INTRODUZIONE

«There was something
I meant to say»

(Howard Devoto)

Perché si effettua l'analisi dei dati? Quali sono le domande alle quali un ricercatore cerca di rispondere attraverso l'analisi dei dati? Quali sono le tecniche più appropriate per rispondere a tali domande, e quale il modo più appropriato per applicare queste tecniche?

La motivazione a scrivere questo libro nasce dal tentativo di rispondere a questi interrogativi: la mia speranza è che, giunto alla fine del libro, il lettore possa darsi delle risposte sufficientemente precise. Certamente non ho voluto scrivere un ennesimo libro di introduzione all'analisi dei dati in psicologia e nelle scienze sociali. Nel testo credo di aver espresso, in maniera sufficientemente non ambigua, la mia posizione rispetto a differenti questioni, che sono oggetto di dibattito e che riguardano l'utilizzo di diverse tecniche di analisi dei dati nella ricerca empirica. Ho cercato di trattare gli argomenti limitando l'utilizzo del formalismo matematico, consapevole del diffuso *analfabetismo matematico* tra i ricercatori e gli studenti che operano nel campo delle discipline psicologico-sociali. La conoscenza di concetti elementari di matematica e algebra, e di nozioni e concetti che solitamente vengono forniti nei corsi introduttivi alla statistica, sono sufficienti per la comprensione di quanto viene trattato nelle pagine seguenti.

Lavorando da molti anni nel campo dell'analisi dei dati mi sono sempre più convinto dell'esistenza di due approcci fonda-

mentali nell'atteggiamento che i ricercatori e più in generale gli utenti hanno verso l'analisi dei dati. C'è un atteggiamento di tipo eminentemente pratico-esplorativo, e un atteggiamento di tipo più teorico-confermativo. L'atteggiamento pratico trova la sua espressione nell'attitudine a farsi suggerire dai dati le risposte ai propri interrogativi. Questo atteggiamento è predominante in alcune tecniche come il *data mining* (vedi Del Cello, Dulli e Saccardi, 2000) e, per certi versi, in alcune applicazioni 'automatiche' di tecniche che invece sono più congeniali all'approccio teorico-confermativo (mi riferisco in particolare alle procedure *stepwise* dell'analisi di regressione multipla e dell'analisi discriminante).

L'atteggiamento teorico-confermativo trova la sua espressione nella messa in atto di strategie di analisi guidate da idee e ipotesi (più o meno precise) che il ricercatore forma in merito ai dati, prima ancora di cominciare l'analisi vera e propria. L'analisi fattoriale (sia nella versione 'classica' o esplorativa, sia nella versione propriamente confermativa), la regressione multipla (soprattutto nell'approccio gerarchico), l'ANOVA e ovviamente i modelli di equazioni strutturali (che integrano gli aspetti più importanti delle tecniche precedenti in un unico modello statistico) rappresentano le tecniche che, più di altre, sono coerenti con questo secondo tipo di atteggiamento.

Il problema più serio si pone, e le conseguenze più gravi si producono, quando tecniche che richiedono al ricercatore solide ipotesi preliminari vengono utilizzate con l'attitudine pratico-esplorativa del genere 'buttiamo i dati dentro il computer e vediamo cosa il computer ci tira fuori'. In un bellissimo lavoro pubblicato nel 1999 su *Psychological Methods*, Fabrigar, Wegener, McCallum e Strahan hanno evidenziato in maniera chiara le conseguenze che questo genere di attitudine ha sull'analisi fattoriale. Thompson (1995) ha fatto altrettanto per l'analisi di regressione e per l'analisi discriminante.

Spesso è la mancanza di conoscenza delle assunzioni e dei principi alla base delle diverse tecniche utilizzate ad essere la causa di errori madornali. A questo riguardo, la disponibilità di

calcolatori super-veloci non solo nei centri di calcolo delle università ma direttamente a casa nostra, e di programmi sempre più *user friendly* per effettuare complesse analisi statistiche, sembra aver prodotto più guasti che altro. La tendenza attuale alla semplificazione dei programmi può generare nell'utente l'illusione di dominare la tecnica. L'utente è in grado ora di realizzare, con un paio di clic del mouse, complesse analisi che, in passato, richiedevano giorni e giorni di programmazione al calcolatore. La mia personale esperienza è che solo una profonda conoscenza del modello statistico può garantire il dominio della tecnica. Se poi il programma che si utilizza è semplice e facile da utilizzare, ancora meglio, ma questa non è sicuramente la condizione necessaria per applicare in maniera corretta una tecnica di analisi dei dati.

E gli sfaceli non vengono compiuti prevalentemente e soprattutto dagli studenti, ma dai loro docenti se, come documentano bene Fabrigar *et al.* (1999) a proposito dell'analisi fattoriale, buona parte dei lavori pubblicati su due importantissime ed elitarie riviste scientifiche sono inficiati da errori a volte veramente marchiani. Il problema è che molte delle tecniche che costituiscono la 'cassetta degli attrezzi' (come dice Luca Ricolfi) del *data analyst* si fondano su delicate assunzioni, il mancato rispetto delle quali comporta conseguenze dannose sui risultati. Quando l'attitudine 'praticona' si coniuga con la scarsa conoscenza del modello statistico, i disastri di cui parlano Fabrigar e Thompson sono assicurati.

Questo libro è diretto *in primis* agli studenti di discipline psicologiche e sociali, e più in generale a chi si occupa di analisi dei dati nella propria attività di ricerca, e/o nel proprio lavoro applicativo. Gli argomenti esaminati sono stati scelti in maniera parziale, nel senso che riflettono i miei interessi di ricerca e le tecniche con le quali ho maggiore dimestichezza e che utilizzo da più tempo. Si tratta di tecniche che spesso rappresentano veri e propri *modelli* di analisi dei dati (nel senso inteso da Ricolfi, 2002). Questi modelli sono spesso riconducibili ad un'equazione come la seguente: $y = a + bx$. Si tratta dunque, nella maggior par-

te dei casi, di *modelli lineari* (l'eccezione principale è sicuramente rappresentata dalla *cluster analysis* che viene discussa nel sesto capitolo). Si tratta di tecniche di solito utilizzate per analizzare simultaneamente gruppi piuttosto ampi di variabili, quindi di tecniche di analisi multivariata.

Molti sono gli argomenti rimasti fuori. Innanzitutto, i modelli di equazioni strutturali, che ritengo così importanti da dedicarvi un libro intero (che nel momento in cui scrivo è ancora in preparazione). Il lettore italiano, nel frattempo, può trovare nel testo di Corbetta (1992) una discreta introduzione alle principali tematiche che riguardano questa tecnica. Il lettore non troverà nulla, poi, che riguarda alcune fondamentali tecniche applicate per l'analisi di dati categoriali, come i modelli loglineari, la regressione logistica e l'analisi delle corrispondenze. Si tratta di tecniche ampiamente discusse in alcuni testi pubblicati recentemente nel nostro paese (ad esempio, Cristante, 2000, e Robusto e Cristante, 2001, per i modelli log-lineari; Bohrnstedt e Knoke, 1994, e Fabbris, 1997, per la regressione logistica; Bolasco, 1999, per l'analisi delle corrispondenze). Un altro argomento *negletto* è quello relativo ai metodi di *scaling*: il testo di Cinanni (1990) rappresenta ancora e sicuramente un importante riferimento per il lettore italiano. È un vero peccato che sia ormai quasi introvabile, ed è molto triste che la vita non abbia concesso a Vincenzo Cinanni il tempo per comunicarci compiutamente la *sua* visione dell'analisi dei dati.

Mancano poi altri modelli veramente importanti come quelli sviluppati nell'ambito della teoria della risposta all'item (*Item Response Theory*, IRT) che stanno finalmente trovando una diffusione più ampia, oltre gli angusti confini delle discipline pedagogiche all'interno dei quali per molti anni hanno trovato ospitalità. Le potenzialità applicative e la potenza innovativa della IRT rispetto alla obsoleta teoria classica dei test sono fortemente trascurate nel nostro paese (eccezioni, limitate al solo modello di Rasch, sono rappresentate da Cristante, 1991, Giampaglia, 1990, e Miceli, 2001). Come quello dei modelli di equazioni strutturali, anche questo tema merita una trattazione a parte. Le potenzialità

della IRT sono state particolarmente evidenziate da fondamentali contributi ad opera dello statistico svedese-americano Bengt Muthén (2000), che recentemente ha unificato IRT, modelli di equazioni strutturali, regressioni logistica e probit, e *modelli di mistura* sotto un unico modello statistico matematico: il suo programma, MPLUS (Muthén e Muthén, 1998), rappresenta un'importantissima novità nell'affollato panorama dell'informatica applicata alla statistica e all'analisi dei dati (alcune sue applicazioni vengono discusse nel terzo capitolo di questo testo). Nulla viene dedicato, inoltre, all'insieme di tecniche e algoritmi di esplorazione dei dati che vanno sotto il nome di *data mining*: si tratta di metodi che stanno avendo una larga diffusione soprattutto nelle ricerche di mercato e nella prassi aziendale (vedi Del Ciello, Dulli e Saccardi, 2000), e che fanno uso sia di tecniche di analisi dei dati più tradizionali (come la *cluster analysis* e l'analisi fattoriale), sia di modelli statistico-matematici, come le reti neurali, applicati finora soprattutto nella psicologia cognitiva (vedi Cammarata, 1997).

Ho evitato infine di entrare nella discussione relativa ai presupposti e ai fondamenti epistemologici dell'analisi dei dati. Il recente libro di Luca Ricolfi (2002) dedica ampio spazio a queste tematiche, oltre a fornire un interessante profilo storico di questa disciplina.

Dopo aver parlato di quello che non c'è nel libro, forse è il caso di spendere qualche parola per descrivere brevemente quello che invece il lettore troverà nelle prossime pagine. Il primo capitolo è dedicato al trattamento preliminare dei dati, ed introduce alcuni temi che verranno poi ripresi nei capitoli successivi (in particolare la verifica delle assunzioni). Il secondo capitolo affronta il tema della regressione lineare, partendo dal modello più semplice (quando c'è una sola variabile indipendente) fino ad arrivare all'esame dei modelli di regressione multipla. Particolare enfasi è posta sul problema delle diverse strategie analitiche (standard, gerarchica, statistica), sulla verifica delle assunzioni e sulle conseguenze della loro violazione. Il terzo capitolo è dedicato all'analisi fattoriale *esplorativa*. Larga parte

del capitolo è dedicata alla presentazione del modello di base, delle equazioni fondamentali, e di alcune problematiche ancora non risolte. Anche in questo caso, la mia preoccupazione è stata quella di mettere sull'avviso il lettore rispetto agli usi distorti che di tale tecnica si possono fare (e spesso si fanno). Il quarto capitolo è dedicato ai modelli di analisi della varianza: spazio rilevante viene dato ai disegni fondamentali, senza tralasciare la presentazione di disegni meno diffusi ma ugualmente importanti. Il quinto capitolo è dedicato all'analisi discriminante e all'analisi multivariata della varianza, due tecniche per analizzare in maniera multivariata le differenze tra gruppi preformati di soggetti. Il sesto capitolo è dedicato alla classificazione, e tratta un insieme di tecniche (definite complessivamente *cluster analysis*) che consentono di classificare unità statistiche in raggruppamenti non noti a priori. Completano il libro tre appendici dedicate rispettivamente all'algebra matriciale, all'algebra dei valori attesi e delle covarianze, e all'interpretazione e all'uso delle tabelle di alcune distribuzioni di probabilità.

Gli esempi discussi nei diversi capitoli sono funzionali alle tecniche esaminate, e non fanno riferimento a specifici programmi informatici per l'analisi dei dati. Questa scelta è dettata dalla considerazione che la conoscenza delle basi teoriche delle diverse tecniche prescinde dalla applicazione delle stesse tramite un particolare software, e ad essa è propedeutica: questo testo, dunque, è diretto a chiunque abbia interesse all'analisi dei dati, indipendentemente dal programma che utilizza per effettuarla. Questo non vuol dire che i programmi non sono importanti, anzi! Sono fermamente convinto che la statistica multivariata sarebbe difficilmente attuabile senza l'ausilio di programmi informatici: le tecniche che sono presentate in questo testo, e altre ancora, sono discusse, considerando il programma SPSS, in un libro scritto da me in collaborazione con Francesca D'Olimpio.

Ci sono diverse persone che sento il dovere di ringraziare. Innanzitutto Gian Vittorio Caprara, con il quale condivido un progetto culturale e scientifico che dura da più di 15 anni, e che si è preso l'onere di leggere la versione *draft* di questo libro: mol-

tissime delle riflessioni che ho cercato di elaborare in questo libro sono state stimolate dalle ricerche che stiamo portando avanti insieme ad altri colleghi del Centro Interuniversitario per lo studio delle condotte prosociali e antisociali, e che ci hanno portato a collaborare, in questi anni, con personaggi come Albert Bandura, Phil Zimbardo, Andy Comrey, Ken Dodge. Devo ringraziare poi Michele Vecchione ed Elena Natali, che in diverse fasi della gestazione e del parto di questo volume mi hanno aiutato a definirne meglio i contenuti e a chiarire quello che poteva risultare ambiguo. Voglio infine ringraziare Francesca D'Olimpio, Anna Paola Ercolani, Fabio Ferlazzo, Roberta Fida, Franco Lucchese, Fabio Lucidi, Laura Picconi e Patrizia Steca, che hanno letto e commentato la versione *draft* del testo. Questa infine non è una chiamata in correo: la responsabilità di tutto quanto si trova nelle prossime pagine è solamente e unicamente mia.

Roma, dicembre 2002 C.B.

1.

TRATTAMENTI PRELIMINARI DEI DATI

In questo capitolo vengono discussi diversi modi di trattare i dati preliminarmente alle elaborazioni di analisi dei dati vere e proprie. Si affronteranno problemi legati alle proprietà distributive delle variabili, e ai modi in cui le variabili possono essere «trasformate» o «ricodificate» per ottenere distribuzioni con proprietà più adatte per il tipo di analisi che si intende effettuare. Si parlerà anche dell'influenza che alcuni casi che presentano punteggi estremi in una variabile, o configurazioni particolarmente «strane» in più variabili, possono avere sui risultati delle analisi, e delle procedure che sono state proposte per individuare tali casi «anomali».

1.1. Dalla distribuzione normale univariata alla distribuzione normale multivariata

In questo primo paragrafo intendiamo riprendere alcuni concetti che dovrebbero essere già stati acquisiti nei corsi di statistica di base, e introdurne di nuovi che risultano particolarmente cruciali per l'analisi multivariata dei dati.

La distribuzione normale «è la distribuzione teorica per eccellenza …; … essa costituisce un punto di riferimento nella ricerca empirica, sia nelle scienze naturali sia nelle scienze umane»

(Albano e Testa, 2002, p. 178). La normalità della distribuzione è alla base di molte analisi univariate o multivariate, e per questo motivo rappresenta un'assunzione che va verificata sui dati. Nel caso della distribuzione normale univariata si fa riferimento alla distribuzione di una singola variabile, nel caso della normalità multivariata si fa riferimento alla generalizzazione della normale quando le variabili che vengono considerate sono k (dove $k > 1$).

Come è noto, la distribuzione normale univariata assume la classica forma «a campana», è unimodale, simmetrica rispetto alla media (quindi media e mediana coincidono, e coincidono anche con la moda), e presenta due punti di flesso per $x = \mu - \sigma$, e $x = \mu + \sigma$ (dove μ e σ sono rispettivamente la media e la deviazione standard della distribuzione). In realtà è bene parlare di «famiglia di distribuzioni normali univariate», alla quale appartengono appunto diverse distribuzioni normali definite da due parametri fondamentali: la media e la deviazione standard della distribuzione [1]. Tutte le distribuzioni normali possono essere ricondotte ad un'unica distribuzione, la normale standardizzata, attraverso le usuali procedure di calcolo dei punteggi standardizzati z. Tale distribuzione è nota e la probabilità dei suoi valori è stata tabulata: qualsiasi testo di statistica di base illustra le procedure per utilizzare i valori tabulati della normale standardizzata (vedi, ad esempio, Albano e Testa, 2002; Ercolani, Areni, e Leone, 2002; Gori, 1999; Vidotto, Xausa e Pedon, 1996): ciò la rende particolarmente utile nella verifica delle ipotesi statistiche.

Ci sono diversi metodi per esaminare se una variabile è normale. Le informazioni di questi diversi metodi vanno integrate per arrivare ad un giudizio finale. È possibile, innanzitutto, osservare l'istogramma della distribuzione di frequenze della variabile, per vedere se ha la tipica forma «a campana» (vedi *Figura*

[1] La funzione di probabilità della distribuzione normale è la seguente:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

1.1.). Un altro tipo di esame grafico è quello basato sulla rappresentazione dei «quantili» (Plot dei quantili, o *Q-Q Plot* o *Cumulative Normal Plot*): in questo esame si confrontano i quantili della distribuzione della variabile, rispetto ai quantili della distribuzione normale (si ricorda che i quantili sono quegli indici di posizione, come la mediana, i quartili, i decili, e i centili, che consentono di suddividere una distribuzione in un certo numero prestabilito di parti uguali). Il grafico in pratica confronta la distribuzione cumulata della variabile in esame con la distribuzione cumulata normale. Se la variabile si distribuisce in forma normale, i punti di tale distribuzione congiunta sono addensati sulla diagonale che va dal basso verso l'alto, e da sinistra verso destra (vedi *Figura 1.2.*).

È possibile utilizzare indici che valutano la forma della distri-

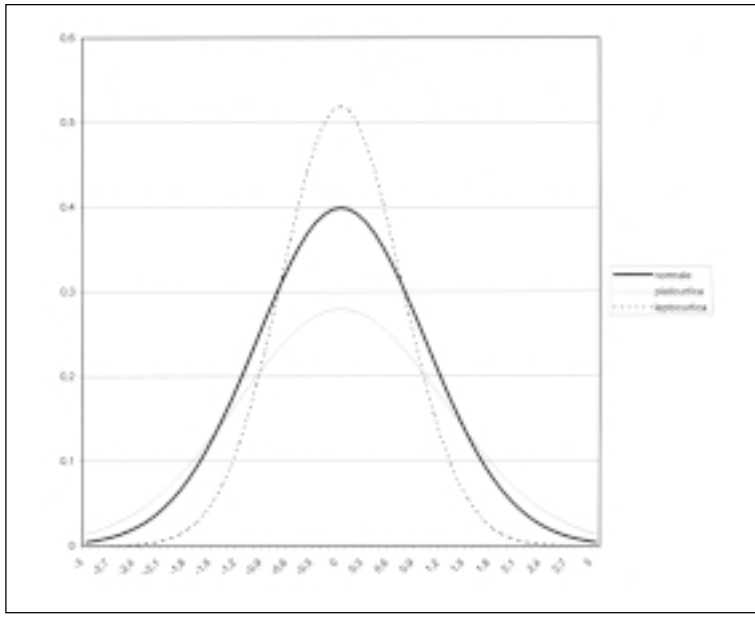

Figura 1.1. – La distribuzione normale univariata. In ascissa sono riportati i valori della distribuzione normale standardizzata, in ordinata la probabilità associata al valore.

buzione, come ad esempio gli indici di asimmetria (o *skewness*) e curtosi [2]. Questi indici possono essere sottoposti a verifica del-

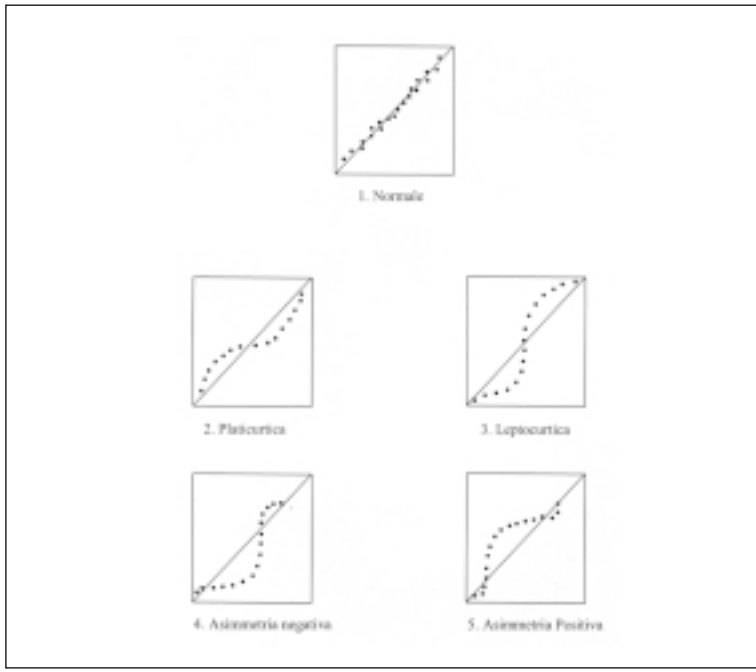

Figura 1.2. – Grafico dei quantili (Q-Q Plot) per la verifica della normalità. In ascissa sono riportati i valori osservati, in ordinata i valori attesi se la distribuzione è normale.

[2] Gli indici di asimmetria e curtosi rappresentano rispettivamente il «momento standardizzato» di 3° e di 4° ordine di una distribuzione, dove per momento si intende il valore medio di una variabile (Piccolo, 1998). L'indice di asimmetria proposto da Pearson si ottiene dalla seguente espressione:

$$\beta_1 = \left(\frac{\sum_{i=1}^{N} (x_i - \bar{x})^3}{N} \right)^2 \Bigg/ \left(\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N} \right)^3 .$$

le ipotesi: la procedura è piuttosto semplice in quanto si tratta di dividere il singolo indice (di asimmetria o di curtosi) per il suo errore standard, ed utilizzare come distribuzione di riferimento la normale standardizzata. Di solito questo test è troppo potente, ovvero risulta significativo quasi sempre, anche in presenza di distribuzioni decisamente normali. Per questo alcuni autori (es., Marcoulides e Hershberger, 1997; Muthén e Kaplan, 1985) considerano accettabili valori degli indici di asimmetria e curtosi compresi tra –1 e 1 (ovvero, valori compresi in questo intervallo indicano che la distribuzione è normale, oppure che la deviazione dalla normalità non è fonte di distorsioni apprezzabili). Altri autori (es., Tabachnick e Fidell, 1989) suggeriscono di essere più conservativi nella verifica delle ipotesi su asimmetria e curtosi, e quindi di utilizzare un livello di α più basso, ad esempio .01 o .001, invece del solito .05.

In generale, un valore di curtosi negativo indica una distribuzione «più schiacciata» verso il basso rispetto alla normale, che viene definita *platicurtica*. Un valore di curtosi positivo invece indica una distribuzione «più appuntita» rispetto alla normale, che viene definita *leptocurtica* (vedi *Figura 1.1.*). Un valore po-

---

Un indice più rapido ma forse più intuitivo per l'asimmetria si ottiene dalla seguente espressione:

$$\text{Asimmetria} = 3\frac{(\bar{x} - \text{Mediana})}{s_x},$$

dove s_x indica la deviazione standard della distribuzione. L'indice di curtosi proposto da Pearson si ottiene invece dalla seguente espressione:

$$\beta_2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{N} \Bigg/ \left(\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}\right)^2.$$

A questo indice di solito viene sottratto il valore 3 per renderlo uguale a 0 nel caso di perfetta distribuzione normale. Mentre l'errore standard della *skewness* è approssimativamente uguale a $(6/N)^{1/2}$, l'errore standard della curtosi è approssimativamente uguale a $(24/N)^{1/2}$, dove N rappresenta la numerosità del campione.

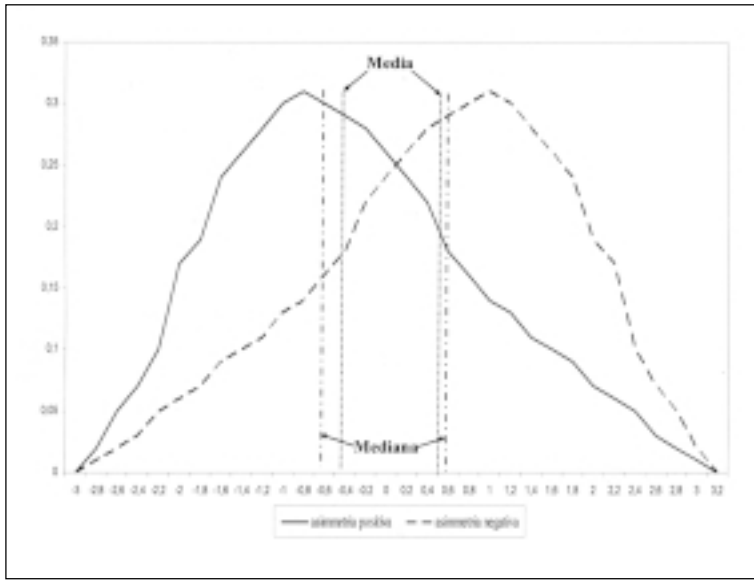

Figura 1.3. – Distribuzioni con asimmetria positiva e negativa.

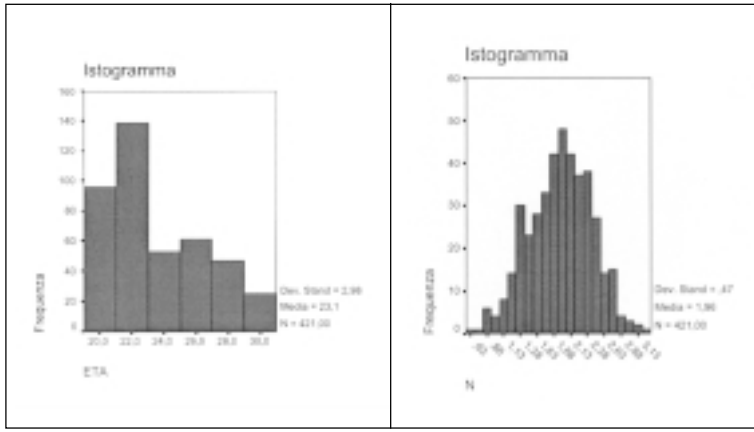

Figura 1.4. – Istogrammi delle distribuzioni.

sitivo di asimmetria indica una distribuzione nella quale i valori bassi hanno frequenza maggiore, e nella quale la media risulta maggiore della mediana. Un valore negativo di asimmetria invece indica una distribuzione nella quale sono i valori alti ad essere più frequenti, e nella quale la media risulta inferiore alla mediana (vedi *Figura 1.3.*).

Ci sono infine due test statistici, il test di Kolmogorov-Smirnov e il test di Shapiro-Wilk, che consentono di valutare se la distribuzione è normale. Se essi risultano significativi si deve rifiutare l'ipotesi nulla che la distribuzione sia normale. Si tratta, tuttavia, di test molto potenti, che, come per quanto detto in merito alla verifica delle ipotesi sui coefficienti di asimmetria e curtosi, conducono troppo spesso al rifiuto dell'ipotesi nulla.

1.1.1. Un esempio di applicazione

I dati di questo esempio riguardano un campione di 421 soggetti sui quali sono state rilevate la variabile età («ETA») espressa in anni, e il punteggio in una scala di personalità («N», Nevroticismo; vedi Costa e McCrae, 1992). Le elaborazioni sono state effettuate con il programma SPSS (vedi Barbaranelli e D'Olimpio, 2003).

La *Figura 1.4.* presenta gli istogrammi di frequenze, relativi alle due variabili, che sembrano evidenziare una piena normalità per la variabile «N», mentre la variabile «ETA» sembra decisamente meno normale. La *Figura 1.5.* presenta il grafico dei quantili delle due distribuzioni rispetto alla distribuzione normale. Mentre nel caso della variabile «ETA» si nota una certa dispersione dei punti intorno alla diagonale, nel caso della variabile «N» i punti sono tutti addensati intorno alla diagonale. Mentre il primo grafico mostra una deviazione dalla normalità soprattutto per i valori inferiori alla media (distribuzione platicurtica con asimmetria positiva), il secondo invece presenta un addensamento di punti sulla diagonale per tutti i valori osservati. Questi risultati dunque sembrano avvalorare quanto emerge dall'esame degli istogrammi.

La *Tabella 1.1.* presenta i risultati relativi alle statistiche descrittive di asimmetria e curtosi. Per quanto riguarda la variabile «ETA», anche se i valori dei due indici sono compresi tra –1 e 1, il rapporto tra statistica ed errore standard risulta uguale a 6.36 per l'asimmetria ed è uguale a –2.79 per la curtosi: in entrambi i casi è possibile rifiutare l'ipotesi nulla che la variabile segua la distribuzione normale.

Per quanto riguarda la variabile «N», i valori dei due indici sono prossimi a zero, e inoltre il rapporto tra statistica ed errore standard risulta uguale a –1.06 per l'asimmetria ed è uguale a

Tabella 1.1. – Indici di asimmetria e curtosi.

Statistiche decrittive					
	N	Asimmetria		Curtosi	
	Statistica	Statistica	Errore std	Statistica	Errore std
ETA	421	.757	.119	–.661	.237
N	421	–.126	.119	–.254	.237
Validi (*listwise*)	421				

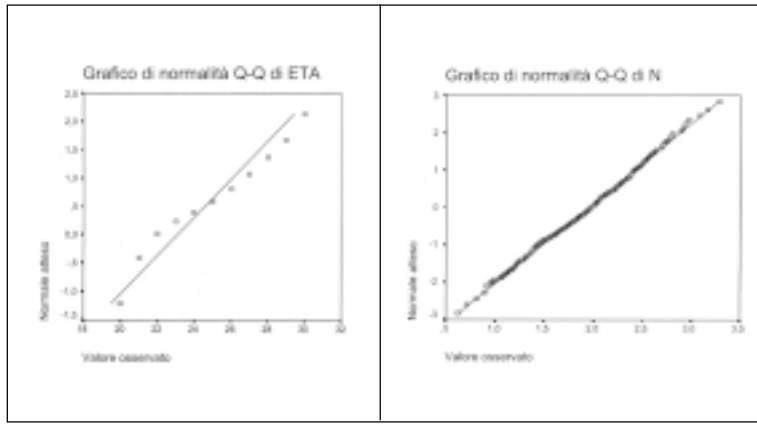

Figura 1.5. – Rappresentazione grafica dei quantili. In ascissa sono riportati i valori osservati, in ordinata i valori attesi se la distribuzione è normale.

Tabella 1.2. – Risultati relativi ai test di normalità.

Test di normalità						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistica	df	Sig.	Statistica	df	Sig.
ETA	.212	421	.000	.868	421	.000

[a] Correzione di significatività di Lilliefors; df = gradi di libertà.

Test di normalità						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistica	df	Sig.	Statistica	df	Sig.
N	.043	421	.063	.996	421	.385

[a] Correzione di significatività di Lilliefors; df = gradi di libertà.

–1.07 per la curtosi: in entrambi i casi non è possibile rifiutare l'ipotesi nulla che la variabile segua la distribuzione normale.

La *Tabella 1.2.* presenta i risultati relativi ai test di normalità. Entrambi i test confermano quanto emerso sia dall'esame grafico sia dagli indici di asimmetria e curtosi: la variabile «ETA» ha una distribuzione che devia in maniera significativa dalla normalità, mentre la variabile «N» risulta normale.

1.2. Valori anomali («outlier») univariati

I valori anomali sono quei valori che si distinguono in maniera particolare rispetto agli altri valori nella distribuzione dei punteggi. Dobbiamo differenziare tra valori anomali univariati e multivariati. I valori anomali, o *outlier*, univariati sono quei casi che in una variabile presentano valori estremamente elevati o estremamente bassi rispetto al resto della distribuzione.

Per individuare gli *outlier* univariati è possibile standardizza-

re i punteggi relativi alla variabile in esame e calcolare una distribuzione delle frequenze. Vengono considerati come possibili valori anomali quei punteggi che corrispondono a un punteggio z maggiore di 3 in valore assoluto (Tabachnick e Fidell, 1989).

In ogni caso, è necessario considerare la distribuzione nella sua interezza, e vedere se i punteggi troppo alti o troppo bassi rappresentano casi isolati dal resto della distribuzione.

I valori anomali possono influenzare molti indicatori, come la media, la deviazione standard, l'asimmetria e la curtosi. Essi possono anche influenzare gli indici di associazione tra le variabili come il coefficiente di correlazione di Pearson. Sono inoltre disponibili alcune statistiche (come la distanza di Cook e l'indice di *leverage*) che consentono di esaminare l'influenza di un caso estremo sui risultati di particolari procedure statistiche (vedi Barbaranelli e D'Olimpio, 2003).

Come esempio, consideriamo la distribuzione relativa alla variabile «N» già esaminata nelle *Tabelle 1.1.* e *1.2.* Dopo aver standardizzato la variabile, non si notano casi con valori standardizzati maggiori di 3 o minori di –3. Ci sono però due casi che si stagliano rispetto al resto della distribuzione: uno con valore pari a –2.8, e uno con un valore pari a 2.8. Soprattutto quest'ultimo sembra essere un valore anomalo. Senza questo caso la media di N scende da 3.29 a 3.17. Quindi questo caso è particolarmente influente. Se invece si elimina il caso con $z = -2.8$ la media rimane 3.29, quindi questo caso non sembra influente.

In presenza di casi anomali univariati che influenzano i risultati delle analisi è possibile utilizzare degli stimatori dei parametri che risultano meno influenzati dalla presenza di tali valori. Ad esempio, la mediana e la moda spesso possono risultare più affidabili della media. Sono inoltre disponibili alcune statistiche che risultano «robuste» alla presenza di tali valori, come ad esempio la media *trimmed* che viene calcolata eliminando il 5% dei casi con punteggi più elevati e più bassi.

(segue)