CAPITOLO

14

Statistica a due variabili

Una nuvola di punti non dice perché, ma dice quanto: misurare un legame è il primo passo per capirlo, e il più facile da fraintendere. (testo della redazione)

OBIETTIVI

1. Calcolare media, deviazione standard e coefficiente di variazione anche per dati raggruppati, e standardizzare un dato.
2. Costruire e leggere una tabella a doppia entrata con le distribuzioni marginali e condizionate.
3. Riconoscere l'indipendenza statistica e misurare l'associazione con le frequenze teoriche, l'indice χ^2 e l'indice V di Cramér.
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, analizzare i residui e valutare le previsioni.
6. Distinguere campione e popolazione, riconoscere le distorsioni e scegliere tra modelli lineari ed esponenziali.

INDICE DEL CAPITOLO

14.1 Media, deviazione standard e dati standardizzati 2
14.2 Distribuzioni doppie 4
14.3 Dipendenza e indipendenza statistica 7
14.4 Due caratteri quantitativi: la correlazione 9
14.5 La retta di regressione 13
14.6 Campioni, popolazioni e modelli 18

Matematica 3 · Parte V · Statistica a due variabili · versione 1.0
Capitolo 14: Statistica a due variabili · pagine 355–380 del volume completo · estratto per lo studio

Chi studia di più prende voti migliori? Chi abita lontano da scuola usa più spesso l'autobus? Una città con più gelaterie ha anche più scottature solari, e se sì, la colpa è dei gelati? Sono domande su *due* caratteri osservati sulle stesse unità statistiche, e la statistica del primo biennio, che descriveva un carattere alla volta, non basta più a rispondere. In questo capitolo impariamo a leggere insieme due caratteri: a organizzare i dati in tabelle a doppia entrata, a misurare quanto un carattere dipende dall'altro, a riassumere con un numero la forza di un legame lineare e a tracciare la retta che meglio descrive una nuvola di punti.

Useremo gli indici della statistica descrittiva di *Matematica 1*, la probabilità condizionata di *Matematica 2*, la parabola (capitolo 6) e i logaritmi (capitolo 13). Tutti i dati del capitolo sono **dati inventati**, costruiti per gli esempi, oppure dati che la classe può raccogliere da sé: altezze, aperture delle braccia, ore di studio, voti, tempi di percorrenza. Chiuderemo con due temi che le *Indicazioni nazionali* mettono al centro della statistica del secondo biennio: il rapporto tra un *campione* e la popolazione da cui è estratto, e il concetto di *modello matematico*.

14.1 Media, deviazione standard e dati standardizzati

14.1.1 Un richiamo

In *Matematica 1* abbiamo descritto una distribuzione di n dati $x_1, x_2, \dots, x_n$ con un indice di posizione, la **media aritmetica** $\overline{x}$, e con un indice di variabilità, la **deviazione standard** σ, radice quadrata della **varianza** σ^2, cioè della media dei quadrati degli scarti dalla media. Il simbolo Σ (sigma maiuscola) abbrevia una somma: $\Sigma\, x_i$ si legge «somma degli x_i» e indica $x_1 + x_2 + \dots + x_n$. Quando i dati sono raccolti in una tabella di frequenze, ogni valore x_k si conta tante volte quanta è la sua frequenza f_k, e le formule diventano medie *ponderate*. Se i dati sono raggruppati in **classi**, non conosciamo più i singoli valori: si sostituisce a ogni classe il suo **valore centrale**, la semisomma degli estremi, e si accetta un'approssimazione.

IN SINTESI

Indici di una distribuzione (popolazione di n dati; valori x_k con frequenze f_k).

- media: $\overline{x} = \frac{\Sigma\, x_i}{n}$, oppure $\overline{x} = \frac{\Sigma\, x_k f_k}{n}$;
- varianza: $\sigma^2 = \frac{\Sigma\, (x_i - \overline{x})^2}{n}$, oppure $\sigma^2 = \frac{\Sigma\, (x_k - \overline{x})^2 f_k}{n}$; deviazione standard $\sigma = \sqrt{\sigma^2}$;
- coefficiente di variazione (dati positivi): $CV = \frac{\sigma}{\overline{x}}$, spesso in percentuale;
- punteggio standardizzato di un dato x: $z = \frac{x - \overline{x}}{\sigma}$.

Il **coefficiente di variazione**, già incontrato nel primo volume, è un numero puro: per questo permette di confrontare la variabilità di grandezze diverse, o della stessa grandezza in gruppi con medie molto diverse. L'ultima formula è nuova. Il **punteggio standardizzato** z (o *punto zeta*) di un dato dice di quante deviazioni standard il dato si trova sopra ($z > 0$) o sotto ($z < 0$) la media. Standardizzare significa misurare ogni distanza con il «metro» σ della propria distribuzione: un dato con $z = 2$ è insolitamente alto in qualunque distribuzione, che si tratti di centimetri, di euro o di punti in una gara. I punteggi standardizzati di tutti i dati hanno sempre media 0 e deviazione standard 1.

ESEMPIO SVOLTO

Dati raggruppati. A 80 studenti di una scuola (dati inventati) si chiede il tempo che impiegano per arrivare a scuola. Le risposte, raggruppate in classi di 10 minuti, sono: da 0 a 10 minuti, 18 studenti; da 10 a 20, 30 studenti; da 20 a 30, 24 studenti; da 30 a 40, 8 studenti. Calcola media, deviazione standard e coefficiente di variazione, e il punteggio standardizzato di uno studente che impiega 35 minuti.

Valori centrali. Le classi hanno valori centrali 5, 15, 25 e 35 minuti.

Media. $\bar{x} = \dfrac{5 \cdot 18 + 15 \cdot 30 + 25 \cdot 24 + 35 \cdot 8}{80} = \dfrac{1420}{80} = 17{,}75$ minuti.

Varianza. Gli scarti dei valori centrali dalla media sono $-12{,}75$; $-2{,}75$; $7{,}25$; $17{,}25$. Quindi

$$\sigma^2 = \frac{12{,}75^2 \cdot 18 + 2{,}75^2 \cdot 30 + 7{,}25^2 \cdot 24 + 17{,}25^2 \cdot 8}{80} = \frac{6795}{80} \approx 84{,}94$$

e $\sigma \approx 9{,}22$ minuti. Il coefficiente di variazione è $CV = \dfrac{9{,}22}{17{,}75} \approx 52\%$: i tempi sono molto dispersi rispetto alla loro media, come ci si aspetta mescolando chi abita accanto alla scuola e chi viene da un altro comune.

Punteggio standardizzato. $z = \dfrac{35 - 17{,}75}{9{,}22} \approx 1{,}87$: lo studente impiega quasi due deviazioni standard più della media.

Attenzione. I risultati sono approssimati: abbiamo finto che tutti gli studenti di una classe impieghino il tempo centrale. Con i dati originali la media e la deviazione standard sarebbero un po' diverse.

14.2 Distribuzioni doppie

14.2.1 La tabella a doppia entrata

Torniamo agli studenti della nostra scuola e osserviamo su ciascuno *due* caratteri qualitativi: la **distanza** tra casa e scuola (meno di 2 km, da 2 a 5 km, oltre 5 km) e il **mezzo** usato più spesso per arrivarci (a piedi o in bici, in autobus, in auto). Ora gli studenti intervistati sono 240 (dati inventati). Ogni studente è una coppia di modalità, e i dati si contano in 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* le due modalità. Questo numero si chiama **frequenza congiunta**, e l'insieme delle frequenze congiunte è la **distribuzione doppia** dei due caratteri (Tabella 14.1).

Tabella 14.1 Distanza da scuola e mezzo di trasporto di 240 studenti (dati inventati): frequenze congiunte, con i totali di riga e di colonna.

Distanza \ mezzo	a piedi o in bici	in autobus	in auto	Totale
meno di 2 km	56	12	12	80
da 2 a 5 km	30	50	20	100
oltre 5 km	6	36	18	60
Totale	**92**	**98**	**50**	**240**

Le caselle si leggono così: 50 studenti abitano tra 2 e 5 km dalla scuola *e* vengono in autobus; 6 abitano oltre 5 km *e* vengono a piedi o in bici. La somma di tutte le frequenze congiunte è il numero $N = 240$ delle unità statistiche.

14.2.2 Distribuzioni marginali

Sommando le frequenze di una riga si ottiene il numero degli studenti con quella distanza, qualunque mezzo usino: 80, 100 e 60. Questi totali, scritti nell'ultima colonna, formano la distribuzione del solo carattere «distanza»; analogamente i totali delle colonne, 92, 98 e 50, formano la distribuzione del solo carattere «mezzo». Si chiamano **distribuzioni marginali**, perché si leggono sui margini della tabella. Ciascuna è una distribuzione semplice, come quelle del primo volume: possiamo calcolarne le frequenze relative (per esempio, $98 : 240 \approx 40{,}8\%$ degli studenti viene in autobus) e, se il carattere è quantitativo, la media e la deviazione standard.

Le due distribuzioni marginali, però, non dicono nulla sul *legame* tra i caratteri: la stessa coppia di margini è compatibile con moltissime tabelle diverse. L'informazione sul legame sta nelle caselle interne.

14.2.3 Distribuzioni condizionate

Chiediamoci: tra gli studenti che abitano oltre 5 km, come si distribuiscono i mezzi? Guardiamo soltanto la terza riga e dividiamo ogni frequenza per il totale della riga: $6 : 60 = 10\%$ a piedi o in bici, $36 : 60 = 60\%$ in autobus, $18 : 60 = 30\%$ in auto. È la **distribuzione condizionata** del mezzo *dato* che la distanza è «oltre 5 km». Ogni riga fornisce una distribuzione condizionata del mezzo, e ogni colonna una distribuzione condizionata della distanza. Per confrontare righe con totali diversi si usano le **percentuali di riga**, che sommano a 100% su ogni riga, come nella Figura 14.1.

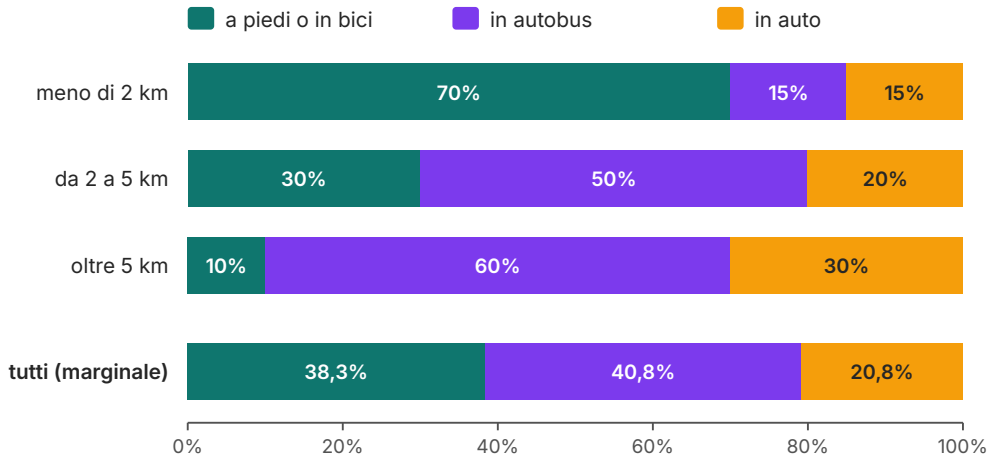


Figura 14.1 Distribuzioni condizionate del mezzo di trasporto per ciascuna classe di distanza (percentuali di riga della Tabella 14.1) e, in basso, la distribuzione marginale del mezzo su tutti i 240 studenti.

La figura risponde alla domanda iniziale con un colpo d'occhio: andando da «meno di 2 km» a «oltre 5 km» la quota di chi va a piedi o in bici crolla dal 70% al 10%, mentre quelle di autobus e auto crescono. Le distribuzioni condizionate sono *diverse* da riga a riga: il mezzo dipende dalla distanza. Si possono anche leggere le percentuali di colonna: tra i 92 studenti che vengono a piedi o in bici, $56 : 92 \approx 60{,}9\%$ abita a meno di 2 km. Attenzione a non confondere le due letture: «il 70% di chi abita vicino va a piedi» e «il 60,9% di chi va a piedi abita vicino» sono informazioni diverse, ottenute dividendo la stessa frequenza 56 per due totali diversi.

COLLEGAMENTI

Frequenze e probabilità condizionate. In *Matematica 2* abbiamo definito la probabilità condizionata $p(A|B) = \dfrac{p(A \cap B)}{p(B)}$. La distribuzione condizionata è la sua versione statistica: se scegliamo a caso uno studente della scuola, la probabilità che venga in auto *sapendo che* abita oltre 5 km è $\dfrac{18/240}{60/240} = \dfrac{18}{60} = 30\%$, proprio la frequenza relativa condizionata. Per la legge empirica del caso, inoltre, in un campione grande le frequenze relative condizionate sono buone stime delle probabilità condizionate nella popolazione. È così che i medici stimano la probabilità di un sintomo in chi ha una certa malattia, e gli epidemiologi confrontano la frequenza di una malattia tra chi è esposto a un fattore di rischio e chi no: sono tabelle a doppia entrata lette per righe.

14.3 Dipendenza e indipendenza statistica

14.3.1 Quando due caratteri sono indipendenti

Se il mezzo non dipendesse affatto dalla distanza, ogni riga avrebbe la stessa distribuzione condizionata, e quindi la stessa distribuzione marginale del mezzo: il 38,3% a piedi o in bici, il 40,8% in autobus, il 20,8% in auto, qualunque sia la distanza. Diciamo che due caratteri sono **statisticamente indipendenti** quando le distribuzioni condizionate (in percentuale) di uno dei due sono tutte uguali tra loro; in caso contrario i caratteri sono **dipendenti** o **associati**. Si dimostra che, se le distribuzioni condizionate per righe sono tutte uguali, lo sono anche quelle per colonne: l'indipendenza è una relazione simmetrica.

Da questa definizione si ricava quale frequenza «dovremmo» trovare in ogni casella se i caratteri fossero indipendenti. Nella prima riga ci sono 80 studenti; se il 92/240 di loro andasse a piedi o in bici, come nella marginale, la casella conterrebbe $80 \cdot \frac{92}{240} = \frac{80 \cdot 92}{240} \approx 30{,}67$ 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 congiunta osservata f_{ij} coincide con la frequenza teorica. È la stessa formula che in *Matematica 2* caratterizzava gli eventi indipendenti, $p(A \cap B) = p(A) \cdot p(B)$: dividendo per N ambo i membri, la frequenza relativa congiunta è il prodotto delle due frequenze relative marginali. Le frequenze teoriche non sono necessariamente numeri interi, e non devono esserlo: descrivono una situazione ideale.

14.3.2 L'indice chi-quadrato

Nei dati reali l'indipendenza perfetta non si osserva quasi mai: anche se due caratteri non hanno alcun legame, il caso produce piccole differenze tra frequenze osservate e teoriche. Serve allora un numero che misuri *quanto* la tabella si allontana dall'indipendenza. Per ogni casella calcoliamo la differenza tra frequenza osservata e teorica, la eleviamo al quadrato (così i segni non si compensano) e la dividiamo per la frequenza teorica (così una differenza di 5 pesa di più in una casella dove ce ne aspettavamo 10 che in una dove ce ne aspettavamo 100). La somma su tutte le caselle è l'**indice chi-quadrato**:

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

La lettera χ è la «chi» greca. L'indice vale 0 se e solo se la tabella è di perfetta indipendenza, ed è tanto più grande quanto più le frequenze osservate si allontanano da quelle teoriche. Ha però

un difetto: dipende dal numero N delle osservazioni. Se raddoppiamo ogni frequenza della tabella, le distribuzioni condizionate non cambiano, ma χ^2 raddoppia. Per confrontare tabelle di dimensioni diverse lo si normalizza. In questo libro usiamo l'**indice V di Cramér**:

$V = \sqrt{\left(\frac{\chi^2}{N \cdot (k-1)}\right)}$, dove k è il minore tra il numero delle righe e il numero delle colonne

Si dimostra che χ^2 non supera mai $N \cdot (k-1)$, per cui $0 \le V \le 1$. $V = 0$ significa indipendenza perfetta; $V = 1$ significa dipendenza perfetta: conoscendo la modalità del carattere che ne ha di più, quella dell'altro ne è determinata, perché ogni riga o colonna di quel carattere ha una sola casella non vuota (in una tabella quadrata, ogni riga e ogni colonna ha una sola casella non vuota). Nella pratica, valori di V sotto 0,1 indicano un'associazione trascurabile, valori sopra 0,5 un'associazione forte; sono solo orientamenti, non regole.

ESEMPIO SVOLTO

Misurare l'associazione. Calcola le frequenze teoriche, l'indice χ^2 e l'indice V per i dati della Tabella 14.1.

Frequenze teoriche. Per ogni casella moltiplichiamo il totale di riga per il totale di colonna e dividiamo per 240. Per esempio, per «da 2 a 5 km, in autobus»: $\frac{100 \cdot 98}{240} \approx 40{,}83$. I risultati, arrotondati al centesimo, sono nella tabella qui sotto, accanto alle frequenze osservate.

Tabella 14.2 Frequenze osservate e teoriche (tra parentesi) per i dati della Tabella 14.1.

Distanza \ mezzo	a piedi o in bici	in autobus	in auto
meno di 2 km	56 (30,67)	12 (32,67)	12 (16,67)
da 2 a 5 km	30 (38,33)	50 (40,83)	20 (20,83)
oltre 5 km	6 (23,00)	36 (24,50)	18 (12,50)

Contributi a χ^2. Per la prima casella: $\frac{(56 - 30{,}67)^2}{30{,}67} \approx 20{,}93$. Procedendo allo stesso modo, i nove contributi sono, riga per riga, 20,93, 13,07, 1,31; 1,81, 2,06, 0,03; 12,57, 5,40, 2,42. La loro somma è $\chi^2 \approx 59{,}59$ (con le frequenze teoriche esatte, non arrotondate).

Indice V. La tabella è 3×3, quindi $k = 3$ e $k - 1 = 2$: $V = \sqrt{\left(\frac{59{,}59}{240 \cdot 2}\right)} \approx 0{,}35$.

Interpretazione. Tra mezzo e distanza c'è un'associazione di intensità media: la distanza conta, ma non determina il mezzo (ci sono studenti lontani che vengono in bici e vicini che vengono in auto). I contributi maggiori vengono dalle caselle «meno di 2 km, a piedi o in bici», dove gli studenti sono molti più del previsto, e «meno di 2 km, in autobus», dove sono molti meno.

Due avvertenze. Primo: χ^2 e V misurano l'*intensità* del legame, non il suo verso né la sua causa; per capire *come* i caratteri sono legati bisogna tornare alle distribuzioni condizionate. Secondo: stabilire se un certo valore di χ^2 è troppo grande per essere dovuto al caso è un problema di **statistica inferenziale**, che richiede strumenti probabilistici fuori dal programma di questo anno; qui usiamo χ^2 e V soltanto per descrivere i dati raccolti.

14.4 Due caratteri quantitativi: la correlazione

14.4.1 Il diagramma a dispersione

Quando entrambi i caratteri sono quantitativi, 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 inglese *scatter plot*). In una classe si è chiesto a sei studenti quante ore studiano matematica a casa in una settimana tipo (x) e si è preso il voto dell'ultima verifica (y); i dati, inventati, sono nella Tabella 14.3 e nella Figura 14.2.

Tabella 14.3 Ore di studio settimanali e voto di sei studenti (dati inventati).

Studente	Ore di studio x	Voto y
1	3	4
2	4	4
3	8	7
4	10	8
5	11	9
6	12	7

La nuvola sale da sinistra a destra: a più ore corrispondono, *in media*, voti più alti, anche se il legame non è rigido (lo studente che studia 12 ore ha preso 7, meno di quello che ne studia 11). Si dice che tra i due caratteri c'è una **correlazione positiva**. Se la nuvola scendesse, parleremmo di **correlazione negativa**. Le medie sono $\overline{x} = 8$ ore e $\overline{y} = 6{,}5$; le rette $x = 8$ e $y = 6{,}5$ dividono il piano in quattro quadranti centrati nel **baricentro** $G(\overline{x}; \overline{y})$ della nuvola.

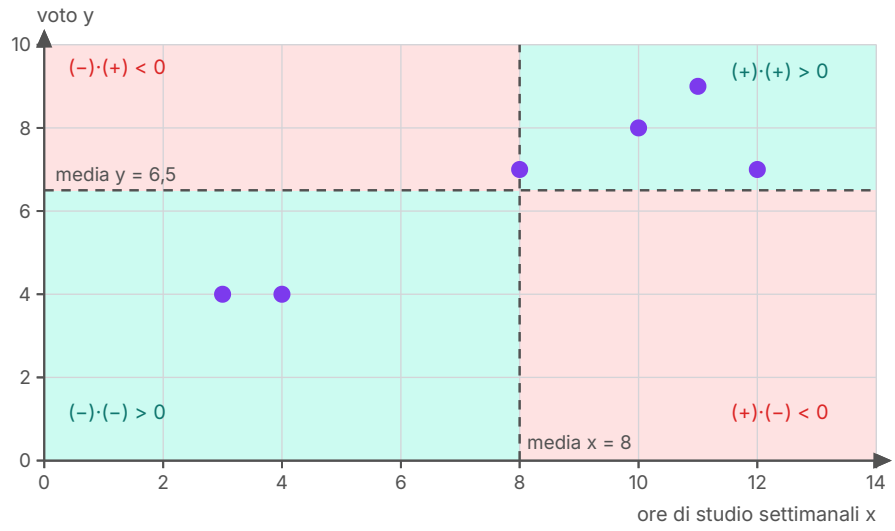


Figura 14.2 Diagramma a dispersione dei dati della Tabella 14.3. Le rette tratteggiate passano per le medie; nei quadranti verdi il prodotto degli scarti è positivo, in quelli rossi è negativo. Tutti i punti stanno nei quadranti verdi o sul loro bordo.

14.4.2 La covarianza

Per ogni punto consideriamo gli scarti dalle medie, $(x_i - \bar{x})$ e $(y_i - \bar{y})$, e il loro prodotto. Nei quadranti 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 (Figura 14.2). Se la nuvola sale, prevalgono i prodotti positivi; se scende, quelli negativi; se non ha alcuna direzione, si compensano. La media dei prodotti degli scarti si chiama **covarianza**:

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

Il segno della covarianza indica il verso del legame. Il suo valore, invece, è difficile da interpretare: si misura in un'unità composta (qui «ore × punti di voto») e cambia se cambiamo unità di misura: esprimendo il tempo in minuti invece che in ore, la covarianza diventa 60 volte più grande, mentre il legame tra i dati è lo stesso. Per misurare l'*intensità* del legame occorre un numero puro.

14.4.3 Il coefficiente di correlazione lineare

Dividiamo la covarianza per il prodotto delle due deviazioni standard. Si ottiene il **coefficiente di correlazione lineare** di Bravais-Pearson:

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

(nella seconda forma i denominatori n si semplificano). Il coefficiente r non ha unità di misura e non cambia se si cambiano le unità o si sposta l'origine dei dati. Dimostreremo nel prossimo paragrafo che vale sempre $-1 \le r \le 1$. Il **segno** di r è quello della covarianza; il **valore assoluto** misura quanto i punti sono vicini a una retta: $|r| = 1$ se e solo se tutti i punti stanno esattamente su una retta (non parallela agli assi), mentre r vicino a 0 significa che nessuna retta descrive bene la nuvola. La Figura 14.3 mostra quattro nuvole di 30 punti, generate al computer, con il loro coefficiente.

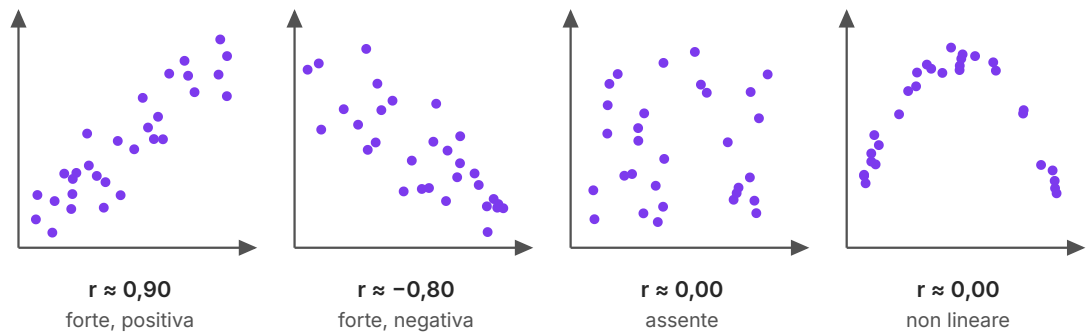


Figura 14.3 Quattro diagrammi a dispersione di 30 punti generati al computer, con il coefficiente di correlazione calcolato sui punti disegnati. Nell'ultimo i punti seguono da vicino una parabola, eppure $r \approx 0$: il coefficiente misura soltanto il legame *lineare*.

Il quarto diagramma è il più istruttivo: tra x e y c'è un legame fortissimo (conoscendo x si prevede y quasi esattamente), ma non è lineare. A sinistra del vertice i punti salgono, a destra scendono, e i prodotti degli scarti si compensano: r risulta quasi nullo. Per questo il diagramma a dispersione va *sempre* disegnato prima di calcolare r: un numero da solo può nascondere la forma della nuvola.

ESEMPIO SVOLTO

Calcolo di r. Calcola la covarianza e il coefficiente di correlazione per i dati della Tabella 14.3.

Medie. $\bar{x} = \frac{3+4+8+10+11+12}{6} = \frac{48}{6} = 8$; $\bar{y} = \frac{4+4+7+8+9+7}{6} = \frac{39}{6} = 6{,}5$.

Tabella 14.4 Calcolo degli scarti e dei loro prodotti per i dati della Tabella 14.3.

x	y	$x - \bar{x}$	$y - \bar{y}$	prodotto	$(x - \bar{x})^2$	$(y - \bar{y})^2$
3	4	−5	−2,5	12,5	25	6,25
4	4	−4	−2,5	10	16	6,25
8	7	0	0,5	0	0	0,25
10	8	2	1,5	3	4	2,25
11	9	3	2,5	7,5	9	6,25
12	7	4	0,5	2	16	0,25
			Somma	**35**	**70**	**21,5**

Covarianza e deviazioni standard. $\operatorname{cov}(x, y) = \frac{35}{6} \approx 5{,}83$; $\sigma_x^2 = \frac{70}{6} \approx 11{,}67$; $\sigma_y^2 = \frac{21{,}5}{6} \approx 3{,}58$.

Coefficiente. Con la seconda forma, che evita gli arrotondamenti intermedi: $r = \frac{35}{\sqrt{70 \cdot 21{,}5}} = \frac{35}{\sqrt{1505}} \approx 0{,}902$.

Interpretazione. r è positivo e vicino a 1: tra ore di studio e voto c'è, in questo gruppo, una correlazione lineare positiva forte. Sei dati sono pochi, però: basterebbe un settimo studente insolito per cambiare parecchio il risultato.

14.4.4 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 (studiare di più fa migliorare). Può darsi che sia y a influire su x (chi ha avuto buoni voti si appassiona e studia di più). 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**.

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 solari, 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 assolate, in cui la spiaggia è affollata. La temperatura è la variabile confondente. Molte correlazioni meno evidenti hanno la stessa struttura: tra le città, il numero di farmacie e il numero di nascite sono 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, come si fa in fisica e in medicina.

«r = 0, quindi non c'è relazione». Anche questo è sbagliato: $r = 0$ esclude soltanto un legame *lineare* (quarto diagramma della Figura 14.3).

14.5 La retta di regressione

14.5.1 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 significa «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 la retta, negativo se sta sotto. Una buona retta ha residui piccoli. Non possiamo semplicemente sommarli, perché residui 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 orbite di comete e pianeti, sceglie la retta che rende *minima la somma dei quadrati dei residui*:

$$S(m, q) = \Sigma\,(y_i - mx_i - q)^2$$

La retta così trovata si chiama **retta di regressione** di y su x. Per trovarla bastano gli strumenti di questo volume: la somma S dipende da due incognite, m e q, ma ragionando su una incognita alla volta diventa la funzione di secondo grado di una sola variabile, e il minimo di una parabola con la concavità verso l'alto è il suo vertice (capitolo 6).

Primo passo: q, per un m fissato. Poniamo $u_i = y_i - mx_i$. Allora

$$S = \Sigma\,(u_i - q)^2 = nq^2 - 2q\,\Sigma\,u_i + \Sigma\,u_i^2$$

è una parabola nella variabile q, con il coefficiente di q^2 uguale a $n > 0$; il suo vertice ha ascissa $q = \frac{2\,\Sigma\, u_i}{2n} = \overline{u}$, la media degli u_i. Poiché $\overline{u} = \overline{y} - m\overline{x}$, per ogni m il valore migliore di q è

$$q = \overline{y} - m\overline{x}$$

DA RICORDARE

La retta dei minimi quadrati passa per il baricentro. L'uguaglianza $q = \overline{y} - m\overline{x}$ si può scrivere $\overline{y} = m\overline{x} + q$: il punto $G(\overline{x}; \overline{y})$ sta sulla retta, qualunque sia la sua pendenza. Il motivo è quello appena visto: per una pendenza fissata, spostare la retta in alto o in basso cambia tutti i residui della stessa quantità, e la somma dei loro quadrati è minima quando i residui hanno media zero, cioè quando la retta passa all'«altezza media» della nuvola. Ne segue anche che *la somma dei residui della retta di regressione è sempre zero*: i punti sopra e i punti sotto si bilanciano esattamente.

Secondo passo: *m*. Con questo q il residuo diventa $e_i = (y_i - \overline{y}) - m(x_i - \overline{x})$. Abbreviamo gli scarti con $X_i = x_i - \overline{x}$ e $Y_i = y_i - \overline{y}$:

$$S(m) = \Sigma\,(Y_i - mX_i)^2 = m^2\,\Sigma\, X_i^2 - 2m\,\Sigma\, X_iY_i + \Sigma\, Y_i^2$$

È di nuovo una parabola, nella variabile m, con il coefficiente di m^2 uguale a $\Sigma\, X_i^2 > 0$ (le x non sono tutte uguali). Il vertice ha ascissa

$$m = \frac{\Sigma\, X_iY_i}{\Sigma\, X_i^2} = \frac{\mathrm{cov}(x, y)}{\sigma_x^{\,2}}$$

(abbiamo diviso numeratore e denominatore per n). Poiché il minimo di S si ottiene scegliendo il q migliore per ogni m e poi il migliore tra questi m, la coppia trovata è proprio quella che rende minima $S(m, q)$.

Il valore minimo di S è l'ordinata del vertice, $-\frac{\Delta}{4a}$ (capitolo 6), che si può scrivere anche $c - \frac{b^2}{4a}$. Si ottiene $S_{\min} = \Sigma\, Y_i^2 - \frac{(\Sigma\, X_iY_i)^2}{\Sigma\, X_i^2}$. Raccogliendo $\Sigma\, Y_i^2$ e ricordando la seconda forma di r, si ottiene

$$S_{\min} = \Sigma\,(y_i - \overline{y})^2 \cdot (1 - r^2)$$

Questa uguaglianza contiene un piccolo teorema. Una somma di quadrati non è mai negativa, quindi $1 - r^2 \geq 0$, cioè $-1 \leq r \leq 1$. Inoltre $r = \pm 1$ se e solo se $S_{\min} = 0$, cioè se e solo se tutti i residui sono nulli: tutti i punti stanno sulla retta. c.v.d.

IN SINTESI

Teorema (retta dei minimi quadrati). 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

$$m = \frac{\text{cov}(x, y)}{\sigma_x^2}, \quad q = \overline{y} - m\overline{x}$$

Essa passa per $G(\overline{x}; \overline{y})$; la somma dei suoi residui è 0; la somma dei quadrati dei residui è $\Sigma\,(y_i - \overline{y})^2(1 - r^2)$, e quindi $-1 \le r \le 1$. Poiché $r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$, si ha anche $m = r \cdot \frac{\sigma_y}{\sigma_x}$: la pendenza ha lo stesso segno di r.

ESEMPIO SVOLTO

Retta di regressione e residui. Per i dati della Tabella 14.3 trova la retta di regressione del voto sulle ore di studio, calcola i residui e prevedi il voto di uno studente che studia 6 ore.

Coefficienti. Dalla tabella dell'esempio precedente: $m = \frac{35}{70} = 0{,}5$ (i denominatori n si semplificano) e $q = 6{,}5 - 0{,}5 \cdot 8 = 2{,}5$. La retta è $y = 0{,}5x + 2{,}5$: secondo il modello, ogni ora di studio in più alla settimana vale mezzo voto.

Residui. I voti previsti per $x = 3, 4, 8, 10, 11, 12$ sono 4; 4,5; 6,5; 7,5; 8; 8,5. I residui (osservato meno previsto) sono 0; −0,5; 0,5; 0,5; 1; −1,5 (Figura 14.4). La loro somma è 0, come deve essere. La somma dei quadrati è $0 + 0{,}25 + 0{,}25 + 0{,}25 + 1 + 2{,}25 = 4$, e infatti $21{,}5 \cdot (1 - r^2) = 21{,}5 \cdot \frac{8}{43} = 4$, perché $r^2 = \frac{35^2}{70 \cdot 21{,}5} = \frac{35}{43}$.

Previsione. Per $x = 6$: $y = 0{,}5 \cdot 6 + 2{,}5 = 5{,}5$. Il valore 6 è compreso tra il minimo (3) e il massimo (12) delle ore osservate, quindi la previsione è ragionevole, purché la si legga come «voto medio atteso», non come voto certo: i residui mostrano scostamenti fino a un voto e mezzo.

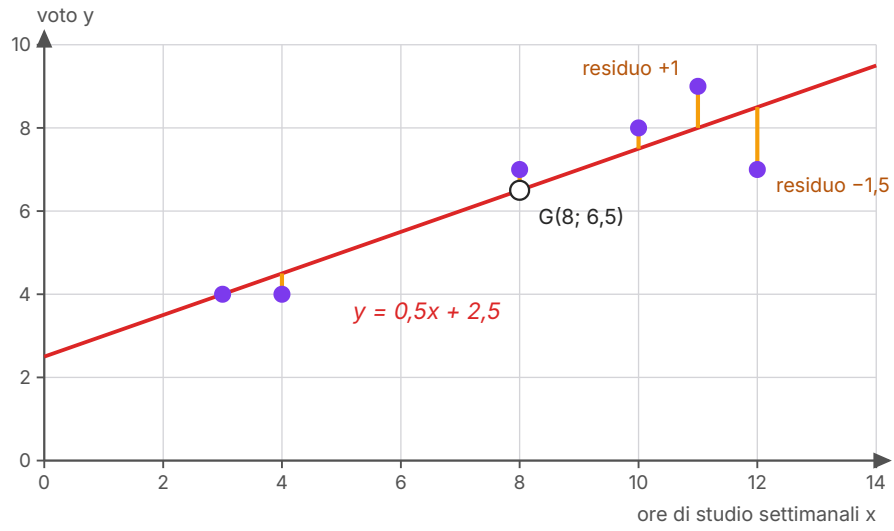


Figura 14.4 La retta di regressione $y = 0{,}5x + 2{,}5$ dei dati della Tabella 14.3, con i residui (segmenti arancioni). La retta passa per il baricentro $G(8;\ 6{,}5)$; il punto con $x = 3$ sta esattamente sulla retta.

14.5.2 Previsioni: interpolazione ed estrapolazione

Come in *Matematica 2*, 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 là dove li abbiamo osservati, e nulla garantisce che il legame resti lineare altrove. Con i nostri dati, uno studente che studiasse 20 ore alla settimana dovrebbe prendere $0{,}5 \cdot 20 + 2{,}5 = 12{,}5$, un voto che non esiste (Figura 14.5).

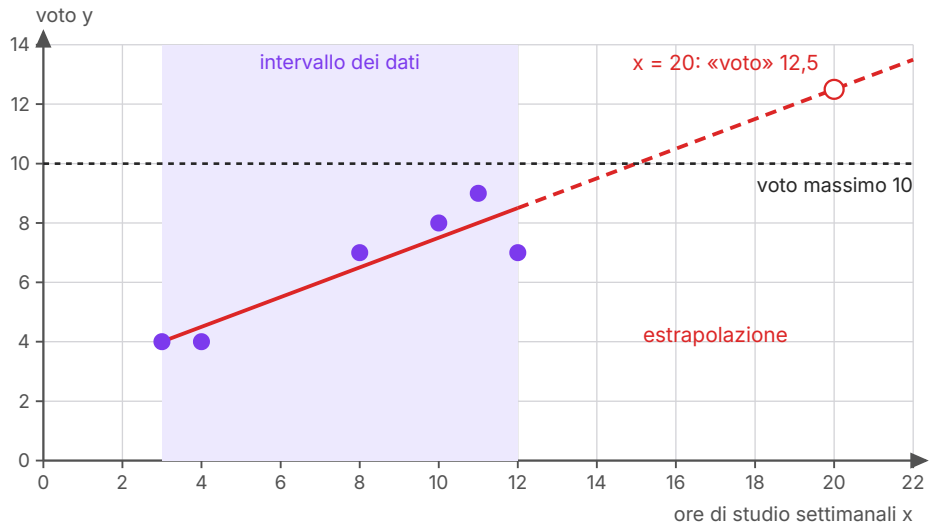


Figura 14.5 Estrapolazione: la retta di regressione, prolungata ben oltre l'intervallo dei dati (fascia viola), prevede per 20 ore di studio un voto di 12,5, superiore al massimo possibile.

14.5.3 Quanto spiega la retta

La varianza dei voti, $\sigma_y^2 \approx 3{,}58$, misura quanto i voti sono diversi tra loro. Una parte di queste differenze è «spiegata» dalla retta, perché studenti con ore diverse hanno voti previsti diversi; la parte rimanente è nei residui. Dalla formula di $S_{\min}$, dividendo per n, la varianza dei residui è $\sigma_y^2(1 - r^2)$. Il numero r^2 si chiama **coefficiente di determinazione** e rappresenta la **quota della varianza di y spiegata dalla retta**. Nel nostro esempio $r^2 = \dfrac{35}{43} \approx 0{,}81$: circa l'81% della variabilità dei voti è descritto dalle ore di studio, il 19% resta ai residui. Un coefficiente $r = 0{,}5$, che sembra discreto, spiega soltanto il 25% della varianza.

Un'ultima osservazione. La retta di regressione di y su x minimizza i residui *verticali*, perché tratta x come il dato da cui si parte e y come la grandezza da prevedere. Se si volesse prevedere x conoscendo y, si dovrebbero minimizzare i residui orizzontali, e si otterrebbe la retta di regressione di x su y, $x = m'y + q'$ con $m' = \dfrac{\operatorname{cov}(x, y)}{\sigma_y^2}$: una retta diversa, che coincide con la prima soltanto se $|r| = 1$.

APPROFONDIMENTO · GALTON E LA «REGRESSIONE»

La parola *regressione* viene da uno studio sull'ereditarietà. Nel 1886 il naturalista inglese Francis Galton, cugino di Charles Darwin, pubblicò un articolo intitolato *Regression towards mediocrity in hereditary stature*, «Regressione verso la mediocrità nella statura ereditaria». Confrontando le stature di centinaia di genitori e figli, Galton osservò che i figli di genitori molto alti sono in media alti, ma meno dei genitori, e che i figli di genitori molto bassi sono in media bassi, ma meno dei genitori: le stature dei figli «regrediscono» verso la media della popolazione. Galton stimò che lo scarto dalla media dei figli fosse circa due terzi di quello dei genitori.

Nel linguaggio di questo capitolo, Galton aveva trovato una retta di regressione con pendenza minore di 1, il che accade, tra variabili con la stessa deviazione standard, ogni volta che $|r| < 1$: è la relazione $m = r \cdot \sigma_y : \sigma_x$. La «regressione verso la media» non è una legge biologica, ma una conseguenza di una correlazione imperfetta: compare anche nei voti di due verifiche successive, nelle prestazioni sportive di due stagioni, nei risultati di un paziente misurati due volte. Il coefficiente di correlazione fu sviluppato da Karl Pearson a partire dalle idee di Galton; la formula era già comparsa, a metà dell'Ottocento, nei lavori del fisico francese Auguste Bravais, e per questo porta il nome di entrambi.

14.6 Campioni, popolazioni e modelli

14.6.1 Campione e popolazione

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

Perché la stima sia attendibile il campione dev'essere **rappresentativo**, cioè riprodurre in piccolo le caratteristiche della popolazione. 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 i numeri con un generatore di numeri casuali. Un campione scelto in modo da favorire sistematicamente alcune unità produce invece stime sbagliate nella stessa direzione, qualunque sia la sua ampiezza: si dice che il campione ha una **distorsione** (in inglese *bias*).

Le fonti di distorsione più comuni sono tre. Il **campione di comodo**: intervistare sulla mobilità soltanto chi passa alla fermata dell'autobus. L'**autoselezione**: nei sondaggi online, o nelle trasmissioni che invitano a telefonare, risponde soltanto chi ha voglia di farlo, 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 correttamente. Il caso più citato è il sondaggio della rivista americana *The Literary Digest* per le elezioni presidenziali del 1936. La rivista spedì il questionario a circa dieci milioni di persone e ne ricevette indietro circa 2,4 milioni, un campione enorme; prevedeva una larga vittoria del candidato repubblicano Alf Landon, con circa il 57% dei voti, ma vinse il presidente in carica Franklin D. Roosevelt, con circa il 61%. Gli indirizzi erano stati presi soprattutto da elenchi di persone benestanti, e a rispondere furono soprattutto gli elettori più ostili a Roosevelt: un campione grande ma distorto sbaglia con grande precisione.

14.6.2 Varianza campionaria e margine d'errore

Come si è accennato nel primo volume, quando i dati sono un campione e vogliamo stimare la varianza della popolazione, al denominatore si mette $n - 1$ invece di n:

$$s^2 = \frac{\Sigma\,(x_i - \overline{x})^2}{n - 1}$$

Si chiama **varianza campionaria**, e la sua radice s è la deviazione standard campionaria. La ragione si intuisce con quanto visto nel paragrafo precedente: la somma $\Sigma\,(x_i - c)^2$, come funzione di c, è una parabola con il vertice in $c = \overline{x}$ (è lo stesso calcolo fatto per q). Quindi gli scarti dalla media del *campione* hanno una somma dei quadrati più piccola di quella degli scarti dalla media vera della popolazione, che non conosciamo: dividendo per n si sottostimerebbe, in media, la variabilità. Si dimostra (ma non lo faremo) che dividere per $n - 1$ corregge esattamente questo difetto. Per n grande la differenza è trascurabile. In questo libro σ indica sempre l'indice calcolato con n su dati trattati come popolazione, s quello calcolato con $n - 1$ su un campione.

Anche un campione casuale senza distorsioni non dà esattamente il parametro: campioni diversi danno stime diverse. I sondaggi seri accompagnano la stima con un **margine d'errore**: «il 42% degli intervistati, con un margine d'errore di 3 punti» significa che la percentuale nella popolazione è stimata tra il 39% e il 45%, con un metodo che, ripetuto su molti campioni, dà un intervallo giusto con un'alta probabilità fissata in anticipo (di solito il 95%). Il margine d'errore diminuisce al crescere dell'ampiezza del campione, ma lentamente: per dimezzarlo occorre un campione circa quattro volte più grande. Il margine misura soltanto l'errore dovuto al caso, non quello dovuto alla distorsione: il sondaggio del 1936 aveva un margine d'errore minuscolo e una distorsione enorme. Calcolarlo richiede la statistica inferenziale, che studierai eventualmente all'università.

14.6.3 Il modello matematico

Una retta di regressione è un esempio di modello matematico, nel senso già incontrato in *Matematica 2*: 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. Per scegliere tra modelli diversi si guardano i residui. Se un modello è adatto, i residui sono piccoli e sparsi a caso, senza una forma riconoscibile; se i residui hanno un andamento regolare, per esempio positivi agli estremi e negativi al centro, il modello trascura qualcosa che i dati mostrano.

Molti fenomeni crescono in modo esponenziale (capitoli 12 e 13), e non linearmente. Come abbiamo visto nel capitolo 13, il logaritmo trasforma un modello esponenziale in uno lineare: se $y = A \cdot b^t$, allora $\ln y = \ln A + t \cdot \ln b$, cioè $\ln y$ è funzione lineare di t. Basta allora calcolare la retta di regressione delle coppie $(t; \ln y)$ e tornare indietro con l'esponenziale.

ESEMPIO SVOLTO

Lineare o esponenziale? In una coltura di lieviti si misura ogni ora la concentrazione N delle cellule, in migliaia per millilitro (dati inventati): per $t = 0, 1, \ldots, 6$ h si trova $N = 2{,}0;\ 2{,}7;\ 3{,}3;\ 4{,}5;\ 5{,}6;\ 7{,}6;\ 9{,}5$. Confronta un modello lineare e un modello esponenziale.

Modello lineare. La retta di regressione di N su t è $N \approx 1{,}236t + 1{,}321$, con $r \approx 0{,}977$: un coefficiente altissimo. Ma i residui, $0{,}68;\ 0{,}14;\ -0{,}49;\ -0{,}53;\ -0{,}66;\ 0{,}10;\ 0{,}76$, sono positivi agli estremi e negativi al centro: i dati curvano verso l'alto e la retta non lo vede.

Modello esponenziale. Calcoliamo $\ln N$: $0{,}693;\ 0{,}993;\ 1{,}194;\ 1{,}504;\ 1{,}723;\ 2{,}028;\ 2{,}251$. La retta di regressione di $\ln N$ su t è $\ln N \approx 0{,}260t + 0{,}705$, con $r \approx 0{,}999$. Tornando alla grandezza originale, $N \approx e^{0{,}705} \cdot (e^{0{,}260})^t \approx 2{,}02 \cdot 1{,}30^t$: la popolazione cresce di circa il 30% ogni ora. I residui sono $-0{,}02;\ 0{,}08;\ -0{,}10;\ 0{,}09;\ -0{,}12;\ 0{,}19;\ -0{,}11$: piccoli e senza una forma regolare (Figura 14.6).

Scelta. Il modello esponenziale descrive meglio i dati, ed è anche quello che ci si aspetta da una popolazione di cellule che si dividono. Entrambi i modelli, però, valgono solo per il periodo osservato: dopo dieci ore il modello lineare prevede 13,7 e quello esponenziale 27,2, e nessuno dei due tiene conto che il nutrimento della coltura prima o poi si esaurisce.

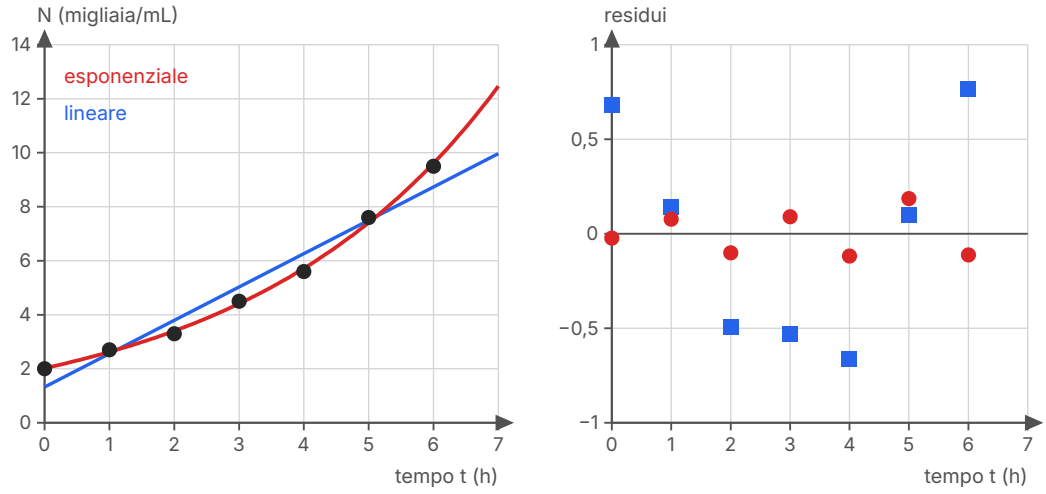


Figura 14.6 A sinistra, i dati della coltura di lieviti con la retta di regressione (blu) e il modello esponenziale ottenuto con i logaritmi (rosso). A destra, i residui dei due modelli: quelli lineari (quadrati blu) hanno un andamento regolare, quelli esponenziali (punti rossi) sono piccoli e sparsi.

L'esempio illustra un'idea che le *Indicazioni nazionali* chiedono di avere chiara: la differenza tra la matematizzazione della fisica classica, in cui si cerca la legge che corrisponde esattamente al fenomeno, e la **modellistica**, in cui la stessa classe di fenomeni può essere rappresentata con approcci diversi, ciascuno utile entro certi limiti. La crescita di una popolazione si può descrivere con una retta su un intervallo breve, con un esponenziale finché le risorse abbondano, con modelli più complessi quando cominciano a scarseggiare. Il modello migliore non è il più elegante, ma quello che descrive i dati con residui piccoli e senza struttura, nell'intervallo in cui ci serve, e che ha un senso rispetto a ciò che sappiamo del fenomeno.

COLLEGAMENTI

Fisica: la retta nei dati di laboratorio. In laboratorio si appendono a una molla masse diverse e se ne misura l'allungamento. La legge di Hooke prevede che l'allungamento sia direttamente proporzionale alla forza peso, ma ogni misura ha un errore: i punti non sono mai perfettamente allineati. La retta dei minimi quadrati è il modo standard di ricavare dai dati la costante elastica, ed è la stessa retta calcolata dalle funzioni del foglio di calcolo. Il fisico, a differenza dello statistico, sa in anticipo quale forma deve avere la legge; i dati servono a stimarne i parametri e a controllare, attraverso i residui, che la legge valga davvero nell'intervallo delle misure: oltre un certo carico la molla si deforma in modo permanente, e la retta smette di funzionare.

SINTESI

- Il punteggio standardizzato $z = (x - \bar{x}) : \sigma$ dice di quante deviazioni standard un dato dista dalla media; il coefficiente di variazione $\sigma : \bar{x}$ confronta la variabilità di grandezze diverse. Con dati in classi si usano i valori centrali.
- 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 percentuale; 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 (totale di riga · totale di colonna) : N.
- L'indice χ^2 somma su tutte le caselle (osservata − teorica)^2 : teorica; vale 0 solo per l'indipendenza. L'indice V di Cramér lo normalizza tra 0 e 1.
- Il diagramma a dispersione rappresenta coppie di dati quantitativi; la covarianza è la media dei prodotti degli scarti e ne indica il verso.
- Il coefficiente di correlazione lineare $r = \text{cov}(x, y) : (\sigma_x \sigma_y)$ è compreso tra −1 e 1; $|r| = 1$ solo se i punti sono allineati. r misura soltanto il legame lineare.
- Correlazione non è causalità: un legame può dipendere da una variabile confondente (correlazione spuria).
- La retta dei minimi quadrati ha $m = \text{cov}(x, y) : \sigma_x^2$ e $q = \bar{y} - m\bar{x}$; passa per il baricentro e la somma dei suoi residui è zero.
- r^2 è la quota della varianza di y spiegata dalla retta. Le previsioni sono ragionevoli dentro l'intervallo dei dati (interpolazione), rischiose fuori (estrapolazione).
- Un campione casuale semplice evita la distorsione; un campione grande ma distorto sbaglia comunque. Per stimare la varianza della popolazione si divide per $n - 1$.
- Un modello matematico si sceglie guardando i residui: un modello esponenziale si adatta ai dati con la regressione su (t; ln y).

ESERCIZI

Comprendere

1. ● In una tabella a doppia entrata, la distribuzione marginale del carattere riportato sulle righe si legge:

   A. nella prima riga della tabella

   B. nella colonna dei totali di riga

   C. nella riga dei totali di colonna

   D. nella casella in basso a destra

2. ● Un coefficiente di correlazione $r = -0{,}95$ indica che:

   A. tra le due grandezze non c'è quasi nessun legame

   B. al crescere di una grandezza l'altra cresce quasi in proporzione

   C. i punti sono vicini a una retta con pendenza negativa

   D. una grandezza è la causa della diminuzione dell'altra

3. ● Due caratteri qualitativi sono statisticamente indipendenti se:

   A. le distribuzioni condizionate di uno dei due sono tutte uguali

   B. le due distribuzioni marginali sono uguali

   C. l'indice χ^2 vale 1

   D. nessuna frequenza congiunta è uguale a zero

4. ●● La retta di regressione di y su x passa sempre per il punto:

   A. O(0; 0)

   B. $(0; \overline{y})$

   C. $(\overline{x}; 0)$

   D. $(\overline{x}; \overline{y})$

5. ●● Si misura l'altezza di un gruppo di piante in centimetri e la loro massa in grammi, e si calcola $r = 0{,}72$. Se si esprimono le altezze in metri e le masse in kilogrammi, il coefficiente diventa:

   A. 0,0072

   B. 0,72

   C. 72

   D. dipende da quante sono le piante

6. ● Un sito di notizie chiede ai lettori di votare online se sono favorevoli a una nuova legge; votano in 40 000. Il campione:

   A. è rappresentativo, perché è molto grande

   B. è un campione casuale semplice dei lettori

   C. è probabilmente distorto, perché i votanti si sono scelti da sé

   D. ha un margine d'errore minore di un punto, quindi è affidabile

7. ●● Vero o falso? Correggi le affermazioni false. a) Se $r = 0$ tra le due grandezze non c'è alcuna relazione. b) $\chi^2 = 0$ se e solo se i due caratteri sono perfettamente indipendenti. c) La retta di regressione di y su x e quella di x su y coincidono sempre. d) Per gli stessi dati (non tutti uguali) la varianza campionaria s^2 è maggiore di σ^2.

8. ● Completa. a) La media dei prodotti degli scarti dalle medie di due caratteri si chiama _____ . b) La differenza tra l'ordinata osservata di un punto e quella prevista dalla retta di regressione si chiama _____ . c) Usare la retta di regressione per un valore di x esterno all'intervallo dei dati si chiama _____ . d) Il numero r^2 si chiama _____ e indica la quota della varianza di y spiegata dalla retta.

Calcolare

9. ● La massa degli zaini di 40 studenti (dati inventati) è raggruppata in classi: da 2 a 4 kg, 6 zaini; da 4 a 6 kg, 14; da 6 a 8 kg, 16; da 8 a 10 kg, 4. Con i valori centrali calcola la media, la deviazione standard e il coefficiente di variazione.

10. ● Nella verifica di matematica Sara ha preso 7,5, in una classe con media 6 e deviazione standard 1; in quella di latino ha preso 8, in una classe con media 7 e deviazione standard 1,25. In quale verifica il suo risultato è più notevole rispetto alla classe?

11. ● A 160 studenti (dati inventati) si chiede se praticano sport regolarmente e quante ore dormono di solito. Sportivi: 18 dormono meno di 7 ore, 48 da 7 a 8 ore, 34 più di 8 ore. Non sportivi: 22, 32 e 6 rispettivamente. a) Scrivi la tabella a doppia entrata con i totali. b) Scrivi le due distribuzioni marginali. c) Calcola le distribuzioni condizionate delle ore di sonno per sportivi e non sportivi, in percentuale, e confrontale.

12. ●● Per i dati dell'esercizio precedente calcola le frequenze teoriche, l'indice χ^2 e l'indice V di Cramér. Commenta.

13. ●● Una tabella a doppia entrata ha due righe (A, B) con totali 60 e 40 e tre colonne (C, D, E) con totali 30, 50 e 20. Completa la tabella in modo che i due caratteri siano perfettamente indipendenti, e verifica che le distribuzioni condizionate delle due righe sono uguali.

14. ● Per le coppie (1; 2), (2; 4), (3; 5), (4; 4), (5; 5) calcola le medie, le deviazioni standard e la covarianza.

15. ●● Per i dati dell'esercizio precedente calcola il coefficiente di correlazione e la retta di regressione di y su x.

16. ●● In un gruppo di ragazzi (dati inventati) l'altezza x ha media 170 cm e deviazione standard 8 cm, la massa y ha media 65 kg e deviazione standard 10 kg, e $r = 0{,}6$. a) Trova la retta di regressione della massa sull'altezza. b) Che massa prevede per un ragazzo alto 180 cm? c) Quale quota della varianza della massa è spiegata dall'altezza?

17. ●● Trova la retta di regressione di y su x per le coppie (2; 3), (4; 4), (6; 8), (8; 9), (10; 11). Calcola i residui e verifica che la loro somma è zero.

18. ● I dati 4, 6, 7, 9, 9 sono un campione estratto da una popolazione più grande. Calcola σ^2 (con n) e s^2 (con $n - 1$). Quale dei due useresti per stimare la varianza della popolazione?

19. ● Per quattro coppie di grandezze si sono trovati i coefficienti di correlazione $r = 0{,}9$, $r = -0{,}7$, $r = 0{,}3$ e $r = -0{,}2$. a) Ordina i legami dal più forte al più debole. b) Per ciascuno calcola r^2 e scrivi quale quota della varianza di y spiega la retta di regressione. c) In quali casi la retta ha pendenza negativa?

20. ●●● Una grandezza è stata misurata agli istanti $t = 0, 1, 2, 3, 4$ e vale $y = 3{,}1; 5{,}9; 12{,}2; 23{,}8; 48{,}5$. a) Calcola $\ln y$ per ogni dato. b) Trova la retta di regressione di $\ln y$ su t. c) Ricava un modello $y = A \cdot b^t$ e interpreta b.

Dimostrare e argomentare

21. ●● Dimostra che la somma dei residui della retta di regressione è zero, usando soltanto l'uguaglianza $q = \overline{y} - m\overline{x}$.

22. ●● Le coppie $(x_i; y_i)$ soddisfano tutte $y_i = ax_i + b$, con $a < 0$ e le x non tutte uguali. Dimostra che $r = -1$.

23. ●● Trova l'errore. Per la tabella sport–sonno dell'esercizio sulle ore di sonno, Marco calcola χ^2 usando, al posto delle frequenze, le percentuali sul totale (per esempio $18 : 160 = 11{,}25\%$ al posto di 18). Ottiene un numero diverso da quello corretto e conclude che «χ^2 non serve, perché cambia a seconda di come si scrivono i dati». Dove sbaglia?

24. ●● Considera i punti $(-2; 4)$, $(-1; 1)$, $(0; 0)$, $(1; 1)$, $(2; 4)$, che stanno sulla parabola $y = x^2$. Calcola la covarianza e spiega perché $r = 0$ anche se y è determinata da x.

25. ●●● Dimostra la «formula di calcolo» della covarianza: $\text{cov}(x, y) = \dfrac{\Sigma\, x_i y_i}{n} - \overline{xy}$. Usala per ricalcolare la covarianza delle coppie $(1; 2)$, $(2; 4)$, $(3; 5)$, $(4; 4)$, $(5; 5)$.

Problemi

26. ●● In laboratorio si appendono a una molla masse di 50, 100, 150, 200 e 250 g e si misurano gli allungamenti 2,1; 3,9; 6,2; 8,0; 9,8 cm. a) Trova la retta di regressione dell'allungamento sulla massa e il coefficiente di correlazione. b) Prevedi l'allungamento con 175 g. c) Ricava la costante elastica della molla, ricordando che l'allungamento è mg/k ($g = 9{,}81$ m/s^2), e trascurando il termine noto.

27. ●● Un bar prova a vendere una bibita a prezzi diversi in cinque settimane: con il prezzo di 1,00 €, 1,50 €, 2,00 €, 2,50 €, 3,00 € vende rispettivamente 120, 100, 85, 62, 48 bottiglie (dati inventati). a) Calcola r e la retta di regressione delle vendite sul prezzo. b) Quante bottiglie prevede il modello a 2,20 €? c) E a 5,00 €? Commenta.

28. ●● Il preside vuole sapere che cosa pensano i 900 studenti della scuola dell'orario delle lezioni. Valuta questi metodi per scegliere 90 studenti, dicendo quali sono campioni casuali semplici e quali rischiano una distorsione, e perché: a) intervistare i primi 90 studenti che entrano a scuola una mattina; b) estrarre 90 numeri a caso dall'elenco numerato di tutti gli studenti; c) mettere un questionario sul sito della scuola e usare le prime 90 risposte; d) scegliere le tre classi del corso A e intervistarne tutti gli studenti.

29. ●● In un'indagine su 200 studenti (dati inventati), 120 dichiarano di usare lo smartphone a letto prima di dormire: di loro, 45 giudicano buona la qualità del proprio sonno e 75 scarsa. Degli altri 80, 54 la giudicano buona e 26 scarsa. a) Scrivi la tabella a doppia entrata e calcola le percentuali di sonno «buono» nei due gruppi. b) Calcola χ^2 e V. c) Un giornale titola: «Lo smartphone a letto rovina il sonno». I dati lo dimostrano?

30. ●● Otto studenti misurano la propria altezza x e l'apertura delle braccia y, da punta a punta delle dita (cm): (158; 156), (163; 165), (167; 166), (170; 172), (172; 171), (176; 178), (181; 180), (185; 187). a) Calcola le medie e il coefficiente di correlazione. b) Trova la retta di regressione dell'apertura sull'altezza. c) Che apertura prevede per uno studente alto 178 cm?

Con il computer

31. ● Nel foglio di calcolo scrivi i dati della tabella delle ore di studio e dei voti nelle colonne A (ore) e B (voti), righe da 2 a 7. a) Calcola le medie con MEDIA, le deviazioni standard con DEV.ST.POP, il coefficiente di correlazione con CORRELAZIONE e i coefficienti della retta con PENDENZA e INTERCETTA, e confrontali con i risultati del capitolo. b) Disegna il grafico a dispersione e aggiungi la linea di tendenza lineare, con l'equazione e il valore di R^2. c) Calcola anche DEV.ST.C: perché dà un risultato diverso da DEV.ST.POP?

32. ●● Riproduci nel foglio di calcolo l'esempio della coltura di lieviti. a) Calcola in una colonna i valori previsti dal modello lineare e in un'altra quelli del modello esponenziale; poi i residui di entrambi e la somma dei loro quadrati. b) Nel grafico a dispersione aggiungi una linea di tendenza esponenziale: che equazione propone il programma? c) Quale modello scegli, e perché?

Verso la verifica

33. ●● A 170 studenti (dati inventati) si chiede se leggono almeno un libro al mese. In terza rispondono sì 30 e no 20; in quarta sì 24 e no 32; in quinta sì 16 e no 48. a) Calcola le distribuzioni marginali e le percentuali di «sì» in ciascuna classe. b) Calcola le frequenze teoriche, χ^2 e V. c) Si può concludere che studiare negli ultimi anni fa leggere di meno?

34. ●● Di 10 coppie di dati si conoscono soltanto le somme: $\Sigma x = 50$, $\Sigma y = 120$, $\Sigma x^2 = 290$, $\Sigma y^2 = 1520$, $\Sigma xy = 640$. Usando le formule di calcolo (la varianza è la media dei quadrati meno il quadrato della media) trova σ_x, σ_y, la covarianza, r e la retta di regressione di y su x.

35. ●●● **Sfida.** Siano m la pendenza della retta di regressione di y su x e m' quella della retta di regressione di x su y, $x = m'y + q'$. a) Dimostra che $m \cdot m' = r^2$. b) Deduci che le due rette coincidono se e solo se $|r| = 1$. c) Verifica a) con i dati delle ore di studio e dei voti.

PER APPROFONDIRE

- [FPP] Freedman, Pisani, Purves, *Statistics*: i capitoli su correlazione e regressione, con l'effetto di regressione di Galton, e quelli sui sondaggi, con la storia del *Literary Digest*.
- [HUF] Huff, *Mentire con le statistiche*: campioni distorti, grafici ingannevoli e correlazioni scambiate per cause, raccontati con ironia.
- [BOY] Boyer, *Storia della matematica*: Legendre, Gauss e il metodo dei minimi quadrati.

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
$r = \text{cov}(x, y) : (\sigma_x \sigma_y)$, numero puro compreso tra −1 e 1 che misura l'intensità e il verso del legame lineare tra due caratteri (Bravais-Pearson).

Coefficiente di determinazione
il numero r^2, quota della varianza di y spiegata dalla retta di regressione.

Correlazione spuria
correlazione tra due grandezze dovuta non a un legame diretto ma a una terza grandezza (variabile confondente) che influisce su entrambe.

Covarianza
media dei prodotti degli scarti dalle medie di due caratteri quantitativi; il suo segno indica il verso del legame.

Diagramma a dispersione
rappresentazione nel piano cartesiano di coppie di dati quantitativi, un punto per unità statistica.

Distorsione
errore sistematico di un campione che favorisce alcune unità della popolazione (campioni di comodo, autoselezione, mancate risposte); non diminuisce aumentando l'ampiezza.

Distribuzione condizionata
distribuzione di un carattere ristretta alle unità che presentano una data modalità dell'altro: una riga o una colonna della tabella, di solito in percentuale.

Distribuzione doppia
insieme delle frequenze congiunte di due caratteri osservati sulle stesse unità statistiche, raccolte in una tabella a doppia entrata.

Distribuzione marginale
distribuzione di uno solo dei due caratteri di una tabella a doppia entrata, letta nei totali di riga o di colonna.

Frequenza congiunta
numero delle unità statistiche che presentano insieme una modalità del primo carattere e una del secondo; è il contenuto di una casella della tabella a doppia entrata.

Frequenza teorica
frequenza che una casella avrebbe se i caratteri fossero indipendenti: (totale di riga · totale di colonna) : N.

Indice chi-quadrato
$\chi^2 = \Sigma\, (f - f')^2 : f'$, somma estesa a tutte le caselle; vale 0 solo per l'indipendenza e cresce con la dipendenza e con N.

Indice V di Cramér
$V = \sqrt{(\chi^2 : (N(k - 1)))}$, con k il minore tra numero di righe e di colonne; misura l'associazione tra 0 (indipendenza) e 1 (dipendenza perfetta).

Indipendenza statistica
relazione tra due caratteri le cui distribuzioni condizionate sono tutte uguali; equivale a frequenze osservate uguali alle frequenze teoriche.

Punteggio standardizzato
numero $z = (x - \text{media}) : \sigma$ che indica di quante deviazioni standard un dato si trova sopra o sotto la media della sua distribuzione.

Residuo
differenza tra il valore osservato di y e quello previsto dalla retta (o dal modello) in corrispondenza della stessa x.

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

Varianza campionaria
$s^2 = \Sigma (x - \bar{x})^2 : (n - 1)$, stima della varianza della popolazione calcolata su un campione.

Soluzioni

1. B. I totali di riga contano le unità con ciascuna modalità del carattere sulle righe, qualunque sia l'altro carattere.

2. C. Il segno dice che una grandezza tende a diminuire quando l'altra cresce; |r| vicino a 1 indica punti vicini a una retta.

3. A. È la definizione; equivale a dire che ogni frequenza osservata è uguale alla frequenza teorica.

4. D. Da $q = \overline{y} - m\overline{x}$ segue $\overline{y} = m\overline{x} + q$.

5. B. Un cambio di unità di misura moltiplica gli scarti di ciascun carattere per una costante positiva: covarianza e deviazioni standard cambiano nella stessa proporzione e r resta invariato.

6. C. Rispondono solo i lettori di quel sito che hanno voglia di farlo (autoselezione): l'ampiezza del campione non corregge la distorsione.

7. a) Falso: r = 0 esclude solo una relazione lineare; i punti possono seguire una parabola. b) Vero: ogni termine della somma è un quadrato diviso per un numero positivo, e la somma è nulla solo se ogni frequenza osservata è uguale a quella teorica. c) Falso: coincidono solo se |r| = 1. d) Vero: stesso numeratore positivo, denominatore n − 1 < n.

8. a) covarianza; b) residuo; c) estrapolazione; d) coefficiente di determinazione.

9. Valori centrali 3, 5, 7, 9 kg. $\overline{x}$ = (18 + 70 + 112 + 36) : 40 = 236 : 40 = 5,9 kg. Scarti −2,9; −0,9; 1,1; 3,1: σ^2 = (8,41 · 6 + 0,81 · 14 + 1,21 · 16 + 9,61 · 4) : 40 = 119,6 : 40 = 2,99; σ ≈ 1,73 kg; CV ≈ 1,73 : 5,9 ≈ 0,29 = 29%.

10. Matematica: z = (7,5 − 6) : 1 = 1,5. Latino: z = (8 − 7) : 1,25 = 0,8. Il risultato di matematica è più notevole: 1,5 deviazioni standard sopra la media contro 0,8, anche se il voto è più basso.

11. a) Righe sportivi (18, 48, 34; totale 100) e non sportivi (22, 32, 6; totale 60); totali di colonna 40, 80, 40; N = 160. b) Sport: sì 100, no 60. Sonno: meno di 7 ore 40, da 7 a 8 ore 80, più di 8 ore 40. c) Sportivi: 18%, 48%, 34%. Non sportivi: 22 : 60 ≈ 36,7%; 32 : 60 ≈ 53,3%; 6 : 60 = 10%. Le distribuzioni sono diverse: tra gli sportivi è molto più alta la quota di chi dorme più di 8 ore (34% contro 10%) e più bassa quella di chi dorme meno di 7 ore.

12. Frequenze teoriche: sportivi 100 · 40 : 160 = 25, 100 · 80 : 160 = 50, 25; non sportivi 15, 30, 15. χ^2 = $7^2/25 + 2^2/50 + 9^2/25 + 7^2/15 + 2^2/30 + 9^2/15$ = 1,96 + 0,08 + 3,24 + 3,27 + 0,13 + 5,4 ≈ 14,08. La tabella è 2 × 3, quindi k − 1 = 1: V = √(14,08 : 160) ≈ 0,30. Associazione debole-moderata: il sonno dipende in parte dalla pratica sportiva (o da altri fattori legati a essa).

13. N = 100. Riga A: 60 · 30 : 100 = 18, 60 · 50 : 100 = 30, 60 · 20 : 100 = 12. Riga B: 12, 20, 8. Percentuali di riga: A 18 : 60 = 30%, 30 : 60 = 50%, 12 : 60 = 20%; B 12 : 40 = 30%, 20 : 40 = 50%, 8 : 40 = 20%: uguali.

14. $\overline{x}$ = 15 : 5 = 3, $\overline{y}$ = 20 : 5 = 4. Scarti di x: −2, −1, 0, 1, 2; di y: −2, 0, 1, 0, 1. $\Sigma(x - \overline{x})^2 = 10$, $\sigma_x^2 = 2$, $\sigma_x \approx 1{,}41$; $\Sigma(y - \overline{y})^2 = 6$, $\sigma_y^2 = 1{,}2$, $\sigma_y \approx 1{,}10$. Prodotti degli scarti: 4, 0, 0, 0, 2; cov(x, y) = 6 : 5 = 1,2.

15. r = 6 : √(10 · 6) = 6 : √60 ≈ 0,775. m = cov : σ_x^2 = 1,2 : 2 = 0,6; q = 4 − 0,6 · 3 = 2,2. Retta: y = 0,6x + 2,2.

16. a) m = r · σ_y : σ_x = 0,6 · 10 : 8 = 0,75 kg/cm; q = 65 − 0,75 · 170 = −62,5. Retta y = 0,75x − 62,5. b) y = 0,75 · 180 − 62,5 = 72,5 kg. c) r^2 = 0,36: il 36%.

17. $\overline{x}$ = 6, $\overline{y}$ = 7. Scarti di x: −4, −2, 0, 2, 4; di y: −4, −3, 1, 2, 4. Σ prodotti = 16 + 6 + 0 + 4 + 16 = 42; $\Sigma(x - \overline{x})^2$ = 40. m = 42 : 40 = 1,05; q = 7 − 1,05 · 6 = 0,7. Retta y = 1,05x + 0,7. Valori previsti 2,8; 4,9; 7; 9,1; 11,2. Residui 0,2; −0,9; 1; −0,1; −0,2: somma 0.

18. $\overline{x}$ = 35 : 5 = 7. Quadrati degli scarti 9, 1, 0, 4, 4: somma 18. σ^2 = 18 : 5 = 3,6; s^2 = 18 : 4 = 4,5. Per stimare la varianza della popolazione si usa s^2 = 4,5.

19. a) |r| decrescente: 0,9; −0,7; 0,3; −0,2. b) r^2 = 0,81; 0,49; 0,09; 0,04: la retta spiega l'81%, il 49%, il 9% e il 4% della varianza. c) Per r = −0,7 e r = −0,2, perché m = r · σ_y : σ_x ha il segno di r.

20. a) $\ln y \approx 1{,}131; 1{,}775; 2{,}501; 3{,}170; 3{,}882$. b) Media di t: 2; media di $\ln y \approx 2{,}492$; $\Sigma(t-2)^2 = 10$; $\Sigma(t-2)(\ln y - \text{media}) \approx 6{,}895$. $m \approx 0{,}690$, $q \approx 1{,}113$: $\ln y \approx 0{,}690t + 1{,}113$. c) $A = e^{1{,}113} \approx 3{,}04$, $b = e^{0{,}690} \approx 1{,}99$: $y \approx 3{,}04 \cdot 1{,}99^t$. La grandezza raddoppia circa a ogni unità di tempo.

21. $\Sigma e_i = \Sigma(y_i - m x_i - q) = \Sigma y_i - m \Sigma x_i - nq = n\bar{y} - mn\bar{x} - n(\bar{y} - m\bar{x}) = 0$. c.v.d.

22. $\bar{y} = a\bar{x} + b$, quindi $y_i - \bar{y} = a(x_i - \bar{x})$. Allora $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = a\,\Sigma(x_i - \bar{x})^2$ e $\Sigma(y_i - \bar{y})^2 = a^2\,\Sigma(x_i - \bar{x})^2$. Con $D = \Sigma(x_i - \bar{x})^2 > 0$: $r = aD : \sqrt{(D \cdot a^2 D)} = aD : (|a|D) = a : |a| = -1$, perché $a < 0$. c.v.d.

23. χ^2 si calcola sulle frequenze assolute: dipende dal numero N delle osservazioni (raddoppiando tutte le frequenze raddoppia). Usando percentuali Marco ha di fatto posto N = 100 invece di 160, e ha ottenuto χ^2 moltiplicato per 100 : 160 ≈ 8,80. Proprio per questo si usa l'indice V, che divide per N e non dipende dalla scala: con le percentuali $V = \sqrt{(8{,}80 : 100)} \approx 0{,}30$, lo stesso valore.

24. $\bar{x} = 0$, $\bar{y} = 10 : 5 = 2$. Scarti di y: 2, −1, −2, −1, 2. Prodotti con gli scarti di x (−2, −1, 0, 1, 2): −4, 1, 0, −1, 4; somma 0, quindi cov = 0 e r = 0. I punti a sinistra del vertice (che scendono) e quelli a destra (che salgono) danno contributi opposti che si annullano: il legame esiste ma non è lineare, e r misura solo il legame lineare.

25. Sviluppando il prodotto: $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - \bar{y}\Sigma x_i - \bar{x}\Sigma y_i + n\bar{x}\bar{y} = \Sigma x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} = \Sigma x_i y_i - n\bar{x}\bar{y}$, perché $\Sigma x_i = n\bar{x}$ e $\Sigma y_i = n\bar{y}$. Dividendo per n si ottiene la formula. c.v.d. Per i dati: $\Sigma x_i y_i = 2 + 8 + 15 + 16 + 25 = 66$; cov = 66 : 5 − 3 · 4 = 13,2 − 12 = 1,2.

26. a) Media delle masse 150 g, media degli allungamenti 6 cm. Scarti delle masse −100, −50, 0, 50, 100; scarti degli allungamenti −3,9; −2,1; 0,2; 2; 3,8. Σ prodotti = 390 + 105 + 0 + 100 + 380 = 975; Σ quadrati delle masse = 25 000; Σ quadrati degli allungamenti = 38,1. m = 975 : 25 000 = 0,039 cm/g; q = 6 − 0,039 · 150 = 0,15 cm. Retta: y = 0,039x + 0,15; r = 975 : √(25 000 · 38,1) ≈ 0,999. b) y = 0,039 · 175 + 0,15 ≈ 6,97 cm. c) 0,039 cm/g = 0,39 m/kg è g/k, quindi k = 9,81 : 0,39 ≈ 25,2 N/m.

27. a) Prezzo medio 2 €, vendite medie 83. Scarti del prezzo −1; −0,5; 0; 0,5; 1; scarti delle vendite 37, 17, 2, −21, −35. Σ prodotti = −37 − 8,5 + 0 − 10,5 − 35 = −91; Σ quadrati del prezzo = 2,5; Σ quadrati delle vendite = 3328. r = −91 : √(2,5 · 3328) ≈ −0,998; m = −91 : 2,5 = −36,4; q = 83 + 36,4 · 2 = 155,8. Retta: y = −36,4x + 155,8. b) y = −36,4 · 2,2 + 155,8 = 75,7, circa 76 bottiglie. c) y = −36,4 · 5 + 155,8 = −26,2: un numero negativo di bottiglie è assurdo. È un'estrapolazione: fuori dall'intervallo dei prezzi provati il legame non può restare lineare.

28. a) Campione di comodo: chi arriva presto (per esempio chi abita vicino o usa un certo autobus) può avere opinioni diverse sull'orario. b) Campione casuale semplice: ogni gruppo di 90 studenti ha la stessa probabilità di essere estratto. c) Autoselezione: risponde chi ha un'opinione forte o frequenta il sito. d) Non è casuale: un solo corso può avere un orario o abitudini particolari; le unità di una stessa classe non sono indipendenti.

29. a) Righe: smartphone sì (45, 75; 120), no (54, 26; 80); colonne buono 99, scarso 101; N = 200. Sonno buono: 45 : 120 = 37,5% contro 54 : 80 = 67,5%. b) Frequenze teoriche: 120 · 99 : 200 = 59,4; 60,6; 39,6; 40,4. $\chi^2 = 14{,}4^2 \cdot (1/59{,}4 + 1/60{,}6 + 1/39{,}6 + 1/40{,}4) \approx 17{,}28$ (in una tabella 2 × 2 tutte le differenze valgono ±14,4); $V = \sqrt{(17{,}28 : 200)} \approx 0{,}29$. c) No: i dati mostrano un'associazione, non una causa. Potrebbe essere il contrario (chi dorme male prende il telefono perché non riesce a dormire), o una variabile confondente (ansia, orari, carico di studio). Per una relazione di causa servirebbe un esperimento controllato.

30. a) $\bar{x} = 1372 : 8 = 171{,}5$ cm; $\bar{y} = 1375 : 8 = 171{,}875$ cm. $\Sigma(x - \bar{x})^2 = 570$; $\Sigma(y - \bar{y})^2 = 666{,}875$; Σ prodotti = 607,5. $r = 607{,}5 : \sqrt{(570 \cdot 666{,}875)} \approx 0{,}985$. b) m = 607,5 : 570 ≈ 1,066; q = 171,875 − m · 171,5 ≈ −10,91. Retta: y ≈ 1,066x − 10,91. c) y ≈ 178,8 cm. L'apertura delle braccia è, in questo gruppo, molto vicina all'altezza.

31. a) MEDIA: 8 e 6,5; DEV.ST.POP: $\sqrt{(70 : 6)} \approx 3{,}42$ e $\sqrt{(21{,}5 : 6)} \approx 1{,}89$; CORRELAZIONE ≈ 0,902; PENDENZA 0,5; INTERCETTA 2,5. b) La linea di tendenza è y = 0,5x + 2,5 con $R^2 \approx 0{,}814$ (= $r^2 = 35/43$). c) DEV.ST.C divide per n − 1 = 5: $\sqrt{(70 : 5)} \approx 3{,}74$ e $\sqrt{(21{,}5 : 5)} \approx$

2,07. È la deviazione standard campionaria s, adatta se i sei studenti sono un campione di una popolazione più grande.

32. a) Somma dei quadrati dei residui: circa 2,04 per il modello lineare, circa 0,09 per quello esponenziale. b) Il programma calcola la tendenza esponenziale proprio con la regressione su (t; ln y): propone $y \approx 2{,}02e^{0{,}26x}$, cioè $2{,}02 \cdot 1{,}30^x$. c) Il modello esponenziale: residui molto più piccoli e senza andamento regolare, e coerenza con la crescita per divisione cellulare.

33. a) Classi: 50, 56, 64. Risposte: sì 70, no 100; N = 170. Percentuali di «sì»: terza 60%, quarta 24 : 56 ≈ 42,9%, quinta 25%. b) Frequenze teoriche (sì; no): terza 20,59; 29,41; quarta 23,06; 32,94; quinta 26,35; 37,65. $\chi^2 \approx 14{,}29$; k = 2, $V = \sqrt{(14{,}29 : 170)} \approx 0{,}29$: associazione debole-moderata. c) No: i dati mostrano un'associazione, non una causa. Potrebbero contare il maggior carico di studio, ma anche il fatto che i gruppi sono studenti diversi (non gli stessi seguiti negli anni), con abitudini diverse.

34. $\overline{x} = 5$, $\overline{y} = 12$. $\sigma_x^2 = 29 - 25 = 4$, $\sigma_x = 2$; $\sigma_y^2 = 152 - 144 = 8$, $\sigma_y = 2\sqrt{2} \approx 2{,}83$; cov = 64 − 60 = 4. $r = 4 : (2 \cdot 2\sqrt{2}) = 1 : \sqrt{2} \approx 0{,}707$. m = 4 : 4 = 1; q = 12 − 5 = 7. Retta: y = x + 7.

35. a) $m = cov : \sigma_x^2$, $m' = cov : \sigma_y^2$, quindi $m \cdot m' = cov^2 : (\sigma_x^2 \sigma_y^2) = r^2$. b) Entrambe le rette passano per $G(\overline{x}; \overline{y})$, quindi coincidono se e solo se hanno la stessa pendenza nel piano xy. Se r = 0 sono la retta $y = \overline{y}$ e la retta $x = \overline{x}$, diverse. Se $r \neq 0$, la seconda si scrive y = x : m′ + costante, con pendenza 1 : m′: le rette coincidono se e solo se m = 1 : m′, cioè $m \cdot m' = 1$, cioè $r^2 = 1$. c) m = 35 : 70 = 0,5; m′ = 35 : 21,5 = 70 : 43; $m \cdot m' = 35 : 43 = r^2 \approx 0{,}814$.