centrale della distribuzione se il primo valore è alto e il secondo basso, oppure se il primo è basso e il secondo alto, oppure se entrambi sono valori intermedi. La distribuzione della somma di due Uniformi è dunque più vicina alla Normale di quanto lo sia la distribuzione Uniforme stessa. Tuttavia, il brutale troncamento nei limiti della distribuzione, 0 e 2, non trova corrispondenza nella distribuzione Normale. La Figura 7.6 mostra anche il risultato della somma di quattro distribuzioni Uniformi e di sei distribuzioni Uniformi. La somiglianza con la distribuzione Normale aumenta all'aumentare del numero di Uniformi sommate, e nell'ultimo caso la somiglianza è così marcata che a fatica si riescono ancora a distinguere le due distribuzioni.

L'approssimazione della distribuzione Binomiale con la Normale è un caso particolare del teorema centrale del limite. La distribuzione di Poisson è un altro caso particolare. Se consideriamo un insieme di variabili aleatorie di Poisson con lo stesso tasso e le sommiamo, otteniamo una variabile che corrisponde al numero di eventi aleatori in un intervallo di tempo più lungo (che è la somma degli intervalli di tempo delle singole variabili), e che dunque è una distribuzione di Poisson con media più grande. Dal momento che è anche somma di variabili aleatorie indipendenti e identicamente distribuite essa tenderà alla distribuzione Normale al crescere della media. Dunque, al crescere della media la distribuzione di Poisson diventa approssimativamente Normale. Per la maggior parte dei nostri scopi, nella pratica questo avviene quando la media è più grande di 10. La somiglianza tra la Poisson e la Binomiale, già sottolineata in § 6.7, è quindi solo un aspetto di una proprietà di convergenza di cui godono molte altre distribuzioni.

7.3 Proprietà della distribuzione Normale

Nella sua forma più semplice, chiamata **distribuzione Normale Standard**, l'equazione della densità di probabilità della distribuzione Normale è di solito indicata con $\phi(z)$, dove ϕ è la lettera greca "phi":

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

dove π è l'usuale costante matematica. Il lettore medico può essere rassicurato sul fatto che nella pratica non abbiamo bisogno di usare questa formula proibitiva. La distribuzione Normale Standard ha media 0, deviazione standard 1, e ha una forma come quella mostrata in Figura 7.7. La curva ha forma simmetrica rispetto alla media, spesso descritta definendola "a campana" (sebbene io non abbia mai visto una campana di questa forma). Possiamo notare che la maggior parte dell'area, ovvero della probabilità, è racchiusa tra -1 e $+1$, che la grande maggioranza è tra -2 e $+2$, e che quasi tutta è tra -3 e $+3$.

Sebbene la densità di probabilità della distribuzione Normale abbia molte proprietà notevoli, ne ha una piuttosto inopportuna: non può essere integrata; in altre parole, non c'è una formula semplice per ottenere la probabilità che una variabile aleatoria distribuita come una Normale giaccia tra due limiti dati. L'area

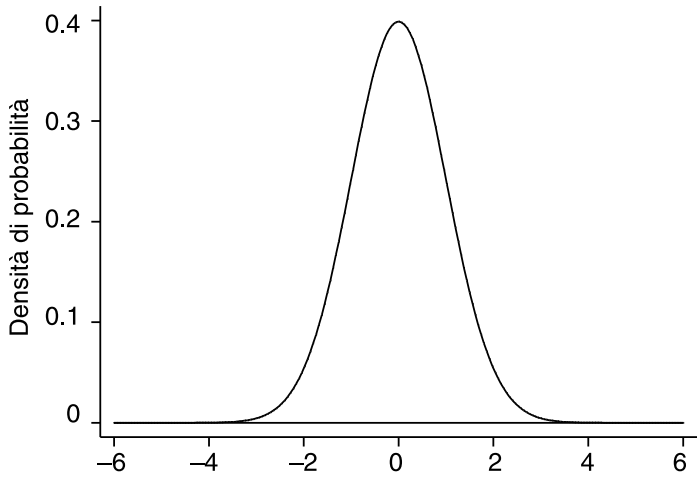

Figura 7.7: La distribuzione Normale Standard

Tabella 7.1: La distribuzione Normale

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
−3.0	0.001	−2.0	0.023	−1.0	0.159	0.0	0.500	1.0	0.841	2.0	0.977
−2.9	0.002	−1.9	0.029	−0.9	0.184	0.1	0.540	1.1	0.864	2.1	0.982
−2.8	0.003	−1.8	0.036	−0.8	0.212	0.2	0.579	1.2	0.885	2.2	0.986
−2.7	0.003	−1.7	0.045	−0.7	0.242	0.3	0.618	1.3	0.903	2.3	0.989
−2.6	0.005	−1.6	0.055	−0.6	0.274	0.4	0.655	1.4	0.919	2.4	0.992
−2.5	0.006	−1.5	0.067	−0.5	0.309	0.5	0.691	1.5	0.933	2.5	0.994
−2.4	0.008	−1.4	0.081	−0.4	0.345	0.6	0.726	1.6	0.945	2.6	0.995
−2.3	0.011	−1.3	0.097	−0.3	0.382	0.7	0.758	1.7	0.955	2.7	0.997
−2.2	0.014	−1.2	0.115	−0.2	0.421	0.8	0.788	1.8	0.964	2.8	0.997
−2.1	0.018	−1.1	0.136	−0.1	0.460	0.9	0.816	1.9	0.971	2.9	0.998
−2.0	0.023	−1.0	0.159	0.0	0.500	1.0	0.841	2.0	0.977	3.0	0.999

Tabella 7.2: Punti percentuali della distribuzione Normale

Unilateri		Bilateri	
$P_1(z)$	z	$P_2(z)$	z
50	0.00		
25	0.67	50	0.67
10	1.28	20	1.28
5	1.64	10	1.64
2.5	1.96	5	1.96
1	2.33	2	2.33
0.5	2.58	1	2.58
0.1	3.09	0.2	3.09
0.05	3.29	0.1	3.29

La Tabella riporta la probabilità $P_1(z)$ che una variabile aleatoria Normale con media 0 e varianza 1 sia maggiore di z, e la probabilità $P_2(z)$ che una variabile aleatoria Normale con media 0 e varianza 1 sia minore di $-z$ oppure maggiore di z.

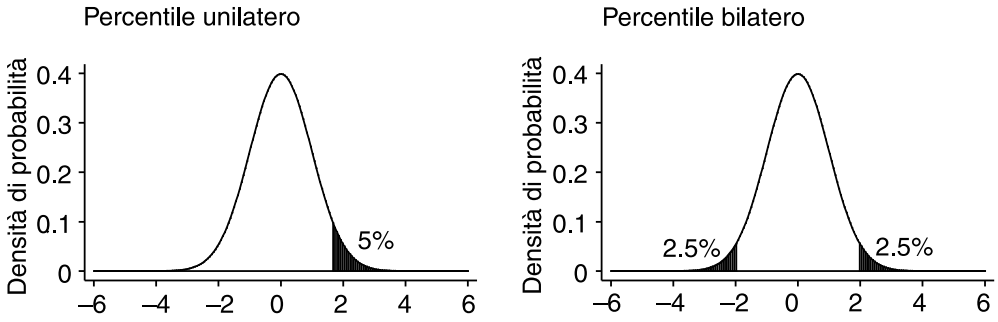

Figura 7.8: Punti percentuali unilateri e bilateri (5%) della distribuzione Normale Standard

sottesa dalla curva può tuttavia essere trovata con algoritmi numerici, ed è stata già calcolata e tabulata. La Tabella 7.1 mostra l'area sottesa dalla densità di probabilità in corrispondenza di differenti valori ammissibili per la distribuzione Normale; per essere più precisi, per un dato valore di z la tabella mostra l'area sottesa dalla curva a sinistra di z, ovvero da $-\infty$ a z (Figura 7.8). Dunque $\Phi(z)$ è la probabilità che un valore scelto a caso dalla distribuzione Normale Standard sia minore di z. Φ è la lettera greca "phi" maiuscola. Si noti che metà di questa tabella non è strettamente necessaria: abbiamo bisogno infatti solo della metà corrispondente ai valori positivi di z, dal momento che $\Phi(-z) + \Phi(z) = 1$. Ciò deriva dalla simmetria della distribuzione. Per trovare la probabilità che z sia compreso tra due valori a e b, con $b > a$, calcoliamo $\Phi(b) - \Phi(a)$; per trovare la probabilità che z sia più grande di a calcoliamo $1 - \Phi(a)$. Queste formule sono tutti esempi della proprietà additiva della probabilità. La Tabella 7.1 riporta solo alcuni dei valori ammissibili per z, ma sono disponibili altre tabelle molto più dettagliate (Lindley e Miller 1955, Pearson e Hartley 1970). Qualsiasi programma di statistica di buon livello è in grado di calcolare questi valori, se necessario.

Esiste un altro modo per tabulare una distribuzione, utilizzando quei punti che sono chiamati punti percentuali. Il **P-esimo punto percentuale unilatero** di una distribuzione è quel valore z tale che c'è una probabilità del $P\%$ che un'osservazione da quella distribuzione sia maggiore o uguale di z (Figura 7.8). Il **P-esimo punto percentuale bilatero** è quel valore z tale che c'è una probabilità del $P\%$ che un'osservazione sia maggiore o uguale di z, oppure minore o uguale di $-z$ (Figura 7.8). La Tabella 7.2 mostra sia i punti pecentuali unilateri che quelli bilateri della distribuzione Normale. La probabilità è riportata sotto forma di percentuale perché quando utilizziamo i punti percentuali stiamo solitamente considerando probabilità piuttosto piccole, come 0.05 o 0.01, e l'utilizzo delle percentuali, in questo caso 5% e 1%, elimina gli zeri superflui.

Finora abbiamo esaminato la distribuzione Normale con media 0 e deviazione standard 1. Se sommiamo una costante μ a una variabile aleatoria Normale Standard, otteniamo una nuova variabile di media μ (si veda § 6.6). La Figura 7.9 mostra la distribuzione Normale di media 0 e la distribuzione ottenuta sommando ad essa il valore 1, in entrambe sono evidenziati i punti percentuali bilateri al 5%.

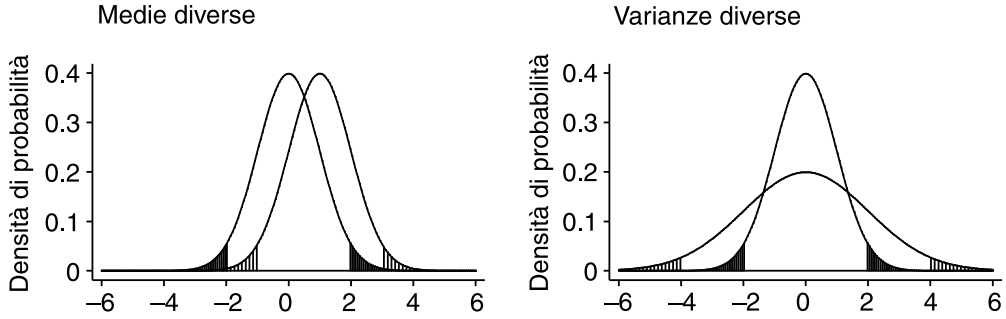

Figura 7.9: Distribuzione Normale per valori diversi della media e della varianza, evidenziando i punti percentuali bilateri al 5%

Le due curve sono identiche, a meno di una traslazione lungo l'asse. La curva di media 0 sottende quasi tutta la probabilità nell'intervallo di estremi -3 e $+3$, mentre la curva di media 1 tra -2 e $+4$, ovvero tra la media -3 e la media $+3$. La probabilità che un valore sia distante un dato numero di unità dalla media è la stessa per entrambe le curve, come è evidente anche dai punti percentuali al 5%.

Se consideriamo una variabile aleatoria Normale Standard, con deviazione standard pari a 1, e la moltiplichiamo per una data costante σ otteniamo una nuova variabile che ha deviazione standard pari a σ. La Figura 7.9 mostra la distribuzione Normale di media 0 e deviazione standard 1, e la distribuzione di una variabile ottenuta moltiplicando per il valore 2 una variabile Normale Standard. Le due curve non sembrano identiche. Per la distribuzione con deviazione standard 2, quasi tutta la probabilità è compresa tra i valori -6 e $+6$, un intervallo molto più ampio rispetto a quello di estremi -3 e $+3$ che gode della stessa proprietà rispetto alla distribuzione Normale Standard. I valori -6 e $+6$ sono rispettivamente pari a -3 e $+3$ deviazioni standard. Possiamo osservare che la probabilità di essere distanti un certo numero di deviazioni standard dalla media è la stessa per entrambe le distribuzioni; questa considerazione può essere fatta anche guardando i punti percentuali al 5%, che in entrambi i casi sono dati dalla media più o meno 1.96 deviazioni standard.

In conclusione, se sommiamo μ a una variabile aleatoria Normale Standard e la moltiplichiamo per σ, otteniamo una variabile aleatoria Normale di media μ e deviazione standard σ. Nell'uso delle Tabelle 7.1 e 7.2 si può applicare direttamente questa proprietà indichando con z la distanza dalla media espressa in numero di deviazioni standard, piuttosto che il valore numerico della variabile; così, per esempio, i punti percentuali bilateri al 5% di una distribuzione Normale di media 10 e deviazione standard 5 si ottengono calcolando $10 - 1.96 \times 5 = 0.2$ e $10 + 1.96 \times 5 = 19.8$, avendo scelto il valore 1.96 in base alla Tabella 7.2.

Questa proprietà della distribuzione Normale, vale a dire il fatto che moltiplicando o sommando costanti si ottiene ancora una distribuzione Normale, non è così scontata come sembra. La distribuzione Binomiale, per esempio, non gode di questa proprietà. Si consideri una variabile aleatoria Binomiale con $n = 3$, i cui valori ammissibili sono dunque 0, 1, 2 e 3, e la si moltiplichi per 2; i valori

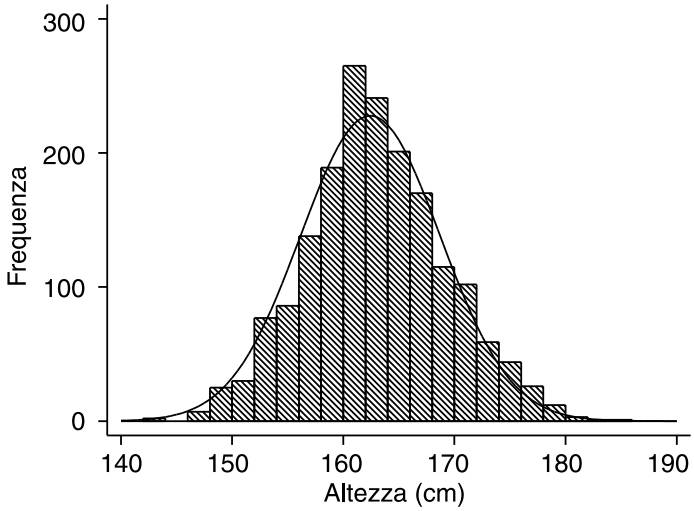

Figura 7.10: Distribuzione delle altezze in un campione di 1794 donne in gravidanza (dati di Brooke *et al.* 1989)

ammissibili sono ora 0, 2, 4 e 6. La distribuzione Binomiale con $n = 6$ ha tuttavia anche 1, 3 e 5 tra i valori ammissibili, dunque le due distribuzioni sono differenti e quella che abbiamo ottenuto non appartiene alla famiglia Binomiale.

Abbiamo visto che sommando una costante ad una variabile con distribuzione Normale si ottiene una variabile che ha ancora distribuzione Normale. Se sommiamo tra loro due variabili con distribuzione Normale, anche se di media e varianza differenti, la somma è ancora una variabile aleatoria Normale; analogamente, anche la differenza tra due variabili con distribuzione Normale ha ancora distribuzione Normale.

7.4 Variabili aleatorie con distribuzione Normale

Finora abbiamo parlato della distribuzione Normale in quanto essa si ottiene da un campionamento come somma o limite di altre distribuzioni. Tuttavia, molte variabili che si incontrano nella pratica, come l'altezza degli esseri umani, sembrano essere ben rappresentate dalla distribuzione Normale. Ci aspettiamo che questo accada ogni volata che la variabile è il risultato della somma di variabilità generate da un insieme di fonti diverse: il processo messo in evidenza dal teorema centrale del limite porta infatti ad un risultato molto vicino alla distribuzione Normale. La Figura 7.10 mostra la distribuzione delle altezze in un campione di donne in gravidanza, e la curva della corrispondente distribuzione Normale. L'adattamento alla distribuzione Normale risulta molto buono.

Se la variabile che misuriamo è il risultato del prodotto di diverse fonti di variabilità, non ci aspetteremo che il risultato sia Normale in base alle proprietà discusse in § 7.2, che erano tutte basate sulla somma di variabili. Tuttavia, se consideriamo la trasformazione logaritmica di tale variabile (Appendice 5A), allora

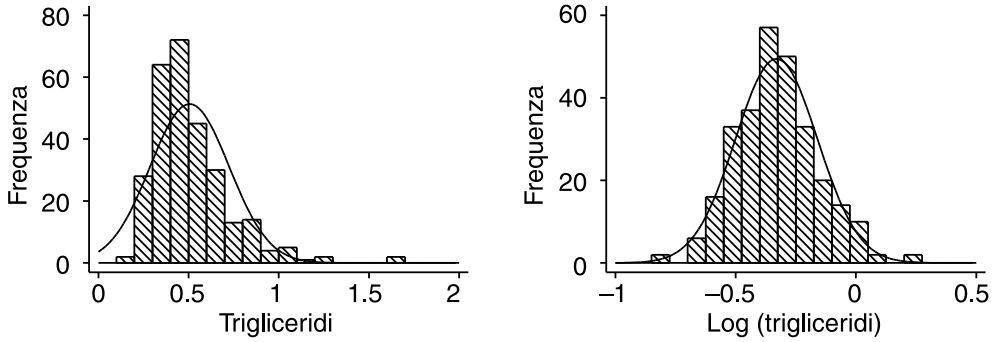

Figura 7.11: Distribuzione della concentrazione del siero trigliceride (Tabella 4.8) e del $\log_{10}$ del siero trigliceride nel cordone ombelicale di 282 bambini, con le corrispondenti curve di distribuzione Normale

otterremo una nuova variabile che è la somma di diverse fonti di variabilità, e che è verosimile abbia distribuzione Normale. Ciò accade di frequente quando abbiamo a che fare con quantità che sono parte di un ciclo metabolico, dal momento che il tasso di concentrazione in corrispondenza del quale può avere luogo la reazione dipende dalla concentrazione di altri composti; per esempio, molte misurazioni di costituenti del sangue si comportano così. In Figura 7.11 è mostrata la distribuzione del siero trigliceride misurata nel cordone ombelicale di 282 bambini (Tabella 4.8); la distribuzione è fortemente asimmetrica, e non c'è molta somiglianza con la curva di distribuzione Normale. Tuttavia, se consideriamo il logaritmo della concentrazione del trigliceride, otteniamo un adattamento decisamente migliore alla distribuzione Normale (Figura 7.11). Se il logaritmo di una variabile aleatoria segue la distribuzione Normale, allora la variabile aleatoria segue la **distribuzione Lognormale**.

Spesso desideriamo modificare la scala su cui analizziamo i nostri dati in modo da ottenere una distribuzione Normale. Definiamo questo procedimento, che consiste nell'analizzare una funzione matematica dei dati piuttosto che i dati stessi, **trasformazione**. Il logaritmo è la trasformazione che si utilizza più frequentemente, altri esempi sono la radice quadrata e il reciproco (si veda anche § 10.4). Per un singolo campione, la trasformazione ci mette in grado di utilizzare la distribuzione Normale per trovare i percentili (§ 4.5). Per esempio, spesso vogliamo stimare il 2.5-esimo e il 97.5-esimo percentile, valori che racchiudono il 95% delle osservazioni; per una distribuzione Normale, questi possono essere stimati da $\bar{x} \pm 1.96s$. Possiamo dunque trasformare i dati in modo che la loro distribuzione sia Normale, calcolare i percentili, e poi tornare alla scala originale.

Si considerino i dati della concentrazione di trigliceride di Figura 7.11 e la Tabella 4.8. La media è 0.51 e la deviazione standard 0.22; la media dei dati trasformati con $\log_{10}$ è -0.33 e la deviazione standard è 0.17. Cosa succede se antitrasformiamo i dati con la funzione inversa del logaritmo? Per la media, otteniamo $10^{-0.33} = 0.47$; questo valore è leggermente minore della media dei dati originali. Il logaritmo inverso della media dei logaritmi è cioè diverso dalla me-

dia aritmetica dei dati non trasformati; infatti, questa quantità si definisce **media geometrica**, che è la radice n-esima del prodotto delle osservazioni. Se sommiamo i logaritmi delle osservazioni otteniamo il logaritmo del loro prodotto (Appendice $5A$), e se moltiplichiamo il logaritmo di un numero per un secondo numero, otteniamo il logaritmo del primo numero elevato al secondo; quindi se dividiamo il logaritmo per n, otteniamo il logaritmo della radice n-esima, da cui si capisce come la media dei logaritmi sia il logaritmo della media geometrica. Anche se antitrasformiamo la media dei dati trasformati prendendone il reciproco otteniamo una media particolare, la **media armonica**, ovvero il reciproco della media dei reciproci.

La media geometrica è espressa nella stessa unità di misura dei dati originari. Se la concentrazione di trigliceride è misurata in $mmol/l$, il logaritmo di una singola osservazione è il logaritmo di una misurazione in $mmol/l$. La somma di n logaritmi è il logaritmo del prodotto di n misurazioni in $mmol/l$, ed è dunque il logaritmo di una misurazione in $mmol/l$ elevato alla n. La radice n-esima è dunque ancora il logaritmo di un numero in $mmol/l$, da cui prendendo il logaritmo inverso si torna all'unità di misura di partenza, $mmol/l$ (si veda Appendice $5A$).

Il logaritmo inverso della deviazione standard, tuttavia, non è espresso nelle unità di misura originarie. Per calcolare la deviazione standard, infatti, prendiamo la differenza tra il logaritmo di ogni osservazione e il logaritmo della media geometrica, utilizzando la formula usuale $\sum(x_i - \bar{x})^2/(n-1)$ (§ 4.8). Otteniamo dunque la differenza tra i logaritmi di due numeri entrambi misurati in $mmol/l$, da cui si ottiene il logaritmo del loro rapporto (§ $5A$), che risulta essere il logaritmo di un numero puro adimensionale. Se la concentrazione di trigliceride fosse stata misurata in $mg/100\,ml$, piuttosto che in $mmol/l$, avremmo ottenuto lo stesso risultato; dunque non possiamo antitrasformare la deviazione standard per tornare alla scala originaria.

Se vogliamo usare la deviazione standard, è più semplice effettuare tutti i calcoli nella scala trasformata, e poi antitrasformare, se necessario, alla fine. Per esempio, il 2.5-esimo percentile in scala logaritmica è $-0.33 - 1.96 \times 0.17 = -0.66$ e il 97.5-esimo percentile è $-0.33 + 1.96 \times 0.17 = 0.00$. Per ottenere questo valore abbiamo preso il logaritmo di qualcosa in $mmol/l$ e abbiamo sommato o sottratto il logaritmo di un numero puro (ovvero moltiplicato o diviso, se pensiamo alla scala naturale), dunque abbiamo ancora il logaritmo di qualcosa in $mmol/l$. Per tornare alla scala originaria applichiamo il logaritmo inverso per ottenere il 2.5-esimo percentile pari a 0.22 e il 97.5-esimo percentile pari a $1.00\,mmol/l$.

Trasformare i dati per ottenere una distribuzione Normale e poi analizzarli sulla scala trasformata potrebbe sembrare come barare. Io non penso sia così. La scala su cui decidiamo di effettuare le misurazioni non deve necessariamente essere lineare, anche se di solito è conveniente che lo sia. Altre scale possono risultare anche più utili; per esempio, il pH si misura di solito su scala logaritmica. È più conveniente misurare l'intensità di un terremoto in mm di ampiezza (lineare), oppure sulla scala Richter (logaritmica)? Le lenti degli occhiali dovrebbero essere misurate in termini di lunghezza focale in cm (lineare), oppure in diottrie (reciproco)? Spesso scegliamo scale non lineari perché si prestano bene ai nostri scopi, e ai

Tabella 7.3: Livelli di vitamina D misurati nel sangue di 26 uomini sani, dati di Hickish *et al.* (1989)

14	25	30	42	54
17	26	31	43	54
20	26	31	46	63
21	26	32	48	67
22	27	35	52	83
24				

fini dell'analisi statistica è spesso opportuno rendere una distribuzione Normale, trovando un'opportuna scala di misurazione in cui i dati siano così distribuiti.

7.5 Il grafico di probabilità Normale

Molte tecniche statistiche possono essere usate solo se i dati seguono la distribuzione Normale (Capitoli 10 e 11). Ci sono molti modi per stabilire se le osservazioni seguono una distribuzione Normale. Avendo a disposizione un campione di ampiezza elevata possiamo osservare l'istogramma, per vedere se si avvicina alla densità di probabilità Normale; ciò non funziona bene con campioni di piccole dimensioni, e un metodo più affidabile risulta essere il **grafico di probabilità normale**. È un metodo grafico, che può essere implementato usando della carta comune e una tavola della distribuzione Normale, oppure con dei fogli di carta specifici per la distribuzione Normale, oppure, più semplicemente, con il calcolatore. Un qualsiasi pacchetto di statistica di buon livello è in grado di generare un grafico di probabiltà Normale; in caso contrario non potrà essere definito un buon pacchetto di statistica. Il metodo del grafico di probabilità Normale può essere utilizzato per indagare l'assunzione di Normalità in campioni di qualsiasi dimensione, ed è un controllo molto utile quando si utilizzano metodi come quello della distribuzione t descritto nel Capitolo 10.

Il grafico di probabilità Normale è un grafico della distribuzione di frequenze cumulata dei dati contro la distribuzione di frequenze cumulata della distribuzione Normale. Innanzitutto, si ordinano i dati in senso crescente; per ogni osservazione ordinata si trova poi il valore che ci aspetteremmo per l'osservazione se i dati seguissero una distribuzione Normale Standard. Ci sono molte formule approssimate per effettuare questo calcolo. Io seguirò quella di Armitage e Berry (1994) e associerò all'i-esima osservazione il valore z, tale che $\Phi(z) = (i - 0.5)/n$; alcuni libri e programmi utilizzano $\Phi(z) = i/(n + 1)$, ed esistono altre formule più complesse. Non fa molta differenza quale formula scegliamo di usare. Troviamo dunque da una tavola della distribuzione Normale i valori di z che corrispondono a $\Phi(z) = 0.5/n$, $1.5/n$, ecc. (la Tabella 7.1 è troppo poco dettagliata per gli scopi della pratica, ma ci servirà per spiegare il procedimento). Per 5 dati, per esempio, otteniamo $\Phi(z) = 0.1$, 0.3, 0.5, 0.7 e 0.9, e $z = -1.3$, -0.5, 0, 0.5 e 1.3. Questi sono i punti della distribuzione Normale Standard che corrispondono ai dati osser-

Tabella 7.4: Calcolo del grafico di probabilità Normale per i dati sulla vitamina D

i	Vit D	$\Phi(z)$	z	i	Vit D	$\Phi(z)$	z
1	14	0.019	−2.07	14	31	0.519	0.05
2	17	0.058	−1.57	15	32	0.558	0.15
3	20	0.096	−1.30	16	35	0.596	0.24
4	21	0.135	−1.10	17	42	0.635	0.34
5	22	0.173	−0.94	18	43	0.673	0.45
6	24	0.212	−0.80	19	46	0.712	0.56
7	25	0.250	−0.67	20	48	0.750	0.67
8	26	0.288	−0.56	21	52	0.788	0.80
9	26	0.327	−0.45	22	54	0.827	0.94
10	26	0.365	−0.34	23	54	0.865	1.10
11	27	0.404	−0.24	24	63	0.904	1.30
12	30	0.442	−0.15	25	67	0.942	1.57
13	31	0.481	−0.05	26	83	0.981	2.07

$\Phi(z) = (i - 0.5)/26$

vati. Ora, se i dati osservati sono estratti da una distribuzione Normale di media μ e varianza σ^2, i punti osservati dovrebbero essere dati da $\sigma z + \mu$, dove z è il punto corrispondente sulla distribuzione Normale Standard. Se dunque tracciamo un grafico dei punti della Normale Standard contro i valori osservati, dovremmo ottenere qualcosa di molto vicino ad una linea retta. Possiamo scrivere l'equazione di questa retta nel seguente modo: $\sigma z + \mu = x$, dove x è la variabile osservata e z il quantile corrispondente sulla distribuzione Normale Standard. Possiamo riscriverla come

$$z = \frac{x}{\sigma} - \frac{\mu}{\sigma},$$

retta che nel piano (x, z) passa per il punto $(\mu, 0)$ e ha pendenza $1/\sigma$ (si veda § 11.1). Se i dati non provengono da una distribuzione Normale, non otterremo una linea retta, ma una curva di qualche genere. Dal momento che stiamo tracciando un grafico dei quantili della distribuzione di frequenze empirica contro i quantili corrispondenti di quella teorica (qui la Normale), questo grafico viene anche chiamato **grafico quantile-quantile** o **q-q plot**.

La Tabella 7.3 mostra i livelli di vitamina misurati nel sangue di 26 uomini sani. I calcoli per il grafico di probabilità Normale sono mostrati nella Tabella 7.4. Si osservi che $\Phi(z) = (i - 0.5)/26$ e che i valori di z sono simmetrici, dato che la seconda metà è esattamente uguale alla prima con il segno cambiato; il valore del quantile della distribuzione Normale Standard, z, può essere trovato interpolando la Tabella 7.1, utilizzando una tavola più completa, oppure con un calcolatore. La Figura 7.12 mostra l'istogramma e il grafico di probabilità Normale per questi dati; la distribuzione è asimmetrica e il grafico di probabilità Normale mostra un'evidente curvatura. La Figura 7.12 mostra anche i dati sulla vitamina D dopo la trasformazione logaritmica; è abbastanza semplice tracciare il grafico di probabilità Normale, dal momento che i valori di z sono immutati. Dobbiamo solo prendere il logaritmo delle osservazioni e poi tracciare un altro grafico. Il grafico di probabilità Normale per i dati trasformati si adatta molto bene alla retta prevista

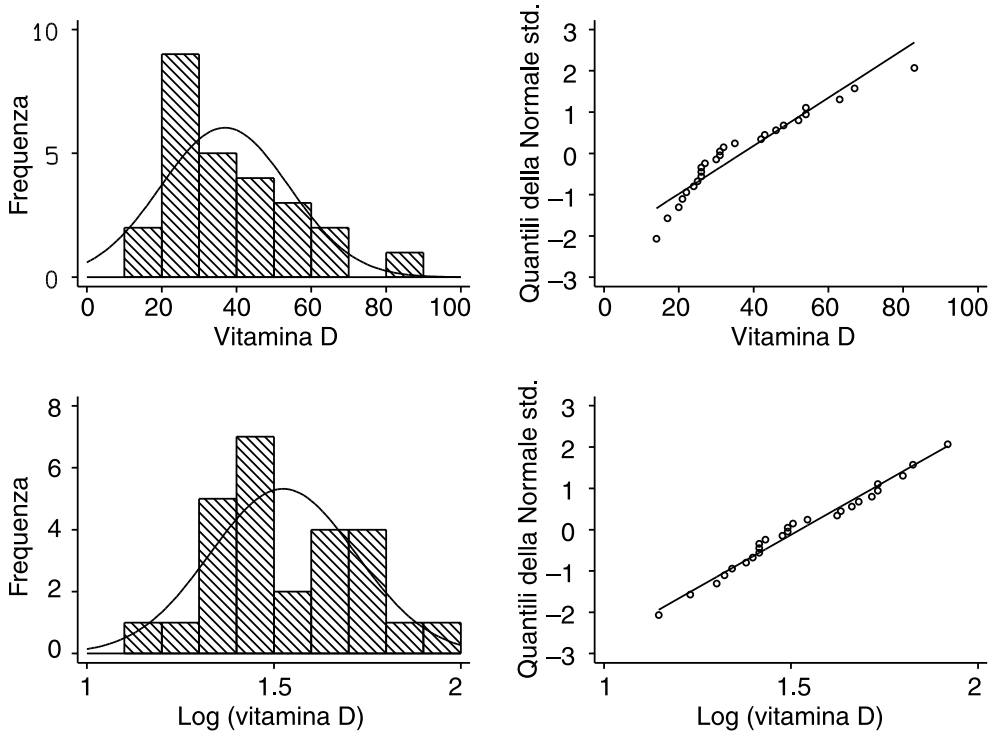

Figura 7.12: Livello di vitamina D nel sangue e $\log_{10}$ del livello di vitamina D per 26 uomini sani, con i corrispondenti grafici di probabilità Normale

dalla teoria, suggerendo che la distribuzione del logaritmo dei dati sulla vitamina D sia approssimativamente Normale.

Un singolo tratto curvilineo nel grafico di probabiltà Normale indica asimmetria. Un duplice tratto curvilineo indica che entrambe le code della distribuzione sono diverse da quelle della Normale, di solito troppo lunghe, e molti tratti curvilinei possono indicare che la distribuzione è bimodale (Figura 7.13). Quando il campione è di piccole dimensioni, ovviamente, sono possibili alcune fluttuazioni aleatorie.

Ci sono diversi modi per tracciare un grafico di probabilità Normale. Alcuni programmi riportano la distribuzione dei dati sull'asse verticale e la distribuzione Normale teorica sull'orizzontale, il che capovolge l'andamento della curva. Altri riportano la distribuzione Normale teorica con media $\bar{x}$, la media campionaria, e deviazione standard s, la deviazione standard campionaria; questo risultato si ottiene calcolando $\bar{x} + sz$. La Figura 7.14(a) mostra proprio questo approccio, ovvero il grafico di probabilità Normale ottenuto con il programma "qnorm" di Stata. La linea retta è la bisettrice; il grafico è esattamente identico al secondo di Figura 7.12, a meno di un cambio di scala e dopo aver scambiato gli assi. Leggermente diverso è il **grafico di probabilità Normale standardizzato** o **p-p plot**, dove le osservazioni vengono standardizzate in modo da avere media 0 e deviazione standard 1, $y = (x - \bar{x})/s$, e poi viene tracciato un grafico delle probabilità Normali cumulate, $\Phi(y)$, contro $(i - 0.5)/n$ oppure $i/(n + 1)$ (Figura

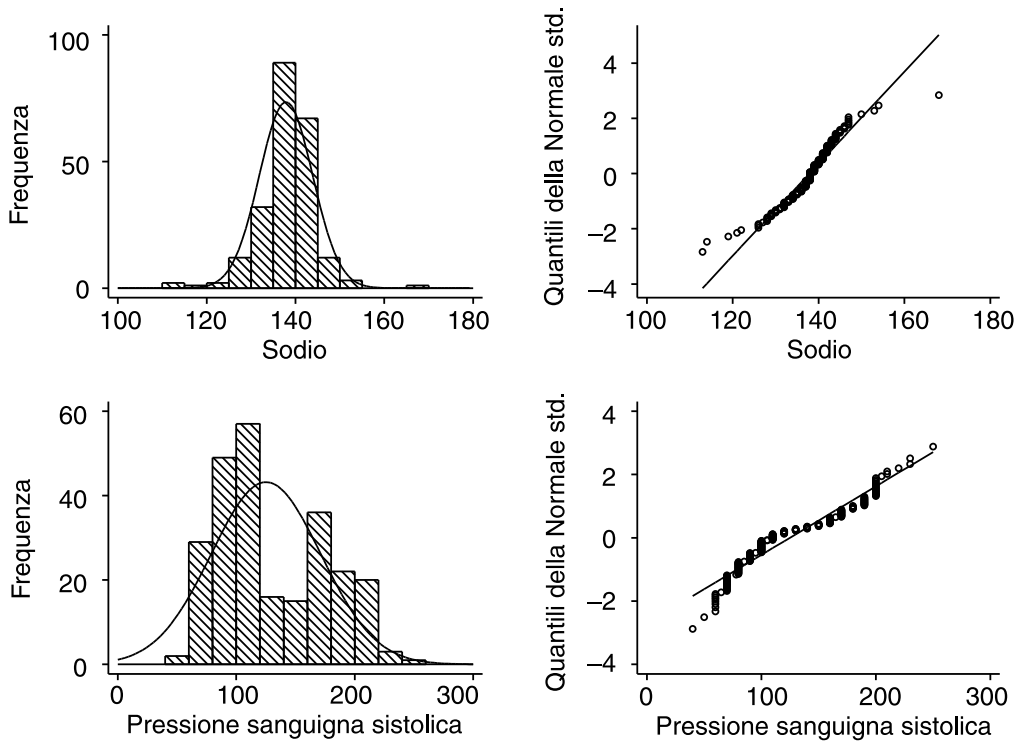

Figura 7.13: Sodio nel sangue e pressione sanguigna sistolica misurati in 250 pazienti del Reparto di Terapia Intensiva del St. George's Hospital, con i corrispondenti grafici di probabilità Normale (dati di Freidland *et al.* 1996)

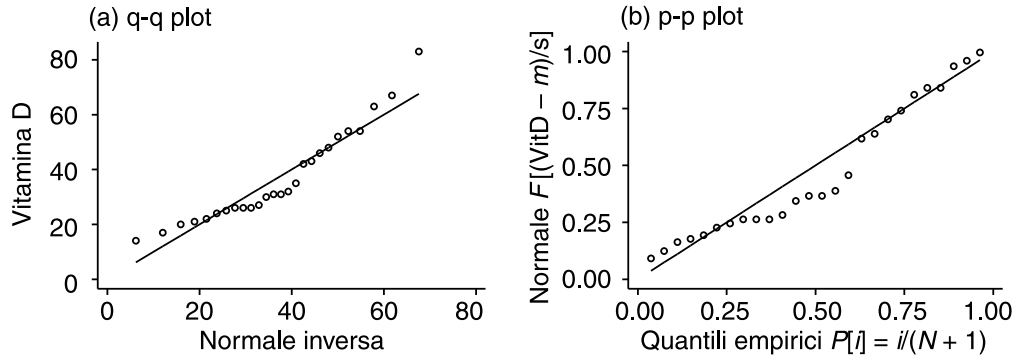

Figura 7.14: Diverse versioni del grafico di probabilità Normale per i dati sulla vitamina D

7.14(b), prodotta grazie al programma "pnorm" di Stata). C'è davvero una differenza molto lieve tra le Figure 7.14(a) e (b), dunque le due versioni del grafico di probabilità Normale, quella con i quantili e quella con le probabilità, dovrebbero essere interpretate nello stesso modo.

7A Appendice: chi-quadro, t e F

I lettori meno inclini alla matematica possono tranquillamente saltare questa sezione, ma coloro che hanno intenzione di proseguire dovrebbero poi trovare molto più comprensibili applicazioni come il test chi-quadro (Capitolo 13).

Molte distribuzioni di probabilità possono essere introdotte a partire da trasformazioni di variabili aleatorie Normali che scaturiscono nell'analisi statistica. Tre di queste sono particolarmente importanti: le distribuzioni chi-quadro, t e F. Queste distribuzioni hanno molte applicazioni, alcune delle quali saranno discusse nei prossimi capitoli.

La distribuzione chi-quadro è definita nel modo seguente. Si supponga che Z sia una variabile aleatoria Normale Standard, dunque di media 0 e varianza 1; allora la variabile aleatoria Z^2 ha distribuzione chi-quadro a 1 grado di libertà. Se consideriamo n variabili aleatorie Normali Standard indipendenti, $Z_1, Z_2, \ldots, Z_n$, allora la variabile data da

$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_n^2$$

si definisce **distribuzione Chi-quadro a *n* gradi di libertà**; χ è la lettera greca "chi". Le curve della densità di probabilità chi-quadro per diversi valori dei gradi di libertà sono riportate in Figura 7.15; la descrizione matematica di questa funzione è piuttosto complicata, ma non abbiamo bisogno di ulteriori approfondimenti.

Alcune proprietà della distribuzione chi-quadro sono semplici da ricavare. Dal momento che la distribuzione è la somma di n variabili aleatorie indipendenti e identicamente distribuite essa tende ad una Normale al crescere di n, per il teorema centrale del limite (§ 7.2). La convergenza, tuttavia, è lenta (Figura 7.15), e la radice quadrata della chi-quadro converge molto più rapidamente. Il valore atteso di Z^2 è la varianza di Z, dato che il valore atteso di Z è 0, dunque $\mathrm{E}(Z^2) = 1$. Il valore atteso di una chi-quadro a n gradi di libertà è, dunque, n:

$$\mathrm{E}(\chi^2) = \mathrm{E}\left(\sum_{i=1}^{n} Z_i^2\right) = \sum_{i=1}^{n} \mathrm{E}(Z_i^2) = \sum_{i=1}^{n} 1 = n.$$

La varianza è $VAR(\chi^2) = 2n$. La radice quadrata di χ^2 ha media circa uguale a $\sqrt{n - 0.5}$ e varianza circa 0.5.

La distribuzione chi-quadro gode di una proprietà molto importante. Supponiamo di limitare la nostra attenzione ad un sottoinsieme dei possibili esiti delle n variabili aleatorie $Z_1, Z_2, \ldots, Z_n$; il sottoinsieme sarà definito da quei valori di $Z_1, Z_2, \ldots, Z_n$ che soddisfano l'equazione $a_1 Z_1 + a_2 Z_2 + \cdots + a_n Z_n = k$, dove $a_1, a_2, \ldots, a_n$ e k sono costanti (questo è chiamato **vincolo lineare**). Allora sotto