978-88-08-18484

A conclusione di questo paragrafo, accenniamo brevemente al fatto che se il campione non è numeroso, la stima della varianza di popolazion tramite la varianza campionaria può essere molto imprecisa, e dunqu è necessario analizzarla in modo più approfondito. Lo statistico ingles W.S. Gossett (1876–1937), che firmava il suo lavoro scientifico con l pseudonimo "Student", ha mostrato che la variabile aleatoria

$$t = \frac{m_N - \mu}{s_N/\sqrt{N}},$$

in cui sia m_N che s_N sono variabili che dipendono dal campione, h una precisa distribuzione di probabilità, che, in suo onore, è detta t d **Student**. In modo analogo a quello con cui si opera con le gaussiane, s possono determinare le probabilità con cui $|t| \geq \ell$, e di conseguenza g intervalli di confidenza.

Rimandiamo a testi più specializzati per i dettagli (*vedi* [4]).

12.4 Ipotesi statistiche

Figura 12.27
Molte delle decisioni che prendiamo dipendono spesso da aspetti caratteriali, che ci portano a fidarci o a diffidare delle informazioni che abbiamo. La statistica inferenziale ci mette invece a disposizione un insieme di tecniche rigorose per valutare dati ed eventi.

La raccolta e l'analisi dei dati relativi a un fenomeno procede di pari pas so con la formulazione di teorie interpretative. È fondamentale, dunqu poter comprendere se la teoria che stiamo formulando sia in accordo co i dati. La statistica inferenziale studia le tecniche di verifica delle ipote che formuliamo sul fenomeno.

Per cominciare, mostriamo un esempio semplice.

Esempio 12.4.1 Una moneta truccata?

In un gioco con una moneta si vince se esce testa, si perde se esce croce. L'orga nizzatore del gioco garantisce che la moneta utilizzata non è truccata e quindi ch $P(T) = P(C) = 0.5$. Prima di giocare, stiamo un po' a guardare e vediamo che su 2 lanci, T esce solo 6 volte, un numero un po' basso rispetto al valore $20\,P(T) = 1$ che corrisponde alla media teorica.

Ci chiediamo se l'organizzatore ci stia ingannando o se il valore che osserviamo si un ragionevole frutto del caso. La risposta che ci diamo è certamente opinabile in molti casi, dipende più dal nostro carattere (sospettoso o fiducioso) che da da oggettivi. È possibile, invece, utilizzare una procedura statistica per decidere sull correttezza dell'organizzatore. Questa procedura prende il nome di **test statistico** ha come scopo quello di verificare se il dato osservato sia probabilisticamente credibil assumendo che la moneta non sia truccata. Illustriamo passo per passo il ragionament che si deve fare.

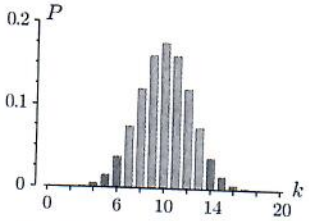

Figura 12.28
Istogramma della distribuzione binomiale di 20 lanci con probabilità di successo 0.5. I valori in rosso corrispondono a $k \leq 6$ e $k \geq 14$. Il totale della probabilità di risultati così "sospetti" è dell'12%. Aver osservato 6 volte T non è in disaccordo con l'ipotesi che la moneta non sia truccata.

Occorre valutare quanto sia credibile che, in 20 lanci di una moneta non truccata, risultato T esca solo 6 volte.

La distribuzione binomiale ci permette di calcolare la probabilità di questo event Indicando infatti con X il numero di volte che esce T in 20 lanci, si ha

$$P(X = 6) = \binom{20}{6} \frac{1}{2^{20}} \approx 0.037.$$

Dunque il risultato è decisamente poco probabile: meno del 4%. Questo valore, per non ci autorizza a trarre nessuna conclusione. Infatti tutti i valori di $P(X = k)$ co $k = 0, 1, \ldots, 20$ sono piuttosto piccoli (fig. 12.28), in particolare la probabilità d risultato ottimale $X = 10$ è solo $P(X = 10) \approx 0.18$.

978-88-08-18484-9

Inoltre su 20 lanci di una moneta non truccata, ci aspettiamo **circa** 10 uscite di T, ma non necessariamente 10.

Per rispondere al nostro dubbio dobbiamo ragionare in modo un po' più profondo, e calcolare la probabilità, nel caso di moneta non truccata, di osservare un risultato "sospetto" o "estremo" quanto o più di quello che abbiamo osservato, cioè 6 volte T. In questo senso, i possibili risultati "estremi" sono $X = 0, 1, 2, 3, 4, 5, 6$, e sono altrettanto estremi i risultati $X = 14, 15, 16, 17, 28, 19, 20$, infatti 6 e 14 distano entrambi 4 dal valor medio 10 (fig. 12.28). La probabilità complessiva di questi risultati è

$$p = P(X = 0) + \ldots + P(T = 6) + P(T = 14) + \ldots + P(T = 20) \approx 0.12\,.$$

Questo risultato ci dice che, assumendo la moneta non truccata, la probabilità di ottenere risultati estremi come $X = 6$ o più estremi ancora è del 12%. In altre parole, osservando molte volte il lancio di 20 monete non truccate, nel 12% dei casi ci troveremmo a osservare risultati dubbi come e più di quello che abbiamo osservato. Poiché 12% è una percentuale piuttosto elevata, non abbiamo forti indizi che suggeriscano che la moneta sia truccata.

La stessa procedura utilizzata nel caso in cui avessimo osservato $X = 6$ volte T su 30 lanci (fig. 12.29), avrebbe dato un valore di probabilità molto molto piccolo; infatti si ha

$$p = P(X = 0) + \ldots + P(X = 6) + P(X = 24) + \ldots + P(X = 30) \approx 0.0014\,.$$

In questo caso saremmo stati fortemente autorizzati a sospettare che la moneta fosse truccata.

■

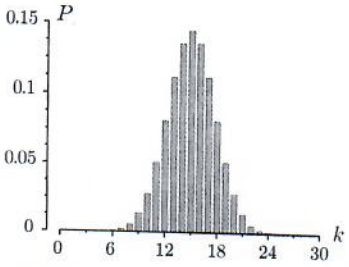

Figura 12.29
Istogramma della distribuzione binomiale per 30 lanci con probabilità di successo 0.5. La probabilità complessiva dei valori $k \leq 6$ e $k \geq 24$ è 1.4 millesimi. Aver osservato 6 volte T su 30 lanci non è in accordo con l'ipotesi che la moneta non sia truccata.

Possiamo introdurre ora qualche definizione generale.

Un **test statistico** è una procedura che serve a verificare se un dato è in accordo con una teoria e si articola nei seguenti passi:

- si formula l'ipotesi da verificare, indicata convenzionalmente con H_0 e chiamata **ipotesi nulla**;
- ipotizzando che H_0 sia vera, si calcola la probabilità p di ottenere un risultato estremo quanto o più di quello effettivamente osservato (p prende il nome di **valore p del test**, *p–value* in inglese);
- se il valore p è troppo piccolo, si rifiuta l'ipotesi H_0, se è grande la si accetta.

Nell'esempio precedente, che descrive un **test binomiale**, l'ipotesi nulla H_0 è che la moneta non sia truccata, e il valore p del test è risultato 0.12, abbastanza grande da non farci dubitare della validità di H_0.

Nella pratica statistica, i valori di p che conducono ad accettare l'ipotesi nulla H_0 o a rifiutarla (cioè a considerarla falsa), sono fissati per convenzione.

Se $p \geq 0.05$, la discrepanza tra il dato osservato e il valore atteso **non è statisticamente significativa**: se è vera l'ipotesi nulla, la discrepanza è un probabile effetto casuale del campionamento.
In questo caso l'ipotesi nulla viene accettata.

978-88-08-18484-9

Se $p < 0.05$, H_0 viene, in genere, rifiutata e in particolare:

- se $0.01 \leq p < 0.05$, la discrepanza tra dato osservato e valore atteso è detta **statisticamente significativa**;
- se $0.001 \leq p < 0.01$, la discrepanza è detta **molto significativa**;
- se $p < 0.001$, la discrepanza è detta **estremamente significativa**.

I valori 0.05, 0.01, 0.001 sono detti **livelli di significatività** del test.

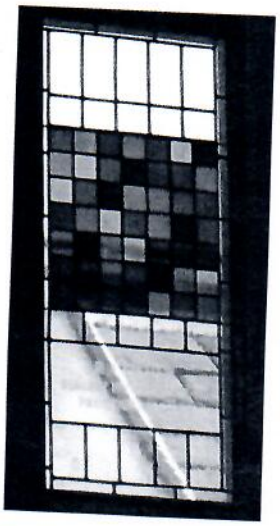

Figura 12.30
Vetrata artistica commemorativa di Ronald Aylmer Fisher (1890–1962) al Gonville and Caius College di Cambridge. L'opera, disegnata da Maria McClafferty e istallata nel 1989, rappresenta un quadrato latino 7x7, una figura sperimentale usata in statistica. Fisher, genetista e statistico inglese, rivoluziona i metodi della ricerca scientifica, mostrando come sottoporre a rigorose procedure di verifica la corrispondenza tra dati e teorie mediante test statistici.

Sempre con riferimento all'esempio 12.4.1, l'evento $X = 6$ su 20 lanci non è statisticamente significativo ($p = 0.12 > 0.05$), mentre l'evento $X = 6$ su 30 lanci è statisticamente molto significativo ($0.001 < p < 0.01$).

● **ATTENZIONE:** Sottolineiamo che il valore p del test non è la probabilità che H_0 sia vera, ma la probabilità del verificarsi di eventi estremi, assumendo che H_0 sia vera: il valore di p è dunque un livello di **fiducia** del test.

Esempio 12.4.2 Una moneta un po' truccata

La conclusione che abbiamo tratto dal test nel caso dei 20 lanci dell'esempio 12.4.1 non ci permette di rifiutare l'ipotesi nulla, cioè che la moneta non sia truccata o, equivalentemente, che sia $P(T) = 0.5$. Vediamo, però, che nemmeno ci permette di escludere che sia "un po' truccata". Assumiamo infatti come ipotesi nulla H_0 che sia $P(T) = 0.52$ (cioè T sia un po' più probabile di C). Su 20 lanci il valore atteso delle uscite T è $20 \cdot 0.52 = 10.4$, dunque i risultati estremi sono quelli inferiori o uguali a 6 e quelli superiori o uguali a 15. Il valore p è in questo caso

$$p = \sum_{i=0}^{6} \binom{20}{i} 0.52^i \, 0.48^{20-i} + \sum_{i=15}^{20} \binom{20}{i} 0.52^i \, 0.48^{20-i} \approx 0.07 \, .$$

Poiché $p > 0.05$ dobbiamo accettare anche l'ipotesi nulla che la moneta sia un po' truccata.

Consideriamo ora l'ipotesi nulla $P(T) = 0.6$. Il valore atteso di X è $20 \cdot 0.6 = 12$, e dunque i risultati estremi sono $X \in \{0, 1, 2, 3, 4, 5, 6, 18, 19, 20\}$. La probabilità complessiva di questi risultati è

$$p = \sum_{i=0}^{6} \binom{20}{i} 0.6^i \, 0.4^{20-i} + \sum_{i=18}^{20} \binom{20}{i} 0.6^i \, 0.4^{20-i} \approx 0.01 \, ,$$

che indica uno scostamento significativo dal valore atteso. In questo caso siamo dunque portati a rifiutare l'ipotesi $P(T) = 0.6$. ■

Questo esempio mostra che esistono valori di $P(T)$, vicini a 0.5, che non possono essere smentiti dall'osservazione $X = 6$.

In generale, fissato il livello di significatività, per esempio 5%, si chiama **intervallo di confidenza** al 5% di fiducia, l'insieme dei valori dei parametri dell'ipotesi nulla che sono compatibili con il dato osservato. Nel caso particolare del lancio della moneta, il valore q è nell'intervallo di confidenza al 5% se non si può rifiutare l'ipotesi $P(T) = q$ con questo livello di fiducia.

978-88-08-18484-9

Il calcolo dell'intervallo fiduciario non è immediato, soprattutto se utilizziamo la formula della distribuzione binomiale. Per risolvere il problema ci viene in aiuto ancora una volta il teorema del limite centrale, infatti la variabile X che conta il numero di T su N lanci ha media qN e varianza $Nq(1-q)$ (*vedi* par. 11.2), dunque la variabile $Z = (X - qN)/\sqrt{Nq(1-q)}$ è approssimativamente gaussiana con media nulla e varianza 1.
Supponiamo di osservare $X = k$ successi. In termini della variabile Z questo equivale a osservare il valore $(k - qN)/\sqrt{Nq(1-q)}$. Questo valore si chiama **statistica del test** e va confrontato con i possibili valori di Z (questo test si chiama infatti **Z–test**).
Il valore p del test è dunque approssimativamente pari alla probabilità che Z disti da 0 più di $|k - qN|/\sqrt{Nq(1-q)}$: cioè

$$p = P\left(|Z| \geq \frac{|k - qN|}{\sqrt{Nq(1-q)}}\right) = \operatorname{erf}\left(\frac{|k - qN|}{\sqrt{2Nq(1-q)}}\right),$$

dove abbiamo usato la (11.17) e la funzione erf è la funzione degli errori definita nella (9.8). Per ottenere il valore estratto è necessario l'utilizzo di opportuni programmi che permettano il calcolo della funzione erf o della funzione di distribuzione per la gaussiana standardizzata. Però, per molti scopi, è sufficiente confrontare il valore $|k - qN|/\sqrt{Nq(1-q)}$ con i valori standard 1.96, 2.58, 3.29 per poter trarre conclusioni sulla significatività dell'osservazione.
Questa stessa approssimazione gaussiana permette di calcolare gli intervalli di confidenza del valore q utilizzando formule analoghe alle (12.6).

Si supponga di osservare una frequenza $\bar{q}$ su un campione di N elementi. Gli intervalli di confidenza per la probabilità di successo q sono

$$\left(\bar{q} - 1.96\sqrt{\frac{\bar{q}(1-\bar{q})}{N}}, \bar{q} + 1.96\sqrt{\frac{\bar{q}(1-\bar{q})}{N}}\right) \quad \text{al } 95\%,$$
$$\left(\bar{q} - 2.58\sqrt{\frac{\bar{q}(1-\bar{q})}{N}}, \bar{q} + 2.58\sqrt{\frac{\bar{q}(1-\bar{q})}{N}}\right) \quad \text{al } 99\%. \tag{12.7}$$

Vediamo un esempio in cui gli intervalli fiduciari hanno un ruolo importante.

Esempio 12.4.3 Sondaggi

Un campione di 100 persone viene intervistato sulle intenzioni di voto a un referendum. Risulta che 42 persone voteranno "no" e 58 persone voteranno "sì".
Il valore osservato della proporzione di "sì" è dunque $\bar{q} = 0.58$. La deviazione standard nell'approssimazione gaussiana è $\sqrt{\bar{q}(1-\bar{q})/N} \approx 0.05$, dunque, utilizzando le (12.7), possiamo affermare che la percentuale q di "sì" verifica

$$q \in (0.48, 0.67) \quad \text{al } 95\%,$$
$$q \in (0.45, 0.71) \quad \text{al } 99\%.$$

978-88-08-18484-9

Questo sondaggio dunque non dà certo risposte conclusive sul risultato del referendum.
Nelle situazioni reali, il numero di intervistati è sempre maggiore di 1000; con lo stesso dato che abbiamo considerato, si otterrebbe in questo caso un'ampiezza di 0.06 per l'intervallo di confidenza al 95%.
Per esercizio, verificare questa affermazione e calcolare anche quanto dovrebbe essere grande N per essere sicuri al 99% che vinceranno i "sì" se si osserva una frequenza del 51% sul campione. ■

A seconda delle grandezze da studiare sono stati ideati vari test statistici. In particolare, se il modello probabilistico riguarda più variabili contemporaneamente, non si può utilizzare uno Z–test, che funziona per analizzare ipotesi su una sola variabile. È il caso delle distribuzioni multinomiali del tipo descritto nella (11.2), per le quali il test più utilizzato è il **test di adattamento del χ^2** ("chi quadro") o **test di Pearson** (dal nome dello statistico inglese Karl Pearson (1857–1936), che lo introdusse). In questo test la statistica (di cui illustreremo il calcolo tra breve) viene confrontata con la variabile aleatoria χ^2. Quest'ultima è definita come la somma dei quadrati di $g \geq 1$ variabili gaussiane standard. Il numero g è chiamato **numero dei gradi di libertà**, perché è il numero delle variabili aleatorie indipendenti che entrano nella definizione di χ^2.
In figura 12.31 mostriamo la densità di probabilità delle variabili χ^2 per diversi valori di g.
Illustriamo i dettagli di questo test con un esempio.

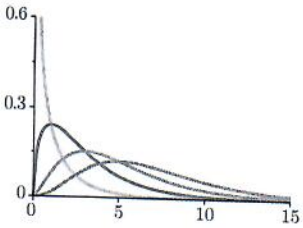

Figura 12.31
Densità di probabilità della variabile χ^2, al variare dei gradi di libertà $g = 1, 3, 5, 7$, dalla curva più alta in arancione alla più schiacciata in verde.

Esempio 12.4.4 Test statistici per i risultati di Mendel – I

Incrociando tra loro piante ottenute da due linee pure, a "fiore rosso" e a "fiore bianco", Mendel osservò 705 piante di piselli a fiore rosso e 224 a fiore bianco.
La prima legge di Mendel (*vedi* es. 10.2.15) stabilisce che la proporzione tra piante di fenotipo dominante (rosso) e piante di fenotipo recessivo (bianco) deve essere 3 : 1, mentre in questo caso, il rapporto è $705/224 \approx 3.15$.
Verifichiamo se i valori ottenuti siano in accordo con l'ipotesi nulla che la probabilità di ottenere piante a fiore rosso sia 3/4 e che la probabilità di ottenere piante a fiore bianco sia 1/4.
Per rispondere a questa domanda potremmo usare un test binomiale, come nell'esempio 12.4.1, oppure usare un'approssimazione gaussiana; mostriamo invece come si costruisce in questo caso il test del χ^2.
Abbiamo due dati: numero delle piante a fiore rosso $R = 705$, e numero di quelle a fiore bianco $B = 224$. La dimensione del campione è $N = R + B = 705 + 224 = 929$. il valore atteso per R è $\langle R \rangle = 929 \cdot 3/4 = 696.75$, per B è $\langle B \rangle = 929 \cdot 1/4 = 232.25$.
Questi risultati e gli scarti tra i dati ottenuti e i valori attesi sono riassunti nella tabella seguente.

	R	B
Dati	705	224
Valori attesi	696.75	232.25
Scarti	8.25	−8.25

La statistica del test si ottiene sommando i quadrati degli scarti dopo averli divisi per i corrispondenti valori attesi:

$$\chi^2 = \frac{8.25^2}{696.75} + \frac{(-8.25)^2}{232.25} \approx 0.39\,.$$

È ragionevole ritenere che, se questo numero è piccolo, vi è un buon accordo con l'ipotesi nulla; se invece il valore è grande, questa ipotesi va rifiutata.

La questione è dunque come stabilire se 0.39 sia un valore grande o piccolo. In questo problema stiamo confrontando valori osservati e valori attesi di due variabili, B e R, che però non sono indipendenti l'una dall'altra, infatti, fissato R, il valore di B è $N - R$. Per questo motivo, la variabile di confronto χ^2 va considerata a un solo grado di libertà, perché c'è realmente una sola variabile indipendente.
Utilizzando un calcolatore, si scopre che la probabilità che una variabile χ^2 a un grado di libertà sia maggiore o uguale a 0.39 è $p \approx 0.53$ (fig. 12.32). Il valore p del test è dunque molto alto e non abbiamo motivo di dubitare della validità dell'ipotesi nulla.
In mancanza di strumenti elettronici che permettano il calcolo del valore p del test, si possono utilizzare opportune tavole che riportano i valori del test per i diversi livelli di confidenza; un esempio del contenuto di queste tavole è riportato in tabella 12.1. Anche da qui si vede che il valore 0.39 è ben al di sotto del valore 2.71 che corrisponde al livello di significatività del 10%.
In un altro esperimento Mendel contò 428 piante a baccello verde e 152 piante a baccello giallo. Verificare che, anche in questo caso, l'ipotesi che valga la prima legge di Mendel non può essere confutata con questi dati. ■

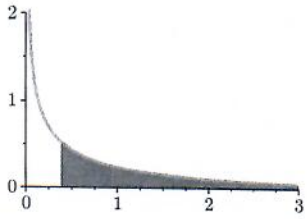

Figura 12.32
Densità di probabilità della variabile χ^2 a un grado di libertà. La probabilità che la variabile venga estratta con valore maggiore di 0.39 è l'area della regione rossa, che vale circa 0.53, cioè, se è vera l'ipotesi nulla, c'è una probabilità superiore al 50% di osservare dati che si discostano più di 0.39.

Descriviamo in generale l'uso del test del χ^2.

Il **test del** χ^2 confronta l'accordo (o adattamento) tra frequenza osservata e frequenza attesa di dati organizzati in n categorie qualitative.
Supponiamo di estrarre da una popolazione un campione di dimensione N e di osservare nel campione le frequenze $F_1, F_2, \ldots, F_n$. Se le frequenze relative delle diverse categorie sono $q_1, q_2, \ldots, q_n$ (ipotesi nulla), i valori attesi delle frequenze sono $E_i = Nq_i$.
La statistica del test è il numero

$$\chi^2 = \sum_{i=1}^{n} \frac{(F_i - E_i)^2}{E_i} .$$

Per valutare se l'ipotesi nulla sia da accettare, bisogna fare riferimento alla legge di distribuzione della variabile χ^2 a $n - 1$ gradi di libertà.

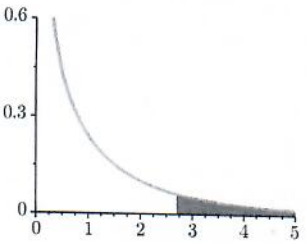

Figura 12.33
Densità di probabilità della variabile χ^2 a un grado di libertà. La probabilità che la variabile venga estratta con valore maggiore di 2.71 è l'area della regione rossa, che vale circa 0.05. Se il valore del test è inferiore a 2.71 l'ipotesi nulla si rifiuta con una significatività del 5%.

Il valore p del test può essere ottenuto utilizzando opportuni programmi, oppure i valori tabulati del χ^2. In particolare, se si richiedono i livelli di significatività del 10%, del 5%, dell'1% e dello 0.1%, si può fare riferimento alla tabella 12.1, che riporta i valori x per i quali si ha $P(\chi^2 > x) = q$, dove q è la significatività (fig. 12.33).

Tabella 12.1
Valori della variabile χ^2 in corrispondenza dei più usati valori della significatività.

g	10%	5%	1%	0.1%
1	2.71	3.84	6.63	10.83
2	4.61	5.99	9.21	13.82
3	6.25	7.81	11.34	16.27
4	7.78	9.49	13.28	18.47
5	9.24	11.07	15.00	20.52
6	10.64	12.59	16.81	22.46
7	12.02	14.07	18.48	24.32
8	13.36	15.51	20.10	26.12
9	14.68	16.92	21.67	27.88

Ricordiamo che, se si considerano n categorie, il numero di gradi di libertà è $n - 1$, infatti, poiché la somma delle frequenze è fissata, i parametri indipendenti che possono variare sono solo $n - 1$. Nella tabella

978-88-08-18484-9

dell'esempio 12.4.4 si nota, in particolare, che la somma dei due scarti è 0, dunque, se varia un dato, deve variare anche l'altro in modo che la somma sia fissata: la variabile indipendente è una sola, quindi il numero di gradi di libertà è 1.

Esempio 12.4.5 Test del χ^2 per gli effetti di una mutazione

Un certo gene di una pianta esiste, allo stato selvatico, in una sola variante, indicata con A. Viene creato artificialmente l'allele mutato a e si ipotizza che la presenza di questo nuovo allele renda maggiore la probabilità di germinazione del seme.
Per verificare questa ipotesi si incrociano tra loro alcune piante eterozigoti Aa e si osserva se effettivamente le piante Aa o aa sono più frequenti.
Su un campione di 800 piante si osservano 221 piante di genotipo aa, 410 piante di genotipo Aa, e 169 piante di tipo AA. Valutiamo quanto questi dati siano statisticamente significativi, cioè quanto siano improbabili, supponendo che sia verificata l'ipotesi nulla che l'allele non influenzi la germinabilità.
Ricordando che $P(AA) = P(aa) = 1/4$ e che $P(Aa) = 1/2$, le frequenze teoriche dei genotipi, nell'ipotesi nulla, sono rispettivamente $800/4 = 200$ per gli omozigoti e $800/2 = 400$ per gli eterozigoti. La statistica del test è

$$\chi^2 = \frac{(221-200)^2}{200} + \frac{(410-400)^2}{400} + \frac{(169-200)^2}{200} = 7.26\,.$$

Confrontando questo valore con i valori in tabella 12.1 della variabile χ^2 a due gradi di libertà, si vede che è superiore a 5.99, che corrisponde al livello 5%. Dunque la discrepanza dai valori attesi è statisticamente significativa e si rifiuta l'ipotesi nulla se la significatività è fissata al 5%, la si accetta se la significatività è fissata all'1%. ■

Il test del χ^2 è usato anche per la verifica dell'indipendenza tra variabili.

Esempio 12.4.6 Test statistici per i risultati di Mendel – II

Incrociando piante eterozigoti sia per il carattere "seme liscio/seme rugoso" sia per il carattere "seme giallo/seme verde" Mendel ottenne i seguenti risultati:

	Liscio	Rugoso	Totale
Giallo	315	101	416
Verde	108	32	140
Totale	423	133	556

(in statistica una tabella di questo tipo prende il nome di **tabella di contingenza**).
Si possono sottoporre questi dati a un test di adattamento, per verificare se siano nelle proporzioni predette dalla seconda legge di Mendel $9:3:3:1$ (*vedi* es. 10.2.16).
È più interessante però sottoporre a test direttamente la domanda importante, se cioè i geni che governano questi caratteri siano indipendenti, senza supporre nulla sulla legge con cui sono estratti.
Secondo questi dati, la frequenza dei semi gialli è $416/556$, mentre quella dei semi lisci è $423/556$. Se fossero indipendenti, la probabilità di trovare un seme giallo e liscio dovrebbe essere il loro prodotto, e dunque la frequenza attesa sarebbe

$$556 \cdot \frac{416}{556} \cdot \frac{423}{556} \approx 316.5\,.$$

Procedendo analogamente si ottengono gli altri valori attesi, che riassumiamo nella tabella seguente.

	Liscio	Rugoso	Totale
Giallo	316.5	99.5	416
Verde	106.5	33.5	140
Totale	423	133	556

Come si vede le frequenze sono praticamente uguali a quelle misurate da Mendel. Valutiamo statisticamente l'indipendenza, determinando il valore del χ^2. Tutti gli scarti al quadrato sono pari a $1.5^2 = 2.25$, dunque il valore del test è

$$\chi^2 = \frac{2.25}{316.5} + \frac{2.25}{106.5} + \frac{2.25}{99.5} + \frac{2.25}{33.5} \approx 0.12\,.$$

Visto che in una tabella di contingenza possiamo modificare liberamente solo un elemento senza modificare i totali sulla riga e sulla colonna (se modifichiamo, infatti, il primo elemento in alto a sinistra, sono necessariamente fissati sia i valori sulla prima riga che quelli sulla prima colonna e da questi valori si ottiene anche il valore sulla seconda riga e seconda colonna), questo valore va confrontato con il χ^2 a un grado di libertà.
Il valore del test è in effetti piccolissimo, molto al di sotto del valore corrispondente alla significatività del 10%. ■

Consideriamo un altro esempio.

Esempio 12.4.7 Efficacia di un farmaco

Una casa farmaceutica vuole testare l'efficacia di un principio attivo, e prepara tre farmaci: V_1 che non contiene il principio attivo, V_2 che lo contiene in una certa quantità e V_3 che lo contiene in quantità doppia rispetto a V_2. Questi preparati vengono sperimentati su pazienti affetti dalla stessa malattia e ritenuti confrontabili per quel che riguarda il loro stato di salute. Si osservano i seguenti risultati della sperimentazione:

	V_1	V_2	V_3	Totale
Pazienti migliorati	12	5	29	46
Pazienti non migliorati	114	80	90	284
Totale	126	85	119	330

Valutiamo se questi dati sono compatibili con l'ipotesi nulla che il farmaco non faccia effetto, cioè che le due righe di dati siano indipendenti.
La tabella dei valori attesi è

	V_1	V_2	V_3	Totale
Pazienti migliorati	17.6	11.8	16.6	46
Pazienti non migliorati	108.4	73.2	102.4	284
Totale	126	85	119	330

Il calcolo del test dà 17.4. Questo valore va confrontato con il χ^2 a due gradi di libertà, infatti si possono modificare liberamente solo due elementi della tabella, tenendo fissi i totali (in generale per tabelle $n \times m$ i gradi di libertà sono $(n-1)(m-1)$).

Figura 12.34
I test di significatività, come quello del χ^2, sono fondamentali per testare l'efficacia dei farmaci.

978-88-08-18484-9

Figura 12.35
Rita Hayworth nella locandina di Gilda, un film statunitense del 1946. Bisogna sempre distinguere tra un'affermazione statistica di correlazione tra variabili e rapporti causa–effetto. Per esempio, una vecchia indagine americana su un campione femminile mostrò l'esistenza di una correlazione positiva tra l'avere il cancro ai polmoni e il portare calze di seta. Una prima grossolana interpretazione potrebbe portare all'assurda conclusione che portare le calze di seta possa contribuire all'insorgere della malattia tumorale. In realtà, portare calze di seta e fumare sigarette erano abitudini corrispondenti a un medesimo modello di comportamento sociale, ed è il fumo di sigarette ad essere in rapporto causa–effetto con il cancro ai polmoni.

Questo valore è più grande di quello che corrisponde al livello di confidenza del 99%, dunque possiamo dire che i risultati osservati sono molto significativi e ci inducono a ritenere che il farmaco abbia effetto. ■

12.5 Correlazione tra variabili

Fin qui abbiamo discusso casi in cui si misura una sola variabile relativa ai dati, ma, in genere, in statistica si analizzano contemporaneamente più variabili, e può essere importante capire se siano legate (correlate) tra loro.

Esempio 12.5.1 Altezza e peso

Consideriamo di nuovo il campione di 120 studentesse di cui abbiamo descritto le altezze nell'esempio 12.1.5. Nel grafico in figura 12.36 abbiamo riportato, per ogni studentessa, i valori X dell'altezza in centimetri in ascissa, e del peso Y in chilogrammi in ordinata. Il valore medio dell'altezza è $m_X = 166.2$, quello del peso è $m_Y \approx 56.2$. Il punto di ascissa m_X e di ordinata m_Y, è il baricentro dei dati.
Nonostante non ci sia una relazione funzionale tra il valore del peso e il valore dell'altezza, vi è evidentemente una relazione tra queste due quantità, infatti, al crescere dell'altezza si osservano pesi maggiori. ■

In questo paragrafo studieremo come si possa stabilire una relazione tra due variabili statistiche attraverso la definizione di opportuni indici. Supponiamo di effettuare un campionamento e di prendere i dati relativi alla coppia di variabili X, Y: alla misura X_1 della variabile X corrisponde la misura Y_1 della variabile Y; a X_2 corrisponde Y_2 e così via. Indicheremo con m_X e m_Y le medie delle due variabili, e con s_X^2 e s_Y^2 le rispettive varianze campionarie. Possiamo chiederci se, all'aumentare dei valori di una distribuzione, quelli dell'altra corrispondentemente aumentino o diminuiscano. In altre parole, vorremmo sapere se i dati X e Y variano congiuntamente, cioè se siano o meno **covarianti**. Per stimare quantitativamente se esista una relazione fra i dati delle due distribuzioni si utilizza un indice detto, appunto, **covarianza**.

Data una popolazione di dimensione N per la quale si misurino i dati $X_1, \ldots, X_N$ e $Y_1, \ldots, Y_N$ relativi a due variabili, si definisce **covarianza** il numero

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^{N} (X_i - m_X)(Y_i - m_Y).$$

Se i dati $X_1, \ldots, X_N$ e $Y_1, \ldots, Y_N$ sono un campione estratto da una popolazione, il numero

$$s_{XY} = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - m_X)(Y_i - m_Y)$$

si chiama **covarianza campionaria**.

Si noti che la covarianza è una grandezza analoga alla varianza, in particolare si ha $\sigma_{XX} = \sigma_X^2$.

978-88-08-18484-9

Nell'espressione della covarianza i termini nella somma sono i prodotti delle deviazioni dalla media delle due variabili; questi prodotti hanno un segno, dunque σ_{XY} può essere positiva, negativa o nulla.
Come nella figura 12.36, possiamo immaginare il piano cartesiano diviso in 4 quadranti centrati nel punto (m_X, m_Y):

$$\text{I} = \{x \geq m_X,\ y \geq m_Y\}, \qquad \text{III} = \{x < m_X,\ y < m_Y\},$$
$$\text{II} = \{x < m_X,\ y \geq m_Y\}, \qquad \text{IV} = \{x \geq m_X,\ y < m_Y\}.$$

Se un punto appartiene al I o al III quadrante, il prodotto $(X_i - m_x)(Y_i - m_y)$ è positivo, altrimenti è negativo. Se al crescere di x i valori di y aumentano, il valore della covarianza s_{XY} sarà positivo. In particolare la covarianza dei dati di altezza e peso dell'esempio 12.5.1 rappresentati in figura 12.36 è circa 16.5. Se invece al crescere di x i valori di y diminuiscono, la covarianza sarà negativa, infine se i punti formano una "nuvola" abbastanza uniformemente sparpagliata intorno al baricentro dei dati, allora σ_{XY} sarà prossima a zero.
Consideriamo un semplice esempio numerico.

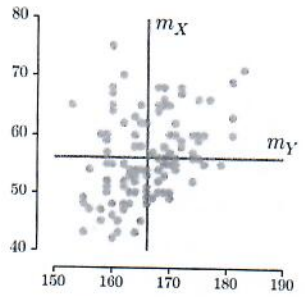

Figura 12.36
Altezza e peso di un campione di 120 studentesse. La linea rossa verticale corrisponde alla media delle altezze, quella orizzontale corrisponde alla media dei pesi. Il punto (m_X, m_Y) coincide con il baricentro dei dati.

Esempio 12.5.2 Covarianza

Consideriamo i dati di una (piccola) popolazione relativi a due variabili statistiche:

(A)	X	1	2	3	0	4	3
	Y	2	3	3	1	2	4

Calcoliamo le medie delle due variabili: si ha $m_X = 13/6$ e $m_Y = 5/2$.
In figura 12.37 mostriamo i corrispondenti punti $P_1 = (1,2)$, $P_2 = (2,3)$, $P_3 = (3,3)$, $P_4 = (0,1)$, $P_5 = (4,2)$ e $P_6 = (3,4)$, che hanno come ascisse e ordinate i valori di X e di Y, e le rette di equazione $x = m_X = 13/6$ e $y = m_Y = 5/2$. Calcoliamo la covarianza: si ha, per definizione,

$$\sigma_{XY} = \frac{1}{6}\left[\left(1-\frac{13}{6}\right)\left(2-\frac{5}{2}\right) + \left(2-\frac{13}{6}\right)\left(3-\frac{5}{2}\right) + \left(3-\frac{13}{6}\right)\left(3-\frac{5}{2}\right)\right.$$
$$\left. + \left(-\frac{13}{6}\right)\left(1-\frac{5}{2}\right) + \left(4-\frac{13}{6}\right)\left(2-\frac{5}{2}\right) + \left(3-\frac{13}{6}\right)\left(4-\frac{5}{2}\right)\right] = \frac{3}{4} > 0.$$

I punti sono distribuiti prevalentemente nei quadranti I e III, infatti la covarianza è positiva. Consideriamo ora altri dati relativi a due variabili

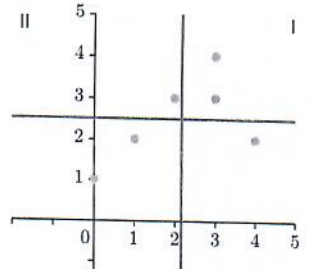

Figura 12.37
I dati (A) sono per la maggior parte nel I e nel III quadrante. La correlazione è positiva.

(B)	X	3	0	1	4	2	0
	Y	1	1	3	1	0	4

Le medie valgono ora $m_X = 5/3$ e $m_Y = 5/3$. Il calcolo della covarianza dà $\sigma_{XY} = -10/9$, che è negativa, in accordo con la rappresentazione grafica (fig. 12.38), in cui si nota che quasi tutti i dati sono nel II e nel IV quadrante.
In questo caso, al crescere dei dati X_i, le Y_i corrispondenti diminuiscono.
Consideriamo, come ultimo caso, i dati seguenti.

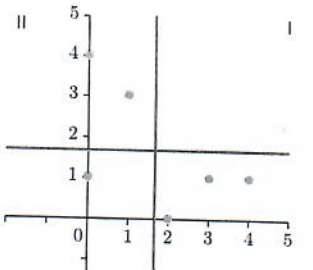

Figura 12.38
I dati (B) sono per la maggior parte nel II e nel IV quadrante. La correlazione è negativa.

(C)	X	1	0	4	2	3	2
	Y	4	1	2	3	4	1

Le medie sono $m_X = 2$ e $m_Y = 5/2$ e la distribuzione dei punti nel piano è quella mostrata in figura 12.39. La covarianza è $1/3$, un valore più piccolo dei precedenti, in accordo con il fatto che i dati sono dispersi in tutte le direzioni.
Per esercizio, studiare la covarianza dei dati

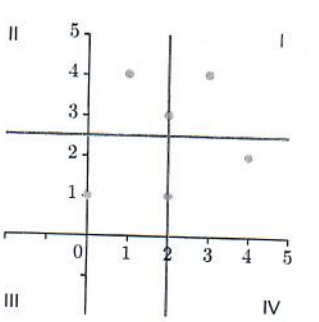

Figura 12.39
I dati (C) sono distribuiti in tutti i quadranti. La correlazione è piccola.

978-88-08-18484-9

X	2	1	0	4	3	2
Y	0	2	4	1	1	5

e spiegare il risultato che si ottiene. Verificare, inoltre, che si ha $\sigma_{XY} = \sigma_{YX}$.

Si può dimostrare (ma non lo faremo) che la covarianza varia in un insieme limitato di valori, infatti soddisfa le seguenti disuguaglianze

$$-\sigma_X\sigma_Y \leq \sigma_{XY} \leq \sigma_X\sigma_Y, \tag{12.8}$$

dove σ_X e σ_Y sono le deviazioni standard delle distribuzioni di X e di Y. Queste disuguaglianze suggeriscono di introdurre un nuovo indice per misurare la variazione congiunta di due collezioni di dati.

Si chiama **coefficiente di correlazione** di due distribuzioni di dati X e Y il numero

$$\rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

Dalla (12.8) si ottiene che

$$-1 \leq \rho \leq 1.$$

La massima correlazione positiva è $\rho = 1$, la massima correlazione negativa è $\rho = -1$, mentre non vi è correlazione se $\rho \approx 0$.

Il coefficiente di correlazione è un indice che misura meglio della covarianza la relazione tra i dati. La covarianza infatti ha come unità di misura il prodotto delle unità di misura di X e Y, dunque cambiare una delle due ne modifica il valore. Al contrario, il coefficiente di correlazione è un numero puro, infatti sia il numeratore che il denominatore della frazione che lo esprime hanno la stessa unità di misura.

Esempio 12.5.3 Correlazione

Consideriamo di nuovo i dati dell'esempio 12.5.2. Le varianze di X e di Y dei dati (A) sono $\sigma_X^2 = 65/36$ e $\sigma_Y^2 = 11/12$. Il coefficiente di correlazione è

$$\rho = \frac{\sigma_{XY}}{\sqrt{\sigma_X\sigma_Y}} \approx 0.58.$$

Le varianze di X e di Y dei dati (B) sono $\sigma_X^2 = 20/9$ e $\sigma_Y^2 = 17/9$. Il coefficiente di correlazione dei dati è $\rho \approx -0.54$. I dati dei due campioni hanno praticamente lo stesso valore di correlazione, ma nel primo caso la correlazione è positiva, nel secondo è negativa.

Le varianze di X e di Y dei dati (C) sono $\sigma_X^2 = 5/3$ e $\sigma_Y^2 = 19/12$. Il coefficiente di correlazione dei dati è $\rho \approx 0.21$. Questo valore è decisamente più piccolo dei precedenti, a conferma che in questo caso le variabili sono meno correlate.

Consideriamo infine i dati

(D)	X	1	0	2	4
	Y	0	−1.5	1.5	4.5

Le medie sono $m_X = 7/4$ e $m_Y = 9/8$. Se rappresentiamo i punti in un grafico è immediatamente evidente che sono allineati, e si può verificare che appartengono alla retta di equazione $y = 1.5x - 1.5$ (fig. 12.40). Calcoliamo le varianze e la covarianza. Risulta

$$\sigma_X^2 = \frac{35}{16}, \quad \sigma_Y^2 = \frac{315}{64}, \quad \sigma_{XY} = \frac{105}{32},$$

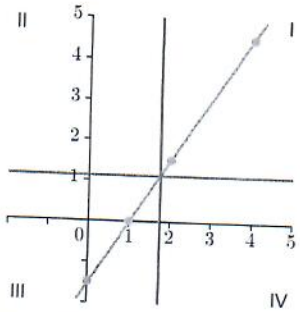

Figura 12.40
Se i dati sono allineati il coefficiente di correlazione è 1 o −1.

quindi si ha

$$\rho = \frac{105}{32}\sqrt{\frac{16}{35}\frac{64}{315}} = 1\,.$$

cioè punti allineati hanno la massima correlazione possibile, cioè una dipendenza lineare tra i dati.

Per esercizio si scelgano 3 punti sulla retta di equazione $y = -2x + 2$ e si mostri che il coefficiente di correlazione tra i valori di X e di Y è -1. ■

L'allineamento dei punti è la caratteristica che contraddistingue la massima correlazione possibile.

La massima correlazione positiva, $\rho = 1$, si ha se i valori delle due variabili sono in relazione lineare crescente, cioè se $Y_i = aX_i + b$, con $a > 0$.

La massima correlazione negativa, $\rho = -1$, si ha se $Y_i = aX_i + b$, con $a < 0$.

Figura 12.41
Esemplari di pulci d'acqua (*Daphnia obtusa*).

Esempio 12.5.4 Dafnie – I

Le popolazioni di *Daphnia obtusa*, un cladocero di acqua dolce detto anche "pulce d'acqua" (fig. 12.41), vivono, in genere, in pozze temporanee come piccoli stagni, fontane o laghetti, che spesso nella stagione calda sono soggetti a secche improvvise. La possibilità di sopravvivenza della popolazione in questi habitat così estremi dipende in gran parte delle modalità riproduttive. Questi organismi, oltre a riprodursi per partenogenesi, si riproducono anche sessualmente, generando uova "dormienti", dotate di una corazza (carapace) di protezione. Queste uova possono sopravvivere anche nel sedimento degli stagni asciutti, in attesa di condizioni più favorevoli alla schiusa. Le modalità, i tempi e gli stimoli ambientali che inducono la produzione di uova dormienti non sono ancora del tutto chiariti.

I dati nella tabella che segue sono relativi a una pozza sul litorale laziale. Le colonne indicano, da sinistra a destra, il volume V d'acqua, misurato in metri cubi, la temperatura T, misurata i gradi Celsius, e il pH; l'ultima colonna invece è la numerosità D, in migliaia, delle femmine portatrici di uova dormienti (che, in genere, sono due per ogni femmina).

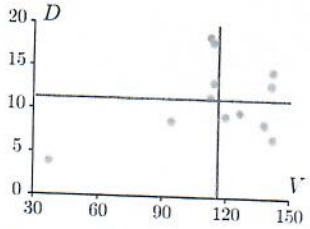

Figura 12.42
Correlazione tra volume V e numerosità D (in migliaia).

V	T	pH	D
37.2	18.0	5.79	3.845
94.0	9.8	6.19	8.673
112.1	8.4	5.97	18.587
112.1	11.0	6.05	11.499
137.9	7.2	6.12	8.502
141.2	5.8	6.04	12.993
141.8	5.8	6.46	14.604
141.8	6.0	6.46	6.827
126.2	7.2	6.53	9.764
119.5	12.0	6.42	9.315
113.5	9.8	6.40	17.934
114.0	12.9	6.88	13.223

Si può stabilire una correlazione tra il numero delle femmine portatrici di uova dormienti e qualcuno dei parametri ambientali?

Si può sicuramente supporre che ci sia una correlazione positiva tra la quantità d'acqua presente nella pozza e il numero di organismi, visto che, in mancanza d'acqua, la sopravvivenza è impossibile. Inoltre, poiché alte temperature possono indicare una secca in arrivo, sarà interessante analizzare la correlazione tra questa variabile e la numerosità D.

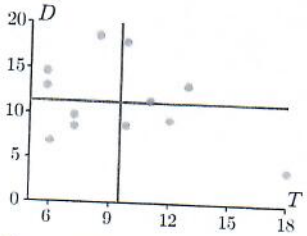

Figura 12.43
Correlazione tra temperatura T e numerosità D (in migliaia).

978-88-08-18484-9

Le medie e le deviazioni standard delle tre variabili sono $m_V \approx 115.94$, $\sigma_V \approx 27.83$, $m_T \approx 9.49$, $\sigma_T \approx 3.46$ e $m_D \approx 11.31$, $\sigma_D \approx 4.20$.
La covarianza tra volume e numerosità vale $\sigma_{VD} \approx 43.69$, quindi il coefficiente di correlazione è $\rho_{VD} \approx 0.37$. Questo numero indica una correlazione positiva, anche se non molto grande, tra il volume e la numerosità delle femmine con uova dormienti.
La covarianza tra temperatura e numerosità vale $\sigma_{TD} \approx -5.10$, quindi il coefficiente di correlazione è $\rho_{TD} \approx -0.35$: se la temperatura aumenta, diminuiscono gli organismi. Anche in questo caso non c'è una grande correlazione.
Per esercizio studiare la correlazione tra il numero di organismi e il pH. ■

La massima correlazione $\rho = \pm 1$ si ottiene solo in casi ideali in cui vi è l'allineamento perfetto. Nei casi pratici, anche se esiste una relazione funzionale lineare tra le variabili, le misure conterranno comunque un certo grado di imprecisione e difficilmente potranno essere rappresentate come punti di una retta. In questi casi è importante trovare la retta "migliore" che descrive la distribuzione dei dati. Così come per tutti gli indici statistici, esistono varie definizioni di "retta migliore", e la scelta dipende dal tipo di problema. Ciò nonostante, la retta più usata è quella che prende il nome di **retta di regressione** e viene determinata attraverso il **metodo dei minimi quadrati**, che ricerca la retta per la quale le deviazioni quadratiche in ordinata tra punti e retta siano le minime possibili.
Vediamo, in un caso molto semplice, come si opera concretamente.

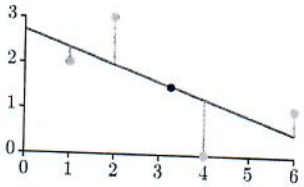

Figura 12.44
La retta dei minimi quadrati (in rosso) è, tra tutte le rette possibili, quella che rende minima la somma dei quadrati degli scarti (i segmenti viola). La retta dei minimi quadrati passa per il baricentro (in nero).

Esempio 12.5.5 Retta di regressione

Consideriamo i dati

(E)	X	1	6	2	4
	Y	2	1	3	0

ai quali corrispondono i punti, non allineati, $P_1 = (1,2)$, $P_2 = (6,1)$, $P_3 = (2,3)$ e $P_4 = (4,0)$ (fig. 12.44).
Vogliamo determinare una retta $y = ax + b$ che "descriva il meglio possibile" la relazione tra i dati. Una tale retta prevede per $x = X_1$ il valore $Y_1' = a \cdot X_1 + b = a + b$, per $x = X_2$ il valore $Y_2' = a \cdot X_2 + b = 6a + b$ e così via. Nella seguente tabella riportiamo gli scarti tra i valori previsti e i valori effettivi.

Ascissa	1	6	2	4
Scarto	$a+b-2$	$6a+b-1$	$2a+b-3$	$4a+b$

La retta di regressione è la retta che rende minima la somma dei quadrati degli scarti, cioè la funzione

$$Q(a,b) = (a+b-2)^2 + (6a+b-1)^2 + (2a+b-3)^2 + (4a+b)^2 .$$

Si tratta di una funzione di due variabili, perché il suo valore cambia sia se cambia il coefficiente angolare a della retta, sia se cambia il termine noto b. Nel paragrafo 8.4 abbiamo mostrato che i minimi di funzioni in più variabili sono raggiunti nei punti in cui le derivate parziali sono tutte nulle. Usiamo questa proprietà per determinare i valori di a e b che rendono minima $Q(a,b)$.

$$\frac{\partial Q}{\partial a} = 2[(a+b-2) + 6(6a+b-1) + 2(2a+b-3) + 4(4a+b)] = 0,$$
$$\frac{\partial Q}{\partial b} = 2[(a+b-2) + (6a+b-1) + (2a+b-3) + (4a+b)] = 0 .$$

Questo sistema di due equazioni in due incognite si risolve facilmente e la soluzione è $a = -22/59$, $b = 160/59$. Con questa scelta di a e b, si ha $Q \approx 2.95$. Ogni altra scelta di a e b dà valori maggiori.

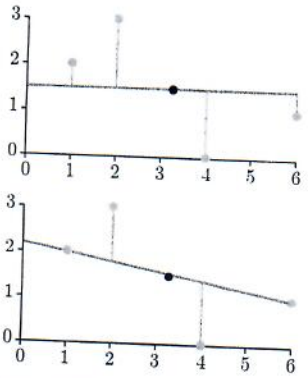

Figura 12.45
Sia per la retta $y = 3/2$ (primo grafico) che per la retta $y = (11 - x)/5$ (secondo grafico) la somma dei quadrati degli scarti è maggiore rispetto al caso della retta di regressione.

Per esercizio si calcoli $Q(a,b)$ nel caso della retta $y = 3/2$ e nel caso della retta $y = (11 - x)/5$ (fig. 12.45). ■

Formalizziamo questo procedimento.

Siano $X_1, \ldots, X_N$ e $Y_1, \ldots, Y_N$ i valori di due variabili statistiche. La **retta di regressione** ha equazione

$$y = \frac{\sigma_{XY}}{\sigma_X^2} x + \left(m_Y - \frac{\sigma_{XY}}{\sigma_X^2} m_X \right) . \tag{12.9}$$

L'espressione (12.9) appare più complessa di quanto in realtà non sia. L'equazione della retta di regressione, infatti, si determina usando il fatto che il coefficiente angolare a verifica

$$a = \frac{\sigma_{XY}}{\sigma_X^2}$$

e che la retta passa per il baricentro (m_X, m_Y), dunque

$$am_X + b = m_Y \implies b = m_Y - am_X .$$

Si noti inoltre che la retta può essere scritta in modo suggestivo usando il coefficiente di correlazione:

$$\frac{y}{\sigma_Y} = \rho \left(\frac{x}{\sigma_X} - \frac{m_X}{\sigma_X} \right) + \frac{m_Y}{\sigma_Y},$$

cioè, il coefficiente di correlazione è proprio il coefficiente angolare della retta dei minimi quadrati se si usano come unità di misura rispettivamente σ_X per le x e σ_Y per le y.
Scrivere l'equazione di una retta di regressione è utile nei casi concreti, perché, assegnati i valori della variabile X, è possibile prevedere, anche se non con certezza, quali saranno i corrispondenti valori della variabile Y.

Esempio 12.5.6 Dafnie – II

Scriviamo l'equazione della retta di regressione relativa ai dati di temperatura e numerosità dell'esempio 12.5.6, utilizzando i valori già calcolati per le medie, le varianze e la covarianza. Si ottiene

$$a = \frac{\sigma_{TD}}{\sigma_T^2} \approx -0.42506, \quad b = m_D - am_T \approx 15.3484 .$$

Usando questa relazione funzionale, possiamo prevedere, per esempio, che quando la temperatura sarà $T = 15$ °C il corrispondente numero di organismi sarà un valore intorno a $15\,a + b \approx 8.972$ migliaia, mentre per $T = 20$ °C sarà un valore intorno a 6847. Per $T = -b/a \approx 36.1$ °C la legge lineare prevede popolazione nulla. Ciò è in accordo con quanto si osserva, visto che le dafnie non sopravvivono a temperature troppo elevate.
Per esercizio, utilizzando i dati dell'esempio 12.5.4, si determini la retta di regressione di D su T e si calcolino i valori che questa retta prevede per $V = 20$ e per $V = 150$. ■

In tutti questi esempi, non abbiamo discusso dei metodi che permettono di quantificare il buon adattamento della legge di regressione ai dati. Per questo argomento rimandiamo a testi di statistica più specialistici.

978-88-08-18484-9

Il metodo di regressione lineare permette di considerare anche dipendenze esponenziali (e a potenza) tra i dati, infatti, come abbiamo visto negli esempi 6.2.14 e 6.2.15 un fenomeno descritto da una legge esponenziale si descrive con una legge lineare, se si usa la scala logaritmica. Questa osservazione permette di utilizzare il calcolo della retta di regressione anche per determinare eventuali leggi esponenziali che leghino le variabili. Infatti, in tal caso, il logaritmo della variabile Y deve dipendere linearmente da X. Vediamo un esempio.

Esempio 12.5.7 Regressione per l'esponenziale

Consideriamo i dati

(F)	X	1	2	3	5	6
	Y	5	9	11	35	48

I valori di Y crescono molto più rapidamente di quelli di X, si può quindi ipotizzare una dipendenza esponenziale di Y da X.
Per ricondurci al caso lineare, consideriamo, per ogni valore di y, il suo logaritmo in base 10, $y' = \log y$. La legge da determinare con i minimi quadrati è la forma della dipendenza lineare di $y' = ax + b$ da x, in accordo con i dati X e $\log Y$.
I valori di $\log Y$ sono

$$0.6990,\ 0.9542,\ 1.0414,\ 1.5441,\ 1.6812.$$

Gli indici statistici di cui abbiamo bisogno sono

$$m_X = 3.4,\quad m_{\log Y} \approx 1.1840,\quad \sigma_X^2 = 4.3,\quad \sigma_{X\ \log Y} \approx 0.6823.$$

Il coefficiente angolare della retta di regressione di $\log Y$ su X è $a = \sigma_{X\ \log Y}/\sigma_X^2 \approx 0.1983$; il valore di b è $b \approx -0.5096$. Dall'espressione della retta di regressione possiamo ottenere la legge esponenziale che lega la variabile Y a X:

$$\log y = 0.1983x - 0.5096 \implies y = 10^{0.1983x - 0.5096} \approx 0.31 \cdot 10^{0.1983x}.$$

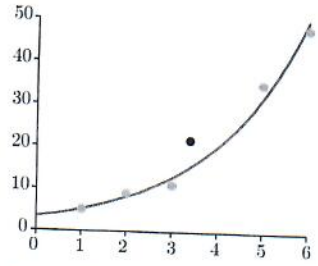

Figura 12.46
Regressione esponenziale. Si noti che la curva approssima bene i dati, ma non passa per il baricentro (in nero).

Gli stessi ragionamenti possono essere applicati anche al caso in cui i dati siano connessi da una relazione a potenza del tipo $y = ax^\beta$, in cui è necessario utilizzare la scala logaritmica per entrambe le variabili.

Esempio 12.5.8 Legge allometrica

Negli organismi di una certa specie si osservano le seguenti misure, calcolate in chili, relative al peso P_c del corpo e a quello dello scheletro P_s

P_c	12.5	17.3	11.2	6.7	3.9	4.7	11.6	10	11.7	9.7
P_s	3.1	3.8	2.5	1.9	1.3	1.6	2.5	2.2	1.8	2.7

Ci si chiede se sia ragionevole ipotizzare che tra le due variabili sussista una relazione di tipo allometrico $P_c = \alpha P_s^\beta$ (*vedi* es. 5.3.5). Se così fosse, introducendo le variabili $P_c' = \ln P_c$, $P_s' = \ln P_s$ e calcolando il logaritmo di entrambi i membri dell'uguaglianza potremmo scrivere la legge nella forma equivalente, ma lineare, nelle nuove variabili: $P_c' = \ln \alpha + \beta P_s'$. La tabella seguente riporta il valore dei logaritmi dei dati sperimentali.

P_c'	2.52	2.85	2.42	1.90	1.36	1.55	2.45	2.40	2.46	2.27
P_s'	1.13	1.36	0.92	0.64	0.26	0.47	0.92	0.79	0.59	0.99

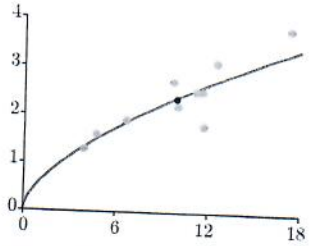

Figura 12.47
Regressione per una legge a potenza.

La media di P_c' è 2.21, la sua deviazione standard è 0.1940, la media di P_s' è 0.804, la sua deviazione standard è 0.09284. La covarianza è 0.1181 e il coefficiente di correlazione è 0.88, che è decisamente alto. Possiamo concludere che vi è una correlazione lineare tra i logaritmi dei dati. Il coefficiente angolare della retta di regressione è $a \approx 0.61$, il termine noto è $b \approx -0.54$. Alla legge lineare tra i logaritmi corrisponde una legge a potenza tra le variabili (fig. 12.47):

$$\ln P_s = a \ln P_c + b \implies P_s = \mathrm{e}^b P_c^a \approx 0.58\, P_c^{0.61}.$$

■

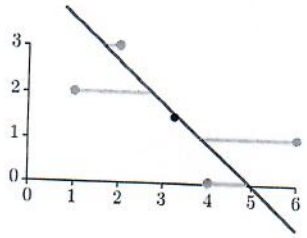

Figura 12.48
Minimizzare la somma dei quadrati degli scarti lungo l'asse x corrisponde a determinare la regressione della variabile X rispetto a Y. La retta che si ottiene è in genere diversa da quella della regressione di X rispetto a Y.

Nella costruzione della retta dei minimi quadrati, abbiamo implicitamente considerato Y come il dato da scrivere in funzione di X, infatti abbiamo misurato lungo l'asse verticale gli scarti tra dati e la retta. È possibile considerare la retta che minimizza gli scarti misurati in orizzontale (fig. 12.48), che corrisponde a determinare la regressione di X su Y, che è in genere diversa da quella che minimizza gli scarti in verticale. Infine, nel caso in cui le due variabili abbiano le stesse unità di misura, è anche possibile scegliere di misurare gli scarti come la distanza tra i dati e la retta, cioè lungo la perpendicolare dal punto alla retta (fig. 12.49). Il calcolo della retta migliore in quest'ultimo caso è più complesso ed è necessario procedere alla diagonalizzazione di una matrice simmetrica (*vedi* par. 3.3).
Illustriamo questo caso con un esempio.

Esempio 12.5.9 Matrice di covarianza

Consideriamo di nuovo i dati dell'esempio 12.5.5

X	1	6	2	4
Y	2	1	3	0

I valori delle medie sono $m_X = 13/4$ e $m_Y = 3/2$. La retta che stiamo cercando passa per il baricentro dei dati. Per determinarne l'inclinazione sono necessari i valori delle varianze e della covarianza: $\sigma_X^2 = 59/16$, $\sigma_Y^2 = 5/4$, $\sigma_{XY} = -11/8$.
Si chiama **matrice di covarianza** la matrice simmetrica

$$C = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} 59/16 & -11/8 \\ -11/8 & 5/4 \end{pmatrix} \tag{12.10}$$

Si tenga presente che $\sigma_{YX} = \sigma_{XY}$ e che $\sigma_X^2 = \sigma_{XX}$, cioè questa matrice è costituita da tutte le covarianze possibili che si possono calcolare per due variabili statistiche.
Calcoliamo autovalori e autovettori di questa matrice, secondo le procedure descritte nel paragrafo 3.3. L'equazione agli autovalori è

$$(59/16 - \lambda)(5/4 - \lambda) - (11/8)^2 = \lambda^2 - \frac{79}{16}\lambda + \frac{87}{32} = 0,$$

che ha soluzioni $\lambda_1 = (79 + \sqrt{3457})/32 \approx 4.31$ e $\lambda_2 = (79 - \sqrt{3457})/32 \approx 0.63$.
L'autovettore corrispondente a λ_1 è $\mathbf{v}_1 = (1.38, -0.62)$, quello corrispondente λ_2 è $\mathbf{v}_2 = (1.38, 3.06)$. In figura 12.49 rappresentiamo i dati e le rette parallele agli autovettori che passano per il baricentro (m_X, m_Y).
La retta che passa per il baricentro e ha come direzione quella individuata dall'autovettore di autovalore massimo (in questo caso λ_1) è la retta cercata, cioè quella più vicina ai dati, nel senso che minimizza la somma dei quadrati delle distanze. ■

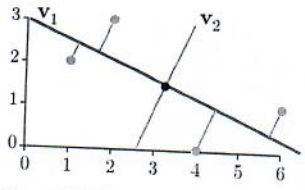

Figura 12.49
La retta in rosso è quella per cui è minima la somma delle distanze al quadrato dai punti. La retta ortogonale in verde è invece quella per cui la somma è massima tra tutte quelle che passano per il baricentro.

978-88-08-18484-9

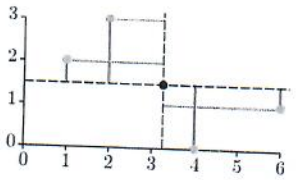

Figura 12.50
La lunghezza dei segmenti orizzontali in verde è la distanza dei punti dalla retta verticale che passa per il baricentro. La media dei quadrati di queste distanze è proprio la varianza nella variabile x. Analogamente, la varianza nella variabile y è la media dei quadrati delle lunghezze dei segmenti verticali in rosso, che rappresentano la distanza dei punti dalla retta orizzontale che passa per il baricentro.

La **matrice di covarianza** è la matrice che ha sulla diagonale i valori delle varianze e fuori diagonale le covarianze.
È una matrice simmetrica e ha autovalori positivi.
L'autovettore di autovalore massimo è la direzione della retta, che passa per il baricentro, per la quale è minima la somma dei quadrati delle distanze dei punti. (fig. 12.49). Questa retta si chiama **prima componente principale**. La retta ortogonale, che ha la direzione del secondo autovettore, è la **seconda componente principale**.

Senza entrare troppo nei dettagli diamo una spiegazione di questo fatto. Notiamo innanzi tutto che, nel calcolo della varianza σ_X^2, gli scarti $X - m_X$ sono le distanze dei punti dall'asse parallelo all'asse y che passa per il baricentro. Dunque σ_X^2 è la media dei quadrati delle distanze dall'asse perpendicolare a x, passante per il baricentro. Un'analoga interpretazione vale per la varianza calcolata rispetto ad altre variabili (fig. 12.50).
Come abbiamo visto alla fine del paragrafo 3.3, usando come nuovi assi coordinati gli autovettori, possiamo scrivere la matrice C in forma diagonale $\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}$, con $\lambda_1 > \lambda_2$. Dunque λ_1 e λ_2 si possono interpretare come σ_1^2 e σ_2^2, dove σ_1^2 è la varianza della variabile che esprime la coordinata del punto lungo la prima componente principale, e σ_2^2 è la varianza della variabile che esprime la coordinata lungo la seconda componente principale. Poiché σ_1^2 è la media delle distanze al quadrato dall'asse della seconda componente, σ_2^2 è la media delle distanze al quadrato dall'asse della prima componente e $\sigma_2^2 < \sigma_1^2$, in effetti la prima componente principale è l'asse che realizza la minima distanza dai dati.
Questa analisi vale in tutte le dimensioni.

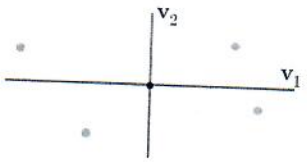

Figura 12.51
Lungo la prima componente principale i dati sono più dispersi, dunque la varianza della prima componente è più rilevante delle varianze delle altre componenti.

La prima componente principale è l'asse che realizza la minima distanza quadratica dai dati, ed è l'asse rispetto al quale la varianza è massima.
La rappresentazione dei dati lungo quest'asse è dunque quella in cui è maggiormente visibile la variazione dei dati (fig. 12.51)

Questo fatto viene anche sintetizzato affermando che le prime componenti principali "spiegano" la maggior parte della varianza complessiva dei dati.
Il metodo delle componenti principali permette quindi di diminuire la dimensione dello spazio in cui rappresentare dati e di quantificare la bontà di questa riduzione (per approfondire quest'ultimo aspetto *vedi* [4]). Per un'analisi di casi scientificamente significativi di applicazione delle componenti principali alle distanze genetiche di popolazione si veda [8] tra le letture consigliate.
Mostriamo un esempio nel caso di tre variabili statistiche.

Esempio 12.5.10 Scomposizione della variazione allelica

Abbiamo già accennato al fatto che gli esseri umani possono essere classificati secondo diverse caratteristiche del loro sangue, sulla base della presenza o assenza di specifici antigeni, che determinano la compatibilità nelle trasfusioni (*vedi* esercizio 11.5.1). Le classi di maggiore importanza sono due: il sistema $AB0$ e il gruppo Rh. I gruppi sanguigni $AB0$ sono determinati da tre alleli di un gene (*vedi* fig. 2.39 ed esercizio 10.14), i, I^A e I^B, mentre il gruppo Rh distingue individui che producono un particolare

978-88-08-18484-9

antigene, che vengono definiti Rh+, e quelli che non lo producono, che vengono detti Rh−. Il fattore Rh è codificato da un allele dominate indicato con D.
Come abbiamo discusso nel paragrafo 11.5, le diverse popolazioni presentano differenti frequenze alleliche. Nella seguente tabella riportiamo alcuni valori per le frequenze alleliche dei gruppi $AB0$ e Rh desunti sulla base di dati reperiti in rete.

Stato	Sigla	I^A	I^B	D
Australia	au	0.23	0.07	0.56
Canada	ca	0.26	0.06	0.61
Danimarca	dk	0.28	0.08	0.60
Finlandia	fi	0.31	0.13	0.64
Francia	fr	0.28	0.07	0.61
Gran Bretagna	uk	0.26	0.08	0.59
Stati Uniti	us	0.25	0.08	0.60
Corea del sud	kr	0.26	0.21	0.94
Svezia	se	0.29	0.09	0.60

(le sigle degli stati sono quelle utilizzate per i domini internet nazionali). Nella tabella non riportiamo le frequenze dell'allele i, poiché la sua frequenza relativa sommata a quelle di I^A e I^B deve essere 1, e quindi i dati relativi a i non aggiungerebbero nessuna informazione a quelle presenti.
Molte e interessanti ricerche sull'evoluzione dell'uomo analizzano i diversi valori delle frequenze alleliche nelle popolazioni. Lo scopo è quantificare la "distanza" tra le popolazioni, per poi poterne ricostruire la storia (*vedi* box 1.2). Per avere un'idea di questa distanza, possiamo rappresentare ogni terna di dati di una popolazione come un punto nello spazio $\mathbb{R}^3$, in cui le coordinate sono proprio le frequenze alleliche. Otterremo una figura poco leggibile (fig. 12.53).
Un metodo per semplificare l'interpretazione dei risultati può essere quello di rappresentare i dati in dimensione più bassa, per esempio potremmo scegliere il piano in cui sono rappresentate le frequenze di I^A e di D, o di I^B e di D. In entrambi i casi, però, useremmo solo parte dei dati. Il metodo delle componenti principali prende in considerazione tutti i dati e permette di determinare rispetto a quale sistema di assi perpendicolari i dati sono meglio rappresentati, cioè è più visibile la loro eventuale struttura. Infatti, le prime componenti principali sono gli assi più vicini ai dati e sono quelli che spiegano la maggior parte della varianza complessiva.
Indichiamo con A, B, D le variabili che specificano le frequenze degli alleli I^A, I^B e D. I valori medi delle variabili sono $m_A = 0.27$, $m_B = 0.10$, $m_D = 0.64$. La matrice di covarianza è

$$C = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} & \sigma_{AC} \\ \sigma_{BA} & \sigma_B^2 & \sigma_{BC} \\ \sigma_{CA} & \sigma_{CB} & \sigma_C^2 \end{pmatrix}.$$

Gli autovalori di C risultano $\lambda_1 = 5.0 \cdot 10^{-1}$, $\lambda_2 = 2.7 \cdot 10^{-3}$, $\lambda_3 = 3.8 \cdot 10^{-4}$.
In figura 12.54 rappresentiamo i dati nel piano che ha come origine il baricentro degli assi, in ascissa la prima componente principale e in ordinata la seconda. I dati sono più separati nella prima coordinata, che ha un intervallo di variabilità pari a circa 0.45, rispetto alla seconda, che ha un intervallo di solo 0.1. Nel grafico di figura 12.55 riportiamo invece il dati nel piano che ha in ascissa la terza componente principale e in ordinata la seconda. La terza componente principale ha un intervallo di variabilità ancora minore, di 0.04. In pratica, nei valori delle prime componenti principali sono racchiuse le maggiori differenze tra i dati e, dunque, guardare queste coordinate permette di estrarre le informazioni più significative.
Nel nostro esempio la figura 12.54 è molto leggibile e concorda con il senso comune: il dato della Corea è lontano da quello dei paesi occidentali, i dati dei paesi anglofoni sono vicini (Australia, Stati Uniti, Canada, Gran Bretagna) e, separati dalla Francia, più in basso ci sono i dati dei due paesi scandinavi. Questi stessi dati rappresentati nel piano formato dalla terza e dalla seconda componente principale (fig. 12.55) appaiono molto diversi e poco significativi: per esempio la terza componente principale (l'ascissa) della Corea si trova tra quella di Gran Bretagna e Svezia. Abbiamo perso la naturalezza "geografica" dei dati: le informazioni sulla terza componente principale sono molto meno importanti rispetto a quelle delle altre due. ■

Figura 12.52
Femmina e giovane esemplare di macaco Rhesus. In questo primate è stato individuato per la prima volta l'antigene Rh (che da lui prende il nome), la cui presenza rappresenta un grave rischio per un feto Rh+ di una madre Rh−, nel caso in cui, in una precedente gravidanza, la madre abbia sviluppato anticorpi contro il fattore Rh.

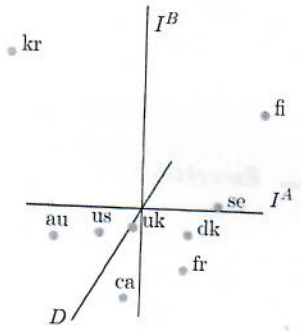

Figura 12.53
Rappresentazione tridimensionale dei dati. L'origine è in corrispondenza del baricentro delle frequenze alleliche.

978-88-08-18484-9

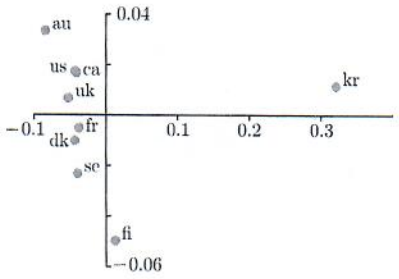

Figura 12.54
In ascissa le coordinate lungo la prima componente principale, in ordinata quelle lungo la seconda.

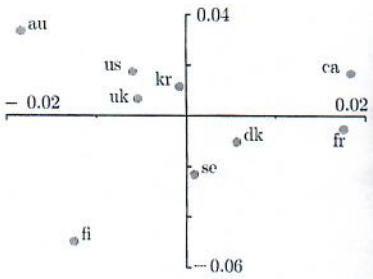

Figura 12.55
In ascissa le coordinate lungo la terza componente principale, in ordinata quelle lungo la seconda.

12

AR

Esercizi

Esercizio 12.1

In un'indagine sul livello di istruzione, si raccolgono i dati relativi al livello di studio raggiunto dai 54 370 abitanti sopra i 30 anni di una piccola città. Questi dati sono riassunti nella tabella che segue.

Analfabeta	215
Nessun titolo di studio	5331
Licenza elementare	6485
Licenza media	22 578
Maturità	10 649
Laurea	9070
Dottorato	42

Che cosa rappresentano i dati sulla seconda colonna? Si calcolino le frequenze relative di ogni categoria e si descriva come costruire un aerogramma che rappresenti i dati.

Esercizio 12.2

Si considerino le due sequenze di dati

(A)	3.1	3.5	3.5	3.8	5	5.2	5.2	5.2	5.5	
(B)	0.25	1.2	0.75	0.75	1.25	9	4.5	0.4	0.1	2.2

e per ciascuna si calcolino la moda, la media, e la mediana.

Esercizio 12.3

I dati dell'esercizio 12.1 sono di una popolazione o di un campione?

Esercizio 12.4

In un campionamento vengono considerati 290 individui scelti a caso in una popolazione, la cui età varia fra i 35 e i 58 anni. Questo intervallo di età viene suddiviso in 4 intervalli di numeri interi:

$$I_1 = [35, 40], \quad I_2 = [41, 46], \quad I_3 = [47, 52], \quad I_4 = [53, 58].$$

Dei 290 individui, 70 hanno un'età appartenente all'intervallo di valori I_1, 50 hanno un'età appartenente a I_2, 120 a I_3 e 50 a I_4. Si disegni un istogramma delle frequenze di questi dati e si determini, a partire da questi dati, l'età media.

Esercizio 12.5

Uno studente ha sostenuto 8 esami riportando, nell'ordine, le votazioni seguenti (in trentesimi): 23, 18, 25, 19, 23, 20, 27, 18. Se nei successivi tre esami lo studente riporta i voti 28, 26, 30, la media aumenta o diminuisce?
Se lo studente volesse raggiungere la media complessiva di 24, sostenendo due successivi esami, che votazioni dovrebbe riportare?
Calcolare la deviazione standard dei voti dei primi 8 esami e spiegare che cosa rappresenta il risultato.

Esercizio 12.6

Si eseguono alcune misure di una grandezza X e si rilevano i seguenti risultati, con le frequenze sottoindicate.

X	0	1.3	1.2	0.3	3.4	0.5	1.6	4.7	0.8	2.9
F	2	6	12	9	19	39	42	39	21	11

Calcolare la media, la varianza e la varianza campionaria.

Esercizio 12.7

Si considerino i dati dell'esercizio 12.6. Se i valori di X vengono moltiplicati per una stessa costante positiva, come cambiano la media, la varianza campionaria, la varianza e la deviazione standard?
Se invece si considerano gli stessi dati X e si moltiplicano per 2 le frequenze, come cambiano la media, la varianza campionaria e la varianza?

Esercizio 12.8

Viene effettuato un test di durata su un campione casuale di 100 lampadine a incandescenza. I dati vengono raggruppati in classi secondo la seguente tabella

T	(0,2.5)	(2.5, 5)	(5, 7.5)	(7.5, 10)	(10,12.5)	(12.5, 15)
F	8	27	15	17	27	6

con la durata T misurata in centinaia di ore. Calcolare la durata media delle lampadine del campione e la deviazione standard campionaria.

Esercizio 12.9

Si vuole fare un'indagine sull'obesità maschile in una piccola città. Si considera, a questo scopo, un campione di 100 individui raggruppati in 4 fasce di età: nella prima fascia ci sono 12 individui da 0 a 25 anni, nella seconda ci sono 37 individui che hanno fra i 25 e i 50 anni, nella terza ci sono 41 individui con età compresa fra i 50 e i 75 anni, infine nell'ultima ci sono 10 individui che hanno più di 75 anni. Da questo campione, si possono ottenere informazioni significative sull'incidenza dell'obesità **indipendentemente** dall'età?
Si considerano, per un'analisi più dettagliata, i 37 individui di età compresa fra i 25 e i 50 anni. I loro pesi in chilogrammi, approssimati ai 5 chili, sono distribuiti secondo la tabella seguente

Peso	65	70	75	80	85	90	105
Frequenza	3	5	11	14	2	1	1

Calcolare moda, media e mediana.

Esercizio 12.10

Si sospetta che un campo di mais sia stato contaminato da semi transgenici oltre la soglia dello 0.1%. Superata questa soglia, è obbligatorio dichiarare la percentuale di OGM presente nelle farine ricavate dal mais.
Viene analizzato un campione di 8000 semi, e di questi 6 risultano della varietà transgenica. A un livello di fiducia del 95%, qual è l'intervallo fiduciario della frazione di semi transgenici sul totale della piantagione?

Esercizio 12.11

Viene analizzato un campione di 1235 semi importati. Di essi 22 risultano transgenici. La ditta produttrice garantisce che la percentuale di semi transgenici tra i suoi prodotti è dell'1%. Si testi l'ipotesi nulla che la ditta affermi il vero.

Esercizio 12.12

Si considerino i dati dell'esercizio 12.8 sul campione di 100 lampadine e si sottoponga a test l'ipotesi nulla che la durata delle lampadine sia uniformemente distribuita negli intervalli di tempo considerati.

Esercizio 12.13

In un esperimento botanico si incrociano piante eterozigoti rispetto a tre coppie di caratteri indipendenti: forma del seme liscia/rugosa, colore del seme giallo/verde, colore del fiore bianco/rosso. Supponendo che i caratteri "seme rugoso", "seme verde", "fiore bianco" siano recessivi, si determinino le frequenze degli 8 fenotipi possibili.
Sperimentalmente si osservano i seguenti dati su 256 piante scelte a caso.

Forma	Col. semi	Col. fiori	Frequenza
liscia	giallo	rosso	118
liscia	giallo	bianco	37
liscia	verde	rosso	42
liscia	verde	bianco	11
rugosa	giallo	rosso	29
rugosa	giallo	bianco	4
rugosa	verde	rosso	15
rugosa	verde	bianco	0

Si valuti se questi dati sono in accordo con le conclusioni teoriche sulle frequenze dei fenotipi.

Esercizio 12.14

Utilizzando i dati dell'esercizio 12.10, riportare in una tabella di contingenza i dati relativi solo alla forma del seme e al colore del fiore in modo che in riga ci siano i fenotipi liscio/rugoso e in colonna i fenotipi rosso/bianco.
Valutare con un test del χ^2 se i dati sono in accordo con l'ipotesi di indipendenza.

Esercizio 12.15

L'immigrazione è uno dei fenomeni che, col passare del tempo, può modificare la frequenza degli alleli in una popolazione.
Per esempio, possiamo supporre che i maschi di una popolazione isolata si presentino con la caratteristica "carnagione chiara" indipendentemente dall'altezza. Se questa popolazione riceve un'immigrazione costante di maschi di carnagione chiara e alti meno di 1.70 m, questa caratteristica di indipendenza dovrebbe venir meno.
Si consideri a questo proposito la seguente tabella di contingenza e si valuti se si può affermare che la caratteristica "carnagione chiara" sia indipendente dalla caratteristica "altezza h inferiore a 1.70 m".

	Chiara	Scura	Totale
$h \leq 1.70$	65	247	312
$h > 1.70$	63	457	520
Totale	128	704	832

Esercizio 12.16

Calcolare le varianze e la covarianza dei seguenti dati

X	2	3	2	5	0	0
Y	6	4	5	2	3	1

978-88-08-18484-9

Esercizio 12.17

Si considerino i seguenti valori per due variabili statistiche relative a un campionamento.

X	0.02	0.80	1.02	0.50	1.08	0.70	0.80
Y	1.08	2.20	0.99	1.60	2.00	1.60	1.70

Dopo aver disegnato in un piano cartesiano i punti (X_i, Y_i), $i = 1, \ldots, 7$, calcolare la covarianza e interpretare il risultato. Calcolare inoltre l'indice di correlazione e spiegare anche questo risultato.

Esercizio 12.18

Scrivere l'equazione della retta di regressione della variabile Y sulla variabile X con i dati dell'esercizio 12.17.
Utilizzando l'equazione della retta trovare il valore corrispondente ai dati $x = 0.67$ e $x = 0.12$.

Esercizio 12.19

Determinare la legge esponenziale che lega la variabile Y alla variabile X per i seguenti dati

X	0.2	1	0.4	2	0.7	2.1	2.2	1.2
Y	1.6	11	2.6	103	27.2	115	120	17.1

Esercizio 12.20

Siano assegnati i valori delle due variabili

X	2	4	1	3	6
Y	48	780	4	240	3800

Determinare, mediante regressione, una relazione del tipo $y = ax^{\beta}$ che lega le due variabili.

978-88-08-18484-9

Capitolo 12

Soluzione 12.1

I dati sulla seconda colonna sono le frequenze assolute con cui si rilevano gli attributi della prima colonna.
Le frequenze relative dei cittadini appartenenti a ognuna delle 7 classi si ottengono dividendo le frequenze assolute per la numerosità della popolazione.

Analfabeta	0.40%
Nessun titolo di studio	9.81%
Licenza elementare	11.93%
Licenza media	41.53%
Maturità	19.59%
Laurea	16.68%
Dottorato	0.08%

(si noti che gli errori di approssimazione portano a una somma di poco differente dal 100%; esistono modi di approssimazioni differenti che evitano questo problema, ma non ne discuteremo qui). Per costruire un aerogramma, si divide l'angolo complessivo di 360 gradi in proporzione alle frequenze relative. Dunque, per esempio, l'angolo che compete alla classe dei cittadini che hanno raggiunto la licenza media ma non hanno proseguito gli studi è $360 \cdot 0.4153 = 149.5\,^\circ$, mentre quello che compete ai dottori di ricerca è di soli 0.3°, praticamente invisibile in un aerogramma.

Soluzione 12.2

La media della prima serie di dati è

$$m_A = \frac{3.1 + 2 \cdot 3.5 + 3.8 + 5 + 3 \cdot 5.2 + 5.5}{9} \approx 4.4\,,$$

mentre la mediana è 5 (si noti che i dati sono già ordinati in ordine crescente e che sono in numero dispari). La moda è 5.2.
I dati (B), riportati in ordine crescente, sono

(B)	0.1	0.25	0.4	0.75	0.75	1.2	1.25	2.2	4.5	9

La media è $m_B = 2.04$, mentre la mediana è $m = (0.75 + 1.2)/2 = 0.975$. La grande differenza tra questi valori è dovuta ai due dati 4 e 9.5, che contribuiscono ad alzare la media. La moda è 0.75.
Per completare l'esercizio, si rappresentino i dati sia in un diagramma a punti (cioè usando il numero indicativo del dato come ascissa e il valore come ordinata) sia in un istogramma delle frequenze, e si riportino gli indici calcolati nei grafici.

Soluzione 12.3

I dati si riferiscono a un'intera città, dunque si tratta di un'indagine su tutti gli elementi di una popolazione. Si potrebbe pensare che la città sia un campione di una nazione, ma in questo caso si tratterebbe di un campione distorto. Si pensi, per esempio, a come incidono sul livello di istruzione dei cittadini la presenza di università o di altre strutture scientifiche e culturali.

Soluzione 12.4

Non avendo a disposizione i dati sull'età esatta degli individui del campionamento, possiamo solo attribuire agli individui con età in un intervallo I il valore medio dell'intervallo.
Le età medie in ciascun intervallo sono 37.5, 43.5, 49.5 e 55.5. L'età media degli individui del campione è $m = (70 \cdot 37.5 + 50 \cdot 43.5 + 120 \cdot 49.5 + 50 \cdot 55.5)/290 \approx 46.6$.
Per completare l'esercizio si disegni l'istogramma delle frequenze.

Soluzione 12.5

La media dei voti dello studente nei primi 8 esami è $m_8 = 21.625$. La media degli ultimi 3 esami è 28, dunque la media complessiva degli 11 esami sarà più grande della media dei primi 8. Risulta infatti $m_{11} \approx 23.36$.
Detti v_1 e v_2 i voti ai successivi esami, deve risultare $24 = (257 + v_1 + v_2)/13$, cioè $v_1 + v_2 = 55$. Quindi il voto medio degli ultimi due esami deve superare $55/2 = 27.5$. I primi 8 voti in ordine sono

$$18,\ 18,\ 19,\ 20,\ 23,\ 23,\ 25,\ 27,$$

dunque la varianza sui primi 8 esami è

$$\sigma^2 = \frac{2(23 - 21.625)^2 + 2(18 - 21.625)^2}{8} + \\ + \frac{(25 - 21.625)^2 + (19 - 21.625)^2}{8} \\ + \frac{(20 - 21.625)^2 + (27 - 21.625)^2}{8} \approx 9.9844\,.$$

La deviazione standard è $\sigma = \sqrt{\sigma^2} \approx 3.16$ e indica che l'oscillazione quadratica media dei voti intorno alla media è di circa 3 punti: lo studente ottiene risultati piuttosto diversi tra loro.

Soluzione 12.6

Somma delle frequenze, si ottiene che il numero dei dati è 200, La media è $m_X = 2.041$, la varianza è $\sigma_X^2 \approx 2.497$, la varianza campionaria è $s_X^2 \approx 2.510$.

Soluzione 12.7

Sia c la costante per la quale vengono moltiplicati i dati. La media viene moltiplicata per c, mentre varianza e varianza campionaria vengono moltiplicate per c^2, infine la deviazione standard viene moltiplicata per c.
Se si moltiplicano per 2 le frequenze assolute, raddoppia anche la dimensione del campione, quindi le frequenze relative non cambiano. Media, varianza e deviazione standard rimangono di conseguenza invariate, perché dipendono solo dalle frequenze relative.

978-88-08-18484-9

Si noti che, invece, la varianza campionaria cambia. Infatti, visto che $s_N^2 = N\sigma_N^2/(N-1)$, se si raddoppia la dimensione del campione, si ha $\sigma_{2N}^2 = \sigma_N^2$, mentre $N/(N-1)$ diventa $2N/(2N-1)$ e dunque $s_{2N}^2 = 2N\sigma_N^2/(2N-1) \neq N\sigma_N^2/(N-1)$.

Soluzione 12.8

I dati sono forniti in classi, in particolare sono specificati intervalli dei dati di durata, ma non i valori misurati. Per poter procedere con il calcolo degli indici statistici sintetizziamo l'informazione legata a un intervallo con il suo valore medio. Riscriviamo dunque la tabella dei dati come:

T	1.25	3.75	6.25	8.75	11.25	13.75
frequenza	8	27	15	17	27	6

La media è $m_T = 7.40$ centinaia di ore. La deviazione standard è $\sigma_T = \sqrt{13.5525} \approx 3.68$. La deviazione standard campionaria è $s_T = \sqrt{13.689394} \approx 3.70$.

Soluzione 12.9

Il campione non è rappresentativo degli individui indipendentemente dall'età, perché il numero degli individui in ciascuna fascia non è lo stesso.
Per quel che riguarda il campione di 37 individui, la moda è 80, il peso medio è circa 77. I dati sono 37, dunque la mediana è il diciannovesimo dato, cioè 75.

Soluzione 12.10

L'intervallo di confidenza al 95% si ottiene calcolando $1.96\sqrt{\bar{q}(1-\bar{q})/N} \approx 0.00060$ dove $\bar{q} = 6/8000 = 0.00075$ è la frequenza osservata. Dunque l'intervallo richiesto, espresso in percentuale, è $(0.015, 0.135)$.
L'estremo superiore di questo intervallo è superiore al limite di legge 0.1%.

Soluzione 12.11

Indichiamo con X la variabile aleatoria che conta il numero di semi transgenici su un campione di $N = 1235$. Se la ditta afferma il vero dicendo che $q = 0.01$ è la frazione di semi transgenici, la variabile $Z = (X/N - q)\sqrt{N/(q(1-q))}$ deve essere gaussiana standardizzata.
Dai dati del problema si ottiene il valore $Z = (22/1235 - 0.01)\sqrt{1235/(0.01 \cdot 0.99)} \approx 2.76$. Confrontandolo con i valori riportanti nelle (12.6) a pag. 453, si vede che la probabilità che $|Z|$ sia maggiore di 2.76 è inferiore all'1% (anche se superiore allo 0.1%), dunque il dato osservato differisce in modo molto significativo dal dato atteso 12.35, e rifiutiamo l'ipotesi nulla.

Soluzione 12.12

I dati del campione sono 100 e risultano distribuiti in 6 classi. Supponendo valida l'ipotesi nulla, la frequenza relativa teorica per ogni classe è 1/6, dunque le frequenze attese sono tutte pari a $100/6 \approx 16.67$. Calcoliamo il valore del χ^2:

$$\frac{1}{16.67}\Big((8-16.67)^2 + (27-16.67)^2 + (15-16.67)^2 + (17-16.67)^2 + (27-16.67)^2 + (6-16.67)^2\Big) \approx 24.32\,.$$

Questo valore va confrontato con la distribuzione del χ^2 a 5 gradi di libertà della tabella 12.1 a pag. 461. Poiché il valore ottenuto è maggiore del valore 20.52, si rifiuta l'ipotesi nulla, con una significatività dello 0.1%.

Soluzione 12.13

Il fenotipo dominante, nel caso di un gene con due alleli, si presenta con frequenza 3/4, e il fenotipo recessivo con frequenza 1/4. Sotto l'ipotesi di indipendenza, per ottenere le frequenze teoriche dei fenotipi complessivi è sufficiente moltiplicare le frequenze dei singoli fenotipi.
Riscriviamo dunque la tabella dei dati, indicando la frequenza relativa teorica P, quella attesa su 256 piante $F_P = 256\,P$, la frequenza osservata F e lo scarto $\delta = F - F_P$.

Forma	Col. semi	Col. fiori	P	F_P	F	δ
liscia	giallo	rosso	27/64	108	118	10
liscia	giallo	bianco	9/64	36	37	1
liscia	verde	rosso	9/64	36	42	6
liscia	verde	bianco	3/64	12	11	−1
rugosa	giallo	rosso	9/64	36	29	−7
rugosa	giallo	bianco	3/64	12	4	−8
rugosa	verde	rosso	3/64	12	15	3
rugosa	verde	bianco	1/64	4	0	−4

Il valore del χ^2 è in questo caso circa 13.48 e va confrontato con i valori del χ^2 a 7 gradi di libertà (*vedi* tab. 12.1). Poiché $13.48 < 14.07$, che è il valore per cui si ha $P(\chi^2 > 14.07) = 0.05$, non rifiutiamo l'ipotesi nulla.

Soluzione 12.14

La frequenza del fenotipo "seme liscio e fiore rosso" si ottiene sommando i dati relativi ai fenotipi "seme liscio giallo e fiore rosso" e "seme liscio verde e fiore rosso", dunque è $118 + 42 = 160$. Analogamente si ottengono i dati relativi agli altri fenotipi.
La tabella richiesta è

	Rosso	Bianco	Totale
Liscio	160	48	208
Rugoso	44	4	48
Totale	204	52	256

978-88-08-18484-9

Nell'ipotesi di indipendenza, il fenotipo liscio è presente con frequenza relativa 208/256, il fenotipo rosso con frequenza 204/256. La frequenza attesa su 256 piante è dunque $208 \cdot 204/256 = 165.75$. In modo analogo si ottengono le frequenze attese degli altri fenotipi, che mostriamo in tabella, in cui riportiamo, tra parentesi, anche gli scarti.

	Rosso	Bianco	Totale
Liscio	165.75 (−5.75)	42.25 (5.75)	208
Rugoso	38.25 (5.75)	9.75 (−5.75)	48
Totale	204	52	256

Il valore del χ^2 è dunque

$$5.75^2 \left(\frac{1}{165.75} + \frac{1}{42.25} + \frac{1}{38.25} + \frac{1}{9.75} \right) \approx 5.24 .$$

Questo valore va confrontato con i valori del χ^2 a un grado di libertà (si noti infatti che i quattro scarti quadratici sono tutti uguali). Esso è più grande di 3.84, che corrisponde al 5% si significatività, ma più piccolo di 6.63, che corrisponde all'1% (*vedi* tab. 12.1). Possiamo concludere che il dato osservato è significativamente distante dal dato atteso, quindi rifiutiamo l'ipotesi nulla che i due caratteri siano indipendenti.

Soluzione 12.15

Nell'ipotesi di indipendenza, la caratteristica "carnagione chiara" dovrebbe presentarsi con frequenza relativa 128/832, la caratteristica "carnagione scura" con frequenza relativa 704/832, la caratteristica $h \leq 1.70$ con frequenza relativa 312/832, la caratteristica $h > 1.70$ con frequenza relativa 520/832.
Da questi valori, utilizzando la regola del prodotto, possiamo scrivere la tabella dei valori attesi nell'ipotesi di indipendenza.

	Chiara	Scura	Totale
$h \leq 1.70$	48	264	312
$h > 1.70$	80	440	520
Totale	128	704	832

Lo scarto è ±17, dunque

$$\chi^2 = 17^2 \left(\frac{1}{48} + \frac{1}{80} + \frac{1}{264} + \frac{1}{440} \right) \approx 11.38 .$$

Questo valore va confrontato con i valori del χ^2 a un grado di libertà (*vedi* tab. 12.1 a pag. 461), e risulta maggiore di 10.83 (che corrisponde a una significatività dello 0.1%), dunque rifiutiamo l'ipotesi di indipendenza, considerando estremamente significativo lo scarto osservato.

Soluzione 12.16

Il campione è costituito di $N = 6$ elementi. Gli indici statistici per le due variabili sono: $m_X = 2$, $m_Y = 7/2$, $\sigma_X^2 = 3$, $\sigma_Y^2 = 35/12$, $\sigma_{XY} = 1/3$.

Soluzione 12.17

Il campione è costituito da $N = 7$ elementi. Gli indici statistici per le due variabili sono: $m_X \approx 0.70$, $m_Y \approx 1.60$, $\sigma_X^2 \approx 0.1099$, $\sigma_Y^2 \approx 0.1675$, $\sigma_{XY} \approx 0.0544$. La covarianza è positiva, dunque al crescere degli X crescono anche i valori di Y. Il coefficiente di correlazione è $\rho \approx 0.40$, valore che indica uno scarso allineamento.

Soluzione 12.18

Il coefficiente angolare della retta di regressione è $\sigma_{XY}/\sigma_X^2 \approx 0.49$, dunque $y = 0.49x + b$. Imponendo che la retta passi per (m_X, m_Y), si ottiene $1.60 = 0.49 \cdot 0.70 + b$, da cui $b \approx 1.26$.
La retta è dunque $y = 0.49x + 1.26$, che prevede per $x = 0.67$ il valore $y \approx 1.59$, per $x = 0.12$ il valore $y \approx 1.32$.

Soluzione 12.19

Per determinare la legge esponenziale, calcoliamo i logaritmi di Y.

X	0.2	1	0.4	2	0.7	2.1	2.2	1.2
Y	1.6	11	2.6	103	27.2	115	120	17.1
$\ln Y$	0.47	2.40	0.96	4.63	3.30	4.74	4.79	2.84

Gli indici statistici valgono $m_X = 1.225$, $m_{\ln Y} \approx 3.02$, $\sigma_X^2 \approx 0.5469$, $\sigma_{\ln Y}^2 \approx 2.4906$, $\sigma_{X \ln Y} \approx 1.0986$. Il coefficiente angolare della retta di regressione è $a = 2.01$, il termine noto è $b = m_{\ln Y} - a m_X \approx 0.55$. La legge cercata è dunque $y = \mathrm{e}^{ax+b} = \mathrm{e}^b \mathrm{e}^{ax} \approx 1.75\,\mathrm{e}^{2.01x}$

Soluzione 12.20

Per determinare una legge a potenza si può considerare la regressione tra i logaritmi delle variabili, che riportiamo approssimati alla seconda cifra decimale.

$\ln X$	0.69	1.39	0.00	1.10	1.79
$\ln Y$	3.87	6.66	1.39	5.48	8.24

Procediamo ora al calcolo del coefficiente di correlazione. Si ha $m_{\ln X} = 0.994$, $m_{\ln Y} = 5.128$, $\sigma_{\ln X}^2 \approx 0.3764$, $\sigma_{\ln Y}^2 \approx 5.5421$, $\sigma_{\ln X \ln Y} \approx 1.4438$.
Dunque il coefficiente di correlazione è $\rho \approx 0.9996$ e i logaritmi dei dati sono praticamente allineati. Il coefficiente angolare della retta di regressione è $a \approx 3.84$, mentre $b \approx 1.32$. Quindi $y = \mathrm{e}^{a \ln x + b} = \mathrm{e}^b x^a \approx 3.74\, x^{3.84}$.