

# Appendice B: Introduzione alla statistica

Quest'appendice vuole essere un aiuto per condurre semplici analisi di statistica e per comprendere la loro applicazione e importanza. Vi aiuterà a completare i quesiti di "Applichiamo il concetto" e "Analisi dei dati" proposti in diversi capitoli del libro. Di seguito verranno esposte le equazioni di alcuni test di statistica, ma lo scopo principale dell'appendice è di aiutarvi a comprendere lo scopo e il ragionamento che si trovano alla base dei vari test. Una volta compreso il ragionamento di base delle analisi statistiche, sarete in grado di consultare uno dei numerosi siti web online gratuiti per effettuare i test e per calcolare test attinenti di statistica (per esempio, <http://faculty.vassar.edu/lowry/VassarStats.html>).

## Per quale motivo si effettuano analisi statistiche?

**Quasi tutto è soggetto a variazioni** Viviamo in un mondo variabile, ma nell'ambito della variabilità che osserviamo tra gli organismi biologici si possono riscontrare quadri prevedibili. Usiamo la statistica per individuare e analizzare questi quadri. Considerate un qualsiasi gruppo di oggetti in natura – per esempio, tutte le donne dell'età di 22 anni, tutte le cellule epatiche del corpo o tutti i fili d'erba in un giardino. Sebbene questi elementi avranno molte caratteristiche in comune, essi mostreranno anche importanti differenze. I maschi dell'età di 22 anni tendono a essere più alti delle donne della stessa età ma, naturalmente, non ogni singolo uomo è più alto di ogni singola donna di questo gruppo di età.

La variabilità naturale può complicare l'individuazione di quadri generali. Per esempio, gli studiosi hanno stabilito che il fumo aumenta il rischio di sviluppare un tumore ai polmoni. Sappiamo, tuttavia, che non tutti i fumatori si ammalano di tumore ai polmoni e che non tutti i non fumatori non svilupperanno un tumore ai polmoni. Pertanto, confrontando soltanto un fumatore con un non fumatore, rischiamo di trarre conclusioni errate. In che modo, allora, gli scienziati hanno stabilito che il fumo aumenta il rischio di tumore al polmone? Quanti fumatori e quanti non fumatori sono stati studiati per poter affermare con una certa sicurezza che il fumo fa male?

*La statistica aiuta a individuare quadri generali anche quando la natura non sempre segue questi quadri generali.*

**Evitare falsi positivi e falsi negativi** Quando una donna si sottopone a un test di gravidanza, vi è una certa possibilità che il test dia un esito positivo anche quando ella non è incinta e vi è anche una certa possibilità che il test dia esito negativo benché ella aspetti effettivamente un figlio. Questi tipi di errore prendono il nome rispettivamente di *falsi positivi* e di *falsi negativi*.

Fare ricerca è un po' come effettuare un test clinico. Osserviamo certi fenomeni nel mondo e in base alle nostre osservazioni cerchiamo di trarre conclusioni su come funziona il mondo. In alcuni casi le nostre osservazioni ci

inducono a trarre conclusioni sbagliate. Potremmo concludere che un fenomeno si realizza quando in realtà non è così oppure potremmo concludere che un fenomeno non si realizza quando in realtà lo fa.

Per esempio, nel corso dell'ultimo secolo il pianeta Terra si stava riscaldando. Gli ecologi stanno cercando di comprendere se le popolazioni di piante e di animali siano state in qualche modo interessate da questo riscaldamento globale. Se disponiamo di informazioni a lungo termine sulla distribuzione delle specie e sulla temperatura in alcune aree del pianeta, allora possiamo verificare se gli spostamenti geografici delle specie coincidono con i cambiamenti climatici. Informazioni di questo tipo, tuttavia, possono essere estremamente complesse. Senza metodi statistici appropriati potremmo non essere in grado di determinare il reale impatto della temperatura sulla distribuzione delle specie, oppure potremmo pensare che esista una correlazione dove in realtà questa non c'è.

*La statistica aiuta a non trarre conclusioni errate.*

## In che modo la statistica aiuta a comprendere il mondo naturale?

La statistica è uno strumento essenziale per le rivelazioni scientifiche. La maggior parte degli studi di biologia comprende cinque tappe fondamentali, ognuna delle quali richiede l'intervento della statistica:

- **Tappa 1: definizione del piano sperimentale**  
Si definiscono con chiarezza la problematica e i metodi necessari per affrontarla.
- **Tappa 2: raccolta dei dati**  
Si raccolgono informazioni relative al mondo naturale attraverso esperimenti e studi sul campo.
- **Tappa 3: organizzazione e visualizzazione dei dati**  
Si realizzano tabelle, grafici e altre rappresentazioni utili per poter analizzare i dati raccolti.
- **Tappa 4: riassunto dei dati**  
Si riassumono i dati con alcuni calcoli chiave di statistica.
- **Tappa 5: statistica inferenziale (deduttiva)**  
Si usano metodi di statistica per trarre conclusioni generali dai dati su come funziona il mondo naturale.

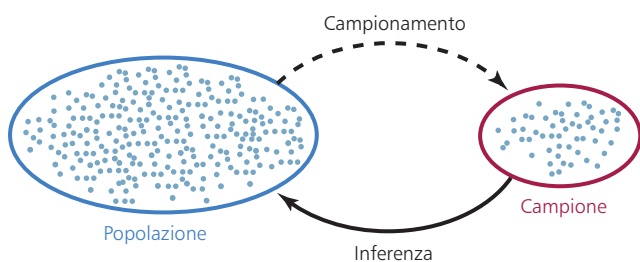
## ■ Tappa 1: definizione del piano sperimentale

Conduciamo esperimenti per acquisire conoscenze sul mondo. I ricercatori si occupano di fenomeni scientifici basati su ricerche precedenti e su osservazioni proprie. Queste idee possono assumere la forma di una domanda, tipo “il fumo provoca tumore?” di un’ipotesi, come “il fumo aumenta il rischio di tumore”; o di una previsione, quale per esempio “quando una persona fuma, la sua probabilità di sviluppare un tumore aumenta”. Gli esperimenti ci permettono di verificare queste domande o ipotesi scientifiche, ma la definizione del piano sperimentale potrebbe non essere così semplice.

Usiamo la statistica come guida per definire il piano sperimentale in modo da ottenere il giusto tipo di dati. Prima di allestire un esperimento ricorriamo alla statistica per determinare la mole di dati necessari per verificare la nostra idea e per evitare che fattori estranei possano indurci in errore. Supponiamo, per esempio, di voler condurre un esperimento sui fertilizzanti per verificare l’ipotesi che l’azoto incrementa la crescita delle piante. Se nello studio includiamo un numero insufficiente di piante, potremmo non essere in grado di determinare se l’azoto ha un effetto sulla crescita delle piante e l’esperimento sarebbe inutile. D’altra parte, lo studio di un numero di piante troppo elevato costituirebbe uno spreco di tempo e di risorse. Inoltre, dovremo allestire l’esperimento in modo tale da poter individuare le differenze che sono effettivamente dovute all’azoto e non ad altre variabili, per esempio, alla luce solare o alle precipitazioni.

## ■ Tappa 2: raccolta dei dati

**Campionamento** Quando i biologi raccolgono informazioni sul mondo naturale, essi raccolgono tipicamente alcuni pezzi rappresentativi di informazione. Per esempio, quando si valuta l’efficacia di un particolare farmaco nella terapia di un tumore cerebrale noto come medulloblastoma, si potrebbe esaminare l’effetto del farmaco su decine o su centinaia di pazienti e trarre successivamente le conclusioni sull’efficacia del farmaco in tutti i pazienti affetti da questo tipo di tumore. In modo simile gli studiosi che valutano il rapporto tra il peso corporeo e il numero di uova deposte dalle femmine di una particolare specie di ragno, potrebbero studiare decine o centinaia di ragni per trarre le proprie conclusioni.



### Figura B1 Campione di una popolazione

I biologi raccolgono campioni rappresentativi di una popolazione, applicano le statistiche descrittive per caratterizzare i campioni raccolti e ricorrono poi alla statistica inferenziale per trarre conclusioni relative alla popolazione originale.

Per descrivere questo metodo generale di prelevare pezzi rappresentativi di informazione dal sistema sotto studio si usa l’espressione “campione di una popolazione” (Figura B1). I pezzi d’informazione contenuti in un **campione** vengono denominati **osservazioni**. Nell’esempio relativo alla terapia del cancro, ogni osservazione corrisponde al cambiamento dimensionale del tumore in un paziente sei mesi dopo aver iniziato il trattamento e la popolazione di interesse corrisponde a tutti gli individui affetti da medulloblastoma. Nell’esempio relativo ai ragni, ogni osservazione riguarda due valori – le dimensioni corporee e il numero di uova deposte – per una singola femmina e la popolazione di interesse è quella delle femmine della specie in questione.

Il campionamento si basa su una necessità, non sul fatto che non si ha voglia di esaminare tutta la popolazione di interesse. Infatti, non possiamo sperare (né lo vogliamo) di raccogliere *tutte* le femmine di ragno della specie in esame presenti sulla Terra! Utilizziamo invece metodi di statistica per determinare quanti ragni devono essere raccolti per poter desumere in modo affidabile qualcosa sull’intera popolazione e ricorriamo di nuovo alla statistica per trarre queste conclusioni.

### I dati possono assumere tutte le forme e dimensioni

In statistica si usa il termine *variabile* per indicare un carattere misurabile di un individuo o di un sistema. Alcune variabili sono espresse in scala metrica, come la temperatura massima giornaliera (un valore numerico basato sulla precisione del termometro utilizzato) oppure il valore numerico delle uova deposte da un ragno (un numero intero: 0, 1, 2, 3, ...). Questi valori prendono il nome di **variabili quantitative**. Le variabili quantitative che esprimono soltanto numeri interi vengono dette **variabili distinte**, mentre le variabili che possono comprendere anche valori frazionari vengono indicate come **variabili continue**.

Altre variabili hanno come valore determinate categorie, come il gruppo sanguigno nell’uomo (A, B, AB o 0) o le caste delle formiche (regina, operaie o maschi). Le variabili di questo tipo vengono indicate come **variabili qualitative**. Le variabili qualitative dotate di un ordine naturale, come il voto finale in Biologia generale (da 0 a 30), vengono definite **variabili ordinali**.

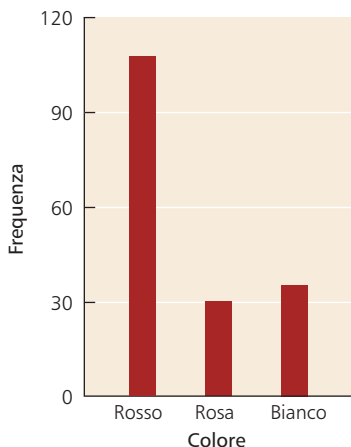
A ogni classe di variabili è associato un proprio corredo di metodi statistici. Alcuni dei metodi più comuni verranno introdotti in quest’appendice con l’intento di permettere allo studente di risolvere i quesiti proposti in questo libro. Tuttavia, per test più avanzati e per le analisi di altri tipi di dati e di problematiche si rimanda a libro di testo di biostatistica.

## ■ Tappa 3: organizzazione e visualizzazione dei dati

Le tabelle e i grafici possono facilitare l’interpretazione dei dati raccolti, permettere di individuare i test di statistica appropriati e anticipare il risultato dell’analisi. Una **distribuzione di frequenze** elenca tutti i possibili valori e il numero di ogni valore nel campione.

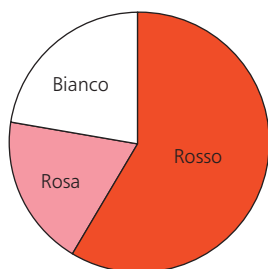
### Figura B2 I diagrammi a barra confrontano dati qualitativi

Questo diagramma a barra illustra la frequenza di tre colori di fiori di poinsettia nella progenie ottenuta da un esperimento di incrocio.



### Figura B3 I diagrammi a torta illustrano le proporzioni delle singole categorie

Questo diagramma a torta illustra le proporzioni dei tre colori di poinsettia riportati nella Tabella B1.



Nella **Tabella B1** è riportata la distribuzione di frequenza dei colori di 182 piante di poinsettia (rosso, rosa o bianco) prodotte da un incrocio sperimentale tra due piante parentali. Per dati qualitativi di questo tipo possiamo visualizzare la distribuzione di frequenze costruendo un **diagramma a barre**. L'altezza delle barre indica il numero di osservazioni per ciascuna categoria (**Figura B2**). Un altro modo per visualizzare questi stessi dati è il **diagramma a torta**, che illustra le percentuali di ogni categoria sotto forma di fette di una torta (**Figura B3**).

**Tabella B1 I colori della poinsettia**

COLORE	FREQUENZA	PROPORZIONE
Rosso	108	0,59
Rosa	34	0,19
Bianco	40	0,22
<b>Totale</b>	<b>182</b>	<b>1,0</b>

Per i dati quantitativi è spesso utile provvedere prima a condensare i dati, raggruppandoli in **classi**. Nella **Tabella B2** è riportata una distribuzione di frequenza raggruppata in questo modo per quanto riguarda il peso corporeo di un campione di 34 pesci (*Abramis brama*) catturati nel lago Laengelmaevesi in Finlandia. La seconda colonna (*Frequenza*) riporta il numero di osservazioni per ciascuna classe, mentre nella terza colonna (*Frequenza relativa*) è riportato per ogni classe il rapporto tra le frequenze assolute e il numero complessivo di osservazioni.

Gli **istogrammi** illustrano le distribuzioni di frequenza per i dati quantitativi. L'istogramma riportato nella **Figura B4** riporta le frequenze relative di ogni classe di peso corporeo presa in considerazione in questo studio. Quando si raggruppano dati quantitativi è necessario decidere il

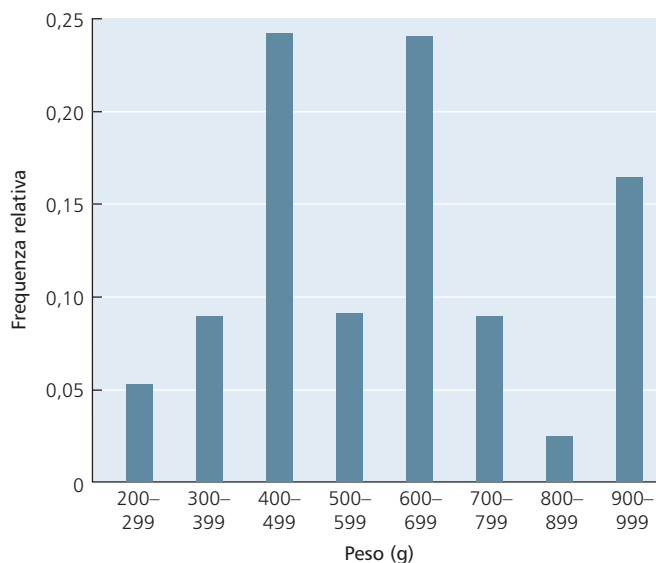
numero di classi che si vogliono includere nello studio. È spesso utile elaborare diversi istogrammi prima di decidere quale raggruppamento offre la migliore rappresentazione grafica dei dati.

**Tabella B2 Il peso di *Abramis brama* nel lago Laengelmaevesi**

PESO (GRAMMI)	FREQUENZA	FREQUENZA RELATIVA
201–300	2	0,06
301–400	3	0,09
401–500	8	0,24
501–600	3	0,09
601–700	8	0,24
701–800	3	0,09
801–900	1	0,03
901–1000	6	0,18
<b>Totale</b>	<b>34</b>	<b>1,0</b>

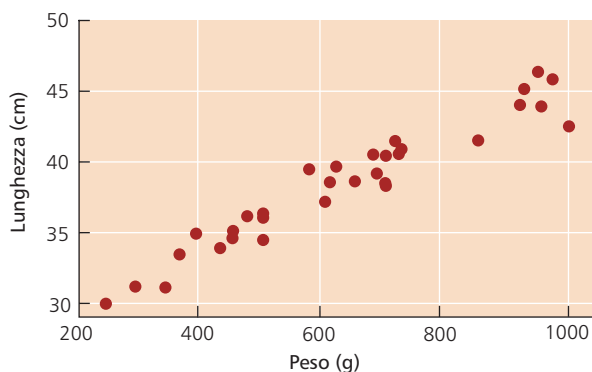
In alcuni casi vogliamo confrontare due variabili quantitative. Per esempio, i ricercatori coinvolti nello studio dei pesci del lago Laengelmaevesi hanno indagato il rapporto tra il peso corporeo e la lunghezza corporea dei pesci e hanno dunque misurato anche la lunghezza di ogni esemplare. Questo rapporto può essere visualizzato usando un **grafico di dispersione** in cui il peso e la lunghezza di ogni pesce costituiscono un singolo punto nel grafico (**Figura B5**). In questo esempio le due variabili sono caratterizzate da un **rapporto lineare**, poiché i singoli punti del grafico cadono approssimativamente su una linea retta.

Le tabelle e i grafici sono estremamente importanti ai fini dell'interpretazione e della comunicazione dei dati raccolti e dovrebbero essere il più possibile autonomi e di facile comprensione. Il contenuto dei grafici e delle tabelle dovrebbe essere facilmente comprensibile anche a prima vista. Gli assi, la denominazione delle variabili e le unità di misura scelte



### Figura B4 Gli istogrammi rappresentano le distribuzioni di frequenza di dati quantitativi

Questo istogramma illustra le frequenze relative di diverse classi di peso corporeo nei pesci (*Abramis brama*).



**Figura B5** I diagrammi diffusi confrontano due variabili

Il diagramma diffuso riporta il peso rispetto alla lunghezza degli esemplari di *Abramis brama* (misurata dalla punta del muso all'estremità caudale della coda). Queste due variabili possiedono un rapporto lineare, poiché i punti si trovano in prossimità di una linea retta.

dovrebbero essere indicati in maniera chiara, dovrebbero essere definiti i termini statistici e quando si tratta di tabelle o di grafici che esprimono dati quantitativi i dati dovrebbero essere raggruppati in maniera appropriata.

#### ■ Tappa 4: riassunto dei dati

Una **statistica** è una quantità numerica calcolata da dati, mentre le **statistiche descrittive** corrispondono a quantità che descrivono quadri generali sotto forma di dati. Le statistiche descrittive permettono di paragonare direttamente diversi insiemi di dati e di comunicare in maniera concisa le caratteristiche dei dati raccolti.

**Descrizione dei dati qualitativi** Le variabili qualitative vengono tipicamente descritte come proporzioni. In altre parole, si allestiscono tabelle con le proporzioni delle osservazioni per ciascuna categoria. Per esempio, la terza colonna nella Tabella B1 riporta le proporzioni tra piante di poinsettia per ogni categoria di colore e il diagramma a torta nella Figura B3 riporta una rappresentazione grafica di tali proporzioni.

**Descrizione dei dati quantitativi** Per i dati quantitativi si inizia spesso calcolando il valore medio o la **media** del campione. Questo termine piuttosto familiare corrisponde semplicemente alla somma di tutti i valori del campione diviso per il numero di osservazioni effettuate sul campione stesso (Figura B6). La media è soltanto uno di numerosi valori che ci dicono approssimativamente dove si trova il **centro** dei nostri dati. Questi valori quantitativi vengono indicati complessivamente come **misurazioni del centro**. Altre misure del centro comunemente usate sono la **mediana** – il valore che si trova letteralmente al centro del campione di dati – e la **norma** (o modo) – ovvero il valore caratterizzato dalla massima frequenza.

La stessa importanza che spetta al calcolo del centro spetta spesso anche alla quantificazione della variabilità dei dati. Esistono numerose statistiche che indicano quanto i valori raccolti differiscono gli uni dagli altri. Si tratta dell'**indice di**

**dispersione**. Quello più semplice da comprendere e da calcolare è l'**escursione**, che corrisponde semplicemente al valore più alto meno il valore più basso. La misura della dispersione più comunemente usata è la **deviazione standard**, che indica quanto i dati si discostano dalla media. Una deviazione è la differenza tra un'osservazione e la media del campione e la deviazione standard è un numero che riassume tutte le deviazioni. Due campioni possono avere la stessa escursione di valori, ma deviazioni standard molto diverse quando uno è raggruppato più vicino alla media rispetto all'altro. Nella **Figura B7**, per esempio, il campione a sinistra possiede una deviazione standard più bassa ( $s = 2,6$ ) rispetto al campione a destra ( $s = 3,6$ ), sebbene entrambi i campioni possiedano la stessa media e la stessa escursione di valori.

Per dimostrare queste statistiche descrittive ritorniamo allo studio realizzato nel lago Laengelmalesi. Infatti, i ricercatori hanno anche catturato e registrato il peso di sei esemplari di *Leusiscus idus*: 270, 270, 306, 540, 800 e 1000 grammi. In questo campione il peso medio (equazione 1 nella Figura B6) è:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{(270 + 270 + 306 + 540 + 800 + 1000)}{6} = 531$$

Poiché il numero di osservazioni nel campione è pari, il peso mediano corrisponde al valore a metà tra i due valori centrali:

$$\frac{306 + 540}{2} = 423$$

**Figura B6** Statistica descrittiva per dati quantitativi

#### METODI DI RICERCA

Sotto si trovano le equazioni usate per calcolare le statistiche descrittive discusse in questa Appendice. Queste statistiche possono essere calcolate da ogni studente oppure si può ricorrere a risorse gratuite presenti in internet per effettuare i calcoli.

Nota:

$x_1, x_2, x_3, \dots, x_n$  sono le  $n$  osservazioni della variabile  $X$  nel campione.

$\sum_{i=1}^n x_i = x_1 + x_2 + x_3 + \dots + x_n$  ... è la somma di tutte le osservazioni. (La lettera greca sigma,  $\Sigma$ , è usata per indicare "somma di.")

Nella regressione la variabile indipendente è  $X$ , mentre la variabile dipendente è  $Y$ .  $b_0$  è l'intercetta verticale di una linea di regressione.  $b_1$  è la pendenza di una linea di regressione.

Equazioni

$$1. \text{ Media: } \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

$$2. \text{ Deviazione standard: } s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

$$3. \text{ Coefficiente di correlazione: } r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$4. \text{ Linea di regressione dei minimi quadrati: } Y = b_0 + b_1 X$$

dove  $b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$  e  $b_0 = \bar{y} - b_1 \bar{x}$

La norma del campione è 270, l'unico valore che compare più di una volta. La deviazione standard (equazione 2 nella Figura B6) è:

$$s = \sqrt{\frac{(270 - 531)^2 + (270 - 531)^2 + (306 - 531)^2 + (540 - 531)^2 + (800 - 531)^2 + (1000 - 531)^2}{5}} = 309,6$$

e l'escursione è pari a  $1000 - 270 = 730$ .

**Descrizione del rapporto tra due variabili quantitative** I biologi sono spesso interessati a comprendere il rapporto tra due diverse variabili quantitative: in che modo l'altezza di un organismo è correlata con il suo peso corporeo? In che modo l'inquinamento dell'aria è collegato con l'incidenza di casi d'asma? In che modo l'abbondanza dei licheni è correlata con i livelli di inquinamento atmosferico? Si ricordi che i grafici di dispersione illustrano visivamente questi rapporti.

Possiamo quantificare la forza del rapporto tra due variabili quantitative usando un singolo valore denominato **coefficiente di correlazione** prodotto-momento di Pearson (equazione 3 nella Figura B6). Questo valore varia tra  $-1$  e  $1$  e indica quanto vicino i punti di un grafico di dispersione si trovano a una linea retta. Un coefficiente di correlazione negativo indica che una variabile diminuisce all'aumentare dell'altra, mentre un coefficiente di correlazione positivo indica che entrambe le variabili aumentano contemporaneamente; infine, un coefficiente di correlazione pari a zero indica che non vi è alcun rapporto lineare tra le due variabili (Figura B8).

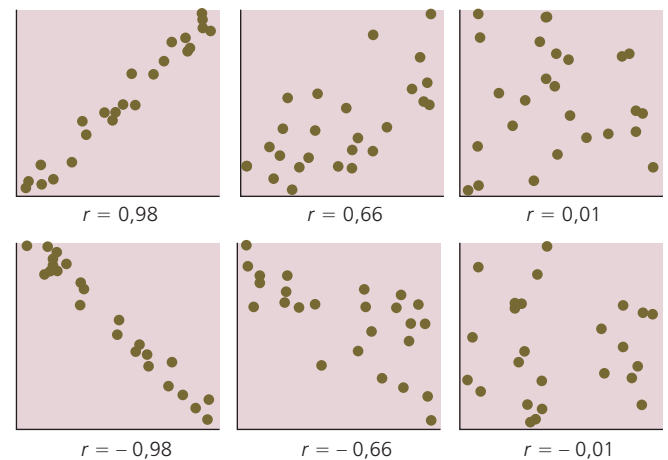
È necessario tener presente che *correlazione non significa causalità*. Due variabili possono essere strettamente correlate senza che una sia la causa dell'altra. Per esempio, il numero di carie di un bambino è correlato positivamente con le dimensioni dei piedi. Naturalmente le carie non favoriscono la crescita dei piedi né la crescita dei piedi induce la caduta dei denti. La correlazione esiste semplicemente per-

ché entrambi i valori tendono ad aumentare con l'aumentare dell'età.

Intuitivamente la linea retta che descrive i punti in un grafico a dispersione dice qualcosa sul rapporto *tipico* tra le due variabili. Tuttavia, in statistica non si adocchiano semplicemente i dati né si disegna una retta a mano. Infatti, in statistica si ricorre spesso al metodo dei minimi quadrati denominato **regressione lineare** per adattare una linea retta ai dati sperimentali (equazione 4 nella Figura B6). Questo metodo permette di calcolare la retta che minimizza le distanze verticali complessive tra i punti del grafico di dispersione e la linea retta stessa. Queste distanze vengono denominate **residui** (Figura B9). La linea di regressione è descritta da due parametri:  $b_0$  (l'intercetta verticale della retta o il valore atteso della variabile  $Y$  quando  $X = 0$ ) e  $b_1$  (la pendenza della retta, ovvero quanto i valori di  $Y$  cambiano con il variare dei valori di  $X$ ).

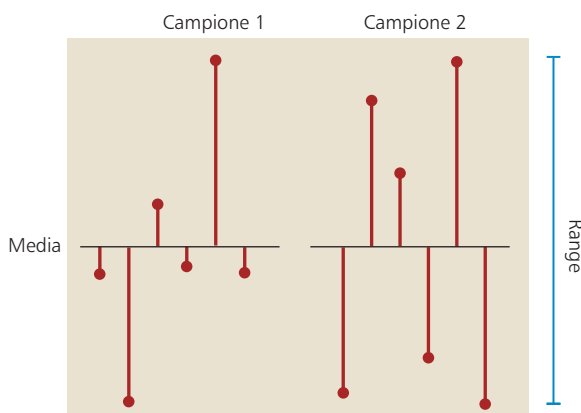
### Figura B8 I coefficienti di correlazione

Il coefficiente di correlazione ( $r$ ) indica sia la forza che la direzione del rapporto.



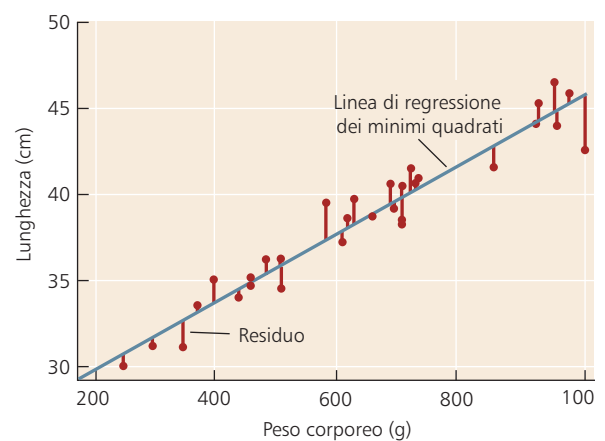
### Figura B7 Le misure della dispersione

Due campioni con la stessa media (linea nera orizzontale) ed escursione di valori (linea blu verticale). Le linee rosse illustrano la deviazione di ogni osservazione dalla media. I campioni con elevate deviazioni possiedono una grande deviazione standard. Il campione a sinistra possiede una deviazione standard minore del campione a destra.



### Figura B9 La regressione lineare stima il tipico rapporto tra le variabili

La retta di regressione lineare dei minimi quadrati per il peso corporeo e la lunghezza corporea (misurata dalla punta del muso all'apice della coda) di *Abramis brama*. La linea di regressione (linea blu) è data dall'equazione  $Y = 26,1 + 0,02X$ . Si tratta della linea che minimizza la somma del quadrato dei residui (linee rosse).



## ■ Tappa 5: statistica inferenziale (deduttiva)

Spesso l'analisi dei dati culmina con l'inferenza statistica – un tentativo di trarre conclusioni generali sul sistema sotto esame. Come illustrato nella Figura B1, la ragione primaria per la quale si raccolgono dati è di ottenere informazioni su un sistema più ampio rispetto a quello dal quale i dati vengono raccolti. Quando esaminiamo dieci pazienti per quanto riguarda l'effetto di un nuovo farmaco contro il medulloblastoma, non vogliamo soltanto conoscere il destino di questi dieci individui, ma speriamo piuttosto di prevedere l'efficacia del farmaco nel gruppo assai più numeroso che comprende tutti i pazienti affetti da medulloblastoma.

**Ipotesi statistiche** Quando si tratta di dedurre qualcosa sul mondo reale a partire dai nostri dati, abbiamo spesso in mente una domanda “Se... oppure no”. Per esempio, vorremo sapere *se* il riscaldamento globale influisce sulla biodiversità *oppure no*; *se* le dimensioni di una nidata di ragni dipendono dalle dimensioni corporee *oppure no* o *se* l'azoto del terreno influisce sulla crescita di una particolare specie di pianta *oppure no*.

Prima di giungere a deduzioni statistiche in base ai dati raccolti, è necessario formalizzare la nostra domanda “Se... oppure no” in un paio di ipotesi contrapposte: un'**ipotesi nulla** (indicata come  $H_0$ ) e un'**ipotesi alternativa** (denominata  $H_A$ ). L'ipotesi alternativa corrisponde al “Se” – essa viene formulata per descrivere l'effetto che ci aspettiamo di dimostrare con i nostri dati; l'ipotesi nulla corrisponde a “oppure no” – essa viene formulata per indicare l'assenza dell'effetto. In altre parole, normalmente conduciamo il nostro esperimento cercando di dimostrare qualcosa di nuovo (l'ipotesi alternativa) rigettando l'idea che l'effetto non si verifichi (l'ipotesi nulla).

Supponiamo, per esempio, di voler sapere *se* un nuovo vaccino è più efficace di uno già esistente per immunizzare i bambini contro l'influenza *oppure no*. Le nostre ipotesi sarebbero le seguenti:

$H_0$ : il nuovo vaccino non è più efficace del vecchio vaccino.  
 $H_A$ : il nuovo vaccino è più efficace di quello vecchio.

Se vogliamo sapere se le radiazioni aumentano il tasso di mutazione nei batteri *Escherichia coli* oppure no, allora formuleremo le seguenti ipotesi:

$H_0$ : la radiazione non aumenta il tasso di mutazione in *E. coli*.  
 $H_A$ : la radiazione aumenta il tasso di mutazione in *E. coli*.

**L'onere della prova in statistica** Nel sistema giurisdizionale degli Stati Uniti una persona è innocente fino a quando non sia stata provata la sua colpevolezza. In statistica, il mondo è *nulla fino a quando non è stata provata un'alternativa*. Le statistiche richiedono prove schiaccianti a favore dell'ipotesi alternativa prima di rigettare l'ipotesi nulla. In altre parole, gli scienziati favoriscono le idee già esistenti e sono restii ad adottare nuove idee fino a quando evidenze impellenti suggeriscono di fare altrimenti.

Quando si analizza l'efficacia di un nuovo vaccino contro l'influenza, l'onere della prova riguarda il nuovo vaccino.

Supponiamo di vaccinare tre bambini con il nuovo vaccino (Gruppo A), tre bambini con il vecchio vaccino (Gruppo B) e di lasciare altri tre bambini senza vaccinazione (Gruppo C). Se nessuno dei bambini del gruppo A, un bambino del gruppo B e un bambino del gruppo C si ammalano di influenza, allora avremmo evidenze a sufficienza per concludere che il nuovo vaccino è più efficace di quello vecchio? No, non avremmo evidenze a sufficienza. Se lo studio fosse ampliato a gruppi più numerosi di bambini e due bambini su 100 del gruppo A, sette su 100 del gruppo B e 22 su 100 del gruppo C si ammalassero, allora avremmo un'evidenza sufficiente per scegliere il nuovo vaccino? Forse, ma dobbiamo ricorrere alla statistica per esserne certi.

Si tratta del tradizionale onere delle prove in campo biologico e, più in generale, in campo scientifico. Di conseguenza, gli scienziati sono più propensi a lasciarsi sfuggire una nuova scoperta (seppur vera) piuttosto che rischiare di fare una scoperta inattendibile. Negli ultimi anni sono nate discussioni riguardo l'utilità di questo approccio e si è iniziato a sviluppare un approccio statistico alternativo, denominato **inferenza bayesiana**, che facilita l'accettazione di nuove ipotesi. Tuttavia, in questa appendice discuteremo soltanto i metodi di statistica tradizionali, che prendono spesso il nome di **statistica frequentistica**.

**Giungere a conclusioni errate** Vi sono due modi in cui un test di statistica può fallire (**Figura B10**). Possiamo rigettare l'ipotesi nulla quando in realtà questa è vera (**errore di tipo 1**) oppure possiamo accettare l'ipotesi nulla quando in realtà essa è falsa (**errore di tipo 2**). Questi tipi di errori sono analoghi rispettivamente ai falsi positivi e ai falsi negativi dei test diagnostici. Se erroneamente si rigetta l'ipotesi nulla quando questa in realtà è vera, allora si approva erroneamente un'ipotesi non corretta. Se siamo incapaci di rigettare l'ipotesi nulla quando questa è in realtà falsa, allora non riconosciamo una verità non ancora nota.

Supponiamo di voler sapere se in una popolazione di 10000 individui esiste un numero maggiore di femmine rispetto ai maschi. Per determinare la distribuzione dei due sessi nella popolazione in questione scegliamo a caso 20 individui e determiniamo la loro appartenenza a uno dei due sessi. La nostra ipotesi nulla è che *non* vi siano più femmine che maschi e la nostra ipotesi alternativa è che nella popolazione vi siano più femmine che maschi. Il seguente scenario illustra i possibili errori che potremmo commettere:

		Il mondo reale	
		Ipotesi nulla vera (non più femmine)	Ipotesi nulla falsa (più femmine)
La nostra conclusione	Ipotesi nulla vera (non più femmine)	✓	Errore di tipo 2 (falso negativo)
	Ipotesi nulla falsa (più femmine)	Errore di tipo 1 (falso positivo)	✓

**Figura B10** Due tipi di errore

Possibili risultati di un test di statistica. L'inferenza statistica può condurre a conclusioni corrette oppure a conclusioni errate sulla popolazione sotto studio.

- **Scenario 1:** la popolazione conta il 40% di femmine e il 60% di maschi. Sebbene il nostro campione di 20 individui scelti a caso abbia elevate probabilità di essere dominato da maschi, per caso scegliamo più femmine che maschi. Se ciò accade, rigettiamo erroneamente l'ipotesi nulla (che *non* vi siano più femmine che maschi) e commettiamo un errore di tipo 1.
- **Scenario 2:** la popolazione comprende il 60% di femmine e il 40% di maschi. Se, per caso, il nostro campione contiene un numero maggior di maschi, allora rigetteremo l'ipotesi nulla, commettendo un errore di tipo 2.

Fortunatamente, le statistiche sono state sviluppate in modo tale da evitare errori di questo tipo e ci permettono di determinare l'affidabilità delle nostre conclusioni. I metodi si basano sul calcolo delle **probabilità** dei possibili risultati. Sebbene potreste aver sentito parlare o anche aver usato il termine "probabilità" in varie occasioni, è importante comprendere il suo significato da un punto di vista matematico. Una probabilità è una quantità numerica che esprime la possibilità che si verifichi un determinato evento. La probabilità che un evento si verifichi può assumere un valore compreso tra zero e uno; zero significa che non vi è alcuna possibilità che l'evento si verifichi, mentre uno significa che l'evento si verificherà con certezza. Ciò ha senso soltanto se esiste un principio di casualità, ovvero se è possibile che l'evento si verifichi e se è possibile che esso non si verifichi. Per esempio, quando si lancia una moneta, la probabilità di ottenere testa è pari a 0,5 e anche la probabilità di ottenere croce è pari a 0,5. Quando selezioniamo individui casualmente in una popolazione costituita per il 60% da femmine e per il 40% da maschi, allora abbiamo una probabilità pari a 0,6 di selezionare una femmina e una probabilità pari a 0,4 di selezionare un maschio.

In statistica alla probabilità spetta un ruolo molto importante. Per trarre conclusioni relative al mondo reale (la popolazione) in base a un nostro campione, calcoliamo prima la probabilità che il nostro campione soddisfi l'ipotesi nulla. In modo specifico, l'inferenza statistica è basata sulle risposte alla seguente domanda:

*Supponendo che l'ipotesi nulla sia vera, qual è la probabilità che un campione casuale differisca, per caso, dall'ipotesi nulla nella stessa misura in cui il nostro campione differisce dall'ipotesi nulla?*

Se è estremamente poco probabile che il nostro campione soddisfi l'ipotesi nulla, allora lo scartiamo a favore dell'ipotesi alternativa. Se, invece, il nostro campione ha una probabilità ragionevole di soddisfare l'ipotesi nulla, allora concludiamo che i nostri dati sono consistenti con l'ipotesi nulla e non la rigettiamo.

Ritornando all'esempio del rapporto tra sessi, consideriamo altri due scenari:

- **Scenario 3:** supponiamo di voler determinare se le femmine costituiscono o meno la maggioranza di una popolazione (la nostra ipotesi alternativa) basandoci su un campione casuale formato da 12 femmine e da 8 maschi. Calcoleremo la probabilità che un campione casuale di

20 individui includa almeno 12 femmine, assumendo che nella popolazione il rapporto tra i sessi sia, in realtà, pari a 50:50 (la nostra ipotesi nulla). Questa probabilità è pari a 0,13, che è un valore troppo alto per scartare l'ipotesi nulla.

- **Scenario 4:** supponiamo ora che il nostro campione comprenda 17 femmine e 3 maschi. Se nella nostra popolazione i due sessi sono davvero equamente rappresentati, allora il campione ha una probabilità assai più bassa rispetto al campione riportato nello scenario 3 di rappresentare la popolazione effettiva. La probabilità di un campione così estremo è pari a 0,0002 e ci indurrebbe a scartare l'ipotesi nulla, concludendo che nella popolazione esistono più femmine che maschi.

Ciò si comprende anche intuitivamente. Quando si scelgono a caso 20 individui da una popolazione in cui i sessi sono equamente rappresentati, allora saremo sorpresi se quasi tutti gli individui del campione fossero femmine, mentre non saremo sorpresi se avessimo scelto un po' più femmine che maschi (oppure un po' più maschi che femmine). Ma esattamente quante femmine dobbiamo avere nel nostro campione per poter dedurre con una certa sicurezza che esse costituiscono la maggioranza della popolazione? E quanto è alta l'affidabilità di questa nostra conclusione? La statistica ci permette di rispondere con precisione a queste domande.

**La significatività statistica: evitare falsi positivi** Ogni qualvolta che verifichiamo un'ipotesi, calcoliamo la probabilità appena discussa e indichiamo questo valore come **valore P** del nostro test. In particolare, il valore *P* corrisponde alla probabilità di ottenere dati così estremi come i nostri dati (soltanto per caso) se l'ipotesi nulla è, in effetti, vera. In altre parole, si tratta della probabilità che il solo caso produrrebbe dati che differiscono dall'ipotesi nulla quanto i nostri dati differiscono dall'ipotesi nulla. Il modo in cui misuriamo la differenza tra i nostri dati e l'ipotesi nulla dipende dal tipo di dati ricavati dal nostro campione (qualitativi o quantitativi, singola variabile o variabili multiple, differenze tra variabili, correlazioni tra variabili ecc.).

Per molti test di statistica i valori di *P* possono essere calcolati matematicamente. Una possibilità è quella di quantificare quanto i dati si discostano dall'ipotesi nulla e verificare poi con appropriate tabelle (disponibili in molti libri di statistica o su internet) la probabilità che il solo caso produca una tale differenza. Tuttavia, la maggior parte degli studiosi individua i valori di *P* principalmente con software di statistica piuttosto che ricorrere a un calcolo a mano o a consultare apposite tabelle. Indipendentemente dalla tecnica e dalla tecnologia usata, le tappe più importanti dell'analisi statistica sono comunque in mano al singolo ricercatore: la formulazione di appropriate ipotesi nulla e alternativa, la scelta del test di statistica da eseguire e le giuste conclusioni da trarre dai risultati ottenuti.

Dopo aver calcolato il valore di *P* dei nostri dati, dobbiamo decidere se esso è sufficientemente piccolo da concludere che i nostri dati sono inconsistenti con l'ipotesi nulla. Ciò accade confrontando il valore di *P* con un valore di so-

glia denominato **livello di significatività**, che viene spesso scelto prima ancora di effettuare un qualsiasi altro tipo di calcolo. Rigettiamo l'ipotesi nulla soltanto quando il valore di  $P$  è inferiore o uguale al livello di significatività, indicato da  $\alpha$ . Ciò assicura che, se l'ipotesi nulla è vera, nel peggiore dei casi abbiamo una probabilità  $\alpha$  di rigettarla accidentalmente. Pertanto, più è piccolo il valore di  $\alpha$  e più è bassa la probabilità di commettere un errore di tipo 1 (cella in basso a sinistra nella Figura B10). Il livello di significatività più comunemente usato è  $\alpha = 0,05$ , che limita al 5% la probabilità di commettere un errore di tipo 1.

Se il nostro test di statistica raggiunge un valore di  $P$  inferiore al nostro livello di significatività  $\alpha$ , allora concludiamo che l'effetto descritto dalla nostra ipotesi alternativa è statisticamente significativo a livello  $\alpha$  e rigettiamo l'ipotesi nulla. Se il valore di  $P$  è maggiore di  $\alpha$ , allora concludiamo di non essere in grado di rigettare l'ipotesi nulla. In questo caso non scartiamo l'ipotesi alternativa, ma concludiamo di non avere ancora elementi sufficienti per supportarla.

**Potenza: evitare falsi negativi** La **potenza** di un test di statistica corrisponde alla probabilità di rigettare giustamente l'ipotesi nulla quando questa è falsa (cella in basso a destra nella Figura B10). Pertanto, più è elevata la potenza del test e minore è la probabilità di commettere un errore di tipo 2 (cella in alto a destra nella Figura B10). La potenza di un test può essere calcolata e i calcoli di questo tipo possono essere usati per migliorare la metodologia usata. In generale, si possono effettuare diversi passaggi per aumentare la potenza, evitando pertanto falsi negativi:

- **Ridurre il livello di significatività  $\alpha$ .** Più è alto il valore di  $\alpha$  e più è difficile rigettare l'ipotesi nulla, persino quando questa è effettivamente falsa.
- **Aumentare la dimensione numerica del campione.** Più dati abbiamo a disposizione e più è probabile trovare evidenze contro l'ipotesi nulla quando questa è effettivamente falsa.
- **Ridurre la variabilità del campione.** Maggiore è la variabilità del campione e più è difficile osservare un chiaro effetto (l'ipotesi alternativa) quando questa effettivamente esiste.

È sempre una buona idea allestire l'esperimento in modo tale da ridurre ogni variabilità che potrebbe oscurare il profilo che si sta cercando di individuare. Per esempio, è possibile che la probabilità che un bambino contragga l'influenza dipenda dal fatto che esso vive in un ambiente affollato (come quello urbano) o in uno meno affollato (rurale). Per ridurre la variabilità del campione, si potrebbe scegliere di analizzare l'efficacia di un nuovo vaccino contro l'influenza prendendo in considerazione soltanto bambini provenienti da uno di questi due ambienti. Dopo aver minimizzato la variabilità di questo tipo, si possono usare calcoli di potenza per scegliere la giusta combinazione di  $\alpha$  e la dimensione numerica del campione nel tentativo di ridurre fino a livelli accettabili i rischi di commettere errori di tipo 1 e 2.

Esiste un compromesso tra errori di tipo 1 ed errori di tipo 2: infatti, con l'aumentare di  $\alpha$  il rischio di commettere errori di tipo 1 diminuisce, mentre aumenta il rischio di

commettere errori di tipo 2. Come discusso in precedenza, gli scienziati tendono a preoccuparsi soprattutto degli errori di tipo 1. In altre parole, si ritiene meno opportuno accettare un'ipotesi sbagliata che rinunciare a fare una nuova scoperta. Pertanto, si preferisce usare valori bassi di  $\alpha$ . Tuttavia, esistono molti scenari nel mondo reale in cui potrebbe essere meno opportuno commettere un errore di tipo 2 piuttosto che uno di tipo 1. Supponiamo, per esempio, di testare un nuovo farmaco contro il raffreddore per quanto riguarda i suoi eventuali effetti collaterali (che mettono a rischio la sopravvivenza degli individui che l'assumono). L'ipotesi nulla è che il farmaco non abbia effetti collaterali. Un errore di tipo 2 potrebbe indurre gli enti regolatori ad approvare un farmaco dannoso, mettendo a rischio la vita degli individui. Al contrario, un errore di tipo 1 porterebbe semplicemente alla presenza di un farmaco contro il raffreddore in meno sugli scaffali delle farmacie. In questi casi i legislatori si accingono a evitare un errore di tipo 2 anche se, così facendo, aumentano il rischio di commettere un errore di tipo 1.

**Inferenza statistica con dati quantitativi** Esistono molti tipi di inferenza statistica per i dati quantitativi. Quando si misura una singola variabile quantitativa, come il peso alla nascita negli agnelli, la concentrazione di calcio nel sangue delle donne gravide o il tasso di migrazione degli uccelli, si desidera spesso determinare il valore medio della popolazione dalla quale è stato raccolto il campione. Tuttavia, la media di un campione casuale non sarà necessariamente la stessa o simile alla media della popolazione. Supponiamo di voler conoscere il peso medio di un agnello alla nascita in una particolare fattoria. Per caso potremmo finire per rac-

**Figura B11** Il test t

**METODI DI RICERCA**

Cos'è il test t? Si tratta di un metodo standard che permette di determinare se le medie di due gruppi differiscono statisticamente tra loro.

**Tappa 1:** Si determinano le ipotesi nulla e alternativa:  
 $H_0$ : le due popolazioni possiedono la stessa media.  
 $H_A$ : le due popolazioni possiedono medie differenti.

**Tappa 2:** Si sceglie il livello di significatività,  $\alpha$ , per limitare il rischio di commettere un errore di tipo 1.

**Tappa 3:** Si calcola il test statistico:  $t_s = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Nota:  $y_1$  e  $y_2$  sono le medie dei campioni;  $s_1$  e  $s_2$  sono le deviazioni standard dei campioni e  $n_1$  e  $n_2$  sono le dimensioni numeriche dei campioni. Attenzione, vedi originale per scrivere  $y_1$  e  $y_2$

**Tappa 4:** Si usa il test di statistica per determinare se i dati sono consistenti con l'ipotesi nulla:  
 Si calcola il valore di  $P$  usando software di statistica oppure lo si calcola a mano consultando apposite tabelle di statistica.

**Tappa 5:** Si traggono le conclusioni dal test:  
 Se  $P \leq \alpha$ , allora si rigetta  $H_0$  e si conclude che la distribuzione della popolazione differisce in maniera significativa.  
 Se  $P > \alpha$ , allora non abbiamo elementi sufficienti per concludere che le medie differiscono tra loro.



colgiere un campione che include un eccesso di agnelli con un peso basso e dunque potremmo ottenere una media inferiore a quella della popolazione complessiva.

Per dedurre la media della popolazione in base ai dati ottenuti dal campione possiamo calcolare un **intervallo di confidenza per la media**. Si tratta di un ambito di valori statisticamente determinati che è centrato sulla media del campione e che include probabilmente la media della popolazione. Per esempio, basandoci sul campione di 34 pesi di pesci di *Abramis brama* prelevati dal lago Laengelmavesi (vedi Tabella B2; Figura B4), l'intervallo di confidenza pari al 95% per il peso medio varia tra 554 grammi e 698 grammi. Il reale peso medio in questa specie ittica cade probabilmente, ma non necessariamente, in questo ambito di valori.

Spesso i biologi vogliono confrontare i valori medi in due o più gruppi; per esempio il peso degli agnelli alla nascita in diverse fattorie, la concentrazione di calcio nelle donne gravide all'inizio e al termine della gravidanza oppure il tasso di migrazione degli uccelli appartenenti a specie diverse. Basandoci sulle medie e sulle deviazioni standard di ogni campione, i biologi deducono se le medie nelle diverse popolazioni differiscono o meno statisticamente tra loro. Esistono numerosi metodi statistici che differiscono l'uno dall'altro e il metodo corretto da usare dipende dal numero di gruppi, dall'allestimento dell'esperimento e dalla natura dei dati.

Nella **Figura B11** sono descritte le singole tappe del test-*t*, un semplice metodo che permette di confrontare le medie di due diversi gruppi. Per illustrare il metodo possiamo applicare il test *t* ai dati raccolti presso il lago Laengelmavesi per cercare di stabilire se le due specie ittiche *Abramis brama* e *Leusiscus idus* possiedono pesi medi che differiscono tra loro in maniera significativa. Iniziamo affermando le nostre ipotesi e scegliendo un livello di significatività:

$H_0$ : *Abramis brama* e *Leusiscus idus* possiedono lo stesso peso medio.

$H_A$ : *Abramis brama* e *Leusiscus idus* possiedono pesi medi differenti.

$\alpha = 0,05$

Il test tiene conto delle medie, delle deviazioni standard e delle dimensioni numeriche dei due campioni:

$$t_s = \frac{626 - 531}{\sqrt{\frac{207^2}{34} + \frac{310^2}{6}}} = 0,724$$

Per individuare il valore di *P* possiamo usare software di statistica oppure consultare uno dei siti gratuiti di statistica presenti in internet; in questo caso  $P = 0,497$ . Poiché *P* è considerevolmente più alto di  $\alpha$ , non rigettiamo l'ipotesi nulla e concludiamo che il nostro studio non fornisce prove a sufficienza per affermare che le due specie possiedono pesi medi differenti.

Per ulteriori dati relativi agli intervalli di confidenza, ai test *t* e ad altri test di statistica si rimanda a un libro di testo di statistica di base.

**Figura B12** Il test del chi quadro per la bontà dell'adattamento

### METODI DI RICERCA

Che cos'è il test del chi quadro per la bontà dell'adattamento? Si tratta di un metodo standard che permette di determinare se un campione proveniva da una popolazione con una specifica distribuzione.

**Tappa 1:** Si determinano le ipotesi nulla e alternativa:

$H_0$ : la popolazione possiede la distribuzione specificata.

$H_A$ : la popolazione non possiede la distribuzione specificata.

**Tappa 2:** Si sceglie il livello di significatività,  $\alpha$ , per limitare il rischio di un errore di tipo 1.

**Tappa 3:** Si determinano la frequenza osservata e la frequenza attesa per ciascuna categoria:

La frequenza osservata di una categoria corrisponde semplicemente al numero di osservazioni nel campione di quel tipo.

La frequenza attesa di una categoria corrisponde alla probabilità della categoria specificata in  $H_0$  moltiplicata per la dimensione numerica complessiva del campione.

**Tappa 4:** Si calcola il test statistico:  $\chi_s^2 = \sum_{i=1}^c \frac{(O_i - E_i)^2}{E_i}$

Nota: *C* è il numero totale di categorie,  $O_i$  è la frequenza osservata della categoria *i* ed  $E_i$  è la frequenza attesa della categoria *i*.

**Tappa 5:** Si usa la statistica del test per determinare se i dati sono consistenti con l'ipotesi nulla:

Si calcola il valore *P* usando software di statistica o a mano con l'uso di apposite tabelle.

**Tappa 6:** Si traggono le conclusioni dal test:

Se  $P \leq \alpha$ , allora si rigetta  $H_0$  e si conclude che la distribuzione della popolazione differisce in maniera significativa dalla distribuzione specificata da  $H_0$ .

Se  $P > \alpha$ , allora non abbiamo elementi sufficienti per concludere che la popolazione possiede una distribuzione differente.

**Inferenza statistica con dati qualitativi** Con i dati qualitativi cerchiamo spesso di dedurre la distribuzione delle diverse categorie all'interno della popolazione dalla quale provengono i nostri campioni. Nel caso più semplice abbiamo una singola variabile qualitativa con due o più categorie di dati qualitativi. Se vi sono soltanto due categorie di dati, allora possiamo costruire un **intervallo di confidenza per la proporzione** della popolazione che appartiene a una delle due categorie in questione. Questa è una gamma di valori ottenuta statisticamente, che è centrata sulla stessa proporzione e che include probabilmente la proporzione nella popolazione. Se vi sono tre o più categorie di dati, allora possiamo usare il test del **chi quadro per la bontà dell'adattamento** per determinare se la distribuzione delle diverse categorie nella popolazione è consistente con una particolare distribuzione.

La **Figura B12** sottolinea le tappe del test del chi quadro per la bontà dell'adattamento. Come esempio consideriamo i dati descritti nella Tabella B1. Molte specie di piante seguono una semplice genetica mendeliana in cui le piante parentali producono una progenie con tre diversi colori dei fiori in un rapporto di 2:1:1. Tuttavia, un botanico ritiene che queste particolari piante di poinsettia seguano un quadro genetico differente, che non produce una progenie

di piante con fiori rispettivamente rossi, rosa e bianchi in un rapporto di 2:1:1. Si può usare il test del chi quadro per la bontà dell'adattamento per determinare se i dati sono o meno consistenti con questo rapporto e pertanto se questa semplice spiegazione genetica è valida oppure no. Iniziamo con la formulazione delle nostre ipotesi e del livello di significatività:

$H_0$ : la progenie di questo tipo di incrocio ha le seguenti probabilità di sviluppare fiori con ognuno dei colori:

$$\Pr(\text{rosso}) = 0,50, \Pr(\text{rosa}) = 0,25, \Pr(\text{bianco}) = 0,25$$

$H_A$ : almeno una delle probabilità di  $H_0$  non è corretta.  
 $\alpha = 0,05$

Usiamo le probabilità di  $H_0$  e la dimensione numerica del campione per calcolare le frequenze attese:

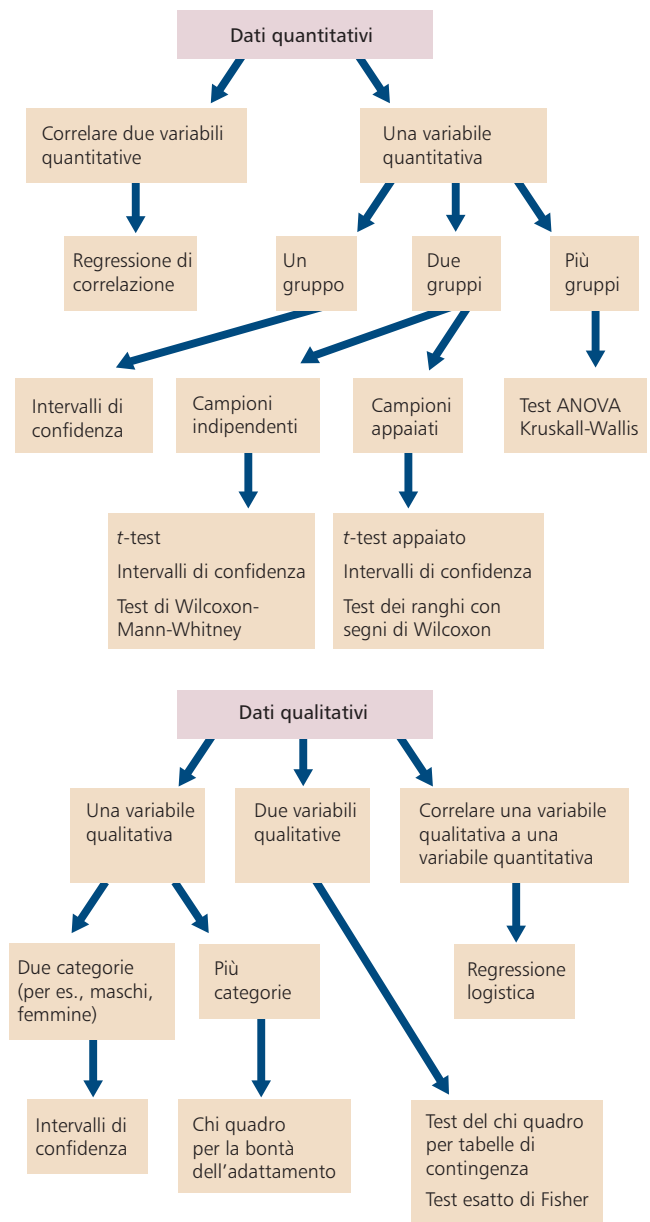
	Rosso	Rosa	Bianco
Osservato	108	34	40
Atteso	$(0,50)(182) = 91$	$(0,25)(182) = 45,5$	$(0,25)(182) = 45,5$

Basato su queste quantità usiamo il test statistico del chi quadro:

$$\chi_s^2 = \sum_{i=1}^c \frac{(O_i - E_i)^2}{E_i} = \frac{(108 - 91)^2}{91} + \frac{(34 - 45,5)^2}{45,5} + \frac{(40 - 45,5)^2}{45,5} = 6,747$$

Per questo risultato, usando software di statistica, il valore di  $P$  è pari a 0,0343. Poiché  $P$  è inferiore ad  $\alpha$ , rigettiamo l'ipotesi nulla e concludiamo che il botanico ha ragione: il quadro di colore dei fiori non può essere spiegato con il semplice modello mendeliano.

Questa appendice intende fornire soltanto una breve introduzione ai concetti dell'analisi statistica, con alcuni test che fungono da esempio. Nella **Figura B13** sono riassunti alcuni dei test di statistica comunemente usati, che si potrebbero incontrare nel corso degli studi di biologia.



**Figura B13** Alcuni metodi comunemente usati nell'inferenza statistica

Questo diagramma di flusso illustra alcuni dei metodi comunemente usati di inferenza statistica per differenti combinazioni di dati. Le descrizioni dettagliate di questi metodi si trovano nella maggior parte dei manuali di base di biostatistica.