12 Dati e ipotesi

La statistica, con il supporto teorico della probabilità, studia come analizzare ed elaborare dati sperimentali. In questo paragrafo introduciamo le nozioni fondamentali di popolazione e campionamento, e mostriamo alcuni metodi per visualizzare i dati, aspetto fondamentale in tutta l'analisi statistica.

La statistica permette di sintetizzare in pochi numeri grandi quantità di dati. Mostriamo l'uso e il significato dei principali indicatori della statistica descrittiva.

Se il campione viene scelto in modo casuale, la media campionaria è una variabile aleatoria, e i metodi della teoria della probabilità permettono di comprendere la relazione tra questa variabile casuale e le caratteristiche statistiche complessive dell'intera popolazione. Nello stesso modo viene analizzata la relazione tra la varianza campionaria e quella di popolazione.

I test statistici sono il principale metodo per verificare l'accordo tra dati e teoria. Ne esistono molti a seconda del tipo di verifica che si deve fare. Tra gli esempi mostriamo la classica verifica statistica delle leggi di Mendel.

Anche le leggi matematiche che possono legare due diverse variabili sono oggetto di analisi statistica. In particolare la covarianza e il coefficiente di correlazione mostrano se e come due variabili sono in relazione tra loro. I più semplici esempi di relazioni matematiche tra due variabili possono essere determinati attraverso la regressione, lineare e non.

978-88-08-18484-

12.1 Dati statistici

Figura 12.1
Un esempio di popolazione che può essere analizzata con metodi statistici è quella costituita dalla cinquantina di lupi della sottospecie appenninica *Canis lupus italicus* (l'unica sopravvissuta in Italia), che vive nel Parco Nazionale d'Abruzzo.

La statistica studia l'insieme delle tecniche necessarie a raccogliere e elaborare dati, ottenuti dall'osservazione di fenomeni collettivi (fenomeni demografici, economici, naturali, genetici ecc.) che coinvolgono molti elementi.

Si chiama **popolazione statistica** l'insieme di tutti gli elementi (individui, geni, cellule ecc.) che si vogliono studiare.
Una popolazione è dunque un insieme composto da un numero finito o infinito di elementi.

Esempio 12.1.1 Popolazioni statistiche

L'insieme dei lupi del Parco Nazionale d'Abruzzo, degli abitanti di Milano, dei valori di temperatura rilevati a Roma alle ore 14 del primo giugno degli anni dal 1998 al 2001, l'altezza degli alunni di una classe di 30 bambini di 6 anni, sono alcuni esempi di popolazioni composte da un numero finito di elementi.
Gli esemplari di *Homo sapiens*, le lunghezze della coda di tutti i gatti, le stelle della Via Lattea sono elementi di insiemi finiti, ma il loro numero è così grande che, a volte, sarà utile considerare la popolazione come infinita.

Figura 12.2
Un invito a registrarsi per il censimento canadese del 2006. La statistica ha avuto origine con i censimenti: già 4000 anni fa i Cinesi utilizzavano tavole di statistiche agricole e nella Bibbia, IV Libro di Mosè, è ricordato un censimento degli uomini in grado di combattere. Molto famoso è il censimento ordinato da Augusto nell'anno della nascita di Cristo. In epoca moderna, documenti di grande interesse statistico sono i registri dello stato civile, prima conservati presso le Chiese, poi presso gli Uffici Civili. In tutti questi casi, si tratta di pure raccolte di dati, senza alcuna analisi. Solo dopo l'introduzione del calcolo delle probabilità la statistica diventa una disciplina a carattere matematico.

Anche se l'oggetto di un'indagine è un'intera popolazione statistica, in genere non è possibile raccogliere dati su ogni singolo elemento della popolazione. Si considera dunque solo un sottoinsieme, che si suppone sia **rappresentativo** dell'intera popolazione; questo sottoinsieme prende il nome di **campione**. Uno dei compiti della statistica è quello di stabilire con quali metodi scegliere campioni che siano rappresentativi. Non tratteremo in dettaglio questo argomento, vasto e complesso, e, per semplicità, considereremo un solo tipo di campione.

Si chiama **campione casuale** una sequenza di elementi scelti a caso dalla popolazione, in modo che ogni elemento abbia la stessa probabilità di far parte del campione.

I metodi e le finalità della statistica sono diversi a seconda che la popolazione sia osservata per intero oppure solo in parte, attraverso un campione. Se l'indagine è totale, come per esempio nel caso del censimento di una città, i metodi statistici forniscono la sintesi quantitativa completa dei fenomeni studiati, e permettono una visione semplificata ma efficace del fenomeno stesso. La parte della statistica che si occupa di questi aspetti si chiama **statistica descrittiva**.
Se invece l'indagine viene effettuata su un campione (come è il caso dei sondaggi d'opinione), e quindi è parziale, la prospettiva di analisi è completamente diversa. È infatti possibile estendere le informazioni dedotte dal campione a tutta la popolazione, ma i risultati ottenuti, contrariamente a quanto accade nel caso precedente, conterranno un certo grado di incertezza. La **statistica inferenziale** studia appunto come e con quale precisione si possono descrivere le caratteristiche di una popolazione partendo dalle informazioni relative a un campione.

Fissata una popolazione, si chiamano **variabili statistiche** tutte quelle caratteristiche che variano al variare dei componenti della popolazione. Le variabili che sono espresse qualitativamente sono dette **attributi**, quelle che sono espresse quantitativamente sono dette **misurabili**.

Esempio 12.1.2 Variabili qualitative e quantitative

Il colore bianco, fulvo, nero ecc. della pelliccia degli esemplari di una certa specie, il sesso (maschio o femmina) sono esempi di attributi. L'età in mesi degli esemplari di lupo del Parco d'Abruzzo, oppure il numero dei cuccioli nati da ogni femmina, sono variabili misurabili discrete. La temperatura di Roma, rilevata alle ore 14 del primo giugno di ogni anno, è un esempio di variabile misurabile continua. ■

Figura 12.3
Le pellicce dei mammiferi possono essere di molti colori diversi. In questo caso il colore è una variabile statistica qualitativa.

Risultati statistici affidabili si basano sempre su dati ottenuti dalle osservazioni e dalle misure, sia di un campione che di tutta la popolazione, ordinati e riassunti in poche informazioni semplici. Queste informazioni sono dette **empiriche**, per distinguerle da quelle **teoriche**, offerte da un eventuale modello che spieghi i dati osservati.

Il passo successivo alla raccolta dei dati è quello della loro rappresentazione grafica. Esistono vari metodi di rappresentazione dei dati, qui ne richiamiamo brevemente qualcuno utilizzando esempi particolari.

Esempio 12.1.3 Diagrammi di punti e istogrammi

Consideriamo un campione di 10 esemplari di gatte, ciascuna identificata da un numero da 1 a 10. A ciascuna possiamo associare il numero di cuccioli partoriti in una fissata stagione di un certo anno. Supponiamo che le osservazioni siano quelle riassunte nella seguente tabella.

Gatta	1	2	3	4	5	6	7	8	9	10
Numero di cuccioli	2	1	3	2	4	2	2	2	1	3

I dati si possono rappresentare in un piano cartesiano disegnando 10 punti le cui coordinate sono, rispettivamente, il numero che identifica ogni esemplare e il numero dei cuccioli generati. In questo modo si ottiene il **diagramma di punti** mostrato nel primo grafico di figura 12.4.

Un'altra rappresentazione grafica, detta **istogramma** o anche "rappresentazione a canne d'organo", consiste nella rappresentazione dei dati come rettangoli: la base, uguale per tutti i rettangoli, rappresenta il numero che identifica ogni gatta e l'altezza è uguale al numero dei cuccioli generati. L'istogramma, rappresentato nel secondo grafico di figura 12.4, è di più facile lettura rispetto al diagramma di punti. In particolare permette di cogliere, con un colpo d'occhio, che il maggior numero delle gatte ha partorito due cuccioli. ■

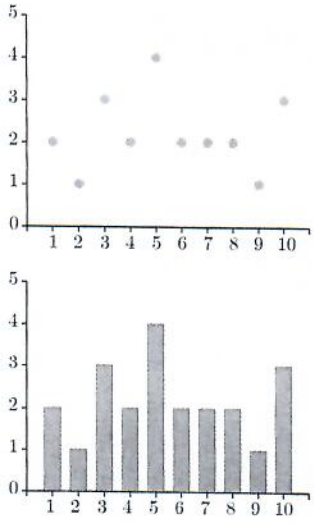


Figura 12.4
Diagramma di punti e istogramma per lo stesso insieme di 10 dati.

Precisiamo ora alcune notazioni relative ai campionamenti.

 978-88-08-18484-

- Indicheremo con N la **dimensione del campione**, cioè il numero totale dei dati raccolti, e con $X_1, X_2, \ldots, X_N$ i dati, cioè i valori assunti nel campione dalla variabile statistica X.
- In molti casi i dati $X_1, X_2, \ldots, X_N$ sono ripetuti, cioè assumono un numero finito di valori discreti distinti $x_1, x_2, \ldots, x_n$, $n \leq N$. In tal caso indicheremo con F_i il numero di dati uguali a x_i (**frequenza assoluta**) mentre $f_i = F_i/N$ indicherà la **frequenza relativa**.

Esempio 12.1.4 Frequenze

I dati nella seconda riga della tabella dell'esempio 12.1.3 indicano i valori che assum la variabile statistica X = "numero di cuccioli":

$$X_1 = X_4 = X_6 = X_7 = X_8 = 2, \qquad X_2 = X_9 = 1, \qquad X_3 = X_{10} = 3, \qquad X_5 =$$

(il numero a pedice di X è il numero identificativo della gatta).
Il numero dei dati è $N = 10$, ma i valori distinti che X assume sono soltanto $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$ e le corrispondenti frequenze assolute sono $F_1 = F_2 = 5$, $F_3 = 2$, $F_4 = 1$. La somma delle frequenze assolute è pari alla dimensio del campione, che è $N = 10$. Dividendo per $N = 10$ i valori delle frequenze assolu si ottengono le frequenze relative $f_1 = 0.2$, $f_2 = 0.5$, $f_3 = 0.2$, $f_4 = 0.1$, la c somma è 1.

Se il campione è molto grande, invece di rappresentare l'istogramma d dati come nell'esempio 12.1.3, è utile rappresentare l'istogramma del frequenze.

Esempio 12.1.5 Istogrammi di frequenze

I seguenti dati sono le altezze in centimetri di un campione di 120 studentesse d primo anno di un corso universitario dell'anno accademico 2012.

170	162	174	160	159	164	168	168	163	175	160	16
168	170	162	171	164	166	155	175	159	168	163	16
159	179	169	159	170	162	162	174	169	158	166	16
166	156	164	165	159	168	169	164	164	158	171	17
176	150	165	174	181	160	169	181	155	164	175	17
161	155	170	172	165	160	156	170	165	166	172	17
164	176	181	165	167	160	171	164	169	166	167	16
168	168	184	159	160	172	175	159	169	161	169	16
166	162	162	172	164	171	169	155	171	170	160	16
174	166	169	165	168	166	174	160	158	168	166	16

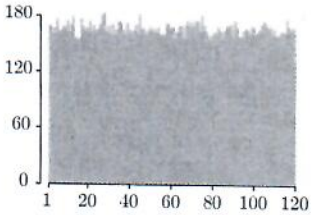


Figura 12.5
Un istogramma piuttosto inutile. In ascissa il numero identificativo della studentessa, in ordinata la sua altezza.

È evidente che la rappresentazione di questi dati in un diagramma di punti o in istogramma è poco utile (fig. 12.5).
Data la grande quantità di dati, conviene invece raggruppare i valori delle altezze **classi** (cioè in sottoinsiemi). Osserviamo che tutti i valori della tabella sono inter compresi tra 150 e 184. Raggruppiamo questi dati in 7 intervalli di 4 centimetri lunghezza:

$$I_1 = [150, 154], \; I_2 = [155, 159], \; I_3 = [160, 164], \; I_4 = [165, 169],$$

$$I_5 = [170, 174], \; I_6 = [175, 179], \; I_7 = [180, 184].$$

I punti medi degli intervalli sono

$$m_1 = 152,\ m_2 = 157,\ m_3 = 162,\ m_4 = 167, m_5 = 172,\ m_6 = 177,\ m_7 = 182.$$

Questi valori rappresentano, approssimativamente, i valori contenuti nel corrispondente intervallo, nel senso che, se un'altezza è per esempio nell'intervallo I_2, sarà approssimativamente pari a $m_2 = 157$ centimetri.
Possiamo ora costruire un istogramma in cui si rappresentano sulla retta reale orizzontale gli intervalli che abbiamo scelto e su quella verticale il numero di dati (la frequenza assoluta) contenuti in ciascun intervallo (fig. 12.6). Dalla tabella dei dati si ottiene che le frequenze sono

$$F_1 = 1,\ F_2 = 17,\ F_3 = 32,\ F_4 = 37,\ F_5 = 21,\ F_6 = 8,\ F_7 = 4.$$

Nella costruzione degli istogrammi di questo tipo, è importante decidere con buon senso la dimensione degli intervalli e il conseguente numero di classi. Nei grafici della figura 12.7 sono rappresentati, per questi stessi dati, l'istogramma ottenuto utilizzando come lunghezza di base un centimetro, e l'istogramma con lunghezza di base 20 cm. Entrambi sono meno utili di quello di figura 12.6, il primo perché ha troppe informazioni poco interessanti, il secondo perché ne ha troppo poche. ■

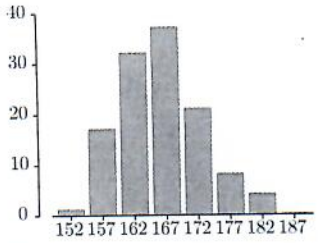


Figura 12.6
Istogramma delle frequenze; in ascissa gli intervalli in cui si raggruppano i dati, in ordinata le frequenze assolute con cui compaiono.

Esempio 12.1.6 Diagramma a fusto e foglie

Una rappresentazione dei dati simile a un istogramma si può ottenere anche senza utilizzare grafici, ma costruendo il diagramma detto **a fusto e foglia** (in inglese *stem and leaf*). In questa rappresentazione ogni dato numerico è scomposto in due parti: il fusto composto da tutte le cifre del dato tranne l'ultima, e la foglia che è l'ultima cifra. Per esempio il numero 158 si scompone in 15 (fusto) e 8 (foglia). Il numero di ripetizioni dell'ultima cifra è uguale alla frequenza del dato. Tutte le cifre del fusto e delle foglie sono poi ordinate in ordine crescente. Il diagramma fusto e foglia dei dati della tabella del precedente esempio è il seguente

15 0	**16** 0000000000	**17** 000000	**18** 111
5555	11	111111	4
66	222222	22222	
888	333	44444	
9999999	444444444	55555	
	555555	66	
	66666666666	9	
	777		
	888888888		
	9999999999		

In questa rappresentazione, in cui si considerano necessariamente intervalli di 10 centimetri, si vede immediatamente che la maggior parte delle altezze ha valori compresi nell'intervallo [160, 169], e che tra essi i più frequenti sono 160, 166 e 169. ■

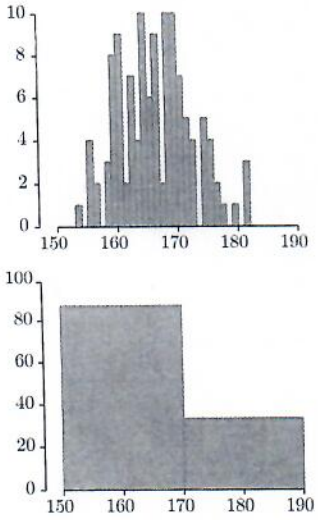


Figura 12.7
Un istogramma con troppe classi e uno che ne ha troppo poche.

Esempio 12.1.7 Aerogramma

Se si vogliono evidenziare i rapporti di proporzione fra i dati, in particolare quelli relativi ad attributi, si usano i cosiddetti **aerogrammi** o **diagrammi a torta**.
Supponiamo di considerare un campione di 100 individui e come variabile il colore dei capelli, che possono essere biondi, rossi, castani, neri, bianchi. Rileviamo 50 individui (la metà del totale) con capelli castani, 25 (un quarto del totale) con capelli neri, 5 (un ventesimo del totale) con capelli bianchi (un decimo del totale), 10 con capelli biondi e 10 con capelli rossi. Nel corrispondente aerogramma (fig. 12.8) si rappresenta una regione circolare divisa in spicchi, in cui ogni spicchio ha l'angolo scelto in proporzione alla frequenza dell'attributo che descrive. Per esempio, i biondi sono 1/10 del campione, dunque l'angolo del relativo spicchio è $2\pi/10$. Si noti che le aree degli spicchi sono proporzionali agli angoli e quindi anch'esse proporzionali alla frequenza dell'attributo che descrivono. ■

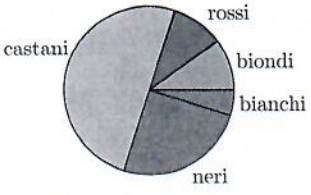


Figura 12.8
Esempio di aerogramma.

978-88-08-18484-

12.2 Riassumere i dati in pochi numeri

Rappresentare graficamente i dati è un modo semplice per osservarne l caratteristiche principali, ma spesso è necessario sintetizzare ancora d più le informazioni. A questo scopo si utilizzano alcuni indici numeri che riassumono le principali caratteristiche matematiche dei dati.

Consideriamo i dati descritti nell'esempio 12.1.5, raggruppati nelle cla individuate dai 7 intervalli, come in figura 12.6. La frequenza massima ottiene per l'intervallo centrato in 167; questo numero dunque è piutto rappresentativo del complesso dei dati. Generalizzando questo esemp definiamo un primo indice riassuntivo.

Se i dati sono espressi mediante la loro appartenenza a diverse class si chiama **classe modale** la classe di frequenza massima. Se le clas sono individuate da numeri, il numero che contraddistingue la class modale prende il nome di **moda**.

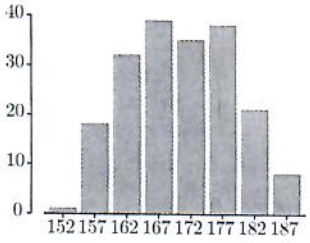


Figura 12.9
Istogramma delle frequenze di un campione di 192 studenti. La distribuzione è bimodale, perché esistono due picchi separati, uno in 167, l'altro in 177.

Per esempio, i dati rappresentati nell'aerogramma della figura 12.8 ha no come classe modale "castani", perché la frequenza di questa class maggiore delle altre.

Nel caso dell'istogramma della figura 12.6, la distribuzione delle frequ ze è una funzione a valori discreti che cresce fino a raggiungere il s massimo nella moda 167, e poi decresce. Una distribuzione di frequ con questa caratteristica è detta **unimodale**. Non tutte le distribuzi sono di questo tipo, come mostriamo nel seguente esempio.

Esempio 12.2.1 Distribuzioni bimodali

In figura 12.9 è riportato l'istogramma dei dati dell'altezza di un campione di studenti, di cui 120 sono quelli già analizzati nell'esempio 12.1.5. Come si vede distribuzione delle frequenze non è unimodale ma **bimodale**, perché esistono picchi, uno nell'intervallo centrato in 167 cm, l'altro nell'intervallo centrato in 177

Non è tipico che una distribuzione sia bimodale (o anche trimodale o, più in gene multimodale). In particolare, nel caso di caratteristiche fisiologiche, ci aspettiamo sola classe modale, con le frequenze assolute che diventano via via più piccole q più ci allontaniamo dalla moda. Infatti una caratteristica fisiologica è spesso l'ef di molte cause indipendenti, e sommare molti effetti indipendenti dà una distrib gaussiana (*vedi* par. 11.4), che ha un solo massimo.

In genere, se una distribuzione è bimodale, possiamo supporre che il campio la sovrapposizione di due campioni con caratteristiche differenti. In questo es pio particolare, in effetti, alle altezze delle 120 studentesse sono state aggiu altezze di 72 studenti maschi, che sono mediamente più alti. In figura 12.10 rap sentiamo affiancati gli istogrammi dei due campioni: sono ben visibili le due differenti.

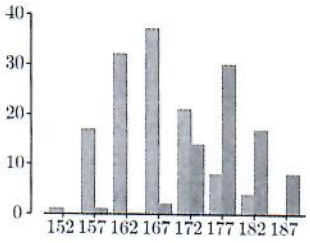


Figura 12.10
Istogrammi separati del campione di femmine (in arancione) e del campione di maschi (in verde) (*vedi anche* fig. 11.42).

Sintetizzare in modo efficace dei dati permette anche il loro confro

Esempio 12.2.2 Confrontare dati – I

All'inizio di ogni anno scolastico un insegnante, per avere indicazioni sullo stato d conoscenze degli studenti nella sua materia di insegnamento, propone un comp

Ripete questa prova per 3 anni di seguito. I risultati dei compiti, espressi in decimi, sono i seguenti.

Anno 2005	6	7	7	5	8	5	4	7	5	6
	5	6	6	6	5	7	5	5	8	3
	2	4	7							
Anno 2006	3	4	7	6	7	8	8	7	6	6
	5	7	6	5	5	7	7	5	4	3
	8	6	7	6	6					
Anno 2007	8	7	4	5	7	6	4	5	5	6
	5	7	6	6	7	7	3	4	8	5

Cosa conclude l'insegnante sullo stato delle conoscenze degli studenti? In quale dei tre anni gli studenti sono più preparati?

È evidente che, in questo caso una rappresentazione grafica dei dati non aiuta molto; si può quindi iniziare osservando che i voti variano tra un voto minimo, che è 2, e un voto massimo, che è 8. In particolare, per ogni anno scolastico, si può compilare la tabella delle frequenze assolute e di quelle relative (che esprimiamo in percentuale).

Anno 2005, 23 studenti

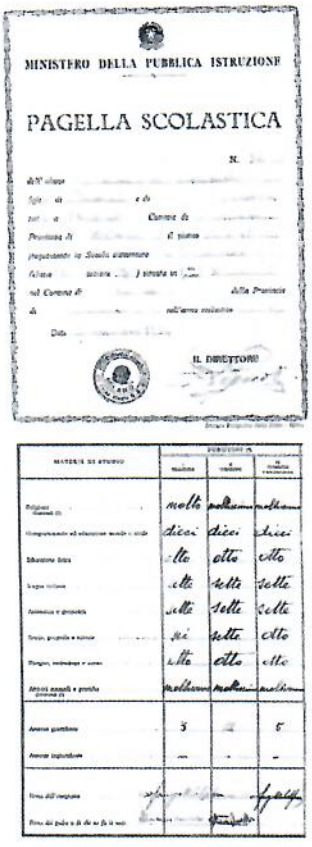
Voto	Frequenza	Percentuale
2	1	$1/23 \approx 0.04 = 4\%$
3	1	$1/23 \approx 0.04 = 4\%$
4	2	$2/23 \approx 0.09 = 9\%$
5	7	$7/23 \approx 0.30 = 30\%$
6	5	$5/23 \approx 0.22 = 22\%$
7	5	$5/23 \approx 0.22 = 22\%$
8	2	$2/23 \approx 0.09 = 9\%$

Anno 2006, 25 studenti

Voto	Frequenza	Percentuale
3	2	$2/25 = 0.08 = 8\%$
4	2	$2/25 = 0.08 = 8\%$
5	4	$4/25 = 0.16 = 16\%$
6	7	$7/25 = 0.28 = 28\%$
7	7	$7/25 = 0.28 = 28\%$
8	3	$3/25 = 0.12 = 12\%$

Anno 2007, 20 studenti

Voto	Frequenza	Percentuale
3	1	$1/20 = 0.05 = 5\%$
4	3	$3/20 = 0.15 = 15\%$
5	5	$5/20 = 0.25 = 25\%$
6	4	$4/20 = 0.20 = 20\%$
7	5	$5/20 = 0.25 = 25\%$
8	2	$2/20 = 0.10 = 10\%$


MINISTERO DELLA PUBBLICA ISTRUZIONE

PAGELLA SCOLASTICA

Figura 12.11
Una pagella di scuola elementare degli anni Cinquanta.

La moda dei dati del 2005 è il voto 5, per il 2006 ci sono due classi modali affiancate, corrispondenti ai voti 6 e 7, per il 2007 ci sono due picchi, una nel voto 5 e una nel voto 7. Confrontando questi valori si conclude che i risultati migliori sono stati ottenuti nel 2006. ■

978-88-08-184

Una descrizione sintetica dei dati si può ottenere anche dal calcolo d cosiddetti **indici di tendenza centrale**. Il più semplice di questi ind è la ben nota **media aritmetica** del campione, cioè il valore che ottiene sommando tutti i valori e dividendo per il loro numero. Co abbiamo mostrato nel paragrafo 11.2, lo stesso risultato si ottiene an sommando i dati diversi, pesati con la loro frequenza relativa.

Se X è una variabile statistica e $X_1, X_2, \ldots, X_N$ sono i valori numerici dei dati relativi a un campionamento, si chiama **media campionaria** di X la media aritmetica dei dati:

$$m_X = \frac{1}{N}\sum_{k=1}^{N} X_k\,.$$

Se i dati distinti sono x_1, x_2, …, x_n (con $n \leq N$) e compaiono con frequenze $F_1, F_2, \ldots, F_n$, si può scrivere equivalentemente

$$m_X = \frac{1}{N}\sum_{i=1}^{n} F_i x_i = \sum_{i=1}^{n} \frac{F_i}{N} x_i = \sum_{i=1}^{n} f_i x_i,$$

dove f_i sono le frequenze relative dei dati.

Esempio 12.2.3 Confrontare dati – II

Calcoliamo le medie dei voti descritti nell'esempio 12.2.2.

Anno 2005: $m = 2\cdot\frac{1}{23} + 3\cdot\frac{1}{23} + 4\cdot\frac{2}{23} + 5\cdot\frac{7}{23} + 6\cdot\frac{5}{23} + 7\cdot\frac{5}{23} + 8\cdot\frac{2}{23} \approx 5.61$

Anno 2006: $m = 3\cdot\frac{2}{25} + 4\cdot\frac{2}{25} + 5\cdot\frac{4}{25} + 6\cdot\frac{7}{25} + 7\cdot\frac{7}{25} + 8\cdot\frac{3}{25} = 5.96$

Anno 2007: $m = 3\cdot\frac{1}{20} + 4\cdot\frac{3}{20} + 5\cdot\frac{5}{20} + 6\cdot\frac{4}{20} + 7\cdot\frac{5}{20} + 8\cdot\frac{2}{20} = 5.75\,.$

Anche utilizzando questo indicatore si deduce che l'anno con i risultati migliori il 2006.

Esempio 12.2.4 Fiori

Consideriamo un campione di 30 piantine stagionali sulle quali i fiori sono distribu con le seguenti frequenze (fig. 12.12).

Fiori per pianta n_i	3	4	5	9
Numero di piante F_i	5	12	6	7

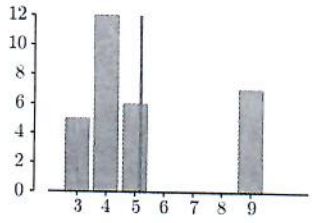


Figura 12.12
Istogramma delle frequenze. In ascissa è rappresentato il numero di fiori per pianta, e la linea rossa è il valore del numero medio di fiori per pianta.

Il numero di piante F_i è esattamente la frequenza assoluta del numero di piante c hanno n_i fiori.
Il numero medio di fiori, per pianta, è dato da

$$m = 3\cdot\frac{5}{30} + 4\cdot\frac{12}{30} + 5\cdot\frac{6}{30} + 9\cdot\frac{7}{30} = 5.2\,.$$

Si noti che non esiste nessuna pianta che porta 5.2 fiori: la media, dunque, non cor sponde necessariamente a un dato rilevato, esprime solo la "tendenza centrale" del distribuzione dei fiori sulle piante.

La media aritmetica è la stima di tendenza centrale più usata, però nc è sempre la migliore. Molti risultati di laboratorio, infatti, si riferiscor a dati di concentrazione di sostanze o a crescite cellulari e, come abbi mo visto nel capitolo 6, queste grandezze hanno solitamente variazio

esponenziali. La media aritmetica di dati di questo tipo non è il valore adatto a riassumere l'andamento di queste misure. Vediamo un esempio concreto.

Esempio 12.2.5 Media geometrica

Si misura la concentrazione molare C di una data sostanza in un campione composto da 6 prelievi di soluzione fisiologica. In accordo con il fatto che stiamo misurando concentrazioni, esprimiamo i dati in notazione scientifica.

C	$1 \cdot 10^{-6}$	$4 \cdot 10^{-6}$	$2 \cdot 10^{-5}$	$6 \cdot 10^{-3}$	$1 \cdot 10^{-2}$	$2 \cdot 10^{-1}$

La concentrazione media è circa $m_C = 3 \cdot 10^{-2}$, che è un numero molto grande se confrontato con i primi due dati, non sembra quindi realmente rappresentativo.
Come abbiamo visto anche nel caso della misura del pH (*vedi* es. 6.2.11), la parte più importante del numero che esprime una concentrazione è l'esponente. È dunque ragionevole calcolare la media degli esponenti piuttosto che la media dei valori. Per eseguire rigorosamente questo calcolo, valutiamo la media dei logaritmi in base 10 dei 6 dati; si ha

$$m_{\log C} = \frac{1}{6}\left[\log 10^{-6} + \log\left(4 \cdot 10^{-6}\right) + \log\left(2 \cdot 10^{-5}\right) + \log\left(6 \cdot 10^{-3}\right) + \log 10^{-2} + \log\left(2 \cdot 10^{-1}\right)\right] \approx -3.5\,.$$

Questo valore è l'esponente medio dei dati. Per risalire a un valore medio di concentrazione dobbiamo considerare l'esponenziale in base 10 di questo numero, che vale $10^{-3.5} \approx 3 \cdot 10^{-4}$. Il valore così ottenuto prende il nome di **media geometrica** dei dati e si indica con GM (dall'inglese *geometric mean*).
Indicando con $C_1, C_2, \ldots, C_6$ i 6 dati, abbiamo in effetti calcolato

$$m_{\log C} = \frac{1}{6}\sum_{i=1}^{6} \log C_i = \log\left(C_1 C_2 \cdots C_6\right)^{1/6}$$

e passando all'esponenziale abbiamo ottenuto

$$GM_C = 10^{m_{\log C}} = \sqrt[6]{C_1 C_2 \cdots C_6}\,. \qquad \blacksquare$$

Diamo la definizione generale di media geometrica.

Siano $X_1, X_2, \ldots, X_N$ i valori campionari di una variabile positiva X. La **media geometrica** è

$$GM_X = \sqrt[N]{X_1 X_2 \cdots X_N}\,.$$

Se i dati assumono i valori distinti $x_1, x_2, \ldots, x_n$ con frequenze assolute $F_1, F_2 \ldots, F_n$, si può scrivere, equivalentemente,

$$GM_X = \sqrt[N]{x_1^{F_1} x_2^{F_2} \cdots x_n^{F_n}}\,.$$

La media geometrica si utilizza, in particolare, se i dati hanno un andamento esponenziale. Esistono altre procedure di calcolo di medie che dipendono dalle caratteristiche dei dati (media armonica, media quadratica ecc.), ma non indugeremo su questo argomento, preferiamo invece tornare a discutere concetti di applicabilità più vasta.

Mostriamo, con un esempio, come si costruisce un altro indice di tendenza centrale generale, la **mediana**, e perché, a volte, il suo uso sia preferibile a quello della media.

978-88-08-18484-9

Figura 12.13
I piccioni viaggiatori (*Columba livia*) sono piccioni addestrati a ritornare nel luogo dal quale sono partiti. Recenti ricerche suggeriscono che per ritrovare la strada del ritorno, essi si servano di una combinazione di risorse, quali la sensibilità al campo magnetico terrestre, alla luce ultravioletta e alla luce polarizzata, nonché la capacità di riconoscere punti di riferimento sulla superficie terrestre.
Le capacità di volo di questi uccelli sono impressionanti: in condizioni di tempo ottimale possono percorrere anche 800 km a una media di 70 km/h.

Esempio 12.2.6 Piccioni viaggiatori

Si consideri un campione di 5 piccioni viaggiatori di un allevamento, numerati da 1 a 5 e sia X la variabile statistica che misura la distanza dall'allevamento, in chilometri, percorsa dagli animali in una certa giornata. Si costruisce la seguente tabella, che descrive quanto lontano è arrivato ciascun esemplare:

Piccione	1	2	3	4	5
Distanza	77	45	1025	101	98

La distanza media percorsa vale $m_X = 269.2$ km; come si vede, però, il valore della media è piuttosto alto perché fortemente influenzato dalla distanza di 1025 km percorsa dal terzo esemplare. La media non è quindi una buona stima riassuntiva dei dati e, in questo caso, conviene usare un'altra valutazione della tendenza centrale, detta **mediana**.
A questo scopo, ordiniamo i dati in ordine crescente:

$$45 \quad 77 \quad 98 \quad 101 \quad 1025\,.$$

La mediana è, per definizione, il dato centrale: in questo caso è il terzo valore, 98. Questo indice prende il nome di mediana proprio perché è il dato che si trova "in mezzo" agli altri lo precede lo stesso numero di dati che lo segue.
Il valore ottenuto, 98, è sicuramente il più rappresentativo della maggioranza dei dati se confrontato con il valore medio.

Siano $X_1, X_2, \ldots, X_N$ i dati di un campione **ordinati in modo crescente**.
La **mediana** è il valore centrale, che si ottiene con la seguente procedura:

- se N è dispari, è il valore del dato che corrisponde all'intero successivo a $N/2$;
- se N è pari, è la media aritmetica dei valori dei dati al posto $N/2$ e al posto successivo.

In entrambi i casi, la mediana separa in due parti uguali l'insieme dei dati.

Per esempio, se si hanno $N = 9$ dati, dopo averli ordinati in ordine crescente, si calcola $N/2 = 4.5$, e l'intero successivo è 5, dunque la mediana è il valore X_5. Se invece si hanno $N = 12$ dati ordinati, si ha che $N/2 =$ e la mediana è il numero $(X_6 + X_7)/2$.

La mediana, al contrario della media, risente poco della presenza di dati "estremi", o anche, eventualmente, errati.

Per esercizio, si consideri la seguente collezione di 8 dati (già ordinati in modo crescente): 2.1, 2.2, 2.2, 3, 4.3, 4.3, 4.3, 5.6 e si verifichi che la media vale 3.5 e la mediana 3.65. Si considerino gli stessi dati dove per un errore di trascrizione, il valore 5.6 è diventato 56. Mostrare che in questo caso la mediana è ancora 3.65 e la media diventa, invece, 9.

Esempio 12.2.7 Confrontare dati – III

Calcoliamo le mediane delle tre serie di dati dell'esempio 12.2.2, iniziando da quelli relativi al 2005. I dati ordinati sono 23, dunque il dato centrale è quello che si trova al dodicesimo posto (infatti $N/2 = 23/2 = 11.5$ e l'intero successivo è 12); prima di X_{12} ci sono 11 dati, e anche dopo ce ne sono 11:

Dati	2	3	4	4	5	5	5	5	5	5	5	6	6	6	6	6	7	7	7	7	7	8	8
Rango	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23

La mediana è dunque 6 (fig. 12.14). Si può ottenere questo stesso valore utilizzando le frequenze relative dei dati, evitando dunque il loro lungo elenco. A questo scopo si riportano in ordine i dati diversi, calcolando la **frequenza cumulata**, cioè la frequenza dei dati che hanno quel valore o un valore inferiore (la frequenza cumulata è un concetto analogo alla funzione di distribuzione in probabilità, *vedi* par. 11.4). La tabella che segue riassume queste informazioni.

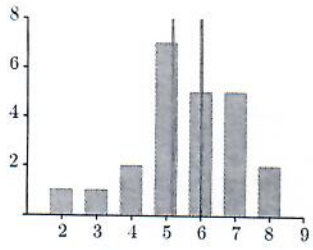


Figura 12.14
Istogramma dei voti del 2005; la moda è 5, la media (in rosso) è 5.6, la mediana (in verde) è 6. La mediana è a destra della media perché il numero di dati a destra della media è superiore al 50%.

Voto	Frequenza relativa	Frequenza cumulata
2	1/23	$1/23 \approx 0.04$
3	1/23	$2/23 \approx 0.09$
4	2/23	$4/23 \approx 0.17$
5	7/23	$11/23 \approx 0.48$
6	5/23	$16/23 \approx 0.70$
7	5/23	$21/23 \approx 0.91$
8	2/23	$23/23 \approx 1.00$

Per esempio, in tabella il valore 3 sulla seconda riga ha frequenza relativa 1/23 perché 3 compare solo una volta su 23, la frequenza cumulata è invece 2/23, infatti la somma delle frequenze di 3 e dei valori precedenti (in questo caso solo uno, il 2) è $1/23+1/23 = 2/23$. Analogamente, la somma delle frequenze relative dei dati inferiori o uguali a 5 è $(1+1+2+7)/23 = 11/23 \approx 0.48$, mentre quella dei dati inferiori o uguali a 6 è $16/23 \approx 0.70$.

La mediana è il dato che separa la metà inferiore dei dati dalla metà superiore. Guardando i valori della frequenza cumulata (fig. 12.15) si comprende che la mediana deve essere, in questo caso, superiore a 5, infatti 5 e i dati inferiori sono meno della metà. Inoltre la mediana non può superare 6, infatti la frequenza cumulata dei dati inferiori o uguali a 6 è maggiore della metà. Se ne deduce che la mediana deve essere 6. Procedendo nello stesso modo calcoliamo la mediana dei dati del 2006.

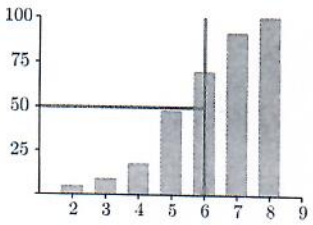


Figura 12.15
Istogramma delle frequenze cumulate. In ordinata sono espresse le percentuali. La retta verticale in verde è $x = 6$. Prima del valore 6 la frequenza cumulata è inferiore al 50%, con il valore 6 supera il 50%, dunque $x = 6$ è la mediana.

Voto	Frequenza relativa	Frequenza cumulata
3	0.08	0.08
4	0.08	0.16
5	0.16	0.32
6	0.28	0.60
7	0.28	0.88
8	0.12	1.00

Anche in questo caso si ottiene 6, infatti la mediana deve superare 5 (solo il 32% dei dati ha valore inferiore o uguale a 5) e non può superare 6 (il 60% dei dati ha valore inferiore o uguale a 6).

Per esercizio, si calcolino le frequenze cumulate per i dati del 2007 e si mostri che, anche in questo caso, la mediana vale 6. ■

Finora abbiamo mostrato come un solo numero (la media o la mediana) possa essere utilizzato per riassumere la tendenza centrale dei dati. In molti casi, però, questa informazione può non essere sufficiente; in particolare può essere necessaria anche una misura dello scostamento dei dati

978-88-08-18484-9

dal valore centrale, cioè un indice di **dispersione** dei dati. Il seguente esempio illustra un caso tipico.

Esempio 12.2.8 Varianza campionaria

Figura 12.16
Una direttiva dell'Unione Europea in vigore dal 2008 obbliga gli Stati membri a prendere sul serio la raccolta differenziata delle pile esauste, che contengono molte sostanze fortemente nocive per gli ecosistemi.

Due ditte di pile pubblicizzano il loro modello "ministilo" a 1.5 volt e affermano, entrambe, che le loro pile hanno una carica complessiva di 1.25 A h (cioè hanno la carica equivalente a un flusso di corrente di 1.25 ampere della durata di un'ora).

Per verificare queste affermazioni, vengono scelti due campioni casuali di 50 pile prodotte da ciascuna delle due ditte, e viene compilata la seguente tabella dei valori delle frequenze assolute con cui sono stati osservati i diversi valori della carica.

Carica in A h	1.15	1.20	1.25	1.30
Prima ditta	1	16	25	8
Seconda ditta	17	0	9	24

Calcoliamo la carica media nei due casi. Si ha

$$\text{prima ditta} \qquad m_1 = 1.15 \cdot \tfrac{1}{50} + 1.20 \cdot \tfrac{16}{50} + 1.25 \cdot \tfrac{25}{50} + 1.30 \cdot \tfrac{8}{50} = 1.24\,,$$

$$\text{seconda ditta} \qquad m_2 = 1.15 \cdot \tfrac{17}{50} + 1.20 \cdot \tfrac{0}{50} + 1.25 \cdot \tfrac{9}{50} + 1.30 \cdot \tfrac{24}{50} = 1.24\,.$$

La carica media, che è di poco più bassa di quella pubblicizzata, è uguale per i due campioni, dunque il solo valore della media non permette di decidere quale dei due prodotti sia migliore. Un ulteriore elemento per giudicare la qualità dei prodotti è la loro affidabilità; ci si può chiedere, in particolare, quanto si discosti la carica di una pila dal valore della carica media.

Nel caso di una variabile aleatoria, la varianza $\sigma^2 = \left\langle (X - \langle X \rangle)^2 \right\rangle$ (*vedi* eq. (11.8)) misura proprio la media del quadrato dello scarto dalla media. Nel caso di dati di un campione, definiamo una quantità analoga, la **varianza campionaria** s^2, che si ottiene calcolando gli scarti dalla media campionaria, moltiplicandoli per la loro eventuale frequenza assoluta e dividendo per $N - 1$, dove N è la dimensione del campione (spiegheremo successivamente il motivo di questa scelta).

Applicando questa definizione, si ottiene, rispettivamente

$$s_1^2 = \frac{(1.15 - 1.24)^2 + 16(1.20 - 1.24)^2 + 25(1.25 - 1.24)^2 + 8(1.30 - 1.24)^2}{49}$$
$$= \frac{0.065}{49} \approx 0.0013,$$

$$s_2^2 = \frac{17(1.15 - 1.24)^2 + 0(1.20 - 1.24)^2 + 9(1.25 - 1.24)^2 + 24(1.30 - 1.24)^2}{49}$$
$$= \frac{0.225}{49} \approx 0.0046\,.$$

Estraendo la radice quadrata della varianza campionaria si ottiene il valore della **deviazione standard campionaria**:

$$s_1 = \sqrt{s_1^2} \approx 0.036, \quad s_2 = \sqrt{s_2^2} \approx 0.068\,.$$

Confrontando questi due valori, si realizza che il secondo è quasi il doppio del primo. Concludiamo che, sebbene la carica media delle pile delle due ditte sia la stessa, questo valore è più attendibile nel caso della prima ditta che nel caso della seconda, infatti i dati di carica della prima ditta sono, in media, più vicini di quelli della seconda al valore medio. ■

Riassumiamo e definiamo in maniera più generale gli indici statistici introdotti nell'esempio precedente.

Sia X una variabile statistica. La **varianza campionaria** di un campione di N dati $X_1, X_2, \ldots, X_N$ di media campionaria m_X è il numero

$$s_X^2 = \frac{(X_1 - m_X)^2 + (X_2 - m_X)^2 + \ldots, +(X_N - m_X)^2}{N-1}, \tag{12.1}$$

che valuta la distanza media al quadrato dei dati dalla media, cioè la loro "dispersione".
Se i dati assumono n valori distinti $x_1, x_2, \ldots, x_n$ con frequenze $F_1, F_2, \ldots, F_n$ si ha, equivalentemente,

$$s_X^2 = \frac{\sum_{i=1}^{n} F_i (x_i - m_X)^2}{N-1}. \tag{12.2}$$

La radice quadrata della varianza $s_X = \sqrt{s_X^2}$ è la **deviazione standard campionaria**.

Esempio 12.2.9 Caramelle

All'uscita di una macchina che produce caramelle, si pesano $N = 1000$ pezzi prodotti e si raggruppano i risultati in classi che differiscono di 0.02 g l'una dall'altra.

Classe	Centro della classe	Frequenza
[1.04, 1.06]	1.05	17
(1.06, 1.08]	1.07	15
(1.08, 1.10]	1.09	27
(1.10, 1.12]	1.11	47
(1.12, 1.14]	1.13	63
(1.14, 1.16]	1.15	85
(1.16, 1.18]	1.17	119
(1.18, 1.20]	1.19	129
(1.20, 1.22]	1.21	126
(1.22, 1.24]	1.23	112
(1.24, 1.26]	1.25	88
(1.26, 1.28]	1.27	69
(1.28, 1.30]	1.29	42
(1.30, 1.32]	1.31	31
(1.32, 1.34]	1.33	16
(1.34, 1.36]	1.35	14

Figura 12.17
La valutazione dei prodotti industriali è un altro campo in cui la statistica ha un ruolo significativo.

Il peso medio di una caramella è 1.2 g; calcoliamo la varianza per valutare le differenze di peso fra i pezzi prodotti.
La varianza campionaria è $s^2 = 4/999 \approx 0.004$, mentre la deviazione standard è $s \approx 0.06$. In media, lo scarto dalla media di ogni pezzo prodotto è dell'ordine di 0.06 g. Per calcolare la varianza abbiamo diviso la somma dei quadrati degli scarti per $1000 - 1 = 999$. Si noti che, se avessimo diviso per 1000, avremmo ottenuto risultati praticamente identici.

978-88-08-18484

Centro	Scarto	Frequenza	Somma del quadrato degli scarti
1.05	−0.15	17	$(-0.15)^2 \cdot 17 = 0.3825$
1.07	−0.13	15	$(-0.13)^2 \cdot 15 = 0.2535$
1.09	−0.11	27	$(-0.11)^2 \cdot 27 = 0.3267$
1.11	−0.09	47	$(-0.09)^2 \cdot 47 = 0.3807$
1.13	−0.07	63	$(-0.07)^2 \cdot 63 = 0.3087$
1.15	−0.05	85	$(-0.05)^2 \cdot 85 = 0.2125$
1.17	−0.03	119	$(-0.03)^2 \cdot 119 = 0.1071$
1.19	−0.01	129	$(-0.01)^2 \cdot 129 = 0.0129$
1.21	0.01	126	$(0.01)^2 \cdot 126 = 0.0126$
1.23	0.03	112	$(0.03)^2 \cdot 112 = 0.1008$
1.25	0.05	88	$(0.05)^2 \cdot 88 = 0.2200$
1.27	0.07	69	$(0.07)^2 \cdot 69 = 0.3381$
1.29	0.09	42	$(0.09)^2 \cdot 42 = 0.3402$
1.31	0.11	31	$(0.11)^2 \cdot 31 = 0.3751$
1.33	0.13	16	$(0.13)^2 \cdot 16 = 0.2704$
1.35	0.15	14	$(0.15)^2 \cdot 14 = 0.3150$
Totale		1000	3.9568

Se invece di un campione si considera un'intera popolazione, è necessari utilizzare simboli differenti per denotare gli indici di tendenza central e di dispersione. Inoltre, come misura di dispersione si usa una divers definizione di varianza, analoga a quella probabilistica.

Se $y_1, y_2, \ldots, y_M$ sono i dati relativi a tutti gli elementi di una popolazione, la media

$$\mu = \frac{1}{M}\sum_{k=1}^{M} y_i$$

è la **media di popolazione**.
Il numero

$$\sigma^2 = \frac{1}{M}\sum_{i=1}^{M}(y_i - \mu)^2 \tag{12.3}$$

è la **varianza di popolazione**, mentre $\sigma = \sqrt{\sigma^2}$ è la **deviazione standard di popolazione**.
La varianza si può anche calcolare con la formula

$$\sigma^2 = \left(\frac{1}{M}\sum_{i=1}^{M} y_i^2\right) - \left(\frac{1}{M}\sum_{k=1}^{M} y_i\right)^2 = \frac{1}{M}\sum_{i=1}^{M} y_i^2 - \mu^2, \tag{12.4}$$

in analogia con la formula (11.8).

La differenza tra la definizione di varianza di popolazione e quella di va rianza campionaria sta nel fatto che, per definire s_X^2, si divide per il nu mero di elementi del campione meno uno, mentre per definire σ^2 si divid per il numero degli elementi. Il motivo di questa differenza è piuttost complesso e lo illustreremo, brevemente, nel prossimo paragrafo.

Esempio 12.2.10 Il teorema fondamentale della selezione naturale

La fitness media di popolazione è il parametro che descrive la capacità delle popolazioni di sopravvivere e riprodursi nel loro ambiente: una popolazione è tanto meglio "adattata" al suo ambiente quanto maggiore è la sua fitness. Il tasso di natalità, che dipende dal genotipo, può dunque essere considerato come espressione della fitness (*vedi* par. 11.5).

Per studiare come varia nel tempo la fitness media, consideriamo una popolazione batterica (che si riproduce in modo asessuato), in cui si distinguono per alcune caratteristiche genetiche n differenti genotipi. Indichiamo con $N_1(t)$, $N_2(t)$, ..., $N_n(t)$ le numerosità dei batteri dei differenti genotipi al tempo t, e, in accordo con l'esempio 8.5.1, indichiamo con $a_1, a_2, \ldots, a_n$ i tassi di variazione specifica delle numerosità dei genotipi (fitness dei genotipi). Le equazioni differenziali che governano queste variabili sono

$$N_1'(t) = a_1 N_1(t), \quad N_2'(t) = a_2 N_2(t), \quad \ldots, \quad N_n'(t) = a_n N_n(t)\,.$$

La numerosità totale $N(t) = \sum_{i=1}^{n} N_i(t)$ verifica l'equazione

$$N'(t) = \sum_{i=1}^{n} a_i N_i(t) = \frac{\sum_{i=1}^{n} a_i N_i(t)}{N(t)} N(t) = a(t)N(t)\,,$$

dove, per definizione,

$$a(t) = \frac{\sum_{i=1}^{n} a_i N_i(t)}{N(t)}$$

è la fitness media della popolazione batterica al tempo t.

Studiamo come varia nel tempo $a(t)$, calcolandone la derivata rispetto a t. Si ha

$$a'(t) = \frac{\mathrm{d}}{\mathrm{d}t} \frac{\sum_{i=1}^{n} a_i N_i(t)}{N(t)} = \frac{(\sum_{i=1}^{n} a_i N_i'(t))N(t) - (\sum_{i=1}^{n} a_i N_i(t))N'(t)}{N^2(t)}\,.$$

Poiché $N_i'(t) = a_i N_i(t)$ e $N'(t) = a(t)N(t)$, sostituendo e semplificando possiamo riscrivere l'equazione precedente come

$$a'(t) = \frac{1}{N(t)} \sum_{i=1}^{n} N_i(t) a_i^2 - a^2(t)\,.$$

Confrontando questa espressione con la (12.4), si ottiene

$$a'(t) = \sigma^2,$$

dove σ^2 è la varianza della fitness dei genotipi a_i.

La precedente relazione può essere vista come un caso particolarmente semplice del famoso Teorema fondamentale della Selezione Naturale di R.A. Fisher, che afferma che il tasso di variazione della fitness media è pari alla varianza della fitness (*vedi* box 12.1). In particolare, a meno che le fitness a_i dei genotipi non siano tutte uguali, risulta $\sigma^2 > 0$, dunque $a'(t) > 0$ e la fitness media aumenta: quanto maggiore è la variazione, tanto più rapido è l'aumento della fitness media. ■

Per valutare la dispersione dei dati, si potrebbe anche pensare di utilizzare l'**intervallo di variabilità** della distribuzione (*range* in inglese),

978-88-08-18484-9

che è la differenza fra il valore più grande della distribuzione e quello più piccolo. Però questo parametro è molto influenzato, ancor più del valore della media, dai valori estremi della distribuzione, che non sono certo rappresentativi della gran parte della distribuzione dei dati. Si usa invece un altro indice di dispersione legato al calcolo della mediana, che, come la mediana, risente poco della presenza di dati estremi.

Esempio 12.2.11 Quartili

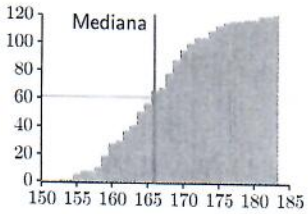


Figura 12.18
Frequenze assolute cumulate per i 120 dati. La mediana, in rosso, separa i dati con indice inferiore a 60 dai dati con indice superiore a 61. La retta orizzontale è esattamente $y = 60.5$.

Consideriamo di nuovo i dati relativi alle altezze del campione di 120 studentesse dell'esempio 12.1.5. Per il calcolo della mediana dovremmo riportare in tabella i dati ordinati. Preferiamo invece disegnare un grafico (fig. 12.18), in cui in ascissa mettiamo il valore del dato e in ordinata il valore della frequenza cumulata, come abbiamo descritto nell'esempio 12.2.7. È facile determinare la mediana su questo grafico: sull'asse verticale troviamo il valore centrale, che corrisponde alla posizione 60.5 (a metà tra la posizione 60 e la posizione 61); sull'asse orizzontale il valore corrispondente è 166.
Questa stessa analisi può essere utilizzata per considerare altre quantità oltre la mediana. In figura 12.19 rappresentiamo lo stesso grafico, ma sull'asse verticale sono segnate le frequenze cumulate in percentuale; alla mediana corrisponde l'ordinata 50%. Il valore 162, che corrisponde all'ordinata 25%, prende il nome di **primo quartile**; il valore 170, che corrisponde all'ordinata 75% prende il nome di **terzo quartile**.
In pratica abbiamo diviso i dati ordinati in 4 gruppi di dimensione pari a un quarto del totale. I valori che separano un gruppo dall'altro sono appunto i **quartili** (la mediana è esattamente il secondo quartile). Il primo quartile lascia alla sua sinistra il 25% dei dati più bassi, il terzo quartile lascia alla sua destra il 25% dei dati più alti. La distanza tra il primo e i terzo quartile è detta **distanza interquartile** ed è una misura di dispersione dei valori intorno alla mediana. Infatti tra il primo e il terzo quartile c'è il 50% dei dati (un quarto sono tra il primo quartile e la mediana, l'altro quarto tra la mediana e il terzo quartile). Più questa distanza è piccola, più i dati sono concentrati intorno alla mediana.
Calcolare, per esercizio, i quartili dei dati relativi ai due campioni di pile descritti nell'esempio 12.2.8. ■

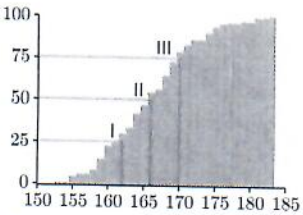


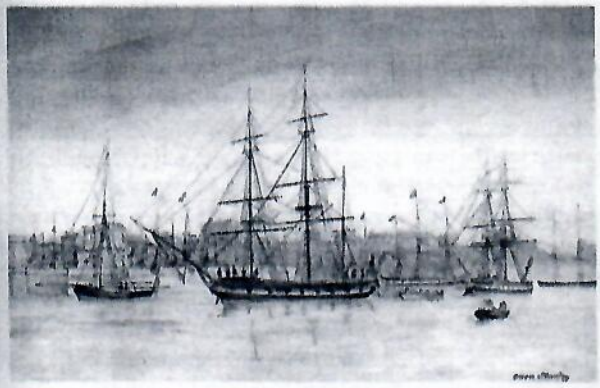
Figura 12.19
Frequenze relative cumulate. La mediana (II) è 166, che corrisponde al 50% dei dati, il primo quartile (I) è 162, e corrisponde alla frequenza cumulata del 25%, il terzo quartile (III) è 170, che corrisponde al 75% dei dati.

In maniera analoga ai quartili, si definiscono i decili e i percentili (più in generale queste quantità prendono il nome di **quantili**). In particolare, il terzo decile è il valore che lascia alla sua sinistra i 3/10 più bassi dei dati ordinati, il 95–esimo percentile è il valore che ha alla sua destra solo il 5% più alto dei dati ordinati.
Diamo la definizione generale di quantile.

Assegnato un campione ordinato di dati e un numero $q \in (0, 1)$, il **quantile** relativo a q è il dato che ha frequenza cumulata pari a q, cioè è il valore x_q per cui la frequenza relativa di dati con valore inferiore è proprio q.
Se $q = 0.5$, il quantile è la mediana, se $q = 0.25$, è il **primo quartile**, se $q = 0.75$, è il **terzo quartile**.
La differenza tra il terzo e il primo quartile è la **distanza interquartile** e dà una misura della dispersione dei dati.

BOX 12.1 Variazione ed evoluzione

Nel 1859, circa 20 anni dopo il suo ritorno dal lungo viaggio di formazione scientifica, Charles Darwin (1809–1882) pubblica l'*Origine delle specie*, l'opera che contiene le idee fondamentali della teoria dell'evoluzione.



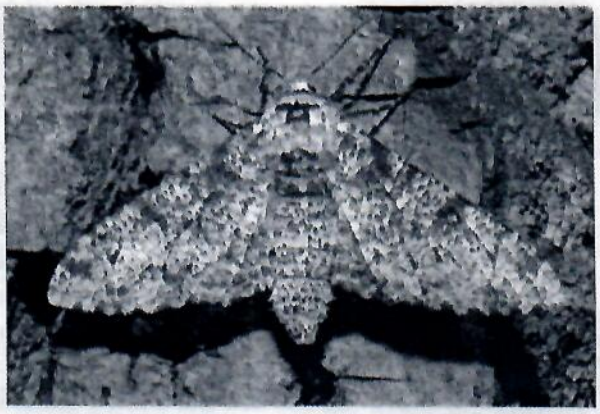
Il brigantino Beagle (al centro) sul quale viaggia Darwin dal 1831 al 1836, in un acquerello del 1841 di Owen Stanley.

Il pensiero di Darwin può essere molto sommariamente riassunto nelle affermazioni seguenti:

- le specie viventi non sono immutabili, ma cambiano gradualmente, evolvendo; in questo processo di cambiamento alcune specie possono anche estinguersi o dare origine a specie differenti;
- il "motore" di questo cambiamento è la **selezione naturale**, detta così per contrasto con quella artificiale operata dall'uomo per migliorare gli esemplari di specie animali o vegetali. Questa "forza", che è il risultato della competizione per la vita, favorisce gli organismi casualmente portatori di caratteristiche più adatte alla sopravvivenza in un dato ambiente e ne sfavorisce altri.

Darwin comprende che, perché la selezione operi, occorre che vi sia **variabilità**, cioè che sussistano tra gli individui differenze ereditabili. Nonostante l'enorme quantità di dati raccolti, le cause e i meccanismi alla base della variabilità rimangono ignote sia a Darwin sia ai suoi contemporanei. Solo all'inizio del Novecento, alla luce dei risultati di Mendel e con il consolidarsi della genetica, si riconosce che la variabilità fenotipica, che è quella su cui agisce l'evoluzione, è il risultato delle differenze alleliche tra individui. Se le leggi di Mendel mostrano, infatti, come il meccanismo della riproduzione sessuale assicuri la conservazione della variabilità, l'esistenza di stati allelici diversi per ogni gene è spiegata dalle mutazioni.

Migrazioni e deriva genetica sono cause di variazione delle frequenze alliche delle popolazioni; ma la massima capacità di generare cambiamento, come previsto da Darwin, la possiede la selezione naturale. Gli individui provvisti di un opportuno genotipo sono più avvantaggiati rispetto ad altri, perché risultano più idonei a riprodursi e sopravvivere nell'ambiente in cui si sviluppano. Molto famoso a questo riguardo è l'esempio, recente, della farfalla *Biston betularia*, dalle ali grigie maculate di bruno. Con l'avvento della rivoluzione industriale di fine Ottocento, che riempì l'aria di scorie di carbone in molte zone dell'Inghilterra, questo organismo fu quasi completamente sostituito da un altro appartenente a una varietà con ali molto scure, cui venne dato il nome di *carbonaria*, dotata di caratteristiche migliori per la sopravvivenza nelle nuove condizioni ambientali.



Le due varietà di farfalla *Biston betularia*.

Molti tentativi sono stati fatti per definire, a partire dalle osservazioni sulla presenza degli organismi negli habitat, un parametro adatto a quantificare l'azione della selezione. Uno dei più usati è la cosiddetta *fitness darwiniana*, che misura proprio l'efficienza riproduttiva dei genotipi. Con l'aiuto di questo parametro, molte ricerche vengono effettuate sia in laboratorio che sul campo per verificare concretamente l'azione della selezione nell'evoluzione.

 978-88-08-18484-9

Esempio 12.2.12 Curva di crescita dei neonati – III

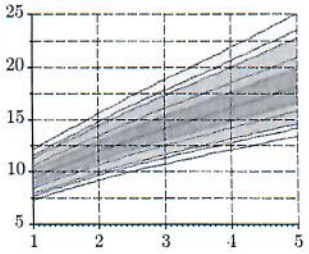


Figura 12.20
Grafico dei percentili del peso in chilogrammi dei bambini tra 1 e 5 anni (dati OMS). Procedendo dall'alto verso il basso, le curve sono i percentili 99, 97, 95, 85, 75 (terzo quartile, in viola), 50 (mediana, in rosso), 25 (primo quartile, in viola), 15, 5, 3, 1.

Negli esempi 5.1.9 e 5.1.16 abbiamo discusso delle caratteristiche di alcuni dati dell'Organizzazione Mondiale della Sanità (OMS) relativi alle curve di crescita dei neonati e dei bambini. I dati dell'OMS sono dati statistici, ottenuti dallo studio di opportuni campioni, e contengono sia i dati centrali (media e mediane) sia gli indici di dispersione.
In particolare, in figura 12.20 riportiamo un grafico dei dati relativi al peso in chilogrammi dei bambini maschi tra 1 e 5 anni. In ascissa è indicata l'età, in ordinata i valori dei percentili. L'OMS fornisce i percentili:

$$1,\ 3,\ 5,\ 15,\ 25,\ 50,\ 75,\ 85,\ 95,\ 97,\ 99\,.$$

I dati espressi con i percentili sono, in effetti, molto più facilmente comprensibili rispetto a quelli espressi tramite media e varianza. Per esempio, supponiamo che un bimbo di 3 anni pesi 12.5 chilogrammi. Nel grafico in figura il punto $(3, 12.5)$ cade sulla curva del 15–esimo percentile. Possiamo concludere che il bimbo è un po' magro, perché solo il 15% dei bambini di 3 anni ha un peso inferiore ai 12.5 chilogrammi. D'altra parte, se un bimbo di 3 anni pesa più di 19 chilogrammi vuol dire che è oltre il 99–esimo percentile, cioè meno dell'uno per cento dei bambini di quell'età pesa così tanto. ■

12.3 Dal campione alla popolazione

L'operazione di raccolta dei dati di una parte di una popolazione si chiama **campionamento**. Se vogliamo usare un campione per ricavare informazioni, anche approssimate, sull'intera popolazione, il primo problema che dobbiamo affrontare è quello di effettuare un campionamento che sia il più possibile rappresentativo della popolazione stessa.
Se formassimo un campione da una popolazione con un criterio preciso, i risultati statistici sarebbero certamente influenzati dal criterio adottato e potrebbe essere sbagliato estendere i risultati a tutta la popolazione. Per esempio, se considerassimo l'altezza media di un campione di 500 individui adulti di uno stesso piccolo comune italiano, difficilmente questo valore sarebbe espressivo dell'altezza media di tutti gli italiani.
Il campionamento, invece, dovrebbe essere sempre casuale, cioè ogni campione dovrebbe avere la stessa possibilità di essere scelto che hanno tutti gli altri possibili campioni della popolazione. Soddisfare questo criterio di scelta equivale a effettuare una "estrazione probabilistica" del campione, che, teoricamente, si può fare proprio con un sorteggio. A ciascuno dei componenti della popolazione può venire idealmente associato un numero progressivo, un'etichetta; tutti questi numeri possono essere messi in un sacchetto, mescolati con cura e poi estratti. A seconda della modalità di estrazione, distinguiamo due diversi procedimenti per formare un campione casuale o, come si dice più precisamente, due diversi **piani di campionamento**.

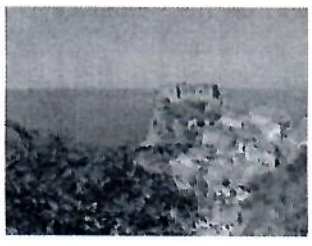


Figura 12.21
Scilla, sullo stretto di Messina. Scegliere individui provenienti da uno stesso luogo per un'indagine statistica dà risultati che non possono essere considerati validi per l'intera popolazione.

In un campionamento **con ripetizione** ogni etichetta estratta viene reinserita tra le altre, dunque un elemento può essere estratto più volte.
In un campionamento **senza ripetizione**, le etichette estratte non vengono rimescolate con le altre, dunque ogni elemento può essere estratto una sola volta.

In entrambi i modi, tutti i campioni possibili della stessa dimensione hanno pari probabilità di essere estratti e ogni singolo elemento della popolazione ha la stessa probabilità di far parte del campione.
Consideriamo un esempio astratto con una popolazione molto piccola, che ci permette di elencare con facilità i campionamenti possibili.

Esempio 12.3.1 Campionamenti – I

Consideriamo come popolazione i valori di una certa grandezza misurati su 3 individui. I valori siano 2, 4 e 12. Elenchiamo tutti e sei i possibili campioni formati da $N = 2$ elementi, estratti senza ripetizione:

Senza ripetizione	$X_1 = 2$ $X_2 = 4$,	$X_1 = 2$ $X_2 = 12$,	$X_1 = 4$ $X_2 = 12$,
	$X_1 = 4$ $X_2 = 2$,	$X_1 = 12$ $X_2 = 2$,	$X_1 = 12$ $X_2 = 4$.

Ogni campione ha la stessa possibilità (uguale a 1/6) di essere estratto. Poiché il calcolo degli indici statistici non dipende dall'ordine dei dati, possiamo considerare come campioni distinti solo quelli della prima riga (infatti quelli della seconda sono uguali a meno dell'ordine). Ciascuno di questi ha probabilità 1/3 di essere estratto.
I campioni di $N = 2$ elementi con ripetizione sono invece

Con ripetizione	$X_1 = 2$ $X_2 = 4$,	$X_1 = 2$ $X_2 = 12$,	$X_1 = 4$ $X_2 = 12$,
	$X_1 = 4$ $X_2 = 2$,	$X_1 = 12$ $X_2 = 2$,	$X_1 = 12$ $X_2 = 4$,
	$X_1 = 2$ $X_2 = 2$,	$X_1 = 4$ $X_2 = 4$,	$X_1 = 12$ $X_2 = 12$,

ognuno dei quali ha probabilità 1/9 di essere estratto, in accordo con il fatto che ogni singolo elemento della popolazione può essere estratto con probabilità 1/3. Poiché l'ordine degli elementi non è importante, possiamo limitarci a considerare solo i 3 campioni dell'ultima riga, che hanno probabilità di estrazione 1/9, e i 3 della prima riga, con probabilità di estrazione 2/9 (pesano il doppio perché sono uguali a quelli della seconda riga, a meno dell'ordine). ■

6	7	8	2	6	3
4	10	3	4	4	3
6	5	3	7	5	10
3	2	8	3	2	7
5	9	2	2	6	6

Figura 12.22
30 numeri interi scelti a caso tra 1 e 10. Prima della diffusione dei computer, l'estrazione di un campione da una popolazione avveniva con l'ausilio di tavole di numeri casuali, compilate in vari modi.
Oggi si utilizzano programmi per computer che generano numeri casuali (*random* in inglese) in modo che ogni numero abbia la stessa probabilità di essere estratto. I numeri così ottenuti sono in realtà solo *pseudorandom*, perché presentano delle regolarità nell'estrazione, che sono però osservabili solo estraendone moltissimi.
La ricerca di efficienti algoritmi di generazione di numeri random ha molte applicazioni concrete, in particolare in crittografia.

Se si escludono le vere e proprie indagini statistiche, i campioni sono spesso semplici raccolte di dati sperimentali, che possiamo assimilare a campioni casuali senza ripetizione. Se infatti la popolazione è molto grande rispetto alla dimensione del campione, in un campione con ripetizione è molto improbabile che ci siano elementi uguali.
Un campione casuale dà certamente un'immagine imparziale, anche se ridotta, dei componenti dell'intera popolazione; i dati relativi al campione hanno dunque qualche speranza di fornire indicazioni anche per l'intera popolazione.
Per comprendere quali indicazioni si possano ottenere, è necessario interpretare in termini probabilistici le quantità statistiche che misuriamo.

In un'estrazione casuale di un campione di dimensione N, la media campionaria e la varianza campionaria sono variabili aleatorie, cioè variano al variare del campione estratto. La probabilità di ciascun valore dipende dalla distribuzione dei dati nella popolazione.

 978-88-08-18484-9

In tutto questo paragrafo, indicheremo con μ la media della popolazione e con σ^2 la varianza. Inoltre, per dare risalto al fatto che consideriamo campioni di dimensione N, indicheremo con m_N la media di un campione e con s_N^2 la varianza campionaria.
Consideriamo un esempio che ci permette di comprendere le proprietà probabilistiche degli indici statistici misurati su un campione.

Esempio 12.3.2 Campionamenti – II

Calcoliamo la media campionaria relativa ai tre campioni senza ripetizione elencati nell'esempio 12.3.1.

Campione	2, 4	2, 12	4, 12
Media campionaria	3	7	8

Osserviamo subito che i tre campioni, di uguale dimensione e estratti dalla stessa popolazione, hanno medie campionarie diverse e ciascun valore ha probabilità 1/3 di essere osservato. Il valore atteso della media campionaria è dunque

$$\langle m_2 \rangle = \frac{1}{3}(3+7+8) = 6\,.$$

Nel caso dei sei distinti campioni con ripetizione, le medie campionarie hanno valori 3, 7 e 8 con probabilità 2/9, e valori 2, 4, 12 con probabilità 1/9. Anche in questo caso il valore atteso è 6, infatti si ha

$$\langle m_2 \rangle = \frac{2}{9}(3+7+8) + \frac{1}{9}(2+4+12) = 6\,.$$

La media di popolazione, che è la somma dei 3 valori 2, 4, 12 divisa per 3, è $\mu = 6$ ed è esattamente uguale al valore atteso $\langle m_2 \rangle$. ■

I risultati di questo esempio sono generali.

Fissata la dimensione N di un campione, il valore atteso della media campionaria è uguale alla media di popolazione

$$\langle m_N \rangle = \mu.$$

Questo risultato vale sia nel caso di campioni con ripetizione, sia nel caso di campioni senza ripetizione.

Nel linguaggio della statistica questa affermazione si esprime dicendo che la media campionaria è uno **stimatore non distorto** della media di popolazione.
Vediamo con un esempio che un'affermazione simile vale anche per la varianza campionaria.

Esempio 12.3.3 Campionamenti – III

Calcoliamo la varianza campionaria relativa ai tre campioni senza ripetizione elencati nell'esempio 12.3.1.

Campione	2, 4	2, 12	4, 12
Varianza campionaria	2	50	32

Il valore atteso della varianza campionaria è

$$\langle s_2^2 \rangle = \frac{1}{3}(2 + 50 + 32) = 28\,.$$

Nel caso del campionamento con ripetizione, si ha

$$\langle s_2^2 \rangle = \frac{2}{9}(2 + 50 + 32) + \frac{1}{9}(0 + 0 + 0) = \frac{168}{9} = \frac{56}{3} \approx 18.67\,,$$

visto che i campioni con elementi uguali danno contributo nullo alla varianza.
Infine, dato che è $\mu = 6$, la varianza di popolazione è

$$\sigma^2 = \frac{(2-6)^2 + (4-6)^2 + (12-6)^2}{3} = \frac{56}{3} \approx 18.67\,.$$

Si ha quindi $\sigma^2 = \langle s_2^2 \rangle$ nel caso che il campionamento sia con ripetizione, mentre, se il campionamento è senza ripetizione, si ottiene $\langle s_2^2 \rangle = 28$, che non è la varianza. Questo valore risulta uguale alla varianza campionaria se il campione coincide con la popolazione:

$$s_3^2 = \frac{(2-6)^2 + (4-6)^2 + (12-6)^2}{3} = 28 = \frac{3}{2}\sigma^2$$

(la differenza tra s_3^2 e σ^2 è solo nel denominatore, che nel primo caso è $M - 1 = 2$ e nel secondo è $M = 3$). ■

Riassumiamo in forma generale i risultati ottenuti.

In campioni di dimensione N, estratti con ripetizione, il **valore atteso della varianza campionaria** è la varianza della popolazione:

$$\langle s_N^2 \rangle = \sigma^2\,.$$

Se la popolazione è formata da M elementi, il valore atteso della varianza per campioni senza ripetizione è invece

$$\langle s_N^2 \rangle = \frac{M}{M-1}\sigma^2;$$

si noti che questo valore è uguale a s_M^2, cioè alla varianza campionaria che si ottiene usando come campione tutta la popolazione.

Se M è molto grande, $\frac{M}{M-1} \approx 1$ e quindi, nel campionamento senza ripetizione, $\langle s_N^2 \rangle$ è praticamente uguale a σ^2. La differenza tra le strategie di campionamento è dunque irrilevante se la popolazione è molto grande.
Queste proprietà sono vere solo se la varianza campionaria è definita esattamente come nella formula (12.2). Questo spiega perché nella formula si divide per $N - 1$ e non per N, come sembrerebbe naturale. Infatti, questa scelta darebbe uno stimatore distorto, perché il valore atteso si discosta dalla varianza di popolazione di una quantità dell'ordine di $1/N$, e questo errore è piccolo solo per campioni molto grandi; nella pratica invece i campioni possono essere anche molto piccoli.

Mostriamo ora un esempio di campionamento in un caso più realistico, in cui, per la dimensione della popolazione, non possiamo effettuare tutti i campionamenti possibili come negli esempi 12.3.1, 12.3.2 e 12.3.3.

978-88-08-18484-9

Esempio 12.3.4 Stambecchi – I

Figura 12.23
Una femmina di stambecco (*Capra ibex ibex*).

La popolazione degli stambecchi del Parco Nazionale dello Stelvio è composta da circa $M = 800$ esemplari adulti (dati del 2011), la cui lunghezza L varia tra i 130 e i 160 centimetri. Più precisamente la lunghezza è distribuita come nella tabella che segue, dove F rappresenta la frequenza assoluta di ogni dato, cioè il numero degli esemplari che hanno la stessa lunghezza.

L	130	135	140	145	150	155	160
F	32	41	158	268	116	70	115

La lunghezza media degli individui della popolazione è quindi $\mu = 146.7$ e la varianza è $\sigma^2 \approx 61.0$.

Costruiamo ora un campione casuale di dimensione $N = 40$ associando a ogni valore L un numero da 1 a 800 e scegliendo a caso 40 numeri (come da un sacchetto ben mescolato) con ripetizione. I valori relativi a questo campione sono

L	130	135	140	145	150	155	160
F	3	0	8	16	5	0	8

La media del campione è $m_{40} = 146.5$, la varianza campionaria è $s^2_{40} \approx 69.5$. Confrontando μ con m_{40} si vede che la differenza è di soli 0.2 cm, quindi m_{40} sembra un'ottima stima di μ. Ma potrebbe essere un caso fortunato! Ripetiamo allora lo stesso procedimento, estraendo un nuovo campione.

L	130	135	140	145	150	155	160
F	1	2	15	11	3	5	3

La media di questo campione è $m_{40} = 145.0$ e, in questo caso, l'errore che si commette stimando μ con m_{40} è 1.7 cm, un po' più grande del precedente, ma ancora contenuto.

Si calcoli, per esercizio, l'errore che si commette se si considera un campione di dimensione $N = 20$ in cui le frequenze delle misure siano

F	0	4	1	7	1	6	1

e se si considera un altro campione in cui le frequenze delle misure siano

F	0	2	3	2	3	6	4

■

Se, nell'esempio precedente, ripetessimo ancora molte volte la procedura descritta, calcolando ogni volta la media campionaria m_N, troveremmo un insieme di valori la cui probabilità di realizzazione si distribuisce seguendo una certa legge, detta **legge di distribuzione della media**, che dipende da come sono realmente distribuiti tutti i dati della popolazione. Se questa distribuzione non è troppo dispersa, c'è una buona possibilità che la media di ogni campione non sia troppo diversa da quella della popolazione intera. È dunque importante avere qualche informazione sulla legge di distribuzione della media campionaria.

Ci viene in aiuto il teorema del limite centrale, di cui abbiamo discusso alla fine del paragrafo 11.4, che ci permette di trarre un'importantissima conclusione pratica.

Se $X_1, X_2, \ldots, X_N$ sono i dati di un campione di dimensione N estratto, con ripetizione, da una popolazione con media di popolazione μ e varianza σ^2, la media campionaria m_N è distribuita, approssimativamente, come una variabile aleatoria gaussiana di media μ, pari alla media di popolazione, e di varianza σ^2/N.

Questa approssimazione vale qualunque sia la distribuzione dei dati nella popolazione ed è tanto più accurata quanto più grande è N.

Come conseguenza, la variabile $Z = (m_N - \mu)/(\sigma/\sqrt{N})$ è distribuita, approssimativamente, come la normale standardizzata $N_{0,1}$. Usando le relazioni 11.18 possiamo quindi trarre le seguenti conclusioni.

Se m_N è la media campionaria di un campione di N elementi estratto da una popolazione di media μ e varianza σ^2, allora la differenza $m_N - \mu$ verifica:

$$
\begin{aligned}
&|m_N - \mu| \leq \frac{\sigma}{\sqrt{N}} \quad \text{con probabilità } 0.682\,, \\
&|m_N - \mu| \leq \frac{2\sigma}{\sqrt{N}} \quad \text{con probabilità } 0.954\,, \\
&|m_N - \mu| \leq \frac{3\sigma}{\sqrt{N}} \quad \text{con probabilità } 0.997\,.
\end{aligned}
\tag{12.5}
$$

Al crescere di N, la stima di μ con m_M è via via più accurata.

Inoltre, sono di uso frequente in statistica le seguenti stime, analoghe alle (12.5)

$$
\begin{aligned}
&|m_N - \mu| \leq 1.96\frac{\sigma}{\sqrt{N}} \quad \text{con probabilità } 0.95\,, \\
&|m_N - \mu| \leq 2.58\frac{\sigma}{\sqrt{N}} \quad \text{con probabilità } 0.99\,, \\
&|m_N - \mu| \leq 3.29\frac{\sigma}{\sqrt{N}} \quad \text{con probabilità } 0.999\,.
\end{aligned}
\tag{12.6}
$$

Vediamo in un esempio come si usano.

Esempio 12.3.5 Stambecchi – II

La popolazione descritta nell'esempio 12.3.4 ha deviazione standard $\sigma \approx 7.8$.

Il primo campionamento di 40 stambecchi dell'esempio 12.3.4 dà come media campionaria $m_{40} = 146.5$. Visto che, in questo caso, risulta $\sigma/\sqrt{N} \approx 1.23$ e $2\sigma/\sqrt{N} \approx 2.46$, utilizzando le (12.5), possiamo concludere che la distanza di m_{40} da μ è, con probabilità 0.682, inferiore a 1.23 centimetri e, con probabilità 0.954, inferiore a 2.46. In effetti, nel caso del secondo campionamento, abbiamo ottenuto un valore di errore di 1.7 centimetri, maggiore di 1.23, ma comunque inferiore a 2.42.

 978-88-08-18484-9

Se ci riferiamo alle stime (12.6), possiamo inoltre affermare che la probabilità che la media campionaria si discosti dalla media di popolazione più di $1.96\,\sigma/\sqrt{N} \approx 2.42$ è del 5%, che si discosti più di $2.58\,\sigma/\sqrt{N} \approx 3.18$ è dell'1%.

In questi esempi la deviazione standard della popolazione è un dato del problema. In contesti concreti ciò non accade quasi mai e la varianza è ignota tanto quanto lo è la media di popolazione. Se però la dimensione del campione è sufficientemente grande (nella pratica $N > 120$ o $N > 150$), non si commette un grande errore se si utilizzano le (12.6), sostituendo alla varianza di popolazione la varianza campionaria.

Esempio 12.3.6 Mele

Figura 12.24
Esistono più di 2000 varietà di mele.

Prima di ingrandire la sua piantagione di mele, un piccolo coltivatore analizza i risultati ottenuti dalle sue 140 piante: in media ha raccolto 163 kg di mele per pianta, con una varianza campionaria di 1312 kg^2. Vediamo che conclusioni può trarre sulla media di un'eventuale piantagione molto più grande.

Il coltivatore può considerare la piantagione estesa come una popolazione di cui la piantagione attuale è solo un campionamento. Visto che il campione di cui dispone è abbastanza grande, il valore della varianza campionaria può essere considerato una buona approssimazione della varianza. Calcolando il rapporto $s/\sqrt{N} \approx 3$, il coltivatore può essere praticamente certo (al 99%) che otterrà dalla sua piantagione (a parità di condizioni di coltivazione) un numero di chili di mele per pianta compreso tra $163 - 2.58 \cdot 3 \approx 155$ e $163 + 2.58 \cdot 3 \approx 171$ chili per pianta.

Per esercizio, determinare quale sarebbe stata l'ampiezza di questo intervallo se il campione fosse stato di $N = 1000$ piante, a parità di parametri.

Abbiamo osservato che la media campionaria è una variabile aleatoria, perché cambia se si sceglie casualmente il campione. Usando le (12.6) si può certamente affermare, come abbiamo fatto, che m_N è, con probabilità 0.95, contenuta nell'intervallo $(\mu - 1.96\sigma/\sqrt{N}, \mu + 1.96\sigma/\sqrt{N})$. Da un punto di vista pratico questa osservazione spesso non è utilizzabile, perché μ non è nota. È più interessante, invece, usare l'affermazione analoga, cioè che la media di popolazione μ (la vera incognita del problema) varia nell'intervallo

$$\left(m_N - 1.96\frac{\sigma}{\sqrt{N}}, m_N + 1.96\frac{\sigma}{\sqrt{N}}\right).$$

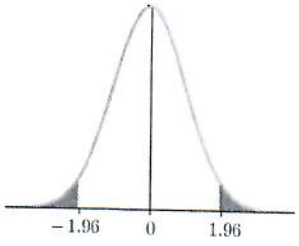


Figura 12.25
Distribuzione normale standard. L'area della regione rossa è 0.05, e rappresenta la probabilità che i dati cadano fuori dall'intervallo $(-1.96, 1.96)$, che è dunque l'intervallo di fiducia al 5%.

Non si può, però, dire che ciò sia vero con probabilità 0.95; infatti, anche se non è nota, la media di popolazione **non è una variabile aleatoria**: solo il campione è scelto probabilisticamente, mentre la popolazione è fissata.

L'affermazione esatta è che, se μ fosse la vera media di popolazione, allora $|m_N - \mu|$ sarebbe inferiore a $1.96\,\sigma/\sqrt{N}$ con probabilità 0.95.

Il cambiamento di punto di vista, dalla probabilità che descrive m_N alla statistica che tenta di indovinare il valore di μ, comporta un cambiamento di linguaggio: nelle affermazioni su m_N, nota μ, si parla correttamente di probabilità, nelle affermazioni su μ ignota, osservata m_N, si parla di **livello di fiducia**.

Gli intervalli in cui si riesce a stimare un parametro della popolazione sono detti **intervalli di confidenza** (o fiduciari).
Nel caso particolare della media, si considerano i seguenti intervalli standard:

$$\mu \in \left(m_N - 1.96\frac{\sigma}{\sqrt{N}}, m_N + 1.96\frac{\sigma}{\sqrt{N}}\right)$$

con un livello di fiducia del 95%,

$$\mu \in \left(m_N - 2.58\frac{\sigma}{\sqrt{N}}, m_N + 2.58\frac{\sigma}{\sqrt{N}}\right)$$

con un livello di fiducia del 99%,

$$\mu \in \left(m_N - 3.29\frac{\sigma}{\sqrt{N}}, m_N + 3.29\frac{\sigma}{\sqrt{N}}\right)$$

con un livello di fiducia del 99.9%.

Il livello di fiducia è dunque pari al valore della probabilità nel caso μ fosse nota e stessimo considerando la scelta di campioni casuali.

Esempio 12.3.7 Intervalli di fiducia sul contenuto di catrame delle sigarette

Il contenuto massimo di catrame delle sigarette è fissato per legge; attualmente è di circa 10 mg per sigaretta. Questa quantità si riferisce al risultato di misure effettuate mediante opportune macchine fumatrici, che simulano l'attività di un fumatore reale. Un produttore che vuole entrare sul mercato verifica che su un campione di 400 sigarette di sua produzione, il contenuto medio di catrame è di 9.62 mg, con una deviazione standard campionaria pari a $s = 0.40$ mg.
Determiniamo gli intervalli di confidenza del contenuto medio di catrame della sua produzione.
Poiché il campione è numeroso, possiamo supporre che la deviazione standard campionaria sia praticamente pari alla deviazione standard della popolazione, dunque $s/\sqrt{N} = 0.02$. L'intervallo fiduciario al 99% per la media è $[9.57, 9.67]$. Ottenuto questo risultato, il produttore può tranquillizzarsi, perché, con un elevato livello di fiducia, il contenuto medio di catrame delle sue sigarette è inferiore al limite di legge.
Determiniamo ora con quale probabilità il contenuto di catrame di una sigaretta scelta a caso supera il limite.
L'intervallo fiduciario che abbiamo determinato non risponde a questa domanda. Abbiamo infatti determinato un intervallo per la media di popolazione, e non la probabilità che una singola sigaretta abbia contenuto di catrame maggiore di 10 mg. Per rispondere, dovremmo invece conoscere la distribuzione del contenuto di catrame in tutta la popolazione, ma di essa non sappiamo quasi nulla, se non, approssimativamente, i valore della media e della varianza.
Se immaginiamo che la fluttuazione del contenuto di catrame sia dovuta alla somma di tante cause diverse e indipendenti, possiamo supporre che la distribuzione di questa grandezza sia gaussiana, di media (circa) 9.62 e di deviazione standard (circa) 0.40. Se così fosse, la probabilità che una sigaretta abbia contenuto di catrame maggiore di 10 sarebbe pari a

$$\int_{10}^{+\infty} N_{9.62.0.40}(x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{0.38/0.40}^{+\infty} \mathrm{e}^{-y^2/2}\,\mathrm{d}y \approx 0.17\,.$$

Per ottenere questo valore abbiamo standardizzato la distribuzione operando il cambio di variabili $(x - m)/\sigma = y$ e successivamente abbiamo ottenuto il valore dell'integrale attraverso un opportuno programma (tutti i programmi statistici al calcolatore permettono questo calcolo). ■

Figura 12.26
Il "catrame" è il residuo particolato della combustione del tabacco e della carta.