CAPITOLO

11

Statistica a due variabili

Due numeri che crescono insieme raccontano una storia; la statistica insegna a chiedersi chi l'abbia scritta. (testo della redazione)

OBIETTIVI

1. Distinguere la deviazione standard σ di una popolazione e quella campionaria s, e standardizzare un dato.
2. Costruire e leggere una tabella a doppia entrata con le distribuzioni marginali e condizionate.
3. Riconoscere l'indipendenza statistica con le frequenze teoriche e diffidare dei confronti che un terzo carattere può alterare.
4. Calcolare covarianza e coefficiente di correlazione lineare, e distinguere correlazione e causalità.
5. Determinare la retta di regressione con il metodo dei minimi quadrati, controllarla con i residui e valutare le previsioni.
6. Riconoscere un campione casuale e le principali distorsioni, e leggere un modello statistico come esempio di modellistica.

INDICE DEL CAPITOLO

11.1 Richiamo: gli indici di un carattere 2
11.2 Distribuzioni doppie 3
11.3 Dipendenza e indipendenza statistica 6
11.4 Due caratteri quantitativi: la correlazione 9
11.5 La retta di regressione 14
11.6 Campione e modello 17

Matematica 4 · Parte IV · Contare, prevedere, misurare la dipendenza · versione 1.0
Capitolo 11: Statistica a due variabili · pagine 291–316 del volume completo · estratto per lo studio

Chi fa sport dorme di più? Chi è più alto pesa di più? Più ore passate davanti a uno schermo la sera significano meno ore di sonno? Sono domande su *due* caratteri osservati sulle stesse persone, e gli indici che conosciamo da *Matematica 1* (cap. 12), la media, la mediana, la deviazione standard, descrivono un carattere alla volta: per rispondere non bastano più. In questo capitolo impariamo a leggere due caratteri insieme. Li raccoglieremo in una tabella a doppia entrata, vedremo quando si può dire che uno «dipende» dall'altro, misureremo con un solo numero quanto una nuvola di punti somiglia a una retta e tracceremo la retta che meglio la descrive.

Tutti i dati del capitolo sono **dati inventati**, costruiti per gli esempi, o **dati simulati** al computer; molti, però, sono del tipo che una classe può raccogliere da sé, con un metro, una bilancia e un questionario anonimo. Useremo la probabilità condizionata del capitolo 10, la retta e la parabola di *Matematica 3*. Chiuderemo con due idee che le *Indicazioni nazionali* mettono al centro della statistica del secondo biennio: il *campione*, cioè la parte che osserviamo di un insieme più grande, e il *modello matematico*.

11.1 Richiamo: gli indici di un carattere

Un carattere quantitativo, osservato su n unità statistiche, produce n dati $x_1, x_2, \ldots, x_n$. Il simbolo Σ (sigma maiuscola) abbrevia una somma: $\Sigma\, x_i$ significa $x_1 + x_2 + \ldots + x_n$. In *Matematica 1* abbiamo riassunto i dati con un indice di posizione, la media aritmetica $\bar{x}$, e con un indice di variabilità, la **deviazione standard** σ, radice quadrata della **varianza** σ^2, che è la media dei quadrati degli scarti dalla media. Gli scarti $x_i - \bar{x}$ hanno sempre somma zero, perché $\Sigma\,(x_i - \bar{x}) = \Sigma\, x_i - n\bar{x} = 0$: per questo, nella varianza, si elevano al quadrato, così che i positivi e i negativi non si cancellino.

Ora serve una distinzione che nel primo volume era solo accennata. Se i dati sono *tutta* la popolazione che ci interessa (i 25 studenti di una classe, quando la domanda riguarda quella classe), la varianza si calcola dividendo per il numero N dei dati: è σ^2. Se invece i dati sono un **campione**, una parte scelta per conoscere un insieme più grande, e vogliamo stimare la variabilità di quell'insieme, si divide per $n - 1$, dove n è il numero dei dati del campione. Si ottiene la **varianza campionaria** s^2, e la sua radice s è la **deviazione standard campionaria**. Il perché lo vedremo nell'ultimo paragrafo; per ora basta sapere che s è sempre un po' più grande di σ e che la differenza diventa trascurabile quando i dati sono molti. In questo libro σ indica sempre l'indice con N, s quello con $n - 1$; le calcolatrici scientifiche hanno di solito due tasti distinti, indicati con σ_n e σ_{n-1} o con σ_x e s_x.

IN SINTESI

Indici di un carattere quantitativo.

- media: $\bar{x} = \frac{\Sigma x_i}{N}$;
- varianza e deviazione standard (popolazione di N dati): $\sigma^2 = \frac{\Sigma (x_i - \bar{x})^2}{N}$, $\sigma = \sqrt{\sigma^2}$;
- varianza e deviazione standard campionarie (n dati di un campione): $s^2 = \frac{\Sigma (x_i - \bar{x})^2}{n - 1}$, $s = \sqrt{s^2}$;
- valore standardizzato di un dato x: $z = \frac{x - \bar{x}}{\sigma}$.

L'ultima formula è nuova. Il **valore standardizzato** z (o *punteggio z*) dice di quante deviazioni standard un dato sta sopra ($z > 0$) o sotto ($z < 0$) la media: misura ogni distanza con il «metro» σ della sua distribuzione. Così diventano confrontabili dati che non lo sono. Giulia ha preso 8 nella verifica di latino, dove la classe ha media 6,4 e $\sigma = 0{,}8$; in quella di matematica ha preso 7, con media 5,8 e $\sigma = 1{,}5$. In latino $z = \frac{8 - 6{,}4}{0{,}8} = 2$, in matematica $z = \frac{7 - 5{,}8}{1{,}5} = 0{,}8$: il voto più basso è un buon risultato, ma l'8 in latino è davvero eccezionale rispetto alla classe. I valori standardizzati di tutti i dati hanno media 0 e deviazione standard 1, qualunque fosse l'unità di misura di partenza.

11.2 Distribuzioni doppie

11.2.1 La tabella a doppia entrata

A 200 studenti di un liceo si chiede, con un questionario anonimo, quanto sport praticano (nessuno, saltuario, regolare: almeno tre allenamenti a settimana) e quante ore dormono di solito in una notte prima di un giorno di scuola (meno di 7, da 7 a 8, più di 8). Ogni studente fornisce una *coppia* di modalità, una per carattere. Per contarle si usa una **tabella a doppia entrata**: le modalità del primo carattere sulle righe, quelle del secondo sulle colonne, e in ogni casella il numero di studenti che presentano *insieme* quelle due modalità. Questo numero è una **frequenza congiunta**, e l'insieme delle frequenze congiunte è la **distribuzione doppia** dei due caratteri (Tabella 11.1).

Tabella 11.1 Pratica sportiva e ore di sonno di 200 studenti (dati inventati): frequenze congiunte, con i totali di riga e di colonna.

Sport \ sonno	meno di 7 h	da 7 a 8 h	più di 8 h	Totale
nessuno	27	24	9	60
saltuario	24	40	16	80
regolare	9	30	21	60
Totale	**60**	**94**	**46**	**200**

Le caselle si leggono con una «e»: 40 studenti fanno sport in modo saltuario *e* dormono da 7 a 8 ore; 9 non fanno sport *e* dormono più di 8 ore. La somma di tutte le frequenze congiunte è il numero $N = 200$ delle unità statistiche. La tabella può raccogliere caratteri qualitativi (il genere musicale preferito), quantitativi divisi in classi (le ore di sonno) o uno per tipo.

11.2.2 Distribuzioni marginali

Sommando le frequenze di una riga si ottiene il numero degli studenti con quel livello di sport, qualunque sia il loro sonno: 60, 80 e 60. Questi totali, nell'ultima colonna, formano la distribuzione del solo carattere «sport»; i totali delle colonne, 60, 94 e 46, formano la distribuzione del solo carattere «sonno». Si chiamano **distribuzioni marginali**, perché si leggono sui margini della tabella. Ciascuna è una distribuzione semplice, come quelle di *Matematica 1*: per esempio il $94 : 200 = 47\%$ degli studenti dorme da 7 a 8 ore.

Le due distribuzioni marginali, però, non dicono nulla sul *legame* tra i caratteri: con gli stessi margini si possono scrivere moltissime tabelle diverse, alcune in cui chi fa sport dorme di più, altre in cui dorme di meno. L'informazione sul legame sta nelle caselle interne.

11.2.3 Distribuzioni condizionate

Chiediamoci: tra gli studenti che fanno sport regolarmente, come si distribuisce il sonno? Guardiamo soltanto la terza riga e dividiamo ogni frequenza per il totale della riga: $9 : 60 = 15\%$ dorme meno di 7 ore, $30 : 60 = 50\%$ da 7 a 8, $21 : 60 = 35\%$ più di 8. È la **distribuzione condizionata** del sonno *dato* che lo sport è «regolare». Ogni riga fornisce una distribuzione condizionata del sonno, ogni colonna una distribuzione condizionata dello sport. Per confrontare righe con totali diversi si usano le **frequenze relative di riga**, che su ogni riga sommano a 100%; il confronto più efficace è un grafico a barre raggruppate, come quello della Figura 11.1.

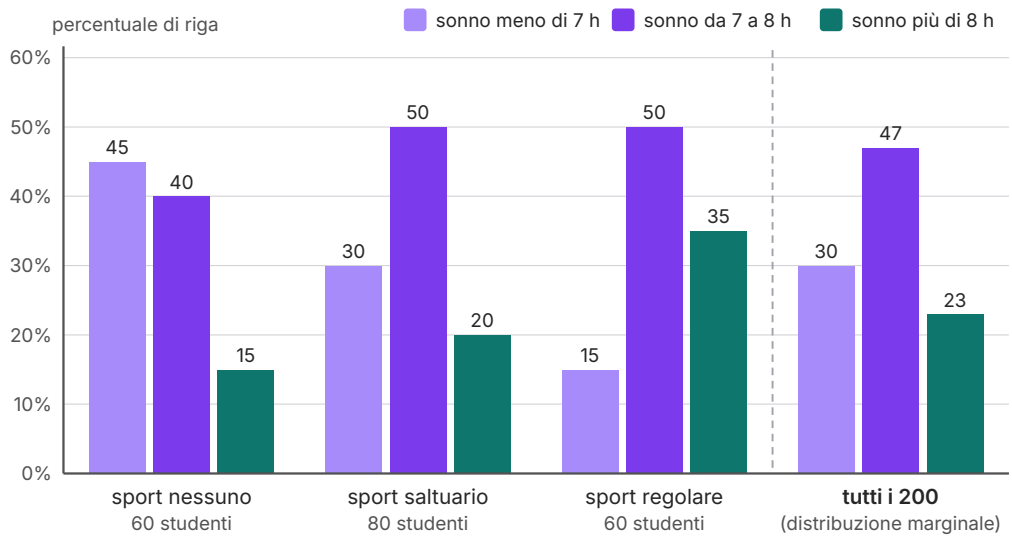

Figura 11.1 Distribuzioni condizionate del sonno per ciascun livello di pratica sportiva (frequenze relative di riga della Tabella 11.1) e, a destra, la distribuzione marginale del sonno su tutti i 200 studenti. Ogni gruppo di tre barre somma a 100%.

Il grafico risponde alla domanda iniziale con un colpo d'occhio. Passando da «nessuno» a «regolare», la quota di chi dorme meno di 7 ore scende dal 45% al 15%, quella di chi dorme più di 8 ore sale dal 15% al 35%: le distribuzioni condizionate sono *diverse* da riga a riga. Si possono anche leggere le frequenze relative di colonna: tra i 46 studenti che dormono più di 8 ore, $21 : 46 \approx 45{,}7\%$ fa sport regolarmente. Attenzione a non confondere le due letture. «Il 35% di chi fa sport regolarmente dorme più di 8 ore» e «il 45,7% di chi dorme più di 8 ore fa sport regolarmente» sono informazioni diverse, ottenute dividendo la stessa frequenza 21 per due totali diversi.

COLLEGAMENTI

Frequenze condizionate e probabilità condizionata. Nel capitolo 10 abbiamo definito $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$, leggendola proprio su una tabella a doppia entrata. La distribuzione condizionata ne è la versione statistica. Se si sceglie a caso uno dei 200 studenti, la probabilità che dorma più di 8 ore *sapendo che* fa sport regolarmente è $\dfrac{21/200}{60/200} = \dfrac{21}{60} = 35\%$, la frequenza relativa condizionata; e l'errore di confondere le due letture della tabella è lo stesso che nel capitolo 10 ci ha messo in guardia contro lo scambio tra $P(A|B)$ e $P(B|A)$. In un campione grande, inoltre, le frequenze condizionate sono buone stime delle probabilità condizionate nella popolazione: è così che un medico stima quanto è frequente un sintomo tra chi ha una certa malattia.

11.3 Dipendenza e indipendenza statistica

11.3.1 Quando due caratteri sono indipendenti

Se il sonno non avesse nulla a che fare con lo sport, ogni riga della tabella avrebbe la stessa distribuzione condizionata, e quindi la stessa distribuzione marginale del sonno: 30% sotto le 7 ore, 47% da 7 a 8, 23% oltre le 8, qualunque sia il livello di sport. Diciamo che due caratteri sono **statisticamente indipendenti** quando le distribuzioni condizionate (in frequenza relativa) di uno dei due sono tutte uguali tra loro; altrimenti sono **dipendenti**, o **associati**. Si dimostra che se sono uguali le distribuzioni condizionate per righe, lo sono anche quelle per colonne: l'indipendenza è una relazione simmetrica (per una tabella 2×2 lo dimostrerai negli esercizi).

Dalla definizione si ricava quale frequenza «dovremmo» trovare in ogni casella se i caratteri fossero indipendenti. Nella prima riga ci sono 60 studenti; se il $\frac{60}{200}$ di loro dormisse meno di 7 ore, come nella marginale, la casella conterrebbe $60 \cdot \frac{60}{200} = 18$ studenti. In generale si chiama **frequenza teorica** della casella nella riga i e nella colonna j il numero

$$f'_{ij} = \frac{(\text{totale della riga } i) \cdot (\text{totale della colonna } j)}{N}$$

Due caratteri sono indipendenti se e solo se ogni frequenza osservata f_{ij} coincide con la frequenza teorica. È la stessa uguaglianza che nel capitolo 10 caratterizza gli eventi indipendenti, $P(A \cap B) = P(A) \cdot P(B)$: dividendo per N i due membri, la frequenza relativa congiunta risulta il prodotto delle due frequenze relative marginali. Le frequenze teoriche possono non essere intere, e non devono esserlo: descrivono una situazione ideale.

ESEMPIO SVOLTO

Frequenze osservate e teoriche. Calcola le frequenze teoriche per i dati della Tabella 11.1 e confrontale con quelle osservate.

Calcolo. Per ogni casella si moltiplica il totale di riga per il totale di colonna e si divide per 200. Per «saltuario, da 7 a 8 h»: $\frac{80 \cdot 94}{200} = 37{,}6$. Le righe «nessuno» e «regolare» hanno lo stesso totale, 60, e quindi le stesse frequenze teoriche.

Tabella 11.2 Frequenze osservate e, tra parentesi, frequenze teoriche per i dati della Tabella 11.1.

Sport \ sonno	meno di 7 h	da 7 a 8 h	più di 8 h
nessuno	27 (18)	24 (28,2)	9 (13,8)
saltuario	24 (24)	40 (37,6)	16 (18,4)
regolare	9 (18)	30 (28,2)	21 (13,8)

Controllo. Su ogni riga e su ogni colonna le frequenze teoriche hanno gli stessi totali di quelle osservate: 18 + 28,2 + 13,8 = 60, 18 + 24 + 18 = 60.

Lettura. Le differenze più grandi sono negli angoli: tra chi non fa sport, chi dorme poco è più frequente del previsto (27 contro 18); tra chi fa sport regolarmente, chi dorme più di 8 ore è più frequente del previsto (21 contro 13,8). La riga «saltuario» è invece quasi uguale alla situazione di indipendenza. I due caratteri sono dipendenti, e lo scarto ha una direzione chiara.

11.3.2 Misurare la distanza dall'indipendenza

Nei dati reali l'indipendenza perfetta non si osserva quasi mai: anche tra due caratteri senza alcun legame il caso produce piccole differenze tra frequenze osservate e teoriche. L'idea, allora, è *misurare* quanto la tabella osservata si allontana dalla tabella teorica, casella per casella, e riassumere le differenze in un solo numero: piccolo se le due tabelle si somigliano, grande se sono molto diverse. Il più usato di questi numeri è l'indice chi-quadrato, presentato nel riquadro.

APPROFONDIMENTO · L'INDICE CHI-QUADRATO

Per ogni casella si calcola la differenza tra frequenza osservata e teorica, la si eleva al quadrato (così i segni non si compensano) e la si divide per la frequenza teorica (così uno scarto di 5 pesa di più dove se ne attendevano 10 che dove se ne attendevano 100). La somma su tutte le caselle è l'**indice chi-quadrato**, dalla lettera greca χ:

$$\chi^2 = \Sigma \, \frac{(f_{ij} - f'_{ij})^2}{f'_{ij}}$$

Per la Tabella 11.2 la prima casella dà $\frac{(27-18)^2}{18} = 4{,}5$, l'ultima $\frac{(21-13{,}8)^2}{13{,}8} \approx 3{,}76$; sommando i nove contributi si ottiene $\chi^2 \approx 15{,}6$. L'indice vale 0 se e solo se la tabella è di perfetta indipendenza. Ha però un difetto: cresce con N. Se tutte le frequenze raddoppiano, le distribuzioni condizionate restano le stesse ma χ^2 raddoppia; per confrontare tabelle diverse lo si divide per una quantità che dipende da N e dalle dimensioni della tabella. Decidere se un certo valore di χ^2 è troppo grande per essere dovuto al caso è un problema di *statistica inferenziale*, affrontato da Karl Pearson nel 1900 e lontano dal nostro programma.

ESEMPIO SVOLTO

Il paradosso di Simpson. Due insegnanti, X e Y, tengono un corso di recupero ciascuno, con 100 studenti; alla fine una verifica dice chi ha recuperato. Gli studenti sono di due tipi: con carenza *lieve* o *grave* (dati inventati, Tabella 11.3). Quale corso funziona meglio?

Tabella 11.3 Esiti di due corsi di recupero (dati inventati): studenti che hanno recuperato, su quelli iscritti.

Corso	carenza lieve	carenza grave	Totale
X	72 su 80 (90%)	10 su 20 (50%)	82 su 100 (82%)
Y	19 su 20 (95%)	48 su 80 (60%)	67 su 100 (67%)

Per gruppi. Tra gli studenti con carenza lieve recupera il 95% con Y e il 90% con X; tra quelli con carenza grave il 60% con Y e il 50% con X. In *entrambi* i gruppi il corso Y ha risultati migliori.

In totale. Sommando i due gruppi, invece, X ha l'82% di successi e Y soltanto il 67%: il confronto si rovescia.

Spiegazione. I due corsi hanno avuto studenti diversi: a X sono toccati soprattutto casi lievi (80 su 100), a Y soprattutto casi gravi. Il totale di X è alto perché i suoi studenti erano più facili da recuperare, non perché il corso fosse migliore. Il terzo carattere, la gravità della carenza, è legato sia al corso sia all'esito, e mescolando i gruppi il suo effetto si sovrappone a quello del corso. È il **paradosso di Simpson**: un'associazione osservata in ogni gruppo può scomparire o invertirsi quando i gruppi vengono riuniti. La morale è che una tabella a doppia entrata va letta chiedendosi sempre se un terzo carattere, non registrato, possa spiegare ciò che si vede.

11.4 Due caratteri quantitativi: la correlazione

11.4.1 Il diagramma a dispersione

Quando entrambi i caratteri sono quantitativi e non vengono divisi in classi, ogni unità statistica è una coppia di numeri $(x_i; y_i)$, cioè un punto del piano cartesiano. L'insieme di questi punti è il **diagramma a dispersione**. In una classe sei studenti hanno misurato la propria altezza x, in centimetri, e il proprio peso y, in chilogrammi (Tabella 11.4, dati inventati); la Figura 11.2 ne mostra il diagramma.

Tabella 11.4 Altezza e peso di sei studenti (dati inventati), con gli scarti dalle medie e i loro prodotti.

Altezza x (cm)	Peso y (kg)	$x - \bar{x}$	$y - \bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
160	55	−10	−7	70	100	49
164	57	−6	−5	30	36	25
168	62	−2	0	0	4	0
172	59	2	−3	−6	4	9
176	67	6	5	30	36	25
180	72	10	10	100	100	100
			Somma	**224**	**280**	**208**

Le medie sono $\bar{x} = \dfrac{1020}{6} = 170$ cm e $\bar{y} = \dfrac{372}{6} = 62$ kg. Il punto $G(\bar{x}; \bar{y}) = G(170; 62)$ si chiama **baricentro** della nuvola, e le rette $x = 170$ e $y = 62$ dividono il piano in quattro quadranti centrati in G. La nuvola sale da sinistra a destra: a un'altezza maggiore corrisponde, *in media*, un peso maggiore, anche se il legame non è rigido (lo studente alto 172 cm pesa meno di quello alto 168). Si dice che tra i due caratteri c'è una **correlazione positiva**; se la nuvola scendesse parleremmo di **correlazione negativa**.

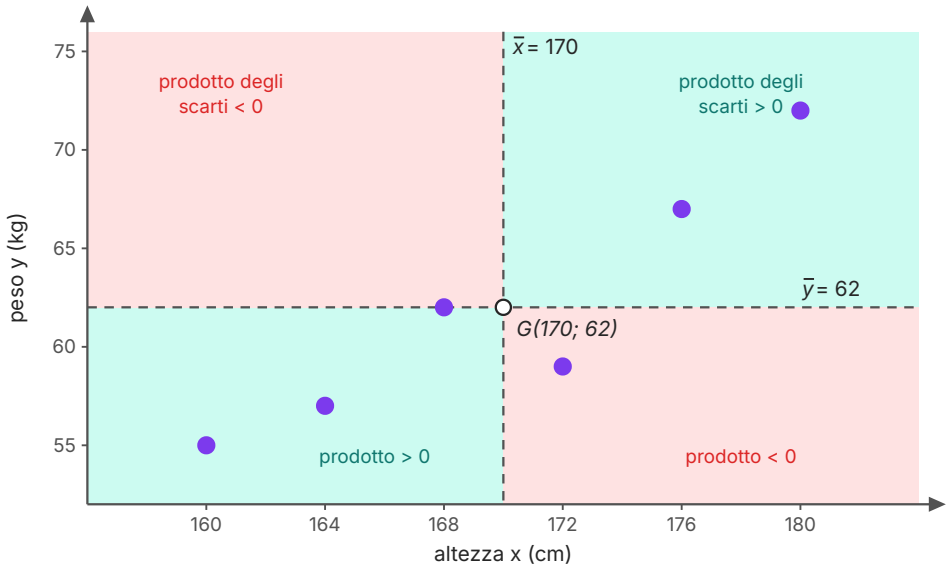

Figura 11.2 Diagramma a dispersione dei dati della Tabella 11.4. Le rette tratteggiate passano per le medie e si incontrano nel baricentro *G*. Nei quadranti verdi il prodotto degli scarti è positivo, in quelli rossi negativo: cinque punti su sei stanno nei quadranti verdi o sul loro bordo.

11.4.2 La covarianza

Per ogni punto consideriamo i due scarti dalle medie e il loro prodotto $(x_i - \bar{x})(y_i - \bar{y})$. In alto a destra e in basso a sinistra i due scarti hanno lo stesso segno e il prodotto è positivo; negli altri due quadranti il prodotto è negativo. Se la nuvola sale prevalgono i prodotti positivi, se scende quelli negativi, se non ha alcuna direzione tendono a compensarsi. La media dei prodotti degli scarti si chiama **covarianza**:

$$\operatorname{cov}(x, y) = \frac{\Sigma\,(x_i - \bar{x})(y_i - \bar{y})}{n}$$

dove n è il numero delle coppie. Per i nostri dati $\operatorname{cov}(x, y) = \frac{224}{6} \approx 37{,}3$. Il segno della covarianza indica il verso del legame; il suo valore, invece, è difficile da interpretare. Si misura in un'unità composta, qui «centimetri per chilogrammi», e cambia con le unità di misura: se esprimiamo l'altezza in metri invece che in centimetri tutti gli scarti delle x diventano 100 volte più piccoli, e così la covarianza, mentre il legame tra i dati è esattamente lo stesso. Per misurare l'*intensità* del legame occorre un numero puro.

11.4.3 Il coefficiente di correlazione lineare

Dividiamo la covarianza per il prodotto delle due deviazioni standard, calcolate anch'esse dividendo per n. Si ottiene il **coefficiente di correlazione lineare**, detto di Bravais–Pearson dal nome di Auguste Bravais, che ne scrisse la formula a metà dell'Ottocento, e di Karl Pearson, che alla fine del secolo ne fece uno strumento della statistica.

IN SINTESI

Coefficiente di correlazione lineare. Per n coppie $(x_i; y_i)$, con le x non tutte uguali e le y non tutte uguali:

$$r = \frac{\operatorname{cov}(x, y)}{\sigma_x \cdot \sigma_y} = \frac{\Sigma\,(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma\,(x_i - x)^2} \cdot \sqrt{\Sigma\,(y_i - y)^2}}$$

- $-1 \le r \le 1$; r è un numero puro e non cambia se si cambiano le unità di misura;
- il **segno** di r è quello della covarianza: $r > 0$ correlazione positiva, $r < 0$ negativa;
- $|r| = 1$ se e solo se tutti i punti stanno su una retta; $|r|$ vicino a 1 indica una nuvola stretta attorno a una retta, r vicino a 0 che nessuna retta la descrive bene.

Nella seconda forma i denominatori n si sono semplificati: per questo conviene usarla nei calcoli a mano, con una tabella come la Tabella 11.4. Che $|r|$ non superi mai 1 si dimostra con la retta di regressione del prossimo paragrafo (è la sfida finale degli esercizi); che valga 1 per punti allineati su una retta crescente lo dimostrerai negli esercizi.

ESEMPIO SVOLTO

Calcolo di r. Calcola il coefficiente di correlazione per i dati della Tabella 11.4.

Somme. Dall'ultima riga della tabella: $\Sigma\,(x_i - \overline{x})(y_i - \overline{y}) = 224$, $\Sigma\,(x_i - \overline{x})^2 = 280$, $\Sigma\,(y_i - \overline{y})^2 = 208$.

Coefficiente. $r = \dfrac{224}{\sqrt{280} \cdot \sqrt{208}} = \dfrac{224}{\sqrt{58\,240}} \approx \dfrac{224}{241{,}3} \approx 0{,}93$.

Deviazioni standard (servono più avanti). $\sigma_x = \sqrt{\dfrac{280}{6}} \approx 6{,}83$ cm, $\sigma_y = \sqrt{\dfrac{208}{6}} \approx 5{,}89$ kg; e infatti $\dfrac{37{,}33}{6{,}83 \cdot 5{,}89} \approx 0{,}93$.

Interpretazione. r è positivo e vicino a 1: nel gruppo c'è una correlazione lineare positiva forte tra altezza e peso. Sei dati, però, sono pochi: un settimo studente insolito potrebbe cambiare parecchio il risultato.

La Figura 11.3 mette a confronto tre nuvole del capitolo. La prima è quella di altezza e peso. La seconda raccoglie, per dieci studenti, le ore passate davanti a uno schermo dopo cena (0,5; 1; 1;

1,5; 2; 2; 2,5; 3; 3; 3,5) e le ore di sonno di quella notte (8,5; 7,5; 8; 7,5; 7; 8,5; 7,5; 7; 7,5; 7): la nuvola scende ma è larga, e $r \approx -0{,}64$. La terza è il caso più istruttivo. In laboratorio un carrellino a molla lancia una pallina con angoli di 15°, 25°, …, 75° e si misura la gittata (2,4; 3,9; 4,7; 5,0; 4,6; 3,8; 2,6 metri): i punti seguono da vicino una curva a forma di parabola rovesciata, il legame tra angolo e gittata è fortissimo, eppure $r \approx 0{,}02$. A sinistra del massimo i punti salgono, a destra scendono, e i prodotti degli scarti si compensano.

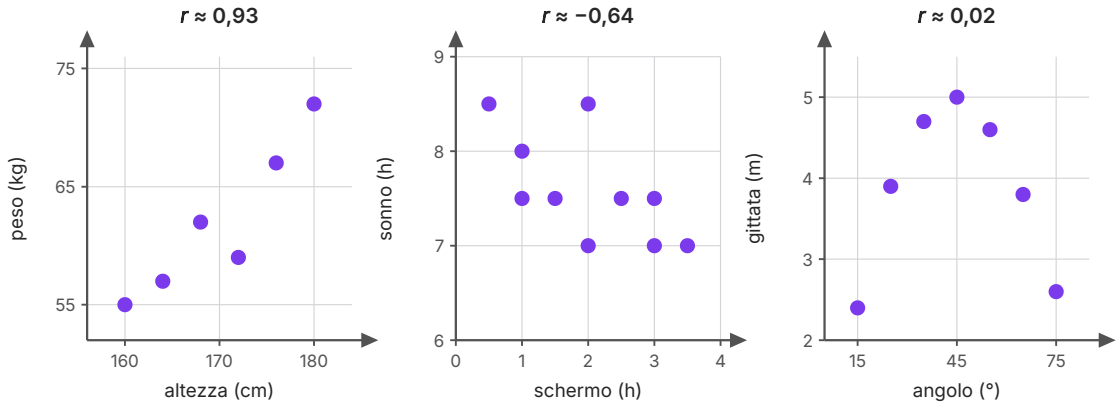

Figura 11.3 Tre diagrammi a dispersione con il loro coefficiente di correlazione, calcolato sui punti disegnati: altezza e peso (Tabella 11.4), ore di schermo e ore di sonno, angolo di lancio e gittata (dati inventati). Nel terzo i punti seguono una curva eppure $r \approx 0$: il coefficiente misura soltanto il legame *lineare*.

11.4.4 Correlazione non è causalità

Un coefficiente di correlazione alto dice che due grandezze variano insieme, non che una sia la *causa* dell'altra. Le possibilità sono almeno tre. Può darsi che x influisca davvero su y (lo schermo acceso ritarda il sonno). Può darsi che sia y a influire su x (chi non ha sonno resta più a lungo al telefono). Oppure entrambe possono dipendere da una terza grandezza che non abbiamo misurato: si parla allora di **correlazione spuria**, e la terza grandezza si chiama **variabile confondente**. Il corso di recupero del paradosso di Simpson ne era un esempio: la gravità della carenza confondeva il confronto tra i corsi.

ERRORE FREQUENTE

«Sono correlati, quindi uno causa l'altro». Registrando, giorno per giorno durante un'estate, i gelati venduti in una località di mare e il numero di persone medicate per scottature, si troverebbe con ogni probabilità una correlazione positiva forte. Nessuno penserebbe che i gelati provochino le scottature: entrambe le grandezze crescono nelle giornate calde e affollate, e la temperatura è la variabile confondente. Molte correlazioni meno evidenti hanno la stessa struttura: tra i comuni di una regione, il numero di farmacie e il numero di nascite in un anno risulterebbero fortemente correlati, perché entrambi crescono con il numero degli abitanti. Per stabilire una relazione di causa servono esperimenti controllati, in cui si fa variare un solo fattore lasciando fermi gli altri.

«$r = 0$, quindi non c'è relazione». Anche questo è sbagliato: $r = 0$ esclude soltanto un legame *lineare*, come mostra il terzo diagramma della Figura 11.3. Prima di calcolare r si disegna *sempre* il diagramma a dispersione.

COLLEGAMENTI

Filosofia: Hume e l'idea di causa. Nella *Ricerca sull'intelletto umano* (1748) David Hume, che incontri nel programma di filosofia di quest'anno, si chiede da dove venga la nostra idea di causa. Quando diciamo che il fuoco *causa* il calore, osserva, non vediamo alcuna «connessione necessaria» tra i due fatti: vediamo soltanto che il secondo segue il primo, e che lo ha sempre seguito in passato. È l'abitudine, nata da questa «congiunzione costante», a farci attendere il calore quando vediamo il fuoco. In termini di questo capitolo, Hume sostiene che anche le nostre certezze causali nascono da qualcosa di molto simile a una correlazione osservata ripetutamente.

La statistica raccoglie la lezione in due modi. Da un lato ricorda che una correlazione, per quanto forte, non dimostra un nesso causale: tra le possibili spiegazioni ci sono la causa inversa e la variabile confondente. Dall'altro, per avvicinarsi a una causa, propone di *intervenire* invece che soltanto osservare: assegnare a caso i trattamenti a due gruppi, così che ogni terzo fattore si distribuisca in modo simile in entrambi. È l'idea dell'esperimento controllato, che nel paradosso di Simpson avrebbe evitato di dare a un corso tutti i casi facili.

11.5 La retta di regressione

11.5.1 Il metodo dei minimi quadrati

Se la nuvola di punti è allungata lungo una retta, è naturale cercare la retta $y = mx + q$ che la rappresenta meglio, per riassumere i dati e per fare previsioni. Ma che cosa vuol dire «meglio»? Per ogni punto confrontiamo l'ordinata osservata y_i con quella prevista dalla retta, $mx_i + q$. La differenza

$$e_i = y_i - (mx_i + q)$$

si chiama **residuo** del punto: è il segmento verticale tra il punto e la retta, positivo se il punto sta sopra, negativo se sta sotto. Una buona retta ha residui piccoli. Sommarli non basta, perché positivi e negativi si compenserebbero; come per la varianza, li eleviamo al quadrato. Il **metodo dei minimi quadrati**, proposto all'inizio dell'Ottocento da Legendre e da Gauss per calcolare le orbite di comete e pianeti, sceglie la retta che rende *minima la somma dei quadrati dei residui*, $S = \Sigma\,(y_i - mx_i - q)^2$. La retta così trovata si chiama **retta di regressione** di y su x.

Una prima proprietà si dimostra con il vertice della parabola (*Matematica 3*, cap. 2). Fissiamo una pendenza m qualsiasi e chiamiamo $u_i = y_i - mx_i$. Allora $S = \Sigma\,(u_i - q)^2 = nq^2 - 2q\,\Sigma\,u_i + \Sigma\,u_i^2$ è una funzione di secondo grado di q, con coefficiente di q^2 uguale a $n > 0$: ha un minimo nel vertice, cioè per $q = \dfrac{2\,\Sigma\,u_i}{2n} = \overline{u}$, la media degli u_i. Poiché $\overline{u} = \overline{y} - m\overline{x}$, il valore migliore è $q = \overline{y} - m\overline{x}$, cioè $\overline{y} = m\overline{x} + q$: **la retta dei minimi quadrati passa per il baricentro $G(\overline{x}; \overline{y})$**, qualunque sia la sua pendenza. Ne segue che i residui hanno media zero, e quindi somma zero: i punti sopra e i punti sotto la retta si bilanciano esattamente. c.v.d.

Resta da scegliere la pendenza tra le rette per G. Con un ragionamento dello stesso tipo, che proponiamo come sfida guidata alla fine degli esercizi, si trova il risultato riassunto nel riquadro.

IN SINTESI

Teorema (retta di regressione). Ipotesi: n coppie $(x_i; y_i)$, con le x non tutte uguali. **Tesi:** la retta che rende minima la somma dei quadrati dei residui è $y = mx + q$ con

$$m = \frac{\operatorname{cov}(x, y)}{\sigma_x^2} = \frac{\Sigma\,(x_i - \overline{x})(y_i - \overline{y})}{\Sigma\,(x_i - x)^2}, \qquad q = \overline{y} - m\overline{x}$$

Essa passa per $G(\overline{x}; \overline{y})$ e la somma dei suoi residui è 0. Poiché $r = \dfrac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}$, si ha anche $m = r \cdot \dfrac{\sigma_y}{\sigma_x}$: la pendenza ha lo stesso segno di r.

ESEMPIO SVOLTO

Retta di regressione e residui. Per i dati della Tabella 11.4 trova la retta di regressione del peso sull'altezza, calcola i residui e prevedi il peso di uno studente alto 174 cm.

Coefficienti. $m = \frac{224}{280} = 0{,}8$ (gli n si semplificano) e $q = 62 - 0{,}8 \cdot 170 = 62 - 136 = -74$. La retta è $y = 0{,}8x - 74$: nel gruppo, ogni centimetro in più corrisponde in media a 0,8 kg in più.

Residui. I pesi previsti per $x = 160, 164, 168, 172, 176, 180$ sono 54; 57,2; 60,4; 63,6; 66,8; 70. I residui, osservato meno previsto, sono 1; −0,2; 1,6; −4,6; 0,2; 2 (Figura 11.4). La loro somma è 0, come deve essere. Il residuo più grande, −4,6, è dello studente alto 172 cm, molto più leggero di quanto la retta prevede.

Previsione. Per $x = 174$: $y = 0{,}8 \cdot 174 - 74 = 65{,}2$ kg. È un peso *medio atteso*, non certo: i residui mostrano scostamenti fino a quasi 5 kg.

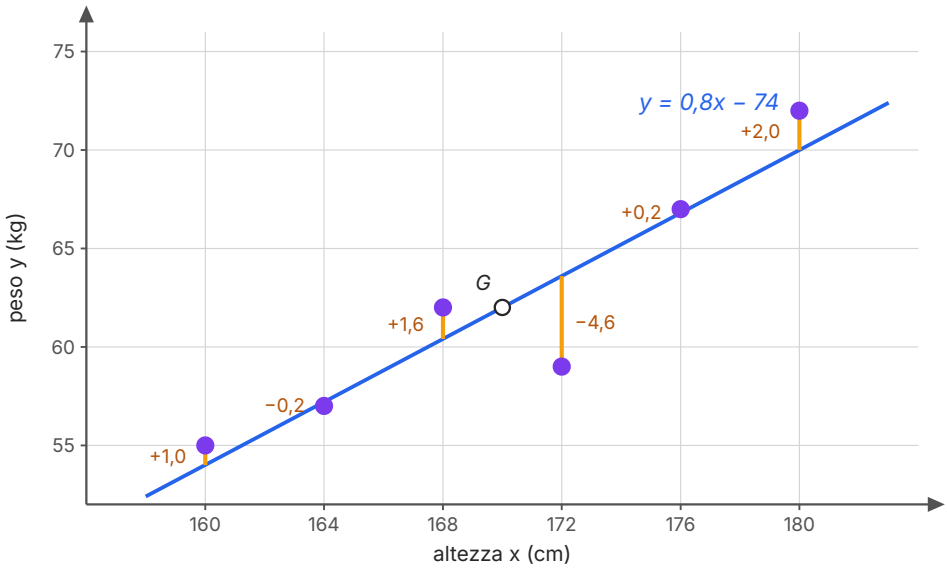

Figura 11.4 La retta di regressione $y = 0{,}8x - 74$ dei dati della Tabella 11.4, con i residui (segmenti arancioni, in kg). La retta passa per il baricentro $G(170; 62)$; la somma dei residui è zero.

11.5.2 Interpolazione ed estrapolazione

Usare la retta per prevedere y in corrispondenza di un valore di x *interno* all'intervallo dei dati osservati si chiama **interpolazione**; usarla per un valore *esterno* si chiama **estrapolazione**. L'estrapolazione è rischiosa: la retta descrive i dati dove li abbiamo osservati, e nulla garantisce che il legame resti lineare altrove. La nostra retta, per un bambino alto 100 cm, prevede

$0{,}8 \cdot 100 - 74 = 6$ kg, un peso assurdo; e la sua ordinata all'origine, $q = -74$, sarebbe il «peso» di una persona alta zero centimetri. Il termine q non ha un significato fisico: serve soltanto a far passare la retta per G nella zona dei dati.

Due osservazioni chiudono il paragrafo. La prima: la retta di regressione di y su x rende minimi i residui *verticali*, perché tratta x come il dato di partenza e y come la grandezza da prevedere. Se volessimo prevedere l'altezza conoscendo il peso dovremmo rendere minimi i residui orizzontali, e otterremmo la **retta di regressione di x su y**, $x = m'y + q'$ con $m' = \dfrac{\text{cov}(x, y)}{\sigma_y^2}$: per i nostri dati $m' = \dfrac{224}{208} \approx 1{,}08$ e la retta è $x \approx 1{,}08y + 103{,}2$, diversa dalla prima. Le due rette coincidono soltanto se $|r| = 1$. La seconda: i residui sono anche un controllo. Se una retta è adatta ai dati, i residui sono piccoli e sparsi a caso, senza una forma riconoscibile; se hanno un andamento regolare, per esempio positivi agli estremi e negativi al centro, la nuvola curva e la retta non lo vede. Fu proprio studiando stature di genitori e figli che Francis Galton, nel 1886, introdusse la parola «regressione»: i figli di genitori molto alti risultavano in media alti, ma meno dei genitori, «regredendo» verso la media.

11.6 Campione e modello

11.6.1 Popolazione e campione

Nel primo volume abbiamo distinto la **popolazione**, l'insieme di tutte le unità che interessano, dal campione, la parte che effettivamente si osserva. Studiare tutta la popolazione è spesso impossibile o troppo costoso: non si intervistano tutti gli elettori prima di un'elezione, non si collaudano fino alla rottura tutte le lampadine prodotte. Un numero che descrive la popolazione (la sua media, la percentuale di chi ha una certa opinione) si chiama **parametro**; lo stesso indice calcolato sul campione è una **stima** del parametro. Il parametro è fisso ma in genere sconosciuto; la stima è nota, ma cambia da un campione all'altro.

Perché la stima sia attendibile il campione deve essere **rappresentativo**, cioè riprodurre in piccolo le caratteristiche della popolazione, e il modo più semplice per ottenerlo è affidarsi al caso. Un **campione casuale semplice** di n unità è scelto in modo che ogni gruppo di n unità della popolazione abbia la stessa probabilità di essere estratto: per esempio numerando tutti gli studenti della scuola ed estraendo n numeri con un generatore di numeri casuali. Un campione scelto in modo da favorire sistematicamente alcune unità produce invece stime sbagliate sempre nella stessa direzione, qualunque sia la sua ampiezza: si dice che ha una **distorsione**.

Le distorsioni più comuni sono tre. Il **campione di comodo**: chiedere quanto si dorme soltanto agli studenti che alle 7.45 sono già in biblioteca. L'**autoselezione**: nei sondaggi online risponde soltanto chi ne ha voglia, spesso chi ha un'opinione più forte. La **mancata risposta**: se una parte degli estratti non risponde, e chi non risponde è diverso da chi risponde, il campione si deforma anche se era stato estratto bene. Il caso più citato è il sondaggio della rivista ameri-

cana *The Literary Digest* per le elezioni presidenziali del 1936. La rivista spedì circa dieci milioni di schede e ne ricevette indietro circa 2,4 milioni, un campione enorme; prevedeva una netta vittoria dello sfidante repubblicano Alf Landon, ma vinse il presidente Franklin D. Roosevelt, con circa il 61% dei voti. Gli indirizzi venivano in buona parte da elenchi telefonici e da registri di proprietari di automobili, più diffusi tra gli elettori benestanti, e secondo gli studi successivi risposero più volentieri gli elettori contrari a Roosevelt: un campione grande ma distorto sbaglia con grande sicurezza.

Ora possiamo spiegare il denominatore $n - 1$ del primo paragrafo. Il calcolo fatto per q mostra che la somma $\Sigma\,(x_i - c)^2$, come funzione di c, è minima per $c = \overline{x}$. Quindi gli scarti dalla media *del campione* hanno una somma dei quadrati più piccola degli scarti dalla media vera della popolazione, che non conosciamo: dividendo per n sottostimeremmo, in media, la variabilità. Si dimostra (non lo faremo) che dividere per $n - 1$ corregge esattamente questo difetto.

11.6.2 Campioni diversi, stime diverse

Anche un campione casuale senza distorsioni non dà esattamente il parametro. Per vedere quanto varia la stima serve una popolazione di cui il parametro si conosce, e un computer. Prendiamo i lanci di un dado regolare: la media della popolazione, cioè il valore atteso di un lancio (*Matematica 2*, cap. 8), è 3,5. Un foglio di calcolo simula 1000 campioni di 5 lanci e ne calcola la media di ciascuno; poi fa lo stesso con 1000 campioni di 20 lanci. La Figura 11.5 mostra gli istogrammi delle 2000 medie ottenute.

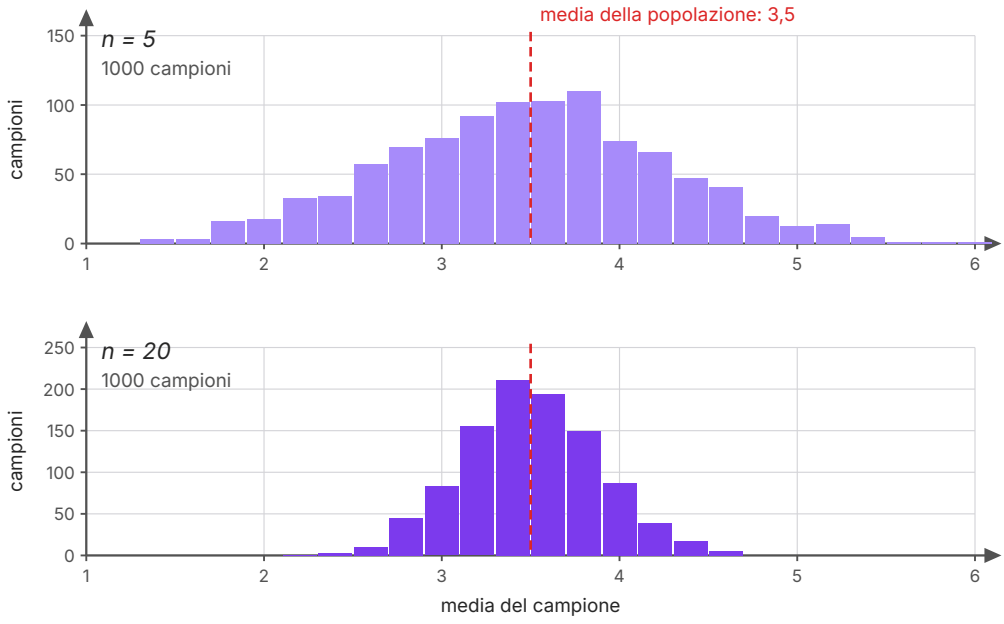

Figura 11.5 Istogrammi delle medie di 1000 campioni simulati di 5 lanci di un dado (sopra) e di 1000 campioni di 20 lanci (sotto), con classi di ampiezza 0,2. La linea rossa è la media della popolazione, 3,5. Le medie dei campioni più grandi sono molto meno disperse (dati simulati).

Le medie dei campioni si raccolgono attorno a 3,5 in entrambi i casi: la media di un campione casuale non sbaglia «in una direzione». Ma la loro dispersione è molto diversa. Con 5 lanci le medie vanno da 1,4 a 6, solo il 56% cade tra 3 e 4 e la loro deviazione standard è circa 0,78; con 20 lanci vanno da 2,25 a 4,6, l'83% cade tra 3 e 4 e la deviazione standard è circa 0,38. **Campioni più grandi danno stime meno variabili**; la simulazione suggerisce anche che, quadruplicando l'ampiezza, la dispersione si dimezza circa, ma i guadagni sono lenti. Per questo i sondaggi accompagnano la stima con un *margine d'errore*, che misura soltanto l'errore dovuto al caso, non quello dovuto alla distorsione: il *Literary Digest* aveva un campione grandissimo, e quindi una stima pochissimo variabile, ma centrata sul valore sbagliato.

COLLEGAMENTI

Storia: censimenti e campioni. Un *censimento* osserva l'intera popolazione: a Roma i censori registravano periodicamente i cittadini e i loro beni, e le *Res gestae* di Augusto ricordano i censimenti dei cittadini romani che l'imperatore fece eseguire. Il primo censimento del Regno d'Italia si tenne il 31 dicembre 1861. Un censimento risponde con esattezza, ma è lento e costoso; per questo, tra un censimento e l'altro, gli istituti di statistica stimano molti parametri della popolazione con indagini su campioni casuali, molto più piccoli, scelti con le cautele di questo paragrafo.

11.6.3 Il modello statistico

Una retta di regressione è un esempio di modello matematico: una funzione scelta per descrivere un fenomeno, con parametri ricavati dai dati. Un modello non è «vero» o «falso»: è più o meno adatto a uno scopo, in un certo intervallo di validità, e si giudica dai residui. Le *Indicazioni nazionali* chiedono di avere chiara la differenza tra la matematizzazione della fisica classica, che cerca la legge corrispondente in modo univoco al fenomeno, e la **modellistica**, in cui la stessa classe di fenomeni può essere rappresentata con approcci diversi. Il fisico che appende masse a una molla sa in anticipo che l'allungamento deve essere proporzionale alla forza (legge di Hooke): usa la retta dei minimi quadrati per stimare la costante elastica e i residui per controllare che la legge valga nell'intervallo delle misure. Lo statistico che studia altezza e peso non ha una legge del genere: la retta è soltanto la descrizione più semplice dei dati di quel gruppo, e un altro gruppo, o un intervallo di altezze più ampio, potrebbe richiedere una curva diversa, come una crescita esponenziale (capitolo 7) per una popolazione di cellule.

Questo è anche il modo migliore di fare statistica in classe: raccogliere dati propri (altezza e apertura delle braccia, ore di sonno, tempi di percorrenza), metterli in una tabella o in un diagramma a dispersione, calcolare gli indici, e poi chiedersi che cosa dicono e che cosa *non* dicono: se il gruppo è un campione rappresentativo di qualcosa, se una variabile confondente può spiegare il legame, fino a dove una previsione è ragionevole.

SINTESI

- La deviazione standard σ si calcola dividendo per N quando i dati sono la popolazione; s si calcola dividendo per $n - 1$ quando sono un campione. Il valore standardizzato $z = \dfrac{x - \bar{x}}{\sigma}$ dice di quante deviazioni standard un dato dista dalla media.
- Una tabella a doppia entrata raccoglie le frequenze congiunte di due caratteri; i totali di riga e di colonna sono le distribuzioni marginali.
- Una distribuzione condizionata è la distribuzione di un carattere in una sola riga (o colonna), di solito in frequenze relative; è la versione statistica della probabilità condizionata.
- Due caratteri sono indipendenti se le distribuzioni condizionate sono tutte uguali, cioè se ogni frequenza osservata è uguale alla frequenza teorica $f' = \dfrac{(\text{totale di riga}) \cdot (\text{totale di colonna})}{N}$.
- Un terzo carattere può alterare o rovesciare un'associazione (paradosso di Simpson).
- Il diagramma a dispersione rappresenta coppie di dati quantitativi; la covarianza, media dei prodotti degli scarti, ne indica il verso.
- Il coefficiente $r = \dfrac{\text{cov}(x, y)}{\sigma_x \sigma_y}$ è compreso tra −1 e 1, vale ±1 solo per punti allineati e misura soltanto il legame lineare: $r \approx 0$ non esclude un legame curvo.
- Correlazione non è causalità: un legame può nascere da una variabile confondente (correlazione spuria).
- La retta di regressione di y su x ha $m = \dfrac{\text{cov}(x, y)}{\sigma_x^2}$ e $q = \bar{y} - m\bar{x}$; passa per il baricentro e i suoi residui hanno somma zero. La regressione di x su y è un'altra retta.
- Le previsioni sono ragionevoli dentro l'intervallo dei dati (interpolazione), rischiose fuori (estrapolazione).
- Un campione casuale semplice evita le distorsioni; un campione grande ma distorto sbaglia comunque. Campioni più grandi danno stime meno variabili.
- Un modello statistico si sceglie e si controlla con i residui; la stessa classe di fenomeni può avere modelli diversi (modellistica).

ESERCIZI

Comprendere

1. ● In una tabella a doppia entrata la distribuzione marginale di un carattere si legge:
   A. nelle caselle interne della tabella
   B. nei totali di riga o di colonna
   C. sulla diagonale della tabella
   D. nella sola casella del totale generale

2. ● Sei punti stanno tutti sulla retta $y = -2x + 5$. Il loro coefficiente di correlazione è:
   A. $r = -2$
   B. $r = 0$
   C. $r = -1$
   D. $r = 5$

3. ● Due caratteri qualitativi sono statisticamente indipendenti quando:
   A. tutte le frequenze congiunte sono diverse da zero
   B. le due distribuzioni marginali sono uguali
   C. il totale generale è 100
   D. le distribuzioni condizionate (in frequenza relativa) di un carattere sono tutte uguali

4. ● La retta di regressione di y su x passa sempre per:
   A. il baricentro $(\overline{x}; \overline{y})$
   B. l'origine degli assi
   C. il primo e l'ultimo punto dei dati
   D. il punto con il residuo più grande

5. ●● Un sito propone ai visitatori un questionario sulle ore di sonno; rispondono 3000 persone. Il campione ottenuto:
   A. è rappresentativo perché è molto grande
   B. è un campione casuale semplice
   C. dà una stima esatta, perché supera le 1000 unità
   D. è affetto da autoselezione

6. ● Vero o falso? Correggi le affermazioni false.
   a) Se $r = 0$ tra x e y non c'è alcuna relazione.
   b) Per gli stessi dati, s è sempre minore di σ.
   c) Se $r < 0$, al crescere di x i valori di y tendono a diminuire.
   d) La retta di regressione di x su y coincide sempre con quella di y su x.

7. ● Completa.
   a) In una tabella con totale N, la frequenza teorica di una casella è $f' =$ _____ .
   b) Il residuo di un punto $(x_i; y_i)$ rispetto alla retta $y = mx + q$ è $e_i =$ _____ .
   c) La covarianza è la _____ dei prodotti degli scarti dalle medie.
   d) Prevedere y per un valore di x esterno all'intervallo dei dati si chiama _____ .

8. ● Con i dati della Tabella 11.1, indica se ciascuna percentuale è una frequenza relativa marginale, condizionata per riga o condizionata per colonna, e calcolala.
   a) La percentuale di studenti che non fanno sport.
   b) La percentuale di chi dorme meno di 7 ore tra chi fa sport in modo saltuario.
   c) La percentuale di chi non fa sport tra chi dorme meno di 7 ore.

Calcolare

9. ● In una verifica di greco la classe ha media 6,2 e $\sigma = 1{,}2$; in una gara di corsa, sulla stessa classe, il tempo medio è 70 s con $\sigma = 8$ s. a) Calcola il valore standardizzato di un 8 in greco e di un 5. b) Marco ha preso 8 in greco e ha corso in 58 s: in quale delle due prove si è distinto di più rispetto alla classe? (Attenzione: nella corsa un tempo minore è un risultato migliore.)

10. ● Per i dati 4, 7, 7, 8, 9 calcola la media, la deviazione standard σ (trattandoli come una popolazione) e la deviazione standard campionaria s (trattandoli come un campione).

11. ● A 120 studenti si chiede il genere di lettura preferito (dati inventati). Nel biennio: romanzi 30, fumetti 20, saggi 10; nel triennio: romanzi 24, fumetti 6, saggi 30. a) Costruisci la tabella a doppia entrata con i totali. b) Scrivi la distribuzione marginale del genere, in percentuale. c) Scrivi le distribuzioni condizionate del genere nel biennio e nel triennio. d) I due caratteri sono indipendenti?

12. ●● Per la tabella dell'esercizio precedente calcola le frequenze teoriche e indica in quali caselle le frequenze osservate si discostano di più da quelle teoriche.

13. ●● In una tabella 2×2 i totali di riga sono 40 e 60, quelli di colonna 30 e 70. Scrivi le frequenze congiunte che renderebbero i due caratteri indipendenti, e verifica che le distribuzioni condizionate per riga sono uguali.

14. ●● Per le cinque coppie (1; 2), (2; 4), (3; 5), (4; 4), (5; 5) calcola le medie, la covarianza e il coefficiente di correlazione.

15. ●● Per le coppie dell'esercizio precedente a) trova la retta di regressione di y su x; b) calcola i residui e verifica che la loro somma è zero; c) prevedi y per $x = 3{,}5$.

16. ●● Di un gruppo di dati si sa che $\overline{x} = 12$, $\overline{y} = 30$, $\sigma_x = 2$, $\sigma_y = 5$, $r = 0{,}8$. a) Scrivi la retta di regressione di y su x. b) Scrivi la retta di regressione di x su y, usando $m' = r \cdot \dfrac{\sigma_x}{\sigma_y}$. c) Verifica che le due rette passano entrambe per (12; 30) ma sono diverse.

17. ●● Si riscrivono le altezze della Tabella 11.4 in metri invece che in centimetri (1,60; 1,64; ...). Come cambiano a) la covarianza, b) il coefficiente r, c) la pendenza della retta di regressione del peso sull'altezza? Calcola i nuovi valori.

18. ●● Con i dati delle ore di schermo e di sonno di dieci studenti (elencati prima della Figura 11.3) verifica che $\bar{x} = 2$ e $\bar{y} = 7{,}6$, calcola r e la retta di regressione del sonno sulle ore di schermo. Quanto sonno prevede la retta per 1,5 ore di schermo?

Dimostrare e argomentare

19. ●● a) Dimostra che la somma degli scarti dalla media è zero: $\Sigma\,(x_i - \bar{x}) = 0$. b) Usando $q = \bar{y} - m\bar{x}$, dimostra che anche la somma dei residui della retta di regressione è zero.

20. ●● Dimostra che, se tutti i punti stanno sulla retta $y = ax + b$ con $a > 0$, allora $r = 1$. (Suggerimento: esprimi gli scarti $y_i - \bar{y}$ per mezzo degli scarti $x_i - \bar{x}$.) Che cosa cambia se $a < 0$?

21. ●● Trova l'errore: «In 30 comuni il numero di chiese è fortemente correlato con il numero di furti denunciati in un anno, con $r = 0{,}9$. Dunque le chiese attirano i ladri.» Indica la variabile confondente più probabile e un modo per tenerne conto.

22. ●●● In una tabella 2×2 con frequenze a, b sulla prima riga e c, d sulla seconda (tutte positive), dimostra che le due distribuzioni condizionate per riga sono uguali se e solo se $ad = bc$. Deduci che allora sono uguali anche le distribuzioni condizionate per colonna.

Problemi

23. ●● **Biologia.** In una classe si misurano altezza x e apertura delle braccia y di cinque studenti, in centimetri: (150; 148), (158; 160), (165; 163), (170; 172), (177; 177). a) Calcola r. b) Trova la retta di regressione di y su x. c) Prevedi l'apertura delle braccia di uno studente alto 162 cm.

24. ●● **Fisica.** A una molla si appendono masse di 100, 200, 300, 400 e 500 g e si misurano gli allungamenti 2,1; 3,9; 6,2; 7,9; 9,9 cm. a) Trova la retta di regressione dell'allungamento sulla massa. b) Quale allungamento prevede per 350 g? c) La retta prevede circa 98 cm per una massa di 5 kg: perché questa previsione non è affidabile?

25. ●● **Trasporti.** In un mese due linee di autobus fanno 200 corse ciascuna (dati inventati). Linea 1: nelle ore di punta 40 corse, di cui 28 in orario; fuori dalle ore di punta 160 corse, di cui 144 in orario. Linea 2: nelle ore di punta 150 corse, di cui 111 in orario; fuori 50, di cui 46 in orario. a) Calcola le percentuali di puntualità di ciascuna linea nei due gruppi e in totale. b) Quale linea è più puntuale? Spiega il paradosso.

26. ●● **Un'indagine a scuola.** Un liceo di 900 studenti vuole stimare le ore di sonno medie. a) Spiega quale distorsione avrebbero un campione degli studenti presenti in biblioteca alle 7.45 e un modulo online facoltativo, e come scegliere invece un campione casuale semplice di 60 studenti. b) Un piccolo campione di 8 studenti dà 7; 6,5; 8; 7,5; 6; 7; 8,5; 7,5 ore. Calcola la media e la deviazione standard campionaria s.

Con il computer

27. ●● Con un foglio di calcolo, simula il lancio di un dado con la funzione che genera un intero casuale tra 1 e 6 (CASUALE.TRA(1;6) nella versione italiana dei fogli di calcolo più diffusi). a) Riempi 500 righe con 20 lanci ciascuna. b) Per ogni riga calcola la media dei primi 5 lanci e la media di tutti i 20. c) Calcola la media e la deviazione standard delle due colonne di medie e disegna i loro istogrammi, con classi di ampiezza 0,2. d) Confronta i risultati con quelli della Figura 11.5.

28. ●● Inserisci in un foglio di calcolo i dati di altezza e peso della Tabella 11.4. a) Usa le funzioni del foglio per la correlazione, la pendenza e l'intercetta (CORRELAZIONE, PENDENZA, INTERCETTA) e verifica i valori dell'esempio svolto sulla retta di regressione. b) Sostituisci il peso dello studente alto 172 cm, 59 kg, con 75 kg: come cambiano r e la pendenza? c) Che cosa insegna il punto b) sull'effetto di un solo dato insolito in un gruppo piccolo?

Verso la verifica

29. ●● A 150 studenti si chiede se suonano uno strumento e se la loro media dei voti è almeno 8 (dati inventati). Suonano uno strumento in 50, e di questi 20 hanno media almeno 8; non suonano in 100, e di questi 25 hanno media almeno 8. a) Costruisci la tabella a doppia entrata. b) Calcola le distribuzioni condizionate della media per chi suona e per chi non suona. c) Calcola le frequenze teoriche e stabilisci se i caratteri sono indipendenti. d) Si può concludere che suonare fa migliorare i voti?

30. ●● In sei giornate d'inverno si registrano la temperatura media esterna x (°C) e il consumo di gas per il riscaldamento di una casa y (m^3): (2; 30), (4; 27), (6; 25), (8; 20), (10; 18), (12; 12) (dati inventati). a) Disegna il diagramma a dispersione. b) Calcola r e la retta di regressione di y su x. c) Prevedi il consumo per $x = 9$ °C. d) Che cosa prevede la retta per $x = 25$ °C? Commenta.

31. ●●● **Sfida.** Siano $m = \dfrac{\text{cov}(x, y)}{\sigma_x^{\,2}}$ e $m' = \dfrac{\text{cov}(x, y)}{\sigma_y^{\,2}}$ le pendenze delle rette di regressione di y su x e di x su y. a) Dimostra che $m \cdot m' = r^2$. b) Per le rette per G di pendenza k, posto $X_i = x_i - \overline{x}$ e $Y_i = y_i - \overline{y}$, la somma dei quadrati dei residui è $S(k) = k^2 \Sigma\, X_i^2 - 2k \Sigma\, X_i Y_i + \Sigma\, Y_i^2$: trova il k che la rende minima e verifica che è m. c) Dimostra che il valore minimo è $\Sigma\, Y_i^2 \cdot (1 - r^2)$ e deduci che $-1 \le r \le 1$. d) Verifica a) con i dati della Tabella 11.4.

PER APPROFONDIRE

- [HUF] Huff, *Mentire con le statistiche*: campioni distorti, medie ingannevoli e correlazioni scambiate per cause, raccontati con molti esempi.
- [BOY] Boyer, *Storia della matematica*: Legendre, Gauss e il metodo dei minimi quadrati nel quadro della matematica dell'Ottocento.
- [MIUR] *Indicazioni nazionali per i licei*, Allegato C: le «distribuzioni doppie condizionate e marginali», la correlazione, la regressione e il campione nel programma del secondo biennio, e il concetto di modello matematico.

Glossario del capitolo

Campione casuale semplice
Campione scelto in modo che ogni gruppo di n unità della popolazione abbia la stessa probabilità di essere estratto.

Coefficiente di correlazione lineare
$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$, compreso tra −1 e 1: misura quanto i punti sono vicini a una retta.

Correlazione spuria
Correlazione tra due grandezze dovuta a una terza grandezza (variabile confondente) che influisce su entrambe.

Covarianza
$\text{cov}(x, y) = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{n}$: media dei prodotti degli scarti; il segno indica il verso del legame.

Deviazione standard campionaria
$s = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n - 1}}$: stima della variabilità di una popolazione a partire da un campione di n dati.

Diagramma a dispersione
Insieme dei punti $(x_i; y_i)$ che rappresentano coppie di dati quantitativi.

Distorsione
Errore sistematico di un campione che favorisce alcune unità; non diminuisce aumentando l'ampiezza del campione.

Distribuzione condizionata
Distribuzione di un carattere tra le sole unità che hanno una data modalità dell'altro: una riga o una colonna della tabella.

Distribuzione marginale
Distribuzione di un solo carattere, letta nei totali di riga o di colonna di una tabella a doppia entrata.

Estrapolazione
Uso di un modello per prevedere fuori dall'intervallo dei dati osservati; il contrario è l'interpolazione.

Frequenza congiunta
Numero delle unità che presentano insieme una modalità del primo carattere e una del secondo.

Frequenza teorica
$f' = \frac{\text{(totale di riga)} \cdot \text{(totale di colonna)}}{N}$: la frequenza che una casella avrebbe se i caratteri fossero indipendenti.

Indipendenza statistica
Due caratteri sono indipendenti se le distribuzioni condizionate di uno di essi sono tutte uguali.

Paradosso di Simpson
Un'associazione presente in ogni gruppo che scompare o si rovescia quando i gruppi vengono riuniti, a causa di un terzo carattere.

Residuo
Differenza tra il valore osservato y_i e quello previsto dalla retta, $e_i = y_i - (mx_i + q)$.

Retta di regressione
La retta che rende minima la somma dei quadrati dei residui: $m = \frac{\text{cov}(x, y)}{\sigma_x^2}$, $q = \bar{y} - m\bar{x}$.

Tabella a doppia entrata
Tabella che conta le unità statistiche secondo le modalità di due caratteri, uno sulle righe e uno sulle colonne.

Valore standardizzato
$z = \frac{x - \bar{x}}{\sigma}$: di quante deviazioni standard un dato sta sopra o sotto la media.

Soluzioni

1. B.

2. C. I punti sono allineati su una retta decrescente: $r = -1$, qualunque sia la pendenza.

3. D.

4. A. Da $q = \bar{y} - m\bar{x}$ segue $\bar{y} = m\bar{x} + q$.

5. D. Rispondono soltanto i visitatori che lo vogliono: l'ampiezza non corregge la distorsione.

6. a) Falso: $r = 0$ esclude soltanto un legame lineare; può esserci un legame curvo. b) Falso: s divide per $n - 1$ invece che per n, quindi $s \geq \sigma$ (uguali solo se i dati sono tutti uguali). c) Vero. d) Falso: coincidono solo se $|r| = 1$.

7. a) $f' = \dfrac{(\text{totale di riga}) \cdot (\text{totale di colonna})}{N}$. b) $e_i = y_i - (mx_i + q)$. c) media. d) estrapolazione.

8. a) Marginale: 60 : 200 = 30%. b) Condizionata per riga: 24 : 80 = 30%. c) Condizionata per colonna: 27 : 60 = 45%.

9. a) $z = \dfrac{8 - 6{,}2}{1{,}2} = 1{,}5$; $z = \dfrac{5 - 6{,}2}{1{,}2} = -1$. b) Nella corsa $z = \dfrac{58 - 70}{8} = -1{,}5$: 1,5 deviazioni standard sotto la media dei tempi, cioè 1,5 deviazioni standard «meglio» della media. Marco si è distinto nella stessa misura nelle due prove.

10. $\bar{x} = \dfrac{35}{5} = 7$; scarti −3, 0, 0, 1, 2; somma dei quadrati 14. $\sigma^2 = \dfrac{14}{5} = 2{,}8$, $\sigma \approx 1{,}67$; $s^2 = \dfrac{14}{4} = 3{,}5$, $s \approx 1{,}87$.

11. a) Righe biennio (30, 20, 10; totale 60) e triennio (24, 6, 30; totale 60); totali di colonna 54, 26, 40; totale 120. b) Romanzi 45%, fumetti ≈ 21,7%, saggi ≈ 33,3%. c) Biennio: 50%, ≈ 33,3%, ≈ 16,7%; triennio: 40%, 10%, 50%. d) No: le distribuzioni condizionate sono diverse (i saggi passano dal 16,7% al 50%).

12. Le due righe hanno lo stesso totale 60, quindi le stesse frequenze teoriche: romanzi $\dfrac{60 \cdot 54}{120} = 27$, fumetti 13, saggi 20. Gli scarti maggiori sono nei saggi (10 e 30 contro 20) e nei fumetti (20 e 6 contro 13).

13. $\dfrac{40 \cdot 30}{100} = 12$, $\dfrac{40 \cdot 70}{100} = 28$, $\dfrac{60 \cdot 30}{100} = 18$, $\dfrac{60 \cdot 70}{100} = 42$. Righe: 12 : 40 = 18 : 60 = 30% e 28 : 40 = 42 : 60 = 70%.

14. $\bar{x} = 3$, $\bar{y} = 4$; scarti di x: −2, −1, 0, 1, 2; di y: −2, 0, 1, 0, 1. Somma dei prodotti 6, $\Sigma (x_i - \bar{x})^2 = 10$, $\Sigma (y_i - \bar{y})^2 = 6$. $\text{cov}(x, y) = \dfrac{6}{5} = 1{,}2$; $r = \dfrac{6}{\sqrt{10 \cdot 6}} = \dfrac{6}{\sqrt{60}} \approx 0{,}77$.

15. a) $m = \dfrac{6}{10} = 0{,}6$, $q = 4 - 0{,}6 \cdot 3 = 2{,}2$: $y = 0{,}6x + 2{,}2$. b) Previsti 2,8; 3,4; 4; 4,6; 5,2; residui −0,8; 0,6; 1; −0,6; −0,2, con somma 0. c) $y = 0{,}6 \cdot 3{,}5 + 2{,}2 = 4{,}3$.

16. a) $m = 0{,}8 \cdot \dfrac{5}{2} = 2$, $q = 30 - 2 \cdot 12 = 6$: $y = 2x + 6$. b) $m' = 0{,}8 \cdot \dfrac{2}{5} = 0{,}32$, $q' = 12 - 0{,}32 \cdot 30 = 2{,}4$: $x = 0{,}32y + 2{,}4$. c) $2 \cdot 12 + 6 = 30$ e $0{,}32 \cdot 30 + 2{,}4 = 12$. Scritta come $y = \dfrac{x - 2{,}4}{0{,}32}$, la seconda retta ha pendenza $\dfrac{1}{0{,}32} = 3{,}125 \neq 2$.

17. Gli scarti delle altezze diventano 100 volte più piccoli. a) La covarianza diventa $\dfrac{37{,}33}{100} \approx 0{,}373$ (m · kg). b) r non cambia: ≈ 0,93. c) La pendenza diventa $\dfrac{224/100}{280/10\,000} = 80$ kg al metro (cioè sempre 0,8 kg al centimetro).

18. Somma delle x: 20, $\bar{x} = 2$; somma delle y: 76, $\bar{y} = 7{,}6$. $\Sigma (x_i - \bar{x})(y_i - \bar{y}) = -3{,}25$, $\Sigma (x_i - \bar{x})^2 = 9$, $\Sigma (y_i - \bar{y})^2 = 2{,}9$. $r = \dfrac{-3{,}25}{\sqrt{9 \cdot 2{,}9}} \approx -0{,}64$.

$m = \frac{-3{,}25}{9} \approx -0{,}361$, $q = 7{,}6 + 0{,}361 \cdot 2 \approx 8{,}32$: $y \approx -0{,}361x + 8{,}32$. Per $x = 1{,}5$: $y \approx 7{,}78$ ore.

19. a) $\Sigma (x_i - \overline{x}) = \Sigma x_i - n\overline{x} = n\overline{x} - n\overline{x} = 0$. b) $\Sigma e_i = \Sigma (y_i - mx_i - q) = n\overline{y} - mn\overline{x} - nq = n(\overline{y} - m\overline{x} - q) = 0$, perché $q = \overline{y} - m\overline{x}$. c.v.d.

20. Da $y_i = ax_i + b$ segue $\overline{y} = a\overline{x} + b$ e $y_i - \overline{y} = a(x_i - \overline{x})$. Allora $\Sigma (x_i - \overline{x})(y_i - \overline{y}) = a \Sigma (x_i - \overline{x})^2$ e $\sqrt{\Sigma (y_i - \overline{y})^2} = |a| \sqrt{\Sigma (x_i - \overline{x})^2}$, quindi $r = \frac{a}{|a|} = 1$ per $a > 0$. Se $a < 0$ si ottiene $r = -1$.

21. La correlazione non dimostra un nesso causale. La variabile confondente più probabile è il numero degli abitanti: i comuni più popolosi hanno più chiese e più furti. Per tenerne conto si confrontano i comuni a parità di popolazione, o si usano i tassi (chiese e furti ogni 1000 abitanti); probabilmente la correlazione si ridurrebbe molto.

22. Le distribuzioni per riga sono uguali se $\frac{a}{a+b} = \frac{c}{c+d}$ (le seconde frequenze relative sono allora uguali per differenza da 1). L'uguaglianza equivale a $a(c+d) = c(a+b)$, cioè $ad = bc$. La condizione $ad = bc$ è simmetrica: per colonne si deve avere $\frac{a}{a+c} = \frac{b}{b+d}$, cioè $a(b+d) = b(a+c)$, ancora $ad = bc$. c.v.d.

23. $\overline{x} = \overline{y} = 164$; $\Sigma (x_i - \overline{x})(y_i - \overline{y}) = 464$, $\Sigma (x_i - \overline{x})^2 = 438$, $\Sigma (y_i - \overline{y})^2 = 506$. a) $r = \frac{464}{\sqrt{438 \cdot 506}} \approx 0{,}99$. b) $m = \frac{464}{438} \approx 1{,}059$, $q = 164 - 1{,}059 \cdot 164 \approx -9{,}74$: $y \approx 1{,}059x - 9{,}74$. c) $y \approx 161{,}9$ cm.

24. $\overline{x} = 300$ g, $\overline{y} = 6$ cm; $\Sigma (x_i - \overline{x})(y_i - \overline{y}) = 1960$, $\Sigma (x_i - \overline{x})^2 = 100\,000$. a) $m = 0{,}0196$ cm/g (1,96 cm ogni 100 g), $q = 6 - 0{,}0196 \cdot 300 = 0{,}12$: $y = 0{,}0196x + 0{,}12$ ($r \approx 0{,}999$). b) $y \approx 6{,}98$ cm. c) È un'estrapolazione molto lontana dai dati (fino a 500 g): con 5 kg la molla potrebbe deformarsi in modo permanente o rompersi, e la legge lineare non valere più.

25. a) Linea 1: punta 70%, fuori punta 90%, totale 172 : 200 = 86%. Linea 2: punta 74%, fuori punta 92%, totale 157 : 200 = 78,5%. b) La linea 2 è più puntuale in entrambi i gruppi, ma in totale sembra peggiore perché fa 150 corse su 200 nelle ore di punta, quando il ritardo è più probabile: la fascia oraria è la variabile confondente (paradosso di Simpson).

26. a) Chi è in biblioteca alle 7.45 si alza presto e probabilmente dorme meno: campione di comodo. Il modulo facoltativo è autoselezionato: rispondono i più interessati. Si numerano i 900 studenti dall'elenco della segreteria e si estraggono 60 numeri distinti con un generatore di numeri casuali. b) Somma 58, $\overline{x} = 7{,}25$ ore; somma dei quadrati degli scarti 4,5; $s^2 = \frac{4{,}5}{7} \approx 0{,}643$, $s \approx 0{,}80$ ore.

27. Traccia: i valori cambiano a ogni ricalcolo. Le medie delle due colonne risultano vicine a 3,5; la deviazione standard delle medie di 5 lanci è circa 0,76, quella delle medie di 20 lanci circa 0,38 (con 1000 campioni la Figura 11.5 dà 0,78 e 0,38). L'istogramma dei campioni di 20 è più stretto: con un campione quattro volte più grande la dispersione si dimezza circa.

28. a) Il foglio dà $r \approx 0{,}928$, pendenza 0,8, intercetta −74. b) Con 75 kg: $y \approx 64{,}67$, $r \approx 0{,}85$, pendenza ≈ 0,914. c) In un gruppo di sei persone un solo dato insolito sposta in modo sensibile sia r sia la retta: le conclusioni tratte da pochi dati sono fragili.

29. a) Strumento sì: 20 (media ≥ 8), 30 (media < 8), totale 50; no: 25, 75, totale 100; totali di colonna 45 e 105; totale 150. b) Chi suona: 40% con media ≥ 8, 60% sotto; chi non suona: 25% e 75%. c) Frequenze teoriche $\frac{50 \cdot 45}{150} = 15, 35, 30, 70$: diverse dalle osservate, i caratteri sono dipendenti. d) No: è un'associazione osservata, non un esperimento; una terza variabile (per esempio l'abitudine allo studio regolare o l'ambiente familiare) potrebbe influire su entrambi i caratteri.

30. a) I punti scendono quasi allineati: correlazione negativa forte. b) $\overline{x} = 7$, $\overline{y} = 22$; $\Sigma (x_i - \overline{x})(y_i - \overline{y}) = -122$, $\Sigma (x_i - \overline{x})^2 = 70$, $\Sigma (y_i - \overline{y})^2 = 218$. $r = \frac{-122}{\sqrt{70 \cdot 218}} \approx -0{,}99$. $m = -\frac{122}{70} \approx -1{,}743$, $q = 22 + 1{,}743 \cdot 7 = 34{,}2$: $y \approx -1{,}743x + 34{,}2$. c) $y \approx 18{,}5$ m³. d) $y \approx -9{,}4$ m³, un consumo negativo, impossibile: oltre i 19,6 °C circa la retta dà valori negativi. È un'estrapolazione: con il caldo il riscaldamento è spento e il legame lineare non vale.

31. a) $m \cdot m' = \dfrac{\operatorname{cov}(x,y)^2}{\sigma_x^{\,2}\sigma_y^{\,2}} = r^2$. b) $S(k)$ è una funzione di secondo grado di k con coefficiente $\Sigma X_i^2 > 0$; il vertice è in $k = \dfrac{\Sigma X_i Y_i}{\Sigma X_i^2} = \dfrac{\operatorname{cov}(x,y)}{\sigma_x^{\,2}} = m$. c) Sostituendo:

$$S(m) = \Sigma Y_i^2 - \frac{(\Sigma X_i Y_i)^2}{\Sigma X_i^2} = \Sigma Y_i^2 \cdot \left(1 - \frac{(\Sigma X_i Y_i)^2}{\Sigma X_i^2 \, \Sigma Y_i^2}\right) = \Sigma Y_i^2 \cdot (1 - r^2).$$

Una somma di quadrati non è negativa, quindi $1 - r^2 \geq 0$, cioè $-1 \leq r \leq 1$; e $r = \pm 1$ solo se tutti i residui sono nulli. d) $m = 0{,}8$, $m' = \dfrac{224}{208} \approx 1{,}0769$, prodotto $\approx 0{,}8615$; $r^2 = \dfrac{224^2}{280 \cdot 208} = \dfrac{50\,176}{58\,240} \approx 0{,}8615$.