unità statistiche. È qualitativo (sconnesso od ordinabile) se le modalità sono parole, quantitativo (discreto o continuo) se sono numeri su cui ha senso calcolare.

Classe di valori
Intervallo [a; b[in cui si raggruppano i dati; ha ampiezza b − a e valore centrale (a + b) : 2.

Densità di frequenza
Rapporto tra la frequenza di una classe e la sua ampiezza; è l'altezza dei rettangoli dell'istogramma quando le classi hanno ampiezze diverse.

Deviazione standard
Radice quadrata della varianza, detta anche scarto quadratico medio (σ); misura la dispersione dei dati nella stessa unità dei dati.

Frequenza assoluta
Numero di volte in cui una modalità compare nei dati.

Frequenza cumulata
Somma delle frequenze di una modalità e di tutte quelle che la precedono; ha senso solo per modalità ordinabili.

Frequenza relativa
Rapporto f : n tra la frequenza assoluta di una modalità e il numero totale dei dati; moltiplicata per 100 dà la frequenza percentuale.

Istogramma
Grafico per dati raggruppati in classi, fatto di rettangoli contigui la cui area è proporzionale alla frequenza della classe.

Media aritmetica
Somma dei dati divisa per il loro numero; la somma degli scarti dei dati dalla media è zero.

Media ponderata
Media in cui ogni valore è moltiplicato per un peso: $(x_1p_1 + \ldots + x_kp_k) : (p_1 + \ldots + p_k)$.

Mediana (in statistica)
Valore centrale dei dati ordinati; se i dati sono in numero pari, media dei due valori centrali. È poco sensibile ai valori anomali.

Moda
Modalità con la frequenza più alta; si può determinare per ogni tipo di carattere.

Modalità
Ciascuno dei diversi valori che un carattere può assumere.

Ortogramma
Diagramma a barre separate, di altezza proporzionale alle frequenze; adatto a caratteri qualitativi e quantitativi discreti.

Popolazione statistica
Insieme di tutti gli elementi (persone, oggetti, eventi) su cui si svolge un'indagine statistica.

Unità statistica
Singolo elemento della popolazione su cui si rileva un carattere.

Varianza
Media dei quadrati degli scarti dei dati dalla media (con n, numero dei dati, al denominatore); si indica con σ^2.

Soluzioni

1. C. Le modalità (azzurri, castani, verdi...) sono parole senza un ordine naturale.

2. A. Si ottiene contando: assume solo valori interi isolati (0, 1, 2, ...).

3. D. $\alpha = 0{,}15 \cdot 360° = 54°$.

4. B. Dati ordinati: 2, 3, 5, 7, 8, 10; con 6 dati la mediana è la media del 3° e del 4°: (5 + 7) : 2 = 6.

5. D. La mediana dipende solo dall'ordine dei dati e si sposta al più di una posizione; media e campo di variazione possono cambiare moltissimo.

6. A. Con classi di ampiezza diversa l'altezza è la densità f : ampiezza, così che l'area sia proporzionale alla frequenza.

7. C. La media aumenta di 5, ma gli scarti dalla media restano gli stessi, quindi varianza e deviazione standard non cambiano.

8. a) V. b) F: la somma delle frequenze relative è 1; è la somma delle percentuali a fare 100. c) F: le cumulate hanno senso solo se le modalità sono ordinabili. d) F: per esempio la media di 1 e 2 è 1,5. e) V (distribuzioni bimodali o plurimodali). f) F: nell'istogramma i rettangoli sono contigui, perché le classi sono intervalli consecutivi; separati sono le barre dell'ortogramma.

9. a) unità statistica; b) modalità; c) frequenza relativa; d) campo di variazione; e) deviazione standard (o scarto quadratico medio).

10. a) Libri 0: $f = 3$, $f_r = 0{,}12$, 12%, cumulata 3 (12%); 1: 5, 0,20, 20%, 8 (32%); 2: 7, 0,28, 28%, 15 (60%); 3: 4, 0,16, 16%, 19 (76%); 4: 4, 0,16, 16%, 23 (92%); 5: 2, 0,08, 8%, 25 (100%). b) Al massimo 2 libri: cumulata del 2, cioè 15 studenti, il 60%. c) Sei barre separate di altezze 3, 5, 7, 4, 4, 2.

11. [150; 160[: $f = 2$, 10%, valore centrale 155; [160; 170[: 7, 35%, 165; [170; 180[: 8, 40%, 175; [180; 190[: 3, 15%, 185. Totale 20 (il dato 170 va nella classe [170; 180[).

12. Angoli esatti: calcio 9/26 · 360° ≈ 124,62°; pallavolo 8/26 · 360° ≈ 110,77°; nuoto 6/26 · 360° ≈ 83,08°; basket 3/26 · 360° ≈ 41,54°. Arrotondati: 125°, 111°, 83°, 42°, con somma 361°. C'è un grado in più: lo si toglie al settore arrotondato per eccesso con la parte decimale più piccola, il basket (41,54° → 41°). Angoli: 125°, 111°, 83°, 41° (somma 360°).

13. a) Densità (persone all'anno di età): 12 : 10 = 1,2; 30 : 10 = 3; 36 : 20 = 1,8; 28 : 20 = 1,4; 24 : 30 = 0,8. b) Il rettangolo più alto è quello della classe [10; 20[(densità 3), mentre la frequenza più alta è della classe [20; 40[(36 persone): non coincidono, perché la classe [20; 40[è larga il doppio. c) Rettangoli contigui su 0, 10, 20, 40, 60, 90, alti 1,2; 3; 1,8; 1,4; 0,8.

14. a) Ordinati: 9, 11, 12, 14, 15, 15, 18. Media 94 : 7 ≈ 13,43; mediana (4° dato) 14; moda 15. b) Ordinati: 19, 21, 22, 25, 25, 26, 28, 30. Media 196 : 8 = 24,5; mediana (25 + 25) : 2 = 25; moda 25.

15. Media (0 · 6 + 1 · 14 + 2 · 13 + 3 · 5 + 4 · 2) : 40 = 63 : 40 = 1,575 figli. Cumulate 6, 20, 33, 38, 40: il 20° dato è 1 e il 21° è 2, quindi mediana (1 + 2) : 2 = 1,5. Moda 1 figlio (14 famiglie).

16. 6,5 · 0,5 + 7 · 0,3 + 8 · 0,2 = 3,25 + 2,1 + 1,6 = 6,95.

17. a) La somma deve essere 7 · 5 = 35; 6 + 7 + 5 + 8 = 26, quindi $x = 9$. b) Somma dei quattro: 4 · 12 = 48; somma dei cinque: 5 · 13 = 65; il quinto numero è 65 − 48 = 17.

18. Media 60 : 6 = 10. Campo 14 − 5 = 9. Scarti −5, −2, −1, 2, 2, 4 (somma 0). Scarto semplice medio (5 + 2 + 1 + 2 + 2 + 4) : 6 = 16 : 6 ≈ 2,67. Varianza (25 + 4 + 1 + 4 + 4 + 16) : 6 = 54 : 6 = 9. Deviazione standard $\sigma = 3$.

19. a) Media 5; scarti −3, −1, 1, 3; quadrati 9, 1, 1, 9; varianza 20 : 4 = 5; $\sigma \approx 2{,}24$. b) Media 10, tutti gli scarti sono 0: varianza e deviazione standard valgono 0, perché i dati non variano affatto.

20. Altezza: 6,6 : 165 = 0,04 = 4%. Massa: 8,8 : 55 = 0,16 = 16%. La massa è relativamente più variabile

(quattro volte tanto).

21. Stima: (155 · 2 + 165 · 7 + 175 · 8 + 185 · 3) : 20 = 3420 : 20 = 171 cm. Media esatta: somma dei 20 dati 3409, quindi 3409 : 20 = 170,45 cm. La stima differisce di 0,55 cm: raggruppando si perde un po' di informazione.

22. $(a - \overline{x}) + (b - \overline{x}) + (c - \overline{x}) = a + b + c - 3\overline{x} = a + b + c - 3 \cdot (a + b + c)/3 = a + b + c - (a + b + c) = 0$.

23. Nuova media: $[(x_1 + k) + \ldots + (x_n + k)] : n = (x_1 + \ldots + x_n + nk) : n = \overline{x} + k$. Nuovo scarto: $(x_i + k) - (\overline{x} + k) = x_i - \overline{x}$, uguale al vecchio. Varianza e deviazione standard dipendono solo dagli scarti, quindi non cambiano (lo stesso vale per lo scarto semplice medio e il campo di variazione).

24. Marco non ha ordinato i dati: la mediana è il valore centrale dei dati *ordinati*. Ordinati: 3, 4, 6, 7, 9; la mediana è 6.

25. L'occhio confronta le aree dei rettangoli, che devono essere proporzionali alle frequenze. Con la frequenza come altezza l'area diventa frequenza · ampiezza e le classi larghe sembrano più importanti. Nella tabella 16.4 la classe [45; 75[(3 studenti, larga 30) avrebbe area 90, quasi come la classe [10; 20[(12 studenti, area 120), pur avendo un quarto degli studenti. Usando la densità (0,1 e 1,2) le aree tornano 3 e 12.

26. No. La somma dei voti è 20 · 7 + 25 · 6 = 140 + 150 = 290; la media generale è 290 : 45 = 58/9 ≈ 6,44. È una media ponderata delle due medie, con pesi 20 e 25. Il ragionamento di Giulia sarebbe giusto solo se le due classi avessero lo stesso numero di studenti.

27. a) Le barre sono alte 46 − 44 = 2 e 52 − 44 = 8 unità: la seconda è 4 volte la prima. b) (52 − 46) : 46 = 6 : 46 ≈ 0,13: aumento del 13% circa. c) Far partire l'asse verticale da 0 (figura 16.5b).

28. $(5 + 6{,}5 + 5{,}5 + x) : 4 \geq 6$, cioè $17 + x \geq 24$, quindi $x \geq 7$. Luca deve prendere almeno 7.

29. a) Somma 8,60 s; valor medio 8,60 : 6 ≈ 1,433 ≈ 1,43 s. b) Campo 1,46 − 1,40 = 0,06 s; semidispersione massima 0,03 s. c) $T = (1{,}43 \pm 0{,}03)$ s.

30. a) Media 18 100 : 7 ≈ 2586 €; mediana 1700 €; moda 1600 €. b) Non è falso (la media supera 2500 €), ma è fuorviante: sei dipendenti su sette guadagnano 1900 € o meno; la media è alzata da un solo stipendio di 8000 €. c) La mediana, perché è robusta rispetto al valore anomalo.

31. Linea A: scarti −2, −1, −1, 0, 0, 0, 1, 1, 1, 1; somma dei quadrati 10; varianza 10 : 10 = 1; σ = 1 min. Linea B: scarti −6, −4, −4, −3, 0, 1, 2, 2, 5, 7; somma dei quadrati 36 + 16 + 16 + 9 + 0 + 1 + 4 + 4 + 25 + 49 = 160; varianza 16; σ = 4 min. Conviene la linea A: la media è la stessa (20 min), ma la variabilità è molto minore; nei 10 giorni non ha mai superato 21 minuti, mentre con la B un giorno (27 minuti) saresti arrivato in ritardo e un altro (25 minuti) appena in tempo. Con la B, media + σ = 24 minuti lascia poco margine rispetto ai 25 disponibili; con la A, media + σ = 21 minuti ne lascia molto.

32. a) Frequenze 3, 4, 5, 3, 2, 1, 1, 1 per 0, 1, 2, 3, 4, 5, 8, 12 assenze (per i valori 6, 7, 9, 10, 11 CONTA.SE dà 0). b) MEDIA = 2,8; MEDIANA = 2; MODA = 2; DEV.ST.POP ≈ 2,82 (la varianza, =VAR.POP, è 7,96). c) La prima è giustificata: meno di 3 assenze hanno 3 + 4 + 5 = 12 studenti su 20, il 60%. La seconda no: i dati riguardano una sola classe, non tutta la scuola, e non si sa se la classe sia rappresentativa. La terza è giustificata: più di 5 assenze hanno 2 studenti su 20 (8 e 12 assenze), cioè uno su dieci.

33. *Traccia.* a) Il mezzo di trasporto è qualitativo non ordinabile; i minuti (o le ore di sonno) sono quantitativi continui: si raggruppano in classi, per esempio di 10 minuti. b) Ortogramma o areogramma per il carattere qualitativo, istogramma per quello quantitativo in classi (con la densità di frequenza se le classi hanno ampiezza diversa). c) Funzioni MEDIA, MEDIANA, MODA, DEV.ST.POP. d) Conclusioni ammissibili descrivono la classe («metà della classe impiega meno di … minuti»); non è ammissibile estenderle a tutti gli studenti della scuola o della città, perché la classe non è un campione scelto per rappresentarli, né leggere nei dati una causa («chi viene in bicicletta dorme di più perché…»).

34. a) f_r: 0,15; 0,20; 0,25; 0,15; 0,10; 0,05; 0,05; 0,05 (percentuali 15, 20, 25, 15, 10, 5, 5, 5). Cumulate: 3, 7, 12, 15, 17, 18, 19, 20 (15%, 35%, 60%, 75%, 85%,

90%, 95%, 100%). b) Somma $0 \cdot 3 + 1 \cdot 4 + 2 \cdot 5 + 3 \cdot 3 + 4 \cdot 2 + 5 + 8 + 12 = 56$; media $56 : 20 = 2{,}8$ assenze. Il 10° e l'11° dato valgono 2 (cumulate 7 e 12): mediana 2. Moda 2. Campo $12 - 0 = 12$. c) Senza il 12 restano 19 dati con somma 44: media $44 : 19 \approx 2{,}32$; mediana 2 (il 10° dei 19 dati ordinati); moda 2; campo $8 - 0 = 8$. Cambiano la media (da 2,8 a 2,32) e il campo di variazione (da 12 a 8), che dipendono dal valore del dato anomalo; mediana e moda restano 2, perché dipendono solo dall'ordine dei dati e dalle frequenze: sono robuste.

35. Ordinati: $a \le b \le 9 \le d \le e$ (la mediana è il 3° dato). La moda 10 compare almeno due volte e solo d ed e possono valere 10: $d = e = 10$. Campo 6: $a = 10 - 6 = 4$. Somma $5 \cdot 8 = 40$: $b = 40 - 4 - 9 - 10 - 10 = 7$. Numeri: 4, 7, 9, 10, 10. Ogni passaggio è obbligato, quindi la soluzione è unica (e 4, 7, 9 compaiono una sola volta, perciò la moda 10 è unica).

36. Ordinati: $a \le b \le 8 \le d \le e$. La moda 7 deve comparire almeno due volte, e solo a e b possono valere 7: $a = b = 7$. La somma è $5 \cdot 10 = 50$, quindi $d + e = 50 - 22 = 28$. $d \ne 8$ (altrimenti 8 comparirebbe due volte come 7 e la moda non sarebbe unica) e $d \ne e$ (14 e 14, stesso motivo): quindi $9 \le d \le 13$. Il campo è $e - 7 = 21 - d$, massimo per $d = 9$: $e = 19$. Numeri 7, 7, 8, 9, 19, campo di variazione massimo 12.