

Montgomery Runger Hubele

# statistica per ingegneria

Edizione italiana a cura di  
**Matteo Gregoratti**  
**Maurizio Verri**

SECONDA EDIZIONE

Montgomery Rungor Hubele

# statistica ingegneria

Edizione italiana a cura di  
**Matteo Gregoratti**  
**Maurizio Verri**

 Egea

**Titolo originale:**  
*Engineering Statistics, 5<sup>th</sup> edition*  
Copyright © 2011 John Wiley & Sons, Inc.

Per l'edizione in lingua italiana  
Copyright © 2012, 2004 EGEA S.p.A.  
Via Salasco, 5 - 20136 Milano  
Tel. 02-58365751 - Fax 02-58365753  
[egea.edizioni@unibocconi.it](mailto:egea.edizioni@unibocconi.it) - [www.egeaonline.it](http://www.egeaonline.it)

*Traduzione:* Giovanni Malafarina  
*Impaginazione:* Imagine, Trezzo sull'Adda (Mi)  
*Copertina:* mStudio, Milano  
*Stampa:* Mediascan, Milano

Tutti i diritti sono riservati, compresi la traduzione,  
l'adattamento totale o parziale, la riproduzione,  
la comunicazione al pubblico e la messa a disposizione  
con qualsiasi mezzo e/o su qualunque supporto  
(ivi compresi i microfilm, i film, le fotocopie, i supporti  
elettronici o digitali), nonché la memorizzazione elettronica  
e qualsiasi sistema di immagazzinamento e recupero  
di informazioni.

Per altre informazioni o richieste di riproduzione  
si veda il sito [www.egeaonline.it/fotocopie.htm](http://www.egeaonline.it/fotocopie.htm)

Date le caratteristiche di Internet, l'Editore non è responsabile  
per eventuali variazioni di indirizzi e contenuti dei siti  
Internet menzionati.

ISBN 978-88-238-2149-1  
Prima edizione italiana: gennaio 2004  
Seconda edizione italiana: febbraio 2012



*Questo volume è stampato su carta FSC® proveniente da foreste gestite  
in maniera responsabile secondo rigorosi standard ambientali, economici  
e sociali definiti dal Forest Stewardship Council®.*

# Prefazione

---

Gli ingegneri rivestono un ruolo significativo nel mondo moderno. Innanzitutto, sono responsabili del progetto e dello sviluppo della maggior parte dei prodotti utilizzati dalla nostra società, nonché dei relativi processi di produzione. Gli ingegneri sono coinvolti anche in molti aspetti manageriali, tanto nelle industrie quanto nelle imprese e nelle organizzazioni del terziario. La preparazione di base in ingegneria, infatti, sviluppa capacità di formulazione, analisi e risoluzione dei problemi spendibili in un'ampia gamma di situazioni pratiche.

La risoluzione di molti tipi di problemi ingegneristici richiede una capacità di valutazione della variabilità e una certa comprensione di come si usano gli strumenti descrittivi e analitici per trattare tale variabilità. La statistica è la branca della matematica applicata che riguarda appunto la variabilità e il suo impatto sui processi decisionali. Il presente manuale è un testo introduttivo per un primo corso in statistica per l'ingegneria; benché molti temi qui presentati siano essenziali per l'applicazione della statistica anche ad altre discipline, abbiamo scelto di concentrarci sulle esigenze proprie degli studenti di ingegneria, presentando le applicazioni pratiche da questo punto di vista. Di conseguenza, i nostri esempi ed esercizi sono riferiti a casi ingegneristici; in quasi tutti abbiamo usato l'impostazione del problema reale oppure i dati ricavati da fonti pubblicate o dalla nostra stessa esperienza di consulenti.

Gli ingegneri di tutte le specializzazioni dovrebbero seguire almeno un corso di statistica. In effetti, la *Accreditation Board on Engineering and Technology* richiede che la statistica e l'uso efficace dei metodi statistici facciano necessariamente parte della formazione degli ingegneri. Questo libro è stato pensato come testo di riferimento per un corso di statistica semestrale, indipendentemente dalla particolare specializzazione in Ingegneria.

La quinta edizione è stata ampiamente revisionata; comprende alcuni nuovi esempi e molti nuovi problemi. Nella revisione abbiamo voluto riscrivere quegli argomenti che, in base alla nostra esperienza di docenti o al feedback ricevuto da altri, si sono rivelati più ostici per gli studenti.

## ORGANIZZAZIONE DEL LIBRO

Il volume è basato su un testo di argomento più generale (Montgomery D.C., Runger G.C., *Applied Statistics and Probability for Engineers*, Fifth Edition, John Wiley & Sons, New York, 2011) che è stato utilizzato dai docenti in un corso di uno o due semestri. Da esso abbiamo estrapolato e posto a base di questo volume gli argomenti fondamentali necessari per un corso di un semestre. Dal lavoro di riduzione e revisione è risultato un libro che richiede un modesto livello di conoscenze di matematica; in particolare, gli studenti di Ingegneria che hanno completato un semestre di Analisi non avranno difficoltà nella lettura di pressoché tutto il testo. Il nostro scopo è di fornire allo studente una comprensione della metodologia statistica e delle possibilità di applicazione alla risoluzione dei problemi ingegneristici, piuttosto che la teoria matematica della statistica. Le note in colonnino sono d'aiuto allo studente in questa attività di comprensione. Abbiamo avuto cura, in tutto il libro, di evidenziare quanto l'approccio statistico sia una parte cruciale del processo di risoluzione dei problemi.

Il Capitolo 1 presenta il ruolo della statistica e della probabilità nella risoluzione dei problemi di ingegneria. Vengono illustrati l'approccio e i metodi della statistica, ponendoli a confronto con altri tipi di approccio per la costruzione e l'utilizzo di modelli nel contesto del *problem solving* in ingegneria. Tramite la presentazione di alcuni semplici esempi viene discusso nei punti essenziali il valore delle metodologie statistiche. Vengono anche introdotte le più semplici statistiche riassuntive.

Il Capitolo 2 illustra le utili informazioni ricavabili da semplici sintesi numeriche e visualizzazioni grafiche. Vengono date le procedure, da eseguirsi al computer, per l'analisi di insiemi di dati numerosi. Sono illustrati metodi di analisi come gli istogrammi, i grafici rami e foglie e le distribuzioni di frequenze. In questo capitolo viene posto l'accento sulla possibilità di usare tali visualizzazioni per ottenere informazioni sul comportamento dei dati o del sistema.

Il Capitolo 3 presenta le variabili aleatorie e le distribuzioni di probabilità atte a descrivere il comportamento. Introduciamo in queste pagine una semplice procedura a tre passi utile per impostare la risoluzione di un problema probabilistico. Ci si concentra sulla distribuzione normale per via del ruolo fondamentale che questa riveste negli strumenti statistici più frequentemente applicati in ingegneria. Abbiamo cercato di evitare l'uso di matematica sofisticata, così come l'approccio dello spazio degli eventi tradizionalmente adottato per presentare questi argomenti agli studenti di ingegneria. Per capire come usare la statistica per l'efficace risoluzione dei problemi ingegneristici non è necessaria una comprensione approfondita del concetto di probabilità. Tra gli altri argomenti di questo capitolo vi sono i valori attesi, le varianze, i grafici dei quantili e il teorema limite centrale.

I Capitoli 4 e 5 introducono gli strumenti di base dell'inferenza statistica: stima puntuale, intervalli di confidenza, verifica delle ipotesi. Nel Capitolo 4 si trovano le tecniche di inferenza relative a un singolo campione, nel Capitolo 5 quelle relative a due campioni. La nostra presentazione è decisamente orientata alle applicazioni, ed enfatizza il fatto che le procedure statistiche sono per loro natura legate al continuo confronto con l'esperimento. Desideriamo che gli studenti di ingegneria comprendano come questi metodi possono venire impiegati per risolvere problemi applicativi, e come utilizzarli in altre situazioni. Delle tecniche forniamo un'esposizione naturale, euristica, piuttosto che una rigorosamente matematica. In questa nuova edizione ci siamo concentrati maggiormente sull'approccio tramite *P*-value alle verifiche di ipotesi, sia perché è abbastanza semplice da comprendere, sia per-

ché rispecchia le modalità con cui i moderni software statistici presentano i risultati delle elaborazioni.

Nel Capitolo 6 è trattata la costruzione dei modelli empirici. Vengono illustrati sia il modello di regressione lineare semplice, sia quello di regressione multipla, e viene discussa l'uso di tali modelli come approssimazioni dei modelli meccanicistici. Mostriamo allo studente come trovare la stima dei minimi quadrati dei coefficienti di regressione, eseguire i test statistici standard e gli intervalli di confidenza e usare i residui dei modelli nella valutazione dell'adeguatezza del modello. In tutto il capitolo si sottolinea il ruolo e l'utilità del computer nell'adattamento e nell'analisi del modello di regressione.

Gli studenti dovrebbero essere incoraggiati a svolgere i problemi in modo da arrivare a padroneggiare meglio la materia. A tale scopo, il libro contiene molti problemi di diversi livelli di difficoltà. Gli esercizi relativi ai singoli paragrafi hanno lo scopo di consolidare i concetti e le tecniche presentate in ciascun paragrafo. Si tratta di esercizi più strutturati di quelli di fine capitolo, che richiedono in genere una maggiore capacità di formulazione e di astrazione e che vengono proposti come problemi di integrazione per rafforzare la padronanza dei concetti teorici, anziché la tecnica analitica. L'uso dei software statistici nella risoluzione dei problemi dovrebbe essere parte integrante del corso.

## IMPIEGO DEL LIBRO

È nostra ferma convinzione che un corso introduttivo di statistica per il Corso di laurea in Ingegneria debba essere, soprattutto e in primo luogo, un *corso applicativo*. L'accento dovrebbe essere posto innanzitutto sulla descrizione dei dati, sull'inferenza (intervalli di confidenza e test) e sulla costruzione dei modelli, *perché queste sono le tecniche che gli studenti dovranno saper impiegare nel mondo del lavoro*. In genere si tende a insegnare questi argomenti soffermandosi a lungo sui concetti di probabilità e di variabili aleatorie (e, in effetti, alcune figure di ingegnere, come gli ingegneri industriali ed elettrici, hanno bisogno di conoscere questi argomenti più approfonditamente che non gli studenti di altre specializzazioni). Ciò può portare il corso di statistica per ingegneria a diventare una sorta di corso di “baby math-stat”. Questo tipo di corso risulta quasi sempre più semplice e divertente da insegnare, dato che è sempre più facile insegnare la teoria che non la pratica, ma non prepara adeguatamente gli studenti alla professione.

Nel corso da noi tenuto all'Arizona State University, gli studenti si riuniscono due volte alla settimana, una in aula e una in laboratorio. Gli studenti sono tenuti a uno studio personale, a risolvere individualmente dei problemi assegnati per casa e a svolgere dei progetti di gruppo. Tra le attività condotte in classe vi sono la pianificazione degli esperimenti, la generazione dei dati e l'effettuazione delle analisi. Rispetto a queste attività, gli esercizi proposti in questo libro possono costituire una buona fonte di spunti. L'intento è di fornire un ambiente di apprendimento attivo, presentando allo studente problemi in grado di sviluppare le capacità di analisi e di sintesi.

## USO DEL COMPUTER

Nella pratica, gli ingegneri usano i computer per applicare i metodi statistici alla risoluzione dei problemi, pertanto raccomandiamo fortemente l'impiego dei software statistici. In

questo libro abbiamo presentato gli output ricavati da Minitab come esempi tipici di ciò che si può ottenere con i più recenti programmi per computer. Nell'attività didattica abbiamo usato Statgraphics, Minitab, Excel e molti altri pacchetti software o fogli elettronici. Non abbiamo inserito nel testo esempi di diversi programmi perché *come* il docente integra il software nelle proprie lezioni è in ultima analisi più importante di *quale* pacchetto viene utilizzato.

Negli incontri in aula con gli studenti, noi abbiamo accesso al software che verrà impiegato in laboratorio; possiamo dunque mostrare alla platea come viene implementata la tecnica in parallelo alla presentazione teorica di quest'ultima. Ci sentiamo di consigliare questo tipo di approccio, che non dovrebbe presentare particolari problemi pratici, considerando il fatto che dei software più popolari sono facilmente reperibili versioni a prezzo ridotto per gli studenti, o che sulla rete locale dell'università è spesso disponibile almeno un software statistico.

 Per la risoluzione di molti degli esercizi proposti si possono usare software statistici. Per alcuni esercizi in particolare, tuttavia, contrassegnati dall'apposita icona, è caldamente consigliato il ricorso al computer.

 L'icona a lato indica che per l'esercizio sono fornite statistiche riassuntive; i dati completi si trovano all'indirizzo [www.egeaonline.it/stating.htm](http://www.egeaonline.it/stating.htm). Alcuni docenti potrebbero scegliere di fare usare agli studenti i dati completi anziché le statistiche di sintesi.

## RINGRAZIAMENTI

Vorremmo esprimere la nostra gratitudine verso chi ha contribuito a sviluppare parte del materiale utilizzato in questo testo a partire dal *Course and Curriculum Development Program* della *Undergraduate Education Division* della *National Science Foundation*. Siamo grati al dottor Dale Kennedy e alla dottorella Mary Anderson-Rowland per i preziosi suggerimenti e i generosi commenti utili all'insegnamento del nostro corso presso la Arizona State University. Ringraziamo anche il dottor Teri Reed Rhoads, della Purdue University, la dottorella Lora Zimmer e la dottorella Sharon Lewis per il lavoro svolto in qualità di assistenti nella realizzazione del corso basato su questo testo. Siamo debitori verso Busaba Laungrungrong, la dottorella Connie Borror per il lavoro svolto sul volume dedicato ai docenti e nei confronti della dottorella Sarah Street, del dottor James C. Ford, del dottor Craig Downing e di Patrick Egbunonu per averci aiutato a verificare l'accuratezza e la completezza del testo, delle soluzioni e degli apparati a corredo.

Ci siamo giovati del supporto dello staff e delle risorse fornite dal programma *Industrial Engineering* presso l'Arizona State University, e del nostro direttore dott. Ronald Askin.

Molti revisori hanno fornito il loro contributo: il dottor Thomas Willemain, Rensselaer, il dottor Hongshik Ahn, SUNY, Stony Brook; il dottor James Simpson, Florida State/FAMU; il dottor John D. O'Neil, California Polytechnic University, Pomona; il dottor Charles Donaghay, University of Houston; il professor Gus Greivel, Colorado School of Mines; il professor Arthur M. Sterling, LSU; il professor David Powers, Clarkson University; il dottor William Warde, Oklahoma State University; il dottor David Mathiason.

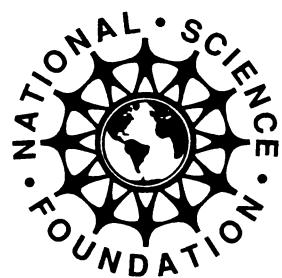
Siamo inoltre debitori verso il dottor Smiley Cheng della University of Manitoba per averci concesso il permesso di adattare molte delle tavole statistiche pubblicate nel suo eccellente libro (assieme al dottor James Fu), *Statistical Tables for Classroom and Exam Room*. John Wiley & Sons, Prentice-Hall, verso la Biometrika Trustees, verso la American

Statistical Association, l'Institute of Mathematical Statistics; i curatori di *Biometrics* ci hanno consentito l'uso di materiale protetto dai diritti d'autore, e per questo siamo loro grati.

Questo progetto è stato realizzato grazie anche al contributo della National Science Foundation.

Le opinioni espresse sono quelle degli autori e non necessariamente quelle della fondazione.

*Douglas C. Montgomery*  
*George C. Runger*  
*Norma Faris Hubele*





# Presentazione della prima edizione italiana

---

Questo libro, rivolto a un pubblico di allievi ingegneri che hanno seguito un primo corso universitario di *Calculus*, si propone di offrire un'introduzione alla statistica con l'obiettivo principale di insegnare allo studente come si affrontano problemi applicativi con l'uso di metodi statistici sia descrittivi sia analitici. Pur rimandando il lettore alla *Prefazione* degli Autori per avere un'esposizione più completa delle motivazioni di merito, ci preme sottolineare che lo scopo dichiarato è far acquisire allo studente sensibilità e autonomia nello *statistical thinking*. La strategia didattica adottata è di presentare concetti e tecniche tramite esempi significativi, la cui valenza generale viene di volta in volta enfatizzata. Gli aspetti di rigore formale nell'esposizione dei fondamenti di probabilità e delle tecniche statistiche di base sono volutamente messi in secondo piano, ma questo perché si possa arrivare ad affrontare lo studio di argomenti relativamente avanzati anche in un corso di breve durata (per esempio di sole trenta ore di lezione ovvero di cinque crediti formativi). Aggiungiamo anche che questo libro è ricco di interessanti osservazioni su come applicare le varie tecniche statistiche e di avvertenze che mettono in guardia l'allievo da un uso acritico di formule e procedure.

Nella sua versione originale in inglese, questo libro, pensato per le specifiche esigenze dei futuri ingegneri, è stato introdotto al Politecnico di Milano dal professor Piercesare Secchi ed è stato poi sperimentato con successo anche da altri colleghi e da noi stessi in diversi corsi di statistica tenuti nell'ambito delle lauree triennali in Ingegneria Meccanica, Energetica, Biomedica, Fisica e Matematica. La nostra esperienza ci dice che i contenuti del libro sono in realtà sovradimensionati rispetto a un corso da cinque crediti, ma questo fatto offre semplicemente l'opportunità al docente di scegliere tra gli argomenti proposti quelli che ritiene più interessanti o adatti al pubblico che ha di fronte.

Per rendere più snella l'edizione italiana rispetto a quella originale abbiamo eliminato i due capitoli finali riguardanti il piano degli esperimenti e il controllo statistico dei processi, due argomenti che per tradizione in Italia sono erogati da insegnamenti più propriamente di ingegneria.

*Alberto Barchielli  
Maurizio Verri*



# Presentazione della seconda edizione italiana

---

Nelle ultime edizioni di “Engineering Statistics”, pubblicate dopo la prima edizione italiana del libro, Montgomery, Runger e Hubele hanno introdotto numerose novità, tutte volte a migliorarne l’efficacia didattica. Ci è sembrato quindi che la proposta di EGEA di curare un aggiornamento anche dell’edizione italiana fosse utile e opportuna per mettere tali novità a disposizioni dei nostri studenti.

Fra le principali novità segnaliamo le introduzioni dei singoli capitoli (che servono a mostrare la rilevanza ingegneristica degli argomenti trattati), le note a margine (per meglio guidare lo studente a interpretare e capire la statistica), la maggiore enfasi data negli esempi alle implicazioni pratiche delle conclusioni statistiche, la nuova presentazione della teoria dei test statistici e del p-value, l’aggiunta di numerosi esercizi. In particolare, molti di questi ultimi sono esplicitamente basati sull’utilizzo del calcolatore al fine di avviare lo studente all’uso di moderni software statistici ogniqualvolta la pratica ingegneristica richieda l’ausilio della Statistica.

*Matteo Gregoratti  
Maurizio Verri*



# Indice

---

## CAPITOLO 1 Il ruolo della statistica in ingegneria 1

---

- 1.1 Il metodo dell'ingegneria e l'approccio statistico 2
- 1.2 Raccolta dei dati in ingegneria 6
  - 1.2.1 Studi retrospettivi 8
  - 1.2.2 Studi osservativi 9
  - 1.2.3 Esperimenti pianificati 10
  - 1.2.4 Campioni casuali 13
- 1.3 Modelli meccanicistici e modelli empirici 16
- 1.4 Osservazione dei processi nel tempo 19

## CAPITOLO 2 Sintesi numerica e presentazione grafica dei dati 25

---

- 2.1 Visualizzazione e sintesi numerica dei dati statistici 27
- 2.2 Diagrammi rami e foglie 31
- 2.3 Istogrammi 36
- 2.4 Box plot 41
- 2.5 Grafici delle serie storiche 43
- 2.6 Dati multivariati 45

## CAPITOLO 3 Variabili aleatorie e distribuzioni di probabilità 59

---

- 3.1 Introduzione 61
- 3.2 Variabili aleatorie 63
- 3.3 Probabilità 64
- 3.4 Variabili aleatorie continue 68
  - 3.4.1 Funzione di densità di probabilità 68
  - 3.4.2 Funzione di distribuzione cumulativa 71
  - 3.4.3 Media e varianza 73
- 3.5 Principali distribuzioni continue 75
  - 3.5.1 Distribuzione normale 75
  - 3.5.2 Distribuzione logonormale 85
  - 3.5.3 Distribuzione gamma 87
  - 3.5.4 Distribuzione di Weibull 88
  - 3.5.5 Distribuzione Beta 90
- 3.6 Grafici dei quantili 92
  - 3.6.1 Grafici dei quantili normali 92
  - 3.6.2 Altri grafici dei quantili 94
- 3.7 Variabili aleatorie discrete 95
  - 3.7.1 Funzione di massa di probabilità 96

- 3.7.2 Funzione di distribuzione cumulativa 97
- 3.7.3 Media e varianza 98
- 3.8 Distribuzione binomiale 99
- 3.9 Processo di Poisson 104
  - 3.9.1 Distribuzione di Poisson 105
  - 3.9.2 Distribuzione esponenziale 108
- 3.10 Approssimazione normale delle distribuzioni binomiale e di Poisson 112
- 3.11 Più variabili aleatorie e indipendenza 116
  - 3.11.1 Distribuzioni congiunte 116
  - 3.11.2 Indipendenza 117
- 3.12 Funzioni di variabili aleatorie 121
  - 3.12.1 Combinazioni lineari di variabili aleatorie indipendenti 122
  - 3.12.2 Combinazioni lineari di variabili aleatorie non indipendenti 123
  - 3.12.3 Funzioni non lineari di variabili aleatorie indipendenti 125
- 3.13 Campioni casuali, statistiche e teorema limite centrale 128

## CAPITOLO 4 Processo decisionale per un singolo campione 147

---

- 4.1 Inferenza statistica 149
- 4.2 Stima puntuale 150
- 4.3 Verifica di ipotesi 155
  - 4.3.1 Ipotesi statistiche 155
  - 4.3.2 Verifica delle ipotesi statistiche 157
  - 4.3.3 Il P-value nella verifica di ipotesi 164
  - 4.3.4 Ipotesi unilaterali e bilaterali 166
  - 4.3.5 Procedura generale per la verifica di ipotesi 168
- 4.4 Inferenza sulla media di una popolazione con varianza nota 168
  - 4.4.1 Verifica di ipotesi sulla media 169
  - 4.4.2 Errore del II tipo e scelta della dimensione campionaria 173
  - 4.4.3 Test con campioni numerosi 177
  - 4.4.4 Considerazioni pratiche sulla verifica di ipotesi 177
  - 4.4.5 Intervallo di confidenza per la media 179
  - 4.4.6 Metodo generale per ricavare un intervallo di confidenza 185
- 4.5 Inferenza sulla media di una popolazione con varianza incognita 186
  - 4.5.1 Verifica di ipotesi sulla media 186
  - 4.5.2 Errore del II tipo e scelta della dimensione campionaria 193
  - 4.5.3 Intervallo di confidenza per la media 195
- 4.6 Inferenza sulla varianza di una popolazione normale 197
  - 4.6.1 Verifica di ipotesi sulla varianza di una popolazione normale 197
  - 4.6.2 Intervallo di confidenza per la varianza di una popolazione normale 201
- 4.7 Inferenza sulla proporzione di una popolazione 202
  - 4.7.1 Verifica di ipotesi su una proporzione binomiale 203
  - 4.7.2 Errore del II tipo e scelta della dimensione campionaria 206
  - 4.7.3 Intervallo di confidenza per una proporzione binomiale 208
- 4.8 Altre stime intervallari per un singolo campione 212
  - 4.8.1 Intervallo di predizione 212
  - 4.8.2 Intervalli di tolleranza per una distribuzione normale 214
- 4.9 Tabelle riassuntive delle procedure di inferenza per un singolo campione 215
- 4.10 Test di adattamento 216

**CAPITOLO 5** Processo decisionale per due campioni 231

---

- 5.1 Introduzione 233
- 5.2 Inferenza sulle medie di due popolazioni con varianze note 233
  - 5.2.1 Verifica di ipotesi sulla differenza tra medie con varianze note 234
  - 5.2.2 Errore del II tipo e scelta della dimensione campionaria 236
  - 5.2.3 Intervallo di confidenza per la differenza tra medie con varianze note 237
- 5.3 Inferenza sulle medie di due popolazioni con varianze incognite 240
  - 5.3.1 Verifica di ipotesi sulla differenza tra medie 240
  - 5.3.2 Errore del II tipo e scelta della dimensione campionaria 247
  - 5.3.3 Intervallo di confidenza per la differenza tra medie 248
- 5.4 Test  $t$  accoppiato 251
- 5.5 Inferenza sul rapporto tra le varianze di due popolazioni normali 256
  - 5.5.1 Verifica di ipotesi sul rapporto tra due varianze 256
  - 5.5.2 Intervallo di confidenza per il rapporto tra due varianze 261
- 5.6 Inferenza sulle proporzioni di due popolazioni 262
  - 5.6.1 Verifica di ipotesi sull'uguaglianza di due proporzioni binomiali 262
  - 5.6.2 Errore del II tipo e scelta della dimensione campionaria 265
  - 5.6.3 Intervallo di confidenza per la differenza tra proporzioni binomiali 266
- 5.7 Tabelle riassuntive delle procedure di inferenza per due campioni 268
- 5.8 Caso di più di due campioni 268
  - 5.8.1 Esperimento completamente casualizzato e analisi della varianza 268
  - 5.8.2 Esperimento a blocchi completi casualizzati 280

**CAPITOLO 6** Costruzione di modelli empirici 299

---

- 6.1 Introduzione ai modelli empirici 300
- 6.2 Regressione lineare semplice 306
  - 6.2.1 Stima dei minimi quadrati 306
  - 6.2.2 Verifica delle ipotesi nella regressione lineare semplice 314
  - 6.2.3 Intervalli di confidenza nella regressione lineare semplice 318
  - 6.2.4 Predizione di nuove osservazioni 321
  - 6.2.5 Controllo dell'adeguatezza del modello 323
  - 6.2.6 Correlazione e regressione 326
- 6.3 Regressione multipla 327
  - 6.3.1 Stima dei parametri nella regressione multipla 327
  - 6.3.2 Inferenze nella regressione multipla 333
  - 6.3.3 Controllo dell'adeguatezza del modello 339
- 6.4 Altri aspetti della regressione 344
  - 6.4.1 Modelli polinomiali 344
  - 6.4.2 Regressori categorici 347
  - 6.4.3 Tecniche di selezione delle variabili 350

**APPENDICE A** Tavole e carte statistiche 365

---

**APPENDICE B** Bibliografia ragionata 379

---

**APPENDICE C** Soluzioni di alcuni esercizi 381

---



# Il ruolo della statistica in ingegneria

## GETTARE UN PONTE

Il compito dell'ingegneria è di gettare un ponte che colleghi i problemi alle loro soluzioni, e tale procedimento richiede di adottare l'approccio del **metodo scientifico**.

Nel 2009 Eileen Huffman, una studentessa universitaria di ingegneria civile del politecnico Virginia Tech, ha applicato il metodo scientifico al suo studio di un antico ponte, l'Ironto Wayside Footbridge; costruito nel 1878, è il più vecchio ponte metallico della Virginia ancora in piedi. Anche se oggi è stato restaurato ed è aperto solo al passaggio pedonale, nella sua vita precedente aveva retto il traffico quotidiano di mezzi per il trasporto di beni e materiali (tre tonnellate o più). Eileen Huffman ha compiuto su di esso studi storici, scoprendo che non era mai stata fatta un'analisi dei carichi. Il suo problema era come realizzare la prima analisi di questo tipo.

Dopo aver raccolto i dati strutturali del ponte disponibili, ha creato un modello computazionale per l'analisi delle sollecitazioni, basandosi sui carichi tipici che si presume il ponte abbia retto in passato. Dopo aver analizzato i risultati ottenuti, li ha sottoposti a test direttamente sulla struttura, per verificare la validità del proprio modello. A tale scopo, ha inserito degli indicatori a quadrante sotto al punto centrale di ogni travatura, quindi ha fatto passare sul ponte un furgone da 3 tonnellate, un carico rappresentativo dei carichi retti in passato dalla struttura.

I risultati di questo test contribuiranno all'Adaptive Bridge Use Project che ha sede presso la University of Massachusetts Amherst, e che ha il supporto della National Science Foundation ([www.ecs.umass.edu/adaptive\\_bridge\\_use/](http://www.ecs.umass.edu/adaptive_bridge_use/)). Gli esiti e le conclusioni di Eileen Huffman saranno utili per la conservazione del ponte e aiuteranno altri a restaurare e studiare i ponti storici. La sua relatrice Cris Moen sottolinea inoltre che il modello computazionale di Huffman può essere impiegato per creare modelli strutturali da applicare al test di altri ponti.

Lo studio di Huffman riflette un uso rigoroso del metodo scientifico nel contesto di un progetto di ingegneria, ed è un eccellente esempio di come si debbano usare i dati campionari per verificare un modello ingegneristico.

## CONTENUTI DEL CAPITOLO

- |  |   |
|--|---|
| <p>1.1 IL METODO DELL'INGEGNERIA<br/>E L'APPROCCIO STATISTICO</p> <p>1.2 RACCOLTA DEI DATI IN INGEGNERIA</p> <ul style="list-style-type: none"><li>1.2.1 Studi retrospettivi</li><li>1.2.2 Studi osservativi</li><li>1.2.3 Pianificazione degli esperimenti</li><li>1.2.4 Campioni casuali</li></ul> | <p>1.3 MODELLI MECCANICISTICI<br/>E MODELLI EMPIRICI</p> <p>1.4 OSSERVAZIONE DEI PROCESSI<br/>NEL TEMPO</p> |
|--|---|
- 

## OBIETTIVI DI APPRENDIMENTO

Dopo uno studio attento di questo capitolo si sarà in grado di:

1. identificare il ruolo della statistica nel processo di risoluzione di problemi ingegneristici
  2. comprendere la variabilità che caratterizza i dati raccolti e utilizzati nei processi decisionali
  3. discutere i metodi usati dagli ingegneri per raccogliere dati
  4. spiegare l'importanza del campionamento casuale
  5. individuare i vantaggi della pianificazione degli esperimenti nella raccolta dei dati
  6. spiegare la differenza fra modelli meccanicistici e modelli empirici
  7. spiegare la differenza fra studi enumerativi e studi analitici
- 

### 1.1 IL METODO DELL'INGEGNERIA E L'APPROCCIO STATISTICO

Gli ingegneri risolvono problemi di interesse per la società mediante l'efficace applicazione di principi scientifici. Il **metodo scientifico o dell'ingegneria** consiste proprio nell'approccio alla formulazione e alla risoluzione di tali problemi. I passi necessari a sviluppare questo approccio sono i seguenti.

1. Sviluppare una descrizione chiara e concisa del problema.
2. Identificare, almeno provvisoriamente, i principali fattori che influenzano il problema o che possono svolgere un ruolo nella sua risoluzione.
3. Proporre un modello che descriva il problema, utilizzando conoscenze scientifiche o tecniche del fenomeno in esame. Esplicitare ogni limitazione o ipotesi insita nel modello.
4. Condurre opportuni esperimenti e raccogliere dati per convalidare il modello provvisorio o le conclusioni dei punti 2 e 3.
5. Raffinare il modello sulla base dei dati osservati.
6. Elaborare il modello in modo da agevolare la risoluzione del problema.

7. Condurre un opportuno esperimento che confermi l'efficacia della soluzione proposta per il problema.
8. Trarre conclusioni o fare raccomandazioni sulla base della soluzione al problema identificata.

I passi del metodo dell'ingegneria sono mostrati in Figura 1.1. Si noti che questo metodo presenta una forte interazione fra il problema, i fattori che ne possono influenzare la soluzione, il modello del fenomeno e la sperimentazione atta a verificare l'adeguatezza del modello e della soluzione proposta. In Figura 1.1 i passi 2-4 sono racchiusi entro una cornice, a indicare che per ottenere la soluzione finale del problema possono essere necessari diversi cicli o iterazioni di tali passi. Di conseguenza, gli ingegneri devono sapere pianificare efficacemente gli esperimenti, raccogliere, analizzare e interpretare i dati, e comprendere la relazione fra i dati osservati e il modello proposto per il problema in esame.

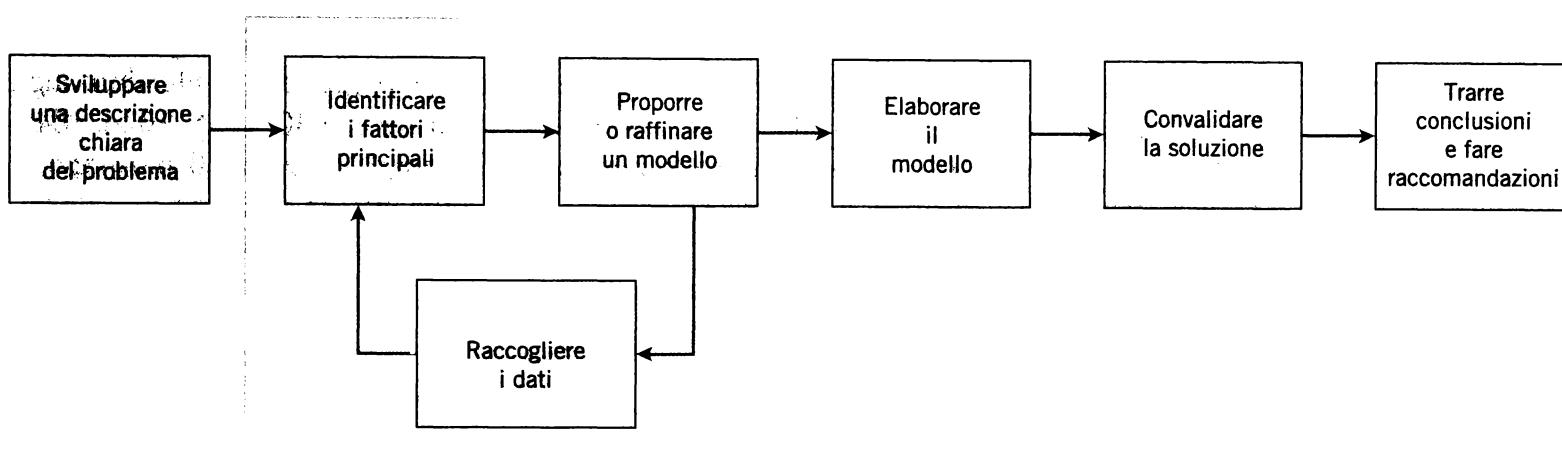


Figura 1.1 Il metodo ingegneristico di risoluzione dei problemi.

La **statistica** si occupa della raccolta, della presentazione, dell'analisi e dell'impiego dei dati ai fini dei processi decisionali e della risoluzione dei problemi.

### Definizione

**La statistica è la scienza che si occupa dei dati.**

Nel metodo di risoluzione dei problemi adottato dall'ingegneria sono coinvolte molte branche dell'ingegneria:

- la meccanica, in particolare la statica e la dinamica
- la scienza dei fluidi
- la termodinamica e la trasmissione del calore
- le scienze elettriche
- la scienza dei materiali
- le scienze chimiche

Poiché per molti aspetti della pratica ingegneristica si deve lavorare con dati, è naturale che una certa conoscenza della statistica sia importante per ogni ingegnere. In particolare, le tecniche della statistica possono rappresentare un valido apporto nella progettazione di nuovi prodotti e sistemi, nel miglioramento di progetti esistenti e nella progettazione, sviluppo e miglioramento dei processi produttivi.

I metodi della statistica vengono utilizzati per aiutarci a descrivere e comprendere la **variabilità**. Con il termine “variabilità” si intende il fatto che osservazioni successive di un sistema o di un fenomeno non producono esattamente lo stesso risultato. Tutti noi incontriamo variabilità nella vita quotidiana, e l'**approccio statistico** può costituire un utile modo di incorporare tale variabilità nei nostri processi decisionali. Per esempio, considerate la percorrenza della vostra auto con un pieno di benzina. Riuscite a percorrere sempre lo stesso numero di chilometri con un pieno? Naturalmente no. In effetti, tale distanza varia a volte in maniera sensibile. Questa variabilità osservata dipende da molti fattori, come il tipo di guida

(guida in città oppure in autostrada), le condizioni del veicolo (pressione delle gomme, rapporto di compressione del motore, usura delle valvole ecc.), la marca e/o il numero di ottani del carburante utilizzato, e perfino le condizioni meteorologiche. Tutti questi fattori rappresentano possibili **cause di variabilità** del sistema. La statistica fornisce un quadro generale per descrivere tale variabilità e per apprendere quali potenziali fattori di variabilità sono più importanti, o quali hanno maggiori conseguenze sul rapporto km/litro.

La variabilità si incontra anche nella maggior parte dei tipi di problemi ingegneristici. Per esempio, supponiamo che un ingegnere stia sviluppando un composto gommoso da utilizzare per gli *O-ring*, le guarnizioni circolari in gomma. Tali O-ring sono destinati all'impiego come guarnizioni entro strumenti per l'incisione al plasma utilizzati nell'industria dei semiconduttori, perciò una caratteristica importante che dovranno possedere è la resistenza agli acidi e ad altre sostanze corrosive. L'ingegnere usa il composto gommoso standard per produrre otto O-ring in un laboratorio di sviluppo e misura la resistenza alla trazione di ogni esemplare dopo l'immersione in una soluzione di acido nitrico a 30°C per 25 minuti [si faccia riferimento allo standard D 1414 dell'ASTM (*American Society for Testing and Materials*) e agli standard associati per numerosi interessanti aspetti dei test conducibili sugli O-ring]. Le resistenze misurate (in psi) per le otto guarnizioni sono 1030, 1035, 1020, 1049, 1028, 1026, 1019 e 1010. Come anticipato, non tutti gli esemplari esaminati mostrano la stessa resistenza alla trazione: c'è dunque una **variabilità** nelle misure effettuate.

Dato che le misure presentano variabilità, si dice che la resistenza alla trazione è una **variabile aleatoria**. Un modo conveniente per vedere una variabile aleatoria  $X$  che rappresenta una grandezza misurata è tramite il **modello**

$$X = \mu + \epsilon$$

dove  $\mu$  è una costante ed  $\epsilon$  è un disturbo casuale, o termine di "rumore". La costante rimane la stessa, ma piccole variazioni delle condizioni ambientali e dell'equipaggiamento usato per il test, differenze nei singoli esemplari di O-ring e potenzialmente molti altri fattori modificano il valore di  $\epsilon$ : se nessuna di queste cause di disturbo fosse presente il valore di  $\epsilon$  sarebbe sempre pari a zero, e  $X$  coinciderebbe con la costante  $\mu$ . Tuttavia, nella pratica ingegneristica ciò non accade mai, perciò le misure effettive che si osservano esibiscono sempre variabilità. Quest'ultima deve spesso essere descritta, quantificata e in definitiva – dal momento che la variabilità può pregiudicare gli obiettivi che si vogliono perseguire – *ridotta*.

In Figura 1.2 è mostrato un **diagramma a punti** della resistenza alla trazione. Questo tipo di diagramma visualizza in modo molto efficace un insieme poco numeroso di dati, per esempio fino a circa 20 osservazioni. Esso consente di vedere a colpo d'occhio due importanti caratteristiche dei dati: la **posizione**, o centro, e la **dispersione** o **variabilità**. Quando il numero di osservazioni è piccolo, in genere è difficile riconoscere uno specifico andamento della variabilità, benché il diagramma a punti sia un modo molto conveniente di osservare comportamenti dei dati come gli **outlier** o **valori erratici** (osservazioni che differiscono notevolmente dall'insieme principale dei dati), o come i **cluster** (raggruppamenti di dati che si presentano molto vicini tra loro).

Nella risoluzione dei problemi di ingegneria sorge di frequente la necessità di assumere un approccio statistico. Consideriamo nuovamente il caso dello sviluppo del materiale per gli O-ring. In base al test condotto sugli esemplari iniziali, l'ingegnere sa che la resistenza media alla trazione è 1027.1 psi. Tuttavia, egli ritiene che questo valore possa essere troppo basso per l'applicazione cui sono destinate le guarnizioni, pertanto decide di prendere in

I metodi grafici aiutano a scoprire degli andamenti tipici nei dati.

esame una formula differente di gomma, inserendovi un additivo Teflon. Vengono realizzati otto O-ring con questo composto modificato, quindi gli esemplari sono sottoposti al test con l'acido nitrico descritto in precedenza. I risultati del test per la resistenza alla trazione sono: 1037, 1047, 1066, 1048, 1059, 1073, 1070 e 1040.

In Figura 1.3 sono riportati i diagrammi a punti per entrambi i gruppi di guarnizioni; essi danno l'impressione visiva diretta che l'aggiunta di Teflon al composto abbia portato a un aumento della resistenza alla trazione. Vi sono però da porsi alcune ovvie domande. Per esempio, come sappiamo che un altro insieme di O-ring campione non darà risultati diversi? Un campione di otto O-ring è sufficiente per dare risultati affidabili? Se usiamo i risultati dei test ottenuti per concludere che l'aggiunta di Teflon alla formula del composto aumenterà la resistenza alla trazione dopo l'esposizione all'acido nitrico, quali rischi sono associati a tale decisione? Per esempio, è possibile (o magari probabile) che l'apparente aumento nella resistenza osservato per le guarnizioni modificate sia dovuto unicamente alla variabilità intrinseca del sistema, e che l'ingrediente addizionale (che comporta un aumento dei costi e una maggiore complessità della produzione) non abbia in realtà alcun effetto sulla resistenza alla trazione? L'approccio e la metodologia della statistica possono aiutare a rispondere a queste domande.



Figura 1.2 Diagramma a punti dei dati relativi alla resistenza alla trazione degli O-ring (composto originale).

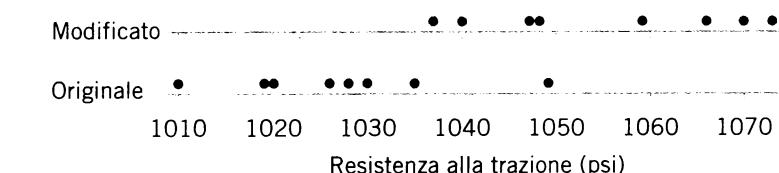


Figura 1.3 Diagramma a punti dei dati relativi alla resistenza alla trazione degli O-ring (composto originale e composto modificato).

Spesso, per aiutare la progettazione di prodotti e processi vengono applicate le leggi fisiche (come la legge di Ohm e quella dei gas perfetti). Abbiamo familiarità con questo approccio che va dalle leggi generali ai casi specifici, ma per rispondere alle precedenti domande è importante anche saper ragionare partendo da specifici insiemi di misure per giungere a casi più generali. Il ragionamento che va da un campione (come gli otto O-ring) a una popolazione (come gli O-ring che verranno venduti alla clientela) viene detto **inferenza statistica** (Figura 1.4). Chiaramente, le deduzioni tratte da misure effettuate su alcuni oggetti e generalizzate a tutti gli oggetti possono dar luogo a errori (detti errori di campionamento). Se però il campione viene accuratamente selezionato, questo rischio può essere quantificato ed è possibile determinare una dimensione campionaria opportuna.

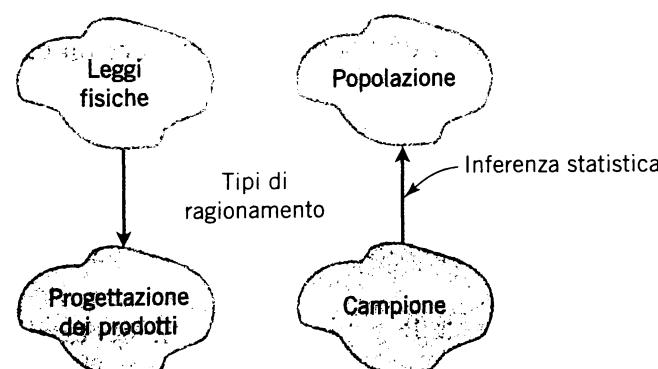


Figura 1.4 L'inferenza statistica è un tipo di ragionamento.

L'inferenza statistica è il processo con cui si decide se le caratteristiche osservate nei dati sono dovute solo al caso.

Spesso, inoltre, gli ingegneri e gli scienziati sono interessati al confronto tra due differenti condizioni, per determinare quale delle due produce un effetto significativo sulla risposta osservata. Tali condizioni vengono chiamate a volte "trattamenti". Il problema della resistenza alla trazione degli O-ring illustra proprio una di queste situazioni; i due diversi trattamenti sono le due formule del composto gommoso, mentre la risposta è la misura della resistenza alla trazione. Scopo dello studio è determinare se la formula modificata dia luogo a un effetto significativo, ossia a un aumento della resistenza alla trazione. Possiamo considerare ogni gruppo di otto O-ring come un campione casuale e rappresentativo di tutti i pezzi che in definitiva saranno prodotti. L'ordine in cui sottoporre al test ciascuna guarnizione è stato anch'esso stabilito in maniera casuale. Si tratta dunque di ciò che viene definito **esperimento pianificato completamente casualizzato**.

Quando in un esperimento casualizzato si osserva una significatività statistica, si può trarre a ragione la conclusione che all'origine delle differenti risposte vi è la diversità dei trattamenti. In altre parole, si può essere sicuri che è stata individuata una relazione causa-effetto.

A volte i soggetti da impiegare nel confronto non vengono assegnati in maniera casuale al trattamento. Per esempio, il numero del settembre 1992 di *Circulation* (una pubblicazione medica dell'American Heart Association) riporta uno studio che collega la presenza nel corpo di alti livelli di ferro a un aumento del rischio di infarto. L'indagine, realizzata in Finlandia, ha seguito 1931 uomini per 5 anni, e ha mostrato un effetto statisticamente significativo dell'innalzamento dei livelli di ferro sull'incidenza degli infarti. In tale studio il confronto non è stato effettuato selezionando casualmente un campione di individui e assegnando alcuni di loro a un trattamento "a basso livello di ferro" e altri a un trattamento "ad alto livello di ferro". I ricercatori hanno semplicemente seguito i soggetti nel corso del tempo. Un'indagine di questo tipo viene definita **studio osservativo**. Gli esperimenti pianificati e gli studi osservativi saranno trattati in maggiore dettaglio nel prossimo paragrafo.

È difficile identificare un rapporto causa-effetto negli studi osservativi, perché la differenza statisticamente significativa nella risposta dei due gruppi può essere dovuta a qualche altro fattore (o insieme di fattori) che non è stato uniformato tramite casualizzazione e che non è dovuto ai trattamenti. Per esempio, la differenza nel rischio di infarto potrebbe essere attribuibile alla differenza tra i livelli di ferro oppure ad altri fattori sottostanti che costituiscono una spiegazione ragionevole dei risultati osservati, come il tasso di colesterolo nel sangue o la presenza di ipertensione.

## 1.2 RACCOLTA DEI DATI IN INGEGNERIA

Nel paragrafo precedente abbiamo illustrato alcuni semplici metodi per riassumere e visualizzare i dati. In ambito ingegneristico i dati sono quasi sempre un **campione** selezionato da qualche **popolazione**.

### Definizione

Una **popolazione** è l'intero insieme di elementi o esiti da cui vengono raccolti i dati.

Un **campione** è un sottoinsieme della popolazione contenente gli elementi osservati o gli esiti e i dati risultanti.

In generale, i dati in ambito ingegneristico vengono raccolti in uno dei seguenti tre modi:

1. uno **studio retrospettivo** basato su dati storici;
2. uno **studio osservativo**;
3. un **esperimento pianificato**.

Una buona procedura di raccolta dati porterà in genere a una semplificazione dell'analisi e garantirà conclusioni più affidabili e più estesamente applicabili. Se non si ragiona abbastanza sulla procedura di raccolta dati si può andare incontro a seri problemi a livello sia di analisi statistica, sia di interpretazione pratica dei risultati.

Montgomery, Peck, Vining (2006) descrivono una colonna di distillazione acetone-butanolo, il cui schema è riportato in Figura 1.5. Useremo questa colonna di distillazione per illustrare le tre modalità di raccolta dei dati in ingegneria identificate poco sopra. Sono tre i fattori che possono influenzare la concentrazione di acetone nel distillato (il prodotto in uscita dalla colonna): la temperatura di riebollizione (controllata dal flusso di vapore), la temperatura del condensato (controllata dal flusso di refrigerante) e la velocità di riflusso. Per la colonna dell'esempio, il personale del reparto produzione registra e archivia quanto segue:

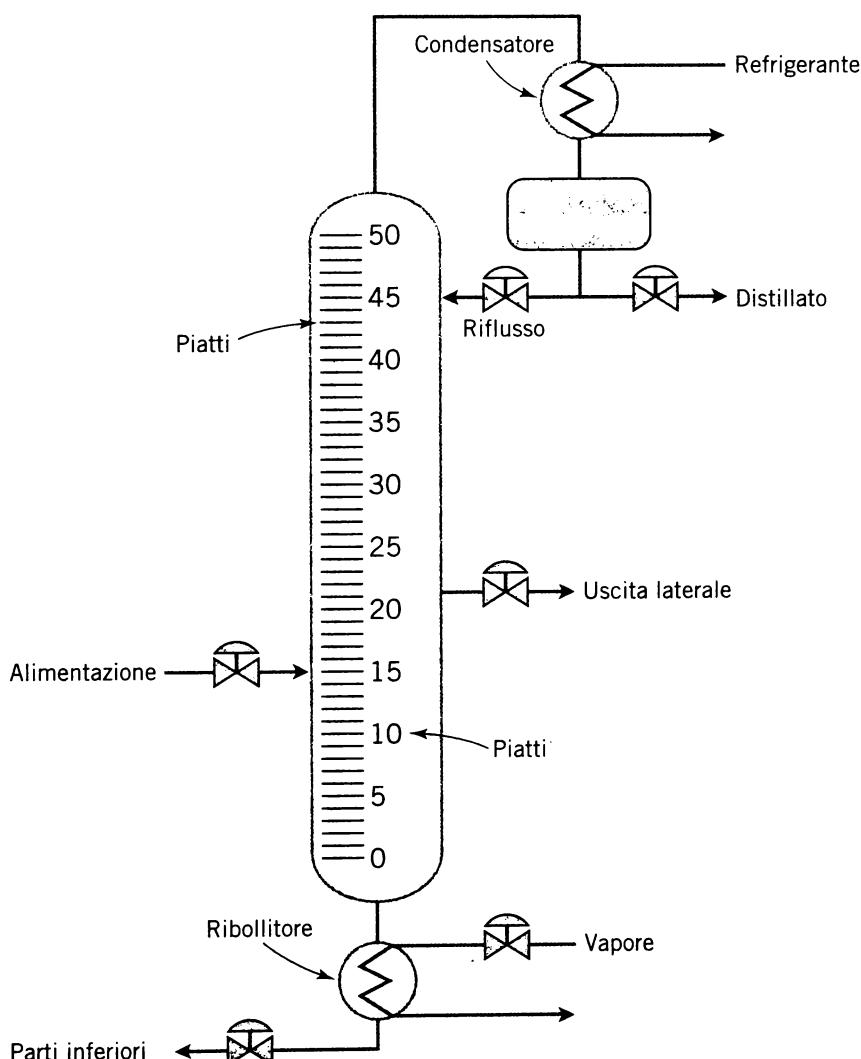


Figura 1.5 Colonna di distillazione acetone-butanolo.

- la concentrazione di acetone in un campione prelevato ogni ora dal distillato;
- le letture della temperatura di riebollizione sul relativo registro dei controlli;
- le letture della temperatura del condensatore sul relativo registro dei controlli;
- la velocità nominale di riflusso, rilevata ogni ora.

Le specifiche del processo richiedono che la velocità nominale di riflusso sia mantenuta costante. Il personale ne modifica molto raramente il valore.

### 1.2.1 Studi retrospettivi

Uno **studio retrospettivo** utilizza tutti i dati **storici** del processo relativi a un dato periodo di tempo, oppure un loro campione. L'obiettivo di uno studio di questo tipo, nel nostro caso, potrebbe essere di determinare le relazioni esistenti fra le due temperature, la velocità di riflusso e la concentrazione di acetone nel distillato in uscita. Nella maggior parte di queste indagini gli ingegneri sono interessati a usare i dati per costruire un **modello** che ponga in relazione le variabili in esame. Per esempio, nel nostro caso il modello potrebbe correlare la concentrazione di acetone (la variabile indipendente) alle tre variabili dipendenti: temperatura di riebollizione, temperatura del condensatore, tasso di riflusso. Questi tipi di modello vengono detti **modelli empirici**, e saranno trattati in maggiore dettaglio nel Paragrafo 1.3.

Uno studio retrospettivo si serve di dati raccolti in precedenza, o dati storici. Ha perciò il vantaggio di minimizzare il costo di una raccolta dei dati per lo studio, ma allo stesso tempo presenta alcuni potenziali problemi:

1. È impossibile isolare l'effetto della velocità di riflusso sulla concentrazione, perché essa, con tutta probabilità, non è variata molto durante il periodo storico.
2. I dati storici sulle due temperature e sulla concentrazione di acetone non sono in corrispondenza diretta. La costruzione di una corrispondenza approssimata richiederebbe probabilmente di assumere parecchie ipotesi e di impegnare un notevole sforzo, senza garanzia di un risultato affidabile.
3. Gli addetti alla produzione mantengono entrambe le temperature il più vicino possibile a specifici valori predefiniti attraverso l'uso di controlli automatici. Dato che tali temperature non variano sensibilmente nel tempo, è molto difficile vedere il loro effettivo impatto sulla concentrazione.
4. Entro i ristretti limiti in cui le temperature variano, quella del condensato tende ad aumentare con quella di riebollizione. Poiché queste due temperature variano insieme, risulta arduo separare i singoli effetti sulla concentrazione di acetone.

Gli studi retrospettivi, benché siano spesso il modo più rapido e semplice di raccogliere dati sui processi ingegneristici, forniscono altrettanto sovente limitate **informazioni** utili per il controllo e l'analisi di un processo. In generale, i loro principali svantaggi sono:

1. mancano sovente alcuni importanti dati del processo;
2. l'affidabilità e la validità dei dati del processo sono spesso discutibili;
3. la natura dei dati del processo a volte può non consentire di inquadrare il problema in esame;

4. gli ingegneri spesso hanno bisogno di utilizzare i dati del processo secondo modalità che non sono state espressamente previste per quei dati;
5. registri, quaderni di appunti e memorie possono non essere in grado di spiegare interessanti fenomeni identificati dall'analisi dei dati.

L'uso di dati storici comporta sempre il rischio che, per qualche ragione, alcuni dei dati essenziali non siano stati raccolti o siano andati persi, o ancora siano stati trascritti o registrati in modo non preciso. Di conseguenza, i dati storici risentono spesso di problemi legati alla qualità dei dati stessi. Tali errori, inoltre, rendono i dati storici inclini a produrre valori erratici. Il solo fatto di essere convenienti da raccogliere non significa che i dati storici siano utili. Spesso vi sono dati ritenuti non essenziali per il monitoraggio di routine del processo e la cui raccolta non è conveniente, ma che nondimeno hanno un impatto significativo sul processo. I dati storici non sono in grado di fornire questo tipo di informazioni se non sono mai stati raccolti dati su qualche variabile essenziale. Per esempio, la temperatura ambientale può influenzare le dispersioni termiche dalla colonna di distillazione: nei giorni freddi la colonna cede più calore all'ambiente che durante i giorni molto caldi. I registri di produzione per la nostra colonna acetone-butanolo non riportano regolarmente la temperatura dell'ambiente. In più, la concentrazione di acetone nel flusso in ingresso si ripercuote sulla concentrazione del flusso in uscita, ma questa variabile, non essendo semplice da misurare, non viene affatto registrata. Pertanto, i dati storici non consentono agli ingegneri di comprendere nell'analisi nessuno dei due fattori appena citati, benché questi ultimi possano essere importanti.

Lo scopo della maggior parte delle analisi dei dati, in ingegneria, è di isolare le cause che stanno alla base dei fenomeni di interesse. Con i dati storici, tali fenomeni possono essersi verificati settimane, mesi o anche anni prima dell'analisi. I registri e gli appunti non sempre fanno luce sulle cause primarie, e i ricordi del personale tecnico coinvolto sono labili. Le analisi basate sui dati storici, perciò, individuano spesso fenomeni interessanti che rimangono senza spiegazione.

Infine, gli studi retrospettivi coinvolgono spesso insiemi di dati molto ampi (in effetti, anche enormi). Gli ingegneri devono padroneggiare saldamente i principi della statistica se vogliono che l'analisi abbia successo.

### 1.2.2 Studi osservativi

Per raccogliere i dati relativi al problema della distillazione è possibile usare anche uno studio osservativo. Come dice il nome, gli **studi osservativi** si limitano all'osservazione del processo o della popolazione durante un periodo di routine operativa. Di solito gli ingegneri interagiscono o disturbano il processo solo nella misura in cui ciò è necessario per ricavare i dati sul sistema, e spesso viene dedicato un particolare sforzo alla raccolta di dati su variabili che non sono registrate secondo routine, se si ritiene che tali dati possano essere utili. Con un'appropriata pianificazione, gli studi osservativi possono assicurare dati completi, accurati e affidabili. D'altro lato, questi studi forniscono sovente informazioni limitate su specifiche relazioni fra le variabili del sistema.

Nell'esempio della colonna di distillazione, gli ingegneri dovrebbero impostare un modulo di raccolta dati che consenta al personale della produzione di registrare le due temperature e l'effettiva velocità di riflusso a determinati istanti, corrispondenti alle osservazioni.

ni della concentrazione di acetone nel flusso in uscita. Il modulo di raccolta dati dovrebbe permettere l'aggiunta di commenti per registrare ogni altro fenomeno di interesse che possa verificarsi, come le variazioni della temperatura dell'ambiente. Durante questo studio a termine relativamente breve si può anche predisporre la misura della concentrazione di acetone nel flusso in ingresso assieme a quella delle altre variabili. Uno studio osservativo condotto in questo modo aiuterebbe ad assicurare una raccolta dati precisa e affidabile, e risolverebbe il problema 2, così come alcuni aspetti del problema 1, associati allo studio retrospettivo. Un approccio di questo tipo, inoltre, minimizza il rischio di riscontrare un valore erratico collegato a qualche errore nei dati. Sfortunatamente, gli studi osservativi non sono in grado di superare i problemi 3 e 4. Vi è infine da tenere conto del fatto che comportano anch'essi la gestione di insiemi molto ampi di dati.

### 1.2.3 Esperimenti pianificati

Il terzo modo di raccogliere i dati in ingegneria è tramite un **esperimento pianificato**, in cui gli ingegneri eseguono deliberate variazioni delle variabili controllabili del sistema (dette **fattori**), osservano l'output risultante del sistema, quindi prendono una decisione o fanno un'inferenza su quali variabili sono responsabili del cambiamento osservato. Un'importante differenza tra un esperimento pianificato e uno studio retrospettivo od osservativo è che le diverse combinazioni dei fattori di interesse sono applicate a caso su un insieme di unità sperimentali. Ciò consente di stabilire relazioni di causa-effetto, cosa che non si può ottenere con gli studi retrospettivi/osservativi.

L'esempio dell'O-ring è un'illustrazione molto semplice di un esperimento pianificato: nella formula del composto è stata introdotta deliberatamente una variazione al fine di scoprire se si potesse ottenere un aumento della resistenza alla trazione. Si tratta in questo caso di un esperimento a un singolo fattore. Possiamo considerare che nei due gruppi di guarnizioni le due formule vengano applicate in maniera casuale ai singoli O-ring. Ciò stabilisce la voluta relazione causa-effetto. L'ingegnere può quindi rispondere alla domanda sulla resistenza alla trazione confrontando le misurazioni di resistenza media per la formula originale con quelle per la formula modificata. Per eseguire questo confronto si possono usare le tecniche statistiche dette **verifica di ipotesi e intervalli di confidenza**; entrambi sono introdotti e illustrati in ampio dettaglio nei Capitoli 4 e 5.

Un esperimento pianificato può venire utilizzato anche nel problema della colonna di distillazione. Supponiamo di avere tre fattori: le due temperature e la velocità di riflusso. Il piano sperimentale deve assicurare la possibilità di separare gli effetti dei tre fattori sulla **variabile risposta**, ossia sulla concentrazione di acetone nel flusso di prodotto in uscita. In un esperimento pianificato, spesso vengono impiegati solo due o tre livelli di ogni fattore. Supponiamo che siano usati due livelli delle temperature e della velocità di riflusso, e che ciascun livello sia codificato come livello  $\pm 1$  (o basso/alto). La migliore strategia sperimentale da usare quando vi sono parecchi fattori di interesse consiste nel condurre un **esperimento fattoriale**, in cui i fattori vengono fatti variare assieme in maniera tale da verificare tutte le possibili combinazioni dei livelli dei fattori.

La Figura 1.6 illustra un esperimento fattoriale per la colonna di distillazione. Poiché tutti e tre i fattori hanno due livelli, vi sono otto possibili combinazioni di livelli dei fattori, mostrati geometricamente come gli otto angoli nel cubo di Figura 1.6a. La rappresentazione tabulare di Figura 1.6b mostra la matrice test per questo esperimento fattoriale; ogni colonna rappresenta

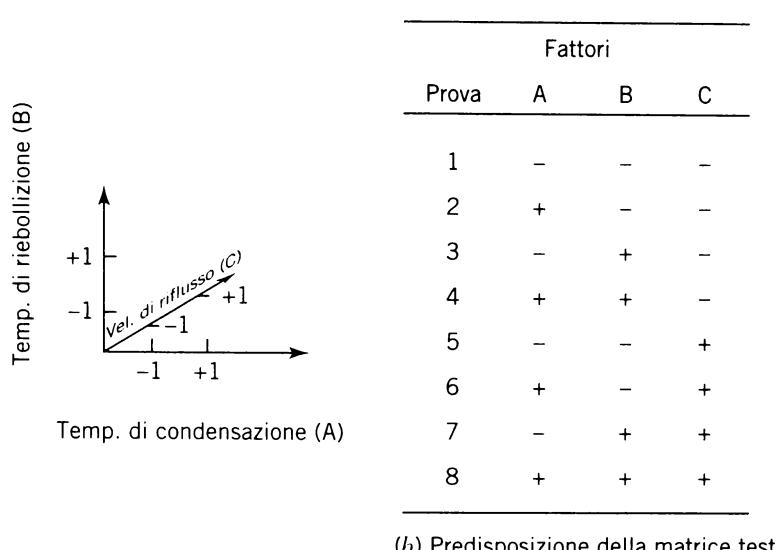
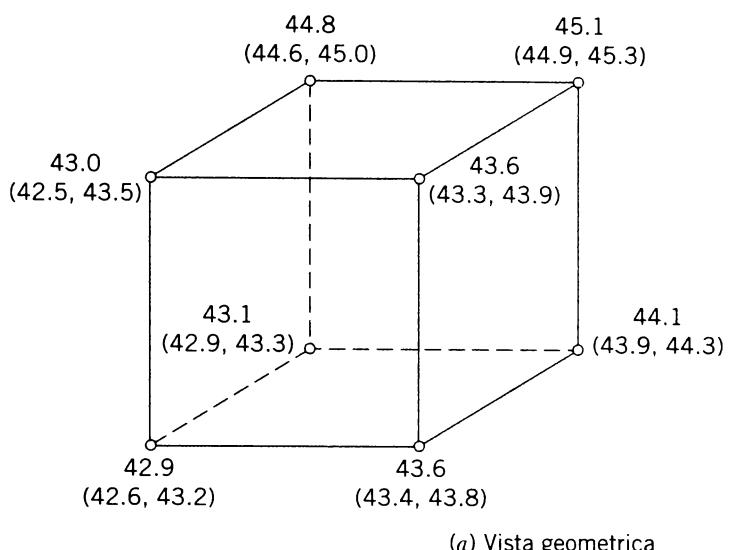


Figura 1.6 Piano fattoriale per la colonna di distillazione.

uno dei tre fattori, e ogni riga corrisponde a una delle otto esecuzioni dell'esperimento. I segni – e + in ciascuna riga indicano le impostazioni basso-alto per i fattori di quella esecuzione. Le esecuzioni sperimentali dovrebbero essere condotte in **ordine casuale**, stabilendo così l'assegnazione casuale alle unità sperimentali delle combinazioni dei livelli dei fattori, che è il principio chiave per un esperimento pianificato. Sono state effettuate due prove, o **repliche**, dell'esperimento (in ordine casuale), dando luogo a sedici esecuzioni ( dette anche **osservazioni**).

Da questo esperimento si possono trarre alcune interessanti conclusioni provvisorie. Innanzitutto, mettiamo a confronto la concentrazione media di acetone per le prime otto osservazioni con temperatura del condensatore a livello alto e la concentrazione media per le otto osservazioni con temperatura del condensatore a livello basso (si tratta delle medie delle otto osservazioni rispettivamente sulle facce di destra e di sinistra del cubo di Figura 1.6a), ossia  $44.1 - 43.45 = 0.65$ . Perciò, incrementando la temperatura del condensatore dal livello basso a quello alto, la concentrazione media aumenta di 0.65 g/l. Poi, per misurare l'effetto di un aumento della velocità di riflusso confrontiamo la media delle otto osservazioni della faccia posteriore del cubo con la media delle otto osservazioni della faccia anteriore, ossia  $44.275 - 43.275 = 1$ . L'effetto di un aumento del tasso di riflusso dal livello basso a quello alto è dunque un aumento della concentrazione di 1 g/l; in altre parole, il riflusso ha apparentemente un effetto maggiore di quello legato alla temperatura del condensatore. L'effetto dovuto alla temperatura di riebolizzazione può essere valutato confrontando la media delle otto osservazioni sulla faccia superiore del cubo con quella delle otto della faccia inferiore, ossia  $44.125 - 43.425 = 0.7$ : al crescere di tale temperatura, la concentrazione sale di 0.7 g/l. Se quindi l'obiettivo è di aumentare la concentrazione di acetone, paiono esserci diversi modi di procedere, associati alle variazioni delle tre variabili del processo.

C'è un'interessante relazione fra velocità di riflusso e temperatura di riebolizzazione, che può essere vista esaminando il grafico di Figura 1.7. Tale grafico è stato costruito calcolando le concentrazioni medie in corrispondenza delle quattro differenti combinazioni di velocità di riflusso e temperatura di riebolizzazione, riportando queste concentrazioni medie in funzione della velocità di riflusso, quindi unendo con linee rette i punti che rappresentano i due livelli di temperature. La pendenza di ciascuna di queste rette non è la stessa, a indicare che l'effetto della velocità di riflusso è diverso in corrispondenza dei due valori di temperatura di rie-

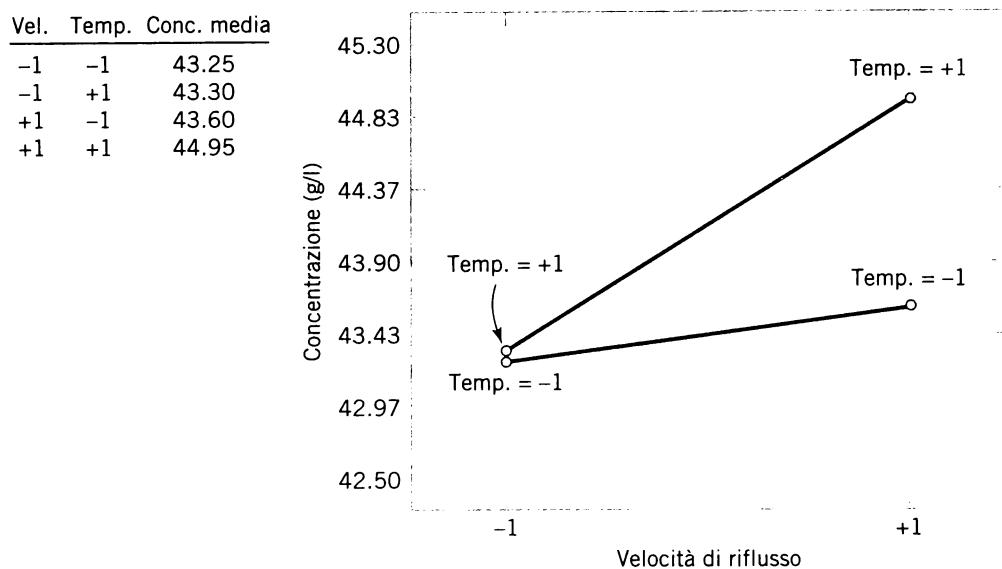


Figura 1.7 Interazione a due fattori fra velocità di riflusso e temperatura di riebollizione.

bollizione. Si tratta di un esempio di **interazione** tra due fattori. La sua interpretazione è immediata: se viene usato il livello basso ( $-1$ ) di velocità di riflusso la temperatura ha scarso effetto, mentre se viene impiegato il livello alto ( $+1$ ) un aumento della temperatura di riebollizione ha grande effetto sulla concentrazione media nel prodotto in uscita. In sistemi fisici e chimici si incontrano spesso interazioni; per studiarle l'unico modo è eseguire esperimenti fattoriali. In effetti, se sono presenti interazioni e non si procede a un esperimento fattoriale, si possono ottenere risultati errati o fuorvianti.

Possiamo facilmente estendere la strategia di un piano fattoriale al caso di più fattori. Supponiamo di voler considerare un quarto fattore, la concentrazione di acetone nel flusso in entrata. La Figura 1.8 mostra come si potrebbero studiare tutti e quattro i fattori in un piano fattoriale. Poiché i quattro fattori sono ancora su due livelli, il piano sperimentale può venire rappresentato geometricamente come un cubo (in effetti è un ipercubo). Si noti che, come in ogni piano fattoriale, sono saggiate tutte le possibili combinazioni dei quattro fattori. L'esperimento richiede 16 esecuzioni. Se ciascuna combinazione dei livelli dei fattori di Figura 1.8 viene provata una volta, questo esperimento ha in effetti lo stesso numero di esecuzioni del piano fattoriale replicato a tre fattori di Figura 1.6.

In generale, se vi sono  $k$  fattori e ciascuno di essi ha due livelli, un piano fattoriale richiederà  $2^k$  esecuzioni. Per esempio, con  $k = 4$  il piano  $2^4$  di Figura 1.8 richiede 16 test. Chiaramente, all'aumentare del numero di fattori il numero di esecuzioni richieste in un esperimento fattoriale cresce rapidamente; per esempio, otto fattori, ciascuno su due livelli, porterebbero a 256 esecuzioni. Tale numero diventa rapidamente irrealizzabile dal punto di vista del tempo e delle altre risorse necessarie. Per fortuna, quando vi sono quattro, cinque o più fattori, in genere non è necessario sottoporre a prova tutte le combinazioni possibili dei livelli dei fattori. Un **esperimento fattoriale frazionario** è una variante della disposizione fattoriale di base in cui viene verificato effettivamente solo un sottoinsieme delle combinazioni di fattori. La Figura 1.9 mostra un piano fattoriale frazionario per la versione a quattro fattori dell'esperimento della colonna di distillazione. Le combinazioni di test cerchiate in

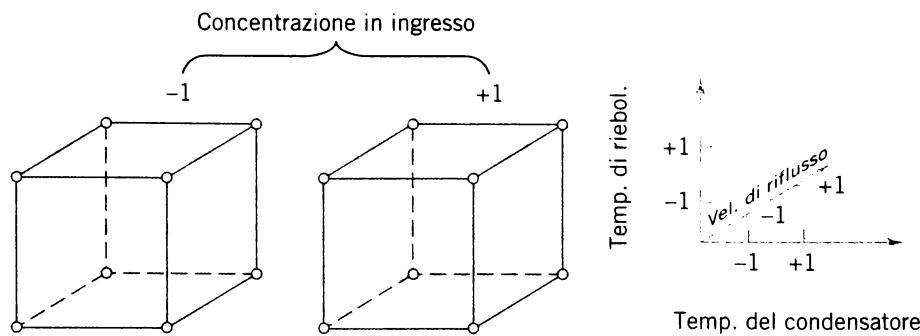


Figura 1.8 Un esperimento fattoriale a quattro fattori per la colonna di distillazione.

figura sono le uniche che è necessario eseguire. L'esperimento richiede cioè solo 8 esecuzioni anziché le originali 16; di conseguenza, sarebbe quello che si definisce una **frazione un mezzo**. Si tratta di un piano sperimentale eccellente con il quale studiare tutti e quattro i fattori; fornirà valide informazioni sui singoli effetti dei quattro fattori e anche qualche informazione sulla loro interazione.

Gli esperimenti fattoriali e quelli fattoriali frazionari sono frequentemente usati dagli ingegneri e dagli scienziati nella ricerca e nello sviluppo industriali, dove vengono progettate e sviluppate nuove tecnologie e nuovi prodotti, e dove si cerca di migliorare i prodotti esistenti. Dal momento che una parte rilevante del lavoro degli ingegneri coinvolge prove e sperimentazioni, è essenziale che tutti gli ingegneri comprendano i principi che sono alla base della pianificazione efficace di un esperimento.

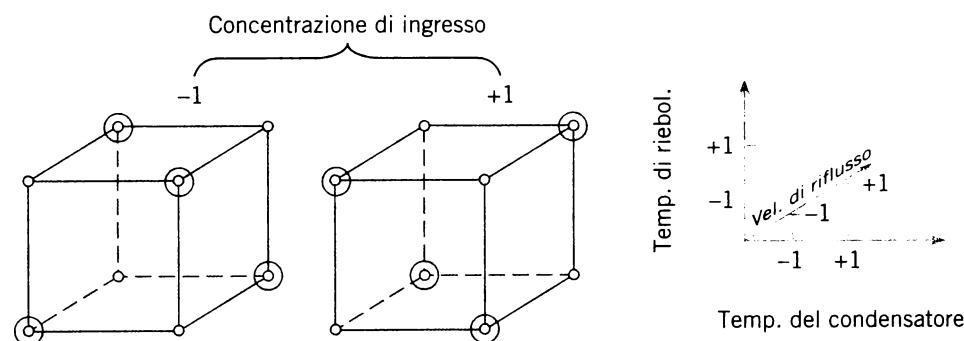


Figura 1.9 Un esperimento fattoriale frazionario per la colonna di distillazione.

#### 1.2.4 Campioni casuali

Come illustrato nei precedenti tre sottoparagrafi, l'analisi statistica è quasi totalmente interamente basata sull'idea di utilizzare un **campione** di dati scelto da qualche **popolazione**. Lo scopo è di usare i dati di tale campione per assumere decisioni o ricavare qualche informazione sulla popolazione intera. Si ricordi che la popolazione è l'insieme completo di elementi da cui è estratto il campione; quest'ultimo è solo un sottoinsieme dell'insieme popolazione.

Per esempio, supponiamo di produrre wafer di semiconduttore, e di volere indagare la resistività dei wafer di un particolare lotto. In questo caso, il lotto costituisce la popolazione. La procedura di indagine prevede di selezionare un campione di (per esempio) tre wafer e di misurarne la resistività. Si tratta di un esempio di **popolazione fisica**; la popolazione, cioè,

consiste in un ben definito gruppo di oggetti, spesso in numero finito, i quali sono disponibili al momento stesso del campionamento.

I dati, spesso, vengono raccolti come risultato di un esperimento ingegneristico. Per esempio, ritorniamo all'esperimento degli O-ring descritto nel Paragrafo 1.1. Inizialmente, si erano prodotti otto O-ring, poi sottoposti a un bagno di acido nitrico, a seguire il quale era stata determinata la resistenza alla trazione di ciascuna guarnizione. In questo caso le resistenze alla trazione degli otto O-ring costituiscono un campione di una popolazione, e quest'ultima è costituita da tutte le misure di resistenza alla trazione che sarebbe possibile ottenere. Questo tipo di popolazione è detto **popolazione concettuale**. Molti problemi di ingegneria chiamano in causa popolazioni concettuali; l'esperimento degli O-ring ne è un esempio semplice ma abbastanza tipico. L'esperimento fattoriale utilizzato per studiare la concentrazione nella colonna di distillazione del Paragrafo 1.2.3 dà anch'essa luogo a dati campionari estratti da una popolazione concettuale.

Anche il modo in cui i campioni vengono scelti è importante. Per esempio, supponiamo di voler ottenere informazioni sulle competenze matematiche degli studenti dell'ASU (*Arizona State University*): abbiamo a che fare con una popolazione fisica, dunque. Supponiamo inoltre di selezionare come campione d'indagine tutti gli studenti che frequentano il corso di statistica per ingegneria. Sarebbe una pessima idea, perché questi specifici studenti avrebbero con tutta probabilità competenze matematiche nettamente diverse da quelle riscontrabili nella maggior parte della popolazione. In generale, è improbabile che i campioni scelti sulla base della convenienza, o attraverso procedimenti che coinvolgono il giudizio soggettivo dell'ingegnere, diano risultati corretti. Per esempio, scegliendo il campione di studenti di statistica per ingegneria si arriverebbe a conclusioni distorte sulle competenze matematiche della popolazione. È quanto accade solitamente con campionamenti soggettivi o di comodo.

Affinché i metodi statistici funzionino correttamente producendo risultati validi si devono utilizzare **campioni casuali**. Il metodo di campionamento casuale più elementare è il **campionamento casuale semplice**. Per comprendere in che cosa consiste, prendiamo nuovamente l'esempio dell'indagine sulle competenze matematiche. Assegniamo a ciascuno studente dell'intera popolazione di studenti dell'ASU un numero intero compreso fra 1 e  $N$ . Supponiamo poi di voler selezionare un campione casuale semplice di 100 studenti. Potremmo allora usare un software per generare a caso 100 numeri interi da 1 a  $N$ , ciascuno dei quali ha la medesima probabilità di estrazione. Selezionando nella popolazione fisica gli studenti corrispondenti a tali numeri estratti, otterremmo il campione casuale semplice. Si noti che in questo modo ogni studente della popolazione ha la stessa probabilità di confluire nel campione. In altri termini, tutti i potenziali campioni di dimensione  $n = 100$  hanno la stessa probabilità di essere scelti.

### Definizione

Un **campione casuale semplice** di dimensione  $n$  è un campione estratto da una popolazione in modo tale che ogni potenziale campione di dimensione  $n$  abbia la stessa probabilità di venire scelto.

### ESEMPIO 1.1

Misure  
di corrente

Un ingegnere elettrico misura più volte la corrente che fluisce in un circuito elementare e osserva che le misure ottenute sono ogni volta diverse. Si possono considerare queste misure come un campione casuale semplice? Qual è la popolazione?

Se il circuito è lo stesso per ogni misura, e se le caratteristiche dell'amperometro non vengono modificate, possiamo considerare le misure di corrente come un campione casuale semplice. La popolazione è una popolazione concettuale: consiste in tutte le misure di corrente che potrebbero essere eseguite su questo circuito con questo amperometro.

### ESEMPIO 1.2

#### Misure nella colonna di distillazione

Si consideri la colonna di distillazione descritta nel Paragrafo 1.2. Si supponga che l'ingegnere faccia funzionare questa colonna per 24 ore consecutive e registri la concentrazione di acetone al termine di ciascuna ora. Si tratta di un campione casuale?

Anche questo è un esempio che coinvolge una popolazione concettuale, costituita da tutte le possibili osservazioni sulla concentrazione oraria. Se si è del tutto sicuri che le letture consecutive sono prese sotto le medesime identiche e costanti condizioni, e che è improbabile che differiscano da osservazioni future, allora, e solo allora, sarebbe ragionevole ritenere i dati raccolti un campione casuale. Nell'esempio abbiamo 24 letture consecutive; è altamente probabile che queste risultino differenti dalle osservazioni future, perché i processi chimici (e non solo quelli) tendono spesso a "spostarsi" nel tempo e a mostrare comportamenti diversi in diversi periodi di tempo; ciò avviene, principalmente, perché gli ingegneri possono apportare migliorie ai reagenti, ai fattori ambientali o alle condizioni operative man mano che apprendono come eseguire al meglio il processo.

Nel Paragrafo 3.13 forniremo una definizione più rigorosa di campione casuale semplice e ne discuteremo alcune proprietà.

Non sempre è facile ottenere un campione casuale. Per esempio, si consideri il lotto di wafer di semiconduttore; se questi sono confezionati e impilati in un contenitore, può essere difficoltoso prelevarli dal fondo o dai lati: si è tentati perciò di prelevare i tre wafer necessari dalla fila superiore, più accessibile. È, questo, un esempio di campionamento di comodo, che può produrre risultati non soddisfacenti, magari perché i pezzi sono stati confezionati in ordine di produzione e quelli in cima sono stati prodotti per ultimi, dopo che può essere occorso qualcosa di inusuale nel processo.

La raccolta di dati retrospettiva non sempre dà luogo a dati considerabili come un campione casuale. Si tratta spesso di dati di comodo, che possono non riflettere le prestazioni del processo in questione. I dati provenienti da studi osservativi è più probabile che riflettano un campionamento casuale, perché di solito si esegue uno studio specifico per la raccolta dei dati stessi. I dati estratti da un esperimento pianificato possono costituire un campione casuale se le singole osservazioni dell'esperimento sono effettuate in ordine casuale. Una casualizzazione totale dell'ordine delle prove dell'esperimento contribuisce a eliminare gli effetti di fattori sconosciuti che potrebbero variare durante l'esecuzione, e garantisce così che si possano considerare i dati estratti come un campione casuale.

La raccolta dei dati retrospettiva su un processo, quella eseguita attraverso uno studio osservativo, e persino quella mediante un esperimento pianificato, comportano quasi sempre il campionamento in una popolazione concettuale. L'obiettivo, in molte di queste indagini sui dati, è di trarre conclusioni sul comportamento futuro del sistema o del processo oggetto di studio. Uno **studio analitico** è uno studio o un esperimento in cui si devono trarre conclusioni per una **popolazione futura**. Per esempio, nell'esperimento della colonna di distillazione vogliamo ricavare un'inferenza sulla concentrazione delle quantità di acetone che verranno prodotte in futuro per essere vendute ai clienti. Si tratta in questo caso di uno studio analitico riguardante una popolazione concettuale che non esiste ancora. Chiaramente, oltre al cam-

pionamento casuale si deve fare qualche ipotesi aggiuntiva di **stabilità** del processo nel tempo. Per esempio, può essere necessario assumere che le cause di variabilità attualmente presenti nella produzione siano le stesse che saranno presenti nella produzione futura.

Il problema dei wafer di semiconduttore estratti da un lotto per determinare la resistività di quest'ultimo rappresenta uno esempio di **studio enumerativo**, in cui viene usato un campione per effettuare un'inferenza sulla popolazione dalla quale è stato selezionato il campione medesimo. Anche l'indagine volta a determinare le competenze matematiche degli studenti dell'ASU è uno studio enumerativo. Si noti che i campioni casuali sono necessari sia negli studi analitici sia in quelli enumerativi, ma lo studio analitico richiede in più un'ipotesi di stabilità. La Figura 1.10 fornisce un'illustrazione schematica dei due tipi di studio.

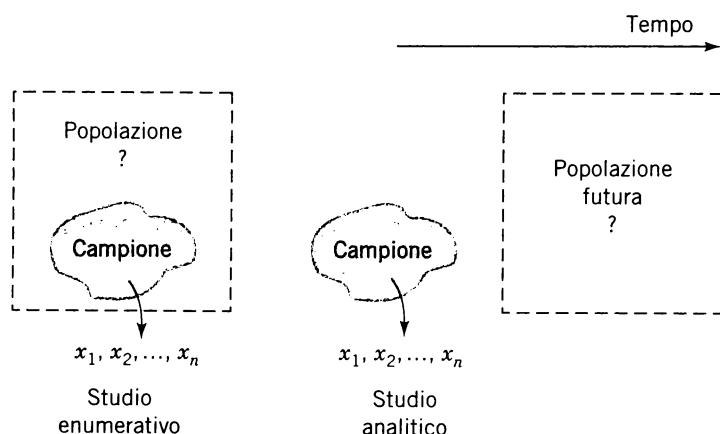


Figura 1.10 Studio enumerativo e studio analitico.

### 1.3 MODELLI MECCANICISTICI E MODELLI EMPIRICI

I modelli rivestono un ruolo importante nell'analisi di quasi tutti i problemi di ingegneria. Gran parte della formazione degli ingegneri riguarda l'apprendimento di modelli importanti in specifici campi e delle tecniche per applicare tali modelli alla formulazione e alla risoluzione dei problemi. Come semplice esempio, supponiamo di stare misurando un flusso di corrente in un filo sottile di rame. Il nostro modello per questo fenomeno è dato dalla legge di Ohm:

$$\text{corrente} = \text{tensione}/\text{resistenza}$$

ovvero

$$I = E/R \quad (1.1)$$

Chiamiamo questo tipo di modello **meccanicistico**, perché è costruito a partire dalla sottostante conoscenza del meccanismo fisico fondamentale che lega tra loro le variabili. Tuttavia, se eseguissimo questa misura più di una volta, magari in diversi momenti o persino in diversi giorni, la corrente rilevata potrebbe essere leggermente diversa da misura a misura a causa di piccole variazioni dei fattori che non sono totalmente sotto controllo, come varia-

zioni della temperatura dell'ambiente, fluttuazioni nel funzionamento dello strumento di misura, piccole impurità presenti in diversi punti del filo e sbalzi nella tensione di alimentazione. Di conseguenza, un modello più realistico per la corrente osservata potrebbe essere:

$$I = E/R + \epsilon \quad (1.2)$$

dove  $\epsilon$  è un termine aggiunto al modello per tenere conto del fatto che i valori osservati di corrente non corrispondono al modello meccanicistico. Possiamo insomma vedere  $\epsilon$  come un termine che ingloba gli effetti di tutte le fonti di variabilità non contemplate dal modello, che influenzano questo sistema.

A volte gli ingegneri hanno a che fare con problemi per i quali non esiste un modello meccanicistico semplice o chiaro che spieghi il fenomeno. Per esempio, supponiamo di essere interessati al peso molecolare medio ( $M_n$ ) di un polimero. Ora, sappiamo che  $M_n$  è legato alla viscosità  $V$  del materiale, e che dipende inoltre dalla quantità  $C$  di catalizzatore e dalla temperatura  $T$  nel reattore di polimerizzazione al momento della produzione del materiale. La relazione tra  $M_n$  e queste variabili è esprimibile come

$$M_n = f(V, C, T) \quad (1.3)$$

dove la forma della funzione  $f$  non è nota. Si potrebbe magari ottenere un modello operativo mediante uno sviluppo in serie di Taylor al primo ordine, che porterebbe a un modello della forma

$$M_n = \beta_0 + \beta_1 V + \beta_2 C + \beta_3 T \quad (1.4)$$

dove i coefficienti  $\beta$  sono incogniti. Come nella legge di Ohm, questo modello non descriverà esattamente il comportamento del sistema, perciò dovremo mettere in conto altre cause di variabilità che possono influenzare il peso molecolare, aggiungendo al modello un termine di disturbo casuale; pertanto, il modello che useremo per la relazione tra il peso molecolare e le altre variabili è

$$M_n = \beta_0 + \beta_1 V + \beta_2 C + \beta_3 T + \epsilon \quad (1.5)$$

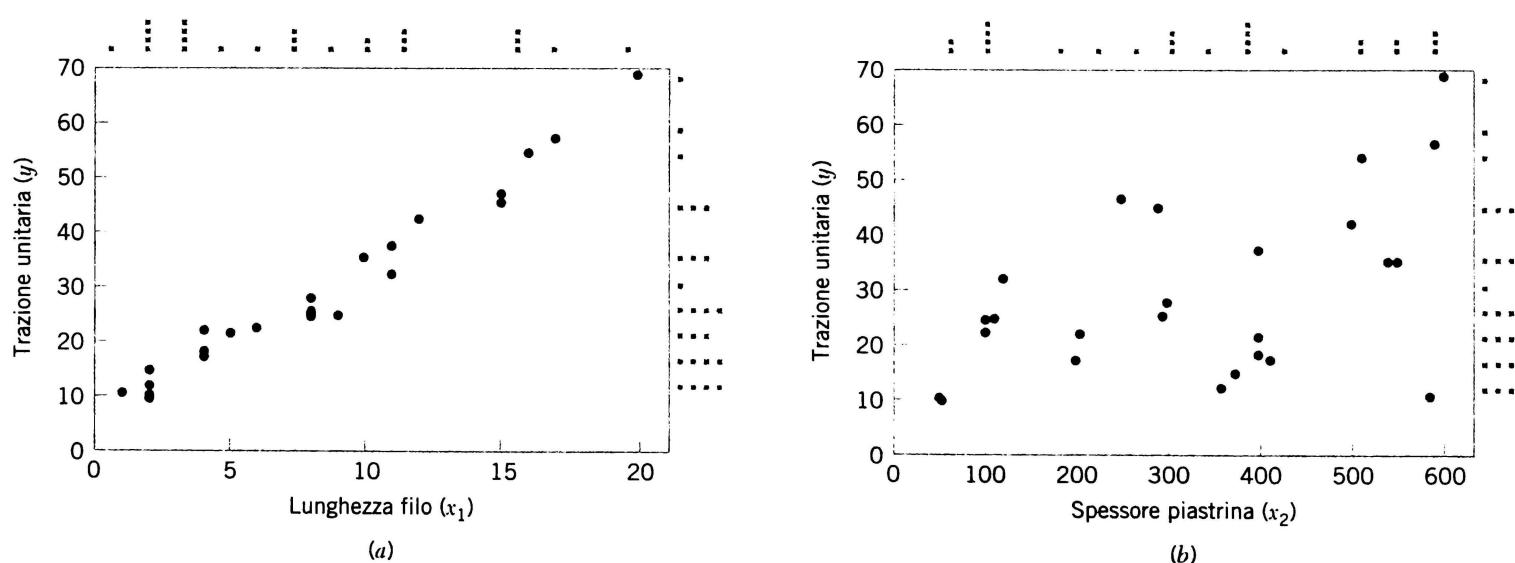
Un modello di questo tipo viene chiamato **modello empirico**, in quanto per costruirlo si ricorre alla conoscenza ingegneristica e scientifica del fenomeno, ma esso non è sviluppato direttamente dalla comprensione teorica o dei principi di base del meccanismo sottostante. Sono necessari dei dati per arrivare a una stima dei coefficienti  $\beta$  nell'Equazione (1.5). Tali dati potrebbero derivare da uno studio retrospettivo od osservativo, oppure li si potrebbe ottenere mediante un esperimento pianificato.

Per illustrare questi concetti con un esempio specifico, si considerino i dati di Tabella 1.1, relativi a tre variabili raccolte in uno studio osservativo condotto in uno stabilimento di produzione di semiconduttori. In tale stabilimento, il semiconduttore finito è collegato a un supporto mediante un filo metallico ricoperto. Le variabili tabulate sono la trazione unitaria (una misura della quantità di forza necessaria a rompere il collegamento), la lunghezza del filo e lo spessore della piastrina. Sarebbe utile trovare un modello che leggi la trazione unitaria alle altre due variabili. Sfortunatamente, non vi è un meccanismo fisico facilmente applicabile, perciò in questo caso un approccio di tipo meccanicistico non avrebbe probabilmente successo.

**Tabella 1.1** Dati relativi al filo metallico ricoperto.

Osservazione Numero	Traz. unitaria $y$	Lungh. filo $x_1$	Spes. piastrina $x_2$	Osservazione Numero	Traz. unitaria $y$	Lunghezza filo $x_1$	Spes. piastrina $x_2$
1	9.95	2	50	14	11.66	2	360
2	24.45	8	110	15	21.65	4	205
3	31.75	11	120	16	17.89	4	400
4	35.00	10	550	17	69.00	20	600
5	25.02	8	295	18	10.30	1	585
6	16.86	4	200	19	34.93	10	540
7	14.38	2	375	20	46.59	15	250
8	9.60	2	52	21	44.88	15	290
9	24.35	9	100	22	54.12	16	510
10	27.50	8	300	23	56.63	17	590
11	17.08	4	412	24	22.13	6	100
12	37.00	11	400	25	21.15	5	400
13	41.95	12	500				

La Figura 1.11a rappresenta un **diagramma di dispersione** della trazione unitaria  $y$  in funzione della lunghezza  $x_1$ . Il grafico è stato costruito semplicemente riportando le coppie di osservazioni  $(y_i, x_{1i})$  con  $i = 1, 2, \dots, 25$ , prese dalle prime due colonne di Tabella 1.1. Abbiamo utilizzato il pacchetto software Minitab per realizzare tale diagramma. Minitab ha un'opzione che consente di produrre un diagramma a punti lungo i lati destro e superiore del diagramma di dispersione, consentendo di vedere la distribuzione di ciascuna variabile individualmente. In questo senso, un diagramma di dispersione è una versione bidimensionale di un diagramma a punti.

**Figura 1.11** Diagrammi di dispersione relativi alla trazione unitaria della giunzione protetta.

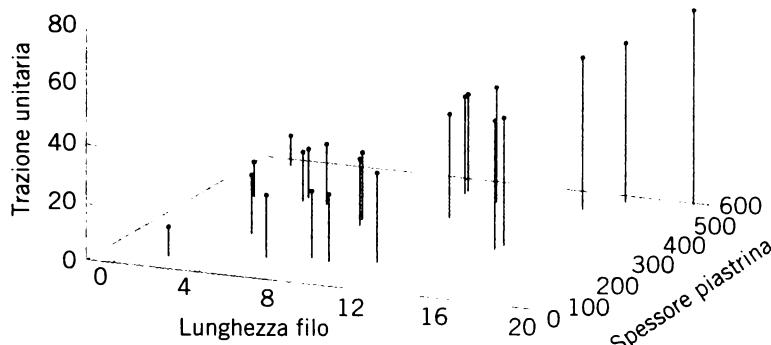


Figura 1.12 Diagramma di dispersione tridimensionale relativo ai dati della giunzione protetta.

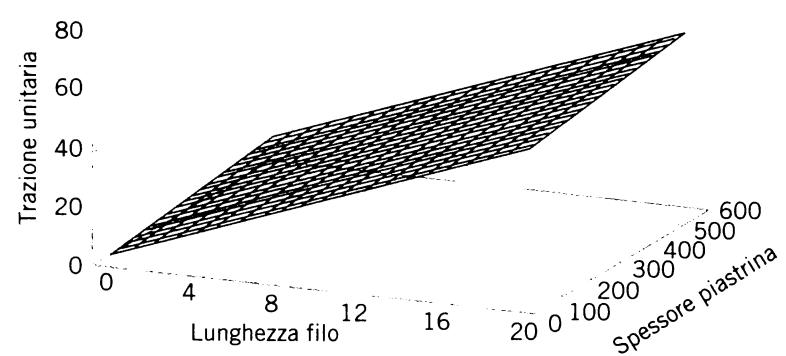


Figura 1.13 Grafico dei valori di trazione unitaria stimati, ricavati dal modello empirico dell'Equazione (1.6).

Il diagramma di dispersione di Figura 1.11a rivela che al crescere della lunghezza del filo cresce anche la trazione unitaria della giunzione. Un'informazione analoga è fornita dal diagramma di dispersione di Figura 1.11b, che riporta la trazione unitaria in funzione dello spessore della piastrina,  $x_2$ . La Figura 1.12 è un **diagramma tridimensionale di dispersione** delle osservazioni effettuate su trazione unitaria, lunghezza del filo e spessore della piastrina. In base a questi grafici sembra ragionevole ritenere che per la relazione sotto esame sia appropriato un modello empirico del tipo

$$\text{Trazione unitaria} = \beta_0 + \beta_1 (\text{lunghezza filo}) + \beta_2 (\text{spessore piastrina}) + \epsilon$$

In generale, questo tipo di modello empirico viene detto **modello di regressione**. Nel Capitolo 6 mostreremo come costruire questi modelli e come verificarne l'adeguatezza come funzioni approssimanti. Verrà anche presentato un metodo per stimare i parametri nei modelli di regressione, detto **metodo dei minimi quadrati**, che ha origine dal lavoro di Karl Gauss. In sintesi, questo metodo sceglie i coefficienti  $\beta$  del modello empirico in modo da minimizzare la somma delle distanze al quadrato fra ciascun punto (ossia ciascun dato) e il piano rappresentato dall'equazione del modello. Applicando questa tecnica ai dati di Tabella 1.1 si ottiene

$$\widehat{\text{traz. unit.}} = 2.26 + 2.74 (\text{lunghezza filo}) + 0.0125 (\text{spessore piastrina}) + \epsilon \quad (1.6)$$

dove il “cappuccio” sulla trazione unitaria indica che si tratta di una grandezza stimata.

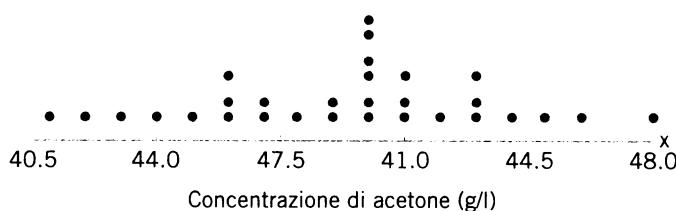
La Figura 1.13 è un grafico dei valori previsti di trazione unitaria in funzione della lunghezza del filo e dello spessore della piastrina, ottenuti dall'Equazione (1.6). Si noti che tali valori predetti giacciono su un piano sopra lo spazio lunghezza filo-spessore piastrina. Osservando il grafico dei dati in Figura 1.12, questo modello sembra ragionevole. Il modello empirico dell'Equazione (1.6) potrebbe venire usato per predire i valori di trazione unitaria per varie combinazioni di lunghezza del filo e di spessore della piastrina. In sostanza, il modello empirico può essere utilizzato dagli ingegneri esattamente allo stesso modo in cui può venire impiegato un modello meccanicistico.

## 1.4 OSSERVAZIONE DEI PROCESSI NEL TEMPO

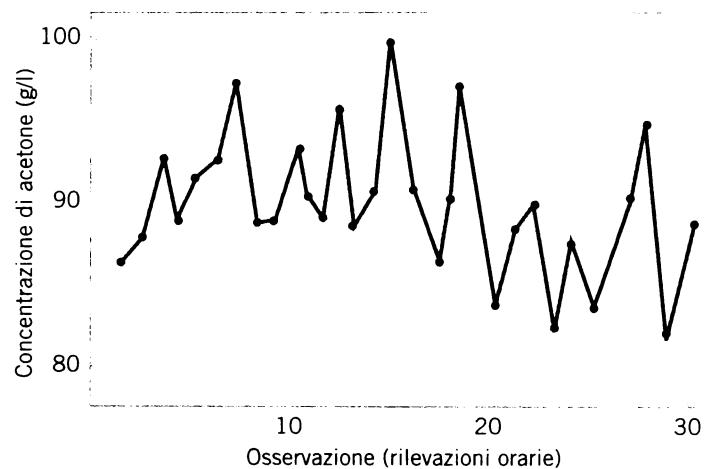
Accade spesso che i dati vengano raccolti su un periodo più o meno lungo di tempo. Ciò avviene in molte situazioni pratiche; probabilmente la situazione più familiare è quella di

dati economici e di impresa che riflettono quotazioni giornaliere di Borsa, tassi di interesse, tassi di inflazione e di disoccupazione mensili, volumi di produzione trimestrali ecc. I giornali e le pubblicazioni economiche come il Wall Street Journal visualizzano tali dati, tipicamente, mediante tabelle e grafici. Anche per molti studi di ingegneria i dati vengono raccolti nel corso del tempo. Con un grafico dei dati rispetto al tempo i fenomeni che possono influenzare il sistema o il processo spesso diventano più visibili, e la stabilità del processo può essere giudicata meglio. Per esempio, la **carta di controllo** è una tecnica che visualizza i dati in funzione del tempo e permette all'ingegnere di valutare la stabilità del processo.

La Figura 1.14 è un **diagramma a punti** delle letture di concentrazione di acetone, prese ogni ora, relative alla colonna di distillazione descritta nel Paragrafo 1.2. La sensibile variazione che si può osservare sul diagramma a punti indica un possibile problema, ma la carta non aiuta a spiegare la ragione di tale variazione. Poiché i dati sono raccolti nel corso del tempo, vengono detti **serie temporali**. Un grafico dei dati in funzione del tempo, come quello di Figura 1.15, è detto **grafico delle serie temporali**. In figura è visibile uno scostamento dal livello medio del processo, e si può ottenere una stima del tempo di tale scostamento.



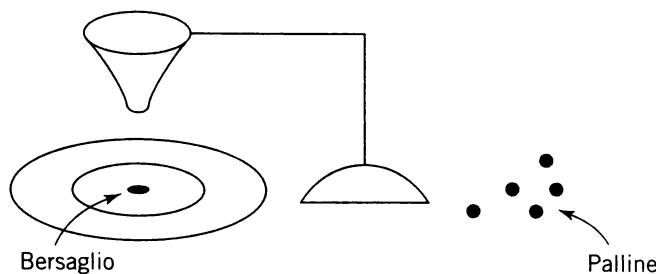
**Figura 1.14** Un diagramma a punti illustra la variazione ma non identifica il problema sottostante.



**Figura 1.15** Un grafico delle serie temporali della concentrazione di acetone fornisce maggiori informazioni di quelle che può fornire un istogramma.

Il famoso esperto di qualità W. Edwards Deming ha sottolineato che è importante comprendere la natura di una variazione nel corso del tempo. Egli condusse un esperimento in cui tentava di far cadere alcune palline il più vicino possibile a un bersaglio posto su un tavolo. Per far ciò, adoperò un imbuto montato su un piedestallo, facendo cadere le palline entro l'imbuto (Figura 1.16). L'imbuto era allineato il più possibile al centro del bersaglio. Deming ricorse quindi a due diverse strategie per condurre l'esperimento. (1) Tenne immobile l'imbuto, lasciando scivolare le palline una dopo l'altra e prendendo nota della loro distanza finale dal bersaglio. (2) Lasciò cadere la prima pallina e registrò la sua posizione rispetto al bersaglio; dopodiché spostò l'imbuto di una distanza uguale, ma nel verso opposto, nel tentativo di compensare l'errore. Proseguì quindi con questo tipo di aggiustamenti dopo ogni lancio di pallina.

Una volta portate a termine entrambe le strategie, Deming notò che la variabilità nella distanza dal bersaglio per la strategia 2 era all'incirca doppia di quella della strategia 1: le regolazioni dell'imbuto avevano aumentato le deviazioni dal bersaglio. La spiegazione di questo fatto è che l'errore per una pallina (la deviazione della posizione della pallina rispetto

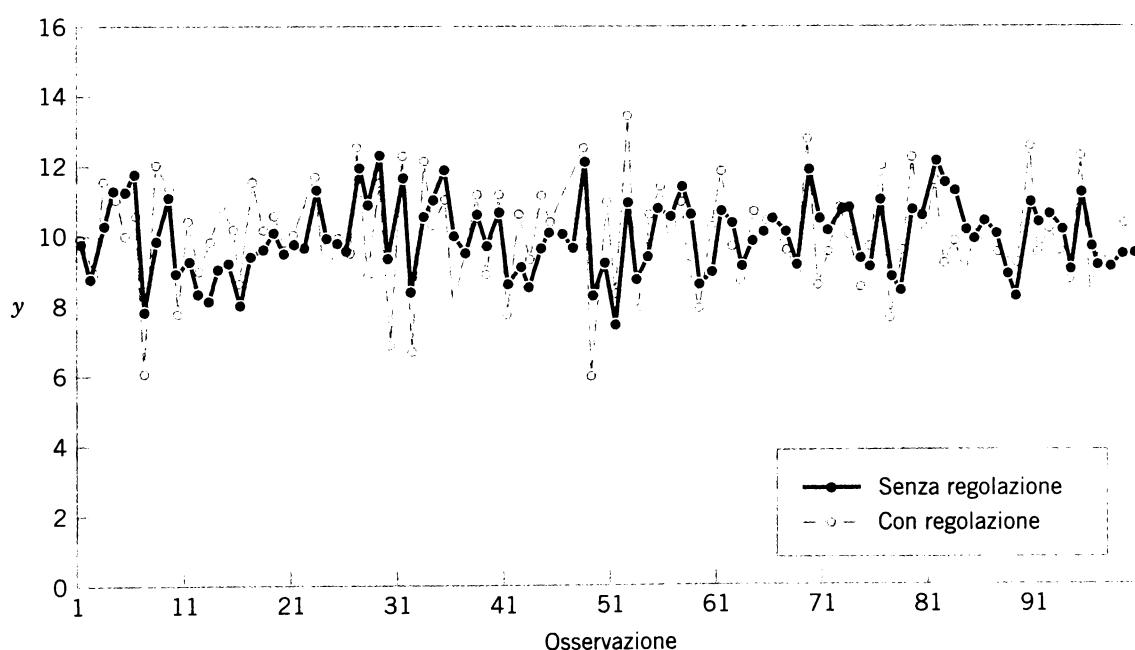


**Figura 1.16**  
L'esperimento  
dell'imbuto  
realizzato da Deming.

al bersaglio) non fornisce alcuna informazione sull'errore cui sarà soggetta la pallina successiva. Di conseguenza, gli aggiustamenti dell'imbuto non diminuiscono gli errori futuri; al contrario, essi tendono a portare l'imbuto più lontano dal bersaglio.

Questo interessante esperimento evidenzia il fatto che le regolazioni effettuate su un processo e basate su disturbi casuali possono in effetti aumentare la variazione del processo medesimo. Ci si riferisce perciò a questo fenomeno con l'espressione **eccesso di controllo**. Le regolazioni andrebbero eseguite solo per compensare spostamenti non casuali nel processo; solo così esse possono effettivamente risultare utili. Per dimostrare la lezione tratta dall'esperimento dell'imbuto può essere sfruttata una simulazione al computer. La Figura 1.17 mostra un grafico temporale di 100 misurazioni (indicate con  $y$ ) di un processo in cui sono presenti solo disturbi casuali. Il valore-bersaglio per il processo è 10 unità. La figura mostra i dati con e senza gli aggiustamenti applicati alla media del processo nel tentativo di ottenere dati più vicini al bersaglio. Ogni regolazione è uguale e opposta alla deviazione della misura precedente rispetto al bersaglio. Per esempio, quando la misurazione è 11 (una unità al di sopra del bersaglio), la media viene ridotta di una unità prima di generare la misurazione successiva. L'eccesso di controllo, come si vede, fa aumentare le deviazioni dal bersaglio.

La Figura 1.18 mostra i dati senza le regolazioni ripresi dalla Figura 1.17, tranne per il fatto che le misurazioni dopo l'osservazione numero 50 sono aumentate di due unità per



**Figura 1.17** Le regolazioni applicate ai disturbi casuali producono un eccesso di controllo sul processo e fanno aumentare le deviazioni dal bersaglio.

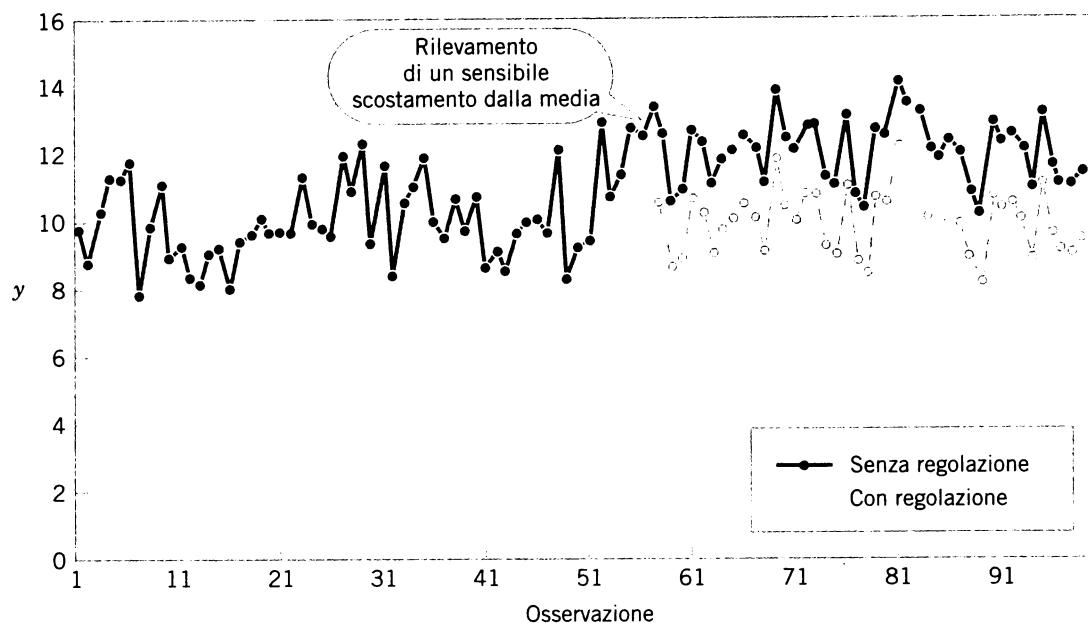


Figura 1.18 Si è rilevato uno scostamento sensibile dalla media in corrispondenza dell’osservazione numero 57, e una regolazione (una diminuzione di due unità) riduce le deviazioni dal bersaglio.

simulare l’effetto di un salto nella media del processo. Quando questo accade veramente, una regolazione può essere utile. La Figura 1.18, inoltre, mostra i dati ottenuti quando viene applicata alla media una regolazione (una diminuzione di due unità) dopo che si è rilevato il salto (all’osservazione 57). Si noti che questo aggiustamento diminuisce la deviazione dal bersaglio.

Il problema di quando applicare (e in quale misura) una regolazione inizia con la comprensione dei tipi di variazione che influenzano un processo. Una **carta di controllo** è uno strumento prezioso per valutare la variabilità nelle serie temporali. La Figura 1.19 rappresenta una carta di controllo per i dati sulla concentrazione della Figura 1.14. La **linea centrale** sulla carta di controllo è semplicemente la media delle misure di concentrazione per i primi 20 campioni ( $= 51.5 \text{ g/l}$ ) quando il processo è stabile. Il **limite di controllo superiore** e il **limite di controllo inferiore** sono una coppia di limiti, ricavati statisticamente, che riflettono la variabilità intrinseca, o naturale, del processo. Tali limiti sono posizionati tre deviazioni standard dei valori di concentrazione rispettivamente al di sopra e al di sotto della linea centrale. Se il processo funziona come dovrebbe, senza che siano presenti nel sistema cause esterne di variabilità, allora le misure di concentrazione dovrebbero fluttuare in maniera casuale intorno alla linea centrale, e quasi tutte dovrebbero cadere entro i limiti di controllo.

Nella carta di controllo di Figura 1.19 la cornice di riferimento costituita dalla linea centrale e dai limiti di controllo indica che qualche disturbo o qualche perturbazione ha influenzato il processo intorno al campione 20, perché tutte le osservazioni successive sono al di sotto della linea centrale e due di esse cadono al di sotto del limite di controllo inferiore. Questo è un segnale molto forte della necessità di un’azione correttiva sul processo. Se si può trovare ed eliminare la causa di tale perturbazione è possibile migliorare considerevolmente le prestazioni del processo.

Le carte di controllo sono essenziali nell’analisi ingegneristica, per le seguenti ragioni. In alcuni casi i dati relativi al campione sono di fatto selezionati dalla popolazione di inte-

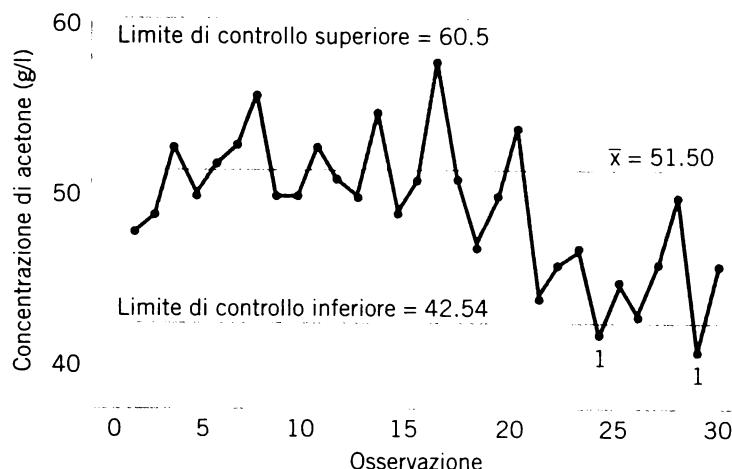


Figura 1.19 Una carta di controllo per i dati della colonna di distillazione.

resse: il campione è cioè un sottoinsieme della popolazione. Per esempio, si può selezionare un campione di tre wafer di silicio da un lotto di produzione; in base ai dati relativi a tale campione, si vuole trarre qualche conclusione sull'intero lotto. Non ci si aspetta per esempio che la media delle misure di resistività nel campione sia esattamente uguale alla media delle resistività del lotto; tuttavia, se la media del campione è alta, si può ragionevolmente ritenere che quella del lotto è troppo alta.

In molti altri casi si usano i dati correnti per trarre conclusioni sulla prestazione futura di un processo. Per esempio, non siamo solo interessati alle misure di concentrazione di acetone prodotto dalla colonna di distillazione: vogliamo anche poter trarre conclusioni sulla concentrazione della produzione futura che sarà venduta ai clienti. Questa popolazione di produzione futura non esiste ancora. Chiaramente questa analisi richiede come ulteriore assunzione la presenza di una certa stabilità. Per esempio, si potrebbe assumere che le cause di variabilità nella produzione attuale siano le stesse di quelle della produzione futura. Una carta di controllo è lo strumento principale per valutare la stabilità di un processo.

Le carte di controllo sono un'applicazione molto importante della statistica ai fini del monitoraggio, del controllo e del miglioramento di un processo. La branca della statistica che fa uso delle carte di controllo viene detta **controllo statistico di processo**, o SPC (*Statistical Process Control*).

## TERMINI E CONCETTI RILEVANTI IN QUESTO CAPITOLO

- Approccio statistico
- Campionamento casuale
- Carta di controllo
- Cause di variabilità
- Diagramma di dispersione
- Diagramma a punti
- Esperimento fattoriale

- Esperimento pianificato
- Metodo scientifico o dell'ingegneria
- Modello empirico
- Studio analitico
- Studio enumerativo
- Studio osservativo
- Studio retrospettivo



# Sintesi numerica e presentazione grafica dei dati

---

## LA TEMPERATURA GLOBALE

In questo capitolo vedremo che per la sintesi numerica dei dati raccolti esistono sia metodi numerici, sia tecniche di presentazione grafica, queste ultime particolarmente importanti: ogni buona analisi statistica dovrebbe sempre iniziare con un **grafico dei dati**.

James Watt inventò il motore a vapore nei primi anni del diciannovesimo secolo, e negli anni Venti del medesimo secolo i combustibili fossili iniziarono a fornire energia per l'industria e i trasporti. La rivoluzione industriale andava letteralmente a tutto vapore. Da allora, come mostra la Figura 2.1, quantità sempre maggiori di anidride carbonica si sono riversate nell'atmosfera. Il grafico mostra inoltre che le temperature del pianeta sono cresciute in palese sincronia con i livelli crescenti di questo gas serra. I dati mostrati in questo singolo grafico hanno contribuito a far prendere coscienza della serietà del problema all'opinione pubblica, ai leader politici e finanziari e agli industriali. Si sta formando un consenso generale sull'idea che il controllo delle emissioni di anidride carbonica dev'essere in cima alla lista degli obiettivi dei prossimi anni, e che occorre sviluppare concreteamente nuove tecnologie che facciano uso di fonti energetiche rinnovabili senza emissioni di gas serra. È un investimento per garantirsi il futuro, per propagandare il quale è opportuno disporre di rappresentazioni grafiche dei dati chiare, coerenti e convincenti. Saranno necessari grafici più aggiornati e con una copertura maggiore per far sì che questa idea continui a circolare e divenga la base per un gran numero di attività imprenditoriali di nuova impostazione.

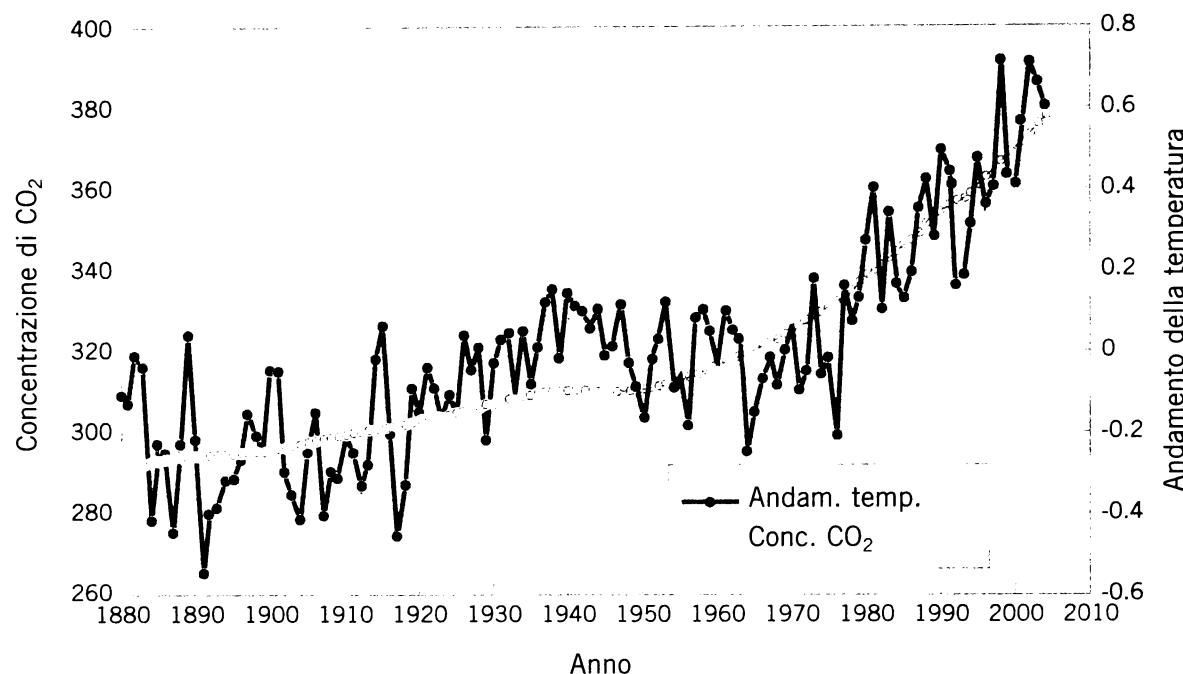


Figura 2.1 Andamento della temperatura media globale e della concentrazione globale di CO<sub>2</sub>, anni 1880-2004.

## CONTENUTI DEL CAPITOLO

- |   |                                  |
|---|----------------------------------|
| 2.1 VISUALIZZAZIONE E SINTESI NUMERICA<br>DEI DATI STATISTICI | 2.4 BOX PLOT                     |
| 2.2 DIAGRAMMI RAMI E FOGLIE                                   | 2.5 GRAFICI DELLE SERIE STORICHE |
| 2.3 ISTOGRAMMI  | 2.6 DATI MULTIVARIATI            |

## OBIETTIVI DI APPRENDIMENTO

Dopo uno studio attento di questo capitolo si sarà in grado di:

1. calcolare e interpretare la media campionaria, la varianza campionaria, la deviazione standard campionaria, la mediana e il range di un campione
2. spiegare i concetti di media campionaria, varianza campionaria, media della popolazione e varianza della popolazione
3. costruire e interpretare rappresentazioni grafiche dei dati, tra cui i diagrammi rami e foglie, gli histogrammi e i box plot, e comprendere l'utilità di queste rappresentazioni nella scoperta e nella sintesi delle caratteristiche dei dati
4. spiegare come si usano i box plot e altre rappresentazioni grafiche per effettuare un confronto visivo fra due o più campioni di dati
5. sapere come usare semplici grafici delle serie storiche per visualizzare le informazioni principali sui dati caratterizzati da un andamento temporale
6. costruire diagrammi di dispersione e calcolare e interpretare un coefficiente di correlazione campionario

## 2.1 VISUALIZZAZIONE E SINTESI NUMERICA DEI DATI STATISTICI

Per una corretta analisi statistica è essenziale poter disporre di visualizzazioni e di sintesi numeriche appropriate, poiché ciò consente di focalizzare l'attenzione su caratteristiche importanti dei dati stessi e suggerisce quale tipo di modello adottare per la risoluzione del problema. Nella presentazione e nell'analisi dei dati il calcolatore è ormai uno strumento indispensabile: nonostante per molte tecniche statistiche possa bastare l'uso di una calcolatrice tascabile, i computer consentono di assolvere ai medesimi compiti con uno sforzo minore e in modo molto più efficiente.

Per la maggior parte delle analisi statistiche si utilizza una libreria di programmi statistici già scritti. L'utente deve solo inserire i dati, quindi selezionare il tipo di analisi e di visualizzazione dell'output che gli serve. Sono disponibili pacchetti software statistici sia per i computer mainframe che per i PC; fra quelli più conosciuti e diffusi vi sono SAS (*Statistical Analysis System*), per server e personal computer, e Minitab (per PC). In questo libro presenteremo alcuni esempi di output di Minitab, ma non spiegheremo le modalità di utilizzo del programma (inserimento e trattamento dei dati, comandi); potrete facilmente trovare i manuali d'uso per Minitab o programmi similari nella vostra università, se non personale di assistenza esperto.

Alcune caratteristiche di un insieme di dati sono esprimibili **numericamente**. Per esempio, è possibile caratterizzare la posizione o tendenza centrale dei dati mediante l'usuale media aritmetica. Dato che penseremo quasi sempre ai nostri dati come a un campione, indicheremo tale media aritmetica come **media campionaria**.

### Media campionaria

Denotate con  $x_1, x_2, \dots, x_n$  le  $n$  osservazioni in un campione, la **media campionaria** è data da

$$\begin{aligned}\bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{\sum_{i=1}^n x_i}{n}\end{aligned}\tag{2.1}$$

**ESEMPIO 2.1**  
Resistenza  
degli O-ring:  
media campionaria

Si consideri l'esperimento relativo alla resistenza alla trazione degli O-ring, descritto nel Capitolo 1. I dati ricavati dal composto gommoso modificato sono mostrati nel **diagramma a punti** di Figura 2.2. La media campionaria (espressa in psi) per le otto osservazioni di resistenza è

$$\begin{aligned}\bar{x} &= \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^8 x_i}{8} = \frac{1037 + 1047 + \cdots + 1040}{8} \\ &= \frac{8440}{8} = 1055.0 \text{ psi}\end{aligned}$$

Una possibile interpretazione fisica della media campionaria come misura della posizione è rappresentata in Figura 2.2. Si noti che si può vedere la media campionaria = 1055 come “punto di equilibrio”; vale a dire: se ciascuna osservazione rappresenta 1 unità di massa posta nel corrispondente punto dell’asse  $x$ , un fulcro posizionato in  $\bar{x}$  equilibrerebbe esattamente questo sistema di pesi.

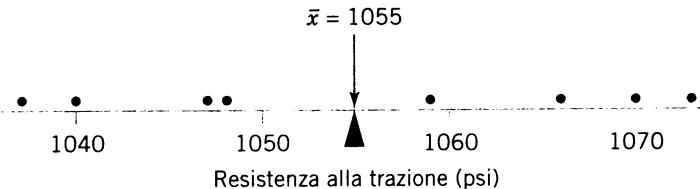


Figura 2.2 Diagramma a punti della resistenza alla trazione degli O-ring. La media campionaria è rappresentata come punto di equilibrio per un sistema di pesi.

La media campionaria è il valor medio di tutte le osservazioni nell’insieme di dati. Questi ultimi, in genere, costituiscono un **campione** di osservazioni selezionate da una più grande **popolazione** di osservazioni; nel caso dell’Esempio 2.1 la popolazione può essere costituita da tutte le guarnizioni che saranno vendute alla clientela. A volte c’è una effettiva popolazione fisica, come un lotto di wafer di silicio prodotti in una fabbrica di semiconduttori. Si può dunque pensare di calcolare il valor medio di tutte le osservazioni in una popolazione: si definisce la media risultante **media della popolazione**, e la si indica con la lettera greca  $\mu$  (si legge “mu”). Quando c’è un numero finito  $N$  di osservazioni nella popolazione, la media della popolazione è data da

$$\mu = \frac{\sum_{i=1}^N x_i}{N} \quad (2.2)$$

La media campionaria  $\bar{x}$  è una stima ragionevole della media  $\mu$  della popolazione. Pertanto, un ingegnere che stesse studiando il composto gommoso modificato per gli O-ring concluderebbe, sulla base dei dati, che una stima della resistenza media alla trazione è 1055 psi.

Benché utile, la media campionaria non fornisce tutte le informazioni su un campione di dati. La variabilità o dispersione presente in questi ultimi potrebbe essere descritta dalla **varianza campionaria** o dalla **deviazione standard campionaria**.

### Varianza e deviazione standard campionarie

Denotate con  $x_1, x_2, \dots, x_n$  le  $n$  osservazioni in un campione, la **varianza campionaria** è data da

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (2.3)$$

La **deviazione standard campionaria**,  $s$ , è la radice quadrata positiva della varianza campionaria.

Le unità di misura della varianza sono il quadrato di quelle della variabile. Se dunque  $x$  è misurata in psi, la varianza è misurata in  $(\text{psi})^2$ . La deviazione standard, al contrario, ha la comoda proprietà di misurare la variabilità nelle unità di misura originali della variabile  $x$  in gioco, nel nostro caso psi.

In che senso la varianza campionaria è una misura della variabilità?

Per vedere come la varianza campionaria misuri la dispersione o variabilità si faccia riferimento alla Figura 2.3, che mostra gli scarti  $x_i - \bar{x}$  per i dati di resistenza alla trazione degli O-ring. Più grande è la variabilità in tali dati, maggiori saranno – in valore assoluto – alcuni degli scarti  $x_i - \bar{x}$ . Poiché gli  $x_i - \bar{x}$ , addizionati tra loro, danno sempre somma zero, dobbiamo usare una misura della variabilità che trasformi gli scarti negativi in quantità positive. Il metodo impiegato nella varianza campionaria consiste nell'elevare al quadrato gli scarti. Di conseguenza, se  $s^2$  è piccola c'è una variabilità relativamente bassa nei dati, mentre se  $s^2$  è grande la variabilità è apprezzabile.

### ESEMPIO 2.2 Resistenza degli O-ring: varianza campionaria

La Tabella 2.1 mostra le grandezze necessarie a calcolare la varianza e la deviazione standard per i dati relativi alla resistenza alla trazione degli O-ring, riportati in Figura 2.3. Il numeratore della formula (2.3) per  $s^2$  è

$$\sum_{i=1}^8 (x_i - \bar{x})^2 = 1348$$

**Tabella 2.1** Calcolo dei termini per la varianza e la deviazione standard campionarie sui dati di Figura 2.2.

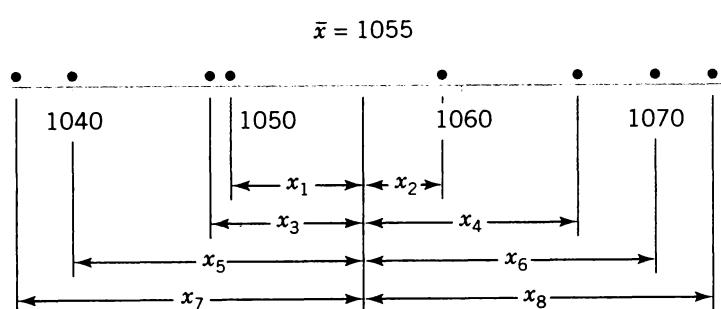


Figura 2.3 Misura della variabilità con la varianza campionaria, tramite gli scarti .

$i$	$x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	1048	-7	49
2	1059	4	16
3	1047	-8	64
4	1066	11	121
5	1046	-15	225
6	1070	15	225
7	1037	-18	324
8	1073	18	324
	8440	0.0	1348

perciò la varianza campionaria è

$$s^2 = \frac{1348}{8 - 1} = \frac{1348}{7} = 192.57 \text{ psi}^2$$

Dunque la deviazione standard campionaria risulta

$$s = \sqrt{192.57} = 13.9 \text{ psi}$$

Il calcolo di  $s^2$  richiede la determinazione di  $\bar{x}$ ,  $n$  sottrazioni e  $n$  elevamenti al quadrato nonché addizioni. Se le osservazioni originali o gli scarti  $x_i - \bar{x}$  non sono interi, può essere noioso lavorare con gli  $x_i - \bar{x}$ , e può essere necessario portarsi dietro parecchi decimali per garantire precisione numerica. Una **formula computazionale** più efficiente e rapida per il calcolo della varianza campionaria è quella che segue

**Una semplice formula per il calcolo della varianza e della deviazione standard campionarie.**

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} = \frac{\sum_{i=1}^n (x_i^2 + \bar{x}^2 - 2\bar{x}x_i)}{n - 1} = \frac{\sum_{i=1}^n x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^n x_i}{n - 1}$$

Poiché  $\bar{x} = (1/n) \sum_{i=1}^n x_i$ , la formula precedente diventa

$$s^2 = \frac{\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 / n}{n - 1} \quad (2.4)$$

Si noti che l'Equazione (2.4) richiede di elevare al quadrato ogni singola  $x_i$ , quindi elevare al quadrato la somma delle  $x_i$ , sottrarre  $(\sum x_i)^2/n$  da  $\sum x_i^2$ , e infine dividere il tutto per  $n - 1$ . Talvolta, questa formula computazionale viene detta **metodo abbreviato** per calcolare  $s^2$  (o  $s$ ).

**ESEMPIO 2.3**  
Resistenza  
degli O-ring:  
calcolo alternativo  
della varianza

Calcoliamo la varianza e la deviazione standard campionarie per i dati relativi alla resistenza alla trazione degli O-ring usando il metodo abbreviato (Equazione (2.4)).

La formula dà

$$s^2 = \frac{\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 / n}{n - 1} = \frac{8905548 - \frac{(8440)^2}{8}}{7} = \frac{1348}{7} = 192.57 \text{ psi}^2$$

e

$$s = \sqrt{192.57} = 13.9 \text{ psi}$$

Questi risultati sono in pieno accordo con quelli ottenuti precedentemente.

Analoga alla varianza campionaria  $s^2$  è la misura della variabilità in una popolazione, la **varianza della popolazione**, denotata con il simbolo  $\sigma^2$  ("sigma quadro"). La sua radice quadrata positiva,  $\sigma$ , rappresenta la **deviazione standard della popolazione**.

Quando la popolazione consiste in un numero finito  $N$  di valori, possiamo definirne la varianza della popolazione come segue

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} \quad (2.5)$$

Una definizione più generale della varianza  $\sigma^2$  verrà data in seguito. In precedenza abbiamo osservato che la media campionaria può venire usata come stima della media della popolazione; allo stesso modo, la varianza campionaria rappresenta una stima della varianza della popolazione.

Si noti che il divisore, nella formula per la varianza campionaria, è pari alla dimensione del campione,  $n$ , meno 1, mentre per la varianza di una popolazione è pari alla dimensione della popolazione,  $N$ . Conoscendo il vero valore della media della popolazione,  $\mu$ , potremmo trovare la varianza *campionaria* come scarto quadratico medio delle osservazioni nel campione intorno a  $\mu$ . In pratica il valore di  $\mu$  non è quasi mai noto, perciò bisogna usare piuttosto la somma degli scarti quadratici intorno alla media campionaria  $\bar{x}$ . Tuttavia le osservazioni  $x_i$  tendono a essere più vicine alla loro media aritmetica  $\bar{x}$  che alla media della popolazione  $\mu$ . È per compensare questo effetto che usiamo come divisore  $n - 1$  anziché  $n$ : se usassimo  $n$ , otterremmo una misura della variabilità che, in media, è nettamente più piccola della vera varianza della popolazione,  $\sigma^2$ .

Da un altro punto di vista, possiamo dire che la varianza campionaria  $s^2$  è basata su  $n - 1$  **gradi di libertà**. L'origine di questa espressione è legata al fatto che gli  $n$  scarti  $x_1 - \bar{x}$ ,  $x_2 - \bar{x}$ , ...,  $x_n - \bar{x}$  danno sempre come somma zero, per cui specificare i valori di  $n - 1$  di queste quantità porta automaticamente a determinare quella rimanente. Tutto ciò è stato illustrato in Tabella 2.1. Pertanto, solo  $n - 1$  degli  $n$  scarti  $x_i - \bar{x}$  sono liberamente determinati.

## 2.2 DIAGRAMMI RAMI E FOGLIE

Il diagramma a punti è un metodo di visualizzazione dei dati utile per campioni poco numerosi, all'incirca sino a 20 osservazioni. Quando il numero di osservazioni è relativamente alto risultano invece più utili altri tipi di visualizzazioni.

Per esempio, si considerino i dati riportati in Tabella 2.2; si tratta delle resistenze alla compressione, misurate in psi, di 80 provini di una nuova lega alluminio-litio sotto studio come possibile materiale per elementi strutturali dei velivoli. I dati sono stati registrati nell'ordine di analisi, e in questo formato non consentono di ricavare molte informazioni sulla resistenza alla compressione. Non è semplice per esempio rispondere a domande come: Quale percentuale dei provini non resiste a una pressione di 120 psi? Dato che vi sono molte osservazioni, costruire un diagramma a punti per questi dati sarebbe alquanto inefficiente; per le situazioni in cui si ha a che fare con insiemi numerosi di dati esistono visualizzazioni più efficaci.

Tramite un **diagramma rami e foglie** (o diagramma *steam and leaf*), per esempio, si riesce a ottenere una visualizzazione ricca di informazioni da un insieme di dati  $x_1, x_2, \dots, x_n$ , dove ogni numero  $x_i$  è ad almeno due cifre. Per costruire un diagramma rami e foglie si opera come segue.

**Fasi  
di costruzione  
di un  
diagramma  
rami e foglie**

1. Si divide ogni numero  $x_i$  in due parti: un **ramo**, costituito da una o più cifre significative, e una **foglia**, costituita dalla rimanente cifra.
2. Si riportano i valori dei rami in colonna.
3. Si registrano le foglie per ciascuna osservazione accanto al corrispondente ramo.
4. Si scrivono sul diagramma le unità per i rami e le foglie.

**Tabella 2.2** Resistenza alla compressione per 80 provini di lega alluminio-litio.

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

A titolo di esempio, se i dati consistono nell'informazione sulla percentuale di elementi difettosi in un lotto di wafer di semiconduttore (la percentuale essendo compresa tra 0 e 100) è possibile dividere il valore 76 nel ramo 7 e nella foglia 6. In generale, si dovrebbero scegliere pochi rami rispetto al numero di osservazioni; di solito è bene optare per un numero di rami compreso fra 5 e 20. Una volta scelto l'insieme dei rami, questi vanno disposti lungo il margine sinistro del diagramma. Accanto a ciascuno di essi, tutte le foglie corrispondenti ai valori osservati vanno elencate nell'ordine in cui si incontrano tali dati nell'insieme esaminato.

**ESEMPIO 2.4**  
**Resistenza  
alla compressione**

Per illustrare la costruzione di un diagramma rami e foglie, si considerino i dati relativi alla resistenza alla compressione di Tabella 2.2. Selezioneremo come valori di ramo i numeri 7, 8, 9, ..., 24. Il diagramma rami e foglie risultante è mostrato in Figura 2.4; l'ultima colonna contiene le frequenze del numero di foglie associate a ciascun ramo. L'osservazione del diagramma rivela immediatamente che la maggior parte delle resistenze alla compressione cade fra 110 e 200 psi, e che tra 150 e 160 psi è presente un valore centrale. Inoltre, le resistenze sono distribuite in maniera approssimativamente simmetrica intorno al valore centrale. Il diagramma rami e foglie consente quindi di determinare rapidamente alcune importanti caratteristiche dei dati, non evidenti all'esame della semplice tabella che li contiene.

Per alcuni insiemi di dati può essere utile fornire più classi o rami. Un modo per far ciò è modificare i rami originali come segue: si divide il ramo 5 (per esempio) in due nuovi rami, 5L ( $L = low = basso$ ) e 5U ( $U = up = alto$ ). Il ramo 5L ha le foglie 0, 1, 2, 3 e 4, mentre il ramo 5U ha le foglie 5, 6, 7, 8 e 9. In questo modo si raddoppia il numero dei rami originali. Si potrebbe quadruplicare quest'ultimo definendo cinque nuovi rami: 5z (foglie 0 e 1), 5t (foglie 2 e 3), 5f (foglie 4 e 5), 5s (foglie 6 e 7) e 5e (foglie 8 e 9), dove le lettere indicano l'iniziale del primo numero (in inglese).

Ramo	Foglia	Frequenza
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

Figura 2.4 Diagramma rami e foglie per i dati di resistenza alla compressione di Tabella 2.2.

### ESEMPIO 2.5 Resa di un processo chimico

La Figura 2.5 mostra il diagramma rami e foglie per 25 osservazioni sulla resa di un processo chimico. In Figura 2.5a abbiamo usato 6, 7, 8 e 9 come rami. Ciò ha portato a un numero di rami troppo basso: in questo modo il diagramma rami e foglie non fornisce sufficienti informazioni sui dati. In Figura 2.5b abbiamo diviso ogni ramo in due parti, dando vita a una visualizzazione dei dati più adeguata. La Figura 2.5c, infine, illustra un diagramma rami e foglie con ciascun ramo diviso in cinque parti. In questo grafico ci sono troppi rami, e risulta difficoltoso visualizzare la forma dei dati.

La Figura 2.6 mostra il diagramma rami e foglie relativo ai dati di resistenza alla compressione di Tabella 2.2 prodotto da Minitab. Il software usa gli stessi rami della Figura 2.4. Si osservi che il computer ordina le foglie all'interno di ogni ramo dalla più piccola alla più grande: questa visualizzazione prende il nome di **diagramma rami e foglie ordinato**. L'ordinamento non viene effettuato, in genere, quando si costruisce il diagramma manualmente, perché può richiedere troppo tempo. Il computer inoltre aggiunge una colonna alla sinistra dei rami, in cui compaiono i conteggi parziali delle osservazioni contenute nel ramo corrispondente e in quelli che lo precedono (per i rami che stanno nella metà superiore del diagramma) o che lo seguono (per quelli che si trovano nella metà inferiore); per il ramo di mezzo (il 16) il conteggio si riferisce alle sole osservazioni del ramo stesso.

Il diagramma rami e foglie ordinato rende piuttosto semplice individuare alcune caratteristiche dei dati, come i percentili, i quartili e la mediana.

La **mediana** è una misura della *tendenza centrale*, o posizione, che divide i dati in due parti uguali, la metà sotto la mediana e la metà sopra. Se il numero di osservazioni è pari, la mediana si trova a metà strada fra i due valori centrali. Dalla Figura 2.6 (in cui le osservazioni sono in totale 80) possiamo vedere che il 40-esimo e 41-esimo valore di resistenza sono 160 e 163, perciò la mediana è  $(160 + 163)/2 = 161.5$ . Se invece il numero di osservazioni è dispari, la mediana è semplicemente il valore centrale.

(a)		(b)		(c)	
Ramo	Foglie	Ramo	Foglie	Ramo	Foglie
6	1 3 4 5 5 6	6L	1 3 4	6z	1
7	0 1 1 3 5 7 8 8 9	6U	5 5 6	6t	3
8	1 3 4 4 7 8 8	7L	0 1 1 3	6f	4 5 5
9	2 3 5	7U	5 7 8 8 9	6s	6
		8L	1 3 4 4	6e	
		8U	7 8 8	7z	0 1 1
		9L	2 3	7t	3
		9U	5	7f	5
				7s	7
				7e	8 8 9
				8z	1
				8t	3
				8f	4 4
				8s	7
				8e	8 8
				9z	
				9t	2 3
				9f	5
				9s	
				9e	

Figura 2.5  
Diagramma rami  
e foglie  
per l'Esempio 2.5.

#### Diagramma rami e foglie

Rami e foglie della resistenza N = 80  
Unità foglia = 1.0

1	7	6
2	8	7
3	9	7
5	10	1 5
8	11	0 5 8
11	12	0 1 3
17	13	1 3 3 4 5 5
25	14	1 2 3 5 6 8 9 9
37	15	0 0 1 3 4 4 6 7 8 8 8
(10)	16	0 0 0 3 3 5 7 7 8 9
33	17	0 1 1 2 4 4 5 6 6 8
23	18	0 0 1 1 3 4 6
16	19	0 3 4 6 9 9
10	20	0 1 7 8

Figura 2.6  
Diagramma rami  
e foglie costruito  
con Minitab.

6	21	8
5	22	1 8 9
2	23	7
1	24	5

Il **range** o **intervallo** è una misura della variabilità facilmente calcolabile a partire dal diagramma rami e foglie ordinato: è infatti dato dalla differenza tra la misura massima e quella minima. In Figura 2.6 si trova che il range è  $245 - 76 = 169$ .

È possibile dividere i dati anche in più di due parti. Quando si divide un insieme ordinato di dati in quattro parti uguali, i punti di suddivisione sono detti **quartili**. Il *primo quartile*,  $q_1$ , è un valore che ha approssimativamente il 25% di osservazioni sotto di sé, e circa il 75% sopra. Il *secondo quartile*,  $q_2$ , ha circa il 50% delle osservazioni sotto di sé, ed è esattamente uguale alla mediana. Il *terzo quartile*,  $q_3$ , ha circa il 75% delle osservazioni sotto di sé. Come nel caso della mediana, i quartili possono non essere unici. I dati di resistenza alla compressione di Figura 2.6 contengono  $n = 80$  osservazioni. Il software Minitab calcola il primo e il terzo quartile prendendo le osservazioni ordinate di rango – ossia di “posizione” –  $(n + 1)/4$  e  $3(n + 1)/4$  ed eseguendo le interpolazioni necessarie. Per esempio, si ha  $(80 + 1)/4 = 20.25$  e  $3(80 + 1)/4 = 60.75$ . Pertanto, Minitab interpola tra la 20-esima e la 21-esima osservazione ordinata ottenendo  $q_1 = 143.50$  e fra la 60-esima e la 61-esima ottenendo  $q_3 = 181.00$ .

La differenza interquartile è una misura della variabilità.

La **differenza interquartile** (IQR, *InterQuartile Range*) è la differenza tra il terzo e il primo quartile, ed è usata a volte come misura della variabilità dei dati.

In generale, il **100k-esimo percentile** è quel valore tale che una percentuale pari a circa il  $100k\%$  delle osservazioni si trova in corrispondenza o al di sotto di esso, mentre circa il  $100(1 - k)\%$  cade al di sopra. Per esempio, per trovare il 95-esimo percentile per i dati di questo campione usiamo la formula  $0.95(80 + 1) = 76.95$  per determinare che è necessario interpolare fra le osservazioni di rango 76 e 77, ossia rispettivamente 221 e 228. Dunque circa il 95% dei dati è inferiore a 227.65 e circa il 5% è superiore. Si osservi che quando il percentile cade fra due osservazioni del campione è pratica comune usare come percentile il punto centrale delle due osservazioni (contrariamente alla procedura di interpolazione usata da Minitab). Con questa procedura semplificata si trova che il primo e terzo quartile e il 95-esimo percentile sono rispettivamente 144, 181 e 224.5. In questo testo, tuttavia, verrà usata sempre la procedura di interpolazione di Minitab.

Molti pacchetti statistici forniscono tabelle di grandezze riassuntive dei dati statistici, tra cui quelle introdotte poco sopra. In Tabella 2.3 è mostrato l'output ottenuto dai dati relativi alla resistenza alla compressione di Tabella 2.2 con l'utilizzo di Minitab. Si noti che i risultati per la mediana e per i quartili coincidono con quelli dati in precedenza. L'abbreviazione “ErrSt” indica l'errore standard della media, che verrà trattato in un capitolo successivo.

**Tabella 2.3** Statistiche riassuntive fornite da Minitab per i dati di resistenza alla compressione.

Variabile	N	Media	Mediana	DevSt	ErrSt
	80	162.66	161.50	33.77	3.78
	Min	Max	Q1	Q3	
	76.00	245.00	143.50	181.00	

## 2.3 ISTOGRAMMI

Un **istogramma** è una sintesi dei dati più compatta di un diagramma rami e foglie. Per costruire un istogramma per dati continui si deve dividere il range dei dati stessi in intervalli, detti in genere **classi** o **celle**. Se possibile, questi ultimi dovrebbero essere di pari ampiezza, in modo da migliorare l'informazione visiva fornita dall'istogramma. Se si vuole sviluppare una visualizzazione ragionevole, bisogna scegliere con attenzione il numero di classi, tenendo conto del numero di osservazioni e della dispersione dei dati: un istogramma che impieghi troppe o troppo poche classi non risulta informativo. Nella maggior parte dei casi una scelta fra 5 e 20 si rivela soddisfacente; il numero di classi dovrebbe inoltre aumentare al crescere di  $n$ . Nella pratica, funziona bene una scelta del numero di classi approssimativamente uguale alla radice quadrata del numero di osservazioni<sup>1</sup>.

Una volta stabilito il numero delle classi e i limiti inferiore e superiore di ciascuna di esse, i dati vengono assegnati alle classi, e si effettua un conteggio del numero di osservazioni in ogni classe. Per costruire l'istogramma si usa l'asse orizzontale per rappresentare la scala di misura dei dati e quello verticale per rappresentare i conteggi, o **frequenze**. A volte le frequenze di ciascuna classe sono divise per il numero totale di osservazioni ( $n$ ), quindi la scala verticale dell'istogramma rappresenta **frequenze relative**. Su ciascuna classe vengono poi costruiti dei rettangoli, di altezza proporzionale alla frequenza assoluta o relativa. La costruzione di istogrammi è una funzione fornita dalla maggior parte dei software statistici.

### ESEMPIO 2.6 Distanza caratteristica di una pallina da golf

La United States Golf Association esegue alcuni test sulle palline da golf per garantirne la conformità alle regole del gioco. Delle palline si esaminano il peso, il diametro, la rotondità e la conformità a uno standard di distanza complessiva. Quest'ultimo test è condotto colpendo le palline con una mazza manovrata da un dispositivo meccanico soprannominato *Iron Byron*, dal nome del leggendario giocatore Byron Nelson, il cui swing la macchina dovrebbe emulare. La Tabella 2.4 fornisce le distanze (in yard) raggiunte colpendo 100 palline da golf di una particolare marca nel corso del test. Poiché i dati contengono 100 osservazioni e la radice quadrata di  $n$  è quindi 10, si ritiene che circa 10 classi forniranno un istogramma soddisfacente; selezioniamo dunque l'opzione di Minitab che permette all'utente di specificare il numero di classi. L'istogramma risultante è mostrato in Figura 2.7. Si noti che il punto centrale della prima classe è 250 yard e che l'istogramma ha solo 9 classi contenenti una frequenza non nulla.

Un istogramma, così come un diagramma rami e foglie, fornisce un'impressione visiva della forma della distribuzione delle misure e informazioni sulla variabilità dei dati. Si osservi la distribuzione ragionevolmente simmetrica o a campana dei dati relativi alle distanze.

<sup>1</sup> Non c'è una regola universalmente adottata per scegliere il numero di classi di un istogramma. Alcuni manuali di statistica di base suggeriscono di adottare la *Regola di Sturges*, che pone il numero di classi uguale a  $h = 1 + \log_2 n$ , dove  $n$  è la dimensione del campione. Di questa regola vi sono molte varianti. I pacchetti di software statistico utilizzano molti differenti algoritmi per determinare il numero e l'ampiezza delle classi, alcuni dei quali non si basano sulla Regola di Sturges.

**Tabella 2.4** Dati relativi alle distanze coperte dalle palline da golf.

291.5	274.4	290.2	276.4	272.0	268.7	281.6	281.6	276.3	285.9
269.6	266.6	283.6	269.6	277.8	287.8	267.6	292.6	273.4	284.4
270.7	274.0	285.2	275.5	272.1	261.3	274.0	279.3	281.0	293.1
277.5	278.0	272.5	271.7	280.8	265.6	260.1	272.5	281.3	263.0
279.0	267.3	283.5	271.2	268.5	277.1	266.2	266.4	271.5	280.3
267.8	272.1	269.7	278.5	277.3	280.5	270.8	267.7	255.1	276.4
283.7	281.7	282.2	274.1	264.5	281.0	273.2	274.4	281.6	273.7
271.0	271.5	289.7	271.1	256.9	274.5	286.2	273.9	268.5	262.6
261.9	258.9	293.2	267.1	255.0	269.7	281.9	269.6	279.8	269.9
282.6	270.0	265.2	277.7	275.5	272.2	270.0	271.0	284.3	268.4

La maggior parte dei software statistici ha impostazioni di default per il numero di celle. In Figura 2.7 è mostrato l'istogramma di Minitab ottenuto con tali impostazioni: le classi sono 16.

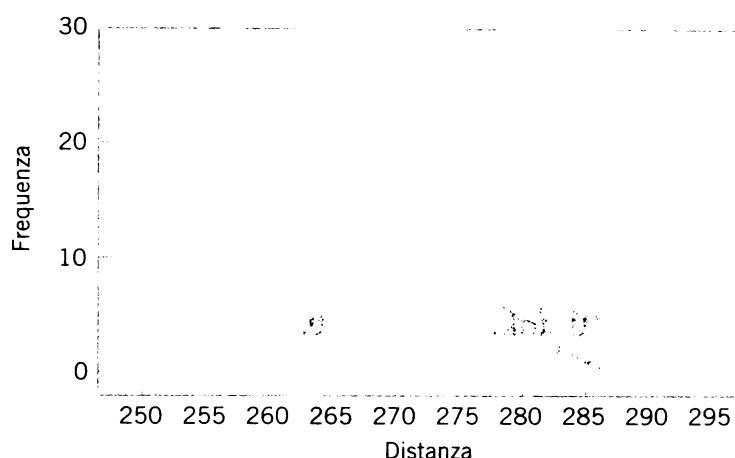


Figura 2.7 Istrogramma costruito con Minitab per i dati delle distanze di Tabella 2.4.

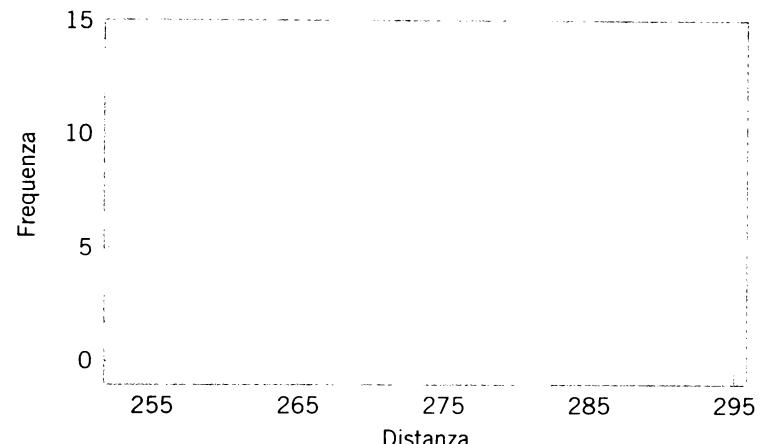


Figura 2.8 Istrogramma di Minitab per i dati delle distanze di Tabella 2.4, costruito con l'opzione di default (16 celle).

Gli istogrammi sono adatti a insiemi di dati numerosi.

Gli istogrammi possono risultare abbastanza sensibili alla scelta del numero e all'ampiezza delle classi. Per piccoli insiemi di dati gli istogrammi possono inoltre cambiare sensibilmente di aspetto se varia il numero e/o l'ampiezza delle classi. Per questa ragione preferiamo pensare all'istogramma come a una tecnica adatta a insiemi di dati **numerosi**, che contengono per esempio da 75 a 100 o più osservazioni. Essendo il numero di osservazioni dell'Esempio 2.6 abbastanza grande ( $n = 100$ ), la scelta del numero di classi non è particolarmente importante, e gli istogrammi delle Figure 2.7-2.8 forniscono informazioni analoghe.

Si noti che passando dai dati o dal diagramma rami e foglie a un istogramma si è persa in un certo senso dell'informazione, perché questo tipo di visualizzazione non preserva le osservazioni originali. Tuttavia, questo inconveniente è di solito ampiamente compensato dalla concisione e dalla facilità di interpretazione dell'istogramma, soprattutto in campioni numerosi.

Gli istogrammi risultano sempre più semplici da interpretare se le classi hanno uguale ampiezza. In caso contrario, è pratica comune disegnare rettangoli la cui *area* (anziché l'altezza) sia proporzionale al numero di osservazioni nella classe.

La Figura 2.9 mostra una variante dell'istogramma disponibile in Minitab: il **grafico delle frequenze cumulate**. In esso, l'altezza di ciascuna barra rappresenta il numero di osservazioni il cui valore è minore o uguale al limite superiore della classe. Le frequenze cumulate sono spesso molto utili nell'interpretazione dei dati. Per esempio, dalla Figura 2.9 possiamo leggere direttamente che circa 15 palline da golf su 100 sottoposte al test hanno percorso una distanza superiore alle 280 yard.

Le distribuzioni di frequenza e gli istogrammi possono venire impiegati anche con dati qualitativi o categorici, o con conteggi (variabili discrete). In alcune applicazioni vi sarà un ordinamento naturale delle categorie (è quanto accade per esempio con i mesi dell'anno), mentre in altre l'ordinamento sarà arbitrario (per esempio: maschi/femmine). Quando si costruisce un istogramma su dati categorici, le ampiezze delle classi devono essere uguali.

Per realizzare un istogramma per dati discreti si determina innanzitutto la frequenza (assoluta o relativa) per ciascun valore di  $x$ , che corrisponde a una classe. Si riportano quindi le frequenze sull'asse verticale e i valori di  $x$  su quello orizzontale. Infine si disegna sopra ogni valore di  $x$  un rettangolo, avente altezza pari alla frequenza corrispondente a quel valore.

### ESEMPIO 2.7 Randy Johnson

Durante il campionato di baseball della stagione 2002, Randy Johnson della squadra Arizona Diamondbacks si è guadagnato la *triple crown* per i lanciatori vincendo 24 incontri, eliminando 334 battitori avversari e realizzando una media di *run* guadagnati pari a 2.32. La Tabella 2.5 riporta un riassunto gara per gara della performance di Johnson per tutti e 35 gli incontri in cui egli è stato il pitcher iniziale. La Figura 2.10 contiene un istogramma degli strikeout dell'atleta. Si noti che il numero di strikeout è una variabile discreta.

In base ai dati tabulati o all'istogramma possiamo calcolare quanto segue

$$\text{frazione di incontri con almeno 10 strikeout} = \frac{15}{35} = 0.4286$$

$$\text{frazione di incontri con } 8 \div 14 \text{ strikeout} = \frac{22}{35} = 0.6286$$

Questi rapporti sono esempi di **frequenze relative**.

Un'importante variante dell'istogramma è la **carta di Pareto**, ampiamente utilizzata negli studi sul miglioramento della qualità e dei processi, in cui i dati rappresentano di solito diversi tipi di difetti, modalità di rottura o altre categorie interessanti per l'analista. Queste categorie vengono ordinate in modo da disporre quella con la maggiore frequenza a sinistra, seguita dalla categoria con la frequenza immediatamente inferiore, e via dicendo. Le carte di questo tipo prendono il nome dall'economista italiano V. Pareto, ed esibiscono in genere la **legge di Pareto**, secondo la quale la maggior parte dei difetti rientra perlopiù in un numero limitato di categorie.

### ESEMPIO 2.8 Incidenti aerei

La Tabella 2.6 riporta i dati relativi agli incidenti aerei tratti da un articolo del Wall Street Journal ("Jet's Troubled History Raises Issues for the FAA and the Manufacturer") del 19 settembre 2000. La tabella mostra il numero totale di incidenti che hanno comportato la perdita del velivolo avvenuti fra il 1959 e il 1999 per 22 tipi di aereo, e il numero di velivoli persi ogni milione di decolli. La Figura 2.11 rappresenta una carta di Pareto delle perdite di velivoli per milione di decolli. Chiaramente, i primi tre modelli di aereo sono coinvolti in un'alta percentuale degli incidenti. Un fatto interessante che li riguarda è che il 707/720 e il

Tabella 2.5 Performance del lanciatore Randy Johnson, stagione 2002.

DATA	AVVERSARIO	PUNTEGGIO	IP	H	R	ER	HR	BB	SO
4/1	San Diego	W, 2-0 (C)	9.0	6	0	0	0	1	8
4/6	@ Milwaukee	W, 6-3	7.0	5	1	1	1	3	12
4/11	@ Colorado	W, 8-4	7.0	3	2	2	0	2	9
4/16	St. Louis	W, 5-3	7.0	8	3	3	1	1	5
4/21	Colorado	W, 7-1(C)	9.0	2	1	0	0	1	17
4/26	@ Florida	W, 5-3	7.0	4	1	1	0	3	10
5/6	Pittsburgh	L, 2-3	7.0	7	3	2	1	0	8
5/11	@ Philadelphia	ND, 6-5 (10)	7.0	8	4	4	2	2	8
5/16	Philadelphia	W, 4-2	7.0	6	1	1	1	4	8
5/21	San Francisco	W, 9-4	7.0	6	3	3	0	3	10
5/26	Los Angeles	ND, 10-9 (10)	5.0	8	7	7	3	2	5
5/31	@ Los Angeles	W, 6-3	8.0	6	3	0	1	1	4
6/5	Houston	ND, 5-4 (13)	8.0	6	3	3	1	0	11
6/10	@ N.Y. Yankees	L, 5-7	7.2	7	5	5	2	3	8
6/15	Detroit	W, 3-1	7.0	7	1	0	0	2	13
6/20	Baltimore	W, 5-1	7.0	5	1	1	1	2	11
6/26	@ Houston	W, 9-1	8.0	3	0	0	0	3	8
7/1	Los Angeles	L, 0-4	7.0	9	4	3	0	0	6
7/6	San Francisco	ND, 2-3	7.0	7	2	2	1	2	10
7/11	@ Los Angeles	ND, 4-3	6.0	6	3	3	2	2	5
7/16	@ San Francisco	W, 5-3	7.0	5	3	3	2	3	7
7/21	@ San Diego	L, 9-11	5.0	8	8	8	1	6	9
7/26	San Diego	W, 12-0	7.0	4	0	0	0	1	8
7/31	@ Montreal	W, 5-1 (C)	9.0	8	1	1	0	3	15
8/5	@ New York	W, 2-0 (C)	9.0	2	0	0	0	2	11
8/10	Florida	W, 9-2	8.0	5	2	2	1	2	14
8/15	@ Cincinnati	W, 7-2	8.0	2	2	1	1	2	11
8/20	Cincinnati	ND, 5-3	7.0	5	2	2	1	3	12
8/25	Chicago	W, 7-0 (C)	9.0	6	0	0	0	2	16
8/30	San Francisco	L, 6-7	5.1	9	7	6	0	3	6
9/4	Los Angeles	W, 7-1 (C)	9.0	3	1	1	1	0	8
9/9	San Diego	W, 5-2	7.0	8	1	1	1	3	7
9/14	Milwaukee	W, 5-0 (C)	9.0	3	0	0	0	2	17
9/19	@ San Diego	W, 3-1	7.0	4	1	1	1	0	9
9/26	Colorado	W, 4-2 (C)	9.0	6	2	0	0	2	8
<b>Totali stagione</b>		24-5, 2.32	260.0	197	78	67	26	71	334

**Legenda:** W = incontri vinti; L = incontri persi; ND = incontri pareggiati; C = incontro completo; IP = inning in battuta; H = battute; R = run; ER = run guadagnati; BB = base su palla; SO = strikeout.

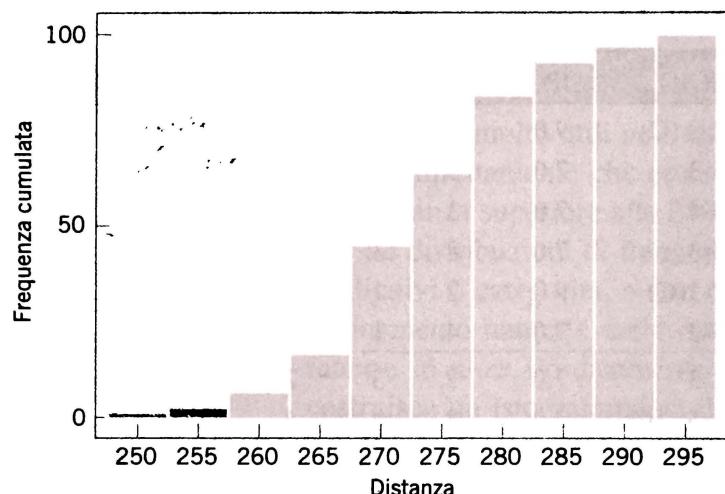


Figura 2.9 Grafico delle frequenze cumulate fornito da Minitab per i dati delle distanze di Tabella 2.4.

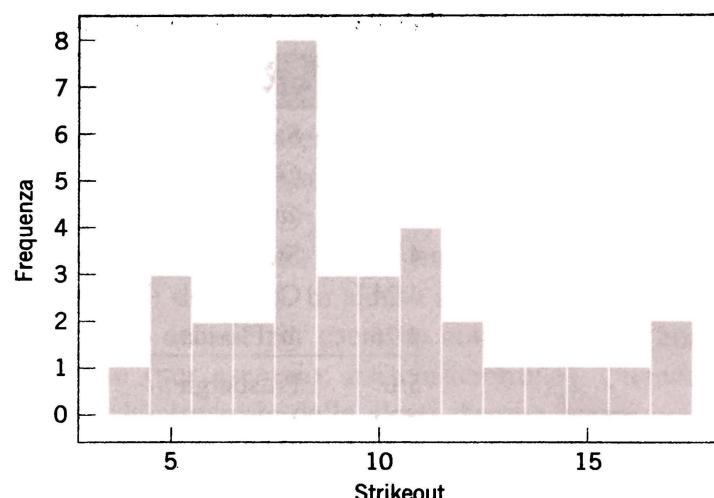


Figura 2.10 Istogramma del numero di strikeout realizzati da Randy Johnson nella stagione di baseball 2002.

Tabella 2.6 Dati statistici sugli incidenti aerei.

Modello di aereo	Numero effettivo di velivoli persi	Velivoli persi/milione di decolli
MD-11	5	6.54
707/720	115	6.46
DC-8	71	5.84
F-28	32	3.94
BAC 1-11	22	2.64
DC-10	20	2.57
747-Prima serie	21	1.90
A310	4	1.40
A300-600	3	1.34
DC-9	75	1.29
A300-Prima serie	7	1.29
737-1 & 2	62	1.23
727	70	0.97
A310/319/321	7	0.96
F100	3	0.80
L-1011	4	0.77
BAe 146	3	0.59
747-400	1	0.49
757	4	0.46
MD-80/90	10	0.43
767	3	0.41
737-3, 4 & 5	12	0.39

DC-8 erano progetti della metà degli anni Cinquanta, non più in servizio passeggeri regolare oggigiorno, mentre l'MD-11 è stato introdotto nel servizio passeggeri nel 1990. Fra il 1990 e il 1999 cinque esemplari della flotta complessiva di 198 MD-11 furono distrutti in incidenti aerei; ne risulta per questo modello un alto tasso di incidenti (un'eccellente discussione delle potenziali cause prime di questi incidenti si trova nell'articolo citato del Wall Street Journal). Lo scopo della maggior parte delle carte di Pareto è di aiutare l'analista a distinguere le poche fonti di difetto o di incidente importanti da quelle – più numerose – poco significative. Esistono molte varianti delle carte di Pareto, per alcuni esempi si veda Montgomery (2009a).

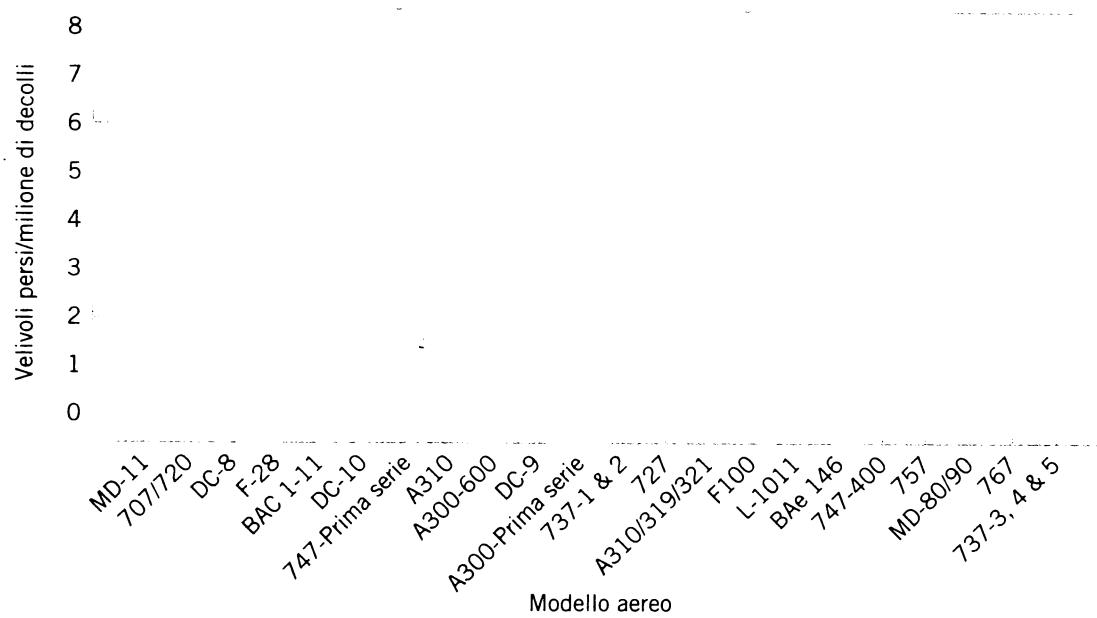


Figura 2.11 Carta di Pareto relativa ai dati statistici sugli incidenti aerei.

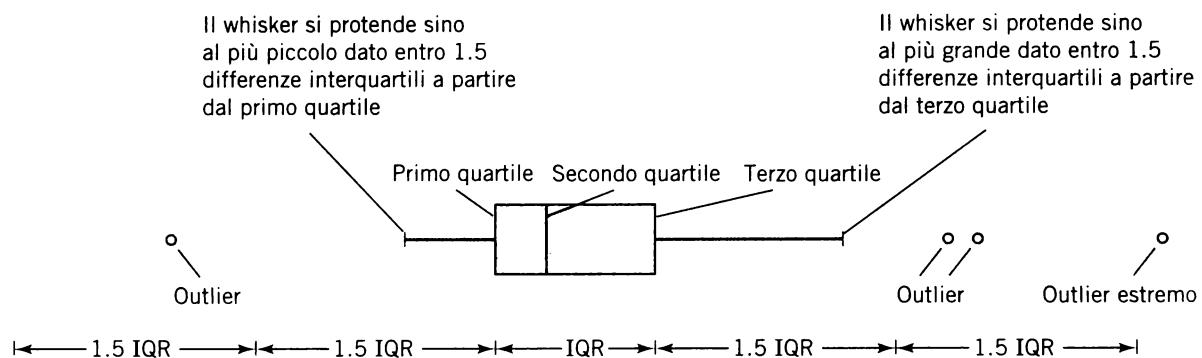
## 2.4 BOX PLOT

I diagrammi rami e foglie e gli istogrammi forniscono informazioni visive generali sull'insieme dei dati in esame, mentre funzioni numeriche come  $\bar{x}$  o  $s$  danno informazioni su una sola caratteristica dei dati. Tramite un **box plot** si è invece in grado di descrivere simultaneamente più caratteristiche importanti degli insiemi di dati, quali il centro, la dispersione, lo scostamento dalla simmetria e l'identificazione di osservazioni che cadono insolitamente lontano dal gruppo principale di dati (queste osservazioni vengono dette *outlier* o *valori erratici*).

Un box plot (Figura 2.12) rappresenta i tre quartili su una “scatola” rettangolare, allineati orizzontalmente oppure verticalmente. Il box racchiude l'intera differenza interquartile, con il lato sinistro (o quello inferiore) in corrispondenza del primo quartile  $q_1$ , e il lato destro (o quello superiore) in corrispondenza del terzo quartile  $q_3$ . Si traccia quindi un segmento trasversale al box in corrispondenza del secondo quartile (che è il 50-esimo percentile o mediana). All'esterno dei lati opposti del box si protendono due segmenti, detti **whisker** o **baffi**. Il whisker sinistro o inferiore va dal primo quartile all'osservazione più piccola entro 1.5 differenze interquartili a partire da  $q_1$ . Il whisker destro o superiore va dal terzo quartile sino all'osservazione più grande entro 1.5 differenze interquartili a partire da  $q_3$ . I

dati che cadono a una distanza superiore a quella dei whisker del box plot vengono rappresentati come singoli punti. Un punto che si trova al di là del whisker, ma a meno di 3 differenze interquartili dal lato del box viene detto **outlier** o **valore erratico**. Un punto che cade al di là di 3 differenze interquartili è detto **outlier estremo**. A volte si possono trovare su questo tipo di grafico simboli di altro tipo, come cerchietti pieni o vuoti, a identificare i due tipi di outlier.

Figura 2.12  
Descrizione  
di un box plot.



La Figura 2.13 mostra il box plot ottenuto da Minitab per i dati relativi alla resistenza alla compressione di Tabella 2.2. Esso indica che la distribuzione delle resistenze alla compressione presenta una certa simmetria intorno al valore centrale, perché i whisker sinistro e destro e le lunghezze delle parti sinistra e destra del box intorno alla mediana hanno circa le stesse dimensioni. Ci sono inoltre outlier da ciascun lato del grafico.

I box plot risultano molto utili nei confronti grafici fra insiemi di dati, poiché hanno un elevato impatto visivo e sono facili da interpretare. Per esempio, la Figura 2.14 mostra i box plot di confronto per un indice di qualità produttiva relativo a dispositivi a semiconduttore in tre stabilimenti di produzione. L'esame di questi grafici rivela che c'è eccessiva variabilità nell'impianto 2 e che gli stabilimenti 2 e 3 devono incrementare le loro performance sugli indici.

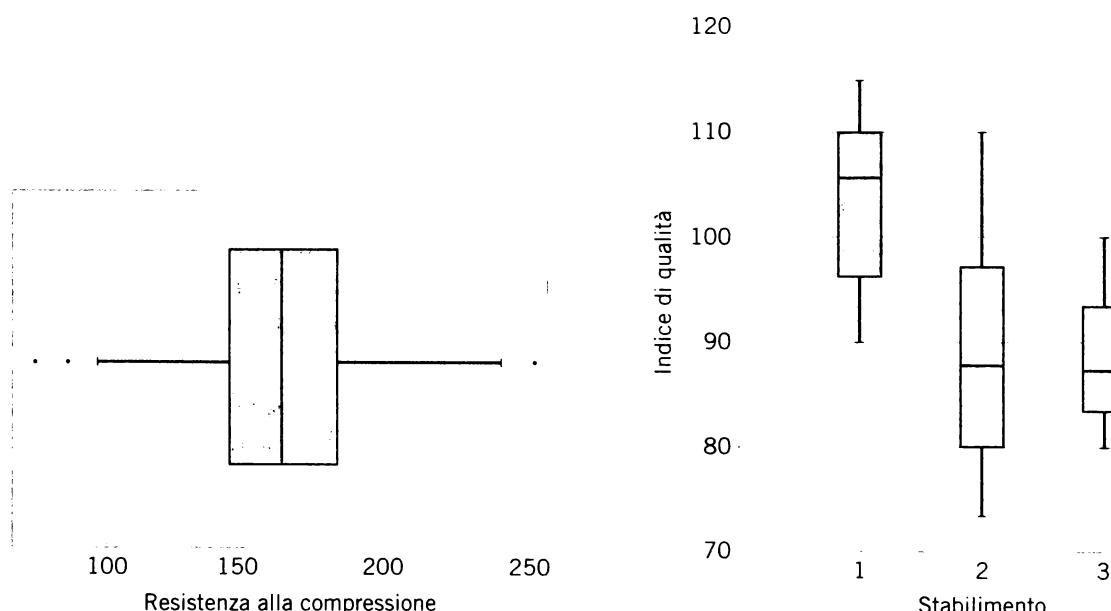


Figura 2.13 Box plot per i dati di Tabella 2.2.

Figura 2.14 Box plot di confronto relativi a un indice di qualità in tre stabilimenti di produzione.

## 2.5 GRAFICI DELLE SERIE STORICHE

Le visualizzazioni grafiche esaminate sinora (istogrammi, diagrammi rami e foglie e box plot) sono metodi visivi molto utili per mostrare la variabilità dei dati raccolti. Tuttavia, abbiamo osservato nel Capitolo 1 che un fattore importante che contribuisce alla variabilità dei dati è il tempo, che i suddetti grafici non tengono in considerazione.

Una **serie storica** o **sequenza temporale** è un insieme di dati in cui le osservazioni sono registrate nell'ordine in cui vengono effettuate. Il **grafico di una serie storica** è un grafico in cui sull'asse verticale sono riportati i valori osservati della variabile, mentre l'asse orizzontale rappresenta la scala dei tempi (minuti, giorni, anni ecc.). Quando le misurazioni sono riportate in grafico come serie storiche è spesso possibile osservare andamenti regolari, cicli o altre caratteristiche dei dati altrimenti non rilevabili.

Per esempio, si consideri la Figura 2.15a, che rappresenta la serie storica delle vendite annuali di un'azienda, relativa agli anni 1982-1991. L'impressione generale che se ne ricava è che le vendite mostrano una **tendenza** alla crescita. C'è una certa variabilità in questo andamento, dal momento che le vendite di alcuni anni crescono rispetto all'anno precedente, quelle di altri diminuiscono. La Figura 2.15b riporta le vendite degli anni 1989-1991 su base trimestrale. Il grafico mostra chiaramente che le vendite annuali presentano una variabilità ciclica trimestrale, nel senso che le vendite del primo e secondo trimestre di ogni anno sono generalmente superiori a quelle del terzo e quarto trimestre.

A volte può risultare utile combinare un grafico di serie storica con qualche altro tipo di grafico tra quelli visti in precedenza. J. Stuart Hunter (*The American Statistician*, vol. 42, 1988, p. 54) ha suggerito di combinare il diagramma rami e foglie con un grafico di serie storica per formare un **grafico digidot**.

La Figura 2.16 mostra un grafico digidot per le osservazioni sulla resistenza alla compressione di Tabella 2.2, nell'ipotesi che le misure siano state registrate nell'ordine in cui sono state effettuate. Esso rappresenta efficacemente la variabilità complessiva dei dati e contemporaneamente ne mostra la variabilità nel tempo. L'impressione generale è che la resistenza alla compressione vari intorno al valore medio di 162.67, e che non vi sia un evidente comportamento regolare della variabilità nel tempo.

Il grafico digidot di Figura 2.17 racconta una storia differente. Esso riassume 30 osservazioni sulla concentrazione di un prodotto di un processo chimico, registrate a inter-

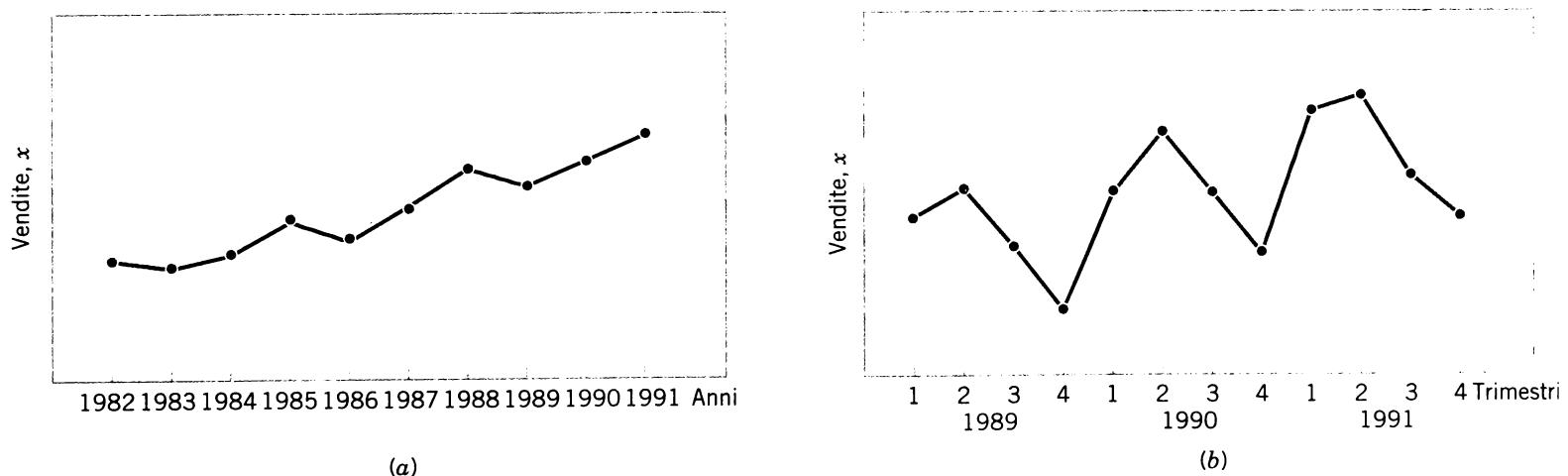


Figura 2.15 Vendite registrate per anno (a) e per trimestre (b).

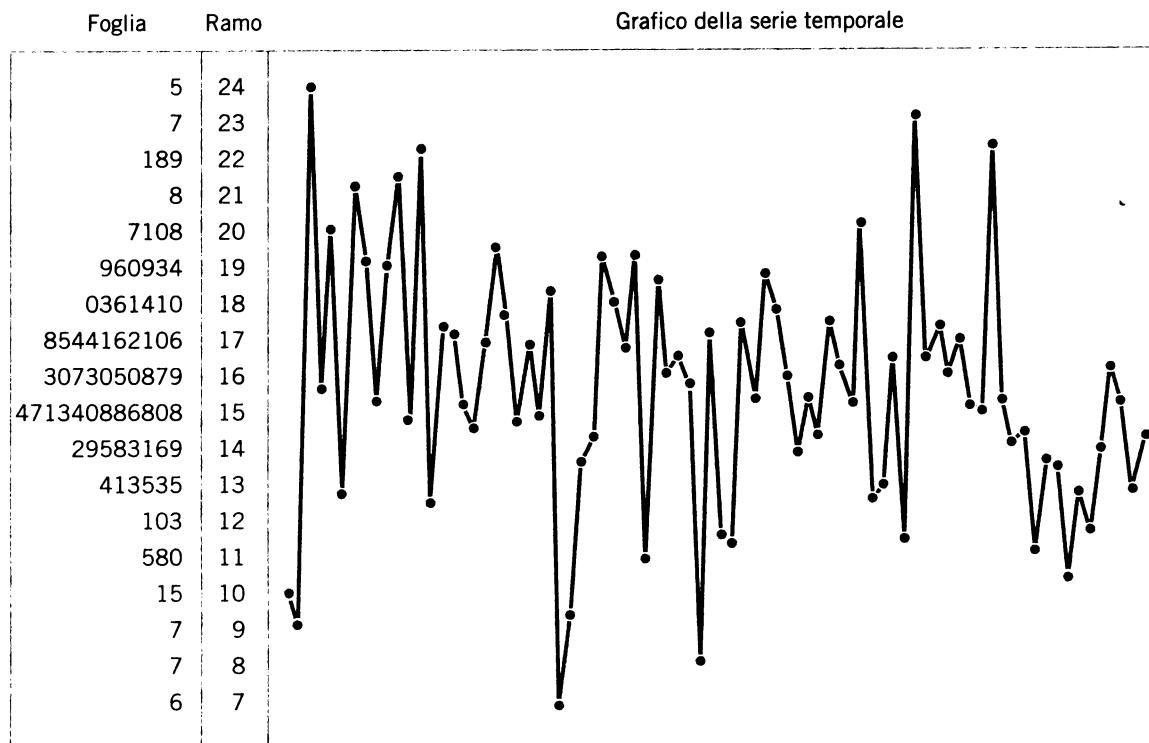


Figura 2.16 Grafico digidot per i dati di resistenza alla compressione (Tabella 2.2).

valli regolari di un'ora. Il grafico indica che durante le prime 20 ore di funzionamento si sono ottenute concentrazioni generalmente al di sopra degli 85 g/l, ma che in seguito è intervenuto qualche fattore che ha portato a concentrazioni più basse. Se si riuscisse a ridurre la variabilità, il processo potrebbe essere reso più efficiente. La carta di controllo relativa a questi dati, che è un tipo particolare di grafico delle serie storiche, è stata mostrata in Figura 1.19.

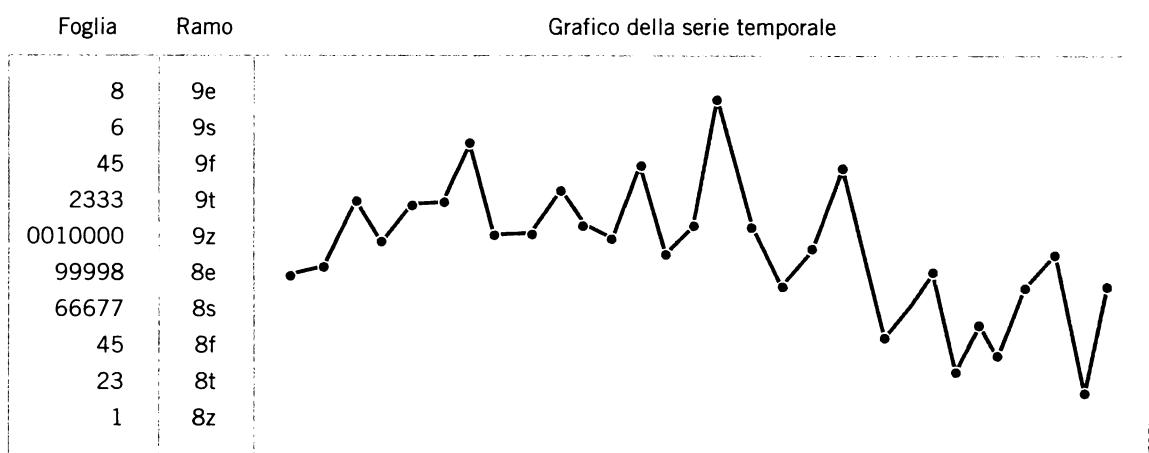


Figura 2.17 Grafico digidot per le osservazioni sulla concentrazione in un processo chimico, rilevate ogni ora.

## 2.6 DATI MULTIVARIATI

Il diagramma a punti, il diagramma rami e foglie, l'istogramma e i box plot sono metodi visivi per descrivere dati **univariati**; in altri termini, essi forniscono informazioni descrittive su una singola variabile. Molti problemi di ingegneria, però, comportano la raccolta e l'analisi di **dati multivariati**, ovvero dati relativi a più variabili. Il problema della colonna di distillazione discusso nel Paragrafo 1.2 e il problema del filo di giunzione visto nel Paragrafo 1.3 sono tipici esempi di studi di ingegneria che coinvolgono dati multivariati. In Tabella 2.7 sono ripetuti per comodità i dati esaminati nel Paragrafo 1.3, relativi alla trazione unitaria delle giunzioni. In ingegneria, spesso l'obiettivo da perseguire è la determinazione delle relazioni fra le variabili o la costruzione di un modello empirico, come discusso nel Paragrafo 1.3.

Tabella 2.7 Dati relativi al filo di giunzione nei semiconduttori (esempio del Paragrafo 1.3).

Osservazione Numero	Traz. unitaria <i>y</i>	Lungh. filo <i>x</i> <sub>1</sub>	Spess. piastrina <i>x</i> <sub>2</sub>	Osservazione Numero	Traz. unitaria <i>y</i>	Lungh. filo <i>x</i> <sub>1</sub>	Spess. piastrina <i>x</i> <sub>2</sub>
1	9.95	2	50	14	11.66	2	360
2	24.45	8	110	15	21.65	4	205
3	31.75	11	120	16	17.89	4	400
4	35.00	10	550	17	69.00	20	600
5	25.02	8	295	18	10.30	1	585
6	16.86	4	200	19	34.93	10	540
7	14.38	2	375	20	46.59	15	250
8	9.60	2	52	21	44.88	15	290
9	24.35	9	100	22	54.12	16	510
10	27.50	8	300	23	56.63	17	590
11	17.08	4	412	24	22.13	6	100
12	37.00	11	400	25	21.15	5	400
13	41.95	12	500				

Il diagramma di dispersione introdotto nel Paragrafo 1.3 è un semplice strumento descrittivo per dati multivariati; è utile per esaminare le relazioni a coppie per le variabili in un insieme di dati multivariati. I diagrammi di dispersione per i dati di Tabella 2.7 sono mostrati in Figura 2.18; sono stati costruiti con Minitab, specificando l'opzione che porta alla visualizzazione, a margine, dei box plot relativi a ciascuna singola variabile.

Come già osservato nel Paragrafo 1.3, entrambi i diagrammi di dispersione trasmettono l'impressione che vi possa essere una relazione approssimativamente lineare sia fra la trazione unitaria della giunzione e la lunghezza del filo, sia fra la trazione unitaria e lo spessore della piastrina; tale relazione appare essere maggiore nel primo caso che nel secondo.

Il grado di linearità di una relazione fra due variabili *y* e *x* può venire espressa mediante il **coefficiente di correlazione campionario** *r*. Supponiamo di avere *n* coppie di osservazioni su due variabili (*y*<sub>1</sub>, *x*<sub>1</sub>), (*y*<sub>2</sub>, *x*<sub>2</sub>), ..., (*y*<sub>*n*</sub>, *x*<sub>*n*</sub>). Sarebbe logico affermare che tra *y* e *x* esiste una relazione positiva se a valori elevati di *y* corrispondono valori elevati di *x* e a valori bassi di *y* corrispondono valori bassi di *x*. Analogamente, esiste una relazione negativa tra le variabili se a valori elevati di *y* corrispondono valori bassi di *x* e a valori bassi di *y* corrispondono valori elevati di *x*. La somma dei prodotti incrociati data da

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right) / n$$

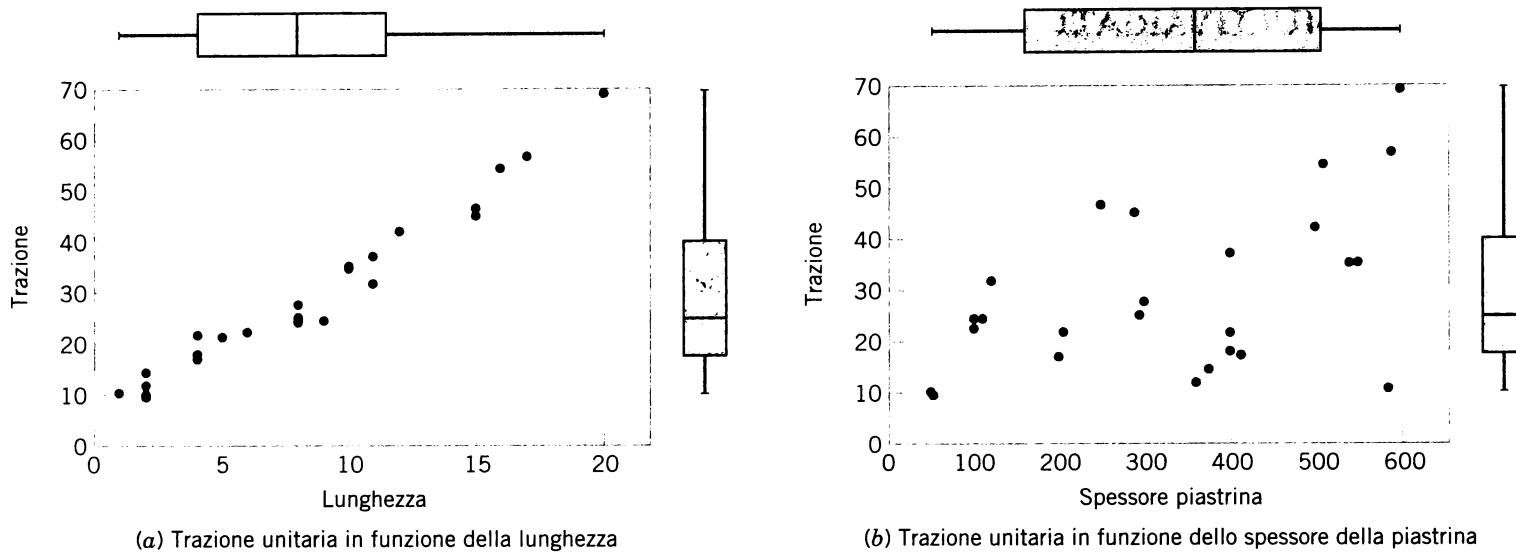


Figura 2.18 Diagrammi di dispersione per i dati relativi al filo di giunzione nei semiconduttori (Tabella 2.7). (a) Trazione unitaria in funzione della lunghezza; (b) Trazione unitaria in funzione dello spessore della piastrina.

è in grado di riflettere questi tipi di relazioni. Per capirne la ragione, si supponga che la relazione tra  $y$  e  $x$  sia fortemente positiva, come nel caso di Figura 2.18a. In questa situazione, un valore di  $x_i$  sopra la media  $\bar{x}$  tenderà a presentarsi assieme a un valore di  $y_i$  sopra la media  $\bar{y}$ , di modo che il prodotto  $(x_i - \bar{x})(y_i - \bar{y})$  sarà positivo. Lo stesso accadrà quando sia  $x_i$  sia  $y_i$  sono al di sotto delle rispettive medie. Di conseguenza, una relazione positiva fra  $y$  e  $x$  implica che  $S_{xy}$  sarà positiva. Un ragionamento simile porta a concludere che quando la relazione è negativa la maggior parte dei prodotti  $(x_i - \bar{x})(y_i - \bar{y})$  sarà negativa, per cui lo sarà anche  $S_{xy}$ .

Ora, le unità di  $S_{xy}$  sono quelle di  $x$  moltiplicate per quelle di  $y$ , perciò sarebbe difficile interpretare  $S_{xy}$  come una misura del grado di linearità di una relazione perché un cambiamento delle unità di  $x$  e/o di  $y$  influenzerebbe il valore di  $S_{xy}$ . Il coefficiente di correlazione campionario  $r$  scala semplicemente  $S_{xy}$  per produrre una quantità adimensionale.

### Coefficiente di correlazione campionario

Siano  $(y_1, x_1), (y_2, x_2), \dots, (y_n, x_n)$   $n$  coppie di dati; il **coefficiente di correlazione campionario  $r$**  è definito da

$$r = \frac{S_{xy}}{\sqrt{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)\left(\sum_{i=1}^n (y_i - \bar{y})^2\right)}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \quad (2.6)$$

con  $-1 \leq r \leq +1$

Il valore  $r = +1$  è ottenibile soltanto se tutte le osservazioni giacciono esattamente su una retta con pendenza positiva; analogamente,  $r = -1$  si ottiene solo se tutte le osservazioni giacciono su una retta con pendenza negativa. Pertanto,  $r$  misura il grado di linearità della relazione fra  $y$  e  $x$ . Quando il suo valore è prossimo a zero può indicare che non esiste relazione fra le due variabili, oppure l'assenza di una relazione lineare (Figura 2.19).

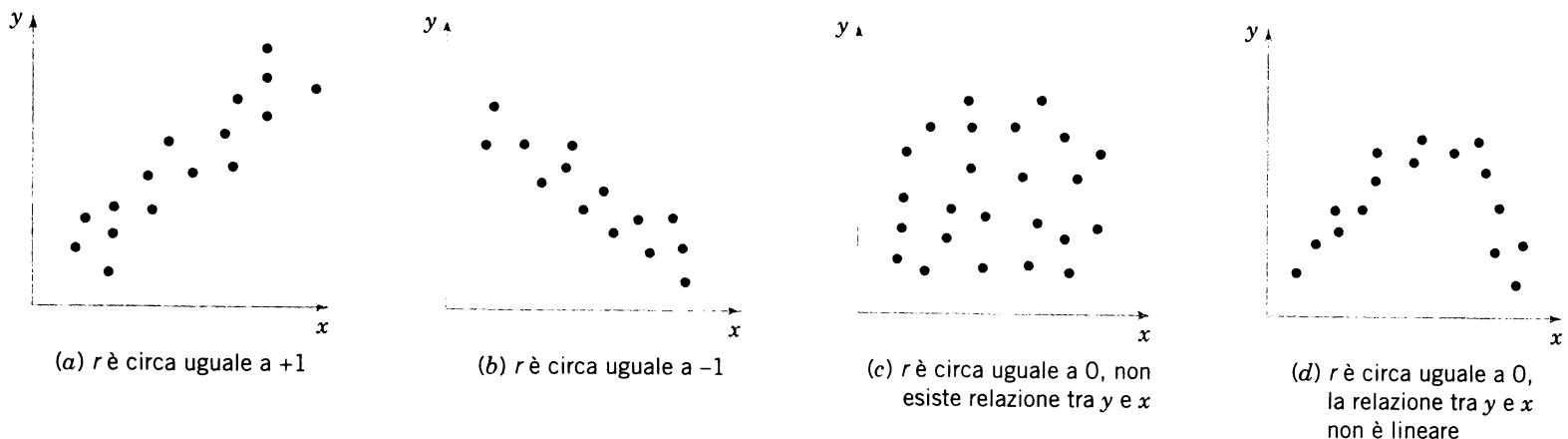


Figura 2.19 Diagrammi di dispersione per diversi valori del coefficiente di correlazione campionario  $r$ . (a)  $r$  è circa uguale a +1; (b)  $r$  è circa uguale a -1; (c)  $r$  è circa uguale a 0; non esiste relazione tra  $y$  e  $x$ ; (d)  $r$  è circa uguale a 0; la relazione tra  $y$  e  $x$  non è lineare.

A titolo di esempio, vediamo i calcoli necessari a ottenere il coefficiente di correlazione campionario fra la trazione unitaria e la lunghezza del filo. In base ai dati di Tabella 2.7 troviamo:

**Calcolo del coefficiente di correlazione campionario.**

$$\sum_{i=1}^{25} x_i^2 = 2396 \quad \sum_{i=1}^{25} x_i = 206 \quad \sum_{i=1}^{25} y_i^2 = 27179 \quad \sum_{i=1}^{25} y_i = 725.82 \quad \sum_{i=1}^{25} x_i y_i = 8008.5$$

$$S_{xx} = \sum_{i=1}^n x_i^2 - \frac{\left( \sum_{i=1}^n x_i \right)^2}{n} = 2396 - \frac{(206)^2}{25} = 698.56$$

$$S_{yy} = \sum_{i=1}^n y_i^2 - \frac{\left( \sum_{i=1}^n y_i \right)^2}{n} = 27179 - \frac{(725.82)^2}{25} = 6106.91$$

$$S_{xy} = \sum_{i=1}^n x_i y_i - \frac{\left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{n} = 8008.5 - \frac{(206)(725.82)}{25} = 2027.74$$

Il coefficiente di correlazione campionario fra le due grandezze è dunque

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{2027.74}{\sqrt{(698.56)(6106.91)}} = 0.982$$

Calcoli simili portano a stabilire che il coefficiente di correlazione fra la trazione unitaria e l'altezza della piastrina è  $r = 0.493$ .

In linea generale si ritiene che esista una forte correlazione fra due variabili se il coefficiente di correlazione è compreso fra 0.8 e 1, una debole correlazione se è compreso fra 0 e 0.5, e una moderata correlazione negli altri casi. Pertanto, possiamo dire che fra la trazione unitaria e la lunghezza del filo esiste una forte correlazione, mentre quella fra trazione unitaria e spessore della piastrina è relativamente debole.

### ESEMPIO 2.9 La temperatura globale

La Tabella 2.8 mostra le variazioni della temperatura media globale e le concentrazioni globali di CO<sub>2</sub> registrate negli anni 1880-2004. Si tratta dei dati mostrati sotto forma di serie storica in Figura 2.1. La Figura 2.20 è un diagramma di dispersione della variazione di temperatura rispetto alla concentrazione di anidride carbonica; tale grafico mostra una correlazione positiva fra le due variabili. Applicando a questi dati l'Equazione (2.6), si trova che il coefficiente di correlazione campionario è  $r = 0.852$ , a indicare una correlazione moderatamente forte. Si deve stare attenti a non dedurre troppo da questa correlazione positiva; la correlazione non implica una **causalità**. Vi sono molti esempi di variabili che sono fortemente correlate ma tra cui non esiste una relazione di causa ed effetto. Per stabilire un rapporto di causalità servono degli esperimenti pianificati, in cui si variano deliberatamente i livelli o i valori di una variabile e si osservano le variazioni di un'altra variabile.

Vi sono molti altri utili metodi grafici per visualizzare dati multivariati. Per illustrarne l'utilizzo, si considerino i dati relativi a uno shampoo riportati in Tabella 2.9, ottenuti in un'indagine scientifica. Le variabili schiuma, profumo, colore e residuo (una misura della capacità di pulizia del prodotto) sono proprietà descrittive valutate, nell'esperimento, con un punteggio da 1 a 10. La qualità è una misura dell'attrattività complessiva dello shampoo, ed è la variabile di risposta nominale che interessa allo sperimentatore. Il parametro "Regione" è un indicatore che specifica se il prodotto è stato realizzato in uno stabilimento dell'Est (1) o dell'Ovest (2) degli Stati Uniti.

La Figura 2.21 rappresenta una matrice di diagrammi di dispersione per i dati di Tabella 2.9, prodotta da Minitab. Esso illustra le relazioni fra tutte le variabili, considerate a coppie, di Tabella 2.9. I singoli diagrammi di dispersione della matrice mostrano che vi può essere una relazione positiva fra la qualità dello shampoo e la schiuma, e relazioni negative fra qualità e profumo e tra qualità e regione.

Vi possono essere anche correlazioni fra alcune delle proprietà, come il colore e il residuo. Minitab è in grado di calcolare anche tutte le correlazioni fra le variabili, prese a due a due. I risultati sono i seguenti:

	Schiuma	Profumo	Colore	Residuo	Regione
Profumo	0.002				
Colore	0.328	0.599			
Residuo	0.193	0.500	0.524		
Regione	-0.032	0.278	0.458	0.165	
Qualità	0.505	-0.240	-0.195	-0.487	-0.512

Si noti che nessuna delle correlazioni è forte.

La Figura 2.22 mostra un diagramma di dispersione qualità/schiuma per lo shampoo oggetto d'indagine. In questa rappresentazione sono stati usati diversi simboli per identificare le osservazioni associate alle differenti aree geografiche di produzione; in tal modo si riescono a fornire informazioni su *tre* variabili su un grafico bidimensionale. La figura rivelava che la relazione tra qualità e schiuma può essere differente nelle due regioni. Un altro modo di asserire ciò è dire che vi può essere un'**interazione** tra schiuma e regione (può essere utile a questo punto rileggere quanto abbiamo detto nel Paragrafo 1.2.3). Ovviamen-

Tabella 2.8 Variazioni della temperatura media globale e concentrazioni di CO<sub>2</sub>, anni 1880-2004.

Anno	Var. temp.	Conc. CO <sub>2</sub>	Anno	Var. temp.	Conc. CO <sub>2</sub>	Anno	Var. temp.	Conc. CO <sub>2</sub>
1880	-0.11	290.7	1922	-0.09	303.8	1964	-0.25	319.2
1881	-0.13	291.2	1923	-0.16	304.1	1965	-0.15	320.0
1882	-0.01	291.7	1924	-0.11	304.5	1966	-0.07	321.1
1883	-0.04	292.1	1925	-0.15	305.0	1967	-0.02	322.0
1884	-0.42	292.6	1926	0.04	305.4	1968	-0.09	322.9
1885	-0.23	293.0	1927	-0.05	305.8	1969	0.00	324.2
1886	-0.25	293.3	1928	0.01	306.3	1970	0.04	325.2
1887	-0.45	293.6	1929	-0.22	306.8	1971	-0.10	326.1
1888	-0.23	293.8	1930	-0.03	307.2	1972	-0.05	327.2
1889	0.04	294.0	1931	0.03	307.7	1973	0.18	328.8
1890	-0.22	294.2	1932	0.04	308.2	1974	-0.06	329.7
1891	-0.55	294.3	1933	-0.11	308.6	1975	-0.02	330.7
1892	-0.40	294.5	1934	0.05	309.0	1976	-0.21	331.8
1893	-0.39	294.6	1935	-0.08	309.4	1977	0.16	333.3
1894	-0.32	294.7	1936	0.01	309.8	1978	0.07	334.6
1895	-0.32	294.8	1937	0.12	310.0	1979	0.13	336.9
1896	-0.27	294.9	1938	0.15	310.2	1980	0.27	338.7
1897	-0.15	295.0	1939	-0.02	310.3	1981	0.40	339.9
1898	-0.21	295.2	1940	0.14	310.4	1982	0.10	341.1
1899	-0.25	295.5	1941	0.11	310.4	1983	0.34	342.8
1900	-0.05	295.8	1942	0.10	310.3	1984	0.16	344.4
1901	-0.05	296.1	1943	0.06	310.2	1985	0.13	345.9
1902	-0.30	296.5	1944	0.10	310.1	1986	0.19	347.2
1903	-0.35	296.8	1945	-0.01	310.1	1987	0.35	348.9
1904	-0.42	297.2	1946	0.01	310.1	1988	0.42	351.5
1905	-0.25	297.6	1947	0.12	310.2	1989	0.28	352.9
1906	-0.15	298.1	1948	-0.03	310.3	1990	0.49	354.2
1907	-0.41	298.5	1949	-0.09	310.5	1991	0.44	355.6
1908	-0.30	298.9	1950	-0.17	310.7	1992	0.16	356.4
1909	-0.31	299.3	1951	-0.02	311.1	1993	0.18	357.0
1910	-0.21	299.7	1952	0.03	311.5	1994	0.31	358.9
1911	-0.25	300.1	1953	0.12	311.9	1995	0.47	360.9
1912	-0.33	300.4	1954	-0.09	312.4	1996	0.36	362.6
1913	-0.28	300.8	1955	-0.09	313.0	1997	0.40	363.8
1914	-0.02	301.1	1956	-0.18	313.6	1998	0.71	366.6
1915	0.06	301.4	1957	0.08	314.2	1999	0.43	368.3
1916	-0.20	301.7	1958	0.10	314.9	2000	0.41	369.5
1917	-0.46	302.1	1959	0.05	315.8	2001	0.56	371.0
1918	-0.33	302.4	1960	-0.02	316.6	2002	0.70	373.1
1919	-0.09	302.7	1961	0.10	317.3	2003	0.66	375.6
1920	-0.15	303.0	1962	0.05	318.1	2004	0.60	377.4
1921	-0.04	303.4	1963	0.03	318.7			

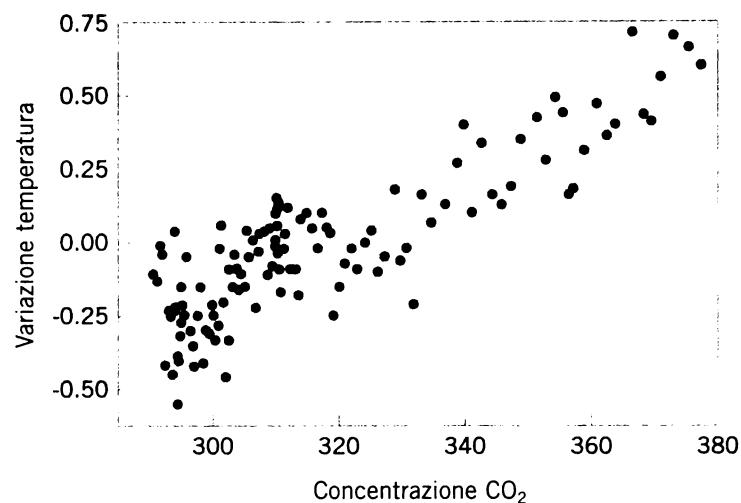
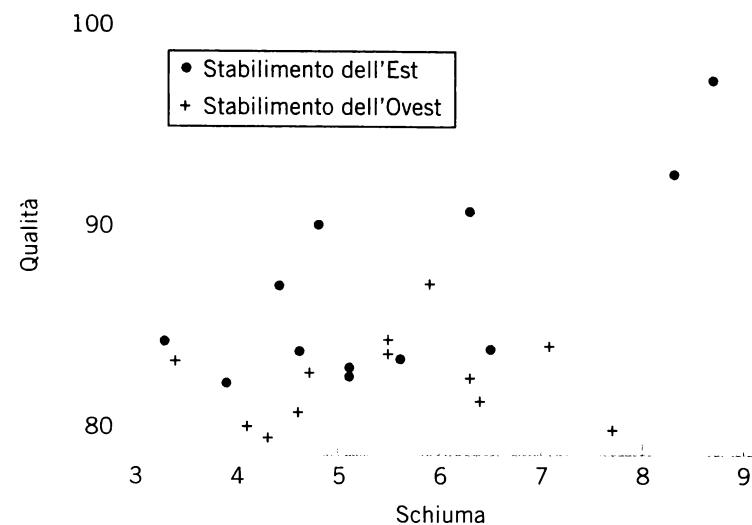
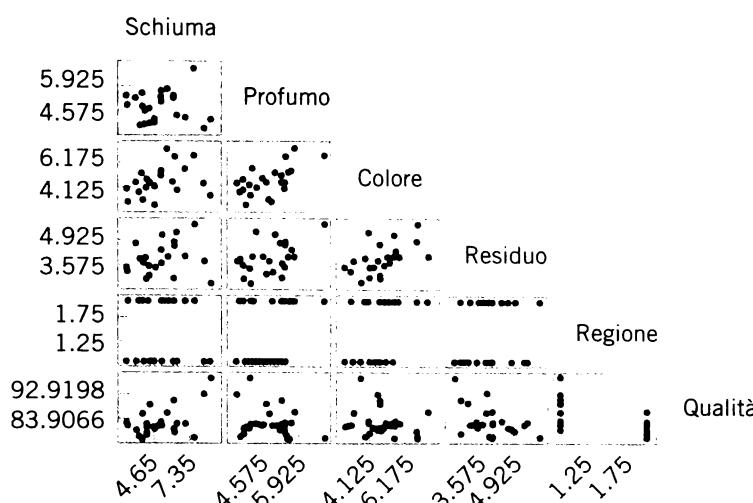


Figura 2.20 Diagramma di dispersione della temperatura media globale rispetto alla concentrazione globale di CO<sub>2</sub>

Tabella 2.9 Dati relativi all'indagine sullo shampoo.

Schiuma	Profumo	Colore	Residuo	Regione	Qualità
6.3	5.3	4.8	3.1	1	91
4.4	4.9	3.5	3.9	1	87
3.9	5.3	4.8	4.7	1	82
5.1	4.2	3.1	3.6	1	83
5.6	5.1	5.5	5.1	1	83
4.6	4.7	5.1	4.1	1	84
4.8	4.8	4.8	3.3	1	90
6.5	4.5	4.3	5.2	1	84
8.7	4.3	3.9	2.9	1	97
8.3	3.9	4.7	3.9	1	93
5.1	4.3	4.5	3.6	1	82
3.3	5.4	4.3	3.6	1	84
5.9	5.7	7.2	4.1	2	87
7.7	6.6	6.7	5.6	2	80
7.1	4.4	5.8	4.1	2	84
5.5	5.6	5.6	4.4	2	84
6.3	5.4	4.8	4.6	2	82
4.3	5.5	5.5	4.1	2	79
4.6	4.1	4.3	3.1	2	81
3.4	5.0	3.4	3.4	2	83
6.4	5.4	6.6	4.8	2	81
5.5	5.3	5.3	3.8	2	84
4.7	4.1	5.0	3.7	2	83
4.1	4.0	4.1	4.0	2	80



te, questa tecnica può venire estesa a più di tre variabili definendo opportuni simboli da utilizzare nel grafico.

La variante del diagramma di dispersione di Figura 2.22 funziona bene quando la terza variabile è **discreta** o **categorica**; quando invece è continua, può essere utile un **coplot**. Un coplot per i dati relativi allo shampoo è mostrato in Figura 2.23. In esso la qualità è riportata in grafico rispetto alla schiuma e, come in Figura 2.22, sono usati differenti simboli per identificare le due regioni di produzione. La variabile descrittiva “residuo” di Tabella 2.9 non è necessariamente una caratteristica desiderabile dello shampoo, e alti livelli di residuo ricevono un punteggio fra 4 e 4.5 o più in Tabella 2.9. Il grafico di Figura 2.23a usa tutte le osservazioni di Tabella 2.9 per le quali la variabile “residuo” è minore o uguale a 4.6, quello di Figura 2.23b usa tutte le osservazioni per le quali tale residuo è maggiore o uguale a 4. Si noti che nella costruzione del coplot c’è

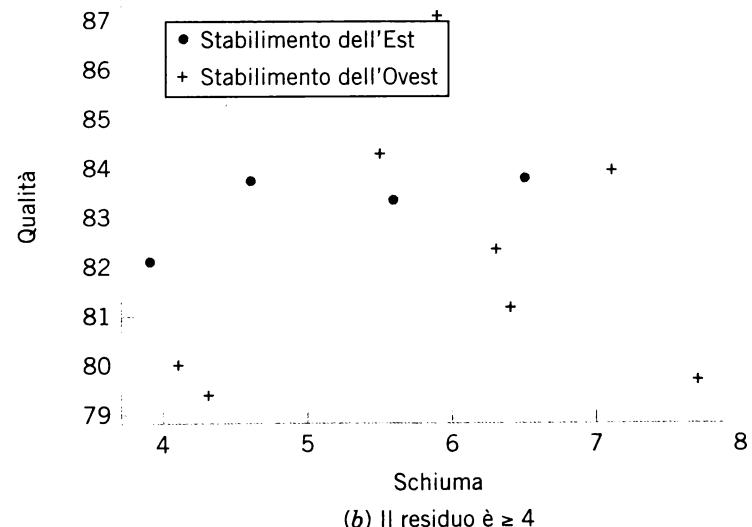
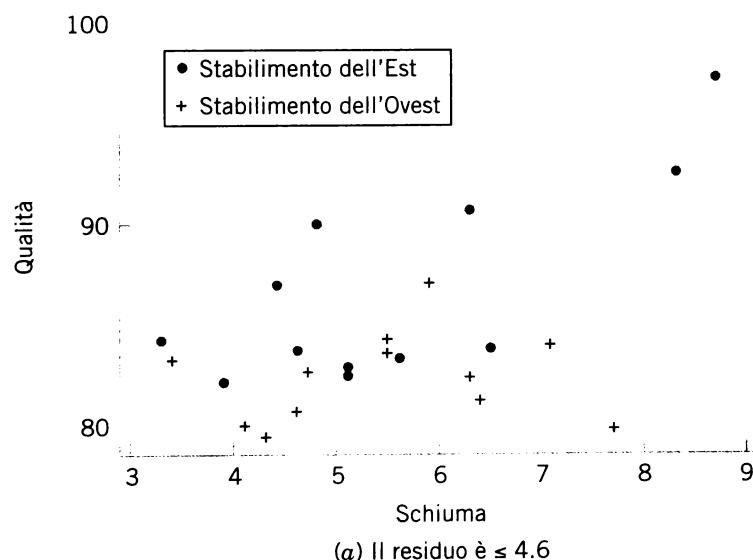


Figura 2.23 Un coplot per i dati relativi allo shampoo. (a) Il residuo è  $\leq 4.6$ ; (b) il residuo è  $\geq 4$ .

una sovrapposizione nei valori del residuo; ciò è assolutamente accettabile. L'uso di più di due categorie può essere altrettanto utile, in alcuni problemi. Il coplot indica che la relazione positiva fra qualità dello shampoo e schiuma è molto più forte per bassi valori di residuo; ciò probabilmente segnala che un residuo eccessivo non dà sempre luogo a un buono shampoo.

## TERMINI E CONCETTI RILEVANTI

---

<b>Box plot</b>	<b>Gradi di libertà</b>
<b>Carta di Pareto</b>	<b>Grafico delle serie storiche</b>
<b>Coefficiente di correlazione campionario</b>	<b>Grafico digidot</b>
<b>Dati multivariati</b>	<b>Istogramma</b>
<b>Dati univariati</b>	<b>Matrice di diagrammi di dispersione</b>
<b>Deviazione standard campionaria, <math>s</math></b>	<b>Media campionaria, <math>\bar{x}</math></b>
<b>Deviazione standard della popolazione, <math>\sigma</math></b>	<b>Media della popolazione, <math>\mu</math></b>
<b>Diagramma a punti</b>	<b>Mediana</b>
<b>Diagramma di dispersione</b>	<b>Percentile</b>
<b>Diagramma rami e foglie</b>	<b>Quartili</b>
<b>Diagramma rami e foglie ordinato</b>	<b>Range</b>
<b>Differenza interquartile (IQR)</b>	<b>Serie storica</b>
<b>Frequenza</b>	<b>Varianza campionaria, <math>s^2</math></b>
<b>Frequenza relativa</b>	<b>Varianza della popolazione, <math>\sigma^2</math></b>

# Esercizi proposti

---

## ESERCIZI PER IL PARAGRAFO 2.1

---

2.1. Un'importante caratteristica dell'acqua è la concentrazione di particelle solide sospese, espressa in mg/l. Dodici misurazioni di tale concentrazione effettuate su un campione d'acqua di un lago hanno dato come esito: 42.4, 65.7, 29.8, 58.7, 52.1, 55.8, 57.0, 68.7, 67.3, 67.3, 54.3, 54.0. Calcolare la media e la deviazione standard campionarie. Costruire quindi un diagramma a punti dei dati.

2.2. Allo scopo di valutare la qualità nel processo produttivo, vengono studiate sette misurazioni dello spessore di ossido in wafer di semiconduttore. I dati (in angstrom) sono: 1264, 1280, 1301, 1300, 1292, 1307 e 1275. Calcolare la media e la deviazione standard campionarie. Costruire quindi un diagramma a punti dei dati.

2.3. Un articolo pubblicato su *Human Factors* nel giugno 1989 presenta i dati relativi all'accomodamento visivo (una funzione del movimento dell'occhio) nell'atto del riconoscimento di una forma a punti luminosi su uno schermo a tubo catodico ad alta risoluzione. I dati rilevati sono: 36.45, 67.90, 38.77, 42.18, 26.72, 50.77, 39.30, 49.71. Calcolare la media e la deviazione standard campionarie. Costruire quindi un diagramma a punti dei dati.

2.4. Si supponga che tutti i dipendenti di un'azienda ricevano un aumento di 200 euro al mese. Che effetti vi sono sulla paga media mensile in tale azienda? Che effetti vi sono sulla deviazione standard della paga mensile?

2.5. La media campionaria è sempre uguale a uno dei valori del campione? Giustificare la risposta o portare un controsempio.

2.6. I risultati di un insieme di misure (esprese in cm) sono i seguenti: 20.1, 20.5, 20.3, 20.5, 20.6, 20.1, 20.2, 20.4. Calcolare la media campionaria e la deviazione standard campionaria. Si supponga quindi che le misure siano convertite in pollici (1 in = 2.54 cm); come vengono modificati tali indici statistici?

2.7. Si supponga che tutti i dipendenti di un'azienda ricevano un aumento annuo della paga pari al 5% dello stipendio attuale. Che effetti vi sono sulla paga media annuale in tale azienda? Che effetti vi sono sulla deviazione standard della paga annuale?

## ESERCIZI PER IL PARAGRAFO 2.2

---

 2.8. I dati che seguono sono i numeri di cicli che portano alla rottura di provini in alluminio sottoposti a ripetuti stress alternati a 21000 psi, 18 cicli al secondo:

1115	1567	1223	1782	1055
1310	1883	375	1522	1764
1540	1203	2265	1792	1330
1502	1270	1910	1000	1608
1258	1015	1018	1820	1535
1315	845	1452	1940	1781
1085	1674	1890	1120	1750
798	1016	2100	910	1501
1020	1102	1594	1730	1238
865	1605	2023	1102	990

2130            706            1315            1578            1468  
1421            2215            1269            758            1512  
1109            785            1260            1416            1750  
1481            885            1888            1560            1642

(a) Costruire un diagramma rami e foglie per tali dati.  
(b) Risulta verosimile che un provino "sopravviva" oltre i 2000 cicli? Giustificare la risposta.

 2.9. I dati che seguono si riferiscono ai risultati di 90 misurazioni consecutive relative a substrati ceramici cui è stato applicato un rivestimento metallico mediante metallizzazione per vaporizzazione. Costruire un diagramma rami e foglie per tali dati.

94.1	87.3	94.1	92.4	84.6	85.4	90.4	86.4	94.7	82.6	96.1	86.4
93.2	84.1	92.1	90.6	83.6	86.6	89.1	87.6	91.1	83.1	98.0	84.5
90.6	90.1	96.4	89.1	85.4	91.7						
91.4	95.2	88.2	88.8	89.7	87.5						
88.2	86.1	86.4	86.4	87.6	84.2						
86.1	94.3	85.0	85.1	85.1	85.1						
95.1	93.2	84.9	84.0	89.6	90.5						
90.0	86.7	78.3	93.7	90.0	95.6						
92.4	83.0	89.6	87.7	90.1	88.3						
87.3	95.3	90.3	90.6	94.3	84.1						
86.6	94.1	93.1	89.4	97.3	83.7						
91.2	97.8	94.6	88.6	96.8	82.9						
86.1	93.1	96.3	84.1	94.4	87.3						

2.10. Trovare la mediana, i quartili e il 5° e 95-esimo percentile per i dati dell'Esercizio 2.8.

2.11. Trovare la mediana, i quartili e il 5° e 95-esimo percentile per i dati dell'Esercizio 2.9.

2.12. Si sono ottenute le seguenti cinque osservazioni: 20.25, 21.38, 22.75, 20.89, 25.50. Si supponga che l'ultima osservazione venga erroneamente registrata come 255.0. Che effetti vi sono sulla paga media annuale in tale azienda? Che effetti ha questo errore sulla media e sulla deviazione standard campionarie? E sulla mediana?

## ESERCIZI PER IL PARAGRAFO 2.3

2.13. Costruire un grafico delle frequenze cumulate e un istogramma con i dati dell'Esercizio 2.8.

2.14. Costruire un grafico delle frequenze cumulate e un istogramma con i dati dell'Esercizio 2.9.

2.15. In un'indagine statistica sui difetti strutturali delle portiere d'automobile si sono ottenuti i seguenti risultati:

incisioni, 4; parti assemblate non in sequenza, 6; parti non sufficientemente rifilate, 21; fori/fenditure mancanti, 8; parti non lubrificate, 5; parti con profilo errato, 30; parti non smussate, 3. Costruire e interpretare la relativa carta di Pareto.

## ESERCIZI PER IL PARAGRAFO 2.4

2.16. I seguenti dati sono le temperature (espresso in gradi Fahrenheit) relative ai giunti degli O-ring presenti nel razzo vettore dello Space Shuttle, misurate in accensioni di prova o in lanci effettivi (tratti dalla *Presidential Commission on the Space Shuttle Challenger Accident*, vol. 1, pp. 129-131):  
84, 49, 61, 40, 83, 67, 45, 66, 70, 69, 80, 58, 68, 60, 67, 72, 73, 70, 57, 63, 70, 78, 52, 67, 53, 67, 75, 61, 70, 81, 76, 79, 75, 76, 58, 31.

- (a) Calcolare la media e la deviazione standard campionarie.
- (b) Trovare il primo e terzo quartile della temperatura.
- (c) Trovare la mediana.
- (d) Scartare l'osservazione più piccola (31 °F) ed eseguire nuovamente i calcoli effettuati ai punti (a), (b) e (c). Commentare i risultati ottenuti. Quanto sono "differenti" le altre temperature da tale valore minimo?
- (e) Costruire un box plot dei dati e commentare l'eventuale presenza di outlier.

2.17. I seguenti dati rappresentano le temperature di scarico in un impianto di trattamento dei fanghi, rilevate in giorni consecutivi:

43	47	51	48	52	50	46	49
45	52	46	51	44	49	46	51
49	45	44	50	48	50	49	50

- (a) Calcolare la media e la mediana campionarie.
- (b) Calcolare la varianza e la deviazione standard campionarie.
- (c) Costruire un box plot dei dati e commentare le informazioni da esso deducibili.
- (d) Trovare il quinto e il 95-esimo percentile delle temperature.

2.18. Vengono registrate e ordinate 18 misure del flusso di prodotto (espresso in  $\text{cm}^3/\text{s}$ ) in un impianto chimico:

6.50	6.77	6.91	7.38	7.64	7.74	7.90	7.91	8.21
8.26	8.30	8.31	8.42	8.53	8.55	9.04	9.33	9.36

- (a) Calcolare la media e la varianza campionarie.
- (b) Trovare il primo e terzo quartile campionari.
- (c) Trovare la mediana.
- (d) Costruire un box plot dei dati.
- (e) Trovare il quinto e il 95-esimo percentile.

## ESERCIZI PER IL PARAGRAFO 2.5

**2.19.** Nel loro manuale *Time Series Analysis, Forecasting, and Control* (Holden-Day, 1976), G.E.P. Box e G.M. Jenkins presentano le letture di concentrazione relative a un processo chimico effettuate ogni 2 ore. Alcuni dei dati sono riportati nella tabella a destra (vanno letti dall'alto verso il basso, quindi da sinistra a destra).

Costruire e interpretare un grafico digidot oppure un diagramma rami e foglie con grafico della serie storica separati.

17.0	16.6	16.3	16.1	17.1	16.9	16.8	17.4
17.1	17.0	16.7	17.4	17.2	17.4	17.4	17.0
17.3	17.2	17.4	16.8	17.1	17.4	17.4	17.5
17.4	17.6	17.4	17.3	17.0	17.8	17.5	18.1
17.5	17.4	17.4	17.1	17.6	17.7	17.4	17.8
17.6	17.5	16.5	17.8	17.3	17.3	17.1	17.4
16.9					17.3		

 **2.20.** In Tabella 2.10 è mostrata la serie storica dal 1770 al 1869 dei *numeri di Wolf* relativi alla quantità annuale delle macchie solari.

- (a) Costruire un grafico della serie storica per tali dati.
- (b) Costruire e interpretare un grafico digidot oppure un diagramma rami e foglie con grafico della serie storica separato.

**Tabella 2.10** Numeri di Wolf.

1770	101	1795	21	1820	16	1845	40
1771	82	1796	16	1821	7	1846	62
1772	66	1797	6	1822	4	1847	98
1773	35	1798	4	1823	2	1848	124
1774	31	1799	7	1824	8	1849	96
1775	7	1800	14	1825	17	1850	66
1776	20	1801	34	1826	36	1851	64
1777	92	1802	45	1827	50	1852	54
1778	154	1803	43	1828	62	1853	39
1779	125	1804	48	1829	67	1854	21
1780	85	1805	42	1830	71	1855	7
1781	68	1806	28	1831	48	1856	4
1782	38	1807	10	1832	28	1857	23
1783	23	1808	8	1833	8	1858	55
1784	10	1809	2	1834	13	1859	94
1785	24	1810	0	1835	57	1860	96
1786	83	1811	1	1836	122	1861	77
1787	132	1812	5	1837	138	1862	59
1788	131	1813	12	1838	103	1863	44
1789	118	1814	14	1839	86	1864	47
1790	90	1815	35	1840	63	1865	30
1791	67	1816	46	1841	37	1866	16
1792	60	1817	41	1842	24	1867	7
1793	47	1818	30	1843	11	1868	37
1794	41	1819	24	1844	15	1869	74

## ESERCIZI PER IL PARAGRAFO 2.6

 2.21. Per individuare un adatto sostituto biodegradabile degli imballaggi per cibi pronti è importante stabilire le proprietà dei materiali. Si considerino i seguenti dati sulla densità del prodotto (in g/cm<sup>3</sup>) e sulla conducibilità termica (in W/mK), pubblicati in *Materials Research and Innovation* (1999, pp. 2-8) riportati nella tabella a destra.

- (a) Costruire un diagramma di dispersione dei dati e fare una previsione sul segno del coefficiente di correlazione.
- (b) Calcolare e interpretare il coefficiente di correlazione campionario.

Conducibilità termica $y$	Densità $x$
0.0480	0.1750
0.0525	0.2200
0.0540	0.2250
0.0535	0.2260
0.0570	0.2500
0.0610	0.2765

 2.22. Per studiare il rendimento di un carburante sono stati raccolti i seguenti dati.

Miglia/gallone	Peso	HP	Miglia/gallone	Peso	HP
$y$	$x_1$	$x_2$	$y$	$x_1$	$x_2$
29.25	2464	130	17.00	4024	394
21.00	3942	235	17.00	3495	294
32.00	2604	110	18.50	4300	362
21.25	3241	260	16.00	4455	389
26.50	3283	200	10.50	3726	485
23.00	2809	240	12.50	3522	550

- (a) Costruire due diagrammi di dispersione dei dati e fare una previsione sul segno di ciascun coefficiente di correlazione.
- (b) Calcolare e interpretare i due coefficienti di correlazione campionari.

## ESERCIZI DI FINE CAPITOLO

2.23. Un operatore misura il pH di una soluzione otto volte, usando lo stesso strumento, ottenendo i seguenti dati: 7.15, 7.20, 7.18, 7.19, 7.21, 7.20, 7.16, 7.18.

- (a) Calcolare la media campionaria. Si supponga che il valore desiderabile per questa soluzione sia 7.20. La media appena calcolata è sufficientemente vicina al valore obiettivo per ritenere che la soluzione sia conforme alla specifica? Giustificare la risposta.
- (b) Calcolare la varianza e la deviazione standard campionarie. Quali sono i principali fattori di variabilità in questo esperimento? Perché è bene avere una varianza bassa delle misure?

2.24. In uno studio sul processo di ossidazione della naftalina in fase di vapore, la percentuale molare di naftalina convertita in anidride maleica è la seguente: 4.2, 4.7, 4.7, 5.0, 3.8, 3.6, 3.0, 5.1, 3.1, 3.8, 4.8, 4.0, 5.2, 4.3, 2.8, 2.0, 2.8, 3.3, 4.8, 5.0.

- (a) Calcolare il range del campione, la varianza e la deviazione standard campionarie.
- (b) Calcolare nuovamente quanto richiesto al punto (a), ma sottraendo prima 1.0 da ciascuna osservazione. Confrontare i risultati ottenuti nei due casi. C'è qualcosa di "speciale" sulla costante 1.0, o qualsiasi altro valore scelto arbitrariamente avrebbe prodotto i medesimi risultati?

2.25. **La media troncata.** Si supponga di disporre di alcuni dati ordinati per valore crescente, e che il  $T\%$  delle osservazioni venga rimosso da ciascun estremo del range, calcolando quindi la media campionaria dei restanti dati. La grandezza che ne risulta si chiama *media troncata*. Essa ha un valore generalmente compreso fra la media e la mediana campionarie. Perché?

- (a) Calcolare la media troncata al 10% per i dati dell'Esercizio 2.9.
- (b) Calcolare la media troncata al 20% per gli stessi dati e confrontare il risultato con quello del punto (a).

- (c) Confrontare i valori ottenuti ai punti (a) e (b) con la media e la mediana campionarie per i medesimi dati. C'è molta differenza fra tali grandezze? Perché?
- (d) Si supponga che la dimensione campionaria  $n$  sia tale per cui  $nT/100$  non è un intero. Sviluppare una procedura per ottenere la media troncata in casi come questo.

2.26. Un canale di comunicazione è sottoposto a controllo per determinare il numero di errori in una sequenza di 1000 bit. Di seguito sono riportati i dati per 20 di tali sequenze.

3	2	4	1	3	1	3	1	0	1
3	2	0	2	0	1	1	1	2	3

- (a) Costruire un diagramma rami e foglie dei dati.  
 (b) Trovare la media e la deviazione standard campionarie.  
 (c) Costruire un grafico della serie storica. Esso indica un aumento o una diminuzione del numero di errori in una sequenza? Giustificare la risposta.

2.27. In un campione, è vero che esattamente metà delle osservazioni cadono sempre al di sotto della media? Giustificare la risposta o portare un controesempio.

2.28. Dato un arbitrario insieme di dati numerici, è possibile che la deviazione standard campionaria sia maggiore della media campionaria? Giustificare la risposta o portare un controesempio.



# Variabili aleatorie e distribuzioni di probabilità

---

## IL PRIMO GOAL SEGNATO

I commentatori sportivi sono soliti discutere se in sport come l'hockey o il calcio, in cui si segna poco, la squadra che realizza un goal per prima abbia o no più possibilità di vincere dell'altra. Due ricercatori del Royal Military College del Canada hanno sviluppato un metodo per studiare questo argomento dal punto di vista statistico (si veda l'articolo “Can Mathematicians Spot the Winning Team Better Than Sports Commentators?”, [www.sciencedaily.com/releases/2009/06/090602112301.htm](http://www.sciencedaily.com/releases/2009/06/090602112301.htm)).

Nei gironi eliminatori dell'hockey esistono meno differenze di livello fra due squadre, rispetto a quanto accade nella stagione regolare, per cui i ricercatori hanno fatto l'ipotesi che in questo contesto ciascuna squadra abbia le stesse possibilità di vittoria, pari al 50%. Hanno tuttavia scoperto che se una squadra riesce a segnare nei primi 5 minuti dell'incontro le sue possibilità di vittoria crescono sino al 70%. Una squadra che segna il primo goal nella seconda frazione di gioco, quando rimangono solo 25 minuti al termine, porta invece le proprie chance di vittoria all'80%.

I ricercatori hanno scoperto inoltre che il numero complessivo di goal segue una distribuzione di Poisson e, poiché i valori in campo sono vicini e la motivazione è alta per tutti, ogni squadra ha la stessa probabilità di segnare un'altra rete dopo la prima. Altri fattori, come la classifica e l'andamento delle prestazioni nella stagione, sono stati presi in considerazione dai ricercatori. La loro indagine ha richiesto la comprensione e l'impiego di molti concetti di statistica, tra cui le distribuzioni esponenziali, di Poisson e binomiali.

In ingegneria ci si trova ad affrontare problemi simili quando si devono creare modelli di eventi discreti o modelli per misure continue. I concetti probabilistici spiegati in questo capitolo sono strumenti importanti per gli ingegneri.

---

## CONTENUTI DEL CAPITOLO

3.1 INTRODUZIONE	3.7.2 Funzione di distribuzione cumulativa
3.2 VARIABILI ALEATORIE	3.7.3 Media e varianza
3.3 PROBABILITÀ	3.8 DISTRIBUZIONE BINOMIALE
3.4 VARIABILI ALEATORIE CONTINUE	3.9 PROCESSO DI POISSON
3.4.1 Funzione di densità di probabilità	3.9.1 Distribuzione di Poisson
3.4.2 Funzione di distribuzione cumulativa	3.9.2 Distribuzione esponenziale
3.4.3 Media e varianza	
3.5 PRINCIPALI DISTRIBUZIONI CONTINUE	3.10 APPROXIMAZIONE NORMALE DELLE DISTRIBUZIONI BINOMIALE E DI POISSON
3.5.1 Distribuzione normale	3.11 PIÙ VARIABILI ALEATORIE E INDIPENDENZA
3.5.2 Distribuzione lognormale	3.11.1 Distribuzioni congiunte
3.5.3 Distribuzione gamma	3.11.2 Indipendenza
3.5.4 Distribuzione di Weibull	3.12 FUNZIONI DI VARIABILI ALEATORIE
3.5.5 Distribuzione Beta	3.12.1 Combinazioni lineari di variabili aleatorie indipendenti
3.6 GRAFICI DEI QUANTILI	3.12.2 Combinazioni lineari di variabili aleatorie non indipendenti
3.6.1 Grafici dei quantili normali	3.12.3 Funzioni non lineari di variabili aleatorie indipendenti
3.6.2 Altri grafici dei quantili	3.13 CAMPIONI CASUALI, STATISTICHE E TEOREMA LIMITE CENTRALE
3.7 VARIABILI ALEATORIE DISCRETE	
3.7.1 Funzione di massa di probabilità	

---

## OBIETTIVI DI APPRENDIMENTO

Dopo uno studio attento di questo capitolo si sarà in grado di:

1. determinare le probabilità per variabili aleatorie discrete servendosi della funzione massa di probabilità, e le probabilità per variabili aleatorie continue servendosi della funzione densità di probabilità; usare in entrambi i casi la funzione di distribuzione cumulativa
  2. calcolare le medie e le varianze per variabili aleatorie discrete e continue
  3. comprendere le ipotesi sottostanti ciascuna distribuzione di probabilità presentata
  4. scegliere un'opportuna distribuzione di probabilità per calcolare le probabilità in specifiche applicazioni
  5. usare le tavole (o un software) per la funzione di distribuzione cumulativa di una distribuzione normale standard per eseguire un calcolo di probabilità
  6. approssimare le probabilità per le distribuzioni binomiale e di Poisson
  7. interpretare e calcolare la covarianza e la correlazione fra variabili aleatorie
  8. calcolare medie e varianze per combinazioni lineari di variabili aleatorie
  9. approssimare la media e la varianza per funzioni generiche di più variabili aleatorie
  10. comprendere il significato e l'applicazione del teorema limite centrale
-

Nei capitoli precedenti abbiamo usato sintesi numeriche e grafiche per riassumere e visualizzare i dati. Una sintesi è spesso necessaria per ricavare informazioni utili dai dati. Altrettanto importanti sono le conclusioni che è possibile trarre sul processo che tali dati ha generato; in altre parole, è importante poter ricavare conclusioni sulla performance a lungo termine di un processo basandosi su un campione relativamente ristretto di dati. Proprio perché viene utilizzato solo un campione, vi è incertezza sulle conclusioni che si traggono. Tuttavia, specificando un modello probabilistico per i dati, si può quantificare tale incertezza e selezionare o modificare le dimensioni del campione in modo da raggiungere un margine di incertezza accettabile. Lo scopo di questo capitolo è di descrivere i modelli probabilistici e presentare qualche importante esempio.

### 3.1 INTRODUZIONE

La misura della corrente in un sottile cavo in rame è un esempio di **esperimento**. Tale misura può produrre risultati leggermente differenti in repliche successive a causa di piccole variazioni nelle variabili non sotto controllo nell'esperimento: variazioni di temperatura, leggere variazioni di spessore e piccole impurità nella composizione chimica allorché si selezionano diverse porzioni del cavo, variazioni nel generatore di corrente e via dicendo. Conseguentemente, si può ritenere che questo esperimento (così come molti di quelli che conduciamo) abbia una componente **casuale**. In alcuni casi le variazioni casuali che si rilevano sono piccole rispetto agli scopi dell'esperimento, tanto da poter essere trascurate. Tuttavia la variabilità è quasi sempre presente e il suo valore può essere così grande da impedire o rendere difficile trarre le conclusioni necessarie sull'esperimento. In questi casi risultano di grande utilità i metodi presentati in questo libro per la modellizzazione e l'analisi dei risultati sperimentali.

Un esperimento che può dar luogo a risultati diversi anche se viene ripetuto ogni volta con le stesse modalità viene detto **esperimento casuale**. Possiamo selezionare una parte della produzione di un determinato giorno e misurare accuratamente una qualche dimensione caratteristica dei pezzi. Anche se si spera che le operazioni di produzione portino a parti identiche, in pratica vi sono spesso piccole variazioni nelle dimensioni effettivamente misurate, dovute a diversi fattori: vibrazioni, fluttuazioni della temperatura, presenza di differenti operatori, calibrazioni dei dispositivi di misura, usura degli strumenti di taglio ecc. Anche la procedura di misurazione può portare a variazioni nei risultati finali.

Per quanto accuratamente venga pianificato e condotto un esperimento, si riscontrano sempre variazioni. Il nostro obiettivo è di comprendere, quantificare e modellizzare i tipi di variabilità che si possono incontrare. Una volta incorporata la variabilità nei ragionamenti e nell'analisi, possiamo trarre dai risultati conclusioni ponderate che non sono invalidate dalla variabilità stessa.

I modelli e le analisi che includono la variabilità non sono diversi da quelli usati in altre aree dell'ingegneria e della scienza. La Figura 3.1 mostra la relazione tra il modello e il sistema fisico che esso rappresenta. Un modello matematico (o un'astrazione matematica) di un sistema fisico non deve necessariamente essere un'astrazione perfetta. Per esempio, le leggi di Newton non sono descrizioni perfette dell'universo fisico, tuttavia costituiscono utili modelli che possono venire studiati e analizzati per quantificare con un certo grado di approssimazione il comportamento di una vasta gamma di prodotti ingegneristici. Se si dispone di un'astrazione matematica confermata da misure condotte sul sistema reale, è possibile utilizzare tale modello per comprendere, descrivere e quantificare importanti aspetti del sistema fisico, fino a prevedere la sua risposta a determinati *input*.

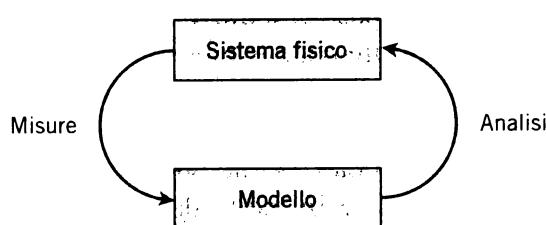


Figura 3.1 Interazione continua fra modello e sistema fisico.

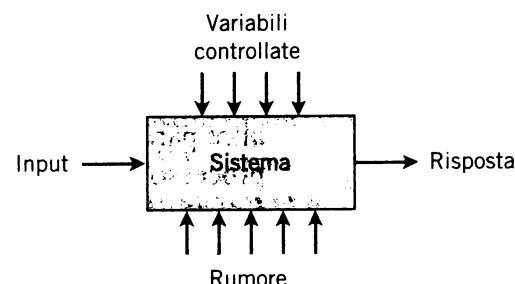


Figura 3.2 Il rumore influenza la trasformazione degli input in risposte del sistema.

Nel presente volume tratteremo modelli che permettono la presenza di variazioni nella risposta di un sistema, anche quando le variabili sotto controllo non sono fatte variare di proposito durante l'esperimento. La Figura 3.2 illustra graficamente un modello che incorpora variabili non controllabili (il *rumore*) che si combinano con quelle controllabili per produrre la risposta del sistema. A causa del rumore, identiche impostazioni delle variabili controllabili non portano al medesimo risultato ogni volta che viene effettuata una misura sul sistema.

Riferendoci all'esempio della misura di corrente condotta su un cavo in rame, il modello del nostro sistema potrebbe essere semplicemente la legge di Ohm

$$\text{corrente} = \text{tensione}/\text{resistenza}$$

Come detto in precedenza, ci si attendono variazioni nelle misure di corrente. La legge di Ohm può essere un'approssimazione accettabile, ma se le variazioni sono grandi rispetto all'uso che si deve fare del dispositivo elettrico in esame, può essere necessaria un'estensione del modello che tenga conto anche di tali variazioni (si veda al proposito la Figura 3.3).

È spesso difficile fare considerazioni sulla grandezza delle variazioni senza effettuare misure empiriche. Disponendo di un numero sufficiente di misure, però, si può approssimare tale grandezza e considerare l'effetto della variabilità sul comportamento di altri dispositivi, come gli amplificatori, presenti nel circuito. La nostra asserzione è dunque che il modello di Figura 3.2 è una descrizione maggiormente utile della misura di corrente.

Come ulteriore esempio, si consideri il progetto di un sistema di comunicazioni, quali una rete di computer o una rete telefonica. In questo sistema un importante fattore da considerare è la capacità di trasmissione disponibile per ciascun utente. Nel caso della comunicazione vocale è necessario disporre di un numero sufficiente di linee esterne per soddisfare le esigenze dell'azienda: supponendo che ciascuna linea possa trasportare una sola conversazione, quante linee si dovrebbero installare? Installandone troppo poche si corre il rischio di perdere o ritardare alcune comunicazioni, installandone un numero eccessivo salgono i costi. La progettazione e lo sviluppo di un prodotto sono fasi sempre più necessarie per raggiungere gli obiettivi *a un costo competitivo*.

Nel progetto del sistema di comunicazione vocale è necessario un modello per il numero e la durata delle chiamate. Anche sapere che, in media, si effettua una telefonata ogni 5 minuti e che ciascuna dura 5 minuti non è sufficiente: se le chiamate venissero fatte a intervalli precisi di 5 minuti e con durata precisa di 5 minuti, allora basterebbe una linea soltanto, ma la minima variazione della frequenza delle chiamate o della loro durata porterebbe al blocco di qualche telefonata da parte di altre (Figura 3.4). Pertanto, un sistema progettato senza tenere conto della variabilità sarebbe purtroppo inutile per un utilizzo pratico.

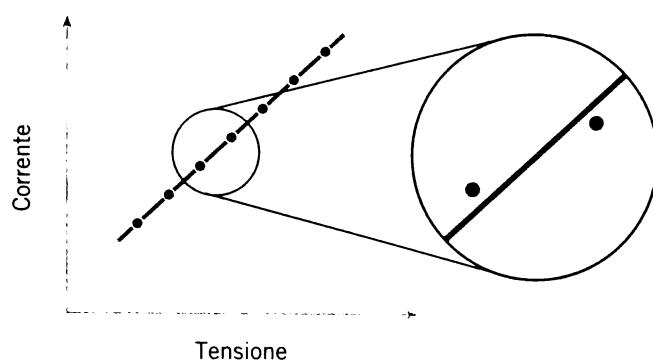


Figura 3.3 Un esame più dettagliato del sistema consente di rilevare deviazioni del modello.

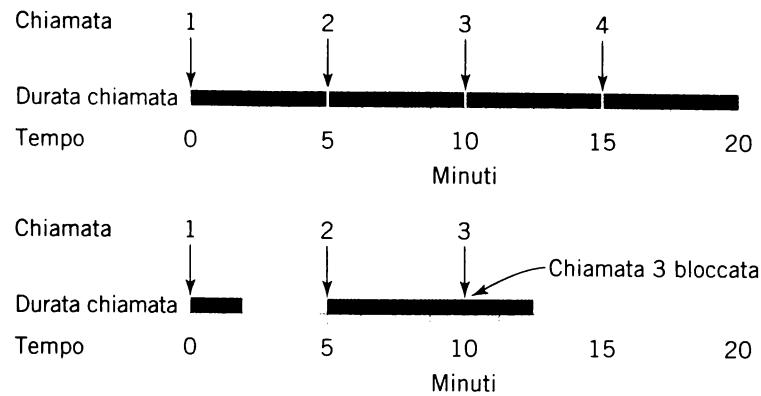


Figura 3.4 Una variazione causa rotture nel sistema.

## 3.2 VARIABILI ALEATORIE

In un esperimento, una misura viene di solito indicata mediante una variabile, per esempio  $X$ . In un esperimento casuale ci si riferisce a una variabile il cui valore misurato può cambiare da una replica dell'esperimento all'altra con l'espressione **variabile aleatoria**. Per esempio,  $X$  può indicare la misura di corrente nell'esperimento della misura di corrente nel cavo in rame.

Una variabile aleatoria non è concettualmente differente da ogni altra variabile di un esperimento; si usa semplicemente il termine "aleatoria" per indicare che diversi disturbi (il cosiddetto "rumore") possono modificarne il valore misurato.

### Variabile aleatoria

**Una variabile aleatoria** è una variabile numerica il cui valore misurato può cambiare da una replica dell'esperimento all'altra.

La notazione per le variabili aleatorie ne prevede l'indicazione con **lettere maiuscole**. Il valore della variabile aleatoria misurato nel corso dell'esperimento viene invece denotato con **lettere minuscole**, come in  $x = 70$  mA. Spesso si sintetizza un esperimento casuale tramite il valore misurato di una variabile aleatoria.

Questo modello può essere collegato ai dati come segue. I dati sono i valori misurati di una variabile aleatoria ottenuti da più repliche di un esperimento. Per esempio, la prima replica può dare come risultato una misura di intensità di corrente (in milliampercere)  $x_1 = 70.1$ , la successiva replica  $x_2 = 71.2$ , la terza  $x_3 = 71.1$  ecc. Questi dati possono venire quindi sintetizzati con i metodi descrittivi visti nel Capitolo 2.

Spesso la misura di interesse – la corrente in un cavo in rame, la lunghezza di un lavorato industriale ecc. – è espressa mediante un numero reale. Di conseguenza, è possibile ottenere una precisione arbitraria nella misura. Nella pratica, naturalmente, si può arrotondare la misura al decimo o al centesimo di una qualche unità di misura. La variabile che rappresenta questa misurazione viene detta **variabile aleatoria continua**.

In altri esperimenti, invece, si può registrare un conteggio, quale il numero di bit trasmessi e ricevuti errati su una rete di PC. In questi casi le misure sono in effetti dei numeri interi. Ancora: possiamo registrare che i bit trasmessi e ricevuti errati sono una frazione di quelli trasmessi (per esempio una proporzione di 0.0042 su 10 000 bit trasmessi). La misura

in questo caso è frazionaria, ma è ancora limitata a punti discreti sulla retta reale. Ogni volta che ci si trova in queste situazioni si dice che la variabile aleatoria è una **variabile aleatoria discreta**.

**Variabili aleatorie discrete e continue.**

Una variabile aleatoria **discreta** è una variabile aleatoria il cui range è costituito da un insieme finito o numerabile di numeri reali.

Una variabile aleatoria **continua** è una variabile aleatoria il cui range è costituito da un intervallo (finito o infinito) di numeri reali.

In alcuni casi la variabile aleatoria  $X$  è effettivamente discreta, ma i valori possibili sono così fitti che conviene analizzarli come se fossero continui. Per esempio, si supponga che le misure di corrente siano lette da uno strumento digitale che visualizza la misura arrotondata al centesimo di milliampere. Dato che le possibili misure sono limitate, la variabile aleatoria è discreta; tuttavia può risultare un'approssimazione più conveniente e più semplice assumere che le misure di corrente siano i valori di una variabile aleatoria continua.

**Esempi di variabili aleatorie.**

**Esempi di variabili aleatorie continue:**

intensità di corrente elettrica, lunghezza, pressione, temperatura, tempo, tensione elettrica, peso

**Esempi di variabili aleatorie discrete:**

numero di graffi su di una superficie, proporzione di parti difettose su 1000 controllate, numero di bit trasmessi e ricevuti errati

### 3.3 PROBABILITÀ

La **probabilità** è un concetto usato per quantificare la verosimiglianza, o possibilità, che una misura cada entro un determinato insieme di valori. Per denotare la misura si impiega una variabile aleatoria. “La possibilità che  $X$ , la lunghezza di un componente lavorato, sia compresa fra 10.8 e 11.2 mm è pari al 25%” è un’affermazione che quantifica una nostra sensazione sulle possibili lunghezze per i componenti. Le affermazioni probabilistiche descrivono insomma la possibilità del verificarsi di particolari valori. Tale possibilità è quantificata mediante l’assegnazione all’insieme dei valori di un numero compreso nell’intervallo  $[0, 1]$  (o di una percentuale variabile tra 0 e 100%). Numeri più alti indicano che un insieme di valori ha maggiori possibilità di verificarsi.

La probabilità di un risultato può venire interpretata come il **grado di fiducia** (soggettivo) nel fatto che tale risultato si presenti. In questo senso, è indubbio che differenti individui attribuiranno diverse probabilità allo stesso risultato.

Un’altra interpretazione della probabilità può basarsi su repliche successive dell’esperimento casuale. In questo senso, la probabilità di un risultato è vista come la frazione di volte che tale risultato si produrrà in esecuzioni ripetute dell’esperimento casuale. Per esempio, se assegniamo una probabilità 0.25 al risultato “lunghezza del componente compresa fra 10.8 e

11.2 mm”, possiamo interpretare questa attribuzione dicendo che se producessimo ripetutamente quel componente (ossia se replicassimo un numero infinito di volte l'esperimento casuale), il 25% dei pezzi prodotti avrebbe una lunghezza compresa in tale intervallo. Questo esempio fornisce una interpretazione della probabilità in termini di **frequenza relativa**. La frazione, o frequenza relativa, di repliche ripetute che cadono nell'intervallo sarà pari a 0.25. Si noti che questa interpretazione si basa su una proporzione sul lungo periodo, in un numero infinito di repliche; con un numero basso di repliche la frazione di lunghezze che cadono effettivamente nell'intervallo indicato potrebbe differire da 0.25.

Si osservi inoltre che se tutte le lunghezze dei componenti lavorati cade nell'intervallo indicato la frequenza relativa, e perciò la probabilità, associata a tale intervallo è 1; se al contrario nessuna delle lunghezze rientra nell'intervallo, la frequenza relativa (e la probabilità) dell'intervallo è 0. Poiché le probabilità sono ristrette all'intervallo  $[0, 1]$ , possono essere interpretate come frequenze relative. La probabilità di un risultato viene in genere espressa in termini di una variabile aleatoria. Per l'esempio svolto in questo paragrafo,  $X$  denota la lunghezza del componente e l'affermazione probabilistica può venire scritta in una delle seguenti forme

$$P(X \in [10.8, 11.2]) = 0.25 \quad \text{ovvero} \quad P(10.8 \leq X \leq 11.2) = 0.25$$

Entrambe le relazioni affermano che la probabilità che la variabile aleatoria  $X$  assuma un valore appartenente all'intervallo  $[10.8, 11.2]$  è 0.25.

Le probabilità associate a una variabile aleatoria vengono in genere determinate a partire da un modello che descrive l'esperimento casuale. Nei prossimi paragrafi prenderemo in esame diversi modelli, ma prima di tutto occorre enunciare alcune proprietà generali relative alla probabilità, che è possibile comprendere alla luce dell'interpretazione frequentista. Allo scopo faremo uso delle seguenti notazioni e definizioni. Dato un insieme  $E$ , il **complementare** di  $E$ , indicato con  $E'$ , è l'insieme degli elementi che non appartengono a  $E$ . L'insieme dei numeri reali è indicato con  $R$ . Gli insiemi  $E_1, E_2, \dots, E_k$  si dicono **incompatibili** o **mutuamente esclusivi** se l'intersezione di ogni coppia è vuota, ossia se ogni elemento contenuto in uno di questi insiemi appartiene a uno e uno solo degli insiemi  $E_1, E_2, \dots, E_k$ .

### Proprietà della probabilità

- |  |       |
|--|-------|
| 1. $P(X \in R) = 1$ , dove $R$ è l'insieme dei numeri reali.<br>2. $0 \leq P(X \in E) \leq 1$ per ogni insieme $E$ .<br>3. Se $E_1, E_2, \dots, E_k$ sono incompatibili, allora:<br>$P(X \in E_1 \cup E_2 \cup \dots \cup E_k) = P(X \in E_1) + \dots + P(X \in E_k).$ | (3.1) |
|--|-------|

La Proprietà 1 può essere usata per dimostrare che il valore massimo di una probabilità è 1. La Proprietà 2 implica che una probabilità non può essere negativa. La Proprietà 3 afferma che la frazione delle misure complessive che cade in  $E_1 \cup E_2 \cup \dots \cup E_k$  è la somma delle frazioni che cadono in  $E_1$  e in  $E_2, \dots$ , e in  $E_k$  ogni volta che questi insiemi sono incompatibili. Per esempio

$$P(X \leq 10) = P(X \leq 0) + P(0 < X \leq 5) + P(5 < X \leq 10)$$

La Proprietà 3, inoltre, può essere usata per porre in relazione la probabilità di un insieme  $E$  con quella del suo complementare  $E'$ . Dato che  $E$  ed  $E'$  sono evidentemente incompatibili, e che  $E \cup E' = R$ , si ha:  $1 = P(X \in R) = P(X \in E \cup E') = P(X \in E) + P(X \in E')$ . Di conseguenza

$$P(X \in E') = 1 - P(X \in E)$$

Per esempio:  $P(X \leq 2) = 1 - P(X > 2)$ . In generale, per ogni numero  $x$  fissato, si ha

$$P(X \leq x) = 1 - P(X > x)$$

Indichiamo con  $\emptyset$  l'insieme vuoto. Dato che il complementare di  $R$  è  $\emptyset$ , si ha  $P(X \in \emptyset) = 0$ .

Si supponga che per la variabile aleatoria  $X$  che rappresenta la durata (espressa in ore) di comuni tubi a fluorescenza siano valide le seguenti probabilità:  $P(X \leq 5000) = 0.1$ ,  $P(5000 < X \leq 6000) = 0.3$ ,  $P(X > 8000) = 0.4$ . Da tali probabilità si possono ricavare i seguenti risultati (può essere utile rappresentare in grafico i diversi insiemi).

La probabilità che la durata sia minore o uguale a 6000 ore è

$$P(X \leq 6000) = P(X \leq 5000) + P(5000 < X \leq 6000) = 0.1 + 0.3 = 0.4$$

per la Proprietà 3. La probabilità che la durata superi 6000 ore è

$$P(X > 6000) = 1 - P(X \leq 6000) = 1 - 0.4 = 0.6$$

La probabilità che la durata sia maggiore di 6000 e minore o uguale a 8000 ore viene determinata a partire dalla considerazione che la somma delle probabilità per questo intervallo e per gli altri tre intervalli deve essere uguale a 1. In altre parole, l'unione degli altri tre intervalli è l'insieme complementare dell'insieme  $\{x \mid 6000 < x \leq 8000\}$ . Pertanto

$$P(6000 < X \leq 8000) = 1 - (0.1 + 0.3 + 0.4) = 0.2$$

La probabilità che la durata delle lampade sia minore o uguale a 5500 ore non può venire calcolata esattamente. Tutto ciò che si può dire è che

$$P(X \leq 5500) \leq P(X \leq 6000) = 0.4 \quad \text{e} \quad 0.1 = P(X \leq 5000) \leq P(X \leq 5500)$$

Se tra i dati vi fosse anche l'indicazione che  $P(5000 < X \leq 6000) = 0.15$ , potremmo dire che

$$\begin{aligned} P(X \leq 5500) &= P(X \leq 5000) + P(5000 < X \leq 6000) - P(5500 < X \leq 6000) \\ &= 0.1 + 0.3 - 0.15 = 0.25 \end{aligned}$$

### Esiti ed eventi

Non sempre da un esperimento si ricava un valore misurato. A volte il risultato è soltanto classificato in una di possibili categorie. Per esempio: la misura di corrente può essere registrata solo come *bassa*, *media*, *alta*; un componente elettronico può essere classificato soltanto come privo di difetti o come difettoso; un bit trasmesso attraverso un canale di comu-

nicazione digitale viene ricevuto errato oppure no. Le categorie possibili sono di solito indicate come **esiti**, e un insieme di uno o più esiti si chiama **evento**. Ebbene, il concetto di probabilità si applica anche agli eventi, e l'interpretazione frequentista è ancora adeguata.

Se l'1% dei bit trasmessi attraverso un canale di comunicazione digitale è ricevuto errato, si dovrebbe assegnare il valore 0.01 alla probabilità di un errore. Se indichiamo con  $E$  l'evento “bit ricevuto errato”, scriveremo

$$P(E) = 0.01$$

Le probabilità assegnate agli eventi soddisfano proprietà analoghe a quelle dell'Equazione (3.1), perciò possono essere interpretate come frequenze relative. Se  $\Omega$  indica l'insieme di tutti i possibili esiti dell'esperimento, allora:

1.  $P(\Omega) = 1$ .
2.  $0 \leq P(E) \leq 1$  per ogni evento  $E$ .
3. Se  $E_1, E_2, \dots, E_k$  sono eventi incompatibili, allora

$$P(E_1 \cup E_2 \cup \dots \cup E_k) = P(E_1 \text{ o } E_2 \text{ o } \dots \text{ o } E_k) = P(E_1) + P(E_2) + \dots + P(E_k).$$

Gli eventi  $E_1, E_2, \dots, E_k$  sono **incompatibili** se l'intersezione di ogni coppia è vuota. Come esempio di eventi incompatibili, si supponga che le probabilità degli esiti *bassa*, *media* e *alta* siano rispettivamente 0.1, 0.7 e 0.2. La probabilità di un esito *media o alta* viene indicata con  $P(\text{media o alta})$ , e vale l'uguaglianza

$$P(\text{media o alta}) = P(\text{media}) + P(\text{alta}) = 0.7 + 0.2 = 0.9$$

### ESEMPIO 3.1 Arrivi al Pronto Soccorso

La seguente tabella riassume gli arrivi ai reparti di Pronto Soccorso di quattro ospedali dell'Arizona. Alcuni pazienti lasciano l'ospedale senza essere stati visitati da un medico, e in questo caso sono indicati in tabella come NV. Gli altri pazienti che si sono recati al Pronto Soccorso vengono visitati, e in questo caso possono essere o meno ricoverati in ospedale.

	Ospedale				
	1	2	3	4	Totali
Totale delle persone	5292	6991	5640	4329	22 252
NV	195	270	246	242	953
Ricoverate	1277	1558	666	984	4485
Non ricoverate	3820	5163	4728	3103	16 814

Indichiamo con  $A$  l'evento “l'arrivo è all'ospedale 1” e con  $B$  l'evento “l'arrivo si conclude con un NV”. Calcoliamo il numero di esiti in  $A \cap B$ ,  $A'$  e  $A \cup B$ .

L'evento  $A \cap B$  consiste nei 195 arrivi all'ospedale 1 che si concludono con una mancata visita (NV). L'evento  $A'$  consiste negli arrivi agli ospedali 2, 3 e 4, e contiene  $6991 + 5640 + 4329 = 16\,690$  arrivi. L'evento  $A \cup B$ , infine, è costituito dagli arrivi all'ospedale 1 o dagli NV o da entrambi, e contiene  $5292 + 270 + 246 + 242 = 6050$  arrivi. Si noti che quest'ultimo numero può essere ottenuto anche come somma degli elementi di  $A$  e degli elementi di  $B$ .

meno il numero degli elementi di  $A \cap B$  (che altrimenti verrebbero contati due volte), ossia  $5292 + 953 - 195 = 6050$ .

Supponiamo i 22 252 esiti in tabella siano equiprobabili. Allora, si può usare il numero di esiti in questi eventi per calcolare le probabilità:

$$P(A \cap B) = \frac{195}{22\,252} = 0.0088 \quad P(A') = \frac{16\,690}{22\,252} = 0.7500 \quad P(A \cup B) = \frac{6050}{22\,252} = 0.2719$$

L'interpretazione pratica è che gli ospedali tracciano gli arrivi che si concludono con una mancata visita per capire quali sono le risorse necessarie all'ospedale e migliorare così i servizi ai pazienti.

## 3.4 VARIABILI ALEATORIE CONTINUE

Come detto nel Paragrafo 3.2, una variabile aleatoria continua è caratterizzata dall'avere come range di valori un intervallo (finito o infinito). In questo paragrafo illustriamo alcune importanti proprietà per questo tipo di variabili aleatorie.

### 3.4.1 Funzione di densità di probabilità

La **distribuzione di probabilità**, o semplicemente **distribuzione**, di una variabile aleatoria  $X$  è una descrizione dell'insieme delle probabilità associate ai possibili valori di  $X$ . La distribuzione di probabilità di una variabile aleatoria può essere specificata in più di un modo.

Le densità sono comunemente usate in ingegneria per descrivere sistemi fisici. Per esempio, si consideri la densità di carico su una trave lunga e sottile (Figura 3.5). La densità (in g/cm) può essere descritta tramite una funzione dei punti  $x$  lungo la trave: zone con alti valori di carico corrispondono allora ad alti valori della funzione. Il carico totale fra i punti  $a$  e  $b$  viene determinato dall'integrale della funzione di densità fra  $a$  e  $b$ ; tale integrale è l'area sottesa dalla funzione di densità in questo intervallo e può essere interpretato come la somma di tutti i carichi sull'intervallo.

Analogamente, si può usare una **funzione di densità di probabilità**  $f(x)$  per descrivere la distribuzione di probabilità di una variabile aleatoria continua  $X$ . La probabilità che  $X$  cada fra  $a$  e  $b$  viene allora determinata come integrale di  $f(x)$  da  $a$  a  $b$  (Figura 3.6):

**Funzione  
di densità  
di probabilità**

La **funzione di densità di probabilità** (a volte abbreviata in pdf, *Probability Density Function*) di una variabile aleatoria continua viene usata per determinare le probabilità nel modo seguente

$$P(a < X < b) = \int_a^b f(x) dx \quad (3.2)$$

Le proprietà della funzione di densità di probabilità sono

- (1)  $f(x) \geq 0$
- (2)  $\int_{-\infty}^{\infty} f(x) dx = 1$

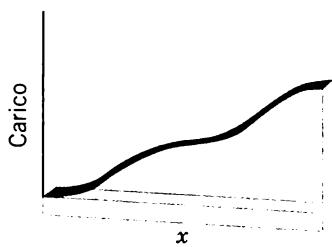


Figura 3.5 Funzione di densità di un carico su una trave lunga e sottile.

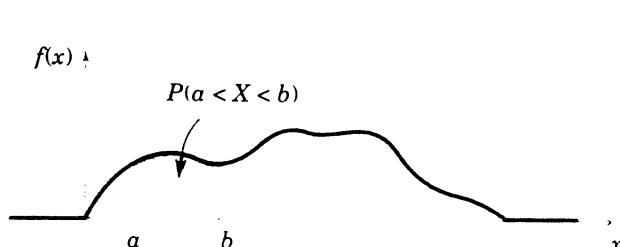


Figura 3.6 Determinazione della probabilità come area sottostante il grafico di  $f(x)$ .

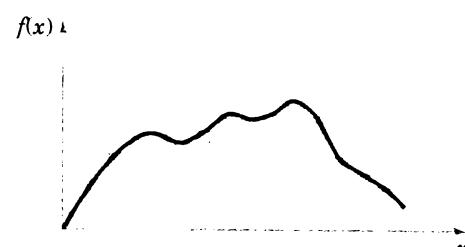


Figura 3.7 L'istogramma approssima una funzione di densità di probabilità. L'area di ciascuna barra è uguale alla frequenza relativa dell'intervallo, l'area sottostante  $f(x)$  su ogni intervallo è uguale alla probabilità dell'intervallo.

Un istogramma è un'approssimazione di una funzione di densità di probabilità (Figura 3.7). Per ogni intervallo dell'istogramma, l'area delle barre è uguale alla frequenza relativa delle misure che cadono in tale intervallo, e la frequenza relativa è una stima della probabilità che una misura cada nell'intervallo considerato. Analogamente, l'area sottostante il grafico di  $f(x)$  su un intervallo è uguale alla vera probabilità che una misura cada nell'intervallo considerato.

Una funzione di densità di probabilità fornisce una semplice descrizione delle probabilità associate a una variabile aleatoria. Siccome  $f(x)$  è positiva o nulla e  $\int_{-\infty}^{\infty} f(x) dx = 1$  si ha  $0 \leq P(a < X < b) \leq 1$ , perciò le probabilità sono limitate in modo appropriato. Una funzione di densità di probabilità vale zero per valori di  $x$  che non possono verificarsi, e si suppone nulla in ogni punto in cui non è specificamente definita.

Il punto importante è che  $f(x)$  viene usata per calcolare un'area che rappresenta la probabilità che  $X$  assuma un valore compreso nell'intervallo  $[a, b]$ . Per le misure di corrente del Paragrafo 3.1 la probabilità che  $X$  assuma un valore appartenente a  $[14 \text{ mA}, 15 \text{ mA}]$  è data dall'integrale della funzione di densità di probabilità  $f(x)$  su questo intervallo. La probabilità che  $X$  assuma un valore in  $[14.5 \text{ mA}, 14.6 \text{ mA}]$  è data dall'integrale della stessa funzione  $f(x)$  su questo intervallo più piccolo. Con una scelta opportuna della forma di  $f(x)$  è possibile rappresentare le probabilità associate a ogni variabile aleatoria  $X$ . La forma di  $f(x)$  permette di confrontare la probabilità che  $X$  assuma un valore in  $[14 \text{ mA}, 15 \text{ mA}]$  con la probabilità di ogni altro intervallo di ampiezza uguale o diversa.

Per la funzione di densità di carico su una trave lunga e sottile, poiché ogni punto ha ampiezza zero, l'integrale che determina il carico in ogni punto è nullo. Analogamente, per una variabile aleatoria continua  $X$  e per ogni valore  $x$  si ha

$$P(X = x) = 0$$

In base a questo risultato potrebbe sembrare che il nostro modello di variabile aleatoria continua sia inutile. Nella pratica, tuttavia, quando si osserva una particolare misura di corrente (per esempio 14.47 mA), si può interpretare questo risultato come il valore arrotondato di una misura di corrente che in effetti cade in un intervallo, per esempio l'intervallo  $14.465 \leq x \leq 14.475$ . Pertanto, la probabilità che come valore di  $X$  si osservi il valore arrotondato 14.47 viene interpretata come la probabilità che  $X$  assuma un valore nell'intervallo  $[14.465, 14.475]$ , che è diversa da zero. Analogamente, il nostro modello di variabile aleatoria continua implica quanto segue.

Se  $X$  è una variabile aleatoria continua, per ogni  $x_1$  e  $x_2$  si ha

$$P(x_1 \leq X \leq x_2) = P(x_1 < X \leq x_2) = P(x_1 \leq X < x_2) = P(x_1 < X < x_2)$$

Quando si deve determinare una probabilità per una variabile aleatoria può essere utile seguire questo schema:

1. Individuare la variabile aleatoria e la relativa distribuzione.
2. Scrivere l'espressione della probabilità di interesse in termini della variabile aleatoria individuata.
3. Calcolare tale probabilità usando la distribuzione individuata.

Questi tre passi vengono seguiti esplicitamente nelle risoluzioni di alcuni esempi di questo capitolo. In altri esempi ed esercizi è possibile utilizzare per proprio conto questa procedura.

### ESEMPIO 3.2 Corrente in un filo conduttore

**Definire la variabile aleatoria e la distribuzione.**

**Scrivere l'espressione della probabilità e calcolare la probabilità.**

Sia  $X$  la variabile aleatoria continua che rappresenta la corrente misurata (in milliampere) in un filo sottile di rame. Si supponga che l'intervallo in cui varia  $X$  sia  $[0, 20 \text{ mA}]$  e che la funzione densità di probabilità di  $X$  sia  $f(x) = 0.05$  per  $0 \leq x \leq 20$ . Qual è la probabilità che una misura di corrente sia minore di 10 mA?

La funzione di densità di probabilità è mostrata in Figura 3.8. Si assume  $f(x) = 0$  laddove non è specificamente definita. La probabilità richiesta è indicata dall'area ombreggiata in figura, ed è pari a

$$P(X < 10) = \int_0^{10} f(x) dx = 0.5$$

Come ulteriore esempio, si avrebbe

$$P(5 < X < 15) = \int_5^{15} f(x) dx = 0.5$$

### ESEMPIO 3.3 Difetto in un disco magnetico

Sia  $X$  la variabile aleatoria continua che rappresenta la distanza (in micron) dall'inizio di una traccia su un disco magnetico sino al primo difetto. I dati storici mostrano che la distribuzione di  $X$  può essere modellizzata da una funzione di densità di probabilità,  $f(x) = \frac{1}{2000} e^{-x/2000}$ , con  $x \geq 0$ . Per quale frazione di dischi la distanza sino al primo difetto è maggiore di 1000  $\mu\text{m}$ ?

In Figura 3.9 sono mostrate la funzione di densità di probabilità e la probabilità richiesta. Si ha

$$P(X > 1000) = \int_{1000}^{\infty} f(x) dx = \int_{1000}^{\infty} \frac{e^{-x/2000}}{2000} dx = -e^{-x/2000} \Big|_{1000}^{\infty} = e^{-1/2} = 0.607$$

Quale proporzione cade fra 1000 e 2000  $\mu\text{m}$ ? Si ha

$$P(1000 < X < 2000) = \int_{1000}^{2000} f(x) dx = -e^{-x/2000} \Big|_{1000}^{2000} = e^{1/2} - e^{-1} = 0.239$$

Poiché l'area totale sotto al grafico di  $f(x)$  è uguale a 1, possiamo anche calcolare  $P(X < 1000) = 1 - P(X > 1000) = 1 - 0.607 = 0.393$ .

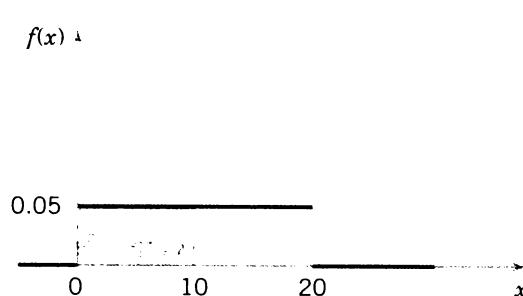


Figura 3.8 Funzione di densità di probabilità per l'Esempio 3.2.

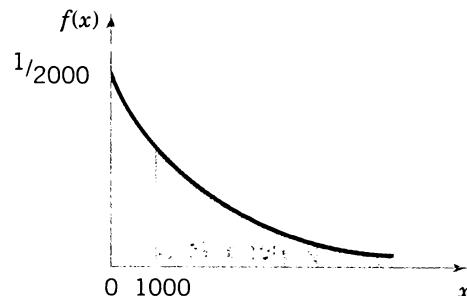


Figure 3.9 Funzione di densità di probabilità per l'Esempio 3.3.

### 3.4.2 Funzione di distribuzione cumulativa

Un altro modo di descrivere la distribuzione di probabilità di una variabile aleatoria consiste nel ricorrere a una funzione di un numero reale  $x$  la quale fornisca la probabilità che  $X$  sia minore o uguale a  $x$ .

**Funzione di distribuzione cumulativa di una variabile aleatoria continua**

La **funzione di distribuzione cumulativa** o **funzione di ripartizione** (a volte abbreviata in cdf, *Cumulative Distribution Function*) di una variabile aleatoria continua  $X$  con funzione di densità di probabilità  $f(x)$  è

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du$$

per  $-\infty < x < \infty$ .

Per una variabile aleatoria continua  $X$ , la definizione può essere anche  $F(x) = P(X < x)$ , perché si ha  $P(X = x) = 0$ .

La funzione di distribuzione cumulativa  $F(x)$  può essere posta in relazione alla funzione di densità di probabilità  $f(x)$  ed essere usata per ottenere le probabilità come segue

$$P(a < X < b) = \int_a^b f(x) dx = \int_{-\infty}^b f(x) dx - \int_{-\infty}^a f(x) dx = F(b) - F(a)$$

Il grafico di una funzione di distribuzione cumulativa ha specifiche proprietà. Dato che fornisce delle probabilità,  $F(x)$  è sempre maggiore o uguale a zero; inoltre, al crescere di  $x$  la funzione  $F(x)$  è non decrescente; per  $x$  che tende all'infinito  $F(x) = P(X \leq x)$  tende a 1.

Infine, si può ricavare la funzione di densità di probabilità dalla funzione di distribuzione cumulativa grazie al teorema fondamentale del calcolo integrale; infatti:

$$\frac{d}{dx} F(x) = \frac{d}{dx} \int_{-\infty}^x f(u) du = f(x)$$

#### ESEMPIO 3.4

Difetto in un  
disco magnetico -  
funzione di  
distribuzione

Si consideri la distanza sino al primo difetto introdotta nell'Esempio 3.3, con funzione di densità di probabilità

$$f(x) = \frac{1}{2000} e^{-x/2000}$$

per  $x \geq 0$ . La funzione di distribuzione cumulativa viene determinata come segue:

$$F(x) = \int_0^x \frac{1}{2000} e^{-u/2000} du = 1 - e^{-x/2000}$$

per  $x \geq 0$ . Si può verificare che  $\frac{d}{dx} F(x) = f(x)$ .

In Figura 3.10 è riportato il grafico di  $F(x)$ . Si può notare che  $F(x) = 0$  per  $x \leq 0$  e che  $F(x)$  cresce sino a 1 come detto.

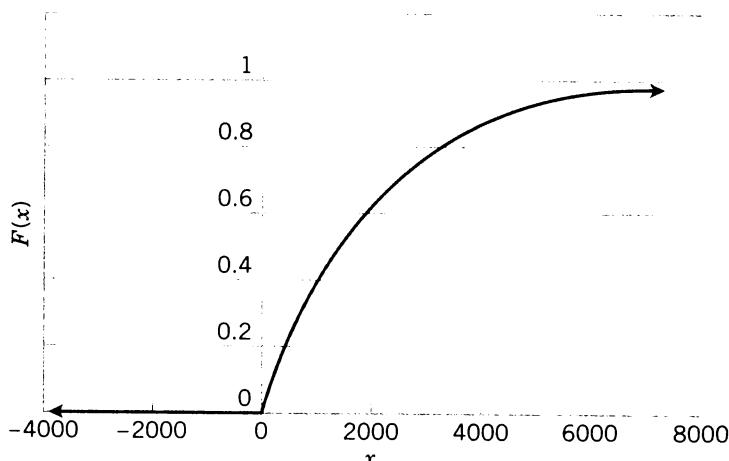


Figura 3.10 Funzione di distribuzione cumulativa per l'Esempio 3.4.

Le seguenti probabilità vanno confrontate con i risultati dell'Esempio 3.3. Determiniamo la probabilità che la distanza sino al primo difetto sia minore di 1000 µm.

La variabile aleatoria è la distanza sino al primo difetto superficiale con distribuzione data da  $F(x)$ . La probabilità richiesta è

Definire la variabile aleatoria e la distribuzione.

**Scrivere l'espressione della probabilità e calcolare la probabilità.**

$$P(X < 1000) = F(1000) = 1 - \exp\left(-\frac{1}{2}\right) = 0.393$$

Calcoliamo ora la probabilità che la suddetta distanza sia maggiore di 2000  $\mu\text{m}$

$$\begin{aligned} P(2000 < X) &= 1 - P(X \leq 2000) = 1 - F(2000) = 1 - [1 - \exp(-1)] \\ &= \exp(-1) = 0.368 \end{aligned}$$

Infine, calcoliamo la probabilità che la distanza sia compresa fra 1000 e 2000  $\mu\text{m}$

$$\begin{aligned} P(1000 < X < 2000) &= F(2000) - F(1000) = 1 - \exp(-1) - [1 - \exp(-0.5)] \\ &= \exp(-0.5) - \exp(-1) = 0.239 \end{aligned}$$

### 3.4.3 Media e varianza

Abbiamo visto che è utile sintetizzare un campione di dati tramite la media e la varianza; allo stesso modo, è possibile sintetizzare la distribuzione di probabilità di  $X$  tramite la sua media e la sua varianza. Per i dati campionari  $x_1, x_2, \dots, x_n$  la media campionaria può essere scritta come

$$\bar{x} = \frac{1}{n} x_1 + \frac{1}{n} x_2 + \dots + \frac{1}{n} x_n$$

In altri termini,  $\bar{x}$  usa pesi uguali, pari a  $1/n$ , come coefficienti moltiplicativi per ciascun valore misurato  $x_i$ . La media di una variabile aleatoria  $X$  usa il modello di probabilità per pesare i possibili valori di  $X$ . La **media o valore atteso** di  $X$ , indicata con  $\mu$  o con  $E(X)$ , è data da

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x) dx$$

L'integrale che compare in  $E(X)$  è analogo alla somma usata per calcolare  $\bar{x}$ .

Si ricordi che  $\bar{x}$  è il punto di equilibrio allorché in ciascun punto corrispondente a una misura sulla retta reale viene posto un uguale peso. Analogamente, se  $f(x)$  è la funzione di densità di un carico su una trave lunga e sottile,  $E(X)$  è il punto in cui la trave risulta in equilibrio. Di conseguenza,  $E(X)$  descrive il “centro” della distribuzione di  $X$  in maniera simile al baricentro di un carico.

Per i dati campionari  $x_1, x_2, \dots, x_n$  la varianza è una sintesi della dispersione dei dati, ed è pari a

$$s^2 = \frac{1}{n-1} (x_1 - \bar{x})^2 + \frac{1}{n-1} (x_2 - \bar{x})^2 + \dots + \frac{1}{n-1} (x_n - \bar{x})^2$$

In altri termini,  $s^2$  usa pesi uguali, pari a  $1/(n - 1)$ , come fattori moltiplicativi per ciascuno scarto quadratico  $(x_i - \bar{x})^2$ . Come detto in precedenza, gli scarti calcolati da  $\bar{x}$  tendono a essere più piccoli di quelli calcolati da  $\mu$ : per compensare, il peso viene cambiato da  $1/n$  a  $1/(n - 1)$ .

La varianza di una variabile aleatoria  $X$  è una misura della dispersione nei valori possibili di  $X$  stessa. La **varianza** di  $X$ , indicata con  $\sigma^2$  o con  $V(X)$  è data da

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

$V(X)$  usa il peso  $f(x)$  come fattore moltiplicativo per ciascun possibile scarto quadratico  $(x - \mu)^2$ . L'integrale nell'espressione di  $V(X)$  è analogo alla somma usata per calcolare  $s^2$ .

Si possono utilizzare le proprietà degli integrali e la definizione di  $\mu$  per dimostrare che

$$\begin{aligned} V(X) &= \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx - 2\mu \int_{-\infty}^{\infty} x f(x) dx + \int_{-\infty}^{\infty} \mu^2 f(x) dx \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx - 2\mu^2 + \mu^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2 \end{aligned}$$

Scrivendo l'ultimo integrale come  $E(x^2)$  otteniamo  $V(x) = E(x^2) - \mu^2$ .

### Media e varianza di una variabile aleatoria continua

Sia  $X$  una variabile aleatoria continua con funzione di densità di probabilità  $f(x)$ . La **media** o **valore atteso** di  $X$ , indicata con  $\mu$  o con  $E(X)$ , è data da

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx \quad (3.3)$$

La **varianza** di  $X$ , indicata con  $V(X)$  o con  $\sigma^2$ , è data da

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$$

La **deviazione standard** di  $X$  è  $\sigma = \sqrt{V(X)}$ .

**ESEMPIO 3.5**  
**Corrente in un  
filo conduttore -  
media**

Per la misurazione di corrente nel filo in rame dell’Esempio 3.2 la media di  $X$  è

$$E(X) = \int_{-\infty}^{\infty} xf(x) dx = \int_0^{20} x \left( \frac{1}{20} \right) dx = 0.05x^2/2 \Big|_0^{20} = 10$$

mentre la varianza di  $X$  è

$$V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_0^{20} (x - 10)^2 \left( \frac{1}{20} \right) dx = 0.05(x - 10)^3/3 \Big|_0^{20} = 33.33$$

**ESEMPIO 3.6**  
**Corrente in un  
filo conduttore -  
varianza**

Per la distanza sino al primo difetto dell’Esempio 3.3 la media di  $X$  è

$$E(X) = \int_{-\infty}^{\infty} xf(x) dx = \int_0^{\infty} x \frac{e^{-x/2000}}{2000} dx$$

Integrando per parti si ottiene:

$$E(X) = -xe^{-x/2000} \Big|_0^{\infty} + \int_0^{\infty} e^{-x/2000} = 0 - 2000 e^{-x/2000} \Big|_0^{\infty} = 2000$$

La varianza di  $X$  è

$$V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_0^{\infty} (x - 2000)^2 \frac{e^{-x/2000}}{2000} dx$$

Una doppia integrazione per parti fornisce

$$V(X) = 2000^2 = 4\,000\,000$$

## 3.5 PRINCIPALI DISTRIBUZIONI CONTINUE

### 3.5.1 Distribuzione normale

Il modello più diffuso per la distribuzione di una variabile aleatoria è senza dubbio la **distribuzione normale**. Nel Capitolo 2 sono stati mostrati diversi istogrammi con forme simili, simmetriche e a campana. Un risultato fondamentale, noto come **teorema limite centrale**, implica che gli istogrammi abbiano spesso questa forma caratteristica, almeno in via approssimata. Ogni volta che viene replicato un esperimento casuale, la variabile aleatoria che è uguale al risultato medio (o totale) sulle repliche tende ad assumere una distribuzione normale man mano che aumenta il numero di repliche. De Moivre presentò questo risultato nel 1733; sfortunatamente, il suo lavoro risultò disperso per un certo tempo, e Gauss sviluppò indipendentemente la distribuzione normale circa un secolo più tardi. Benché la scoperta sia

stata in seguito riconosciuta a De Moivre, la distribuzione normale è oggi nota con il nome alternativo di **distribuzione gaussiana**.

Quando facciamo la media (o la somma) dei risultati? Spessissimo. Nell'Esempio 2.1 abbiamo calcolato la media di otto resistenze alla trazione, e ottenuto 1055.0 psi. Se ipotizziamo che ogni misura risulti da una replica di un esperimento casuale, si può usare la distribuzione normale per trarre conclusioni approssimate sulla media. Tali conclusioni sono l'argomento primario dei successivi capitoli di questo libro.

A volte il teorema limite centrale è meno ovvio. Per esempio, si supponga che la deviazione (o l'errore) nella lunghezza di un manufatto sia la somma di un elevato numero di effetti infinitesimali, quali le alterazioni dovute alla temperatura e all'umidità, le vibrazioni, le variazioni dell'angolo di taglio, l'usura degli strumenti di taglio e dei supporti, le variazioni della velocità di rotazione, le variazioni di montaggio e fissaggio, le variazioni in diverse caratteristiche del materiale grezzo e quelle dei livelli dei contaminanti. Se gli errori componenti sono indipendenti ed è ugualmente verosimile che siano positivi o negativi, si può dimostrare che l'errore totale ha una distribuzione approssimativamente normale.

Inoltre, si incontra la distribuzione normale nello studio di numerosi fenomeni fisici fondamentali. Per esempio, il fisico scozzese Maxwell ottenne una distribuzione normale in base a semplici assunzioni sulle velocità delle molecole di un gas.

Abbiamo citato le basi teoriche della distribuzione normale per giustificare la forma alquanto complessa della funzione di densità di probabilità. Il nostro obiettivo è ora di calcolare le probabilità per una variabile aleatoria normale. Il teorema limite centrale verrà enunciato in maniera più rigorosa verso la fine del capitolo.

Le variabili aleatorie con diverse medie e varianze possono venire modellizzate da funzioni di densità di probabilità normali con opportune scelte del centro e della larghezza della curva. Il valore di  $E(X) = \mu$  determina il centro della funzione di densità di probabilità, quello di  $V(X) = \sigma^2$  ne determina la larghezza. La Figura 3.11 mostra diverse funzioni di densità di probabilità normali per determinati valori di  $\mu$  e  $\sigma^2$ : ognuna di esse ha la caratteristica forma simmetrica a campana, ma i centri e le dispersioni sono differenti da curva a curva. La seguente definizione fornisce la formula per le funzioni di densità di probabilità normali.

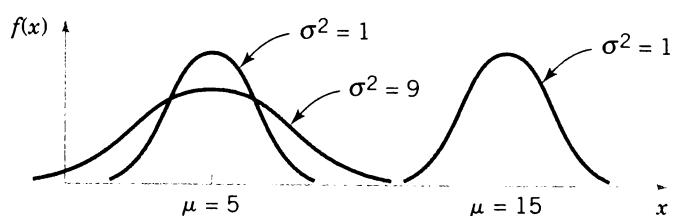


Figura 3.11 Funzioni di densità di probabilità normali per alcuni valori dei parametri  $\mu$  e  $\sigma^2$ .

### Distribuzione normale

Una variabile aleatoria  $X$  con funzione di densità di probabilità

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{per } -\infty < x < \infty \quad (3.4)$$

ha una **distribuzione normale** (e si dice perciò **variabile aleatoria normale**) con parametri  $\mu$  e  $\sigma$ , dove  $-\infty < \mu < \infty$  e  $\sigma > 0$ . Inoltre, si ha

$$E(X) = \mu \quad \text{e} \quad V(X) = \sigma^2$$

La media e la varianza della distribuzione normale verranno ricavate alla fine di questo paragrafo.

Spesso, per indicare una distribuzione normale con media  $\mu$  e varianza  $\sigma^2$  si usa la notazione  $N(\mu, \sigma^2)$ .

### ESEMPIO 3.7 Corrente in un filo conduttore: distribuzione normale

Si supponga che le misure di corrente in una striscia di materiale conduttore seguano una distribuzione normale con media 10 mA e varianza 4 mA<sup>2</sup>. Qual è la probabilità che una misura superi i 13 mA?

Indichiamo con  $X$  la corrente (misurata in milliampercere). La probabilità richiesta può essere rappresentata con  $P(X > 13)$  ed è mostrata in Figura 3.12 come area ombreggiata sottostante il grafico della funzione di densità di probabilità normale. Sfortunatamente non c'è un'espressione analitica per l'integrale di una funzione di densità di probabilità normale: le probabilità basate sulla distribuzione normale vengono ricavate, tipicamente, per via numerica o mediante opportune tabelle (che presenteremo in seguito).

In Figura 3.13 sono riassunti alcuni utili risultati relativi a una distribuzione normale. Per ogni variabile aleatoria normale si ha

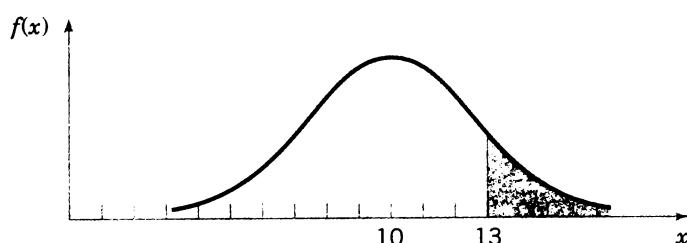


Figura 3.12 Probabilità che sia  $X > 13$  per una variabile aleatoria normale con  $\mu = 10$  e  $\sigma^2 = 4$  (Esempio 37).

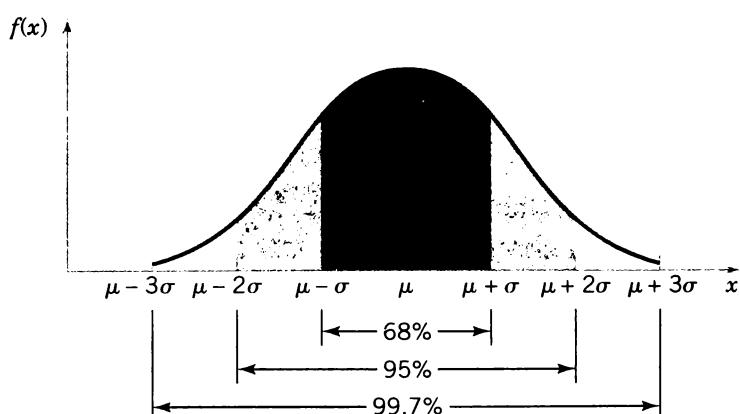


Figura 3.13 Probabilità associate a una distribuzione normale.

$$P(\mu - \sigma < X < \mu + \sigma) = 0.6827$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$$

Dalla simmetria di  $f(x)$  si ottiene  $P(X > \mu) = P(X < \mu) = 0.5$ . Poiché  $f(x)$  è positiva per tutte le  $x$ , questo modello assegna una probabilità non nulla a ogni intervallo della retta reale. Tuttavia la funzione di densità di probabilità decresce all'allontanarsi di  $x$  da  $\mu$ . Di conseguenza, la probabilità che una misura cada lontano da  $\mu$  è bassa, e da una certa distanza rispetto a  $\mu$  in poi può essere approssimata a zero. L'area sottesa dal grafico di una distribuzione normale oltre una distanza pari a  $3\sigma$  dalla media è decisamente piccola; si può sfruttare questo fatto per tracciare un andamento approssimato della funzione densità di probabilità normale, che consente di determinare le probabilità. Dato che più del 99.73% della probabilità di una distribuzione normale è all'interno dell'intervallo  $(\mu - 3\sigma, \mu + 3\sigma)$ , come **ampiezza della distribuzione gaussiana** si indica spesso il valore  $6\sigma$ . Si può dimostrare analiticamente che l'area sottesa dalla funzione di densità di probabilità normale da  $-\infty$  a  $\infty$  è uguale a 1.

### Variabile aleatoria normale standard

Una variabile aleatoria normale con media  $\mu = 0$  e varianza  $\sigma^2 = 1$  è detta **variabile aleatoria normale standard**; la si denota con il simbolo  $Z$ .

La Tavola I dell'Appendice A fornisce le probabilità cumulate per una variabile aleatoria normale standard; l'uso della tavola è illustrato dal seguente esempio.

### ESEMPIO 3.8 Distribuzione normale standard

Sia  $Z$  una variabile aleatoria normale standard. La Tavola I dell'Appendice A fornisce le probabilità nella forma  $P(Z \leq z)$ . La Figura 3.14 mostra l'uso della tavola per ricavare  $P(Z \leq 1.5)$ . Scendiamo nella colonna  $z$  sino alla riga con il valore 1.5, quindi leggiamo la probabilità nella colonna adiacente, etichettata con 0.00: la probabilità risulta 0.93319.

Le intestazioni di colonna si riferiscono ai centesimi nel valore di  $z$  in  $P(Z \leq z)$ . Per esempio,  $P(Z \leq 1.53)$  si trova all'incrocio fra la riga corrispondente a 1.5 nella colonna  $z$  e la colonna con l'intestazione 0.03: vale 0.93699.

### Distribuzione normale standard cumulata

La funzione

$$\Phi(z) = P(Z \leq z)$$

(tabulata nella Tavola I dell'Appendice A) indica la **funzione di distribuzione cumulata** di una variabile aleatoria normale standard. Per essa è necessario ricorrere ai valori tabulati o a un software statistico, poiché tale probabilità non può venire determinata con metodi elementari.

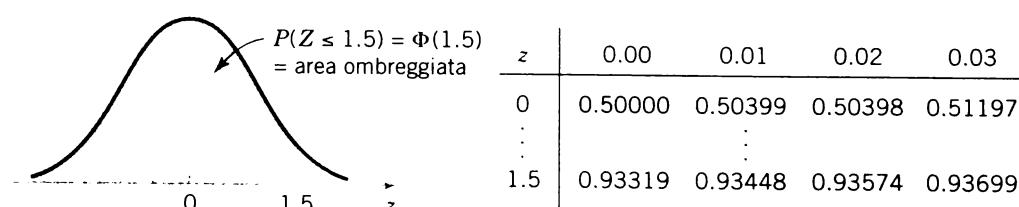


Figura 3.14 Funzione di densità della normale standard.

Le probabilità che non sono nella forma  $P(Z \leq z)$  si trovano usando le regole di base della probabilità e la simmetria della distribuzione normale, assieme alla Tavola I dell'Appendice A (o un software). I seguenti esempi illustrano il metodo da applicare.

### ESEMPIO 3.9

#### Probabilità di una distribuzione normale

I calcoli di questo esempio sono rappresentati graficamente in Figura 3.15. Nella pratica, una probabilità viene spesso arrotondata a una o due cifre significative.

- (1)  $P(Z > 1.26) = 1 - P(Z \leq 1.26) = 1 - 0.89616 = 0.10384$
- (2)  $P(Z < -0.86) = 0.19490$
- (3)  $P(Z > -1.37) = P(Z < 1.37) = 0.91465$
- (4) La probabilità  $P(-1.25 < Z < 0.37)$  si può trovare dalla differenza fra due aree:  $P(Z < 0.37) - P(Z < -1.25)$ .

Ora, si ha

$$P(Z < 0.37) = 0.64431 \quad \text{e} \quad P(Z < -1.25) = 0.10565$$

Pertanto

$$P(-1.25 < Z < 0.37) = 0.64431 - 0.10565 = 0.53866$$

- (5) La probabilità  $P(Z \leq -4.6)$  non si può trovare esattamente ricorrendo alla Tavola I, però si può utilizzare l'ultima voce della tavola per ricavare  $P(Z \leq -3.99) = 0.00003$ . Dal momento che  $P(Z \leq -4.6) < P(Z \leq -3.99)$ , si può dire che  $P(Z \leq -4.6)$  vale circa zero.
- (6) Per trovare il valore di  $z$  tale che  $P(Z > z) = 0.05$  riscriviamo l'equazione come  $P(Z \leq z) = 0.95$ , quindi usiamo la Tavola I all'inverso: scorriamo cioè le probabilità per trovare il valore corrispondente a 0.95; la soluzione è mostrata in Figura 3.15. Non riusciamo a trovare esattamente 0.95, ma solo il valore a esso più prossimo: 0.95053, che corrisponde a  $z = 1.65$ .
- (7) Troviamo infine il valore di  $z$  tale che  $P(-z < Z < z) = 0.99$ . Data la simmetria della distribuzione normale, se l'area della regione ombreggiata in Figura 3.15(7) deve essere uguale a 0.99, l'area in ciascuna coda della gaussiana deve essere uguale a 0.005. Pertanto, il valore di  $z$  cercato corrisponde a una probabilità pari a 0.995 in Tavola I. La probabilità che più vi si avvicina è 0.99506, corrispondente a  $z = 2.58$ .

Gli esempi precedenti mostrano come calcolare le probabilità per variabili aleatorie normali standard. Volendo impiegare lo stesso approccio per una variabile aleatoria normale arbitraria, sarebbe necessario disporre di tavole distinte per ogni possibile coppia di valori di  $\mu$  e  $\sigma$ . Per fortuna tutte le distribuzioni di probabilità normali sono legate algebricamente, e basta una semplice trasformazione per poter usare la Tavola I dell'Appendice A per trovare le probabilità associate a una qualsiasi variabile aleatoria normale.

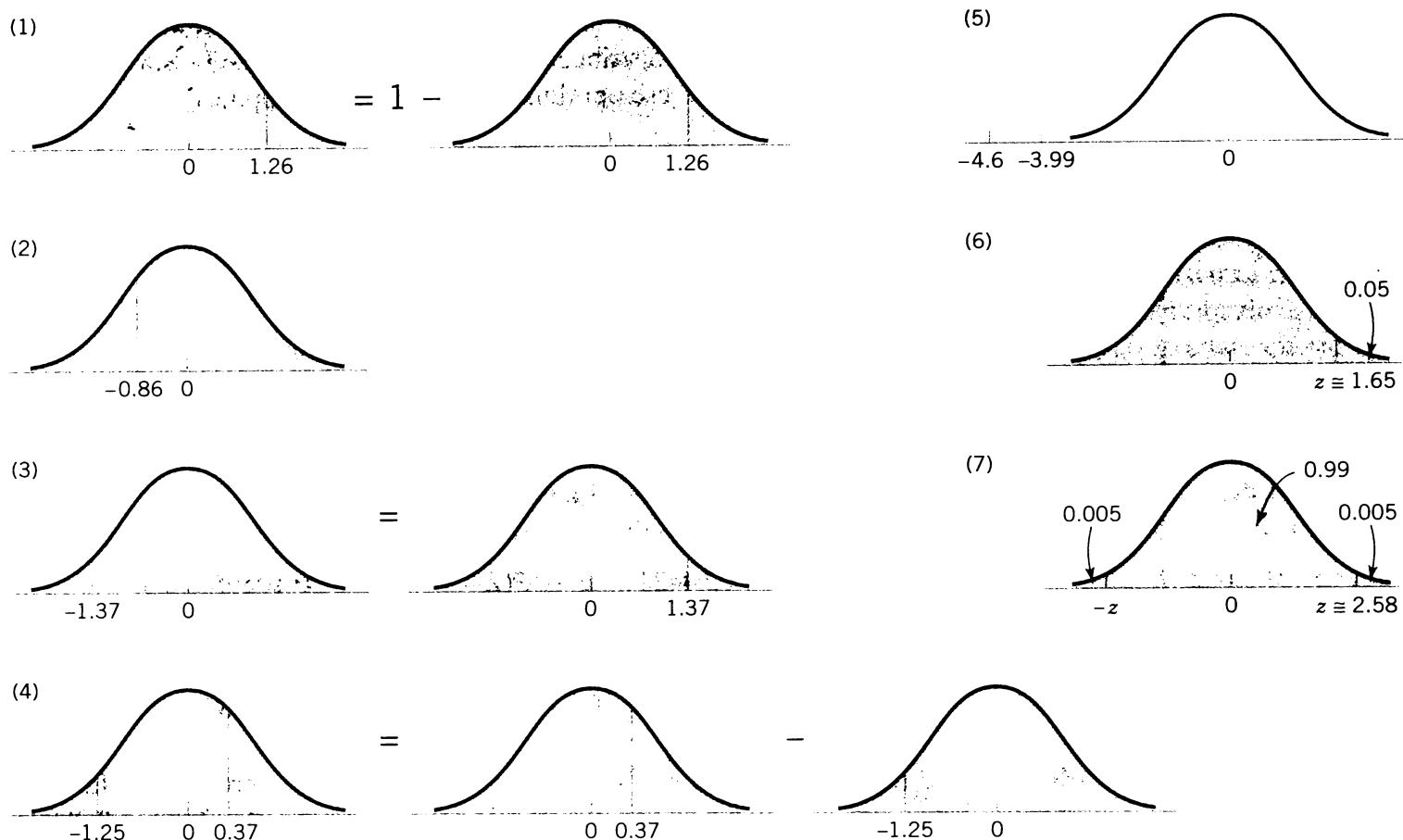


Figura 3.15 Visualizzazioni grafiche per l'Esempio 3.9.

**Variabile aleatoria normale standard**

Se  $X$  è una variabile aleatoria normale con media  $E(X) = \mu$  e varianza  $V(X) = \sigma^2$ , la variabile aleatoria

$$Z = \frac{X - \mu}{\sigma}$$

è una variabile aleatoria normale con  $E(Z) = 0$  e varianza  $V(Z) = 1$ , dunque è una variabile aleatoria **normale standard**.

La creazione di una nuova variabile aleatoria mediante questa trasformazione si chiama **standardizzazione**. La variabile aleatoria  $Z$  rappresenta la distanza di  $X$  dalla sua media in termini di deviazioni standard. Questo è il passo cruciale nella procedura di calcolo di una probabilità per una variabile aleatoria normale arbitraria.

**ESEMPIO 3.10**

Corrente in un filo conduttore: probabilità di una distribuzione normale

Si supponga che le misure di corrente in una striscia di materiale conduttore seguano una distribuzione normale con media  $10 \text{ mA}$  e varianza  $4 \text{ mA}^2$ . Quale è la probabilità che una misura superi i  $13 \text{ mA}$ ?

Indichiamo con  $X$  la corrente (misurata in millampere). La probabilità richiesta può essere rappresentata con  $P(X > 13)$ . Sia  $Z = (X - 10)/2$ . La relazione fra i valori di  $X$  e i val-

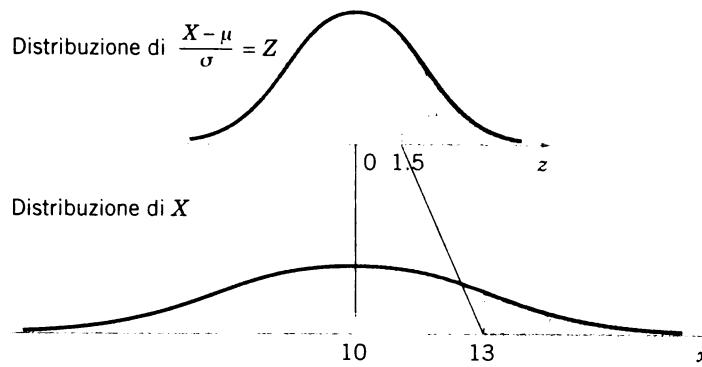


Figura 3.16 Standardizzazione di una variabile aleatoria normale.

ri trasformati di  $Z$  è illustrata in Figura 3.16. Possiamo osservare che  $X > 13$  corrisponde a  $Z > 1.5$ . Pertanto, dalla Tavola I si ricava

$$P(X > 13) = P(Z > 1.5) = 1 - P(Z \leq 1.5) = 1 - 0.93319 = 0.06681$$

Anziché usare la Figura 3.16 si può ricavare la probabilità dalla disequazione  $X > 13$

$$P(X > 13) = P\left(\frac{X - 10}{2} > \frac{13 - 10}{2}\right) = P(Z > 1.5) = 0.06681$$

Nel precedente esempio il valore 13 è stato trasformato in 1.5 mediante standardizzazione; 1.5 è allora quello che spesso viene detto il **valore  $z$**  associato alla probabilità.

Il seguente riquadro riassume i calcoli delle probabilità ricavati da variabili aleatorie normali.

### Standardizzazione

Sia  $X$  una variabile aleatoria normale con media  $\mu$  e varianza  $\sigma^2$ . Allora

$$P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P(Z \leq z) \quad (3.5)$$

dove  $Z$  è una variabile aleatoria **normale standard** e  $z = (x - \mu)/\sigma$  è il **valore  $z$**  ottenuto mediante **standardizzazione** di  $x$ .

La probabilità è ottenuta cercando  $z = (x - \mu)/\sigma$  nella **Tavola I dell'Appendice A**.

**ESEMPIO 3.11**  
Corrente in un  
filo conduttore:  
probabilità di  
una distribuzione  
normale

Definire la variabile  
aleatoria e la  
distribuzione.  
Scrivere l'espressione  
della probabilità.  
Calcolare la probabilità.

Proseguendo l'Esempio 3.10, qual è la probabilità che una misura di corrente sia compresa fra 9 e 11 mA?

Procedendo per via algebrica, si ha

$$\begin{aligned} P(9 < X < 11) &= P\left(\frac{9 - 10}{2} < \frac{X - 10}{2} < \frac{11 - 10}{2}\right) = P(-0.5 < Z < 0.5) \\ &= P(Z < 0.5) - P(Z < -0.5) = 0.69146 - 0.30854 = 0.38292 \end{aligned}$$

**Definire la variabile aleatoria e la distribuzione.**

**Scrivere l'espressione della probabilità**

**Calcolare la probabilità.**

Determiniamo il valore per cui vale 0.98 la probabilità che una misura di corrente sia al di sotto di esso. Il valore richiesto è mostrato graficamente in Figura 3.17. Ci serve il valore di  $x$  tale per cui  $P(X < x) = 0.98$ . Standardizzando, possiamo riscrivere questa espressione di probabilità come

$$P(X < x) = P\left(\frac{X - 10}{2} < \frac{x - 10}{2}\right) = P\left(Z < \frac{x - 10}{2}\right) = 0.98$$

Per trovare il valore di  $z$  per cui  $P(Z < z) = 0.98$  ricorriamo alla Tavola I; usando la probabilità più prossima abbiamo

$$P(Z < 2.05) = 0.97982$$

Pertanto,  $(x - 10)/2 = 2.05$ , e si usa all'inverso la trasformazione di standardizzazione per ricavare  $x$

$$x = 2(2.05) + 10 = 14.1 \text{ mA}$$

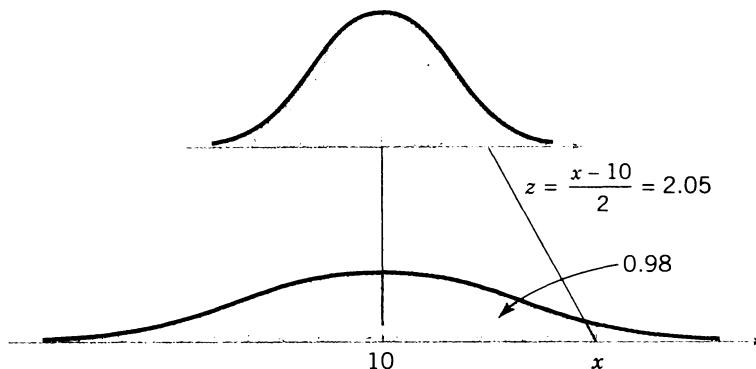


Figura 3.17 Determinazione del valore di  $x$  che soddisfa una specifica richiesta di probabilità (Esempio 3.11).

**ESEMPIO 3.12**  
**Tensione del rumore**

Si supponga che nella trasmissione di un segnale digitale il rumore di fondo seguia una distribuzione normale con media pari a 0 V e deviazione standard uguale a 0.45 V. Se il sistema assume che sia stata trasmessa la cifra 1 quando la tensione supera 0.9 V, qual è la probabilità di rilevare una cifra 1 quando non ne è stata inviata alcuna?

Sia  $N$  la variabile aleatoria che denota la tensione del rumore. La probabilità richiesta (che possiamo descrivere come la probabilità di un falso rilevamento) è allora

$$P(N > 0.9) = P\left(\frac{N}{0.45} > \frac{0.9}{0.45}\right) = P(Z > 2) = 1 - 0.97725 = 0.02275$$

Calcoliamo i limiti simmetrici intorno allo 0 che comprendono il 99% di tutte le letture di rumore. Dobbiamo trovare  $x$  per cui  $P(-x < N < x) = 0.99$ . In Figura 3.18 è mostrato un grafico esplicativo. Ora

$$P(-x < N < x) = P\left(-\frac{x}{0.45} < \frac{N}{0.45} < \frac{x}{0.45}\right) = P\left(-\frac{x}{0.45} < Z < \frac{x}{0.45}\right) = 0.99$$

Dalla Tavola I

$$P(-2.58 < Z < 2.58) = 0.99$$

Pertanto

$$\frac{x}{0.45} = 2.58$$

e

$$x = 2.58(0.45) = 1.16$$

Si supponga che una cifra 1 sia rappresentata come uno spostamento di 1.8 V della media della distribuzione del rumore. Qual è la probabilità che una cifra 1 non venga rilevata? Sia  $S$  la variabile aleatoria che rappresenta la tensione quando viene trasmessa una cifra 1. Allora

$$P(S < 0.9) = P\left(\frac{S - 1.8}{0.45} < \frac{0.9 - 1.8}{0.45}\right) = P(Z < -2) = 0.02275$$

Questa probabilità può essere interpretata come la probabilità di un segnale non rilevato.

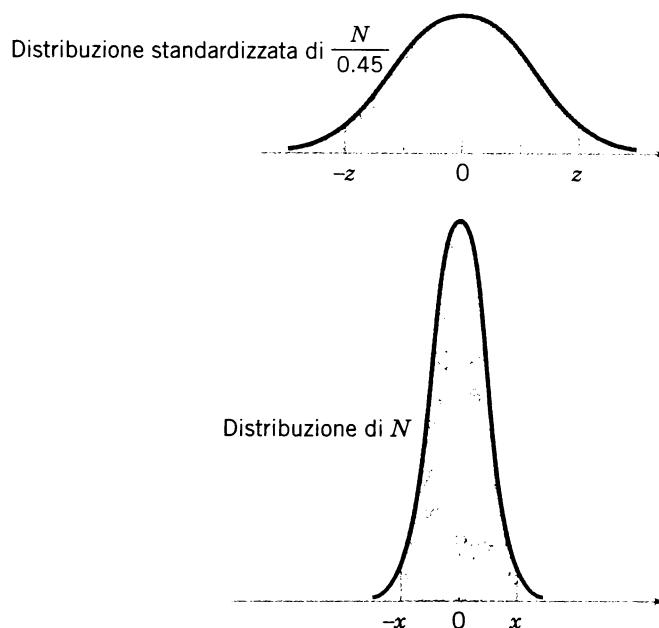


Figura 3.18 Determinazione del valore di  $x$  che soddisfa una specifica richiesta di probabilità (Esempio 3.12).

### ESEMPIO 3.13

Diametro  
di un alberino

Il diametro dell'alberino in un lettore di memorie ottiche è una variabile distribuita normalmente con media 0.2508 pollici e deviazione standard 0.0005 pollici. Le specifiche relative all'alberino sono  $0.2500 \pm 0.0015$  pollici. Quale frazione di alberini è conforme alle specifiche?

Sia  $X$  la variabile che rappresenta il diametro in pollici. La probabilità richiesta è mostrata in Figura 3.19; si ha

$$\begin{aligned} P(0.2485 < X < 0.2515) &= P\left(\frac{0.2485 - 0.2508}{0.0005} < Z < \frac{0.2515 - 0.2508}{0.0005}\right) \\ &= P(-4.6 < Z < 1.4) = P(Z < 1.4) - P(Z < -4.6) \\ &= 0.91924 - 0.00000 = 0.91924 \end{aligned}$$

La maggior parte degli alberini non conformi ha un diametro eccessivo, perché la media del processo si pone molto vicino al limite di specifica superiore. Se il processo viene centrato in modo che la media del processo sia uguale al valore desiderato, 0.2500, si ha

$$\begin{aligned} P(0.2485 < X < 0.2515) &= P\left(\frac{0.2485 - 0.2500}{0.0005} < Z < \frac{0.2515 - 0.2500}{0.0005}\right) \\ &= P(-3 < Z < 3) = P(Z < 3) - P(Z < -3) \\ &= 0.99865 - 0.00135 = 0.9973 \end{aligned}$$

Ricentrando il processo, dunque, la resa cresce sino a circa il 99.73%.

### Media e varianza della distribuzione normale

Dimostriamo che la media e la varianza di una variabile aleatoria normale sono rispettivamente  $\mu$  e  $\sigma^2$ . La media di  $x$  è

$$E(X) = \int_{-\infty}^{\infty} x \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} dx$$

Con la sostituzione di variabile  $y = (x - \mu)/\sigma$ , l'integrale diventa

$$E(X) = \mu \int_{-\infty}^{\infty} \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy + \sigma \int_{-\infty}^{\infty} y \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy$$

Il primo integrale che compare nella precedente espressione è pari a 1, poiché  $\frac{e^{-y^2/2}}{\sqrt{2\pi}}$  è la distribuzione di probabilità normale standard, mentre il secondo integrale è nullo, come si può vedere eseguendo formalmente la sostituzione di variabile  $u = -y^2/2$  oppure osservando la simmetria della funzione integranda intorno a  $y = 0$ . Pertanto,  $E(X) = \mu$ .

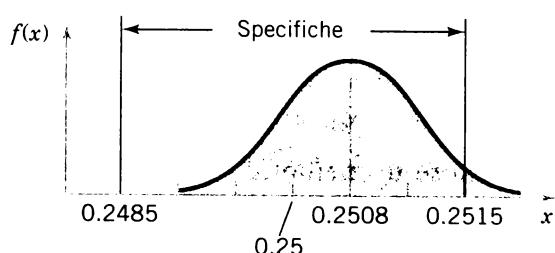


Figura 3.19 Distribuzione dell'Esempio 3.13.

La varianza di  $X$  è

$$V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma} dx$$

Con la sostituzione di variabile  $y = (x - \mu)/\sigma$  l'integrale diventa

$$V(X) = \sigma^2 \int_{-\infty}^{\infty} y^2 \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy$$

Integrando per parti con  $u = y$  e  $dv = y \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy$ , si trova che  $V(X)$  è uguale a  $\sigma^2$ .

### 3.5.2 Distribuzione lognormale

Le variabili di un sistema seguono a volte una relazione esponenziale del tipo  $x = \exp(w)$ . Se l'esponente è una variabile aleatoria, per esempio  $W$ , allora  $X = \exp(W)$  è anch'essa una variabile aleatoria ed è di interesse la distribuzione di  $X$ . Un importante caso speciale si presenta quando  $W$  ha distribuzione normale; in questo caso la distribuzione di  $X$  è detta **distribuzione lognormale**. Il nome di questa distribuzione deriva dalla trasformazione  $\ln(X) = W$ ; in altre parole, il logaritmo neperiano o naturale di  $X$  è normalmente distribuito.

Le probabilità per  $X$  si ottengono dalla trasformazione in  $W$ , ma è necessario rendersi conto che il range di  $X$  è  $(0, \infty)$ . Si supponga che  $W$  sia normalmente distribuita con media  $\theta$  e varianza  $\omega^2$ ; allora la funzione di distribuzione cumulativa per  $X$  è

$$\begin{aligned} F(x) &= P[X \leq x] = P[\exp(W) \leq x] = P[W \leq \ln(x)] \\ &= P\left[Z \leq \frac{\ln(x) - \theta}{\omega}\right] = \Phi\left[\frac{\ln(x) - \theta}{\omega}\right] \end{aligned}$$

per  $x > 0$ , dove  $Z$  è una variabile aleatoria normale standard. Pertanto, si può usare la Tavola I dell'Appendice per determinare la probabilità. Inoltre,  $F(x) = 0$  per  $x \leq 0$ .

Si può ottenere la funzione di densità di probabilità lognormale dalla derivata di  $F(x)$ , applicata all'ultimo termine nell'espressione per  $F(x)$ , l'integrale della funzione di densità normale standard. Una volta che la funzione di densità di probabilità è nota, si possono ricavare la media e la varianza di  $X$ . Omettiamo i dettagli del calcolo, riassumendo di seguito i soli risultati.

#### Distribuzione lognormale

Sia  $W$  una variabile aleatoria con distribuzione normale con media  $\theta$  e varianza  $\omega^2$ ; allora  $X = \exp(W)$  è una **variabile aleatoria lognormale** con funzione densità di probabilità

$$f(x) = \frac{1}{x\omega\sqrt{2\pi}} \exp\left[-\frac{(\ln(x) - \theta)^2}{2\omega^2}\right] \quad 0 < x < \infty \quad (3.6)$$

La media e la varianza di  $X$  sono

$$E(X) = e^{\theta + \omega^2/2} \quad \text{e} \quad V(X) = e^{2\theta + \omega^2}(e^{\omega^2} - 1) \quad (3.7)$$

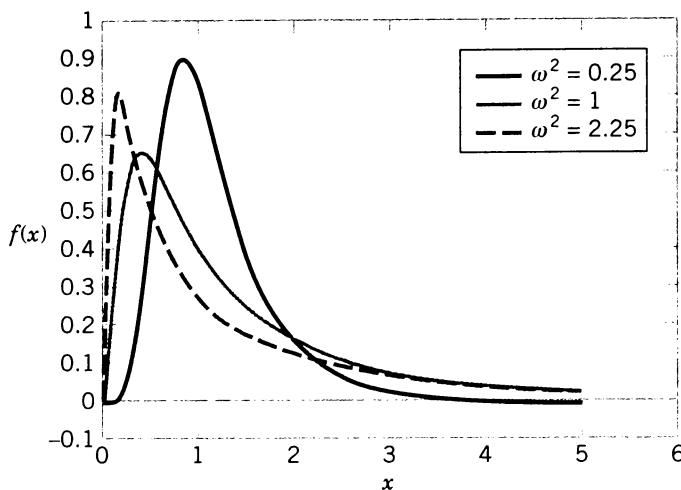


Figura 3.20 Funzioni di densità di probabilità lognormali con  $\theta = 0$  per alcuni valori di  $\omega^2$ .

I parametri di una distribuzione lognormale sono  $\theta$  e  $\omega^2$ , ma è necessario prestare attenzione al fatto che queste sono la media e la varianza della variabile aleatoria normale  $W$ , mentre la media e la varianza di  $X$  sono le funzioni di questi parametri che compaiono nella (3.7). La Figura 3.20 illustra le distribuzioni lognormali per alcuni valori dei parametri.

La durata di un prodotto che si degrada nel tempo viene spesso modellizzata da una variabile aleatoria lognormale. È il caso, per esempio, del tempo di vita di un laser a semiconduttore. In questo tipo di applicazioni si possono usare anche altre distribuzioni continue, ma poiché la distribuzione lognormale viene ricavata da una semplice funzione esponenziale di una variabile aleatoria normale è facile da visualizzare e da usare per il calcolo delle probabilità.

#### ESEMPIO 3.14

Tempo di vita  
di un laser

Il tempo di vita di un laser a semiconduttore ha una distribuzione lognormale con  $\theta = 10$  e  $\omega = 1.5$  ore. Qual è la probabilità che il tempo di vita superi le 10 000 ore?

Dalla funzione di distribuzione cumulativa per  $X$  si ha

$$\begin{aligned} P(X > 10\,000) &= 1 - P[\exp(W) \leq 10\,000] = 1 - P[W \leq \ln(10\,000)] \\ &= \Phi\left(\frac{\ln(10\,000) - 10}{1.5}\right) = 1 - \Phi(-0.52) = 1 - 0.30 = 0.70 \end{aligned}$$

Quale tempo di vita è superato dal 99% dei laser? Il problema ora è determinare la  $x$  per cui  $P(X > x) = 0.99$ ; pertanto

$$P(X > x) = P[\exp(W) > x] = P[W > \ln(x)] = 1 - \Phi\left(\frac{\ln(x) - 10}{1.5}\right) = 0.99$$

Con la Tabella I si trova che  $1 - \Phi(z) = 0.99$  per  $z = -2.33$ . Pertanto

$$\frac{\ln(x) - 10}{1.5} = -2.33 \quad \text{e} \quad x = \exp(6.505) = 668.48 \text{ ore}$$

Troviamo la media e la deviazione standard del tempo di vita

$$E(X) = e^{\theta + \omega^2/2} = \exp(10 + 1.125) = 67\,846.3$$

$$V(X) = e^{2\theta + \omega^2} (e^{\omega^2} - 1) = \exp(20 + 2.25)[\exp(2.25) - 1] = 39\,070\,059\,886.6$$

perciò la deviazione standard di  $X$  è 197 661.5 ore. Si noti che la deviazione standard della vita del laser è grande, rispetto alla media.

### 3.5.3 Distribuzione gamma

Per definire la distribuzione gamma abbiamo bisogno di una generalizzazione della funzione fattoriale:

#### Funzione gamma

**La funzione gamma** è definita da

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx, \quad \text{per } r > 0 \quad (3.8)$$

Si può dimostrare che l'integrale nella definizione di  $\Gamma(r)$  è finito. Inoltre, usando l'integrazione per parti si ricava

$$\Gamma(r) = (r - 1)\Gamma(r - 1)$$

Lasciamo questo risultato al lettore per esercizio. Pertanto, se  $r$  è un intero positivo,  $\Gamma(r) = (r - 1)!$ . Inoltre,  $\Gamma(1) = 0! = 1$  e (lo si può dimostrare)  $\Gamma(1/2) = \sqrt{\pi}$ .

Ora siamo in grado di definire la funzione di densità di probabilità gamma.

#### Distribuzione gamma

**La variabile aleatoria  $X$  con funzione di densità di probabilità**

$$f(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)}, \quad \text{per } x > 0 \quad (3.9)$$

**è una variabile aleatoria gamma** con parametri  $\lambda > 0$  e  $r > 0$ . La media e la varianza sono

$$\mu = E(X) = r/\lambda \quad \text{e} \quad \sigma^2 = V(X) = r/\lambda^2 \quad (3.10)$$

In Figura 3.21 sono mostrati alcuni grafici di distribuzione gamma per diversi valori di  $\lambda$  e di  $r$ .

La distribuzione gamma è molto utile nella modellizzazione di molteplici esperimenti casuali. Inoltre, la **distribuzione chi-quadro** è un caso speciale della distribuzione gamma, in cui  $\lambda$  vale  $1/2$  e  $r$  è uguale a uno dei valori  $1/2, 1, 3/2, 2, \dots$ . Questa distribuzione viene usata frequentemente nella stima per intervalli e nella verifica di ipotesi, argomenti trattati nei Capitoli 4 e 5. Quando il parametro  $r$  è un intero, la distribuzione gamma viene chiamata *distribuzione Erlang* (dal nome di A.K. Erlang, che per primo la impiegò nel campo delle telecomunicazioni).

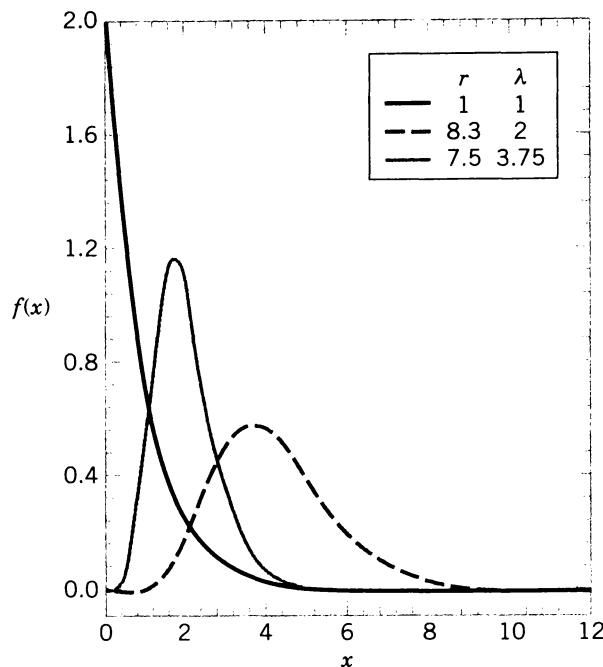


Figura 3.21 Funzioni di densità di probabilità gamma per alcuni valori di  $\lambda$  e di  $r$ .

### 3.5.4 Distribuzione di Weibull

La distribuzione di Weibull viene spesso utilizzata per modellizzare il tempo sino al guasto di vari sistemi fisici. I parametri della distribuzione offrono grande flessibilità nella costruzione di modelli in cui il numero di guasti cresce nel tempo (per esempio per l'usura dei cuscinetti), decresce nel tempo (come avviene in alcuni semiconduttori) o rimane costante (si pensi ai guasti causati da shock esterni al sistema).

#### Distribuzione di Weibull

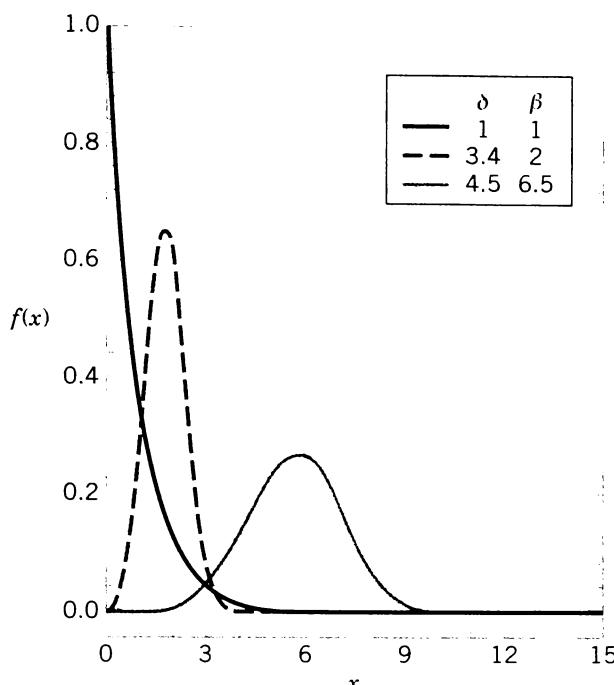
La variabile aleatoria  $X$  con funzione di densità di probabilità

$$f(x) = \frac{\beta}{\delta} \left( \frac{x}{\delta} \right)^{\beta-1} \exp \left[ -\left( \frac{x}{\delta} \right)^\beta \right], \text{ per } x > 0 \quad (3.11)$$

è una **variabile aleatoria di Weibull** con parametro di scala  $\delta > 0$  e parametro di forma  $\beta > 0$ .

La flessibilità della distribuzione di Weibull è illustrata tramite i grafici di Figura 3.22, relativi ad alcune funzioni di densità di probabilità.

Per calcolare le probabilità si usa spesso la funzione di distribuzione cumulativa, per la quale valgono i seguenti risultati.



**Figura 3.22** Funzioni di densità di probabilità di Weibull per alcuni valori di  $\delta$  e di  $\beta$ .

**Funzione di distribuzione di Weibull cumulativa**

Se  $X$  ha una distribuzione di Weibull con parametri  $\delta$  e  $\beta$ , la funzione di distribuzione cumulativa di  $X$  è

$$F(x) = 1 - \exp\left[-\left(\frac{x}{\delta}\right)^\beta\right]$$

La media e la varianza della distribuzione di Weibull sono le seguenti.

Se  $X$  ha una distribuzione di Weibull con parametri  $\delta$  e  $\beta$ , allora

$$\mu = E(X) = \delta\Gamma\left(1 + \frac{1}{\beta}\right) \quad \text{e} \quad \sigma^2 = V(X) = \delta^2\Gamma\left(1 + \frac{2}{\beta}\right) - \delta^2\left[\Gamma\left(1 + \frac{1}{\beta}\right)\right]^2 \quad (3.12)$$

**ESEMPIO 3.15**  
Tempo di vita di un cuscinetto

Il tempo sino al guasto (espresso in ore) di un cuscinetto in un albero meccanico è efficacemente modellizzato come una variabile aleatoria di Weibull con  $\beta = 1/2$  e  $\delta = 5000$  ore. Calcoliamo la durata media.

Dall'espressione per la media ricaviamo

$$E(X) = 5000\Gamma[1 + (1/0.5)] = 5000\Gamma[3] = 5000 \times 2! = 10\,000 \text{ ore}$$

Calcoliamo la probabilità che un cuscinetto duri almeno 6000 ore

$$P(X > 6000) = 1 - F(6000) = \exp\left[-\left(\frac{6000}{5000}\right)^{1/2}\right] = e^{-1.095} = 0.334$$

Di conseguenza, solo il 33.4% di tutti i cuscinetti durerà almeno 6000 ore.

### 3.5.5 Distribuzione Beta

Una distribuzione continua flessibile, ma limitata su un intervallo finito è utile per i modelli probabilistici. La quota di radiazioni solari assorbita da un materiale e la quota del tempo complessivo necessaria per portare a termine un'attività facente parte di un progetto sono due esempi di variabili aleatorie continue sull'intervallo  $[0, 1]$ .

La variabile aleatoria  $X$  con funzione di densità di probabilità

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \text{ per } x \in [0, 1]$$

è una **variabile aleatoria beta** con parametri  $\alpha > 0$  e  $\beta > 0$ .

I parametri di forma  $\alpha$  e  $\beta$  consentono alla funzione di densità di probabilità di assumere molte forme differenti. La Figura 3.23 ne illustra qualche esempio. Se  $\alpha = \beta$ , la distribuzione beta è simmetrica intorno al punto  $x = 0.5$ ; se  $\alpha = \beta = 1$  la distribuzione beta coincide con una funzione di distribuzione continua uniforme. La figura mostra come altre scelte dei parametri generino distribuzioni non simmetriche.

In generale, non esiste una espressione in forma chiusa per la funzione di distribuzione cumulativa; è necessario calcolare le probabilità per le variabili aleatorie beta numericamente. Negli esercizi è proposto qualche caso particolare in cui la funzione di densità di probabilità è più facilmente trattabile.

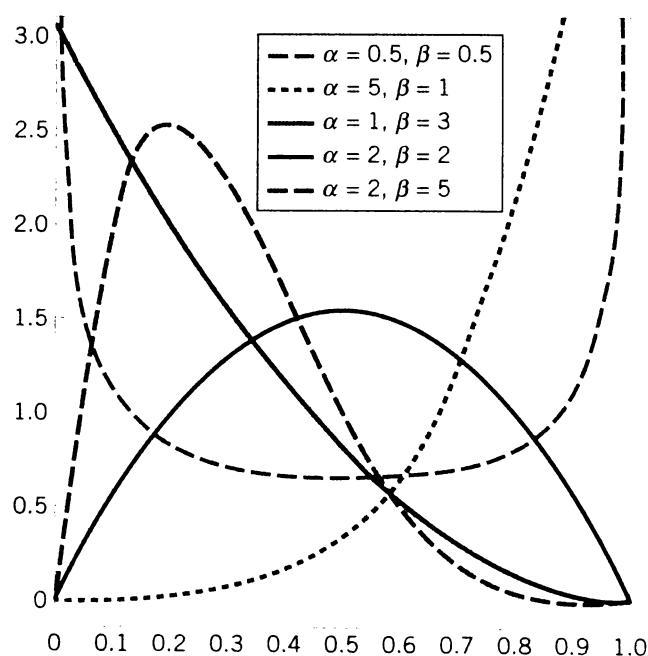


Figura 3.23 Funzioni di densità di probabilità per alcuni valori dei parametri  $\alpha$  e  $\beta$ .

**ESEMPIO 3.16**  
**Quota-tempo per un'attività**

Si consideri il tempo necessario a sviluppare un progetto commerciale di grandi dimensioni. La quota del tempo massimo consentito per portare a termine un'attività facente parte di questo progetto viene modellizzata da una variabile aleatoria beta con  $\alpha = 2.5$  e  $\beta = 1$ . Qual è la probabilità che la quota temporale sia maggiore di 0.7?

Indichiamo con  $X$  la quota del tempo massimo necessario per completare l'attività. La probabilità è

$$\begin{aligned} P(X > 0.7) &= \int_{0.7}^1 \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \\ &= \int_{0.7}^1 \frac{\Gamma(3.5)}{\Gamma(2.5) \Gamma(1)} x^{1.5} \\ &= \frac{2.5(1.5)(0.5)\sqrt{\pi}}{(1.5)(0.5)\sqrt{\pi}} \frac{1}{2.5} x^{2.5} \Big|_0^1 \\ &= 1 - 0.7^{2.5} = 0.59 \end{aligned}$$

Se  $X$  ha una distribuzione beta con parametri  $\alpha$  e  $\beta$ , allora

$$\mu = E(X) = \frac{\alpha}{\alpha + \beta} \quad \text{e} \quad \sigma^2 = V(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

**ESEMPIO 3.17**  
**Media e varianza della quota-tempo per un'attività**

Si consideri la quota del tempo necessario a portare a termine l'attività dell'esempio precedente. Calcoliamo la media e la varianza di tale variabile aleatoria beta.

Dall'espressione per la media e la varianza, sostituendo i dati, abbiamo

$$\mu = \frac{2.5}{2.5+1} = 0.71 \quad \sigma^2 = \frac{2.5}{3.5^2(4.5)} = 0.045$$

Se  $\alpha > 1$  e  $\beta > 1$ , la moda (il picco di densità) cade all'interno dell'intervallo  $[0, 1]$  ed è data da

$$\text{Moda} = \frac{\alpha - 1}{\alpha + \beta - 2}$$

Questa espressione è utile perché stabilisce una relazione fra il picco di densità e i parametri. Per la distribuzione dei due esempi precedenti si ha  $\alpha = 2.5$  e  $\beta = 1$ , quindi la moda di tale distribuzione è  $(2.5 - 1)/(3.5 - 2) = 1$ . Inoltre, benché una variabile aleatoria  $X$  sia definita sull'intervallo  $[0, 1]$ , è possibile costruire una variabile aleatoria  $W$  sull'intervallo finito  $[a, b]$  definita da  $W = a + (b - a)X$ .

## 3.6 GRAFICI DEI QUANTILI

### 3.6.1 Grafici dei quantili normali

Come si può sapere se una distribuzione normale costituisce un modello ragionevole per i dati? La costruzione di **grafici dei quantili** è un metodo grafico per stabilire se i dati campionari sono conformi a una distribuzione ipotizzata basandosi su un esame visivo soggettivo dei dati stessi. La procedura generale è molto semplice e rapida da eseguire. Per i grafici dei quantili si usa tipicamente una carta speciale per grafici, chiamata “carta di probabilità”, pensata specificamente per la distribuzione ipotizzata. Sono disponibili carte di probabilità per le distribuzioni normali, lognormali, di Weibull e per varie distribuzioni chi-quadro e gamma. In questo paragrafo illustreremo il **grafico dei quantili normali**, mentre prenderemo in considerazione i grafici dei quantili per altre distribuzioni continue nel Paragrafo 3.6.2.

Per costruire un grafico dei quantili, innanzitutto si ordinano le osservazioni del campione in senso crescente: il campione  $x_1, x_2, \dots, x_n$  viene dunque disposto come  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ , dove  $x_{(1)}$  è l’osservazione più piccola,  $x_{(2)}$  è l’osservazione immediatamente più grande e via dicendo, sino all’osservazione  $x_{(n)}$ , che è la più grande. Le osservazioni ordinate  $x_{(j)}$  vengono quindi riportate in grafico in funzione della loro frequenza cumulata osservata  $(j - 0.5)/n$  su un’opportuna carta di probabilità. Se la distribuzione ipotizzata descrive adeguatamente i dati, i punti tracciati in grafico cadranno approssimativamente lungo una linea retta; se si scostano significativamente e sistematicamente da una retta il modello ipotizzato non è appropriato. In genere, è soggettivo stabilire se i dati sono rappresentabili o meno come punti di una retta. La procedura è illustrata nel seguente esempio.

**ESEMPIO 3.18**  
Durata di servizio  
di una batteria

Quelle che seguono sono dieci osservazioni sull’effettiva durata di servizio delle batterie che alimentano un computer portatile, espresse in minuti: 176, 191, 214, 220, 205, 192, 201, 190, 183, 185. Si ipotizza che un modello adeguato per la durata delle batterie sia una distribuzione normale. Per impiegare i grafici dei quantili al fine di studiare questa ipotesi si dispongono innanzitutto le osservazioni in ordine crescente, e si calcolano le loro frequenze cumulate  $(j - 0.5)/n$ :

$j$	$x_{(j)}$	$(j - 0.5)/10$
1	176	0.05
2	183	0.15
3	185	0.25
4	190	0.35
5	191	0.45
6	192	0.55
7	201	0.65
8	205	0.75
9	214	0.85
10	220	0.95

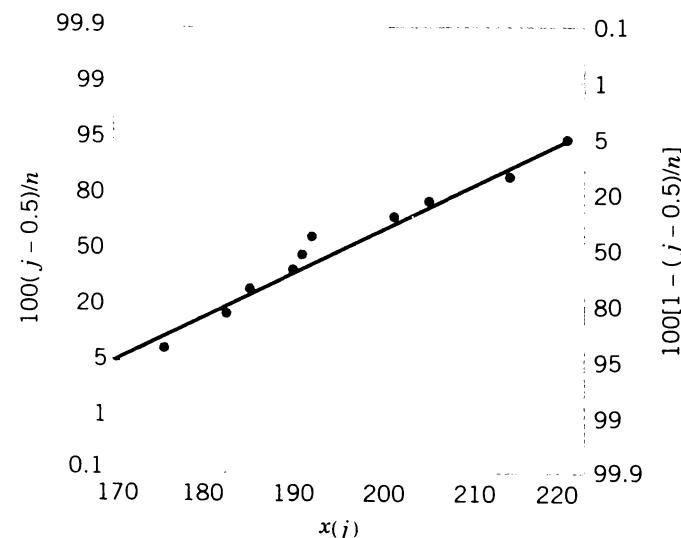


Figura 3.24 Grafico dei quantili normali per la durata delle batterie.

Fatto ciò, si riportano in grafico, su carta di probabilità normale, le coppie di valori  $x_{(j)}$  e  $(j - 0.5)/n$ . Il grafico è quello mostrato in Figura 3.24. La maggior parte delle carte di probabilità riportano sulla scala verticale di sinistra i valori  $100(j - 0.5)/n$  e su quella di destra i valori  $100[1 - (j - 0.5)/n]$ , mentre i valori della variabile sono rappresentati sull'asse orizzontale. In figura, attraverso l'insieme dei punti è stata tracciata una retta, scelta in maniera soggettiva come curva di “best fit”. Nel disegnare questa retta si dovrebbe essere più attenti ai punti vicino al centro del grafico che a quelli presso le estremità dell'insieme di osservazioni. Una buona regola pratica consiste nel tracciare la retta approssimativamente fra i punti relativi al 25-esimo e 75-esimo percentile, come è stato fatto in Figura 3.24 per valutare la deviazione sistematica dei punti dalla retta si immagini di disporre una matita lungo la retta: se tutti i punti risultano coperti da tale matita immaginaria, significa che una distribuzione normale descrive adeguatamente i dati. In Figura 3.24 i punti rappresentati superano questo criterio empirico, perciò concludiamo che il modello di distribuzione normale è appropriato.

Si può costruire un grafico dei quantili normali anche su carta comune, riportando in funzione di  $x_{(j)}$  i quantili della normale standard  $z_j$ , dove gli  $z_j$  sono tali che

$$\frac{j - 0.5}{n} = P(Z \leq z_j) = \Phi(z_j)$$

Per esempio, se  $(j - 0.5)/n = 0.05$ ,  $\Phi(z_j) = 0.05$  implica  $z_j = -1.64$ . Per comprendere quanto asserito, si considerino i dati dell'esempio precedente. Nella tabella sottostante i quantili della normale standard sono elencati nell'ultima colonna:

$j$	$x_{(j)}$	$(j - 0.5)/10$	$z_j$
1	176	0.05	-1.64
2	183	0.15	-1.04
3	185	0.25	-0.67
4	190	0.35	-0.39
5	191	0.45	-0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205	0.75	0.67
9	214	0.85	1.04
10	220	0.95	1.64

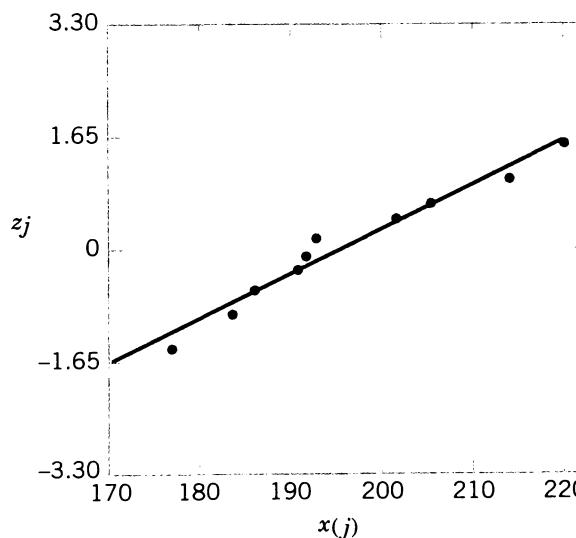


Figura 3.25 Grafico dei quantili ottenuto a partire dai quantili della normale standard.

La Figura 3.25 presenta il grafico di  $z_j$  in funzione di  $x_{(j)}$ ; tale grafico è equivalente a quello di Figura 3.25.

Un'applicazione molto importante dei grafici dei quantili normali si ha nella *verifica delle assunzioni* allorché si usano procedure di inferenza statistica che richiedono l'ipotesi di normalità.

### 3.6.2 Altri grafici dei quantili

I grafici dei quantili sono utilissimi, e rappresentano spesso la prima tecnica che si usa quando si tratta di stabilire quale distribuzione di probabilità fornirà verosimilmente un modello ragionevole per i dati a disposizione. Nell'uso di un grafico dei quantili la distribuzione viene in genere scelta in base a valutazioni soggettive sul grafico stesso. In associazione a questi grafici, tuttavia, si possono impiegare più formali tecniche di **bontà dell'adattamento**. Nel Paragrafo 4.10 descriveremo un **test di adattamento** molto semplice.

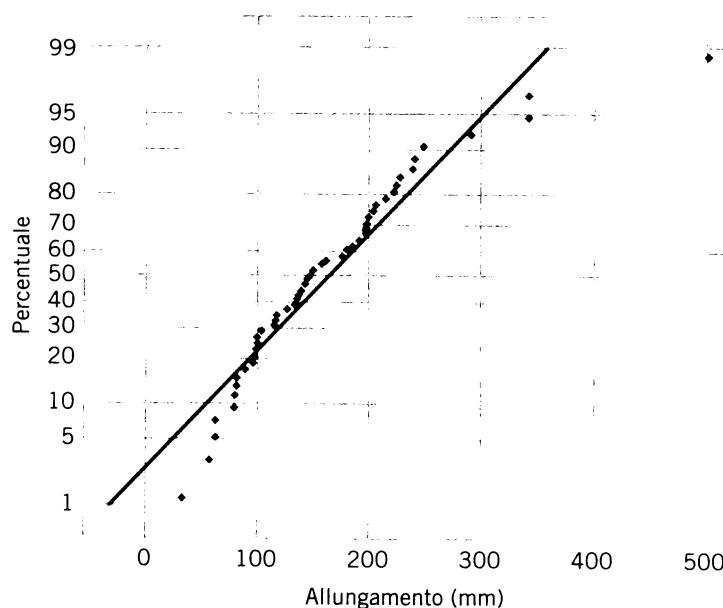
**Interpretazione di un grafico dei quantili.**

Per mostrare l'utilità dei grafici dei quantili nella determinazione di un'appropriata distribuzione per i dati, si considerino le osservazioni sull'allungamento massimo per una lega di alluminio (Tabella 3.1). Quello in Figura 3.26 è un grafico dei quantili normali relativo a tali dati; si noti come le code dei dati si scostino dalla retta: è un indizio del fatto che la distribuzione normale non è un buon modello per questi dati. La Figura 3.27 mostra un grafico dei quantili lognormali sui dati di allungamento, ottenuto con Minitab. I dati cadono ora molto più vicino alla retta, in particolare le osservazioni nelle code: ciò suggerisce che la distribuzione lognormale fornirà verosimilmente, per i dati a disposizione, un modello più ragionevole della distribuzione normale.

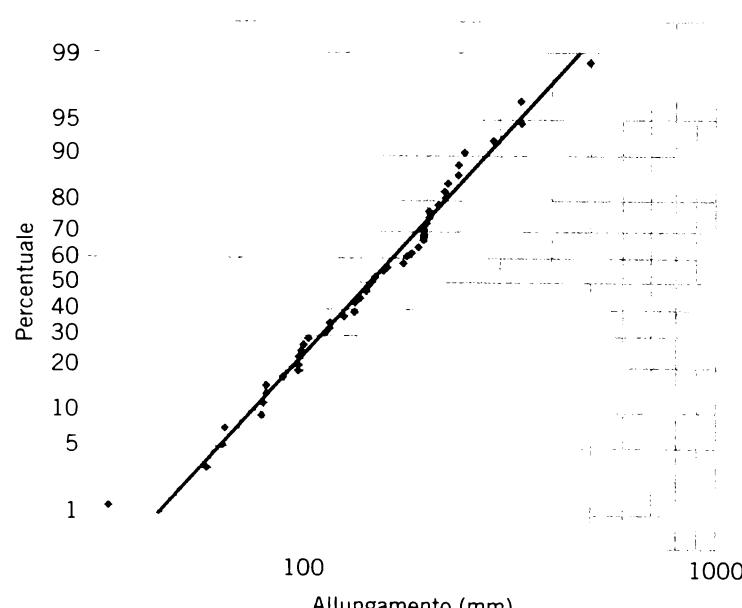
**Tabella 3.1** Allungamenti massimi (in mm) per una lega di alluminio.

81	98	291	101	98	118	158	197	139	249
249	135	223	205	80	177	82	64	137	149
117	149	127	115	198	342	83	34	342	185
227	225	185	240	161	197	98	65	144	151
134	59	181	151	240	146	104	100	215	200

Infine, la Figura 3.28 mostra un grafico dei quantili di Weibull per gli stessi dati di Tabella 3.1, anch'esso generato da Minitab. Le osservazioni nella coda inferiore del grafico non



**Figura 3.26** Grafico dei quantili normali per i dati di allungamento massimo di Tabella 3.1.



**Figura 3.27** Grafico dei quantili lognormali per i dati di allungamento massimo di Tabella 3.1.

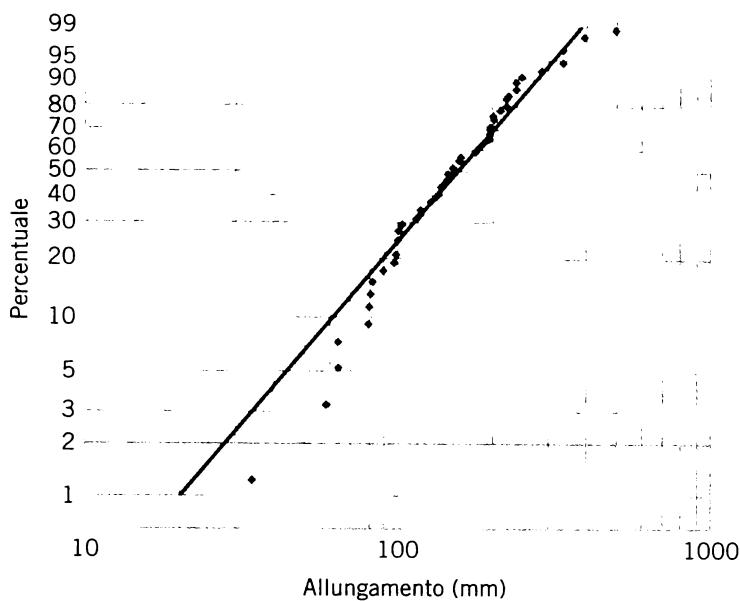


Figura 3.28 Grafico di probabilità di Weibull per i dati di allungamento massimo di Tabella 3.1.

sono molto vicine alla retta, a indicare che quello di Weibull non è un modello molto valido per i dati. Pertanto, in base all’osservazione dei tre grafici che abbiamo costruito, la scelta più appropriata per i dati di allungamento massimo sembra essere quella della distribuzione lognormale.

### 3.7 VARIABILI ALEATORIE DISCRETE

Come si è detto nel Paragrafo 3.2, per una variabile aleatoria discreta sono possibili solo misure con valori discreti. In questo capitolo abbiamo già incontrato esempi di variabili aleatorie discrete, e altri ne incontreremo. In questo paragrafo vengono presentate le proprietà di cui gode una variabile aleatoria discreta; sono proprietà analoghe a quelle viste per le variabili aleatorie continue.

#### ESEMPIO 3.19

Centralino telefonico

Il centralino di un’azienda è dotato di 48 linee esterne. In un certo istante, il sistema viene osservato, e vengono rilevate le linee in uso. Se  $X$  è la variabile aleatoria che indica il numero di linee in uso, allora  $X$  può assumere uno qualunque dei valori interi compresi fra 0 e 48.

#### ESEMPIO 3.20

Contaminazione di un wafer semiconduttore

L’analisi della superficie di un wafer semiconduttore porta a registrare il numero di particelle contaminanti che superano una determinata dimensione, rappresentato con la variabile aleatoria  $X$ . I possibili valori di  $X$  sono allora gli interi compresi fra 0 e un valore massimo, grande, che rappresenta il numero massimo di particelle contaminanti che possono essere presenti su un wafer. Se questo numero è molto alto, può essere conveniente assumere che i valori possibili di  $X$  siano tutti quelli compresi fra zero e infinito.

### 3.7.1 Funzione di massa di probabilità

Come detto in precedenza, la distribuzione di probabilità di una variabile aleatoria  $X$  è una descrizione delle probabilità associate ai possibili valori di  $X$ . Per una variabile aleatoria discreta la distribuzione viene sovente specificata mediante un semplice elenco dei valori possibili, affiancati dalle probabilità di ciascun valore. In alcuni casi conviene esprimere tali probabilità tramite una formula.

#### ESEMPIO 3.21

Errori  
di trasmissione  
dei bit

È possibile che un bit trasmesso lungo un canale di trasmissione digitale sia ricevuto in maniera errata. Detto  $X$  il numero di bit errati nei prossimi 4 bit trasmessi, i possibili valori di  $X$  sono  $\{0, 1, 2, 3, 4\}$ . Le probabilità associate ai suddetti valori verranno determinate in base a un modello degli errori, presentato nel seguente paragrafo. Si supponga che tali probabilità siano

$$\begin{aligned} P(X = 0) &= 0.6561 & P(X = 1) &= 0.2916 & P(X = 2) &= 0.0486 \\ P(X = 3) &= 0.0036 & P(X = 4) &= 0.0001 \end{aligned}$$

La distribuzione di probabilità di  $X$  è specificata dai possibili valori di  $X$  accompagnati dalla probabilità di ciascuno di essi. Una descrizione grafica di tale distribuzione è mostrata in Figura 3.29.

Si supponga che una trave lunga e sottile sia sottoposta a carichi discreti, ossia che su di essa poggiino alcune masse, poste solo in certi punti (si veda la Figura 3.30). Il carico può essere allora descritto da una funzione che specifica la massa che preme su ogni punto interessato dal carico. Analogamente, la distribuzione di una variabile aleatoria discreta  $X$  può venire descritta da una funzione che specifica le probabilità associate a ogni possibile valore discreto di  $X$ .

#### Funzione di massa di probabilità

Data una variabile aleatoria discreta  $X$  con possibili valori  $x_1, x_2, \dots, x_n$ , la **funzione di massa di probabilità** (o pmf, *probability mass function*) è definita da

$$f(x_i) = P(X = x_i) \quad (3.13)$$

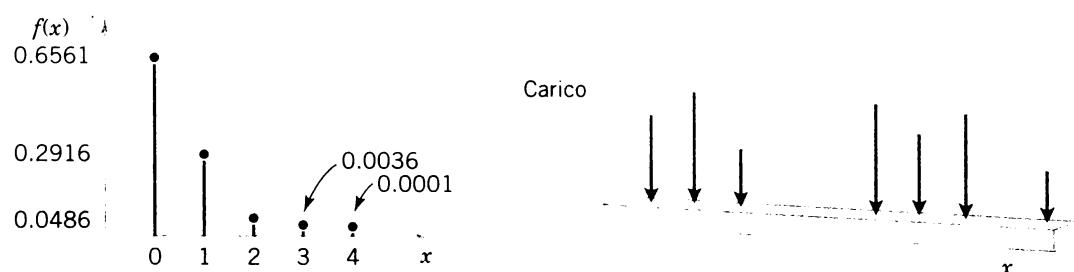
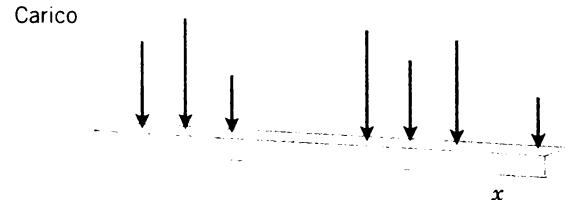


Figura 3.29 Distribuzione di probabilità per la variabile aleatoria  $X$  dell'Esempio 3.21.

Figura 3.30 Carichi discreti su una trave lunga e sottile.



Poiché  $f(x_i)$  è una probabilità, si ha  $f(x_i) \geq 0$  per ogni  $x_i$ , e

$$\sum_{i=1}^n f(x_i) = 1$$

Si lascia al lettore il compito di verificare che la somma delle probabilità dell'esempio precedente è uguale a 1.

La procedura in tre passi utile a determinare la probabilità di una variabile aleatoria descritta nel Paragrafo 3.4.1 si adatta anche alle variabili aleatorie discrete. Ripetiamo qui sotto quali sono i tre passi:

1. Individuare la variabile aleatoria e la relativa distribuzione.
2. Scrivere l'espressione della probabilità di interesse in termini della variabile aleatoria individuata.
3. Calcolare tale probabilità usando la distribuzione individuata.

Questi tre passi vengono seguiti esplicitamente nelle risoluzioni di alcuni esempi di questo capitolo. In altri esempi ed esercizi è possibile utilizzare per proprio conto questa procedura.

### 3.7.2 Funzione di distribuzione cumulativa

Per fornire la distribuzione di probabilità di una variabile aleatoria discreta si può usare anche una funzione di distribuzione cumulativa. La funzione di distribuzione cumulativa per un valore  $x$  è la somma delle probabilità associate a tutti i punti minori o uguali a  $x$ .

**Funzione  
di distribuzione  
cumulativa  
di una variabile  
aleatoria  
discreta**

**La funzione di distribuzione cumulativa** di una variabile aleatoria discreta  $X$  è definita da

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

#### ESEMPIO 3.22

Errori  
di trasmissione  
dei bit

La funzione di massa di probabilità di  $X$  per l'esempio precedente è

$$\begin{aligned} P(X = 0) &= 0.6561 & P(X = 1) &= 0.2916 & P(X = 2) &= 0.0486 \\ P(X = 3) &= 0.0036 & P(X = 4) &= 0.0001 \end{aligned}$$

Pertanto

$$F(0) = 0.6561 \quad F(1) = 0.9477 \quad F(2) = 0.9963 \quad F(3) = 0.9999 \quad F(4) = 1$$

Anche se la variabile aleatoria può assumere solo valori interi, la funzione di distribuzione cumulativa è definita anche per valori non interi. Per esempio

$$F(1.5) = P(X \leq 1.5) = P(X \leq 1) = 0.9477$$

Il grafico di  $F(x)$  è mostrato in Figura 3.31; si noti che ha delle discontinuità (salti) in corrispondenza dei valori discreti di  $X$ . L'ampiezza del salto in un punto  $x$  è uguale alla probabilità in  $x$ . Per esempio, si consideri  $x = 1$ : in tale punto si ha  $F(1) = 0.9477$ , ma per  $0 \leq x < 1$ ,  $F(x) = 0.6561$ . Il salto è  $P(X = 1) = 0.2916$ .

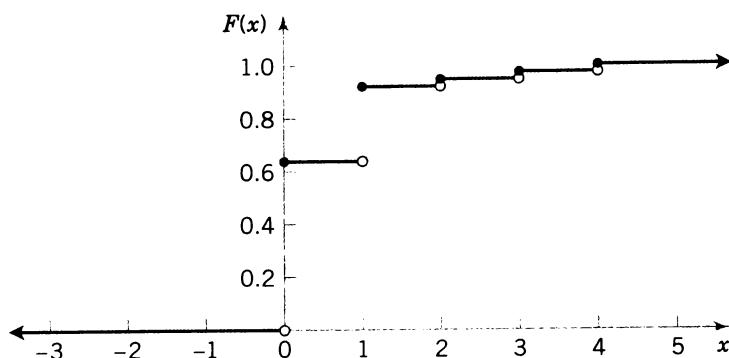


Figura 3.31 Funzione di distribuzione cumulativa per  $X$  (Esempio 3.22).

### 3.7.3 Media e varianza

La media e la varianza di una variabile aleatoria discreta sono definite in modo analogo a quanto avviene per una variabile aleatoria continua, solo che al posto delle sommatorie nelle definizioni compaiono gli integrali.

#### Media e varianza di una variabile aleatoria discreta

Siano  $x_1, x_2, \dots, x_n$  i possibili valori della variabile aleatoria  $X$ . La funzione densità di massa di  $X$  è  $f(x)$ , per cui  $f(x_i) = P(X = x_i)$ .

La **media** o **valore atteso** della variabile aleatoria  $X$ , indicata con  $\mu$  o con  $E(X)$ , è

$$\mu = E(X) = \sum_{i=1}^n x_i f(x_i) \quad (3.14)$$

La **varianza** di  $X$ , indicata con  $\sigma^2$  o  $V(X)$ , è

$$\sigma^2 = V(X) = E(X - \mu)^2 = \sum_{i=1}^n (x_i - \mu)^2 f(x_i) = \sum_{i=1}^n x_i^2 f(x_i) - \mu^2$$

La **deviazione standard** di  $X$  è  $\sigma = \sqrt{V(X)}$ .

La media di  $X$  può essere interpretata come il baricentro dei valori di  $X$ : se in ogni punto  $x_i$  della retta reale collichiamo una massa uguale a  $f(x_i)$ , allora  $E(X)$  rappresenta il punto di equilibrio della retta reale. È possibile interpretare per mezzo di questa analogia meccanica l'espressione "funzione di massa di probabilità".

**ESEMPIO 3.23**  
Errori  
di trasmissione  
dei bit: media  
e varianza

Per la variabile aleatoria dell'esempio precedente si ha

$$\begin{aligned} \mu &= E(X) = 0f(0) + 1f(1) + 2f(2) + 3f(3) + 4f(4) \\ &= 0(0.6561) + 1(0.2916) + 2(0.0486) + 3(0.0036) + 4(0.0001) = 0.4 \end{aligned}$$

Benché  $X$  non assuma mai il valore 0.4, la media pesata dei possibili valori è 0.4.

Per calcolare  $V(X)$  conviene ricorrere a una tabella.

$x$	$x - 0.4$	$(x - 0.4)^2$	$f(x)$	$f(x)(x - 0.4)^2$
0	-0.4	0.16	0.6561	0.104976
1	0.6	0.36	0.2916	0.104976
2	1.6	2.56	0.0486	0.124416
3	2.6	6.76	0.0036	0.024336
4	3.6	12.96	0.0001	0.001296

$$V(X) = \sigma^2 = \sum_{i=1}^5 f(x_i)(x_i - 0.4)^2 = 0.36$$

**ESEMPIO 3.24**  
Ricavi  
di un prodotto

Si devono confrontare due progetti di nuovi prodotti sulla base dei potenziali ricavi. L'ufficio commerciale ritiene che i ricavi derivanti dal progetto A possano essere previsti abbastanza accuratamente in 3 milioni di euro. I ricavi potenziali del progetto B sono invece più difficili da stimare. I responsabili del reparto commerciale giungono alla conclusione che c'è una probabilità pari a 0.3 che i ricavi derivanti dal progetto B siano pari a 7 milioni di euro, ma che vi è una probabilità di 0.7 che siano invece di soli 2 milioni di euro. Quale progetto scegliereste?

Sia  $X$  la variabile che rappresenta i ricavi derivanti dal progetto A. Dato che non vi è incertezza su tali ricavi, possiamo modellizzare la distribuzione della variabile aleatoria  $X$  come 3 milioni di euro con probabilità 1. Perciò:  $E(X) = 3000000 \text{ €}$ .

Sia invece  $Y$  la variabile che rappresenta i ricavi derivanti dal progetto B. Il valore atteso di  $Y$ , in milioni di euro, è

$$E(Y) = € 7(0.3) + € 2(0.7) = € 3.5$$

Poiché  $E(Y)$  è maggiore di  $E(X)$ , potremmo scegliere il progetto B. Tuttavia la variabilità del risultato del progetto B è maggiore; si ha cioè

$$\sigma^2 = (7 - 3.5)^2 \times (0.3) + (2 - 3.5)^2 \times (0.7) = 5.25 \text{ milioni di euro al quadrato}$$

e  $\sigma = \sqrt{5.25} = 2.29$  milioni di euro

### 3.8 DISTRIBUZIONE BINOMIALE

In questo paragrafo presentiamo una variabile aleatoria discreta ampiamente utilizzata. Si considerino i seguenti esperimenti casuali e le relative variabili aleatorie.

1. Si lancia in aria una moneta equilibrata 10 volte. Sia  $X$  il numero di volte che esce testa.
2. Un macchinario usurato produce l'1% di parti difettose. Sia  $X$  il numero di parti difettose nei prossimi 25 pezzi prodotti.
3. Campioni di acqua contengono alti livelli di solido organico nel 10% dei test. Sia  $X$  il numero di campioni con alto contenuto di impurità organiche nei prossimi 18 test.

4. Di tutti i bit trasmessi lungo un canale di trasmissione digitale, il 10% viene ricevuto errato. Sia  $X$  il numero di bit errati nei prossimi 4 bit trasmessi.
5. Un questionario a scelta multipla contiene 10 domande con 4 risposte per ciascuna; l'esaminando risponde a caso a ciascuna domanda. Sia  $X$  il numero di domande cui ha risposto correttamente.
6. Sia  $X$  il numero di neonati di sesso femminile nelle prossime 20 nascite in una clinica.
7. Di tutti i pazienti che soffrono di una particolare malattia, il 35% trae giovamento da una determinata cura. Sia  $X$  il numero di pazienti che risentono di benefici tra i prossimi 30 cui è somministrata la cura.

Tutti questi esempi illustrano l'utilità di una modello probabilistico generale in cui essi possono rientrare come casi particolari.

Ognuno degli esperimenti casuali precedenti può essere visto come costituito da una successione di prove casuali ripetute: 10 lanci della moneta nell'esperimento 1, la produzione di 25 pezzi nell'esperimento 2, e via dicendo. In ciascun caso la variabile aleatoria è un conteggio del numero di prove che soddisfano un determinato criterio. L'esito di ciascuna prova soddisfa il criterio, e viene conteggiato da  $X$ , oppure non lo soddisfa; di conseguenza, di ogni prova si può dire che o ha avuto successo oppure non ha avuto successo. Per esempio, nell'esperimento del questionario a scelta multipla, per ciascuna domanda solo la scelta corretta è considerata un **successo**. La scelta di una qualsiasi delle altre tre risposte errate viene invece registrata come insuccesso.

Il termine "successo" e il termine "insuccesso" sono semplicemente delle etichette. Potremmo usare anche "A" e "B", o "0" e "1". Purtroppo, queste etichette usate abitualmente possono talvolta essere fuorvianti; per esempio, nell'esperimento 2, dato che  $X$  conta le parti difettose, la produzione di una parte difettosa viene chiamata successo.

Una prova con due soli possibili esiti è usata così spesso come elemento di base di un esperimento che per essa è stata coniata l'espressione **prova di Bernoulli**. Si assume di solito che le prove che costituiscono l'esperimento casuale siano **indipendenti**; ciò comporta che l'esito di una prova non ha effetto sull'esito di ogni altra prova. Inoltre, è spesso ragionevole assumere che la **probabilità di un successo in ciascuna prova sia costante**.

Al punto 5, l'esperimento a scelta multipla, se l'esaminando non possiede nozioni sulla materia trattata e prova semplicemente a indovinare le risposte, si può assumere che la probabilità di risposta corretta sia  $1/4$  per ogni domanda. Per analizzare  $X$ , si ricordi l'interpretazione della probabilità come frequenza relativa. La frazione di volte che ci si aspetta che la risposta 1 sia corretta è  $1/4$ , così come per la risposta 2. Nel caso ci si limiti a tirare a indovinare, la frazione di volte che entrambe le risposte sono corrette è

$$(1/4)(1/4) = 1/16$$

La frazione di volte che la risposta 1 è corretta e quella 2 è errata è

$$(1/4)(3/4) = 3/16$$

La frazione di volte che la risposta 1 è errata e la risposta 2 è corretta è

$$(3/4)(1/4) = 3/16$$

Infine, la frazione di volte che la risposta 1 e la risposta 2 sono entrambe errate è

$$(3/4)(3/4) = 9/16$$

In questo modo abbiamo preso in considerazione tutte le possibili combinazioni di risposta corretta o errata per le prime due domande; le quattro probabilità associate a tali combinazioni, sommate fra loro, danno 1

$$1/16 + 3/16 + 3/16 + 9/16 = 1$$

Questo approccio viene usato nel seguente esempio per ricavare la distribuzione binomiale.

**ESEMPIO 3.25**  
Errori  
di trasmissione  
dei bit

In riferimento all’Esempio 3.21, si supponga che la probabilità di ricevere un bit errato quando questo è trasmesso lungo un canale di trasmissione digitale sia pari a 0.1. Inoltre, si assuma che le prove di trasmissione siano indipendenti. Sia  $X$  il numero di bit errati su quattro bit trasmessi. Calcoliamo  $P(X = 2)$ .

Denotiamo con  $E$  un bit errato, e con  $O$  il bit ricevuto senza errori. Possiamo rappresentare gli esiti di questo esperimento come stringa di 4 lettere indicante i bit errati e quelli no. Per esempio, l’esito  $OEOE$  indica che il secondo e il quarto bit sono errati, mentre gli altri due sono ricevuti correttamente. I corrispondenti valori di  $x$  sono:

Esito	$x$	Esito	$x$
$OOOO$	0	$EOOO$	1
$OOOE$	1	$EOOE$	2
$OOEO$	1	$EOEO$	2
$OOEE$	2	$EOEE$	3
$OEOO$	1	$EEOO$	2
$OEOE$	2	$EEOE$	3
$OEEE$	2	$EEEO$	3
$OEEO$	3	$EEEE$	4

L’evento  $X = 2$  consiste nei sei esiti

$$\{EEOO, EOEO, EOOE, OEEO, OEOE, OOEE\}$$

Usando l’ipotesi di indipendenza delle prove, la probabilità di  $\{EEOO\}$  è

$$P(EEOO) = P(E)P(E)P(O)P(O) = (0.1)^2(0.9)^2 = 0.0081$$

Inoltre, i sei esiti incompatibili per cui  $X = 2$  hanno tutti la stessa probabilità di verificarsi. Pertanto

$$P(X = 2) = 6(0.0081) = 0.0486$$

e per questo esempio

$$P(X = x) = (\text{numero di esiti che contengono } x \text{ errori}) \times (0.1)^x \times (0.9)^{4-x}$$

Per completare una formula probabilistica generale (e completare l'Esempio 3.25) è necessaria a questo punto un'espressione per il numero di esiti che contengono esattamente  $x$  successi in  $n$  prove. Si può costruire un esito che contiene  $x$  successi scegliendo dalle  $n$  prove (per esempio le prove 1, 2, 3 e 4) le  $x$  prove (per esempio le prove 2 e 4) che danno i successi.

Il numero di modi in cui si possono selezionare  $x$  oggetti in un gruppo di  $n$ , senza sostituzione, è

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

e questo è il numero di possibili esiti con  $x$  successi. Perciò, per completare l'esempio, abbiamo

$$P(X=x) = \binom{4}{x} (0.1)^x (0.9)^{4-x}$$

Si noti che  $\binom{4}{2} = \frac{4!}{2!2!} = 6$ , come visto in precedenza. La funzione di massa di probabilità di  $X$  è stata mostrata in Figura 3.29.

L'esempio appena illustrato giustifica il seguente risultato.

### Distribuzione binomiale

Un esperimento casuale consistente in  $n$  prove ripetute tali che:

1. le prove sono indipendenti
2. ogni prova ha due possibili esiti, etichettati come “successo” e “insuccesso”
3. la probabilità  $p$  di successo rimane costante per ogni prova

viene detto *esperimento binomiale*.

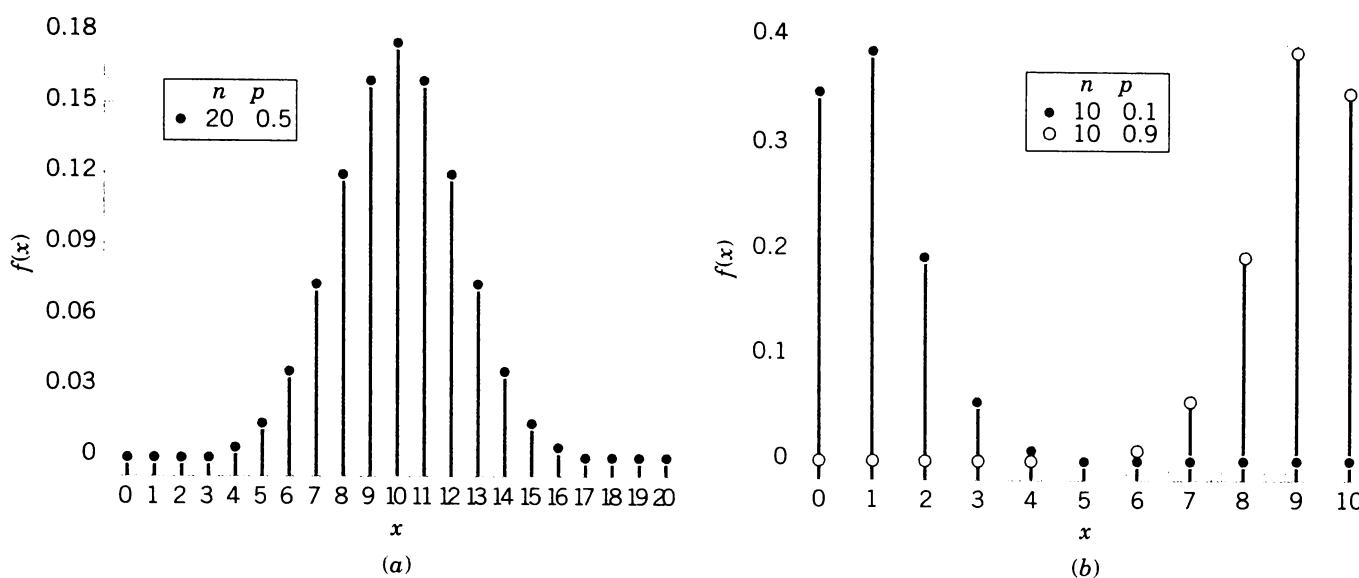
La variabile aleatoria  $X$  uguale al numero di prove che hanno come esito un successo ha una **distribuzione binomiale** con parametri  $p$  e  $n$ , dove  $0 < p < 1$  e  $n = \{1, 2, 3, \dots\}$ . La funzione di massa di probabilità di  $X$  è

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n \quad (3.15)$$

Come in precedenza,  $\binom{n}{x}$  è il numero di successioni od ordinamenti di esiti che contengono  $x$  successi e  $n - x$  insuccessi. Il numero di successioni contenenti  $x$  successi e  $n - x$  insuccessi moltiplicato per la probabilità di ogni successione è pari a  $P(X=x)$ .

Si può dimostrare (usando la formula dello sviluppo binomiale) che la somma delle probabilità per una variabile aleatoria binomiale è 1. Inoltre, dato che ogni prova dell'esperimento ha uno di due esiti possibili, la distribuzione viene detta “binomiale”. Una distribuzione più generale con due o più esiti viene detta per analogia una distribuzione *multinomiale*. L'argomento è trattato per esempio in Montgomery, Runger (2011).

Esempi di distribuzioni binomiali sono mostrati in Figura 3.32. Per  $n$  fissato la distribuzione diventa più simmetrica al crescere di  $p$  da 0 a 0.5 o al decrescere di  $p$  da 1 a 0.5. Per  $p$  fissato, la distribuzione diventa più simmetrica al crescere di  $n$ .

Figura 3.32 Distribuzioni binomiali per alcuni valori di  $n$  e di  $p$ .
**ESEMPIO 3.26**  
 Coefficiente  
 binomiale

Quelli qui presentati sono diversi esempi di calcolo del coefficiente binomiale  $\binom{n}{x}$ .

$$\binom{10}{3} = 10!/[3! \cdot 7!] = (10 \cdot 9 \cdot 8)/(3 \cdot 2) = 120$$

$$\binom{15}{10} = 15!/[10! \cdot 5!] = (15 \cdot 14 \cdot 13 \cdot 12 \cdot 11)/(5 \cdot 4 \cdot 3 \cdot 2) = 3003$$

$$\binom{100}{4} = 100!/[4! \cdot 96!] = (100 \cdot 99 \cdot 98 \cdot 97)/(4 \cdot 3 \cdot 2) = 3\,921\,225$$

**ESEMPIO 3.27**  
 Solidi organici

Ogni campione di acqua ha una probabilità del 10% di contenere alti livelli di solido organico. Si assume che i campioni siano indipendenti rispetto alla presenza del solido. Trovare la probabilità che nei prossimi 18 campioni esattamente 2 contengano alti livelli di solido organico.

Sia  $X$  il numero di campioni che contengono alti livelli di solido nei prossimi 18 campioni analizzati. Allora  $X$  è una variabile aleatoria binomiale con  $p = 0.1$  e  $n = 18$ . Perciò

$$P(X = 2) = \binom{18}{2}(0.1)^2(0.9)^{16}$$

Ora  $\binom{18}{2} = (18!/[2! \cdot 16!]) = 18(17)/2 = 153$ . Pertanto

$$P(X = 2) = 153(0.1)^2(0.9)^{16} = 0.284$$

Troviamo la probabilità che almeno 4 campioni contengano alti livelli di solido

$$P(X \geq 4) = \sum_{x=4}^{18} \binom{18}{x}(0.1)^x(0.9)^{18-x}$$

Definire la variabile aleatoria e la distribuzione.

Scrivere l'espressione della probabilità e calcolare la probabilità.

È tuttavia più semplice usare l'evento complementare

$$\begin{aligned} P(X \geq 4) &= 1 - P(X < 4) = 1 - \sum_{x=0}^3 \binom{18}{x} (0.1)^x (0.9)^{18-x} \\ &= 1 - [0.150 + 0.300 + 0.284 + 0.168] = 0.098 \end{aligned}$$

Inoltre, la probabilità che sia  $3 \leq X < 7$  è

$$\begin{aligned} P(3 \leq X < 7) &= \sum_{x=3}^6 \binom{18}{x} (0.1)^x (0.9)^{18-x} \\ &= 0.168 + 0.070 + 0.022 + 0.005 = 0.265 \end{aligned}$$

La media e la varianza di una variabile aleatoria binomiale dipendono solo dai parametri  $p$  e  $n$ . Si possono dimostrare i risultati seguenti.

Se  $X$  è una variabile aleatoria binomiale con parametri  $p$  e  $n$

$$\mu = E(X) = np \quad \text{e} \quad \sigma^2 = V(X) = np(1-p) \quad (3.16)$$

### ESEMPIO 3.28

Errori  
di trasmissione  
dei bit: media e  
varianza binomiali

Per il numero di bit trasmessi ricevuti errati (Esempio 3.21), si ha  $n = 4$  e  $p = 0.1$ , per cui

$$E(X) = 4(0.1) = 0.4$$

La varianza del numero di bit difettosi è

$$V(X) = 4(0.1)(0.9) = 0.36$$

Questi risultati combaciano con quelli calcolati direttamente dalle probabilità nell'Esempio 3.23.

## 3.9 PROCESSO DI POISSON

Si considerino i messaggi di posta elettronica che arrivano su un server di posta in una rete di computer. Gli arrivi di messaggi sono un esempio di eventi che si verificano in maniera casuale in un intervallo (per esempio un intervallo temporale). Il numero di eventi in un intervallo (come il numero di mail arrivate in 1 ora) è una variabile aleatoria discreta che viene spesso modellizzata tramite una distribuzione di Poisson. L'ampiezza dell'intervallo fra gli eventi (come il tempo che intercorre fra un messaggio e l'altro) viene spesso modellizzata tramite una distribuzione esponenziale. Queste due distribuzioni sono collegate; esse forniscono le probabilità per variabili aleatorie differenti nello stesso esperimento casuale. In Figura 3.33 è descritto graficamente un processo di Poisson.

**Figura 3.33** In un processo di Poisson gli eventi si verificano in maniera casuale in un intervallo.



### 3.9.1 Distribuzione di Poisson

Introduciamo la distribuzione di Poisson con un esempio.

**ESEMPIO 3.29**  
Limite per il numero di bit errati

Si consideri la trasmissione di  $n$  bit su un canale di comunicazione digitale, e sia  $X$  la variabile aleatoria che rappresenta il numero di bit errati. Quando la probabilità di avere un bit errato è costante e le trasmissioni sono indipendenti,  $X$  ha una distribuzione binomiale. Sia  $p$  la probabilità che un bit sia errato. Allora,  $E(X) = pn$ . Si supponga ora che il numero di bit trasmessi aumenti e che la probabilità di un errore diminuisca esattamente quanto basta per mantenere  $pn$  costante, per esempio pari a  $\lambda$ ; in altre parole,  $n$  aumenta e  $p$  diminuisce di conseguenza, di modo che  $E(X)$  rimanga costante. Allora:

$$\begin{aligned} P(X = x) &= \binom{n}{x} p^x (1 - p)^{n-x} \\ &= \frac{n(n-1)(n-2)\cdots(n-x+1)}{n^x x!} (np)^x (1-p)^n (1-p)^{-x} \end{aligned}$$

Eseguendo qualche calcolo si può dimostrare che

$$\lim_{n \rightarrow \infty} P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

Inoltre, poiché il numero di bit trasmessi tende all'infinito, il numero di errori può essere uguale a qualsiasi intero non negativo; i possibili valori di  $X$  sono dunque gli interi da zero a infinito.

La distribuzione ottenuta come limite nell'esempio precedente è più utile di quanto dica il modo in cui è stata ricavata. Il prossimo esempio ne illustra la più vasta applicabilità.

**ESEMPIO 3.30**  
Difetti in un filo di rame

In un filo di rame sottile i difetti si presentano in maniera casuale. Sia  $X$  la variabile aleatoria che conta il numero di difetti in una lunghezza pari a  $L$  millimetri di filo, e si supponga che il numero medio di difetti in  $L$  millimetri sia  $\lambda$ .

Si può trovare la distribuzione di probabilità di  $X$  con un ragionamento analogo a quello dell'Esempio 3.29. Si divide la lunghezza del filo in  $n$  sottointervalli di piccola lunghezza, per esempio un micron ciascuno. Se i sottointervalli sono abbastanza piccoli, la probabilità che si riscontri più di un difetto in uno di essi è trascurabile. Inoltre, l'assunzione che i difetti si presentano a caso si traduce nel dire che ogni sottointervalllo ha la stessa probabilità  $p$  di contenere un difetto. Infine, se supponiamo che un sottointervalllo contenga un difetto indipendentemente da quello che succede negli altri sottointervallli, possiamo modellizzare la distribuzione di  $X$  approssimativamente come quella di una variabile aleatoria binomiale. Poiché

$$E(X) = \lambda = np$$

otteniamo

$$p = \lambda/n$$

In altri termini: la probabilità che un sottointervallo contenga un difetto è  $\lambda/n$ . Con sottointervalli abbastanza piccoli  $n$  è molto grande e  $p$  molto piccola. Perciò la distribuzione di  $X$  si ottiene al tendere a zero della lunghezza del sottointervallo, come nel precedente esempio.

Chiaramente, l'Esempio 3.29 può essere generalizzato in modo da includere una vasta gamma di esperimenti casuali. L'intervallo che abbiamo suddiviso era la lunghezza di un filo, ma lo stesso ragionamento potrebbe essere applicato a qualsiasi intervallo, compresi quelli temporali, le aree o i volumi. Per esempio, i conteggi (1) delle particelle contaminanti nella produzione di semiconduttori, (2) dei difetti in rotoli di tessuto, (3) delle chiamate a un centralino, (4) delle interruzioni della corrente elettrica e (5) delle particelle atomiche emesse da un campione sono stati tutti modellizzati con successo dalla funzione di massa di probabilità della definizione che segue.

In generale, si consideri un intervallo  $T$  di numeri reali suddiviso in piccoli sottointervalli di ampiezza  $\Delta t$  e si ipotizzi che per  $\Delta t$  che tende a zero:

1. la probabilità di più di un evento in un sottointervallo tenda a zero;
2. la probabilità di un evento in un sottointervallo sia asintotica a  $\lambda\Delta t / T$ ;
3. l'evento in ogni sottointervallo sia indipendente dagli altri sottointervalli.

Un esperimento casuale con queste proprietà viene detto **processo di Poisson**.

Le ipotesi precedenti implicano che i sottointervalli possono essere visti come prove di Bernoulli approssimativamente indipendenti con probabilità di successo  $p = \lambda\Delta t / T$  e che il numero di prove è uguale a  $n = T / \Delta t$ . Pertanto,  $pn = \lambda$ , e per  $\Delta t$  che tende a zero  $n$  tende a infinito; risulta evidente la somiglianza con il limite dell'Esempio 3.30. Tutto ciò porta alla seguente definizione.

### Distribuzione di Poisson

La variabile aleatoria  $X$  uguale al numero di eventi in un processo di Poisson è una **variabile aleatoria di Poisson** con parametro  $\lambda > 0$ , e la funzione di massa di probabilità di  $X$  è

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \dots \quad (3.17)$$

La media e la varianza di  $X$  sono

$$E(X) = \lambda \quad \text{e} \quad V(X) = \lambda \quad (3.18)$$

Storicamente, il termine “processo” è stato adoperato per suggerire l’idea dell’osservazione di un sistema nel tempo. Nel nostro esempio relativo al filo in rame abbiamo mostrato che la distribuzione di Poisson si potrebbe applicare anche a intervalli quali le lunghezze. La Figura 3.34 mostra i grafici di alcune distribuzioni di Poisson.

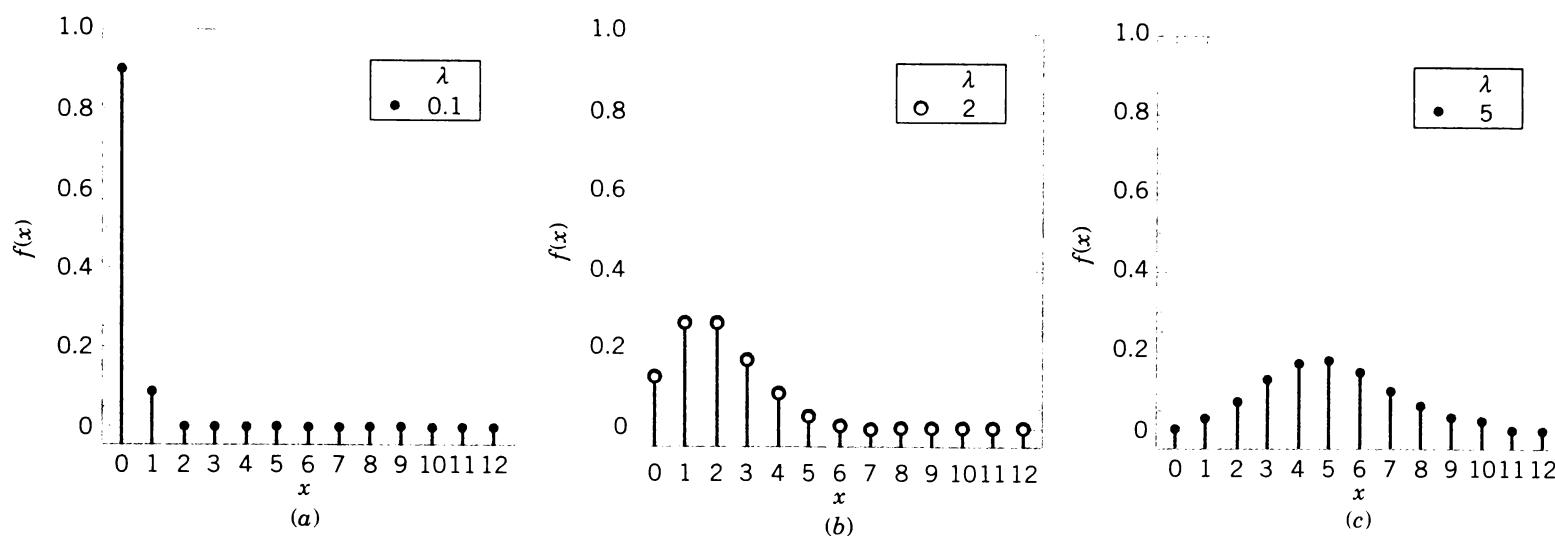


Figura 3.34 Distribuzione di Poisson per alcuni valori del parametro  $\lambda$ .

È importante usare **unità di misura coerenti** nel calcolo di probabilità, medie e varianze che coinvolgono variabili aleatorie di Poisson. Il seguente esempio illustra le conversioni delle unità di misura. Per esempio, se il

numero medio di difetti per millimetro di filo è 3.4, allora il  
numero medio di difetti in 10 mm di filo è 34, e il  
numero medio di difetti in 100 mm di filo è 340

Se la variabile aleatoria di Poisson rappresenta il numero di conteggi in qualche intervallo, la media della variabile aleatoria è uguale al numero atteso di conteggi nello stesso intervallo.

### ESEMPIO 3.31 Probabilità per i difetti in un filo di rame

In riferimento all'esempio del filo in rame, si supponga che il numero di difetti segua una distribuzione di Poisson con media di 2.3 difetti per millimetro. Calcoliamo la probabilità di avere esattamente 2 difetti in un millimetro di filo.

Indichiamo con  $X$  il numero di difetti in un millimetro di filo. Allora  $E(X) = 2.3$  difetti e

$$P(X = 2) = \frac{e^{-2.3} 2.3^2}{2!} = 0.265$$

Calcoliamo la probabilità di 10 difetti in 5 mm di filo; indicando il numero di difetti in 5 mm con  $X$ , quest'ultima ha una distribuzione di Poisson con

$$E(X) = 5 \text{ mm} \times 2.3 \text{ difetti/mm} = 11.5 \text{ difetti}$$

Pertanto

$$P(X = 10) = e^{-11.5} 11.5^{10}/10! = 0.113$$

Calcoliamo ora la probabilità di avere almeno un difetto in 2 mm di filo. Se  $X$  denota il numero di difetti in 2 mm, allora  $X$  ha una distribuzione di Poisson con

$$E(X) = 2 \text{ mm} \times 2.3 \text{ difetti/mm} = 4.6 \text{ difetti}$$

Pertanto:

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ &= 1 - e^{-4.6} = 0.9899 \end{aligned}$$

Il prossimo esempio fa uso di un software per addizionare le probabilità di Poisson.

### ESEMPIO 3.32

#### Contaminazione di dischi ottici

Nella realizzazione di memorie ottiche la contaminazione costituisce un serio problema. Il numero di particelle contaminanti che si presentano in un disco ottico ha una distribuzione di Poisson, e il numero medio di particelle per centimetro quadrato di superficie del supporto è 0.1. L'area di un disco sotto esame è  $100 \text{ cm}^2$ . Calcoliamo la probabilità di trovare 12 particelle nell'area del disco esaminato.

Sia  $X$  il numero di particelle nell'area esaminata. Dato che il numero medio di particelle è 0.1 particelle per centimetro quadrato, si ha

$$\begin{aligned} E(X) &= 100 \text{ cm}^2 \times 0.1 \text{ particelle/cm}^2 \\ &= 10 \text{ particelle} \end{aligned}$$

Perciò

$$P(X = 12) = \frac{e^{-10} 10^{12}}{12!} = 0.095$$

Calcoliamo la probabilità di trovare zero particelle nell'area del disco esaminata. Si ha:  $P(X = 0) = e^{-10} = 4.54 \times 10^{-5}$ .

Calcoliamo invece la probabilità di trovare 12 o meno particelle nell'area del disco esaminata. Tale probabilità vale

$$\begin{aligned} P(X \leq 12) &= P(X = 0) + P(X = 1) + \cdots + P(X = 12) \\ &= \sum_{i=0}^{12} \frac{e^{-10} 10^i}{i!} \end{aligned}$$

Essendo questa sommatoria noiosa da calcolare, si può ricorrere a uno dei numerosi software che calcolano le probabilità cumulate di Poisson. Con Minitab si ricava  $P(X \leq 12) = 0.7916$ .

Si è detto che la varianza di una variabile aleatoria di Poisson è uguale alla sua media. Per esempio, se i conteggi delle particelle seguono una distribuzione di Poisson con media 25 particelle per centimetro quadrato, la deviazione standard dei conteggi è 5 per centimetro quadrato. Di conseguenza, si possono ottenere facilmente informazioni sulla variabilità. Viceversa, se la varianza dei dati di conteggio è molto maggiore della media dei dati stessi, la distribuzione di Poisson non è un buon modello per la distribuzione della variabile aleatoria.

### 3.9.2 Distribuzione esponenziale

Parlando della distribuzione di Poisson abbiamo definito una variabile aleatoria che conta il numero di difetti lungo un filo in rame. La distanza fra i difetti è un'altra variabile che risul-

ta spesso interessante. Sia  $X$  la variabile aleatoria che rappresenta la lunghezza a partire da un qualsiasi punto sul filo sino al primo difetto. Come ci si potrebbe attendere, si può ottenere la distribuzione di  $X$  dalla conoscenza della distribuzione del numero di difetti. La chiave per la relazione fra le due distribuzioni è fornita dal seguente concetto. La distanza sino al primo difetto supera i 3 mm se e solo se non ci sono difetti entro una lunghezza di 3 mm: semplice, ma sufficiente per un'analisi della distribuzione di  $X$ .

In generale, sia  $N$  la variabile aleatoria che rappresenta il numero di difetti in  $x$  millimetri di filo. Se il numero medio di difetti è  $\lambda$  per millimetro,  $N$  ha una distribuzione di Poisson con media  $\lambda x$ . Ora

$$P(X > x) = P(N = 0) = \frac{e^{-\lambda x} (\lambda x)^0}{0!} = e^{-\lambda x}$$

e

$$P(X \leq x) = 1 - e^{-\lambda x}$$

per  $x \geq 0$ . Se  $f(x)$  è la funzione di densità di probabilità di  $X$ , la funzione di distribuzione cumulativa è

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du$$

Per il teorema fondamentale del calcolo integrale, la derivata di  $F(x)$  rispetto a  $x$  è  $f(x)$ . Però, la funzione di densità di probabilità di  $X$  è

$$f(x) = \frac{d}{dx} (1 - e^{-\lambda x}) = \lambda e^{-\lambda x} \quad \text{per } x \geq 0$$

La distribuzione di  $X$  dipende solo dall'ipotesi che i difetti nel filo seguano un processo di Poisson. Inoltre, il punto iniziale per la misura di  $X$  non ha rilevanza, perché la probabilità del numero di difetti in un intervallo di un processo di Poisson dipende solo dalla lunghezza dell'intervallo, e non dalla sua posizione. Per tutti i processi di Poisson vale il seguente risultato generale.

### Distribuzione esponenziale

La variabile aleatoria  $X$  pari alla distanza fra successivi conteggi di un processo di Poisson con media per unità di misura  $\lambda > 0$  ha una distribuzione esponenziale con parametro  $\lambda$ . La funzione di densità di probabilità di  $X$  è

$$f(x) = \lambda e^{-\lambda x}, \quad \text{per } 0 \leq x < \infty \tag{3.19}$$

La media e la varianza di  $X$  sono

$$E(X) = \frac{1}{\lambda} \quad \text{e} \quad V(X) = \frac{1}{\lambda^2} \tag{3.20}$$

Il nome della distribuzione esponenziale deriva dalla funzione esponenziale che compare nella funzione di densità di probabilità. In Figura 3.35 sono mostrati alcuni grafici della distribuzione esponenziale per diversi valori di  $\lambda$ . Per ogni valore di  $\lambda$  la distribuzione esponenziale è alquanto asimmetrica. Le formule per la media e la varianza si ottengono mediante integrazione per parti. Si noti, poi, che la distribuzione esponenziale è un caso particolare di due distribuzioni continue che abbiamo studiato nelle pagine precedenti. La distribuzione di Weibull con  $\beta = 1$  si riduce alla distribuzione esponenziale, e la distribuzione gamma con  $\gamma = 1$  è una distribuzione esponenziale. La distribuzione gamma è ricavabile come somma di  $\gamma$  variabili aleatorie esponenziali indipendenti.

È importante usare **unità di misura coerenti** nel calcolo di probabilità, medie e varianze che coinvolgono variabili aleatorie esponenziali. Il seguente esempio illustra le conversioni delle unità di misura.

### ESEMPIO 3.33

Connessioni  
alla rete

Individuare la variabile  
aleatoria e la  
distribuzione.

Scrivere l'espressione  
della probabilità

In una ampia rete aziendale di computer, le connessioni degli utenti al sistema possono venire modellizzate come un processo di Poisson con media di 25 connessioni all'ora. Qual è la probabilità che non vi siano connessioni in un intervallo di 6 minuti?

Indichiamo con  $X$  il tempo (in ore) dall'inizio dell'intervallo sino alla prima connessione.  $X$  ha allora una distribuzione esponenziale con  $\lambda = 25$  connessioni/ora. Vogliamo trovare la probabilità che  $X$  superi i 6 minuti. Dato che  $\lambda$  è espressa in connessioni all'ora, esprimiamo tutte le unità di tempo in ore: abbiamo ovviamente  $6 \text{ min} = 0.1 \text{ h}$ . La probabilità richiesta è mostrata come area ombreggiata al di sotto del grafico della funzione di densità di probabilità in Figura 3.36.

Pertanto

$$P(X > 0.1) = \int_{0.1}^{\infty} 25e^{-25x} dx = e^{-25(0.1)} = 0.082$$

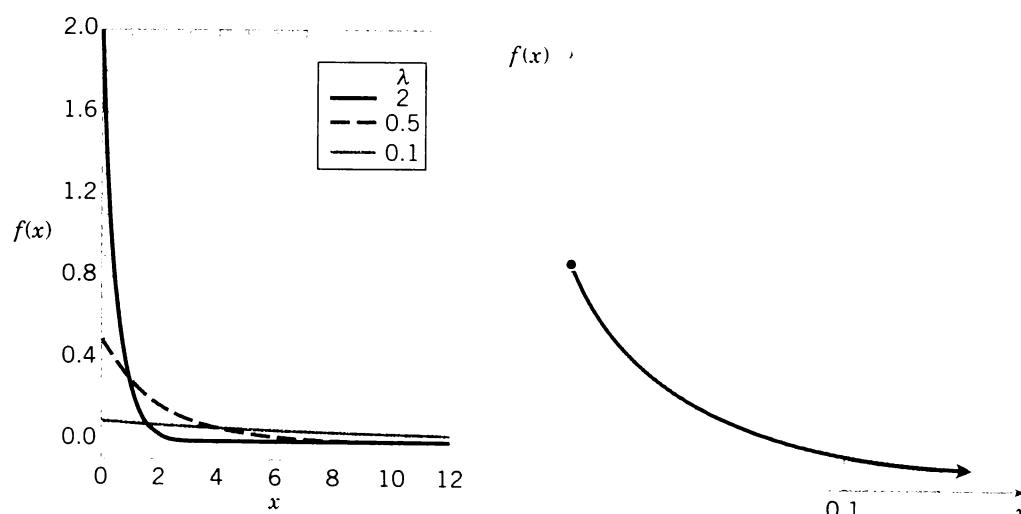


Figura 3.35 Funzione di densità di probabilità di una variabile aleatoria esponenziale per alcuni valori di  $\lambda$ .

Figura 3.36 Probabilità per la distribuzione esponenziale dell'Esempio 3.33.

**Calcolare la probabilità.**

Un risultato identico si ottiene esprimendo il numero medio di connessioni come 0.417 connessioni al minuto e calcolando la probabilità che il tempo occorrente sino alla successiva connessione superi i 6 minuti. Provate!

Qual è la probabilità che il tempo occorrente sino alla successiva connessione sia compreso fra 2 e 3 minuti? Convertendo tutte le unità in ore, si ha

$$P(0.033 < X < 0.05) = \int_{0.033}^{0.05} 25e^{-25x} dx = -e^{-25x} \Big|_{0.033}^{0.05} = 0.152$$

Troviamo l'intervallo di tempo in cui la probabilità di zero connessioni è 0.90. Dobbiamo allo scopo ricavare il tempo  $x$  tale che  $P(X > x) = 0.90$ . All'inizio di questo paragrafo abbiamo trovato che  $P(X > x) = e^{-\lambda x}$ . Ora

$$P(X > x) = e^{-25x} = 0.90$$

Pertanto, passando ai logaritmi in entrambi i membri, ricaviamo

$$x = 0.00421 \text{ ore} = 0.25 \text{ minuti}$$

Inoltre, il tempo medio sino alla successiva connessione è

$$E(X) = 1/25 \text{ ore} = 0.04 \text{ ore} = 2.4 \text{ minuti}$$

La deviazione standard del tempo sino alla prossima connessione è

$$\sigma_X = 1/25 \text{ ore} = 2.4 \text{ minuti}$$

**Proprietà di assenza di memoria**

Nel precedente esempio, la probabilità di zero connessioni in un intervallo di 6 minuti valeva 0.082 indipendentemente dall'istante iniziale dell'intervallo. In un processo di Poisson vale l'assunzione che gli eventi si verifichino uniformemente in tutto l'intervallo di osservazione, ossia che non vi sia raggruppamento degli eventi. Se le connessioni sono modellizzate da un processo di Poisson, la probabilità che la prima connessione dopo mezzogiorno si verifichi dopo le 12:06 è uguale alla probabilità che la prima connessione dopo le 15 si verifichi dopo le 15:06. Se qualcuno si connette alle 14:22, la probabilità che la successiva connessione avvenga dopo le 14:28 è ancora 0.082.

Il nostro punto di partenza per l'osservazione del sistema, dunque, non ha importanza. Tuttavia, se vi è un periodo di intenso utilizzo durante il giorno, come l'orario immediatamente successivo alle 8:00, seguito da un periodo di scarso impiego del sistema, un processo di Poisson non è un modello appropriato per le connessioni, e la distribuzione non è adatta per il calcolo delle probabilità. Può essere ragionevole modellizzare separatamente ciascun periodo di intenso o scarso utilizzo mediante processi di Poisson distinti, impiegando un valore maggiore di  $\lambda$  durante il periodo di maggiore utilizzo e uno minore durante quello di minore impiego. Si può quindi utilizzare una distribuzione esponenziale con il corrispondente valore di  $\lambda$  per calcolare le probabilità di connessione per i periodi di intenso e scarso utilizzo.

Una proprietà ancora più interessante di una variabile aleatoria esponenziale è la **proprietà di assenza di memoria**. Si supponga che non vi siano connessioni dalle 12:00 alle 12:15; la probabilità che non vi siano connessioni dalle 12:15 alle 12:21 è ancora 0.082. Poiché abbiamo già atteso 15 minuti, abbiamo la sensazione di “essere in debito”, ossia che la probabilità di una connessione nei successivi 6 minuti dovrebbe essere maggiore di 0.082. In effetti, per una distribuzione esponenziale ciò non è vero.

La proprietà di assenza di memoria non sorprende più di tanto se si considera come abbiammo ricavato un processo di Poisson. In quel procedimento abbiamo assunto che un intervallo potesse venire suddiviso in piccoli sottointervalli indipendenti. La presenza o l'assenza di eventi in tali sottointervalli è analoga alle prove indipendenti di Bernoulli in un processo binomiale; la conoscenza dei risultati precedenti non influenza le probabilità degli eventi nei sottointervalli futuri.

La distribuzione esponenziale viene spesso utilizzata negli **studi di affidabilità** come modello per il tempo sino al guasto di un dispositivo. Per esempio, la durata di un chip può essere modellizzata da una variabile aleatoria esponenziale con media 40 000 ore. La proprietà di assenza di memoria della distribuzione esponenziale comporta il fatto che il dispositivo non si usura; cioè, indipendentemente da quanto tempo ha operato il dispositivo, la probabilità di un guasto nelle successive 1000 ore è la stessa di un guasto nelle prime 1000 ore di funzionamento. La durata di un dispositivo con guasti causati da shock casuali può essere appropriatamente modellizzata come variabile aleatoria esponenziale; tuttavia, per la durata di un dispositivo che è soggetto a una leggera usura meccanica, quale un cuscinetto, è più adatto un modello costituito da una distribuzione che non perde la memoria, come la distribuzione di Weibull (con  $\beta \neq 1$ ).

### 3.10 APPROSSIMAZIONE NORMALE DELLE DISTRIBUZIONI BINOMIALE E DI POISSON

#### Approssimazione della distribuzione binomiale

Essendo una variabile aleatoria binomiale un conteggio ricavato da prove indipendenti ripetute, si può applicare a essa il teorema limite centrale. Di conseguenza, non dovrebbe sorprendere che per approssimare le probabilità binomiali nei casi in cui  $n$  è elevato si usi la distribuzione normale. Il seguente esempio mostra che per molti sistemi fisici è appropriato il modello binomiale con un valore altissimo di  $n$ . In questi casi è difficile calcolare le probabilità usando la distribuzione binomiale. Fortunatamente, in tali casi l'approssimazione normale risulta assai efficace (si faccia riferimento alla Figura 3.37). Ogni barra in figura ha ampiezza unitaria, per cui l'area della barra su un valore  $x$  è uguale alla probabilità binomiale di  $x$ . Alle barre è sovrapposta una distribuzione normale con  $\mu = np = 5$  e  $\sigma^2 = np(1 - p) = 2.5$ . Si noti che l'area delle barre (la probabilità binomiale) può essere approssimata dall'area sottesa dalla curva normale (la probabilità ottenuta dalla distribuzione normale).

**ESEMPIO 3.34**  
Errori  
di trasmissione  
dei bit:  
dimensione  
campionaria  
elevata

Si supponga che in un canale di comunicazione digitale il numero di bit ricevuti errati possa venire modellizzato da una variabile aleatoria binomiale, e che la probabilità che un bit venga ricevuto errato sia  $10^{-5}$ . Se sono trasmessi 16 milioni di bit, qual è la probabilità che si verifichino più di 150 errori?

Sia  $X$  la variabile aleatoria che rappresenta il numero di errori.  $X$  è allora una variabile aleatoria binomiale e

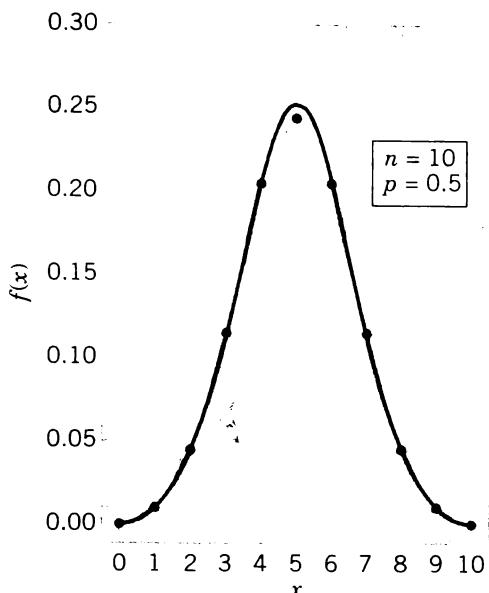


Figura 3.37 Approssimazione normale della distribuzione binomiale.

$$\begin{aligned} P(X > 150) &= 1 - P(X \leq 150) \\ &= 1 - \sum_{x=0}^{150} \binom{16\,000\,000}{x} (10^{-5})^x (1 - 10^{-5})^{16\,000\,000 - x} \end{aligned}$$

È chiaro che la probabilità del precedente esempio è difficile da calcolare. Per fortuna, in questo caso si può usare la distribuzione normale per ricavare un'eccellente approssimazione.

Se  $X$  è una variabile aleatoria binomiale,

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \quad (3.21)$$

è approssimativamente una variabile aleatoria normale standard. Di conseguenza, si possono usare le probabilità calcolate in base a  $Z$  per approssimare le probabilità di  $X$ .

Rammentiamo che per una variabile binomiale  $X$  si ha  $E(X) = np$  e  $V(X) = np(1 - p)$ . Pertanto, l'approssimazione normale non è altro che la formula per la standardizzazione della variabile aleatoria  $X$ . Le probabilità che coinvolgono  $X$  possono essere approssimate usando una variabile aleatoria normale standard. L'approssimazione normale della distribuzione binomiale è valida se  $n$  è sufficientemente grande rispetto a  $p$ ; in particolare, ogni volta che

$$np > 5 \quad \text{e} \quad n(1 - p) > 5$$

Per migliorare ulteriormente l'approssimazione è possibile utilizzare un fattore correttivo (detto **fattore di continuità**). Si noti, in Figura 3.37 che l'area delle barre verticali che rap-

presenta una probabilità binomiale come  $P(4 < X \leq 7) = P(X = 5) + P(X = 6) + P(X = 7)$  è bene approssimata dall'area sottesa dalla curva normale fra 4.5 e 7.5. Si noti altresì che  $P(X = 6)$  è bene approssimata dall'area sottesa dalla curva normale fra 6.5 e 7.5. Di conseguenza, si aggiunge  $\pm \frac{1}{2}$  ai valori binomiali per migliorare l'approssimazione. Una regola pratica prevede di applicare il fattore correttivo  $\pm \frac{1}{2}$  in modo che aumenti la probabilità binomiale da approssimare.

Il problema della comunicazione digitale viene risolto come segue

$$\begin{aligned} P(X > 150) &= P(X \geq 151) \cong P\left(\frac{X - 160}{\sqrt{160(1 - 10^{-5})}} > \frac{150,5 - 160}{\sqrt{160(1 - 10^{-5})}}\right) \\ &= P(Z > -0.75) = P(Z < 0.75) = 0.773 \end{aligned}$$

Si osservi che dopo aver scritto la probabilità binomiale con il simbolo di maggiore o uguale come  $P(X \geq 151)$ , il fattore correttivo porta a sottrarre  $\frac{1}{2}$  da  $X$  per aumentare la probabilità.

**ESEMPIO 3.35**  
Errori  
di trasmissione  
dei bit:  
approssimazione  
normale

Consideriamo nuovamente la trasmissione di bit del precedente esempio. Per valutare quanto efficacemente funziona l'approssimazione normale si supponga che solo  $n = 50$  bit debbano essere trasmessi, e che la probabilità di un errore sia  $p = 0.1$ . L'esatta probabilità che si verifichino 2 o meno errori è

$$P(X \leq 2) = \binom{50}{0} 0.9^{50} + \binom{50}{1} 0.1(0.9^{49}) + \binom{50}{2} 0.1^2(0.9^{48}) = 0.11$$

In base all'approssimazione normale si ha

$$P(X \leq 2) = P\left(\frac{X - 5}{\sqrt{50(0.1)(0.9)}} < \frac{2.5 - 5}{\sqrt{50(0.1)(0.9)}}\right) \cong P(Z < -1.18) = 0.12$$

Per un campione di 50 bit, con  $np = 5$ , l'approssimazione normale è ragionevole.

Tuttavia, se  $np$  o  $n(1-p)$  sono piccoli, la distribuzione binomiale è alquanto asimmetrica e la distribuzione normale simmetrica non è una buona approssimazione. Due casi di questo tipo sono illustrati in Figura 3.38.

#### Approssimazione della distribuzione di Poisson

Si ricordi che la distribuzione di Poisson è stata sviluppata come limite di una distribuzione binomiale in cui il numero di prove tende a infinito. Di conseguenza, anche per approssimare le probabilità di una variabile aleatoria di Poisson si può usare la distribuzione normale. L'approssimazione è buona per

$$\lambda > 5$$

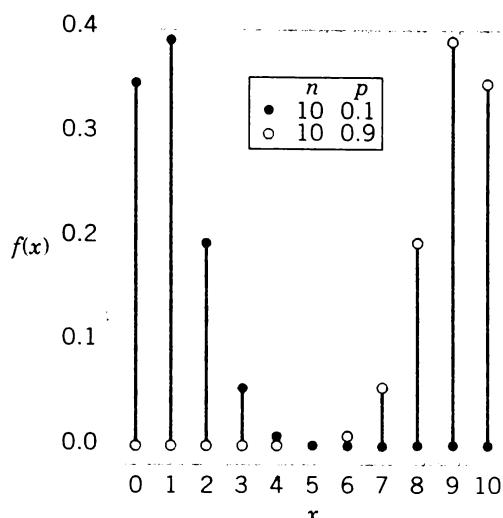


Figura 3.38 La distribuzione binomiale non è simmetrica se  $p$  è prossimo a 0 o a 1.

Se  $X$  è una variabile aleatoria di Poisson con  $E(X) = \lambda$  e  $V(X) = \lambda$ ,

$$Z = \frac{X - \lambda}{\sqrt{\lambda}} \quad (3.22)$$

è approssimativamente una variabile aleatoria normale standard.

### ESEMPIO 3.36 Contaminanti nell'acqua

Si supponga che il numero di particelle contaminanti in un litro d'acqua segua una distribuzione di Poisson con media 1000. Se viene analizzato un campione, qual è la probabilità di trovare meno di 950 particelle?

La probabilità richiesta può essere espressa esattamente da

$$P(X \leq 950) = \sum_{x=0}^{950} \frac{e^{-1000} 1000^x}{x!}$$

È evidente la difficoltà del calcolo. La probabilità può essere approssimata come segue

$$P(X \leq x) \cong P\left(Z \leq \frac{950.5 - 1000}{\sqrt{1000}}\right) = P(Z \leq -1.57) = 0.059$$

Approssimiamo ora la probabilità di trovare più di 25 particelle in 20 millilitri d'acqua.

Se il numero medio di particelle per litro è 1000, la media per millilitro è 1, e quella per 20 millilitri è 20. Denotiamo con  $X$  il numero di particelle in 20 millilitri; allora  $X$  ha una distribuzione di Poisson con media 20 e la probabilità richiesta è

$$P(X > 25) \cong P\left(Z > \frac{25.5 - 20}{\sqrt{20}}\right) = P(Z > 1.22) = 0.109$$

## 3.11 PIÙ VARIABILI ALEATORIE E INDIPENDENZA

### 3.11.1 Distribuzioni congiunte

In molti esperimenti vengono misurate più variabili. Per esempio, si supponga che di un disco stampato a iniezione siano misurati sia il diametro che lo spessore, indicati rispettivamente con  $X$  e  $Y$ . Queste due variabili aleatorie sono spesso collegate: se aumenta la pressione nello stampo, vi può essere un aumento del riempimento dello stampo stesso che dà luogo a valori più alti di  $X$  e  $Y$ . Analogamente, una diminuzione della pressione può portare a valori più bassi sia per  $X$  che per  $Y$ . Immaginiamo che le misure di diametro e di spessore eseguite su vari pezzi siano riportate su un grafico di dispersione nel piano  $X-Y$ . Come mostra la Figura 3.39, la relazione fra  $X$  e  $Y$  comporta che in alcune regioni del piano  $X-Y$  vi saranno verosimilmente più misure che non in altre. Questo concetto è stato discusso nel Paragrafo 2.6, allorché abbiamo definito il coefficiente di correlazione campionario.

Questa tendenza può venire modellizzata da una funzione di densità di probabilità [indicata con  $f(x, y)$ ] definita sul piano  $X-Y$  (Figura 3.40). Le analogie che ponevano in relazione una funzione di densità di probabilità al carico su una trave lunga e sottile possono venire applicate per porre in relazione questa funzione di densità di probabilità bidimensionale alla densità di un carico su una superficie piatta e larga. La probabilità che l'esperimento casuale (la produzione di pezzi) generi misure contenute in una regione del piano  $X-Y$  viene determinata dall'integrale di  $f(x, y)$  sulla regione, come mostra la Figura 3.41. L'integrale è uguale al volume racchiuso da  $f(x, y)$  sulla regione. Poiché  $f(x, y)$  determina le probabilità per due variabili aleatorie, viene chiamata **funzione di densità di probabilità congiunta**. Dalla Figura 3.41 si vede che la probabilità che un pezzo sia prodotto con misure appartenenti alla regione indicata è

$$P(a < X < b, c < Y < d) = \int_a^b \int_c^d f(x, y) dy dx$$

Riflessioni analoghe possono essere applicate alle variabili aleatorie discrete. Per esempio, si supponga che la qualità di ogni bit ricevuto attraverso un canale di comunicazione digitale sia classificata nelle quattro categorie “eccellente”, “buona”, “discreta” e “scarsa”, contrassegnate rispettivamente con le lettere E, B, D e S. Siano  $X$ ,  $Y$ ,  $W$  e  $Z$  i numeri di bit rispettivamente di categoria E, B, D e S in una trasmissione di 20 bit. In questo esempio siamo interessati alla

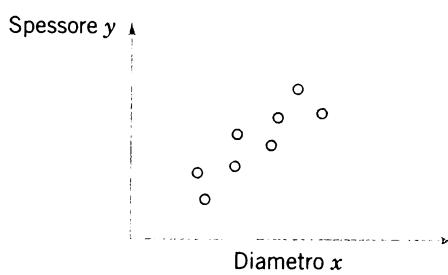


Figura 3.39 Diagramma di dispersione delle misure di diametro e spessore.

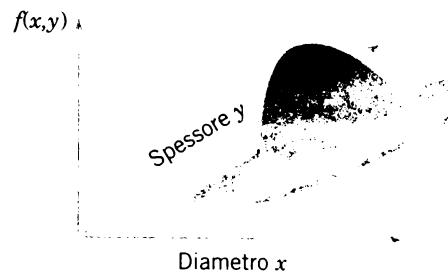


Figura 3.40 Funzione di densità di probabilità congiunta nelle variabili  $x$  e  $y$ .

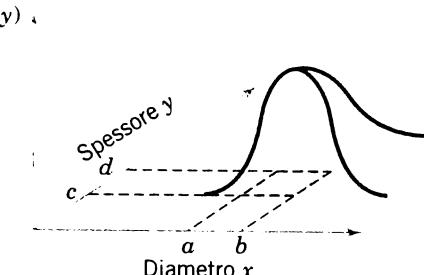


Figura 3.41 La probabilità di una regione è data dal volume racchiuso da  $f(x, y)$  sulla regione.

distribuzione di probabilità congiunta di quattro variabili aleatorie. Per semplificare, consideriamo solo  $X$  e  $Y$ . La distribuzione di probabilità congiunta di  $X$  e  $Y$  può essere specificata da una funzione di massa di probabilità congiunta  $f(x, y) = P(X = x, Y = y)$ . Poiché ciascuno dei 20 bit è inserito in una delle quattro classi, si ha  $X + Y + W + Z = 20$ , perciò solo gli interi per cui  $X + Y \leq 20$  hanno probabilità positiva nella funzione di massa di probabilità congiunta di  $X$  e  $Y$ . Quest'ultima è nulla altrove. Per una discussione generale delle distribuzioni congiunte il lettore interessato può consultare Montgomery, Runger (2011). Qui concentriamo l'attenzione, piuttosto, sull'importante caso speciale di variabili aleatorie indipendenti.

### 3.11.2 Indipendenza

Facendo qualche assunzione sui nostri modelli probabilistici, si riesce spesso a semplificare una probabilità che coinvolge più di una variabile aleatoria. Nell'Esempio 3.13 si è determinato il valore 0.919 come probabilità che un diametro soddisfi le specifiche. Che cosa possiamo dire su 10 di tali diametri? Qual è la probabilità che tutti quanti soddisfino le specifiche? Questo è il tipo di domande che interessano gli acquirenti di lettori ottici.

Le domande precedenti conducono a un importante concetto e alla relativa definizione. Per tenere conto di un numero di variabili superiore a due adottiamo la notazione  $X_1, X_2, \dots, X_n$  per rappresentare  $n$  variabili aleatorie.

#### Indipendenza

Le variabili aleatorie  $X_1, X_2, \dots, X_n$  sono **indipendenti** se

$$P(X_1 \in E_1, X_2 \in E_2, \dots, X_n \in E_n) = P(X_1 \in E_1)P(X_2 \in E_2) \cdots P(X_n \in E_n)$$

per *ogni* scelta degli insiemi  $E_1, E_2, \dots, E_n$ .

L'importanza del concetto di indipendenza è illustrata dal seguente esempio.

#### ESEMPIO 3.37

##### Diametri dei lettori ottici

Nell'Esempio 3.13 si è determinato il valore 0.919 come probabilità che un diametro soddisfi le specifiche. Qual è la probabilità che tutti i 10 diametri soddisfino le specifiche, supponendo che i diametri siano indipendenti?

Indichiamo il diametro del primo alberino con  $X_1$ , quello del secondo con  $X_2$  e via dicendo, sino a quello del decimo alberino,  $X_{10}$ . La probabilità richiesta può essere scritta come

$$P(0.2485 < X_1 < 0.2515, 0.2485 < X_2 < 0.2515, \dots, 0.2485 < X_{10} < 0.2515)$$

In questo esempio l'unico insieme che ci interessa è

$$E_1 = (0.2485, 0.2515)$$

Con le notazioni usate nella definizione di indipendenza si ha

$$E_1 = E_2 = \cdots = E_{10}$$

Si ricordi l'interpretazione frequentista della probabilità. La frazione di volte che ci si attende che l'alberino 1 rispetti le specifiche è 0.919, per l'alberino 2 è ancora 0.919 ecc. Se le variabili aleatorie sono indipendenti, la frazione di volte in cui ci si aspetta che tutti i 10 alberini rispettino la specifiche è

$$\begin{aligned} & P(0.2485 < X_1 < 0.2515, 0.2485 < X_2 < 0.2515, \dots, 0.2485 < X_{10} < 0.2515) \\ & = P(0.2485 < X_1 < 0.2515) \times P(0.2485 < X_2 < 0.2515) \times \dots \times P(0.2485 < X_{10} \\ & < 0.2515) = 0.919^{10} = 0.430 \end{aligned}$$

Le variabili aleatorie indipendenti sono essenziali nelle analisi condotte nel resto di questo libro. Si assume spesso che le variabili aleatorie che registrano le repliche di un esperimento casuale siano indipendenti, come nell'esempio appena illustrato. In effetti ciò che si assume è che i termini di disturbo  $\epsilon_i$  (per  $i = 1, 2, \dots, n$  repliche) nel modello

$$X_i = \mu_i + \epsilon_i$$

siano indipendenti, perché sono i disturbi che generano la casualità e le probabilità associate alle misure.

Si noti che l'indipendenza implica che le probabilità possono essere moltiplicate per *ogni* scelta degli insiemi  $E_1, E_2, \dots, E_n$ . Pertanto, non dovrebbe costituire una sorpresa l'apprendere che una definizione equivalente di indipendenza è che la funzione di densità di probabilità congiunta delle variabili aleatorie è uguale al prodotto delle funzioni di densità di probabilità di ogni variabile aleatoria. Questa definizione è valida anche per la funzione di massa di probabilità nel caso di variabili aleatorie discrete.

### ESEMPIO 3.38 Spessore dei rivestimenti

Si supponga che  $X_1, X_2$  e  $X_3$  rappresentino rispettivamente lo spessore (in micron) di un substrato, di uno strato attivo e dello strato di rivestimento di un prodotto chimico. Si assuma che  $X_1, X_2$  e  $X_3$  siano indipendenti e normalmente distribuiti con  $\mu_1 = 10\ 000, \mu_2 = 1000, \mu_3 = 80, \sigma_1 = 250, \sigma_2 = 20, \sigma_3 = 4$ . Le specifiche per gli spessori dei tre strati sono rispettivamente  $9200 < x_1 < 10\ 800, 950 < x_2 < 1050$  e  $75 < x_3 < 85$ . Quale frazione di prodotti chimici soddisfa tutte le specifiche sugli spessori? Quale dei tre spessori ha la minore probabilità di soddisfare le specifiche?

La probabilità richiesta è  $P(9200 < X_1 < 10\ 800, 950 < X_2 < 1050, 75 < X_3 < 85)$ . Usando la notazione impiegata nella definizione di indipendenza si ha in questo caso  $E_1 = (9200, 10\ 800), E_2 = (950, 1050), E_3 = (75, 85)$ . Essendo le variabili aleatorie indipendenti, si ha

$$\begin{aligned} & P(9200 < X_1 < 10800, 950 < X_2 < 1050, 75 < X_3 < 85) \\ & = P(9200 < X_1 < 10800)P(950 < X_2 < 1050)P(75 < X_3 < 85) \end{aligned}$$

In seguito alla standardizzazione l'espressione precedente diventa

$$P(-3.2 < Z < 3.2)P(-2.5 < Z < 2.5)P(-1.25 < Z < 1.25)$$

dove  $Z$  è una variabile aleatoria normale standard. Dalla tabella della distribuzione normale standard si ha che l'espressione precedente diventa a sua volta

$$(0.99862)(0.98758)(0.78870) = 0.7778$$

Lo spessore dello strato di rivestimento ha la minore probabilità di soddisfare le specifiche. Di conseguenza, si dovrebbe prestare particolare attenzione a ridurre la variabilità in questo segmento del processo di produzione.

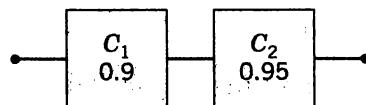
Il concetto di indipendenza può essere applicato anche agli esperimenti che classificano i risultati, come fatto per ricavare la distribuzione binomiale. Si ricordi che un esaminando che si limiti a indovinare la risposta corretta fra le quattro proposte per ogni domanda ha una probabilità pari a 1/4 di rispondere esattamente. Se si fa l'ipotesi che l'esito corretto o errato di una domanda sia indipendente dagli altri, la probabilità di avere, per esempio, 5 risposte corrette può essere determinata mediante moltiplicazione; risulta

$$(1/4)^5 = 0.00098$$

Alcune ulteriori applicazioni dell'indipendenza si incontrano frequentemente nell'area dell'analisi dei sistemi. Si consideri un sistema costituito da dispositivi che funzionano o sono guasti. Si assume che tali dispositivi siano indipendenti.

#### ESEMPIO 3.39 Sistema in serie

Il sistema schematizzato in figura è operativo solo se i dispositivi sul percorso da sinistra a destra funzionano tutti. La probabilità che ciascun componente funzioni è indicata nel diagramma. Si supponga che i componenti funzionino o siano guasti in maniera indipendente. Qual è la probabilità che il sistema sia operativo?



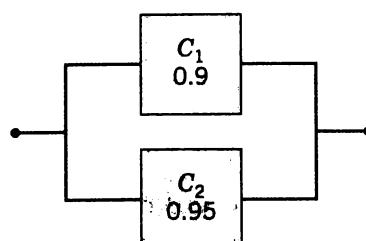
Siano  $C_1$  e  $C_2$  gli eventi corrispondenti rispettivamente al funzionamento del componente 1 e del componente 2. Perché il sistema possa essere operativo devono essere entrambi funzionanti. La probabilità che il sistema sia operativo è dunque

$$P(C_1, C_2) = P(C_1)P(C_2) = (0.9)(0.95) = 0.855$$

Si noti che tale probabilità è minore di quella associata al funzionamento dei singoli componenti. Il sistema è guasto ogni volta che *uno qualsiasi* dei componenti non funziona. Un sistema di questo tipo è detto **sistema in serie**.

#### ESEMPIO 3.40 Sistema in parallelo

Il sistema schematizzato in figura è operativo solo se vi è un percorso di dispositivi funzionanti da sinistra a destra. La probabilità che ciascun componente funzioni è indicata nel diagramma. Si supponga che i componenti funzionino o siano guasti in maniera indipendente. Qual è la probabilità che il sistema sia operativo?



Siano  $C_1$  e  $C_2$  gli eventi corrispondenti rispettivamente al funzionamento del componente 1 e del componente 2. Inoltre, indichiamo con  $C'_1$  e  $C'_2$ , rispettivamente, gli eventi che i componenti 1 e 2 non funzionino, con probabilità associate  $P(C'_1) = 1 - 0.9 = 0.1$  e  $P(C'_2) = 1 - 0.95 = 0.05$ . Perché il sistema possa essere operativo almeno uno dei componenti deve essere funzionante. La probabilità che il sistema sia operativo è 1 meno la probabilità che il sistema sia guasto, e ciò si verifica quando *entrambi* i componenti indipendenti non funzionano. Perciò la probabilità richiesta è

$$P(C_1 \text{ o } C_2) = 1 - P(C'_1, C'_2) = 1 - P(C'_1)P(C'_2) = 1 - (0.1)(0.05) = 0.995$$

Si noti che la probabilità che il sistema sia operativo è maggiore di quella associata al funzionamento dei singoli componenti. Questa disposizione costituisce un'utile strategia di progettazione per ridurre i guasti nei sistemi. Il sistema è guasto solo se *tutti* i componenti non funzionano. Un sistema di questo tipo è detto **sistema in parallelo**.

Si possono ricavare risultati più generali. La probabilità che un componente non si guasti nel tempo in cui deve operare è detta **affidabilità**. Si supponga che  $r_i$  indichi l'affidabilità del componente  $i$  in un sistema costituito da  $k$  componenti e che  $r$  indichi la probabilità che il sistema non si guasti durante il periodo programmato;  $r$  è allora l'affidabilità del sistema. Gli esempi precedenti possono venire estesi, ottenendo i seguenti risultati. Per un sistema in serie vale

$$r = r_1 r_2 \cdots r_k$$

mentre per un sistema in parallelo

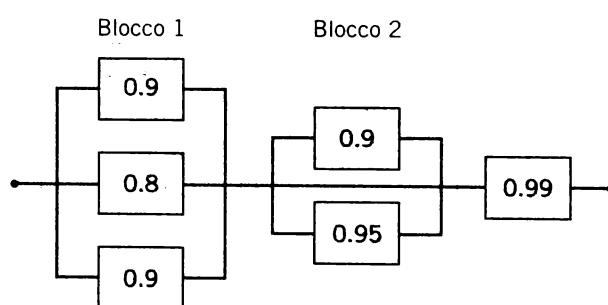
$$r = 1 - (1 - r_1)(1 - r_2) \cdots (1 - r_k)$$

Si può condurre l'analisi di un sistema complesso operando una suddivisione in sottosistemi, detti a volte *blocchi*.

#### ESEMPIO 3.41

Sistema  
complesso

Il sistema schematizzato in figura è operativo solo se vi è un percorso di dispositivi funzionanti da sinistra a destra. La probabilità che ciascun componente funzioni è indicata nel diagramma. Si supponga che i componenti funzionino o siano guasti in maniera indipendente. Qual è la probabilità che il sistema sia operativo?



Il sistema può essere suddiviso in blocchi, in modo che in ogni blocco la disposizione dei componenti sia esclusivamente quella in parallelo. A ciascun blocco si può allora applicare il

risultato valido per i sistemi in parallelo, combinando poi i risultati dei singoli blocchi mediante l'analisi valida per un sistema in serie. Per il blocco 1 si ottiene l'affidabilità

$$1 - (0.1)(0.2)(0.1) = 0.998$$

Analogamente, per il blocco 2 l'affidabilità è

$$1 - (0.1)(0.5) = 0.995$$

Come detto, si ricava l'affidabilità complessiva del sistema dal risultato valido per i sistemi in serie

$$(0.998)(0.995)(0.99) = 0.983$$

### 3.12 FUNZIONI DI VARIABILI ALEATORIE

In molti problemi pratici una variabile aleatoria è definita come funzione di una o più variabili aleatorie. Esistono opportuni metodi per determinare la distribuzione di probabilità di una funzione di una o più variabili aleatorie e per trovare importanti proprietà, come la media e la varianza. Una trattazione esaustiva di questo argomento si può trovare in Montgomery, Runger (2011). In questo paragrafo ci limitiamo a presentare alcuni dei risultati più utili.

Iniziamo con alcune semplici proprietà. Sia  $X$  una variabile aleatoria con media  $\mu$  e varianza  $\sigma^2$ , e sia  $c$  una costante. Definiamo una nuova variabile aleatoria  $Y$  come segue

$$Y = X + c$$

Dalla definizione di valore atteso e varianza (Equazione (3.3)) segue che

$$E(Y) = E(X) + c = \mu + c \quad (3.23)$$

$$V(Y) = V(X) + 0 = \sigma^2 \quad (3.24)$$

Vale a dire: aggiungendo una costante a una variabile aleatoria la media aumenta del valore della costante, mentre la varianza della variabile aleatoria rimane la stessa.

Si supponga ora di moltiplicare la variabile aleatoria  $X$  per una costante

$$Y = cX$$

In questo caso abbiamo

$$E(Y) = E(cX) = cE(X) = c\mu \quad (3.25)$$

$$V(Y) = V(cX) = c^2V(X) = c^2\sigma^2 \quad (3.26)$$

Pertanto, la media di una variabile aleatoria che è moltiplicata per una costante è uguale al prodotto della costante per la media della variabile aleatoria originaria, mentre la varianza di una variabile aleatoria che è moltiplicata per una costante è uguale il *quadrato* del prodotto della costante per la varianza della variabile aleatoria originaria.

Consideriamo ora qualche ulteriore caso che coinvolge più di una variabile aleatoria. Ci serviremo dei risultati delle Equazioni dalla (3.23) alla (3.26).

### 3.12.1 Combinazioni lineari di variabili aleatorie indipendenti

In molte situazioni si ha a che fare con una combinazione lineare di variabili aleatorie. Per esempio, si supponga che le variabili aleatorie  $X_1$  e  $X_2$  rappresentino rispettivamente la lunghezza e la larghezza di un pezzo lavorato. Per  $X_1$  si supponga di conoscere  $\mu_1 = 2$  cm e  $\sigma_1 = 0.1$  cm, mentre per  $X_2$  si sa che  $\mu_2 = 5$  cm e  $\sigma_2 = 0.2$  cm. Inoltre, si assuma che  $X_1$  e  $X_2$  siano indipendenti. Vogliamo calcolare la media e la deviazione standard del perimetro del pezzo. Faremo l'ipotesi che i lati opposti siano uguali di modo che il pezzo sia sempre rettangolare.

Il perimetro del pezzo è

$$Y = 2X_1 + 2X_2$$

Dobbiamo trovare la media e la deviazione standard di  $Y$ . Questo problema è un caso particolare della determinazione di media e varianza (o, equivalentemente, deviazione standard) di una combinazione lineare di variabili aleatorie **indipendenti**.

Siano  $c_0, c_1, c_2, \dots, c_n$  delle costanti, e  $X_1, X_2, \dots, X_n$  delle variabili aleatorie indipendenti con medie  $E(X_i) = \mu_i, i = 1, 2, \dots, n$  e varianze  $V(X_i) = \sigma_i^2, i = 1, 2, \dots, n$ .

**Media e varianza  
di una  
combinazione  
lineare  
di variabili  
aleatorie  
indipendenti**

La media e la varianza della combinazione lineare di variabili aleatorie **indipendenti**

$$Y = c_0 + c_1X_1 + c_2X_2 + \dots + c_nX_n$$

sono

$$E(Y) = c_0 + c_1\mu_1 + c_2\mu_2 + \dots + c_n\mu_n \quad (3.27)$$

e

$$V(Y) = c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \dots + c_n^2\sigma_n^2 \quad (3.28)$$

**ESEMPIO 3.42**  
Perimetro di un  
pezzo lavorato

Riprendiamo in considerazione il pezzo lavorato descritto poco sopra, dove le variabili aleatorie  $X_1$  e  $X_2$  rappresentano rispettivamente la lunghezza e la larghezza del pezzo. Per la lunghezza conosciamo  $\mu_1 = 2$  cm e  $\sigma_1 = 0.1$  cm, mentre per  $X_2$  sappiamo che  $\mu_2 = 5$  cm e  $\sigma_2 = 0.2$  cm. Il perimetro del pezzo,  $Y = 2X_1 + 2X_2$ , è semplicemente una combinazione lineare della lunghezza e della larghezza. Usando le Equazioni (3.27) e (3.28) si ricava che la media del perimetro è

$$E(Y) = 2E(X_1) + 2E(X_2) = 2(2) + 2(5) = 14 \text{ cm}$$

mentre la varianza del perimetro è

$$V(Y) = 2^2(0.1^2) + 2^2(0.2^2) = 0.2 \text{ cm}^2$$

Perciò la deviazione standard del perimetro del pezzo è

$$\sigma_Y = \sqrt{V(Y)} = \sqrt{0.2} = 0.447 \text{ cm}$$

Un caso molto importante è quello in cui tutte le variabili aleatorie  $X_1, X_2, \dots, X_n$  nella combinazione lineare sono **indipendenti e normalmente distribuite**.

### Combinazione lineare di variabili aleatorie normali indipendenti

Siano  $X_1, X_2, \dots, X_n$  variabili aleatorie indipendenti e normalmente distribuite con medie  $E(X_i) = \mu_i$ ,  $i = 1, 2, \dots, n$  e varianze  $V(X_i) = \sigma_i^2$ ,  $i = 1, 2, \dots, n$ . Allora la combinazione lineare

$$Y = c_0 + c_1X_1 + c_2X_2 + \dots + c_nX_n$$

è normalmente distribuita con media

$$E(Y) = c_0 + c_1\mu_1 + c_2\mu_2 + \dots + c_n\mu_n$$

e varianza

$$V(Y) = c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \dots + c_n^2\sigma_n^2$$

### ESEMPIO 3.43 Perimetro di un pezzo lavorato: distribuzione normale

Consideriamo ancora una volta la produzione del pezzo descritta in precedenza. Supponiamo ora che la lunghezza  $X_1$  e la larghezza  $X_2$  siano indipendenti e indipendentemente distribuite con  $\mu_1 = 2 \text{ cm}$ ,  $\sigma_1 = 0.1 \text{ cm}$ ,  $\mu_2 = 5 \text{ cm}$  e  $\sigma_2 = 0.2 \text{ cm}$ . Nel precedente esempio abbiamo trovato che la media e la varianza del perimetro del pezzo,  $Y = 2X_1 + 2X_2$  erano rispettivamente  $E(Y) = 14 \text{ cm}$  e  $V(Y) = 0.2 \text{ cm}^2$ . Calcoliamo la probabilità che il perimetro del pezzo sia maggiore di 14.5 cm.

In base al risultato precedente,  $Y$  è anch'essa una variabile aleatoria normalmente distribuita, perciò possiamo calcolare tale probabilità come segue

$$P(Y > 14.5) = P\left(\frac{Y - \mu_Y}{\sigma_Y} > \frac{14.5 - 14}{0.447}\right) = P(Z > 1.12) = 0.13$$

Vi è perciò una probabilità uguale a 0.13 che il perimetro del pezzo superi i 14.5 cm.

### 3.12.2 Combinazioni lineari di variabili aleatorie non indipendenti

Dopo aver letto il precedente paragrafo sorge naturale una domanda: cosa succede se le variabili aleatorie nella combinazione lineare non sono indipendenti? L'assunzione di indipendenza è in realtà molto importante. Vediamo perché considerando un caso molto semplice

$$Y = X_1 + X_2$$

dove le due variabili aleatorie  $X_1$  e  $X_2$  hanno medie  $\mu_1$  e  $\mu_2$  e varianze  $\sigma_1^2$  e  $\sigma_2^2$  ma non sono indipendenti. La media di  $Y$  è ancora

$$E(Y) = E(X_1 + X_2) = E(X_1) + E(X_2) = \mu_1 + \mu_2$$

ossia è semplicemente la somma delle medie delle due variabili aleatorie  $X_1$  e  $X_2$ . La varianza di  $Y$ , in base all'Equazione (3.3), risulta

$$\begin{aligned} V(Y) &= E(Y^2) - E(Y)^2 \\ &= E[(X_1 + X_2)^2] - [E(X_1 + X_2)]^2 \end{aligned}$$

Ora, si ha  $E(X_1 + X_2) = \mu_1 + \mu_2$ , perciò la precedente equazione diventa

$$\begin{aligned} V(Y) &= E(X_1^2 + X_2^2 + 2X_1X_2) - \mu_1^2 - \mu_2^2 - 2\mu_1\mu_2 \\ &= E(X_1^2) + E(X_2^2) + 2E(X_1X_2) - \mu_1^2 - \mu_2^2 - 2\mu_1\mu_2 \\ &= [E(X_1^2) - \mu_1^2] + [E(X_2^2) - \mu_2^2] + 2E(X_1X_2) - 2\mu_1\mu_2 \\ &= \sigma_1^2 + \sigma_2^2 + 2[E(X_1X_2) - \mu_1\mu_2] \end{aligned}$$

La quantità  $E(X_1X_2) - \mu_1\mu_2$  viene detta **covarianza** delle variabili aleatorie  $X_1$  e  $X_2$ . Quando queste ultime sono indipendenti la covarianza è uguale a zero, e si ritorna al risultato familiare che è un caso speciale dell'Equazione (3.28), ossia  $V(Y) = \sigma_1^2 + \sigma_2^2$ .

La covarianza è una misura della relazione lineare fra le due variabili aleatorie  $X_1$  e  $X_2$ . Quando la covarianza è diversa da zero le variabili aleatorie  $X_1$  e  $X_2$  non sono indipendenti. La covarianza è strettamente legata alla **correlazione** fra  $X_1$  e  $X_2$ , che viene definita come segue:

### Correlazione

La correlazione fra due variabili aleatorie  $X_1$  e  $X_2$  è

$$\rho_{X_1X_2} = \frac{E(X_1X_2) - \mu_1\mu_2}{\sqrt{\sigma_1^2\sigma_2^2}} = \frac{\text{Cov}(X_1, X_2)}{\sqrt{\sigma_1^2\sigma_2^2}} \quad (3.29)$$

con  $-1 \leq \rho_{X_1X_2} \leq +1$ , e  $\rho_{X_1X_2}$  è chiamato di solito **coefficiente di correlazione**.

Poiché le varianze sono sempre positive, se la covarianza fra  $X_1$  e  $X_2$  è negativa, nulla o positiva la correlazione fra  $X_1$  e  $X_2$  è anch'essa rispettivamente negativa, nulla o positiva. Tuttavia, essendo compreso tra  $-1$  e  $+1$ , il coefficiente di correlazione è più semplice da interpretare della covarianza. Inoltre, per stimare il coefficiente di correlazione a partire dai dati campionari si usa di solito il **coefficiente di correlazione campionario** introdotto nel Paragrafo 2.6 (si veda l'Equazione (2.6)). Può essere utile rileggere la discussione sul coefficiente di correlazione campionario nel citato Paragrafo 2.6.

Possiamo dare un risultato valido in generale per le combinazioni lineari di variabili aleatorie.

**Media  
e varianza  
di una  
combinazione  
lineare: caso  
generale**

Siano  $X_1, X_2, \dots, X_n$  variabili aleatorie con medie  $E(X_i) = \mu_i$ , e varianze  $V(X_i) = \sigma_i^2$ ,  $i = 1, 2, \dots, n$ , e covarianze  $\text{Cov}(X_i, X_j)$ ,  $i, j = 1, 2, \dots, n$  con  $i < j$ . Allora la media della combinazione lineare

$$Y = c_0 + c_1 X_1 + c_2 X_2 + \dots + c_n X_n$$

è

$$E(Y) = c_0 + c_1 \mu_1 + c_2 \mu_2 + \dots + c_n \mu_n \quad (3.30)$$

e la varianza è

$$V(Y) = c_1^2 \sigma_1^2 + c_2^2 \sigma_2^2 + \dots + c_n^2 \sigma_n^2 + 2 \sum_{i < j} c_i c_j \text{Cov}(X_i, X_j) \quad (3.31)$$

### 3.12.3 Funzioni non lineari di variabili aleatorie indipendenti

In molti problemi ingegneristici si ha a che fare con funzioni non lineari di variabili aleatorie. Per esempio, la potenza  $P$  dissipata dalla resistenza  $R$  in un circuito elettrico è data dalla relazione

$$P = I^2 R$$

dove  $I$  è l'intensità di corrente. Se la resistenza è una costante nota e la corrente è una variabile aleatoria, la potenza è una variabile aleatoria, funzione non lineare della corrente. Come altro esempio si consideri il periodo  $T$  di un pendolo, dato da

$$T = 2\pi \sqrt{L/g}$$

dove  $L$  è la lunghezza del pendolo e  $g$  è l'accelerazione di gravità. Se  $g$  si può considerare costante e  $L$  è una variabile aleatoria, il periodo del pendolo è una funzione non lineare di una variabile aleatoria. Infine, possiamo misurare sperimentalmente l'accelerazione di gravità lanciando in alto una palla e misurando il tempo  $T$  che le occorre per percorrere una distanza fissata  $d$ . La relazione da adottare è

$$G = 2d/T^2$$

In questo esperimento il tempo  $T$  viene misurato con un errore, sicché è una variabile aleatoria. Pertanto, l'accelerazione di gravità è una funzione non lineare della variabile aleatoria  $T$ .

In termini generali, si supponga che la variabile aleatoria  $Y$  sia funzione della variabile aleatoria  $X$  tramite una funzione generica  $h$

$$Y = h(X)$$

In questo caso può essere difficile trovare una soluzione per la media e la varianza di  $Y$ ; dipende dalla complessità della forma della funzione  $h(X)$ . Tuttavia, se si può usare un'approssimazione lineare di  $h(X)$ , allora è a portata di mano una soluzione approssimata.

**Media  
e varianza  
approssimate  
di una funzione  
non lineare**

Se  $X$  ha media  $\mu_X$  e varianza  $\sigma_X^2$ , si possono calcolare la media e la varianza approssimate di  $Y$  con le formule

$$E(Y) = \mu_Y \approx h(\mu_X) \quad (3.32)$$

$$V(Y) = \sigma_Y^2 \approx \left( \frac{dh}{dX} \right)^2 \sigma_X^2 \quad (3.33)$$

dove la derivata  $dh/dX$  è valutata in  $\mu_X$ .

Gli ingegneri chiamano di solito l'Equazione (3.33) **formula di trasmissione dell'errore** o **formula di propagazione dell'errore**.

**ESEMPIO 3.44**

Potenza di un circuito

La potenza  $P$  dissipata dalla resistenza  $R$  in un circuito elettrico è data da  $P = I^2R$ , dove  $I$ , l'intensità di corrente, è una variabile aleatoria con media  $\mu_I = 20$  A e deviazione standard  $\sigma_I = 0.1$  A. La resistenza  $R = 80 \Omega$  è una costante. Vogliamo trovare la media e la deviazione standard approssimate della potenza. In questo problema la funzione  $h$  è  $h = I^2R$ , perciò calcolando la derivata si ricava  $dh/dI = 2IR = 2I(80)$ ; applicando le Equazioni (3.33) e (3.34) troviamo che la potenza media approssimata è

$$E(P) = \mu_P \approx h(\mu_I) = \mu_I^2 R = 20^2(80) = 3200 \text{ W}$$

mentre la varianza approssimata della potenza è

$$V(P) = \sigma_P^2 \approx \sigma_I^2 = [2(20)(80)]^2 0.1^2 = 102400 \text{ W}^2$$

Perciò la deviazione standard della potenza è  $\sigma_P \approx 320$  W. Si ricordi che la derivata  $dh/dI$  è valutata in  $\mu_I = 20$  A.

Le Equazioni (3.32) e (3.33) sono state ricavate approssimando la funzione non lineare  $h$  con una funzione lineare. Si trova l'approssimazione lineare usando una serie di Taylor al primo ordine. Supponendo che  $h(X)$  sia derivabile, l'approssimazione con serie di Taylor al primo ordine per  $Y = h(X)$  intorno al punto  $\mu_X$  è

$$Y \approx h(\mu_X) + \frac{dh}{dX}(X - \mu_X) \quad (3.34)$$

Ora,  $dh/dX$  è una costante quando viene valutata in  $\mu_X$ ,  $h(\mu_X)$  è una costante ed  $E(X) = \mu_X$ , perciò quando prendiamo il valore atteso di  $Y$  il secondo termine nell'Equazione (3.35) è zero. Di conseguenza

$$E(Y) \approx h(\mu_X)$$

La varianza approssimata di  $Y$  è

$$V(Y) \approx V[h(\mu_X)] + V\left[\frac{dh}{dX}(X - \mu_X)\right] = \left(\frac{dh}{dX}\right)^2 \sigma_X^2$$

che è la formula di propagazione dell'errore dell'Equazione (3.34). Il metodo della serie di Taylor che abbiamo usato per trovare la media e la varianza approssimate di  $Y$  viene detto di solito **metodo delta**.

A volte la variabile  $Y$  è una funzione non lineare di più variabili aleatorie, ossia

$$Y = h(X_1, X_2, \dots, X_n) \quad (3.35)$$

dove si assume che  $X_1, X_2, \dots, X_n$  siano variabili aleatorie indipendenti con medie  $E(X_i) = \mu_i$ ,  $i = 1, 2, \dots, n$  e varianze  $V(X_i) = \sigma_i^2$ ,  $i = 1, 2, \dots, n$ . Il metodo delta può essere utilizzato per trovare espressioni approssimate per la media e la varianza di  $Y$ . Lo sviluppo in serie di Taylor al primo ordine dell'Equazione (3.36) è

$$\begin{aligned} Y &\approx h(\mu_1, \mu_2, \dots, \mu_n) + \frac{\partial h}{\partial X_1}(X_1 - \mu_1) + \frac{\partial h}{\partial X_2}(X_2 - \mu_2) + \dots + \frac{\partial h}{\partial X_n}(X_n - \mu_n) \\ &= h(\mu_1, \mu_2, \dots, \mu_n) + \sum_{i=1}^n \frac{\partial h}{\partial X_i}(X_i - \mu_i) \end{aligned} \quad (3.36)$$

Prendendo il valore atteso e la varianza di  $Y$  nell'Equazione (3.36) e facendo uso delle formule per la combinazione lineare (3.27) e (3.28) si ottengono i seguenti risultati

**Data la variabile**

$$Y = h(X_1, X_2, \dots, X_n)$$

dove  $X_i$ , con  $i = 1, 2, \dots, n$ , sono  $n$  variabili aleatorie indipendenti, ciascuna di media  $\mu_i$  e varianza  $\sigma_i^2$ , allora la media e la varianza approssimate di  $Y$  sono

$$E(Y) = \mu_Y \approx h(\mu_1, \mu_2, \dots, \mu_n) \quad (3.37)$$

$$V(Y) = \sigma_Y^2 \approx \sum_{i=1}^n \left( \frac{\partial h}{\partial X_i} \right)^2 \sigma_i^2 \quad (3.38)$$

dove le derivate parziali  $\partial h / \partial X_i$  sono valutate in  $\mu_1, \mu_2, \dots, \mu_n$ .

### ESEMPIO 3.45 Resistenze in parallelo

Due resistori sono collegati in parallelo. Le resistenze  $R_1$  e  $R_2$  sono variabili aleatorie con  $E(R_1) = \mu_{R_1} = 20 \Omega$ ,  $V(R_1) = \sigma_{R_1}^2 = 0.5 \Omega^2$  ed  $E(R_2) = \mu_{R_2} = 50 \Omega$ ,  $V(R_2) = \sigma_{R_2}^2 = 1 \Omega^2$ . Vogliamo trovare la media e la deviazione standard della resistenza equivalente al parallelo, che è data da

$$R = \frac{R_1 R_2}{R_1 + R_2}$$

La media approssimata di  $R$  è

$$E(R) = \mu_R \simeq \frac{20(50)}{20 + 50} = 14.29 \Omega$$

Le derivate parziali valutate in  $\mu_{R_1}$  e  $\mu_{R_2}$  sono

$$\begin{aligned}\frac{\partial R}{\partial R_1} &= \left(\frac{R_2}{R_1 + R_2}\right)^2 = \left(\frac{50}{20 + 50}\right)^2 = 0.5102 \\ \frac{\partial R}{\partial R_2} &= \left(\frac{R_1}{R_1 + R_2}\right)^2 = \left(\frac{20}{20 + 50}\right)^2 = 0.0816\end{aligned}$$

Dall'Equazione (3.39) ricaviamo che la varianza approssimata di  $R$  è

$$\begin{aligned}V(R) = \sigma_R^2 &\simeq \left(\frac{\partial R}{\partial R_1}\right)^2 \sigma_{R_1}^2 + \left(\frac{\partial R}{\partial R_2}\right)^2 \sigma_{R_2}^2 \\ &\simeq (0.5102)^2(0.5) + (0.0812)^2(1) \\ &\simeq 0.1367 \Omega^2\end{aligned}$$

La deviazione standard di  $R$  è dunque  $\sigma_R \simeq 0.3698 \Omega$ .

### 3.13 CAMPIONI CASUALI, STATISTICHE E TEOREMA LIMITE CENTRALE

In precedenza, in questo stesso capitolo, abbiamo citato il fatto che i dati sono i valori osservati di variabili aleatorie ottenute dalle repliche di un esperimento casuale. Indichiamo con  $X_1, X_2, \dots, X_n$  le variabili aleatorie che rappresentano le osservazioni ricavate dalle  $n$  repliche. Poiché le repliche sono identiche, ogni variabile aleatoria ha la medesima distribuzione. Inoltre, spesso si assume che le variabili aleatorie siano indipendenti, ossia che i risultati delle repliche non si influenzino a vicenda. Nel resto del volume faremo frequentemente uso di modelli dove le osservazioni sono variabili aleatorie indipendenti con la stessa distribuzione. In altre parole, i dati saranno costituiti da osservazioni derivanti da repliche indipendenti di un esperimento casuale. Il modello è così frequente che ne diamo una definizione.

**Campione casuale**

Delle variabili aleatorie indipendenti  $X_1, X_2, \dots, X_n$  con la stessa distribuzione sono dette un campione casuale.

L'espressione “campione casuale” ha origine storiche, per l'uso che in passato è stato fatto dei metodi statistici. Si supponga di selezionare a caso un campione di  $n$  oggetti da una popolazione numerosa. In questo contesto, “a caso” significa che ogni sottoinsieme di dimensione  $n$  ha la stessa probabilità di essere selezionato. Se il numero di oggetti nella popolazione è molto maggiore di  $n$ , si può dimostrare che le variabili aleatorie  $X_1, X_2, \dots, X_n$  che rappresentano le osservazioni tratte dal campione sono variabili aleatorie approssimati-

vamente indipendenti con la stessa distribuzione. È per questo motivo che le variabili aleatorie indipendenti aventi la medesima distribuzione sono denominate campione casuale.

### ESEMPIO 3.46 Resistenza degli O-ring

Nell’Esempio 2.1 del Capitolo 2 la resistenza media alla trazione di otto O-ring era 1055 psi. Sorgono spontanee, allora, le domande: Che cosa possiamo concludere sulla resistenza media alla trazione degli O-ring che saranno prodotti in futuro? Quanto possiamo sbagliare sostenendo che la resistenza media alla trazione di questa popolazione futura è 1055?

Vi sono due punti essenziali da considerare per poter rispondere a queste domande.

1. Innanzitutto, poiché occorre trarre una conclusione su una **popolazione futura**, quello preso in considerazione è un esempio di **studio analitico**. Certamente dobbiamo assumere che gli esemplari attualmente a disposizione siano rappresentativi degli O-ring che saranno prodotti. Ciò si collega alla questione della stabilità che abbiamo affrontato nel Capitolo 1. L’approccio adottato di solito consiste nell’assumere che questi O-ring siano un campione casuale estratto dalla popolazione futura. Indicando la media di tale popolazione futura con  $\mu$ , l’obiettivo sarà di stimare  $\mu$ .
2. In secondo luogo, anche se assumiamo che questi O-ring siano un campione casuale estratto dalla produzione futura, la media aritmetica di questi otto esemplari potrebbe non essere uguale alla media della produzione futura. Tuttavia è possibile quantificare l’errore che si commette.

Il concetto centrale è che la media aritmetica è una funzione delle singole resistenze alla trazione degli otto O-ring. In altri termini, la media aritmetica è una funzione del campione casuale. Di conseguenza, essa è una variabile aleatoria con una sua propria distribuzione. Si ricordi che la distribuzione di una singola variabile aleatoria può venire usata per determinare la probabilità che una misura cada oltre una, due o tre deviazioni standard dalla media della distribuzione. Allo stesso modo, la distribuzione di una media aritmetica fornisce la probabilità che tale media cada oltre una distanza specificata da  $\mu$ . Perciò, se stimiamo che  $\mu$  sia 1055 psi, l’errore viene determinato dalla distribuzione della media aritmetica. Affrontiamo questo argomento nella parte finale del paragrafo.

L’Esempio 3.46 mostra che si può considerare una tipica sintesi numerica dei dati, quale una media aritmetica, come funzione del campione casuale. Si usano spesso molte altre sintesi numeriche, e ciò porta a un’importante definizione.

#### Statistica

**Una statistica è una funzione delle variabili aleatorie di un campione casuale.**

Abbiamo calcolato già numerose statistiche a partire dai dati. Tutte le sintesi numeriche del Capitolo 2, quali la media campionaria la varianza campionaria  $S^2$  e la deviazione standard campionaria  $S$  sono statistiche. Se la definizione di statistica può sembrare eccessivamente complessa, è solo perché in genere non consideriamo la distribuzione di una statistica. Una volta, però, che ci chiediamo quanto possiamo sbagliare, siamo costretti a pensare alla statistica come funzione di variabili aleatorie. Di conseguenza, ogni statistica ha una distribuzione, ed è quest’ultima che determina la bontà di una stima come quella di  $\mu$ . Spesso si può determinare la distribuzione di probabilità di una statistica a partire dalla distribuzione di probabilità di un elemento del campione casuale e dalla dimensione del campione. È opportuna un’altra definizione.

### Distribuzione campionaria

La distribuzione di probabilità di una statistica viene detta la sua **distribuzione campionaria**.

Si consideri la distribuzione campionaria della media campionaria. Si supponga di prelevare un campione casuale di dimensione  $n$  da una popolazione normale con media  $\mu$  e varianza  $\sigma^2$ . Ora, tutte le variabili aleatorie di questo campione – siano esse  $X_1, X_2, \dots, X_n$  – sono variabili aleatorie normali e indipendenti con media  $\mu$  e varianza  $\sigma^2$ . In base ai risultati del Sottoparagrafo 3.12.1 sulle combinazioni lineari di variabili aleatorie normali e indipendenti, concludiamo che la media campionaria

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

ha una distribuzione normale con media

$$E(\bar{X}) = \frac{\mu + \mu + \cdots + \mu}{n} = \mu$$

e varianza

$$V(\bar{X}) = \frac{\sigma^2 + \sigma^2 + \cdots + \sigma^2}{n^2} = \frac{\sigma^2}{n}$$

La media e la varianza di  $\bar{X}$  vengono indicate rispettivamente anche con  $\mu_{\bar{X}}$  e  $\sigma_{\bar{X}}^2$ .

### ESEMPIO 3.47 Volumi di riempimento

Un macchinario automatico è preposto al riempimento delle lattine di bibite. Il volume medio di riempimento è 12.1 once fluide e la deviazione standard è 0.05 once fluide. Si assume che i volumi di riempimento delle lattine siano variabili aleatorie normali e indipendenti. Qual è la probabilità che il volume medio di 10 lattine selezionate da questo processo sia inferiore a 12 once fluide?

Indichiamo con  $X_1, X_2, \dots, X_n$  i volumi di riempimento delle 10 lattine. Il volume medio di riempimento (indicato con  $\bar{X}$ ) è una variabile aleatoria normale con

$$E(\bar{X}) = 12.1 \quad \text{e} \quad V(\bar{X}) = \frac{0.05^2}{10} = 0.00025$$

Di conseguenza,  $\sigma_{\bar{X}} = \sqrt{0.00025} = 0.0158$  e

$$\begin{aligned} P(\bar{X} < 12) &= P\left(\frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{12 - 12.1}{0.0158}\right) \\ &= P(Z < -6.32) \cong 0 \end{aligned}$$

Se stiamo prelevando un campione da una popolazione che ha una distribuzione di probabilità non nota, la distribuzione campionaria della media campionaria sarà ancora approssimativamente normale con media  $\mu$  e varianza  $\sigma^2/n$ , se la dimensione del campione,  $n$ , è elevata. Questo è quanto afferma uno dei più utili teoremi della statistica, il **teorema limite centrale**:

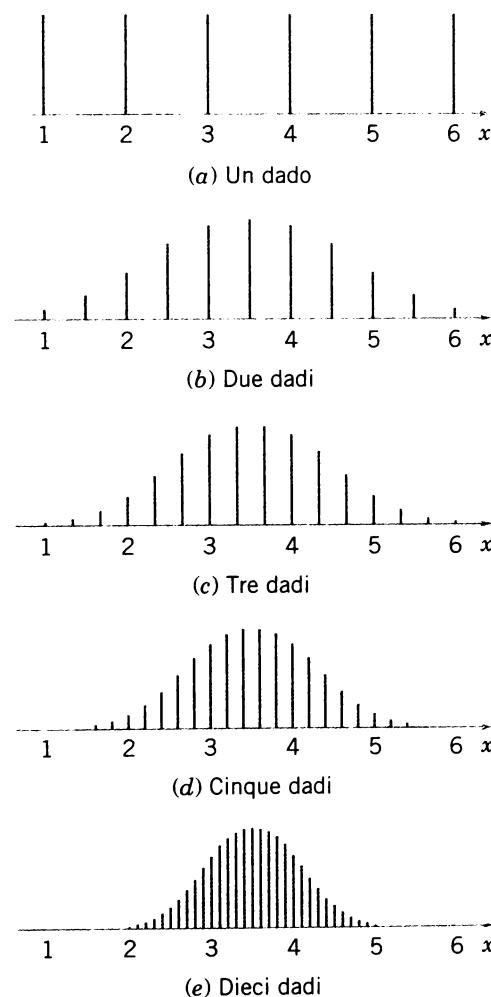
**Teorema  
limite centrale**

Se  $X_1, X_2, \dots, X_n$  è un campione casuale di dimensione  $n$  prelevato da una popolazione con media  $\mu$  e varianza  $\sigma^2$ , e se  $\bar{X}$  è la media campionaria, allora, per  $n \rightarrow \infty$ , la distribuzione limite di

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad (3.39)$$

è la distribuzione normale standard.

L'approssimazione normale per  $\bar{X}$  dipende dalla dimensione  $n$  del campione. La Figura 3.42a mostra la distribuzione che si ottiene per i lanci di un dado da gioco regolare a sei facce. Le probabilità sono uguali ( $1/6$ ) per tutti i valori ottenuti: 1, 2, 3, 4, 5 o 6. La Figura 3.42b mostra la distribuzione del punteggio medio ottenuto lanciando due dadi, mentre le Figure 3.42c, 3.42d e 3.42e mostrano le distribuzioni dei punteggi medi ottenuti dal lancio di 3, 5 e 10 dadi, rispettivamente. Si noti che mentre la distribuzione di un singolo dado è relativamente distante dalla distribuzione normale, quella delle medie è approssimata bene dalla



**Figura 3.42** Distribuzioni  
dei punteggi medi ottenuti con il lancio  
di dadi. [Adattato da Box, Hunter,  
Hunter (1978), su permesso degli autori.]

distribuzione normale per dimensioni campionarie pari almeno a 5. (Le distribuzioni del lancio dei dadi sono discrete, comunque, mentre la normale è continua.)

Anche se il teorema limite centrale funziona bene nella maggior parte dei casi per piccoli campioni ( $n = 4, 5$ ) – specialmente quelli in cui la popolazione è continua, unimodale e simmetrica – in altre situazioni sono richiesti campioni più numerosi, a seconda della forma della popolazione. In molti casi di interesse pratico, se  $n \geq 30$  l'approssimazione normale è soddisfacente, indipendentemente dalla forma della popolazione. Se  $4 \leq n$ , il teorema limite centrale funziona se la distribuzione della popolazione non è decisamente non normale.

### ESEMPIO 3.48

Un'azienda di dispositivi elettronici produce resistori che hanno resistenza media pari a  $100 \Omega$  e deviazione standard  $10 \Omega$ . Trovare la probabilità che un campione casuale di  $n = 25$  resistori abbia resistenza media minore di  $95 \Omega$ .

Osserviamo che la distribuzione campionaria di  $\bar{X}$  è approssimativamente normale, con media  $\mu_{\bar{X}} = 100 \Omega$  e deviazione standard

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$

Pertanto, la probabilità cercata corrisponde all'area ombreggiata in Figura 3.43. Standardizzando il punto  $\bar{X}$  in Figura 3.43 troviamo che

$$z = \frac{95 - 100}{2} = -2.5$$

da cui

$$P(\bar{X} < 95) = P(Z < -2.5) = 0.0062$$

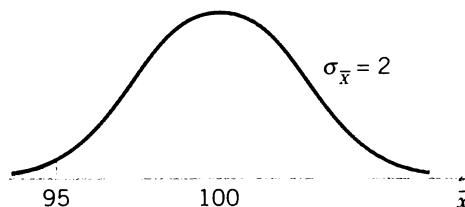


Figura 3.43 Funzione di densità di probabilità della resistenza media.

### TERMINI E CONCETTI RILEVANTI

Approssimazioni normali alle distribuzioni binomiale  
e di Poisson  
Campione casuale  
Deviazione standard di una variabile aleatoria  
Distribuzione beta  
Distribuzione binomiale  
Distribuzione campionaria  
Distribuzione di Poisson  
Distribuzione di probabilità  
Distribuzione di probabilità congiunta  
Distribuzione di Weibull  
Distribuzione esponenziale  
Distribuzione gamma  
Distribuzione lognormale  
Distribuzione normale  
Distribuzione normale standard  
Esperimento casuale  
Eventi

Fattore di continuità  
Funzione di densità di probabilità  
Funzione di distribuzione cumulativa  
Funzione di massa di probabilità  
Grafici dei quantili  
Grafici dei quantili normali  
Indipendenza  
Media di una variabile aleatoria  
Metodo delta  
Probabilità  
Processo di Poisson  
Propagazione dell'errore  
Statistica  
Teorema limite centrale  
Variabile aleatoria  
Variabile aleatoria continua  
Variabile aleatoria discreta  
Varianza di una variabile aleatoria

# Esercizi proposti

---

## ESERCIZI PER IL PARAGRAFO 3.2

---

Decidere, per ciascuna delle seguenti grandezze, se è più adatto un modello basato su variabili aleatorie discrete oppure uno basato su variabili aleatorie continue.

3.1. La vita di un dispositivo biomedico impiantato in un paziente.

3.2. Il numero di volte che un transistor installato nella memoria di un computer cambia stato in un'operazione.

3.3. La resistenza di un manufatto in cemento.

3.4. Il numero di molecole in un determinato volume di gas.

3.5. L'energia liberata in una reazione.

3.6. La concentrazione di solidi organici in un campione d'acqua.

## ESERCIZI PER IL PARAGRAFO 3.3

---

3.7. Dati due insiemi  $A$  e  $B$  tali che  $A \cap B = \emptyset$ ,  $P(X \in A) = 0.4$  e  $P(X \in B) = 0.6$ , rispondere alle seguenti domande.

(a) Gli insiemi  $A$  e  $B$  sono incompatibili?

(b) Quanto vale  $P(X \in A')$ ? (c) Quanto vale  $P(X \in B')$ ?

(d) Quanto vale  $P(X \in A \cup B)$ ?

3.8. Sia  $P(X \leq 15) = 0.3$ ,  $P(15 < X \leq 24) = 0.6$  e  $P(X > 20) = 0.5$ . Trovare:

(a)  $P(X > 15)$  (b)  $P(X \leq 24)$

(c)  $P(15 < X \leq 20)$

(d)  $P(X \leq 18)$ , sapendo che  $P(18 < X \leq 24)$

3.9. La durata di un laser a semiconduttore (espressa in ore) è rappresentata dalla variabile  $X$ , con le seguenti probabilità:

$$P(X \leq 5000) = 0.05$$

$$P(X > 7000) = 0.45$$

(a) Qual è la probabilità che la durata sia minore o uguale a 7000 ore?

(b) Qual è la probabilità che la durata sia maggiore di 5000 ore?

(c) Quanto vale  $P(5000 < X \leq 7000)$ ?

3.10. Si consideri i dati relativi ai reparti Pronto Soccorso dell'Esempio 3.1. Sia  $A$  l'evento "l'arrivo è all'ospedale 4" e  $B$  l'evento "l'arrivo si conclude con un NV" (in uno qualsiasi degli ospedali considerati). Calcolare le seguenti probabilità.

(a)  $P(A \cap B)$

(b)  $P(A')$

(c)  $P(A \cup B)$

(d)  $P(A \cup B')$

(e)  $P(A' \cup B')$

## ESERCIZI PER IL PARAGRAFO 3.4

---

3.11. Dimostrare che le seguenti funzioni sono densità di probabilità per qualche valore di  $k$  (trovare quale). Calcolare quindi la media e la varianza di  $X$ .

(a)  $f(x) = kx^2$  per  $0 < x < 4$

(b)  $f(x) = k(1 + 2x)$  per  $0 < x < 2$

(c)  $f(x) = ke^{-x}$  per  $x > 0$

(d)  $f(x) = k$  con  $k > 0$  e  $100 < x < 100 + k$

3.12. Sia data la funzione  $f(x)$  avente la seguente forma:

$$f(x) = \begin{cases} e^{-(x-6)} & x > 6 \\ 0 & x \leq 6 \end{cases}$$

Calcolare le seguenti probabilità.

(a)  $P(X > 6)$  (b)  $P(6 \leq X < 8)$

(c)  $P(X < 8)$  (d)  $P(X > 8)$

(e) Determinare  $x$  tale che  $P(X < x) = 0.95$

3.13. La funzione di densità di probabilità del tempo di rottura di un componente elettronico in una fotocopiatrice (espresso in ore) è

$$f(x) = \begin{cases} \frac{e^{-x/3000}}{3000} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Calcolare la probabilità che:

- (a) un componente duri più di 1000 ore prima della rottura;
- (b) un componente si rompa nell'intervallo temporale 1000 ÷ 2000 ore;
- (c) un componente si rompa prima di 3000 ore.
- (d) Calcolare il numero di ore entro le quali si rompe il 10% di tutti i componenti.
- (e) Calcolare la media.

3.14. Le temperature lette su una termocoppia in una fornace fluttuano secondo la funzione di distribuzione cumulativa

$$F(x) = \begin{cases} 0 & x < 800^\circ\text{C} \\ 0.1x - 80 & 800^\circ\text{C} \leq x < 810^\circ\text{C} \\ 1 & x > 810^\circ\text{C} \end{cases}$$

Calcolare quanto segue.

- (a)  $P(X < 805)$
- (b)  $P(800 < X \leq 805)$
- (c)  $P(X > 808)$
- (d) Se le specifiche del processo richiedono che la temperatura della fornace sia compresa fra 802 e 808 °C, qual è la probabilità che la fornace operi al di fuori delle specifiche?

3.15. Si supponga che il diametro (in micron) di particelle contaminanti possa venire modellizzato dalla funzione

$$f(x) = \begin{cases} 2x^{-3} & x > 1 \\ 0 & x \leq 1 \end{cases}$$

- (a) Verificare che  $f(x)$  è una densità di probabilità.
- (b) Trovare la funzione di distribuzione cumulativa.
- (c) Calcolare la media.
- (d) Qual è la probabilità che il diametro di una particella selezionata a caso sia inferiore a 5 μm?
- (e) Viene commercializzato un dispositivo ottico capace di rilevare le particelle di impurità il cui diametro supera i 7 μm. Quale frazione delle particelle sarebbe rilevata?

3.16. Si supponga che la funzione di distribuzione cumulativa della lunghezza (espressa in millimetri) dei cavi per computer sia

$$F(x) = \begin{cases} 0 & x \leq 1200 \\ 0.1x - 120 & 1200 < x \leq 1210 \\ 1 & x > 1210 \end{cases}$$

(a) Calcolare  $P(X < 1208)$ .

(b) Se le specifiche relative alla lunghezza sono 1195 mm <  $x$  < 1205 mm, qual è la probabilità che un cavo selezionato a caso le soddisfi?

3.17. Un acceleratore lineare utilizzato in medicina accelera gli elettroni per creare fasci ad alta energia in grado di distruggere le cellule tumorali con un impatto minimo sui tessuti sani circostanti. L'energia del fascio fluttua fra 200 e 210 MeV. Data la funzione di distribuzione cumulativa

$$F(x) = \begin{cases} 0 & x < 200 \\ 0.1x - 20 & 200 \leq x \leq 210 \\ 1 & x > 210 \end{cases}$$

calcolare quanto segue.

- (a)  $P(X < 209)$
- (b)  $P(200 < X < 208)$
- (c)  $P(X > 209)$
- (d) Qual è la funzione di densità di probabilità?
- (e) Riportare in grafico la funzione di densità di probabilità e quella di distribuzione cumulativa.
- (f) Calcolare la media e la varianza dell'energia del fascio di elettroni.

3.18. La funzione di densità di probabilità per il peso in libbre dei pacchi consegnati da un ufficio postale statunitense è

$$f(x) = \frac{70}{69x^2}, \text{ con } x \in (1, 70).$$

- (a) Qual è la probabilità che un pacco pesi meno di 10 libbre?
- (b) Calcolare la media e la varianza del peso dei pacchi.
- (c) Se le spese di spedizione sono 3 dollari per libbra, qual è il costo di spedizione medio di un pacco?

3.19. Il tempo di attesa prima di essere visitati presso il Pronto Soccorso di un ospedale è modellizzato dalla funzione di distribuzione di probabilità

$$f(x) = \begin{cases} \frac{1}{9}x & 0 < x < 3 \\ \frac{2}{3} - \frac{1}{9}x & 3 < x < 6 \end{cases}$$

dove  $x$  è misurata in ore. Determinare:

- (a) la probabilità che l'attesa sia minore di 4 ore;
- (b) la probabilità che l'attesa superi le 5 ore;
- (c) la probabilità che l'attesa sia minore o uguale a 30 minuti;
- (d) il tempo di attesa superato da solo il 10% dei pazienti;
- (e) il tempo di attesa medio.

## ESERCIZI PER IL PARAGRAFO 3.5

**(3.20).** Supponendo che  $Z$  abbia una distribuzione normale standard, usare la Tavola I dell'Appendice A per determinare i valori di  $z$  soluzioni delle seguenti equazioni.

- (a)  $P(Z < z) = 0.50000$
- (b)  $P(Z < z) = 0.001001$
- (c)  $P(Z > z) = 0.881000$
- (d)  $P(Z > z) = 0.866500$
- (e)  $P(-1.3 < Z < z) = 0.863140$

**(3.21).** Supponendo che  $X$  sia normalmente distribuita con media 20 e deviazione standard 2, calcolare quanto richiesto di seguito.

- (a)  $P(X < 24)$
- (b)  $P(X > 18)$
- (c)  $P(18 < X < 22)$
- (d)  $P(14 < X < 26)$
- (e)  $P(16 < X < 20)$
- (f)  $P(20 < X < 26)$

**(3.22).** La resistenza alla compressione di alcuni provini di cemento può venire modellizzata mediante una distribuzione normale con media  $6000 \text{ kg/cm}^2$  e deviazione standard  $100 \text{ kg/cm}^2$ .

- (a) Qual è la probabilità che la resistenza di un provino sia inferiore a  $6250 \text{ kg/cm}^2$ ?
- (b) Qual è la probabilità che la resistenza di un provino sia compresa fra  $5800$  e  $5900 \text{ kg/cm}^2$ ?
- (c) Quale valore di resistenza è superato dal 95% dei provini?

**(3.23)** L'ampiezza dell'incisione di uno strumento utilizzato per la produzione di semiconduttori è normalmente distribuita con media  $0.5 \mu\text{m}$  e deviazione standard  $0.05 \mu\text{m}$ .

- (a) Qual è la probabilità che un'ampiezza sia maggiore di  $0.62 \mu\text{m}$ ?
- (b) Qual è la probabilità che un'ampiezza sia compresa fra  $0.47$  e  $0.63 \mu\text{m}$ ?
- (c) L'ampiezza di incisione del 90% del campione è al di sotto di tale valore?

**3.24.** Il volume di riempimento di una macchina per il riempimento automatico utilizzata per riempire lattine di bibite è distribuito normalmente con media 12.4 once fluide e deviazione standard 0.1 once fluide.

- (a) Qual è la probabilità che un volume di riempimento sia inferiore a 12 once fluide?
- (b) Se tutte le lattine con contenuto minore di 12.1 o maggiore di 12.6 once fluide devono venire scartate, quale frazione di lattine sarà scartata?
- (c) Determinare le specifiche, simmetriche rispetto alla media, che comprendono il 99% di tutte le lattine.

**(3.25).** Riferendosi ancora all'esercizio precedente, la media dell'operazione di riempimento può venire facilmente corretta, ma la deviazione standard rimane 0.1 once.

(a) Su quale valore dovrebbe essere impostata la media affinché il 99.9% di tutte le lattine superi le 12 once?

(b) Su quale valore dovrebbe essere impostata la media affinché il 99.9% di tutte le lattine superi le 12 once se la deviazione standard può essere ridotta a 0.05 once fluide?

**3.26.** La lunghezza di un involucro di plastica stampato a iniezione, usato come supporto per nastri, è distribuito normalmente con lunghezza media 90.2 mm e deviazione standard 0.1 mm.

- (a) Qual è la probabilità che una parte sia più lunga di 90.3 mm o più corta di 89.7 mm?
- (b) Su quale valore dovrebbe venire impostata la media del processo per ottenere il numero massimo di parti comprese fra 89.7 e 90.3 mm?
- (c) Se le parti non comprese fra 89.7 e 90.3 mm vengono scartate, quale sarà la resa per la media del processo scelta al punto (b)?

**3.27.** Un dispositivo usato per registrare i livelli di inquinanti è dotato di un sensore che rileva la quantità di CO presente nell'aria. Avendolo posto in una determinata località, si trova che la quantità di CO è normalmente distribuita con media  $6.23 \text{ ppm}$  e varianza  $4.26 \text{ ppm}^2$ .

- (a) Qual è la probabilità che il livello di CO superi 9 ppm?
- (b) Qual è la probabilità che il livello di CO sia compreso fra 5.5 e 8.5 ppm?
- (c) Se il livello di CO supera una certa soglia si deve attivare un allarme. Qual è il valore soglia che si trova 3.75 deviazioni standard sopra la media?

**3.28.** Il diametro dei punti prodotti da una stampante è distribuito normalmente con media  $0.002 \text{ in}$  e deviazione standard  $0.0004 \text{ in}$ .

- (a) Qual è la probabilità che il diametro di un punto superi  $0.0026 \text{ in}$ ?
- (b) Qual è la probabilità che il diametro di un punto sia compreso fra  $0.0014$  e  $0.0026 \text{ in}$ ?
- (c) Quale deviazione standard dei diametri è necessaria per avere al punto (b) una probabilità pari a 0.995?

**3.29.** Supponendo che  $X$  abbia una distribuzione lognormale con parametri  $\theta = 2$  e  $\omega^2 = 4$ , calcolare:

- (a)  $P(X < 500)$
- (b)  $P(500 < X < 1000)$
- (c)  $P(1500 < X < 2000)$
- (d) Che cosa implica la differenza fra le probabilità trovate ai punti (a), (b) e (c), relativamente alla durata delle variabili aleatorie lognormali?

3.30. Il tempo (espresso in secondi) per il quale il visitatore della pagina di un sito Web rimane su tale pagina prima di passare a un'altra è una variabile aleatoria lognormale con parametri  $\theta = 0.5$  e  $\omega^2 = 1$ .

- (a) Qual è la probabilità che una pagina venga letta per più di 10 secondi?
- (b) Qual è il tempo di lettura del 50% dei visitatori?
- (c) Quali sono la media e la deviazione standard del tempo dopo il quale un utente si sposta su un'altra pagina?

3.31. Si supponga che la durata di un cuscinetto a sfera segua una distribuzione di Weibull con parametri  $\beta = 2$  e  $\delta = 10\,000$  ore.

- (a) Trovare la probabilità che un cuscinetto duri almeno 8000 ore.
- (b) Calcolare la durata media del cuscinetto sino al guasto.
- (c) Se vi sono 10 cuscinetti in esercizio e i guasti si verificano in maniera indipendente, qual è la probabilità che tutti i 10 cuscinetti durino almeno 8000 ore?

3.32. La durata (espressa in ore) di un processore è modellizzata da una distribuzione di Weibull con parametri  $\beta = 3$  e  $\delta = 900$  ore.

- (a) Determinare la durata media del processore.
- (b) Determinare la varianza della durata del processore.
- (c) Qual è la probabilità che il processore si guasti prima di 500 ore?

3.33. Un articolo della rivista *Journal of the Indian Geophysical Union*, intitolato “Weibull and Gamma Distributions for wave parameter Predictions” (Vol. 9, 2005, pp. 55-64) ha fatto uso della distribuzione di Weibull per modellizzare l'altezza delle onde oceaniche. Si ipotizzi che l'altezza media dell'onda presso la stazione di rilevamento sia di 2.5 m e che il parametro di forma sia uguale a 2. Calcolare la deviazione standard dell'altezza dell'onda.

3.34. Supponendo che  $X$  abbia una distribuzione gamma con parametri  $\lambda = 2.5$  e  $r = 3.2$ , calcolare la media e la varianza di  $X$ .

3.35. Supponendo che  $X$  rappresenti le misure di diametro estratte da una distribuzione gamma con media 3 mm e varianza 1.5 mm<sup>2</sup>, trovare i parametri  $\lambda$  e  $r$ .

3.36. Un valore europeo standard per i vetri delle finestre a bassa emissione indica 0.59 come la frazione di energia solare che entra nella stanza. Si supponga che la distribuzione della frazione di energia solare che entra in una stanza sia una variabile aleatoria beta.

- (a) Calcolare la moda, la media e la varianza della distribuzione per  $\alpha = 3$  e  $\beta = 1.4$ .
- (b) Calcolare la moda, la media e la varianza della distribuzione per  $\alpha = 10$  e  $\beta = 6.25$ .
- (c) Commentare la differenza di dispersione nelle distribuzioni calcolate ai punti precedenti.

3.37. Il tempo di permanenza presso un reparto di Pronto Soccorso è la somma del tempo di attesa e del tempo di servizio. Sia  $X$  la frazione di tempo trascorsa in attesa, e si assuma che abbia una distribuzione beta con  $\alpha = 10$  e  $\beta = 1$ . Calcolare:

- (a)  $P(X > 0.9)$
- (b)  $P(X > 0.5)$
- (c) Media e varianza

3.38. Il tempo massimo per portare a termine un progetto è 2.5 giorni. Supponendo che il tempo di completamento come frazione di questo tempo massimo sia una variabile aleatoria beta con  $\alpha = 2$  e  $\beta = 3$ , qual è la probabilità che siano necessari più di due giorni per concludere il progetto?

3.39. Un articolo pubblicato sulla rivista *Air Quality, Atmosphere & Health*, intitolato “Linking Particulate Matter (PM10) and Childhood Asthma in Central Phoenix”, ha fatto uso dei rilevamenti orari di PM10 (particolato di diametro inferiore a 10 micron) nell'aria effettuati dalle centraline automatiche della città di Phoenix, Arizona. La media giornaliera (sulle 24 ore) di PM10 per una centralina era di 50.9  $\mu\text{g}/\text{m}^3$  con deviazione standard pari a 25.0. Si supponga che la media giornaliera di PM10 abbia distribuzione normale.

- (a) Qual è la probabilità di rilevare una media giornaliera di PM10 maggiore di 100  $\mu\text{g}/\text{m}^3$ ?
- (b) Qual è la probabilità di rilevare una media giornaliera di PM10 minore di 25  $\mu\text{g}/\text{m}^3$ ?
- (c) Quale media giornaliera di PM10 viene superata con una probabilità del 5%?

## ESERCIZI PER IL PARAGRAFO 3.7

Verificare che le funzioni degli Esercizi 3.40-3.41 sono funzioni di massa di probabilità e determinare i valori richiesti.

3.40)	$x$	1	2	3	4
	$f(x)$	0.326	0.088	0.019	0.251
	$x$	5	6	7	
	$f(x)$	0.158	0.140	0.018	

- (a)  $P(X \leq 3)$
- (b)  $P(3 < X < 5.1)$
- (c)  $P(X > 4.5)$
- (d) Media e varianza
- (e) Disegnare il grafico di  $F(x)$

3.41)  $f(x) = (8/7)(1/2)^x, x = 1, 2, 3$

- (a)  $P(X \leq 1)$
- (b)  $P(X > 1)$
- (d) Media e varianza
- (e) Disegnare il grafico di  $F(x)$

3.42) Un'auto viene venduta con una certa gamma di optional. La funzione di massa di probabilità del numero di optional scelti è

$x$	7	8	9	10
$f(x)$	0.040	0.130	0.190	0.240
$x$	11	12	13	
$f(x)$	0.300	0.050	0.050	

- (a) Qual è la probabilità che un acquirente scelga meno di 9 optional?
- (b) Qual è la probabilità che un acquirente scelga più di 11 optional?
- (c) Qual è la probabilità che un acquirente scelga un numero di optional compreso fra 8 e 12, estremi inclusi?

- (d) Qual è il numero atteso di optional scelti? Qual è la varianza?

3.43) Sia  $X$  il numero di tacche del segnale del vostro cellulare quando vi trovate a un incrocio, caratterizzato dalle seguenti probabilità:

$x$	0	1	2	3	4	5
$P(X = x)$	0.1	0.15	0.25	0.25	0.15	0.1

Calcolare:

- (a)  $F(x)$
- (b) Media e varianza
- (c)  $P(X < 2)$
- (d)  $P(X \leq 3.5)$

3.44) Sia  $X$  il tempo di attesa (in secondi, arrotondato al primo decimale) necessario per l'aggiornamento di un grande database. La funzione di massa di probabilità per  $X$  è:

$x$	0.1	0.2	0.3	0.4	0.5	0.6
$f(x)$	0.1	0.1	0.3	0.2	0.2	0.1

Calcolare:

- (a)  $P(X < 0.25)$
- (b)  $P(0.15 < X \leq 4.5)$
- (c)  $F(x)$
- (d)  $E(x)$

## ESERCIZI PER IL PARAGRAFO 3.8

3.45) Per ciascuna delle situazioni descritte di seguito, stabilire se la distribuzione binomiale è un modello ragionevole per la variabile aleatoria, e giustificare la risposta. Enunciare esplicitamente ogni assunzione adottata.

- (a) In un processo di produzione vengono realizzati migliaia di trasduttori di temperatura. Sia  $X$  il numero di trasduttori non conformi in un campione di dimensione 30 selezionato in maniera casuale dal processo.
- (b) Da un lotto di 50 trasduttori di temperatura viene estratto senza reimmissione un campione di 30 elementi. Sia  $X$  il numero di trasduttori non conformi in tale campione.

- (c) Quattro componenti elettronici identici sono collegati a un regolatore. Sia  $X$  il numero di componenti che hanno subito un guasto dopo un determinato periodo di operatività.
- (d) Sia  $X$  il numero di plichi espresso ricevuti dall'ufficio postale in 24 ore.
- (e) Sia  $X$  il numero di risposte corrette di uno studente che svolge una prova d'esame a risposta multipla, nella quale egli è in grado di scartare alcune risposte errate in qualche domanda, e tutte le risposte in altre domande.
- (f) Quaranta chip selezionati in maniera casuale vengono sottoposti a test. Sia  $X$  il numero di chip per i quali il test rileva almeno una particella contaminante.

- (g) Sia  $X$  il numero di particelle contaminanti rilevate su un campione di 40 chip selezionati in maniera casuale.
- (h) Nell'operazione di riempimento di contenitori per detergenti si deve raggiungere un peso predeterminato. Sia  $X$  il numero di contenitori il cui peso è al di sotto della specifica.
- (i) In un canale di comunicazione digitale gli errori si verificano in sequenze che danneggiano parecchi bit consecutivi. Sia  $X$  il numero di bit errati in una trasmissione di 100 000 bit.
- (j) Sia  $X$  il numero di difetti superficiali in una grande bobina di acciaio galvanizzato.

3.46. La variabile aleatoria  $X$  ha una distribuzione binomiale con  $n = 20$  e  $p = 0.5$ . Calcolare le seguenti probabilità.

- (a)  $P(X = 15)$  (b)  $P(X \leq 12)$   
 (c)  $P(X \geq 19)$  (d)  $P(13 \leq X < 15)$   
 (e) Disegnare un grafico approssimativo della funzione di distribuzione.

3.47. La variabile aleatoria  $X$  ha una distribuzione binomiale con  $n = 10$  e  $p = 0.1$ . Calcolare le seguenti probabilità.

- (a)  $P(X = 5)$  (b)  $P(X \leq 2)$   
 (c)  $P(X \geq 9)$  (d)  $P(3 \leq X < 5)$

3.48. Una protesi per l'articolazione dell'anca viene sottoposta a un test di stress in laboratorio. La probabilità di superamento del test è uguale a 0.80. Se vengono esaminate 7 protesi scelte in maniera casuale e indipendente, qual è la probabilità che superino il test esattamente 2 delle 7?

3.49. Vengono sottoposti a controllo lotti di 50 molle a spirale selezionati da un processo di produzione per verificarne l'adeguatezza alle richieste dell'acquirente. Il numero medio di pezzi non conformi in un lotto è 5. Si assume che  $X$ , il numero di molle non conformi in un lotto, sia una variabile aleatoria binomiale.

- (a) Quanto valgono  $n$  e  $p$ ?  
 (b) Che cos'è  $P(X \leq 2)$ ?  
 (c) Che cos'è  $P(X \geq 49)$ ?

3.50. Poiché non tutti i passeggeri di una linea aerea si presentano a ritirare i biglietti prenotati, la compagnia vende 125 biglietti per un volo che può trasportarne solo 120. La probabilità che un passeggero non ritiri il biglietto prenotato è 0.10, e i passeggeri si comportano in maniera indipendente.

- (a) Qual è la probabilità che ogni passeggero che si presenta a ritirare il biglietto riesca a salire sul volo?  
 (b) Qual è la probabilità che il volo avvenga con posti vuoti?  
 (c) Quali sono la media e la deviazione standard del numero di passeggeri che ritirano la prenotazione?

3.51. La probabilità di atterrare con successo con un simulatore di volo è 0.80. A 9 allievi piloti selezionati in maniera casuale e indipendente viene chiesto di effettuare un volo di prova con il simulatore.

- (a) Qual è la probabilità che tutti gli allievi atterrino con successo?  
 (b) Qual è la probabilità che nessun allievo atterri con successo?  
 (c) Qual è la probabilità che esattamente 8 allievi su 9 atterrino con successo?

3.52. Un articolo della rivista *Information Security Technical Report*, intitolato "Malicious Software – Past, present and Future" (Vol. 9, 2004, pp. 6-18), ha fornito i seguenti dati sulla diffusione di malware nel 2002. Quello di gran lunga più diffuso è stato il worm "Klez", tuttora una delle minacce più diffuse. Si tratta di un virus rilevato per la prima volta il 26 ottobre 2001, che è stato presente per il più lungo tempo, nella storia dei virus informatici, nella classifica dei 10 malware più diffusi.

Posizione	Nome	% di rilevamenti
1	I-Worm.Klez	61.22
2	I-Worm.Lentin	20.52
3	I-Worm.Tanatos	20.9
4	I-Worm.BadtransII	1.31
5	Macro.Word97.Thus	1.19
6	I-Worm.Hybris	0.60
7	I-Worm.Bridex	0.32
8	I-Worm.Magistr	0.30
9	Win95.CIH	0.27
10	I-Worm.Sircam	0.24

(Fonte: Kaspersky Labs)

Si supponga che siano stati documentati 20 rilevamenti di malware e che le fonti dei virus siano indipendenti.

- (a) Qual è la probabilità che almeno uno dei rilevamenti sia di "Klez"?  
 (b) Qual è la probabilità che tre o più rilevamenti siano di "Klez"?  
 (c) Quali sono la media e la deviazione standard del numero di rilevamenti di "Klez" fra i 20 rilevamenti documentati?

---

## ESERCIZI PER IL PARAGRAFO 3.9

---

**3.53.** Il numero di chiamate telefoniche che arrivano a un centralino viene spesso modellizzato come variabile aleatoria di Poisson. Si assume che vi siano in media 20 chiamate all'ora.

- (a) Qual è la probabilità che vi siano esattamente 18 telefonate in un'ora?
- (b) Qual è la probabilità che vi siano al più 3 chiamate in 30 minuti?
- (c) Qual è la probabilità che vi siano esattamente 30 chiamate in 2 ore?
- (d) Qual è la probabilità che vi siano esattamente 10 chiamate in 30 minuti?

**3.54.** Il numero di malfunzionamenti di un apparecchio di misura dovuti a particelle contaminanti presenti nel prodotto è una variabile aleatoria di Poisson con una media di 0.04 malfunzionamenti all'ora.

- (a) Qual è la probabilità che lo strumento non subisca guasti in un intervallo di 8 ore?
- (b) Qual è la probabilità che si verifichino almeno 3 malfunzionamenti in 24 ore?

**3.55.** Una stazione per le telecomunicazioni è progettata per ricevere un massimo di 10 chiamate ogni 0.5 s. Se il numero di chiamate verso la stazione è modellizzato come variabile aleatoria di Poisson con una media di 9 chiamate ogni 0.5 s, qual è la probabilità che il numero di chiamate superi il vincolo massimo di progetto?

**3.56.** I messaggi inviati a un computer server seguono una distribuzione di Poisson al ritmo medio di 10/ora.

- (a) Qual è la probabilità che arrivino 3 messaggi in un'ora?
- (b) Qual è la probabilità che arrivino 6 messaggi in mezz'ora?

**3.57.** Sia  $X$  una variabile aleatoria con distribuzione esponenziale di media uguale a 5. Calcolare quanto segue.

- (a)  $P(X > 5)$
- (b)  $P(X > 15)$
- (c)  $P(X > 20)$
- (d) Trovare il valore di  $x$  tale che  $P(X < x) = 0.95$ .

**3.58.** Si supponga che i conteggi rilevati da un contatore Geiger seguano un processo di Poisson con una media di 3 conteggi/min.

- (a) Qual è la probabilità che non vi siano conteggi in un intervallo di 30 secondi?
- (b) Qual è la probabilità che il primo conteggio avvenga entro meno di 10 secondi?
- (c) Qual è la probabilità che il primo conteggio avvenga tra 1 e 2 minuti dopo l'avvio della misurazione?

**3.59.** Continuazione dell'Esercizio 3.58.

- (a) Qual è il tempo medio fra conteggi successivi?
- (b) Qual è la deviazione standard del tempo fra conteggi successivi?
- (c) Determinare  $x$  tale che la probabilità di registrare almeno un conteggio prima del tempo  $x$  sia uguale a 0.95.

**3.60.** La distanza fra le crepe rilevanti del manto stradale di un'autostrada seguono una distribuzione esponenziale con media di 5 km.

- (a) Qual è la probabilità che non vi siano crepe rilevanti in un tratto di 10 km?
- (b) Qual è la probabilità che vi siano 2 crepe rilevanti in un tratto di 10 km?
- (c) Qual è la deviazione standard delle distanza tra crepe rilevanti?

**3.61.** Continuazione dell'Esercizio 3.60.

- (a) Qual è la probabilità che la prima crepa rilevante si trovi fra 12 e 15 km dall'inizio dell'ispezione?
- (b) Qual è la probabilità che non vi siano crepe rilevanti in due tratti distinti di autostrada, lunghi 5 km?
- (c) Supposto che non vi siano crepe nei primi 5 km ispezionati, qual è la probabilità che non ve ne siano anche nei successivi 10 km ispezionati?

---

## ESERCIZI PER IL PARAGRAFO 3.10

---

**3.62.** Un grande dispositivo elettronico contiene 2000 componenti. Si assume che ciascun componente abbia una probabilità di operare senza malfunzionamenti durante la vita del dispositivo pari a 0.995, e che i componenti si guastino indipendentemente l'uno dall'altro. Approssimare la probabilità che 5 o più componenti tra i 2000 originali subisca un malfunzionamento durante la vita del dispositivo.

**3.63.** Nel 2000, negli Stati Uniti vi erano 49.7 milioni di persone con qualche tipo di disabilità permanente, un numero corrispondente al 19.3% della popolazione di età superiore ai 5 anni (<http://factfinder.census.gov>). Si immagini di selezionare un campione di 1000 persone, sotto l'ipotesi che le condizioni di disabilità di queste persone siano indipendenti.

- (a) Approssimare la probabilità che nel campione manifestino una disabilità più di 200 persone.

## ESERCIZI PER IL PARAGRAFO 3.10

(b) Approssimare la probabilità che nel campione manifestino una disabilità fra le 180 e le 300 persone.

**3.64.** A Phoenix, l'acqua viene erogata a circa 1.4 milioni di persone, che insieme possiedono più di 362 000 utenze presso la società dell'acqua potabile (<http://phoenix.gov/WATER/wtrfacts.html>). I consumi di tutte le utenze sono misurati e fatturati mensilmente. La probabilità che un'utenza sia soggetta a un errore in un mese è pari a 0.001; le utenze possono essere considerate indipendenti.

(a) Quali sono la media e la deviazione standard del numero di errori sulle utenze ogni mese?

(b) Approssimare la probabilità che si verifichino meno di 350 errori in un mese.

(c) Fornire un valore approssimato per cui la probabilità che gli errori siano maggiori di tale valore è uguale a 0.05.

**3.65.** Si supponga che il numero di particelle di amianto in un provino di 1 cm<sup>2</sup> di polveri sia una variabile aleatoria di

Poisson con media 1000. Approssimare la probabilità che 10 cm<sup>2</sup> di polveri contengano più di 10 000 particelle.

**3.66.** Il numero di e-mail spam ricevute in un giorno segue una distribuzione di Poisson con media 50. Approssimare le seguenti probabilità.

- (a) Si ricevono più di 40 e meno di 60 spam in un giorno.
- (b) Si ricevono almeno 40 spam in un giorno.
- (c) Si ricevono meno di 40 spam in un giorno.
- (d) Il numero totale di spam supera 340 in una settimana.

**3.67.** Il numero di chiamate a un servizio di assistenza sanitaria segue una distribuzione di Poisson con media 36 chiamate all'ora. Approssimare le seguenti probabilità.

- (a) Si ricevono più di 42 chiamate in un'ora.
- (b) Si ricevono meno di 30 chiamate in un'ora.
- (c) Si ricevono più di 300 chiamate in otto ore.

## ESERCIZI PER IL PARAGRAFO 3.11

**3.68.** Sia  $X$  una variabile aleatoria normale con  $\mu = 10$  e  $\sigma = 1.5$ , e sia  $Y$  una variabile aleatoria normale con  $\mu = 2$  e  $\sigma = 0.25$ . Si assuma che  $X$  e  $Y$  siano indipendenti. Calcolare le seguenti probabilità.

- (a)  $P(X < 9, Y < 2.5)$
- (b)  $P(X > 8, Y < 2.25)$
- (c)  $P(8.5 \leq X \leq 11.5, Y > 1.75)$
- (d)  $P(X < 13, 1.5 \leq Y \leq 1.8)$

**3.69.** Due vendori indipendenti forniscono cemento a un appaltatore di autostrade. Grazie all'esperienza precedente, si sa che la resistenza alla compressione di provini in cemento può essere modellizzata da una distribuzione normale con  $\mu_1 = 6000 \text{ kg/cm}^2$  e  $\sigma_1 = 100 \text{ kg/cm}^2$  per il venditore 1 e  $\mu_2 = 5825 \text{ kg/cm}^2$  e  $\sigma_2 = 90 \text{ kg/cm}^2$  per il venditore 2. Qual è la probabilità che entrambi i vendori forniscono una partita di cemento con resistenza alla compressione:

- (a) minore di 6100 kg/cm<sup>2</sup>?
- (b) compresa tra 5800 e 6050 kg/cm<sup>2</sup>?
- (c) superiore a 6200 kg/cm<sup>2</sup>?

**3.70.** Il tempo intercorrente fra il riscontro di problemi di finitura superficiale in un processo di galvanizzazione è distribuito esponenzialmente con media di 40 ore. In un singolo impianto operano tre linee di galvanizzazione che si assume lavorino indipendentemente l'una dall'altra.

(a) Qual è la probabilità che in nessuna delle tre linee si riscontrino problemi di finitura superficiale in 40 ore di operatività?

(b) Qual è la probabilità che in tutte e tre le linee si riscontrino problemi di finitura superficiale fra le 20 e le 40 ore di operatività?

**3.71.** Si consideri il sistema in serie descritto nell'Esempio 3.39. Si supponga che la probabilità che il componente  $C_1$  funzioni sia 0.95 e che quella del componente  $C_2$  sia 0.92.

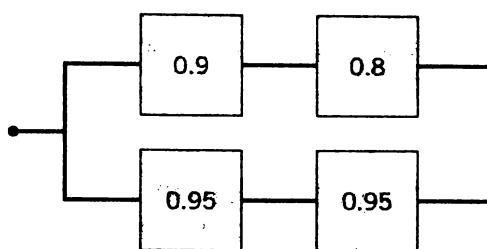
- (a) Qual è la probabilità che il sistema sia operativo?
- (b) Qual è la probabilità che il sistema non sia operativo?

**3.72.** Si consideri il sistema in parallelo descritto nell'Esempio 3.40. Si supponga che la probabilità che il componente  $C_1$  funzioni sia 0.85 e che quella del componente  $C_2$  sia 0.92.

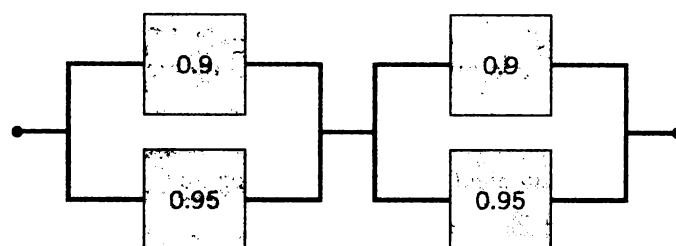
- (a) Qual è la probabilità che il componente  $C_1$  non funzioni?
- (b) Qual è la probabilità che il componente  $C_2$  non funzioni?
- (c) Qual è la probabilità che il sistema sia operativo?
- (d) Qual è la probabilità che il sistema non sia operativo?

**3.73.** Il seguente circuito opera se e solo se vi è un percorso di dispositivi funzionanti da sinistra a destra. In ciascun blocco è indicata la probabilità di funzionamento del dispositi-

tivo rappresentato. Supponendo che la probabilità che un dispositivo funzioni non dipenda dal funzionamento degli altri dispositivi, qual è la probabilità che il circuito sia operativo?



3.74. Il seguente circuito opera se e solo se vi è un percorso di dispositivi funzionanti da sinistra a destra. In ciascun blocco è indicata la probabilità di funzionamento del dispositivo rappresentato. Supponendo che la probabilità che un dispositivo funzioni non dipenda dal funzionamento degli altri dispositivi, qual è la probabilità che il circuito sia operativo?



3.75. Un negozio di alimentari registra la distribuzione congiunta del numero di mele e arance in ogni acquisto (con

qualche arrotondamento). Siano  $X$  e  $Y$  il numero di mele e arance rispettivamente, e si assuma la seguente distribuzione congiunta:

		$x$		
		0	6	12
$y$	0	0.5	0.05	0.1
	6	0.05	0.1	0.05
	12	0.1	0.05	0

Calcolare:

- (a)  $P(X = 6, Y = 6)$
- (b)  $P(X \leq 6, Y \leq 6)$
- (c)  $P(X \geq 6, Y \geq 6)$
- (d)  $P(X = 6)$
- (e)  $P(X \leq 6)$
- (f) Stabilire se  $X$  e  $Y$  sono indipendenti.

3.76. Si supponga che la distribuzione congiunta di  $X$  e  $Y$  abbia una funzione di densità di probabilità  $f(x,y) = 0.25xy$  per  $0 < x < 2$  e  $0 < y < 2$ . Calcolare:

- (a)  $P(X < 1, Y < 1)$
- (b)  $P(X < 1, Y > 1)$
- (c)  $P(X > 1, Y > 1)$
- (d)  $P(X < 1)$
- (e) Stabilire se  $X$  e  $Y$  sono indipendenti.

## ESERCIZI PER IL PARAGRAFO 3.12

3.77. Se  $X_1$  e  $X_2$  sono variabili aleatorie indipendenti con  $\mu_1 = 6$ ,  $\mu_2 = 1$ ,  $\sigma_1 = 2$ ,  $\sigma_2 = 4$ , e  $Y = 4X_1 - 2X_2$ , calcolare quanto segue.

- (a)  $E(Y)$
- (b)  $V(Y)$
- (c)  $E(2Y)$
- (d)  $V(2Y)$

3.78. L'involucro in plastica di un disco magnetico è composto da due metà; lo spessore di ciascuna di esse è distribuito normalmente con media 1.5 mm e deviazione standard 0.1 mm. Gli spessori delle due metà sono indipendenti.

- (a) Trovare la media e la deviazione standard dello spessore complessivo delle due metà.
- (b) Qual è la probabilità che lo spessore totale superi i 3.3 mm?

3.79. Si considerino le variabili aleatorie definite nell'Esercizio 3.77. Si assuma che le variabili non siano indipen-

denti e che  $\text{Cov}(X_1, X_2) = 5$ . Calcolare la media e la varianza di  $Y$ .

3.80. Si consideri l'Esempio 3.45. Se i due resistori sono collegati in serie anziché in parallelo,  $R = R_1 + R_2$ . Si assume che le due resistenze siano indipendenti. Calcolare la media e la varianza di  $R$ .

3.81. Si consideri l'Esempio 3.44. La corrente abbia intensità media 40 A e deviazione standard 0.5 A. Sapendo che il circuito elettrico ha una resistenza di 100  $\Omega$ , calcolare la media e la varianza di  $P$ .

3.82. Si consideri l'equazione per il periodo  $T$  del pendolo (Paragrafo 3.12.3). Si supponga che la lunghezza  $L$  sia una variabile aleatoria con media 30 piedi e deviazione standard 0.02. Calcolare la media e la varianza di  $T$ .

(3.83) Si consideri l'equazione per l'accelerazione di gravità  $G$  (Paragrafo 3.12.3). Si supponga che  $E(T) = 5.2$  s e  $V(T) = 0.0004$  s<sup>2</sup>. Calcolare la media e la varianza di  $G$ .

(3.84). Si considerino le variabili aleatorie dell'Esercizio 3.75. Calcolare:

- (a)  $E(2X + Y)$
- (b)  $\text{Cov}(X, Y)$
- (c)  $V(X + 3Y)$
- (d)  $\rho_{XY}$

3.85. Si considerino le variabili aleatorie dell'Esercizio

3.75. Calcolare:

- (a)  $E(2X + Y)$
- (b)  $\text{Cov}(X, Y)$
- (c)  $V(X + 3Y)$
- (d)  $\rho_{XY}$

### ESERCIZI PER IL PARAGRAFO 3.13

---

(3.86). Lo spessore della gelatina nella produzione di semi-conduttori ha media 10 μm e deviazione standard 1 μm. Si assuma che lo spessore sia distribuito normalmente e che gli spessori di wafer distinti siano indipendenti.

- (a) Calcolare la probabilità che lo spessore medio di 10 wafer sia maggiore di 11 o minore di 9 μm.
- (b) Calcolare il numero di wafer da misurare per avere una probabilità pari a 0.01 che la media dello spessore superi 11 μm.

(3.87). Una fibra sintetica utilizzata nella produzione di tappeti ha una resistenza alla trazione che è distribuita normalmente con media 75.5 psi e deviazione standard 3.5 psi. Trovare la probabilità che un campione casuale di  $n = 6$  esemplari di fibra abbiano resistenza media campionaria superiore a 75.75 psi.

(3.88). Il tempo che un cliente passa al check-in di un aeroporto è una variabile aleatoria con media 8.2 min e deviazione standard 1.5 min. Si supponga di osservare un campione casuale di  $n = 49$  clienti. Trovare la probabilità che il tempo medio di attesa in coda per tali clienti sia:

- (a) minore di 8 min;
- (b) compreso tra 8 e 9 min;
- (c) minore di 7.5 min.

(3.89). Si può misurare la viscosità di un fluido con un esperimento in cui si lascia cadere una sferetta in una provetta graduata contenente tale fluido e osservando la variabile aleatoria  $X$ , che rappresenta il tempo necessario alla sferetta per coprire una distanza prefissata. Si assuma che  $X$  sia distribuita normalmente con media 20 s e deviazione standard 0.5 s per un dato tipo di liquido.

(a) Qual è la deviazione standard del tempo medio di 40 esperimenti?

(b) Qual è la probabilità che il tempo medio di 40 esperimenti superi 20.1 s?

(c) Si supponga che l'esperimento venga ripetuto solo 20 volte. Qual è la probabilità che il valore medio di  $X$  superi 20.1 s?

(d) La probabilità calcolata al punto (b) è maggiore o minore di quella calcolata al punto (c)? Spiegare il perché di questa differenza.

(3.90). Si supponga che il tempo per preparare un letto d'ospedale sia modellizzabile mediante una variabile aleatoria con media 20 minuti e varianza 16 minuti. Approssimare le probabilità dei seguenti eventi.

- (a) Il tempo medio per preparare 100 letti è minore di 21 minuti.
- (b) Il tempo totale per preparare 100 letti è minore di 2200 minuti.

(3.91). La media e la deviazione standard del tempo di vita di una batteria in un computer portatile sono rispettivamente 3.5 e 1.0 ore.

- (a) Approssimare la probabilità che il tempo di vita medio di 25 batterie superi 3.25 ore.
- (b) Approssimare la probabilità che il tempo di vita medio di 100 batterie superi 3.25 ore.
- (c) Spiegare perché le risposte ai punti precedenti sono diverse tra loro.

## ESERCIZI DI FINE CAPITOLO

3.92. La vita media di un certo tipo di compressore è 10 anni, con deviazione standard 1 anno. Il produttore sostituisce gratuitamente tutti i compressori che si guastano entro il termine di garanzia; se è disposto a sostituire il 3% di tutti i compressori venduti, quanto dovrà durare la garanzia? Assumere una distribuzione normale.

3.93. Il numero di messaggi inviati a una BBS è una variabile aleatoria di Poisson con media di 5 messaggi all'ora.

- (a) Qual è la probabilità che vengano ricevuti 5 messaggi in un'ora?
- (b) Qual è la probabilità che vengano ricevuti 10 messaggi in 1.5 ore?
- (c) Qual è la probabilità che vengano ricevuti meno di 2 messaggi in mezz'ora?

3.94. Continuazione dell'Esercizio 3.93. Sia  $Y$  la variabile aleatoria definita come il tempo tra messaggi successivi in arrivo alla BBS.

- (a) Qual è la distribuzione di  $Y$ ? Qual è la sua media?
- (b) Qual è la probabilità che il tempo fra messaggi successivi superi i 15 min?
- (c) Qual è la probabilità che il tempo fra messaggi successivi sia minore di 5 min?
- (d) Supposto che siano passati 10 minuti senza arrivo di messaggi, qual è la probabilità che non vi siano messaggi nemmeno nei successivi 10 min?

3.95. Le durate dei sei componenti principali di una fotocopiatrice sono variabili aleatorie esponenziali indipendenti con medie rispettivamente di 8000, 10 000, 10 000, 20 000, 20 000 e 25 000 ore.

- (a) Qual è la probabilità che la durata di tutti i sei componenti superi 5000 ore?
- (b) Qual è la probabilità che nessun componente duri più di 5000 ore?
- (c) Qual è la probabilità che la durata di tutti i sei componenti sia minore di 3000 ore?

3.96. In base alle commesse contrattuali e a numerosi test sperimentali del passato, si sa che le misure di resistenza alla compressione sono distribuite normalmente con resistenza media alla compressione  $\mu = 5500$  psi e deviazione standard  $\sigma = 100$  psi. Presso il cliente destinatario viene sottoposto a test un campione casuale di elementi strutturali, al fine di misurarne la resistenza alla compressione.

- (a) Qual è la deviazione standard della distribuzione della media campionaria se  $n = 9$ ?
- (b) Qual è la deviazione standard della distribuzione della media campionaria se  $n = 20$ ?

(c) Confrontare i risultati dei punti (a) e (b) e spiegare perché sono uguali o perché sono differenti.

3.97. Un organo meccanico di un motore per auto è costituito dall'assemblaggio di quattro elementi principali. I pesi dei componenti sono indipendenti e distribuiti normalmente con le seguenti medie e deviazioni standard (in once):

Componente	Media	Deviazione standard
Parte sinistra dell'involucro	4	0.4
Parte destra dell'involucro	5.5	0.5
Assemblaggio cuscinetto	10	0.2
Assemblaggio bulloni	8	0.5

- (a) Qual è la probabilità che il peso di un assemblaggio superi 29.5 once?
- (b) Qual è la probabilità che il peso medio di 8 assemblaggi indipendenti superi 29 once?

3.98. Si dice che un processo è di **qualità sei sigma** se la sua media è distante almeno 6 deviazioni standard dalla specifica più vicina. Si assuma una misura distribuita normalmente.

- (a) Se una media di processo è centrata fra le specifiche superiore e inferiore a una distanza di 6 sigma da ciascuna di esse, qual è la probabilità che un prodotto non soddisfi le specifiche? Usando il fatto che 0.000001 è uguale a una parte per milione (ppm), esprimere la risposta in ppm.
- (b) Dato che è difficile mantenere la media di un processo centrata fra le specifiche, spesso si calcola la probabilità che un prodotto non soddisfi le specifiche dopo aver tenuto conto degli scostamenti del processo. Se la media del processo come al punto (a) slitta all'insù di 1.5 deviazioni standard, qual è la probabilità che un prodotto non soddisfi le specifiche? Esprimere la risposta in ppm.

3.99. Lo spessore delle lastre di vetro prodotte in un certo processo sono distribuite normalmente con media  $\mu = 3.00$  mm e deviazione standard  $\sigma = 0.12$  mm. Qual è il valore di  $c$  per cui c'è il 99% di probabilità che una lastra di vetro abbia uno spessore compreso nell'intervallo  $[3.00 - c, 3.00 + c]$ ?

3.100. I costruttori di un acceleratore lineare a utilizzo biomedico devono stabilire se ogni apparecchio funziona entro appropriati parametri, prima di inviarli agli ospedali. Si sa che una singola macchina ha una probabilità di guasto durante il test iniziale pari a 0.10. Vengono sottoposti a test 8 acceleratori.

- (a) Qual è la probabilità che si guastino al massimo due acceleratori?  
 (b) Qual è la probabilità che non si guasti alcun acceleratore?

**3.101.** Un'industria di cartucce di inchiostro sviluppa le cartucce per un produttore di stampanti, fornendo sia l'inchiostro che l'involucro. Quelle che seguono sono le funzioni di massa di probabilità del numero di cartucce usate nella vita della stampante:

$x$	5	6	7	8	9
$f(x)$	0.04	0.19	0.61	0.13	0.03

- (a) Qual è il numero atteso di cartucce utilizzate?  
 (b) Qual è la probabilità che vengano usate più di 6 cartucce?  
 (c) Qual è la probabilità che 9 stampanti su 10 selezionate in maniera casuale usino più di 6 cartucce?

**3.102.** Si consideri il seguente sistema, costituito da componenti in serie e in parallelo. La probabilità che ciascun componente funzioni è mostrata in Figura 3.44.

- (a) Qual è la probabilità che il sistema sia operativo?  
 (b) Qual è la probabilità che il sistema non sia operativo a causa dei componenti in serie, supposto che quelli in parallelo funzionino regolarmente?  
 (c) Qual è la probabilità che il sistema non sia operativo a causa dei componenti in parallelo, supposto che quelli in serie funzionino regolarmente?  
 (d) Calcolare la probabilità che il sistema non sia operativo usando la seguente formula:

$$\begin{aligned} & [1 - P(C_1) \cdot P(C_4)] \cdot [1 - P(C'_2)P(C'_4)] \\ & + P(C_1) \cdot P(C_4) \cdot P(C'_2) \cdot P(C'_3) \\ & + [1 - P(C_1)P(C_4)] \cdot P(C'_2) \cdot P(C'_3). \end{aligned}$$

- (e) Descrivere a parole il significato di ciascun termine della formula al punto (d).  
 (f) Usare il punto (a) per calcolare la probabilità che il sistema si guasti.  
 (g) Migliorando il componente  $C_1$ , portandone la probabilità di funzionamento a 0.95, ricalcolare i punti (a), (b), (c) e (f).  
 (h) In alternativa, non cambiare la probabilità originaria associata a  $C_1$ , ma portare la probabilità di funzione di  $C_2$  a un valore di 0.95 e ricalcolare i punti (a), (b), (c) e (f).  
 (i) In base alle risposte date ai punti (g) e (h), discutere se è meglio aumentare l'affidabilità di un componente in serie oppure quella di un componente in parallelo al fine di aumentare l'affidabilità del sistema complessivo.

**3.103.** Si considerino i seguenti dati, che rappresentano il numero di ore di operatività di una telecamera di sorveglianza sino al primo guasto:

675 3650 175 1150 290 2000 100 375.

- (a) Costruire un grafico dei quantili normali e commentare l'adeguatezza dell'adattamento.  
 (b) Trasformare i dati usando i logaritmi, ossia porre  $y^*$  (valore nuovo) =  $\text{Log } y$  (valore vecchio). Costruire un grafico dei quantili normali e commentare l'adeguatezza dell'adattamento.  
 (c) Il costruttore delle telecamere è interessato a definire un limite di garanzia tale per cui non più del 2% dei pezzi debba essere sostituito. Usare il modello del punto (b) per proporre un limite di garanzia per il tempo al guasto di una telecamera di sorveglianza scelta a caso. (*Suggerimento*: assicurarsi di esprimere il tempo al guasto in ore; nei calcoli usare la media e la deviazione standard campionarie per stimare i parametri della popolazione.)

**3.104.** Si considerino i seguenti dati, rappresentanti la durata degli involucri rigidi di dischi magnetici esposti a gas corrosivi (espressa in ore):

4, 86, 335, 746, 80, 1510, 195, 562, 137, 1574, 7600, 4394, 4, 98, 1196, 15, 934, 11

- (a) Costruire un grafico di probabilità di Weibull e stabilire l'adeguatezza dell'adattamento.  
 (b) Usando il parametro di forma stimato 0.53 e il parametro di scala stimato 604, stimare la probabilità che un disco si rompa entro 150 ore.  
 (c) Se una garanzia deve coprire non più del 10% dei pezzi lavorati, a quale limite temporale bisogna porla?

**3.105. Non unicità dei modelli probabilistici.** È possibile adattare più di modello a un determinato insieme di dati. Si considerino i tempi di durata forniti nell'Esercizio 3.101.

- (a) Trasformare i dati usando i logaritmi, ossia porre  $y^*$  (valore nuovo) =  $\text{Log } y$  (valore vecchio). Costruire un grafico dei quantili normali e commentare l'adeguatezza dell'adattamento.  
 (b) Usare la distribuzione normale stimata al punto (a) per stimare la probabilità che il disco si rompa entro 150 ore e confrontare il risultato trovato con quello calcolato al punto (b) dell'Esercizio 3.104.

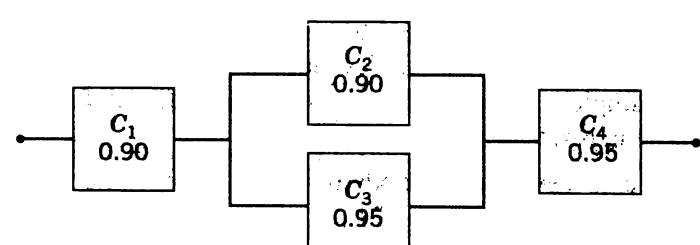


Figura 3.44 Figura per l'Esercizio 3.102.

**3.106.** Il colesterolo è una sostanza grassa che costituisce una parte importante della membrana esterna delle cellule animali. Si supponga che la media e la deviazione standard per una popolazione di individui siano rispettivamente 180 mg/dl e 20 mg/dl. Si prelevano dei campioni da 25 individui, da considerarsi indipendenti.

- (a) Quale è la probabilità che la media delle 25 misurazioni superi 185 mg/dl?
- (b) Determinare un intervallo centrato in 180 in modo che la probabilità che la media campionaria cada in tale intervallo sia uguale a 0.95.

**3.107.** Si supponga che il tempo per preparare un letto d'ospedale sia modellizzabile mediante una distribuzione esponenziale con  $\lambda = 3$  letti/ora. Determinare:

- (a) La probabilità che un letto venga preparato in meno di 10 minuti.
- (b) La probabilità che il tempo per preparare un letto sia maggiore di 30 minuti.
- (c) La probabilità che ogni paziente in un gruppo di 10 abbia il letto preparato in meno di 30 minuti. Assumere l'indipendenza dei tempi necessari a preparare i letti.
- (d) La probabilità che almeno 8 pazienti su 10 abbiano il letto preparato in meno di 30 minuti. Assumere l'indipendenza dei tempi necessari a preparare i letti.



# Processo decisionale per un singolo campione

## IL DISASTRO DELLO SPACE SHUTTLE CHALLENGER

Il 28 gennaio 1986 la mancata adozione delle norme precauzionali consigliate da un ingegnere causò il disastro del *Challenger* e la tragica morte di sette astronauti. Roger M. Boisjoly, l'ingegnere che si occupava dei giunti presenti nei razzi a propellente solido, aveva richiesto con forza il rinvio del lancio. La sua ipotesi, fondata su una grande mole di dati, era che il lancio alle basse temperature di quel giorno avrebbe potuto portare a un cedimento di alcune guarnizioni essenziali per il corretto funzionamento dei razzi. Lo staff dirigenziale della NASA, tuttavia, non prese in considerazione il suo avvertimento e, così facendo, sottopose la sua ipotesi a test direttamente nella realtà! Il *Challenger* si disintegrò 73 secondi dopo il lancio, e i resti della navetta precipitarono in mare; non vi furono sopravvissuti.

Della commissione di indagine incaricata di studiare le cause dell'incidente faceva parte, fra gli altri, Richard Feynman, uno dei fisici più di spicco dei suoi tempi, uomo dalla curiosità inesauribile. Proprio questa sua curiosità lo portò a interrogare a lungo gli ingegneri e a formulare un'ipotesi: gli O-ring, progettati per sigillare eventuali perdite dei giunti presenti sui razzi a propellente solido, non avevano funzionato a dovere. Secondo le conclusioni di Feynman, la resilienza del materiale da cui erano composti gli O-ring non sarebbe stata sufficiente a causa delle basse temperature che si erano registrate intorno alla rampa di lancio dello shuttle quel giorno. La rottura di una guarnizione avrebbe potuto lasciar fuoriuscire gas incandescenti dai razzi a propellente solido. Un attento esame dei filmati del lancio rivelò che, subito prima della disintegrazione, da uno dei due razzi laterali si era sprigionata una fiammata, che si era diretta sul serbatoio a propellente liquido.

In una conferenza stampa rimasta famosa, Feynman condusse un semplice esperimento con il materiale degli O-ring, fornendo una verifica della sua ipotesi direttamente di fronte agli organi di informazione. Egli mise un campione dell'O-ring in una morsa per simulare la pressione cui è sottoposta la guarnizione sui razzi, quindi lo immerse in un bicchiere d'acqua con ghiaccio, tenendovelo alcuni secondi. Quando lo estrasse, Feynman dimostrò che il pezzo aveva perso la sua resilienza, una proprietà essenziale per lo scopo cui deve assolvere

l’O-ring. Con frase a effetto, disse semplicemente: “Credo che questo abbia qualcosa a che fare con il nostro problema”.

La verifica di ipotesi è una fase fondamentale nella risoluzione di problemi di ingegneria e scientifici.

## CONTENUTI DEL CAPITOLO

<b>4.1 INFERENZA STATISTICA</b>	<b>4.5.2 Errore del II tipo e scelta della dimensione campionaria</b>
<b>4.2 STIMA PUNTUALE</b>	<b>4.5.3 Intervallo di confidenza per la media</b>
<b>4.3 VERIFICA DI IPOTESI</b>	<b>4.6 INFERENZA SULLA VARIANZA DI UNA POPOLAZIONE NORMALE</b>
4.3.1 Ipotesi statistiche	4.6.1 Verifica di ipotesi sulla varianza di una popolazione normale
4.3.2 Verifica delle ipotesi statistiche	4.6.2 Intervallo di confidenza per la varianza di una popolazione normale
4.3.3 Il P-value nella verifica di ipotesi	
4.3.3 Ipotesi unilaterali e bilaterali	
4.3.4 Procedura generale per la verifica di ipotesi	
<b>4.4 INFERENZA SULLA MEDIA DI UNA POPOLAZIONE CON VARIANZA NOTA</b>	<b>4.7 INFERENZA SULLA PROPORZIONE DI UNA POPOLAZIONE</b>
4.4.1 Verifica di ipotesi sulla media	4.7.1 Verifica di ipotesi su una proporzione binomiale
4.4.2 Errore del II tipo e scelta della dimensione campionaria	4.7.2 Errore del II tipo e scelta della dimensione campionaria
4.4.3 Test con campioni numerosi	4.7.3 Intervallo di confidenza per una proporzione binomiale
4.4.4 Considerazioni pratiche sulla verifica di ipotesi	
4.4.5 Intervallo di confidenza per la media	<b>4.8 ALTRE STIME INTERVALLARI PER UN SINGOLO CAMPIONE</b>
4.4.6 Metodo generale per ricavare un intervallo di confidenza	4.8.1 Intervallo di predizione
<b>4.5 INFERENZA SULLA MEDIA DI UNA POPOLAZIONE CON VARIANZA INCOGNITA</b>	4.8.2 Intervalli di tolleranza per una distribuzione normale
4.5.1 Verifica di ipotesi sulla media	<b>4.9 TABELLE RIASSUNTIVE DELLE PROCEDURE DI INFERENZA PER UN SINGOLO CAMPIONE</b>
	<b>4.10 TEST DI ADATTAMENTO</b>

---

## OBIETTIVI DI APPRENDIMENTO

Dopo uno studio attento di questo capitolo si sarà in grado di:

1. eseguire verifiche di ipotesi e costruire intervalli di confidenza sulla media di una distribuzione normale
  2. eseguire verifiche di ipotesi e costruire intervalli di confidenza sulla varianza di una distribuzione normale
  3. eseguire verifiche di ipotesi e costruire intervalli di confidenza sulla proporzione di una popolazione
  4. calcolare la potenza e l'errore del II tipo e assumere decisioni sulla scelta della dimensione campionaria per le verifiche di ipotesi e gli intervalli di confidenza
  5. spiegare e utilizzare la relazione fra intervalli di confidenza e verifiche di ipotesi
  6. costruire un intervallo di predizione per un'osservazione futura
  7. costruire un intervallo di tolleranza per una distribuzione normale
  8. spiegare la differenza fra intervalli di confidenza, di predizione e di tolleranza
  9. usare il test di adattamento chi-quadro per verificare gli assunti su una distribuzione
- 

### 4.1 INFERENZA STATISTICA

Il campo dell'inferenza statistica è costituito da quei metodi utilizzati per assumere decisioni o per trarre conclusioni su una **popolazione** e che per tale scopo si basano sull'informazione contenuta in un **campione** di tale popolazione. La relazione esistente fra una popolazione e un campione è illustrata in Figura 4.1. Con questo capitolo iniziamo dunque il nostro studio dei metodi statistici impiegati nell'inferenza e nei processi decisionali.

È possibile suddividere l'inferenza statistica in due aree principali: la **stima dei parametri** e la **verifica di ipotesi**. Come esempio di problema di stima dei parametri si supponga che un ingegnere strutturale stia analizzando la resistenza alla trazione di un componente impiegato nel telaio di un'automobile. Dato che fra i singoli componenti è naturalmente presente una variabilità della resistenza alla trazione, a causa (per esempio) delle differenze tra i lotti di materia prima, tra i processi di produzione e tra le procedure di misura, l'ingegnere è interessato a stimare la resistenza media alla trazione dei componenti. La conoscenza delle proprietà statistiche dello stimatore usato metterebbe l'ingegnere in grado di stabilire la precisione della stima.

Si consideri ora un processo chimico, in cui si possono usare due diverse temperature di reazione,  $t_1$  e  $t_2$ . L'ingegnere ritiene che  $t_1$  porti a una resa più alta di  $t_2$ . La verifica statistica delle ipotesi costituisce un quadro di riferimento per risolvere problemi di questo tipo. In questo caso l'ipotesi sarebbe che la resa media dovuta all'impiego della temperatura  $t_1$  è maggiore di quella dovuta all'impiego della temperatura  $t_2$ . Si noti che l'accento non è sulla stima delle rese, ma piuttosto sulla possibilità di trarre conclusioni su un'ipotesi enunciata.

Questo capitolo inizia con la discussione dei metodi per la stima dei parametri. Vengono quindi introdotti i principi fondamentali della verifica di ipotesi. Una volta presentate queste basi statistiche, potremo applicarle a diverse situazioni che si incontrano di frequente nella pratica ingegneristica, tra cui l'inferenza sulla media, sulla varianza e sulla proporzione di una popolazione.

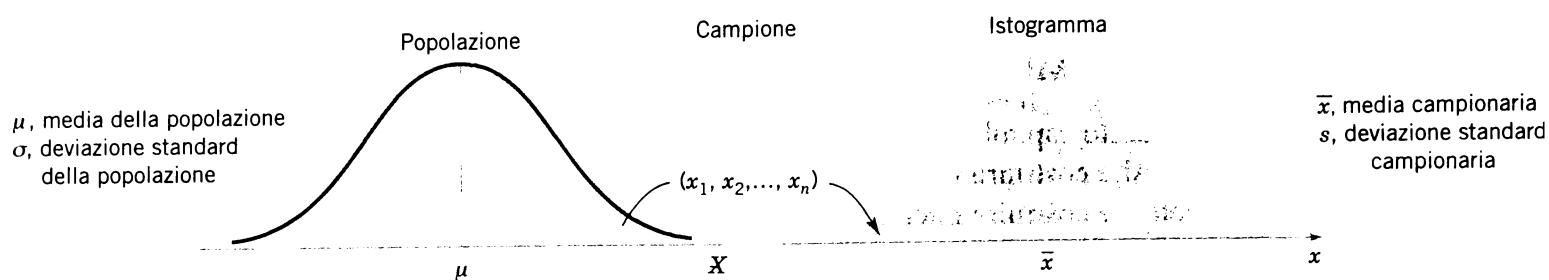


Figura 4.1 Relazione fra una popolazione e un campione.

## 4.2 STIMA PUNTUALE

Un'applicazione molto importante della statistica si ha nella procedura per ottenere le **stime puntuale** di parametri quali la media e la varianza della popolazione. Quando si trattano i problemi di inferenza conviene disporre di un simbolo generale che rappresenti i parametri di interesse. Useremo la dunque lettera greca  $\theta$  (theta) per rappresentare il generico parametro. L'obiettivo della stima puntuale è di selezionare un singolo numero, basandosi sui dati campionari, che sia il valore più plausibile per  $\theta$ . Come stima puntuale verrà usato un valore numerico di una statistica campionaria.

**Media campionaria.**

Per esempio, si supponga che la variabile aleatoria  $X$  sia distribuita normalmente con media incognita  $\mu$ . La media campionaria è uno stimatore puntuale della media della popolazione,  $\mu$ , che appunto non si conosce; in altri termini:  $\hat{\mu} = \bar{X}$ . Dopo che si è selezionato il campione, il valore numerico  $\bar{x}$  costituisce la stima puntuale di  $\mu$ . Perciò, se  $x_1 = 25$ ,  $x_2 = 30$ ,  $x_3 = 29$  e  $x_4 = 31$ , la stima puntuale di  $\mu$  è

$$\bar{x} = \frac{25 + 30 + 29 + 31}{4} = 28.75$$

Analogamente, se la varianza della popolazione,  $\sigma^2$ , è anch'essa incognita, uno stimatore puntuale per  $\sigma^2$  è la varianza campionaria  $S^2$ , e il valore numerico  $s^2 = 6.9$ , calcolato a partire dai dati campionari, viene detto stima puntuale di  $\sigma^2$ .

**Stimatore puntuale.**

In generale, se  $X$  è una variabile aleatoria con distribuzione di probabilità  $f(x)$ , caratterizzata dal parametro incognito  $\theta$ , e se  $X_1, X_2, \dots, X_n$  è un campione casuale di dimensione  $n$  estratto da  $f(x)$ , la statistica  $\hat{\Theta} = h(X_1, X_2, \dots, X_n)$ , dove  $h$  è una generica funzione delle osservazioni sul campione casuale, viene detta **stimatore puntuale** di  $\theta$ . Si noti che  $\hat{\Theta}$  è una variabile aleatoria, in quanto funzione di variabili aleatorie. Dopo che si è selezionato il campione,  $\hat{\Theta}$  assume un particolare valore numerico  $\hat{\theta}$  detto **stima puntuale** di  $\theta$ .

**Stima  
puntuale**

Una **stima puntuale** di un parametro  $\theta$  della popolazione è un singolo valore numerico  $\hat{\theta}$  di una statistica  $\hat{\Theta}$ .

In ingegneria si incontrano di frequente problemi di stima. Spesso è necessario stimare:

- la media  $\mu$  di una singola popolazione;
- la varianza  $\sigma^2$  (o la deviazione standard  $\sigma$ ) di una singola popolazione;

- la proporzione  $p$  degli elementi di una popolazione che appartengono a una classe di interesse;
- la differenza tra le medie di due popolazioni,  $\mu_1 - \mu_2$ ;
- la differenza tra le proporzioni di due popolazioni,  $p_1 - p_2$ .

Ragionevoli stime puntuali dei suddetti parametri sono:

- per  $\mu$  la stima è  $\hat{\mu} = \bar{x}$ , la media campionaria;
- per  $\sigma^2$  la stima è  $\hat{\sigma}^2 = s^2$ , la varianza campionaria;
- per  $p$  la stima è  $\hat{p} = x/n$ , la proporzione campionaria, dove  $x$  è il numero di elementi in un campione casuale di dimensione  $n$  che appartengono alla classe di interesse;
- per  $\mu_1 - \mu_2$  la stima è  $\hat{\mu}_1 - \hat{\mu}_2 = \bar{x}_1 - \bar{x}_2$ , la differenza fra le medie campionarie di due campioni casuali indipendenti;
- per  $p_1 - p_2$  la stima è  $\hat{p}_1 - \hat{p}_2$ , la differenza fra due proporzioni campionarie calcolate a partire da due campioni casuali indipendenti.

La seguente tabella riassume le relazioni fra i parametri incogniti e le tipiche statistiche e stime puntuali a essi associate.

Parametro incognito $\theta$	Statistica $\hat{\theta}$	Stima puntuale $\hat{\theta}$
$\mu$	$\bar{X} = \frac{\sum X_i}{n}$	$\bar{x}$
$\sigma^2$	$S^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1}$	$s^2$
$p$	$\hat{P} = \frac{X}{n}$	$\hat{p}$
$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2 = \frac{\sum X_{1i}}{n_1} - \frac{\sum X_{2i}}{n_2}$	$\bar{x}_1 - \bar{x}_2$
$p_1 - p_2$	$\hat{P}_1 - \hat{P}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$	$\hat{p}_1 - \hat{p}_2$

Ci possono essere molte scelte differenti per lo stimatore puntuale di un parametro. Per esempio, se vogliamo stimare la media di una popolazione possiamo considerare come estimatori puntuali la media campionaria, la mediana campionaria o magari la media aritmetica della più piccola e della più grande osservazione campionaria. Per stabilire quale stimatore puntuale di uno specifico parametro è il migliore da usare, abbiamo bisogno di esaminarne le proprietà statistiche e di sviluppare qualche criterio che ci consenta di confrontare gli estimatori.

Uno stimatore dovrebbe in qualche modo essere “vicino” al vero valore del parametro incognito. Formalmente, si dice che  $\hat{\theta}$  è uno stimatore **non distorto** di  $\theta$  se il valore atteso di  $\hat{\theta}$  è uguale a  $\theta$ . Ciò equivale a dire che la media della distribuzione di probabilità di (o la media della distribuzione campionaria di  $\hat{\theta}$ ) è uguale a  $\theta$ .

**Stimatore  
non distorto**

Lo stimatore puntuale  $\hat{\Theta}$  è uno **stimatore non distorto** (o corretto, non polarizzato, *unbiased*) per il parametro  $\theta$  se

$$E(\hat{\Theta}) = \theta \quad (4.1)$$

Se lo stimatore non è corretto, la differenza

$$E(\hat{\Theta}) - \theta \quad (4.2)$$

viene detta **distorsione** dello stimatore .

Quando uno stimatore è non distorto,  $E(\hat{\Theta}) = \theta$ , ossia la distorsione vale zero.

**ESEMPIO 4.1**
**Stimatori  
non distorti**

Si supponga che  $X$  sia una variabile aleatoria con media  $\mu$  e varianza  $\sigma^2$ . Sia  $X_1, X_2, \dots, X_n$  un campione casuale di dimensione  $n$  estratto dalla popolazione rappresentata da  $X$ . Dimostrare che la media campionaria  $\bar{X}$  e la varianza campionaria  $S^2$  sono estimatori non distorti di  $\mu$  e  $\sigma^2$ , rispettivamente.

Consideriamo innanzitutto la media campionaria. Nel Capitolo 3 abbiamo mostrato che  $E(\bar{X}) = \mu$ . Pertanto, la media campionaria  $\bar{X}$  è uno stimatore non distorto della media della popolazione,  $\mu$ .

Consideriamo ora la varianza campionaria. Abbiamo

$$\begin{aligned} E(S^2) &= E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}\right] = \frac{1}{n-1} E\sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \frac{1}{n-1} E\sum_{i=1}^n (X_i^2 + \bar{X}^2 - 2\bar{X}X_i) = \frac{1}{n-1} E\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right) \\ &= \frac{1}{n-1} \left[ \sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2) \right] \end{aligned}$$

L'ultima uguaglianza segue dall'Equazione (3.28). Tuttavia, essendo  $E(X_i^2) = \mu^2 + \sigma^2$  e  $E(\bar{X}^2) = \mu^2 + \sigma^2/n$  abbiamo

$$\begin{aligned} E(S^2) &= \frac{1}{n-1} \left[ \sum_{i=1}^n (\mu^2 + \sigma^2) - n\left(\mu^2 + \frac{\sigma^2}{n}\right) \right] \\ &= \frac{1}{n-1} (n\mu^2 + n\sigma^2 - n\mu^2 - \sigma^2) \\ &= \sigma^2 \end{aligned}$$

Perciò, la varianza campionaria  $S^2$  è uno stimatore non distorto della varianza della popolazione,  $\sigma^2$ . Tuttavia, è possibile dimostrare che la deviazione standard campionaria  $S$  è uno stimatore distorto della deviazione standard della popolazione; per campioni numerosi questa distorsione è trascurabile.

A volte vi è più di uno stimatore non distorto del parametro della popolazione. Per esempio, si supponga di prelevare un campione casuale di dimensione  $n = 10$  da una popolazione normale, ottenendo i dati:  $x_1 = 12.8, x_2 = 9.4, x_3 = 8.7, x_4 = 11.6, x_5 = 13.1, x_6 = 9.8, x_7 = 14.1, x_8 = 8.5, x_9 = 12.1, x_{10} = 10.3$ . Ora, la media campionaria è

**Media campionaria e mediana campionaria.**

$$\bar{x} = \frac{12.8 + 9.4 + 8.7 + 11.6 + 13.1 + 9.8 + 14.1 + 8.5 + 12.1 + 10.3}{10} = 11.04$$

la mediana campionaria è

$$\tilde{x} = \frac{10.3 + 11.6}{2} = 10.95$$

e una singola osservazione da questa popolazione normale, per esempio  $x_1$ , è 12.8.

È possibile dimostrare che tutti questi valori provengono da stimatori non distorti di  $\mu$ . Dato che non vi è un unico stimatore corretto, non possiamo affidarci solo alla proprietà di non distorsione per scegliere lo stimatore da usare: abbiamo bisogno di un metodo per effettuare tale scelta.

Si supponga che  $\hat{\Theta}_1$  e  $\hat{\Theta}_2$  siano stimatori non distorti di  $\theta$ . Ciò significa che la distribuzione di ogni stimatore è centrata sul vero valore di  $\theta$ . Le varianze di queste distribuzioni possono però essere tra loro differenti. La situazione è rappresentata dalla Figura 4.2. Poiché  $\hat{\Theta}_1$  ha una varianza minore di  $\hat{\Theta}_2$  è più verosimile che lo stimatore  $\hat{\Theta}_1$  produca una stima vicina al vero valore di  $\theta$ . Un principio logico della stima, allorché si deve effettuare una scelta fra parecchi stimatori, consiste nello scegliere lo stimatore a varianza minima.

**Stimatore non distorto a varianza minima**

Se si considerano tutti gli stimatori non distorti di  $\theta$ , quello con la minima varianza viene detto **stimatore non distorto a varianza minima** (MVUE, Minimum Variance Unbiased Estimator).

I concetti di stimatore non distorto e di stimatore a varianza minima sono estremamente importanti. Esistono metodi per ricavare formalmente le stime dei parametri di una distribuzione di probabilità. Uno di questi, il **metodo della massima verosimiglianza**, fornisce stimatori puntuali che sono approssimativamente non distorti e molto vicini allo stimatore a varianza minima. Per ulteriori informazioni su tale metodo si veda Montgomery, Runger (2011).

Nella pratica, a volte si deve utilizzare uno stimatore distorto (come  $S$  per  $\sigma$ ). In questi casi può essere importante l'errore quadratico medio dello stimatore. L'**errore quadratico medio** di uno stimatore  $\hat{\Theta}$  è la differenza al quadrato attesa fra  $\hat{\Theta}$  e  $\theta$ .

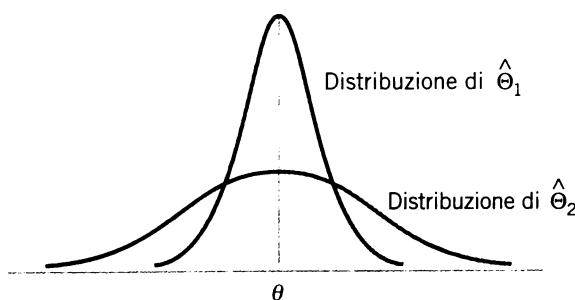


Figura 4.2 Le distribuzioni campionarie di due stimatori non distorti  $\hat{\Theta}_1$  e  $\hat{\Theta}_2$ .

**Errore quadrattico medio di uno stimatore**

L'errore quadrattico medio (MSE, *Mean Square Error*) di uno stimatore  $\hat{\Theta}$  del parametro  $\theta$  è definito da

$$\text{MSE}(\hat{\Theta}) = E[(\hat{\Theta} - \theta)^2] \quad (4.3)$$

L'espressione per l'errore quadrattico medio può venire riscritta come segue

$$\begin{aligned} \text{MSE}(\hat{\Theta}) &= E[\hat{\Theta} - E(\hat{\Theta})]^2 + [\theta - E(\hat{\Theta})]^2 \\ &= V(\hat{\Theta}) + (\text{distorsione})^2 \end{aligned}$$

Ossia: l'errore quadrattico medio di  $\hat{\Theta}$  è uguale alla varianza dello stimatore più la distorsione al quadrato. Se  $\hat{\Theta}$  è uno stimatore non distorto di  $\theta$ , l'errore quadrattico medio di  $\hat{\Theta}$  è uguale alla varianza di  $\hat{\Theta}$ .

L'errore quadrattico medio è un criterio importante per confrontare due stimatori. Siano  $\hat{\Theta}_1$  e  $\hat{\Theta}_2$  due stimatori del parametro  $\theta$ , e siano  $\text{MSE}(\hat{\Theta}_1)$  e  $\text{MSE}(\hat{\Theta}_2)$  gli errori quadrattici medi di  $\hat{\Theta}_1$  e  $\hat{\Theta}_2$ . Allora l'**efficienza relativa** di  $\hat{\Theta}_2$  rispetto a  $\hat{\Theta}_1$  è definita da

$$\frac{\text{MSE}(\hat{\Theta}_1)}{\text{MSE}(\hat{\Theta}_2)} \quad (4.4)$$

Se questa efficienza relativa è minore di 1 possiamo concludere che  $\hat{\Theta}_1$  è uno stimatore di  $\theta$  più efficiente di  $\hat{\Theta}_2$ , nel senso che ha un errore quadrattico medio minore.

In precedenza abbiamo suggerito diversi stimatori di  $\mu$ : la media campionaria, la mediana campionaria, e una singola osservazione. Poiché lavorare con la mediana campionaria è abbastanza complicato, consideriamo soltanto la media campionaria  $\hat{\Theta}_1 = \bar{X}$  e  $\hat{\Theta}_2 = X_i$ . Si noti che  $\bar{X}$  e  $X_i$  sono entrambi stimatori non distorti di  $\mu$ ; di conseguenza, l'errore quadrattico medio è per tutti e due semplicemente la varianza. Per quanto riguarda la media campionaria, abbiamo  $\text{MSE}(\bar{X}) = V(\bar{X}) = \sigma^2/n$  dall'Equazione (3.29). Perciò, l'**efficienza relativa** di  $X_i$  rispetto a  $\bar{X}$  è

$$\frac{\text{MSE}(\hat{\Theta}_1)}{\text{MSE}(\hat{\Theta}_2)} = \frac{\sigma^2/n}{\sigma^2} = \frac{1}{n}$$

Essendo  $(1/n) < 1$  per dimensioni campionarie  $n \geq 2$ , possiamo concludere che la media campionaria è uno stimatore di  $\mu$  migliore di una singola osservazione  $X_i$ . Si tratta di un risultato importante, in quanto spiega perché, in generale, per molti tipi di problemi statistici è preferibile disporre di campioni numerosi anziché di piccoli campioni.

La varianza di uno stimatore,  $V(\hat{\Theta})$  può essere vista come la varianza della distribuzione campionaria di  $\hat{\Theta}$ . La sua radice quadrata  $\sqrt{V(\hat{\Theta})}$  viene chiamata di solito **errore standard dello stimatore**.

**Errore standard**

L'**errore standard** di una statistica è la deviazione standard della sua distribuzione campionaria. Se esso coinvolge parametri incogniti i cui valori possono essere stimati, la sostituzione di tali stime nell'errore standard dà luogo a un **errore standard stimato**.

L'errore standard dà un'idea della **precisione della stima**. Per esempio, se come stimatore della media della popolazione,  $\mu$ , viene usata la media campionaria  $\bar{X}$ , l'errore standard di  $\bar{X}$  misura con quanta precisione  $\bar{X}$  stima  $\mu$ .

Si supponga di estrarre un campione da una distribuzione normale con media  $\mu$  e varianza  $\sigma^2$ . Ora, la distribuzione di  $\bar{X}$  è normale con media  $\mu$  e varianza  $\sigma^2/n$ , per cui l'errore standard di  $\bar{X}$  è

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Se non conoscessimo  $\sigma$  ma mettessimo al suo posto, nella precedente equazione, la deviazione standard campionaria  $S$ , l'errore standard stimato di  $\bar{X}$  sarebbe

$$\hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{n}}$$

#### Calcolo dell'errore standard.

Per illustrare questa definizione, un articolo pubblicato sul *Journal of Heat Transfer* (*Trans. ASME*, Ses. C, 96, 1974, p. 59) descrisse un nuovo metodo per misurare la conducibilità termica del ferro Armco. Usando una temperatura di 100 °F e una potenza in ingresso di 550 W, si ottennero le seguenti 10 misure di conducibilità termica (espresso in Btu/h-ft-°F): 41.60, 41.48, 42.34, 41.95, 41.86, 42.18, 41.72, 42.26, 41.81, 42.04. Una stima puntuale della conducibilità termica media a 100 °F e 550 W è la media campionaria, ovvero

$$\bar{x} = 41.924 \text{ Btu/h-ft-°F}$$

L'errore standard della media campionaria è  $\sigma_{\bar{X}} = \sigma/\sqrt{n}$  ed essendo  $\sigma$  incognita, possiamo inserire al suo posto la deviazione standard campionaria  $s = 0.284$  per ricavare l'errore standard stimato di  $\bar{X}$

$$\hat{\sigma}_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{0.284}{\sqrt{10}} = 0.0898$$

Si noti che l'errore standard è circa lo 0.2% della media campionaria; ciò significa che abbiamo ottenuto una stima puntuale abbastanza precisa della conducibilità termica.

## 4.3 VERIFICA DI IPOTESI

### 4.3.1 Ipotesi statistiche

Nel precedente paragrafo abbiamo mostrato come si può stimare un parametro a partire dai dati campionari. Molti problemi dell'ingegneria, però, richiedono di decidere se accettare o rifiutare un'asserzione relativa a qualche parametro. L'asserzione viene chiamata **ipotesi**, mentre il processo decisionale su di essa prende il nome di **verifica di ipotesi**. Si tratta di uno dei più utili aspetti dell'inferenza statistica, perché molte categorie di problemi decisionali, test o esperimenti del mondo ingegneristico possono venire formulati come problemi di verifica di ipotesi. Possiamo vedere le verifiche di ipotesi statistiche come lo stadio di analisi dei dati di un **esperimento comparativo** in cui l'ingegnere è interessato, per esempio, a confrontare la media di una popolazione con un determinato valore. Questi semplici esperimenti comparativi sono frequenti nella pratica, e forniscono una buona base per i più com-

plessi problemi di pianificazione degli esperimenti. In questo capitolo tratteremo gli esperimenti comparativi che interessano una sola popolazione; l'attenzione sarà rivolta in particolare alla verifica delle ipotesi relative ai parametri della popolazione medesima.

Un'ipotesi statistica può avere origine da leggi fisiche, conoscenze teoriche, esperienza passata o considerazioni esterne, quali i requisiti ingegneristici. Diamo ora una definizione formale di ipotesi statistica.

### Ipotesi statistica

**Un'ipotesi statistica** è un'asserzione relativa ai parametri di una o più popolazioni.

Poiché usiamo le distribuzioni di probabilità per rappresentare le popolazioni, un'ipotesi statistica può essere vista anche come asserzione sulla distribuzione di probabilità di una variabile aleatoria; l'ipotesi coinvolgerà in genere uno o più parametri di tale distribuzione. A titolo di esempio, si supponga di essere interessati alla velocità di combustione di un propellente solido usato per fornire energia ai seggiolini a espulsione per i piloti; tale grandezza è una variabile aleatoria che può essere descritta da una distribuzione di probabilità. Poniamo che il nostro interesse sia centrato sulla velocità media di combustione (un parametro di questa distribuzione). In particolare, vogliamo poter stabilire se la velocità media di combustione è uguale a 50 cm/s oppure no. Possiamo esprimere questo fatto a livello formale come segue

$$\begin{aligned} H_0: \mu &= 50 \text{ cm/s} \\ H_1: \mu &\neq 50 \text{ cm/s} \end{aligned} \quad (4.5)$$

L'asserzione  $H_0: \mu = 50 \text{ cm/s}$  dell'Equazione (4.5) viene chiamata **ipotesi nulla**, mentre l'asserzione  $H_1: \mu \neq 50 \text{ cm/s}$  è detta **ipotesi alternativa**. Poiché quest'ultima indica valori di  $\mu$  che potrebbero essere maggiori o minori di 50 cm/s, la si chiama **ipotesi alternativa bilaterale**. In alcune situazioni si può voler formulare invece un'**ipotesi alternativa unilaterale\***, come:

$$H_0: \mu = 50 \text{ cm/s} \quad H_1: \mu < 50 \text{ cm/s} \quad \text{ovvero} \quad H_0: \mu = 50 \text{ cm/s} \quad H_1: \mu > 50 \text{ cm/s} \quad (4.6)$$

È importante tenere a mente che le ipotesi sono sempre asserzioni sulla popolazione o sulla distribuzione oggetto di studio, non sul campione. Il valore del parametro della popolazione specificato nell'ipotesi nulla (50 cm/s, nell'esempio precedente) viene di solito determinato in tre modi. In primo luogo, esso può risultare dall'esperienza passata o dalla conoscenza del processo o anche da precedenti test o esperimenti. L'obiettivo della verifica di ipotesi, allora, è in genere di stabilire se il valore del parametro è cambiato. In secondo luogo, il suddetto valore può essere determinato da qualche teoria o modello riguardante il processo sotto studio. In questo caso l'obiettivo della verifica di ipotesi è di verificare la teoria o il modello. Una terza situazione si ha quando il valore del parametro della popolazione deriva da consi-

---

\* Vi sono due modelli utilizzabili per l'ipotesi alternativa unilaterale. Se  $H_1: \mu > 50 \text{ cm/s}$  (per esempio), allora possiamo scrivere l'ipotesi nulla come  $H_0: \mu = 50$  o come  $H_0: \mu \leq 50$ . Nel primo caso stiamo limitando  $\mu$  a essere uguale a 50 (il valore nullo), mentre nel secondo consentiamo al valore nullo di essere minore di 50. Tuttavia, entrambe le espressioni di  $H_0$  danno luogo alle medesime procedure di verifica e di decisione, cioè entrambe conducono a una procedura basata sull'uguaglianza  $\mu = 50$ . Man mano che il lettore si familiarizzerà con le procedure della verifica di ipotesi, diventerà chiaro che una decisione che porta al rifiuto dell'ipotesi nulla quando  $H_0: \mu = 50$ , porterà necessariamente al rifiuto dell'ipotesi nulla anche quando  $H_0: \mu < 50$ . Di conseguenza, noi scriveremo di solito l'ipotesi nulla con il segno di uguale, ma si devono considerare appropriati anche " $\leq$ " o " $\geq$ ", a seconda dei casi.

derazioni esterne, come le specifiche ingegneristiche o di progetto, o da obblighi contrattuali. In questo caso la verifica di ipotesi è di solito una verifica di conformità.

Una procedura che porta a una decisione inerente una particolare ipotesi viene chiamata **test di ipotesi**. Questo tipo di procedure si basa sull'uso delle informazioni contenute in un campione casuale estratto dalla popolazione di interesse. Se tali informazioni sono coerenti con l'ipotesi, concluderemo che quest'ultima è vera; in caso contrario concluderemo che è falsa. Sottolineiamo che la verità o falsità di una particolare ipotesi non può mai essere nota con certezza a meno che sia possibile esaminare l'intera popolazione (cosa non praticabile nella maggior parte dei casi). Perciò, si deve sviluppare una procedura di verifica di ipotesi tenendo presente la probabilità di arrivare a una conclusione errata.

La struttura dei problemi di verifica di ipotesi è identica in tutte le applicazioni che considereremo. L'ipotesi nulla è quella da sottoporre a verifica, e il suo rifiuto porta sempre ad accettare l'ipotesi alternativa. Nella nostra trattazione l'ipotesi nulla verrà sempre enunciata in modo che specifichi un valore esatto del parametro (come nell'asserzione  $H_0: \mu = 50$  cm/s dell'Equazione (4.5)). L'ipotesi alternativa consentirà al parametro di assumere diversi valori (come nell'asserzione  $H_1: \mu \neq 50$  cm/s dell'Equazione (4.5)). La verifica di ipotesi comporta la selezione di un campione casuale, il calcolo di una **statistica test** a partire dai dati campionari e l'utilizzo di tale statistica per assumere una decisione sull'ipotesi nulla.

### 4.3.2 Verifica delle ipotesi statistiche

Per illustrare i concetti generali, consideriamo il problema della velocità di combustione del propellente introdotto in precedenza. L'ipotesi nulla è che la velocità di combustione media sia 50 cm/s, e l'alternativa è che non sia uguale a 50 cm/s. Vogliamo dunque verificare

$$H_0: \mu = 50 \text{ cm/s}$$

$$H_1: \mu \neq 50 \text{ cm/s}$$

Supponiamo che venga sottoposto a test un campione di  $n = 10$  provini e che sia osservata una media campionaria per la velocità di combustione  $\bar{x}$ . La media campionaria è una stima della vera media della popolazione,  $\mu$ . Un valore della media campionaria  $\bar{x}$  che cade vicino al valore di  $\mu$  ipotizzato costituisce una prova che la vera media  $\mu$  è effettivamente 50 cm/s; in altre parole, tale prova va a sostegno dell'ipotesi nulla  $H_0$ . D'altro canto, una media campionaria che sia notevolmente differente dal valore 50 cm/s è una prova a sostegno dell'ipotesi alternativa  $H_1$ . Pertanto, in questo caso la statistica test è la media campionaria.

La media campionaria può assumere molti valori diversi. Supponiamo che se  $48.5 \leq \bar{x} \leq 51.5$ , non rifiuteremo l'ipotesi nulla  $H_0: \mu = 50$ , e che se  $\bar{x} < 48.5$  oppure  $\bar{x} > 51.5$  rifiuteremo l'ipotesi nulla a favore di quella alternativa,  $H_1: \mu \neq 50$ . La situazione è illustrata in Figura 4.3. I valori di  $\bar{x}$  minori di 48.5 e maggiori di 51.5 costituiscono la **regione critica** per la verifica, mentre tutti i valori compresi nell'intervallo [48.5, 51.5] formano una regione per cui non verrà rifiutata l'ipotesi nulla. Gli estremi della regione critica sono detti **valori critici**. Nel nostro esempio i valori critici sono dunque 48.5 e 51.5. È uso comune enunciare le conclusioni rispetto all'ipotesi nulla  $H_0$ ; pertanto, rifiuteremo  $H_0$  a favore di  $H_1$  se la statistica test cade nella regione critica, la accetteremo in caso contrario.

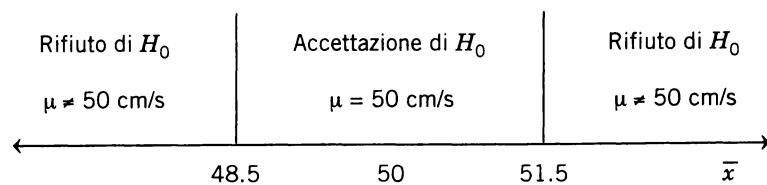


Figura 4.3 Criteri decisionali per la verifica  $H_0: \mu = 50$  cm/s contro  $H_1: \mu \neq 50$  cm/s.

Questa procedura decisionale può portare a due tipi di conclusioni errate. Per esempio, la vera velocità media di combustione del propellente potrebbe essere uguale a 50 cm/s. Tuttavia, per i provini di propellente selezionati a caso che vengono sottoposti a test, potremmo osservare un valore della statistica test  $\bar{x}$  che cade nella regione critica. Rifiuteremmo allora l'ipotesi nulla  $H_0$  a favore di  $H_1$  quando in effetti è vera  $H_0$ . Questo tipo di conclusione errata viene detto **errore del I tipo**.

**Errore  
del I tipo**

Il rifiuto dell'ipotesi nulla  $H_0$  quando essa è vera viene definito errore del I tipo.

Si supponga ora che la vera velocità media di combustione sia diversa da 50 cm/s, e tuttavia la media campionaria  $\bar{x}$  non cada nella regione critica. In questo caso commetteremmo l'errore di non rifiutare  $H_0$  quando questa è falsa. Questo tipo di conclusione errata viene detto **errore del II tipo**.

**Errore  
del II tipo**

Il mancato rifiuto dell'ipotesi nulla  $H_0$  quando essa è falsa viene definito errore del II tipo.

Perciò, nella verifica di ogni ipotesi statistica, quattro differenti situazioni determinano se la decisione finale è corretta oppure errata (si veda la Tabella 4.1).

Tabella 4.1 Decisioni nella verifica di ipotesi

Decisione	$H_0$ è vera	$H_0$ è falsa
Mancato rifiuto di $H_0$	Nessun errore	Errore del II tipo
Rifiuto di $H_0$	Errore del I tipo	Nessun errore

Basandosi la nostra decisione su variabili aleatorie, agli errori del I e del II tipo di Tabella 4.1 si possono associare delle probabilità. La probabilità di commettere un errore del I tipo è indicata con la lettera greca  $\alpha$  (alfa)

$$\alpha = P(\text{errore I tipo}) = P(\text{rifiuto di } H_0 \text{ quando } H_0 \text{ è vera}) \quad (4.7)$$

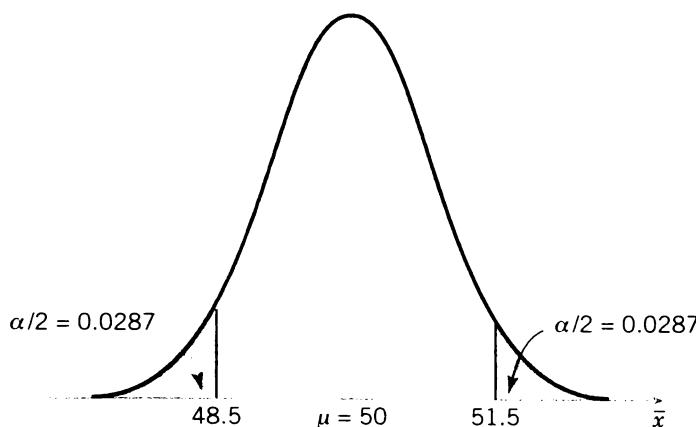


Figura 4.4 La regione critica per  $H_0: \mu = 50$  contro  $H_1: \mu \neq 50$  e  $n = 10$ .

A volte, la probabilità di un errore del I tipo viene detta **livello o ampiezza di significatività del test**. Nell'esempio della velocità di combustione del propellente, si ha un errore del I tipo quando  $\bar{x} > 51.5$  o  $\bar{x} < 48.5$  per una vera velocità media di combustione  $\mu = 50$  cm/s. Si supponga che la deviazione standard della velocità di combustione sia  $\sigma = 2.5$  cm/s e che la velocità di combustione abbia una distribuzione alla quale si applicano le condizioni del teorema limite centrale: la distribuzione della media campionaria è dunque approssimativamente normale con media  $\mu = 50$  e deviazione standard  $\sigma/\sqrt{n} = 2.5/\sqrt{10} = 0.79$ . La probabilità di commettere un errore del I tipo (ossia il livello di significatività del nostro test) è uguale alla somma delle aree ombreggiate nelle code della distribuzione normale in Figura 4.4. Possiamo trovare tale probabilità come segue

$$\alpha = P(\bar{X} < 48.5 \text{ dove } \mu = 50) + P(\bar{X} > 51.5 \text{ dove } \mu = 50)$$

**Calcolo del livello di significatività  $\alpha$ .**

I valori  $z$  che corrispondono ai valori critici 48.5 e 51.5 sono

$$z_1 = \frac{48.5 - 50}{0.79} = -1.90 \quad \text{e} \quad z_2 = \frac{51.5 - 50}{0.79} = 1.90$$

Pertanto

$$\begin{aligned} \alpha &= P(Z < -1.90) + P(Z > 1.90) \\ &= 0.0287 + 0.0287 = 0.0574 \end{aligned}$$

Ciò comporta che il 5.74% di tutti i campioni casuali porterebbe al rifiuto dell'ipotesi  $H_0: \mu = 50$  cm/s quando la vera velocità media di combustione è effettivamente 50 cm/s.

Da un esame della Figura 4.4 si nota che è possibile ridurre  $\alpha$  spingendo ancor più le regioni critiche verso le code della distribuzione. Per esempio, se poniamo come valori critici 48 e 52, il valore di  $\alpha$  risulta

$$\begin{aligned} \alpha &= P\left(Z < \frac{48 - 50}{0.79}\right) + P\left(Z > \frac{52 - 50}{0.79}\right) = P(Z < -2.53) + P(Z > 2.53) \\ &= 0.0057 + 0.0057 = 0.0114 \end{aligned}$$

**Misurazione dell'effetto della dimensione campionaria.**

Potremmo ridurre  $\alpha$  anche aumentando la dimensione campionaria. Se  $n = 16$ , si ha  $\sigma/\sqrt{n} = 2.5/\sqrt{16} = 0.625$  e usando la regione critica di Figura 4.3, troviamo

$$z_1 = \frac{48.5 - 50}{0.625} = -2.40 \quad \text{e} \quad z_2 = \frac{51.5 - 50}{0.625} = 2.40$$

Pertanto

$$\alpha = P(Z < -2.40) + P(Z > 2.40) = 0.0082 + 0.0082 = 0.0164$$

Nel valutare una procedura di verifica di ipotesi, è importante anche esaminare la probabilità di un errore del II tipo, che indicheremo con la lettera  $\beta$  (beta). Si ha dunque

$$\beta = P(\text{errore II tipo}) = P(\text{mancato rifiuto di } H_0 \text{ quando } H_0 \text{ è falsa}) \quad (4.8)$$

Per calcolare  $\beta$  dobbiamo avere una specifica ipotesi alternativa, ossia uno specifico valore di  $\mu$ . Per esempio, si supponga che sia importante rifiutare l'ipotesi nulla  $H_0: \mu = 50$  ogni volta che la velocità media di combustione è maggiore di 52 cm/s o minore di 48 cm/s. Potremmo calcolare la probabilità di un errore del II tipo  $\beta$  per i valori  $\mu = 52$  e  $\mu = 48$ , e vedere se questo risultato ci dice qualcosa su come funzionerebbe la procedura di verifica. In particolare, ci chiediamo come funzionerebbe la procedura se volessimo rifiutare  $H_0$  per un valore medio  $\mu = 52$  o  $\mu = 48$ . A causa della simmetria, è necessario valutare solo uno dei due casi, per esempio trovare la probabilità di accettazione dell'ipotesi nulla  $H_0: \mu = 50$  quando la vera media è  $\mu = 52$  cm/s.

**Calcolo della probabilità dell'errore del II tipo  $\beta$ .**

La Figura 4.5 ci aiuterà a calcolare la probabilità dell'errore di II tipo  $\beta$ . La distribuzione normale sulla sinistra della figura è la distribuzione della statistica test  $\bar{X}$  quando l'ipotesi nulla  $H_0: \mu = 50$  è vera (è questo che significa l'espressione "sotto  $H_0: \mu = 50$ "); la distribuzione normale sulla destra è la distribuzione della statistica test  $\bar{X}$  quando è vera l'ipotesi alternativa e il valore della media è 52 (ossia si è "sotto  $H_1: \mu = 52$ "). Ora, verrà commesso un errore del II tipo se la media campionaria cade fra 48.5 e 51.5 (gli estremi della regione critica) quando  $\mu = 52$ . Come si vede in Figura 4.5, questa è semplicemente la probabilità che  $48.5 \leq \bar{X} \leq 51.5$  quando la vera media è  $\mu = 52$ , ovvero l'area ombreggiata sotto la distribuzione normale di destra. Pertanto, riferendosi alla Figura 4.5, si trova

$$\beta = P(48.5 \leq \bar{X} \leq 51.5 \text{ dove } \mu = 52)$$

I valori  $z$  che corrispondono ai valori critici 48.5 e 51.5 quando  $\mu = 52$  sono

$$z_1 = \frac{48.5 - 52}{0.79} = -4.43 \quad \text{e} \quad z_2 = \frac{51.5 - 52}{0.79} = -0.63$$

Pertanto

$$\begin{aligned} \beta &= P(-4.43 \leq Z \leq -0.63) = P(Z \leq -0.63) - P(Z \leq -4.43) \\ &= 0.2643 - 0.000 = 0.2643 \end{aligned}$$

Così, se stiamo verificando l'ipotesi  $H_0: \mu = 50$  contro l'ipotesi alternativa  $H_1: \mu \neq 50$  con  $n = 16$ , e il vero valore della media è  $\mu = 52$ , la probabilità di mancato rifiuto dell'ipotesi

nulla falsa è 0.2643. Per simmetria, se il vero valore della media è  $\mu = 48$ , il valore di  $\beta$  sarà ancora 0.2643.

La probabilità di commettere un errore del II tipo aumenta all'avvicinarsi della vera media  $\mu$  al valore ipotizzato. Si veda per esempio la Figura 4.6, dove il vero valore della media è  $\mu = 50.5$  e quello ipotizzato è  $H_0: \mu = 50$ . Il vero valore di  $\mu$  è molto vicino a 50, e il valore di  $\beta$  risulta

$$\beta = P(48.5 \leq \bar{X} \leq 51.5 \text{ dove } \mu = 50.5)$$

I valori  $z$  che corrispondono ai valori critici 48.5 e 51.5 quando  $\mu = 50.5$  sono

$$z_1 = \frac{48.5 - 50.5}{0.79} = -2.53 \quad \text{e} \quad z_2 = \frac{51.5 - 50.5}{0.79} = 1.27$$

Pertanto

$$\begin{aligned} \beta &= P(-2.53 \leq Z \leq 1.27) = P(Z \leq 1.27) - P(Z \leq -2.53) \\ &= 0.8980 - 0.0057 = 0.8923 \end{aligned}$$

La probabilità dell'errore del II tipo è perciò molto maggiore nel caso in cui la vera media è 50.5 cm/s che non nel caso in cui questa vale 52 cm/s. Naturalmente, in molte situazioni pratiche, non saremmo così interessati all'errore del II tipo se la vera media fosse "vicina" al valore ipotizzato; molto più interessante sarebbe rilevare grandi differenze fra la vera media e il valore specificato nell'ipotesi nulla.

La probabilità dell'errore del II tipo dipende anche dalla dimensione campionaria  $n$ . Si supponga che l'ipotesi nulla sia  $H_0: \mu = 50$  cm/s e che il vero valore della media sia  $\mu = 52$  cm/s. Se la dimensione campionaria viene aumentata da 10 a 16, si ottiene la situazione illustrata in Figura 4.7. La distribuzione normale sulla sinistra è la distribuzione di  $\bar{X}$  quando la media è  $\mu = 50$ , quella sulla destra è la distribuzione di  $\bar{X}$  quando la media è  $\mu = 52$ . Come mostra la figura, la probabilità dell'errore del II tipo è

$$\beta = P(48.5 \leq \bar{X} \leq 51.5 \text{ dove } \mu = 52)$$

Quando  $n = 16$ , la deviazione standard di  $\bar{X}$  è  $\sigma/\sqrt{n} = 2.5/\sqrt{16} = 0.625$ , e i valori  $z$  corrispondenti a 48.5 e 51.5 quando  $\mu = 52$  sono

$$z_1 = \frac{48.5 - 52}{0.625} = -5.60 \quad \text{e} \quad z_2 = \frac{51.5 - 52}{0.625} = -0.80$$

Pertanto

$$\begin{aligned} \beta &= P(-5.60 \leq Z \leq -0.80) = P(Z \leq -0.80) - P(Z \leq -5.60) \\ &= 0.2119 - 0.000 = 0.2119 \end{aligned}$$

Si ricordi che per  $n = 10$  e  $\mu = 52$  avevamo trovato  $\beta = 0.2643$ ; perciò l'aumento della dimensione campionaria fa diminuire la probabilità dell'errore del II tipo.

Di seguito vengono riassunti i risultati di questo paragrafo e altri calcoli simili.

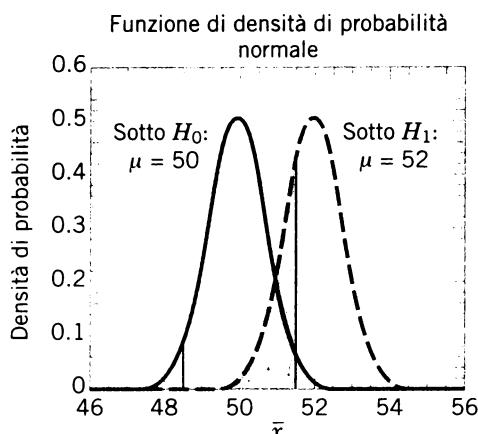


Figura 4.5 Probabilità dell'errore del II tipo quando  $\mu = 52$  e  $n = 10$ .

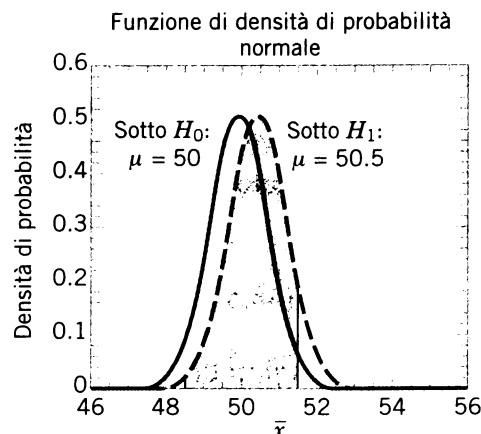


Figura 4.6 Probabilità dell'errore del II tipo quando  $\mu = 50.5$  e  $n = 10$ .

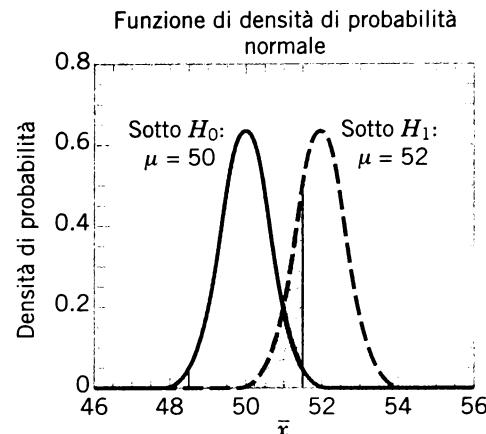


Figura 4.7 Probabilità dell'errore del II tipo quando  $\mu = 52$  e  $n = 16$ .

Mancato rifiuto di $H_0$ quando	Dimensione campionaria	$\alpha$	$\beta$ per $\mu = 52$	$\beta$ per $\mu = 50.5$
$48.5 < \bar{x} < 51.5$	10	0.0576	0.2643	0.8923
$48 < \bar{x} < 52$	10	0.0114	0.5000	0.9705
$48.5 < \bar{x} < 51.5$	16	0.0164	0.2119	0.9445
$48 < \bar{x} < 52$	16	0.0014	0.5000	0.9918

#### Comprensione della relazione fra $\alpha$ e $\beta$ .

I risultati racchiusi in rettangoli non sono stati calcolati nel testo, ma lo loro validità si può facilmente verificare. Questo schema e la precedente discussione rivelano quattro punti importanti.

1. La dimensione della regione critica, e conseguentemente la probabilità  $\alpha$  di un errore del I tipo, può sempre essere ridotta tramite un'opportuna scelta dei valori critici.
2. Gli errori del I e del II tipo sono collegati. Una diminuzione della probabilità dell'errore di un tipo porta sempre a un aumento della probabilità dell'errore dell'altro tipo, purché la dimensione campionaria non vari.
3. Una aumento della dimensione campionaria farà diminuire, in generale, sia  $\alpha$  sia  $\beta$ , purché i valori critici siano mantenuti costanti.
4. Quando l'ipotesi nulla è falsa,  $\beta$  aumenta all'avvicinarsi del vero valore del parametro al valore ipotizzato nell'ipotesi nulla stessa. Il valore di  $\beta$  diminuisce all'aumentare della differenza fra la vera media e il valore ipotizzato.

#### Relazione fra $\beta$ e la dimensione campionaria e fra $\beta$ e la vera differenza tra $\mu$ e $\mu_0$

In generale, l'analista controlla la probabilità  $\alpha$  dell'errore del I tipo quando sceglie i valori critici. Perciò è di solito facile, per l'analista, impostare la probabilità dell'errore del I tipo a un qualsiasi valore desiderato (o a un valore a esso prossimo). Poiché è possibile controllare direttamente la probabilità di rifiutare erroneamente  $H_0$ , si considera sempre il rifiuto dell'ipotesi nulla come una conclusione forte.

Dato che è possibile impostare la probabilità  $\alpha$  di commettere un errore del I tipo (ovvero il livello di significatività), sorge spontaneo chiedersi, allora, quale valore debba essere

usato. La probabilità di un errore del I tipo è la misura di un rischio, precisamente del rischio di concludere che l'ipotesi nulla è falsa quando invece è vera. Pertanto il valore di  $\alpha$  andrebbe scelto in modo da riflettere le conseguenze (economiche, sociali ecc.) di tale rifiuto erroneo. Valori di  $\alpha$  più piccoli sono da scegliersi quando le possibili conseguenze sono più serie, valori più grandi sono adatti quando le conseguenze sono meno severe. È una scelta ardua da compiere; per questo motivi nella pratica scientifica e ingegneristica si utilizza spesso il valore  $\alpha = 0.05$ , a meno che le informazioni di cui si dispone indichino che si tratta di una scelta inappropriate. Nel problema del propellente con  $n = 10$ , questa scelta corrisponderebbe ai valori critici 48.45 e 51.55.

Una procedura ampiamente utilizzata nella verifica di ipotesi prevede di adottare il livello di significatività  $\alpha = 0.05$ . Questo valore è stato ricavato dalle esperienze passate, e può non essere appropriato per tutte le situazioni.

Dall'altro lato, la probabilità  $\beta$  dell'errore del II tipo non è una costante: essa dipende dal vero valore del parametro, nonché dalla dimensione campionaria scelta. Poiché la probabilità dell'errore del II tipo è una funzione sia della dimensione campionaria sia dell'estensione della regione su cui l'ipotesi nulla è falsa, si usa considerare la decisione di non rifiutare  $H_0$  una conclusione debole, a meno che si sappia che  $\beta$  è accettabilmente piccola. Per questo motivo, anziché dire “accettazione di  $H_0$ ” si preferisce usare l'espressione “mancato rifiuto di  $H_0$ ”. Il mancato rifiuto di  $H_0$  implica che non si hanno sufficienti prove per rifiutare l'ipotesi nulla, ovvero per arrivare a una conclusione forte. Non significa necessariamente, invece, che vi è un'elevata probabilità che  $H_0$  sia vera; può significare semplicemente che sono necessari più dati per raggiungere tale conclusione forte. Queste considerazioni possono avere ripercussioni significative sulla formulazione delle ipotesi.

Un concetto importante di cui faremo uso è la potenza di un test statistico.

#### Potenza di un test

La **potenza** di un test statistico è la probabilità di rifiutare l'ipotesi nulla  $H_0$  quando l'ipotesi alternativa è vera.

La potenza viene calcolata come  $1 - \beta$ , e può essere interpretata come la probabilità di rifiutare correttamente un'ipotesi nulla falsa. Spesso si confrontano i test statistici sulla base delle proprietà delle rispettive potenze. Per esempio, si consideri il problema della velocità di combustione del propellente in cui si verifica l'ipotesi  $H_0: \mu = 50$  cm/s contro  $H_1: \mu \neq 50$  cm/s. Si supponga che il vero valore della media sia  $\mu = 52$ . Per  $n = 10$  abbiamo trovato  $\beta = 0.2643$ , perciò la potenza di questo test è  $1 - \beta = 1 - 0.2643 = 0.7357$  per  $\mu = 52$ .

La potenza è una misura molto descrittiva e concisa della **sensibilità** di un test statistico, dove con il termine “sensibilità” si intende la capacità del test di rilevare le differenze. Nel nostro caso la sensibilità del test nel rilevamento della differenza fra una velocità media di combustione pari a 50 e una pari a 52 è 0.7357. In altre parole, se la vera media è proprio 52 cm/s, questo test rifiuterà correttamente l'ipotesi  $H_0: \mu = 50$  e “rileverà” questa differenza il 73.57% delle volte. Se l'analista ritiene questo livello di potenza troppo basso, può aumentare  $\alpha$  o la dimensione campionaria  $n$ .

### 4.3.3 Il P-value nella verifica di ipotesi

L'approccio alla verifica di ipotesi che abbiamo schematizzato nei paragrafi precedenti ha fatto uso di un **livello di significatività fissato**  $\alpha$ , dove  $\alpha$  viene di solito scelto pari a 0.05. Si tratta di un approccio comodo in quanto porta direttamente a definire l'errore del II tipo e la potenza, concetti molto utili e di notevole utilità quando si tratta di determinare una dimensione campionaria appropriata per una data verifica di ipotesi.

Fissare il livello di significatività, tuttavia, comporta degli svantaggi. Per esempio, si supponga di venire a sapere che l'ipotesi nulla relativa alla velocità di combustione del propellente è stata rifiutata al livello di significatività  $\alpha = 0.05$ . Questo tipo di conclusione è spesso inadeguata, perché non indica al decisore se la velocità di combustione media è di poco dentro alla regione critica o ben addentro a essa, ossia non informa sulla consistenza degli indizi contro l'ipotesi nulla. Inoltre, enunciare i risultati in questo modo costringe altri utenti dell'informazione a riferirsi al livello di significatività predefinito, e alcuni responsabili dei processi decisionali potrebbero non trovare accettabile una probabilità di commettere un errore del I tipo pari a 0.05.

Per evitare questi potenziali problemi, nella pratica si è diffusa l'adozione di un **approccio basato sul P-value**. Quest'ultimo rappresenta la probabilità che la media campionaria assuma un valore almeno altrettanto estremo quanto il valore osservato quando l'ipotesi nulla  $H_0$  è vera. In altri termini, il  $P$ -value porta con sé un'informazione sul peso degli indizi contro  $H_0$ : più è piccolo, maggiore è l'indizio contro  $H_0$ . Quando il  $P$ -value è sufficientemente piccolo, finiamo per rifiutare l'ipotesi nulla in favore dell'ipotesi alternativa. Questo tipo di approccio consente a chi deve decidere di trarre conclusioni per *ogni* livello di significatività ritenuto opportuno.

Diamo ora una definizione formale di  $P$ -value.

#### Definizione

**Il  $P$ -value è il più piccolo livello di significatività che porterebbe al rifiuto dell'ipotesi nulla  $H_0$ .**

Per illustrare il concetto di  $P$ -value, consideriamo il problema della velocità di combustione del propellente, per il quale si hanno le ipotesi

$$H_0: \mu = 50 \text{ cm/s}$$

$$H_1: \mu \neq 50 \text{ cm/s}$$

Sappiamo inoltre che  $\sigma = 2.5 \text{ cm/s}$ . Supponiamo che un campione casuale costituito da  $n = 10$  provini dia luogo a una media campionaria  $\bar{x} = 51.8 \text{ cm/s}$ . La Figura 4.8 mostra come viene calcolato il  $P$ -value.

La curva normale in questa figura è la distribuzione della media campionaria sotto l'ipotesi nulla; la media è  $\mu = 50$  e la deviazione standard è  $\sigma/\sqrt{n} = 2.5/\sqrt{10} = 0.79$ . Il valore 51.8 è il valore osservato della media campionaria. La probabilità di osservare un valore della media campionaria che sia come minimo uguale a 51.8 viene determinato calcolando

$$z = \frac{51.8 - 50}{0.79} = 2.28$$

e la probabilità che la variabile aleatoria normale standard sia maggiore o uguale a 2.28 è 0.0113. Essendo l'ipotesi nulla bilaterale, 0.0113 è metà del  $P$ -value. Dobbiamo anche con-

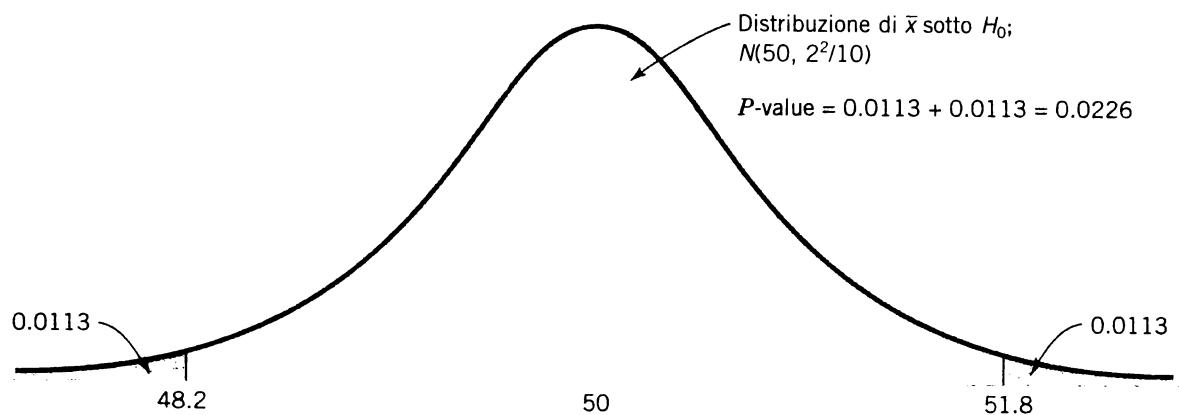


Figura 4.8 Calcolo del  $P$ -value per il problema della velocità di combustione del propellente.

siderare il caso in cui  $z$  possa assumere valori negativi, ossia  $z = -2.28$  (la corrispondenza sarebbe allora con il punto 48.2 mostrato in Figura 4.8). Dato che la curva normale è simmetrica, la probabilità che la variabile aleatoria standard sia minore o uguale a  $-2.28$  è anch'essa 0.0113. Pertanto, il  $P$ -value per questa verifica di ipotesi è

$$P = 0.0113 + 0.0113 = 0.0226$$

#### Interpretazione del $P$ -value

Il  $P$ -value ci dice che se l'ipotesi nulla  $H_0$  è vera, la probabilità di ottenere un campione casuale la cui media dista da 50 almeno quanto dista da tale valore il punto 51.8 (o il punto 48.2) vale 0.0226. Perciò, l'osservazione di una media campionaria uguale a 51.8 è un evento raro, se l'ipotesi nulla è effettivamente vera. In confronto al livello di significatività “standard” di 0.05, il nostro  $P$ -value osservato è più piccolo, per cui se stessimo utilizzando il livello di significatività 0.05 saremmo portati a rifiutare l'ipotesi nulla. In effetti  $H_0$  verrebbe rifiutata a *qualsiasi* livello di significatività maggiore o uguale a 0.0226. È quanto riportato nel precedente riquadro che contiene la definizione di  $P$ -value: il  $P$ -value è il minore livello di significatività che porterebbe al rifiuto dell'ipotesi nulla  $H_0$ .

Da un punto di vista operativo, una volta determinato il  $P$ -value lo si confronta tipicamente con un livello di significatività predefinito per assumere una decisione. Spesso tale livello predefinito è uguale a 0.05, ma nella presentazione dei risultati e delle conclusioni è pratica usuale riportare il  $P$ -value osservato, assieme alla decisione presa riguardo all'ipotesi nulla.

È chiaro che il  $P$ -value fornisce una misura della **credibilità** dell'ipotesi nulla, nel senso che misura il **peso degli indizi** contro  $H_0$ . Più precisamente, quantifica il rischio di prendere la decisione sbagliata che si corre rifiutando  $H_0$ .

**Il  $P$ -value non è la probabilità che l'ipotesi nulla sia vera, né  $1 - P$  è la probabilità che sia falsa. L'ipotesi nulla o è vera o è falsa (non vi è una probabilità associata a ciò); il  $P$ -value va correttamente interpretato in termini di rischio associato a un rifiuto erroneo di  $H_0$ .**

In questo volume faremo ampio uso dell'approccio basato sul  $P$ -value. I software statistici moderni riportano i risultati delle verifiche di ipotesi quasi esclusivamente in termini di  $P$ -value.

#### 4.3.4 Ipotesi unilaterali e bilaterali

Una verifica di ipotesi come

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

viene detta **bilaterale** perché è importante rilevare le differenze rispetto al valore della media ipotizzato,  $\mu_0$ , che giacciono sia a destra sia a sinistra di  $\mu_0$ . In una simile verifica la regione critica è suddivisa in due parti, di solito aventi uguale probabilità, poste in ciascuna coda della distribuzione della statistica test.

Molti problemi di verifica di ipotesi coinvolgono in modo naturale un'ipotesi alternativa **unilaterale**, come

$$\begin{array}{ll} H_0: \mu = \mu_0 & H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 & H_1: \mu < \mu_0 \end{array}$$

Se l'ipotesi alternativa è  $H_1: \mu > \mu_0$  la regione critica dovrebbe cadere nella coda superiore della distribuzione della statistica test, mentre se l'ipotesi alternativa è  $H_1: \mu < \mu_0$ , la regione critica dovrebbe cadere nella coda inferiore della distribuzione. Di conseguenza, queste verifiche vengono dette a volte test a una coda. La collocazione della regione critica per le verifiche unilaterali è di solito facile da stabilire. È sufficiente visualizzare il comportamento della statistica test se l'ipotesi nulla è vera e collocare la regione critica nell'estremità o coda appropriata della distribuzione. In generale la disuguaglianza nell'ipotesi alternativa "punta" nel verso della regione critica.

**Formulazione dell'ipotesi nulla.**

Nel costruire le ipotesi formuleremo sempre l'ipotesi nulla come uguaglianza, in modo da poter controllare a uno specifico valore la probabilità dell'errore del I tipo  $\alpha$  (si faccia riferimento alla precedente nota a piè pagina). L'ipotesi alternativa può essere sia unilaterale sia bilaterale, a seconda delle conclusioni da trarre nel caso  $H_0$  venga rifiutata. Se l'obiettivo è di fare un'asserzione basata su espressioni come "maggiore di", "minore di", "superiore a", "supera", "almeno" e via dicendo, è appropriata un'alternativa unilaterale. Se l'asserzione non implica alcuna direzione o verso, o se si deve utilizzare l'espressione "diverso da", si deve adottare un'alternativa bilaterale.

**ESEMPIO 4.2**  
Propellente per razzi

Si consideri il problema della velocità di combustione del propellente. Si supponga che se tale velocità è minore di 50 cm/s si voglia mostrare ciò con una conclusione forte. Le ipotesi dovrebbero allora venire formulate come segue

$$H_0: \mu = 50 \text{ cm/s}$$

$$H_1: \mu < 50 \text{ cm/s}$$

In questo caso il  $P$ -value va calcolato determinando la probabilità che la variabile aleatoria normale abbia valore minore del valore osservato di  $\bar{X}$ . Ovvero, il  $P$ -value viene calcolato dalla coda inferiore della distribuzione di  $\bar{X}$  sotto l'ipotesi nulla. Poiché il rifiuto di  $H_0$  è sempre una conclusione forte, questa enunciazione dell'ipotesi produrrà la risposta desiderata.

ta se  $H_0$  viene rifiutata. Si noti che benché l'ipotesi nulla sia scritta con il segno di uguale, si sottintende che comprende ogni valore di  $\mu$  non specificato dall'ipotesi alternativa. Perciò il mancato rifiuto di  $H_0$  non significa che  $\mu$  è esattamente uguale a 50 cm/s, ma solo che non abbiamo sufficienti prove per sostenere  $H_1$ .

In alcuni problemi applicativi in cui sono indicate procedure di verifica unilaterali è a volte difficile scegliere un'appropriata formulazione dell'ipotesi alternativa. Per esempio, si supponga che un imbottigliatore di bevande analcoliche acquisti bottiglie da 2 litri da una fabbrica di contenitori in vetro. L'imbottigliatore vuole essere sicuro che le bottiglie soddisfino le specifiche sulla pressione media interna o sulla resistenza allo scoppio, che per le bottiglie da 2 litri deve essere almeno pari a 200 psi. L'imbottigliatore ha deciso di formulare la procedura decisionale per uno specifico lotto di contenitori sotto forma di problema di verifica di ipotesi. Vi sono due possibili formulazioni per questo problema:

$$\begin{aligned} H_0: \mu &= 200 \text{ psi} \\ H_1: \mu &> 200 \text{ psi} \end{aligned} \quad (4.9)$$

oppure

$$\begin{aligned} H_0: \mu &= 200 \text{ psi} \\ H_1: \mu &< 200 \text{ psi} \end{aligned} \quad (4.10)$$

Si consideri la formulazione (4.9). Se l'ipotesi nulla viene rifiutata le bottiglie saranno valutate come soddisfacenti, mentre il mancato rifiuto di  $H_0$  significa che le bottiglie non sono conformi alle specifiche e dovrebbero essere scartate. Essendo il rifiuto di  $H_0$  una conclusione forte, questa formulazione costringe il produttore delle bottiglie a "dimostrare" che la resistenza media allo scoppio supera le specifiche. Si consideri ora l'altra formulazione, la (4.10); in questa seconda situazione le bottiglie saranno ritenute soddisfacenti a meno che  $H_0$  sia rifiutata, vale a dire che si concluderà che le bottiglie sono soddisfacenti a meno di forti prove a sostegno del contrario.

Qual è dunque la formulazione corretta, la (4.9) o la (4.10)? La risposta è: "dipende". Per la formulazione (4.9) vi è qualche probabilità che  $H_0$  non venga rifiutata (si deciderebbe allora di non usare le bottiglie) anche se la vera media è leggermente maggiore di 200 psi. La formulazione implica quindi che si vuole sia il produttore a dimostrare che il suo prodotto soddisfa le specifiche minime. Potrebbe allora essere appropriata se si sa che tale produttore ha avuto difficoltà, nel passato, a fornire prodotti soddisfacenti le specifiche, o se considerazioni a livello di sicurezza impongono di attenersi strettamente alla specifica dei 200 psi. Dall'altro lato, per la formulazione (4.10) vi è qualche probabilità che  $H_0$  venga accettata, e le bottiglie giudicate soddisfacenti, anche se la vera media è leggermente minore di 200 psi. Concluderemmo insomma che le bottiglie sono insoddisfacenti solo quando vi fosse una forte evidenza del fatto che la media non supera i 200 psi, ossia quando  $H_0: \mu = 200$  psi venisse rifiutata. Questa formulazione assume che noi siamo abbastanza soddisfatti dalle prestazioni passate del produttore, e che piccole deviazioni dalla specifica psi non sono dannose.

Nel formulare ipotesi alternative unilaterali dovremmo ricordare che il rifiuto di  $H_0$  è sempre una conclusione forte. Di conseguenza, **dovremmo porre l'asserzione su cui è importante trarre una conclusione forte nell'ipotesi alternativa**. Nei problemi che trattano questioni del mondo reale, la scelta dipenderà spesso dal nostro punto di vista e dall'esperienza passata con la situazione.

#### 4.3.5 Procedura generale per la verifica di ipotesi

In questo capitolo sono sviluppate procedure di verifica di ipotesi per molti problemi pratici. Si raccomanda a tale proposito di usare la seguente successione di sette passi, che verrà illustrata nei prossimi paragrafi.

1. **Parametro di interesse:** identificare dal contesto del problema il parametro di interesse.
2. **Ipotesi nulla  $H_0$ :** formulare l'ipotesi nulla,  $H_0$ .
3. **Ipotesi alternativa  $H_1$ :** specificare un'opportuna ipotesi alternativa,  $H_1$ .
4. **Statistica test:** scegliere un'appropriata statistica test.
5. **Rifiutare  $H_0$  se:** stabilire il criterio che condurrà al rifiuto di  $H_0$ .
6. **Calcoli:** calcolare ogni quantità del campione necessaria, sostituire i valori ricavati nell'equazione per la statistica test e calcolarne il valore.
7. **Conclusioni:** decidere se  $H_0$  dovrebbe o meno essere rifiutata e riportare tale decisione nel contesto del problema. Ciò può comportare il calcolo di un  $P$ -value o il confronto della statistica test con un insieme di valori critici.

I passi 1-5 dovrebbero venire completati prima dell'esame dei dati campionari.

#### 4.4 INFERENZA SULLA MEDIA DI UNA POPOLAZIONE CON VARIANZA NOTA

In questo paragrafo consideriamo le inferenze sulla media  $\mu$  di una singola popolazione la cui varianza  $\sigma^2$  è nota.

##### Assunzioni

1.  $X_1, X_2, \dots, X_n$  è un campione casuale di dimensione  $n$  estratto da una popolazione;
2. la popolazione è normale o, se non lo è, si applicano le condizioni del teorema limite centrale.

In base alla discussione del Paragrafo 4.2, la media campionaria  $\bar{X}$  è uno **stimatore puntuale non distorto** di  $\mu$ . Con queste assunzioni la distribuzione di  $\bar{X}$  è approssimativamente normale con media  $\mu$  e varianza  $\sigma^2/n$ .

Sotto le precedenti assunzioni, la quantità

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad (4.11)$$

ha una distribuzione normale standard  $N(0, 1)$ .

#### 4.4.1 Verifica di ipotesi sulla media

Si supponga di volere verificare le ipotesi

$$\begin{aligned} H_0: \mu &= \mu_0 \\ H_1: \mu &\neq \mu_0 \end{aligned} \quad (4.12)$$

dove  $\mu_0$  è una determinata costante. Abbiamo un campione casuale  $X_1, X_2, \dots, X_n$  estratto dalla popolazione. Poiché  $\bar{X}$  ha una distribuzione approssimativamente normale (ossia la **distribuzione campionaria** di  $\bar{X}$  è approssimativamente normale) con media  $\mu_0$  e deviazione standard  $\sigma/\sqrt{n}$  se l'ipotesi nulla è vera, potremo calcolare un  $P$ -value o, nel caso vogliamo usare un livello di significatività fissato, potremo costruire una regione critica per il valore calcolato della media campionaria  $\bar{x}$  come nei Paragrafi 4.3.2 e 4.3.3.

Di solito è più conveniente **standardizzare** la media campionaria e usare una statistica test basata sulla distribuzione normale standard, la procedura è spesso detta test  $z$ ; la procedura di verifica per  $H_0: \mu = \mu_0$  usa cioè la **statistica test**

**Statistica test  
per il test  $z$**

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \quad (4.13)$$

Se l'ipotesi nulla  $H_0: \mu = \mu_0$  è vera,  $E(\bar{X}) = \mu_0$ , e si ha che la distribuzione di  $Z_0$  è la distribuzione normale standard [indicata, come sappiamo, con  $N(0, 1)$ ].

Il denominatore nell'Equazione (4.13),  $\sigma/\sqrt{n}$ , è l'**errore standard** della media campionaria  $\bar{X}$ . Perciò la forma generale della statistica test è

$$\frac{\text{differenza fra media campionaria e media ipotizzata}}{\text{errore standard}}$$

ed è quella che si incontra in quasi tutte le statistiche test per le medie.

Si supponga di selezionare un campione casuale di dimensione  $n$  e di osservare una media campionaria  $\bar{x}$ . Per verificare l'ipotesi nulla con l'approccio del  $P$ -value dobbiamo trovare la probabilità di osservare, posto che  $H_0$  sia vera, un valore della media campionaria almeno pari a  $\bar{x}$ . Il valore  $z$  della distribuzione normale standard che corrisponde a  $\bar{x}$  viene determinato a partire dalla statistica test dell'Equazione (4.13):

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

In termini di funzione di distribuzione cumulativa della normale standard, stiamo cercando la probabilità  $1 - \Phi(|z_0|)$ ; nell'argomento di questa funzione compare il valore assoluto perché  $z_0$  potrebbe essere sia positivo sia negativo, a seconda del valore della media campionaria osservato. Trattandosi di un test a due code, questa espressione fornisce solo metà del  $P$ -value. Perciò, per l'ipotesi alternativa bilaterale, il  $P$ -value è

$$P = 2[1 - \Phi(|z_0|)] \quad (4.14)$$

Il  $P$ -value relativo a questo caso è illustrato in Figura 4.9a.

Consideriamo adesso le alternative unilaterali. Supponiamo di sottoporre a verifica

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu > \mu_0 \end{aligned} \quad (4.15)$$

Ancora una volta, immaginiamo di avere un campione casuale di dimensione  $n$  e che la media campionaria sia  $\bar{x}$ . Calcoliamo la statistica test mediante l'Equazione (4.13), ottenendo  $z_0$ . Trattandosi di un test sulla coda superiore, solo i valori di  $\bar{x}$  maggiori di  $\mu_0$  sono concordi con l'ipotesi alternativa. Pertanto, il  $P$ -value misura la probabilità che la variabile aleatoria normale standard sia maggiore del valore  $z_0$  della statistica test; lo si calcola mediante

$$P = 1 - \Phi(z_0) \quad (4.16)$$

Il  $P$ -value relativo a questo caso è illustrato in Figura 4.9b.

Il test sulla coda inferiore riguarda le ipotesi

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu < \mu_0 \end{aligned} \quad (4.17)$$

Supponiamo di avere un campione casuale di dimensione  $n$  e che la media campionaria sia  $\bar{x}$ . Calcoliamo la statistica test mediante l'Equazione (4.13), ottenendo  $z_0$ . Trattandosi di un test sulla coda inferiore, solo i valori di  $\bar{x}$  minori di  $\mu_0$  sono concordi con l'ipotesi alternativa. Pertanto, il  $P$ -value misura la probabilità che la variabile aleatoria normale standard sia minore del valore  $z_0$  della statistica test; lo si calcola mediante

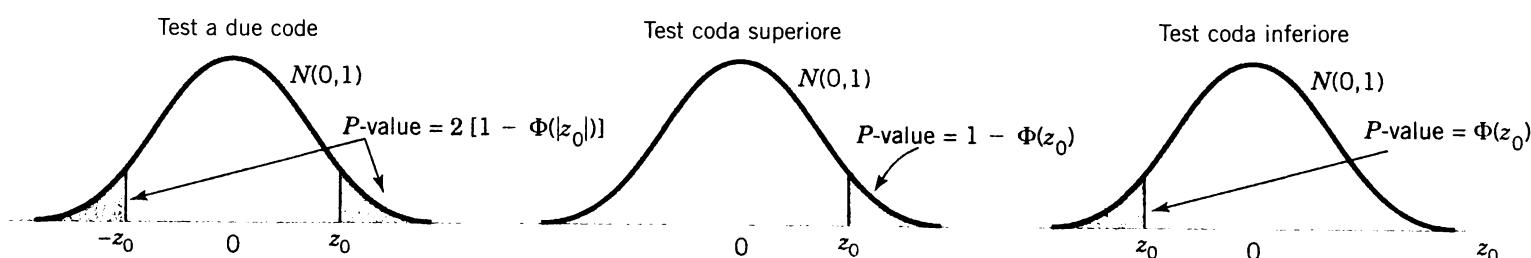


Figura 4.9 Il  $P$ -value per un test  $z$ . (a) l'alternativa bilaterale  $H_1: \mu \neq \mu_0$ ; (b) l'alternativa unilaterale  $H_1: \mu > \mu_0$ ; (c) l'alternativa unilaterale  $H_1: \mu < \mu_0$ .

$$P = \Phi(z_0) \quad (4.16)$$

Il  $P$ -value relativo a questo caso è illustrato in Figura 4.9c.

Non è sempre facile calcolare il  $P$ -value esatto per una statistica test, ma la maggior parte dei software di analisi statistica più recenti ha una funzione che permette di calcolarlo, e lo stesso vale per alcune calcolatrici tascabili scientifiche. Più avanti mostreremo anche come approssimare i  $P$ -value.

È possibile altresì eseguire verifiche di ipotesi con livelli di significatività fissati mediante il **test  $z$** . Tutto ciò che dobbiamo fare è decidere dove collocare le regioni critiche per le ipotesi alternative bilaterali e unilaterali.

Consideriamo innanzitutto l'alternativa bilaterale espressa dall'Equazione (4.12).

Ora, se  $H_0: \mu = \mu_0$  è vera, si ha una probabilità pari a  $1 - \alpha$  che la statistica test  $Z_0$  cada fra  $-z_{\alpha/2}$  e  $z_{\alpha/2}$ , dove  $z_{\alpha/2}$  è il  $100\alpha/2$ -esimo punto percentuale della distribuzione normale standard. Le regioni associate con  $z_{\alpha/2}$  e  $-z_{\alpha/2}$  sono illustrate in Figura 4.10a. Si noti che si ha una probabilità  $\alpha$  che la statistica test  $Z_0$  cada nella regione  $Z_0 > z_{\alpha/2}$  o  $Z_0 < -z_{\alpha/2}$  quando  $H_0: \mu = \mu_0$  è vera. Chiaramente, un campione che produca un valore della statistica test che cada nelle code della distribuzione di  $Z_0$  sarebbe insolito se  $H_0: \mu = \mu_0$  fosse vera; perciò, si ha un'indicazione che  $H_0$  è falsa. Pertanto, dovremmo rifiutare  $H_0$  se

$$\underline{z_0 > z_{\alpha/2}} \quad (4.19)$$

o se

$$\underline{z_0 < -z_{\alpha/2}} \quad (4.20)$$

e dovremmo non rifiutare  $H_0$  se

$$\underline{-z_{\alpha/2} \leq z_0 \leq z_{\alpha/2}} \quad (4.21)$$

Le formule (4.19) e (4.20) definiscono la **regione critica** o **regione di rifiuto** per la verifica. La **probabilità dell'errore del I tipo** per questa procedura è  $\alpha$ .

Possiamo anche sviluppare delle procedure per la verifica di ipotesi sulla media  $\mu$  in cui l'ipotesi alternativa è unilaterale. Si consideri il caso della coda superiore espresso dall'Equazione (4.15).

Nel definire la regione critica per questa verifica osserviamo che un valore negativo della statistica test  $Z_0$  non ci porterebbe mai a concludere che  $H_0: \mu = \mu_0$  è falsa. Perciò, por-

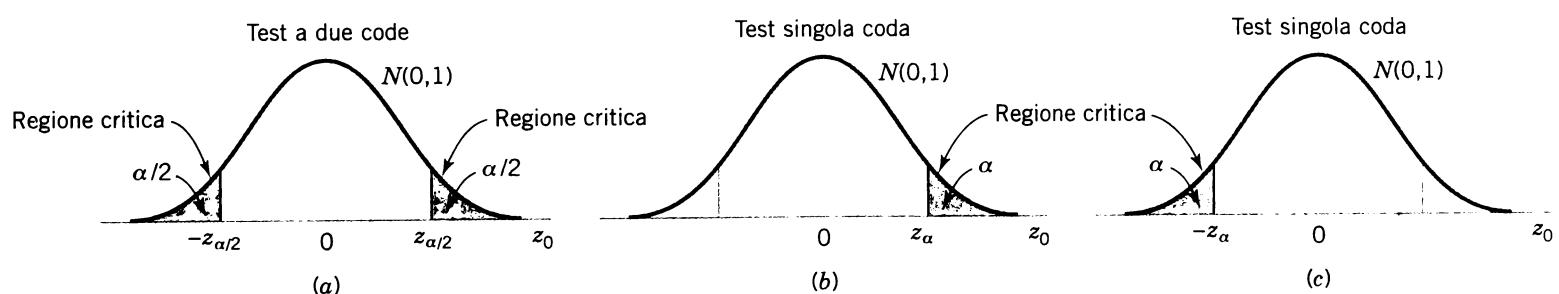


Figura 4.10 La distribuzione di  $Z_0$  quando  $H_0: \mu = \mu_0$  è vera, con regione critica per (a) l'alternativa bilaterale  $H_1: \mu \neq \mu_0$ ; (b) l'alternativa unilaterale  $H_1: \mu > \mu_0$ ; (c) l'alternativa unilaterale  $H_1: \mu < \mu_0$ .

remmo la regione critica nella coda superiore della distribuzione normale standard e rifiuteremmo  $H_0$  se il valore di  $Z_0$  calcolato fosse troppo grande. (Si faccia riferimento alla Figura 4.10b.) In altre parole, rifiuteremmo  $H_0$  se

$$z_0 > z_\alpha \quad (4.22)$$

Analogamente, per verificare il caso della coda inferiore dell'Equazione (4.17) calcoleremmo la statistica test  $Z_0$  e rifiuteremmo  $H_0$  se il valore di  $Z_0$  calcolato fosse troppo piccolo. In altri termini, la regione critica è nella coda inferiore della distribuzione normale standard come in Figura 4.10c, e rifiutiamo  $H_0$  se

$$z_0 < -z_\alpha \quad (4.23)$$

<b>Verifica di ipotesi sulla media, varianza nota (test z)</b>		
<b>Ipotesi nulla:</b>	$H_0: \mu = \mu_0$	<b>Criterio di rifiuto per test con livello fissato</b>
Statistica test:	$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$	
<b>Ipotesi alternativa</b>	<b>P-value</b>	
$H_1: \mu \neq \mu_0$	Probabilità a destra di $ z_0 $ e a sinistra di $- z_0 $ , $P = 2[1 - \Phi( z_0 )]$	$z_0 > z_{\alpha/2}$ o $z_0 < -z_{\alpha/2}$
$H_1: \mu > \mu_0$	Probabilità a destra di $z_0$ , $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1: \mu < \mu_0$	Probabilità a sinistra di $z_0$ , $P = \Phi(z_0)$	$z_0 < -z_\alpha$

I P-value e le re regioni critiche per queste situazioni sono mostrate nelle Figure 4.9 e 4.10.

### ESEMPIO 4.3 Velocità di combustione del propellente

I sistemi di espulsione automatica di sicurezza per i piloti di aerei da caccia sono alimentati con propellente solido. La velocità di combustione di questo propellente è un'importante caratteristica del prodotto. Le specifiche richiedono che la velocità media di combustione sia 50 cm/s. Si sa che la deviazione standard della velocità di combustione è  $\sigma = 2$  cm/s. Lo sperimentatore decide di specificare una probabilità dell'errore di I tipo, o livello di significatività,  $\alpha = 0.05$ ; seleziona quindi un campione casuale di dimensione  $n = 25$  e ottiene una velocità di combustione media campionaria  $\bar{x} = 51.3$  cm/s. Quali conclusioni dovrebbe trarre?

Possiamo risolvere questo problema seguendo la procedura a sette passi schematizzata nel Paragrafo 4.3.5, ottenendo quanto segue.

1. **Parametro di interesse:** il parametro di interesse è  $\mu$ , la velocità media di combustione.
2. **Ipotesi nulla  $H_0$ :**  $\mu = 50$  cm/s
3. **Ipotesi alternativa  $H_1$ :**  $\mu \neq 50$  cm/s
4. **Statistica test:** la statistica test è

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

0,05

5. **Rifiutare  $H_0$  se:** si rifiuta  $H_0$  se il  $P$ -value è più piccolo di 0.02 (si noti che i limiti della corrispondente regione critica per il livello di significatività fissato sarebbero  $-z_{0.025} = -1.96$  e  $z_{0.025} = 1.96$ ).
6. **Calcoli:** poiché  $\bar{x}$  e  $\sigma = 2$ , si ha

$$z_0 = \frac{51.3 - 50}{2/\sqrt{25}} = \frac{1.3}{0.4} = 3.25$$

7. **Conclusioni:** il  $P$ -value è  $P = 2 [1 - \Phi(3.25)] = 0.0012$ . Essendo minore di 0.05, rifiutiamo  $H_0 : \mu = 50$ . Dal **punto di vista ingegneristico**, concludiamo che la velocità media di combustione è diversa da 50 cm/s, sulla base di un campione di 25 misure. In effetti, vi è una forte evidenza che la velocità media di combustione superi i 50 cm/s.

Utilizziamo Minitab per eseguire il test  $z$  per il problema della velocità di combustione dell'esempio 4.3. Otteniamo l'output mostrato qui sotto. Si noti che Minitab riporta l'**errore standard** della media ( $\sigma/\sqrt{n} = 0.4$ ), indicato come *SE Mean*, oltre al  $P$ -value e a un **intervallo di confidenza** (CI) per la velocità di combustione media. Nel Paragrafo 4.4.5 spiegheremo come calcolare e come interpretare questo intervallo.

### One-Sample Z

Test of mu = 50 vs not = 50  
The assumed standard deviation = 2

N	Mean	SE Mean	95% CI	Z	P
25	51.3000	0.4000	(50.5160, 52.0840)	3.25	0.0012

## 4.4.2 Errore del II tipo e scelta della dimensione campionaria

Nella verifica di ipotesi, l'analista sceglie direttamente la probabilità dell'errore del I tipo. Tuttavia, la probabilità dell'errore del II tipo dipende dalla scelta della dimensione campionaria. In questo paragrafo mostreremo come calcolare la probabilità dell'errore del II tipo e come selezionare la dimensione campionaria in modo da ottenere un determinato valore di  $\beta$ .

Trovare la probabilità dell'errore del II tipo  $\beta$

Si consideri l'ipotesi bilaterale

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

Si supponga che l'ipotesi nulla sia falsa e che il vero valore della media sia  $\mu = \mu_0 + \delta$ , per esempio, con  $\delta > 0$ . Il valore atteso della statistica test  $Z_0$  è

$$E(Z_0) = E\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right) = \frac{\mu_0 + \delta - \mu_0}{\sigma/\sqrt{n}} = \frac{\delta\sqrt{n}}{\sigma}$$

e la varianza di  $Z_0$  è uguale a 1. Pertanto, la distribuzione di  $Z_0$  quando  $H_1$  è vera è

$$Z_0 \sim N\left(\frac{\delta\sqrt{n}}{\sigma}, 1\right) \quad (4.23)$$

dove la notazione  $\sim$  significa “è distribuita come”. La distribuzione della statistica test  $Z_0$  sotto entrambe le ipotesi, nulla e alternativa, è mostrata in Figura 4.11. Esaminando tale figura si nota che se  $H_1$  è vera, verrà commesso un errore del II tipo solo se  $-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}$ , dove  $Z_0 \sim N(\delta\sqrt{n}/\sigma, 1)$ . In altri termini, la probabilità dell'errore del II tipo  $\beta$  è la probabilità che  $Z_0$  cada fra  $-z_{\alpha/2}$  e  $z_{\alpha/2}$  posto che  $H_1$  sia vera. Questa probabilità è mostrata in Figura 4.11 come porzione ombreggiata, ed è espressa matematicamente dalla seguente equazione.

**Probabilità di un errore del II tipo per l'ipotesi alternativa bilaterale  
sulla media, varianza nota**

$$\beta = \Phi\left(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) \quad (4.24)$$

dove  $\Phi(z)$  indica la probabilità a sinistra di  $z$  nella distribuzione normale standard. Si noti che l'Equazione (4.24) è stata ricavata valutando la probabilità che  $Z_0$  cada nell'intervallo

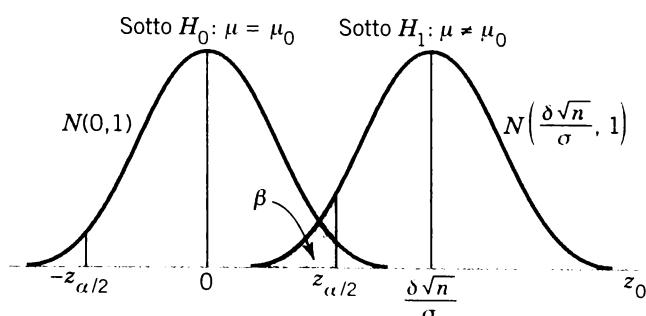


Figura 4.11 La distribuzione di  $Z_0$  sotto  $H_0$  e  $H_1$ .

$[-z_{\alpha/2}, z_{\alpha/2}]$  quando  $H_1$  è vera. Inoltre, si osservi che l'Equazione (4.24) vale anche se  $\delta < 0$ , per via della simmetria della distribuzione normale. È possibile anche ricavare un'equazione simile alla (4.24) per un'ipotesi alternativa unilaterale.

#### Formule per la dimensione campionaria

Si possono facilmente ricavare formule che determinano la dimensione campionaria appropriata per ricavare un particolare valore di  $\beta$ , dati  $\alpha$  e  $\delta$ . Per l'ipotesi alternativa bilaterale, sappiamo dall'Equazione (4.24) che

$$\beta = \Phi\left(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right)$$

o, se  $\delta > 0$

$$\beta \approx \Phi\left(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) \quad (4.25)$$

perché  $\Phi(-z_{\alpha/2} - \delta\sqrt{n}/\sigma) \approx 0$  quando  $\delta$  è positiva. Sia  $z_\beta$  il  $100\beta$ -esimo percentile della distribuzione normale standard. Allora  $\beta = \Phi(-z_\beta)$ . Dall'Equazione (4.25)

$$-z_\beta \approx z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}$$

che porta alla seguente equazione.

#### Dimensione campionaria per l'ipotesi alternativa bilaterale sulla media, varianza nota

Per l'ipotesi alternativa bilaterale sulla media con varianza nota e livello di significatività  $\alpha$ , la dimensione campionaria necessaria per rilevare una differenza fra la vera media e quella ipotizzata pari a  $\delta$  con potenza uguale ad almeno  $1 - \beta$  è

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\delta^2} \quad (4.26)$$

dove

$$\delta = \mu - \mu_0$$

Se  $n$  non è intero, per convenzione si arrotonda sempre la dimensione campionaria all'intero successivo.

Questa approssimazione è buona quando  $\Phi(-z_{\alpha/2} - \delta\sqrt{n}/\sigma) \approx 0$  è piccolo in confronto a  $\beta$ . Per ognuna delle ipotesi alternative unilaterali, la dimensione campionaria necessaria per produrre un determinato errore del II tipo con probabilità  $\beta$  date  $\delta$  e  $\alpha$  è la seguente.

### Dimensione campionaria per l'ipotesi alternativa unilaterale sulla media, varianza nota

Per l'ipotesi alternativa unilaterale sulla media con varianza nota e livello di significatività  $\alpha$ , la dimensione campionaria necessaria per rilevare una differenza fra la vera media e quella ipotizzata pari a  $\delta$  con potenza uguale ad almeno  $1 - \beta$  è

$$n = \frac{(z_\alpha + z_\beta)^2 \sigma^2}{\delta^2} \quad (4.27)$$

dove

$$\delta = \mu - \mu_0$$

Se  $n$  non è intero, per convenzione si arrotonda sempre la dimensione campionaria all'intero successivo.

#### ESEMPIO 4.4 Dimensione campionaria per il problema della velocità di combustione

Si consideri il problema del propellente dell'Esempio 4.3. Si supponga che l'analista desideri pianificare il test in modo che se la vera velocità media di combustione differisce dal valore 50 cm/s per 1 cm/s, il test sia in grado di rilevare tale differenza (ovvero rifiutare  $H_0: \mu = 50$ ) con un'alta probabilità, per esempio 0.90. Ora, notiamo che  $\sigma = 2$ ,  $\delta = 51 - 50 = 1$ ,  $\alpha = 0.05$  e  $\beta = 0.10$ . Poiché  $z_{\alpha/2} = z_{0.025} = 1.96$  e  $z_\beta = z_{0.10} = 1.28$ , la dimensione campionaria necessaria per rilevare questo scostamento da  $H_0: \mu = 50$  risulta, per l'Equazione (4.26)

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\delta^2} = \frac{(1.96 + 1.28)^2 2^2}{(1)^2} \approx 42$$

L'approssimazione è in questo caso buona, perché  $\Phi(-z_{\alpha/2} - \delta \sqrt{n}/\sigma) = \Phi(-1.96 - (1)\sqrt{42}/2) =$  che è piccolo in confronto a  $\beta$ .

#### Calcolo della dimensione campionaria e della potenza mediante Minitab.

Molti software statistici calcolano le dimensioni campionarie e le probabilità degli errori del II tipo. A titolo di illustrazione, la Tabella 4.2 mostra alcuni output di Minitab per il problema della velocità di combustione.

Nella prima parte della tabella abbiamo chiesto a Minitab di lavorare sull'Esempio 4.4, cioè di trovare la dimensione campionaria necessaria per rilevare uno scostamento di 1 cm/s da  $\mu_0 = 50$  cm/s con potenza pari a 0.9 e  $\alpha = 0.05$ . La risposta ottenuta,  $n = 43$ , è in stretto accordo con quella calcolata tramite l'Equazione (4.26) nell'Esempio 4.4, che era 42. La differenza è dovuta al fatto che Minitab usa un valore di  $z_\beta$  che ha più di due cifre decimali. La seconda parte dell'output allenta il requisito sulla potenza a 0.75. Si noti che l'effetto è di ridurre la dimensione campionaria a  $n = 28$ . La terza parte dell'output rappresenta la situazione dell'Esempio 4.4, ma ora si vuole determinare la probabilità dell'errore di II tipo  $\beta$  o la potenza  $1 - \beta$  per la dimensione campionaria  $n = 25$ .

**Tabella 4.2** Calcoli eseguiti con Minitab**1-Sample Z Test**

Testing mean = null (versus not = null)

Calculating power for mean = null + difference

Alpha = 0.05 Sigma = 2

	Sample	Target	Actual
Difference	Size	Power	Power
1	43	0.9000	0.9064

**1-Sample Z Test**

Testing mean = null (versus not = null)

Calculating power for mean = null + difference

Alpha = 0.05 Sigma = 2

	Sample	Target	Actual
Difference	Size	Power	Power
1	28	0.7500	0.7536

**1-Sample Z Test**

Testing mean = null (versus not = null)

Calculating power for mean = null + difference

Alpha = 0.05 Sigma = 2

	Sample		
Difference	Size		Power
1	25	0.7054	

**4.4.3 Test con campioni numerosi**

Anche se abbiamo sviluppato la procedura di verifica per l'ipotesi nulla  $H_0: \mu = \mu_0$  assumendo che  $\sigma^2$  fosse nota, in molte situazioni pratiche, se non nella maggior parte,  $\sigma^2$  sarà incognita. In generale, se  $n \geq 30$ , la varianza campionaria  $s^2$  sarà prossima a  $\sigma^2$  per la maggior parte dei campioni, per cui si potrà sostituire  $s$  al posto di  $\sigma$  nella procedura di verifica senza effetti dannosi. Perciò, anche se abbiamo fornito un test per  $\sigma^2$  nota, esso può venire facilmente convertito in una *procedura di verifica per campioni numerosi con  $\sigma^2$  incognita*. La trattazione rigorosa del caso in cui  $\sigma^2$  è incognita e  $n$  è piccola richiede l'uso della *distribuzione t* e viene rinviata al Paragrafo 4.5.

**4.4.4 Considerazioni pratiche sulla verifica di ipotesi****La procedura a sette passi**

Nel Paragrafo 4.3.5 abbiamo descritto una procedura a sette passi per la verifica delle ipotesi statistiche. Tale procedura è stata adottata nell'Esempio 4.3 e verrà utilizzata molte altre volte in questo capitolo. In pratica, però, una procedura così formale e (apparentemente) rigida non è sempre indispensabile. In generale, una volta che lo sperimentatore (o il responsabile delle decisioni) ha deciso qual è il problema di interesse e ha stabilito il **piano dell'esperimento**, la procedura a sette passi non è più necessaria.

**sperimento** (cioè come devono essere raccolti i dati, come devono essere effettuate le misure e quante osservazioni sono necessarie), servono effettivamente soltanto tre passi:

1. Specificare le ipotesi (a due code, a coda superiore o inferiore).
2. Specificare la statistica test da adottare (come  $z_0$ ).
3. Specificare i criteri per il rifiuto (tipicamente, il valore di  $\alpha$ , o il  $P$ -value al quale si dovrebbe operare il rifiuto).

Questi passi vengono spesso completati simultaneamente nella risoluzione dei problemi concreti, anche se si sottolinea che è importante ragionare attentamente su ciascuna fase. Ecco perché presentiamo e usiamo la procedura a sette passi: è utile per consolidare le basi di un approccio corretto. Anche se non la si usa in ogni risoluzione di problemi concreti, serve da utile quadro di riferimento per cominciare a imparare la verifica di ipotesi.

#### Significatività statistica e significatività pratica

Abbiamo osservato in precedenza che riportare i risultati di una verifica di ipotesi in termini di  $P$ -value è molto utile, in quanto tale valore contiene più informazioni delle semplici asserzioni "rifiuto di  $H_0$ " o "mancato rifiuto di  $H_0$ ". In altri termini, il rifiuto di  $H_0$  al livello 0.05 di significatività ha un significato molto maggiore se il valore della statistica test è ben addentro alla regione critica, superando di molto il valore critico 5%, che non se supera a malapena tale valore.

Anche un  $P$ -value molto piccolo può essere difficile da interpretare da un punto di vista pratico, quando si deve assumere una decisione; anche se esso indica comunque una significatività statistica, nel senso che  $H_0$  dovrebbe venire rifiutata a favore di  $H_1$ , l'effettivo scostamento da  $H_0$  rilevato può avere una significatività pratica scarsa, se non addirittura nulla (agli ingegneri piace dire significatività "ingegneristica" anziché "pratica"). Ciò è vero soprattutto quando la dimensione campionaria è alta.

Per esempio, si consideri il problema della velocità di combustione dell'Esempio 4.3, dove si sottopone a verifica  $H_0: \mu = 50$  cm/s contro  $H_1: \mu \neq 50$  cm/s con  $\sigma = 2$ . Se supponiamo che la vera velocità media sia 50.5 cm/s, non si tratta di uno scostamento sensibile dall'ipotesi nulla, nel senso che in questo caso non c'è un effetto pratico osservabile sul comportamento del sistema di espulsione. Detto in altro modo, concludere che  $\mu = 50$  cm/s quando in effetti è 50.5 cm/s costituisce un errore trascurabile ai fini pratici. Per una dimensione campionaria abbastanza alta un valore di 50.5 porterà a una media campionaria  $\bar{x}$  vicina a 50.5 cm/s, e non vorremmo che questo valore di  $\bar{x}$  portasse al rifiuto di  $H_0$ . La seguente tabella mostra il  $P$ -value per la verifica di  $H_0: \mu = 50$  quando si osserva  $\bar{x} = 50.5$  cm/s e la potenza del test per  $\alpha = 0.05$  quando la vera media è 50.5, per diverse dimensioni campionarie  $n$ .

Relazione fra potenza e dimensione campionaria per il test z.

I  $P$ -value diminuiscono al crescere della dimensione campionaria per un valore di  $\bar{x}$  fissato.

Dimensione campionaria $n$	$P$ -value quando $\mu = 50.5$	Potenza (per $\alpha = 0.05$ ) quando $\mu = 50.5$
10	0.4295	0.1241
25	0.2113	0.2396
50	0.0767	0.4239
100	0.0124	0.7054
400		0.9988
1000		1.0000

La colonna con il *P*-value indica che per dimensioni campionarie elevate il valore di  $\bar{x}$  per il campione osservato suggerirebbe con forza il rifiuto di  $H_0: \mu = 50$ , anche se i risultati del campione osservato implicano che da un punto di vista pratico la vera media non è molto diversa dal valore ipotizzato, 50. La colonna della potenza indica che se verifichiamo un'ipotesi a un dato livello di significatività  $\alpha$ , e anche se c'è una piccola differenza pratica fra la vera media e quella ipotizzata, una dimensione campionaria elevata porterà quasi sempre al rifiuto di  $H_0$ . La morale di questa dimostrazione è chiara:

Occorre prestare attenzione a interpretare i risultati della verifica di ipotesi quando la dimensione campionaria è alta, perché ogni piccolo scostamento dal valore  $\mu_0$  ipotizzato verrà verosimilmente rilevato, anche quando la differenza ha poca o nulla significatività pratica.

#### 4.4.5 Intervallo di confidenza per la media

In molte situazioni, una stima puntuale non fornisce abbastanza informazioni su di un parametro. Per esempio, nel problema del propellente abbiamo rifiutato l'ipotesi nulla  $H_0: \mu = 50$ , e la nostra stima puntuale della velocità media di combustione è  $\bar{x} = 51.3$  cm/s. Tuttavia, gli ingegneri preferirebbero avere un **intervallo** in cui ci si aspetterebbe di trovare la vera velocità media di combustione, perché  $\mu = 51.3$  è un valore non verosimile. Un modo per ottenerne questo è di utilizzare una stima intervallare detta intervallo di confidenza (*CI, Confidence Interval*).

Una stima intervallare del parametro incognito  $\mu$  è un intervallo della forma  $l \leq \mu \leq u$ , dove gli estremi  $l$  e  $u$  dipendono dal valore numerico della media campionaria  $\bar{X}$  per un particolare campione. Dato che diversi campioni produrranno diversi valori di  $\bar{x}$  e, perciò, diversi valori degli estremi dell'intervallo, questi ultimi sono valori di variabili aleatorie, per esempio  $L$  e  $U$  rispettivamente. A partire dalla distribuzione campionaria della media campionaria  $\bar{X}$  saremo in grado di determinare valori di  $L$  e  $U$  per i quali è valida la seguente asserzione probabilistica

$$\underline{P(L \leq \mu \leq U) = 1 - \alpha} \quad (4.28)$$

dove  $0 < \alpha < 1$ . Pertanto, abbiamo una probabilità  $1 - \alpha$  di selezionare un campione che produrrà un intervallo contenente il vero valore di  $\mu$ .

L'intervallo risultante

$$l \leq \mu \leq u, \quad (4.29)$$

viene chiamato intervallo di confidenza al  $100(1 - \alpha)\%$  per il parametro  $\mu$ . Le quantità  $l$  e  $u$  sono dette rispettivamente **limiti di confidenza inferiore e superiore**, e  $1 - \alpha$  è detto **livello di confidenza o coefficiente di confidenza**. L'interpretazione di un intervallo di confidenza è che, se si è raccolto un numero infinito di campioni casuali e si è calcolato un intervallo di confidenza al  $100(1 - \alpha)\%$  per  $\mu$  a partire da ciascun campione, allora il  $100(1 - \alpha)\%$  di tali intervalli conterrà il vero valore di  $\mu$ .

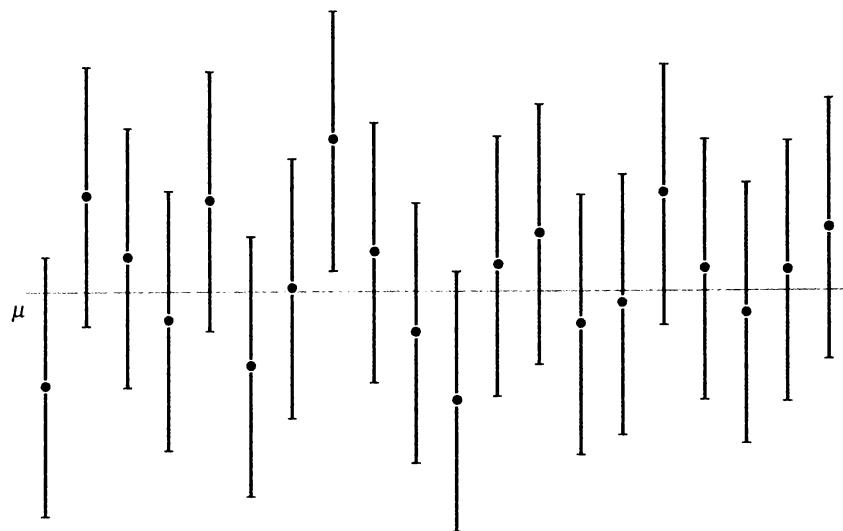


Figura 4.12 Costruzione ripetuta di un intervallo di confidenza per  $\mu$ .

La situazione è illustrata in Figura 4.12, che mostra diversi intervalli di confidenza al  $100(1 - \alpha)\%$  per la media  $\mu$  di una distribuzione. I punti al centro di ogni intervallo indicano la stima puntuale di  $\mu$  (cioè  $\bar{x}$ ). Si noti che 1 dei 20 intervalli non contiene il vero valore di  $\mu$ ; se questo fosse un intervallo di confidenza al 95%, nel lungo periodo solo il 5% degli intervalli non conterrebbe  $\mu$ .

Ora, in pratica, noi otteniamo solo un campione casuale e calcoliamo un intervallo di confidenza. Dato che questo intervallo conterrà oppure non conterrà il vero valore di  $\mu$ , non è ragionevole associare un livello di probabilità a questo specifico evento. L'asserzione appropriata è che l'intervallo  $[l, u]$  osservato contiene il vero valore di  $\mu$  con una confidenza  $100(1 - \alpha)$ . Tale asserzione ha una interpretazione frequentista: non sappiamo se è vera per questo specifico campione, ma il metodo usato per ottenere l'intervallo  $[l, u]$  porta ad asserzioni corrette il  $100(1 - \alpha)\%$  delle volte.

L'intervallo di confidenza dell'Equazione (4.29) viene chiamato più propriamente un **intervallo di confidenza bilaterale**, perché specifica sia un limite inferiore che un limite superiore su  $\mu$ . A volte potrebbe essere più appropriato un **limite di confidenza unilaterale**. Un limite inferiore di confidenza unilaterale al  $100(1 - \alpha)\%$  su  $\mu$  è dato da

$$l \leq \mu \quad (4.30)$$

dove il limite inferiore di confidenza  $l$  è scelto in modo che

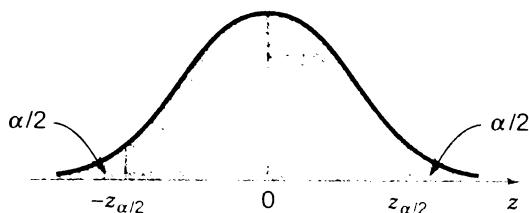
$$P(L \leq \mu) = 1 - \alpha \quad (4.31)$$

Analogamente, il limite superiore di confidenza unilaterale al  $100(1 - \alpha)\%$  su  $\mu$  è dato da

$$\mu \leq u \quad (4.32)$$

dove il limite superiore di confidenza  $u$  è scelto in modo che

$$P(\mu \leq U) = 1 - \alpha \quad (4.33)$$

Figura 4.13 La distribuzione di  $Z$ .

La lunghezza  $u - l$  dell'intervallo di confidenza bilaterale osservato è un'importante misura della qualità dell'informazione ottenuta dal campione. La semilunghezza dell'intervallo,  $\mu - l$  o  $u - \mu$ , viene detta **precisione** dello stimatore. Più è ampio l'intervallo di confidenza, più siamo confidenti nel fatto che l'intervallo contenga effettivamente il vero valore di  $\mu$ . Per contro, più è ampio l'intervallo, meno informazioni abbiamo sul vero valore di  $\mu$ . In una situazione ideale dovremmo ottenere un intervallo relativamente corto con un alto livello di confidenza.

È molto facile trovare le quantità  $L$  e  $U$  che definiscono l'intervallo di confidenza bilaterale per  $\mu$ . Sappiamo che la distribuzione campionaria di  $\bar{X}$  è normale con media  $\mu$  e varianza  $\sigma^2/n$ . Perciò la statistica

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

ha una distribuzione normale standard.

La distribuzione di  $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$  è mostrata in Figura 4.13. Esaminando questa figura si vede che

$$\underline{P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha}$$

per cui

$$\underline{P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha}$$

Si può riscrivere questa espressione come

$$P\left(\bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right) = 1 - \alpha \quad (4.34)$$

Considerando l'Equazione (4.28), i limiti inferiore e superiore delle diseguaglianze nell'Equazione (4.34) sono i limiti inferiore e superiore di confidenza  $L$  e  $U$ , rispettivamente. Ciò porta alla seguente definizione.

### Intervallo di confidenza per la media, varianza nota

Se  $\bar{x}$  è la media campionaria di un campione casuale di dimensione  $n$  estratto da una popolazione con varianza nota  $\sigma^2$ , un **intervallo di confidenza al 100(1 –  $\alpha$ )%** per  $\mu$  è dato da

$$\bar{x} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \quad (4.35)$$

dove  $z_{\alpha/2}$  è il punto percentuale superiore  $100\alpha/2$  e  $-z_{\alpha/2}$  è il punto percentuale inferiore  $100\alpha/2$  della distribuzione normale standard (Appendice A, Tavola I).

Per campioni estratti da una popolazione normale o per campioni di dimensione  $n \geq 40$  e indipendentemente dalla forma della popolazione, l'intervallo di confidenza dell'Equazione (4.35) fornirà ottimi risultati. Tuttavia, per piccoli campioni estratti da popolazioni non normali non possiamo aspettarci che il livello di confidenza  $1 - \alpha$  sia esatto.

#### ESEMPIO 4.5 Velocità di combustione del propellente

Si consideri il problema del propellente dell'Esempio 4.3. Si supponga di voler trovare un intervallo di confidenza al 95% sulla velocità media di combustione. Possiamo usare l'Equazione (4.35) per costruire l'intervallo di confidenza. Un intervallo al 95% significa che  $1 - \alpha = 0.95$ , per cui  $\alpha = 0.05$  e, dalla Tavola I dell'Appendice A,  $z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$ .

Il limite inferiore di confidenza è

$$\begin{aligned} l &= \bar{x} - z_{\alpha/2}\sigma/\sqrt{n} \\ &= 51.3 - 1.96(2)/\sqrt{25} \\ &= 51.3 - 0.78 \\ &= 50.52 \end{aligned}$$

e quello superiore è

$$\begin{aligned} u &= \bar{x} + z_{\alpha/2}\sigma/\sqrt{n} \\ &= 51.3 + 1.96(2)/\sqrt{25} \\ &= 51.3 + 0.78 \\ &= 52.08 \end{aligned}$$

perciò l'intervallo di confidenza bilaterale al 95% è

$$50.52 \leq \mu \leq 52.08$$

Si ricordi di interpretare l'intervallo di confidenza; questo specifico intervallo contiene  $\mu$  oppure non lo contiene (e noi non sappiamo in quale caso ci troviamo), ma grazie alla procedura usata per costruire l'intervallo di confidenza, in ripetuti campionamenti il 95% degli intervalli che calcoleremo conterrà il vero valore di  $\mu$ . Questo intervallo di confidenza è presente anche nell'output di Minitab presentato nel Paragrafo 4.4.1.

### Relazioni fra verifiche di ipotesi e intervalli di confidenza

Esiste una stretta relazione fra la verifica di un'ipotesi su un qualsiasi parametro  $\theta$  e l'intervalle di confidenza per  $\theta$ . Se  $[l, u]$  è un intervallo di confidenza al  $100(1 - \alpha)\%$  per il parametro  $\theta$ , il test di livello di significatività  $\alpha$  delle ipotesi

$$H_0: \theta = \theta_0$$

$$H_1: \theta \neq \theta_0$$

porterà al rifiuto di  $H_0$  se e solo se  $\theta_0$  non è nell'intervalle di confidenza al  $100(1 - \alpha)\%$ ,  $[l, u]$ . A titolo esemplificativo, si consideri il solito problema del propellente più volte discusso. L'ipotesi nulla  $H_0: \mu = 50$  è stata rifiutata, usando  $\alpha = 0.05$ . L'intervalle di confidenza bilaterale al 95% su  $\mu$  è  $50.52 \leq \mu \leq 52.08$ . Cioè, l'intervalle  $[l, u]$  è  $[50.52, 52.08]$ , e poiché  $\mu_0 = 50$  non appartiene a tale intervallo, l'ipotesi nulla viene rifiutata.

### Livelli di confidenza e precisione della stima

Si noti, nel precedente esempio, che la nostra scelta del livello di confidenza 95% era sostanzialmente arbitraria. Che cosa sarebbe successo se avessimo scelto un livello di confidenza più alto, per esempio 99%? In effetti, non è ragionevole richiedere il livello di confidenza più alto? Per  $\alpha = 0.01$ , troviamo  $z_{\alpha/2} = z_{0.01/2} = z_{0.005} = 2.58$ , mentre per  $\alpha = 0.05$  troviamo  $z_{0.025/2} = 1.96$ . Pertanto, l'ampiezza dell'intervalle di confidenza al 95% è

$$2(1.96 \sigma/\sqrt{n}) = 3.92 \sigma/\sqrt{n}$$

mentre l'ampiezza dell'intervalle di confidenza al 99% è

$$2(2.58 \sigma/\sqrt{n}) = 5.16 \sigma/\sqrt{n}$$

L'intervalle di confidenza al 99% è più ampio di quello al 95%, e ciò spiega perché abbiamo un livello di confidenza maggiore nel primo intervallo di confidenza. In generale, per una fissata dimensione campionaria  $n$  e una data deviazione standard  $\sigma$ , più è alto il livello di confidenza, più è ampio l'intervalle di confidenza risultante.

Poiché la semiampiezza dell'intervalle di confidenza misura la precisione della stima, vediamo che la precisione è inversamente proporzionale al livello di confidenza. Come abbiamo già osservato, è desiderabile ottenere un intervallo di confidenza che sia abbastanza corto per gli scopi decisionali e abbia al contempo un'adeguata confidenza. Un modo per raggiungere questo obiettivo è scegliere la dimensione campionaria  $n$  abbastanza alta da dare un intervallo di confidenza di ampiezza specificata con confidenza stabilita.

In molte situazioni pratiche il livello di confidenza scelto è il 95%. Si tratta di un compromesso spesso ragionevole fra la precisione della stima e la confidenza (ossia l'affidabilità della procedura). È raro vedere intervalli di confidenza più piccoli del 90% o maggiori del 99.5%.

### Scelta della dimensione campionaria

La precisione dell'intervalle di confidenza nell'Equazione (4.35) è  $z_{\alpha/2}\sigma/\sqrt{n}$ . Ciò significa che usando  $\bar{x}$  per stimare  $\mu$ , l'errore  $E = |\bar{x} - \mu|$  è minore o uguale a  $z_{\alpha/2}\sigma/\sqrt{n}$  con confidenza  $100(1 - \alpha)$ , come mostra graficamente la Figura 4.14. In situazioni in cui la dimen-

Gli intervalli di confidenza si allargano se il livello di confidenza aumenta e si accorciano se il livello di confidenza diminuisce.

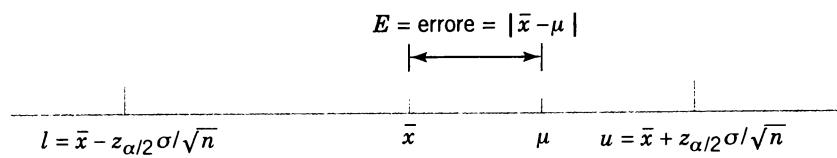


Figura 4.14 Errore nella stima di  $\mu$  con  $\bar{x}$

sione campionaria può essere controllata possiamo scegliere  $n$  in modo da essere confidenti al  $100(1 - \alpha)\%$  che l'errore nella stima di  $\mu$  sia minore di un errore specificato  $E$ . L'appropriata dimensione campionaria si trova scegliendo  $n$  tale per cui  $z_{\alpha/2}\sigma/\sqrt{n} = E$ . La soluzione di questa equazione dà la seguente formula per  $n$ .

#### **Dimensione campionaria per un errore E specificato sulla media, varianza nota**

Se si usa  $\bar{x}$  come stima di  $\mu$ , si può essere confidenti al  $100(1 - \alpha)\%$  che l'errore  $|\bar{x} - \mu|$  non supererà un valore specificato  $E$  quando la dimensione campionaria è

$$n = \left( \frac{z_{\alpha/2}\sigma}{E} \right)^2 \quad (4.36)$$

Se il membro di destra dell'Equazione (4.36) non è un intero, deve essere arrotondato all'intero successivo, il che garantirà che il livello di confidenza non scenda al di sotto del  $100(1 - \alpha)\%$ . Si noti che  $2E$  è l'ampiezza dell'intervallo di confidenza risultante.

#### **ESEMPIO 4.6** Velocità di combustione del propellente

Per mostrare l'uso di questa procedura, si supponga di volere che l'errore nella stima della velocità media di combustione del propellente (ci riferiamo all'esempio sviluppato in tutto il capitolo) sia inferiore a 1.5 cm/s, con il 95% di confidenza. Poiché  $\sigma = 2$  e  $z_{0.025} = 1.96$ , possiamo trovare la dimensione campionaria richiesta con l'Equazione (4.36)

$$n = \left( \frac{z_{\alpha/2}\sigma}{E} \right)^2 = \left[ \frac{(1.96)2}{1.5} \right]^2 = 6.83 \approx 7$$

Si noti la relazione generale fra dimensione campionaria, ampiezza voluta dell'intervallo di confidenza ( $2E$ ), livello di confidenza al  $100(1 - \alpha)\%$  e deviazione standard  $\sigma$ :

- al diminuire dell'ampiezza voluta dell'intervallo di confidenza ( $2E$ ), la dimensione campionaria  $n$  richiesta aumenta per un valore di  $\sigma$  fissato e una specifica confidenza;
- all'aumentare di  $\sigma$ , la dimensione campionaria  $n$  richiesta aumenta per un'ampiezza  $2E$  fissata e una specifica confidenza;
- all'aumentare del livello di confidenza, la dimensione campionaria  $n$  richiesta aumenta per un'ampiezza  $2E$  fissata e una specifica deviazione standard  $\sigma$ .

### Limiti di confidenza unilaterali

È possibile anche ottenere limiti unilaterali di confidenza per  $\mu$  ponendo  $l = -\infty$  oppure  $u = \infty$  e sostituendo  $z_{\alpha/2}$  con  $z_\alpha$ .

Come abbiamo visto per il caso bilaterale, possiamo anche usare il limite unilaterale di confidenza per eseguire verifiche di ipotesi con un'ipotesi alternativa unilaterale. Precisamente, se  $u$  è il limite superiore di un intervallo di confidenza unilaterale al  $100(1 - \alpha)\%$  per il parametro  $\theta$ , il test di livello di significatività  $\alpha$  delle ipotesi

$$\begin{aligned} H_0: \theta &= \theta_0 \\ H_1: \theta &< \theta_0 \end{aligned}$$

porterà al rifiuto di  $H_0$  se e solo se  $\theta_0 > u$ . Analogamente, se  $l$  è il limite inferiore di un intervallo di confidenza unilaterale al  $100(1 - \alpha)\%$ , il test di livello di significatività  $\alpha$  delle ipotesi

$$\begin{aligned} H_0: \theta &= \theta_0 \\ H_1: \theta &> \theta_0 \end{aligned}$$

porterà al rifiuto se e solo se  $\theta_0 < l$ .

Poiché  $z_\alpha$  è sempre minore di  $z_{\alpha/2}$ , il limite inferiore di confidenza al  $100(1 - \alpha)\%$  unilaterale sarà sempre maggiore del limite inferiore dell'intervallo di confidenza bilaterale al  $100(1 - \alpha)\%$ , e il limite superiore di confidenza al  $100(1 - \alpha)\%$  unilaterale sarà sempre minore del limite superiore dell'intervallo di confidenza bilaterale al  $100(1 - \alpha)\%$ . Di conseguenza, se si rifiuta l'ipotesi  $H_0: \mu = \mu_0$  con un'alternativa bilaterale, a maggior ragione la si rifiuta con un'alternativa unilaterale.

### Limiti unilaterali di confidenza per la media, varianza nota

**Il limite superiore di confidenza al  $100(1 - \alpha)\%$   $u$  per  $\mu$  è**

$$\mu \leq u = \bar{x} + z_\alpha \sigma / \sqrt{n} \quad (4.37)$$

**e il limite inferiore di confidenza al  $100(1 - \alpha)\%$   $l$  per  $\mu$  è**

$$\bar{x} - z_\alpha \sigma / \sqrt{n} = l \leq \mu \quad (4.38)$$

#### 4.4.6 Metodo generale per ricavare un intervallo di confidenza

È semplice fornire un metodo generale per trovare un intervallo di confidenza per un parametro incognito  $\theta$ . Sia  $X_1, X_2, \dots, X_n$  un campione casuale di  $n$  osservazioni. Si supponga di poter trovare una variabile aleatoria  $g(X_1, X_2, \dots, X_n; \theta)$  con le seguenti proprietà:

1.  $g(X_1, X_2, \dots, X_n; \theta)$  dipende sia dal campione, sia da  $\theta$ ;
2. la distribuzione di probabilità di  $g(X_1, X_2, \dots, X_n; \theta)$  non dipende da  $\theta$  o da altri parametri non noti.

Nel caso preso in esame in questo paragrafo il parametro è  $\theta = \mu$ . La variabile aleatoria è  $g(X_1, X_2, \dots, X_n; \mu) = (\bar{X} - \mu) / (\sigma / \sqrt{n})$  e soddisfa entrambe le condizioni: dipende dal

campione e da  $\mu$  e ha una distribuzione normale standard perché  $\sigma$  è nota. Ora si devono trovare le costanti  $C_L$  e  $C_U$  tali che

$$P[C_L \leq g(X_1, X_2, \dots, X_n; \theta) \leq C_U] = 1 - \alpha$$

Per via della Proprietà 2,  $C_L$  e  $C_U$  non dipendono da  $\theta$ . Nel nostro esempio,  $C_L = -z_{\alpha/2}$  e  $C_U = z_{\alpha/2}$ . Infine, bisogna portare le disuguaglianze nell'asserzione probabilistica nella forma

$$P[L(X_1, X_2, \dots, X_n) \leq \theta \leq U(X_1, X_2, \dots, X_n)] = 1 - \alpha$$

Si ottengono così  $L(X_1, X_2, \dots, X_n)$  e  $U(X_1, X_2, \dots, X_n)$  come limiti inferiore e superiore di confidenza che definiscono l'intervallo di confidenza al  $100(1 - \alpha)\%$  per  $\theta$ . Per il nostro esempio troviamo  $L(X_1, X_2, \dots, X_n) = \bar{X} - z_{\alpha/2}\sigma/\sqrt{n}$  e  $U(X_1, X_2, \dots, X_n) = \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}$ .

## 4.5 INFERENZA SULLA MEDIA DI UNA POPOLAZIONE CON VARIANZA INCOGNITA

Quando si effettuano verifiche di ipotesi o si costruiscono intervalli di confidenza sulla media  $\mu$  di una popolazione con  $\sigma^2$  nota, si possono usare le procedure di verifica viste nel Paragrafo 4.4, purché la dimensione campionaria sia elevata (per esempio maggiore di 40). Tali procedure sono approssimativamente valide (per via del teorema limite centrale) indipendentemente dal fatto che la sottostante popolazione sia o meno normale. Tuttavia, quando il campione è poco numeroso e non si conosce  $\sigma^2$ , si deve fare un'assunzione sulla forma della distribuzione, per ottenere una procedura di verifica. In molti casi un'assunzione ragionevole è che la distribuzione sottostante sia normale.

Molte popolazioni che si incontrano nella pratica sono bene approssimate dalla distribuzione normale, perciò questa assunzione porterà a procedure di inferenza di ampia applicabilità. In effetti, uno scostamento moderato dalla normalità avrà scarsi effetti sulla validità dei risultati. Quando invece l'assunzione non è ragionevole un'alternativa consiste nell'adottare procedure non parametriche che siano valide per ogni distribuzione sottostante. Si veda Montgomery, Runger (2011) per un'introduzione a queste ultime tecniche.

### 4.5.1 Verifica di ipotesi sulla media

Si supponga che la popolazione di interesse abbia distribuzione normale con media  $\mu$  e varianza  $\sigma^2$  incognite. Vogliamo verificare l'ipotesi che  $\mu$  sia uguale a una costante  $\mu_0$ . Si noti che questa situazione è simile a quella del Paragrafo 4.4, eccetto per il fatto che ora non si conoscono né  $\mu$  né  $\sigma^2$ . Si assuma che sia disponibile un campione casuale di dimensione  $n$ , per esempio  $X_1, X_2, \dots, X_n$ , e siano  $\bar{X}$  e  $S^2$ , rispettivamente, la media e la varianza campionarie.

Vogliamo sottoporre a verifica l'ipotesi alternativa bilaterale

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

Se la varianza  $\sigma^2$  è nota, la statistica test è l'Equazione (4.13)

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Quando  $\sigma^2$  non è nota, una procedura ragionevole prevede di sostituire  $\sigma$  nell'espressione precedente con la deviazione standard campionaria  $S$ . La statistica test è ora

$$T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \quad (4.39)$$

Una domanda ovvia è: che effetto ha la sostituzione di  $\sigma$  con  $S$  sulla distribuzione della statistica  $T_0$ ? Se  $n$  è grande, la risposta a tale domanda è: molto piccolo, e possiamo usare la procedura di verifica basata sulla distribuzione normale (Paragrafo 4.4). Tuttavia,  $n$  è di solito piccola nella maggior parte dei problemi di ingegneria, e in tal caso va impiegata una distribuzione differente.

**Sia  $X_1, X_2, \dots, X_n$  un campione casuale estratto da una distribuzione normale con media  $\mu$  e varianza  $\sigma^2$  incognite. La quantità**

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

**ha una distribuzione  $t$  con  $n - 1$  gradi di libertà.**

La funzione di densità di probabilità  $t$  è

**Proprietà della distribuzione  $t$ .**

$$f(x) = \frac{\Gamma[(k + 1)/2]}{\sqrt{\pi k} \Gamma(k/2)} \cdot \frac{1}{[(x^2/k) + 1]^{(k+1)/2}} \quad -\infty < x < \infty \quad (4.40)$$

dove  $k$  è il numero di gradi di libertà. La media e la varianza della distribuzione  $t$  sono rispettivamente 0 e  $k/(k - 2)$ , per  $k > 2$ . La funzione  $\Gamma(m) = \int_0^\infty e^{-x} x^{m-1} dx$  è la funzione gamma, introdotta nel Paragrafo 3.5.3. Anche se è definita per  $m \geq 0$ , nel caso particolare in cui  $m$  sia un intero si ha  $\Gamma(m) = (m - 1)!$ . Inoltre,  $\Gamma(1) = \Gamma(0) = 1$ .

In Figura 4.15 sono mostrate diverse distribuzioni  $t$ . L'aspetto generale della distribuzione  $t$  è simile a quello della distribuzione normale standard, in quanto entrambe sono simmetriche e unimodali, e il massimo valore dell'ordinata è raggiunto per  $x = 0$ . Tuttavia, la distribuzione  $t$  ha code più alte di quella normale, ossia ha più probabilità nelle code della distribuzione normale. Al tendere del numero di gradi di libertà all'infinito, la forma limite della distribuzione  $t$  è la distribuzione normale standard. Quando si visualizza la distribuzione  $t$ , è utile a volte sapere che l'ordinata della densità in corrispondenza della media  $x = 0$  è approssimativamente 4 o 5 volte maggiore dell'ordinata al quinto e 95-esimo percentile. Per esem-

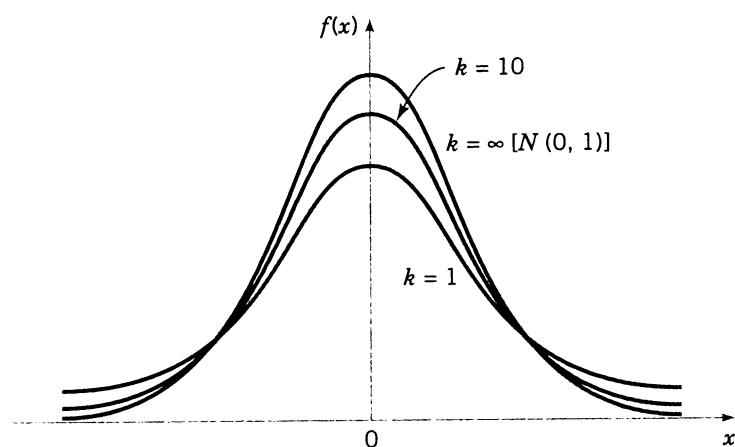


Figura 4.15 Funzioni di densità di probabilità di alcune distribuzioni  $t$ .

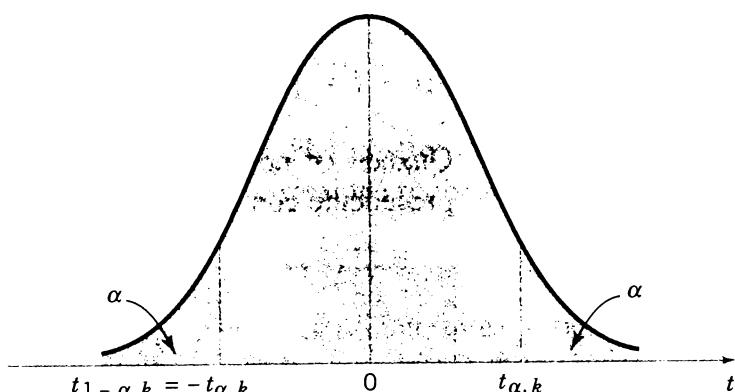


Figura 4.16 Punti percentuali della distribuzione  $t$ .

pio, con 10 gradi di libertà per  $t$  questo rapporto vale 4.8, con 20 vale 4.3 e con 30 vale 4.1, valori da confrontare con l'analogo fattore della distribuzione normale, pari a 3.9.

La Tavola II dell'Appendice A fornisce i **punti percentuali** della distribuzione  $t$ . Porremo  $t_{\alpha, k}$  come valore della variabile aleatoria  $T$  con  $k$  gradi di libertà a destra del quale troviamo un'area (o una probabilità)  $\alpha$ . Perciò,  $t_{\alpha, k}$  è un  $100\alpha$ -esimo punto percentuale della coda superiore della distribuzione  $t$  con  $k$  gradi di libertà. Questo punto percentuale è mostrato in Figura 4.16. Nella Tavola II dell'Appendice A i valori  $\alpha$  sono le intestazioni di colonna, i gradi di libertà sono elencati nella colonna a sinistra. Per illustrare l'uso di questa tavola si noti che il valore  $t$  con 10 gradi di libertà avente un'area di 0.05 alla destra è  $t_{0.05, 10} = 1.812$ , cioè:

$$P(T_{10} > t_{0.05, 10}) = P(T_{10} > 1.812) = 0.05$$

Poiché la distribuzione  $t$  è simmetrica intorno allo 0, abbiamo  $t_{1-\alpha} = -t_\alpha$ , cioè il valore  $t$  avente alla sua destra un'area pari a  $1 - \alpha$  (e, quindi, un'area  $\alpha$  alla sua sinistra) è uguale all'opposto del valore  $t$  avente area  $\alpha$  nella coda di destra della distribuzione. Pertanto,  $t_{0.95, 10} = -t_{0.05, 10} = -1.812$ .

È ora immediato vedere che la distribuzione della statistica test dell'Equazione (4.39) è  $t$  con  $n - 1$  gradi di libertà se l'ipotesi nulla  $H_0: \mu = \mu_0$  è vera. La procedura è detta **test  $t$** . Per verificare tale ipotesi nulla rispetto all'alternativa bilaterale  $H_1: \mu \neq \mu_0$ , viene calcolato il valore della statistica test,  $t_0$ , nell'Equazione (4.39) e si trova il  $P$ -value dalla distribuzione  $t$  con  $n - 1$  gradi di libertà.

Poiché il test è bilaterale, il  $P$ -value è la somma delle probabilità nelle due code della distribuzione  $t$  (Figura 4.17a). Dunque, se la statistica test è positiva il  $P$ -value è la probabilità a destra del valore della statistica  $t_0$  più la probabilità a sinistra dell'opposto di  $t_0$ ; se invece la statistica test è negativa il  $P$ -value è la probabilità a sinistra del valore della statistica  $t_0$  più la probabilità a destra del suo valore assoluto,  $|t_0| = -t_0$ . Essendo la distribuzione simmetrica intorno allo zero, si possono riassumere questi fatti con la scrittura

$$P = 2P(T_{n-1} > |t_0|) \quad (4.41)$$

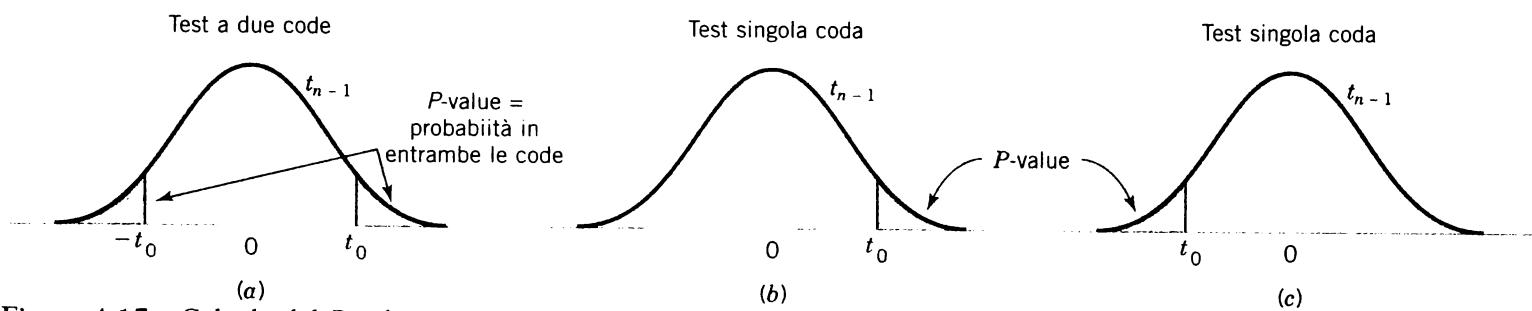


Figura 4.17 Calcolo del  $P$ -value per un test  $t$ : (a)  $H_1: \mu \neq \mu_0$ ; (b)  $H_1: \mu > \mu_0$ ; (c)  $H_1: \mu < \mu_0$

Un basso  $P$ -value è un indizio contro  $H_0$ , perciò se  $P$  è sufficientemente piccolo (tipicamente minore di 0.05) si dovrebbe rifiutare l'ipotesi nulla.

Per l'ipotesi alternativa unilaterale

$$\begin{aligned} H_0: \mu &= \mu_0 \\ H_1: \mu &> \mu_0 \end{aligned} \quad (4.42)$$

calcoliamo la statistica test  $t_0$  dall'Equazione (4.39) e calcoliamo il  $P$ -value come

$$P = P(T_{n-1} > t_0) \quad (4.43)$$

Per l'altra ipotesi alternativa unilaterale

$$\begin{aligned} H_0: \mu &= \mu_0 \\ H_1: \mu &< \mu_0 \end{aligned} \quad (4.44)$$

calcoliamo il  $P$ -value come

$$P = P(T_{n-1} < t_0) \quad (4.45)$$

La Figura 4.17b e c illustrano i calcoli per i  $P$ -value.

I software statistici calcolano e visualizzano i  $P$ -value. Tuttavia, se si risolvono i problemi a mano è utile essere in grado di trovare il  $P$ -value per un test  $t$ . Poiché la Tavola II, Appendice A, contiene solo 10 valori critici per ogni distribuzione  $t$ , è di solito impossibile calcolare l'esatto  $P$ -value direttamente dalla tavola. È però facile trovare limiti inferiori e superiori sul  $P$ -value nella medesima tavola.

A titolo di esempio, si consideri un test  $t$  a singola coda superiore (per cui  $H_1: \mu > \mu_0$ ) con 14 gradi di libertà. I valori critici essenziali letti sulla Tavola II sono

#### Approssimazione del $P$ -value per un test $t$ .

Valore critico:	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
Area della coda:	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005

Dopo aver calcolato la statistica test si trova  $t_0 = 2.8$ ; questo valore si trova fra i due valori tabulati 2.624 e 2.977. Perciò, il  $P$ -value deve essere compreso fra 0.01 e 0.005 (Figura 4.18). Si tratta effettivamente dei limiti superiore e inferiore sul  $P$ -value.

Questa è la procedura per una verifica a coda superiore. Se fosse a coda inferiore, basterebbe cambiare il segno sui limiti di  $t_0$  e procedere come prima. Si ricordi che per una veri-

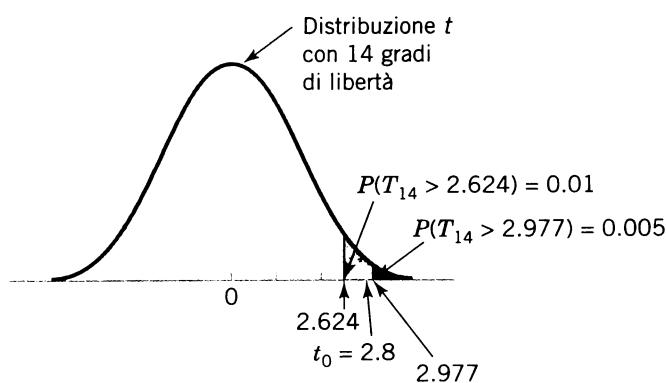


Figura 4.18  $P$ -value per  $t_0 = 2.8$ ; è mostrata una verifica a coda superiore compresa fra 0.005 e 0.01.

fica bilaterale il livello di significatività associato a un particolare valore critico è il doppio della corrispondente area di coda nell'intestazione di colonna. Bisogna tenere presente questa considerazione quando si calcola il limite sul  $P$ -value. Per esempio, si supponga di avere  $t_0 = 2.8$  per un'alternativa bilaterale basata su 14 gradi di libertà. Il valore è compreso tra  $t_0 > 2.624$  (corrispondente ad  $\alpha = 2 \times 0.01 = 0.02$ ) e  $t_0 < 2.977$  (corrispondente ad  $\alpha = 2 \times 0.005 = 0.01$ ), per cui i limiti inferiore e superiore sul  $P$ -value sarebbero in questo caso  $0.01 < P < 0.02$ .

#### Uso di Minitab per calcolare il $P$ -value

Alcuni software statistici, come detto, permettono di calcolare il  $P$ -value. Minitab, per esempio, è in grado di determinare le probabilità cumulate per molte distribuzioni standard, compresa la distribuzione  $t$ . Nel menu *Calc* si seleziona questo tipo di distribuzione e si inserisce il valore della statistica test  $t_0$ , assieme all'opportuno numero di gradi di libertà, v. Minitab visualizzerà allora la probabilità  $P = P(T_v \leq t_0)$ . Una volta nota la probabilità cumulata, si può ricavare il  $P$ -value.

Il test  $t$  per singolo campione appena descritto può essere condotto anche usando l'approccio basato sul livello di significatività fissato.

Si consideri in primo luogo l'ipotesi alternativa bilaterale: l'ipotesi nulla viene rifiutata se il valore della statistica  $t_0$  cade nella regione critica definita dagli  $\alpha/2$ -esimi punti percentuali inferiore e superiore della distribuzione  $t$  con  $n - 1$  gradi di libertà. Ovvero si rifiuta  $H_0$  se

$$t_0 > t_{\alpha/2, n-1} \quad \text{o} \quad t_0 < t_{\alpha/2, n-1}$$

Per i test a singola coda, la posizione della regione critica è determinata dal verso della diseguaglianza nell'ipotesi alternativa: se quest'ultima è  $H_1: \mu > \mu_0$ , si rifiuta  $H_0$  se

$$t_0 > t_{\alpha, n-1}$$

mentre se l'ipotesi alternativa è  $H_1: \mu < \mu_0$ , si rifiuta  $H_0$  se

$$t_0 < -t_{\alpha, n-1}$$

La Figura 4.19 mostra queste regioni critiche.

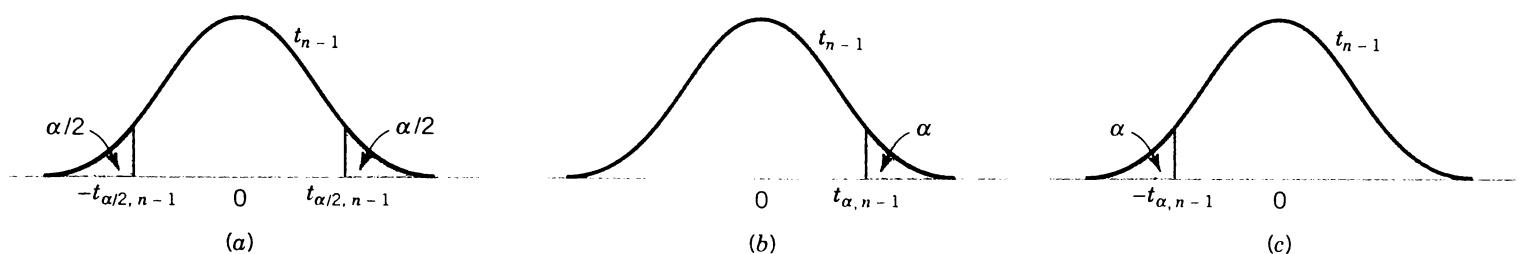


Figura 4.19 La distribuzione di  $T_0$  quando  $H_0: \mu = \mu_0$  è vera, con la regione critica per (a)  $H_1: \mu \neq \mu_0$ ; (b)  $H_1: \mu > \mu_0$ ; (c)  $H_1: \mu < \mu_0$ .

### Sintesi

#### Verifica di ipotesi sulla media di una distribuzione normale, varianza incognita

Ipotesi nulla:  $H_0: \mu = \mu_0$

Statistica test:  $T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$

Ipotesi alternativa	P-value	Criterio di rifiuto
$H_1: \mu \neq \mu_0$	Somma della probabilità a destra di $ t_0 $ , e di quella a sinistra di $- t_0 $	$t_0 > t_{\alpha/2, n-1}$ o $t_0 < -t_{\alpha/2, n-1}$
$H_1: \mu > \mu_0$	Probabilità a destra di $t_0$	$t_0 > t_{\alpha, n-1}$
$H_1: \mu < \mu_0$	Probabilità a sinistra di $t_0$	$t_0 < -t_{\alpha, n-1}$

Le regioni critiche per queste situazioni sono mostrate nelle Figure 4.19a, b, c, rispettivamente.

Come abbiamo osservato in precedenza, il test  $t$  è relativamente **robusto** rispetto all'assunzione di normalità, ossia scostamenti bassi o modesti dalla normalità hanno poco effetto sulla procedura. Si può sempre utilizzare un grafico dei quantili normali per verificare l'assunzione di normalità.

### ESEMPIO 4.7 Mazze da golf

L'aumentata disponibilità di materiali leggeri ad alta resistenza ha rivoluzionato la progettazione e la realizzazione di mazze da golf, in particolare dei *driver*. Mazze con testa concava e facce molto sottili permettono di eseguire colpi molto più lunghi sul *tee*, soprattutto se maneggiate da giocatori alle prime armi. Ciò è dovuto in parte all'“effetto molla” che la faccia sottile impedisce alla pallina. Tale effetto può venire quantificato “sparando” una pallina contro la testa della mazza e misurando il rapporto tra la velocità della pallina dopo e prima dell'urto con la mazza. Tale rapporto viene chiamato *coefficiente di restituzione* della mazza.

Si è eseguito un esperimento in cui sono stati selezionati in maniera casuale 15 driver di una particolare marca e sono stati misurati i relativi coefficienti di restituzione. Nell'esperimento, le palline sono state lanciate da un cannone ad aria compressa, in modo da controllare con precisione la velocità e la rotazione delle palline stesse. Ci interessa stabilire se vi sono prove (con  $\alpha = 0.05$ ) a sostegno dell'ipotesi che il coefficiente medio di restituzione sia maggiore di 0.82. Le osservazioni sono le seguenti:

0.8411	0.8191	0.8182	0.8125	0.8750
0.8580	0.8532	0.8483	0.8276	0.7983
0.8042	0.8730	0.8282	0.8359	0.8660

La media e la deviazione standard campionarie sono  $\bar{x} = 0.83725$  e  $s = 0.02456$ . Il grafico dei quantili di Figura 4.20 supporta l'assunzione che il coefficiente di restituzione è distribuito normalmente. Poiché l'obiettivo dello sperimentatore è di dimostrare che il coefficiente medio di restituzione è maggiore di 0.82, è appropriata un'ipotesi alternativa unilaterale.

La soluzione mediante la procedura a sette passi è la seguente:

1. **Parametro di interesse:** il parametro di interesse è il coefficiente medio di restituzione,  $\mu$ .
2. **Ipotesi nulla  $H_0$ :**  $\mu = 0.82$
3. **Ipotesi alternativa  $H_1$ :**  $\mu > 0.82$ . Vogliamo rifiutare  $H_0$  se il coefficiente medio di restituzione supera il valore 0.82.
4. **Statistica test:** la statistica test è

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

5. **Rifiutare  $H_0$  se:** si rifiuta  $H_0$  se il  $P$ -value è minore di 0.05.
6. **Calcoli:** poiché  $\bar{x} = 0.83725$ ,  $s = 0.02456$ ,  $\mu_0 = 0.82$  e  $n = 15$ , abbiamo

$$t_0 = \frac{0.83725 - 0.82}{0.02456/\sqrt{15}} = 2.72$$

7. **Conclusioni:** nella Tavola II dell'Appendice A troviamo, per una distribuzione  $t$  con 14 gradi di libertà, che  $t_0 = 2.72$  cade fra due valori: 2.624 per cui  $\alpha = 0.01$ , e 2.977

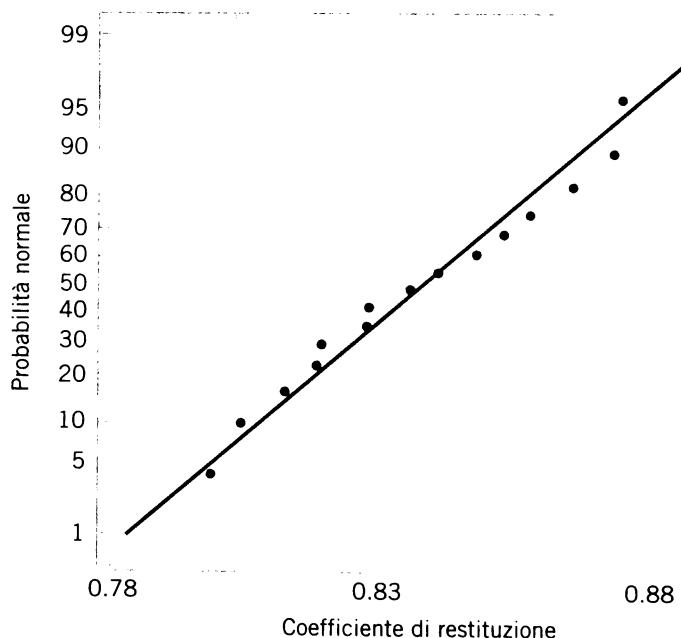


Figura 4.20 Grafico dei quantili dei dati per il coefficiente di restituzione dell'Esempio 4.7.

per cui  $\alpha = 0.005$ . Trattandosi di un test a singola coda, sappiamo che il  $P$ -value è compreso fra questi valori:  $0.005 < P < 0.01$ . Essendo dunque  $P$  minore di 0.05, rifiutiamo l'ipotesi nulla e concludiamo che il coefficiente di restituzione medio supera 0.82. Operando come indicato in precedenza, si arriva a ottenere da Minitab la probabilità  $P(T_{14} \leq 2.72) = 0.991703$ . Il  $P$ -value è  $P(T_{14} \geq 2.72)$  o  $P = 1 - (T_{14} \geq 2.72) = 1 - 0.991703 = 0.008297$ .

Dal **punto di vista pratico**, si conclude che vi sono forti indizi che questo tipo di mazza da golf abbia un coefficiente di restituzione maggiore di 0.82; se tale valore rappresenta una specifica di produzione o una soglia da non superare, è possibile che il produttore debba modificare il progetto dell'attrezzo.

Conducendo il test con Minitab, si ottiene un output come il seguente:

One-Sample T: COR					
Test of mu = 0.82 vs mu > 0.82					
Variable	N	Mean	StDev	SE Mean	
COR	15	0.83725	0.02456	0.00634	
Variable	95.0%	Lower Bound	T	P	
COR		0.82608	2.72	0.008	

Si noti che Minitab calcola sia la statistica test  $T_0$  sia un limite inferiore di confidenza al 95% per il coefficiente di restituzione. Daremo le formule per i limiti di confidenza nel Paragrafo 4.5.3. Tuttavia, sulla scorta della discussione del Paragrafo 4.4.5 sulla relazione tra verifiche di ipotesi e intervalli di confidenza, osserviamo che poiché il limite inferiore di confidenza al 95% è maggiore di 0.82, dovremmo rifiutare l'ipotesi  $H_0: \mu = 0.82$  e concludere che è più appropriata l'ipotesi alternativa  $H_1: \mu > 0.82$ . Minitab calcola anche un  $P$ -value per la statistica test  $T_0$ .

Il valore riportato è  $P = 0.008$ , che si trova fra il limite inferiore e il limite superiore che avevamo ricavato dalla tavola della distribuzione  $t$  ed è in buon accordo con il valore trovato direttamente mediante la funzione di Minitab che restituisce la distribuzione  $t$  cumulativa.

#### 4.5.2 Errore del II tipo e scelta della dimensione campionaria

La probabilità dell'errore del II tipo per verifiche di ipotesi sulla media di una distribuzione normale con varianza incognita dipende dalla distribuzione della statistica test dell'Equazione (4.39) quando l'ipotesi nulla  $H_0: \mu = \mu_0$  è falsa. Quando il vero valore della media è  $\mu = \mu_0 + \delta$ , la distribuzione per  $T_0$  viene detta **distribuzione  $t$  non centrale** con  $n - 1$  gradi di libertà e parametro di non centralità  $\delta\sqrt{n}/\sigma$ . Si noti che se  $\delta = 0$  la distribuzione  $t$  non centrale si riduce all'usuale **distribuzione  $t$  centrale**. Pertanto, l'errore del II tipo dell'alternativa bilaterale (per esempio) sarebbe

$$\begin{aligned}\beta &= P(-t_{\alpha/2, n-1} \leq T_0 \leq t_{\alpha/2, n-1} \text{ quando } \delta \neq 0) \\ &= P(-t_{\alpha/2, n-1} \leq T'_0 \leq t_{\alpha/2, n-1})\end{aligned}$$

dove  $T'_0$  indica la variabile aleatoria non centrale  $t$ . Trovare la probabilità  $\beta$  dell'errore del II tipo per il test  $t$  comporta il trovare la probabilità compresa fra due punti sulla distribuzione  $t$  non centrale. Poiché la variabile aleatoria non centrale  $t$  ha una funzione densità complessa, l'integrazione va eseguita numericamente.

#### Utilizzo delle curve OC

Per fortuna questo compito non piacevole è già stato assolto, e i risultati sono rappresentati in una serie di grafici in Appendice A, Carte Va, Vb, Vc e Vd, che riportano  $\beta$  per il test  $t$  in funzione di un parametro  $d$  per varie dimensioni campionarie  $n$ . Tali grafici sono chiamati **curve caratteristiche operative** (od OC). Le curve vengono fornite per alternative bilaterali sulle Carte Va e Vb. Il fattore di scala in ascissa è definito da

$$d = \frac{|\mu - \mu_0|}{\sigma} = \frac{|\delta|}{\sigma} \quad (4.46)$$

Per l'alternativa unilaterale  $\mu > \mu_0$  come nell'Equazione (4.42) usiamo le Carte Vc e Vd con

$$d = \frac{\mu - \mu_0}{\sigma} = \frac{\delta}{\sigma} \quad (4.47)$$

mentre se  $\mu < \mu_0$  come nell'Equazione (4.4)

$$d = \frac{\mu_0 - \mu}{\sigma} = \frac{\delta}{\sigma} \quad (4.48)$$

Osserviamo che  $d$  dipende dal parametro incognito  $\sigma^2$ . Possiamo evitare questa difficoltà in molti modi. In alcuni casi possiamo usare i risultati di un esperimento precedente o le informazioni note a priori per fare una stima iniziale grossolana di  $\sigma^2$ . Se ci interessa valutare il comportamento del test dopo che i dati sono stati raccolti potremmo usare la varianza campionaria  $s^2$  per stimare  $\sigma^2$ . Se non disponiamo di esperienze precedenti con le quali stimare  $\sigma^2$ , definiamo rispetto a  $\sigma$  la differenza tra le medie  $d$  che vogliamo rilevare. Per esempio, se vogliamo rilevare una piccola differenza tra le medie, possiamo usare un valore  $d = |\delta|/\sigma \leq 1$  (per esempio), mentre se vogliamo rilevare solo differenze abbastanza grandi tra le medie, possiamo usare un valore  $d = |\delta|/\sigma = 2$  (per esempio). In altre parole, è il valore del rapporto  $|\delta|/\sigma$  a essere importante nella determinazione della dimensione campionaria, e se è possibile specificare la dimensione relativa della differenza tra medie che siamo interessati a rilevare, allora si può di solito scegliere un valore di  $d$  appropriato.

#### ESEMPIO 4.8 Mazze da golf

Si consideri il problema del test sulle mazze da golf dell'Esempio 4.7. Se il coefficiente medio di restituzione differisce da 0.82 per una quantità pari a 0.2, la dimensione campionaria  $n = 15$  è adeguata per assicurare che l'ipotesi  $H_0: \mu = 0.82$  venga rifiutata con probabilità almeno di 0.8?

Per risolvere questo problema useremo la deviazione campionaria standard  $s = 0.02456$  per stimare  $\sigma$ . Allora  $d = |\delta|/\sigma = 0.2/0.02456 = 0.81$ . Riferendoci alle curve caratteristiche operative della Carta Vc (per  $\alpha = 0.05$ ) con  $d = 0.81$  e  $n = 15$ , troviamo che  $\beta$  è approssimativamente uguale a 0.10. Perciò la probabilità di rifiutare  $H_0: \mu = 0.82$  se la vera media supera questo valore di 0.02 vale approssimativamente  $1 - \beta = 1 - 0.10 = 0.90$ , e concludiamo che una dimensione campionaria  $n = 15$  è sufficiente per fornire la sensibilità richiesta.

Minitab esegue anche i calcoli della potenza e della dimensione campionaria per il test  $t$  a un campione. Il seguente riquadro mostra diversi calcoli effettuati sui dati del problema relativo alle mazze da golf.

<b>1-Sample t Test</b>
Testing mean = null (versus > null)
Calculating power for mean = null + difference
Alpha = 0.05 Sigma = 0.02456
Sample
Difference      Size      Power
0.02            15        0.9117
<b>1-Sample t Test</b>
Testing mean = null (versus > null)
Calculating power for mean = null + difference
Alpha = 0.05 Sigma = 0.02456
Sample
Difference      Size      Power
0.01            15        0.4425
<b>1-Sample t Test</b>
Testing mean = null (versus > null)
Calculating power for mean = null + difference
Alpha = 0.05 Sigma = 0.02456
Sample              Target              Actual
Difference      Size      Power      Power
0.01            39        0.8000      0.8029

#### Interpretazione dell'output di Minitab.

Nella prima parte dell'output Minitab riproduce la soluzione dell'Esempio 4.8, verificando che una dimensione campionaria  $n = 15$  è sufficiente per dare una potenza almeno uguale a 0.8 se il coefficiente medio di restituzione supera 0.82 di almeno 0.02. Nella parte centrale è stato chiesto a Minitab di calcolare la potenza se la differenza tra le medie che si vuole rilevare è 0.01. Si noti che con  $n = 15$  la potenza scende sensibilmente, sino a 0.4435. La terza parte è la dimensione campionaria necessaria per dare una potenza almeno pari a 0.8 se la differenza tra le medie di interesse è in effetti 0.01. È necessaria una dimensione campionaria molto maggiore ( $n = 39$ ) per rilevare questa piccola differenza.

#### 4.5.3 Intervallo di confidenza per la media

È facile trovare un intervallo di confidenza al  $100(1 - \alpha)\%$  per la media di una distribuzione normale con varianza incognita procedendo come nel Paragrafo 4.4.5 In generale, la distribuzione di  $T = (\bar{X} - \mu)/(S/\sqrt{n})$  è una distribuzione  $t$  con  $n - 1$  gradi di libertà. Se  $t_{\alpha/2,n-1}$  è il  $100\alpha/2$ -esimo punto percentuale della distribuzione  $t$  con  $n - 1$  gradi di libertà, possiamo scrivere

$$P(-t_{\alpha/2,n-1} \leq T \leq t_{\alpha/2,n-1}) = 1 - \alpha$$

ovvero

$$P\left(-t_{\alpha/2,n-1} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2,n-1}\right) = 1 - \alpha$$

Ridisponendo i termini dell'ultima equazione si ottiene

$$P(\bar{X} - t_{\alpha/2,n-1}S/\sqrt{n} \leq \mu \leq \bar{X} + t_{\alpha/2,n-1}S/\sqrt{n}) = 1 - \alpha \quad (4.49)$$

che conduce alla seguente definizione dell'intervallo di confidenza bilaterale al  $100(1 - \alpha)\%$  per  $\mu$ .

**Intervallo di confidenza per la media di una distribuzione normale, varianza incognita**

Se  $\bar{x}$  e  $s$  sono la media e la deviazione standard campionarie di un campione casuale estratto da una popolazione con varianza incognita  $\sigma^2$ , un **intervallo di confidenza al  $100(1 - \alpha)\%$**  per  $\mu$  è dato da

$$\bar{x} - t_{\alpha/2,n-1}s/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}s/\sqrt{n} \quad (4.50)$$

dove  $t_{\alpha/2,n-1}$  è il  $100\alpha/2$ -esimo punto percentuale superiore della distribuzione  $t$  con  $n - 1$  gradi di libertà.

### Limite di confidenza unilaterale

Per trovare un limite inferiore al  $100(1 - \alpha)\%$  per  $\mu$  con varianza incognita  $\sigma^2$ , è sufficiente sostituire  $-t_{\alpha/2,n-1}$  con  $-t_{\alpha,n-1}$  nell'estremo inferiore dell'Equazione (4.50) e porre quello superiore a  $\infty$ . Analogamente, per trovare un limite superiore al  $100(1 - \alpha)\%$  su  $\mu$  con varianza incognita  $\sigma^2$ , è sufficiente sostituire  $t_{\alpha/2,n-1}$  con  $t_{\alpha,n-1}$  nel limite superiore e porre quello inferiore a  $-\infty$ . Queste formule sono riassunte nelle tabelle in seconda e terza pagina di copertina.

### ESEMPIO 4.9 Mazze da golf

Si riconsideri il problema del coefficiente di restituzione dell'Esempio 4.7. Sappiamo che  $n = 15$ ,  $\bar{x} = 0.83725$  e  $s = 0.02456$ . Troveremo un intervallo di confidenza al 95% per  $\mu$ . Dall'Equazione (4.50) troviamo ( $t_{\alpha/2,n-1} = t_{0.025,14} = 2.145$ )

$$\begin{aligned} \bar{x} - t_{\alpha/2,n-1}s/\sqrt{n} &\leq \mu \leq \bar{x} + t_{\alpha/2,n-1}s/\sqrt{n} \\ 0.83725 - 2.145(0.02456)/\sqrt{15} &\leq \mu \leq 0.83725 + 2.145(0.02456)/\sqrt{15} \\ 0.83725 - 0.01360 &\leq \mu \leq 0.83725 + 0.01360 \\ 0.82365 &\leq \mu \leq 0.85085 \end{aligned}$$

Nell’Esempio 4.7 abbiamo verificato un’ipotesi alternativa unilaterale su  $\mu$ . Alcuni ingegneri possono essere interessati a un limite di confidenza unilaterale. Si ricordi che l’output di Minitab ha in effetti calcolato un limite di confidenza inferiore. Il limite inferiore di confidenza al 95% per il coefficiente medio di restituzione è

$$\begin{aligned}\bar{x} - t_{0.05,n-1}s/\sqrt{n} &\leq \mu \\ 0.83725 - 1.761(0.02456)/\sqrt{15} &\leq \mu \\ 0.82608 &\leq \mu\end{aligned}$$

Perciò possiamo stabilire con una confidenza del 95% che il coefficiente medio di restituzione supera 0.82608, che è il risultato riportato anche da Minitab.

## 4.6 INFERENZA SULLA VARIANZA DI UNA POPOLAZIONE NORMALE

A volte sono necessarie verifiche di ipotesi e intervalli di confidenza sulla varianza della popolazione o sulla deviazione standard. Se abbiamo un campione casuale  $X_1, X_2, \dots, X_n$ , la varianza campionaria  $S^2$  è uno stimatore puntuale non distorto di  $\sigma^2$ . Quando la popolazione è modellizzata da una distribuzione normale, sono applicabili i test e gli intervalli descritti in questo paragrafo.

### 4.6.1 Verifica di ipotesi sulla varianza di una popolazione normale

Si supponga di volere verificare l’ipotesi che la varianza di una popolazione normale  $\sigma^2$  è uguale a uno specifico valore, per esempio  $\sigma_0^2$ . Sia  $X_1, X_2, \dots, X_n$  un campione casuale composto da  $n$  osservazioni ricavate da questa popolazione. Per verificare

$$\begin{aligned}H_0: \sigma^2 = \sigma_0^2 \\ H_1: \sigma^2 \neq \sigma_0^2\end{aligned}\tag{4.51}$$

useremo la seguente statistica test

$$X_0^2 = \frac{(n - 1)S^2}{\sigma_0^2}\tag{4.52}$$

Per definire la procedura di verifica, dobbiamo conoscere la distribuzione della statistica test  $X_0^2$  nell’Equazione (4.52) quando l’ipotesi nulla è vera.

Sia  $X_1, X_2, \dots, X_n$  un campione casuale estratto da una distribuzione normale con media  $\mu$  e varianza  $\sigma^2$  incognite. La quantità

$$X^2 = \frac{(n-1)S^2}{\sigma^2} \quad (4.53)$$

ha una distribuzione chi-quadro con  $n-1$  gradi di libertà, che indichiamo in maniera abbreviata con  $\chi_{n-1}^2$ . In generale, la funzione di densità di probabilità di una variabile aleatoria chi-quadro è

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2} \quad x > 0 \quad (4.54)$$

dove  $k$  è il numero di gradi di libertà e  $\Gamma(k/2)$  è stata definita nel Paragrafo 4.5.1.

La media e la varianza della distribuzione  $\chi^2$  sono, rispettivamente,

$$\mu = k \quad \text{e} \quad \sigma^2 = 2k \quad (4.55)$$

In Figura 4.21 sono mostrate diverse distribuzioni chi-quadro. Si noti che la variabile aleatoria chi-quadro è non negativa e che la distribuzione di probabilità è asimmetrica verso destra. Tuttavia, all'aumentare di  $k$ , la distribuzione diventa più simmetrica. Per  $k \rightarrow \infty$ , la forma limite della distribuzione chi-quadro è la distribuzione normale.

I punti percentuali della distribuzione  $\chi^2$  sono dati nella Tavola III dell'Appendice A. Definiamo  $\chi_{\alpha,k}^2$  come il punto percentuale o valore della variabile aleatoria chi-quadro con  $k$  gradi di libertà tale che la probabilità che  $X^2$  superi questo valore è  $\alpha$ , ovvero

$$P(X^2 > \chi_{\alpha,k}^2) = \int_{\chi_{\alpha,k}^2}^{\infty} f(u) du = \alpha$$

**Utilizzo della Tavola III dell'Appendice A per la distribuzione  $\chi^2$ .**

Questa probabilità è rappresentata dall'area ombreggiata di Figura 4.22. Per illustrare l'utilizzo della Tavola III, si noti che le aree  $\alpha$  sono le intestazioni di colonna, mentre i gradi di libertà  $k$  sono riportati nella colonna di sinistra, etichettata con  $v$ .

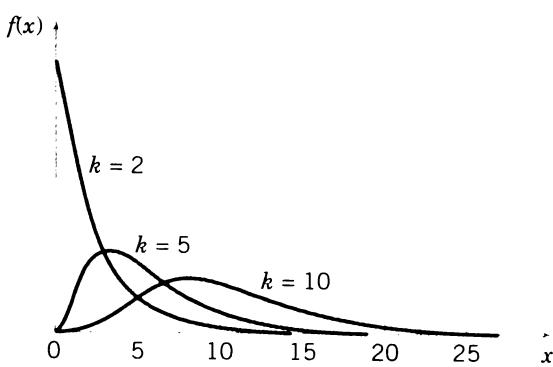


Figura 4.21 Funzioni di densità di probabilità di diverse distribuzioni  $\chi^2$ .

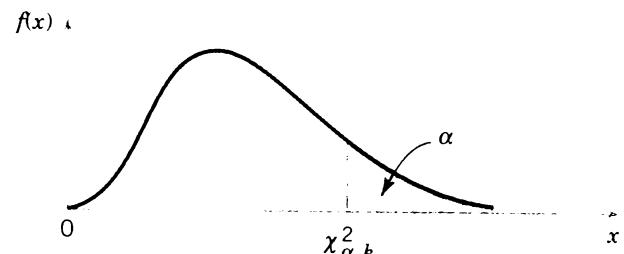


Figura 4.22 Il punto percentuale  $\chi_{\alpha,k}^2$  della distribuzione  $\chi^2$ .

Pertanto con 10 gradi di libertà il valore avente alla destra un'area (una probabilità) pari a 0.05 è  $\chi^2_{0.05,10} = 18.31$ . Questo valore viene spesso chiamato punto 5% superiore di chi-quadro con 10 gradi di libertà. Possiamo scrivere quanto detto sotto forma di asserzione probabilistica

$$P(X^2 > \chi^2_{0.05,10}) = P(X^2 > 18.31) = 0.05$$

È relativamente semplice costruire un test per l'ipotesi dell'Equazione (4.51). Se l'ipotesi nulla  $H_0: \sigma^2 = \sigma_0^2$  è vera, la statistica test  $X_0^2$  definita nell'Equazione (4.52) segue la distribuzione chi-quadro con  $n - 1$  gradi di libertà. Calcoliamo quindi il valore della statistica test  $X_0^2$  e rifiutiamo l'ipotesi  $H_0: \sigma^2 = \sigma_0^2$  se

$$X_0^2 > \chi^2_{\alpha/2,n-1} \quad \text{o se} \quad X_0^2 < \chi^2_{1-\alpha/2,n-1} \quad (4.56)$$

dove  $\chi^2_{\alpha/2,n-1}$  e  $\chi^2_{1-\alpha/2,n-1}$  sono, rispettivamente, i  $100\alpha/2$ -esimi punti percentuali superiore e inferiore della distribuzione chi-quadro con  $n - 1$  gradi di libertà. La regione critica è mostrata in Figura 4.23a.

La medesima statistica test è usata per ipotesi alternative unilaterali. Per l'ipotesi unilaterale

$$\begin{aligned} H_0: \sigma^2 &= \sigma_0^2 \\ H_1: \sigma^2 &> \sigma_0^2 \end{aligned} \quad (4.57)$$

rifiuteremmo  $H_0$  se

$$X_0^2 < \chi^2_{1-\alpha,n-1} \quad (4.58)$$

Per l'altra ipotesi unilaterale

$$\begin{aligned} H_0: \sigma^2 &= \sigma_0^2 \\ H_1: \sigma^2 &< \sigma_0^2 \end{aligned} \quad (4.59)$$

rifiuteremmo  $H_0$  se

$$X_0^2 < \chi^2_{\alpha,n-1} \quad (4.60)$$

Le regioni critiche unilaterali sono mostrate nelle Figure 4.23b e c.

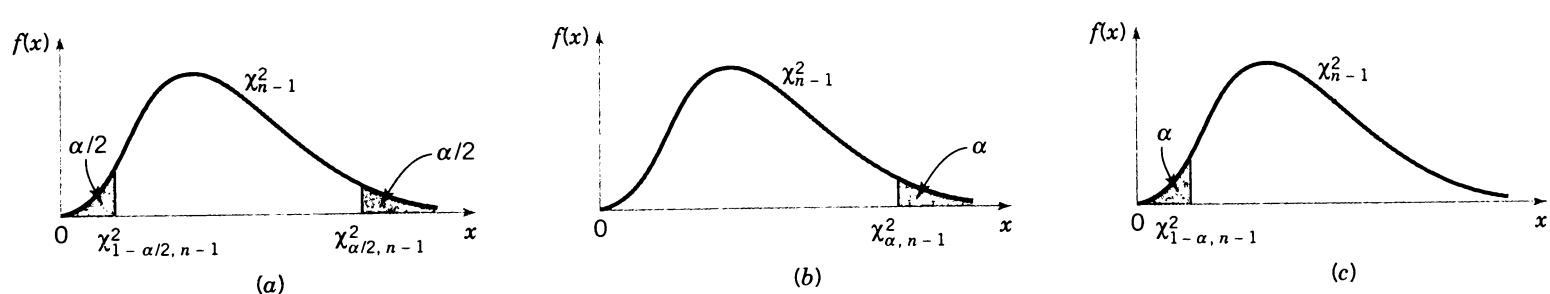


Figura 4.23 Distribuzione della statistica test per  $H_0: \sigma^2 = \sigma_0^2$  con valori la regione critica per (a)  $H_1: \sigma^2 \neq \sigma_0^2$  (b)  $H_1: \sigma^2 > \sigma_0^2$  (c)  $H_1: \sigma^2 < \sigma_0^2$ .

### Verifica di ipotesi sulla varianza di una distribuzione normale

Ipotesi nulla:  $H_0: \sigma^2 = \sigma_0^2$

Statistica test:  $\chi_0^2 = \frac{(n - 1)S^2}{\sigma_0^2}$

**Ipotesi alternativa**

$$H_1: \sigma^2 \neq \sigma_0^2$$

$$H_1: \sigma^2 > \sigma_0^2$$

$$H_1: \sigma^2 < \sigma_0^2$$

**Criterio di rifiuto**

$$\chi_0^2 > \chi_{\alpha/2, n-1}^2 \text{ o } \chi_0^2 < \chi_{1-\alpha/2, n-1}^2$$

$$\chi_0^2 > \chi_{\alpha, n-1}^2$$

$$\chi_0^2 < \chi_{1-\alpha, n-1}^2$$

Le collocazioni della regione critica sono mostrate in Figura 4.23.

#### ESEMPIO 4.10

##### Riempimento di bottiglie

Una macchina imbottigliatrice è impiegata per riempire bottiglie con un liquido detergente. Un campione casuale di 20 bottiglie dà luogo a una varianza campionaria del volume riempito pari a  $s^2 = 0.0153$  once fluide al quadrato. Se la varianza del volume riempito supera 0.01 once fluide al quadrato, una frazione non accettabile di bottiglie sarà sotto-riempita e sovra-riempita. Vi è qualche indicazione nei dati del campione che suggerisca che l'imbottigliatore ha un problema con bottiglie sotto-riempite e sovra-riempite? Usare  $\alpha = 0.05$  e assumere che il volume riempito abbia una distribuzione normale.

Per la risoluzione usiamo la procedura composta da sette passaggi.

1. **Parametro di interesse:** il parametro di interesse è la varianza della popolazione,  $\sigma^2$ .

2. **Ipotesi nulla  $H_0$ :**  $\sigma^2 = 0.01$

3. **Ipotesi alternativa  $H_1$ :**  $\sigma^2 > 0.01$

4. **Statistica test:** la statistica test è

$$\chi_0^2 = \frac{(n - 1)s^2}{\sigma_0^2}$$

5. **Rifiutare  $H_0$  se:** usando un test a livello di significatività fissato, si rifiuta  $H_0$  se  $\chi_0^2 > \chi_{0.05, 19}^2 = 30.14$ .

6. **Calcoli:**

$$\chi_0^2 = \frac{19(0.0153)}{0.01} = 29.07$$

7. **Conclusioni:** poiché  $\chi_0^2 = 29.07 < \chi_{0.05, 19}^2 = 30.14$ , concludiamo che non c'è una prova che la varianza del volume riempito superi 0.01 once fluide al quadrato.

**Conclusioni pratiche:** non vi sono solide ragioni per rifiutare l'asserzione che  $\sigma^2 = 0.01$ . Tuttavia, come vedremo più avanti, il  $P$ -value è circa uguale a 0.065, perciò la consueta “conclusione debole” associata con il mancato rifiuto di  $H_0$  è ancora più debole. Forse si dovrebbe prendere in considerazione un altro esperimento con una dimensione campionaria maggiore.

**Calcolo del *P*-value per un test chi-quadro.**

I *P*-value possono venire calcolati anche con test chi-quadro. Si consideri l’Esempio 4.10, che comportava un test a singola coda superiore. Il *P*-value è la probabilità a destra del valore calcolato della statistica test nella distribuzione  $\chi^2_{n-1}$ . Dato che la Tavola III contiene solo 11 aree di coda (colonne), di solito si devono trovare i limiti inferiore e superiore per *P*. È semplice: il valore calcolato della statistica test nell’Esempio 4.10 è  $\chi^2_0 = 29.07$ .

Analizzando la tabella, troviamo che  $\chi^2_{0.05,19} = 27.20$  e  $\chi^2_{0.05,19} = 30.14$ . Siccome  $27.20 < 29.07 < 30.14$ , concludiamo che il *P*-value per il test dell’Esempio 4.10 è nell’intervallo  $0.05 < P < 0.10$ .

Il *P*-value per il test a coda inferiore viene determinato come area (probabilità) della coda inferiore della distribuzione  $\chi^2_{n-1}$ , a sinistra del valore calcolato  $\chi^2_0$ . Per l’alternativa bilaterale, si trova l’area della coda associata al valore calcolato della statistica test e lo si raddoppia per ottenere il *P*-value.

#### 4.6.2 Intervallo di confidenza per la varianza di una popolazione normale

È stato fatto osservare nel precedente paragrafo che se la popolazione è normale la distribuzione campionaria di

$$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

è una distribuzione chi-quadro con  $n-1$  gradi di libertà. Per sviluppare l’intervallo di confidenza, scriviamo innanzitutto

$$P(\chi^2_{1-\alpha/2,n-1} \leq X^2 \leq \chi^2_{\alpha/2,n-1}) = 1 - \alpha$$

in modo che

$$P\left[\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right] = 1 - \alpha$$

Quest’ultima equazione può essere riscritta come

$$P\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right] = 1 - \alpha \quad (4.61)$$

e porta alla seguente definizione dell’intervallo di confidenza per  $\sigma^2$ .

#### Intervallo di confidenza per la varianza di una distribuzione normale

Se  $s^2$  è la varianza campionaria ricavata da un campione casuale di  $n$  osservazioni da una distribuzione normale con varianza incognita  $\sigma^2$ , un intervallo di confidenza al 100(1 -  $\alpha$ )% per  $\sigma^2$  è

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}} \quad (4.62)$$

dove  $\chi^2_{\alpha/2,n-1}$  e  $\chi^2_{1-\alpha/2,n-1}$  sono, rispettivamente, i  $100\alpha/2$ -esimi punti percentuali superiore e inferiore della distribuzione chi-quadro con  $n-1$  gradi di libertà. Per trovare un intervallo di confidenza per la deviazione standard,  $\sigma$ , si estraе semplicemente la radice quadrata dei termini nella doppia diseguaglianza (4.62).

### Limiti di confidenza unilaterali

Per trovare un limite inferiore di confidenza al  $100(1 - \alpha)\%$  su  $\sigma^2$ , poniamo il limite superiore di confidenza nell'Equazione (4.62) uguale a  $\infty$ , e sostituiamo  $\chi_{\alpha/2,n-1}^2$  con  $\chi_{\alpha,n-1}^2$ . Il limite di confidenza al  $100(1 - \alpha)\%$  superiore è trovato ponendo il limite di confidenza inferiore nell'Equazione (4.62) uguale a zero e sostituendo  $\chi_{1-\alpha/2,n-1}^2$  con  $\chi_{1-\alpha,n-1}^2$ . Per comodità, queste equazioni per la costruzione degli intervalli di confidenza unilaterali superiore e inferiore sono forniti nelle tabelle in seconda e terza pagina di copertina.

#### ESEMPIO 4.11 Riempimento di bottiglie

Si riconsideri la macchina imbottigliatrice dell'Esempio 4.10. Continueremo ad assumere che il volume riempito sia in via approssimativa distribuito normalmente. Un campione casuale di 20 bottiglie dà luogo a una varianza campionaria  $s^2 = 0.0153$  (once fluide)<sup>2</sup>. Un limite superiore di confidenza al 95% si ricava dall'Equazione (4.62) nel seguente modo

$$\sigma^2 \leq \frac{(n - 1)s^2}{\chi_{0.95,19}^2}$$

ovvero

$$\sigma^2 \leq \frac{(19)0.0153}{10.12} = 0.0287 \text{ (once fluide)}^2$$

Quest'ultima affermazione può essere trasformata in un limite di confidenza per la deviazione standard  $\sigma$  estraendo la radice quadrata di entrambi i membri dell'equazione. Otteniamo così

$$\sigma \leq 0.17 \text{ once fluide}$$

Pertanto, a un livello di confidenza del 95%, i dati indicano che la deviazione standard del processo potrebbe arrivare a 0.17 once fluide.

**Conclusioni pratiche:** L'intervallo di confidenza indica che vi è una ragionevole possibilità che la deviazione standard possa valere fino a 0.17 once fluide; dal punto di vista ingegneristico, è ora necessario stabilire se ciò possa portare a un rischio inaccettabile di riempire più del dovuto o meno delle bottiglie.

## 4.7 INFERENZA SULLA PROPORZIONE DI UNA POPOLAZIONE

Spesso è necessario verificare ipotesi e costruire intervalli di confidenza su una proporzione di popolazione. Per esempio, si supponga di aver estratto un campione casuale di dimensione  $n$  da una popolazione numerosa (possibilmente infinita), e che  $X$  ( $\leq n$ ) osservazioni del campione appartengano a una classe di interesse. Allora  $\hat{P} = X/n = X/n$  è uno stimatore puntuale della proporzione  $p$  della popolazione che appartiene a questa classe. Si noti che  $n$  e  $p$  sono i parametri di una distribuzione binomiale. Inoltre, dal Capitolo 3 sappiamo che la distribuzione campionaria di  $\hat{P}$  è approssimativamente normale con media  $p$  e varianza  $p(1-p)/n$ , se  $p$  non è troppo vicina a 0 o a 1 e se  $n$  è relativamente grande. Tipicamente, per applicare questa approssimazione si richiede che  $np$  e  $n(1-p)$  siano maggiori o uguali a 5. In questo paragrafo faremo uso dell'approssimazione normale.

### 4.7.1 Verifica di ipotesi su una proporzione binomiale

In molti problemi di ingegneria abbiamo a che fare con una variabile aleatoria che segue la distribuzione binomiale. Per esempio, si consideri un processo produttivo che realizza articoli classificati in due categorie: articoli accettabili e articoli difettosi. È in genere ragionevole modellizzare il presentarsi di articoli difettosi con la distribuzione binomiale, dove il parametro binomiale  $p$  rappresenta la proporzione di articoli difettosi prodotti. Di conseguenza, molti problemi decisionali, in ingegneria, comportano la verifica di ipotesi su  $p$ .

Considereremo la verifica di

$$\begin{aligned} H_0: p &= p_0 \\ H_1: p &\neq p_0 \end{aligned} \quad (4.63)$$

Sarà data una verifica approssimata basata sull'approssimazione normale alla binomiale. Come abbiamo osservato in precedenza, questa procedura approssimativa sarà valida fino a che  $p$  non è molto vicina a 0 o a 1, e se la dimensione campionaria è relativamente grande. Per eseguire la verifica dell'ipotesi e per costruire intervalli di confidenza su  $p$  verrà usato il seguente risultato.

Sia  $X$  il numero di osservazioni in un campione casuale di dimensione  $n$  che appartengono alla classe associata a  $p$ . Allora la quantità

$$Z = \frac{X - np}{\sqrt{np(1 - p)}} \quad (4.64)$$

ha approssimativamente una distribuzione normale standard,  $N(0, 1)$ .

Pertanto, se l'ipotesi nulla  $H_0: p = p_0$  è vera, abbiamo approssimativamente  $X \sim N[np_0, np_0(1 - p_0)]$ . Per verificare  $H_0: p = p_0$ , calcoliamo la **statistica test**

$$Z_0 = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}}$$

e determiniamo il  $P$ -value. Poiché la statistica test segue una distribuzione normale standard se  $H_0$  è vera, il  $P$ -value viene calcolato esattamente come quello per il test  $z$  (Paragrafo 4.4). Dunque, per l'ipotesi alternativa bilaterale, il  $P$ -value è la somma delle probabilità secondo la distribuzione normale standard a destra del valore positivo calcolato  $|z_0|$  e a sinistra del valore negativo  $-|z_0|$ , ossia

$$P = 2[1 - \Phi(|z_0|)]$$

Per l'ipotesi alternativa unilaterale  $H_1: p > p_0$  il  $P$ -value è la probabilità a destra di  $z_0$ , ossia

$$P = 1 - \Phi(z_0)$$

e per l'ipotesi alternativa unilaterale  $H_1: p < p_0$  il *P-value* è la probabilità a sinistra di  $z_0$ , ossia

$$P = \Phi(z_0)$$

È anche possibile eseguire un test a **livello di significatività fissato**. Per l'ipotesi alternativa bilaterale rifiutiamo  $H_0: p = p_0$  se

$$z_0 > z_{\alpha/2} \quad \text{o} \quad z_0 < -z_{\alpha/2}$$

Le regioni critiche per le ipotesi alternative unilaterali verranno costruite nel modo usuale.

### Sintesi

<b>Verifica di ipotesi su una proporzione binomiale</b>		
<b>Ipotesi nulla:</b>	$H_0: p = p_0$	<b>Criterio di rifiuto per test a livello fissato</b>
Statistica test:	$Z_0 = \frac{X - np_0}{\sqrt{np_0(1 - p_0)}}$	
<b>Ipotesi alternativa</b>	<b>P-value</b>	
$H_1: p \neq p_0$	Somma della probabilità a destra di $ z_0 $ e di quella a sinistra di $- z_0 $ , $P = 2[1 - \Phi( z_0 )]$	$z_0 > z_{\alpha/2} \quad \text{o} \quad z_0 < -z_{\alpha/2}$
$H_1: p > p_0$	Probabilità a destra di $t_0$ , $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1: p < p_0$	Probabilità a sinistra di $t_0$ , $P = \Phi(z_0)$	$z_0 < -z_\alpha$

### ESEMPIO 4.12

Dispositivi  
di controllo  
per motori

Un'azienda di semiconduttori produce i dispositivi di controllo impiegati nella motoristica per automobili. Il cliente richiede che lo scarto del processo o la frazione di pezzi difettosi in una fase di realizzazione critica non superi 0.05, e che l'azienda dimostri una capacità di processo a questo livello di qualità usando  $\alpha = 0.05$ . L'azienda di semiconduttori preleva un campione casuale di 200 dispositivi e trova che 4 di essi sono difettosi. È in grado il produttore di dimostrare al cliente la capacità di processo richiesta?

Possiamo risolvere questo problema usando la procedura di verifica di ipotesi a sette passaggi:

1. **Parametro di interesse:** il parametro di interesse è la frazione di pezzi difettosi del processo,  $p$ .
2. **Ipotesi nulla**  $H_0: p = 0.05$
3. **Ipotesi alternativa**  $H_1: p < 0.05$

Questa formulazione del problema permetterà al produttore di fare una dichiarazione forte riguardo la capacità di processo se è rifiutata l'ipotesi nulla  $H_0: p = 0.05$ .

**4. Statistica test:** la statistica test è (dall'Equazione (4.64))

$$z_0 = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}}$$

dove  $x = 4$ ,  $n = 200$  e  $p_0 = 0.05$ .

**5. Rifiutare  $H_0$  se:** si rifiuta  $H_0: p = 0.05$  se il  $P$ -value è minore di 0.05.

**6. Calcoli:** la statistica test è

$$z_0 = \frac{4 - 200(0.05)}{\sqrt{200(0.05)(0.95)}} = -1.95$$

**7. Conclusioni:** poiché  $z_0 = -1.95$ , il  $P$ -value è  $\Phi(-1.95) = 0.0256$ ; essendo minore di 0.05 rifiutiamo  $H_0$  e concludiamo che la frazione di pezzi difettosi del processo,  $p$ , è minore di 0.05. La conclusione pratica è che il processo soddisfa le richieste del cliente.

Occasionalmente incontreremo un'altra forma della statistica test  $Z_0$  dell'Equazione (4.64). Si noti che se  $X$  è il numero di osservazioni in un campione casuale di dimensione  $n$  che appartengono a una classe di interesse,  $\hat{P} = X/n$  è la proporzione campionaria che appartiene a questa classe. Dividiamo ora per  $n$  sia il numeratore che il denominatore di  $Z_0$  nell'Equazione (4.64). Otteniamo

$$Z_0 = \frac{X/n - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

ovvero

$$Z_0 = \frac{\hat{P} - p_0}{\sqrt{p_0(1 - p_0)/n}} \quad (4.65)$$

In questa equazione la statistica test è in funzione della proporzione campionaria invece che del numero di elementi  $X$  nel campione che appartengono alla classe di interesse.

Per eseguire il test su una proporzione binomiale si può utilizzare Minitab. Usando i dati dell'Esempio 4.12 si ottiene il seguente output:

<b>Test and CI for One Proportion</b>							
Test of $p = 0.05$ vs $p > 0.05$							
Sample	X	N	Sample p	95%		Z-Value	<i>P</i> -Value
				Upper Bound			
1	4	200	0.020000	0.036283		-1.95	0.026

\*Note\* The normal approximation may be inaccurate for small samples.

È mostrato fra l'altro anche un limite di confidenza superiore unilaterale al 95% per  $P$ . Nel Paragrafo 4.7.3 mostreremo come si calcolano gli intervalli di confidenza su una proporzione binomiale. Nell'output precedente il risultato visualizzato si basa sull'uso dell'approssimazione normale per i test e gli intervalli di confidenza; per piccole dimensioni campionarie questo approccio può rivelarsi inadeguato.

#### Verifiche su una proporzione binomiale per piccole dimensioni campionarie

Quando si ha un campione poco ampio, le verifiche sulla proporzione sono basate sulla distribuzione binomiale, anziché sull'approssimazione normale della binomiale. Per mostrare come avviene ciò, si immagini di voler sottoporre a verifica l'ipotesi  $H_0: p \leq p_0$ . Sia  $X$  il numero di successi nel campione. Il  $P$ -value di questo test si ricava dalla coda inferiore di una distribuzione binomiale con parametri  $n$  e  $p_0$ . Precisamente, è la probabilità che una variabile aleatoria binomiale con parametri  $n$  e  $p_0$  sia minore o uguale a  $X$ . Analogamente si procede per il test a coda superiore e per l'alternativa bilaterale.

Minitab è in grado di calcolare il  $P$ -value esatto per una verifica binomiale. Il seguente output riporta il  $P$ -value esatto per l'Esempio 4.12.

Test of $p = 0.05$ vs $p > 0.05$					
Sample	X	N	Sample p	95%	
				Upper Bound	Exact P-Value
1	4	200	0.020000	0.045180	0.026

Si tratta dello stesso valore ottenuto con l'approssimazione normale, perché nell'esempio la dimensione campionaria non è bassa. Si noti che invece l'intervalllo di confidenza è diverso da quello che si trova usando l'approssimazione normale.

#### 4.7.2 Errore del II tipo e scelta della dimensione campionaria

È possibile ottenere equazioni chiuse per l'errore  $\beta$  approssimato per i test del Paragrafo 4.7.1. Si supponga che  $p$  sia il vero valore della proporzione di una popolazione.

L'errore  $\beta$  approssimato per l'alternativa bilaterale  $H_1: p \neq p_0$  è

$$\begin{aligned}\beta &= \Phi\left[\frac{p_0 - p + z_{\alpha/2} \sqrt{p_0(1 - p_0)/n}}{\sqrt{p(1 - p)/n}}\right] \\ &\quad - \Phi\left[\frac{p_0 - p - z_{\alpha/2} \sqrt{p_0(1 - p_0)/n}}{\sqrt{p(1 - p)/n}}\right]\end{aligned}\quad (4.66)$$

Se l'alternativa è  $H_1: p < p_0$ ,

$$\beta = 1 - \Phi\left[\frac{p_0 - p - z_{\alpha} \sqrt{p_0(1 - p_0)/n}}{\sqrt{p(1 - p)/n}}\right]\quad (4.67)$$

mentre se l'alternativa è  $H_1: p > p_0$ ,

$$\beta = \Phi\left[\frac{p_0 - p + z_{\alpha} \sqrt{p_0(1 - p_0)/n}}{\sqrt{p(1 - p)/n}}\right]\quad (4.68)$$

Si possono risolvere queste equazioni per trovare la dimensione campionaria approssimata,  $n$ , che fornisce una verifica di livello  $\alpha$  avente uno specifico rischio  $\beta$ . L'equazione per la dimensione campionaria è la seguente.

#### **Dimensione campionaria per una verifica di ipotesi bilaterale su una proporzione binomiale**

---

$$n = \left[ \frac{z_{\alpha/2} \sqrt{p_0(1 - p_0)} + z_{\beta} \sqrt{p(1 - p)}}{p - p_0} \right]^2\quad (4.69)$$

Se  $n$  non è un intero occorre arrotondare la dimensione campionaria all'intero successivo.

Per un'alternativa unilaterale occorre sostituire  $z_{\alpha/2}$  con  $z_{\alpha}$  nell'Equazione (4.69).

**ESEMPIO 4.13**  
Dispositivi  
di controllo

Si consideri di nuovo l'Esempio 4.12. Si supponga che lo scarto del processo sia realmente  $p = 0.03$ . Qual è l'errore  $\beta$  per questa verifica della capacità di processo, che usa  $n = 200$  e  $\alpha = 0.05$ ?

L'errore  $\beta$  può essere calcolato usando l'Equazione (4.67)

$$\beta = 1 - \Phi\left[\frac{0.05 - 0.03 - (1.645) \sqrt{0.05(0.95)/200}}{\sqrt{0.03(1 - 0.03)/200}}\right] = 1 - \Phi(-0.44) = 0.67$$

Perciò, si ha una probabilità pari a circa 0.7 che il produttore mancherà di concludere che il processo soddisfa le richieste se la vera frazione di parti difettose del processo è  $p = 0.03$  (3%). Sembra un errore  $\beta$  elevato, ma la differenza tra  $p = 0.05$  e  $p = 0.03$  è piuttosto piccola, e la dimensione campionaria  $n = 200$  non è particolarmente elevata.

Si supponga che l'azienda di semiconduttori sia disposta ad accettare un errore  $\beta$  pari a 0.10 se la vera frazione di parti difettose è  $p = 0.03$ . Se continua a usare  $\alpha = 0.05$ , quale dimensione campionaria sarebbe necessaria?

La dimensione campionaria richiesta può venire calcolata dall'Equazione (4.69)

$$n = \left[ \frac{1.645\sqrt{0.05(0.95)} + 1.28\sqrt{0.03(0.97)}}{0.03 - 0.05} \right]^2 \approx 832$$

dove abbiamo usato  $p = 0.03$  nell'Equazione (4.69) e abbiamo sostituito  $z_{\alpha/2}$  con  $z_\alpha$  per un'alternativa unilaterale. Si noti che  $n = 832$  è una dimensione campionaria molto elevata. Tuttavia, stiamo provando a rilevare una deviazione abbastanza piccola dal valore nullo  $p_0 = 0.05$ .

#### 4.7.3 Intervallo di confidenza per una proporzione binomiale

Usando l'approssimazione normale, è immediato trovare un intervallo approssimato di confidenza al  $100(1 - \alpha)\%$  su una proporzione binomiale. Ricordiamo che la distribuzione campionaria di  $\hat{P}$  è approssimativamente normale con media  $p$  e varianza  $p(1 - p)/n$ , se  $p$  non è troppo vicino a 0 o a 1, e se  $n$  è relativamente grande. Allora la distribuzione di

$$Z = \frac{X - np}{\sqrt{np(1 - p)}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1 - p)}{n}}} \quad (4.70)$$

è approssimativamente normale standard.

Per costruire l'intervallo di confidenza su  $p$ , notiamo che

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \approx 1 - \alpha$$

cosicché

$$P\left[-z_{\alpha/2} \leq \frac{\hat{P} - p}{\sqrt{\frac{p(1 - p)}{n}}} \leq z_{\alpha/2}\right] \approx 1 - \alpha$$

Questa equazione può essere riscritta come

$$P\left[\hat{P} - z_{\alpha/2}\sqrt{\frac{p(1 - p)}{n}} \leq p \leq \hat{P} + z_{\alpha/2}\sqrt{\frac{p(1 - p)}{n}}\right] \approx 1 - \alpha \quad (4.71)$$

La quantità  $\sqrt{p(1 - p)/n}$  nell'Equazione (4.71) è detta **errore standard dello stimatore puntuale  $\hat{P}$** . Sfortunatamente, i limiti superiore e inferiore dell'intervallo di confidenza ottenuti dall'Equazione (4.71) contengono il parametro incognito  $p$ . Tuttavia, una soluzione soddisfacente consiste nel sostituire  $p$  con  $\hat{P}$  nell'errore standard, il che porta a

$$P\left[\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1 - \hat{P})}{n}} \leq p \leq \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1 - \hat{P})}{n}}\right] \approx 1 - \alpha \quad (4.72)$$

L'Equazione (4.72) porta all'intervallo approssimato di confidenza al  $100(1 - \alpha)\%$  per  $p$ .

### Intervallo di confidenza per una proporzione binomiale

Se  $\hat{p}$  è la proporzione di osservazioni in un campione casuale di dimensione  $n$  che appartengono a una classe di interesse, un intervallo approssimato di confidenza al  $100(1 - \alpha)\%$  per la proporzione  $p$  della popolazione che appartiene a questa classe è

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad (4.73)$$

dove  $z_{\alpha/2}$  è il punto percentuale superiore  $100\alpha/2$  della distribuzione normale standard.

Questa procedura dipende dall'adeguatezza dell'approssimazione normale alla binomiale. Per essere ragionevolmente garantiti, ciò richiede che  $np$  e  $n(1 - p)$  siano maggiori o uguali a 5. Nelle situazioni in cui questa approssimazione è inappropriata, in particolare nei casi in cui  $n$  è piccola, si devono usare altri metodi. Si potrebbero usare le tabelle di distribuzione binomiale per ottenere un intervallo di confidenza per  $p$ , ma preferiamo usare metodi numerici basati sulla funzione di massa di probabilità che sono implementati in software statistici. È così che opera Minitab, come illustrato nel riquadro relativo all'Esempio 4.12.

#### ESEMPIO 4.14 Cuscinetti per alberi di trasmissione

In un campione casuale di 85 cuscinetti per alberi di trasmissione delle auto, 10 hanno una finitura superficiale più grossolana di quanto permettono le specifiche. Pertanto, una stima puntuale della proporzione di cuscinetti nella popolazione che superano la specifica di rugosità superficiale è  $\hat{p} = x/n = 10/85 = 0.12$ . Un intervallo di confidenza bilaterale al 95% per  $p$  è calcolato dall'Equazione (4.73) come segue

$$\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq p \leq \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

ovvero

$$0.1176 - 1.96\sqrt{\frac{0.1176(0.88)}{85}} \leq p \leq 0.1176 + 1.96\sqrt{\frac{0.1176(0.88)}{85}}$$

che, semplificato, dà

$$0.0491 \leq p \leq 0.1861$$

#### Scelta della dimensione campionaria

Siccome  $\hat{P}$  è lo stimatore puntuale di  $p$ , possiamo definire l'errore commesso stimando  $p$  con  $\hat{P}$  come  $E = |\hat{P} - p|$ . Si noti che siamo approssimativamente confidenti al

$100(1 - \alpha)\%$  che questo errore sia minore di  $z_{\alpha/2} \sqrt{p(1 - p)/n}$ . Per esempio, nel problema dell'Esempio 4.14, siamo confidenti al 95% che la proporzione campionaria  $\hat{p} = 0.12$  differisce dalla vera proporzione  $p$  di una quantità che non supera 0.07.

Nelle situazioni in cui la dimensione campionaria può essere selezionata, possiamo scegliere  $n$  in modo da essere confidenti al  $100(1 - \alpha)\%$  che l'errore sia minore di qualche valore specifico di  $E$ . Se poniamo  $E = z_{\alpha/2} \sqrt{p(1 - p)/n}$  e risolviamo rispetto a  $n$ , otteniamo la seguente formula.

#### Dimensione campionaria per uno specifico errore $E$ per una proporzione binomiale

Se si usa  $\hat{P}$  come stima di  $p$ , si può ottenere una confidenza del  $100(1 - \alpha)\%$  sul fatto che l'errore  $|\hat{P} - p|$  non supererà una specifica quantità  $E$  quando la dimensione campionaria è

$$n = \left( \frac{z_{\alpha/2}}{E} \right)^2 p(1 - p) \quad (4.74)$$

Per usare l'Equazione (4.74) è necessaria una stima di  $p$ . Se è disponibile una stima di  $\hat{p}$  da un precedente campione, essa può venire sostituita al posto di  $p$  nell'Equazione (4.74), oppure se ne può fare una stima soggettiva. Se queste alternative non sono soddisfacenti, si può prendere in considerazione un campione preliminare, calcolare  $\hat{p}$ , e infine usare l'Equazione (4.74) per determinare quante osservazioni aggiuntive sono necessarie per stimare  $p$  con l'accuratezza desiderata. Un altro approccio per la scelta di  $n$  sfrutta il fatto che la dimensione campionaria calcolata con l'Equazione (4.74) sarà massima per  $p = 0.5$  [cioè,  $p(1 - p) \leq 0.25$  che diventa uguaglianza per  $p = 0.5$ ], e questo può essere usato per ottenere un limite superiore su  $n$ . In altre parole, siamo confidenti almeno al  $100(1 - \alpha)\%$  che l'errore nella stima di  $p$  con  $\hat{p}$  è minore di  $E$  se la dimensione campionaria è selezionata come descritto di seguito.

Per uno specifico errore  $E$ , un limite superiore sulla dimensione campionaria per la stima di  $p$  è

$$n = \left( \frac{z_{\alpha/2}}{E} \right)^2 \frac{1}{4} \quad (4.75)$$

#### ESEMPIO 4.15 Cuscinetti per alberi di trasmissione

Si consideri la situazione descritta nell'Esempio 4.14. Quanto deve essere grande il campione se vogliamo essere confidenti al 95% che l'errore commesso stimando  $p$  con  $\hat{p}$  sia minore di 0.05? Usando  $\hat{p} = 0.1176$  come stima iniziale di  $p$ , troviamo dall'Equazione (4.74) che la dimensione campionaria richiesta è

$$n = \left( \frac{z_{0.025}}{E} \right)^2 \hat{p}(1 - \hat{p}) = \left( \frac{1.96}{0.05} \right)^2 0.1176(0.8824) \approx 160$$

Se volevamo essere confidenti *almeno* al 95% che la nostra stima per la vera proporzione  $p$  fosse entro il valore 0.05, indipendentemente dal valore di  $p$ , avremmo dovuto usare l'Equazione (4.75) per trovare la dimensione campionaria

$$n = \left( \frac{z_{0.025}}{E} \right)^2 (0.25) = \left( \frac{1.96}{0.05} \right)^2 (0.25) \approx 385$$

Si noti che se disponessimo di informazioni riguardanti il valore di  $p$ , o da un campione preliminare o da una precedente esperienza, useremmo un campione più piccolo, mantenendo sia la precisione della stima sia il livello di confidenza desiderati.

### Limiti di confidenza unilaterali

Per trovare un limite inferiore approssimato di confidenza al  $100(1 - \alpha)\%$  su  $p$ , si sostituisce semplicemente  $-z_{\alpha/2}$  con  $-z_\alpha$  nel limite inferiore dell'Equazione (4.73) e si pone il limite superiore a 1. Analogamente, per trovare un limite superiore approssimato di confidenza al  $100(1 - \alpha)\%$  su  $p$  si sostituisce  $z_{\alpha/2}$  con  $z_\alpha$  nel limite superiore dell'Equazione (4.73) e si pone il limite inferiore a 0. Queste formule sono riassunte nelle tabelle in seconda e terza pagina di copertina. Analogamente, quando determiniamo la dimensione campionaria nel caso di limiti di confidenza unilaterali, sostituiamo semplicemente  $z_{\alpha/2}$  con  $z_\alpha$  nelle Equazioni (4.74) e (4.75).

### Un intervallo di confidenza alternativo per una proporzione binomiale

Vi è un modo di costruire un intervallo di confidenza per una proporzione binomiale alternativo a quello tradizionale espresso dall'Equazione (4.73). Partiamo dall'Equazione (4.71) e sostituiamo le disuguaglianze con uguaglianze, quindi risolviamo l'equazione quadratica rispetto a  $p$

$$p = \frac{\hat{P} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + \frac{z_{\alpha/2}^2}{n}}$$

Questo comporta che un intervallo di confidenza bilaterale sulla proporzione  $p$  sia

$$\begin{aligned} u &= \frac{\hat{P} + \frac{z_{\alpha/2}^2}{2n} + z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + \frac{z_{\alpha/2}^2}{n}} \\ l &= \frac{\hat{P} + \frac{z_{\alpha/2}^2}{2n} - z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + \frac{z_{\alpha/2}^2}{n}} \end{aligned} \quad (4.76)$$

L'articolo di Agresti e Coull pubblicato nella rivista *The American Statistician* (intitolato “Approssimazione migliore di quella ‘esatta’ per la stima intervallare per una proporzione binomiale”, 1988, pp. 119-126) indica che l'effettivo livello di confidenza per l'intervallo

dell'Equazione (4.76) è più prossimo a quello "dichiarato" o nominale per quasi tutti i valori di  $\alpha$  e  $p$  rispetto all'intervallo tradizionale dell'Equazione (4.73). Gli autori sostengono inoltre che questo intervallo alternativo si può usare con quasi tutte le dimensioni campionarie, per cui non sono strettamente necessari i requisiti  $n\hat{P} \geq 5$  o 10 o  $n(1-\hat{P}) \geq 5$  o 10. Se la dimensione campionaria è elevata, la quantità  $z_{\alpha/2}^2/n$  sarà piccola, per cui l'**intervallo di confidenza di Agresti-Coull** (Equazione (4.76)) si ridurrà all'intervallo di confidenza tradizionale.

**ESEMPIO 4.16**  
CI di Agresti-Coull su una proporzione

Si considerino nuovamente i dati relativi ai cuscinetti introdotti nell'Esempio 4.14, in cui si aveva  $\hat{p} = 0.12$  e  $n = 85$ . L'intervallo di confidenza al 95% era

$$0.0491 \leq p \leq 0.1861$$

Per costruire quello alternativo usiamo l'Equazione (4.76)

$$u = \frac{\hat{P} + \frac{z_{\alpha/2}^2}{2n} + z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + \frac{z_{\alpha/2}^2}{n}} = \frac{0.12 + \frac{1.96^2}{2(85)} + 1.96 \sqrt{\frac{0.12(0.88)}{85} + \frac{1.96^2}{4(85^2)}}}{1 + \frac{1.96^2}{85}} = 0.2024$$

$$l = \frac{\hat{P} + \frac{z_{\alpha/2}^2}{2n} - z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + \frac{z_{\alpha/2}^2}{n}} = \frac{0.12 + \frac{1.96^2}{2(85)} - 1.96 \sqrt{\frac{0.12(0.88)}{85} + \frac{1.96^2}{4(85^2)}}}{1 + \frac{1.96^2}{85}} = 0.0654$$

I due intervalli di confidenza sarebbero più prossimi l'uno all'altro se la dimensione campionaria fosse più grande.

## 4.8 ALTRE STIME INTERVALLARI PER UN SINGOLO CAMPIONE

### 4.8.1 Intervallo di predizione

In alcune situazioni, siamo interessati a **predire** un'osservazione futura di una variabile aleatoria. Possiamo inoltre volere trovare un intervallo di valori verosimili per la variabile associata alla una predizione. Si tratta di un problema diverso dalla stima della media di tale variabile aleatoria, per cui un intervallo di confidenza per la media non è appropriato. A titolo di esempio, si considerino le mazze da golf trattate nell'Esempio 4.7. Si supponga di volere pianificare l'acquisto di un nuovo driver della marca sottoposta a test nell'esempio. Qual è una ragionevole predizione del coefficiente di restituzione per il driver che acquistiamo (che *non* è tra quelli studiati nell'esempio), e qual è un intervallo di valori verosimili per il coefficiente di restituzione? La media campionaria  $\bar{X}$  delle mazze sottoposte a test è una ragionevole predizione puntuale del coefficiente di restituzione della nuova mazza da golf; mostreremo come ottenere un **intervallo di predizione** (PI, *Prediction Interval*) al  $100(1 - \alpha)\%$  sulla nuova osservazione.

Si supponga che  $X_1, X_2, \dots, X_n$  sia un campione casuale estratto da una popolazione normale con media e varianza incognite. Vogliamo predire il valore di una singola osservazione futura, per esempio  $X_{n+1}$ . Come osservato in precedenza, la media del campione originale,  $\bar{X}$ ,

è una ragionevole predizione puntuale di  $X_{n+1}$ . Il valore atteso dell'errore di predizione è  $E(X_{n+1} - \bar{X}) = \mu - \mu = 0$  e la varianza dell'errore di predizione è

$$V(X_{n+1} - \bar{X}) = \sigma^2 + \frac{\sigma^2}{n} = \sigma^2 \left(1 + \frac{1}{n}\right)$$

essendo l'osservazione futura  $X_{n+1}$  indipendente dalla media campionaria corrente. L'errore di predizione è normalmente distribuito perché le osservazioni originali sono distribuite normalmente.

Pertanto

$$Z = \frac{X_{n+1} - \bar{X}}{\sigma \sqrt{1 + \frac{1}{n}}}$$

ha una distribuzione normale standard. Sostituendo  $\sigma$  con la deviazione standard campionaria  $S$  otteniamo

$$T = \frac{X_{n+1} - \bar{X}}{S \sqrt{1 + \frac{1}{n}}}$$

che ha una distribuzione  $t$  con  $n - 1$  gradi di libertà. Effettuando opportuni passaggi su questo rapporto  $T$  come abbiamo fatto precedentemente nello sviluppo degli intervalli di confidenza, otteniamo un intervallo di predizione per l'osservazione futura  $X_{n+1}$ .

### Intervallo di predizione

**Un intervallo di predizione al  $100(1 - \alpha)\%$  per una singola osservazione futura estratta da una distribuzione normale è dato da**

$$\bar{x} - t_{\alpha/2, n-1} s \sqrt{1 + \frac{1}{n}} \leq X_{n+1} \leq \bar{x} + t_{\alpha/2, n-1} s \sqrt{1 + \frac{1}{n}} \quad (4.77)$$

L'intervallo di predizione per  $X_{n+1}$  sarà sempre più ampio di un intervallo di confidenza per  $\mu$  perché vi è maggiore variabilità associata all'errore di predizione per  $X_{n+1}$  che non all'errore di stima per  $\mu$ . Questo è facilmente intuibile, essendo l'errore di predizione la differenza tra due variabili aleatorie ( $X_{n+1} - \bar{X}$ ), ed essendo l'errore di stima usato per costruire un intervallo di confidenza la differenza tra una variabile aleatoria e una costante ( $\bar{X} - \mu$ ). Per  $n$  che diventa grande ( $n \rightarrow \infty$ ), l'ampiezza dell'intervallo di confidenza tende a zero, divenendo il vero valore della media  $\mu$ , mentre l'ampiezza dell'intervallo di predizione si avvicina a  $2z_{\alpha/2}\sigma$ . Di conseguenza, al crescere di  $n$  l'incertezza della stima  $\mu$  va a zero, ma esisterà sempre un'incertezza riguardo l'osservazione futura  $X_{n+1}$  anche quando non occorre stimare alcuno dei parametri della distribuzione.

Infine, si ricordi che gli intervalli di confidenza e le verifiche di ipotesi sulla media sono relativamente insensibili all'assunzione di normalità. Gli intervalli di predizione, viceversa,

non condividono questa comoda caratteristica: sono piuttosto sensibili all'assunzione di normalità essendo essi associati a un singolo valore futuro estratto a caso da una distribuzione normale.

### ESEMPIO 4.17

#### Mazze da golf

Si riconsiderino le mazze da golf che sono state sottoposte a test nell'Esempio 4.7. Il coefficiente di restituzione è stato misurato per  $n = 15$  driver selezionati in maniera casuale, e si è trovato  $\bar{x} = 0.83725$  e  $s = 0.02456$ . Pianifichiamo l'acquisto di una nuova mazza della marca esaminata. Il grafico dei quantili in Figura 4.17 non indica alcun problema con l'assunzione di normalità. Una ragionevole predizione puntuale del suo coefficiente di restituzione è la media campionaria, 0.83725. Un intervallo di predizione al 95% del coefficiente di restituzione per il nuovo driver viene calcolato con l'Equazione (4.77)

$$\begin{aligned}\bar{x} - t_{\alpha/2, n-1}s\sqrt{1 + \frac{1}{n}} &\leq X_{n+1} \leq \bar{x} + t_{\alpha/2, n-1}s\sqrt{1 + \frac{1}{n}} \\ 0.83725 - 2.145(0.02456)\sqrt{1 + \frac{1}{15}} &\leq X_{16} \leq 0.83725 + 2.145(0.02456)\sqrt{1 + \frac{1}{15}} \\ 0.78284 &\leq X_{16} \leq 0.89166\end{aligned}$$

Perciò potremmo logicamente aspettarci che la nuova mazza avrà un coefficiente di restituzione compreso tra 0.78284 e 0.89166. Al confronto, l'intervallo bilaterale di confidenza al 95% per il coefficiente di restituzione medio risulta  $0.82365 \leq \mu \leq 0.85085$ . Si noti che l'intervallo di predizione è considerevolmente più esteso dell'intervallo di confidenza per la media.

## 4.8.2 Intervalli di tolleranza per una distribuzione normale

Benché gli intervalli di confidenza e di predizione siano molto utili, esiste un terzo tipo di intervallo che trova molte applicazioni. Si consideri la popolazione delle mazze da golf dalla quale è stato estratto il campione di dimensione  $n = 15$  usato negli Esempi 4.7 e 4.17. Si supponga di sapere con certezza che il coefficiente medio di restituzione per i driver in questa popolazione era  $\mu = 0.83$  e che la deviazione standard era  $\sigma = 0.025$ . Allora l'intervallo da  $0.83 - 1.96(0.025) = 0.781$  a  $0.83 + 1.96(0.025) = 0.879$  cattura i coefficienti di restituzione del 95% dei driver di questa popolazione, dato che l'intervallo da  $-1.96$  a  $+1.96$  cattura il 95% dell'area (la probabilità) sottesa dalla curva normale standard. In generale, l'intervallo da  $\mu - z_{\alpha/2}\sigma$  a  $\mu + z_{\alpha/2}\sigma$  è detto **intervallo di tolleranza al 100(1 -  $\alpha$ )%**.

Se i parametri della distribuzione normale  $\mu$  e  $\sigma$  sono incogniti, possiamo usare i dati ricavati da un campione casuale di dimensione  $n$  per calcolare  $\bar{x}$  e  $s$ , e formare quindi l'intervallo  $(\bar{x} - 1.96s, \bar{x} + 1.96s)$ . Tuttavia, data la variabilità campionaria in  $\bar{x}$  e  $s$ , è verosimile che questo intervallo conterrà meno del 95% dei valori della popolazione. La soluzione consiste nel sostituire 1.96 con qualche valore che porterà la proporzione della popolazione a essere contenuta nell'intervallo al 95% con qualche livello di confidenza. Fortunatamente, si tratta di un compito semplice.

### Intervallo di tolleranza

Un **intervallo di tolleranza** che contiene almeno  $\gamma\%$  dei valori di una popolazione normale con livello di confidenza al  $100(1 - \alpha)\%$  è

$$\bar{x} - ks, \bar{x} + ks$$

dove  $k$  è il fattore dell'intervallo di tolleranza per una distribuzione normale, tabulato nella Tavola VI dell'Appendice A. I valori di  $k$  sono dati per livelli di confidenza  $1 - \alpha = 0.90, 0.95, 0.99$  e per  $\gamma = 90, 95$  e  $99\%$ .

Possiamo calcolare anche i limiti di tolleranza unilaterali. I fattori di tolleranza per questi limiti sono anch'essi tabulati nella Tavola VI.

#### ESEMPIO 4.18 Mazze da golf

Si considerino nuovamente le mazze da golf dell'Esempio 4.7. Ricordiamo che la media campionaria e la deviazione standard dei coefficienti di restituzione per le  $n = 15$  mazze sottoposte a test sono  $\bar{x} = 0.83725$  e  $s = 0.02456$ . Vogliamo trovare un intervallo di tolleranza per il coefficiente di restituzione che includa il 95% delle mazze della popolazione con il 90% di confidenza. Dalla Tavola VI dell'Appendice A il fattore di tolleranza risulta  $k = 2.713$ . L'intervallo di tolleranza desiderato è

$$(\bar{x} - ks, \bar{x} + ks) \quad \text{ovvero} \quad [0.83725 - (2.713)0.02456, 0.83725 + (2.713)0.02456]$$

che si riduce a  $(0.77062, 0.90388)$ . Pertanto, possiamo essere confidenti al 90% che almeno il 95% delle mazze da golf in questa popolazione ha un coefficiente di restituzione compreso tra 0.77062 e 0.90388.

Dalla Tabella VI dell'Appendice A notiamo che, per  $n \rightarrow \infty$ , il valore del fattore di tolleranza  $k$  della distribuzione normale tende al valore  $z$  associato al livello desiderato di contenimento per la distribuzione normale. Per esempio, se vogliamo che il 95% della popolazione sia compreso nell'intervallo di tolleranza bilaterale,  $k$  tende a  $z_{0.05} = 1.96$  per  $n \rightarrow \infty$ . Si noti che per  $n \rightarrow \infty$  un intervallo di predizione al  $100(1 - \alpha)\%$  su un'osservazione futura si avvicina a un intervallo di tolleranza che contiene il  $100(1 - \alpha)\%$  della distribuzione.

## 4.9 TABELLE RIASSUNTIVE DELLE PROCEDURE DI INFERENZA PER UN SINGOLO CAMPIONE

Le tabelle in seconda e terza pagina di copertina riassumono tutte le procedure per le verifiche di ipotesi gli intervalli di confidenza a campione singolo trattate in questo capitolo. Esse contengono l'ipotesi nulla, la statistica test, le varie ipotesi alternative e il criterio per il rifiuto di  $H_0$ , nonché le formule per costruire gli intervalli di confidenza al  $100(1 - \alpha)\%$ .

## 4.10 TEST DI ADATTAMENTO

Le procedure di verifica di ipotesi che abbiamo discusso nei precedenti paragrafi sono progettate per problemi in cui la popolazione o la distribuzione di probabilità sono note e le ipotesi coinvolgono i parametri della distribuzione. Si incontra spesso un altro tipo di ipotesi: non conosciamo la sottostante distribuzione della popolazione, e vogliamo verificare l'ipotesi che una particolare distribuzione sarà soddisfacente come modello di popolazione. Per esempio, potremmo voler verificare l'ipotesi che la popolazione è normale.

Nel Capitolo 3 abbiamo discusso una tecnica grafica molto utile per questo problema, i **grafici dei quantili**, e abbiamo mostrato come applicarla alle distribuzioni normale, lognormale e di Weibull. In questo paragrafo descriviamo una procedura formale per il test di adattamento che si basa sulla distribuzione chi-quadro.

La procedura del test richiede un campione casuale di dimensione  $n$  estratto dalla popolazione la cui distribuzione di probabilità è incognita. Queste  $n$  osservazioni sono organizzate sotto forma di istogramma, con  $k$  classi. Sia  $O_i$  la frequenza osservata nell' $i$ -esima classe. Dalla distribuzione di probabilità ipotizzata calcoliamo la frequenza attesa nella  $i$ -esima classe, indicata con  $E_i$ . La statistica test è

$$X_0^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (4.78)$$

Si può dimostrare che se la popolazione segue la distribuzione ipotizzata,  $\chi_0^2$  ha approssimativamente una distribuzione chi-quadro con  $k - p - 1$  gradi di libertà, dove  $p$  rappresenta il numero di parametri della distribuzione ipotizzata stimato mediante statistiche campionarie. Questa approssimazione migliora all'aumentare di  $n$ . Rifiuteremo l'ipotesi che la distribuzione della popolazione è la distribuzione ipotizzata se il valore calcolato della statistica test è  $\chi_0^2 > \chi_{\alpha, k-p-1}^2$ .

Un aspetto da prendere in considerazione nell'applicazione di questa procedura di verifica riguarda il valore delle frequenze attese. Se queste frequenze attese sono troppo piccole, la statistica test  $\chi_0^2$  non rifletterà lo scostamento delle frequenze osservate da quelle attese, ma solo la piccola ampiezza delle frequenze attese. Non esiste un accordo generale sul valore minimo che devono assumere le frequenze attese, comunque si impiegano comunemente come minimi i valori 3, 4 e 5. Alcuni autori suggeriscono che una frequenza attesa potrebbe valere anche 1 o 2, fintanto che la maggior parte di esse supera 5. Se una frequenza attesa dovesse essere troppo piccola, la si può combinare con la frequenza attesa in una classe adiacente. Le frequenze osservate corrispondenti devono essere anch'esse combinate, e  $k$  deve essere ridotto a 1. Non è necessario che gli intervalli abbiano uguale larghezza.

Diamo ora un esempio della procedura di verifica.

**ESEMPIO 4.19**  
Schede circuitali

### Una distribuzione di Poisson

Si ritiene che il numero di difetti nelle schede circuitali segua una distribuzione di Poisson. È stato raccolto un campione casuale di  $n = 60$  schede circuitali, ed è stato osservato il numero di difetti presente in ognuna. Si sono ottenuti i seguenti dati:

Numero di difetti	Frequenza osservata
0	32
1	15
2	9
3	4

La media della distribuzione di Poisson assunta in questo esempio è incognita e deve essere stimata a partire dai dati campionari. La stima del numero medio di difetti per scheda è la media campionaria, ossia  $(32 \cdot 0 + 15 \cdot 1 + 9 \cdot 2 + 4 \cdot 3)/60 = 0.75$ . Dalla distribuzione di Poisson con parametro 0.75 possiamo calcolare  $p_i$ , la probabilità teorica ipotizzata associata all' $i$ -esima classe. Siccome ciascuna classe corrisponde a un particolare numero di difetti, possiamo trovare il parametro  $p_i$  nel seguente modo:

$$p_1 = P(X = 0) = \frac{e^{-0.75}(0.75)^0}{0!} = 0.472$$

$$p_2 = P(X = 1) = \frac{e^{-0.75}(0.75)^1}{1!} = 0.354$$

$$p_3 = P(X = 2) = \frac{e^{-0.75}(0.75)^2}{2!} = 0.133$$

$$p_4 = P(X \geq 3) = 1 - (p_1 + p_2 + p_3) = 0.041$$

Le frequenze attese vengono calcolate moltiplicando la dimensione campionaria  $n = 60$  per la probabilità  $p_i$ , cioè  $E_i = np_i$ . Le frequenze attese sono mostrate nella seguente tabella.

Numero di difetti	Probabilità	Frequenza attesa
0	0.472	28.32
1	0.354	21.24
2	0.133	7.98
3 (o più)	0.041	2.46

Siccome la frequenza attesa nell'ultima cella è minore di 3, combiniamo le ultime due celle:

Numero di difetti	Frequenza osservata	Frequenza attesa
0	32	28.32
1	15	21.24
2 (o più)	13	10.44

La statistica test chi-quadro nell'Equazione (4.78) avrà  $k - p - 1 = 3 - 1 - 1 = 1$  gradi di libertà perché la media della distribuzione di Poisson è stata stimata dai dati.

Applichiamo la procedura di verifica di ipotesi a sette passi, usando  $\alpha = 0.05$ .

1. **Parametro di interesse:** la variabile di interesse è la forma della distribuzione dei difetti presenti nelle schede circuitali.
2. **Ipotesi nulla  $H_0$ :** la forma della distribuzione dei difetti è quella di Poisson.
3. **Ipotesi alternativa  $H_1$ :** la forma della distribuzione dei difetti non è quella di Poisson.
4. **Statistica test:** la statistica test è

$$\chi^2_0 = \sum_{i=1}^k \frac{(o_i - E_i)^2}{E_i}$$

5. **Rifiutare  $H_0$  se:** si rifiuta  $H_0$  se il  $P$ -value è minore di 0.05.

6. **Calcoli:**

$$\chi^2_0 = \frac{(32 - 28.32)^2}{28.32} + \frac{(15 - 21.24)^2}{21.24} + \frac{(13 - 10.44)^2}{10.44} = 2.94$$

7. **Conclusioni:** dalla tavola III dell'Appendice A si ricava  $\chi^2_{0.10,1} = 2.71$  e  $\chi^2_{0.5,1} = 3.84$ . Poiché  $\chi^2_0 = 2.94$  cade tra questi due valori, concludiamo che il  $P$ -value è tale che  $0.05 < P < 0.10$ . pertanto, essendo maggiore di 0.05, non siamo in grado di rifiutare l'ipotesi nulla che la distribuzione dei difetti nelle schede circuitali è di Poisson. Il  $P$ -value esatto può essere calcolato con Minitab, ed è uguale a 0.0864.

## TERMINI E CONCETTI RILEVANTI

---

Coefficiente di confidenza

Copertura

Curve caratteristiche operative

Determinazione della dimensione campionaria

Distorsione della stima

Distribuzione chi-quadro

Distribuzione  $t$

Efficienza relativa di uno stimatore

Errore del I tipo

Errore del II tipo

Errore standard

Errore standard stimato

Eperimento comparativo

Inferenza statistica

Intervallo di confidenza

Intervallo di predizione

Intervallo di tolleranza

Ipotesi alternativa

Ipotesi alternativa bilaterale

Ipotesi alternativa unilaterale

Ipotesi nulla

Ipotesi statistica

Limiti di confidenza

Livello di confidenza

Livello di significatività

Potenza

Precisione della stima

Probabilità di un errore del I tipo

Probabilità di un errore del II tipo

Procedura per la verifica di ipotesi

$P$ -value

Regione critica

Significatività pratica e significatività statistica

Statistica test

Stima dei parametri

Stima puntuale

Stimatore non distorto a varianza minima

Test di adattamento

Verifica di ipotesi

# Esercizi proposti

---

## ESERCIZI PER IL PARAGRAFO 4.2

---

4.1. Nel seguente output di Minitab per un campione casuale di dati vi sono alcuni valori mancanti; è chiesto di calcolarli.

Variable	N	Mean	SE Mean	StDev	Variance	Min.	Max.
X	9	19.96	?	3.12	?	15.94	27.16

4.2. Nel seguente output di Minitab per un campione casuale di dati vi sono alcuni valori mancanti; è chiesto di calcolarli.

Variable	N	Mean	Variance	Sum of		Minimum	Maximum
				Sum	Squares		
X	10	?	?	109.891	1258.899	6.451	13.878

4.3. Si supponga di disporre di un campione casuale di dimensione  $2n$  estratto da una popolazione  $X$ , e che  $E(X) = \mu$ ,  $V(X) = \sigma^2$ . Siano

$$\bar{X}_1 = \frac{1}{2n} \sum_{i=1}^{2n} X_i \quad \text{e} \quad \bar{X}_2 = \frac{1}{n} \sum_{i=1}^n X_i$$

due stimatori di  $\mu$ . Quale stimatore di  $\mu$  è migliore? Spiegare i motivi della scelta.

4.4. Sia  $X_1, X_2, \dots, X_9$  un campione casuale estratto da una popolazione con media  $\mu$  e varianza  $\sigma^2$ . Si considerino i seguenti stimatori di  $\mu$ :

$$\hat{\Theta}_1 = \frac{X_1 + X_2 + \dots + X_9}{9}$$

$$\hat{\Theta}_2 = \frac{3X_1 - X_6 + 2X_4}{2}$$

- (a) Gli stimatori sono non distorti?
- (b) Quale stimatore è “il migliore”? In quale senso?

4.5. Si supponga che  $\hat{\Theta}_1$  e  $\hat{\Theta}_2$  siano stimatori non distorti del parametro  $\theta$ . Si sa che  $V(\hat{\Theta}_1) = 2$  e  $V(\hat{\Theta}_2) = 4$ . Quale stimatore è migliore, e in che senso?

4.6. Calcolare l’efficienza relativa dei due stimatori dell’Esercizio 4.4.

4.7. Calcolare l’efficienza relativa dei due stimatori dell’Esercizio 4.5.

4.8. Si supponga che  $\hat{\Theta}_1$  e  $\hat{\Theta}_2$  siano stimatori del parametro  $\theta$ . Si sa che  $E(\hat{\Theta}_1) = \theta$ ,  $E(\hat{\Theta}_2) = \theta/2$ ,  $V(\hat{\Theta}_1) = 10$  e  $V(\hat{\Theta}_2) = 4$ . Quale stimatore è “il migliore”, e in che senso?

4.9. Si supponga che  $\hat{\Theta}_1$ ,  $\hat{\Theta}_2$  e  $\hat{\Theta}_3$  siano stimatori del parametro  $\theta$ . Si sa che  $E(\hat{\Theta}_1) = E(\hat{\Theta}_2) = \theta$ ,  $E(\hat{\Theta}_3) \neq \theta$ ,  $V(\hat{\Theta}_1) = 16$  e  $V(\hat{\Theta}_2) = 11$  ed  $E(\hat{\Theta}_3 - \theta)^2 = 6$ . Confrontare questi tre stimatori. Quale è preferibile? Perché?

4.10. Si abbiano tre campioni di dimensioni  $n_1 = 20$ ,  $n_2 = 10$  e  $n_3 = 8$ , estratti da una popolazione con media  $\mu$  e varianza  $\sigma^2$ . Siano inoltre  $S_1^2$ ,  $S_2^2$  e  $S_3^2$  le rispettive varianze campionarie. Dimostrare che  $S^2 = (20S_1^2 + 10S_2^2 + 8S_3^2)/38$  è uno stimatore non distorto di  $\sigma^2$ .

4.11. (a) Dimostrare che  $\sum_{i=1}^n (X_i - \bar{X})^2/n$  è uno stimatore distorto di  $\sigma^2$ .  
 (b) Trovare la distorsione dello stimatore.  
 (c) Che cosa accade alla distorsione al crescere della dimensione campionaria?

4.12. Sia  $X_1, X_2, \dots, X_n$  un campione casuale di dimensione  $n$ .  
 (a) Dimostrare che  $\bar{X}^2$  è uno stimatore distorto per  $\mu^2$ .  
 (b) Trovare la distorsione dello stimatore.  
 (c) Che cosa accade alla distorsione al crescere della dimensione campionaria?

## ESERCIZI PER IL PARAGRAFO 4.3

4.13. Un fabbricante di fibra tessile sta studiando un nuovo filato, la cui forza di allungamento, di media  $\mu$ , ha una deviazione standard di 0.3 kg. L'azienda desidera verificare l'ipotesi  $H_0: \mu = 14$  contro  $H_1: \mu < 14$ , usando un campione casuale di cinque esemplari.

- (a) Qual è il  $P$ -value se la media campionaria è  $\bar{x} = 13.7$  kg?
- (b) Trovare  $\beta$  per il caso in cui la vera forza media di allungamento vale 13.5 kg, assumendo  $\alpha = 0.05$ .
- (c) Qual è la potenza del test al punto (b)?

4.14. Ripetere l'Esercizio 4.13 usando una dimensione campionaria  $n = 16$  e la stessa regione critica.

4.15. Nell'Esercizio 4.13, con  $n = 5$ :

- (a) Trovare i limiti della regione critica se è specificato che la probabilità dell'errore del I tipo deve essere  $\alpha = 0.01$ .
- (b) Trovare  $\beta$  per il caso in cui la vera forza media di allungamento vale 13.0 kg.
- (c) Qual è la potenza del test al punto (b)?

4.16. Nell'Esercizio 4.14, con  $n = 16$ :

(a) Trovare i limiti della regione critica se è specificato che la probabilità dell'errore del I tipo deve essere 0.05.

(b) Trovare  $\beta$  per il caso in cui la vera forza media di allungamento vale 13.0 kg.

(c) Qual è la potenza del test al punto (b)?

4.17. Un produttore industriale è interessato alla tensione di uscita di un alimentatore elettrico per PC. Si assume che tale tensione sia distribuita normalmente, con deviazione standard pari a 0.25 V; il produttore vuole verificare l'ipotesi  $H_0: \mu = 9$  V contro  $H_1: \mu \neq 9$  V usando  $n = 10$  unità.

- (a) La regione critica è  $\bar{x} < 8.85$  o  $\bar{x} > 9.15$ . Trovare il valore di  $\alpha$ .
- (b) Trovare la potenza del test nel rilevamento di una vera tensione media di uscita pari a 9.1 V.

4.18. Rielaborare l'Esercizio 4.17 nel caso in cui si hanno  $n = 16$  unità e i limiti della regione di accettazione non variano.

4.19. Si consideri l'Esercizio 4.17, e si supponga che l'ingegnere responsabile del processo voglia che la probabilità dell'errore del I tipo sia  $\alpha = 0.05$ . Dove dovrebbe venire collocata la regione critica?

## ESERCIZI PER IL PARAGRAFO 4.4

4.20. Si consideri il seguente output di Minitab.

One-Sample Z						
Test of mu = 30 vs not = 30						
The assumed standard deviation = 1.2						
N	Mean	SE Mean	95% CI	Z	P	
16	31.2000	0.3000	(30.6120, 31.7880)	?	?	

- (a) Completare inserendo i valori mancanti. Quali conclusioni si possono trarre?
- (b) Si tratta di un test unilaterale o bilaterale?
- (c) Usare l'output e la tavola della distribuzione normale per trovare un intervallo di confidenza di livello 99% per la media.
- (d) Qual è il  $P$ -value se l'ipotesi alternativa è  $H_1: \mu > 30$ ?

4.21. Si consideri il seguente output di Minitab.

One-Sample Z						
Test of mu = 20 vs not = 20						
The assumed standard deviation = 4						
N	Mean	SE Mean	95% CI	Z	P	
25	21.400	?	(19.832, 22.968)	?	?	?

- (a) Completare inserendo i valori mancanti. Si può rifiutare l'ipotesi nulla al livello di significatività 0.05? Perché?
- (b) Se l'ipotesi alternativa fosse  $H_1: \mu > 20$ , quale sarebbe il  $P$ -value? Si potrebbe rifiutare l'ipotesi nulla al livello di significatività 0.05? Perché?
- (c) Usare l'output e la tavola della distribuzione normale per trovare un intervallo di confidenza di livello 99% per la media.

4.22. Si consideri il seguente output di Minitab.

**One-Sample Z**

Test of mu = 100 vs > 100  
The assumed standard deviation = 5

N	Mean	SE Mean	95% Lower Bound		Z	P
			?	?		
8	105.20	1.77				

- (a) Completare inserendo i valori mancanti. Si può rifiutare l'ipotesi nulla al livello di significatività 0.05? Perché?  
 (b) Se l'ipotesi alternativa fosse  $H_1: \mu \neq 100$ , quale sarebbe il P-value? Si potrebbe rifiutare l'ipotesi nulla al livello di significatività 0.05? Perché?  
 (c) Supponendo di dover trovare un intervallo di confidenza bilaterale di livello 95% per la media, il limite inferiore di tale intervallo sarebbe maggiore del limite inferiore dell'intervallo unilaterale calcolato al punto (a)?

4.23. Data una popolazione normale con varianza  $\sigma^2$  nota, rispondere ai seguenti quesiti.

- (a) Qual è il livello di confidenza per l'intervallo

$$\bar{x} - 2.14\sigma\sqrt{n} \leq \mu \leq \bar{x} + 2.14\sigma\sqrt{n}?$$

- (b) Qual è il livello di confidenza per l'intervallo

$$\bar{x} - 2.49\sigma\sqrt{n} \leq \mu \leq \bar{x} + 2.49\sigma\sqrt{n}?$$

- (c) Qual è il livello di confidenza per l'intervallo

$$\bar{x} - 1.85\sigma\sqrt{n} \leq \mu \leq \bar{x} + 1.85\sigma\sqrt{n}?$$

4.24. La ricerca medica ha sviluppato un nuovo cuore artificiale composto principalmente in titanio e plastica. Una volta impiantato nel corpo del paziente, la sua durata e il suo funzionamento sono pressoché infiniti, ma è necessario ricaricare le batterie ogni 4 ore circa. Viene selezionato un campione di 50 batterie, poi sottoposto a un test di durata. La vita media risulta 4.05 ore. Si supponga che la vita delle batterie seguia una distribuzione normale con deviazione standard  $\sigma = 0.2$  ore.

- (a) Vi sono indizi che supportano l'ipotesi che la vita media delle batterie superi le 4 ore? Utilizzare  $\alpha = 0.05$ .  
 (b) Qual è il P-value per il test del punto (a)?  
 (c) Calcolare la potenza del test nel caso in cui la reale vita media delle batterie è 4.5 ore.  
 (d) Quale dimensione campionaria sarebbe necessaria per rilevare una vita media reale di 4.5 ore se si volesse che la potenza del test fosse almeno uguale a 0.9?  
 (e) Spiegare come si potrebbe rispondere al punto (a) costruendo un limite di confidenza unilaterale per la vita media.

4.25. La resa di un processo chimico viene sottoposta a studio. In base all'esperienza precedente con tale processo si sa che la deviazione standard del prodotto vale 3. Gli ultimi 5 giorni di operatività dell'impianto hanno dato le seguenti resse: 91.6, 88.75, 90.8, 89.95 e 91.3%. Usare  $\alpha = 0.05$ .

- (a) Vi è qualche indicazione che la resa media non sia del 90%? Usare l'approccio del P-value.  
 (b) Quale dimensione campionaria sarebbe necessaria per rilevare una vera resa media del 85% con una probabilità pari a 0.95?  
 (c) Qual è la probabilità dell'errore del II tipo se la vera resa media è del 92%?  
 (d) Trovare un intervallo bilaterale di confidenza al 95% sulla vera resa media.  
 (e) Usare l'intervallo di confidenza trovato al punto (d) per verificare l'ipotesi.

4.26. Nella produzione dei dispositivi di gonfiamento degli airbag per automobili, un'azienda è interessata ad assicurare che la distanza media dalla lamina all'estremità del dispositivo sia almeno 2.00 cm. Misure effettuate su 20 dispositivi hanno dato un valore medio della distanza pari a 2.02 cm. Si assuma una deviazione standard di 0.05 sulle misure di distanza e un livello di significatività di 0.01.

- (a) Effettuare un test di conformità al requisito dell'azienda. Usare l'approccio del P-value.  
 (b) Qual è il valore di  $\beta$  se la vera media è 2.03?  
 (c) Quale dimensione campionaria sarebbe necessaria per rilevare una vera resa media pari a 2.03 con una probabilità pari almeno a 0.90?  
 (d) Trovare un limite inferiore unilaterale di confidenza al 99% sulla vera media.  
 (e) Usare l'intervallo trovato al punto (e) per verificare l'ipotesi.

4.27. Un ingegnere civile analizza la resistenza alla compressione del cemento. Tale resistenza è distribuita in maniera approssimativamente normale con varianza  $\sigma^2 = 1000 \text{ psi}^2$ . Un campione casuale di 12 provini ha una resistenza media alla compressione  $\bar{x} = 3255.42 \text{ psi}$ .

- (a) Verificare l'ipotesi che la resistenza media alla compressione è 3500 psi. Usare  $\alpha = 0.01$ .  
 (b) Qual è il livello di significatività minimo al quale si vorrebbe rifiutare l'ipotesi nulla?  
 (c) Costruire un intervallo di confidenza bilaterale di livello 95% sulla resistenza media alla compressione.  
 (d) Costruire un intervallo di confidenza bilaterale di livello 99% sulla resistenza media alla compressione e confrontarne l'ampiezza con quella dell'intervallo trovato al punto (c). Commentare l'esito del confronto.

4.28. Si supponga che nell'Esercizio 4.26 volessimo essere confidenti al 95% che l'errore nella stima della distanza media è inferiore a 0.01. Quale dimensione campionaria dovremmo usare?

## ESERCIZI PER IL PARAGRAFO 4.5

4.30. Si consideri il seguente output di Minitab.

One-Sample T: X						
Test of mu = 91 vs not = 91						
		SE				
Variable	N	Mean	StDev	95% CI	T	P
X	25	92.5805	?	0.4673 (91.6160, ?)	3.38	0.002

- (a) Completare inserendo i valori mancanti. Si può rifiutare l'ipotesi nulla al livello 0.05? Perché?
- (b) Si tratta di un test unilaterale o bilaterale?
- (c) Se l'ipotesi fosse  $H_0: \mu = 90$  contro  $H_1: \mu \neq 90$ , si rifiuterebbe l'ipotesi nulla al livello 0.05?
- (d) Usare l'output e la tavola della distribuzione  $t$  per trovare un intervallo di confidenza bilaterale di livello 99% per la media.
- (e) Qual è il  $P$ -value se l'ipotesi alternativa è  $H_1: \mu > 91$ ?

4.31. Si consideri il seguente output di Minitab.

One-Sample T: X						
Test of mu = 25 vs > 25						
		95%				
Variable	N	Mean	StDev	Mean Bound	T	P
X	12	25.6818	?	0.3360 ?	?	0.034

- (a) Quanti gradi di libertà vi sono nella statistica test  $t$ ?
- (b) Completare inserendo le informazioni mancanti.

4.32. Si consideri il seguente output di Minitab.

Test of mu = 50 vs not = 50						
		SE				
Variable	N	Mean	StDev	95% CI	T	P
C1	10	49.30	3.62	? (46.71, 51.89)	?	?

- (a) Quanti gradi di libertà vi sono nella statistica test  $t$ ?
- (b) Completare inserendo le informazioni mancanti. Si possono calcolare i limiti per il  $P$ -value.

4.33. Un articolo pubblicato nella rivista *Computers in Electrical Engineering* ("Parallel Simulation of Cellular Neu-

ral Networks", 1996, Vol. 22, pp. 61-84) ha preso in considerazione la rapidità delle reti neurali cellulari (CNN) per l'architettura di calcolo in parallelo. I tempi sono i seguenti:

3.775302	3.350679	4.217981	4.030324	4.639692
4.139665	4.395575	4.824257	4.268119	4.584193
4.930027	4.315973	4.600101		

- (a) Vi sono indizi sufficienti per rifiutare l'asserzione secondo cui la rapidità media supera 4.0? Assumere  $\alpha = 0.05$ .
- (b) I dati hanno una distribuzione approssimativamente normale? Fare una rappresentazione grafica per supportare la risposta.
- (c) Trovare un intervallo di confidenza bilaterale di livello 95% per la rapidità media.
- (d) Quale dimensione campionaria sarebbe necessaria per rilevare una rapidità media reale a 4.75 se si volesse che la potenza del test fosse almeno 0.8? Usare la deviazione campionaria standard  $s$  calcolata al punto (a) come stima di  $\sigma$ .

4.34. Un articolo pubblicato sulla rivista *ASCE Journal of Energy Engineering* (Vol. 125, 1999, pp. 59-75) descrive uno studio sulle proprietà di inerzia termica di un particolare calcestruzzo utilizzato come materiale edile. Sono stati sottoposti a test cinque campioni di materiale, ottenendo le seguenti temperature interne medie (in °C): 23.01, 22.22, 22.04, 22.62 e 22.59.

- (a) Eseguire una verifica dell'ipotesi  $H_0: \mu = 22.5$  contro  $H_1: \mu \neq 22.5$ , utilizzando  $\alpha = 0.05$ . Seguire l'approccio basato sul  $P$ -value.
- (b) Verificare l'assunzione che la temperatura interna è normalmente distribuita.
- (c) Trovare un intervallo di confidenza al 95% per la temperatura media interna.
- (d) Quale dimensione campionaria sarebbe necessaria per rilevare una reale temperatura interna media pari a 22.75 se si volesse che la potenza del test fosse almeno 0.9? Usare la deviazione campionaria standard  $s$  come stima di  $\sigma$ .

4.35. Un ingegnere che lavora nel reparto Ricerca in un'azienda di pneumatici sta studiando la durata degli pneumatici

realizzati con una nuova mescola. Ha costruito 10 pneumatici e li ha sottoposti a un test stradale per capirne la durata della loro utilizzabilità. La media e la deviazione standard campionarie sono rispettivamente 61 492 e 3035 km.

- (a) L'ingegnere vorrebbe dimostrare che la vita media di questo nuovo pneumatico è superiore ai 60 000 km. Formulare e verificare le ipotesi appropriate, assicurandosi di stabilire (e verificare, se possibile) le assunzioni, quindi trarre le conclusioni usando  $\alpha = 0.05$ .
- (b) Si supponga che se la vita media è uguale a 61 000 km, l'ingegnere desideri rilevare tale differenza con una probabilità pari almeno a 0.90. La dimensione campionaria  $n = 10$  usata al punto (a) era adeguata? Usare la deviazione standard campionaria  $s$  come stima di  $\sigma$  per arrivare alla decisione.
- (c) Trovare un limite inferiore di confidenza al 95% sulla vita media degli pneumatici.
- (d) Usare il limite trovato al punto (c) per verificare l'ipotesi.

 4.36. Si sa che le ore di vita di un dispositivo biomedico in fase di sviluppo presso un laboratorio sono distribuite normalmente. Viene selezionato un campione di 15 dispositivi e ne risulta una vita media di 5625.1 ore e una deviazione standard di 226.1 ore.

- (a) Usando l'approccio del  $P$ -value, verificare l'ipotesi che la vera vita media del dispositivo biomedico è maggiore di 5500.
- (b) Trovare un limite inferiore di confidenza al 95% sulla media.
- (c) Usare il limite trovato al punto (b) per verificare l'ipotesi.

4.37. Nella costruzione di circuiti elettrici la tensione di rottura dei diodi è una caratteristica qualitativa importante. Si sono registrate le seguenti tensioni di rottura di 12 diodi: 9.099, 9.174, 9.327, 9.377, 8.471, 9.575, 9.514, 8.928, 8.800, 8.920, 9.913, 8.306

- (a) Controllare l'assunzione di normalità per i dati.
- (b) Sottoporre a verifica l'asserzione per cui la tensione di rottura è minore di 9 V con un livello di significatività pari a 0.05.
- (c) Costruire un limite superiore unilaterale di livello 95% sulla tensione media di rottura.

## ESERCIZI PER IL PARAGRAFO 4.6

 4.40. Si deve inserire un rivetto in un foro. Se la deviazione standard del diametro del foro supera 0.02 mm, vi è una probabilità non accettabile che il rivetto non entri nel foro. Viene selezionato un campione casuale di  $n = 15$  parti, e si misura il diametro del foro. La deviazione standard campionaria della misura del diametro del foro è  $s = 0.016$  mm.

- (d) Usare il limite trovato al punto (c) per verificare l'ipotesi.
- (e) Si supponga che la vera tensione di rottura sia 8.8 V; è importante rilevarla con una probabilità pari ad almeno 0.95. Usando la deviazione standard campionaria per stimare la deviazione standard della popolazione e un livello di significatività di 0.05, determinare la dimensione campionaria necessaria.

4.38. Un articolo pubblicato sul *Journal of Composite Materials* (December 1989, Vol. 23, p. 1200) descrive l'effetto della separazione degli strati sulla frequenza naturale di travi fatte da laminati compositi. Cinque di tali travi sono state sottoposte a carichi, ottenendo le seguenti frequenze (in Hz):

230.66, 233.05, 232.58, 229.48, 232.58

Trovare un intervallo di confidenza bilaterale di livello 90% sulla frequenza naturale media. I risultati dei calcoli supportano l'asserzione che la frequenza naturale media vale 235 Hz? Discutere quanto si è trovato e specificare ogni assunzione necessaria.

4.39. L'inseminazione delle nuvole (*cloud seeding*) è stata studiata per decenni come procedura di modifica del tempo atmosferico (per un interessante studio sull'argomento si veda l'articolo pubblicato in *Technometrics*, Vol. 17, 1975, pp. 161-166). Di seguito sono elencati i dati di piovosità, misurati in piedi-acro, ricavati dall'inseminazione con nitrato d'argento di 20 nuvole selezionate a caso: 18.0, 30.7, 19.8, 27.1, 22.3, 18.8, 31.8, 23.4, 21.2, 27.9, 31.9, 27.1, 25.0, 24.7, 26.9, 21.8, 29.2, 34.8, 26.7, 31.6.

- (a) Si è in grado di supportare l'asserzione per cui la piovosità media supera i 25 piedi-acro? Assumere  $\alpha = 0.05$ . Trovare il  $P$ -value.
- (b) Verificare che la piovosità è normalmente distribuita.
- (c) Calcolare la potenza del test nel caso in cui la vera piovosità media sia 27 piedi-acro.
- (d) Quale dimensione campionaria sarebbe necessaria per rilevare una reale piovosità media di 27.5 piedi-acro, se si volesse che la potenza del test fosse almeno 0.9?
- (e) Spiegare come si potrebbe rispondere al punto (a) costruendo un limite di confidenza unilaterale per la piovosità media.

- (a) Vi sono forti indizi che indichino che la deviazione standard del diametro del foro supera 0.02 mm? Calcolare il  $P$ -value e usarlo per trarre delle conclusioni. Specificare ogni assunzione necessaria sulla sottostante distribuzione dei dati.
- (b) Costruire un limite inferiore di confidenza di livello 95% per  $\sigma$ .
- (c) Usare il limite trovato al punto (b) per verificare l'ipotesi.

4.41. Si consideri l'Esercizio 4.35, riguardante la durata di uno pneumatico.

- (a) Usando  $\alpha = 0.05$ , è possibile concludere che la deviazione standard della durata dello pneumatico supera 3000 km? Specificare ogni assunzione necessaria sulla sottostante distribuzione dei dati.
- (b) Trovare il  $P$ -value per questo test.
- (c) Trovare un limite inferiore di confidenza al 95% per  $\sigma$ .
- (d) Usare il limite di confidenza trovato al punto (c) per verificare l'ipotesi.

 4.42. La percentuale di titanio presente in una lega usata per la realizzazione di veicoli aerospaziali è misurata su

## ESERCIZI PER IL PARAGRAFO 4.7

4.43. Si consideri il seguente output di Minitab.

Test and CI for One Proportion						
Test of $p = 0.3$ vs $p \neq 0.3$						
Sample	X	N	Sample p	95% CI	Z-Value	P-Value
1	95	250	0.380000	(0.319832, 0.440168)	2.76	0.006

- (a) Si tratta di un test unilaterale o bilaterale?
- (b) Il test è stato condotto usando l'approssimazione normale per la binomiale? È stata una scelta adeguata?
- (c) Si può rifiutare l'ipotesi nulla al livello 0.05?
- (d) Si può rifiutare l'ipotesi nulla  $H_0: p = 0.4$  contro  $H_1: p \neq 0.4$ , utilizzando  $\alpha = 0.05$ ? Come si può rispondere senza eseguire ulteriori calcoli?
- (e) Costruire un intervallo di confidenza tradizionale approssimato di livello 90% per  $p$ .

4.44. Si consideri il seguente output di Minitab.

Test and CI for One Proportion						
Test of $p = 0.65$ vs $p > 0.65$						
Sample	X	N	Sample p	95%	Z-Value	P-Value
				Lower		
1	553	800	?	?	2.45	?

- (a) Si tratta di un test unilaterale o bilaterale?
- (b) Il test è stato condotto usando l'approssimazione normale per la binomiale? È stata una scelta adeguata?
- (c) Completare con i valori mancanti.

4.45. Si ritiene che i grandi furgoni per il trasporto passeggeri abbiano un'alta propensione al capottamento quando

sono a pieno carico. Sono stati esaminati trenta incidenti che hanno coinvolto questi furgoni, e in 11 di essi è avvenuto il capottamento.

- (a) Verificare l'asserzione per cui la proporzione di capottamenti supera 0.25 con  $\alpha = 0.10$ .
- (b) Si supponga che la vera  $p$  sia 0.35 e  $\alpha = 0.10$ . Qual è l'errore  $\beta$  per questo test?
- (c) Si supponga che la vera  $p$  sia 0.35 e  $\alpha = 0.10$ . Quanto dovrebbe essere numeroso il campione se si vuole avere  $\beta = 0.10$ ?
- (d) Trovare un limite inferiore di confidenza al 90% per il tasso di capottamenti.
- (e) Usare il limite trovato al punto (d) per verificare l'ipotesi.
- (f) Quanto dovrebbe essere numeroso il campione per essere confidenti almeno al 95% che l'errore su  $p$  sia minore di 0.02? Usare una stima iniziale di  $p$  ricavata da questo problema.



4.46. Un produttore di calcolatrici elettroniche vuole stimare la frazione di unità difettose prodotte. Un campione casuale di 800 calcolatrici contiene 10 unità difettose.

- (a) Formulare e verificare un'appropriata ipotesi per determinare se la frazione di pezzi difettosi supera 0.01 al livello di significatività 0.05.
- (b) Si supponga che la vera  $p$  sia 0.02 e  $\alpha = 0.05$ . Qual è l'errore  $\beta$  per questo test?
- (c) Si supponga che la vera  $p$  sia 0.02 e  $\alpha = 0.05$ . Quanto dovrebbe essere numeroso il campione se si vuole avere  $\beta = 0.10$ ?



4.47. Si sta studiando la frazione di circuiti integrati difettosi prodotti da un processo fotolitografico. Viene sottoposto a test un campione casuale di 300 circuiti integrati, e si rilevano 18 circuiti difettosi.

- (a) Usare i dati per verificare l'ipotesi che la proporzione non è 0.04. Usare  $\alpha = 0.05$ .

- (b) Trovare il  $P$ -value per questo test.  
 (c) Trovare un intervallo di confidenza bilaterale di livello 95% per la proporzione di pezzi difettosi.  
 (d) Usare l'intervallo trovato al punto (d) per verificare l'ipotesi.

4.48. Si considerino i dati riguardanti i circuiti difettosi e l'ipotesi dell'Esercizio 4.47.

- (a) Si supponga che la frazione di pezzi difettosi sia effettivamente  $p = 0.05$ . Qual è l'errore  $\beta$  per questo test?  
 (b) Si supponga che il produttore voglia accettare un errore  $\beta$  di 0.10 se il vero valore di  $p$  è 0.05. Con  $\alpha = 0.05$ , quale dimensione campionaria è necessaria?

4.49. A un campione casuale di 500 elettori della città di Phoenix è stato chiesto se sono favorevoli all'uso per un anno di carburante ossigenato per ridurre l'inquinamento dell'aria. Se più di 315 elettori rispondono affermativamente, si concluderà che almeno il 60% dei votanti è favorevole all'uso di questo carburante.

- (a) Trovare la probabilità dell'errore del I tipo se esattamente il 60% dei votanti è favorevole all'uso di questo carburante.  
 (b) Qual è la probabilità dell'errore  $\beta$  del II tipo se il 75% dei votanti è favorevole a questa strategia antinquinamento?



4.50. Il periodo di garanzia per le batterie dei telefoni cellulari è di 400 ore se le procedure di ricarica sono eseguite correttamente. Viene eseguito uno studio su 2000 batterie, che rivela come 3 batterie si siano guastate prima delle 400 ore di garanzia previste. Questi risultati sperimentali supportano l'affermazione per cui meno dello 0.2% delle batterie si guasterà durante il periodo di garanzia, con le operazioni di carica rispettate? Usare una procedura di verifica di ipotesi con  $\alpha = 0.01$ .

4.51. Un articolo pubblicato in *Knee Surgery, Sports Traumatology, Arthroscopy* ("Arthroscopic Meniscal Repair with an Absorbable Screw: Results and Surgical Technique", 2005, Vol. 13 pp. 273-279) ha presentato dati che illustrano la guarigione di 25 su 37 rotture del menisco di lunghezza compresa fra 3 e 6 mm.

- (a) Calcolare un intervallo di confidenza bilaterale tradizionale sulla proporzione delle ferite che guariranno.  
 (b) Calcolare un limite di confidenza unilaterale tradizionale di livello 95% sulla proporzione delle ferite che guariranno.

4.52. Si considerino i dati dell'Esercizio 4.51. Calcolare l'intervallo di confidenza bilaterale di Agresti-Coull (Equazione 4.76) e confrontarlo con l'intervallo tradizionale trovato al punto (a) dell'esercizio precedente.

## ESERCIZI PER IL PARAGRAFO 4.8

4.53. Si consideri il problema riguardante la durata degli pneumatici descritto nell'Esercizio 4.35.

- (a) Costruire un intervallo di predizione al 95% sulla vita di un singolo pneumatico.  
 (b) Trovare un intervallo di tolleranza per la vita dello pneumatico che comprenda il 90% degli pneumatici della popolazione con il 95% di confidenza.

4.54. Si consideri la vita del dispositivo biomedico descritto nell'Esercizio 4.36.

- (a) Costruire un intervallo di predizione al 99% per la vita di un singolo dispositivo.

- (b) Trovare un intervallo di tolleranza per la vita del dispositivo che comprenda il 99% dei dispositivi della popolazione con il 90% di confidenza.

4.55. Si consideri la tensione di rottura dei diodi descritta nell'Esercizio 4.37.

- (a) Costruire un intervallo di predizione al 99% per la tensione di rottura di un singolo diodo.  
 (b) Trovare un intervallo di tolleranza per la tensione di rottura che comprenda il 99% dei diodi con il 99% di confidenza.

## ESERCIZI PER IL PARAGRAFO 4.10

4.56. Sia  $X$  il numero di incrinature riscontrate su una grande bobina di acciaio galvanizzato. Sono state ispezionate 75 bobine, ottenendo i seguenti valori di  $X$ .

Valori	1	2	3	4	5	6	7	8
Frequenza osservata	1	11	8	13	11	12	10	9

- (a) Appare appropriato assumere la distribuzione di Poisson con media 6.0 come modello di probabilità per questi dati? Usare  $\alpha = 0.01$ .

- (b) Calcolare il  $P$ -value per questo test.

4.57. Il numero di chiamate in arrivo a un centralino tra le 12 e le 13 durante i giorni feriali (da lunedì a venerdì) viene

## ESERCIZI PER IL PARAGRAFO 4.10

monitorato per 4 settimane (cioè 30 giorni). Sia  $X$  il numero di chiamate che arrivano nell'intervallo di un'ora. La frequenza osservata delle chiamate è stata registrata ed è riportata nella seguente tabella.

Valori	5	7	8	9	10
Frequenza osservata	4	4	4	5	1

Valori	11	12	13	14	15
Frequenza osservata	3	3	1	4	1

- (a) È appropriato assumere la distribuzione di Poisson come modello di probabilità per questi dati? Usare  $\alpha = 0.05$ .  
(b) Calcolare il  $P$ -value per questo test.

4.58 Sia  $X$  il numero di bottiglie sotto-riempite durante l'imbottigliamento, in un cartone composto da 12 bottiglie. Vengono ispezionati 80 cartoni; le relative osservazioni su  $X$  sono registrate nella seguente tabella.

Valori	0	1	2	3	4
Frequenza osservata	21	30	22	6	1

- (a) In base alle 80 osservazioni, la distribuzione binomiale può essere ritenuta un modello appropriato? Eseguire un test di adattamento con  $\alpha = 0.10$ .  
(b) Calcolare il  $P$ -value per questo test.

## ESERCIZI DI FINE CAPITOLO

4.59 Si consideri l'intervallo di confidenza per  $\mu$  con deviazione standard nota,  $\sigma$ :

$$\bar{x} - z_{\alpha_1}\sigma/\sqrt{n} \leq \mu \leq \bar{x} + z_{\alpha_2}\sigma/\sqrt{n}$$

dove  $\alpha_1 + \alpha_2 = \alpha$ . Porre  $\alpha = 0.05$  e trovare l'intervallo per  $\alpha_1 = \alpha_2 = \alpha/2 = 0.025$ . Trovare quindi l'intervallo per il caso  $\alpha_1 = 0.01$  e  $\alpha_2 = 0.04$ . Qual è l'intervallo più corto? Comporta qualche vantaggio avere un intervallo di confidenza "simmetrico"?

4.60 L'articolo "Mix Design for Optimal Strength Development of Fly Ash Concrete", apparso su *Cement and Concrete Research* (1989, Vol. 19, No. 4, pp. 634-640) studia la resistenza alla compressione del cemento quando questo è miscelato con altri elementi come silice, allumina, ferro, ossido di magnesio e altro. Le resistenze alla compressione per nove provini in condizione di asciutto al 28-esimo giorno sono le seguenti (in Mpa):

40.2	30.4	28.9	30.5	22.4
25.8	18.4	14.2	15.3	

- (a) Dato il grafico dei quantili dei dati di Figura 4.24, qual è l'assunzione logica riguardo la sottostante distribuzione dei dati?

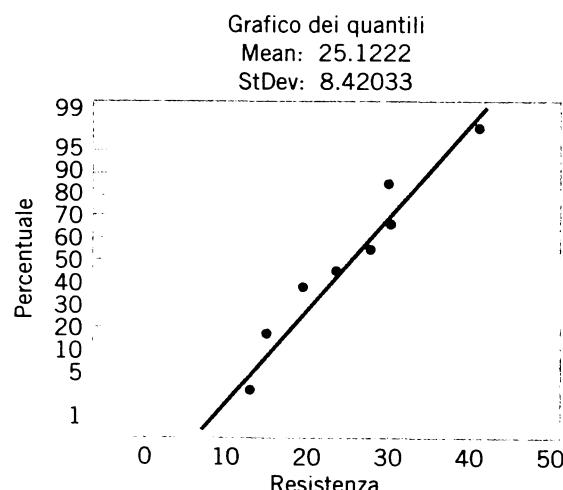


Figura 4.24 Grafico dei quantili normali per i dati dell'Esercizio 4.60.

- (b) Trovare un limite inferiore unilaterale di confidenza di livello 99% per la resistenza media alla compressione. Fornire un'interpretazione pratica di questo limite.  
(c) Trovare un intervallo bilaterale di confidenza al 98% per la resistenza media alla compressione. Fornire un'interpretazione pratica di questo intervallo e spiegare perché il limite inferiore dell'intervallo è o non è uguale a quello riscontrato al punto (b).

- (d) Trovare un limite superiore unilaterale di confidenza di livello 99% per la varianza della resistenza alla compressione. Fornire un'interpretazione pratica per questo limite.
- (e) Trovare un intervallo bilaterale di confidenza al 98% per la varianza della resistenza alla compressione. Fornire un'interpretazione pratica di questo intervallo e spiegare perché l'estremo superiore dell'intervallo è o non è uguale a quello riscontrato al punto (d).
- (f) Si supponga di avere scoperto che l'osservazione più grande, 40.2, è stata registrata non correttamente e che in realtà sia 20.4. A questo punto la media campionaria è  $\bar{x} = 22.9$  e la varianza campionaria è  $s^2 = 39.83$ . Usare questi nuovi valori e ripetere i punti (c) ed (e). Confrontare gli intervalli originali calcolati e gli intervalli calcolati con i valori corretti. Quanto influisce questo errore sui valori della media campionaria, della varianza campionaria e sull'ampiezza degli intervalli di confidenza bilaterali?
- (g) Si supponga, invece, di avere scoperto che l'osservazione più grande, 40.2, è corretta, ma che non lo sia l'osservazione 25.8, che in realtà è 24.8; conseguentemente, la media campionaria è  $\bar{x} = 25.0$  e la varianza campionaria è  $s^2 = 70.84$ . Usare questi nuovi valori e ripetere i punti (c) ed (e). Confrontare gli intervalli calcolati originariamente con quelli calcolati con il valore di osservazione corretto. Quanto influisce questo errore sui valori della media campionaria, della varianza campionaria e sull'ampiezza degli intervalli di confidenza bilaterali?
- (h) Si usino i risultati dei punti (f) e (g) per spiegare l'effetto dei valori registrati in modo errato sulle stime campionarie. Commentare questo effetto quando i valori errati sono vicini alla media campionaria oppure non lo sono.
- (i) Usando i dati originali, costruire un intervallo di predizione al 99% per la resistenza alla compressione di un singolo provino.
- (j) Trovare un intervallo di tolleranza per la resistenza alla compressione che includa il 95% della popolazione con il 99% di confidenza.

**4.61** Un ispettore per il controllo della qualità di dosatori di flusso per iniezioni in vena esegue una verifica di ipotesi per determinare se la velocità media di flusso è differente dalla velocità preimpostata, pari a 200 ml/h. In base a informazioni precedenti, la deviazione standard della velocità del flusso viene assunta uguale a 12 ml/h. Per ciascuna delle seguenti dimensioni campionarie e per  $\alpha = 0.05$  fissato, trovare la probabilità dell'errore del II tipo se la vera media è 205 ml/h.

- (a)  $n = 25$   
 (b)  $n = 60$   
 (c)  $n = 100$   
 (d) La probabilità dell'errore del II tipo aumenta o diminuisce se la dimensione campionaria aumenta? Spiegare la risposta.

**4.62** Si supponga che nell'esercizio precedente lo sperimentatore abbia ritenuto che  $\sigma$  fosse uguale a 14. Per ciascuna delle seguenti dimensioni campionarie e per  $\alpha = 0.05$  fissata, trovare la probabilità dell'errore del II tipo se la vera media è 205 ml/h.

- (a)  $n = 20$   
 (b)  $n = 50$   
 (c)  $n = 100$   
 (d) Confrontando la risposta con quella dell'esercizio precedente, la probabilità dell'errore del II tipo aumenta o diminuisce con l'aumento della deviazione standard? Spiegare la risposta.

 **4.63** Si sa che la durata, in ore, di un elemento di riscaldante usato in una fornace è distribuito in maniera approssimativamente normale. Viene selezionato un campione casuale di 15 elementi e si trova una durata media di 598.14 ore e una deviazione standard campionaria di 16.93 ore.

- (a) Al livello di significatività  $\alpha = 0.05$ , usare l'appropriata procedura a otto passi per verificare l'ipotesi  $H_0: \mu = 550$  contro  $H_1: \mu > 550$ . Completata la verifica dell'ipotesi, è possibile ritenere che la vera durata media dell'elemento riscaldante sia maggiore di 550 ore? Rispondere in modo chiaro e completo.  
 (b) Trovare il  $P$ -value della statistica test.  
 (c) Costruire un limite inferiore di confidenza al 95% sulla media e descrivere come questo intervallo può essere usato per verificare l'ipotesi alternativa del punto (a).  
 (d) Costruire un intervallo bilaterale di confidenza al 95% per la relativa varianza.

**4.64** Si consideri l'esperimento riguardante il dispositivo biomedico descritto nell'Esercizio 4.36.

- (a) Per la dimensione campionaria  $n = 15$ , i dati supportano l'asserzione che la deviazione standard della durata operativa è minore di 280 ore?  
 (b) Si supponga che la dimensione campionaria sia invece 51. Ripetere l'analisi eseguita al punto (a) usando  $n = 51$ .  
 (c) Confrontare le risposte e commentare l'influenza della dimensione campionaria sulle conclusioni ottenute ai punti (a) e (b).

**4.65** Un articolo pubblicato in *Food Testing and Analysis* ("Improving Reproducibility of refractometry Measurements of Fruit Juices", Vo. 4, No. 4, 1999, pp. 13-17) riporta i risultati di uno studio volto a misurare la concentrazione di zucchero (Brix) nel succo di mela. Tutte le letture sono state prese a 20 °C:

11.48	11.45	11.48	11.47	11.48
11.50	11.42	11.49	11.45	11.44
11.45	11.47	11.46	11.47	11.43
11.50	11.49	11.45	11.46	11.47

- (a) Verificare l'ipotesi  $H_0: \mu = 11.5$  contro l'ipotesi  $H_1: \mu \neq 11.5$  usando  $\alpha = 0.05$ . Trovare il  $P$ -value.
- (b) Calcolare la potenza del test nel caso in cui la vera media sia 11.4.
- (c) Quale dimensione campionaria sarebbe necessaria per rilevare una reale concentrazione media di 11.45, se si volesse che la potenza del test fosse almeno 0.9?
- (d) Spiegare come si potrebbe rispondere al punto (a) costruendo un intervallo di confidenza bilaterale per la concentrazione di zucchero media.
- (e) Vi sono indizi che supportino l'assunzione di una distribuzione normale per la concentrazione?

**4.66** Un articolo pubblicato in *British American Journal* presenta una ricerca tramite la quale si è scoperto che la nefrolitotomia percutanea (PN) ha avuto successo nella rimozione dei calcoli renali con 289 pazienti su 350. Il metodo tradizionale ha un'efficacia del 78%.

- (a) Vi sono indizi per sostenere che il tasso di successo della PN è maggiore di quello della procedura tradizionale? Trovare il  $P$ -value.
- (b) Spiegare come si potrebbe rispondere al punto (a) costruendo un intervallo di confidenza.

**4.67** Il seguente elenco riporta il numero annuale di terremoti di magnitudo 7.0 o maggiore, a partire dal 1900. (*Fonte:* U.S. Geological Survey, National Earthquake Information Center, Golden, CO).

1900	13	1928	22	1956	15	1984	8
1901	14	1929	19	1957	34	1985	15
1902	8	1930	13	1958	10	1986	6
1903	10	1931	26	1959	15	1987	11
1904	16	1932	13	1960	22	1988	8
1905	26	1933	14	1961	18	1989	7
1906	32	1934	22	1962	15	1990	18
1907	27	1935	24	1963	20	1991	16
1908	18	1936	21	1964	15	1992	13
1909	32	1937	22	1965	22	1993	12
1910	36	1938	26	1966	19	1994	13
1911	24	1939	21	1967	16	1995	20
1912	22	1940	23	1968	30	1996	15
1913	23	1941	24	1969	27	1997	16
1914	22	1942	27	1970	29	1998	12
1915	18	1943	41	1971	23	1999	18
1916	25	1944	31	1972	20	2000	15
1917	21	1945	27	1973	16	2001	16
1918	21	1946	35	1974	21	2002	13
1919	14	1947	26	1975	21	2003	15
1920	8	1948	28	1976	25	2004	15
1921	11	1949	36	1977	16	2005	11

1922	14	1950	39	1978	18	2006	11
1923	23	1951	21	1979	15	2007	18
1924	18	1952	17	1980	18	2008	12
1925	17	1953	22	1981	14	2009	15
1926	19	1954	17	1982	10		
1927	20	1955	19	1983	15		

- (a) Usare un software statistico per riassumere questi dati in una distribuzione di frequenze. Verificare l'ipotesi che il numero annuale di terremoti di magnitudo 7.0 o più segue una distribuzione di Poisson per  $\alpha = 0.05$ .
- (b) Trovare il  $P$ -value per il test.

 **4.68.** Un articolo apparso su *The Engineer* ("Redesign for Suspect Wiring", June 1990) riporta i risultati di uno studio eseguito sugli errori di cablaggio che avvengono su aerei commerciali e che possono fornire informazioni sbagliate al personale di bordo. Si ritiene che un simile errore di cablaggio possa essere stato la causa di un incidente aereo che ha coinvolto un velivolo della British Midland Airways nel gennaio 1989, in cui è accertato che il pilota ha spento il motore sbagliato. Su 1660 aerei selezionati in maniera casuale, 8 hanno presentato errori di cablaggio che avrebbero potuto fornire informazioni sbagliate al personale di bordo.

- (a) Trovare un intervallo bilaterale di confidenza al 99% sulla proporzione di aerei in cui sono presenti errori di cablaggio.
- (b) Si supponga di usare l'informazione di questo esempio per fornire una stima preliminare di  $p$ . Quanto grande dovrebbe essere un campione per produrre una stima di  $p$  per la quale siamo dati confidenti al 99% che differisce dal vero valore al più di 0.008?
- (c) Si supponga di non possedere una stima preliminare di  $p$ . Quanto grande deve essere un campione se si ricerca una confidenza almeno del 99% sul fatto che la proporzione campionaria differisce dalla vera proporzione al più di 0.008, indipendentemente dal vero valore di  $p$ ?
- (d) Commentare il vantaggio di avere un'informazione preliminare ai fini del calcolo della dimensione campionaria necessaria.

**4.69** Si consideri il seguente output di Minitab

Test and CI for One Proportion							
Test of p = 0.2 vs p < 0.2							
Sample	X	N	Sample p	95%		Z-Value	P-Value
				Upper	Bound		
1	146	850	0.171765	0.193044	-2.06	-0.020	

- (a) Si tratta di un test unilaterale o bilaterale?
- (b) Il test è stato condotto usando l'approssimazione normale per la binomiale? È stata una scelta adeguata?
- (c) Si può rifiutare l'ipotesi nulla al livello 0.05? E al livello 0.01?
- (d) Si può rifiutare l'ipotesi nulla  $H_0: p = 0.2$  contro  $H_1: p \neq 0.2$ , utilizzando  $\alpha = 0.05$ ?
- (e) Costruire un limite di confidenza unilaterale approssimato di livello 90% per  $p$ .



# Processo decisionale per due campioni

---

## CALCESTRUZZO O GRANITO LIQUIDO?

Un nuovo materiale da costruzione, chiamato granito liquido, sembra offrire vantaggi significativi rispetto al calcestruzzo; per gli ingegneri edili ciò rappresenta una nuova sfida, che richiederà l'adozione di un processo decisionale per due campioni al fine di scegliere la soluzione migliore per componenti strutturali come muri, pilastri e architravi, finora costruiti in calcestruzzo.

Anche se il calcestruzzo non è infiammabile, è sensibile agli effetti di un intenso riscaldamento e presenta alcuni limiti di resistenza e integrità quando viene sottoposto ad alte temperature. Il cemento che funge da legante nel calcestruzzo quando è asciutto forma una sostanza rocciosa grazie ai legami con le molecole d'acqua; un riscaldamento intenso fa sì che il cemento si disidrati e tenda a ritornare allo stato di polvere di cemento. Questa disidratazione riduce la resistenza e il modulo di elasticità del calcestruzzo. Inoltre l'acqua rilasciata sotto forma di vapore, a volte in modo violento, causa delle cricche e altri danni strutturali fisici. Il calcestruzzo non brucia, ma può dunque cedere strutturalmente a causa degli effetti del riscaldamento. È da sottolineare che l'interesse verso gli effetti del calore sul cemento si sono grandemente intensificati dopo i fatti dell'11 settembre 2001.

Il granito liquido è molto meno soggetto a cedimenti strutturali prodotti da cause termiche. Essendo in grado di reggere molto più a lungo al calore rispetto al cemento, se utilizzato per costruire le strutture portanti garantisce più tempo per evacuare gli edifici che dovessero andare in fiamme.

Il granito liquido è anche una sostanza più ecologica del calcestruzzo. Innanzitutto contiene solo il 30% del cemento presente in quest'ultimo, e poiché la produzione di cemento è responsabile di circa il 5% delle emissioni di anidride carbonica imputabili all'uomo, la sua impronta carbonica sull'ambiente è inferiore a quella del calcestruzzo. In secondo luogo, l'impronta del granito liquido si riduce ulteriormente grazie all'impiego di una percentuale fra il 30 e il 70% di materiali industriali riciclati, con conseguente riduzione della quantità di energia necessaria alla sua produzione.

Gli ingegneri, oggi, si trovano quindi ad assumere decisioni basate sul confronto fra due materiali: il calcestruzzo e il granito liquido.

## CONTENUTI DEL CAPITOLO

5.1 INTRODUZIONE	5.5.1 Verifica di ipotesi sul rapporto tra due varianze
5.2 INFERENZA SULLE MEDIE DI DUE POPOLAZIONI CON VARIANZE NOTE	5.5.2 Intervallo di confidenza per il rapporto tra due varianze
5.2.1 Verifica di ipotesi sulla differenza tra medie con varianze note	5.6 INFERENZA SULLE PROPORZIONI DI DUE POPOLAZIONI
5.2.2 Errore del II tipo e scelta della dimensione campionaria	5.6.1 Verifica di ipotesi sull'uguaglianza di due proporzioni binomiali
5.2.3 Intervallo di confidenza per la differenza tra medie con varianze note	5.6.2 Errore del II tipo e scelta della dimensione campionaria
5.3 INFERENZA SULLE MEDIE DI DUE POPOLAZIONI CON VARIANZE INCOGNITE	5.6.3 Intervallo di confidenza per la differenza tra proporzioni binomiali
5.3.1 Verifica di ipotesi sulla differenza tra medie	5.7 TABELLE RIASSUNTIVE DELLE PROCEDURE DI INFERENZA PER DUE CAMPIONI
5.3.2 Errore del II tipo e scelta della dimensione campionaria	5.8 CASO DI PIÙ DI DUE CAMPIONI
5.3.3 Intervallo di confidenza per la differenza tra medie	5.8.1 Esperimento completamente casualizzato e analisi della varianza
5.4 TEST t ACCOPPIATO	5.8.2 Esperimento a blocchi
5.5 INFERENZA SUL RAPPORTO TRA LE VARIANZE DI DUE POPOLAZIONI NORMALI	

## OBIETTIVI DI APPRENDIMENTO

Dopo uno studio attento di questo capitolo si sarà in grado di:

1. impostare esperimenti di comparazione in cui si utilizza la verifica di ipotesi per due campioni
2. eseguire verifiche di ipotesi e costruire intervalli di confidenza sulla differenza tra le medie di due distribuzioni normali
3. eseguire verifiche di ipotesi e costruire intervalli di confidenza sul rapporto tra le varianze di due distribuzioni normali
4. eseguire verifiche di ipotesi e costruire intervalli di confidenza sulla differenza tra le proporzioni di due popolazioni
5. calcolare la potenza e l'errore del II tipo e assumere decisioni sulla scelta della dimensione campionaria per le verifiche di ipotesi e gli intervalli di confidenza
6. comprendere come si può utilizzare l'analisi della varianza in un esperimento volto a confrontare fra loro più medie
7. valutare l'adeguatezza di un modello ANOVA con grafici dei residui
8. comprendere il principio d'uso dei blocchi e come può essere sfruttato per isolare l'effetto dei fattori di disturbo in un esperimento
9. progettare e condurre esperimenti a blocchi completamente casualizzati

## 5.1 INTRODUZIONE

Nel precedente capitolo abbiamo presentato le verifiche di ipotesi e gli intervalli di confidenza per un singolo parametro della popolazione (la media  $\mu$ , la varianza  $\sigma^2$  o una proporzione  $p$ ). Questo capitolo estende quei risultati al caso di due popolazioni indipendenti.

La situazione generale è mostrata in Figura 5.1. La popolazione 1 ha media  $\mu_1$  e varianza  $\sigma_1^2$ , la popolazione 2 ha media  $\mu_2$  e varianza  $\sigma_2^2$ . Le inferenze si baseranno su due campioni casuali rispettivamente di dimensione  $n_1$  e  $n_2$ . In altri termini,  $X_{11}, X_{12}, \dots, X_{1n_1}$ , è un campione casuale di  $n_1$  osservazioni sulla popolazione 1, e  $X_{21}, X_{22}, \dots, X_{2n_2}$ , è un campione casuale di  $n_2$  osservazioni sulla popolazione 2.

## 5.2 INFERENZA SULLE MEDIE DI DUE POPOLAZIONI CON VARIANZE NOTE

In questo paragrafo consideriamo le inferenze statistiche sulla differenza tra le medie,  $\mu_1 - \mu_2$ , delle popolazioni mostrate in Figura 5.1, mentre le varianze  $\sigma_1^2$  e  $\sigma_2^2$  sono note. Le assunzioni adottate in questo paragrafo sono elencate nel seguente riquadro.

### Assunzioni

1.  $X_{11}, X_{12}, \dots, X_{1n_1}$  costituiscono un campione casuale di dimensione  $n_1$  estratto dalla popolazione 1.
2.  $X_{21}, X_{22}, \dots, X_{2n_2}$  costituiscono un campione casuale di dimensione  $n_2$  estratto dalla popolazione 2.
3. Le due popolazioni rappresentate da  $X_1$  e  $X_2$  sono indipendenti.
4. Entrambe le popolazioni sono normali, o, in caso contrario, sono valide le condizioni del teorema limite centrale.

Uno stimatore puntuale naturale di  $\mu_1 - \mu_2$  è la differenza tra le medie campionarie,  $\bar{X}_1 - \bar{X}_2$ . Basandoci sulle proprietà dei valori attesi viste nel Capitolo 3, abbiamo

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

e la varianza di  $\bar{X}_1 - \bar{X}_2$  risulta

$$V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Grazie alle assunzioni e ai precedenti risultati, possiamo affermare quanto segue

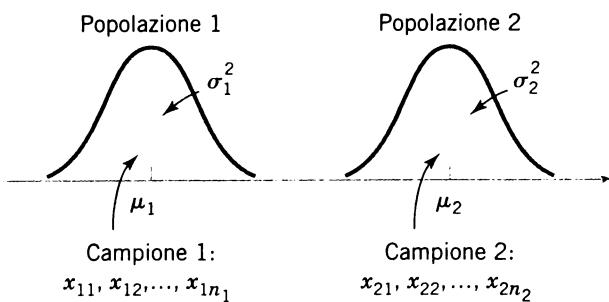


Figura 5.1 Due popolazioni indipendenti.

Sotto le precedenti assunzioni, la quantità

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (5.1)$$

ha una distribuzione normale standard,  $N(0, 1)$ .

Questo risultato verrà usato per costruire le verifiche di ipotesi e gli intervalli di confidenza su  $\mu_1 - \mu_2$ . Essenzialmente, possiamo pensare alla differenza  $\mu_1 - \mu_2$  come a un parametro  $\theta$ , il cui stimatore è  $\hat{\Theta} = \bar{X}_1 - \bar{X}_2$  con varianza  $\sigma_{\hat{\Theta}}^2 = \sigma_1^2/n_1 + \sigma_2^2/n_2$ . Se  $\theta_0$  è il valore di ipotesi nulla specificato per  $\theta$ , la statistica test sarà  $= (\hat{\Theta} - \theta_0)/\sigma_{\hat{\Theta}}$ . Si noti come questa sia simile alla statistica test per una media singola usata nel Capitolo 4.

### 5.2.1 Verifica di ipotesi sulla differenza tra medie con varianze note

Consideriamo ora la verifica di ipotesi sulla differenza tra le medie,  $\mu_1 - \mu_2$ , delle due popolazioni di Figura 5.1. Supponiamo di essere interessati a verificare che tale differenza sia uguale a un valore specifico  $\Delta_0$ . L'ipotesi nulla sarà dunque  $H_0: \mu_1 - \mu_2 = \Delta_0$ . Ovviamente, in molti casi imporremo  $\Delta_0 = 0$  per verificare l'uguaglianza tra le due medie (vale a dire,  $H_0: \mu_1 = \mu_2$ ). La statistica test appropriata verrebbe trovata sostituendo nell'Equazione (5.1)  $\mu_1 - \mu_2$  con  $\Delta_0$ , e questa statistica test avrebbe una distribuzione normale standard sotto l'ipotesi  $H_0$ . Supponiamo che l'ipotesi alternativa sia  $H_1: \mu_1 - \mu_2 \neq \Delta_0$ . Ora, un valore campionario di  $\bar{x}_1 - \bar{x}_2$  considerevolmente differente da  $\Delta_0$  è un indizio del fatto che  $H_1$  è vera. Siccome  $Z_0$  ha la distribuzione  $N(0, 1)$  quando  $H_0$  è vera, dovremmo calcolare il  $P$ -value come somma delle probabilità oltre il valore della statistica test  $|z_0|$  e  $-|z_0|$  nella distribuzione normale standard; ovvero,  $P = 2[1 - \Phi(|z_0|)]$ . È esattamente ciò che abbiamo fatto nel problema della verifica di ipotesi per singolo campione del Paragrafo 4.4.1. Se volessimo eseguire un test per un livello di significatività fissato, dovremmo prendere  $-z_{\alpha/2}$  e  $z_{\alpha/2}$  come limiti della regione critica, come abbiamo fatto nella verifica di ipotesi per singolo campione. Ciò darebbe un test con livello di significatività  $\alpha$ . I  $P$ -value o le regioni critiche per le alternative unilaterali sarebbero ricavate in modo analogo. Riassumiamo in

maniera formale questi risultati per la **verifica di ipotesi per due campioni** nel seguente riquadro.

<b>Verifica di ipotesi sulla differenza tra medie con varianze note</b>		
<b>Ipotesi nulla:</b>	$H_0: \mu_1 - \mu_2 = \Delta_0$	
<b>Statistica test:</b>	$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	
<b>Ipotesi alternativa</b>	<b>P-value</b>	<b>Criterio di rifiuto</b>
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	Probabilità a destra di $ z_0 $ e a sinistra di $- z_0 $ , $P = 2[1 - \Phi( z_0 )]$	$z_0 > z_{\alpha/2}$ o $z_0 < -z_{\alpha/2}$
$H_1: \mu_1 - \mu_2 > \Delta_0$	Probabilità a destra di $z_0$ , $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1: \mu_1 - \mu_2 < \Delta_0$	Probabilità a sinistra di $z_0$ , $P = \Phi(z_0)$	$z_0 < -z_\alpha$

### ESEMPIO 5.1 Tempo di asciugatura di una vernice

Il responsabile dello sviluppo di prodotto è interessato a ridurre il tempo necessario a una vernice di imprimitura per asciugarsi. Vengono sottoposte a test due formulazioni della vernice: la formulazione 1 è il prodotto standard, la formulazione 2 contiene un nuovo ingrediente asciugante che dovrebbe ridurre il tempo di asciugatura. In base all'esperienza si sa che la deviazione standard del tempo di asciugatura è 8 min, e questa variabilità intrinseca non dovrebbe essere influenzata dall'aggiunta del nuovo ingrediente. Vengono verniciati 10 provini con la vernice di formulazione 1, e altri 10 vengono verniciati con quella di formulazione 2; i 20 provini sono verniciati in ordine casuale. I tempi medi di asciugatura dei due campioni sono rispettivamente  $\bar{x}_1 = 121$  min e  $\bar{x}_2 = 112$  min. Quali conclusioni può trarre il responsabile dello sviluppo di prodotto riguardo l'efficacia del nuovo ingrediente, usando  $\alpha = 0.05$ ?

Applichiamo la procedura a sette passaggi per la risoluzione di questo problema, come segue.

- 1. Parametro di interesse:** la quantità di interesse è la differenza tra le medie dei tempi di asciugatura,  $\mu_1 - \mu_2$ , e  $\Delta_0 = 0$ .
- 2. Ipotesi sulla  $H_0$ :**  $\mu_1 - \mu_2 = 0$ , ovvero  $H_0: \mu_1 = \mu_2$ .
- 3. Ipotesi alternativa  $H_1$ :**  $\mu_1 > \mu_2$ . Vogliamo rifiutare  $H_0$  se il nuovo ingrediente riduce il tempo medio di asciugatura.
- 4. Statistica test:** la statistica test è

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

dove  $\sigma_1^2 = \sigma_2^2 = (8)^2 = 64$  e  $n_1 = n_2 = 10$ .

**5. Rifiutare  $H_0$  se:** si rifiuta  $H_0: \mu_1 = \mu_2$  se il  $P$ -value è minore di 0.05.

**6. Calcoli:** siccome  $\bar{x}_1 = 121$  min e  $\bar{x}_2 = 112$  min, la statistica test è

$$z_0 = \frac{121 - 112}{\sqrt{\frac{(8)^2}{10} + \frac{(8)^2}{10}}} = 2.52$$

**7. Conclusioni:** siccome il  $P$ -value è  $P = 1 - \Phi(2.52) = 0.0059$ , rifiutiamo  $H_0$ . Essendo il  $P$ -value uguale a 0.0059, l'ipotesi nulla sarebbe rifiutata a **qualsiasi** livello di significatività  $\alpha \geq 0.0059$ . La conclusione pratica ingegneristica è che l'aggiunta del nuovo ingrediente alla vernice riduce significativamente il tempo di asciugatura.

### 5.2.2 Errore del II tipo e scelta della dimensione campionaria

Supponiamo che l'ipotesi nulla  $H_0: \mu_1 - \mu_2 = \Delta_0$  sia falsa e che la differenza vera tra le medie sia  $\mu_1 - \mu_2 = \Delta$ , dove  $\Delta > \Delta_0$ . Possiamo trovare le formule per la dimensione campionaria necessaria per ottenere un valore specifico della probabilità  $\beta$  dell'errore del II tipo per una data differenza tra le medie  $\Delta$  e per un dato livello di significatività  $\alpha$ .

#### Dimensione campionaria per ipotesi alternativa bilaterale sulla differenza tra medie, con varianze note, quando $n_1 = n_2$

Per un livello di significatività  $\alpha$  con ipotesi alternativa bilaterale, la dimensione campionaria  $n_1 = n_2 = n$  richiesta per rilevare una differenza vera tra le medie pari a  $\Delta$  con una potenza almeno di  $1 - \beta$  è

$$n \simeq \frac{(z_{\alpha/2} + z_{\beta})^2 (\sigma_1^2 + \sigma_2^2)}{(\Delta - \Delta_0)^2} \quad (5.2)$$

Se  $n$  non è un intero si arrotonda la dimensione campionaria per eccesso all'intero immediatamente superiore.

Questa approssimazione è valida quando  $\Phi(-z_{\alpha/2} - (\Delta - \Delta_0)\sqrt{n}/\sqrt{\sigma_1^2 + \sigma_2^2})$  è piccola in confronto a  $\beta$ .

**Dimensione campionaria per ipotesi alternativa unilaterale  
sulla differenza tra medie, con varianze note, quando  $n_1 = n_2$**

Per un livello di significatività  $\alpha$  con ipotesi alternativa unilaterale, la dimensione campionaria  $n_1 = n_2 = n$  necessaria per rilevare una differenza vera tra le medie pari a  $\Delta$  ( $\neq \Delta_0$ ) con una potenza almeno di  $1 - \beta$  è

$$n = \frac{(z_\alpha + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{(\Delta - \Delta_0)^2} \quad (5.3)$$

Il modo in cui si ricavano le Equazioni (5.2) e (5.3) segue da vicino il caso di singolo campione visto nel Paragrafo 4.4.3. Per esempio, per ottenere l'Equazione (5.2) scriviamo innanzitutto l'espressione per l'errore  $\beta$  per l'alternativa bilaterale, che è

$$\beta = \Phi\left(z_{\alpha/2} - \frac{\Delta - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) - \Phi\left(-z_{\alpha/2} - \frac{\Delta - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right)$$

dove  $\Delta$  è la differenza vera tra le medie di interesse, e  $\Delta_0$  è specificato nell'ipotesi nulla. Allora, seguendo una procedura simile a quella usata per ottenere l'Equazione (4.26), si può ricavare l'espressione per  $n$  per il caso in cui  $n_1 = n_2 = n$ .

**ESEMPIO 5.2**  
**Tempo**  
**di asciugatura**  
**di una vernice**

Per illustrare l'impiego delle precedenti equazioni ai fini della determinazione della dimensione campionaria, si consideri la situazione descritta nell'Esempio 5.1 e si supponga che se la differenza vera tra i tempi di asciugatura vale 10 min, vogliamo rilevarla con una probabilità uguale almeno a 0.90. Sotto l'ipotesi nulla,  $\Delta_0 = 0$ . Abbiamo un'ipotesi alternativa unilaterale con  $\Delta = 10$ ,  $\alpha = 0.05$  (per cui  $z_\alpha = z_{0.05} = 1.645$ ), e siccome la potenza è 0.9,  $\beta = 0.10$  (per cui  $z_\beta = z_{0.10} = 1.28$ ). Pertanto, possiamo trovare la dimensione campionaria richiesta grazie all'Equazione (5.3), come segue

$$n = \frac{(z_\alpha + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{(\Delta - \Delta_0)^2} = \frac{(1.645 + 1.28)^2[(8)^2 + (8)^2]}{(10 - 0)^2}$$

$$\simeq 11$$

### 5.2.3 Intervallo di confidenza per la differenza tra medie con varianze note

L'intervallo di confidenza al  $100(1 - \alpha)\%$  per la differenza tra due medie,  $\mu_1 - \mu_2$ , quando le varianze sono note, può essere trovato direttamente dai risultati dati poco sopra. Ricordiamo che  $X_{11}, X_{12}, \dots, X_{1n}$  è un campione casuale di  $n_1$  osservazioni estratto dalla prima popolazione, e  $X_{21}, X_{22}, \dots, X_{2n_2}$  è un campione casuale di  $n_2$  osservazioni estratto dalla seconda popolazione. La differenza tra le medie campionarie,  $\bar{X}_1 - \bar{X}_2$ , è uno stimatore puntuale di  $\mu_1 - \mu_2$ , e

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

ha una distribuzione normale standard se le due popolazioni sono normali, o una distribuzione approssimativamente normale standard se si applicano le condizioni del teorema limite centrale. Ciò implica che

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

ovvero

$$P\left[-z_{\alpha/2} \leq \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2}\right] = 1 - \alpha$$

Questa espressione può essere riscritta come

$$P\left(\bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

Perciò, l'intervallo di confidenza al  $100(1 - \alpha)\%$  per  $\mu_1 - \mu_2$  è definito come segue.

### Intervallo di confidenza sulla differenza tra medie, varianze note

Se  $\bar{x}_1$  e  $\bar{x}_2$  sono le medie di campioni casuali indipendenti di dimensione  $n_1$  e  $n_2$  estratti da popolazioni con varianze note, rispettivamente uguali a  $\sigma_1^2$  e  $\sigma_2^2$ , un intervallo di confidenza al  $100(1 - \alpha)\%$  per  $\mu_1 - \mu_2$  è

$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (5.4)$$

dove  $z_{\alpha/2}$  è il punto percentuale superiore  $100 \alpha/2$  e  $-z_{\alpha/2}$  è il punto percentuale inferiore  $100 \alpha/2$  della distribuzione normale standard della Tavola I in Appendice A.

Il livello di confidenza  $1 - \alpha$  è esatto quando le popolazioni sono normali. Per popolazioni non normali, il livello di confidenza è valido approssimativamente per dimensioni campionarie elevate.

**ESEMPIO 5.3**  
Longheroni  
per aerei

Sono state eseguite verifiche di resistenza alla trazione su longheroni composti da due differenti qualità di alluminio, usati nella costruzione dell'ala di un aereo per trasporti commerciali. Grazie all'esperienza passata sui processi di produzione dei longheroni e alla procedura di test, si assumono note le deviazioni standard della resistenza alla trazione. I dati ottenuti sono mostrati in Tabella 5.1. Se  $\mu_1$  e  $\mu_2$  indicano le vere resistenze medie alla trazione per le due qualità di alluminio, possiamo trovare un intervallo di confidenza di livello 90% per la differenza tra le resistenze medie  $\mu_1$  e  $\mu_2$  nel seguente modo:

$$l = \bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 87.6 - 74.5 - 1.645 \sqrt{\frac{(1.0)^2}{10} + \frac{(1.5)^2}{12}} \\ = 13.1 - 0.88 = 12.22 \text{ kg/mm}^2$$

$$u = \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 87.6 - 74.5 + 1.645 \sqrt{\frac{(1.0)^2}{10} + \frac{(1.5)^2}{12}} \\ = 13.1 + 0.88 = 13.98 \text{ kg/mm}^2$$

Pertanto, l'intervallo di confidenza al 90% per la differenza tra le resistenze medie alla trazione è

$$12.22 \text{ kg/mm}^2 \leq \mu_1 - \mu_2 \leq 13.98 \text{ kg/mm}^2$$

**Conclusione pratica:** l'intervallo di confidenza non comprende lo zero, il che comporta che la resistenza media dell'alluminio di qualità 1 ( $\mu_1$ ) sia maggiore della resistenza media dell'alluminio di qualità 2 ( $\mu_2$ ). In effetti, possiamo affermare di essere confidenti al 90% che la resistenza media alla trazione dell'alluminio di qualità 1 supera quella dell'alluminio di qualità 2 per un valore compreso fra 12.22 e 13.98 kg/mm<sup>2</sup>.

**Tabella 5.1** Risultati per la verifica della resistenza alla trazione dei longheroni di alluminio.

Qualità alluminio longherone	Dimensione campionaria	Resistenza media campionaria alla traz. (kg/mm <sup>2</sup> )	Deviazione Standard (kg/mm <sup>2</sup> )
1	$n_1 = 10$	$\bar{x}_1 = 87.6$	$\sigma_1 = 1.0$
2	$n_2 = 12$	$\bar{x}_2 = 74.5$	$\sigma_2 = 1.5$

### Limiti di confidenza unilaterali

Per trovare un limite inferiore di confidenza al livello  $100(1 - \alpha)\%$  per  $\mu_1 - \mu_2$ , con  $\sigma^2$  nota, sostituiamo semplicemente  $-z_{\alpha/2}$  con  $-z_\alpha$  nel limite inferiore dell'Equazione (5.4) e poniamo il limite superiore a  $\infty$ . Analogamente, per trovare un limite superiore di confidenza al  $100(1 - \alpha)\%$  per  $\mu_1 - \mu_2$ , con  $\sigma^2$  nota, sostituiamo  $z_{\alpha/2}$  con  $z_\alpha$  nel limite superiore e poniamo il limite inferiore a  $-\infty$ .

### Scelta della dimensione campionaria

Se le deviazioni standard  $\sigma_1$  e  $\sigma_2$  sono note (almeno approssimativamente) e le due dimensioni campionarie  $n_1$  e  $n_2$  sono uguali ( $n_1 = n_2 = n$ , per esempio), possiamo determinare la dimensione campionaria necessaria a che l'errore nella stima di  $\mu_1 - \mu_2$  con  $\bar{x}_1 - \bar{x}_2$  sia minore di  $E$  con confidenza  $100(1 - \alpha)\%$ . La dimensione necessaria per i campioni estratti da ciascuna popolazione è riportata di seguito.

**Dimensione campionaria per un dato valore  $E$  dell'errore  
sulla differenza tra medie, con varianze note, quando  $n_1 = n_2$**

Se  $\bar{x}_1$  e  $\bar{x}_2$  sono usate come stime rispettivamente di  $\mu_1$  e  $\mu_2$ , possiamo essere confidenzi al livello  $100(1 - \alpha)\%$  che l'errore  $|(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)|$  non supererà un valore specificato  $E$  quando la dimensione campionaria  $n_1 = n_2 = n$  è

$$n = \left( \frac{z_{\alpha/2}}{E} \right)^2 (\sigma_1^2 + \sigma_2^2) \quad (5.5)$$

Si ricordi di arrotondare per eccesso nel caso in cui  $n$  non sia intero; ciò assicurerà che il livello di confidenza non scenda sotto il livello  $100(1 - \alpha)\%$ .

### 5.3 INFERENZA SULLE MEDIE DI DUE POPOLAZIONI CON VARIANZE INCOGNITE

Estendiamo i risultati del precedente paragrafo sulla differenza tra le medie delle due distribuzioni di Figura 5.1 quando sono incognite le varianze  $\sigma_1^2$  e  $\sigma_2^2$  di entrambe le distribuzioni. Se le dimensioni campionarie  $n_1$  e  $n_2$  superano il valore 30, è possibile usare le procedure per distribuzione normale del Paragrafo 5.2. Tuttavia, nel caso di piccoli campioni, assumiamo che le popolazioni siano normalmente distribuite, e baseremo le nostre verifiche di ipotesi e i nostri intervalli di confidenza sulla distribuzione  $t$ . Ciò segue esattamente la falsariga del caso di inferenza sulla media di un singolo campione con varianza incognita.

#### 5.3.1 Verifica di ipotesi sulla differenza tra medie

Consideriamo ora le verifiche di ipotesi sulla differenza  $\mu_1 - \mu_2$ , tra le medie di due distribuzioni normali dove le varianze  $\sigma_1^2$  e  $\sigma_2^2$  sono incognite. Per verificare queste ipotesi verrà usata una statistica  $t$ . Come abbiamo notato poco sopra e nel Paragrafo 4.6, l'assunzione di normalità è necessaria per sviluppare la procedura di verifica, ma moderati scostamenti dalla normalità non influiscono negativamente sulla procedura. Bisogna trattare due differenti casi. Nel primo caso assumiamo che le varianze delle distribuzioni normali siano incognite ma uguali, cioè  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ . Nel secondo caso assumiamo che  $\sigma_1^2$  e  $\sigma_2^2$  siano incognite e non necessariamente uguali.

Caso 1:  $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Supponiamo di avere due popolazioni normali indipendenti con medie  $\mu_1$  e  $\mu_2$  incognite, e varianze incognite ma uguali, cioè  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ . Vogliamo verificare le ipotesi

$$\begin{aligned} H_0: \mu_1 - \mu_2 &= \Delta_0 \\ H_1: \mu_1 - \mu_2 &\neq \Delta_0 \end{aligned} \quad (5.6)$$

Sia  $X_{11}, X_{12}, \dots, X_{1n_1}$ , un campione casuale di  $n_1$  osservazioni estratto dalla prima popolazio-

ne, e  $X_{21}, X_{22}, \dots, X_{2n_2}$  un campione casuale di  $n_2$  osservazioni estratto dalla seconda popolazione. Siano  $\bar{X}_1, \bar{X}_2, S_1^2, S_2^2$  le rispettive medie e varianze campionarie. Ora, il valore atteso della differenza tra le medie campionarie  $\bar{X}_1 - \bar{X}_2$  è  $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$ , quindi  $\bar{X}_1 - \bar{X}_2$  è uno stimatore non distorto della differenza tra le medie. La varianza di  $\bar{X}_1 - \bar{X}_2$  è

$$V(\bar{X}_1 - \bar{X}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)$$

Sembra ragionevole combinare le due varianze campionarie  $S_1^2$  e  $S_2^2$  per formare uno stimatore di  $\sigma^2$ . Lo **stimatore pooled**, o *raggruppato*, di  $\sigma^2$  è definito come segue.

### Stimatore pooled di $\sigma^2$

**Lo stimatore pooled** di  $\sigma^2$ , indicato con  $S_p^2$ , è definito da

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \quad (5.7)$$

È facile vedere che lo stimatore pooled  $S_p^2$  può essere scritto come

$$\begin{aligned} S_p^2 &= \frac{n_1 - 1}{n_1 + n_2 - 2} S_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2^2 \\ &= wS_1^2 + (1 - w)S_2^2 \end{aligned}$$

dove  $0 < w \leq 1$ . Quindi  $S_p^2$  è una **media pesata** delle due varianze campionarie  $S_1^2$  e  $S_2^2$ , dove i pesi  $w$  e  $1 - w$  dipendono dalle due dimensioni campionarie  $n_1$  e  $n_2$ . Ovviamente, se  $n_1 = n_2 = n$ ,  $w = 0.5$  e  $S_p^2$  è semplicemente la media aritmetica di  $S_1^2$  e  $S_2^2$ . Se  $n_1 = 10$  ed  $n_2 = 20$  (per esempio),  $w = 0.32$  e  $1 - w = 0.68$ . Il primo campione apporta  $n_1 - 1$  gradi di libertà a  $S_p^2$  e il secondo campione apporta  $n_2 - 1$  gradi di libertà. Perciò  $S_p^2$  ha  $n_1 + n_2 - 2$  gradi di libertà.

Ora, sappiamo che

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

ha una distribuzione  $N(0, 1)$ . Sostituendo  $\sigma$  con  $S_p$  otteniamo quanto affermato nel seguente riquadro.

Date le assunzioni di questo paragrafo, la quantità

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (5.8)$$

ha una distribuzione  $t$  con  $n_1 + n_2 - 2$  gradi di libertà.

L'uso di questo risultato per verificare le ipotesi nell'Equazione (5.6) è immediato: semplicemente si sostituisce  $\mu_1 - \mu_2$  con  $\Delta_0$ , e la **statistica test** risultante ha una distribuzione  $t$  con  $n_1 + n_2 - 2$  gradi di libertà sotto l'ipotesi  $H_0: \mu_1 - \mu_2 = \Delta_0$ . La posizione della regione critica per entrambe le alternative, bilaterale e unilaterale, ricalca quella del caso a singolo campione. Questa procedura è spesso chiamata **test  $t$  pooled**.

**Caso 1: Verifica di ipotesi sulla differenza tra medie di due distribuzioni normali, varianze incognite e uguali**

Ipotesi nulla:	$H_0: \mu_1 - \mu_2 = \Delta_0$	
Statistica test:	$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	
Ipotesi alternativa	P-value	Criterio di rifiuto
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	Somma probabilità a destra di $ t_0 $ e a sinistra di $- t_0 $ ,	$t_0 > t_{\alpha/2, n_1+n_2-2}$ o $t_0 < -t_{\alpha/2, n_1+n_2-2}$
$H_1: \mu_1 - \mu_2 > \Delta_0$	Probabilità a destra di $t_0$ ,	$t_0 > t_{\alpha, n_1+n_2-2}$
$H_1: \mu_1 - \mu_2 < \Delta_0$	Probabilità a sinistra di $t_0$ ,	$t_0 < -t_{\alpha, n_1+n_2-2}$

**ESEMPIO 5.4**  
Resa di un processo chimico

Vengono analizzati due catalizzatori (1 e 2) per determinare come influiscono sulla resa media di un processo chimico. Precisamente, il catalizzatore 1 è quello impiegato attualmente, ma il catalizzatore 2 è accettabile. Essendo quest'ultimo più conveniente, lo si vorrebbe adottare, purché non apporti variazioni sulla resa del processo. Nell'impianto pilota viene eseguito un test, i cui risultati sono mostrati in Tabella 5.2. Esiste qualche differenza tra le rese medie? Assumere varianze uguali.

**Tabella 5.2** Dati (in percentuale) relativi alla resa dei catalizzatori dell'Esempio 5.4.

Osservazione	Catalizzatore 1	Catalizzatore 2
1	91.50	89.19
2	94.18	90.95
3	92.18	90.46
4	95.39	93.21
5	91.79	97.19
6	89.07	97.04
7	94.72	91.07
8	89.21	92.75
	$\bar{x}_1 = 92.255$	$\bar{x}_2 = 92.733$
	$s_1 = 2.39$	$s_2 = 2.98$
	$n_1 = 8$	$n_2 = 8$

<sup>1</sup> Anche se abbiamo sviluppato questa procedura per il caso in cui le dimensioni campionarie potrebbero essere differenti, è vantaggioso usare dimensioni campionarie uguali  $n_1 = n_2 = n$ . Infatti, quando le dimensioni campionarie sono le medesime per entrambe le popolazioni, il test  $t$  è molto robusto o insensibile all'assunzione di varianze uguali.

Per la soluzione usiamo la procedura per le verifiche di ipotesi a sette passi:

1. **Parametro di interesse:** i parametri di interesse sono  $\mu_1$  e  $\mu_2$ , cioè la resa media del processo con uso, rispettivamente, del catalizzatore 1 e del catalizzatore 2; vogliamo sapere se  $\mu_1 - \mu_2 = 0$ .
2. **Ipotesi nulla  $H_0$ :**  $\mu_1 - \mu_2 = 0$ , ovvero  $H_0: \mu_1 = \mu_2$
3. **Ipotesi alternativa  $H_1$ :**  $\mu_1 \neq \mu_2$
4. **Statistica test:** la statistica test è

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

5. **Rifiutare  $H_0$  se:** si rifiuta  $H_0$  se il  $P$ -value è minore di 0.05.
6. **Calcoli:** dalla Tabella 5.2 abbiamo  $\bar{x}_1 = 92.255$ ,  $s_1 = 2.39$ ,  $n_1 = 8$ ,  $\bar{x}_2 = 92.733$ ,  $s_2 = 2.98$  e  $n_2 = 8$ . Perciò

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(7)(2.39)^2 + 7(2.98)^2}{8 + 8 - 2} = 7.30$$

$$s_p = \sqrt{7.30} = 2.70$$

e

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{2.70 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{92.255 - 92.733}{2.70 \sqrt{\frac{1}{8} + \frac{1}{8}}} = -0.35$$

7. **Conclusioni:** dall'Appendice A, Tabella II, troviamo che  $t_{0.40,14} = 0.258$  e  $t_{0.25,14} = 0.692$ . Quindi, siccome  $|t_0| = 0.35$  e  $0.258 < 0.35 < 0.692$ , concludiamo che i limiti inferiore e superiore del  $P$ -value sono  $0.50 < P < 0.80$ . Pertanto, dato che il  $P$ -value è maggiore di  $\alpha = 0.05$ , l'ipotesi nulla non può essere rifiutata. Cioè, al livello di significatività 0.05 non si hanno ragioni evidenti per concludere che il catalizzatore 2 comporta una resa media differente da quella che si ottiene con il catalizzatore 1. Il  $P$ -value esatto è  $P = 0.73$ , ottenibile con un software statistico.

Quello che segue è la procedura di Minitab per il test  $t$  a due campioni e per l'intervallo di confidenza per l'Esempio 5.4:

## Two-sample T for Cat 1 vs Cat 2

	N	Mean	StDev	SE Mean
Cat 1	8	92.26	2.39	0.84
Cat 2	8	92.73	2.99	1.1

Difference = mu Cat 1 - mu Cat 2

Estimate for difference: -0.48

95% CI for difference: (-3.37, 2.42)

T-Test of difference = 0 (vs not =): T-Value = -0.35 P-Value = 0.730 DF = 14

Both use Pooled StDev = 2.70

Verifica dell'assunzione  
di normalità.

Si noti che i risultati numerici sono essenzialmente uguali a quelli ottenuti con il calcolo manuale nell'Esempio 5.4. Sono riportati il  $P$ -value 0.73 e l'intervallo di confidenza bilaterale per  $\mu_1 - \mu_2$ ; daremo la formula per il calcolo dell'intervallo di confidenza nel Paragrafo 5.3.3. La Figura 5.2 mostra il grafico dei quantili normali dei due campioni relativi ai dati della resa dei prodotti e i box plot comparativi. I grafici dei quantili indicano che l'assunzione di normalità non comporta alcun problema. Inoltre, le linee rette assumono una pendenza quasi uguale, fornendo una qualche verifica dell'assunzione di varianze uguali. I box plot comparativi indicano che non esiste una evidente differenza tra i due catalizzatori, benché il catalizzatore 2 abbia una variabilità campionaria leggermente superiore.

Caso 2:  $\sigma_1^2 \neq \sigma_2^2$ 

In alcune situazioni non possiamo ragionevolmente assumere che le varianze incognite  $\sigma_1^2$  e  $\sigma_2^2$  siano uguali. In questo caso non esiste una statistica  $t$  esatta disponibile per verificare  $H_0: \mu_1 - \mu_2 = \Delta_0$ . Tuttavia, è possibile usare la seguente statistica.

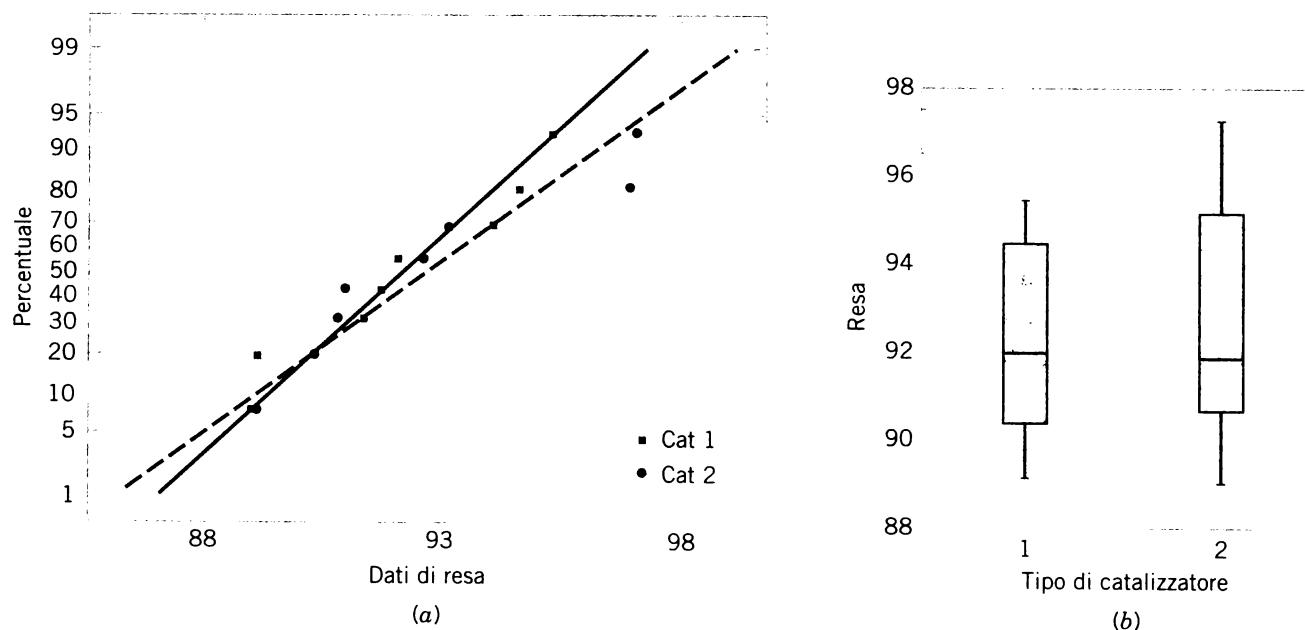


Figura 5.2 Grafico dei quantili normali (a) e box plot comparativi (b) per i dati di resa dei catalizzatori (Esempio 5.4).

**Caso 2: Statistica test per la differenza tra le medie di due distribuzioni normali, varianze incognite e non necessariamente uguali**

$$T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (5.10)$$

è distribuita approssimativamente come una  $t$  con gradi di libertà dati da

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}} \quad (5.11)$$

se risulta vera l'ipotesi nulla  $H_0: \mu_1 - \mu_2 = \Delta_0$ . Se  $v$  non è un intero, occorre arrotondare per difetto all'intero più vicino.

Perciò, quando  $\sigma_1^2 \neq \sigma_2^2$ , le ipotesi sulle differenze tra le medie di due distribuzioni normali sono verificate come nel caso di varianze uguali, ma si usa come statistica Test  $T_0^*$  e si sostituisce  $n_1 + n_2 - 2$  con  $v$  nella determinazione dei gradi di libertà.

**ESEMPIO 5.5**  
Arsenico  
nell'acqua  
potabile

La concentrazione di arsenico presente nell'acqua potabile è un potenziale rischio per la salute. Un articolo apparso su *Arizona Republic* (domenica 27 maggio 2001) riporta la concentrazione di arsenico nelle acque potabili espressa in parti per miliardo (ppb, *part per billion*), per 10 comunità che risiedono nella città di Phoenix e per 10 comunità abitanti nell'Arizona rurale. I dati sono i seguenti:

Territorio urbano di Phoenix ( $\bar{x}_1 = 12.5$ , $s_1 = 7.63$ )	Territorio rurale dell'Arizona ( $\bar{x}_2 = 27.5$ , $s_2 = 15.3$ )
Phoenix, 3	Rimrock, 48
Chandler, 7	Goodyear, 44
Gilbert, 25	New River, 40
Glendale, 10	Apache Junction, 38
Mesa, 15	Buckeye, 33
Paradise Valley, 6	Nogales, 21
Peoria, 12	Black Canyon City, 20
Scottsdale, 25	Sedona, 12
Tempe, 15	Payson, 1
Sun City, 7	Casa Grande, 18

Vogliamo determinare se esiste qualche differenza tra la concentrazione media di arsenico presente nelle comunità urbane di Phoenix e quella presente nelle comunità dell'Arizona rurale. Per i nostri scopi illustrativi assumiamo che questi due insiemi di dati siano campioni casuali rappresentativi di due comunità. La Figura 5.3 mostra un grafico dei quantili per i due campioni di concentrazione di arsenico. L'assunzione di normalità sembra abbastanza

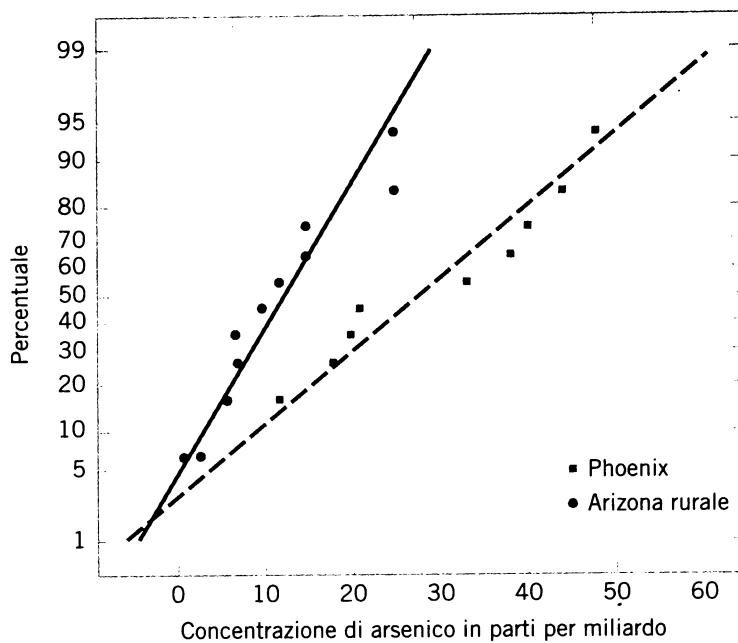


Figura 5.3 Grafico dei quantili per i dati relativi alla concentrazione di arsenico (Esempio 5.5).

ragionevole, ma poiché i coefficienti angolari delle due rette sono tra loro molto diversi, non è verosimile che le varianze della popolazione siano le medesime.

Applichiamo la consueta procedura di risoluzione a sette passi:

1. **Parametro di interesse:** i parametri di interesse sono le concentrazioni medie di arsenico per le due regioni geografiche (le indichiamo con  $\mu_1$  e  $\mu_2$ ); vogliamo stabilire se  $\mu_1 - \mu_2 = 0$ .
2. **Ipotesi nulla  $H_0$ :**  $\mu_1 - \mu_2 = 0$ , ovvero  $H_0: \mu_1 = \mu_2$
3. **Ipotesi alternativa  $H_1$ :**  $\mu_1 \neq \mu_2$
4. **Statistica test:** la statistica test è

$$t_0^* = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

5. **Rifiutare  $H_0$  se:** rifiuteremo  $H_0: \mu_1 = \mu_2$  se il  $P$ -value è minore di 0.05.
6. **Calcoli:** i gradi di libertà per  $t_0^*$  si ricavano dall'Equazione (5.11):

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} = \frac{\left[\frac{(7.63)^2}{10} + \frac{(15.3)^2}{10}\right]^2}{\frac{[(7.63)^2/10]^2}{9} + \frac{[(15.3)^2/10]^2}{9}} = 13.2 \approx 13$$

Il valore della statistica test è:

$$t_0^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{12.5 - 27.5}{\sqrt{\frac{(7.63)^2}{10} + \frac{(15.3)^2}{10}}} = -2.77$$

7. **Conclusioni:** siccome  $t_{0.01,13} = 2.650$ ,  $t_{0.005,13} = 3.012$  e  $|t_0| = 2.77$ , troviamo che  $0.01 < P < 0.02$ . Perciò il  $P$ -value è minore di 0.05 e rifiutiamo l'ipotesi nulla. Pertanto,

è possibile concludere che la concentrazione media di arsenico nell'acqua potabile nell'Arizona rurale è differente dalla concentrazione media di arsenico nell'acqua della città di Phoenix. Inoltre, la concentrazione di arsenico media più alta è quella tra le comunità rurali dell'Arizona.

Nel seguente riquadro è riportato l'output di Minitab per questo esempio.

**Output di Minitab  
per un test *t* per  
due campioni  
e intervallo  
di confidenza**

Two-sample T for PHX vs RuralAZ				
	<i>N</i>	Mean	StDev	SE Mean
PHX	10	12.50	7.63	2.4
RuralAZ	10	27.5	15.3	4.9
<b>Difference = mu PHX – mu RuralAZ</b>				
<b>Estimate for difference: -15.00</b>				
<b>95% CI for difference: (-26.71, -3.29)</b>				
<b>T-Test of difference = 0 (vs not =): T-Value = -2.77 P-Value = 0.016 DF = 13</b>				

I risultati numerici ottenuti con Minitab corrispondono esattamente ai calcoli dall'Esempio 5.5. Si noti che è riportato anche un intervallo di confidenza bilaterale al 95% per  $\mu_1 - \mu_2$ ; discuteremo il suo calcolo nel Paragrafo 5.3.3. Inoltre, l'intervallo non comprende lo zero. Infatti, il limite di confidenza superiore di livello 95% è -3.29 parti per miliardo, cioè ben sotto lo zero, e la differenza media osservata è  $\bar{x}_1 - \bar{x}_2 = 12.5 - 17.5 = -15$  parti per miliardo.

### 5.3.2 Errore del II tipo e scelta della dimensione campionaria

Per valutare l'errore del II tipo per il caso in cui  $\sigma_1^2 = \sigma_2^2 = \sigma^2$  vengono usate le curve caratteristiche operative dell'Appendice A, Carte Va, Vb, Vc e Vd. Sfortunatamente, quando  $\sigma_1^2 \neq \sigma_2^2$  la distribuzione di  $T_0^*$  è incognita se l'ipotesi nulla è falsa, e per questo caso non sono disponibili curve caratteristiche operative.

Per l'alternativa bilaterale  $H_1: \mu_1 - \mu_2 \neq \Delta_0$ , quando  $\sigma_1^2 = \sigma_2^2 = \sigma^2$  e  $n_1 = n_2 = n$  vengono utilizzate le Carte Va e Vb con

$$d = \frac{|\Delta - \Delta_0|}{2\sigma} \quad (5.12)$$

dove  $\Delta$  è la vera differenza tra le medie di interesse. Per usare queste curve, si deve inserire la dimensione campionaria  $n^* = 2n - 1$ . Per l'ipotesi alternativa unilaterale usiamo le Carte Vc e Vd, e definiamo  $d$  e  $\Delta$  come nell'Equazione (5.12). Si noti che il parametro  $d$  è funzione di  $\sigma$ , che è incognita. Come nel test *t* per singolo campione, possiamo dover far riferimento a una precedente stima di  $\sigma$  o usare una stima soggettiva. In alternativa, potremmo definire rispetto a  $\sigma$  le differenze tra le medie che vogliamo rilevare.

**ESEMPIO 5.6**

Resa di un processo chimico

Si consideri l'esperimento dei catalizzatori dell'Esempio 5.4. Si supponga che se il catalizzatore 2 comporta una resa media differente per il 4.0% da quella del catalizzatore 1, vorremmo rifiutare l'ipotesi nulla con una probabilità uguale almeno a 0.85. Quale dimensione campionaria è necessaria?

Usando  $s_p = 2.70$  come stima grossolana della deviazione standard comune  $\sigma$ , abbiamo  $d = |\Delta|/2\sigma = |4.0|/[(2)(2.70)] = 0.74$ . Dalla Carta Va in Appendice A con  $d = 0.74$  e  $\beta = 0.15$ , troviamo approssimativamente  $n^* = 20$ . Quindi, siccome  $n^* = 2n - 1$

$$n = \frac{n^* + 1}{2} = \frac{20 + 1}{2} = 10.5 \approx 11$$

e useremmo dimensioni campionarie  $n_1 = n_2 = n = 11$ .

Il programma Minitab eseguirà anche i calcoli della potenza e della dimensione campionaria per il test  $t$  a due campioni (varianze uguali). L'output di Minitab relativo all'Esempio 5.6 è presentato nel seguente riquadro.

**Calcoli della potenza e della dimensione campionaria eseguiti da Minitab per il test  $t$  per due campioni (varianze uguali)**

2-Sample t Test			
Testing mean 1 = mean 2 (versus not = )			
Calculating power for mean 1 = mean 2 + difference			
Alpha = 0.05 Sigma = 2.7			
Difference	Sample Size	Target Power	Actual Power
4	10	0.8500	0.8793

I risultati sono in buon accordo con i risultati ottenuti dalla curva OC.

### 5.3.3 Intervallo di confidenza per la differenza tra medie

Caso 1:  $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Per sviluppare l'intervallo di confidenza per la differenza tra le medie,  $\mu_1 - \mu_2$ , quando le due varianze sono uguali, notiamo che la distribuzione della statistica

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

è la distribuzione  $t$  con  $n_1 + n_2 - 2$  gradi di libertà. Pertanto

$$P(-t_{\alpha/2, n_1+n_2-2} \leq T \leq t_{\alpha/2, n_1+n_2-2}) = 1 - \alpha$$

ovvero

$$P \left[ -t_{\alpha/2, n_1+n_2-2} \leq \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{\alpha/2, n_1+n_2-2} \right] = 1 - \alpha$$

La manipolazione delle quantità all'interno delle parentesi quadre porta al seguente intervallo di confidenza al  $100(1 - \alpha)\%$  per  $\mu_1 - \mu_2$ .

### Caso 1: Intervallo di confidenza per la differenza tra medie di due distribuzioni normali, varianze non note e uguali

Siano  $\bar{x}_1$ ,  $\bar{x}_2$ ,  $s_1^2$  e  $s_2^2$  le medie e le varianze di due campioni casuali rispettivamente di dimensione  $n_1$  e  $n_2$ , estratti da due popolazioni normali indipendenti con varianze incognite ma uguali. Un intervallo di confidenza di livello  $100(1 - \alpha)\%$  per la differenza tra le medie  $\mu_1 - \mu_2$  è

$$\begin{aligned} \bar{x}_1 - \bar{x}_2 - t_{\alpha/2, n_1 + n_2 - 2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\alpha/2, n_1 + n_2 - 2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \end{aligned} \quad (5.13)$$

dove  $s_p = \sqrt{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)}$  è la stima pooled della deviazione standard comune delle popolazioni, e  $t_{\alpha/2, n_1 + n_2 - 2}$  è il punto percentuale superiore  $100\alpha/2$  della distribuzione  $t$  con  $n_1 + n_2 - 2$  gradi di libertà.

#### ESEMPIO 5.7 Calcio presente nel cemento drogato

Un articolo apparso sul giornale *Hazardous Waste and Hazardous Materials* (Vol. 6, 1989) riportava i risultati di un'analisi del peso del calcio presente nel cemento standard e nel cemento drogato con piombo. Livelli ridotti di calcio indicherebbero che il meccanismo di idratazione nel cemento è bloccato e consentirebbero all'acqua di attaccare vari punti nella struttura in cemento. Dieci provini di cemento standard avevano un peso medio percentuale di calcio pari a  $\bar{x}_1 = 90.0$ , con una deviazione standard campionaria  $s_1 = 5.0$ , e 15 provini di cemento con piombo avevano un peso medio percentuale di calcio pari a  $\bar{x}_2 = 87.0$ , con una deviazione standard campionaria  $s_2 = 4.0$ .

Assumeremo che la percentuale di peso di calcio sia distribuita normalmente e troveremo un intervallo di confidenza al 95% per la differenza tra le medie,  $\mu_1 - \mu_2$ , per i due tipi di cemento. Inoltre, assumeremo che entrambe le popolazioni normali abbiano la medesima deviazione standard.

Si trova la stima pooled della deviazione standard comune usando l'Equazione (5.7) come segue

$$\begin{aligned} s_p^2 &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \\ &= \frac{9(5.0)^2 + 14(4.0)^2}{10 + 15 - 2} \\ &= 19.52 \end{aligned}$$

Perciò la stima della deviazione standard pooled  $s_p = \sqrt{19.52} = 4.4$ . L'intervallo di confidenza al 95% si ricava dell'Equazione (5.13)

$$\bar{x}_1 - \bar{x}_2 - t_{0.025, 23} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{0.025, 23} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

o, sostituendo i valori del campione e usando  $t_{0.025,23} = 2.069$

$$90.0 - 87.0 - 2.069(4.4)\sqrt{\frac{1}{10} + \frac{1}{15}} \leq \mu_1 - \mu_2 \leq 90.0 - 87.0 + 2.069(4.4)\sqrt{\frac{1}{10} + \frac{1}{15}}$$

che si riduce a

$$-0.72 \leq \mu_1 - \mu_2 \leq 6.72$$

Si noti che l'intervallo di confidenza al 95% comprende lo zero; quindi, a questo livello di confidenza, non possiamo concludere che esiste una differenza tra le medie. In altre parole, non possiamo affermare che drogando il cemento con piombo venga influenzata la percentuale media del peso di calcio; perciò non possiamo asserire che la presenza di piombo influisca su questo aspetto del meccanismo di idratazione a un livello di confidenza del 95%.

### Caso 2: $\sigma_1^2 \neq \sigma_2^2$

In molte situazioni non è ragionevole assumere  $\sigma_1^2 = \sigma_2^2$ . Quando questa assunzione non è plausibile, possiamo comunque trovare un intervallo di confidenza di livello  $100(1 - \alpha)\%$  su  $\mu_1 - \mu_2$  usando il fatto che

$$T^* = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

è distribuita approssimativamente come una  $t$  con  $v$  gradi di libertà dati dall'Equazione (5.11). Pertanto

$$P(-t_{\alpha/2,v} \leq T^* \leq t_{\alpha/2,v}) \cong 1 - \alpha$$

e, se sostituiamo l'espressione di  $T^*$  in questa formula e isoliamo  $\mu_1 - \mu_2$  tra le diseguaglianze, possiamo ottenere il seguente intervallo di confidenza per  $\mu_1 - \mu_2$ .

### Caso 2: Intervallo di confidenza per la differenza tra le medie di due distribuzioni normali, varianze incognite e diverse

Se  $\bar{x}_1$ ,  $\bar{x}_2$ ,  $s_1^2$  e  $s_2^2$  sono le medie e le varianze di due campioni casuali, rispettivamente di dimensione  $n_1$  e  $n_2$ , estratti da due popolazioni normali indipendenti con varianze incognite e diverse, allora un intervallo di confidenza al  $100(1 - \alpha)\%$  per la differenza tra le medie,  $\mu_1 - \mu_2$ , approssimato è

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (5.14)$$

dove  $v$  è dato dall'Equazione (5.11) e  $t_{\alpha/2,v}$  è il punto percentuale superiore  $100\alpha/2$  della distribuzione  $t$  con  $v$  gradi di libertà.

### Limiti di confidenza unilaterali

Per trovare un limite inferiore di confidenza al  $100(1 - \alpha)\%$  per  $\mu_1 - \mu_2$  con valori di  $\sigma^2$  incogniti, sostituiamo semplicemente  $-t_{\alpha/2, n_1 + n_2 - 2}$  con  $-t_{\alpha, n_1 + n_2 - 2}$  nel limite inferiore dell'Equazione (5.13) per il caso 1, e  $-t_{\alpha/2, v}$  con  $-t_{\alpha, v}$  nel limite inferiore dell'Equazione (5.14) per il caso 2; il limite superiore viene impostato a  $\infty$ . Analogamente, per trovare un limite superiore di confidenza al  $100(1 - \alpha)\%$  per  $\mu_1 - \mu_2$  con valori di  $\sigma^2$  incogniti, sostituiamo semplicemente  $t_{\alpha/2, n_1 + n_2 - 2}$  con  $t_{\alpha, n_1 + n_2 - 2}$  nel limite superiore dell'Equazione (5.13) per il caso 1 e  $t_{\alpha/2, v}$  con  $t_{\alpha, v}$  nel limite superiore dell'Equazione (5.14) per il caso 2; il limite inferiore viene impostato a  $-\infty$ .

## 5.4 TEST $t$ ACCOPPIATO

Un caso speciale del test  $t$  per due campioni visto nel Paragrafo 5.3 si ha quando le osservazioni su due popolazioni di interesse sono raccolte in **coppia**. Ciascuna coppia di osservazioni, per esempio  $(X_{1j}, X_{2j})$ , viene presa sotto condizioni omogenee, ma queste condizioni possono cambiare da una coppia all'altra. Per esempio, supponiamo di voler confrontare due differenti tipi di punte per una macchina che deve saggiare la durezza dei metalli. Questa macchina preme la punta su un provino di metallo con una forza nota. Misurando la profondità della depressione causata dalla punta, si determina la durezza del metallo del provino. Se sono stati selezionati casualmente diversi provini, sottoposti a prova metà con la punta 1 e metà con la punta 2, ed è stato applicato il test  $t$  indipendente o pooled del Paragrafo 5.3, i risultati del test potrebbero essere errati. I provini di metallo potrebbero essere stati tagliati da uno stock di barre prodotte a calore differente, o potrebbero non essere omogenei per altri motivi, tali da influire sulla durezza. In tal caso la differenza osservata tra le misure di durezza media per i due tipi di punte include anche le differenze di durezza tra i provini.

Una procedura sperimentale più efficace consiste nel raccogliere i dati in coppia, cioè effettuare due letture di durezza su ciascun provino, una per ogni punta. La procedura di verifica dovrebbe quindi consistere nell'analisi delle *differenze* tra le letture di durezza su ciascun provino. Se non esistono differenze tra le punte, la media delle differenze dovrebbe essere zero. Questa procedura di verifica viene chiamata **test  $t$  accoppiato**.

Siano  $(X_{11}, X_{21}), (X_{12}, X_{22}), \dots, (X_{1n}, X_{2n})$  un insieme di  $n$  osservazioni accoppiate, dove assumiamo che la media e la varianza della popolazione rappresentata da  $X_1$  siano  $\mu_1$  e  $\sigma_1^2$ , e la media e la varianza della popolazione rappresentata da  $X_2$  siano  $\mu_2$  e  $\sigma_2^2$ . Definiamo le differenze tra ogni coppia di osservazioni come  $D_j = X_{1j} - X_{2j}, j = 1, 2, \dots, n$ . I termini  $D_j$  si assumono distribuiti normalmente con media

$$\mu_D = E(X_1 - X_2) = E(X_1) - E(X_2) = \mu_1 - \mu_2$$

e varianza  $\sigma_D^2$ . In questo modo, per verificare le ipotesi riguardanti la differenza tra  $\mu_1$  e  $\mu_2$  si può eseguire un test  $t$  a singolo campione su  $\mu_D$ . Specificatamente, verificare  $H_0: \mu_1 - \mu_2 = \Delta_0$  contro  $H_1: \mu_1 - \mu_2 \neq \Delta_0$  è equivalente a verificare

$$\begin{aligned} H_0: \mu_D &= \Delta_0 \\ H_1: \mu_D &\neq \Delta_0 \end{aligned} \tag{5.15}$$

La statistica test è data nel seguente riquadro.

<b>Test <math>t</math> accoppiato</b>		
Ipotesi nulla:	$H_0: \mu_D = \Delta_0$	
Statistica test:	$T_0 = \frac{\bar{D} - \Delta_0}{S_D/\sqrt{n}}$	
<b>Ipotesi alternativa</b>	<b>P-value</b>	<b>Regione di rifiuto</b>
$H_1: \mu_D \neq \Delta_0$	Somma probabilità a destra di $ t_0 $ e a sinistra di $- t_0 $ ,	$t_0 > t_{\alpha/2,n-1}$ o $t_0 < -t_{\alpha/2,n-1}$
$H_1: \mu_D > \Delta_0$	Probabilità a destra di $t_0$ ,	$t_0 > t_{\alpha,n-1}$
$H_1: \mu_D < \Delta_0$	Probabilità a sinistra di $t_0$ ,	$t_0 < -t_{\alpha,n-1}$

Nell'Equazione (5.16),  $\bar{D}$  è la media campionaria delle  $n$  differenze  $D_1, D_2, \dots, D_n$  e  $S_D$  è la deviazione standard campionaria di queste differenze.

### ESEMPIO 5.8 Resistenza al taglio di travi d'acciaio

Un articolo apparso sul *Journal of Strain Analysis* (1983, Vol. 18, No. 2) confronta diversi metodi per predire la resistenza al taglio di travi a lamine di acciaio. I dati per due di questi metodi, la procedura di Karlsruhe e la procedura di Lehigh, riguardanti nove specifiche travi, sono mostrati in Tabella 5.3. Vogliamo determinare se esiste qualche differenza (in media) tra i due metodi.

Applichiamo la solita procedura a sette passi.

- Parametro di interesse:** il parametro di interesse è la differenza tra la resistenza al taglio media tra i due metodi:  $\mu_D = \mu_1 - \mu_2$ .
- Ipotesi nulla  $H_0$ :**  $\mu_D = 0$
- Ipotesi alternativa  $H_1$ :**  $\mu_D \neq 0$
- Statistica test:** la statistica test è

$$t_0 = \frac{\bar{d}}{s_d/\sqrt{n}}$$

- Rifiutare  $H_0$  se:** si rifiuta  $H_0$  se il P-value è minore di 0.05.
- Calcoli:** la media e la deviazione standard campionarie delle differenze  $d_j$  sono  $\bar{d} = 0.2769$  e  $s_d = 0.1350$ , per cui la statistica test è

$$t_0 = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{0.2769}{0.1350/\sqrt{9}} = 6.15$$

- Conclusioni:** siccome  $t_{0.0005,8} = 5.041$  e  $t = 6.15$  supera questo valore, il P-value è minore di  $2(0.0005) = 0.001$ . Perciò, concludiamo che i metodi di predizione della resistenza danno risultati differenti. Più precisamente, i dati indicano che il metodo di Karlsruhe produce, in media, predizioni di resistenza più elevate rispetto a quelle prodotte dal metodo di Lehigh.

**Tabella 5.3** Predizione della resistenza per nove travi a lamine di acciaio (carico predetto/carico osservato).

Trave	Metodo Karlsruhe	Metodo Lehigh	Differenza $d_j$
S1/1	1.186	1.061	0.125
S2/1	1.151	0.992	0.159
S3/1	1.322	1.063	0.259
S4/1	1.339	1.062	0.277
S5/1	1.200	1.065	0.135
S2/1	1.402	1.178	0.224
S2/2	1.365	1.037	0.328
S2/3	1.537	1.086	0.451
S2/4	1.559	1.052	0.507

Nel seguente riquadro è riportato l'output di Minitab per questo esempio.

Output  
di Minitab  
per un test  $t$   
accoppiato  
e intervallo  
di confidenza  
per l'Esempio  
5.8

<u>Paired t-Test and CI: Karlsruhe, Lehigh</u>				
Paired T for Karlsruhe-Lehigh				
	N	Mean	StDev	SE Mean
Karlsruhe	9	1.34011	0.14603	0.04868
Lehigh	9	1.06322	0.05041	0.01680
Difference	9	0.276889	0.135027	0.045009
95% CI for mean difference: (0.173098, 0.380680)				
T-Test of mean difference = 0 (vs not = 0): T-Value = 6.15, P-Value = 0.000				

I risultati ottenuti tramite software coincidono sostanzialmente con quelli ottenuti tramite calcoli manuali. Oltre ai risultati della verifica di ipotesi, Minitab riporta un intervallo di confidenza bilaterale sulla differenza fra le medie. Questo intervallo di confidenza è stato ricavato costruendo un intervallo di confidenza a campione singolo per  $\mu_D$ . Forniamo i dettagli di seguito.

#### Confronto tra test accoppiato e test non accoppiato

Nell'eseguire un esperimento comparativo, l'analista può a volte scegliere tra l'esperimento accoppiato e l'esperimento a due campioni (o non accoppiato). Se si devono effettuare  $n$  misurazioni su ciascuna popolazione, la statistica  $t$  a due campioni è

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p \sqrt{\frac{1}{n} + \frac{1}{n}}}$$

e dovrebbe essere confrontata con  $t_{2n-2}$ ; naturalmente, la statistica  $t$  accoppiata è

$$T_0 = \frac{\bar{D} - \Delta_0}{S_D/\sqrt{n}}$$

da confrontarsi con  $t_{n-1}$ . Si noti che essendo

$$\bar{D} = \sum_{j=1}^n \frac{D_j}{n} = \sum_{j=1}^n \frac{(X_{1j} - X_{2j})}{n} = \sum_{j=1}^n \frac{X_{1j}}{n} - \sum_{j=1}^n \frac{X_{2j}}{n} = \bar{X}_1 - \bar{X}_2$$

i numeratori delle due statistiche sono identici. Tuttavia, il denominatore del test  $t$  a due campioni si basa sull'assunzione che  $X_1$  e  $X_2$  siano *indipendenti*. In molti esperimenti accoppiati esiste una forte correlazione positiva  $\rho$  tra  $X_1$  e  $X_2$ . Si può quindi dimostrare che

$$V(\bar{D}) = V(\bar{X}_1 - \bar{X}_2 - \Delta_0) = V(\bar{X}_1) + V(\bar{X}_2) - 2 \operatorname{cov}(\bar{X}_1, \bar{X}_2) = \frac{2\sigma^2(1 - \rho)}{n}$$

assumendo che entrambe le popolazioni  $X_1$  e  $X_2$  abbiano identiche varianze  $\sigma^2$ . Inoltre  $S_D^2/n$ , stima la varianza di  $\bar{D}$ . Tutte le volte che esiste una correlazione positiva all'interno delle coppie, il denominatore del test  $t$  accoppiato sarà più piccolo del denominatore del test  $t$  a due campioni. Ciò può far sì che il test  $t$  a due campioni sottostimi considerevolmente la significatività dei dati se viene applicato erroneamente ai campioni accoppiati.

Benché l'accoppiamento porti spesso a un valore più piccolo della varianza di  $\bar{X}_1 - \bar{X}_2$  ha uno svantaggio: il test  $t$  accoppiato porta a una perdita di  $n - 1$  gradi di libertà rispetto al test  $t$  a due campioni. In generale, sappiamo che all'aumento dei gradi di libertà di un test aumenta la potenza rispetto a ogni valore alternativo fissato del parametro.

Come possiamo decidere di condurre l'esperimento, dunque? Dobbiamo accoppiare le osservazioni o no? Benché questa domanda non abbia una risposta valida in generale, possiamo fornire alcune linee guida sulla base della precedente discussione:

1. Se le unità sperimentali sono relativamente omogenee ( $\sigma$  piccola) e la correlazione all'interno delle coppie è piccola, il guadagno in precisione attribuibile all'accoppiamento sarà annullato dalla perdita di gradi di libertà; si dovrebbe quindi usare un esperimento con campioni indipendenti.
2. Se le unità sperimentali sono relativamente eterogenee ( $\sigma$  grande) e vi è una grande correlazione positiva all'interno delle coppie, si dovrebbe ricorrere all'esperimento accoppiato. Tipicamente, questo caso si presenta quando le unità sperimentali sono le *medesime* per entrambi i trattamenti, come nell'Esempio 5.8, dove per verificare i due metodi sono state usate le medesime travi.

L'applicazione di queste regole generali richiede comunque una certa attenzione, perché  $\sigma$  e  $\rho$  non sono mai noti con precisione. Inoltre, se il numero di gradi di libertà è elevato (per esempio 40 o 50), la perdita di  $n - 1$  gradi di libertà per l'accoppiamento può non essere preoccupante. Se invece il numero di gradi di libertà è basso (per esempio 10 o 20), la perdita di metà dei gradi di libertà può essere potenzialmente preoccupante se non viene compensata da una maggiore precisione dovuta all'accoppiamento.

Intervallo di confidenza per  $\mu_D$

Per costruire l'intervallo di confidenza per  $\mu_D$ , si osservi che

$$T = \frac{\bar{D} - \mu_D}{S_D / \sqrt{n}}$$

segue una distribuzione  $t$  con  $n - 1$  gradi di libertà. Perciò, essendo

$$P(-t_{\alpha/2,n-1} \leq T \leq t_{\alpha/2,n-1}) = 1 - \alpha$$

possiamo sostituire l'espressione di  $T$  nell'argomento della probabilità ed eseguire i passi necessari per isolare  $\mu_D = \mu_1 - \mu_2$  tra le disuguaglianze. Questo porta al seguente intervallo di confidenza al  $100(1 - \alpha)\%$  per  $\mu_D = \mu_1 - \mu_2$ .

#### Intervallo di confidenza per $\mu_D$ per osservazioni accoppiate

Se  $\bar{d}$  e  $s_d$  sono, rispettivamente, la media e la deviazione standard campionarie della differenza distribuita normalmente di  $n$  coppie casuali di misure, un intervallo di confidenza al  $100(1 - \alpha)\%$  per la differenza tra le medie  $\mu_D = \mu_1 - \mu_2$  è

$$\bar{d} - t_{\alpha/2,n-1}s_d/\sqrt{n} \leq \mu_D \leq \bar{d} + t_{\alpha/2,n-1}s_d/\sqrt{n} \quad (5.17)$$

dove  $t_{\alpha/2,n-1}$  è il punto percentuale superiore  $100\alpha/2$  della distribuzione  $t$  con  $n - 1$  gradi di libertà.

Questo intervallo di confidenza è valido anche per il caso in cui  $\sigma_1^2 \neq \sigma_2^2$ , dato che  $s_D^2$  stima  $\sigma_D^2 = V(X_1 - X_2)$ . Inoltre, per campioni numerosi (per esempio  $n \geq 40$  coppie), l'assunzione esplicita di normalità non è necessaria grazie al teorema limite centrale.

L'Equazione (5.17) è stata impiegata per calcolare l'intervallo di confidenza nell'esperimento dell'Esempio 5.8.

#### ESEMPIO 5.9 Parcheggi in parallelo

La pubblicazione *Human Factors* (1962, pp. 375-380) riporta uno studio nel quale a  $n = 14$  soggetti è stato richiesto di parcheggiare in parallelo due automobili aventi pneumatici di spessore molto diverso e raggi di sterzo molto differenti. Sono stati registrati i tempi, espressi in secondi, impiegati da ciascun soggetto: sono quelli elencati in Tabella 5.4. Dalla colonna delle differenze osservate calcoliamo  $\bar{d} = 1.21$  e  $s_d = 12.68$ . L'intervallo di confidenza al 90% per  $\mu_D = \mu_1 - \mu_2$  si ricava dall'Equazione (5.17) nel seguente modo:

$$\begin{aligned} \bar{d} - t_{0.05,13}s_d/\sqrt{n} &\leq \mu_D \leq \bar{d} + t_{0.05,13}s_d/\sqrt{n} \\ 1.21 - 1.771(12.68)/\sqrt{14} &\leq \mu_D \leq 1.21 + 1.771(12.68)/\sqrt{14} \\ -4.79 &\leq \mu_D \leq 7.21 \end{aligned}$$

Si noti che l'intervallo di confidenza per  $\mu_D$  comprende lo zero. Questo implica che, a un livello di confidenza del 90%, i dati non supportano l'asserzione che le due automobili abbiano tempi medi di parcheggio  $\mu_1$  e  $\mu_2$  differenti. Il valore  $\mu_D = \mu_1 - \mu_2 = 0$ , cioè, non è inconsistente con i dati osservati.

**Tabella 5.4** Tempo in secondi necessario per parcheggiare in parallelo due automobili.

Soggetto	Automobile		Differenza ( $d_j$ )
	1 ( $x_{1j}$ )	2 ( $x_{2j}$ )	
1	37.0	17.8	19.2
2	25.8	20.2	5.6
3	16.2	16.8	-0.6
4	24.2	41.4	-17.2
5	22.0	21.4	0.6
6	33.4	38.4	-5.0
7	23.8	16.8	7.0
8	58.2	32.2	26.0
9	33.6	27.8	5.8
10	24.4	23.2	1.2
11	23.4	29.6	-6.2
12	21.2	20.6	0.6
13	36.2	32.2	4.0
14	29.8	53.8	-24.0

## 5.5 INFERENZA SUL RAPPORTO TRA LE VARIANZE DI DUE POPOLAZIONI NORMALI

Presentiamo ora le verifiche e gli intervalli di confidenza per le varianze delle due popolazioni mostrate in Figura 5.1. Assumeremo che entrambe le popolazioni siano normali. Sia le verifiche di ipotesi, sia le procedure per costruire gli intervalli di confidenza sono relativamente sensibili all'assunzione di normalità.

### 5.5.1 Verifica di ipotesi sul rapporto tra due varianze

Supponiamo di avere a che fare con due popolazioni normali indipendenti; le medie e le varianze delle popolazioni,  $\mu_1$ ,  $\sigma_1^2$ ,  $\mu_2$ , e  $\sigma_2^2$ , sono incognite. Vogliamo verificare l'ipotesi dell'uguaglianza delle due varianze, cioè  $H_0: \sigma_1^2 = \sigma_2^2$ . Assumiamo che siano disponibili due campioni casuali, uno di dimensione  $n_1$  estratto dalla popolazione 1, e uno di dimensione  $n_2$  estratto dalla popolazione 2; infine, siano  $S_1^2$  e  $S_2^2$  le varianze campionarie. Vogliamo verificare le ipotesi

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Lo sviluppo di una procedura di verifica per queste ipotesi richiede una nuova distribuzione di probabilità.

### La distribuzione $F$

Una delle distribuzioni più utili nella statistica è la distribuzione  $F$ . La variabile aleatoria  $F$  è definita come il rapporto tra due variabili aleatorie chi-quadro indipendenti, ciascuna divisa per il proprio numero di gradi di libertà, ossia

$$F = \frac{W/u}{Y/v}$$

dove  $W$  e  $Y$  sono variabili aleatorie chi-quadro indipendenti, rispettivamente con  $u$  e  $v$  gradi di libertà. Definiamo formalmente la distribuzione campionaria di  $F$ .

#### Distribuzione $F$

Siano  $W$  e  $Y$  variabili aleatorie chi-quadro indipendenti, rispettivamente con  $u$  e  $v$  gradi di libertà. Allora il rapporto

$$F = \frac{W/u}{Y/v} \tag{5.18}$$

ha la funzione densità di probabilità

$$f(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{u/2}x^{(u/2)-1}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)x + 1\right]^{(u+v)/2}}, \quad 0 < x < \infty \tag{5.19}$$

e si dice che segue la distribuzione  $F$  con  $u$  gradi di libertà al numeratore e  $v$  gradi di libertà al denominatore. Tale distribuzione viene indicata di solito con la notazione abbreviata  $F_{u,v}$ .

La media e la varianza della distribuzione  $F$  sono  $\mu = v/(v - 2)$  per  $v > 2$ , e

$$\sigma^2 = \frac{2v^2(u + v - 2)}{u(v - 2)^2(v - 4)}, \quad v > 4$$

In Figura 5.4 sono mostrate due distribuzioni  $F$ . La variabile aleatoria  $F$  è non negativa, e la distribuzione è asimmetrica verso destra. La distribuzione  $F$  è molto simile alla distribuzione chi-quadro di Figura 4.21; tuttavia, i due parametri  $u$  e  $v$  forniscono una flessibilità di forma supplementare.

I punti percentuali della distribuzione  $F$  sono dati nella Tavola IV dell'Appendice A. Sia  $f_{\alpha,u,v}$  il punto percentuale della distribuzione  $F$ , con  $u$  gradi di libertà al numeratore e  $v$  gradi di libertà al denominatore, cosicché la probabilità che la variabile aleatoria  $F$  superi questo valore è

Punti percentuali  
della coda superiore  
della distribuzione  $F$ .

$$P(F > f_{\alpha,u,v}) = \int_{f_{\alpha,u,v}}^{\infty} f(x) dx = \alpha$$

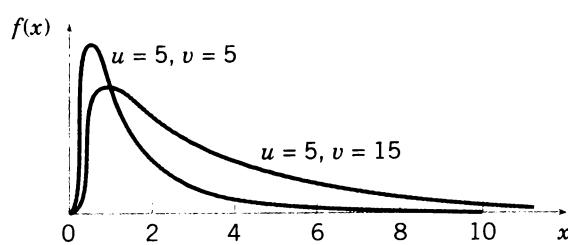


Figura 5.4 Funzione di densità di probabilità di due distribuzioni  $F$ .

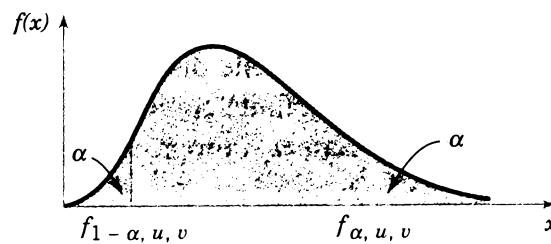


Figura 5.5 Punti percentuali superiore e inferiore della distribuzione  $F$ .

Questo risultato è illustrato in Figura 5.5. Per esempio, se  $u = 5$  e  $v = 10$ , troviamo dalla Tavola IV dell'Appendice A che

$$P(F > f_{0.05,5,10}) = P(F_{5,10} > 3.33) = 0.05$$

Insomma, il punto percentuale superiore 5 di  $F_{5,10}$  è  $f_{0.05,5,10} = 3.33$ .

La Tavola IV contiene solo punti percentuali della coda superiore della distribuzione (per valori selezionati di  $f_{\alpha,u,v}$  per  $\alpha=0.25$ )  $F$ . I punti percentuali della coda inferiore  $f_{1-\alpha,u,v}$  si possono trovare nel seguente modo:

$$f_{1-\alpha,u,v} = \frac{1}{f_{\alpha,v,u}} \quad (5.20)$$

Per esempio, per trovare il punto percentuale della coda inferiore  $f_{0.95,5,10}$  notiamo che

**Determinazione  
di un punto percentuale  
della coda inferiore  
della distribuzione  $F$ .**

$$f_{0.95,5,10} = \frac{1}{f_{0.05,10,5}} = \frac{1}{4.74} = 0.211$$

La procedura di verifica

Una procedura per la verifica dell'ipotesi dell'uguaglianza di due varianze si basa sul seguente risultato.

Sia  $X_{11}, X_{12}, \dots, X_{1n_1}$  un campione casuale estratto da una popolazione normale con media  $\mu_1$  e varianza  $\sigma_1^2$ , e sia  $X_{21}, X_{22}, \dots, X_{2n_2}$  un campione casuale estratto da una seconda popolazione normale con media  $\mu_2$  e varianza  $\sigma_2^2$ . Si assuma che le due popolazioni normali siano indipendenti. Siano  $S_1^2$  e  $S_2^2$  le varianze campionarie. Allora il rapporto

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

ha una distribuzione  $F$  con  $n_1 - 1$  gradi di libertà al numeratore e  $n_2 - 1$  gradi di libertà al denominatore.

Questo risultato si basa sul fatto che  $(n_1 - 1)S_1^2/\sigma_1^2$  è una variabile aleatoria chi-quadro con  $n_1 - 1$  gradi di libertà, che  $(n_2 - 1)S_2^2/\sigma_2^2$  è una variabile aleatoria chi-quadro con  $n_2 - 1$  gradi di libertà, e che le due popolazioni normali sono indipendenti. Chiaramente, sotto l'ipotesi nulla:  $H_0: \sigma_1^2 = \sigma_2^2$  il rapporto  $F_0 = S_1^2/S_2^2$  ha una distribuzione  $F_{n_1-1, n_2-1}$ . Su ciò si basa la seguente procedura di verifica.

## Sintesi

**Verifica di ipotesi sull'uguaglianza delle varianze di due distribuzioni normali**

$$\text{Ipotesi nulla: } H_0: \sigma_1^2 = \sigma_2^2$$

$$\text{Statistica test: } F_0 = \frac{S_1^2}{S_2^2} \quad (5.21)$$

**Ipotesi alternative**

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

$$H_1: \sigma_1^2 > \sigma_2^2$$

$$H_1: \sigma_1^2 < \sigma_2^2$$

**Criterio di rifiuto**

$$f_0 > f_{\alpha/2, n_1-1, n_2-1} \text{ o } f_0 < f_{1-\alpha/2, n_1-1, n_2-1}$$

$$f_0 > f_{\alpha, n_1-1, n_2-1}$$

$$f_0 < f_{1-\alpha, n_1-1, n_2-1}$$

Le regioni critiche sono mostrate in Figura 5.6.

**ESEMPIO 5.10**  
 Riduzione  
 chimica dello  
 spessore  
 di un wafer  
 semiconduttore

Gli strati di ossido su wafer di semiconduttore vengono attaccati con una miscela di gas per ridurne lo spessore al valore appropriato. La variabilità dello spessore di questi strati di ossido è una caratteristica critica del wafer, e per le fasi successive del processo è desiderabile una bassa variabilità. Vengono studiate due diverse miscele di gas per stabilire quale delle due è più efficace nella riduzione della variabilità dello spessore dell'ossido. Con ciascuna miscela di gas si attaccano sedici wafer. Le deviazioni standard campionarie dello spessore di ossido sono rispettivamente  $s_1 = 1.96$  angstrom e  $s_2 = 2.13$  angstrom. Vi è indicazione che uno dei due gas sia preferibile all'altro? Usare  $\alpha = 0.05$ .

A questo problema si può applicare la solita procedura di verifica di ipotesi a sette passi.

1. **Parametro di interesse:** i parametri di interesse sono le varianze dello spessore di ossido  $\sigma_1^2$  e  $\sigma_2^2$ . Assumeremo che lo spessore di ossido sia una variabile aleatoria normale per entrambe le miscele di gas.
2. **Ipotesi nulla  $H_0$ :**  $\sigma_1^2 = \sigma_2^2$
3. **Ipotesi alternativa  $H_1$ :**  $\sigma_1^2 \neq \sigma_2^2$
4. **Statistica test:** la statistica test è data dall'Equazione (5.21):

$$f_0 = \frac{s_1^2}{s_2^2}$$

5. **Rifiutare  $H_0$  se:** Essendo  $n_1 = n_2 = 16$  e  $\alpha = 0.05$ , rifiuteremo  $H_0: \sigma_1^2 = \sigma_2^2$  se  $f_0 > f_{0.025, 15, 15} = 2.86$  o se  $f_0 < f_{0.975, 15, 15} = 1/f_{0.025, 15, 15} = 1/2.86 = 0.35$ . Si faccia riferimento alla Figura 5.6a.

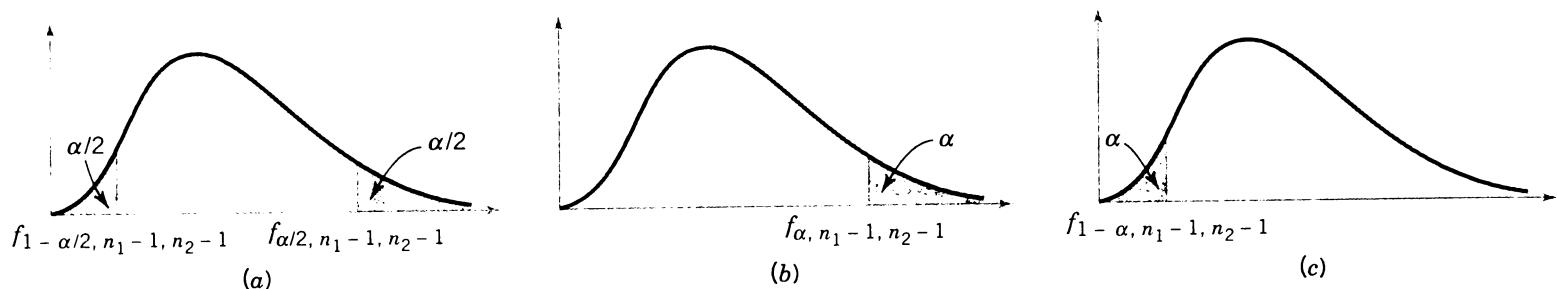


Figura 5.6 La distribuzione  $F$  per la verifica di  $H_0: \sigma_1^2 = \sigma_2^2$  con i valori della regione critica per (a)  $H_1: \sigma_1^2 \neq \sigma_2^2$ , (b)  $H_1: \sigma_1^2 > \sigma_2^2$ , e (c)  $H_1: \sigma_1^2 < \sigma_2^2$ .

6. **Calcoli:** essendo  $s_1^2 = (1.96)^2 = 3.84$  e  $s_2^2 = (2.13)^2 = 4.54$ , la statistica test è

$$f_0 = \frac{s_1^2}{s_2^2} = \frac{3.84}{4.54} = 0.85$$

7. **Conclusioni:** siccome  $f_{0.975, 15, 15} = 0.35 < 0.85 < f_{0.25, 15, 15} = 2.86$ , non possiamo rifiutare l'ipotesi nulla  $H_0: \sigma_1^2 = \sigma_2^2$  al livello di significatività 0.05. **Conclusione pratica ingegneristica:** non vi è una forte evidenza del fatto che una delle due miscele di gas dia luogo a una minore varianza dello spessore dell'ossido. Possiamo allora scegliere il gas a minore costo o più facile da usare.

### P-value per il test $F$

L'approccio mediante  $P$ -value può essere adottato anche per i test  $F$ . Per capire come procedere, si consideri il test a una coda relativo alla coda superiore. Il  $P$ -value è l'area (la probabilità) al di sotto della curva di distribuzione  $F$  con  $n_1 - 1$  e  $n_2 - 2$  gradi di libertà che si trova al di là del valore della statistica test calcolato,  $f_0$ . Nella Tavola IV dell'Appendice A si possono reperire i limiti superiore e inferiore per il  $P$ -value. Per esempio, si consideri un test  $F$  con 9 gradi di libertà al numeratore e 14 gradi di libertà al denominatore, per il quale  $f_0 = 3.05$ . Dalla tavola IV dell'Appendice A ricaviamo  $f_{0.05, 9, 14} = 2.65$  e  $f_{0.025, 9, 14} = 3.21$ ; pertanto, dato che  $f_0 = 3.05$  cade entro questi estremi, si ha:  $0.025 < P < 0.05$ . Si può trovare in maniera analoga il  $P$ -value per un test a una coda relativo alla coda inferiore; poiché però la Tavola IV dell'Appendice A riporta solo punti della coda superiore per la distribuzione  $F$ , si dovrà utilizzare l'Equazione (5.20). Per un test a due code si dovranno invece raddoppiare i limiti ottenuti per il test a una coda se si vuole determinare il  $P$ -value corretto.

Per illustrare il calcolo dei limiti per il  $P$ -value in un test  $F$  a due code, si consideri nuovamente l'Esempio 5.10. In tale esempio il valore della statistica test calcolato è  $f_0 = 0.85$ , e cade nella coda inferiore della distribuzione  $F_{15, 15}$ . Il punto della coda inferiore che ha una probabilità pari a 0.25 alla sua sinistra è  $f_{0.75, 15, 15} = 1/f_{0.25, 15, 15} = 1/1.43 = 0.70$ , e dato che  $0.70 < 0.85$ , la probabilità che giace alla sinistra di 0.85 è maggiore di 0.25. Di conseguenza, dovremmo concludere che il  $P$ -value per  $f_0 = 0.85$  è maggiore di  $2(0.25) = 0.5$ , perciò non vi sono indizi sufficienti per rifiutare l'ipotesi nulla. Ciò è coerente con le conclusioni originali dell'Esempio 5.10. Il  $P$ -value effettivo è 0.7570; lo si può ottenere tramite un calcolatore, da cui si ricava  $P(F_{15, 15} \leq 0.85) = 0.3785$  e perciò  $2(0.3785) = 0.7570$ . Per calcolare le probabilità richieste si può anche usare Minitab.

### 5.5.2 Intervallo di confidenza per il rapporto tra due varianze

Per trovare l'intervallo di confidenza, si ricordi che la distribuzione campionaria di

$$F = \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2}$$

è una distribuzione  $F$  con  $n_2 - 1$  e  $n_1 - 1$  gradi di libertà. *Nota:* Partiamo con  $S_2^2$  al numeratore e  $S_1^2$  al denominatore per semplificare i calcoli algebrici usati per ottenere un intervallo per  $\sigma_1^2/\sigma_2^2$ . Pertanto

$$P(f_{1-\alpha/2, n_2-1, n_1-1} \leq F \leq f_{\alpha/2, n_2-1, n_1-1}) = 1 - \alpha$$

La sostituzione di  $F$  con la sua espressione e la manipolazione delle diseguaglianze porterà al seguente intervallo di confidenza al  $100(1 - \alpha)\%$  per  $\sigma_1^2/\sigma_2^2$ .

#### Definizione

##### **Intervallo di confidenza per il rapporto tra le varianze di due distribuzioni normali**

Se  $s_1^2$  e  $s_2^2$  sono le varianze campionarie di due campioni casuali, rispettivamente di dimensione  $n_1$  e  $n_2$ , estratti da due popolazioni normali indipendenti con varianze incognite  $\sigma_1^2$  e  $\sigma_2^2$ , un intervallo di confidenza di livello  $100(1 - \alpha)\%$  per il rapporto  $\sigma_1^2/\sigma_2^2$  è

$$\frac{s_1^2}{s_2^2} f_{1-\alpha/2, n_2-1, n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2} f_{\alpha/2, n_2-1, n_1-1} \quad (5.22)$$

dove  $f_{\alpha/2, n_2-1, n_1-1}$  e  $f_{1-\alpha/2, n_2-1, n_1-1}$  sono i punti percentuali superiore e inferiore  $100 \alpha/2$  della distribuzione  $F$ , con  $n_2 - 1$  gradi di libertà al numeratore e  $n_1 - 1$  gradi di libertà al denominatore.

#### ESEMPIO 5.11

##### Finitura superficiale

Un'azienda realizza rotori impiegati nei motori a turbina per jet. Una delle fasi di costruzione consiste nella molatura della finitura superficiale di un componente in lega di titanio. Si possono usare due differenti processi di molatura: entrambi sono in grado di fornire parti con identica ruvidezza superficiale media. L'ingegnere addetto alla produzione vorrebbe scegliere il tipo di processo con la minore variabilità della ruvidezza superficiale. Un campione casuale composto da  $n_1 = 11$  parti trattato con il primo processo dà luogo a una deviazione standard campionaria  $s_1 = 5.1 \mu\text{in}$ ; un campione casuale composto da  $n_2 = 16$  parti trattato con il secondo processo dà luogo invece a una deviazione standard campionaria  $s_2 = 4.7 \mu\text{in}$ . Troviamo un intervallo di confidenza al 90% per il rapporto tra le due varianze  $\sigma_1^2/\sigma_2^2$ .

Assumendo che i due processi siano indipendenti e che la ruvidezza superficiale sia distribuita normalmente, possiamo usare l'Equazione (5.22) come segue:

$$\frac{s_1^2}{s_2^2} f_{0.95, 15, 10} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2} f_{0.05, 15, 10}$$

$$\frac{(5.1)^2}{(4.7)^2} 0.39 \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{(5.1)^2}{(4.7)^2} 2.85$$

ovvero

$$0.46 \leq \frac{\sigma_1^2}{\sigma_2^2} \leq 3.36$$

Si noti che abbiamo usato l'Equazione (5.20) per trovare  $f_{0.95, 15, 10} = 1/f_{0.05, 10, 15} = 1/2.54 = 0.39$ . Dato che questo intervallo di confidenza comprende l'unità, non possiamo affermare che le deviazioni standard della ruvidezza superficiale per i due processi siano differenti a un livello di confidenza del 90%.

#### Limiti di confidenza unilaterali

Per trovare un limite inferiore di confidenza di livello  $100(1 - \alpha)\%$  per  $\sigma_1^2/\sigma_2^2$  sostituiamo semplicemente  $f_{\alpha/2, n_2 - 1, n_1 - 1}$  con  $f_{1 - \alpha/2, n_2 - 1, n_1 - 1}$  nel limite inferiore dell'Equazione (5.22); il limite superiore viene impostato a  $\infty$ . Analogamente, per trovare un limite superiore di confidenza al livello  $100(1 - \alpha)\%$  per  $\sigma_1^2/\sigma_2^2$  sostituiamo  $f_{\alpha/2, n_2 - 1, n_1 - 1}$  con  $f_{1 - \alpha/2, n_2 - 1, n_1 - 1}$  come limite superiore nell'Equazione (5.22); il limite inferiore viene impostato a 0. Per trovare l'intervallo di confidenza o i limiti di confidenza di  $\sigma_1/\sigma_2$  estraiamo semplicemente la radice quadrata degli estremi dell'intervallo o dei limiti di confidenza.

## 5.6 INFERENZA SULLE PROPORZIONI DI DUE POPOLAZIONI

Consideriamo ora il caso in cui vi siano due parametri binomiali di interesse, per esempio  $p_1$  e  $p_2$ , e si vogliano trarre inferenze su queste proporzioni. Presenteremo procedure per le verifiche di ipotesi e per gli intervalli di confidenza per grandi campioni, basate sull'approssimazione normale alla distribuzione binomiale.

### 5.6.1 Verifica di ipotesi sull'uguaglianza di due proporzioni binomiali

Si supponga che i due campioni casuali indipendenti di dimensione  $n_1$  e  $n_2$  siano estratti da due popolazioni, e che  $X_1$  e  $X_2$  rappresentino il numero di osservazioni appartenenti, rispettivamente, alla classe di interesse nel campione 1 e nel campione 2. Inoltre, si supponga che l'approssimazione normale della binomiale sia valida per ciascuna popolazione, di modo che gli stimatori delle proporzioni delle popolazioni  $\hat{P}_1 = X_1/n_1$  e  $\hat{P}_2 = X_2/n_2$  abbiano approssimativamente distribuzioni normali. Siamo interessati alla verifica delle ipotesi

$$H_0: p_1 = p_2$$

$$H_1: p_1 \neq p_2$$

**La quantità**

$$Z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \quad (5.23)$$

ha approssimativamente una distribuzione normale standard,  $N(0, 1)$ .

Questo risultato è la base di un test per  $H_0: p_1 = p_2$ . Nello specifico, se l'ipotesi nulla  $H_0: p_1 = p_2$  è vera, sfruttando l'uguaglianza  $p_1 = p_2 = p$ , la variabile aleatoria

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

è approssimativamente distribuita come una  $N(0, 1)$ . Uno stimatore del parametro comune  $p$  è

$$\hat{P} = \frac{X_1 + X_2}{n_1 + n_2}$$

La statistica test per  $H_0: p_1 = p_2$  è allora

$$Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Questo porta alle seguenti procedure di verifica.

### Verifica di ipotesi sull'uguaglianza di due proporzioni binomiali

**Ipotesi nulla:**

$$H_0: p_1 = p_2$$

**Statistica test:**

$$Z_0 = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}(1-\hat{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

**Ipotesi alternativa**

$$H_1: p_1 \neq p_2$$

**P-value**

Probabilità a destra  
di  $|z_0|$  e a sinistra di  $-|z_0|$ ,  
 $P = 2[1 - \Phi(|z_0|)]$

**Criterio di rifiuto**

$$z_0 > z_{\alpha/2} \text{ o } z_0 < -z_{\alpha/2}$$

$$H_1: p_1 > p_2$$

Probabilità a destra di  $z_0$ ,  
 $P = 1 - \Phi(z_0)$

$$z_0 > z_\alpha$$

$$H_1: p_1 < p_2$$

Probabilità a sinistra di  $z_0$ ,  
 $P = \Phi(z_0)$

$$z_0 < -z_\alpha$$

**ESEMPIO 5.12**  
Lenti speciali  
per operati  
di cataratta

Per un possibile impiego in operazioni di finitura nella realizzazione di particolari lenti impiegate dopo l'asportazione di cataratta, vengono valutati due differenti tipi di soluzioni per finitura. Si sono rifinite trecento lenti usando la prima soluzione e, di queste, 253 lenti non presentano difetti indotti dalla lucidatura. Altre 300 lenti sono state rifinite usando la seconda soluzione e, di queste, 196 lenti sono risultate alla fine soddisfacenti. Esiste qualche ragione per affermare che le due soluzioni differiscono?

La procedura di verifica di ipotesi a sette passi porta ai seguenti risultati.

1. **Parametro di interesse:** i parametri di interesse sono  $p_1$  e  $p_2$ , le proporzioni di lenti che sono risultate soddisfacenti dopo l'operazione di finitura con le soluzioni 1 o 2
2. **Ipotesi nulla  $H_0$ :**  $p_1 = p_2$
3. **Ipotesi alternativa  $H_1$ :**  $p_1 \neq p_2$
4. **Statistica test:** la statistica test è

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

dove  $\hat{p}_1 = 253/300 = 0.8433$ ,  $\hat{p}_2 = 196/300 = 0.6533$ ,  $n_1 = n_2 = 300$  e

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{253 + 196}{300 + 300} = 0.7483$$

5. **Rifiutare  $H_0$  se:** si rifiuta  $H_0: p_1 = p_2$  se il  $P$ -value è minore di 0.05.
6. **Calcoli:** il valore della statistica test è

$$z_0 = \frac{0.8433 - 0.6533}{\sqrt{0.7483(0.2517)\left(\frac{1}{300} + \frac{1}{300}\right)}} = 5.36$$

7. **Conclusioni:** siccome  $z_0 = 5.36$ , il  $P$ -value è  $P = 2[1 - \Phi(5.36)] \approx 0$ , e rifiutiamo l'ipotesi nulla. Questo è il valore più prossimo al valore esatto che possiamo ottenere dalla Tavola I dell'Appendice A. Usando una calcolatrice otteniamo  $P \approx 8.32 \times 10^{-8}$ . **Conclusione pratica ingegneristica:** esiste una forte evidenza del fatto che le due soluzioni sono differenti. La soluzione 1 produce una frazione maggiore di lenti non difettose.

Nel seguente riquadro è riportato l'output di Minitab per questo esempio.

**Output di Minitab  
per un test per  
due proporzioni  
e intervallo  
di confidenza per  
l'Esempio 5.12**

<u>Test and CI for Two Proportions</u>			
Sample	X	N	Sample P
1	253	300	0.843333
2	196	300	0.653333
Difference = p(1) – p(2)			
Estimate for difference: 0.19			
95% CI for difference: (0.122236, 0.257764)			
Test for difference = 0 (vs not = 0): Z = 5.36, P-Value = 0.000			

I risultati ottenuti tramite software coincidono con quelli ottenuti tramite calcoli manuali. Oltre ai risultati della verifica di ipotesi, Minitab riporta un intervallo di confidenza bilaterale sulla differenza fra le proporzioni. Forniamo i dettagli nel Paragrafo 5.6.3.

### 5.6.2 Errore del II tipo e scelta della dimensione campionaria

Il calcolo dell'errore  $\beta$  per il precedente test è in una certa misura più involuto rispetto al caso a singolo campione. Il problema è che il denominatore di  $Z_0$  è una stima della deviazione standard di  $\hat{P}_1 - \hat{P}_2$  sotto l'assunzione  $p_1 = p_2 = p$ . Quando  $H_0: p_1 = p_2$  è falsa, la deviazione standard di  $\hat{P}_1 - \hat{P}_2$  è

$$\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}} \quad (5.25)$$

**Errore  $\beta$ :  
test bilaterale  
per la differenza  
tra proporzioni**

Se l'ipotesi alternativa è bilaterale, l'errore  $\beta$  è

$$\begin{aligned} \beta &= \Phi \left[ \frac{z_{\alpha/2} \sqrt{pq(1/n_1 + 1/n_2)} - (p_1 - p_2)}{\sigma_{\hat{P}_1 - \hat{P}_2}} \right] \\ &\quad - \Phi \left[ \frac{-z_{\alpha/2} \sqrt{pq(1/n_1 + 1/n_2)} - (p_1 - p_2)}{\sigma_{\hat{P}_1 - \hat{P}_2}} \right] \end{aligned} \quad (5.26)$$

dove

$$\begin{aligned} \bar{p} &= \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \\ \bar{q} &= \frac{n_1(1 - p_1) + n_2(1 - p_2)}{n_1 + n_2} = 1 - \bar{p} \end{aligned}$$

e  $\sigma_{\hat{P}_1 - \hat{P}_2}$  è data dall'Equazione (5.25).

**Errore  $\beta$ :  
test unilaterale  
per la differenza  
tra proporzioni**

Se l'ipotesi alternativa è  $H_1: p_1 > p_2$ , allora

$$\beta = \Phi \left[ \frac{z_\alpha \sqrt{pq(1/n_1 + 1/n_2)} - (p_1 - p_2)}{\sigma_{\hat{P}_1 - \hat{P}_2}} \right] \quad (5.27)$$

Se l'ipotesi alternativa è  $H_1: p_1 < p_2$ , allora

$$\beta = 1 - \Phi \left[ \frac{-z_\alpha \sqrt{pq(1/n_1 + 1/n_2)} - (p_1 - p_2)}{\sigma_{\hat{P}_1 - \hat{P}_2}} \right] \quad (5.28)$$

Per una determinata coppia di valori  $p_1$  e  $p_2$ , possiamo trovare le dimensioni campionarie  $n_1 = n_2 = n$  necessarie per fornire il test di ampiezza  $\alpha$  che ha un errore del II tipo  $\beta$  specificato. La formula è presentata nel seguente riquadro.

**Dimensione campionaria per una verifica di ipotesi bilaterale  
sulla differenza tra due proporzioni binomiali**

Per l'alternativa bilaterale, la dimensione campionaria comune è

$$n = \frac{(z_{\alpha/2} \sqrt{(p_1 + p_2)(q_1 + q_2)/2} + z_\beta \sqrt{p_1 q_1 + p_2 q_2})^2}{(p_1 - p_2)^2} \quad (5.29)$$

dove  $q_1 = 1 - p_1$  e  $q_2 = 1 - p_2$ .

Per un'alternativa unilaterale, dobbiamo sostituire  $z_{\alpha/2}$  con  $z_\alpha$  nell'Equazione (5.29).

### 5.6.3 Intervallo di confidenza per la differenza tra proporzioni binomiali

L'intervallo di confidenza per  $p_1 - p_2$  può essere trovato direttamente, sapendo che

$$z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}}$$

è una variabile aleatoria normale standard. Pertanto

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \approx 1 - \alpha$$

per cui possiamo inserire l'espressione di  $Z$  nell'ultima formula e usare un approccio simile a quello precedentemente impiegato; troviamo il seguente intervallo di confidenza al  $100(1 - \alpha)\%$  approssimato per  $p_1 - p_2$ .

**Intervallo di confidenza tradizionale per la differenza tra proporzioni binomiali**

Se  $\hat{p}_1$  e  $\hat{p}_2$  sono le proporzioni campionarie delle osservazioni di due campioni casuali indipendenti di dimensione  $n_1$  e  $n_2$  che appartengono a una classe di interesse, un intervallo di confidenza al  $100(1 - \alpha)\%$  approssimato per la differenza tra le vere proporzioni,  $p_1 - p_2$ , è

$$\begin{aligned} \hat{p}_1 - \hat{p}_2 - z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \\ \leq p_1 - p_2 \leq \hat{p}_1 - \hat{p}_2 + z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \end{aligned} \quad (5.30)$$

dove  $z_{\alpha/2}$  è il punto percentuale superiore  $\alpha/2$  della distribuzione normale standard.

**ESEMPIO 5.13**  
**Cuscinetti  
per alberi  
di trasmissione**

Si considerino i cuscinetti per l'albero a gomiti descritti nell'Esempio 4.14. Si supponga che venga apportata una modifica al processo di finitura superficiale, e che, conseguentemente, si ottenga un secondo campione casuale di 85 alberi. Il numero di alberi difettosi in questo secondo campione è 8. Perciò, essendo  $n_1 = 85$ ,  $\hat{p}_1 = 0.12$ ,  $n_2 = 85$ , e  $\hat{p}_2 = 8/85 = 0.09$ , possiamo ottenere un intervallo di confidenza al livello 95% approssimato per la differenza tra le proporzioni di cuscinetti difettosi prodotti nei due processi, usando l'Equazione (5.30) come segue

$$\begin{aligned}\hat{p}_1 - \hat{p}_2 &= z_{0.025} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \\ &\leq p_1 - p_2 \leq \hat{p}_1 - \hat{p}_2 + z_{0.025} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}\end{aligned}$$

ovvero

$$\begin{aligned}0.1176 - 0.0941 - 1.96 \sqrt{\frac{0.1176(0.8824)}{85} + \frac{0.0941(0.9059)}{85}} \\ \leq p_1 - p_2 \leq 0.1176 - 0.0941 + 1.96 \sqrt{\frac{0.1176(0.8824)}{85} + \frac{0.0941(0.9059)}{85}}\end{aligned}$$

Semplificando, otteniamo

$$-0.0685 \leq p_1 - p_2 \leq 0.01155$$

Questo intervallo di confidenza comprende lo zero; perciò, basandoci sui dati campionari, non sembra verosimile che le variazioni apportate al processo di finitura superficiale abbiano ridotto la proporzione di cuscinetti per albero a gomiti difettosi.

L'intervallo di confidenza fornito dall'Equazione (5.30) è quello tradizionalmente utilizzato per la differenza di due proporzioni binomiali. Tuttavia, il livello di confidenza effettivo per questo intervallo può differire notevolmente dal valore nominale o dichiarato. Perciò, se si vuole (per esempio) un intervallo di confidenza al 95% e si usa  $z_{0.025} = 1.96$  nell'Equazioni (5.30), il livello di confidenza effettivo che si rileva può essere alquanto differente dal 95%. Questa situazione può essere fronteggiata con un banale ritocco della procedura: si aggiungono un successo e un insuccesso ai dati di ciascun campione e si calcola quindi

$$\begin{aligned}\tilde{p}_1 &= \frac{X_1 + 1}{n_1 + 2} \quad \text{e} \quad \tilde{n}_1 = n_1 + 2 \\ \tilde{p}_2 &= \frac{X_2 + 1}{n_2 + 2} \quad \text{e} \quad \tilde{n}_2 = n_2 + 2\end{aligned}$$

Dopodiché nell'Equazione (5.30) si sostituiscono  $\hat{p}_1$ ,  $\hat{p}_2$ ,  $\hat{n}_1$  e  $\hat{n}_2$  con  $\tilde{p}_1$ ,  $\tilde{p}_2$ ,  $\tilde{n}_1$  e  $\tilde{n}_2$ .

Per mostrare in concreto come si procede, si consideri nuovamente l'Esempio 5.13. Usando la procedura corretta appena illustrata, troviamo

$$\begin{aligned}\tilde{p}_1 &= \frac{X_1 + 1}{n_1 + 2} = \frac{10 + 1}{85 + 2} = 0.1264 \quad \text{e} \quad \tilde{n}_1 = n_1 + 2 = 85 + 2 = 87 \\ \tilde{p}_2 &= \frac{X_2 + 1}{n_2 + 2} = \frac{8 + 1}{85 + 2} = 0.1034 \quad \text{e} \quad \tilde{n}_2 = n_2 + 2 = 85 + 2 = 87\end{aligned}$$

Se ora sostituiamo  $\hat{p}_1$ ,  $\hat{p}_2$ ,  $\hat{n}_1$  e  $\hat{n}_2$  nell'Equazione (5.30) con i valori di  $\tilde{p}_1$ ,  $\tilde{p}_2$ ,  $\tilde{n}_1$  e  $\tilde{n}_2$  appena calcolati troviamo che il nuovo intervallo di confidenza perfezionato è  $-0.0730 \leq p_1 - p_2 \leq 0.1190$ , simile a quello tradizionale determinato nell'Esempio 5.13. L'ampiezza dell'intervallo tradizionale è 0.1840, quella dell'intervallo ricalcolato è 0.1920. Un'ampiezza leggermente maggiore è probabilmente una riflessione del fatto che la copertura dell'intervallo perfezionato è più vicina al livello consigliato del 95%. Tuttavia, siccome anche questo intervallo di confidenza comprende lo zero, le conclusioni sarebbero le stesse indipendentemente da quale intervallo di confidenza venisse usato.

## 5.7 TABELLE RIASSUNTIVE DELLE PROCEDURE DI INFERNZA PER DUE CAMPIONI

Nelle tabelle in seconda e terza pagina di copertina sono riepilogate tutte le procedure di inferenza a due campioni date in questo capitolo. Tali tabelle contengono gli enunciati delle ipotesi nulle, le statistiche test, i criteri per il rifiuto delle varie ipotesi alternative e le formule per costruire intervalli di confidenza al  $100(1 - \alpha)\%$ .

## 5.8 CASO DI PIÙ DI DUE CAMPIONI

Come è stato illustrato nel Capitolo 4 e in questo capitolo, la verifica e la sperimentazione sono parti naturali dell'analisi ingegneristica e del processo decisionale. Si supponga, per esempio, che un ingegnere civile stia studiando l'effetto di differenti metodi per trattare e migliorare la resistenza media alla compressione di un certo tipo di cemento. L'esperimento consisterebbe nel preparare diversi provini di cemento usando tutti i metodi di trattamento proposti e saggiando quindi la resistenza alla compressione di ciascun provino. I dati di questo esperimento potrebbero quindi essere usati per determinare quale trattamento dovrebbe venire adottato per ottenere la massima resistenza media alla compressione.

Se vi sono solamente due metodi di trattamento di interesse, questo esperimento potrebbe essere progettato e analizzato usando il test  $t$  a due campioni presentato in questo capitolo. In altre parole, lo sperimentatore ha un **singolo fattore** di interesse, il metodo di trattamento, e ci sono solo due **livelli** di tale fattore.

Molti esperimenti a singolo fattore richiedono di prendere in considerazione più di due livelli del fattore. Per esempio, l'ingegnere civile può voler studiare cinque differenti metodi di trattamento del cemento. In questo capitolo mostriamo come usare l'**analisi della varianza** (ANOVA) per confrontare le medie quando vi sono più di due livelli di un singolo fattore. Discuteremo inoltre la **casualizzazione** delle esecuzioni sperimentali e il ruolo importante che questo concetto ha nella strategia sperimentale complessiva.

### 5.8.1 Esperimento completamente casualizzato e analisi della varianza

Un fabbricante di carta impiegata nei sacchetti per alimenti è interessato a migliorare la resistenza alla trazione dei suoi prodotti. Il responsabile della produzione ritiene che la resistenza alla trazione sia funzione della concentrazione di legno di latifoglio presente nell'impasto

con cui si realizza la carta, e che l'intervallo di tale concentrazione di interesse pratico è tra il 5 e il 20%. Una squadra di ingegneri responsabile dello studio decide di analizzare quattro livelli di concentrazione di legno di latifoglio: 5, 10, 15 e 20%, e di preparare sei provini per ciascun livello di concentrazione, usando un impianto pilota. Tutti i 24 provini vengono sottoposti al test di resistenza di laboratorio, in ordine casuale. I dati di questo esperimento sono mostrati in Tabella 5.5.

Tabella 5.5 Resistenza alla trazione della carta (psi).

Concentrazione di legno di latifoglio (%)	Osservazioni						Totali	Medie
	1	2	3	4	5	6		
5	7	8	15	11	9	10	60	10.00
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17.00
20	19	25	22	23	18	20	127	21.17
							383	15.96

Quello appena esposto è un esempio di esperimento a singolo fattore completamente casualizzato con quattro livelli del fattore. I livelli del fattore sono chiamati a volte **trattamenti** e, in questo caso, ogni trattamento ha sei osservazioni o **repliche**. Il ruolo della **casualizzazione** in questo esperimento è estremamente importante. Casualizzando l'ordine delle 24 esecuzioni, si bilancia in via approssimativa l'effetto di ogni variabile di disturbo che possa influenzare la resistenza alla trazione osservata. Per esempio, supponiamo che esista un effetto di riscaldamento della macchina che misura la tensione: per esempio, più a lungo la macchina rimane accesa, più alta è la resistenza alla trazione osservata. Se tutte le 24 esecuzioni si svolgono in ordine di concentrazione di legno di latifoglio crescente (cioè si sottopongono a test prima tutti i sei provini con concentrazione al 5%, poi tutti i sei provini con concentrazione al 10%, e così via), ogni differenza riscontrata tra le resistenze alla trazione potrebbe anche essere imputabile all'effetto di riscaldamento.

È importante analizzare graficamente i dati ricavati da un esperimento pianificato. La Figura 5.7a presenta i box plot della resistenza alla trazione per i quattro livelli di concentrazione di legno di latifoglio. La figura indica che la variazione di tale concentrazione influenza la resistenza alla trazione; più precisamente, più alte concentrazioni di legno di latifoglio portano a osservare valori più elevati di resistenza alla trazione. Inoltre, la distribuzione della resistenza alla trazione a un particolare livello di legno di latifoglio è ragionevolmente simmetrica, e la variabilità della resistenza non cambia drammaticamente al variare della concentrazione di legno di latifoglio.

L'interpretazione grafica dei dati è sempre una buona idea. I box plot mostrano sia la variabilità delle osservazioni *entro* un trattamento (cioè un livello del fattore) sia la variabilità *fra* i trattamenti.

Vediamo ora come i dati ricavati da un esperimento casualizzato a singolo fattore possono venire analizzati statisticamente.

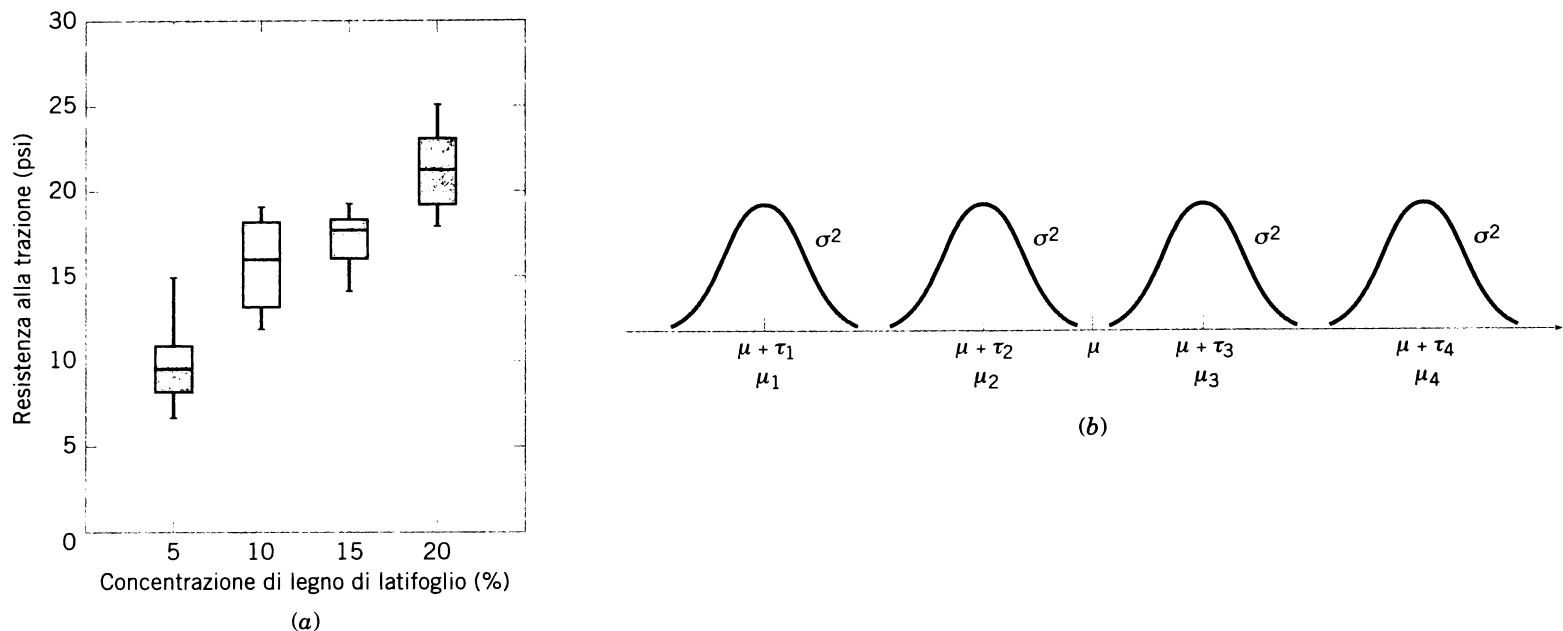


Figura 5.7 (a) Box plot per i dati relativi alla concentrazione di legno di latifoglio. (b) Rappresentazione del modello dell'Equazione (5.31) per l'esperimento a singolo fattore completamente casualizzato.

### Analisi della varianza

Supponiamo di avere  $a$  differenti livelli di un singolo fattore, che vogliamo confrontare. A volte ciascun livello del fattore viene chiamato *trattamento*, un termine molto generale che probabilmente risale alle prime applicazioni della metodologia della pianificazione degli esperimenti alle scienze agrarie. La risposta per ciascuno degli  $a$  trattamenti è una variabile aleatoria. I dati osservati dovrebbero apparire come in Tabella 5.6, dove la voce  $y_{ij}$ , per esempio, indica la  $j$ -esima osservazione rilevata sotto il trattamento  $i$ . Considereremo inizialmente il caso in cui vi è un numero uguale di osservazioni,  $n$ , per ciascun trattamento.

Possiamo descrivere le osservazioni elencate in Tabella 5.6 con il **modello statistico lineare**

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij} \quad \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, n \end{cases} \quad (5.31)$$

Tabella 5.6 Dati tipici per un esperimento a fattore singolo

Trattamento	Osservazioni			Totali	Medie
1	$y_{11}$	$y_{12}$		$y_{1n}$	$y_{1\cdot}$
2	$y_{21}$	$y_{22}$		$y_{2n}$	$\bar{y}_{2\cdot}$
.	.	.		.	.
$a$	$y_{a1}$	$y_{a2}$	..	$y_{an}$	$y_{a\cdot}$
					$\bar{y}_{\cdot\cdot}$

dove  $Y_{ij}$  è una variabile aleatoria che indica la  $ij$ -esima osservazione,  $\mu$  è un parametro comune a tutti i trattamenti chiamato **media complessiva**,  $\tau_i$  è un parametro associato all' $i$ -esimo trattamento chiamato *i*-esimo **effetto del trattamento**, e  $\epsilon_{ij}$  è una componente di errore casuale. Si noti che avremmo potuto scrivere il modello come

**Modello statistico per un esperimento a un singolo fattore.**

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, n \end{cases} \quad (5.31)$$

dove  $\mu_i = \mu + \tau_i$  è la media dell' $i$ -esimo trattamento. Ponendo il modello in questa forma, vediamo che ciascun trattamento definisce una popolazione che ha media  $\mu_i$ , composta dalla media complessiva  $\mu$  più un effetto  $\tau_i$  dovuto a quel particolare trattamento. Assumeremo che gli errori  $\epsilon_{ij}$  siano indipendenti e distribuiti normalmente con media nulla e varianza  $\sigma^2$ . Pertanto, ciascun trattamento può essere pensato come una popolazione normale con media  $\mu_i$  e varianza  $\sigma^2$  (si veda la Figura 5.7b).

L'Equazione (5.31) costituisce il modello sottostante per un esperimento a singolo fattore. Inoltre, siccome si richiede che le osservazioni vengano rilevate in ordine casuale e che le condizioni dell'ambiente (spesso chiamate unità sperimentali) nel quale i trattamenti sono applicati siano il più possibile uniformi, questo piano viene detto **esperimento completamente casualizzato**.

Presentiamo ora l'analisi della varianza per la verifica dell'uguaglianza delle  $a$  medie delle popolazioni. Ricordiamo comunque che l'ANOVA è una tecnica molto più utile e generale; sarà usata in modo estensivo nel prosieguo del volume. In questo paragrafo mostriamo come può essere usata per saggiare l'uguaglianza degli effetti dei trattamenti. Nella nostra applicazione gli effetti dei trattamenti  $\tau_i$  sono usualmente definiti come scarti dalla media complessiva  $\mu$ , per cui

$$\sum_{i=1}^a \tau_i = 0 \quad (5.32)$$

Indichiamo con il termine  $y_{i\cdot}$  il totale delle osservazioni sotto l' $i$ -esimo trattamento, e con  $\bar{y}_{i\cdot}$  la media delle osservazioni sotto l' $i$ -esimo trattamento. Analogamente, indichiamo con  $y_{..}$  il totale globale di tutte le osservazioni, e con  $\bar{y}_{..}$  la media globale di tutte le osservazioni. In termini matematici, abbiamo

$$\begin{aligned} y_{i\cdot} &= \sum_{j=1}^n y_{ij} & \bar{y}_{i\cdot} &= y_{i\cdot}/n & i &= 1, 2, \dots, a \\ y_{..} &= \sum_{i=1}^a \sum_{j=1}^n y_{ij} & \bar{y}_{..} &= y_{..}/N \end{aligned} \quad (5.33)$$

dove  $N = an$  è il numero totale di osservazioni. Quindi il “punto” posto a pedice implica una sommatoria rispetto all'indice che sostituisce.

Siamo interessati a sottoporre a test l'uguaglianza delle  $a$  medie dei trattamenti,  $\mu_1, \mu_2, \dots, \mu_a$ . Usando l'Equazione (5.32), troviamo che questo equivale a verificare le ipotesi

$$\begin{aligned} H_0: \tau_1 &= \tau_2 = \dots = \tau_a = 0 \\ H_1: \tau_i &\neq 0 \quad \text{per almeno una } i \end{aligned} \quad (5.34)$$

Perciò, se l'ipotesi nulla è vera, ciascuna osservazione è composta dalla media complessiva  $\mu$  più una realizzazione della componente di errore casuale  $\epsilon_{ij}$ . Ciò equivale a dire che tutte le  $N$  osservazioni sono prese da una distribuzione normale con media  $\mu$  e varianza  $\sigma^2$ . Pertanto, se l'ipotesi nulla è vera, la variazione dei livelli del fattore non ha influenza sulla risposta media.

L'analisi della varianza suddivide la variabilità totale dei dati campionari in due componenti, quindi la verifica dell'ipotesi dell'Equazione (5.34) è basata su un confronto di due stime indipendenti della varianza della popolazione. La variabilità totale dei dati è descritta dalla **somma totale dei quadrati**

$$SS_T = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2$$

La suddivisione della somma totale dei quadrati è data nella seguente formula.

**L'identità della somma dei quadrati dell'ANOVA è**

$$\sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 = n \sum_{i=1}^a (\bar{y}_{i\cdot} - \bar{y}_{..})^2 + \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{i\cdot})^2 \quad (5.35)$$

La dimostrazione di questa identità è immediata, ed è possibile trovarla in Montgomery, Runger (2011).

L'identità dell'Equazione (5.35) mostra che la variabilità totale dei dati, misurata dalla somma totale dei quadrati, può essere suddivisa in una somma dei quadrati delle differenze tra le medie del trattamento e la media globale, più una somma dei quadrati delle differenze tra le osservazioni entro un trattamento e la media del trattamento. Le differenze tra le medie del trattamento osservate e la media globale misura le differenze fra i trattamenti, mentre le differenze delle osservazioni entro un trattamento e la media del trattamento possono essere dovute solo all'errore casuale. In conclusione, scriviamo l'Equazione (5.35) in modo simbolico come

$$SS_T = SS_{\text{Trattamenti}} + SS_E \quad (5.36)$$

dove

$$SS_T = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{..})^2 = \text{somma totale dei quadrati}$$

$$SS_{\text{Trattamenti}} = n \sum_{i=1}^a (\bar{y}_{i\cdot} - \bar{y}_{..})^2 = \text{somma dei quadrati del trattamento}$$

e

$$SS_E = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \bar{y}_{i\cdot})^2 = \text{somma dei quadrati degli errori}$$

Possiamo comprendere più approfonditamente come funziona l'analisi della varianza esaminando i valori attesi di  $SS_{\text{Trattamenti}}$  e di  $SS_E$ . Questo ci porterà a una statistica appropriata per la verifica dell'ipotesi di nessuna differenza tra le medie dei trattamenti (ovvero  $\tau_i = 0$ ).

Si può dimostrare che

$$E\left(\frac{SS_{\text{Trattamenti}}}{a - 1}\right) = \sigma^2 + \frac{n \sum_{i=1}^a \tau_i^2}{a - 1} \quad (5.37)$$

Il rapporto

$$MS_{\text{Trattamenti}} = SS_{\text{Trattamenti}}/(a - 1)$$

**Media quadratica per i trattamenti.**

è chiamato **media quadratica (mean square) per i trattamenti**. Quindi, se  $H_0$  è vera,  $MS_{\text{Trattamenti}}$  è uno stimatore non distorto di  $\sigma^2$  perché sotto  $H_0$  ogni  $\tau_i = 0$ . Se  $H_1$  è vera,  $MS_{\text{Trattamenti}}$  stima  $\sigma^2$  più un termine positivo che incorpora la variazione dovuta alla differenza sistematica tra le medie dei trattamenti.

**Errore quadratico medio.**

Possiamo anche dimostrare che il valore atteso della somma dei quadrati degli errori è  $E(SS_E) = a(n - 1)\sigma^2$ . L'**errore quadratico medio**

$$MS_E = SS_E/[a(n - 1)]$$

è dunque uno stimatore non distorto di  $\sigma^2$  indipendentemente dal fatto che  $H_0$  sia vera o no.

Esiste anche una suddivisione del numero di gradi di libertà che corrisponde all'identità della somma dei quadrati dell'Equazione (5.35). Cioè, vi sono  $an = N$  osservazioni;  $SS_T$  ha allora  $an - 1$  gradi di libertà. Vi sono  $a$  livelli del fattore, perciò  $SS_{\text{Trattamenti}}$  ha  $a - 1$  gradi di libertà. Infine, entro ogni trattamento vi sono  $n$  repliche che forniscono  $n - 1$  gradi di libertà con cui stimare l'errore sperimentale. Siccome vi sono  $a$  trattamenti, abbiamo  $a(n - 1)$  gradi di libertà per l'errore. Pertanto, la suddivisione dei gradi di libertà è

$$an - 1 = a - 1 + a(n - 1)$$

Assumiamo ora che ciascuna delle  $a$  popolazioni possa essere modellizzata mediante una distribuzione normale. Usando questa assunzione, possiamo dimostrare che se l'ipotesi nulla  $H_0$  è vera, il rapporto

$$F_0 = \frac{SS_{\text{Trattamenti}}/(a - 1)}{SS_E/[a(n - 1)]} = \frac{MS_{\text{Trattamenti}}}{MS_E} \quad (5.38)$$

ha una distribuzione  $F$  con  $a - 1$  e  $a(n - 1)$  gradi di libertà. Inoltre, dai valori attesi delle medie quadratiche, sappiamo che  $MS_E$  è uno stimatore non distorto di  $\sigma^2$ . Sotto l'ipotesi nulla, anche  $MS_{\text{Trattamenti}}$  è uno stimatore non distorto di  $\sigma^2$ . Tuttavia, se l'ipotesi nulla è falsa, il valore atteso di  $MS_{\text{Trattamenti}}$  è più grande di  $\sigma^2$ , perciò sotto l'ipotesi alternativa, il valore atteso del numeratore della statistica test (Equazione (5.38)) è maggiore del valore atteso del denominatore. Di conseguenza, dovremmo rifiutare  $H_0$  se la statistica è grande. Questo implica una regione critica con una singola coda superiore. Rifiuteremmo quindi  $H_0$

se  $f_0 > f_{\alpha,a-1,a(n-1)}$ , dove  $f_0$  è calcolato dall'Equazione (5.38). Questi risultati sono riepilogati di seguito.

### Sintesi

#### Verifica di ipotesi su più di due medie (ANOVA)

$$MS_{\text{Trattamenti}} = \frac{SS_{\text{Trattamenti}}}{a - 1} \quad E(MS_{\text{Trattamenti}}) = \sigma^2 + \frac{n \sum_{i=1}^a \tau_i^2}{a - 1}$$

$$MS_E = \frac{SS_E}{a(n - 1)} \quad E(MS_E) = \sigma^2$$

Ipotesi nulla:  $H_0: \tau_1 = \tau_2 = \dots = \tau_a = 0$

Ipotesi alternativa:  $H_1: \tau_i \neq 0 \quad \text{per almeno una } i$

Statistica test:  $F_0 = \frac{MS_{\text{Trattamenti}}}{MS_E}$

Criterio di rifiuto:  $f_0 > f_{\alpha,a-1,a(n-1)}$

Si possono ricavare formule di calcolo efficienti per la somma dei quadrati sviluppando e semplificando le definizioni di  $SS_{\text{Trattamenti}}$  e di  $SS_T$ . Questo porta ai risultati riassunti di seguito.

#### Esperimento completamente casualizzato con dimensioni campionarie uguali

Le formule per il calcolo della somma dei quadrati nell'analisi della varianza per un esperimento completamente casualizzato con dimensioni campionarie uguali in ogni trattamento sono:

$$SS_T = \sum_{i=1}^a \sum_{j=1}^n y_{ij}^2 - \frac{y_{..}^2}{N}$$

e

$$SS_{\text{Trattamenti}} = \sum_{i=1}^a \frac{y_{i..}^2}{n} - \frac{y_{..}^2}{N}$$

La somma dei quadrati degli errori si ottiene in genere per sottrazione:

$$SS_E = SS_T - SS_{\text{Trattamenti}}$$

I calcoli per questa procedura di verifica sono generalmente riepilogati in forma tabulare come mostrato in Tabella 5.7; tale forma viene chiamata **tabella dell'analisi della varianza** (o ANOVA).

Tabella 5.7 L'analisi della varianza per un esperimento a singolo fattore.

Causa della variazione	Somma dei quadrati	Gradi di libertà	Media quadratica	$F_0$
Trattamenti	$SS_{\text{Trattamenti}}$	$a - 1$	$MS_{\text{Trattamenti}}$	$\frac{MS_{\text{Trattamenti}}}{MS_E}$
Errore	$SS_E$	$a(n - 1)$	$MS_E$	
Totale	$SS_T$	$an - 1$		

### ESEMPIO 5.14

Resistenza alla trazione

Si consideri l'esperimento relativo alla resistenza alla trazione per la carta, descritto nel Paragrafo 5.8.1. Possiamo usare l'analisi della varianza per verificare l'ipotesi che differenti concentrazioni di legno di latifoglio non influiscono sulla resistenza media alla trazione della carta.

La procedura di verifica delle ipotesi a sette passi porta ai seguenti risultati.

1. **Parametro di interesse:** i parametri di interesse sono  $\tau_1, \tau_2, \tau_3$  e  $\tau_4$ , le resistenze medie alla trazione della carta delle quattro differenti concentrazioni di legno duro.
2. **Ipotesi nulla  $H_0$ :**  $\tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$
3. **Ipotesi alternativa  $H_1$ :**  $\tau_i \neq 0$  per almeno una  $i$
4. **Statistica test:** la statistica test è

$$f_0 = \frac{MS_{\text{Trattamenti}}}{MS_E}$$

5. **Rifiutare  $H_0$  se:** si rifiuta  $H_0$  se il  $P$ -value è minore di 0.05.
6. **Calcoli:** le somme dei quadrati per l'ANOVA sono calcolate dalle Equazioni (5.39), (5.40) e (5.41) come segue:

$$\begin{aligned} SS_T &= \sum_{i=1}^4 \sum_{j=1}^6 y_{ij}^2 - \frac{y_{..}^2}{N} \\ &= (7)^2 + (8)^2 + \dots + (20)^2 - \frac{(383)^2}{24} = 512.96 \end{aligned}$$

$$\begin{aligned} SS_{\text{Trattamenti}} &= \sum_{i=1}^4 \frac{y_{i..}^2}{n} - \frac{y_{..}^2}{N} \\ &= \frac{(60)^2 + (94)^2 + (102)^2 + (127)^2}{6} - \frac{(383)^2}{24} = 382.79 \end{aligned}$$

$$\begin{aligned} SS_E &= SS_T - SS_{\text{Trattamenti}} \\ &= 512.96 - 382.79 = 130.17 \end{aligned}$$

Di solito non si eseguono questi calcoli a mano. In Tabella 5.8 è presentato il calcolo dell'ANOVA effettuato con Minitab.

Tabella 5.8 Output di Minitab dell'analisi della varianza per l'esperimento di resistenza alla trazione della carta.

One-Way Analysis of Variance					
Analysis of Variance					
Source	DF	SS	MS	F	P
Factor	3	382.79	127.60	19.61	0.000
Error	20	130.17	6.51		
Total	23	512.96			
Individual 95% CIs For Mean Based on Pooled StDev					
Level	N	Mean	StDev	----- + ----- + ----- + ----- + -	
5	6	10.000	2.828	(---*---)	
10	6	15.667	2.805	(---*---)	
15	6	17.000	1.789	(---*---)	
20	6	21.167	2.639	(---*---)	
----- + ----- + ----- + ----- + -					
Pooled StDev = 2.551				10.0	15.0
				20.0	25.0

7. **Conclusioni:** in Tabella 5.8 notiamo che il valore della statistica test calcolato è  $f_0 = 19.61$  e il  $P$ -value è  $P = 0.000$  (il  $P$ -value non può essere veramente nullo; Minitab produce per default questo valore di output quando il  $P$ -value è minore di 0.001). Essendo il  $P$ -value nettamente minore di  $\alpha = 0.05$ , vi sono forti indizi per concludere che l'ipotesi nulla non è vera. Vale a dire, la concentrazione di legno di latifoglio influenza la resistenza alla trazione della carta. Trattandosi di un test  $F$  a coda superiore, potremmo trovare un limite per il  $P$ -value usando la tavola IV dell'Appendice A; vedremmo allora che  $f_{0.01,3,20} = 4.94$ . Poiché  $f_0 = 19.61$  supera questo valore, sappiamo che il  $P$ -value è minore di 0.01. Il valore effettivo (determinabile con una calcolatrice) è  $3.59 \times 10^{-6}$ . Si noti inoltre che Minitab fornisce qualche informazione riassuntiva su ciascun livello di concentrazione di legno di latifoglio, compreso l'intervallo di confidenza su ciascuna media.

In alcuni esperimenti a singolo fattore, il numero di osservazioni rilevate sotto ciascun trattamento può essere differente. Si dice allora che il piano sperimentale è **non bilanciato**. L'analisi della varianza descritta in precedenza è ancora valida, ma si devono apportare leggere modifiche alle formule delle somme dei quadrati. Indichiamo con  $n_i$  il numero di osservazioni che sono prese sotto il trattamento  $i$  ( $i = 1, 2, \dots, a$ ), e sia  $N = \sum_{i=1}^a n_i$  il numero totale di osservazioni. Le formule di calcolo per  $SS_T$  e  $SS_{\text{Trattamenti}}$  sono quelle riportate nel seguente riquadro.

### Esperimento completamente casualizzato con dimensioni campionarie disuguali

Le formule di calcolo per le somme dei quadrati nell'analisi della varianza per un esperimento completamente casualizzato con dimensioni campionarie disuguali  $n_i$  in ciascun trattamento sono

$$SS_T = \sum_{i=1}^a \sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N}$$

$$SS_{\text{Trattamenti}} = \sum_{i=1}^a \frac{y_{i..}^2}{n_i} - \frac{y_{..}^2}{N}$$

e

$$SS_E = SS_T - SS_{\text{Trattamenti}}$$

#### Quali medie differiscono?

In conclusione, si osservi che l'analisi della varianza ci dice se esiste una differenza tra le medie, ma non ci dice quali medie differiscono. Se l'analisi della varianza indica che esiste una differenza statisticamente significativa tra le medie, si può usare una semplice procedura grafica per isolare le specifiche differenze. Supponiamo che,  $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_a$ , siano le medie osservate per questi livelli del fattore. Ciascuna media di trattamento ha una deviazione standard  $\sigma/\sqrt{n}$ , dove  $\sigma$  è la deviazione standard di una osservazione individuale. Se tutte le medie del trattamento sono uguali, le medie osservate  $\bar{y}_i$  si comporterebbero come se fossero un insieme di osservazioni estratte a caso da una distribuzione normale con media  $\mu$  e deviazione standard  $\sigma/\sqrt{n}$ .

Si immagini questa densità normale libera di muoversi lungo un asse sotto cui sono riportate le medie di trattamento  $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_a$ . Se tutte le medie di trattamento sono uguali, ci dovrebbe essere qualche posizione per questa densità che rende evidente che i valori  $\bar{y}_i$  sono stati estratti dalla medesima distribuzione. Se ciò non avviene, i valori  $\bar{y}_i$  che non appaiono essere stati estratti da questa distribuzione vengono associati ai trattamenti che producono differenti risposte medie.

Il solo difetto in questa logica è che  $\sigma$  è incognita. Tuttavia, possiamo usare  $\sqrt{MS_E}$  dall'analisi della varianza per stimare  $\sigma$ . Ciò implica dover usare una distribuzione  $t$  invece di quella normale per fare il grafico, ma poiché la distribuzione  $t$  assomiglia molto a quella normale, generalmente funziona molto bene anche schizzare una curva normale che sia approssimativamente ampia  $6\sqrt{MS_E}/n$  unità.

La Figura 5.8 mostra quanto suggerito sopra per l'esperimento riguardante la concentrazione di legno di latifoglio. La deviazione standard di questa distribuzione normale è

$$\sqrt{MS_E/n} = \sqrt{6.51/6} = 1.04$$

Se immaginiamo di traslare questa distribuzione lungo l'asse orizzontale, notiamo che non esiste una posizione che suggerisca che tutte le quattro osservazioni (le medie riportate in grafico) siano valori tipici, selezionati in maniera casuale dalla distribuzione. Questo, ovviamente, era prevedibile, dato che l'analisi della varianza ha indicato che le medie differiscono.

**Esame delle differenze fra le medie.**

no; la rappresentazione in Figura 5.8 è solo una rappresentazione grafica dei risultati dell'analisi della varianza. La figura indica che il trattamento 4 (20% di legno di latifoglio) produce la carta con la più alta resistenza media alla trazione rispetto agli altri trattamenti, e che il trattamento 1 (5% di legno di latifoglio) comporta la più bassa resistenza media alla trazione fra i vari trattamenti. Le medie dei trattamenti 2 e 3 (rispettivamente 10 e 15% di legno di latifoglio) non differiscono.

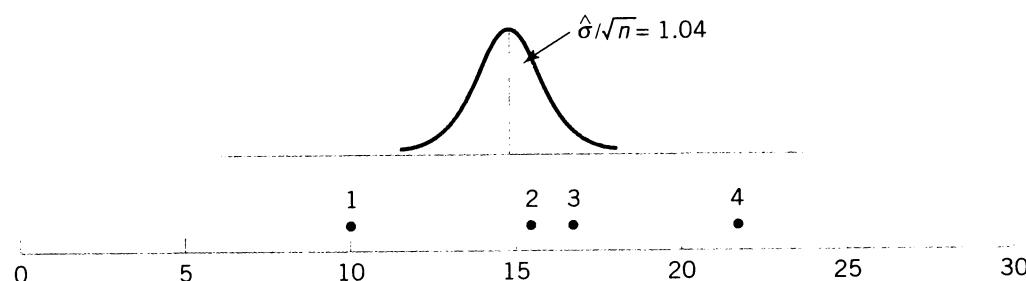


Figura 5.8 Medie della resistenza all trazione tratte dall'esperimento relativo alla concentrazione di legno di latifoglio in relazione a una distribuzione normale con deviazione standard  $\sqrt{MS_E/n} = \sqrt{6.51/6} = 1.04$ .

Questa semplice procedura, pur essendo grossolana, è una tecnica molto utile ed efficace per confrontare le medie dopo un'analisi della varianza. Esistono molte tecniche quantitative, chiamate **procedure di confronto multiplo**, per valutare le differenze tra le medie specifiche dopo un'analisi della varianza. Siccome queste procedure coinvolgono in genere una serie di test, l'errore del tipo I si combina per produrre un **tasso di errore “della famiglia di test” o errore experiment-wise**. Per maggiori dettagli su queste procedure si veda Montgomery (2009).

#### Analisi dei residui e verifica del modello

L'analisi della varianza "a una via" (o a un singolo fattore) assume che le osservazioni siano indipendenti e distribuite normalmente con la medesima varianza per ciascun trattamento o livello del fattore. Queste assunzioni dovrebbero essere controllate esaminando i residui. Un residuo è la differenza tra un'osservazione  $y_{ij}$  e il suo valore stimato ricavato dal modello statistico studiato, indicato con  $\hat{y}_{ij}$ . Per il piano completamente casualizzato  $\hat{y}_{ij} = \bar{y}_i$ , e ogni residuo è  $e_{ij} = y_{ij} - \bar{y}_i$ , cioè la differenza tra un'osservazione e la corrispondente media del trattamento osservata. I residui per l'esperimento riguardante la percentuale di legno di latifoglio sono mostrati in Tabella 5.9. Usando  $\bar{y}_i$  per calcolare ciascun residuo, si rimuove essenzialmente l'effetto della concentrazione di legno di latifoglio da quei dati; conseguentemente, i residui contengono informazioni riguardo la variabilità non spiegata.

#### Grafici dei residui.

L'assunzione di normalità può essere controllata costruendo un grafico dei quantili per i residui. Per controllare l'assunzione di varianze uguali per ciascun livello del fattore, riportiamo i residui in funzione dei livelli del fattore e confrontiamo l'allargamento nei residui. È utile anche riportare i residui in funzione di  $\bar{y}_i$  (detto a volte valore stimato); la variabilità nei residui non dovrebbe dipendere in alcun modo dal valore di  $\bar{y}_i$ . La maggior parte dei pacchetti di software statistico sono in grado di costruire questi tipi di grafici. Quando si

Tabella 5.9 Residui per l'esperimento di resistenza alla trazione.

Centrazione di legno (%)	Residui						
	5	-3.00	-2.00	5.00	1.00	-1.00	0.00
10	-3.67	1.33	-2.67	2.33	3.33	-0.67	
15	-3.00	1.00	2.00	0.00	-1.00	1.00	
20	-2.17	3.83	0.83	1.83	-3.17	-1.17	

rileva un andamento riconoscibile in questi grafici, esso suggerisce in genere la necessità di una trasformazione, ossia di analizzare i dati in una metrica differente. Per esempio, se la variabilità nei residui aumenta con  $\bar{y}_i$ , si dovrebbe prendere in considerazione una trasformazione come  $\log y$  o  $\sqrt{y}$ . In alcuni problemi, la dipendenza della dispersione dei residui rispetto dalla media osservata  $\bar{y}_i$  è un'informazione molto importante; può essere utile selezionare il livello di fattore che dà luogo alla risposta massima; tuttavia, questo livello può anche causare ulteriore variazione nella risposta da esecuzione a esecuzione.

L'assunzione di indipendenza può essere controllata rappresentando graficamente i residui in funzione del tempo o in funzione dell'ordine di esecuzione in cui l'esperimento è stato svolto. Un andamento riconoscibile in questo grafico, come successioni di residui positivi e negativi, può indicare che le osservazioni non sono indipendenti. Ciò suggerisce che il tempo o l'ordine di esecuzione sono importanti, o che vi sono importanti quantità che variano nel tempo non incluse nel piano sperimentale.

Un grafico dei quantili per i residui ricavati dall'esperimento relativo alla resistenza alla trazione della carta è mostrato in Figura 5.9. Le Figure 5.10 e 5.11 presentano i grafici dei residui in funzione, rispettivamente, dei livelli del fattore e del valore stimato  $\bar{y}_i$ . Questi grafici non rivelano alcuna inadeguatezza del modello o problema inusuale con le assunzioni.

Interpretazione  
del grafico dei quantili  
normali.

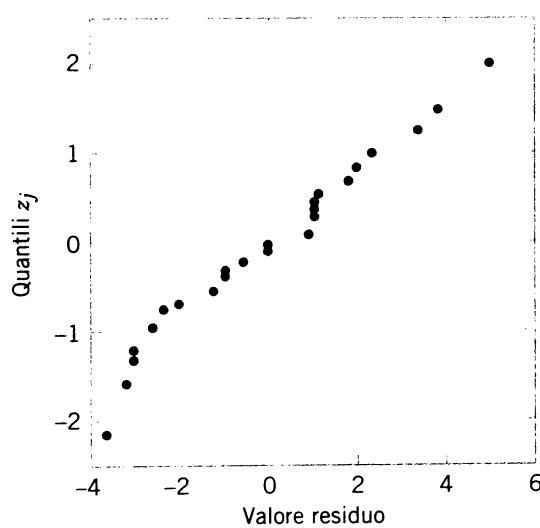


Figura 5.9 Grafico dei quantili per i residui per l'esperimento relativo alla concentrazione di latifoglio.

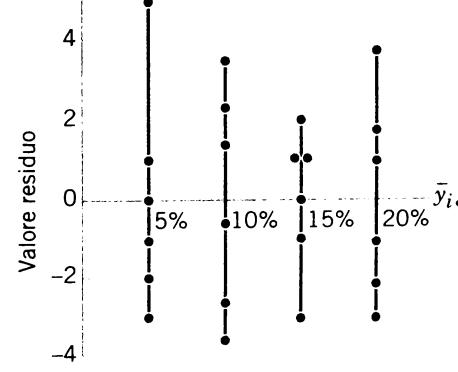


Figura 5.10 Grafico dei residui in funzione dei livelli del fattore (concentrazione di legno di latifoglio).

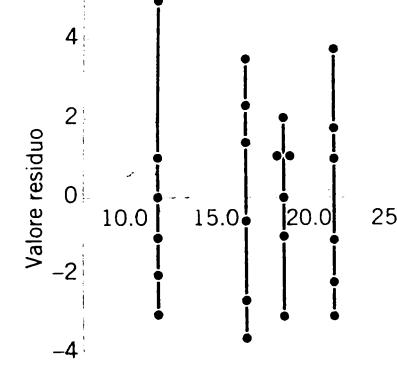


Figura 5.11 Grafico dei residui in funzione di  $\bar{y}_i$ .

### 5.8.2 Esperimento a blocchi completi casualizzati

In molti problemi concernenti i piani sperimentali è necessario pianificare l'esperimento in modo che la variabilità proveniente da un fattore di disturbo possa essere controllata. Per esempio, si consideri la situazione dell'Esempio 5.8, dove sono stati usati due metodi differenti per predire la resistenza al taglio delle travi a lamine di acciaio. Siccome ciascuna trave ha (potenzialmente) differenti resistenze e questa variabilità nella resistenza non era di interesse diretto, abbiamo pianificato l'esperimento usando i due metodi di test su ciascuna trave e confrontando quindi con lo zero la differenza media tra le misure di resistenza su ciascuna trave mediante il test  $t$  accoppiato. Il test  $t$  accoppiato è una procedura per confrontare due medie di trattamento quando le esecuzioni sperimentali non possono essere eseguite tutte sotto condizioni omogenee. In alternativa, possiamo considerare il test  $t$  accoppiato come un metodo per ridurre il rumore di fondo dell'esperimento neutralizzando l'effetto di un **fattore di disturbo**. Il blocco è il fattore di disturbo; in questo caso, si tratta dell'**unità sperimentale** effettiva, cioè dei provini di trave di acciaio usati nell'esperimento.

Il piano a blocchi casualizzato è un'estensione del test  $t$  accoppiato alle situazioni in cui il fattore di interesse ha più di due livelli; in altre parole, si devono confrontare più di due trattamenti. Per esempio, si supponga che possano venire usati tre metodi per valutare la resistenza delle travi a lamine di acciaio. Possiamo vedere questi metodi come tre trattamenti, che indichiamo per esempio con  $t_1$ ,  $t_2$  e  $t_3$ . Se usiamo quattro travi come unità sperimentali, un **piano a blocchi completi casualizzati** dovrebbe apparire come in Figura 5.12. Il piano è chiamato piano a blocchi completi casualizzati perché ciascun blocco è abbastanza grande da contenere tutti i trattamenti e perché l'assegnazione effettiva di ciascuno dei tre trattamenti all'interno di ogni blocco è eseguita casualmente. Una volta che è stato condotto l'esperimento, i dati sono registrati in una tabella come quella di Tabella 5.10. Le osservazioni in questa tabella, cioè i termini  $y_{ij}$ , rappresentano la risposta ottenuta con l'impiego del metodo  $i$  sulla trave  $j$ .

La procedura generale per un piano a blocchi completi casualizzati consiste nella selezione di  $b$  blocchi e nell'esecuzione di una replica completa dell'esperimento in ciascun blocco. I dati che risultano dall'esecuzione di un piano a blocchi completi casualizzati per lo

Tabella 5.10 Un piano a blocchi completi casualizzati.

Treattamento (Metodo)	Blocco (Trave)			
	1	2	3	4
1	$y_{11}$	$y_{12}$	$y_{13}$	$y_{14}$
2	$y_{21}$	$y_{22}$	$y_{23}$	$y_{24}$
3	$y_{31}$	$y_{32}$	$y_{33}$	$y_{34}$

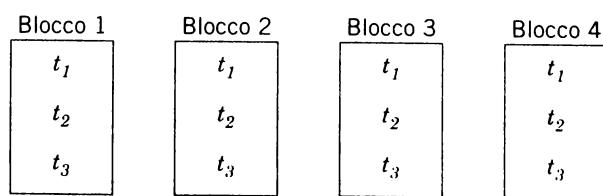


Figura 5.12  
Un piano a blocchi completi casualizzati.

studio di un singolo fattore con  $a$  livelli e  $b$  blocchi sono mostrati in Tabella 5.11. Vi saranno  $a$  osservazioni (una per livello di fattore) in ciascun blocco, e l'ordine nel quale queste osservazioni sono eseguite è assegnato in modo casuale dentro il blocco.

**Tabella 5.11** Un piano a blocchi completi casualizzati con  $a$  trattamenti e  $b$  blocchi.

<b>Trattamenti</b>	<b>Blocchi</b>				<b>Totali</b>	<b>Medie</b>
	1	2	...	$b$		
1	$y_{11}$	$y_{12}$		$y_{1b}$	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	$y_{21}$	$y_{22}$		$y_{2b}$	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
.	.	.		.	.	.
$a$	$y_{a1}$	$y_{a2}$		$y_{ab}$	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
Totali	$y_{\cdot 1}$	$y_{\cdot 2}$		$y_{\cdot b}$	$y_{\cdot \cdot}$	
Medie	$\bar{y}_{\cdot 1}$	$\bar{y}_{\cdot 2}$	...	$\bar{y}_{\cdot b}$		$\bar{y}_{\cdot \cdot}$

Descriviamo ora l'ANOVA per un piano a blocchi completi casualizzati. Si supponga che vi sia un singolo fattore di interesse con  $a$  livelli e che l'esperimento sia eseguito in  $b$  blocchi. Le osservazioni possono essere rappresentate dal modello statistico lineare

$$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij} \begin{cases} i = 1, 2, \dots, a \\ j = 1, 2, \dots, b \end{cases} \quad (5.39)$$

dove  $\mu$  è una media complessiva,  $\tau_i$  è l'effetto dell' $i$ -esimo trattamento,  $\beta_j$  è l'effetto del  $j$ -esimo blocco e gli  $\epsilon_{ij}$  sono i termini di errore casuale che si assumono indipendenti e distribuiti normalmente con media nulla e varianza  $\sigma^2$ . Per il nostro scopo, i trattamenti e i blocchi saranno considerati come fattori fissati. Inoltre, gli effetti dei trattamenti e dei blocchi sono definiti come scarti dalla media complessiva, per cui  $\sum_{i=1}^a \tau_i = 0$  e  $\sum_{j=1}^b \beta_j = 0$ . Inoltre, assumiamo che i trattamenti e i blocchi non interagiscano, cioè che l'effetto del trattamento  $i$  sia lo stesso indipendentemente dal blocco (o dai blocchi) in cui è saggiato. Siamo interessati a verificare l'uguaglianza degli effetti del trattamento, cioè a verificare le ipotesi

$$\begin{aligned} H_0: \tau_1 &= \tau_2 = \dots = \tau_a = 0 \\ H_1: \tau_i &\neq 0 \text{ per almeno una } i \end{aligned} \quad (5.40)$$

Come nell'esperimento completamente casualizzato, verificare l'ipotesi che tutti gli effetti di trattamento  $\tau_i$  sono uguali a zero è equivalente a verificare l'ipotesi che le medie dei trattamenti sono uguali.

La procedura dell'ANOVA per il piano a blocchi completi casualizzati impiega una identità della somma dei quadrati che suddivide la somma totale dei quadrati in tre componenti.

**L'identità della somma dei quadrati per il piano a blocchi completi casualizzati è**

$$\begin{aligned} \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y}_{..})^2 &= b \sum_{i=1}^a (\bar{y}_{i\cdot} - \bar{y}_{..})^2 + a \sum_{j=1}^b (\bar{y}_{\cdot j} - \bar{y}_{..})^2 \\ &\quad + \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y}_{\cdot j} - \bar{y}_{i\cdot} + \bar{y}_{..})^2 \end{aligned} \quad (5.41)$$

Tale identità della somma dei quadrati può essere rappresentata simbolicamente come

$$SS_T = SS_{\text{Trattamenti}} + SS_{\text{Blocchi}} + SS_E$$

dove

$$SS_T = \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y}_{..})^2 = \text{somma totale dei quadrati}$$

$$SS_{\text{Trattamenti}} = b \sum_{i=1}^a (\bar{y}_{i\cdot} - \bar{y}_{..})^2 = \text{somma dei quadrati di trattamento}$$

$$SS_{\text{Blocchi}} = a \sum_{j=1}^b (\bar{y}_{\cdot j} - \bar{y}_{..})^2 = \text{somma dei quadrati di blocco}$$

$$SS_E = \sum_{i=1}^a \sum_{j=1}^b (\bar{y}_{ij} - \bar{y}_{\cdot j} - \bar{y}_{i\cdot} + \bar{y}_{..})^2 = \text{somma dei quadrati degli errori}$$

Inoltre, la suddivisione dei gradi di libertà corrispondente a queste somme dei quadrati è

$$ab - 1 = (a - 1) + (b - 1) + (a - 1)(b - 1)$$

Per il piano a blocchi casualizzati, le medie quadratiche rilevanti sono

$$MS_{\text{Trattamenti}} = \frac{SS_{\text{Trattamenti}}}{a - 1} \quad MS_{\text{Blocchi}} = \frac{SS_{\text{Blocchi}}}{b - 1} \quad MS_E = \frac{SS_E}{(a - 1)(b - 1)} \quad (5.42)$$

Si può dimostrare che i valori attesi di queste medie quadratiche sono uguali a

$$\begin{aligned} E(MS_{\text{Trattamenti}}) &= \sigma^2 + \frac{b \sum_{i=1}^a \tau_i^2}{a - 1} \\ E(MS_{\text{Blocchi}}) &= \sigma^2 + \frac{a \sum_{j=1}^b \beta_j^2}{b - 1} \\ E(MS_E) &= \sigma^2 \end{aligned}$$

Pertanto, se l'ipotesi nulla  $H_0$  è vera, in modo che tutti gli effetti di trattamento  $\tau_i = 0$ ,  $MS_{\text{Trattamenti}}$  è uno stimatore non distorto di  $\sigma^2$ , mentre se  $H_0$  è falsa,  $MS_{\text{Trattamenti}}$  sovrastima  $\sigma^2$ . La media quadratica degli errori è sempre uno stimatore non distorto di  $\sigma^2$ . Per verificare l'ipotesi nulla che gli effetti di trattamento sono tutti nulli, calcoliamo il rapporto

$$F_0 = \frac{MS_{\text{Trattamenti}}}{MS_E} \quad (5.43)$$

che ha una distribuzione  $F$  con  $a - 1$  e  $(a - 1)(b - 1)$  gradi di libertà se è vera l'ipotesi nulla. Rifiuteremo l'ipotesi nulla al livello di significatività  $\alpha$  se il valore calcolato della statistica test nell'Equazione (5.43) è

$$f_0 > f_{\alpha, a-1, (a-1)(b-1)}$$

In pratica, calcoliamo  $SS_T$ ,  $SS_{\text{Trattamenti}}$  e  $SS_{\text{Blocchi}}$ , quindi otteniamo la somma dei quadrati degli errori  $SS_E$  tramite sottrazione. Le formule di calcolo appropriate sono le seguenti.

### Esperimento a blocchi completi casualizzati

Le formule di calcolo per le somme dei quadrati nell'analisi della varianza per un piano a blocchi completi casualizzati sono

$$\begin{aligned} SS_T &= \sum_{i=1}^a \sum_{j=1}^b y_{ij}^2 - \frac{y_{..}^2}{ab} \\ SS_{\text{Trattamenti}} &= \frac{1}{b} \sum_{i=1}^a y_{i.}^2 - \frac{y_{..}^2}{ab} \\ SS_{\text{Blocchi}} &= \frac{1}{a} \sum_{j=1}^b y_{.j}^2 - \frac{y_{..}^2}{ab} \end{aligned}$$

e

$$SS_E = SS_T - SS_{\text{Trattamenti}} - SS_{\text{Blocchi}}$$

I calcoli vengono in genere raccolti in una tabella di analisi della varianza, come mostrato in Tabella 5.12. Di solito per eseguire l'ANOVA per il piano a blocchi completi casualizzati si impiega un pacchetto software. Nel seguente esempio diamo solo i risultati del calcolo anziché elencare esplicitamente la procedura a sette passi.

**Tabella 5.12** Analisi della varianza per un piano a blocchi completi casualizzati

Causa della variazione	Somma dei quadrati	Gradi di libertà	Media quadratica	$F_0$
Trattamenti	$SS_{\text{Trattamenti}}$	$a - 1$	$\frac{SS_{\text{Trattamenti}}}{a - 1}$	$\frac{MS_{\text{Trattamenti}}}{MS_E}$
Blocchi	$SS_{\text{Blocchi}}$	$b - 1$	$\frac{SS_{\text{Blocchi}}}{b - 1}$	
Errori	$SS_E$ (per sottrazione)	$(a - 1)(b - 1)$	$\frac{SS_E}{(a - 1)(b - 1)}$	
Totali	$SS_T$	$ab - 1$		

**ESEMPIO 5.15**  
**Resistenza  
di un tessuto**

Si è eseguito un esperimento per determinare l'effetto di quattro differenti sostanze chimiche sulla resistenza di un tessuto. Queste sostanze chimiche vengono impiegate nel processo finale di stampa del tessuto. Sono stati selezionati cinque esemplari di tessuto, ed è stato eseguito un piano a blocchi completi casualizzati, provando i diversi tipi di sostanza su ciascun campione di tessuto, in ordine casuale. I dati ottenuti sono mostrati in Tabella 5.13. Valuteremo le differenze tra le medie usando l'analisi della varianza con  $\alpha = 0.01$ . Le somme dei quadrati per l'ANOVA sono calcolate come segue:

$$\begin{aligned}
 SS_T &= \sum_{i=1}^4 \sum_{j=1}^5 y_{ij}^2 - \frac{y_{..}^2}{ab} \\
 &= (1.3)^2 + (1.6)^2 + \dots + (3.4)^2 - \frac{(39.2)^2}{20} = 25.69 \\
 SS_{\text{Trattamenti}} &= \sum_{i=1}^4 \frac{y_{i.}^2}{b} - \frac{y_{..}^2}{ab} \\
 &= \frac{(5.7)^2 + (8.8)^2 + (6.9)^2 + (17.8)^2}{5} - \frac{(39.2)^2}{20} = 18.04
 \end{aligned}$$

**Tabella 5.13** Dati relativi alla resistenza del tessuto, piano a blocchi completi casualizzati.

Tipo di sostanza chimica	Esemplare di tessuto					Totali trattamento	Medie trattamento
	1	2	3	4	5		
1	1.3	1.6	0.5	1.2	1.1	5.7	1.14
2	2.2	2.4	0.4	2.0	1.8	8.8	1.76
3	1.8	1.7	0.6	1.5	1.3	6.9	1.38
4	3.9	4.4	2.0	4.1	3.4	17.8	3.56
Totali blocco $y_{ij}$	9.2	10.1	3.5	8.8	7.6	39.2(y..)	
Medie blocco $\bar{y}_{ij}$	2.30	2.53	0.88	2.20	1.90		1.96(\bar{y}..)

$$\begin{aligned}
 SS_{\text{Blocchi}} &= \sum_{j=1}^5 \frac{y_{ij}^2}{a} - \frac{y_{..}^2}{ab} \\
 &= \frac{(9.2)^2 + (10.1)^2 + (3.5)^2 + (8.8)^2 + (7.6)^2}{4} - \frac{(39.2)^2}{20} = 6.69 \\
 SS_E &= SS_T - SS_{\text{Blocchi}} - SS_{\text{Trattamenti}} \\
 &= 25.69 - 6.69 - 18.04 = 0.96
 \end{aligned}$$

L'ANOVA è riepilogata in Tabella 5.14. Siccome  $f_0 = 75.13 > f_{0.01,3,12} = 5.95$ , il  $P$ -value è minore di 0.01, perciò concludiamo che vi è una differenza significativa tra le sostanze chimiche per quanto concerne il loro effetto sulla resistenza media del tessuto. Il valore effettivo del  $P$ -value è  $4.79 \times 10^{-8}$ .

**Tabella 5.14** Analisi della varianza per l'esperimento a blocchi completi casualizzati

Causa della variazione	Somma dei quadrati	Gradi di libertà	Media quadratica	$f_0$	$P$ -value
Sostanze chimiche (trattamenti)	18.04	3	6.01	75.13	4.79 E-8
Esemplari tessuto (blocchi)	6.69	4	1.67		
Errore	0.96	12	0.08		
Totale	25.69	19			

### Quando sono necessari i blocchi?

Si consideri un esperimento condotto con un piano a blocchi casualizzati, in una situazione però, in cui la procedura dei blocchi non era in effetti necessaria. Vi sono  $ab$  osservazioni e  $(a - 1)(b - 1)$  gradi di libertà per l'errore. Se l'esperimento è stato eseguito come piano a singolo fattore completamente casualizzato con  $b$  repliche, avremmo avuto  $a(b - 1)$  gradi di libertà per l'errore, quindi la procedura dei blocchi è costata  $a(b - 1) - (a - 1)(b - 1) = b - 1$  gradi di libertà per l'errore. Concludendo, siccome la perdita di gradi di libertà per l'errore è generalmente piccola, se esiste una ragionevole possibilità che gli effetti del blocco possano essere importanti, lo sperimentatore dovrebbe usare il piano a blocchi casualizzati.

### Soluzione tramite software statistico

La Tabella 5.15 presenta l'output di Minitab per il piano a blocchi completi casualizzati dell'Esempio 5.15. Per risolvere questo problema abbiamo usato il menu dell'analisi della varianza per i piani bilanciati. I risultati appaiono in stretto accordo con quelli eseguiti a mano (Tabella 5.14). Si noti che Minitab calcola una statistica  $F$  per i blocchi (gli esemplari di tessuto). La validità di questo rapporto come statistica test per l'ipotesi nulla di nessun effetto di blocco è dubbia, poiché i blocchi rappresentano una **restrizione sulla casualizzazione**; cioè, abbiamo casualizzato solo all'interno dei blocchi. Se i blocchi non sono scelti a caso, o se non sono eseguiti in ordine casuale, il rapporto  $F$  per i blocchi può non fornire un'informazione affidabile riguardo gli effetti del blocco. Per un'analisi più approfondita si veda Montgomery (2009, Capitolo 4).

**Tabella 5.15** Analisi della varianza condotta con Minitab per il piano a blocchi completi casualizzati dall'Esempio 5.15

Analysis of Variance (Balanced Designs)						
Factor	Type	Levels	Values			
Chemical	fixed	4	1	2	3	4
Fabric S	fixed	5	1	2	3	4
Analysis of Variance for strength						
Source	DF	SS	MS	F	P	
Chemical	3	18.0440	6.0147	75.89	0.000	
Fabric S	4	6.6930	1.6733	21.11	0.000	
Error	12	0.9510	0.0792			
Total	19	25.6880				
F-test with denominator: Error						
Denominator MS = 0.079250 with 12 degrees of freedom						
Numerator	DF	MS	F	P		
Chemical	3	6.015	75.89	0.000		
Fabric S	4	1.673	21.11	0.000		

Quali medie differiscono?

Quando l'ANOVA indica che esiste una differenza tra le medie del trattamento, possiamo trovarci nella necessità di eseguire alcuni test supplementari per isolare le specifiche differenze. Per questo scopo può essere usato il metodo grafico descritto precedentemente. Le medie delle quattro sostanze chimiche sono

$$\bar{y}_1 = 1.14 \quad \bar{y}_2 = 1.76 \quad \bar{y}_3 = 1.38 \quad \bar{y}_4 = 3.56$$

Ciascuna media di trattamento usa  $b = 5$  osservazioni (una da ciascun blocco). Quindi, la deviazione standard di una media di trattamento è  $\sigma/\sqrt{b}$ . La stima di  $\sigma$  è  $\sqrt{MS_E}$ . Perciò, la deviazione standard usata per la distribuzione normale è

$$\sqrt{MS_E/b} = \sqrt{0.0792/5} = 0.126$$

**Esame delle differenze  
tra le medie.**

In Figura 5.13 è mostrata la rappresentazione di una distribuzione normale ampia  $6\sqrt{MS_E/b} = 0.755$  unità. Se immaginiamo di traslare questa densità lungo l'asse orizzontale, notiamo che non esiste una posizione per la distribuzione che suggerisca che tutte le quattro medie sono valori tipici selezionati casualmente da quella distribuzione. Questo non dovrebbe sorprendere, dato che l'analisi della varianza ha indicato che le medie differiscono. Le coppie di medie sottolineate non sono differenti. La sostanza chimica di tipo 4 dà luogo a una resistenza che differisce in modo significativo da quella delle altre tre sostanze. Le sostanze di tipo 2 e 3 e quelle di tipo 1 e 3 non differiscono tra loro. Può sussistere una piccola differenza nella resistenza tra la sostanza di tipo 1 e quella di tipo 2.

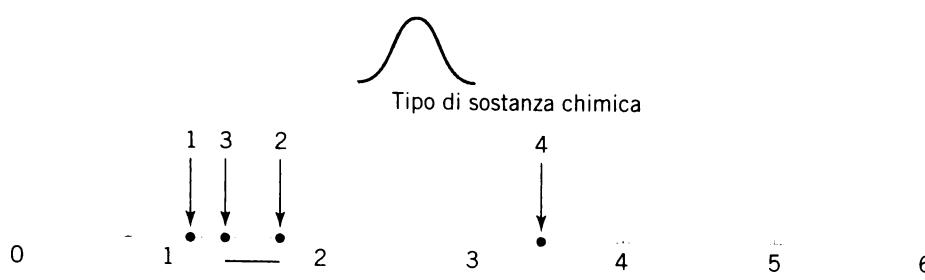


Figura 5.13 Medie della resistenza relativa all'esperimento del tessuto in relazione alla distribuzione normale con deviazione standard  $\sqrt{MS_E/b} = \sqrt{0.0792/5} = 0.126$ .

### Analisi dei residui e verifica del modello

In ogni esperimento pianificato, è sempre importante esaminare i residui e cercare se vi è una violazione delle assunzioni di base che potrebbe invalidare i risultati. Al solito, i residui per il piano a blocchi completi casualizzati sono solo la differenza tra i valori osservati e quelli stimati ricavati dal modello statistico; per esempio:

$$e_{ij} = y_{ij} - \hat{y}_{ij}$$

e i valori stimati sono

$$\hat{y}_{ij} = \bar{y}_{i\cdot} + \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot} \quad (5.44)$$

Il valore stimato rappresenta la stima della risposta media quando l' $i$ -esimo trattamento è eseguito nel  $j$ -esimo blocco. I residui dall'esperimento riguardante le sostanze chimiche sono mostrati in Tabella 5.16.

Tabella 5.16 Residui dal piano a blocchi completi casualizzati.

Sostanza chimica	Esemplare di tessuto				
	1	2	3	4	5
1	-0.18	-0.10	0.44	-0.18	0.02
2	0.10	0.08	-0.28	0.00	0.10
3	0.08	-0.24	0.30	-0.12	-0.02
4	0.00	0.28	-0.48	0.30	-0.10

Grafici dei residui per l'esperimento sulla resistenza del tessuto.

Le Figure 5.14, 5.15, 5.16 e 5.17 presentano i grafici dei residui importanti per l'esperimento. Questi grafici dei residui vengono usualmente costruiti mediante pacchetti software. Vi è qualche indicazione che l'esemplare di tessuto 3 abbia una variabilità maggiore della resistenza quando è trattato con le quattro sostanze chimiche, rispetto agli altri esemplari. La sostanza chimica di tipo 4, che fornisce la resistenza maggiore, ha una certa variabilità in più a livello di resistenza. Possono essere necessari ulteriori esperimenti per confermare questi risultati, se sono potenzialmente importanti.

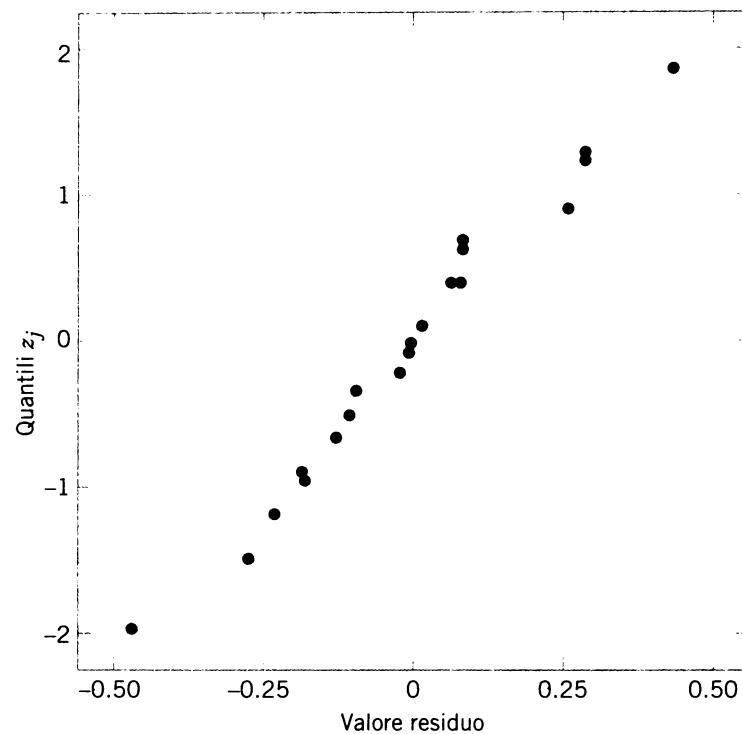


Figura 5.14 Grafico dei quantili per i residui del piano a blocchi completi casualizzati.

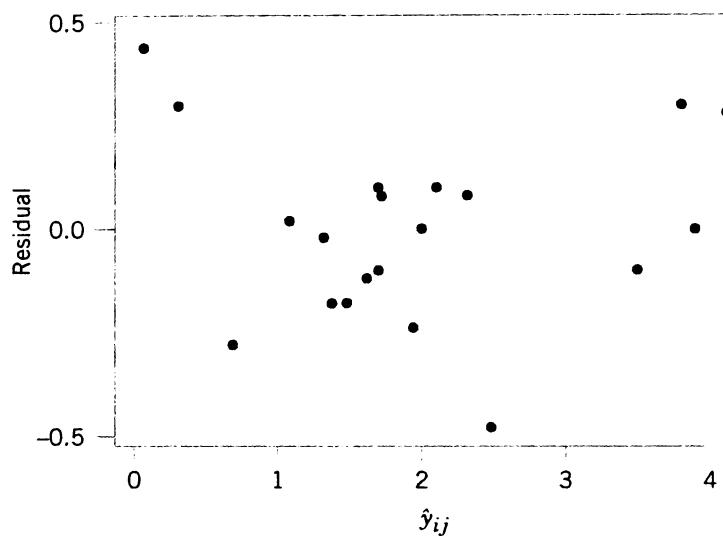


Figura 5.15 Residui in funzione di  $\hat{y}_{ij}$ .

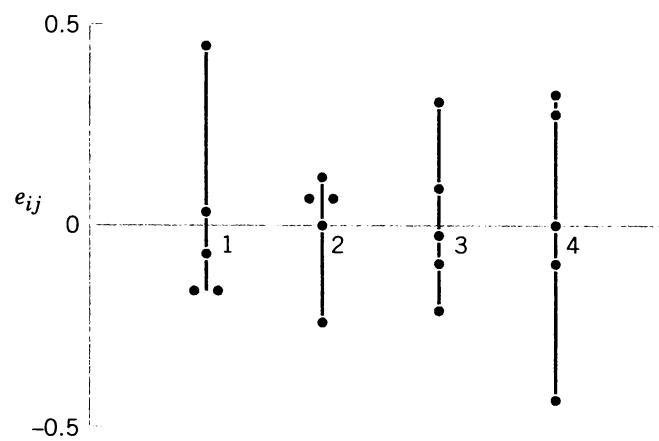


Figura 5.16 Residui in funzione del tipo di sostanza chimica.

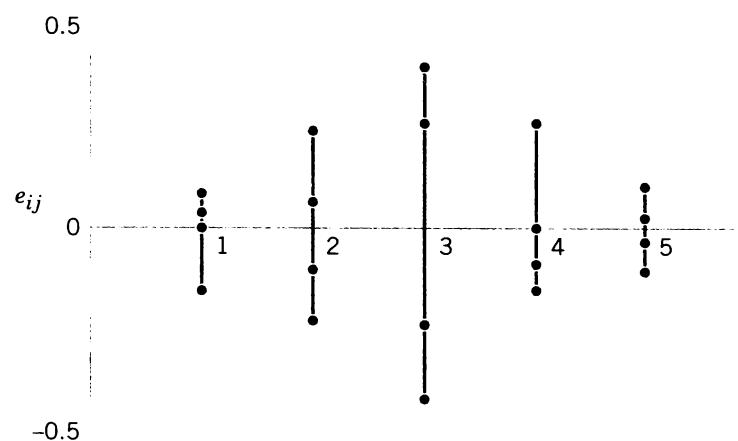


Figura 5.17 Residui in funzione dei blocchi.

## TERMINI E CONCETTI RILEVANTI

---

Analisi della varianza (ANOVA)

Blocchi

Collegamento fra verifiche di ipotesi e intervalli  
di confidenza

Curve operative caratteristiche

Determinazione della dimensione campionaria  
per gli intervalli di confidenza

Determinazione della dimensione campionaria  
per le verifiche di ipotesi

Distribuzione chi-quadro

Distribuzione F

Distribuzione t

Errore del I tipo

Errore del II tipo

Intervalli di confidenza

Ipotesi alternativa

Ipotesi alternative unilaterali e bilaterali

Ipotesi nulla

Ipotesi statistiche

Limiti di confidenza unilaterali

Livello di confidenza

Piano a blocchi completi casualizzati

Piano completamente casualizzato

Potenza di un test

P-value

Regione critica per una statistica test

Significatività statistica e significatività pratica

Statistica test

Test t a due campioni

Test t accoppiato

Test t pooled

# Esercizi proposti

## ESERCIZI PER IL PARAGRAFO 5.2

5.1. Un software statistico ha prodotto il seguente output per un problema di verifica di ipotesi:

Differenza tra le medie campionarie: 2.35.

Errore standard della differenza tra le medie: ?

Statistica test:  $z_0 = 2.01$

P-value: 0.0222

- (a) Qual è il valore mancante relativo all'errore standard?
- (b) Si tratta di un test unilaterale o bilaterale?
- (c) Se  $\alpha = 0.05$ , quali conclusioni si possono trarre?
- (d) Trovare un intervallo di confidenza dilivello 90% bilaterale per la differenza tra le medie.

5.2. Un software statistico ha prodotto il seguente output per un problema di verifica di ipotesi:

Differenza tra le medie campionarie: 11.5

Errore standard della differenza tra le medie: ?

Statistica test:  $z_0 = -1.88$

P-value: 0.0601

- (a) Qual è il valore mancante relativo all'errore standard?
- (b) Si tratta di un test unilaterale o bilaterale?
- (c) Se  $\alpha = 0.05$ , quali conclusioni si possono trarre?
- (d) Trovare un intervallo di confidenza al 95% bilaterale per la differenza tra le medie.

5.3. Un produttore di componenti elettronici deve scegliere tra due tipi di plastica. La resistenza alla rottura di tali materiali è un parametro molto importante. Si sa che  $\sigma_1 = \sigma_2 = 1.0$  psi. Da due campioni casuali di dimensioni  $n_1 = 10$  e

$n_2 = 12$  si ottengono le medie  $\bar{x}_1 = 162.7$  e  $\bar{x}_2 = 155.4$ . La compagnia non impiegherà la plastica 1 a meno che la sua resistenza media alla rottura non superi quella della plastica 2 di almeno 10 psi. In base all'informazione sul campione, la compagnia utilizzerà la plastica 1? Usare l'approccio basato sul P-value per rispondere.

5.4. Due differenti formule di carburante ossigenato (formula 1 e formula 2) vengono sottoposte a test per studiarne i numeri di ottani. La varianza del numero di ottani relativa al carburante con formula 1 è  $\sigma_1^2 = 1.5$ , quella del carburante con formula 2 è  $\sigma_2^2 = 1.2$ . Vengono studiati due campioni casuali di dimensioni  $n_1 = 15$  e  $n_2 = 20$ ; i numeri medi di ottani osservati sono rispettivamente  $\bar{x}_1 = 88.85$  e  $\bar{x}_2 = 92.54$ . Assumere la normalità.

- (a) Costruire un intervallo di confidenza bilaterale al 95% per la differenza tra i numeri medi di ottani.
- (b) Il produttore desidera stabilire se la formula 2 produce un più alto numero di ottani rispetto alla formula 1. Formulare e verificare un'ipotesi appropriata usando l'approccio basato sul P-value.
- (c) Qual è il P-value del test condotto al punto (b)?

5.5. Si consideri il test condotto nell'Esercizio 5.4. Quale dimensione campionaria è necessaria per ciascuna popolazione se si desidera un livello di confidenza del 95% sul fatto che l'errore nella stima della differenza tra i numeri medi di ottani sia minore di 1?

## ESERCIZI PER IL PARAGRAFO 5.3

5.6. Si consideri il seguente output di Minitab:

<b>Two-Sample T-Test and CI: X1, X2</b>				
Two-Sample T for X1 vs X2				
	N	Mean	StDev	SE Mean
X1	20	50.19	1.71	0.38
X2	20	52.52	2.48	0.55

Difference = mu (X1) – mu (X2)  
Estimate for difference: -2.33341  
95% CI for difference: (-3.69547, -0.97135)  
T-Test of difference = 0 (vs not =):  
T-Value = -3.47, P-Value = 0.001, DF = 38  
Both use Pooled StDev = 2.1277

- (a) Si può rifiutare l'ipotesi nulla al livello 0.05?
- (b) Si tratta di un test unilaterale o bilaterale?
- (c) Se l'ipotesi alternativa a  $H_0: \mu_1 - \mu_2 = 2$  fosse  $H_1: \mu_1 - \mu_2 \neq 2$ , si dovrebbe rifiutare l'ipotesi nulla al livello 0.05?
- (d) Se l'ipotesi alternativa a  $H_0: \mu_1 - \mu_2 = 2$  fosse  $H_1: \mu_1 - \mu_2 < 2$ , si dovrebbe rifiutare l'ipotesi nulla al livello 0.05? Si può rispondere a questo quesito senza fare ulteriori calcoli? Perché?
- (e) Usare l'output e la tavola per il test  $t$  per trovare un limite superiore di confidenza al 95% per la differenza tra le medie.
- (f) Qual è il  $P$ -value se l'ipotesi alternativa ad  $H_0: \mu_1 - \mu_2 = 2$  è  $H_1: \mu_1 - \mu_2 \neq 2$ ?

5.7. Si consideri il seguente output di Minitab:

<b>Two-Sample T-Test and CI: X1, X2</b>				
Two-Sample T for X1 vs X2				
	N	Mean	StDev	SE Mean
X1	15	75.47	1.63	?
X2	25	76.06	1.99	0.40

Difference = mu (X1) – mu (X2)  
Estimate for difference: -0.590171  
95% CI for difference: ?  
T-Test of difference = 0 (vs<): T-Value = -0.97,  
P-Value = 0.170, DF = ?  
Both use Pooled StDev = ?

- (a) Completare l'output con i valori mancanti. Si può rifiutare l'ipotesi nulla al livello 0.05? Perché?
- (b) Si tratta di un test unilaterale o bilaterale?

(c) Usare l'output e la tavola per il test  $t$  per trovare un limite di confidenza unilaterale superiore al 99% per la differenza tra le medie.

(d) Qual è il  $P$ -value se l'ipotesi alternativa a  $H_0: \mu_1 - \mu_2 = 1$  è  $H_1: \mu_1 - \mu_2 < 1$ ?

5.8. Si vuole studiare il diametro di aste di acciaio prodotte su due differenti macchine per estrusione. Vengono selezionati due campioni casuali di dimensioni  $n_1 = 15$  e  $n_2 = 17$ ; la media e la varianza campionarie sono rispettivamente  $\bar{x}_1 = 8.73$ ,  $s_1^2 = 0.35$ ,  $s_2^2 = 0.40$ . Si assuma che  $\sigma_1^2 = \sigma_2^2$  e che i dati siano ricavati da una distribuzione normale.

- (a) C'è qualche evidenza a sostegno dell'asserzione che le due macchine producono aste con diametri medi differenti? Usare il  $P$ -value per arrivare a questa conclusione.
- (b) Costruire un intervallo di confidenza al 95% per la differenza tra i diametri medi delle aste. Interpretare questo intervallo.

5.9. Nella produzione dei semiconduttori, per rimuovere il silicio dal retro dei wafer prima della metallizzazione viene spesso usato l'attacco chimico a umido. In tale processo, la velocità di attacco è una caratteristica essenziale, e si sa che segue una distribuzione normale. Sono state poste a confronto due differenti soluzioni di attacco usando due campioni casuali composti da 10 wafer per ciascuna soluzione. Le velocità di attacco osservate, espresse in ml/min, sono elencate nella seguente tabella:

Soluzione 1	Soluzione 2
9.9	10.6
9.4	10.3
9.3	10.0
9.6	10.3
10.2	10.1

- (a) I dati giustificano l'asserzione secondo cui la velocità media di attacco è la medesima per entrambe le soluzioni? Nel rispondere a questo quesito, usare  $\alpha = 0.05$  e assumere che le varianze delle popolazioni siano uguali.
- (b) Calcolare il  $P$ -value per il test al punto (a).
- (c) Trovare un intervallo di confidenza al 95% per la differenza tra le velocità medie di attacco.
- (d) Costruire grafici dei quantili per i due campioni. Tali grafici si accordano con le assunzioni di normalità e di uguali varianze? Scrivere un'interpretazione pratica di questi grafici.

 5.10. Si ritiene che lo spessore di un film di plastica (espresso in ml) depositato su di un materiale di substrato sia influenzato dalla temperatura alla quale si applica il rivestimento. Si esegue dunque un esperimento completamente casualizzato. Vengono rivestiti con tale film undici substrati a 125 °F, ottenendo uno spessore medio del rivestimento  $\bar{x}_1 = 101.28$  e una deviazione standard campionaria  $s_1 = 5.08$ . Vengono poi rivestiti altri 13 substrati a 150 °F, ottenendo  $\bar{x}_2 = 101.70$  e  $s_2 = 20.15$ . In partenza, si sospettava che l'aumento della temperatura del processo riducesse lo spessore di rivestimento medio.

I dati confermano questa ipotesi? Usare l'approccio del *P*-value e assumere che le deviazioni standard delle due popolazioni non siano uguali.

5.11. Si riconsideri l'esperimento dell'esercizio precedente. Rispondere alla domanda posta riguardo l'effetto della temperatura sullo spessore del rivestimento usando un intervallo di confidenza. Giustificare la risposta.

5.12. Una compagnia lirica regionale ha provato due approcci diversi per richiedere finanziamenti a 16 potenziali sostenitori. Questi ultimi sono stati selezionati a caso in una popolazione di possibili sostenitori e divisi, sempre a caso, in due gruppi di otto elementi ciascuno. Su ciascun gruppo è stato poi tentato uno degli approcci. Le somme di denaro raccolte, in dollari, sono le seguenti:

Approccio 1	\$1000	\$1500	\$1200	\$1800	\$1600	\$1100	\$1000	\$1250
Approccio 2	\$1500	\$1000	\$1200	\$1500	\$1200	\$1250	\$1100	\$1000

- (a) Vi sono indizi che indicano una differenza fra i due approcci rispetto alla media delle somme donate?
- (b) Costruire un intervallo di confidenza al 95% bilaterale per la differenza tra le medie.
- (c) Occorre dire qualcosa sull'assunto di normalità in questo problema?

## ESERCIZI PER IL PARAGRAFO 5.4

5.13. Si consideri il seguente output di Minitab:

Paired T-Test and CI: X1, X2				
Paired for X1 – X2				
	N	Mean	StDev	SE Mean
X1	12	74.2430	1.8603	0.5370
X2	12	73.4797	1.9040	0.5496
Difference	12	?	2.905560	0.838763

95% CI for mean difference: (-1.082805, 2.609404))  
T-Test of mean difference = 0 (vs not = 0):  
T-Value = ? P-Value = ?

- (a) Completare l'output con i valori mancanti, compreso un limite per il *P*-value. Si può rifiutare l'ipotesi nulla al livello 0.05? Perché?
- (b) Si tratta di un test unilaterale o bilaterale?
- (c) Usare l'output e la tavola per il test *t* per trovare un intervallo bilaterale di confidenza al 99% per la differenza tra le medie.
- (d) Come si potrebbe dire che non vi sono indizi sufficienti per rifiutare l'ipotesi nulla semplicemente osservando l'output, senza fare ulteriori calcoli?

5.14. Si consideri il seguente output di Minitab:

Paired T-Test and CI: X1, X2				
Paired for X1 – X2				
	N	Mean	StDev	SE Mean
X1	10	100.642	?	0.488
X2	10	105.574	2.867	0.907
Difference	10	-4.93262	3.66736	?

95% CI for mean difference: (-7.55610, -2.30915)  
T-Test of mean difference = 0 (vs not = 0):  
T-Value = ? P-Value = 0.002

- (a) Completare l'output con i valori mancanti. Si può rifiutare l'ipotesi nulla al livello 0.05? Perché?
- (b) Si tratta di un test unilaterale o bilaterale?
- (c) Usare l'output e la tavola per il test *t* per trovare un intervallo bilaterale di confidenza al 99% per la differenza tra le medie.
- (d) Qual è il *P*-value per la statistica test se l'obiettivo è di dimostrare che la media della popolazione 1 è minore della media della popolazione 2?
- (e) Qual è il *P*-value per la statistica test se l'obiettivo è di dimostrare che la differenza tra le medie è uguale a 4?

5.15. Un informatico sta studiando l'utilità di due differenti linguaggi di progetto rispetto al miglioramento dei compiti di programmazione. A dodici programmati esperti, aventi familiarità con tali linguaggi, viene richiesto di codificare una funzione standard in entrambi i linguaggi; viene quindi registrato il tempo (in minuti) impiegato in ciascuna prova. I dati sono elencati nella seguente tabella:

Programmatore	Tempo	
	Linguaggio 1	Linguaggio 2
1	17	18
2	16	14
3	21	19
4	14	11
5	18	23
6	24	21
7	16	10
8	14	13
9	21	19
10	23	24
11	13	15
12	18	20

- (a) Trovare un intervallo di confidenza al 95% sulla differenza tra i tempi medi di codifica. Esiste qualche indicazione che porti a preferire un determinato linguaggio di progettazione?
- (b) È ragionevole assumere che la differenza tra i tempi di codifica sia distribuita normalmente? Addurre prove a sostegno della risposta.

5.16. Un articolo apparso sul *Journal of Aircraft* (Vol. 23, 1986, pp. 859-864) descrive una nuova formulazione del metodo di analisi della lastra equivalente, in grado di modellizzare strutture di aeromobili come l'ala con parti esterne ad angolo di un aereo, e che produce risultati simili al metodo di

analisi a elementi finiti intensivo. Le frequenze naturali di vibrazione per tale struttura sono calcolate usando entrambi i metodi di analisi; nella seguente tabella sono elencate i risultati per le prime sette frequenze naturali.

Fre- quenza cicli/s	Elementi finiti equivalente	Lastra	Fre- quenza cicli/s	Elementi finiti equivalente	Lastra
	cicli/s	cicli/s		cicli/s	cicli/s
1	14.58	14.76	5	174.73	181.22
2	48.52	49.10	6	212.72	220.14
3	97.22	99.99	7	277.38	294.80
4	113.99	117.53			

- (a) I dati suggeriscono che i due metodi forniscono il medesimo valore medio per la frequenza di vibrazione naturale? Usare l'approccio basato sul *P*-value.
- (b) Trovare un intervallo di confidenza al 95% per la differenza media tra i due metodi e usarlo per rispondere alla domanda del punto (a).

5.17. Per determinare il livello di impurità presente in leghe di acciaio si possono impiegare due differenti test di analisi. Vengono sottoposti a test otto provini, usando entrambe le procedure; i risultati sono mostrati nella seguente tabella. Vi è sufficiente evidenza per concludere che i due test forniscono il medesimo livello medio di impurità, se si usa  $\alpha = 0.01$ ?

Provino	Test 1	Test 2	Provino	Test 1	Test 2
1	1.2	1.4	5	1.7	2.0
2	1.3	1.7	6	1.8	2.1
3	1.5	1.5	7	1.4	1.7
4	1.4	1.3	8	1.3	1.6

5.18. Si considerino i livelli di impurità dell'esercizio precedente. Costruire un intervallo di confidenza al 99% per la differenza media tra le due procedure di analisi. Usare tale intervallo per rispondere alla domanda posta nell'esercizio precedente.

## ESERCIZI PER IL PARAGRAFO 5.5

5.19. Si considerino i dati della velocità di attacco dell'Esercizio 5.9. Verificare l'ipotesi  $H_0: \sigma_1^2 = \sigma_2^2$  contro  $H_1: \sigma_1^2 \neq \sigma_2^2$  usando  $\alpha = 0.05$ , e trarre le opportune conclusioni.

5.20. Si considerino i dati del diametro delle aste dell'Esercizio 5.8. Costruire:

- (a) un intervallo di confidenza bilaterale al 90% per  $\sigma_1/\sigma_2$ ;

- (b) un intervallo di confidenza bilaterale al 95% per  $\sigma_1/\sigma_2$ . Confrontare l'ampiezza di questo intervallo con quella dell'intervallo al punto (a) e commentare il risultato;
- (c) un limite inferiore di confidenza al 90% per  $\sigma_1/\sigma_2$ .

5.21. L'Esercizio 5.10 riguardava la misura dello spessore del rivestimento di plastica a due differenti temperature di

applicazione. Verificare l'ipotesi che la varianza dello spessore sia minore per il processo a 125 °F di quella per il processo a 150 °F; usare  $\alpha = 0.10$ .

 5.22. È stato eseguito uno studio per determinare se tra uomo e donna esiste una differenza a livello di ripetibilità delle azioni necessarie per assemblare i componenti su un circuito stampato. Sono stati selezionati due campioni di 25 uomini e 21 donne, e ciascun soggetto ha assemblato le unità circuitali. Le due deviazioni standard campionarie del tempo di assemblaggio

sono state:  $s_{\text{uomo}} = 0.914 \text{ min}$  e  $s_{\text{donna}} = 1.093 \text{ min}$ . Esiste qualche indicazione che porti ad affermare che gli uomini sono meno predisposti alla ripetitività nell'operazione di assemblaggio rispetto alle donne? Usare  $\alpha = 0.01$  ed esplicitare le necessarie assunzioni riguardo alla sottostante distribuzione dei dati.

5.23. Si consideri nuovamente l'esperimento di ripetitività dell'esercizio precedente. Trovare un limite inferiore di livello 99% per il rapporto tra le due varianze. Fornire un'interpretazione dell'intervallo.

## ESERCIZI PER IL PARAGRAFO 5.6

5.24. Completare il seguente output di Minitab:

### Paired T-Test and CI: X1, X2

Sample	X	N	Sample p
1	285	500	0.570000
2	521	?	0.651250

Difference = p (1) – p (2)

Estimate for difference: ?

95% CI for difference: (-0.135782, -0.0267185)

Test for difference = 0 (vs not = 0): Z = ?

P-Value = 0.003

- (a) Si tratta di un test unilaterale o bilaterale?
- (b) Si può rifiutare l'ipotesi nulla al livello 0.05?
- (c) Se l'ipotesi alternativa a  $H_0: p_1 = p_2$  fosse  $H_1: p_1 < p_2$ , si dovrebbe rifiutare l'ipotesi nulla al livello 0.05? Come si potrebbe farlo senza eseguire ulteriori calcoli?
- (d) Se l'ipotesi alternativa a  $H_0: p_1 - p_2 = -0.02$  fosse  $H_1: p_1 - p_2 \neq -0.02$  si dovrebbe rifiutare l'ipotesi nulla al livello 0.05? Come si potrebbe farlo senza eseguire ulteriori calcoli?
- (e) Costruire un intervallo bilaterale di confidenza al 90%, approssimato, per  $p_1 - p_2$ .

5.25. Completare il seguente output di Minitab:

### Test and CI for Two Proportions

Sample	X	N	Sample p
1	190	250	0.760000
2	240	350	0.685714

Difference = p (1) – p (2)

Estimate for difference: ?

95% lower bound for difference: 0.0139543

Test for difference = 0 (vs > 0): Z = ? P-Value = ?

(a) Si tratta di un test unilaterale o bilaterale?

(b) Si può rifiutare l'ipotesi nulla al livello 0.05?

(c) Se l'ipotesi alternativa a  $H_0: p_1 = p_2$  fosse  $H_1: p_1 > p_2$ , si dovrebbe rifiutare l'ipotesi nulla al livello 0.05? Come si potrebbe farlo senza eseguire ulteriori calcoli?

(d) Costruire un intervallo bilaterale di confidenza tradizionale al 95%, approssimato, per  $p_1 - p_2$ .

 5.26. Per realizzare parti in plastica vengono utilizzati due differenti tipi di macchinari per lo stampaggio a iniezione. Si considera un pezzo difettoso se presenta un eccessivo restringimento o se è scolorito. Vengono selezionati due campioni casuali, ciascuno di dimensione 300; nel campione della macchina sono trovate 15 parti difettose, nel campione della macchina 2 sono trovate 8 parti difettose. È ragionevole concludere che le due macchine producono la medesima frazione di parti difettose, usando  $\alpha = 0.05$ ? Trovare il  $P$ -value per questo test.

5.27. Si consideri la situazione descritta nell'esercizio precedente. Si supponga che  $p_1 = 0.05$  e  $p_2 = 0.01$ .

- (a) Con le dimensioni campionarie di prima, qual è la potenza del test bilaterale per questa alternativa?
- (b) Determinare la dimensione campionaria necessaria per rilevare questa differenza con una probabilità pari ad almeno 0.9. Usare  $\alpha = 0.05$ .

5.28. Si consideri la situazione descritta nell'Esercizio 5.26. Si supponga che  $p_1 = 0.05$  e  $p_2 = 0.02$ .

- (a) Date queste dimensioni campionarie, qual è la potenza del test bilaterale per questa alternativa?
- (b) Determinare la dimensione campionaria necessaria per rilevare questa differenza con una probabilità pari ad almeno 0.9. Usare  $\alpha = 0.05$ .

5.29. Costruire un intervallo di confidenza tradizionale al 95% per la differenza tra le due frazioni difettose dell'Esercizio 5.26.

5.30. Costruire un intervallo di confidenza al 95% per la differenza tra le due frazioni difettose dell'Esercizio 5.26

usando un intervallo perfezionato. Confrontare i risultati ottenuti con i risultati ricavati con l'intervallo tradizionale.

## ESERCIZI PER IL PARAGRAFO 5.8

5.31. Quello che segue è l'output ANOVA di Minitab. Completarlo inserendo i valori mancanti. Si possono fornire limiti per il  $P$ -value.

### One-way ANOVA:

Source	DF	SS	MS	F	P
Factor	3	36.15	?	?	?
Error	?	?	?		
Total	19	196.04			

5.32. Quello che segue è l'output ANOVA di Minitab. Completarlo inserendo i valori mancanti. Si possono fornire limiti per il  $P$ -value.

### One-way ANOVA:

Source	DF	SS	MS	F	P
Factor	?	?	246.93	?	?
Error	25	186.53	?		
Total	29	1174.24			

 5.33. Nel volume *Design and Analysis of Experiments*, 7<sup>a</sup> edizione (John Wiley & Sons, 2009), D.C. Montgomery descrive un esperimento nel quale viene studiata la resistenza alla trazione di una fibra sintetica. Si sospetta che tale resistenza sia in relazione alla percentuale di cotone presente nella fibra. Si impiegano cinque livelli di percentuale di cotone, e vengono eseguite cinque repliche in ordine casuale; i dati ottenuti sono elencati nella seguente tabella.

Percentuale di cotone	Osservazioni				
	1	2	3	4	5
15	7	7	15	11	9
20	12	17	12	18	18
25	14	18	18	19	19
30	19	25	22	19	23
35	7	10	11	15	11

- (a) La percentuale di cotone influisce sulla resistenza alla trazione? Tracciare i box plot comparativi ed eseguire un'analisi della varianza. Usare l'approccio basato sul  $P$ -value.
- (b) Tracciare il diagramma della resistenza media alla trazione in funzione della percentuale di cotone e interpretare i risultati.
- (c) Quali medie specifiche sono differenti?
- (d) Eseguire l'analisi dei residui e il controllo del modello.

5.34. È stato eseguito uno studio per determinare se la temperatura di vulcanizzazione influenza la resistenza alla trazione di una gomma siliconata. Per misurare la resistenza alla trazione (in megapascal, MPa) di ciascun provino si è impiegato un dispositivo automatico controllato assialmente che applica una determinata forza idraulica. I risultati ottenuti sono elencati nella seguente tabella.

Temperatura, gradi Celsius		
25	40	55
2.09	2.22	2.03
2.14	2.09	2.22
2.18	2.10	2.10
2.05	2.02	2.07
2.18	2.05	2.03
2.11	2.01	2.15

- (a) Verificare l'ipotesi che la temperatura di vulcanizzazione influisce sulla resistenza alla trazione della gomma siliconata. Usare l'approccio basato sul  $P$ -value.
- (b) Costruire i box plot dei dati. Tali grafici supportano le conclusioni? Spiegare perché.
- (c) Eseguire l'analisi dei residui e il controllo del modello.

 5.35. Un articolo apparso su *American Industrial Hygiene Association Journal* (Vol. 37, 1976, pp. 418-422) descrive una verifica sul campo da usare per la determinazione della presenza di arsenico in campioni di urina. Il test è stato proposto per l'impiego sui lavoratori forestali, a causa dell'aumento di utilizzo di arsenici organici in tale settore industriale. L'esperimento ha confrontato i test eseguiti da un

apprendista, da un tecnico esperto e da un laboratorio esterno. Sono stati selezionati quattro soggetti per l'analisi, considerati come blocchi. La variabile risposta è il contenuto di arsenico (in ppm) nell'urina del soggetto. I dati sono elencati nella seguente tabella.

Verifica	Soggetto			
	1	2	3	4
Apprendista	0.05	0.05	0.04	0.15
Esperto	0.05	0.05	0.04	0.17
Laboratorio	0.04	0.04	0.03	0.10

- (a) Esiste qualche differenza nella procedura per la verifica dell'arsenico?  
 (b) Eseguire l'analisi dei residui per questo esperimento.

 5.36. È stato condotto un esperimento per studiare la corrente di dispersione in un dispositivo SOS MOSFET adottato nei micrometri. Scopo dell'esperimento è di studiare come varia tale corrente al variare della lunghezza di canale. Sono state selezionate quattro lunghezze di canale. Per ciascuna lunghezza sono state usate cinque differenti larghezze,

e la larghezza è da considerarsi un fattore di disturbo. I dati sono elencati nella seguente tabella.

Lunghezza di canale	Larghezza				
	1	2	3	4	5
1	0.7	0.8	0.8	0.9	1.0
2	0.8	0.8	0.9	0.9	1.0
3	0.9	1.0	1.7	2.0	4.0
4	1.0	1.5	2.0	3.0	20.0

- (a) Verificare l'ipotesi che la corrente media di dispersione non dipende dalla lunghezza del canale; usare  $\alpha = 0.05$ .  
 (b) Eseguire l'analisi dei residui per questo esperimento. Commentare i diagrammi dei residui.

5.37. Si consideri l'esperimento della corrente di dispersione descritto nell'esercizio precedente. La corrente di dispersione osservata per il canale di lunghezza 4 e larghezza 5 è stata registrata in modo errato. La corretta osservazione è 4.0. Analizzare i dati corretti di questo esperimento. Esiste motivo di affermare che la corrente media di dispersione aumenta con la lunghezza del canale?

## ESERCIZI DI FINE CAPITOLO

 5.38. Un responsabile degli acquisti ha comprato 25 resistori dal venditore 1 e 35 dal venditore 2. Di ciascun resistore viene misurata la resistenza. I risultati (in ohm) sono elencati nella seguente tabella.

Venditore 1						
96.8	100.0	100.3	98.5	98.3	98.2	99.6
99.4	99.9	101.1	103.7	97.7	99.7	101.1
97.7	98.6	101.9	101.0	99.4	99.8	99.1
99.6	101.2	98.2			98.6	

Venditore 2					
106.8	106.8	104.7	104.7	108.0	102.2
103.2	103.7	106.8	105.1	104.0	106.2
102.6	100.3	104.0	107.0	104.3	105.8
104.0	106.3	102.2	102.8	104.2	103.4
104.6	103.5	106.3	109.2	107.2	105.4
106.4	106.8	104.1	107.1		107.7

- (a) Quale assunzione sulla distribuzione è necessaria per verificare l'asserzione che la varianza della resistenza dei prodotti acquistati dal venditore 1 non è significativamente differente dalla varianza della resistenza dei prodotti acquistati dal venditore 2? Eseguire una procedura grafica per controllare questa assunzione.

- (b) Eseguire un'appropriata verifica di ipotesi per determinare se il responsabile degli acquisti può affermare che la varianza della resistenza dei resistori del venditore 1 è significativamente differente da quella dei resistori del venditore 2.

5.39. È stata studiata la resistenza alla rottura di due tipi di filo di lana. Sappiamo in base all'esperienza precedente che  $\sigma_1 = 5$  e  $\sigma_2 = 4$  psi. Per ciascun tipo di filo di lana si è costituito un campione casuale di 20 provini e si è ottenuto, rispettivamente,  $\bar{x}_1 = 88$  psi e  $\bar{x}_2 = 91$  psi.

- (a) Usando un intervallo di confidenza al 90% per la differenza tra le medie della resistenza alla rottura, commentare se vi è o meno evidenza per affermare che il filo di lana di tipo 2 ha una più alta resistenza media alla rottura.  
 (b) Usando un intervallo di confidenza al 98% per la differenza tra le medie della resistenza alla rottura, commentare se vi è o meno evidenza per affermare che il filo di lana di tipo 2 ha una più alta resistenza media alla rottura.  
 (c) Spiegare perché i risultati dei punti (a) e (b) sono differenti, o perché sono uguali. Quale scegliereste per prendere una decisione, e perché?

5.40. Si consideri il precedente esercizio. Si supponga che, prima di raccogliere i dati, si decida di volere che l'errore

commesso stimando  $\mu_1 - \mu_2$  con  $\bar{x}_1 - \bar{x}_2$  sia inferiore a 1.5 psi. Specificare la dimensione campionaria per le seguenti confidenze percentuali:

- (a) 90%
- (b) 98%
- (c) Commentare l'effetto dell'incremento della percentuale di confidenza sulla dimensione campionaria necessaria.
- (d) Ripetere i punti (a), (b) e (c) con un errore inferiore a 0.75 psi invece di 1.5 psi.
- (e) Commentare l'effetto della diminuzione dell'errore sulla dimensione campionaria necessaria.

**5.41.** L'esperimento di Salk del vaccino per la poliomielite (1954) si focalizzò sull'efficacia del vaccino nella lotta alla poliomielite paralitica. Poiché si ritenne che senza un gruppo di controllo composto da bambini non vi fosse una base solida per valutare l'efficacia del vaccino, quest'ultimo venne somministrato a un primo gruppo, mentre a un secondo fu somministrato un placebo (esteriormente identico al vaccino originale ma con nessun effetto terapeutico). Per ragioni etiche, e per evitare che la conoscenza da parte del somministratore del tipo di vaccino impiegato potesse influire sulla diagnosi, l'esperimento è stato condotto in modalità a doppio cieco: sia i bambini, sia i somministratori non sapevano cioè chi avesse ricevuto il vaccino e chi il placebo. I dati reali per questo esperimento sono i seguenti:

Gruppo con placebo:  $n = 201$  299: 110 casi di polio osservati  
Gruppo con vaccino:  $n = 200$  745: 33 casi di polio osservati

- (a) Usare una procedura di verifica di ipotesi per stabilire se la proporzione di bambini, nei due gruppi, che contrassero la poliomielite è statisticamente differente. Usare una probabilità di errore del I tipo uguale a 0.05.
- (b) Ripetere il punto (a) usando una probabilità di errore del I tipo uguale a 0.01.
- (c) Confrontare le conclusioni ottenute nei punti (a) e (b) e spiegare perché è la medesima, o perché è differente.

**5.42.** Un articolo pubblicato sulla rivista *Journal of the Environmental Engineering Division* (intitolato "Distribution of Toxic Substances in Rivers", Vol. 108, 1982, pp. 639-649) descrive uno studio sulla concentrazione di numerose sostanze organiche idrofobe nel Wolf River, in Tennessee. Sono state fatte misure di esaclorobenzene (HCB), in nanogrammi per litro, a differenti profondità a valle di un sito industriale abbandonato. I dati relativi a due profondità sono i seguenti:

Superficie: 3.74, 4.61, 4.00, 4.67, 4.87, 5.12, 4.52, 5.29, 5.74, 5.48  
Fondo: 5.44, 6.88, 5.37, 5.44, 5.03, 6.48, 3.89, 5.85, 6.85, 7.16

- (a) Quali ipotesi sono necessarie per verificare l'asserzione che la concentrazione di HCB è la stessa alle due profon-

dità? Verificare le ipotesi per le quali si dispone di informazioni.

- (b) Applicare un'opportuna procedura per stabilire se i dati supportano l'asserzione del punto (a).
- (c) Si supponga che la vera differenza tra le concentrazioni medie sia di 2.0 nanogrammi per litro. Per  $\alpha = 0.05$ , qual è la potenza di una statistica test per  $H_0: \mu_1 = \mu_2$  contro  $H_1: \mu_1 \neq \mu_2$ ?
- (d) Quale dimensione campionaria sarebbe necessaria per rilevare una differenza di 1.0 nanogrammi per litro al livello  $\alpha = 0.05$ , se la potenza deve essere non inferiore a 0.9?

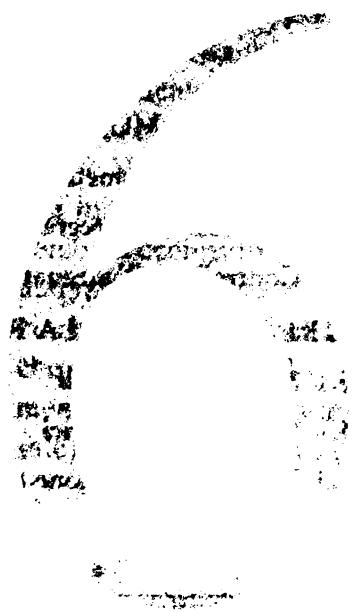
**5.43.** Si considerino nuovamente i dati dell'Esercizio 5.38 e si supponga che ogni resistore con resistenza inferiore a 100 ohm sia da considerarsi difettoso.

- (a) Stimare la frazione di resistori difettosi prodotti da ciascun venditore.
- (b) Costruire un intervallo di confidenza bilaterale tradizionale al 95% per la differenza tra le proporzioni di resistori difettosi prodotti dai due venditori.
- (c) Costruire un intervallo di confidenza bilaterale perfezionato al 95% per la differenza tra le proporzioni di resistori difettosi prodotti dai due venditori.
- (d) Confrontare gli intervalli trovati ai due punti precedenti.

**5.44.** Un articolo apparso sul *Wall Street Journal* il 27 aprile 2010, intitolato "Eating Chocolate is Linked to Depression", riportava uno studio finanziato dal National Heart, Lung and Blood Institute (una divisione dei National Institutes of Health) e dalla University of California di San Diego. Sono stati sottoposti a indagine statistica 931 adulti che non facevano uso di antidepressivi e non erano affetti da disturbi cardiovascolari o diabete. Il gruppo era composto per il 70% circa da uomini; l'età media si aggirava intorno ai 58 anni. Ai partecipanti sono state poste domande sul consumo di cioccolato, quindi hanno dovuto compilare un questionario volto a rilevare una eventuale depressione. Con un punteggio inferiore a 16 si era considerati non depressi, con un punteggio fra 16 e 22 si era considerati potenzialmente depressi e con un punteggio superiore a 22 si era classificati come depressi. Dall'indagine è emerso che le persone non depresse consumano 5.4 porzioni di cioccolato al mese, quelle potenzialmente depressi 8.4 e quelle depresse 11.8. Non sono state fatte differenze fra cioccolato al cacao e cioccolato bianco. Sono stati esaminati altri alimenti, ma non è emerso nessun tratto caratteristico che li legasse alla depressione.

Questo tipo di studio stabilisce un rapporto di causa ed effetto fra depressione e consumo di cioccolato? Come avrebbe dovuto essere condotto per stabilire tale rapporto?





# Costruzione di modelli empirici

## UN MODELLO PER LA PRODUZIONE DI IDROGENO

Le pile che usano l'idrogeno come combustibile sono state prodotte per il settore aerospaziale fin dagli anni Sessanta, e si profila un loro diffuso utilizzo anche qui sulla Terra. Tuttavia l'atmosfera del nostro pianeta contiene solo piccole tracce di idrogeno libero, perciò questo elemento deve essere estratto da altre fonti, come il metano. Un'importante area di ricerca ingegneristica è proprio quella che riguarda lo sviluppo di modelli empirici per migliorare il processo di estrazione dell'idrogeno dal metano. Un aspetto critico di tale processo concerne le piastre di catalizzazione dove hanno luogo le reazioni necessarie.

Alcuni ricercatori dell'Università di Salerno stanno usando modelli computazionali bidimensionali e tridimensionali per studiare i flussi di reagenti e di calore all'interno e intorno alle suddette piastre. Hanno così scoperto che l'uso del modello tridimensionale fornisce molte più informazioni di quello bidimensionale, ma rispetto a quest'ultimo presenta costi enormemente maggiori in termini di energia e di tempo. I due modelli vengono usati dai ricercatori per esaminare l'influenza dello spessore delle piastre sulle prestazioni complessive. Il loro lavoro rappresenta un valido contributo al miglioramento della produzione di idrogeno in vista della sua applicazione come combustibile.

Uno dei vantaggi dell'idrogeno, sotto questo punto di vista, è che il suo unico prodotto di emissione è l'acqua. Tuttavia, il processo di estrazione dell'idrogeno da un combustibile fossile lascia in effetti un'impronta ambientale sotto forma di monossido di carbonio, che finisce per ritrovarsi in atmosfera come anidride carbonica, un gas serra. La buona notizia è che questa via per l'estrazione di idrogeno produce un'energia doppia di quella ottenibile tramite la combustione diretta della stessa quantità di metano. Notizie ancora migliori provengono dai modelli per l'uso delle biomasse come fonte di idrogeno e per il riutilizzo nei fertilizzanti del carbone prodotto da questo metodo di estrazione dell'idrogeno; tali fertilizzanti servono quindi a concimare le piantagioni, in un ciclo virtuoso. I modelli mostrano che con l'adozione di questi metodi la quantità di anidride carbonica nell'atmosfera verrebbe ridotta significativamente. I modelli empirici sono strumenti formidabili per il progresso tecnologico nel settore dell'energia.

## CONTENUTI DEL CAPITOLO

- 6.1 INTRODUZIONE AI MODELLI EMPIRICI
- 6.2 REGRESSIONE LINEARE SEMPLICE
  - 6.2.1 Stima dei minimi quadrati
  - 6.2.2 Verifica delle ipotesi nella regressione lineare semplice
  - 6.2.3 Intervalli di confidenza nella regressione lineare semplice
  - 6.2.4 Predizione di nuove osservazioni
  - 6.2.5 Controllo dell'adeguatezza del modello
  - 6.2.6 Correlazione e regressione

- 6.3 REGRESSIONE MULTIPLA
  - 6.3.1 Stima dei parametri nella regressione multipla
  - 6.3.2 Inferenze nella regressione multipla
  - 6.3.3 Controllo dell'adeguatezza del modello
- 6.4 ALTRI ASPETTI DELLA REGRESSIONE
  - 6.4.1 Modelli polinomiali
  - 6.4.2 Regressori categorici
  - 6.4.3 Tecniche di selezione delle variabili

## OBIETTIVI DI APPRENDIMENTO

Dopo uno studio attento di questo capitolo si sarà in grado di:

1. usare la regressione lineare semplice o quella multipla per costruire modelli empirici di dati scientifici e ingegneristici
2. eseguire l'analisi dei residui per stabilire se il modello di regressione si adatta bene ai dati o per capire se qualche assunto隐式的 è stato violato
3. eseguire verifiche di ipotesi statistiche e costruire intervalli di confidenza per i parametri di un modello di regressione
4. usare la regressione allo scopo di stimare una media o di predire osservazioni future
5. usare gli intervalli di confidenza o di predizione per descrivere l'errore nella stima basata su una regressione
6. commentare i punti di forza e i punti deboli del modello adottato

## 6.1 INTRODUZIONE AI MODELLI EMPIRICI

Gli ingegneri, nella formulazione e nella risoluzione di un problema, fanno frequentemente uso di **modelli**. A volte, questi ultimi si basano sulle nostre conoscenze di fisica, di chimica o di ingegneria relative al fenomeno in esame: in questi casi i modelli vengono chiamati **modelli meccanicistici**. Tra gli esempi di modelli meccanicistici si possono citare la legge di Ohm, le leggi dei gas e le leggi di Kirchhoff. Tuttavia, si incontrano molte situazioni in cui vi sono relazioni tra due o più variabili di interesse, ma il modello meccanicistico che descrive tali relazioni è sconosciuto. In questi casi è necessario costruire un modello che ponga in relazione le variabili in base a dati osservati; questo tipo di modello viene chiamato **modello empirico**. Un modello empirico può essere manipolato e analizzato esattamente come un modello meccanicistico.

Come esempio, si considerino i dati riportati in Tabella 6.1:  $y$  rappresenta la concentrazione di sale (espressa in milligrammi litro) rilevata in corsi d'acqua superficiali in un bacino idrografico, mentre  $x$  rappresenta la percentuale di area del bacino coperta da strade pavimentate. I dati rispecchiano quelli apparsi in un articolo del *Journal of Environmental Engineering* (1989, Vol. 115, No. 3). In Figura 6.1 è mostrato un diagramma di dispersione dei dati elencati in tabella (con associati i diagrammi a punti delle singole variabili). Non esiste un evidente meccanismo fisico che leggi la concentrazione di sale all'area stradale carrabile, ma il diagramma di dispersione segnala l'esistenza di qualche relazione, magari lineare. Una relazione lineare non passerà esattamente per tutti i punti di Figura 6.1, tuttavia il grafico indica che tali punti sono dispersi in modo casuale intorno a una linea retta. Pertanto, è probabilmente ragionevole assumere che tra la media della variabile aleatoria  $Y$  (la concentrazione di sale) e l'area stradale  $x$  sussista la seguente relazione lineare

$$E(Y|x) = \mu_{Y|x} = \beta_0 + \beta_1 x$$

dove il coefficiente angolare, o pendenza, e l'ordinata all'origine, o intercetta, della retta sono parametri incogniti. La notazione  $E(Y|x)$  indica il valore atteso della **variabile risposta**  $Y$  per un particolare valore del **regressore**  $x$ . Benché la media di  $Y$  sia una funzione lineare di  $x$ , l'effettivo valore osservato  $y$  non cade esattamente su una linea retta. Il modo corretto di generalizzare questo risultato a un **modello probabilistico lineare** consiste nell'assumere che il valore atteso di  $Y$  sia una funzione lineare di  $x$ , ma che per un valore fissato di  $x$  il valore reale di  $Y$  sia determinato dalla funzione valor medio (ossia dal modello lineare), *più un termine di errore casuale*  $\epsilon$ .

### Modello di regressione lineare semplice

Nel **modello di regressione lineare semplice** la variabile dipendente, o **risposta**, è in relazione con una variabile indipendente, o **regressore**, tramite l'equazione

$$Y = \beta_0 + \beta_1 x + \epsilon \quad (6.1)$$

dove  $\epsilon$  è il termine di errore casuale. I parametri  $\beta_0$  e  $\beta_1$  sono detti **coefficienti di regressione**.

Per comprendere meglio questo modello, supponiamo di riuscire a fissare il valore di  $x$  e osservare il valore della variabile aleatoria  $Y$ . Se  $x$  è fissato, la componente casuale  $\epsilon$  nel membro di destra dell'equazione del modello (Equazione (6.1)) determina le proprietà di  $Y$ . Supponiamo che la media e la varianza di  $\epsilon$  siano, rispettivamente, 0 e  $\sigma^2$ ; allora

$$E(Y|x) = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x + E(\epsilon) = \beta_0 + \beta_1 x$$

Si noti che questa è la medesima relazione che abbiamo dedotto inizialmente in modo empirico dall'analisi del diagramma di dispersione di Figura 6.1. La varianza di  $Y$  data  $x$  è

$$V(Y|x) = V(\beta_0 + \beta_1 x + \epsilon) = V(\beta_0 + \beta_1 x) + V(\epsilon) = 0 + \sigma^2 = \sigma^2$$

**Tabella 6.1** Concentrazione di sale in corsi d'acqua superficiali e area stradale.

Osservazione	Concentrazione di sale (y)	Area stradale (x)
1	3.8	0.19
2	5.9	0.15
3	14.1	0.57
4	10.4	0.40
5	14.6	0.70
6	14.5	0.67
7	15.1	0.63
8	11.9	0.47
9	15.5	0.75
10	9.3	0.60
11	15.6	0.78
12	20.8	0.81
13	14.6	0.78
14	16.6	0.69
15	25.6	1.30
16	20.9	1.05
17	29.9	1.52
18	19.6	1.06
19	31.3	1.74
20	32.7	1.62

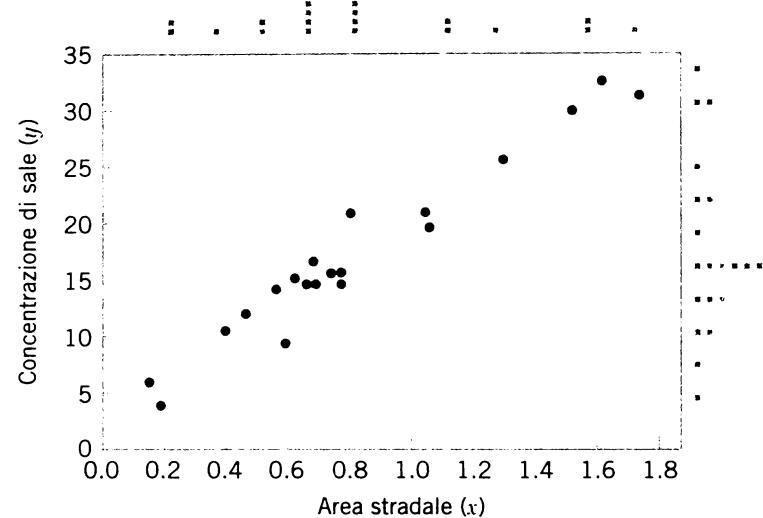


Figura 6.1 Diagramma di dispersione dei dati di Tabella 6.1.

Pertanto, il modello di regressione reale  $\mu_{Y|x} = \beta_0 + \beta_1 x$  è una retta di valori medi; cioè, l'ordinata del punto sulla retta di regressione per ogni valore di  $x$  è semplicemente il valore atteso di  $Y$  per quella  $x$ . Il coefficiente angolare della retta,  $\beta_1$ , può essere interpretato come l'incremento della media di  $Y$  per un incremento unitario di  $x$ . Inoltre, la variabilità di  $Y$  per un particolare valore di  $x$  è determinata dalla varianza dell'errore,  $\sigma^2$ . Ciò comporta l'esistenza di una distribuzione dei valori di  $Y$  per ciascun valore di  $x$ , la cui varianza è la stessa per ogni  $x$ .

Per esempio, si supponga che il modello di regressione reale che lega la concentrazione di sale all'area stradale sia  $\mu_{Y|x} = 3 + 15x$ , e che la varianza sia  $\sigma^2 = 2$  (Figura 6.2). Si noti

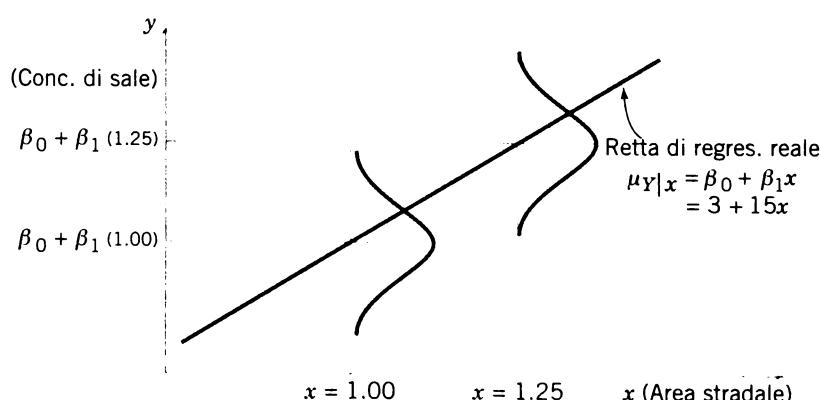


Figura 6.2  
La distribuzione di  $Y$   
per un determinato  
valore di  $x$  relativa  
ai dati di Tabella 6.1.

che abbiamo usato una distribuzione normale per descrivere la variazione casuale dell'errore  $\epsilon$ . Poiché  $Y$  è la somma di un termine costante  $\beta_0 + \beta_1 x$  (la media) e di una variabile aleatoria distribuita normalmente, risulta anch'essa una variabile aleatoria distribuita normalmente. La varianza  $\sigma^2$  determina la variabilità delle osservazioni  $Y$  sulla concentrazione di sale. Perciò, quando  $\sigma^2$  è piccola, i valori osservati di  $Y$  cadranno in prossimità della retta, mentre quando  $\sigma^2$  è grande, i valori osservati di  $Y$  potranno scostarsi notevolmente dalla retta. Essendo  $\sigma^2$  costante, la variabilità di  $Y$  per ogni valore di  $x$  è la medesima.

Il modello di regressione descrive la relazione tra la concentrazione di sale  $Y$  e l'area stradale  $x$ , quindi, per ogni valore dell'area stradale, la concentrazione di sale ha una distribuzione normale con media  $3 + 15x$  e varianza 2. Per esempio, se  $x = 1.25$ , allora  $Y$  ha valor medio  $= \mu_{Y|x} = 3 + 15(1.25) = 21.75$  e varianza pari a 2.

Esistono molte situazioni, in cui si deve costruire un modello empirico, nelle quali è presente più di un regressore. Anche in questi casi si può usare un modello di regressione per descrivere la relazione. Un modello di regressione che contiene più di una variabile regressore viene detto **modello di regressione multipla**.

A titolo di esempio, si supponga che la durata operativa di uno strumento da taglio dipenda dalla velocità e dall'angolo di taglio. Un modello di regressione multipla in grado di descrivere questa relazione è

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \quad (6.2)$$

dove  $Y$  rappresenta la durata dello strumento,  $x_1$  la velocità di taglio,  $x_2$  l'angolo di taglio ed  $\epsilon$  è il termine di errore casuale. Si tratta di un **modello di regressione lineare multipla** con due regressori. Il termine *lineare* deriva dal fatto che l'Equazione (6.2) è una funzione lineare dei parametri incogniti  $\beta_0$ ,  $\beta_1$  e  $\beta_2$ .

Il modello di regressione dell'Equazione (6.2) descrive un piano nello spazio tridimensionale di  $Y$ ,  $x_1$  e  $x_2$ . La Figura 6.3a mostra questo piano per il modello di regressione

$$E(Y) = 50 + 10x_1 + 7x_2$$

dove abbiamo assunto che il valore atteso del termine di errore sia zero, cioè  $E(\epsilon) = 0$ . Il parametro  $\beta_0$  è l'**intercetta** del piano. A volte  $\beta_1$  e  $\beta_2$  si dicono **coefficienti di regressione parziali** perché  $\beta_1$  misura l'incremento atteso di  $Y$  per incremento unitario di  $x_1$  con  $x_2$  costante,  $\beta_2$  misura l'incremento atteso di  $Y$  per incremento unitario di  $x_2$  con  $x_1$  costante. La Figura 6.3b mostra un **grafico delle curve di livello** del modello di regressione, cioè delle curve in cui  $E(Y)$  come funzione di  $x_1$  e  $x_2$  è costante. Si noti che in questo grafico le curve di livello sono rette.

### Modello di regressione lineare multipla

In un **modello di regressione lineare multipla** la **variabile dipendente o risposta** è in relazione con  $k$  **variabili indipendenti o regressori**. Il modello è

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \quad (6.3)$$

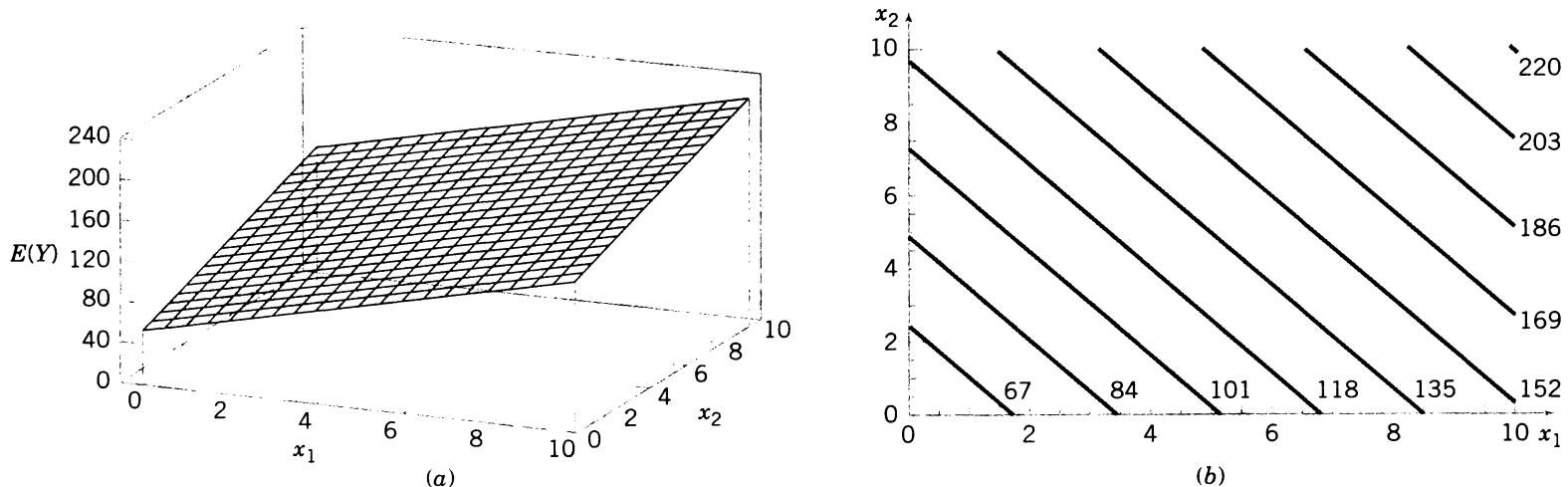


Figura 6.3 (a) Piano di regressione per il modello  $E(Y) = 50 + 10x_1 + 7x_2$ . (b) Curve di livello.

I parametri  $\beta_j$ , con  $j = 0, 1, \dots, k$ , sono detti coefficienti di regressione. Il modello descrive un **iperpiano** nello spazio delle variabili regressore  $\{x_j\}$  e risposta  $Y$ . Il parametro  $\beta_j$  rappresenta l'incremento atteso della risposta  $Y$  per incremento unitario di  $x_j$  quando tutti i rimanenti regressori  $x_i$  (con  $i \neq j$ ) sono mantenuti costanti.

I modelli di regressione lineare multipla vengono spesso impiegati come **modelli empirici**, nel senso che il modello meccanicistico che lega  $Y$  a  $x_1, x_2, \dots, x_k$  non è noto, ma su una certa regione delle variabili indipendenti il modello di regressione lineare risulta un'approssimazione adeguata.

Questi modelli empirici sono in relazione con la ben nota e importante approssimazione mediante serie di Taylor di una funzione non lineare, introdotta nel Capitolo 3; per esempio, l'approssimazione con serie di Taylor al primo ordine della funzione ignota  $f(x)$  intorno alla media  $\mu_x$

$$\begin{aligned} f(x) &\approx f(\mu_x) + \frac{df(x)}{dx} \Big|_{x=\mu_x} (x - \mu_x) + R \\ &\approx \beta_0 + \beta_1 (x - \mu_x) \end{aligned}$$

che, quando il resto viene trascurato, è un semplice modello lineare intorno alla media senza il termine di errore. Inoltre, modelli con struttura più complessa, approssimabili con serie di Taylor di ordine superiore della funzione  $f(x)$  o di una funzione  $f(x_1, x_2, \dots, x_k)$  di  $k$  variabili indipendenti, possono essere analizzati con tecniche di regressione lineare multipla. Per esempio, si consideri il modello polinomiale del terzo ordine in un regressore

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon \quad (6.4)$$

Se poniamo  $x_1 = x, x_2 = x^2, x_3 = x^3$ , l'Equazione (6.4) può essere scritta come

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \quad (6.5)$$

che rappresenta un modello di regressione lineare multipla con tre variabili regressore.

Anche modelli che includono effetti di **interazione** possono venire analizzati con metodi di regressione lineare multipla. Si può rappresentare un'interazione tra due variabili mediante un termine di prodotto incrociato, come nella seguente formula

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \quad (6.6)$$

Se poniamo  $x_3 = x_1 x_2$  e  $\beta_3 = \beta_{12}$ , l'Equazione (6.6) può essere scritta come

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

che costituisce un modello di regressione lineare.

Le Figure 6.4a e b mostrano il grafico tridimensionale del modello di regressione

$$Y = 50 + 10x_1 + 7x_2 + 5x_1 x_2$$

e il corrispondente grafico delle curve di livello in due dimensioni. Benché esso sia un modello di regressione lineare, la forma della superficie da esso generata non è lineare. In generale, **qualsiasi modello di regressione che è lineare nei parametri  $\beta$  è un modello di regressione lineare, indipendentemente dalla forma che assume la superficie da esso generata.**

La Figura 6.4 fornisce un'interessante interpretazione grafica di un'interazione. Di solito l'interazione implica che l'effetto prodotto facendo variare una variabile (per esempio  $x_1$ ) dipende dal livello dell'altra variabile ( $x_2$ ). Per esempio, la Figura 6.4 mostra che variando  $x_1$  da 2 a 8, in  $E(Y)$  si produce una variazione molto più piccola per  $x_2 = 2$  che non per  $x_2 = 10$ . Si incontrano frequentemente effetti di interazione nella progettazione del prodotto e del processo, nell'ottimizzazione del processo e in altre attività ingegneristiche; per la descrizione di tutte queste situazioni è possibile ricorrere, fra le varie tecniche, ai metodi di regressione.

Come esempio finale, si consideri il modello del secondo ordine con interazione

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 \quad (6.7)$$

Se poniamo  $x_3 = x_1^2$ ,  $x_4 = x_2^2$ ,  $x_5 = x_1 x_2$ ,  $\beta_3 = \beta_{11}$ ,  $\beta_4 = \beta_{22}$  e  $\beta_5 = \beta_{12}$ , l'Equazione (6.7) può essere scritta come modello di regressione lineare multipla, come segue

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$$

Le Figure 6.5a e b mostrano il grafico tridimensionale e le corrispondenti curve di livello per

$$E(Y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1 x_2$$

Tali grafici indicano che la variazione attesa di  $Y$  quando  $x_1$  varia, per esempio, di un'unità è una funzione di *entrambe* le variabili  $x_1$  e  $x_2$ . In questo modello, il termine quadratico e quello di interazione generano una funzione a grafico concavo. A seconda dei valori dei coefficienti di regressione, il modello del secondo ordine con interazione è in grado di generare un'ampia varietà di forme; è dunque un modello di regressione molto flessibile.

Nella maggior parte dei problemi del mondo reale, i valori dei parametri (i coefficienti di regressione  $\beta_i$ ) e la varianza dell'errore  $\sigma^2$  non sono noti; vanno quindi stimati a partire da dati campionari. L'**analisi di regressione** è un insieme di strumenti statistici impiegati per trovare le stime dei parametri del modello di regressione. Tipicamente, si usa quindi l'equazione (o modello) di regressione stimata allo scopo di predire osservazioni future di  $Y$  o di stimare la risposta media a un particolare livello di  $x$ . Per illustrare quanto detto con un esempio di modello di regressione lineare semplice, un ingegnere ambientale potrebbe

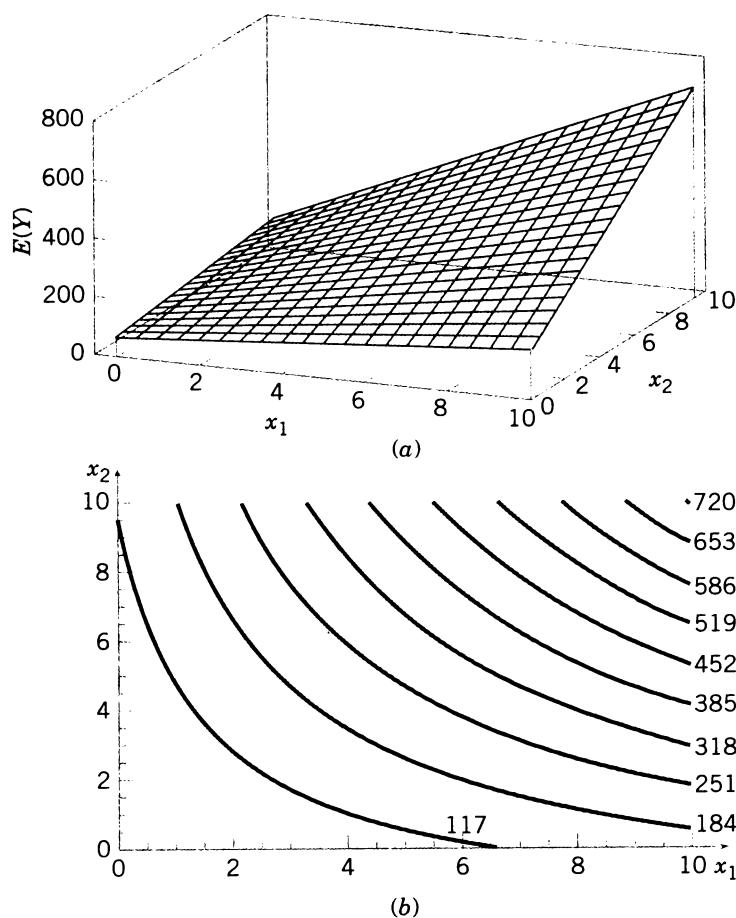


Figura 6.4 (a) Grafico tridimensionale del modello di regressione  $E(Y) = 50 + 10x_1 + 7x_2 + 5x_1x_2$ .  
 (b) Curve di livello.

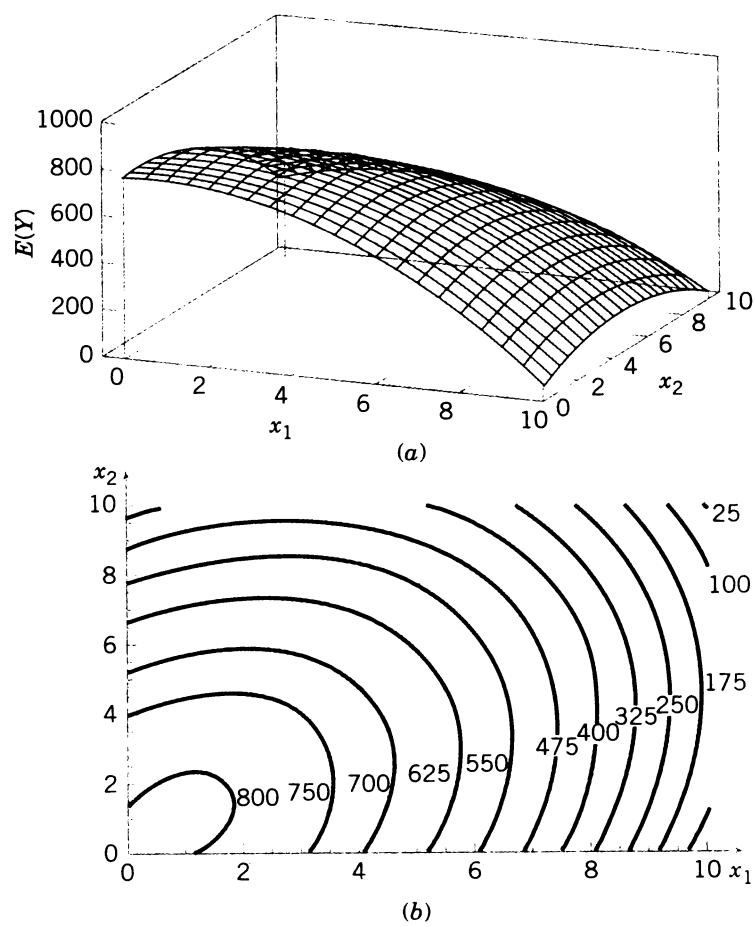


Figura 6.5 (a) Grafico tridimensionale del modello di regressione  $E(Y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1x_2$ . (b) Curve di livello.

essere interessato a stimare la concentrazione media di sale in corsi d'acqua superficiali quando la percentuale dell'area pavimentata del bacino idrografico è  $x = 1.25\%$ . In questo capitolo verranno discusse queste procedure e queste applicazioni per i modelli di regressione lineare.

## 6.2 REGRESSIONE LINEARE SEMPLICE

### 6.2.1 Stima dei minimi quadrati

La **regressione lineare semplice** considera un *singolo regressore* o *predittore*  $x$  e una variabile dipendente o variabile *risposta*  $Y$ . Supponiamo che l'effettiva relazione tra  $Y$  e  $x$  sia rappresentata da una retta, e che l'osservazione  $Y$  per ciascun valore di  $x$  sia una variabile aleatoria. Come abbiamo osservato in precedenza, il valore atteso di  $Y$  per ogni valore di  $x$  è

$$E(Y|x) = \beta_0 + \beta_1 x$$

dove l'intercetta  $\beta_0$  e il coefficiente angolare  $\beta_1$  sono coefficienti di regressione incogniti. Assumiamo che ogni osservazione  $Y$  possa essere descritta dal modello

$$\cdot \quad Y = \beta_0 + \beta_1 x + \epsilon \quad (6.8)$$

dove  $\epsilon$  è un errore casuale con media nulla e varianza  $\sigma^2$ . Si assume che gli errori casuali corrispondenti a differenti osservazioni siano variabili aleatorie non correlate.

Supponiamo di avere  $n$  coppie di osservazioni  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . La Figura 6.6 mostra un tipico grafico di dispersione di dati osservati e una possibile retta di regressione stimata. Le stime di  $\beta_0$  e  $\beta_1$  dovrebbero portare a una retta che è (in un certo senso) il “miglior adattamento” (*best fit*) ai dati. Lo scienziato tedesco Karl Gauss (1777-1855) propose di stimare i parametri  $\beta_0$  e  $\beta_1$  nell’Equazione (6.8) minimizzando la somma dei quadrati degli scarti verticali in Figura 6.6, con un approccio detto **metodo dei minimi quadrati**.

Usando l’Equazione (6.8) possiamo esprimere le  $n$  osservazioni nel campione come

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n \quad (6.9)$$

La somma dei quadrati degli scarti delle osservazioni dalla vera retta di regressione è

$$L = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (6.10)$$

Gli estimatori dei minimi quadrati di  $\beta_0$  e  $\beta_1$ , che indichiamo come  $\hat{\beta}_0$  e  $\hat{\beta}_1$ , devono soddisfare le condizioni

$$\frac{\partial L}{\partial \beta_0} \Big|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (6.11)$$

$$\frac{\partial L}{\partial \beta_1} \Big|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

che semplificate danno

$$\begin{aligned} n \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i &= \sum_{i=1}^n y_i \\ \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n y_i x_i \end{aligned} \quad (6.12)$$

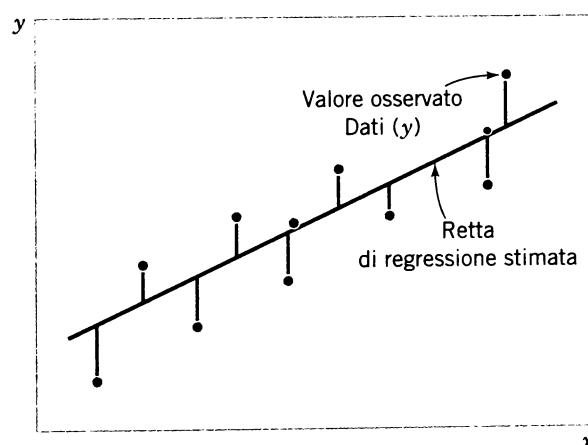


Figura 6.6 Scarti dei dati dal modello di regressione stimato.

Le Equazioni (6.12) sono dette **equazioni normali dei minimi quadrati**; la loro soluzione porta alle stime dei minimi quadrati  $\hat{\beta}_0$  e  $\hat{\beta}_1$ .

### Formule di calcolo per la regressione lineare semplice

Le **stime dei minimi quadrati** dell'intercetta e del coefficiente angolare del modello di regressione lineare semplice sono

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (6.13)$$

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n y_i x_i - \frac{\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \end{aligned} \quad (6.14)$$

dove  $\bar{y} = (1/n) \sum_{i=1}^n y_i$  e  $\bar{x} = (1/n) \sum_{i=1}^n x_i$ .

La **retta di regressione stimata** è quindi

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (6.15)$$

Si noti che ogni coppia di osservazioni soddisfa la relazione

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i, \quad i = 1, 2, \dots, n$$

dove  $e_i = y_i - \hat{y}_i$  è un termine chiamato **residuo**. Il residuo descrive l'errore nell'adattamento del modello alla  $i$ -esima osservazione  $y_i$ . Nel prosieguo useremo i residui per fornire informazioni riguardo l'**adeguatezza** del modello stimato.

**ESEMPIO 6.1**  
Concentrazione  
di sale e area  
stradale

Adattiamo un semplice modello di regressione lineare ai dati di Tabella 6.1 relativi alla concentrazione di sale e all'area stradale del bacino idrografico. Possiamo calcolare le seguenti quantità

$$n = 20 \quad \sum_{i=1}^{20} x_i = 16.480 \quad \sum_{i=1}^{20} y_i = 342.70 \quad \bar{x} = 0.824 \quad \bar{y} = 17.135$$

$$\sum_{i=1}^{20} y_i^2 = 7060.00 \quad \sum_{i=1}^{20} x_i^2 = 17.2502 \quad \sum_{i=1}^{20} x_i y_i = 346.793$$

$$S_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20} = 17.2502 - \frac{(16.486)^2}{20} = 3.67068$$

e

$$S_{xy} = \sum_{i=1}^{20} x_i y_i - \frac{\left(\sum_{i=1}^{20} x_i\right)\left(\sum_{i=1}^{20} y_i\right)}{20} = 346.793 - \frac{(16.480)(342.70)}{20} = 64.4082$$

Pertanto, le stime dei minimi quadrati del coefficiente angolare e dell'intercetta sono

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{64.4082}{3.67068} = 17.5467$$

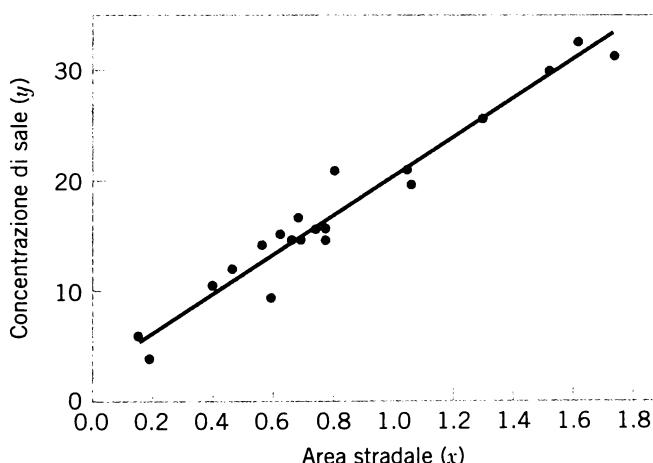
e

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 17.135 - (17.5467)0.824 = 2.6765$$

Il modello di regressione lineare semplice stimato, rappresentato graficamente in Figura 6.7 insieme ai dati campionari, è

$$\hat{y} = 2.6765 + 17.5467x$$

Usando il modello di regressione lineare dell'Esempio 6.1, avremmo predetto come concentrazione di sale presente nei corsi d'acqua superficiali il valore  $\hat{y} = 2.6765 + 17.5467(1.25) = 24.61$  mg/l, con una percentuale di strade pavimentate pari all'1.25%. Il valore predetto può essere interpretato o come stima della concentrazione media di sale quando l'area stra-



**Figura 6.7**  
Diagramma  
di dispersione  
della concentrazione di  
sale  $y$  in funzione  
dell'area stradale  $x$   
e modello  
di regressione stimato.

dale è  $x = 1.25\%$ , o come stima di una *nuova* osservazione quando  $x = 1.25\%$ . Tali stime, naturalmente, sono soggette a errore; non è verosimile cioè che la reale concentrazione media di sale o un'osservazione futura siano *esattamente* uguali a 24.61 mg/l quando l'area carrabile è dell'1.25%. Più avanti vedremo come usare gli intervalli di confidenza e gli intervalli di predizione per descrivere l'errore nella stima a partire da un modello di regressione.

Per svolgere i calcoli coinvolti nei modelli di regressione si usano comunemente i pacchetti software. In Tabella 6.2 è mostrato l'output di Minitab per il modello di regressione riferito alla concentrazione di sale e all'area carrabile.

Abbiamo evidenziato diverse voci nell'output di Minitab, comprese le stime di  $\beta_0$  e  $\beta_1$  (nella colonna a intestazione "Coef" nella parte superiore della Tabella 6.2). Si noti che Minitab elabora i **residui** del modello: sostituisce cioè via via ciascun valore di  $x_i$  ( $i = 1, 2, \dots, n$ ) del campione nel modello di regressione stimato, calcola i valori stimati  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ , quindi trova i residui per differenza  $e_i = y_i - \hat{y}_i$ ,  $i = 1, 2, \dots, n$ . Per esempio, la nona osservazione ha  $x_9 = 0.75$  e  $y_9 = 15.5$ ; il modello di regressione porta alla predizione = 15.837, per cui il residuo corrispondente è  $e_9 = 15.5 - 15.837 = -0.377$ . In tabella sono elencati i residui per tutte le 20 osservazioni.

I residui vengono utilizzati per stimare la varianza  $\sigma^2$  degli errori del modello. Ricordiamo che  $\sigma^2$  determina la variabilità delle osservazioni della risposta  $y$  per un dato valore del regressore  $x$ . Per ottenere la stima di  $\sigma^2$  si calcola la somma dei residui elevati al quadrato.

### Definizione

**La somma dei quadrati dei residui** o, talvolta, **somma dei quadrati degli errori** (*error sum of squares*) è definita da

$$SS_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 \quad (6.16)$$

e la stima di  $\sigma^2$  è

$$\hat{\sigma}^2 = \frac{SS_E}{n - 2} \quad (6.17)$$

Benché i residui siano  $n$ , l'Equazione (6.11) mostra che i residui soddisfano due equazioni. Di conseguenza, si può sfruttare la conoscenza di  $n - 2$  residui per calcolare i rimanenti due. Ecco perché nell'Equazione (6.17) compare al denominatore l'espressione  $n - 2$ .

### ESEMPIO 6.1 (proseguzione)

In Tabella 6.2 sono evidenziati sia  $SS_E = 57.7$ , sia  $\hat{\sigma}^2 = 3.2$  per il modello di regressione concentrazione sale-area carrabile. La quantità  $s = 1.791$  è una stima della deviazione standard degli errori del modello. (Si osservi che  $s$  non è esattamente uguale a  $\sqrt{\hat{\sigma}^2} = \sqrt{3.2}$  per via degli arrotondamenti effettuati da Minitab.)

**Tabella 6.2** Output di Minitab per l'analisi di regressione.

Regression Analysis: Salt conc ( $y$ ) versus Roadway area ( $x$ )

The regression equation is

$$\text{Salt conc } (y) = 2.68 + 17.5 \text{ Roadway area } (x)$$

Predictor	Coef	SE Coef	T	P
Constant	2.6765 $\leftarrow \hat{\beta}_0$	0.8680	3.08	0.006
Roadway area	17.5467 $\leftarrow \hat{\beta}_1$	0.9346	18.77	0.000
S = 1.791	R-Sq = 95.1%		R-Sq(adj) = 94.9%	

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1130.1	1130.1	352.46	0.000
Residual Error	18	57.7 $\leftarrow SS_E$	3.2 $\leftarrow \hat{\sigma}^2$		
Total	19	1187.9			

Obs	Roadway area	Salt con	Fit	SE Fit	Residual
1	0.19	3.800	6.010	0.715	-2.210
2	0.15	5.900	5.309	0.746	0.591
3	0.57	14.100	12.678	0.465	1.422
4	0.40	10.400	9.695	0.563	0.705
5	0.70	14.600	14.959	0.417	-0.359
6	0.67	14.500	14.433	0.425	0.067
7	0.63	15.100	13.731	0.440	1.369
8	0.47	11.900	10.923	0.519	0.977
9	0.75	15.500	15.837	0.406	-0.337
10	0.60	9.300	13.205	0.452	-3.905
11	0.78	15.600	16.363	0.403	-0.763
12	0.81	20.800	16.889	0.401	3.911
13	0.78	14.600	16.363	0.403	-1.763
14	0.69	16.600	14.784	0.420	1.816
15	1.30	25.600	25.487	0.599	0.113
16	1.05	20.900	21.101	0.453	-0.201
17	1.52	29.900	29.347	0.764	0.553
18	1.06	19.600	21.276	0.457	-1.676
19	1.74	31.300	33.208	0.945	-1.908
20	1.62	32.700	31.102	0.845	1.598

Predicted Values for New Observations

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	24.610	0.565	(23.424, 25.796)	(20.665, 28.555)

Values of Predictors for New Observations

New Obs	Roadway area
1	1.25

### Assunzioni per la regressione e proprietà del modello

Nella regressione lineare si assume in genere che gli errori del modello  $\epsilon_i$ , con  $i = 1, 2, \dots, n$  siano indipendenti e distribuiti normalmente con media nulla e varianza  $\sigma^2$ . I valori della variabile regressore  $x_i$  si assumono fissi prima della raccolta dei dati, di modo che la variabile risposta  $Y_i$  ha una distribuzione normale con media  $\beta_0 + \beta_1 x_i$  e varianza  $\sigma^2$ . Inoltre, sia  $\beta_0$  sia  $\beta_1$  possono venire scritti come combinazioni lineari delle  $Y_i$ . Le proprietà delle funzioni lineari di variabili aleatorie indipendenti e normali portano ai seguenti risultati.

#### Stimatori dei coefficienti, regressione lineare semplice

1.  $\hat{\beta}_0$  e  $\hat{\beta}_1$  sono **stimatori non distorti** rispettivamente dell'intercetta e della pendenza. Vale a dire: la distribuzione di  $\hat{\beta}_1$  (e di  $\hat{\beta}_0$ ) è centrata sul valore vero di  $\beta_1$  (e di  $\beta_0$ ).

2. Le varianze di  $\hat{\beta}_0$  e di  $\hat{\beta}_1$  sono rispettivamente

$$V(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \quad \text{e} \quad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

3. Le distribuzioni di  $\hat{\beta}_0$  e di  $\hat{\beta}_1$  sono **normali**.

Se nelle espressioni per le varianze della pendenza e dell'intercetta sostituiamo  $\sigma^2$  con  $\hat{\sigma}^2$  ricavata dall'Equazione (6.17), ed estraiamo la radice quadrata, otteniamo gli **errori standard** della pendenza e dell'intercetta.

#### Errori standard della pendenza e dell'intercetta, regressione lineare semplice

Gli errori standard della pendenza e dell'intercetta nella regressione lineare semplice sono rispettivamente

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \tag{6.18}$$

e

$$se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \tag{6.19}$$

#### ESEMPIO 6.1 (prosecuzione)

Minitab calcola gli errori standard della pendenza e dell'intercetta e li riporta in output (si veda la Tabella 6.2) accanto alle stime  $\hat{\beta}_0$  e  $\hat{\beta}_1$  dei coefficienti nella colonna intestata "SE Coef". Dall'output di Minitab troviamo che  $se(\hat{\beta}_0) = 0.8680$  ed  $se(\hat{\beta}_1) = 0.9346$ . Questi errori standard saranno usati per trovare gli intervalli di confidenza e per la verifica di ipotesi riguardanti la pendenza e l'intercetta.

### Regressione e analisi della varianza

La **somma totale dei quadrati** (*total sum of squares*),  $SS_T$ , dei valori  $y$  osservati

$$SS_T = S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} \quad (6.20)$$

è una misura della variabilità totale della risposta. Tale somma può venire riscritta come

#### ANOVA per l'analisi di regressione

$$SS_T = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 = SS_R + SS_E \quad (6.21)$$

Questa è un'**analisi della varianza** (ANOVA), simile all'ANOVA che abbiamo incontrato nel Paragrafo 5.8. Essa scomponete la variabilità totale della risposta in due componenti. Una di queste è la **somma dei quadrati dei residui o degli errori**  $SS_E$  (Equazione (6.16)), che rappresenta una misura della variabilità delle  $y$  non spiegata dal modello di regressione; l'altra componente,  $SS_R$ ,

#### Somma dei quadrati di regressione

$$SS_R = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

misura la variabilità spiegata dal modello di regressione. La componente  $SS_R$  è chiamata in genere **somma dei quadrati di regressione, o somma dei quadrati del modello**. Tali somme di quadrati sono riportate in Tabella 6.2 nella sezione dell'output intitolata "Analisi della varianza." Di solito si considera il rapporto  $SS_E/SS_T$  come la proporzione di variabilità della variabile risposta di cui non si può dare conto con il modello di regressione. Di conseguenza,  $1 - SS_E/SS_T$  è la frazione di variabilità della risposta che è giustificata dal modello.

#### Coefficiente di determinazione ( $R^2$ )

Il coefficiente di determinazione è definito da

$$R^2 = 1 - \frac{SS_E}{SS_T} \quad (6.22)$$

Lo si interpreta come la proporzione di variabilità nella variabile risposta osservata che è spiegata dal modello di regressione lineare. A volte è riportato  $100R^2$ , che viene interpretato come la percentuale di variabilità spiegata dal modello.

Minitab è in grado di calcolare e visualizzare anche la statistica  $R^2$ . Per esempio, per il modello di regressione relativo ai dati di Tabella 6.2, il programma riporta per  $100R^2$  il valore 95.1%, a indicare che il modello di regressione dà conto del 95.1% della variabilità osservata nei dati.

La scomposizione ANOVA implica  $0 \leq R^2 \leq 1$ . Un valore alto di  $R^2$  suggerisce che il modello ha avuto successo nello spiegare la variabilità della risposta; il fatto che  $R^2$  sia invece piccolo può significare che occorre trovare un modello alternativo, quale un modello di regressione multipla, che sia in grado di dare maggiormente conto della variabilità di  $y$ .

### Altri aspetti della regressione

I modelli di regressione sono usati prevalentemente per l'**interpolazione**. In altri termini, quando prediciamo una nuova osservazione sulla risposta (o stimiamo la risposta media) per un particolare valore del regressore  $x$ , dovremmo usare solo i valori di  $x$  appartenenti all'intervallo delle  $x$  usate per stimare il modello. Per esempio, nel problema dell'Esempio 6.1, i valori dell'area carrabile compresi tra 0.19% e 1.62% sono appropriati, ma un valore di  $x = 2.5$  non sarebbe ragionevole perché cade molto al di fuori dell'intervallo originale dei regressori. In sostanza, man mano che ci si allontana dall'intervallo dei dati originali diminuisce l'affidabilità dell'approssimazione lineare come modello empirico della reale relazione.

In tutto questo paragrafo abbiamo supposto che la variabile regressore  $x$  fosse controllabile e impostata a livelli scelti dall'analista, e che la variabile risposta  $Y$  fosse una variabile aleatoria. Esistono molte situazioni, però, in cui ciò non accade. In effetti, i dati di concentrazione di sale-area carrabile non erano controllati. L'analista ha selezionato un gruppo di 20 bacini e sia la concentrazione di sale *sia* l'area carrabile erano variabili aleatorie. I metodi di regressione che descriviamo in questo capitolo possono essere impiegati sia quando i valori dei regressori sono preventivamente fissati, sia quando sono aleatori; ci soffermeremo tuttavia sul caso dei regressori fissati, perché risulta abbastanza più semplice da descrivere. Quando  $Y$  e  $X$  sono entrambe aleatorie, possiamo anche usare la **correlazione** come misura del legame tra le due variabili: ne discuteremo brevemente nel Paragrafo 6.2.6.

L'espressione "analisi di regressione" è stata usata per la prima volta alla fine del diciannovesimo secolo da Sir Francis Galton, che studiò la relazione tra le altezze dei padri e dei figli. Galton mise a punto un modello per predire l'altezza di un figlio a partire dalla conoscenza di quella del padre, scoprendo che se il padre era di altezza superiore alla media, anche quella del figlio tendeva a esserlo, ma non quanto quella del padre. In altre parole, l'altezza dei figli regrediva verso la media.

## 6.2.2 Verifica delle ipotesi nella regressione lineare semplice

Spesso è utile sottoporre a verifica delle ipotesi che riguardano la pendenza e l'intercetta in un modello di regressione lineare. Continuano a valere le assunzioni di normalità effettuate sugli errori del modello, e di conseguenza sulla variabile risposta, che abbiamo introdotto nel Paragrafo 6.2.1.

### Impiego dei test $t$

Si supponga di voler verificare l'ipotesi che la pendenza sia uguale a una certa costante  $\beta_{1,0}$ . Le ipotesi appropriate sono:

$$\begin{aligned} H_0: \beta_1 &= \beta_{1,0} \\ H_1: \beta_1 &\neq \beta_{1,0} \end{aligned} \quad (6.23)$$

Siccome le risposte  $Y_i$  sono variabili aleatorie normali e indipendenti,  $\hat{\beta}_1$  è  $N(\beta_1, \sigma^2/S_{xx})$ . Come risultato, la statistica

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)} \quad (6.24)$$

segue la distribuzione  $t$  con  $n - 2$  gradi di libertà sotto l'ipotesi  $H_0: \beta_1 = \beta_{1,0}$ . Dovremmo rifiutare l'ipotesi  $H_0: \beta_1 = \beta_{1,0}$  se per il valore calcolato della statistica test valesse

$$|t_0| > t_{\alpha/2, n-2} \quad (6.25)$$

dove  $t_0$  viene calcolato dall'Equazione (6.24). Si può usare una procedura analoga per verificare le ipotesi riguardanti l'intercetta. Per verificare

$$\begin{aligned} H_0: \beta_0 &= \beta_{0,0} \\ H_1: \beta_0 &\neq \beta_{0,0} \end{aligned} \quad (6.26)$$

dovremmo usare la statistica

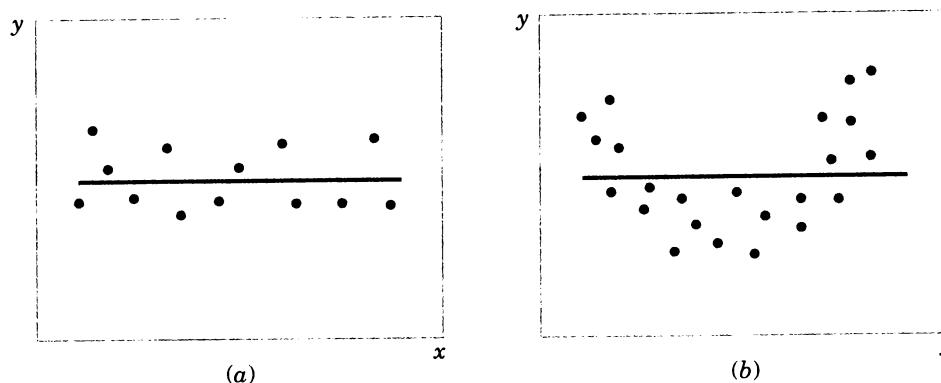
$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}} = \frac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)} \quad (6.27)$$

Il  $P$ -value è da calcolare come in qualsiasi test  $t$ . Per un livello fissato, rifiuteremmo l'ipotesi nulla se il valore calcolato di questa statistica test,  $t_0$ , fosse tale che  $|t_0| > t_{\alpha/2, n-2}$ .

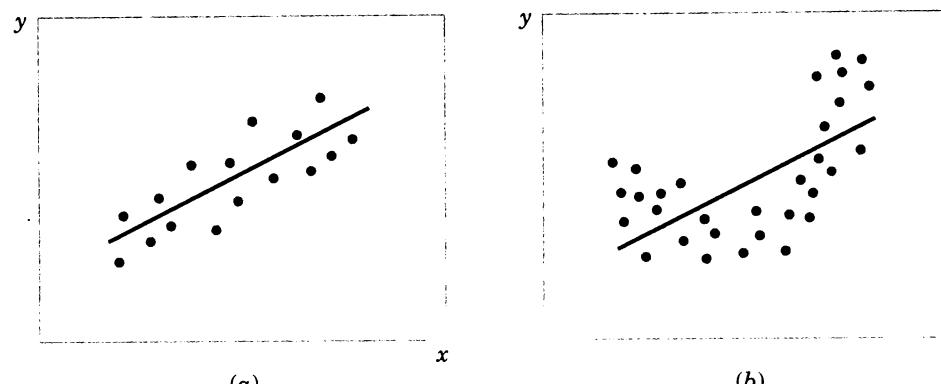
Un caso speciale molto importante delle ipotesi dell'Equazione (6.23) è

$$\begin{aligned} H_0: \beta_1 &= 0 \\ H_1: \beta_1 &\neq 0 \end{aligned} \quad (6.28)$$

Queste ipotesi sono collegate alla **significatività della regressione**. Non rifiutare  $H_0: \beta_1 = 0$  equivale a concludere che non esiste una relazione lineare tra  $x$  e  $Y$ . Questa situazione è illustrata in Figura 6.8. Si noti che ciò può implicare o che  $x$  è di scarso peso nella spiegazione della variazione di  $Y$  e che il migliore stimatore di  $Y$  per ogni  $x$  è  $\hat{y} = \bar{Y}$  (Figura 6.8a), o che la vera relazione tra  $x$  e  $Y$  non è lineare (Figura 6.8b). Dall'altro lato, rifiutare l'ipotesi nulla  $H_0: \beta_1 = 0$  significa che  $x$  è di peso non trascurabile nel giustificare la variabilità di  $Y$  (Figura 6.9). Rifiutare  $H_0: \beta_1 = 0$  potrebbe significare o che il modello della retta è adeguato (Figura 6.9a), o che, pur esistendo un effetto lineare di  $x$ , si potrebbero ottenere risultati migliori con l'aggiunta di termini polinomiali in  $x$  di grado più elevato (Figura 6.9b).



**Figura 6.8** L'ipotesi  $H_0: \beta_1 = 0$  non è rifiutata.



**Figura 6.9** L'ipotesi  $H_0: \beta_1 = 0$  è rifiutata.

### ESEMPIO 6.2 Concentrazione di sale e area stradale

Verifichiamo la significatività della regressione usando il modello adottato per i dati su concentrazione salina e superficie carrabile dell'Esempio 6.1. Le ipotesi sono:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Useremo  $\alpha = 0.01$ . Dall'Esempio 6.1 e dall'output di Minitab in Tabella 6.2 si ricava

$$\hat{\beta}_1 = 17.5467 \quad n = 20, \quad S_{xx} = 3.67068, \quad \hat{\sigma}^2 = 3.2$$

per cui la statistica  $t$  dell'Equazione (6.24) è

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{17.5467}{\sqrt{3.2/3.67068}} = 18.77$$

Essendo il valore critico di  $t$  uguale a  $t_{0.005, 18} = 2.88$ , il valore della statistica test è nettamente dentro la regione critica, e di conseguenza  $H_0: \beta_1 = 0$  dovrebbe essere rifiutata. Il  $P$ -value per questo test è prossimo a zero. Questo risultato è stato ottenuto manualmente con una calcolatrice e indica una forte significatività di  $\beta_1$ .

In Tabella 6.2 è mostrato l'output di Minitab per questo problema. Si osservi che il valore della statistica  $t$  per la pendenza calcolato è 18.77, e che il  $P$ -value riportato è  $P = 0.000$ . Minitab riporta anche la statistica  $t$  per la verifica dell'ipotesi  $H_0: \beta_0 = 0$ . Questa statistica è calcolata mediante l'Equazione (6.27) con  $\beta_{0,0} = 0$ : si ottiene  $t_0 = 3.08$ . Poiché il  $P$ -value è 0.006, l'ipotesi che l'intercetta sia zero viene rifiutata. Ciò significa che si è rilevata una relazione fra l'area stradale e la concentrazione di sale.

### Approccio mediante analisi della varianza

L'analisi della varianza può essere impiegata anche per valutare la significatività della regressione. Se è vera l'ipotesi nulla per la significatività della regressione,  $H_0: \beta_1 = 0$ ,  $SS_R/\sigma^2$  è una variabile aleatoria chi-quadro con 1 grado di libertà. Si noti che il numero di gradi di libertà per questa variabile aleatoria chi-quadro è uguale al numero di variabili regressore del modello. Si può anche dimostrare che  $SS_E/\sigma^2$  è una variabile aleatoria chi-quadro con  $n - 2$  gradi di libertà, e che  $SS_E$  e  $SS_R$  sono indipendenti.

#### Verifica della significatività della regressione nella regressione lineare semplice

$$MS_R = \frac{SS_R}{1} \quad MS_E = \frac{SS_E}{n - 2} \quad (6.29)$$

Ipotesi nulla:	$H_0: \beta_1 = 0$
Ipotesi alternativa:	$H_1: \beta_1 \neq 0$
Statistica test:	$F_0 = \frac{MS_R}{MS_E}$
Criterio di rifiuto:	$f_0 > f_{\alpha/2, n-2}$
P-value:	Probabilità oltre $f_0$ nella distribuzione $F_{1, n-2}$

Il test dell'ANOVA per la significatività della regressione è solitamente riassunto in una tabella, come la Tabella 6.3.

Tabella 6.3 Analisi della varianza per la verifica della significatività della regressione.

Causa della variazione	Somma dei quadrati	Gradi di libertà	Media quadratica	$F_0$
Regressione	$SS_R$	1	$MS_R$	$MS_R/MS_E$
Errore o residuo	$SS_E$	$n - 2$	$MS_E$	
Totale	$SS_T$	$n - 1$		

#### ESEMPIO 6.2 (proseguimento)

L'output di Minitab mostrato in Tabella 6.2 contiene il test di analisi della varianza per la significatività della regressione. Il valore calcolato della statistica  $F$  per la significatività della regressione è  $f_0 = MS_R/MS_E = 1130.1/3.2 = 352.46$ . Minitab riporta per questo test un P-value di 0.000 (il P-value effettivo è  $2.87 \times 10^{-13}$ ). Pertanto rifiutiamo l'ipotesi nulla che la pendenza della linea di regressione sia zero e concludiamo che esiste una relazione lineare tra la concentrazione di sale e l'area carrabile.

Il test  $t$  per la significatività della regressione è in stretta relazione con il test  $F$  dell'ANOVA; in effetti, i due test forniscono risultati identici. Ciò non dovrebbe sorprendere, perché tali procedure verificano le medesime ipotesi. Risulta che il quadrato del valore calcolato dalla statistica test,  $t_0$ , è uguale al valore calcolato dalla statistica test dell'ANOVA,  $f_0$  (a meno di

arrotondamenti che possono influire sui risultati). Per rendersene conto, si osservi l'output di Minitab (Tabella 6.2), dove si può notare che  $t_0^2 = 18.77^2 = 352.3$ : a meno di arrotondamenti nei numeri riportati da Minitab, tale valore è uguale a quello della statistica  $F$  dell'ANOVA. In generale, il quadrato di una variabile aleatoria  $t$  con  $r$  gradi di libertà è uguale a una variabile aleatoria  $F$  con un grado di libertà a numeratore e  $r$  gradi di libertà a denominatore.

### 6.2.3 Intervalli di confidenza nella regressione lineare semplice

#### Intervalli di confidenza per la pendenza e l'intercetta

In aggiunta alle stime puntuali della pendenza e dell'intercetta, per questi parametri è possibile costruire intervalli di confidenza. L'ampiezza di questi intervalli di confidenza è una misura della qualità complessiva della retta di regressione. Se i termini di errore  $\epsilon_i$  nel modello di regressione sono indipendenti e distribuiti normalmente, si ha che

$$(\hat{\beta}_1 - \beta_1)/se(\hat{\beta}_1) \quad \text{e} \quad (\hat{\beta}_0 - \beta_0)/se(\hat{\beta}_0)$$

sono entrambi distribuiti come variabili aleatorie  $t$  con  $n - 2$  gradi di libertà. Ciò porta alla seguente definizione di intervalli di confidenza di livello  $100(1 - \alpha)\%$  per la pendenza e l'intercetta.

#### Intervalli di confidenza per i parametri del modello, regressione lineare semplice

Sotto l'assunzione che le osservazioni siano indipendenti e distribuite normalmente, un **intervallo di confidenza di livello  $100(1 - \alpha)\%$  per la pendenza  $\beta_1$**  nella regressione lineare semplice è

$$\hat{\beta}_1 - t_{\alpha/2,n-2} se(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2} se(\hat{\beta}_1) \quad (6.31)$$

Analogamente, un **intervallo di confidenza di livello  $100(1 - \alpha)\%$  per l'intercetta  $\beta_0$**  è

$$\hat{\beta}_0 - t_{\alpha/2,n-2} se(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2} se(\hat{\beta}_0) \quad (6.32)$$

dove  $se(\hat{\beta}_1)$  e  $se(\hat{\beta}_0)$  sono definiti rispettivamente nelle Equazioni (6.18) e (6.19).

**ESEMPIO 6.3**  
Concentrazione  
di sale e area  
stradale

Troviamo un intervallo di confidenza di livello 95% per la pendenza della retta di regressione usando i dati dell'Esempio 6.1. Ricordiamo che  $\hat{\beta}_1 = 17.5467$  e che  $se(\hat{\beta}_1) = 0.9346$  (si veda la Tabella 6.2). Pertanto, dall'Equazione (6.30) ricaviamo

$$\hat{\beta}_1 - t_{0.025,18} se(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{0.025,18} se(\hat{\beta}_1)$$

ovvero

$$17.5467 - 2.101(0.9346) \leq \beta_1 \leq 17.5467 + 2.101(0.9346)$$

Svolgendo i conti, si ha

$$15.5831 \leq \beta_1 \leq 19.5103$$

L'intervallo di confidenza indica che una variazione dell'1% dell'area carrabile corrisponde a un aumento della concentrazione di sale compreso fra 15.5 e 19.5 milligrammi/litro.

### Intervalli di confidenza per la risposta media

Si può costruire un intervallo di confidenza per la risposta media a un valore specifico di  $x$ , per esempio  $x_0$ . Si tratta di un intervallo di confidenza per  $E(Y|x_0) = \mu_{Y|x_0}$ , spesso chiamato intervallo di confidenza per la retta di regressione. Siccome  $E(Y|x_0) = \mu_{Y|x_0} = \beta_0 + \beta_1 x_0$ , possiamo ottenere una stima puntuale della media di  $Y$  in  $x = x_0$  (ossia  $\hat{\mu}_{Y|x_0}$ ) dal modello stimato come

$$\hat{\mu}_{Y|x_0} = \hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Ora,  $\hat{\mu}_{Y|x_0}$  è uno stimatore puntuale non distorto di  $\mu_{Y|x_0}$ , essendo  $\hat{\beta}_0$  e  $\hat{\beta}_1$  stimatori non distorti di  $\beta_0$  e  $\beta_1$ . La varianza di  $\hat{\mu}_{Y|x_0}$  è

$$V(\hat{\mu}_{Y|x_0}) = \sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] \quad (6.33)$$

Inoltre,  $\hat{\mu}_{Y|x_0}$  è distribuito normalmente, perché lo sono  $\hat{\beta}_1$  e  $\hat{\beta}_0$ ; se usiamo  $\hat{\sigma}^2$  come stima di  $\sigma^2$ , è semplice mostrare che

$$\frac{\hat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\sqrt{\hat{\sigma}^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}} = \frac{\hat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{se(\hat{\mu}_{Y|x_0})}$$

ha una distribuzione  $t$  con  $n - 2$  gradi di libertà. La grandezza  $\hat{\mu}_{Y|x_0}$  è chiamata a volte errore standard del valore stimato. Tutto ciò porta alla seguente definizione dell'intervallo di confidenza.

#### Intervallo di confidenza per la risposta media, regressione lineare semplice

Un **intervallo di confidenza di livello  $100(1 - \alpha)\%$  per la risposta media** in corrispondenza del valore  $x = x_0$ , cioè  $\mu_{Y|x_0}$ , è dato da

$$\hat{\mu}_{Y|x_0} - t_{\alpha/2,n-2} se(\hat{\mu}_{Y|x_0}) \leq \mu_{Y|x_0} \leq \hat{\mu}_{Y|x_0} + t_{\alpha/2,n-2} se(\hat{\mu}_{Y|x_0}) \quad (6.34)$$

dove  $\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$  è calcolato dal modello di regressione stimato.

Si noti che l'ampiezza dell'intervallo di confidenza per  $\mu_{Y|x_0}$  è funzione del valore specificato per  $x_0$ ; essa assume valore minimo per  $x_0 = \bar{x}$  e si allarga all'aumentare di  $|x_0 - \bar{x}|$ .

**ESEMPIO 6.4**  
Concentrazione  
di sale e area  
stradale

Costruiamo un intervallo di confidenza di livello 95% per la risposta media per i dati dell’Esempio 6.1. La stima è  $\hat{\mu}_{Y|x_0} = 2.6765 + 17.5467x_0$ , e dall’Equazione (6.34) si ricava l’intervallo di confidenza di livello 95% per  $\mu_{Y|x_0}$

$$\hat{\mu}_{Y|x_0} \pm 2.101 \text{ se}(\hat{\mu}_{Y|x_0}) \quad \text{se}(\hat{\mu}_{Y|x_0}) = \sqrt{3.2 \left[ \frac{1}{20} + \frac{(x_0 - 0.824)^2}{3.67068} \right]}$$

Supponiamo di essere interessati a predire la concentrazione media di sale quando l’area carrabile  $x_0$  vale 1.25%. Allora

$$\hat{\mu}_{Y|1.25} = 2.6765 + 17.5467(1.25) = 24.61$$

e l’intervallo di confidenza di livello 95% è

$$\left\{ 24.61 \pm 2.101 \sqrt{3.2 \left[ \frac{1}{20} + \frac{(1.25 - 0.824)^2}{3.67068} \right]} \right\}$$

ovvero

$$24.61 \pm 2.101(0.564)$$

Pertanto, l’intervallo di confidenza di livello 95% per  $\mu_{Y|1.25}$  è

$$23.425 \leq \mu_{Y|1.25} \leq 24.795$$

Minitab esegue anche questi calcoli; si veda la Tabella 6.2, in cui è mostrato il valore predetto di  $y$  in  $x = 1.25$  insieme a  $\text{se}(\hat{\mu}_{Y|1.25})$  e all’intervallo di confidenza al 95% per la media di  $y$  a questo livello di  $x$ . Minitab contrassegna l’errore standard  $\text{se}(\hat{\mu}_{Y|1.25})$  con “SE Fit”.

Ripetendo questi calcoli per differenti valori di  $x_0$ , possiamo ottenere i limiti di confidenza per ciascun valore corrispondente di  $\mu_{Y|x_0}$ . Minitab calcola l’errore standard  $\text{se}(\hat{\mu}_{Y|x_0})$  per ogni valore  $x$  del campione. In Tabella 6.2 questi errori standard sono riportati nella colonna contrassegnata “SE Fit”. La Figura 6.10 mostra il diagramma di dispersione prodot-

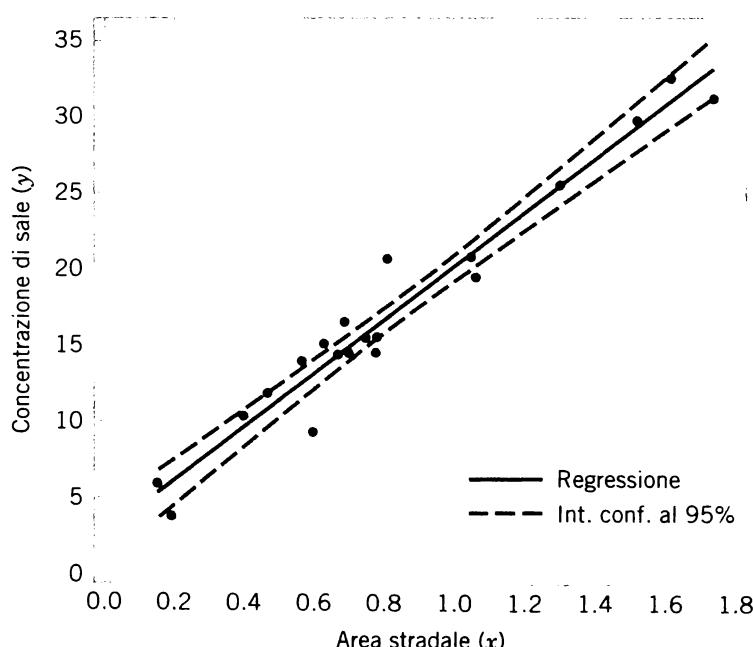


Figura 6.10  
Diagramma  
di dispersione  
della concentrazione di  
sale e dell’area carra-  
bile (Esempio 6.1),  
con la retta  
di regressione stimata  
e i limiti di confidenza  
al 95% per .

to da Minitab insieme al modello stimato e i corrispondenti limiti di confidenza al 95% (le curve tratteggiate). Il livello di confidenza del 95% si applica solo all'intervallo ottenuto per un valore di  $x$ , e non all'intero insieme dei livelli  $x$ . Si noti che l'ampiezza dell'intervallo di confidenza per  $\mu_{Y|x_0}$  aumenta al crescere di  $|x_0 - \bar{x}|$ .

#### 6.2.4 Predizione di nuove osservazioni

Un'importante applicazione del modello di regressione consiste nel predire nuove o future osservazioni  $Y$  corrispondenti a uno specifico livello della variabile regressore  $x$ . Se  $x_0$  è il valore della variabile regressore di interesse, si ha che

$$\hat{Y}_0 = \hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \quad (6.35)$$

è lo stimatore puntuale del nuovo o futuro valore della risposta  $Y_0$ .

Vogliamo ora ottenere una stima intervallare per l'osservazione futura  $Y_0$ ; quest'ultima è indipendente dalle osservazioni usate per sviluppare il modello di regressione, perciò l'intervallo di confidenza per  $\mu_{Y|x_0}$  nell'Equazione (6.34) è inappropriato, dato che si basa solo sui dati usati nella stima della retta di regressione. L'intervallo di confidenza per  $\mu_{Y|x_0}$  si riferisce alla risposta media reale in  $x = x_0$  (vale a dire a un parametro della popolazione), e non a osservazioni future.

Sia  $Y_0$  l'osservazione futura in  $x = x_0$ , e sia  $\hat{Y}_0$  lo stimatore di  $Y_0$  dato dall'Equazione (6.35). Si noti che l'errore nella predizione  $Y_0 - \hat{Y}_0$  è una variabile aleatoria distribuita normalmente con media zero e varianza

$$V(Y_0 - \hat{Y}_0) = \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]$$

poiché  $Y_0$  è indipendente da  $\hat{Y}_0$ . Se usiamo  $\hat{\sigma}^2$  per stimare  $\sigma^2$ , possiamo dimostrare che

$$\frac{Y_0 - \hat{Y}_0}{\sqrt{\hat{\sigma}^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}}$$

ha una distribuzione  $t$  con  $n - 2$  gradi di libertà. A partire da questo risultato possiamo sviluppare la seguente definizione di **intervallo di predizione**.

### Intervallo di predizione per un'osservazione futura, regressione lineare semplice

Un intervallo di predizione di livello  $100(1 - \alpha)\%$  per un'osservazione futura  $Y_0$  in corrispondenza del valore  $x_0$  è dato da

$$\begin{aligned} \hat{y}_0 &= t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]} \\ &\leq Y_0 \leq \hat{y}_0 + t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]} \end{aligned} \quad (6.36)$$

Il valore  $\hat{y}_0$  è calcolato dal modello di regressione  $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ .

Si noti che l'ampiezza dell'intervallo di predizione assume il minimo in  $x_0 - \bar{x}$  e si allarga all'aumentare di  $|x_0 - \bar{x}|$ . Confrontando l'Equazione (6.36) con l'Equazione (6.34), osserviamo che l'intervallo di predizione per il punto  $x_0$  è sempre più ampio dell'intervallo di confidenza per il medesimo punto. Ciò avviene perché l'intervallo di predizione dipende sia dall'errore legato alla stima del modello, sia dall'errore associato a osservazioni future. L'intervallo di predizione nell'Equazione (6.35) è analogo all'intervallo di predizione per una futura osservazione estratta da una distribuzione normale, introdotto nel Paragrafo 4.8.1; l'unica differenza consiste nel fatto che in questo caso nella determinazione del valore futuro è coinvolta una variabile regressore.

**ESEMPIO 6.5**  
Concentrazione  
di sale e area  
stradale

Per illustrare la costruzione di un intervallo di predizione, supponiamo di usare i dati dell'Esempio 6.1 e di trovare un intervallo di predizione di livello 95% per un'osservazione futura della concentrazione di sale quando l'area carrabile è  $x_0 = 1.25\%$ . Usando l'Equazione (6.35) e ricordando dall'Esempio 6.4 che  $\hat{y} = 24.61$ , troviamo che l'intervallo di predizione è

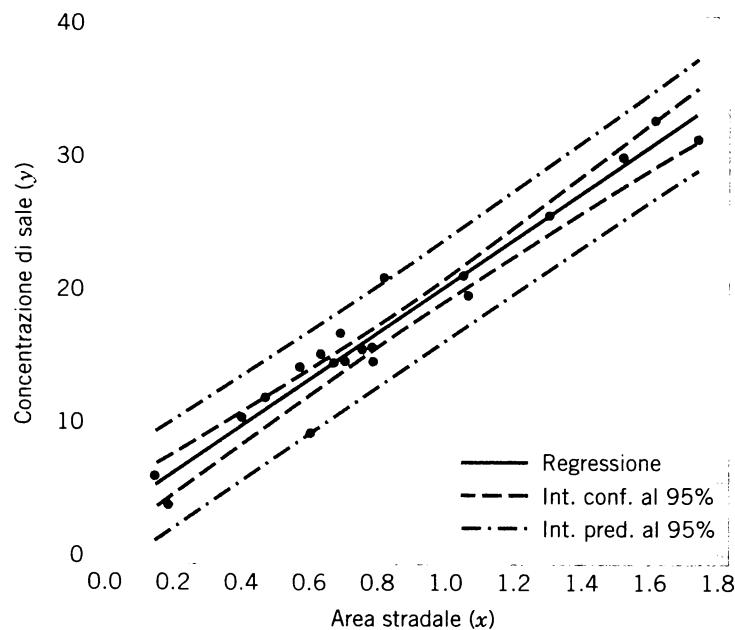
$$\begin{aligned} 24.61 &- 2.101 \sqrt{3.2 \left[ 1 + \frac{1}{20} + \frac{(1.25 - 0.824)^2}{3.67068} \right]} \\ &\leq Y_0 \leq 24.61 + 2.101 \sqrt{3.2 \left[ 1 + \frac{1}{20} + \frac{(1.25 - 0.824)^2}{3.67068} \right]} \end{aligned}$$

che diventa

$$20.66 \leq y_0 \leq 28.55$$

Minitab calcola anche gli intervalli di predizione. In Tabella 6.2 è mostrato l'intervallo di predizione al 95% su una futura osservazione quando  $x_0 = 1.25\%$ .

Ripetendo i calcoli precedenti per differenti livelli di  $x_0$ , possiamo ottenere gli intervalli di predizione al 95%, mostrati graficamente in Figura 6.11 come la curva superiore e quella inferiore intorno alla retta di regressione stimata. Si noti che il grafico mostra anche i limiti di confidenza al 95% per calcolati nell'Esempio 6.4. Questo illustra il fatto che i limiti di predizione sono sempre più ampi di quelli di confidenza.



**Figura 6.11**  
Diagramma  
di dispersione dei dati  
concentrazione  
di sale-area carrabile  
dell'Esempio 6.1,  
con la retta  
di regressione stimata,  
i limiti di predizione  
al 95% (curve esterne)  
e i limiti di confidenza  
al 95% per  $\mu_{Y|x_0}$ .

### 6.2.5 Controllo dell'adeguatezza del modello

L'adattamento ai dati di un modello di regressione richiede diverse assunzioni. La stima dei parametri del modello richiede l'assunzione che gli errori siano variabili aleatorie non correlate con media nulla e varianza costante. Le verifiche di ipotesi e la stima intervallare richiedono che gli errori siano distribuiti normalmente. Inoltre, si assume che l'ordine del modello sia corretto: se cioè utilizziamo un modello di regressione semplice, assumiamo che il fenomeno si comporti realmente in modo lineare.

L'analista dovrebbe sempre dubitare della validità di queste assunzioni, e condurre opportune analisi per esaminare l'adeguatezza del modello provvisoriamente preso in considerazione.

I residui del modello di regressione, definiti da  $e_i = y_i - \hat{y}_i$ ,  $i = 1, 2, \dots, n$ , sono spesso utili per il controllo delle assunzioni su normalità e costanza della varianza, nonché per determinare l'utilità o meno di aggiungere ulteriori termini al modello.

Come controllo approssimato della normalità, lo sperimentatore può costruire un istogramma delle frequenze dei residui o un **grafico dei quantili per i residui**. Molti programmi per computer sono in grado di fornire un tale grafico, e considerato anche che le dimensioni del campione nella regressione sono spesso troppo piccole perché l'istogramma possa essere significativo, si preferisce il metodo del grafico dei quantili. Esso richiede cautela nella valutazione della "anormalità" di tali grafici. (Si faccia riferimento alla discussione del metodo "della matita" del Capitolo 3.)

Possiamo anche **standardizzare** i residui calcolando  $d_i = e_i / \sqrt{\hat{\sigma}^2}$ , con  $i = 1, 2, \dots, n$ . Se gli errori sono distribuiti normalmente, circa il 95% dei residui standarizzati dovrebbe cadere nell'intervallo  $(-2, +2)$ . I residui che sono molto al di fuori di questo intervallo possono indicare la presenza di un **outlier**, cioè di un'osservazione anomala rispetto al gruppo di dati rimanenti. Sono state proposte varie regole per scartare gli outlier; questi ultimi, tuttavia, alcune volte forniscono importanti informazioni riguardo a condizioni insolite che possono interessare lo sperimentatore, nel qual caso non dovrebbero essere scartati. Per un'ulteriore discussione sugli outlier si veda Montgomery, Peck, Vining (2006).

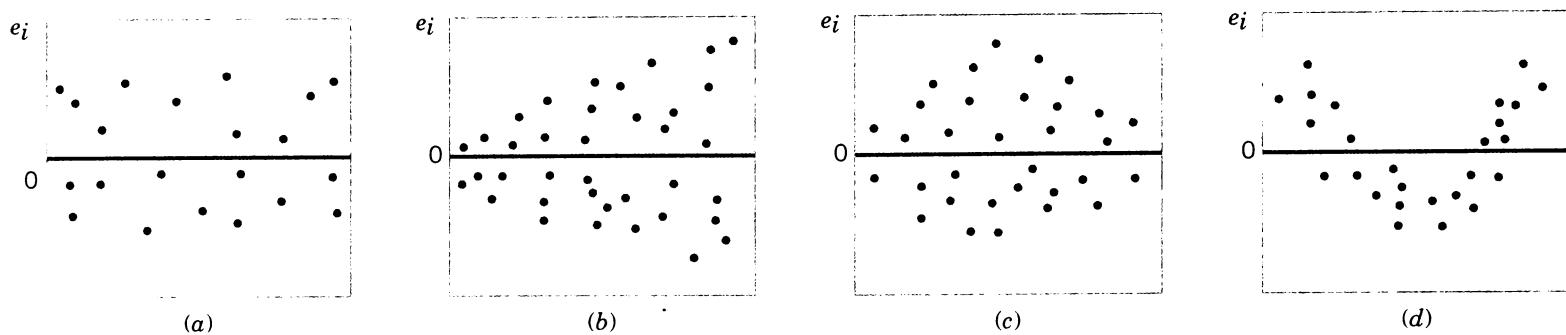


Figura 6.12 Andamenti tipici per i grafici dei residui: (a) soddisfacente, (b) a imbuto, (c) a doppio arco, (d) non lineare. Sull'asse orizzontale vi può essere la variabile tempo, la variabile  $\hat{y}_i$  o la variabile  $x_i$ .

È spesso utile rappresentare graficamente i residui: (1) in sequenza temporale (se nota), (2) in funzione di  $\hat{y}_i$  e (3) in funzione della variabile indipendente  $x$ . Questi grafici assomiglieranno in genere a uno dei quattro andamenti mostrati in Figura 6.12. La forma (a) in Figura 6.12 rappresenta la situazione ideale, mentre le forme (b), (c) e (d) rappresentano situazioni anomale. Se i residui appaiono come in (b), la varianza delle osservazioni può aumentare con il tempo o con il valore di  $y_i$  o  $x_i$ . Spesso si usa operare una trasformazione sulla risposta  $y$  per eliminare questo problema. Le più diffuse trasformazioni volte a stabilizzare la varianza prevedono l'impiego delle funzioni  $\sqrt{y}$ ,  $\ln y$  o  $1/y$  come risposta. [Si veda Montgomery, Peck, Vining (2006) per maggiori dettagli sui criteri di scelta di un'appropriata trasformazione.] Se un grafico dei residui in funzione del tempo appare come il grafico in (b), la varianza delle osservazioni aumenta con il tempo. Grafici dei residui come quello in (c) indicano anch'essi disomogeneità della varianza. Grafici dei residui che appaiono come quello in (d) indicano una inadeguatezza del modello; in quest'ultimo caso occorre aggiungere al modello termini di ordine più alto, o prendere in considerazione una trasformazione della variabile  $x$  o della variabile  $y$  (o di entrambe), oppure considerare altri regressori. Gli outlier possono avere un effetto enorme su un modello di regressione. Come si farà notare più avanti, un residuo alto è spesso indizio della presenza di un outlier.

### ESEMPIO 6.6 Concentrazione di sale e area stradale

In Tabella 6.2 sono mostrati i residui per il modello di regressione dei dati concentrazione di sale-area carrabile. Si analizzino i residui per verificare se il modello di regressione è adeguato per i dati raccolti o se qualcuna delle ipotesi soggiacenti viene violata.

**Soluzione.** Un grafico dei quantili per questi residui è mostrato in Figura 6.13. Non si riscontrano evidenti deviazioni dalla normalità, anche se i due residui più estremi non risultano molto vicini alla retta che attraversa gli altri residui. Il grafico dei residui in funzione di  $\hat{y}$  è mostrato in Figura 6.14. Non vi sono indicazioni che contrastino con l'assunzione di varianza costante.

Nel modello di regressione dei dati concentrazione di sale-area carrabile, i due residui maggiori sono  $e_{10} = -3.905$  ed  $e_{12} = -3.911$  (si faccia riferimento alla Tabella 6.2). I residui standardizzati sono  $d_{10} = e_{10}/\sqrt{\hat{\sigma}^2} = -3.905/\sqrt{3.2} = -2.183$  e  $d_{12} = e_{12}/\sqrt{\hat{\sigma}^2} = 3.911/\sqrt{3.2} = 2.186$ ; essi non sono abbastanza lontani dall'intervallo nominale  $-2 \div +2$ , dove ci si aspetterebbe che cadesse la maggior parte dei residui standardizzati, per causare qualche preoccupazione.

È facile mostrare l'impatto che può avere un outlier. Supponiamo che la concentrazione di sale per l'osservazione 12 sia  $y_{12} = 28.8$  (invece di 20.8). La Figura 6.15 mostra un grafico di dispersione di questo insieme di dati modificato con la relativa retta dei minimi quadrati. Usando Minitab, si può facilmente verificare che il valore stimato corrispondente all'osservazione

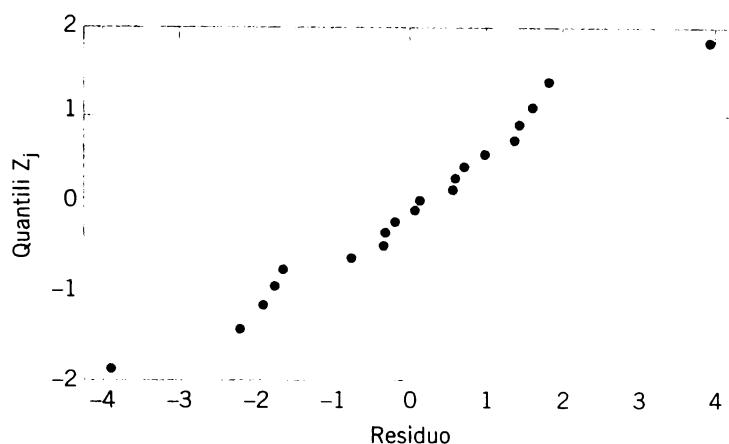


Figura 6.13 Grafico dei quantili per i residui per il modello di regressione concentrazione di sale-area carrabile.

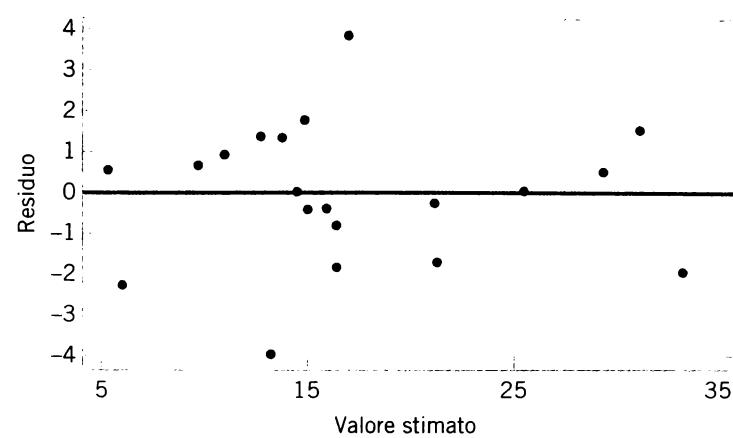


Figura 6.14 Grafico dei residui in funzione dei valori stimati per il modello di regressione concentrazione di sale-area carrabile.

12 è ora  $\hat{y}_{12} = 17.29$ , e che il corrispondente residuo è  $v_{12} - \hat{v}_{12} = 28.8 - 17.29 = 11.51$ . Il valore standardizzato di questo residuo è  $d_{12} = e_{12}/\sqrt{\hat{\sigma}^2} = 11.51/\sqrt{10.1} = 3.62$  ( $MS_E$  o  $\hat{\sigma}^2$  nel nuovo modello di regressione), che è sufficientemente lontano dall'intervallo nominale  $-2 \div +2$  per classificare l'osservazione 12 come outlier. L'impatto effettivo sulla retta di regressione di questo outlier appare comunque decisamente modesto. Confrontando le Figure 6.15 e 6.7 (che mostra la retta dei minimi quadrati per i dati originali) si rileva che la pendenza del modello di regressione non è stata seriamente influenzata dall'outlier (17.5467 rispetto a 17.5467); l'intercetta è invece aumentata di molto, proporzionalmente: da 2.6765 a 3.102. L'outlier ha sostanzialmente alzato l'altezza media della retta stimata.

Supponiamo ora che la risposta per l'osservazione 19 sia 61.3 invece di 31.3. In Figura 6.16 sono mostrati il grafico di dispersione e la retta stimata. Questo outlier ha avuto un impatto più deciso, tanto da allontanare la retta stimata dal resto dei dati. Ciò è dovuto sia alla dimensione dell'outlier, sia alla sua posizione lungo l'asse  $x$ . I punti campionari prossimi

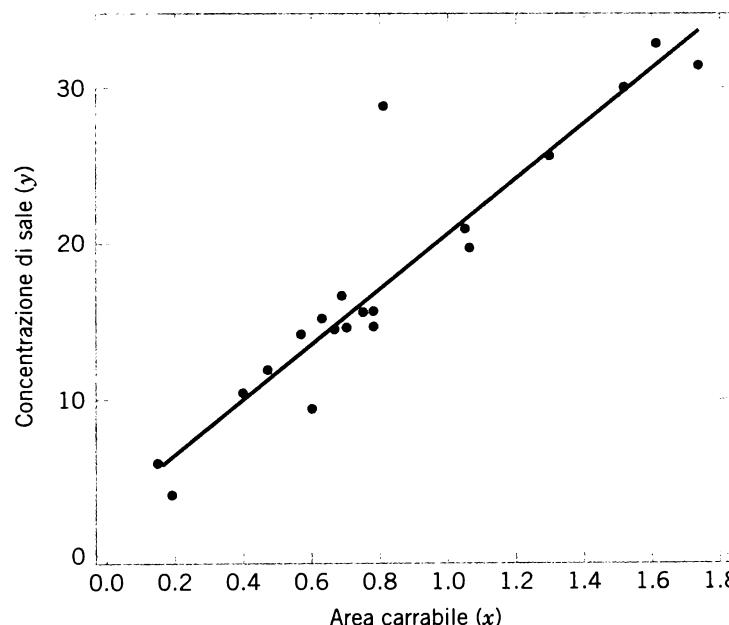


Figura 6.15 Effetto di un outlier.

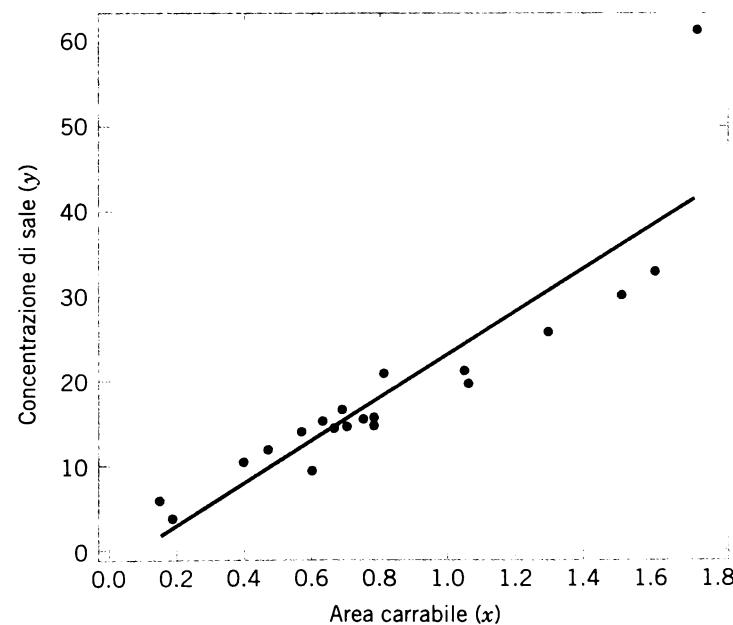


Figura 6.16 Effetto di una osservazione influente.

agli estremi dell'intervallo delle  $x$  hanno un impatto potenzialmente maggiore sulla retta stimata che non quelli posti vicino al centro dell'intervallo delle  $x$ . I punti che sono verso gli estremi dell'intervallo delle  $x$  e che hanno residui elevati sono spesso chiamati **osservazioni influenti**; per identificarli nella regressione lineare semplice è utile un grafico di dispersione. Nella regressione multipla, invece, la dimensionalità del problema può rendere difficile la loro identificazione. Diremo qualcosa in più sulle osservazioni influenti nel Paragrafo 6.3.3.

### 6.2.6 Correlazione e regressione

Abbiamo sottolineato nel Paragrafo 6.2.1 che nel nostro sviluppo della regressione vi era l'assunto che la variabile regressore  $x$  fosse fissata o scelta preventivamente, e che la variabile risposta  $Y$  fosse una variabile aleatoria, ma che i risultati per la stima del parametro e l'inferenza del modello si potevano ancora applicare anche quando  $Y$  e  $X$  erano entrambe variabili aleatorie. In questo Paragrafo discutiamo ulteriormente questo punto e mostriamo alcune delle connessioni tra regressione e correlazione.

Si supponga che  $X$  e  $Y$  siano variabili aleatorie congiuntamente normali con coefficiente di correlazione  $\rho$  (le distribuzioni congiunte sono state introdotte nel Paragrafo 3.11). Il coefficiente  $\rho$ , detto **coefficiente di correlazione della popolazione**, misura l'intensità della relazione lineare tra  $X$  e  $Y$  nella popolazione o nella distribuzione congiunta. Abbiamo anche un insieme di coppie campionarie  $(x_i, y_i)$ , con  $i = 1, 2, \dots, n$ ; il **coefficiente di correlazione campionario** tra  $Y$  e  $X$  è dato da

$$r = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (6.37)$$

Abbiamo illustrato il calcolo di  $r$  nel Paragrafo 2.6 e ne abbiamo discusso l'interpretazione. Il coefficiente di correlazione campionario per i dati concentrazione di sale-area carrabile è  $r = 0.975$ , il che indica una forte correlazione positiva. Si noti che  $r$  è semplicemente la radice quadrata del coefficiente di determinazione,  $R^2$ , con il segno preso uguale al segno della pendenza.

Il coefficiente di correlazione campionario è anche in stretta relazione con la pendenza nel modello di regressione lineare; infatti

$$r = \hat{\beta}_1 \left( \frac{S_{xy}}{SS_T} \right)^{1/2} \quad (6.38)$$

per cui verificare l'ipotesi che la pendenza sia nulla (significatività della regressione) equivale a verificare che il coefficiente di correlazione della popolazione sia nullo. Possiamo anche condurre questa verifica direttamente, vale a dire verificare

$$\begin{aligned} H_0: \rho &= 0 \\ H_1: \rho &\neq 0 \end{aligned} \quad (6.39)$$

Il valore calcolato della statistica test è

$$t_0 = \frac{r\sqrt{n-3}}{\sqrt{1-r^2}} \quad (6.40)$$

e se  $|t_0| > t_{\alpha/2,n-2}$  l'ipotesi nulla nell'Equazione (6.39) viene rifiutata.

### ESEMPIO 6.6 (prosecuzione)

Usando i dati relativi alla concentrazione e all'area carrabile e l'output di Minitab si ha

$$r = 17.5467 \left( \frac{3.67068}{1187.9} \right)^{1/2} = 0.9754$$

Pertanto la statistica test è

$$t_0 = \frac{0.9754\sqrt{20-3}}{\sqrt{1-(0.9754)^2}} = 18.24$$

Il valore della statistica test, 18.24, supera il valore critico  $t_{0.005,18} = 2.88$  e ha un  $P$ -value pari a 0. Si noti che questo valore della statistica test è uguale (eccetto che per gli arrotondamenti) alla statistica test per  $\beta_1$ , 18.77. Entrambi i test portano alla stessa conclusione: la correlazione fra  $X$  e  $Y$  è significativa, ovvero il modello di regressione è significativo.

Vi sono altre verifiche di ipotesi e costruzioni di intervalli di confidenza associate al coefficiente di correlazione campionario  $\rho$  [si veda Montgomery, Ranger (2011)].

## 6.3 REGRESSIONE MULTIPLA

In questo paragrafo consideriamo il modello di regressione lineare multipla introdotto nel Paragrafo 6.1. Come abbiamo fatto per la regressione lineare semplice, mostreremo come stimare i parametri del modello usando il metodo dei minimi quadrati, come verificare ipotesi e costruire intervalli di confidenza per i parametri del modello, come predire osservazioni future e stabilire se il modello è adeguato o meno.

### 6.3.1 Stima dei parametri nella regressione multipla

Per stimare i coefficienti di regressione nel modello di regressione multipla si può usare il metodo dei minimi quadrati (Equazione (6.3)). Si supponga di avere  $n$  osservazioni, con  $n > k$ , e sia  $x_{ij}$  l' $i$ -esima osservazione, o livello, della variabile  $x_j$ . Le osservazioni sono

$$(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) \quad i = 1, 2, \dots, n > k$$

Si è soliti presentare i dati per la regressione multipla in una tabella come la Tabella 6.4.

Tabella 6.4 Dati per la regressione lineare multipla.

$y$	$x_1$	$x_2$	$\dots$	$x_k$
$y_1$	$x_{11}$	$x_{12}$	$\dots$	$x_{1k}$
$y_2$	$x_{21}$	$x_{22}$	$\dots$	$x_{2k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$y_n$	$x_{n1}$	$x_{n2}$	$\dots$	$x_{nk}$

Ciascuna osservazione  $(x_{i1}, x_{i2}, \dots, x_{ik}, y_i)$  soddisfa il modello dell'Equazione (6.3)

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i \\ &= \beta_0 + \sum_{j=1}^k \beta_j x_{ij} + \epsilon_i \quad i = 1, 2, \dots, n \end{aligned} \quad (6.41)$$

La funzione del metodo dei minimi quadrati è

$$L = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2 \quad (6.42)$$

Vogliamo minimizzare  $L$  rispetto a  $\beta_0, \beta_1, \dots, \beta_k$ . Le **stime dei minimi quadrati** di  $\beta_0, \beta_1, \dots, \beta_k$  devono allora soddisfare le equazioni

$$\frac{\partial L}{\partial \beta_0} \Big|_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} = -2 \sum_{i=1}^n \left( y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij} \right) = 0 \quad (6.43a)$$

e

$$\frac{\partial L}{\partial \beta_j} \Big|_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} = -2 \sum_{i=1}^n \left( y_i - \hat{\beta}_0 - \sum_{j=1}^k \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \quad j = 1, 2, \dots, k \quad (6.43b)$$

Semplificando l'Equazione (6.43), otteniamo le **equazioni normali dei minimi quadrati**

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n x_{ik} &= \sum_{i=1}^n y_i \\ \hat{\beta}_0 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1} x_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n x_{i1} x_{ik} &= \sum_{i=1}^n x_{i1} y_i \\ \vdots &\quad \vdots &\quad \vdots &\quad \vdots \\ \hat{\beta}_0 \sum_{i=1}^n x_{ik} + \hat{\beta}_1 \sum_{i=1}^n x_{ik} x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{ik} x_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n x_{ik}^2 &= \sum_{i=1}^n x_{ik} y_i \end{aligned} \quad (6.44)$$

Si noti che vi sono  $p = k + 1$  equazioni normali, una per ogni coefficiente di regressione incognito. Le soluzioni delle equazioni normali daranno gli **stimatori dei minimi quadrati** dei coefficienti di regressione,  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ . Le equazioni normali possono essere risolte con un qualsiasi metodo di risoluzione applicabile a un sistema di equazioni lineari.

**ESEMPIO 6.7**  
**Resistenza  
alla trazione  
di un legame  
a filo**

Nel Capitolo 1, per illustrare la costruzione di un modello empirico, abbiamo considerato in un processo di produzione di semiconduttori la resistenza alla trazione di un legame a filo metallico, la lunghezza del filo e lo spessore della piastrina. Useremo i medesimi dati, che per comodità riproponiamo in Tabella 6.5, e mostreremo i dettagli della stima dei parametri del modello. I grafici di dispersione dei dati sono presentati nelle Figure 1.11a e 1.11b. La Figura 6.17 mostra invece una matrice composta dai grafici di dispersione bidimensionali dei dati. Questo tipo di grafico può essere d'aiuto nel visualizzare le relazioni tra le variabili in un insieme di dati a più variabili.

Useremo il modello di regressione lineare multipla

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

dove  $Y$  = resistenza alla trazione,  $x_1$  = lunghezza del filo,  $x_2$  = spessore della piastrina. Dai dati in Tabella 6.5 calcoliamo

$$\begin{aligned} n &= 25, \sum_{i=1}^{25} y_i = 725.82, \sum_{i=1}^{25} x_{i1} = 206, \sum_{i=1}^{25} x_{i2} = 8294 \\ \sum_{i=1}^{25} x_{i1}^2 &= 2396, \sum_{i=1}^{25} x_{i2}^2 = 3531848 \\ \sum_{i=1}^{25} x_{i1} x_{i2} &= 77177, \sum_{i=1}^{25} x_{i1} y_i = 8008.47, \sum_{i=1}^{25} x_{i2} y_i = 274816.71 \end{aligned}$$

Per il modello  $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ , le equazioni normali 6.44 sono

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_{i1} + \hat{\beta}_2 \sum_{i=1}^n x_{i2} = \sum_{i=1}^n y_i$$

Tabella 6.5 Dati per l'Esempio 6.7.

Osservazione Numero	Resist. traz. $y$	Lunghezza filo $x_1$	Spess. piastr. $x_2$	Osservazione Numero	Resist. traz. $y$	Lunghezza filo $x_1$	Spess. piastr. $x_2$
1	9.95	2	50	14	11.66	2	360
2	24.45	8	110	15	21.65	4	205
3	31.75	11	120	16	17.89	4	400
4	35.00	10	550	17	69.00	20	600
5	25.02	8	295	18	10.30	1	585
6	16.86	4	200	19	34.93	10	540
7	14.38	2	375	20	46.59	15	250
8	9.60	2	52	21	44.88	15	290
9	24.35	9	100	22	54.12	16	510
10	27.50	8	300	23	56.63	17	590
11	17.08	4	412	24	22.13	6	100
12	37.00	11	400	25	21.15	5	400
13	41.95	12	500				

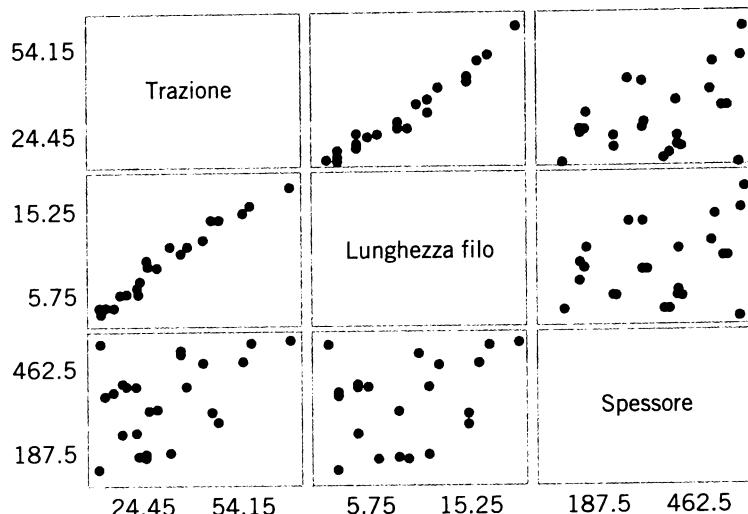


Figura 6.17 Matrice dei grafici di dispersione (prodotta con Minitab) per i dati di Tabella 6.5.

$$\hat{\beta}_0 \sum_{i=1}^n x_{i1} + \hat{\beta}_1 \sum_{i=1}^n x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n x_{i1} x_{i2} = \sum_{i=1}^n x_{i1} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^n x_{i2} + \hat{\beta}_1 \sum_{i=1}^n x_{i1} x_{i2} + \hat{\beta}_2 \sum_{i=1}^n x_{i2}^2 = \sum_{i=1}^n x_{i2} y_i$$

Inserendo le somme calcolate nelle equazioni normali otteniamo:

$$25\hat{\beta}_0 + 206\hat{\beta}_1 + 8294\hat{\beta}_2 = 725.82$$

$$206\hat{\beta}_0 + 2396\hat{\beta}_1 + 77177\hat{\beta}_2 = 8008.47$$

$$8294\hat{\beta}_0 + 77177\hat{\beta}_1 + 3531848\hat{\beta}_2 = 274816.71$$

La soluzione di questo sistema di equazioni è

$$\hat{\beta}_0 = 2.26379, \hat{\beta}_1 = 2.74427, \hat{\beta}_2 = 0.01253$$

Quindi l'equazione di regressione stimata è

$$\hat{y} = 2.26379 + 2.74427x_1 + 0.01253x_2$$

Questa equazione può essere usata per prevedere la resistenza alla trazione per coppie di valori delle variabili regressore lunghezza filo ( $x_1$ ) e spessore piastrina ( $x_2$ ). Si tratta essenzialmente del medesimo modello di regressione dell'Equazione (1.6), Paragrafo 1.3. La Figura 1.13 mostra un grafico tridimensionale del piano dei valori  $\hat{y}$  previsti, generati da questa equazione.

Per svolgere i calcoli coinvolti nei modelli di regressione multipla si usano quasi sempre applicazioni software. In Tabella 6.6 è mostrato l'output di Minitab per i dati dell'Esempio 6.7.

**Tabella 6.6** Output di Minitab per l'analisi di regressione.

Regression Analysis: Strength versus Wire Length, Die Height

The regression equation is

$$\text{Strength} = 2.26 + 2.74 \text{ Wire Ln} + 0.0125 \text{ Die Ht}$$

Predictor	Coef	SE Coef	T	P
Constant	2.264	$\leftarrow \hat{\beta}_0$	1.060	2.14
Wire Ln	2.74427	$\leftarrow \hat{\beta}_1$	0.09352	29.34
Die Ht	0.012528	$\leftarrow \hat{\beta}_2$	0.002798	4.48
$S = 2.288 \leftarrow \hat{\sigma}$		R-Sq = 98.1%		R-Sq (adj) = 97.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	5990.8 $\leftarrow SS_R$	2995.4	572.17	0.000
Residual Error	22	115.2 $\leftarrow SS_E$	5.2 $\leftarrow \hat{\sigma}^2$		
Total	24	6105.9 $\leftarrow SS_T$			

Obs	Strength	Fit	SE Fit	Residual	St Resid
1	9.950	8.379	0.907	1.571	0.75
2	24.450	25.596	0.765	-1.146	-0.53
3	31.750	33.954	0.862	-2.204	-1.04
4	35.000	36.597	0.730	-1.597	-0.74
5	25.020	27.914	0.468	-2.894	-1.29
6	16.860	15.746	0.626	1.114	0.51
7	14.380	12.450	0.786	1.930	0.90
8	9.600	8.404	0.904	1.196	0.57
9	24.350	28.215	0.819	-3.865	-1.81
10	27.500	27.976	0.465	-0.476	-0.21
11	17.080	18.402	0.696	-1.322	-0.61
12	37.000	37.462	0.525	-0.462	-0.21
13	41.950	41.459	0.655	0.491	0.22
14	11.660	12.262	0.769	-0.602	-0.28
15	21.650	15.809	0.621	5.841	2.65
16	17.890	18.252	0.679	-0.362	-0.17
17	69.000	64.666	1.165	4.334	2.20
18	10.300	12.337	1.238	-2.037	-1.06
19	34.930	36.472	0.710	-1.542	-0.71
20	46.590	46.560	0.878	0.030	0.01
21	44.880	47.061	0.824	-2.181	-1.02
22	54.120	52.561	0.843	1.559	0.73
23	56.630	56.308	0.977	0.322	0.16
24	22.130	19.982	0.756	2.148	0.99
25	21.150	20.996	0.618	0.154	0.07

Nell'output di Minitab di Tabella 6.6 abbiamo  $x_1$  = lunghezza filo e  $x_2$  = spessore piastrina. Le stime dei coefficienti di regressione  $\hat{\beta}_0$ ,  $\hat{\beta}_1$  e  $\hat{\beta}_2$  sono evidenziate con una freccia. I valo-

ri stimati  $\hat{y}_i$ , corrispondenti a ciascuna osservazione, sono elencati nella colonna con l'intestazione "Fit". Sono inoltre elencati i residui calcolati con  $e_i = y_i - \hat{y}_i$ ,  $i = 1, 2, \dots, 25$ .

Molte delle procedure di calcolo e di analisi che abbiamo introdotto per la regressione lineare semplice si possono applicare anche al caso della regressione multipla. Per esempio, la somma dei quadrati dei residui viene usata per stimare la varianza dell'errore  $\sigma^2$ . Tale somma è  $SS_E = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ , mentre la stima di  $\sigma^2$  per un modello di regressione lineare multipla con  $p$  parametri è

### Stima della varianza

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p} = \frac{SS_E}{n - p} \quad (6.45)$$

Le Equazioni (6.43a) e (6.43b) mostrano che i residui soddisfano  $p$  equazioni. Di conseguenza,  $p$  residui possono venire calcolati dai rimanenti  $n - p$  residui. Ecco perché nell'Equazione (6.45) compare al denominatore l'espressione  $n - p$ .

**ESEMPIO 6.7**  
(proseguo)  
Stima della varianza  
dell'errore (tramite  
software).

L'output di Minitab in Tabella 6.6 mostra che la somma dei quadrati dei residui per il modello di regressione per la resistenza alla trazione è  $SS_E = 115.2$ , che ci sono  $n = 25$  osservazioni, e che il modello ha  $p = 3$  parametri ( $\beta_0$ ,  $\beta_1$  e  $\beta_2$ ). Pertanto, dall'Equazione (6.45) risulta che la stima della varianza dell'errore è  $\hat{\sigma}^2 = SS_E/n - p = 115.2/(25 - 3) = 5.2$ , come riportato da Minitab in Tabella 6.6.

La scomposizione, nell'analisi della varianza, della somma totale dei quadrati (Equazione (6.21)) è valida anche per la regressione multipla. L'output di Minitab in Tabella 6.6 contiene i risultati di ANOVA; la somma totale dei quadrati ha  $n - 1$  gradi di libertà, la somma dei quadrati di regressione o del modello ha  $k = p - 1$  gradi di libertà (ricordiamo che  $k$  è il numero dei regressori), e la somma dei quadrati dei residui ha  $n - p$  gradi di libertà.

Il coefficiente di determinazione  $R^2$ , nella regressione multipla, viene calcolato esattamente come nella regressione lineare semplice:  $R^2 = SS_R/SS_T = 1 - (SS_E/SS_T)$ . In un modello di regressione lineare multipla si è soliti chiamare  $R^2$  coefficiente di determinazione *multipla*. Per il modello di regressione riguardante la resistenza alla trazione, Minitab calcola  $R^2 = 1 - (115.2/6105.9) = 0.981$ , e l'output riportato è  $R^2 \times 100\% = 98.1\%$ . Questo viene interpretato come indicazione del fatto che il modello contenente la lunghezza del filo e lo spessore della piastrina rende conto di circa il 98.1% della variabilità osservata nella resistenza alla trazione.

Il valore numerico di  $R^2$  non può diminuire con l'aggiunta di variabili al modello di regressione. Per esempio, se per i dati riguardanti la resistenza del legame a filo usassimo solo il regressore  $x_1$  = lunghezza del filo, avremmo  $R^2 = 0.964$ ; aggiungendo il secondo regressore  $x_2$  = spessore della piastrina, il valore di  $R^2$  diventa 0.981. Per valutare meglio l'aggiunta di un altro regressore al modello si può usare la statistica  **$R^2$  corretto**.

Calcolo e  
interpretazione di  $R^2$ .

### Coefficiente di determinazione multipla corretto ( $R^2_{\text{Adjusted}}$ )

Il **coefficiente di determinazione multipla corretto (adjusted)** per un modello di regressione multipla con  $k$  regressori è

$$R^2_{\text{Adjusted}} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} = \frac{(n-1)R^2 - k}{n-p} \quad (6.46)$$

La statistica  $R^2$  corretto, in sostanza, abbassa la statistica  $R^2$  usuale prendendo in considerazione il numero di regressori nel modello. In generale, la statistica  $R^2$  corretto non aumenterà sempre all'aggiunta di una variabile al modello; lo farà solo se tale aggiunta produce una sufficiente riduzione della somma dei quadrati dei residui per compensare la perdita di un grado di libertà.

#### ESEMPIO 6.7 (proseguo)

Per illustrare questo punto, consideriamo il modello di regressione per i dati di resistenza alla trazione del filo con la sola variabile regressore  $x_1$  = lunghezza del filo. Il valore della somma dei quadrati dei residui per questo modello è  $SS_E = 220.09$ . Dall'Equazione (6.45), risulta che la statistica  $R^2$  corretto è

$$R^2_{\text{Adjusted}} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} = 1 - \frac{220.09/(25-2)}{6105.9/(25-1)} = 0.962$$

Quando si includono entrambe le variabili regressore nel modello, il valore di  $R^2_{\text{Adjusted}}$  diventa 0.979 (si veda la Tabella 6.6). Poiché la statistica  $R^2$  corretto è aumentata con l'aggiunta di una seconda variabile regressore, potremmo concludere che aggiungere al modello questa nuova variabile è stata probabilmente una buona idea, visto che il modello riesce ora a spiegare maggiormente la variabilità della risposta.

Un altro modo per valutare il contributo dovuto all'aggiunta di un regressore al modello consiste nell'esaminare la variazione della media dei quadrati dei residui. Per i dati relativi alla resistenza alla trazione del legame a filo con solo  $x_1$  = lunghezza del filo come regressore, la somma dei quadrati dei residui è  $SS_E = 220.09$ , e la media dei quadrati dei residui è  $SS_E/(n-p) = 9.57$ . Con entrambi i regressori compresi nel modello la media dei quadrati dei residui vale 5.2. Poiché la media dei quadrati dei residui dà una stima di  $\sigma^2$  (la varianza della variabilità non spiegata nella risposta), concludiamo che il modello con due regressori è superiore. Si noti che usando la media dei quadrati dei residui come stima di  $\sigma^2$  si ottiene una stima **modello-dipendente**. Comunque, un modello di regressione con un piccolo valore di tale grandezza è quasi sempre superiore a un modello con un valore elevato.

#### Contributo di un secondo regressore.

#### Calcolo e interpretazione di $R^2_{\text{Adjusted}}$ .

#### 6.3.2 Inferenze nella regressione multipla

Come per la regressione lineare semplice, anche nella regressione multipla è importante verificare ipotesi e costruire intervalli di confidenza. In questo paragrafo descriveremo queste procedure; nella maggior parte dei casi si tratta di modifiche dirette delle procedure usate per la regressione lineare semplice.

### Test per la significatività della regressione

Il test per la significatività della regressione nella regressione lineare semplice controllava se vi era una relazione lineare utile tra la risposta  $y$  e il singolo regressore  $x$ . Nella regressione multipla, questo test controlla l'ipotesi che non vi sia alcuna relazione lineare utile tra la risposta  $y$  e *nessuno* dei regressori  $x_1, x_2, \dots, x_k$ . Le ipotesi sono dunque

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \text{almeno un } \beta_j \neq 0$$

### Verifica della significatività della regressione nella regressione multipla

$$MS_R = \frac{SS_R}{k} \quad MS_E = \frac{SS_E}{n - p}$$

Ipotesi nulla:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

Ipotesi alternativa:

$$H_1: \text{almeno un } \beta_j \neq 0$$

Statistica test:

$$F_0 = \frac{MS_R}{MS_E} \quad (6.47)$$

P-value:

Probabilità a destra di  $f_0$  nella distribuzione  $F_{k,n-p}$

Criterio di rifiuto per un  
test con livello fissato:

$$f_0 > f_{\alpha,k,n-p}$$

Perciò, rifiutare l'ipotesi nulla significa che almeno una delle variabili regressore del modello è in relazione lineare con la risposta. Per verificare queste ipotesi si usa la scomposizione ANOVA della variabilità totale della risposta  $y$  (Equazione (6.21)).

### ESEMPIO 6.7 (proseguo)

La procedura del test viene di solito riassunta in una tabella ANOVA ed è prevista nell'output dei programmi di regressione multipla. Per il modello di regressione della resistenza alla trazione (Tabella 6.6), le ipotesi sono

$$H_0: \beta_1 = \beta_2 = 0$$

$$H_1: \text{almeno un } \beta_j \neq 0$$

Nell'output di Minitab presentato nella sezione intitolata "Analisi della varianza" troviamo i valori della media dei quadrati per la regressione e per il residuo; il valore calcolato della statistica test dall'Equazione (6.46) è  $f_0 = 572.17$ . Essendo il P-value molto piccolo, concluderemmo che almeno uno dei regressori è in relazione con la risposta  $y$ .

Questo test è un primo passo; poiché  $H_0$  è rifiutata, l'interesse si concentra ora sui singoli coefficienti di regressione.

### Inferenza sui singoli coefficienti di regressione

Poiché le stime dei coefficienti di regressione,  $\hat{\beta}_j, j = 0, 1, \dots, k$ , in un modello di regressione multipla sono semplicemente combinazioni lineari delle  $y$ , e per queste ultime si è

### Interpretazione della tabella ANOVA.

assunta una distribuzione normale, i coefficienti  $\hat{\beta}_j$  sono distribuiti normalmente; sono inoltre stimatori non distorti dei veri coefficienti del modello, e i loro errori standard,  $se(\hat{\beta}_j)$ ,  $j = 0, 1, \dots, k$ , possono venire calcolati come il prodotto di  $\hat{\sigma}$  e di una funzione delle  $x$ . L'errore standard è dato da un'espressione piuttosto complicata, ma viene calcolato da tutti i software di regressione multipla. Nell'output di Minitab di Tabella 6.6 gli errori standard dei coefficienti di regressione del modello sono elencati nella colonna intestata "SE Coef".

Le inferenze su un singolo coefficiente di regressione si basano sulla quantità

$$T = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}$$

che ha la distribuzione  $t$  con  $n - p$  gradi di libertà. Ciò porta alle seguenti verifiche di ipotesi e ai seguenti intervalli di confidenza per un singolo coefficiente di regressione  $\beta_j$ .

#### Inferenze sui parametri del modello nella regressione multipla

1. La verifica di  $H_0: \beta_j = \beta_{j,0}$  contro  $H_1: \beta_j \neq \beta_{j,0}$  impiega la **statistica test**

$$T_0 = \frac{\hat{\beta}_j - \beta_{j,0}}{se(\hat{\beta}_j)} \quad (6.48)$$

e l'ipotesi nulla è rifiutata se  $|t_0| > t_{\alpha/2, n-p}$ . Si possono verificare anche ipotesi alternative unilaterali.

2. Un intervallo di confidenza di livello  $100(1 - \alpha)\%$  per un singolo coefficiente di regressione è dato da

$$\hat{\beta}_j - t_{\alpha/2, n-p} se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2, n-p} se(\hat{\beta}_j) \quad (6.49)$$

Un caso speciale molto importante di test su un singolo coefficiente di regressione è quello che prevede un'ipotesi della forma  $H_0: \beta_j = 0$  contro  $H_1: \beta_j \neq 0$ . La maggior parte dei software di regressione calcola la statistica test per questa ipotesi per ciascuna variabile presente nel modello. Si tratta di una misura del contributo apportato da ciascuna *singola* variabile regressore al modello complessivo. La statistica test è

$$T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \quad (6.50)$$

Tale verifica viene spesso chiamata test **parziale** o **marginale** perché valuta il contributo che ogni variabile regressore offre al modello *supponendo* che tutti gli *altri* regressori siano anch'essi inclusi.

**ESEMPIO 6.7  
(proseuzione)**

**Interpretazione dei valori della statistica  $t$ .**

L'output di Minitab riportato in Tabella 6.6 mostra i valori della statistica test calcolati dall'Equazione (6.50) per ciascuno dei regressori: lunghezza del filo e spessore della piastrina. Il valore della statistica  $t$  per la lunghezza del filo è  $t_0 = 29.34$ , e ciò indica che il regressore lunghezza del filo contribuisce in modo significativo al modello, posto che l'altro regressore, lo spessore della piastrina, sia anch'esso incluso nel modello. Il valore della statistica  $t$  per lo spessore della piastrina è  $t_0 = 4.48$ , e ciò indica che questo regressore contribuisce in modo significativo al modello, posto che l'altro regressore, la lunghezza del filo, sia anch'esso incluso nel modello. In genere, se la statistica  $t$  per qualche singolo coefficiente di regressione non è significativa, cioè se l'ipotesi  $H_0: \beta_j = 0$  non viene rifiutata, si ha una indicazione del fatto che il regressore  $x_j$  dovrebbe essere rimosso dal modello. In alcune situazioni, questi test  $t$  possono indicare che più regressori non sono importanti. L'approccio corretto, allora, consiste nel rimuovere il regressore meno significativo e ricalcolare tutte le stime; successivamente, si eseguono di nuovo i test  $t$  per i regressori del nuovo modello così ottenuto, in modo da stabilire quali regressori sono eventualmente ancora non significativi. Si procede in questo modo fino a che tutti i regressori non risultano significativi.

### Intervalli di confidenza sulla risposta media e intervalli di predizione

Un modello di regressione multipla è spesso usato per ottenere una **stima puntuale** della risposta media per un particolare insieme delle variabili  $x$ , che possiamo indicare con  $x_1 = x_{10}$ ,  $x_2 = x_{20}$ , ...,  $x_k = x_{k0}$ . La risposta media reale in tale punto è  $\hat{\mu}_{Y|x_{10},x_{20},\dots,x_{k0}} = \beta_0 + \beta_1 x_{10} + \beta_2 x_{20} + \dots + \beta_k x_{k0}$ , e la corrispondente stima puntuale è

$$\hat{\mu}_{Y|x_{10},x_{20},\dots,x_{k0}} = \hat{\beta}_0 + \hat{\beta}_1 x_{10} + \hat{\beta}_2 x_{20} + \dots + \hat{\beta}_k x_{k0} \quad (6.51)$$

L'errore standard di questa stima puntuale è una funzione complicata, che dipende dalle  $x$  usate per stimare il modello di regressione e dalle coordinate in cui viene calcolata la stima. Esistono comunque molti pacchetti software per la regressione che forniscono tale funzione, indicata con  $se(\hat{\mu}_{Y|x_{10},x_{20},\dots,x_{k0}})$ . L'intervalllo di confidenza per la risposta media nel punto  $x_1 = x_{10}$ ,  $x_2 = x_{20}$ , ...,  $x_k = x_{k0}$  è dato dalla seguente espressione.

#### Intervallo di confidenza per la risposta media nella regressione multipla

Un intervallo di confidenza di livello  $100(1 - \alpha)\%$  per la risposta media nel punto  $x_1 = x_{10}$ ,  $x_2 = x_{20}$ , ...,  $x_k = x_{k0}$  in un modello di regressione multipla è dato da

$$\hat{\mu}_{Y|x_{10},x_{20},\dots,x_{k0}} - t_{\alpha/2,n-p} se(\hat{\mu}_{Y|x_{10},x_{20},\dots,x_{k0}}) \leq \mu_{Y|x_{10},x_{20},\dots,x_{k0}} \leq \hat{\mu}_{Y|x_{10},x_{20},\dots,x_{k0}} + t_{\alpha/2,n-p} se(\hat{\mu}_{Y|x_{10},x_{20},\dots,x_{k0}}) \quad (6.52)$$

dove  $\hat{\mu}_{Y|x_{10},x_{20},\dots,x_{k0}}$  è calcolato mediante l'Equazione (6.51).

**ESEMPIO 6.7  
(proseuzione)**

Minitab calcolerà l'intervalllo di confidenza dell'Equazione (6.51) per un punto di interesse. Per esempio, si supponga di voler trovare una stima della resistenza media alla trazione in due punti: (1) il punto in cui la lunghezza del filo,  $x_1$ , vale 11 e lo spessore della piastrina,  $x_2$ , vale 35; (2) il punto in cui la lunghezza del filo,  $x_1$ , vale 5 e lo spessore della piastrina,  $x_2$ , vale

**Interpretazione della risposta media e CI al 95%.**

20. Si trova la stima puntuale sostituendo dapprima  $x_1 = 11$  e  $x_2 = 35$ , quindi  $x_1 = 5$  e  $x_2 = 20$  nel modello di regressione stimato, e calcolando infine i valori stimati della risposta nei due punti. I risultati prodotti da Minitab (stima puntuale e intervalli di confidenza al 95% associati) sono riportati in Tabella 6.7.

**Tabella 6.7** Output di Minitab.

Predicted Values for New Observations				
New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	32.889	1.062	(30.687, 35.092)	(27.658, 38.121)
Values of Predictors for New Observations				
New Obs	Wire Ln	Die Ht		
1	11.0	35.0		
New Obs	Wire Ln	Die Ht		
2	5.00	20.0		

La risposta media stimata  $\hat{\mu}_{y|11,35} = 32.899$  ha un intervallo di confidenza al 95% uguale a (30.687, 35.092), mentre la risposta media stimata  $\hat{\mu}_{y|5,20} = 16.236$  ha un intervallo di confidenza al 95% uguale a (14.310, 18.161).

Si noti che l'intervallo di confidenza al 95% per il secondo punto è più stretto dell'intervallo al 95% per il primo punto. Come nella regressione lineare semplice, l'ampiezza dell'intervallo di confidenza per la risposta media aumenta all'allontanarsi del punto dal centro della regione spazzata dalle variabili  $x$ , e in questo senso il punto 1 è più lontano del punto 2 dal centro di tale regione.

In un modello di regressione multipla è possibile anche determinare un intervallo di predizione al  $100(1 - \alpha)\%$  per un'osservazione futura nel punto  $x_1 = x_{10}, x_2 = x_{20}, \dots, x_k = x_{k0}$ . La risposta nel punto di interesse è

$$Y_0 = \beta_0 + \beta_1 x_{10} + \beta_2 x_{20} + \cdots + \beta_k x_{k0} + \epsilon$$

e il corrispondente valore predetto è

$$\hat{Y}_0 = \hat{\mu}_{Y|x_{10}, x_{20}, \dots, x_{k0}} = \hat{\beta}_0 + \hat{\beta}_1 x_{10} + \hat{\beta}_2 x_{20} + \cdots + \hat{\beta}_k x_{k0} \quad (6.53)$$

L'errore di predizione è  $Y_0 - \hat{Y}_0$ , e la sua deviazione standard è

$$\sqrt{\hat{\sigma}^2 + [se(\hat{\mu}_{Y|x_{10}, x_{20}, \dots, x_{k0}})]^2}$$

Pertanto, per l'intervallo di predizione su un'osservazione futura vale quanto riassunto di seguito.

### Intervallo di predizione su un'osservazione futura

Un intervallo di predizione (PI) di livello  $100(1 - \alpha)\%$  su un'osservazione futura nel punto  $x_1 = x_{10}, x_2 = x_{20}, \dots, x_k = x_{k0}$  in un modello di regressione multipla è dato da

$$\begin{aligned}\hat{y}_0 - t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 + [se(\hat{\mu}_{Y|x_{10}, x_{20}, \dots, x_{k0}})]^2} &\leq Y_0 \\ \hat{y}_0 + t_{\alpha/2, n-p} \sqrt{\hat{\sigma}^2 + [se(\hat{\mu}_{Y|x_{10}, x_{20}, \dots, x_{k0}})]^2} &\end{aligned}\quad (6.54)$$

dove  $\hat{y}_0 = \hat{\mu}_{Y|x_{10}, x_{20}, \dots, x_{k0}}$  è calcolato mediante l'Equazione (6.53).

**ESEMPIO 6.7  
(proseguimento)**

**Interpretazione  
di un'osservazione  
futura e PI al 95%.**

L'output di Minitab in Tabella 6.7 mostra gli intervalli di predizione al 95% per la resistenza alla trazione in due nuovi punti, nei quali si ha rispettivamente:  $x_1 = 11, x_2 = 35$  e  $x_1 = 5, x_2 = 20$ . L'osservazione futura  $\hat{y}_0 = \hat{\mu}_{y|11,35} = 32.889$  ha un intervallo di predizione al 95% uguale a (27.658, 38.121), mentre l'osservazione futura  $\hat{y}_0 = \hat{\mu}_{y|5,20} = 16.236$  ha un intervallo di predizione al 95% uguale a (11.115, 21.357). Si noti che questi intervalli di predizione sono più ampi dei corrispondenti intervalli di confidenza e che la loro ampiezza cresce all'allontanarsi del punto in cui viene effettuata la predizione dal centro della regione delle  $x$ .

#### Un test per la significatività di un gruppo di regressori

Nella costruzione di un modello di regressione esistono alcune situazioni in cui l'interesse è centrato su un sottoinsieme di regressori del modello completo. Per esempio, si consideri un modello del secondo ordine

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$$

per il quale siamo incerti sul contributo apportato al modello da parte dei termini di secondo ordine. Siamo perciò interessati a verificare l'ipotesi

$$H_0: \beta_{12} = \beta_{11} = \beta_{22} = 0$$

$$H_1: \text{almeno uno dei } \beta \neq 0$$

Per verificare queste ipotesi possiamo usare un test  $F$ .

In generale, supposto che il **modello completo** abbia  $k$  regressori, siamo interessati a stabilire se gli ultimi  $k - r$  regressori possono essere eliminati dal modello. L'eventuale modello più piccolo che si ottiene viene chiamato **modello ridotto**. Se supponiamo che il modello completo sia

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r + \beta_{r+1} x_{r+1} + \cdots + \beta_k x_k + \epsilon$$

e che la riduzione del modello comporti  $\beta_{r+1} = \beta_{r+2} = \cdots = \beta_k = 0$ , il modello ridotto risulta

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_r x_r + \epsilon$$

Il test viene eseguito stimando sia il modello completo sia quello ridotto e confrontando le somme dei quadrati dei residui per i due modelli. Sia  $SS_E(FM)$  la somma dei quadrati dei residui per il modello completo ( $FM = full\ model$ ) e  $SS_E(RM)$  quella per il modello ridotto ( $RM = reduced\ model$ ). Allora, per verificare le ipotesi

$$\begin{aligned} H_0: \beta_{r+1} &= \beta_{r+2} = \cdots = \beta_k = 0 \\ H_1: \text{almeno uno dei } \beta &\neq 0 \end{aligned} \quad (6.55)$$

dovremmo usare la statistica test

#### Statistica test per la significatività di un gruppo di regressori

$$F_0 = \frac{[SS_E(RM) - SS_E(FM)]/(k - r)}{SS_E(FM)/(n - p)} \quad (6.56)$$

L'ipotesi nulla nell'Equazione (6.55) viene rifiutata se  $f_0 > f_{a,k-r,n-p}$ . Si può usare anche un approccio basato sul  $P$ -value.

### 6.3.3 Controllo dell'adeguatezza del modello

#### Analisi dei residui

Uno dei metodi per controllare l'adeguatezza di un modello di regressione lineare multipla è l'analisi grafica dei residui  $e_i = y_i - \hat{y}_i$ ,  $i = 1, 2, \dots, n$ . Per controllare l'assunzione di normalità si impiega un grafico dei quantili per i residui, mentre i grafici dei residui in funzione dei valori stimati, e possibilmente in funzione di ciascuno dei singoli regressori, possono rivelare altre inadeguatezze del modello, come una varianza non costante e l'eventuale necessità di aggiungere al modello ulteriori regressori.

#### ESEMPIO 6.7 (proseguo)

##### Interpretazione dei grafici dei residui.

La Figura 6.18 presenta il grafico dei quantili per i residui, le Figure 6.19, 6.20 e 6.21 i grafici dei residui in funzione dei valori stimati  $\hat{y}$  e di ciascun regressore  $x_1$  e  $x_2$  per il modello di regressione relativo alla resistenza alla trazione. Il grafico dei quantili è soddisfacente, ma i grafici dei residui in funzione di  $\hat{y}$  e di  $x_1$  rivelano una leggera curvatura, che suggerisce la necessità di aggiungere al modello un'altra variabile regressore. In generale, comunque, nessuno dei grafici evidenzia seri problemi per quanto riguarda il modello.

Nella regressione multipla, spesso si prendono in esame residui riscalati. Un comune residuo riscalato è il **residuo standardizzato**,  $d_i = e_i / \sqrt{\hat{\sigma}^2}$ ,  $i = 1, 2, \dots, n$ . Abbiamo discusso il residuo standardizzato nella regressione lineare semplice e abbiamo osservato che esso può risultare utile nella ricerca degli outlier. Un altro residuo riscalato è il **residuo studentizzato**, molto utile nella regressione multipla. Il residuo studentizzato si ottiene a partire dal consueto residuo dei minimi quadrati dividendo quest'ultimo per il suo esatto errore standard.

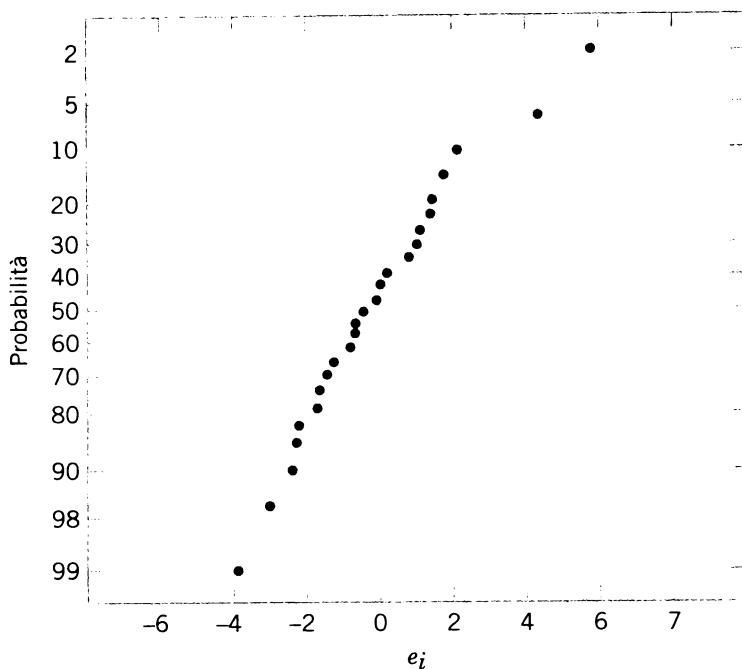
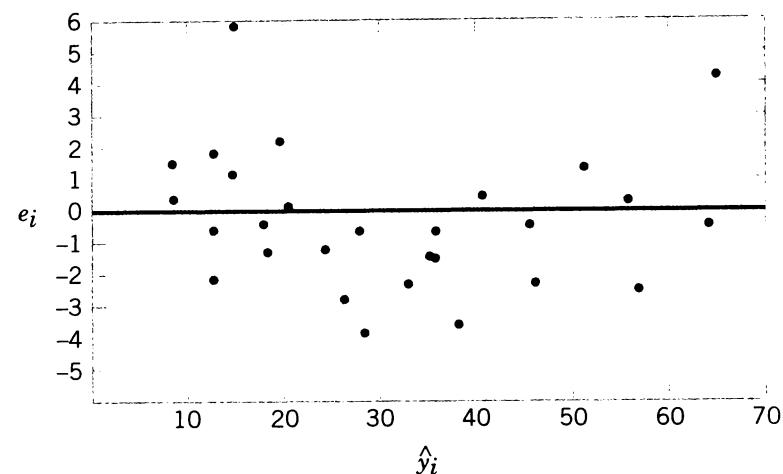


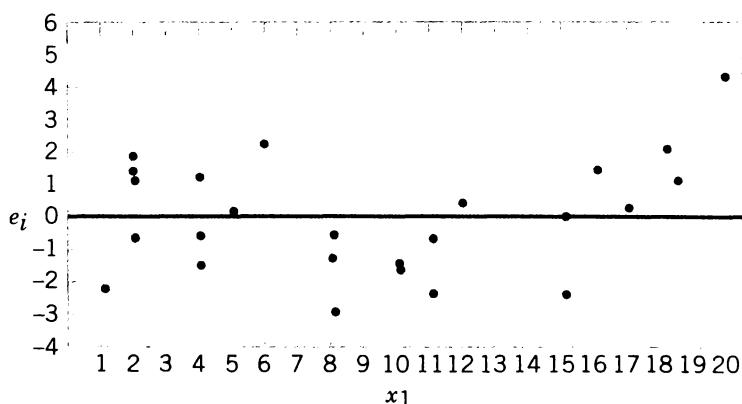
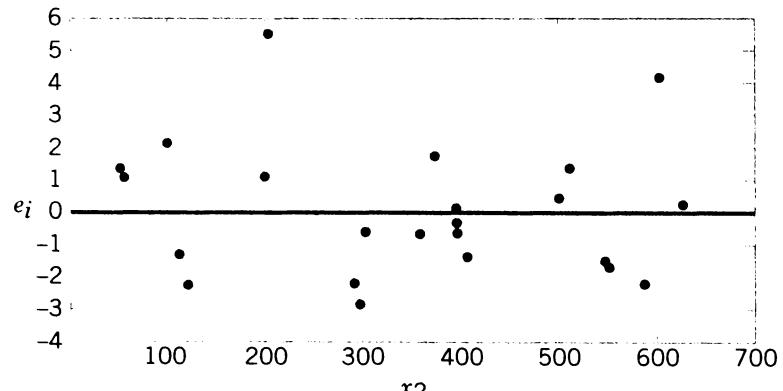
Figura 6.18 Grafico dei quantili per i residui.

Figura 6.19 Grafico dei residui in funzione dei valori  $\hat{y}_i$ .

Mostriamo ora come calcolare i residui studentizzati. I coefficienti di regressione sono combinazioni lineari delle osservazioni  $y$ . Poiché i valori predetti  $\hat{y}$  sono combinazioni lineari dei coefficienti di regressione, sono anche combinazioni lineari delle osservazioni  $y_i$ . Possiamo dunque scrivere le relazioni tra i valori  $\hat{y}_i$  e i valori  $y_i$  come segue

$$\begin{aligned}\hat{y}_1 &= h_{11}y_1 + h_{12}y_2 + \cdots + h_{1n}y_n \\ \hat{y}_2 &= h_{21}y_1 + h_{22}y_2 + \cdots + h_{2n}y_n \\ &\vdots \\ \hat{y}_n &= h_{n1}y_1 + h_{n2}y_2 + \cdots + h_{nn}y_n\end{aligned}\tag{6.57}$$

I termini  $h_{ij}$  sono funzioni delle sole variabili  $x$  usate per stimare il modello, e sono effettivamente molto semplici da calcolare [per i dettagli cfr. Montgomery, Peck, Vining (2006)]. Inoltre, si può dimostrare che  $h_{ij} = h_{ji}$  e che i termini diagonali  $h_{ii}$  del sistema di equazioni sono tali che  $0 < h_{ii} \leq 1$ . Passiamo quindi a definire i residui studentizzati.

Figura 6.20 Grafico dei residui in funzione di  $x_1$ .Figura 6.21 Grafico dei residui in funzione di  $x_2$ .

### Residui studentizzati

I residui studentizzati sono definiti da

$$r_i = \frac{e_i}{se(e_i)} = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}, i = 1, 2, \dots, n \quad (6.58)$$

Essendo i termini  $h_{ii}$  sempre compresi tra 0 e 1, un residuo studentizzato è sempre maggiore del corrispondente residuo standardizzato; di conseguenza, è più adatto a diagnosticare la presenza di outlier.

**ESEMPIO 6.7**  
(prosecuzione)

Interpretazione  
dei residui studentizzati.

L'output di Minitab riportato in Tabella 6.6 elenca i residui studentizzati (nella colonna intestata "St Resid") per il modello di regressione relativo ai dati di resistenza alla trazione. Nessuno di questi residui studentizzati è abbastanza elevato da indicare la presenza di outlier.

#### Osservazioni influenti

Usando la regressione multipla, si può scoprire che alcuni sottoinsiemi delle osservazioni sono influenti in maniera insolita. Alcune volte queste osservazioni sono relativamente lontane dal resto dei dati. In Figura 6.22 è rappresentata un'ipotetica situazione per due variabili: si nota che un'osservazione nella regione delle  $x$  è lontana dal resto dei dati. La disposizione dei punti nella regione delle  $x$  si rivela essenziale nel determinare le proprietà del modello. Per esempio, il punto  $(x_{i1}, x_{i2})$  in Figura 6.22 può essere molto influente nel determinare  $R^2$ , le stime dei coefficienti di regressione e la grandezza dell'errore quadratico medio.

Vogliamo ora esaminare i punti influenti per stabilire se da essi dipendono molte proprietà del modello. Se questi punti influenti sono "cattivi" punti, o in qualche modo punti erronei, vanno eliminati. D'altra parte, potrebbe non esservi nulla di sbagliato in questi pun-



Figura 6.22 Punto lontano nella regione delle  $x$ .

ti, ma è utile in ogni caso stabilire almeno se questi punti producono o meno risultati consistenti con il resto dei dati: anche se un punto influente è un punto valido, dobbiamo infatti sapere se controlla proprietà importanti del modello, perché ciò potrebbe avere ripercussioni sull'uso del modello stesso.

A tale scopo, è utile esaminare i termini  $h_{ii}$  definiti nell'Equazione (6.57). Il valore di  $h_{ii}$  può essere interpretato come una misura della distanza del punto  $(x_{i1}, x_{i2}, \dots, x_{ik})$  dalla media di tutti i punti appartenenti all'insieme dei dati. Il valore di  $h_{ii}$  non è l'usuale distanza, ma possiede proprietà simili; di conseguenza, un valore di  $h_{ii}$  elevato implica che il punto dell' $i$ -esimo dato è lontano dal centro dei dati (come in Figura 6.22). Una regola pratica consiste nell'esaminare se il valore di  $h_{ii}$  è maggiore di  $2p/n$ . Un punto per cui  $h_{ii}$  supera questo valore viene detto **punto di leva**. Il motivo di tale denominazione è dovuto al fatto che, essendo un punto lontano, ha un'alta capacità di modificare l'analisi della regressione, come se "facesse leva". Qualunque sia l'insieme dei dati, il valor medio di  $h_{ii}$  è  $p/n$ , perciò la regola evidenzia i valori che superano il doppio della media.

Montgomery, Peck, Vining (2006) e Myers (1990) descrivono diversi altri metodi per rilevare osservazioni influenti. Un'eccellente strumento diagnostico è rappresentato dalla misura della **distanza di Cook**. Si tratta di una misura della distanza al quadrato tra l'usuale stima dei minimi quadrati ( $\hat{\beta}$ ) basata su tutte le  $n$  osservazioni e la stima che si ottiene eliminando l' $i$ -esimo punto.

### Misura della distanza di Cook

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{(1 - h_{ii})} \quad i = 1, 2, \dots, n \quad (6.59)$$

Chiaramente, se il punto  $i$ -esimo è influente, la sua rimozione comporterà per alcuni coefficienti di regressione una notevole variazione rispetto al valore assunto allorché si usano tutte le  $n$  osservazioni. Perciò, un valore elevato di  $D_i$  implica che l' $i$ -esimo punto è influente. Osserviamo che  $D_i$  contiene il residuo studentizzato al quadrato, che indica la bontà dell'adattamento del modello all' $i$ -esima osservazione  $y_i$  [si ricordi che  $r_i = e_i / \sqrt{\hat{\sigma}^2(1 - h_{ii})}$  nell'Equazione (6.57)] e un fattore che misura quanto è lontano il punto dal resto dei dati [ $h_{ii}/(1 - h_{ii})$  è una misura della distanza dell' $i$ -esimo punto nella regione delle  $x$  dal centroide degli  $n - 1$  punti rimanenti]. Un valore di  $D_i > 1$  indicherebbe che il punto è influente. Ciascun fattore (o entrambi) può contribuire a produrre un elevato valore di  $D_i$ .

#### ESEMPIO 6.8 Resistenza alla trazione di un legame a filo

In Tabella 6.8 sono elencati i valori di  $h_{ii}$  e le misure delle distanze di Cook  $D_i$  per i dati relativi alla resistenza alla trazione. Per illustrare i calcoli, consideriamo l'osservazione 1

$$\begin{aligned} D_1 &= \frac{r_1^2}{p} \cdot \frac{h_{11}}{(1 - h_{11})} = -\frac{[e_1 / \sqrt{\hat{\sigma}^2(1 - h_{11})}]^2}{p} \cdot \frac{h_{11}}{(1 - h_{11})} \\ &= \frac{[1.571 / \sqrt{5.2(1 - 0.1573)}]^2}{3} \cdot \frac{0.1573}{(1 - 0.1573)} = 0.035 \end{aligned}$$

**Interpretazione dell'influenza.**

Il doppio della media dei valori  $h_{ii}$  è  $2p/n = 2(3)/25 = 0.2400$ . Due dati (17 e 18) hanno valori di  $h_{ii}$  che superano questa soglia, e potrebbero essere classificati come punti di leva. Tuttavia, siccome i valori di  $D_{17}$  e  $D_{18}$  sono minori di 1, questi punti non sono influenti in maniera inusuale. Si noti che nessuno dei valori  $D_i$  è abbastanza elevato da richiedere attenzione.

**Tabella 6.8** Rilevamento dell'influenza delle osservazioni per l'esempio della resistenza alla trazione.

Osservazioni		Distanza di Cook		Osservazioni		Distanza di Cook	
<i>i</i>	$h_{ii}$		$D_i$	<i>i</i>	$h_{ii}$		$D_i$
1	0.1573		0.035	14	0.1129		0.003
2	0.1116		0.012	15	0.0737		0.187
3	0.1419		0.060	16	0.0879		0.001
4	0.1019		0.021	17	0.2593		0.565
5	0.0418		0.024	18	0.2929		0.155
6	0.0749		0.007	19	0.0962		0.018
7	0.1181		0.036	20	0.1473		0.000
8	0.1561		0.020	21	0.1296		0.052
9	0.1280		0.160	22	0.1358		0.028
10	0.0413		0.001	23	0.1824		0.002
11	0.0925		0.013	24	0.1091		0.040
12	0.0526		0.001	25	0.0729		0.000
13	0.0820		0.001				

**Multicollinearità**

Nei problemi di regressione multipla ci si aspetta di trovare una dipendenza tra le variabili regressore e la risposta. Tuttavia, in molti problemi di regressione, si trovano anche dipendenze tra i regressori. Quando queste dipendenze sono forti si dice che esiste una **multicollinearità**. La multicollinearità può avere seri effetti sulle stime dei parametri di un modello di regressione, portando a parametri stimati male (varianza o errore standard elevati) e instabili, nel senso che un diverso campione del medesimo processo può produrre stime molto differenti dei coefficienti  $\beta$ . Modelli con forte multicollinearità danno spesso luogo a equazioni di predizione non affidabili.

Esistono diversi strumenti diagnostici in grado di stabilire la presenza o meno di multicollinearità; il più semplice è il **fattore di inflazione della varianza** (VIF, *Variance Inflation Factor*).

**Fattori di inflazione della varianza**

I VIF per un modello di regressione lineare multipla sono dati da

$$VIF(\beta_j) = \frac{1}{1 - R_j^2}, \quad j = 1, 2, \dots, k \quad (6.60)$$

dove  $R_j^2$  è il coefficiente di determinazione multipla che risulta dalla regressione di  $x_j$  sugli altri  $k - 1$  regressori.

È facile comprendere perché il VIF definito dall'Equazione (6.60) è una buona misura della multicollinearità. Se il regressore  $x_j$  ha una forte dipendenza lineare da qualche sottoinsieme degli altri regressori del modello,  $R_j^2$  sarà elevato, per esempio prossimo a 1, e il corrispondente VIF sarà grande. Per contro, se il regressore  $x_j$  non è quasi linearmente dipendente da altri regressori, il valore di  $R_j^2$  sarà piccolo, così come il corrispondente VIF. In generale, se il VIF associato a un regressore è maggiore di 10, dobbiamo sospettare la presenza di multicollinearità.

### ESEMPIO 6.8 (proseguimento)

#### Interpretazione dei VIF.

Molti software per la regressione sono in grado di calcolare i VIF. Il seguente riquadro mostra i risultati di Minitab per il modello di regressione relativo alla resistenza alla trazione.

Predictor	Coef	SE Coef	T	P	VIF
Constant	2.264	1.060	2.14	0.044	
Wire Ln	2.74427	0.09352	29.34	0.000	1.2
Die Ht	0.01258	0.002798	4.48	0.000	1.2

Poiché i VIF sono in questo caso piuttosto bassi, non esiste un evidente problema di multicollinearità nei dati.

Nei casi in cui è presente una forte multicollinearità, molti analisti raccomandano di prendere in considerazione una o più contromisure, compresa l'eliminazione di alcune variabili regressore dal modello, o l'uso di una tecnica alternativa al metodo dei minimi quadrati per stimare i parametri del modello. Una discussione esaustiva della multicollinearità si può trovare in Montgomery, Peck, Vining (2006).

## 6.4 ALTRI ASPETTI DELLA REGRESSIONE

In questo paragrafo passiamo brevemente in rassegna altri tre aspetti dell'utilizzo della regressione multipla: la costruzione di modelli con termini polinomiali, l'uso di variabili qualitative o categoriche come regressori, la scelta delle variabili per un modello di regressione. Per una discussione più dettagliata di questi (e altri) argomenti, si può consultare Montgomery, Runger (2011) o Montgomery, Peck, Vining (2006).

### 6.4.1 Modelli polinomiali

Nel Paragrafo 6.1 abbiamo osservato che i modelli con termini polinomiali nei regressori, quale il modello del secondo ordine

$$Y = \beta_0 + \beta_1 x_1 + \beta_{11} x_1^2 + \epsilon$$

sono in realtà modelli di regressione lineare e possono essere stimati e analizzati usando i metodi discussi nel Paragrafo 6.3. I modelli polinomiali si incontrano frequentemente nelle scienze e in ingegneria, e questo contribuisce in modo rilevante alla diffusione della regressione lineare in tali campi.

**ESEMPIO 6.9**  
 Resa  
 dell'acetilene

Per illustrare come si tratta un modello di regressione polinomiale, consideriamo i dati relativi alla produzione di acetilene e a due variabili di processo: la temperatura del reattore e il rapporto tra H<sub>2</sub> ed n-eptano [per una discussione e un'analisi più dettagliate di questi dati e per i riferimenti bibliografici si veda Montgomery, Peck, Vining (2006)] (Tabella 6.9). Gli ingegneri, di fronte a questo tipo di dati di processo chimico, ricorrono spesso a un modello del secondo ordine.

**Tabella 6.9** Dati relativi alla produzione di acetilene.

Osservazione	Prod., Y	Temp., T	Rapp., R	Osservazione	Prod., Y	Temp., T	Rapp., R
1	49.0	1300	7.5	9	34.5	1200	11.0
2	50.2	1300	9.0	10	35.0	1200	13.5
3	50.5	1300	11.0	11	38.0	1200	17.0
4	48.5	1300	13.5	12	38.5	1200	23.0
5	47.5	1300	17.0	13	15.0	1100	5.3
6	44.5	1300	23.0	14	17.0	1100	7.5
7	28.0	1200	5.3	15	20.5	1100	11.0
8	31.5	1200	7.5	16	29.5	1100	17.0

Il modello del secondo ordine con due regressori è

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$$

Quando si stima un modello polinomiale, è in genere buona norma **centrare** i regressori (sottraendo la media  $\bar{x}_j$  da ciascuna osservazione) e usare i regressori centrati per costruire il modello. Ciò riduce la **multicollinearità** dei dati, e porta sovente a un modello di regressione più affidabile, perché caratterizzato da coefficienti stimati con maggiore precisione. Per i dati dell'acetilene ciò comporta sottrarre 1212.5 da ciascuna osservazione su  $x_1$  = temperatura, e 12.444 da ciascuna osservazione su  $x_2$  = rapporto. Pertanto, il modello di regressione che adatteremo è

$$Y = \beta_0 + \beta_1(T - 1212.5) + \beta_2(R - 12.444) \\ + \beta_{12}(T - 1212.5)(R - 12.444) + \beta_{11}(T - 1212.5)^2 + \beta_{22}(R - 12.444)^2 + \epsilon$$

Calcol usando i dati  
centrati.

Di seguito è mostrata una parte dell'output di Minitab per questo modello.

The regression equation is

$$\text{Yield} = 36.4 + 0.130 \text{ Temp} + 0.480 \text{ Ratio} - 0.00733 \text{ T} \times \text{R} + 0.000178 \text{ T}^2 \\ - 0.0237 \text{ R}^2$$

Predictor	Coef	SE Coef	T	P	VIF
Constant	36.4339	0.5529	65.90	0.000	
Temp	0.130476	0.003642	35.83	0.000	1.1
Ratio	0.48005	0.05860	8.19	0.000	1.5
T × R	-0.0073346	0.0007993	-9.18	0.000	1.4
T^2	0.00017820	0.00005854	3.04	0.012	1.2
R^2	-0.02367	0.01019	-2.32	0.043	1.7

$$S = 1.066 \quad R-\text{Sq} = 99.5\% \quad R-\text{Sq} (\text{adj}) = 99.2\%$$

#### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	2112.34	422.47	371.49	0.000
Residual Error	10	11.37	1.14		
Total	15	2123.71			

Si ricordi che i coefficienti di regressione in questo output si riferiscono ai regressori *centrati* nel modello presentato più sopra. Il test dell'ANOVA per la significatività della regressione suggerisce che almeno alcune delle variabili nel modello sono importanti; i test *t* sulle singole variabili indicano che nel modello tutti i termini sono necessari. I VIF sono tutti piccoli, e non esiste quindi un evidente problema con la multicollinearità.

Supponiamo di volere valutare il contributo apportato a questo modello dai termini di secondo ordine. In altre parole, è valsa la pena di complicare il modello includendo i termini del secondo ordine? Le ipotesi che occorre verificare sono

$$H_0: \beta_{r+1} = \beta_{r+2} = \cdots = \beta_k = 0$$

$$H_1: \text{almeno uno dei } \beta \neq 0$$

#### ESEMPIO 6.9 (proseguo)

Abbiamo mostrato come verificare queste ipotesi nel Paragrafo 6.3.2. Ricordiamo che la procedura richiede di considerare il modello quadratico come **modello completo** e di stimare quindi un **modello ridotto**, che in questo caso sarebbe il modello del primo ordine

$$Y = \beta_0 + \beta_1(T - 1212.5) + \beta_2(R - 12.444) + \epsilon$$

#### Calcolo del modello ridotto.

L'output della regressione prodotto da Minitab per questo modello ridotto è il seguente.

The regression equation is

$$\text{Yield} = 36.1 + 0.134 \text{ Temp} + 0.351 \text{ Ratio}$$

Predictor	Coef	SE Coef	T	P	VIF
Constant	36.1063	0.9060	39.85	0.000	
Temp	0.13396	0.01191	11.25	0.000	1.1
Ratio	0.3511	0.1696	2.07	0.059	1.1

$$S = 3.624 \quad R-\text{Sq} = 92.0\% \quad R-\text{Sq}(\text{adj}) = 90.7\%$$

#### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	1952.98	976.49	74.35	0.000
Residual Error	13	170.73	13.13		
Total	15	2123.71			

La statistica test per le ipotesi sopra citate è già stata data nell'Equazione (6.56), ma la ripetiamo per comodità

$$F_0 = \frac{[SS_E(RM) - SS_E(FM)]/(k - r)}{SS_E(FM)/(n - p)}$$

**ESEMPIO 6.9**  
(proseguo)  
Interpretazione  
dell'adeguatezza  
del modello ridotto.

Nella statistica test,  $SS_E(RM) = 170.73$  è la somma dei quadrati dei residui per il modello ridotto,  $SS_E(FM) = 11.37$  è la somma dei quadrati dei residui per il modello completo,  $n = 16$  è il numero di osservazioni,  $p = 6$  è il numero di parametri nel modello completo,  $k = 5$  è il numero di regressori nel modello completo, e  $r = 2$  è il numero di regressori nel modello ridotto. Pertanto, il valore calcolato della statistica test è

$$f_0 = \frac{[SS_E(RM) - SS_E(FM)]/(k - r)}{SS_E(FM)/(n - p)} = \frac{(170.3 - 11.37)/(5 - 2)}{11.37/(16 - 6)} = 46.72$$

Questo valore dovrebbe essere confrontato con  $f_{\alpha/2, 10}$ . In alternativa, il  $P$ -value è  $3.49 \times 10^{-6}$ ; essendo molto piccolo, dovremmo rifiutare l'ipotesi nulla  $H_0: \beta_{12} = \beta_{11} = \beta_{22} = 0$  e concludere che almeno uno dei termini di secondo ordine contribuisce in modo significativo al modello. In realtà, sappiamo dai test  $t$  nell'output di Minitab che tutti e tre i termini del secondo ordine sono importanti.

#### 6.4.2 Regressori categorici

Nei modelli di regressione studiati sinora tutte le variabili regressore erano variabili **quantitative**, ossia variabili di tipo numerico oppure misurabili su una scala ben definita. A volte, però, è necessario includere nel modello di regressione variabili **categoriche** o **qualitative**.

**ESEMPIO 6.10**  
Consumo  
di benzina  
per chilometro

Supponiamo di studiare il consumo di benzina per chilometro di un gruppo di automobili. La variabile risposta è  $Y$  = consumo di benzina per chilometro, e due regressori di interesse sono  $x_1$  = cilindrata del motore (in  $\text{cm}^3$ ) e  $x_2$  = potenza del motore (in hp). La maggior parte delle auto esaminate hanno il cambio automatico, ma alcune hanno il cambio manuale.

È semplice incorporare un'informazione categorica come quella che indica il tipo di cambio in un modello di regressione. Sia  $x_3$  = tipo di cambio; la definiamo nel modo seguente

$$x_3 = \begin{cases} 0 & \text{se l'auto ha il cambio automatico} \\ 1 & \text{se l'auto ha il cambio manuale} \end{cases}$$

A volte una variabile che assume valore 0 oppure 1 viene detta **variabile indicatore**. Il modello di regressione per l'analisi del consumo di carburante è quindi

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Questo modello in effetti descrive due differenti modelli di regressione. Quando  $x_3 = 0$ , e l'automobile ha quindi un cambio automatico, il modello per il consumo di benzina è

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

ma quando l'automobile ha un cambio manuale ( $x_3 = 1$ ), il modello è

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(1) + \epsilon \\ &= (\beta_0 + \beta_3) + \beta_1 x_1 + \beta_2 x_2 + \epsilon \end{aligned}$$

Si noti che i due modelli hanno differenti intercette, ma i parametri del modello che contengono informazioni sulla cilindrata e sulla potenza non sono influenzati dal tipo di cambio dell'auto. Ciò potrebbe apparire non ragionevole; in effetti, ci si potrebbe aspettare un'**interazione** tra le variabili regressore cilindrata del motore e tipo di cambio, e anche tra potenza e tipo di cambio.

Includere termini di interazione come quello qui considerato è facile. Il modello appropriato è

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \epsilon$$

Ora, quando l'auto ha un cambio automatico ( $x_3 = 0$ ), il modello per il consumo di benzina è

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

ma quando l'auto ha un cambio manuale ( $x_3 = 1$ ) il modello diventa

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(1) + \beta_{13} x_1(1) + \beta_{23} x_2(1) + \epsilon \\ &= (\beta_0 + \beta_3) + (\beta_1 + \beta_{13}) x_1 + (\beta_2 + \beta_{23}) x_2 + \epsilon \end{aligned}$$

Perciò tutti e tre i coefficienti del modello sono influenzati dalla presenza del cambio manuale. Questo potrebbe avere un forte impatto sulla forma della funzione di regressione.

**Uso delle variabili indicatori per modellizzare l'interazione.**

**ESEMPIO 6.11**  
Dati relativi allo shampoo

Come esempio, riconsideriamo i dati di Tabella 2.9 relativi a uno shampoo prodotto negli Stati Uniti. Una delle variabili, la Regione (Est, Ovest), è categorica, e può essere incorporata nel modello di regressione esattamente come abbiamo fatto nell'esempio del consumo di carburante. Se lo shampoo è prodotto nell'Est porremo Regione = 0, se è prodotto nell'Ovest porremo Regione = 1. Di seguito è mostrato l'output di Minitab per un modello di regressione lineare che utilizza come regressori *Schiuma*, *Residuo* e *Regione*.

The regression equation is

$$\text{Quality} = 89.9 + 1.79 \text{ Foam} - 3.34 \text{ Residue} - 3.45 \text{ Region}$$

Predictor	Coef	SE Coef	T	P
Constant	89.856	3.035	29.61	0.000
Foam	1.7927	0.3308	5.42	0.000
Residue	-3.3441	0.6960	-4.80	0.000
Region	-3.4488	0.9331	-3.70	0.001

S = 2.249

R-Sq = 76.8%

R-Sq (adj) = 73.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	335.63	111.88	22.11	0.000
Residual Error	20	101.19	5.06		
Total	23	436.82			

#### Interpretazione dell'output e aggiunta di interazione.

Si noti che tutti e tre i regressori sono importanti. In questo modello, il regressore Regione ha l'effetto di spostare l'intercetta di una quantità uguale a -3.70 unità quando si fanno predizioni sulla qualità per uno shampoo prodotto nell'Ovest.

Si possono studiare i potenziali effetti di interazione in questi dati includendo nel modello i termini di interazione *Schiuma × Regione* e *Residuo × Regione*. Di seguito è mostrato l'output di Minitab per questo modello.

The regression equation is

$$\text{Quality} = 88.6 + 1.96 \text{ Foam} - 3.25 \text{ Residue} - 2.50 \text{ Region} - 0.779 \text{ F} \times \text{R} + 0.80 \text{ Res} \times \text{R}$$

Predictor	Coef	SE Coef	T	P
Constant	88.571	4.888	18.12	0.000
Foam	1.9602	0.4335	4.52	0.000
Residue	-3.2534	0.9621	-3.38	0.003
Region	-2.503	6.638	-0.38	0.710
F × R	-0.7790	0.9529	-0.82	0.424
Res × R	0.797	1.913	0.42	0.682

S = 2.328

R-Sq = 77.7%

R-Sq (adj) = 71.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	339.263	67.853	12.52	0.000
Residual Error	18	97.562	5.420		
Total	23	436.825			

È evidente che in questo modello non sono necessari termini di interazione.

Le variabili indicatore possono essere usate quando esistono più di due livelli di una variabile categorica. Per esempio, si supponga che lo shampoo sia prodotto in tre regioni: Est, Centro e Ovest. Due variabili indicatore (per esempio,  $x_1$  e  $x_2$ ) verrebbero definite come segue:

Regione	$x_1$	$x_2$
Est	0	0
Centro	1	0
Ovest	0	1

In generale, se esistono  $r$  livelli per una variabile categorica, occorreranno  $r - 1$  indicatori per incorporare la variabile categorica nel modello di regressione.

#### 6.4.3 Tecniche di selezione delle variabili

In molte applicazioni della regressione si ha a che fare con un insieme di dati caratterizzato da un numero abbastanza elevato di regressori. Si desidera ovviamente costruire, se possibile, un modello con un numero minore di regressori, in modo da rendere il modello più semplice a livello di operatività e di interpretazione, e tale da consentire predizioni più affidabili rispetto a un modello contenente tutti i regressori.

Se il numero di regressori non è troppo elevato, un modo per selezionare il sottoinsieme di regressori per il modello consiste nello stimare tutti i possibili sottomodelli e nell'operare la scelta finale valutando questi modelli candidati con appropriati criteri. Può sembrare complesso, ma in realtà è un metodo molto pratico e semplice da applicare in numerosi problemi. La limitazione pratica che impone Minitab è di circa 20 regressori candidati; la limitazione effettiva dipende dallo specifico software usato e da come è stato implementato quel particolare pacchetto.

Due criteri spesso utilizzati per valutare un sottomodello di regressione sono  $R^2$  e la media dei quadrati dei residui  $MS_E$ . L'obiettivo è trovare un modello per il quale  $R^2$  sia elevato e  $MS_E$  sia piccolo. Ora,  $R^2$  non può diminuire quando si aggiungono variabili, quindi si deve trovare un sottomodello per il quale il valore di  $R^2$  sia all'incirca uguale al valore assunto da  $R^2$  quando tutte le variabili sono nel modello. È preferibile avere un modello con valore minimo di  $MS_E$ , perché ciò implica che il modello stesso ha potuto rendere conto ampiamente della variabilità della risposta.

Un terzo criterio è basato sull'errore di stima quadratico totale standardizzato

$$\Gamma_p = \frac{E \left\{ \sum_{i=1}^n [\hat{Y}_i - E(Y_i)]^2 \right\}}{\sigma^2} = \frac{E[SS_E(p)]}{\sigma^2} - n + 2p$$

dove  $\hat{Y}_i$  è la risposta predetta dal sottomodello con  $p$  parametri,  $SS_E(p)$  è la somma dei quadrati dei residui associata a questo modello, ed  $E(Y_i)$  è la risposta attesa dal modello “reale”, ossia dal modello con il corretto sottoinsieme di regressori. Le quantità  $E[SS_E(p)]$  e  $\sigma^2$  sono ignote, ma possono venire stimate dal valore osservato di  $SS_E(p)$  e dalla stima  $\hat{\sigma}^2$  ottenuta dal modello completo, cioè il modello contenente tutti i regressori candidati. Il criterio diventa quindi

$$C_p = \frac{SS_E(p)}{\hat{\sigma}^2(FM)} - n + 2p$$

Un modello con un piccolo valore di  $C_p$  è considerato preferibile.

**ESEMPIO 6.12**  
Tutte le possibili regressioni.

Per illustrare l'approccio “tutte le possibili regressioni”, applicheremo la tecnica ai dati relativi allo shampoo (Tabella 2.9). Minitab fornirà i migliori sottomodelli a  $m$  regressori ( $1 \leq m \leq 10$ ) su un totale di regressori candidati che può arrivare sino a 20, dove con “migliore” si intende il modello con il massimo valore di  $R^2$  o il minimo valore di  $MS_E$ . L'output di Minitab con  $m = 5$  è mostrato in Tabella 6.10.

**Tabella 6.10** I migliori sottomodelli di regressione, come scelti da Minitab.  
Qualità in funzione di Schiuma, Profumo, Colore, Residuo, Regione.

Response is Quality					
Vars	R-Sq	R-Sq(adj)	C-p	S	R e R S C s e F c o i g o e l d i a n o u o m t r e n
1	26.2	22.8	45.4	3.8286	X
1	25.5	22.2	45.9	3.8451	X
1	23.7	20.2	47.6	3.8925	X
1	5.7	1.5	63.5	4.3263	X
1	3.8	0.0	65.2	4.3704	X
2	61.0	57.3	16.5	2.8478	X X
2	50.1	45.3	26.2	3.2219	X X
2	42.8	37.4	32.6	3.4486	X X
2	40.2	34.5	35.0	3.5284	X X
2	31.3	24.8	42.8	3.7795	X X
3	76.8	73.4	4.5	2.2493	X X X
3	62.0	56.3	17.6	2.8810	X X X
3	61.5	55.8	18.1	2.8984	X X X
3	52.6	45.4	26.0	3.2188	X X X
3	51.5	44.2	27.0	3.2551	X X X
4	79.6	75.3	4.1	2.1667	X X X X
4	77.9	73.3	5.6	2.2532	X X X X
4	64.0	56.5	17.8	2.8753	X X X X
4	52.6	42.6	28.0	3.3014	X X X X
4	51.5	41.3	28.9	3.3390	X X X X
5	79.7	74.0	6.0	2.2212	X X X X X

Nell'output di Minitab, "S" è la radice quadrata della media dei quadrati dei residui. Il modello con il minor valore della media dei quadrati dei residui è il modello a quattro variabili contenente *Schiuma*, *Profumo*, *Residuo* e *Regione*, che ha anche il minor valore di  $C_p$ . Perciò, assumendo che l'analisi dei residui sia soddisfacente, questo modello sarebbe un buon candidato per la migliore equazione di regressione che descrive le relazioni nell'insieme di dati considerato.

Un altro approccio alla selezione dei sottomodelli di regressione è rappresentato dalla **regressione a passi** (*stepwise*). Si tratta in effetti di un insieme di metodi similari, progettati per lavorare efficacemente con grandi insiemi di dati. Una procedura a passi tra le più usate è l'**eliminazione a ritroso** (*backward elimination*): questo metodo parte con tutti i regressori nel modello, e successivamente li elimina in base al valore della statistica test  $t \hat{\beta}_j/se(\hat{\beta}_j)$ . Se il valore assoluto più piccolo di questo rapporto  $t$  è minore di un valore di soglia  $t_{\text{out}}$ , il regressore associato a questo rapporto  $t$  viene rimosso dal modello. Il modello viene quindi ricalcolato, e il processo di eliminazione a ritroso continua fino a quando nessun altro regressore può essere eliminato. Minitab usa un valore di soglia per  $t_{\text{out}}$  con un livello di significatività di 0.1.

Un'altra variante della regressione a passi è l'**inclusione progressiva** (*forward selection*). Questa procedura inizia con il modello senza variabili e le aggiunge successivamente una alla volta. A ogni passaggio viene inserita nel modello la variabile che risulta avere il più elevato valore della statistica  $t$ , finché il valore della statistica test supera il valore di soglia  $t_{\text{in}}$ . Minitab usa un livello di significatività di 0.25 per determinare il valore di soglia  $t_{\text{in}}$ . La procedura termina quando non rimangono variabili candidate che soddisfino il criterio di ingresso nel modello.

La variante più comune della regressione a passi usa una combinazione di passaggi a ritroso e progressivi ed è usualmente indicata semplicemente come regressione a passi. La procedura inizia con un passo di inclusione, ma immediatamente dopo avere inserito una variabile si effettua un passo di eliminazione per stabilire se qualche variabile aggiunta nei precedenti passaggi possa essere rimossa. Si devono selezionare due valori di soglia:  $t_{\text{in}}$  e  $t_{\text{out}}$ . In genere si pone  $t_{\text{in}} = t_{\text{out}}$ . Minitab usa un livello di significatività di 0.15 sia per  $t_{\text{in}}$  sia per  $t_{\text{out}}$ .

### ESEMPIO 6.12 Regressione a passi

In Tabella 6.11 è mostrata la procedura di Minitab di eliminazione a ritroso applicata ai dati relativi all'esempio dello shampoo. Il modello finale contiene *Schiuma*, *Residuo* e *Regione*. Si noti che questo modello è leggermente differente da quello che si è trovato usando tutte le regressioni possibili.

In Tabella 6.12 è mostrato invece l'output per l'inclusione progressiva. Il modello finale contiene *Schiuma*, *Profumo*, *Residuo* e *Regione*, e coincide con il modello trovato mediante tutte le regressioni possibili.

Infine, in Tabella 6.13 è riportato l'output di Minitab per la regressione a passi riguardante i dati dello shampoo, in cui per il modello finale sono scelti come regressori *Profumo*, *Residuo* e *Regione*. Il modello coincide con quello ottenuto mediante la procedura di eliminazione a ritroso.

**Tabella 6.11** Regressione a passi: qualità in funzione di Schiuma, Profumo, Colore, Residuo, Regione.

Backward elimination. Alpha-to-Remove: 0.1			
Response is Quality on 5 predictors, with N = 24			
Step	1	2	3
Constant	86.04	85.87	89.86
Foam	1.80	1.84	1.79
T-Value	4.99	5.76	5.42
P-Value	0.000	0.000	0.000
Scent	1.16	1.28	
T-Value	1.25	1.60	
P-Value	0.229	0.126	
Color	0.20		
T-Value	0.28		
P-Value	0.783		
Residue	-4.00	-3.94	-3.34
T-Value	-4.90	-5.14	-4.80
P-Value	0.000	0.000	0.000
Region	-3.91	-3.78	-3.45
T-Value	-3.72	-4.10	-3.70
P-Value	0.002	0.001	0.001
S	2.22	2.17	2.25
R-Sq	79.67	79.58	76.83
R-Sq (adj)	74.02	75.28	73.36
C-p	6.0	4.1	4.5

Molti analisti considerano l'approccio "tutte le possibili regressioni" il migliore dei metodi disponibili, perché è possibile valutare implicitamente tutte le equazioni candidate. Di conseguenza, si può essere sicuri di trovare il modello che minimizza la media dei quadrati dei residui o che minimizza  $C_p$ . I metodi a passi sono "miopi", perché cambiano una variabile alla volta in ogni successiva equazione; non garantiscono un'equazione finale che ottimizzi un particolare criterio. Tuttavia, una ricca esperienza pratica con i metodi a passi porta ad affermare che l'equazione che ne risulta è generalmente molto buona.

**Tabella 6.12** Regressione a passi: qualità in funzione di Schiuma, Profumo, Colore, Residuo, Regione.

Forward selection. Alpha-to-Enter: 0.25

Response is Quality on 5 predictors, with N = 24

Step	1	2	3	4
Constant	86.71	78.53	89.86	85.87
Region	-4.37	-4.23	-3.45	-3.78
T-Value	-2.79	-3.21	-3.70	-4.10
P-Value	0.011	0.004	0.001	0.001
Foam		1.47	1.79	1.84
T-Value		3.17	5.42	5.76
P-Value		0.005	0.000	0.000
Residue			-3.34	-3.94
T-Value			-4.80	-5.14
P-Value			0.000	0.000
Scent				1.28
T-Value				1.60
P-Value				0.126
S	3.83	3.22	2.25	2.17
R-Sq	26.18	50.10	76.83	79.58
R-Sq (adj)	22.82	45.34	73.36	75.28
C-p	45.4	26.2	4.5	4.1

**Tabella 6.13** Regressione a passi: qualità in funzione di Schiuma, Profumo, Colore, Residuo, Regione.

---

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is Quality on 5 predictors, with N = 24

Step	1	2	3	4
Constant	86.71	78.53	89.86	85.87
Region	-4.37	-4.23	-3.45	-3.78
T-Value	-2.79	-3.21	-3.70	-4.10
P-Value	0.011	0.004	0.001	0.001
Foam		1.47	1.79	1.84
T-Value		3.17	5.42	5.76
P-Value		0.005	0.000	0.000
Residue			-3.34	-3.94
T-Value			-4.80	-5.14
P-Value			0.000	0.000
Scent				1.28
T-Value				1.60
P-Value				0.126
S	3.83	3.22	2.25	2.17
R-Sq	26.18	50.10	76.83	79.58
R-Sq (adj)	22.82	45.34	73.36	75.28
C-p	45.4	26.2	4.5	4.1

---

## TERMINI E CONCETTI RILEVANTI

---

Adeguatezza del modello	Modello di regressione
Analisi dei residui	Modello empirico
Analisi della regressione	Modello meccanicistico
Analisi della varianza (ANOVA)	Multicollinearità
Approccio “tutte le possibili regressioni”	Osservazioni influenti
Coefficiente di correlazione campionaria, $r$	Outlier
Coefficiente di correlazione della popolazione, $\rho$	$R^2$ corretto
Coefficiente di determinazione, $R^2$	Regressione lineare semplice
Coefficienti di regressione	Regressione multipla
Eliminazione a ritroso	Regressione polinomiale
Equazioni normali dei minimi quadrati	Regressione a passi
Errori standard dei coefficienti del modello	Residui
Fattore di inflazione della varianza	Residui standardizzati
Grafico delle curve di livello	Residui studentizzati
Inclusione progressiva	Significatività della regressione
Interazione	Somma dei quadrati di regressione
Intercetta	Somma dei quadrati dei residui
Intervallo di confidenza per i coefficienti di regressione	Statistica $C_p$
Intervallo di confidenza per la risposta media	Stimatori non distorti
Intervallo di predizione	Test $t$ sui coefficienti di regressione
Metodo dei minimi quadrati	Variabile indicatore
Misura della distanza di Cook, $D_i$	Variabile regressore
Modello	Variabile risposta

# Esercizi proposti

---

## ESERCIZI PER IL PARAGRAFO 6.2

---

Per gli esercizi dal 6.1 al 6.3, svolgere i seguenti punti.

- Stimare l'intercetta  $\beta_0$  e la pendenza  $\beta_1$ . Scrivere la retta di regressione stimata.
- Calcolare i residui
- Calcolare  $SS_E$  e stimare la varianza.
- Trovare l'errore standard dei coefficienti di pendenza e intercetta.
- Dimostrare che  $SS_T = SS_R + SS_E$
- Calcolare il coefficiente di determinazione  $R^2$ , e commentarne il valore.
- Eseguire un test  $t$  per la significatività dei coefficienti di intercetta e pendenza per  $\alpha = 0.05$ . Commentare i risultati.
- Costruire la tabella ANOVA ed eseguire il test di significatività della regressione. Commentare i risultati e la loro relazione con quelli ottenuti al punto (g).
- Costruire un intervallo di confidenza al 95% per l'intercetta e la pendenza. Commentare le relazioni tra questi intervalli e i risultati ottenuti ai punti (g) e (h).
- Controllare l'adeguatezza del modello. Il modello si adatta bene ai dati?
- Calcolare il coefficiente di correlazione campionario ed eseguire il relativo test di significatività per  $\alpha = 0.05$ . Commentare i risultati.



6.1. Per individuare un opportuno sostituto biodegradabile degli imballaggi per cibi pronti è importante stabilire le proprietà dei materiali. Si considerino i seguenti dati sulla densità del prodotto (in  $g/cm^3$ ) e sulla conducibilità termica (in  $W/mK$ ), pubblicati in *Materials Research and Innovation* (1999, pp. 2-8):

Conducibilità Termica	Densità del prodotto
0.0480	0.1750
0.0525	0.2200
0.0540	0.2250
0.0535	0.2260
0.0570	0.2500
0.0610	0.2765

- 6.2. Per analizzare i dati di uno studio sulla relazione tra la temperatura superficiale della strada ( $x$ ) e la curvatura della pavimentazione ( $y$ ) si sono impiegati dei metodi di regressione (si veda la tabella).

Temperatura $x$	Curvatura $y$	Temperatura $x$	Curvatura $y$
70.0	0.621	72.7	0.637
77.0	0.657	67.8	0.627
72.1	0.640	76.6	0.652
72.8	0.623	73.4	0.630
78.3	0.661	70.5	0.627
74.5	0.641	72.1	0.631
74.0	0.637	71.2	0.641
72.4	0.630	73.0	0.631
75.2	0.644	72.7	0.634
76.0	0.639	71.4	0.638

- 6.3. Un articolo apparso sul *Concrete Research* ("Near Surface Characteristics of Concrete: Intrinsic Permeability," vol. 41, 1989) presenta i dati relativi alla resistenza alla compressione  $x$  e alla permeabilità intrinseca  $y$  di alcuni esemplari in cemento prodotti in diverse miscele e sottoposti a diversi trattamenti.

Resistenza $x$	Permeabilità $y$	Resistenza $x$	Permeabilità $y$
3.1	33.0	2.4	35.7
4.5	31.0	3.5	31.9
3.4	34.9	1.3	37.3
2.5	35.6	3.0	33.8
2.2	36.1	3.3	32.8
1.2	39.0	3.2	31.6
5.3	30.1	1.8	37.7
4.8	31.2		

6.4. Si considerino i dati e la regressione lineare semplice dell'Esercizio 6.1.

- Trovare la conducibilità termica media sapendo che la densità del prodotto è 0.2350.
- Calcolare un intervallo di confidenza al 95% per tale risposta media.
- Calcolare un intervallo di predizione al 95% per un'osservazione futura quando la densità del prodotto è 0.2350.
- Commentare l'ampiezza relativa di questi due intervalli. Qual è maggiore, e perché?

6.5. Si considerino i dati e la regressione lineare semplice dell'Esercizio 6.3.

- Trovare la permeabilità media sapendo che la resistenza alla compressione è 2.1.
- Calcolare un intervallo di confidenza al 99% per tale risposta media.
- Calcolare un intervallo di predizione al 99% per un'osservazione futura quando la resistenza alla compressione è 2.1.
- Commentare l'ampiezza di questi due intervalli. Qual è maggiore, e perché?

6.6. Usare il seguente output incompleto di Minitab per rispondere ai seguenti quesiti.

- Determinare tutti i valori mancanti.
- Trovare una stima di  $\sigma^2$ .
- Eseguire un test per la significatività della regressione e per  $\beta_1$ , quindi commentare i due risultati. Utilizzare  $\alpha = 0.05$ .
- Costruire un intervallo di confidenza al 95% su  $\beta_1$ , quindi usarlo per eseguire il test sulla significatività della regressione.
- Commentare i risultati ottenuti ai punti (c) e (d).
- Scrivere il modello di regressione e usarlo per calcolare il residuo per  $x = 0.58$  e  $y = -3.30$ .
- Usare il modello di regressione per calcolare la risposta media e per predire la risposta futura per  $x = 1.5$ . Posto  $\bar{x} = 1.76$  e  $S_{xx} = 5.326191$ , costruire un intervallo di confidenza al 95% per la risposta media e un intervallo di predizione al 95% per la risposta futura. Quale dei due intervalli è più ampio? Perché?

Predictor	Coef	SE Coef	T	P
Constant	0.6649	0.1594	4.17	0.001
X	0.83075	0.08552	?	?
S = ?	$R - Sq = 88.7\%$		$R - Sq(adj) = ?$	
<b>Analysis of Variance</b>				
Source	DF	SS	MS	F P
Regression	1	3.6631	3.6631	?
Residual Error	12	0.4658	?	?
Total	13	?		

6.7. Usare il seguente output incompleto di Minitab per rispondere ai seguenti quesiti.

- Determinare tutti i valori mancanti.
- Trovare una stima di  $\sigma^2$ .
- Eseguire un test per la significatività della regressione e per  $\beta_1$ , quindi commentare i due risultati. Utilizzare  $\alpha = 0.05$ .
- Costruire un intervallo di confidenza al 95% su  $\beta_1$ , quindi usarlo per eseguire il test sulla significatività della regressione.
- Commentare i risultati ottenuti ai punti (c) e (d).
- Scrivere il modello di regressione e usarlo per calcolare il residuo per  $x = 0.58$  e  $y = -3.30$ .
- Usare il modello di regressione per calcolare la media e la predizione della risposta futura per  $x = 0.6$ . Posto  $\bar{x} = 0.52$  e  $S_{xx} = 1.218294$ , costruire un intervallo di confidenza al 95% per la risposta media e un intervallo di predizione al 95% per la risposta futura. Quale dei due intervalli è più ampio? Perché?

Predictor	Coef	SE Coef	T	P
Constant	0.9798	0.3367	2.91	0.011
X	-8.3088	0.5725	?	?
S = ?	$R - Sq = 93.8\%$		$R - Sq(adj) = ?$	
<b>Analysis of Variance</b>				
Source	DF	SS	MS	F P
Regression	1	84.106	84.106	?
Residual Error	14	5.590	?	?
Total	15	?		

## ESERCIZI PER IL PARAGRAFO 6.3

Per gli esercizi 6.8 e 6.9, svolgere i seguenti punti con l'ausilio di Minitab o di un altro software.

- Stimare i coefficienti di regressione. Scrivere il modello di regressione lineare multipla. Commentare la relazione trovata tra l'insieme delle variabili indipendenti e la variabile dipendente.
- Calcolare i residui.
- Calcolare  $SS_E$  e stimare la varianza.
- Calcolare il coefficiente di determinazione  $R^2$  e il coefficiente di determinazione multipla corretto  $R^2_{\text{Adjusted}}$ . Commentare i loro valori.
- Costruire una tabella ANOVA e valutare la significatività della regressione. Commentare i risultati.
- Trovare l'errore standard dei singoli coefficienti.
- Usare un test  $t$  per valutare la significatività dei coefficienti individuali per  $\alpha = 0.05$ . Commentare i risultati.
- Costruire intervalli di confidenza al 95% per i coefficienti individuali. Commentare le relazioni tra questi intervalli e i risultati ottenuti al punto (g).
- Valutare l'adeguatezza del modello, includendo il calcolo dei residui studentizzati e la misura della distanza di Cook per ciascuna osservazione. Commentare i risultati.
- Calcolare il fattore di inflazione della varianza e discutere la presenza di multicollinearità.

6.8. Si considerino i dati relativi ai consumi di carburante dell'Esercizio 2.22.

 6.9. Si ritiene che la potenza elettrica assorbita ogni mese da un'industria chimica sia legata alla temperatura media dell'ambiente ( $x_1$ ), al numero di giorni del mese ( $x_2$ ), alla purezza media del prodotto ( $x_3$ ) e alle tonnellate di sostanza chimica prodotte ( $x_4$ ). Nella seguente tabella sono presentati i dati relativi all'anno passato.

$y$	$x_1$	$x_2$	$x_3$	$x_4$
240	25	24	91	100
236	31	21	90	95
290	45	24	88	110
274	60	25	87	88
301	65	25	91	94
316	72	26	94	99
300	80	25	87	97
296	84	25	86	96
267	75	24	88	110
276	60	25	91	105
288	50	25	90	100
261	38	23	89	98

6.10. Si considerino i dati relativi al consumo di carburante e la regressione lineare multipla dell'Esercizio 6.8.

- Trovare il consumo medio, posto che il peso sia 2650 e la potenza sia 120.
- Calcolare un intervallo di confidenza al 95% per tale risposta media.
- Calcolare un intervallo di predizione al 95% per un'osservazione futura quando il peso è 2650 e la potenza è 120.
- Commentare la dimensione relativa di questi due intervalli. Qual è maggiore, e perché?

6.11. Si considerino i dati relativi all'assorbimento di potenza e la regressione lineare multipla nell'Esercizio 6.9.

- Trovare il consumo medio di potenza, nel caso in cui  $x_1 = 75 \text{ } ^\circ\text{F}$ ,  $x_2 = 24$  giorni,  $x_3 = 90\%$  e  $x_4 = 98$  t.
- Calcolare un intervallo di confidenza al 95% per tale risposta media.
- Calcolare un intervallo di predizione al 95% quando  $x_1 = 75 \text{ } ^\circ\text{F}$ ,  $x_2 = 24$  giorni,  $x_3 = 90\%$  e  $x_4 = 98$  t.
- Commentare la dimensione relativa di questi due intervalli. Qual è maggiore, e perché?

6.12. Usare il seguente output incompleto di Minitab per rispondere ai seguenti quesiti.

- Determinare tutti i valori mancati.
- Trovare una stima di  $\sigma^2$ .
- Eseguire un test per la significatività della regressione. Utilizzare  $\alpha = 0.05$ .
- Eseguire un test per la significatività di  $\beta_1$  e  $\beta_2$  ricorrendo a un test  $t$  con  $\alpha = 0.05$ . Commentare i due risultati.
- Costruire un intervallo di confidenza al 95% su  $\beta_1$ , quindi usarlo per eseguire il test sulla significatività di  $\beta_1$ .
- Costruire un intervallo di confidenza al 95% su  $\beta_2$ , quindi usarlo per eseguire il test sulla significatività di  $\beta_2$ .
- Commentare i risultati ottenuti ai punti (c)-(f). Il modello di regressione è appropriato? Che cosa raccomandereste come successivo passo dell'analisi?

Predictor	Coef	SE Coef	T	P
Constant	3.318	1.007	3.29	0.003
x1	0.7417	0.5768	?	?
x2	9.1142	0.6571	?	?
$S = ?$ $R - Sq = ?$ $R - Sq(\text{adj}) = 87.6\%$				
<b>Analysis of Variance</b>				
Source	DF	SS	MS	F
Regression	2	133.366	66.683	?
Residual Error	?	17.332	?	
Total	27	150.698		

6.13. Usare il seguente output incompleto di Minitab per rispondere ai seguenti quesiti.

- Determinare tutti i valori mancanti.
- Trovare una stima di  $\sigma^2$ .
- Eseguire un test per la significatività della regressione. Utilizzare  $\alpha = 0.05$ .
- Eseguire un test per la significatività di  $\beta_1$ ,  $\beta_2$  e  $\beta_3$  ricorrendo a un test  $t$  con  $\alpha = 0.05$ . Commentare i risultati.
- Costruire un intervallo di confidenza al 95% su  $\beta_1$ , quindi usarlo per eseguire il test sulla significatività di  $\beta_1$ .
- Costruire un intervallo di confidenza al 95% su  $\beta_2$ , quindi usarlo per eseguire il test sulla significatività di  $\beta_2$ .
- Costruire un intervallo di confidenza al 95% su  $\beta_3$ , quindi usarlo per eseguire il test sulla significatività di  $\beta_3$ .
- Commentare i risultati ottenuti ai punti (c)-(g). Il modello di regressione è appropriato? Che cosa raccomandereste come successivo passo dell'analisi?

Predictor	Coef	SE Coef	T	P
Constant	6.188	2.704	2.29	0.027
x1	9.6864	0.4989	?	?
x2	-0.3796	0.2339	?	?
x3	2.9447	0.2354	?	?
$S = ? \quad R - Sq = ? \quad R - Sq(\text{adj}) = 90.2\%$				
Analysis of Variance				
Source	DF	SS	MS	F
Regression	3	363.01	121.00	?
Residual Error	44	36.62		?
Total	47	399.63		

#### ESERCIZI PER IL PARAGRAFO 6.4

6.14. Si considerino i dati relativi al consumo di carburante per chilometro e il problema di regressione lineare multipla degli Esercizi 6.8 e 6.9. Svolgere i seguenti punti con l'ausilio di Minitab di altro software.

- Calcolare tutti i modelli polinomiali del secondo ordine.
- Controllare la multicollinearità nei dati per ciascuno dei modelli polinomiali. Commentare i risultati.
- Valutare il contributo dei termini del secondo ordine nei modelli rispetto al modello ridotto del primo ordine. Commentare i risultati.

6.15. Si considerino i dati relativi all'assorbimento di potenza dell'Esercizio 6.9.

Usando solo termini del primo ordine, costruire i modelli di regressione sfruttando le seguenti tecniche:

- Tutte le regressioni possibili. Trovare i valori di  $C_p$  e di  $S$ .
- Inclusione progressiva.
- Eliminazione a ritroso.
- Commentare i modelli ottenuti. Quale modello è preferibile?

6.16. Un ingegnere meccanico sta analizzando la relazione esistente tra la finitura superficiale delle parti metalliche prodotta con un tornio e la velocità (numero di giri al minuto) del tornio stesso. I dati sono mostrati in Tabella 6.14.

- Costruire un modello di regressione usando variabili indicatore e commentare la significatività della regressione.
- Costruire un modello separato per ogni tipo di utensile e commentare la significatività della regressione per ciascun modello.

Tabella 6.14 Dati riguardanti la finitura superficiale per l'Esercizio 6.12.

Osservazione Numero, $i$	Finitura superficiale $y_i$	Tipo di utensile giri/min	Tipo di utensile Tool	Osservazione Number, $i$	Finitura superficiale $y_i$	Tipo di utensile giri/min	Tipo di utensile Tool
1	45.44	225	302	11	33.50	224	416
2	42.03	200	302	12	31.23	212	416
3	50.10	250	302	13	37.52	248	416
4	48.75	245	302	14	37.13	260	416
5	47.92	235	302	15	34.70	243	416
6	47.79	237	302	16	33.92	238	416
7	52.26	265	302	17	32.13	224	416
8	50.52	259	302	18	35.47	251	416
9	45.58	221	302	19	33.49	232	416
10	44.78	218	302	20	32.29	216	416

## ESERCIZI DI FINE CAPITOLO

**6.17. Residui studentizzati.** Dimostrare che in un modello di regressione lineare semplice la varianza dell' $i$ -esimo residuo è

$$V(e_i) = \sigma^2 \left[ 1 - \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \right]$$

*Suggerimento:*

$$\text{cov}(Y_i, \hat{Y}_i) = \sigma^2 \left[ \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right]$$

L' $i$ -esimo residuo studentizzato per questo modello è definito da

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2 \left[ 1 - \left( \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \right]}}$$

- (a) Spiegare perché  $r_i$  ha deviazione standard unitaria (per  $\sigma$  nota).
- (b) I residui standardizzati hanno deviazione standard unitaria?
- (c) Discutere il comportamento dei residui studentizzati quando il valore campionario  $x_i$  è molto vicino al centro dell'intervallo delle  $x$ .
- (d) Discutere il comportamento del residuo studentizzato quando il valore campionario  $x_i$  è molto vicino a un estremo dell'intervallo delle  $x$ .



6.18. I dati mostrati in Tabella 6.15 si riferiscono alla spinta di un motore a turbina ( $y$ ) e ai seguenti sei candidati regressori:  $x_1$  = velocità di rotazione primaria,  $x_2$  = velocità di rotazione secondaria,  $x_3$  = velocità del flusso di carburante,  $x_4$  = pressione,  $x_5$  = temperatura dei gas di scarico,  $x_6$  = temperatura dell'ambiente al momento del test.

- (a) Stimare un modello di regressione lineare usando come regressori:  $x_3$ ,  $x_4$  e  $x_5$ .
- (b) Valutare la significatività della regressione usando  $\alpha = 0.01$ . Trovare il  $P$ -value per questo test. Quali conclusioni si possono trarre?
- (c) Trovare la statistica test  $t$  per ogni regressore. Usando  $\alpha = 0.01$ , spiegare in dettaglio le conclusioni che si possono trarre da queste statistiche.
- (d) Trovare i coefficienti  $R^2$  e  $R^2_{\text{Adjusted}}$  per questo modello. Commentare il significato di ciascun valore e la loro utilità nella valutazione del modello.

- (e) Costruire un grafico dei quantili per i residui e interpretarlo.
- (f) Rappresentare graficamente i residui in funzione di . Esiste qualche indicazione di disomogeneità della varianza o di non linearità?
- (g) Rappresentare graficamente i residui in funzione di  $x_3$ . Esiste qualche indicazione di non linearità?
- (h) Predire la spinta di un motore per cui  $x_3 = 1670$ ,  $x_4 = 170$  e  $x_5 = 1589$ .

6.19. Si considerino i dati relativi alla spinta del motore a turbina dell'Esercizio 6.18. Stimare nuovamente il modello usando  $y^* = \ln y$  come variabile risposta e  $x_3^* = \ln x_3$  come regressore (assieme a  $x_4$  e  $x_5$ ).

- (a) Valutare la significatività della regressione usando  $\alpha = 0.01$ . Trovare il  $P$ -value per questo test e trarre delle conclusioni.
- (b) Usare la statistica  $t$  per verificare  $H_0: \beta_j = 0$  rispetto a  $H_1: \beta_j \neq 0$  per ciascuna variabile del modello. Se  $\alpha = 0.01$ , che conclusioni si possono trarre?
- (c) Rappresentare graficamente i residui in funzione di e in funzione di  $x_3^*$ . Commentare tali grafici e confrontarli con la loro controparte ottenuta nell'Esercizio 6.18 punti (f) e (g).



6.20. Un motore di missile viene realizzato unendo due tipi di propulsori, un igniter e un sustainer. Si ritiene che la resistenza al taglio della giunzione,  $y$ , sia funzione lineare dell'età  $x$  del propulsore al momento della realizzazione del motore. Nella seguente tabella sono elencate venti osservazioni.

- (a) Tracciare un diagramma di dispersione dei dati. La linea retta del modello di regressione sembra plausibile?
- (b) Trovare le stime dei minimi quadrati della pendenza e dell'intercetta nel modello di regressione lineare semplice.
- (c) Stimare la resistenza al taglio media di un motore in cui il propulsore ha 20 settimane.
- (d) Ricavare i valori stimati  $\hat{y}_i$  corrispondenti a ciascun valore di  $y_i$ . Rappresentare graficamente  $\hat{y}_i$  in funzione di  $y_i$ , e dire quale sarebbe la forma di questo grafico se la relazione lineare tra resistenza al taglio ed età del propulsore fosse perfettamente deterministica (nessun errore). Il grafico indica che è corretto scegliere l'età come variabile regressore di questo modello?

Tabella 6.15 Dati per l'Esercizio 6.18.

Osservazione Numero	$y$	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$
1	4540	2140	20640	30250	205	1732	99
2	4315	2016	20280	30010	195	1697	100
3	4095	1905	19860	29780	184	1662	97
4	3650	1675	18980	29330	164	1598	97
5	3200	1474	18100	28960	144	1541	97
6	4833	2239	20740	30083	216	1709	87
7	4617	2120	20305	29831	206	1669	87
8	4340	1990	19961	29604	196	1640	87
9	3820	1702	18916	29088	171	1572	85
10	3368	1487	18012	28675	149	1522	85
11	4445	2107	20520	30120	195	1740	101
12	4188	1973	20130	29920	190	1711	100
13	3981	1864	19780	29720	180	1682	100
14	3622	1674	19020	29370	161	1630	100
15	3125	1440	18030	28940	139	1572	101
16	4560	2165	20680	30160	208	1704	98
17	4340	2048	20340	29960	199	1679	96
18	4115	1916	19860	29710	187	1642	94
19	3630	1658	18950	29250	164	1576	94
20	3210	1489	18700	28890	145	1528	94
21	4330	2062	20500	30190	193	1748	101
22	4119	1929	20050	29960	183	1713	100
23	3891	1815	19680	29770	173	1684	100
24	3467	1595	18890	29360	153	1624	99
25	3045	1400	17870	28960	134	1569	100
26	4411	2047	20540	30160	193	1746	99
27	4203	1935	20160	29940	184	1714	99
28	3968	1807	19750	29760	173	1679	99
29	3531	1591	18890	29350	153	1621	99
30	3074	1388	17870	28910	133	1561	99
31	4350	2071	20460	30180	198	1729	102
32	4128	1944	20010	29940	186	1692	101
33	3940	1831	19640	29750	178	1667	101
34	3480	1612	18710	29360	156	1609	101
35	3064	1410	17780	28900	136	1552	101
36	4402	2066	20520	30170	197	1758	100
37	4180	1954	20150	29950	188	1729	99
38	3973	1835	19750	29740	178	1690	99
39	3530	1616	18850	29320	156	1616	99
40	3080	1407	17910	28910	137	1569	100

Osservazione Numero	Resistenza al taglio $y$ (psi)	Età $x$ (settimane)
1	2158.70	15.50
2	1678.15	23.75
3	2316.00	8.00
4	2061.30	17.00
5	2207.50	5.00
6	1708.30	19.00
7	1784.70	24.00
8	2575.00	2.50
9	2357.90	7.50
10	2277.70	11.00
11	2165.20	13.00
12	2399.55	3.75
13	1779.80	25.00
14	2336.75	9.75
15	1765.30	22.00
16	2053.50	18.00
17	2414.40	6.00
18	2200.50	12.50
19	2654.20	2.00
20	1753.70	21.50

Dati per l'Ex. 6.20



6.21. Si consideri il modello di regressione lineare semplice  $Y = \beta_0 + \beta_1 x + \epsilon$  e si supponga che l'analista desideri usare  $z = x - \bar{x}$  come variabile regressore.

- (a) Usando i dati dell'Esercizio 6.20 costruire un grafico di dispersione dei punti  $(x_i, y_i)$  e un altro grafico per i punti  $(z_i = x_i - \bar{x}, y_i)$ . Usare i due grafici per spiegare intuitivamente la relazione fra i due modelli,  $Y = \beta_0 + \beta_1 x + \epsilon$  e  $Y = \beta_0^* + \beta_1^* z + \epsilon$ .
- (b) Trovare le stime dei minimi quadrati di  $\beta_0^*$  e  $\beta_1^*$  nel modello  $Y = \beta_0^* + \beta_1^* z + \epsilon$ . Qual è la relazione con le stime dei minimi quadrati  $\hat{\beta}_0$  e  $\hat{\beta}_1$ ?

6.22. Si supponga che ogni valore di  $x_i$  venga moltiplicato per una costante positiva  $a$ , e che ciascun valore di  $y_i$  venga moltiplicato per un'altra costante positiva  $b$ . Dimostrare che la statistica  $t$  per la verifica di  $H_0: \beta_1 = 0$  rispetto a  $H_1: \beta_1 \neq 0$  non cambia valore.



# Appendice A

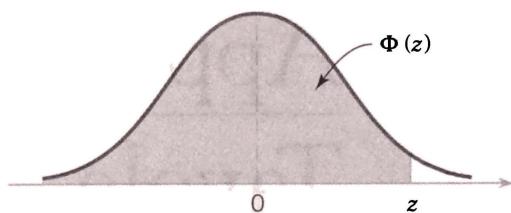
## Tavole

## e carte

## statistiche

Tavola I Distribuzione normale standard cumulativa

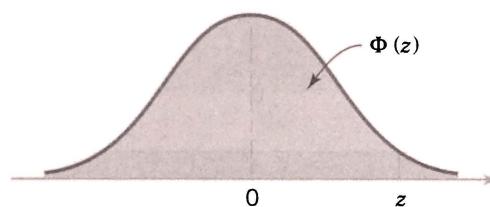
$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$



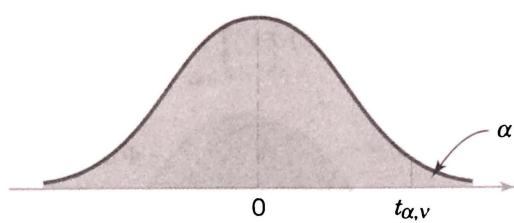
$z$	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01	-0.00	$z$
-3.9	0.000033	0.000034	0.000036	0.000037	0.000039	0.000041	0.000042	0.000044	0.000046	0.000048	-3.9
-3.8	0.000050	0.000052	0.000054	0.000057	0.000059	0.000062	0.000064	0.000067	0.000069	0.000072	-3.8
-3.7	0.000075	0.000078	0.000082	0.000085	0.000088	0.000092	0.000096	0.000100	0.000104	0.000108	-3.7
-3.6	0.000112	0.000117	0.000121	0.000126	0.000131	0.000136	0.000142	0.000147	0.000153	0.000159	-3.6
-3.5	0.000165	0.000172	0.000179	0.000185	0.000193	0.000200	0.000208	0.000216	0.000224	0.000233	-3.5
-3.4	0.000242	0.000251	0.000260	0.000270	0.000280	0.000291	0.000302	0.000313	0.000325	0.000337	-3.4
-3.3	0.000350	0.000362	0.000376	0.000390	0.000404	0.000419	0.000434	0.000450	0.000467	0.000483	-3.3
-3.2	0.000501	0.000519	0.000538	0.000557	0.000577	0.000598	0.000619	0.000641	0.000664	0.000687	-3.2
-3.1	0.000711	0.000736	0.000762	0.000789	0.000816	0.000845	0.000874	0.000904	0.000935	0.000968	-3.1
-3.0	0.001001	0.001035	0.001070	0.001107	0.001144	0.001183	0.001223	0.001264	0.001306	0.001350	-3.0
-2.9	0.001395	0.001441	0.001489	0.001538	0.001589	0.001641	0.001695	0.001750	0.001807	0.001866	-2.9
-2.8	0.001926	0.001988	0.002052	0.002118	0.002186	0.002256	0.002327	0.002401	0.002477	0.002555	-2.8
-2.7	0.002635	0.002718	0.002803	0.002890	0.002980	0.003072	0.003167	0.003264	0.003364	0.003467	-2.7
-2.6	0.003573	0.003681	0.003793	0.003907	0.004025	0.004145	0.004269	0.004396	0.004527	0.004661	-2.6
-2.5	0.004799	0.004940	0.005085	0.005234	0.005386	0.005543	0.005703	0.005868	0.006037	0.006210	-2.5
-2.4	0.006387	0.006569	0.006756	0.006947	0.007143	0.007344	0.007549	0.007760	0.007976	0.008198	-2.4
-2.3	0.008424	0.008656	0.008894	0.009137	0.009387	0.009642	0.009903	0.010170	0.010444	0.010724	-2.3
-2.2	0.011011	0.011304	0.011604	0.011911	0.012224	0.012545	0.012874	0.013209	0.013553	0.013903	-2.2
-2.1	0.014262	0.014629	0.015003	0.015386	0.015778	0.016177	0.016586	0.017003	0.017429	0.017864	-2.1
-2.0	0.018309	0.018763	0.019226	0.019699	0.020182	0.020675	0.021178	0.021692	0.022216	0.022750	-2.0
-1.9	0.023295	0.023852	0.024419	0.024998	0.025588	0.026190	0.026803	0.027429	0.028067	0.028717	-1.9
-1.8	0.029379	0.030054	0.030742	0.031443	0.032157	0.032884	0.033625	0.034379	0.035148	0.035930	-1.8
-1.7	0.036727	0.037538	0.038364	0.039204	0.040059	0.040929	0.041815	0.042716	0.043633	0.044565	-1.7
-1.6	0.045514	0.046479	0.047460	0.048457	0.049471	0.050503	0.051551	0.052616	0.053699	0.054799	-1.6
-1.5	0.055917	0.057053	0.058208	0.059380	0.060571	0.061780	0.063008	0.064256	0.065522	0.066807	-1.5
-1.4	0.068112	0.069437	0.070781	0.072145	0.073529	0.074934	0.076359	0.077804	0.079270	0.080757	-1.4
-1.3	0.082264	0.083793	0.085343	0.086915	0.088508	0.090123	0.091759	0.093418	0.095098	0.096801	-1.3
-1.2	0.098525	0.100273	0.102042	0.103835	0.105650	0.107488	0.109349	0.111233	0.113140	0.115070	-1.2
-1.1	0.117023	0.119000	0.121001	0.123024	0.125072	0.127143	0.129238	0.131357	0.133500	0.135666	-1.1
-1.0	0.137857	0.140071	0.142310	0.144572	0.146859	0.149170	0.151505	0.153864	0.156248	0.158655	-1.0
-0.9	0.161087	0.163543	0.166023	0.168528	0.171056	0.173609	0.176185	0.178786	0.181411	0.184060	-0.9
-0.8	0.186733	0.189430	0.192150	0.194894	0.197662	0.200454	0.203269	0.206108	0.208970	0.211855	-0.8
-0.7	0.214764	0.217695	0.220650	0.223627	0.226627	0.229650	0.232695	0.235762	0.238852	0.241964	-0.7
-0.6	0.245097	0.248252	0.251429	0.254627	0.257846	0.261086	0.264347	0.267629	0.270931	0.274253	-0.6
-0.5	0.277595	0.280957	0.284339	0.287740	0.291160	0.294599	0.298056	0.301532	0.305026	0.308538	-0.5
-0.4	0.312067	0.315614	0.319178	0.322758	0.326355	0.329969	0.333598	0.337243	0.340903	0.344578	-0.4
-0.3	0.348268	0.351973	0.355691	0.359424	0.363169	0.366928	0.370700	0.374484	0.378281	0.382089	-0.3
-0.2	0.385908	0.389739	0.393580	0.397432	0.401294	0.405165	0.409046	0.412936	0.416834	0.420740	-0.2
-0.1	0.424655	0.428576	0.432505	0.436441	0.440382	0.444330	0.448283	0.452242	0.456205	0.460172	-0.1
0.0	0.464144	0.468119	0.472097	0.476078	0.480061	0.484047	0.488033	0.492022	0.496011	0.500000	0.0

Tavola I Distribuzione normale standard cumulativa (seguito)

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$$

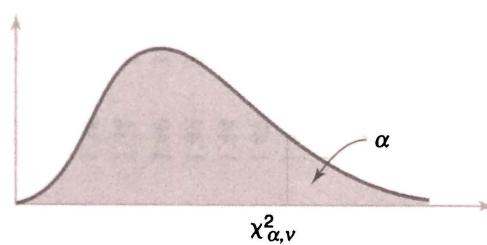


<i>z</i>	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	<i>z</i>
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.523922	0.527903	0.531881	0.535856	0.0
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345	0.1
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092	0.2
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732	0.3
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933	0.4
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715661	0.719043	0.722405	0.5
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903	0.6
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236	0.7
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267	0.8
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913	0.9
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143	1.0
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977	1.1
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475	1.2
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736	1.3
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888	1.4
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083	1.5
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486	1.6
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273	1.7
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621	1.8
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705	1.9
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691	2.0
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738	2.1
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989	2.2
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576	2.3
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613	2.4
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201	2.5
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427	2.6
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365	2.7
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074	2.8
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605	2.9
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999	3.0
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289	3.1
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499	3.2
3.3	0.999517	0.999533	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999650	3.3
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758	3.4
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999821	0.999828	0.999835	3.5
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888	3.6
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925	3.7
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950	3.8
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967	3.9

Tavola II Punti percentuali  $t_{\alpha,v}$  della distribuzione  $t$ 

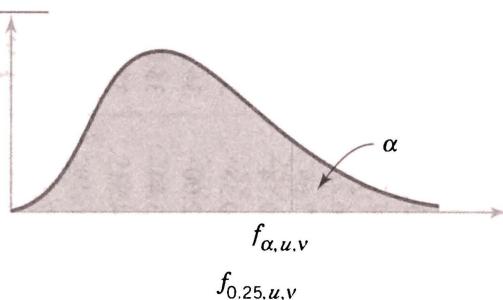
$\nu \backslash \alpha$	.40	.25	.10	.05	.025	.01	.005	.0025	.001	.0005
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	.289	.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	.277	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	.271	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.265	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	.262	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	.260	.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	.260	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	.258	.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	.257	.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	.256	.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	.256	.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	.255	.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	.254	.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	.254	.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
$\infty$	.253	.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

 $\nu$  = gradi di libertà.

Tavola III Punti percentuali  $\chi^2_{\alpha,v}$  della distribuzione chi-quadro

$v \backslash \alpha$	.995	.990	.975	.950	.900	.500	.100	.050	.025	.010	.005
1	.00+	.00+	.00+	.00+	.02	.45	2.71	3.84	5.02	6.63	7.88
2	.01	.02	.05	.10	.21	1.39	4.61	5.99	7.38	9.21	10.60
3	.07	.11	.22	.35	.58	2.37	6.25	7.81	9.35	11.34	12.84
4	.21	.30	.48	.71	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	.41	.55	.83	1.15	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	.68	.87	1.24	1.64	2.20	5.35	10.65	12.59	14.45	16.81	18.55
7	.99	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.28	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.65
28	12.46	13.57	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	79.33	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	89.33	107.57	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	99.33	118.50	124.34	129.56	135.81	140.17

 $v =$  gradi di libertà.

Tavola IV Punti percentuali  $f_{\alpha,u,v}$  della distribuzione F

$v$	$u$	Gradi di libertà per il numeratore ( $u$ )														$\infty$			
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
1	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.41	9.49	9.58	9.63	9.67	9.71	9.76	9.80	9.85
2	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39	3.41	3.43	3.43	3.44	3.45	3.46	3.47	3.48
3	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45	2.46	2.46	2.46	2.47	2.47	2.47	2.47	2.47
4	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08
5	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.88	1.88	1.88	1.88	1.87	1.87	1.87
6	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77	1.76	1.76	1.75	1.75	1.75	1.74	1.74	1.74
7	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.70	1.69	1.68	1.68	1.67	1.67	1.66	1.66	1.65	1.65	1.65
8	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63	1.62	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.58
9	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.54	1.53	1.53
10	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.53	1.52	1.52	1.51	1.51	1.50	1.49	1.48
11	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.47	1.46	1.45
12	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.45	1.44	1.43	1.42
13	1.45	1.55	1.55	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.42	1.41	1.40
14	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.41	1.41	1.40	1.39	1.38
15	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45	1.44	1.43	1.41	1.41	1.40	1.39	1.38	1.37	1.36
16	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34
17	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33
18	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33	1.32
19	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41	1.40	1.38	1.37	1.36	1.35	1.34	1.33	1.32	1.30
20	1.40	1.49	1.48	1.47	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.32	1.31	1.29
21	1.40	1.48	1.48	1.46	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28
22	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39	1.37	1.36	1.34	1.33	1.32	1.31	1.30	1.29	1.28
23	1.39	1.47	1.47	1.45	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28	1.27
24	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38	1.36	1.35	1.33	1.32	1.31	1.30	1.29	1.28	1.26
25	1.39	1.47	1.46	1.44	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.32	1.31	1.29	1.28	1.27	1.25
26	1.38	1.46	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.37	1.35	1.34	1.32	1.31	1.30	1.29	1.28	1.26	1.25
27	1.38	1.46	1.45	1.43	1.42	1.40	1.39	1.38	1.37	1.36	1.35	1.33	1.32	1.31	1.30	1.28	1.27	1.26	1.24
28	1.38	1.46	1.45	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.31	1.30	1.29	1.28	1.27	1.25	1.24
29	1.38	1.45	1.45	1.43	1.41	1.40	1.38	1.37	1.36	1.35	1.34	1.32	1.31	1.30	1.29	1.27	1.26	1.25	1.23
30	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35	1.34	1.32	1.30	1.29	1.28	1.27	1.26	1.24	1.23
40	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.31	1.30	1.28	1.26	1.25	1.24	1.22	1.21	1.19
60	1.35	1.42	1.41	1.38	1.37	1.35	1.33	1.32	1.31	1.30	1.29	1.27	1.25	1.24	1.22	1.21	1.19	1.17	1.15
120	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28	1.26	1.24	1.22	1.21	1.19	1.18	1.16	1.13	1.10
$\infty$	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25	1.24	1.22	1.19	1.18	1.16	1.14	1.12	1.08	1.00

**Tavola IV** Punti percentuali  $f_{\alpha,u,v}$  della distribuzione F (seguito)

$f_{0.10,u,v}$

v	u	Gradi di libertà per il numeratore (u)																	
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.03	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
$\infty$	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

Tavola IV Punti percentuali  $f_{\alpha,u,v}$  della distribuzione F (seguito) $f_{0.05,u,v}$ 

v	u	Gradi di libertà per il numeratore (u)																		
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	$\infty$
	1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
	26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
	28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
	29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
	120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.55	1.43	1.35	1.25
	$\infty$	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

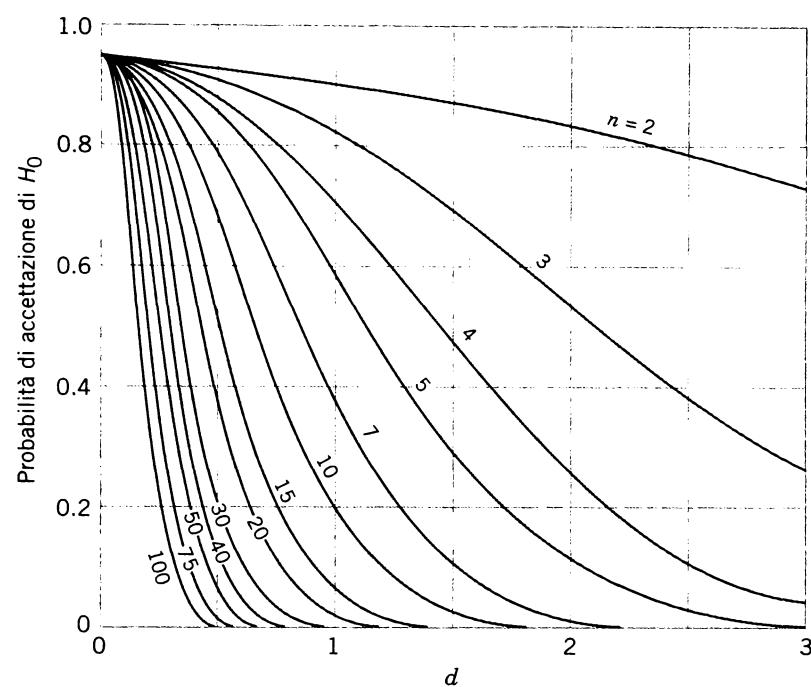
Tavola IV Punti percentuali  $f_{\alpha,u,v}$  della distribuzione F (seguito)

$$f_{0.025,u,v}$$

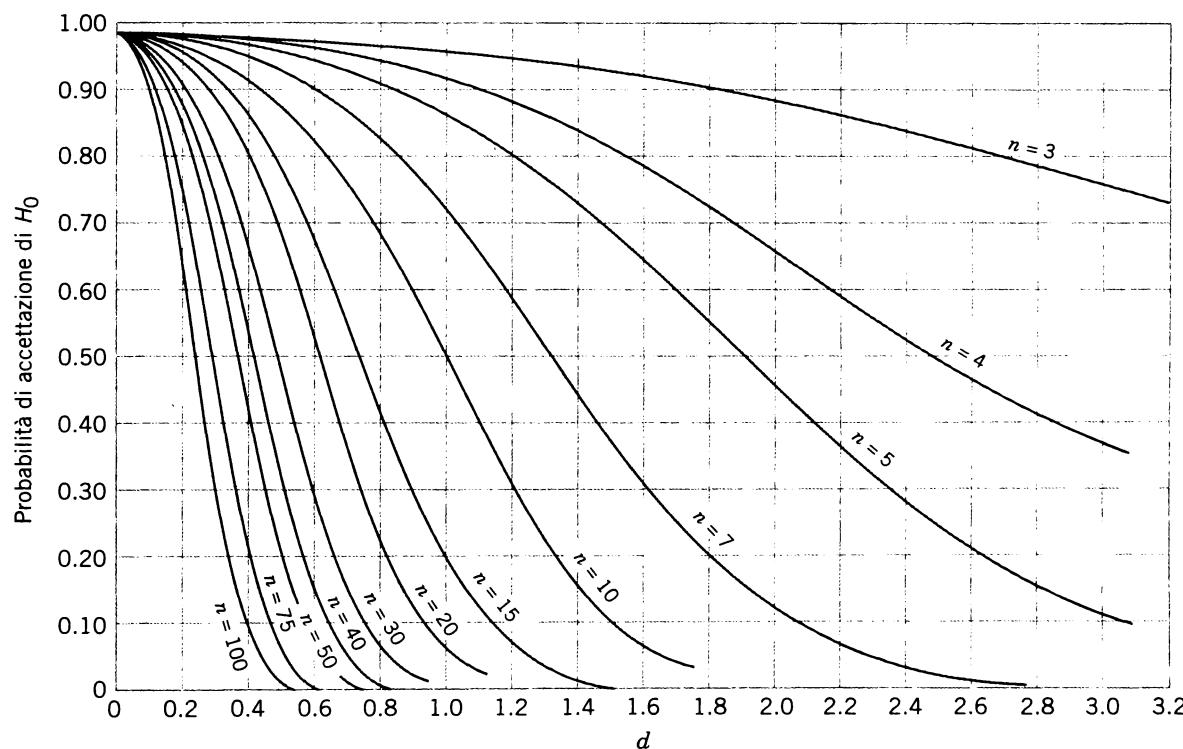
$v$	$u$	Gradi di libertà per il numeratore ( $u$ )																		
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	$\infty$
	1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
	2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
	3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
	4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
	5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
	6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
	7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
	8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
	9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
	10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
	11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
	12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
	13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
	14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
	15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
	16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
	17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
	18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
	19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
	20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
	21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
	22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
	23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
	24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
	25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
	26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
	27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
	28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
	29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
	30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
	40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
	60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
	120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
	$\infty$	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

Tavola IV Punti percentuali  $f_{\alpha,u,v}$  della distribuzione F (seguito) $f_{0.01,u,v}$ 

v	u	Gradi di libertà per il numeratore (u)																		
		1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	$\infty$
	1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
	2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
	3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.00	26.50	26.41	26.32	26.22	26.13
	4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
	5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
	6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
	7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
	8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.46
	9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
	10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
	11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
	12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
	13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
	14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
	15	8.68	6.36	5.42	4.89	4.36	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
	16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
	17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
	18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
	19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.59
	20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
	21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
	22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
	23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
	24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
	25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
	26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
	27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
	28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
	29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
	30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
	40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
	60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
	120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
	$\infty$	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

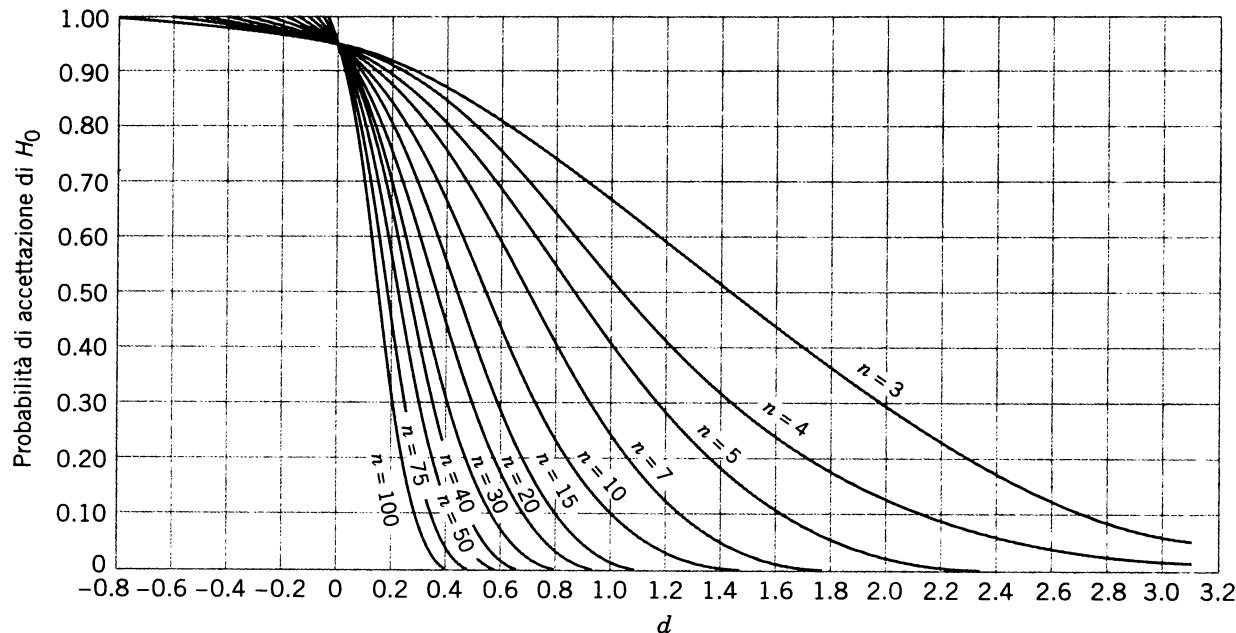
Carta V Curve operative caratteristiche per il test  $t$ 

(a) Curve operative caratteristiche, per diversi valori di  $n$ , per il test  $t$  bilaterale per un livello di significatività  $\alpha = 0.05$ .

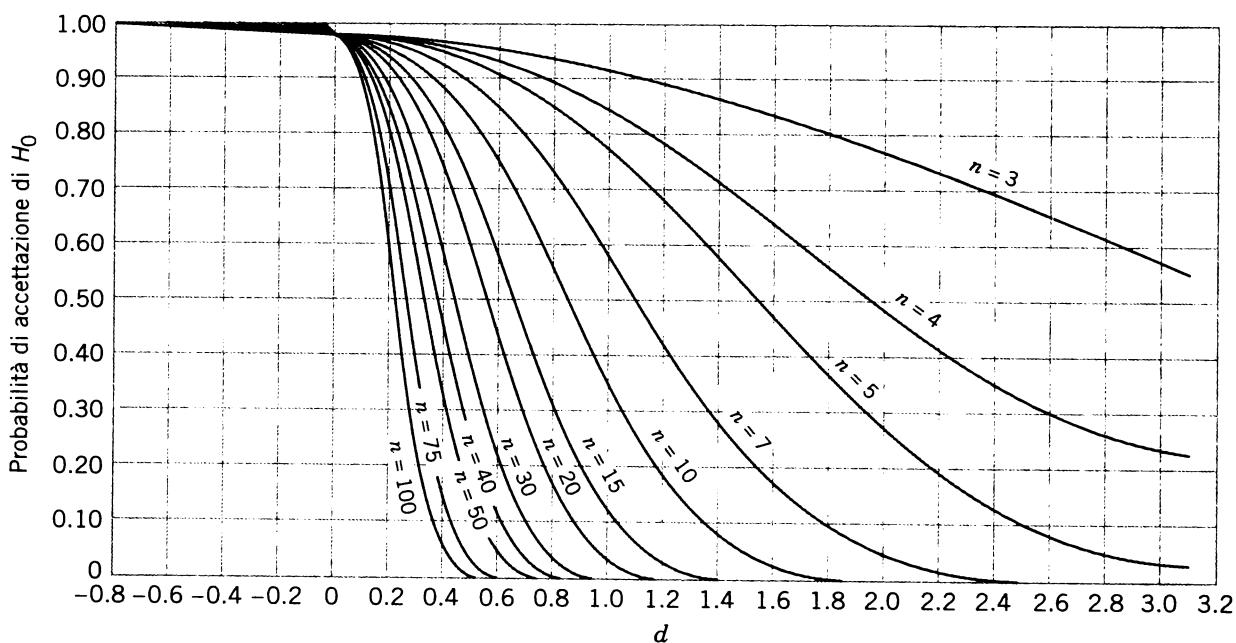


(b) Curve operative caratteristiche, per diversi valori di  $n$ , per il test  $t$  bilaterale per un livello di significatività  $\alpha = 0.01$ .

*Fonte:* Queste carte sono riprodotte su autorizzazione da C.L. Ferris, F.E. Grubbs, C.L. Weaver, "Operating Characteristics for the Common Statistical Tests of Significance", *Annals of Mathematical Statistics*, June 1946, e da A.H. Bowker, G.J. Lieberman, *Engineering Statistics*, 2nd Edition, Prentice-Hall, 1972.

Carta V Curve operative caratteristiche per il test  $t$  (seguito)

(c) Curve operative caratteristiche, per diversi valori di  $n$ , per il test  $t$  unilaterale per un livello di significatività  $\alpha = 0.05$ .



(d) Curve operative caratteristiche, per diversi valori di  $n$ , per il test  $t$  unilaterale per un livello di significatività  $\alpha = 0.01$ .

Tavola VI Fattori per i limiti di tolleranza

Livello di confidenza	Valori di $k$ per intervalli bilaterali								
	0.90			0.95			0.99		
	90	95	99	90	95	99	90	95	99
Percentuale $\gamma$ contenuta									
Dimensione campionaria									
2	15.978	18.800	24.167	32.019	37.674	48.430	160.193	188.491	242.300
3	5.847	6.919	8.974	8.380	9.916	12.861	18.930	22.401	29.055
4	4.166	4.943	6.440	5.369	6.370	8.299	9.398	11.150	14.527
5	3.949	4.152	5.423	4.275	5.079	6.634	6.612	7.855	10.260
6	3.131	3.723	4.870	3.712	4.414	5.775	5.337	6.345	8.301
7	2.902	3.452	4.521	3.369	4.007	5.248	4.613	5.488	7.187
8	2.743	3.264	4.278	3.136	3.732	4.891	4.147	4.936	6.468
9	2.626	3.125	4.098	2.967	3.532	4.631	3.822	4.550	5.966
10	2.535	3.018	3.959	2.839	3.379	4.433	3.582	4.265	5.594
11	2.463	2.933	3.849	2.737	3.259	4.277	3.397	4.045	5.308
12	2.404	2.863	3.758	2.655	3.162	4.150	3.250	3.870	5.079
13	2.355	2.805	3.682	2.587	3.081	4.044	3.130	3.727	4.893
14	2.314	2.756	3.618	2.529	3.012	3.955	3.029	3.608	4.737
15	2.278	2.713	3.562	2.480	2.954	3.878	2.945	3.507	4.605
16	2.246	2.676	3.514	2.437	2.903	3.812	2.872	3.421	4.492
17	2.219	2.643	3.471	2.400	2.858	3.754	2.808	3.345	4.393
18	2.194	2.614	3.433	2.366	2.819	3.702	2.753	3.279	4.307
19	2.172	2.588	3.399	2.337	2.784	3.656	2.703	3.221	4.230
20	2.152	2.564	3.368	2.310	2.752	3.615	2.659	3.168	4.161
21	2.135	2.543	3.340	2.286	2.723	3.577	2.620	3.121	4.100
22	2.118	2.524	3.315	2.264	2.697	3.543	2.584	3.078	4.044
23	2.103	2.506	3.292	2.244	2.673	3.512	2.551	3.040	3.993
24	2.089	2.489	3.270	2.225	2.651	3.483	2.522	3.004	3.947
25	2.077	2.474	3.251	2.208	2.631	3.457	2.494	2.972	3.904
30	2.025	2.413	3.170	2.140	2.529	3.350	2.385	2.841	3.733
40	1.959	2.334	3.066	2.052	2.445	3.213	2.247	2.677	3.518
50	1.916	2.284	3.001	1.996	2.379	3.126	2.162	2.576	3.385
60	1.887	2.248	2.955	1.958	2.333	3.066	2.103	2.506	3.293
70	1.865	2.222	2.920	1.929	2.299	3.021	2.060	2.454	3.225
80	1.848	2.202	2.894	1.907	2.272	2.986	2.026	2.414	3.173
90	1.834	2.185	2.872	1.889	2.251	2.958	1.999	2.382	3.130
100	1.822	2.172	2.854	1.874	2.233	2.934	1.977	2.355	3.096
$\infty$	1.645	1.960	2.576	1.645	1.960	2.576	1.645	1.960	2.576

Tavola VI Fattori per i limiti di tolleranza (*seguito*)

Livello di confidenza	Valori di $k$ per intervalli bilaterali									
	0.90			0.95			0.99			
	Percentuale $\gamma$ contenuta	90	95	99	90	95	99	90	95	99
Dimensione campionaria	2	10.253	13.090	18.500	20.581	26.260	37.094	103.029	131.426	185.617
	3	4.258	5.311	7.340	6.155	7.656	10.553	13.995	17.370	23.896
	4	3.188	3.957	5.438	4.162	5.144	7.042	7.380	9.083	12.387
	5	2.742	3.400	4.666	3.407	4.203	5.741	5.362	6.578	8.939
	6	2.494	3.092	4.243	3.006	3.708	5.062	4.411	5.406	7.335
	7	2.333	2.894	3.972	2.755	3.399	4.642	3.859	4.728	6.412
	8	2.219	2.754	3.783	2.582	3.187	4.354	3.497	4.285	5.812
	9	2.133	2.650	3.641	2.454	3.031	4.143	3.240	3.972	5.389
	10	2.066	2.568	3.532	2.355	2.911	3.981	3.048	3.738	5.074
	11	2.011	2.503	3.443	2.275	2.815	3.852	2.898	3.556	4.829
	12	1.966	2.448	3.371	2.210	2.736	3.747	2.777	3.410	4.633
	13	1.928	2.402	3.309	2.155	2.671	3.659	2.677	3.290	4.472
	14	1.895	2.363	3.257	2.109	2.614	3.585	2.593	3.189	4.337
	15	1.867	2.329	3.212	2.068	2.566	3.520	2.521	3.102	4.222
	16	1.842	2.299	3.172	2.033	2.524	3.464	2.459	3.028	4.123
	17	1.819	2.272	3.137	2.002	2.486	3.414	2.405	2.963	4.037
	18	1.800	2.249	3.105	1.974	2.453	3.370	2.357	2.905	3.960
	19	1.782	2.227	3.077	1.949	2.423	3.331	2.314	2.854	3.892
	20	1.765	2.028	3.052	1.926	2.396	3.295	2.276	2.808	3.832
	21	1.750	2.190	3.028	1.905	2.371	3.263	2.241	2.766	3.777
	22	1.737	2.174	3.007	1.886	2.349	3.233	2.209	2.729	3.727
	23	1.724	2.159	2.987	1.869	2.328	3.206	2.180	2.694	3.681
	24	1.712	2.145	2.969	1.853	2.309	3.181	2.154	2.662	3.640
	25	1.702	2.132	2.952	1.838	2.292	3.158	2.129	2.633	3.601
	30	1.657	2.080	2.884	1.777	2.220	3.064	2.030	2.515	3.447
	40	1.598	2.010	2.793	1.697	2.125	2.941	1.902	2.364	3.249
	50	1.559	1.965	2.735	1.646	2.065	2.862	1.821	2.269	3.125
	60	1.532	1.933	2.694	1.609	2.022	2.807	1.764	2.202	3.038
	70	1.511	1.909	2.662	1.581	1.990	2.765	1.722	2.153	2.974
	80	1.495	1.890	2.638	1.559	1.964	2.733	1.688	2.114	2.924
	90	1.481	1.874	2.618	1.542	1.944	2.706	1.661	2.082	2.883
	100	1.470	1.861	2.601	1.527	1.927	2.684	1.639	2.056	2.850
	$\infty$	1.28	1.645	1.960	1.28	1.645	1.960	1.28	1.645	1.960

# Appendice B

---

## Bibliografia

---

### ragionata

---

#### TESTI INTRODUTTIVI E METODI GRAFICI

Chambers, J., Cleveland, W., Kleiner, B., Tukey, P. (1983), *Graphical Methods for Data Analysis*, Wadsworth & Brooks/Cole, Pacific Grove, CA. Un'ottima presentazione dei metodi grafici utilizzati in statistica.

Freedman, D., Pisani, R., Purves R., Adzikari, A. (1991), *Statistics*, 2nd ed., Norton, New York. Un'eccellente introduzione all'approccio della statistica, che richiede una minima preparazione matematica.

Hoaglin, D., Mosteller, F., Tukey, J. (1983), *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, New York. Una buona disamina illustrativa delle tecniche quali i diagrammi rami e foglie e i box plot.

Tanur, J., et al. (eds.) (1989), *Statistics: A Guide to the Unknown*, 3rd ed., Wadsworth & Brooks/Cole, Pacific Grove, CA. Contiene una raccolta di brevi articoli di carattere non matematico che descrivono diverse applicazioni della statistica.

#### PROBABILITÀ

Derman, C., Olkin, I., Gleser, L. (1980), *Probability Models and Applications*, 2nd ed., Macmillan, New York. Una trattazione generale della probabilità, svolta a un livello matematico superiore a quello adottato nel presente libro.

Hoel, P.G., Port, S. C., Stone, C.J. (1971), *Introduction to Probability Theory*, Houghton Mifflin, Boston. Una trattazione generale, ben scritta, della teoria della probabilità e delle distribuzioni continue e discrete più comuni.

Mosteller, F., Rourke, R., Thomas, G. (1970), *Probability with Statistical Applications*, 2nd ed., Addison-Wesley, Reading, MA. Un'introduzione alla probabilità che non fa uso dell'analisi matematica, con molti esempi eccellenti.

Ross, S. (1998), *A First Course in Probability*, 5th ed., Macmillan, New York. Più formale dal punto di vista matematico di quanto non lo sia il presente libro, ma con molti eccellenti esempi ed esercizi.

## STATISTICA PER L'INGEGNERIA

Montgomery, D.C., Runger, G.C. (2011), *Applied Statistics and Probability for Engineers*, 5th ed., John Wiley & Sons, New York. Un libro più completo sulla statistica per l'ingegneria, più o meno allo stesso livello del presente volume.

Ross, S. (1987), *Introduction to Probability and Statistics for Engineers and Scientists*, John Wiley & Sons, New York. [Tr. it. *Probabilità e statistica per l'ingegneria e le scienze*, Apogeo, Milano, 2003.] Di livello più alto e con un maggiore formalismo matematico; contiene però alcuni ottimi esempi.

## PIANI DEGLI ESPERIMENTI

Box, G.E.P., Hunter, W.G., Hunter, J.S. (2005), *Statistics for Experimenters*, 2nd ed., John Wiley & Sons, New York. Un'eccellente introduzione all'argomento, rivolto agli studenti che desiderano una trattazione di taglio statistico. Contiene molti utili suggerimenti per l'analisi dei dati.

Montgomery, D.C. (2009a), *Design and Analysis of Experiments*, 7th ed., John Wiley & Sons, New York. Scritto allo stesso livello di Box, Hunter, Hunter, ma concentrato sulle applicazioni ingegneristiche e scientifiche.

## CONTROLLO STATISTICO DELLA QUALITÀ E RELATIVI METODI

Duncan, A.J. (1974), *Quality Control and Industrial Statistics*, 4th ed., Richard D. Irwin, Homewood, IL. Un classico sull'argomento.

Grant, E.L., Leavenworth, R.S. (1996), *Statistical Quality Control*, 7th ed., McGraw-Hill, New York. Uno dei primi libri che hanno affrontato la disciplina, contiene numerosi esempi.

Montgomery, D.C. (2009b), *Introduction to Statistical Quality Control*, 6th ed., John Wiley & Sons, New York. [Tr.

it. *Controllo statistico della qualità*, McGraw-Hill Libri Italia, Milano, 2000.] Una trattazione moderna e completa del soggetto, scritta allo stesso livello del presente volume.

Western Electric Company (1956), *Statistical Quality Control Handbook*, Western Electric Company, Inc., Indianapolis, IN.

# Appendice C

---

## Soluzioni

---

### di alcuni esercizi

---

#### CAPITOLO 2

##### Paragrafo 2.1

- 2.1.**  $\bar{x} = 56.09, s = 11.33$   
**2.2.**  $\bar{x} = 1288.43, s = 15.80$   
**2.3.**  $\bar{x} = 43.98, s = 12.29$   
**2.5.** No. Se le osservazioni sono 1, 2, 3, 8, e 10,  $\bar{x} = 4.8$ .  
**2.7.** Entrambe  $\bar{x}$  e  $s$  aumentano di 5%.

##### Paragrafo 2.2

	Mediana	$Q_1$	$Q_3$	$5^{\circ}$	95-esimo
Cicli	1436.5	1097.8	1735.0	772.85	2113.5
Resa	89.25	86.10	93.125	83.055	96.58

- 2.12.** La media campionaria e la deviazione standard cambiano, ma la mediana non cambia.

##### Paragrafo 2.4

- 2.16.** (a)  $\bar{x} = 65.86, s = 12.16$   
(b)  $Q_1 = 58.5, Q_3 = 75$   
(c) Mediana = 67.5  
(d)  $\bar{x} = 66.86, s = 10.74, Q_1 = 60, Q_3 = 75$

#### CAPITOLO 3

##### Paragrafo 3.2

- 3.1.** Continue

- 3.3.** Continue

- 3.4.** Discrete

- 3.6.** Continue

##### Paragrafo 3.3

- 3.7.** (a) Sì (b) 0.6 (c) 0.4 (d) 1  
**3.8.** (a) 0.7 (b) 0.9 (c) 0.2 (d) 0.5  
**3.9.** (a) 0.55 (b) 0.95 (c) 0.50

##### Paragrafo 3.4

- 3.11.** (a)  $3/64, E(X) = 3, V(X) = 0.6$   
(b)  $1/6, E(X) = 11/9, V(X) = 0.284$   
(c)  $1, E(X) = 1, V(X) = 1$   
(d)  $1, E(X) = 100.5, V(X) = 0.08333$   
**3.12.** (a) 1 (b) 0.8647 (c) 0.8647 (d) 0.1353 (e) 9  
**3.13.** (a) 0.7165 (b) 0.2031 (c) 0.6321 (d) 316.2  
(e)  $E(X) = 3000, V(X) = (3000)^2$   
**3.14.** (a) 0.5 (b) 0.5 (c) 0.2 (d) 0.4  
**3.15.** (b)  $1 - x^{-2}$  (c) 2.0 (d) 0.96 (e) 0.0204  
**3.16.** (a) 0.8 (b) 0.5  
**3.17.** (a) 0.9 (b) 0.8 (c) 0.1 (d) 0.1  
(f)  $E(X) = 205, V(X) = 8.333$   
**3.18.** (a) 0.913 (b)  $E(X) = 4.3101, V(X) = 51.4230$   
(c) 12.9303  
**3.19.** (a) 0.778 (b) 0.056 (c) 0.014 (d) 4.658 (e) 3

##### Paragrafo 3.5

- 3.20.** (a) 0 (b) -3.09 (c) -1.18 (d) -1.11 (e) 1.75

- 3.21.** (a) 0.97725 (b) 0.84134 (c) 0.68268  
 (d) 0.9973 (e) 0.47725 (f) 0.49865
- 3.22.** (a) 0.9938 (b) 0.1359 (c) 5835.51
- 3.23.** (a) 0.0082 (b) 0.7211 (c) 0.5641
- 3.25.** (a) 12.309 (b) 12.155
- 3.26.** (a) 0.1587 (b) 90.0 (c) 0.9973
- 3.27.** (a) 0.09012 (b) 0.501165 (c) 13.97
- 3.28.** (a) 0.0668 (b) 0.8664 (c) 0.000214
- 3.33.**  $E(X) = 2.5$ ,  $V(X) = 1.7077$
- 3.36.** (a) moda = 0.8333,  $\mu = 0.6818$ ,  $\sigma^2 = 0.0402$   
 (b) moda = 0.6316,  $\mu = 0.6154$ ,  $\sigma^2 = 0.0137$
- 3.38.** 0.0136
- 3.39.** (a) 0.0248 (b) 0.1501 (c) 92.02

#### Paragrafo 3.7

- 3.40.** (a) 0.433 (b) 0.409 (c) 0.316  
 (d)  $E(X) = 3.319$ ,  $V(X) = 3.7212$
- 3.41.** (a)  $4/7$  (b)  $3/7$   
 (c)  $E(X) = 11/7$ ,  $V(X) = 26/49$
- 3.42.** (a) 0.170 (b) 0.10 (c) 0.91  
 (d)  $E(X) = 9.98$ ,  $V(X) = 2.02$
- 3.43** (b)  $E(X) = 2.5$ ,  $V(X) = 2.05$  (c) 0.5 (d) 0.75

#### Paragrafo 3.8

- 3.46.** (a) 0.0148 (b) 0.8684 (c) 0 (d) 0.1109
- 3.47.** (a) 0.0015 (b) 0.9298 (c) 0 (d) 0.0686
- 3.48.** 0.0043
- 3.49.** (a)  $n = 50$ ,  $p = 0.1$  (b) 0.1117 (c) 0
- 3.50.** (a) 0.9961 (b) 0.989  
 (c)  $E(X) = 112.5$ ,  $\sigma_X = 3.354$
- 3.51.** (a) 0.13422 (b) 0.000001 (c) 0.30199
- 3.52.** (a) 1 (b) 0.999997  
 (c)  $E(X) = 12.244$ ,  $\sigma = 2.179$

#### Paragrafo 3.9

- 3.53.** (a) 0.0844 (b) 0.0103 (c) 0.0185  
 (d) 0.1251
- 3.54.** (a) 0.7261 (b) 0.0731
- 3.55.** 0.2941
- 3.56.** (a) 0.0076 (b) 0.1462
- 3.57.** (a) 0.3679 (b) 0.0498 (c) 0.0183  
 (d) 14.9787
- 3.60.** (a) 0.1353 (b) 0.2707 (c) 5
- 3.61.** (a) 0.0409 (b) 0.1353 (c) 0.1353

#### Paragrafo 3.10

- 3.64.** (a)  $E(X) = 362$ ,  $\sigma = 19.0168$  (b) 0.2555  
 (c) 392.7799
- 3.66** (a) 0.819754 (b) 0.930563  
 (c) 0.069437 (d) 0.694974

#### Paragrafo 3.11

- 3.68.** (a) 0.2457 (b) 0.7641 (c) 0.5743 (d) 0.1848
- 3.69.** (a) 0.8404 (b) 0.4033 (c) 0
- 3.71.** 0.8740
- 3.72.** 0.988
- 3.73** 0.973675
- 3.75** (a) 0.1 (b) 0.7 (c) 0.2 (d) 0.2 (e) 0.85  
 (f) no

#### Paragrafo 3.12

- 3.77.** (a) 22 (b) 128 (c) 44 (d) 512
- 3.78.** (a)  $E(T) = 3$ ,  $\sigma_T = 0.141$  (b) 0.0169
- 3.83.**  $E(G) = 0.07396$ ,  $V(G) = 3.23 \times 10^{-7}$
- 3.84** (a) 9 (b) 1.8 (c) 19.8 (d) 0.091

#### Paragrafo 3.13

- 3.86.** (a) 0.0016 (b) 6
- 3.87.** 0.4306
- 3.88.** (a) 0.1762 (b) 0.8237 (c) 0.0005
- 3.89.** (a) 0.0791 (b) 0.1038 (c) 0.1867
- 3.90.** (a) 0.9938 (b) 1

#### Esercizi di fine capitolo

- 3.94.** (a) Esponenziale con media 12 min. (b) 0.2865  
 (c) 0.341 (d) 0.436
- 3.95.** (a) 0.0978 (b) 0.0006 (c) 0.00005
- 3.96.** (a) 33.3 (b) 22.36
- 3.97.** (a) 0.0084 (b) 0
- 3.98.** (a) 0.0018 ppm (b) 3.4 ppm
- 3.99.** 0.309
- 3.100.** (a) 0.9619 (b) 0.4305
- 3.101.** (a) 6.92 (b) 0.77 (c) 0.2188
- 3.103.** (c) 312.825 ore
- 3.104.** (b) 0.38 (c) 8.65
- 3.106.** (a) 0.105650 (b) (172.16, 187.84)

## CAPITOLO 4

#### Paragrafo 4.2

- 4.1.**  $SE = 1.04$ , Varianza = 9.7344
- 4.2.** Media = 10.989, Varianza = 51.296
- 4.5.**  $\hat{\theta}_1$  è migliore
- 4.7.** 0.5

#### Paragrafo 4.3

- 4.13.** (a) 0.0129 (b) 0.0681 (c) 0.9319
- 4.14.** (a) 13.687 (b) 0.0923 (c) 0.9077
- 4.17.** (a) 0.0574 (b) 0.265
- 4.19.** 8.85, 9.16

## Paragrafo 4.4

- 4.20.** (a)  $Z = 4$ ,  $P\text{-value} = 0$  (b) Bilaterale  
 (c)  $(30.426, 31.974)$  (d) 0

- 4.21.** (a) SE mean = 0.80,  $z = 1.75$ ,  $P = 0.0802$ ,  
 non rifiuto  $H_0$

- 4.23.** (a) 0.0324 (b) 0.0128 (c) 0.0644

- 4.24.** (a)  $z_0 = 1.77 > 1.65$ , rifiuto  $H_0$  (b) 0.04  
 (c) 0 (d) 2 (e)  $(4.003, \infty)$

- 4.25.** (a)  $P\text{-value} = 0.7188$ , non rifiuto  $H_0$   
 (b) 5 (c) 0.680542 (d)  $(87.85, 93.11)$   
 (e) Non rifiuto  $H_0$

- 4.26.** (a)  $P\text{-value} = 0.0367$ , non rifiuto  $H_0$   
 (b) 0.3632 (c) 37 (d)  $(1.994, \infty)$

- (e) Non rifiuto  $H_0$

- 4.27.** (a)  $z_0 = -26.79$ , rifiuto  $H_0$  (b)  $P\text{-value} = 0$ ,  
 rifiuto  $H_0$  (c)  $(3237.53, 3273.31)$   
 (d)  $(3231.96, 3278.88)$

- 4.28.** 97

## Paragrafo 4.5

- 4.31.** (a) 11

- (b) StDev = 1.1639, 95%  $L = (26.2853, \infty)$ ,  
 $T = 2.029$

- 4.33.** (a)  $t = 2.61$ , rifiuto  $H_0$  (c)  $(4.07, 4.56)$  (d) 4

- 4.35.** (a)  $t_0 = 1.55$ , non rifiuto  $H_0$   
 (b) No,  $d = 0.3295$  Potenza = 0.22  
 (c)  $59732.78, \infty$  (d) non rifiuto  $H_0$

- 4.36.** (a)  $t_0 = 2.14 > 1.761$ , rifiuto  $H_0$  (b)  $(5522.3, \infty)$   
 (c) Rifiuto  $H_0$

- 4.37.** (a) Normale (b)  $t_0 = 0.8735 > -1.796$ ,  
 non rifiuto  $H_0$  (c)  $(-\infty, 9.358)$   
 (d) Non rifiuto (e) 60

- 4.39.** (a)  $t_0 = 0.97 < 2.539$ , non rifiuto  $H_0$   
 (b) Normale (c)  $d = 0.42$ , potenza = 0.3  
 (d)  $d = 0.52$ , potenza = 0.9,  $n = 50$  (e)  $(23.326, \infty)$

## Paragrafo 4.6

- 4.40.** (a)  $X_0^2 = 8.96 < 23.685$ , non rifiuto  $H_0$   
 (b)  $(0.00015, \infty)$  (c) Non rifiuto  $H_0$

- 4.41.** (a)  $X_0^2 = 9.2112 < 16.919$ , non rifiuto  $H_0$   
 (b)  $0.10 < P\text{-value} < 0.50$   
 (c)  $(4.899, 877.36, \infty)$  (d) Non rifiuto  $H_0$

- 4.42.** (a)  $32.36 < X_0^2 = 55.88 < 71.42$ , non rifiuto  $H_0$   
 (b)  $0.20 < P\text{-value} < 1$  (c) 0.096, 0.2115)  
 (d) Non rifiuto  $H_0$

## Paragrafo 4.7

- 4.44.** (a) Unilaterale (b) Sì  
 (c)  $p$  campionaria = 0.69125, 95%  $L = (0.66445, \infty)$ ,  
 $P\text{-value} = 0.014286$

- 4.45.** (a)  $z_0 = 1.48$ , rifiuto  $H_0$  (b) 0.5055 (c) 136  
 (d)  $(0.254, \infty)$  (e) rifiuto  $H_0$  (f) 1572

- 4.48.** (a) 0.8288 (b) 4397

- 4.49.** (a) 0.08535 (b) 0

- 4.51.** (a)  $(0.5249, 0.8265)$  (b)  $0.5491 \leq p$

## Paragrafo 4.8

- 4.53.** (a)  $(54291.75, 68692.25)$  (b)  $(52875.64, 70108.37)$

- 4.54.** (a)  $(4929.93, 6320.27)$  (b)  $(4819.73, 6430.47)$

- 4.55.** (a)  $(7.617, 10.617)$  (b)  $(6.760, 11.474)$

## Paragrafo 4.10

- 4.57.** (a)  $X_0^2 = 5.79 < 9.49$ , non rifiuto  $H_0$  (b) 0.2154

## Esercizi di fine capitolo

- 4.60.** (a) Normale (b)  $(16.99, \infty)$  (c)  $(16.99, 33.25)$

- (d)  $(-\infty, 343.76)$  (e)  $(28.23, 343.74)$

- (f)  $(15.81, 192.44)$  (g) media:  $(16.88, 33.14)$ ,  
 varianza:  $(28.23, 343.74)$  (i)  $(-4.657, 54.897)$   
 (j)  $(-13.191, 63.431)$

- 4.61.** (a) 0.452 (b) 0.102 (c) 0.014

- 4.63.** (a)  $t_0 = 11.01$ , rifiuto  $H_0$  (b)  $P\text{-value} < 0.0005$

- (c)  $(590.95, \infty)$  (d)  $(153.63, 712.74)$

- 4.64.** (a)  $0.1 < P\text{-value} < 0.5$ , non rifiuto  $H_0$

- (b)  $0.025 < P\text{-value} < 0.05$ , rifiuto  $H_0$

- 4.67.** Media = 19.514, df = 14

## CAPITOLO 5

## Paragrafo 5.2

- 5.1.** (a) SE = 1.1692 (b) Unilaterale  
 (c) Rifiuto  $H_0$  (d)  $(0.4270, 4.2730)$

## Paragrafo 5.3

- 5.6.** (a)  $P\text{-value} = 0.001 < 0.05$ , rifiuto  $H_0$

- (b) Bilaterale (c) Rifiuto  $H_0$  (d) Rifiuto  $H_0$

- (e)  $-1.196$  (f)  $P\text{-value} < 0.001$

- 5.8.** (a)  $P\text{-value} > 0.80$ , non rifiuto  $H_0$

- (b)  $(-0.394, 0.494)$

- 5.9.** (a)  $t_0 = -2.82, < -2.101$ , rifiuto  $H_0$

- (b)  $0.01 < P\text{-value} < 0.02$  (c)  $(-0.749, -0.111)$

- 5.10.**  $P\text{-value} > 0.40$ , non rifiuto  $H_0$

## Paragrafo 5.4

- 5.14.** (a)  $\text{stDev}_{x1} = 1.5432$ ,  $\text{SE Mean}_{\text{Diff}} = 1.160$ ,  
 $t = -4.2534$ ,  $P\text{-value} = 0.002 < 0.05$ , rifiuto  $H_0$

- (b) Bilaterale (c)  $(-8.7017, -1.1635)$

- (d)  $0.001 < P\text{-value} < 0.0025$

- (e)  $P\text{-value} < 0.001$

- 5.15.** (a)  $(-1.216, 2.55)$

- 5.16.** (a)  $0.01 < P\text{-value} < 0.025$ , rifiuto  $H_0$

- (b)  $(-10.97, -0.011)$

- 5.17.**  $t_0 = -3.48, > -3.499$ , non rifiuto  $H_0$

## Paragrafo 5.5

- 5.19.  $0.248 < f_0 = 3.34 < 4.03$ , non rifiuto  $H_0$   
 5.21.  $f_0 = 0.064 < 0.4386$ , rifiuto  $H_0$   
 5.23.  $(0.245, \infty)$

## Paragrafo 5.6

- 5.25. (a) Unilaterale  
 (b)  $P\text{-value} = 0.023 < 0.05$ , rifiuto  $H_0$   
 (c) Rifiuto  $H_0$  (d)  $(0.0024, 0.14618)$   
 5.27. (a) 0.819 (b) 383

## Paragrafo 5.8

- 5.32.  $DF_{\text{Factor}} = 4$ ,  $SS_{\text{Factor}} = 987.71$ ,  $MS_{\text{Error}} = 7.46$ ,  
 $F = 33.1$ ,  $P\text{-value} < 0.01$   
 5.33. (a)  $P\text{-value} = 0$ , rifiuto  
 5.34. (a)  $P\text{-value} = 0.559$ , non rifiuto  $H_0$

## Esercizi di fine capitolo

- 5.38. (b)  $f_0 = 0.609$ , non rifiuto  $H_0$   
 5.39. (a)  $1.167 < \mu_2 - \mu_1$  (b)  $0.065 < \mu_2 - \mu_1$   
 5.41. (a)  $z_0 = 6.55 > 1.96$ , rifiuto  $H_0$   
 (b)  $z_0 = 6.55 > 2.57$ , rifiuto  $H_0$

## CAPITOLO 6

## Paragrafo 6.2

- 6.1. (a)  $\hat{y} = 0.0249 + 0.129x$   
 (c)  $SS_E = 0.000001370$ ,  $\hat{\sigma}^2 = 0.000000342$   
 (d)  $se(\hat{\beta}_1) = 0.007738$ ,  $se(\hat{\beta}_0) = 0.001786$   
 (f) 98.6% (i)  $\beta_0: (0.02, 0.03)$ ,  $\beta_1: (0.107, 0.15)$   
 (k)  $r = 0.993$ ,  $P\text{-value} = 0$   
 6.2. (a)  $\hat{y} = 0.393 + 0.00333x$   
 (c)  $SS_E = 0.0007542$ ,  $\hat{\sigma}^2 = 0.0000419$   
 (d)  $se(\hat{\beta}_1) = 0.0005815$ ,  $se(\hat{\beta}_0) = 0.04258$   
 (f) 64.5% (i)  $\beta_0: (0.304, 0.483)$ ,  
 $\beta_1: (0.00211, 0.00455)$   
 (k)  $r = 0.803$ ,  $P\text{-value} = 0$   
 6.3. (a)  $\hat{y} = 40.6 - 2.12x$  (c)  $SS_E = 13.999$ ,  $\hat{\sigma}^2 = 1.077$   
 (d)  $se(\hat{\beta}_1) = 0.2313$ ,  $se(\hat{\beta}_0) = 0.7509$

(f) 86.6% (i)  $\beta_0: (38.93, 41.18)$ , $\beta_1: (-2.62, -1.62)$ (k)  $r = -0.931$ ,  $P\text{-value} = 0$ 

- 6.4. (a) 0.055137 (b)
- $(0.054460, 0.055813)$

(c)  $(0.053376, 0.056897)$ 

- 6.5. (a) 36.095 (b)
- $(35.059, 37.131)$

(c)  $(32.802, 39.388)$ 

- 6.6. (a)
- $t_x = 9.7141$
- ,
- $P\text{-value}_x < 0.001$
- ,
- 
- $R^2_{\text{Adjusted}} = 87.78\%$
- ,
- $SS_{\text{Total}} = 4.1289$
- 
- $MS_{\text{Error}} = 0.0388$
- ,
- $S = 0.197$
- ,
- $F = 94.41$
- ,
- 
- $P\text{-value}_{\text{regression}} < 0.02$

## Paragrafo 6.3

- 6.9. (a)
- $\hat{y} = -103 + 0.605x_1 + 8.92x_2 + 1.44x_3 + 0.014x_4$

(c)  $SS_E = 1699.0$ ,  $\hat{\sigma}^2 = 242.7$ (d)  $R^2 = 74.5\%$ ,  $R^2_{\text{Adjusted}} = 59.9\%$ (f)  $se(\hat{\beta}_0) = 207.9$ ,  $se(\hat{\beta}_1) = 0.3689$ ,  
 $se(\hat{\beta}_2) = 5.301$ ,  $se(\hat{\beta}_3) = 2.392$ ,  
 $se(\hat{\beta}_4) = 0.7338$ (h)  $\beta_0: (-594.38, 388.98)$ ,  $\beta_1: (-0.267, 1.478)$ ,  
 $\beta_2: (-3.613, 21.461)$ ,  $\beta_3: (-4.22, 7.094)$ ,  
 $\beta_4: (-1.722, 1.75)$  (k)  $VIFs < 10$ 

- 6.11. (a) 287.56 (b)
- $(263.77, 311.35)$
- (c)
- $(243.69, 331.44)$

- 6.12. (a)
- $t_{x1} = 1.2859$
- ,
- $t_{x2} = 13.8682$
- ,
- 
- $0.20 < P\text{-value}_{x1} < 0.50$
- ,
- $P\text{-value}_{x2} < 0.001$
- ,
- 
- $R^2 = 88.50\%$
- ,
- $SS_{\text{Total}} = 4.1289$
- ,
- $DF_{\text{Error}} = 25$
- ,
- 
- $MS_{\text{Error}} = 0.6933$
- ,
- $s = 0.832646$
- ,
- 
- $f = 96.182$
- ,
- $P\text{-value}_{\text{regression}} < 0.02$
- 
- (b)
- $\hat{\sigma}^2 = 0.693$
- (c)
- $F_0 = 96.18 > 3.39$
- ,
- 
- rifiuto
- $H_0$
- 
- (d)
- $\beta_1: t_0 = 1.29 < 2.060$
- , non significativo
- 
- $\beta_2: t_0 = 13.87 > 2.060$
- , significativo
- 
- (e)
- $(-0.4463, 1.9297)$
- (f)
- $(7.7602, 10.45682)$

## Esercizi di fine capitolo

- 6.19. (a)
- $F_0 = 1323.62 > 4.38$
- , rifiuto
- $H_0$

- 6.20. (b)
- $\hat{y} = 2625.39 - 36.9618x$
- (c) 1886.154