Indice generale

Ross

Presentazione dell'edizione italiana

L'analisi statistica è una scienza affascinante. Essa fornisce la chiave di lettura per interpretare dati a prima vista "rumorosi" e imperscrutabili, ricavandone informazioni reali, o quanto meno attendibili. In qualche senso la statistica concilia l'esattezza dei risultati teorici con la realtà del mondo fisico, risolvendo il loro (spesso frustrante) rapporto.

Essendo io per formazione un probabilista, il lavoro di Ross mi ha molto colpito. Non si tratta ovviamente di un testo rivolto ai teorici, e, come ci si può aspettare, non indugia in un eccesso di rigore e di formalismo; tuttavia, non si concede affatto a "ricette" pronte all'uso che possano essere applicate senza avere una buona comprensione dei fenomeni statistici. La comprensione stessa dei fenomeni è il *leitmotiv* del testo. Anche il risultato più sofisticato, pure in assenza di una dimostrazione che sarebbe fuori luogo, è sempre affiancato da considerazioni sul suo significato, sulla sua plausibilità e sulla sua portata in contesti più ampi.

Allo studente, quindi, non è concesso di procedere senza capire. I problemi di fine capitolo (che sono molto numerosi), contribuiscono a conferire questo taglio. Molti di essi sono casi pratici presi dalle branche dell'ingegneria e dal mondo delle scienze pure (soprattutto la biologia); questi problemi sono caratterizzati da una grande concretezza, e richiedono oltre agli strumenti tecnici una certa visione di insieme e una dose di buon senso. Non mancano anche problemi di natura più teorica, alcuni dei quali guidano lo studente a dimostrare rigorosamente risultati di probabilità anche non banali, che vengono poi usati nel testo. Vi sono infine esercizi di livello più difficile del normale, che permettono anche al migliore degli studenti di mettere alla prova il suo livello di comprensione. (Segnalo solo il Problema 32 del Capitolo 2, che è la più ingegnosa versione del problema delle tre porte che abbia mai trovato.)

Particolarmente significativa è infine la presenza degli esempi, anch'essi molto concreti, e raramente volti alla mera illustrazione di tecniche standard. Essi sono spesso anzi arricchiti da considerazioni generali (come il riquadro sull'effetto placebo che segue l'Esempio 8.3.7) o sono di per sé utili (come l'Esempio 4.4.3, dedicato al concetto di entropia dell'informazione), contribuendo a dare al lettore una "filosofia" del corretto ragionamento statistico.

Nella traduzione italiana viene riportata sovente la terminologia inglese originale, soprattutto per i concetti di introduzione più recente, che tendono nella universalità delle applicazioni ad adottare questa lingua come standard.

Occorre menzionare delle minori variazioni di notazione, introdotte in questo adattamento, che sono il simbolo $\cap$ per l'intersezione tra eventi o insiemi, il simbolo $:=$ per definire grandezze matematiche, e l'uso delle parentesi tonde per la funzione di probabilità, come in $P(A)$, che sostituisce il meno diffuso $P\{A\}$ che era usato in originale.

Francesco Morandin

Prefazione

Questo libro è scritto per un corso introduttivo di statistica o di probabilità e statistica per studenti di ingegneria, informatica, matematica, statistica o scienze naturali. Si presuppone che lo studente possegga le basi dell'analisi matematica.

Il Capitolo 1 presenta la statistica dal punto di vista storico, e ne illustra le due branche principali, la statistica descrittiva e quella inferenziale. La prima di esse è sviluppata nel Capitolo 2, che spiega come rappresentare efficacemente un campione di dati in forma grafica o tabellare. Vengono pure introdotte delle quantità che sintetizzano i dati in un numero contenuto di informazioni significative: le statistiche campionarie.

In tutti i casi in cui si cercano informazioni su una popolazione numerosa tramite l'esame di un campione casuale ridotto, vi è una certa aleatorietà nell'esperimento, e di conseguenza anche nelle conclusioni a cui si giunge. La teoria della probabilità è quindi indispensabile a formalizzare le conclusioni dell'inferenza statistica, ed è necessario che lo studente ne acquisisca le basi. Quest'ultimo è l'obiettivo del Capitolo 3, che introduce l'idea di esperimento probabilistico, illustra il concetto di probabilità di un evento e presenta gli assiomi della probabilità. Tale studio prosegue e viene sviluppato nel Capitolo 4, che si occupa dei fondamentali concetti di variabile aleatoria e di speranza matematica, e nel Capitolo 5, che passa in rassegna alcuni tipi speciali di variabili aleatorie che emergono spesso nelle applicazioni. Vengono definite le variabili aleatorie binomiali, di Poisson, ipergeometriche, normali, uniformi, gamma, chi-quadro, le t di Student e le F di Fisher.

Nel Capitolo 6 studiamo la distribuzione di statistiche campionarie come la media e la varianza campionarie. Mostriamo come usare un notevole risultato della teoria della probabilità, il teorema del limite centrale, per approssimare la distribuzione di probabilità della media campionaria. Inoltre discutiamo la distribuzione congiunta di media e varianza campionaria nel caso fondamentale in cui i dati provengano da una popolazione gaussiana.

Il Capitolo 7 mostra come usare i dati per stimare parametri di interesse. Pensiamo ad uno studioso che voglia determinare la frazione dei laghi statunitensi soggetta a piogge acide. Vi sono due tipologie di stimatori sostanzialmente diverse, che si possono considerare. Nel primo caso si stima la quantità in questione con un singolo numero (per esempio si potrebbe ottenere che il 47% circa dei laghi è interessato da piogge acide), mentre nel secondo si ricava una stima che ha la forma di un intervallo

di valori (nel nostro esempio si potrebbe trovare che la percentuale di laghi colpiti da piogge acide cade tra il 45% ed il 49%). Il secondo tipo di stimatori ci dice anche il "livello di confidenza" che possiamo avere sulla loro validità. Infatti mentre è quasi impossibile che il valore reale coincida *precisamente* con quello da noi stimato inizialmente (47%), un intervallo di valori ci consente una maggiore sicurezza, e possiamo avere una certa confidenza (ad esempio del 95%) che la percentuale effettiva sia compresa tra il 45% ed il 49%.

Il Capitolo 8 presenta i test di ipotesi, un settore importante che riguarda l'utilizzo dei dati per verificare la plausibilità di ipotesi definite in precedenza. Un esempio di ipotesi statistica valida potrebbe essere che meno del 44% dei laghi americani sia soggetto a piogge acide, e il test su un campione di quei laghi potrebbe permettere di escluderla, oppure accettarla. Viene quindi introdotto il concetto di *p*-dei-dati, una grandezza che misura il grado di plausibilità dell'ipotesi assegnata, dopo l'osservazione dei dati. Vengono presi in considerazione diversi tipi di test di ipotesi, in particolare quelli riguardanti media e varianza di una o due popolazioni normali, e quelli sui parametri delle distribuzioni di Bernoulli e di Poisson.

Il Capitolo 9 si occupa della regressione. Vengono trattate sia la regressione lineare semplice, sia quella multipla, approfondite con lo studio dei residui, tecniche di linearizzazione, minimi quadrati pesati e cenni storici sul fenomeno del regressione alla media di Galton.

Il Capitolo 10 introduce l'analisi della varianza. Vengono considerati sia i problemi ad una via sia quelli a due vie (con o senza interazione).

Il Capitolo 11 riguarda i test di adattamento, che possono essere usati per verificare se il modello proposto sia compatibile coi dati. Il test classico del chi-quadro viene presentato e applicato alla verifica dell'indipendenza in tabelle di contingenza. La sezione finale del capitolo presenta il test di Kolmogorov-Smirnov, che si usa per verificare se i dati provengano da una distribuzione continua assegnata.

Il Capitolo 12 affronta i test di ipotesi non parametrici, che possono essere impiegati quando non si è in grado di stabilire la particolare classe (ad esempio normale, o esponenziale) della distribuzione originale dei dati.

Il Capitolo 13 considera il controllo di qualità, una tecnica statistica fondamentale per i processi di fabbricazione e produzione. Vengono affrontate diverse carte di controllo di Shewhart, e anche alcune più sofisticate, basate sulle medie mobili e le somme cumulate.

Il Capitolo 14 affronta l'inferenza sul tempo di vita dei sistemi. In questo ambito è la distribuzione esponenziale piuttosto che la normale ad avere un ruolo chiave.

Sul sito web dedicato a questo libro (www.apogeonline.com/libri/00897/allegati/) è disponibile un software statistico liberamente scaricabile e che può essere usato per risolvere la gran parte dei problemi di statistica del testo. Il software è formato

da una collezione di programmi[1]. Una prima parte di essi consente di calcolare il *p*-dei-dati per la maggior parte dei test di ipotesi, compresi quelli sull'analisi della varianza e la regressione. Altri permettono di ottenere le probabilità che definiscono le più importanti distribuzioni[2]. Un ultimo programma infine ha lo scopo di illustrare il Teorema del Limite Centrale; esso considera variabili aleatorie che assumono i valori 0, 1, 2, 3 e 4 con probabilità che sono assegnate dall'utente assieme ad un intero n, e visualizza la funzione di massa di probabilità della somma di n variabili aleatorie indipendenti con questa distribuzione. Facendo crescere n si può "vedere" la funzione di massa convergere alla forma tipica di una densità di probabilità gaussiana.

[1] Per il corretto funzionamento del software statistico abbinato al libro, è necessario impostare Microsoft Windows in modo che il separatore decimale sia il punto, e non la virgola, che è l'impostazione predefinita nell'installazione del sistema operativo in italiano, [*N.d.T.*]

[2] Per chi non ha accesso ad un personal computer o al world wide web, nell'Appendice in fondo al libro sono comunque incluse tabelle che possono essere usate per risolvere quasi tutti i problemi del testo.

1 Una introduzione alla statistica

Contenuto

La raccolta dei dati e la loro analisi sono strumenti indispensabili per capire a fondo la complessa realtà che ci circonda. La *statistica* è l'arte di apprendere dai dati. Essa si occupa della loro raccolta, della loro descrizione e della loro analisi, guidandoci nel trarre le conclusioni.

1.1 Raccolta dei dati e statistica descrittiva

Alcune volte l'analisi statistica parte da un campione di dati definito. Ad esempio, lo stato raccoglie e pubblicizza regolarmente i dati riguardanti le precipitazioni, le scosse telluriche, il livello di disoccupazione, il prodotto interno lordo ed il tasso di inflazione. La statistica permette di descrivere, sintetizzare ed analizzare questi dati.

In altri casi il lavoro dello statistico inizia prima che i dati siano stati ottenuti, e il suo primo obiettivo consiste nell'ideare un procedimento ottimale per la loro raccolta. Immaginiamo ad esempio che un docente voglia determinare quale sia il più efficace tra due diversi metodi per insegnare la programmazione a dei neofiti. Un possibile approccio consiste nel dividere gli studenti in due gruppi e usare un diverso metodo didattico per ciascun gruppo. Alla fine del corso gli studenti vengono esaminati e i punteggi dei membri dei due gruppi sono confrontati. Se i risultati di uno dei due gruppi risultassero notevolmente più alti, sarebbe ragionevole pensare che il corrispondente metodo di insegnamento sia migliore.

È importante notare, a questo proposito, che per poter trarre delle conclusioni valide dai dati è essenziale che gli studenti siano divisi in modo che in nessuno dei due gruppi si vengano a trovare elementi con una maggiore predisposizione alla programmazione. Quindi, ad esempio, il docente dovrebbe evitare di mettere i maschi

in un gruppo e le femmine nell'altro, perché in tal caso, anche dove risultasse che le femmine hanno ottenuto punteggi più alti, non sarebbe chiaro se questo sia dovuto al metodo usato per istruirle o ad una loro innata predisposizione nella capacità di programmare.

Il metodo accettato per superare questo problema consiste nel dividere "a caso" gli studenti in due gruppi. Più precisamente la suddivisione va scelta tra tutte quelle possibili, con uguale probabilità.

Ad esperimento concluso, i dati devono essere raccolti e commentati. Nel nostro esempio saranno presentati i punteggi dei due gruppi, congiuntamente a quantità riassuntive, come le medie relative a ciascun gruppo. Quella parte della statistica che si occupa di illustrare e sintetizzare i dati è detta *statistica descrittiva.*

1.2 Inferenza statistica e modelli probabilistici

Continuando l'esempio, dopo che la prova si è conclusa ed i dati sono stati illustrati e sintetizzati, vorremmo poter trarre una conclusione su quale dei due metodi di insegnamento sia superiore. La parte della statistica che si occupa di questo aspetto è detta *inferenza statistica.*

Per dedurre enunciati formalmente validi dai dati raccolti, è necessario che prendiamo in considerazione l'influenza del caso. Ad esempio, potrebbe darsi che il punteggio medio dei membri del primo gruppo sia superiore, ma di poco, a quello del secondo gruppo. È corretto allora concludere che questa differenza sia dovuta al metodo didattico utilizzato? Oppure è possibile che così non sia, ed essa vada piuttosto imputata ad una casualità? Citando un altro esempio, il fatto che una moneta lanciata 10 volte abbia dato 7 volte testa, non significa necessariamente che ci si debba aspettare che nei prossimi lanci sia più probabile ottenere testa piuttosto che croce. È chiaramente plausibile che si tratti di una moneta perfettamente normale che, per caso, ha dato testa in 7 tiri su 10. (D'altro canto, se la moneta avesse realizzato 47 teste su 50 tiri, potremmo essere quasi certi che non si tratti di una moneta del tutto normale.)

Per poter giungere a conclusioni pienamente giustificate, è allora necessario fare alcune assunzioni sulla *probabilità* che i dati che andiamo a misurare assumano i diversi valori possibili. L'insieme di queste ipotesi è detto *modello probabilistico* per i dati.

A volte la natura dei dati suggerisce la forma del modello probabilistico da adottare. Consideriamo ad esempio un ingegnere della produzione che voglia scoprire la frazione di circuiti integrati difettosi riscontrata con un nuovo metodo produttivo. Egli potrebbe selezionere un certo numero di questi chip e testarli; il numero di quelli difettosi costituirà il dato sperimentale. Se la scelta del campione da testare è stata fatta "a caso", è ragionevole supporre che ciascuno dei chip sarà difettoso con probabilità p pari alla frazione incognita di chip difettosi sul totale di quelli prodotti. Il dato misurato può allora essere usato per fare delle inferenze su p.

In altre situazioni potrebbe non essere evidente quale sia il modello di probabilità appropriato per un certo campione di dati. Molto spesso tuttavia, una attenta descrizione e presentazione dei dati ci permette di inferire quale sia un modello accettabile, che può eventualmente essere messo alla prova raccogliendo nuovi dati.

Siccome l'inferenza statistica si basa sull'individuazione del corretto modello di probabilità che descrive i dati, una certa conoscenza della teoria della probabilità risulta indispensabile alla comprensione della statistica stessa. L'inferenza statistica si basa sul presupposto che importanti aspetti del fenomeno sotto studio possano essere descritti in termini di probabilità; utilizza quindi i dati per fare inferenze su queste probabilità.

1.3 Popolazioni e campioni

La statistica è normalmente interessata ad ottenere informazioni su un insieme completo di oggetti che viene detto *popolazione.* Esso è spesso troppo grande perché sia possibile un esame esaustivo: esempi comuni sono i residenti di una certa regione, i televisori prodotti da una azienda, oppure i nuclei familiari con un certo livello di reddito. In tutti questi casi, si cerca di imparare qualcosa sulle popolazioni scegliendo e poi esaminando dei sottogruppi di loro elementi. Un sottogruppo della popolazione è detto *campione.*

Siccome il campione deve contenere informazioni sulla popolazione complessiva, deve essere (in qualche senso) rappresentativo di quella popolazione. Se ad esempio fossimo interessati alla distribuzione delle età degli abitanti di un certo comune, e, intervistati i primi 100 che entrano in una biblioteca, trovassimo una media di 46.2 anni, saremmo giustificati a concludere che questa è approssimativamente l'età media dell'intera popolazione? Probabilmente no; infatti si può obiettare che il campione prescelto non è rappresentativo della popolazione in esame, essendo gli utenti della biblioteca più facilmente studenti ed anziani che non persone in età lavorativa.

A volte, come nell'esempio della biblioteca, ci viene fornito un campione, e sta a noi stabilire se sia rappresentativo o meno dell'intera popolazione. Si tenga presente che in generale, solo campioni scelti completamente a caso sono certamente rappresentativi; infatti ogni criterio di selezione non casuale finisce con il produrre campioni che sono automaticamente sbilanciati verso valori particolari.

Perciò, anche se sembra paradossale, abbiamo le migliori possibilità di ottenere un campione rappresentativo quando scegliamo i suoi membri in modo completamente casuale, senza alcuna considerazione a priori sugli elementi da prendere. In particolare non è opportuno costruire deliberatamente un campione che contenga, ad esempio, la stessa percentuale di femmine e la stessa percentuale di occupati per

Tabella 1.1 Numero totale di decessi in Inghilterra

Anno	Decessi	Di cui per la peste
1592	25 886	11 503
1593	17 844	10 662
1603	37 294	30 561
1625	51 758	35 417
1636	23 359	10 400

Fonte: John Graunt, Observations Made upon the Bills of Mortality. 3rd ed. London: John Martyn and James Allestry (1st ed. 1662).

ciascun impiego che troveremmo nella popolazione totale. È preferibile piuttosto lasciare che il caso ci faccia ottenere approssimativamente le percentuali corrette.

1.4 Una breve storia della statistica

La raccolta sistematica di dati sulla popolazione e sull'economia ebbe origine a Venezia e a Firenze durante il Rinascimento. Il termine *statistica* deriva dalla parola *stato*, in quanto indicava una raccolta di fatti di interesse per lo stato. L'idea di raccogliere dati si diffuse dall'Italia a tutta l'Europa occidentale, ed entro la prima metà del sedicesimo secolo era generalmente diffusa la consuetudine, presso i governi europei, di richiedere alle parrocchie di registrare nascite, matrimoni e morti. A causa delle tragiche condizioni di salute pubbliche, quest'ultima statistica era di particolare importanza.

Fino al diciannovesimo secolo, l'alta mortalità registrata in Europa era principalmente dovuta a malattie epidemiche, guerre e carestie. Tra le epidemie, la peggiore era la peste. A cominciare dalla Peste Nera del 1348, la peste comparve spesso per quasi 400 anni. Nel 1562 la città di Londra cominciò a pubblicare settimanalmente dei bollettini di mortalità, nel tentativo di tenere aggiornata la corte reale, che stava considerando un trasferimento in campagna. All'inizio questi bollettini elencavano solo il luogo dei decessi e se si trattasse di morte per peste. Dal 1625 però furono estesi a comprendere anche le altre cause di decesso.

Nel 1662 il commerciante inglese John Graunt pubblicò un libro dal titolo *Natural and Political Observation Made upon the Bills of Mortality*. La Tabella 1.1 è stata estratta da tale libro; elenca il numero annuale di decessi in Inghilterra e quanti di essi furono imputati alla peste, per cinque diversi anni di diffusione del contagio.

Graunt pensò di utilizzare i bollettini di mortalità per stimare la popolazione di Londra. Per stimare quella del 1660, ad esempio, Graunt fece delle ricerche in alcune parrocchie e sulle famiglie di vari quartieri, e scoprì che in media c'erano stati quell'anno circa 3 morti ogni 88 persone. Dividendo per 3 si trova un decesso ogni 88/3 abitanti. Siccome i bollettini riportavano 13 200 morti per Londra quell'anno,

Graunt stimò che la popolazione complessiva di Londra fosse di circa

$$13\,200 \times 88/3 = 387\,200$$

abitanti. Graunt impiegò questa stima per fare proiezioni sull'intera Inghilterra. Nel suo libro annotò che queste cifre potevano interessare ai governanti del paese, in quanto indicatori sia del numero di uomini che potevano essere coscritti, sia del numero di quelli che potevano essere tassati.

Graunt riuscì anche ad impiegare questi dati – ed un po' di intelligenti supposizioni su quali malattie sono mortali alle diverse età – per stimare le età al momento dei decessi. (Si ricordi che i bollettini elencavano solo luoghi e cause delle morti e non le età dei deceduti.) Utilizzò quindi queste informazioni per compilare delle tabelle che davano la percentuale di popolazione che muore alle diverse età. La Tabella 1.2 è una di queste tabelle di mortalità. Essa dice che su 100 nati, 36 morivano prima di arrivare a 6 anni, 24 morivano tra i 6 ed i 15 anni e così via.

La stima della speranza di vita era di grande interesse per coloro che si occupavano di rendite vitalizie. Queste ultime sono l'opposto delle assicurazioni sulla vita, poiché inizialmente si versa una somma come investimento e si ha poi diritto alla riscossione di pagamenti regolari per tutta la durata della vita rimanente.

Il lavoro di Graunt sulle tabelle di mortalità ispirò nel 1693 le ricerche di Edmund Halley. Halley, lo scopritore dell'omonima cometa (nonché la persona che permise, con incoraggiamenti e supportandola finanziariamente, la pubblicazione dei *Principia Mathematica* di Isaac Newton), usò le tabelle di mortalità per stabilire con che probabilità una persona di una data età sarebbe vissuta fino ad un qualunque numero di anni. Halley con la sua influenza riuscì a convincere le compagnie assicuratrici che i premi delle assicurazioni dovevano dipendere dall'età dell'assicurato.

Dopo Graunt e Halley, la raccolta di dati si accrebbe stabilmente per tutto il resto del diciassettesimo e durante il diciottesimo secolo. Anche Parigi nel 1667 iniziò

Tabella 1.2 Tabella delle mortalità di John Graunt

(Le classi di età arrivano fino all'estremo destro escluso. Ad esempio 0–6 significa tutte le età dagli 0 ai 5 anni.)

Tempo di vita	Numero di decessi su 100 nascite
0–6	36
6–16	24
16–26	15
26–36	9
36–46	6
46–56	4
56–66	3
66–76	2
76 o più	1

a registrare i decessi e nel 1730 era ormai pratica comune in tutta Europa annotare anche le età in cui avvenivano.

Il termine *statistica*, che per tutto il diciottesimo secolo veniva usato come abbreviazione di scienza descrittiva dello stato, dal secolo successivo iniziò ad essere associato ai numeri. Entro il 1830 era diventato sinonimo di "scienza numerica" della società. Questo cambiamento di significato fu consentito dalla vasta disponibilità di registrazioni censuarie ed altri dati che, a partire dal 1800 circa, vennero raccolti sistematicamente dai governi dell'Europa occidentale e dagli Stati Uniti.

Durante il diciannovesimo secolo, anche se la teoria della probabilità era stata sviluppata da matematici come Jacob Bernoulli, Karl Friedrich Gauss e Pierre-Simon Laplace, il suo uso per studiare risultati statistici era praticamente inesistente, dato che molti statistici di quel tempo sostenevano l'autoevidenza dei dati. In particolare essi non erano tanto interessati a fare inferenza su singoli, quanto sulla società nel suo insieme, e per questo non studiavano campioni statistici, ma cercavano di ottenere dati sempre più completi dell'intera popolazione. L'inferenza probabilistica da un campione alla popolazione era quasi del tutto ignota alla statistica sociale di quel secolo.

Negli ultimi anni dell'800, la statistica iniziò ad occuparsi di inferire conclusioni a partire da dati numerici. Tra i fautori di questo approccio vanno ricordati Francis Galton, il cui lavoro di analisi sull'ereditarietà dell'intelligenza introdusse ciò che ora chiamiamo regressione e analisi della correlazione (si veda il Capitolo 9), e Karl Pearson. Pearson sviluppò il test del chi-quadro per verificare la bontà di un fit (si veda il Capitolo 11), e fu il primo direttore del Laboratorio Galton, fondato per donazione di Francis Galton nel 1904. Qui Pearson organizzò un programma di ricerca mirato allo sviluppo di nuovi metodi per la statistica e l'inferenza. Vi si accoglievano studenti avanzati di materie scientifiche ed industriali che venivano ad imparare le tecniche statistiche per poterle poi applicare nei loro campi. Uno dei primi ricercatori ospiti dell'istituto fu W. S. Gosset, un chimico di formazione, che dimostrò la sua devozione a Pearson pubblicando i propri lavori sotto lo pseudonimo di "Student". (Altri sostengono che Gosset non volesse pubblicare con il suo vero nome per timore che i suoi datori di lavoro alla fabbrica di birra Guinness non avrebbero approvato che uno dei loro chimici facesse ricerche di statistica.) Gosset è celebre per aver sviluppato la teoria del test t (si veda il Capitolo 8).

I due campi di maggiore importanza per la statistica applicata dell'inizio del ventesimo secolo erano la biologia delle popolazioni e l'agricoltura, e ciò era dovuto al personale interesse dello stesso Pearson e di altri nel laboratorio, come pure ai notevoli risultati dello scienziato inglese Ronald A. Fisher. La teoria dell'inferenza sviluppata da questi pionieri (tra i quali citiamo anche il figlio di Karl Pearson, Egon, ed il matematico di origini polacche Jerzy Neyman) era abbastanza generale da adattarsi ad un gran numero di problemi quantitativi e pratici. Per questo, dopo i primi

Tabella 1.3 L'evoluzione nelle definizioni di statistica

La statistica ha quindi per suo oggetto quello di presentare una fedele rappresentazione di uno stato in una determinata epoca. (Quetelet, 1849)
Le statistiche sono gli unici strumenti tramite i quali è possibile aprire una breccia nella formidabile barriera di difficoltà che blocca il cammino di chi ricerca la Scienza dell'uomo. (Galton, 1889)
La statistica può essere vista come (i) lo studio delle popolazioni, (ii) lo studio della variabilità, (iii) lo studio dei metodi di riduzione dei dati. (Fisher, 1925)
La statistica è una disciplina scientifica che si occupa della raccolta, analisi ed interpretazione dei dati ottenuti da osservazioni sperimentali. Questa materia ha una struttura coerente che si basa sulla teoria della probabilità e include molte tecniche differenti che si affiancano alla ricerca e allo sviluppo in tutti i campi della Scienza e della Tecnologia. (E. Pearson, 1936)
Statistica è il nome della scienza nonché arte che si occupa delle inferenze non certe - che impiega i numeri per dare risposte sulla natura e sull'esperienza. (Weaver, 1952)
La statistica è stata riconosciuta nel ventesimo secolo come lo strumento matematico capace di analizzare i dati degli esperimenti e quelli osservati in ogni contesto. (Porter, 1986)
La statistica è l'arte di apprendere dai dati. (Questo libro, 1999)

anni del secolo, un numero rapidamente crescente di persone che si occupavano di scienze, affari e governo incominciarono a considerare la statistica come il principale strumento capace di fornire risposte quantitative a problemi scientifici e pratici (si veda la Tabella 1.3).

Attualmente gli accenni alla statistica sono ovunque. In tutti i quotidiani e le riviste vi sono esempi di statistica descrittiva. L'inferenza statistica invece è divenuta indispensabile per la salute dell'uomo e la ricerca medica, per l'ingegneria e gli studi scientifici, per il marketing ed il controllo di qualità, per l'istruzione, per la contabilità, l'economia, le previsioni meteorologiche, per i sondaggi e le inchieste, per gli sport, le assicurazioni, il gioco e per tutti i tipi di ricerca che abbiano delle pretese di scientificità. La statistica è senza dubbio divenuta parte integrante della nostra eredità culturale.

Problemi

1. La prossima settimana si terranno le elezioni presidenziali americane, ed intervistando un campione di elettori vorremmo stabilire se prevarrà il candidato repubblicano o quello democratico. Quale dei seguenti metodi di selezione produrrà più facilmente un campione rappresentativo?

   (a) Intervistare tutti gli spettatori di maggiore età ad una partita di basket tra college.

(b) Intervistare tutte le persone di maggiore età che escono da un lussuoso ristorante del centro.

(c) Ottenere una copia dell'elenco degli elettori, sceglierne 100 a caso ed intervistarli.

(d) Usare i risultati di un sondaggio televisivo basato sulle telefonate dei telespettatori.

(e) Scegliere dei nomi dall'elenco telefonico e intervistare queste persone.

2. L'approccio suggerito nel punto (e) del Problema 1 portò ad una predizione disastrosa in occasione delle elezioni presidenziali americane del 1936, quando Franklin Roosevelt ottenne una vittoria schiacciante su Alfred Landon. La rivista *Literary Digest* predisse infatti la vittoria di Landon, basandosi sulle preferenze di un campione di elettori che era stato scelto sugli elenchi del telefono e tra i proprietari di automobili.

   (a) Per quale motivo la predizione fu tanto scorretta?

   (b) Dal 1936 ad oggi è intervenuto qualche cambiamento che autorizzi a pensare che un approccio di questo tipo sia ora più affidabile di allora?

3. Un ricercatore vuole determinare l'aspettativa di vita attuale negli Stati Uniti. Come campione di dati, legge per 30 giorni i necrologi di un importante quotidiano nazionale e annota le età di tutti i decessi. È rappresentativo il campione ottenuto con questo criterio?

4. Per determinare la percentuale di fumatori tra i residenti di un comune, si decide di fare un sondaggio, intervistando i frequentatori di uno dei luoghi seguenti:

   (a) una piscina;

   (b) un bowling;

   (c) un centro commerciale;

   (d) una biblioteca.

   Quale di essi ha più chances di fornire una buona approssimazione della percentuale in esame? Perché?

5. Una università conduce un'indagine per determinare il reddito annuale medio dei suoi laureati recenti. Vengono selezionati 200 laureati degli ultimi anni ai quali viene inviato un questionario con domande sul loro impiego attuale. Dei 200 questionari però, solo 86 vengono restituiti. Il reddito annuale medio che ne risulterà è di 75 000 dollari.

   (a) È corretto pensare che 75 000 dollari sia una buona approssimazione del reddito medio di tutti i suoi laureati recenti? Giustifica la risposta.

   (b) Se la risposta data al punto (a) è stata no, di che diversa categoria di laureati questa cifra è allora una approssimazione rappresentativa del reddito?

6. Su un articolo di giornale compare la seguente statistica: l'80% dei pedoni vittime di incidenti stradali notturni indossava abiti scuri, mentre il restante 20% indossava abiti chiari. L'autore dell'articolo conclude che è più sicuro vestirsi in abiti chiari se si esce a piedi la sera.

   (a) Questa conclusione è giustificata? Perché?

   (b) Se la risposta data al punto (a) è stata no, di quali altre informazioni dovremmo poter disporre prima di trarre qualunque conclusione?

7. Analizza criticamente il metodo usato da Graunt per stimare la popolazione di Londra. Che cosa viene assunto implicitamente?

8. I bollettini di mortalità di Londra riportano 12 246 decessi per il 1658. Supponendo che una inchiesta nelle parrocchie avesse rivelato che quell'anno era deceduto circa il 2% della popolazione, usa il metodo di Graunt per stimare la popolazione complessiva della città.

9. Immagina di impersonare un venditore di rendite vitalizie del 1662, l'anno di pubblicazione del libro di Graunt. Spiega come si potrebbero utilizzare i dati di Graunt sulle età alla morte.

10. Basandoti sulla tabella di mortalità di Graunt, rispondi alle seguenti domande.

    (a) Quale frazione della popolazione raggiungeva l'età di 6 anni?

    (b) Quale frazione raggiungeva i 46 anni?

    (c) Quale frazione moriva tra i 6 ed i 45 anni?

2 Statistica descrittiva

Contenuto

2.1 Introduzione

In questo capitolo presentiamo e sviluppiamo la statistica descrittiva, la branca delle scienze statistiche che si occupa dei metodi di esposizione e sintesi dei dati. La Sezione 2.2 è dedicata alla rappresentazione degli insiemi di dati; la 2.2.1 e la 2.2.2 si occupano di campioni poco numerosi, discutendo i tipi di grafici e di tabelle utili alla loro presentazione; la 2.2.3 spiega secondo quali criteri conviene raggruppare in intervalli di valori i campioni più numerosi. La Sezione 2.3 discute come si possono ottenere informazioni sintetiche su un campione sperimentale introducendo le *statistiche*, che sono grandezze numeriche calcolabili dai dati. Tra le statistiche più utili vi sono le tre che indicano il "centro" dei dati (media, mediana e moda campionarie, descritte nella Sezione 2.3.1) e le due che quantificano la loro dispersione (varianza e deviazione standard campionarie, nella Sezione 2.3.2). La Sezione 2.3.3 definisce i percentili, statistiche che dicono – ad esempio – quale valore è maggiore del 95% dei dati. Nella Sezione 2.4 viene presentata la disuguaglianza di Chebyshev (nella versione campionaria). Questa celebre disuguaglianza fornisce un limite superiore alla frazione di dati di un campione che si allontanano dalla loro media campionaria più di un multiplo della deviazione standard. La disugualgianza di Chebyshev vale per tutti gli insiemi di dati, e ci sono situazioni in cui questo limite può essere notevolmente migliorato; nella Sezione 2.5 discutiamo infatti i campioni normali, caratterizzati da

Tabella 2.1 Stipendi annuali iniziali. Dati in migliaia di dollari.

Stipendio iniziale	Frequenza
27	4
28	1
29	3
30	5
31	8
32	10
34	5
36	2
37	3
40	1

un grafico della distribuzione a forma di campana, e da una regola empirica che fornisce stime più precise della citata disuguaglianza. La Sezione 2.6 si occupa infine dei campioni formati da coppie di valori tra loro (eventualmente) legate. Vengono presentate due semplici tecniche per valutare la relazione esistente tra i due tipi di dati: il diagramma a dispersione, che è un approccio visivo, e il coefficiente di correlazione campionaria, una statistica che misura il grado di corrispondenza di valori elevati del primo tipo di dati con valori elevati del secondo.

2.2 Organizzazione e descrizione dei dati

I risultati numerici di una ricerca dovrebbero essere sempre presentati in maniera chiara, concisa, e in modo da dare rapidamente al lettore un'idea generale delle loro caratteristiche globali. Nel corso degli anni sono state selezionate un certo numero di tecniche di rappresentazione tabellari e grafiche che sono ormai accettate universalmente e che hanno il pregio di evidenziare aspetti come il supporto, la simmetria e il grado di concentrazione dei dati. In questa sezione saranno affrontate alcune di quelle più diffuse.

2.2.1 Tabelle e grafici delle frequenze

Dei dati che si suddividano in un numero relativamente basso di valori distinti possono essere convenientemente rappresentati in una tabella tramite le loro *frequenze*. Ad esempio, la Tabella 2.1 raccoglie lo stipendio annuale iniziale di 42 ingegneri neolaureati. Possiamo evincerne, tra le altre cose, che lo stipendio minimo è stato di 27 000 dollari, e ha interessato 4 ingegneri, mentre lo stipendio massimo, di 40 000 dollari, è toccato a uno solo. La cifra più comune è stata di 32 000 dollari, ed è stata percepita da 10 ingegneri.

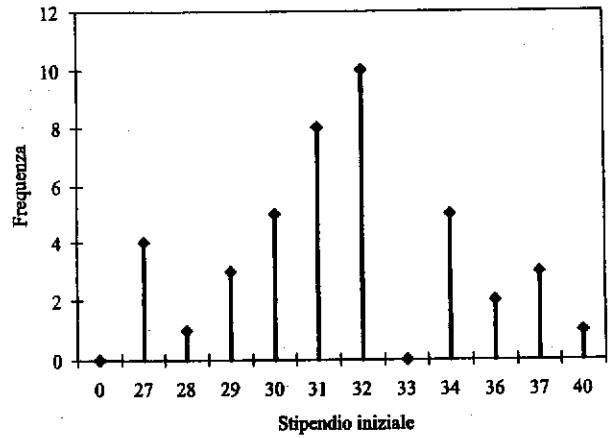


Figura 2.1 Grafico a bastoncini – *line graph*

Per rappresentare graficamente la distribuzione delle frequenze di un insieme di dati di questo tipo si usano, tra gli altri, i tre tipi di grafici di cui sono illustrati esempi nelle Figure 2.1, 2.2, 2.3, che sono basati sui dati salariali della Tabella 2.1.

In tutti i casi, sull'asse delle ascisse sono indicati i diversi valori che compaiono come dati; sull'asse delle ordinate vi è invece la frequenza di ciascun valore. Se essa è rappresentata da linee verticali, la figura prende il nome di *grafico a bastoncini*, come in Figura 2.1. Se alle linee viene dato spessore fino a farle divenire rettangoli adiacenti, si parla di *grafico a barre*, come in Figura 2.2. Infine se, come in Figura 2.3 i punti del grafico sono uniti in una spezzata, si parla di *grafico a linee* oppure di *poligonale*[1].

2.2.2 Tabelle e grafici delle frequenze relative

Consideriamo un insieme di n dati numerici. Se f è la frequenza di uno dei valori che vi compaiono, allora il rapporto f/n si dice la sua *frequenza relativa*. Quindi la frequenza relativa di un valore è la frazione di volte che esso compare nell'insieme di dati. È possibile rappresentare la distribuzione di un campione di dati tramite un grafico delle loro frequenze relative, esattamente come si fa per le frequenze assolute; si possono in particolare usare grafici a bastoncini, a barre e a linee. Come ci si

[1] La terminologia inglese è sempre importante, e quella corrispondente a queste definizioni va menzionata in quanto si presta particolarmente a confusione, poiché con *line graph* si intende il grafico a bastoncini, mentre quello che noi chiamiamo a linee si dice *polygon graph*. Uguale invece l'ultima dicitura, che diviene *bar graph*, [*N.d.T.*]

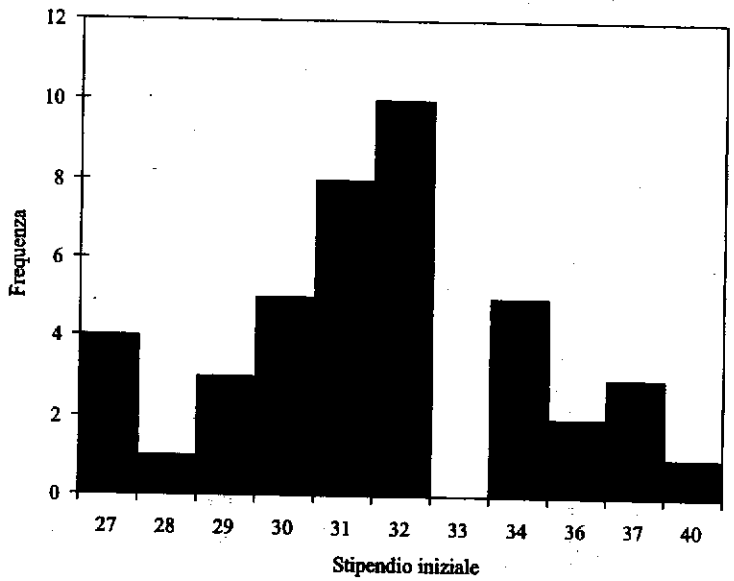


Figura 2.2 Grafico a barre

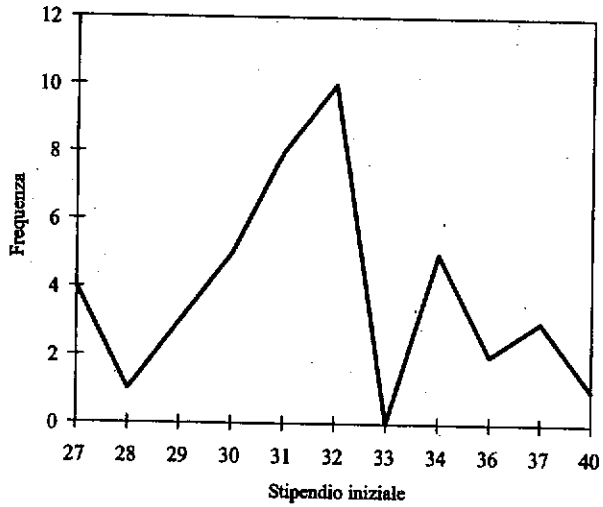


Figura 2.3 Grafico a linee – *polygon*

aspetta, questi grafici saranno identici a quelli delle frequenze assolute, ma con i valori delle ordinate riscalati di un fattore $1/n$.

Esempio 2.2.1. La Tabella 2.2 riporta le frequenze relative dei dati della Tabella 2.1. In ogni riga, si è semplicemente diviso il valore di frequenza assoluta per il numero totale di dati, che era 42, ottenendo le frequenze relative corrispondenti. □

Tabella 2.2 Redditi annuali iniziali. Dati in migliaia di dollari.

Stipendio iniziale	Frequenza relativa
27	4/42 ≈ 0.0952 = 9.52%
28	1/42 ≈ 0.0238 = 2.38%
29	3/42 ≈ 0.0714 = 7.14%
30	5/42 ≈ 0.1190 = 11.90%
31	8/42 ≈ 0.1905 = 19.05%
32	10/42 ≈ 0.2381 = 23.81%
34	5/42 ≈ 0.1190 = 11.90%
36	2/42 ≈ 0.0476 = 4.76%
37	3/42 ≈ 0.0714 = 7.14%
40	1/42 ≈ 0.0238 = 2.38%

Un altro tipo di rappresentazione grafica tra le più comuni è il *grafico a torta*, utile in particolare quando i dati non sono numerici ma categorici. Si costruisce tracciando un cerchio e suddividendolo in tanti settori circolari (le fette o spicchi) quante sono le categorie distinte di dati, ogni settore con un angolo al centro proporzionale alla frequenza (relativa o assoluta è lo stesso) della categoria corrispondente.

Esempio 2.2.2. I dati nella tabella che segue riguardano i vari tipi di tumore riscontrati negli ultimi 200 pazienti entrati in una clinica oncologica. Essi sono rappresentati nel grafico a torta in Figura 2.4. □

Tipo di tumore	Numero di casi	Frequenza relativa
Polmoni	42	0.210
Seno	50	0.250
Colon	32	0.160
Prostata	55	0.275
Melanoma	9	0.045
Vescica	12	0.060

2.2.3 Raggruppamento dei dati, istogrammi, ogive e diagrammi stem and leaf

Le metodologie sviluppate nelle Sezioni 2.2.1 e 2.2.2 sono utili soprattutto nel caso che i dati da esaminare abbiano un numero di valori distinti non troppo numeroso.

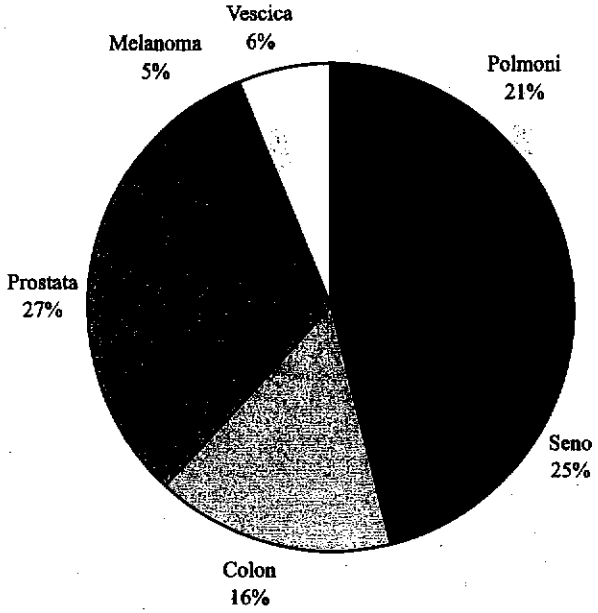


Figura 2.4 Grafico a torta

Quando il campione di dati non soddisfa questo requisito, sorge spontanea l'idea di dividere i dati in gruppi di valori contigui, o *classi*, e poi presentare con grafici e tabelle il numero di dati che cadono nell'intervallo di valori assegnato a ciascuna classe. La scelta di quante classi adottare è un fattore importante, infatti (1) da un lato se si prendono poche classi si perde troppa informazione sulla posizione che avevano i dati all'interno degli intervalli di classe, (2) dall'altro, con troppe classi le frequenze di ciascuna assumerebbero valori troppo piccoli e diventerebbe difficile riconoscere la forma della distribuzione. Anche se valori tipici per il numero di classi sono tra 5 e 10, la scelta migliore deve essere fatta in ogni situazione in maniera soggettiva ed empirica, anche provando varie soluzioni, fino a trovare il numero di classi che porta ai grafici più significativi. È pratica comune, anche se non essenziale, prendere intervalli di classe tutti della stessa larghezza.

I *bordi* di una classe sono gli estremi del suo intervallo. Noi adottiamo la convenzione di includere i bordi di sinistra, intendendo con questo che ogni intervallo di classe contiene il suo estremo sinistro e non il suo estremo destro. Ad esempio l'intervallo 20–30 contiene tutti i valori che sono contemporaneamente maggiori o uguali a 20 e minori stretti di 30.

La Tabella 2.3 riporta i tempi di funzionamento di 200 lampadine a incandescenza. La Tabella 2.4 ne sintetizza la distribuzione tramite la frequenza di 10 intervalli

di classe di lunghezza 100, con il primo che comincia a 500.

Il grafico a barre delle frequenze (rispettivamente frequenze relative) delle classi prende il nome di *istogramma* (rispettivamente *istogramma delle frequenze relative*). La Figura 2.5 mostra l'istogramma dei dati della Tabella 2.4.

Un diverso tipo di rappresentazione di un insieme di dati è il grafico delle frequenze (relative o assolute) *cumulative*. Con ciò si intende una curva sul piano cartesiano

Tabella 2.3 Un insieme di dati numeroso: tempi di vita in ore di 200 lampadine ad incandescenza

1 067	919	1 196	785	1 126	936	918	1 156	920	948
855	1 092	1 162	1 170	929	950	905	972	1 035	1 045
1 157	1 195	1 195	1 340	1 122	938	970	1 237	956	1 102
1 022	978	832	1 009	1 157	1 151	1 009	765	958	902
923	1 333	811	1 217	1 085	896	958	1 311	1 037	702
521	933	928	1 153	946	858	1 071	1 069	830	1 063
930	807	954	1 063	1 002	909	1 077	1 021	1 062	1 157
999	932	1 035	944	1 049	940	1 122	1 115	833	1 320
901	1 324	818	1 250	1 203	1 078	890	1 303	1 011	1 102
996	780	900	1 106	704	621	854	1 178	1 138	951
1 187	1 067	1 118	1 037	958	760	1 101	949	992	966
824	653	980	935	878	934	910	1 058	730	680
844	814	1 103	1 000	788	1 143	935	1 069	1 170	1 067
1 037	1 151	863	990	1 035	1 112	931	970	932	904
1 026	1 147	883	867	990	1 258	1 192	922	1 150	1 091
1 039	1 083	1 040	1 289	699	1 083	880	1 029	658	912
1 023	984	856	924	801	1 122	1 292	1 116	880	1 173
1 134	932	938	1 078	1 180	1 106	1 184	954	824	529
998	996	1 133	765	775	1 105	1 081	1 171	705	1 425
610	916	1 001	895	709	860	1 110	1 149	972	1 002

Tabella 2.4 Frequenze assolute per classi di valori

Intervallo di classe	Frequenza (numero di dati che appartengono all'intervallo)
500–600	2
600–700	5
700–800	12
800–900	25
900–1 000	58
1 000–1 100	41
1 100–1 200	43
1 200–1 300	7
1 300–1 400	6
1 400–1 500	1

per cui le ascisse rappresentano i possibili valori dei dati, e le ordinate indicano il numero o la frazione di dati che sono minori o uguali ai valori in ascissa. Questo tipo di tracciato è anche detto ogiva (è soprattutto usato l'equivalente inglese *ogive*), e un esempio, relativo ai dati della Tabella 2.4 è dato in Figura 2.6. Studiando il grafico possiamo dedurre che il 100% dei dati sono inferiori a 1 500, il 40% circa sono minori o uguali a 900, l'80% circa sono minori o uguali a 1 100, e così via.

Una maniera efficiente di organizzare un numero non troppo grande di dati è il diagramma *stem and leaf* (in italiano, *ramo-foglia*, ma generalmente è usata la dicitura inglese). Per costruirlo, occorre dividere le cifre di ogni dato numerico in due

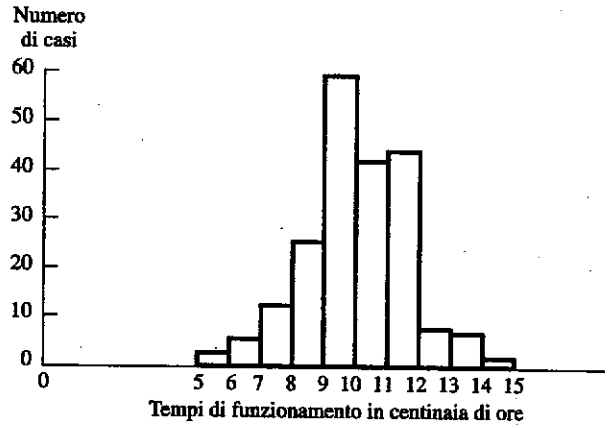


Figura 2.5 Istogramma

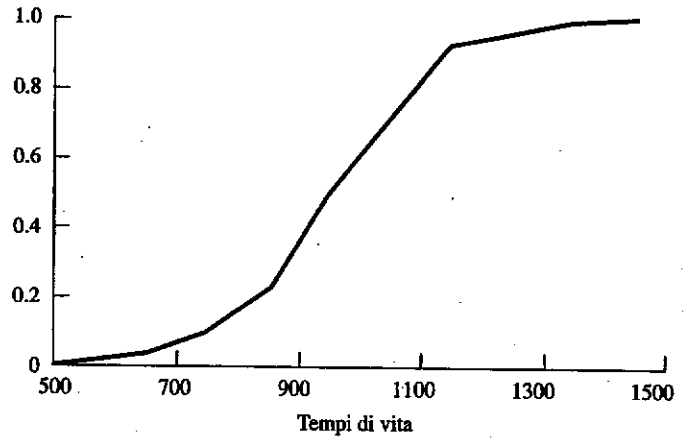


Figura 2.6 Grafico delle frequenze relative cumulative, anche detto ogiva.

parti, una più significativa (lo stem), e una meno significativa (la leaf). Ad esempio se tutti i dati fossero numeri di due cifre, sarebbe naturale scegliere le decine come stem e le unità come leaf. Con questa scelta il numero 62 diventa

Stem	Leaf
6	2

e i due dati 62 e 67 si possono scrivere insieme in questo modo

Stem	Leaf
6	2, 7

Esempio 2.2.3. La Tabella 2.5 fornisce le medie mensili e annuali delle temperature minime giornaliere (in gradi Fahrenheit) in 35 città americane. Le medie annuali sono riportate nel seguente diagramma stem and leaf.

7	0.0
6	9.0
5	1.0, 1.3, 2.0, 5.5, 7.1, 7.4, 7.6, 8.5, 9.3
4	0.0, 1.0, 2.4, 3.6, 3.7, 4.8, 5.0, 5.2, 6.0, 6.7, 8.1, 9.0, 9.2
3	3.1, 4.1, 5.3, 5.8, 6.2, 9.0, 9.1, 9.5, 9.5
2	9.0, 9.8

□

2.3 Le grandezze che sintetizzano i dati

Al giorno d'oggi non è raro dover trattare quantità anche notevoli di dati. Già nel 1951, gli statistici R. Doll e A. B. Hill, nel tentativo di scoprire le conseguenze sulla salute di alcune abitudini sociali, inviarono dei questionari a tutti i medici del Regno Unito, ricevendo circa 40 000 risposte. Le domande erano molteplici, riguardavano età, abitudini alimentari, fumo. Coloro di cui si ebbe risposta vennero seguiti per i successivi dieci anni, e si registrarono le cause di decesso di quelli tra loro che morirono. Per avere una sensazione di un così vasto campione di dati, è utile saperli sintetizzare in qualche misura. In questa sezione presentiamo alcune statistiche sintetiche, dove con il termine *statistica* si intende una grandezza calcolata a partire dai dati.

2.3.1 Media, mediana e moda campionarie

Le statistiche affrontate in questa sezione sono usate per descrivere il centro di un insieme di dati, ovvero un valore attorno al quale si forma la rosa dei dati. Siccome non vi è un modo univoco di intendere questa dicitura (il valore più tipico? il valore più centrale?) non vi è una definizione unica che chiuda il problema, ve ne sono tre, tra le quali scegliere a seconda degli aspetti che ci interessano.

Tabella 2.5 Temperature minime giornaliere tipiche – alcune città

Stato	Stazione meteo	Gen.	Feb.	Mar.	Apr.	Mag.	Giu.	Lug.	Ago.	Set.	Ott.	Nov.	Dic.	Media annuale
AL	Mobile	40.0	42.7	50.1	57.1	64.4	70.7	73.2	72.9	68.7	57.3	49.1	43.1	57.4
AK	Juneau	19.0	22.7	26.7	32.1	38.9	45.0	48.1	47.3	42.9	37.2	27.2	22.6	34.1
AZ	Phoenix	41.2	44.7	48.8	55.3	63.9	72.9	81.0	79.2	72.8	60.8	48.9	41.8	59.3
AR	Little Rock	29.1	33.2	42.2	50.7	59.0	67.4	71.5	69.8	63.5	50.9	41.5	33.1	51.0
CA	Los Angeles	47.8	49.3	50.5	52.8	56.3	59.5	62.8	64.2	63.2	59.2	52.8	47.9	55.5
	Sacramento	37.7	41.4	43.2	45.5	50.3	55.3	58.1	58.0	55.7	50.4	43.4	37.8	48.1
	San Diego	48.9	50.7	52.8	55.6	59.1	61.9	65.7	67.3	65.6	60.9	53.9	48.8	57.6
	San Francisco	41.8	45.0	45.8	47.2	49.7	52.6	53.9	55.0	55.2	51.8	47.1	42.7	49.0
CO	Denver	16.1	20.2	25.8	34.5	43.6	52.4	58.6	56.9	47.6	36.4	25.4	17.4	36.2
CT	Hartford	15.8	18.6	28.1	37.5	47.6	56.9	62.2	60.4	51.8	40.7	32.8	21.3	39.5
DE	Wilmington	22.4	24.8	33.1	41.8	52.2	61.6	67.1	65.9	58.2	45.7	37.0	27.6	44.8
DC	Washington	26.8	29.1	37.7	46.4	56.6	66.5	71.4	70.0	62.5	50.3	41.1	31.7	49.2
FL	Jacksonville	40.5	43.3	49.2	54.9	62.1	69.1	71.9	71.8	69.0	59.3	50.2	43.4	57.1
	Miami	59.2	60.4	64.2	67.8	72.1	75.1	76.2	76.7	75.9	72.1	66.7	61.5	69.0
GA	Atlanta	31.5	34.5	42.5	50.2	58.7	66.2	69.5	69.0	63.5	51.9	42.8	35.0	51.3
HI	Honolulu	65.6	65.4	67.2	68.7	70.3	72.2	73.5	74.2	73.5	72.3	70.3	67.0	70.0
ID	Boise	21.6	27.5	31.9	36.7	43.9	52.1	57.7	56.8	48.2	39.0	31.1	22.5	39.1
IL	Chicago	12.9	17.2	28.5	38.6	47.7	57.5	62.6	61.6	53.9	42.2	31.6	19.1	39.5

Tabella 2.5 (continua)

Stato	Stazione meteo	Gen.	Feb.	Mar.	Apr.	Mag.	Giu.	Lug.	Ago.	Set.	Ott.	Nov.	Dic.	Media annuale
IL	Peoria	13.2	17.7	29.8	40.8	50.9	60.7	65.4	63.1	55.2	43.1	32.5	19.3	41.0
IN	Indianapolis	17.2	20.9	31.9	41.5	51.7	61.0	65.2	62.8	55.6	43.5	34.1	23.2	42.4
IA	Des Moines	10.7	15.6	27.6	40.0	51.5	61.2	66.5	63.6	54.5	42.7	29.9	16.1	40.0
KS	Wichita	19.2	23.7	33.6	44.5	54.3	64.6	69.9	67.9	59.2	46.6	33.9	23.0	45.0
KY	Louisville	23.2	26.5	36.2	45.4	54.7	62.9	67.3	65.8	58.7	45.8	37.3	28.6	46.0
LA	New Orleans	41.8	44.4	51.6	58.4	65.2	70.8	73.1	72.8	69.5	58.7	51.0	44.8	58.5
ME	Portland	11.4	13.5	24.5	34.1	43.4	52.1	58.3	57.1	48.9	38.3	30.4	17.8	35.8
MD	Baltimora	23.4	25.9	34.1	42.5	52.6	61.8	66.8	65.7	58.4	45.9	37.1	28.2	45.2
MA	Boston	21.6	23.0	31.3	40.2	48.8	59.1	65.1	64.0	56.8	46.9	38.3	26.7	43.6
MI	Detroit	15.6	17.6	27.0	36.8	47.1	56.3	61.3	59.6	52.5	40.9	32.2	21.4	39.0
	Sault Ste. Marie	4.6	4.8	15.3	28.4	38.4	45.5	51.3	51.3	44.3	36.2	25.9	11.8	29.8
MN	Duluth	-2.2	2.8	15.7	28.9	39.6	48.5	55.1	53.3	44.5	35.1	21.5	4.9	29.0
	Minneapolis-St. Paul	2.8	9.2	22.7	36.2	47.6	57.6	63.1	60.3	50.3	38.8	25.2	10.2	35.3
MS	Jackson	32.7	35.7	44.1	51.9	60.0	67.1	70.5	69.7	63.7	50.3	42.3	36.1	52.0
MO	Kansas City	16.7	21.8	32.6	43.8	53.9	63.1	68.2	65.7	56.9	45.7	33.6	21.9	43.7
	St. Louis	20.8	25.1	35.5	46.4	56.0	65.7	70.4	67.9	60.5	48.3	37.7	26.0	46.7
MT	Great Falls	11.6	17.2	22.8	31.9	40.9	48.6	53.2	52.2	43.5	35.8	24.3	14.6	33.1

Fonte: U.S. National Oceanic and Atmospheric Administration, Climatography of the United States, No. 81.

Supponiamo di avere un insieme $x_1, x_2, \ldots, x_n$ di n dati (o come anche si dice, un campione di *ampiezza* o *numerosità* pari a n). La media campionaria è la media aritmetica di questi valori.

Definizione 2.3.1. Si dice *media campionaria* e si denota con $\bar{x}$, la quantità

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.3.1}$$

Il calcolo manuale di questa grandezza può essere notevolmente semplificato se si nota che, prese comunque due costanti a e b, se si considera il nuovo insieme di dati

$$y_i := ax_i + b, \qquad i = 1, \ldots, n \tag{2.3.2}$$

allora la media campionaria di $y_1, y_2, \ldots, y_n$ è legata a quella dei dati iniziali dalla stessa relazione lineare:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) = \frac{1}{n}\sum_{i=1}^{n} ax_i + \frac{1}{n}\sum_{i=1}^{n} b = a\bar{x} + b \tag{2.3.3}$$

Esempio 2.3.1. Quelli che seguono sono i punteggi vincenti del torneo di golf U.S. Masters negli anni dal 1982 al 1991:

284 280 277 282 279 285 281 283 278 277

Se ne vuole trovare la media campionaria.

Invece che applicare direttamente la definizione, si può usare la considerazione fatta sopra, costruendo ad esempio il nuovo insieme di dati $y_i = x_i - 280$, che è più maneggevole da trattare:

4 0 −3 2 −1 5 1 3 −2 −3

La media campionaria dei dati trasformati si calcola molto facilmente,

$$\bar{y} = \frac{4+0-3+2-1+5+1+3-2-3}{10} = \frac{6}{10}$$

Ne segue che

$$\bar{x} = \bar{y} + 280 = 280.6 \quad \square$$

Merita menzione l'aritmetica necessaria a calcolare la media campionaria di un insieme di dati che sia fornito tramite le frequenze dei suoi valori. Siano $v_1, v_2, \ldots, v_k$ i k valori distinti assunti dai dati, e siano $f_1, f_2, \ldots, f_k$ le relative frequenze assolute. Siccome il numero complessivo di dati è $n = \sum_{i=1}^{k} f_i$, e per

$i = 1, 2, \ldots, k$ il valore v_i compare f_i volte nel campione di dati, segue che la media campionaria degli n dati è

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} f_i v_i \tag{2.3.4}$$

Se si riscrive l'ultima formula come

$$\bar{x} = \frac{f_1}{n}v_1 + \frac{f_2}{n}v_2 + \cdots + \frac{f_k}{n}v_k$$

si può notare come la media campionaria non sia altro che una *media pesata* dei valori assunti dai dati. Ogni valore usa come peso la sua frequenza relativa, ovvero la frazione dei dati uguali a tale valore.

Esempio 2.3.2. Quelle che seguono sono le frequenze delle età dei membri di una orchestra sinfonica giovanile.

Età	15	16	17	18	19	20
Frequenza	2	5	11	9	14	13

Si vuole trovare la media campionaria dei 54 dati.

$$\bar{x} = \frac{15\cdot 2 + 16\cdot 5 + 17\cdot 11 + 18\cdot 9 + 19\cdot 14 + 20\cdot 13}{54} \approx 18.24 \quad \square$$

Una seconda statistica che indica il centro di un insieme di dati è la mediana campionaria; sinteticamente, si tratta del valore centrale una volta che i dati siano messi in ordine crescente.

Definizione 2.3.2. Assegnato un insieme di dati di ampiezza n, lo si ordini dal minore al maggiore. Se n è dispari, si dice *mediana campionaria* il valore del dato in posizione $(n+1)/2$; se n è pari invece, è la media aritmetica tra i valori dei dati che occupano le posizioni $n/2$ e $n/2+1$.

Così la mediana di un campione di tre dati è quello che ha valore intermedio, mentre per un insieme di quattro dati è la media aritmetica tra i due valori intermedi.

Esempio 2.3.3. Cerchiamo la mediana campionaria dei dati forniti nell'Esempio 2.3.2.

Poiché i dati sono 54, un numero pari, si prendono i due che occupano la posizione 27 e la 28 in ordine crescente, in questo caso un 18 e un 19. La mediana campionaria è la loro media aritmetica, ovvero 18.5. $\square$

Media e mediana campionaria sono entrambe statistiche utili per descrivere i valori centrali dei dati. La media fa uso di tutti i dati e in particolare è influenzata in maniera sensibile da valori eccezionalmente alti o bassi. La mediana invece dipende

direttamente solo da uno o due valori in centro alla distribuzione e non risente dei dati estremi. La decisione di quale statistica scegliere dipende dall'uso che se ne intende fare. Per esempio, un'amministrazione comunale che volesse avere una stima del gettito fiscale complessivo (supponendo un'aliquota fiscale costante), dovrebbe scegliere di utilizzare la media campionaria dei redditi dei residenti. La stessa amministrazione potrebbe invece trovare più utile la mediana campionaria dei redditi, nel caso fosse interessata a costruire abitazioni popolari e volesse stabilire quale sia il potere di acquisto del ceto residenziale medio.

Esempio 2.3.4. In uno studio scientifico[2], un gruppo di topi di cinque settimane fu sottoposto a una dose di radiazione di 300 rad. I topi furono quindi divisi in due gruppi, il primo dei quali venne tenuto in ambiente sterile, mentre il secondo in normali condizioni di laboratorio. I seguenti diagrammi stem and leaf riportano i giorni di vita dei topi che in seguito morirono di linfoma del timo.

	Topi in ambiente sterile		Topi in ambiente normale
1	58, 92, 93, 94, 95	1	59, 89, 91, 98
2	02, 12, 15, 29, 30, 37, 40, 44, 47, 59	2	35, 45, 50, 56, 61, 65, 66, 80
3	01, 01, 21, 37	3	43, 56, 83
4	15, 34, 44, 85, 96	4	03, 14, 28, 32
5	29, 37		
6	24		
7	07		
8	00		

È evidente dai diagrammi stem and leaf che la media campionaria del primo campione sarà sensibilmente maggiore di quella del secondo; infatti eseguendo i calcoli si trovano 344.07 giorni di media per i topi in ambiente sterile, e 292.32 giorni di media nell'altro caso. Determiniamo ora le mediane campionarie. Il primo insieme di osservazioni ha numerosità 29, quindi la mediana è il 15-esimo dato in ordine crescente, 259. Il secondo campione è formato da 19 dati, e il decimo in ordine crescente, 265, è la sua mediana. Quindi, anche se la media del primo campione è notevolmente maggiore di quella del secondo, le mediane campionarie sono molto vicine. La spiegazione di questo fatto è che la media campionaria del primo gruppo risente fortemente dei cinque valori maggiori di 500, che hanno però molta meno influenza sulla mediana campionaria. Infatti, essa resterebbe invariata anche se sostituissimo quei cinque dati con numeri molto più piccoli, purché non inferiori a 259. Sembra perciò che l'ambiente sterile abbia allungato la vita dei cinque topi che vissero più a lungo, ma non è chiaro che effetto abbia avuto, se ne ha avuto, sul tempo di vita degli altri topi. □

[2] D. G. Hoel, "A representation of mortality data by competing risks", *Biometrics*, vol. 28, pp. 475–488, 1972.

La terza statistica che viene impegata per descrivere il centro di una distribuzione di dati è la moda campionaria.

Definizione 2.3.3. La *moda campionaria* di un insieme di dati, se esiste, è l'unico valore che ha frequenza massima. Se non vi è un solo valore con frequenza massima, ciascuno di essi è detto *valore modale*.

Esempio 2.3.5. La seguente tabella riporta la frequenza di uscita delle sei facce di un dado, su 40 lanci.

Valore	1	2	3	4	5	6
Frequenza	9	8	5	5	6	7

Vogliamo calcolare: **(a)** la media campionaria, **(b)** la mediana campionaria e **(c)** la moda campionaria.

(a) La media campionaria è

$$\bar{x} = \frac{9+16+15+20+30+42}{40} = 3.05$$

(b) La mediana campionaria è la media aritmetica del 20-esimo e del 21-valore, che sono entrambi 3, quindi è essa stessa pari a 3. **(c)** La moda campionaria è 1, il valore che è comparso più di frequente. □

2.3.2 Varianza e deviazione standard campionarie

Le statistiche presentate nella sezione precedente forniscono sotto diversi punti di vista i valori centrali della distribuzione dei dati. Un'altra questione di chiaro interesse è quanto i dati siano concentrati o viceversa dispersi attorno a tali valori tipici. Una strategia impiegabile a questo scopo potrebbe essere allora considerare le distanze dei dati dalla media campionaria, elevarle al quadrato e farne la media aritmetica. In effetti questa è quasi la definizione di varianza campionaria, che però, per ragioni tecniche, si ottiene dividendo per $n-1$ anziché per n.

Definizione 2.3.4. Assegnato un insieme di dati $x_1, x_2, \ldots, x_n$, si dice *varianza campionaria* e si denota con s^2 la quantità

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.3.5}$$

Esempio 2.3.6. Si trovi la varianza campionaria dei due insiemi di dati seguenti.

A: $3, 4, 6, 7, 10$ **B:** $-20, 5, 15, 24$

La media del campione **A** è $\bar{x} = (3+4+6+7+10)/5 = 6$; dalla definizione di varianza campionaria allora

$$s^2 = [(-3)^2 + (-2)^2 + 0^2 + 1^2 + 4^2]/4 = 7.5$$

Anche per **B** la media campionaria è 6; tuttavia la sua varianza risulta

$$s^2 = [(-26)^2 + (-1)^2 + 9^2 + 12^2]/3 \approx 360.67$$

Perciò, anche se entrambi gli insiemi di dati hanno la stessa media campionaria, vi è una variabilità molto maggiore nei valori di **B** che non in quelli di **A**. □

La seguente identità algebrica è usata spesso per velocizzare il calcolo manuale della varianza campionaria.

Proposizione 2.3.1. Sia dato un insieme di dati $x_1, x_2, \ldots, x_n$, e sia $\bar{x}$ la sua media campionaria, allora

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \tag{2.3.6}$$

Dimostrazione.

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) && \text{sviluppando il quadrato}\\
&= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\bar{x}^2 && \text{spezzando la sommatoria}\\
&= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 && \text{per la definizione di } \bar{x}\\
&= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \quad \square
\end{aligned}$$

Il calcolo della varianza campionaria può anche essere semplificato se si nota che, prese comunque due costanti a e b, se si considera il nuovo insieme di dati $y_i := ax_i + b$, dove $i = 1, 2, \ldots, n$, allora per quanto già detto a pagina 22, $\bar{y} = a\bar{x} + b$, e quindi

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = a^2\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Perciò, se s_y^2 e s_x^2 sono le rispettive varianze campionarie, si ha che

$$s_y^2 = a^2 s_x^2 \tag{2.3.7}$$

Per riassumere, sommare una costante a ciascuno dei dati non fa cambiare la varianza, mentre moltiplicarli per un fattore costante fa sì che la varianza campionaria risulti moltiplicata per il quadrato di tale fattore.

Esempio 2.3.7. È qui di seguito riportato il numero di incidenti aerei mortali in tutto il mondo negli anni dal 1985 al 1993. Questi dati si riferiscono a voli commerciali.

Anno	1985	1986	1987	1988	1989	1990	1991	1992	1993
Incidenti	22	22	26	28	27	25	30	29	24

Fonte: Civil Aviation Statistics of the World, annual.

Si trovi la varianza campionaria di questi dati.

Per cominciare, sottraiamo 22 ai valori di partenza, ottenendo il nuovo campione:

$$0\ \ 0\ \ 4\ \ 6\ \ 5\ \ 3\ \ 8\ \ 7\ \ 2$$

denotiamo questi dati con $y_1, y_2, \ldots, y_9$ e calcoliamo

$$\sum_{i=1}^{9} y_i = 35, \qquad \sum_{i=1}^{9} y_i^2 = 16 + 36 + 25 + 9 + 64 + 49 + 4 = 203$$

da cui, ricordando che la varianza dei dati trasformati è in questo caso uguale a quella dei dati iniziali, e usando la Proposizione 2.3.1, otteniamo

$$s^2 = \frac{203 - 9(35/9)^2}{8} \approx 8.361 \quad \square$$

Il Programma 2.3, disponibile sul sito web dedicato a questo volume, può essere usato per calcolare la varianza di campioni più numerosi.

La radice quadrata della varianza è detta deviazione standard.

Definizione 2.3.5. Assegnato un insieme di dati $x_1, x_2, \ldots, x_n$, si dice *deviazione standard campionaria* e si denota con s la quantità

$$s := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.3.8}$$

Questa grandezza ha le stesse unità di misura dei dati sperimentali.

2.3.3 Percentili campionari e box plot

In sintesi, il percentile k-esimo di un campione di dati è un valore che è maggiore di una percentuale k dei dati, e minore della restante percentuale $100 - k$, dove k è un numero tra 0 e 100. Più formalmente diamo la seguente definizione.

Definizione 2.3.6. Sia k un numero intero con $0 \leq k \leq 100$. Assegnato un insieme di dati numerici, ne esiste sempre uno che è contemporaneamente maggiore o uguale di almeno il k percento dei dati, e minore o uguale di almeno il $100-k$ percento dei dati. Se il dato con queste caratteristiche è unico, esso è per definizione il *percentile k-esimo* dell'insieme di dati considerato. Se invece non è unico, allora sono esattamente due, e in questo caso il *percentile k-esimo* è definito come la loro media aritmetica.

Quindi per determinare il percentile k-esimo di un campione di numerosità n occorre trovare quello o quei dati tali che, detto p il rapporto $k/100$,

1. almeno np tra tutti i dati dell'insieme siano minori o uguali a loro;
2. almeno $n(1-p)$ tra tutti i dati dell'insieme siano maggiori o uguali a loro.

Per prima cosa disponiamo i dati in ordine crescente. Notiamo poi che, se il numero np non è intero, l'unico dato che soddisfa le richieste è quello che occupa la posizione data da np arrotondato all'intero successivo. Ad esempio, supponiamo che siano $n = 22$ e $k = 80$, e di conseguenza $p = 0.8$ e $np = 17.6$; ci viene chiesto di trovare un dato che sia maggiore o uguale di almeno 17.6 (ovvero almeno 18) delle osservazioni e minore o uguale di almeno 4.4 (ovvero almeno 5) di esse; ovviamente, solo il 18-esimo dato in ordine crescente soddisfa questa richiesta, ed esso è il percentile 80-esimo. Se invece np è un numero intero, è facile vedere che sia esso sia il suo successivo soddisfano le richieste, e quindi la quantità cercata è la media di questi due valori.

Esempio 2.3.8. La Tabella 2.6 riporta la popolazione delle 30 maggiori città americane per il 1990. Calcoliamo **(a)** il decimo percentile e **(b)** il 95-esimo percentile di questi dati.

(a) Poiché la numerosità del campione è $n = 30$, e $np = 30 \cdot 0.1 = 3$ è un numero intero, il decimo percentile è la media aritmetica del terzo e del quarto dato dal più piccolo, ovvero

$$\frac{447\,619 + 465\,648}{2} = 456\,633.5$$

(b) Poiché $30 \cdot 0.95 = 28.5$, il 95-esimo percentile è il 29-esimo dato dal più piccolo, ovvero 3 485 557. □

Il 50-esimo percentile coincide ovviamente con la mediana campionaria. Assieme al 25-esimo e al 75-esimo percentile, forma i quartili campionari.

Definizione 2.3.7. Il 25-esimo percentile si dice *primo quartile*; il 50-esimo si dice mediana campionaria o *secondo quartile*; il 75-esimo è il *terzo quartile*.

I quartili dividono il campione in quattro parti: i dati minori del primo quartile, quelli maggiori del terzo, quelli compresi tra il primo e il secondo e quelli tra il secondo e il terzo sono sempre circa il 25%.

Tabella 2.6 Popolazione delle 30 maggiori città degli Stati Uniti

Posizione	Città	Residenti
1	New York, NY	7 322 564
2	Los Angeles, CA	3 485 557
3	Chicago, IL	2 783 726
4	Houston, TX	1 629 902
5	Philadelphia, PA	1 585 577
6	San Diego, CA	1 110 623
7	Detroit, MI	1 027 974
8	Dallas, TX	1 007 618
9	Phoenix, AZ	983 403
10	San Antonio, TX	935 393
11	San Jose, CA	782 224
12	Indianapolis, IN	741 952
13	Baltimora, MD	736 014
14	San Francisco, CA	723 959
15	Jacksonville, FL	672 971
16	Columbus, OH	632 945
17	Milwaukee, WI	628 088
18	Memphis, TN	610 337
19	Washington, DC	606 900
20	Boston, MA	574 283
21	Seattle, WA	516 259
22	EI Paso, TX	515 342
23	Nashville-Davidson, TN	510 784
24	Cleveland, OH	505 616
25	New Orleans, LA	496 938
26	Denver, CO	467 610
27	Austin, TX	465 648
28	Fort Worth, TX	447 619
29	Oklahoma City, OK	444 724
30	Portland, OR	438 802

Fonte: Bureau of the Census, U.S. Dept. of Commerce (100 most populous cities ranked by April 1990 census; revised April 1994).

Esempio 2.3.9. Il rumore si misura il decibel, indicati dal simbolo dB. Un decibel è circa la soglia di udibilità in condizioni ideali per una persona con un ottimo udito; 30 dB sono il livello sonoro di un sussurro; un tono di conversazione normale può misurare 70 dB; una radio ad alto volume arriva a 100 dB; la soglia di tollerabilità è intorno ai 120 dB. I valori seguenti sono i livelli di rumore misurati in 36 differenti occasioni in prossimità della stazione centrale di Manhattan.

82	89	94	110	74	122	112	95	100	78	65	60
90	83	87	75	114	85	69	94	124	115	107	88
97	74	72	68	83	91	90	102	77	125	108	65

Per determinare i quartili campionari, riportiamo i dati in un diagramma stem and leaf:

6	0, 5, 5, 8, 9
7	2, 4, 4, 5, 7, 8
8	2, 3, 3, 5, 7, 8, 9
9	0, 0, 1, 4, 4, 5, 7
10	0, 2, 7, 8
11	0, 2, 4, 5
12	2, 4, 5

Il primo quartile è la media del nono e del decimo dato, vale 76. Il secondo è la media del 18-esimo e 19-esimo dato, vale 89.5. Il terzo è la media del 27-esimo e del 28-esimo dato, e vale 104.5. □

Uno strumento utile a visualizzare alcune delle statistiche rappresentative dei dati è il *box plot*. Si ottiene sovrapponendo ad una linea orizzontale che va dal minore al maggiore dei dati, un rettangolo (il *box*) che va dal primo al terzo quartile, con una linea verticale che lo divide al livello del secondo quartile. Per esempio, i 42 dati della Tabella 2.1 vanno da un minimo di 27 ad un massimo di 40, i quartili campionari sono nell'ordine 30, 31.5 e 34; il box plot corrispondente è quello di Figura 2.7.

La lunghezza della linea orizzontale del box plot, pari alla distanza tra il minimo e il massimo dei suoi valori, si dice *campo di variazione* (oppure *range*, che è l'espressione inglese corrispondente). La lunghezza del solo rettangolo invece, pari alla distanza tra il primo e il terzo quartile, è detta *scarto interquartile*.

2.4 La disuguaglianza di Chebyshev

Siano $\bar{x}$ e s media e deviazione standard campionarie di un insieme di dati. Nell'ipotesi che $s > 0$, la disuguaglianza di Chebyshev afferma che per ogni reale $k \geq 1$, almeno una frazione $(1-1/k^2)$ dei dati cade nell'intervallo che va da $\bar{x}-ks$ a $\bar{x}+ks$. Così ad esempio, con $k = 1.5$ scopriamo che almeno i $5/9$ – pari al 55.56% circa – di un qualunque campione di dati stanno entro una distanza di $1.5s$ dalla loro media campionaria. Con $k = 2$ calcoliamo che almeno il 75% dei dati sta entro $2s$ dalla media campionaria. Con $k = 3$ troviamo che almeno l'88.9% dei dati sta entro una distanza di $3s$ da $\bar{x}$.

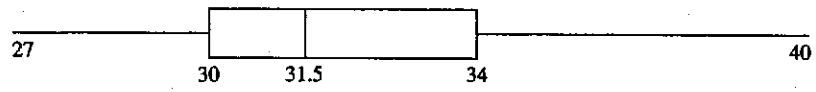


Figura 2.7 Un box plot

Quando l'ampiezza n del campione è nota, la disuguaglianza si può migliorare, come si evince dall'enunciato formale e dalla dimostrazione che seguono.

Proposizione 2.4.1 (Disuguaglianza di Chebyshev). Sia assegnato un insieme di dati $x_1, x_2, \ldots, x_n$, con media campionaria $\bar{x}$ e deviazione standard campionaria $s > 0$. Denotiamo con S_k l'insieme degli indici corrispondenti a dati compresi tra $\bar{x} - ks$ e $\bar{x} + ks$:

$$S_k := \{i, 1 \leq i \leq n : |x_i - \bar{x}| < ks\} \tag{2.4.1}$$

e sia $\#S_k$ il numero di elementi o *cardinalità* dell'insieme S_k. Allora, per ogni $k \geq 1$,

$$\frac{\#S_k}{n} \geq 1 - \frac{n-1}{nk^2} > 1 - \frac{1}{k^2} \tag{2.4.2}$$

Dimostrazione.

$$\begin{aligned}
(n-1)s^2 &= \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i \in S_k}(x_i - \bar{x})^2 + \sum_{i \notin S_k}(x_i - \bar{x})^2 \\
&\geq \sum_{i \notin S_k}(x_i - \bar{x})^2 && \text{perché gli addendi sono non negativi} \\
&\geq \sum_{i \notin S_k} k^2 s^2 && \text{perché se } i \notin S_k,\ (x_i - \bar{x})^2 \geq k^2 s^2 \\
&= k^2 s^2 (n - \#S_k)
\end{aligned}$$

Dividendo entrambi i membri per nk^2s^2 si trova che

$$\frac{n-1}{nk^2} \geq 1 - \frac{\#S_k}{n}$$

da cui segue l'enunciato. □

L'ipotesi $s > 0$ non è in realtà fondamentale. Infatti poiché $s \geq 0$ per definizione, l'unico caso che resta escluso è quando $s = 0$. Tuttavia guardando alla definizione di deviazione standard campionaria, è facile convincersi che l'unico modo in cui può essere nulla, è se tutti i dati sono uguali, $x_1 = x_2 = \cdots = x_n = \bar{x}$, nel qual caso la disuguaglianza è ancora vera, anche se in modo triviale.

Poiché la disuguaglianza di Chebyshev vale per tutti gli insiemi di numeri, è lecito aspettarsi che in molti casi la percentuale di dati che cadono entro ks dalla media $\bar{x}$, sia in realtà molto maggiore di quella stimata.

Tabella 2.7 Automobili più vendute negli Stati Uniti. Anno solare 1993 (nazionali e importate).

1.	Ford Taurus	380 448
2.	Honda Accord	330 030
3.	Toyota Camry	299 737
4.	Chevrolet Cavalier	273 617
5.	Ford Escort	269 034
6.	Honda Civic	255 579
7.	Saturn	229 356
8.	Chevrolet Lumina	219 683
9.	Ford Tempo	217 644
10.	Pontiac Grand Am	214 761
11.	Toyota Corolla	193 749
12.	Chevrolet Corsica/Beretta	171 794
13.	Nissan Sentra	167 351
14.	Buick LeSabre	149 299

American Automobile Manufacturers Assn.

Esempio 2.4.1. La Tabella 2.7 elenca le 14 auto più vendute negli Stati Uniti nel 1993. Un calcolo diretto di media e deviazione standard campionarie, ad esempio con il software abbinato al testo, fornisce i seguenti valori,

$$\bar{x} \approx 239\,434, \qquad s \approx 62\,235$$

La disuguaglianza di Chebyshev afferma che almeno il 55.56% dei dati (o almeno il 58.73%, usando la versione più raffinata che suppone n nota), devono stare nell'intervallo

$$(\bar{x} - 1.5s, \bar{x} + 1.5s) = (146\,082, 332\,787)$$

quando invece i valori che cadono entro questi limiti sono 13 su 14, ovvero il 92.1% circa. □

2.5 Campioni normali

Osservando gli istogrammi dei campioni numerici forniti da esperimenti reali, si può notare come vi sia una forma caratteristica che compare molto spesso, e accomuna un gran numero di campioni di dati, provenienti dai contesti più disparati. Questi grafici hanno un solo massimo, in corrispondenza della mediana, e decrescono da entrambi i lati simmetricamente, secondo una curva a campana. Un campione di dati che rispetta questi requisiti di dice *normale*. La Figura 2.8 presenta un ideale istogramma di questo tipo.

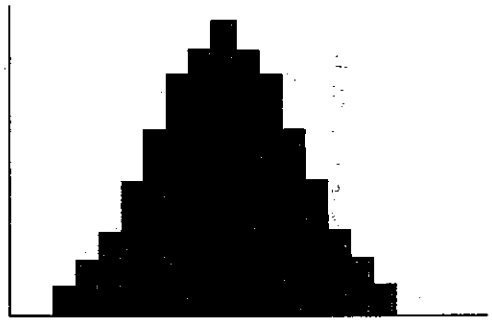


Figura 2.8 Un istogramma normale perfetto

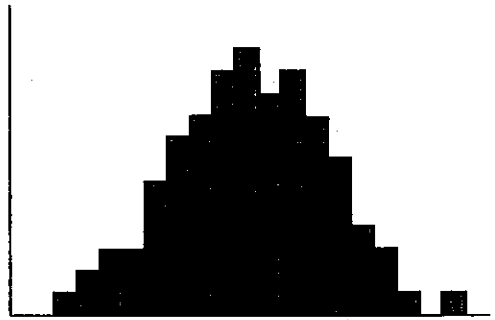


Figura 2.9 Un istogramma approssimativamente normale

In realtà, pur mantenendo un aspetto simile a quello descritto, non capita mai che un istogramma reale rispetti perfettamente la simmetria e la monotonia. Si può parlare allora di campione approssimativamente normale, e l'istogramma in Figura 2.9 ne costituisce un esempio. Se un insieme di dati presenta un istogramma che è sensibilmente asimmetrico rispetto alla mediana, come quelli nelle Figure 2.10 e 2.11, si parla di campione *skewed* (ovvero *sbilanciato*), a sinistra o a destra, a seconda del lato in cui ha la coda più lunga.

Dalla simmetria degli istogrammi normali segue che un campione approssimativamente normale avrà media e mediana campionaria circa uguali.

Supponiamo che $\bar{x}$ e s siano media e deviazione standard di un campione approssimativamente normale. La seguente regola empirica specifica che percentuale dei dati ci si aspetta di trovare entro s, $2s$ e $3s$ dalla media campionaria. Essa rispetta i limiti imposti dalla disuguaglianza di Chebyshev, ma ne migliora grandemente la precisione, valendo non per ogni insieme di dati, ma solo per campioni approssimativamente normali, e fornendo risultati non esatti.

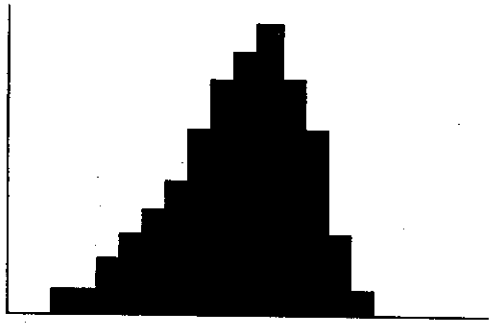


Figura 2.10 L'istogramma di un campione *skewed* a sinistra.

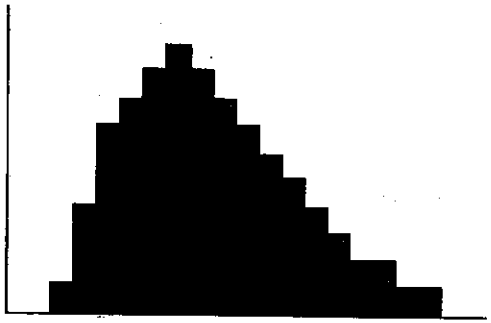


Figura 2.11 L'istogramma di un campione *skewed* a destra.

Osservazione 2.5.1 (La Regola Empirica). Se un campione numerico è approssimativamente normale, ha media campionaria $\bar{x}$ e deviazione standard campionaria s, allora

1. Circa il 68% dei dati cade nell'intervallo $\bar{x} \pm s$
2. Circa il 95% dei dati cade nell'intervallo $\bar{x} \pm 2s$
3. Circa il 99.7% dei dati cade nell'intervallo $\bar{x} \pm 3s$

Esempio 2.5.1. Il diagramma stem and leaf che segue riporta i punteggi ottenuti in un esame di statistica da alcuni studenti di ingegneria industrale.

9	0, 1, 4
8	3, 5, 5, 7, 8
7	2, 4, 4, 5, 7, 7, 8
6	0, 2, 3, 4, 6, 6
5	2, 5, 5, 6, 8
4	3, 6

Ruotando il diagramma in senso antiorario, si può notare che il corrispondente istogramma è approssimativamente normale. Mettiamo alla prova la regola empirica.

Facendo i conti si trova

$$\bar{x} \approx 70.6, \qquad s \approx 14.4$$

La regola empirica dice che i punteggi compresi tra 56.2 e 85.0 dovrebbero essere circa il 68%. In effetti, essi sono 17/28, pari al 60.7%. Analogamente, quelli compresi tra 41.8 e 99.4 dovrebbero essere il 95%, e in realtà sono il 100%. □

Un insieme di dati ottenuto campionando da una popolazione non omogenea, ma costituita da sottogruppi eterogenei, di solito non risulta normale. Piuttosto, l'istogramma di un tale campione, presenta spesso l'aspetto di una sovrapposizione di istogrammi normali, e in particolare può avere due o più massimi locali. Siccome questi picchi sono analoghi alla moda, un campione di questo tipo si dice *bimodale* se ne possiede due e *multimodale* in generale. I dati rappresentati in Figura 2.12 sono appunto bimodali.

2.6 Insiemi di dati bivariati e coefficiente di correlazione campionaria

Talvolta non abbiamo a che fare con sequenze di dati singoli, ma con sequenze di coppie di numeri, tra i quali esiste qualche relazione. In questi casi ogni coppia è da considerarsi una osservazione; se scegliamo di denominare con x e y i due tipi di grandezze che compaiono in ciascun dato, possiamo denotare con (x_i, y_i) la coppia di valori che costituisce la osservazione i-esima. Dati di questa forma prendono il nome di campione *bivariato*. Ad esempio, un'azienda che vuole indagare il rapporto tra la temperatura ambientale e il numero di parti difettose che escono dalla sua linea

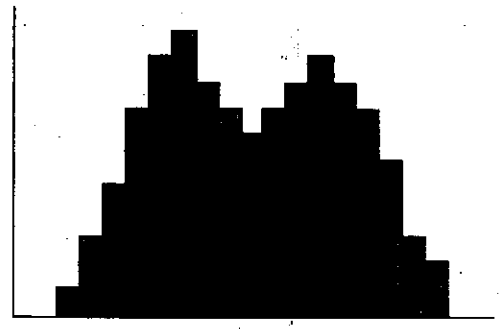


Figura 2.12 L'istogramma di un campione bimodale

di produzione, può registrare per un certo numero di giorni le temperature massime e il numero di difetti riscontrati. Dei dati esemplificativi sono riportati in Tabella 2.8; in questo caso x_i e y_i denotano rispettivamente la temperatura e i difetti del giorno i-esimo.

Uno strumento utile a visualizzare campioni bivariati è il *diagramma di dispersione*, ovvero la rappresentazione sul piano cartesiano di tanti punti quante sono le osservazioni, ciascuno tracciato alle coordinate corrispondenti ai suoi due valori x e y. La Figura 2.13 mostra il diagramma ottenuto dai dati della Tabella 2.8.

Una questione di grande interesse quando si studiano campioni bivariati è se vi sia una *correlazione* tra i valori x e y, ovvero se si verifica che le osservazioni che hanno un alto valore di x tendano tipicamente ad avere anche un alto valore di y, o viceversa tendano ad averne uno basso, e analogamente, si chiede che le osservazioni che hanno un basso livello di x abbiano abbinato pure un basso (o viceversa alto) livello di y. Se numeri elevati corrispondono a numeri elevati e valori bassi corrispondono a valori bassi, la correlazione è *positiva*, se invece quando x è grande y è tipicamente piccolo e viceversa, allora si parla di correlazione *negativa*. Una risposta grossolana alla questione della correlazione si può ottenere osservando il diagramma di dispersione;

Tabella 2.8 Temperature massime giornaliere in gradi Celsius e numero di parti difettose

Giorno	Temperatura	Difetti
1	24.2	25
2	22.7	31
3	30.5	36
4	28.6	33
5	25.5	19
6	32.0	24
7	28.6	27
8	26.5	25
9	25.3	16
10	26.0	14
11	24.4	22
12	24.8	23
13	20.6	20
14	25.1	25
15	21.4	25
16	23.7	23
17	23.9	27
18	25.2	30
19	27.4	33
20	28.3	32
21	28.8	35
22	26.6	24

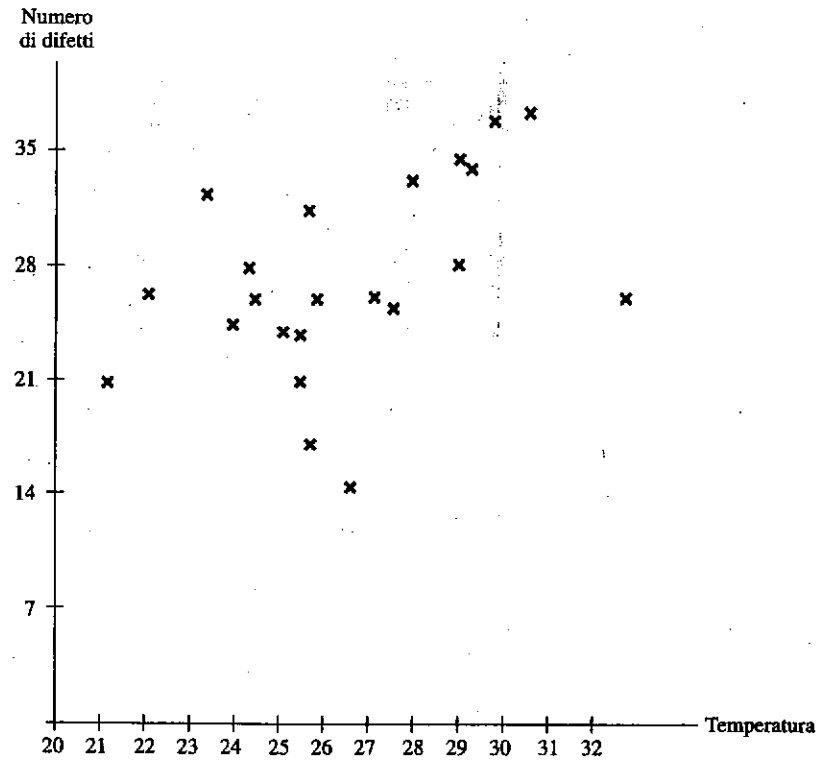


Figura 2.13 Un diagramma di dispersione

ad esempio la Figura 2.13 mostra qualche tipo di corrispondenza tra alte temperature e elevato numero di difetti. Per ottenere una misura quantitativa di questa relazione, costruiamo una nuova statistica.

Consideriamo un campione bivariato (x_i, y_i), per $i = 1, 2, \ldots, n$. Siano $\bar{x}$ e $\bar{y}$ le medie campionarie relative ai valori x e y rispettivamente. Possiamo senz'altro dire che se un valore x_i è grande rispetto a quelli tipici, allora la differenza $x_i - \bar{x}$ sarà positiva, mentre se x_i è piccolo, essa sarà negativa; possiamo ragionare analogamente per i valori y. Quindi, se consideriamo il prodotto $(x_i - \bar{x})(y_i - \bar{y})$, esso sarà maggiore di zero per le osservazioni in cui x_i e y_i sono correlate positivamente, e minore di zero per quelle in cui vi è correlazione negativa. Quindi se l'intero campione mostra una forte correlazione c'è da aspettarsi che la somma $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ lo percepisca, a seconda del tipo, assumendo un valore molto positivo o molto negativo.

Per dare senso all'affermazione che quella sommatoria sia "molto" positiva, si usa normalizzarla, dividendo per $n - 1$ e per il prodotto delle deviazioni standard campionarie dei valori x e y.

La statistica che si ottiene è il coefficiente di correlazione campionaria.

Definizione 2.6.1. Sia dato un campione bivariato (x_i, y_i), per $i = 1, 2, \ldots, n$, con medie campionarie $\bar{x}$ e $\bar{y}$ e deviazioni standard campionarie s_x e s_y, per i soli dati x e per i soli dati y rispettivamente. Allora di dice *coefficiente di correlazione campionaria* e si denota con r la quantità

$$r := \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.6.1}$$

Quando $r > 0$ i dati sono *correlati positivamente*, mentre se $r < 0$ sono *correlati negativamente*.

Proposizione 2.6.1. Di seguito diamo alcune delle proprietà del coefficiente di correlazione campionaria.

1. $-1 \leq r \leq 1$.
2. Se per oppurtune costanti a e b, con $b > 0$, sussiste la relazione lineare
$$y_i = a + bx_i, \qquad \forall i = 1, 2, \ldots, n$$
allora $r = 1$.
3. Se per oppurtune costanti a e b, con $b < 0$ sussiste la relazione lineare
$$y_i = a + bx_i, \qquad \forall i = 1, 2, \ldots, n$$
allora $r = -1$.
4. Se r è il coefficiente di correlazione del campione (x_i, y_1), $i = 1, \ldots, n$, allora lo è anche per il campione
$$(a + bx_i, c + dy_i) \qquad \forall i = 1, 2, \ldots, n$$
purché le costanti b e d abbiano lo stesso segno.

La Proprietà 1 dice che che r è sempre compreso tra -1 e $+1$, inoltre le Proprietà 2 e 3 precisano che i valori limite $+1$ e -1 sono effettivamente raggiunti solo quando tra x e y sussiste una relazione lineare (ovvero i punti del diagramma di dispersione giacciono esattamente su una retta). La Proprietà 4 afferma che il coefficiente di correlazione non cambia se sommiamo costanti o moltiplichiamo per costanti tutti i valori di x e/o tutti i valori di y. Ciò significa ad esempio che r non dipende dalle

unità di misura scelte per i dati. Il coefficiente di correlazione tra peso e altezza di un gruppo di individui non cambia se si decide di misurare il peso in libbre piuttosto che in chilogrammi, o la statura in pollici piuttosto che in centimetri o anche in metri. Analogamente, se uno dei valori di interesse è una temperatura, è lo stesso usare dati in gradi Celsius o Fahrenheit o Kelvin.

Il valore assoluto di r è una misura della forza della correlazione esistente. Come si è già detto, quando $|r| = 1$ vi è relazione lineare perfetta, e i punti del diagramma di dispersione stanno tutti su una retta; valori intorno a 0.8 indicano una correlazione molto intensa, e anche se i punti del grafico non stanno tutti su una retta, ve n'è una (la retta *interpolante*) che passa non passa troppo lontana da nessuno di essi; valori di r intorno a 0.3 denotano una relazione molto debole.

Il segno di r indica la direzione della retta. È positivo se x e y tendono a essere grandi e piccoli assieme, nel qual caso la retta interpolante punta verso l'alto. È negativo invece se, quando x è grande y è tipicamente piccolo e viceversa; allora l'approssimante punta in basso. La Figura 2.14 mostra diagrammi di dispersione corrispondenti a diversi valori di r.

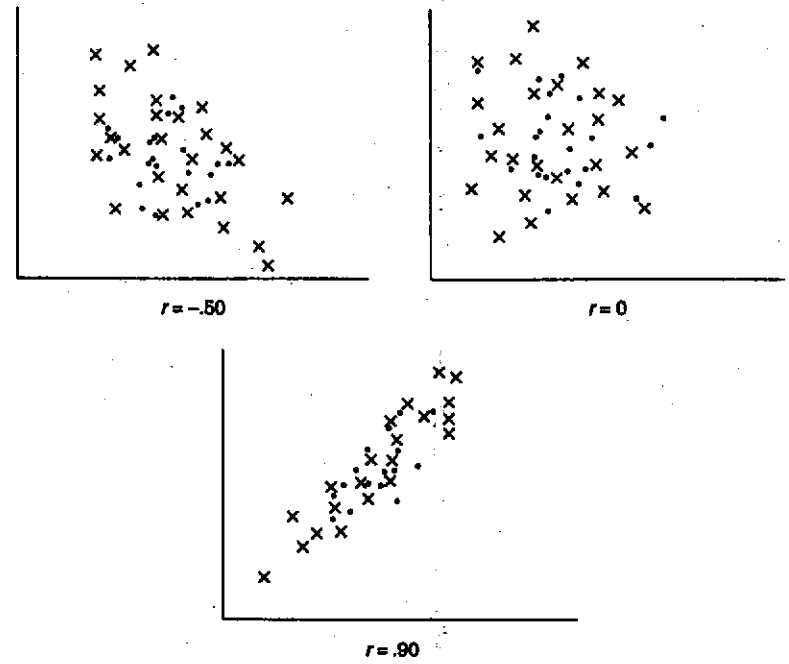


Figura 2.14 Diagrammi di dispersione corrispondenti a diversi valori di r.

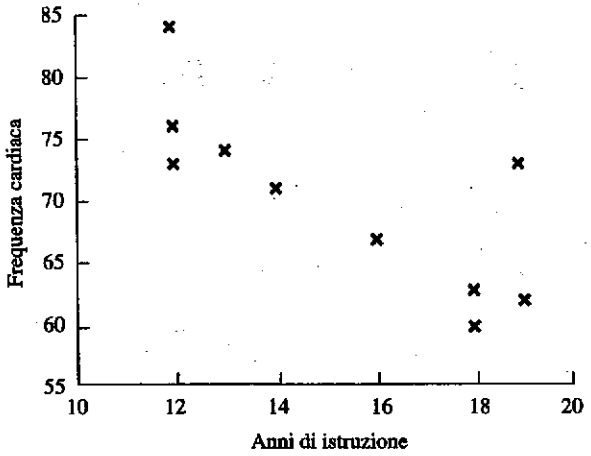


Figura 2.15 Diagramma di dispersione per la frequenza cardiaca e gli anni di scuola di un campione di 10 individui.

Esempio 2.6.1. Per quanto riguarda i dati riportati nella Tabella 2.8, un calcolo diretto mostra che $r = 0.4189$. Questo valore denota una correlazione positiva debole tra la temperatura giornaliera massima e il numero di parti difettose prodotte. □

Esempio 2.6.2. Riportiamo di seguito la frequenza cardiaca a riposo (in battiti al minuto) e gli anni complessivi di istruzione di dieci individui.

Soggetto	1	2	3	4	5	6	7	8	9	10
Anni di istruzione	12	16	13	18	19	12	18	19	12	14
Frequenza cardiaca	73	67	74	63	73	84	60	62	76	71

Il diagramma di dispersione corrispondente è illustrato in Figura 2.15, il coefficiente di correlazione lineare risulta essere $r = -0.7638$. Questa correlazione negativa indica che vi è una forte associazione tra una scolarizzazione lunga e una bassa frequenza cardiaca e viceversa. □

Correlazione, non rapporto di causa-effetto

Le conclusioni dell'Esempio 2.6.2 indicano una forte correlazione tra gli anni di istruzione e la frequenza cardiaca a riposo per gli individui del campione. Tuttavia questo non significa che gli ulteriori anni di studio ne abbiano effettivamente ridotto i battiti cardiaci. Quella che abbiamo trovato è infatti una associazione tra le due caratteristiche, non un rapporto di causa-effetto. Spesso, la spiegazione di associazioni di questo tipo dipende da un terzo fattore implicito, legato a entrambe le variabili in esame. Nel nostro caso, potrebbe darsi che le persone con una migliore istruzione siano più informate nel campo della salute, e quindi ad esempio più conscie dell'importanza di fare esercizio regolarmente di una sana alimentazione. Oppure è possibile che non sia la conoscenza a fare la differenza, ma piuttosto il fatto che persone con titoli di studio più elevati possono accedere a impieghi che lasciano più tempo per fare attività fisica ed essendo meglio pagati favoriscono l'acquisto di cibi migliori. La forte correlazione trovata è certamente dovuta a una combinazione di questi e probabilmente altri fattori inespressi.

Problemi

1. Quello che segue è un campione dei prezzi della benzina praticati nel giugno del 1997 nella zona di San Francisco. I dati sono in centesimi di dollaro per gallone.

   137 139 141 137 144 141 139 137 144 141 143 143 141

   (a) Organizza questi dati in una tabella delle frequenze.

   (b) Rappresenta la loro frequenza relativa con un grafico a bastoncini.

2. Spiega come si costruisce un grafico a torta. Se uno dei valori del campione ha frequenza relativa r, che angolo al centro avrà il settore circolare corrispondente?

3. Quelle che seguono sono delle stime non aggiornate – in milioni di barili – delle riserve di petrolio di quattro regioni del continente americano.

Stati Uniti	38.7
Sud America	22.6
Canada	8.8
Messico	60.0

   Traccia un grafico a torta per questi dati.

4. La tabella a pagina seguente riporta, per i 50 Stati degli USA, il tempo medio necessario per raggiungere il posto di lavoro e la percentuale di lavoratori che usa mezzi pubblici.

Mĕzzi impiegati e tempi necessari per raggiungere il posto di lavoro

Regione, Divisione e Stato	Percentuale che si serve dei mezzi pubblici	Tempo medio di spostamento
Stati Uniti	**5.3**	**22.4**
Northeast	**12.8**	**24.5**
New England	5.1	21.5
Maine	0.9	19.0
New Hampshire	0.7	21.9
Vermont	0.7	18.0
Massachusetts	8.3	22.7
Rhode Island	2.5	19.2
Connecticut	3.9	21.1
Middle Atlantic	15.7	25.7
New York	24.8	28.6
New Jersey	8.8	25.3
Pennsylvania	6.4	21.6
Midwest	**3.5**	**20.7**
East North Central	4.3	21.7
Ohio	2.5	20.7
Indiana	1.3	20.4
Illinois	10.1	25.1
Michigan	1.6	21.2
Wisconsin	2.5	18.3
West North Central	1.9	18.4
Minnesota	3.6	19.1
Iowa	1.2	16.2
Missouri	2.0	21.6
North Dakota	0.6	13.0
South Dakota	0.3	13.8
Nebraska	1.2	15.8
Kansas	0.6	17.2
South	**2.6**	**22.0**
South Atlantic	3.4	22.5
Delaware	2.4	20.0
Maryland	8.1	27.0
Virginia	4.0	24.0
West Virginia	1.1	21.0
North Carolina	1.0	19.8
South Carolina	1.1	20.5
Georgia	2.8	22.7
Florida	2.0	21.8
East South Central	1.2	21.1
Kentucky	1.6	20.7
Tennessee	1.3	21.5
Alabama	0.8	21.2
Mississippi	0.8	20.6
West South Central	2.0	21.6
Arkansas	0.5	19.0
Luisiana	3.0	22.3
Oklahoma	0.6	19.3
Texas	2.2	22.2
West	**4.1**	**22.7**
Mountain	2.1	19.7
Montana	0.6	14.8
Idaho	1.9	17.3
Wyoming	1.4	15.4
Colorado	2.9	20.7
New Mexico	1.0	19.1
Arizona	2.1	21.6
Utah	2.3	18.9
Nevada	2.7	19.8
Pacific	4.8	23.8
Washington	4.5	22.0
Oregon	3.4	19.6
California	4.9	24.6
Alaska	2.4	16.7
Hawaii	7.4	23.8

Fonte: U.S. Bureau of the Census. Census of population and housing, 1990.

(a) Rappresenta i tempi medi di spostamento con un istogramma.

(b) Organizza i dati sulla percentuale di lavoratori che usa mezzi pubblici con un diagramma stem and leaf.

5. Scegli un libro oppure un articolo e conta il numero di parole in ciascuna delle prime 100 frasi, quindi presenta i valori osservati tramite un diagramma stem and leaf. Successivamente, ripeti l'esercizio su un testo di un autore differente. I due diagrammi stem and leaf ottenuti si assomigliano? È ragionevole pensare di impiegare questa tecnica per stabilire se due articoli sono stati scritti da autori differenti?

6. La Tabella a pagina seguente riporta il numero di incidenti aerei mortali all'anno e il numero delle vittime, per i voli commerciali effettuati negli Stati Uniti dal 1980 al 1995. Per quanto riguarda il numero di incidenti all'anno:

(a) costruisci la tabella delle frequenze;

(b) traccia il grafico a linee delle frequenze;

(c) traccia il grafico delle frequenze cumulative relative;

Sicurezza dei voli negli USA, veicoli commeciali, 1980–1995

Anno	Voli (milioni)	Incidenti mortali	Vittime
1980	5.4	0	0
1981	5.2	4	4
1982	5.0	4	233
1983	5.0	4	5
1984	5.4	1	4
1985	5.8	4	197
1986	6.4	2	5
1987	6.6	4	231
1988	6.7	3	285
1989	6.6	11	278
1990	6.9	6	39
1991	6.8	4	62
1992	7.1	4	33
1993	7.2	1	1
1994	7.5	4	239
1995	8.1	2	166

Fonte: National Transportation Safety Board

(d) calcola la media campionaria;

(e) calcola la mediana campionaria;

(f) calcola la moda campionaria;

(g) calcola la deviazione standard campionaria.

7. Con riferimento alla Tabella del Problema 6, considera il numero di vittime all'anno:

(a) rappresenta i dati in un istogramma;

(b) riorganizzali in un diagramma stem and leaf;

(c) calcola la media campionaria;

(d) calcola la mediana campionaria;

(e) calcola la deviazione standard campionaria.

8. Usa i dati della tabella di pagina 45 per

(a) realizzare un diagramma stem and leaf e

(b) trovare la mediana campionaria

del numero di linee telefoniche su 100 persone nelle diverse nazioni.

9. Usando la tabella del Problema 4, trova le medie e le mediane campionarie dei tempi di spostamento per gli stati che fanno parte delle seguenti regioni.

(a) northeast;

Numero di linee telefoniche attive ogni 100 persone (dati del 1994)

Paese	Linee	Paese	Linee
Algeria	4	Kuwait	23
Arabia Saudita	10	Libano	9
Argentina	14	Lussemburgo	54
Australia	50	Malaysia	15
Austria	47	Marocco	4
Belgio	45	Messico	9
Brasile	7	Norvegia	55
Bulgaria	34	Nuova Zelanda	47
Canada	58	Olanda	51
Cile	11	Pakistan	1
Cina	2	Panama	11
Cipro	45	Paraguay	3
Colombia	9	Perù	4
Corea del Sud	40	Polonia	13
Costarica	13	Portogallo	35
Cuba	3	Portorico	33
Danimarca	60	Regno Unito	47
Ecuador	5	Repubblica Ceca	21
Egitto	4	Repubblica Dominicana	8
Filippine	2	Repubblica Sudafricana	9
Finlandia	55	Romania	12
Francia	55	Russia	16
Germania	48	Singapore	47
Giappone	48	Siria	5
Grecia	48	Spagna	37
Guatemala	2	Stati Uniti	59
Honduras	2	Svezia	68
Hong Kong	54	Svizzera	60
India	1	Tailandia	4
Indonesia	1	Taiwan	40
Iran	7	Trinidad e Tobago	16
Iraq	3	Tunisia	5
Irlanda	33	Turchia	20
Islanda	56	Ungheria	17
Israele	37	Uruguay	17
Italia	43	Venezuela	11
Jamaica	10		

Fonte: International Telecommunication Union, Ginevra.

(b) midwest;

(c) south;

(d) west.

10. I valori della tabella di pagina 47 sono le mediane dei prezzi per le abitazioni monofamiliari in diverse città americane nel 1992 e nel 1994.

(a) Rappresenta i dati del 1992 con un istogramma.

(b) Rappresenta i dati del 1992 con un diagramma stem and leaf.

(c) Calcola la mediana campionaria delle mediane dei prezzi del 1992.

(d) Calcola la mediana campionaria delle mediane dei prezzi del 1994.

11. La tabella che segue riporta il numero di pedoni – classificati secondo età e sesso – che sono morti in incidenti stradali in Inghilterra nel 1922.

(a) Trova media e mediana campionaria dell'età al decesso per i maschi.

(b) Trova media e mediana campionaria dell'età al decesso per le femmine.

(c) Calcola i quartili per i maschi.

(d) Calcola i quartili per le femmine.

Età	Maschi	Femmine
0–5	120	67
5–10	184	120
10–15	44	22
15–20	24	15
20–30	23	25
30–40	50	22
40–50	60	40
50–60	102	76
60–70	167	104
70–80	150	90
80–100	49	27

12. I valori che seguono sono le percentuali di ceneri residue per 12 campioni di carbone trovati in uno stesso sito.

9.2 14.1 9.8 12.4 16.0 12.6 22.7 18.9 21.0 14.5 20.4 16.9

Trova media e deviazione standard campionarie di queste percentuali.

13. Usando i dati del Problema 4, calcola la varianza campionaria dei tempi di spostamento per gli stati che si trovano nelle divisioni:

(a) South Atlantic;

(b) Mountain.

14. La media e la varianza di un campione di 5 dati sono rispettivamente $\bar{x} = 104$ e $s^2 = 4$. Sapendo che tre dati sono 102, 100 e 105, quali sono gli altri due dati?

Mediane dei prezzi delle case da abitazione monofamiliari

Città	Aprile 1992	Aprile 1994
Akron, OH	75 500	81 600
Albuquerque, NM	86 700	103 100
Anaheim/Santa Ania, CA	235 100	209 500
Atlanta, GA	85 800	93 200
Baltimora, MD	111 500	115 700
Baton Rouge, LA	71 800	78 400
Birmingham, LA	89 500	99 500
Boston, MA	168 200	170 600
Bradenton, FL	80 400	86 400
Buffalo, NY	79 700	82 400
Charleston, SC	82 000	91 300
Chicago, IL	131 100	135 500
Cincinnati, OH	87 500	93 600
Cleveland, OH	88 100	94 200
Columbia, SC	85 100	82 900
Columbus, OH	90 300	92 800
Corpus Christi, TX	62 500	71 700
Dallas, TX	90 500	95 100
Daytona Beach, FL	63 600	66 200
Denver, CO	91 300	111 200
Des Moines, IA	71 200	77 400
Detroit, MI	77 500	84 500
El Paso, TX	65 900	73 600
Grand Rapids, MI	73 000	76 600
Hartford, CT	141 500	132 900
Honolulu, HI	342 000	355 000
Houston, TX	78 200	84 800
Indianapolis, IN	80 100	90 500
Jacksonville, FL	75 100	79 700
Kansas City, MO	76 100	84 900
Knoxville, TN	78 300	88 600
Las Vegas, NV	101 400	110 400
Los Angeles, CA	218 000	188 500

Fonte: National Association of Realtors: Dati di metà 1994.

15. La tabella di pagina 48 riporta il reddito annuale medio pro capite negli stati americani per il 1992 e il 1993.

(a) Ti aspetti che la media campionaria dei dati dei 51 stati sia uguale al dato degli interi Stati Uniti?

(b) Se la risposta al punto **(a)** è negativa, spiega che informazioni servirebbero, oltre alle medie relative ai singoli stati, per calcolare la media campionaria dell'intera

Reddito annuale medio per stato: 1992 e 1993

(Dati espressi in dollari. Sono esclusi i piccoli coltivatori, i militari, le cariche politiche, gli impiegati delle ferrovie, i lavoratori a domicilio, gli studenti lavoratori, gli impiegati di alcune organizzazioni no profit e la maggior parte degli imprenditori. Il reddito include i bonus, il controvalore di vitto e alloggio, le mance e altre gratifiche.)

Stato	1992	1993	Stato	1992	1993
Stati Uniti	**25 897**	**26 362**	Missouri	23 550	23 898
Alabama	22 340	22 786	Montana	19 378	19 932
Alaska	31 825	32 336	Nebraska	20 355	20 815
Arizona	23 153	23 501	Nevada	24 743	25 461
Arkansas	20 108	20 337	New Hampshire	24 866	24 962
California	28 902	29 468	New Jersey	32 073	32 716
Colorado	25 040	25 682	New Mexico	21 051	21 731
Connecticut	32 603	33 169	New York	32 399	32 919
Delaware	26 596	27 143	North Carolina	22 249	22 770
District of Columbia	37 951	39 199	North Dakota	18 945	19 382
Florida	23 145	23 571	Ohio	24 845	25 339
Georgia	24 373	24 867	Oklahoma	21 698	22 003
Hawaii	25 538	26 325	Oregon	23 514	24 093
Idaho	20 649	21 188	Pennsylvania	25 785	26 274
Illinois	27 910	28 420	Rhode Island	24 351	24 889
Indiana	23 570	24 109	South Carolina	21 398	21 928
Iowa	20 937	21 441	South Dakota	18 016	18 613
Kansas	21 982	22 430	Tennessee	22 807	23 368
Kentucky	21 858	22 170	Texas	25 088	25 545
Louisiana	22 342	22 632	Utah	21 976	22 250
Maine	21 808	22 026	Vermont	22 360	22 704
Maryland	27 145	27 684	Virginia	24 940	25 496
Massachusetts	29 664	30 229	Washington	25 553	25 760
Michigan	27 463	28 260	West Virginia	22 168	22 373
Minnesota	25 324	25 711	Wisconsin	23 008	23 610
Mississippi	19 237	19 694	Wyoming	21 215	21 745

Fonte: U.S. Bureau of Labor Statistics, Employment and Wages Annual Averages 1993; and USDL News Release 94-451, Average Annual Pay by State and Industry, 1993.

nazione. Spiega anche come impiegare quelle informazioni a questo scopo.

(c) Calcola le mediane campionarie dei dati relativi al 1992 e dei dati relativi al 1993.

(d) Calcola la media campionaria dei redditi del 1992 per i primi dieci stati elencati.

(e) Calcola la deviazione standard campionaria dei redditi del 1993 per gli ultimi dieci stati in elenco.

16. I dati seguenti rappresentano i tempi di vita (in ore) di un campione di 40 transistor.

112	121	126	108	141	104	136	134	121	118
143	116	108	122	127	140	113	117	126	130
134	120	131	133	118	125	151	147	137	140
132	119	110	124	132	152	135	130	136	128

(a) Determina media, mediana, e moda campionarie.

(b) Traccia un grafico delle frequenze culumative relative per questi dati.

17. Un esperimento volto a misurare la percentuale di restringimento tramite essiccazione di 50 campioni di argilla, ha dato i seguenti valori:

18.2	21.2	23.1	18.5	15.6	20.8	19.4	15.4	21.2	13.4
16.4	18.7	18.2	19.6	14.3	16.6	24.0	17.6	17.8	20.2
17.4	23.6	17.5	20.3	16.6	19.3	18.5	19.3	21.2	13.9
20.5	19.0	17.6	22.3	18.4	21.2	20.4	21.4	20.3	20.1
19.6	20.6	14.8	19.7	20.5	18.0	20.8	15.8	23.1	17.0

(a) Crea un diagramma stem and leaf con questi dati.

(b) Calcola media, mediana e moda campionarie.

(c) Determina la varianza campionaria.

(d) Raggruppa i dati in intervalli di classe di larghezza pari a un punto percentuale a iniziare dal 13.0%; traccia poi l'istogramma corrispondente.

(e) Utilizzando le frequenze delle classi ottenute al punto (d), e facendo finta che i dati all'interno di ogni intervallo di classe siano localizzati nel punto medio, calcola media e varianza campionarie, e confrontale con i valori trovati nei punti (b) e (c). Come mai sono diversi?

18. Un metodo computazionalmente efficiente per calcolare media e varianza campionaria dell'insieme di dati $x_1, x_2, \ldots, x_n$ è il seguente. Sia

$$\bar{x}_j := \frac{1}{j}\sum_{i=1}^{j} x_i, \qquad j = 1, 2, \ldots, n$$

la media campionaria dei primi j dati; e sia

$$s_j^2 := \frac{1}{j-1}\sum_{i=1}^{j}(x_i - \bar{x}_j)^2, \qquad j = 2, \ldots, n$$

la varianza campionaria dei primi j dati (con $j \geq 2$, attenzione!). Allora se si pone $s_1^2 := 0$ è possibile dimostrare che

$$\bar{x}_{j+1} = \bar{x}_j + \frac{x_{j+1} - \bar{x}_j}{j+1}$$

$$s_{j+1}^2 = \left(1 - \frac{1}{j}\right)s_j^2 + (j+1)(\bar{x}_{j+1} - \bar{x}_j)^2$$

(a) Utilizza queste due formule ricorsive per calcolare media e varianza campionarie dei dati 3, 4, 7, 2, 9, 6.

(b) Verifica la correttezza del risultato trovato al punto (a) usando i metodi usuali.

(c) Dimostra la prima delle due formule citate.

19. Utilizza i dati del Problema 10 per calcolare, sia per il 1992, sia per il 1994, (a) il decimo percentile, (b) il 40-esimo percentile e (c) il 90-esimo percentile delle mediane dei prezzi.

20. Analizza la tabella a pagina 51 trovando i quartili dei redditi medi sia per il 1992, sia per il 1993.

21. Utilizza la Figura 2.16, che riporta gli stanziamenti federali per la ricerca che vennero assegnati nel 1992 a 15 università, per rispondere alle seguenti domande.

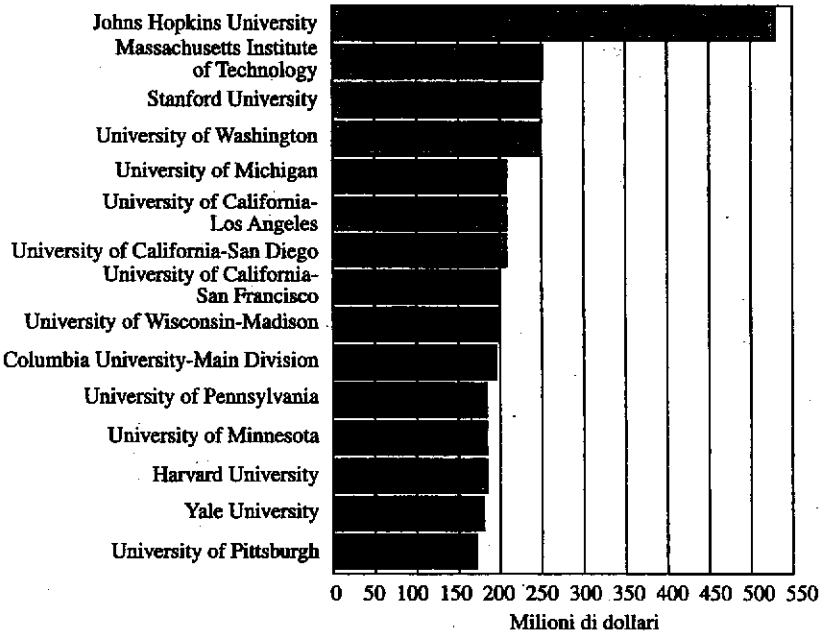
(a) Quali università ricevettero più di 225 milioni di dollari di stanziamenti?

(b) Quanto vale approssimativamente la media campionaria di quegli stanziamenti?

(c) E la varianza campionaria?

(d) Determina i quartili campionari.

22. Disegna il box plot dei dati sulla percentuale di lavoratori che fa uso di mezzi pubblici per recarsi sul posto di lavoro. Usa la tabella del Problema 4.



Fonte: U.S. Bureau of Census.

Figura 2.16 Fondi federali per la ricerca – primi 15 centri universitari.

23. La tabella di pagina 52 riporta il numero di cani delle diverse razze che furono iscritti nel 1995 nell'American Kennel Club. Rappresenta questi numeri in un box plot.

24. La misurazione della concentrazione di particelle in sospensione in un complesso petrolchimico in 36 diversi momenti, fornisce (in microgrammi per metro cubo) i valori seguenti:

Reddito annuale medio in alcune aree metropolitane: 1992 e 1993

(Dati in dollari. Aree metropolitane ordinate per reddito medio del 1993. Comprende i dati di Metropolitan Statistical Areas e Primary Metropolitan Statistical Areas, secondo le definizioni valide al 30 giugno 1993. Nelle aree del New England sono state utilizzate le definizioni del New England County Metropolitan Area (NECMA). Vedere le fonti per dettagli. Vedere anche le precisazioni sulla tabella di pagina 48.)

Area Metropolitana	1992	1993
Tutte le aree metropolitane	**27 051**	**27 540**
New York, NY	38 802	39 381
San Jose, CA	37 068	38 040
Middlesex-Somerset-Hunterdon, NJ	34 796	35 573
San Francisco, CA	34 364	35 278
Newark, NJ	34 302	35 129
New Haven-Bridgeport-Stamford-Danbury-Waterbury, CT	34 517	35 058
Trenton, NJ	33 960	34 365
Bergen-Passaic, NJ	33 555	34 126
Anchorage, AK	33 007	33 782
Washington, DC-MD-VA-WV	32 337	33 170
Jersey City, NJ	31 638	32 815
Hartford, CT	31 967	32 555
Los Angeles-Long Beach, CA	31 165	31 760
Oakland, CA	30 623	31 701
Detroit, MI	30 534	31 622
Chicago, IL	30 210	30 720
Boston-Worcester-Lawrence-Lowell-Brockton, MA-NH	30 100	30 642
Flint, MI	29 672	30 512
Nassau-Suffolk, NY	29 708	30 226
Houston, TX	29 794	30 069
Orange County, CA	29 353	29 916
Philadelphia, PA-NJ	29 392	29 839
Dutchess County, NY	29 262	29 730
Kokomo, IN	28 676	29 672
Dallas, TX	28 813	29 489
Seattle-Bellevue-Everett, WA	29 466	29 399
Huntsville, AL	28 944	29 243
Wilmington-Newark, DE-MD	28 635	29 232
New London-Norwich, CT	27 926	28 630

Fonte: U.S. Bureau of Labor Statistics, USDL New Release 94–516, Average Annual Pay Levels in Metropolitan Areas.

Le 25 razze più diffuse secondo l'American Kennel Club

Posizione	Razza	Cani iscritti
1	Labrador Retriever	132 051
2	Rottweiler	93 656
3	Pastore tedesco	78 088
4	Golden Retriever	64 107
5	Beagle	57 063
6	Barboncino	54 784
7	Cocker Spaniel	48 065
8	Bassotto	44 680
9	Pomeranian	37 894
10	Yorkshire Terrier	36 881
11	Dalmata	36 714
12	Shih Tzu	34 947
13	Pastore delle Shetland	33 721
14	Chihuahua	33 542
15	Boxer	31 894
16	Schnauzer Nano	30 256
17	Siberian Husky	24 291
18	Dobermann Pinscher	18 141
19	Pinscher Nano	17 810
20	Chow Chow	17 722
21	Maltese	16 179
22	Basset Hound	16 055
23	Boston Terrier	16 031
24	Carlino	15 927
25	English Springer Spaniel	15 039

Fonte: American Kennel Club, New York, NY: cani iscritti nel 1995.

5	18	15	7	23	220	130	85	103	25	80	7
24	6	13	65	37	25	24	65	82	95	77	15
70	110	44	28	33	81	29	14	45	92	17	53

(a) Rappresenta questi dati in un istogramma.

(b) Si tratta di un campione approssimativamente normale?

25. Un ingegnere chimico che vuole studiare la velocità di evaporazione dell'acqua dalle vasche di una salina, dispone di 55 osservazioni giornaliere fatte nei mesi di luglio nell'arco di 4 anni. I dati, in pollici di acqua evaporata in 24 ore, sono riportati nel diagramma stem and leaf che segue, e vanno da un minimo di 0.02 ad un massimo di 0.56 pollici.

0.0	2, 6
0.1	1, 4
0.2	1, 1, 1, 3, 3, 4, 5, 5, 5, 6, 9
0.3	0, 0, 2, 2, 2, 3, 3, 3, 3, 4, 4, 5, 5, 5, 6, 6, 7, 8, 9
0.4	0, 1, 2, 2, 2, 3, 4, 4, 4, 5, 5, 5, 6, 7, 8, 8, 8, 9, 9
0.5	2, 5, 6

(a) Trova la media campionaria.

(b) Trova la mediana campionaria.

(c) Calcola la deviazione standard campionaria.

(d) Secondo te questo campione è approssimativamente normale?

(e) Che percentuale dei dati sta entro una deviazione standard dalla media campionaria?

26. Quelle che seguono sono le medie alla laurea di 30 studenti ammessi al programma di studi post laurea presso il dipartimento di ingegneria industriale all'Università della California (Berkeley).

3.46	3.72	3.95	3.55	3.62	3.80	3.86	3.71	3.56	3.49
3.96	3.90	3.70	3.61	3.72	3.65	3.48	3.87	3.82	3.91
3.69	3.67	3.72	3.66	3.79	3.75	3.93	3.74	3.50	3.83

(a) Rappresenta i dati in un diagramma stem and leaf.

(b) Trova la mediana campionaria $\bar{x}$.

(c) Calcola la deviazione standard campionaria s.

(d) Determina la frazione di dati che sta nell'intervallo $\bar{x} \pm 1.5s$ e confrontala con il limite inferiore fornito dalla disuguaglianza di Chebyshev.

(e) Ripeti il punto precedente per l'intervallo $\bar{x} \pm 2s$.

27. Il campione di dati del Problema 26 è approssimativamente normale? Confronta il valore trovato al punto **(e)** di quel problema con la stima fornite dalla regola empirica.

28. Pensi che l'istogramma dei pesi corporei delle persone che frequentano un fitness club sarà approssimativamente normale? Perché?

29. Usa i dati del Problema 16.

(a) Calcola media e mediana del campione.

(b) La distribuzione dei dati è approssimativamente normale?

(c) Calcola la deviazione standard campionaria.

(d) Che percentuale dei dati cade entro $\bar{x} \pm 2s$?

(e) Confronta il risultato del punto **(d)** con la stima data dalla regola empirica.

(f) Confronta il risultato del punto **(d)** con il limite inferiore dato dalla disuguaglianza di Chebyshev.

30. Usa i dati riguardanti i primi 10 stati che compaiono nella tabella del Problema 15.

(a) Disegna il diagramma di dispersione che mette in relazione i dati del 1992 con quelli del 1993.

(b) Calcola il coefficiente di correlazione campionaria.

31. La tabella di pagina 56 riporta il 50-esimo percentile dei redditi per soggetti che hanno conseguito una laurea oppure un master nei campi della scienza pura e dell'ingegneria.

(a) Traccia il diagramma di dispersione e

(b) calcola il coefficiente di correlazione campionaria tra i redditi dei laureati semplici e quelli con master.

32. Usa la tabella di pagina 55 per trovare i coefficienti di correlazione campionaria tra i salari dei seguenti settori:

(a) ingegneria e finanza;

(b) ingegneria e informatica;

(c) ingegneria e insegnamento;

(d) marketing e chimica.

33. Utilizzando i dati delle prime 10 città elencate nella Tabella 2.5, traccia un diagramma di dispersione e determina il coefficiente di correlazione campionaria tra le temperature di Gennaio e quelle di Luglio.

34. Dimostra le Proprietà 2 e 3 della Proposizione 2.6.1 sul coefficiente di correlazione campionaria.

35. Dimostra la Proprietà 4 della Proposizione 2.6.1 sul coefficiente di correlazione campionaria.

36. In uno studio su bambini dalla seconda alla quarta elementare, un ricercatore sottopose i soggetti ad un test sulle capacità di lettura e ne risultò che esisteva una correlazione positiva tra il punteggio del test e la statura. Egli concluse che i bambini più alti leggevano meglio perché potevano vedere meglio la lavagna. Cosa ne pensi?

Settore e posizione	1975	1980	1985	1988	1989	1990	1991	1992	1993
Salari (dollari americani)									
Insegnanti scolastici	8 233	10 764	15 460	19 400	NP	20 486	21 481	22 171	22 505
Laureati di primo livello:									
Ingegneria	12 744	20 136	26 880	29 856	30 852	32 304	34 236	34 620	35 004
Ragioneria	11 880	15 720	20 628	25 140	25 908	27 408	27 924	28 404	28 020
Commercio e marketing	10 344	15 936	20 616	23 484	27 768	27 828	26 580	26 532	28 536
Amministrazione aziendale	9 768	14 100	19 896	23 880	25 344	26 496	26 256	27 156	27 564
Settore letterario	9 312	13 296	18 828	23 508	25 608	26 364	25 560	27 324	27 216
Chimica	11 904	17 124	24 216	27 108	27 552	29 088	29 700	30 360	30 456
Matematica e statistica	10 980	17 604	22 704	25 548	28 416	28 944	29 244	29 472	30 756
Economia e finanza	10 212	14 472	20 964	23 928	25 812	26 712	26 424	27 708	28 584
Informatica	NP	17 712	24 156	26 904	28 608	29 100	30 924	30 888	31 164
Indice (1975 = 100)									
Insegnanti scolastici	100	131	187	236	NP	249	261	269	273
Laureati di primo livello:									
Ingegneria	100	158	211	234	242	253	268	271	275
Ragioneria	100	132	174	212	218	230	235	239	236
Commercio e marketing	100	154	199	227	268	269	257	256	276
Amministrazione aziendale	100	144	204	244	259	271	268	278	282
Settore letterario	100	143	202	252	275	283	274	293	292
Chimica	100	144	203	228	231	244	249	255	256
Matematica e statistica	100	160	207	233	258	263	266	268	280
Economia e finanza	100	142	205	234	252	261	258	271	280
Informatica (1978 = 100)	NP	125	171	190	202	205	218	217	218

Fonte dati sui laureati: Northwestern University Placement Center, Evanston, IL. The Northwestern University Lindquist-Endicott Report ©.
Fonte dati sugli insegnanti della scuola pubblica: National Education Association, Washington, DC, dati non pubblicati.

Livello e categorie di occupazione per vari titoli di studio in campo scientifico e tecnologico.
(Studenti che hanno conseguito un titolo in campo scientifico e tecnologico nel 1991 o nel 1992. Le occupazioni sono quelle registrate nell'aprile del 1993. Nella categoria *proseguono gli studi* entrano solo gli studenti a tempo pieno. Con *S&I* si intende il settore scientifico-ingegneristico. I dati sui redditi escludono gli studenti titolari di borse e gli imprenditori.)

Titolo e settore	Titolati (migliaia)	Proseguono gli studi	Occupati settore S&I	Altro lavoro	Non occupati	Mediana reddito ($ 1 000)
Diploma di laurea	**639.4**	**22**	**22**	**50**	**6**	**24.0**
Scienze pure	521.1	24	13	57	6	22.1
Matematica e Informatica	77.6	11	32	51	5	28.5
Scienze della vita	99.7	38	14	43	6	21.0
Fisica	33.8	39	28	29	4	26.0
Scienze sociali	310.0	21	6	66	6	21.0
Ingegnerie	118.4	15	60	20	5	33.8
Ingegneria aerospaziale	7.3	23	35	37	6	29.0
Ingegneria chimica	6.7	16	70	10	4	40.0
Ingegneria civile e architettura	15.6	12	69	15	4	31.0
Ingegneria elettrica	41.8	16	59	18	7	35.0
Ingegneria industriale	7.7	7	59	30	3	33.0
Ingegneria meccanica	25.1	13	65	19	3	35.0
Altre	14.1	17	53	25	5	33.0
Master	**115.6**	**23**	**48**	**24**	**5**	**38.1**
Scienze pure	74.6	26	37	31	5	33.8
Matematica e Informatica	24.1	16	48	31	5	40.0
Scienze della vita	13.2	28	35	31	6	29.0
Fisica	10.6	38	46	13	4	34.0
Scienze sociali	26.7	31	24	39	6	28.0
Ingegnerie	41.0	17	68	11	4	42.9
Ingegneria aerospaziale	1.9	26	56	16	3	40.0
Ingegneria chimica	1.7	33	56	7	4	44.0
Ingegneria civile e architettura	4.9	15	74	7	5	38.8
Ingegneria elettrica	15.7	15	71	10	4	44.0
Ingegneria industriale	2.6	13	63	20	4	42.5
Ingegneria meccanica	6.4	17	72	6	4	42.0
Altre	7.9	18	61	18	3	43.0

Fonte: National Science Foundation/SRS. National Survey of Recent College Graduates: 1993.

3 Elementi di probabilità

Contenuto

3.1 Introduzione

Il concetto di probabilità di un evento, quando si effettua un esperimento, è passibile di diverse interpretazioni. Per fare un esempio, immaginiamo che un geologo affermi che in una certa regione vi è il 60% di probabilità che vi sia del petrolio. Tutti probabilmente abbiamo un'idea di cosa questo significhi, e in particolare, la maggior parte delle persone dà una delle due interpretazioni seguenti.

1. Il geologo crede che, trovando molte regioni con caratteristiche esterne simili a quella in esame, circa nel 60% dei casi vi sarà presenza di petrolio.

2. Il geologo crede che sia più verosimile che vi sia petrolio, piuttosto che non vi sia; inoltre 0.6 rappresenta la misura della sua fiducia nell'ipotesi che nella regione in esame vi sia il petrolio.

Queste due interpretazioni del concetto di probabilità di un evento sono note come interpretazione frequentista e interpretazione soggettivistica (o personale). Nell'*interpretazione frequentista* la probabilità di un esito è considerata una proprietà dell'esito stesso. In particolare si pensa che essa possa essere determinata operativamente ripetendo in continuazione l'esperimento, come rapporto tra il numero di cas

in cui si è registrato l'esito sul totale. Questo è il punto di vista prevalente tra gli scienziati.

Nell'*interpretazione soggettivistica*, non si crede che la probabilità di un esito sia una proprietà oggettiva, ma piuttosto la precisazione del livello di fiducia che lo studioso ripone nel verificarsi dell'esito. Questo punto di vista è preferito da alcuni filosofi e analisti finanziari.

Qualunque interpretazione si favorisca, vi è comunque un consenso generale sulla matematica della probabilità, nel senso che – ad esempio – se si stima che vi sia una probabilità di 0.3 che domani piova, e una probabilità di 0.2 che la giornata sia coperta, ma senza pioggia, allora, indipendentemente dall'interpretazione adottata, vi è una probabilità di 0.5 che vi sia pioggia o il cielo sia coperto. In questo capitolo presentiamo le regole e gli assiomi della teoria della probabilità.

3.2 Spazio degli esiti ed eventi

Preliminarmente all'enunciare gli assiomi, occorre introdurre il concetto di spazio degli esiti, e quello di evento.

Si consideri un esperimento il cui esito non sia prevedibile con certezza. Quello che normalmente si può fare comunque, è individuare la rosa degli esiti plausibili. L'insieme di tutti gli esiti possibili si dice *spazio degli esiti* (in inglese, *sample space*), e normalmente si denota con S o con Ω. Quelli che seguono sono alcuni esempi.

Esempio 3.2.1. Se l'esito dell'esperimento consiste nella determinazione del sesso di un neonato, allora poniamo

$$S = \{f, m\}$$

dove si intende che l'esito f rappresenta la nascita di una femmina, e l'esito m quella di un maschio. □

Esempio 3.2.2. Se l'esperimento consiste in una gara tra sette cavalli denotati dai numeri 1, 2, 3, 4, 5, 6 e 7, allora

$$S = \{\text{tutti gli ordinamenti di } (1, 2, 3, 4, 5, 6, 7)\}$$

In questo caso l'esito (2, 3, 7, 6, 5, 4, 1) è quello in cui il cavallo 2 arriva primo, il 3 arriva secondo, il 7 terzo, e così via. □

Esempio 3.2.3. Supponiamo di voler determinare il minimo dosaggio di un farmaco al quale un paziente reagisce positivamente. Una possibile scelta per lo spazio degli esiti di questo esperimento potrebbe essere l'insieme di tutti i numeri positivi, ovvero

$$S = (0, \infty)$$

intendendo ovviamente che l'esito sarebbe x se il paziente reagisse a un dosaggio pari a x e a nessun dosaggio inferiore. □

Spazio degli esiti: insieme degli esiti possibili

I sottoinsiemi dello spazio degli esiti si dicono *eventi*, quindi un evento E è un insieme i cui elementi sono esiti possibili. Se l'esito dell'esperimento è contenuto in E, diciamo che l'evento E si è verificato. Diamo di seguito alcuni esempi.

Nell'Esempio 3.2.1, se poniamo $E = \{f\}$, significa che E è l'evento che il nascituro sia una bambina; se poniamo $F = \{m\}$, F è l'evento che si tratti di un bambino.

Nell'Esempio 3.2.2, se

$$E = \{\text{tutti gli esiti in } S \text{ che incominciano con } 3\}$$

allora E è l'evento che il cavallo 3 risulti vincitore.

La *unione* $E \cup F$ di due eventi E e F dello stesso spazio degli esiti S, è definita come l'insieme formato dagli esiti che stanno o in E o in F. Quindi l'evento $E \cup F$ si verifica se *almeno uno* tra E e F si verifica. Perciò nell'Esempio 3.2.1, se $E = \{f\}$ e $F = \{m\}$, allora $E \cup F = \{f, m\}$, ovvero $E \cup F$ coincide con l'intero spazio degli esiti S. Nell'Esempio 3.2.2, se $E = \{$tutti gli esiti che cominciano con 6$\}$ è l'evento in cui il cavallo 6 arriva primo e $F = \{$tutti gli esiti che hanno 6 in seconda posizione$\}$ è l'evento in cui arriva secondo, allora $E \cup F$ è l'evento in cui il cavallo 6 arriva primo o secondo.

In maniera simile è utile definire la *intersezione* $E \cap F$ di due eventi E e F. Essa è l'insieme formato dagli esiti che sono presenti sia in E, sia in F. Come evento, rappresenta il verificarsi di *entrambi* gli eventi E e F. Quindi nell'Esempio 3.2.3, se $E = (0, 5)$ è l'evento in cui il dosaggio cercato è minore di 5, e $F = (2, 10)$ è l'evento in cui esso è compreso tra 2 e 10, allora $E \cap F = (2, 5)$ è l'evento in cui esso è compreso tra 2 e 5. Nell'Esempio 3.2.2, se $E = \{$tutti gli esiti che terminano con 5$\}$ è l'evento "il cavallo 5 arriva ultimo" e $F = \{$tutti gli esiti che cominciano con 5$\}$ è l'evento "il cavallo 5 arriva primo", allora chiaramente l'evento $E \cap F$ non contiene esiti possibili e non può avvenire mai. Per dare una denominazione ad un tale evento, ci riferiremo ad esso come l'*evento vuoto* e lo rappresenteremo con il simbolo $\varnothing$. Esso è quindi un evento che non contiene esiti possibili per l'esperimento. Se $E \cap F = \varnothing$, ovvero se E e F non possono verificarsi entrambi, li diremo eventi *mutuamente esclusivi* o eventi *disgiunti*.

Per ogni evento E, definiamo l'evento E^c, che diciamo il *complementare* di E, come l'insieme formato dagli esiti di S che non stanno in E. Quindi E^c si verifica se e solo se non si verifica E. Nell'Esempio 3.2.1, se l'evento $E = \{m\}$ si verifica quando il neonato è maschio, allora $E^c = \{f\}$ è l'evento che il neonato sia femmina. Si noti infine come valga la ovvia relazione $S^c = \varnothing$.

Se, per una coppia di eventi E e F accade che tutti gli esiti di E appartengono anche a F, si dice che E è contenuto in F, e si scrive $E \subset F$ (o, in modo equivalente, $F \supset E$). Chiaramente questo significa che se si verifica E, si verifica necessariamente anche F. Se valgono entrambe le relazioni $E \subset F$ e $F \subset E$, allora diciamo che E e F sono uguali, e scriviamo $E = F$.

È anche possibile definire l'unione o l'intersezione di più di due eventi. In particolare, l'unione degli eventi $E_1, E_2, \ldots, E_n$, che indichiamo con $E_1 \cup E_2 \cup \cdots \cup E_n$ o con $\bigcup_{i=1}^{n} E_i$ è l'evento formato da tutti gli esiti che appartengono ad almeno uno degli E_i. L'intersezione degli stessi eventi viene indicata con $E_1 \cap E_2 \cap \cdots \cap E_n$ o con $\bigcap_{i=1}^{n} E_i$, ed è l'evento formato dagli esiti che appartengono a tutti gli E_i, per $i = 1, 2, \ldots, n$. In altre parole, l'unione degli E_i si verifica se *almeno uno* degli eventi E_i si verifica, mentre l'intersezione degli E_i si verifica solo se *tutti* gli eventi E_i si verificano.

3.3 I diagrammi di Venn e l'algebra degli eventi

Un tipo di rappresentazione grafica degli eventi, molto utile per illustrare le relazioni logiche che li legano, sono i *diagrammi di Venn*. Lo spazio degli esiti S è rappresentato da un grande rettangolo che contiene il resto della figura, oppure dal foglio stesso. Gli eventi da prendere in considerazione, invece, sono rappresentati da cerchi o altre curve chiuse disegnate all'interno del rettangolo. A questo punto, tutti gli eventi complessi di nostro interesse possono essere evidenziati colorando opportune regioni del diagramma. Ad esempio nei tre diagrammi di Venn illustrati in Figura 3.1, le regioni scurite rappresentano, nell'ordine, gli eventi $E \cup F$, $E \cap F$ ed E^c. Il diagramma di Venn della Figura 3.2 invece, mostra che $E \subset F$

Gli operatori unione, intersezione e complementare, obbediscono a regole non dissimili da quelle dell'algebra dell'addizione e della moltiplicazione dei numeri reali. Ne elenchiamo solo alcune: si tratta delle proprietà commutative (3.3.1), associative (3.3.2) e distributive (3.3.3).

$$E \cup F = F \cup E \qquad E \cap F = F \cap E \tag{3.3.1}$$

$$(E \cup F) \cup G = E \cup (F \cup G) \qquad (E \cap F) \cap G = E \cap (F \cap G) \tag{3.3.2}$$

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G) \qquad (E \cap F) \cup G = (E \cup G) \cap (F \cup G) \tag{3.3.3}$$

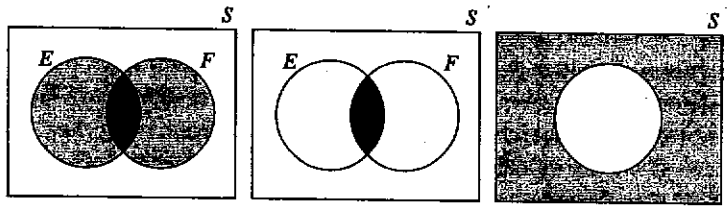


Figura 3.1 Diagrammi di Venn.

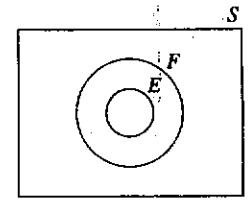


Figura 3.2 Diagramma di Venn che illustra la relazione $E \subset F$.

Il modo rigoroso per dimostrare queste identità consiste nel verificare che ogni esito appartenente all'evento al primo membro è anche contenuto nell'evento al secondo membro, e viceversa. Un diverso approccio, meno formale e più intuitivo, consiste nell'usare i diagrammi di Venn. Ad esempio, la prima delle due proprietà distributive può essere verificata dalla sequenza di diagrammi che compare in Figura 3.3.

Esistono due relazioni particolarmente utili che mettono in gioco tutte e tre le operazioni base che si possono fare sugli eventi. Sono le *leggi di De Morgan*:

$$\begin{aligned} (E \cup F)^c &= E^c \cap F^c \\ (E \cap F)^c &= E^c \cup F^c \end{aligned} \Rightarrow \text{De Morgan} \tag{3.3.4}$$

3.4 Assiomi della probabilità

Se si ripete molte volte un esperimento mettendosi sempre nelle stesse condizioni, si verifica empiricamente che la frazione di casi sul totale in cui si realizza un qualunque evento E tende – al crescere dei tentativi – ad un valore costante che dipende solo da E. Tutti sanno ad esempio, che se si lancia tante volte una moneta, il rapporto tra il numero di risultati *testa* e il numero di tentativi, man mano che aumentiamo il numero di lanci, tende ad un valore costante (cioè 0.5). Il valore limite della fre-

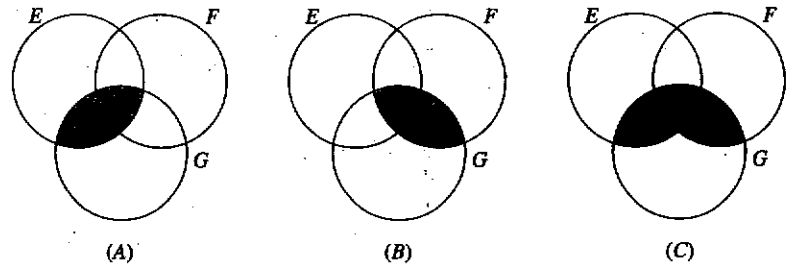


Figura 3.3 Illustrazione di una proprietà distributiva per mezzo dei diagrammi di Venn.

quenza empirica di un evento è quello che molti hanno in mente quando cercano di descrivere la probabilità di quell'evento.

Quale che sia la definizione di probabilità che vogliamo abbracciare, vi è un comune accordo sulle regole che tali probabilità devono rispettare: da qui in poi il modo di procedere diviene allora esclusivamente astratto. Si associa ad ogni evento E sullo spazio degli esiti S, un numero che si denota con $P(E)$ e che si dice *probabilità* dell'evento E. Ciò non può essere fatto in maniera completamente libera: le probabilità dei vari eventi devono rispettare alcuni assiomi dal significato intuitivo.

$$0 \le P(E) \le 1 \qquad \text{(Assioma 1)}$$

$$P(S) = 1 \qquad \text{(Assioma 2)}$$

Inoltre per ogni successione di eventi mutuamente esclusivi $E_1, E_2, \ldots$ (cioè tali che $E_i \cap E_j = \varnothing$ quando $i \neq j$),

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i), \qquad n = 1, 2, \ldots, \infty \qquad \text{(Assioma 3)}$$

Il primo assioma afferma che ogni probabilità è un numero compreso tra 0 e 1. Il secondo stabilisce che l'evento S si verifica con probabilità 1, ovvero vi è assoluta certezza che si realizzi un esito contenuto in S, o ancora, S contiene necessariamente tutti gli esiti possibili del nostro esperimento. Il terzo assioma infine afferma che, preso un insieme finito o numerabile di eventi mutuamente esclusivi, la probabilità che se ne verifichi almeno uno è uguale alla somma delle loro probabilità.

Si può a questo punto notare che se si interpreta $P(E)$ come la frequenza relativa dell'evento E quando l'esperimento è ripetuto un gran numero di volte, questa definizione soddisfa i predetti assiomi. Infatti è certo che la frequenza relativa di un evento sia sempre compresa tra 0 e 1; è altrettanto sicuro che l'evento S si verifica ad ogni esperimento, e quindi ha una frequenza relativa sempre uguale a 1; si può anche notare che se E e F sono eventi che non hanno esiti in comune, il numero di casi in cui si verifica $E \cup F$ è pari alla somma di quelli in cui si verificano E e F, quindi la frequenza relativa dell'unione è pari alla somma delle frequenze relative. Per illustrare meglio questo concetto, supponiamo che l'esperimento in questione sia il lancio di una coppia di dadi, e denotiamo con E l'evento che la loro somma sia pari a 2, 3 o 12, mentre l'evento F sarà composto dagli esiti in cui la somma vale 7 o 11. Allora se dopo molte prove, E si è verificato nell'11% dei casi, e F nell 22%, è facile accettare che nel 33% dei casi la somma dei dadi è stata 2, 3, 12, 7 o 11.

Gli assiomi permettono di dedurre un gran numero di proprietà delle probabilità degli eventi. Ad esempio, possiamo notare che E e E^c sono eventi disgiunti, e quindi usando gli Assiomi 2 e 3,

$$1 = P(S) = P(E \cup E^c) = P(E) + P(E^c)$$

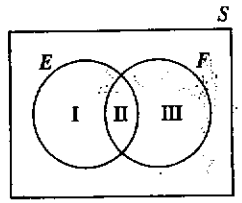


Figura 3.4 Illustrazione della Proposizione 3.4.2 con un diagramma di Venn.

ovvero:

Proposizione 3.4.1. Per ogni evento $E \subset S$, vale la relazione

$$P(E^c) = 1 - P(E) \qquad (3.4.1)$$

La probabilità che un evento qualsiasi non si verifichi è pari a uno meno la probabilità che si verifichi. Ad esempio, se sappiamo che la probabilità di ottenere *testa* lanciando una cerca moneta è 3/8, allora evidentemente la probabilità di ottenere *croce* dalla stessa moneta è 5/8.

La prossima proposizione fornisce la probabilità dell'unione di due eventi in termini delle loro probabilità singole e di quella dell'intersezione. (Si noti che questa rappresenta una estensione dell'Assioma 3 che funziona anche con eventi non mutuamente esclusivi.)

Proposizione 3.4.2. Se E e F sono due eventi qualsiasi, allora

$$P(E \cup F) = P(E) + P(F) - P(E \cap F) \qquad (3.4.2)$$

Dimostrazione. I diagrammi di Venn forniscono una dimostrazione molto intuitiva. Si osservi la Figura 3.4; poiché le regioni I, II e III sono disgiunte, si può applicare tre volte l'Assioma 3 per ottenere

$$P(E \cup F) = P(\mathrm{I}) + P(\mathrm{II}) + P(\mathrm{III})$$
$$P(E) = P(\mathrm{I}) + P(\mathrm{II})$$
$$P(F) = P(\mathrm{II}) + P(\mathrm{III})$$

Confrontando le tre identità si vede che

$$P(E \cup F) = P(E) + P(F) - P(\mathrm{II})$$

e la dimostrazione è conclusa, poiché $\mathrm{II} = E \cap F$ □

Esempio 3.4.1. La percentuale di maschi americani che fuma la sigaretta è del 28%; quelli che fumano il sigaro sono il 7%; quelli che fumano entrambi sono il 5%. Qual è la percentuale di chi non fuma né la sigaretta né il sigaro?.

Immaginiamo di selezionare un individuo a caso nella categoria degli Statunitensi di sesso maschile. Sia E l'evento che egli fumi la sigaretta e F l'evento che sia un fumatore di sigari. La probabilità che si realizzi almeno uno dei due eventi è data da

$$P(E \cup F) = P(E) + P(F) - P(E \cap F) = 0.07 + 0.28 - 0.05 = 0.3$$

Perciò la probabilità che l'individuo selezionato non sia un fumatore è pari a 0.7 o al 70%. Se ne deduce che questa deve essere anche la percentuale cercata. □

3.5 Spazi di esiti equiprobabili

Per tutta una serie di esperimenti è naturale assumere che ogni esito di uno spazio S abbia la stessa probabilità di realizzarsi. Ciò può accadere solo se S è un insieme finito (perché?), e in questo caso, si può assumere senza perdita di generalità che sia $S = \{1, 2, \ldots, N\}$; in queste ipotesi l'equiprobabilità degli esiti si scrive

$$P(\{1\}) = P(\{2\}) = \cdots = P(\{N\}) =: p$$

Dagli Assiomi 2 e 3 segue che

$$1 = P(S) = P(\{1\}) + P(\{2\}) + \cdots + P(\{N\}) = Np$$

da cui si deduce che $P(\{i\}) = p = 1/N$, per tutti gli $i = 1, 2, \ldots, N$. Da questo risultato e ancora dall'Assioma 3 si conclude che per ogni evento E,

$$P(E) = \frac{\#E}{N} \tag{3.5.1}$$

dove con $\#E$ si intende il numero di elementi di E. In altre parole se si assume che ogni esito di S abbia la medesima probabilità, allora la probabilità di un qualunque evento E è pari al rapporto tra il numero di esiti contenuti in E e il numero totale di esiti di S.

Una conseguenza notevole di questo risultato è che occorre sapere contare efficacemente il numero di esiti differenti appartenenti ad un evento. A questo scopo faremo uso della regola seguente.

Osservazione 3.5.1 (Principio di enumerazione). Consideriamo la realizzazione di due diversi esperimenti (detti 1 e 2), che possono avere rispettivamente m e n esiti differenti. Allora complessivamente vi sono mn diversi risultati se si considerano entrambi gli esperimenti contemporaneamente.

Dimostrazione. L'enunciato si dimostra enumerando tutte le possibili coppie di risultati dei due esperimenti, che sono:

$$\begin{array}{cccc} (1,1) & (1,2) & \ldots & (1,n) \\ (2,1) & (2,2) & \ldots & (2,n) \\ \vdots & \vdots & & \vdots \\ (m,1) & (m,2) & \ldots & (m,n) \end{array}$$

dove si intende che si ottiene il risultato (i, j) se nell'esperimento 1 si realizza l'esito i-esimo tra gli m possibili, e nell'esperimento 2 quello j-esimo tra gli n possibili. Siccome la tabella ottenuta ha m righe e n colonne, vi sono complessivamente mn esiti possibili. □

Esempio 3.5.1. Si estraggono a caso due palline da un'urna che ne contiene 6 di bianche e 5 di nere. Qual è la probabilità che le due estratte siano una bianca e una nera?

Se consideriamo le due estrazioni con il loro ordine, la prima pallina viene scelta tra le 11 presenti nell'urna all'inizio, mentre la seconda tra le 10 che restano dopo la prima estrazione. Lo spazio degli esiti ha quindi in tutto $10 \times 11 = 110$ elementi. Inoltre, vi sono $6 \times 5 = 30$ casi in cui la prima estratta è bianca e la seconda nera, e similmente 5×6 casi in cui la prima è nera e la seconda bianca. Quindi se assumiamo che l'ipotesi di "estrazione casuale" stia a significare che i 110 esiti devono intendersi equiprobabili, concludiamo che la probabilità cercata è

$$\frac{30 + 30}{110} = \frac{6}{11} \quad \square$$

Generalizzazione del principio di enumerazione

Se si eseguono r esperimenti, ed è noto che il primo esperimento ammette n_1 esiti possibili, per ognuno dei quali il secondo esperimento ammette n_2 esiti diversi, inoltre se per ogni combinazione di esiti dei primi due esperimenti il terzo ammette n_3 esiti diversi, e così via, allora vi sono un totale di $n_1 \times n_2 \times \cdots \times n_r$ combinazioni di esiti degli r esperimenti considerati tutti insieme.

Il principio di enumerazione ammette una utile generalizzazione, descritta nel riquadro presentato in queste pagine. Per illustrarne un'applicazione, proviamo a determinare il numero di modi diversi in cui si possono ordinare n oggetti. Per esempio, il numero di modi in cui si possono ordinare i tre simboli a, b e c sono sei, ovvero esplicitamente, *abc*, *acb*, *bac*, *bca*, *cab* e *cba*. Ciascuno di questi ordinamenti prende

il nome di *permutazione* dei tre simboli considerati; le permutazioni di tre elementi sono perciò sei. Vediamo come questo risultato fosse deducibile dal principio di enumerazione generalizzato. Il primo simbolo della permutazione può essere scelto in tre modi diversi; per ogni scelta del primo simbolo, il secondo può essere preso tra i due restanti; il terzo e ultimo viene individuato per esclusione (una sola scelta). Quindi vi sono $3 \times 2 \times 1 = 6$ possibili permutazioni.

Supponiamo ora di avere n oggetti. Se ragioniamo in modo analogo, scopriamo che vi sono

$$n(n-1)(n-2)\cdots 3\cdot 2\cdot 1 =: n!$$

diverse permutazioni degli n oggetti. Tale valore viene normalmente denotato con $n!$ e viene detto "n fattoriale". Alcuni esempi sono, $1! = 1$, $2! = 2$, $3! = 6$, $4! = 24$, $5! = 120$ e così via. Risulterà anche conveniente porre $0! = 1$.

Esempio 3.5.2. Una corso di probabilità è frequentato da 10 studenti: 6 maschi e 4 femmine. Viene effettuato un esame, e i punteggi degli studenti sono tutti diversi. **(a)** Quante diverse classifiche sono possibili? **(b)** Se tutte le classifiche si pensano equiprobabili, qual è la probabilità che le quattro studentesse ottengano i punteggi migliori?

(a) Siccome ogni classifica è associata ad una precisa permutazione dei dieci studenti, esse in tutto sono $10! = 3\,628\,800$. **(b)** Poiché vi sono $4!$ diverse classifiche delle studentesse tra di loro e $6!$ classifiche dei maschi, segue dal principio di enumerazione che vi sono $4! \times 6! = 24 \times 720 = 17\,280$ possibili classifiche in cui le studentesse occupano le prime 4 posizioni. Quindi la probabilità cercata è

$$\frac{4!\cdot 6!}{10!} = \frac{4\cdot 3\cdot 2\cdot 1}{10\cdot 9\cdot 7\cdot 6} = \frac{1}{210} \quad \square$$

Esempio 3.5.3. Se in una stanza sono radunate n persone, qual è la probabilità che non ve ne siano due che compiono gli anni lo stesso giorno dell'anno? Quanto grande deve essere n affinché tale probabilità sia minore di $1/2$?

Siccome ogni persona può celebrare il compleanno in uno qualsiasi dei 365 giorni, vi sono in tutto 365^n diversi esiti dell'esperimento consistente nel domandare a ciascun partecipante la data di nascita. (Sì, stiamo ignorando la possibilità che qualcuno sia nato il 29 febbraio di un anno bisestile.) Secondariamente, vi sono in tutto $365\cdot 364\cdot 363\cdots(365-n+1)$ esiti che fanno sì che tutte le persone abbiano date di compleanno diverse. Infatti la prima persona può compiere gli anni in uno qualsiasi dei 365 giorni dell'anno; la seconda – non potendo usare la stessa data – può essere nata in uno dei 364 giorni rimanenti; la terza in uno dei 363 giorni diversi da quelli delle prime due, e così via fino all'ultima persona, che ha $365 - n + 1$ date libere in cui può compiere gli anni. Allora, assumendo che ciascun esito sia equiprobabile, la probabilità cercata è pari a

$$\frac{365\cdot 364\cdot 363\cdots(365-n+1)}{365^n}$$

Anche se può sembrare sorprendente, già con $n = 23$, questo prodotto diviene minore di $1/2$. Ovvero, se si riuniscono almeno 23 persone, la probabilità che tra di loro ve ne siano due che compiono gli anni lo stesso giorno supera il 50%. Molte persone trovano questo risultato inaspettato e antiintuitivo, ma forse è ancora più straordinario il fatto che se $n = 50$ la probabilità raggiunge 0.970, e che se $n = 100$ addirittura la probabilità che vi siano due compleanni coincidenti è di più di tre milioni a uno (con questa locuzione si intende che essa è maggiore di $(3 \times 10^6)/(3 \times 10^6 + 1)$). $\square$

3.5.1 Il coefficiente binomiale

Ci rivolgiamo ora ad un diverso problema di calcolo combinatorio. Vogliamo infatti determinare il numero di diversi gruppi di r oggetti che si possono formare scegliendoli da un insieme di n. Ad esempio, quanti diversi gruppi di tre lettere si possono formare usando le cinque lettere A, B, C, D, E? Si può ragionare nel modo seguente. Vi sono 5 scelte per la prima lettera, 4 per la seconda e 3 per la terza, vi sono quindi $5 \times 4 \times 3$ modi per scegliere tre lettere su cinque, tenendo conto dell'ordine. Tuttavia, ogni gruppo di tre lettere viene contato più volte, perché stiamo tenendo conto dell'ordine. Ad esempio la tripletta A, C, D, compare come ACD, ADC, CAD, CDA, DAC e DCA, ovvero in tutte le sue 6 permutazioni. Poiché stiamo contando $6 = 3!$ volte ogni gruppo di tre lettere, se ne deduce che il numero di gruppi diversi di tre lettere può essere ricavato come $(5 \times 4 \times 3)/(3 \times 2 \times 1) = 10$.

Più in generale, poiché il numero di modi diversi di scegliere r oggetti su n tenendo conto dell'ordine è dato da $n(n-1)\cdots(n-r+1)$, e poiché ogni gruppo di lettere fissato viene contato $r!$ volte (una per ogni sua permutazione), il numero di diversi gruppi di r elementi, scelti in un insieme di n oggetti è dato dalla formula

$$\frac{n(n-1)\cdots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!} =: \binom{n}{r} \tag{3.5.2}$$

Questo valore si dice il numero di *combinazioni* di n elementi presi r alla volta e si indica con il simbolo $\binom{n}{r}$, che prende il nome di *coefficiente binomiale*.

Per fare qualche esempio, vi sono

$$\binom{8}{2} = \frac{8\times 7}{2\times 1} = 28$$

gruppi diversi di due elementi su un insieme di 8, e

$$\binom{10}{2} = \frac{10\times 9}{2\times 1} = 45$$

coppie diverse di individui in un gruppo di 10 persone. Poiché inoltre $0! = 1$, si noti che vale

$$\binom{n}{0} = 1 = \binom{n}{n} \tag{3.5.3}$$

Esempio 3.5.4. Una commissione di 5 elementi deve essere selezionata da un gruppo di 6 uomini e 9 donne. Se la scelta viene fatta a caso, che probabilità vi è che vengano presi 3 uomini e due donne?

Cominciamo con il supporre che con "scelta fatta a caso" si intenda che le $\binom{15}{5}$ possibili combinazioni sono tutte equiprobabili. Ci sono allora $\binom{6}{3}$ possibili scelte per i tre uomini e $\binom{9}{2}$ scelte per le due donne. Ne segue che la probabilità cercata è data da

$$\frac{\binom{6}{3}\binom{9}{2}}{\binom{15}{5}} = \frac{240}{1001} \quad \square$$

Esempio 3.5.5. Da un insieme di n elementi si estrare a caso un sottoinsieme di cardinalità k. Qual è la probabilità che un elemento fissato precedentemente tra gli n iniziali si trovi tra i k estratti?

Il numero di gruppi di cardinalità k che contiene l'elemento fissato è $\binom{1}{1}\binom{n-1}{k-1}$. La probabilità cercata è quindi

$$\binom{n-1}{k-1}\binom{n}{k}^{-1} = \frac{(n-1)!}{(n-k)!(k-1)!}\frac{(n-k)!k!}{n!} = \frac{k}{n} \quad \square$$

Esempio 3.5.6. Una squadra di basket è composta di 6 giocatori di colore e 6 bianchi. Essi devono essere divisi a coppie per occupare sei camere doppie. Se la suddivisione viene fatta a caso qual è la probabilità che nessun nero sia in camera con un bianco?

Inizialmente immaginiamo che le sei coppie di compagni di stanza siano numerate, ovvero che si distingua tra la coppia 1, la coppia 2, eccetera. Per la prima coppia vi sono $\binom{12}{2}$ possibili scelte; per ognuna di esse ve ne sono $\binom{10}{2}$ per la seconda coppia; per ogni scelta delle prime due coppie vi sono $\binom{8}{2}$ possibilità per la terza coppia e così via. Per il principio di enumerazione generalizzato si deduce che vi sono

$$\binom{12}{2}\binom{10}{2}\binom{8}{2}\binom{6}{2}\binom{4}{2}\binom{2}{2} = \frac{12!}{2^6}$$

modi di dividere i dodici giocatori in sei coppie distinte. Quindi vi sono $12!/\{2^6 6!\}$ suddivisioni in coppie senza tenere conto dell'ordine. Analogamente, vi sono $6!/\{2^3 3!\}$ modi di appaiare i sei giocatori di colore tra di loro (senza ordine) e altrettanti per i bianchi. Poiché si sceglie a caso tra suddivisioni equiprobabili, il valore cercato è dato da

$$\left(\frac{6!}{2^3 3!}\right)^2 \cdot \frac{2^6 6!}{12!} = \frac{5}{231} \approx 0.0216$$

Quindi, vi sono solo circa due probabilità su cento che sorteggiando le camere non capiti che un bianco e un nero dividano la stessa stanza. $\square$

3.6 Probabilità condizionata

In questa sezione presentiamo e sviluppiamo uno dei concetti fondamentali della teoria della probabilità – quello di probabilità condizionata. L'importanza che ha è duplice. In primo luogo, accade spesso di volere calcolare delle probabilità quando si è in possesso di informazioni parziali sull'esito dell'esperimento, o di volerle ricalcolare una volta ottenute nuove informazioni. Quelle di questo tipo sono probabilità condizionate. Secondariamente vi è una sorta di bonus nel fatto che a volte il modo più semplice di determinare la probabilità di un evento complesso, consiste nel condizionarlo al realizzarsi o meno di un evento accessorio.

Per illustrare questo concetto, immaginiamo di tirare due dadi. Lo spazio degli esiti di questo esperimento può essere descritto da

$$S = \{(i,j), \quad i = 1,2,\ldots,6, \quad j = 1,2,\ldots,6\}$$

dove si intende che si ottiene l'esito (i,j) se il risultato del primo dado è i e quello del secondo j. Supponiamo che ciascuno dei 36 esiti di S abbia la stessa probabilità, ovvero $1/36$. (In queste ipotesi si dice che i due dadi sono onesti.) Supponiamo infine che il primo dado sia risultato in un 3. Allora, possedendo questa informazione, qual è la probabilità che la somma dei due dadi valga 8? Dato che il primo dado ha totalizzato un 3, vi sono solo 6 risultati possibili per l'esperimento, che sono $(3,1)$, $(3,2)$, $(3,3)$, $(3,4)$, $(3,5)$ e $(3,6)$. Inoltre, siccome in origine ciascuno di questi esiti aveva la stessa probabilità di realizzarsi, essi dovrebbero essere ancora equiprobabili. Ciò significa che, se il primo dado ha dato un 3, allora la probabilità (condizionata) di ciascuno degli esiti possibili $(3,1)$, $(3,2)$, $(3,3)$, $(3,4)$, $(3,5)$, $(3,6)$ è $1/6$, mentre la probabilità (condizionata) degli altri 30 elementi di S è 0. Se ne conclude che la probabilità cercata è $1/6$.

Se denotiamo con E e F rispettivamente l'evento che la somma dei due dadi valga 8 e l'evento che il primo dado risulti in un 3, allora la probabilità che abbiamo appena calcolato si dice *probabilità condizionata di E dato F*, e si denota con

$$P(E|F)$$

Con un ragionamento analogo a quello dell'esempio è possibile trovare una formula generale per $P(E|F)$, valida per qualunque coppia di eventi (si veda la Figura 3.5). Infatti, se si è verificato l'evento F, affinché si verifichi anche E, il caso avere favorito un elemento che sta sia in E sia in F, ovvero che appartiene all'intersezione $E \cap F$. In secondo luogo essendosi verificato F, questo evento diviene il nuovo (ridotto) spazio degli esiti e per questo la probabilità condizionata dell'evento $E \cap F$ sarà pari al rapporto tra la sua probabilità e quella di F. In formula,

$$P(E|F) := \frac{P(E \cap F)}{P(F)} \tag{3.6.1}$$

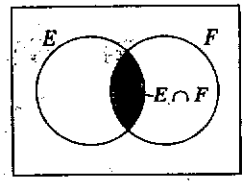


Figura 3.5 $P(E|F) := P(E \cap F)/P(F)$.

Si noti che l'Equazione (3.6.1) ha senso solo se $P(F) > 0$ e infatti in caso contrario $P(E|F)$ non si definisce.

La definizione di probabilità condizionata che compare nell'Equazione (3.6.1) è compatibile con l'interpretazione frequentista della probabilità degli eventi. Supponiamo di realizzare un numero molto elevato n di ripetizioni di un esperimento. Poiché $P(F)$ è il limite della frazione di prove in cui si verifica F, su un numero elevato n di tentativi, saranno circa $nP(F)$ quelli in cui si realizza F. Analogamente saranno approssimativamente $nP(E \cap F)$ quelli in cui si realizzano sia E sia F. Perciò limitatamente agli esperimenti che hanno visto la realizzazione F, la frazione di quelli per i quali ha avuto luogo anche l'evento E è circa uguale a

$$\frac{nP(E \cap F)}{nP(F)} = \frac{P(E \cap F)}{P(F)}$$

Le approssimazioni fatte divengono esatte quando n tende all'infinito, e quindi la (3.6.1) è la corretta definizione di probabilità di E qualora di sia verificato F.

Esempio 3.6.1. Una confezione contiene 5 transistor guasti (non funzionano per niente), 10 difettosi (funzionano correttamente per qualche ora e poi si guastano) e 25 accettabili. Si sceglie un transistor a caso. Qual è la probabilità che sia accettabile se inizialmente funziona?

Sappiamo che non si tratta di uno dei 5 guasti, perché per il momento sta funzionando. Consentendoci un rilassamento nella notazione[1], la quantità cercata si può esprimere come

$$\begin{aligned} P(\text{accettabile}|\text{non guasto}) &= \frac{P(\text{accettabile, non guasto})}{P(\text{non guasto})} \\ &= \frac{P(\text{accettabile})}{P(\text{non guasto})} \end{aligned}$$

[1] Sarebbe infatti più corretto scrivere $P(\{\text{accettabile}\}|\{\text{non guasto}\})$, ma alla lunga esagerare con le parentesi distrae l'attenzione. Si noti anche che la virgola nell'argomento di $P(\cdot)$ denota l'intersezione degli eventi descritti ai suoi lati. Questo tipo di notazione è assai comune e sarà usata ancora.

dove la seconda uguaglianza segue perché i transistor contemporaneamente accettabili e non guasti sono esattamente quelli accettabili. Assumendo allora che i 40 transistor possano essere scelti con uguale probabilità, si ottiene

$$P(\text{accettabile}|\text{non guasto}) = \frac{25/40}{35/40} = \frac{5}{7}$$

È utile notare che si sarebbe arrivati al medesimo risultato operando direttamente sullo spazio degli esiti ridotto. Infatti, sapendo che il pezzo scelto non è guasto, il problema si riduce a calcolare con che probabilità un transistor scelto da una confezione con 25 pezzi accettabili e 10 difettosi, risulti accettabile. Questa probabilità è ovviamente $25/35$. □

Esempio 3.6.2. La organizzazione per cui lavora il signor Jones organizza una cena tra uomini per i dipendenti e i loro figli. Sono invitati i dipendenti padri di figli maschi, assieme al minore fra i loro figli maschi. Jones ha due figli, ed è invitato alla cena. Qual è la probabilità condizionata che entrambi i suoi figli siano maschi?

Lo spazio degli esiti è $S := \{(m,m),(m,f),(f,m),(f,f)\}$, dove con (m,f) si intende che il figlio maggiore è maschio e la minore è femmina; prima di condizionare, tutti gli esiti sono equiprobabili. L'informazione che Jones è invitato alla cena equivale a sapere che almeno uno dei suoi figli è maschio, quindi che non si è verificato l'evento (f,f). Denotando con A e B gli eventi "almeno un figlio è maschio" e "entrambi i figli sono maschi", la quantità cercata è $P(B|A)$, ovvero (si noti come, volendo essere precisi, ciascuna parentesi sia necessaria):

$$\begin{aligned} P(B|A) &= \frac{P(A \cap B)}{P(A)} \\ &= \frac{P(\{(m,m)\})}{P(\{(m,m),(m,f),(f,m)\})} \\ &= \frac{1/4}{3/4} = \frac{1}{3} \end{aligned}$$

Molte persone pensano erroneamente che la probabilità che entrambi i figli siano maschi sia $1/2$, anziché $1/3$, essendo convinte che il figlio di Jones che non partecipa alla cena abbia la stessa probabilità di essere maschio o femmina. Si rammenti tuttavia che inizialmente i quattro esiti erano equiprobabili, e il sapere che almeno un figlio è maschio equivale a escludere l'esito (f,f). Questo ci lascia con tre esiti equiprobabili, mostrando che vi sono il doppio delle possibilità che l'altro figlio di Jones sia femmina piuttosto che maschio. La risposta sarebbe stata $1/2$ ad esempio se avessimo avuto l'informazione che il minore dei figli di Jones è maschio. (Ci si convinca di questa affermazione, quindi si affronti il Problema 32.) □

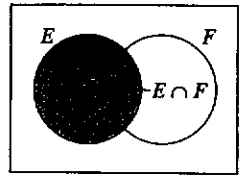


Figura 3.6 $E = (E \cap F) \cup (E \cap F^c)$.

Se si moltiplicano entrambi i membri dell'Equazione (3.6.1) per $P(F)$, si trova

$$P(E \cap F) = P(E|F)P(F) \tag{3.6.2}$$

Parafrasandola, l'Equazione (3.6.2) dice che la probabilità che E e F si verifichino entrambi è pari quella che si verifichi F per la probabilità condizionata di E dato che si è verificato F. Questa formula mostra la sua utilità quando si vuole calcolare la probabilità di una intersezione, come illustra l'esempio seguente.

Esempio 3.6.3. Il signor Perez è convinto che vi sia il 30% di probabilità che la sua azienda apra un nuovo ufficio a Phoenix. Nel caso ciò si verifichi, egli stima di avere un 60% di probabilità di assumere il ruolo dirigenziale nella nuova filiale. Che probabilità vi è che egli divenga il manager nel nuovo ufficio di Phoenix?

Se denotiamo con U l'evento "viene aperto un nuovo ufficio a Phoenix" e con M l'evento "Perez viene promosso manager a Phoenix", allora la probabilità cercata è $P(U \cap M)$, ovvero,

$$\begin{aligned} P(U \cap M) &= P(M|U)P(U) \\ &= 0.6 \times 0.3 = 0.18 \end{aligned}$$

Quindi vi è una probabilità del 18% che Perez divenga il manager a Phoenix. □

3.7 Fattorizzazione di un evento e formula di Bayes

Siano E ed F due eventi qualsiasi. È possibile esprimere E come

$$E = (E \cap F) \cup (E \cap F^c)$$

Infatti ogni punto che appartiene all'evento E, o sta sia in E sia in F, oppure sta in E ma non in F (si veda la Figura 3.6). Inoltre, visto che $E \cap F$ e $E \cap F^c$ sono eventi disgiunti, si ha per l'Assioma 3,

$$\begin{aligned} P(E) &= P(E \cap F) + P(E \cap F^c) \\ &= P(E|F)P(F) + P(E|F^c)P(F^c) \\ &= P(E|F)P(F) + P(E|F^c)[1 - P(F)] \end{aligned} \tag{3.7.1}$$

L'Equazione (3.7.1) afferma che la probabilità dell'evento E si può ricavare come media pesata delle probabilità condizionali di E sapendo: **(1)** che F si è verificato e **(2)** che non si è verificato. I pesi corretti sono le probabilità degli eventi rispetto a cui si condiziona. Questa formula è estremamente utile, in quanto in molte situazioni non è possibile calcolare una probabilità complessa direttamente, mentre essa è facilmente ricavabile dalla (3.7.1), condizionando al verificarsi o meno di un secondo evento. L'evento accessorio va scelto in modo che, una volta che si sappia se esso si è verificato o meno, risulti evidente la probabilità dell'evento complesso di partenza, tenendo conto di questa informazione.

Esempio 3.7.1. Una società di assicurazioni ritiene che la popolazione possa essere divisa in due categorie: quella delle persone inclini a provocare incidenti e quella delle persone non inclini. I rilevamenti statistici effettuati mostrano che una persona incline agli incidenti ha un incidente in un anno con probabilità 0.4, mentre questa probabilità si riduce a 0.2 per l'altra categoria. Assumendo che il 30% della popolazione sia incline agli incidenti, quanto vale la probabilità che un nuovo assicurato abbia un incidente entro un anno dalla stipula del contratto assicurativo?

Otteniamo la probabilità richiesta condizionando alla categoria di appartenenza del nuovo assicurato. Se denotiamo con A_1 l'evento "avrà un incidente entro un anno" e con H l'evento "è incline ad avere incidenti", otteniamo per $P(A_1)$,

$$\begin{aligned} P(A_1) &= P(A_1|H)P(H) + P(A_1|H^c)P(H^c) \\ &= 0.4 \times 0.3 + 0.2 \times 0.7 = 26\% \end{aligned}$$ □

Negli esempi che seguono mostriamo come rivalutare la probabilità dell'evento condizionante (F nella notazione della (3.7.1)), alla luce di informazioni addizionali (come il verificarsi dell'evento E).

Esempio 3.7.2. Riconsideriamo l'Esempio 3.7.1 e supponiamo che il nuovo assicurato abbia un incidente entro un anno dalla stipula del contratto. Qual è la probabilità che appartenga alla categoria delle persone inclini agli incidenti?

Nell'esempio iniziale assumevamo per un nuovo assicurato una probabilità del 30% che fosse incline ad avere incididenti, quindi, $P(H) = 0.3$. Tuttavia con la nuova informazione che A_1 si è verificato, possiamo stimare più correttamente questa probabilità, nel modo seguente.

$$\begin{aligned} P(H|A_1) &= \frac{P(H \cap A_1)}{P(A_1)} \\ &= \frac{P(A_1|H)P(H)}{P(A_1)} \\ &= \frac{0.3 \times 0.4}{0.26} = \frac{6}{13} \approx 0.4615 \end{aligned}$$ □

Esempio 3.7.3. In una prova a risposte multiple, nel rispondere ad una domanda uno studente può conoscere la risposta, oppure provare a indovinarla. Sia p la probabilità che conosca la risposta e $1-p$ la probabilità che tiri a indovinare. Si assuma che, se prova ad indovinare, risponda correttamente con probabilità $1/m$, dove m è il numero di alternative nelle scelte multiple. Qual è la probabilità condizionata che egli conoscesse la risposta a una domanda alla quale ha risposto correttamente?

Siano C e K rispettivamente gli eventi "sceglie la risposta giusta" e "conosce la risposta giusta". Per calcolare

$$P(K|C) = \frac{P(K \cap C)}{P(C)}$$

Notiamo subito che

$$P(K \cap C) = P(C|K)P(K) = 1 \times p = p$$

Per trovare $P(C)$, condizioniamo al fatto che sapesse la risposta o meno.

$$\begin{aligned} P(C) &= P(C|K)P(K) + P(C|K^c)P(K^c) \\ &= p + (1/m)(1-p) \end{aligned}$$

Quindi la quantità richiesta è

$$P(K|C) = \frac{p}{p + (1/m)(1-p)} = \frac{mp}{1 + (m-1)p}$$

Così ad esempio, se $p = 1/2$ e $m = 5$, la probabilità che lo studente conoscesse la risposta, considerato il fatto che ha risposto correttamente è pari a $5/6$. □

Esempio 3.7.4. Una particolare analisi del sangue è efficace al 99% nell'individuare una certa malattia quando essa è presente. Si possono però anche verificare dei "falsi positivi" con probabilità dell'1% (ovvero una persona sana che si sottoponga al test, con una probabilità di 0.01 risulta erroneamente affetta dalla malattia in questione). Se l'incidinza di questo male sulla popolazione è dello 0.5%, qual è la probabilità che un soggeto sia malato, condizionata al fatto che le analisi abbiano dato esito positivo?

Sia M l'evento "il soggetto è malato" ed E l'evento "il risultato dell'analisi è positivo". Allora $P(M|E)$ si trova trmite

$$\begin{aligned} P(M|E) &= \frac{P(M \cap E)}{P(E)} \\ &= \frac{P(E|M)P(M)}{P(E|M)P(M) + P(E|M^c)P(M^c)} \\ &= \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995} \approx 0.3322 \end{aligned}$$

Perciò solo il 33% delle persone che risultano positive alle analisi sono realmente affette dalla malattia. Siccome molti studenti si stupiscono di questo risultato (infatti le caratteristiche del test sembrano buone e ci si aspetterebbe un valore più elevato), vale forse la pena di presentare una seconda argomentazione che anche se meno rigorosa può aiutare a chiarirsi le idee.

Se lo $0.5\% = 1/200$ della popolazione soffre di questo male, in media su 200 persone vi sarà un solo malato. Se egli si sottopone alle analisi, verrà trovato positivo quasi certamente (con probabilità 0.99), così che su 200 individui testati ve ne saranno in media 0.99 che saranno correttamente individuati come malati. D'altro canto le (in media) 199 persone sane hanno una probabilità di 0.01 di risultare positive, e quindi in media su 200 analisi vi saranno $199 \times 0.01 = 1.99$ falsi positivi. Se consideriamo che ogni 0.99 positivi veri vi sono in media 1.99 positivi falsi, ricaviamo nuovamente che la frazione di malati reali tra i soggetti positivi alle analisi è di

$$\frac{0.99}{0.99 + 1.99} \approx 0.3322 \quad \square$$

L'Equazione (3.7.1) è utile anche quando si voglia riconsiderare il proprio (personale) convincimento o livello di confidenza su un fatto, alla luce di nuove informazioni. Si vedano i prossimi esempi.

Esempio 3.7.5. Ad un certo stadio delle indagini su un crimine, l'investigatore capo è convinto al 60% della colpevolezza di un certo sospetto. Supponiamo che si scopra un nuovo indizio che mostra che il colpevole deve possedere una certa caratteristica distintiva (come ad esempio essere mancino, calvo, o avere i capelli castani); inoltre anche il sospettato la possiede. Se tale particolarità interessa il 20% della popolazione, quanto sicuro deve essere l'investigatore della colpevolezza del sospettato?

Denotiamo con G e C i due eventi "il sospetto è colpevole" e "il sospetto possiede il tratto distintivo del colpevole". Abbiamo,

$$P(G|C) = \frac{P(G \cap C)}{P(C)}$$

dove

$$P(G \cap C) = P(C|G)P(G) = 1 \times 0.6 = 0.6$$

e dove la probabilità di C si trova condizionando alla colpevolezza o meno del sospetto, nel modo seguente.

$$\begin{aligned} P(C) &= P(C|G)P(G) + P(C|G^c)P(G^c) \\ &= 1 \times 0.6 + 0.2 \times 0.4 = 0.68 \end{aligned}$$

Qui abbiamo stabilito che la probabilità che il sospetto abbia la caratteristica rilevante se non è colpevole sia quella generale della popolazione, 0.2. Concludendo,

$$P(G|C) = 0.6/0.68 \approx 0.882$$

e l'ispettore dovrebbe alzare all'88% la sua confidenza sulla colpevolezza del sospetto. □

Esempio 3.7.6 (continua). Cosa fare se l'indizio rinvenuto non è univoco? Supponiamo ad esempio che esso dica che non è certo, ma vi è il 90% di probabilità che il colpevole possieda questa caratteristica. Come si modifica la risoluzione per tenere conto di questa complicazione?

In questo caso, la probabilità che il sospetto possegga la caratteristica rilevante, supponendo che sia colpevole è di 0.9, mentre prima era pari a 1. Allora,

$$\begin{aligned} P(G|C) &= \frac{P(G \cap C)}{P(C)} \\ &= \frac{P(C|G)P(G)}{P(C|G)P(G) + P(C|G^c)P(G^c)} \\ &= \frac{0.9 \times 0.6}{0.9 \times 0.6 + 0.2 \times 0.4} = \frac{0.54}{0.62} \approx 0.871 \end{aligned}$$

che è un valore un po' inferiore a quello ottenuto precedentemente (perché?). □

L'Equazione (3.7.1) può essere generalizzata nel modo seguente. Siano assegnati una quantità finita (o numerabile) di eventi mutuamente esclusivi $F_1, F_2, \ldots, F_n$ tali che

$$\bigcup_{i=1}^{n} F_i = S$$

Questa proprietà si cita dicendo che gli eventi F_i *ricoprono* S e significa che si verifica sempre almeno uno di essi (esattamente uno, se – come nel nostro caso – sono anche disgiunti). Consideriamo un ulteriore evento E, che riscriviamo come

$$E = \bigcup_{i=1}^{n} (E \cap F_i)$$

notando che anche gli eventi $E \cap F_i$, per $i = 1, 2, \ldots, n$ sono mutuamente esclusivi. Si ottiene dall'Assioma 3 che

$$\begin{aligned} P(E) &= \sum_{i=1}^{n} P(E \cap F_i) \\ &= \sum_{i=1}^{n} P(E|F_i)P(F_i) \end{aligned} \tag{3.7.2}$$

Questa formula – detta *formula di fattorizzazione* o di *disintegrazione* – mostra che è possibile calcolare la probabilità di un evento E condizionando rispetto a quale si verifichi tra un gruppo di eventi accessori mutuamente esclusivi e che ricoprono S. Di nuovo $P(E)$ può essere vista come la media pesata delle probabilità condizionate $P(E|F_i)$, usando come pesi le corrispondenti $P(F_i)$.

Si immagini ora di disporre dell'ulteriore informazione che si sia effettivamente verificato l'evento E. Che probabilità avranno gli eventi F_j tenendone conto?

$$\begin{aligned} P(F_j|E) &= \frac{P(F_j \cap E)}{P(E)} \\ &= \frac{P(E|F_j)P(F_j)}{\sum_{i=1}^{n} P(E|F_i)P(F_i)} \end{aligned} \tag{3.7.3}$$

L'Equazione (3.7.3) prende il nome di *formula di Bayes*, in onore del filosofo inglese Thomas Bayes. Se pensiamo agli eventi F_j come a possibili "ipotesi" alternative che abbiano influenza su un qualche esperimento, si può immaginare che la formula di Bayes ci mostri come è necessario modificare le opinioni su tali ipotesi da prima a dopo l'esperimento stesso, con le loro probabilità che passano da $P(F_j)$ a $P(F_j|E)$.

Esempio 3.7.7. Un aereo è scomparso, e si suppone che possa essere caduto in una qualsiasi di tre regioni, con uguale probabilità. Per $i = 1, 2, 3$, sia $1 - \alpha_i$ la probabilità di rintracciare un velivolo che cada nella regione i-esima. (Le costanti α_i rappresentano la probabilità di non rinvenire il velivolo; sono normalmente dovute alle condizioni geografiche e ambientali delle regioni.) Qual è la probabilità che l'aereo si trovi in ciascuna delle tre regioni se una ricerca della regione 1 ha dato esito negativo?

Per $i = 1, 2, 3$, denotiamo con R_i l'evento "il velivolo si trova nella regione i-esima"; sia E l'evento "la ricerca nella regione 1 non ha successo". Dalla formula di Bayes otteniamo per R_1

$$\begin{aligned} P(R_1|E) &= \frac{P(E|R_1)P(R_1)}{\sum_{i=1}^{3} P(E|R_i)P(R_i)} \\ &= \frac{\alpha_1/3}{\alpha_1/3 + 1/3 + 1/3} = \frac{\alpha_1}{\alpha_1 + 2} \end{aligned}$$

mentre per $j = 2, 3$,

$$\begin{aligned} P(R_j|E) &= \frac{P(E|R_j)P(R_j)}{\sum_{i=1}^{3} P(E|R_i)P(R_i)} \\ &= \frac{1/3}{\alpha_1/3 + 1/3 + 1/3} = \frac{1}{\alpha_1 + 2} \end{aligned}$$

Quindi se ad esempio fosse $\alpha_1 = 0.4$, la probabilità che il velivolo sia nella prima regione nonostante cercandolo lì non sia stato trovato sarebbe di $1/6$. □

3.8 Eventi indipendenti

Gli esempi dati nella sezione precedente illustrano bene il fatto che $P(E|F)$, la probabilità di E condizionata ad F, è generalmente diversa dalla probabilità non condizionata, $P(E)$. Insomma, sapere che l'evento F si è verificato, modifica di solito la probabilità che si sia verificato E. Nel caso particolare in cui invece $P(E|F)$ e $P(E)$ siano uguali, diciamo che E è indipendente da F. Quindi E è indipendente da F se la conoscenza che F si è avverato non cambia la probabilità di E.

Siccome $P(E|F) = P(E \cap F)/P(F)$, si vede che E è indipendente da F se

$$P(E \cap F) = P(E)P(F) \tag{3.8.1}$$

Poiché questa equazione è simmetrica in E e F, quando E è indipendente da F, è anche vero che F è indipendente da E. Si dà allora la seguente definizione.

Definizione 3.8.1. Due eventi E e F si dicono *indipendenti* se vale l'Equazione (3.8.1), altrimenti si dicono *dipendenti*.

Esempio 3.8.1. Si pesca una carta a caso da un mazzo da 52 carte da gioco. Se A è l'evento che la carta sia un asso e C l'evento che il seme sia cuori, allora A e C sono indipendenti, infatti $P(A \cap C) = 1/52$, mentre $P(A) = 4/52$ e $P(C) = 13/52$:

$$\frac{4}{52} \cdot \frac{13}{52} = \frac{52}{52^2} = \frac{1}{52} \quad \square$$

Esempio 3.8.2. Se denotiamo con E l'evento che la prossima presidenza statunitense sia repubblicana e con F l'evento che ci sarà un terremoto eccezionale nel prossimo anno, pare del tutto convincente che E e F siano indipendenti. Si noti però come sarebbe invece fonte di controversie la decisione se E sia dipendente o indipendente da G, dove G è l'evento che nei prossimi due anni vi sia un periodo di recessione. $\square$

Diamo ora un utile risultato sull'indipendenza di eventi.

Proposizione 3.8.1. Se E e F sono indipendenti, lo sono anche E e F^c.

Dimostrazione. Dobbiamo dimostrare che $P(E \cap F^c) = P(E)P(F^c)$. Siccome E è l'unione disgiunta di $E \cap F$ e $E \cap F^c$,

$$\begin{aligned} P(E \cap F^c) &= P(E) - P(E \cap F) \\ &= P(E) - P(E)P(F) \qquad \text{per l'indipendenza di } E \text{ e } F \\ &= P(E)[1 - P(F)] = P(E)P(F^c) \quad \square \end{aligned}$$

Quindi, se E e F sono indipendenti, la probabilità che E si realizzi non è modificata dall'informazione se F si sia verificato oppure no.

Se E è indipendente sia da F sia da G, possiamo concludere che E è indipendente da $F \cap G$? Sorprendentemente, la risposta è no: si veda il prossimo esempio.

Esempio 3.8.3. Si tirano due dadi non truccati. Sia E_7 l'evento "la somma dei due punteggi è pari a 7", sia F l'evento "il primo dado totalizza un 4" e sia G l'evento "il secondo dado totalizza un 3". Si può dimostrare che E_7 è indipendente da F come pure da G (si svolga il Problema 36 adesso!). Tuttavia chiaramente E_7 non è indipendente da $F \cap G$, poiché $P(E_7|F \cap G) = 1$. $\square$

Da esempi come il precedente si capisce che per estendere la definizione di indipendenza a tre eventi non basta imporre quella due a due delle $\binom{3}{2}$ coppie di eventi. Siamo allora portati alla seguente definizione.

Definizione 3.8.2. I tre eventi E, F e G si dicono indipendenti se valgono tutte e quattro le equazioni seguenti:

$$\begin{aligned} P(E \cap F \cap G) &= P(E)P(F)P(G) \\ P(E \cap F) &= P(E)P(F) \\ P(E \cap G) &= P(E)P(G) \\ P(F \cap G) &= P(F)P(G) \end{aligned} \tag{3.8.2}$$

Si noti che se tre eventi E, F e G sono indipendenti, allora ciascuno di essi è indipendente da qualunque evento si possa costruire con gli altri due. Ad esempio E risulta indipendente da $F \cup G$, infatti

$$\begin{aligned} &P[E \cap (F \cup G)] = \\ &\quad = P[(E \cap F) \cup (E \cap G)] \\ &\quad = P(E \cap F) + P(E \cap G) - P(E \cap F \cap G) && \text{per la Proposizione 3.4.2} \\ &\quad = P(E)P(F) + P(E)P(G) - P(E)P(F \cap G) && \text{per l'indipendenza} \\ &\quad = P(E)[P(F) + P(G) - P(F \cap G)] \\ &\quad = P(E)P(F \cup G) && \text{per la Proposizione 3.4.2} \end{aligned}$$

Chiaramente la definizione precedente si può estendere senza sforzo ad un numero finito arbitrario di eventi. Gli eventi $E_1, E_2, \ldots, E_n$ si dicono indipendenti se per ogni loro sottogruppo $E_{a_1}, E_{a_2}, \ldots, E_{a_r}$, con $1 \le a_1 < \cdots < a_r \le n$, vale l'equazione

$$P\left(\bigcap_{i=1}^{r} E_{a_i}\right) = \prod_{i=1}^{r} P(E_{a_i}) \tag{3.8.3}$$

Accade spesso che un esperimento casuale (in particolare quelli di interesse statistico) consista di una successione di prove, come il lancio ripetuto di una moneta. In molte di tali situazioni è ragionevole assumere che gli esiti di qualunque gruppo di queste prove non influenzino quelli delle altre. In questi casi gli eventi che dipendono dai singoli sottoesperimenti sono indipendenti, e l'intero ambito prende il nome di *schema delle prove indipendenti*.

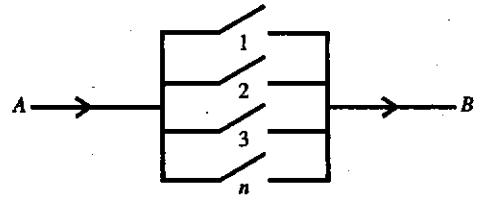


Figura 3.7 Un sistema in parallelo. Funziona se la corrente passa da A a B.

Esempio 3.8.4. Un sistema composto di n componenti distinti si dice *in parallelo* se funziona fino a che almeno uno dei componenti funziona (si veda la Figura 3.7). Sia dato un sistema di questo tipo, per il quale, per $i = 1, 2, \ldots, n$ il componente i-esimo funziona – indipendentemente da tutti gli altri – con probabilità p_i. Qual è la probabilità che l'intero sistema funzioni?

Denotiamo con A_i l'evento che il componente i funzioni. Allora

$$\begin{aligned} P(\text{il sistema funziona}) &= 1 - P(\text{il sistema non funziona}) \\ &= 1 - P(\text{nessun componente funziona}) \\ &= 1 - P(A_1^c \cap A_2^c \cap \cdots \cap A_n^c) \\ &= 1 - \prod_{i=1}^{n}(1 - p_i) \quad \square \end{aligned}$$

Problemi

1. Una scatola contiene una biglia rossa, una verde e una blu.

- **(a)** Descrivi lo spazio degli esiti dell'esperimento che consiste nell'estrarre una biglia, rimetterla nella scatola ed estrarre una seconda volta.
- **(b)** Ripeti l'esercizio senza la rimessa della prima biglia.

2. Si tira tre volte una moneta. Qual è lo spazio degli esiti di questo esperimento casuale? Scrivi esplicitamente l'evento "si ottengono più teste che croci".

3. Siano $S := \{1,2,3,4,5,6,7\}$, $E := \{1,3,5,7\}$, $F := \{7,4,6\}$, $G := \{1,4\}$. Scrivi gli elementi dei seguenti eventi.

(a) $E \cap F$; **(b)** $E \cap G^c$; **(c)** $E^c \cap (F \cup G)$;
(d) $E \cup (F \cap G)$; **(e)** $(E \cap F^c) \cup G$; **(f)** $(E \cap G) \cup (F \cap G)$.

4. Si tirano due dadi. Sia E l'evento che la somma dei punteggi sia pari, F che il primo dado realizzi un 1, e G che la somma sia 5. Si descrivano gli eventi

(a) $E \cap F$; **(b)** $E \cup F$; **(c)** $F \cap G$; **(d)** $E \cap F^c$; **(e)** $E \cap F \cap G$.

5. Un sistema è composto da 4 componenti, ciascuno dei quali funziona oppure è guasto. Si osserva lo stato dei componenti, ottenendo un vettore (x_1, x_2, x_3, x_4), dove x_i è 1 oppure 0 a seconda che il componente i-esimo funzioni oppure no.

- **(a)** Da quanti elementi è formato lo spazio degli esiti?
- **(b)** Il sistema nel suo insieme funziona fintantoché entrambi i componenti 1 e 2 oppure quelli 3 e 4 funzionano. Specifica tutti gli esiti dell'evento "il sistema funziona".
- **(c)** Sia E l'evento "i componenti 1 e 3 sono guasti". Quanti esiti contiene?

6. Siano E, F e G tre eventi qualsiasi. Trova le espressioni algebriche, in termini di intersezioni, unioni e complementazione, per gli eventi costituiti dal fatto che, tra E, F e G, si verifichino

- **(a)** soltanto E;
- **(b)** sia E sia G, ma non F;
- **(c)** almeno uno dei tre;
- **(d)** almeno due dei tre;
- **(e)** tutti e tre;
- **(f)** nessuno;
- **(g)** non più di un evento;
- **(h)** non più di due eventi;
- **(i)** esattamente due eventi;
- **(j)** non più di tre eventi.

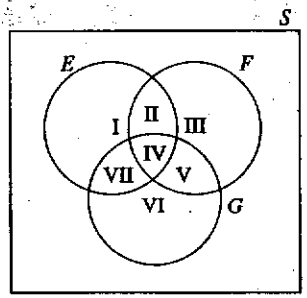
7. Semplifica, dove possibile, le espressioni che seguono.

- **(a)** $E \cup E^c$;
- **(b)** $E \cap E^c$;
- **(c)** $(E \cup F) \cap (E \cup F^c)$;
- **(d)** $(E \cup F) \cap (E^c \cup F) \cap (E \cup F^c)$;
- **(e)** $(E \cup F) \cap (F \cup G)$.

8. Usa i diagrammi di Venn (o un metodo a piacere) per mostrare che

- **(a)** $E \cap F \subset E, E \subset E \cup F$;
- **(b)** se $E \subset F$, allora $F^c \subset E^c$;
- **(c)** le due proprietà commutative (3.3.1) di pagina 62 sono valide;
- **(d)** le due proprietà associative (3.3.2) di pagina 62 sono valide;
- **(e)** $F = (F \cap E) \cup (F \cap E^c)$;
- **(f)** $E \cup F = E \cup (E^c \cap F)$;
- **(g)** le leggi di De Morgan (3.3.4) di pagina 63 sono valide.

9. Studia la figura seguente e descrivi gli eventi denominati con i numeri romani da I a VI, in termini dei tre eventi E, F e G.



10. Dimostra che se $E \subset F$, allora $P(E) \leq P(F)$. (*Suggerimento:* Scrivi F come unione disgiunta di E e un altro evento.)

11. Dimostra la *proprietà subadditiva* di P, ovvero che se $E_1, E_2, \ldots, E_n$ sono eventi qualsiasi,

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} P(E_i)$$

12. Dimostra che se $P(E) = 0.9$ e $P(F) = 0.9$, allora $P(E \cap F) \geq 0.8$. Poi dimostra che in generale vale la disuguaglianza seguente

$$P(E \cap F) \geq P(E) + P(F) - 1$$

13. Dimostra la due equazioni seguenti

(a) $P(E \cap F^c) = P(E) - P(E \cap F)$;

(b) $P(E^c \cap F^c) = 1 - P(E) - P(F) + P(E \cap F)$.

14. Dimostra che la probabilità che si realizzi uno e uno solo degli eventi E e F è pari a $P(E) + P(F) - 2P(E \cap F)$.

15. Calcola i coefficienti binomiali $\binom{9}{3}$, $\binom{9}{6}$, $\binom{7}{2}$, $\binom{7}{5}$ e $\binom{10}{7}$.

16. Dimostra che, per ogni scelta di $0 \leq r \leq n$,

$$\binom{n}{r} = \binom{n}{n-r}$$

Poi trova un argomento combinatorio che illustri la stessa equazione spiegando in che senso scegliere r elementi da un insieme di n è equivalente a scegliere $n - r$ elementi dallo stesso insieme.

17. Dimostra che

$$\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r}$$

Per trovare una spiegazione combinatoria, considera un insieme di n elementi, di cui uno fissato: quanti sono i diversi sottoinsiemi di r che contengono l'elemento fissato? E quanti quelli che non lo contengono?

18. Un gruppo di 5 bambini e 10 bambine è in fila in ordine casuale, nel senso che tutte le 15! possibili permutazioni si suppongono equiprobabili.

(a) Qual è la probabilità che il quarto della fila sia un bambino?

(b) E il dodicesimo?

(c) Qual è la probabilità che un determinato bambino occupi la terza posizione?

19. In un comune vi sono 5 alberghi. Se 3 persone devono scegliere un albergo in cui pernottare, qual è la probabilità che finiscano tutte in alberghi differenti? Che cosa stiamo assumendo senza dirlo esplicitamente?

20. In un paese vi sono 4 tecnici che riparano televisori. Se si guastano 4 TV, qual è la probabilità che vengano chiamati esattamente 2 tecnici? Che cosa stiamo assumendo senza dirlo esplicitamente?

21. Una donna ha un mazzo con n chiavi, una delle quali apre la sua porta. Se le prova a caso scartando quelle che non aprono, qual è la probabilità che trovi la chiave giusta al k-esimo tentativo? E se non scartasse le chiavi già provate?

22. Una scarpiera contiene 8 paia di scarpe. Se si prendono a caso 4 calzature, qual è la probabilità **(a)** di non formare nessun paio di scarpe uguali; **(b)** di formarne esattamente uno?

23. Il re non è figlio unico: ha un fratello oppure una sorella. Qual è la probabilità che si tratti di una sorella?

24. Una coppia ha due figli. Qual è la probabilità che si tratti di due maschi, se il primogenito è un maschio?

25. Tra gli studenti di un college americano, le femmine sono il 52%, quelli che studiano informatica sono il 5%, le femmine che studiano informatica sono il 2%. Se si sceglie a caso uno studente, quali sono le probabilità condizionate che:

(a) sia una femmina, sapendo che studia informatica;

(b) studi informatica, sapendo che è una femmina?

26. Intervistando un totale di 500 coppie di coniugi, entrambi lavoratori, si sono ottenuti i seguenti dati sui loro redditi annuali.

	Marito	
Moglie	Meno di \$ 25 000	Più di \$ 25 000
Meno di \$ 25 000	212	198
Più di \$ 25 000	36	54

Se si sceglie a caso una di queste coppie, qual è

(a) la probabilità che il marito guadagni meno di \$ 25 000;

(b) la probabilità condizionata che la moglie guadagni più di $ 25 000, se questo è vero per il marito;

(c) la probabilità condizionata che la moglie guadagni più di $ 25 000, se il marito quadagna meno di quella cifra?

27. In una certa regione vi sono due ditte che producono apparecchi radiofonici. Quelle della fabbrica A sono difettose con probabilità 0.05, mentre quelle della fabbrica B, con probabilità 0.01. Supponi di avere acquistato due radio prodotte dalla stessa ditta, che può essere la A o la B con probabilità del 50%. Se la prima delle due radio è difettosa, qual è la probabilità condizionata che sia difettosa anche la seconda?

28. Dimostra che

$$\frac{P(H|E)}{P(G|E)} = \frac{P(E|H)}{P(E|G)} \frac{P(H)}{P(G)}$$

Supponi che prima di ottenere una nuova informazione l'ipotesi H fosse tre volte più probabile della G. Se l'informazione aggiuntiva è due volte più probabile quando è vera G rispetto a quando è vera H, qual è l'ipotesi più credibile tenendo conto della nuova informazione?

29. Hai chiesto ad un vicino di innaffiare una piantina delicata mentre sei in vacanza. Pensi che senza acqua la piantina muoia con probabilità 0.8, mentre se innaffiata questa probabilità si ridurrebbe a 0.15. La tua fiducia che il vicino si ricordi di innaffiarla è del 90%.

(a) Qual è la probabilità che la pianta sia ancora viva al tuo ritorno?

(b) Se fosse morta, quale sarebbe la probabilità che il vicino si sia dimenticato di innaffiarla?

30. In un'urna vengono inserite due palline, ciascuna delle quali può essere rossa o blu con la stessa probabilità. Si estrae a caso una pallina che viene reinserita, quindi si estrae di nuovo a caso una pallina: se entrambe le estratte sono risultate rosse, con che probabilità

(a) entrambe le palline nell'urna erano rosse?

(b) estraendo nuovamente una delle due palline si trova una rossa?

31. Su 1 000 membri di una associazione di pensionati americani, 600 si dichiarano repubblicani, mentre gli altri democratici. In occasione di una elezione interna in cui hanno votato tutti, 60 repubblicani hanno dato la loro preferenza al candidato democratico e 50 democratici hanno votato il candidato repubblicano. Se un membro dell'associazione scelto a caso ha votato il repubblicano, con che probabilità si tratta di un democratico?

32. Due palline vengono tinte con vernice nera o dorata, ciascuna con probabilità $1/2$ e indipendentemente l'una dall'altra. Esse vengono poi inserite in un'urna.

(a) Supponi di sapere per certo che la vernice dorata sia stata usata (e quindi vi è almeno una pallina di questo colore). Calcola la probabilità condizionata che entrambe le palline siano dorate.

(b) Supponi adesso che l'urna venga scossa violentemente, e ne esca una pallina dorata. Qual è la probabilità condizionata che anche l'altra pallina lo sia?

(c) Spiega come mai nei due punti precedenti hai ottenuto lo stesso numero/un numero diverso.

33. Due cassettiere esternamente identiche dispongono di due cassetti ciascuna; quelli della prima contengono una moneta d'argento ciascuno, mentre i due cassetti della seconda contengono l'uno, una moneta d'argento e l'altro, una moneta d'oro. Si seleziona una cassettiera a caso, quindi si sceglie a caso uno dei suoi cassetti, e apertolo vi si trova una moneta d'argento. Qual è la probabilità che anche nell'altro cassetto vi sia una moneta d'argento?

34. Supponi che vi sia un test per diagnosticare un certo tipo di tumore con affidabilità che è pari al 95% sia per le persone malate, sia per quelle sane. Se lo 0.4% della popolazione soffre di questa forma di tumore, calcola la probabilità che un soggetto che è risultato positivo al test sia realmente malato.

35. Una compagnia di assicurazioni classifica i suoi clienti in tre fasce – basso rischio, medio rischio e alto rischio. Le sue statistiche indicano che le probabilità che un cliente delle tre fasce abbia un incidente entro un periodo di un anno sono rispettivamente 0.05, 0.15 e 0.30. Se il 20% dei clienti sono a basso rischio, il 50% a medio rischio e il 30% ad alto rischio, che percentuale dei clienti avrà mediamente incidenti in un lasso di un anno? Se un cliente non ha avuto incidenti nel 1987, qual è la probabilità che appartenga a ciascuna delle tre fasce?

36. Si tirano due dadi non truccati. Sia E l'evento "la somma dei punteggi realizzati è 7". Dimostra che E è indipendente sia dall'evento che il primo dado realizzi un 4, sia dall'evento che il secondo dado realizzi un 3.

37. Le probabilità di chiusura dei cinque relè in ciascuna delle tre figure della pagina seguente sono p_1, p_2, p_3, p_4 e p_5. Tutti i relè sono indipendenti. Quali sono le probabilità che passi corrente tra gli estremi A e B dei tre circuiti?

38. In ingegneria un sistema composto da n componenti si dice "sistema k-su-n" se funziona quando almeno k dei suoi n componenti sono efficienti. Supponi che tutti i componenti funzionino indipendentemente l'uno dall'altro, e che l'i-esimo componente funzioni con probabilità p_i, per $i = 1, 2, \ldots, n$.

(a) Qual è la probabilità che un sistema 2-su-4 funzioni?

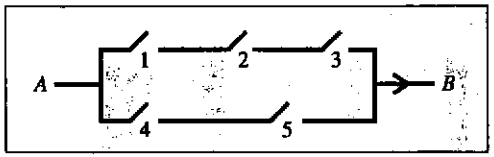
(b) E per un sistema 3-su-5?

39. Si tira cinque volte una moneta non truccata. Trova le probabilità degli eventi seguenti.

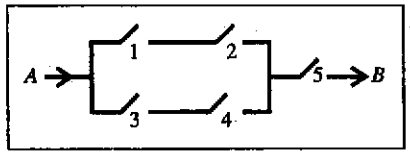
(a) I primi tre risultati sono uguali.

(b) I primi tre o gli ultimi tre risultati sono uguali.

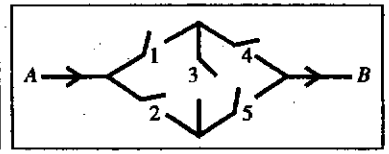
(c) Vi sono almeno due teste nei primi tre lanci e almeno due croci negli ultimi tre lanci.



(a)



(b)



(c)

40. Si ripete n volte in maniera indipendente un esperimento che può dare esiti 0, 1 o 2 con probabilità 0.3, 0.5 e 0.2 rispettivamente. Calcola la probabilità che vi sia almeno un 1 e almeno un 2 nella serie di n ripetizioni. (*Suggerimento:* Considera la probabilità dell'evento complementare.)

41. Un sistema di n componenti in parallelo funziona se non tutti i suoi componenti sono guasti. Considera un sistema di questo tipo in cui il funzionamento di ogni componente è indipendente da tutti gli altri, e ciascuno funziona con probabilità $1/2$. Qual è la probabilità che il primo componente non sia guasto condizionata al funzionamento dell'intero sistema?

42. Prendiamo in considerazione 5 differenti geni[2] di un dato organismo (li denotiamo con le prime cinque lettere dell'alfabeto). Ogni gene appare in due forme (che denotiamo con lettere maiuscole e minuscole), e ogni esemplare possiede un paio di ciascuno dei 5 geni, che possono esser uguali o diversi (per il primo gene quindi le alternative sono aa, aA e AA). Assumiamo la convenzione che la forma maiuscola sia quella dominante, mentre la minuscola sia recessiva. Ciò significa che se un organismo possiede la coppia xX esprimerà le caratteristiche del gene X. Ad esempio, se X è il gene degli occhi castani e x quello degli occhi azzurri, gli esemplari con le coppie xX e XX avranno gli occhi castani, e solo quelli con la coppia xx avranno gli occhi azzurri. Le manifestazioni fisiche dei caratteri genetici di un organismo costituiscono il suo fenotipo, mentre il suo patrimonio genetico costituisce il genotipo. (Quindi due organismi con le coppie di geni aA, bB, cc, dD, ee e AA, BB, cc, DD, ee hanno genotipi diversi ma lo stesso fenotipo.) Quando si incrociano due organismi, ciascuno contribuisce con uno a caso dei due geni di ciascuna delle sue cinque coppie, in maniera indipendente tra loro e con l'altro genitore. Se si incrociano due organismi con le coppie di geni aA, bB, cC, dD, eE e aa, bB, cc, dD, ee, qual è la probabilità che la progenie corrisponda (limitatamente a questi cinque geni) (1) nel fenotipo, e (2) nel genotipo,

(a) al primo genitore

[2] Si veda ad esempio: Peter J. Russell *Genetica*, seconda edizione, EdiSES 1996.

(b) al secondo genitore;

(c) a uno dei genitori;

(d) a nessuno dei genitori?

43. Tre prigionieri condannati a morte vengono informati da un secondino che due di loro, scelti a caso, saranno graziati. Uno di essi gli chiede allora di essere informato privatamente almeno su quale dei suoi due compagni verrà graziato, sostenendo che non vi sia alcun male nel divulgare questa informazione, poiché è chiaro a tutti che comunque almeno uno dei due sarà graziato. Il secondino però si rifiuta di dare risposta, perché in tal modo la probabilità di essere giustiziato del prigioniero curioso salirebbe da $1/3$ a $1/2$, restando solo due prigionieri dal destino celato. Cosa pensi del ragionamento del secondino?

44. Anche se i miei genitori hanno entrambi gli occhi castani, io ho gli occhi azzurri. Qual è la probabilità che anche mia sorella abbia gli occhi azzurri (si veda il Problema 42)?

45. Quante persone è necessario riunire affinché sia almeno del 50% la probabilità che qualcuno sia nato un 29 di Febbraio? Quali assunzioni hai fatto per dare questa risposta?

4 Variabili aleatorie e valore atteso

Contenuto

4.1 Variabili aleatorie

Quando si realizza un esperimento casuale, non sempre si è interessati in ugual modo a tutte le informazioni ricavabili dal suo esito. Spesso si può individuare una singola quantità numerica (ricavabile dall'esito stesso) che racchiude tutto ciò che in realtà vogliamo sapere. Se tiriamo due dadi, ad esempio, può accadere che ci interessi solamente il valore della loro somma, e non ciascuno dei punteggi. Potremmo volere registrare che il totale realizzato è 7, senza dare importanza a quale sia l'esito vero e proprio dell'esperimento, tra i sei possibili, che sono $(1,6)$, $(2,5)$, $(3,4)$, $(4,3)$, $(5,2)$ e $(6,1)$. Un ingegnere civile che segue il livello di un bacino idrico, allo stesso modo, potrebbe decidere di prendere delle misurazioni solo alla fine di ogni stagione delle piogge, perché magari le oscillazioni giornaliere non aggiungono informazioni rilevanti.

Quantità di interesse che, come queste, sono determinate dal risultato di un esperimento casuale sono dette *variabili aleatorie*. Siccome il valore di una variabile aleatoria è determinato dall'esito dell'esperimento, possiamo assegnare delle probabilità ai suoi valori possibili.

Esempio 4.1.1. Si tirano due dadi indipendenti e non truccati, e si denota con la lettera X la variabile aleatoria definita dalla loro somma. Ha senso domandarsi quanto vale la probabilità che $X = 3$, ovvero la probabilità dell'evento $\{s \in S : X(s) = 3\}$. Vi sono due elementi dello spazio degli esiti di questo esperimento che danno ad X il valore 3. Essi sono $(1,2)$ e $(2,1)$. Perciò, con una notazione più leggera,

$$\{X = 3\} = \{s \in S : X(s) = 3\} = \{(1,2),(2,1)\} \tag{4.1.1}$$

e di conseguenza la probabilità che $X = 3$ è pari a $2/36$ perché abbiamo a che fare con esiti equiprobabili. Il modo corretto di scrivere questo risultato sarebbe, $P(\{X = 3\}) = 2/36$, ma è invalso l'uso di scrivere, con leggero abuso di notazione $P(X = 3) = 2/36$. Ricorrendo a questa convenzione elenchiamo le probabilità per tutti i valori possibili di X.

$$\begin{aligned}
P(X = 2) &= P\{(1,1)\} = \tfrac{1}{36}\\
P(X = 3) &= P\{(1,2),(2,1)\} = \tfrac{2}{36}\\
P(X = 4) &= P\{(1,3),(2,2),(3,1)\} = \tfrac{3}{36}\\
P(X = 5) &= P\{(1,4),(2,3),(3,2),(4,1)\} = \tfrac{4}{36}\\
P(X = 6) &= P\{(1,5),(2,4),(3,3),(4,2),(5,1)\} = \tfrac{5}{36}\\
P(X = 7) &= P\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\} = \tfrac{6}{36}\\
P(X = 8) &= P\{(2,6),(3,5),(4,4),(5,3),(6,2)\} = \tfrac{5}{36}\\
P(X = 9) &= P\{(3,6),(4,5),(5,4),(6,3)\} = \tfrac{4}{36}\\
P(X = 10) &= P\{(4,6),(5,5),(6,4)\} = \tfrac{3}{36}\\
P(X = 11) &= P\{(5,6),(6,5)\} = \tfrac{2}{36}\\
P(X = 12) &= P\{(6,6)\} = \tfrac{1}{36}
\end{aligned} \tag{4.1.2}$$

La variabile aleatoria X può assumere tutti i valori interi che vanno da 2 a 12, con probabilità specificate dalle Equazioni (4.1.2). Siccome X deve assumere uno di questi valori, ne segue che $S = \bigcup_{i=2}^{12}\{X = i\}$, e di conseguenza

$$1 = P(S) = P\left(\bigcup_{i=2}^{12}\{X = i\}\right) = \sum_{i=2}^{12} P(X = i)$$

come si verifica facilmente dalle (4.1.2).

Un'altra variabile aleatoria di possibile interesse all'interno di questo esperimento è il valore del primo dado. La denotiamo con Y e notiamo che

$$P(Y = i) = 1/6, \qquad i = 1,2,3,4,5,6$$

Ovvero Y può assumere ciascuno dei valori interi da 1 a 6 con la stessa probabilità. □

Esempio 4.1.2. Un tizio acquista due componenti elettronici, ciascuno dei quali può essere *a*ccettabile o *d*ifettoso. Supponiamo che le probabilità dei 4 esiti possibili – $(d,d),(d,a),(a,d),(a,a)$ – siano rispettivamente 0.09, 0.21, 0.21 e 0.49. Sia X il numero di componenti accettabili; allora X è una variabile aleatoria che può assumere i valori 0, 1 o 2 con probabilità

$$\begin{aligned}
P(X = 0) &= 0.09\\
P(X = 1) &= 0.42\\
P(X = 2) &= 0.49
\end{aligned}$$

Se vogliamo limitarci a registrare se vi sia almeno un componente accettabile, possiamo definire una variabile aleatoria I come segue,

$$I := \begin{cases} 1 & \text{se } X = 1 \text{ o } 2\\ 0 & \text{se } X = 0 \end{cases}$$

Se con A si denota l'evento che vi sia almeno un componente accettabile, allora I è detta la *funzione indicatrice* dell'evento A, infatti I assume i valori 1 o 0 a seconda se l'evento A si verifica o meno. Le probabilità corrispondenti ai valori possibili di I sono

$$\begin{aligned}
P(I = 1) &= 0.91\\
P(I = 0) &= 0.09 \quad \square
\end{aligned}$$

Negli esempi precedenti tutte le variabili aleatorie disponevano di un insieme finito di valori possibili. Variabili aleatorie con un numero finito o numerabile di valori possibili sono dette *discrete*. Esistono comunque anche variabili aleatorie dette appunto *continue*, che possono assumere un insieme continuo di valori possibili, come può essere un intervallo di numeri reali. Un esempio è il tempo di vita di una automobile, che può assumere qualunque valore di un qualche intervallo (a,b).

Definizione 4.1.1. La *funzione di ripartizione* F di una variabile aleatoria X, è definita, per ogni numero reale x, tramite

$$F(x) := P(X \le x) \tag{4.1.3}$$

Quindi $F(x)$ esprime la probabilità che la variabile aleatoria X assuma un valore minore o uguale a x. Useremo la notazione $X \sim F$ per indicare che F è la funzione di ripartizione di X.

Tutte le questioni di probabilità che si possano sollevare su una variabile aleatoria, ammettono una risposta in termini della sua funzione di ripartizione. Ad esempio,

volendo calcolare $P(a < X \leq b)$, basta notare che $\{X \leq b\}$ è l'unione dei due eventi disgiunti $\{X \leq a\}$ e $\{a < X \leq b\}$. Quindi per l'Assioma 3,

$$P(X \leq b) = P(X \leq a) + P(a < X \leq b)$$

da cui

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) \tag{4.1.4}$$

Esempio 4.1.3. Sia assegnata una variabile aleatoria X con funzione di ripartizione

$$F(x) = \begin{cases} 0 & x \leq 0 \\ 1 - \exp(-x^2) & x > 0 \end{cases}$$

Qual è la probabilità che X sia maggiore di 1? Si procede come segue:

$$\begin{aligned} P(X > 1) &= 1 - P(X \leq 1) \\ &= 1 - F(1) = e^{-1} \approx 0.368 \quad \square \end{aligned}$$

4.2 Variabili aleatorie discrete e continue

Come è già stato detto, si dice discreta una variabile aleatoria che può assumere una quantità finita o numerabile di valori.

Definizione 4.2.1. Se X è una variabile aleatoria discreta, la sua *funzione di massa di probabilità* o *funzione di massa* si definisce nel modo seguente,

$$p(a) := P(X = a) \tag{4.2.1}$$

La funzione $p(a)$ è non nulla su un insieme al più numerabile di valori. Infatti se $x_1, x_2, \ldots$ sono i valori possibili di X, allora

$$\begin{aligned} p(x_i) &> 0, \qquad i = 1, 2, \ldots \\ p(x) &= 0, \qquad \text{tutti gli altri valori di } x \end{aligned}$$

Siccome X deve assumere uno dei valori $x_1, x_2, \ldots$, necessariamente la funzione di massa di probabilità deve soddisfare la seguente equazione:

$$\sum_{i=1}^{\infty} p(x_i) = 1 \tag{4.2.2}$$

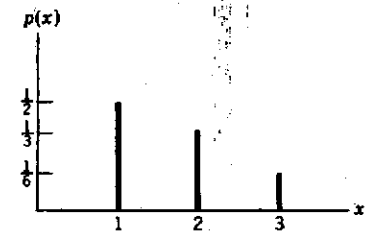


Figura 4.1 Grafico di $p(x)$ per la variabile aleatoria descritta dall'Esempio 4.2.1.

Esempio 4.2.1. Consideriamo una variabile aleatoria X che può assumere i valori 1, 2 o 3. Se sappiamo che

$$p(1) = \frac{1}{2} \qquad \text{e} \qquad p(2) = \frac{1}{3}$$

allora, dato che $p(1) + p(2) + p(3) = 1$, ne segue che $p(3) = 1/6$. La Figura 4.1 mostra il grafico di questa funzione di massa. $\square$

Per una variabile aleatoria discreta, la funzione di ripartizione F può essere espressa in funzione della funzione di massa di probabilità p, tramite

$$F(a) = \sum_{x \leq a} p(x) \tag{4.2.3}$$

dove si intende che la serie è limitata ai soli valori possibili di X minori o uguali ad a. Si noti che la F che ne risulta è una funzione a gradini, e più precisamente, se $x_1 < x_2 < \ldots$ sono i valori possibili di X, allora F è costante su ciascuno degli intervalli $[x_{i-1}, x_i)$ e in x_i fa un salto di ampiezza $p(x_i)$, passando da

$$p(x_1) + p(x_2) + \cdots + p(x_{i-1}) \qquad \text{a} \qquad p(x_1) + p(x_2) + \cdots + p(x_{i-1}) + p(x_i)$$

Supponendo che X abbia la stessa funzione di massa di probabilità dell'Esempio 4.2.1, con

$$p(1) = \frac{1}{2}, \qquad p(2) = \frac{1}{3}, \qquad p(3) = \frac{1}{6}$$

la funzione di ripartizione F di X è data da

$$F(a) = \begin{cases} 0 & a < 1 \\ \frac{1}{2} & 1 \leq a < 2 \\ \frac{5}{6} & 2 \leq a < 3 \\ 1 & 3 \leq a \end{cases}$$

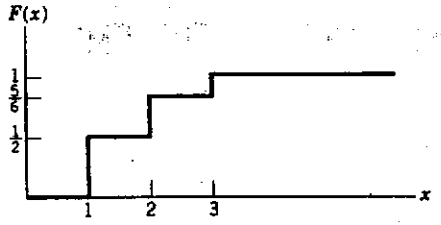


Figura 4.2 Grafico di $F(x)$.

Il grafico di tale funzione F è illustrato in Figura 4.2.

Una variabile aleatoria che possa assumere una infinità non numerabile di valori, non potrà essere discreta. Si dirà invece *continua* se[1] esite una funzione non negativa f, definita su tutto $\mathbb{R}$, avente la proprietà che per ogni insieme B di numeri reali,

$$P(X \in B) = \int_B f(x)\,dx \tag{4.2.4}$$

Definizione 4.2.2. La funzione f che compare nell'Equazione (4.2.4) è la *funzione di densità di probabilità* o più semplicemente la *densità* della variabile aleatoria X.

L'Equazione (4.2.4) dice che la probabilità che una variabile aleatoria continua X appartenga a un insieme B si può trovare integrando la sua densità su tale insieme. Poiché X deve assumere un qualche valore di $\mathbb{R}$, la sua densità deve soddisfare:

$$1 = P(X \in \mathbb{R}) = \int_{-\infty}^{\infty} f(x)\,dx \tag{4.2.5}$$

Tutte le probabilità che riguardano una variabile aleatoria continua possono essere espresse in termini di integrali della sua densità. Ad esempio, se poniamo $B = [a, b]$, ricaviamo dalla (4.2.4) che

$$P(a \le X \le b) = \int_a^b f(x)\,dx \tag{4.2.6}$$

e se in quest'ultima equazione poniamo $b = a$, troviamo che

$$P(X = a) = \int_a^a f(x)\,dx = 0$$

ovvero, la probabilità che una variabile aleatoria continua assuma un qualunque valore particolare a è nulla (si veda anche la Figura 4.3).

[1] Non sfuggirà al lettore attento che non essendo vero che tutte le variabili aleatorie che non sono discrete sono continue, questa classificazione non può essere completa. Effettivamente stiamo per semplicità omettendo di presentare anche quelle dette *miste*, che oltre a complicare notevolmente la trattazione sono piuttosto infrequenti, nella teoria come nella pratica.

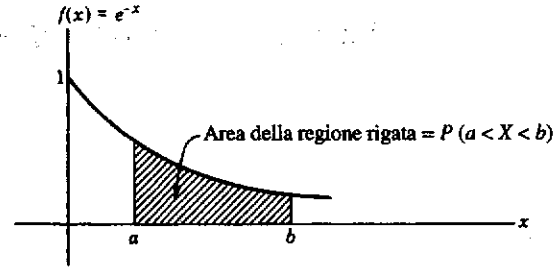


Figura 4.3 La funzione di densità di probabilità $f(x) = e^{-x}, \quad x \ge 0$.

Una relazione che lega la funzione di ripartizione F alla densità f è la seguente,

$$F(a) := P(X \in (-\infty, a]) = \int_{-\infty}^{a} f(x)\,dx \tag{4.2.7}$$

Derivando entrambi i membri si ottiene allora la relazione fondamentale:

$$\frac{d}{da}F(a) = f(a) \tag{4.2.8}$$

La densità è la derivata della funzione di ripartizione. Una interpretazione forse meno astratta della funzione di densità di probabilità si può ricavare dall'Equazione (4.2.6) nel modo che segue: se $\varepsilon > 0$ è piccolo si può approssimare l'integrale con il teorema del valore medio,

$$P\Big(a - \frac{\varepsilon}{2} \le X \le a + \frac{\varepsilon}{2}\Big) = \int_{a-\frac{\varepsilon}{2}}^{a+\frac{\varepsilon}{2}} f(x)\,dx \approx \varepsilon f(a) \tag{4.2.9}$$

Si scopre così che la probabilità che X stia in un intorno di a di ampiezza ε è approssimativamente uguale a $\varepsilon f(a)$, e quindi $f(a)$ rappresenta una indicazione di quanto è probabile che X cada "vicino" ad a (si rammenti che $\{X = a\}$ ha probabilità nulla).

Esempio 4.2.2. Sia assegnata una variabile aleatoria X con densità data da

$$f(x) = \begin{cases} C(4x - 2x^2) & 0 < x < 2 \\ 0 & \text{altrimenti} \end{cases}$$

(a) Quanto vale C? **(b)** Quanto vale $P(X > 1)$?

(a) Siccome f è una densità, deve valere l'Equazione (4.2.5), e quindi

$$\begin{aligned} 1 &= C\int_0^2 (4x - 2x^2)\,dx \\ &= C\left[2x^2 - \frac{2x^3}{3}\right]\Big|_{x=0}^{x=2} = C \cdot \frac{8}{3} \end{aligned}$$

da cui $C = 3/8$. (b) Ora che conosciamo completamente f, la probabilità di $\{X > 1\}$ si può trovare con un integrale.

$$P(X > 1) = \int_1^\infty f(x)\,dx = \frac{3}{8}\int_1^2 (4x - 2x^2)\,dx = \frac{1}{2} \quad \square$$

Osservazione 4.2.1. Quando conosciamo la funzione di massa di probabilità di una variabile aleatoria discreta, oppure la funzione di densità di probabilità di una continua, oppure ancora quando conosciamo la funzione di ripartizione di una variabile aleatoria qualsiasi, abbiamo abbastanza informazioni da poter calcolare la probabilità di ogni evento che dipenda solo da tale variabile aleatoria. Si dice in questo caso che conosciamo la *distribuzione* o *legge* della variabile aleatoria considerata. Perciò, affermare ad esempio che X e Y hanno la stessa distribuzione, vuole dire che le rispettive funzioni di ripartizione sono identiche, $X \sim F_X \equiv F_Y \sim Y$, e quindi anche che $P(X \in A) = P(Y \in A)$ per ogni insieme di valori $A \subset \mathbb{R}$.

4.3 Coppie e vettori di variabili aleatorie

Ci sono situazioni in cui la scelta (descritta all'inizio del capitolo) di ridurre un esperimento casuale allo studio di una sola variabile aleatoria, è destinata a fallire a priori, perché l'oggetto di interesse sono proprio le relazioni presenti tra due o più grandezze numeriche. Ad esempio, in un esperimento sulle possibili cause di tumore, potremmo voler indagare il rapporto tra il numero medio di sigarette fumate quotidianamente e l'età in cui viene riscontrata questa patologia. Analogamente, un ingegnere meccanico che si occupi del montaggio di un tipo laminati in acciaio, potrebbe volere conoscere la relazione tra il diametro dei punti di saldatura e la loro sollecitazione di taglio.

Per specificare la relazione tra due variabili aleatorie X e Y, il punto di partenza è estendere il concetto di funzione di ripartizione.

Definizione 4.3.1. Siano X e Y due variabili aleatorie che riguardano lo stesso esperimento casuale. Si dice *funzione di ripartizione congiunta* di X e Y – e si indica normalmente con la lettera F – la funzione di due variabili seguente.

$$F(x, y) := P(X \le x, Y \le y) \tag{4.3.1}$$

dove la virgola nell'argomento di $P()$ denota l'intersezione tra eventi.

La conoscenza di questa funzione permette, almeno in teoria, di calcolare le probabilità di tutti gli eventi che dipendono, singolarmente o congiuntamente, da X e Y. Ad esempio la funzione di ripartizione di X – che denotiamo questa volta con F_X –

può essere ottenuta dalla funzione di ripartizione congiunta F così:

$$\begin{aligned} F_X(x) &:= P(X \le x) \\ &= P(X \le x, Y < \infty) && \text{perché } Y < \infty \text{ sempre} \\ &= F(x, \infty) && \text{nel senso del limite } \lim_{y\to\infty} F(x, y) \end{aligned}$$

E analogamente la funzione di ripartizione di Y,

$$F_Y(y) = F(\infty, y)$$

4.3.1 Distribuzione congiunta per variabili aleatorie discrete

Come nel caso scalare, se sappiamo che un vettore aleatorio è di tipo discreto, possiamo definire e utilizzare la funzione di massa di probabilità.

Definizione 4.3.2. Se X e Y sono variabili aleatorie discrete che assumono i valori $x_1, x_2, \ldots$ e $y_1, y_2, \ldots$ rispettivamente, la funzione

$$p(x_i, y_j) := P(X = x_i, Y = y_j), \qquad i = 1, 2, \ldots, \quad j = 1, 2, \ldots \tag{4.3.2}$$

è la loro *funzione di massa di probabilità congiunta.*

Le funzioni di massa individuali di X e Y si possono ricavare da quella congiunta notando che, siccome Y deve assumere uno dei valori y_j, l'evento $\{X = x_i\}$ può essere visto come l'unione al variare di j degli eventi $\{X = x_i, Y = y_j\}$, che sono mutuamente esclusivi; in formule,

$$\{X = x_i\} = \bigcup_j \{X = x_i, Y = y_j\}$$

da cui, grazie all'Assioma 3,

$$\begin{aligned} p_X(x_i) &:= P(X = x_i) \\ &= P\Big(\bigcup_j \{X = x_i, Y = y_j\}\Big) \\ &= \sum_j P(X = x_i, Y = y_j) \\ &=: \sum_j p(x_i, y_j) \end{aligned} \tag{4.3.3}$$

Analogamente per p_Y,

$$p_Y(y_j) = \sum_i p(x_i, y_j) \tag{4.3.4}$$

Tabella 4.1 Funzione di massa congiunta $p(i,j) := P(X = i, Y = j)$ per le variabili aleatorie dell'Esempio 4.3.1

		j				totali righe
		0	**1**	**2**	**3**	$P(X = i)$
i	**0**	$\frac{10}{220}$	$\frac{40}{220}$	$\frac{30}{220}$	$\frac{4}{220}$	$\frac{84}{220}$
	1	$\frac{30}{220}$	$\frac{60}{220}$	$\frac{18}{220}$	0	$\frac{108}{220}$
	2	$\frac{15}{220}$	$\frac{12}{220}$	0	0	$\frac{27}{220}$
	3	$\frac{1}{220}$	0	0	0	$\frac{1}{220}$
totali colonne $P(Y = j)$		$\frac{56}{220}$	$\frac{112}{220}$	$\frac{48}{220}$	$\frac{4}{220}$	

Anche se abbiamo mostrato che le funzioni di massa individuali (un altro termine usato è *marginali*) si possono sempre ricavare da quella congiunta, il viceversa è falso. Quindi, conoscere $P(X = x_i)$ e $P(Y = y_j)$ non permette di ricavare $P(X = x_y, Y = y_j)$.

Esempio 4.3.1. Da un gruppo di 12 batterie – di cui 3 nuove, 4 usate e 5 difettose – ne vengono scelte tre a caso. Siano X e Y rispettivamente il numero di batterie nuove e usate tra quelle scelte. La funzione di massa di probabilità congiunta, $p(i,j)$, è data dai valori seguenti, come il lettore può verificare facilmente, con ragionamenti simili a quelli della Sezione 3.5.

$$p(0,0) = \frac{\binom{5}{3}}{\binom{12}{3}} = \frac{10}{220} \qquad p(0,1) = \frac{\binom{4}{1}\binom{5}{2}}{\binom{12}{3}} = \frac{40}{220}$$

$$p(0,2) = \frac{\binom{4}{2}\binom{5}{1}}{\binom{12}{3}} = \frac{30}{220} \qquad p(0,3) = \frac{\binom{4}{3}}{\binom{12}{3}} = \frac{4}{220}$$

$$p(1,0) = \frac{\binom{3}{1}\binom{5}{2}}{\binom{12}{3}} = \frac{30}{220} \qquad p(1,1) = \frac{\binom{3}{1}\binom{4}{1}\binom{5}{1}}{\binom{12}{3}} = \frac{60}{220}$$

$$p(1,2) = \frac{\binom{3}{1}\binom{4}{2}}{\binom{12}{3}} = \frac{18}{220} \qquad p(2,0) = \frac{\binom{3}{2}\binom{5}{1}}{\binom{12}{3}} = \frac{15}{220}$$

$$p(2,1) = \frac{\binom{3}{2}\binom{4}{1}}{\binom{12}{3}} = \frac{12}{220} \qquad p(3,0) = \frac{\binom{3}{3}}{\binom{12}{3}} = \frac{1}{220}$$

Queste probabilità possono essere convenientemente presentate in forma tabellare, come illustrato nella Tabella 4.1

Si può notare come le funzioni di massa di X e Y si possano ottenere facendo le somme lungo le righe e lungo le colonne, in accordo con le Equazioni (4.3.3) e (4.3.4). Il fatto che questo tipo di tabella sia piuttosto comune, e le funzioni di massa individuali vi compaiano lungo i margini, giustifica il termine già introdotto di funzioni di massa di probabilità *marginali*. Una verifica veloce che la tabella non contenga errori grossolani consiste nel controllare che le somme dei valori sulla riga e sulla colonna marginale siano pari a 1. (Perché?) □

Esempio 4.3.2. All'interno di una certa popolazione, il 15% delle coppie non ha figli, il 20% ne ha uno, il 35% ne ha due e il 30% ne ha tre. Inoltre ogni bambino, indipendentemente da tutti gli altri, può essere maschio o femmina con pari probabilità. Se si seleziona una famiglia a caso e si denotano con X e Y il numero di femmine e di maschi presenti tra i figli in tale famiglia, si ottiene la funzione di massa di probabilità mostrata in Tabella 4.2.

Le probabilità sono state ricavate come segue.

$$\begin{aligned}
P(X = 0, Y = 0) &= P(\text{nessun figlio}) = 0.15 \\
P(X = 1, Y = 0) &= P(\text{un totale di 1 figlio, femmina}) \\
&= P(1 \text{ figlio})P(1 \text{ femmina}|1 \text{ figlio}) = 0.20 \times 0.5 = 0.1 \\
P(X = 2, Y = 0) &= P(\text{un totale di 2 figli, entrambe femmine}) \\
&= P(2 \text{ figli})P(2 \text{ femmine}|2 \text{ figli}) = 0.35 \times 0.5^2 = 0.0875 \\
P(X = 3, Y = 0) &= P(\text{un totale di 3 figli, tutte femmine}) \\
&= P(3 \text{ figli})P(3 \text{ femmine}|3 \text{ figli}) = 0.30 \times 0.5^3 = 0.0375
\end{aligned}$$

Lasciamo al lettore la verifica che anche gli altri valori della Tabella 4.2 sono corretti. Si noti anche come sia possibile usare la tabella in maniera più sofisticata, scoprendo ad esempio (in che modo?) che la probabilità che vi sia almeno una bambina è pari a 0.625. □

Tabella 4.2 Funzione di massa congiunta per le variabili aleatorie X e Y dell'Esempio 4.3.2

		j				totali righe
		0	**1**	**2**	**3**	$P(X = i)$
i	**0**	0.1500	0.1000	0.0875	0.0375	0.3750
	1	0.1000	0.1750	0.1125	0	0.3875
	2	0.0875	0.1125	0	0	0.2000
	3	0.0375	0	0	0	0.0375
totali colonne $P(Y = j)$		0.3750	0.3875	0.2000	0.0375	

4.3.2 Distribuzione congiunta per variabili aleatorie continue

Due variabili aleatorie X e Y sono *congiuntamente continue* se esiste una funzione non negativa $f(x,y)$, definita per tutti gli x e y, avente la proprietà che per ogni sottoinsieme C del piano cartesiano,

$$P\big((X,Y)\in C\big) = \iint_{(x,y)\in C} f(x,y)\,dx\,dy \tag{4.3.5}$$

Definizione 4.3.3. La funzione di due variabili f, che compare nell'Equazione (4.3.5) è la *densità congiunta* delle variabili aleatorie X e Y.

Se A e B sono sottoinsiemi qualsiasi di $\mathbb{R}$, e se si denota con $C := A \times B$ il loro prodotto cartesiano su $\mathbb{R}^2$, ovvero

$$C := \{(x,y)\in\mathbb{R}^2 : x\in A, y\in B\}$$

si vede dall'Equazione (4.3.5) che la densità congiunta f soddisfa

$$P(X\in A, Y\in B) = \int_B\int_A f(x,y)\,dx\,dy \tag{4.3.6}$$

e quindi, ponendo $A=(-\infty,a]$, $B=(-\infty,b]$, si può riscrivere la funzione di ripartizione congiunta di X e Y come,

$$\begin{aligned} F(a,b) &:= P(X\le a, Y\le b)\\ &= P(X\in A, Y\in B)\\ &= \int_B\int_A f(x,y)\,dx\,dy\\ &= \int_{-\infty}^{b}\int_{-\infty}^{a} f(x,y)\,dx\,dy \end{aligned} \tag{4.3.7}$$

da cui derivando, nelle due direzioni

$$f(a,b) = \frac{\partial^2 F(a,b)}{\partial a\,\partial b} \tag{4.3.8}$$

in tutti i punti in cui le derivate parziali sono definite. Anche qui, come nel caso scalare (si veda a pagina 97), è possibile ottenere dall'Equazione (4.3.6) una formula approssimata che motiva la scelta del nome di *densità di probabilità*:

$$\begin{aligned} P(a\le X\le a+da, b\le Y\le b+db) &= \int_b^{b+db}\int_a^{a+da} f(x,y)\,dx\,dy\\ &\approx f(a,b)\,da\,db \end{aligned} \tag{4.3.9}$$

L'approssimazione finale è valida (per il teorema del valore medio) se gli incrementi da e db sono piccoli e f è continua nel punto (a,b). Se ne deduce che $f(a,b)$ è circa pari al rapporto tra la probabilità di un rettangolino attorno al punto (a,b), e l'area $da\,db$ del rettangolino stesso, è insomma una densità di probabilità nel senso comune che questo termine assume, e una indicazione di quanto è probabile che (X,Y) cada vicino ad (a,b).

Se X e Y sono congiuntamente continue, allora prese individualmente, sono variabili aleatorie continue nel senso usuale; inoltre le loro *densità marginali* si ricavano come segue. Per ogni insieme A di numeri reali,

$$\begin{aligned} \int_A f_X(x)\,dx &= P(X\in A) && \text{per la (4.2.4)}\\ &= P(X\in A, Y\in\mathbb{R})\\ &= \int_A\int_{-\infty}^{\infty} f(x,y)\,dy\,dx && \text{per la (4.3.6)} \end{aligned}$$

Da questa equazione, visto che A è un insieme arbitrario, si ricava (con teoremi generali) che deve valere per forza l'uguaglianza degli integrandi:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \tag{4.3.10}$$

Analogamente, si può ricavare la funzione di densità marginale di Y che è,

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx \tag{4.3.11}$$

Esempio 4.3.3. Siano X e Y due variabili aleatorie congiuntamente continue con densità di probabilità congiunta data da

$$f(x,y) = \begin{cases} 2e^{-x}e^{-2y} & x>0, y>0\\ 0 & \text{altrimenti}\end{cases}$$

Si calcolino **(a)** $P(X>1, Y<1)$; **(b)** $P(X<Y)$; **(c)** $P(X<a)$.

(a) Occorre integrare $f(x,y)$ nella regione in cui $x>1$ e $y<1$, ma la seconda disuguaglianza si riduce a $0<y<1$ perché $f(x,y)$ è nulla quando $y\le 0$.

$$\begin{aligned} P(X>1, Y<1) &= \int_0^1\int_1^{\infty} 2e^{-x}e^{-2y}\,dx\,dy\\ &= \int_0^1 2e^{-2y}\left(\int_1^{\infty} e^{-x}\,dx\right)dy && \text{si integra prima in una variabile...}\\ &= \int_0^1 2e^{-2y}\{-e^{-x}\}\big|_{x=1}^{\infty}\,dy \end{aligned}$$

$$= e^{-1} \int_0^1 2e^{-2y}\,dy \qquad \text{...e poi nell'altra}$$
$$= e^{-1}(1 - e^{-2})$$

(b) In questo caso la regione su cui integrare è quella dove $x < y$. Gli estremi di integrazione che corrispondono a questo dominio possono essere scelti in due modi: **(1)** o si integra internamente in dx tra gli estremi 0 e y (infatti $x > 0$ altrimenti f è nulla, mentre $x < y$ è la definizione della regione che stiamo considerando), ed esternamente in dy tra 0 e ∞ (infatti basta porre la condizione $x < y$ sull'integrale interno); **(2)** o si integra internamente in dy tra x e ∞ (per rispettare $x < y$), ed esternamente in dx tra 0 e ∞. Scegliamo la prima strada.

$$P(X < Y) = \iint_{(x,y):0<x<y} 2e^{-x}e^{-2y}\,dx\,dy \qquad \text{a questa regione...}$$
$$= \int_0^\infty \int_0^y 2e^{-x}e^{-2y}\,dx\,dy \qquad \text{...corrispondono questi estremi}$$
$$= \int_0^\infty 2e^{-2y}\left(\int_0^y e^{-x}\,dx\right) dy \qquad \text{si integra prima nella variabile i cui estremi dipendono dall'altra}$$
$$= \int_0^\infty 2e^{-2y}(1 - e^{-y})\,dy$$
$$= \int_0^\infty 2e^{-2y}\,dy - \int_0^\infty 2e^{-3y}\,dy$$
$$= 1 - \frac{2}{3} = \frac{1}{3}$$

(c) Nell'ultimo caso gli estremi di integrazione sono semplici. La variabile aleatoria Y può assumere un valore qualsiasi, quindi y si integra su tutto $\mathbb{R}$. X deve invece essere minore di a. Supponendo che sia $a > 0$, questo significa integrare in dx tra 0 e a. (Se a è minore o uguale a zero invece, $\{X < a\}$ è un evento di probabilità nulla.)

$$P(X < a) = \int_0^a e^{-x}\left(\int_0^\infty 2e^{-2y}\,dy\right) dx$$
$$= \int_0^a e^{-x}\,dx$$
$$= 1 - e^{-a} \quad \square$$

4.3.3 Variabili aleatorie indipendenti

In analogia con quanto definito a pagina 80 per gli eventi, due variabili aleatorie sono indipendenti se tutti gli eventi relativi alla prima sono indipendenti da tutti quelli relativi alla seconda.

Definizione 4.3.4. Due variabili aleatorie che riguardano lo stesso esperimento casuale si dicono *indipendenti* se, per ogni coppia di insiemi di numeri reali A e B, è soddisfatta l'equazione

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \tag{4.3.12}$$

ovvero, se per ogni scelta di A e B, gli eventi $\{X \in A\}$ e $\{Y \in B\}$ risultano indipendenti. In caso contrario X e Y si dicono *dipendenti*.

Usando gli assiomi della probabilità è possibile dimostrare che questa definizione è equivalente alla richiesta che per ogni coppia di reali a e b,

$$P(X \le a, Y \le b) = P(X \le a)P(Y \le b)$$

ovvero che la funzione di ripartizione congiunta sia il prodotto delle marginali:

$$F(a,b) = F_X(a)F_Y(b), \qquad \forall a, b \in \mathbb{R} \tag{4.3.13}$$

dove si intende che $F_X \sim X$, $F_Y \sim Y$ e F è la funzione di ripartizione congiunta di X e Y.

Se le variabili aleatorie considerate sono discrete, l'indipendenza è anche equivalente a chiedere che la funzione di massa congiunta sia il prodotto delle marginali:

$$p(x,y) = p_X(x)p_Y(y), \qquad \forall x, y \in \mathbb{R} \tag{4.3.14}$$

Tale equivalenza si prova facilmente. Per una direzione basta notare che la (4.3.12) implica la (4.3.14) non appena si pone $A = \{x\}$ e $B = \{y\}$. Per l'altra direzione è necessario dimostrare che l'Equazione (4.3.12) è soddisfatta per ogni scelta di insiemi reali A e B.

$$P(X \in A, Y \in B) = \sum_{x \in A} \sum_{y \in B} p(x,y)$$
$$= \sum_{x \in A} \sum_{y \in B} p_X(x)p_Y(y) \qquad \text{perché stiamo supponendo vera l'Equazione (4.3.14)}$$
$$= \sum_{x \in A} p_X(x) \sum_{y \in B} p_Y(y)$$
$$= P(X \in A)P(Y \in B)$$

Nel caso di variabili aleatorie congiuntamente continue invece, X e Y sono indipendenti se e solo se la densità congiunta è il prodotto delle marginali:

$$f(x,y) = f_X(x)f_Y(y), \qquad \forall x, y \in \mathbb{R} \tag{4.3.15}$$

Questa ulteriore equivalenza può essere provata con passaggi simili a quelli qui sopra, sfruttando le Equazioni (4.3.7) e (4.3.8).

Il senso della definizione e delle molte forme equivalenti che abbiamo dato è che due variabili aleatorie sono indipendenti se conoscere il valore di una non cambia la distribuzione dell'altra.

Esempio 4.3.4. Siano assegnate due variabili aleatorie, X e Y, indipendenti e con la stessa funzione di densità,

$$f_X(t) = f_Y(t) = \begin{cases} e^{-t} & t > 0 \\ 0 & \text{altrimenti} \end{cases}$$

Qual è la densità di probabilità della variabile aleatoria data dal rapporto X/Y?

Occorre per prima cosa calcolare la funzione di ripartizione di X/Y. Per $a > 0$,

$$\begin{aligned}
F_{X/Y}(a) &:= P(X/Y \le a) && \text{per la definizione di } F \\
&= \iint_{(x,y):x/y\le a} f(x,y)\,dx\,dy && \text{per la definizione di } f \\
&= \iint_{(x,y):x\le ay} f(x)f(y)\,dx\,dy && \text{usando l'indipendenza} \\
&= \int_0^\infty \int_0^{ay} f(x)f(y)\,dx\,dy && \text{sostituendo gli estremi di integrazione corretti, come nell'Esempio 4.3.3} \\
&= \int_0^\infty e^{-y}\left(\int_0^{ay} e^{-x}\,dx\right)dy \\
&= \int_0^\infty e^{-y}(1-e^{-ay})dy \\
&= \left[-e^{-y} + \frac{e^{-(a+1)y}}{a+1}\right]\Bigg|_{y=0}^{\infty} \\
&= 1 - \frac{1}{a+1}
\end{aligned}$$

La funzione di densità si ricava infine derivando la funzione di ripartizione rispetto la suo argomento.

$$f_{X/Y}(a) = \frac{d}{da}\left(1 - \frac{1}{a+1}\right) = \frac{1}{(a+1)^2}, \quad a > 0 \quad \square$$

4.3.4 Generalizzazione a più di due variabili aleatorie

Tutti gli argomenti della Sezione 4.3 si possono estendere in maniera più o meno naturale ad un numero arbitrario n di variabili aleatorie. La funzione di ripartizione

congiunta di $X_1, X_2, \ldots, X_n$ è la funzione di n variabili F, definita da

$$F(a_1, a_2, \ldots, a_n) := P(X_1 \le a_1, X_2 \le a_2, \ldots, X_n \le a_n) \tag{4.3.16}$$

Se queste variabili aleatorie sono discrete, è possibile definire la funzione di massa di probabilità congiunta p, che è data da

$$p(x_1, x_2, \ldots, x_n) := P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \tag{4.3.17}$$

Altrimenti, le variabili aleatorie $X_1, X_2, \ldots, X_n$ sono congiuntamente continue, se esiste una densità di probabilità congiunta f, funzione di n variabili a valori positivi tale che, per ogni sottoinsieme C di $\mathbb{R}^n$,

$$P\big((X_1, X_2, \ldots, X_n) \in C\big) = \underset{(x_1,x_2,\ldots,x_n)\in C)}{\iint\cdots\int} f(x_1, x_2, \ldots, x_n)\,dx_1\,dx_2\cdots dx_n \tag{4.3.18}$$

Ciò significa in particolare che se $A_1, A_2, \ldots, A_n$ sono insiemi di numeri reali, allora

$$\begin{aligned}
&P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) \\
&\qquad = \int_{A_1}\int_{A_2}\cdots\int_{A_n} f(x_1, x_2, \ldots, x_n)\,dx_1\,dx_2\cdots dx_n
\end{aligned} \tag{4.3.19}$$

Anche il concetto di indipendenza si estende a più di due dimensioni. In generale n variabili aleatorie $X_1, X_2, \ldots, X_n$ si dicono indipendenti se per ogni n-upla $A_1, A_2, \ldots, A_n$ di sottoinsiemi di $\mathbb{R}$, è soddisfatta l'equazione

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i)$$

Di nuovo, si può dimostrare che ciò è equivalente a chiedere che per ogni n-upla $a_1, a_2, \ldots, a_n$ di numeri reali, sia soddisfatta l'equazione

$$F(a_1, a_2, \ldots, a_n) = \prod_{i=1}^{n} F_{X_i}(a_i) \tag{4.3.20}$$

Per concludere, collezioni *infinite* di variabili aleatorie si dicono indipendenti se ogni loro sottogruppo finito è formato da variabili aleatorie tutte indipendenti.

Esempio 4.3.5. Assumiamo per semplicità che le variazioni giornaliere del prezzo di un titolo azionario siano variabili aleatorie indipendenti e identicamente distribuite,

con funzione di massa data da

$$P(X_i = k) = \begin{cases} 0.05 & \text{se } k = -3 \\ 0.10 & \text{se } k = -2 \\ 0.20 & \text{se } k = -1 \\ 0.30 & \text{se } k = 0 \\ 0.20 & \text{se } k = +1 \\ 0.10 & \text{se } k = +2 \\ 0.05 & \text{se } k = +3 \end{cases} \qquad \forall i = 1, 2, \ldots$$

dove con X_i abbiamo indicato la variazione di prezzo nel giorno i-esimo. La probabilità con cui si osservano in tre giorni consecutivi degli incrementi successivi di 1, 2 e 0 punti, è data da

$$P(X_1 = 1, X_2 = 2, X_3 = 0) = 0.20 \times 0.10 \times 0.30 = 0.006 \quad \square$$

4.3.5 * Distribuzioni condizionali

Le relazioni esistenti tra due variabili aleatorie possono essere chiarite dallo studio della distribuzione condizionale di una delle due, dato il valore dell'altra. Si ricorda che presi comunque due eventi E e F con $P(F) > 0$, la probabilità di E condizionata a F è data dall'espressione

$$P(E|F) := \frac{P(E \cap F)}{P(F)}$$

È naturale applicare questo schema alle variabili aleatorie *discrete*.

Definizione 4.3.5. Siano X e Y due variabili aleatorie discrete con funzione di massa congiunta $p(\cdot, \cdot)$. Si dice *funzione di massa di probabilità condizionata* di X dato Y, e si indica con $p_{X|Y}(\cdot|\cdot)$, la funzione di due variabili così definita:

$$\begin{aligned} p_{X|Y}(x|y) &:= P(X = x|Y = y) \\ &= \frac{P(X = x, Y = y)}{P(Y = y)} \\ &= \frac{p(x, y)}{p_Y(y)}, \qquad \forall x, \forall y \text{ con } p_Y(y) > 0 \end{aligned} \tag{4.3.21}$$

Se y non è un valore possibile di Y, ovvero se $P(Y = y) = 0$, la quantità $p_{X|Y}(x|y)$ non è definita.

Esempio 4.3.6. Riguardo all'Esempio 4.3.2, aggiungendo l'informazione che la famiglia selezionata ha esattamente una figlia, qual è la funzione di massa condizionata del numero di figli maschi?

Notiamo intanto dalla Tabella 4.2 che $P(X = 1) = 0.3875$, informazione che useremo più volte.

$$\begin{aligned} P(Y = 0|X = 1) &= \frac{P(Y = 0, X = 1)}{P(X = 1)} = \frac{0.1}{0.3875} = \frac{8}{31} \\ P(Y = 1|X = 1) &= \frac{P(Y = 1, X = 1)}{P(X = 1)} = \frac{0.175}{0.3875} = \frac{14}{31} \\ P(Y = 2|X = 1) &= \frac{P(Y = 2, X = 1)}{P(X = 1)} = \frac{0.1125}{0.3875} = \frac{9}{31} \\ P(Y = 3|X = 1) &= \frac{P(Y = 3, X = 1)}{P(X = 1)} = 0 \end{aligned}$$

Quindi, per fare un esempio, data la presenza di una figlia, vi sono 23 possibilità su 31 che vi sia anche almeno un maschio. $\square$

Esempio 4.3.7. Siano X e Y due variabili aleatorie discrete con funzione di massa congiunta p, data da

$$p(0,0) = 0.4, \qquad p(0,1) = 0.2, \qquad p(1,0) = 0.1, \qquad p(1,1) = 0.3$$

Qual è la funzione di massa di X condizionata a $Y = 1$?

Per prima cosa, calcoliamo $P(Y = 1)$,

$$P(Y = 1) = \sum_x p(x, 1) = p(0,1) + p(1,1) = 0.5$$

Quindi,

$$\begin{aligned} P(X = 0|Y = 1) &= \frac{p(0,1)}{P(Y = 1)} = \frac{2}{5} \\ P(X = 1|Y = 1) &= \frac{p(1,1)}{P(Y = 1)} = \frac{3}{5} \end{aligned} \quad \square$$

Se X e Y sono variabili congiuntamente continue, non è possibile utilizzare la definizione di distribuzione condizionata valida per quelle discrete, infatti sappiamo che $P(Y = y) = 0$ per tutti i valori di y (si veda a pagina 96).

Definizione 4.3.6. Siano X e Y due variabili aleatorie con funzione di densità congiunta f. Si dice *densità condizionale* di X rispetto a Y, e si indica con $f_{X|Y}(\cdot|\cdot)$, la funzione di due variabili seguente, che è definita per ogni x e per tutte le y per le quali $f_Y(y) > 0$:

$$f_{X|Y}(x, y) := \frac{f(x, y)}{f_Y(y)} \tag{4.3.22}$$

Tale definizione è giustificata dalle Equazioni (4.2.9) e (4.3.9). Infatti moltiplicando il lato sinistro della (4.3.22) per dx e quello destro per $\frac{dx\,dy}{dy}$, si ottiene

$$\begin{aligned} f_{X|Y}(x|y)\,dx &= \frac{f(x,y)\,dx\,dy}{f_Y(y)\,dy} \\ &\approx \frac{P(x \le X \le x+dx, y \le Y \le y+dy)}{P(y \le Y \le y+dy)} \\ &= P(x \le X \le x+dx \,|\, y \le Y \le y+dy) \end{aligned}$$

In altre parole, per valori piccoli di dx e di dy, $f_{X|Y}\,dx$ rappresenta la probabilità condizionata che X stia nell'intervallo $[x, x+dx]$, sapendo che Y appartiene all'intervallo $[y, y+dy]$.

La densità condizionale ci permette di definire la probabilità di eventi relativi a una variabile aleatoria quando conosciamo il valore di una seconda. Più precisamente se X e Y sono congiuntamente continue e A è un sottoinsieme dei numeri reali, per ogni y si può definire

$$P(X \in A|Y=y) := \int_A f_{X|Y}(x|y)dx \tag{4.3.23}$$

La grandezza $P(X \in A|Y=y)$ non è una probabilità condizionata nel senso usuale del termine, in quanto l'evento $\{Y=y\}$ ha sempre probabilità zero. Ciononostante, sfruttando la densità condizionale di X rispetto a Y siamo riusciti a dare un senso e persino un valore numerico a questo oggetto di sicuro interesse pratico[2].

Si noti che se X e Y sono indipendenti, allora

$$f_{X|Y}(x,y) = f_X(x), \qquad P(X \in A|Y=y) = P(X \in A)$$

e quindi l'indipendenza si comporta nei confronti del condizionamento rispetto a variabili aleatorie continue, esattamente come nel caso più semplice di condizionamento rispetto a eventi di probabilità positiva.

Esempio 4.3.8. È data la seguente densità congiunta di X e Y:

$$f(x,y) = \begin{cases} \frac{12}{5}x(2-x-y) & 0<x<1, 0<y<1 \\ 0 & \text{altrimenti} \end{cases}$$

[2] Per distinguere i condizionamenti "veri" (fatti cioè rispetto ad eventi di probabilità positiva) da quelli "impropri" come quello dell'Equazione (4.3.23), in italiano si usa nel primo caso l'aggettivo "condiziona-*to/a*", e nel secondo l'aggettivo "condiziona-*le*". Non tutti concordano su questa nomenclatura, e in molti testi questi termini sono utilizzati indifferentemente, tuttavia l'importanza concettuale della distinzione è straordinaria, [*N.d.T.*]

Si calcoli la densità condizionale di X rispetto a $Y=y$, per $0<y<1$.

Se x e y sono compresi tra 0 e 1, abbiamo:

$$\begin{aligned} f_{X|Y}(x|y) &:= \frac{f(x,y)}{f_Y(y)} \\ &= \frac{f(x,y)}{\int_{-\infty}^{\infty} f(x',y)\,dx'} && \text{sfruttando la (4.3.11); nota: attenzione a non riutilizzare } x \text{ per l'integrale!} \\ &= \frac{x(2-x-y)}{\int_0^1 x'(2-x'-y)\,dx'} \\ &= \frac{x(2-x-y)}{\frac{2}{3}-\frac{y}{2}} \\ &= \frac{6x(2-x-y)}{4-3y} \quad \square \end{aligned}$$

4.4 Valore atteso

Uno dei concetti più importanti in tutta la teoria della probabilità è quello di valore atteso.

Definizione 4.4.1. Sia X una variabile aleatoria discreta che può assumere i valori $x_1, x_2, \ldots$; il *valore atteso* di X, che si indica con $E[X]$, è (se esiste[3]) il numero

$$E[X] := \sum_i x_i P(X = x_i) \tag{4.4.1}$$

In altri termini, si tratta della media pesata dei valori possibili di X, usando come pesi le probabilità che tali valori vengano assunti da X. Per questo $E[X]$ è anche detta *media* di X (anche se questo termine è poco consigliato perché può assumere anche altri significati), oppure *aspettazione* (dal termine inglese *expectation*).

[3] Il valore atteso di X è definito solo se la serie (4.4.1) convergere in valore assoluto, ovvero deve valere

$$\sum_i |x_i| P(X = x_i) < \infty$$

In caso contrario si dice che X non ha valore atteso. Tutte le variabili aleatorie che tratteremo nel seguito si supporranno dotate di valore atteso finito. Esempi di distribuzioni per le quali il valore atteso non ha senso sono dati dalle funzioni di massa seguenti:

$$p_1(k) = \begin{cases} 2^{-n-1} & \text{se } k = \pm 2^n, n = 1,2,\ldots \\ 0 & \text{altrimenti} \end{cases} \quad \text{e} \quad p_2(k) = \begin{cases} \dfrac{1}{k^2+k} & k = 1,2,\ldots \\ 0 & \text{altrimenti} \end{cases}$$

Per illustrare il concetto di media pesata, facciamo un semplice esempio. Se X è una variabile aleatoria con funzione di massa

$$p(0) = \frac{1}{2} = p(1)$$

allora,

$$E[X] = 0 \times \frac{1}{2} + 1 \times \frac{1}{2} = \frac{0+1}{2} = \frac{1}{2}$$

è semplicemente la media aritmetica dei valori che X può assumere. Però, se

$$p(0) = \frac{1}{3}, \quad p(1) = \frac{2}{3}$$

allora,

$$E[X] = 0 \times \frac{1}{3} + 1 \times \frac{2}{3} = \frac{0+1\times 2}{3} = \frac{2}{3}$$

è una media pesata degli stessi valori 0 e 1, dove al secondo è stato dato un peso che è il doppio di quello del primo.

L'interpretazione frequentista della probabilità fornisce una importante giustificazione del concetto di valore atteso. Da tale punto di vista la probabilità di un evento è definita come il limite a cui tende – empiricamente – il rapporto tra il numero di ripetizioni in cui si è realizzato l'evento e il numero totale di ripetizioni di un esperimento. Consideriamo una variabile aleatoria X che può assumere i valori $x_1, x_2, \ldots, x_n$, con funzione di massa di probabilità p. Immaginando che X sia la vincita in una singola mano di un gioco casuale, qual è la vincita media (nel senso comune del termine) se giochiamo molte mani? Su un numero N di ripetizioni dell'esperimento, ciascuno degli valori x_i si verificherà un certo numero N_i di volte. L'interpretazione frequentista afferma che se N è molto grande, $N_i \approx Np(x_i)$. D'altronde ci si convince facilmente[4] che la vincita media è data da

$$\frac{x_1N_1 + x_2N_2 + \cdots + x_nN_n}{N} = \sum_{i=1}^{n} x_i \frac{N_i}{N} \approx \sum_{i=1}^{n} x_i p(x_i) =: E[X]$$

e quindi coincide approssimativamente con la definizione di valore atteso di X.

Esempio 4.4.1. Sia X il punteggio che si ottiene lanciando un dado non truccato. Quanto vale $E[X]$?

Siccome $p(1) = p(2) = p(3) = p(4) = p(5) = p(6) = 1/6$, ricaviamo che

$$E[X] := 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{7}{2} = 3.5$$

[4] Il ragionamento da fare è analogo a quello che ci ha portati all'Equazione (2.3.4) di pagina 23, a proposito della media di un campione di dati fornito tramite le frequenze assolute dei suoi valori.

È utile notare che in questo esempio, il valore atteso di X non è uno dei valori che X può assumere. (Tirando un dado non c'è modo di ottenere un punteggio di 3.5.) Perciò, anche se $E[X]$ è chiamato *valore atteso* si X, non vuole affatto dire che noi ci attendiamo di vedere questo valore, ma piuttosto che ci aspettiamo che sia il limite a cui tende il punteggio medio del dado su un numero crescente di ripetizioni. In effetti su molti lanci di dado la media aritmetica di tutti i valori ottenuti tende a $7/2$. (Lo studente curioso dovrebbe cimentarsi in questo esperimento.) □

Esempio 4.4.2. Se I è la funzione indicatrice di un evento A, ovvero se

$$I := \begin{cases} 1 & \text{se } A \text{ si verifica} \\ 0 & \text{se } A \text{ non si verifica} \end{cases}$$

allora

$$E[I] := 1 \cdot P(I = 1) + 0 \cdot P(I = 0) = P(I = 1) = P(A)$$

Quindi il valore atteso della funzione indicatrice di un evento è la probabilità di quest'ultimo. □

Esempio 4.4.3 (Entropia). Sia assegnata una variabile aleatoria discreta X. Quanta informazione è contenuta nel verificarsi dell'evento $\{X = x\}$? Questa domanda nasce all'interno della teoria dell'informazione, una branca della probabilità che studia i flussi di dati in vari tipi di comunicazioni. La variabile aleatoria X rappresenta qui un frammento del messaggio ricevuto dal destinatario (ad esempio la prima parola di una e-mail). Vogliamo avviarci a dare una risposta al quesito proposto per piccoli passi. In primo luogo sembra ragionevole che la quantità di informazione portata dal *messaggio* $\{X = x\}$ dipenda dalla probabilità che X sia uguale a x. Inoltre è sensato che il messaggio contenga tanta più informazione quanto più rara è la sua occorrenza. Infatti se X è la somma di due dadi, sembra esserci più informazione nel messaggio $\{X = 12\}$ di quanta ve ne sia in $\{X = 7\}$, nel primo caso la probabilità è $1/36$ e sappiamo esattamente quanto hanno totalizzato entrambi i dadi (6 entrambi), nel secondo caso la probabilità è $1/6$ e non abbiamo idea di quanto abbiano totalizzato i singoli dadi. Se invece X è la prima parola di una e-mail, sembra esserci più informazione nel messaggio "Domani" che nel messaggio, molto più frequente "Caro".

Denotiamo allora con $I(p)$ la quantità di informazione contenuta nel realizzarsi di un evento di probabilità p. È chiaro che $I(p)$ dovrà essere non negativa, e decrescente in p. Per determinarne l'espressione, aggiungiamo un requisito, ovvero che sia additiva rispetto alla somma di messaggi. Cosa ciò significhi è illustrato in quanto segue. Supponiamo che X e Y siano due variabili aleatorie indipendenti, e che $\{X = x\}$ e $\{Y = y\}$ siano due messaggi di probabilità p e q rispettivamente. Quanta informazione è contenuta nel messaggio che X è pari a x e Y è pari a y? Per prima cosa,

$I(p)$ è l'informazione contenuta nel solo messaggio $\{X = x\}$; poi, siccome vale l'indipendenza di X e Y, il valore di X non influenza la distribuzione di Y, e perciò pare sensato che l'informazione aggiunta da $\{Y = y\}$ sia $I(q)$ indipendentemente dal valore di X. Concludendo, appare ragionevole che l'informazione contenuta in $\{X = x, Y = y\}$ sia pari a $I(p) + I(q)$. Siccome poi

$$P(X = x, Y = y) = P(X = x)P(Y = y) = pq$$

se ne deduce che deve valere

$$I(pq) = I(p) + I(q)$$

Ora, se costruiamo la funzione $G(a) := I(e^a)$, si vede che essa è ancora monotona, e inoltre è *additiva*, infatti:

$$G(a + b) := I(e^{a+b}) = I(e^a e^b) = I(e^a) + I(e^b) =: G(a) + G(b)$$

Ma è noto che le uniche funzioni monotone e additive sono quelle della forma $G(a) = ca$ per qualche costante a. Perciò, siccome $I(p) = G(\log p)$,

$$I(p) = G(\log p) = c \log p$$

La convenzione è di porre $c = -\frac{1}{\log 2}$, in modo tale che risulti

$$I(p) = -\log_2(p)$$

Con questa scelta di c l'informazione viene misurata in *bit*, ovvero in cifre binarie (in inglese, *binary digits* di cui *bit* è l'abbreviazione).

Si consideri adesso una variabile aleatoria X che possa assumere i valori $x_1, x_2, \ldots, x_n$ con probabilità $p_1, p_2, \ldots, p_n$ rispettivamente. Siccome tutte le volte che $X = x_i$, l'informazione ricevuta è pari a $-\log_2(p_i)$, il valore atteso dell'informazione contenuta in X sarà pari a

$$H(X) := -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{4.4.2}$$

Il valore $H(X)$ è noto in teoria dell'informazione con il nome di *entropia* della variabile aleatoria X.

Si noti che l'entropia di un bit casuale è ... 1 bit. (Lo studente verifichi che se X assume i valori 0 o 1 con probabilità $1/2$, allora $H(X) = 1$.) □

È anche possibile definire il valore atteso di una variabile aleatoria continua. Se X ha densità di probabilità f e dx è abbastanza piccolo,

$$f(x)\,dx \approx P(x < X < x + dx)$$

Ne segue che una media pesata dei valori di X con il peso di ciascun x dato dalla probabilità che X sia vicino a x, è semplicemente l'integrale su tutto $\mathbb{R}$ di $xf(x)$.

Definizione 4.4.2. Sia X una variabile aleatoria continua con funzione di densità f; il *valore atteso*, o *aspettazione* o anche *media* di X, che si indica con $E[X]$, è (se esiste[5]) la quantità

$$E[X] := \int_{-\infty}^{\infty} x f(x)\,dx \tag{4.4.3}$$

Esempio 4.4.4. Siamo in attesa di una comunicazione che deve arrivare dopo le ore 17. Dall'esperienza passata è noto che il numero di ore X che è necessario aspettare a partire dalle 17 è una variabile aleatoria con funzione di densità data da

$$f(x) = \begin{cases} \frac{1}{1.5} & \text{se } 0 < x < 1.5 \\ 0 & \text{altrimenti} \end{cases}$$

Il valore atteso del tempo che trascorre tra le 17 e il momento di arrivo della comunicazione è quindi

$$E[X] = \int_0^{1.5} \frac{x}{1.5}\,dx = 0.75$$

Quindi, in media, sarà necessario aspettare tre quarti d'ora. □

Osservazione 4.4.1. Il concetto di valore atteso è analogo in fisica al concetto di centro di gravità o *baricentro* di una distribuzione di massa. Consideriamo una variabile aleatoria discreta X con funzione di massa di probabilità $P(x_i)$, per $i \geq 1$. Se immaginiamo un'asta ideale, priva di peso, graduata e dotata, in corrispondenza dei valori di ascissa x_i, di pesi di massa $P(x_i)$, per $i \geq 1$ (si veda la Figura 4.4), allora il suo baricentro, l'unico punto in cui l'asta con i pesi si potrebbe sostenere rimanendo in equilibrio, si trova al valore di ascissa $E[X]$. Ciò può essere provato se il lettore conosce i rudimenti della statica, notando che, se il fulcro è posto in $\bar{x}$, il momento totale delle forze peso agenti è dato da $\sum_i P(x_i)(x_i - \bar{x})$, che chiaramente è nullo se e solo se $\bar{x} = E[X]$.

Osservazione 4.4.2. $E[X]$ ha le stesse unità di misura della variabile aleatoria X.

[5] Di nuovo, si richiede una convergenza in valore assoluto; deve valere

$$\int_{-\infty}^{\infty} |x| f(x)\,dx < \infty$$

4.5 Proprietà del valore atteso

Consideriamo una variabile aleatoria X di cui conosciamo la distribuzione (si veda l'Osservazione 4.2.1). Se anziché volere calcolare il valore atteso di X, ci interessasse determinare quello di una sua qualche funzione $g(X)$, come potremmo fare? Una prima strada è notare che $g(X)$ stessa è una variabile aleatoria, e quindi ha una sua distribuzione che in qualche modo si può ricavare; dopo averla ottenuta, il valore atteso $E[g(X)]$ si calcola con la definizione usuale applicata alla nuova variabile aleatoria.

Esempio 4.5.1. Quanto vale il valore atteso del quadrato di una variabile aleatoria X con funzione di massa seguente?

$$p(0) = 0.2, \qquad p(1) = 0.5, \qquad p(2) = 0.3$$

Poniamo $Y := X^2$. Questa è una variabile aleatoria che può assumere i valori 0^2, 1^2 e 2^2, con probabilità

$$\begin{aligned} p_Y(0) &:= P(Y = 0^2) = 0.2 \\ p_Y(1) &:= P(Y = 1^2) = 0.5 \\ p_Y(4) &:= P(Y = 2^2) = 0.3 \end{aligned}$$

Quindi,

$$E[X^2] = E[Y] = 0 \cdot 0.2 + 1 \cdot 0.5 + 4 \cdot 0.3 = 1.7 \quad \square$$

Esempio 4.5.2. Il tempo – in ore – necessario per localizzare un guasto nell'impianto elettrico di una fabbrica è una variabile aleatoria X con funzione di densità

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{altrimenti} \end{cases}$$

Se il danno economico provocato da una interruzione di x ore è x^3, qual è il valore atteso di questo costo?

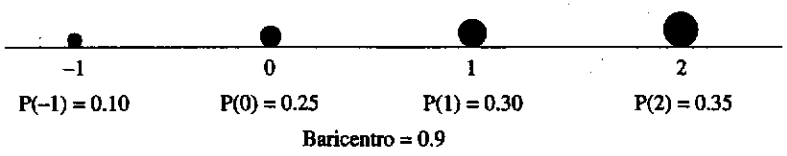


Figura 4.4

Poniamo $Y := X^3$, che rappresenta la variabile aleatoria "costo di una interruzione". La sua distribuzione si può ricavare in maniera molto efficace tramite la funzione di ripartizione. Sia $0 < a < 1$,

$$\begin{aligned} F_Y(a) &:= P(Y \le a) \\ &= P(X^3 \le a) \\ &= P(X \le a^{1/3}) && \text{perché la funzione } x \mapsto x^{1/3} \text{ è crescente} \\ &= \int_0^{a^{1/3}} 1\,dx && \text{l'integrale parte da 0 perché } f \text{ è nulla sui negativi} \\ &= a^{1/3} \end{aligned}$$

Derivando F_Y si trova la densità di Y,

$$f_Y(a) = \frac{1}{3}a^{-2/3}, \qquad 0 \le a < 1$$

Infine, otteniamo $E[X^3]$ come $E[Y]$, visto che coincidono.

$$\begin{aligned} E[Y] &:= \int_{-\infty}^{\infty} a f_Y(a)\,da \\ &= \int_0^1 a \cdot \frac{1}{3}a^{-2/3}\,da \\ &= \frac{1}{3}\int_0^1 a^{1/3}\,da \\ &= \frac{1}{3}\left[\frac{3}{4}a^{4/3}\right]\Big|_{a=0}^{1} = \frac{1}{4} \quad \square \end{aligned}$$

Anche se la procedura descritta permette in principio di calcolare il valore atteso di qualunque funzione di una variabile aleatoria di cui conosciamo la distribuzione, esiste un approccio più semplice che porta agli stessi risultati. Supponiamo infatti di volere determinare il valore atteso di $g(X)$: siccome questa variabile aleatoria assume il valore $g(x)$ quando $X = x$, sembra intuitivo che $E[g(X)]$ coincida con la media pesata dei valori possibili di $g(X)$, usando come peso da dare a $g(x)$ la probabilità (o densità di probabilità nel caso continuo) che X sia pari a x. Quanto detto può essere dimostrato in maniera rigorosa, e l'enunciato formale che ne risulta è il seguente.

Proposizione 4.5.1 (Valore atteso di una funzione di variabile aleatoria).

1. Se X è una variabile aleatoria discreta con funzione di massa di probabilità p, allora, per ogni funzione reale g,

$$E[g(X)] = \sum_x g(x)p(x) \tag{4.5.1}$$

2. Se X è una variabile aleatoria continua con funzione di densità di probabilità f, allora, per ogni funzione reale g,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx \tag{4.5.2}$$

Anche in questo caso si richiede, affinché $E[g(X)]$ abbia senso, che la serie (4.5.1) e l'integrale (4.5.2) convergano in valore assoluto[6]. Nel seguito questa questione di esistenza e buona definizione non verrà più approfondita. Si tenga comunque presente che ogni volta che una grandezza numerica è definita tramite il valore atteso di una quantità aleatoria, la questione si pone, e in principio sarebbe necessario verificare la convergenza in valore assoluto caso per caso. Nella pratica sono poche (ma non assenti) le variabili aleatorie che non soddisfano tali verifiche.

Esempio 4.5.3. Applicando la Proposizione 4.5.1 alla situazione dell'Esempio 4.5.1, si trova immediatamente,

$$E[X^2] = 0^2 \cdot 0.2 + 1^2 \cdot 0.5 + 2^2 \cdot 0.3 = 1.7$$

che ovviamente conferma il valore già trovato. □

Esempio 4.5.4. Applicando la Proposizione 4.5.1 alla situazione dell'Esempio 4.5.2, si ottiene, ricordando che $f(x) = 1$ per $0 < x < 1$, che

$$E[X^3] = \int_0^1 x^3\,dx = \frac{1}{4} \quad \square$$

Quello che segue è un facile corollario della Proposizione 4.5.1.

Corollario 4.5.2. Per ogni coppia di costanti reali a e b,

$$E[aX + b] = aE[X] + b \tag{4.5.3}$$

Dimostrazione. Nel caso discreto,

$$\begin{aligned} E[aX + b] &= \sum_x (ax + b)p(x) \\ &= a\sum_x xp(x) + b\sum_x p(x) \\ &= aE[X] + b \qquad \text{usando la (4.2.2)} \end{aligned}$$

[6] Ovvero, deve essere rispettivamente

$$\sum_x |g(x)|p(x) < \infty \qquad \text{o} \qquad \int_{-\infty}^{\infty} |g(x)|f(x)\,dx < \infty$$

Nel caso continuo,

$$\begin{aligned} E[aX + b] &= \int_{-\infty}^{\infty} (ax + b)f(x)\,dx \\ &= a\int_{-\infty}^{\infty} xf(x)\,dx + b\int_{-\infty}^{\infty} f(x)\,dx \\ &= aE[X] + b \qquad \text{usando la (4.2.5)} \quad \square \end{aligned}$$

Se nel Corollario 4.5.2 si pone $a = 0$, si scopre che

$$E[b] = b$$

In altri termini, il valore atteso di una costante, è semplicemente il suo valore stesso. (Il lettore si convinca del significato di questa affermazione!) Se invece si pone $b = 0$, si ottiene che

$$E[aX] = aE[X]$$

Ovvero, il valore atteso di un fattore costante moltiplicato per una variabile aleatoria, è pari alla costante per il valore atteso della variabile aleatoria.

Come già accennato, il termine *valore atteso* ha tra i suoi sinonimi *aspettazione* e *media*. Un'ulteriore denominazione è quella di *momento primo*, con riferimento alla definizione seguente.

Definizione 4.5.1. Se $n = 1, 2, \ldots$, la quantità $E[X^n]$, quando esiste, è detta *momento n-esimo* della variabile aleatoria X.

Volendo essere più espliciti, si può applicare il Corollario 4.5.2 per ricavare,

$$E[X^n] = \begin{cases} \displaystyle\sum_x x^n p(x) & \text{se } X \text{ è discreta} \\ \displaystyle\int_{-\infty}^{\infty} x^n f(x)\,dx & \text{se } X \text{ è continua} \end{cases}$$

4.5.1 Valore atteso della somma di variabili aleatorie

La versione in due dimensioni della Proposizione 4.5.1 afferma che se X e Y sono due variabili aleatorie e g è una qualunque funzione di due variabili, allora, se $E[g(X, Y)]$ esiste,

$$E[g(X,Y)] = \begin{cases} \displaystyle\sum_x \sum_y g(x,y)p(x,y) & \text{nel caso discreto} \\ \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\,dx\,dy & \text{nel caso continuo} \end{cases} \tag{4.5.4}$$

Si può applicare questo enunciato a $g(X,Y) = X + Y$, ottenendo che

$$E[X+Y] = E[X] + E[Y] \tag{4.5.5}$$

Tale risultato è valido sia nel caso discreto (che si lascia al lettore), sia in quello continuo, come è dimostrato dai passaggi seguenti.

$$\begin{aligned} E[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty} x\left[\int_{-\infty}^{\infty} f(x,y)\,dy\right]dx + \int_{-\infty}^{\infty} y\left[\int_{-\infty}^{\infty} f(x,y)\,dx\right]dy \\ &= \int_{-\infty}^{\infty} x f_X(x)\,dx + \int_{-\infty}^{\infty} y f_Y(y)\,dy, \qquad \text{usando le Equazioni (4.3.10) e (4.3.11)} \\ &= E[X] + E[Y] \end{aligned}$$

Applicando ricorsivamente l'Equazione (4.5.5) si può estenderne la portata alla somma di un numero finito di variabili aleatorie. Ad esempio,

$$\begin{aligned} E[X+Y+Z] &= E[(X+Y)+Z] \\ &= E[X+Y] + E[Z] \qquad \text{applicando la (4.5.5) a } (X+Y) \text{ e } Z \\ &= E[X] + E[Y] + E[Z] \qquad \text{applicando la (4.5.5) a } X \text{ e } Y \end{aligned}$$

E in generale, per ogni n,

$$E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n] \tag{4.5.6}$$

L'Equazione (4.5.6) costituisce una formula di grande utilità, come è illustrato dai prossimi esempi.

Esempio 4.5.5. Un'impresa edile ha recentemente sottoposto i suoi preventivi per tre gare, per degli appalti che le darebbero profitti per 10 000, 20 000 e 40 000 mila dollari. Se le probabilità di vittoria dei singoli appalti sono rispettivamente 0.2, 0.8 e 0.3, qual è il profitto totale medio che farà l'azienda?

Siano X_1, X_2 e X_3 i profitti (in migliaia di dollari) percepiti per i tre lavori. Il profitto totale Y sarà dato da $Y := X_1 + X_2 + X_3$, e quindi

$$E[Y] = E[X_1] + E[X_2] + E[X_3]$$

Siccome ciascuno degli X_i può essere nullo o pari a un valore fissato con probabilità specificate dal problema, si trova che

$$\begin{aligned} E[X_1] &= 10 \times 0.2 + 0 \times 0.8 = 2 \\ E[X_2] &= 20 \times 0.8 + 0 \times 0.2 = 16 \\ E[X_3] &= 40 \times 0.3 + 0 \times 0.7 = 12 \end{aligned}$$

Perciò il profitto totale medio dell'azienda è di 30 000 dollari. □

Esempio 4.5.6. Una segretaria ha finito di scrivere una pila di N lettere, e ha appena compilato le buste con gli indirizzi, quando tutto il materiale le cade per terra e si mischia. Se si inseriscono le lettere nelle buste in maniera del tutto casuale (nel senso che ciascuna lettera può finire in ogni busta con pari probabilità), qual è il numero medio di lettere che capitano nella busta corretta?

Sia X il numero di lettere che finiscono nella busta giusta. Il valore atteso $E[X]$ può essere calcolato molto facilmente notando che $X = X_1 + X_2 + \cdots + X_N$, dove

$$X_i := \begin{cases} 1 & \text{se la lettera } i\text{-esima viene inserita nella propria busta} \\ 0 & \text{altrimenti} \end{cases}$$

Siccome l'i-esima lettera può finire in una qualunque delle N buste con pari probabilità,

$$P(X_i = 1) = P(\text{la lettera } i\text{-esima è nella sua busta}) = 1/N$$

e quindi

$$E[X_i] = 1 \cdot P(X_i = 1) + 0 \cdot P(X_i = 0) = 1/N$$

Perciò, otteniamo dall'Equazione (4.5.6) che

$$E[X] = E[X_1] + \cdots + E[X_N] = N\frac{1}{N} = 1$$

Quindi, indipendentemente dal numero di lettere presenti, in media vi sarà una sola lettera nella busta giusta. □

Esempio 4.5.7. In un prodotto commerciale vengono inseriti dei buoni sconto in regalo. Vi sono 20 tipi diversi di buoni, e in ogni confezione se ne trova uno qualsiasi con pari probabilità. Se si aprono 10 confezioni, quant'è il valore atteso del numero di tipi diversi di buoni sconto che si trovano?

Sia X il numero di tipi diversi di buoni che troviamo nelle 10 confezioni. Allora $X = X_1 + \cdots + X_{20}$, dove

$$X_i := \begin{cases} 1 & \text{se il tipo } i\text{-esimo di buoni è presente nelle 10 confezioni} \\ 0 & \text{altrimenti} \end{cases}$$

Le X_i si studiano facilmente,

$$\begin{aligned} E[X_i] &= P(X_i = 1) \\ &= P(\text{il tipo } i\text{-esimo di buoni è presente nelle 10 confezioni}) \\ &= 1 - P(\text{il tipo } i\text{-esimo di buoni non è presente nelle 10 confezioni}) \\ &= 1 - \left(\frac{19}{20}\right)^{10} \end{aligned}$$

dove l'ultima uguaglianza segue dal fatto che ciascuno dei 10 buoni sarà di tipo diverso da quello i-esimo (indipendentemente) con probabilità $19/20$. Concludendo,

$$E[X] = E[X_1] + \cdots + E[X_{20}] = 20\left[1 - \left(\frac{19}{20}\right)^{10}\right] \approx 8.025 \quad \square$$

Osservazione 4.5.1. Vi è una interessante proprietà della media che emerge quando si vuole *predire* con il minore errore possibile il valore che verrà assunto da una variabile aleatoria. Supponiamo di voler predire il valore di X. Se scegliamo un numero reale c e diciamo che X sarà uguale a c, il quadrato dell'errore che commetteremo è $(X-c)^2$. Mostriamo di seguito che la media dell'errore al quadrato[7] è minimizzata se per c scegliamo il valore della media di X. Infatti, detta $\mu := E[X]$,

$$\begin{aligned} E[(X-c)^2] &= E[(X-\mu+\mu-c)^2] \\ &= E[(X-\mu)^2 + 2(X-\mu)(\mu-c) + (\mu-c)^2] \\ &= E[(X-\mu)^2] + 2(\mu-c)E[X-\mu] + (\mu-c)^2 \\ &= E[(X-\mu)^2] + (\mu-c)^2 \qquad \text{infatti } E[X-\mu] = E[X] - \mu = 0 \\ &\geq E[(X-\mu)^2] \end{aligned}$$

Perciò la migliore previsione di X, in termini di minimizzazione dell'errore quadratico medio, è la sua aspettazione.

4.6 Varianza

Data una variabile aleatoria X, di cui sia nota la distribuzione, sarebbe molto utile se si potessero riassumere le caratteristiche fondamentali della sua distribuzione con quantità sintetiche come è la media $E[X]$. Tuttavia $E[X]$ è il "baricentro" dei valori possibili di X, e non coglie la variabilità, la dispersione di questi valori. Ad esempio, se W, Y e Z sono definite come segue,

$$W := 0 \text{ con probabilità } 1$$

$$Y := \begin{cases} -1 & \text{con probabilità } 1/2 \\ 1 & \text{con probabilità } 1/2 \end{cases}$$

$$Z := \begin{cases} -100 & \text{con probabilità } 1/2 \\ 100 & \text{con probabilità } 1/2 \end{cases}$$

allora tutte hanno media nulla, ma vi è molta più variabilità in Y che non in W (che è addirittura costante), e ancora di più in Z.

[7] Più comunemente nota come *errore quadratico medio*.

Siccome i valori di X sono distribuiti comunque attorno alla sua media $\mu := E[X]$, un approccio per misurare la loro variabilità potrebbe essere quantificare la loro distanza da μ, ad esempio calcolando quanto vale $E[|X-\mu|]$. Questo metodo in linea di principio funziona, nel senso che variabili aleatorie che assumono valori sparsi su un supporto più largo, sono associate a valori più elevati di questa grandezza, tuttavia le difficoltà matematiche che sorgono a causa del valore assoluto sono notevoli, e in realtà se lo si sostituisce con un elevamento al quadrato, si ottiene una definizione molto più fruttuosa.

Definizione 4.6.1. Sia X è una variabile aleatoria con media μ. La *varianza* di X, che si denota con $\mathrm{Var}(X)$, è (se esiste) la quantità

$$\mathrm{Var}(X) := E[(X-\mu)^2] \tag{4.6.1}$$

Esiste una formula alternativa per la varianza, che si ricava in questo modo:

$$\begin{aligned} \mathrm{Var}(X) &:= E[(X-\mu)^2] \\ &= E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - \mu^2 \end{aligned}$$

Ovvero,

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 \tag{4.6.2}$$

In altri termini, la varianza di X è uguale al valore atteso del quadrato di X (anche detto il *momento secondo*, si veda la Definizione 4.5.1), meno il quadrato della media di X. Nella pratica questa formula è spesso il miglior modo di calcolare $\mathrm{Var}(X)$.

Esempio 4.6.1. Si calcoli la varianza del punteggio di un dado non truccato.

Sia X il punteggio realizzato dal dado. Siccome $P(X=i) = 1/6$, per $i = 1, 2, \ldots, 6$, otteniamo

$$\begin{aligned} E[X^2] &= \sum_{i=1}^{6} i^2 P(X=i) \\ &= 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} = \frac{91}{6} \end{aligned}$$

Da cui, ricordando dall'Esempio 4.4.1 che $E[X] = 7/2$,

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12} \quad \square$$

Esempio 4.6.2 (Varianza della funzione indicatrice di un evento). Sia I la funzione indicatrice di un evento A:

$$I := \begin{cases} 1 & \text{se } A \text{ si verifica} \\ 0 & \text{se } A \text{ non si verifica} \end{cases}$$

Allora, notando che $I^2 = I$ sempre (infatti i valori possibili di I sono solamente 0 e 1, che soddisfano $1^2 = 1$ e $0^2 = 0$),

$$\begin{aligned} \operatorname{Var}(I) &= E[I^2] - E[I]^2 \\ &= E[I] - E[I]^2 && \text{perché } I^2 = I \text{ con probabilità 1} \\ &= E[I](1 - E[I]) \\ &= P(A)(1 - P(A)) && \text{perché } E[I] = P(A) \text{ dall'Esempio 4.4.2} \quad \square \end{aligned}$$

Una utile identità che riguarda la varianza è la seguente. Per ogni coppia di costanti reali a e b,

$$\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X) \tag{4.6.3}$$

Per dimostrarla, poniamo $\mu := E[X]$ e ricordiamo che $E[aX + b] = aE[X] + b = a\mu + b$, in modo tale che

$$\begin{aligned} \operatorname{Var}(aX + b) &:= E[(aX + b - E[aX + b])^2] \\ &= E[(aX + b - a\mu - b)^2] && \text{usando } E[aX + b] = a\mu + b \\ &= E[a^2(X - \mu)^2] && \text{semplificando e raccogliendo} \\ &= a^2 E[(X - \mu)^2] && \text{usando la (4.5.3)} \\ &= a^2 \operatorname{Var}(X) \end{aligned}$$

Se si sostituiscono valori particolari di a e b nell'Equazione (4.6.3), si ottengono diversi risultati interessanti. Ad esempio se poniamo $a = 0$ troviamo che

$$\operatorname{Var}(b) = 0$$

cioè che le costanti hanno varianza nulla. (Il lettore si convinca che è una cosa ragionevole.) Scegliendo $a = 1$ invece, si ottiene che

$$\operatorname{Var}(X + b) = \operatorname{Var}(X)$$

ovvero, che sommare una costante non cambia la varianza di una variabile aleatoria. (Si ragioni su questo risultato!) Infine, con $b = 0$, la (4.6.3) diviene

$$\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$$

Definizione 4.6.2. La quantità $\sqrt{\operatorname{Var}(X)}$ è detta *deviazione standard* della variabile aleatoria X.

Osservazione 4.6.1. Proseguendo l'analogia con la statica iniziata con l'Osservazione 4.4.1 di pagina 115, se la media è in termini fisici il baricentro di una distribuzione di masse, la varianza è il suo *momento di inerzia*.

4.7 La covarianza e la varianza della somma di variabili aleatorie

Come abbiamo visto nella Sezione 4.5.1, la media della somma di variabili aleatorie coincide con la somma delle loro medie. Per la varianza questo in generale non è vero. Ad esempio,

$$\begin{aligned} \operatorname{Var}(X + X) &= \operatorname{Var}(2X) \\ &= 2^2 \operatorname{Var}(X) && \text{usando la (4.6.3)} \\ &= 4 \operatorname{Var}(X) \neq \operatorname{Var}(X) + \operatorname{Var}(X) \end{aligned}$$

Vi è tuttavia un caso importante in cui la varianza della somma di due variabili aleatorie è pari alla somma delle loro varianze, ovvero quando le variabili aleatorie sono indipendenti. Prima di dimostrare questo risultato, però dobbiamo definire il concetto di covarianza di due variabili aleatorie.

Definizione 4.7.1. Siano assegnate due variabili aleatorie X e Y di media μ_X e μ_Y rispettivamente. La loro *covarianza*, che si indica con $\operatorname{Cov}(X, Y)$ è (se esiste) la quantità

$$\operatorname{Cov}(X, Y) := E[(X - \mu_X)(Y - \mu_Y)] \tag{4.7.1}$$

Si può ottenere anche una formula alternativa più semplice, analoga alla (4.6.2) per la varianza. Si trova espandendo il prodotto al secondo membro.

$$\begin{aligned} \operatorname{Cov}(X, Y) &= E[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y] \\ &= E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X \mu_Y \\ &= E[XY] - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y \\ &= E[XY] - E[X]E[Y] \end{aligned} \tag{4.7.2}$$

Dalla Definizione 4.7.1 si deducono alcune semplici proprietà, quali la simmetria,

$$\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X) \tag{4.7.3}$$

e il fatto che la covarianza generalizza il concetto di varianza,

$$\mathrm{Cov}(X, X) = \mathrm{Var}(X) \tag{4.7.4}$$

Un'altro enunciato interessante, la cui semplice dimostrazione lasciamo al lettore, è che per ogni costante a

$$\mathrm{Cov}(aX, Y) = a\,\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, aY) \tag{4.7.5}$$

Come la media, la covarianza è additiva, nel senso specificato dalla Proposizione 4.7.2. Premettiamo un risultato parziale in questa direzione.

Lemma 4.7.1. Se X, Y e Z sono variabili aleatorie qualsiasi,

$$\mathrm{Cov}(X + Y, Z) = \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z) \tag{4.7.6}$$

Dimostrazione.

$$\begin{aligned}
\mathrm{Cov}(X+Y, Z) &= E[(X+Y)Z] - E[X+Y]E[Z] && \text{per la (4.7.2)}\\
&= E[XZ + YZ] - \{E[X] + E[Y]\}E[Z]\\
&= E[XZ] - E[X]E[Z] + E[YZ] - E[Y]E[Z]\\
&= \mathrm{Cov}(X, Z) + \mathrm{Cov}(Y, Z) \quad \square
\end{aligned}$$

Il Lemma 4.7.1 può essere facilmente generalizzato a più di due variabili aleatorie (si svolga il Problema 48), ottenendo che, se $X_1, \ldots, X_n$ e Y sono variabili aleatorie qualsiasi,

$$\mathrm{Cov}\left(\sum_{i=1}^{n} X_i,\ Y\right) = \sum_{i=1}^{n} \mathrm{Cov}(X_i, Y) \tag{4.7.7}$$

In questo modo siamo in grado di dimostrare l'enunciato seguente.

Proposizione 4.7.2. Se $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ sono variabili aleatorie qualsiasi,

$$\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(X_i, Y_j) \tag{4.7.8}$$

Dimostrazione.

$$\begin{aligned}
\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) &= \sum_{i=1}^{n} \mathrm{Cov}\left(X_i, \sum_{j=1}^{m} Y_j\right) && \text{per la (4.7.7)}\\
&= \sum_{i=1}^{n} \mathrm{Cov}\left(\sum_{j=1}^{m} Y_j, X_i\right) && \text{per la simmetria, (4.7.3)}\\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(Y_j, X_i) && \text{di nuovo per la (4.7.7)}
\end{aligned}$$

Il risultato richiesto segue allora applicando una seconda volta la proprietà di simmetria data dall'Equazione (4.7.3). $\square$

Utilizzando a questo punto l'Equazione (4.7.4) sulla variabile aleatoria $\sum_i X_i$, si ottiene finalmente la formula che fornisce la varianza di una somma di variabili aleatorie.

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) &= \mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j)\\
&= \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} \mathrm{Cov}(X_i, X_j)
\end{aligned} \tag{4.7.9}$$

Nel caso in cui $n = 2$, la (4.7.9) si riduce a

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + \mathrm{Cov}(X, Y) + \mathrm{Cov}(Y, X)\\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)
\end{aligned} \tag{4.7.10}$$

Teorema 4.7.3. Se X e Y sono variabili aleatorie indipendenti, allora

$$E[XY] = E[X]E[Y] \tag{4.7.11}$$

Questo inoltre implica che

$$\mathrm{Cov}(X, Y) = 0 \tag{4.7.12}$$

e quindi che, se $X_1, \ldots, X_n$ sono indipendenti,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) \tag{4.7.13}$$

Dimostrazione. Proviamo che $E[XY] = E[X]E[Y]$. Se X e Y sono entrambe discrete,

$$\begin{aligned}
E[XY] &= \sum_i \sum_j x_i y_j P(X = x_i, Y = y_j) && \text{per la (4.5.4)}\\
&= \sum_i \sum_j x_i y_j P(X = x_i) P(Y = y_j) && \text{per l'indipendenza}\\
&= \sum_i x_i P(X = x_i) \sum_j y_j P(Y = y_j)\\
&=: E[X]E[Y]
\end{aligned}$$

I casi in cui una o entrambe le variabili aleatorie siano continue si provano in maniera analoga. Che la covarianza di X e Y sia nulla, segue poi dall'Equazione (4.7.2), mentre l'ultima parte dell'enunciato è una conseguenza dell'Equazione (4.7.9). □

Esempio 4.7.1. Si calcoli la varianza della somma di 10 lanci indipendenti di un dado non truccato.

Se denotiamo con X_i il punteggio realizzato dal dado i-esimo, allora grazie all'indipendenza degli X_i e al Teorema 4.7.3, abbiamo che

$$\begin{aligned}\operatorname{Var}\left(\sum_{i=1}^{10} X_i\right) &= \sum_{i=1}^{10} \operatorname{Var}(X_i) \\ &= 10 \cdot \frac{35}{12} \qquad \text{dall'Esempio 4.6.1} \\ &= \frac{175}{6} \quad \square\end{aligned}$$

Esempio 4.7.2. Si determini la varianza del numero di teste su 10 lanci indipendenti di una moneta non truccata.

Sia I_j la funzione indicatrice dell'evento "il lancio j-esimo è testa",

$$I_j := \begin{cases} 1 & \text{se il lancio } j\text{-esimo è testa} \\ 0 & \text{se il lancio } j\text{-esimo è croce} \end{cases}$$

Allora, il numero totale di teste è $\sum_{j=1}^{10} I_j$, e quindi grazie all'indipendenza,

$$\operatorname{Var}\left(\sum_{j=1}^{10} I_j\right) = \sum_{j=1}^{10} \operatorname{Var}(I_j)$$

Siccome I_j è la funzione indicatrice di un evento di probabilità $1/2$, segue dall'Esempio 4.6.2, che la varianza di una singola I_j e della somma di tutte e 10 sono,

$$\operatorname{Var}(I_j) = \frac{1}{2}\left(1 - \frac{1}{2}\right) = \frac{1}{4}$$

$$\operatorname{Var}\left(\sum_{j=1}^{10} I_j\right) = 10 \cdot \frac{1}{4} = \frac{5}{2} \quad \square$$

Se due variabili aleatorie non sono indipendenti, la loro covarianza è un importante indicatore della relazione che sussiste tra loro. Come esempio, si consideri la situazione in cui X e Y sono le funzioni indicatrici di due eventi A e B, ovvero

$$X := \begin{cases} 1 & \text{se } A \text{ si verifica} \\ 0 & \text{altrimenti} \end{cases}, \qquad Y := \begin{cases} 1 & \text{se } B \text{ si verifica} \\ 0 & \text{altrimenti} \end{cases}$$

Si noti intanto che anche XY è una funzione indicatrice:

$$XY = \begin{cases} 1 & \text{se } X = 1, Y = 1 \\ 0 & \text{altrimenti} \end{cases}$$

Si ottiene quindi che

$$\begin{aligned}\operatorname{Cov}(X,Y) &= E[XY] - E[X]E[Y] \\ &= P(X=1, Y=1) - P(X=1)P(Y=1)\end{aligned}$$

da cui deduciamo che

$$\begin{aligned}\operatorname{Cov}(X,Y) > 0 &\Leftrightarrow P(X=1,Y=1) > P(X=1)P(Y=1) \\ &\Leftrightarrow \frac{P(X=1,Y=1)}{P(Y=1)} > P(X=1) \\ &\Leftrightarrow P(X=1|Y=1) > P(X=1)\end{aligned}$$

Perciò la covarianza di X e Y è positiva se condizionando a $\{Y=1\}$, è più probabile che $X=1$ (si noti che vale anche l'enunciato simmetrico).

In generale si può mostrare che un valore positivo di $\operatorname{Cov}(X,Y)$ indica che X e Y tendenzialmente assumono valori grandi o piccoli contemporaneamente. La forza della relazione tra X e Y è misurata più propriamente dal *coefficiente di correlazione lineare*, un numero puro (senza unità di misura) che tiene conto anche delle deviazioni standard di X e Y[8]. Esso si indica con $\operatorname{Corr}(X,Y)$ ed è definito come

$$\operatorname{Corr}(X,Y) := \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} \tag{4.7.14}$$

Si può dimostrare (si svolga il Problema 49), che questa quantità è sempre compresa tra -1 e $+1$.

4.8 La funzione generatrice dei momenti

Definizione 4.8.1. La *funzione generatrice dei momenti*, o più semplicemente *funzione generatrice* ϕ, di una variabile aleatoria X, è definita, per tutti i t reali per i quali il valore atteso di e^{tX} ha senso, dall'espressione

$$\phi(t) := E[e^{tX}] = \begin{cases} \displaystyle\sum_x e^{tx} p(x) & \text{se } X \text{ è discreta} \\ \displaystyle\int_{-\infty}^{\infty} e^{tx} f(x)\,dx & \text{se } X \text{ è continua} \end{cases} \tag{4.8.1}$$

[8] Si noti infatti come la covarianza tra $2X$ e $2Y$ sia sempre molto più forte (quattro volte maggiore, in effetti) di quella tra X e Y. Per il coefficiente di correlazione lineare invece, le due situazioni portano al medesimo valore.

Il nome adottato deriva dal fatto che tutti i momenti di cui è dotata X possono essere ottenuti derivando più volte nell'origine la funzione $\phi(t)$. Ad esempio,

$$\phi'(t) = \frac{d}{dt}E[e^{tX}] = E\left[\frac{d}{dt}e^{tX}\right] = E[Xe^{tX}]$$

da cui $\phi'(0) = E[X]$. Analogamente,

$$\phi''(t) = \frac{d^2}{dt^2}E[e^{tX}] = E\left[\frac{d^2}{dt^2}e^{tX}\right] = E[X^2e^{tX}]$$

da cui $\phi''(0) = E[X^2]$, è il momento secondo di X. Più in generale, la derivata n-esima di $\phi(t)$ calcolata in 0 fornisce il momento n-esimo di X:

$$\phi^{(n)}(0) = E[X^n], \qquad n \geq 1 \tag{4.8.2}$$

Un'altra importante proprietà di ϕ è che la funzione generatrice dei momenti della somma di variabili aleatorie indipendenti è il prodotto delle funzioni generatrici delle singole variabili aleatorie.

Proposizione 4.8.1. Se X e Y sono variabili aleatorie indipendenti con funzioni generatrici ϕ_X e ϕ_Y rispettivamente, e se ϕ_{X+Y} è la funzione generatrice dei momenti di $X + Y$, allora

$$\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t) \tag{4.8.3}$$

Dimostrazione. Si noti intanto che se X e Y sono indipendenti, lo sono anche le variabili aleatorie e^{tX} ed e^{tY}. Infatti per verificare l'Equazione (4.3.12) di pagina 105, occorre mostrare che, comunque si scelgano A e B,

$$P\left(e^{tX} \in A, e^{tY} \in B\right) = P\left(e^{tX} \in A\right)P\left(e^{tY} \in B\right)$$

D'altra parte, se A' è l'insieme formato dai numeri z tali che $e^{tz} \in A$, allora $e^{tX} \in A \Leftrightarrow X \in A'$. Se si definisce analogamente B', si vede che

$$\begin{aligned} P\left(e^{tX} \in A, e^{tY} \in B\right) &= P(X \in A', Y \in B') && \text{per la definizione di } A' \text{ e } B' \\ &= P(X \in A')P(Y \in B') && \text{per l'indipendenza di } X \text{ e } Y \\ &= P\left(e^{tX} \in A\right)P\left(e^{tY} \in B\right) \end{aligned}$$

A questo punto, basta sfruttare il fatto che l'indipendenza implica che la media del prodotto è il prodotto delle medie, per concludere che

$$\begin{aligned} \phi_{X+Y}(t) &:= E[e^{t(X+Y)}] \\ &= E[e^{tX}e^{tY}] \\ &= E[e^{tX}]E[e^{tY}] \\ &= \phi_X(t)\phi_Y(t) \quad \square \end{aligned}$$

Osservazione 4.8.1. Un ulteriore risultato che mostra l'importanza della funzione generatrice dei momenti è che essa *determina la distribuzione*, nel senso che due variabili aleatorie con identica funzione generatrice hanno necessariamente la stessa legge (e quindi la stessa funzione di ripartizione, e la stessa funzione di massa, ovvero la stessa densità).

4.9 La legge debole dei grandi numeri

Cominciamo con un risultato preliminare.

Proposizione 4.9.1 (Disuguaglianza di Markov). Se X è una variabile aleatoria che non è mai negativa, allora per ogni $a > 0$,

$$P(X \geq a) \leq \frac{E[X]}{a} \tag{4.9.1}$$

Dimostrazione. Diamo la dimostrazione nel caso che X sia continua con densità f.

$$\begin{aligned} E[X] &:= \int_0^\infty xf(x)\,dx \\ &= \int_0^a xf(x)\,dx + \int_a^\infty xf(x)\,dx \\ &\geq \int_a^\infty xf(x)\,dx && \text{perché il primo addendo è positivo} \\ &\geq \int_a^\infty af(x)\,dx && \text{perché } x \geq a \text{ nella regione di integrazione} \\ &= a\int_a^\infty f(x)\,dx \\ &= aP(X \geq a) \end{aligned}$$

E l'enunciato segue dividendo entrambi i termini per a. $\square$

Come corollario, ricaviamo la proposizione seguente.

Proposizione 4.9.2 (Disuguaglianza di Chebyshev). Se X è una variabile aleatoria con media μ e varianza σ^2, allora per ogni $r > 0$,

$$P(|X - \mu| \geq r) \leq \frac{\sigma^2}{r^2} \tag{4.9.2}$$

Dimostrazione. Gli eventi $\{|X-\mu| \geq r\}$ e $\{(X-\mu)^2 \geq r^2\}$ coincidono e sono quindi equiprobabili. Visto che $(X-\mu)^2$ è una variabile aleatoria non negativa, possiamo applicarle la disuguaglianza di Markov con $a = r^2$, ottenendo che

$$P(|X-\mu| \geq r) = P\big((X-\mu)^2 \geq r^2\big) \leq \frac{E[(X-\mu)^2]}{r^2} = \frac{\sigma^2}{r^2} \quad \square$$

L'importanza delle disuguaglianze di Markov e di Chebyshev, sta nel fatto che permettono di limitare le probabilità di eventi rari che riguardano variabili aleatorie di cui conosciamo solo la media, oppure la media e la varianza. Naturalmente, quando la distribuzione è nota, tali probabilità possono essere calcolate esattamente e non vi è necessità di ridursi all'utilizzo di maggiorazioni.

Esempio 4.9.1. Il numero di pezzi prodotti da una fabbrica durante una settimana è una variabile aleatoria di media 50. **(a)** Cosa si può dire sulla probabilità che la produzione superi occasionalmente i 75 pezzi? **(b)** Se si suppone nota anche la varianza, pari a 25, cosa si può dire sulla probabilità che la produzione sia compresa tra i 40 e i 60 pezzi?

Denotiamo con X la variabile aleatoria che indica il numero di pezzi prodotti in una settimana. **(a)** Per la disuguaglianza di Markov,

$$P(X \geq 75) \leq \frac{E[X]}{75} = \frac{50}{75} = \frac{2}{3}$$

(b) Applicando la disuguaglianza di Chebyshev,

$$P(|X-50| \geq 10) \leq \frac{25}{10^2} = \frac{1}{4}$$

Quindi

$$P(40 \leq X \leq 60) = P(|X-50| \leq 10) \geq 1 - \frac{1}{4} = \frac{3}{4}$$

Perciò la probabilità che la produzione sia compresa tra i 40 e i 60 pezzi è almeno del 75%. $\square$

Se nella disuguaglianza di Chebyshev si pone $r = k\sigma$, essa assume la forma seguente:

$$P(|X-\mu| > k\sigma) \leq \frac{1}{k^2} \tag{4.9.3}$$

In altri termini, la probabilità che una variabile aleatoria differisca dalla sua media per più di k volte la deviazione standard, non può superare il valore $1/k^2$.

Concludiamo questa sezione provando, grazie alla disuguaglianza di Chebyshev, la legge debole dei grandi numeri, un enunciato che afferma che la media aritmetica di

n copie indipendenti di una variabile aleatoria tende al valore atteso di quest'ultima per n che tende all'infinito. Tale convergenza si precisa dicendo che scelto un ε comunque piccolo, la media aritmetica si discosta dal valore atteso per più di ε con probabilità che tende a zero, quando n tende all'infinito.

Teorema 4.9.3 (Legge debole dei grandi numeri). Sia $X_1, X_2, \ldots$ una successione di variabili aleatorie i.i.d. (indipendenti e identicamente distribuite), tutte con media $E[X_i] =: \mu$. Allora per ogni $\varepsilon > 0$,

$$P\left(\left|\frac{X_1+\cdots+X_n}{n} - \mu\right| > \varepsilon\right) \longrightarrow 0 \qquad \text{quando } n \longrightarrow \infty \tag{4.9.4}$$

Dimostrazione. Proveremo il risultato solo sotto l'ipotesi aggiuntiva che le X_i abbiano varianza finita σ^2. Dalle proprietà di media e varianza segue che

$$E\left[\frac{X_1+\cdots+X_n}{n}\right] = \mu \qquad \text{e} \qquad \operatorname{Var}\left(\frac{X_1+\cdots+X_n}{n}\right) = \frac{\sigma^2}{n}$$

La seconda ad esempio si prova in questo modo:

$$\begin{aligned} \operatorname{Var}\left(\frac{X_1+\cdots+X_n}{n}\right) &= \frac{1}{n^2}\operatorname{Var}(X_1+\cdots+X_n) && \text{per la (4.6.3)} \\ &= \frac{\operatorname{Var}(X_1)+\cdots+\operatorname{Var}(X_n)}{n^2} && \text{per l'indipendenza e il Teorema 4.7.3} \\ &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

Segue allora dalla disuguaglianza di Chebyshev applicata alla variabile aleatoria $(X_1+\cdots+X_n)/n$, che

$$P\left(\left|\frac{X_1+\cdots+X_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}$$

Poiché il secondo membro tende a zero per n che tende all'infinito, l'enunciato è provato. $\square$

Una applicazione di questo teorema è la seguente, che permette anche di giustificare l'interpretazione frequentista della probabilità di un evento. Supponiamo di ripetere in successione molte copie indipendenti di un esperimento, in ciascuna delle quali può verificarsi un certo evento E. Ponendo

$$X_i := \begin{cases} 1 & \text{se } E \text{ si realizza nell'esperimento } i\text{-esimo} \\ 0 & \text{se } E \text{ non si realizza nell'esperimento } i\text{-esimo} \end{cases}$$

la sommatoria $X_1 + X_2 + \cdots + X_n$ rappresenta il numero di prove – tra le prime n – in cui si è verificato l'evento E. Poiché

$$E[X_i] = P(X_i = 1) = P(E)$$

si deduce che la frazione delle n prove nelle quali si realizza E, tende (nel senso della legge debole dei grandi numeri) alla probabilità $P(E)$.

Problemi

1. Si forma la classifica dei punteggi di un gruppo di 10 studenti – 5 studenti maschi e 5 femmine – dopo un esame. Non vi sono ex aequo, e tutte le 10! possibili classifiche diverse hanno pari probabilità. Sia X la migliore posizione ottenuta da una studentessa (ad esempio $X = 2$ se il primo in classifica è maschio e la seconda è femmina). Calcola, per $i = 1, 2, \ldots, 10$, quanto vale $P(X = i)$.

2. Sia X la differenza tra il numero di teste e il numero di croci ottenute in una sequenza di n lanci di una moneta. Quali sono i valori possibili di X?

3. Se nel Problema 2 si suppone che la moneta non sia truccata e si pone $n = 3$, quali sono le probabilità associate ai diversi valori che X può assumere?

4. Supponiamo di disporre della funzione di ripartizione F di una variabile aleatoria X. Come faresti per determinare la probabilità $P(X = 1)$? (*Suggerimento:* Serve il concetto di limite per dare la risposta.)

5. La funzione di ripartizione di X è definita come segue.

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{2} & 0 \le x < 1 \\ \frac{2}{3} & 1 \le x < 2 \\ \frac{11}{12} & 2 \le x < 3 \\ 1 & 3 \le x \end{cases}$$

   (a) Se ne tracci il grafico.

   (b) Quanto vale $P(X > 1/2)$?

   (c) Quanto vale $P(2 < X \le 4)$?

   (d) Quanto vale $P(X < 3)$?

   (e) Quanto vale $P(X = 1)$?

   Per rispondere ai punti (d) e (e) occorre ragionare in modo analogo a quanto fatto nel Problema 4.

6. Supponiamo che il tempo (in ore) di funzionamento ininterrotto di un computer, prima che sia necessario riavviarlo a causa di un crash di sistema sia una variabile aleatoria continua con funzione di densità data da

$$f(x) = \begin{cases} \lambda e^{-x/100} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

   (a) Dopo avere determinato il valore della costante λ, (b) calcola quanto vale la probabilità che il computer funzioni tra le 50 e le 150 ore prima di bloccarsi. (c) Qual è invece la probabilità che funzioni meno di 100 ore?

7. Il tempo di vita in ore di un certo tipo di valvola termoionica (quelle usate per amplificare i segnali nei vecchi impianti stereofonici) è una variabile aleatoria con funzione di densità come segue.

$$f(x) = \begin{cases} 0 & x \le 100 \\ 100 \cdot x^{-2} & x > 100 \end{cases}$$

   Qual è la probabilità che esattamente 2, su 5 esemplari di tali valvole, debbano essere sostituiti nelle prime 150 ore di funzionamento? Si supponga che i 5 eventi: "la valvola i-esima viene sostituita entro 150 ore", per $i = 1, 2, 3, 4, 5$, siano tutti indipendenti.

8. È data una variabile aleatoria X con funzione di densità

$$f(x) = \begin{cases} ce^{-2x} & x > 0 \\ 0 & x \le 0 \end{cases}$$

   (a) Quanto vale c?

   (b) Quanto vale $P(X \ge 2)$?

9. Un gruppo di 5 transistor ne contiene 3 di difettosi. I transistor vengono testati uno alla volta, per vedere quali funzionino e quali no. Denotiamo con N_1 il numero di transistor testati prima di incorrere nel primo pezzo difettoso, e con N_2 il numero di ulteriori pezzi testati per trovare il secondo difettoso. Si scriva la funzione di massa di probabilità congiunta di N_1 e N_2.

10. La densità di probabilità congiunta di X e Y è data da

$$f(x, y) \frac{6}{7}\left(x^2 + \frac{xy}{2}\right), \qquad 0 < x < 1, \qquad 0 < y < 2$$

    (a) Verifica che questa sia effettivamente una densità congiunta valida.

    (b) Calcola la densità di probabilità della variabile aleatoria X.

    (c) Determina $P(X > Y)$.

11. Siano $X_1, X_2, \ldots, X_n$ variabili aleatorie indipendenti e tutte con distribuzione data dalla seguente funzione di ripartizione:

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x \le 1 \\ 1 & 1 < x \end{cases}$$

(Variabili aleatorie siffatte si dicono *uniformi sull'intervallo* $[0, 1]$.) Sia M uguale alla massima tra tutte le X_i,

$$M = \max(X_1, X_2, \ldots, X_n)$$

(a) Dimostra che la funzione di ripartizione di M è data da

$$F_M(x) = x^n, \qquad 0 \le x \le 1$$

(b) Qual è la funzione di densità di M?

12. La densità congiunta di X e Y è data da

$$f(x, y) = \begin{cases} xe^{-(x+y)} & x > 0, y > 0 \\ 0 & \text{altrimenti} \end{cases}$$

(a) Calcola la densità di X.

(b) Calcola la densità di Y.

(c) Le due variabili aleatorie sono indipendenti?

13. La densità congiunta di X e Y è

$$f(x, y) = \begin{cases} 2 & \text{se } 0 < x < y < 1 \\ 0 & \text{altrimenti} \end{cases}$$

(a) Calcola la densità di X.

(b) Calcola la densità di Y.

(c) Le due variabili aleatorie sono indipendenti?

14. Dimostra che, se la densità congiunta di due variabili aleatorie X e Y è il prodotto di un termine che dipende solo da x e uno che dipende solo da y, allora X e Y sono indipendenti. In altri termini devi dimostrare che se

$$f(x, y) = g(x)h(y), \qquad -\infty < x < \infty, \qquad -\infty < y < \infty$$

allora X e Y sono indipendenti.

15. Confronta l'enunciato del Problema 14 con i risultati ottenuti per i Problemi 12 e 13. Sono compatibili?

16. Siano X e Y due variabili aleatorie continue. Si dimostri che

(a) $P(X + Y \le a) = \displaystyle\int_{-\infty}^{\infty} F_X(a - y) f_Y(y)\, dy$

(b) $P(X \le Y) = \displaystyle\int_{-\infty}^{\infty} F_X(y) f_Y(y)\, dy$

dove F_X denota la funzione di ripartizione di X e f_Y la funzione di densità di Y.

17. Quando una corrente I (misurata in ampere) scorre attraverso una resistenza R (misurata in ohm), la potenza dissipata (in watt) è data da $W = I^2R$. Supponiamo che I e R siano variabili aleatorie indipendenti con densità

$$f_I(x) = 6x(1 - x) \qquad 0 \le x \le 1$$
$$f_R(x) = 2x \qquad 0 \le x \le 1$$

Determina la densità di probabilità di W.

18. Nell'Esempio 4.3.2, calcola la funzione di massa di probabilità del numero di figli di una famiglia scelta a caso, condizionata al fatto che abbia due bambine.

***19.** Calcola la densità di probabilità condizionale di X dato Y, per i Problemi 10 e 13.

***20.** Prova che X e Y sono indipendenti se e solo se per ogni x e y,

(a) $p_{X|Y}(x|y) = p_X(x)$ nel caso discreto.

(b) $f_{X|Y}(x|y) = f_X(x)$ nel caso continuo.

21. Calcola il valore atteso della variabile aleatoria X del Problema 1.

22. Calcola il valore atteso della variabile aleatoria X del Problema 3.

23. Ogni sera i diversi meteorologi alla televisione ci danno le loro "probabilità" che il giorno successivo ci sia pioggia. Per giudicare se le loro previsioni siano attendibili, decidiamo di dare dei punteggi come segue: se un meteorologo dice che pioverà con probabilità p, riceve un punteggio di

$$1 - (1 - p)^2 \qquad \text{se pioverà}$$
$$1 - p^2 \qquad \text{se non pioverà}$$

Registriamo i punteggi per un certo periodo di tempo, e alla fine concludiamo che il meteorologo con il più alto punteggio medio sia il più attendibile. Supponiamo però che uno dei concorrenti sia al corrente del metodo di valutazione utilizzato e voglia massimizzare in media il suo punteggio. Se questa persona fosse realmente convinta che pioverà con probabilità p^*, che valore p le converrebbe dichiarare per massimizzare il valore atteso del punteggio che riceverà?

24. Una compagnia di assicurazioni emette una polizza che garantisce che verrà pagata una cifra A, in caso si verifichi un evento E entro l'anno. Se la compagnia stima che questo evento accada entro l'anno con probabilità p, quanto deve fare pagare la polizza al cliente per avere un ricavo il cui valore atteso sia il 10% di A?

25. Il trasporto di 148 alunni di una scuola presso un campo sportivo viene realizzato tramite 4 autobus, sui quali salgono 40, 33, 25 e 50 ragazzini. Si sceglie un alunno a caso, e si denota con X il numero totale di quelli saliti sul suo stesso autobus. Si sceglie poi, indipendentemente, uno dei quattro autisti e si denota con Y il numero totale di alunni saliti sull'autobus da lui portato.

(a) Quale pensi che sarà il maggiore, tra $E[X]$ ed $E[Y]$? Perché?

(b) Calcola $E[X]$ ed $E[Y]$.

26. Due giocatori disputano una serie di partite che termina solo quando uno dei due arriva a vincerne i. Supponiamo che ogni partita venga vinta (indipendentemente dalle altre) dal primo giocatore con probabilità p e dal secondo con probabilità $1-p$. Trova il numero medio di partite disputate se $i=2$. Dimostra poi che questo valore è massimo se si pone $p=1/2$.

27. La funzione di densità di X è data da

$$f(x)=\begin{cases} a+bx^2 & \text{se } 0\le x\le 1\\ 0 & \text{altrimenti}\end{cases}$$

Determina il valore di a e b, sapendo che $E[X]=3/5$.

28. Il tempo di vita di un fusibile è una variabile aleatoria X con funzione di densità

$$f(x)=a^2xe^{-ax},\qquad x\ge 0$$

Calcola il tempo di vita medio.

29. Siano $X_1,X_2,\ldots,X_n$ variabili aleatorie indipendenti, tutte con densità

$$f(x)=\begin{cases} 1 & \text{se } 0<x<1\\ 0 & \text{altrimenti}\end{cases}$$

Calcola $E[\max(X_1,\ldots,X_n)]$ e $E[\min(X_1,\ldots,X_n)]$.

30. Supponiamo che X abbia densità

$$f(x)=\begin{cases} 1 & \text{se } 0<x<1\\ 0 & \text{altrimenti}\end{cases}$$

Calcola il momento n-esimo di X, $E[X^n]$, sia trovando la distribuzione di X^n, sia applicando la Proposizione 4.5.1.

31. Supponiamo che il tempo necessario per riparare un personal computer sia una variabile aleatoria (misurata in ore) la cui densità è data da

$$f(x)=\begin{cases} 1/2 & \text{se } 0<x<2\\ 0 & \text{altrimenti}\end{cases}$$

Il costo del lavoro è variabile: se sono necessarie x ore per la riparazione, il relativo costo è pari a $40+30\sqrt{x}$ dollari. Calcola il valore atteso del costo di una riparazione.

32. Sapendo che $E[X]=2$ e $E[X^2]=8$, calcola

(a) $E[(2+4X)^2]$ e (b) $E[X^2+(X+1)^2]$

33. Da un'urna contenente 17 palline bianche e 23 nere, si estraggono a caso e senza rimessa 10 palline. Sia X il numero di palline bianche estratte. Calcola quanto vale $E[X]$, sfruttando alternativamente i due suggerimenti seguenti:

(a) Definisci delle oppurtune funzioni indicatrici X_i, con $i=1,2,\ldots,10$, in modo tale che X ne sia la somma.

(b) Definisci delle oppurtune funzioni indicatrici Y_i, con $i=1,2,\ldots,17$, in modo tale che X ne sia la somma.

34. Se X è una variabile aleatoria continua con funzione di ripartizione F, la sua *mediana* è quel valore m per cui si ha

$$F(m)=\frac{1}{2}$$

Determina la mediana delle variabili aleatorie definite dalle funzioni densità seguenti.

(a) $f(x)=e^{-x}$, $\quad x\ge 0$

(b) $f(x)=1$, $\quad 0\le x\le 1$

35. La mediana (definita nel Problema 34), come la media è utile per predire il valore di una variabile aleatoria. Nell'Osservazione 4.5.1 abbiamo provato che la media è la migliore previsione di X, in termini di minimizzazione dell'errore quadratico medio; la mediana invece è la migliore se si vuole minimizzare il valore atteso del modulo dell'errore. In altre parole, $E[|X-c|]$ è minimo se per c si sceglie la mediana di X. Dai una dimostrazione di questo risultato, nell'ipotesi che X sia continua, con densità f e funzione di ripartizione F. (*Suggerimento:* verifica che

$$\begin{aligned} E[|X-c|] &= \int_{-\infty}^{\infty}|x-c|f(x)\,dx\\ &= \int_{-\infty}^{c}|x-c|f(x)\,dx+\int_{c}^{\infty}|x-c|f(x)\,dx\\ &= \int_{-\infty}^{c}(c-x)f(x)\,dx+\int_{c}^{\infty}(x-c)f(x)\,dx\\ &= cF(c)-\int_{-\infty}^{c}xf(x)\,dx+\int_{c}^{\infty}xf(x)\,dx-c[1-F(c)]\end{aligned}$$

Poi usa l'analisi per determinare quale valore di c minimizza questa espressione.)

36. Se k è un numero tra 0 e 100, e poniamo $p:=k/100$, il *k-esimo quantile* di una variabile aleatoria con funzione di ripartizione F, è un valore m_p tale che

$$F(m_p)=p$$

Determina m_p in funzione di p, per la variabile aleatoria che ha densità

$$f(x)=2e^{-2x},\qquad x\ge 0$$

37. Una piccola comunità è composta da 100 coppie di coniugi. Se durante un certo periodo di tempo muoiono 50 membri della comunità, qual è il valore atteso del numero di matrimoni intatti alla fine? Assumiamo che il gruppo di 50 persone che muoiono possa essere con pari probabilità uno dei $\binom{200}{50}$ gruppi possibili di 50 persone. (*Suggerimento:* Per $i = 1, 2, \ldots, 100$, sia X_i la funzione indicatrice dell'evento "nessun coniuge della coppia i muore".)

38. Calcola media e varianza del numero di successi su n ripetizioni indipendenti di un esperimento, in ciascuna delle quali si ha un successo con probabilità p. È necessaria l'indipendenza?

39. Supponi che X possa assumere i valori 1, 2, 3 e 4 con pari probabilità. Trova media e varianza di X.

40. Supponi che X possa assumere i valori 1, 2 e 3 con probabilità p_1, p_2 e p_3, e inoltre che $E[X] = 2$. Quali sono i valori di p_1, p_2 e p_3 che massimizzano e minimizzano $\mathrm{Var}(X)$?

41. Calcola media e varianza del numero di teste, in tre lanci di una moneta non truccata.

42. Spiega perché, per ogni variabile aleatoria X,

$$E[X^2] \geq E[X]^2$$

In che casi si ha l'uguaglianza?

43. Una variabile aleatoria X, che rappresenta il peso in once (oz) di un articolo, ha densità di probabilità data da

$$f(z) = \begin{cases} z - 8 & \text{se } 8 \leq x \leq 9 \\ 10 - z & \text{se } 9 < x \leq 10 \\ 0 & \text{altrimenti} \end{cases}$$

(a) Calcola media e varianza di X.

(b) Il produttore vende questi articoli a 2 dollari l'uno, con la garanzia di restituire i soldi a tutti i clienti che ne trovassero uno da meno di 8.25 oz. Il suo costo di produzione in dollari è legato al peso x del pezzo dalla relazione $x/15 + 0.35$. Determina il profitto medio.

44. Supponiamo che la durezza X misurata con il metodo di Rockwell, e la perdita per abrasione Y (in una scala opportuna) di un materiale abbiano densità congiunta seguente.

$$f(u, v) = \begin{cases} u + v & \text{per } 0 \leq u \leq 1, 0 \leq v \leq 1 \\ 0 & \text{altrimenti} \end{cases}$$

(a) Trova le densità marginali di X e Y.

(b) Calcola media e varianza di X.

45. Un tipo di prodotti vengono classificati a seconda dei loro difetti e della fabbrica che li ha prodotti. Sia X_1 il numero (1 o 2) della fabbrica, e sia X_2 il numero di difetti per pezzo (che possono essere da 0 a 3), di un prodotto scelto a caso tra la totalità di quelli esistenti. La tabella seguente riporta la funzione di massa di probabilità congiunta per queste due variabili aleatorie discrete.

		X_2 = 0	1	2	3
X_1	1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{1}{8}$
	2	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{4}$

(a) Trova le distribuzioni marginali di X_1 e X_2.

(b) Calcola media e varianza di entrambe le X_i.

(c) Calcola la covarianza delle due variabili aleatorie.

46. Un macchinario produce pezzi che subiscono uno *screening* completo (vengono ispezionati tutti) prima di essere spediti. Lo strumento di misurazione utilizzato a questo scopo, è tale che la lettura di valori compresi tra 1 e $1\frac{1}{3}$ è difficoltosa, e conseguentemente la densità di probabilità dei valori che sono stati misurati, dopo lo screening è la seguente,

$$f(z) = \begin{cases} kz^2 & \text{per } 0 \leq z \leq 1 \\ 1 & \text{per } 1 < z \leq 1\frac{1}{3} \\ 0 & \text{altrimenti} \end{cases}$$

(a) Trova il valore di k.

(b) Che frazione delle misurazioni cadrà al di fuori della zona di imprecisione (e quindi tra 0 e 1)?

(c) Determina media e varianza di questa variabile aleatoria.

47. Verifica la correttezza dell'Equazione (4.7.5) di pagina 126.

48. Dimostra l'Equazione (4.7.7) a pagina 126, usando l'induzione matematica.

49. Siano X e Y due variabili aleatorie di varianza σ_X^2 e σ_Y^2 rispettivamente. Partendo dal fatto che

$$0 \leq \mathrm{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right)$$

dimostra che $\mathrm{Corr}(X, Y) \geq -1$. Poi, usando la disuguaglianza

$$0 \leq \mathrm{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right)$$

concludi che $-1 \leq \mathrm{Corr}(X, Y) \leq 1$. Infine, sfruttando il fatto che $\mathrm{Var}(Z) = 0$ se e solo se Z è costante, dimostra che se $\mathrm{Corr}(X, Y) = \pm 1$, allora X e Y sono legati da una relazione lineare

$$Y = a + bX$$

dove il segno di b è positivo se la correlazione era $+1$, e negativo se essa era -1.

50. Consideriamo n ripetizioni indipendenti di un esperimento che può risultare nei tre esiti 1, 2 e 3, con probabilità p_1, p_2 e p_3 rispettivamente, dove $p_1+p_2+p_3 = 1$. Per $i = 1, 2, 3$, sia N_i il numero di esperimenti con risultato i. Mostra che $\mathrm{Cov}(N_1, N_2) = -np_1p_2$. Spiega inoltre come mai è intuitivo che tale covarianza sia negativa. (*Suggerimento:* Per $i = 1, 2, \ldots, n$, si ponga

$$X_i := \begin{cases} 1 & \text{se la prova } i\text{-esima ha dato esito 1} \\ 0 & \text{se la prova } i\text{-esima non ha dato esito 1} \end{cases}$$

e analogamente per $j = 1, 2, \ldots, n$, si ponga

$$Y_j := \begin{cases} 1 & \text{se la prova } j\text{-esima ha dato esito 2} \\ 0 & \text{se la prova } j\text{-esima non ha dato esito 2} \end{cases}$$

Mostra che

$$N_1 = \sum_{i=1}^{n} X_i, \qquad N_2 = \sum_{j=1}^{n} Y_j$$

quindi utilizza la Proposizione 4.7.2 e il Teorema 4.7.3.)

51. Nell'Esempio 4.5.6 di pagina 121, calcola $\mathrm{Cov}(X_i, X_j)$ e usa il risultato per dimostrare che $\mathrm{Var}(X) = 1$.

52. Dimostra che se X_1 e X_2 hanno la stessa distribuzione, allora

$$\mathrm{Cov}(X_1 + X_2, X_1 - X_2) = 0$$

Nota che non è necessario supporre che siano indipendenti.

53. Sia X una variabile aleatoria continua con funzione di densità data da

$$f(x) = e^{-x}, \qquad x > 0$$

Calcola la funzione generatrice dei momenti di X e impiegala per determinare valore atteso e varianza di X. Verifica il risultato ottenuto per la media con un calcolo diretto.

54. La funzione di densità di una variabile aleatoria X è

$$f(x) = 1, \qquad 0 < x < 1$$

Determina un'espressione per la funzione $E[e^{tX}]$. Derivala per ottenere $E[X^n]$ e verifica il risultato calcolando i momenti di X in modo usuale.

55. Supponiamo che X sia una variabile aleatoria con media e varianza entrambe uguali a 20. Che si può dire di $P(0 \le X \le 40)$?

56. Dall'esperienza passata, un docente sa che se si sceglie uno studente a caso, il suo punteggio all'esame di fine corso sarà una variabile aleatoria di media 75.

(a) Dai un limite superiore alla probabilità che un punteggio superi gli 85 punti.

Supponiamo ora che sia nota anche la varianza di tale variabile aleatoria, pari a 25.

(b) Cosa si può dire sulla probabilità che uno studente ottenga un punteggio compreso tra 65 e 85?

(c) Quanti studenti devono sostenere l'esame affinché vi sia una probabilità almeno di 0.9 che la media dei punteggi della sessione non disti più di 5 da 75?

5 Modelli di variabili aleatorie

Contenuto

Alcuni tipi di variabili aleatorie compaiono molto frequentemente in natura o negli studi tecnologici. In questo capitolo, presentiamo dei modelli di variabili aleatorie particolari, che sono caratterizzati dalla grande generalità dei campi applicativi nei quali compaiono.

5.1 Variabili aleatorie di Bernoulli e binomiali

Supponiamo che venga realizzata una prova, o un esperimento, il cui esito può essere solo un "successo" o un "fallimento". Se definiamo la variabile aleatoria X in modo che sia $X = 1$ nel primo caso e $X = 0$ nel secondo, la funzione di massa di probabilità di X è data da

$$\begin{aligned} P(X = 0) &= 1 - p \\ P(X = 1) &= p \end{aligned} \tag{5.1.1}$$

dove con p abbiamo indicato la probabilità che l'esperimento registri un "successo". Ovviamente dovrà essere $0 \leq p \leq 1$.

Definizione 5.1.1. Una variabile aleatoria X si dice *di Bernoulli*[1] o *bernoulliana* se la sua funzione di massa di probabilità è del tipo dell'Equazione (5.1.1), per una scelta opportuna del parametro p.

[1] In onore del matematico svizzero Jacques Bernoulli.

In altri termini, una variabile aleatoria è bernoulliana se può assumere solo i valori 0 e 1. Il suo valore atteso è dato da

$$E[X] := 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = p \tag{5.1.2}$$

ed è quindi pari alla probabilità che la variabile aleatoria assuma il valore 1.

Definizione 5.1.2. Supponiamo di realizzare n ripetizioni indipendenti di un esperimento, ciascuna delle quali può concludersi in un "successo" con probabilità p, o in un "fallimento" con probabilità $1 - p$. Se X denota il numero totale di successi, X di dice variabile aleatoria *binomiale* di parametri (n, p).

La funzione di massa di probabilità per una variabile aleatoria binomiale di parametri (n, p) è data da

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}, \qquad i = 0, 1, \ldots, n \tag{5.1.3}$$

dove il coefficiente binomiale:

$$\binom{n}{i} := \frac{n!}{i!(n-i)!}$$

(si veda la Sezione 3.5.1), rappresenta il numero di combinazioni differenti che possiamo ottenere scegliendo i elementi da un insieme di n oggetti.

La correttezza dell'Equazione (5.1.3) può essere verificata nel modo seguente: innanzitutto, fissata una qualunque sequenza di esiti con i successi e $n - i$ fallimenti, la probabilità che si realizzi esattamente tale sequenza è $p^i(1-p)^{n-i}$ per l'indipendenza delleripetizioni. L'Equazione (5.1.3) segue quindi dal contare quante sono le diverse sequenze di esiti con questa caratteristica. Esse sono $\binom{n}{i}$ perché corrispondono a tutti i modi in cui si possono scegliere gli i esperimenti che hanno dato esito positivo sugli n in totale. Perciò, per $n = 5$ e $i = 2$ vi sono $\binom{5}{2} = 10$ scelte possibili, ovvero

$$\begin{array}{lllll} (s,s,f,f,f) & (s,f,s,f,f) & (s,f,f,s,f) & (s,f,f,f,s) & (f,s,s,f,f) \\ (f,s,f,s,f) & (f,s,f,f,s) & (f,f,s,s,f) & (f,f,s,f,s) & (f,f,f,s,s) \end{array}$$

dove, ad esempio, si intende che l'esito (f, s, f, s, f) è quello in cui i due successi si sono verificati nelle prove numero 2 e numero 4. Si noti che la somma delle probabilità di tutti i valori possibili di una variabile aleatoria binomiale, è pari a 1 per

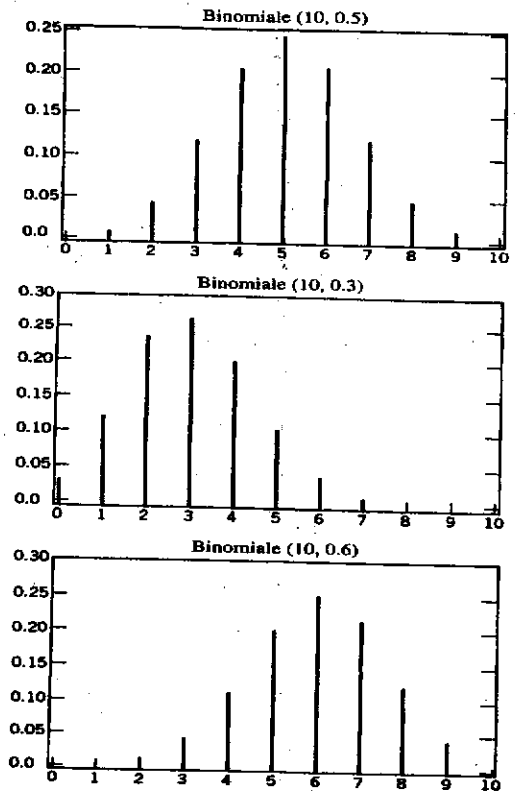


Figura 5.1 La funzione di massa di probabilità per tre variabili aleatorie binomiali.

la formula delle potenze del binomio[2]:

$$\sum_i P(X = i) = \sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = [p + (1-p)]^n = 1$$

Le funzioni di massa per le variabili aleatorie binomiali di parametri $(10, 0.5)$, $(10, 0.3)$ e $(10, 0.6)$ sono rappresentate in Figura 5.1. Si noti come la prima sia simmetrica attorno a 0.5 mentre le altre due pesino di più i valori piccoli o quelli grandi.

[2] Tale formula afferma che

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}$$

Esempio 5.1.1. Una azienda produce dischetti per PC che sono difettosi con probabilità 0.01, indipendentemente l'uno dall'altro. Questi dischetti sono poi venduti in confezioni da 10 pezzi, con la garanzia di rimborso in caso vi sia più di un pezzo difettoso. Che percentuale delle confezioni viene ritornata? Se si comprano tre confezioni, qual è la probabilità di ritornarne esattamente una?

Se X è il numero di pezzi difettosi in una scatola da 10 dischetti, X è una variabile aleatoria binomiale di parametri $(10, 0.01)$. Perciò, assumendo che tutti i clienti che ne hanno la possibilità sfruttino la garanzia, la probabilità che una scatola sia ritornata è pari a

$$\begin{aligned} P(X > 1) &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - \binom{10}{0} \cdot 0.01^0 \cdot 0.99^{10} - \binom{10}{1} \cdot 0.01^1 \cdot 0.99^9 \approx 0.0043 \end{aligned}$$

Poiché ogni scatola – indipendentemente dalle altre – viene resa con probabilità di circa 0.43%, a lungo andare sarà reso circa lo 0.43% delle confezioni. Da quanto detto segue inoltre, che acquistando 3 scatole, il numero di quelle che verranno rese è una variabile aleatoria binomiale di parametri $(3, 0.0043)$, quindi la probabilità richiesta è

$$\binom{3}{1} \cdot 0.0043^1 \cdot 0.9957^2 \approx 0.013 \quad \square$$

Esempio 5.1.2. Supponiamo per semplicità che il colore degli occhi di ogni persona sia determinato da una sola coppia di geni, con il fenotipo "occhi castani" dominante rispetto a quello "occhi azzurri". Ciò significa che un individuo con due geni per gli occhi azzurri presenta occhi azzurri, mentre uno che abbia almeno un gene per gli occhi castani ce li avrà di quel colore (si veda anche il Problema 42 del Capitolo 4). Quando due individui procreano, ciascuno dei figli prende a caso uno dei due geni da ciascuno dei due genitori. Se il figlio maggiore di una coppia di persone con gli occhi castani ha gli occhi azzurri, qual è la probabilità che esattamente 2 degli altri 4 figli (che non comprendono gemelli) abbiano gli occhi azzurri?

Per prima cosa, si noti che poiché il figlio più anziano ha gli occhi azzurri, entrambi i genitori devono necessariamente possedere un gene per gli occhi azzurri e uno per quelli castani. (Si spieghi perché.) La probabilità che un figlio di questa coppia abbia gli occhi azzurri, equivale allora alla probabilità che egli riceva il gene corrispondente da entrambi i genitori, ovvero $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. Siccome ciascuno degli altri 4 figli indipendentemente, ha gli occhi azzurri con probabilità $1/4$, la probabilità richiesta è data da

$$\binom{4}{2} \cdot \left(\frac{1}{4}\right)^2 \cdot \left(\frac{3}{4}\right)^2 = \frac{27}{128} \approx 0.2109 \quad \square$$

Esempio 5.1.3. Un sistema di comunicazione è costituito da n elementi, ciascuno dei quali, indipendentemente, funziona con probabilità p. Affinché l'intero sistema sia in grado di funzionare, almeno la metà dei suoi elementi deve farlo.

(a) Per quali valori di p un sistema a 5 componenti funziona con maggiore probabilità di uno a 3 componenti?

(b) In generale, quando è che un sistema a $2k+1$ componenti si comporta meglio di uno a $2k-1$ componenti?

(a) Siccome il numero di componenti funzionanti è una variabile aleatoria binomiale di parametri (n, p), la probabilità che un sistema a 5 componenti funzioni è data da

$$\sum_{i=3}^{5} \binom{5}{i} p^i (1-p)^{5-i} = \binom{5}{3} p^3 (1-p)^2 + \binom{5}{4} p^4 (1-p) + p^5$$

mentre per un sistema a 3 componenti essa è pari a

$$\sum_{i=2}^{3} \binom{3}{i} p^i (1-p)^{3-i} = \binom{3}{2} p^2 (1-p) + p^3$$

perciò il primo sistema è migliore del secondo se

$$10p^3(1-p)^2 + 5p^4(1-p) + p^5 \geq 3p^2(1-p) + p^3$$

Con un po' di conti, la disuguaglianza precedente si riduce a

$$3p(p-1)^2(2p-1) \geq 0$$

che è soddisfatta se e solo se $p \geq \frac{1}{2}$.

(b) Denotiamo con q_n la probabilità che un sistema di n componenti funzioni. Consideriamo quindi un sistema con $2k+1$ componenti, e sia X il numero di componenti funzionanti tra i primi $2k-1$. Il sistema suddetto funziona **(1)** se $X \geq k+1$; **(2)** se $X = k$ e almeno uno degli ultimi due componenti funziona; **(3)** se $X = k-1$ e entrambi gli ultimi due componenti funzionano. In formule,

$$q_{2k+1} = P(X \geq k+1) + P(X = k)\{1 - (1-p)^2\} + P(X = k-1)p^2$$

Per un sistema di $2k-1$ componenti, d'altra parte,

$$q_{2k-1} = P(X \geq k) = P(X \geq k+1) + P(X = k)$$

da cui, usando anche il fatto che $\binom{2k-1}{k-1} = \binom{2k-1}{k}$,

$$\begin{aligned} q_{2k+1} - q_{2k-1} &= P(X = k-1)p^2 - P(X = k)(1-p)^2 \\ &= \binom{2k-1}{k-1} p^{k-1}(1-p)^k p^2 - \binom{2k-1}{k} p^k (1-p)^{k-1}(1-p)^2 \\ &= \binom{2k-1}{k} p^{k+1}(1-p)^k - \binom{2k-1}{k} p^k (1-p)^{k+1} \\ &= \binom{2k-1}{k} p^k (1-p)^k (2p-1) \end{aligned}$$

Siccome questa grandezza è positiva se e solo se $p \geq \frac{1}{2}$, quest'ultima è precisamente la condizione cercata. □

Esempio 5.1.4. Un produttore di componenti elettronici fabbrica dei chip il 10% dei quali sono difettosi. Se ordiniamo 100 di questi chip e denotiamo con X il numero di quelli difettosi che riceviamo, possiamo affermare che X sia una variabile aleatoria binomiale?

La variabile aleatoria X è binomiale di parametri $(100, 0.1)$ solo se il funzionamento di ciascuno dei 100 chip acquistati è indipendente da tutti gli altri. Se questa sia una assunzione sensata dipende da fattori ulteriori. Ad esempio, se sapessimo che i circuiti prodotti in una singola giornata sono tutti funzionanti o tutti difettosi (e il 90% dei giorni si producono chip funzionanti), e se i 100 chip ordinati fossero stati prodotti nello stesso giorno, X avrebbe funzione di massa data da

$$\begin{aligned} P(X = 100) &= 0.1 \\ P(X = 0) &= 0.9 \end{aligned}$$

e non sarebbe quindi binomiale a causa della mancanza di indipendenza. □

Per come è stata definita la variabile aleatoria binomiale di parametri (n, p) (il numero di esperimenti con esito positivo, su n ripetizioni indipendenti, ciascuna con probabilità di successo p), essa può essere rappresentata come somma di bernoulliane. Più precisamente, se X è binomiale di parametri (n, p), si può scrivere

$$X = \sum_{i=1}^{n} X_i \tag{5.1.4}$$

dove X_i è la funzione indicatrice del successo dell'i-esimo esperimento:

$$X_i := \begin{cases} 1 & \text{se la prova } i\text{-esima ha successo} \\ 0 & \text{altrimenti} \end{cases}$$

È evidente che le X_i sono tutte bernoulliane di parametro p, quindi abbiamo che

$$\begin{aligned} E[X_i] &= p && \text{per la (5.1.2)} \\ E[X_i^2] &= p && \text{infatti } X_i \equiv X_i^2 \\ \mathrm{Var}(X_i) &= E[X_i^2] - E[X_i]^2 \\ &= p - p^2 = p(1-p) \end{aligned}$$

Per quanto riguarda X, poi, dalle proprietà di media e varianza e dalla rappresentazione fornita dall'Equazione (5.1.4), otteniamo che

$$E[X] = \sum_{i=1}^{n} E[X_i] = np \tag{5.1.5}$$

$$\begin{aligned} \mathrm{Var}(X) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) \quad \text{per l'indipendenza delle } X_i \\ &= np(1-p) \end{aligned} \tag{5.1.6}$$

Osservazione 5.1.1. Se X_1 e X_2 sono binomiali di parametri (n_1, p) e (n_2, p) e sono indipendenti, la loro somma $X_1 + X_2$ è binomiale di parametri $(n_1 + n_2, p)$. Questo può essere facilmente dedotto dal fatto che se si effettuano n_1 e poi ancora n_2 prove indipendenti dello stesso esperimento con probabilità di successo p, se X_1 e X_2 rappresentano il numero di successi nelle due *tranche* di prove, $X_1 + X_2$ rappresenta il numero di successi sul totale delle $n_1 + n_2$ prove. È quindi binomiale con i parametri precedentemente citati per costruzione.

5.1.1 Calcolo esplicito della distribuzione binomiale

Supponiamo che X sia binomiale di parametri (n, p). Per potere calcolare operativamente la funzione di ripartizione

$$P(X \leq i) = \sum_{k=0}^{i} \binom{n}{k} p^k (1-p)^{n-k}, \qquad i = 0, 1, \ldots, n$$

o la funzione di massa

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}, \qquad i = 0, 1, \ldots, n$$

è molto utile la seguente relazione tra $P(X = k+1)$ e $P(X = k)$:

$$P(X = k+1) = \frac{p}{1-p} \frac{n-k}{k+1} P(X = k) \tag{5.1.7}$$

la cui dimostrazione è lasciata come esercizio.

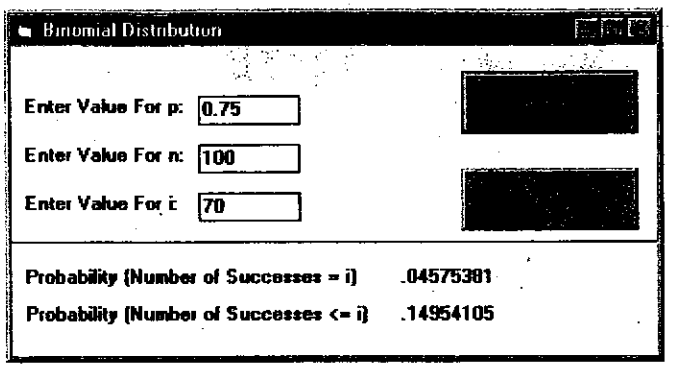


Figura 5.2 La schermata del software per il calcolo della distribuzione binomiale.

Esempio 5.1.5. Sia X una variabile aleatoria binomiale di parametri $n = 6$ e $p = 0.4$. Allora, iniziando da $P(X = 0) = 0.6^6$ e applicando ricorsivamente l'Equazione (5.1.7), si trova

$$
\begin{aligned}
P(X = 0) &= 0.6^6 \approx 0.0467 \\
P(X = 1) &= \tfrac{4}{6} \cdot \tfrac{6}{1} \cdot P(X = 0) \approx 0.1866 \\
P(X = 2) &= \tfrac{4}{6} \cdot \tfrac{5}{2} \cdot P(X = 1) \approx 0.3110 \\
P(X = 3) &= \tfrac{4}{6} \cdot \tfrac{4}{3} \cdot P(X = 2) \approx 0.2765 \\
P(X = 4) &= \tfrac{4}{6} \cdot \tfrac{3}{4} \cdot P(X = 3) \approx 0.1382 \\
P(X = 5) &= \tfrac{4}{6} \cdot \tfrac{2}{5} \cdot P(X = 4) \approx 0.0369 \\
P(X = 6) &= \tfrac{4}{6} \cdot \tfrac{1}{6} \cdot P(X = 5) \approx 0.0041 \quad \square
\end{aligned}
$$

Il Programma 5.1 del pacchetto software abbinato a questo libro (disponibile online) utilizza l'Equazione (5.1.7) per calcolare la distribuzione delle variabili aleatorie binomiali. Il programma accetta in input i parametri n e p e un numero i, e restituisce le probabilità che una binomiale (n, p) sia uguale, oppure minore o uguale, al dato i.

Esempio 5.1.6. Se X è una variabile aleatoria binomiale di parametri $n = 100$ e $p = 0.75$, quanto valgono le probabilità $P(X = 70)$ e $P(X \leq 70)$?

Usando il software, si ottiene la schermata di Figura 5.2 $\square$

5.2 Variabili aleatorie di Poisson

Proseguiamo la panoramica con un'altra importante variabile aleatoria discreta che assume solo valori interi non negativi.

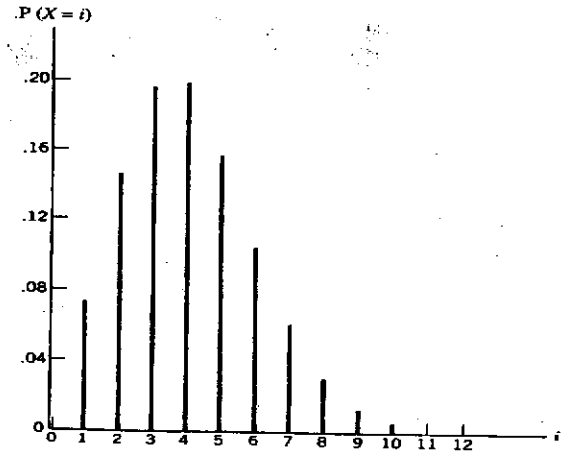


Figura 5.3 La funzione di massa di probabilità della distribuzione di Poisson con parametro $\lambda = 4$.

Definizione 5.2.1. Una variabile aleatoria X che assuma i valori $0, 1, 2, \ldots$, è una variabile aleatoria *di Poisson* o *poissoniana* di parametro λ, $\lambda > 0$, se la sua funzione di massa di probabilità è data da

$$
P(X = i) = \frac{\lambda^i}{i!} e^{-\lambda}, \qquad i = 0, 1, 2, \ldots \tag{5.2.1}
$$

Storicamente, tale distribuzione fu introdotta da Poisson in un libro sulle applicazioni della teoria della probabilità alla risoluzione di cause e processi giudiziari[3].

È immediato verificare che l'Equazione (5.2.1) rappresenta una funzione di massa accettabile, infatti[4]

$$
\sum_{i=0}^{\infty} P(X = i) = e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{-\lambda} e^{\lambda} = 1
$$

La Figura 5.3 mostra il grafico della funzione di massa di una poissoniana con $\lambda = 4$.

---

[3] S. D. Poisson, *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*, 1837.

[4] Si rammenti lo sviluppo in serie di potenze dell'esponenziale: per tutti i numeri y, vale

$$
\sum_{i=0}^{\infty} \frac{y^i}{i!} = e^y
$$

Sia X una variabile aleatoria di Poisson. Per determinarne la media e la varianza, calcoliamo la sua funzione generatrice dei momenti.

$$\begin{aligned}\phi(t) &:= E[e^{tX}] \\ &= \sum_{i=0}^{\infty} e^{ti} P(X=i) \\ &= e^{-\lambda} \sum_{i=0}^{\infty} e^{ti} \frac{\lambda^i}{i!} \\ &= e^{-\lambda} \sum_{i=0}^{\infty} \frac{(\lambda e^t)^i}{i!} \\ &= e^{-\lambda} e^{\lambda e^t} = \exp\{\lambda(e^t-1)\}\end{aligned} \tag{5.2.2}$$

Derivando si trova allora

$$\begin{aligned}\phi'(t) &= \lambda e^t \exp\{\lambda(e^t-1)\} \\ \phi''(t) &= (\lambda e^t)^2 \exp\{\lambda(e^t-1)\} + \lambda e^t \exp\{\lambda(e^t-1)\}\end{aligned}$$

e valutando le due espressioni in $t=0$, si ottiene

$$E[X] = \phi'(0) = \lambda \tag{5.2.3}$$

$$\begin{aligned}\mathrm{Var}(X) &= \phi''(0) - E[X]^2 \\ &= \lambda^2 + \lambda - \lambda^2 = \lambda\end{aligned} \tag{5.2.4}$$

Quindi, sia il valore atteso, sia la varianza delle poissoniane coincidono con il parametro λ.

La variabile aleatoria di Poisson ha un vasto campo di applicazioni, in aree numerose e diverse, anche perché può essere utilizzata come approssimazione di una binomiale di parametri (n,p), quando n è molto grande e p molto piccolo. Per convincerci di questo fatto, sia X una variabile aleatoria binomiale di parametri (n,p), e si ponga $\lambda = np$. Allora

$$\begin{aligned}P(X=i) &= \frac{n!}{(n-i)!i!} p^i (1-p)^{n-i} \\ &= \frac{n(n-1)\cdots(n-i+1)}{i!} \left(\frac{\lambda}{n}\right)^i \left(1-\frac{\lambda}{n}\right)^{n-i} \\ &= \frac{n(n-1)\cdots(n-i+1)}{n^i} \cdot \frac{\lambda^i}{i!} \cdot \frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}\end{aligned}$$

Se si suppone che n sia molto grande e p molto piccolo, valgono le seguenti approssimazioni,

$$\left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda} \qquad \frac{n}{n}\cdot\frac{n-1}{n}\cdots\frac{n-i+1}{n} \approx 1 \qquad \left(1-\frac{\lambda}{n}\right)^i \approx 1$$

E quindi, se n è grande, p piccolo, e $\lambda = np$,

$$P(X=i) \approx \frac{\lambda^i}{i!} e^{-\lambda} \tag{5.2.5}$$

In altri termini, il totale dei "successi" in un gran numero n di ripetizioni indipendenti di un esperimento che ha una piccola probabilità di riuscita p, è una variabile aleatoria con distribuzione approssimativamente di Poisson, con media $\lambda = np$.

Quelli che seguono sono alcuni esempi di variabili aleatorie che seguono con buona approssimazione la legge di Poisson (ovvero che rispettano approssimativamente l'Equazione (5.2.1), per una qualche scelta di λ):

1. Il numero di refusi in una pagina (o un insieme di pagine) di un libro.
2. Il numero di individui, all'interno di una certa categoria di persone, che raggiungono i cento anni di età.
3. La quantità di numeri telefonici errati che vengono composti in una giornata.
4. Il numero di transistor che si guastano nel loro primo giorno di utilizzo.
5. Il numero di clienti che entrano in un ufficio postale nell'arco di una giornata.
6. La quantità di particelle alfa emesse in un periodo di tempo fissato da un campione di materiale radioattivo.

Ciascuna delle variabili aleatorie dei precedenti, come di numerosi altri esempi, è approssimativamente di Poisson per lo stesso motivo – ovvero, perché alcune variabili aleatorie binomiali si possono approssimare con poissoniane. Ad esempio, possiamo supporre che ciascuna lettera tipografata nella pagina di un libro abbia una probabilità p molto piccola di essere sbagliata, e così il numero totale di refusi è circa poissoniano con media $\lambda = np$, dove n è il (presumibilmente elevato) numero di lettere in una pagina di testo. Analogamente, possiamo immaginare che all'interno di una certa categoria di persone, ciascuno indipendentemente dagli altri abbia una piccola probabilità p di superare i cento anni di età, e quindi il numero di individui ai quali capiterà è approssimativamente una variabile aleatoria di Poisson di media $\lambda = np$, dove n è il numero (elevato) di persone di quel gruppo. Lasciamo al lettore interessato di ragionare sul perché le restanti variabili aleatorie degli esempi dal 3 al 6, debbano avere distribuzione approssimativamente poissoniana.

Esempio 5.2.1. Supponendo che il numero medio di incidenti settimanali in un particolare tratto di autostrada sia pari a 3, si vuole calcolare la probabilità che la prossima settimana vi sia almeno un incidente.

Denotiamo con X il numero di incidenti in quel tratto di autostrada nella settimana in esame. Poiché si può ragionevolmente supporre che in una settimana passino un gran numero di autovetture, e che ciascuna abbia una piccola probabilità di essere coinvolta in un incidente, il numero di tali incidenti sarà approssimativamente distribuito come una variabile aleatoria di Poisson di media 3. Quindi

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ &= 1 - \frac{3^0}{0!}e^{-3} \\ &= 1 - e^{-3} \approx 0.9502 \quad \square \end{aligned}$$

Esempio 5.2.2. Un macchinario produce oggetti che hanno una probabilità di essere difettosi pari a 0.1. Supponendo l'indipendenza nella qualità dei pezzi successivi, con che probabilità un campione di 10 oggetti ne conterrà al più uno di difettoso?

Il numero di pezzi difettosi è una variabile aleatoria binomiale di parametri $(10, 0.1)$. La probabilità richiesta è quindi $\binom{10}{0}\cdot 0.1^0 \cdot 0.9^{10} + \binom{10}{1}\cdot 0.1^1 \cdot 0.9^9 \approx 0.7361$. Usando l'approssimazione di Poisson, si ottiene invece,

$$\frac{1^0}{0!}e^{-1} + \frac{1^1}{1!}e^{-1} \approx 0.7358 \quad \square$$

Esempio 5.2.3. Consideriamo un esperimento che consiste nel contare il numero di particelle alfa emesse in un secondo da un grammo di un certo materiale radioattivo. Sappiamo dall'esperienza passata che il valore medio di questa variabile aleatoria è 3.2; qual è una buona approssimazione della probabilità che nell'esperimento in esame non vengano emesse più di 2 particelle?

Se pensiamo alla sorgente come a un numero n (grande) di atomi radioattivi, ciascuno dei quali ha una probabilità di $3.2/n$ (piccola) di emettere una particella alfa in un secondo, ci convinciamo che, con eccellente livello di precisione la variabile aleatoria di interesse si può approssimare con una poissoniana di parametro $\lambda = 3.2$. Quindi la probabilità richiesta è data da

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= e^{-3.2} + \frac{3.2^1}{1!}e^{-3.2} + \frac{3.2^2}{2!}e^{-3.2} \\ &= \left(1 + 3.2 + \frac{3.2^2}{2}\right)e^{-3.2} \approx 0.3799 \quad \square \end{aligned}$$

Esempio 5.2.4. Una compagnia di assicurazioni riceve in media 5 richieste di rimborso al giorno. **(a)** Che frazione delle giornate vedrà arrivare meno di 3 richieste? **(b)** Con che probabilità in una settimana lavorativa di 5 giorni, in esattamente 3 giorni arrivano 4 richieste? Si può assumere l'indipendenza del numero di richieste che arrivano in giorni successivi.

(a) Poiché il numero di assicurati è elevato, ma la probabilità che essi mandino una richiesta di rimborso in un dato giorno è piccola, il numero totale di richieste al giorno, che denotiamo con X, è approssimativamente una poissoniana. Siccome $E[X] = 5$, la probabilità che vi siano meno di 3 richieste in un giorno è data da

$$P(X < 3) = \left(1 + 5 + \frac{5^2}{2}\right)e^{-5} \approx 0.1247$$

Siccome in ciascuna giornata arrivano meno di 3 richieste con probabilità 0.125 circa, a lungo andare, nel 12.5% delle giornate vi saranno meno di 3 richieste.

(b) A causa dell'indipendenza tra le richeste arrivate nei vari giorni, il numero di giorni in una serie di 5, nei quali arriveranno 4 richieste è una variabile aleatoria binomiale Y, di parametri $n = 5$ e $p = P(X = 4)$. Poiché

$$p = P(X = 4) = \frac{5^4}{4!}e^{-5} \approx 0.1755$$

si ottiene che la probabilità cercata è data da

$$P(Y = 3) = \binom{5}{3}(0.1755)^3(0.8245)^2 \approx 0.0367 \quad \square$$

La approssimazione con variabili aleatorie di Poisson è valida anche in condizioni più generali di quelle in cui è stata dimostrata in questa sede. Ad esempio, se si eseguono n esperimenti indipendenti, in cui la probabilità di successo dell'i-esimo è p_i, allora il numero totale di successi è circa una poissoniana di media $\sum_{i=1}^{n} p_i$, anche se le p_i non sono tutte uguali, purché siano tutte piccole, e n sia grande. A volte addirittura, è possibile far cadere la richiesta dell'indipendenza, a patto che tra gli esperimenti vi sia una dipendenza "debole", nel senso spiegato dal prossimo esempio.

Esempio 5.2.5. (Si veda anche l'Esempio 4.5.6 di pagina 121.) A una festa viene organizzato un passatempo; n persone gettano il loro cappello in centro alla stanza, e poi ciascuna ne riprende uno a caso. Denotiamo con X il numero di persone che finisce con il riappropriarsi del suo cappello. Si può dimostrare che se n è grande, X è approssimativamente di Poisson con media 1. Per vedere che quanto affermato è plausibile, poniamo

$$X_i := \begin{cases} 1 & \text{se la persona } i\text{-esima sceglie il proprio cappello} \\ 0 & \text{altrimenti} \end{cases}$$

così da potere esprimere X come $X = X_1 + X_2 + \cdots + X_n$. In questo modo, si può pensare a X come al numero di "successi" su n "prove", dove ovviamente la prova i-esima ha successo se l'i-esimo partecipante finisce con l'impossessarsi del proprio cappello. Siccome i cappelli con i quali può finire sono n, segue che

$$P(X_i = 1) = \frac{1}{n} \tag{5.2.6}$$

Cosa si può dire sull'eventuale indipendenza delle X_i? Consideriamo due indici diversi, i e j: la probabilità condizionata $P(X_i = 1|X_j = 1)$, che la persona i scelga il proprio cappello sapendo che la persona j lo ha fatto, è data da

$$P(X_i = 1|X_j = 1) = \frac{1}{n-1} \tag{5.2.7}$$

infatti $n - 1$ sono i cappelli rimasti disponibili per i quando sappiamo che j ottiene il suo. Confrontando le due Equazioni (5.2.6) e (5.2.7) possiamo notare che X_i e X_j non sono indipendenti (perché altrimenti le equazioni avrebbero avuto il medesimo valore), e tuttavia la dipendenza è abbastanza "debole", soprattutto per n grande, perché $1/n$ e $1/(n-1)$ non sono molto diversi. Non stupisce allora che la distribuzione di X sia approssimativamente di Poisson. Il fatto che $E[X] = 1$ segue poi perché la (5.2.6) implica che $E[X_i] = 1/n$, e da

$$\begin{aligned} E[X] &= E[X_1 + X_2 + \cdots + X_n] \\ &= E[X_1] + E[X_2] + \cdots + E[X_n] \\ &= nE[X_1] = 1 \quad \square \end{aligned}$$

La distribuzione di Poisson è *riproducibile*, nel senso che la somma di due poissoniane indipendenti è ancora una poissoniana. Per dimostrarlo, siano assegnate due variabili aleatorie di Poisson e indipendenti, X_1 e X_2, di parametri rispettivamente λ_1 e λ_2, e calcoliamo la funzione generatrice dei momenti della loro somma:

$$\begin{aligned} \phi_{X_1+X_2}(t) &= \phi_{X_1}(t)\phi_{X_2}(t) && \text{per la Proposizione 4.8.1} \\ &= \exp\{\lambda_1(e^t - 1)\}\exp\{\lambda_2(e^t - 1)\} && \text{per l'Equazione (5.2.2)} \\ &= \exp\{(\lambda_1 + \lambda_2)(e^t - 1)\} \end{aligned}$$

Siccome $\exp\{(\lambda_1 + \lambda_2)(e^t - 1)\}$ è la funzione genetratrice di una poissoniana di media $\lambda_1 + \lambda_2$, e $\phi_{X_1+X_2}$ determina la distribuzione di $X_1 + X_2$ (si veda l'Osservazione 4.8.1), si deduce che $X_1 + X_2$ è una variabile aleatoria di Poisson di media $\lambda_1 + \lambda_2$.

Esempio 5.2.6. Si è stabilito che il numero di apparecchi difettosi prodotti giornalmente da uno stabilimento che assembla impianti stereo, è una variabile aleatoria

di Poisson di media 4. Qual è la probabilità che nell'arco di 2 giorni non vengano prodotti più di 3 stereo difettosi?

Denotiamo con X_1 e X_2 il numero di impianti difettosi prodotti nei due giorni. Nell'ipotesi che queste due variabili aleatorie siano indipendenti, $X_1 + X_2$ è una poissoniana di media 8, e allora

$$P(X_1 + X_2 \leq 3) = \sum_{i=0}^{3} \frac{8^i}{i!} e^{-8} \approx 0.04238 \quad \square$$

5.2.1 Calcolo esplicito della distribuzione di Poisson

Se X è una variabile aleatoria di Poisson di media λ, allora

$$\frac{P(X = i+1)}{P(X = i)} = \frac{\lambda^{i+1}e^{-\lambda}}{(i+1)!}\frac{i!}{\lambda^i e^{-\lambda}} = \frac{\lambda}{i+1} \tag{5.2.8}$$

È possibile utilizzare l'Equazione (5.2.8) ricorsivamente, a partire da $P(X = 0) = e^{-\lambda}$, per calcolare successivamente

$$\begin{aligned} P(X = 1) &= \lambda P(X = 0) \\ P(X = 2) &= \frac{\lambda}{2}P(X = 1) \\ &\cdots \\ P(X = i+1) &= \frac{\lambda}{i+1}P(X = i) \end{aligned}$$

Il Programma 5.2 del software abbinato al libro calcola le probabilità relative alle distribuzioni di Poisson, usando precisamente questa strategia.

5.3 Variabili aleatorie ipergeometriche

Una scatola contiene N batterie accettabili e M difettose. Si estraggono senza rimessa e in maniera casuale n batterie, dando pari probabilità a ciascuno degli $\binom{N+M}{n}$ sottoinsiemi possibili. Se denotiamo con X il numero di batterie accettabili contenute

nel campione estratto, non è difficile convincersi che[5]

$$P(X=i)=\frac{\binom{N}{i}\binom{M}{n-i}}{\binom{N+M}{n}}, \qquad i=0,1,\ldots,n \tag{5.3.1}$$

Definizione 5.3.1. Una variabile aleatoria X che abbia massa di probabilità data dall'Equazione (5.3.1) si dice *ipergeometrica* di parametri N, M e n.

Esempio 5.3.1. Per assemblare un sistema, si prendono a caso 6 componenti da una cassa contenente 20 componenti usati. Il sistema montato funziona solo se tra i 6 componenti impiegati, quelli guasti non sono più di 2. Se nella cassa vi erano 15 componenti efficienti e 5 guasti, qual è la probabilità che il sistema funzioni?

Se diciamo X il numero di componenti funzionanti tra i 6 estratti, X è ipergeometrica di parametri 15, 5 e 6. La probabilità richiesta è quindi

$$\begin{aligned}P(X\geq 4)&=\sum_{i=4}^{6}P(X=i)\\&=\frac{\binom{15}{4}\binom{5}{2}+\binom{15}{5}\binom{5}{1}+\binom{15}{6}\binom{5}{0}}{\binom{20}{6}}\approx 0.8687 \quad \square\end{aligned}$$

Volendo determinare media e varianza di una variabile aleatoria di questo tipo, immaginiamo che le batterie siano estratte una alla volta, e sia

$$X_i:=\begin{cases}1 & \text{se la } i\text{-esima batteria estratta è accettabile}\\ 0 & \text{altrimenti}\end{cases}$$

Siccome se non sappiamo nulla delle altre, la batteria i-esima può essere una qualunque delle $N+M$ disponibili con pari probabilità,

$$P(X_i=1)=\frac{N}{N+M} \tag{5.3.2}$$

D'altronde, se $i\neq j$ ed è noto che la batteria i-esima è accettabile, allora quella j-esima può essere una qualunque delle $N+M-1$ disponibili, di cui $N-1$ sono

[5] Stiamo qui adottando la convenzione che, se $r>m$ oppure $r<0$, allora $\binom{m}{r}=0$, in modo da permettere che alcune delle probabilità $P(X=i)$ siano in effetti nulle. Diversamente dovremmo notare che se $n>N$, X non può assumere i valori da $N+1$ a n (non vi sono batterie efficienti a sufficienza), e analogamente se $n>M$, X non può assumere i valori da 0 a $n-M-1$ (perché non vi sono abbastanza batterie difettose). Per la precisione così i valori possibili per X non vanno da 0 a n, ma da $\max(0,n-M)$ a $\min(n,N)$.

accettabili, per cui

$$\begin{aligned}P(X_i=1,X_j=1)&=P(X_j=1|X_i=1)P(X_i=1)\\&=\frac{N-1}{N+M-1}\cdot\frac{N}{N+M}\end{aligned} \tag{5.3.3}$$

Si può notare che ciascuna delle X_i è una bernoulliana, quindi in particolare,

$$E[X_i]=P(X_i=1)=\frac{N}{N+M} \tag{5.3.4}$$

$$\mathrm{Var}(X_i)=P(X_i=1)P(X_i=0)=\frac{NM}{(N+M)^2} \tag{5.3.5}$$

Utilizziamo a questo punto il fatto che X è la somma delle X_i, per ottenere la sua media:

$$E[X]=E\left[\sum_{i=1}^{n}X_i\right]=\sum_{i=1}^{n}E[X_i]=n\frac{N}{N+M} \tag{5.3.6}$$

Per quanto riguarda la varianza, l'Equazione (4.7.9) di pagina 127, fornisce una formula per il calcolo della varianza della somma di variabili aleatorie anche quando esse non siano indipendenti. Nel nostro caso essa diventa

$$\mathrm{Var}(X)=\sum_{i=1}^{n}\mathrm{Var}(X_i)+2\sum_{j=2}^{n}\sum_{i<j}\mathrm{Cov}(X_i,X_j) \tag{5.3.7}$$

Per determinare il valore del termine $\mathrm{Cov}(X_i,X_j)=E[X_iX_j]-E[X_i]E[X_j]$, ci serve $E[X_iX_j]$. Si noti che X_iX_j è ancora una bernoulliana (infatti può valere solo 0 oppure 1), e quindi

$$\begin{aligned}E[X_iX_j]&=P(X_iX_j=1)\\&=P(X_i=1,X_j=1)\\&=\frac{N(N-1)}{(N+M)(N+M-1)} \qquad \text{per la (5.3.3)}\end{aligned} \tag{5.3.8}$$

da cui si ricava, sostituendo la (5.3.8) e la (5.3.4) e svolgendo i calcoli, che

$$\begin{aligned}\mathrm{Cov}(X_i,X_j)&=\frac{N(N-1)}{(N+M)(N+M-1)}-\left(\frac{N}{N+M}\right)^2\\&=\frac{-NM}{(N+M)^2(N+M-1)}\end{aligned}$$

Sostituendo questo risultato in ciascuno degli $\binom{n}{2}=n(n-1)/2$ addendi della sommatoria doppia nel secondo membro dell'Equazione (5.3.7), e quindi anche la

varianza delle X_i, trovata con la (5.3.5), si ottiene che

$$\begin{aligned}\mathrm{Var}(X) &= \frac{nNM}{(N+M)^2} - \frac{n(n-1)NM}{(N+M)^2(N+M-1)} \\ &= \frac{nNM}{(N+M)^2}\left(1 - \frac{n-1}{N+M-1}\right)\end{aligned} \tag{5.3.9}$$

Ovvero, se indichiamo con p la frazione di batterie efficienti, la (5.3.6) e la (5.3.9) divengono

$$p := \frac{N}{N+M} \quad \Longrightarrow \quad \begin{aligned} E[X] &= np \\ \mathrm{Var}(X) &= np(1-p)\left[1 - \frac{n-1}{N+M-1}\right]\end{aligned}$$

È interessante notare che se si fissa p e si fa tendere $N+M$ all'infinito, $\mathrm{Var}(X)$ tende a $np(1-p)$, che è la varianza di una variabile aleatoria binomiale di parametri (n,p). (Perché questo comportamento non ci deve stupire?)

Esempio 5.3.2. Sia N il numero incognito di animali che popolano una certa regione. Per stimare le dimensioni della popolazione, gli ecologi spesso realizzano il seguente esperimento. Catturano una prima volta un certo numero r di animali, li marcano in qualche modo e li liberano. Dopo avere lasciato passare un tempo sufficiente perché tali esemplari si mischino nuovamente con l'intera popolazione, si esegue una nuova cattura di n animali. Sia X il numero di prede che vengono trovate marcate. Se si accetta che il numero totale di animali non sia cambiato tra le due catture, e che nella seconda ogni animale della popolazione aveva pari probabilità di essere preso, X è una variabile aleatoria ipergeometrica con funzione di massa data da

$$P(X=i) = \frac{\binom{r}{i}\binom{N-r}{n-i}}{\binom{N}{n}}$$

Supponiamo allora che si osservi un valore di X pari a i. Ciò significa che nella seconda cattura, la frazione di animali marcati è stata di i/n. Assumendo che questa sia approssimativamente uguale alla frazione r/N di animali marcati nell'intera popolazione, e risolvendo la semplice proporzione $i : n = r : N$, si ottiene che rn/i è una stima del numero di animali della regione. Quindi, se nella prima battuta si catturano $r = 50$ animali, che vengono marcati e poi liberati, e nella seconda se ne prendono $n = 100$ di cui $X = 25$ marcati, si stima che la popolazione complessiva sia intorno ai 200 esemplari. □

Esiste una relazione tra le variabili aleatorie binomiali e quelle ipergeometriche. Essa ci sarà utile nello sviluppare test statistici che riguardano due popolazioni binomiali.

Esempio 5.3.3. Siano X e Y due variabili aleatorie binomiali e indipendenti, di parametri (n,p) e (m,p) rispettivamente. La funzione di massa di X, condizionata all'evento che $X+Y=k$, è come segue.

$$\begin{aligned} P(X=i|X+Y=k) &= \\ &= \frac{P(X=i, X+Y=k)}{P(X+Y=k)} \\ &= \frac{P(X=i, Y=k-i)}{P(X+Y=k)} && X+Y=k \text{ e } X=i, \text{ quindi } Y=k-i \\ &= \frac{P(X=i)P(Y=k-i)}{P(X+Y=k)} && \text{per l'indipendenza di } X \text{ e } Y \\ &= \frac{\binom{n}{i}p^i(1-p)^{n-i}\binom{m}{k-i}p^{k-i}(1-p)^{m-k+i}}{\binom{n+m}{k}p^k(1-p)^{n+m-k}} && \text{per l'Osservazione 5.1.1} \\ &= \frac{\binom{n}{i}\binom{m}{k-i}}{\binom{n+m}{k}} && \text{si semplifica tutto} \end{aligned}$$

Scopriamo quindi che la distribuzione di X condizionata al valore di $X+Y$ è ipergeometrica.

Questo risultato può anche essere ottenuto con un ragionamento astratto. Supponiamo infatti di eseguire $n+m$ ripetizioni indipendenti di un esperimento che ha probabilità p di avere successo. Siano X i successi nei primi n tentativi e Y quelli nei restanti m. Se sappiamo che il numero totale di successi è stato k, ovvero se condizioniamo all'evento $\{X+Y=k\}$, ciò non modifica l'omogeneità delle prove, che (pur non più indipendenti) hanno tutte le stesse probabilità di avere successo; è quindi intuitivo che ciascun sottoinsieme di k prove abbia la stessa probabilità di costituire l'insieme delle prove riuscite. Le k prove riuscite, sono perciò distribuite come se fossero estratte a caso tra le $n+m$ disponibili. Per questo il numero di prove riuscite che fanno parte delle prime n è una variabile aleatoria ipergeometrica. □

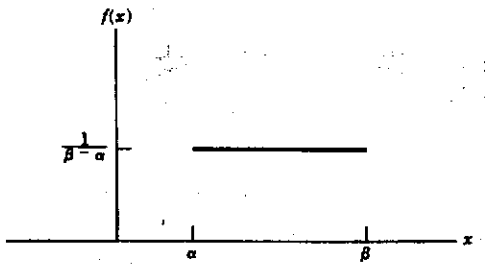


Figura 5.4 Densità di probabilità per una variabile aleatoria uniforme su $[\alpha, \beta]$.

5.4 Variabili aleatorie uniformi

Definizione 5.4.1. Una variabile aleatoria continua si dice *uniforme* sull'intervallo $[\alpha, \beta]$, se ha funzione di densità data da

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha} & \text{se } \alpha \le x \le \beta \\ 0 & \text{altrimenti} \end{cases} \tag{5.4.1}$$

Il grafico di una densità di questo tipo è illustrato in Figura 5.4. Si noti che essa soddisfa le condizioni per essere una densità di probabilità, in quanto

$$\int_{-\infty}^{\infty} f(x)\,dx = \frac{1}{\beta - \alpha}\int_{\alpha}^{\beta} dx = 1$$

Per potere assumere la distribuzione uniforme, nella pratica, occorre che la variabile aleatoria abbia come valori possibili i punti di un intervallo limitato $[\alpha, \beta]$; inoltre si deve poter supporre che essa abbia le stesse probabilità di cadere vicino ad un qualunque punto dell'intervallo.

La probabilità che una variabile aleatoria X, uniforme su $[\alpha, \beta]$, appartenga ad un dato intervallo contenuto in $[\alpha, \beta]$ è pari al rapporto tra le lunghezze dei due intervalli. Infatti, se $[a, b]$ è contenuto in $[\alpha, \beta]$ (si veda la Figura 5.5),

$$P(a < X < b) = \frac{1}{\beta - \alpha}\int_{a}^{b} dx = \frac{b - a}{\beta - \alpha} \tag{5.4.2}$$

Esempio 5.4.1. Sia X una variabile aleatoria uniforme sull'intervallo $[0, 10]$. Si trovino le probabilità che **(a)** $2 < X < 9$, **(b)** $1 < X < 4$, **(c)** $X < 5$, **(d)** $X > 6$.

Le rispettive risposte sono **(a)** $7/10$, **(b)** $3/10$, **(c)** $5/10$, **(d)** $4/10$. □

Esempio 5.4.2. Ad una certa fermata passa un autobus ogni 15 minuti a cominciare dalle 7 (quindi alle 7.00, alle 7.15, alle 7.30, e così via). Se un passeggero arriva alla fermata in un momento casuale con distribuzione uniforme tra le 7 e le 7.30, si calcoli con che probabilità dovrà aspettare il prossimo autobus per **(a)** meno di 5 minuti; **(b)** almeno 12 minuti.

(a) Sia X l'istante (espresso in termini di minuti dopo le 7) in cui questa persona arriva alla fermata. X è ovviamente uniforme sull'intervallo $[0, 30]$. Siccome il passeggero deve aspettare meno di 5 minuti solo se arriva tra le 7.10 e le 7.15, oppure tra le 7.25 e le 7.30, la probabilità richiesta è data da

$$P(10 < X < 15) + P(25 < X < 30) = \frac{5}{30} + \frac{5}{30} = \frac{1}{3}$$

(b) Analogamente, egli deve attendere per almeno 12 minuti se arriva tra le 7 e le 7.03 o tra le 7.15 e le 7.18, quindi la probabilità cercata è pari a

$$P(0 < X < 3) + P(15 < X < 18) = \frac{3}{30} + \frac{3}{30} = \frac{1}{5} \quad \square$$

Determiniamo ora la media di una variabile aleatoria X, uniforme su $[\alpha, \beta]$:

$$\begin{aligned} E[X] &:= \int_{\alpha}^{\beta} \frac{x\,dx}{\beta - \alpha} \\ &= \frac{\beta^2 - \alpha^2}{2(\beta - \alpha)} \\ &= \frac{(\beta - \alpha)(\beta + \alpha)}{2(\beta - \alpha)} = \frac{\alpha + \beta}{2} \end{aligned} \tag{5.4.3}$$

Perciò il valore atteso di una variabile aleatoria uniforme è il punto medio del suo intervallo di definizione, come si poteva intuire direttamente senza fare i calcoli. (Perché?)

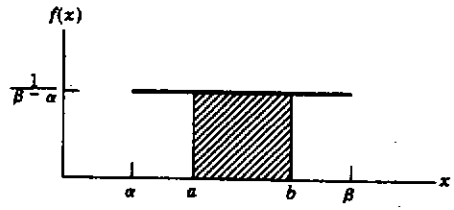


Figura 5.5 La probabilità di un intervallo di valori, per una variabile aleatoria uniforme.

Per ottenere la varianza ci serve il momento secondo.

$$\begin{aligned} E[X^2] &= \int_\alpha^\beta \frac{x^2\,dx}{\beta-\alpha} \\ &= \frac{\beta^3-\alpha^3}{3(\beta-\alpha)} \\ &= \frac{\alpha^2+\alpha\beta+\beta^2}{3} \end{aligned}$$

quindi

$$\begin{aligned} \mathrm{Var}(X) &= \frac{\alpha^2+\alpha\beta+\beta^2}{3} - \left(\frac{\alpha+\beta}{2}\right)^2 \\ &= \frac{\alpha^2-2\alpha\beta+\beta^2}{12} \\ &= \frac{(\beta-\alpha)^2}{12} \end{aligned} \tag{5.4.4}$$

Esempio 5.4.3. La corrente I che attraversa un diodo a semiconduttore è determinata dall'equazione di Shockley

$$I = I_0(e^{aV}-1)$$

dove V è la tensione ai capi del diodo, I_0 la corrente di inversione, ed a è una costante. Si trovi $E[I]$ quando $a=5$, $I_0=10^{-6}$, e V è uniforme sull'intervallo $(1,3)$.

$$\begin{aligned} E[I] &= E[I_0(e^{aV}-1)] \\ &= I_0(E[e^{aV}]-1) && \text{per la linearità di } E \\ &= 10^{-6}\int_1^3 e^{5x}\frac{1}{2}\,dx - 10^{-6} && \text{si noti che } f_V(x)=\frac{1}{2},\ 1<x<3 \\ &= 10^{-7}(e^{15}-e^5)-10^{-6} \approx 0.3269 \quad \square \end{aligned}$$

L'esempio che segue fornisce un'illustrazione di come si possano usare i semplici numeri generati dal calcolatore per simulare esperimenti probabilistici anche complessi. Consideriamo una clinica sperimentale che desidera testare l'efficacia di un nuovo farmaco per ridurre il livello di colesterolo nel sangue. Vengono assunti 1000 volontari che si sottoporranno al test. Per non trascurare la possibilità che il livello di colesterolo durante il periodo di somministrazione possa cambiare per fattori esterni (come i cambiamenti climatici), si decide di dividere i volontari in 2 gruppi di 500: quello di *trattamento*, a cui viene somministrato il farmaco e quello di *controllo*, a cui viene dato un placebo. Sia ai volontari, sia a coloro che somministrano il farmaco non viene rivelata la composizione dei gruppi, per evitare reazioni emotive.

Numeri pseudocasuali generati tramite personal computer

Le variabili aleatorie uniformi su $[0,1]$ rivestono particolare importanza nella pratica, perché sono quelle più direttamente generabili al calcolatore. In effetti la quasi totalità dei sistemi informatici ha delle funzioni interne per generare quelle che, con un buon grado di approssimazione, sono successioni di variabili aleatorie uniformi su $[0,1]$ e indipendenti. La Tabella 5.1 è un esempio di questo genere: presenta un insieme di 240 numeri casuali di questo tipo, generati tramite un comune personal computer. La generazione di variabili aleatorie con il calcolatore è importante in probabilità e statistica, in quanto permette di stimare empiricamente, tramite delle simulazioni, diverse probabilità e valori attesi.

È chiaramente di fondamentale importanza il modo in cui vengono formati i due gruppi. Si desidera infatti che essi siano più simili possibile in tutti gli aspetti tranne la composizione della sostanza somministrata: in questo modo si può senz'altro concludere che ogni differenza significativa nella risposta dei due gruppi sia realmente dovuta al farmaco. Vi è accordo in generale sul fatto che il miglior modo per ottenere questo risultato sia quello di sceglire i 500 volontari di un gruppo in maniera completamente casuale, ovvero la scelta dovrebbe essere fatta in modo che ciascuno dei $\binom{1000}{500}$ sottoinsiemi di 500 volontari abbia la stessa probabilità di essere scelto come gruppo di trattamento. Come si può realizzare questo esperimento casuale?

Esempio 5.4.4 (* Scelta di un sottoinsieme casuale). Consideriamo un insieme di n elementi, numerati con gli interi $1,2,\ldots,n$. Si vuole scegliere a caso uno dei suoi $\binom{n}{k}$ sottoinsiemi di cardinalità k, in modo che abbiano tutti la medesima probabilità di essere selezionati.

Per risolvere questo problema a prima vista complesso, partiamo dalla fine, e supponiamo di avere effettivamente generato nel modo richiesto uno dei sottoinsiemi di k elementi. Per $j=1,2,\ldots,n$, poniamo

$$I_j := \begin{cases} 1 & \text{se l'elemento } j\text{-esimo è nel sottoinsieme} \\ 0 & \text{altrimenti} \end{cases}$$

e calcoliamo la ditribuzione condizionata di I_j dati $I_1, I_2, \ldots, I_{j-1}$. Per prima cosa notiamo che la probabilità che l'elemento 1 stia nel sottoinsieme è k/n (lo si può vedere **(1)** o perché vi è una probabilità di $1/n$ che l'elemento 1 sia il j-esimo elemento estratto, per $j=1,2,\ldots,k$; **(2)** o perché la frazione di esiti della selezione casuale che contengono l'elemento 1 è data da $\binom{1}{1}\binom{n-1}{k-1}/\binom{n}{k} = k/n$). Per questo abbiamo

Tabella 5.1 Numeri casuali generati da un computer

0.6287	0.1304	0.0694	0.2071	0.1494	0.0373	0.6140	0.6661	0.8396	0.8321
0.2878	0.8574	0.1152	0.1937	0.3201	0.4293	0.6524	0.6799	0.0002	0.0125
0.3292	0.6378	0.4862	0.6797	0.2026	0.3157	0.0295	0.9514	0.5085	0.5453
0.4719	0.4071	0.7671	0.5883	0.4498	0.3682	0.5668	0.1206	0.4755	0.8426
0.3353	0.6691	0.0880	0.9331	0.7707	0.1458	0.7114	0.7318	0.9625	0.9029
0.5553	0.2042	0.7008	0.5509	0.2435	0.6768	0.4588	0.0831	0.9798	0.4409
0.1196	0.8310	0.1879	0.8040	0.2126	0.5262	0.4720	0.8021	0.0785	0.8332
0.7614	0.0122	0.2017	0.1074	0.1099	0.4003	0.0623	0.0290	0.9150	0.7234
0.4791	0.4884	0.4062	0.7403	0.6981	0.0029	0.0854	0.6503	0.6172	0.4377
0.2817	0.9549	0.4096	0.5610	0.4150	0.3068	0.0134	0.7427	0.9964	0.3080
0.2380	0.0587	0.1769	0.7661	0.5029	0.7902	0.3543	0.2176	0.0468	0.8749
0.3294	0.8258	0.3312	0.7830	0.7511	0.9578	0.6719	0.9788	0.9245	0.5355
0.2306	0.2980	0.0518	0.1438	0.9940	0.6689	0.1360	0.8925	0.9689	0.3086
0.2136	0.0775	0.4149	0.1647	0.1828	0.2929	0.2119	0.3511	0.4916	0.3354
0.4055	0.5846	0.7221	0.3177	0.3021	0.8223	0.4015	0.4745	0.2977	0.2342
0.3095	0.7528	0.0774	0.5026	0.3785	0.0179	0.4036	0.7699	0.0603	0.2589
0.6763	0.0517	0.5855	0.6920	0.7153	0.8710	0.5628	0.0734	0.6313	0.8521
0.9706	0.5958	0.3707	0.7006	0.9524	0.3181	0.5531	0.5894	0.0241	0.4821
0.5441	0.3833	0.2116	0.8870	0.4703	0.5724	0.0769	0.2379	0.1527	0.6095
0.0204	0.4900	0.1903	0.6979	0.1870	0.5738	0.5360	0.4076	0.9481	0.9872
0.8941	0.5272	0.5608	0.6799	0.2557	0.3492	0.0900	0.4304	0.2744	0.9811
0.3490	0.0688	0.9424	0.3615	0.4435	0.7067	0.6218	0.0370	0.4794	0.3303
0.1105	0.8843	0.6817	0.2674	0.7234	0.3599	0.0001	0.6404	0.4855	0.3589
0.2023	0.7191	0.2734	0.0773	0.8761	0.4052	0.7219	0.4130	0.6764	0.2780

che

$$P(I_1 = 1) = \frac{k}{n} \tag{5.4.5}$$

Calcoliamo adesso la probabilità che l'elemento 2 appartenga al sottoinsieme, condizionata ad I_1. Se $I_1 = 1$, a parte il primo, i restanti $k - 1$ elementi del sottoinsieme vengono scelti a caso tra gli $n - 1$ elementi disponibili dell'insieme di partenza. Perciò in analogia con quanto già detto per l'elemento 1, otteniamo che

$$P(I_2 = 1|I_1 = 1) = \frac{k-1}{n-1} \tag{5.4.6}$$

Similmente, se $I_1 = 0$, allora il primo elemento non appartiene al sottoinsieme, e i k elementi di quest'ultimo vengono scelti a caso tra gli altri $n - 1$ elementi, così che

$$P(I_2 = 1|I_1 = 0) = \frac{k}{n-1} \tag{5.4.7}$$

Mettendo assieme le Equazioni (5.4.6) e (5.4.7), si può dire che

$$P(I_2 = 1|I_1) = \frac{k - I_1}{n-1}$$

e, generalizzando questo procedimento, si arriva a scoprire che

$$P(I_{j+1} = 1|I_1, I_2, \ldots, I_j) = \frac{k - \sum_{i=1}^{j} I_i}{n-j}, \qquad j = 1, \ldots, n \tag{5.4.8}$$

infatti $\sum_{i=1}^{j} I_i$ rappresenta il numero di elementi tra i primi j che appartengono al sottoinsieme, così che condizionando ai valori di $I_1, I_2, \ldots, I_j$, restano $k - \sum_{i=1}^{j} I_i$ elementi del sottoinsieme che devono essere scelti tra gli $n - j$ che rimangono dell'insieme di partenza.

Riconsideriamo il problema dall'inizio. Se U è una variabile aleatoria uniforme su $[0, 1]$, e $0 \le a \le 1$, allora $P(U < a) = a$. Si possono perciò utilizzare le Equazioni (5.4.5) e (5.4.8) per costruire un sottoinsieme casuale con le caratteristiche richieste: si genera una successione $U_1, U_2, \ldots$ di (al più n) variabili aleatorie uniformi su $[0, 1]$ e indipendenti, e quindi si pone

$$I_1 := \begin{cases} 1 & \text{se } U_1 < \dfrac{k}{n} \\ 0 & \text{altrimenti} \end{cases}$$

e, per $j = 1, 2, \ldots$

$$I_{j+1} := \begin{cases} 1 & \text{se } U_{j+1} < \dfrac{k - I_1 - \cdots - I_j}{n-j} \\ 0 & \text{altrimenti} \end{cases}$$

Il procedimento termina non appena $I_1 + \cdots + I_j = k$, e a quel punto il sottoinsieme casuale consiste dei k elementi le cui corrispondenti funzioni indicatrici I sono pari a 1. In formule, $S := \{i : I_i = 1\}$.

Per esemplificare, se $k = 2$ e $n = 5$, il diagramma ad albero della Figura 5.6 illustra la tecnica appena descritta. Il sottoinsieme casuale S è dato dalla posizione finale sull'albero, che viene percorso dalla radice alle foglie, scegliendo ad ogni bivio a seconda del valore di una nuova variabile aleatoria uniforme. Si noti come la probabilità di finire in una qualsiasi delle posizioni finali è sempre pari a $1/10$, come si può vedere moltiplicando le probabilità di muoversi lungo l'albero fino al punto desiderato. Ad esempio, la probabilità di terminare nel punto etichettato $S = \{2, 4\}$ è $P(U_1 > 0.4) \cdot P(U_2 < 0.5) \cdot P(U_3 > \frac{1}{3}) \cdot P(U_4 > \frac{1}{2}) = 0.6 \cdot 0.5 \cdot \frac{2}{3} \cdot \frac{1}{2} = 0.1$. □

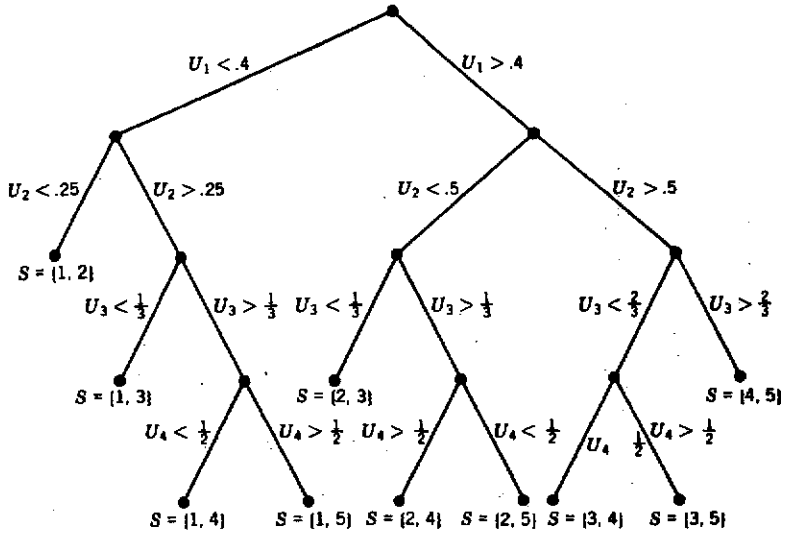


Figura 5.6 Diagramma ad albero per la generazione di un sottoinsieme casuale di 2 elementi, partendo da un insieme di 5. Si noti come il prodotto delle probabilità degli eventi che caratterizzano i rami (dalla radice a una qualsiasi foglia) sia sempre pari a $1/10$.

5.5 Variabili aleatorie normali o gaussiane

Definizione 5.5.1. Una variabile aleatoria X si dice *normale* oppure *gaussiana* di parametri μ e σ^2, e si scrive $X \sim \mathcal{N}(\mu, \sigma^2)$, se X ha funzione di densità data da[6]

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad \forall x \in \mathbb{R} \tag{5.5.1}$$

La densità normale è una curva a campana simmetrica rispetto all'asse $x = \mu$, dove ha il massimo pari a $(\sigma\sqrt{2\pi})^{-1} \approx 0.399/\sigma$ (si veda la Figura 5.7).

La distribuzione normale venne introdotta nel 1733 dal matematico francesce Abraham De Moivre, che la utilizzò per approssimare le probabilità associate a variabili aleatorie binomiali quando il parametro n è grande. Il suo risultato fu poi esteso da Laplace e altri, fino ad essere incluso in un enunciato di teoria della probabilità noto come *teorema del limite centrale* (si veda la Sezione 6.3). Quest'ultimo fornisce la giustificazione teorica di un fatto evidente dall'esperienza empirica, ovvero che molti fenomeni casuali seguono una legge approssimativamente normale.

[6] Per la verifica che questa è una funzione di densità valida, si veda il Problema 29.

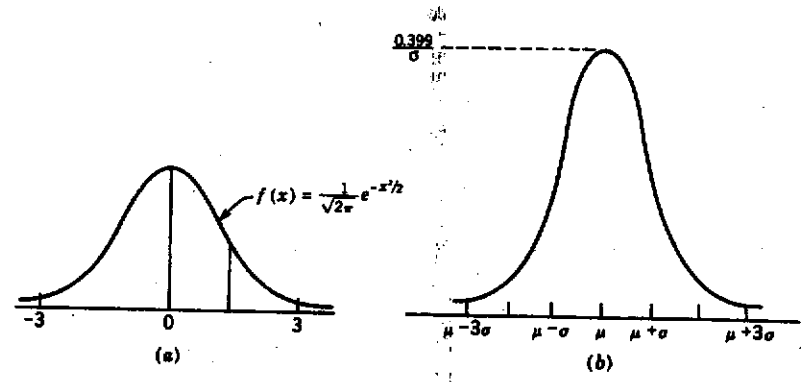


Figura 5.7 Grafici di densità gaussiane, (a) la normale standard, con $\mu = 0$ e $\sigma = 1$ e (b) una generica di parametri μ e σ.

Alcuni esempi di tale comportamento sono la statura delle persone, la velocità in ciascuna direzione di una molecola di gas, gli errori di misurazione delle grandezze fisiche.

La funzione generatrice dei momenti di una variabile aleatoria gaussiana di parametri μ e σ^2 si deduce come segue:

$$\begin{aligned}
\phi(t) &:= E[e^{tX}] \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{tx} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{t(\sigma y+\mu)} e^{-y^2/2}\, dx \qquad \text{ponendo } y = \frac{x-\mu}{\sigma} \\
&= \frac{e^{\mu t}}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left\{\frac{2\sigma t y - y^2}{2}\right\} dx \\
&= \frac{e^{\mu t}}{\sqrt{2\pi}\sigma} e^{\sigma^2 t^2/2} \int_{-\infty}^{\infty} \exp\left\{-\frac{y^2 - 2\sigma t y + \sigma^2 t^2}{2}\right\} dx \\
&= \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y-\sigma t)^2}{2}\right\} dx \\
&= \exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}
\end{aligned} \tag{5.5.2}$$

dove l'ultima uguaglianza segue perché l'espressione dentro l'integrale rappresenta la densità di probabilità di una variabile aleatoria normale di parametri σt e 1, e come tale il suo integrale su tutto $\mathbb{R}$ è pari a 1.

Derivando l'espressione della funzione generatrice data dall'Equazione (5.5.2) si ottiene

$$\phi'(t) = (\mu + \sigma^2 t)\exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}$$

$$\phi''(t) = [\sigma^2 + (\mu + \sigma^2 t)^2]\exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}$$

da cui ricaviamo i primi due momenti e la varianza di una variabile aleatoria gaussiana:

$$E[X] = \phi'(0) = \mu \tag{5.5.3}$$

$$E[X^2] = \phi''(0) = \sigma^2 + \mu^2$$

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \sigma^2 \tag{5.5.4}$$

Così che i parametri μ e σ^2 rappresentano rispettivamente la media e la varianza della distribuzione normale.

Un risultato importante riguardo questo tipo di variabili aleatorie è che se X è gaussiana e Y è una trasformazione lineare di X, allora Y è a sua volta gaussiana. L'enunciato seguente precisa quanto detto.

Proposizione 5.5.1. Sia $X \sim \mathcal{N}(\mu, \sigma^2)$, e sia $Y = \alpha X + \beta$, dove α e β sono due costanti reali e $\alpha \neq 0$. Allora Y è una variabile aleatoria normale con media $\alpha\mu + \beta$ e varianza $\alpha^2\sigma^2$.

Dimostrazione. Calcoliamo la funzione generatrice di Y:

$$\begin{aligned} E[e^{t(\alpha X+\beta)}] &= e^{\beta t}E[e^{\alpha t X}] \\ &= e^{\beta t}\phi_X(\alpha t) \\ &= e^{\beta t}\exp\left\{\mu\alpha t + \frac{\sigma^2\alpha^2 t^2}{2}\right\} \qquad \text{per la (5.5.2)} \\ &= \exp\left\{(\alpha\mu + \beta)t + \frac{(\alpha^2\sigma^2)t^2}{2}\right\} \end{aligned}$$

L'Equazione (5.5.2) afferma che l'espressione ottenuta è la funzione generatrice di una variabile aleatoria gaussiana di media $\alpha\mu + \beta$ e varianza $\alpha^2\sigma^2$. Siccome la funzione generatrice di Y ne determina la distribuzione (si veda l'Osservazione 4.8.1), quanto detto dimostra l'enunciato. □

Un corollario della precedente proposizione è che se $X \sim \mathcal{N}(\mu, \sigma^2)$, allora

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

è una variabile aleatoria normale con media 0 e varianza 1. Una tale variabile aleatoria si dice *normale standard*; la sua funzione di ripartizione riveste un ruolo importante in statistica ed è normalmente indicata con il simbolo Φ:

$$\Phi(x) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-y^2/2}\,dy, \qquad \forall x \in \mathbb{R} \tag{5.5.5}$$

Il fatto che $Z := (X - \mu)/\sigma$ abbia distribuzione normale standard quando X è gaussiana di media μ e varianza σ^2 ci permette di esprimere le probabilità relative a X in termini di probabilità su Z. Ad esempio per trovare $P(X < b)$, notiamo che $X < b$ se e solo se

$$\frac{X - \mu}{\sigma} < \frac{b - \mu}{\sigma}$$

così che

$$\begin{aligned} P(X < b) &= P\left(\frac{X - \mu}{\sigma} < \frac{b - \mu}{\sigma}\right) \\ &= P\left(Z < \frac{b - \mu}{\sigma}\right) \\ &=: \Phi\left(\frac{b - \mu}{\sigma}\right) \end{aligned} \tag{5.5.6}$$

Analogamente, per ogni $a < b$, si ha che

$$\begin{aligned} P(a < X < b) &= P\left(\frac{a - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{b - \mu}{\sigma}\right) \\ &= P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right) \\ &= P\left(Z < \frac{b - \mu}{\sigma}\right) - P\left(Z < \frac{a - \mu}{\sigma}\right) \\ &=: \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right) \end{aligned} \tag{5.5.7}$$

In entrambi i casi ci siamo ricondotti a determinare un valore di $\Phi(x)$. L'integrale dell'Equazione (5.5.5) che definisce questa funzione non si può risolvere analiticamente; è comunque possibile calcolare $\Phi(x)$ usando delle approssimazioni, come i valori tabulati con 4 cifre di precisione in Appendice nella Tabella A.1; in alternativa si può fare approssimare il risultato da un calcolatore, ad esempio usando il Programma 5.5a del software di questo libro.

Nonostante la tabella in Appendice riporti $\Phi(x)$ solo per valori non negativi di x, è possibile ottenere $\Phi(-x)$ usando la simmetria della distribuzione rispetto a 0.

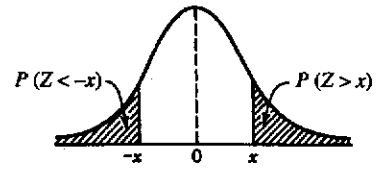


Figura 5.8 Probabilità di eventi simmetrici per una variabile aleatoria normale standard Z.

Infatti, sia $x > 0$ e supponiamo che Z rappresenti una variabile aleatoria normale standard, allora (si veda la Figura 5.8),

$$\begin{aligned}\Phi(-x) &= P(Z < -x)\\ &= P(Z > x) \qquad \text{per simmetria}\\ &= 1 - P(Z < x) = 1 - \Phi(x) \qquad (5.5.8)\end{aligned}$$

Così che ad esempio

$$P(Z < -1) = \Phi(-1) = 1 - \Phi(1) \approx 1 - 0.8413 \approx 0.1587$$

Esempio 5.5.1. Sia X una variabile aleatoria normale con media $\mu = 3$ e varianza $\sigma^2 = 16$. Si trovino **(a)** $P(X < 11)$; **(b)** $P(X > -1)$; **(c)** $P(2 < X < 7)$.

(a) Poniamo al solito $Z := (X - \mu)/\sigma$,

$$\begin{aligned}P(X < 11) &= P\left(\frac{X-3}{4} < \frac{11-3}{4}\right)\\ &= P(Z < 2)\\ &= \Phi(2) \approx 0.9972\end{aligned}$$

(b) In modo del tutto analogo,

$$\begin{aligned}P(X > -1) &= P\left(\frac{X-3}{4} > \frac{-1-3}{4}\right)\\ &= P(Z > -1)\\ &= P(Z < 1)\\ &= \Phi(1) \approx 0.8413\end{aligned}$$

(c) Infine

$$\begin{aligned}P(2 < X < 7) &= P\left(\frac{2-3}{4} < \frac{X-3}{4} < \frac{7-3}{4}\right)\\ &= P(-1/4 < Z < 1)\\ &= \Phi(1) - \Phi(-0.25)\\ &= \Phi(1) - 1 + \Phi(0.25) \approx 0.4400 \quad \square\end{aligned}$$

Esempio 5.5.2. Per trasmettere un messaggio binario ("0", oppure "1") da una sorgente A ad un ricevente B tramite un canale (ad esempio un filo elettrico), si decide di mandare un segnale elettrico di 2 volt se il messaggio era "1" e di -2 volt se il messaggio era "0". A causa dei disturbi nel canale, se A invia il segnale x, $x = \pm 2$, il ricevente B riceve un segnale $R = x + N$, dove la variabile aleatoria N rappresenta il rumore (*noise*) del canale. Alla ricezione di un qualunque segnale R, si decodifica il messaggio con la seguente regola:

$$\begin{aligned}&\text{se } R \geq 0.5, \quad \text{si decodifica "1"}\\ &\text{se } R < 0.5, \quad \text{si decodifica "0"}\end{aligned}$$

Giustificati dal fatto che solitamente il rumore del canale ha distribuzione normale, determiniamo le probabilità di decodificare erroneamente il messaggio nell'ipotesi che $N \sim \mathcal{N}(0, 1)$.

Vi sono due possibili tipi di errore: **(1)** decodificare "0" quando è stato trasmesso "1"; **(2)** decodificare "1" quando è stato trasmesso "0". Il primo si verifica quando il messaggio è "1" e $2 + N < 0.5$, mentre il secondo si verifica quando il messaggio è "0" e $-2 + N \geq 0.5$. Perciò

$$\begin{aligned}P(\text{errore}|\text{il messaggio è "1"}) &= P(N < -1.5)\\ &= 1 - \Phi(1.5) \approx 0.0668\\ P(\text{errore}|\text{il messaggio è "0"}) &= P(N > 2.5)\\ &= 1 - \Phi(2.5) \approx 0.0062 \quad \square\end{aligned}$$

Esempio 5.5.3. La potenza W dissipata da una resistenza è proporzionale al quadrato della differenza di potenziale V ai suoi capi. Ovvero,

$$W = rV^2$$

dove r è una costante. Sia $r = 3$ e si supponiamo che V sia (con buona approssimazione) normale di media 6 e deviazione standard 1. Si trovino **(a)** $E[W]$; **(b)** $P(W > 120)$.

(a) Si usa in maniera inusuale la formula $\text{Var}(V) = E[V^2] - E[V]^2$:

$$\begin{aligned} E[W] &= E[rV^2] \\ &= 3E[V^2] \\ &= 3(\text{Var}(V) + E[V]^2) \\ &= 3(1 + 6^2) = 111 \end{aligned}$$

(b) Di nuovo poniamo $Z := (V - E[V])/\sqrt{\text{Var}(V)} = V - 6$, in modo che sia $Z \sim \mathcal{N}(0,1)$:

$$\begin{aligned} P(W > 120) &= P(rV^2 > 120) \\ &= P(V > \sqrt{40}) \\ &= P\left(\frac{V-6}{1} > \sqrt{40} - 6\right) \\ &\approx P(Z > 0.3246) \\ &= 1 - \Phi(0.3246) \\ &\approx 0.3727 \quad \square \end{aligned}$$

La distribuzione normale è *riproducibile*, nel senso che la somma di variabili aleatorie normali e indipendenti ha essa stessa distribuzione normale. Siano infatti $X_1, X_2, \ldots, X_n$ delle variabili aleatorie normali e indipendenti, dove X_i ha media μ_i e varianza σ_i^2. La funzione generatrice di $\sum_{i=1}^n X_i$ è data da

$$\begin{aligned} \phi(t) &= E[\exp\{tX_1 + tX_2 + \cdots + tX_n\}] \\ &= E[e^{tX_1} e^{tX_2} \ldots e^{tX_n}] \\ &= \prod_{i=1}^n E[e^{tX_i}] && \text{per l'indipendenza} \\ &= \prod_{i=1}^n \exp\left\{\mu_i t + \frac{\sigma_i^2 t^2}{2}\right\} && \text{per la (5.5.2)} \\ &= \exp\left\{\bar{\mu} t + \frac{\bar{\sigma}^2 t^2}{2}\right\} \end{aligned}$$

dove si è posto

$$\bar{\mu} := \sum_{i=1}^n \mu_i, \qquad \bar{\sigma}^2 := \sum_{i=1}^n \sigma_i^2$$

Poiché $\sum_{i=1}^n X_i$ ha la medesima funzione generatrice di una variabile aleatoria $\mathcal{N}(\bar{\mu}, \bar{\sigma}^2)$, e la funzione generatrice determina in maniera univoca la distribuzione, si conclude che $\sum_{i=1}^n X_i$ è gaussiana con media $\sum_{i=1}^n \mu_i$ e varianza $\sum_{i=1}^n \sigma_i^2$.

Esempio 5.5.4. I dati a disposizione dei meteorologi indicano che le precipitazioni annuali a Los Angeles hanno distribuzione normale con media 12.08 pollici e deviazione standard 3.1 pollici. Assumiamo anche che le precipitazioni di anni successivi siano indipendenti.

(a) Si trovi la probabilità che le precipitazioni dei prossimi 2 anni superino complessivamente i 25 pollici.

(b) Si trovi la probabilità che le precipitazioni dell'anno prossimo superino quelle dell'anno successivo per più di 3 pollici.

(a) Sia $Z \sim \mathcal{N}(0,1)$ e siano X_1 e X_2 le precipitazioni dei prossimi due anni. La somma $X_1 + X_2$ è normale con media $2 \times 12.08 = 24.16$ e varianza $2 \times (3.1)^2 = 19.22$. Ne segue che

$$\begin{aligned} P(X_1 + X_2 > 25) &= P\left(\frac{X_1 + X_2 - 24.16}{\sqrt{19.22}} > \frac{25 - 24.16}{\sqrt{19.22}}\right) \\ &\approx P(Z > 0.1916) \approx 0.4240 \end{aligned}$$

(b) Siccome $-X_2$ è gaussiana con media -12.08 e varianza $(-1)^2 \times (3.1)^2$ (per la Proposizione 5.5.1, applicata con $\alpha = -1$ e $\beta = 0$), si ha che $X_1 - X_2$ è gaussiana con media nulla e varianza 19.22. Quindi

$$\begin{aligned} P(X_1 > X_2 + 3) &= P(X_1 - X_2 > 3) \\ &= P\left(\frac{X_1 - X_2}{\sqrt{19.22}} > \frac{3}{\sqrt{19.22}}\right) \\ &\approx P(Z > 0.6843) \approx 0.2469 \end{aligned}$$

Riassumendo, vi è una probabilità del 42.4% che nei prossimi due anni cadano a Los Angeles più di 25 pollici di pioggia, e vi è una probabiltà del 24.69% che le precipitazioni dell'anno prossimo superino quelle dell'anno successivo per almeno 3 pollici. $\square$

Introduciamo ora una notazione che semplificherà molte delle formule per gli intervalli di confidenza del Capitolo 7 e per i test statistici del Capitolo 8. Per ogni $\alpha \in (0,1)$, definiamo il numero z_α in modo che sia

$$P(Z > z_\alpha) = 1 - \Phi(z_\alpha) = \alpha \tag{5.5.9}$$

Ovvero, definiamo $z_\alpha := \Phi^{-1}(1-\alpha)$, in modo che la probabilità che una normale standard assuma un valore maggiore di z_α sia esattamente α (si veda la Figura 5.9).

Il valore di z_α al variare di α può essere ottenuto dalla Tabella A.1. Ad esempio, siccome

$$1 - \Phi(1.645) \approx 0.05 \qquad 1 - \Phi(1.96) \approx 0.025 \qquad 1 - \Phi(2.33) \approx 0.01$$

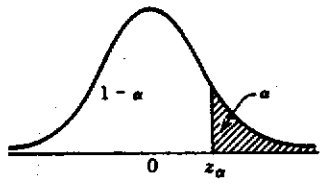


Figura 5.9 Definizione dei quantili gaussiani: $P(Z > z_\alpha) = \alpha$.

si trova immediatamente che

$$z_{0.05} \approx 1.645 \qquad z_{0.025} \approx 1.96 \qquad z_{0.01} \approx 2.33$$

Per calcolare i valori di z_α si può anche impiegare il Programma 5.5b, disponibile online sulla pagina di questo libro, oppure riferirsi all'ultima riga della Tabella A.3, come illustrato più avanti nella nota a pié di pagina 194.

Si noti infine che, prendendo in considerazione il Problema 36 del Capitolo 4, se definiamo il quantile gaussiano k-esimo come quel valore m tale che

$$\Phi(m) = \frac{k}{100}$$

allora posto $k = 100(1 - \alpha)$, si ha che tale quantile è dato da z_α. Il senso di quanto detto è che una gaussiana standard sarà inferiore a z_α nel $k\%$ dei casi.

5.6 Variabili aleatorie esponenziali

Definizione 5.6.1. Una variabile aleatoria continua la cui funzione di densità di probabilità è data da

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{se } x \geq 0 \\ 0 & \text{se } x < 0 \end{cases} \tag{5.6.1}$$

per un opportuno valore della costante $\lambda > 0$, si dice *esponenziale* con parametro (o *intensità*) λ.

La funzione di ripartizione di una tale variabile aleatoria è data da

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= \int_0^x \lambda e^{-\lambda y}\, dy \\ &= 1 - e^{-\lambda x}, \qquad x \geq 0 \end{aligned} \tag{5.6.2}$$

Nella pratica, la distribuzione esponenziale può rappresentare il tempo di attesa prima che si verifichi un certo evento casuale. Ad esempio il tempo che trascorrerà (a partire da questo momento) fino al verificarsi di un terremoto, o allo scoppiare di un nuovo conflitto, o al giungere della prossima telefonata di qualcuno che ha sbagliato numero, sono tutte variabili aleatorie che in pratica tendono ad avere distribuzioni esponenziali (si veda la Sezione 5.6.1 per una spiegazione).

La funzione generatrice dei momenti di una variabile aleatoria esponenziale di intensità λ è data da

$$\begin{aligned} \phi(t) &:= E[e^{tX}] \\ &= \int_0^\infty e^{tx} \lambda e^{-\lambda x}\, dx \\ &= \lambda \int_0^\infty e^{-(\lambda - t)x}\, dx \\ &= \frac{\lambda}{\lambda - t}, \qquad t < \lambda \end{aligned} \tag{5.6.3}$$

Derivando si trova che

$$\phi'(t) = \frac{\lambda}{(\lambda - t)^2}$$

$$\phi''(t) = \frac{2\lambda}{(\lambda - t)^3}$$

e da cui è facile ottenere i primi due momenti e la varianza.

$$E[X] = \phi'(0) = \frac{1}{\lambda} \tag{5.6.4}$$

$$E[X^2] = \phi''(0) = \frac{2}{\lambda^2}$$

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \frac{1}{\lambda^2} \tag{5.6.5}$$

Per una variabile aleatoria esponenziale, λ è il reciproco del valore atteso, e la varianza è il quadrato di quest'ultimo.

La proprietà centrale della distribuzione esponenziale è la sua *assenza di memoria*. Con questa espressione, riferita ad una variabile aleatoria positiva X si intende che

$$P(X > s + t | X > t) = P(X > s) \qquad \forall s, t \geq 0 \tag{5.6.6}$$

Per capire perché l'Equazione (5.6.6) è detta proprietà di assenza di memoria, si immagini che X rappresenti il tempo di vita di un certo oggetto prima di guastarsi. Sapendo che tale oggetto è già in funzione da un tempo t e non si è ancora rotto, qual

è la probabilità che esso continui a funzionare almeno per un ulteriore intervallo di tempo s? Chiaramente la probabilità richiesta è quella espressa dal membro sinistro dell'Equazione (5.6.6), ovvero $P(X > s+t|X > t)$. Infatti dire che l'oggetto non si è ancora guastato al tempo t equivale a dire che il tempo in cui avverrà la rottura (X), è superiore a t, mentre affermare che l'oggetto funzionerà per un ulteriore tempo s a partire dal tempo t, significa che il tempo X dovrà essere maggiore di $t+s$. In questo senso, l'Equazione (5.6.6) afferma che la distribuzione del tempo di vita rimanente dell'oggetto considerato, è la medesima sia nel caso in cui esso stia funzionando da un tempo t, sia nel caso in cui esso sia nuovo, o, in altri termini, se l'Equazione (5.6.6) è soddisfatta, non vi è alcun bisogno di tenere presente l'età dell'oggetto, perché fino a che esso funziona, si comporta esattamente come se fosse "nuovo di zecca".

La condizione di assenza di memoria è equivalente a chiedere che

$$\frac{P(X > s+t, X > t)}{P(X > t)} = P(X > s)$$

e quindi anche a

$$P(X > s+t) = P(X > s)P(X > t)$$

Quest'ultima formulazione è facilmente verificabile se X è esponenziale, visto che, per $x > 0$, $P(X > x) = e^{-\lambda x}$ e ovviamente, $e^{-\lambda(s+t)} = e^{-\lambda s}e^{-\lambda t}$. Abbiamo quindi provato che le variabile aleatorie esponenziali sono prive di memoria. (In realtà è possibile dimostrare che esse sono *le uniche* ad avere questa proprietà.)

Esempio 5.6.1. Supponiamo che il numero di miglia percorse da una automobile prima che la sua batteria sia esausta sia una variabile aleatoria esponenziale di media 10 000 miglia. Se una persona intende intraprendere un viaggio di 5 000 miglia, qual è la probabilità che lo porti a termine senza dovere sostituire la batteria? Cosa si può dire quando la distribuzione non è esponenziale?

La proprietà di assenza di memoria della distribuzione esponenziale implica che il tempo di vita residuo (in migliaia di miglia) della batteria all'inizio del viaggio è esponenziale con intensità $\lambda = 1/10$. La probabilità cercata è data quindi da

$$\begin{aligned} P(\text{vita residua } > 5) &= 1 - F(5) \\ &= e^{-5\lambda} \\ &= e^{-0.5} \approx 0.607 \end{aligned}$$

Se non sapessimo che la distribuzione è esponenziale, la probabilità richiesta sarebbe data da

$$\begin{aligned} P(\text{vita residua } > 5) &= P(\text{vita totale } > t+5 | \text{vita totale } > t) \\ &= \frac{1 - F(t+5)}{1 - F(t)} \end{aligned}$$

dove t è il numero di miglia di funzionamento della batteria fino al momento del viaggio. Perciò, se la distribuzione non è esponenziale, è necessario ottenere nuove informazioni (il valore di t e la forma di F) per potere calcolare la probabilità desiderata. □

L'esempio seguente fornisce un'altra applicazione della proprietà di assenza di memoria.

Esempio 5.6.2. Una squadra di operai ha 3 macchine interscambiabili, e ha bisogno di 2 di esse per potere lavorare. Una volta messa in funzione, ciascuna delle macchine (indipendenntemente dalle altre) sarà efficiente per un tempo esponenziale di parametro λ, prima di guastarsi. Gli operai decidono inizialmente di usare le macchine A e B, tenendo quella denominata C come riserva da mettere in funzione non appena si guasti una delle altre due. In questo modo, saranno in grado di lavorare sino al momento in cui resterà in funzione una macchina sola. Qual è la probabilità che quest'ultima macchina sia la C?

Si può rispondere facilmente alla domanda, senza bisogno di fare alcun calcolo, semplicemente invocando l'assenza di memoria. Consideriamo il momento in cui viene messa in funzione la macchina C. In quell'istante di tempo, esattamente una, tra le macchine A e B, si è guastata e l'altra – chiamiamola 0 – funziona ancora. Nonostante la macchina 0 sia già in funzione da un po', siccome la distribuzione esponenziale non ha memoria, il suo tempo di vita residuo ha la stessa distribuzione di quello di una macchina che venga messa in funzione per la prima volta. Per questo, i tempi di funzionamento residui di 0 e C avranno la stessa distribuzione, e quindi per simmetria, la probabilità che 0 si guasti prima di C è del 50%. □

Una ulteriore utile proprietà della distribuzione esponenziale è enunciata nella seguente proposizione.

Proposizione 5.6.1. Se $X_1, X_2, \ldots, X_n$ sono variabili aleatorie esponenziali e indipendenti, di parametri $\lambda_1, \lambda_2, \ldots, \lambda_n$ rispettivamente, allora la variabile aleatoria $Y := \min(X_1, X_2, \ldots, X_n)$ è esponenziale di parametro $\sum_{i=1}^n \lambda_i$.

Dimostrazione. Basta dimostrare che $P(Y \leq x) = 1 - \exp\{-x\sum_{i=1}^n \lambda_i\}$, ovvero che $P(Y > x) = \exp\{-x\sum_{i=1}^n \lambda_i\}$. Siccome il minore di un insieme di numeri è più grande di x se e solo se ciascuno dei numeri in questione è maggiore di x,

abbiamo che

$$\begin{aligned} P(Y > x) &= P(\min(X_1, X_2, \ldots, X_n) > x) \\ &= P(X_1 > x, X_2 > x, \ldots, X_n > x) \\ &= \prod_{i=1}^{n} P(X_i > x) \qquad \text{per l'indipendenza} \\ &= \prod_{i=1}^{n} (1 - F_{X_i}(x)) \\ &= \prod_{i=1}^{n} e^{-\lambda_i x} \\ &= e^{-x \sum_{i=1}^{n} \lambda_i} \quad \square \end{aligned}$$

Esempio 5.6.3. Un *sistema in serie* è un dispositivo fabbricato in modo tale che il suo corretto funzionamento richiede che tutti i suoi componenti siano efficienti. Consideriamo un sistema di n componenti in serie, tutti indipendenti e ciascuno con tempo di vita esponenziale. Denotiamo con $\lambda_1, \lambda_2, \ldots, \lambda_n$ i rispettivi parametri: qual è la probabilità che il sistema funzioni almeno per un tempo t?

Il tempo di vita del sistema è il minore tra i tempi di vita dei componenti, infatti appena si guasta il primo componente il dispositivo smette di funzionare. Applicando la Proposizione 5.6.1 ti ottiene che

$$P(\text{tempo di vita del sistema } > t) = \exp(-t \textstyle\sum_i \lambda_i) \quad \square$$

Una ulteriore proprietà della distribuzione esponenziale è la seguente: se $c > 0$, e X è esponenziale di intensità λ, allora la variabile aleatoria cX è a sua volta esponenziale, con intensità λ/c. Per dimostrarlo basta scrivere la funzione di ripartizione,

$$\begin{aligned} P(cX \leq x) &= P(X \leq x/c) \\ &= 1 - e^{-\lambda \frac{x}{c}} \\ &= 1 - e^{-\frac{\lambda}{c} x} \end{aligned}$$

5.6.1 * Il processo di Poisson

In questa sezione costruiamo un primo esempio di *processo stocastico*, ovvero una famiglia di variabili aleatorie (non necessariamente indipendenti) parametrizzata da qualche indice (in questo caso, un tempo t).

Definizione 5.6.2. Consideriamo una serie di "eventi" istantanei che avvengono a intervalli di tempo casuali, e sia $N(t)$ il numero di quanti se ne sono verificati nell'intervallo di tempo $[0, t]$. $N(t)$ si dice *processo di Poisson* di intensità λ, $\lambda > 0$, se

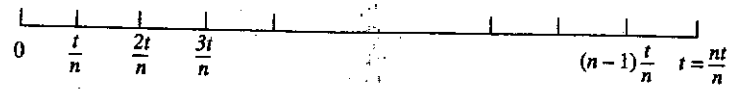


Figura 5.10

1. $N(0) = 0$.
2. Il numero degli eventi che hanno luogo in intervalli di tempo disgiunti son indipendenti.
3. La distribuzione del numero di eventi che si verifica in un dato intervallo di tempo dipende solo dalla lunghezza dell'intervallo, e non dalla sua posizione
4. $\lim_{h \to 0} \frac{P(N(h) = 1)}{h} = \lambda$.
5. $\lim_{h \to 0} \frac{P(N(h) \geq 2)}{h} = 0$.

La condizione 1 stabilisce che si iniziano a contare gli eventi dal tempo 0. La condizione 2 – la *indipendenza degli incrementi* – afferma ad esempio che il numero di eventi fino al tempo t [ovvero $N(t)$] è indipendente dal numero di eventi tra il tempo t e il tempo $t + s$ [ovvero $N(t + s) - N(t)$]. La condizione 3 – la *stazionarietà degli incrementi* – dice che la distribuzione di $N(t + s) - N(t)$ è la stessa per tutti i valori di t. Le condizioni 4 e 5, infine, affermano che se si considera un intervallo di tempo molto piccolo (sia h la sua lunghezza), vi è approssimativamente una probabilità λh che vi occorra un evento solo, e circa una probabilità nulla che se ne verifichino due o più.

Con queste sole ipotesi (qualitative e del tutto sensate) è possibile dimostrare un fatto quantitativo molto preciso, ovvero che il numero di eventi che si verificano in un qualsiasi intervallo di tempo lunghezza t è una variabile aleatoria di Poisson di media λt. Diamo di seguito uno *sketch* di una possibile dimostrazione.

Consideriamo il numero $N(t)$ di eventi che si presentano nell'intervallo $[0, t]$. Vorremmo ottenere una espressione per $P(N(t) = k)$; procediamo dividendo $[0, t]$ in n sottointervalli adiacenti di lunghezza t/n, come in Figura 5.10, con l'intenzione di fare tendere n all'infinito. L'evento $\{N(t) = k\}$ può verificarsi in due modi: **(1)** o vi sono k sottointervalli con un evento ciascuno e gli altri $n - k$ non ne contengono; **(2)** o $N(t) = k$, e almeno un sottointervallo contiene 2 o più degli eventi. Le due possibilità sono mutuamente esclusive, e se n è molto grande, in modo che i sottintervalli siano molto piccoli, la probabilità della seconda possibilità è prossima a zero, per la condizione 5. Quindi se n è grande

$$P(N(t) = k) \approx P(k \text{ sottointervalli con 1 evento}, n - k \text{ con 0 eventi})$$

Sempre per n grande, la condizione 4 e le condizioni 4 e 5 insieme implicano che

$$P(\text{1 evento in un sottointervallo fissato}) \approx \frac{\lambda t}{n}$$

$$P(\text{0 eventi in un sottointervallo fissato}) \approx 1 - \frac{\lambda t}{n}$$

Quindi, usando l'indipendenza data dalla condizione 2, il numero totale di eventi è assimilabile ad una variabile aleatoria binomiale, se si trascura la possibilità che se ne verifichino due o più in un solo sottointervallo,

$$P(k \text{ sottointervalli con 1 evento}, n-k \text{ con 0 eventi}) \approx \binom{n}{k}\left(\frac{\lambda t}{n}\right)^k\left(1-\frac{\lambda t}{n}\right)^{n-k}$$

Se si fa tendere n all'infinito, tale distribuzione può essere approssimata con quella di Poisson di media λt; infatti i parametri della binomiale sono n e $p := \lambda t/n$, che tendono all'infinito e a zero rispettivamente, e il cui prodotto è uguale a λt per ogni n. (Si veda l'Equazione (5.2.5), sull'approssimazione di distribuzioni binomiali con poissoniane.) Si ottiene di conseguenza:

$$P(N(t) = k) \approx \frac{(\lambda t)^k}{k!}e^{-\lambda t}$$

Il risultato trovato non dipende più da n, inoltre le approssimazioni fatte divengono esatte al limite, quindi il simbolo $\approx$ può essere sostituito dall'uguale nell'equazione precedente. Abbiamo quindi mostrato che:

Proposizione 5.6.2. Se $N(t)$ è un processo di Poisson di intensità λ, allora

$$P(N(t) = k) = \frac{(\lambda t)^k}{k!}e^{-\lambda t}, \qquad k = 0, 1, 2, \ldots \tag{5.6.7}$$

Ovvero, il numero di eventi che si verificano in $[0, t]$, come in un qualsiasi altro intervallo di tempo di lunghezza t, ha distribuzione di Poisson di media λt.

Sia X_1 l'istante di tempo in cui si realizza il primo evento, e siano $X_2, X_3, \ldots$ gli intervalli di tempo che intercorrono tra il primo evento e il secondo, tra il secondo e il terzo, e così via. (Quindi ad esempio, se $X_1 = 5$ e $X_2 = 8$, il primo evento avviene all'istante 5 e il secondo all'istante 13.)

Vogliamo determinare la distribuzione delle X_i. L'evento (nel senso probabilistico) $\{X_1 > t\}$ si verifica se e soltanto se nell'intervallo $[0, t]$ non si sono realizzati eventi (nel senso del processo di Poisson). Quindi

$$P(X_1 > t) = P(N(t) = 0) = e^{-\lambda t}$$

Questo significa che $F_{X_1}(t) := P(X_1 \le t) = 1 - e^{-\lambda t}$, e quindi X_1 è una variabile aleatoria esponenziale di intensità λ. Per trovare la distribuzione di X_2, si noti che qualunque valore s assuma la variabile aleatoria X_1, la probabilità condizionale $P(X_2 > t | X_1 = s)$, nel senso dato a pagina 110 con l'Equazione (4.3.23), è sempre data da

$$\begin{aligned} P(X_2 > t | X_1 = s) &= P(\text{0 eventi in } (s, s+t] | X_1 = s) \\ &= P(\text{0 eventi in } (s, s+t]) && \text{per la condizione 2} \\ &= e^{-\lambda t} && \text{per la Proposizione 5.6.2} \end{aligned}$$

Siccome $P(X_2 > t | X_1 = s)$ non dipende da s, dovrà essere uguale a $P(X_2 > t)$. Questo prova sia che la variabile aleatoria X_2 è esponenziale di intensità λ, sia la sua indipendenza da X_1. Ragionando analogamente per $X_3, X_4, \ldots$ si dimostra il seguente enunciato:

Proposizione 5.6.3. I tempi che separano gli eventi di un processo di Poisson di intensità λ sono una successione di variabili aleatorie esponenziali di intensità λ e tra loro indipendenti.

5.7 * Variabili aleatorie di tipo Gamma

Definizione 5.7.1. Una variabile aleatoria continua si dice avere distribuzione *di tipo gamma* di parametri (α, λ), con $\alpha > 0$ e $\lambda > 0$, se la sua funzione di densità di probabilità è data da

$$f(x) = \begin{cases} \dfrac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & \text{se } x > 0 \\ 0 & \text{se } x \le 0 \end{cases} \tag{5.7.1}$$

dove $\Gamma(\,\cdot\,)$ denota la funzione *gamma di Eulero*, che è definita in modo da normalizzare l'integrale di f:

$$\begin{aligned} \Gamma(\alpha) &:= \int_0^\infty \lambda^\alpha x^{\alpha-1} e^{-\lambda x}\, dx \\ &= \int_0^\infty y^{\alpha-1} e^{-y}\, dy \qquad \text{ponendo } y = \lambda x \end{aligned} \tag{5.7.2}$$

La funzione gamma ha un'importante proprietà. Usando la formula di integrazione per parti, se $\alpha > 1$, si può scrivere

$$\begin{aligned} \int_0^\infty y^{\alpha-1} e^{-y}\, dy &= -y^{\alpha-1} e^{-y}\Big|_{y=0}^\infty + \int_0^\infty (\alpha-1) y^{\alpha-2} e^{-y}\, dy \\ &= (\alpha-1) \int_0^\infty y^{\alpha-2} e^{-y}\, dy \end{aligned}$$

dove il termine $-y^{\alpha-1}e^{-y}\Big|_{y=0}^{\infty}$ è nullo perché $\alpha > 1$ implica che $\lim_{y\to 0} y^{\alpha-1} = 0$.

Abbiamo quindi dimostrato che

$$\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1) \tag{5.7.3}$$

Questa proprietà permette di calcolare per induzione il valore che la funzione gamma assume sugli interi, infatti

$$\Gamma(1) = \int_0^\infty e^{-y}\,dy = 1$$

e, per $n \geq 1$,

$$\begin{aligned}\Gamma(n) &= (n-1)\Gamma(n-1)\\ &= (n-1)(n-2)\Gamma(n-2)\\ &\cdots\\ &= (n-1)!\Gamma(1)\end{aligned}$$

Da cui

$$\Gamma(n) = (n-1)! \tag{5.7.4}$$

Si noti che per $\alpha = 1$ la distribuzione gamma coincide con quella esponenziale.

La funzione generatrice dei momenti di una variabile aleatoria X di tipo gamma con parametri (α, λ) si ottiene come segue:

$$\begin{aligned}\phi(t) &:= E[e^{tX}]\\ &= \int_0^\infty e^{tx}\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}\,dx\\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha-1}e^{-(\lambda-t)x}\,dx\\ &= \left(\frac{\lambda}{\lambda-t}\right)^\alpha\frac{1}{\Gamma(\alpha)}\int_0^\infty y^{\alpha-1}e^{-y}\,dy \qquad \text{ponendo } y=(\lambda-t)x\\ &= \left(\frac{\lambda}{\lambda-t}\right)^\alpha \qquad \text{per la (5.7.2)}\end{aligned} \tag{5.7.5}$$

Derivando la funzione generatrice si ottiene che

$$\phi'(t) = \frac{\alpha\lambda^\alpha}{(\lambda-t)^{\alpha+1}}$$

$$\phi''(t) = \frac{\alpha(\alpha+1)\lambda^\alpha}{(\lambda-t)^{\alpha+2}}$$

e quindi

$$E[X] = \phi'(0) = \frac{\alpha}{\lambda} \tag{5.7.6}$$

$$E[X^2] = \phi''(0) = \frac{\alpha(\alpha+1)}{\lambda^2}$$

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = \frac{\alpha}{\lambda^2} \tag{5.7.7}$$

Come altre distribuzioni che abbiamo studiato, anche le gamma, se si fissa λ, sono riproducibili. In particolare se X_1 e X_2 sono due variabili aleatorie gamma indipendenti, di parametri rispettivamente (α_1, λ) e (α_2, λ), allora $X_1 + X_2$ è una gamma di parametri $(\alpha_1+\alpha_2, \lambda)$. Ciò può essere desunto dal calcolo della funzione generatrice:

$$\begin{aligned}\phi(t) &= E[e^{t(X_1+X_2)}]\\ &= E[e^{tX_1}e^{tX_2}]\\ &= E[e^{tX_1}]E[e^{tX_2}] \qquad \text{per l'indipendenza}\\ &= \left(\frac{\lambda}{\lambda-t}\right)^{\alpha_1}\left(\frac{\lambda}{\lambda-t}\right)^{\alpha_2} \qquad \text{per la (5.7.5)}\\ &= \left(\frac{\lambda}{\lambda-t}\right)^{\alpha_1+\alpha_2}\end{aligned}$$

La funzione generatrice trovata concide con quella di una distribuzione gamma di parametri $(\alpha_1 + \alpha_2, \lambda)$; l'enunciato segue quindi dal fatto che ϕ determina univocamente la distribuzione della variabile aleatoria. Non è difficile inoltre generalizzare alla somma di più di due variabili aleatorie, vale infatti il seguente risultato.

Proposizione 5.7.1. Se X_i, $i = 1, 2, \ldots, n$ sono variabili aleatorie indipendenti, di tipo gamma con parametri (α_i, λ), allora $\sum_{i=1}^n X_i$ è una gamma di parametri $(\sum_{i=1}^n \alpha_i, \lambda)$.

Poiché una gamma di parametri $(1, \lambda)$ non è altro che una esponenziale di intensità λ, siamo in grado di determinare la legge della somma di questo tipo di variabili aleatorie.

Corollario 5.7.2. Se X_i, $i = 1, 2, \ldots, n$ sono variabili aleatorie esponenziali indipendenti, tutte di intensità λ, allora $\sum_{i=1}^n X_i$ è una gamma di parametri (n, λ).

Esempio 5.7.1. Se il tempo di vita di un tipo di batterie è una variabile aleatoria esponenziale di intensità λ, volendo fare funzionare un walkman che richiede una sola batteria, e avendone n a disposizione, il tempo totale di riproduzione che si può ottenere ha distribuzione gamma di parametri (n, λ). □

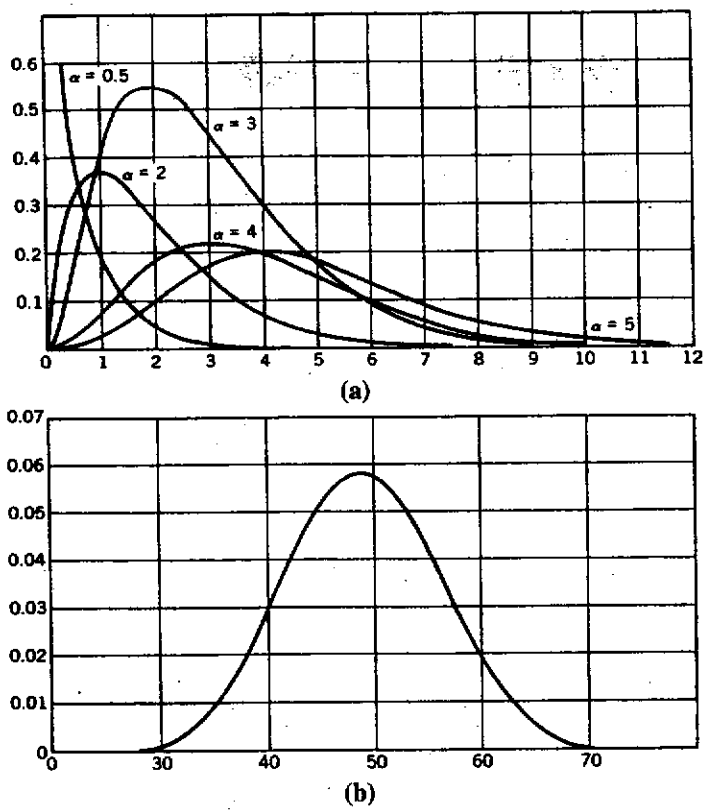


Figura 5.11 Densità di probabilità di varie distribuzioni di tipo gamma $(\alpha, 1)$, per (a) $\alpha = .5, 2, 3, 4, 5$ e (b) $\alpha = 50$.

La Figura 5.11 presenta le funzioni di densità della distribuzione gamma $(\alpha, 1)$, per diversi valori di α. Si noti come, quando α diventa grande, la densità tende a somigliare a quella normale. La giustificazione teorica di questo fatto risiede nel teorema del limite centrale, che sarà presentato nel prossimo capitolo.

5.8 Distribuzioni che derivano da quella normale

5.8.1 Le distribuzioni chi-quadro

Definizione 5.8.1. Se $Z_1, Z_2, \ldots, Z_n$ sono variabili aleatorie normali standard e indipendenti, allora la somma dei loro quadrati,

$$X := Z_1^2 + Z_2^2 + \cdots + Z_n^2 \tag{5.8.1}$$

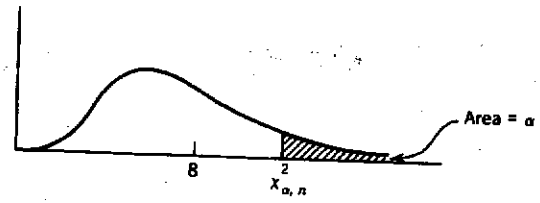


Figura 5.12 Probabilità di coda di un chi-quadro con 8 gradi di libertà.

è una variabile aleatoria che prende il nome di *chi-quadro a n gradi di libertà*. La notazione che useremo per indicare questo fatto è la seguente:

$$X \sim \chi^2_n \tag{5.8.2}$$

La distribuzione chi-quadro è riproducibile, nel senso che se X_1 e X_2 sono due chi-quadro indipendenti, con n_1 e n_2 gradi di libertà rispettivamente, allora la loro somma $X_1 + X_2$ è un chi-quadro con $n_1 + n_2$ gradi di libertà. Per dimostrare questo fatto non è necessario ricorrere alle funzioni generatrici, perché dalla definizione è evidente che $X_1 + X_2$ è la somma dei quadrati di $n_1 + n_2$ normali standard indipendenti, e quindi è una chi-quadro con altrettanti gradi di libertà.

In analogia con l'Equazione (5.5.9), per la distribuzione normale standard, se X è una chi-quadro con n gradi di libertà e α è un reale compreso tra 0 e 1, si definisce la quantità $\chi^2_{\alpha,n}$ tramite l'equazione seguente

$$P(X \geq \chi^2_{\alpha,n}) = \alpha \tag{5.8.3}$$

Ciò è illustrato in Figura 5.12.

Nella Tabella A.2 dell'Appendice, sono tabulati i valori di $\chi^2_{\alpha,n}$ per numerose combinazioni dei parametri α e n (incluse quelle utili a risolvere gli esempi e i problemi del testo). Inoltre i Programmi 5.8.1a e 5.8.1b del pacchetto software disponibile online, consentono di calcolare le probabilità inerenti alle distribuzioni chi-quadro, come pure i valori di $\chi^2_{\alpha,n}$.

Esempio 5.8.1. Si determini $P(X \leq 30)$ quando X è una variabile aleatoria chi-quadro con 26 gradi di libertà.

Usando il Programma 5.8.1a si trova immediatamente il risultato

$$P(X \leq 30) \approx 0.7325 \quad \square$$

Esempio 5.8.2. Si trovi quanto vale $\chi^2_{0.05,15}$.

Il Programma 5.8.2a fornisce il valore

$$\chi^2_{0.05,15} \approx 24.996 \quad \square$$

Esempio 5.8.3. Si vuole localizzare un oggetto nello spazio tridimensionale, tuttavia la misurazione che viene effettuata porta un errore sperimentale in ciascuna delle tre direzioni che è una variabile aleatoria normale di media 0 e deviazione standard 2 metri. Supponendo che questi tre errori siano indipendenti, si determini la probabilità che la distanza tra la posizione misurata e quella reale sia maggiore di 3 metri.

Se denotiamo con X_i, $i = 1,2,3$ gli errori nelle tre coordinate, e con D la distanza tra misurazione e posizione reale, per il teorema di Pitagora,

$$D^2 = X_1^2 + X_2^2 + X_3^2$$

D^2 non è una chi-quadro perché le X_i non sono normali standard: hanno deviazione standard pari a 2. Tuttavia $Z_i := X_i/2$ sono normali standard, quindi, $Y := Z_1^2 + Z_2^2 + Z_3^2$ è un chi-quadro a 3 gradi di libertà, e otteniamo che

$$\begin{aligned} P(D > 3) &= P(D^2 > 9) \\ &= P(X_1^2 + X_2^2 + X_3^2 > 9) \\ &= P(Z_1^2 + Z_2^2 + Z_3^2 > 9/4) \\ &= P(Y > 9/4) \approx 0.5222 \end{aligned}$$

dove il valore numerico finale è stato ottenuto con il Programma 5.8.1a □

5.8.1.1 * La relazione tra le distribuzioni chi-quadro e gamma

Vogliamo calcolare la funzione generatrice dei momenti delle distribuzioni chi-quadro. Iniziamo con 1 grado di libertà: sia $X \sim \chi_1^2$. Allora per definizione $X = Z^2$, dove $Z \sim \mathcal{N}(0,1)$, così che

$$\begin{aligned} E[e^{tX}] &= E[e^{tZ^2}] \\ &= \int_{-\infty}^{\infty} e^{tx^2} f_Z(x)\,dx \\ &= \int_{-\infty}^{\infty} e^{tx^2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-(1-2t)\frac{x^2}{2}\right\} dx \\ &= (1-2t)^{-1/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\bar{\sigma}}} \exp\left\{-\frac{x^2}{2\bar{\sigma}^2}\right\} dx \quad \text{ponendo } \bar{\sigma} = (1-2t)^{-1/2} \\ &= (1-2t)^{-1/2} \quad \text{perché l'integrando è la densita di una } \mathcal{N}(0,\bar{\sigma}^2) \end{aligned}$$

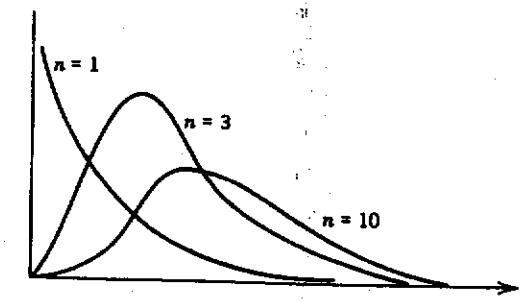


Figura 5.13 Grafici delle densità di probabilità di alcune leggi chi-quadro; n rappresenta il numero dei gradi di libertà.

Nel caso generale in cui $X \sim \chi_n^2$, si ottiene

$$\begin{aligned} E[e^{tX}] &= E[\exp\{tZ_1^2 + tZ_2^2 + \cdots + tZ_n^2\}] \\ &= E[e^{tZ_1^2} e^{tZ_2^2} \cdots e^{tZ_n^2}] \\ &= \prod_{i=1}^{n} E[e^{tZ_i^2}] \quad \text{per l'indipendenza delle } Z_i \\ &= (1-2t)^{-n/2} \quad \text{si deduce dal caso } n = 1 \end{aligned}$$

Siccome

$$(1-2t)^{-n/2} = \left(\frac{1}{1-2t}\right)^{n/2} = \left(\frac{1/2}{1/2-t}\right)^{n/2}$$

riconosciamo nella precedente la funzione generatrice di una distribuzione gamma di parametri $(n/2, 1/2)$. Quindi per l'unicità della distribuzione di probabilità corrispondende ad una data funzione generatrice, concludiamo che la distribuzione chi-quadro con n gradi di libertà coincide con la distribuzione gamma di parametri $n/2$ e $1/2$. La densità di probabilità di X è perciò data da

$$f(x) = \frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)}, \qquad x > 0 \tag{5.8.4}$$

Le densità delle distribuzioni chi-quadro con 1, 3 e 10 gradi di libertà sono rappresentate in Figura 5.13.

Riconsideriamo di seguito l'Esempio 5.8.3, ambientandolo nel piano anziché nello spazio tridimensionale.

Esempio 5.8.4. Nel tentativo di localizzare un oggetto nel piano selezioniamo un punto, e gli errori lungo le due coordinate sono normali indipendenti di media 0 e

deviazione standard 2. Vogliamo trovare la probabilità che la distanza dall'obiettivo sia maggiore di 3.

Denotiamo con D la distanza e con X_1, X_2 gli errori nelle due coordinate, in modo che sia

$$D^2 = X_1^2 + X_2^2$$

Poiché $Z_i := X_i/2$, per $i = 1, 2$ sono normali standard, $Y := Z_1^2 + Z_2^2$ è una chi-quadro con 2 gradi di libertà, ovvero una gamma di parametri $(1, 1/2)$, ovvero una esponenziale di intensità $1/2$, così che

$$\begin{aligned} P(D > 3) &= P(X_1^2 + X_2^2 > 9) \\ &= P(Z_1^2 + Z_2^2 > 9/4) \\ &= P(Y > 9/4) \\ &= e^{-9/8} \approx 0.3247 \quad \square \end{aligned}$$

Siccome la distribuzione chi-quadro con n gradi di libertà coincide con la gamma di parametri $\alpha = n/2$ e $\lambda = 1/2$, si può dedurre dalle Equazioni (5.7.6) e (5.7.7), che se $X \sim \chi_n^2$, allora

$$E[X] = n, \qquad \mathrm{Var}(X) = 2n \tag{5.8.5}$$

5.8.2 Le distribuzioni t

Definizione 5.8.2. Se Z e C_n sono variabili aleatorie indipendenti, la prima normale standard e la seconda chi-quadro con n gradi di libertà, allora la variabile aleatoria T_n, definita come

$$T_n := \frac{Z}{\sqrt{C_n/n}} \tag{5.8.6}$$

si dice avere *distribuzione t con n gradi di libertà*, cosa che si denota sinteticamente con

$$T_n \sim t_n \tag{5.8.7}$$

Tale variabile aleatoria viene anche detta *t di Student*[7] con n gradi di libertà.

In Figura 5.14 sono rappresentati i grafici delle densità di t_n per $n = 1, 5, 10$. La densità delle distribuzioni t, proprio come quella normale standard, è simmetrica rispetto all'asse di ascissa 0. In realtà è possibile mostrare che al crescere di n, la densità di t_n converge a quella della normale standard. Per capirne il motivo, ricordiamo che $C_n \sim \chi_n^2$ può essere espressa come somma dei quadrati di n gaussiane

[7] Si tratta dello pseudonimo usato da W. S. Gosset, si veda a pagina 6.

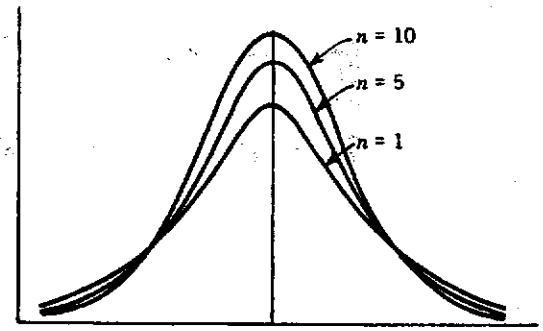


Figura 5.14 Densità di probabilità di alcune distribuzioni t; n rappresenta il numero di gradi di libertà.

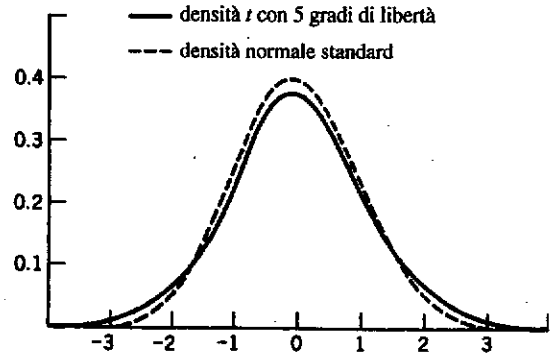


Figura 5.15 Confronto tra la densità normale standard e quella di t_5.

standard indipendenti, ovvero

$$\frac{C_n}{n} = \frac{Z_1^2 + \cdots + Z_n^2}{n}$$

dove $Z_1, \ldots, Z_n$ sono appunto $\mathcal{N}(0, 1)$ e indipendenti. La legge dei grandi numeri applicata a questa espressione, ci dice però che per n grande, C_n/n sarà, con probabilità prossima al 100%, molto vicino a $E[Z_i^2] = 1$. Quindi, per n grande, $T_n := Z/\sqrt{C_n/n}$ avrà circa la stessa distribuzione di Z.

La Figura 5.15 mette a confronto la densità di una distribuzione t con 5 gradi di libertà con quella della normale standard. Si noti che la t è caratterizzata da "code" più spesse (il termine esatto è *pesanti*), a indicare una variabilità maggiore rispetto alla gaussiana.

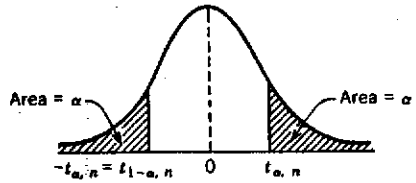


Figura 5.16 Dimostrazione grafica che $-t_{\alpha,n} = t_{1-\alpha,n}$.

Anche se in questa sede non approfondiamo questo argomento, è possibile dimostrare che valore atteso e varianza di T_n sono dati da

$$E[T_n] = 0, \qquad n \geq 2 \tag{5.8.8}$$

$$\text{Var}(T_n) = \frac{n}{n-2}, \qquad n \geq 3 \tag{5.8.9}$$

Si noti che, al crescere di n, la varianza di t_n decresce, convergendo a 1 dall'alto (cioè alla varianza della gaussiana standard).

In analogia con quanto fatto in precedenza per la distribuzione normale standard e per le chi-quadro, se T_n è una t con n gradi di libertà e $\alpha \in (0,1)$, si definisce la quantità $t_{\alpha,n}$ in modo che sia

$$P(T_n \geq t_{\alpha,n}) = \alpha \tag{5.8.10}$$

Dalla simmetria rispetto allo zero della densità t, segue che $-T_n$ ha la stessa distribuzione di T_n, cosicché

$$\begin{aligned} \alpha &= P(-T_n \geq t_{\alpha,n}) \\ &= P(T_n \leq -t_{\alpha,n}) \\ &= 1 - P(T_n > -t_{\alpha,n}) \end{aligned}$$

quindi

$$P(T_n \geq -t_{\alpha,n}) = 1 - \alpha$$

da cui si ottiene che

$$-t_{\alpha,n} = t_{1-\alpha,n} \tag{5.8.11}$$

come è illustrato in Figura 5.16.

I valori di $t_{\alpha,n}$ per diverse combinazioni di α e n sono tabulati nella Tabella A.3 in Appendice[8]. Inoltre, i Programmi 5.8.2a e 5.8.2b disponibili online sul sito di

[8] La Tabella A.3 riporta per ultima una riga di valori relativi ad un numero "infinito" di gradi di libertà. Come abbiamo avuto modo di rilevare, il limite della legge di t_n per n che tende all'infinito è la distribuzione $\mathcal{N}(0,1)$, e infatti i valori della tabella sono quelli di z_α.

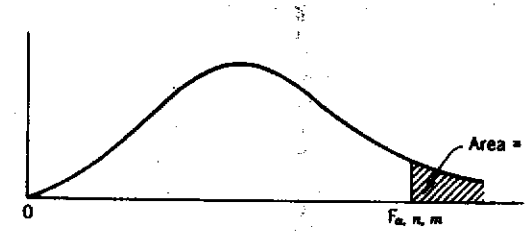


Figura 5.17 Grafico della densità di $F_{n,m}$.

questo libro permettono di calcolare la funzione di ripartizione delle t di Student e i valori di $t_{\alpha,n}$ rispettivamente.

Esempio 5.8.5. Sia $T \sim t_{12}$; si trovino **(a)** $P(T \leq 1.4)$ e **(b)** $t_{0.025,9}$.

Eseguendo i Programmi 5.8.2a e 5.8.2b si ottengono immediatamente i risultati seguenti: **(a)** 0.9066, **(b)** 2.263. □

Osservazione 5.8.1. Si noti che la Tabella A.3 riporta per $t_{0.025,9}$ un valore di 2.262, che è leggermente diverso da quello ottenuto qui. Ciò è dovuto al fatto che il software in dotazione fornisce una approssimazione del valore cercato. Per questo motivo, se si richiede un risultato molto preciso, è preferibile quando possibile usare i valori tabulati, oppure un software professionale.

5.8.3 Le distribuzioni F

Definizione 5.8.3. Se C_n e C_m sono variabili aleatorie indipendenti, di tipo chi-quadro con n e m gradi di libertà rispettivamente, allora la variabile aleatoria $F_{n,m}$, definita da

$$F_{n,m} := \frac{C_n/n}{C_m/m} \tag{5.8.12}$$

si dice avere *distribuzione F con n e m gradi di libertà*; $F_{n,m}$ prende anche il nome di variabile aleatoria di tipo F, oppure di *F di Fisher*, oppure di *Z di Fisher*, con n e m gradi di libertà.

In analogia con quanto fatto in precedenza per altre distribuzioni, per ogni $\alpha \in (0,1)$, si definisce la quantità $F_{\alpha,n,m}$ in modo che sia

$$P(F_{n,m} > F_{\alpha,n,m}) = \alpha \tag{5.8.13}$$

Ciò è rappresentato in Figura 5.17.

Le quantità $F_{\alpha,n,m}$ sono tabulate nella Tabella A.4 in Appendice per diversi valori di n e m, per $\alpha = 0.05$. In effetti tipicamente le tavole di valori per $F_{\alpha,n,m}$, contengono solo valori di α minori di 0.5. Se si vuole invece un valore corrispondente ad

un $\alpha > 0.5$, è possibile ottenerlo con i passaggi seguenti:

$$\begin{aligned}\alpha &= P\left(\frac{C_n/n}{C_m/m} > F_{\alpha,n,m}\right)\\ &= P\left(\frac{C_m/m}{C_n/n} < \frac{1}{F_{\alpha,n,m}}\right)\\ &= 1 - P\left(\frac{C_m/m}{C_n/n} \geq \frac{1}{F_{\alpha,n,m}}\right)\end{aligned}$$

o, equivalentemente,

$$P\left(\frac{C_m/m}{C_n/n} > \frac{1}{F_{\alpha,n,m}}\right) = 1 - \alpha$$

Siccome però la variabile aleatoria ottenuta all'interno della $P(\cdot)$ è di tipo F con m e n gradi di libertà, per la definizione di $F_{1-\alpha,n,m}$,

$$P\left(\frac{C_m/m}{C_n/n} \geq F_{1-\alpha,n,m}\right) = 1 - \alpha$$

Confrontando le ultime due equazioni si vede subito che deve essere

$$\frac{1}{F_{\alpha,n,m}} = F_{1-\alpha,n,m} \tag{5.8.14}$$

Quindi, ad esempio, $F_{0.9,5,7} = F^{-1}_{0.1,7,5} \approx 1/3.37 \approx 0.297$, dove il valore di $F_{0.1,7,5}$ è stato ricavato dalla Tabella A.4 dell'Appendice.

Il Programma 5.8.3 sul sito web del libro, permette di calcolare la funzione di ripartizione di $F_{n,m}$.

Esempio 5.8.6. Si determini $P(F_{6,14} \leq 1.5)$.

Eseguendo il Programma 5.8.3 si trova che la soluzione è 0.752. □

Problemi

1. Uno dei sistemi installati su un satellite è costituito da 4 componenti, e riesce a funzionare correttamente se almeno 2 di essi sono efficienti. Se ciascuno dei componenti, indipendentemente dagli altri, funziona bene con una probabilità di 0.6, qual è la probabilità che l'intero sistema funzioni?

2. Un canale di comunicazione trasmette dei *bit*, ovvero cifre binarie che possono essere 0 oppure 1. A causa del rumore elettrostatico, vi è una probabilità di 0.2 che il bit ricevuto sia tanto disturbato da essere decodificato erroneamente. Supponiamo in queste condizioni di volere trasmettere un messaggio importante, costituito da una sola cifra. Per

ridurre la probabilità di errore potremmo trasmettere 00000 al posto di 0 e 11111 al posto di 1. Introducendo questa ridondanza, e decodificando il messaggio "a maggioranza" (si decodifica 1 se si ricevono più cifre 1 che cifre 0, e viceversa), qual è la probabilità di decodificare erroneamente il messaggio? Quali ipotesi di indipendenza stai implicitamente assumendo?

3. Se un votante scelto a caso è favorevole ad una certa riforma con probabilità di 0.7, qual è la probabilità che su 10 votanti, esattamente 7 siano favorevoli?

4. Supponiamo che un particolare tratto somatico (come il colore degli occhi o l'essere mancini) sia governato da una sola coppia di geni, ciascuno dei quali può essere d, dominante, oppure r, recessivo. Un individuo con la coppia dd è dominante puro, uno con la coppia rr è recessivo puro e uno con la coppia rd è ibrido. È noto inoltre che i soggetti ibridi presentano lo stesso tratto somatico dei dominanti puri e che ogni nascituro riceve un gene a caso da ciascun genitore (si veda anche il Problema 42 del Capitolo 3, a pagina 88). Se due genitori ibridi rispetto ad un certo tratto, hanno 4 figli, qual è la probabilità che esattamente 3 di essi presentino il tratto dominante?

5. Un moderno aereo civile è in grado di restare in volo se almeno la metà dei suoi motori è in funzione. Supponiamo che ogni motore indipendentemente dagli altri abbia una probabilità p di funzionare correttamente. Per quali valori di p un aereo a 4 motori ha più probabilità di successo di un aereo a 2 motori?

6. Sia X una variabile aleatoria binomiale con media 7 e varianza 2.1. Quanto valgono **(a)** $P(X = 4)$ e **(b)** $P(X > 12)$?

7. Siano X e Y due variabili aleatorie binomiali di parametri (n, p) e $(n, 1-p)$. Verifica e commenta le seguenti identità:

   (a) $P(X \leq i) = P(Y \geq n - i)$, per ogni $i = 0, 1, \ldots, n$;

   (b) $P(X = k) = P(Y = n - k)$, per ogni $k = 0, 1, \ldots, n$.

8. Sia X una variabile aleatoria binomiale di parametri n e p con $0 < p < 1$. Dimostra che

   (a) $P(X = k+1) = \dfrac{p}{1-p}\dfrac{n-k}{k+1}P(X = k)$, per $k = 0, 1, \ldots, n-1$.

   (b) Al crescere di k da 0 a n, $P(X = k)$ prima cresce, poi decresce, toccando il suo massimo quando k è il più grande intero minore o uguale a $(n+1)p$.

9. Determina la funzione generatrice dei momenti della distribuzione binomiale e poi usala per verificare le formule per la media e la varianza ricavate nel testo.

10. Confronta le probabilità esatte con l'approssimazione di Poisson nei casi seguenti. Si intende che X è binomiale di parametri n e p.

    (a) $P(X = 2)$ quando $n = 10$ e $p = 0.1$;

    (b) $P(X = 0)$ quando $n = 10$ e $p = 0.1$;

    (c) $P(X = 4)$ quando $n = 9$ e $p = 0.2$.

11. Se tu acquistassi un biglietto di ciascuna di 50 diverse lotterie, e in ognuna la probabilità di vittoria fosse 1/100, quale sarebbe la probabilità (approssimativa) che tu risulti vincitore **(a)** almeno una volta, **(b)** esattamente una volta, e **(c)** almeno due volte?

12. Supponiamo che il numero di raffreddori contratti da ogni persona in un anno solare sia una variabile aleatoria di Poisson di media 3. Viene presentato un nuovo miracoloso farmaco che – efficacie sul 75% della popolazione – abbassa la media della poissoniana a 2. Nel restante 25% dei casi non ha invece alcun effetto apprezzabile. Se un individuo prova il farmaco per un anno, e in quel periodo di tempo non si ammala di raffreddore nemmeno una volta, qual è la probabilità che il farmaco su di lui sia stato efficace?

13. Negli Stati Uniti, durante gli anni 80 del secolo scorso, ogni settimana sono morte sul lavoro una media di 121.95 persone. Dai una stima delle seguenti quantità:

(a) la frazione di settimane con 130 vittime o più;

(b) la frazione di settimane con 100 vittime o meno.

Spiega il tuo ragionamento.

14. In un anno, nella città di New York, si celebrano circa 80 000 matrimoni. Dai una stima della probabilità che per una almeno delle coppie

(a) entrambi gli sposi siano nati il 30 aprile;

(b) i due sposi celebrino il compleanno nella medesima data.

Giustifica le risposte date.

15. Il numero medio di errori tipografici per pagina di una certa rivista è di 0.2. Qual è la probabilità che la pagina che ti accingi a leggere contenga **(a)** nessun refuso oppure **(b)** 2 o più refusi? Spiega il tuo ragionamento.

16. La probabilità di errore nella trasmissione di una cifra binaria attraverso un certo canale di comunicazione è di 10^{-3}.

(a) Scrivi un'espressione esatta per la probabilità di totalizzare più di tre errori trasmettendo un blocco di 1000 bit.

(b) Calcola una approssimazione di tale probabilità.

Puoi assumere l'indipendenza degli errori.

17. Sia X una variabile aleatoria di Poisson di media λ. Devi dimostrare che, al crescere di i, $P(X = i)$ prima aumenta, poi diminuisce, toccando il suo massimo quando i è il più grande intero minore o uguale a λ.

18. Un commerciante fa una ordinazione di 100 transistor. La sua politica consiste nel provarne 10 scelti a caso e rifiutare tutta l'ordinazione se almeno 2 di essi sono difettosi. Se effettivamente essa contiene 20 pezzi difettosi, qual è la probabilità che venga accettata?

19. Sia X una variabile aleatoria ipergeometrica di parametri n, m, e k. Tale cioè che

$$P(X = i) = \frac{\binom{n}{i}\binom{m}{k-i}}{\binom{n+m}{k}}, \qquad i = 0, 1, \ldots, n$$

(a) Deduci una formula per $P(X = i)$ in termini di $P(X = i - 1)$.

(b) Poni $n = m = 10$ e $k = 5$. Calcola $P(X = i)$ per $i = 0, 1, \ldots, 5$, partendo da $P(X = 0)$ e utilizzando la formula trovata al punto **(a)**.

(c) Scrivi un programma per computer che utilizzi la ricorsione di cui al punto **(a)** per calcolare la funzione di ripartizione di una generica variabile aleatoria ipergeometrica.

(d) Usa il programma scritto al punto **(c)** per calcolare $P(X \leq 10)$ quando $n = m = 30$, e $k = 15$.

20. Si effettua una successione di prove indipendenti, ciascuna delle quali ha probabilità di successo pari a p. Sia X il numero della prima prova che risulta in un successo, ovvero X vale k se le prime $k - 1$ prove hanno esito negativo ma la k-esima ha esito positivo Una variabile aleatoria di questo tipo si dice *geometrica* di parametro p. Calcola

(a) $P(X = k)$, per $k = 1, 2, \ldots$

(b) $E[X]$.

Fissato poi un numero intero $r \geq 1$, sia Y il numero della r-esima prova che risulta in un successo, ovvero, Y rappresenta quante prove dobbiamo attendere per ottenere r successi. Questo tipo di variabile aleatoria si dice *di Pascal* o anche *binomiale negativa*.

(c) Calcola $P(Y = k)$, per $k = r, r + 1, \ldots$ (*Suggerimento:* Affinché Y sia pari a k, quanti successi e quanti insuccessi devono realizzarsi nelle prime $k - 1$ prove? E quale deve essere il risultato della k-esima?).

(d) Dimostra che $E[Y] = r/p$ (*Suggerimento:* Decomponi Y nella somma $Y_1 + \cdots + Y_r$, dove Y_i è il numero di prove che vengono realizzate successivamente al successo $(i - 1)$-esimo, e fino al verificarsi del successo i-esimo, inclusa quest'ultima).

21. Dimostra che, se U è uniforme su $(0, 1)$, allora $a + (b - a)U$ è uniforme su (a, b).

22. Arrivi alla fermata dell'autobus alle 10, e sei certo che ne passerà uno in un momento distribuito uniformemente tra le 10 e le 10.30.

(a) Qual è la probabilità che tu debba aspettare più di 10 minuti?

(b) Se alle 10.15 l'autobus non è ancora arrivato, qual è la probabilità che tu debba aspettare almeno altri 10 minuti?

23. Sia X una variabile aleatoria normale di parametri $\mu = 10$ e $\sigma^2 = 36$. Calcola **(a)** $P(X > 5)$; **(b)** $P(4 < X < 16)$; **(c)** $P(X < 8)$; **(d)** $P(X < 20)$; **(e)** $P(X > 16)$.

24. Un certo test nazionale di matematica viene proposto in tutte le ultime classi delle scuole secondarie. Esso produce punteggi che hanno distribuzione normale con media 500 e deviazione standard 100. Si scelgono poi a caso 5 studenti che hanno affrontato il test; calcola le probabilità che **(a)** i loro punteggi siano tutti inferiori a 600; **(b)** esattamente 3 punteggi siano superiori a 640.

25. Il livello (in pollici) delle precipitazioni annuali in una certa regione, ha distribuzione normale con $\mu = 40$ e $\sigma = 4$. Qual è la probabilità che in 2 dei prossimi 4 anni le precipitazioni superino i 50 pollici? Puoi assumere che i livelli di pioggia di anni successivi siano indipendenti.

26. La larghezza di una scanalatura in un trafilato di duralluminio è (espressa in pollici) una variabile aleatoria normale con $\mu = 0.9000$ e $\sigma = 0.0030$. Le specifiche di fabbricazione assegnate impongono il limite 0.9000 ± 0.0050.

(a) Che percentuale dei trafilati sarà difettosa?

(b) Qual è il più alto valore di σ accettabile, per avere una percentuale di difettosi non superiore all'1%?

27. Un certo tipo di lampadine ha una luminosità che ha distribuzione normale con media 2 000 e deviazione standard 85. Determina un limite inferiore di luminosità da dichiarare affinché non più del 5% delle lampadine prodotte non lo rispetti. (Ovvero, determina L tale che $P(X \geq L) = 0.95$, dove X è la luminosità di una lampadina scelta a caso.)

28. Una azienda produce bulloni con diametro dichiarato tra 1.19 e 1.21 pollici. Se i bulloni che escono dalla linea di produzione hanno un diametro che è una variabile aleatoria gaussiana con media 1.20 pollici e deviazione standard 0.005, che percentuale dei bulloni non soddisfa le specifiche?

29. Sia $I := \int_{-\infty}^{\infty} e^{-x^2/2}\,dx$.

(a) Dimostra che, se $I = \sqrt{2\pi}$, allora per ogni μ e σ, con $\sigma > 0$,

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx = 1$$

(b) Dimostra che $I = \sqrt{2\pi}$ procedendo come segue.

$$I^2 = \int_{-\infty}^{\infty} e^{-x^2/2}\,dx \int_{-\infty}^{\infty} e^{-y^2/2}\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left\{-\frac{x^2+y^2}{2}\right\} dx\,dy$$

Valuta l'integrale doppio tramite un cambiamento di coordinate, da cartesiane a polari. (Ovvero, poni $x = r\cos\theta$, $y = r\sin\theta$ e $dx\,dy = r\,dr\,d\theta$.)

30. Una variabile aleatoria X ha distribuzione *lognormale* se $\log X$ ha distribuzione normale. Supponendo che $\log X \sim \mathcal{N}(\mu, \sigma^2)$, calcola la funzione di ripartizione di X: quanto vale $P(X \leq x)$?

31. I tempi di vita dei circuiti integrati, fabbricati da un produttore di semiconduttori, hanno distribuzione normale con media di 4.4×10^6 ore e deviazione standard di 3×10^5 ore. Se un produttore di mainframe necessita che almeno il 90% di una grossa ordinazione di circuiti abbia un tempo di vita non inferiore a 4.0×10^6 ore, è il caso che si rivolga a queso produttore?

32. Con riferimento al Problema 31, qual è la probabilità che su un'ordinazione di 100 pezzi ve ne siano almeno 4 con tempo di vita inferiore a 3.8×10^6 ore?

33. Il tempo di vita del tubo catodico di un televisore a colori ha distribuzione gaussiana con media 8.2 anni e deviazione standard 1.4 anni. Quale percentuale di questi tubi catodici dura **(a)** più di 10 anni; **(b)** meno di 5 anni; **(c)** tra i 5 e i 10 anni?

34. Le precipitazioni annuali a Cincinnati hanno distribuzione normale con media 40.14 pollici e deviazione standard 8.7 pollici.

(a) Qual è la probabilità che quest'anno si superino i 42 pollici?

(b) Con che probabilità nei prossimi due anni cadranno in totale più di 84 pollici di pioggia?

(c) Con che probabilità nei prossimi tre anni cadranno in totale più di 126 pollici di pioggia?

(d) Per i punti (b) e (c), che ipotesi di indipendenza stai assumendo?

35. La statura delle donne adulte negli Stati Uniti, ha una distribuzione normale con media 64.5 pollici e deviazione standard 2.4 pollici.

(a) Trova la probabilità che una donna scelta a caso sia alta meno di 63 pollici;

(b) meno di 70 pollici;

(c) tra i 63 e i 70 pollici.

(d) Alice è alta 72 pollici. Che percentuale della popolazione femminile adulta è più bassa di lei?

(e) Trova la probabilità che la media aritmetica della statura di due donne scelte a caso sia superiore a 66 pollici.

(f) Ripeti il punto (e) per 4 donne.

36. Un test per il Q.I. produce punteggi con distribuzione normale di media 100 e deviazione standard 14.2. Che intervallo di punteggi raggiunge l'1% della popolazione formato dalle persone più intelligenti?

37. Il tempo (in ore) necessario per riparare un macchinario è una variabile aleatoria esponenziale con parametro $\lambda = 1$.

(a) Qual è la probabilità che la riparazione superi le 2 ore di tempo?

(b) Qual è la probabilità condizionata che la riparazione richieda almeno 3 ore, sapendo che ne richiede più di 2?

38. Il numero di anni di funzionamento di una radio ha distribuzione esponenziale di parametro $\lambda = 1/8$. Se si compra una radio usata, qual è la probabilità che funzioni per altri 10 anni o più?

39. Il signor Jones è convinto che il tempo di vita di una automobile (in migliaia di miglia percorse) sia una variabile aleatoria esponenziale di parametro $1/20$. Il signor Smith ha una macchina usata da vendere, che ha percorso circa 10 000 miglia.

(a) Se Jones decide di comprarla, che probabilità ha di farle fare almeno altre 20 000 miglia, prima che sia da buttare?

(b) Rispondi nuovamente, nell'ipotesi che il tempo di vita dell'auto (in migliaia di miglia percorse), abbia distribuzione uniforme sull'intervallo $(0, 40)$.

***40.** Siano $X_1, X_2, \ldots, X_n$ i primi n tempi che separano gli eventi di un processo di Poisson di intensità λ, e poniamo $S_n := \sum_{i=1}^{n} X_i$.

(a) Qual è l'interpretazione di S_n?

(b) Spiega perché i due eventi $\{S_n \leq t\}$ e $\{N(t) \geq n\}$ sono identici.

(c) Usa il risultato del punto **(b)** per mostrare che

$$P(S_n \leq t) = 1 - \sum_{j=0}^{n-1} \frac{(\lambda t)^j}{j!} e^{-\lambda t}$$

(d) Derivando la formula del punto **(c)** per la funzione di ripartizione di S_n, mostra che S_n ha distribuzione gamma con parametri n e λ. (Questo risultato segue anche dal Corollario 5.7.2.)

***41.** In una certa regione, i terremoti si susseguono secondo un processo di Poisson di intensità pari a 5 all'anno.

(a) Qual è la probabilità che vi siano almeno 2 terremoti nella prima metà del 2015?

(b) Assumendo che l'evento del punto **(a)** si verifichi, qual è la probabilità che nei primi 9 mesi del 2016 non vi siano terremoti?

(c) Assumendo ancora che l'evento del punto **(a)** si verifichi, qual è la probabilità che nei primi 9 mesi del 2015 vi siano almeno 4 terremoti?

42. Stiamo sparando ad un bersaglio che si trova su un piano bidimensionale. Le distanze in orizzontale e in verticale del punto che colpiamo rispetto al bersaglio sono variabili aleatorie normali e indipendenti con media 0 e varianza 4. Sia D la distanza tra il bersaglio e il punto colpito. Quanto vale $E[D]$?

43. Sia X una variabile aleatoria chi-quadro con 6 gradi di libertà. Trova **(a)** $P(X \leq 6)$; **(b)** $P(3 \leq X \leq 9)$.

44. Siano X e Y due chi-quadro indipendenti, con 3 e 6 gradi di libertà rispettivamente. Determina la probabilità che $X + Y$ sia superiore a 10.

45. Dimostra che $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. (*Suggerimento:* Calcola il valore di $\int_0^\infty x^{-1/2} e^{-x}\, dx$, con la sostituzione $x = \frac{1}{2}y^2$, $dx = y\, dy$.)

46. Sia T una t di Student con con 8 gradi di libertà. Trova **(a)** $P(T \geq 1)$, **(b)** $P(T \leq 2)$, e **(c)** $P(-1 < T < 1)$.

47. Dimostra che, se T_n ha distribuzione t con n gradi di libertà, allora T_n^2 ha distribuzione F con 1 e n gradi di libertà.

6 La distribuzione delle statistiche campionarie

Contenuto

6.1 Introduzione

La statistica è la scienza che si occupa di trarre conclusioni dai dati sperimentali. Una situazione tipica con la quale bisogna spesso confrontarsi negli ambiti tecnologici, è quella in cui si studia un insieme molto grande, detto *popolazione*, di oggetti a cui sono associate delle quantità misurabili. L'approccio statistico consiste nel selezionare un sottoinsieme ridotto di oggetti, che viene detto *campione*, e analizzarlo sperando di essere in grado di trarre da esso delle conclusioni valide per la popolazione nel suo insieme.

Per basare sui dati del campione delle inferenze che riguardino l'intera popolazione, è necessario assumere qualche condizione sulle relazioni che legano questi due insiemi. Un'ipotesi fondamentale – in molti casi del tutto ragionevole – è che vi sia una (implicita) distribuzione di probabilità della popolazione, nel senso che se da essa si estraggono degli oggetti in maniera casuale, le quantità numeriche loro associate possono essere pensate come variabili aleatorie indipendenti, tutte con tale distribuzione. Se tutto il campione viene selezionato in maniera casuale, sembra ragionevole supporre che i suoi dati siano valori indipendenti provenienti da tale distribuzione.

Definizione 6.1.1. Un insieme $X_1, X_2, \ldots, X_n$ di variabili aleatorie indipendenti, tutte con la stessa distribuzione F, si dice *campione* o *campione aleatorio* della distribuzione F.

In pratica la distribuzione F non è mai completamente nota, però è possibile usare i dati per fare dell'*inferenza* su F. In alcuni casi è possibile che F sia nota eccetto che per dei parametri incogniti (si potrebbe ad esempio sapere che F è una distribuzione normale, ma non conoscerne la media e la varianza; oppure F potrebbe essere di Poisson, ma con parametro incognito); in altri casi potremmo non sapere praticamente nulla di F (tranne forse assumere che essa sia continua, oppure discreta). I problemi in cui la distribuzione F è nota a meno di un insieme di parametri incogniti sono detti problemi di inferenza *parametrica*; quelli in cui nulla si sa sulla distribuzione F sono invece problemi di inferenza *non parametrica*.

Esempio 6.1.1. È stato da poco introdotto un nuovo sistema di produzione dei circuiti integrati; i chip prodotti hanno tempi di vita che si pensano essere variabili aleatorie indipendenti con distribuzione F incognita.

È possibile che si individuino delle ragioni fisiche che convincano a priori che F deve avere una particolare forma parametrica; ad esempio potremmo essere portati a pensare che F sia normale, o forse esponenziale. Se questo è il caso abbiamo a che fare con un problema di statistica parametrica, e si possono usare i dati di un campione per stimare i parametri di F. Se F fosse una distribuzione normale incognita, vorremmo stimare la sua media e la sua varianza; se invece presumessimo che F sia di tipo esponenziale, vorremmo stimare la sua media o (ma sarebbe equivalente) la sua intensità.

In altre situazioni invece potrebbe non esserci alcuna ragione fisica per supporre che F abbia una forma particolare; in quel caso, fare dell'inferenza su F costituirebbe un problema non parametrico. □

In questo capitolo ci occupiamo delle distribuzioni di probabilità di alcune statistiche. Il termine *statistica* indica una *variabile aleatoria* che è semplicemente una funzione dei dati di un campione; i due principali esempi di statistiche che affrontiamo, sono la media campionaria e la varianza campionaria. Nella Sezione 6.2 prendiamo in considerazione la media campionaria e ne determiniamo valore atteso e varianza. È un fatto notevole che quando la numerosità del campione è anche solo moderatamente elevata, la distribuzione della media campionaria diviene approssimativamente normale (per quasi ogni forma di F!). Questa è una conseguenza del teorema del limite centrale, uno dei risultati teorici più rilevanti in probabilità, che è discusso nella Sezione 6.3. Nella Sezione 6.4 presentiamo la varianza campionaria e ne calcoliamo il valore atteso. Nella Sezione 6.5 ci restringiamo al caso che la popolazione abbia distribuzione normale e determiniamo la legge congiunta di media e varianza campionarie. Nella Sezione 6.6, infine, approfondiamo il concetto di campionamento da una popolazione finita e illustriamo cosa si intende con "campione aleatorio"; in pratica quando le dimensioni della popolazione sono grandi rispetto all'ampiezza del campione, essa viene trattata come se fosse infinita: questo approccio viene illustrato e se ne discutono le conseguenze.

6.2 La media campionaria

Consideriamo una popolazione di elementi, a ciascuno dei quali è associata una grandezza numerica. La popolazione potrebbe ad esempio essere costituita dagli individui adulti facenti parte di una qualche categoria di persone, e la grandezza numerica di interesse potrebbe essere il reddito annuale, la statura, l'età o altro. Sia $X_1, X_2, \ldots, X_n$ un campione di dati estratto da questa popolazione. È comune supporre che i valori numerici associati a ciascuno degli elementi del campione, siano variabili aleatorie indipendenti e identicamente distribuite. Denotiamo con μ e σ^2 la loro media e la loro varianza, che prendono il nome di *media e varianza della popolazione*. In analogia con la Definizione 2.3.1 di pagina 22, definiamo la *media campionaria* come

$$\overline{X} := \frac{X_1 + X_2 + \cdots + X_n}{n} \tag{6.2.1}$$

Si noti che $\overline{X}$ è una funzione delle variabili aleatorie $X_1, X_2, \ldots, X_n$. In quanto tale è una *statistica*, e in particolare è a sua volta una variabile aleatoria. Ha senso quindi domandarsi quanto valgano il valore atteso della media campionaria e la sua varianza. È facile vedere che

$$\begin{aligned} E[\overline{X}] &= E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] \\ &= \frac{E[X_1] + E[X_2] + \cdots + E[X_n]}{n} \\ &= \frac{n\mu}{n} = \mu \end{aligned} \tag{6.2.2}$$

e, per la varianza,

$$\begin{aligned} \mathrm{Var}(\overline{X}) &= \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\ &= \frac{\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)}{n^2} \qquad \text{per l'indipendenza} \\ &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned} \tag{6.2.3}$$

La media campionaria ha quindi lo stesso valore atteso della distribuzione da stimare, mentre la sua varianza risulta ridotta di un fattore n. Da questo possiamo dedurre che $\overline{X}$ è centrata attorno a μ, e la sua variabilità si riduce sempre di più con l'aumentare di n. Una esemplificazione di questo comportamento è illustrata nella Figura 6.1, che riporta, per diversi valori di n, le densità di probabilità per le medie campionarie di una popolazione normale standard.

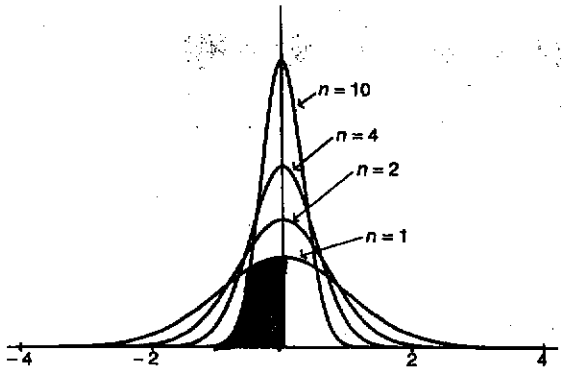


Figura 6.1 Densità delle medie campionarie di una popolazione normale standard.

6.3 Il teorema del limite centrale

In questa sezione affrontiamo uno dei risultati più notevoli della teoria della probabilità, il *teorema del limite centrale*[1]. In termini semplicistici, esso afferma che la somma di un numero elevato di variabili aleatorie indipendenti, tende ad avere distribuzione approssimativamente normale. L'importanza è duplice: da un lato siamo in grado di ottenere stime approssimative delle probabilità che riguardano la somma di variabili aleatorie indipendenti, dall'altro abbiamo giustificato il fatto notevole che la distribuzione empirica delle frequenze di un gran numero di popolazioni naturali esibisca forme a campana (in realtà, gaussiane).

L'enunciato, presentato nella sua versione più semplice, è il seguente:

Teorema 6.3.1 (Teorema del limite centrale). Siano $X_1, X_2, \ldots, X_n$ delle variabili aleatorie i.i.d. (indipendenti e identicamente distribuite), tutte con media μ e varianza σ^2. Allora se n è grande, la somma

$$X_1 + X_2 + \cdots + X_n$$

è approssimativamente normale con media $n\mu$ e varianza $n\sigma^2$.

Si può anche normalizzare la somma precedente in modo da ottenere una distribuzione approssimativamente normale *standard*. Si ha infatti che

$$\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \sim \mathcal{N}(0,1) \tag{6.3.1}$$

[1] Spesso lo si trova abbreviato negli acronimi TLC o *CLT*, dove il secondo deriva ovviamente dall'espressione inglese corrispondente, *central limit theorem*.

dove con il simbolo $\sim$ si intende "è approssimativamente distribuito come". Ciò significa che per n grande e x qualsiasi vale l'approssimazione

$$P\left(\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < x\right) \approx \Phi(x)$$

dove Φ denota la funzione di ripartizione della normale standard che è stata introdotta con l'Equazione (5.5.5).

Esempio 6.3.1. Una compagnia di assicurazioni ha 25 000 polizze auto attive. Il risarcimento dovuto annualmente per ogni singolo assicurato è una variabile aleatoria con media 320 e deviazione standard 540. Quanto vale approssimativamente la probabilità che in un determinato anno le richieste di indennizzi superino 8.3 milioni?

Sia X la richiesta annuale complessiva di indennizzi. Numeriamo gli assicurati, e sia X_i il risarcimento dovuto all'assicurato i-esimo, per $i = 1, 2, \ldots, n$, con $n = 25\,000$. È chiaro che $X = \sum_{i=1}^{n} X_i$, e segue dal teorema del limite centrale, che X ha approssimativamente distribuzione normale con media $320 \times 25\,000 = 8 \times 10^6$ e deviazione standard $540\sqrt{25\,000} \approx 8.54 \times 10^4$. Perciò, se Z denota una variabile aleatoria con distribuzione $\mathcal{N}(0,1)$,

$$\begin{aligned} P(X > 8.3 \times 10^6) &= P\left(\frac{X - 8 \times 10^6}{8.54 \times 10^4} > \frac{8.3 \times 10^6 - 8 \times 10^6}{8.54 \times 10^4}\right) \\ &\approx P\left(Z > \frac{0.3 \times 10^6}{8.54 \times 10^4}\right) \\ &\approx P(Z > 3.51) \approx 0 \end{aligned}$$

Quindi la probabilità che la compagnia debba pagare in un anno più di 8.3 milioni è trascurabile. □

Esempio 6.3.2. Gli ingegneri che stanno studiando un ponte sono convinti che il numero di tonnellate W, che una singola campata può sostenere senza subire danni strutturali, sia una variabile aleatoria normale di media 200 e deviazione standard 20. Supponiamo che il peso in tonnellate degli autoveicoli che vi passano sia una variabile aleatoria di media 1.5 e deviazione standard 0.15. Quante automobili dovrebbero essere contemporaneamente sulla campata, affinché la probabilità di danno strutturale superi il 10%?

Sia P_n la probabilità di un danno strutturale, quando vi sono n autoveicoli.

$$\begin{aligned} P_n &= P(X_1 + X_2 + \cdots + X_n \geq W) \\ &= P(X_1 + X_2 + \cdots + X_n - W \geq 0) \end{aligned}$$

dove $X_1, X_2, \ldots, X_n$ sono i pesi delle auto. Per il teorema del limite centrale, $\sum_{i=1}^{n} X_i$ è approssimativamente normale, $\mathcal{N}(1.5n, 0.0225n)$. Quindi, siccome W è indipendente da tutte le X_i ed è normale, ne segue che $\sum_{i=1}^{n} X_i - W$ è

approssimativamente normale con media e varianza date da

$$E\left[\sum_{i=1}^{n} X_i - W\right] = 1.5n - 200$$

$$\text{Var}\left(\sum_{i=1}^{n} X_i - W\right) = \text{Var}\left(\sum_{i=1}^{n} X_i\right) + \text{Var}(W) = 0.0225n + 400$$

Perciò, se poniamo

$$Z := \frac{\sum_{i=1}^{n} X_i - W - (1.5n - 200)}{\sqrt{0.0225n + 400}}$$

allora

$$P_n = P\left(Z \geq \frac{-(1.5n - 200)}{\sqrt{0.0225n + 400}}\right)$$

dove Z è approssimativamente normale standard. Dalle Tabella A.1 in Appendice si può notare che $P(Z \geq 1.28) \approx 0.1$, quindi se il numero di autoveicoli n è tale che

$$\frac{200 - 1.5n}{\sqrt{0.0225n + 400}} \leq 1.28$$

ovvero quando $n \geq 117$ (si trova ricavando n, o per tentativi), vi è almeno 1 probabilità su 10 che il ponte subisca danni strutturali. □

Il teorema del limite centrale è illustrato dal Programma 6.1 del software del libro. Questo programma rappresenta la funzione di massa della somma di n variabili aleatorie i.i.d. che assumono i valori 0, 1, 2, 3 e 4. Quando lo si esegue è necessario inserire le probabilità dei cinque numeri, e il valore desiderato di n. Le Figure 6.2(a)–(f) illustrano i grafici ottenuti per una fissata configurazione delle probabilità quando n vale 1, 3, 5, 10, 25 e 100.

Una delle più dirette applicazioni del teorema del limite centrale riguarda le variabili aleatorie binomiali. Siccome una binomiale X di parametri (n, p) rappresenta il numero di successi in n prove indipendenti, ciascuna con probabilità p di riuscita, possiamo scrivere

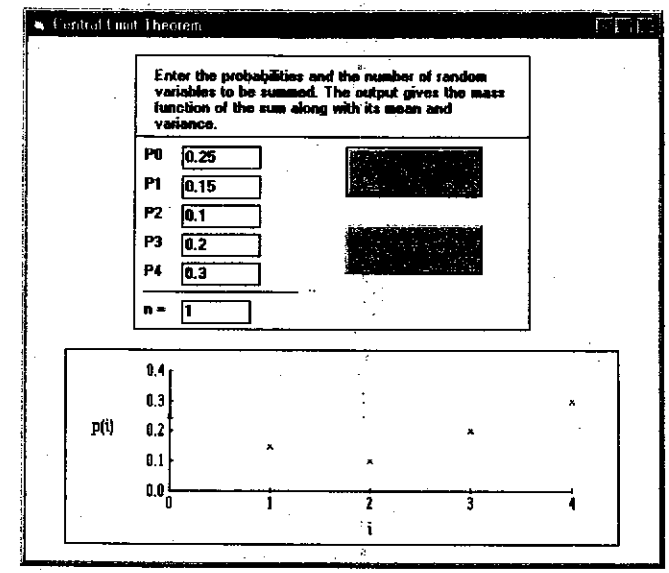
$$X = X_1 + X_2 + \cdots + X_n$$

dove

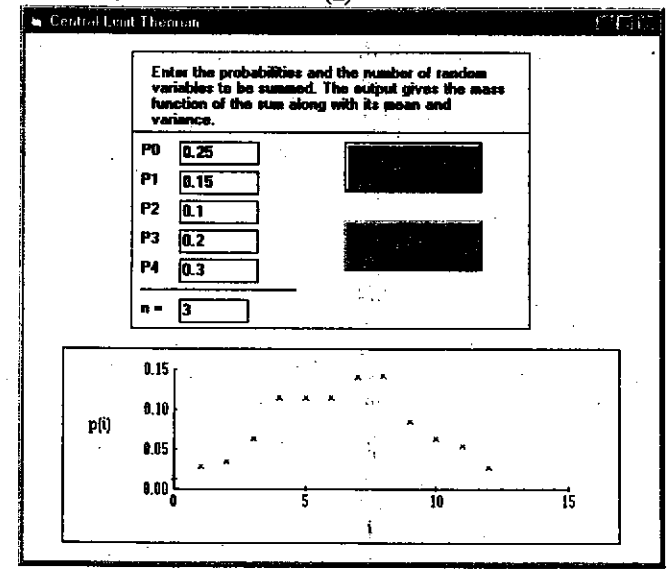
$$X_i := \begin{cases} 1 & \text{se l'}i\text{-esima prova ha successo} \\ 0 & \text{altrimenti} \end{cases}$$

Poiché, come sappiamo,

$$E[X_i] = p, \qquad \text{Var}(X_i) = p(1-p)$$

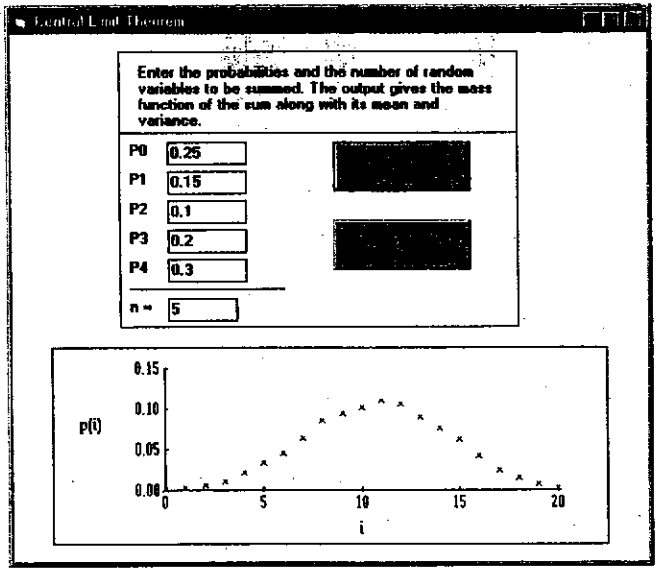


(a)

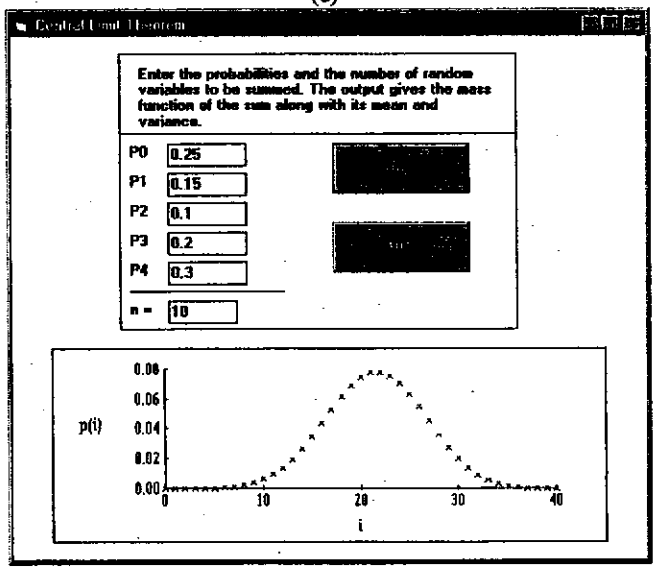


(b)

Figura 6.2 (a) $n = 1$, (b) $n = 3$, (c) $n = 5$, (d) $n = 10$, (e) $n = 25$, (f) $n = 100$.

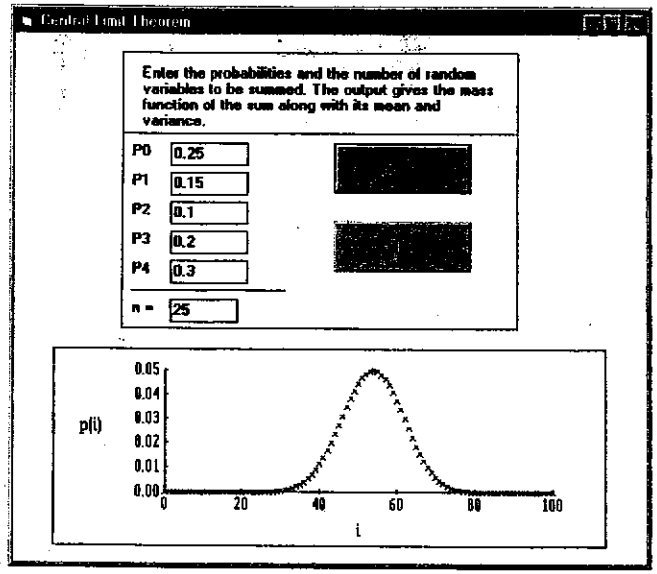


(c)

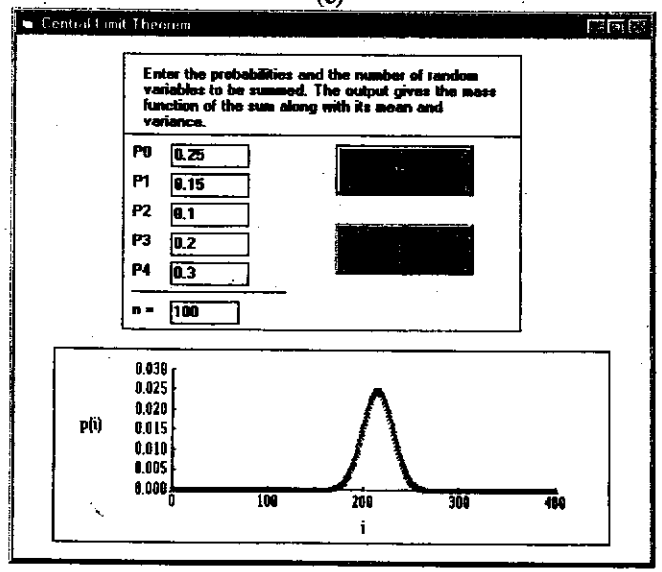


(d)

Figura 6.2 (continua)



(e)



(f)

Figura 6.2 (continua)

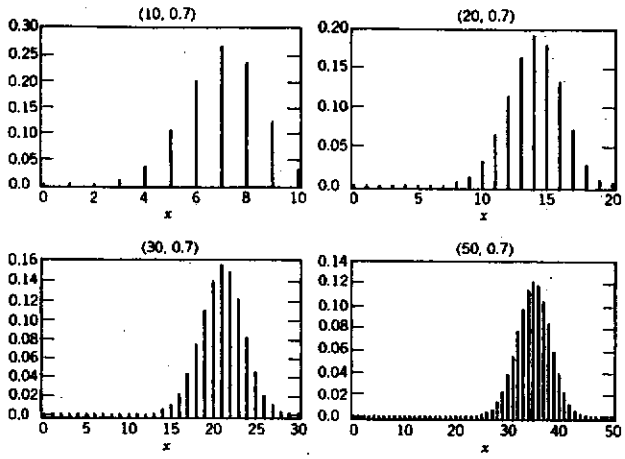


Figura 6.3 Funzioni di massa binomiali che convergono ad una densità normale.

segue dal teorema del limite centrale che, per n grande,

$$\frac{X - np}{\sqrt{np(1-p)}} \dot{\sim} \mathcal{N}(0,1) \tag{6.3.2}$$

ovvero vale una *approssimazione normale* delle variabili aleatorie binomiali. La Figura 6.3 illustra graficamente come la funzione di massa di una variabile aleatoria binomiale di parametri (n, p) tenda a divenire gaussiana al crescere di n.

Esempio 6.3.3. Il numero ideale di studenti per il primo anno di un corso in un certo college è di 150. Il college, sapendo dall'esperienza passata che solo il 30% degli studenti ammessi segue le lezioni, adotta la politica di accettare le iscrizioni di 450 studenti. Si calcoli la probabilità che più di 150 studenti del primo anno frequentino le lezioni.

Sia X il numero degli studenti che frequantano. Se assumiamo che ogni studente ammesso decida o meno di seguire le lezioni indipendentemente da tutti gli altri, allora X ha distribuzione binomiale di parametri $n = 450$ e $p = 0.3$. La probabilità richiesta è

$$P(X > 150) = \sum_{i=151}^{\infty} P(X = i)$$

Siccome vorremmo approssimare la variabile aleatoria discreta X con una normale, che è continua, è conveniente scrivere $P(X = i)$ come $P(i - 0.5 < X < i + 0.5)$

(questo passaggio si chiama *correzione di continuità*). In tal modo,

$$P(X > 150) = \sum_{i=151}^{\infty} P(i - 0.5 < X < i + 0.5) = P(X > 150.5)$$

E infatti l'approssimazione con il teorema del limite centrale fornisce un risultato più preciso usando 150.5 come estremo dell'intervallo.

$$\begin{aligned} P(X > 150.5) &= P\left(\frac{X - 450 \times 0.3}{\sqrt{450 \times 0.3 \times 0.7}} > \frac{150.5 - 450 \times 0.3}{\sqrt{450 \times 0.3 \times 0.7}}\right) \\ &\approx P(Z > 1.59) \approx 0.06 \end{aligned}$$

Quindi, solo il 6% circa[2] degli anni gli studenti che decidono di seguire superano il numero raccomandato di 150. □

È bene notare che a questo punto disponiamo di due diverse approssimazioni per le variabili aleatorie binomiali: quella di Poisson, che è valida quando n è grande e p piccolo, e quella normale, che (si può dimostrare) è valida quando $np(1-p)$ è grande (in effetti, per ottenere risultati accettabili, basta che $np(1-p)$ sia almeno 10).

6.3.1 Distribuzione approssimata della media campionaria

Sia $X_1, X_2, \ldots, X_n$ un campione proveniente da una popolazione di media μ e varianza σ^2. Vediamo come il teorema del limite centrale ci permette di approssimare la distribuzione della media campionaria,

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \tag{6.3.3}$$

Siccome il prodotto di una variabile aleatoria normale per una costante è ancora normale, ne segue che, quando n è grande, $\overline{X}$ è approssimativamente gaussiana. Poiché inoltre la media campionaria ha valore atteso μ e deviazione standard $\sigma/\sqrt{n}$, otteniamo che

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \dot{\sim} \mathcal{N}(0,1) \tag{6.3.4}$$

[2] Lo studente attento noterà che qui il numero di cifre di precisione che ci permettiamo di mantenere è più basso del solito. Ciò è dovuto al fatto che l'approssimazione con una normale, per quanto utile, non consente in genere una precisione molto alta. Per questo esempio, se si facessero i calcoli tenendo tutte le cifre decimali, si troverebbe che il valore di $P(X > 150)$ con la distribuzione binomiale è circa 0.0565, mentre con l'approssimazione normale, $P(X > 150.5) \approx 0.0554$ e $P(X > 150) \approx 0.0614$. Si vede da questi valori che (1) usare 150.5 come estremo fornisce un risultato più preciso di 150, e (2) per evitare di tenere più cifre significative di quelle esatte, conviene nel caso dell'approssimazione di una binomiale con una gaussiana limitare la precisione all'1% circa.

Esempio 6.3.4. Una popolazione formata da operai maschi, presenta dei pesi corporei (in libbre) di media 167 e deviazione standard 27.

(a) Se si seleziona un campione di 36 elementi, quanto vale circa la probabilità che la media campionaria dei loro pesi stia tra 163 e 171?

(b) E se si selezionano 144 operai?

(a) Sia Z una variabile aleatoria normale standard. Dal teorema del limite centrale segue che la media campionaria è approssimativamente normale con media 167 e deviazione standard $27/\sqrt{36} = 4.5$. Quindi

$$\begin{aligned} P(163 < \overline{X} < 171) &= P\left(\frac{163-167}{4.5} < \frac{\overline{X}-167}{4.5} < \frac{171-167}{4.5}\right) \\ &\approx P(-0.8889 < Z < 0.8889) \\ &= 2P(Z < 0.8889) - 1 \approx 0.63 \end{aligned}$$

(b) Con una ampiezza del campione di 144, $\overline{X}$ sarà approssimativamente normale di media 167 e deviazione standard $27/\sqrt{144} = 2.25$. Quindi

$$\begin{aligned} P(163 < \overline{X} < 171) &= P\left(\frac{163-167}{2.25} < \frac{\overline{X}-167}{2.25} < \frac{171-167}{2.25}\right) \\ &\approx P(-1.7778 < Z < 1.7778) \\ &= 2P(Z < 1.7778) - 1 \approx 0.92 \end{aligned}$$

Aumentando la numerosità del campione da 36 a 144, la probabilità richiesta è salita dal 63% al 92% circa. □

Esempio 6.3.5. Un astronomo vuole misurare la distanza di una stella lontana. Tuttavia, a causa dei disturbi dovuti all'atmosfera, le misurazioni effettuate dal suo osservatorio non restituiscono la distanza esatta d. Per questo motivo, egli ha deciso di fare una serie di misurazioni in condizioni diverse, e di usare la media campionaria come stimatore di d. È infatti convinto che misurazioni successive siano variabili aleatorie indipendenti, di media d, e deviazione standard 2 (l'unità di misura è l'anno-luce). Quante misurazioni deve effettuare per avere il 95% di probabilità che la sua stima sia accurata entro ± 0.5 anni-luce?

Se l'astronomo effettua un numero sufficientemente elevato n di misurazioni, allora la loro media campionaria $\overline{X}$ avrà distribuzione approssimativamente normale con media d e deviazione standard $2/\sqrt{n}$. La probabilità che questo stimatore cada entro $d \pm 0.5$ si ottiene come segue,

$$\begin{aligned} P(-0.5 < \overline{X} - d < 0.5) &= P\left(\frac{-0.5}{2/\sqrt{n}} < \frac{\overline{X}-d}{2/\sqrt{n}} < \frac{0.5}{2/\sqrt{n}}\right) \\ &\approx P(-\sqrt{n}/4 < Z < \sqrt{n}/4) \\ &= 2P(Z < \sqrt{n}/4) - 1 \end{aligned}$$

dove $Z \sim \mathcal{N}(0,1)$. Se ne deduce che n è un numero di osservazioni sufficiente solo se vale

$$2P(Z < \sqrt{n}/4) - 1 \geq 0.95$$

o equivalentemente,

$$P(Z < \sqrt{n}/4) \geq 0.975$$

Siccome $P(Z < 1.96) \approx 0.975$ si ottiene che n deve soddisfare

$$\sqrt{n}/4 \geq 1.96$$

e quindi si rendono necessarie almeno 62 osservazioni. □

6.3.2 Quando un campione è abbastanza numeroso?

Il teorema del limite centrale lascia aperta la questione di quanto grande debba essere la numerosità del campione n, affinché l'approssimazione normale sia valida. In effetti la risposta dipende dalla distribuzione da cui vengono campionati i dati. Ad esempio, se la distribuzione della popolazione è normale, allora $\overline{X}$ sarà a sua volta normale indipendentemente dall'ampiezza del campione (questo perché la distribuzione normale è riproducibile: si veda a pagina 176). Una buona regola empirica è che si può essere confidenti nella validità dell'approssimazione se n è almeno 30. Questo vuole dire che, per quanto "poco gaussiana" sia la distribuzione considerata, la media campionaria di un gruppo di dati di numerosità 30 risulta comunque approssimativamente normale. Si tenga presente comunque che in molti casi è possibile che questo accada anche per n molto più piccolo, e in effetti spesso $n = 5$ è sufficiente ad ottenere approssimazioni non troppo sbagliate. La Figura 6.4 presenta la distribuzione delle medie campionarie di una popolazione esponenziale, per n pari a 1, 5 e 10.

6.4 La varianza campionaria

Sia $X_1, X_2, \ldots, X_n$ un campione aleatorio, proveniente da una distribuzione di media μ e varianza σ^2. Sia $\overline{X}$ la sua media campionaria. In analogia con la Definizione 2.3.4 di pagina 25, introduciamo una seconda statistica.

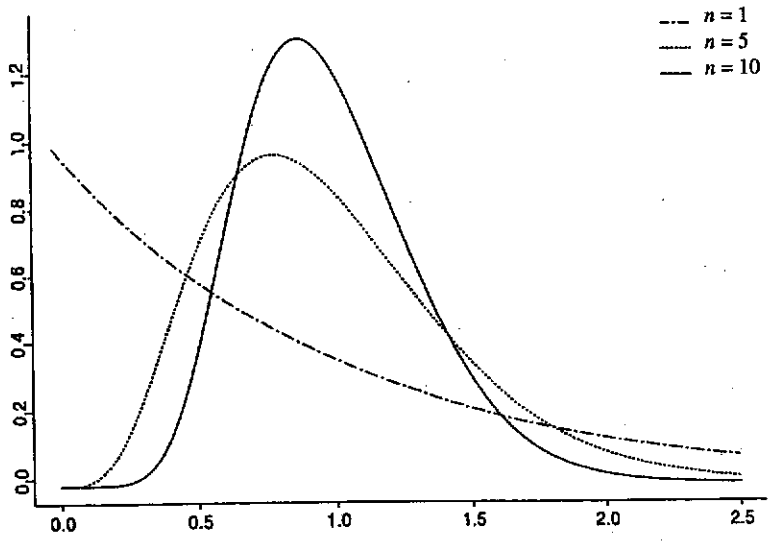


Figura 6.4 Densità della media aritmetica di n variabili aleatorie esponenziali di parametro unitario e indipendenti.

Definizione 6.4.1. La statistica S^2, definita da

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{6.4.1}$$

si dice *varianza campionaria*. La sua radice quadrata, $S = \sqrt{S^2}$ prende invece il nome di *deviazione standard campionaria*.

Volendo calcolare $E[S^2]$, sfruttiamo la Proposizione 2.3.1 di pagina 26 che afferma che per una qualsiasi n-upla di numeri $x_1, x_2, \ldots, x_n$,

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

dove $\bar{x} = \sum_{i=1}^{n} x_i/n$. Applicato a $X_1, X_2, \ldots, X_n$, questo enunciato implica che

$$S^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\Big) \tag{6.4.2}$$

ovvero che

$$(n-1)S^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$$

Prendendo il valore atteso di entrambi i membri di quest'ultima equazione, e ricordando che il momento secondo di una qualunque variabile aleatoria W si può ottenere come $E[W^2] = \mathrm{Var}(W) + E[W]^2$, deduciamo che

$$\begin{aligned}(n-1)E[S^2] &= E\Big[\sum_{i=1}^{n} X_i^2\Big] - E[n\overline{X}^2] \\ &= nE[X_1^2] - nE[\overline{X}^2] \\ &= n\,\mathrm{Var}(X_1) + nE[X_1]^2 - n\,\mathrm{Var}(\overline{X}) - nE[\overline{X}]^2 \\ &= n\sigma^2 + n\mu^2 - n\frac{\sigma^2}{n} - n\mu^2 \\ &= (n-1)\sigma^2\end{aligned}$$

da cui

$$E[S^2] = \sigma^2 \tag{6.4.3}$$

Il valore atteso della varianza campionaria coincide con la varianza della popolazione.

6.5 Le distribuzioni delle statistiche di popolazioni normali

In questa sezione ci restringiamo al caso in cui la distribuzione di popolazione sia di tipo normale.

Sia $X_1, X_2, \ldots, X_n$ un campione estratto da una distribuzione normale di media μ e varianza σ^2, intendendo con questo che tali variabili aleatorie sono tra loro indipendenti e $X_i \sim \mathcal{N}(\mu, \sigma^2)$, per $i = 1, 2, \ldots, n$. Denotiamo al solito con

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{e} \qquad S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{6.5.1}$$

la media e la varianza campionarie, rispettivamente. Ci proponiamo di determinare le loro distribuzioni.

6.5.1 La distribuzione della media campionaria

Siccome la somma di variabili aleatorie normali e indipendenti ha ancora distribuzione gaussiana, anche $\overline{X}$ è normale. La sua media e la sua varianza, come nel caso generale, sono μ e σ^2/n rispettivamente, e quindi

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \tag{6.5.2}$$

è una variabile aleatoria normale standard[3].

6.5.2 La distribuzione congiunta di $\overline{X}$ e S^2

In questa sezione, non solo deriviamo la distribuzione della varianza campionaria S^2, ma enunciamo anche il fatto fondamentale che $\overline{X}$ e S^2 sono variabili aleatorie indipendenti, con

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{6.5.3}$$

Per iniziare, si noti che, assegnati dei numeri $x_1, x_2, \ldots, x_n$, e posto $y_i := x_i - \mu$ per $i = 1, 2, \ldots, n$, dall'identità

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

si deduce che

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

Se applichiamo questa seconda identità ad un campione $X_1, X_2, \ldots, X_n$ di una popolazione normale con media μ e varianza σ^2, otteniamo che

$$\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} - \frac{n(\overline{X} - \mu)^2}{\sigma^2}$$

o equivalentemente,

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} + \left[\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}\right]^2 \tag{6.5.4}$$

Poiché le variabili aleatorie $(X_i - \mu)/\sigma$, per $i = 1, 2, \ldots, n$ sono normali standard indipendenti, il primo membro dell'Equazione (6.5.4) è una chi-quadro con n gradi di libertà. Per quanto detto nella Sezione 6.5.1, anche $\sqrt{n}(\overline{X} - \mu)/\sigma$ è normale standard, e quindi il suo quadrato è una chi-quadro con 1 grado di libertà. In conclusione, l'Equazione (6.5.4) esprime una χ^2_n come somma di due variabili aleatorie, una delle quali è una χ^2_1. Poiché sappiamo che la somma due chi-quadro indipendenti è un'altra chi-quadro i cui gradi di libertà sono la somma di quelli partenza, sembra decisamente plausibile che i due addendi al secondo membro della (6.5.4) siano una χ^2_{n-1} e una χ^2_1 indipendenti.

Anche se in questa sede non lo faremo, è possibile dimostrare la validità di questa nostra congettura, che è formalizzata nell'enunciato seguente.

[3] Si faccia attenzione a distinguere questa affermazione da quanto detto a pagina 215. In questo caso non vi sono approssimazioni: il risultato ottenuto è esatto, grazie all'ipotesi aggiuntiva che le X_i fossero gaussiane. Inoltre, quanto detto qui vale anche quando n è piccolo.

Teorema 6.5.1. Se $X_1, X_2, \ldots, X_n$ è un campione proveniente da una distribuzione normale di media μ e varianza σ^2, allora $\overline{X}$ e S^2 sono variabili aleatorie indipendenti. Inoltre, $\overline{X}$ è normale con media μ e varianza σ^2/n, e $(n-1)S^2/\sigma^2$ è una chi-quadro con $n-1$ gradi di libertà.

Questo teorema non solo ci fornisce le distribuzioni di $\overline{X}$ e S^2 per le popolazioni gaussiane, ma stabilisce anche l'importante proprietà – unica della distribuzione normale – che queste statistiche sono indipendenti. L'importanza di quanto detto emergerà con evidenza nei capitoli successivi.

Esempio 6.5.1. Il tempo impiegato da un microprocessore ad eseguire alcuni processi è una variabile aleatoria normale con media di 30 secondi e deviazione standard di 3 secondi. Se si osserva l'esecuzione di un campione di 15 processi, qual è la probabilità che la varianza campionaria risultante sia maggiore di 12?

Siccome l'ampiezza del campione è $n = 15$, e $\sigma^2 = 9$, scriviamo

$$\begin{aligned} P(S^2 > 12) &= P\left((n-1)\frac{S^2}{\sigma^2} > 14 \cdot \frac{12}{9}\right) \\ &\approx P(\chi^2_{14} > 18.67) \\ &\approx 1 - 0.8221 = 0.1779 \quad \square \end{aligned}$$

Il seguente corollario del Teorema 6.5.1 sarà di una certa utilità nei prossimi capitoli.

Corollario 6.5.2. Sia $X_1, X_2, \ldots, X_n$ un campione proveniente da una popolazione gaussiana di media μ. Se $\overline{X}$ e S^2 denotano la media e la varianza campionaria, allora

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \tag{6.5.5}$$

Quindi, se si normalizza $\overline{X}$ sottraendo la sua media μ e dividendo per la sua deviazione standard $\sigma/\sqrt{n}$, si ottiene una normale standard (è il risultato della Sezione 6.5.1). Se invece si divide per $S/\sqrt{n}$, si ha una distribuzione t con $n-1$ gradi di libertà.

Dimostrazione. Si ricordi che la t di Student con m gradi di libertà è, per la Definizione 5.8.2, la distribuzione del rapporto

$$\frac{Z}{\sqrt{\chi^2_m/m}}$$

dove $Z \sim \mathcal{N}(0,1)$, χ^2_m è una chi-quadro con m gradi di libertà, e queste due variabili aleatorie sono prese indipendenti. Allora, usando il fatto che

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \qquad (n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$$

e inoltre che queste due statistiche sono indipendenti per il Teorema 6.5.1, si ottiene che

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sqrt{\frac{\sigma^2}{S^2}\frac{n-1}{n-1}} = \frac{\overline{X}-\mu}{S/\sqrt{n}}$$

è una t di Student con $n-1$ gradi di libertà. □

6.6 Campionamento da insiemi finiti

Consideriamo una popolazione di N elementi. Con il concetto di campione aleatorio (di numerosità n) estratto da questa popolazione, si intende la scelta di un sottoinsieme di n elementi, fatta in modo tale che tutti i $\binom{N}{n}$ sottoinsiemi candidati abbiano le stesse probabilità di essere selezionati. Per esempio, se la popolazione di partenza consiste dei tre elementi a, b e c, un campione casuale di 2 elementi è un sottoinsieme scelto con pari probabilità tra $\{a,b\}$, $\{a,c\}$ e $\{b,c\}$. Un sottoinsieme casuale può essere individuato in pratica scegliendo uno alla volta i suoi elementi: il primo con pari probabilità tra gli N possibili, il secondo con pari probabilità tra gli $N-1$ restanti, e così via.

Supponiamo ora che alcuni elementi della popolazione di partenza abbiano una certa caratteristica, e denotiamo con p la frazione di questi rispetto al totale. Vi sono complessivamente pN elementi che posseggono questa caratteristica e $(1-p)N$ che non ce l'hanno. Selezioniamo un campione casuale di ampiezza n, e dopo avere numerato i suoi elementi, poniamo, per i che va da 1 a n:

$$X_i := \begin{cases} 1 & \text{se l'elemento } i \text{ del campione possiede la caratteristica} \\ 0 & \text{altrimenti} \end{cases}$$

Consideriamo la somma di queste variabili aleatorie,

$$X := X_1 + X_2 + \cdots + X_n$$

Siccome ognuna delle X_i contribuisce con 1 o con 0 alla somma, a seconda che l'elemento i possieda la caratteristica saliente o meno, X conta quanti sono in tutto quelli che la possiedono. Inoltre la media campionaria

$$\overline{X} = \frac{X}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

è pari alla frazione degli elementi del campione che mostrano tale caratteristica.

Passiamo ora ad analizzare le probabilità associate alle statistiche X e $\overline{X}$. Per cominciare, si noti che, siccome ciascuno degli N elementi di partenza ha le stesse possibilità di essere selezionato come membro i-esimo del campione, si ottiene

$$P(X_i = 1) = \frac{pN}{N} = p$$

Da cui ovviamente segue che

$$P(X_i = 0) = 1 - p$$

Le X_i sono variabili aleatorie di Bernoulli di parametro p.

È bene notare che $X_1, X_2, \ldots, X_n$ non sono indipendenti. Nonostante infatti sia p la probabilità che nella seconda selezione capiti un elemento con la caratteristica,

$$P(X_2 = 1) = p$$

ciò è vero solo se non si sa nulla di cosa sia successo nelle altre estrazioni. Supponendo ad esempio di sapere che $X_1 = 1$, ovvero che nella prima è stato selezionato un elemento tra i pN con la caratteristica, è chiaro che

$$P(X_2 = 1 | X_1 = 1) = \frac{pN-1}{N-1}$$

perché nella popolazione restano $N-1$ elementi, di cui $nP-1$ con la caratteristica. In maniera del tutto analoga, se si sa che $X_1 = 0$,

$$P(X_2 = 1 | X_1 = 0) = \frac{pN}{N-1}$$

Perciò il sapere se il primo membro selezionato per entrare a fare parte del campione abbia la caratteristica, modifica le probabilità per quelli successivi. Tuttavia, se la numerosità della popolazione N è molto grande rispetto a quella del campione n, questa variazione nelle probabilità sarà in ogni caso molto piccola. Per fare un esempio, se $N = 1\,000$ e $p = 0.4$, si ottengono le probabilità

$$P(X_2 = 1 | X_1 = 1) = \frac{399}{999} \approx 0.3994$$

$$P(X_2 = 1 | X_1 = 0) = \frac{400}{999} \approx 0.4004$$

entrambe molto vicine a

$$P(X_2 = 1) = 0.4$$

In effetti è possibile dimostrare che quando l'ampiezza della popolazione N è molto maggiore di quella del campione n, allora $X_1, X_2, \ldots, X_n$ sono approssimativamente indipendenti. Siccome la somma di bernoulliane indipendenti e identicamente distribuite è una variabile aleatoria binomiale, ne segue – sempre nell'ipotesi che N sia grande rispetto a n – che $X := \sum_i X_i$ è approssimativamente distribuita come una binomiale di parametri n e p.

Osservazione 6.6.1. Per la precisione, X è una variabile aleatoria ipergeometrica di parametri pN, $(1-p)N$ e n (si veda la Sezione 5.3). Quanto detto sopra implica che questo tipo di variabili aleatorie possono essere approssimate con binomiali quando il numero di elementi scelti è piccolo rispetto al numero degli elementi di partenza.

Da qui in poi supporremo sempre che la popolazione sia molto numerosa rispetto al campione estratto, e che la distribuzione di X sia binomiale.

La media e la varianza di X sono determinate dalle Equazioni (5.1.5) e (5.1.6) di pagina 151:

$$E[X] = np \qquad \text{e} \qquad \text{Var}(X) = np(1-p)$$

Poiché inoltre $\overline{X} = X/n$, si ottiene che

$$E[\overline{X}] = E[X]/n = p \tag{6.6.1}$$

e che

$$\text{Var}(\overline{X}) = \text{Var}(X)/n^2 = p(1-p)/n \tag{6.6.2}$$

Esempio 6.6.1. Supponiamo che alle prossime elezioni, il 45% della popolazione favorisca un certo candidato. Si seleziona un campione di 200 persone da intervistare. Si trovino

(a) valore atteso e deviazione standard del numero di intervistati che preferiscono quel candidato;

(b) la probabilità che essi siano più della metà degli interpellati.

(a) Detto X il numero di intervistati che voterà per il candidato considerato, la sua media e la sua deviazione standard sono

$$E[X] = 200 \times 0.45 = 90, \qquad \sqrt{\text{Var}(X)} = \sqrt{200 \times 0.45 \times 0.55} \approx 7.0356$$

(b) Poiché X è binomiale di parametri 200 e 0.45, il Programma 5.1 fornisce la soluzione

$$P(X \geq 101) = 1 - P(X \leq 100) \approx 0.0681$$

Se per qualche ragione il software non fosse disponibile, con l'approssimazione normale della distribuzione binomiale e la Tabella A.1 in Appendice, si trova che

$$\begin{aligned} P(X \geq 101) &= P(X \geq 100.5) && \text{correzione di continuità} \\ &\approx P\left(\frac{X - E[X]}{\sqrt{\text{Var}(X)}} \geq \frac{100.5 - 90}{7.0356}\right) \\ &\approx 1 - \Phi(1.49) \\ &\approx 1 - 0.9319 = 0.0681 && \text{usando la Tabella A.1} \end{aligned}$$

Si noti che abbiamo arrotondato $(100.5 - 90)/7.0356$ a due sole cifre decimali per ottenere un valore $x = 1.49$ il cui corrispondente valore di $\Phi(x)$ fosse presente nella Tabella A.1. A questo punto, anche se il risultato finale che troviamo, 0.0681, è corretto fino alla terza cifra significativa, questa non può essere che una coincidenza, e tenendo conto delle approssimazioni fatte, è più serio considerare 0.068 o addirittura 0.07 il risultato finale. A riprova di ciò, se teniamo un maggior numero di cifre decimali, troviamo $x = 1.4924$. Usando a questo punto il Programma 5.5a e non la Tabella A.1, per calcolare $\Phi(x)$, otteniamo che $1 - \Phi(1.4924) \approx 0.0678$. □

Anche quando gli elementi della popolazione possono avere più di due valori possibili, è ancora vero che i dati campionari possono essere pensati come variabili aleatorie indipendenti, e distribuite come la popolazione. È precisamente da questa considerazione che discende la Definizione 6.1.1

Esempio 6.6.2. Secondo il dipartimento dell'agricoltura statunitense, la nazione con il più elevato consumo pro-capite di carne di maiale è la Danimarca. Nel 1994 ad esempio, il consumo annuale per persona è stato una variabile aleatoria di media 147 e deviazione standard 62 (in libbre). Selezionando in maniera casuale 25 Danesi, qual è la probabilità che la media campionaria del loro consumo del 1994 abbia superato le 150 libbre?

Se per i che va da 1 a 25, denotiamo con X_i il consumo di carne di maiale durante tutto il 1994 del membro i-esimo del campione, la probabilità richiesta è data da

$$P\left(\frac{X_1 + X_2 + \cdots + X_n}{25} > 150\right) = P(\overline{X} > 150)$$

dove $\overline{X}$ è la media campionaria dei 25 dati. Siccome le X_i possono essere pensate come variabili aleatorie indipendenti di media 147 e deviazione standard 62, si deduce dal teorema del limite centrale che la loro media campionaria sarà approssimativamente normale, con media 147 e deviazione standard $62/5 = 12.4$. Così, con Z che indica una variabile aleatoria normale standard, abbiamo

$$\begin{aligned} P(\overline{X} > 150) &= P\left(\frac{\overline{X} - 147}{12.4} > \frac{150 - 147}{12.4}\right) \\ &\approx P(Z > 0.242) \approx 0.404 \quad \square \end{aligned}$$

Problemi

1. È data una popolazione con distribuzione seguente:

$$P(X = 0) = 0.2, \qquad P(X = 1) = 0.3, \qquad P(X = 2) = 0.5$$

Determina la funzione di massa di probabilità della media campionaria di un campione casuale $X_1, X_2, \ldots, X_n$ proveniente da questa popolazione e tracciane il grafico, quando **(a)** $n = 2$ e **(b)** $n = 3$. In entrambi i casi calcola anche media e varianza di $\overline{X}$.

2. Si tirano 10 dadi non truccati. Determina approssimativamente quanto vale la probabilità che la somma dei loro punteggi sia compresa tra 30 e 40 inclusi.

3. Calcola approssimativamente la probabilità che la somma di 16 variabili aleatorie indipendenti e uniformi su $(0, 1)$ sia superiore a 10.

4. La roulette di un casinò ha 38 settori, numerati con 0, 00, e da 1 a 36. Scommettendo 1 su un certo numero, si vince 35 se quel numero esce, e si perde 1 altrimenti. Supponendo di continuare a scommettere in questo modo, determina approssimativamente la probabilità di stare vincendo: **(a)** dopo 34, **(b)** dopo 1 000, e **(c)** dopo 100 000 scommesse. Puoi assumere che tutti i 38 risultati escano con la stessa probabilità, e che quelli di giocate diverse siano indipendenti.

5. L'ente che gestisce un tratto di autostrada conserva sale a sufficienza per eliminare un totale di 80 pollici di neve. Supponiamo che la quantità di neve che cade al giorno sia una variabile aleatoria di media 1.5 pollici e deviazione standard 0.3 pollici.

(a) Trova la probabilità approssimativa che il sale a disposizione basti per 50 giorni.

(b) Quali sono le ipotesi che hai assunto per rispondere al punto **(a)**?

(c) Ti sembra che tali ipotesi siano giustificate? Spiega brevemente.

6. Si prendono 50 numeri, che vengono arrotondati all'intero più vicino e poi sommati tutti. Se gli errori di arrotondamento individuali sono variabili aleatorie indipendenti e uniformi su $(-0.5, 0.5)$, quanto vale approssimativamente la probabilità che la somma così ottenuta differisca da quella esatta per più di 3 unità?

7. Un normale dado da gioco non truccato viene tirato ripetutamente, fino a che la somma di tutti i punteggi ottenuti non superi 400. Determina in maniera approssimata la probabilità che siano necessari più di 140 lanci.

8. Il numero di settimane di funzionamento di un certo tipo di batterie è una variabile aleatoria con media 5 e deviazione standard 1.5. Quando una batteria si esaurisce, viene immediatamente sostituita con una nuova. Calcola approssimativamente la probabilità che in un anno si debbano impiegare 13 o più batterie.

9. Il tempo di vita di un certo componente elettrico è una variabile aleatoria di media 100 ore e deviazione standard 20 ore. Se si provano 16 componenti di questo tipo, quanto vale la probabilità che la media campionaria delle loro durate sia **(a)** minore di 104; **(b)** compresa tra 98 e 104?

10. Un produttore di sigarette dichiara che la quantità di nicotina contenuta in ciascuna delle sue sigarette è una variabile aleatoria di media 2.2 mg e deviazione standard 0.3 mg. Tuttavia, analizzando un campione casuale di 100 sigarette si trova una media campionaria di 3.1 mg. Se le affermazioni della ditta fossero veritiere, quale sarebbe approssimativamente la probabilità di trovare una media campionaria così elevata (3.1 o più)?

11. Il tempo di vita (in ore) di un tipo di lampadine ha valore atteso 500 e deviazione standard 80. Preso un campione di ampiezza n, e detta $\overline{X}$ la media campionaria dei rispettivi tempi di vita, quanto vale la probabilità che $\overline{X}$ sia maggiore di 525? Calcola un valore che approssimi la risposta **(a)** quando $n = 4$; **(b)** quando $n = 16$; **(c)** quando $n = 36$; **(d)** quando $n = 64$.

12. Un docente sa dall'esperienza passata che il punteggio all'esame finale degli studenti del suo corso è distribuito con media 77 e deviazione standard 15. Attualmente egli ha due classi diverse, una di 64 e una di 25 studenti.

(a) Quanto vale la probabilità che la media aritmetica dei punteggi (o punteggio medio) della classe di 25 studenti sia compresa tra 72 e 82?

(b) E per l'altra classe?

(c) Quanto vale approssimativamente la probabilità che il punteggio medio della classe da 25, superi quello della classe da 64?

(d) Supponiamo che i punteggi medi delle due classi siano 76 e 83. Quale delle due classi è più probabile abbia ottenuto il punteggio di 83?

13. Sia X una variabile aleatoria binomiale di parametri $n = 150$ e $p = 0.6$. Calcola il valore di $P(X \leq 80)$:

(a) in modo esatto;

(b) con l'approssimazione normale;

(c) con l'approssimazione normale ma senza la correzione di continuità.

14. I circuiti integrati prodotti da un certo impianto sono difettosi con probabilità di 0.25, tutti indipendentemente l'uno dall'altro. Se si testa un campione di 1 000 pezzi, con che probabilità se ne troveranno meno di 200 di difettosi?

15. Una squadra di basket ha di fronte una stagione con 60 incontri. Di queste partite, 32 sono con squadre di livello A e 28 con squadre di livello B. I risultati delle partite sono tutti indipendenti; le probabilità di vittoria sono del 50% con una squadra di livello A, e del 70% negli altri casi. Sia X il numero totale di vittorie ottenute durante la stagione.

(a) La distribuzione di X è binomiale?

Siano X_A e X_B il numero di vittorie contro squadre di livello A e B rispettivamente.

(b) Che tipo di variabili aleatorie sono X_A e X_B?

(c) Quale relazione lega X_A, X_B e X?

(d) Quanto vale approssimativamente la probabilità che vi siano almeno 40 vittorie?

16. Giustifica con un ragionamento basato sul teorema del limite centrale, il fatto che una variabile aleatoria di Poisson di media λ si possa approssimare con una normale di media e varianza entrambe pari a λ, quando questo parametro è grande. Se X è una poissoniana di media 100, determina in modo esatto la probabilità che $X \leq 116$ e confrontala con i risultati ottenuti con l'approssimazione normale, con e senza la correzione di continuità. La convergenza delle variabili aleatorie di Poisson alla distribuzione gaussiana è illustrata in Figura 6.5.

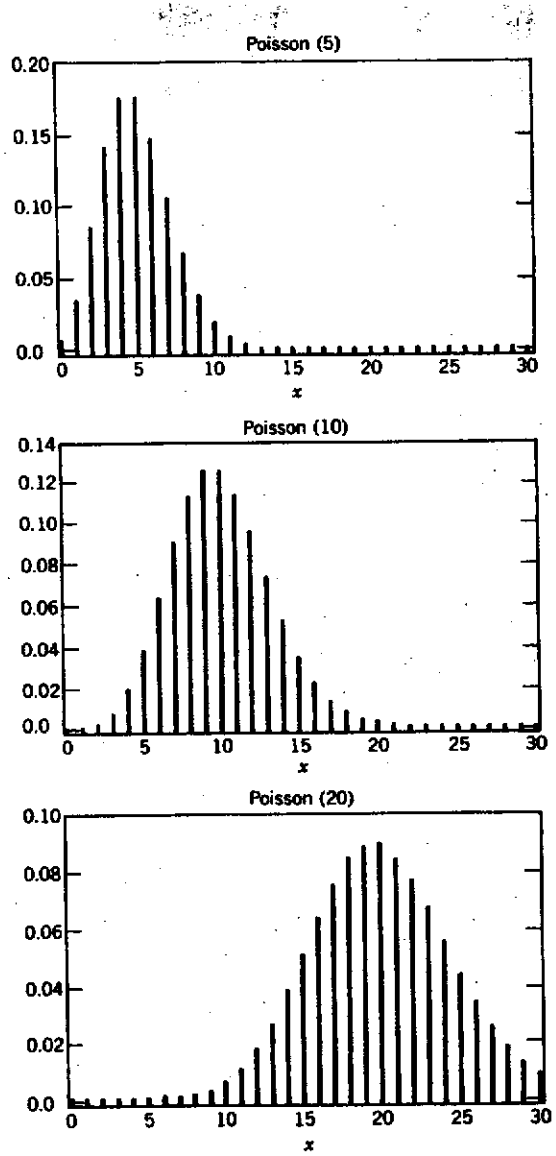


Figura 6.5 Funzioni di massa di probabilità poissoniane.

17. Usa il software abbinato al testo per calcolare in maniera esatta $P(X \leq 10)$, dove X è una variabile aleatoria binomiale di parametri $(100, 0.1)$. Confronta il valore ottenuto con le sue approssimazioni di Poisson e normale. Nel caso della approssimazione normale, scrivi la probabilità richiesta come $P(X < 10.5)$, utilizzando così la correzione di continuità.

18. La temperatura alla quale un termostato scatta, ha distribuzione gaussiana con varianza σ^2. Considerato che lo strumento viene testato 5 volte, calcola

(a) $P(S^2/\sigma^2 \leq 1.8)$

(b) $P(0.85 \leq S^2/\sigma^2 \leq 1.15)$

dove S^2 è la varianza campionaria dei cinque dati misurati.

19. Con riferimento al Problema 18, a quante prove occorre sottoporre il termostato affinché la probabilità del punto **(a)** sia almeno del 95%?

20. Consideriamo due campioni indipendenti – il primo ha ampiezza 10 e proviene da una popolazione normale di varianza 4, il secondo ha ampiezza 5 e proviene da una popolazione normale di varianza 2. Calcola la probabilità che la varianza campionaria del secondo campione sia maggiore di quella del primo. (*Suggerimento:* Collega le quantità cercate ad una distribuzione F.)

21. Il 12% della popolazione mondiale è mancina. Trova la probabilità che in un campione aleatorio di 100 persone vi sia un numero di mancini tra i 10 e i 14.

22. La tabella seguente riporta la percentuale di adulti soggetta ad alcune abitudini negative per la salute. Supponiamo di selezionare un campione di 300 maschi. Determina approssimativamente la probabilità che

(a) quelli che fanno colazione raramente siano almeno 150;

(b) i fumatori siano meno di 100.

	Dorme meno di sei ore per notte	**Fuma**	**Fa colazione raramente**	**È sovrappeso del 20% o più**
Maschi	22.7	28.4	45.4	29.6
Femmine	21.4	22.8	42.0	25.6

Fonte: U.S. National Center for Health Statistics, Health Promotion and Disease Prevention, 1990.

23. Osserva la tabella del Problema 22. Supponiamo di selezionare un campione di 300 femmine. Determina approssimativamente la probabilità che

(a) quelle sovrappeso del 20% o più siano almeno 60;

(b) quelle che dormono meno di sei ore per notte siano meno di 50.

24. Osserva la tabella del Problema 22. Supponiamo di selezionare un campione formato da 300 maschi e 300 femmine. Determina approssimativamente la probabilità che nel campione, le femmine che fanno colazione raramente siano più dei maschi.

25. La tabella seguente si riferisce a dati del 1989. Essa suddivide i lavoratori a tempo pieno a seconda del sesso e della categoria di reddito annuale. Supponiamo di selezionare un campione di 1 000 uomini e 1 000 donne che nel 1989 avevano un lavoro a tempo pieno. Usa la tabella per calcolare le probabilità che

(a) almeno la metà delle donne guadagnasse meno di \$ 20 000;

(b) più della metà degli uomini guadagnasse \$ 20 000 o più;

(c) più di metà sia degli uomini, sia delle donne, guadagnasse \$ 20 000 o più;

(d) le donne che percepivano almeno \$ 25 000 fossero 250 o meno;

(e) gli uomini che percepivano \$ 50 000 o più fossero almeno 200;

(f) nella categoria tra \$ 20 000 e \$ 24 999 vi fossero più donne che uomini.

Intervallo di reddito	Percentuale delle donne	Percentuale degli uomini
\$ 4 999 o meno	2.8	1.8
da \$ 5 000 a \$ 9 999	10.4	4.7
da \$ 10 000 a \$ 19 999	41.0	23.1
da \$ 20 000 a \$ 24 999	16.5	13.4
da \$ 25 000 a \$ 49 999	26.3	42.1
\$ 50 000 e oltre	3.0	14.9

Fonte: U.S. Department of Commerce, Bureau of the Census.

26. Nel 1995 il 14.9% della forza lavoro era iscritta a qualche sindacato. Se in quell'anno si fossero scelti a caso 5 lavoratori, quale sarebbe stata la probabilità che nessuno di essi avesse un sidacato? Confronta la tua risposta con quella che avresti dato per l'anno 1945, quando i lavoratori con un sindacato hanno toccato il massimo storico del 35.5%.

27. In una prova di matematica proposta di recente in tutte le scuole superiori di San Francisco, la media e la deviazione standard dei punteggi di tutti gli studenti sono stati 517 e 120. Trova la probabilità approssimata che un campione di 144 studenti abbia un punteggio medio che superi **(a)** 507; **(b)** 517; **(c)** 537; **(d)** 550.

28. Il reddito medio dei neolaureati in ingegneria chimica è di \$ 35 600, con una deviazione standard di \$ 3 200. Determina la probabilità approssimata che un campione di 12 si essi presenti uno stipendio medio superiore a \$ 37 000.

7 Stima parametrica

Contenuto

7.1 Introduzione

Consideriamo un campione aleatorio $X_1, X_2, \ldots, X_n$ estratto da una distribuzione F_θ che dipende da un vettore di parametri incogniti θ. Potrebbe ad esempio trattarsi di variabili aleatorie di Poisson, delle quali ignoriamo il valore di λ; oppure potremmo avere a che fare con un campione normale, della cui distribuzione ignoriamo media e varianza. Mentre quando si fa della probabilità è normale suppore che le distribuzioni in gioco siano completamente note, in statistica è vero il contrario, e il problema centrale è quello di dire qualcosa (ovvero *fare dell'inferenza*) sui parametri sconosciuti, usando i dati osservati.

Nella Sezione 7.2 presentiamo il metodo della *massima verosimiglianza*, per individuare degli stimatori dei parametri incogniti. Quelli ottenuti in tal modo sono detti *stimatori puntuali*, perché forniscono un singolo valore come stima di θ. Nella Sezione 7.3 affrontiamo invece il problema degli stimatori non puntuali – o di intervallo – meglio noti come *intervalli di confidenza*. Con questi strumenti siamo in grado di ottenere non un singolo punto, come stima del parametro θ, ma un intervallo di valori plausibili per θ. A ciascuno di questi intervalli è associato un *livello di confidenza* nei

confronti dell'ipotesi che θ vi appartenga. Questi concetti vengono illustrati nel caso della media di una distribuzione gaussiana di cui sia nota la varianza; successivamente ci rivolgiamo ad una varietà di altri problemi di stima non puntuale. In particolare vengono determinati gli intervalli di confidenza: per la media di una normale con la varianza incognita (nella Sezione 7.3.1); per la varianza di una normale (nella Sezione 7.3.2); per la differenza delle medie di due normali (nella Sezione 7.4), sia nel caso che le varianze siano note, sia nel caso che siano identiche ma incognite; per il parametro delle distribuzioni di Bernoulli (nella Sezione 7.5); ed infine per la media di una esponenziale (nella Sezione facoltativa 7.6).

Con la Sezione facoltativa 7.7 ritorniamo al problema di individuare i possibili stimatori puntuali dei parametri incogniti, e spieghiamo come valutare la bontà di uno stimatore in termini del suo errore quadratico medio. Discutiamo poi del *bias* degli stimatori e analizziamo la sua relazione con l'errore quadratico medio.

Nella Sezione facoltativa 7.8 affrontiamo il problema di determinare la stima di un parametro sfruttando le informazioni a priori che possono essere disponibili. Questo è il cosiddetto approccio *bayesiano*, che richiede che prima di osservare i dati, si disponga di alcune informazioni su θ. Tali conoscenze si suppongono essere nella forma di una distribuzione di probabilità sui possibili valori di θ. In questo contesto è possibile calcolare lo *stimatore bayesiano*, che è quello per cui il valore atteso del quadrato della distanza da θ è minimo.

7.2 Stimatori di massima verosimiglianza

Una qualunque statistica il cui scopo sia quello di dare una stima di un parametro θ si dice *stimatore* di θ; gli stimatori sono quindi variabili aleatorie. Il valore deterministico assunto da uno stimatore è detto invece *stima*. Ad esempio, come avremo modo di vedere, la media campionaria $\overline{X} := \sum_i X_i/n$ di un campione normale $X_1, X_2, \ldots, X_n$ costituisce lo stimatore abituale della media μ della distribuzione.

Consideriamo delle variabili aleatorie $X_1, X_2, \ldots, X_n$, la cui distribuzione congiunta sia nota a meno di un parametro incognito θ. Un problema di interesse consiste nello stimare θ usando i valori che vengono assunti da queste variabili aleatorie. Per esemplificare, potremmo immaginare che le X_i siano variabili aleatorie esponenziali e indipendenti, tutte di media θ incognita. In questo caso la loro densità congiunta sarebbe data da

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n) \\ &= \frac{1}{\theta} e^{-x_1/\theta} \frac{1}{\theta} e^{-x_2/\theta} \cdots \frac{1}{\theta} e^{-x_n/\theta}, \quad x_i > 0, \quad i = 1, \ldots, n \\ &= \frac{1}{\theta^n} \exp\left\{ -\sum_{i=1}^{n} \frac{x_i}{\theta} \right\}, \qquad\qquad x_i > 0, \quad i = 1, \ldots, n \end{aligned}$$

e il nostro obiettivo consisterebbe nello stimare θ partendo dai valori osservati $X_1, X_2, \ldots, X_n$.

Vi è una classe particolare di stimatori, detti *stimatori di massima verosimiglianza*[1], che è largamente utilizzata in statistica. Uno stimatore di questo tipo si ottiene con il ragionamento seguente. Denotiamo con $f(x_1, x_2, \ldots, x_n|\theta)$ la funzione di massa congiunta di $X_1, X_2, \ldots, X_n$ oppure la loro densità congiunta, a seconda che siano variabili aleatorie discrete o continue. Poiché stiamo supponendo che θ sia una incognita, mostriamo esplicitamente che f dipende da θ. Se interpretiamo $f(x_1, x_2, \ldots, x_n|\theta)$ come la verosimiglianza (o plausibilità, o credibilità, in un italiano più diretto) che si realizzi la n-upla di dati $x_1, x_2, \ldots, x_n$, quando θ è il vero valore assunto dal parametro, sembra ragionevole adottare come stima di θ quel valore che rende massima la verosimiglianza per i dati osservati. In altri termini, la stima di massima verosimiglianza $\hat{\theta}$ è definita come il valore di θ che rende massima $f(x_1, x_2, \ldots, x_n|\theta)$, quando i valori osservati sono $x_1, x_2, \ldots, x_n$. La funzione $f(x_1, x_2, \ldots, x_n|\theta)$ è detta funzione di *likelihood* (il termine inglese per verosimiglianza).

Nel calcolare il valore di θ che massimizza f, conviene spesso usare il fatto che le due funzioni $f(x_1, x_2, \ldots, x_n|\theta)$ e $\log[f(x_1, x_2, \ldots, x_n|\theta)]$ assumono il massimo in corrispondenza dello stesso valore di θ (perché?). Quindi è possibile ottenere $\hat{\theta}$ anche massimizzando $\log[f(x_1, x_2, \ldots, x_n|\theta)]$, che è detta funzione di log-*likelihood*.

Esempio 7.2.1 (Stimatore di massima verosimiglianza del parametro di una bernoulliana). Supponiamo che vengano realizzate n prove indipendenti, ciascuna con probabilità p di successo. Qual è lo stimatore di massima verosimiglianza per p?

I dati a disposizione consistono nei valori di $X_1, X_2, \ldots, X_n$, dove

$$X_i = \begin{cases} 1 & \text{se la prova } i\text{-esima ha successo} \\ 0 & \text{altrimenti} \end{cases}$$

La distribuzione delle X_i è determinata da

$$P(X_i = 1) = p = 1 - P(X_i = 0)$$

o, in maniera più compatta,

$$P(X_i = k) = p^k (1-p)^{1-k}, \qquad k = 0, 1 \tag{7.2.1}$$

[1] È di uso molto comune l'acronimo MLE, dall'inglese *maximum likelihood estimator*, [*N.d.T.*]

Quindi, sfruttando l'indipendenza delle prove, la likelihood (ovvero la funzione di massa congiunta) del campione è data da

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|p) &:= P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n|p) \\ &= p^{x_1}(1-p)^{1-x_1} \cdots p^{x_n}(1-p)^{1-x_n} \\ &= p^{\sum_i x_i}(1-p)^{n-\sum_i x_i}, \qquad x_i = 0, 1, \quad i = 1, \ldots, n \end{aligned} \tag{7.2.2}$$

Per determinare il valore di p che massimizza questa funzione, prima prendiamo i logaritmi,

$$\log f(x_1, x_2, \ldots, x_n|p) = \sum_{i=1}^{n} x_i \log p + \left(n - \sum_{i=1}^{n} x_i\right) \log(1-p)$$

quindi deriviamo rispetto a p

$$\frac{d}{dp} \log f(x_1, x_2, \ldots, x_n|p) = \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p}\left(n - \sum_{i=1}^{n} x_i\right)$$

ponendo il secondo termine uguale a zero e risolvendo rispetto a p, otteniamo un'espressione per la stima $\widehat{p}$,

$$\frac{1}{\widehat{p}} \sum_{i=1}^{n} x_i = \frac{1}{1-\widehat{p}}\left(n - \sum_{i=1}^{n} x_i\right)$$

da cui

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Perciò lo stimatore di massima verosimiglianza di una distribuzione di Bernoulli di media incognita è dato da

$$d(X_1, X_2, \ldots, X_n) := \frac{1}{n} \sum_{i=1}^{n} X_i \tag{7.2.3}$$

Siccome $\sum_{i=1}^{n} X_i$ è il numero di prove che hanno avuto successo, si vede che lo stimatore di massima verosimiglianza di p coincide con la frazione di prove che hanno avuto successo. Per vedere una applicazione, supponiamo che ogni circuito di RAM[2] prodotto in un certo stabilimento sia – indipendentemente da tutti gli altri – accettabile con probabilità p. Se su un campione di 1 000 pezzi quelli accettabili sono 921, si ottiene che la stima di massima verosimiglianza per p è 0.921. □

[2] *Random Access Memory*, ovvero memoria ad accesso causale. Si tratta della memoria volatile principale di un qualunque personal computer.

Esempio 7.2.2 (Stimatore di massima verosimiglianza del parametro di una poissoniana). Supponiamo che $X_1, X_2, \ldots, X_n$ siano variabili aleatorie di Poisson indipendenti, ciascuna con valore atteso λ. Si determini lo stimatore di massima verosimiglianza per λ.

La funzione di likelihood è data da

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|\lambda) &= \frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \cdots \frac{\lambda^{x_n} e^{-\lambda}}{x_n!} \\ &= \frac{\lambda^{\sum_i x_i} e^{-n\lambda}}{x_1! \cdots x_n!} \end{aligned} \tag{7.2.4}$$

ovvero,

$$\log f(x_1, x_2, \ldots, x_n|\lambda) = \sum_{i=1}^{n} x_i \log \lambda - n\lambda - \log c$$

dove $c := x_1! \cdots x_n!$ non dipende da λ. Derivando si trova che

$$\frac{d}{d\lambda} \log f(x_1, x_2, \ldots, x_n|\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n$$

Uguagliando infine a zero questa espressione si ottiene una formula per la stima $\widehat{\lambda}$ in funzione delle osservazioni $x_1, x_2, \ldots, x_n$,

$$\widehat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

La stessa formula applicata al campione $X_1, X_2, \ldots, X_n$, fornisce lo stimatore desiderato.

$$d(X_1, X_2, \ldots, X_n) := \frac{1}{n} \sum_{i=1}^{n} X_i \tag{7.2.5}$$

Volendo citare un caso pratico, supponiamo che il numero di persone che ogni giorno entra in un negozio sia una variabile aleatoria di Poisson avente una certa media λ che vogliamo stimare. Se in 20 giorni il numero totale di persone entrate nel negozio è di 857, allora la stima di massima verosimiglianza per λ è $857/20 = 42.85$. Quindi stimiamo che in media ogni giorno entreranno 42.85 persone. □

Esempio 7.2.3. Nel 1998 a Berkeley in California, il numero di incidenti stradali in 10 giornate senza pioggia scelte a caso è stato di

4 0 6 5 2 1 2 0 4 3

Si usino questi dati per stimare per quell'anno la frazione di giornate senza pioggia con 2 incidenti o meno.

Siccome vi è un elevato numero di automobilisti, ciascuno dei quali ha solo una piccola probabilità di essere coinvolto in un incidente stradale, è ragionevole assumere che il numero di incidenti quotidiani sia una variabile aleatoria di Poisson. Visto che

$$\overline{X} = \frac{1}{10}\sum_{i=1}^{10} X_i = 2.7$$

si ottiene che la stima di massima verosimiglianza della media della poissoniana è 2.7. Siccome a lungo andare la frazione di giornate senza pioggia con 2 incidenti o meno sarà pari a $P(X \leq 2)$, dove X è il numero di incidenti stradali in un giorno, si ottiene che la stima desiderata è data da

$$(1 + 2.7 + (2.7)^2/2)e^{-2.7} \approx 0.4936$$

Quindi la nostra stima è che in poco meno della metà dei giorni senza pioggia vi siano fino a 2 incidenti stradali. □

Esempio 7.2.4 (Stimatore di massima verosimiglianza per una popolazione normale). Siano $X_1, X_2, \ldots, X_n$ variabili aleatorie normali e indipendenti, con media μ e deviazione standard σ, entrambe incognite. La densità congiunta, e quindi la likelihood, è data da

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|\mu, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \end{aligned} \tag{7.2.6}$$

La log-likelihood corrispondente è data da

$$\log f(x_1, x_2, \ldots, x_n|\mu, \sigma) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Per trovare le stime $\widehat{\mu}$ e $\widehat{\sigma}$ che contemporaneamente massimizzino la log-likelihood, occorre porre uguali a zero le due derivate parziali, e mettere a sistema le due equazioni trovate.

$$\frac{\partial}{\partial\mu}\log f(x_1, x_2, \ldots, x_n|\mu, \sigma) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

$$\frac{\partial}{\partial\sigma}\log f(x_1, x_2, \ldots, x_n|\mu, \sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2$$

da cui il sistema

$$\begin{cases} \dfrac{1}{\widehat{\sigma}^2}\displaystyle\sum_{i=1}^{n}(x_i-\widehat{\mu}) = 0 \\ -\dfrac{n}{\widehat{\sigma}} + \dfrac{1}{\widehat{\sigma}^3}\displaystyle\sum_{i=1}^{n}(x_i-\widehat{\mu})^2 = 0 \end{cases}$$

la cui risoluzione ci porta alle seguenti formule per le stime,

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\widehat{\sigma} = \left\{\frac{1}{n}\sum_{i=1}^{n}(x_i-\widehat{\mu})^2\right\}^{1/2}$$

Quindi, gli stimatori di massima verosimiglianza di μ e σ sono dati rispettivamente da

$$\overline{X} \quad \text{e} \quad \sqrt{\frac{1}{n}\sum_i (X_i - \overline{X})^2} \tag{7.2.7}$$

È bene notare che lo stimatore di massima verosimiglianza per la deviazione standard non coincide con la deviazione standard campionaria,

$$S := \sqrt{\frac{1}{n-1}\sum_i (X_i - \overline{X})^2} \tag{7.2.8}$$

in quanto nel primo si divide per $\sqrt{n}$ e nel secondo per $\sqrt{n-1}$. In ogni caso, per n non troppo piccolo questi due stimatori di σ saranno approssimativamente uguali. □

Esempio 7.2.5. Una legge dovuta a Kolmogorov sulla frammentazione dei corpi afferma che le dimensioni di una particella presa a caso tra quelle formatesi dalla frammentazione di un pezzo di minerale, hanno distribuzione lognormale. (Si ricorda che X si dice avere distribuzione lognormale se $\log X$ ha una distribuzione normale.) Questa legge, che è stata prima ottenuta empiricamente e poi dedotta teoricamente da Kolmogorov, è stata applicata a una varietà di studi di ingegneria. Ad esempio è stata usata nell'analisi delle dimensioni delle particelle d'oro facenti parti di una polvere d'oro. Una applicazione meno ovvia di questa legge riguarda lo studio del rilascio di energia presso le faglie sismiche[3].

Supponiamo che un campione di 10 granelli presi da una grossa pila di polvere metallica abbiano le seguenti lunghezze (in millimetri)

2.2 3.4 1.6 0.8 2.7 3.3 1.6 2.8 2.5 1.9

[3] C. Lomnitz, "Global tectonics and earthquake risk", *Developments in Geotectonics*, Elsevier, 1979.

Si stimi la percentuale di granelli nella pila la cui lunghezza è compresa tra 2 e 3 mm.

Se si prendono i logaritmi naturali dei 10 dati del campione, si ottiene un campione normale, nel nostro caso,

$$0.7885 \quad 1.2238 \quad 0.4700 \quad -0.2231 \quad 0.9933$$
$$1.1939 \quad 0.4700 \quad 1.0296 \quad 0.9163 \quad 0.6419$$

Poiché media e deviazione standard campionarie di questi dati sono

$$\bar{x} \approx 0.7504, \qquad s \approx 0.4351$$

si ottiene che il logaritmo naturale della lunghezza di un granello della pila è una variabile aleatoria normale di media e deviazione standard approssimativamente pari a 0.7504 e 0.4351. Allora, se X è la lunghezza di un granello preso a caso,

$$\begin{aligned} P(2 < X < 3) &= P(\log 2 < \log X < \log 3) \\ &= P\left(\frac{\log 2 - 0.7504}{0.4351} < \frac{\log X - 0.7504}{0.4351} < \frac{\log 3 - 0.7504}{0.4351}\right) \\ &\approx P(-0.1316 < Z < 0.8003) \\ &\approx \Phi(0.8003) - \Phi(-0.1316) \approx 0.3405 \quad \square \end{aligned}$$

In tutti gli esempi precedenti, lo stimatore di massima verosimiglianza della media della popolazione è risultato coincidere con la media campionaria $\overline{X}$. Per verificare che non sempre è così, si consideri l'esempio seguente.

Esempio 7.2.6 (Stimatore di massima verosimiglianza per la media di una distribuzione uniforme). Sia $X_1, X_2, \ldots, X_n$ un campione proveniente da una distribuzione uniforme sull'intervallo $(0, \theta)$, con θ incognita. La densità congiunta è data da

$$f(x_1, x_2, \ldots, x_n|\theta) = \begin{cases} \dfrac{1}{\theta^n} & 0 < x_i < \theta, \qquad i = 1, \ldots, n \\ 0 & \text{altrimenti} \end{cases} \tag{7.2.9}$$

Questa densità si massimizza scegliendo θ il più piccolo possibile. Siccome θ deve essere comunque maggiore di tutti i valori osservati x_i, ne segue che la più piccola scelta possibile per θ è $\max(x_1, x_2, \ldots, x_n)$. Lo stimatore di massima verosimiglianza per θ è quindi

$$\hat{\theta} = \max(X_1, X_2, \ldots, X_n) \tag{7.2.10}$$

da cui segue subito che lo stimatore di massima verosimiglianza della media della distribuzione (media che è pari a $\theta/2$, si veda l'Equazione (5.4.3) di pagina 165) è $\max(X_1, X_2, \ldots, X_n)/2$. $\square$

7.3 Intervalli di confidenza

Sia $X_1, X_2, \ldots, X_n$ un campione estratto da una popolazione normale di media incognita μ e varianza nota σ^2. Abbiamo in precedenza dimostrato che $\overline{X} := \sum_i X_i/n$ è lo stimatore di massima verosimiglianza per μ. Ciò non significa che possiamo aspettarci che la media campionaria sia esattamente uguale a μ, ma solo che le sarà "vicina". Perciò, rispetto ad uno stimatore puntuale, è a volte preferibile potere produrre un intervallo per il quale abbiamo un certo livello di fiducia (confidenza), che il parametro μ vi appartenga. Per ottenere un tale *intervallo di confidenza*, dobbiamo fare uso della distribuzione di probabilità dello stimatore puntuale. Illustriamo di seguito il procedimento in questa situazione particolare.

Ricordiamo intanto che nelle ipotesi in cui ci siamo messi, $\overline{X}$ è normale di media μ e varianza σ^2/n. Ne segue che

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \tag{7.3.1}$$

Perciò

$$P\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \approx 0.95$$

o equivalentemente,

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

da cui, moltiplicando le disuguaglianze per -1,

$$P\left(1.96\frac{\sigma}{\sqrt{n}} > \mu - \overline{X} > -1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95$$

ovvero, finalmente,

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95 \tag{7.3.2}$$

Il 95% circa delle volte μ starà a una distanza non superiore a $1.96\sigma/\sqrt{n}$ dalla media aritmetica dei dati. Se osserviamo il campione, e registriamo che $\overline{X} = \bar{x}$, allora possiamo dire che "con il 95% di confidenza"

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \tag{7.3.3}$$

Stiamo quindi affermando che, con il 95% di confidenza, la media vera della distribuzione appartiene all'intervallo

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) \tag{7.3.4}$$

Questo intervallo è detto *intervallo di confidenza ad un livello del 95%*, o più semplicemente *intervallo di confidenza al 95%* per μ.

Esempio 7.3.1. Supponiamo che quando un segnale elettrico di valore μ viene trasmesso dalla sorgente A, il ricevente B registri un valore distribuito come una normale di media μ e varianza 4. Altrimenti detto, se μ è il segnale inviato, quello ricevuto è $\mu + N$, dove N denota il rumore, ed è $N \sim \mathcal{N}(0,4)$. Immaginiamo che per ridurre l'errore, lo stesso segnale sia stato trasmesso 9 volte. I valori registrati da B in ricezione sono stati

$$5 \quad 8.5 \quad 12 \quad 15 \quad 7 \quad 9 \quad 7.5 \quad 6.5 \quad 10.5$$

Cerchiamo di ottenere un intervallo di confidenza al 95% per μ.

Siccome

$$\bar{x} = \frac{81}{9} = 9$$

ne segue, sotto l'ipotesi aggiuntiva che i valori ricevuti siano indipendenti, che un intervallo di confidenza al 95% per μ è

$$\left(9 - 1.96\frac{2}{3}, \quad 9 + 1.96\frac{2}{3}\right) = (7.69, 10.31)$$

Perciò possiamo dire di avere "il 95% di fiducia" che il vero messaggio fosse compreso tra 7,69 e 10.31. □

Gli intervalli di confidenza trovati fin qui sono detti in particolare *bilaterali*, perché hanno due estremi finiti. Altre volte invece, siamo interessati a determinare un singolo valore che ci permetta ad esempio di affermare con il 95% di confidenza che μ gli è superiore.

Per trovare un valore siffatto, si noti che se Z è $\mathcal{N}(0,1)$, allora

$$\begin{aligned} 0.95 &\approx P(Z < 1.645) \\ &\approx P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.645\right) \\ &\approx P\left(\overline{X} - 1.645\frac{\sigma}{\sqrt{n}} < \mu\right) \end{aligned}$$

così che un *intervallo di confidenza unilaterale destro ad un livello del 95%* per μ è il seguente,

$$\left(\bar{x} - 1.645\frac{\sigma}{\sqrt{n}}, \quad \infty\right) \tag{7.3.5}$$

dove $\bar{x}$ è il valore che si osserva per la media campionaria.

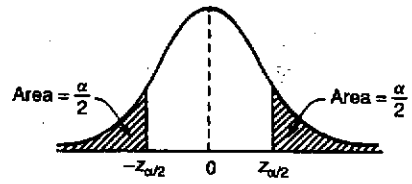


Figura 7.1 Illustrazione grafica di: $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$.

Si possono analogamente definire anche gli intervalli di confidenza unilaterali sinistri, e ad esempio quello al 95% per μ è

$$\left(-\infty, \quad \bar{x} + 1.645\frac{\sigma}{\sqrt{n}}\right) \tag{7.3.6}$$

Esempio 7.3.2. Si determinino al 95% di confidenza degli intervalli unilaterali destro e sinitro per il parametro μ dell'Esempio 7.3.1.

Siccome

$$1.645\frac{\sigma}{\sqrt{n}} = \frac{3.29}{3} \approx 1.097$$

l'intervallo destro al 95% è

$$(9 - 1.097, \infty) = (7.903, \infty)$$

mentre quello sinistro è

$$(-\infty, 9 + 1.097) = (-\infty, 10.097) \quad \square$$

Si possono ottenere intervalli di confidenza per ogni livello di confidenza assegnato. Per riuscirci, si ricordi che (a pagina 177) avevamo definito i numeri z_α in modo tale che

$$P(Z > z_\alpha) = \alpha \tag{7.3.7}$$

dove $Z \sim \mathcal{N}(0,1)$. Questo implica (si veda la Figura 7.1) che per ogni $\alpha \in (0,1)$

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

Da questa equazione si deduce facilmente che

$$\begin{aligned}1-\alpha &= P\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right)\\ &= P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\\ &= P\left(z_{\alpha/2}\frac{\sigma}{\sqrt{n}} > \mu-\overline{X} > -z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\\ &= P\left(\overline{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\end{aligned}$$

Quindi un intervallo di confidenza bilaterale ad un livello di $1-\alpha$ per μ è

$$\left(\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\quad \bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{7.3.8}$$

dove $\bar{x}$ è il valore che si osserva per la media campionaria.

In maniera del tutto analoga, dal fatto che

$$Z := \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$$

è una normale standard, e dalle identità

$$P(Z > z_\alpha) = \alpha$$
$$P(Z < -z_\alpha) = \alpha$$

si deducono intervalli di confidenza unilaterali per qualunque livello di confidenza. In particolare si ottiene che

$$\left(\bar{x}-z_\alpha\frac{\sigma}{\sqrt{n}},\quad \infty\right) \qquad \text{e} \qquad \left(-\infty,\quad \bar{x}+z_\alpha\frac{\sigma}{\sqrt{n}}\right) \tag{7.3.9}$$

sono gli intervalli di confidenza unilaterali (rispettivamente destro e sinistro) ad un livello di $1-\alpha$ per μ.

Esempio 7.3.3. Usiamo i dati dell'Esempio 7.3.1 per calcolare vari intervalli di confidenza al 99% per la media μ: quello bilaterale, e i due unilaterali.

Siccome $z_{0.005} \approx 2.58$ (si usi ad esempio il Programma 5.5b), e

$$2.58\frac{\sigma}{\sqrt{n}} = 1.72$$

ne segue che l'intervallo bilaterale al 99% per μ è

$$9 \pm 1.72$$

ovvero, è l'intervallo $(7.28, 10.72)$.

Inoltre, visto che $z_{0.01} \approx 2.33$, l'intervallo di confidenza unilaterale destro è

$$(9 - 2.33 \times 2/3, \infty) \approx (7.447, \infty)$$

mentre quello sinistro è

$$(-\infty, 9 + 2.33 \times 2/3) \approx (-\infty, 10.553) \quad \square$$

In alcune situazioni ci è richiesto che un intervallo di confidenza di un certo livello $1-\alpha$, abbia una larghezza prescritta, e noi dobbiamo determinare qual è la ampiezza n del campione che garantisce questo risultato. Ad esempio supponiamo di volere un intervallo di lunghezza non superiore a 0.1 che contenga μ con un livello di confidenza del 99%. Quanto grande deve essere n? Ci annotiamo intanto che $z_{0.005} \approx 2.58$ (trovato con il Programma 5.5b). Ne segue che per un campione di ampiezza n, l'intervallo di confidenza al 99% per μ è dato da

$$\left(\bar{x}-2.58\frac{\sigma}{\sqrt{n}},\quad \bar{x}+2.58\frac{\sigma}{\sqrt{n}}\right)$$

La sua lunghezza è quindi pari a $2 \cdot 2.58 \cdot \sigma/\sqrt{n}$. Imponendo allora che

$$5.16\frac{\sigma}{\sqrt{n}} = 0.1$$

si trova che n deve essere almeno pari a

$$n = (51.6 \cdot \sigma)^2$$

Si tenga infine presente che n deve comunque essere intero, quindi se fosse $\sigma = 0.2$, visto che $(51.6 \cdot \sigma)^2 \approx 106.5$, la risposta al quesito iniziale dovrebbe essere che n è almeno pari a 107.

Esempio 7.3.4. Dall'esperienza passata si sa che il peso dei salmoni cresciuti in un allevamento commerciale ha distribuzione normale con media che varia da stagione a stagione, e con deviazione standard sempre pari a 0.3 libbre. Quanto grande occorre prendere il campione, se vogliamo essere sicuri al 95% che la nostra stima del peso medio dei salmoni di quest'anno sia precisa entro ± 0.1 libbre?

Un intervallo di confidenza al 95% per μ, basato su un campione di ampiezza n è dato da

$$\mu \in \left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}},\quad \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right)$$

Poiché la stima $\bar{x}$ non dista più di $1.96 \cdot \sigma/\sqrt{n} = 0.588/\sqrt{n}$ da qualunque punto dell'intervallo, possiamo essere certi al 95% che $\bar{x}$ stia entro 0.1 da μ se

$$\frac{0.588}{\sqrt{n}} \leq 0.1$$

Confidenza, non probabilità

L'espressione "vi è un livello di confidenza del 95% che μ stia nell'intervallo" può portare a interpretazioni erronee. È bene notare che *non stiamo affermando* che la probabilità che $\mu \in (\bar{x} - 1.96\sigma/\sqrt{n}, \bar{x} + 1.96\sigma/\sqrt{n})$ è di 0.95, infatti in questo enunciato non compaiono variabili aleatorie. Quello che affermiamo, invece, è che la tecnica adottata per arrivare a questo intervallo, nel 95% dei casi in cui viene impiegata, produce un intervallo che contiene il valore vero di μ. In altri termini, prima di osservare i dati possiamo dire che vi è il 95% di probabilità che l'intervallo che otterremo contenga μ, mentre dopo l'osservazione dei dati possiamo solo asserire che l'intervallo trovato contiene μ "col 95% di confidenza".

ovvero

$$\sqrt{n} \geq 5.88$$

o ancora

$$n \geq 34.57$$

Concludendo, sarà sufficiente un campione di 35 salmoni. □

7.3.1 Intervalli di confidenza per la media di una distribuzione normale, quando la varianza non è nota

Sia ora $X_1, X_2, \ldots, X_n$ un campione di una popolazione $\mathcal{N}(\mu, \sigma^2)$, con entrambi i parametri ignoti. Vogliamo nuovamente costruire un intervallo di confidenza per μ ad un livello prescritto di $1 - \alpha$. Siccome la deviazione standard σ non è nota, non possiamo più basarci sul fatto che $\sqrt{n}(\overline{X} - \mu)/\sigma$ è una normale standard. Tuttavia, se

$$S^2 := \frac{1}{n-1} \sum_i (X_i - \overline{X})^2 \tag{7.3.10}$$

denota la varianza campionaria, allora segue dal Corollario 6.5.2 di pagina 221 che

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \tag{7.3.11}$$

è una variabile aleatoria di tipo t con $n - 1$ gradi di libertà. Allora, poiché la densità delle distribuzioni t è simmetrica rispetto a zero come quella della normale standard, abbiamo per $\alpha \in (0, 1/2)$ (si veda la Figura 7.2),

$$P\left(-t_{\frac{\alpha}{2},n-1} < \frac{\overline{X} - \mu}{S/\sqrt{n}} < t_{\frac{\alpha}{2},n-1}\right) = 1 - \alpha$$

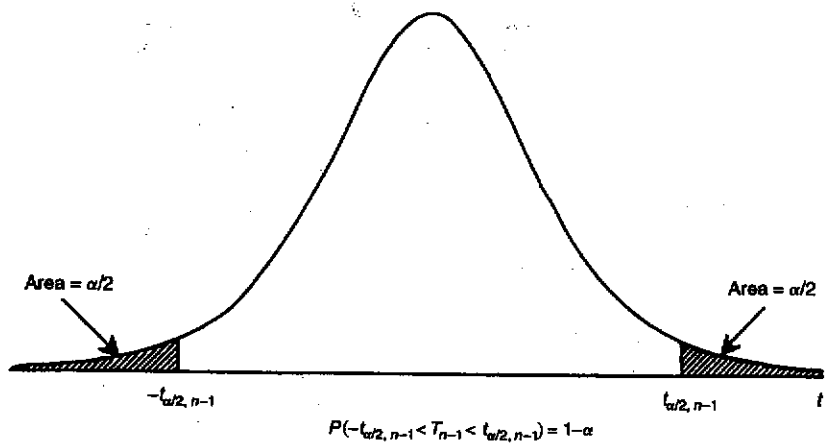


Figura 7.2 Densità di T_n.

o equivalentemente,

$$P\left(\overline{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

Così, se i valori osservati sono $\overline{X} = \bar{x}$ e $S = s$, possiamo dire "con un livello di confidenza di $1 - \alpha$" che

$$\mu \in \left(\bar{x} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}, \quad \bar{x} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}\right) \tag{7.3.12}$$

Esempio 7.3.5. Consideriamo di nuovo l'Esempio 7.3.1, ma questa volta immaginiamo di non conoscere σ. Determiniamo un intervallo di confidenza al 95% per μ, usando i 9 dati ricevuti

5 8.5 12 15 7 9 7.5 6.5 10.5

Un calcolo diretto permette di verificare che

$$\bar{x} = 9$$
$$s^2 = \frac{\sum_i x_i^2 - 9\bar{x}^2}{8} = 9.5$$
$$s \approx 3.082$$

Quindi, poiché $t_{0.025,8} \approx 2.306$ (usando la Tabella A.3 in Appendice, o il Programma 5.8.2b), un intervallo di confidenza al 95% per μ è quello dato da

$$9 \pm 2.306 \cdot \frac{3.082}{3} \quad \text{ovvero} \quad (6.63, 11.37)$$

che è un intervallo più largo di quello ottenuto nell'Esempio 7.3.1.

Il motivo per cui abbiamo ottenuto un intervallo di confidenza più ampio è duplice. In primo luogo abbiamo usato qui una stima di σ maggiore del valore accettato in precedenza. Infatti in quella sede ci era dato che la varianza era 4, mentre qui abbiamo dovuto usare la stima fornita dai dati che è di 9.5. In secondo luogo, è bene notare che anche se avessimo trovato una stima della varianza pari a 4, l'intervallo di confidenza sarebbe risultato comunque più largo; infatti disponendo solo di una stima della varianza, siamo tenuti ad usare una distribuzione di tipo t anziché quella normale standard, che avrebbe una varianza minore (si veda ancora la Figura 5.15 di pagina 193: la distribuzione di tipo t ha le code pesanti). Per chiarire, se avessimo trovato $\bar{x} = 9$ e $s^2 = 4$, il nostro intervallo ci confidenza sarebbe stato

$$9 \pm 2.306 \cdot \frac{2}{3} \quad \text{ovvero} \quad (7.46, 10.54)$$

che è ancora un poco più ampio di quello di pagina 240. □

Osservazione 7.3.1.

(a) Gli intervalli di confidenza per μ quando σ è nota si basano sul fatto che $\sqrt{n}(\overline{X} - \mu)/\sigma$ ha distribuzione normale standard. Quando invece σ non è conosciuta, la si stima con S e poi si usa il fatto che $\sqrt{n}(\overline{X} - \mu)/S$ ha distribuzione di tipo t con $n - 1$ gradi di libertà.

(b) L'ampiezza di un intervallo di confidenza ad un livello fissato, non è per forza maggiore quando non si conosce la varianza. La sua misura infatti è pari a $2z_\alpha\sigma/\sqrt{n}$ quando σ è nota, ed a $2t_{\alpha,n-1}S/\sqrt{n}$ in caso contrario, ed è certamente possibile che la deviazione standard campionaria risulti molto minore di σ. Tuttavia è anche possibile dimostrare che la lunghezza *media* dell'intervallo è maggiore quando la varianza è incognita. Ovvero si può dimostrare rigorosamente che

$$t_{\alpha,n-1}E[S] \geq z_\alpha\sigma$$

Nel Capitolo 13 valuteremo $E[S]$ (si vedano l'Equazione (13.2.11) e la Tabella 13.1), e mostreremo che ad esempio,

$$E[S] \approx \begin{cases} 0.94\sigma & \text{quando } n = 5 \\ 0.97\sigma & \text{quando } n = 9 \end{cases}$$

Siccome però

$$z_{0.025} \approx 1.96 \qquad t_{0.025,4} \approx 2.78 \qquad t_{0.025,8} \approx 2.31$$

l'ampiezza di un intervallo di confidenza al 95% per un campione di 5 dati è di $2 \cdot 1.96 \cdot \sigma/\sqrt{5} \approx 1.75\sigma$ quando si conosce σ, mentre il suo valore atteso è

$2 \cdot 2.78 \cdot 0.94 \cdot \sigma/\sqrt{5} \approx 2.34\sigma$ quando non la si conosce – un aumento del 33.7%. Se si prende invece un campione di 9 dati, i due valori da confrontare sono 1.31σ e 1.49σ, e qui l'aumento è solo del 13.7%.

Gli intervalli di confidenza unilaterali si possono dedurre notando che

$$\begin{aligned} 1 - \alpha &= P\left(\frac{\overline{X} - \mu}{S/\sqrt{n}} < t_{\alpha,n-1}\right) \\ &= P\left(\overline{X} - \mu < \frac{S}{\sqrt{n}} t_{\alpha,n-1}\right) \\ &= P\left(\mu > \overline{X} - \frac{S}{\sqrt{n}} t_{\alpha,n-1}\right) \end{aligned}$$

Ciò infatti significa che se osserviamo $\overline{X} = \bar{x}$ e $S = s$, possiamo affermare con un livello di confidenza di $1 - \alpha$ che

$$\mu \in \left(\bar{x} - \frac{s}{\sqrt{n}} t_{\alpha,n-1}, \quad \infty\right) \tag{7.3.13}$$

e analogamente possiamo dire con lo stesso livello di confidenza che

$$\mu \in \left(-\infty, \quad \bar{x} + \frac{s}{\sqrt{n}} t_{\alpha,n-1}\right) \tag{7.3.14}$$

Il Programma 7.3.1 permette di calcolare gli intervalli di confidenza bilaterali e unilaterali per la media di una popolazione gaussiana, quando non sia nota la varianza.

Esempio 7.3.6. Si determini un intervallo di confidenza al 95% per la media della frequenza cardiaca a riposo degli iscritti di una palestra, nell'ipotesi che un campione casuale di 15 di queste persone abbia fornito i seguenti dati:

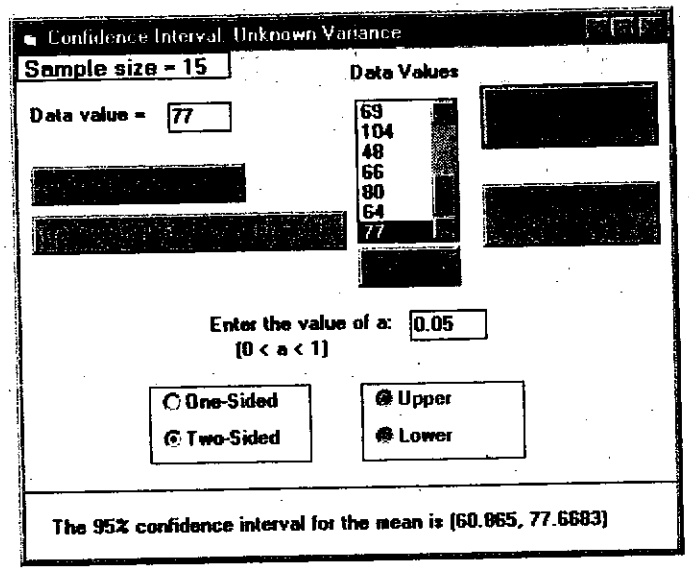
54 63 58 72 49 92 70 73 69 104 48 66 80 64 77

Si trovi anche un intervallo di confidenza sinistro, sempre al 95%.

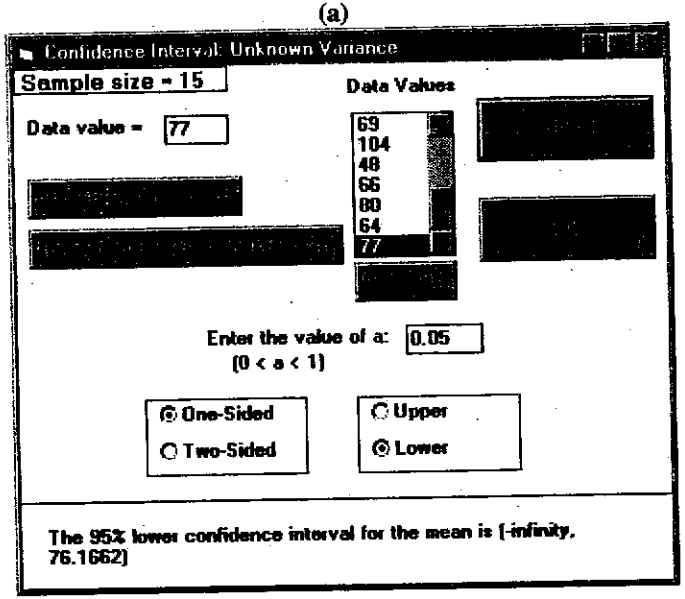
La soluzione si ottiene direttamente dal Programma 7.3.1 (Figura 7.3). □

Nel ricavare gli intervalli di confidenza per la media forniti fino a qui, abbiamo sempre ipotizzato che la distribuzione della popolazione fosse normale. Nel caso in cui questa ipotesi non fosse più valida, le espressioni trovate forniscono comunque delle approssimazioni degli intervalli di confidenza esatti, a condizione però che il campione aleatorio sia sufficientemente numeroso. Infatti, per il teorema del limite centrale,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \stackrel{\cdot}{\sim} \mathcal{N}(0,1) \qquad \text{e} \qquad \frac{\overline{X} - \mu}{S/\sqrt{n}} \stackrel{\cdot}{\sim} t_{n-1} \tag{7.3.15}$$



(a)



(b)

Figura 7.3 Intervalli di confidenza (a) bilaterale e (b) unilaterale per l'Esempio 7.3.6.

saranno approssimativamente distribuite come una normale standard e una t di Student.

Esempio 7.3.7 (Simulazione Monte Carlo). La simulazione al calcolatore costituisce un metodo molto potente per valutare gli integrali mono- e multidimensionali. Supponiamo infatti che f sia una funzione da $\mathbb{R}^r$ in $\mathbb{R}$, e che siamo interessati a stimare la quantità θ, definita da

$$\theta := \int_0^1 \int_0^1 \cdots \int_0^1 f(y_1, y_2, \ldots, y_n)\, dy_1\, dy_2 \cdots dy_n \tag{7.3.16}$$

Notiamo subito, che se $U_1, U_2, \ldots, U_r$ sono variabili aleatorie uniformi su $(0,1)$, allora (grazie alla seconda parte della Proposizione 4.5.1 di pagina 117),

$$\theta = E[f(U_1, U_2, \ldots, U_r)] \tag{7.3.17}$$

Supponiamo ora di fare generare ad un computer r numeri casuali, uniformi su $(0,1)$ e indipendenti[4], e di valutare f a quelle coordinate. Questo produrrà un numero casuale distribuito come $f(U_1, U_2, \ldots, U_r)$ che denotiamo con X_1. Si noti che $E[X_1] = \theta$. Se ripetiamo il procedimento un numero n di volte, otteniamo una successione $X_1, X_2, \ldots, X_n$ di variabili aleatorie i.i.d. che hanno media θ; possiamo allora impiegare questo campione per stimare θ. Questo metodo di approssimazione degli integrali è detto *simulazione Monte Carlo* o *metodo Monte Carlo*.

Pensiamo ad esempio alla stima dell'integrale seguente:

$$\theta := \int_0^1 \sqrt{1-y^2}\, dy = E[\sqrt{1-U^2}]$$

dove U ha distribuzione uniforme su $(0,1)$. Siano $U_1, \ldots, U_{100}$ delle variabili aleatorie con tale distribuzione e indipendenti, generate da un calcolatore. Ponendo

$$X_i := \sqrt{1-U_i^2}, \qquad i = 1, 2, \ldots, 100$$

otteniamo un campione di 100 variabili aleatorie di media θ. Realizzando questa simulazione abbiamo trovato una media campionaria di 0.786 e una deviazione standard campionaria di 0.23. Allora, siccome $t_{0.025,99} \approx 1.985$, si ottiene che un intervallo di confidenza al 95% per θ è il seguente,

$$0.786 \pm 1.985 \cdot 0.023$$

Quindi possiamo affermare con il 95% di confidenza, che θ (il cui valore esatto si può dimostrare essere $\pi/4 \approx 0.7854$) è compreso tra 0.740 e 0.832. □

[4] Si ricordi che questo tipo di variabili aleatorie sono le uniche direttamente riproducibili al calcolatore. Ogni altro tipo di distribuzione desiderata deve essere ricostruita a partire da essi. Si veda anche il riquadro a pagina 167, e il successivo Esempio 5.4.4

7.3.2 Intervalli di confidenza per la varianza di una distribuzione normale

Se $X_1, X_2, \ldots, X_n$ è un campione proveniente da una distribuzione normale con parametri μ e σ^2 entrambi incogniti, possiamo contruire degli intervalli di confidenza per σ^2 basandoci sul fatto che

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{7.3.18}$$

Infatti,

$$\begin{aligned} 1-\alpha &= P\left(\chi^2_{1-\frac{\alpha}{2},n-1} < (n-1)\frac{S^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2},n-1}\right) \\ &= P\left(\frac{\chi^2_{1-\frac{\alpha}{2},n-1}}{(n-1)S^2} < \frac{1}{\sigma^2} < \frac{\chi^2_{\frac{\alpha}{2},n-1}}{(n-1)S^2}\right) \\ &= P\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right) \end{aligned}$$

Quindi, se $S^2 = s^2$, il seguente costituisce un intervallo di confidenza (bilaterale) per σ^2 ad un livello di confidenza di $1-\alpha$:

$$\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \quad \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right) \tag{7.3.19}$$

Esempio 7.3.8. Una certa procedura automatizzata deve produrre rondelle con una variabilità di spessore molto ridotta. Supponiamo di scegliere a caso 10 rondelle e misurarne lo spessore, che risulta, in pollici,

0.123 0.133 0.124 0.125 0.126 0.128 0.120 0.124 0.130 0.126

qual è l'intervallo di confidenza al 90% per la deviazione standard dello spessore delle rondelle?

Un calcolo diretto mostra che $s^2 \approx 1.366 \times 10^{-5}$. Consultando la Tabella A.2 in Appendice, o eseguendo il Programma 5.8.1b si trova che $\chi^2_{0.05,9} \approx 16.917$ e $\chi^2_{0.95,9} \approx 3.334$, quindi

$$\begin{aligned} \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} &\approx \frac{9 \times 1.366 \times 10^{-5}}{16.917} \approx 7.26 \times 10^{-6} \\ \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}} &\approx \frac{9 \times 1.366 \times 10^{-5}}{3.334} \approx 36.87 \times 10^{-6} \end{aligned}$$

Tabella 7.1 Intervalli con livello di confidenza $1-\alpha$ per campioni normali.

$$X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$$

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S := \left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^{1/2}$$

Ipotesi	θ	Intervallo bilaterale	Intervallo sinistro	Intervallo destro
σ^2 nota	μ	$\overline{X} \pm z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$	$\left(-\infty, \overline{X} + z_\alpha\frac{\sigma}{\sqrt{n}}\right)$	$\left(\overline{X} - z_\alpha\frac{\sigma}{\sqrt{n}}, \infty\right)$
σ^2 non nota	μ	$\overline{X} \pm t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}$	$\left(-\infty, \overline{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n}}\right)$	$\left(\overline{X} - t_{\alpha,n-1}\frac{S}{\sqrt{n}}, \infty\right)$
μ non nota	σ^2	$\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right)$	$\left(0, \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}\right)$	$\left(\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}}, \infty\right)$

per cui

$$\sigma^2 \in (7.26 \times 10^{-6}, 36.87 \times 10^{-6})$$

con il 90% di confidenza, o equivalentemente, prendendo le radici quadrate,

$$\sigma \in (2.69 \times 10^{-3}, 6.07 \times 10^{-3})$$

sempre con il 90% di confidenza. □

Gli intervalli di confidenza unilaterali per σ^2 si ottengono in maniera del tutto analoga, e sono presentati nella Tabella 7.1, che riassume tutti i risultati di questa sezione.

7.4 Stime per la differenza delle medie di due popolazioni normali

Siano $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_m$ due campioni estratti da popolazioni normali differenti, e denotiamo con μ_1 e σ_1^2 i parametri della prima, e con μ_2 e σ_2^2 quelli della seconda. Supponiamo che i due campioni siano tra loro indipendenti, e tentiamo di stimare $\mu_1 - \mu_2$.

Siccome $\overline{X} := \sum_{i=1}^{n} X_i/n$ e $\overline{Y} := \sum_{j=1}^{m} Y_j/m$ sono gli stimatori di massima verosimiglianza di μ_1 e μ_2, sembra ragionevole (e infatti può essere dimostrato) che $\overline{X} - \overline{Y}$ sia lo stimatore di massima verosimiglianza di $\mu_1 - \mu_2$.

Per ottenere uno stimatore non puntuale, nella forma di un intervallo di confidenza, occorre conoscere la distribuzione di $\overline{X} - \overline{Y}$. Poiché

$$\overline{X} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n}\right) \qquad \text{e} \qquad \overline{Y} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{m}\right)$$

si può dedurre, dal fatto che la somma di normali indipendenti è ancora una variabile aleatoria normale, che

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right) \tag{7.4.1}$$

dove abbiamo sfruttato che $E[\overline{X} - \overline{Y}] = E[\overline{X}] - E[\overline{Y}]$ e che $\mathrm{Var}(\overline{X} - \overline{Y}) = \mathrm{Var}(\overline{X} + (-1)\overline{Y}) = \mathrm{Var}(\overline{X}) + (-1)^2\,\mathrm{Var}(\overline{Y})$. Perciò, ipotizzando di conoscere σ_1^2 e σ_2^2, abbiamo che

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \sim \mathcal{N}(0,1) \tag{7.4.2}$$

e possiamo dedurre, con i passaggi che ci sono ormai familiari, che

$$1 - \alpha = P\left(-z_{\frac{\alpha}{2}} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} < z_{\frac{\alpha}{2}}\right)$$

$$= P\left(\overline{X} - \overline{Y} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} < \mu_1 - \mu_2 < \overline{X} - \overline{Y} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right)$$

Se $\overline{X}$ e $\overline{Y}$ dopo l'osservazione di dati risultano uguali a $\bar{x}$ e $\bar{y}$ rispettivamente, allora con un livello di confidenza di $1 - \alpha$,

$$\mu_1 - \mu_2 \in \left(\bar{x} - \bar{y} - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}, \quad \bar{x} - \bar{y} + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right) \tag{7.4.3}$$

Gli intervalli di confidenza unilaterali si ottengono in maniera analoga, e lasciamo al lettore la verifica che vi è un livello di confidenza di $1 - \alpha$ che

$$\mu_1 - \mu_2 \in \left(-\infty, \quad \bar{x} - \bar{y} + z_\alpha\sqrt{\sigma_1^2/n + \sigma_2^2/m}\right) \tag{7.4.4}$$

Il Programma 7.4.1, disponibile online sul sito web di questo libro è in grado di calcolare tutti gli intervalli di confidenza bilaterali e unilaterali per $\mu_1 - \mu_2$.

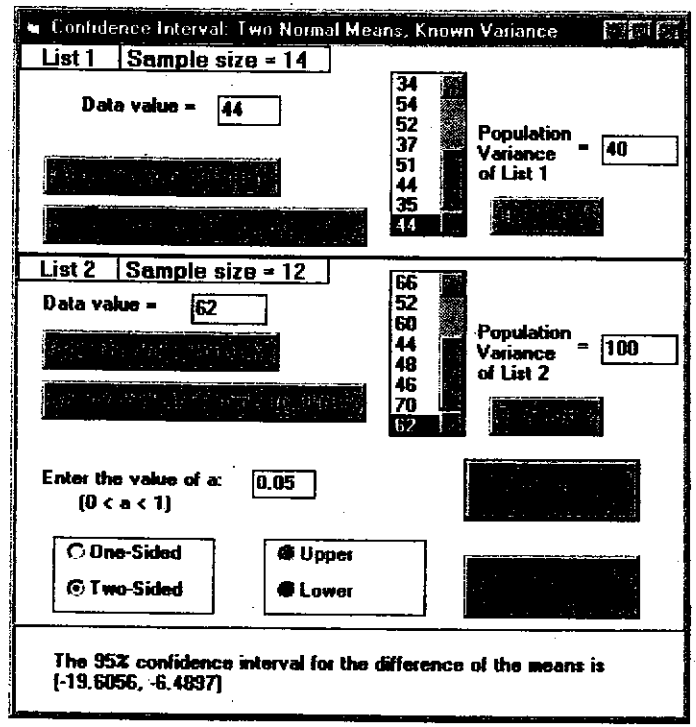
Esempio 7.4.1. Due tipi diversi di guaine isolanti per cavi elettrici vengono testati per determinare a che voltaggio cominciano a rovinarsi. Sottoponendo gli esemplari a livelli crescenti di tensione si registrano i guasti alle tensioni seguenti:

Tipo A	36	44	41	53	38	36	34	54	52	37	51	44	35	44
Tipo B	52	64	38	68	66	52	60	44	48	46	70	62		

Supponiamo di sapere che il voltaggio tollerato dai cavi abbia distribuzione normale: per quelli di tipo A, con media incognita μ_A e varianza $\sigma_A^2 = 40$, mentre per quelli di tipo B i parametri sono μ_B e $\sigma_B^2 = 100$. Si determini un intervallo bilaterale con il 95% di confidenza per $\mu_A - \mu_B$. Si determini anche un valore che permetta di affermare che $\mu_A - \mu_B$ gli è superiore, con il 95% di confidenza.

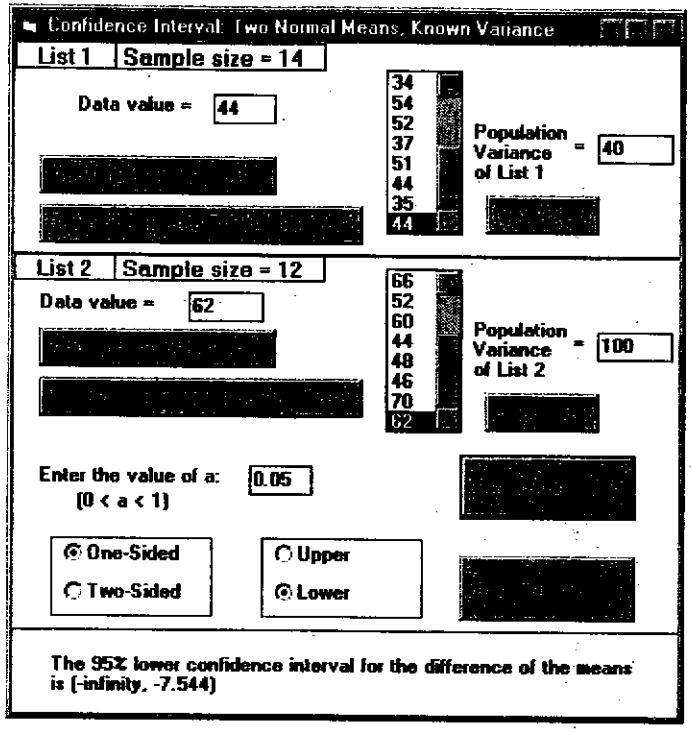
Eseguiamo il Programma 7.4.1 per ottenere la soluzione (in Figura 7.4). □

Vogliamo ora stimare nuovamente $\mu_1 - \mu_2$ con un intervallo di confidenza, questa volta però nell'ipotesi che σ_1^2 e σ_2^2 non siano note. È abbastanza naturale tentare di sostituire le varianze reali, che sono incognite, con quelle campionarie, che sono



(a)

Figura 7.4 Intervalli di confidenza (a) bilaterale e (b) unilaterale per l'Esempio 7.4.1.



(b)

Figura 7.4 (continua)

stimatori delle prime:

$$S_1^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$
$$S_2^2 := \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y})^2 \qquad (7.4.5)$$

Vorremmo quindi basarci su una statistica come la seguente,

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n + S_2^2/m}}$$

Tuttavia, per potere utilizzare l'espressione precedente per ricavarne degli intervalli di confidenza, occorrerebbe conoscere la sua distribuzione, ed essa non deve dipendere da σ_1^2 e σ_2^2. Sfortunatamente, questa distribuzione è molto complicata e dipende

effettivamente dai parametri incogniti σ_1^2 e σ_2^2: soltanto nel caso particolare in cui $\sigma_1^2 = \sigma_2^2$ siamo in grado di ottenere uno stimatore non puntuale. Supponiamo quindi che le varianze delle popolazioni, anche se incognite, siano identiche, e denotiamo con σ^2 il loro comune valore. Segue dal Teorema 6.5.1 che

$$(n-1)\frac{S_1^2}{\sigma^2} \sim \chi_{n-1}^2 \qquad \text{e} \qquad (m-1)\frac{S_2^2}{\sigma^2} \sim \chi_{m-1}^2 \qquad (7.4.6)$$

Inoltre, poiché i due campioni sono indipendenti, le due chi-quadro precedenti sono indipendenti, e quindi la loro somma ha a sua volta una distribuzione di tipo chi-quadro, con un numero di gradi di libertà che è la somma di quelli di partenza:

$$(n-1)\frac{S_1^2}{\sigma^2} + (m-1)\frac{S_2^2}{\sigma^2} \sim \chi_{n+m-2}^2 \qquad (7.4.7)$$

Abbiamo già notato che

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2/n + \sigma^2/m}} \sim \mathcal{N}(0,1) \qquad (7.4.8)$$

e sappiamo che il rapporto tra una normale standard e $\sqrt{\chi_k^2/k}$ (χ_k^2 è una chi-quadro con k gradi di libertà, indipendente dalla normale) è per definizione una distribuzione di tipo t con k gradi di libertà. Nel nostro caso la chi-quadro è quella data dall'Equazione (7.4.7), e $k = n + m - 2$. L'indipendenza è garantita dal fatto che $\overline{X}$, $\overline{Y}$, S_1^2 e S_2^2 sono indipendenti per il Teorema 6.5.1. Se poniamo allora

$$S_p^2 := \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2} = \frac{n-1}{n+m-2}S_1^2 + \frac{m-1}{n+m-2}S_2^2 \qquad (7.4.9)$$

otteniamo che

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(1/n + 1/m)}}\left(\frac{S_p^2}{\sigma^2}\right)^{-1/2} = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{1/n + 1/m}} \sim t_{n+m-2} \qquad (7.4.10)$$

siamo quindi in grado di determinare gli intervalli di confidenza per $\mu_1 - \mu_2$. Infatti

$$P\left(-t_{\frac{\alpha}{2},n+m-2} \le \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{1/n + 1/m}} \le t_{\frac{\alpha}{2},n+m-2}\right) = 1 - \alpha$$

quindi se dopo l'osservazione dei dati $\overline{X} = \bar{x}$, $\overline{Y} = \bar{y}$ e $S_p = s_p$, possiamo affermare con un livello di confidenza di $1 - \alpha$, che $\mu_1 - \mu_2$ appartiene all'intervallo

$$\bar{x} - \bar{y} \pm t_{\frac{\alpha}{2},n+m-2}\, s_p\sqrt{1/n + 1/m} \qquad (7.4.11)$$

Gli intervalli di confidenza unilaterali si trovano in maniera analoga.

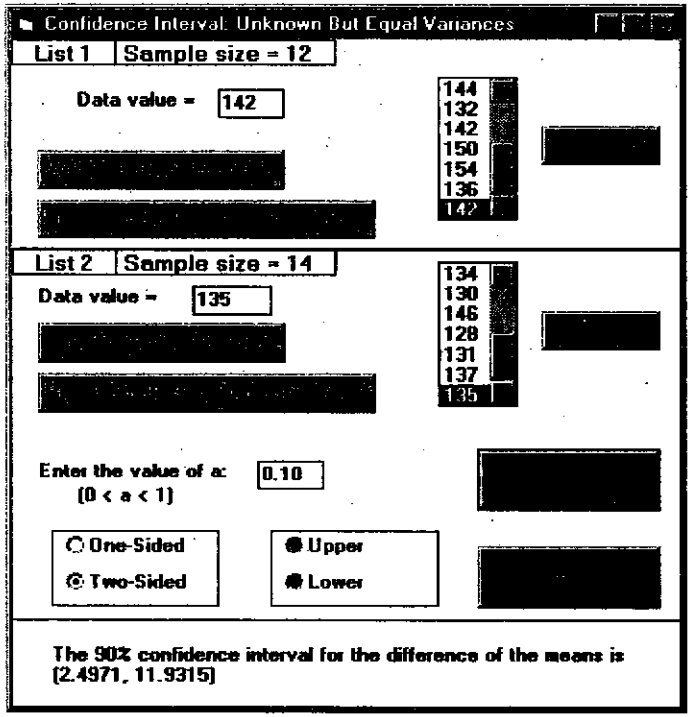
Il Programma 7.4.2 permette di calcolare gli intervalli di confidenza bilaterali e unilaterali per la differenza tra le medie di due popolazioni gaussiane di varianze sconosciute ma coincidenti.

Esempio 7.4.2. Un produttore di batterie dispone di due tecniche di fabbricazione differenti. Due gruppi di batterie scelti a caso, 12 prodotte con la tecnica I e 14 con la tecnica II, sono risultate avere le seguenti capacità (in ampere-ora):

Tecnica I	140	136	138	150	152	144	132	142	150	154	136	142	
Tecnica II	144	132	136	140	128	150	130	134	130	146	128	131	137	135

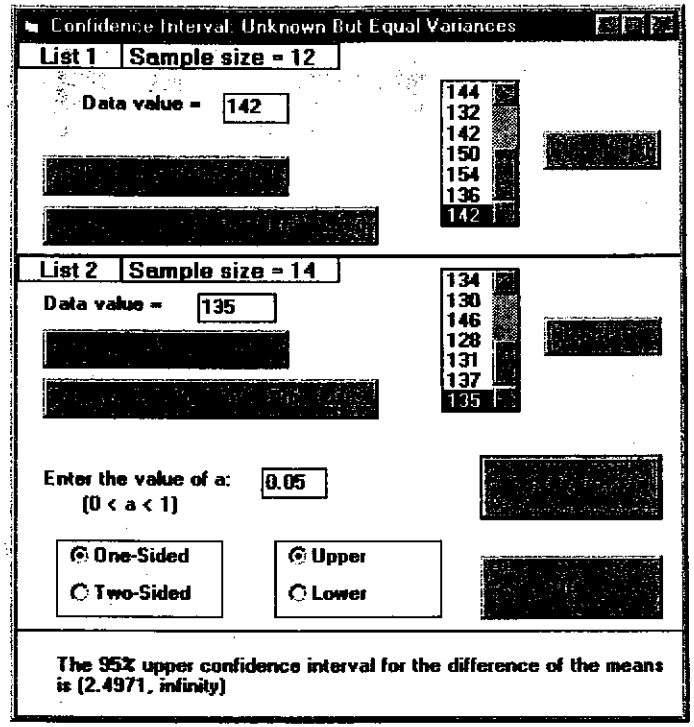
Si determini un intervallo di confidenza la 90%, bilaterale, per la differenza delle medie, ipotizzando che le varianze delle due popolazioni siano uguali. Si calcoli poi un intervallo unilaterale destro per $\mu_{\mathrm{I}} - \mu_{\mathrm{II}}$ ad un livello di confidenza del 95%.

Eseguiamo il Programma 7.4.2 per ottenere la soluzione in Figura 7.5. □



(a)

Figura 7.5 Intervalli di confidenza (a) bilaterale e (b) unilaterale per l'Esempio 7.4.2.



(b)

Figura 7.5 (continua)

Osservazione 7.4.1. L'intervallo di confidenza dell'Equazione (7.4.11) è stato ottenuto sotto l'ipotesi che le varianze delle due popolazioni fossero uguali; avendo denotato con σ^2 il loro comune valore, la statistica che compare nell'Equazione (7.4.8) risulta avere distribuzione normale standard. Siccome però σ non è noto, questo risultato non poteva essere usato direttamente per trovare gli intervalli di confidenza: era necessario prima stimare σ^2. Per farlo, notando che entrambe le varianze campionarie S_1^2 e S_2^2 sono stimatori di σ^2, le abbiamo usate tutte e due, costruendo lo stimatore S_p^2 che è una loro media pesata a seconda dei gradi di libertà (Equazione (7.4.9)). La statistica S_p^2 è a volte detta stimatore *pooled*; essa ci ha permesso di riscalare l'espressione dell'Equazione (7.4.8), ottenendo una nuova statistica la cui distribuzione non dipende più da σ, ovvero quella che compare nell'Equazione (7.4.10).

I risultati di questa sezione sono riassunti nella Tabella 7.2

Tabella 7.2 Intervalli di confidenza ad un livello di $1-\alpha$ per $\mu_1-\mu_2$, cioè la differenza tra le medie di due popolazioni normali.

$$X_i \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right), i = 1, \ldots, n \qquad Y_j \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right), j = 1, \ldots, m$$

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \overline{Y} := \frac{1}{m}\sum_{j=1}^{m} Y_j$$

$$S_1^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \qquad S_2^2 := \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y})^2$$

$$N := n + m - 2 \qquad S_p := \sqrt{\frac{(n-1)S_1^2 + (m-1)S_2^2}{N}}$$

Si assume	Intervallo bilaterale	Intervallo sinistro
σ_1 e σ_2 note	$\overline{X} - \overline{Y} \pm z_{\frac{\alpha}{2}}\sqrt{\sigma_1^2/n + \sigma_2^2/m}$	$\left(-\infty, \quad \overline{X} - \overline{Y} + z_\alpha\sqrt{\sigma_1^2/n + \sigma_2^2/m}\right)$
σ_1 e σ_2 non note ma uguali	$\overline{X} - \overline{Y} \pm t_{\frac{\alpha}{2},N} \cdot S_p\sqrt{\frac{1}{n} + \frac{1}{m}}$	$\left(-\infty, \quad \overline{X} - \overline{Y} + t_{\alpha,N} \cdot S_p\sqrt{\frac{1}{n} + \frac{1}{m}}\right)$

Nota: gli intervalli unilaterali destri per $\mu_1 - \mu_2$ si possono ricavare da quelli sinistri per $\mu_2 - \mu_1$.

7.5 Intervalli di confidenza approssimati per la media di una distribuzione di Bernoulli

Consideriamo una popolazione di oggetti, ognuno dei quali indipendentemente da tutti gli altri soddisfa certi requisiti con probabilità incognita p. Nel caso vengano testati n di questi oggetti, rilevando quanti di essi raggiungono tali requisiti, come possiamo usare questa grandezza per ottenere un intervallo di confidenza per p?

Se X denota quanti oggetti, sugli n testati, soddisfano i requisiti di interesse, è facile convincersi che X ha distribuzione binomiale di parametri n e p. Quindi nel caso n sia un numero elevato, X è approssimativamente normale con media np e varianza $np(1-p)$, e di conseguenza

$$\frac{X - np}{\sqrt{np(1-p)}} \dot{\sim} \mathcal{N}(0,1) \tag{7.5.1}$$

Preso allora un qualunque valore $\alpha \in (0,1)$,

$$P\left(-z_{\frac{\alpha}{2}} < \frac{X - np}{\sqrt{np(1-p)}} < z_{\frac{\alpha}{2}}\right) \approx 1 - \alpha$$

Se x è il valore assunto da X, quella che segue è una *regione* che contiene p co livello di confidenza di $1-\alpha$,

$$\left\{p : -z_{\frac{\alpha}{2}} < \frac{x - np}{\sqrt{np(1-p)}} < z_{\frac{\alpha}{2}}\right\}$$

Tale regione non è un intervallo. Se vogliamo ottenere un intervallo di confidenza vero e proprio, denotiamo con $\hat{p} := X/n$ la frazione degli oggetti del campione che soddisfa i requisiti in esame. Sappiamo dall'Esempio 7.2.1 che $\hat{p}$ è lo stimatore di massima verosimiglianza di p, ed è una buona approssimazione di p. Per questo motivo $\sqrt{n\hat{p}(1-\hat{p})}$ è approssimativamente uguale a $\sqrt{np(1-p)}$, e quindi dall'Equazione (7.5.1) deduciamo che

$$\frac{X - np}{\sqrt{n\hat{p}(1-\hat{p})}} \dot{\sim} \mathcal{N}(0,1) \tag{7.5.2}$$

Questa statistica al contrario della precedente ci consente di arrivare rapidamente a un intervallo di confidenza. Sia $\alpha \in (0,1)$, allora

$$\begin{aligned} 1 - \alpha &\approx P\left(-z_{\frac{\alpha}{2}} < \frac{X - np}{\sqrt{n\hat{p}(1-\hat{p})}} < z_{\frac{\alpha}{2}}\right) \\ &= P\left(-z_{\frac{\alpha}{2}}\sqrt{n\hat{p}(1-\hat{p})} < np - X < z_{\frac{\alpha}{2}}\sqrt{n\hat{p}(1-\hat{p})}\right) \\ &= P\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n}\right) \end{aligned} \tag{7.5.3}$$

e l'ultima formula fornisce un intervallo di confidenza approssimato per p.

Esempio 7.5.1. Un campione di 100 transistor viene estratto da una grossa fornitura e testato. In tutto 80 pezzi sono adeguati ai requisiti; volendo trovare un intervallo di confidenza al 95% per la percentuale p di transistor accettabili della fornitura, scriviamo

$$\left(0.8 - 1.96\sqrt{0.8 \cdot 0.2/100}, \quad 0.8 + 1.96\sqrt{0.8 \cdot 0.2/100}\right) = (0.7216, \quad 0.8784)$$

Quindi possiamo affermare con il 95% di confidenza che sarà accettabile una percentuale di transistor compresa tra il 72.16% e il 87.84% □

Esempio 7.5.2. Il 14 ottobre del 1997 il *New York Times* riportò un sondaggio recente che indicava che il 52% della popolazione con un margine di errore di ±4% era soddisfatta dell'operato del presidente Clinton. Cosa significa? È possibile stabilire quante persone furono intervistate?

È pratica comune per i mezzi di informazione presentare intervalli di confidenza al 95%. Sia p la percentuale della popolazione favorevole all'operato del presidente. Siccome $z_{0.025} \approx 1.96$, un intervallo di confidenza per p al 95% è dato da

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n} = 0.52 \pm 1.96\sqrt{0.52 \cdot 0.48/n}$$

dove n è il numero degli intervistati. Siccome il margine di errore è di 4%, si può dedurre che

$$1.96\sqrt{0.52 \cdot 0.48/n} \approx 0.04$$

ovvero

$$n \approx \frac{1.96^2 \cdot 0.52 \cdot 0.48}{0.04^2} \approx 599.29$$

Perciò gli intervistati furono circa 599, e il 52% di essi si dichiarò a favore dell'operato del presidente. □

Spesso ci viene richiesto di ottenere un intervallo di confidenza per p non più ampio di una lunghezza b assegnata. Il problema consiste nel determinare un valore appropriato dell'ampiezza n del campione. Notiamo a questo scopo che ad un livello di confidenza di $1-\alpha$, l'ampiezza dell'intervallo di confidenza approssimato per p è

$$2z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n} \approx 2z_{\frac{\alpha}{2}}\sqrt{p(1-p)/n}$$

Sfortunatamente, né p né $\hat{p}$ sono noti in anticipo, e quindi non possiamo imporre che una delle espressioni qui sopra sia uguale a b, risolvendo poi rispetto a n. Quello che possiamo fare allora, è raccogliere un campione preliminare per ottenere almeno una stima grossolana p^* di p, e usare questa stima per determinare n risolvendo l'equazione

$$2z_{\frac{\alpha}{2}}\sqrt{p^*(1-p^*)/n} = b$$

che, elevando al quadrato e moltiplicando per n/b^2 entrambi i membri ci porta a

$$n = \frac{4z^2_{\alpha/2}}{b^2}p^*(1-p^*) \tag{7.5.4}$$

Così, se il campione preliminare era costituito da k elementi, è necessario raccogliere altri $n-k$ dati (se n non è minore di k, ovviamente) per ottenere un intervallo di confidenza che avrà approssimativamente l'ampiezza richiesta.

Esempio 7.5.3. Una azienda produce circuiti integrati, ciascuno dei quali risulta accettabile indipendentemente da tutti gli altri con probabilità incognita p. Si vuole ottenere un intervallo di confidenza per p ad un livello del 99%, la cui ampiezza sia approssimativamente di 0.05. Si raccoglie allora un primo campione di 30 chip, 26 dei quali risultano accettabili, fornendo una prima stima grossolana di p che è $p^* = 26/30$. Usando questo valore, un intervallo di confidenza approssimato di ampiezza 0.05 richiederebbe un campione di

$$n = 4\frac{(z_{0.005})^2}{0.05^2} \cdot \frac{26}{30}\left(1-\frac{26}{30}\right) \approx 4\frac{2.58^2}{0.05^2} \cdot \frac{26}{30} \cdot \frac{4}{30} \approx 1231$$

Tabella 7.3 Intervalli di confidenza ad un livello di $1-\alpha$ per il parametro di una distribuzione di Bernoulli.

$\hat{p} := \dfrac{X}{n}$, X è il numero di valori 1 nel campione bernoulliano

Tipo di intervallo	Intervallo di confidenza
Bilaterale	$\hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n}$
Unilaterale sinistro	$\left(-\infty, \quad \hat{p} + z_\alpha\sqrt{\hat{p}(1-\hat{p})/n}\right)$
Unilaterale destro	$\left(\hat{p} - z_\alpha\sqrt{\hat{p}(1-\hat{p})/n}, \quad \infty\right)$

chip. Dobbiamo allora testarne altri 1 201; immaginando di trovarne, per esempio, 1 040 di accettabili, l'intervallo di confidenza finale che ne risulta è dato da

$$\frac{1\,066}{1\,231} \pm \frac{z_{0.005}}{1\,231}\sqrt{1\,066\left(1-\frac{1\,066}{1\,231}\right)}$$

ovvero

$$(0.8409, \quad 0.8910)$$

che ha effettivamente una ampiezza di 0.0501. □

Osservazione 7.5.1. Come abbiamo visto, l'intervallo bilaterale con livello di confidenza $1-\alpha$, ha lunghezza approssimativamente b quando il numero di elementi del campione è

$$n = \frac{4z^2_{\alpha/2}}{b^2}p(1-p)$$

La parabola $g(p) := p(1-p)$ tocca il suo massimo pari a $1/4$ quando $p = 1/2$. Qualunque sia il valore di p, quindi, si avrà sempre

$$n \leq \frac{z^2_{\alpha/2}}{b^2} \tag{7.5.5}$$

perciò scegliendo un campione di ampiezza $z^2_{\alpha/2}/b^2$, siamo sicuri di ottenere un intervallo di confidenza non più grande di b senza bisogno di procurarci un campione preliminare. Si tenga presente che questa sovrastima di n è tanto peggiore quanto più p è vicino a 0 oppure a 1.

Gli intervalli di confidenza unilaterali per p si ottengono altrettanto facilmente; la Tabella 7.3 riporta le espressioni finali.

7.6 * Intervalli di confidenza per la media della distribuzione esponenziale

Consideriamo un campione $X_1, X_2, \ldots, X_n$ di variabili aleatorie esponenziali i.i.d. tutte con media θ incognita. È possibile dimostrare che lo stimatore di massima verosimiglianza per θ è costituito dalla media campionaria $\sum_{i=1}^n X_i/n$. Per ottenere gli intervalli di confidenza, ricordiamo dal Corollario 5.7.2 di pagina 187 che $\sum_{i=1}^n X_i$ ha distribuzione gamma con parametri n e $1/\theta$. Deduciamo allora dalla Sezione 5.8.1.1 a pagina 190 che

$$\frac{2}{\theta}\sum_{i=1}^{n} X_i \sim \chi^2_{2n} \tag{7.6.1}$$

quindi per ogni $\alpha \in (0,1)$,

$$\begin{aligned} 1-\alpha &= P\left(\chi^2_{1-\frac{\alpha}{2},2n} < \frac{2}{\theta}\sum_{i=1}^{n} X_i < \chi^2_{\frac{\alpha}{2},2n}\right) \\ &= P\left(\frac{2\sum_{i=1}^n X_i}{\chi^2_{\frac{\alpha}{2},2n}} < \theta < \frac{2\sum_{i=1}^n X_i}{\chi^2_{1-\frac{\alpha}{2},2n}}\right) \end{aligned}$$

Dopo che il campione di dati viene osservato, e si trova che $X_i = x_i$, per $i = 1, \ldots, n$ si può affermare con un livello di confidenza di $1-\alpha$ che

$$\theta \in \left(\frac{2\sum_{i=1}^n x_i}{\chi^2_{\frac{\alpha}{2},2n}}, \quad \frac{2\sum_{i=1}^n x_i}{\chi^2_{1-\frac{\alpha}{2},2n}}\right) \tag{7.6.2}$$

Esempio 7.6.1. Si pensa che gli oggetti prodotti da una azienda abbiano tempi di vita in ore che sono variabili aleatorie esponenziali indipendenti di media θ. La loro densità è quindi

$$f(x) = \frac{1}{\theta}e^{-x/\theta}; \qquad x > 0$$

Se la somma dei tempi di vita di 10 esemplari è pari a 1 740 ore, che intervallo di confidenza al 95% ne risulta, per la media della popolazione θ?

Usando il Programma 5.8.1b o la Tabella A.2, scopriamo che

$$\chi^2_{0.025,20} \approx 34.170, \qquad \chi^2_{0.975,20} \approx 9.591$$

Possiamo quindi concludere che, con il 95% di confidenza, θ appartiene all'intervallo

$$\left(\frac{2\times 1\,740}{34.170}, \quad \frac{2\times 1\,740}{9.591}\right) \approx (101.84, \quad 362.84) \quad \square$$

7.7 * Valutare l'efficienza degli stimatori puntuali

Sia $\boldsymbol{X} := (X_1, X_2, \ldots, X_n)$ un campione casuale estratto da una popolazione di distribuzione nota eccetto che per un parametro incognito θ, e sia $d = d(\boldsymbol{X})$ uno stimatore di θ. Come possiamo valutare la sua efficacia come stimatore? Un criterio potrebbe essere quello di considerare il quadrato della differenza tra $d(\boldsymbol{X})$ e θ, però $(d(\boldsymbol{X})-\theta)^2$ è una variabile aleatoria, quindi stabiliamo di adoperare $r(d,\theta)$, l'*errore quadratico medio* dello stimatore d, che è per definizione

$$r(d,\theta) := E[(d(\boldsymbol{X})-\theta)^2] \tag{7.7.1}$$

Sarà questo il nostro indicatore del valore di d come stimatore di θ.

Sarebbe ideale se esistesse un singolo stimatore d che minimizzasse $r(d,\theta)$ per tutti i valori di θ, però questo non accade tranne che in situazioni comunque banali. Infatti se definiamo lo stimatore d^* in modo che sia sempre uguale a 4,

$$d^*(\boldsymbol{X}) \equiv 4$$

anche se questa scelta può sembrare assurda (ad esempio perché lo stimatore non fa alcun uso dei dati), è certamente vero che quando $\theta = 4$, questo stimatore, con il suo errore quadratico medio nullo, si comporta meglio di qualunque altro.

Anche se stimatori con errore quadratico medio minimo esistono raramente, a volte si può trovarne uno che minimizzi $r(d,\theta)$ tra tutti quelli che soddisfano una certa proprietà, come ad esempio quella di essere non distorti.

Definizione 7.7.1. Sia $d = d(\boldsymbol{X})$ uno stimatore del parametro θ. Allora

$$b_\theta(d) := E[d(X)] - \theta \tag{7.7.2}$$

è detto il *bias* di d come stimatore di θ. Se esso è nullo, diciamo che d è uno stimatore *corretto* o anche *non distorto*.

In altri termini, uno stimatore è corretto se il suo valore atteso coincide con il parametro che esso deve stimare.

Esempio 7.7.1. Sia $X_1, X_2, \ldots, X_n$ un campione proveniente da una popolazione di media incognita θ. Allora le due statistiche seguenti,

$$\begin{aligned} d_1(X_1, X_2, \ldots, X_n) &= X_1 \\ d_2(X_1, X_2, \ldots, X_n) &= \frac{X_1 + X_2 + \cdots + X_n}{n} \end{aligned}$$

sono entrambe degli stimatori non distorti di θ; la verifica è immediata,

$$E[X_1] = E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \theta$$

Più in generale, $d_3(X_1, X_2, \ldots, X_n) := \sum_{i=1}^n \lambda_i X_i$ è uno stimatore corretto di θ ogni volta che $\sum_{i=1}^n \lambda_i = 1$. Infatti

$$E\left[\sum_{i=1}^n \lambda_i X_i\right] = \sum_{i=1}^n \lambda_i E[X_i]$$
$$= \sum_{i=1}^n \lambda_i \theta = \theta \quad \square$$

Se $d = d(\boldsymbol{X})$ è uno stimatore corretto, allora il suo errore quadratico medio è

$$\begin{aligned} r(d,\theta) &= E[(d-\theta)^2] \\ &= E[(d-E[d])^2] \\ &= \mathrm{Var}(d) \end{aligned}$$

Quindi l'errore quadratico medio di uno stimatore corretto è pari alla sua varianza.

Esempio 7.7.2 (Combinazione di stimatori corretti indipendenti). Consideriamo due stimatori corretti e indipendenti di un parametro θ, denotati d_1 e d_2, e siano σ_1^2 e σ_2^2 le rispettive varianze. Quindi per $i = 1, 2$,

$$E[d_i] = \theta \qquad \mathrm{Var}(d_i) = \sigma_i^2$$

Qualunque statistica della forma

$$d := \lambda d_1 + (1-\lambda) d_2$$

sarà comunque uno stimatore corretto di θ. Vogliamo allora trovare il valore di λ che produce lo stimatore d con il minore errore quadratico medio. Notiamo intanto che

$$\begin{aligned} r(d,\theta) &= \mathrm{Var}(d) \\ &= \lambda^2 \mathrm{Var}(d_1) + (1-\lambda)^2 \mathrm{Var}(d_2) \qquad \text{per l'indipendenza di } d_1 \text{ e } d_2 \\ &= \lambda^2 \sigma_1^2 + (1-\lambda)^2 \sigma_2^2 \end{aligned}$$

Per minimizzare questa espressione, ne calcoliamo la derivata,

$$\frac{d}{d\lambda} r(d,\theta) = 2\lambda\sigma_1^2 - 2(1-\lambda)\sigma_2^2$$

e ne studiamo il segno, denotando con $\hat{\lambda}$ il valore di λ che produce il minimo,

$$2\hat{\lambda}\sigma_1^2 - 2(1-\hat{\lambda})\sigma_2^2 = 0$$

da cui

$$\hat{\lambda} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2} \tag{7.7.3}$$

Altrimenti detto, il peso ottimale da dare a uno stimatore deve essere inversamente proporzionale alla sua varianza (solo nell'ipotesi che tutti gli stimatori siano corretti e indipendenti).

Per vedere una applicazione di quanto detto, immaginiamo che una associazione per la conservazione ambientale voglia determinare l'acidità delle acque di un certo lago. Raccoglie quindi dei campioni d'acqua che invia a n diversi laboratori di analisi. Questi ultimi effettueranno la titolazione indipendentemente l'uno dagli altri, ciascuno con le proprie attrezzature, dotate di livelli di precisione diversi. In particolare, ipotizziamo che per i che va da 1 a n, d_i sia il risultato delle analisi del laboratorio i – una variabile aleatoria con media pari al livello vero di acidità θ, e con varianza σ_i^2. Se le varianze sono conosciute, l'associazione dovrebbe stimare l'acidità dei campioni d'acqua con

$$d = \frac{\sum_{i=1}^n d_i/\sigma_i^2}{\sum_{i=1}^n 1/\sigma_i^2} \tag{7.7.4}$$

che è la migliore combinazione lineare delle d_i per quanto riguarda l'errore quadratico medio:

$$\begin{aligned} r(d,\theta) &= \mathrm{Var}(d) \qquad \text{perché } d \text{ è non distorto} \\ &= \left(\frac{1}{\sum_{i=1}^n 1/\sigma_i^2}\right)^2 \sum_{i=1}^n \left(\frac{1}{\sigma_i^2}\right)^2 \sigma_i^2 \\ &= \frac{1}{\sum_{i=1}^n 1/\sigma_i^2} \quad \square \end{aligned}$$

Il fatto che per uno stimatore non distorto l'errore quadratico medio coincida con la varianza si può generalizzare ad uno stimatore qualsiasi: la formula viene corretta sommando il quadrato del bias, come si deduce dai passaggi seguenti.

$$\begin{aligned} r(d,\theta) &= E[(d-\theta)^2] \\ &= E[(d-E[d]+E[d]-\theta)^2] \\ &= E[(d-E[d])^2 + 2(d-E[d])(E[d]-\theta) + (E[d]-\theta)^2] \\ &= E[(d-E[d])^2] + 2(E[d]-\theta)E[d-E[d]] + E[(E[d]-\theta)^2] \\ &= \mathrm{Var}(d) + 0 + E[b_\theta(d)^2] \qquad \text{perché } d - E[d] \text{ ha media nulla} \\ &= \mathrm{Var}(d) + b_\theta(d)^2 \end{aligned} \tag{7.7.5}$$

Esempio 7.7.3. Sia $X_1, X_2, \ldots, X_n$ un campione aleatorio estratto da una popolazione con distribuzione uniforme su $(0, \theta)$, dove θ è un parametro incognito. Poiché $E[X_i] = \theta/2$, uno stimatore corretto "naturale" per θ è dato da

$$d_1 = d_1(\boldsymbol{X}) := 2\overline{X} := \frac{2}{n}\sum_{i=1}^{n} X_i \tag{7.7.6}$$

Siccome $E[d_1] = \theta$, si ottiene che

$$\begin{aligned} r(d_1, \theta) &= \mathrm{Var}(d_1) \\ &= \frac{4}{n}\mathrm{Var}(X_i) \\ &= \frac{4}{n}\frac{\theta^2}{12} \qquad \text{per l'Equazione (5.4.4)} \\ &= \frac{\theta^2}{3n} \end{aligned}$$

Un secondo stimatore possibile per θ è quello di massima verosimiglianza, che, nell'Esempio 7.2.6 abbiamo dimostrato essere

$$d_2 = d_2(\boldsymbol{X}) := \max_i X_i \tag{7.7.7}$$

Per calcolare l'errore quadratico medio di d_2 occorre prima conoscere la sua media (per ottenere il bias) e la sua varianza. Cerchiamo per cominciare la funzione di ripartizione.

$$\begin{aligned} F_2(x) &= P(d_2(\boldsymbol{X}) \le x) \\ &= P(\max_i X_i \le x) \\ &= P(X_1 \le x, X_2 \le x, \ldots, X_n \le x) \\ &= \prod_{i=1}^{n} P(X_i \le x) \qquad \text{per l'indipendenza} \\ &= F_{X_i}(x)^n = \left(\frac{x}{\theta}\right)^n, \qquad 0 \le x \le \theta \end{aligned}$$

Derivando la funzione di ripartizione si trova la densità di d_2,

$$f_2(x) = \frac{nx^{n-1}}{\theta^n}, \qquad 0 \le x \le \theta$$

e quindi possiamo calcolare i primi due momenti e la varianza di d_2,

$$E[d_2] = \int_0^\theta x\frac{nx^{n-1}}{\theta^n}\,dx = \frac{n}{n+1}\theta \tag{7.7.8}$$

$$\begin{aligned} E[d_2^2] &= \int_0^\theta x^2\frac{nx^{n-1}}{\theta^n}\,dx = \frac{n}{n+2}\theta^2 \\ \mathrm{Var}(d_2) &= E[d_2^2] - E[d_2]^2 \\ &= \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 \\ &= n\theta^2\left[\frac{1}{n+2} - \frac{n}{(n+1)^2}\right] \\ &= \frac{n\theta^2}{(n+2)(n+1)^2} \end{aligned} \tag{7.7.9}$$

Quindi

$$\begin{aligned} r(d_2, \theta) &= \mathrm{Var}(d_2) + (E[d_2] - \theta)^2 \\ &= \frac{n\theta^2}{(n+2)(n+1)^2} + \frac{\theta^2}{(n+1)^2} \\ &= \frac{\theta^2}{(n+1)^2}\left[\frac{n}{n+2} + 1\right] \\ &= \frac{2\theta^2}{(n+1)(n+2)} \end{aligned} \tag{7.7.10}$$

Possiamo ora confrontare i due valori trovati per gli errori quadratici medi di d_1 e d_2, e siccome per ogni $n = 1, 2, \ldots,$

$$\frac{2\theta^2}{(n+1)(n+2)} \le \frac{\theta^2}{3n}$$

ne segue che d_2 è migliore di d_1 come stimatore per θ.

L'espressione per il valore atteso di d_2 fornita dall'Equazione (7.7.8), suggerisce ancora un altro stimatore, infatti se la media di d_2 è $n \cdot \theta/(n+1)$, allora

$$\frac{n+1}{n}d_2 = \frac{n+1}{n}\max_i X_i$$

è sicuramente uno stimatore corretto. Comunque, piuttosto che calcolare l'errore quadratico medio di questo stimatore particolare, consideriamo tutti quelli della forma

$$d_c(\boldsymbol{X}) := c \cdot \max_i X_i = c \cdot d_2(\boldsymbol{X}) \tag{7.7.11}$$

dove c è una costante assegnata. Il corrispondente errore quadratico medio è

$$\begin{aligned} r(d_c,\theta) &= \mathrm{Var}(d_c) + (E[d_c]-\theta)^2 \\ &= c^2\,\mathrm{Var}(d_2) + (cE[d_2]-\theta)^2 && \text{per la (7.7.11)} \\ &= \frac{c^2 n\theta^2}{(n+2)(n+1)^2} + \theta^2\left(c\frac{n}{n+1}-1\right)^2 && \text{per la (7.7.9) e la (7.7.8)} \end{aligned} \tag{7.7.12}$$

Per determinare la costante c^* cui corrisponda lo stimatore con il minore errore quadratico medio tra tutti quelli del tipo $d_c(\boldsymbol{X})$, deriviamo l'espressione di $r(d_c,\theta)$,

$$\begin{aligned} \frac{d}{dc}r(d_c,\theta) &= \frac{2cn\theta^2}{(n+2)(n+1)^2} + \frac{2n\theta^2}{n+1}\left(c\frac{n}{n+1}-1\right) \\ &= \frac{2n\theta^2}{(n+1)^2}\left[\frac{c}{n+2} + cn - (n+1)\right] \end{aligned}$$

quindi la poniamo uguale a zero,

$$\frac{c^*}{n+2} + c^* n - (n+1) = 0$$

ricaviamo c^*,

$$c^* = \frac{(n+1)(n+2)}{n^2+2n+1} = \frac{n+2}{n+1}$$

e infine scopriamo che il migliore stimatore tra quelli del tipo $d_c(\boldsymbol{X})$ è costituito da

$$\frac{n+2}{n+1}\max_i X_i \tag{7.7.13}$$

Si tratta di uno stimatore distorto con errore quadratico medio che (sostituendo c^* nell'Equazione (7.7.12)) è dato da

$$\begin{aligned} r(d_{c^*},\theta) &= \frac{n(n+2)\theta^2}{(n+1)^4} + \theta^2\left(\frac{n(n+2)}{(n+1)^2}-1\right)^2 \\ &= \frac{n(n+2)\theta^2}{(n+1)^4} + \frac{\theta^2}{(n+1)^4} \\ &= \frac{\theta^2}{(n+1)^2} \end{aligned} \tag{7.7.14}$$

Un confronto con l'Equazione (7.7.10) ci permette di concludere che anche se l'ultimo stimatore trovato non è corretto (ha un bias non nullo), il suo errore quadratico medio è poco più della metà di quello dello stimatore di massima verosimiglianza. □

7.8 * Stimatori bayesiani

Vista la indeterminazione del parametro incognito θ, in alcune situazioni può essere ragionevole considerarlo assumere la forma una variabile aleatoria: il valore vero del parametro da stimare diviene quindi il numero realizzato dalla variabile aleatoria. Tale approccio, che viene detto *bayesiano*, è di norma giustificato quando, prima di osservare gli esiti del campione di dati $X_1, X_2, \ldots, X_n$, abbiamo delle informazioni su quelli che possono essere i valori assunti da θ, e magari sulla loro plausibità. Se queste informazioni a priori assumono la forma di una distribuzione di probabilità, questa prende appropriatamente il nome di distribuzione *a priori* per θ (in inglese è la *prior distribution*). Per esempio supponiamo che, dall'esperienza passata, ci si aspetti che θ possa avere un qualunque valore compreso tra 0 e 1, ma non valori esterni a quell'intervallo. Se inoltre θ ha le stesse possibilità di essere vicino a qualunque punto di $(0,1)$, possiamo ragionevolmente assumere che si tratti di una variabile aleatoria uniforme su $(0,1)$.

Supponiamo allora di potere esprimere le nostre considerazioni a priori su θ nella forma di una distribuzione continua, con densità di probabilità $p(\theta)$; osserviamo i valori di un campione di dati la cui distribuzione dipende da θ, e denotiamo con $f(x|\theta)$ la funzione di likelihood – si tratta quindi della funzione di massa di probabilità nel caso discreto, oppure della funzione di densità di probabilità nel caso continuo – che esprime la plausibilità che uno dei dati sia uguale a x quando θ è il valore del parametro. Se i valori osservati sono $X_i = x_i$, per $i = 1, 2, \ldots, n$, allora la densità di probabilità condizionale di θ è data da

$$\begin{aligned} f(\theta|x_1,x_2,\ldots,x_n) &= \frac{f(x_1,x_2,\ldots,x_n,\theta)}{f(x_1,x_2,\ldots,x_n)} \\ &= \frac{f(x_1,x_2,\ldots,x_n|\theta)p(\theta)}{\int f(x_1,x_2,\ldots,x_n|\theta')p(\theta')\,d\theta'} \end{aligned} \tag{7.8.1}$$

La densità condizionale $f(\theta|x_1,x_2,\ldots,x_n)$ è detta densità di probabilità *a posteriori*. (Prima dell'osservazione dei dati la nostra previsione di θ è espressa dalla distribuzione a priori; dopo di essa la distribuzione viene aggiornata divenendo quella a posteriori.)

Come il lettore attento ricorderà, abbiamo dimostrato nell'Osservazione 4.5.1 di pagina 122 che quando conosciamo la distribuzione di una variabile aleatoria, la migliore stima del suo valore (in termini di errore quadratico medio) è data dalla media. Quindi, la migliore stima di θ, assegnati i valori dei dati $X_i = x_i$, per $i = 1, \ldots, n$, è data dalla media della distribuzione a posteriori $f(\theta|x_1,x_2,\ldots,x_n)$. Lo stimatore appena descritto è detto *stimatore bayesiano*, si indica con $E[\theta|X_1,X_2,\ldots,X_n]$ e il suo valore si calcola nel modo usuale:

$$E[\theta|X_1 = x_1,\ldots,X_n = x_n] = \int_{-\infty}^{\infty} \theta f(\theta|x_1,x_2,\ldots,x_n)\,d\theta \tag{7.8.2}$$

Esempio 7.8.1. Supponiamo che $X_1, X_2, \ldots, X_n$ siano variabili aleatorie i.i.d. di Bernoulli, con funzione di massa di probabilità

$$f(x|\theta) = \theta^x(1-\theta)^{1-x}, \qquad x = 0, 1$$

dove θ è un parametro sconosciuto. Supponiamo che la distribuzione a priori di θ sia uniforme su $(0, 1)$, e calcoliamo lo stimatore bayesiano di θ.

Denotiamo con p la densità a priori di θ,

$$p(\theta) = 1, \qquad 0 < \theta < 1$$

La densità condizionale di θ date $x_1, x_2, \ldots, x_n$ è data da

$$\begin{aligned} f(\theta|x_1, x_2, \ldots, x_n) &= \frac{f(x_1, x_2, \ldots, x_n, \theta)}{f(x_1, x_2, \ldots, x_n)} \\ &= \frac{f(x_1, x_2, \ldots, x_n|\theta)p(\theta)}{\int_0^1 f(x_1, x_2, \ldots, x_n|\vartheta)p(\vartheta)\,d\vartheta} \\ &= \frac{\theta^{\sum x_i}(1-\theta)^{n-\sum x_i}}{\int_0^1 \vartheta^{\sum x_i}(1-\vartheta)^{n-\sum x_i}\,d\vartheta} \end{aligned}$$

Non è difficile provare (integrando per parti un certo numero di volte) che per ogni valore intero di m e r,

$$\int_0^1 \theta^m(1-\theta)^r\,d\theta = \frac{m!r!}{(m+r+1)!} \tag{7.8.3}$$

Quindi ponendo $x := \sum_{i=1}^n x_i$,

$$f(\theta|x_1, x_2, \ldots, x_n) = \frac{(n+1)!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}, \qquad 0 < \theta < 1 \tag{7.8.4}$$

Siamo ora in grado di calcolare la stima bayesiana.

$$\begin{aligned} E[\theta|x_1, x_2, \ldots, x_n] &= \frac{(n+1)!}{x!(n-x)!}\int_0^1 \theta^{1+x}(1-\theta)^{n-x}\,d\theta \\ &= \frac{(n+1)!}{x!(n-x)!}\,\frac{(1+x)!(n-x)!}{(n+2)!} \qquad \text{usando la (7.8.3)} \\ &= \frac{x+1}{n+2} \end{aligned}$$

di conseguenza lo stimatore bayesiano è dato da

$$E[\theta|X_1, X_2, \ldots, X_n] = \frac{1+\sum_{i=1}^n X_i}{n+2} \tag{7.8.5}$$

Per illustrare il risultato, se raccogliendo un campione di 10 bernoulliane trovassimo 6 successi, lo stimatore bayesiano di θ con distribuzione a priori uniforme su $(0, 1)$, fornirebbe un valore di $7/12$. Si noti che lo stimatore di massima verosimiglianza varrebbe invece $6/10$. □

Osservazione 7.8.1. La distribuzione condizionale di θ dati $x_1, x_2, \ldots, x_n$, la cui densità compare nell'Equazione (7.8.4), è detta *distribuzione beta* di parametri $x+1$ e $n-x+1$.

Esempio 7.8.2. Supponiamo che $X_1, X_2, \ldots, X_n$ sia un campione proveniente da una distribuzione normale di media incognita θ e varianza nota σ_0^2. Se la distribuzione a priori di θ è pensata essere normale di media μ e varianza σ^2, qual è lo stimatore bayesiano per θ?

Per determinare lo stimatore bayesiano $E[\theta|X_1, X_2, \ldots, X_n]$, dobbiamo prima ottenere la densità condizionale di θ dati i valori di $X_1, X_2, \ldots, X_n$:

$$f(\theta|x_1, x_2, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n|\theta)p(\theta)}{f(x_1, x_2, \ldots, x_n)}$$

dove

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|\theta) &= \frac{1}{(2\pi)^{n/2}\sigma_0^n}\exp\left\{-\sum_{i=1}^n \frac{(x_i-\theta)^2}{2\sigma_0^2}\right\} \\ p(\theta) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(\theta-\mu)^2}{2\sigma^2}\right\} \\ f(x_1, x_2, \ldots, x_n) &= \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n|\theta)p(\theta)\,d\theta \end{aligned}$$

Con l'aiuto di un po' di algebra è possibile dimostrare che questa distribuzione condizionale è anch'essa di tipo normale; in particolare ha media

$$\begin{aligned} E[\theta|X_1, X_2, \ldots, X_n] &= \frac{n\sigma^2}{n\sigma^2+\sigma_0^2}\overline{X} + \frac{\sigma_0^2}{n\sigma^2+\sigma_0^2}\mu \\ &= \frac{n/\sigma_0^2}{n/\sigma_0^2+1/\sigma^2}\overline{X} + \frac{1/\sigma^2}{n/\sigma_0^2+1/\sigma^2}\mu \end{aligned} \tag{7.8.6}$$

e varianza

$$\mathrm{Var}(\theta|X_1, X_2, \ldots, X_n) = \frac{\sigma_0^2\sigma^2}{n\sigma^2+\sigma_0^2} = \frac{1}{n/\sigma_0^2+1/\sigma^2} \tag{7.8.7}$$

L'espressione della media condizionale nella seconda formulazione data qui sopra è molto significativa, in quanto ha la forma di una media pesata della media campionaria $\overline{X}$, e della media *a priori* μ. I pesi inoltre sono proporzionali all'inverso di σ_0^2/n (la varianza condizionale della media campionaria $\overline{X}$ data θ) e σ^2 (la varianza della distribuzione a priori). □

Sulla scelta di una distribuzione a priori normale

Come è evidenziato dall'Esempio 7.8.2, è computazionalmente molto conveniente scegliere una distribuzione a priori normale per la media incognita θ di un'altra distribuzione normale – in tal modo infatti lo stimatore bayesiano è semplicemente dato dall'Equazione (7.8.6). Questo solleva la questione di come si possa capire se vi sia una distribuzione normale che può rappresentare le nostre supposizioni a priori sulla media incognita.

Per cominciare, sembra ragionevole individuare un valore μ, che a priori pensiamo essere vicino a θ. Ciò equivale a fissare la moda della distribuzione a priori (per una distribuzione normale media e moda coincidono). Secondariamente dovremmo chiarirci se pensiamo che la distribuzione a priori sia simmetrica rispetto a μ. Dobbiamo domandarci se per ogni valore di $a > 0$ siamo convinti che sia altrettanto plausibile trovare θ nell'intervallo $(\mu - a, \mu)$ che nell'intervallo $(\mu, \mu + a)$. Se la risposta è positiva, possiamo accettare come ipotesi di lavoro, che le nostre idee a priori su θ possano essere espresse in termini di una distribuzione a priori normale con media μ. Per determinare la deviazione standard a priori σ, cerchiamo un intervallo centrato su μ che crediamo a priori che abbia il 90% di chances di contenere θ. Ad esempio, supponiamo di esserci convinti che vi sia il 90% di possibilità (non di meno e non di più) che θ starà in un certo intervallo $(\mu - a, \mu + a)$. Allora, visto che per una normale $\theta \sim \mathcal{N}(\mu, \sigma^2)$ vale

$$P\left(-1.645 < \frac{\theta - \mu}{\sigma} < 1.645\right) \approx 0.90$$

ovvero

$$P(\mu - 1.645\sigma < \theta < \mu + 1.645\sigma) \approx 0.90$$

sembra ragionevole porre $a = 1.645\sigma$ e ricavare $\sigma = a/1.645$

Se le nostre convinzioni a priori devono essere compatibili con una distribuzione normale, essa dovrà perciò avere media μ e deviazione standard $\sigma = a/1.645$. Questa ipotesi può essere ulteriormente verificata ponendosi successivamente altre domande, come ad esempio se vi sia il 95% di confidenza che θ appartenga a $\mu \pm 1.96\sigma$ e il 99% che appartenga a $\mu \pm 2.58\sigma$; questi intervalli sono determinati dalle probabilità seguenti, che sono valide nell'ipotesi che θ sia normale con media μ e varianza σ^2.

$$P\left(-1.96 < \frac{\theta - \mu}{\sigma} < 1.96\right) \approx 0.95 \quad P\left(-2.58 < \frac{\theta - \mu}{\sigma} < 2.58\right) \approx 0.99$$

Esempio 7.8.3. Sia data una funzione di likelihood $f(x_1, x_2, \ldots, x_n|\theta)$, e supponiamo che la distribuzione a priori di θ sia uniforme su un certo intervallo (a, b). La densità a posteriori di θ dati i valori $x_1, x_2, \ldots, x_n$ del campione $X_1, X_2, \ldots, X_n$ è data da

$$\begin{aligned} f(\theta|x_1, x_2, \ldots, x_n) &= \frac{f(x_1, x_2, \ldots, x_n|\theta)p(\theta)}{\int_a^b f(x_1, x_2, \ldots, x_n|\theta')p(\theta')\,d\theta'} \\ &= \frac{f(x_1, x_2, \ldots, x_n|\theta)}{\int_a^b f(x_1, x_2, \ldots, x_n|\theta')\,d\theta'}, \qquad a < \theta < b \end{aligned}$$

Siccome la *moda* di una densità $f(\theta)$ è stata definita come quel valore di θ che massimizza la densità stessa, si vede bene che la moda della densità a posteriori $f(\theta|x_1, x_2, \ldots, x_n)$ è anche il valore di θ che massimizza $f(x_1, x_2, \ldots, x_n|\theta)$, e per questo è uguale allo stimatore di massima verosimiglianza (a patto che si imponga a θ di stare tra a e b). In conclusione, se si prende una distribuzione a priori uniforme, la moda della distribuzione a posteriori coincide con lo stimatore di massima verosimiglianza. □

Se invece di uno stimatore puntuale desideriamo trovare un intervallo in cui θ stia con una probabilità assegnata, diciamo $1 - \alpha$, possiamo ottenerlo prendendo due valori a e b in modo tale che

$$\int_a^b f(\theta|x_1, x_2, \ldots, x_n)\,d\theta = 1 - \alpha \tag{7.8.8}$$

Esempio 7.8.4. Consideriamo la trasmissione da una sorgente A di un segnale di valore s. Il segnale ricevuto da B ha distribuzione $\mathcal{N}(s, 60)$, a causa del rumore del canale di trasmissione. Supponiamo anche di sapere a priori che il segnale inviato sia normale $\mathcal{N}(50, 100)$. Si determini un intervallo che contenga il valore inviato col 90% di probabilità, nel caso in cui il valore ricevuto da B sia 40.

Segue dall'Esempio 7.8.2 che la distribuzione condizionale del segnale inviato S, sapendo di avere ricevuto 40, è normale con media e varianza date da

$$\begin{aligned} E[S|\text{dati}] &= \frac{1/60}{1/60 + 1/100}40 + \frac{1/100}{1/60 + 1/100}50 = 43.75 \\ \text{Var}(S|\text{dati}) &= \frac{1}{1/60 + 1/100} = 37.5 \end{aligned}$$

Quindi, condizionando al ricevimento del valore 40, $(S - 43.75)/\sqrt{37.5}$ ha distribuzione normale standard, e

$$\begin{aligned} 0.90 &\approx P\left(-1.645 < \frac{S - 43.75}{\sqrt{37.5}} < 1.645 \,\middle|\, \text{dati}\right) \\ &= P\left(43.75 - 1.645\sqrt{37.5} < S < 43.75 + 1.645\sqrt{37.5} \,\middle|\, \text{dati}\right) \end{aligned}$$

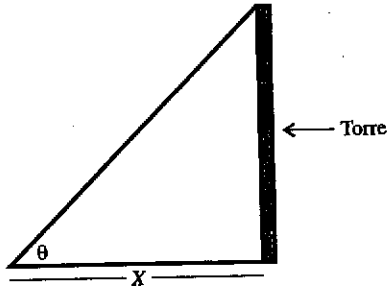


Figura 7.6 Come misurare l'altezza di una torre.

Perciò, con *probabilità* di 0.90, il segnale realmente inviato appatiene all'intervallo (33.68, 53.82). □

Problemi

1. Sia $X_1, X_2, \ldots, X_n$ un campione proveniente da una distribuzione di densità

$$f(x) = \begin{cases} e^{-(x-\theta)} & x \geq \theta \\ 0 & \text{altrimenti} \end{cases}$$

Determina lo stimatore di massima verosimiglianza di θ.

2. Sia $X_1, X_2, \ldots, X_n$ un campione proveniente da una distribuzione di densità

$$f(x) = \frac{1}{2} e^{-|x-\theta|}$$

Determina lo stimatore di massima verosimiglianza di θ.

3. Sia $X_1, X_2, \ldots, X_n$ un campione proveniente da una popolazione normale $\mathcal{N}(\mu, \sigma^2)$.

(a) Determina lo stimatore di massima verosimiglianza per σ^2 nel caso in cui la media μ sia nota.

(b) Qual è il valore atteso di tale stimatore?

4. Vogliamo misurare l'altezza di una torre per le telecomunicazioni sfruttando la distanza orizzontale X tra la sua base e la nostra posizione, e l'angolo verticale θ sotto cui la torre viene vista a tale distanza (si faccia riferimento alla Figura 7.6). Le 5 misurazioni della distanza X hanno dato (in piedi) i valori seguenti

150.42 150.45 150.49 150.52 150.40

Le 4 misurazioni dell'angolo θ hanno dato in gradi

40.26 40.27 40.29 40.26

Stima l'altezza della torre.

5. Fai girare una moneta sul suo bordo (come una trottola) per 100 volte, e usa i risultati ottenuti per stimare quale sia la probabilità di ottenere testa tirandola in questo modo.

6. Le piene dei fiumi vengono misurate tramite la loro portata (espressa di seguito in piedi cubi al secondo). Un numero v è detto valore di una piena secolare se

$$P(D \geq v) = 0.01$$

dove D è la portata della più grande piena in un anno a caso. La tabella seguente riporta le portate delle maggiori piene del fiume Blackstone River, a Woonsocket nel Rhode Island, negli anni da 1929 al 1965. Assumendo che la distribuzione di questi dati sia lognormale, stima il valore di una piena secolare.

Anno	Portata	Anno	Portata
1929	4 570	1948	5 810
1930	1 970	1949	2 030
1931	8 220	1950	3 620
1932	4 530	1951	4 920
1933	5 780	1952	4 090
1934	6 560	1953	5 570
1935	7 500	1954	9 400
1936	15 000	1955	32 900
1937	6 340	1956	8 710
1938	15 100	1957	3 850
1939	3 840	1958	4 970
1940	5 860	1959	5 398
1941	4 480	1960	4 780
1942	3 330	1961	4 020
1943	5 310	1962	5 790
1944	3 830	1963	4 510
1945	3 410	1964	5 520
1946	3 830	1965	5 300
1947	3 150		

7. Un produttore di scambiatori di calore richiede che la distanza tra le piastre degli scambiatori sia compresa tra 0.240 e 0.260 pollici. Un ingegnere che si occupa di controllo di qualità campiona 20 scambiatori e misura questa distanza, trovando che la media e la deviazione standard campionarie sono rispettivamente di 0.254 e 0.005 pollici. Stima la frazione di scambiatori che cadrà al di fuori dell'intervallo richiesto, assumendo che la distribuzione dei dati sia gaussiana.

8. Il peso misurato da una bilancia elettronica è quello reale dell'oggetto più un errore casuale che ha distribuzione normale di media 0 e deviazione standard 0.01 (in milligrammi). Supponiamo che i risultati di 5 pesate successive dello stesso oggetto abbiano dato i valori

3.142 3.163 3.155 3.150 3.141

Determina un intervallo di confidenza per il peso reale dell'oggetto ad un livello di confidenza (a) del 95%; (b) del 99%.

9. La concentrazione di PCB presenti nei pesci del lago Michigan viene misurata con una tecnica che porta ad un errore che ha distribuzione normale di deviazione standard 0.8 ppm (parti per milione). Supponiamo che i risultati di 10 analisi indipendenti su una quantità di pesce abbiano dato i seguenti valori,

11.2 12.4 10.8 11.6 12.5 10.1 11.0 12.2 12.4 10.6

Trova, per la concentrazione di PCB nel pesce, gli intervalli di confidenza al 95% seguenti: (a) quello bilaterale, (b) quello unilaterale sinistro, (c) quello unilaterale destro.

10. La deviazione standard per i punteggi dei candidati ad un certo esame pubblico ha tipicamente un valore di 11.3. Se quest'anno un primo campione di 81 candidati presenta una punteggio medio di 74.6, qual è l'intervallo di confidenza bilaterale al 90% per il punteggio medio di tutti i candidati?

11. Volendo determinare un intervallo di confidenza per la media di una popolazione normale di varianza nota, quanto numeroso deve essere il campione se vogliamo che l'intervallo risultante abbia ampiezza pari ad un terzo di quello che si ottiene con un campione di numerosità n?

12. Dimostra che $(-\infty, \overline{X} + z_\alpha \cdot \sigma/\sqrt{n})$ è l'intervallo di confidenza unilaterale sinistro con livello di confidenza $1-\alpha$ per la media di una popolazione normale di varianza nota σ^2, avendo a disposizione un campione $X_1, X_2, \ldots, X_n$.

13. Si analizza un campione di 20 sigarette per determinarne il contenuto di nicotina, e il valore medio dei dati ottenuti è di 1.2 mg. Calcola un intervallo di confidenza bilaterale al 99% per il contenuto medio di nicotina di quel tipo di sigarette, sapendo che la deviazione standard è di 0.2 mg.

14. Con riferimento al Problema 13, supponiamo di non conoscere la varianza della popolazione e che quella campionaria proveniente dall'esperimento sia risultata essere 0.04. Calcola un intervallo di confidenza bilaterale al 99% per il contenuto medio di nicotina di una sigaretta.

15. Con riferimento al Problema 14, determina un valore c che permetta di affermare con il 99% di confidenza che c è maggiore del contenuto medio di nicotina di una sigaretta.

16. Supponiamo di volere stimare la media di una popolazione normale che ha entrambi i parametri incogniti. In particolare cerchiamo di determinare che numerosità deve avere il campione affinché ad un livello di confidenza $1-\alpha$, l'intervallo di confidenza bilaterale abbia ampiezza non più grande di A. Spiega come si possa realizzare approssimativamente questo progetto tramite un doppio campionamento che preveda di raccogliere un campione preliminare di ampiezza 30 e usarne i dati per dimensionare il campione definitivo.

17. I dati seguenti sono il risultato di 24 misurazioni indipendenti del punto di fusione del piombo (espressi in gradi Celsius),

330 328.6 342.4 334 337.5 341 343.3 329.5
322 331 340.4 326.5 327.3 340 331 332.3
345 342 329.7 325.8 322.6 333 341 340

Assumendo che questi dati possano essere pensati come un campione normale la cui media è il vero punto di fusione del piombo, determina gli intervalli di confidenza bilaterali per questo valore: (a) al 95% di confidenza; (b) al 99% di confidenza.

18. Quelli che seguono sono i punteggi dei test del Q.I. di un campione casuale di 18 studenti di una certa università.

130 122 119 142 136 127 120 152 141
132 127 118 150 141 133 137 129 142

Costruisci, per il punteggio di Q.I. medio degli studenti di quella università, gli intervalli di confidenza al 95% seguenti: (a) quello bilaterale, (b) quello unilaterale sinistro, (c) quello unilaterale destro.

19. Un campione di 9 prezzi di abitazioni vendute recentemente in una certa città, ha media campionaria di \$ 122 000 e deviazione standard campionaria di \$ 12 000. Determina un intervallo unilaterale destro che contenga il prezzo medio attuale delle abitazioni, con un livello di confidenza del 95%.

20. Una compagnia vuole assicurare il suo vasto parco auto contro i tamponamenti. Per determinare il costo medio di riparazione per collisione, vengono scelti a caso 16 incidenti, e ne risultano una media campionaria di \$ 2 200 e una deviazione standard campionaria di \$ 800. Trova un intervallo di confidenza al 90% per il costo medio delle riparazioni di un tamponamento.

21. Nello stato di Washington ogni anno gli alunni del sesto anno della scuola dell'obbligo vengono sottoposti ad un esame. Un sovraintendente all'istruzione che vuole conoscere il punteggio medio degli alunni del suo distretto, seleziona un campione casuale di 100 studenti, e ottiene una media e una deviazione standard campionarie di 320 e 16 punti rispettivamente. Fornisci un intervallo di confidenza bilaterale al 95% per il punteggio medio degli alunni del distretto.

22. Venti studenti di scienze misurano il punto di fusione del piombo. La media e la deviazione standard campionarie dei dati ottenuti sono 330.2 e 15.4 gradi Celsius rispettivamente. Costruisci degli intervalli di confidenza bilaterali per il punto di fusione del piombo, ad un livello di confidenza (a) del 95%; e (b) del 99%.

23. Controllando un campione aleatorio di 300 titolari di carte di credito si evince dai loro conti che il debito medio è di \$ 1 220, con una deviazione standard campionaria di \$ 840. Costruisci un intervallo di confidenza al 95% per stimare il debito medio della totalità dei possessori di carte di credito.

24. Con riferimento al Problema 23, trova il più piccolo valore v che permetta di affermare con il 90% di confidenza che il debito medio di tutti i possessori di carte di credito gli sia inferiore.

25. Verifica la formula presentata nella Tabella 7.1 per l'intervallo di confidenza sinistro per μ quando σ^2 non è nota.

26. Si investiga la gittata di un nuovo tipo di proiettile da mortaio. Le gittate in metri, che vengono osservate testando 20 proiettili sono le seguenti,

2100	1950	2043	2210	2018	1984	1992
2218	2152	2106	2072	2096	2244	1962
1938	1898	2103	2206	2007	1956	

Assumendo che la distribuzione delle gittate sia normale, si determini

(a) un intervallo di confidenza al 95% per la gittata media dei proiettili;

(b) un intervallo di confidenza al 99% dello stesso tipo;

(c) il più grande valore v che con il 95% di confidenza è inferiore alla gittata media indagata.

27. A Los Angeles sono stati condotti degli studi per determinare la concentrazione di monossido di carbonio vicino alle autostrade. La tecnica base utilizzata consiste nel catturare campioni di aria in speciali borse e poi misurarne il contenuto di monossido usando uno spettrofotometro. Le misurazioni in ppm (parti per milione) durante il periodo di campionamento sono state

102.2	98.4	104.1	101	102.2	100.4	98.6	88.2	78.8	83
84.7	94.8	105.1	106.2	111.2	108.3	105.2	103.2	99	98.8

Calcola un intervallo di confidenza al 95% per la concentrazione media di monossido di carbonio nell'aria.

28. Un insieme di 10 determinazioni della percentuale di acqua contenuta in una soluzione di metanolo, eseguite secondo un metodo ideato dal chimico Karl Fischer, hanno riportato i valori seguenti,

0.50	0.55	0.53	0.56	0.54	0.57	0.52	0.60	0.55	0.58

Supponendo che la distribuzione fosse normale, usa questi dati per costruire un intervallo di confidenza al 95% per la percentuale reale.

29. Assegnata una successione $U_1, U_2, \ldots$ di variabili i.i.d. e uniformi su $(0,1)$, si pone

$$N := \min\{n : U_1 + U_2 + \cdots + U_n > 1\}$$

In tal modo N denota il numero di variabili aleatorie uniformi su $(0,1)$ che è necessario sommare per superare il valore di 1. Realizza una simulazione al calcolatore per generare 36 variabili distribuite come N e tra loro indipendenti, quindi usa questi dati per ottenere un intervallo di confidenza al 95% per $E[N]$. Basandoti infine sull'intervallo trovato, prova a indovinare il valore esatto di $E[N]$.

30. Una questione importante per i venditori al dettaglio è come decidere quando sia il momento di ordinare la merce dal distributore. Una pratica molto comune per prendere questa decisione è quella detta di tipo s, S; essa consiste nel fare l'ordinazione quando la quantità di merce in magazzino scende al di sotto di s, richiedendone abbastanza da portarla fino a S. I valori appropriati dei parametri s e S dipendono da diversi fattori di costo, come lo stoccaggio, il profitto per pezzo venduto, e la distribuzione della domanda in un periodo di tempo. È quindi fondamentale per il venditore raccogliere dati collegati ai parametri della distribuzione della domanda. Supponiamo che i valori di seguito riportati rappresentino il numero di oggetti di un certo tipo, venduti in ciascuna di 30 settimane,

14	8	12	9	5	22	15	12	16	7	10	9	15	15	12
9	11	16	8	7	15	13	9	5	18	14	10	13	7	11

Assumendo che i numeri delle vendite delle diverse settimane siano variabili aleatorie indipendenti, provenienti dalla stessa distribuzione, usa questi dati per ottenere un intervallo di confidenza al 95% per il numero medio di vendite alla settimana.

31. Un campione casuale di 16 professori ordinari di una grande università privata ha una media campionaria del reddito annuale di \$ 90 450, con una deviazione standard campionaria di \$ 9 400. Determina un intervallo di confidenza al 95% per lo stipendio medio di tutti i professori ordinari di quella università.

32. Sia $X_1, X_2, \ldots, X_n, X_{n+1}$ un campione casuale proveniente da una popolazione normale di media μ e varianza σ^2, entrambe incognite. Siamo interessati a utilizzare i valori osservati di $X_1, X_2, \ldots, X_n$ per determinare un intervallo – detto *di predizione* – che conterrà il valore di X_{n+1} con un livello di confidenza $1-\alpha$. Denotiamo con $\overline{X}_n$ e S_n^2 la media e la varianza campionarie di $X_1, X_2, \ldots, X_n$.

(a) Trova la distribuzione di $X_{n+1} - \overline{X}_n$.

(b) Trova la distribuzione di

$$\frac{X_{n+1} - \overline{X}_n}{S_n\sqrt{1+n^{-1}}}$$

(c) Ottieni l'intervallo di predizione per X_{n+1}.

(d) L'intervallo trovato al punto (c) conterrà il valore di X_{n+1} con un livello di confidenza di $1-\alpha$. Chiarisci il significato di questa affermazione.

33. I dati ufficiali mostrano che i decessi per annegamento accidentale negli Stati Uniti per gli anni dal 1990 al 1993 sono stati (in migliaia) 5.2, 4.6, 4.3 e 4.8. Usa questa informazione per fornire un intervallo che, con il 95% di confidenza, conterrà il numero di morti per annegamento del 1994.

34. La concentrazione di ossigeno disciolto in un corso d'acqua è stata registrata per 30 giorni, ottenendo una media campionaria di 2.5 mg/l e una deviazione standard campionaria di 2.12 mg/l. Determina un valore che sia superiore alla concentrazione media giornaliera con un livello di confidenza del 90%.

35. Verifica le formule riportate nella Tabella 7.1 per gli intervalli di confidenza unilaterali per σ^2.

36. Le capacità (in ampere-ora) di 10 batterie sono risultate:

140 136 150 144 148 152 138 141 143 151

(a) Stima la varianza σ^2 della popolazione.

(b) Calcola un intervallo di confidenza al 99% per σ^2.

(c) Trova un valore v che permetta di dire con il 90% di confidenza, che $\sigma^2 < v$.

37. Trova un intervallo di confidenza bilaterale al 95% per la varianza del diametro di un rivetto, basandoti sui dati seguenti.

6.68 6.76 6.78 6.76 6.74 6.64 6.81 6.74 6.70 6.66 6.67 6.66

Puoi assumere che la popolazione sia normale.

38. I tempi di combustione (in secondi) di 10 unità di un tipo di combustibile sono risultati i seguenti

50.6 54.8 54.4 44.9 42.1 69.8 53.6 66.1 48.0 37.8

Costruisci un intervallo di confidenza bilaterale al 90% per la varianza del tempo di combustione. Puoi supporre che la distribuzione considerata sia gaussiana.

39. La quantità di berillio in una sostanza può essere determinata con metodi di filtrazione fotometrica. Se il peso del berillio è indicato con μ, il valore restituito da una misurazione di questo tipo ha distribuzione normale di media μ e deviazione standard σ. I valori seguenti sono misurazioni indipendenti di 3.180 mg di berillio.

3.166 3.192 3.175 3.180 3.182 3.171 3.184 3.177

Usa i dati precedenti per

(a) stimare σ;

(b) trovare un intervallo di confidenza al 90% per σ.

40. Sia $X_1, X_2, \ldots, X_n$ un campione proveniente da una popolazione $\mathcal{N}(\mu, \sigma^2)$. Spiega come si possa ottenere un intervallo di confidenza per σ^2, con un livello di confidenza di $1 - \alpha$, nel caso che μ sia nota. Chiarisci in quale senso la conoscenza di μ porti a un intervallo di confidenza migliore di quello che si ha quando μ non è nota.

Ripeti il Problema 38 supponendo che sia noto che la media della popolazione dei tempi di combustione è di 53.6 secondi.

41. Un ingegnere civile vuole misurare la resistenza alla compressione di due diversi tipi di calcestruzzo. Viene provato un campione di 10 esemplari per ciascuno dei due tipi di materiale, ottenendo i dati seguenti (in libbre per pollice quadrato)

Tipo 1	3250	3268	4302	3184	3266	3297	3332	3502	3064	3116
Tipo 2	3094	3106	3004	3066	2984	3124	3316	3212	3380	3018

Ipotizziamo che i campioni vengano da due popolazioni normali con la medesima varianza. Determina per la differenza delle medie delle due popolazioni, gli intervalli di confidenza al 95% seguenti: (a) quello bilaterale, (b) quello unilaterale sinistro, (c) quello unilaterale destro.

42. Si studiano campioni indipendenti di oggetti prodotti da due macchine di una linea di produzione. Siamo interessati a confrontare il loro peso. Dalla prima macchina si estrae un campione di 36 oggetti, ottenendo una media campionaria di 120 grammi e varianza campionaria 4. Dalla seconda macchina si pesano 64 oggetti, che hanno media campionaria di 130 grammi e varianza campionaria 5. Assumendo che entrambe le distribuzioni siano normali, con medie rispettivamente μ_1 e μ_2 e identica varianza σ^2, determina un intervallo di confidenza al 99% per $\mu_1 - \mu_2$.

43. Risolvi il Problema 42 con l'ipotesi aggiuntiva che 4 e 5 siano le varianze reali delle due popolazioni normali.

44. Quelli che seguono sono i tempi di combustione in secondi di alcuni esemplari di due diversi tipi di candelotti fumogeni:

Tipo I	481	506	527	661	501	572	561	501	487	524
Tipo II	526	511	556	542	491	537	582	605	558	578

Costruisci un intervallo di confidenza al 99% per la differenza media dei tempi di combustione, assumendo che le popolazioni siano normali con la stessa varianza.

45. Siano $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_m$ due campioni gaussiani indipendenti, con medie note μ_1 e μ_2 e varianze ignote σ_1^2 e σ_2^2. Determina un intervallo di confidenza ad un livello $1 - \alpha$ per il rapporto delle varianze, σ_1^2/σ_2^2.

46. Due analisti di laboratorio prendono ripetutamente delle misure sulla durezza dell'acqua di una città. Assumendo che i dati di ciascuno abbiano distribuzione normale, con varianze σ_1^2 e σ_2^2, calcola un intervallo di confidenza bilaterale al 95% per σ_1^2/σ_2^2, usando i dati seguenti,

Analista 1	0.46	0.62	0.37	0.40	0.44	0.58	0.48	0.53		
Analista 2	0.82	0.61	0.89	0.51	0.33	0.48	0.23	0.25	0.67	0.88

47. Un campione casuale di 1200 ingegneri è risultato contenere 48 ispanoamericani, 60 afroamericani e 204 femmine. Determina gli intervalli di confidenza al 90% per la frazione di ingegneri che sono (a) femmine; (b) ispanoamericani o afroamericani.

48. Per stimare la frazione p di quanti neonati siano maschi, si registra il sesso di un campione casuale di 10000 bambini appena nati. Sapendo che 5106 di essi sono risultati maschi, determina degli intervalli di confidenza per p (a) al 90% e (b) al 99%.

49. Una compagnia aerea vuole determinare qual è la percentuale dei suoi passeggeri che vola per affari. Se si volesse il 90% di confidenza che la stima abbia un errore entro il 2%, quanto numeroso dovrebbe essere il campione utilizzato?

50. Un sondaggio elettorale effettuato da un quotidiano riporta il candidato A in vantaggio sul candidato B con il 53% contro il 47% delle preferenze, con un margine di errore di ±4%. Il giornale continua dicendo che siccome la differenza di 6 punti tra i due candidati è maggiore del margine di errore, i lettori possono ritenersi certi della vittoria del candidato A. Questo ragionamento è completamente corretto?

51. Una compagnia di ricerche di mercato vuole determinare la percentuale di famiglie che stanno assistendo ad un particolare evento sportivo. Per riuscirci, effettua un sondaggio telefonico. Quante famiglie dovranno essere intervistate come minimo, se si vuole avere il 90% di confidenza che la stima non porti un errore superiore a ±0.02?

52. In uno studio recente è stato verificato che su 140 meteoriti osservati, 79 sono entrati nell'atmosfera a una velocità non superiore alle 25 miglia al secondo. Se prendiamo $\widehat{p} := 79/140$ come stima della probabilità che un qualsiasi meteorite che entra nell'atmosfera lo faccia a una velocità inferiore alle 25 miglia al secondo, cosa possiamo dire con il 99% di confidenza, sul massimo errore della nostra stima?

53. Un campione aleatorio di 100 pezzi di una linea di produzione ne conteneva 17 di difettosi. Calcola un intervallo di confidenza bilaterale al 95% per la probabilità che un pezzo qualsiasi sia difettoso. Che ipotesi stai implicitamente facendo?

54. Su 100 casi di tumore ai polmoni selezionati a caso, 67 pazienti sono deceduti entro 5 anni dalla diagnosi.

(a) Stima la probabilità che una persona che si ammala di tumore ai polmoni, muoia entro 5 anni.

(b) Quanto grande dovrebbe essere un ulteriore campione di casi, per acquisire il 95% di confidenza che la probabilità stimata nel punto (a) non sia sbagliata per più di 0.02?

55. Scrivi delle formule per gli intervalli di confidenza unilaterali per il parametro p di una distribuzione di Bernoulli, quando si conoscano i valori di n variabili aleatorie indipendenti con tale distribuzione.

***56.** Supponiamo che i tempi di vita di un tipo di batterie abbiano distribuzione esponenziale di media θ. Un campione di 10 di esse ha fornito una media campionaria di 36 ore. Trova un intervallo di confidenza bilaterale al 95% per θ.

***57.** Costruisci entrambi i tipi di intervalli di confidenza unilaterali, ad un livello di confidenza di $1-\alpha$, per il parametro θ del Problema 56.

***58.** Sia $X_1, X_2, \ldots, X_n$ un campione estratto da una popolazione di media μ incognita. Utilizza i risultati dell'Esempio 7.7.2 per dimostrare che, fra tutti gli stimatori di μ della forma $\sum_{i=1}^{n} \lambda_i X_i$, con i coefficienti che soddisfano $\sum_{i=1}^{n} \lambda_i = 1$, quello con il minimo errore quadratico medio è la *media campionaria*, che si ottiene ponendo $\lambda_i \equiv \frac{1}{n}$, per ogni i.

***59.** Consideriamo due campioni indipendenti $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_m$ provenienti da due popolazioni normali con la stessa varianza σ^2. Siano S_x^2 e S_y^2 le rispettive varianze campionarie, che sono stimatori non distorti di σ^2. Usando i risultati dell'Esempio 7.7.2 e il fatto che la varianza di una chi-quadro con k gradi di libertà è pari a $2k$, dimostra che lo stimatore di σ^2 che presenta il minore errore quadratico medio, tra tutti quelli della forma $\lambda S_x^2 + (1-\lambda) S_y^2$, è il seguente,

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

Ovvero lo stesso stimatore pooled, che abbiamo discusso nell'Osservazione 7.4.1.

***60.** Consideriamo due stimatori del parametro θ, indicati da d_1 e d_2. Quale dei due dobbiamo preferire, se $E[d_1] = \theta$, $\mathrm{Var}(d_1) = 6$ e $E[d_2] = 2 + \theta$, $\mathrm{Var}(d_2) = 2$?

***61.** Supponiamo che il numero di incidenti che si verificano quotidianamente in un certo impianto abbia distribuzione di Poisson con media λ incognita. Basandosi sulla sua esperienza con stabilimenti analoghi, uno statistico ha la convinzione a priori che i valori plausibili per λ possano essere descritti da una distribuzione esponenziale con parametro unitario, ovvero che la densità a priori sia,

$$p(\lambda) = e^{-\lambda}, \qquad \lambda > 0$$

Determina la stima bayesiana per λ ipotizzando che vi siano stati un totale di 83 incidenti negli ultimi 10 giorni. Qual è la stima di massima verosimiglianza?

***62.** I tempi di vita in ore dei circuiti integrati prodotti da una certa fabbrica di semiconduttori sono variabili aleatorie esponenziali di media $1/\lambda$. Supponiamo che la distribuzione a priori per λ sia di tipo gamma, con funzione di densità data da

$$p(\lambda) = \frac{\lambda^2}{2} e^{-\lambda}, \qquad \lambda > 0$$

Se i primi 20 chip testati mostrano un tempo di vita medio di 4.6 ore, qual è la stima bayesiana per λ?

***63.** Gli oggetti prodotti da una macchina sono difettosi, indipendentemente gli uni dagli altri, con probabilità p. Sapendo che la distribuzione a priori per p è uniforme su $(0,1)$, calcola la probabilità a posteriori che p sia minore di 0.2, sapendo che su un campione di 10 oggetti ne sono stati trovati (a) 2, (b) 1, (c) 10 di difettosi.

***64.** Si misura la resistenza allo strappo per 10 esemplari di un certo tipo di tessuto. La distribuzione di popolazione è normale con media incognita μ e deviazione standard 3 psi (libbre per pollice quadrato). Supponiamo di aspettarci dall'esperienza passata che la distribuzione a priori sia normale con media 200 e deviazione standard 2. La media campionaria dei dati del campione è risultata di 182 psi; determina una regione che contenga θ con probabilità del 95%.

8 Verifica delle ipotesi

Contenuto

8.1 Introduzione

Come nel capitolo precedente, supponiamo anche qui di disporre di un campione aleatorio proveniente da una distribuzione che ci è nota tranne che per uno o più parametri incogniti. La nuova chiave di lettura non prevede più di stimare direttamente questi parametri, ma piuttosto di utilizzare il campione raccolto per verificare qualche ipotesi che li coinvolga. Per chiarire il concetto, pensiamo ad una impresa edile che acquisti una grossa partita di cavi con una resistenza media alla rottura che è garantita maggiore di 7 000 psi (libbre per pollice quadrato). La ditta potrebbe volere verificare se è vero che questi cavi hanno quella resistenza, e a questo scopo prendere un campione di 10 esemplari e testarli. I dati così ottenuti possono essere utilizzati per stabilire se accettare o meno l'ipotesi del produttore che la resistenza media dei cavi sia almeno pari a 7 000 psi.

Una *ipotesi statistica* è normalmente una affermazione su uno o più parametri della distribuzione di popolazione. Si parla di ipotesi perché a priori non sappiamo se sia vera o meno: il problema primario è quello di sviluppare una procedura per determinare se i valori di un campione aleatorio e l'ipotesi fatta siano compatibili oppure no. Un esempio potrebbe essere una popolazione gaussiana con varianza

unitaria e media θ incognita; l'affermazione "θ è minore di 1" è una ipotesi statistica che possiamo provare a verificare osservando un campione di questa popolazione. Se esso sarà giudicato compatibile con l'ipotesi considerata, diremo che quest'ultima è "accettata", altrimenti diremo che è "rifiutata".

Si noti che quando accettiamo una ipotesi, non stiamo affermando che sia necessariamente vera, ma solo che i dati raccolti sono accettabilmente in accordo con essa: che non la escludono. Continuando l'esempio della popolazione $\mathcal{N}(\theta, 1)$, se un campione di 10 dati presenta una media campionaria di 1.25, anche se tale risultato non è certo un indizio a favore dell'ipotesi "$\theta < 1$", non è nemmeno incompatibile con questa ipotesi, che quindi dovrebbe essere accettata. D'altra parte, se la media di un campione di 10 dati fosse stata pari a 3, anche se un valore così elevato è possibile anche con $\theta < 1$, diventa talmente improbabile da sembrare incompatibile con l'ipotesi fatta, che verrebbe senz'altro rifiutata.

8.2 Livelli di significatività

Consideriamo una popolazione avente distribuzione F_θ che dipende da un parametro incognito θ, e supponiamo di volere verificare una qualche ipotesi su θ, che chiameremo *ipotesi nulla*, e denoteremo con H_0. Se F_θ è ad esempio una distribuzione normale con media θ e varianza 1, due possibili ipotesi nulle su θ sono

1. $H_0 : \theta = 1$
2. $H_0 : \theta \leq 1$

La prima di queste ipotesi afferma che la popolazione ha distribuzione $\mathcal{N}(1, 1)$, mentre la seconda sostiene che essa è normale con varianza 1 e media non superiore a 1. Si noti che l'ipotesi nulla 1, quando è vera, caratterizza completamente la distribuzione della popolazione, mentre questo non è vero per l'ipotesi nulla 2. Nel primo caso si parla allora di ipotesi *semplice*, mentre nel secondo caso si parla di ipotesi *composta*.

Supponiamo di disporre di un campione aleatorio $X_1, X_2, \ldots, X_n$, proveniente da questa popolazione, e di volerlo utilizzare per eseguire una verifica o *test* di una certa ipotesi nulla H_0. Siccome dobbiamo decidere se accettare o meno H_0 basandoci esclusivamente sugli n valori dei dati, il test sarà definito da una regione C nello spazio a n dimensioni, con l'intesa che se il vettore $(X_1, X_2, \ldots, X_n)$ cade all'interno di C l'ipotesi viene rifiutata, mentre viene accettata in caso contrario. Una regione C con queste caratteristiche viene detta *regione critica* del test. Schematizzando quanto detto, il test statistico determinato dalla regione critica C è quello che

$$\text{accetta } H_0 \text{ se } (X_1, X_2, \ldots, X_n) \notin C$$

e

$$\text{rifiuta } H_0 \text{ se } (X_1, X_2, \ldots, X_n) \in C$$

Per anticipare un esempio concreto, una verifica molto comune dell'ipotesi che una popolazione gaussiana di varianza 1 abbia media 1, si ottiene con la regione critica seguente,

$$C = \left\{ (X_1, X_2, \ldots, X_n) : \left| 1 - \frac{1}{n} \sum_{i=1}^{n} X_i \right| > \frac{1.96}{\sqrt{n}} \right\} \qquad (8.2.1)$$

Bisogna quindi rifiutare l'ipotesi nulla "$\theta = 1$", quando la media campionaria dista da 1 più di 1.96 diviso per la radice quadrata dell'ampiezza del campione.

È importante notare che in qualunque test per verificare una ipotesi nulla, il risultato può essere sbagliato in due modi differenti. Si ha infatti un *errore di prima specie* quando i dati ci portano a rifiutare una ipotesi H_0 che in realtà è corretta, e un *errore di seconda specie* quando finiamo con l'accettare H_0 ed essa è falsa. Non vi è simmetria tra i due tipi di errori. Ricordiamo infatti che l'obiettivo di una verifica di H_0 non è quello di dire se questa ipotesi sia vera o falsa, ma piuttosto di dire se l'ipotesi fatta sia anche solo compatibile con i dati raccolti. In effetti vi è un ampio livello di tolleranza nell'accettare H_0, mentre per rifiutarla occorre che i dati campionari siano molto improbabili quando H_0 è soddisfatta.

Questo bilanciamento si ottiene specificando un valore α, detto *livello di significatività*, e imponendo che il test abbia la proprietà che quando l'ipotesi H_0 è vera, la probabilità che venga rifiutata non possa superare α. Il livello di significatività del test viene normalmente fissato in anticipo, con valori tipici dell'ordine di 0.1, 0.05 o 0.005. Detto in altri termini, un test con livello di significatività α deve avere una probabilità di errore di prima specie minore o uguale ad α.

Per chiarire un po' come viene costruita la regione critica, immaginiamo di volere verificare l'ipotesi nulla

$$H_0 : \theta \in w$$

dove con w stiamo indicando un insieme di valori possibili per il parametro. Un approccio naturale per formulare una verifica di H_0, ad un livello di significatività α prescritto, consiste nell'individuare uno stimatore puntuale di θ, che denotiamo con $d(\boldsymbol{X})$, e quindi rifiutare l'ipotesi quando $d(\boldsymbol{X})$ è "lontano" dalla regione w. Per capire quanto "lontano" deve essere per giustificare un rifiuto di H_0 ad un livello di significatività pari ad α, occorre conoscere la distribuzione dello stimatore $d(\boldsymbol{X})$ nel caso in cui H_0 sia vera. Questo ci permetterebbe infatti di usare il fatto che l'errore di prima specie deve avere probabilità inferiore ad α, per capire quando lo stimatore deve considerarsi abbastanza "lontano" da w, e quindi per determinare la regione critica del test. Ad esempio la verifica dell'ipotesi che la media di una popolazione $\mathcal{N}(\theta, 1)$ sia pari a 1 (l'Equazione (8.2.1) ne specifica la regione critica), impone di rifiutare l'ipotesi quando lo stimatore puntuale di θ (ovvero, la media campionaria),

dista da 1 (il valore di θ a cui corrisponde l'ipotesi nulla) più di $1.96/\sqrt{n}$. Come vedremo nella prossima sezione, quest'ultimo valore è stato scelto in modo da dare al test un livello di significatività del 5%.

8.3 La verifica di ipotesi sulla media di una popolazione normale

8.3.1 Il caso in cui la varianza è nota

Supponiamo che $X_1, X_2, \ldots, X_n$ sia un campione aleatorio proveniente da una popolazione normale di parametri μ e σ^2, con la varianza nota e media incognita. Fissata una costante μ_0, vogliamo verificare l'ipotesi nulla

$$H_0 : \mu = \mu_0$$

contro l'ipotesi alternativa

$$H_1 : \mu \neq \mu_0$$

Siccome $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ è lo stimatore puntuale naturale per μ, sembra ragionevole accettare H_0 quando $\overline{X}$ non è troppo lontano da μ_0. Perciò la regione critica del test sarà del tipo

$$C := \{(X_1, X_2, \ldots, X_n) : |\overline{X} - \mu_0| > c\}$$

per una scelta opportuna della costante c.

Se vogliamo che il test abbia livello di significatività α, dobbiamo individuare quel valore di c nell'equazione precedente che rende pari ad α la probabilità di errore di prima specie. Ciò significa che c deve soddisfare la relazione seguente,

$$\begin{aligned} \alpha &= P(\text{errore di I specie}) \\ &= P_{\mu_0}(|\overline{X} - \mu_0| > c) \end{aligned} \tag{8.3.1}$$

dove scriviamo P_{μ_0} per intendere che la probabilità precedente viene calcolata con l'assunzione che $\mu = \mu_0$. Infatti la definizione di errore di prima specie prevede che esso si verifichi quando i dati ci portano a rifiutare H_0 (quindi quando $(X_1, X_2, \ldots, X_n) \in C$) mentre in realtà essa è vera (quindi nel caso in cui $\mu = \mu_0$).

Quando però $\mu = \mu_0$, sappiamo che $\overline{X}$ ha distribuzione normale con media μ_0 e varianza σ^2/n, e quindi se Z denota una variabile aleatoria $\mathcal{N}(0,1)$, allora

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \overset{\mu_0}{\sim} Z \tag{8.3.2}$$

dove la relazione $\sim$ è condizionata all'ipotesi $H_0 : \mu = \mu_0$. Possiamo allora riscrivere l'Equazione (8.3.1) nella forma seguente,

$$\begin{aligned} \alpha &= P_{\mu_0}\left(\left|\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\right| > \frac{c\sqrt{n}}{\sigma}\right) \\ &= P\left(|Z| > \frac{c\sqrt{n}}{\sigma}\right) \\ &= 2P\left(Z > \frac{c\sqrt{n}}{\sigma}\right) \end{aligned}$$

e quindi $P(Z > c\sqrt{n}/\sigma) = \alpha/2$. Siccome però per definizione di $z_{\frac{\alpha}{2}}$ vale,

$$P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

si deduce che

$$\frac{c\sqrt{n}}{\sigma} = z_{\frac{\alpha}{2}}$$

e quindi che

$$c = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \tag{8.3.3}$$

Il test con livello di significatività α dovrà allora rifiutare H_0 se $|\overline{X} - \mu_0| > z_{\frac{\alpha}{2}} \cdot \sigma/\sqrt{n}$, ovvero

$$\begin{aligned} &\text{si rifiuta } H_0 \text{ se } \left|\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\right| > z_{\frac{\alpha}{2}} \\ &\text{si accetta } H_0 \text{ se } \left|\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\right| \leq z_{\frac{\alpha}{2}} \end{aligned} \tag{8.3.4}$$

La regione di accettazione per la statistica del test[1] è un intervallo simmetrico rispetto allo zero, come è illustrato in Figura 8.1, dove si è riportata in sovrapposizione la densità della distribuzione normale standard (che è la densità della statistica del test quando H_0 è vera).

Esempio 8.3.1. Un segnale di valore μ trasmesso da una sorgente A, viene raccolto dal ricevente B con un rumore normale di media nulla e varianza 4; il segnale ricevuto da B ha quindi distribuzione $\mathcal{N}(\mu, 4)$. Per ridurre il rumore, viene inviato per 5 volte lo stesso segnale: la media campionaria dei segnali ricevuti è $\overline{X} = 9.5$. Si sa infine che B aveva motivo di supporre che il valore inviato dovesse essere 8. Si verifichi questa ipotesi.

[1] Ogni verifica di ipotesi si basa fondamentalmente su una statistica particolare. In questo caso si intende la variabile aleatoria $\sqrt{n}(\overline{X} - \mu_0)/\sigma$.

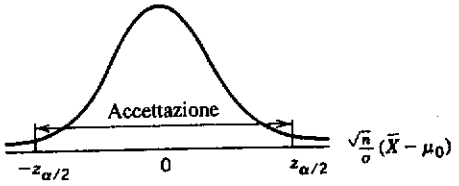


Figura 8.1 Densità della statistica del test e regione di accettazione.

Verifichiamo l'ipotesi ad un livello di significatività del 5%. Per prima cosa calcoliamo la statistica del test,

$$\frac{\sqrt{n}}{\sigma}|\overline{X}-\mu_0| = \frac{\sqrt{5}}{2}\cdot 1.5 \approx 1.68$$

Siccome questo valore è minore di $z_{0.025} \approx 1.96$, l'ipotesi va accettata. I dati non sono incompatibili con l'ipotesi fatta, nel senso che, se fosse $\mu = 8$, la media campionaria verrebbe osservata ad una distanza altrettanto grande (più distante di 1.5 da 8) più del 5% delle volte. Si noti comunque che, con un livello di significatività meno stringente, ad esempio $\alpha = 0.1$, l'ipotesi nulla sarebbe stata rifiutata. Questo perché $z_{0.05} = 1.645$ è inferiore a 1.68. Quindi se avessimo chiesto una verifica che avesse il 10% di probabilità di rifiutare H_0 quando essa è vera, avremmo effettivamente ottenuto un rifiuto.

Il livello di significatività "corretto" da usare nelle varie situazioni dipende di volta in volta dalle circostanze ed è influenzato da diversi fattori. Ad esempio se la decisione di rifiutare l'ipotesi H_0 portasse ad un costo elevato, che risulterebbe quindi perduto se H_0 fosse in realtà valida, potremmo forse decidere di essere abbastanza cauti, scegliendo un livello di significatività di 0.05 o 0.01. O ancora, se ci sentissimo a priori molto convinti della correttezza di H_0, potremmo richiedere una forte evidenza sperimentale contraria, per rifutare questa ipotesi, scegliendo di nuovo un valore di α molto basso. □

La regola fornita dall'Equazione (8.3.4) può essere riformulata come segue. Dopo avere calcolato il valore assunto dalla statistica del test, $\sqrt{n}|\overline{X}-\mu_0|/\sigma$, che denotiamo con v, valutiamo la probabilità (condizionata alla validità di H_0) che la statistica stessa assumesse un valore come v o più estremo ancora. Se tale probabilità è minore del livello di α, rifiutiamo l'ipotesi H_0, altrimenti la accettiamo. In altri termini dobbiamo calcolare prima il valore della statistica del test, poi la probabilità che una normale standard, in valore assoluto, superi tale quantità. Questa probabilità, detta il p-dei-dati del test, fornisce il livello di significatività critico, scendendo al di sotto del quale la decisione cambia da rifiuto ad accettazione.

In pratica spesso non si fissa in anticipo il livello di significatività, ma si osservano i dati e si ricava il p-dei-dati corrispondente. Se esso risulta molto maggiore di quanto siamo disposti ad accettare come probabilità di un errore di prima specie, conviene accettare l'ipotesi nulla; se invece esso è molto piccolo, possiamo rifiutarla.

Esempio 8.3.2. Con riferimento all'Esempio 8.3.1, supponiamo che la media campionaria dei 5 segnali ricevuti fosse 8.5. In quel caso

$$\frac{\sqrt{n}}{\sigma}|\overline{X}-\mu_0| = \frac{\sqrt{5}}{2}\cdot 0.5 \approx 0.559$$

Siccome

$$\begin{aligned} P(|Z| > 0.559) &= 2P(Z > 0.559) \\ &\approx 2 \times 0.288 = 0.576 \end{aligned}$$

si ha che il p-dei-dati è 0.576 e quindi l'ipotesi nulla che il segnale inviato fosse 8, viene accettata per ogni $\alpha < 0.576$. Poiché sarebbe assurdo eseguire un test con un livello di significatività elevato come 0.576, è senz'altro opportuno accettare H_0.

D'altra parte, se avessimo ottenuto che $\overline{X} = 11.5$, il corrispondente valore del p-dei-dati sarebbe stato

$$\begin{aligned} P\left(|Z| > \frac{\sqrt{5}}{4}\cdot 3.5\right) &\approx 2P(Z > 0.3913) \\ &\approx 0.00005 \end{aligned}$$

e con un valore così piccolo, l'ipotesi che il messaggio fosse stato 8, va rifiutata. □

Non abbiamo ancora discusso la probabilità degli errori di seconda specie – cioè la probabilità di accettare H_0 quando in realtà essa non è valida. Tale probabilità dipende da μ, e in particolare vale:

$$\begin{aligned} \beta(\mu) &:= P_\mu(\text{accettare } H_0) \\ &= P_\mu\left(\left|\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}\right| \le z_{\frac{\alpha}{2}}\right) \\ &= P_\mu\left(-z_{\frac{\alpha}{2}} \le \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \le z_{\frac{\alpha}{2}}\right) \end{aligned}$$

La funzione $\beta(\mu)$ è detta curva OC (che sta per *curva operativa caratteristica*, o più propriamente per il suo equivalente inglese, *operating characteristic curve*), e rappresenta la probabilità di accettare H_0 quando la media reale è μ.

Per calcolare questa probabilità, usiamo il fatto che $\overline{X} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$ e quindi

$$Z := \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

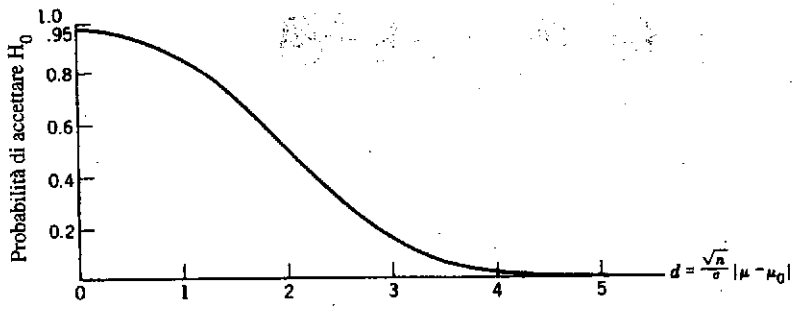


Figura 8.2 Curva OC di un test bilaterale per la media di una popolazione normale, con $\alpha = 0.05$.

Da cui

$$\begin{aligned}\beta(\mu) &= P_\mu\left(-z_{\frac{\alpha}{2}} \le \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \le z_{\frac{\alpha}{2}}\right)\\ &= P_\mu\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}} - z_{\frac{\alpha}{2}} \le \frac{\mu_0-\mu}{\sigma/\sqrt{n}} + \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \le \frac{\mu_0-\mu}{\sigma/\sqrt{n}} + z_{\frac{\alpha}{2}}\right)\\ &= P_\mu\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}} - z_{\frac{\alpha}{2}} \le Z \le \frac{\mu_0-\mu}{\sigma/\sqrt{n}} + z_{\frac{\alpha}{2}}\right)\\ &= \Phi\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}} + z_{\frac{\alpha}{2}}\right) - \Phi\left(\frac{\mu_0-\mu}{\sigma/\sqrt{n}} - z_{\frac{\alpha}{2}}\right) \qquad (8.3.5)\end{aligned}$$

dove Φ indica la funzione di ripartizione della distribuzione normale standard.

Per un livello di significatività α fissato, la curva OC è simmetrica rispetto a μ_0, e in effetti dipende da μ solo tramite $\sqrt{n}|\mu-\mu_0|/\sigma$. In Figura 8.2 è rappresentata la curva OC per $\alpha = 0.05$, con l'ascissa trasformata da μ a $d := \sqrt{n}|\mu-\mu_0|/\sigma$.

Esempio 8.3.3. Con riferimento all'Esempio 8.3.1, quanto vale la probabilità di accettare $\mu = 8$, quando in realtà $\mu = 10$? Calcoliamo

$$\frac{\sqrt{n}}{\sigma}(\mu_0-\mu) = \frac{\sqrt{5}}{2}(-2) = -\sqrt{5}$$

Poiché $z_{0.025} \approx 1.96$, sostituendo nell'Equazione (8.3.5) ricaviamo la probabilità cercata,

$$\begin{aligned}\beta(10) &\approx \Phi(-\sqrt{5}+1.96) - \Phi(-\sqrt{5}-1.96)\\ &\approx \Phi(-0.276) - \Phi(-4.196)\\ &= 1 - \Phi(0.276) - 1 + \Phi(4.196)\\ &\approx -0.609 + 1 = 0.391 \quad \square\end{aligned}$$

Osservazione 8.3.1. La funzione $1-\beta(\mu)$ viene detta *funzione di potenza* del test. Per un valore di μ fissato, la potenza del test è la probabilità di rifiutare (correttamente) H_0 quando μ è il valore vero.

La curva OC permette di dimensionare il campione in modo che l'errore di seconda specie soddisfi delle condizioni specifiche. Supponiamo ad esempio, di cercare il valore di n con il quale la probabilità di accettare $H_0 : \mu = \mu_0$ quando il valore vero è μ_1, sia approssivamente pari a un valore β fissato. Vogliamo insomma n tale che

$$\beta(\mu_1) \approx \beta$$

Per l'Equazione (8.3.5), questo è equivalente a chiedere che

$$\Phi\left(\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} + z_{\frac{\alpha}{2}}\right) - \Phi\left(\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} - z_{\frac{\alpha}{2}}\right) \approx \beta \qquad (8.3.6)$$

Anche se l'equazione precedente non può essere risolta analiticamente in funzione di n, si può arrivare ad una soluzione usando i tabulati di Φ. Inoltre, un valore molto approssimato per n si può ricavare dall'Equazione (8.3.6), nel modo seguente. Supponiamo che $\mu_1 > \mu_0$ (il viceversa è analogo e viene lasciato come esercizio). Ciò significa che la seconda $\Phi(\cdot)$ che compare nella (8.3.6) vale certamente meno di $\alpha/2$, e quindi in molti casi può essere trascurata, infatti:

$$\mu_1 > \mu_0 \Leftrightarrow \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} - z_{\frac{\alpha}{2}} < -z_{\frac{\alpha}{2}}$$

da cui, visto che Φ è monotona crescente,

$$\begin{aligned}\Phi\left(\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} - z_{\frac{\alpha}{2}}\right) &\le \Phi(-z_{\frac{\alpha}{2}})\\ &= P(Z \le -z_{\frac{\alpha}{2}})\\ &= P(Z \ge z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}\end{aligned}$$

Per cui si può considerare trascurabile il termine

$$\Phi\left(\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} - z_{\frac{\alpha}{2}}\right) \approx 0$$

ottenendo quindi dall'Equazione (8.3.6) che

$$\beta \approx \Phi\left(\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} + z_{\frac{\alpha}{2}}\right)$$

Quest'ultima equazione è finalmente risolvibile rispetto a n, visto che

$$\beta = P(Z > z_\beta) = P(Z < -z_\beta) = \Phi(-z_\beta)$$

e quindi possiamo uguagliare

$$\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{\frac{\alpha}{2}} \approx -z_\beta$$

ricavando

$$n \approx \left[\frac{(z_{\frac{\alpha}{2}} + z_\beta)\sigma}{\mu_1 - \mu_0}\right]^2 \tag{8.3.7}$$

È bene notare che anche nel caso in cui $\mu_1 < \mu_0$, si perviene esattamente alla stessa formula.

Esempio 8.3.4. Con riferimento all'Esempio 8.3.1, quante volte è necessario inviare il segnale affinché la verifica dell'ipotesi $H_0 : \mu = 8$ ad un livello di significatività di 0.05, abbia almeno il 75% di probabilità di rifiutare l'ipotesi nulla quando $\mu = 9.2$?

Siccome $z_{0.025} \approx 1.96$ e $z_{0.25} \approx 0.67$, per l'approssimazione descritta qui sopra,

$$n \approx \left(\frac{1.96 + 0.67}{1.2}\right)^2 4 \approx 19.21$$

Per cui è necessario un campione di 20 segnali. Dall'Equazione (8.3.5) vediamo che con $n = 20$,

$$\begin{aligned}\beta(9.2) &\approx \Phi\left(-\frac{1.2\sqrt{20}}{2} + 1.96\right) - \Phi\left(-\frac{1.2\sqrt{20}}{2} - 1.96\right)\\ &\approx \Phi(-0.723) - \Phi(-4.643)\\ &\approx 1 - \Phi(0.723) \approx 0.235\end{aligned}$$

Perciò se il segnale viene trasmesso 20 volte vi è il 76.5% di probabilità che l'ipotesi nulla $\mu = 8$ sia rifiutata se la media reale è 9.2. □

8.3.1.1 I test unilaterali

Nel verificare l'ipotesi nulla $\mu = \mu_0$ abbiamo costruito un test che porta ad un rifiuto quando $\overline{X}$ è lontana da μ_0, ovvero, valori di $\overline{X}$ troppo bassi o troppo elevati rispetto a μ_0 sembrano smentire che μ (stimata da $\overline{X}$) sia proprio uguale a μ_0. Cosa accade invece quando μ può essere solo maggiore a μ_0, quando non sono uguali? Ovvero cosa occorre fare se l'ipotesi alternativa a $H_0 : \mu = \mu_0$, è $H_1 : \mu > \mu_0$? Chiaramente quando il contesto è questo, valori molto bassi di $\overline{X}$ non ci dovrebbero fare rifiutare l'ipotesi nulla (visto che è più probabile ottenere una $\overline{X}$ piccola quando è vera H_0 che non quando è vera H_1). Perciò, nel verificare l'ipotesi

$$H_0 : \mu = \mu_0 \qquad \text{contro} \qquad H_1 : \mu > \mu_0$$

dovremmo rifiutare l'ipotesi nulla quando $\overline{X}$, lo stimatore di μ, è molto più grande di μ_0, e quindi la regione critica dovrebbe essere del tipo seguente:

$$C := \{(X_1, X_2, \ldots, X_n) : \overline{X} - \mu_0 > c\}$$

per una scelta opportuna della costante c. In particolare, siccome la probabilità di rifiuto dovrebbe essere α quando H_0 è vera (cioè quando $\mu = \mu_0$), occorre che c soddisfi la relazione,

$$P_{\mu_0}(\overline{X} - \mu_0 > c) = \alpha \tag{8.3.8}$$

Di nuovo, poiché stiamo supponendo che $\mu = \mu_0$, $\overline{X}$ ha media μ_0, e quindi la statistica Z definita qui sotto ha distribuzione normale standard,

$$Z := \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

Perciò la (8.3.8) è equivalente a

$$P\left(Z > \frac{c\sqrt{n}}{\sigma}\right) = \alpha$$

che si risolve in funzione di c ricordando che $P(Z > z_\alpha) = \alpha$, ottenendo quindi che

$$c = z_\alpha \frac{\sigma}{\sqrt{n}} \tag{8.3.9}$$

Il test con livello di significatività α dovrà allora rifiutare H_0 se $\overline{X} - \mu_0 > z_\alpha \cdot \sigma/\sqrt{n}$, ovvero

$$\begin{aligned}&\text{si rifiuta } H_0 \text{ se } \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha\\ &\text{si accetta } H_0 \text{ se } \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \leq z_\alpha\end{aligned} \tag{8.3.10}$$

Quella trovata è detta regione critica *unilaterale*, o *a una coda* (a differenza delle regioni critiche trovate nella sezione precedente che erano *bilaterali* o *a due code*). In accordo con quanto detto, anche il problema di verificare le ipotesi alternative

$$\begin{aligned}H_0 &: \mu = \mu_0\\ H_1 &: \mu > \mu_0\end{aligned}$$

si dice problema di test unilaterale.

Per ottenere il p-dei-dati di questo tipo di test, si calcola innanzitutto il valore della statistica del test,

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

in funzione dei dati raccolti; il p-dei-dati è quindi uguale alla probabilità che una normale standard superi questo valore.

Esempio 8.3.5. Supponiamo, nell'Esempio 8.3.1, di sapere in anticipo che il segnale inviato non è inferiore a 8. Cosa possiamo concludere in questo caso?

Per vedere se i dati siano compatibili con l'ipotesi che la media sia 8, verifichiamo

$$H_0 : \mu = 8$$

contro l'alternativa a una coda

$$H_1 : \mu > 8$$

Il valore della statistica del test è di

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \sqrt{5}(9.5 - 8)/2 \approx 1.68$$

quindi il p-dei-dati è la probabilità che una normale standard superi 1.68, ovvero

$$p\text{-dei-dati} = 1 - \Phi(1.68) \approx 0.0465$$

Siccome la verifica impone un rifiuto a tutti i livelli di significatività maggiori o uguali a 0.0465, l'ipotesi nulla sarebbe rifiutata se si ponesse ad esempio $\alpha = 0.05$. □

La curva OC del test unilaterale (8.3.10) si può ricavare come segue. Visto che per $i = 1, 2, \ldots, n$, si ha che $X_i \sim \mathcal{N}(\mu, \sigma^2)$, e quindi che $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, se poniamo $Z := \sqrt{n}(\overline{X} - \mu)/\sigma$, questa statistica è normale standard, per cui

$$\begin{aligned}
\beta(\mu) &:= P_\mu(\text{accettare } H_0) \\
&= P_\mu\left(\overline{X} \le \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\right) \\
&= P_\mu\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right) \\
&= P_\mu\left(Z \le \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right) \\
&= \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_\alpha\right) \qquad (8.3.11)
\end{aligned}$$

Siccome Φ, in quanto funzione di ripartizione, è crescente, è chiaro che $\beta(\mu)$ è una funzione decrescente. Questo risultato appare incoraggiante, visto che è ragionevole che, al crescere di μ, sia sempre meno facile concludere che $\mu < \mu_0$. Si noti anche che, siccome $\Phi(z_\alpha) = 1 - \alpha$, si ha che

$$\beta(\mu_0) = 1 - \alpha$$

La regola fornita dalla (8.3.10), che abbiamo utilizzato per verificare l'ipotesi $H_0 : \mu = \mu_0$ contro l'ipotesi $H_1 : \mu > \mu_0$, vale anche per verificare, ad un livello di significatività α, l'ipotesi unilaterale

$$H_0 : \mu \le \mu_0$$

contro l'alternativa

$$H_1 : \mu > \mu_0$$

Per accertarci che il livello di significatività sia rimasto α, dobbiamo dimostrare che la probabilità di un errore di prima specie non superi mai questo valore. Al variare di μ, la probabilità di rifiuto è data da $1 - \beta(\mu)$. Siccome si commette un errore di prima specie se H_0 è vera e i dati ci impongono di rifiutarla, dobbiamo verificare che, per ogni μ compatibile con H_0, quindi per ogni $\mu \le \mu_0$,

$$1 - \beta(\mu) \le \alpha, \qquad \text{per ogni } \mu \le \mu_0$$

ovvero che

$$\beta(\mu) \ge 1 - \alpha, \qquad \text{per ogni } \mu \le \mu_0$$

Ma avendo già dimostrato che $\beta(\mu)$ è una funzione decrescente, che vale proprio $1 - \alpha$ quando $\mu = \mu_0$, è chiaro che per valori di μ più piccoli, il valore di $\beta(\mu)$ sarà superiore a $1 - \alpha$ come richiesto.

Osservazione 8.3.2. È anche possibile verificare l'ipotesi

$$H_0 : \mu = \mu_0$$

contro l'ipotesi alternativa

$$H_1 : \mu < \mu_0$$

ad un livello di significatività α, decidendo che

$$\begin{aligned}
\text{si rifiuta } H_0 \text{ se } & \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha \\
\text{si accetta } H_0 \text{ se } & \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \ge -z_\alpha
\end{aligned} \qquad (8.3.12)$$

Anche questo test può essere in alternativa effettuato calcolando la statistica $\sqrt{n}(\overline{X} - \mu_0)/\sigma$ in funzione dei dati, poi trovando il p-dei-dati che è la probabilità che una normale standard sia inferiore a quel valore, e concludendo che a qualunque livello di significatività α, maggiore o uguale al p-dei-dati, il test impone di rifiutare l'ipotesi nulla.

Esempio 8.3.6. Tutti i tipi di sigarette attualmente presenti sul mercato hanno un contenuto medio di nicotina non inferiore a 1.6 mg. Una marca di tabacchi afferma però di avere individuato un particolare trattamento delle foglie di tabacco che permette di abbassare il livello medio di nicotina al di sotto di 1.6 mg. Per verificare questa affermazione, si analizza un campione di 20 sigarette di questa marca, trovando una media campionaria del contenuto di nicotina di 1.54 mg. Supponendo che la deviazione standard della popolazione sia[2] di 0.8 mg e fissando il livello di significatività al 5%, cosa decide il test?

Nel risolvere questo esercizio, il primo passo consiste nell'individuare quale sia l'ipotesi nulla appropriata. Si tenga presente infatti che non vi è simmetria (nemmeno nel caso unilaterale!) tra ipotesi nulla e alternativa, nel senso che passando da

$$H_0 : \mu \leq \mu_0 \qquad \text{contro} \qquad H_1 : \mu > \mu_0$$

a

$$H_0 : \mu \geq \mu_0 \qquad \text{contro} \qquad H_1 : \mu < \mu_0$$

non è affatto detto che se uno dei due test accetta H_0, l'altro la rifiuti. Come mai? Si ricordi che α per definizione non è mai inferiore alla probabilità di rifiutare H_0 quando essa è vera; per questo se il test decide di rifiutare H_0, si è certi che la probabilità di errore non supera α (che è un valore piccolo e per di più fissato a priori), quindi il rifiutare l'ipotesi nulla è una affermazione "forte", nel senso che abbiamo un eccellente controllo sulla probabilità di sbagliare. Non è invece possibile accettare l'ipotesi H_0 con lo stesso livello di controllo: la probabilità di errore è incerta, essendo pari a $\beta(\mu)$ che dipende da μ, e può essere anche un valore molto elevato (fino a $1 - \alpha$, come abbiamo visto).

Se si accetta l'ipotesi nulla, significa che non vi è evidenza sperimentale sufficiente ad escluderla: non significa che i dati la avvalorino con decisione.

Ciò premesso, siccome vogliamo avvalorare l'affermazione del produttore solo in presenza di una chiara evidenza sperimentale in questa direzione, dobbiamo prendere tale affermazione come ipotesi alternativa, e perciò dobbiamo verificare

$$H_0 : \mu \geq 1.6 \qquad \text{contro} \qquad H_1 : \mu < 1.6$$

[2] Quanto proposto solleva la questione di come si possa affermare di conoscere la deviazione standard di un *nuovo* tipo di sigarette. Una possibile giustificazione si avrebbe se la variabilità del contenuto di nicotina non fosse alterata dal trattamento usato sulle foglie, ma dipendesse solo dal contenuto di tabacco di ogni sigaretta. Se così fosse, si potrebbe affermare che la deviazione standard deve essere la stessa degli altri tipi di sigarette, e potrebbe quindi essere nota dall'esperienza passata. In ogni caso, anche i casi in cui la varianza di popolazione non sia nota si possono affrontare con successo, come è descritto nella sezione successiva.

Il valore della statistica del test è di

$$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{1.54 - 1.6}{0.8/\sqrt{20}} \approx -0.335$$

Così che il p-dei-dati è dato da

$$p\text{-dei-dati} \approx P(Z < -0.335) = \Phi(-0.335) \approx 0.369$$

dove Z ha distribuzione normale standard. Siccome il risultato ottenuto è maggiore di 0.05 (e in effetti è maggiore di qualunque livello di significatività sensato), non si può rifiutare l'ipotesi nulla. In altri termini, il dato in nostro possesso, anche se avvalora la tesi del produttore, non è abbastanza forte da farci escludere che il contenuto medio di nicotina di quel tipo di sigarette sia maggiore o uguale a 1.6 mg. □

Osservazione 8.3.3. Vi è una evidente analogia tra la stima di parametri con gli intervalli di confidenza e la verifica delle ipotesi. Ad esempio abbiamo dimostrato nella Sezione 7.3 (con l'Equazione (7.3.8) di pagina 242), che un intervallo di confidenza bilaterale ad un livello di $1 - \alpha$ per la media di una distribuzione normale di varianza nota σ^2, è dato da

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

dove $\bar{x}$ era il valore della media campionaria. In maniera più rigorosa, ciò significa che

$$P\left\{\mu \in \left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\right\} = 1 - \alpha$$

Nel compiere una verifica sulle ipotesi bilaterali $H_0 : \mu = \mu_0$ contro $H_1 : \mu \neq \mu_0$ ad un livello di significatività di α, quello che facciamo è di accettare l'ipotesi nulla quando

$$\mu_0 \in \left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

e se $\mu = \mu_0$, questo evento è lo stesso di prima, e infatti la sua probabilità sotto P_{μ_0} è ancora di $1 - \alpha$.

Similmente, siccome un intervallo di confidenza unilaterale destro per μ è dato da

$$\left(\overline{X} - z_{\alpha}\frac{\sigma}{\sqrt{n}}, \quad \infty\right)$$

ne segue che un test con livello di significatività α per le ipotesi $H_0 : \mu \leq \mu_0$ contro $H_1 : \mu > \mu_0$ si ottiene accettando l'ipotesi nulla quando $\mu_0 \in (\overline{X} - z_\alpha \cdot \sigma/\sqrt{n}, \infty)$ in accordo con quanto dimostrato in questa sezione.

La Tabella 8.1 riassume i test di questa sezione.

Sulla robustezza di un test

Un test che si comporta bene anche quando alcune delle assunzioni su cui si basa non sono valide si dice *robusto*. Per esempio i test introdotti nelle Sezioni 8.3.1 e 8.3.1.1 sono stati ottenuti assumendo che la distribuzione della popolazione fosse normale, con varianza nota σ^2; tuttavia, anche se essa è una qualunque altra distribuzione con varianza σ^2, la media campionaria $\overline{X}$ è comunque *approssimativamente* normale (purché il campione sia numeroso), per il teorema del limite centrale, e quindi i risultati trovati saranno approssimativamente corretti (mostrando così la robustezza di quei test).

Tabella 8.1 $X_1, X_2, \ldots, X_n$ è un campione estratto da una popolazione $\mathcal{N}(\mu, \sigma^2)$.

σ^2 nota $\qquad \overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$

H_0	H_1	Statistica del test, X_{ts}	Si rifiuta H_0 con livello di significatività α se...	p-dei-dati se $X_{ts} = t$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\ldots \lvert X_{ts}\rvert > z_{\frac{\alpha}{2}}$	$2P(Z > \lvert t\rvert)$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\ldots X_{ts} > z_\alpha$	$P(Z > t)$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\ldots X_{ts} < -z_\alpha$	$P(Z < t)$

Nota: Z ha distribuzione normale standard.

8.3.2 Quando la varianza non è nota: il test t

Fino ad ora abbiamo supposto che l'unico parametro incognito della distribuzione di popolazione fosse la media. Più comunemente però, né la media μ, né la varianza σ^2 sono note. Supponiamo di essere in tale situazione e consideriamo di nuovo come si possa verificare l'ipotesi nulla che μ sia uguale ad un valore assegnato μ_0, contro l'ipotesi alternativa $\mu \neq \mu_0$,

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0 .$$

È bene notare che l'ipotesi nulla non è semplice (nel senso della definizione data a pagina 286, ovvero che supporre vera H_0 non specifica completamente la distribuzione),

perché non fornisce il valore di σ^2.

Come in precedenza, sembra ragionevole rifiutare l'ipotesi nulla quando $\overline{X}$ cade lontano da μ_0; tuttavia la distanza a cui deve essere da μ_0 per giustificare questo rifiuto, dipende dalla deviazione standard σ che in quella sede era nota; in particolare $|\overline{X} - \mu_0|$ doveva essere maggiore di $z_{\frac{\alpha}{2}} \cdot \sigma/\sqrt{n}$, o equivalentemente,

$$\left|\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\right| > z_{\frac{\alpha}{2}}$$

Qui σ non è più conosciuta. Possiamo allora pensare di sostituirla con il suo stimatore, la deviazione standard campionaria S

$$S := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{8.3.13}$$

rifiutando l'ipotesi nulla quando

$$\left|\frac{\overline{X} - \mu_0}{S/\sqrt{n}}\right|$$

è troppo grande.

Quanto grande è "troppo grande"? Affinché il test alla fine abbia livello di significatività pari ad α, dobbiamo conoscere la distribuzione della statistica del test quando H_0 è vera, e imporre che la probabilità di rifiutare l'ipotesi nulla sia (non più grande di) α. Sappiamo (per il Corollario 6.5.2 di pagina 221), che la variabile aleatoria

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \tag{8.3.14}$$

ha distribuzione t. Se si denota con T la statistica di questo test, ovvero

$$T := \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \tag{8.3.15}$$

allora quando H_0 è vera, visto che $\mu = \mu_0$, T ha distribuzione t con $n-1$ gradi di libertà. Imponiamo ora che la probabilità di errore di prima specie sia α, ovvero passando agli eventi complementari, che sia $1-\alpha$ la probabilità di accettare l'ipotesi nulla quando $\mu = \mu_0$:

$$P_{\mu_0}\left(-c \leq \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \leq c\right) = 1 - \alpha$$

Per ricavare c, si noti che, siccome la densità della distribuzione t è simmetrica rispetto allo zero,

$$\begin{aligned}\alpha &= 1 - P(-c < T < c)\\ &= P(T \leq -c) + P(T \geq c)\\ &= 2P(T \geq c)\end{aligned}$$

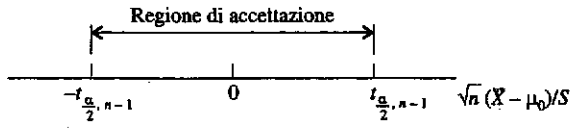


Figura 8.3 Il test t bilaterale.

Per cui $P(T > c) = \frac{\alpha}{2}$ e quindi deve valere $c = t_{\frac{\alpha}{2},n-1}$ per definizione di $t_{\alpha,k}$. Concludendo, diamo la regola per usare il test:

$$\begin{aligned} &\text{si rifiuta } H_0 \text{ se } \left|\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right| > t_{\frac{\alpha}{2},n-1} \\ &\text{si accetta } H_0 \text{ se } \left|\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right| \leq t_{\frac{\alpha}{2},n-1} \end{aligned} \tag{8.3.16}$$

Il test descritto qui sopra è detto *test t* bilaterale, ed è illustrato in Figura 8.3.

Se si denota con t il valore assunto da T – la statistica del test – calcolata in funzione dei dati del campione, il valore del p-dei-dati corrispondente è la probabilità che $|T|$ superi $|t|$, quando H_0 è vera. Si tratta quindi della probabilità che una t di Student con $n-1$ gradi di libertà abbia valore assoluto maggiore di $|t|$. Come nei casi precedenti, si deve rifiutare l'ipotesi nulla a tutti i livelli di significatività maggiori del p-dei-dati, mentre la si accetta a tutti i livelli inferiori.

Il Programma 8.3.2 del software abbinato a questo libro, calcola il valore della statistica del test t e del p-dei-dati corrispondente; può essere usato sia per i test t a due code, sia per quelli ad una coda. Questi ultimi saranno presentati brevemente dopo i due esempi seguenti.

Esempio 8.3.7. Tra quei pazienti di una clinica che hanno un livello di colesterolo da medio a elevato (al di sopra di 220 millilitri per decilitro di siero), vengono cercati dei volontari per sperimentare un nuovo farmaco che dovrebbe aiutare a ridurre il tasso di colesterolo. Si sceglie un gruppo di 50 volontari a cui viene somministrato il farmaco per un mese, alla fine si registra la variazione nel tasso di colesterolo e si trova una riduzione media di 14.8, con una deviazione standard campionaria di 6.4. Che conclusioni si possono trarre?

Verifichiamo se è possibile che tale diminuzione sia dovuta esclusivamente ad un caso fortuito – testiamo quindi l'ipotesi che le 50 variazioni siano normali con media nulla. Poiché il valore della statistica del test t, calcolata con $\mu_0 = 0$ è

$$T = \sqrt{n} \cdot \overline{X}/S = \sqrt{50} \cdot 14.8/6.4 \approx 16.35$$

è chiaro che dobbiamo rifiutare l'ipotesi nulla che avevamo fatto. □

Test clinici ed effetto placebo

Nell'Esempio 8.3.7 si è determinato che la diminuzione di colesterolo riscontrata non poteva essere casuale; tuttavia non si è comunque giustificati a concludere che il merito sia stato del farmaco. In effetti è ben noto che la somministrazione di una qualunque sostanza che il paziente pensa che possa avere un effetto benefico, tende a migliorarne le condizioni anche se non dovrebbe avere nessun effetto fisiologico (è il cosiddetto *effetto placebo*). Inoltre vi è la possibilità che agenti esterni come le condizioni meteorologiche possano influire sull'esperimento.
In effetti, un esperimento congegnato con intelligenza dovrebbe cercare di neutralizzare tutte le cause esterne, per ottenere una chiara indicazione sull'efficacia del farmaco. L'approccio a cui si ricorre comunenente consiste nel dividere i volontari in due gruppi, somministrando ad uno il farmaco vero, e all'altro un placebo (ovvero una sostanza con lo stesso aspetto e sapore del farmaco, che però non ha alcun effetto fisiologico), senza comunicare a nessuno come sono formati i gruppi, e possibilmente tenendo anche i medici che sono a contatto con i volontari all'oscuro, per evitare che con il loro atteggiamento provochino qualche effetto. Se i volontari sono suddivisi in modo casuale possiamo aspettarci che in media tutti gli altri fattori che influiscono sui due gruppi siano gli stessi, e quindi che ogni differenza riscontrata sia da attribuirsi al farmaco.

Esempio 8.3.8. Si vuole verificare l'ipotesi che il consumo medio di acqua per abitazione sia di 350 galloni al giorno. Si misurano i consumi medi di un campione di 20 abitazioni, trovando i seguenti dati

340 356 332 362 318 344 386 402 322 360
362 354 340 372 338 375 364 355 324 370

Cosa si conclude?

Dobbiamo verificare le due ipotesi seguenti

$$H_0 : \mu = 350 \qquad \text{contro} \qquad H_1 : \mu \neq 350$$

Ciò può essere ottenuto usando il Programma 8.3.2 o, in alternativa, calcolando prima la media e la deviazione standard campionarie dei dati, che sono

$$\overline{X} = 353.8 \qquad S \approx 21.85$$

trovando quindi il valore della statistica del test,

$$T \approx \frac{\sqrt{20} \cdot 3.8}{21.85} \approx 0.778$$

Siccome 0.778 è minore di $t_{0.05,19} \approx 1.729$, l'ipotesi nulla è accettata ad un livello di significatività del 5%. In realtà un calcolo del p-dei-dati fornisce il valore

$$p\text{-dei-dati} \approx P(|T_{19}| > 0.778) = 2P(T_{19} > 0.778) \approx 0.446$$

che è così grande che l'ipotesi nulla viene accettata a qualunque livello di significatività ragionevole, e quindi i dati non sono in disaccordo con l'ipotesi che il consumo medio per abitazione sia di 350 galloni al giorno. □

Si può costruire un test t a una coda per verificare l'ipotesi

$$H_0 : \mu = \mu_0 \qquad (\text{o } H_0 : \mu \leq \mu_0)$$

contro l'ipotesi alternativa

$$H_1 : \mu > \mu_0$$

ad un livello di significatività α, decidendo che

$$\begin{aligned} &\text{si rifiuta } H_0 \text{ se } \frac{\overline{X} - \mu_0}{S/\sqrt{n}} > t_{\alpha,n-1} \\ &\text{si accetta } H_0 \text{ se } \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \leq t_{\alpha,n-1} \end{aligned} \tag{8.3.17}$$

Se il valore di $\sqrt{n}(\overline{X} - \mu_0)/S$ realizzato dai dati è v, allora il p-dei-dati corrispondente è la probabilità che una t di Student con $n-1$ gradi di libertà sia maggiore o uguale a v.

Analogamente la verifica ad un livello di significatività α dell'ipotesi

$$H_0 : \mu = \mu_0 \qquad (\text{o } H_0 : \mu \geq \mu_0)$$

contro l'ipotesi alternativa

$$H_1 : \mu < \mu_0$$

si ottiene decidendo che

$$\begin{aligned} &\text{si rifiuta } H_0 \text{ se } \frac{\overline{X} - \mu_0}{S/\sqrt{n}} < -t_{\alpha,n-1} \\ &\text{si accetta } H_0 \text{ se } \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \geq -t_{\alpha,n-1} \end{aligned} \tag{8.3.18}$$

Il p-dei-dati in questo caso è la probabilità che una t di Student con $n-1$ gradi di libertà sia minore o uguale del valore osservato di $\sqrt{n}(\overline{X} - \mu_0)/S$.

Esempio 8.3.9. Il produttore di un nuovo tipo di pneumatico in fibra di vetro afferma che la vita media del suo prodotto è di almeno 40 000 miglia. Si prende un campione di 12 pneumatici per verificare questa affermazione, e i tempi di vita trovati (in unità di 1 000 miglia) sono i seguenti,

Gomma	1	2	3	4	5	6	7	8	9	10	11	12
Vita	36.1	40.2	33.8	38.5	42	35.8	37	41	36.8	37.2	33	36

Verifichiamo quanto affermato dal produttore ad un livello di significatività del 5%.

Per determinare se i dati raccolti siano compatibili con l'ipotesi che la vita media sia superiore alle 40 000 miglia, verifichiamo l'ipotesi

$$H_0 : \mu \geq 40 \qquad \text{contro} \qquad H_1 : \mu < 40$$

Un calcolo diretto fornisce

$$\overline{X} \approx 37.2833 \qquad S \approx 2.7319$$

e il valore della statistica del test risultante è

$$T \approx \frac{\sqrt{12}(37.2833 - 40}{2.7319} \approx -3.445$$

Siccome questo numero è inferiore a $-t_{0.05,11} \approx -1.796$, l'ipotesi nulla è rifiutata ad un livello di significatività del 5%. In effetti il p-dei-dati di questo test risulta essere

$$p\text{-dei-dati} \approx P(T_{11} < -3.445) = P(T_{11} > 3.445) \approx 0.0028$$

indicando che l'affermazione del produttore deve essere rifiutata ad ogni livello di significatività superiore a 0.28%. □

Il risultato precedente si sarebbe ottenuto anche utilizzando il Programma 8.3.2, come illustrato in Figura 8.4.

Esempio 8.3.10. Consideriamo un problema di teoria delle code. Un sistema con un unico server impiega un tempo con media μ e varianza σ^2 per servire un cliente. I clienti arrivano in tempi casuali, secondo un processo di Poisson di intensità λ. È possibile dimostrare che a lungo andare il tempo medio di attesa in coda dei clienti è dato da

$$\frac{\lambda(\mu^2 + \sigma^2)}{2(1 - \lambda\mu)} \tag{8.3.19}$$

dove si intende che $\lambda\mu < 1$, perché in caso contrario la coda si allunga all'infinito, e anche il tempo d'attesa diverge. Come si vede dalla formula, inoltre, il tempo medio d'attesa è piuttosto grande quando μ è solo di poco inferiore a $\frac{1}{\lambda}$, dove, visto che λ è la frequenza degli arrivi, $\frac{1}{\lambda}$ indica il tempo medio tra due arrivi consecutivi.

Supponiamo allora che il gestore del server voglia affittarne un secondo se si stabilisce che il tempo medio di servizio μ, è superiore a 8 minuti. I dati seguenti rappresentano i tempi di servizio per 28 clienti. Si può dire che essi vengano da una distribuzione con media superiore a 8?

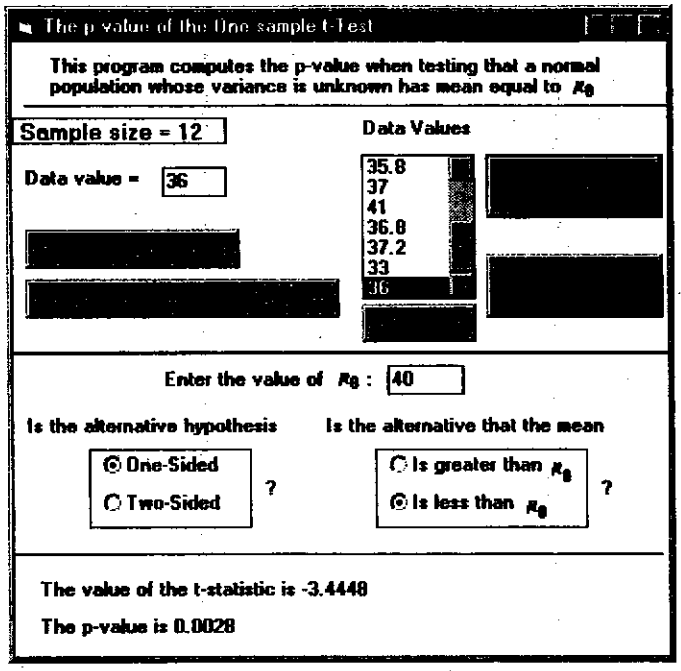


Figura 8.4 Verifica dell'ipotesi a una coda per l'Esempio 8.3.9.

8.6 9.4 5.0 4.4 3.7 11.4 10.0 7.6 14.4 12.2 11.0 14.4 9.3 10.5
10.3 7.7 8.3 6.4 9.2 5.7 7.9 9.4 9.0 13.3 11.6 10.0 9.5 6.6

Utilizziamo i dati precedenti per verificare l'ipotesi nulla che il tempo di servizio sia minore o uguale a 8 minuti. Un p-dei-dati molto piccolo sarebbe una forte indicazione a favore dell'ipotesi che il tempo medio di servizio sia superiore agli 8 minuti. Eseguendo il Programma 8.3.2 su questi dati si vede che il valore della statistica del test è pari a 2.257, con un p-dei-dati risultante di 0.016. Un valore così piccolo è una prova molto forte che il tempo medio di servizio supera gli 8 minuti. □

La Tabella 8.2 riassume le verifiche di questa sezione.

8.4 Verificare se due popolazioni normali hanno la stessa media

Una situazione che si presenta comunemente nella statistica applicata all'ingegneria è quando occorre decidere se due differenti approcci allo stesso problema hanno portato

Tabella 8.2 $X_1, X_2, \ldots, X_n$ è un campione con distribuzione $\mathcal{N}(\mu, \sigma^2)$ e σ^2 non è nota.

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(\overline{X} - X_i)^2$$

H_0	H_1	Statistica del test, X_{ts}	Si rifiuta H_0 con livello di significatività α se...	p-dei-dati se $X_{ts} = t$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\ldots \lvert X_{ts}\rvert > t_{\frac{\alpha}{2},n-1}$	$2P(T_{n-1} > \lvert t\rvert)$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\ldots X_{ts} > t_{\alpha,n-1}$	$P(T_{n-1} > t)$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\frac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\ldots X_{ts} < -t_{\alpha,n-1}$	$P(T_{n-1} < t)$

Nota: T_{n-1} ha distribuzione t con $n-1$ gradi di libertà. Inoltre $P(T_{n-1} > t_{\alpha,n-1}) = \alpha$.

al medesimo risultato, oppure no. Tale problematica si riconduce spesso alla verifica dell'ipotesi che due popolazioni normali abbiano la stessa media.

8.4.1 Il caso in cui le varianze sono note

Supponiamo che $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_m$ siano campioni indipendenti provenienti da due popolazioni normali di medie incognite μ_x e μ_y e varianze note σ_x^2 e σ_y^2. Consideriamo il problema di verificare l'ipotesi

$$H_0 : \mu_x = \mu_y$$

contro l'ipotesi alternativa

$$H_1 : \mu_x \neq \mu_y$$

Siccome $\overline{X}$ è uno stimatore di μ_x e $\overline{Y}$ è uno stimatore di μ_y, segue che $\overline{X} - \overline{Y}$ può essere usato per stimare $\mu_x - \mu_y$. Perciò, pensando di riscrive l'ipotesi nulla come $H_0 : \mu_x - \mu_y = 0$, sembra ragionevole rifiutarla quando $\overline{X} - \overline{Y}$ è lontano da zero. Ovvero la forma del test dovrebbe essere la seguente

$$\begin{aligned} &\text{si rifiuta } H_0 \text{ se } |\overline{X} - \overline{Y}| > c \\ &\text{si accetta } H_0 \text{ se } |\overline{X} - \overline{Y}| \leq c \end{aligned} \tag{8.4.1}$$

per un opportuno valore di c.

In analogia con quanto fatto in precedenza, si può trovare il valore di c che rende questo test di livello di significatività α, se si conosce la distribuzione di $\overline{X} - \overline{Y}$

quando H_0 è valida. Possiamo allora riutilizzare i risultati della Sezione 7.4, e in particolare ricordiamo che

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right)$$

per cui

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}} \sim \mathcal{N}(0,1) \tag{8.4.2}$$

Allora quando H_0 è vera (e $\mu_x - \mu_y = 0$), si ha che la statistica del test,

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}} \tag{8.4.3}$$

ha distribuzione normale standard, e quindi, per ogni $\alpha \in [0,1]$,

$$P_{H_0}\left(-z_{\frac{\alpha}{2}} \leq \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

Da quest'ultima equazione si deduce facilmente che un test con livello di significatività α per verificare l'ipotesi nulla $H_0 : \mu_x = \mu_y$ contro l'ipotesi alternativa $H_1 : \mu_x \neq \mu_y$ è dato dalla regola seguente,

$$\begin{aligned} &\text{si rifiuta } H_0 \text{ se } \frac{|\overline{X} - \overline{Y}|}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}} > z_{\frac{\alpha}{2}} \\ &\text{si accetta } H_0 \text{ se } \frac{|\overline{X} - \overline{Y}|}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}} \leq z_{\frac{\alpha}{2}} \end{aligned} \tag{8.4.4}$$

Il Programma 8.4.1 può essere usato per calcolare il valore della statistica del test, che compare nell'Equazione (8.4.3).

Esempio 8.4.1. Vengono proposti due nuovi metodi di produzione per pneumatici. Per accertare quale dei due sia superiore, un produttore ottiene 10 gomme del primo tipo, e 8 dell'altro, e le prova in due sedi diverse, denotate con A e B rispettivamente. È noto dall'esperienza passata che i tempi di vita degli pneumatici hanno una distribuzione con media che dipende quasi escusivamente dalla fattura della gomma, e varianza che dipende quasi esclusivamente dalla sede di prova. In particolare, per la sede A la deviazione standard è di 4 000 chilometri, mentre per la sede B è di 6 000. Se il produttore vuole verificare l'ipotesi che le due medie di popolazione siano sostanzialmente identiche ad un livello di significatività del 5%, che conclusioni si traggono con i dati della Tabella 8.3?

Tabella 8.3 Tempi di vita degli pneumatici dell'Esempio 8.4.1.

Pneumatici testati nella sede A	Pneumatici testati nella sede B
61.1	62.2
58.2	56.6
62.3	66.4
64.0	56.2
59.7	57.4
66.2	58.4
57.8	57.6
61.4	65.4
62.2	
63.6	

Un calcolo diretto (o l'impiego del Programma 8.4.1), mostra che il valore della statistica del test è di 0.066. Un valore così piccolo (si ricordi che la distribuzione della statistica è normale standard), significa che l'ipotesi nulla viene accettata ad ogni livello di significatività ragionevole. In particolare per $\alpha = 0.05$, si ha che $z_{0.025} \approx 1.96$, un valore enormemente maggiore di 0.066. □

Si possono anche in questo contesto verificare ipotesi a una coda. Valga come unico esempio il seguente: se si vuole verificare l'ipotesi nulla $H_0 : \mu_x = \mu_y$ (oppure l'ipotesi nulla $H_0 : \mu_x \leq \mu_y$) contro l'ipotesi alternativa $H_1 : \mu_x > \mu_y$ la regola da usare è:

$$\begin{aligned} &\text{si rifiuta } H_0 \text{ se } \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}} > z_\alpha \\ &\text{si accetta } H_0 \text{ se } \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}} \leq z_\alpha \end{aligned} \tag{8.4.5}$$

8.4.2 Il caso in cui le varianze non sono note ma si suppongono uguali

Prendiamo nuovamente in considerazione i campioni indipendenti $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_m$, estratti da due popolazioni normali $\mathcal{N}(\mu_x, \sigma_x^2)$ e $\mathcal{N}(\mu_y, \sigma_y^2)$; supponiamo che i quattro parametri siano tutti incogniti e studiamo nuovamente come si possa verificare

$$H_0 : \mu_x = \mu_y \qquad \text{contro} \qquad H_1 : \mu_x \neq \mu_y$$

Possiamo dare una risposta se supponiamo che le due varianze incognite siano uguali tra di loro, ovvero imponiamo che

$$\sigma^2 := \sigma_x^2 = \sigma_y^2 \tag{8.4.6}$$

Come in precedenza, desideriamo rifiutare H_0 quando $\overline{X} - \overline{Y}$ è "lontano" da zero. Per capire quanto, calcoliamo le due varianze campionarie,

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$
$$S_y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y})^2$$

e ricordiamo che, per quanto detto nella Sezione 7.4 (in particolare con l'Equazione (7.4.10) di pagina 255),

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{S_p\sqrt{1/n + 1/m}} \sim t_{n+m-2} \tag{8.4.7}$$

dove S_p^2 è lo stimatore *pooled* di σ^2, definito da

$$S_p^2 := \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2} \tag{8.4.8}$$

Perciò quando H_0 è vera, e quindi $\mu_x - \mu_y = 0$, la statistica del test,

$$T := \frac{\overline{X} - \overline{Y}}{S_p\sqrt{1/n + 1/m}} \tag{8.4.9}$$

ha distribuzione t con $n + m - 2$ gradi di libertà. Ne segue che possiamo verificare l'ipotesi $\mu_x = \mu_y$ come segue,

$$\begin{aligned} &\text{si rifiuta } H_0 \text{ se } |T| > t_{\frac{\alpha}{2},n+m-2} \\ &\text{si accetta } H_0 \text{ se } |T| \le t_{\frac{\alpha}{2},n+m-2} \end{aligned} \tag{8.4.10}$$

dove $t_{\frac{\alpha}{2},n+m-2}$, come ricorda la Figura 8.5, è il valore di ascissa a cui – per la distribuzione t con $n + m - 2$ gradi di libertà – corrisponde una probabilità della coda destra di $\frac{\alpha}{2}$.

In alternativa si può eseguire il test determinando il p-dei-dati. Se si osservano i dati e si denota con v, il valore assunto da T, il p-dei-dati corrispondente è

$$\begin{aligned} p\text{-dei-dati} &= P(|T_{n+m-2}| \ge |v|) \\ &= 2P(T_{n+m-2} \ge |v|) \end{aligned} \tag{8.4.11}$$

dove T_{n+m-2} è una variabile aleatoria t di Student con $n + m - 2$ gradi di libertà.

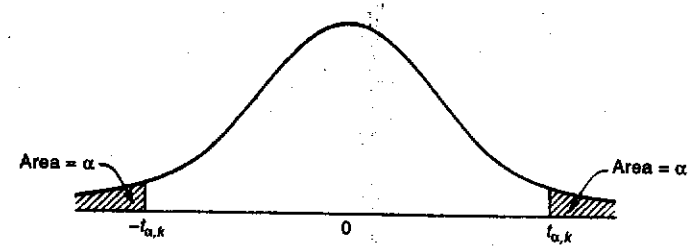


Figura 8.5 Densità di una distribuzione t con k gradi di libertà.

Se vogliamo invece verificare una ipotesi unilaterale come

$$H_0 : \mu_x \le \mu_y \qquad \text{contro} \qquad H_1 : \mu_x > \mu_y$$

allora H_0 deve essere rifiutata per valori elevati di T, e in particolare il test ha livello di significatività α quando

$$\begin{aligned} &\text{si rifiuta } H_0 \text{ se } T > t_{\alpha,n+m-2} \\ &\text{si accetta } H_0 \text{ se } T \le t_{\alpha,n+m-2} \end{aligned} \tag{8.4.12}$$

Se v è il valore assunto dalla statistica T, allora il p-dei-dati corrispondente è

$$p\text{-dei-dati} = P(T_{n+m-2} \ge v) \tag{8.4.13}$$

Il Programma 8.4.2, infine, permette di calcolare il valore della statistica del test e i relativi p-dei-dati.

Esempio 8.4.2. Un gruppo di 22 volontari presso un centro di ricerca media, viene esposto a vari tipi di virus influenzali e tenuto sotto controllo medico. Ad un campione casuale di 10 volontari viene somministrato un grammo di vitamina C quattro volte al giorno. Agli altri 12 volontari viene somministrato un placebo non distinguibile dal farmaco. I volontari vengono poi visitati spesso da un medico che non conosce la divisione in gruppi, e non appena uno di essi viene trovato guarito si registra la durata della malattia.

Alla fine dell'esperimento si possiedono i seguenti dati:

Trattati con vitamina C	Trattati con un placebo
5.5	6.5
6.0	6.0
7.0	8.5
6.0	7.0
7.5	6.5
6.0	8.0
7.5	7.5
5.5	6.5
7.0	7.5
6.5	6.0
	8.5
	7.0

Si può concludere che l'assunzione di 4 grammi di vitamina C al giorno abbia accorciato il decorso medio della malattia? A che livello di significatività?

Per *provare* l'ipotesi fatta, dobbiamo necessariamente assumerla come ipotesi alternativa, e riuscire a rifiutare l'ipotesi nulla corrispondente al livello di significatività desiderato. Eseguiamo quindi un test su

$$H_0 : \mu_p \leq \mu_c \qquad \text{contro} \qquad H_1 : \mu_p > \mu_c$$

dove μ_c e μ_p indicano i tempi medi di decorso dell'influenza assumendo la vitamina C e assumendo un placebo rispettivamente. Sembra ragionevole supporre che le varianze della durata della malattia nei due casi siano uguali, quindi eseguiamo il Programma 8.4.2, ottenendo il risultato della Figura 8.6. E quindi l'ipotesi nulla viene rifiutata ad un livello di significatività del 5%.

Naturalmente se non volessimo per qualche motivo impiegare il software, potremmo eseguire il test manualmente, determinando per prima cosa le statistiche $\overline{X}$, $\overline{Y}$, S_x^2 e S_y^2, per le quali otteniamo i risultati seguenti,

$$\overline{X} = 6.450 \qquad \overline{Y} = 7.125$$
$$S_x^2 \approx 0.581 \qquad S_y^2 \approx 0.778$$

calcolando poi lo stimatore S_p^2,

$$S_p^2 = \frac{9}{20} S_x^2 + \frac{11}{20} S_y^2 \approx 0.689$$

e la statistica del test,

$$v = \frac{-0.675}{\sqrt{0.689(1/10 + 1/12)}} \approx -1.90$$

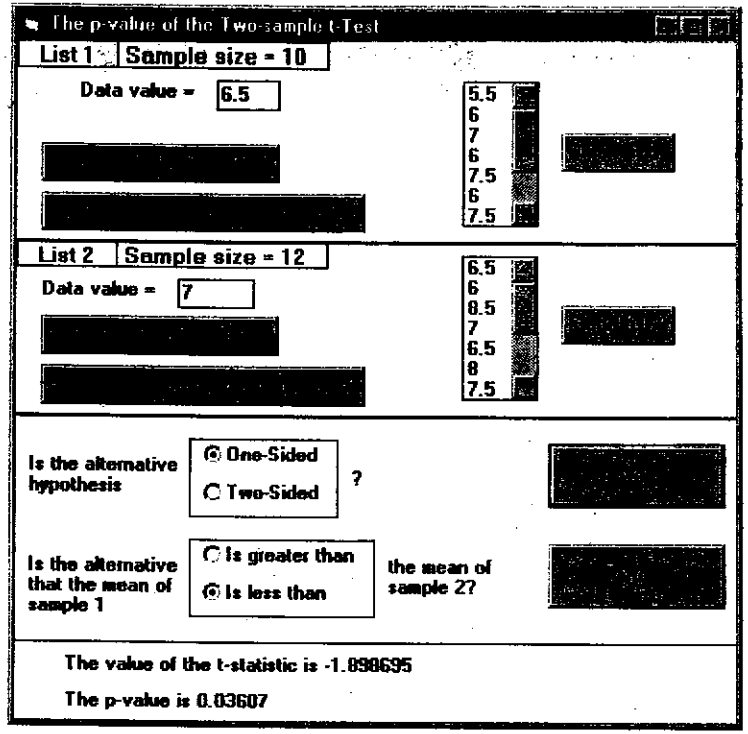


Figura 8.6 Verifica dell'ipotesi a una coda per l'Esempio 8.4.2.

Siccome $t_{0.05,20} \approx 1.725$, l'ipotesi nulla viene rifiutata ad un livello di significatività del 5%: quindi a questo livello di significatività i dati raccolti evidenziano un accorciamento del decorso dell'influenza, somministrando vitamina C. □

Esempio 8.4.3. Riconsideriamo l'Esempio 8.4.1, questa volta supponendo che le varianze siano ignote ma identiche.

Usando il Programma 8.4.2 si trova che il valore della statistica del test è 1.028, e il p-dei-dati relativo è

$$p\text{-dei-dati} = P(T_{16} > 1.028) \approx 0.3192$$

Perciò l'ipotesi nulla viene accettata a ogni livello di significatività minore di 31.92%, e quindi per ogni valore ragionevole di α. □

8.4.3 Il caso in cui le varianze sono ignote e diverse

Cosa possiamo dire se le varianze delle popolazioni, σ_x^2 e σ_y^2 oltre ad essere incognite non si possono assumere uguali? In tale situazione, siccome S_x^2 e S_y^2 sono gli stimatori naturali delle varianze, sembra sensato basare la nostra verifica della solita ipotesi

$$H_0 : \mu_x = \mu_y \qquad \text{contro} \qquad H_1 : \mu_x \neq \mu_y$$

sulla statistica

$$\frac{\overline{X} - \overline{Y}}{\sqrt{S_x^2/n + S_y^2/m}} \tag{8.4.14}$$

Questa statistica però ha una distribuzione complicata, che inoltre dipende dai parametri incogniti anche se assumiamo che H_0 sia valida. Per questi motivi, non può essere usata in generale, tuttavia, almeno nel caso in cui n e m sono entrambi dei numeri elevati, si può dimostrare che essa ha distribuzione approssimativamente normale standard. Perciò, quando n e m sono entrambi molto grandi, per verificare *approssimativamente* ad un livello di significatività α, l'ipotesi nulla $\mu_x = \mu_y$ contro l'ipotesi alternativa $\mu_x \neq \mu_y$,

$$\begin{aligned} &\text{si accetta } H_0 \text{ se } -z_{\frac{\alpha}{2}} \leq \frac{\overline{X} - \overline{Y}}{\sqrt{S_x^2/n + S_y^2/m}} \leq z_{\frac{\alpha}{2}} \\ &\text{si rifiuta } H_0 \text{ negli altri casi} \end{aligned} \tag{8.4.15}$$

Il problema di individuare un test di livello α esatto, per l'ipotesi che due popolazioni normali abbiano la stessa media è noto come problema di Behrens-Fisher, e a tutt'oggi non se ne conoscono soluzioni soddisfacenti.

La Tabella 8.4 presenta un riepilogo dei test a due code di questa sezione.

8.4.4 Il test t per campioni di coppie di dati

Ipotizziamo di essere interessati a determinare se l'installazione di un particolare dispositivo contro l'inquinamento possa influire sui consumi di una automobile. Un modo per realizzare questo progetto, consiste nel radunare un campione di n auto prive del dispositivo, e provare i consumi di ciascuna prima e dopo l'installazione.

I dati che raccogliamo alla fine sono descritti da n coppie di valori (X_i, Y_i), per $i = 1, 2, \ldots, n$, dove X_i e Y_i sono i consumi dell'auto i prima e dopo l'installazione. È importante notare che poiché le n automobili sono intrinsecamente diverse, non possiamo trattare $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_n$ come se si trattasse di campioni indipendenti. Infatti se sappiamo che X_1 è molto grande, ci aspetteremo che anche Y_1 lo sia, quindi non possiamo usare i metodi fin qui sviluppati per rispondere alla domanda.

Tabella 8.4 $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_n$ sono due campioni indipendenti, provenienti da popolazioni $\mathcal{N}(\mu_x, \sigma_x^2)$ e $\mathcal{N}(\mu_y, \sigma_y^2)$ rispettivamente.

$$H_0 : \mu_x = \mu_y \qquad H_1 : \mu_x \neq \mu_y$$

$$S_p^2 := \frac{n-1}{n+m-2} S_x^2 + \frac{m-1}{n+m-2} S_y^2$$

Si assume	Statistica del test, D_{ts}	Si rifiuta H_0 con livello di significatività α se...	p-dei-dati se $D_{ts} = t$
σ_x e σ_y note	$\dfrac{\overline{X} - \overline{Y}}{\sqrt{\sigma_x^2/n + \sigma_y^2/m}}$	$\ldots \lvert D_{ts} \rvert > z_{\frac{\alpha}{2}}$	$2P(Z > \lvert t \rvert)$
$\sigma_x = \sigma_y$ ignote	$\dfrac{\overline{X} - \overline{Y}}{S_p\sqrt{1/n + 1/m}}$	$\ldots \lvert D_{ts} \rvert > t_{\frac{\alpha}{2}, n+m-2}$	$2P(T_{n+m-2} > \lvert t \rvert)$
n e m grandi	$\dfrac{\overline{X} - \overline{Y}}{\sqrt{S_x^2/n + S_y^2/m}}$	$\ldots \lvert D_{ts} \rvert > z_{\frac{\alpha}{2}}$	$2P(Z > \lvert t \rvert)$

Un possibile approccio per verificare l'ipotesi che il dispositivo non influisca sui consumi è di prendere come dati le variazioni nel consumo di carburante, ponendo quindi $W_i := X_i - Y_i$, per $i = 1, 2, \ldots, n$. Se non vi fosse nessuna influenza del dispositivo, le W_i avrebbero media nulla, perciò possiamo verificare l'ipotesi che ci interessa con il test di

$$H_0 : \mu_W = 0 \qquad \text{contro} \qquad H_1 : \mu_W \neq 0$$

dove stiamo pensando che $W_1, W_2, \ldots, W_n$ sia un campione proveniente da una popolazione $\mathcal{N}(\mu_W, \sigma_W^2)$. Il test t presentato nella Sezione 8.3.2 ci fornisce la regola cercata:

$$\begin{aligned} &\text{si accetta } H_0 \text{ se } -t_{\frac{\alpha}{2}, n-1} \leq \sqrt{n}\frac{\overline{W}}{S_W} \leq t_{\frac{\alpha}{2}, n-1} \\ &\text{si rifiuta } H_0 \text{ negli altri casi} \end{aligned} \tag{8.4.16}$$

Esempio 8.4.4. Di recente nell'industria dei semiconduttori è stato introdotto un programma di sicurezza sul lavoro. Nella tabella seguente sono riportate le medie settimanali delle ore-uomo perse a causa di incidenti, per 10 stabilimenti dalle caratteristiche simili. Le medie sono state calcolate nel corso di un mese prima e un mese dopo la riforma.

Stabilimento	Prima	Dopo	Differenza
1	30.5	23.0	−7.5
2	18.5	21.0	+2.5
3	24.5	22.0	−2.5
4	32.0	28.5	−3.5
5	16.0	14.5	−1.5
6	15.0	15.5	+0.5
7	23.5	24.5	+1.0
8	25.5	21.0	−4.5
9	28.0	23.5	−4.5
10	18.0	16.5	−1.5

Determiniamo ad un livello di significatività del 5% se il programma di sicurezza è risultato efficace.

Dobbiamo verificare l'ipotesi

$$H_0 : \mu_d - \mu_p \geq 0 \qquad \text{contro} \qquad H_1 : \mu_d - \mu_p < 0$$

infatti questo ci permetterà di stabilire se vi sia nei dati una forte evidenza che le ore-uome perse siano diminuite. Per eseguire il test utilizziamo il Programma 8.3.2, che ci fornisce un valore per la statistica del test di -2.266, con

$$p\text{-dei-dati} = P(T_9 \leq -2.266) \approx 0.025$$

Siccome il p-dei-dati è inferiore a 0.05, l'ipotesi che il programma non abbia avuto effetto viene rifiutata e concludiamo che esso sembra essere efficace se si giudica con livello di significatività superiore al 2.5%. □

Si noti che il test t per campioni dipendenti[3] vale anche se le varianze delle due popolazioni non sono uguali.

8.5 La verifica delle ipotesi sulla varianza di una popolazione normale

Sia $X_1, X_2, \ldots, X_n$ un campione proveniente da una popolazione normale con media incognita μ e varianza incognita σ^2, e supponiamo di volere verificare l'ipotesi nulla

$$H_0 : \sigma^2 = \sigma_0^2 \qquad \text{contro l'alternativa} \qquad H_1 : \sigma^2 \neq \sigma_0^2$$

per un valore di σ_0^2 fissato.

[3] Questo tipo di test in italiano è chiamato anche test t per dati appaiati, oppure con la dicitura inglese, *paired t-test*, [*N.d.T.*]

Per ottenere un test, ricordiamo dalla Sezione 6.5.2 che $(n-1)S^2/\sigma^2$ ha distribuzione chi-quadro con $n-1$ gradi di libertà, così quando H_0 è vera,

$$\frac{S^2}{\sigma_0^2}(n-1) \sim \chi^2_{n-1} \tag{8.5.1}$$

e quindi

$$P_{H_0}\left(\chi^2_{1-\frac{\alpha}{2},n-1} \leq \frac{S^2}{\sigma_0^2}(n-1) \leq \chi^2_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$$

Perciò la regola da adottare è la seguente:

$$\begin{aligned} &\text{si accetta } H_0 \text{ se } \chi^2_{1-\frac{\alpha}{2},n-1} \leq \frac{S^2}{\sigma_0^2}(n-1) \leq \chi^2_{\frac{\alpha}{2},n-1} \\ &\text{si rifiuta } H_0 \text{ negli altri casi} \end{aligned} \tag{8.5.2}$$

Il test precedente può anche essere implementato come segue: si osservano i dati; si calcola il valore c assunto dalla statistica del test, ovvero $(n-1)S^2/\sigma_0^2$; si determina poi la probabilità che una chi-quadro con $n-1$ gradi di libertà sia (1) più piccola di c, (2) più grande di c. L'ipotesi nulla viene rifiutata se una di queste due probabilità è inferiore ad $\frac{\alpha}{2}$. Altrimenti detto, il p-dei-dati del test è

$$p\text{-dei-dati} = 2\min\{P(\chi^2_{n-1} \leq c), \quad 1 - P(\chi^2_{n-1} \leq c)\} \tag{8.5.3}$$

La quantità $P(\chi^2_{n-1} \leq c)$ può essere ricavata col Programma 5.8.1a. Il p-dei-dati per un test a una coda si trova analogamente.

Esempio 8.5.1. È stata appena installata una nuova macchina che deve controllare la quantità di nastro su un rocchetto. Questa macchina si può considerare efficiente se la deviazione standard della quantità di nastro selezionata non supera i 0.15 cm. Se un campione di 20 pezzi fornisce una varianza campionaria $S^2 = 0.025$ cm^2, è giustificato concludere che la macchina non è efficiente?

Prendiamo come H_0 l'ipotesi che la macchina sia efficiente: visto che un rifiuto di H_0 è una scelta forte, questo ci garantisce un ottimo controllo nell'eventuale conclusione che la macchina non è efficiente. Le due ipotesi sono,

$$H_0 : \sigma^2 \leq 0.0225 \qquad \text{e} \qquad H_1 : \sigma^2 > 0.0225$$

perciò dovremo rifiutare l'ipotesi nulla quando S^2 è troppo grande. Da questo vediamo che il p-dei-dati di questo test è pari alla probabilità che una chi-quadro con 19 gradi di libertà sia maggiore del valore osservato, $19 \cdot S^2/0.0225 \approx 21.11$, e quindi

$$\begin{aligned} p\text{-dei-dati} &\approx P(\chi^2_{19} > 21.11) \\ &\approx 1 - 0.6693 = 0.3307 \end{aligned}$$

dove abbiamo usato il Programma 5.8.1a. Si conclude allora che il valore osservato $S^2 = 0.025$ non è così grande da precludere la possibilità che $\sigma^2 \leq 0.0225$, e l'ipotesi nulla va accettata. □

8.5.1 Verificare se due popolazioni normali hanno la stessa varianza

Siano $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_m$ due campioni normali indipendenti, provenienti da popolazioni di parametri (incogniti) rispettivamente μ_x, σ_x^2 e μ_y, σ_y^2, e si consideri la verifica dell'ipotesi

$$H_0 : \sigma_x^2 = \sigma_y^2 \qquad \text{contro} \qquad H_1 : \sigma_x^2 \neq \sigma_y^2$$

Se definiamo le varianze campionarie come al solito,

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$S_y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \overline{Y})^2$$

allora, come sappiamo dalla Sezione 6.5.2, $(n-1)S_x^2/\sigma_x^2$ e $(m-1)S_y^2/\sigma_y^2$ sono due chi-quadro indipendenti con $n-1$ e $m-1$ gradi di libertà rispettivamente. Quindi $(S_x^2/\sigma_x^2)/(S_y^2/\sigma_y^2)$ ha distribuzione F con parametri $n-1$ e $m-1$, e quando H_0 è vera,

$$\frac{S_x^2}{S_y^2} \sim F_{n-1,m-1} \tag{8.5.4}$$

da cui si deduce che

$$P_{H_0}\left(F_{1-\frac{\alpha}{2},n-1,m-1} \leq \frac{S_x^2}{S_y^2} \leq F_{\frac{\alpha}{2},n-1,m-1}\right) = 1-\alpha$$

Perciò la regola da adottare è la seguente:

$$\begin{aligned} &\text{si accetta } H_0 \text{ se } F_{1-\frac{\alpha}{2},n-1,m-1} \leq \frac{S_x^2}{S_y^2} \leq F_{\frac{\alpha}{2},n-1,m-1} \\ &\text{si rifiuta } H_0 \quad \text{negli altri casi} \end{aligned} \tag{8.5.5}$$

In alternativa, si calcola il valore v assunto dalla statistica del test, S_x^2/S_y^2; si determina $P(F_{n-1,m-1} \leq v)$, dove $F_{n-1,m-1}$ ha distribuzione F di parametri $n-1$ e $m-1$; se tale probabilità risulta minore di $\frac{\alpha}{2}$ (indicando che S_x^2 è molto minore di S_y^2) o maggiore di $1-\frac{\alpha}{2}$ (indicando che S_x^2 è molto maggiore di S_y^2), l'ipotesi nulla deve essere rifiutata. In altri termini, il p-dei-dati del test è dato da

$$p\text{-dei-dati} = 2\min\left\{P(F_{n-1,m-1} \leq v),\ 1 - P(F_{n-1,m-1} \leq v)\right\} \tag{8.5.6}$$

e il test impone di rifiutare H_0 ogni volta che il livello di significatività α è maggiore o uguale al p-dei-dati.

Esempio 8.5.2. Per facilitare una certa reazione chimica si deve scegliere tra due catalizzatori diversi. Per verificare se la varianza nella quantità di prodotto con i due catalizzatori sia la stessa, si fanno 10 esperimenti con il primo e 12 con il secondo, ottenendo delle varianze campionarie di $S_1^2 = 0.14$ e $S_2^2 = 0.28$. Possiamo rifiutare ad un livello di significatività del 5% l'ipotesi che le varianze siano uguali?

Il Programma 5.8.3, che calcola la funzione di ripartizione delle distribuzioni F, ci dice che

$$P(F_{9,11} \leq 0.5) \approx 0.154$$

per cui

$$p\text{-dei-dati} \approx 2\min(0.154,\ 0.846) = 0.308$$

Quindi l'ipotesi nulla deve essere accettata. □

8.6 La verifica di ipotesi su una popolazione di Bernoulli

La distribuzione binomiale compare frequentemente nei problemi dell'ingegneria. Un esempio tipico è un processo produttivo dal quale si ottengono oggetti che possono appartenere a due categorie, come "accettabili" o "difettosi". Una ipotesi di lavoro che spesso viene assunta è che ogni oggetto prodotto sia difettoso in maniera indipendente da tutti gli altri con probabilità p. In questo modo il numero di difetti in un campione di n pezzi ha distribuzione binomiale di parametri (n, p). Consideriamo allora la verifica dell'ipotesi

$$H_0 : p \leq p_0 \qquad \text{contro l'alternativa} \qquad H_1 : p > p_0$$

dove p_0 è un valore assegnato.

Se denotiamo con X il numero di pezzi difettosi in un campione di n, dobbiamo certamente rifiutare H_0 quando X è troppo grande. Per calcolare poi quanto grande deve essere per giustificare un rifiuto dell'ipotesi nulla ad un livello di significatività pari ad α, notiamo che

$$\begin{aligned} P(X \geq k) &= \sum_{i=k}^{n} P(X = i) \\ &= \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \end{aligned}$$

È certamente intuitivo (e può essere dimostrato facilmente) che $P(X \geq k)$ è una funzione crescente di p; infatti la probabilità che un campione contenga k o più pezzi difettosi cresce con p. Usando questo fatto, è immediato che, quando H_0 è vera (e

quindi $p \leq p_0$),

$$P_{H_0}(X \geq k) \leq \sum_{i=k}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i}$$

Per verificare le ipotesi suddette ad un livello di significatività α, si deve rifiutare H_0 quando

$$X \geq k^*$$

dove con k^* si è denotato il più piccolo numero intero k tale che $\sum_{i=k}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \alpha$. In formule,

$$k^* := \min\left\{k : \sum_{i=k}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \alpha\right\} \tag{8.6.1}$$

Un modo migliore per implementare il test consiste ne determinare prima il valore x della statistica del test, X, e poi calcolare il p-dei-dati come segue,

$$p\text{-dei-dati} = \sum_{i=x}^{n} \binom{n}{i} p_0^i (1-p_0)^{n-i} \tag{8.6.2}$$

Esempio 8.6.1. Un produttore di circuiti integrati afferma che non più del 2% dei pezzi da lui venduti sono difettosi. Una compagnia di prodotti elettronici, colpita da una dichiarazione così forte, acquista una grossa quantità di tali circuiti. Per determinare se tale affermazione sia davvero completamente rispettata, la compagnia decide di provare un campione di 300 pezzi, e ne trova 10 di difettosi. Questo risultato è tale da negare quanto annunciato dal produttore?

Verifichiamo l'ipotesi nulla $p \leq 0.02$ ad un livello di significatività del 5%. Per capire se dobbiamo rifiutare H_0, ipotizziamo che p sia 0.02 e calcoliamo la probabilità che in un campione di 300 pezzi se ne trovino 10 o più di difettosi. Siccome questa grandezza è esattamente il p-dei-dati, se troviamo un valore inferiore a 0.05 dobbiamo rifiutare l'affermazione del produttore:

$$\begin{aligned} P_{0.02}(X \geq 10) &= 1 - P_{0.02}(X < 10) \\ &= 1 - \sum_{i=0}^{9} \binom{300}{i} (0.02)^i (0.98)^{n-i} \\ &\approx 0.0818 \qquad \text{dal Programma 3.1} \end{aligned}$$

perciò quanto dichiarato dal produttore non può essere rifiutato con il 5% di significatività. □

Quando la numerosità del campione è elevata, possiamo ottenere un test approssimativo con significatività α, utilizzando la distribuzione normale. Poiché infatti quando n è molto grande X è approssimativamente normale, con media e varianza,

$$E[X] = np \qquad \text{Var}(X) = np(1-p)$$

ne segue che

$$\frac{X - np}{\sqrt{np(1-p)}} \sim \mathcal{N}(0,1) \tag{8.6.3}$$

sarà approssimativamente normale standard, e quindi per ottenere un test che confronti le ipotesi $H_0 : p \leq p_0$ e $H_1 : p > p_0$, si deve rifiutare l'ipotesi nulla quando

$$\frac{X - np_0}{\sqrt{np_0(1-p_0)}} \geq z_\alpha$$

Esempio 8.6.2. Per i dati dell'Esempio 8.6.1, il valore della statistica del test, $(X - np_0)/\sqrt{np_0(1-p_0)}$ è di $(10 - 300 \times 0.02)/\sqrt{300 \times 0.02 \times 0.98} \approx 1.65$, da cui usando l'approssimazione normale segue che l'ipotesi nulla va rifiutata per tutti i livelli di significatività maggiori o uguali al valore del p-dei-dati,

$$p\text{-dei-dati} \approx P(Z \geq 1.65) \approx 0.0495$$

Così ad esempio H_0 verrebbe rifiutata al 5% di significatività, al contrario di quanto ottenuto con il test esatto realizzato nell'Esempio 8.6.1. Quanto detto mette in luce i rischi in cui si incorre utilizzando il test approssimato: se il campione non è sufficientemente numeroso si può infatti pervenire a una conclusione diversa da quella del test esatto. Una buona regola pratica per stabilire l'applicabilità dell'approssimazione gaussiana consiste nel valutare n anche in relazione con p_0: il p-dei-dati del test esatto e quello del test approssimato saranno davvero vicini solo quando $np_0(1-p_0)$ è 20 o più. Si noti che nell'esempio in questione, $np_0(1-p_0) \approx 6$ e quindi non stupisce troppo che il test approssimato abbia portato a una conclusione diversa rispetto a quello esatto. □

In questo contesto è più frequente imbattersi in test a una coda, comunque quelli bilaterali non presentano difficoltà ulteriori. Proviamo a verificare

$$H_0 : p = p_0 \qquad \text{contro l'alternativa} \qquad H_1 : p \neq p_0$$

Se la variabile aleatoria X, che è binomiale con parametri n e p, viene ossevata ed assume il valore x, sarà necessario rifiutare l'ipotesi nulla quando x cadrà molto lontano da quello che è il valore atteso quando p è uguale a p_0. Più precisamente, il test rifiuterà H_0 quando

$$P_{p_0}(X \geq x) \leq \frac{\alpha}{2} \qquad \text{oppure} \qquad P_{p_0}(X \leq x) \leq \frac{\alpha}{2}$$

Il valore del p-dei-dati corrispondente è quindi

$$p\text{-dei-dati} = 2\min\{P_{p_0}(X \geq x),\ P_{p_0}(X \leq x)\} \tag{8.6.4}$$

Esempio 8.6.3. I dati storici di uno stabilimento industriale mostrano che la percentuale di pezzi difettosi prodotti è del 4%. Essendosi di recente concluso uno scontro sindacale particolarmente astioso, il management dell'azienda è curioso di capire se questo porterà a un cambiamento apprezzabile di tale cifra. Preso un campione di 500 pezzi, se ne trovano 16 di difettosi (pari al 3.2%). Si può affermare con livello di significatività del 5% che vi sia stato qualche cambiamento?

Per potere concludere che è cambiato qualcosa, i dati dovrebbero essere abbastanza forti da rifiutare l'ipotesi nulla $H_0 : p = 0.04$, in favore dell'ipotesi alternativa $H_1 : p \neq 0.04$, dove p è la probabilità che un pezzo sia difettoso. Il p-dei-dati calcolato per 16 difettosi su 500 è dato da

$$p\text{-dei-dati} = 2\min\{P(X \geq 16),\ P(X \leq 16)\}$$

dove X è binomiale di parametri $n = 500$ e $p_0 = 0.04$. Siccome $E[X] = 20$, si deduce che $P(X \geq 16) > P(X \leq 16)$ e quindi il p-dei-dati è $2P(X \leq 16)$. Poiché X ha media 20 e deviazione standard $\sqrt{20 \times 0.96} \approx 4.38$, è chiaro che il doppio della probabilità che X sia minore di 16 – un valore che dista meno di una deviazione standard dalla sua media – non sarà così piccola da giustificare un rifiuto. In effetti è possibile calcolare che

$$p\text{-dei-dati} = 2P(X \leq 16) \approx 0.432$$

chiarendo oltre ogni dubbio che non vi è evidenza sufficiente a rifiutare l'ipotesi che la percentuale di pezzi difettosi sia rimasta invariata. □

8.6.1 Verificare se due popolazioni di Bernoulli hanno lo stesso parametro

Immaginiamo di volere confrontare due diversi metodi di fabbricazione per transistor. Indichiamo con p_1 e p_2 le probabilità (incognite) che un pezzo prodotto con i metodi 1 e 2 sia difettoso; raccogliamo poi campioni di numerosità n_1 e n_2 di transistor fabbricati nei due modi, e indichiamo con X_1 e X_2 il numero di pezzi difettosi trovati. In questo modo X_1 e X_2 sono variabili aleatorie binomiali indipendenti, con parametri (n_1, p_1) e (n_2, p_2) rispettivamente. In questa sezione sviluppiamo il cosiddetto *test di Fisher-Irwin*, che permette di confrontare p_1 e p_2.

Se desideriamo vagliare l'ipotesi nulla $p_1 = p_2$, pare sensato che essa venga rifiutata quando la frazione di pezzi difettosi prodotti col primo e col secondo metodo è molto diversa, ovvero quando X_1/n_1 e X_2/n_2 sono distanti tra loro. Per quantificare meglio il test, si noti che quando H_0 è valida, e quindi p_1 e p_2 sono uguali,

gli $n_1 + n_2$ pezzi prodotti complessivamente hanno tutti la medesima probabilità di essere difettosi. Se indichiamo con $k := X_1 + X_2$ il numero totale dei difettosi, essi saranno allora distribuiti come una selezione casuale di k elementi all'interno di un gruppo di $n_1 + n_2$ oggetti. Se a questo punto distinguiamo i due tipi di transitor (è come se stessimo estrendo k palline da un'urna che ne contiene n_1 di bianche e n_2 di nere), quelli difettosi tra gli n_1 prodotti col primo metodo (ovvero il numero delle palline bianche estratte dall'urna, proseguendo l'analogia) avrà distribuzione *ipergeometrica*[4] di parametri n_1, n_2 e k. In altri termini, la distribuzione del numero di pezzi difettosi nel primo campione X_1, condizionata all'evento che il numero totale di pezzi difettosi nei due campioni sia k, è la seguente

$$P_{H_0}(X_1 = i | X_1 + X_2 = k) = \frac{\binom{n_1}{i}\binom{n_2}{k-i}}{\binom{n_1+n_2}{k}}, \quad \begin{matrix} k = 0, 1, \ldots, n_1 + n_2 \\ i = 0, 1, \ldots, k \end{matrix} \tag{8.6.5}$$

Perciò, volendo realizzare un test per verificare l'ipotesi nulla

$$H_0 : p_1 = p_2 \qquad \text{contro l'alternativa} \qquad H_1 : p_1 \neq p_2$$

si osserva quanto valgono $X_1 = x_1$ e $X_2 = x_2$ e si calcola la somma $k = x_1 + x_2$; denotata quindi con X una variabile aleatoria ipergeometrica di parametri n_1, n_2 e k, si conclude che se $P(X \leq x_1)$ è molto piccola, la frazione di pezzi difettosi è significativamente minore nel primo campione, mentre se $P(X \geq x_1)$ è molto piccola, accade il viceversa. Perciò la regola del test deve essere la seguente:

$$\begin{aligned} &\text{si rifiuta } H_0 \text{ se } P(X \leq x_1) < \tfrac{\alpha}{2} \text{ o } P(X \geq x_1) < \tfrac{\alpha}{2} \\ &\text{si accetta } H_0 \text{ negli altri casi} \end{aligned} \tag{8.6.6}$$

Il p-dei-dati relativo a questo test si può quindi calcolare tramite

$$p\text{-dei-dati} = 2\min\{P(X \leq x_1),\ P(X \geq x_1)\} \tag{8.6.7}$$

8.6.1.1 Calcoli relativi al test di Fisher-Irwin

Per utilizzare il test di Fisher-Irwin, dobbiamo essere in grado di calcolare le probabilità relative alla distribuzione ipergeometrica. Si può usare il fatto che se X è una variabile aleatoria ipergeometrica di parametri n_1, n_2 e k, allora $P(X = i)$ può essere calcolata ricorsivamente. Si tenga presente che X non può essere minore di

[4] Si veda anche l'Esempio 5.3.3 a pagina 163 per una derivazione formale di questo risultato.

$k - n_2$ se questo numero è positivo, e non può essere minore di zero altrimenti, quindi il punto da cui fare partire la ricorsione è variabile. Il passo da i a $i+1$ è invece semplicemente dato da,

$$\frac{P(X=i+1)}{P(X=i)} = \frac{\binom{n_1}{i+1}\binom{n_2}{k-i-1}}{\binom{n_1}{i}\binom{n_2}{k-i}} = \frac{(n_1-i)(k-i)}{(i+1)(n_2-k+i+1)} \tag{8.6.8}$$

Il Programma 8.6.1 del software abbinato al libro, usa esattamente questo procedimento per calcolare il p-dei-dati per il test di Fisher-Irwin sull'uguaglianza dei parametri di due popolazioni bernouilliane. Per come è fatto, questo programma funziona al meglio se la probabilità che un pezzo sia difettoso risulta minore di 0.5, quindi in caso più della metà dei pezzi prodotti sia difettoso, conviene scambiare le quantità di oggetti difettosi ed accettabili, in modo da ottenere un risultato più preciso.

Esempio 8.6.4. Supponiamo che su 100 transistor prodotti, il metodo 1 abbia dato 20 pezzi non accettabili, mentre il metodo 2 ne ha dati 12. Possiamo concludere al 10% di significatività che i due metodi sono equivalenti?

Eseguendo il Programma 8.6.1 otteniamo che

$$p\text{-dei-dati} \approx 0.1763$$

per cui l'ipotesi che i due metodi siano in realtà equivalenti non va rifiutata. □

Quando n_1 e n_2 sono molto grandi è possibile usare l'approssimazione normale delle variabili aleatorie binomiali, per ottenere un test semplificato dell'ipotesi $H_0 : p_1 = p_2$. Il Problema 59 enuncia il risultato nei particolari.

8.7 Ipotesi sulla media di una distribuzione di Poisson

Sia X una variabile aleatoria di Poisson con media λ, e supponiamo di volere confrontare le ipotesi

$$H_0 : \lambda = \lambda_0 \qquad \text{oppure} \qquad H_1 : \lambda \neq \lambda_0$$

Se x è il valore osservato per X, un test con livello di significatività α deve rifiutare l'ipotesi nulla se

$$P_{\lambda_0}(X \geq x) \leq \frac{\alpha}{2} \qquad \text{o se} \qquad P_{\lambda_0}(X \leq x) \leq \frac{\alpha}{2}$$

dove con P_{λ_0} intendiamo come al solito la probabilità calcolata assumendo che X abbia media λ_0. Di conseguenza, il p-dei-dati è dato da

$$p\text{-dei-dati} = 2\min\{P_{\lambda_0}(X \leq x),\ P_{\lambda_0}(X \geq x)\} \tag{8.7.1}$$

Il calcolo di tutte probabilità necessarie può essere effettuato con l'ausilio del Programma 5.2.

Esempio 8.7.1. La direzione dice che il numero medio di circuiti integrati difettosi prodotti ogni giorno non è superiore a 25, ma questa affermazione è in discussione. Verifichiamo quanto dichiarato al 5% di significatività, sapendo che un campione di 5 giorni ha registrato 28, 34, 32, 38 e 22 chip difettosi.

Poiché ogni giorno vengono prodotti un gran numero di circuiti integrati, e ciascuno ha una piccola probabilità di risultare difettoso, è naturale supporre che la distribuzione del numero di pezzi difettosi prodotti ogni giorno sia di Poisson. Sia λ la sua media. Per decidere se l'affermazione del produttore sia credibile, eseguiamo un test per confrontale le ipotesi

$$H_0 : \lambda \leq 25 \qquad \text{e} \qquad H_1 : \lambda > 25$$

Se H_0 fosse valida, la distribuzione del numero *totale* di pezzi difettosi nei 5 giorni sarebbe poissoniana di media non maggiore di 125 (la somma di poissoniane indipendenti è una poissoniana). Il numero totale di difetti riscontrati è di 154, e il p-dei-dati che ne risulta è dato da

$$\begin{aligned} p\text{-dei-dati} &= P_{125}(X \geq 154) \\ &= 1 - P_{125}(X \leq 154) \approx 0.0066 \end{aligned}$$

dove per l'ultimo passaggio abbiamo utilizzato il Programma 5.2. Perciò la tesi sostenuta dal produttore va rifiutata al 5% di significatività, e in realtà sarebbe stata rifiutata anche all'1%. □

Si tenga sempre presente che, in assenza di un software di calcolo come il Programma 5.2, una distribuzione di Poisson con media λ molto grande, è approssimativamente normale con media e varianza entrambe pari a λ.

8.7.1 Testare la relazione tra i parametri di due popolazioni di Poisson

Siano X_1 e X_2 due variabili aleatorie di Poisson e indipendenti, con medie rispettivamente λ_1 e λ_2; supponiamo di volere vagliare l'ipotesi

$$H_0 : \lambda_2 = c\lambda_1 \qquad \text{contro} \qquad H_1 : \lambda_2 \neq c\lambda_1$$

per una costante c assegnata. Il test che costruiremo è di tipo condizionale (di spirito simile a quello di Fisher-Irwin della Sezione 8.6.1), ed è basato sul fatto che la distribuzione di X_1 condizionata al valore della somma di X_1 e X_2, è binomiale. In particolare vale l'enunciato seguente.

Proposizione 8.7.1. Se X_1 e X_2 sono due variabili aleatorie di Poisson indipendenti, con media λ_1 e λ_2 rispettivamente, allora per ogni valore di $n = 1, 2, \ldots$, la distribuzione di X_1 condizionata all'evento $\{X_1 + X_2 = n\}$ è binomiale, con parametri n e $\lambda_1/(\lambda_1 + \lambda_2)$.

Dimostrazione. Occorre provare che, per ogni $n = 1, 2, \ldots$ e ogni $k = 0, 1, \ldots, n$, vale la relazione

$$P(X_1 = k | X_1 + X_2 = n) = \binom{n}{k} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{n-k} \tag{8.7.2}$$

Ciò è provato dai semplici passaggi seguenti:

$$\begin{aligned}
&P(X_1 = k | X_1 + X_2 = n) \\
&\quad = \frac{P(X_1 = k, X_1 + X_2 = n)}{P(X_1 + X_2 = n)} \\
&\quad = \frac{P(X_1 = k, X_2 = n - k)}{\frac{(\lambda_1 + \lambda_2)^n}{n!} e^{-(\lambda_1 + \lambda_2)}} && X_1 + X_2 \text{ è di Poisson con media } \lambda_1 + \lambda_2, \text{ per quanto detto a pagina 158} \\
&\quad = P(X_1 = k) P(X_2 = n - k) \frac{n! e^{\lambda_1 + \lambda_2}}{(\lambda_1 + \lambda_2)^n} && \text{per l'indipendenza} \\
&\quad = \frac{\lambda_1^k}{k!} e^{-\lambda_1} \cdot \frac{\lambda_2^{n-k}}{(n-k)!} e^{-\lambda_2} \cdot \frac{n! e^{\lambda_1 + \lambda_2}}{(\lambda_1 + \lambda_2)^n} \\
&\quad = \frac{n!}{k!(n-k)!} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{n-k} && \square
\end{aligned}$$

La Proposizione 8.7.1 afferma che quando l'ipotesi nulla è verificata, la distribuzione di X_1 condizionata al valore n osservato per $X_1 + X_2$, è binomiale di parametri n e $p_c := 1/(1 + c)$. Da questo concludiamo che, detto x_1 il valore di X_1, e n quello assunto da $X_1 + X_2$, si deve rifiutare H_0 qualora

$$P(Y \geq x_1) \leq \frac{\alpha}{2} \qquad \text{oppure} \qquad P(Y \leq x_1) \leq \frac{\alpha}{2} \tag{8.7.3}$$

dove Y ha distribuzione binomiale di parametri n e p_c.

Esempio 8.7.2. Un'azienda manifatturiera possiede due grossi stabilimenti. Se gli incidenti nel primo di essi per le ultime 8 settimane sono stati 16, 18, 9, 22, 17, 19, 24 e 8, mentre quelli dell'altro impianto nelle ultime 6 settimane sono stati 22, 18, 26, 30, 25 e 28, possiamo concludere al 5% di significatività che le condizioni di sicurezza dei due impianti sono diverse?

Poiché in ogni minuto vi è una piccola probabilità che vi sia un incidente, sembra plausibile che il numero di incidenti settimanali possa essere descritto da una popolazione di Poisson. Dette quindi X_1 e X_2 le due variabili aleatorie indicanti il numero complessivo di incidenti dei periodi considerati nei due stabilimenti, e indicate con λ_1 e λ_2 le relative medie, si potrà dire che le condizioni di sicurezza sono le stesse se

$$\lambda_2 = \frac{3}{4}\lambda_1$$

dove si è tenuto conto della diversa lunghezza dei periodi in esame. Di conseguenza, vogliamo verificare l'ipotesi nulla

$$H_0 : \lambda_2 = \frac{3}{4}\lambda_1 \qquad \text{contro} \qquad H_1 : \lambda_2 \neq \frac{3}{4}\lambda_1$$

Essendo $X_1 = 133$ e $X_2 = 149$, si pone $n = 133 + 149 = 282$; essendo $c = \frac{3}{4}$, si pone $p_c = \frac{1}{1+3/4} = \frac{4}{7}$; detta quindi Y una variabile aleatoria con distribuzione binomiale di parametri $(282, 4/7)$, il p-dei-dati del test risulta dato da

$$\begin{aligned}
p\text{-dei-dati} &= 2\min\{P(Y \geq 133),\ P(Y \leq 133)\} \\
&\approx 2\min\{1 - 0.00072,\ 0.00047\} && \text{usando il Programma 5.1} \\
&= 0.00094
\end{aligned}$$

Concludiamo che si può senz'altro escludere che le condizioni di sicurezza dei due impianti siano le stesse. □

Problemi

1. In un processo il giudice (o una giuria) deve decide se l'imputato è innocente o colpevole.
   (a) Nell'ambito della verifica delle ipotesi, e per un sistema giuridico che sostenga l'innocenza dell'imputato fino a che non sia stato dimostrato il contrario, quale dovrebbe essere l'ipotesi nulla?
   (b) Quale pensi sarebbe un livello di significatività appropriato?

2. Un laboratorio di analisi possiede una colonia di diverse migliaia di topi, usate come cavie. È noto che il peso medio dei topi è di 32 grammi, con una deviazione standard di 4 grammi. Uno scienziato chiede ad un assistente di selezionare un campione casuale di 25 cavie; decide poi di pesarle, per controllare che la casualità della scelta dell'assistente

non sia stata falsata da qualche criterio inconscio (se ad esempio i topi scelti fossero quelli più lenti nell'evitare la mano dell'assistente, questo potrebbe indicare una certa inferiorità fisica di questo gruppo). Se le 25 cavie risultano in un peso medio di 30.4 grammi, si può dire che questo evidenzi al 5% di significatività che il campione non è stato scelto in maniera casuale?

3. Una distribuzione di popolazione ha deviazione standard 20. Calcola il p-dei-dati per il test dell'ipotesi che la media sia 50, supponendo che la media campionaria su 64 osservazioni sia stata di **(a)** 52.5; **(b)** 55.0; **(c)** 57.5.

4. In un certo procedimento chimico, è di fondamentale importanza che il pH di uno dei reagenti sia esattamente 8.20. Si sa che il metodo usato per misurare tale pH fornisce valori con distribuzione normale con media pari al valore autentico e deviazione standard 0.02. Supponiamo che 10 misurazioni indipendenti abbiano dato i seguenti valori:

   8.18 8.16 8.17 8.22 8.19 8.17 8.15 8.21 8.16 8.18

   Che conclusioni si possono trarre con livello di significatività pari ad **(a)** $\alpha = 0.10$ e **(b)** $\alpha = 0.05$?

5. Si richiede che la pressione di rottura media di un certo tipo di fibra sia almeno pari a 200 psi. La nostra esperienza passata ci dice che la deviazione standard per questo genere di fibre è di 5 psi. Un campione di 8 esemplari ha fornito i valori seguenti:

   210 195 197.4 199 198 202 196 195.5

   Concluderesti **(a)** al 5% o **(b)** al 10% di significatività che la fibra non è accettabile?

6. Supponiamo di sapere che negli Stati Uniti la statura media di un maschio adulto è di 70 pollici, con una deviazione standard di 3 pollici. Per verificare che gli uomini di una città sono "nella media", si sceglie un campione di 20 maschi adulti e se ne misura la statura, ottenendo i risultati seguenti:

   72 68.1 69.2 72.8 71.2 72.2 70.8 74 66 70.3
   70.4 76 72.5 74 71.8 69.6 75.6 70.6 76.2 77

   Cosa concludi? Spiega quali assunzioni stai facendo.

7. Supponiamo di volere nuovamente affrontare il Problema 4, con le richieste seguenti: se il pH è realmente pari a 8.20, il test deve affermarlo con probabilità del 95%; d'altra parte, se il pH vero differisce da 8.20 di 0.03 (in una direzione qualsiasi), tale differenza deve essere evidenziata nel 95% almeno dei casi.

   (a) Come si può realizzare una verifica di questo tipo?

   (b) Quanto numeroso dovrà essere il campione scelto?

   (c) Se $\bar{x} = 8.31$, che conclusioni trai?

   (d) Se il pH vero fosse 8.32, quale sarebbe la probabilità di concludere che esso è diverso da 8.20, usando la procedura precedente?

8. Verifica che l'Equazione (8.3.7) resti valida anche quando $\mu_1 < \mu_0$.

9. Una compagnia farmaceutica vuole mettere in commercio un nuovo farmaco per la cura sintomatica delle emicranie, basato su un principio attivo particolarmente rapido a entrare in circolo. Per convincere l'ente preposto al controllo dei nuovi medicinali che il tempo medio che il farmaco impiega a raggiungere il sangue è inferiore ai 10 minuti, questa ditta raduna un campione di persone soggette ad emicranie e conduce un esperimento. Come vanno scelte l'ipotesi nulla e quella alternativa?

10. I salmoni cresciuti ogni anno in un allevamento commerciale hanno dei pesi con distribuzione normale di deviazione standard 1.2 libbre. La ditta dichiara che il peso medio dei suoi pesci quest'anno è superiore alle 7.6 libbre. Supponi che un campione casuale di 16 pesci sia risultato in un peso medio di 7.2 libbre. Si può dire che questo dato sia abbastanza forte da farci respingere l'affermazione dell'azienda **(a)** al 5% di significatività? **(b)** All'1% di significatività? **(c)** Quanto vale il p-dei-dati di questo test?

11. Si vuole verificare $H_0 : \mu \leq 100$ contro l'alternativa $H_1 : \mu > 100$. Supponiamo che un campione di 20 dati abbia dato una media campionaria pari a 105. Determina il p-dei-dati nel caso in cui la deviazione standard della popolazione sia nota e pari a **(a)** 5; **(b)** 10; **(c)** 15.

12. Il messaggio pubblicitario di un nuovo dentifricio afferma che esso è in grado di ridurre la frequenza delle carie dei bambini negli anni in cui ne sono soggetti. Supponiamo che il numero di carie all'anno per un bambino di quell'età abbia distribuzione con media 3 e varianza 1 e che uno studio dell'efficacia del nuovo prodotto, condotto su 2 500 bambini abbia rivelato un numero medio di carie all'anno pari a 2.95. Ipotizziamo che la varianza usando il dentifricio reclamizzato non sia diversa da quella naturale.

    (a) Questi dati sono abbastanza forti da convalidare al 5% di significatività l'annuncio pubblicitario?

    (b) Ti convincono a cambiare dentifricio?

13. La quantità di fenobarbitale contenuta nelle pillole vendute da una ditta farmaceutica può avere una certa variabilità, comunque il suo valore medio è dichiarato in 20.0 mg. Per convalidare questa affermazione, si analizza un campione di 25 pillole, trovando una media campionaria di 19.7 mg e una deviazione standard campionaria di 1.3 mg. Che conclusioni si possono trarre dai dati? Si può dire in particolare che i risultati di questo esperimento dimostrino che l'affermazione della ditta non era vera? Usa un livello di significatività del 5%.

14. Venti anni fa i maschi del primo anno di una certa scuola superiore erano in grado di fare in media 24 flessioni in 60 secondi. Per vedere se questo sia ancora vero al giorno d'oggi, si sceglie un campione casuale di 36 maschi del primo anno, e si trova una media campionaria di 22.5, con una deviazione standard di 3.1. Possiamo concludere che la media non è più pari a 24? Usa un livello di significatività del 5%.

15. Il tempo medio di risposta per una varietà di suini ad un particolare stimolo è di 0.8 secondi. Si somministrano 2 once di soluzione alcolica ad un campione di 28 suini e li si sottopone al medesimo stimolo, registrando un tempo medio di risposta di 1.0 secondi con una deviazione standard campionaria di 0.3 secondi. Si può concludere che l'alcool ha avuto un qualche effetto sui tempi di risposta dei suini? Usa il 5% di significatività.

16. Un medico ricercatore è convinto che la temperatura basale media delle persone (esteriormente) sane sia cresciuta nel tempo, e non sia più pari a 98.6 gradi Fahrenheit. Per dimostrarlo, egli misura la temperatura di 100 soggetti sani selezionati a caso, trovando una temperatura media di 98.74 gradi e una deviazione standard campionaria di 1.1 gradi. È vero che questi dati provano la sua congettura al 5% di significatività? E all'1%?

17. La pubblicità di una nuova auto afferma che essa è in grado di fare 30 miglia di guida in autostrada con un gallone di benzina. Volendo verificare questo fatto, si fanno 10 esperimenti indipendenti, e con quella quantità di carburante l'automobile copre 26, 24, 20, 25, 27, 25, 28, 30, 26 e 33 miglia. Si può credere all'annuncio? Che ipotesi stai facendo?

18. Un produttore afferma che la carica media di un certo tipo di batterie è di almeno 240 ampere-ora. Un campione di 18 batterie di questo tipo che è stato analizzato ha fornito i dati valori seguenti.

237 242 244 262 225 218 242 248 243
234 236 228 232 230 254 220 232 240

Assumendo che la distribuzione della carica sia approssimativamente normale, si può dire che i dati contraddicono le specifiche delle batterie?

19. Usa i dati dell'Esempio 2.3.9 di pagina 29 per verificare l'ipotesi nulla che il livello di rumore medio in prossimità della stazione centrale di Manhattan sia minore o uguale a 80 dB.

20. Una compagnia petrolifera dichiara che il contenuto di zolfo del suo carburante diesel non supera lo 0.15%. Per verificare questa ipotesi se ne analizzano 40 campioni, trovando un contenuto medio di 0.162% con deviazione standard campionaria di 0.40%. Usando il 5% di significatività possiamo confutare le affermazioni della compagnia?

21. Una azienda produce laminati plastici per uso industriale. Viene sviluppato un nuovo tipo di materiale, e si vorrebbe poterlo pubblicizzare dicendo che la resistenza media alla rottura del nuovo prodotto non è inferiore a 30.0 psi. I dati seguenti sono le pressioni di rottura di esemplari presi dalla linea di produzione. Si può dire basandosi su questi valori che tale dichiarazione sarebbe chiaramente ingiustificata?

30.1 27.7 31.2 29.1 32.7 29.8 24.3 33.4
22.5 28.9 26.4 32.5 27.5 31.4 22.8 21.7

Assumi che la popolazione sia normale e usa il 5% di significatività.

22. È stato affermato che un certo tipo di transistor bipolare ha un valore medio del guadagno di 210 almeno. Si prova un campione di questi transistor trovando una media campionaria di 200 e una deviazione standard campionaria di 35. Al 5% di significatività si dovrebbe rifiutare quanto affermato, **(a)** se l'ampiezza del campione era 25? **(b)** E se era 64?

23. Un produttore di condensatori afferma che la tensione di breakdown di un certo modello è mediamente superiore a 100 volt. Provando 12 di questi elementi si sono trovate le seguenti tensioni di breakdown,

96 98 105 92 111 114 99 103 95 101 106 97

Si può dire che questi dati confermino oppure che confutino quanto detto?

24. Si è pescato un campione di 10 pesci del lago A, misurandone la concentrazione di PCB con una certa tecnica. I valori trovati (in parti per milione) sono riportati nella tabella qui sotto, assieme a quelli di 8 pesci presi nel lago B, e il cui contenuto di PCB è stato misurato con una tecnica differente.

Lago A	11.5	10.8	11.6	9.4	12.4	11.4	12.2	11.0	10.6	10.8
Lago B	11.8	12.6	12.2	12.5	11.7	12.1	10.4	12.6		

Sapendo che i due metodi di misurazione portano ad errori statistici di varianza 0.09 e 0.16 rispettivamente, si può concludere ad un livello di significatività del 5% che i due laghi sono ugualmente inquinati?

25. Uno scienziato che si occupa di inquinamento ambientale vuole verificare se due campioni di soluzioni in suo possesso possono provenire dalla stessa sorgente. Se fosse così, i pH delle due soluzioni dovrebbero coincidere, e per stabilire se questo sia vero, vengono fatte 10 misurazioni indipendenti per ciascuna soluzione. Il metodo usato garantisce che i valori misurati hanno distribuzione normale con media pari al pH vero e deviazione standard di 0.05. I dati ottenuti sono i seguenti.

Soluzione A	6.24	6.31	6.28	6.30	6.25	6.26	6.24	6.29	6.22	6.28
Soluzione B	6.27	6.25	6.33	6.27	6.24	6.31	6.28	6.29	6.34	6.27

(a) Tali dati mostrano una apprezzabile differenza nei pH al 5% di significatività?

(b) Quanto vale il p-dei-dati di questo test?

26. Quelli che seguono sono due campioni indipendenti di due popolazioni diverse.

Campione 1	122	114	130	165	144	133	139	142	150
Campione 2	108	125	122	140	132	120	137	128	138

Denota con μ_1 e μ_2 le medie di popolazione rispettive, e determina il p-dei-dati del test di $H_0 : \mu_1 \leq \mu_2$ rispetto ad $H_1 : \mu_1 > \mu_2$, quando le deviazioni standard di popolazione sono rispettivamente $\sigma_1 = 10$ e **(a)** $\sigma_2 = 5$; **(b)** $\sigma_2 = 10$; **(c)** $\sigma_2 = 20$.

27. I dati presentati qui sotto costituiscono i tempi di vita (in centinaia di ore) di due tipi di valvole termoioniche. Lo studio passato di questo tipo di dati ci permette di dire che la loro distribuzione deve essere lognormale, ovvero i logaritmi dei tempi di vita hanno distribuzione normale. Assumendo che le varianze dei logaritmi siano uguali per i due campioni, verifica al 5% di significatività l'ipotesi che le distribuzioni coincidano interamente.

Tipo 1	32	84	37	42	78	62	59	74
Tipo 2	39	111	55	106	90	87	85	

28. Si misurano le viscosità di due diverse marche di olio per macchine, ottenendo i dati seguenti:

Marca 1	10.62	10.58	10.33	10.72	10.44	10.74	
Marca 2	10.50	10.52	10.58	10.62	10.55	10.51	10.53

Controlla l'ipotesi che la viscosità media delle due marche sia la stessa, assumendo che le popolazioni abbiano distribuzione normale con identica varianza.

29. Si suppone che la resistenza del cavo A sia maggiore di quella del cavo B. Dopo avere fatto varie prove su entrambi, trovi questi risultati (in ohm):

Cavo A	0.140	0.138	0.143	0.142	0.144	0.137
Cavo B	0.135	0.140	0.136	0.142	0.138	0.140

Che conclusioni puoi trarre ad un livello di significatività del 10%? Spiega quali assunzioni stai facendo.

Nei Problemi dal 30 al 37 puoi assumere che le distribuzioni delle popolazioni siano normali con la medesima varianza.

30. Un gruppo di 25 uomini di età compresa fra i 25 e i 30 anni, è stato selezionato per partecipare ad uno studio sul cuore. Di questi 11 erano fumatori e 14 no. I dati seguenti si riferiscono alla misurazione della loro pressione sistolica.

Fumatori	Non fumatori
124	130
134	122
136	128
125	129
133	118
127	122
135	116
131	127
133	135
125	120
118	122
	120
	115
	123

Usa questi dati per verificare l'ipotesi che la pressione sanguigna dei fumatori e dei non fumatori sia la stessa.

31. In un esperimento[5] del 1943, 10 ratti albini furono usati per studiare l'efficacia del tetracloruro di carbonio nel trattamento dei vermi. Le cavie furono infettate con larve, e dopo dieci giorni divise a caso in due gruppi di 5: il primo venne trattato con 0.032 cc della

[5] Whitlock and Bliss, "A bioassay technique for antihelminthics", *Journal of Parasitology*, vol. 29, pp. 48–58, 1943.

sostanza e il secondo con 0.063 cc. Due giorni dopo i ratti furono soppressi e venne contato il numero di vermi adulti formatisi nei loro corpi. Il gruppo con dosaggio inferiore ne aveva

421 462 400 378 413

mentre l'altro ne aveva

207 17 412 74 116

Questi dati dimostrano che il dosaggio superiore è stato più efficace?

32. Un docente è convinto che lo stipendio iniziale medio di un neolaureato in ingegneria industriale sia superiore a quello di un neolaureato in ingegneria civile. Per studiare questa ipotesi si intervista un campione di 16 elementi di entrambe le categorie, scelti a caso tra i laureati del 1993. I risultati dell'inchiesta sono una media campionaria di 47 700 dollari con una deviazione standard campionaria di 2 400 per i primi, e 46 400 di media con 2 200 di deviazione standard per i secondi. Confermeresti l'opinione del docente? Quanto vale il *p*-dei-dati?

33. In un laboratorio sperimetale si sta studiando un metodo (A) per produrre benzina a partire dal petrolio greggio. Prima di completare la sperimentazione, viene individuato un nuovo metodo di produzione B. Essendo comparabili tutti gli altri fattori, si decide che si abbandonerebbe il metodo A in favore del metodo B solo se il rendimento medio si dimostrasse chiaramente più alto. Si suppone che il rendimento dei due metodi abbia distribuzione normale; le deviazioni standard vere non sono state ancora ottenute per mancanza di tempo, ma non sembra ci siano motivi particolari per non assumerle uguali. Gli alti costi impongono limiti severi all'ampiezza dei campioni che possono essere ottenuti. Se non ci si può permettere un livello di significatività meno stringente dell'1%, che cosa consiglieresti, basandoti sui campioni aleatori seguenti? Le cifre rappresentano il rendimento in percentuale di petrolio greggio.

A	23.2	26.6	24.4	23.5	22.6	25.7	25.5
B	25.7	27.7	26.2	27.9	25.0	21.4	26.1

34. È stato condotto uno studio su come le abitudini alimentari delle donne si modifichino tra l'inverno e l'estate. Si è tenuto sotto osservazione un campione aleatorio di 12 femmine durante il mese di luglio, misurando tra le altre cose quale percentuale delle calorie da loro assunte provenisse dai grassi. Successivamente, osservazioni del tutto analoghe sono state compiute su un altro campione di 12 donne, nel mese di gennaio. I risultati sono riassunti qui sotto.

Luglio	32.2	27.4	28.6	32.4	40.5	26.2	29.4	25.8	36.6	30.3	28.5	32.0
Gennaio	30.5	28.4	40.2	37.6	36.5	38.8	34.7	29.5	29.7	37.2	41.5	37.0

Verifica l'ipotesi che la percentuale media di calorie ricavate dai grassi sia la stessa in entrambi i mesi. Usa **(a)** il 5% di significatività e **(b)** l'1% di significatività.

35. Per studiare le abitudini di caccia dei pipistrelli, 22 esemplari sono stati muniti di un segnalatore, e monitorati via radio. Di questi 22 pipistrelli, 12 erano femmine e 10 erano maschi. Nell'esperimento sono state misurate le distanze percorse (in metri) tra un pasto e il successivo, ottenendo i dati riassunti nella tabella seguente,

Pipistrelli femmine	Pipistrelli maschi
$n = 12$	$m = 10$
$\overline{X} = 180$	$\overline{Y} = 136$
$S_x = 92$	$S_y = 86$

Verifica con un livello di significatività del 5%, l'ipotesi che la distanza media percorsa tra i pasti sia la stessa per maschi e femmine.

36. I dati seguenti sono stati ottenuti da una comparazione tra le tracce di piombo contenute nei capelli presi da individui morti tra il 1880 e il 1920 e il contenuto di piombo negli adulti di oggi. I dati sono espressi in microgrammi (10^{-6} g).

	1880–1920	Oggi
Ampiezza del campione	30	100
Media campionaria	48.5	26.6
Deviazione standard campionaria	14.5	12.3

(a) È vero che questi dati provano all'1% di significatività, che il contenuto medio di piombo nei capelli dell'uomo è oggi minore di quanto fosse negli anni tra il 1880 e il 1920? Chiarisci bene quali sono l'ipotesi nulla e quella alternativa.

(b) Qual è il p-dei-dati di questo test?

37. I pesi in libbre per dei campioni di neonati appartenenti a due contee adiacenti nella Western Pennsylvania hanno fornito i seguenti valori:

$n = 53$	$m = 44$
$\overline{X} = 6.9$	$\overline{Y} = 7.2$
$S_x^2 = 5.2$	$S_y^2 = 4.9$

Costruisci un test per verificare l'ipotesi che il peso medio dei neonati delle due contee sia lo stesso. Quanto vale il p-dei-dati?

38. Risolvi nuovamente il Problema 34, questa volta con l'assunzione che le 12 donne monitorate fossero le stesse in entrambi i mesi, e che i due dati di ciascuna colonna si riferiscano alla stessa donna.

39. A 10 donne incinte è stata somministrata una iniezione di pitocina (una forma sintetica dell'ossitocina) per stimolarne il travaglio. Le pressioni sanguigne sistoliche immediatamente prima e dopo la somministrazione sono state:

Paziente	1	2	3	4	5	6	7	8	9	10
Prima	134	122	132	130	128	140	118	127	125	142
Dopo	140	130	135	126	134	138	124	126	132	144

Ti sembra che i dati indichino che l'iniezione provochi un cambiamento della pressione sanguigna?

40. Una questione di interesse medico è se fare jogging porti a un miglioramento della frequenza cardiaca a riposo. Per verificare questa ipotesi, 8 volontari che non hanno mai fatto questo tipo di esercizio fisico hanno accettato di iniziare un programma di un mese di jogging. Alla fine si è stati in grado di confrontare la frequenza cardiaca a riposo prima e dopo il mese di pratica. Se i dati sono quelli riportati qui sotto, possiamo concludere che questo tipo di esercizio abbia modificato la frequenza cardiaca media a riposo?

Soggetto	1	2	3	4	5	6	7	8
Frequenza cardiaca precedente	74	86	98	102	78	84	79	70
Frequenza cardiaca successiva	70	85	90	110	71	80	69	74

41. Sia $X_1, X_2, \ldots, X_n$ un campione proveniente da una popolazione normale di parametri incogniti μ e σ^2. Costruisci un test ad un livello α di significatività per verificare l'ipotesi

$$H_0 : \sigma^2 \leq \sigma_0^2$$

in alternativa a

$$H_1 : \sigma^2 > \sigma_0^2$$

per un valore positivo assegnato σ_0^2.

42. Con riferimento al Problema 41, spiega come andrebbe modificato il procedimento se la media di popolazione μ fosse nota.

43. È stata di recente sviluppata una "pistola" senza aghi che dovrebbe sostituire le siringhe nel somministrare i vaccini. Questo strumento può essere settato per iniettare diverse quantità di liquido, ma a causa delle fluttuazioni casuali, il valore esatto è una variabile aleatoria con media pari al valore stabilito e varianza σ^2 incognita. È stato deciso che lo strumento sarebbe da considerarsi troppo pericoloso se σ fosse più grande di 0.10, e si è misurato un campione di 50 iniezioni, trovando una deviazione standard campionaria di 0.08.

(a) Al 10% di significatività cosa si deciderebbe?

(b) Spiega secondo te quale sarebbe il livello di significatività più indicato per questo problema, e quale dovrebbe essere l'ipotesi nulla.

44. Una casa farmaceutica produce un certo farmaco in dosi che risultano avere una deviazione standard di 0.5 mg. Il settore di ricerca e sviluppo dell'azienda ha sviluppato un nuovo metodo che, con un costo maggiore, dovrebbe permettere di diminuire tale valore. Considerati i pro e i contro si decide di adottare il nuovo metodo solo se vi sarà una forte evidenza che la deviazione standard di una dose sia divenuta minore di 0.4 mg. Se un campione di 10 dosi mostra i pesi seguenti, è il caso di adottare il nuovo metodo?

572.8 572.2 572.7 571.8 572.3 573.1 571.9 572.4 572.6 572.2

45. La produzione di grossi trasformatori elettrici e condensatori richiede l'impiego di sostanze tossiche (le PCB), molto pericolose quando vengono disperse nell'ambiente. Si vogliono confrontare due metodi per misurare il livello di PCB nel pesce di un lago nelle cui prossimità vi è un impianto di grande dimensioni. Si pensa che ciascun metodo produca misurazioni con una sua propria distribuzione normale. Verifica al 10% di significatività l'ipotesi che le relative varianze siano uguali, avendo a disposizione 8 misurazioni eseguite con ciascun metodo su uno stesso pesce:

Metodo 1	6.2	5.8	5.7	6.3	5.9	6.1	6.2	5.7
Metodo 2	6.3	5.7	5.9	6.4	5.8	6.2	6.3	5.5

46. Nel Problema 28, verifica se le popolazioni possono avere la stessa varianza.

47. Siano $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_m$ campioni indipendenti provenienti da due popolazioni normali con varianze σ_x^2 e σ_y^2 rispettivamente. Costruisci un test statistico per verificare l'ipotesi

$$H_0 : \sigma_x^2 \leq \sigma_y^2 \qquad \text{contro} \qquad H_1 : \sigma_x^2 > \sigma_y^2$$

48. Lo spessore del rivestimento superficiale (interno ed esterno) per dei sacchetti di carta cerata, ha distribuzione normale. Vi sono ragioni per credere che vi sia una maggiore variabilità nel rivestimento interno che in quello esterno. Si realizzano 75 osservazioni raccogliendo i dati seguenti (in unità di peso su unità di superficie),

Superficie esterna	Superficie interna
$\bar{x} \approx 0.948$	$\bar{y} \approx 0.652$
$\sum x_i^2 = 91$	$\sum y_i^2 = 82$

Costruisci un test al 5% di significatività per stabilire se la variabilità dello spessore interno possa ritenersi maggiore di quella dello spessore esterno.

49. In un celebre esperimento per determinare l'efficacia dell'acido acetilsalicilico nella prevenzione degli infarti, 22 000 uomini di mezza età vennero divisi casualmente in due gruppi, e fu loro somministrata una dose giornaliera del farmaco o di un placebo. Quando l'esperimento venne concluso erano stati colpiti da infarto 104 uomini nel gruppo principale e 189 in quello di controllo. Usa questi dati per vagliare l'ipotesi che l'assunzione preventiva di questo principio attivo non modifichi la probabilità di subire un infarto.

50. Nello studio svolto nel Problema 49 risultò che 119 uomini del gruppo principale e 98 di quello di controllo subirono un ictus celebrale durante lo stesso periodo. Questi valori sono abbastanza diversi da mostrare un'influenza dell'acido acetilsalicilico sull'occorrenza degli ictus?

51. Tre agenzie di stampa indipendenti stanno conducendo sondaggi per determinare se più della metà della popolazione sia favorevole alla proposta di limitare il traffico veicolare nel centro cittadino. Tutte e tre desiderano scoprire se vi sono elementi a sufficienza per poter dire che i favorevoli sono più della metà. Di conseguenza le ipotesi scelte sono in ogni caso

$$H_0 : p \leq 0.5 \qquad \text{contro} \qquad H_1 : p > 0.5$$

dove p indica la percentuale di cittadinanza favorevole all'iniziativa.

(a) La prima agenzia ottiene 100 risposte, 56 delle quali sono favorevoli. Si può rifiutare al 5% di significatività l'ipotesi che i favorevoli non siano più della metà?

(b) La seconda agenzia ottiene 120 risposte, 68 delle quali sono favorevoli. Si può rifiutare al 5% di significatività l'ipotesi nulla?

(c) La terza agenzia ottiene 110 risposte, 62 delle quali sono favorevoli. Si può rifiutare al 5% di significatività l'ipotesi nulla?

(d) Se le agenzie mettessero in comune i loro dati, avrebbero un campione di 330 intervistati, 186 dei quali favorevoli all'iniziativa. Si potrebbe rifiutare al 5% di significatività l'ipotesi nulla?

52. Secondo dati ufficiali del governo, nel 1990 il 25.5% della popolazione adulta americana era composto da fumatori. Una ricercatrice ha di recente sostenuto che questo dato è in crescita, e per supportare le sue affermazioni ha campionato 500 individui da questa popolazione, scoprendo che tra loro i fumatori erano 138. Si può confermare la sua convinzione al 5% di significatività?

53. Un servizio di ambulanze sostiene che almeno il 45% delle chiamate che riceve riguarda casi di vita o di morte. Per verificare questa ipotesi, si seleziona un campione di 200 chiamate tra quelle in archivio e si trova che 70 di esse riguardavano emergenze possibilmente mortali. Decidi se l'ipotesi fatta è confermata dai dati ad un livello di significatività **(a)** del 5% e **(b)** dell'1%.

54. Si sa che un certo farmaco molto usato è efficace nel 72% dei casi in cui viene impiegato per curare delle infezioni. È stato ora sviluppato un farmaco alternativo che si è mostrato efficace in 42 casi su 50. Questi dati sono abbastanza rilevanti da dimostrare con il 5% di significatività che la nuova sostanza sia più efficace di quella vecchia? Calcola il *p*-dei-dati.

55. Risolvi il Problema 54 con un test basato sull'approssimazione normale della distribuzione binomiale.

56. In un sondaggio effettuato recentemente negli Stati Uniti, 54 intervistati su 200 hanno dichiarato di possedere un'arma da fuoco. In una indagine molto simile condotta in precedenza erano invece stati 30 su 150. È possibile che non vi sia differenza nella percentuale della popolazione che possiede un'arma e che questo risultato sia esclusivamente dovuto ad oscillazioni casuali?

57. Siano X_1 e X_2 due variabili aleatorie binomiali indipendenti, con parametri (n_1, p_1) e (n_2, p_2). Costruisci un test statistico con lo stesso approccio di quello di Fisher-Irwin, per le ipotesi

$$H_0 : p_1 \leq p_2 \qquad \text{contro} \qquad H_1 : p_1 > p_2$$

58. Verifica i passaggi che portano all'Equazione (8.6.8).

59. Siano X_1 e X_2 due variabili aleatorie binomiali indipendenti, con parametri (n_1, p_1) e (n_2, p_2). Mostra che quando n_1 e n_2 sono grandi e si è interessati all'ipotesi nulla $H_0 : p_1 = p_2$, si può ottenere un test approssimato ad un livello di significatività α nel modo seguente:

$$\text{si rifiuta } H_0 \text{ se } \frac{\left|\frac{X_1}{n_1} - \frac{X_2}{n_2}\right|}{\sqrt{\frac{X_1+X_2}{n_1+n_2}\left(1 - \frac{X_1+X_2}{n_1+n_2}\right)\left(\frac{1}{n_1} - \frac{1}{n_2}\right)}} > z_{\frac{\alpha}{2}}$$

Suggerimento:

(a) Chiarisci perché quando n_1 e n_2 sono grandi, la variabile aleatoria

$$\frac{\frac{X_1}{n_1} - \frac{X_2}{n_2} - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

ha distribuzione approssimativamente normale standard.

(b) Mostra che quando H_0 è valida, e quindi $p_1 = p_2$, il loro comune valore può essere stimato in maniera ottimale con la statistica

$$\frac{X_1 + X_2}{n_1 + n_2}$$

60. Risolvi il Problema 56 impiegando il test approssimato del Problema 59.

61. Molti malati di cancro devono confrontarsi con la decisione se ricorrere alla chirurgia o alla radioterapia per curare il loro male. Un'informazione che può aiutarli a decidere è la percentuale di sopravvivenza dopo cinque anni per i due tipi di trattamenti. Si è però scoperto che – sorprendentemente – la decisione presa sembra essere influenzata a seconda se viene comunicata la percentuale di sopravvissuti o quella di deceduti (anche se il significato delle cifre è lo stesso). Ad esempio, in un esperimento un gruppo di 200 malati di tumore alla prostata è stato diviso in due gruppi di 100. Ai primi è stato comunicato che con un intervento chirurgico la percentuale di sopravvivenza dopo cinque anni era del 77%; ai secondi che la percentuale di decessi era del 23%. Le informazioni date sulla radioterapia sono invece state le stesse. Sapendo che 24 pazienti del primo gruppo e 12 del secondo hanno deciso di sottoporsi all'intervento chirurgico, che conclusioni trai?

62. Verifica l'ipotesi che il numero medio di terremoti all'anno su una certa isola sia 52, sapendo che negli ultimi 8 anni ve ne sono stati

46 62 60 58 47 50 59 49

Usa il 5% di significatività, supponi che la distribuzione sia di Poisson e spiega il perché di tale assunzione.

63. La Tabella 2.6 di pagina 44 riporta il numero di incidenti aerei mortali all'anno e il numero delle vittime, per i voli commerciali effettuati negli Stati Uniti nei 16 anni che vanno dal 1980 al 1995. Determina se al 5% di significatività questi dati confutano l'ipotesi che il numero medio di incidenti all'anno sia maggiore o uguale al 4.5%. Qual è il p-dei-dati? (*Suggerimento:* Prima formula un modello per il numero di incidenti all'anno.)

64. I due campioni seguenti provengono da popolazioni di Poisson di media λ_1 e λ_2. Verifica l'ipotesi che $\lambda_1 = \lambda_2$.

Campione 1	24	32	29	33	40	28	34	36
Campione 2	42	36	41					

9 Regressione

Contenuto

9.1 Introduzione

Molti problemi dell'ingegneria e della scienza hanno a che fare con la determinazione delle relazioni tra due o più insiemi di variabili. In un processo chimico, per esempio, è interessante studiare le dipendenze tra la quantità di catalizzatore impiegato, la temperatura e il rendimento. La conoscenza di queste relazioni ci consentirebbe di predire il rendimento per diversi valori della temperatura e della quantità di catalizzatore.

Le situazioni più comuni prevedono una singola variabile Y di *risposta*, e un certo numero di variabili $x_1, x_2, \ldots, x_r$ di *ingresso* (o di *input*). Il modello suppone che la risposta sia in funzione degli ingressi; per questo Y è anche detta *variabile dipendente*, mentre le x_i sono le *variabili indipendenti*. La più semplice relazione che è possibile immaginare è quella lineare; essa si presenta quando per delle opportune costanti $\beta_0, \beta_1, \ldots, \beta_r$ vale l'equazione

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r \tag{9.1.1}$$

Se la relazione che lega le variabili fosse questa sarebbe possibile (una volta scoperte le β_i), predire esattamente la risposta per qualunque combinazione delle variabili di ingresso. In pratica comunque questo livello di precisione non può essere raggiunto, e il massimo che ci si può aspettare è che l'Equazione (9.1.1) sia valida *salvo per un errore casuale*. Con questo intendiamo che la relazione concreta è

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + e \tag{9.1.2}$$

dove e, che rappresenta l'errore casuale, si suppone essere una variabile aleatoria di media nulla. In effetti un secondo modo per esprimere l'Equazione (9.1.2) è il seguente:

$$E[Y|\boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r \tag{9.1.3}$$

dove $\boldsymbol{x} = (x_1, x_2, \ldots, x_r)$ è il vettore delle variabili indipendenti, e $E[Y|\boldsymbol{x}]$ denota il valore atteso della risposta, condizionato all'ingresso $\boldsymbol{x}$.

L'Equazione (9.1.2) è chiamata *equazione di regressione lineare*; diciamo che essa esprime la regressione di Y rispetto alle variabili indipendenti $x_1, x_2, \ldots, x_r$. Le costanti $\beta_0, \beta_1, \ldots, \beta_r$ sono dette *coefficienti di regressione*, e vanno normalmente stimati a partire da un campione di dati. Un'equazione di regressione si dice *semplice* se $r = 1$, e quindi vi è una sola variabile indipendente; negli altri casi si parla di regressione *multipla*.

Un modello lineare semplice presuppone quindi una relazione lineare tra la risposta media e il valore di una singola variabile indipendente x. L'equazione di regressione diviene perciò

$$Y = \alpha + \beta x + e \tag{9.1.4}$$

Esempio 9.1.1. Per $i = 1, 2, \ldots, 10$, consideriamo le 10 coppie di valori (x_i, y_i), che legano y (il rendimento percentuale di un esperimento di laboratorio), a x (la temperatura a cui è stato condotto l'esperimento):

i	1	2	3	4	5	6	7	8	9	10
x_i	100	110	120	130	140	150	160	170	180	190
y_i	45	52	54	63	62	68	75	76	92	88

Quello rappresentato in Figura 9.1 è un *diagramma di dispersione* delle coppie di dati raccolti. In pratica, si tratta di tracciare un segno per ogni coppia, con le due coordinate pari ai valori di x e y rispettivamente (si veda anche quanto detto a proposito di statistica descrittiva nella Sezione 2.6). Poiché il grafico mostra, a meno di errori casuali, una relazione lineare tra y e x, sembra che la scelta di un modello di regressione lineare sia in questo caso appropriata. □

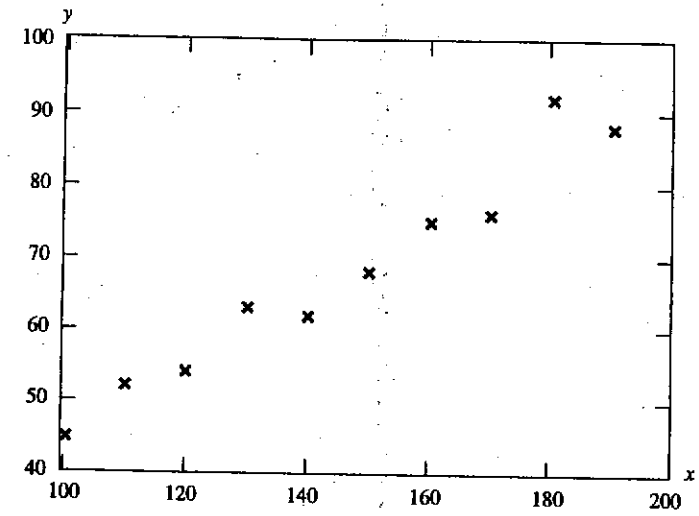


Figura 9.1 Diagramma di dispersione.

9.2 Stima dei parametri di regressione

Supponiamo di osservare, per i che va da 1 a n, le risposte Y_i corrispondenti a certi valori di ingresso x_i, e di volerle usare per stimare α e β in un modello di regressione lineare semplice. Se A e B sono gli stimatori cercati, allora $A + Bx_i$ è lo stimatore della risposta corrispondente all'ingresso x_i. Poiché la risposta realmente ottenuta con quel livello di ingresso è Y_i, la quantità $(Y_i - A - Bx_i)^2$ rappresenta il quadrato della differenza tra predizione e valore osservato, e quindi dovrebbe idealmente essere resa più piccola possibile. Denotiamo con SS la somma dei quadrati degli scarti tra risposte stimate e reali:

$$SS := \sum_{i=1}^{n} (Y_i - A - Bx_i)^2 \tag{9.2.1}$$

Il metodo dei *minimi quadrati* consiste nello scegliere come stimatori di α e β i due valori A e B che minimizzano SS. Per calcolarli, deriviamo SS rispetto ad A e B:

$$\frac{\partial SS}{\partial A} = -2 \sum_{i=1}^{n} (Y_i - A - Bx_i)$$

$$\frac{\partial SS}{\partial B} = -2 \sum_{i=1}^{n} x_i (Y_i - A - Bx_i)$$

Per cercare i punti critici di SS, ed in particolare il minimo, occorre uguagliare a zero le due espressioni, ottenendo il sistema

$$\begin{cases} \sum_{i=1}^{n} Y_i = nA + B\sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i Y_i = A\sum_{i=1}^{n} x_i + B\sum_{i=1}^{n} x_i^2 \end{cases} \tag{9.2.2}$$

Le (9.2.2) sono dette *equazioni normali*. Se si pone

$$\overline{Y} := \frac{1}{n}\sum_i Y_i \qquad \text{e} \qquad \bar{x} := \frac{1}{n}\sum_i x_i$$

la prima equazione normale diventa

$$A = \overline{Y} - B\bar{x}$$

Sostituendo questa formula al posto di A nella seconda otteniamo

$$\sum_i x_i Y_i = (\overline{Y} - B\bar{x})n\bar{x} + B\sum_i x_i^2$$

ovvero

$$B\Big(\sum_i x_i^2 - n\bar{x}^2\Big) = \sum_i x_i Y_i - n\bar{x}\overline{Y}$$

da cui si ricava che

$$B = \frac{\sum_i x_i Y_i - n\bar{x}\overline{Y}}{\sum_i x_i^2 - n\bar{x}^2}$$

Quanto detto prova l'enunciato seguente.

Proposizione 9.2.1. Gli stimatori dei minimi quadrati di β e α corrispondenti alle variabili x_i e Y_i, $i = 1, 2, \ldots, n$ sono rispettivamente,

$$\begin{aligned} B &= \frac{\sum_i x_i Y_i - \bar{x}\sum_i Y_i}{\sum_i x_i^2 - n\bar{x}^2} \\ A &= \overline{Y} - B\bar{x} \end{aligned} \tag{9.2.3}$$

La retta $y = A + Bx$ è la *stima della retta di regressione*, ovvero la retta che interpola[1] meglio i dati. Il Programma 9.2 calcola gli stimatori dei minimi quadrati A e B, e fornisce altre statistiche la cui utilità sarà chiara nelle prossime sezioni.

[1] Un termine di derivazione inglese usato anche in italiano è *fit*, che può essere tradotto con interpolazione. Possiamo dire ad esempio che la retta $y = A + Bx$ è il migliore fit lineare dei dati. [*N.d.T.*]

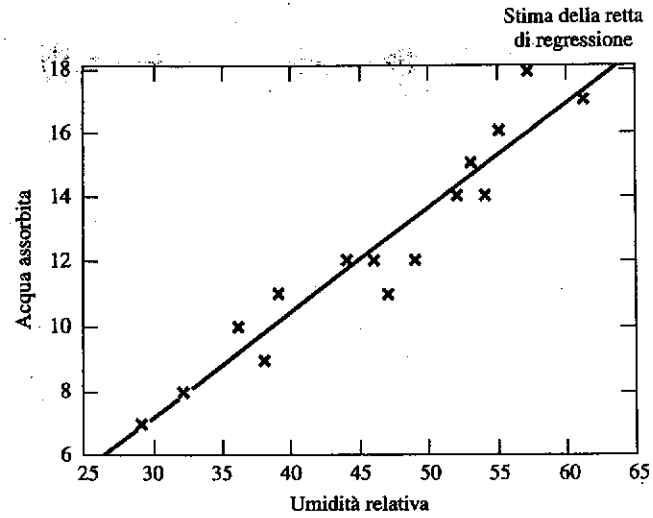


Figura 9.2 Diagramma di dispersione dei dati dell'Esempio 9.2.1.

Esempio 9.2.1. Il materiale grezzo usato per la produzione di una particolare fibra sintetica è immagazzinato in un ambiente che non dispone di controllo dell'umidità. Per 15 giorni vengono prese misurazioni abbinate dell'umidità atmosferica e dell'acqua assorbita dal materiale, ottenendo i risultati seguenti (in punti percentuali),

Umidità atmosferica	46	53	29	61	36	39	47	49	52	38	55	32	57	54	44
Acqua assorbita	12	15	7	17	10	11	11	12	14	9	16	8	18	14	12

Questi dati sono rappresentati nella Figura 9.2. Per calcolare gli stimatori dei minimi quadrati e la stima della retta di regressione utilizziamo il Programma 9.2, ottenendo la schermata che compare in Figura 9.3. □

9.3 Distribuzione degli stimatori

Se fino ad ora è stato sufficiente supporre che gli errori casuali avessero media nulla, per ottenere la distribuzione degli stimatori A e B è necessario fare delle assunzioni ulteriori. Il punto di vista comune è di ipotizzare che essi siano normali indipendenti di media nulla e varianza costante σ^2. Di conseguenza, se per $i = 1, 2, \ldots, n$, Y_i è la risposta data all'ingresso x_i, supporremo che $Y_1, Y_2, \ldots, Y_n$ siano indipendenti e che

$$Y_i \sim \mathcal{N}\left(\alpha + \beta x_i, \sigma^2\right) \tag{9.3.1}$$

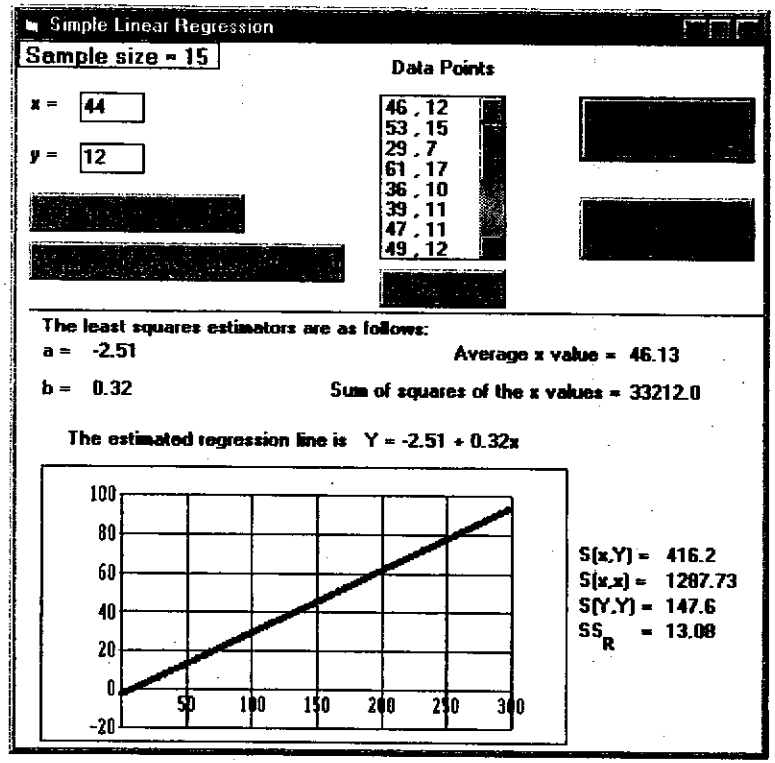


Figura 9.3 Regressione lineare semplice per l'Esempio 9.2.1.

Si noti che stiamo confidando in particolare nel fatto che la varianza dell'errore casuale non dipenda dal livello di ingresso. Il valore di σ^2 non si assume noto e può anzi essere stimato a partire dai dati.

Possiamo riscrivere B, lo stimatore dei minimi quadrati di β, come

$$B = \frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i x_i^2 - n\bar{x}^2} \tag{9.3.2}$$

scoprendo così che esso è in effetti una combinazione lineare delle variabili aleatorie normali e indipendenti $Y_1, Y_2, \ldots, Y_n$, e quindi ha anch'esso distribuzione normale. Ne calcoliamo i parametri.

$$\begin{aligned} E[B] &= \frac{\sum_i (x_i - \bar{x}) E[Y_i]}{\sum_i x_i^2 - n\bar{x}^2} && \text{usando la (9.3.2) e la linearità} \\ &= \frac{\sum_i (x_i - \bar{x})(\alpha + \beta x_i)}{\sum_i x_i^2 - n\bar{x}^2} && \text{per la (9.3.1)} \end{aligned}$$

$$\begin{aligned} &= \frac{\alpha \sum_i (x_i - \bar{x}) + \beta \sum_i x_i (x_i - \bar{x})}{\sum_i x_i^2 - n\bar{x}^2} \\ &= \frac{0 + \beta \sum_i x_i^2 - \beta \bar{x} \sum_i x_i}{\sum_i x_i^2 - n\bar{x}^2} = \beta \qquad \text{perché } \textstyle\sum_i (x_i - \bar{x}) = 0 \end{aligned}$$

Quindi $E[B] = \beta$, e di conseguenza B è uno stimatore non distorto.

$$\begin{aligned} \mathrm{Var}(B) &= \frac{\mathrm{Var}\{\sum_i (x_i - \bar{x}) Y_i\}}{(\sum_i x_i^2 - n\bar{x}^2)^2} && \text{per la (9.3.2)} \\ &= \frac{\sum_i (x_i - \bar{x})^2 \mathrm{Var}(Y_i)}{(\sum_i x_i^2 - n\bar{x}^2)^2} && \text{per l'indipendenza} \\ &= \frac{\sigma^2 \sum_i (x_i - \bar{x})^2}{(\sum_i x_i^2 - n\bar{x}^2)^2} \\ &= \frac{\sigma^2}{\sum_i x_i^2 - n\bar{x}^2} \end{aligned}$$

dove l'ultimo passaggio segue dall'identità

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

che abbiamo usato più volte ed è oggetto della Proposizione 2.3.1 di pagina 26.

Passando alla distribuzione di A, visto che B è una combinazione lineare di $Y_1, Y_2, \ldots, Y_n$ e A si può scrivere come

$$A = \frac{1}{n} \sum_i Y_i - B\bar{x}$$

segue che anche A è una combinazione lineare di variabili aleatorie normali e indipendenti, e quindi ha distribuzione normale. Quali sono i suoi parametri?

$$\begin{aligned} E[A] &= \frac{1}{n} \sum_i E[Y_i] - \bar{x} E[B] \\ &= \frac{1}{n} \sum_i (\alpha + \beta x_i) - \bar{x}\beta \\ &= \alpha + \beta\bar{x} - \bar{x}\beta = \alpha \end{aligned}$$

Perciò anche A è uno stimatore corretto. La varianza può essere ottenuta esprimendo A come combinazione lineare di $Y_1, Y_2, \ldots, Y_n$, applicando le proprietà della varianza. Il risultato (i cui dettagli sono lasciati come esercizio) è che

$$\mathrm{Var}(A) = \frac{\sigma^2 \sum_i x_i^2}{n\left(\sum_i x_i^2 - n\bar{x}^2\right)} \tag{9.3.3}$$

Volgiamo ora la nostra attenzione alle quantità $Y_i - A - Bx_i$, per $i = 1, 2, \ldots, n$, che rappresentano le differenze tra le risposte osservate (le Y_i) e i loro stimatori dei minimi quadrati (ovvero, $A + Bx_i$), e sono chiamate i *residui*. La somma dei quadrati dei residui

$$SS_R := \sum_{i=1}^{n} (Y_i - A - Bx_i)^2 \tag{9.3.4}$$

può essere usata per stimare la varianza degli errori, σ^2. Si può in effetti dimostrare che

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2} \tag{9.3.5}$$

e inoltre SS_R è indipendente da A e B. Il fatto che SS_R/σ^2 abbia distribuzione chi-quadro con $n-2$ gradi di libertà implica tra le altre cose che

$$E\left[\frac{SS_R}{\sigma^2}\right] = n - 2 \qquad \text{e quindi che} \qquad E\left[\frac{SS_R}{n-e}\right] = \sigma^2$$

Così che $SS_R/(n-2)$ è uno stimatore non distorto del parametro incognito σ^2.

Osservazione 9.3.1. Anche se non dimostreremo che SS_R/σ^2 è una chi-quadro con $n-2$ gradi di libertà indipendente da A e B, vogliamo giustificarne brevemente la plausibilità. Siccome le Y_i sono normali indipendenti, si ha che le $(Y_i - E[Y_i])/\sqrt{\mathrm{Var}(Y_i)}$ sono normali *standard* indipendenti, e quindi la somma dei loro quadrati ha distribuzione χ^2_n:

$$\sum_{i=1}^{n} \frac{(Y_i - \alpha - \beta x_i)^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(Y_i - E[Y_i])^2}{\mathrm{Var}(Y_i)} \sim \chi^2_n \tag{9.3.6}$$

Se in tale espressione sostituiamo α e β con i rispettivi stimatori A e B, si ha un risultato analogo a quanto accadeva sostituendo nell'equazione

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2_n$$

lo stimatore $\overline{X}$ al posto di μ. In quel caso si perdeva un grado di libertà, ottenendo che

$$\frac{S^2}{\sigma^2}(n-1) = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

valeva inoltre l'indipendenza di S^2 e $\overline{X}$. Qui SS_R^2/σ^2 si ottiene sostituendo due stimatori nell'Equazione (9.3.6), non stupisce quindi che si perdano *due* gradi di libertà e che SS_R^2, A e B siano indipendenti.

Dovendo trattare con diverse sommatorie, è di grande utilità sviluppare una notazione sintetica. Poniamo allora (Lo studente giustifichi le uguaglianze.)

$$\begin{aligned} S_{xY} &:= \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \overline{Y}) = \sum_{i=1}^{n} x_i Y_i - n\bar{x}\overline{Y} \\ S_{xx} &:= \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \\ S_{YY} &:= \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 \end{aligned} \tag{9.3.7}$$

Gli stimatori dei minimi quadrati possono essere sinteticamente espressi tramite

$$B = \frac{S_{xY}}{S_{xx}} \qquad\qquad A = \overline{Y} - B\bar{x} \tag{9.3.8}$$

Si può ottenere anche una formulazione compatta per SS_R, la somma dei quadrati dei residui. Vale infatti l'equazione:

$$SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}} \tag{9.3.9}$$

La seguente proposizione riassume i risultati della sezione.

Proposizione 9.3.1. Nell'ipotesi che le risposte Y_i, $i = 1, 2, \ldots, n$ siano normali indipendenti con media $\alpha + \beta x_i$ e varianza σ^2, gli stimatori dei minimi quadrati per β e α sono

$$B = \frac{S_{xY}}{S_{xx}} \qquad\qquad A = \overline{Y} - B\bar{x}$$

e hanno distribuzione

$$B \sim \mathcal{N}\left(\beta, \frac{\sigma^2}{S_{xx}}\right) \qquad\qquad A \sim \mathcal{N}\left(\alpha, \frac{\sigma^2 \sum_i x_i^2}{nS_{xx}}\right)$$

Se inoltre denotiamo con

$$SS_R := \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

la somma dei quadrati dei residui, essa può essere calcolata tramite la formula

$$SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

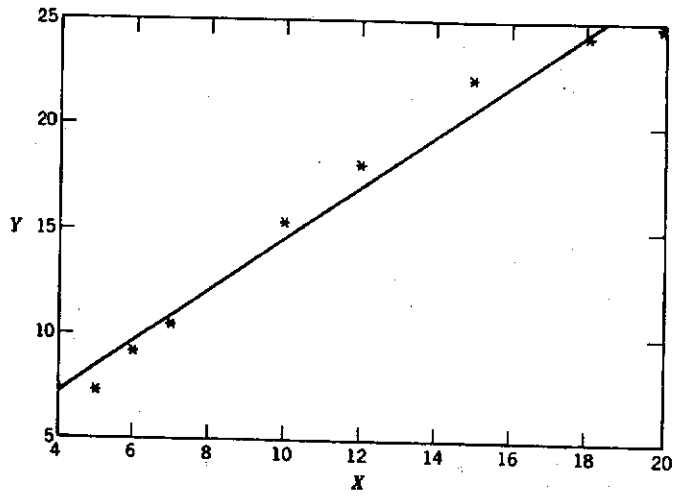


Figura 9.4 Interpolazione lineare dei dati dell'Esempio 9.3.1.

ha distribuzione

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

e infine SS_R, A e B sono indipendenti.

Il Programma 9.2 del software abbinato al libro permette di calcolare A, B come anche $\bar{x}$, $\sum_i x_i^2$, S_{xx}, S_{xY}, S_{YY} e SS_R.

Esempio 9.3.1. I dati seguenti mettono in relazione x, la percentuale d'acqua durante la lavorazione di un certo materiale, con Y, la densità del prodotto finito.

x_i	5	6	7	10	12	15	18	20
y_i	7.4	9.3	10.6	15.4	18.1	22.2	24.1	24.8

Si trovi una retta che interpoli questi dati e si determini il valore di SS_R.

Un grafico dei dati con la stima della retta di regressione compare in Figura 9.4. I coefficienti di quest'ultima sono stati trovati eseguendo il Programma 9.2, che fornisce anche il valore di SS_R. La schermata è riportata in Figura 9.5. □

9.4 Inferenza statistica sui parametri di regressione

Grazie alla Proposizione 9.3.1, costruire test statistici e intervalli di confidenza per i parametri di regressione diventa una questione relativamente semplice.

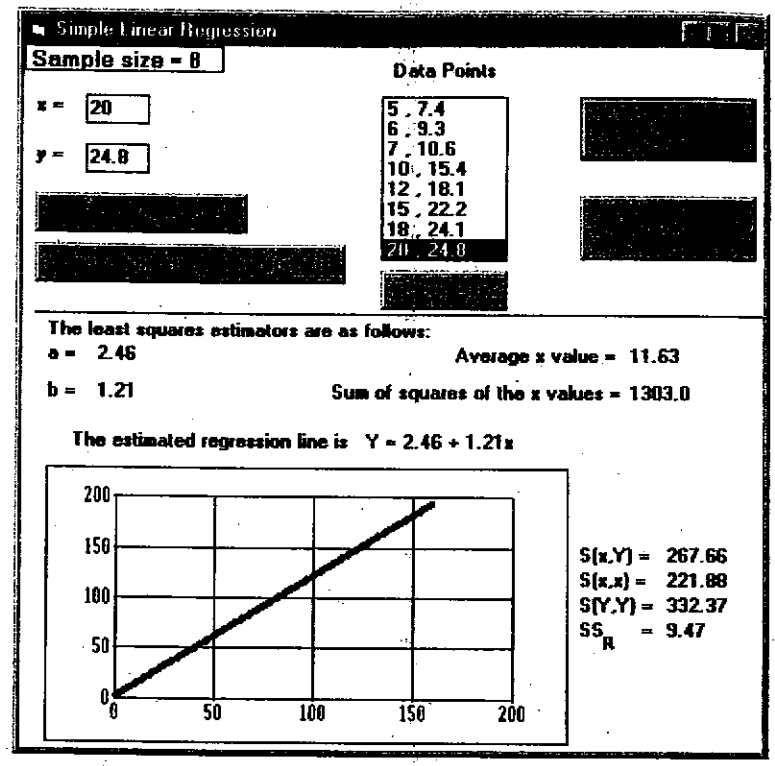


Figura 9.5 Regressione lineare semplice per l'Esempio 9.3.1.

9.4.1 Inferenza su β

Una ipotesi che è molto importante verificare, riguardo il modello di regressione lineare semplice

$$Y = \alpha + \beta X + e$$

è l'ipotesi che β sia pari a zero. Questo ruolo privilegiato è dovuto al fatto che se $\beta = 0$ la risposta non dipende dall'ingresso, ovvero non vi è correlazione tra le due variabili. Per verificare

$$H_0 : \beta = 0 \qquad \text{contro} \qquad H_1 : \beta \neq 0$$

notiamo dalla Proposizione 9.3.1 che

$$\frac{B - E[B]}{\sqrt{\text{Var}(B)}} = \frac{B - \beta}{\sigma/\sqrt{S_{xx}}} \sim \mathcal{N}(0,1) \tag{9.4.1}$$

e inoltre tale variabile aleatoria è indipendente da

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

Perciò dalla definizione di distribuzione t segue che

$$\frac{B-\beta}{\sigma/\sqrt{S_{xx}}}\sqrt{\frac{\sigma^2(n-2)}{SS_R}} = \sqrt{\frac{(n-2)S_{xx}}{SS_R}}(B-\beta) \sim t_{n-2} \qquad (9.4.2)$$

Abbiamo in tal modo individuato una statistica per il test che ci interessa; essa ha distribuzione t con $n-2$ gradi di libertà. Quando l'ipotesi nulla è valida, $\beta = 0$ e quindi

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}B \sim t_{n-2}$$

Questo ci porta a definire la seguente regola che permette di verificare le ipotesi di nostro interesse ad un livello di significatività γ:

$$\begin{aligned} &\text{si rifiuta } H_0 \text{ se } \sqrt{\frac{(n-2)S_{xx}}{SS_R}}|B| > t_{\frac{\gamma}{2},n-2} \\ &\text{si accetta } H_0 \text{ negli altri casi} \end{aligned} \qquad (9.4.3)$$

Si può anche procedere calcolando il valore v assunto da $\sqrt{(n-2)S_{xx}/SS_R}|B|$, e rifiutando quindi H_0 se il livello di significatività è maggiore o uguale a

$$\begin{aligned} p\text{-dei-dati} &= P(|T_{n-2}| > v) \\ &= 2P(T_{n-2} > v) \end{aligned} \qquad (9.4.4)$$

dove T_{n-2} ha distribuzione t con $n-2$ gradi di libertà. Questa probabilità può essere ottenuta impiegando il Programma 5.8.2a del software del libro.

Esempio 9.4.1. Un tale è convinto che il consumo di carburante della sua vettura non dipenda dalla velocità di guida, ma solo dalla distanza percorsa. Per verificare se questa ipotesi sia plausibile, si misurano i consumi dell'automobile a diverse velocità tra le 45 e le 70 miglia orarie. Le miglia percorse con un gallone di carburante sono state le seguenti,

Velocità	45	50	55	60	65	70	75
Miglia con un gallone	24.2	25.0	23.3	22.0	21.5	20.6	19.8

Questi dati confermano l'idea che la velocità non influenzi il consumo di carburante?

Supponendo che un modello di regressione lineare semplice

$$Y = \alpha + \beta x + e$$

leghi le miglia Y, percorse con un gallone di carburante, alla velocità di percorrenza x, l'ipotesi fatta che x e Y non siano legate è equivalente a dire che $\beta = 0$. Per stabilire se i dati sono abbastanza forti da negare questa ipotesi, occorre sceglierla come ipotesi nulla. Verifichiamo perciò

$$H_0 : \beta = 0 \qquad \text{contro} \qquad H_1 : \beta \neq 0$$

Per valutare la statistica del test, calcoliamo S_{xx}, S_{YY} e S_{xY}. Un rapido conto manuale stabilisce che

$$S_{xx} = 700, \qquad S_{YY} \approx 21.757, \qquad S_{xY} = -119$$

Il valore di SS_R può essere determinato usando l'Equazione (9.3.9),

$$SS_R \approx \frac{700 \times 21.757 - 119^2}{700} \approx 1.527$$

mentre per B, si trova

$$B = S_{xY}/S_{xx} = -119/700 = -0.17$$

in modo tale che il valore della statistica di questo test è

$$|-0.17|\sqrt{5 \times 700/1.527} \approx 8.139$$

Dalla Tabella A.3 dell'Appendice, si ricava che $t_{0.005,5} \approx 4.032$, quindi l'ipotesi nulla va rifiutata all'1% significatività. Concludendo, l'affermazione che i consumi della vettura non dipendano dalla velocità è decisamente confutata dai dati, vi sono anzi prove a sufficienza per stabilire che i consumi aumentano con la velocità. □

Dall'Equazione (9.4.2) si possono anche ricavare gli intervalli di confidenza per β. Infatti, per ogni γ appartenente all'intervallo $(0,1)$, si ha che

$$P\left(-t_{\frac{\gamma}{2},n-2} < \sqrt{\frac{(n-2)S_{xx}}{SS_R}}(B-\beta) < t_{\frac{\gamma}{2},n-2}\right) = 1-\gamma$$

o equivalentemente,

$$P\left(B - t_{\frac{\gamma}{2},n-2}\cdot\sqrt{\frac{SS_R}{(n-2)S_{xx}}} < \beta < B + t_{\frac{\gamma}{2},n-2}\cdot\sqrt{\frac{SS_R}{(n-2)S_{xx}}}\right) = 1-\gamma$$

in tal modo un intervallo che contiene β con livello di confidenza $1-\gamma$ è dato da

$$\left(B - t_{\frac{\gamma}{2},n-2}\cdot\sqrt{\frac{SS_R}{(n-2)S_{xx}}},\quad B + t_{\frac{\gamma}{2},n-2}\cdot\sqrt{\frac{SS_R}{(n-2)S_{xx}}}\right) \qquad (9.4.5)$$

Osservazione 9.4.1. È bene notare che anche se abbiamo dimostrato che

$$\frac{B-\beta}{\sigma/\sqrt{S_{xx}}} \sim \mathcal{N}(0,1)$$

questo risultato non può essere utilizzato direttamente per fare dell'inferenza su β, in quanto la statistica dipende dal parametro incognito σ^2. Sostituendo σ^2 con il suo stimatore $SS_R/(n-2)$, la distribuzione della statistica passa da normale standard a t con $n-2$ gradi di libertà.

Esempio 9.4.2. Con riferimento all'Esempio 9.4.1, si calcoli un intervallo di confidenza al 95% per il parametro β.

Siccome $t_{0.025,5} \approx 2.571$, si deduce dai calcoli fatti in quell'esempio che l'intervallo cercato è dato da

$$-0.170 \pm 2.571\sqrt{\frac{1.527}{3500}} \approx -0.170 \pm 0.054$$

E quindi abbiamo il 95% di confidenza che β sia compreso fra -0.224 e -0.116. □

9.4.1.1 Regressione alla media

Il termine *regressione* fu adoperato per la prima volta da Francis Galton a proposito delle leggi dell'ereditarietà. Galton pensava che tali leggi prescrivessero per la progenie degli estremi della popolazione una "regressione verso la media", intendendo con questo che i figli di individui con caratteristiche eccezionalmente alte o basse tendono ad essere più nella media dei loro genitori.

Se assumiamo che vi sia una relazione lineare tra il valore della caratteristica in esame per il figlio (Y) e per il genitore (x), si avrà una regressione verso la media ogni volta che che il parametro β è compreso tra 0 e 1. Ovvero, se

$$E[Y] = \alpha + \beta x$$

e $0 < \beta < 1$, allora $E[Y]$ sarà più piccolo di x quando x è molto grande e più grande di x quando x è molto piccolo. Ci si può convincere di questo fatto dimostrandolo algebricamente (guidati dal Problema 13), oppure anche disegnando i grafici delle due rette

$$y = \alpha + \beta x \qquad \text{e} \qquad y = x$$

che mostrano chiaramente come la prima stia sopra la seconda per valori piccoli di x mentre accade il contrario per valori grandi di x.

Esempio 9.4.3. Per dimostrare la tesi di Galton sulla regressione verso la media dei caratteri ereditari, lo statistico britannico Karl Pearson confrontò le stature di 10 figli maschi scelti a caso con quelle dei loro padri. I dati ottenuti (in pollici) furono i seguenti.

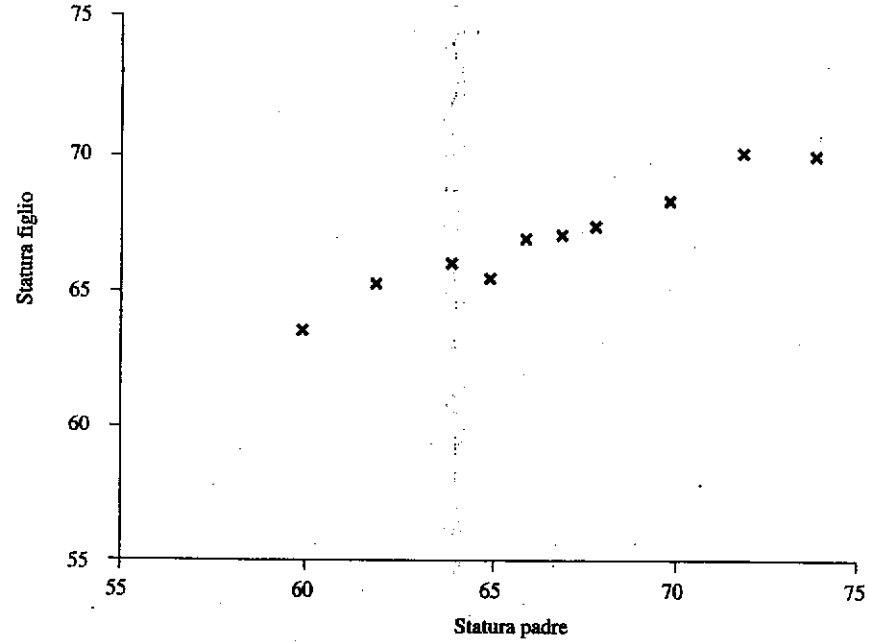


Figura 9.6 Diagramma di dispersione delle stature dei figli rispetto a quelle dei padri.

Padre	60	62	64	65	66	67	68	70	72	74
Figlio	63.6	65.2	66	65.5	66.9	67.1	67.4	68.3	70.1	70

La Figura 9.6 mostra un diagramma di dispersione per questi dati. Si noti che anche se il grafico mostra che padri alti tendono ad avere figli alti, sembra anche indicare come i figli di padri estremamente alti o bassi tendano a essere più "nella media" dei loro genitori, sembra quindi esserci davvero una "regressione verso la media".

Se questo sia confermato anche *quantitativamente* dai dati sarà chiaro verificando

$$H_0 : \beta \geq 1 \qquad \text{contro} \qquad H_1 : \beta < 1$$

o, in maniera equivalente,

$$H_0 : \beta = 1 \qquad \text{contro} \qquad H_1 : \beta < 1$$

Procediamo come in precedenza notando che per l'Equazione (9.4.2), quando $\beta = 1$, la statistica del test, che denotiamo con D_{ts},

$$D_{ts} := \sqrt{8S_{xx}/SS_R}(B-1)$$

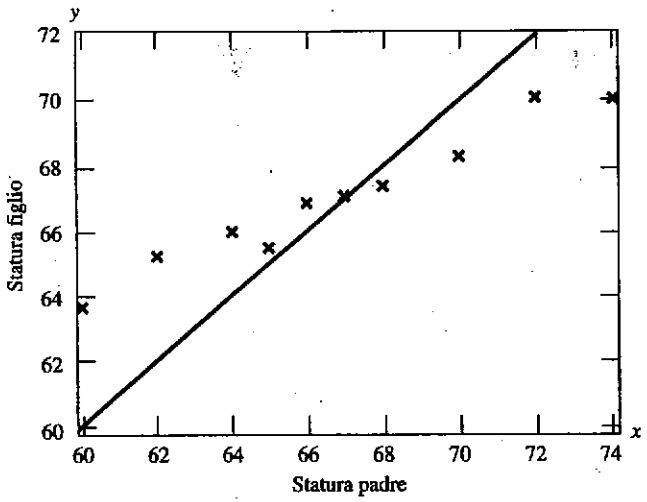


Figura 9.7 La regressione verso la media di Galton. Per x piccolo, $y > x$. Per x grande, $y < x$.

ha distribuzione t con 8 gradi di libertà. Fissato perciò un livello γ di significatività, il test dovrebbe rifiutare H_0 quando il valore di D_{ts} è abbastanza piccolo (infatti ciò si verifica quando B, lo stimatore di β, è sufficientemente minore di 1). In particolare, rifiuteremo l'ipotesi nulla se

$$D_{\text{ts}} < -t_{\gamma,8}$$

Il Programma 9.2 fornisce i seguenti valori,

$$\sqrt{8S_{xx}/SS_{\text{R}}}(B-1) \approx 30.3 \times (0.46-1) \approx -16.4$$

Siccome $t_{0.01,8} \approx 2.896$, otteniamo subito che $D_{\text{ts}} < -t_{0.01,8}$ e quindi l'ipotesi nulla che β fosse maggiore o uguale a 1 viene rifiutata con l'1% di significatività. In effetti, il p-dei-dati è circa nullo:

$$p\text{-dei-dati} \approx P(T_8 \leq -16.4) \approx 0$$

per cui H_0 va rifiutata ad ogni livello di significatività ragionevole, provando così che la regressione verso la media è un fenomeno reale (si veda la Figura 9.7).

Una giustificazione biologica moderna del fenomeno della regressione alla media dovrebbe basarsi sul fatto che ogni figlio ottiene una selezione casuale di metà dei geni di ciascuno dei genitori; banalizzando un poco potremmo dire che in questo modo, il figlio di un individuo molto alto avrà tipicamente meno geni "della statura" di suo padre. □

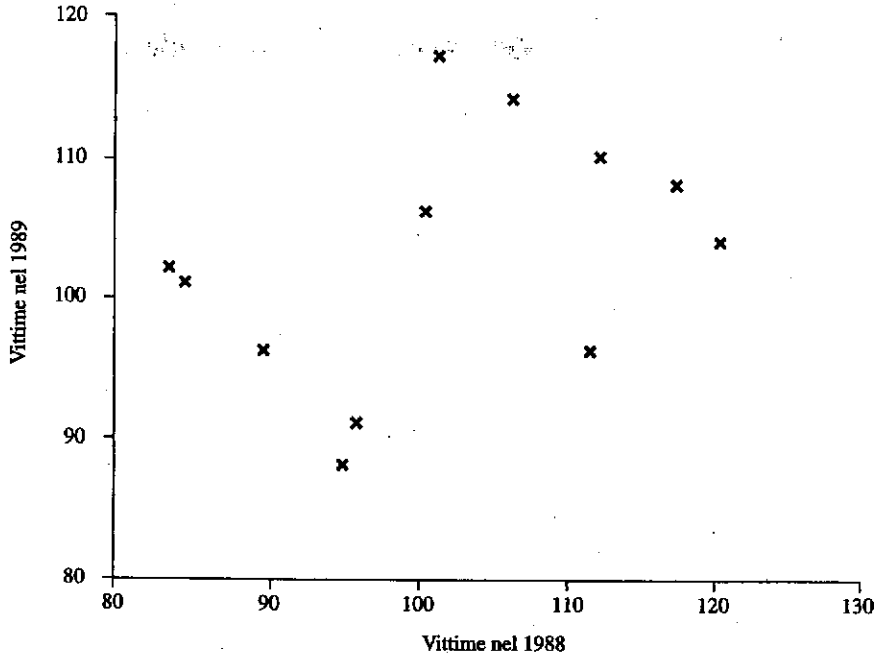


Figura 9.8 Diagramma di dispersione delle vittime nel 1989 rispetto a quelle nel 1988.

Anche se il principale campo di applicazione della regressione alla media è sicuramente quello biologico, e in particolare nell'ambito della relazione tra le caratteristiche mostrate da genitori e figli, questo fenomeno compare anche in altre situazioni, in particolare quando abbiamo due insiemi di dati che si riferiscono alle stesse variabili.

Esempio 9.4.4. I dati della tabella seguente mostrano il numero di vittime di incidenti stradali in 12 contee degli Stati Uniti nordoccidentali, per gli anni 1988 e 1989.

Contea	1	2	3	4	5	6	7	8	9	10	11	12
Vittime nel 1988	121	96	85	113	102	118	90	84	107	112	95	101
Vittime nel 1989	106	91	101	110	117	108	96	102	114	96	88	106

Un'occhiata alla Figura 9.8 indica che nel 1989 vi fu nella gran parte dei casi una riduzione nel numero di vittime per le contee che ne ebbero molte nel 1988, e un aumento in quelle che ne avevano avute di meno. Per verificare se sia in atto un fenomeno di regressione alla media, eseguiamo il Programma 9.2 ottendo l'equazione di regressione stimata

$$y = 74.59 + 0.28x$$

la quale mostra un valore stimato per β che è effettivamente molto minore di 1.

Occorre essere prudenti nel considerare la ragioni che stanno dietro al fenomeno di regressione in questo caso. Certamente sembra naturale immaginare che le contee che ebbero un elevato numero di incidenti nel 1988 siano corse ai ripari con miglioramenti nella sicurezza delle strade e campagne di sensibilizzazione ai pericoli di una guida imprudente. Si può pure ipotizzare che le contee che avevano avuto pochi incidenti si siano "adagiate sugli allori" e non si siano sforzate attivamente di tenere basso il numero di vittime, ottenendone anzi un certo aumento nell'anno seguente.

Anche se è del tutto possibile che le ragioni espresse siano corrette e che abbiano giocato un ruolo nei dati in nostro possesso, è importante rendersi conto che si sarebbe probabilmente notata una regressione verso la media anche se nessuna delle contee avesse fatto niente di particolare. Infatti può accadere che le contee che ebbero un elevato numero di vittime nel 1988, attraversassero semplicemente un anno sfortunato. In questo caso una diminuzione per il 1989 indicherebbe solo che vi fu un ritorno ad un risultato più normale. (Per avere una analogia, si pensi di avere ottenuto 9 teste lanciando 10 volte una moneta. Se si effettuano altri 10 lanci, è piuttosto probabile che il numero di teste sia inferiore.) Analogamente, le contee che nel 1988 ebbero poche vittime potrebbero essere state "fortunate", e quindi un valore nella media nel 1989 sarebbe risultato in un aumento rispetto all'anno precedente.

L'errata convinzione che la regressione alla media sia sempre dovuta a qualche fattore esterno quando in realtà è spesso opera del "caso", si incontra abbastanza spesso che è sembrato opportuno darle un nome: viene detta *regression fallacy*. □

9.4.2 Inferenza su α

La determinazione degli intervalli di confidenza e dei test statistici che riguardano il parametro α si ottiene in modo analogo a quanto fatto per β. In particolare si può usare la Proposizione 9.3.1 per mostrare che

$$\sqrt{\frac{n(n-2)S_{xx}}{SS_{\mathrm{R}} \cdot \sum_i x_i^2}}(A-\alpha) \sim t_{n-2} \tag{9.4.6}$$

di conseguenza, ad un livello di $1-\gamma$, l'intervallo di confidenza bilaterale è dato da

$$A \pm t_{\frac{\gamma}{2},n-2} \cdot \sqrt{\frac{SS_{\mathrm{R}} \cdot \sum_i x_i^2}{n(n-2)S_{xx}}} \tag{9.4.7}$$

I test statistici che riguardano α si ottengono facilmente a partire dall'Equazione (9.4.6) e la loro costruzione è lasciata come esercizio.

9.4.3 Inferenza sulla risposta media $\alpha + \beta X$

Una questione certamente interessante è l'utilizzo delle coppie di dati (x_i, Y_i), $i = 1, 2, \ldots, n$ per stimare $\alpha + \beta x_0$, vale a dire la risposta media per un livello di ingresso assegnato x_0. Se si desidera uno stimatore puntuale, la scelta naturale è $A + Bx_0$ che è uno stimatore non distorto, visto che A e B lo sono entrambi:

$$E[A + Bx_0] = E[A] + x_0 E[B] = \alpha + \beta x_0$$

Se invece vogliamo ottenere degli intervalli di confidenza, oppure verificare delle ipotesi sulla risposta media, è necessario prima determinare la distribuzione dello stimatore $A + Bx_0$. Procediamo.

Usando l'espressione per B data dall'Equazione (9.3.2), si ha che

$$B = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{S_{xx}} Y_i$$

dove si è usato che $S_{xx} = \sum_i x_i^2 - n\bar{x}^2$. Siccome poi

$$A = \overline{Y} - B\bar{x}$$

si può scrivere $A + Bx_0$ come combinazione lineare di $Y_1, Y_2, \ldots, Y_n$:

$$\begin{aligned} A + Bx_0 &= \overline{Y} - B(\bar{x} - x_0) \\ &= \sum_{i=1}^{n} \frac{1}{n} Y_i - \sum_{i=1}^{n} \frac{(x_i - \bar{x})(\bar{x} - x_0)}{S_{xx}} Y_i \\ &= \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(x_i - \bar{x})(\bar{x} - x_0)}{S_{xx}} \right] Y_i \end{aligned}$$

Poiché $Y_1, Y_2, \ldots, Y_n$ sono variabili aleatorie normali indipendenti, anche ogni loro combinazione lineare – e in particolare $A + Bx_0$ – ha distribuzione normale. Per determinare la legge esatta ci servono la media (che conosciamo già) e la varianza, che è data da

$$\begin{aligned} \mathrm{Var}(A + Bx_0) &= \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(x_i - \bar{x})(\bar{x} - x_0)}{S_{xx}} \right]^2 \mathrm{Var}(Y_i) \\ &= \sigma^2 \sum_{i=1}^{n} \left[\frac{1}{n^2} - \frac{2(x_i - \bar{x})(\bar{x} - x_0)}{nS_{xx}} + \frac{(x_i - \bar{x})^2 (\bar{x} - x_0)^2}{S_{xx}^2} \right] \\ &= \sigma^2 \left[\frac{1}{n} - \frac{2(\bar{x} - x_0)}{nS_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) + \frac{(\bar{x} - x_0)^2}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right] \end{aligned}$$

$$= \sigma^2\left[\frac{1}{n} - 0 + \frac{(\bar{x}-x_0)^2}{S_{xx}^2}S_{xx}\right] \qquad \text{perché } \sum_{i=1}^n (x_i - \bar{x}) = 0$$
$$= \sigma^2\left[\frac{1}{n} + \frac{(\bar{x}-x_0)^2}{S_{xx}}\right]$$

dove si è usata la definizione di S_{xx} e il fatto che $\sum_i (x_i - \bar{x}) = n\bar{x} - n\bar{x} = 0$. Abbiamo in tal modo dimostrato che

$$A + Bx_0 \sim \mathcal{N}\left(\alpha + \beta x_0, \quad \sigma^2\left[\frac{1}{n} + \frac{(\bar{x}-x_0)^2}{S_{xx}}\right]\right) \tag{9.4.8}$$

Non possiamo usare direttamente questa statistica per fare dell'inferenza perché σ^2 è incognita. Notiamo però che $A + Bx_0$ è indipendente da

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

per cui

$$\frac{A + Bx_0 - (\alpha + \beta x_0)}{\sqrt{\frac{1}{n} + \frac{(\bar{x}-x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}} \sim t_{n-2} \tag{9.4.9}$$

Usando l'Equazione precedente è immediato ricavare gli intervalli di confidenza per $\alpha + \beta x_0$. Se $1 - \gamma$ è il livello di confidenza richiesto, si ottiene,

$$A + Bx_0 \pm t_{\frac{\gamma}{2},n-2} \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x}-x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}} \tag{9.4.10}$$

Esempio 9.4.5. Usando i dati dell'Esempio 9.4.3, si determi un intervallo che contenga con il 95% di confidenza la statura media di tutti i maschi il cui padre è alto 68 pollici.

I dati che ci servono sono

$$n = 10, \qquad x_0 = 68, \qquad \bar{x} = 66.8, \qquad S_{xx} = 171.6, \qquad SS_R \approx 1.49$$

Si ha quindi che

$$\sqrt{\frac{1}{n} + \frac{(\bar{x}-x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}} \approx 0.142$$

Poiché inoltre

$$t_{0.025,8} \approx 2.306, \qquad A + Bx_0 \approx 67.3$$

L'intervallo di confidenza cercato è

$$\alpha + \beta x_0 \in (66.9,\ 67.6) \quad \square$$

9.4.4 Intervallo di predizione di una risposta futura

In alcuni casi, è più importante stimare il valore che sarà assunto da una risposta futura che non il suo valore medio (come ci si aspetta e come mostreremo, la differenza sta nelle stime tramite intervalli e non in quelle puntuali). Ad esempio volendo realizzare un procedimento chimico ad una temperatura assegnata x_0, saremmo più interessati a predire $Y(x_0)$, il rendimento di questo esperimento, che non il rendimento medio $E[Y(x_0)] = \alpha + \beta x_0$. Al contrario il rendimento medio potrebbe essere più interessante se si dovessero realizzare una *serie* di esperimenti alla stessa temperatura x_0.

Per prima cosa consideriamo cerchiamo un valore *singolo* (analogo a uno stimatore puntuale) che predica la risposta $Y(x_0)$ che si ottiene con un livello di ingresso x_0. Il migliore predittore per $Y(x_0)$ è il suo valore medio[2] $\alpha + \beta x_0$. Siccome α e β sono incognite, il predittore puntuale appropriato sarà $A + Bx_0$.

Immaginiamo ora di volere non una stima puntuale, ma un intervallo di valori che conterrà la risposta con un certo livello di confidenza. Denotiamo semplicemente con Y la risposta futura con un livello di ingresso x_0, e consideriamo la distribuzione di probabilità di $Y - A - Bx_0$, cioè la differenza tra risposta e valore predetto. Sappiamo per ipotesi che

$$Y \sim \mathcal{N}(\alpha + \beta x_0, \sigma^2)$$

Sappiamo inoltre dalla Sezione 9.4.3 che

$$A + Bx_0 \sim \mathcal{N}\left(\alpha + \beta x_0, \quad \sigma^2\left[\frac{1}{n} + \frac{(\bar{x}-x_0)^2}{S_{xx}}\right]\right)$$

Y è indipendente da $Y_1, Y_2, \ldots, Y_n$, e quindi anche da $A + Bx_0$, che è una loro combinazione lineare. Di conseguenza

$$Y - A - Bx_0 \sim \mathcal{N}\left(0,\ \sigma^2\left[1 + \frac{1}{n} + \frac{(\bar{x}-x_0)^2}{S_{xx}}\right]\right)$$

o, equivalentemente,

$$\frac{Y - A - Bx_0}{\sigma\sqrt{\frac{n+1}{n} + \frac{(\bar{x}-x_0)^2}{S_{xx}}}} \sim \mathcal{N}(0,1) \tag{9.4.11}$$

[2] In realtà si potrebbe obiettare che il miglior predittore di una variabile aleatoria può essere (1) la sua media – che minimizza il valore atteso del quadrato della differenza tra predizione e osservazione (si veda l'Osservazione 4.5.1 a pagina 122); o (2) la sua mediana – che minimizza la media del valore assoluto della differenza tra predizione e osservazione (si veda il Problema 35 a pagina 139); o (3) la sua moda – che rappresenta il valore che ha più possibilità di essere osservato. Siccome stiamo supponendo che la risposta abbia distribuzione normale, e per tali variabili aleatorie, media, mediana e moda coincidono, il problema in questo caso non si pone.

Usando adesso il fatto che SS_R è indipendente da A e B, come pure da Y, e che

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

otteniamo, sostituendo σ con il suo stimatore, che

$$\frac{Y - A - Bx_0}{\sqrt{\frac{n+1}{n} + \frac{(\bar{x}-x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}} \sim t_{n-2} \tag{9.4.12}$$

e quindi per ogni valore γ, $0 < \gamma < 1$, si ha che

$$P\left(-t_{\frac{\gamma}{2},n-2} < \frac{Y - A - Bx_0}{\sqrt{\frac{n+1}{n} + \frac{(\bar{x}-x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}} < t_{\frac{\gamma}{2},n-2}\right) = 1 - \gamma$$

Abbiamo in tal modo dimostrato che se ci si basa sull'osservazione delle risposte Y_i corrispondenti ai livelli di ingresso x_i, con $i = 1, 2, \ldots, n$; allora la risposta Y ad un livello di ingresso x_0 apparterrà con un livello di confidenza di $1 - \gamma$ all'intervallo

$$A + Bx_0 \pm t_{\frac{\gamma}{2},n-2} \cdot \sqrt{\left[1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right]\frac{SS_R}{n-2}} \tag{9.4.13}$$

Esempio 9.4.6. Con riferimento all'Esempio 9.4.3, supponiamo di volere trovare un intervallo per il quale abbiamo il 95% di fiducia che conterrà la statura di un maschio adulto il cui padre sia alto 68 pollici. Un veloce calcolo fornisce l'intervallo di predizione

$$Y(68) \in 67.3 \pm 1.0$$

quindi con il 95% di confidenza, l'altezza della persona in questione sarà compresa tra 66.3 e 68.3 □

Osservazione 9.4.2. Si fa spesso un po' di confusione tra intervalli di confidenza e di predizione. Un intervallo di confidenza contiene con un certo livello di confidenza un *parametro* di interesse. Un intervallo di predizione invece, contiene con un certo livello di confidenza il valore di una *variabile aleatoria*.

Osservazione 9.4.3. Non si dovrebbero fare predizioni su una risposta che corrisponde a un livello di ingresso distante da quelli usati per ottenere la retta di regressione stimata. Non ha ad esempio alcun senso usare i dati dell'Esempio 9.4.3 per predire l'altezza di un maschio il cui padre è alto 42 pollici (circa 105 cm).

9.4.5 Sommario dei risultati

Riassumiamo qui di seguito le distribuzioni ottenute nella sezione.

$$\textit{modello: } Y = \alpha + \beta x + e, \qquad e \sim \mathcal{N}(0, \sigma^2)$$
$$\textit{dati: } (x_i, Y_i), \qquad i = 1, 2, \ldots, n$$

Inferenze su	Risultato da utilizzare
β	$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}(B - \beta) \sim t_{n-2}$
α	$\sqrt{\frac{n(n-2)S_{xx}}{SS_R \cdot \sum_i x_i^2}}(A - \alpha) \sim t_{n-2}$
$\alpha + \beta x_0$	$\frac{A + Bx_0 - (\alpha + \beta x_0)}{\sqrt{\frac{1}{n} + \frac{(\bar{x}-x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}} \sim t_{n-2}$
$Y(x_0)$	$\frac{Y - A - Bx_0}{\sqrt{\frac{n+1}{n} + \frac{(\bar{x}-x_0)^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}}} \sim t_{n-2}$

9.5 Coefficiente di determinazione e coefficiente di correlazione campionaria

Supponiamo di volere esprimere la variabilità o dispersione dell'insieme di risposte $Y_1, Y_2, \ldots, Y_n$, ottenute con livelli di ingresso $x_1, x_2, \ldots, x_n$. Una comune misura statistica della variabilità[3] è costituita da

$$S_{YY} := \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{9.5.1}$$

una quantità che rappresenta, a meno di un fattore moltiplicativo, la varianza campionaria delle Y_i. Se esse fossero ad esempio tutte uguali tra loro – e quindi tutte uguali a $\overline{Y}$– il valore di S_{YY} sarebbe pari a 0.

La variabilità nei valori delle Y_i viene però da due contributi. Per prima cosa, se le x_i non sono tutte uguali, le Y_i hanno valori attesi diversi, e questo disperderà le loro realizzazioni. Secondariamente, una volta che si tenga conto della variabilità delle x_i, ogni Y_i ha distribuzione con varianza σ^2 attorno al suo valore atteso e non coinciderà quindi esattamente con le nostre predizioni.

[3] La somma di quadrati che segue, in alcuni contesti prende il nome di *devianza* dei dati (si veda anche la nota a pagina 417). [*N.d.T.*]

Cerchiamo di quantificare quale parte della variabilità delle Y_i sia dovuta ai diversi livelli di ingresso e quale alla varianza propria delle risposte una volta che si tenga conto del valore degli ingressi. Notiamo che la quantità

$$SS_R := \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

misura quella parte di variabilità intrinseca nelle risposte quando si tenga conto delle x_i. Di conseguenza

$$S_{YY} - SS_R$$

rappresenta l'altra parte, cioè quella che si spiega con la diversità dei livelli di ingresso. La statistica R^2, definita da

$$R^2 := \frac{S_{YY} - SS_R}{S_{YY}} = 1 - \frac{SS_R}{S_{YY}} \tag{9.5.2}$$

è la frazione della variabilità totale che è giustificata dalla diversità dei livelli di ingresso, e prende il nome di *coefficiente di determinazione*.

Questo coefficiente è sempre compreso tra 0 e 1; valori di R^2 prossimi a 1 indicano che la gran parte della variazione nei dati delle risposte si spiega con la dispersione dei livelli di ingresso, mentre quando R^2 è prossimo a zero è vero il contrario.

Esempio 9.5.1. Nell'Esempio 9.4.3, l'output del Programma 9.2 aveva fornito i valori seguenti,

$$S_{YY} \approx 38.53, \qquad SS_R \approx 1.49$$

e quindi

$$R^2 \approx 1 - \frac{1.49}{38.53} = 0.961$$

In altri termini, il 96% circa della variabilità delle altezze dei 10 soggetti si spiega con le altezze dei loro padri. Il restante 4% (non giustificato) è dovuto alla varianza propria nella statura dei figli quando anche si sappia quella dei padri. (È quindi dovuta a σ^2, la varianza dell'errore casuale.) □

Il valore di R^2 è spesso usato come un indicatore di quanto quanto bene il modello di regressione interpreti i dati, con valori vicini a 1 che indicano una buona aderenza, e valori prossimi a 0 che indicano una cattiva aderenza. In altri termini il modello di regressione viene considerato interpretare bene i dati se riesce a spiegare la maggior parte della variabilità nelle risposte.

Ricordiamo che nella Sezione 2.6 avevamo definito il coefficiente di correlazione campionaria r, di un insieme di coppie di dati (x_i, Y_i), per $i = 1, 2, \ldots, n$. La sua espressione è la seguente:

$$r := \frac{\sum_i (x_i - \bar{x})(Y_i - \overline{Y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (Y_i - \overline{Y})^2}} \tag{9.5.3}$$

Avevamo in quella sede notato che r fornisce una misura del grado di corrispondenza tra i valori estremi di x e quelli di Y. In particolare un valore prossimo a $+1$ indica che valori elevati di x sono fortemente associati a valori grandi di Y e similmente valori piccoli con valori piccoli; viceversa un valore prossimo a -1 indica che vi è corrispondenza tra valori grandi di x e piccoli di Y nonché tra valori piccoli di x e grandi di Y.

Con la notazione di questo capitolo possiamo scrivere che

$$r = \frac{S_{xY}}{\sqrt{S_{xx} S_{YY}}}$$

e usando l'identità dell'Equazione (9.3.9),

$$SS_R = \frac{S_{xx} S_{YY} - S_{xY}^2}{S_{xx}}$$

otteniamo che

$$\begin{aligned} r^2 &= \frac{S_{xY}^2}{S_{xx} S_{YY}} \\ &= \frac{S_{xx} S_{YY} - SS_R S_{xx}}{S_{xx} S_{YY}} \\ &= 1 - \frac{SS_R}{S_{YY}} = R^2 \end{aligned}$$

Quindi

$$|r| = \sqrt{R^2} \tag{9.5.4}$$

e così, eccetto al più per il segno, il coefficiente di correlazione lineare è uguale alla radice quadrata del coefficiente di determinazione. Il segno di r coincide con quello di B.

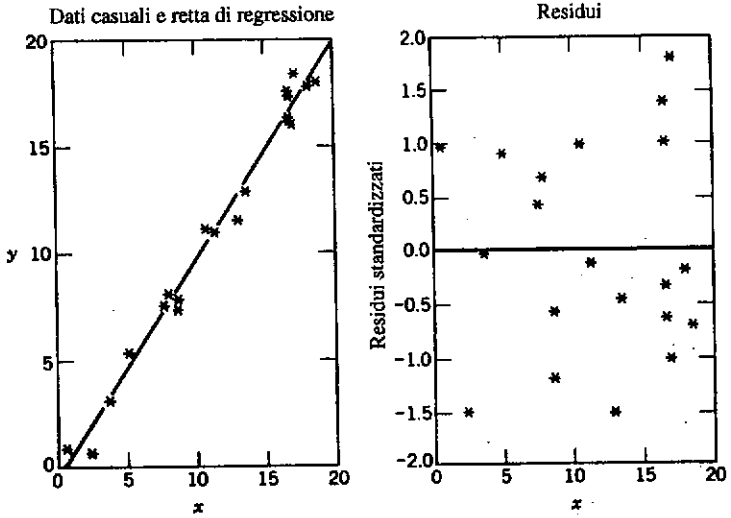
Quanto detto arricchisce di un significato ulteriore il coefficiente di correlazione lineare. Se ad esempio un campione di dati ha $r = 0.9$ ciò significa che il modello di regressione lineare semplice giustifica l'81% (visto che $0.9^2 = 0.81$) della variabilità nei valori delle risposte.

9.6 Analisi dei residui: verifica del modello

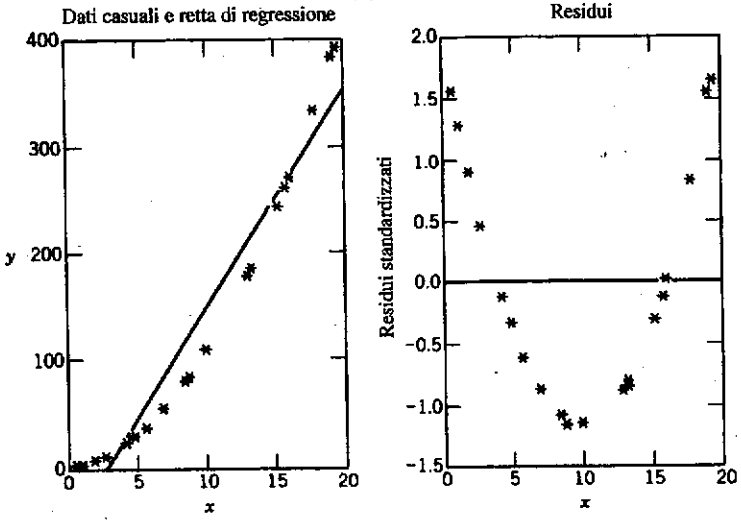
Il primo passo per chiarire se un modello di regressione lineare semplice quale

$$Y = \alpha + \beta x + e, \qquad e \sim \mathcal{N}(0, \sigma^2)$$

si adatti o meno ai dati, consiste nello studio del diagramma di dispersione: spesso anzi esso è sufficiente a convincerci in un senso o nell'altro. Quando però il diagramma di dispersione non è tale da escludere il modello suddetto, è bene calcolare gli

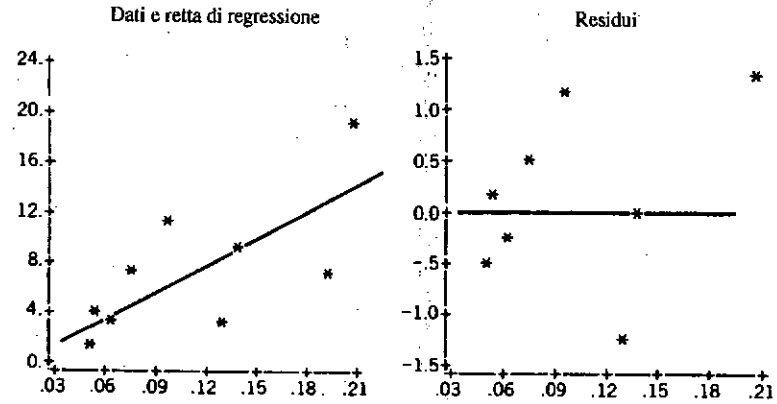


(a)



(b)

Figura 9.9



(c)

Figura 9.9 (continua)

stimatori dei minimi quadrati A e B e quindi analizzare i *residui*, $Y_i-(A+Bx_i)$, per $i=1,2,\ldots,n$. Per prima cosa essi vanno normalizzati, dividendoli per lo stimatore $\sqrt{SS_R/(n-2)}$ della deviazione standard delle Y_i. Le quantità risultanti,

$$\frac{Y_i-(A+Bx_i)}{\sqrt{SS_R/(n-2)}}, \qquad i=1,2,\ldots,n \tag{9.6.1}$$

sono chiamate *residui standardizzati*.

Quando il modello di regressione lineare semplice è corretto, i residui standardizzati sono approssimativamente variabili aleatorie normali standard indipententi, essendo quindi distribuiti attorno allo zero, con il 95% circa dei valori compresi tra -2 e $+2$ (più precisamente, $P(-1.96 < Z < 1.96) \approx 0.95$). Inoltre, un grafico di questi valori non deve mostrare alcuna regolarità geometrica, perché esse sono un forte indizio che il modello lineare semplice non è valido.

La Figura 9.9 presenta tre diversi diagrammi di dispersione, con i loro corrispondenti residui standardizzati. Il primo diagramma sembra adattarsi piuttosto bene alla stima della retta di regressione e questo si evince sia dalla dispersione casuale dei residui, sia da quella dei dati. La seconda coppia di grafici mostra una forte regolarità nei residui, che sono prima decrescenti e poi crescenti all'aumentare del livello di ingresso. Questo di solito significa che per descrivere la relazione tra ingresso e risposta si rendono necessari termini di grado più elevato (rispetto a quelli lineari), e ciò in questo caso è ben visibile anche dal diagramma di dispersione dei dati (i quali, più che una retta, sembrano seguire una parabola). Anche il terzo diagramma dei residui standardizzati mostra una certa regolarità: in questo caso il loro valore assoluto sembra crescere con il livello di ingresso. Ciò può voler dire ad esempio che la

varianza delle Y_i non è costante, ma cresce con x_i e anche in questo caso il modello di regressione lineare semplice non interpreta correttamente i dati.

9.7 Linearizzazione

In certe situazioni può essere evidente che la risposta media non sia una funzione lineare del livello di ingresso. Se la forma di questa relazione può essere determinata si può a volte riportarsi al caso lineare con un cambiamento di variabili. Ad esempio in certi ambiti l'intensità $W(t)$ di un segnale dopo un tempo t dall'emissione si sa seguire un decadimento approssimativemente esponenziale,

$$W(t) \approx ce^{-dt}$$

Se prendiamo i logaritmi naturali, ciò può essere espresso come

$$\log W(t) \approx \log c - dt$$

se ora poniamo

$$\begin{aligned} Y &= \log W(t) \\ \alpha &= \log c \\ \beta &= -d \end{aligned}$$

la relazione iniziale può essere modellizzata da

$$Y = \alpha + \beta t + e$$

permettendoci di stimare α e β con l'usuale metodo dei minimi quadrati. Si possono perciò fare predizioni sulla relazione studiata tramite

$$W(t) \approx e^{A+Bt}$$

Esempio 9.7.1. È stato dimostrato che la probabilità che un quarantenne che fuma da dieci anni si ammali di tumore ai polmoni entro i venti anni successivi è una funzione del numero medio di sigarette che consuma. Quelli riportati in Tabella 9.1 sono i risultati di uno studio estensivo (fatto sui topi ed estrapolato agli esseri umani). Usando questi dati vorremmo stimare la probabilità di contrarre il cancro per una persona che consumi 35 sigarette al giorno.

Denotiamo con P_i la probabilità di contrarre il cancro ai polmoni nei prossimi venti anni, nell'ipotesi che continuiamo a fumare i sigarette al giorno. Nonostante un grafico di P_i possa sembrare grosso modo lineare (si veda la Figura 9.10), possiamo ottenere una corrispondenza migliore considerando una relazione nonlineare. Per

Tabella 9.1

Numero medio di sigarette al giorno	Probabilità di contrarre il cancro ai polmoni
5	0.061
10	0.113
20	0.192
30	0.259
40	0.339
50	0.401
60	0.461
80	0.551

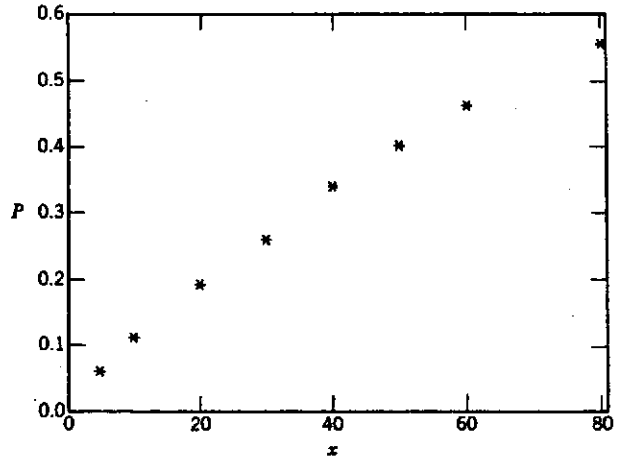


Figura 9.10 Esempio 9.7.1.

trovare un modello che descriva come P_i è legato a i, ragioniamo come segue (anche se potrà sembrare uno schema superficiale e semplificato sarà alla fine giustificato se troveremo una buona corrispondenza con i dati in nostro possesso).

Supponiamo che ogni sigaretta fumata – indipendentemente dalle altre – abbia una piccola probabilità fissata di causare la malattia (ad esempio danneggiando il DNA di una cellula polmonare). Fumando i sigarette al giorno, la probabilità di non contrarre il cancro con nessuna di queste è il prodotto delle probabilità che ciascuna delle $i \times 365 \times 20$ sigarette fumate in vent'anni non abbia avuto conseguenze. Aggiungiamo anche un fattore incognito c per la probabilità di ammalarsi per ragioni indipendenti dal fumo, ottenendo che

$$1 - P_i = P(\text{niente tumore fumando } i \text{ sigarette al giorno})$$

$$= c \cdot P(\text{una sigaretta fumata non causa il tumore})^{20 \cdot 365 \cdot i}$$

Questa relazione può essere scritta come

$$1 - P \approx cd^x$$

ovvero

$$\log(1 - P) \approx \log c + x \log d$$

Da cui, ponendo

$$Y = -\log(1 - P), \qquad \alpha = -\log c, \qquad \beta = -\log d$$

otteniamo l'equazione di regressione

$$Y = \alpha + \beta x + e$$

Per vedere se i dati confermano questo modello, tracciamo il diagramma di dispersione di $-\log(1 - P)$ rispetto a x. I dati trasformati sono riportati in Tabella 9.2, e il grafico è rappresentato in Figura 9.11.

Eseguendo il Programma 9.2 o facendo i calcoli a mano, troviamo che

$$A \approx 0.0154 \qquad B \approx 0.00989$$

Ritornando alle variabili originali con la trasformazione inversa otteniamo poi che le stime di c e d sono

$$\widehat{c} = e^{-A} \approx 0.9847$$
$$\widehat{d} = e^{-B} \approx 0.9901$$

e quindi la relazione nonlineare stimata è

$$\widehat{P} \approx 1 - 0.9847 \cdot (0.9901)^x$$

I residui $P - \widehat{P}$ sono presentati nella Tabella 9.3 □

Tabella 9.2

Numero medio di sigarette al giorno	$-\log(1 - P)$
5	0.063
10	0.120
20	0.213
30	0.300
40	0.414
50	0.512
60	0.618
80	0.801

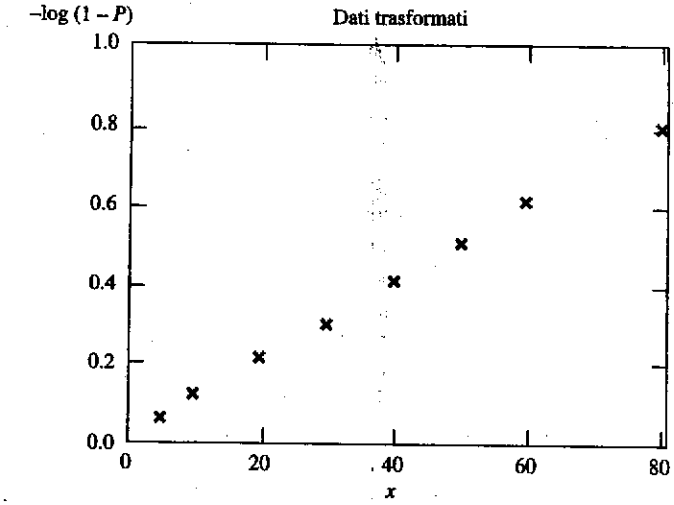


Figura 9.11

Osservazione 9.7.1. Quando P è la frazione di una popolazione che contrae un male, e il livello di esposizione è indicato da x, possiamo, come nell'Esempio 9.7.1, usare il modello

$$-\log(1 - P) = \alpha + \beta x + e \tag{9.7.1}$$

Un secondo modello frequentemente utilizzato e detto *modello logistico* è basato sulla relazione

$$\log\left(\frac{P}{1 - P}\right) = \alpha + \beta x + e \tag{9.7.2}$$

La quantità $\frac{P}{1-P}$ è detta *odds-ratio*. Il suo senso è questo: se un evento ha probabilità $P = \frac{3}{4}$ di verificarsi, allora il suo odds-ratio è $\frac{P}{1-P} = 3/1$, ovvero un bookmaker onesto lo dovrebbe "dare 3 a 1".

Tabella 9.3

x	P	$\widehat{P}$	$P - \widehat{P}$
5	0.061	0.063	-0.002
10	0.113	0.109	0.040
20	0.192	0.193	-0.001
30	0.259	0.269	-0.010
40	0.339	0.339	0.000
50	0.401	0.401	0.000
60	0.461	0.458	0.003
80	0.551	0.556	-0.005

9.8 Minimi quadrati pesati

Nel modello di regressione

$$Y = \alpha + \beta x + e$$

può capitare che la varianza della risposte non sia costante ma dipenda dal livello di ingresso. Se queste dipendenze sono note – oppure se sono note a meno di un fattore moltiplicativo – i parametri di regressione si possono stimare minimizzando una *somma pesata* dei residui al quadrato. In particolare, se

$$\mathrm{Var}(Y_i) = \frac{\sigma^2}{w_i} \tag{9.8.1}$$

con le w_i note e σ^2 eventualmente ignota, allora gli stimatori A e B vanno scelti in modo da minimizzare

$$\sum_{i=1}^{n} \frac{[Y_i - (A + Bx_i)]^2}{\mathrm{Var}(Y_i)} = \frac{1}{\sigma^2} \sum_{i=1}^{n} w_i (Y_i - A - Bx_i)^2$$

Calcolando le derivate parziali rispetto ad A e a B e ponendole uguali a zero, si trova il sistema seguente, per i parametri A e B cercati.

$$\begin{cases} \displaystyle\sum_{i=1}^{n} w_i Y_i = A \sum_{i=1}^{n} w_i + B \sum_{i=1}^{n} w_i x_i \\ \displaystyle\sum_{i=1}^{n} w_i x_i Y_i = A \sum_{i=1}^{n} w_i x_i + B \sum_{i=1}^{n} w_i x_i^2 \end{cases} \tag{9.8.2}$$

Queste equazioni possono essere facilmente risolte per trovare gli stimatori dei minimi quadrati.

Esempio 9.8.1. Per maturare una comprensione del perché gli stimatori giusti si trovino minimizzando la somma pesata dei quadrati, anziché la somma semplice, consideriamo la seguente situazione. Siano $X_1, X_2, \ldots, X_n$ variabili aleatorie $\mathcal{N}(\mu, \sigma^2)$ e indipendenti. Supponiamo inoltre che le X_i non siano osservabili, e che disponiamo solo del valore di Y_1 e Y_2, definite da

$$Y_1 := X_1 + \cdots + X_k, \qquad Y_2 := X_{k+1} + \cdots + X_n, \qquad k < n$$

Basandoci solo su Y_1 e Y_2, come possiamo stimare μ?

Anche se sappiamo che il miglior stimatore per μ è la media campionaria

$$\overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i = \frac{Y_1 + Y_2}{n}$$

vediamo di calcolare quale stimatore si otterrebbe usando metodo dei minimi quadrati ordinario. Siccome

$$E[Y_1] = k\mu, \qquad E[Y_2] = (n-k)\mu$$

lo stimatore dei minimi quadrati per μ si trova minimizzando al variare di μ l'espressione

$$(Y_1 - k\mu)^2 + (Y_2 - (n-k)\mu)^2$$

Derivando rispetto a μ e uguagliando a zero, troviamo che lo stimatore cercato deve soddisfare

$$-2k(Y_1 - k\widehat{\mu}) - 2(n-k)(Y_2 - (n-k)\widehat{\mu}) = 0$$

ovvero

$$[k^2 + (n-k)^2]\widehat{\mu} = kY_1 + (n-k)Y_2$$

e quindi

$$\widehat{\mu} = \frac{kY_1 + (n-k)Y_2}{k^2 + (n-k)^2}$$

Quello che abbiamo costruito è uno stimatore non distorto, infatti

$$\begin{aligned} E[\widehat{\mu}] &= \frac{kE[Y_1] + (n-k)E[Y_2]}{k^2 + (n-k)^2} \\ &= \frac{k^2\mu + (n-k)^2\mu}{k^2 + (n-k)^2} = \mu \end{aligned}$$

e tuttavia non è lo stimatore ottimale $\overline{X}$.

Proviamo ora a calcolare lo stimatore che si ottiene minimizzando la somma pesata dei quadrati. Cerchiamo quindi il valore μ_w che rende minima la seguente espressione al variare di μ:

$$\frac{(Y_1 - k\mu)^2}{\mathrm{Var}(Y_1)} + \frac{(Y_2 - (n-k)\mu)^2}{\mathrm{Var}(Y_2)}$$

Siccome

$$\mathrm{Var}(Y_1) = k\sigma^2, \qquad \mathrm{Var}(Y_2) = (n-k)\sigma^2$$

ciò è equivalente a minimizzare

$$\frac{(Y_1 - k\mu)^2}{k} + \frac{(Y_2 - (n-k)\mu)^2}{n-k}$$

Calcoliamo la derivata rispetto a μ e poniamola pari a zero, ottendo che μ_w deve soddisfare

$$-2k\frac{Y_1 - k\mu_w}{k} - 2(n-k)\frac{Y_2 - (n-k)\mu_w}{n-k} = 0$$

ovvero

$$Y_1 + Y_2 = n\mu_w$$

e cioè

$$\mu_w = \frac{Y_1 + Y_2}{n}$$

Perciò lo stimatore dei minimi quadrati pesati coincide con la media campionaria, che è ottimale tra tutti gli stimatori possibili. □

Osservazione 9.8.1.

(a) La somma pesata dei quadrati può anche essere vista come la naturale quantità da minimizzare quando l'equazione di regressione

$$Y = \alpha + \beta x + e$$

viene moltipicata per $\sqrt{w}$. Infatti nell'equazione

$$Y\sqrt{w} = \alpha\sqrt{w} + \beta x\sqrt{w} + e\sqrt{w}$$

il termine di errore $e\sqrt{w}$, ha media nulla e varianza costante $\frac{\sigma^2}{w_i} w_i = \sigma^2$, per cui gli stimatori dei minimi quadrati di α e β sono quei valori A e B che rendono minima l'espressione

$$\sum_{i=1}^{n} (Y_i\sqrt{w_i} - A\sqrt{w_i} - Bx_i\sqrt{w_i})^2 = \sum_{i=1}^{n} w_i(Y_i - A - Bx_i)^2$$

(b) L'approccio dei minimi quadrati pesati dà grande rilevanza ai dati con i pesi maggiori (ovvero quelli con la minore varianza nel termine di errore).

Potrebbe sembrare che il metodo dei minimi quadrati pesati non sia utile nella pratica, visto che richiede (a meno di una costante) la conoscenza della varianza delle risposte a livelli di ingresso arbitrari. Tuttavia, analizzando il modello che ha generato i dati è spesso possibile determinare questi valori, come sarà evidenziato dai prossimi due esempi.

Esempio 9.8.2. I dati seguenti rappresentano dei tempi di percorrenza in una zona centrale di una grande città. La variabile indipendente è la distanza percorsa.

Distanza (miglia)	0.5	1	1.5	2	3	4	5	6	8	10
Tempo (minuti)	15.0	15.1	16.5	19.9	27.7	29.7	26.7	35.9	42.0	49.4

Assumendo una relazione lineare del tipo

$$Y = \alpha + \beta x + e$$

tra il tempo di percorrenza Y e la distanza x, come possiamo stimare α e β? Per impiegare il metodo dei minimi quadrati pesati dovremmo conoscere la varianza di Y in funzione di x, a meno di una costante di proporzionalità. Siamo convinti che la varianza sia proporzionale a x, e di seguito ne diamo una argomentazione.

Sia d la lunghezza di un isolato del centro. Uno spostamento di una distanza x consiste allora di x/d isolati, e se denotiamo con Y_i, per $i = 1, 2, \ldots, x/d$ i tempi di percorrenza dei singoli isolati attraversati, allora vale la relazione

$$Y = Y_1 + Y_2 + \cdots + Y_{x/d}$$

Sembra ragionevole per molte applicazioni supporre che le Y_i siano indipendenti e abbiano varianza comune. In questo modo

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}(Y_1) + \cdots + \mathrm{Var}(Y_{x/d}) \\ &= \frac{x}{d}\,\mathrm{Var}(Y_1) && \text{perché le varianze sono uguali} \\ &= x\sigma^2 && \text{ponendo } \sigma^2 := \mathrm{Var}(Y_1)/d \end{aligned}$$

Perciò non sembra azzardato prendere come stimatori dei parametri di regressione i valori A e B che rendono minima l'espressione

$$\sum_{i=1}^{n} \frac{(Y_i - A - Bx_i)^2}{x_i}$$

Usando i dati precedenti con i pesi $w_i = 1/x_i$, le Equazioni (9.8.2) divengono

$$\begin{cases} 104.22 = 5.34A + 10B \\ 277.9 = 10A + 41B \end{cases}$$

che hanno come soluzione

$$A \approx 12.56, \qquad B \approx 3.71$$

Un grafico della retta di regressione stimata $12.56 + 3.71x$, unitamente ai punti osservati è illustrato in Figura 9.12. Come verifica qualitativa della soluzione trovata, si noti che la linea di regressione interpola bene i dati con livello di ingresso piccolo, che è quello che ci si aspetta, visto che i pesi sono inversamente proporzionali agli ingressi. □

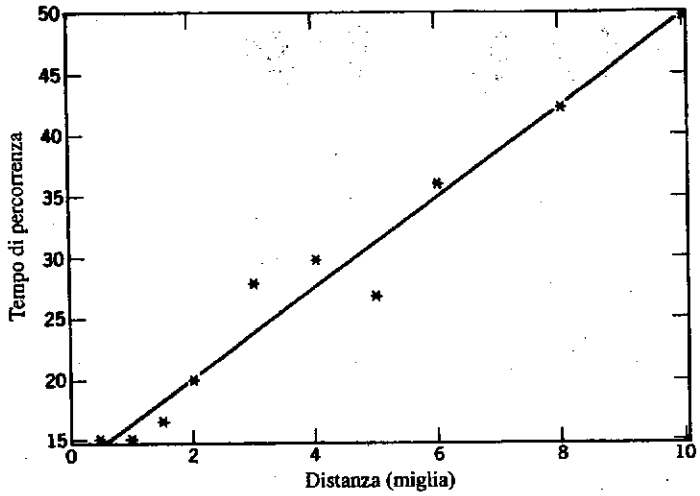


Figura 9.12 Retta di regressione e dati dell'Esempio 9.8.2

Esempio 9.8.3. Consideriamo la relazione tra il numero x delle vetture che percorrono un tratto autostradale molto trafficato in un certo intervallo di tempo, e il numero Y degli incidenti che hanno luogo nello stesso periodo. Dopo un po' di riflessione potremmmo essere d'accordo che il modello lineare

$$Y = \alpha + \beta x + e$$

sia appropriato alla circostanza. Non sembra però esserci alcuna ragione a priori perché $\mathrm{Var}(Y)$ non dipenda dal livello di ingresso x, e quindi non è chiaro se siamo giustificati nello stimare α e β con il metodo dei minimi quadrati ordinario. In effetti, proveremo ora a giustificare l'approccio dei minimi quadrati pesati, con scelta dei pesi $1/x$, ovvero A e B andranno presi in modo da rendere minima l'espressione

$$\sum_{i=1}^{n} \frac{(Y_i - A - Bx_i)^2}{x_i}$$

La ragione per questa scelta va cercata nel fatto che Y ha approssimativamente distribuzione di Poisson. Infatti possiamo pensare che vi sia un grande numero di automobili x, ciascuna delle quali con una piccola probabilità di essere coinvolta in un incidente. Siccome la varianza di una poissoniana coincide con la sua media,

otteniamo che

$$\begin{aligned} \mathrm{Var}(Y) &\approx E[Y] && \text{perché } Y \text{ è approssimativamente di Poisson} \\ &= \alpha + \beta x \\ &\approx \beta x && \text{per } x \text{ grande } \square \end{aligned}$$

Osservazione 9.8.2.

(a) Un'altra tecnica impiegata spesso quando la varianza della risposta dipende dal livello di ingresso consiste nel tentare di stabilizzare la prima con un'opportuna trasformazione. Ad esempio, se Y è di Poisson con media λ, si può dimostrare che $\sqrt{Y}$ ha approssimativamente varianza $1/4$, indipendentemente dal valore di λ (si veda la parte **(b)** più avanti). Basandoci su questo fatto, potremmo cercare ragionamenti che giustifichino una relazione lineare tra il livello di ingresso e $E[\sqrt{Y}]$, considerando poi un modello di regressione del tipo

$$\sqrt{Y} = \alpha + \beta x + e$$

Il problema di questo approccio è che nelle situazioni in cui è ragionevole immaginare una relazione approssimativamente lineare tra ingresso e risposta media, non è assolutamente chiaro perché dovrebbe esistere una simile relazione anche tra la media della radice quadrata della risposta e il livello di ingresso. Per questa ragione l'autore predilige l'approccio dei minimi quadrati pesati.

(b) Se Y ha distribuzione di Poisson di media λ, allora $\mathrm{Var}(\sqrt{Y}) \approx 0.25$, e l'approssimazione è tanto migliore quanto più grande è λ. Abbozziamo una dimostrazione di questo fatto[4].

Sia $g(y) := \sqrt{y}$, e consideriamo l'espansione in serie di Taylor di g nel punto λ. Ignorando i termini successivi a quello del secondo ordine otteniamo che

$$g(y) \approx g(\lambda) + g'(\lambda)(y - \lambda) + \frac{1}{2} g''(\lambda)(y - \lambda)^2$$

da cui, sostituendo $g'(\lambda) = \frac{1}{2}\lambda^{-1/2}$ e $g''(\lambda) = -\frac{1}{4}\lambda^{-3/2}$ otteniamo, valutando l'espressione nel punto casuale Y (che cadrà però vicino a $\lambda = E[Y]$),

$$\sqrt{Y} \approx \sqrt{\lambda} + \frac{1}{2}\lambda^{-1/2}(Y - \lambda) - \frac{1}{8}\lambda^{-3/2}(Y - \lambda)^2$$

Prendendo quindi i valori attesi e ricordando che

$$E[Y - \lambda] = 0, \qquad E[(Y - \lambda)^2] =: \mathrm{Var}(Y) = \lambda$$

[4] Il lettore tenga presente che i passaggi seguenti possono essere resi rigorosi.

si ha che

$$E[\sqrt{Y}] \approx \sqrt{\lambda} - \frac{1}{8\sqrt{\lambda}}$$

e quindi

$$E[\sqrt{Y}]^2 \approx \lambda - \frac{1}{4} + \frac{1}{64\lambda} \approx \lambda - \frac{1}{4}$$

da cui

$$\begin{aligned} \mathrm{Var}(\sqrt{Y}) &= E[Y] - E[\sqrt{Y}]^2 \\ &\approx \lambda - \left(\lambda - \frac{1}{4}\right) = \frac{1}{4} \end{aligned}$$

9.9 Regressione polinomiale

Nei casi in cui la relazione che lega la variabile di risposta Y con quella indipendente x non possa essere approssimata adeguatamente con modelli lineari, si può a volte ottenere un buon fit, prendendo in considerazione anche le relazioni polinomiali. In particolare, possiamo studiare se si adatti bene ai dati un modello come il seguente,

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r + e \tag{9.9.1}$$

dove $\beta_0, \beta_1, \ldots, \beta_r$ sono i coefficienti di regressione che è necessario stimare. Supponendo che i dati consistano di n coppie di valori, (x_i, Y_i), $i = 1, 2, \ldots, n$, gli stimatori dei minimi quadrati di $\beta_0, \beta_1, \ldots, \beta_r$, che denotiamo con $B_0, B_1, \ldots, B_r$ sono quei valori che rendono minima l'espressione seguente,

$$\sum_{i=1}^{n} (Y_i - B_0 - B_1 x_i - \cdots - B_r x_i^r)^2$$

Per determinarli calcoliamo le derivate parziali rispetto a $B_0, B_1, \ldots, B_r$ della somma di quadrati precedente, e le poniamo uguali a zero. Riarrangiando le equazioni che si ottengono, arriviamo al seguente sistema di $r+1$ equazioni lineari[5], che

[5] Sono lineari rispetto alle B_i che sono le incognite.

sono dette equazioni normali.

$$\begin{cases} \displaystyle\sum_{i=1}^{n} Y_i = B_0 n + B_1 \sum_{i=1}^{n} x_i + B_2 \sum_{i=1}^{n} x_i^2 + \cdots + B_r \sum_{i=1}^{n} x_i^r \\ \displaystyle\sum_{i=1}^{n} x_i Y_i = B_0 \sum_{i=1}^{n} x_i + B_1 \sum_{i=1}^{n} x_i^2 + B_2 \sum_{i=1}^{n} x_i^3 + \cdots + B_r \sum_{i=1}^{n} x_i^{r+1} \\ \cdots \\ \displaystyle\sum_{i=1}^{n} x_i^r Y_i = B_0 \sum_{i=1}^{n} x_i^r + B_1 \sum_{i=1}^{n} x_i^{r+1} + B_2 \sum_{i=1}^{n} x_i^{r+2} + \cdots + B_r \sum_{i=1}^{n} x_i^{2r} \end{cases} \tag{9.9.2}$$

Nel cercare il polinomio che meglio interpola i dati, la scelta del grado necessario va ponderata studiando il diagramma di dispersione, che spesso ce ne può dare un'idea (ad esempio la Figura 9.9 (b) mostra dei dati che suggeriscono di usare polinomi di secondo grado). È bene sottolineare che si deve sempre scegliere il grado più basso[6] tra quelli che permettono di descrivere adeguatamente i dati.

Ancora di più che nel caso lineare, è estremamente rischioso usare un fit polinomiale per predire il valore della risposta corrispondente ad un livello di ingresso x_0 che non sia molto vicino ai livelli $x_1, x_2, \ldots, x_n$, usati per ottenere il fit stesso. (È addirittura possibile che il fit polinomiale sia valido solo in una regione ristretta, che contiene $x_1, x_2, \ldots, x_n$ ma non x_0.)

Esempio 9.9.1. Si trovi un polinomio che interpoli i dati seguenti.

x	1	2	3	4	5	6	7	8	9	10
Y	20.6	30.8	55	71.4	97.3	131.8	156.5	197.3	238.7	291.7

Un grafico di questi dati (come quello in Figura 9.13), suggerisce che potrebbe valere una relazione quadratica del tipo

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

Possiamo a questo punto calcolare le somme di prodotti che ci occorrono:

$$\sum_{i=1}^{n} x_i = 55, \quad \sum_{i=1}^{n} x_i^2 = 385, \quad \sum_{i=1}^{n} x_i^3 = 3\,025, \quad \sum_{i=1}^{n} x_i^4 = 25\,333$$

$$\sum_{i=1}^{n} Y_i = 1\,291.1, \quad \sum_{i=1}^{n} x_i Y_i = 9\,549.3, \quad \sum_{i=1}^{n} x_i^2 Y_i = 77\,758.9$$

[6] Si noti infatti che se r è troppo alto (pari al numero n di dati o più), esiste un polinomio di grado r che passa *esattamente* per tutti i punti del diagramma, tuttavia non si può dare molta fiducia ad una tale "interpolazione".

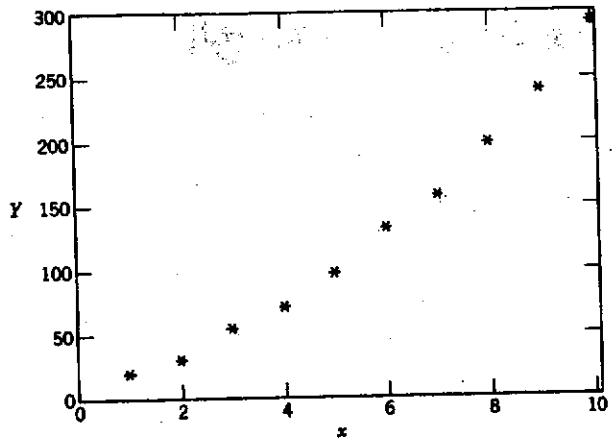


Figura 9.13

gli stimatori dei minimi quadrati sono le soluzioni del seguente sistema lineare,

$$\begin{cases} 1\,291.1 = 10B_0 + 55B_1 + 385B_2 \\ 9\,549.3 = 55B_0 + 385B_1 + 3\,025B_2 \\ 77\,758.9 = 385B_0 + 3\,025B_1 + 25\,333B_2 \end{cases} \tag{9.9.3}$$

Risolvendo queste equazioni (si veda eventualmente l'Osservazione 9.9.1 di seguito), si trova che

$$B_0 \approx 12.593, \qquad B_1 \approx 6.326, \qquad B_2 \approx 2.123$$

Quindi l'equazione di regressione quadratica stimata è

$$Y = 12.59 + 6.33x + 2.12x^2$$

Essa è rappresentata, in sovrapposizione ai dati, in Figura 9.14 □

Osservazione 9.9.1. In notazione matriciale l'Equazione (9.9.3) si può scrivere come

$$\begin{bmatrix} 1\,291.1 \\ 9\,549.3 \\ 77\,758.9 \end{bmatrix} = \begin{bmatrix} 10 & 55 & 385 \\ 55 & 385 & 3\,025 \\ 385 & 3\,025 & 25\,333 \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix}$$

che ha per soluzione

$$\begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} 10 & 55 & 385 \\ 55 & 385 & 3\,025 \\ 385 & 3\,025 & 25\,333 \end{bmatrix}^{-1} \begin{bmatrix} 1\,291.1 \\ 9\,549.3 \\ 77\,758.9 \end{bmatrix}$$

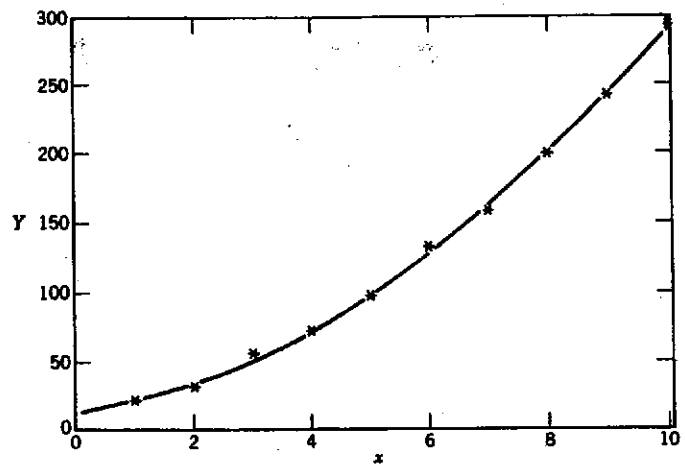


Figura 9.14

9.10 * Regressione lineare multipla

Nella gran parte delle applicazioni la risposta di un esperimento può essere predetta e modellizzata più accuratamente se invece di basarsi su di una singola variabile indipendente se ne utilizzano diverse. Studiamo il modello di regressione in cui vi sono k variabili indipendenti, e la risposta è legata loro tramite una relazione lineare:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e \tag{9.10.1}$$

dove per j che va da 1 a k, x_j è il livello della j-esima variabile di ingresso ed e è un errore casuale che noi assumeremo abbia distribuzione normale con media nulla e varianza σ^2 costante. I parametri $\beta_0, \beta_1, \ldots, \beta_k$, così come σ^2 si suppongono incogniti e devono essere stimati dai dati. Questi ultimi consisteranno di n osservazioni di risposte $Y_1, Y_2, \ldots, Y_n$, unitamente ai rispettivi livelli di ingresso, infatti per ogni $i = 1, 2, \ldots, n$ la risposta Y_i corrisponde a k livelli di ingresso, che denotiamo con $x_{i1}, x_{i2}, \ldots, x_{ik}$. Le variabili Y_i sono legate agli ingressi tramite

$$E[Y_i] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \tag{9.10.2}$$

Se denotiamo con $B_0, B_1, \ldots, B_k$ gli stimatori di $\beta_0, \beta_1, \ldots, \beta_k$, allora la somma dei residui al quadrato è

$$\sum_{i=1}^{n} (Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik})^2$$

ed è precisamente l'espressione che deve essere minimizzata dagli stimatori dei minimi quadrati, $B_0, B_1, \ldots, B_k$.

Per determinarli calcoliamo le derivate parziali rispetto a $B_0, B_1, \ldots, B_r$ della somma di quadrati precedente, e le poniamo uguali a zero. Le $r+1$ equazioni che si ottengono sono

$$\sum_{i=1}^{n}(Y_i - B_0 - B_1x_{i1} - B_2x_{i2} - \cdots - B_kx_{ik}) = 0$$

$$\sum_{i=1}^{n}x_{i1}(Y_i - B_0 - B_1x_{i1} - B_2x_{i2} - \cdots - B_kx_{ik}) = 0$$

$$\cdots$$

$$\sum_{i=1}^{n}x_{ik}(Y_i - B_0 - B_1x_{i1} - B_2x_{i2} - \cdots - B_kx_{ik}) = 0$$

Riarrangiando queste equazioni si trova che gli stimatori dei minimi quadrati $B_0, B_1, \ldots, B_k$ devono soddisfare il seguente sistema di equazioni normali:

$$\begin{cases} \sum_{i=1}^{n}Y_i = nB_0 + B_1\sum_{i=1}^{n}x_{i1} + B_2\sum_{i=1}^{n}x_{i2} + \cdots + B_k\sum_{i=1}^{n}x_{ik} \\ \sum_{i=1}^{n}x_{i1}Y_i = B_0\sum_{i=1}^{n}x_{i1} + B_1\sum_{i=1}^{n}x_{i1}^2 + B_2\sum_{i=1}^{n}x_{i1}x_{i2} + \cdots + B_k\sum_{i=1}^{n}x_{i1}x_{ik} \\ \quad\cdots \\ \sum_{i=1}^{n}x_{ik}Y_i = B_0\sum_{i=1}^{n}x_{ik} + B_1\sum_{i=1}^{n}x_{ik}x_{i1} + B_2\sum_{i=1}^{n}x_{ik}x_{i2} + \cdots + B_k\sum_{i=1}^{n}x_{ik}^2 \end{cases} \tag{9.10.3}$$

Prima di rivolvere le equazioni normali, conviene introdurre una notazione matriciale sintetica. Poniamo allora

$$\boldsymbol{Y} := \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{X} := \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} := \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{e} := \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \tag{9.10.4}$$

Si noti che $\boldsymbol{Y}$ è una matrice $n \times 1$, $\boldsymbol{X}$ è una $n \times p$, $\boldsymbol{\beta}$ una $p \times 1$ ed $\boldsymbol{e}$ una $n \times 1$, dove ovviamente si è posto $p = k + 1$.

Con questa notazione il modello di regressione multipla può essere scritto nella forma

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{9.10.5}$$

Se inoltre denotiamo con

$$\boldsymbol{B} := \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_k \end{bmatrix} \tag{9.10.6}$$

la matrice $n \times 1$ degli stimatori di minimi quadrati, allora le equazioni normali (9.10.3) prendono la forma

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B} = \boldsymbol{X}'\boldsymbol{Y} \tag{9.10.7}$$

dove $\boldsymbol{X}'$ è la trasposta di $\boldsymbol{X}$.

Per vedere che l'Equazione (9.10.7) è equivalente alla (9.10.3), si noti che

$$\boldsymbol{X}'\boldsymbol{X} := \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_{11} & x_{21} & \ldots & x_{n1} \\ \vdots & \vdots & & \vdots \\ x_{1k} & x_{2k} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum_i x_{i1} & \sum_i x_{i2} & \ldots & \sum_i x_{ik} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \ldots & \sum_i x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_i x_{ik} & \sum_i x_{ik}x_{i1} & \sum_i x_{ik}x_{i2} & \ldots & \sum_i x_{ik}^2 \end{bmatrix}$$

e anche che

$$\boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} \sum_i Y_i \\ \sum_i x_{i1}Y_i \\ \vdots \\ \sum_i x_{ik}Y_i \end{bmatrix}$$

da qui è facile convincersi che la (9.10.7) è proprio la versione matriciale delle Equazioni (9.10.3). Se poi $\boldsymbol{X}'\boldsymbol{X}$ è invertibile, cosa che accade quasi sempre, si possono ricavare gli stimatori dei minimi quadrati $\boldsymbol{B}$, moltiplicando a sinistra ambo i membri dell'equazione precedente per la matrice inversa $(\boldsymbol{X}'\boldsymbol{X})^{-1}$:

$$\boldsymbol{B} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{9.10.8}$$

Il Programma 9.10 del software abbinato al libro permette dei calcolare gli stimatori dei minimi quadrati, la matrice inversa $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, e SS_R.

Esempio 9.10.1. I dati nella Tabella 9.4 mettono in relazione il tasso di suicidi con l'ampiezza della popolazione e il tasso di divorzi in 8 posti diversi.

Vogliamo individuare un modello di regressione lineare multipla che interpoli questi dati; usiamo in particolare un modello della forma

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

dove Y è il tasso di suicidi, x_1 è la popolazione e x_2 è il tasso di divorzi.

Eseguiamo il Programma 9.10, ottenendo le schermate riportate nelle Figure 9.15. L'equazione di regressione stimata è perciò

$$Y = 3.507 - 0.2477 \cdot 10^{-3} \cdot x_1 + 0.2609 \cdot x_2$$

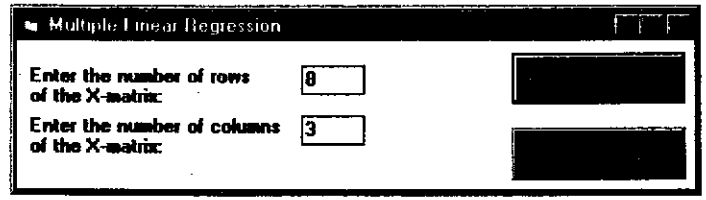
Il valore di B_1 indica che la popolazione non gioca un ruolo essenziale nel predire il tasso di suicidi (almeno nel caso in cui sia dato il tasso di divorzi). Magari la *densità* di popolazione avrebbe potuto rivelarsi un'informazione più utile. □

Osservando l'Equazione (9.10.8) si può notare che gli stimatori $B_0, B_1, \ldots, B_k$ (che compaiono come elementi della matrice B), sono combinazioni lineari delle $Y_1, Y_2, \ldots, Y_n$, che stiamo supponendo essere variabili aleatorie normali e indipendenti. Di conseguenza anche ciascuno di tali stimatori ha distribuzione normale, e, considerati nel loro insieme costituiscono una *variabile aleatoria normale multivariata*. Cerchiamo di ricavare i loro parametri. Per quanto riguarda le medie, si dimostra che gli stimatori dei minimi quadrati sono corretti:

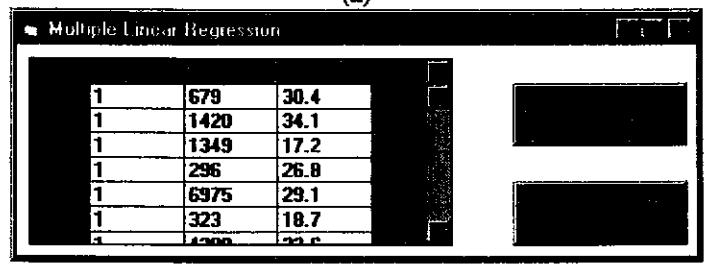
$$\begin{aligned} E[B] &= E[(X'X)^{-1}X'Y] \\ &= E[(X'X)^{-1}X'(X\beta + e)] \qquad \text{per l'Equazione (9.10.5)} \\ &= E[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'e] \\ &= E[\beta + (X'X)^{-1}X'e] \\ &= \beta + (X'X)^{-1}X'E[e] = \beta \end{aligned} \qquad (9.10.9)$$

Tabella 9.4

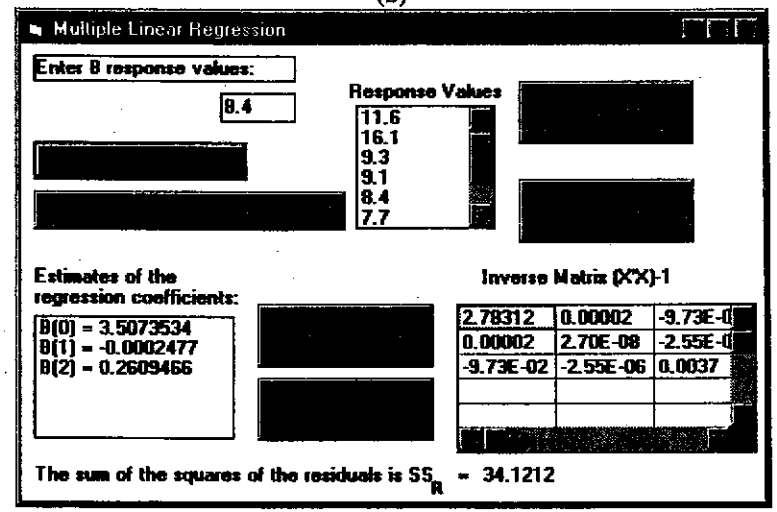
Luogo	Popolazione in migliaia	Divorzi su 100 000	Suicidi su 100 000
Akron, Ohio	679	30.4	11.6
Anaheim, California	1 420	34.1	16.1
Buffalo, New York	1 349	17.2	9.3
Austin, Texas	296	26.8	9.1
Chicago, Illinois	6 975	29.1	8.4
Columbia, South Carolina	323	18.7	7.7
Detroit, Michigan	4 200	32.6	11.3
Gary, Indiana	633	32.5	8.4



(a)



(b)



(c)

Figura 9.15

Per quanto riguarda le varianze, o meglio le covarianze delle B_j, mostreremo che esse possono essere ottenute dalla matrice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. In particolare l'elemento che si trova nella riga $i+1$ e nella colonna $j+1$ di tale matrice vale $\mathrm{Cov}(B_i, B_j)/\sigma^2$.

Per dimostrarlo, poniamo

$$\boldsymbol{C} := (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \tag{9.10.10}$$

Siccome $\boldsymbol{X}$ è $n \times p$, $\boldsymbol{X}'$ è $p \times n$, quindi $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ è $p \times p$ e così $\boldsymbol{C}$ è $p \times n$. Se denotiamo con C_{ij} l'elemento che si trova nella riga i e nella colonna j di questa matrice, possiamo riscrive $\boldsymbol{B}$ nella forma

$$\begin{bmatrix} B_0 \\ \vdots \\ B_{i-1} \\ \vdots \\ B_k \end{bmatrix} = \boldsymbol{B} = \boldsymbol{CY} = \begin{bmatrix} C_{11} & \dots & C_{1n} \\ \vdots & & \vdots \\ C_{i1} & \dots & C_{in} \\ \vdots & & \vdots \\ C_{p1} & \dots & C_{pn} \end{bmatrix} \begin{bmatrix} Y_1 \\ \\ \vdots \\ \\ Y_n \end{bmatrix}$$

Si ha quindi che

$$\begin{aligned} B_{i-1} &= \sum_{l=1}^{n} C_{il} Y_l \\ B_{j-1} &= \sum_{r=1}^{n} C_{jr} Y_r \end{aligned} \tag{9.10.11}$$

La covarianza di questi due stimatori è data da

$$\begin{aligned} \mathrm{Cov}(B_{i-1}, B_{j-1}) &= \mathrm{Cov}\left(\sum_{l=1}^{n} C_{il} Y_l, \sum_{r=1}^{n} C_{jr} Y_r\right) \\ &= \sum_{l=1}^{n}\sum_{r=1}^{n} C_{il} C_{jr}\, \mathrm{Cov}(Y_l, Y_r) \end{aligned}$$

Siccome quando $l \neq r$, Y_l e Y_r sono indipendenti,

$$\mathrm{Cov}(Y_l, Y_r) = \begin{cases} 0 & \text{se } l \neq r \\ \mathrm{Var}(Y_r) & \text{se } l = r \end{cases}$$

visto inoltre che $\mathrm{Var}(Y_r) = \sigma^2$, otteniamo che

$$\begin{aligned} \mathrm{Cov}(B_{i-1}, B_{j-1}) &= \sigma^2 \sum_{r=1}^{n} C_{ir} C_{jr} \\ &= \sigma^2 (\boldsymbol{CC}')_{ij} \end{aligned} \tag{9.10.12}$$

dove si intende che $(\boldsymbol{CC}')_{ij}$ è l'elemento della riga i, colonna j, di $\boldsymbol{CC}'$. Se si denota con $\mathrm{Cov}(\boldsymbol{B})$ la matrice delle covarianze, vale a dire,

$$\mathrm{Cov}(\boldsymbol{B}) := \begin{bmatrix} \mathrm{Cov}(B_0, B_0) & \dots & \mathrm{Cov}(B_0, B_k) \\ \vdots & & \vdots \\ \mathrm{Cov}(B_k, B_0) & \dots & \mathrm{Cov}(B_k, B_k) \end{bmatrix} \tag{9.10.13}$$

l'Equazione (9.10.12) si riscrive come

$$\mathrm{Cov}(\boldsymbol{B}) = \sigma^2 \boldsymbol{CC}' \tag{9.10.14}$$

Questa espressione può essere semplificata. Calcoliamo la trasposta di $\boldsymbol{C}$:

$$\begin{aligned} \boldsymbol{C}' &:= \left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)' \\ &= \boldsymbol{X}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\right)' \\ &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned}$$

dove l'ultima uguaglianza segue dal fatto che $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ è una matrice simmetrica (visto che anche $\boldsymbol{X}'\boldsymbol{X}$ lo è). Di conseguenza

$$\begin{aligned} \boldsymbol{CC}' &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned}$$

e quindi l'Equazione (9.10.14) diventa

$$\mathrm{Cov}(\boldsymbol{B}) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{9.10.15}$$

che era ciò che ci eravamo proposti di dimostrare. Si noti in particolare che, siccome $\mathrm{Cov}(B_i, B_i) = \mathrm{Var}(B_i)$, le varianze degli stimatori dei minimi quadrati sono date da σ^2 moltiplicato per gli elementi sulla diagonale di $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

La quantità σ^2 può essere stimata usando la somma dei quadrati dei residui. Infatti se poniamo

$$SS_{\mathrm{R}} := \sum_{i=1}^{n} (Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik})^2 \tag{9.10.16}$$

è possibile dimostrare che

$$\frac{SS_{\mathrm{R}}}{\sigma^2} \sim \chi^2_{n-(k+1)} \tag{9.10.17}$$

da cui deriva che

$$E\left[\frac{SS_{\mathrm{R}}}{\sigma^2}\right] = n - k - 1 \qquad \text{e anche} \qquad E\left[\frac{SS_{\mathrm{R}}}{n-k-1}\right] = \sigma^2$$

per cui $SS_{\mathrm{R}}/(n-k-1)$ è uno stimatore corretto di σ^2. Come nel caso della regressione lineare semplice, SS_{R} risulta indipendente dagli stimatori dei minimi quadrati $B_0, B_1, \dots, B_k$.

Osservazione 9.10.1. Denotiamo con r_i il residuo i-esimo, vale a dire

$$r_i := Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik}, \qquad i = 1, \ldots, n \tag{9.10.18}$$

e sia $\boldsymbol{r}$ la matrice (o vettore colonna) di questi residui,

$$\boldsymbol{r} := \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix} \tag{9.10.19}$$

in modo che

$$\boldsymbol{r} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B} \tag{9.10.20}$$

Questa notazione consente di scrivere SS_R in una nuova forma.

$$\begin{aligned} SS_R &:= \sum_{i=1}^{n} r_i^2 \\ &= \boldsymbol{r}'\boldsymbol{r} \\ &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}) \\ &= (\boldsymbol{Y}' - \boldsymbol{B}'\boldsymbol{X}')(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}) \\ &= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}\boldsymbol{B} - \boldsymbol{B}'(\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}) \\ &= \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X}\boldsymbol{B} \qquad \text{per la (9.10.7)} \end{aligned}$$

dove l'ultima uguaglianza segue dalla forma matriciale delle equazioni normali. Come SS_R, anche $\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{B}$ è uno scalare (anche perché visto che $\boldsymbol{Y}'$ è una matrice $1 \times n$, $\boldsymbol{X}$ è $n \times p$ e $\boldsymbol{B}$ è $p \times 1$, il loro prodotto è una matrice 1×1), ed è quindi uguale alla sua trasposta:

$$\begin{aligned} \boldsymbol{Y}'\boldsymbol{X}\boldsymbol{B} &= (\boldsymbol{Y}'\boldsymbol{X}\boldsymbol{B})' \\ &= \boldsymbol{B}'\boldsymbol{X}'\boldsymbol{Y} \end{aligned}$$

Abbiamo quindi dimostrato l'identità seguente:

$$SS_R = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{B}'\boldsymbol{X}'\boldsymbol{Y} \tag{9.10.21}$$

Questa è una formula per il calcolo di SS_R di una certa utilità (anche se occorre fare attenzione ai possibili problemi di instabilità numerica).

Esempio 9.10.2. Usando i dati dell'Esempio 9.10.1 avevamo calcolato che $SS_R \approx 34.12$. Siccome $n = 8$ e $k = 2$, la stima per σ^2 è $34.12/5 = 6.824$. □

Tabella 9.5

Albero	Età	Altitudine (1 000 piedi)	Precipitazioni (pollici)	Densità del legno	Diametro massimo (pollici)
1	44	1.3	250	0.63	18.1
2	33	2.2	115	0.59	19.6
3	33	2.2	75	0.56	16.6
4	32	2.6	85	0.55	16.4
5	34	2.0	100	0.54	16.9
6	31	1.8	75	0.59	17.0
7	33	2.2	85	0.56	20.0
8	30	3.6	75	0.46	16.6
9	34	1.6	225	0.63	16.2
10	34	1.5	250	0.60	18.5
11	33	2.2	255	0.63	18.7
12	36	1.7	175	0.58	19.4
13	33	2.2	75	0.55	17.6
14	34	1.3	85	0.57	18.3
15	37	2.6	90	0.62	18.8

Fonte: R. G. Skolmen, "Shrinkage and specific gravity variation in Robusta Eucalyptus wood grown in Hawaii", USDA Forest Service PSW-298, 1975.

Esempio 9.10.3. Il diametro massimo del tronco di un albero è influenzato da molti fattori. I dati della Tabella 9.5 mettono in relazione quello di una particolare varietà di eucalipto con la sua età, l'altitudine a cui cresce, la piovosità media annuale e la densità del legno.

Supponiamo che sussista un modello di regressione lineare della forma

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$$

dove x_1 indica l'età, x_2 l'altitudine, x_3 le precipitazioni, x_4 la densità del legno e Y è il diametro del tronco. Verifichiamo l'ipotesi che $\beta_2 = 0$, ovvero che conoscendo gli altri tre fattori, l'altitudine a cui l'albero cresce non influisca sul diametro del tronco.

Per verificare tale ipotesi eseguiamo il Programma 9.10 che fornisce, tra le altre, le statistiche seguenti,

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}_{3,3} \approx 0.379, \qquad SS_R \approx 19.34, \qquad B_2 \approx 0.0744$$

Dall'Equazione (9.10.15) segue allora che

$$\mathrm{Var}(B_2) \approx 0.379\sigma^2$$

e quindi

$$\frac{B_2 - \beta_2}{\sigma\sqrt{0.379}} \sim \mathcal{N}(0, 1)$$

Sostituendo σ^2 col suo stimatore $SS_R/10$, la variabile aleatoria precedente diviene una t di Student con 10 (vale a dire $n-k-1$) gradi di libertà:

$$\frac{B_2-\beta_2}{\sqrt{0.379\cdot SS_R/10}}\sim t_{10}$$

per cui, supponendo vera H_0, e quindi che $\beta_2=0$, si avrebbe che

$$\frac{B_2\sqrt{10}}{\sqrt{0.379\cdot SS_R}}\sim t_{10}$$

Siccome il valore assunto da questa statistica è $0.0744\sqrt{10}/\sqrt{0.379\times 19.34}\approx 0.087$, il p-dei-dati del test dell'ipotesi che $\beta_2=0$ vale

$$\begin{aligned} p\text{-dei-dati} &= P(|T_{10}|>0.087)\\ &= 2P(T_{10}>0.087)\\ &\approx 0.932 \qquad \text{grazie al Programma 5.8.2a}\end{aligned}$$

L'ipotesi viene quindi accettata a qualunque livello di significatività inferiore a 0.932, e in particolare a qualunque livello di significatività ragionevole. □

Osservazione 9.10.2. La quantità

$$R^2:=1-\frac{SS_R}{\sum_i(Y_i-\overline{Y})^2} \tag{9.10.22}$$

che misura la diminuzione di variabilità nelle risposte quando si tenga conto del valore degli ingressi, usando un modello del tipo

$$Y=\beta_0+\beta_1x_1+\cdots+\beta_kx_k+e$$

è detta *coefficiente di determinazione multipla.*

9.10.1 Predizione di risposte future

Supponiamo di essere prossimi a realizzare una serie di esperimenti, tutti con livelli di ingresso fissati, $x_1,x_2,\dots,x_k$. Basandoci su dati precedenti, che consistono nelle risposte $Y_1,Y_2,\dots,Y_n$, vorremmo stimare la risposta media di questi nuovi esperimenti. Siccome tale parametro incognito è dato da

$$E[Y|x]=\beta_0+\beta_1x_1+\cdots+\beta_kx_k \tag{9.10.23}$$

il naturale stimatore puntuale è $\sum_{i=0}^k B_ix_i$ (da qui in poi si intende che $x_0\equiv 1$).

Per ottenere gli intervalli di confidenza dobbiamo determinare la distribuzion della statistica $\sum_{i=0}^k B_ix_i$, che notiamo subito essere una variabile aleatoria normal in quanto esprimibile come combinazione lineare delle variabili aleatorie normali indipendenti $Y_1,Y_2,\dots,Y_n$. Resta solo da calcolarne media e varianza:

$$\begin{aligned} E\Big[\sum_{i=0}^k x_iB_i\Big] &= \sum_{i=0}^k x_iE[B_i]\\ &= \sum_{i=0}^k x_i\beta_i \qquad \text{perché } E[B_i]=\beta_i\\ &= E[Y|x] \end{aligned} \tag{9.10.24}$$

Si tratta perciò di uno stimatore corretto. Ricordando poi che la varianza di una variabile aleatoria coincide con la sua covarianza con sé stessa, si ha che

$$\begin{aligned} \operatorname{Var}\Big(\sum_{i=0}^k x_iB_i\Big) &= \operatorname{Cov}\Big(\sum_{i=0}^k x_iB_i,\ \sum_{j=0}^k x_jB_j\Big)\\ &= \sum_{i=0}^k\sum_{j=0}^k x_ix_j\operatorname{Cov}(B_i,B_j)\\ &= \sigma^2\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x} \end{aligned} \tag{9.10.25}$$

dove si è posto

$$\boldsymbol{x}:=\begin{bmatrix}x_0\\x_1\\\vdots\\x_k\end{bmatrix}=\begin{bmatrix}1\\x_1\\\vdots\\x_k\end{bmatrix} \tag{9.10.26}$$

e si è usato il fatto che l'elemento di coordinate $i+1$ e $j+1$ della matrice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ è $\operatorname{Cov}(B_i,B_j)/\sigma^2$. Con i risultati (9.10.24) e (9.10.25), che forniscono la media e la varianza della statistica studiata, otteniamo che

$$\frac{\sum_i x_iB_i-\sum_i x_i\beta_i}{\sigma\sqrt{\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}}}\sim\mathcal{N}(0,1)$$

ovvero, sostituendo σ^2 con il suo stimatore $SS_R/(n-k-1)$ analogamente a quanto fatto in precedenza, otteniamo che

$$\frac{\sum_i x_iB_i-\sum_i x_i\beta_i}{\sqrt{\frac{SS_R}{n-k-1}}\sqrt{\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}}}\sim t_{n-k-1} \tag{9.10.27}$$

Questo risultato ci consente di formulare gli intervalli di confidenza per la risposta media. In particolare si può affermare con livello di confidenza $1-\gamma$ che $\sum_{i=0}^{k} x_i\beta_i$ appartiene all'intervallo bilaterale

$$\sum_{i=0}^{k} x_i B_i \pm t_{\frac{\gamma}{2},n-k-1} \cdot \sqrt{\frac{SS_R}{n-k-1}} \sqrt{x'(X'X)^{-1}x} \qquad (9.10.28)$$

Esempio 9.10.4. Una acciaieria sta valutando la produzione di lamine ridotte a freddo con lo 0.15% di carbonio per una temperatura di ricottura di 1 150 gradi Fahrenheit. Se ne vuole stimare la durezza media (metodo Rockwell 30 T). Per riuscirci si dispone dei dati mostrati nella Tabella 9.6, ottenuti da 10 differenti esemplari, ottenuti con percentuali di carbonio e temperature di ricottura diverse.

Tabella 9.6

Durezza	Percentuale di carbonio	Temperatura di ricottura (1000 F)
79.2	0.02	1.05
64.0	0.03	1.20
55.7	0.03	1.25
56.3	0.04	1.30
58.6	0.10	1.30
84.3	0.15	1.00
70.4	0.15	1.10
61.3	0.09	1.20
51.3	0.13	1.40
49.8	0.09	1.40

Si stimi la durezza media delle lamine che si progetta di realizzare, tramite un intervallo di confidenza al 95%.

Per prima cosa eseguiamo il Programma 9.10, che fornisce i risultati mostrati nelle Figure 9.16, 9.17 e 9.18. Ne deduciamo che la stima puntuale della durezza media

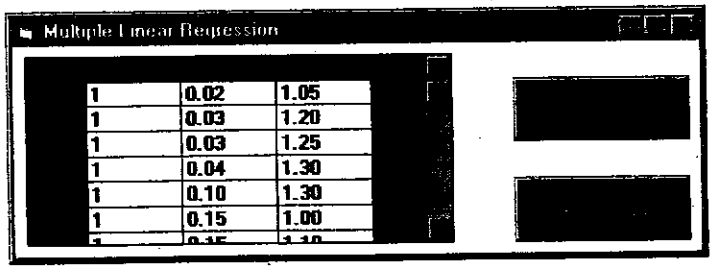


Figura 9.16

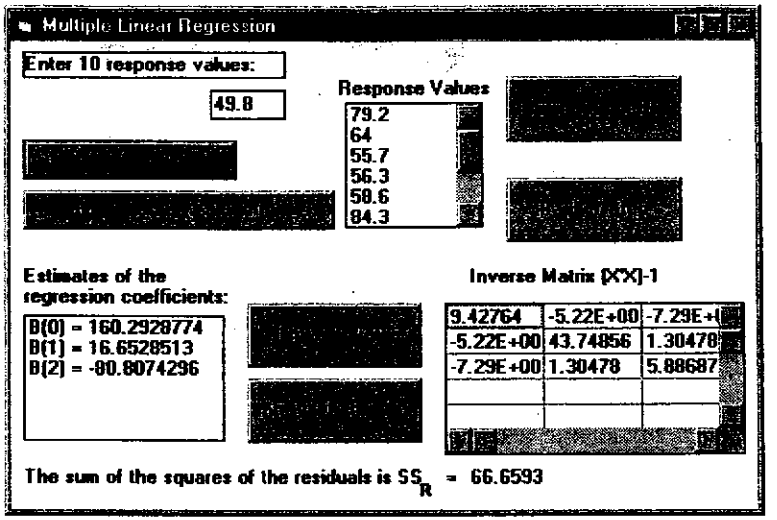


Figura 9.17

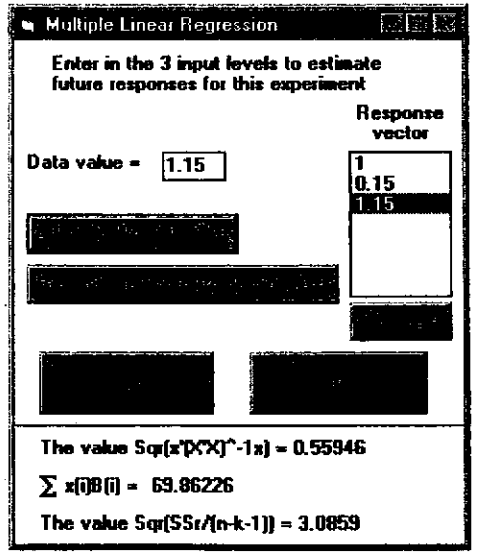


Figura 9.18

per delle lamine con lo 0.15% di carbonio e una temperatura di ricottura di 1 150 è di 69.86. Secondariamente, visto che $t_{0.025,7} \approx 2.365$, un intervallo di confidenza al 95% è dato da

$$69.86 \pm 4.08 \quad \square$$

Nel caso si voglia realizzare un singolo esperimento ai livelli di ingresso $x_1, x_2, \ldots, x_k$ (e non tutta una serie di prove), è solitamente più utile ottenere un predittore della risposta, piuttosto uno stimatore della risposta media. Siamo quindi interessati a utilizzare il campione di dati $Y_1, Y_2, \ldots, Y_n$ per predire nel modo migliore il valore che verrà assunto dalla variabile aleatoria

$$Y(\boldsymbol{x}) = \sum_{i=0}^{k} \beta_i x_i + e, \qquad \text{dove } x_0 = 1$$

Un predittore puntuale è dato da $\sum_{i=0}^{k} B_i x_i$, dove B_i per $i = 0, 1, \ldots, k$, è lo stimatore dei minimi quadrati di β_i. Per determinare un intervallo di predizione per $Y(\boldsymbol{x})$, notiamo intanto che tale risposta è indipendente da $B_0, B_1, \ldots, B_k$, che sono basate su risposte precedenti. Quindi $Y(\boldsymbol{x}) - \sum_{i=0}^{k} B_i x_i$ è normale con media nulla e varianza data da

$$\begin{aligned} \operatorname{Var}\left[Y(\boldsymbol{x}) - \sum_{i=0}^{k} B_i x_i\right] &= \operatorname{Var}\left[Y(\boldsymbol{x})\right] + \operatorname{Var}\left(\sum_{i=0}^{k} B_i x_i\right) && \text{per l'indipendenza} \\ &= \sigma^2 + \sigma^2 \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x} && \text{per la (9.10.25)} \end{aligned}$$

motivo per cui

$$\frac{Y(\boldsymbol{x}) - \sum_i B_i x_i}{\sigma\sqrt{1 + \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}}} \sim \mathcal{N}(0,1)$$

ovvero, tramite la solita sostituzione di σ con il relativo stimatore,

$$\frac{Y(\boldsymbol{x}) - \sum_i B_i x_i}{\sqrt{\frac{SS_R}{n-k-1}}\sqrt{1 + \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}}} \sim t_{n-k-1} \tag{9.10.29}$$

Concludendo, con livello di confidenza $1 - \gamma$, la risposta $Y(\boldsymbol{x})$ cadrà entro

$$\sum_{i=0}^{k} B_i x_i \pm t_{\frac{\gamma}{2}, n-k-1} \cdot \sqrt{\frac{SS_R}{n-k-1}}\sqrt{1 + \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}} \tag{9.10.30}$$

Esempio 9.10.5. Torniamo all'Esempio 9.10.4 e immaginiamo di essere interessati a determinare un intervallo di valori che contenga con il 95% di confidenza la durezza di un singolo esemplare di lamina d'acciaio con lo 0.15% di carbonio e una temperatura di ricottura di 1 150 gradi Fahrenheit. Il punto medio di tale intervallo è lo

stesso trovato nell'Esempio 9.10.4, mentre il suo raggio differisce da quello usato precedenza per un fattore

$$\frac{\sqrt{1 + \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}}}{\sqrt{\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}}} \approx \frac{\sqrt{1.313}}{\sqrt{0.313}}$$

quindi l'intervallo di predizione cercato è dato da

$$69.86 \pm 8.36 \quad \square$$

Problemi

1. I dati seguenti mettono in relazione la percentuale di acqua x, contenuta in un certo materiale in una delle fasi di lavorazione, con la densità Y del prodotto finito.

x	5	6	7	10	12	15	18	29
Y	7.4	9.3	10.6	15.4	18.1	22.2	24.1	24.8

(a) Traccia il diagramma di dispersione.

(b) Trova la retta di regressione che interpola questi dati.

2. I dati seguenti illustrano la relazione esistente tra il prezzo unitario di un certo bene in luoghi differenti e il numero di unità dello stesso bene che sono state ordinate.

Pezzi ordinati	88	112	123	136	158	172
Prezzo	50	40	35	30	20	15

Secondo te quante unità verrebbero ordinate se il prezzo fosse 25?

3. Si studia il livello di corrosione di una certa sostanza metallica esponendola ad una atmosfera di ossigeno puro, ad una temperatura di 500 gradi Celsius. L'aumento relativo di massa della sostanza viene utilizzato come indicatore della quantità di ossigeno che ha reagito. I dati raccolti sono i seguenti:

Ore di esposizione	1.0	2.0	2.5	3.0	3.5	4.0
Incremento percentuale	0.02	0.03	0.035	0.042	0.05	0.054

(a) Traccia il diagramma di dispersione.

(b) Trova la relazione lineare che interpola meglio i dati.

(c) Fornisci una previsione dell'incremento di massa dopo 3.2 ore di esposizione.

4. I dati che seguono mostrano la relazione tra la densità x di certi campioni di legname e Y, la massima resistenza alla compressione opposta dal legno nella direzione della fibra (misurata in psi).

x	0.41	0.46	0.44	0.47	0.42	0.39	0.41	0.44	0.43	0.44
Y	1 850	2 620	2 340	2 690	2 160	1 760	2 500	2 750	2 730	3 120

(a) Traccia il diagramma di dispersione. Pensi che sussista una relazione lineare?

(b) Stima i coefficienti di regressione.

(c) Predici la resistenza alla compressione per un campione di legname con una densità di 0.43.

5. I dati seguenti mostrano l'incremento nella velocità di lettura (misurata in parole al minuto) dopo un numero diverso di settimane per 10 individui iscritti ad un corso di lettura veloce.

Numero di settimane	2	3	8	11	4	5	9	7	5	7
Aumento di velocità	21	42	102	130	52	57	105	85	62	90

(a) Traccia il diagramma di dispersione per capire se può sussistere una relazione lineare.

(b) Trova le stime dei minimi quadrati dei coefficienti di regressione.

(c) Stima il guadagno nel quale può mediamente sperare un iscritto che intenda seguire il corso per 7 settimane.

6. La spettroscopia infrarossa è spesso impiegata per determinare la percentuale di gomma naturale in misture di gomma naturale e sintetica. Per esemplari di composizione nota, lo strumento ha fornito le letture seguenti:

Percentuale	0	20	40	60	80	100
Lettura	0.734	0.885	1.050	1.191	1.314	1.432

Se una nuova miscela dà una lettura di 1.15 allo spettroscopio, qual è la percentuale di gomma naturale stimata?

7. La tabella che segue fornisce i punteggi medi per le parti linguistica e matematica del SAT[7] del 1996, in ciascuno degli stati americani. Viene anche riportata la percentuale di studenti diplomati che hanno sostenuto il test.

(a) Usa i dati dei primi 20 stati (da Alabama a Maine) per ottenere una predizione del punteggio medio in matematica in funzione della percentuale di studenti che sostengono il test.

(b) Confronta i valori predetti con quelli riscontrati nei 5 stati successivi.

[7] *Scholastic Aptitude Test*. Si tratta di un esame pubblico che devono superare gli studenti che finite le scuole secondarie desiderano iscriversi alla gran parte dei college americani, [*N.d.T.*]

Punteggi medi del SAT, ordinati per stato, 1996 (scala ricentrata)

	Linguistico	Matematico	Percentuale di partecipazione
Alabama	565	558	8
Alaska	521	513	47
Arizona	525	521	28
Arkansas	566	550	6
California	495	511	45
Colorado	536	538	30
Connecticut	507	504	79
Delaware	508	495	66
Dist. of Columbia	489	473	50
Florida	498	496	48
Georgia	484	477	63
Hawaii	485	510	54
Idaho	543	536	15
Illinois	564	575	14
Indiana	494	494	57
Iowa	590	600	5
Kansas	579	571	9
Kentucky	549	544	12
Louisiana	559	550	9
Maine	504	498	68
Maryland	507	504	64
Massachusetts	507	504	80
Michigan	557	565	11
Minnesota	582	593	9
Mississipi	569	557	4
Missouri	570	569	9
Montana	546	547	21
Nebraska	567	568	9
Nevada	508	507	31
New Hampshire	520	514	70
New Jersey	498	505	69
New Mexico	554	548	12
New York	497	499	73
North Carolina	490	486	59
North Dakota	596	599	5
Ohio	536	535	24
Oklahoma	566	557	8
Oregon	523	521	50
Pennsylvania	498	492	71
Rhode Island	501	491	69
South Carolina	480	474	57
South Dakota	574	566	5

Tennessee	563	552	14
Texas	495	500	48
Utah	583	575	4
Vermont	506	500	70
Virginia	507	496	68
Washington	519	519	47
West Virginia	526	506	17
Wisconsin	577	586	8
Wyoming	544	544	11
Media Nazionale	505	508	41

Fonte: The College Board

8. Verifica l'Equazione (9.3.3) che afferma che

$$\mathrm{Var}(A) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i^2 - \bar{x}^2)}$$

9. Considera nuovamente il Problema 4.

(a) Stima la varianza di una singola risposta.

(b) Determina un intervallo di confidenza al 90% per tale parametro.

10. Verifica l'identità seguente.

$$SS_{\mathrm{R}} = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

11. I dati seguenti riguardano 12 studenti di uno stesso corso di studi in legge. Tutti riportarono punteggi simili nella prova finale, inoltre ciascuno di essi entrò a lavorare in uno studio legale. La tabella mette a confronto i loro redditi in migliaia di dollari con le loro stature in pollici.

Statura	64	65	66	67	69	70	72	72	74	74	75	76
Reddito	91	94	88	103	77	96	105	88	122	102	90	114

(a) Ad un livello di significatività del 5%, questi dati confermano che vi sia un legame tra salario e altezza?

(b) Qual è stata la tua scelta per l'ipotesi nulla nel punto **(a)**? Motiva la risposta.

12. I dati che seguono rappresentano il numero di macchie solari apparse e il numero di vittime di incidenti stradali che si sono verificati negli anni dal 1970 al 1983. Verifica l'ipotesi che il numero delle vittime della strada non sia influenzato dalle macchie solari.

Anno	Macchie solari	Vittime di incidenti stradali (1 000)
1970	165	54.6
1971	89	53.3
1972	55	56.3
1973	34	49.6
1974	9	47.1
1975	30	45.9
1976	59	48.5
1977	83	50.1
1978	109	52.4
1979	127	52.5
1980	153	53.2
1981	112	51.4
1982	80	46.0
1983	45	44.6

Fonti: per le macchie solari, Jastrow and Thompson, Fundamentals and frontiers of astronomy; *per gli incidenti,* General Statistics of the U.S. *1985.*

13. Considera il modello di regressione lineare semplice

$$Y = \alpha + \beta x + e$$

e supponi che $0 < \beta < 1$.

(a) Dimostra che se $x < \frac{\alpha}{1-\beta}$, allora

$$x < E[Y] < \frac{\alpha}{1-\beta}$$

(b) Dimostra che se $x > \frac{\alpha}{1-\beta}$, allora

$$x > E[Y] > \frac{\alpha}{1-\beta}$$

E concludi che $E[Y]$ è sempre compresa tra x e $\frac{\alpha}{1-\beta}$.

14. È stato affermato, da istrutturi di volo con grande esperienza, che gli apprezzamen per un atterraggio particolarmente ben riuscito portano tipicamente ad un atterraggi immediatamente successivo che si rivela peggiore, mentre le critiche per un pessim atterraggio spesso sono seguite da una prestazione migliore. Dobbiamo concluderne ch i complimenti tendono ad abbassare il livello dell'esecuzione, mentre le critiche tendon ad elevarlo? Esiste qualche altra spiegazione?

15. Verifica la correttezza dell'Equazione (9.4.6):

$$\sqrt{\frac{n(n-2)S_{xx}}{SS_{\mathrm{R}} \cdot \sum_i x_i^2}}(A - \alpha) \sim t_{n-2}$$

16. I dati seguenti rappresentano la relazione esistente tra il numero di errori di allineamento Y e numero di rivetti mancanti x, per 10 differenti aeromobili.

Rivetti mancanti	13	15	10	22	30	7	25	16	20	15
Errori di allineamento	7	7	5	12	15	2	13	9	11	8

(a) Disegna il diagramma di dispersione.

(b) Stima i coefficienti di regressione.

(c) Verifica l'ipotesi che $\alpha = 1$.

(d) Stima il numero medio di errori di allineamento per un aeroplano cui manchino 24 rivetti.

(e) Calcola un intervallo di confidenza al 90% per la quantità del punto **(d)**.

17. Le cifre che seguono sono le medie annuali dei prezzi di tutti i libri recensiti dalla rivista *Science*, dal 1990 al 1996. Dai un intervallo che con il 95% di confidenza contenga la media dei prezzi di tutti i libri che sono stati recensiti nel 1997.

Anno	Prezzo medio (dollari)
1990	54.43
1991	54.08
1992	57.58
1993	51.21
1994	59.96
1995	60.52
1996	62.13

I Problemi dal 18 al 22 si riferiscono alla tabella di pagina seguente, che mette in relazione il livello di fumo con i tassi di morte per 4 tipi di tumore in 14 stati americani.

18. **(a)** Disegna il diagramma di dispersione dei decessi per tumore alla vescica rispetto ai consumi di sigarette.

(b) Diresti che è possibile che vi sia una relazione lineare?

(c) Trova il miglior fit lineare.

(d) Se il consumo medio pro capite in un certo stato fosse di 2 500 sigarette, quale sarebbe la tua previsione di decessi per questo tipo di cancro?

19. **(a)** Disegna il diagramma di dispersione dei decessi per cancro ai polmoni, in funzione del consumo di sigarette.

(b) Stima i parametri di regressione α e β.

(c) Verifica al 5% di significatività l'ipotesi che il consumo di sigarette non influisca sulla frequenza dei decessi per cancro ai polmoni.

(d) Qual è il p-dei-dati del test del punto **(c)**?

20. **(a)** Disegna il diagramma di dispersione dei decessi per cancro ai reni rispetto al consumo di sigarette.

(b) Stima la retta di regressione.

(c) Qual è il p-dei-dati del test che tale retta abbia pendenza nulla?

(d) Determina un intervallo di confidenza al 90% per il tasso medio di morte per cancro ai reni per gli stati in cui il consumo medio di sigarette per cittadino sia di 3 400 all'anno.

21. **(a)** Disegna il diagramma di dispersione dei decessi per leucemia rispetto al consumo di sigarette.

(b) Stima i coefficienti di regressione.

(c) Verifica l'ipotesi che non vi sia correlazione tra il tasso di morti per leucemia e il numero di sigarette fumate, ovvero che $\beta = 0$.

(d) Determina un intervallo di predizione al 90% per il tasso di morte per leucemia in uno stato in cui il consumo medio di sigarette per cittadino sia di 2 500 all'anno.

22. **(a)** Stima la varianza delle variabili dipendenti nei Problemi dal 18 al 21.

(b) Determina un intervallo di confidenza al 95% per la varianza nei dati sul cancro ai polmoni.

(c) Dividi i dati sul cancro ai polmoni in due parti, a seconda se il consumo di sigarette sia inferiore o superiore alle 2 300 unità. Assumi che per entrambi i gruppi di dati sussista un modello di regressione lineare. Come verificheresti l'ipotesi che nei due gruppi la varianza delle risposte sia la stessa?

(d) Effettua il test del punto **(c)** al 5% di significatività.

Fumo di sigarette e tassi di morte per cancro

		Decessi all'anno su 100 000 persone			
Stato	**Sigarette pro capite**	**Cancro alla vescica**	**Cancro ai polmoni**	**Cancro ai reni**	**Leucemia**
California	2 860	4.46	22.07	2.66	7.06
Idaho	2 010	3.08	13.58	2.46	6.62
Illinois	2 791	4.75	22.80	2.95	7.27
Indiana	2 618	4.09	20.30	2.81	7.00
Iowa	2 212	4.23	16.59	2.90	7.69
Kansas	2 184	2.91	16.84	2.88	7.42
Kentucky	2 344	2.86	17.71	2.13	6.41
Massachusetts	2 692	4.69	22.04	3.03	6.89
Minnesota	2 206	3.72	14.20	3.54	8.28
New York	2 914	5.30	25.02	3.10	7.23
Alaska	3 034	3.46	25.88	4.32	4.90
Nevada	4 240	6.54	23.03	2.85	6.67
Utah	1 400	3.31	12.01	2.20	6.71
Texas	2 257	3.21	20.74	2.69	7.02

23. Disegna i residui standardizzati per i dati del Problema 1. Cosa indica tale grafico riguardo alla nostra assunzione che sia valido un modello di regressione lineare?

24. Misurare direttamente il contenuto di proteine nei campioni di fegato richiede un procedimento lungo e difficile. Per questo motivo i laboratori di medicina fanno spesso uso della spettrofotometria, grazie al fatto che la luce assorbita dal campione è legata alla quantità di proteine presenti. La procedura di misurazione consiste nel preparare una sospensione del campione in acqua e registrarne l'assorbimento luminoso tramite uno spettrofotometro; essa è stata effettuata su 5 campioni con un contenuto di proteine noto, ottenendo i risultati seguenti.

Luce assorbita	0.44	0.82	1.20	1.61	1.83
Contenuto di proteine	2	16	30	46	55

(a) Calcola il coefficiente di determinazione.

(b) Ti sembra che questo sia un modo ragionevole di misurare le proteine nei campioni di fegato?

(c) Qual è la stima del contenuto di proteine se l'assorbimento di luce è 1.5?

(d) Determina un intervallo di predizione al 90% per la stima del punto (c).

25. Determinare la sollecitazione di taglio di un punto di saldatura è relativamente difficile: misurarne il diametro è molto più semplice. Sarebbe molto vantaggioso perciò se la prima grandezza potesse essere predetta da una misurazione della seconda. I dati trovati in una sperimentazione sono i seguenti.

Sollecitazione di taglio (psi)	Diametro della saldatura (10^{-4} pollici)
370	400
780	800
1 210	1 250
1 560	1 600
1 980	2 000
2 450	2 500
3 070	3 100
3 550	3 600
3 940	4 000
3 950	4 000

(a) Traccia il diagramma di dispersione.

(b) Determina gli stimatori dei minimi quadrati dei coefficienti di regressione.

(c) Verifica al 5% di significatività l'ipotesi che il coefficiente angolare della retta di regressione sia 1.

(d) Stima il valore atteso della sollecitazione di taglio quando il diametro è di 0.25 pollici.

(e) Trova un intervallo di predizione che contenga con il 95% di confidenza la sollecitazione di taglio di un punto di saldatura del diametro di 0.225 pollici.

(f) Traccia il grafico dei residui standardizzati.

(g) Il grafico ottenuto al punto **(f)** è in accordo con le assunzioni del modello?

26. Un produttore di viti vuole fornire ai suoi clienti dei dati sulla relazione tra lunghezze nominali ed effettive dei suoi prodotti. Vengono osservati i dati (in pollici) che sono riportati nella tabella alla fine del problema.

(a) Stima i coefficienti di regressione.

(b) Stima la varianza che risulta nella fabbricazione di una vite.

(c) Trova un intervallo di confidenza al 90% per la lunghezza media di un elevato numero di viti di 1 pollice nominale.

(d) Determina un intervallo di predizione al 90% per la lunghezza di una singola vite di 1 pollice nominale.

(e) Traccia il grafico dei residui standardizzati.

(f) Il grafico ottenuto al punto **(e)** fa sorgere qualche dubbio sul modello di regressione?

(g) Calcola il coefficiente di correlazione lineare.

Lunghezza nominale x	Lunghezza effettiva y		
$\frac{1}{4}$	0.262	0.262	0.245
$\frac{1}{2}$	0.496	0.512	0.490
$\frac{3}{4}$	0.743	0.744	0.751
1	0.976	1.010	1.004
$1\frac{1}{4}$	1.265	1.254	1.252
$1\frac{1}{2}$	1.498	1.518	1.504
$1\frac{3}{4}$	1.738	1.759	1.750
2	2.005	1.992	1.992

27. Il vetro gioca un ruolo importante nelle indagini criminali, infatti l'attività criminale finisce spesso col causare la rottura di finestre e altri oggetti di vetro, e siccome piccoli frammenti tendono a rimanere attaccati ai vestiti del colpevole, è fondamentale riuscire a identificare i diversi tipi di vetro e collegarli con il luogo del delitto. Due proprietà fisiche del vetro che sono utili per l'identificazione sono l'indice di rifrazione e la densità. Il primo è di facile misurazione, mentre il secondo è molto più complicato; siccome inoltre la misurazione esatta della densità è molto facilitata se si possiede almeno una sua buona stima prima di approntare l'esperimento, sarebbe piuttosto utile se si potesse impiegare l'indice di rifrazione per stimare l'altro parametro.

I dati seguenti mettono in relazione l'indice di rifrazione di 18 tipi di vetro con la loro densità.

Indice di rifrazione	Densità	Indice di rifrazione	Densità
1.5139	2.4801	1.5161	2.4843
1.5153	2.4819	1.5165	2.4858
1.5155	2.4791	1.5178	2.4950
1.5155	2.4796	1.5181	2.4922
1.5156	2.4773	1.5191	2.5035
1.5157	2.4811	1.5227	2.5086
1.5158	2.4765	1.5227	2.5117
1.5159	2.4781	1.5232	2.5146
1.5160	2.4909	1.5233	2.5187

(a) Predici la densità di un frammento di vetro che abbia un indice di rifrazione di 1.52.

(b) Determina un intervallo che con il 95% di confidenza contenga la densità cercata al punto **(a)**.

28. Il modello di regressione

$$Y = \beta x + e, \qquad e \sim \mathcal{N}(0, \sigma^2)$$

è detto regressione attraverso l'origine, perché suppone che la risposta media quando il livello di ingresso è $x = 0$ sia nulla. Supponi che (x_i, Y_i), per $i = 1, 2, \ldots, n$ sia un campione di coppie di dati provenienti da questo modello.

(a) Determina lo stimatore dei minimi quadrati B, di β.

(b) Qual è la distribuzione di B?

(c) Definisci SS_R e trova la sua distribuzione.

(d) Costruisci un test per verificare $H_0 : \beta = \beta_0$ di contro a $H_1 : \beta \neq \beta_0$.

(e) Determina un intervallo di predizione con un livello di confidenza di $1 - \gamma$ per $Y(x_0)$, la riposta al livello di ingresso x_0.

29. Dimostra l'identità seguente:

$$R^2 = \frac{S_{xY}^2}{S_{xx} S_{YY}}$$

30. La tabella che segue riporta il peso e la pressione sistolica per un campione casuale di 20 uomini americani di età compresa tra i 25 e 30 anni.

Soggetto	Peso (libbre)	Pressione	Soggetto	Peso (libbre)	Pressione
1	165	130	11	172	153
2	167	133	12	159	128
3	180	150	13	168	132
4	155	128	14	174	149
5	212	151	15	183	158
6	175	146	16	215	150
7	190	150	17	195	163
8	210	140	18	180	156
9	200	148	19	143	124
10	149	125	20	240	170

(a) Stima i coefficienti di regressione.

(b) Ti sembra che i dati supportino la tesi che la pressione del sangue non dipenda dal peso corporeo?

(c) Scelto un campione numeroso di soggetti del peso di 182 libbre, trova un intervallo che con il 95% di confidenza contenga la media delle loro pressioni sistoliche.

(d) Analizza i residui standardizzati.

(e) Determina il coefficiente di correlazione campionaria.

31. Si è determinato che la relazione tra la tensione T e il numero di cicli N prima di una rottura, per una particolare lega metallica è dato da

$$T = \frac{A}{N^m}$$

dove A e m sono costanti da determinare. Stimale, sapendo che una sperimentazione che è stata effettuata ha ottenuto i dati seguenti.

Tensione (1 000 psi)	55.0	50.5	43.5	42.5	42.0	41.0	35.7	34.5	33.0	32.0
N (milioni di cicli)	.223	.925	6.75	18.1	29.1	50.5	126	215	445	420

32. Nel 1957 l'ingegnere olandese J. R. DeJong propose un modello per il tempo necessario per svolgere una semplice operazione manuale, in funzione del numero di volte che era stata praticata. La formula era

$$T \approx ts^{-n}$$

dove T è il tempo necessario, n è il numero di volte che si è praticata l'operazione e t e s sono parametri che dipendono dal tipo di lavoro e dalla persona coinvolta. Stima t e s per il campione di dati seguente.

T	22.4	21.3	19.7	15.6	15.2	13.9	13.7
n	0	1	2	3	4	5	6

33. Il residuo di cloro in una piscina in diversi momenti successivi alla pulitura più recente è il seguente:

Tempo (ore)	2	4	6	8	10	12
Cloro (ppm)	1.8	1.5	1.45	1.42	1.38	1.36

(a) Interpola una relazione del tipo $Y \approx ae^{-bx}$

(b) Che residuo di cloro prevedi si avrà 15 ore dopo la pulitura?

34. La frazione di eccedenza termica che viene dissipata da un corpo dopo un tempo t da quando si rimuove la sorgente di calore, segue la legge

$$P = 1 - e^{-\alpha t}$$

per una opportuna costante α. Avendo a disposizione i dati

P	0.07	0.21	0.32	0.38	0.4	0.45	0.51
t	0.1	0.2	0.3	0.4	0.5	0.6	0.7

(a) stima il valore di α;

(b) stima il valore di t al quale risulta dissipata la metà dell'eccedenza termica.

35. I dati seguenti rappresentano la conta batterica nei campioni di sangue di 5 cavie in momenti diversi dopo un'inoculazione con batteri vitali.

Giorni	3	6	7	8	9
Conta batterica (migliaia)	121	134	147	210	330

(a) Interpola una curva.

(b) Stima la conta batterica per un'altra cavia dopo 8 giorni.

36. I dati seguenti rappresentano l'ammontare di idrogeno (in parti per milione) presente in trapanature del nucleo di una colata metallica sotto vuoto, a varie distanze dalla base.

Distanza	1	2	3	4	5	6	7	8	9	10
Idrogeno	1.28	1.50	1.12	0.94	0.82	0.75	0.60	0.72	0.95	1.20

(a) Disegna il diagramma di dispersione.

(b) Interpola questi dati con una curva della forma

$$Y = \alpha + \beta x + \gamma x^2 + e$$

37. Un nuovo farmaco per la cura dei tumori viene sperimentato su 10 topi da laboratorio, ciascuno dei quali presentava inizialmente una massa tumorale di 4 grammi. Dopo un trattamento a dosaggi differenti, si riscontrano le seguenti riduzioni delle masse tumorali:

Dose di farmaco	1	2	3	4	5	6	7	8	9	10
Riduzione tumore (g)	0.5	0.9	1.2	1.35	1.5	1.6	1.53	1.38	1.21	0.65

Usa un modello di regressione quadratico del tipo

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

per stimare la massima riduzione mediamente ottenibile, e il dosaggio di farmaco che la raggiunge.

38. Nella tabella seguente è riportato il numero di fusti trasportati in container che sono risultati danneggiati in occasione di impatti a diverse velocità.

Velocità	3	3	3	5	5	5	6	7	7	8
Fusti danneggiati	54	62	65	94	122	84	142	139	184	254

(a) Analizza i dati con un modello di regressione lineare semplice.

(b) Disegna il grafico dei residui standardizzati.

(c) Ti sembra che ciò che hai ottenuto al punto **(b)** indichi qualche difetto nel modello?

(d) Se la risposta al punto **(c)** è positiva, individua un modello migliore e stima i parametri corrispondenti.

39. Affronta nuovamente il Problema 5 sotto l'ipotesi che la varianza dell'incremento nella velocità di lettura sia proporzionale alle settimane di preparazione.

40. I dati che seguono sono stati generati con il modello

$$Y = 20 + 4x + e$$

dove e è normale con media 0 e varianza $15/(5 + x)$:

x	1	2	3	4	5	6	7	8	9	10
y	23.9	27.9	31	36.8	41.8	43.6	48	49.9	56	59.7

(a) Traccia un grafico dei dati.

(b) Interpola i dati con una retta usando il metodo dei minimi quadrati ordinario.

(c) Interpola i dati con il metodo dei minimi quadrati pesati.

(d) Traccia le due rette dei punti **(b)** e **(c)** in sovrapposizione ai dati.

41. I dati seguenti si riferiscono all'Esempio 9.8.3:

Autovetture (al giorno)	Incidenti (al mese)
2 000	15
2 300	27
2 500	20
2 600	21
2 800	31
3 000	16
3 100	22
3 400	23
3 700	40
3 800	39
4 000	27
4 600	43
4 800	53

(a) Stima il numero di incidenti al mese in un tratto di autostrada percorso da 3 500 autoveicoli al giorno.

(b) Usa il modello

$$\sqrt{Y} = \alpha + \beta x + e$$

e rispondi nuovamente al punto (a).

***42.** La portata massima dei fiumi è un parametro importante per molti problemi di progettazione. Per ottenere delle stime di tale valore, si possono utilizzare dati quali l'estensione (x_1) e la pendenza media (x_2) del bacino idrografico. Stima la relazione tra queste grandezze usando i dati seguenti (l'area del bacino è espressa in miglia quadrate, e la portata in piedi cubi al secondo).

x_1	36	37	45	87	450	550	1 200	4 000
x_2	0.005	0.04	0.004	0.002	0.004	0.001	0.002	0.0005
Portata massima	50	40	45	110	490	400	650	1 550

***43.** I sedimenti trasportati dai corsi d'acqua dipendono dalle dimensioni del bacino idrografico e dalla portata media. Stima la relazione esistente, usando i dati seguenti.

Bacino ($\times 1\,000$ mi^2)	Portata media (piedi cubi al secondo)	Sedimenti (milioni di tonnellate all'anno)
8	65	1.8
19	625	6.4
31	1 450	3.3
16	2 400	1.4
41	6 700	10.8
24	8 500	15.0
3	1 550	1.7
3	3 500	0.8
3	4 300	0.4
7	12 100	1.6

***44.** Stima i coefficienti di regressione lineare multipla per i dati seguenti.

x_1	1	2	3	4	5	6	7	8	9	10
x_2	11	10	9	8	7	6	5	4	3	2
x_3	16	9	4	1	2	1	4	9	16	25
x_4	4	3	2	1	1	−1	−2	−3	−4	−5
y	275	183	140	82	97	122	146	246	359	482

***45.** I dati che seguono si riferiscono ad alcuni trapianti di cuore eseguiti a Stanford. In particolare vi compaiono il tempo di sopravvivenza (in giorni), il *mismatch score*, che è un indicatore dell'incompatibilità fisiologica tra donatore e ricevente, e l'età del ricevente.

Giorni di sopravvivenza	Mismatch score	Età
624	1.32	51.0
46	0.61	42.5
64	1.89	54.6
1 350	0.87	54.1
280	1.12	49.5
10	2.76	55.3
1 024	1.13	43.4
39	1.38	42.8
730	0.96	58.4
136	1.62	52.0
836	1.58	45.0
60	0.69	64.5

(a) Usando come variabile dipendente il logaritmo del tempo di sopravvivenza, interpola un modello di regressione lineare multipla sulle variabili indipendenti costituite dal mismatch score e dall'età.

(b) Stima la varianza del termine di errore.

***46.** (a) Stima l'equazione di regressione lineare multipla per i dati seguenti.

(b) Verifica l'ipotesi che $\beta_0 = 0$.

(c) Verifica l'ipotesi che $\beta_3 = 0$.

(d) Verifica l'ipotesi che sia di 8.5 la risposta media ai livelli di ingresso $x_1 = x_2 = x_3 = 1$.

x_1	x_2	x_3	y
7.1	0.68	4	41.53
9.9	0.64	1	63.75
3.6	0.58	1	16.38
9.3	0.21	3	45.54
2.3	0.89	5	15.52
4.6	0.00	8	28.55
0.2	0.37	5	5.65
5.4	0.11	3	25.02
8.2	0.87	4	52.49
7.1	0.00	6	38.05
4.7	0.76	0	30.76
5.4	0.87	8	39.69
1.7	0.52	1	17.59
1.9	0.31	3	13.22
9.2	0.19	5	50.98

***47.** La resistenza alla trazione riscontrata in un certo tipo di fibra sintetica sembra essere legata alla percentuale di cotone nella fibra e al tempo di asciugatura della fibra stessa. Una sperimentazione su 10 esemplari prodotti in condizioni differenti ha dato i risultati qui sotto:

Resistenza alla trazione	213	220	216	225	235	218	239	243	233	240
Percentuale di cotone	13	15	14	18	19	20	22	17	16	18
Tempo di asciugatura	2.1	2.3	2.2	2.5	3.2	2.4	3.4	4.1	2.0	4.3

(a) Interpola i dati con una equazione di regressione multipla.

(b) Determina un intervallo di confidenza al 90% per la resistenza media alla trazione di una fibra sintetica con il 21% di cotone e il cui tempo di asciugatura sia stato pari a 3.6.

***48.** I minuti di funzionamento senza guasti y di un componente di una macchina sono legati al voltaggio di funzionamento x_1, alla velocità del motore (in giri al minuto) x_2, e alla temperatura di funzionamento x_3. Nel reparto di ricerca e sviluppo si realizzano una serie di esperimenti, ottenendo i dati seguenti.

y	2 145	2 155	2 220	2 225	2 260	2 266	2 334	2 340	2 212	2 180
x_1	110	110	110	110	120	120	120	130	115	115
x_2	750	850	1 000	1 100	750	850	1 000	1 000	840	880
x_3	140	180	140	180	140	180	140	180	150	150

(a) Trova il fit lineare multiplo per questi dati.

(b) Stima la varianza dell'errore.

(c) Determina un intervallo di confidenza al 95% per la media del tempo di funzionamento ad una tensione di 125 volt, una velocità di 900 giri al minuto e una temperatura di 160 gradi Fahrenheit.

49. Spiega perché, mantenendo gli stessi dati, ogni intervallo di predizione di una risposta futura contiene il corrispondente intervallo di confidenza della risposta media.

***50.** Considera il seguente campione di dati.

x_1	x_2	y
5.1	2	55.42
5.4	8	100.21
5.9	−2	27.07
6.6	12	169.95
7.5	−6	−17.93
8.6	16	197.77
9.9	−10	−25.66
11.4	20	264.18
13.1	−14	−53.88
15.0	24	317.84
17.1	−18	−72.53
19.4	28	385.53

(a) Interpola una relazione lineare tra y e x_i.

(b) Calcola la varianza del termine di errore.

(c) Determina un intervallo che con il 95% di confidenza contenga la risposta che si otterrebbe con ingressi $x_1 = 10.2$ e $x_2 = 17$.

***51.** Il costo di produzione energetica per kilowatt-ora è una funzione del fattore di carico e del costo del carbone in centesimi di dollaro per milione di Btu. I dati seguenti sono stati ottenuti da 12 centrali.

Fattore di carico	84	81	73	74	67	87	77	76	69	82	90	88
Costo del carbone	14	16	22	24	20	29	26	15	29	24	25	13
Costo energetico	4.1	4.4	5.6	5.1	5.0	5.3	5.4	4.8	6.1	5.5	4.7	3.9

(a) Stima l'equazione di regressione.

(b) Verifica l'ipotesi che il coefficiente del fattore di carico sia nullo.

(c) Determina un intervallo di predizione al 95% per il costo di produzione dell'energia quando il fattore di carico sia 85 e il costo del carbone 20.

***52.** I dati seguenti mettono in relazione la pressione sistolica di un gruppo di individui con la loro età e il loro peso. I soggetti dell'esperimento hanno stili di vita e corporature simili.

Età	25	25	42	55	30	40	66	60	38
Peso (libbre)	162	184	166	150	192	155	184	202	174
Pressione	112	144	138	145	152	110	118	160	108

(a) Verifica l'ipotesi che, conoscendo il peso di un individuo, la sua età non dia informazioni ulteriori nel predirne la pressione.

(b) Determina un intervallo che, con il 95% di confidenza, contenga la media delle pressioni di tutti gli individui (simili ai precedenti) di 45 anni che pesano 180 libbre.

(c) Determina un intervallo che, con il 95% di confidenza, contenga la pressione di una persona di 45 anni che pesa 180 libbre.

***53.** Uno studio completato di recente ha tentato di mettere in relazione la soddisfazione nel lavoro con il reddito annuale (in migliaia di dollari) e l'anzianità, di un campione di 9 dipendenti municipali. La soddisfazione per il proprio impiego (in una scala da 1 a 10) è il valore dichiarato dai singoli soggetti:

Reddito annuale	27	22	34	28	36	39	33	42	46
Anni in quell'impiego	8	4	12	9	16	14	10	15	22
Soddisfazione	5.6	6.3	6.8	6.7	7.0	7.7	7.0	8.0	7.8

(a) Stima i parametri di regressione.

(b) Che considerazioni qualitative puoi trarre su come cambia il valore di soddisfazione quando si aumentarno gli anni di servizio tenendo fisso il reddito?

(c) Predici la soddisfazione nel suo lavoro di un impiegato assunto da 5 anni con un reddito di 31 000 dollari.

***54.** Considera il Problema 53 senza i dati sul reddito; supponi quindi che la soddisfazione nel lavoro sia legata solamente agli anni di servizio.

(a) Stima i parametri di regressione α e β.

(b) Qual è la relazione qualitativa tra le due variabili? In altre parole, come sembra cambiare la soddisfazione all'aumentare dell'anzianità di servizio?

(c) Confronta le due risposte date ai punti (b) di questo problema e del 53.

(d) Commenta il risultato del punto (c). Che conclusioni se ne devono trarre?

10 Analisi della varianza

Contenuto

10.1 Introduzione

Una società molto grande sta valutando l'acquisto in quantità di un pacchetto software per insegnare un nuovo linguaggio di programmazione. Sono disponibili sul mercato quattro prodotti differenti, che alcuni personaggi influenti all'interno dell'azienda ritengono essere sostanzialmente equivalenti, nel senso che la scelta di uno piuttosto di un altro non avrà una apprezzabile influenza sul livello di apprendimento dell'utente. Per verificare questa ipotesi si scelgono 160 ingegneri, che vengono divisi in 4 gruppi di 40, e si assegna a ogni gruppo un pacchetto differente per imparare il linguaggio di programmazione in questione. Alla fine del periodo di studio, si sottopongono gli ingegneri ad un esame molto approfondito, e si desidera utilizzare i risultati ottenuti per stabilire se davvero i pacchetti fossero equivalenti. Come si può effettuare questa analisi?

La prima cosa da notare è che quando i punteggi medi dei quattro gruppi di ingegneri sono molto simili, è auspicabile concludere che i pacchetti siano interscambiabili, mentre quando i quattro valori sono troppo distanti dovrà essere possibile rifiutare questa l'ipotesi. Affinché questi ragionamenti siano validi è però necessario fare molta attenzione al criterio con cui formiamo i gruppi. Infatti nel caso che i membri di un gruppo realizzino punteggi decisamente più alti dei colleghi, cosa abbiamo dimostrato? È il pacchetto software utilizzato a essere migliore, o sono i

soggetti dell'esperimento che sono più capaci degli altri? Per escludere la seconda alternativa occorre che la suddivisione sia fatta in modo tale da rendere estremamente improbabile che si formi una concentrazione di elementi migliori o peggiori in un gruppo. Il metodo che si è appurato essere più indicato per queste finalità è la formazione dei gruppi in modo assolutamente casuale, vale a dire scegliendo con pari probabilità una qualsiasi suddivisione tra tutte quelle possibili[1].

Quando la suddivisione in gruppi sia casuale, è probabilmente ragionevole supporre che (1) i punteggi dei singoli soggetti all'esame finale siano variabili aleatorie normali e indipendenti; (2) i parametri di tali distribuzioni dipendano solo dal pacchetto software utilizzato, e anzi, mentre le medie μ_1, μ_2, μ_3 e μ_4 possono effettivamente cambiare da un pacchetto all'altro, si può supporre che la varianza delle distribuzioni sia dovuta alla variabilità nell'apprendimento delle persone, e quindi sia una costante (incognita) σ^2. Si denota quindi con X_{ij}, per $i = 1,2,3,4$ e $j = 1,2,\ldots,40$ il punteggio totalizzato dal membro j-esimo del gruppo i, e le X_{ij} si suppongono essere indipendenti e avere distribuzione normale di parametri incogniti μ_i e σ^2. L'ipotesi in esame, che i pacchetti siano equivalenti, si scrive allora come $\mu_1 = \mu_2 = \mu_3 = \mu_4$.

In questo capitolo presentiamo una tecnica che può essere usata per verificare tale ipotesi. Essa si rivela molto generale, e può essere impiegata per fare inferenze su un gran numero di parametri legati alle medie delle popolazioni. Tale tecnica prende il nome di *analisi della varianza*[2].

10.2 Lo schema generale

La verifica delle ipotesi sulle medie di *due* distribuzioni normali è stata affrontata nel Capitolo 8; qui ci occupiamo del caso generale in cui il numero di distribuzioni da confrontare sia arbitrario. Nella Sezione 10.3 studiamo il caso in cui si dispone di m campioni provenienti da popolazioni diverse, ciascuno di n elementi, e usiamo questi dati per verificare l'ipotesi che le m medie di popolazione siano tutte uguali. Poiché la media di queste variabili aleatorie dipende da un solo fattore, vale a dire la popolazione $1,2,\ldots,m$ da cui sono estratte, questo ambito prende il nome di *analisi della varianza a una via* (o anche *one-way*). Nella Sezione 10.3.1 presentiamo una tecnica per confrontare contemporaneamente tutte le $\binom{m}{2}$ coppie (μ_i, μ_j) di medie delle diverse popolazioni, per poter dire qualcosa di più, specialmente quando si rifiuta l'ipotesi

[1] Non è affatto ovvio come realizzare una tale scelta casuale, comunque una procedura che si rivela molto efficiente consiste nel numerare da 1 a 160 i soggetti dell'esperimento, generare una permutazione casuale degli interi $1,2,\ldots,160$, e infine mettere nel primo gruppo gli ingegneri i cui numeri occupano le prime 40 posizioni, nel secondo gruppo quelli delle ulteriori 40 posizioni e così via.

[2] In inglese è detta *analysis of variance*, da cui l'usatissimo acronimo *ANOVA*, [*N.d.T.*]

che siano tutte uguali. Nella Sezione 10.3.2 illustriamo come procedere quando gli m campioni non hanno tutti la stessa numerosità.

Nelle Sezioni 10.4 e 10.5 consideriamo dei modelli in cui vi siano due fattori che determinano la media delle variabili aleatorie. Queste ultime si immaginano costituire una matrice, e il valore atteso di ogni elemento si assume dipendere sia dalla riga sia dalla colonna a cui appartiene. Questo modello prende il nome di *analisi della varianza a due vie* (oppure *two-way*). L'ipotesi di lavoro più semplice è che la media delle variabili aleatorie dipenda dalla riga e dalla colonna in modo additivo, e quindi che il valore atteso di X_{ij} assuma la forma $\mu + \alpha_i + \beta_j$. In questo caso sviluppiamo stimatori dei parametri (Sezione 10.4) e costruiamo i test per verificare l'ipotesi che o la riga o la colonna non influiscano in realtà sulle medie (Sezione 10.5). Nella Sezione 10.6 abbandoniamo l'assunzione di additività e ci mettiamo nel caso in cui la media delle variabili aleatorie dipenda in maniera anche nonlineare dalla riga e dalla colonna in cui si trova; si rende possibile così la presenza di *interazioni* tra i due fattori. Mostriamo come verificare l'ipotesi che non vi siano interazioni, come pure quella che non vi sia effetto di riga oppure di colonna.

In tutti i modelli considerati in questo capitolo assumiamo che i dati abbiano distribuzione normalecon la medesima varianza σ^2, che non si suppone nota. Per verificare una ipotesi nulla H_0 riguardante dei parametri legati alle medie delle popolazioni, l'approccio dell'analisi della varianza si basa sul confronto di due stimatori di σ^2. Tali stimatori sono costruiti in modo che il primo sia valido indipendentemente dalla correttezza di H_0, mentre il secondo si comporta bene solo nel caso che H_0 sia vera, e altrimenti tende ad errare per eccesso. I test vengono perciò costruiti in base al principio che l'ipotesi nulla va rifiutata se il rapporto tra il secondo stimatore e il primo è troppo alto. In altre parole, siccome i due stimatori dovrebbero essere vicini quando H_0 è valida (infatti in quel caso entrambi stimano σ^2), è naturale rifiutare l'ipotesi nulla quando essi non sono affatto vicini.

Gli stimatori di σ^2 che esibiremo fanno uso di un importante proprietà – sulla quale ora ci soffermiamo – delle distribuzioni chi-quadro. Siano $X_1, X_2, \ldots, X_N$ delle variabili aleatorie normali indipendenti con medie eventualmente diverse $\mu_1, \mu_2, \ldots, \mu_N$, e varianza in comune σ^2. Poiché le variabili aleatorie

$$Z_i := \frac{X_i - \mu_i}{\sigma}, \qquad i = 1,2,\ldots,N$$

sono normali standard, segue dalla definizione della distribuzione chi-quadro che

$$\sum_{i=1}^{N} Z_i^2 = \sum_{i=1}^{N} \frac{(X_i - \mu_i)^2}{\sigma^2} \sim \chi^2_N$$

è una chi-quadro con N gradi di libertà. Supponiamo ora di non stimare direttamente le μ_i, ma usare il fatto che esse sono combinazioni lineari di k parametri incogniti,

i quali possono essere stimati; costruendo le medesime combinazioni lineari con gli stimatori dei parametri, si determinano degli stimatori $\widehat{\mu}_i$ per le medie vere μ_i, per $i = 1, 2, \ldots, N$. In queste ipotesi è possibile dimostrare che

$$\sum_{i=1}^{N} \frac{(X_i - \widehat{\mu}_i)^2}{\sigma^2} \sim \chi^2_{N-k}$$

In altre parole, si comincia notando che

$$\sum_{i=1}^{N} \frac{(X_i - E[X_i])^2}{\sigma^2} \sim \chi^2_N$$

Se si scrive ciascuna delle $E[X_i]$ come combinazione lineare dei k parametri e quindi si sostituiscono questi ultimi con gli stimatori corrispondenti, l'espressione risultante ha ancora distribuzione chi-quadro, ma i gradi di libertà vanno diminuiti di uno per ogni parametro che viene sostituito col suo stimatore.

Per dare un esempio di questo comportamento, si consideri il caso in cui tutte le medie sono uguali, ovvero

$$E[X_i] = \mu, \qquad i = 1, 2, \ldots, N$$

Prendiamo μ come unico parametro da stimare, così che $k = 1$. Se sostituiamo μ con $\overline{X}$ che è il suo stimatore, troviamo quella che era l'espressione di $(N-1)S^2/\sigma^2$:

$$\sum_{i=1}^{N} \frac{(X_i - \overline{X})^2}{\sigma^2} = \frac{N-1}{\sigma^2} \cdot \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \overline{X})^2 = \frac{S^2}{\sigma^2}(N-1)$$

che sappiamo dalla Sezione 6.5.2 avere distribuzione chi-quadro con $N-1$ gradi di libertà in accordo con il risultato generale enunciato poco fa.

10.3 Analisi della varianza ad una via

Consideriamo m campioni indipendenti, ciascuno formato da n variabili aleatorie normali con media che dipende dal campione e varianza fissata. Denotiamo tali dati con X_{ij}, dove $i = 1, \ldots, m$ indica il campione e $j = 1, \ldots, n$ indica la posizione all'interno del campione stesso. L'ipotesi di gaussianità appena espressa si riformula in questi termini:

$$X_{ij} \sim \mathcal{N}\left(\mu_i, \sigma^2\right), \qquad i = 1, \ldots, m, \qquad j = 1, \ldots, n$$

dove i parametri $\mu_1, \mu_2, \ldots, \mu_m$ e σ sono incogniti. Il nostro obiettivo è la verifica dell'ipotesi nulla

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_m$$

di contro all'ipotesi alternativa H_1 che non tutte le medie siano identiche. Una situazione pratica che può illustrare questo modello si ha quando disponiamo di m trattamenti diversi, e il risultato dell'applicazione del trattamento i ad un oggetto è una variabile aleatoria $\mathcal{N}\left(\mu_i, \sigma^2\right)$. Applichiamo ciascun trattamento a n oggetti diversi e alla fine vogliamo stabilire se è vero o no che tutti i trattamenti hanno (mediamente) lo stesso effetto.

Siccome vi sono in tutto nm variabili aleatorie normali e indipendenti, la somma dei quadrati delle loro versioni standardizzate avrà distribuzione chi-quadro con nm gradi di libertà:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(X_{ij} - E[X_{ij}])^2}{\sigma^2} = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(X_{ij} - \mu_i)^2}{\sigma^2} \sim \chi^2_{nm} \tag{10.3.1}$$

Come stimatori degli m parametri incogniti $\mu_1, \mu_2, \ldots, \mu_m$, usiamo le medie campionarie dei singoli campioni di dati; in particolare X_{i*} denoterà quella del campione i-esimo:

$$X_{i*} := \frac{1}{n} \sum_{j=1}^{n} X_{ij} \tag{10.3.2}$$

Siccome X_{i*} è uno stimatore di μ_i, per $i = 1, 2, \ldots, m$, se li sostituiamo tutti al posto dei parametri nell'Equazione (10.3.1), l'espressione che otteniamo,

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(X_{ij} - X_{i*})^2}{\sigma^2} = \frac{SS_{\mathrm{W}}}{\sigma^2} \sim \chi^2_{nm-m} \tag{10.3.3}$$

rappresenta una chi-quadro con $nm - m$ gradi di libertà. (Si ricordi che si perde un grado di libertà per ogni parametro sostituito da un suo stimatore). Nella precedente si è posto

$$SS_{\mathrm{W}} := \sum_{i=1}^{m} \sum_{j=1}^{n} (X_{ij} - X_{i*})^2 \tag{10.3.4}$$

Poiché il valore atteso di una variabile aleatoria chi-quadro è il numero dei suoi gradi di libertà, calcolando la media di SS_{W} si ha che

$$E\left[\frac{SS_{\mathrm{W}}}{\sigma^2}\right] = nm - m \qquad \text{ovvero} \qquad E\left[\frac{SS_{\mathrm{W}}}{nm - m}\right] = \sigma^2$$

Abbiamo così trovato il primo stimatore di σ^2, ovvero $SS_{\mathrm{W}}/(nm - m)$. Si noti che fino a qui non abbiamo dovuto supporre che H_0 fosse vera o meno.

Definizione 10.3.1. La statistica SS_{W} definita nell'Equazione (10.3.4) è chiamata somma dei quadrati[3] *entro* i campioni (*within*), perché si ottiene sostituendo al posto

[3] Useremo spesso in questo capitolo le somme dei quadrati degli scarti tra un certo numero di valori e la loro media aritmetica. Queste quantità, che sono evidentemente molto vicine a varianze campionarie, vengono a volte dette *devianze*, [*N.d.T.*]

delle medie di popolazione gli stimatori calcolati entro ogni campione. La statistica

$$\frac{SS_W}{nm-m}$$

è uno stimatore corretto di σ^2

Il secondo stimatore di σ^2 deve essere valido solo nel caso che l'ipotesi nulla sia vera. Assumiamo allora H_0 e quindi che tutte le medie siano uguali, ovvero $\mu_i = \mu$, per tutti gli indici i. Sotto questa ipotesi tutti gli stimatori $X_{1*}, X_{2*}, \dots X_{m*}$ sono normali di media μ e varianza σ^2/n, quindi la somma dei quadrati delle loro versioni normalizzate è una chi-quadro con m gradi di libertà:

$$\sum_{i=1}^{m} \frac{(X_{i*} - E[X_{i*}])^2}{\mathrm{Var}(X_{i*})} = \sum_{i=1}^{m} \frac{(X_{i*} - \mu)^2}{\sigma^2/n} \sim \chi^2_m \tag{10.3.5}$$

Ci occorre uno stimatore di μ, ed avendo tutti i dati valore atteso μ, la loro media campionaria costituisce la scelta migliore, perciò lo stimatore è dato da

$$X_{**} := \frac{1}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij} = \frac{1}{m} \sum_{i=1}^{m} X_{i*} \tag{10.3.6}$$

Se ora sostituiamo μ con X_{**} nell'Equazione (10.3.5), la quantità ottenuta ha distribuzione χ^2_{m-1}, quando H_0 è vera:

$$\sum_{i=1}^{m} \frac{(X_{i*} - X_{**})^2}{\sigma^2/n} = \frac{SS_b}{\sigma^2} \sim \chi^2_{m-1} \tag{10.3.7}$$

dove si è posto

$$SS_b := n \sum_{i=1}^{m} (X_{i*} - X_{**})^2 \tag{10.3.8}$$

Di conseguenza, quando H_0 è vera,

$$E\left[\frac{SS_b}{\sigma^2}\right] = m-1 \qquad \text{ovvero} \qquad E\left[\frac{SS_b}{m-1}\right] = \sigma^2$$

Definizione 10.3.2. La statistica SS_b definita nell'Equazione (10.3.8) è chiamata somma dei quadrati *tra* i campioni (*between*). Quando H_0 è valida, la statistica

$$\frac{SS_b}{m-1}$$

è uno stimatore corretto di σ^2.

Riassumendo, fino a qui abbiamo provato che:

$$\frac{SS_W}{nm-m} \quad \text{stima } \sigma^2 \text{ in ogni caso.}$$
$$\frac{SS_b}{m-1} \quad \text{stima } \sigma^2 \text{ se } H_0 \text{ è vera.}$$

Siccome si può anche dimostrare che, quando H_0 non è vera, il secondo stimatore tende a superare σ^2, è naturale usare come statistica del test l'espressione

$$D_{ts} := \frac{SS_b/(m-1)}{SS_W/(nm-m)} \tag{10.3.9}$$

e rifiutare l'ipotesi nulla quando D_{ts} è abbastanza grande.

Per quantificare questo valore sfruttiamo un altro importante risultato che non dimostriamo: quando H_0 è vera, SS_W e SS_b sono indipendenti, e quindi D_{ts} ha distribuzione$F$ con $m-1$ gradi di libertà al numeratore e $nm-m$ al denominatore. Denotiamo come usuale con $F_{m-1,nm-m}$ una variabile aleatoria di questo tipo, e per ogni $\alpha \in (0,1)$ definiamo $F_{\alpha,m-1,nm-m}$ in modo che valga

$$P(F_{m-1,nm-m} > F_{\alpha,m-1,nm-m}) = \alpha$$

Con questa notazione un test ad un livello α di significatività deve

$$\begin{aligned} &\text{rifiutare } H_0 \text{ se } \frac{SS_b/(m-1)}{SS_W/(nm-m)} > F_{\alpha,m-1,nm-m} \\ &\text{accettare } H_0 \text{ se } \frac{SS_b/(m-1)}{SS_W/(nm-m)} \leq F_{\alpha,m-1,nm-m} \end{aligned} \tag{10.3.10}$$

La Tabella A.4 in Appendice riporta il valore di $F_{\alpha,n,m}$ per $\alpha = 0.05$ e per diverse scelte di n e m. Una parte di quei valori è presentata anche nella Tabella 10.1, che ad esempio ci dice che vi è una probabilità del 5% che una F di Fisher con 3 gradi di libertà al numeratore e 10 al denominatore superi 3.71.

Tabella 10.1 Valori di $F_{0.05,n,m}$, dove n è il numero di gradi di libertà del numeratore, e m del denominatore

		n			
		1	**2**	**3**	**4**
m	**4**	7.71	6.94	6.59	6.39
	5	6.61	5.79	5.41	5.19
	10	4.96	4.10	3.71	3.48

Un metodo alternativo per verificare l'ipotesi che tutte le medie siano uguali consiste nel calcolare il p-dei-dati e confrontarlo con il livello di significatività desiderato. Se v denota il valore assunto dalla statistica del test, allora il p-dei-dati vale

$$p\text{-dei-dati} = P(F_{m-1,nm-m} \geq v) \tag{10.3.11}$$

e può essere ad esempio calcolato con il Programma 10.3, che fornisce anche il valore della statistica D_{ts}.

Esempio 10.3.1. Una azienda di noleggio auto vuole valutare l'efficienza di 3 tipi diversi di benzina. Predispone 15 auto identiche per viaggiare a una stessa velocità fissata, e mette 10 galloni di carburante in ciascun serbatoio, dividendo le auto in 3 gruppi da 5. I dati seguenti sono le miglia percorse fino all'esaurimento di tutto il carburante.

Tipo 1	220	251	226	246	260
Tipo 2	244	235	232	242	225
Tipo 3	252	272	250	238	256

Si verifichi l'ipotesi che l'autonomia media ottenuta non dipenda dal tipo di carburante. Si usi il 5% di significatività.

Eseguiamo il Programma 10.3 ottenendo i risultati della Figura 10.1. Siccome il p-dei-dati è maggiore di 0.05 non possiamo escludere l'ipotesi che i tre tipi di benzina siano equivalenti. □

Nel caso si svolgano i calcoli a mano, è utile la seguente identità algebrica.

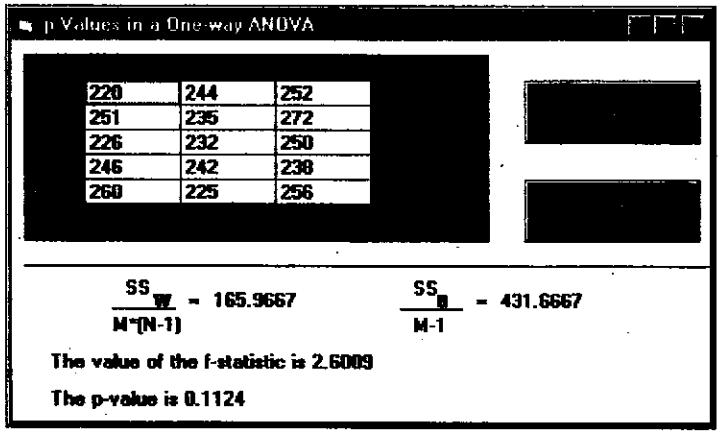


Figura 10.1

Proposizione 10.3.1 (Identità delle somme dei quadrati). Siano dati nm numeri $X_{ij}, i = 1, \ldots, m, j = 1, \ldots, n$, e si definiscano le grandezze SS_W, X_{**} e SS_b come nelle Equazioni (10.3.4), (10.3.6) e (10.3.8) delle pagine precedenti. Allora

$$\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2 = nmX_{**}^2 + SS_b + SS_W \tag{10.3.12}$$

Nello svolgere i conti a mano, conviene calcolare nell'ordine

$$X_{i*} = \frac{1}{n}\sum_{j=1}^{n} X_{ij}, \qquad i = 1, \ldots, m$$

$$X_{**} = \frac{1}{m}\sum_{i=1}^{m} X_{i*}$$

$$SS_b = n\sum_{i=1}^{m}(X_{i*} - X_{**})^2$$

si deve poi trovare il valore di $\sum_i \sum_j X_{ij}^2$, e infine

$$SS_W = \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}^2 - nmX_{**}^2 - SS_b$$

Esempio 10.3.2. Eseguiamo a mano i calcoli dell'Esempio 10.3.1. Per prima cosa notiamo che sottrarre una costante a tutti i dati non cambia il valore della statistica del test. Decidiamo allora di sottrarre 220, ottenendo i risultati seguenti:

Tipo	Autonomia					$\sum_j X_{ij}$	$\sum_j X_{ij}^2$
1	0	31	6	26	40	103	3 273
2	24	15	12	22	5	78	1 454
3	32	52	30	18	36	168	6 248

Per cui

$$X_{1*} = 103/5 = 20.6 \qquad X_{2*} = 78/5 = 15.6 \qquad X_{3*} = 168/5 = 33.6$$

$$X_{**} = (X_{1*} + X_{2*} + X_{3*})/3 \approx 23.267, \qquad X_{**}^2 \approx 541.334$$

$$SS_b \approx 5[(20.6 - 23.267)^2 + (15.6 - 23.267)^2 + (33.6 - 23.267)^2] \approx 863.33$$

$$\sum_i \sum_j X_{ij}^2 = 3\,273 + 1\,454 + 6\,248 = 10\,975$$

e infine

$$SS_W \approx 10\,975 - 15 \times 541.334 - 863.33 \approx 1991.6$$

Di conseguenza la statistica rilevante risulta pari a

$$D_{\text{ts}} = \frac{863.33/2}{1991.6/12} \approx 2.60$$

A questo punto, consultando la Tabella A.4 in Appendice otteniamo che $F_{0.05,2,12} \approx 3.89$, che essendo maggiore di 2.60 non ci autorizza a rifiutare l'ipotesi nulla al 5% di significatività. □

La Tabella 10.2 riassume i risultati di questa sezione.

Tabella 10.2 ANOVA a una via.

Variazione	**Somma di quadrati**	**Gradi di libertà**
Tra i campioni	$SS_{\text{b}} := n\sum_i (X_{i*} - X_{**})^2$	$m-1$
Entro i campioni	$SS_{\text{W}} := \sum_i \sum_j (X_{ij} - X_{i*})^2$	$nm-m$

Ipotesi nulla	**Statistica del test**	**Un test con significatività α deve**	***p*-dei-dati se $D_{\text{ts}} = v$**
Tutte le μ_i uguali	$D_{\text{ts}} := \dfrac{SS_{\text{b}}/(m-1)}{SS_{\text{W}}/(nm-m)}$	rifiutare H_0 se $D_{\text{ts}} > F_{\alpha,m-1,nm-m}$	$P(F_{m-1,nm-m} \geq v)$

10.3.1 Confronti multipli delle medie

Quando rifiutamo l'ipotesi nulla che le medie delle popolazioni siano uguali, vorremmo spingerci oltre e poter confrontare $\mu_1, \mu_2, \ldots, \mu_m$, ad esempio per dire qual è la popolazione con la media più elevata. Una procedura che permette di compiere questa analisi è il cosiddetto metodo T di Tukey. Esso, per un qualunque valore $0 < \alpha < 1$ fornisce intervalli di confidenza congiunti per le $\binom{m}{2}$ possibili differenze $\mu_i - \mu_j$, con $1 \leq i < j \leq m$, nel senso che vi è una probabilità di $1-\alpha$ che tutte le differenze contemporaneamente appartengano ai rispettivi intervalli. Il metodo T è infatti basato sul risultato seguente:

Proposizione 10.3.2. Per ogni scelta degli indici i, j diversi tra loro, e per ogni $\alpha \in (0,1)$, con probabilità $1-\alpha$,

$$X_{i*} - X_{j*} - W < \mu_i - \mu_j < X_{i*} - X_{j*} + W \tag{10.3.13}$$

dove si è posto

$$W := \frac{1}{\sqrt{n}} C(m, nm-m, \alpha)\sqrt{SS_{\text{W}}/(nm-m)} \tag{10.3.14}$$

I valori dei coefficienti $C(m,d,\alpha)$ per $\alpha = 0.01$ e $\alpha = 0.05$ sono riportati nella Tabella A.5 in Appendice.

Esempio 10.3.3. Il direttore di un college si domanda se vi sia differenza nel livello di preparazione degli studenti del primo anno provenienti da 3 diverse scuole superiori. Scelti 4 studenti a caso da ciascuna scuola, se ne confrontano le medie alla fine del primo anno di università (i dati sono riportati nella tabella qui sotto). Al 5% di significatività si rifiuta o si accetta l'ipotesi che le tre scuole superiori siano equivalenti? Nel caso di un rifiuto, si determinino degli intervalli di confidenza al 95% per le differenze dei punteggi medi degli studenti provenienti dalle diverse scuole.

Scuola 1	3.2,	3.4,	3.3,	3.5
Scuola 2	3.4,	3.0,	3.7,	3.3
Scuola 3	2.8,	2.6,	3.0,	2.7

Notiamo intanto che $m = 3$ e $n = 4$; eseguiamo quindi il Programma 10.3, che ci fornisce i seguenti valori:

$$SS_{\text{W}}/9 \approx 0.0431, \qquad p\text{-dei-dati} \approx 0.0046$$

quindi l'ipotesi che i punteggi medi degli studenti delle diverse scuole superiori siano gli stessi va rifiutata decisamente.

Per determinare gli intervalli di confidenza congiunti, notiamo intanto che

$$X_{1*} \approx 3.350, \qquad X_{2*} \approx 3.350, \qquad X_{3*} \approx 2.775$$

Dalla Tabella A.5 in Appendice ricaviamo che $C(3,9,0.05) \approx 3.95$, e quindi $W \approx \frac{1}{\sqrt{4}} 3.95 \cdot \sqrt{0.431} \approx 0.410$. Gli intervalli di confidenza al 95% sono allora

$$\begin{aligned} -0.410 < \mu_1 - \mu_2 < 0.410 \\ 0.165 < \mu_1 - \mu_3 < 0.985 \\ 0.165 < \mu_2 - \mu_3 < 0.985 \end{aligned}$$

Possiamo concludere, con il 95% di confidenza, che la media dei punteggi di fine anno per le matricole provenienti dalla scuola 3, è inferiore a quella delle altre due per un ammontare di punti tra 0.165 e 0.985, e che la differenza tra quelle delle scuole 1 e 2 è inferiore a 0.410 punti. □

10.3.2 Campioni con numerosità diverse

Fino a qui abbiamo sempre supposto di disporre di m campioni ciascuno dei quali con lo stesso numero n di elementi. Anche se questa situazione è certamente preferibile (si veda l'Osservazione 10.3.1 alla fine della sezione), non è sempre possibile ottenerla. Vediamo allora come modificare l'analisi della varianza a una via nel caso che gli m campioni abbiano numerosità $n_1, n_2, \ldots, n_m$. Denotiamo ancora i dati con X_{ij}, questa volta con $i = 1, \ldots, m$ e $j = 1, 2, \ldots, n_i$, e supponiamo che $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$. Siamo interessati all'ipotesi H_0 che tutte le medie siano uguali.

In primo luogo notiamo che

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{(X_{ij}-E[X_{ij}])^2}{\sigma^2}=\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{(X_{ij}-\mu_i)^2}{\sigma^2}\sim\chi^2_N$$

ha distribuzione chi-quadro con $N:=\sum_{i=1}^m n_i$ gradi di libertà. Di conseguenza, sostituendo le medie μ_i con i rispettivi stimatori

$$X_{i*}:=\frac{1}{n_i}\sum_{j=1}^{n_i}X_{ij} \tag{10.3.15}$$

otteniamo che

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{(X_{ij}-X_{i*})^2}{\sigma^2}=\frac{SS_{\mathrm{W}}}{\sigma^2}\sim\chi^2_{N-n} \tag{10.3.16}$$

è una chi-quadro con $\sum_{i=1}^m n_i-n$ gradi di libertà. Siccome abbiamo posto

$$SS_{\mathrm{W}}:=\sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij}-X_{i*})^2 \tag{10.3.17}$$

se ne ricava che $SS_{\mathrm{W}}/(\sum_{i=1}^m n_i-n)$ è uno stimatore non distorto di σ^2.

Secondariamente se H_0 è vera, e denotiamo con μ la media comune di tutte le X_{ij}, allora le medie campionarie X_{i*}, per $i=1,2,\dots,m$ sono normali indipendenti con parametri

$$E[X_{i*}]=\mu,\qquad \mathrm{Var}(X_{i*})=\frac{\sigma^2}{n_i}$$

quindi

$$\sum_{i=1}^{m}\frac{(X_{i*}-\mu)^2}{\sigma^2/n_i}\sim\chi^2_m$$

e sostituendo X_{**}, la media campionaria di tutte le X_{ij}, al posto di μ,

$$\sum_{i=1}^{m}\frac{(X_{i*}-X_{**})^2}{\sigma^2/n_i}\sim\chi^2_{m-1} \tag{10.3.18}$$

per [illegible], ponendo

$$SS_{\mathrm{b}}:=\sum_{i=1}^{m}n_i(X_{i*}-X_{**})^2 \tag{10.3.19}$$

ricaviamo che quando H_0 è vera, $SS_{\mathrm{b}}/(m-1)$ è un altro stimatore non distorto di σ^2. È possibile anche dimostrare che quando H_0 è vera, SS_{W} e SS_{b} sono indipendenti,

e che quando H_0 è falsa SS_{b} è tendenzialmente più grande di σ^2. Quindi, posto $N=\sum_{i=1}^m n_i$, un test per $H_0:\mu_1=\mu_2=\cdots=\mu_m$ che abbia significatività α deve

$$\begin{aligned}&\text{rifiutare } H_0 \text{ se } \frac{SS_{\mathrm{b}}/(m-1)}{SS_{\mathrm{W}}/(N-m)}>F_{\alpha,m-1,N-m}\\ &\text{accettare } H_0 \text{ se } \frac{SS_{\mathrm{b}}/(m-1)}{SS_{\mathrm{W}}/(N-m)}\le F_{\alpha,m-1,N-m}\end{aligned} \tag{10.3.20}$$

Osservazione 10.3.1. Quando i campioni hanno ampiezze differenti, si dice che ci troviamo in un caso *non bilanciato*. Si tenga presente che quando ciò sia possibile, è sempre preferibile mettersi in una situazione bilanciata. Uno dei motivi è che un esperimento bilanciato è più robusto di uno che non lo sia, nel senso che è meno sensibile a piccole deviazioni dall'ipotesi (che assumiamo sempre) che la varianza sia costante.

10.4 Analisi della varianza a due vie: introduzione e stima parametrica

Il modello introdotto nella Sezione 10.3 ci ha permesso di studiare l'effetto di un singolo fattore sulla distribuzione dei dati. In questa sezione e nelle successive mostriamo come si possa estendere questo tipo di approccio al caso più generale in cui vi siano diversi fattori influenti, e in particolare ci concentriamo sull'ambito a due fattori, che è l'oggetto di studio dell'analisi della varianza *a due vie*.

Esempio 10.4.1. Un gruppo di 5 studenti viene sottoposto a 4 diversi esami scritti, tutti basati sulla comprensione di un testo, e di difficoltà analoga. I risultati sono:

		Studente				
		1	2	3	4	5
Esame	1	75	73	60	70	86
	2	78	71	64	72	90
	3	80	69	62	70	85
	4	73	67	63	80	92

Ciascun valore in questo campione di 20 dati è influenzato da due fattori, l'esame e lo studente. Il fattore-esame ha 4 possibili valori o *livelli*, mentre il fattore-studente ne ha 5. □

Più in generale, supponiamo che vi siano m diversi livelli del primo fattore e n del secondo fattore, e denotiamo con X_{ij} il valore ottenuto quando il primo fattore ha

livello i e il secondo j. Una buona abitudine consiste nel rappresentare il campione di dati in una tabella rettangolare come la seguente,

$$\begin{array}{ccccccc} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2j} & \cdots & X_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mj} & \cdots & X_{mn} \end{array}$$

Per questo motivo il primo e il secondo fattore vengono detti anche fattori "riga" e "colonna" rispettivamente.

Assumiamo come nella Sezione 10.3 che le X_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$ siano normali indipendenti, tutte con la medesima varianza σ^2. In quel caso la media dei dati dipendeva da un solo fattore (il campione di appartenenza), invece qui supporremo che il valore atteso di X_{ij} dipenda in maniera additiva sia dalla riga sia dalla colonna. Vediamo perché.

Nell'analisi della varianza ad una via, il modello può essere sintetizzato da

$$E[X_{ij}] = \mu_i, \qquad i = 1, 2, \ldots, m$$

Se determiniamo la media (aritmetica) delle μ_i, $\mu := \frac{1}{m}\sum_{i=1}^{m} \mu_i$, e poniamo $\alpha_i := \mu_i - \mu$, il modello si riformula come

$$E[X_{ij}] = \mu + \alpha_i, \qquad i = 1, 2, \ldots, m$$

e si nota subito che $\sum_{i=1}^{m} \alpha_i = 0$ per come sono definiti gli scarti α_i.

Un modello a due fattori *additivo* può similmente essere espresso in termini di deviazioni di riga e di colonna. Se μ_{ij} denota il valore atteso di X_{ij}, allora il modello additivo suppone che esistano delle costanti a_i, $i = 1, 2, \ldots, m$ e b_j, $j = 1, 2 \ldots, n$ tali che

$$\mu_{ij} = a_i + b_j, \qquad i = 1, 2, \ldots, m, \qquad j = 1, 2, \ldots, n$$

Continuando con la nostra notazione per la media aritmetica, poniamo

$$\mu_{i*} := \frac{1}{n}\sum_{j=1}^{n} \mu_{ij}, \qquad \mu_{*j} := \frac{1}{m}\sum_{i=1}^{m} \mu_{ij}, \qquad \mu_{**} := \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n} \mu_{ij}$$

$$a_* := \frac{1}{m}\sum_{i=1}^{m} a_i, \qquad b_* := \frac{1}{n}\sum_{j=1}^{n} b_j$$

Notiamo subito che

$$\begin{aligned} \mu_{i*} &= \frac{1}{n}\sum_{j=1}^{n} (a_i + b_j) \\ &= a_i + b_* \end{aligned}$$

e similmente,

$$\mu_{*j} = a_* + b_j, \qquad \mu_{**} = a_* + b_*$$

Allora se poniamo

$$\begin{aligned} \mu &:= \mu_{**} = a_* + b_* \\ \alpha_i &:= \mu_{i*} - \mu = a_i - a_* \\ \beta_j &:= \mu_{*j} - \mu = b_j - b_* \end{aligned}$$

il modello si riformula come

$$E[X_{ij}] = \mu_{ij} = \mu + \alpha_i + \beta_j, \qquad i = 1, 2, \ldots, m, \qquad j = 1, 2, \ldots, n \quad (10.4.1)$$

e abbiamo come in precedenza che

$$\sum_{i=1}^{m} \alpha_i = \sum_{j=1}^{n} \beta_j = 0 \quad (10.4.2)$$

Il valore μ è a volte detto *media generale*, perciò α_i è la *deviazione dalla media generale dovuta alla riga i*, mentre β_j è la *deviazione dalla media generale dovuta alla colonna j*.

Gli stimatori dei parametri μ, α_i e β_j, al variare di i e j, si ricavano come medie campionarie di insiemi di dati opportuni. Poniamo in particolare

$$X_{i*} := \frac{1}{n}\sum_{j=1}^{n} X_{ij} \qquad \text{la media dei valori nella riga } i \quad (10.4.3)$$

$$X_{*j} := \frac{1}{m}\sum_{i=1}^{m} X_{ij} \qquad \text{la media dei valori nella colonna } j \quad (10.4.4)$$

$$X_{**} := \frac{1}{nm}\sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij} \qquad \text{la media di tutti i valori} \quad (10.4.5)$$

Si vede facilmente che

$$E[X_{i*}] = \frac{1}{n}\sum_{j=1}^{n} E[X_{ij}]$$
$$= \frac{1}{n}\sum_{j=1}^{n}(\mu + \alpha_i + \beta_j)$$
$$= \mu + \alpha_i + \frac{1}{n}\sum_{j=1}^{n}\beta_j = \mu + \alpha_i$$

dove l'ultimo passaggio è dovuto al fatto che la somma dei β_j è nulla per l'Equazione (10.4.2). Analogamente si trova che

$$E[X_{*j}] = \mu + \beta_j, \qquad E[X_{**}] = \mu$$

I valori attesi appena calcolati possono alternativamente essere espressi tramite

$$E[X_{**}] = \mu$$
$$E[X_{i*} - X_{**}] = \alpha_i$$
$$E[X_{*j} - X_{**}] = \beta_j$$

e in questo modo abbiamo individuato degli stimatori non distorti di μ, α_i e β_j, vale a dire

$$\widehat{\mu} := X_{**}$$
$$\widehat{\alpha_i} := X_{i*} - X_{**} \qquad (10.4.6)$$
$$\widehat{\beta_j} := X_{*j} - X_{**}$$

Esempio 10.4.2. La tabella che segue riporta gli stessi dati dell'Esempio 10.4.1, con il valore di alcune statistiche (1 480 è il totale generale e 74.0 il valore di X_{**}, ovvero $\widehat{\mu}$). La si impieghi per stimare i parametri del modello.

		Studente					Totali	
		1	**2**	**3**	**4**	**5**	**per riga**	X_{i*}
Esame	**1**	75	73	60	70	86	364	72.8
	2	78	71	64	72	90	375	75.0
	3	80	69	62	70	85	366	73.2
	4	73	67	63	80	92	375	75.0
Totali per colonna		306	280	249	292	353	1 480	
X_{*j}		76.5	70	62.25	73	88.25		74.0

Come già detto, $\widehat{\mu} = 74$. Gli altri stimatori, come risulta dalla tabella, sono

$$\widehat{\alpha_1} = 72.8 - 74 = -1.2 \qquad \widehat{\beta_1} = 76.5 - 74 = 2.5$$
$$\widehat{\alpha_2} = 75 - 74 = 1 \qquad \widehat{\beta_2} = 70 - 74 = -4$$
$$\widehat{\alpha_3} = 73.2 - 74 = -0.8 \qquad \widehat{\beta_3} = 62.25 - 74 = -11.75$$
$$\widehat{\alpha_4} = 75 - 74 = 1 \qquad \widehat{\beta_4} = 73 - 74 = -1$$
$$\widehat{\beta_5} = 88.25 - 74 = 14.25$$

Questo significa, ad esempio che se si sceglie a caso uno studente e un tipo di esame, allora la stima del punteggio medio è $\widehat{\mu}$ =74. Se invece si fissa l'esame i e lo si sottopone ad uno studente scelto a caso, dovremo aumentare la stima del punteggio medio della quantità $\widehat{\alpha_i}$. Se infine si fissa lo studente j e lo si valuta in un esame a caso, la stima del punteggio medio andrà aumentata della quantità $\widehat{\beta_j}$. Quindi, fissati l'esame di tipo 1 e lo studente 2, stimeremo che il punteggio ottenuto sia il valore assunto da una variabile aleatoria normale di media $\widehat{\mu} + \widehat{\alpha_1} + \widehat{\beta_2} = 74 - 1.2 - 4 = 68.8$ □

10.5 Analisi della varianza a due vie: verifica di ipotesi

Consideriamo nuovamente un modello a due fattori, in cui i dati sono le variabili aleatorie normali e indipendenti X_{ij}, per $i = 1, \ldots, m$ e $j = 1, \ldots, n$, tutte con varianza σ^2, e con

$$E[X_{ij}] = \mu + \alpha_i + \beta_j$$

dove

$$\sum_{i=1}^{m}\alpha_i = \sum_{j=1}^{n}\beta_j = 0$$

In questa sezione siamo interessati a due tipi di test. In primo luogo vogliamo verificare l'ipotesi

$$H_0 : \alpha_i \equiv 0, \qquad i = 1, \ldots, m$$

in alternativa a

$$H_1 : \text{ non tutte le } \alpha_i \text{ sono uguali a zero}$$

In altre parole, vogliamo capire se vi sia o meno effetto di riga, e se quindi la media dei dati dipenda dal fattore riga.

In secondo luogo vogliamo verificare l'ipotesi

$$H_0 : \beta_j \equiv 0, \qquad j = 1, \ldots, n$$

in alternativa a

$$H_1 : \text{ non tutte le } \beta_j \text{ sono uguali a zero}$$

Per capire se vi sia effetto di colonna.

Come nella Sezione 10.3 per realizzare il test applichiamo l'analisi della varianza, esibendo due stimatori di σ^2, il primo dei quali è valido in ogni caso, mentre il secondo si comporta bene se l'ipotesi nulla è valida, e tende a sovrastimare σ^2 negli altri casi.

Cominciamo con il notare che

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(X_{ij}-E[X_{ij}])^2}{\sigma^2}=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(X_{ij}-\mu-\alpha_i-\beta_j)^2}{\sigma^2}\sim\chi^2_{nm}$$

Se in questa espressione sostituiamo i parametri $\mu, \alpha_1, \alpha_2, \ldots, \alpha_m$ e $\beta_1, \beta_2, \ldots, \beta_n$ con i loro stimatori $\widehat{\mu}, \widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_m$ e $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_n$, la statistica ottenuta perde un grado di libertà per ogni parametro sostituito con il suo stimatore. Occorre però fare grande attenzione al conteggio dei parametri stimati, infatti visto che $\sum_{i=1}^{m}\alpha_i = 0$, una volta che siano stati stimate $m-1$ delle α_i, quella restante può essere ottenuta per differenza (questo si esprime dicendo che gli m stimatori sono *linearmente dipendenti*). Per questo motivo il numero di stimatori effettivamente utilizzati nel sostituire gli $m+n+1$ parametri è di $m-1$ per le $\widehat{\alpha}_i$, di $n-1$ per le $\widehat{\beta}_j$, più lo stimatore $\widehat{\mu}$; sono in tutto $m+n-1$. Quindi, visto che $nm-m-n+1=(m-1)(n-1)$,

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(X_{ij}-\widehat{\mu}-\widehat{\alpha}_i-\widehat{\beta}_j)^2}{\sigma^2}\sim\chi^2_{(m-1)(n-1)}$$

Siccome $\widehat{\mu}:=X_{**}$, $\widehat{\alpha}_i:=X_{i*}-X_{**}$, $\widehat{\beta}_j:=X_{*j}-X_{**}$, si ha che $\widehat{\mu}+\widehat{\alpha}_i+\widehat{\beta}_j = X_{i*}+X_{*j}-X_{**}$, e quindi l'espressione precedente si riformula come

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(X_{ij}-X_{i*}-X_{*j}+X_{**})^2}{\sigma^2}\sim\chi^2_{(m-1)(n-1)} \qquad (10.5.1)$$

Definizione 10.5.1. La statistica SS_e, definita da

$$SS_e:=\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}-X_{i*}-X_{*j}+X_{**})^2 \qquad (10.5.2)$$

è chiamata *somma dei quadrati degli errori.*

Effettivamente, se pensiamo alla differenza tra il valore osservato per X_{ij}, e quello stimato, $\widehat{\mu}+\widehat{\alpha}_i+\widehat{\beta}_j = X_{i*}+X_{*j}-X_{**}$, come ad un "errore", SS_e risulta la somma dei quadrati di tali errori. Dall'Equazione (10.5.1) deduciamo che SS_e/σ^2 ha distribuzione chi-quadro con $(m-1)(n-1)$ gradi di libertà, e quindi

$$E\left[\frac{SS_e}{\sigma^2}\right]=(m-1)(n-1) \qquad \text{ovvero} \qquad E\left[\frac{SS_e}{(m-1)(n-1)}\right]=\sigma^2$$

Ne segue che $SS_e/(m-1)(n-1)$ è uno stimatore corretto di σ^2.

Per verificare l'ipotesi nulla che non vi sia effetto di riga:

$$H_0 : \alpha_i \equiv 0, \qquad i=1,\ldots,m$$

abbiamo bisogno di un secondo stimatore di σ^2 che sia valido solo quando H_0 è vera. Consideriamo allora le medie per riga, X_{i*}, $i=1,\ldots,m$ e notiamo che quando l'ipotesi nulla è vera,

$$E[X_{i*}]=\mu+\alpha_i=\mu$$

Essendo inoltre X_{i*} la media campionaria di n variabili aleatorie i.i.d. di varianza σ^2, si ha che

$$\mathrm{Var}(X_{i*})=\frac{\sigma^2}{n}$$

e quindi

$$\sum_{i=1}^{m}\frac{(X_{i*}-E[X_{i*}])^2}{\mathrm{Var}(X_{i*})}=\sum_{i=1}^{m}\frac{(X_{i*}-\mu)^2}{\sigma^2/n}\sim\chi^2_m$$

Se si sostituisce μ con il suo stimatore X_{**}, si ottiene una statistica che quando H_0 è vera ha perciò distribuzione chi-quadro con $m-1$ gradi di libertà:

$$\sum_{i=1}^{m}\frac{(X_{i*}-X_{**})^2}{\sigma^2/n}\sim\chi^2_{m-1} \qquad (10.5.3)$$

Definizione 10.5.2. La statistica SS_r, definita da

$$SS_r:=n\sum_{i=1}^{m}(X_{i*}-X_{**})^2 \qquad (10.5.4)$$

è chiamata *somma dei quadrati delle righe.*

Per quanto affermato dall'Equazione (10.5.3), se H_0 è vera, SS_r/σ^2 è una chi-quadro con $m-1$ gradi di libertà e quindi

$$E\left[\frac{SS_r}{\sigma^2}\right]=m-1 \qquad \text{ovvero} \qquad E\left[\frac{SS_r}{m-1}\right]=\sigma^2$$

Se ne deduce che, nelle stesse ipotesi, $SS_r/(m-1)$ è uno stimatore di σ^2. Si può anche dimostrare che questo secondo stimatore tende a sovrastimare σ^2 quando H_0 non è soddisfatta, ed è indipendente da SS_r in caso contrario. Avendo ottenuto due stimatori di σ^2 con le caratteristiche desiderate, possiamo costruire il test dell'ipotesi che le α_i siano tutte nulle, usando come statistica rilevante il loro rapporto, che grazie

Tabella 10.3 ANOVA a due fattori.

	Somma di quadrati	Gradi di libertà
Riga	$SS_r := n\sum_i (X_{i*} - X_{**})^2$	$m-1$
Colonna	$SS_c := m\sum_j (X_{*j} - X_{**})^2$	$n-1$
Errore	$SS_e := \sum_i \sum_j (X_{ij} - X_{i*} - X_{*j} + X_{**})^2$	$(n-1)(m-1)$

Sia $N = (n-1)(m-1)$.

Ipotesi nulla	Statistica del test	Un test con significatività α deve	p-dei-dati se $D_{ts} = v$
Tutte le $\alpha_i = 0$	$D_{ts} := \frac{SS_r}{SS_e}(n-1)$	rifiutare H_0 se $D_{ts} > F_{\alpha,m-1,N}$	$P(F_{m-1,N} \geq v)$
Tutte le $\beta_j = 0$	$D_{ts} := \frac{SS_c}{SS_e}(m-1)$	rifiutare H_0 se $D_{ts} > F_{\alpha,n-1,N}$	$P(F_{n-1,N} \geq v)$

all'indipendenza ha distribuzione F quando H_0 è vera, e altrimenti tende ad assumere valori maggiori.

$$D_{ts} := \frac{SS_r/(m-1)}{SS_e/(n-1)(m-1)} = \frac{SS_r}{SS_e}(n-1) \tag{10.5.5}$$

Allora, ponendo $N := (n-1)(m-1)$, un test dell'ipotesi H_0, che abbia livello di significatività α, deve

$$\begin{aligned} &\text{rifiutare } H_0 \text{ se } \frac{SS_r}{SS_e}(n-1) > F_{\alpha,m-1,N} \\ &\text{accettare } H_0 \text{ se } \frac{SS_r}{SS_e}(n-1) \leq F_{\alpha,m-1,N} \end{aligned} \tag{10.5.6}$$

In alternativa, la verifica può essere effettuata calcolando il p-dei-dati. Se v è il valore assunto dalla statistica D_{ts}, il suo valore è dato da

$$p\text{-dei-dati} = P(F_{m-1,N} \geq v) \tag{10.5.7}$$

Si può ottenere un test del tutto analogo per verificare l'ipotesi che tutte le β_j siano nulle, ovvero che non vi sia effetto di colonna. I risultati sono sintetizzati nella Tabella 10.3. Il Programma 10.5 rende automatici i calcoli necessari e fornisce il p-dei-dati.

Esempio 10.5.1. I dati seguenti[4] rappresentano il numero di specie di invertebrati di dimensioni macroscopiche, individuati nei pressi di 6 diversi luoghi con scarichi termici, dal 1970 al 1977.

[4] Wartz and Skinner, "A 12 year macroinvertebrate study in the vicinity of two thermal discharges to the Susquehanna River near York, Haven, PA.", *Jour. of Testing and Evaluation*, vol. 12, pp. 157–163, Maggio 1984.

		Stazione 1	2	3	4	5	6
Anno	1970	53	35	31	37	40	43
	1971	36	34	17	21	30	18
	1972	47	37	17	31	45	26
	1973	55	31	17	23	43	37
	1974	40	32	19	26	45	37
	1975	52	42	20	27	26	32
	1976	39	28	21	21	36	28
	1977	40	32	21	21	36	35

Eseguiamo il Programma 10.5 per verificare se i dati siano influenzati: (1) dall'anno e (2) dalla stazione di rilevamento. I risultati sono presentati dalla schermata in Figura 10.2, che fornisce due p-dei-dati così piccoli che sia l'ipotesi che la distribuzione non dipenda dall'anno, sia quella che non dipenda dalla stazione vengono rifiutate a qualunque livello di significatività ragionevole. □

10.6 Analisi della varianza a due vie con interazioni

Nelle Sezioni 10.4 e 10.5 abbiamo sempre supposto che l'influenza del fattore riga e del fattore colonna fosse di tipo additivo, ovvero che X_{ij} fosse normale di varianza σ^2 fissata e media $\mu + \alpha_i + \beta_j$ costituita da una parte di media generale μ, e da due

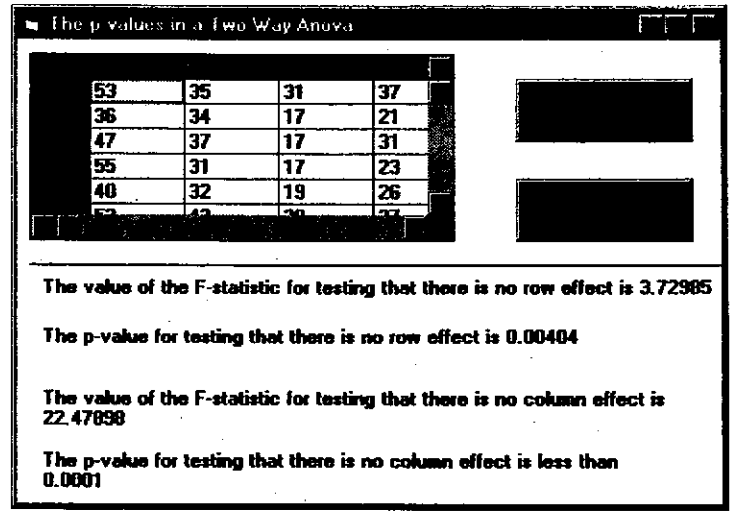


Figura 10.2

contributi, dovuti ai due fattori (la riga i e la colonna j) che venivano semplicemente sommati. Il punto debole di questo modello è che supponendo i contributi additivi, non contempla casi in cui vi siano *interazioni* tra i due fattori.

Si consideri ad esempio un esperimento volto ad analizzare il numero medio di pezzi difettosi prodotti da quattro operai utilizzando tre macchinari differenti. Se la macchina è il fattore di riga, il contributo α_i rappresenta quanti pezzi difettosi in più o in meno vengono prodotti in media dalla macchina i. Non è assurdo assumere che questo contributo sia lo stesso per ogni operaio, ma è anche possibile che un particolare operaio sia più efficiente nell'adoperare tale macchinario (magari perché lo conosce meglio) e che quindi il contributo corrispondente non sia lo stesso per ogni operaio j, ma si differenzi, e sensatamente lo faccia in modo diverso da macchina a macchina. Vi potrebbe insomma essere una interazione uomo-macchina che il modello additivo non contempla.

Per permettere questo tipo di interazioni tra fattore riga e fattore colonna, poniamo come in precedenza per $i = 1, \dots, m$ e $j = 1, \dots, n$,

$$\begin{aligned} \mu_{ij} &:= E[X_{ij}] & \mu &:= \mu_{**} \\ \alpha_i &:= \mu_{i*} - \mu_{**} & \beta_j &:= \mu_{*j} - \mu_{**} \end{aligned} \tag{10.6.1}$$

e introduciamo il parametro di interazione definito per differenza:

$$\gamma_{ij} := \mu_{ij} - \mu_{i*} - \mu_{*j} + \mu_{**}$$

in modo che valga l'identità fondamentale

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \tag{10.6.2}$$

Non è difficile verificare che queste definizioni sono fatte in modo che

$$\sum_{i=1}^{m} \alpha_i = \sum_{j=1}^{n} \beta_j = \sum_{i=1}^{m} \gamma_{ij} = \sum_{j=1}^{n} \gamma_{ij} = 0 \tag{10.6.3}$$

I parametri che compaiono nell'Equazione (10.6.2) meritano qualche commento. La media aritmetica dei valori attesi di tutti gli nm dati è la *media generale* μ; il parametro α_i rappresenta la differenza tra la media aritmetica dei valori attesi dei dati sulla riga i e la media generale, ed è quindi detta *effetto della riga* i; analogamente β_j è l'*effetto della colonna* j. Il parametro γ_{ij}, pari a $\mu_{ij} - (\mu + \alpha_i + \beta_j)$, è quindi lo scostamento tra la media vera μ_{ij} e il valore che si ottiene tenendo conto della media generale e degli effetti di riga e di colonna. Questo residuo rappresenta quanto la media μ_{ij} si discosta dal valore che si otterrebbe con un modello additivo ed è detto *interazione tra la riga i e la colonna j*.

Come risulterà chiaro in seguito, se si vuole verificare l'ipotesi che non vi sia interazione, ovvero che $\gamma_{ij} \equiv 0$ per tutte le coppie (i, j), non è sufficiente una sola

osservazione per ogni coppia di fattori (i, j). Supponiamo quindi di disporre di osservazioni indipendenti per ogni combinazione dei due fattori, e denotiamo con X_{ijk} la k-esima osservazione alla riga i e colonna j. Poiché tutti i dati si suppongono avere distribuzione normale con varianza costante σ^2, il modello è rappresentato dall variabili aleatorie indipendenti

$$\begin{gathered} X_{ijk} \sim \mathcal{N}\left(\mu + \alpha_i + \beta_j + \gamma_{ij}, \sigma^2\right) \\ i = 1, 2, \dots, m, \qquad j = 1, 2, \dots, n, \qquad k = 1, 2, \dots, l \end{gathered}$$

dove i coefficienti devono soddisfare l'Equazione (10.6.3). I problemi che affronteremo sono la stima dei parametri precedenti e la verifica di tre tipi di ipotesi statistiche:

$$\begin{aligned} H_0^{\mathrm{r}} &: \alpha_i \equiv 0, && \text{per ogni } i \\ H_0^{\mathrm{c}} &: \beta_j \equiv 0, && \text{per ogni } j \\ H_0^{\mathrm{int}} &: \gamma_{ij} \equiv 0, && \text{per ogni } i, j \end{aligned}$$

ovvero rispettivamente l'assenza di effetto di riga, l'assenza di effetto di colonna e l'assenza di interazioni tra le righe e le colonne.

La stima dei parametri non presenta alcuna difficoltà ed è condotta come nelle sezioni precedenti, sfruttando l'Equazione (10.6.3) e la seguente,

$$E[X_{ijk}] = \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

Infatti calcolando le medie aritmetiche degli $E[X_{ijk}]$ al variare di k ed eventualmente di altri indici si ha che:

$$\begin{aligned} E[X_{ij*}] &= \mu + \alpha_i + \beta_j + \gamma_{ij} \\ E[X_{i**}] &= \mu + \alpha_i \\ E[X_{*j*}] &= \mu + \beta_j \\ E[X_{***}] &= \mu \end{aligned}$$

Perciò denotando con $\hat{\theta}$ lo stimatore di un generico parametro θ, costruiamo per μ, α_i, β_j e γ_{ij} gli stimatori seguenti,

$$\begin{aligned} \widehat{\mu} &:= X_{***} \\ \widehat{\alpha_i} &:= X_{i**} - X_{***} \\ \widehat{\beta_j} &:= X_{*j*} - X_{***} \\ \widehat{\gamma_{ij}} &:= X_{ij*} - \widehat{\mu} - \widehat{\alpha_i} - \widehat{\beta_j} = X_{ij*} - X_{i**} - X_{*j*} + X_{***} \end{aligned} \tag{10.6.4}$$

che soddisfano per costruzione le uguaglianze

$$\sum_{i=1}^{m} \widehat{\alpha_i} = \sum_{j=1}^{n} \widehat{\beta_j} = \sum_{i=1}^{m} \widehat{\gamma_{ij}} = \sum_{j=1}^{n} \widehat{\gamma_{ij}} = 0 \tag{10.6.5}$$

Per sviluppare i test per le ipotesi H_0^{int}, H_0^{r} e H_0^{c}, notiamo intanto che

$$\sum_{k=1}^{l}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(X_{ijk}-\mu-\alpha_i-\beta_j-\gamma_{ij})^2}{\sigma^2}\sim\chi^2_{nml}$$

Se sostituiamo i parametri con i loro stimatori otteniamo la statistica

$$\sum_{k=1}^{l}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(X_{ijk}-\widehat{\mu}-\widehat{\alpha}_i-\widehat{\beta}_j-\widehat{\gamma_{ij}})^2}{\sigma^2} \tag{10.6.6}$$

che è una chi-quadro con un numero di gradi di libertà pari a nml meno il numero di stimatori linearmente indipendenti usati. Sappiamo già che per avere tutte le $\widehat{\alpha}_i$ basta calcolarne $m-1$ dai dati, perché l'ultima si può ricavare per differenza dall'Equazione (10.6.5). Similmente basta stimare $n-1$ delle β_j. Per quanto riguarda gli stimatori $\widehat{\gamma_{ij}}$, si noti che, sempre per l'equazione citata, se li si dispone in una tabella $m\times n$, allora la somma dei valori su tutte le righe e su tutte le colonne è nulla, e quindi basta conoscere $(n-1)(m-1)$ di questi stimatori (tolta una riga e una colonna) per ricavare i restanti $m+n-1$ per differenza. Tenendo conto anche di $\widehat{\mu}$, allora in tutto gli stimatori che devono essere calcolati a partire dai dati sono

$$n-1+m-1+(n-1)(m-1)+1=nm$$

per cui i gradi di libertà residui sono $nml-nm=nm(l-1)$. Se definiamo

$$SS_{\text{e}}:=\sum_{k=1}^{l}\sum_{j=1}^{n}\sum_{i=1}^{m}(X_{ijk}-X_{ij*})^2 \tag{10.6.7}$$

e notiamo che

$$\widehat{\mu}+\widehat{\alpha}_i+\widehat{\beta}_j+\widehat{\gamma_{ij}}=X_{ij*}$$

otteniamo che SS_{e}/σ^2 coincide con la statistica (10.6.6) e quindi

$$\frac{SS_{\text{e}}}{\sigma^2}\sim\chi^2_{nm(l-1)} \tag{10.6.8}$$

per cui $SS_{\text{e}}/nm(l-1)$ è uno stimatore non distorto di σ^2.

Supponiamo di dover vagliare l'ipotesi $H_0^{\text{int}}:\gamma_{ij}\equiv 0$, ovvero l'assenza di interazioni tra righe e colonne. Quando questa ipotesi è soddisfatta, le X_{ij*} sono normali con media e varianza date da

$$E[X_{ij*}]=\mu+\alpha_i+\beta_j \qquad \text{e} \qquad \text{Var}(X_{ij*})=\frac{\sigma^2}{l}$$

infatti X_{ij*} è la media campionaria di $X_{ij1},X_{ij2},\ldots,X_{ijl}$, ciascuna delle quali ha distribuzione $\mathcal{N}(\mu+\alpha_i+\beta_j,\sigma^2)$. Quindi nell'ipotesi che non vi siano interazioni,

$$\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{(X_{ij*}-\mu-\alpha_i-\beta_j)^2}{\sigma^2/l}\sim\chi^2_{nm}$$

e, visto che è necessario stimare dai dati esattamente $1+m-1+n-1=n+m-1$ parametri, ne segue che se definiamo

$$\begin{aligned}SS_{\text{int}}&:=l\sum_{j=1}^{n}\sum_{i=1}^{m}(X_{ij*}-\widehat{\mu}-\widehat{\alpha}_i-\widehat{\beta}_j)^2\\&=l\sum_{j=1}^{n}\sum_{i=1}^{m}(X_{ij*}-X_{i**}-X_{*j*}+X_{***})^2\end{aligned} \tag{10.6.9}$$

otteniamo che, quando H_0^{int} è vera,

$$\frac{SS_{\text{int}}}{\sigma^2}\sim\chi^2_{(n-1)(m-1)} \tag{10.6.10}$$

per cui, quando H_0^{int} è soddisfatta, $SS_{\text{int}}/(n-1)(m-1)$ è uno stimatore non distorto di σ^2, e inoltre SS_{int} e SS_{e} sono indipendenti, perciò

$$F_{\text{int}}:=\frac{SS_{\text{int}}/(n-1)(m-1)}{SS_{\text{e}}/nm(l-1)}\sim F_{(n-1)(m-1),nm(l-1)} \tag{10.6.11}$$

ha distribuzione F con $(n-1)(m-1)$ gradi libertà al numeratore e $nm(l-1)$ al denominatore. Un test per la verifica di H_0^{int} con significatività α deve quindi

$$\begin{aligned}&\text{rifiutare } H_0^{\text{int}} \text{ se } \frac{SS_{\text{int}}/(n-1)(m-1)}{SS_{\text{e}}/nm(l-1)}>F_{\alpha,(n-1)(m-1),nm(l-1)}\\&\text{accettare } H_0^{\text{int}} \text{ se } \frac{SS_{\text{int}}/(n-1)(m-1)}{SS_{\text{e}}/nm(l-1)}\le F_{\alpha,(n-1)(m-1),nm(l-1)}\end{aligned} \tag{10.6.12}$$

In alternetiva si può calcolare il p-dei-dati. Sia v il valore assunto dalla statistica F_{int}, allora il p-dei-dati del test dell'ipotesi che tutte le interazioni siano nulle è:

$$p\text{-dei-dati}=P(F_{(n-1)(m-1),nm(l-1)}>v) \tag{10.6.13}$$

Consideriamo ora l'ipotesi nulla H_0^{r} che non vi sia effetto di riga, ovvero che i parametri α_i siano tutti nulli. Supponiamo che questa ipotesi sia soddisfatta. Allora

$X_{ijk} \sim \mathcal{N}\left(\mu + \beta_j + \gamma_{ij}, \sigma^2\right)$, inoltre visto che $\sum_{j=1}^{n} \beta_j = \sum_{j=1}^{n} \gamma_{ij} = 0$,

$$\begin{aligned} E[X_{i**}] &= E\left[\frac{1}{nl}\sum_{k=1}^{l}\sum_{j=1}^{n} X_{ijk}\right] \\ &= \frac{1}{nl}\sum_{k=1}^{l}\sum_{j=1}^{n}(\mu + \beta_j + \gamma_{ij}) \\ &= \frac{1}{l}\sum_{k=1}^{l}\mu = \mu \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(X_{i**}) &= \mathrm{Var}\left(\frac{1}{nl}\sum_{k=1}^{l}\sum_{j=1}^{n} X_{ijk}\right) \\ &= \frac{1}{n^2 l^2}\sum_{k=1}^{l}\sum_{j=1}^{n}\mathrm{Var}(X_{ijk}) \\ &= \frac{nl\sigma^2}{n^2 l^2} = \frac{\sigma^2}{nl} \end{aligned}$$

per cui la somma dei quadrati standardizzati al variare di i, è una chi-quadro con m gradi di libertà,

$$\sum_{i=1}^{m} \frac{(X_{i**} - \mu)^2}{\sigma^2/nl} \sim \chi^2_m \tag{10.6.14}$$

Poniamo allora

$$SS_{\mathrm{r}} := nl\sum_{i=1}^{m}(X_{i**} - \widehat{\mu})^2 \tag{10.6.15}$$

In tal modo SS_{r}/σ^2, che coincide con la (10.6.14) se si sostituisce l'unico parametro μ con il suo stimatore, perde un grado di libertà e ha distribuzione χ^2_{m-1}; quindi $SS_{\mathrm{r}}/(m-1)$ è uno stimatore corretto di σ^2. Stiamo supponendo vera H_0^{r} e in queste ipotesi si può dimostrare che SS_{r} e SS_{e} sono indipendenti, per cui

$$F_{\mathrm{r}} := \frac{SS_{\mathrm{r}}/(m-1)}{SS_{\mathrm{e}}/nm(l-1)} \sim F_{m-1,nm(l-1)} \tag{10.6.16}$$

Utilizzando questa statistica, un test di H_0^{r} con significatività α deve

$$\begin{aligned} &\text{rifiutare } H_0^{\mathrm{r}} \text{ se } \frac{SS_{\mathrm{r}}/(m-1)}{SS_{\mathrm{e}}/nm(l-1)} > F_{\alpha,m-1,nm(l-1)} \\ &\text{accettare } H_0^{\mathrm{r}} \text{ se } \frac{SS_{\mathrm{r}}/(m-1)}{SS_{\mathrm{e}}/nm(l-1)} \leq F_{\alpha,m-1,nm(l-1)} \end{aligned} \tag{10.6.17}$$

Tabella 10.4 ANOVA a due fattori, con interazioni e l osservazioni per cella. Si è posto $N := nm(l-1)$ e $M := (n-1)(m-1)$.

Fonte di variabilità	Somma di quadrati	Gradi di libertà
Riga	$SS_{\mathrm{r}} := nl\sum_{i=1}^{m}(X_{i**} - X_{***})^2$	$m-1$
Colonna	$SS_{\mathrm{c}} := ml\sum_{j=1}^{n}(X_{*j*} - X_{***})^2$	$n-1$
Interazioni	$SS_{\mathrm{int}} := l\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij*} - X_{i**} - X_{*j*} + X_{***})^2$	M
Errore	$SS_{\mathrm{e}} := \sum_{k=1}^{l}\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ijk} - X_{***})^2$	N

Ipotesi nulla	Statistica del test	Un test con significatività α deve	p-dei-dati se $F = v$
H_0^{r}: Le α_i sono tutte nulle	$F_{\mathrm{r}} := \frac{SS_{\mathrm{r}}/(m-1)}{SS_{\mathrm{e}}/N}$	rifiutare H_0^{r} se $F_{\mathrm{r}} > F_{\alpha,m-1,N}$	$P(F_{m-1,N} \geq v)$
H_0^{c}: Le β_j sono tutte nulle	$F_{\mathrm{c}} := \frac{SS_{\mathrm{c}}/(n-1)}{SS_{\mathrm{e}}/N}$	rifiutare H_0^{c} se $F_{\mathrm{c}} > F_{\alpha,n-1,N}$	$P(F_{n-1,N} \geq v)$
H_0^{int}: Le γ_{ij} sono tutte nulle	$F_{\mathrm{int}} := \frac{SS_{\mathrm{int}}/M}{SS_{\mathrm{e}}/N}$	rifiutare H_0^{int} se $F_{\mathrm{int}} > F_{\alpha,M,N}$	$P(F_{M,N} \geq v)$

Il p-dei-dati è sempre un'alternativa percorribile, e in particolare quello relativo all'ipotesi che non vi sia effetto di riga è dato da

$$p\text{-dei-dati} = P(F_{m-1,nm(l-1)} > v) \tag{10.6.18}$$

dove v è il valore assunto dalla statistica F_{r}.

Lo studio dell'ipotesi H_0^{c} è del tutto analogo a quello di H_0^{r}. I passaggi deduttivi vengono lasciati al lettore, mentre i risultati vengono presentati assieme agli altri di questa sezione nella Tabella 10.4.

È bene ricordare che tutti i test citati portano ad un rifiuto quando la statistica corrispondente è grande. Il motivo sta nel fatto che quando l'ipotesi nulla non è valida la distribuzione delle statistiche che stanno al numeratore delle varie F_{int}, F_{r} e F_{c} si sposta verso valori più grandi, mentre la distribuzione di SS_{e} al denominatore non cambia.

Il Programma 10.6 del software abbinato al libro permette di calcolare le tre statistiche in questione, nonché i corrispondenti valori del p-dei-dati.

Esempio 10.6.1. Si pensa che il tempo di vita di un tipo di generatori possa essere influenzato sia dal materiale con cui sono costruiti, sia dalla temperatura dell'ambiente di lavoro. I dati che seguono rappresentano i tempi di vita di 24 generatori, fabbricati con tre diversi materiali e messi in funzione in due ambienti a temperature diverse.

		Temperatura di funzionamento	
		10 gradi	18 gradi
Materiale	1	135, 150, 176, 85	50, 55, 64, 38
	2	150, 162, 171, 120	76, 88, 91, 57
	3	138, 111, 140, 106	68, 60, 74, 51

Vi è qualche indicazione che il materiale e/o la temperatura siano davvero fattori influenti? Sembra che vi siano delle interazioni in atto?

La risoluzione può essere ricavata dal Programma 10.6, come illustrato nelle Figure 10.3 e 10.4. □

Problemi

1. Uno dei processi di purificazione impiegati per una certa sostanza chimica prevede di metterla in soluzione e filtrarla con una resina che ne fissi le impurità. Un ingegnere chimico vuole provare l'efficienza di 3 tipi di resine. Divide allora una piccola quantità della soluzione in 15 campioni, che filtra con le tre resine, 5 per tipo. Le concentrazioni di impurità dopo il filtraggio sono risultate le seguenti:

Resina I	0.046	0.025	0.014	0.017	0.043
Resina II	0.038	0.035	0.031	0.022	0.012
Resina III	0.031	0.042	0.020	0.018	0.039

Verifica l'ipotesi che non vi siano differenze tra le efficienze delle tre resine.

2. Siamo interessati a determinare quale sia il filtro più adatto ad essere applicato sullo schermo di un radar a tubo catodico per far sì che l'operatore individui facilmente gli obiettivi. Realizziamo l'esperimento seguente: rappresentiamo sullo schermo un segnale di solo rumore di fondo, sovrapponendogli poi un singolo obiettivo, la cui intensità viene fatta aumentare da zero fino a quando l'operatore lo individua. Si ripete questo esperimento per 20 volte, con ciascuno dei 3 filtri a disposizione, segnando il livello di intensità al quale l'operatore individua l'obiettivo. I dati trovati sono quelli che seguono.

Filtro 1		Filtro 2		Filtro 3	
90	90	88	95	95	92
87	82	90	86	95	85
93	93	97	89	89	97
96	90	87	92	98	90
94	96	90	98	96	87
88	87	96	95	81	90
90	99	90	102	92	101
84	101	90	105	79	100
101	79	100	85	105	84
96	98	93	97	98	102

Verifica al 5% di significatività l'ipotesi che che i filtri siano equivalenti.

3. Spiega come mai l'ipotesi $H_0 : \mu_1 = \mu_2 = \cdots = \mu_m$ non può essere verificata semplicemente eseguendo un test t per ciascuna del $\binom{m}{2}$ coppie di campioni.

4. Una fabbrica contiene 3 forni che vengono usati per riscaldare esemplari di vari metalli. È richiesto che le temperature mantenute da tutti e tre siano uguali a meno di fluttuazioni casuali. Per verificare questa ipotesi si annotano 15 misurazioni delle temperature, ottenendo la tabella seguente.

Forno 1	492.4	493.6	498.5	488.6	494
Forno 2	488.5	485.3	482	479.4	478
Forno 3	502.1	492	497.5	495.3	486.7

Diresti che i forni funzionano alla stessa temperatura? Verifica questa ipotesi al 5% di significatività.

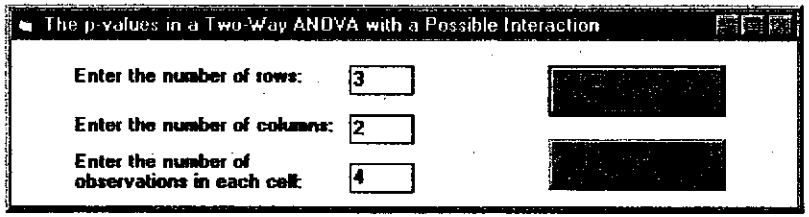


Figura 10.3

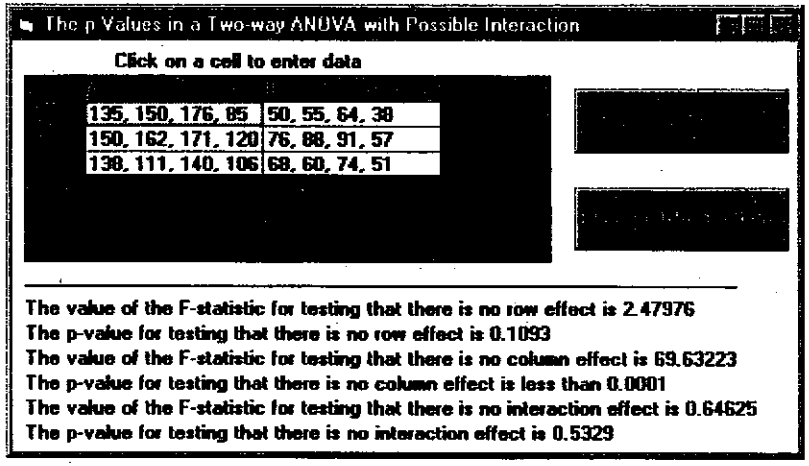


Figura 10.4

5. Si provano 4 diverse tecniche che permettono di misurare il contenuto di magnesio di un composto chimico. Ripetendo l'analisi 4 volte con ciascuna tecnica su uno stesso campione si trovano questi dati:

Metodo 1	76.42	78.62	80.40	78.20
Metodo 2	80.41	82.26	81.15	79.20
Metodo 3	74.20	72.68	78.84	80.32
Metodo 4	86.20	86.04	84.36	80.68

Ti sembra che i dati indichino che i diversi metodi di misurazione siano equivalenti?

6. Per confrontare l'efficacia di due diete, si scelgono 20 individui sovrappeso di almeno 40 libbre, e li si divide a caso in due gruppi da 10, ciascuno dei quali viene sottoposto a una delle due diete. Dopo 10 settimane le diminuzioni di peso riscontrate sono state (in libbre):

Dieta 1	22.2	23.4	24.2	16.1	9.4	12.5	18.6	32.2	8.8	7.6
Dieta 2	24.2	16.8	14.6	13.7	19.5	17.6	11.2	9.5	30.1	21.5

Verifica al 5% di significatività l'ipotesi che le due diete abbiano uguale effetto.

7. Nello sperimentare l'efficacia di un certo polimero nel rimuovere delle scorie tossiche dall'acqua, sono state condotte prove a 3 diverse temperature. I dati che seguono riportano le percentuali di scorie rimosse in 21 esperimenti indipendenti.

Bassa temperatura	42	41	37	29	35	40	32
Media temperatura	36	35	32	38	39	42	34
Alta temperatura	33	44	40	36	44	37	45

Verifica l'ipotesi che il polimero sia altrettanto efficacie a tutte e tre le temperature. Usa **(a)** il 5% di significatività; **(b)** l'1% di significatività.

8. Considera l'analisi della varianza ad un fattore con n elementi per campione. Sia S_i^2 la varianza campionaria del campione i, per $i = 1, 2, \ldots, m$. Dimostra che

$$SS_W = (n-1)\sum_{i=1}^{m} S_i^2$$

9. I dati che seguono si riferiscono ai mesi di vita di 30 ratti di una razza poco longeva che sono stati divisi a caso in 3 campioni di 10 esemplari e nutriti con 3 tipi di diete diverse.

	Basso livello calorico	**Medio livello calorico**	**Alto livello calorico**
Media campionaria	22.4	16.8	13.7
Varianza campionaria	24.0	23.2	17.1

Verifica l'ipotesi che la vita media dei ratti non sia influenzata dalla dieta: **(a)** al 5% di significatività; **(b)** all'1% di significatività.

10. Il livello di chininogeno nel plasma è legato alla capacità del corpo umano di resister alle infiammazioni. In uno studio[5] si sono riscontrate le seguenti concentrazioni del la sostanza (in microgrammi per millilitro), in soggetti normali e in soggetti affetti da sindrome di Hodgkin, conclamata o meno.

Soggetto sano	**Sindrome conclamata**	**Sindrome non conclamata**
5.37	3.96	5.37
5.80	3.04	10.60
4.70	5.28	5.02
5.70	3.40	14.30
3.40	4.10	9.90
8.60	3.61	4.27
7.48	6.16	5.75
5.77	3.22	5.03
7.15	7.48	5.74
6.49	3.87	7.85
4.09	4.27	6.82
5.94	4.05	7.90
6.38	2.40	8.36

Verifica al 5% di significatività l'ipotesi che il livello medio di chininogeno dei tre grupp. sia lo stesso.

11. In uno studio[6] del 1984 sui muscoli flessori del tronco, sono state visitate 75 bambine d. 3 ai 7 anni, divise in base all'età in 5 gruppi da 15. La forza muscolare è stata misurat. in una scala da 0 a 5, e la tabella seguente riassume i risultati ottenuti.

Età	3	4	5	6	7
Media campionaria	3.3	3.7	4.1	4.4	4.8
Varianza campionaria	0.9	1.1	1.1	0.9	0.5

Verifica ad un livello di significatività del 5% se la forza media dei flessori del tronco si la stessa a tutte le età.

12. Un medico che lavora in un pronto soccorso vuole confrontare 3 tipi di steroidi usa per curare delle leggere crisi asmatiche, per vedere quale sia più rapido nel liberare polmoni. Per un certo periodo egli somministra una delle tre sostanze a caso ai pazienti che ne hanno bisogno, e alla fine nota che ha testato ciascuno steroide su 12 pazienti diversi, ottenendo dei campioni di dati (in minuti) le cui statistiche sono riassunte qu sotto.

[5] N. Eilam, P. K. Johnson, N. L. Johnson, W. Creger, "Bradykininogen levels in Hodgkin's disease" *Cancer*, vol. 22, pp. 631–634, 1968.

[6] K. Baldauf, D. Swenson, J. Medeiros, S. Radtka, "Clinical assessment of trunk flexor muscle strength in healthy girls 3 to 7", *Physical Therapy*, vol. 64, pp. 1203–1208, 1984.

Steroide	A	B	C
Media campionaria	32	40	30
Varianza campionaria	145	138	150

(a) Verifica l'ipotesi che il tempo medio per uscire da una crisi asmatica sia lo stesso per tutti e tre gli steroidi. Usa il 5% di significatività.

(b) Trova degli intervalli di valori per le differenze $\mu_i - \mu_j$ che siano validi con il 95% di confidenza.

13. Si analizza l'apporto di grassi di 3 marche di carni lavorate. Si usano 5 confezioni di ciascun tipo, trovando i dati seguenti (in percentuale sul peso):

Marca 1	32	34	31	35	33
Marca 2	41	32	33	29	35
Marca 3	36	37	30	28	33

(a) Il contenuto medio di grassi di una confezione cambia da marca a marca?

(b) Trova degli intervalli di valori per tutte le differenze $\mu_i - \mu_j$ che siano contemporaneamente validi con il 95% di confidenza.

14. Un nutrizionista divide a caso 15 ciclisti in 3 gruppi di 5. Poi per 3 settimane ne modifica l'alimentazione come segue: al primo gruppo vengono fatte assumere delle vitamine con tutti i pasti; il secondo riceve istruzioni di consumare dei cereali integrali ad alto contenuto di fibre; il terzo è il gruppo di controllo e si alimenta normalmente. Alla fine di questo periodo di tempo, tutti i ciclisti vengono cronometrati su un percorso di 6 miglia, ottenendo i tempi seguenti:

Vitamine	15.6	16.4	17.2	15.5	16.3
Cereali integrali	17.1	16.3	15.8	16.4	16.0
Controllo	15.9	17.2	16.4	15.4	16.8

(a) Questi dati sono compatibili con l'ipotesi che né le vitamine né i cereali ad alto contenuto di fibre influenzino le prestazioni dei ciclisti? Usa il 5% di significatività.

(b) Trova degli intervalli di valori per tutte le differenze $\mu_i - \mu_j$ che siano contemporaneamente validi con il 95% di confidenza.

15. Verifica l'ipotesi che questi tre campioni indipendenti provengano tutti dalla stessa popolazione normale.

Campione 1	35	37	29	27	30
Campione 2	29	38	34	30	32
Campione 3	44	52	56		

16. Assegnati dei numeri reali x_{ij}, per $i = 1, 2, \ldots, m$ e $j = 1, 2, \ldots, n$, dimostra che

$$x_{**} = \frac{1}{m}\sum_{i=1}^{m} x_{i*} = \frac{1}{n}\sum_{j=1}^{n} x_{*j}$$

17. Ponendo $x_{ij} = i + j^2$, determina

(a) $\sum_{j=1}^{3}\sum_{i=1}^{2} x_{ij}$,

(b) $\sum_{i=1}^{2}\sum_{j=1}^{3} x_{ij}$.

18. Ponendo $x_{ij} = a_i + b_j$, dimostra che

$$\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} = n\sum_{i=1}^{m} a_i + m\sum_{j=1}^{n} b_j$$

19. Si conduce uno studio sull'estrazione della piretrina – una sostanza impiegata come pesticida – dai fiori di piretro. Vengono provati 4 metodi di estrazione su campioni ottenuti da fiori in 3 stati di conservazione: (1) fiori freschi, (2) conservati un anno, (3) trattati e conservati per un anno. Il contenuto percentuale di piretrina che ne è risultato è il seguente.

Metodo di estrazione	A	B	C	D
Stato di conservazione 1	1.35	1.13	1.06	0.98
Stato di conservazione 2	1.40	1.23	1.26	1.22
Stato di conservazione 3	1.49	1.46	1.40	1.35

Assumi che non vi siano interazioni. Suggerisci un modello che descriva le informazioni precedenti e usa i dati per stimare i parametri.

20. I dati seguenti riportano il numero di decessi ogni 10 000 adulti in una grande città degli Stati Uniti orientali, divisi per anno e per stagione.

Anno	Inverno	Primavera	Estate	Autunno
1982	33.6	31.4	29.8	32.1
1983	32.5	30.1	28.5	29.9
1984	35.3	33.2	29.5	28.7
1985	34.4	28.6	33.9	30.1
1986	37.3	34.1	28.5	29.4

(a) Assumi un modello a due fattori e stimane i parametri.

(b) Verifica al 5% di significatività l'ipotesi che la mortalità non sia influenzata dalla stagione.

(c) Verifica al 5% di significatività l'ipotesi che la mortalità non sia influenzata dal passare degli anni.

21. Fai riferimento al Problema 19.

(a) Puoi dire che i metodi di estrazione abbiano efficacia diversa?

(b) Al 5% di significatività diresti che la quantità di sostanza estratta dipende dallo stato di conservazione?

22. Si provano 3 diverse macchine pulitrici con 4 tipi di detergenti. La tabella seguente riporta l'efficacia di pulitura in una scala opportuna.

Macchina	1	2	3
Detergente 1	53	50	59
Detergente 2	54	54	60
Detergente 3	56	58	62
Detergente 4	50	45	57

(a) Stima l'incremento di punteggio medio se si utilizza il detergente 1, rispetto al 2, al 3 e al 4.

(b) Stima l'incremento di punteggio medio utilizzando la macchina numero 3, rispetto alla numero 1 e alla numero 2.

(c) Verifica al 5% di significatività l'ipotesi che il punteggio non sia influenzato dal detergente scelto.

(d) Verifica al 5% di significatività l'ipotesi che il punteggio non sia influenzato dalla macchina impiegata.

23. Si effettua una sperimentazione su 3 tipi di benzine, ciascuna delle quali viene provata in combinazione con 3 additivi diversi. Vengono impiegati in totale 9 motori identici, ogni volta con 5 galloni di carburante, e si ottengono i dati seguenti.

Miglia percorse

Additivo	1	2	3
Benzina 1	124.1	131.5	127.0
Benzina 2	126.4	130.6	128.4
Benzina 3	127.2	132.7	125.6

(a) Verifica l'ipotesi che la benzina scelta non influenzi l'autonomia.

(b) Verifica l'ipotesi che i diversi additivi siano equivalenti.

(c) Che cosa stai implicitamente assumendo?

24. Supponi che nel Problema 6 i 10 soggetti nei due campioni fossero 5 maschi e 5 femmine, con i dati suddivisi in questo modo:

	Femmine					Maschi				
Dieta 1	7.6	8.8	12.5	16.1	18.6	22.2	23.4	24.2	32.2	9.4
Dieta 2	19.5	17.6	16.8	13.7	21.5	30.1	24.2	9.5	14.6	11.2

(a) Verifica l'ipotesi che non vi sia interazione tra il tipo di dieta e il sesso del soggetto.

(b) Verifica se la dieta ha lo stesso effetto su maschi e femmine.

25. Un ricercatore vuole confrontare la resistenza alla rottura dei laminati prodotti con varietà di legno e 3 tipi diversi di colla. Per riuscirci, produce 5 esemplari per ciascun delle 9 combinazioni di legno e colla, quindi li sottopone ad un test di sollecitazione, e misura i seguenti valori della pressione di rottura:

		Tipo di colla 1			2			3		
Legno	**A**	196	208	247	214	216	235	258	250	264
		216	221		240	252		248	272	
	B	216	228	240	215	217	235	246	247	261
		224	236		219	241		250	255	
	C	230	242	232	212	218	216	255	251	261
		244	228		224	222		258	247	

(a) Verifica l'ipotesi che gli effetti del legno e della colla siano additivi.

(b) Verifica l'ipotesi che la scelta del legno non influenzi la resistenza alla rottura del laminato finale.

(c) Verifica se il tipo di colla influenza la pressione di rottura.

26. Si effettua uno studio per determinare la capacità di smaltimento di un certo farmaco da parte dell'organismo umano. Si misura la sua concentrazione nel sangue 24 ore dopo l somministrazione, in varie fascie di età e distinguendo tra maschi e femmine. Vengon riscontrati i valori seguenti (in milligrammi per centimetro cubo).

	Fascia di età 11–25		26–40		41–65		oltre 65	
Maschi	52.0	56.6	52.5	49.6	53.2	53.6	82.4	86.2
	68.2	82.5	48.7	44.6	49.8	50.0	101.3	92.4
	85.6		43.4		51.2		78.6	
Femmine	68.6	80.4	60.2	58.4	58.7	55.9	82.2	79.6
	86.2	81.3	56.2	54.2	56.0	57.2	81.4	80.6
	77.2		61.1		60.0		82.2	

(a) Verifica l'ipotesi che non vi siano interazioni in atto tra sesso ed età.

(b) Verifica se il sesso del soggetto influenza la concentrazione media.

(c) Verifica l'ipotesi che l'età non influenzi la concentrazione media.

27. Nel Problema 23, supponiamo che vi siano state delle controversie sull'assunzione fatta che non vi siano interazioni tra benzine e additivi. Per contemplare anche il caso non additivo, si ripete l'esperimento con 36 motori, 4 per ciascuna combinazione benzina-additivo, trovando i risultati presentati qui sotto.

		Additivo 1		Additivo 2		Additivo 3	
Benzina	1	126.2	124.8	130.4	131.6	127.0	126.6
		125.3	127.0	132.5	128.6	129.4	130.1
	2	127.2	126.6	142.1	132.6	129.5	142.6
		125.8	128.4	128.5	131.2	140.5	138.7
	3	127.1	128.3	132.3	134.1	125.2	123.3
		125.1	124.9	130.6	133.0	122.6	120.9

(a) Puoi concludere che vi sia effetto di interazione?

(b) Ti sembra che le benzine diano risultati analoghi?

(c) Verifica se gli additivi abbiano effetti diversi.

(d) Che conclusioni trai?

28. Si realizza un esperimento per studiare se cure a base di ossigeno possano migliorare la capacità di memorizzazione delle persone anziane. Si scelgono 20 donne e 20 uomini anziani, che vengono divisi in 4 gruppi di 5, e sottoposti a trattamenti di 0, 1, 2 e 3 settimane rispettivamente. Nessun soggetto è in grado di stabilire di che gruppo fa parte, perché tutti sono convinti di ricevere i trattamenti per tutte e tre le settimane. Gli uomini e le donne che ricevono "zero" settimane di trattamenti sono il gruppo di controllo. I risultati trovati sono riportati nella seguente tabella.

	Settimane di trattamento 0			1			2			3		
Maschi	42	54	46	39	52	51	38	50	47	42	55	39
	38	51		50	47		45	43		38	51	
Femmine	49	44	50	48	51	52	27	42	47	61	55	45
	45	43		54	40		53	58		40	42	

(a) Verifica se vi sia effetto di interazione oppure no.

(b) Verifica l'ipotesi che la durata dei trattamenti non abbia influenza sulla capacità di memorizzazione.

(c) Si nota qualche differenza tra maschi e femmine?

(d) Un gruppo di 5 maschi anziani scelto a caso viene sottoposto al test sulla memorizzazione senza ricevere alcun trattamento. I punteggi registrati sono 37, 35, 33, 39, 29. Che conclusioni puoi trarre?

29. In uno studio[7] sull'influenza di fattori come l'altitudine sulla produzione di piastrine, 16 ratti vennero tenuti in un laboratorio a 15 000 piedi di altitudine e altri 16 al livello del mare. La metà dei ratti di ciascun gruppo era stata privata della milza. I dati qui sotto rappresentano il livello di fibrinogeno (in centesimi di milligrammo) riscontrati il giorno 21.

[7] K. Rand, T. Anderson, G. Lukis, W. Creger, "Effect of hypoxia on platelet level in the rat", *Clinical Research*, vol. 18, p. 178, 1970.

	Privi di milza				Normali			
In altitudine	528	444	338	342	434	331	312	575
	338	331	288	319	472	444	575	384
Al livello del mare	294	254	352	241	272	275	350	350
	291	175	241	238	466	388	425	344

(a) Verifica l'ipotesi che non vi siano interazioni.

(b) Verifica se vi sia qualche effetto dovuto all'altitudine.

(c) Verifica l'ipotesi che non vi sia alcun effetto dovuto alla rimozione della milza.

Usa in tutti e tre i casi il 5% di significatività.

11 Verifica del modello e test di indipendenza

Contenuto

11.1 Introduzione

In questo capitolo vogliamo imparare a riconoscere quando un modello probabilistico si adatta ad un certo fenomeno casuale. Questa ricerca consiste spesso nel verificare se un campione aleatorio assegnato possa realisticamente provenire da una certa distribuzione di probabilità. Per fare un esempio, potrebbe esserci motivo di credere (a priori) che il numero di incidenti che si verificano giornalmente in un impianto industriale sia una variabile aleatoria di Poisson: questa convinzione può essere verificata osservando per un certo periodo il numero di incidenti, ed eseguendo quindi un test che sia in grado di stabilire se la popolazione possa avere questo tipo di distribuzione. I test statistici che servono a verificare se un dato modello probabilistico sia compatibile con i dati sono detti *test sulla bontà di adattamento*[1].

L'approccio classico per verificare l'ipotesi nulla che un campione provenga da una distribuzione di probabilità assegnata, consiste nel partizionare i valori possibili in un numero finito di regioni (in maniera analoga agli intervalli di classe della Sezione 2.2.3); si determina poi quanti elementi del campione appartengono a ciascuna

[1] È molto usata pure la forma inglese, *goodness of fit tests*, [*N.d.T.*]

regione e si confrontano questi valori con le previsioni teoriche nell'ipotesi che la distribuzione fosse quella in esame. L'ipotesi nulla viene rifiutata quando le differenze che si riscontrano sono significative.

I dettagli su questo tipo di test sono affrontati nella Sezione 11.2, dove si assume che l'ipotesi nulla consista di una specificazione completa della distribuzione. Nella Sezione 11.3 generalizziamo l'analisi ai casi in cui l'ipotesi nulla specifica la famiglia parametrica della distribuzione, senza fissarne tutti i parametri; ad esempio ci si potrebbe domandare se una popolazione sia normale, senza volersi limitare ad una particolare scelta di media e varianza. All'interno delle Sezioni 11.4 e 11.5 consideriamo le situazioni in cui gli elementi di una popolazione sono classificabili secondo due variabili, eventualmente collegate tra loro (come la *statura* e il *peso* della popolazione dei maschi americani adulti); l'analisi precedente viene impiegata per verificare l'ipotesi che scegliendo un membro a caso della popolazione, le due caratteristiche risultino tra loro indipendenti. Il test per stabilire se m popolazioni distinte abbiano la stessa distribuzione discreta si ottiene come applicazione di questo formalismo. La Sezione 11.6, che chiude il capitolo, è facoltativa, e torna al problema iniziale di verificare la bontà di adattamento tra il campione ed una distribuzione continua assegnata; anziché usare la discretizzazione e le metodologie della Sezione 11.2, viene introdotto il test di *Kolmogorov-Smirnov*.

11.2 Test di adattamento ad una distribuzione completamente specificata

Consideriamo un esperimento che consista nell'osservare n variabili aleatorie indipendenti $Y_1, Y_2, \ldots, Y_n$, che possono assumere i valori $1, 2, \ldots, k$. Siamo interessati a verificare l'ipotesi nulla che $\{p_i, i = 1, \ldots, k\}$ sia la funzione di massa di probabilità delle Y_j, quindi se Y rappresenta una qualunque delle Y_j, l'ipotesi nulla e quella alternativa sono:

$$\begin{aligned} H_0 &: P(Y = i) = p_i, \qquad \text{per ogni } i = 1, 2, \ldots, k \\ H_1 &: P(Y = i) \neq p_i, \qquad \text{per qualche } i = 1; 2, \ldots, k \end{aligned} \tag{11.2.1}$$

Per realizzare questo test denotiamo con X_i, per $i = 1, 2, \ldots, k$, il numero delle Y_j che sono uguali ad i. Se H_0 è soddisfatta, ciascuna delle Y_j assume il valore i con probabilità p_i indipendentemente da tutte le altre, quindi X_i è binomiale di parametri n e p_i, e il suo valore atteso è np_i. Di conseguenza, $(X_i - np_i)^2$ è un indicatore di quanto sia verosimile che p_i sia davvero la probabilità dell'evento $\{Y = i\}$. Quando questi quadrati hanno valori troppo elevati, ci suggeriscono che H_0 può non essere corretta; è quindi naturale che la statistica per il test sia una somma pesata di questi k contributi; quali siano i pesi giusti non è ovvio, ma la conclusione (sulla quale

torniamo nelle osservazioni) è che la statistica da adottare è la seguente:

$$T := \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \tag{11.2.2}$$

L'ipotesi nulla va rifiutata quando T è troppo grande, e il valore di soglia dipende dal livello di significatività richiesto. Sia infatti α il livello di significatività del test, allora per trovare la regione critica, dobbiamo calcolare un valore c tale che

$$P_{H_0}(T \geq c) = \alpha$$

ovvero tale che quando H_0 è vera, T sia superiore a c con probabilità α. Fatto questo, il test dovrà rifiutare l'ipotesi nulla quando il valore osservato per T sia superiore a c.

Il valore critico che cerchiamo si trova usando il fatto che quando n è grande, la distribuzione di T è approssimativamente quella di una chi-quadrocon $k - 1$ gradi di libertà, e l'approssimazione migliora con il crescere di n. Allora nell'ipotesi che n sia un numero abbastanza elevato, $c \sim \chi^2_{\alpha,k-1}$ e quindi un test approssimato con significatività α deve

$$\begin{aligned} &\text{rifiutare } H_0 \text{ se } T > \chi^2_{\alpha,k-1} \\ &\text{accettare } H_0 \text{ se } T \leq \chi^2_{\alpha,k-1} \end{aligned} \tag{11.2.3}$$

Ovvero, se si vuole usare il p-dei-dati, si denota con t il valore assunto da T, e si calcola

$$p\text{-dei-dati} \approx P(\chi^2_{k-1} \geq t) \tag{11.2.4}$$

Una regola empirica comunemente accettata per sapere quando n è sufficientemente grande da rendere utile questa approssimazione, è che almeno l'80% delle np_i dovrebbero essere maggiori di 5, e le restanti dovrebbero essere tutte maggiori di 1.

Osservazione 11.2.1.

(a) Una formula computazionalmente valida per il calcolo di T può essere ottenuta dall'Equazione (11.2.2) svolgendo il quadrato e struttando le due identità (lo studente si convinca della seconda) $\sum_i p_i = 1$ e $\sum_i X_i = n$:

$$\begin{aligned} T &= \sum_{i=1}^{k} \frac{X_i^2 - 2np_iX_i + n^2p_i^2}{np_i} \\ &= \sum_{i=1}^{k} \frac{X_i^2}{np_i} - 2\sum_{i=1}^{k} X_i + n\sum_{i=1}^{k} p_i \\ &= \sum_{i=1}^{k} \frac{X_i^2}{np_i} - n \end{aligned} \tag{11.2.5}$$

(b) Il fatto che T, nonostante sia costruita sulle k variabili aleatorie $X_1, X_2, \ldots, X_k$, tenda ad una chi-quadrocon soli $k-1$ gradi di libertà, è dovuto alla relazione lineare $\sum_i X_i = n$, che fa "perdere" un grado di libertà.

(c) La dimostrazione che T ha asintoticamente distribuzione χ^2_{k-1} è piuttosto avanzata, con l'eccezione del caso $k = 2$, che illustriamo rapidamente. In tali ipotesi, visto che $X_1 + X_2 = n$ e $p_1 + p_2 = 1$, si ha che

$$\begin{aligned} T &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(n - X_1 - n + np_1)^2}{n(1-p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{n(1-p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1(1-p_1)} \qquad \text{infatti } \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)} \end{aligned}$$

Ma poiché X_1 è binomiale di media np_1 e varianza $np_1(1-p_1)$, per l'approssimazione normale, quando n tende all'infinito, $(X_1 - np_1)/\sqrt{np_1(1-p_1)}$ tende ad avere distribuzione $\mathcal{N}(0,1)$, e quindi T che è il suo quadrato, tende ad una chi-quadro con 1 grado di libertà.

Esempio 11.2.1. Negli anni recenti è sempre più studiata la correlazione tra il benessere mentale e fisico nell'uomo. L'analisi che segue può essere vista come una prova di questo legame; studiamo infatti le date di nascita e di morte di persone scelte nella categoria di quelle "famose". È ragionevole supporre che l'attesa di un lieto evento migliori lo stato d'animo delle persone, e un uomo o una donna famosi vedono probabilmente nel loro compleanno un evento sostanzialmente gradevole, a causa delle attenzioni e dell'affetto che li circondano in tali occasioni. Se una persona famosa fosse gravemente malata e prossima a morire, l'attesa per il proprio compleanno potrebbe sollevarne il morale, migliorarne il benessere mentale (e forse di conseguenza anche quello fisico), abbastanza da diminuire sensibilmente la probabilità di morire poco prima di compiere gli anni. È quindi possibile che i dati mostrino che una persona famosa abbia meno possibilità di morire nei mesi immediatamente precedenti a quelli del suo compleanno, che in quelli successivi.

Per verificare questa ipotesi, si è scelto dal *Who Was Who in America* un campione casuale di 1 251 americani deceduti, e si sono annotate date di nascita e di morte. I dati sono riassunti nella Tabella 11.1, che ci dice per esempio che solo 86 soggetti morirono nel mese precedente al loro compleanno.

Se il giorno della morte non dipendesse da quello di nascita, sembrerebbe ragionevole che ciascuno dei 1 251 individui abbia avuto le stesse probabilità di cadere

Tabella 11.1 Numero di decessi nei mesi precedenti e successivi a quello di nascita

(I mesi di differenza sono stati ottenuti sottraendo quello del decesso da quello del compleanno: un valore negativo indica che il decesso ha preceduto il compleanno di qualche mese.)

Mesi di differenza	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5
Numero di decessi	90	100	87	96	101	86	119	118	121	114	113	106

nelle 12 categorie. Verifichiamo allora l'ipotesi nulla seguente:

$$H_0 : p_i \equiv \frac{1}{12}, \qquad i = 1, 2, \ldots, 12$$

Siccome $np_i = 1\,251/12 = 104.25$, la statistica di questo test è

$$\begin{aligned} T &= \frac{90^2 + 100^2 + 87^2 + \cdots + 106^2}{104.25} - 1\,251 \\ &\approx 17.192 \end{aligned}$$

Il p-dei-dati è allora

$$\begin{aligned} p\text{-dei-dati} &\approx P(\chi^2_{11} \geq 17.192) \\ &\approx 1 - 0.8977 = 0.1023 \qquad \text{usando il Programma 5.8.1a} \end{aligned}$$

Il risultato del test appena eseguito suggerisce che il compleanno non influisca sulla data di morte, ma non è del tutto convincente. Infatti, anche se i dati non sono forti abbastanza (ad esempio non lo sono al 10% di significatività) da escludere l'ipotesi nulla, ci lasciano il dubbio di una sua possibile falsità. Potremmo allora pensare di usare meno di 12 categorie, in modo da ottenere forse un test più potente. In effetti, se avessimo codificato in 4 categorie in questo modo:

$$\begin{aligned} \text{esito } 1 &= \{-6, -5, -4\} \\ \text{esito } 2 &= \{-3, -2, -1\} \\ \text{esito } 3 &= \{0, 1, 2\} \\ \text{esito } 4 &= \{3, 4, 5\} \end{aligned}$$

i dati avrebbero assunto la frequenza seguente,

Esito	1	2	3	4
Frequenza	277	283	358	333

La statistica del test sarebbe stata

$$\begin{aligned} T &= \frac{277^2 + 283^2 + 358^2 + 333^2}{1\,251/4} - 1\,251 \\ &\approx 14.775 \end{aligned}$$

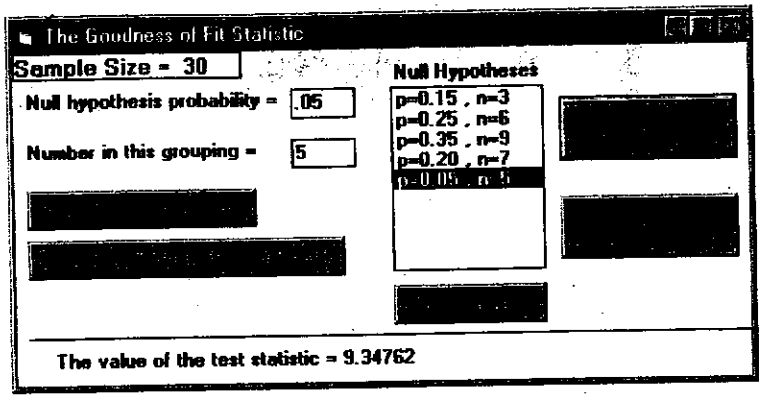


Figura 11.1

Poiché $\chi^2_{0.01,3} \approx 11.345$, l'ipotesi nulla verrebbe in questo caso rifiutata anche all'1% di significatività. E infatti il Programma 5.8.1a ci dice che

$$\begin{aligned} p\text{-dei-dati} &\approx P(\chi^2_3 \geq 14.775) \\ &\approx 1 - 0.998 = 0.002 \end{aligned}$$

L'analisi appena conclusa è però suscettibile di critiche, in quanto l'ipotesi nulla è stata scelta dopo avere osservato i dati. In effetti, mentre non vi è nulla di sbagliato nell'utilizzare un campione di dati per individuare il modo "corretto" di formulare l'ipotesi nulla, usare poi quello stesso campione per eseguire il test di tale ipotesi è quanto meno opinabile. Perciò, per essere ragionevolmente sicuri delle conclusioni che vorremmo trarre, sarebbe opportuno scegliere un secondo campione aleatorio, codificarlo come in precedenza in 4 regioni e verificare nuovamente l'ipotesi H_0 che $p_i = \frac{1}{4}$, per $i = 1, 2, 3, 4$ (si veda il Problema 3). □

Il Programma 11.2.1 serve a calcolare il valore di T.

Esempio 11.2.2. Un produttore di lampade a incandescenza informa i suoi clienti che la qualità dei suoi prodotti non è uniforme, e che ogni lampadina può essere indipendentemente di qualità A, B, C, D o E con probabilità del 15%, 25%, 35%, 20% e 5% rispettivamente. Tuttavia uno dei clienti, acquistando grossi volumi di merce, ha l'impressione di ricevere troppi pezzi di qualità E (la peggiore), e quindi decide di verificare l'affermazione del produttore investendo tempo e denaro per stabilire il livello qualitativo di 30 lampade. Supponiamo che ve ne siano 3 di qualità A, 6 di qualità B, 9 di qualità C, 7 di qualità D e 5 di qualità E. Al 5% di significatività cosa si decide?

Il Programma 11.2.1 con la schermata in Figura 11.1 fornisce per la statistica del test un valore di 9.348 circa. Il p-dei-dati corrispondente può essere ottenuto dal

Programma 5.8.1a nel modo usuale:

$$\begin{aligned} p\text{-dei-dati} &\approx P(\chi^2_4 \geq 9.348) \\ &\approx 1 - 0.947 = 0.053 \end{aligned}$$

Facendoci concludere che l'ipotesi nulla non può essere rifiutata al 5% di significatività (ma siccome essa sarebbe rifiutata a qualunque livello di significatività superiore al 5.3%, il cliente dovrà certamente rimanere scettico). □

11.2.1 Determinazione della regione critica per simulazione

Dal 1900 quando Karl Pearson dimostrò che T ha approssimativamente distribuzione chi-quadrocon $k-1$ gradi di libertà (approssimazione che diventa esatta al tendere di n all'infinito), fino a molto recentemente, questa approssimazione era l'unico metodo disponibile per determinare il p-dei-dati di un test di adattamento. Tuttavia con l'avvento della potenza di calcolo degli elaboratori moderni (economici, veloci, e diffusissimi), si è aperta una seconda strada che permette la determinazione del p-dei-dati con una precisione potenzialmente migliore: il metodo della *simulazione*.

L'approccio è il seguente. Per prima cosa si determina il valore t assunto dalla statistica del test T. Per calcolare il p-dei-dati, è necessario determinare la probabilità che, essendo valida H_0, T assuma valori superiori a t. Si simulano perciò n variabili aleatorie indipendenti $Y_1^{(1)}, Y_2^{(1)}, \ldots, Y_n^{(1)}$, ciascuna con funzione di massa di probabilità $\{p_i, i = 1, 2, \ldots, k\}$, ovvero

$$P(Y_j^{(1)} = i) = p_i, \qquad i = 1, 2, \ldots, k, \qquad j = 1, 2, \ldots, n$$

e si pone, per $i = 1, 2, \ldots, k$,

$$X_i^{(1)} := \text{numero degli indici } j \text{ tali che } Y_j^{(1)} = i$$

$$T^{(1)} := \sum_{i=1}^{k} \frac{\left(X_i^{(1)} - np_i\right)^2}{np_i}$$

Si ripete quindi la procedura simulando un secondo campione $Y_1^{(2)}, Y_2^{(2)}, \ldots, Y_n^{(2)}$, indipendente dal primo e con le stesse caratteristiche, e si calcola $T^{(2)}$. Iterando il procedimento un numero r di volte, otteniamo r variabili aleatorie indipendenti $T^{(1)}, T^{(2)}, \ldots, T^{(r)}$, ciascuna delle quali ha la distribuzione di T quando H_0 è soddisfatta. Perciò per la legge dei grandi numeri, la percentuale di tali variabili aleatorie che supera t sarà molto prossima alla probabilità che $T > t$, sotto H_0. Vale a dire,

$$p\text{-dei-dati} = P_{H_0}(T > t) \approx \frac{\text{numero degli indici } l \text{ tali che } T^{(l)} > t}{r}$$

In effetti se r è abbastanza grande, questa approssimazione può essere considerata a tutti i fini pratici un'uguaglianza, e quindi l'ipotesi nulla va rifiutata se la percentuale di variabili $T^{(i)}$ che sono maggiori di t è minore o uguale al livello di significatività α.

Osservazione 11.2.2.

(a) Per poter usare il metodo della simulazione al calcolatore descritto qui sopra, occorre essere in grado di simulare o generare al calcolatore una variabile aleatoria Y tale che $P(Y = i) = p_i$, per $i = 1, 2, \ldots, k$. Quello illustrato di seguito è uno dei possibili metodi per riuscirci partendo dalle variabili aleatorie uniformi sull'intervallo $(0, 1)$, che tutti i computer possono generare.

Passo 1: Si genera un numero casuale U.

Passo 2: Se $U < p_1$ si pone $Y = 1$; se $p_1 \leq U < p_1 + p_2$ si pone $Y = 2$; in generale, se

$$p_1 + p_2 + \cdots + p_{i-1} \leq U < p_1 + p_2 + \cdots + p_{i-1} + p_i$$

si pone $Y = i$. Siccome U ha distribuzione uniforme sull'intervallo $(0, 1)$, per ogni scelta di $0 \leq a < b \leq 1$,

$$P(a < X < b) = b - a$$

poiché inoltre $0 \leq p_1 + p_2 + \cdots + p_i \leq 1$ per ogni scelta di i,

$$\begin{aligned} P(Y = i) &= P(p_1 + p_2 + \cdots + p_{i-1} \leq U < p_1 + p_2 + \cdots + p_{i-1} + p_i) \\ &= (p_1 + p_2 + \cdots + p_{i-1} + p_i) - (p_1 + p_2 + \cdots + p_{i-1}) = p_i \end{aligned}$$

esattamente come desiderato.

(b) Una domanda importante a cui non abbiamo ancora risposto è quanti cicli di simulazione siano in effetti necessari. È stato dimostrato che per un livello di significatività del 5%, un valore di r intorno al centinaio è normalmente sufficiente[2].

Esempio 11.2.3. Consideriamo nuovamente i dati dell'Esempio 11.2.2. Una simulazione al calcolatore fornisce questo risultato:

$$P_{H_0}(T \leq 9.52381) = 0.95$$

[2] A. Hope, "A simplified Monte Carlo significance test procedure", *J. of Royal Statist. Soc.*, vol. B 30, pp. 582–598, 1968.

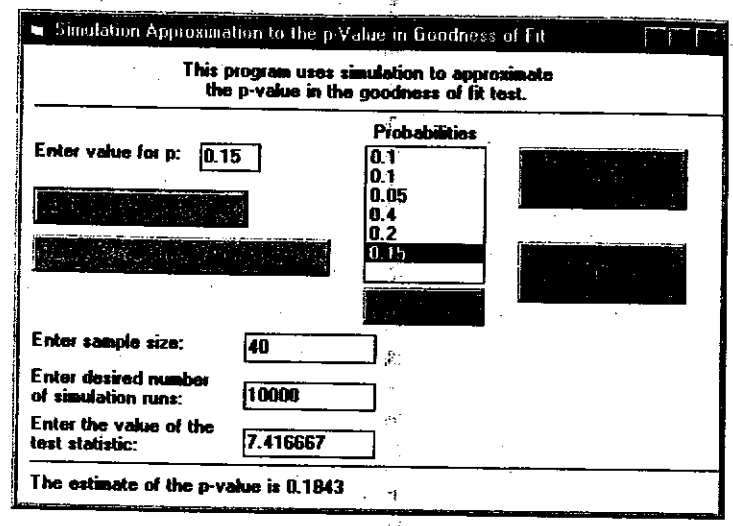


Figura 11.2

Quindi l'estremo della regione critica dovrebbe essere 9.52381, che è assai vicino a $\chi^2_{0.05,4} \approx 9.488$ (il valore critico approssimato che si ottiene dalla distribuzione chi-quadro). Questo risultato è molto interessante, in quanto in questo esempio la regola empirica per applicare l'approssimazione, che l'80% dei valori np_i sia almeno pari a 5, non vale, fornendo un indicazione che le richieste di tale regola siano piuttosto prudenziali. □

Il Programma 11.2.2 permette di ottenere il p-dei-dati per un test di questo tipo, usando il metodo della simulazione.

Esempio 11.2.4. Consideriamo un esperimento che ha 6 possibili esiti, le cui rispettive probabilità sono ipotizzate valere 0.1, 0.1, 0.05, 0.4, 0.2 e 0.15. Si effettua un test replicando 40 volte l'esperimento, e si ottiene che gli esiti nell'ordine si realizzano 3, 3, 5, 18, 4 e 7 volte. Va accettata l'ipotesi nulla?

Un calcolo diretto, ovvero l'impiego del Programma 11.2.1 ci dice che il valore della statistica del test è 7.4167. Usando il Programma 5.8.1a otteniamo il risultato che

$$P(\chi^2_5 \leq 7.4167) \approx 0.8088$$

e quindi il p-dei-dati vale approssimativamente 0.1912. Per controllare la bontà di questa approssimazione, lanciamo il Programma 11.2.2, facendogli eseguire 10 000 cicli di simulazione; in questo modo otteniamo una stima del p-dei-dati di 0.1843 (si veda la Figura 11.2).

Poiché il numero dei valori simulati che superano 7.4167 è una variabile aleatoria binomiale di parametri $n = 10^4$ e $p = p$-dei-dati, ne segue che un intervallo di confidenza al 90% per il p-dei-dati stesso è il seguente,

$$0.1843 \pm 1.645\sqrt{0.1843 \times 0.8157/10^4}$$

Perciò con il 90% di confidenza,

$$p\text{-dei-dati} \in (0.1779,\ 0.1907) \quad \square$$

11.3 Test di adattamento ad una distribuzione specificata a meno di parametri

Si può effettuare un test di adattamento anche se le probabilità $\{p_i, i = 1, 2, \ldots, k\}$ non sono completamente specificate. Ne è un esempio la situazione citata all'inizio del capitolo, in cui si voleva capire se il numero di incidenti quotidiani in un impianto fosse una variabile aleatoria di Poisson. Non si chiede quindi se la distribuzione sia di Poisson con una media λ in particolare (una tale H_0 specificherebbe tutte le p_i), ma ci si domanda in generale se si possa trattare di una *qualsiasi* distribuzione poissoniana. Supponiamo allora di raccogliere dei dati per n giorni, e denotiamo con $Y_1, Y_2, \ldots, Y_n$ il numero di incidenti registrati. La prima difficoltà è che se la distribuzione deve essere di Poisson, non esiste un k che limiti i valori delle Y_j, che possono essere arbitrariamente alti. Si codificano quindi gli esiti delle Y_j in un numero finito di regioni, ad esempio regione 1 se vi sono stati 0 incidenti, regione 2 con 1 incidente, regione 3 con 2 o 3 incidenti, regione 4 con 4 o 5 incidenti e regione 5 se vi sono stati 6 o più incidenti. Se la distribuzione è realmente di Poisson con media λ, le probabilità delle diverse regioni sono allora:

$$\begin{aligned}
p_1 &= P(Y = 0) = e^{-\lambda} \\
p_2 &= P(Y = 1) = \lambda e^{-\lambda} \\
p_3 &= P(Y = 2) + P(Y = 3) = \frac{\lambda^2}{2}e^{-\lambda} + \frac{\lambda^3}{6}e^{-\lambda} \\
p_4 &= P(Y = 4) + P(Y = 5) = \frac{\lambda^4}{24}e^{-\lambda} + \frac{\lambda^5}{120}e^{-\lambda} \\
p_5 &= P(Y \geq 6) = 1 - e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2} + \frac{\lambda^3}{6} + \frac{\lambda^4}{24} + \frac{\lambda^5}{120}\right)
\end{aligned}$$

La seconda difficoltà è che il valore medio λ non è specificato da H_0. La strada più intuitiva in questo caso è anche quella giusta: supponendo vera H_0 si può produrre una stima $\widehat{\lambda}$ del parametro incognito λ usando metodi parametrici (come il criterio di massima verosimiglianza), ricavare i corrispondenti valori per le $\widehat{p}_i$, sostituendo $\widehat{\lambda}$ al posto di λ nelle equazioni precedenti, e quindi calcolare la statistica del test, definita come

$$T := \sum_{i=1}^{k} \frac{(X_i - n\widehat{p}_i)^2}{n\widehat{p}_i} \tag{11.3.1}$$

dove X_i indica, come già in precedenza, il numero di dati Y_j che appartengono alla regione i.

L'approccio qui descritto può in generale essere impiegato quando l'ipotesi nulla non specifica dei parametri che sono indispensabili al calcolo delle p_i. Supponiamo che vi siano m parametri non specificati, che abbiamo comunque stimato con il metodo della massima verosimiglianza. Se si usano queste stime per calcolare le probabilità $\widehat{p}_i$, si può dimostrare che, sotto H_0, la statistica T tende, al crescere di n, ad avere approssimativamente distribuzione chi-quadro con $k - 1 - m$ gradi di libertà. (Si perdono tanti gradi di libertà quanti sono gli stimatori indipendenti usati al posto dei parametri.)

Un test di adattamento con livello di significatività α deve quindi

$$\begin{aligned}
&\text{rifiutare } H_0 \text{ se } T > \chi^2_{\alpha,k-1-m} \\
&\text{accettare } H_0 \text{ se } T \leq \chi^2_{\alpha,k-1-m}
\end{aligned} \tag{11.3.2}$$

Un modo equivalente di realizzare il test consiste – come al solito – nel calcolare il valore t assunto dalla statistica T, e quindi definire il p-dei-dati come

$$p\text{-dei-dati} \approx P(\chi^2_{k-1-m} \geq t) \tag{11.3.3}$$

Se α è maggiore del p-dei-dati, si rifiuta l'ipotesi nulla, altrimenti la si accetta.

Esempio 11.3.1. Supponiamo che il numero di incidenti settimanali in un periodo di 30 settimane sia stato il seguente:

8	0	0	1	3	4	0	2	12	5	1	8	0	2	0
1	9	3	4	5	3	3	4	7	4	0	1	2	1	2

Si verifichi l'ipotesi che la distribuzione del numero di incidenti settimanali sia di Poisson.

Poiché il numero totale di incidenti nelle 30 settimane risulta essere 95, lo stimatore di massima verosimiglianza per la media λ della eventuale distribuzione di Poisson è $\widehat{\lambda} = 95/30 \approx 3.16667$. Di conseguenza la stima della funzione di massa è data da

$$P(Y = i) \overset{\text{stima}}{=} \frac{\widehat{\lambda}^i}{i!}e^{-\widehat{\lambda}}$$

e, usando ad esempio le cinque regioni descritte all'inizio della sezione, si trova con qualche calcolo che

$$\widehat{p_1} \approx 0.04214 \qquad \widehat{p_2} \approx 0.13346 \qquad \widehat{p_3} \approx 0.43435$$
$$\widehat{p_4} \approx 0.28841 \qquad \widehat{p_5} \approx 0.10164$$

Usando successivamente i dati codificati, $X_1 = 6$, $X_2 = 5$, $X_3 = 8$, $X_4 = 6$, $X_5 = 5$, si trova per la statistica del test il valore

$$T = \sum_{i=1}^{5} \frac{(X_i - 30\widehat{p_i})^2}{30\widehat{p_i}} \approx 21.99$$

Per determinare il p-dei-dati possiamo usare il Programma 5.8.1a, ottenendo che:

$$\begin{aligned} p\text{-dei-dati} &\approx P(\chi^2_3 > 21.99) \\ &\approx 1 - 0.999936 = 0.000064 \end{aligned}$$

e quindi l'ipotesi che la distribuzione di provenienza fosse poissoniana deve chiaramente essere rifiutata. (Il motivo è che vi sono troppe settimane senza incidenti per poter accettare che la distribuzione fosse di Poisson con media 3.167.) □

11.4 Test per l'indipendenza e tabelle di contingenza

In questa sezione consideriamo situazioni in cui ogni membro di una popolazione può essere classificato secondo due criteri, ovvero in base a due caratteristiche, che vengono denotate con X e Y. Supponiamo che la caratteristica X abbia r valori possibili e la Y abbia s valori possibili; indichiamo allora con P_{ij} la probabilità che per un elemento a caso della popolazione, X assuma il valore i e Y assuma il valore j, con $i = 1, 2, \ldots, r$ e $j = 1, 2, \ldots, s$:

$$P_{ij} := P(X = i, Y = j) \tag{11.4.1}$$

Elementi diversi della popolazione vengono supposti indipendenti come al solito; le due caratteristiche di un singolo elemento invece non sono in generale indipendenti, anzi il nostro obiettivo consiste precisamente nel verificare se esse lo siano oppure no. Denotiamo quindi con p_i e q_j le funzioni di massa marginali di questa distribuzione congiunta:

$$\begin{aligned} p_i &:= P(X = i) = \sum_{j=1}^{s} P_{ij} \\ q_j &:= P(Y = j) = \sum_{i=1}^{r} P_{ij} \end{aligned} \tag{11.4.2}$$

La nostra ipotesi nulla consisterà nell'indipendenza di X e Y, e quindi (si veda l'Equazione (4.3.14) a pagina 105):

$$\begin{aligned} &H_0 : P_{ij} = p_i q_j, \qquad \text{per ogni } i = 1, 2, \ldots, r, \quad j = 1, 2, \ldots, s \\ &H_1 : P_{ij} \neq p_i q_j, \qquad \text{per qualche } i = 1, 2, \ldots, r, \quad j = 1, 2, \ldots, s \end{aligned} \tag{11.4.3}$$

È bene notare che questo tipo di ipotesi nulla rientra nella casistica trattata nella Sezione 11.3, in quanto p_i e q_j sono parametri incogniti.

Supponiamo che i dati consistano in un campione di n elementi provenienti dalla popolazione in esame, e al variare di i e j, denotiamo con N_{ij} quanti di essi soddisfano contemporaneamente le condizioni $X = i$ e $Y = j$.

Occorre intanto stimare le quantità p_i e q_j. Sia $1 \leq i \leq r$ qualsiasi; lo stimatore di massima verosimiglianza di p_i è pari alla frazione di elementi del campione la cui caratteristica X vale i. Le due grandezze

$$N_i := \sum_{j=1}^{s} N_{ij} \qquad \widehat{p_i} := \frac{N_i}{n} \tag{11.4.4}$$

rappresentano rispettivamente il numero dei membri del campione per i quali $X = i$, e lo stimatore cercato. Analogamente se $1 \leq j \leq s$, e indichiamo con M_j il numero di elementi del campione la cui caratteristica Y vale j, e con $\widehat{q_j}$ lo stimatore di massima verosimiglianza di q_j, allora

$$M_j := \sum_{i=1}^{r} N_{ij} \qquad \widehat{q_j} := \frac{M_j}{n} \tag{11.4.5}$$

Con queste posizioni, la statistica del test è data da

$$\begin{aligned} T &:= \sum_{j=1}^{s} \sum_{i=1}^{r} \frac{(N_{ij} - n\widehat{p_i}\widehat{q_j})^2}{n\widehat{p_i}\widehat{q_j}} \\ &= \sum_{j=1}^{s} \sum_{i=1}^{r} \frac{N_{ij}^2}{n\widehat{p_i}\widehat{q_j}} - n \end{aligned} \tag{11.4.6}$$

e infatti $E[N_{ij}] = nP_{ij}$ che è uguale a $n\widehat{p_i}\widehat{q_j}$ se H_0 è soddisfatta.

A prima vista potrebbe sembrare che i parametri che devono essere stimati dai dati siano $r + s$, tuttavia, siccome le somme $\sum_{i=1}^{r} p_i$ e $\sum_{j=1}^{s} q_j$ sono pari a 1, proprio come quelle dei corrispondenti stimatori, occorre determinare solo $r - 1$ dei primi e $s - 1$ dei secondi, perché gli ultimi due possono essere ricavati per differenza. Per questo motivo, i gradi di libertà della distribuzione chi-quadro che approssima T

quando n è grande sono $rs - 1 - (r-1) - (s-1) = (r-1)(s-1)$, e quindi un test con significatività α dovrebbe

$$\begin{aligned} &\text{rifiutare } H_0 \text{ se } T > \chi^2_{\alpha,(r-1)(s-1)} \\ &\text{accettare } H_0 \text{ se } T \leq \chi^2_{\alpha,(r-1)(s-1)} \end{aligned} \tag{11.4.7}$$

Esempio 11.4.1. Si sono scelti a caso 300 statunitensi adulti, che sono stati suddivisi per sesso e convinzioni politiche. Una tabella come quella qui sotto è detta *tabella di contingenza*:

	Democratici	**Repubblicani**	**Indipendenti**	
Donne	68	56	32	156
Uomini	52	72	20	144
	120	128	52	300

Una tabella di contingenza riporta normalmente anche i totali per riga e per colonna ed è quindi lo strumento più indicato per studiare l'indipendenza delle categorie di dati. Quella qui presentata ci dice ad esempio che su 300 intervistati, 156 erano donne, e di queste 68 si sono dichiarate democratiche, 56 repubblicane e 32 indipendenti. Volendo usare la notazione delle variabili N_{ij}, detta X la categoria sesso, e Y la categoria politica, ciò significa che $N_{11} = 68$, $N_{12} = 56$ e $N_{13} = 32$. Analogamente si ha che $N_{21} = 52$, $N_{22} = 72$ e $N_{23} = 20$, e anche che $N_1 = 156$, $N_2 = 144$, $M_1 = 120$, $M_2 = 128$ e $M_3 = 52$. Usiamo questi dati per verificare se il sesso e le convinzioni politiche di un americano adulto scelto a caso siano o no indipendenti.

Dai dati della tabella ricaviamo che i sei coefficienti $n\widehat{p_i}\widehat{q_j} = N_i M_j / n$ valgono

$$\frac{N_1 M_1}{n} = \frac{156 \times 120}{300} = 62.40 \qquad \frac{N_2 M_1}{n} = \frac{144 \times 120}{300} = 57.60$$

$$\frac{N_1 M_2}{n} = \frac{156 \times 128}{300} = 66.56 \qquad \frac{N_2 M_2}{n} = \frac{144 \times 128}{300} = 61.44$$

$$\frac{N_1 M_3}{n} = \frac{156 \times 52}{300} = 27.04 \qquad \frac{N_2 M_3}{n} = \frac{144 \times 52}{300} = 24.96$$

Per cui la statistica del test è la seguente,

$$\begin{aligned} T = &\frac{(68-62.40)^2}{62.40} + \frac{(56-66.56)^2}{66.56} + \frac{(32-27.04)^2}{27.04} \\ &+ \frac{(52-57.60)^2}{57.60} + \frac{(72-61.44)^2}{61.44} + \frac{(20-24.96)^2}{24.96} \\ &\approx 6.433 \end{aligned}$$

Siccome $(r-1)(s-1) = 2$, volendo un livello di significatività del 5%, dobbiamo confrontare il valore di T con quello di $\chi^2_{0.05,2}$. La Tabella A.2 ci dice che

$$\chi^2_{0.05,2} \approx 5.991$$

e siccome $T > 5.991$, l'ipotesi nulla viene rifiutata e concludiamo che al 5% di significatività non si può accettare con questi dati l'ipotesi che il sesso e le convinzioni politiche degli americani siano indipendenti. □

Anche di questo tipo di test si può calcolare il p-dei-dati, infatti,

$$\begin{aligned} p\text{-dei-dati} &= P_{H_0}(T > t) \\ &\approx P(\chi^2_{(r-1)(s-1)} > t) \end{aligned} \tag{11.4.8}$$

Un test di H_0 con significatività α deve rifiutare l'ipotesi nulla ogni volta che il p-dei-dati risulta minore di α.

Il Programma 11.4 del software abbinato al libro permette di calcolare il valore di T per i test di indipendenza.

Esempio 11.4.2. Una azienda tiene in funzione 4 macchine (denotate con A, B, C e D) per 3 turni di lavoro ogni giorno. La tabella di contingenza seguente presenta il numero di fermi macchina risultati in un periodo di 6 mesi.

	A	**B**	**C**	**D**	
Turno 1	10	12	6	7	35
Turno 2	10	24	9	10	53
Turno 3	13	20	7	10	50
	33	56	22	27	138

Supponiamo di essere interessati a capire se tutte le macchine tendono a rompersi con elevata probabilità nei medesimi turni, o se piuttosto qualcuna di esse abbia turni critici che le sono propri, e non sono altrettanto problematici per le altre macchine. In altri termini, ci chiediamo se, per una rottura generica, la macchina che l'ha provocata e il turno in cui è avvenuta siano variabili aleatorie indipendenti.

Possiamo calcolare la statistica di questo test direttamente o tramite il Programma 11.4, che ritorna il valore $T \approx 1.8148$ (si veda la Figura 11.3). Usando poi il Programma 5.8.1a otteniamo che

$$\begin{aligned} p\text{-dei-dati} &\approx P(\chi^2_6 > 1.8148) \\ &\approx 1 - 0.0641 = 0.9359 \end{aligned}$$

Perciò va senz'altro accettata l'ipotesi nulla che la macchina e il turno relativi a ogni blocco siano indipendenti. □

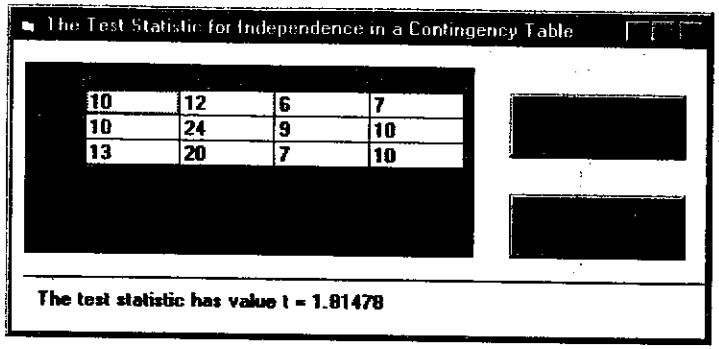


Figura 11.3

11.5 Tabelle di contingenza con i marginali fissati

Consideriamo nuovamente l'Esempio 11.4.1, nel quale eravano interessati a determinare se le convinzioni politiche fossero indipendenti dal sesso degli elettori americani. In quella sede disponevamo di dati ideali: un campione di 300 elementi scelto in maniera completamente casuale dalla popolazione totale. Tuttavia in molte situazioni pratiche ci possiamo trovare di fronte a dati raccolti in maniera diversa: ad esempio non sarebbe strano se il numero di uomini e donne da intervistare venisse deciso in anticipo, e poi si selezionassero con qualche criterio due campioni aleatori dalle sottopopolazioni maschile e femminile. Siccome nella tabella di contingenza risultante i totali delle righe sono decisi a priori, e quindi non contengono informazioni, tale tabella è detta avere i *marginali fissati*.

È possibile dimostrare che anche nel caso i dati vengano raccolti come descritto qui sopra, è possibile utilizzare il test di indipendenza e le strategie costruite nella Sezione 11.4 senza modifiche. In particolare la statistica da utilizzare è sempre

$$T := \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - \widehat{e}_{ij})^2}{\widehat{e}_{ij}} \tag{11.5.1}$$

dove:

- N_{ij} è il numero di membri del campione per i quali vale contemporaneamente $X = i$ e $Y = j$;
- $N_i = \sum_{j=1}^{s} N_{ij}$ è il numero di quelli per i quali $X = i$;
- $M_j = \sum_{i=1}^{r} N_{ij}$ è il numero di quelli per i quali $Y = j$;
- n è la numerosità del campione e si è posto

$$\widehat{e}_{ij} := n\widehat{p}_i\widehat{q}_j = \frac{N_i M_j}{n} \tag{11.5.2}$$

Inoltre è ancora vero che quando H_0 è soddisfatta, al crescere di n, T tende ad avere distribuzione chi-quadro con $(r-1)(s-1)$ gradi di libertà. In sostanza il test di indipendenza basato sulle tabelle di contingenza rimane lo stesso sia se i marginali di una delle caratteristiche sono fissati a priori, sia se sono liberi e vengono ottenuti campionando dall'intera popolazione.

Esempio 11.5.1. In un esperimento vennero scelti a caso un gruppo di 20 000 non fumatori e uno di 10 000 fumatori. Queste persone furono seguite per dieci anni; i dati seguenti illustrano quante di esse svilupparono in tale periodo il cancro ai polmoni.

	Fumatori	**Non fumatori**	
Cancro ai polmoni	62	14	76
Niente cancro ai polmoni	9 938	19 986	29 924
	10 000	20 000	30 000

Si verifichi l'ipotesi che il cancro ai polmoni e il fumo siano indipendenti. Si impieghi un livello di significatività dell'1%.

Le stime del numero di persone che ci si aspetterebbe di trovare nelle diverse celle se valesse l'indipendenza ipotizzata da H_0 sono:

$$\widehat{e}_{11} = \frac{76 \times 10000}{30000} \approx 25.33 \qquad \widehat{e}_{12} = \frac{76 \times 20000}{30000} \approx 50.67$$

$$\widehat{e}_{21} = \frac{29924 \times 10000}{30000} \approx 9974.67 \qquad \widehat{e}_{22} = \frac{29924 \times 20000}{30000} \approx 19949.33$$

Quindi la statistica del test vale

$$\begin{aligned} T &\approx \frac{(62 - 25.33)^2}{25.33} + \frac{(14 - 50.67)^2}{50.67} + \frac{(9\,938 - 9\,974.67)^2}{9\,974.67} \\ &\quad + \frac{(19\,986 - 19\,949.33)^2}{19\,949.33} \\ &\approx 53.09 + 26.54 + 0.13 + 0.07 = 79.83 \end{aligned}$$

Siccome il risultato è molto maggiore di $\chi^2_{0.01,1} \approx 6.635$, possiamo senz'altro rifiutare l'ipotesi che se una persona a caso contrae un tumore ai polmoni, questo sia indipendente dal fatto che fumi o meno. □

Il formalismo che abbiamo sviluppato in questa sezione si può adattare alla verifica dell'uguaglianza di m popolazioni discrete.

Supponiamo infatti che siano date m popolazioni con distribuzione discreta e valori possibili i numeri da 1 a n. Per $i = 1, 2, \ldots, m$ e $j = 1, 2, \ldots, n$, sia p_{ij} la

probabilità che un elemento a caso della popolazione i assuma il valore j. L'ipotesi nulla che tutte le popolazioni siano uguali si formalizza nel seguente sistema di equazioni:

$$H_0 : p_{1j} = p_{2j} = p_{3j} = \cdots = p_{mj}, \qquad j = 1, 2, \ldots, n \tag{11.5.3}$$

Se si prende in considerazione la popolazione complessiva che consiste degli elementi di ciascuna delle m popolazioni in esame, si può pensare che ognuno dei suoi membri abbia due caratteristiche: la prima che indica da quale delle m sottopopolazioni proviene, e la seconda che ne specifica il valore. L'ipotesi che le m distribuzioni siano tutte uguali è equivalente a quella che le percentuali di elementi di ciascuna popolazione che assumono i diversi valori siano le stesse. Siccome questa riformulazione equivale all'indipendenza delle due caratteristiche di un membro che sia scelto a caso dalla popolazione totale, possiamo verificare H_0 scegliendo campioni aleatori delle diverse sottopopolazioni ed eseguendo un test di indipendenza.

Estraiamo campioni aleatori di ampiezze $M_1, M_2, \ldots, M_m$ dalle m popolazioni in esame, e denotiamo con N_{ij} il numero di elementi del campione i che hanno valore j. Questa operazione corrisponde – letta sulla popolazione complessiva – a costruire una tabella di contingenza a marginali fissati, come quella riportata in Tabella 11.2.

La verifica di H_0 si otterrà quindi con un test di indipendenza di tale tabella.

Esempio 11.5.2. In uno studio compiuto di recente, da ciascuno di quattro paesi si è scelto un campione aleatorio di 500 impiegate, che hanno risposto ad un questionario. Una delle domande era se queste donne subissero spesso abusi verbali o sessuali sul lavoro; i dati seguenti rappresentano le risposte ottenute.

Tabella 11.2 Schema di tabella di contingenza per il confronto delle distribuzioni di m popolazioni discrete

		Popolazione						
		1	**2**	$\cdots$	i	$\cdots$	m	
Valore	**1**	N_{11}	N_{21}	$\cdots$	N_{i1}	$\cdots$	N_{m1}	N_1
	2	N_{12}	N_{22}	$\cdots$	N_{i2}	$\cdots$	N_{m2}	N_2
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	j	N_{1j}	N_{2j}	$\cdots$	N_{ij}	$\cdots$	N_{mj}	N_j
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	n	N_{1n}	N_{2n}	$\cdots$	N_{in}	$\cdots$	N_{mn}	N_n
		M_1	M_2	$\cdots$	M_i	$\cdots$	M_m	

Paese	**Casi di abuso frequente (su 500)**
Australia	28
Germania	30
Giappone	51
Stati Uniti	55

Basandosi su questi dati, è possibile che la percentuale di impiegate che si sente spesso oggetto di abusi sul lavoro, sia la stessa per le quattro nazioni?

Codificando con i numeri da 1 a 4 i paesi coinvolti e riportando i dati su di una tabella di contingenza si ottiene:

	1	**2**	**3**	**4**	
Abusi frequenti	28	30	58	55	171
Altre	472	470	442	445	1 829
	500	500	500	500	2 000

L'ipotesi nulla può essere verificata eseguendo un test di indipendenza su questa tabella di contingenza. Eseguendo il Programma 11.4 e poi calcolando il p-dei-dati si ottiene che

$$T \approx 19.52, \qquad p\text{-dei-dati} \approx 0.0002$$

si può quindi affermare che la percentuale di donne con lavoro d'ufficio che si sente spesso oggetto di abusi dipende effettivamente dal paese in esame, infatti l'ipotesi nulla viene rifiutata con l'1% di significatività, come pure con ogni livello di significatività superiore allo 0.02%. □

11.6 * Il test di adattamento di Kolmogorov-Smirnov per i dati continui

Nelle sezioni precedenti abbiamo sempre studiato distribuzioni discrete. Consideriamo invece adesso un campione di dati $Y_1, Y_2, \ldots, Y_n$ proveniente da una distribuzione continua, e ragioniamo su come si possa verificare l'ipotesi nulla che la relativa funzione di ripartizione sia una certa F assegnata.

Un possibile approccio consiste nel dividere i valori possibili delle Y_j (di solito tutto $\mathbb{R}$) in k intervalli disgiunti, ad esempio

$$(y_0, y_1), (y_1, y_2), \ldots, (y_{k-1}, y_k)$$

dove $-\infty = y_0 < y_1 < y_2 < \cdots < y_{k-1} < y_k = \infty$; successivamente si possono considerare al posto di $Y_1, Y_2, \ldots, Y_n$, le variabili discretizzate $Y_1^d, Y_2^d, \ldots, Y_n^d$, definite tramite

$$Y_j^d := i \qquad \text{se } Y_j \text{ appartiene all'intervallo } (y_{i-1}, y_i)$$

La validità dell'ipotesi nulla implicherebbe allora in questo caso che

$$P(Y_j^d = i) = F(y_i) - F(y_{i-1}), \qquad i = 1, 2, \ldots, k$$

e questo può essere verificato facilmente con il test di adattamento per variabili aleatorie discrete presentato nella Sezione 11.2.

Esiste però un altro metodo per verificare se le Y_j provengano da una distribuzione con funzione di ripartizione F, e questo metodo risulta più efficiente della discretizzazione.

Dopo avere osservato il campione $Y_1, Y_2, \ldots, Y_n$, denotiamo con F_e la funzione di distribuzione empirica corrispondente:

$$F_e(x) := \frac{\#\{i : Y_i \le x\}}{n} \tag{11.6.1}$$

Il valore di $F_e(x)$ rappresenta la percentuale di dati del campione minori o uguali a x (si rammenti che con la notazione $\#A$ si intende la cardinalità o numero di elementi dell'insieme A), e quindi la funzione F_e è la funzione di ripartizione della variabile aleatoria discreta che può assumere con uguale probabilità gli n valori osservati.

Poiché $F_e(x)$ è lo stimatore naturale della probabilità che un'osservazione sia minore o uguale a x, ovvero della funzione di ripartizione *vera* dei dati, ne segue che se H_0 è valida F_e dovrebbe essere piuttosto vicina a F. La quantità su cui si basa il test che intendiamo costruire è infatti

$$D := \max_{-\infty<x<\infty} |F_e(x) - F(x)| \tag{11.6.2}$$

La statistica D è la *statistica del test di Kolmogorov-Smirnov*.

Studiamo per prima cosa come si possa calcolare il valore di questa statistica. Denotiamo con $y_1, y_2, \ldots, y_n$ i valori assunti dal campione aleatorio, e sia $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ la loro permutazione che li mette in ordine crescente[3], cioè

$$y_{(j)} := \text{il } j\text{-esimo più piccolo tra } y_1, y_2, \ldots, y_n \tag{11.6.3}$$

Con questa notazione, la funzione F_e può essere riscritta così:

$$F_e(x) = \begin{cases} 0 & \text{se } x < y_{(1)} \\ \frac{1}{n} & \text{se } y_{(1)} \le x < y_{(2)} \\ \vdots & \\ \frac{j}{n} & \text{se } y_{(j)} \le x < y_{(j+1)} \\ \vdots & \\ 1 & \text{se } y_{(n)} \le x \end{cases} \tag{11.6.4}$$

[3] Ad esempio, con $n = 3$ e $y_1 = 3$, $y_2 = 5$, $y_3 = 1$, si avrebbe $y_{(1)} = 1$, $y_{(2)} = 3$ e $y_{(3)} = 5$.

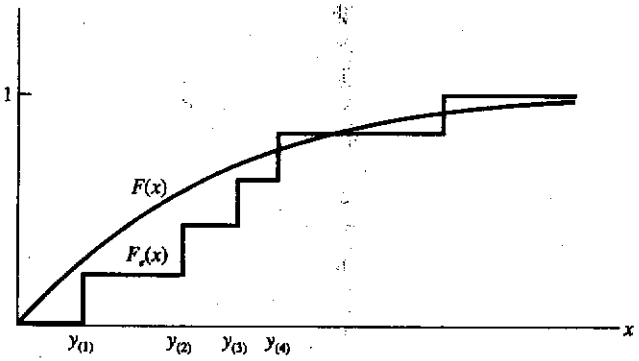


Figura 11.4 Confronto tra la funzione di ripartizione assegnata F e quella empirica F_e, per un campione di 5 dati.

Si tratta quindi di una funzione a gradini: costante in ciascuno degli intervalli $(y_{(j)}, y_{(j+1)})$, compie un salto di ampiezza $1/n$ nei punti $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$. Dovendo studiare il massimo di $|F_e(x) - F(x)|$, analizziamo separatamente $F_e(x) - F(x)$ e $F(x) - F_e(x)$.

Siccome F è una funzione non decrescente e minore o uguale a 1, il massimo al variare di x di $F_e(x) - F(x)$ è non negativo e viene raggiunto in uno dei punti $y_{(j)}$, $j = 1, 2, \ldots, n$ (si veda la Figura 11.4). Quindi

$$\max_{-\infty<x<\infty} \big(F_e(x) - F(x)\big) = \max_{j=1,\ldots,n} \left(\frac{j}{n} - F(y_{(j)})\right)$$

Analogamente il valore massimo di $F(x) - F_e(x)$ è non negativo e viene assunto subito prima di uno dei punti di salto $y_{(j)}$, quando $F(x)$ tende a valere $F(y_{(j)})$ per continuità, mentre $F_e(x)$ vale ancora $(j-1)/n$:

$$\max_{-\infty<x<\infty} \big(F(x) - F_e(x)\big) = \max_{j=1,\ldots,n} \left(F(y_{(j)}) - \frac{j-1}{n}\right)$$

Combinando le due equazioni precedenti si ottiene che

$$D = \max\left\{\frac{j}{n} - F(y_{(j)}),\ F(y_{(j)}) - \frac{j-1}{n},\ j = 1, \ldots, n\right\} \tag{11.6.5}$$

e questa formula può essere usata per calcolare la statistica di Kolmogorov-Smirnov.

Sia d il valore assunto dalla statistica D. È chiaro che un valore troppo elevato di d sarebbe incompatibile con l'ipotesi nulla. Per questo motivo il p-dei-dati va definito come

$$p\text{-dei-dati} := P_F(D \ge d) \tag{11.6.6}$$

dove si è scritto P_F per rendere esplicito il fatto che tale probabilità va calcolata nell'ipotesi che H_0 sia soddisfatta, e quindi F sia la vera distribuzione della popolazione.

Il p-dei-dati precedente può essere approssimato tramite delle simulazioni al calcolatore, e queste ultime sono semplificate dal fatto che la distribuzione di D, e quindi la probabilità $P_F(D \geq d)$, non dipendono in realtà dalla scelta di F. Questo ci permette di stimare il p-dei-dati simulandouna qualsiasi distribuzione continua F: per esempio quella uniforme su $(0, 1)$.

Proposizione 11.6.1. Sia $Y_1, Y_2, \ldots, Y_n$ un campione di variabili aleatorie indipendenti, tutte con funzione di ripartizione continua F, e si definiscano F_e e D come nelle Equazioni (11.6.1) e (11.6.2).

Allora per ogni scelta di d la quantità $P(D \geq d)$ non dipende da F.

Dimostrazione.

$$\begin{aligned} P(D \geq d) &= P\left(\max_x \left|\frac{\#\{i : Y_i \leq x\}}{n} - F(x)\right| \geq d\right) \\ &= P\left(\max_x \left|\frac{\#\{i : F(Y_i) \leq F(x)\}}{n} - F(x)\right| \geq d\right) \\ &= P\left(\max_x \left|\frac{\#\{i : U_i \leq F(x)\}}{n} - F(x)\right| \geq d\right) \end{aligned}$$

La prima uguaglianza è giustificata dal fatto che F è una funzione crescente[4] e quindi $Y \leq x$ è equivalente a $F(Y) \leq F(x)$. Nella seconda uguaglianza si sono indicate con $U_1, U_2, \ldots, U_n$ delle variabili aleatorie indipendenti uniformi su $(0, 1)$. Essa è giustificata dal risultato (la cui dimostrazione è lasciata come esercizio) che se Y ha funzione di ripartizione continua F, allora $F(Y)$ ha distribuzione uniforme su $(0, 1)$.

Continuando le uguaglianze precedenti, e notando che se x varia da $-\infty$ a ∞, allora $F(x)$ varia da 0 a 1, possiamo dire che

$$P(D \geq d) = P\left(\max_{0<y<1} \left|\frac{\#\{i : U_i \leq y\}}{n} - y\right| \geq d\right)$$

che mostra come la distribuzione di D non dipenda da F. □

Dalla proposizione precedente si può dedurre che, una volta ricavato dai dati il valore d della statistica D, il p-dei-dati può essere ottenuto simulando variabili aleatorie

[4] In realtà F è solo non decrescente, però se Y è generata con distribuzione F, i valori per cui F è costante sono impossibili per Y (perché?), quindi con probabilità 1, F è strettamente crescente in Y.

uniformi su $(0, 1)$. In pratica, si genera un campione di n copie indipendenti di queste variabili aleatorie, $U_1, U_2, \ldots, U_n$, e si verifica se è verificata questa disuguaglianza:

$$\max_{0<y<1} \left|\frac{\#\{i : U_i \leq y\}}{n} - y\right| \geq d$$

Si ripete poi un gran numero di volte questo procedimento: la percentuale delle prove in cui la disuguaglianza è soddisfatta è una stima del p-dei-dati.

Come è già stato evidenziato, il primo membro della disuguaglianza può essere più facilmente determinato usando l'identità

$$\max_{0<y<1} \left|\frac{\#\{i : U_i \leq y\}}{n} - y\right| = \max\left\{\frac{j}{n} - U_{(j)},\ U_{(j)} - \frac{j-1}{n},\ j = 1, \ldots, n\right\}$$

dove $U_{(1)}, U_{(2)}, \ldots, U_{(n)}$ non sono altro che le stesse $U_1, U_2, \ldots, U_n$, riordinate dalla più piccola alla più grande. Ad esempio se $n = 3$ e $U_1 = 0.7$, $U_2 = 0.6$, $U_3 = 0.4$, allora $U_{(1)} = 0.4$, $U_{(2)} = 0.6$ e $U_{(3)} = 0.7$ e il corrispondente valore di D è

$$D = \max\left\{\frac{1}{3} - 0.4,\ \frac{2}{3} - 0.6,\ 1 - 0.7,\ 0.4,\ 0.6 - \frac{1}{3},\ 0.7 - \frac{2}{3}\right\} = 0.4$$

Per ottenere un test con significatività α che (in prima approssimazione) non dipenda da n, si definisce di solito la quantità D^*:

$$D^* := (\sqrt{n} + 0.12 + 0.11/\sqrt{n})D \tag{11.6.7}$$

I corrispondenti valori critici d^*_α, sono per definizione i numeri che soddisfano, al variare di $\alpha \in (0, 1)$,

$$P_F(D^* \geq d^*_\alpha) = \alpha \tag{11.6.8}$$

Quelle che seguono sono approssimazioni accurate di d^*_α per i valori più frequentemente utilizzati di α:

$$d^*_{0.1} \approx 1.224, \quad d^*_{0.05} \approx 1.358, \quad d^*_{0.025} \approx 1.480, \quad d^*_{0.01} \approx 1.626 \tag{11.6.9}$$

Un test con significatività α deve rifiutare l'ipotesi nulla che la distribuzione di popolazione sia F quando il valore osservato per D^* risulta maggiore di d^*_α.

Esempio 11.6.1. Supponiamo di volere verificare l'ipotesi che una certa popolazione abbia distribuzione esponenziale con media 100, ovvero che $F(x) = 1 - e^{-x/100}$ per tutte le x positive. Che conclusioni si possono trarre se un campione di numerosità 10 (riordinato) mostra i valori seguenti?

66 72 81 94 112 116 124 140 145 155

Per rispondere a questa domanda, usiamo l'Equazione (11.6.5) per calcolare la statistica D del test di Kolmogorv-Smirnov. Dopo qualche calcolo si ottiene che $D \approx 0.48315$, da cui

$$D^* \approx 0.48315(\sqrt{10} + 0.12 + 0.11/\sqrt{10}) \approx 1.60$$

Siccome tale valore è compreso tra $d^*_{0.025} \approx 1.480$ e $d^*_{0.01} \approx 1.626$, ne segue che l'ipotesi nulla che i dati provenissero da una distribuzione esponenziale di media 100 va rifiutata al 2.5% di significatività (ma andrebbe accettata ad esempio all'1% di significatività). □

Problemi

1. Secondo la teoria mendeliana, incrociando due piante di piselli a fiori rosa di una particolare varietà, si dovrebbero ottenere piantine con fiori bianchi, rosa o rossi con probabilità 1/4, 1/2 e 1/4. Per sperimentare questa teoria si è studiato un campione di 564 piselli, ed è risultato che 141 hanno prodotto fiori bianchi, 291 rosa e 132 rossi. Che conclusioni trai al 5% di significatività, usando l'approssimazione con una chi-quadro?

2. Per stabilire se un dado sia regolare o truccato, si eseguono 1 000 lanci, annotando i risultati seguenti:

Punteggio	1	2	3	4	5	6
Frequenza	158	172	164	181	160	165

Verifica l'ipotesi che il dado sia bilanciato (ovvero che le facce siano equiprobabili) al 5% di significatività. Usa l'approssimazione con una chi-quadro.

3. Procurati le date di nascita e di morte di 100 persone famose e, usando l'approccio con sole quattro categorie, individuato alla fine dell'Esempio 11.2.1, verifica l'ipotesi che il giorno della morte non sia influenzato dalla data di compleanno. Usa l'approssimazione con una chi-quadro.

4. Si pensa che il numero delle interruzioni quotidiane di potenza elettrica in una certa città degli Stati Uniti abbia distribuzione di Poisson di media 4.2. Verifica questa ipotesi se raccogliendo dati per 150 giorni si è trovato il risultato seguente:

Interruzioni	0	1	2	3	4	5	6	7	8	9	10	11
Numero di giorni	0	5	22	23	32	22	19	13	6	4	4	0

5. Su 100 valvole termoioniche testate, 41 hanno avuto una vita inferiore alle 30 ore, 31 l'hanno avuta tra le 30 e le 60 ore, 13 tra le 60 e le 90 ore e 15 oltre le 90 ore. Questi dati sono compatibili con l'ipotesi che il tempo di vita di queste valvole abbia distribuzione esponenziale con media di 50 ore?

6. La produzione passata di una macchina indica che le unità da essa fabbricate si rivelano di qualità eccellente, alta, media o bassa con probabilità rispettivamente di 0.4, 0.3, 0.2 e 0.1. Viene messa in prova una nuova macchina, concepita per eseguire lo stesso compito e su 500 pezzi prodotti se ne ottengono 234 di qualità eccellente, 117 di qualità alta, 81 di qualità media e 68 di qualità bassa. È plausibile che le differenze di prestazioni siano dovute solo al caso?

7. Si attiva un esperimento in grado di individuare i neutrini provenienti dallo spazio esterno, e lo si mantiene attivo per diversi giorni annotando il numero totale di segnali per ogni ora siderale. I risultati trovati sono i seguenti:

Frequenza di neutrini provenienti dallo spazio esterno

Ora	Segnali	Ora	Segnali	Ora	Segnali
0	24	8	37	16	37
1	24	9	37	17	28
2	36	10	49	18	43
3	32	11	51	19	30
4	33	12	29	20	40
5	36	13	26	21	22
6	41	14	38	22	30
7	24	15	26	23	42

Verifica se i segnali siano distribuiti uniformemente nell'arco delle 24 ore.

8. In un altro esperimento di rilevazione dei neutrini, si è annotato per parecchi giorni il numero totale di segnali ricevuti in ciascuna ora. La tabella delle frequenze seguente riassume i risultati:

Numero di segnali in un'ora	Ore con quel numero di segnali
0	1 924
1	541
2	103
3	17
4	1
5	1
6 o più	0

Verifica l'ipotesi che le osservazioni provengano da una distribuzione di Poisson di media 0.3.

9. In una certa zona, i dati in possesso delle assicurazioni dicono che in un anno, l'82% degli automobilisti non ha alcun incidente, il 15% ha esattamente un incidente, e il 3% ne ha 2 o più. Su un campione aleatorio di 440 automobilisti laureati in ingegneria nell'ultimo anno 366 non hanno avuto incidenti, 68 ne hanno avuto uno, 6 ne hanno avuti 2 o più. Puoi concludere che questa sottopopolazione presenta un profilo di rischi diverso da quello generale della zona?

10. Tempo fa è stato condotto uno studio per capire se i terremoti di intensità almeno moderata (4.4 gradi della scala Richter o più) che hanno coinvolto il sud della California tendono a verificarsi in giorni particolari della settimana. I cataloghi hanno permesso di ricavare informazioni su 1 100 terremoti:

Giorno della settimana	Lun	Mar	Mer	Gio	Ven	Sab	Dom
Numero di terremoti	144	170	158	172	148	152	156

Verifica al 5% di significatività l'ipotesi che un terremoto di media intensità abbia le stesse probabilità di verificarsi in un qualsiasi giorno della settimana.

11. Alcune volte i dati raccolti sono in così buon accordo con il modello proposto, da generare il sospetto che non siano stati ottenuti in maniera corretta. Ad esempio un mio amico sostiene di avere sperimentato una moneta lanciandola 40 000 volte e ottenendo 20 004 teste e 19 996 croci; ti sembra che questo risultato sia credibile? Giustifica la tua risposta.

12. Usa delle simulazioni al calcolatore per determinare il p-dei-dati del Problema 1 e confrontalo con quello ottenuto approssimando la statistica con una chi-quadro. Usa simulazioni con numero di iterazioni pari a (a) 1 000; (b) 5 000; (c) 10 000.

13. Un campione di ampiezza 120 ha media campionaria 100 e varianza campionaria 15. Dei 120 dati, 3 sono minori di 70, 18 sono compresi tra 70 e 85, 30 tra 85 e 100, 35 tra 100 e 115, 32 tra 115 e 130, e 2 sono maggiori di 130. Verifica l'ipotesi che la distribuzione da cui è stato estratto il campione fosse normale.

14. Nel Problema 4, verifica l'ipotesi che il numero di interruzioni al giorno abbia distribuzione di Poisson.

15. Un campione aleatorio di 500 nuclei familiari degli Stati Uniti è stato classificato per regione e reddito (in migliaia di dollari), ottenendo i risultati seguenti.

Reddito	**Sud**	**Nord**
0–10	42	53
10–20	55	90
20–30	47	88
30 o più	36	89

Determina il p-dei-dati del test di indipendenza tra reddito e regione di una famiglia scelta a caso.

16. I dati seguenti legano il peso alla nascita di un campione di neonati, con l'età della loro madre.

	Neonati fino a 2.5 Kg	**Neonati oltre i 2.5 Kg**
Madre fino a 20 anni	10	40
Madre oltre i 20 anni	15	135

Verifica l'ipotesi che il peso del bambino sia indipendente dall'età della madre.

17. Risolvi nuovamente il Problema 16 con tutti i dati raddoppiati, ovvero con questi valori:

20	80
30	270

18. La tabella che segue riporta la mortalità infantile in funzione del peso del neonato alla nascita, per 72 730 nati vivi a New York nel 1974.

	Vivi dopo un anno	**Deceduti entro un anno**
Neonati fino a 2.5 Kg	4 597	618
Neonati oltre i 2.5 Kg	67 093	422

Verifica l'ipotesi che il peso alla nascita sia indipendente dall'evento che il neonato viva per più di un anno.

19. Un esperimento congegnato per studiare la relazione tra ipertensione e fumo ha fornito i dati seguenti:

	Non fumatori	**Fumatori moderati**	**Grandi fumatori**
Soggetti a ipertensione	20	38	28
Non soggetti a ipertensione	50	27	18

Verifica l'ipotesi che l'essere affetto o meno da ipertensione sia indipendente da quanto una persona fuma.

20. La tabella seguente riporta il numero di pezzi difettosi, accettabili e qualitativamente superiori prodotti in in impianto, prima e dopo l'introduzione di una modifica del processo di fabbricazione.

	Difettosi	**Accettabili**	**Superiori**
Prima della modifica	25	218	22
Dopo la modifica	9	103	14

Si notano cambiamenti apprezzabili, al 5% di significatività?

21. Un campione di 300 automobili dotate di telefono cellulare e un campione di 400 automobili che ne erano prive, sono stati monitorati per un anno. La tabella seguente riporta quante di queste auto sono state coinvolte in incidenti stradali in quell'arco di tempo.

	Coinvolte in incidenti	**Nessun incidente**
Con telefono cellulare	22	278
Senza telefono cellulare	26	374

Utilizza i dati forniti per verificare l'ipotesi che avere il cellulare in auto non abbia influenza sulla possibilità di essere coinvolti in incidenti. Usa il 5% di significatività.

22. Per studiare l'effetto delle acque arricchite di fluoro sui problemi dentali, si sono scelte due zone dalle caratteristiche socioeconomiche molto simili, una delle quali ha l'acqua potabile arricchita di fluoro, mentre l'altra no. Sono stati selezionati dei campioni casuali di 200 adolescenti da entrambe le popolazioni, e se ne è determinato il numero di carie, ottenendo i dati seguenti.

Numero di carie	Acqua arricchita di fluoro	Acqua normale
0	154	133
1	20	18
2	14	21
3 o più	12	28

(a) Puoi affermare che questi dati, al 5% di significatività stabiliscono che il numero di carie non sia indipentente dalla presenza di fluoro nell'acqua potabile? **(b)** Cosa si conclude all'1% di significatività?

23. Con lo scopo di determinare se le cause per negligenza intentate contro i medici siano più frequenti per certi tipi di interventi che per altri, si sono studiati dei campioni casuali di tre tipi di interventi, ottenendo i dati seguenti.

Tipo di intervento	Casi campionati	Cause intentate
Chirurgia cardiaca	400	16
Chirurgia celebrale	300	19
Appendicectomia	300	7

Verifica l'ipotesi che la percentuale di operazioni che porta ad una causa giudiziaria sia la stessa per i tre tipi di interventi. Usa **(a)** il 5% di significatività; **(b)** l'1% di significatività.

24. In un famoso articolo[5] pubblicato in Inghilterra nel 1926, sono stati riportati i dati seguenti sul colore del cielo la sera e la presenza eventuale di pioggia il giorno successivo.

Colore del cielo	Numero di osservazioni	Osservazioni seguite da pioggia
Rosso	61	26
Principalmente rosso	194	52
Giallo	159	81
Principalmente giallo	188	86
Rosso e giallo	194	52
Grigio	302	167

Verifica se il colore del cielo la sera abbia influenza sul fatto che il giorno seguente vi sia pioggia o meno.

[5] S. Russell, "A red sky at night...", *Metropolitan Magazine London*, vol. 61, p. 15, 1926.

***25.** Dei dati si dicono *lognormali* di parametri μ e σ se i loro logaritmi naturali hann distribuzione $\mathcal{N}(\mu, \sigma^2)$. I valori seguenti rappresentano i giorni di vita di un campion di topi affetti da cancro e curati con una terapia sperimentale:

24 12 36 40 16 10 12 30 38 14 22 18

Utilizza un test di Kolmogorov-Smirnov con il 5% di significatività, per stabilire s queste osservazioni possano provenire da una popolazione lognormale di parametri μ = 3 e $\sigma = 4$.

12 Test statistici non parametrici

Contenuto

12.1 Introduzione

In questo capitolo presentiamo alcune tecniche per verificare ipotesi su distribuzioni la cui forma o classe di appartenenza non sia nota. Per questo, diversamente da solito, non assumiamo che la popolazione studiata sia normale, o esponenziale, o qualunque altro tipo di classe parametrica, e i test che introdurremo sono di conseguenza detti *non parametrici*.

Il vantaggio delle strategie non parametriche è che possono essere applicate senza particolari conoscenze sulla distribuzione in esame; tuttavia, quando vi siano buone ragioni per supporre qualche distribuzione particolare, vanno sempre preferiti i relativi metodi parametrici, che si rivelano più potenti.

Nella Sezione 12.2 prendiamo in esame una classe di ipotesi sulla mediana di una distribuzione continua, e presentiamo il *test dei segni*, che può essere impiegato per la loro verifica. Nella Sezione 12.3 costruiamo il *test dei segni per ranghi*, che permette di verificare l'ipotesi che una distribuzione continua sia simmetrica rispetto ad un valore assegnato. Nella Sezione 12.4 studiamo il confronto di due campioni, e il problema di stabilire se sia plausibile che essi provengano dalla stessa distribuzione; il *test della somma dei ranghi* permette di fornire una risposta. Nella Sezione 12.5, infine, presentiamo il *test delle successioni*, che è usato per stabilire se i dati di un campione siano realmente indipendenti, oppure vi sia evidenza che il loro oscillare segue un qualche schema.

12.2 Il test dei segni

Sia $X_1, X_2, \ldots, X_n$ un campione estratto da una popolazione continua con funzione di ripartizione F, e supponiamo di essere interessati a fare dell'inferenza sulla mediana[1] m; in particolare vogliamo discernere tra le due ipotesi

$$H_0 : m = m_0 \qquad \text{contro} \qquad H_1 : m \neq m_0 \tag{12.2.1}$$

dove m_0 è un valore assegnato qualsiasi.

La strategia che adottiamo è basata sul fatto che ognuna delle osservazioni è minore di m_0, indipendentemente da tutte le altre, con probabilità $F(m_0)$. Quindi se poniamo

$$I_i := \begin{cases} 1 & \text{se } X_i < m_0 \\ 0 & \text{se } X_i \geq m_0 \end{cases}$$

si ha che le variabili aleatorie $I_1, I_2, \ldots, I_n$ risultano indipendenti e bernoulliane di parametro $F(m_0)$, perciò l'ipotesi nulla è equivalente ad affermare che la media di questo nuovo campione sia $\frac{1}{2}$. Le metodologie per verificare questa ipotesi sono già state sviluppate nella Sezione 8.6: sia W una variabile aleatoria binomiale di parametri n e $\frac{1}{2}$. Detto v il numero complessivo di osservazioni inferiori ad m_0, ovvero $\sum_{i=1}^{n} I_i$, segue dall'Equazione (8.6.4) di pagina 322 che il p-dei-dati del test dell'ipotesi che ci interessa è dato da

$$p\text{-dei-dati} = 2\min\{P(W \leq v),\ P(W \geq v)\} \tag{12.2.2}$$

Siccome il parametro p di W è pari a $\frac{1}{2}$, è facile vedere che, per ogni k compreso tra 0 e n, $P(W = k) = P(W = n - k)$, e di conseguenza

$$P(W \geq v) = P(W \leq n - v)$$

per cui il p-dei-dati può anche essere calcolato con la formula

$$\begin{aligned} p\text{-dei-dati} &= 2\min\{P(W \leq v),\ P(W \leq n - v)\} \\ &= \begin{cases} 2P(W \leq v) & \text{se } v \leq n/2 \\ 2P(W \leq n - v) & \text{se } v > n/2 \end{cases} \end{aligned} \tag{12.2.3}$$

Siccome il valore di $v := \sum_{i=1}^{n} I_i$ dipende da quante delle osservazioni X_i sono minori di m_0, ovvero da quanti dei termini $X_i - m_0$ sono negativi, il test precedente prende il nome di *test dei segni*.

[1] La mediana, definita nel Problema 35 a pagina 139, è quel particolare valore m per cui $F(m) = \frac{1}{2}$.

Esempio 12.2.1. Se su di un campione di 200 dati ve ne sono 120 minori di m_0 e 80 maggiori, quanto vale il p-dei-dati del test che m_0 sia la mediana della popolazione?

Usando l'Equazione (12.2.3), con $n = 200$ e $v = 120$, si ha che

$$p\text{-dei-dati} = 2P(W \leq 80) \approx 0.00568$$

dove si è fatto uso del Programma 5.1; l'ipotesi nulla va rifiutata persino con un livello di significatività dell'1%. □

Il test dei segni può essere applicato alle stesse situazioni in cui si usa il test t per i dati appaiati, sviluppato nella Sezione 8.4.4. Riconsideriamo infatti l'Esempio 8.4.4 a pagina 315, in cui si analizzava l'effetto della recente introduzione di un programma di sicurezza industriale, in termini di ore-uomo perse per gli incidenti. Indichiamo con X_i e Y_i i valori relativi alla fabbrica i prima e dopo la modifica. Se fosse vera l'ipotesi H_0 che il programma non ha avuto effetti, X_i e Y_i avrebbero la stessa distribuzione, e quindi la loro differenza $Z_i = Y_i - X_i$, dovrebbe avere mediana nulla (perché?). I valori riscontrati per $Z_1, Z_2, \ldots, Z_{10}$ erano stati:

$$-7.5 \quad 2.5 \quad -2.5 \quad -3.5 \quad -1.5 \quad 0.5 \quad 1.0 \quad -4.5 \quad -4.5 \quad -1.5$$

Siccome questi dati contengono tre valori positivi e sette negativi, l'ipotesi che provengano da una popolazione di mediana nulla va rifiutata con significatività α se

$$\sum_{i=0}^{3} \binom{10}{i} \left(\frac{1}{2}\right)^{10} \leq \frac{\alpha}{2}$$

Visto che la sommatoria al primo membro vale $176/1024 \approx 0.172$, l'ipotesi nulla non può essere rifiutata al 5% di significatività (e in effetti verrebbe accettata a qualunque livello di significatività α minore di 34.4%).

In conclusione il test dei segni non ci permette di affermare che il programma di sicurezza abbia avuto effetti statisticamente rilevanti, e questo risultato è in contraddizione con quanto ottenuto nell'Esempio 8.4.4; in quella sede avevamo però assunto che le differenze avessero distribuzione normale, e questa ipotesi di lavoro ci consentiva di prendere in considerazione non solo il segno delle differenze, ma anche le loro ampiezze. (Il test che introdurremo nella prossima sezione, pur restando di tipo non parametrico, migliorerà le prestazioni del test dei segni, tenendo conto anche di queste ampiezze, facendo pesare maggiormente i segni delle differenze con elevato valore assoluto.)

Il test dei segni può essere applicato ad ipotesi unilaterali con poche modifiche. Supponiamo di volere decidere tra le ipotesi

$$H_0 : m \leq m_0 \qquad \text{e l'alternativa} \qquad H_1 : m > m_0 \tag{12.2.4}$$

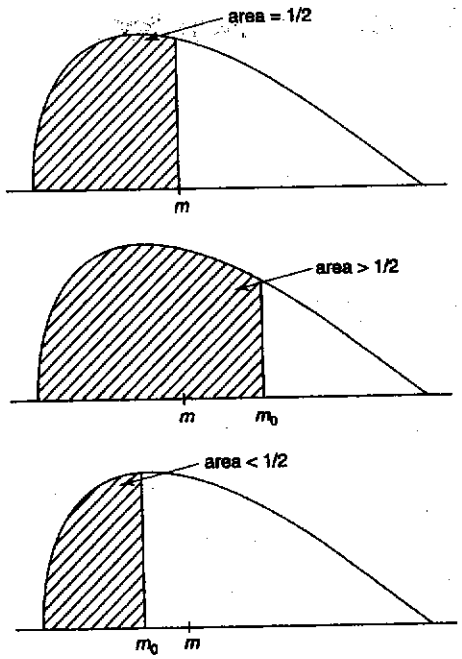


Figura 12.1 Una funzione di densità, la sua mediana e la probabilità di ottenere un valore minore di m_0 quando $m_0 > m$ e quando $m_0 < m$.

dove m è la mediana della popolazione e m_0 un valore assegnato qualsiasi. Sia p la probabilità che un dato sia minore di m_0; se H_0 è vera, $p \geq \frac{1}{2}$, mentre se è falsa, $p < \frac{1}{2}$ (la Figura 12.1 dovrebbe chiarire questo fatto).

Per verificare H_0 con un test dei segni, si estrae dalla popolazione un campione casuale di n elementi: se v di essi hanno un valore inferiore a m_0, il p-dei-dati corrispondente è pari alla probabilità di ottenere un valore come v o più piccolo per puro caso, nonostante ogni elemento avesse probabilità $\frac{1}{2}$ di essere minore di m_0. Perciò

$$p\text{-dei-dati} = P(W \leq v) \tag{12.2.5}$$

dove W è binomiale di parametri n e $\frac{1}{2}$.

Esempio 12.2.2. Un istituto finanziario sta considerando l'apertura di una filiale in una nuova zona. La decisione è condizionata al fatto che la mediana dei redditi delle famiglie della zona sia di almeno 90 000 dollari. Si intervistano 80 famiglie, e si trova che 28 di esse hanno un reddito inferiore a questa cifra, mentre 52 ce l'hanno superiore. Si può affermare con questi dati e al 5% di significatività, che la mediana dei redditi annuali dei nuclei familiari della zona sia superiore ai 90 000 dollari?

Vediamo se i dati sono tali da rifiutare l'ipotesi nulla che la mediana in questione m sia inferiore a 90 000 dollari. Ciò è equivalente a verificare l'ipotesi unilaterale $H_0 : p \geq \frac{1}{2}$, dove p è la probabilità che una famiglia scelta a caso abbia reddito inferiore a quello richiesto. Il p-dei-dati è quindi dato da

$$p\text{-dei-dati} = P(W \leq 28) \approx 0.0048$$

dove W è binomiale di parametri $\frac{1}{2}$ e 80. Si conclude che l'ipotesi che la mediana non superi i 90 000 dollari va rifiutata, e quindi non vi sono controindicazioni all'apertura della nuova filiale. □

Il test dell'ipotesi unilaterale che la mediana sia maggiore o uguale ad un certo valore m_0, si ottiene in maniera analoga al suo simmetrico: se su un campione di numerosità n, i dati che sono risultati minori di m_0 sono v, allora

$$p\text{-dei-dati} = P(W \geq v) \tag{12.2.6}$$

dove W è ha distribuzione binomiale di parametri $\frac{1}{2}$ e n.

12.3 Il test dei segni per ranghi

Il test dei segni permette di verificare l'ipotesi che la mediana di una distribuzione continua sia un valore m_0 assegnato; tuttavia in molte applicazioni pratiche si richiede di sapere se la distribuzione in esame sia non solo centrata, ma anche *simmetrica* rispetto a m_0 (si veda la Figura 12.2). In formule ciò significherebbe verificare l'ipotesi che

$$H_0 : P(X < m_0 - a) = P(X > m_0 + a), \qquad \text{per tutti gli } a > 0 \tag{12.3.1}$$

dove X è un valore estratto dalla popolazione sotto studio.

Anche se tecnicamente si può pure usare il test dei segni per verificare questa ipotesi, esso presenta il difetto di contare solo quanti dati cadono a sinistra e quanti a destra di m_0, ignorando ad esempio la distanza che li separa da tale valore. Un test non parametrico che tenga conto di queste informazioni ulteriori è quello che prende il nome di *test dei segni per ranghi*, o anche di *test del rango segnato* (in inglese è il *signed rank test*), e costituisce l'argomento di questa sezione.

Sia $X_1, X_2, \ldots, X_n$ il campione di dati raccolto, e denotiamo con $Y_i := X_i - m_0$ per $i = 1, 2, \ldots, n$, gli scarti da m_0. Dopo avere ordinato dal più piccolo al più grande i valori assoluti $|Y_1|, |Y_2|, \ldots, |Y_n|$, definiamo le funzioni indicatrici I_j come

segue:

$$I_j := \begin{cases} 1 & \text{se il } j\text{-esimo dei dati nel nuovo ordine è minore di } m_0 \\ 0 & \text{altrimenti} \end{cases}$$

La somma $\sum_{j=1}^{n} I_j$ è di nuovo la statistica del test dei segni (l'unica differenza è l'ordine degli addendi); il test dei segni per ranghi usa invece una nuova statistica, che pesa di più i segni dei dati più lontani da m_0:

$$T := \sum_{j=1}^{n} j I_j \tag{12.3.2}$$

Quando i valori assoluti degli scarti $|Y_1|, |Y_2|, \ldots, |Y_n|$, vengono ordinati dal minore al maggiore, la posizione occupata da $|Y_i|$ è detta *rango* dell'osservazione X_i. Quindi il rango di X_i vale 1 se $|Y_i|$ è il più piccolo, vale 2 se è il secondo più piccolo e così via. Con questa notazione è facile vedere che la statistica del test è la somma dei ranghi dei dati minori di m_0:

$$T = \sum_{i:X_i<m_0} (\text{rango di } X_i)$$

Il nome del test deriva ovviamente da questa formulazione.

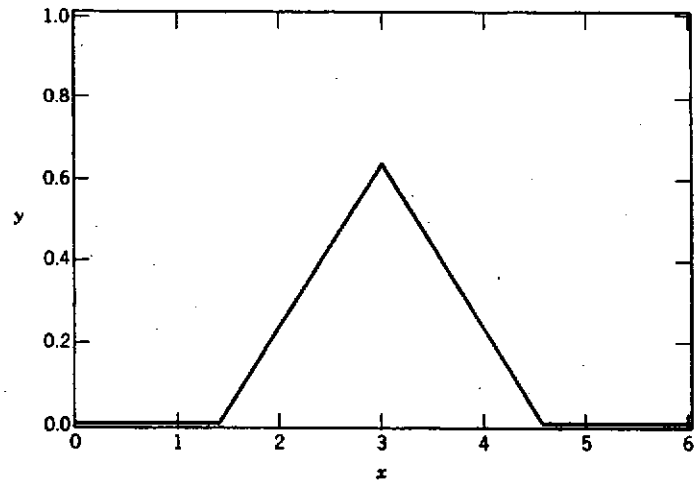


Figura 12.2 Una funzione di densità simmetrica. La mediana è $m = 3$. La formula analitica usata è $f(x) := \max\{0,\ \sqrt{0.4} - 0.4|x-3|\}$

Esempio 12.3.1. Facciamo una prova con i dati

$$4.2 \quad 1.8 \quad 5.3 \quad 1.7$$

e con $m_0 = 2$. Le Y_i corrispondenti sono 2.2, −0.2, 3.3, −0.3, ovvero prese in ordine di valore assoluto:

$$-0.2 \quad -0.3 \quad 2.2 \quad 3.3$$

I segni ci dicono che $I_1 = I_2 = 1$ e $I_3 = I_4 = 0$. Quindi $T = 1 + 2 + 0 + 0 = 3$. □

Supponiamo ora che l'ipotesi H_0 sia soddisfatta, e calcoliamo media e varianza di T. Occorre notare che, siccome le X_i hanno distribuzione simmetrica rispetto a m_0, le Y_i hanno distribuzione simmetrica rispetto a 0. Per questo motivo, qualunque sia il valore y assunto da $|Y_j|$, vi è la stessa probabilità che $Y_j = y$ e $Y_j = -y$, ovvero, il modulo e il segno delle Y_j sono indipendenti. Per questo motivo le variabili aleatorie $I_1, I_2, \ldots, I_n$ sono delle bernoulliane di parametro $\frac{1}{2}$, tra loro indipendenti,

$$P(I_j = 1) = \frac{1}{2} = P(I_j = 0), \qquad j = 1, 2, \ldots, n$$

Con queste premesse, il calcolo di media e varianza di T è un esercizio analogo ad altri simili svolti nei capitoli precedenti.

$$\begin{aligned} E[T] &= E\left[\sum_{j=1}^{n} j I_j\right] \\ &= \sum_{j=1}^{n} \frac{j}{2} = \frac{n(n+1)}{4} && \text{perché } E[I_j] = \frac{1}{2} \qquad (12.3.3) \\ \operatorname{Var}(T) &= \operatorname{Var}\left(\sum_{j=1}^{n} j I_j\right) \\ &= \sum_{j=1}^{n} j^2 \operatorname{Var}(I_j) && \text{per l'indipendenza} \\ &= \sum_{j=1}^{n} \frac{j^2}{4} = \frac{n(n+1)(2n+1)}{24} && \text{perché } \operatorname{Var}(I_j) = \frac{1}{4} \qquad (12.3.4) \end{aligned}$$

dove si è usato il fatto che la varianza di una bernoulliana di parametro p è data da $p(1-p)$, e si è applicata la formula per la somma dei primi n quadrati perfetti[2].

[2] Non essendo un risultato completamente elementare, conviene ricordarlo qui brevemente.

$$1 + 4 + 9 + 16 + \cdots + n^2 = \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

Come utile esercizio, si provi a dimostrare questa formula per induzione.

È possibile dimostrare che quando n è grande (di solito si chiede $n > 25$) la distribuzione di T è approssimativamente normale con media e varianza date dalle due espressioni precedenti. Sebbene questa sia la strada che è stata storicamente usata per compiere questo test, la recente disponibilità di potenza di calcolo a basso costo ci permette di usare un diverso approccio, e ottenere il p-dei-dati esatto tramite calcoli espliciti.

Supponiamo di volere un test con significatività α dell'ipotesi H_0 che la distribuzione sia simmetrica rispetto a m_0. Siccome l'ipotesi nulla sembra poco verosimile sia se vi sono pochi valori (molto) minori di m_0, sia se ve ne sono troppi, essa va rifiutata sia quando la statistica T è molto grande, sia quando è molto piccola: la regione critica deve perciò essere di tipo bilaterale, e H_0 va rifiutata se

$$P_{H_0}(T \le t) < \frac{\alpha}{2} \qquad \text{o} \qquad P_{H_0}(T \ge t) < \frac{\alpha}{2}$$

dove abbiamo indicato con t il valore assunto dalla statistica del test calcolata sui dati. Con la stessa notazione il p-dei-dati è dato da

$$p\text{-dei-dati} = 2\min\{P_{H_0}(T \le t),\ P_{H_0}(T \ge t)\} \tag{12.3.5}$$

I calcoli necessari a determinare il p-dei-dati sono notevolmente ridotti usando la seguente identità, che è dimostrata alla fine di questa sezione:

$$P_{H_0}(T \ge t) = P_{H_0}\left(T \le \frac{n(n+1)}{2} - t\right) \tag{12.3.6}$$

Grazie ad essa, il p-dei-dati si riscrive come

$$\begin{aligned} p\text{-dei-dati} &= 2\min\left\{P_{H_0}(T \le t),\ P_{H_0}\left(T \le \frac{n(n+1)}{2} - t\right)\right\} \\ &= 2P_{H_0}(T \le t^*) \end{aligned} \tag{12.3.7}$$

dove si è posto

$$t^* := \min\left\{t,\ \frac{n(n+1)}{2} - t\right\} \tag{12.3.8}$$

Rimane soltanto da calcolare $P_{H_0}(T \le t^*)$. Sia allora $P_k(i)$ la probabilità, condizionata ad H_0, dell'evento $\{T \le i\}$, quando il campione ha numerosità k. Mostreremo come costruire una formula ricorsiva per $P_k(i)$, partendo da $k = 1$.

Quando $k = 1$ vi è un solo dato, che sotto H_0 può essere minore o maggiore di m_0 con probabilità $\frac{1}{2}$; ne segue che T è bernoulliana di parametro $\frac{1}{2}$, e quindi

$$P_1(i) = \begin{cases} 0 & i = -1, -2, \dots \\ \frac{1}{2} & i = 0 \\ 1 & i = 1, 2, \dots \end{cases} \tag{12.3.9}$$

Supponiamo adesso che la numerosità del campione sia k, e calcoliamo $P_k(i)$ condizionando al valore di I_k come segue:

$$P_k(i) := P_{H_0}(T \le i)$$

... usando la formula di fattorizzazione, Equazione (3.7.1)...

$$= P_{H_0}(T \le i | I_k = 1) P_{H_0}(I_k = 1) + P_{H_0}(T \le i | I_k = 0) P_{H_0}(I_k = 0)$$

... usando la definizione di T, e il fatto che se H_0 è vera, $P(I_k = 1) = P(I_k = 0) = 1/2$...

$$\begin{aligned} &= \frac{1}{2} P_{H_0}\left(\sum_{j=0}^{k} j I_j \le i \,\middle|\, I_k = 1\right) + \frac{1}{2} P_{H_0}\left(\sum_{j=0}^{k} j I_j \le i \,\middle|\, I_k = 0\right) \\ &= \frac{1}{2} P_{H_0}\left(\sum_{j=1}^{k-1} j I_j \le i - k \,\middle|\, I_k = 1\right) + \frac{1}{2} P_{H_0}\left(\sum_{j=1}^{k-1} j I_j \le i \,\middle|\, I_k = 0\right) \end{aligned}$$

... usando l'indipendenza delle $I_1, I_2, \dots, I_k$ ...

$$\begin{aligned} &= \frac{1}{2} P_{H_0}\left(\sum_{j=1}^{k-1} j I_j \le i - k\right) + \frac{1}{2} P_{H_0}\left(\sum_{j=1}^{k-1} j I_j \le i\right) \\ &= \frac{1}{2} P_{k-1}(i-k) + \frac{1}{2} P_{k-1}(i) \end{aligned}$$

Partendo da $P_1(i)$ che abbiamo già calcolato, la formula ricorsiva appena trovata,

$$P_k(i) = \frac{P_{k-1}(i-k) + P_{k-1}(i)}{2} \tag{12.3.10}$$

permette di calcolare successivamente $P_2(\cdot)$, $P_3(\cdot)$, eccetera, fino ad arrivare al valore desiderato di $P_n(t^*)$.

Esempio 12.3.2. Con i dati dell'Esempio 12.3.1 troviamo:

$$t^* := \min\left(3,\ \frac{4 \times 5}{2} - 3\right) = 3$$

quindi il p-dei-dati coincide con $2P_4(3)$, che si può calcolare come segue (si tenga presente che $P_k(i)$ è sempre nullo se $i < 0$):

$$P_2(0) = \frac{P_1(-2)+P_1(0)}{2} = \frac{1}{4} \qquad P_2(1) = \frac{P_1(-1)+P_1(1)}{2} = \frac{1}{2}$$
$$P_2(2) = \frac{P_1(0)+P_1(2)}{2} = \frac{3}{4} \qquad P_2(3) = \frac{P_1(1)+P_1(3)}{2} = 1$$
$$P_3(0) = \frac{P_2(-3)+P_2(0)}{2} = \frac{1}{8} \qquad P_3(1) = \frac{P_2(-2)+P_2(1)}{2} = \frac{1}{4}$$
$$P_3(2) = \frac{P_2(-1)+P_2(2)}{2} = \frac{3}{8} \qquad P_3(3) = \frac{P_2(0)+P_2(3)}{2} = \frac{5}{8}$$
$$P_4(0) = \frac{P_3(-4)+P_3(0)}{2} = \frac{1}{16} \qquad P_4(1) = \frac{P_3(-3)+P_3(1)}{2} = \frac{1}{8}$$
$$P_4(2) = \frac{P_3(-2)+P_3(2)}{2} = \frac{3}{16} \qquad P_4(3) = \frac{P_3(-1)+P_3(3)}{2} = \frac{5}{16} \quad \square$$

Il Programma 12.3 del software abbinato a questo libro usa esattamente questo metodo ricorsivo per calcolare il p-dei-dati del test del rango segnato. I dati che è necessario immettere sono l'ampiezza n del campione e il valore t della statistica del test.

Esempio 12.3.3. Supponiamo di essere interessati a verificare se una certa popolazione ha distribuzione simmetrica rispetto allo zero. Che conclusioni si possono trarre al 10% di significatività, se un campione di 20 dati presenta un valore di 142 per la statistica del test dei segni per ranghi?

Eseguendo il Programma 12.3 otteniamo che il p-dei-dati vale circa 0.177. Perciò l'ipotesi che la distribuzione sia simmetrica rispetto allo zero viene accettata al 10% di significatività. $\square$

Concludiamo questa sezione dando una dimostrazione dell'identità (12.3.6):

$$P_{H_0}(T \geq t) = P_{H_0}\left(T \leq \frac{n(n+1)}{2} - t\right)$$

Ricordiamo che I_j vale 1 se il dato con rango j (il j-esimo dato in ordine crescente di distanza da m_0) è minore di m_0, e vale 0 altrimenti. Di conseguenza, $1 - I_j$ vale 1 se il dato con rango j è *maggiore* di m_0, e vale 0 altrimenti. Perciò se poniamo

$$\begin{aligned} T' &:= \sum_{j=1}^{n} j(1-I_j) \\ &= \sum_{j=1}^{n} j - \sum_{j=1}^{n} jI_j \\ &= \frac{n(n+1)}{2} - T \end{aligned}$$

questa quantità rappresenta la somma dei ranghi delle osservazioni maggiori di m_0. Se H_0 è soddisfatta, per la simmetria della distribuzione rispetto a m_0, T e T' devono avere la stessa distribuzione, e quindi

$$\begin{aligned} P_{H_0}(T \geq t) &= P_{H_0}(T' \geq t) \\ &= P_{H_0}\left(\frac{n(n+1)}{2} - T \geq t\right) \\ &= P_{H_0}\left(T \leq \frac{n(n+1)}{2} - t\right) \end{aligned}$$

Osservazione 12.3.1 (Sugli ex aequo o *ties*). Siccome abbiamo supposto che la distribuzione della popolazione fosse continua, non è in teoria possibile che, mettendo in ordine i valori assoluti delle differenze vi siano due o più valori equidistanti da m_0: tale evento ha infatti probabilità zero. Accade però nella pratica che le osservazioni siano quantizzate, e quindi dei pareggi (in inglese *ties*) siano possibili. Nel caso si verifichi una di queste situazioni, i ranghi vanno ridistribuiti in modo che tutti i dati che si trovano alla stessa distanza da m_0 abbiano lo stesso rango, pari alla media dei ranghi che avrebbero se i loro valori venissero modificati leggermente.

Ad esempio, se $m_0 = 0$, e i dati sono 2, 4, 7, -5 e -7, i valori assoluti riordinati sono 2, 4, 5, 7 e 7. Siccome il valore assoluto 7 compare nelle posizioni 4 e 5, il rango che viene assegnato a -7 e a 7 è di 4.5, e quindi il valore della statistica del test è $T = 3 + 4.5 = 7.5$. Il p-dei-dati va poi calcolato come nel caso in cui tutti i valori fossero stati distinti. (Anche se teoricamente questo metodo non è del tutto corretto, l'errore che si commette è normalmente piccolo.)

12.4 Il confronto di due campioni

Pensiamo ad una situazione in cui sono disponibili due metodi per fabbricare degli oggetti, questi oggetti hanno delle caratteristiche misurabili, e siamo interessati a verificare se i due metodi siano statisticamente equivalenti.

Selezioniamo allora n pezzi prodotti con il primo metodo e m con il secondo. Siano $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_m$ i valori corrispondenti a questi oggetti, e denotiamo con F e G le funzioni di ripartizione (incognite) delle due popolazioni, che supponiamo essere continue. L'ipotesi nulla che vogliamo verificare è $H_0 : F = G$.

Presentiamo di seguito una tecnica per eseguire questa verifica, che prende vari nomi, tra cui test di Mann-Whitney, test di Wilcoxon o test della somma dei ranghi. Per prima cosa occorre ordinare dal minore al maggiore le $n + m$ osservazioni $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$; poiché F e G sono assunte continue, con probabilità 1 non vi sono valori uguali, e quindi l'ordinamento è ben definito; si denota a questo punto con R_i, per $i = 1, 2, \ldots, n$ il rango di X_i, ovvero la posizione del dato

con la somma dei ranghi minore, t potrebbe valere addirittura 40 100, e di conseguenza potrebbero rendersi necessari fino a $200 \times 200 \times 40\,100 = 1.604 \times 10^9$ diversi valori di $P(N, M, K)$ per calcolare il p-dei-dati. Perciò, per campioni di grosse dimensioni, il calcolo esatto, basato sull'Equazione (12.4.3) non è percorribile. Si aprono allora due possibilità, che sono il metodo classico basato sull'approssimazione della distribuzione di T, e la simulazione al calcolatore.

12.4.1 Approssimazione classica

Se l'ipotesi nulla è vera, e quindi $F = G$, gli $n + m$ valori osservati provengono da una sola distribuzione, e quindi tutti i $(n + m)!$ modi di assegnare i ranghi ai dati $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$ sono equiprobabili. Ne segue che la scelta degli n ranghi per il primo campione è equivalente alla estrazione casuale di n valori da un'urna che contenga i numeri $1, 2, \ldots, n+m$. Usando questo fatto si può dimostrare che

$$\begin{aligned} E_{H_0}[T] &= \frac{n(n+m+1)}{2} \\ \mathrm{Var}_{H_0}[T] &= \frac{nm(n+m+1)}{12} \end{aligned} \tag{12.4.7}$$

È inoltre possibile dimostrare che quando n e m sono entrambi non troppo piccoli (di solito si chiede che siano maggiori di 7), la distribuzione di T sotto H_0 è approssimativamente normale. Quindi quando H_0 è soddisfatta, la statistica

$$\frac{T - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} \sim \mathcal{N}(0,1) \tag{12.4.8}$$

ha approssimativamente distribuzione normale standard, perciò se si denota con d il modulo della differenza tra il valore osservato per T e la sua media data dall'Equazione (12.4.7), allora il p-dei-dati di questo test è dato da

$$\begin{aligned} p\text{-dei-dati} &= P_{H_0}(|T - E_{H_0}[T]| > d) \\ &\approx P\left(|Z| > d/\sqrt{\frac{nm(n+m+1)}{12}}\right) \\ &= 2P\left(Z > d/\sqrt{\frac{nm(n+m+1)}{12}}\right) \end{aligned} \tag{12.4.9}$$

dove Z è una variabile aleatoria $\mathcal{N}(0,1)$.

Esempio 12.4.5. Riconsideriamo l'Esempio 12.4.1. Abbiamo $n = 5$ e $m = 6$, e il valore della statistica del test è 21. Visto che

$$\frac{n(n+m+1)}{2} = 30 \quad \text{e} \quad \frac{nm(n+m+1)}{12} = 30$$

si trova $d = 9$ e quindi

$$\begin{aligned} p\text{-dei-dati} &\approx 2P\left(Z > \frac{9}{\sqrt{30}}\right) \\ &\approx 2P(Z > 1.643) \\ &\approx 2(1 - 0.9498) = 0.1004 \end{aligned}$$

che può essere confrontato con il valore esatto trovato nell'Esempio 12.4.1, che è 0.1225. □

Esempio 12.4.6. Nell'Esempio 12.4.4, $n = 9$ e $m = 13$, cosicché

$$\frac{n(n+m+1)}{2} = 103.5 \quad \text{e} \quad \frac{nm(n+m+1)}{12} = 224.25$$

Siccome $T = 72$, risulta che $d = |72 - 103.5| = 31.5$, quindi il p-dei-dati approssimato è dato da

$$\begin{aligned} p\text{-dei-dati} &\approx 2P\left(Z > \frac{31.5}{\sqrt{224.25}}\right) \\ &\approx 2P(Z > 2.104) \\ &\approx 2(1 - 0.9823) = 0.0354 \end{aligned}$$

che è piutosto vicino al valore esatto trovato nell'Esempio 12.4.4, ovvero 0.0364. □

I due esempi appena discussi confermano la regola empirica che con campioni di ampiezze maggiori di 7 si trova una approssimazione che è già piuttosto buona (ad un costo computazionale trascurabile), mentre con campioni più piccoli si può sbagliare anche di parecchio.

12.4.2 Simulazione

Se indichiamo con t il valore osservato per la statistica del test, allora il p-dei-dati è dato da

$$p\text{-dei-dati} = 2\min\{P_{H_0}(T \le t),\ P_{H_0}(T \ge t)\}$$

Questo valore può essere approssimato simulando una serie di volte la somma di n elementi estratti casualmente dall'insieme $\{1, 2, \ldots, n+m\}$. La frazione delle prove nelle quali la somma così ottenuta risulta minore o uguale a t approssima $P_{H_0}(T \le t)$, e analogamente la frazione delle prove in cui la somma è maggiore o uguale a t approssima $P_{H_0}(T \ge t)$.

Nel software abbinato al libro e disponibile online è incluso (nella parte relativa al Capitolo 12) un programma che utilizza questa strategia per simulare il p-dei-dati del test della somma dei ranghi. L'efficienza di questo programma è maggiore se come primo campione viene scelto il meno numeroso.

Esempio 12.4.7. Simulando con il programma suddetto il p-dei-dati degli Esempi 12.4.1 e 12.4.4, si ottengono le schermate delle Figure 12.4 e 12.5, che forniscono valori piuttosto vicini a quelli esatti (che sono 0.1225 e 0.0364). □

L'approccio della simulazione richiede molto più tempo di calcolo dell'approssimazione classica. Esso tuttavia presenta il vantaggio di poter fornire risultati arbitrariamente accurati, semplicemente aumentando il numero delle iterazioni.

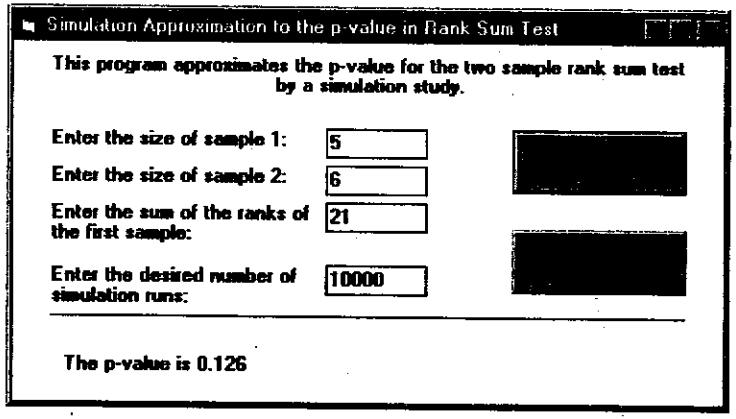


Figura 12.4

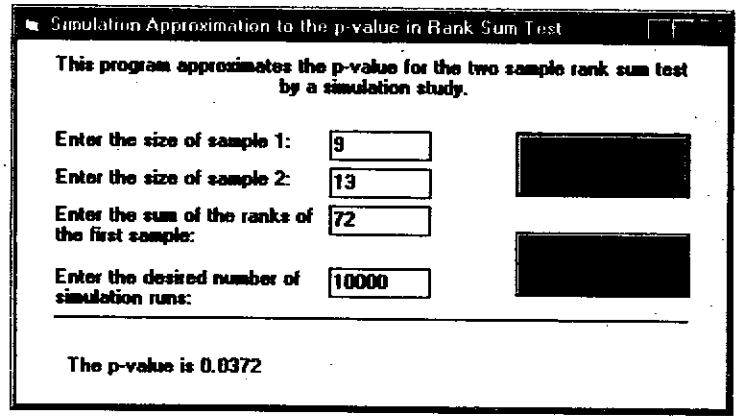


Figura 12.5

12.5 Test delle successioni per la casualità di un campione

Una delle assunzioni che stanno alla base di tutta l'analisi statistica è che il campione di osservazioni sia formato da variabili aleatorie *indipendenti*, provenienti tutte dalla stessa distribuzione. Può però anche succedere che i dati non siano generati in maniera completamente casuale, ma seguendo una tendenza, o delle configurazioni cicliche particolari. In questa sezione presentiamo il test delle successioni (in inglese *runs test*), che permette di verificare l'ipotesi H_0 che il campione sia effettivamente casuale.

Per iniziare, supponiamo che i dati osservati $X_1, X_2, \ldots, X_N$, siano semplicemente delle cifre 0 o 1. (Questo accade, ad esempio, ogni volta che l'esito delle prove viene catalogato in due categorie, come "successo" e "fallimento".) Si chiama *successione*, ogni sequenza di cifre consecutive uguali presente nel campione. Se ad esempio i dati fossero

$$1\ 0\ 0\ 1\ 1\ 1\ 0\ 0\ 1\ 0\ 1\ 1\ 1\ 1\ 0\ 1\ 0\ 0\ 0\ 0\ 1\ 1$$

potremmo contare l'alternarsi di 11 successioni: 6 successioni di uno e 5 successioni di zeri.

Supponiamo che il campione $X_1, X_2, \ldots, X_N$ sia formato da n dati 1 ed m dati 0, con $n + m = N$, e sia R il numero delle sue successioni ("runs"). Se H_0 è soddisfatta, l'alternarsi di 0 e 1 può essere, con uguale probabilità, una qualsiasi delle $\binom{N}{n}$ combinazioni; perciò condizionando all'ipotesi nulla e al numero complessivo di 0 e 1, la funzione massa di probabilità di R è data da

$$P_{H_0}(R = k) = \frac{\begin{array}{c}\text{(numero delle combinazioni di } n \text{ dati 1} \\ \text{e } m \text{ dati 0, che mostrano } k \text{ successioni)}\end{array}}{\binom{n+m}{n}}$$

Tale numero di combinazioni può essere determinato esplicitamente, mostrando che

$$\begin{aligned} P_{H_0}(R = 2k) &= 2\frac{\binom{m-1}{k-1}\binom{n-1}{k-1}}{\binom{n+m}{n}} \\ P_{H_0}(R = 2k+1) &= \frac{\binom{m-1}{k-1}\binom{n-1}{k} + \binom{m-1}{k}\binom{n-1}{k-1}}{\binom{n+m}{n}} \end{aligned} \tag{12.5.1}$$

Il test delle successioni prescrive di rifiutare l'ipotesi nulla quando il valore osservato per R è troppo grande o troppo piccolo per potere essere stato ottenuto casualmente dalla distribuzione definita dall'Equazione (12.5.1). In particolare, se si

X_i nell'ordinamento appena ottenuto,

$$R_i := \text{posizione del dato } X_i \tag{12.4.1}$$

La statistica utilizzata dal test è la somma dei ranghi delle osservazioni X_i,

$$T := \sum_{i=1}^{n} R_i \tag{12.4.2}$$

Esempio 12.4.1. In un esperimento ideato per confrontare due tipi di trattamenti anti-corrosione si sono ottenuti i risultati seguenti:

Trattamento 1	65.2	67.1	69.4	78.2	74	80.3
Trattamento 2	59.4	72.1	68	66.2	58.5	

(I dati rappresentano le profondità massime – in millesimi di pollice – dei microsolchi formatisi su campioni di filo di ferro trattati nei due modi.)

I valori riordinati sono:

58.5 59.4 **65.2** 66.2 **67.1** 68 **69.4** 72.1 **74** **78.2** **80.3**

quelli che sono stati incorniciati provengono dal primo campione; il corrispondente valore di T è $3+5+7+9+10+11=45$. □

Supponiamo di volere verificare $H_0 : F = G$ con livello di significatività α; se il valore assunto dalla statistica del test è t, allora l'ipotesi nulla va rifiutata se

$$P_{H_0}(T \le t) < \frac{\alpha}{2} \quad \text{o} \quad P_{H_0}(T \ge t) < \frac{\alpha}{2}$$

ovvero se il valore riscontrato per t è troppo grande o troppo piccolo perché si possa pensare che sia una deviazione casuale.

Siccome T assume solo valori interi,

$$\begin{aligned} P(T \ge t) &= 1 - P(T < t) \\ &= 1 - P(T \le t-1) \end{aligned}$$

Quindi si può anche dire che H_0 va rifiutata se

$$P_{H_0}(T \le t) < \frac{\alpha}{2} \quad \text{o} \quad P_{H_0}(T \le t-1) > 1 - \frac{\alpha}{2}$$

Abbiamo quindi bisogno della funzione di ripartizione di T sotto l'ipotesi che H_0 sia soddisfatta. Sia allora $P(N, M, K)$ la probabilità, condizionata ad H_0, dell'evento $\{T \le K\}$, quando i campioni hanno numerosità N e M. Otterremo di seguito

una formula ricorsiva computazionalmente valida, che ci consentirà di ricavare le probabilità necessarie al test:

$$P_{H_0}(T \le t) = P(n, m, t) \qquad \text{e} \qquad P_{H_0}(T \le t-1) = P(n, m, t-1)$$

La probabilità $P_{H_0}(T \le t)$ può essere calcolata condizionando ai due eventi complementari che l'osservazione con rango massimo $N+M$ appartenga al primo o al secondo campione (stiamo usando qui la formula di fattorizzazione, Equazione (3.7.1), discussa a pagina 74). Siccome supponiamo vera H_0, tutte le $N+M$ osservazioni possono essere quella di rango massimo con pari probabilità, quindi le probabilità di questi due eventi sono

$$\begin{aligned} P_{H_0}(\text{è una delle } X_i \text{ ad avere rango } N+M) &= \frac{N}{N+M} \\ P_{H_0}(\text{è una delle } Y_j \text{ ad avere rango } N+M) &= \frac{M}{N+M} \end{aligned}$$

Se condizioniamo al primo caso, la somma dei ranghi del primo campione vale $N+M$ più i ranghi degli altri $N-1$ membri. Questa somma è minore o uguale a K se la somma degli $N-1$ ranghi diversi da $N+M$ è minore di $K-(N+M)$, ma siccome i restanti $N-1+M$ valori – cioè tutti tranne il maggiore – provengono tutti dalla stessa distribuzione (stiamo supponendo vera H_0), ne segue che la somma dei ranghi di $N-1$ elementi è minore di $K-N-M$ con probabilità data da $P(N-1, M, K-N-M)$. Con un ragionamento analogo si prova che condizionando al secondo caso la somma dei ranghi del primo campione è minore o uguale a K con probabilità $P(N, M-1, K)$. Mettendo assieme i risultati otteniamo che

$$P(N, M, K) = \frac{N}{N+M} P(N-1, M, K-N-M) + \frac{M}{N+M} P(N, M-1, K) \tag{12.4.3}$$

Ad iniziare dalle condizioni di bordo

$$P(1,0,K) = \begin{cases} 0 & K \le 0 \\ 1 & K > 0 \end{cases}, \qquad P(0,1,K) = \begin{cases} 0 & K < 0 \\ 1 & K \ge 0 \end{cases} \tag{12.4.4}$$

l'Equazione (12.4.3) può essere applicata ricorsivamente fino ad ottenere $P(n, m, t-1)$ e $P(n, m, t)$.

Esempio 12.4.2. Supponendo di volere calcolare $P(2,1,3)$, possiamo applicare l'Equazione (12.4.3) come segue:

$$\begin{aligned} P(2,1,3) &= \frac{2}{3} P(1,1,0) + \frac{1}{3} P(2,0,3) \\ P(1,1,0) &= \frac{1}{2} P(0,1,-2) + \frac{1}{2} P(1,0,0) = 0 \\ P(2,0,3) &= P(1,0,1) + 0 = 1 \end{aligned}$$

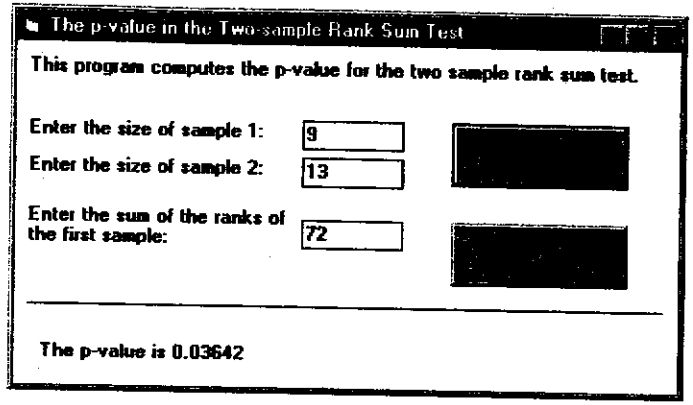


Figura 12.3

Quindi $P(2,1,3) = \frac{1}{3}$, come ci si aspettava; infatti i dati sono X_1, X_2, Y_1, e affinché la somma dei ranghi di X_1 e X_2 non superi 3, occorre che il maggiore sia Y_1, e questo evento, quando H_0 è vera, ha probabilità $\frac{1}{3}$. □

Siccome il test della somma dei ranghi rifiuta l'ipotesi nulla quando

$$2P(n,m,t) < \alpha \qquad \text{o} \qquad \alpha > 2(1 - P(n,m,t-1))$$

ne segue che, se t è il valore della statistica calcolato sui dati,

$$p\text{-dei-dati} = 2\min\{P(n,m,t),\ 1 - P(n,m,t-1)\} \tag{12.4.5}$$

Il Programma 12.4 usa la ricorsione descritta in questa sezione per calcolare il p-dei-dati per il test della somma dei ranghi. I dati che occorre immettere sono le ampiezze dei due campioni e la somma dei ranghi del primo campione. Sebbene si possa scegliere come primo campione uno qualsiasi dei due, il programma termina più rapidamente se si sceglie quello a cui corrisponde la somma dei ranghi minore.

Esempio 12.4.3. Nell'Esempio 12.4.1 le ampiezze dei due campioni sono 5 e 6, e la somma dei ranghi del campione di 5 osservazioni è 21. Eseguendo il Programma 12.4 troviamo quindi:

$$p\text{-dei-dati} \approx 0.1255 \quad \square$$

Esempio 12.4.4. Si sta verificando se i risultati ottenuti con due diversi metodi di produzione siano analoghi. Vengono fabbricati 9 pezzi con un metodo, e 13 con l'altro. Una volta misurata la caratteristica rilevante dei 22 pezzi, risulta che la somma dei ranghi del campione di 9 elementi vale 72. Che conclusioni si possono trarre?

Eseguiamo il Programma 12.4 ottenendo la schermata in Figura 12.3. L'ipotesi che le distribuzioni siano identiche va quindi rifiutata al 5% di significatività. □

Resta il problema di calcolare la statistica T. Un metodo piuttosto efficiente consiste nell'ordinare i dati con uno degli algoritmi standard dell'informatica (come il quicksort), e poi determinare la somma dei ranghi direttamente. Un diverso approccio, facile da implementare anche se risulta efficiente solo per valori piccoli di n e m, sfrutta la seguente identità.

Proposizione 12.4.1. Per $i = 1,2,\ldots,n$ e $j = 1,2,\ldots,m$ sia

$$W_{ij} := \begin{cases} 1 & \text{se } X_i > Y_j \\ 0 & \text{altrimenti} \end{cases}$$

Allora

$$T = \frac{n(n+1)}{2} + \sum_{i=1}^{n}\sum_{j=1}^{m} W_{ij} \tag{12.4.6}$$

Dimostrazione. Notiamo intanto che $\sum_{j=1}^{m} W_{ij}$ è il numero degli indici j per cui $Y_j < X_i$. Secondariamente, il rango di X_i è pari a uno più il numero di dati che gli sono minori, quindi:

$$\begin{aligned} \text{rango di } X_i &= \#\{j : Y_j < X_i\} + \#\{k : X_k < X_i\} + 1 \\ &= \sum_{j=1}^{m} W_{ij} + \#\{k : X_k \le X_i\} \end{aligned}$$

Per cui

$$T := \sum_{i=1}^{n}(\text{rango di } X_i) = \sum_{i=1}^{n}\sum_{j=1}^{m} W_{ij} + \sum_{i=1}^{n} \#\{k : X_k \le X_i\}$$

Per concludere basta dimostrare che

$$\sum_{i=1}^{n} \#\{k : X_k \le X_i\} = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

e questo è vero, perché ognuno degli insiemi $\{k : X_k \le X_i\}$, per $i = 1,2,\ldots,n$, ha un numero di elementi diverso e compreso tra 1 e n, quindi questi valori sono semplicemente una permutazione degli interi $1,2,\ldots,n$, e la loro somma non dipende dall'ordine. □

Il metodo ricorsivo per il calcolo del p-dei-dati che usa l'Equazione (12.4.3) presenta il problema che il tempo di calcolo aumenta molto velocemente con le ampiezze dei campioni. Ad esempio, se $n = m = 200$, siccome la somma di tutti i ranghi è $1+2+\cdots+400 = \frac{400\times401}{2} = 80\,200$, anche scegliendo come primo campione quello

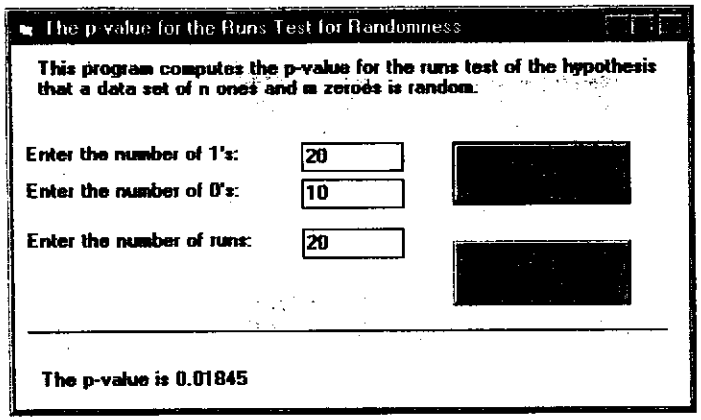


Figura 12.6

osservano r successioni, il p-dei-dati del test è dato da

$$p\text{-dei-dati} = 2\min\{P_{H_0}(R \geq r),\ P_{H_0}(R \leq r)\} \tag{12.5.2}$$

Il Programma 12.5 usa l'Equazione (12.5.1) per calcolare il p-dei-dati.

Esempio 12.5.1. Quelli che seguono sono i risultati delle ultime 30 partite giocate da una squadra di baseball: ogni v indica una vittoria e ogni S una sconfitta.

$$v\ v\ v\ S\ v\ v\ S\ v\ v\ S\ v\ S\ v\ v\ S\ v\ v\ v\ v\ S\ v\ S\ v\ v\ v\ S\ v\ S\ v\ S$$

Si può accettare con questi dati che il campione sia completamente casuale?

Il campione è composto da 20 v e 10 S, e contiene 20 successioni. Eseguiamo il Programma 12.5, ottenendo la schermata in Figura 12.6. L'ipotesi nulla che il campione sia casuale viene rifiutata al 5% e al 2%, ma non all'1% di significatività. (La stranezza di questi dati è che la squadra in questione ha vinto dopo ogni sconfitta, cosa che è piuttosto rara se tutti gli esiti con 20 vittorie e 10 sconfitte sono equiprobabili.) □

La stessa strategia permette di verificare la casualità di un campione anche se i dati non sono composti da sole cifre 0 e 1. Per verificare se i dati $X_1, X_2, \ldots, X_N$ siano veramente casuali, denotiamo con c la mediana campionaria (si veda la Definizione 2.3.2 a pagina 23), con n il numero di dati minori o uguali a c, e con m il numero di quelli maggiori di c. (Si noti che se N è pari e tutti i dati sono distinti $n = m = N/2$.) Definiamo poi, per $j = 1, 2, \ldots, N$ le funzioni indicatrici

$$I_j := \begin{cases} 1 & \text{se } X_j \leq c \\ 0 & \text{altrimenti} \end{cases}$$

Se l'ipotesi nulla è vera, il numero di successioni riscontrabili nella sequenza $I_1, I_2, \ldots, I_N$ ha funzione di massa data dall'Equazione (12.5.1). In particolare è possibile verificare H_0 applicando il test precedente al campione $I_1, I_2, \ldots, I_N$.

Esempio 12.5.2. I tempi di vita di 19 batterie prodotte in successione sono stati i seguenti:

142 152 148 155 176 134 184 132 145 162
165 185 174 198 179 194 201 169 182

La mediana campionaria è il decimo valore dal più piccolo, ovvero 169. Usando 169 come soglia, e associando 1 ai valori inferiori o uguali, e 0 a quelli superiori, si trova,

1 1 1 1 0 1 0 1 1 1 1 0 0 0 0 0 0 1 0

Le successioni sono 8. Per stabilire se questo valore sia statisticamente significativo eseguiamo il Programma 12.5 con $n = 10$ e $m = 9$, ottenendo il risultato

$$p\text{-dei-dati} \approx 0.357$$

L'ipotesi di casualità viene in questo caso accettata. □

È possibile dimostrare che, quando n e m sono grandi e H_0 è valida, R ha distribuzione approssimativamente normale, con media e deviazione standard date da

$$\mu = \frac{2nm}{n+m} + 1 \qquad \text{e} \qquad \sigma = \sqrt{\frac{2nm(2nm-n-m)}{(n+m)^2(n+m-1)}} \tag{12.5.3}$$

Perciò quando n e m sono numeri elevati, denotando con Z una variabile aleatoria con distribuzione $\mathcal{N}(0,1)$,

$$\begin{aligned} P_{H_0}(R \leq r) &= P\left(\frac{R-\mu}{\sigma} \leq \frac{r-\mu}{\sigma}\right) \\ &\approx P\left(Z \leq \frac{r-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{r-\mu}{\sigma}\right) \end{aligned}$$

e analogamente

$$P_{H_0}(R \geq r) \approx 1 - \Phi\left(\frac{r-\mu}{\sigma}\right)$$

Da queste espressioni si ricava immediatamente una approssimazione del p-dei-dati di questo test quando i dati sono grandi, ovvero

$$p\text{-dei-dati} \approx 2\min\left\{\Phi\left(\frac{r-\mu}{\sigma}\right),\ 1 - \Phi\left(\frac{r-\mu}{\sigma}\right)\right\} \tag{12.5.4}$$

dove μ e σ sono dati dall'Equazione (12.5.3) e r è il numero di successioni osservate.

Esempio 12.5.3. Supponiamo che in una sequenza di 60 dati 1 e 60 dati 0 vi siano 75 successioni. Calcoliamo prima media e deviazione standard di R

$$\mu = 61 \qquad \text{e} \qquad \sigma = \sqrt{\frac{3\,540}{119}} \approx 5.454$$

Siccome $(r - \mu)/\sigma \approx 2.567$, il p-dei-dati approssimato è dato da

$$\begin{aligned} p\text{-dei-dati} &\approx 2\min\{\Phi(2.567),\ 1 - \Phi(2.567)\} \\ &\approx 2 \times (1 - 0.9949) = 0.0102 \end{aligned}$$

D'altra parte, usando il Programma 12.5 possiamo trovare il valore esatto che è 0.130.

Cambiando dati, se il numero di successioni fosse stato 70 anziché 75, avremmo trovato un p-dei-dati approssimato di $2(1 - \Phi(1.650)) \approx 0.0990$, mentre il valore esatto è 0.1189. □

Problemi

1. Una nuova medicina contro l'ipertensione viene sperimentata su 18 pazienti. Dopo 40 giorni di trattamento si osservano queste variazioni nella pressione diastolica:

$$\begin{array}{cccccccccc} -5 & -1 & +2 & +8 & -25 & +1 & +5 & -12 & -16 \\ -9 & -8 & -18 & -5 & -22 & +4 & -21 & -15 & -11 \end{array}$$

   (a) Usa il test dei segni per stabilire se la medicina abbia avuto qualche effetto.

   (b) Quanto vale il p-dei-dati?

2. Uno studio di ingegneria vuole stabilire il sistema informatico adatto alle sue esigenze. Quando la scelta si è ristretta a due produttori, l'azienda sottopone loro 8 problemi di calcolo e misura i tempi necessari per risolverli con le architetture e i software da loro proposti. I tempi ottenuti sono:

Problema	1	2	3	4	5	6	7	8
Sistema A	15	32	17	26	42	29	12	38
Sistema B	22	29	1	23	46	25	19	47

   Determina il p-dei-dati del test dei segni sull'ipotesi nulla che non vi siano differenze nelle distribuzioni dei tempi necessari ai due calcolatori per risolvere i problemi.

3. Il valore ufficiale per la mediana della pressione sistolica negli uomini di mezza età è di 128. Volendo controllare se questo dato sia ancora valido, si misura la pressione di un campione di 100 individui di questa popolazione. Verifica l'ipotesi che la mediana sia 128 sapendo che i soggetti che hanno mostrato una pressione superiore a questo valore sono stati: **(a)** 60; **(b)** 70; **(c)** 80. In ciascuno di questi casi, determina il p-dei-dati.

4. Per verificare l'ipotesi che la mediana del peso della popolazione femminile di sedici anni di Los Angeles sia di almeno 110 libbre, si seleziona un campione di 200 di queste giovani, e si osserva che 120 di loro pesano meno di 110 libbre.

   (a) Cosa si conclude al 5% di significatività?

   (b) Quanto vale il p-dei-dati?

5. Nel 1987 la mediana nazionale del reddito dei medici degli Stati Uniti era di 124 400 dollari. Un campione casuale dei redditi del 1990, mostra, in migliaia di dollari, i valori:

$$\begin{array}{ccccccc} 125.5 & 130.3 & 133.0 & 102.6 & 198.0 & 232.5 & 106.8 \\ 114.5 & 122.0 & 100.0 & 118.8 & 108.6 & 312.7 & 125.5 \end{array}$$

   Usa questi dati per verificare l'ipotesi che la mediana dei redditi dei medici nel 1990 non sia cresciuta rispetto al 1987. Quanto vale il p-dei-dati?

6. Viene condotto un esperimento per studiare l'influenza che un nuovo additivo per benzina (un detergente) ha sui consumi. I dati che seguono rappresentano le miglia percorse con un gallone di benzina per 8 diverse automobili, con e senza l'additivo.

Auto	1	2	3	4	5	6	7	8
Senza additivo	24.2	30.4	32.7	19.8	25.0	24.9	22.2	21.5
Con additivo	23.5	29.6	32.3	17.6	25.3	25.4	20.6	20.7

   Determina il p-dei-dati dell'ipotesi che i consumi non siano influenzati dall'additivo usando **(a)** il test dei segni; **(b)** il test dei segni per ranghi.

7. Ricalcola il p-dei-dati dei Problemi 1 e 2 usando il test dei segni per ranghi.

8. In una clinica si somministra un farmaco a 12 pazienti che hanno un elevato livello di albumina. La concentrazione della sostanza prima e dopo il trattamento (in grammi per 100 millilitri) è quella seguente.

Paziente	1	2	3	4	5	6	7	8	9	10	11	12
Prima	5.02	5.08	4.75	5.25	4.80	5.77	4.85	5.09	6.05	4.77	4.85	5.24
Dopo	4.66	5.15	4.30	5.07	5.38	5.10	4.80	4.91	5.22	4.50	4.85	4.56

   Si può concludere che l'effetto del farmaco sia apprezzabile al 5% di significatività? Usa **(a)** il test dei segni; **(b)** il test dei segni per ranghi.

9. Un ingegnere è convinto che la vernice usata su un particolare tipo di aeroplani influisca sulla velocità di crocera. Per accertare questo fatto si fanno volare 10 esemplari appena usciti dalla linea di produzione e prima di verniciarli; successivamente si stende la vernice e si ripete l'esperimento, ottenendo (in nodi) i dati qui sotto:

Velivolo	1	2	3	4	5	6	7	8	9	10
Senza vernice	426.1	418.4	424.4	438.5	440.6	421.8	412.2	409.8	427.5	441.2
Verniciato	416.7	403.2	420.1	431.0	432.6	404.2	398.3	405.4	422.8	444.8

Si può affermare che questi dati supportino l'idea dell'ingegnere?

10. Presentiamo si seguito 10 coppie di determinazioni spettrochimiche per il nichel. Le due serie di dati sono ottenute con due strumenti diversi.

Campione	1	2	3	4	5	6	7	8	9	10
Strumento 1	1.94	1.99	1.98	2.07	2.03	1.96	1.95	1.96	1.92	2.00
Strumento 2	2.00	2.09	1.95	2.03	2.08	1.98	2.03	2.03	2.01	2.12

Verifica al 5% di significatività l'ipotesi che i due strumenti di misurazione siano equivalenti.

11. Sia $X_1, X_2, \ldots, X_n$ un campione estratto da una distribuzione continua F e denotiamo con m la sua mediana; supponiamo di volere verificare l'ipotesi $H_0 : m = m_0$ in alternativa all'ipotesi a una coda $H_1 : m > m_0$. Sviluppa l'analogo a una coda del test dei segni per ranghi. Spiega come calcolare il p-dei-dati.

12. In uno studio sul bilinguismo furono selezionati 12 studenti universitari, ciascuno dei quali mostrava un perfetto bilinguismo inglese-francese; dopo averli divisi a caso in due gruppi da 6, venne dato a tutti un articolo in francese e un questionario con 25 domande a risposta multipla. Per un gruppo le domande erano in francese, mentre per l'altro in inglese; il numero di risposte corrette date dagli studenti è riportato di seguito.

Esame in francese	11	12	16	22	25	25
Esame in inglese	10	13	17	19	21	24

Questi dati, provano al 5% di significatività che esiste una difficoltà nel trasferire le informazioni da una lingua all'altra?

13. Per uno studio sulla sicurezza stradale vengono selezionate 15 città di dimensioni molto simili. Un campione casuale di 8 di esse viene scelto per una campagna giornalistica di informazione sulla sicurezza stradale della durata di un mese. Alla fine di tale periodo, per un altro mese, si registra il numero di incidenti stradali in ciascuna delle 15 città. I dati osservati sono questi:

Gruppo di trattamento	19	31	39	45	47	66	74	81
Gruppo di controllo	28	36	44	49	52	52	60	

Calcola il p-dei-dati esatto nel verificare l'ipotesi che gli articoli non abbiano sortito alcun effetto apprezzabile.

14. Determina nuovamente il p-dei-dati del Problema 13: **(a)** usando l'approssimazione normale; **(b)** con una simulazione.

15. Usa i dati del Problema 44 del Capitolo 7 per verificare con un test non parametrico l'ipotesi che le distribuzioni dei tempi di combustione siano uguali.

(a) Determina il p-dei-dati esatto.

(b) Calcola il p-dei-dati con l'approssimazione normale.

(c) Realizza una simulazione per stimare il valore del p-dei-dati.

16. Risolvi con tecniche non parametriche il Problema 31 del Capitolo 8.

17. In uno studio sugli schemi di diffusione dei castori, nell'arco di 10 anni, nel Parco Nazionale di Allegany (New York), sono stati catturati e marcati 332 di questi roditori, 32 dei quali (9 femmine e 23 maschi) sono poi stati ritrovati stanziati in altre zone. I dati seguenti riportano le distanze (in chilometri) tra il primo sito di cattura e quello successivo di stanziamento:

Femmine				**Maschi**			
0.660	0.984	0.984	1.992	0.288	0.312	0.456	0.528
4.368	6.960	10.656	21.600	0.576	0.720	0.792	0.984
31.680				1.224	1.584	2.304	2.328
				2.496	2.688	3.096	3.408
				4.296	4.884	5.928	6.192
				6.384	13.224	27.600	

Questi dati provano che vi sia una correlazione tra le distanze di dispersione e il sesso?

18. Il confronto di m campioni. Siano dati m campioni indipendenti, di ampiezze rispettivamente $n_1, n_2, \ldots, n_m$, estratti da delle distribuzioni continue $F_1, F_2, \ldots, F_m$; si desidera verificare l'ipotesi nulla $H_0 : F_1 = F_2 = \cdots = F_m$. Per realizzare un test, si raggruppano tutti i dati, si assegnano i ranghi, quindi, per $i = 1, 2, \ldots, m$ si denota con R_i la somma dei ranghi associati agli n_i elementi che provengono dal campione i-esimo.

(a) Dimostra che, quando H_0 è soddisfatta, $E[R_i] = n_i(N+1)/2$, dove si è posto $N = \sum_i n_i$.

(b) Trova una statistica adatta a questo test, usando il risultato del punto (a), e ispirandoti a quella che si usa per il test della somma dei ranghi.

(c) Chiarisci come si possa impiegare un algoritmo che genera una permutazione casuale dei numeri $1, 2, \ldots, N$, per realizzare una simulazione che determini il p-dei-dati relativo alla statistica individuata nel punto (b).

19. Si controllano 50 pezzi usciti consecutivamente da una linea di produzione; quelli che risultano difettosi sono 11, e occupano le posizioni

8 12 13 14 31 32 37 38 40 41 42

Si può concludere che questa successione di pezzi non sia completamente casuale?

20. I livelli qualitativi misurati per 25 articoli sono:

100 110 122 132 99 96 88 75 45 211 154 143 161
142 99 111 105 133 142 150 153 121 126 117 155

Si può pensare che questi dati siano un campione estratto *in maniera casuale* da una qualche popolazione?

21. È possibile modificare il test delle successioni usando come livello di soglia (per assegnare una cifra 0 o 1 a ogni dato), non la mediana campionaria ma un qualunque valore prefissato?

22. La tabella seguente, presa da un articolo[3] del 1987, riporta il livello (alto o basso) di intensità del fenomeno atmosferico "El Nino", nei principali anni in cui si è presentato, dal 1800 al 1987. Usala per vagliare l'ipotesi che le intensità delle manifestazioni del fenomeno si succedano in maniera casuale.

Anno e intensità (0=moderata, 1=forte) per le maggiori manifestazioni di El Nino, 1800–1987

Anno	Intensità	Anno	Intensità	Anno	Intensità	Anno	Intensità
1803	1	1854	0	1896	0	1939	0
1806	0	1857	0	1899	1	1940	1
1812	0	1860	0	1902	0	1943	0
1814	1	1864	1	1905	0	1951	0
1817	0	1866	0	1907	0	1953	0
1819	0	1867	0	1911	1	1957	1
1821	0	1871	1	1914	0	1965	0
1824	0	1874	0	1917	1	1972	1
1828	1	1877	1	1918	0	1976	0
1832	0	1880	0	1923	0	1982	1
1837	0	1884	1	1925	1	1984	0
1844	1	1887	0	1930	0		
1850	0	1891	1	1932	1		

[3] W. H. Quinn, T. V. Neal, Antunez de Mayolo, "El Nino occurrences over the past four-and-a-half centuries", *Journal of Geophysical Research*, vol. 92 (C13), pp. 14449–14461, 1987.

13 Controllo della qualità

Contenuto

13.1 Introduzione

È cosa ben nota che praticamente tutti i processi produttivi introducono una certa variabilità casuale negli oggetti fabbricati: indipendentemente da quanto severamente vengono tenuti sotto controllo i vari stadi, è impossibile ottenere pezzi esattamente uguali al modello, o anche solo identici tra di loro. Questo fenomeno è detto *variazione casuale* e viene considerato inscindibile dal processo. Vi è tuttavia un altro tipo di variazione che può verificarsi: quella dovuta a qualche *causa speciale*, o *assegnabile*, che spesso si traduce in effetti negativi sulla qualità del prodotto. Una configurazione imprecisa delle macchine, una bassa qualità delle materie prime, una limitazione del software, o un errore umano, sono tutte possibili cause assegnabili che si concretizzano in variazioni di questo tipo. Quando non sono presenti cause speciali, e le uniche variazioni tra i pezzi prodotti e il modello sono dovute al caso, diciamo che il processo è *in controllo statistico*. Il problema chiave a cui cercheremo di rispondere in questo capitolo è determinare se e quando un processo sia *fuori controllo*.

Operativamente questo tipo di verifica viene eseguita tramite le *carte di controllo* (*control charts*), le quali consistono di due numeri, che sono detti limiti di controllo inferiore e superiore. I dati generati dal processo produttivo vengono divisi in sottogruppi, dei quali si calcolano alcune statistiche rilevanti, come possono essere la media e la deviazione standard campionarie; poi si traccia un punto sulla carta per ogni sottogruppo, e se tale valore non cade entro i limiti stabiliti, il processo viene dichiarato fuori controllo.

Nelle prossime due sezioni ci concentriamo sulle caratteristiche misurabili (numeriche) degli oggetti prodotti. Assumiamo che quando il processo è in controllo statistico una di tali caratteristiche abbia media e varianza fissate, e quindi mostriamo come costruire carte di controllo basate sulle medie campionarie (Sezione 13.2), e sulle deviazioni standard campionarie (Sezione 13.3). Nella Sezione 13.4 affrontiamo i casi in cui la qualità di ogni pezzo è descritta da un *attributo* che può essere presente o assente, invece che da un numero ("non accettabile" è un esempio di attributo di notevole interesse); anche in tali ipotesi sviluppiamo carte di controllo per determinare delle variazioni nella qualità del processo. Nella Sezione 13.5, costruiamo la carta di controllo per le situazioni in cui ogni oggetto prodotto ha un numero casuale di difetti. Infine, nella Sezione 13.6, discutiamo degli esempi di carte di controllo più sofisticate, che non trattano ogni sottogruppo come una osservazione isolata, ma lo integrano con informazioni provenienti dagli altri sottogruppi. Le tipologie di carte introdotte sono la media mobile (con o senza pesi esponenziali) e le somme cumulate.

13.2 La carta di controllo $\overline{X}$ per il valore medio

Consideriamo la produzione di oggetti che abbiano caratteristiche qualitative misurabili, e supponiamo di sapere che quando il processo è in controllo statistico, i valori di tali caratteristiche sono variabili aleatorie normali di media μ e varianza σ^2. Siccome ammettiamo la possibiltà che il processo vada fuori controllo e gli oggetti prodotti seguano una diversa distribuzione, ci proponiamo di cercare un metodo che permetta di riconoscere tali situazioni, consentendoci quando opportuno di fermare la produzione, cercare il problema e risolverlo.

Siano $X_1, X_2, \ldots$ i valori relativi alle caratteristiche degli oggetti che escono dal processo produttivo. La prima cosa da fare è dividere i dati in sottogruppi di ampiezza n fissata. Il valore di n e la composizione dei sottogruppi devono essere scelti in modo da assicurare l'omogeneità dei dati di ciascun sottogruppo[1]; ad esempio potrebbero essere stati ottenuti nello stesso giorno, durante lo stesso turno, o usando le stesse impostazioni, in modo tale che si possa supporre che le alterazioni nella distribuzione in esame possano avvenire tra un sottogruppo di dati e l'altro ma non all'interno di essi.

[1] La cura necessaria nel formare questi campioni rispettando tali condizioni fa sì che, in italiano, vengano normalmente chiamati "sottogruppi *razionali*", [*N.d.T.*]

Per $i = 1, 2, 3, \ldots$, denotiamo con $\overline{X}_i$ la media campionaria del sottogruppo i-esimo. Quindi ad esempio:

$$\begin{aligned} \overline{X}_1 &:= \frac{X_1 + X_2 + \cdots + X_n}{n} \\ \overline{X}_2 &:= \frac{X_{n+1} + X_{n+2} + \cdots + X_{2n}}{n} \\ &\cdots \\ \overline{X}_i &:= \frac{X_{in-n+1} + X_{in-n+2} + \cdots + X_{in}}{n} \\ &\cdots \end{aligned} \tag{13.2.1}$$

Supponiamo ora che il processo sia sotto controllo durante la produzione di questi sottogruppi. Ciò significa che ciascuna delle X_i ha distribuzione $\mathcal{N}(\mu, \sigma^2)$, quindi

$$E[\overline{X}_i] = \mu \qquad \text{e} \qquad \text{Var}(\overline{X}_i) = \frac{\sigma^2}{n} \tag{13.2.2}$$

ovvero,

$$\frac{\overline{X}_i - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

Sia Z una variabile aleatoria normale standard; sappiamo già che Z è quasi sempre compresa tra -3 e $+3$ (in effetti, $P(-3 < Z < 3) \approx 0.9973$), quindi ci aspettiamo di osservare che

$$-3 < \frac{\overline{X}_i - \mu}{\sigma/\sqrt{n}} < 3$$

o equivalentemente

$$\mu - \frac{3\sigma}{\sqrt{n}} < \overline{X}_i < \mu + \frac{3\sigma}{\sqrt{n}}$$

I valori

$$\text{UCL} := \mu + \frac{3\sigma}{\sqrt{n}} \qquad \text{e} \qquad \text{LCL} := \mu - \frac{3\sigma}{\sqrt{n}} \tag{13.2.3}$$

sono detti rispettivamente *limite di controllo superiore* e *limite di controllo inferiore*[2].

La *carta di controllo* $\overline{X}$ ha lo scopo di determinare una alterazione nel valore medio della distribuzione; essa si ottiene tracciando le diverse medie campionarie $\overline{X}_i$ e dichiarando il processo fuori controllo non appena uno di questi valori non cade tra LCL e UCL (si veda la Figura 13.1).

Esempio 13.2.1. Un'azienda produce aste in acciaio con diametro distribuito con media di 3 mm e deviazione standard di 0.1 mm. Campioni successivi di 4 aste ciascuno hanno fornito le seguenti medie campionarie:

[2] Queste sigle derivano ovviamente dalle espressioni inglesi, *upper* e *lower control limit*, [*N.d.T.*]

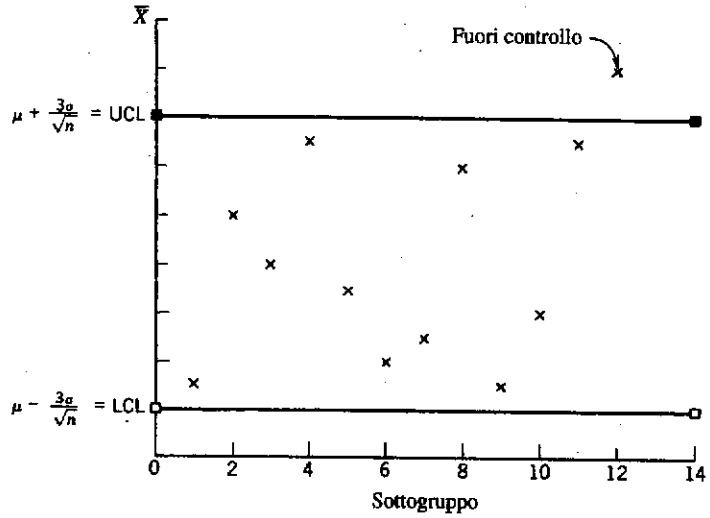


Figura 13.1 Carta di controllo per $\overline{X}$, n è la numerosità dei sottogruppi.

Campione	1	2	3	4	5	6	7	8	9	10
$\overline{X}$	3.01	2.97	3.12	2.99	3.03	3.02	3.10	3.14	3.09	3.20

Che conclusione si deve trarre?

Nello stato di controllo, i diametri hanno media $\mu = 3$ e deviazione standard $\sigma = 0.1$; i campioni hanno numerosità $n = 4$, quindi i limiti di controllo sono

$$\text{LCL} = 3 - \frac{3 \times 0.1}{\sqrt{4}} = 2.85, \qquad \text{UCL} = 3 + \frac{3 \times 0.1}{\sqrt{4}} = 3.15$$

Siccome il decimo campione ha media 3.20 e cade oltre il limite di controllo superiore, vi è ragione di sospettare che il diametro medio delle aste sia cambiato. (A giudicare dai risultati dei campioni dal 5 al 10, μ potrebbe avere superato i 3 mm.) □

Osservazione 13.2.1. Anche se abbiamo supposto nei paragrafi precedenti che la distribuzione delle singole osservazioni X_i fosse normale, i ragionamenti fatti sono approssimativamente corretti anche quando questa ipotesi non sussiste; infatti in virtù del teorema del limite centrale le $\overline{X}_i$ sono comunque approssimativamente normali, e quindi non è probabile che si discostino dalla loro media per più di 3 deviazioni standard.

Osservazione 13.2.2. È frequente che non si disponga dei valori misurati di tutti i pezzi prodotti, ma solo di campioni casuali ristretti. In questo caso è naturale scegliere, come sottogruppi, oggetti prodotti in momenti vicini. Questa scelta va fatta comunque tenendo conto che n deve valere tipicamente almeno 4, 5 o 6.

È opportuno notare che anche quando il processo è sotto controllo vi è una piccola probabilità (per la precisione: 0.0027), che la media campionaria di un sottogruppo cada esternamente ai limiti di controllo, costringendoci a fermare il processo e a cercare un difetto inesistente.

Supponiamo ora che il processo sia appena andato fuori controllo perché la media è passata μ a $\mu + a$, con $a > 0$. Quanto tempo ci può volere perché la carta rilevi che il processo è fuori controllo (sempre che non vi siano altre variazioni nella media)? Sappiamo che la media di un sottogruppo cade entro i limiti di controllo se

$$-3 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 3$$

o equivalentemente, se

$$-3 - \frac{a\sqrt{n}}{\sigma} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} - \frac{a\sqrt{n}}{\sigma} < 3 - \frac{a\sqrt{n}}{\sigma}$$

ovvero

$$-3 - \frac{a\sqrt{n}}{\sigma} < \frac{\overline{X} - \mu - a}{\sigma/\sqrt{n}} < 3 - \frac{a\sqrt{n}}{\sigma}$$

Siccome $\overline{X}$ è normale con media $\mu + a$ e varianza σ^2/n, si ha che $\sqrt{n}(\overline{X} - \mu - a)/\sigma$ ha distribuzione $\mathcal{N}(0,1)$, e quindi la probabilità che l'osservazione cada entro i limiti di controllo si riscrive come

$$\begin{aligned} P\left(-3 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 3\right) &= P\left(-3 - \frac{a\sqrt{n}}{\sigma} < Z < 3 - \frac{a\sqrt{n}}{\sigma}\right) \\ &= \Phi\left(3 - \frac{a\sqrt{n}}{\sigma}\right) - \Phi\left(-3 - \frac{a\sqrt{n}}{\sigma}\right) \\ &\approx \Phi\left(3 - \frac{a\sqrt{n}}{\sigma}\right) \end{aligned}$$

dove abbiamo indicato con Z una variabile aleatoria normale standard. Di conseguenza la probabilità che la media di un sottogruppo cada fuori dai limiti di controllo è approssimativamente $1 - \Phi(3 - a\sqrt{n}/\sigma)$. Se l'ampiezza dei sottogruppi fosse ad esempio $n = 4$ e l'aumento della media fosse stato di 1 deviazione standard (intendendo con questo che $a = \sigma$), questa probabilità sarebbe perciò prossima a $1 - \Phi(1) \approx 0.159$. Poiché in ciascun sottogruppo – indipendentemente dagli altri – si rileva lo stato di fuori controllo con probabilità $1 - \Phi(3 - a\sqrt{n}/\sigma)$, si ha che il numero di sottogruppi da controllare prima che questo accada è una variabile aleatoria geometrica con media

$$\lambda := \left[1 - \Phi\left(3 - \frac{a\sqrt{n}}{\sigma}\right)\right]^{-1} \tag{13.2.4}$$

Nell'esempio proposto con $n = 4$ e $a = \sigma$, il numero di sottogruppi da ispezionare prima di notare che il processo è fuori controllo ha distribuzione geometrica di media 6.3.

13.2.1 Il caso in cui μ e σ siano incognite

Se all'inizio della compilazione di una carta di controllo non si dispone di dati storici affidabili, c'è il problema di stimare μ e σ, visto che queste due quantità non sono in tal caso note. Si riserva quindi inizialmente un certo numero k di sottogruppi per eseguire questa stima; k va scelto piuttosto grande se si desidera ottenere risultati precisi: di solito si chiede che $k \geq 20$ e $nk \geq 100$. Lo stimatore naturale di μ è media aritmetica delle medie campionarie dei sottogruppi:

$$\overline{\overline{X}} := \frac{\overline{X}_1 + \overline{X}_2 + \cdots + \overline{X}_k}{k} \tag{13.2.5}$$

Per stimare σ usiamo le deviazioni standard campionarie dei sottogruppi: poniamo infatti

$$\begin{aligned} S_1 &:= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_1)^2} \\ S_2 &:= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{n+i} - \overline{X}_2)^2} \\ &\cdots \\ S_k &:= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_{kn-n+i} - \overline{X}_k)^2} \end{aligned} \tag{13.2.6}$$

e denotiamo con $\overline{S}$ la media aritmetica di questi valori.

$$\overline{S} := \frac{S_1 + S_2 + \cdots + S_k}{k} \tag{13.2.7}$$

La statistica $\overline{S}$ *non* è uno stimatore corretto di σ. Infatti è immediato che

$$E[\overline{S}] = \frac{E[S_1] + E[S_2] + \cdots + E[S_k]}{k} = E[S_1]$$

tuttavia $E[S_1] \neq \sigma$. Di seguito calcoliamo il valore esatto di $E[\overline{S}]$ per mostrare che è diverso da σ e per trovare il coefficiente moltiplicativo che permetterà di trasformare $\overline{S}$ in uno stimatore corretto di σ.

Ricordiamo intanto che per i campioni normali vale il risultato

$$Y := (n-1)\frac{S_1^2}{\sigma^2} = \sum_{i=1}^{n}\frac{(X_i - \overline{X}_1)^2}{\sigma^2} \sim \chi^2_{n-1} \tag{13.2.8}$$

Non è inoltre difficile provare (si veda il Problema 3) che se $Y \sim \chi^2_{n-1}$,

$$E[\sqrt{Y}] = \frac{\Gamma(\frac{n}{2})\sqrt{2}}{\Gamma(\frac{n-1}{2})} \tag{13.2.9}$$

dove la funzione $\Gamma(\cdot)$ è la gamma di Eulero, definita a pagina 185. Siccome d'altra parte

$$E[\sqrt{Y}] = E\left[\sqrt{(n-1)\frac{S_1^2}{\sigma^2}}\right] = \sqrt{n-1}\frac{E[S_1]}{\sigma}$$

otteniamo che

$$\begin{aligned} E[\overline{S}] &= E[S_1] \\ &= E[\sqrt{Y}]\frac{\sigma}{\sqrt{n-1}} \\ &= \frac{\sigma\Gamma(\frac{n}{2})\sqrt{2}}{\Gamma(\frac{n-1}{2})\sqrt{n-1}} \end{aligned}$$

Perciò se si pone

$$c(n) := \frac{\Gamma(\frac{n}{2})\sqrt{2}}{\Gamma(\frac{n-1}{2})\sqrt{n-1}} \tag{13.2.10}$$

si ottiene che

$$E[S_1] = \sigma c(n) \tag{13.2.11}$$

e $\overline{S}/c(n)$ risulta uno stimatore corretto di σ.

La Tabella 13.1 presenta i valori di $c(n)$ per n che va da 2 a 10.

Osservazione 13.2.3. Come si calcolano i valori $\Gamma(\frac{n}{2})$ e $\Gamma(\frac{n-1}{2})$ necessari a determinare i coefficienti $c(n)$? Per compilare la Tabella 13.1 è stata usata la formula ricorsiva

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

che è stata provata nella Sezione 5.7. Essa permette di stabilire il valore della funzione gamma sugli interi:

$$\begin{aligned} \Gamma(n) &= (n-1)\Gamma(n-1) \\ &= (n-1)(n-2)\Gamma(n-2) \\ &= (n-1)!\Gamma(1) \\ &= (n-1)! \qquad \text{perché } \Gamma(1) = \textstyle\int_0^\infty e^{-t}dt = 1 \end{aligned}$$

Tabella 13.1 Valori del coefficiente $c(n)$, definito dall'Equazione (13.2.10)

n	$c(n)$
2	0.797885
3	0.886227
4	0.921318
5	0.939986
6	0.951533
7	0.959369
8	0.965031
9	0.969311
10	0.972659

e sugli interi più $\frac{1}{2}$:

$$\Gamma\left(n+\frac{1}{2}\right) = \left(n-\frac{1}{2}\right)\cdot\Gamma\left(n-\frac{1}{2}\right)$$
$$= \left(n-\frac{1}{2}\right)\left(n-\frac{3}{2}\right)\cdots\frac{3}{2}\cdot\frac{1}{2}\cdot\Gamma\left(\frac{1}{2}\right)$$

dove $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. (Si veda a pagina 203 il Problema 45 del Capitolo 5.)

Le stime precedenti per μ e σ fanno uso dei valori di k sottogruppi, e quindi sono sensate solo se il processo è rimasto stabile durante la produzione di questi oggetti. Per esercitare un controllo su questa assunzione molto importante, si possono calcolare i valori di UCL ed LCL basati sulle stime, ovvero:

$$\text{LCL} := \overline{\overline{X}} - \frac{3\overline{S}}{c(n)\sqrt{n}} \quad \text{e} \quad \text{UCL} := \overline{\overline{X}} + \frac{3\overline{S}}{c(n)\sqrt{n}} \tag{13.2.12}$$

e poi verificare che le medie campionarie di tutti i k sottogruppi usati cadano tra il limite inferiore e quello superiore. Se così non è, occorre scartare i valori anomali (immaginando che il processo sia stato fuori controllo momentaneamente), e ricalcolare le stime e i limiti, verificando poi nuovamente che tutti i sottogruppi questa volta cadano entro tali valori, iterando ancora il procedimento se necessario. Chiaramente, se le medie che escono dai limiti di controllo fossero troppe, si dovrebbe concludere che il processo è fuori controllo.

Esempio 13.2.2. Riconsideriamo l'Esempio 13.2.1, questa volta immaginando che μ e σ non siano note. Forniamo anche i valori delle deviazioni standard campionarie:

Campione	1	2	3	4	5	6	7	8	9	10
$\overline{X}$	3.01	2.97	3.12	2.99	3.03	3.02	3.10	3.14	3.09	3.20
S	0.12	0.14	0.08	0.11	0.09	0.08	0.15	0.16	0.13	0.16

Visto che $\overline{\overline{X}} = 3.067$, $\overline{S} = 0.122$ e $c(4) \approx 0.9213$, i limiti di controllo risultano

$$\text{LCL} = 3.067 - \frac{3\times 0.122}{2\times 0.9213} \approx 2.868$$
$$\text{UCL} = 3.067 + \frac{3\times 0.122}{2\times 0.9213} \approx 3.266$$

Tutti i valori $\overline{X}_i$ cadono entro questi limiti, quindi facciamo l'assunzione che il processo sia in controllo statistico, con $\mu \approx 3.067$ e $\sigma = \overline{S}/c(4) \approx 0.1324$.

Ipotizziamo che siano accettabili solo i pezzi a cui corrispondono valori che rientrano nelle specifiche 3 ± 0.1; assumendo che il processo non vada fuori controllo, e che le stime date siano accurate, che percentuale degli oggetti prodotti soddisfa le richieste?

Sfruttando il fatto che $X \sim \mathcal{N}(\mu, \sigma^2)$, e che $\mu \approx 3.067$ e $\sigma \approx 0.1324$, troviamo:

$$P(2.9 \le X \le 3.1) \approx P\left(\frac{2.9-3.067}{0.1324} \le \frac{X-3.067}{0.1324} \le \frac{3.1-3.067}{0.1324}\right)$$
$$\approx \Phi(0.2492) - \Phi(-1.2613)$$
$$\approx 0.5984 - (1 - 0.8964) = 0.4948$$

Per cui il 49% degli oggetti prodotti soddisferà le specifiche. □

Osservazione 13.2.4. In passato, per ridurre la quantità di calcoli necessari, è stato molto usato uno stimatore di σ che si basava sul range dei sottogruppi (definito come la differenza tra l'osservazione maggiore e la minore). Comunque con la potenza di calcolo dei giorni nostri non ha alcun senso prediligere tale stimatore solo perché è più semplice da calcolare; lo stimatore basato sulle deviazioni standard campionarie ha una varianza minore ed è più robusto (nel senso che fornisce una stima ragionevolmente corretta anche quando si perde l'ipotesi di normalità). Per questo motivo l'altro stimatore non viene affrontato in questo testo.

13.3 La carta di controllo S

Le carte di controllo $\overline{X}$, presentate nella sezione precedente, sono concepite con lo scopo di rilevare cambiamenti nella media della popolazione. Nel caso che si sia interessati anche a possibili alterazioni nella varianza, si devono usare anche le carte di controllo S.

Come in precedenza, supponiamo che quando il processo è sotto controllo, le caratteristiche misurabili dei pezzi prodotti abbiano distribuzione normale $\mathcal{N}(\mu, \sigma^2)$.

Quale stimatore per σ?

Lo stimatore $\overline{\overline{X}}$ è uguale alla media aritmetica di tutte le nk osservazioni, e quindi è lo stimatore più naturale per μ. Potrebbe invece non essere chiaro perché per stimare la deviazione standard σ non si sia usata la deviazione standard campionaria dell'intera collezione di dati,

$$S := \sqrt{\frac{1}{n-1}\sum_{i=1}^{nk}(X_i - \overline{\overline{X}})^2}$$

Il motivo è che il processo potrebbe non essere stato sotto controllo in corrispondenza di tutti i k sottogruppi, e in tal caso quest'ultimo stimatore sarebbe molto distante dal valore reale di σ. L'andare fuori controllo del processo in un sottogruppo, infatti, consiste spesso in un cambiamento della media μ, con la deviazione standard che rimane invariata. Quando si presenta questa situazione, le deviazioni standard campionarie dei sottogruppi sono ancora dei buoni stimatori di σ, mentre S tende necessariamente a sovrastimare.
Persino quando il processo sembra essere rimasto in controllo statistico in corrispondenza di tutti i sottogruppi, si preferisce $\overline{S}/c(n)$ alla deviazione standard campionaria di tutte le osservazioni. Infatti anche se le medie relative a tutti i sottogruppi cadono entro i limiti, e quindi abbiamo concluso che il processo è sotto controllo, ciò non significa che questo sia vero (possono esservi cause speciali di variazione che hanno causato un cambiamento che non è ancora stato rilevato dalla carta); significa solamente che la nostra strategia non prevede di eseguire un blocco e una revisione fino a che non saremo relativamente certi di essere fuori controllo; nel frattempo conviene comportarci come se il processo fosse in controllo statistico e lasciare ancora che produca oggetti.
In conclusione, siccome ammettiamo che potrebbe in ogni caso essere presente una causa speciale di variazione, prediligiamo $\overline{S}/c(n)$, che è uno stimatore più "prudente" della deviazione standard campionaria: anche se non è uno stimatore altrettanto buono quando il processo è rimasto stabile tutto il tempo, può diventare molto migliore quando vi siano state delle variazioni della media, anche non riscontrate.

Sia S_i la deviazione standard campionaria delle osservazioni nel sottogruppo i, ovvero

$$S_i := \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_{in-n+j} - \overline{X}_i)^2} \tag{13.3.1}$$

Il valore atteso di S_i è stato calcolato nella Sezione 13.2.1, e vale

$$E[S_i] = c(n)\sigma \tag{13.3.2}$$

Per quanto riguarda la varianza, invece, si noti che

$$(n-1)\frac{S_i^2}{\sigma^2} \sim \chi^2_{n-1}$$

e quindi

$$(n-1)\frac{E[S_i^2]}{\sigma^2} = E[\chi^2_{n-1}] = n-1$$

Usando allora che $E[S_i^2] = \sigma^2$, si trova immediatamente

$$\begin{aligned} \mathrm{Var}(S_i) &= E[S_i^2] - E[S_i]^2 \\ &= \sigma^2 - c^2(n)\sigma^2 \\ &= \sigma^2(1 - c^2(n)) \end{aligned} \tag{13.3.3}$$

A partire dall'ipotesi che, quando il processo è in controllo statistico, la distribuzione di S_i è quella di un multiplo fissato della radice di una chi-quadro con $n-1$ gradi di libertà, è possibile dimostrare che S_i cade entro 3 deviazioni standard dalla sua media, con probabilità prossima a uno:

$$P\big(E[S_i] - 3\sqrt{\mathrm{Var}(S_i)} < S_i < E[S_i] + 3\sqrt{\mathrm{Var}(S_i)}\big) \approx 0.99$$

Perciò, usando le espressioni per $E[S_i]$ e $\mathrm{Var}(S_i)$ date dalle Equazioni (13.3.2) e (13.3.3), è naturale fissare i limiti di controllo della carta S ai valori:

$$\begin{aligned} \mathrm{UCL} &:= \sigma\{c(n) + 3\sqrt{1 - c^2(n)}\} \\ \mathrm{LCL} &:= \sigma\{c(n) - 3\sqrt{1 - c^2(n)}\} \end{aligned} \tag{13.3.4}$$

L'uso della carta S è analogo a quello della carta $\overline{X}$. I valori successivi degli stimatori S_i vanno tracciati sul piano cartesiano, e non appena uno di essi non rientra nei limiti di controllo stabiliti, il processo produttivo va interrotto e dichiarato fuori controllo.

Se quando si avvia la carta di controllo σ non è nota, è possibile stimarne il valore tramite $\overline{S}/c(n)$, ottenendo i limiti di controllo,

$$\begin{aligned} \mathrm{UCL} &:= \overline{S}\{1 + 3\sqrt{c^{-2}(n) - 1}\} \\ \mathrm{LCL} &:= \overline{S}\{1 - 3\sqrt{c^{-2}(n) - 1}\} \end{aligned} \tag{13.3.5}$$

Analogamente a quanto detto per la carta $\overline{X}$, è necessario accertarsi che tutti le k deviazioni standard campionarie $S_1, S_2, \ldots, S_k$, usate per stimare σ cadano entro questi limiti. Se qualcuno di questi valori cade al di fuori, il sottogruppo corrispondente va scartato, e occorre ricalcolare $\overline{S}$.

Esempio 13.3.1. Quelli che seguono sono i valori di $\overline{X}$ e S per 20 sottogruppi di ampiezza 5 per un processo avviato recentemente.

Sottogruppo	1	2	3	4	5	6	7	8	9	10
$\overline{X}$	35.1	33.2	31.7	35.4	34.5	36.4	35.9	38.4	35.7	27.2
S	4.2	4.4	2.5	3.2	2.6	4.5	3.4	5.1	3.8	6.2

Sottogruppo	11	12	13	14	15	16	17	18	19	20
$\overline{X}$	38.1	37.6	38.8	34.3	43.2	41.3	35.7	36.3	35.4	34.6
S	4.2	3.9	3.2	4.0	3.5	8.2	8.1	4.2	4.1	3.7

Visto che $\overline{\overline{X}} = 35.94$, $\overline{S} = 4.35$ e $c(5) \approx 0.9400$, usando le Equazioni (13.2.12) e (13.3.5), i limiti di controllo per $\overline{X}$ e S risultano

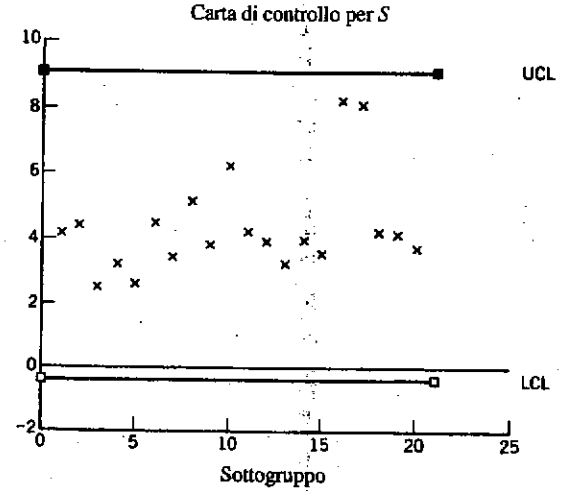
$$\begin{aligned} \mathrm{LCL}(\overline{X}) &\approx 29.73 \\ \mathrm{UCL}(\overline{X}) &\approx 42.15 \\ \mathrm{LCL}(S) &\approx -0.386 \\ \mathrm{UCL}(S) &\approx 9.087 \end{aligned}$$

Le carte di controllo per $\overline{X}$ e S con i limiti di controllo precedenti sono rappresentate nelle Figure 13.2 **(a)** e **(b)**. Poiché $\overline{X}_{10}$ e $\overline{X}_{15}$ cadono fuori dai limiti di controllo, questi sottogruppi vanno eliminati e i limiti di controllo ricalcolati. Questo compito viene affidato allo studente e costituisce il Problema 5. □

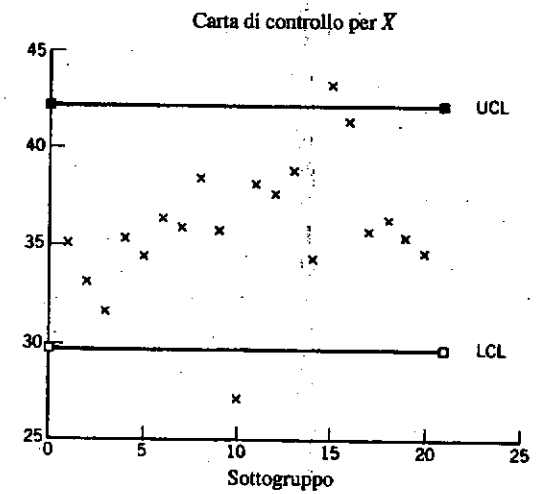
13.4 Carte di controllo per attributi

Le carte di controllo $\overline{X}$ e S sono pensate per dati che possono assumere un intervallo continuo di valori; altre volte gli oggetti prodotti possono avere invece caratteristiche *qualitative* – anche denominate *attributi* – come ad esempio l'essere difettosi o accettabili. Anche in tali situazioni è possibile ricorrere a delle specifiche carte di controllo.

Supponiamo che quando il processo è sotto controllo i pezzi prodotti possano essere difettosi con probabilità p, indipendentemente l'uno dall'altro. Sia X il numero di elementi difettosi all'interno di un sottogruppo di n oggetti, e sia $F := X/n$ la frazione corrispondente, indicante quale parte degli elementi del sottogruppo sia difettosa; quando il processo è in controllo statistico, X ha distribuzione binomiale di



(a)



(b)

Figura 13.2

parametri (n, p), e quindi F ha media e deviazione standard date da

$$E[F] = \frac{E[X]}{n} = \frac{np}{n} = p$$
$$\sqrt{\mathrm{Var}(F)} = \sqrt{\frac{\mathrm{Var}(X)}{n^2}} = \sqrt{\frac{np(1-p)}{n^2}} = \sqrt{\frac{p(1-p)}{n}} \tag{13.4.1}$$

Perciò quando il processo è sotto controllo la frazione di difettosi in un sottogruppo di ampiezza n dovrebbe cadere entro i limiti

$$\mathrm{LCL} := p - 3\sqrt{\frac{p(1-p)}{n}}$$
$$\mathrm{UCL} := p + 3\sqrt{\frac{p(1-p)}{n}} \tag{13.4.2}$$

con elevata probabilità.

L'ampiezza n dei sottogruppi per questa carta di controllo, va normalmente scelta molto maggiore dei valori tipici tra 4 e 10 utilizzati per le carte $\overline{X}$ e S. Infatti, se p è piccola e n non è sufficientemente grande, una gran parte dei sottogruppi non presenterà alcun elemento difettoso, persino nel caso che il processo sia fuori controllo, e quindi sarà necessario attendere molto tempo prima che la carta rilevi una variazione nella qualità.

Per inizializzare una carta di controllo di questo tipo, occorre prima stimare p; scegliendo k sottogruppi a questo scopo (per ottenere risultati precisi è preferibile che sia $k \geq 20$), e denotando con F_i la frazione di difettosi nel sottogruppo i, la stima di p è data da

$$\overline{F} := \frac{F_1 + F_2 + \cdots + F_k}{k} \tag{13.4.3}$$

Siccome $nF_i = X_i$ è il numero di elementi difettosi nel sottogruppo i, si vede che $\overline{F}$ può anche essere espressa nella forma

$$\overline{F} = \frac{nF_1 + nF_2 + \cdots + nF_k}{nk}$$
$$= \frac{\text{numero di pezzi difettosi in tutti i sottogruppi}}{\text{numero totale di pezzi nei sottogruppi}}$$

quindi la stima di p non è altro che la frazione di tutti i pezzi esaminati che è risultata difettosa.

Le stime dei limiti di controllo basate su $\overline{F}$ sono naturalmente:

$$\mathrm{LCL} := \overline{F} - 3\sqrt{\frac{\overline{F}(1-\overline{F})}{n}}$$
$$\mathrm{UCL} := \overline{F} + 3\sqrt{\frac{\overline{F}(1-\overline{F})}{n}} \tag{13.4.4}$$

ed è necessario verificare che tutti i valori $F_1, F_2, \ldots, F_k$ cadano entro questi estremi; nel caso alcuni di essi non soddisfino questa condizione, vanno esclusi e si deve ricalcolare $\overline{F}$.

Esempio 13.4.1. Per controllare una macchina automatica per la fabbricazione di viti di ferro, ogni ora si preleva un campione di 50 viti consecutive, che vengono ispezionate e catalogate come accettabili o difettose. I dati seguenti sono relativi a 20 sottogruppi di questo tipo.

Sottogruppo	1	2	3	4	5	6	7	8	9	10
Viti difettose	6	5	3	0	1	2	1	0	2	1
F	0.12	0.10	0.06	0.00	0.02	0.04	0.02	0.00	0.04	0.02

Sottogruppo	11	12	13	14	15	16	17	18	19	20
Viti difettose	1	3	2	0	1	1	0	2	1	2
F	0.02	0.06	0.04	0.00	0.02	0.02	0.00	0.04	0.02	0.04

I limiti di controllo possono essere calcolati come segue:

$$\overline{F} = \frac{\text{numero totale di viti difettose}}{\text{numero totale di viti ispezionate}} = \frac{34}{1\,000} = 0.034$$
$$\mathrm{UCL} = 0.34 + 3\sqrt{\frac{0.34 \times 0.968}{50}} \approx 0.1109$$
$$\mathrm{LCL} = 0.34 - 3\sqrt{\frac{0.34 \times 0.968}{50}} \approx -0.0429$$

Notiamo che $F_1 = 0.12$ è l'unico valore a non rientrare nei limiti di controllo. Una volta che sia stato rimosso, si ricalcolano $\overline{F}$ e i limiti, trovando che

$$\overline{F} = \frac{34 - 6}{1\,000 - 50} \approx 0.0295$$

e quindi i limiti di controllo sono dati da $0.0295 \pm \sqrt{0.0295 \times 0.9705/50}$, ovvero

$$\mathrm{UCL} \approx 0.1013, \qquad \mathrm{LCL} \approx -0.0423$$

Siccome i sottogruppi rimanenti hanno tutti valori entro questi estremi, accettiamo che, quando il processo è sotto controllo, la frazione di oggetti difettosi in un sottogruppo non deve superare 0.1013. □

13.5 Carte di controllo per il numero di non conformità

In questa sezione consideriamo dati che rappresentano il numero di non conformità o difetti riscontrati in unità che possono essere costituite da un oggetto singolo o da

Fuori controllo per eccesso di pezzi accettabili

Si noti che, in linea di principio, una carta di questo tipo rileva le variazioni di qualità sia positive sia negative; quindi il processo viene considerato "fuori controllo" anche se per qualche motivo la percentuale di pezzi difettosi è diminuita. Non si tratta di un errore: in realtà è importante percepire ogni variazione sensibile nel livello qualitativo, per potere trovare la sua causa. Se si riscontra un miglioramento nella qualità dei prodotti, è molto utile analizzare il processo di produzione per scoprirne il motivo (ed eventualmente applicare tale scoperta ad altre linee di produzione).

gruppi di oggetti. Esempi potrebbero essere il numero di rivetti difettosi sull'ala di un aereo, o il numero di circuiti integrati difettosi che vengono prodotti giornalmente da una compagnia. Poiché normalmente in questi casi vi è un numero elevato di possibili parti difettose, ciascuna delle quali ha solo una piccola probabilità di esserlo davvero, si considera ragionevole supporre che la distribuzione del numero totale di difetti riscontrati sia di Poisson[3]. Facciamo quindi l'assunzione che, quando il processo è sotto controllo, il numero di difetti per unità sia una variabile aleatoria poissoniana di media λ.

Denotiamo con X_i il numero di difetti riscontrati nella unità i; siccome la varianza di una poissoniana coincide con la sua media, quando il processo è sotto controllo,

$$E[X_i] = \lambda, \qquad \mathrm{Var}(X_i) = \lambda \tag{13.5.1}$$

e di conseguenza ogni singolo valore X_i dovrebbe cadere entro $\lambda \pm 3\sqrt{\lambda}$ con elevata probabilità. I limiti di controllo vengono quindi definiti nel modo seguente:

$$\mathrm{LCL} := \lambda - 3\sqrt{\lambda}, \qquad \mathrm{UCL} := \lambda + 3\sqrt{\lambda} \tag{13.5.2}$$

Qualora all'inizializzazione della carta di controllo il valore di λ non fosse noto, è ancora possibile stimarlo tramite k osservazioni. Lo stimatore naturale è:

$$\overline{X} := \frac{X_1 + X_2 + \cdots + X_k}{k} \tag{13.5.3}$$

e le stime dei limiti di controllo risultanti sono:

$$\mathrm{LCL} := \overline{X} - 3\sqrt{\overline{X}}, \qquad \mathrm{UCL} := \overline{X} + 3\sqrt{\overline{X}} \tag{13.5.4}$$

[3] Si veda la Sezione 5.2 per una spiegazione di questo fatto.

Se tutte le osservazioni $X_1, X_2, \ldots, X_k$ cadono entro questi estremi, possiamo supporre che il processo sia sotto controllo e assumere $\lambda = \overline{X}$. Altrimenti è necessario escludere i valori anomali, ricalcolare le stime e così via.

Nel caso che il numero medio di difetti per unità sia piccolo, risulta molto conveniente combinare un certo numero n di unità e usare come dati il numero totale di difetti riscontrati in questi raggruppamenti. Siccome la somma di variabili aleatorie di Poisson indipendenti è ancora una poissoniana (con una media maggiore), i dati trasformati in questo modo avranno ancora distribuzione di Poisson. Nella pratica questa tecnica si rivela effettivamente utile quando il numero medio di difetti per unità è inferiore a 25. L'esempio seguente illustra in dettaglio i vantaggi del raggruppamento.

Esempio 13.5.1. Supponiamo che quando il processo è in controllo statistico, il numero medio di difetti per oggetto sia 4; accade quindi qualcosa, e questo valore cambia improvvisamente da 4 a 6: si ha quindi un incremento di 1 deviazione standard. Immaginiamo di raggruppare gli oggetti prodotti n alla volta e vediamo quanti oggetti vengono prodotti, in media, prima che il processo sia dichiarato fuori controllo. Alla fine stabiliremo quali sono i valori di n che rendono minima questa quantità.

Il numero di difetti in un gruppo di n oggetti è, sotto controllo, una variabile aleatoria di Poisson con media e varianza $4n$, quindi i limiti di controllo da adottare sono $4n \pm 3\sqrt{4n} = 4n \pm 6\sqrt{n}$. Siccome in realtà il processo è fuori controllo e il numero medio di difetti per oggetto è 6, i dati hanno distribuzione di Poisson con media e varianza $6n$. Sia Y una variabile aleatoria con tale distribuzione. Se denotiamo con $p(n)$ la probabilità che un dato cada all'esterno dei limiti di controllo, si ha

$$\begin{aligned} p(n) &:= P(Y < 4n - 6\sqrt{n}) + P(Y > 4n + 6\sqrt{n}) \\ &\approx P(Y > 4n + 6\sqrt{n}) \\ &= P\left(\frac{Y - 6n}{\sqrt{6n}} > \frac{4n + 6\sqrt{n} - 6n}{\sqrt{6n}}\right) \\ &\approx P\left(Z > \frac{6\sqrt{n} - 2n}{\sqrt{6n}}\right) \qquad \text{dove } Z \sim \mathcal{N}(0,1) \\ &= 1 - \Phi\left(\sqrt{6} - \sqrt{\frac{2n}{3}}\right) \end{aligned}$$

Poiché ogni dato ha probabilità $p(n)$ di cadere fuori dai limiti di controllo, il numero di dati che devono essere analizzati prima di dichiarare il processo fuori controllo è una variabile aleatoria geometrica di media $1/p(n)$. Siccome servono n oggetti per fare un dato, il numero medio di oggetti prodotti prima che venga rilevata la

variazione di λ è $n/p(n)$:

$$\text{numero medio di oggetti prodotti fuori controllo} \approx \frac{n}{1-\Phi(\sqrt{6}-\sqrt{2n/3})}$$

La Tabella 13.2 riporta i valori di questa espressione per diverse scelte di n. Si noti che quando il processo è in controllo statistico, conviene che n sia più grande possibile (perché il numero medio di oggetti prodotti prima che venga erroneamente rilevato lo stato di fuori controllo è circa $n/0.0027$). Perciò consultando la Tabella 13.2 appare evidente che conviene combinare almeno 9 oggetti. Ciò significa che ogni dato (ottenuto raggruppando n oggetti) avrà media almeno pari a $9 \times 4 = 36$. □

Esempio 13.5.2. I dati seguenti rappresentano il numero di difetti trovati su unità successive di 10 automobili ciascuna.

141	162	150	111	92	74	85	95	76	68
63	74	103	81	94	68	95	81	102	73

Secondo questi dati, il processo è rimasto sotto controllo per tutto il tempo?

Siccome $\overline{X} = 94.4$, segue che i limiti di controllo sono

$$\text{LCL} = 94.4 - 3\sqrt{94.4} \approx 65.25$$
$$\text{UCL} = 94.4 + 3\sqrt{94.4} \approx 123.55$$

I primi tre dati sono superiori a UCL, quindi vanno esclusi. La nuova media campionaria è

$$\overline{X} = \frac{94.4 \times 20 - (141 + 162 + 150)}{17} \approx 84.41$$

Tabella 13.2

n	Numero medio di oggetti
1	19.60
2	20.66
3	19.80
4	19.32
5	18.80
6	18.18
7	18.13
8	18.02
9	18.00
10	18.18
11	18.33
12	18.51

e i nuovi limiti di controllo risultano

$$\text{LCL} = 84.41 - 3\sqrt{84.41} \approx 56.85$$
$$\text{UCL} = 84.41 + 3\sqrt{84.41} \approx 111.97$$

A questo punto tutti i 17 dati restanti cadono entro i limiti e potremmo dichiarare che il processo è in controllo statistico, con valore medio 84.41. Siccome però sembra di capire dai dati che il numero medio di difetti fosse più elevato in una prima fase, per poi stabilizzarsi in stato di controllo, si può pensare che anche il quarto dato, che è piuttosto alto, sia stato generato prima che il processo fosse sotto controllo. È quindi consigliabile eliminare anche quel valore, e ricalcolare con i 16 dati restanti,

$$\overline{X} = 82.56$$
$$\text{LCL} = 82.56 - 3\sqrt{82.56} \approx 55.30$$
$$\text{UCL} = 82.56 + 3\sqrt{82.56} \approx 109.82$$

concludendo quindi che il processo appare ora sotto controllo, con un valore medio di 82.56. □

13.6 Altre carte di controllo per la media

La principale debolezza della carta di controllo $\overline{X}$ presentata nella Sezione 13.2 è che essa si dimostra relativamente insensibile a piccole variazioni nella media di popolazione. Infatti quando si ha una modesta variazione della media, siccome ogni punto tracciato si basa su un solo sottogruppo, e tende quindi ad avere una varianza notevole, serve un elevato numero di osservazioni per rendersi conto di quello che è accaduto. Un modo per ovviare a tale debolezza consiste nel consentire che i punti tracciati sulla carta dipendano non solo dal sottogruppo più recente, ma anche da alcuni altri. Tra i metodi che mettono in pratica questa idea e si sono dimostrati efficaci, ne trattiamo tre basati su (1) medie mobili, (2) medie mobili con pesi esponenziali (EWMA) e (3) carte di controllo a somme cumulate (CuSum).

13.6.1 Carte per le medie mobili

La carta di controllo a media mobile con finestra di lunghezza k si ottiene tracciando di volta in volta la media aritmetica dei k sottogruppi più recenti. Quindi, denotando con M_t la media mobile al tempo t, essa, sui tempi $t \geq k$ è definita come

$$M_t := \frac{\overline{X}_t + \overline{X}_{t-1} + \cdots + \overline{X}_{t-k+2} + \overline{X}_{t-k+1}}{k}, \qquad \text{se } t \geq k \tag{13.6.1}$$

dove $\overline{X}_i$ è la media campionaria delle osservazioni del sottogruppo i. I valori successivi $M_{t+1}, M_{t+2}, \ldots$ possono essere ottenuti facilmente, sfruttando il fatto che

$$kM_t = \overline{X}_t + \overline{X}_{t-1} + \cdots + \overline{X}_{t-k+1}$$

per cui vale anche

$$kM_{t+1} = \overline{X}_{t+1} + \overline{X}_t + \cdots + \overline{X}_{t-k+2}$$

e quindi, sottraendo membro a membro,

$$kM_{t+1} - kM_t = \overline{X}_{t+1} - \overline{X}_{t-k+1}$$

ovvero,

$$M_{t+1} = M_t + \frac{\overline{X}_{t+1} - \overline{X}_{t-k+1}}{k} \tag{13.6.2}$$

In altri termini, la media mobile all'istante $t+1$ è uguale quella all'istante t più $\frac{1}{k}$ della differenza tra il dato appena entrato e quello appena uscito dalla finestra della media mobile. Per valori di t inferiori a k, la media mobile è definita come media aritmetica dei primi t sottogruppi, ovvero:

$$M_t := \frac{\overline{X}_1 + \overline{X}_2 + \cdots + \overline{X}_t}{t}, \qquad \text{se } t < k \tag{13.6.3}$$

Supponiamo che quando il processo è in controllo statistico, i valori delle osservazioni provengano da una popolazione normale con media μ e varianza σ^2, e sia n l'ampiezza dei sottogruppi; i dati $\overline{X}_i$ sono allora normali con media μ e varianza σ^2/n. Se si calcola la media aritmetica di m di questi dati, si ottiene ancora una variabile aleatoria gaussiana con media μ, ma questa volta la varianza risulta σ^2/nm, e quindi, quando il processo è sotto controllo,

$$E[M_t] = \mu$$

$$\mathrm{Var}(M_t) = \begin{cases} \dfrac{\sigma^2}{nt} & \text{se } t < k \\ \dfrac{\sigma^2}{nk} & \text{se } t \geq k \end{cases} \tag{13.6.4}$$

Poiché una variabile aleatoria normale è quasi sempre meno distante di 3 deviazioni standard dalla media, i limiti di controllo superiore ed inferiore per M_t vengono definiti così:

$$\mathrm{UCL} := \begin{cases} \mu + 3\sigma/\sqrt{nt} & \text{se } t < k \\ \mu + 3\sigma/\sqrt{nk} & \text{se } t \geq k \end{cases}$$

$$\mathrm{LCL} := \begin{cases} \mu - 3\sigma/\sqrt{nt} & \text{se } t < k \\ \mu - 3\sigma/\sqrt{nk} & \text{se } t \geq k \end{cases} \tag{13.6.5}$$

Quindi, a parte le prime $k-1$ medie mobili, il processo viene dichiarato fuori controllo se una delle successive dista da μ più di $3\sigma/\sqrt{nk}$.

Esempio 13.6.1. Gli oggetti che escono da un certo processo produttivo hanno, sotto controllo, valori con distribuzione $\mathcal{N}(10, 4)$. I dati mostrati nella Tabella 13.3 sono le medie campionarie di 25 sottogruppi di dimensione 5, simulati però da una distribuzione di media 11 e varianza 4: rappresentano cioè dei possibili valori ottenuti dopo che il processo sia andato fuori controllo perché la media è passata da 10 a 11. Nella tabella sono state calcolate anche le medie mobili basate su $k = 8$ dati, e i limiti di controllo per M_t. In particolare quelli validi per $t \geq 8$ sono 9.051 e 10.949.

Tabella 13.3 Dati dell'Esempio 13.6.1. Il simbolo * indica lo stato di fuori controllo

t	$\overline{X}_t$	M_t	LCL	UCL
1	9.617728	9.617728	7.316719	12.68328
2	10.25437	9.936049	8.102634	11.89737
3	9.876195	9.913098	8.450807	11.54919
4	10.79338	10.13317	8.658359	11.34164
5	10.60699	10.22793	8.8	11.2
6	10.48396	10.2706	8.904554	11.09545
7	13.33961	10.70903	8.95815	11.01419
8	9.462969	10.55328	9.051318	10.94868
9	10.14556	10.61926	9.051318	10.94868
10	11.66342	10.79539	⋮	⋮
*11	11.55484	11.00634		
*12	11.26203	11.06492		
*13	12.31473	11.27839		
*14	9.220009	11.1204		
15	11.25206	10.85945		
*16	10.48662	10.98741		
17	9.025091	10.84735		
18	9.693386	10.6011		
19	11.45989	10.58923		
20	12.44213	10.73674		
21	11.18981	10.59613		
22	11.56674	10.88947		
23	9.869849	10.71669		
24	12.11311	10.92		
*25	11.48656	11.22768		

Come il lettore può notare, la prima media mobile a cadere fuori da questi limi si ha all'istante 11, mentre le successive sono agli istanti 12, 13, 14, 16 e 25. È anch interessante notare che in questo caso la carta di controllo $\overline{X}$ avrebbe dichiarato processo fuori controllo già all'istante 7, perché $\overline{X}_7$ è molto grande. Comunqu

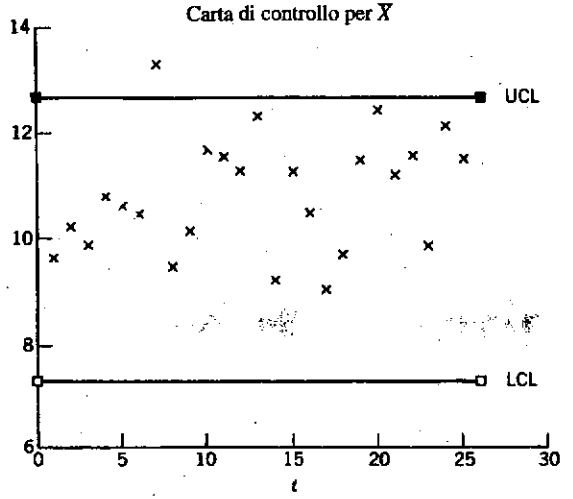


Figura 13.3

questo è l'unico punto dove quella carta avrebbe mostrato una mancanza di controllo (si veda la Figura 13.3). □

Vi è una relazione inversa tra la variazione della media percepibile e la lunghezza k della finestra utilizzata per la media mobile; più piccola è la variazione che si vuole poter rilevare, più grande deve essere preso k.

13.6.2 Carte per le medie mobili con pesi esponenziali (EWMA)

La media mobile impiegata nelle carte di controllo della Sezione 13.6.1 può essere vista, per ogni istante t, come una media pesata di *tutti* i dati precedenti, dando peso $\frac{1}{k}$ ai k valori più recenti, e 0 agli altri. Poiché questa strategia si rivela piuttosto efficace per individuare delle variazioni di media anche piccole, emerge la possibilità di impiegare con lo stesso fine altri tipi di medie pesate. Una scelta dei pesi utilizzata spesso si ottiene riducendo ad ogni passo di un fattore costante i pesi di tutte le osservazioni precedenti.

Assegnata una costante $0 < \alpha < 1$, definiamo ricorsivamente la quantità W_t:

$$W_t := \alpha\overline{X}_t + (1-\alpha)W_{t-1} \tag{13.6.6}$$

per $t > 0$, mentre per $t = 0$ si pone

$$W_0 := \mu \tag{13.6.7}$$

La sequenza dei valori $W_0, W_1, W_2, \ldots$ rappresenta la *media mobile con pesi esponenziali* (in inglese *exponentially weighted moving-average*, da cui l'acronimo EWMA) delle quantità $\mu, \overline{X}_1, \overline{X}_2, \overline{X}_3, \ldots$. Per comprendere il perché di tale nome è sufficiente sostituire ripetutamente la relazione data dall'Equazione (13.6.6), trovando che

$$\begin{aligned} W_t &= \alpha\overline{X}_t + (1-\alpha)\{\alpha\overline{X}_{t-1} + (1-\alpha)W_{t-2}\} \\ &= \alpha\overline{X}_t + \alpha(1-\alpha)\overline{X}_{t-1} + (1-\alpha)^2 W_{t-2} \\ &= \alpha\overline{X}_t + \alpha(1-\alpha)\overline{X}_{t-1} + (1-\alpha)^2\{\alpha\overline{X}_{t-2} + (1-\alpha)W_{t-3}\} \\ &= \alpha\overline{X}_t + \alpha(1-\alpha)\overline{X}_{t-1} + \alpha(1-\alpha)^2\overline{X}_{t-2} + (1-\alpha)^3 W_{t-3} \\ &\ldots \\ &= \sum_{i=0}^{t-1} \alpha(1-\alpha)^i \overline{X}_{t-i} + (1-\alpha)^t \mu \end{aligned} \tag{13.6.8}$$

dove si è usato il fatto che $W_0 = \mu$. Dall'Equazione (13.6.8) si deduce che W_t è una media pesata dei tutti i dati fino al tempo t; il più recente di essi ha peso α, e i precedenti hanno pesi via via minori, ognuno ridotto rispetto al precedente di un fattore $1-\alpha$. L'ultimo termine è μ, che ha peso $(1-\alpha)^t$. I pesi successivi assegnati ai valori dei sottogruppi sempre meno recenti possono essere scritti come

$$\alpha(1-\alpha)^{i-1} = \overline{\alpha}e^{-\beta i}$$

dove si è posto

$$\overline{\alpha} := \frac{\alpha}{1-\alpha}, \qquad \beta := -\log(1-\alpha)$$

da cui l'espressione "pesi esponenziali" (si veda la Figura 13.4).

Minore è il valore di α, più simili saranno i pesi assegnati ai vari dati. Ad esempio, se $\alpha = 0.1$, il primo peso è 0.1, e quelli successivi vanno moltiplicati per un fattore 0.9, per cui risultano 0.9, 0.81, 0.73, 0.66, 0.59, e così via. D'altra parte, con $\alpha = 0.4$ i pesi che si ottengono sono 0.4, 0.24, 0.144, 0.087, 0.052, che decrescono molto più velocemente.

Nell'ipotesi che il processo sia in controllo statistico, calcoliamo ora media e varianza di W_t. Le medie campionarie $\overline{X}_i$ sono variabili aleatorie normali indipendenti, di media μ e varianza σ^2/n. Sfruttando l'Equazione (13.6.8) otteniamo che

$$\begin{aligned} E[W_t] &= \mu\{\alpha + \alpha(1-\alpha) + \alpha(1-\alpha)^2 + \cdots + \alpha(1-\alpha)^{t-1}\} + \mu(1-\alpha)^t \\ &= \mu\alpha\frac{1-(1-\alpha)^t}{1-(1-\alpha)} + \mu(1-\alpha)^t = \mu \end{aligned}$$

$$\mathrm{Var}(W_t) = \frac{\sigma^2}{n}\{\alpha^2 + \alpha^2(1-\alpha)^2 + \alpha^2(1-\alpha)^4 + \cdots + \alpha^2(1-\alpha)^{2t-2}\}$$

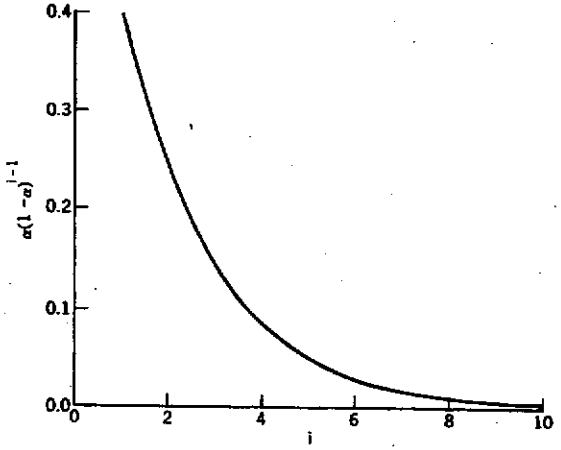


Figura 13.4 Grafico di $\alpha(1-\alpha)^{i-1}$, in funzione di i, con $\alpha = 0.4$.

$$= \frac{\sigma^2}{n}\alpha^2\{1+\beta+\beta^2+\cdots+\beta^{t-1}\} \qquad \text{ponendo } \beta := (1-\alpha)^2$$
$$= \frac{\sigma^2\alpha^2}{n}\cdot\frac{1-(1-\alpha)^{2t}}{1-(1-\alpha)^2}$$
$$= \frac{\sigma^2\alpha}{n}\cdot\frac{1-(1-\alpha)^{2t}}{2-\alpha}$$

Perciò per t grande, se il processo è rimasto sotto controllo per tutto il tempo,

$$E[W_t] = \mu$$
$$\text{Var}(W_t) \approx \frac{\sigma^2\alpha}{n(2-\alpha)} \qquad \text{visto che } (1-\alpha)^{2t} \approx 0 \tag{13.6.9}$$

I limiti di controllo asintotici per W_t sono dati da

$$\text{UCL} := \mu + 3\sigma\sqrt{\frac{\alpha}{n(2-\alpha)}}$$
$$\text{LCL} := \mu - 3\sigma\sqrt{\frac{\alpha}{n(2-\alpha)}} \tag{13.6.10}$$

Si noti che tali limiti di controllo coincidono con quelli della carta a media mobile basata su una finestra di lunghezza k, se vale la condizione

$$\frac{3\sigma}{\sqrt{nk}} = 3\sigma\sqrt{\frac{\alpha}{n(2-\alpha)}}$$

o equivalentemente se

$$k = \frac{2-\alpha}{\alpha} \qquad \text{oppure} \qquad \alpha = \frac{2}{k+1}$$

Esempio 13.6.2. Presso un laboratorio per la riparazione di elettrodomestici, ogni volta che un tecnico viene inviato a fare un intervento a domicilio, telefona in sede alla fine del lavoro, e il tempo trascorso viene annotato. I dati storici mostrano che il tempo che passa dall'uscita del tecnico alla telefonata, è una variabile aleatoria normale con media di 62 minuti e deviazione standard di 24. Per monitorare eventuali variazioni nella distribuzione, il laboratorio traccia una carta di controllo a media mobile con pesi esponenziali, usando come dati le medie di gruppi di 4 osservazioni successive, con un fattore di peso $\alpha = 0.25$. Il valore della carta in un dato momento è 60, e le medie dei 16 sottogruppi successivi sono:

48 52 70 62 57 81 56 59 77 82 78 80 74 82 68 84

Cosa si può concludere?

Ad iniziare da $W_0 = 60$, i valori successivi $W_1, W_2, \ldots, W_{16}$ possono essere ottenuti dalla formula

$$W_t = 0.25\overline{X}_t + 0.75W_{t-1}$$

ottenendo

$$W_1 = 0.25 \times 48 + 0.75 \times 60 = 57$$
$$W_2 = 0.25 \times 52 + 0.75 \times 57 = 55.75$$
$$W_3 = 0.25 \times 70 + 0.75 \times 55.75 \approx 59.31$$
$$W_4 = 0.25 \times 62 + 0.75 \times 59.31 \approx 59.98$$
$$W_5 = 0.25 \times 57 + 0.75 \times 59.98 \approx 59.24$$
$$W_6 = 0.25 \times 81 + 0.75 \times 59.24 \approx 68.68$$

e così via. I valori successivi, da W_7 a W_{16}, risultano

62.51 61.63 65.47 69.61 71.70 73.78 73.83 75.88 73.91 76.43

Visto che

$$3\sigma\sqrt{\frac{\alpha}{n(2-\alpha)}} = 3\sqrt{\frac{0.25}{1.75}}\frac{24}{\sqrt{4}} \approx 13.61$$

si trovano i limiti di controllo:

$$\text{LCL} = 62 - 13.61 \approx 48.39$$
$$\text{UCL} = 62 + 13.61 \approx 75.61$$

Quindi la carta EWMA dichiara il sistema fuori controllo in corrispondenza di W_{14}, come anche di W_{16}. È interessante notare che in questo caso, poiché la deviazione standard dei sottogruppi è $\sigma/\sqrt{n} = 12$, nessun dato dista da $\mu = 62$ di più di 2 deviazioni standard, e quindi la carta $\overline{X}$ non avrebbe dichiarato il sistema fuori controllo. □

Esempio 13.6.3. Consideriamo i dati dell'Esempio 13.6.1, ma usiamo questa volta una carta di controllo basata sulle medie mobili con pesi esponenziali con $\alpha = 2/9$. Si ottiene la successione di valori nella tabella qui sotto, che permette di dichiarare il processo fuori controllo già per $t = 7$, infatti i limiti di controllo asintotici sono (si veda anche la Figura 13.5),

$$\text{LCL} \approx 9.051$$
$$\text{UCL} \approx 10.949 \quad \square$$

t	$\overline{X}_t$	W_t	t	$\overline{X}_t$	W_t
1	9.617728	9.915051	14	9.220009	10.84522
2	10.25437	9.990456	15	11.25206	10.93563
3	9.867195	9.963064	16	10.48662	10.83585
4	10.79338	10.14758	17	9.025091	10.43346
5	10.60699	10.24967	18	9.693386	10.269
6	10.48396	10.30174	19	11.45989	10.53364
*7	13.33961	10.97682	*20	12.44213	10.95775
8	9.462969	10.64041	*21	11.18981	11.00932
9	10.14556	10.53044	*22	11.56674	11.13319
10	11.66342	10.78221	23	9.869849	10.85245
*11	11.55484	10.95391	*24	12.11311	11.13259
*12	11.26203	11.02238	*25	11.48656	11.21125
*13	12.31473	11.30957			

13.6.3 Carte di controllo per le somme cumulate

Quando è importante distinguere variazioni non molto grandi della media, la principale alternativa alle carte basate sulle medie mobili, sono quelle basate sulle *somme cumulate* (*cumulative sum*), spesso abbreviate in "carte CuSum".

Supponiamo come in precedenza che $\overline{X}_1, \overline{X}_2, \ldots$ denotino le medie campionarie di sottogruppi successivi di n elementi, e ammettiamo che quando il processo è in controllo statistico, esse abbiano distribuzione normale con media μ e deviazione standard $\sigma/\sqrt{n}$. Ci concentriamo inizialmente sull'evidenziare soltanto un eventuale incremento della media del processo (la carta di controllo che otterremo sarà

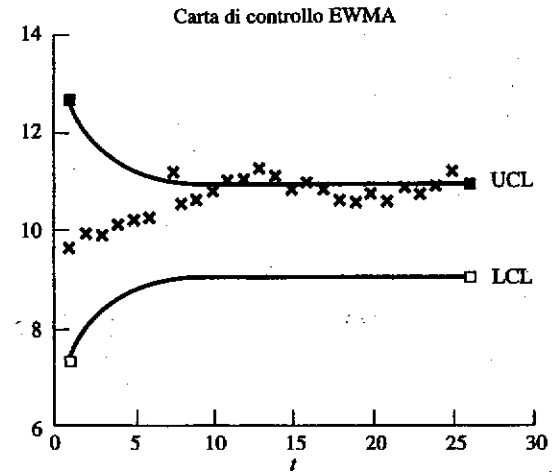


Figura 13.5

unilaterale). Scelte due costanti positive d e B, definiamo, per $j = 1, 2, 3, \ldots$,

$$Y_j := \overline{X}_j - \mu - d\sigma/\sqrt{n} \tag{13.6.11}$$

e

$$\begin{aligned} S_0 &:= 0 \\ S_{j+1} &:= \max\{S_j + Y_{j+1}, 0\}, \qquad j \geq 0 \end{aligned} \tag{13.6.12}$$

La carta di controllo CuSum di parametri d e B si ottiene tracciando i valori successivi di S_j, e dichiarando che la media del processo è aumentata, la prima volta che

$$S_j > B\sigma/\sqrt{n}$$

Per comprendere la logica che sta dietro queste definizioni, immaginiamo di volta in volta di tracciare non S_j, ma la somma di tutte le Y_i fino all'istante attuale; denotiamo tale somma con P_j:

$$P_j := \sum_{i=1}^{j} Y_i$$

è anche possibile dare una definizione ricorsiva, infatti

$$\begin{aligned} P_0 &:= 0 \\ P_{j+1} &:= P_j + Y_{j+1}, \qquad j \geq 0 \end{aligned}$$

Quando il processo è in controllo statistico, la media di ciascuna delle $\overline{X}_i$ è pari a μ e di conseguenza quella delle Y_i è negativa. Si ha infatti,

$$E[Y_i] = -d\sigma/\sqrt{n} < 0$$

Ci si deve quindi aspettare che la somma di un certo numero delle Y_i sia negativa con elevata probabilità (per la legge dei grandi numeri). Perciò se il valore di P_j fosse fortemente positivo per qualche j – diciamo maggiore di $B\sigma\sqrt{n}$ – questo sarebbe una ragionevole indicazione che la media è aumentata, portando il processo fuori controllo. Tuttavia, se un tale aumento della media si verificasse solo dopo lungo tempo, P_j avrebbe a quel punto un valore negativo molto grande (essendo la somma di parecchie variabili aleatorie indipendenti di media negativa), e quindi si renderebbe necessario molto tempo perché essa arrivi a superare $B\sigma/\sqrt{n}$. Proprio per evitare che la somma divenga troppo negativa quando il processo è in controllo statistico, la carta di controllo delle somme cumulate adotta il semplice espediente di resettare il suo valore a zero, non appena esso diventi negativo. La quantità S_j rappresenta infatti la sommatoria di tutte le Y_i fino al tempo j, con la correzione che ogni volta che essa diventa negativa viene azzerata.

Esempio 13.6.4. Supponiamo che le medie campionarie dei sottogruppi di osservazioni relative ad un processo produttivo, abbiano valore atteso $\mu = 30$ e deviazione standard $\sigma/\sqrt{n} = 8$; consideriamo la carta di controllo delle somme cumulate con $d = 0.5$ e $B = 5$. Se le medie dei primi 8 sottogruppi sono state

$$29 \quad 33 \quad 35 \quad 42 \quad 36 \quad 44 \quad 43 \quad 45$$

allora i valori corrispondenti delle statistiche $Y_j = \overline{X}_j - 30 - 4 = \overline{X}_j - 34$ sono

$$Y_1 = -5 \quad Y_2 = -1 \quad Y_3 = 1 \quad Y_4 = 8 \quad Y_5 = 2 \quad Y_6 = 10 \quad Y_7 = 9 \quad Y_8 = 11$$

e quindi

$$\begin{aligned} S_1 &= \max\{-5, 0\} = 0 & S_2 &= \max\{-1, 0\} = 0 \\ S_3 &= \max\{1, 0\} = 1 & S_4 &= \max\{9, 0\} = 9 \\ S_5 &= \max\{11, 0\} = 11 & S_6 &= \max\{21, 0\} = 21 \\ S_7 &= \max\{30, 0\} = 30 & S_8 &= \max\{41, 0\} = 41 \end{aligned}$$

Il limite di controllo è

$$B\sigma/\sqrt{n} = 5 \times 8 = 40$$

La carta di controllo stabilisce quindi che la media è aumentata; tale conclusione viene raggiunta dopo l'osservazione dell'ottavo sottogruppo. □

Nel caso si desiderino rilevare variazioni della media sia positive sia negative, si possono impiegare simultaneamente due carte di controllo di questo tipo. Si noti infatti che una diminuzione di $E[X_i]$ equivale ad un aumento di $E[-X_i]$; per questo, applicando una carta di controllo CuSum ai dati dei sottogruppi *cambiati di segno*, si possono mettere in evidenza le eventuali diminuzioni della media. In concreto, per dei valori fissati di d e B, non ci dobbiamo più limitare a tracciare i valori di S_j, ma dobbiamo anche calcolare le quantità W_i, date da

$$W_i := -\overline{X}_i - (-\mu) - d\sigma/\sqrt{n} = \mu - \overline{X}_i - d\sigma/\sqrt{n} \tag{13.6.13}$$

che sono l'analogo delle Y_i, e poi le somme cumulate T_j, definite da

$$\begin{aligned} T_0 &:= 0 \\ T_{j+1} &:= \max\{T_j + W_{j+1}, 0\}, \qquad j \geq 0 \end{aligned} \tag{13.6.14}$$

che sono analoghi alle S_j. Il processo viene dichiarato fuori controllo la prima volta che S_j o T_j superano $B\sigma/\sqrt{n}$.

Riassumendo, per realizzare una carta di controllo CuSum sono necessari i passi seguenti: **(1)** scegliere due costanti positive d e B; **(2)** determinare le quantità S_j e T_j per i differenti valori di j, utilizzando le medie campionarie dei sottogruppi e le Equazioni (13.6.12) e (13.6.14); **(3)** dichiarare il sistema fuori controllo non appena uno di questi valori superi il limite di controllo $B\sigma/\sqrt{n}$.

Tre scelte comuni per le costanti di definizione sono

$$\begin{aligned} d &= 0.25, & B &= 8.00 \\ d &= 0.50, & B &= 4.77 \\ d &= 1.00, & B &= 2.49 \end{aligned} \tag{13.6.15}$$

Ciascuna di queste scelte porta ad un criterio di controllo che ha circa lo stesso tasso di falsi allarmi (lo 0.27%) di una carta di controllo $\overline{X}$ con limiti di controllo a $\mu \pm 3\sigma/\sqrt{n}$. Si noti anche che in generale, più piccola è la variazione della media che si vuole potere rilevare, più piccolo dovrà essere il valore scelto per d.

Problemi

1. Assumi che una caratteristica dei pezzi che produciamo abbia distribuzione normale con media 35 e deviazione standard 3. Per sorvegliare questo processo si estraggono, come sottogruppi, dei campioni di 5 osservazioni. Se quelle che seguono sono (nell'ordine) le medie campionarie dei primi 20 sottogruppi, si può dire che il processo sia in controllo statistico?

34.0 31.6 30.8 33.0 35.0 32.2 33.0 32.6 33.8 35.8
35.8 35.8 34.0 35.0 33.8 31.6 33.0 33.2 31.8 35.6

2. Supponi che un processo sia in controllo statistico con $\mu = 14$ e $\sigma = 2$. Si impiega una carta di controllo $\overline{X}$ basata su sottogruppi razionali di 5 elementi. Se la media subisce una variazione di 2.2 unità, qual è la probabilità che il sottogruppo successivo abbia una media campionaria fuori dai limiti di controllo? In media, quanti sottogruppi occorre valutare prima di dichiarare il processo fuori controllo?

3. Sia Y una variabile aleatoria con distribuzione chi-quadro con $n-1$ gradi di libertà. Dimostra che

$$E[\sqrt{Y}] = \frac{\Gamma(\frac{n}{2})\sqrt{2}}{\Gamma(\frac{n-1}{2})}$$

(*Suggerimento:* Verifica i seguenti passaggi:

$$\begin{aligned} E[\sqrt{Y}] &= \int_0^\infty \sqrt{y} f_{\chi^2_{n-1}}(y)\, dy \\ &= \int_0^\infty \sqrt{y} \frac{e^{-y/2} \cdot y^{(n-1)/2} \cdot y^{-1}}{2^{(n-1)/2} \cdot \Gamma[(n-1)/2]}\, dy \\ &= \int_0^\infty \frac{e^{-y/2} \cdot y^{n/2-1}}{2^{(n-1)/2} \cdot \Gamma[(n-1)/2]}\, dy \end{aligned}$$

Quindi esegui il cambio di varibili $x := y/2$.)

4. Da un processo di produzione vengono estratti a intervalli regolari dei campioni di 5 pezzi, per i quali si calcolano media e deviazione standard campionarie. Le somme di queste statistiche per i primi 25 campioni risultano

$$\sum_{i=1}^{25} \overline{X}_i = 357.2 \qquad \sum_{i=1}^{25} S_i = 4.88$$

   (a) Supponendo lo stato di controllo, determina i limiti di controllo per una carta $\overline{X}$.

   (b) Assumendo che il processo rimanga in controllo statistico e approssimando i parametri veri con quelli stimati al punto **(a)**, che percentuale dei pezzi prodotti rientrerà nelle specifiche di accettabilità, che sono stabilite in 14.3 ± 0.45?

5. Completa l'Esempio 13.3.1, ricalcolando i limiti di controllo per le carte $\overline{X}$ e S, dopo avere escluso i dati anomali.

6. Nel Problema 4, determina i limiti di controllo per una carta S.

7. Quelli che seguono sono i valori di $\overline{X}$ e di S per 20 sottogruppi di ampiezza 5.

Sottogruppo	1	2	3	4	5	6	7	8	9	10
$\overline{X}$	33.8	37.2	40.4	39.3	41.1	40.4	35.0	36.1	38.2	32.4
S	5.1	5.4	6.1	5.5	5.2	4.8	5.0	4.1	7.3	6.6

Sottogruppo	11	12	13	14	15	16	17	18	19	20
$\overline{X}$	29.7	31.6	38.4	40.2	35.6	36.4	37.2	31.3	33.6	36.7
S	5.1	5.3	5.8	6.4	4.8	4.6	6.1	5.7	5.5	4.2

   (a) Determina i limiti di controllo per una carta $\overline{X}$.

   (b) Quali sono i limiti di controllo S?

   (c) Ti sembra che il processo sia rimasto per tutto il tempo in controllo statistico?

   (d) Se la tua risposta al punto (c) è negativa, suggerisci quali valori dei limiti di controllo andrebbero usati per i sottogruppi successivi.

   (e) Se i limiti di tollerabilità dei pezzi prodotti sono 35 ± 10, quale stimi sia la percentuale degli oggetti accettabili che escono dalla linea di produzione?

8. Presso una azienda si mantengono carte di controllo per $\overline{X}$ e S per la sollecitazione di taglio dei punti di saldatura. Dopo 30 sottogruppi di ampiezza 4, i totali delle statistiche campionarie sono $\sum \overline{X}_i = 12\,660$ e $\sum S_i = 500$. Assumi che il processo sia in controllo statistico.

   (a) Quali sono i limiti di controllo $\overline{X}$?

   (b) Determina i limiti di controllo per una carta S.

   (c) Stima la deviazione standard del processo.

   (d) Se la sollecitazione minima accettabile è di 400 libbre, che percentuale delle saldature non soddisfa questa richiesta?

9. Nel redigere le carte di controllo per $\overline{X}$ e S per i resistori prodotti in un impianto, si usano sottogruppi razionali di 4 osservazioni. Avendo raccolto i dati di 20 di essi, si trova che $\sum \overline{X}_i = 8\,620$ e $\sum S_i = 450$.

   (a) Calcola i valori dei limiti di controllo per le carte $\overline{X}$ e S.

   (b) Stima il valore di σ nell'ipotesi che il processo sia sempre rimasto in controllo statistico.

   (c) Se le specifiche commerciali richiedono che i valori di resistenza siano compresi nell'intervallo 430 ± 30, che conclusioni puoi trarre sulla capacità di questo processo produttivo di rispettare le specifiche?

   (d) Se la media μ aumenta di 60, qual è la probabilità che la media campionaria di un sottogruppo cada al di fuori dei limiti di controllo?

10. I dati seguenti si riferiscono alla differenza – in millesi di pollice – tra il diametro effettivo e quello nominale di 15 campioni di cuscinetti a sfera.

Sottogruppo	Osservazioni				
1	2.5	0.5	2.0	−1.2	1.4
2	0.2	0.3	0.5	1.1	1.5
3	1.5	1.3	1.2	−1.0	0.7
4	0.2	0.5	−2.0	0.0	−1.3
5	−0.2	0.1	0.3	−0.6	0.5
6	1.1	−0.5	0.6	0.5	0.2
7	1.1	−1.0	−1.2	1.3	0.1
8	0.2	−1.5	−0.5	1.5	0.3
9	−2.0	−1.5	1.6	1.4	0.1
10	−0.5	3.2	−0.1	−1.0	−1.5
11	0.1	1.5	−0.2	0.3	2.1
12	0.0	−2.0	−0.5	0.6	−0.5
13	−1.0	−0.5	−0.5	−1.0	0.2
14	0.5	1.3	−1.2	−0.5	−2.7
15	1.1	0.8	1.5	−1.5	1.2

(a) Stabilisci i limiti di controllo per le carte $\overline{X}$ e S.

(b) Ti sembra che il processo sia rimasto in controllo statistico per tutta la durata del campionamento?

(c) Se la risposta al punto **(b)** è negativa, trova dei limiti dei controllo più precisi.

11. Dei campioni di 6 oggetti vengono estratti ad intervalli regolari da un processo manifatturiero. Si misura una caratteristica che si sa avere distribuzione normale, e si calcolano le statistiche $\overline{X}$ e S di ogni campione. Dopo l'esame di 50 sottogruppi si ottiene che

$$\sum_{i=1}^{50} \overline{X}_i = 970 \quad \text{e} \quad \sum_{i=1}^{50} S_i = 85$$

(a) Calcola i limiti di controllo per le carte $\overline{X}$ e S. Puoi assumere che tutti i punti di entrambe le carte cadano all'interno dei limiti trovati.

(b) Se i limiti di accettabilità specificati sono 19 ± 4.0, quali sono le tue conclusioni sulla capacità di questo processo di produrre oggetti conformi alle richieste?

12. I dati che seguono rappresentano il numero di assemblaggi difettosi di cuscinetto e guarnizione, su campioni di ampiezza 100.

5 2 1 5 9 4 3 3 2 5
4 10 0 8 3 6 2 1 6 10

Si può dire che il processo sia rimasto sotto controllo durante tutto il campionamento? In caso contrario trova i limiti di controllo corretti se possibile.

13. I dati seguenti rappresentano il risultato di un esame approfondito di tutti i personal computer prodotti in un certo laboratorio negli ultimi 12 giorni:

Giorno	1	2	3	4	5	6	7	8	9	10	11	12
Unità	80	110	90	80	100	90	80	70	80	90	90	110
Difettose	5	7	4	9	12	10	4	3	5	6	5	7

Si può dire che il processo sia rimasto in controllo statistico tutto il tempo? Determina i limiti di controllo per la produzione futura.

14. Supponiamo che quando un certo processo è sotto controllo, la probabilità che un pezzo sia difettoso sia di 0.04; supponiamo inoltre che si testino quotidianamente campioni di ampiezza 500. Qual è la probabilità che, nel caso il tasso di difettosi salisse a 0.08, la carta di controllo rilevi questa variazione già al campione successivo?

15. I dati qui sotto rappresentano il numero di integrati difettosi prodotti negli ultimi 15 giorni in uno stabilimento:

121 133 98 85 101 78 66 82 90 78 85 81 100 75 89

Si può concludere che il processo sia rimasto in controllo statistico per tutto il periodo? Che limiti di controllo consiglieresti per la produzione futura?

16. Si è proceduto a contare il numero di difetti superficiali riscontrabili su 25 lastre di acciaio; i valori trovati sono stati:

2 3 4 3 1 2 5 0 2 5 1 7 8
10 2 2 6 5 4 6 3 7 0 2 4

Realizza una carta di controllo. Il processo di produzione di queste lastre risulta in stato di controllo?

17. La tabella che segue riporta le medie campionarie di 25 sottogruppi razionali, unitamente alle corrispondenti medie mobili basate su una finestra di 5 dati. Le osservazioni sono state generate da un processo che, quando è in controllo, produce pezzi con distribuzione normale di media 30 e varianza 40; i sottogruppi sono composti da 4 elementi ciascuno. Ti risulta che il processo sia rimasto in controllo per tutto il tempo?

$\overline{X}_t$	M_t	$\overline{X}_t$	M_t
35.62938	35.62938	35.80945	32.34106
39.13018	37.37978	30.9136	33.1748
29.45974	34.73976	30.54829	32.47771
32.5872	34.20162	36.39414	33.17019
30.06041	33.37338	27.62703	32.2585
26.54353	31.55621	34.02624	31.90186
37.75199	31.28057	27.81629	31.2824
26.88128	30.76488	26.99926	30.57259
32.4807	30.74358	32.44703	29.78317
26.7449	30.08048	38.53433	31.96463
34.03377	31.57853	28.53698	30.86678
32.93174	30.61448	28.65725	31.03497
32.18547	31.67531		

18. I dati riportati qui di seguito sono le medie campionarie dei sottogruppi, e le medie mobili con finestra di lunghezza $k = 8$, per dei sottogruppi di 4 osservazioni, che in controllo statistico dovrebbero avere media 50 e varianza 5. Cosa concludi?

$\overline{X}_t$	M_t	$\overline{X}_t$	M_t
50.79806	50.79806	53.08497	52.2036
46.21413	48.50609	55.02968	52.79759
51.85793	49.62337	54.25338	52.85237
50.27771	49.78696	50.48405	52.82834
53.81512	50.59259	50.34928	52.69814
50.67635	50.60655	50.86896	52.6002
51.39083	50.71859	52.03695	52.58531
51.65246	50.83533	53.23255	52.41748
52.15607	51.00508	48.12588	51.79759
54.57523	52.05022	52.23154	51.44783

19. Affronta nuovamente il Problema 17, impiegando una carta EWMA, con $\alpha = 1/3$.

20. Analizza i dati del Problema 18, usando una carta EWMA con $\alpha = 2/9$.

21. Spiega come mai impiegando carte per le medie mobili con finestra di k sottogruppi, si devono usare dei limiti di controllo differenti per le prime $k - 1$ medie mobili, mentre le carte per le medie mobili con pesi esponenziali consentono di usare sempre gli stessi limiti. (*Suggerimento:* Mostra che $\mathrm{Var}(M_t)$ è decrescente in t, mentre $\mathrm{Var}(W_t)$ è crescente, e spiega perché questo fatto giustifica la differenza.)

22. Ripeti il Problema 17, questa volta usando una carta delle somme cumulate, **(a)** con $d = 0.25$ e $B = 8$; **(b)** con $d = 0.5$ e $B = 4.77$.

23. Ripeti il Problema 18 usando una carta CuSum, con $d = 1$ e $B = 2.49$.

14 Affidabilità dei sistemi

Contenuto

14.1 Introduzione

In questo capitolo prendiamo in considerazione una popolazione di oggetti i cui tempi di vita sono variabili aleatorie indipendenti con una distribuzione comune. Tale distribuzione si suppone nota a meno di un parametro incognito; il nostro obiettivo sarà di usare tutti i dati a disposizione per stimare tale parametro.

Nella Sezione 14.2 viene introdotto il concetto di funzione di rischio (o intensità di rotture), uno strumento ingegneristico che permette di esprimere la distribuzione dei tempi di vita in maniera più significativa delle funzioni di ripartizione e di densità. Nella Sezione 14.3 l'attenzione si concentra sulla legge esponenziale, e viene illustrato come ottenere stime puntuali, intervalli di confidenza e stime bayesiane della media, sotto una serie di schemi di prova. La Sezione 14.4 sviluppa un test per verificare l'ipotesi che due popolazioni esponenziali indipendenti abbiano lo stesso parametro. Nella Sezione 14.5, infine, viene presentata la distribuzione di Weibull, con due approcci che permettono di stimarne i parametri.

14.2 Funzione di intensità di rotture

Consideriamo una variabile aleatoria X, continua e positiva, che rappresenta il tempo di vita di un certo tipo di oggetti. Sia F la sua funzione di ripartizione, e f la densità

di probabilità. La *funzione di rischio*, o *intensità di rotture* è la funzione λ definita da

$$\lambda(t) := \frac{f(t)}{1 - F(t)} \tag{14.2.1}$$

È importante capire il significato pratico di $\lambda(t)$. Supponiamo di studiare un elemento che è soggetto a rotture, e che funziona ininterrottamente da un tempo t; vogliamo sapere la probabilità che si guasti nell'immediato futuro, nel prossimo intervallo di tempo di durata dt. Quella che cerchiamo è una probabilità condizionata, che è espressa da

$$\begin{aligned} P(X \in (t, t+dt) | X > t) &:= \frac{P(X \in (t, t+dt), X > t)}{P(X > t)} \\ &= \frac{P(X \in (t, t+dt))}{1 - F(t)} \\ &\approx \frac{f(t)\,dt}{1 - F(t)} =: \lambda(t)\,dt \end{aligned}$$

Perciò $\lambda(t)$ rappresenta la densità condizionale di probabilità, che un oggetto di età t si guasti "nel prossimo istante".

Nel caso particolare in cui la distribuzione dei tempi di vita sia esponenziale, per la proprietà di assenza di memoria (si veda il Capitolo 5, a pagina 179), la distribuzione della vita residua di un oggetto di età t è identica a quella di un oggetto nuovo. L'intensità di rotture deve quindi avere un valore costante, come si può verificare facilmente:

$$\begin{aligned} \lambda(t) &:= \frac{f(t)}{1 - F(t)} \\ &= \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda \end{aligned}$$

Il valore trovato è l'*intensità* della distribuzione esponenziale, che coincide quindi con la sua intensità di rotture.

Non è difficile dimostrare che la funzione λ determina univocamente la F, e quindi la distribuzione della variabile aleatoria. Infatti, per definizione:

$$\begin{aligned} \lambda(s) &:= \frac{f(s)}{1 - F(s)} \\ &= \frac{F'(s)}{1 - F(s)} \\ &= -\frac{d}{ds} \log\bigl(1 - F(s)\bigr) \end{aligned} \tag{14.2.2}$$

Integrando entrambi i membri tra 0 e t si ottiene che

$$\begin{aligned} \int_0^t \lambda(s)\,ds &= -\log\bigl(1 - F(t)\bigr) + \log\bigl(1 - F(0)\bigr) \\ &= -\log\bigl(1 - F(t)\bigr) \qquad \text{perchè } F(0) = 0 \end{aligned}$$

e quindi si ha che

$$1 - F(t) = \exp\left\{ -\int_0^t \lambda(s)\,ds \right\} \tag{14.2.3}$$

Ciò significa che la funzione di ripartizione di una variabile aleatoria continua può essere specificata tramite la corrispondente funzione di intensità di rotture. Ad esempio, se sappiamo che l'intensità di rotture è una funzione lineare di t, come

$$\lambda(t) = a + bt$$

allora la funzione di ripartizione è necessariamente data da

$$F(t) = 1 - e^{-at - bt^2/2}$$

e derivando troviamo che la densità è fornita dalla seguente espressione:

$$f(t) = (a + bt) e^{-at - bt^2/2}$$

Nel caso che nell'esempio qui sopra si prenda $a = 0$, si ottiene la cosiddetta distribuzione di probabilità di *Rayleigh*.

Esempio 14.2.1. Si sente dire spesso che il tasso di mortalità di un fumatore è, ad ogni età, il doppio di quello di un non fumatore. Cosa significa? Vuole dire ad esempio che la probabilità di sopravvivere negli anni successivi per un non fumatore è il doppio di quella di un fumatore della stessa età?

Denotiamo con $\lambda_f(t)$ e con $\lambda_n(t)$ le intensità (o tassi) di mortalità all'età t, per un fumatore e per un non fumatore. Stiamo ipotizzando che valga la relazione:

$$\lambda_f(t) = 2\lambda_n(t)$$

La probabilità che un non fumatore di età a sopravviva fino all'età $b > a$ è data da

$$\begin{aligned}
&P(\text{Non fumatore di età } a \text{ arriva all'età } b)\\
&\quad = P(\text{Non fumatore vive fino a } b \,|\, \text{È vissuto almeno fino ad } a)\\
&\quad = \frac{P(\text{Non fumatore vive almeno fino a } b)}{P(\text{Non fumatore vive almeno fino ad } a)}\\
&\quad = \frac{1 - F_{\text{n}}(b)}{1 - F_{\text{n}}(a)}\\
&\quad = \frac{\exp\{-\int_0^b \lambda_{\text{n}}(t)\,dt\}}{\exp\{-\int_0^a \lambda_{\text{n}}(t)\,dt\}} \qquad \text{per l'Equazione (14.2.3)}\\
&\quad = \exp\left\{-\int_a^b \lambda_{\text{n}}(t)\,dt\right\}
\end{aligned}$$

Lo stesso ragionamento applicato ad un fumatore porta al seguente risultato:

$$\begin{aligned}
&P(\text{Un fumatore di età } a \text{ arriva all'età } b)\\
&\quad = \exp\left\{-\int_a^b \lambda_{\text{f}}(t)\,dt\right\}\\
&\quad = \exp\left\{-2\int_a^b \lambda_{\text{n}}(t)\,dt\right\}\\
&\quad = \left[\exp\left\{-\int_a^b \lambda_{\text{n}}(t)\,dt\right\}\right]^2
\end{aligned}$$

Per cui affermare che il tasso di mortalità di chi fuma sia doppio porta alla conclusione che se si confrontano un fumatore e un non fumatore della stessa età, la probabilità che il primo sopravviva per un certo numero di anni è il *quadrato*, e non la metà, della probabilità corrispondente per il secondo. Ad esempio se misuriamo i tempi in anni, e se $\lambda_{\text{n}}(t)$ fosse circa uguale a $1/20$, per $50 \le t \le 60$, allora la probabilità che un non fumatore cinquantenne arrivi ai 60 anni sarebbe $e^{-0.5} \approx 0.607$, mentre la corrispondente probabilità per un fumatore sarebbe $e^{-1} \approx 0.368$. □

14.3 Il ruolo della distribuzione esponenziale

14.3.1 Prove simultanee – interruzione al fallimento r-esimo

In questa sezione ci occupiamo dell'esame simultaneo di un campione di n oggetti con tempi di vita esponenziali e indipendenti, con media incognita θ. Pensiamo di interrompere l'esperimento quando il numero di oggetti che si guastano raggiunge un numero fissato $r \le n$; ci domandiamo come si possa stimare θ. I dati a nostra disposizione sono gli r tempi di vita registrati, che denotiamo, nell'ordine con

$$x_1 \le x_2 \le \cdots \le x_r$$

unitamente ai numeri di identificazione degli oggetti guasti, $i_1, i_2, \ldots, i_r$; intendiamo con questo che, se gli n oggetti erano numerati in progressione, per $j = 1, 2, \ldots, r$, il numero i_j indica quale oggetto si è guastato per j-esimo, e precisamante all'istante x_j.

Se denotiamo con X_i il tempo di vita dell'oggetto i, dove $1 \le i \le n$, i dati precedenti possono anche essere riassunti tramite

$$\begin{gathered}
X_{i_1} = x_1,\ X_{i_2} = x_2,\ \ldots,\ X_{i_r} = x_r\\
\text{le restanti } n - j \text{ delle } X_j \text{ sono tutte maggiori di } x_r
\end{gathered} \tag{14.3.1}$$

La densità di probabilità delle X_{i_j} è

$$f_{X_{i_j}}(x_j) = \frac{1}{\theta} e^{-x_j/\theta}, \qquad j = 1, 2, \ldots, r$$

e quindi, grazie all'indipendenza, la densità congiunta delle X_{i_j}, $j = 1, 2, \ldots, r$ è data da

$$f_{X_{i_1},\ldots,X_{i_r}}(x_1, \ldots, x_r) = \prod_{j=1}^{r} \frac{1}{\theta} e^{-x_j/\theta}$$

Inoltre la probabilità che le restanti $n - r$ tra le X_j siano tutte maggiori di x_r è data, usando sempre l'indipendenza, da

$$P(X_j > x_r \text{ per } j \notin \{i_1, i_2, \ldots, i_r\}) = (e^{-x_r/\theta})^{n-r}$$

Di conseguenza la likelihood (o verosimiglianza) dei dati osservati, che viene denotata con $L(x_1, x_2, \ldots, x_r, i_1, i_2, \ldots, i_r|\theta)$, è data da

$$\begin{aligned}
&L(x_1, x_2, \ldots, x_r, i_1, i_2, \ldots, i_r|\theta)\\
&\quad = f_{X_{i_1},\ldots,X_{i_r}}(x_1, \ldots, x_r) P(X_j > x_r, j \notin \{i_1, i_2, \ldots, i_r\})\\
&\quad = \frac{1}{\theta^r} e^{-x_1/\theta} e^{-x_2/\theta} \cdots e^{-x_r/\theta} (e^{-x_r/\theta})^{n-r}\\
&\quad = \frac{1}{\theta^r} \exp\left\{-\frac{1}{\theta}\sum_{i=1}^{r} x_i - \frac{(n-r)x_r}{\theta}\right\}
\end{aligned} \tag{14.3.2}$$

Osservazione 14.3.1. La funzione di likelihood ottenuta con l'equazione precedente non è condizionata solo agli istanti delle rotture, $x_1, x_2, \ldots, x_r$, ma anche a quali sono gli r oggetti che si sono guastati e all'ordine $i_1, i_2, \ldots, i_r$ con cui ciò si è verificato.

Se fosse richiesta la verosimiglianza in funzione solo degli r tempi di rottura, visto che vi sono

$$n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}$$

diverse scelte ordinate per gli r oggetti, e visto che esse sono tutte equiprobabili, la densità di probabilità congiunta e quindi la funzione di likelihood risulterebbe, per $x_1 \le x_2 \le \cdots \le x_r$, pari a

$$f(x_1, x_2, \ldots, x_r) = \frac{n!}{(n-r)!}\frac{1}{\theta^r}\exp\left\{\frac{1}{\theta}\sum_{i=1}^{r} x_i - \frac{(n-r)x_r}{\theta}\right\}$$

Per ottenere lo stimatore di massima verosimiglianza di θ, calcoliamo e poniamo uguale a zero la derivata rispetto a θ del logaritmo di L.

$$\log L(x_1, x_2, \ldots, x_r, i_1, i_2, \ldots, i_r|\theta) = -r\log\theta - \frac{1}{\theta}\sum_{i=1}^{r} x_i - \frac{(n-r)x_r}{\theta}$$

$$\frac{\partial}{\partial\theta}\log L(x_1, x_2, \ldots, x_r, i_1, i_2, \ldots, i_r|\theta) = -\frac{r}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{r} x_i + \frac{(n-r)x_r}{\theta^2}$$

L'unica scelta di θ che annulla l'espressione precedente è

$$\frac{\sum_{i=1}^{r} x_i + (n-r)x_r}{r}$$

Se si denota con $X_{(i)}$ l'istante in cui si guasta l'i-esimo oggetto ($X_{(i)}$ viene detta *statistica di ordine* i), allora sostituendo nella formula precedente le statistiche $X_{(i)}$ alle loro realizzazioni x_i, si trova lo stimatore di massima verosimiglianza $\hat{\theta}$:

$$\hat{\theta} := \frac{\sum_{i=1}^{r} X_{(i)} + (n-r)X_{(r)}}{r} =: \frac{\tau}{r} \tag{14.3.3}$$

dove si è denotato con τ il numeratore dell'espressione precedente, che viene detto *total time on test statistic*[1]. Infatti quando l'esperimento viene interrotto, i primi r oggetti a guastarsi hanno vissuto per dei tempi $X_{(1)}, X_{(2)}, \ldots, X_{(r)}$ (e poi si sono rotti), mentre gli altri $n-r$ hanno vissuto per un tempo $X_{(r)}$ (fino alla conclusione dell'esperimento).

Per determinare gli intervalli di confidenza per θ, è necessario ottenere la distribuzione di τ. Ricordando che $X_{(i)}$ denota il tempo di vita dell'oggetto che si guasta

[1] In italiano sarebbe *statistica del tempo totale di funzionamento*, ma è più usata l'espressione inglese, anche nell'acronimo TTT. [*N.d.T.*]

per i-esimo, riscriviamo τ come somma delle statistiche Y_i, per $i = 1, 2, \ldots, r$, che indicano il tempo totale di funzionamento racchiuso tra la rottura dell'oggetto $(i-1)$-esimo e quella dell'i-esimo. Dal tempo 0 al tempo $X_{(1)}$, sono in funzione n oggetti, quindi il total time on test in questo intervallo è

$$Y_1 := nX_{(1)}$$

Nell'intervallo tra gli istanti $X_{(1)}$ e $X_{(2)}$ sono in funzione $n-1$ oggetti, quindi

$$Y_2 := (n-1)(X_{(2)} - X_{(1)})$$

Analogamente ma più in generale, ponendo $X_{(0)} := 0$,

$$Y_j := (n-j+1)(X_{(j)} - X_{(j-1)}), \qquad j = 1, 2, \ldots, r \tag{14.3.4}$$

E ovviamente vale

$$\tau = \sum_{j=1}^{r} Y_j \tag{14.3.5}$$

L'importanza di questa rappresentazione di τ risiede nel fatto che la distribuzione delle Y_j si ottiene facilmente. La statistica $X_{(1)}$, in quanto tempo di vita del primo oggetto che si guasta, è il minimo di n variabili aleatorie esponenziali i.i.d. di intensità $1/\theta$, e quindi è a sua volta esponenziale, ma con intensità n/θ (si veda la Proposizione 5.6.1 a pagina 181), perciò $Y_1 = nX_{(1)}$ ha distribuzione esponenziale con intensità $1/\theta$ e media θ. Nel momento in cui l'oggetto i_1 si guasta, ne restano $n-1$, che per la assenza di memoria della distribuzione esponenziale sono "come nuovi"; ciascuno di essi avrà un ulteriore tempo di vita che è una variabile aleatoria esponenziale di media θ, perciò il tempo che trascorre tra $X_{(1)}$ e $X_{(2)}$ è esponenziale di intensità $(n-1)/\theta$, e di conseguenza $Y_2 = (n-1)(X_{(2)} - X_{(1)})$ è esponenziale con media θ.

Proseguendo su questa linea si dimostra che le variabili aleatorie $Y_1, Y_2, \ldots, Y_r$ sono esponenziali indipendenti di media θ. Siccome il Corollario 5.7.2 afferma che la somma di variabili aleatorie esponenziali i.i.d. ha distribuzione gamma, otteniamo che τ è una gamma con parametri r e $1/\theta$.

Un metodo economico per determinare le probabilità relative alle variabili aleatorie di tipo gamma, consiste nel ricordare che una gamma di parametri r e $1/\theta$ è anche una chi-quadro con $2r$ gradi di libertà, moltiplicata per $\theta/2$ (si veda la Sezione 5.8.1.1, a partire da pagina 190), e infatti

$$\frac{2\tau}{\theta} \sim \chi^2_{2r} \tag{14.3.6}$$

Sfruttando questa relazione si vede subito che

$$P\left(\chi^2_{1-\frac{\alpha}{2},2r} < 2\tau/\theta < \chi^2_{\frac{\alpha}{2},2r}\right) = 1-\alpha$$

E quindi vi è un livello di confidenza $1-\alpha$ nell'affermare che

$$\theta \in \left(\frac{2\tau}{\chi^2_{\frac{\alpha}{2},2r}}, \quad \frac{2\tau}{\chi^2_{1-\frac{\alpha}{2},2r}} \right) \tag{14.3.7}$$

Gli intervalli di confidenza unilaterali si ottengono in maniera del tutto analoga.

Esempio 14.3.1. Un totale di 50 transistor vengono messi in funzione simultaneamente; l'esperimento si concluderà quando il 15-esimo di essi si guasta. Il total time on test che si ottiene è di 525 ore. Si trovi un intervallo di confidenza al 95% per la vita media di un componente di questo tipo. Si assuma che la distribuzione della popolazione sia esponenziale.

Dalla Tabella A.2 in Appendice si ottiene che

$$\chi^2_{0.025,30} \approx 46.98, \qquad \chi^2_{0.975,30} \approx 16.79$$

perciò, sostituendo i dati nell'Equazione (14.3.7), si può affermare con il 95% di confidenza che

$$\theta \in (22.35,\ 62.54) \quad \square$$

Dovendo verificare delle ipotesi su θ, l'Equazione (14.3.6) permette di calcolare facilmente il p-dei-dati. Supponiamo per esempio di volere confrontare l'ipotesi nulla

$$H_0 : \theta \geq \theta_0$$

con una alternativa unilaterale

$$H_1 : \theta < \theta_0$$

Ciò può essere ottenuto calcolando prima il valore v della statistica del test, che è $2\tau/\theta_0$, e poi determinando la probabilità che una chi-quadro con $2r$ gradi di libertà assuma un valore piccolo come v. Tale grandezza coincide con il p-dei-dati di questo test statistico, in quanto rappresenta la probabilità che con H_0 soddisfatta, si osservi un valore estremo come v. L'ipotesi nulla va poi rifiutata a tutti i livelli di significatività superiori al p-dei-dati.

Esempio 14.3.2. Un produttore di batterie sostiene che la vita media dei suoi prodotti sia di almeno 150 ore. Per verificare questa affermazione, si mettono in funzione simultaneamente 100 batterie, con l'intenzione di fermare l'esperimento quando si siano riscontrati 20 difetti. Se alla fine il tempo di funzionamento complessivo è stato di 1 800 ore, cosa si conclude?

Calcoliamo il valore della statistica del test, che è $2\tau/\theta_0 = 3\,600/150 = 24$. Il p-dei-dati è la probabilità che una chi-quadro con $2r = 40$ gradi di libertà assuma un valore inferiore a 24. Il Programma 5.8.1a ci fornisce allora:

$$p\text{-dei-dati} = P(\chi^2_{40} \leq 24) \approx 0.021$$

e quindi l'affermazione del produttore va rifiutata – ad esempio – al 5% di significatività. $\square$

Una conseguenza dell'Equazione (14.3.6) è che l'accuratezza dello stimatore τ/r dipende solo da r e non dal numero di componenti messi in prova, n. L'importanza di n risiede nel fatto che se questo valore è grande, si ha l'assicurazione che con elevata probabilità l'esperimento avrà breve durata. Questo si può evincere dal seguente calcolo dei momenti di $X_{(r)}$, che è la durata della prova.

Siccome, ponendo $X_{(0)} := 0$, si ha che

$$X_{(j)} - X_{(j-1)} = \frac{Y_j}{n-j+1}, \qquad j = 1, 2, \ldots, r$$

sommando su tutti gli indici j si trova che

$$X_{(r)} = \sum_{j=1}^{r} \frac{Y_j}{n-j+1}$$

Ricordando a questo punto che $Y_1, Y_2, \ldots, Y_r$ sono esponenziali indipendenti di media θ, e perciò hanno varianza θ^2, è facile dedurre che

$$\begin{aligned} E[X_{(r)}] &= \sum_{j=1}^{r} \frac{\theta}{n-j+1} \\ &= \theta \sum_{j=n-r+1}^{n} \frac{1}{j} \end{aligned} \tag{14.3.8}$$

$$\begin{aligned} \mathrm{Var}(X_{(r)}) &= \sum_{j=1}^{r} \frac{\mathrm{Var}(Y_j)}{(n-j+1)^2} \\ &= \theta^2 \sum_{j=n-r+1}^{n} \frac{1}{j^2} \end{aligned} \tag{14.3.9}$$

Quando n è grande, le due formule esatte qui sopra possono essere approssimate grazie al fatto che

$$\sum_{j=k}^{n} \frac{1}{j} \approx \int_k^n \frac{dx}{x} = \log \frac{n}{k} \qquad \text{e} \qquad \sum_{j=k}^{n} \frac{1}{j^2} \approx \int_k^n \frac{dx}{x^2} = \frac{1}{k} - \frac{1}{n}$$

Gli andamenti asintotici di media e varianza sono quindi

$$E[X_{(r)}] \sim \theta \log \left(\frac{n}{n-r+1} \right) \tag{14.3.10}$$

$$\mathrm{Var}(X_{(r)}) \sim \frac{\theta^2 (r-1)}{n(n-r+1)} \tag{14.3.11}$$

Come ci si attendeva, media e varianza convergono a zero per n che tende all'infinito.

In pratica, se nell'Esempio 14.3.2 la vita media di una batteria fosse stata di 120 ore, l'esperimento descritto si sarebbe concluso dopo un tempo aleatorio $X_{(20)}$ molto minore:

$$E[X_{(20)}] \approx 120 \log\left(\frac{100}{81}\right) \approx 25.29$$

$$\mathrm{Var}(X_{(20)}) \approx \frac{120^2 \times 19}{100 \times 81} \approx 33.78$$

14.3.2 Prove sequenziali

In questa sezione ci occupiamo di un diverso tipo di situazione. Immaginiamo di disporre di una riserva infinita (o semplicemente molto grande) di oggetti, ciascuno con tempo di vita esponenziale con media sconosciuta θ, e di esaminarli sequenzialmente, mettendone in funzione uno nuovo ogni volta che il precedente si guasta; l'esperimento viene concluso dopo un tempo prefissato T. I dati a nostra disposizione sono il numero r di oggetti che si sono guastati entro l'istante T, e i tempi di vita dei primi r oggetti, che indichiamo con $x_1, x_2, \ldots, x_r$.

Se denotiamo con X_i il tempo di vita dell'oggetto i-esimo, si ottengono i dati precedenti solo se

$$X_i = x_i, \qquad i = 1, 2, \ldots, r$$
$$\sum_{i=1}^{r} x_i < T \qquad \text{e} \qquad X_{r+1} > T - \sum_{i=1}^{r} x_i \tag{14.3.12}$$

Infatti affinché il numero di guasti sia esattamente pari a r, deve accadere che il guasto r-esimo si verifichi entro T, (e quindi $\sum_{i=1}^{r} X_i < T$), mentre il tempo di vita dell'oggetto $(r+1)$-esimo deve essere maggiore di $T - \sum_{i=1}^{r} X_i$ (si veda la Figura 14.1).

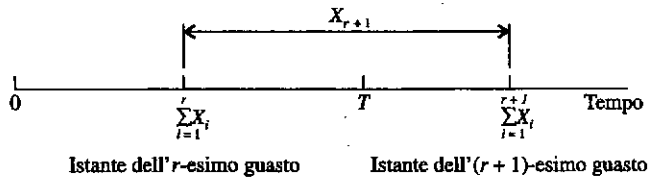


Figura 14.1 Un totale di r guasti entro il tempo T.

Non è difficile a questo punto rendersi conto che la likelihood corrispondente ai dati $r, x_1, x_2, \ldots, x_r$, con $\sum x_i < T$, è data da

$$\begin{aligned} f(r, x_1, x_2, \ldots, x_r|\theta) &= f_{X_1, X_2, \ldots, X_r}(x_1, x_2, \ldots, x_r) \cdot P\left(X_{r+1} > T - \sum_{i=1}^{r} x_i\right) \\ &= \frac{1}{\theta^r} \exp\left\{-\sum_{i=1}^{r} \frac{x_i}{\theta}\right\} \exp\left\{-\frac{1}{\theta}\left(T - \sum_{i=1}^{r} x_i\right)\right\} \\ &= \frac{1}{\theta^r} e^{-T/\theta} \end{aligned} \tag{14.3.13}$$

Calcoliamo la derivata rispetto al parametro, del logaritmo di questa espressione,

$$\log f(r, x_1, x_2, \ldots, x_r|\theta) = -r \log \theta - \frac{T}{\theta}$$
$$\frac{\partial}{\partial \theta} \log f(r, x_1, x_2, \ldots, x_r|\theta) = -\frac{r}{\theta} + \frac{T}{\theta^2}$$

La scelta di $\hat{\theta}$ che annulla l'ultima espressione è

$$\hat{\theta} := \frac{T}{r} \tag{14.3.14}$$

Visto che T rappresenta il tempo di funzionamento totalizzato da tutti gli oggetti in esame, anche in questo caso, come in quello delle prove simultanee, lo stimatore di massima verosimiglianza della media cercata è il rapporto tra il total time on test e il numero di guasti osservati durante tale periodo.

Se si denota con $N(T)$ la statistica che indica il numero di guasti osservati fino all'istante T, lo stimatore di massima verosimiglianza di θ è dato da $T/N(T)$, ma come si può trovare un intervallo di valori con un generico livello di confidenza $1-\alpha$? Sia r il valore assunto da $N(T)$, e immaginiamo di determinare due valori θ_I e θ_S tali che

$$P_{\theta_S}(N(T) \geq r) = \frac{\alpha}{2} \qquad \text{e} \qquad P_{\theta_I}(N(T) \leq r) = \frac{\alpha}{2}$$

dove si è indicata con $P_\theta(A)$ la probabilità che si verifichi l'evento A, nell'ipotesi che la media reale sia θ. In queste ipotesi vi è un livello di confidenza $1 - \alpha$ che

$$\theta \in (\theta_I, \theta_S)$$

Per capire il motivo di tale affermazione, si noti intanto che $P_\theta(N(T) \leq r)$ cresce con θ mentre $P_\theta(N(T) \geq r)$ decresce (perché?). Di conseguenza

$$\text{se } \theta < \theta_I, \qquad \text{allora } P_\theta(N(T) \leq r) < P_{\theta_I}(N(T) \leq r) = \frac{\alpha}{2}$$
$$\text{se } \theta > \theta_S, \qquad \text{allora } P_\theta(N(T) \geq r) < P_{\theta_S}(N(T) \geq r) = \frac{\alpha}{2}$$

Quindi se θ fosse esterna all'intervallo (θ_I, θ_S), il valore osservato r sarebbe così estremo da richiedere il verificarsi di un evento di probabilità inferiore ad α.

Resta solo da determinare il valore di θ_I e θ_S. L'evento $\{N(T) \geq r\}$ si verifica quando il guasto r-esimo avviene prima dell'istante T. Ovvero,

$$N(T) \geq r \Leftrightarrow X_1 + X_2 + \cdots + X_r \leq T \tag{14.3.15}$$

e quindi, se W ha distribuzione gamma di parametri r e $1/\theta$,

$$\begin{aligned} P_\theta(N(T) \geq r) &= P_\theta(X_1 + X_2 + \cdots + X_r \leq T) \\ &= P(W \leq T) \\ &= P\left(\frac{\theta}{2}\chi^2_{2r} \leq T\right) \\ &= P\left(\chi^2_{2r} \leq \frac{2T}{\theta}\right) \end{aligned}$$

Valutando l'equazione precedente in $\theta = \theta_S$ si ottiene che

$$\frac{\alpha}{2} = P_{\theta_S}(N(T) \geq r) = P\left(\chi^2_{2r} \leq \frac{2T}{\theta_S}\right)$$

per cui

$$\frac{2T}{\theta_S} = \chi^2_{1-\frac{\alpha}{2},2r}$$

ovvero

$$\theta_S = \frac{2T}{\chi^2_{1-\frac{\alpha}{2},2r}}$$

In maniera analoga è possibile dimostrare che

$$\theta_I = \frac{2T}{\chi^2_{\frac{\alpha}{2},2r}}$$

e quindi si può asserire con livello di confidenza $1 - \alpha$ che

$$\theta \in \left(\frac{2T}{\chi^2_{\frac{\alpha}{2},2r}}, \quad \frac{2T}{\chi^2_{1-\frac{\alpha}{2},2r}}\right) \tag{14.3.16}$$

Esempio 14.3.3. In una prova sequenziale la cui durata è fissata in $T = 500$ ore, si verificano 10 guasti. Se i tempi di vita dei singoli esemplari hanno distribuzione esponenziale di media θ, la stima di massima verosimiglianza per θ è di $500/10 = 50$ ore. Si può ottenere un intervallo di confidenza al 95%, calcolando

$$\theta \in \left(\frac{1\,000}{\chi^2_{0.025,20}}, \quad \frac{1\,000}{\chi^2_{0.975,20}}\right)$$

La Tabella A.2 in Appendice fornisce, per le chi-quadro che ci interessano, i valori

$$\chi^2_{0.025,20} \approx 34.17, \qquad \chi^2_{0.975,20} \approx 9.59$$

e quindi si può affermare, con il 95% di confidenza, che

$$\theta \in (29.27, 104.28) \quad \square$$

Nel caso si desideri verificare, con livello di significatività α, l'ipotesi nulla

$$H_0 : \theta = \theta_0$$

in alternativa all'ipotesi

$$H_1 : \theta \neq \theta_0$$

si denota con r il valore assunto dalla statistica $N(T)$, e quindi si rifiuta l'ipotesi nulla se accade che

$$P_{\theta_0}(N(T) \leq r) \leq \frac{\alpha}{2} \qquad \text{o} \qquad P_{\theta_0}(N(T) \geq r) \leq \frac{\alpha}{2}$$

Detto in altri termini, l'ipotesi H_0 va rifiutata a tutti i livelli di significatività maggiori o uguali al p-dei-dati, che è dato da

$$\begin{aligned} p\text{-dei-dati} &= 2\min\{P_{\theta_0}(N(T) \geq r),\ P_{\theta_0}(N(T) \leq r)\} \\ &= 2\min\{P_{\theta_0}(N(T) \geq r),\ 1 - P_{\theta_0}(N(T) \geq r+1)\} \\ &= 2\min\left\{P\left(\chi^2_{2r} \leq \frac{2T}{\theta_0}\right),\ 1 - P\left(\chi^2_{2r+2} \leq \frac{2T}{\theta_0}\right)\right\} \end{aligned}$$

Il p-dei-dati per un test statistico unilaterale può essere trovato in maniera analoga.

Si rammenti che le probabilità delle distribuzioni chi-quadro che compaiono nelle espressioni precedenti, possono essere ottenute usando il Programma 5.8.1a.

Esempio 14.3.4. Una compagnia sostiene che il tempo di vita medio dei semiconduttori che produce è almeno di 25 ore. Per avvalorare questa affermazione, una società di certificazione indipendente mette in prova sequenziale questi componenti per un tempo complessivo di 600 ore. Si contano in tutto 30 guasti. Cosa si può dire al 10% di significatività sull'affermazione del produttore?

Si tratta di un test statistico unilaterale delle ipotesi

$$H_0 : \theta \geq 25 \qquad \text{contro} \qquad H_1 : \theta < 25$$

Il p-dei-dati rappresenta la probabilità che avvengano 30 o più guasti, nell'ipotesi che la vita media sia 25; ovvero:

$$\begin{aligned} p\text{-dei-dati} &= P_{25}(N(600) \geq 30) \\ &= P(\chi^2_{60} \leq 1\,200/25) \\ &= P(\chi^2_{60} \leq 48) \\ &\approx 0.132 \qquad \text{grazie al Programma 5.8.1a} \end{aligned}$$

Quindi con livello di significatività del 10%, l'ipotesi nulla viene accettata. □

14.3.3 Test simultaneo – interruzione ad un tempo fissato

Consideriamo un diverso tipo di esame per componenti con tempi di vita esponenziali. Come per la Sezione 14.3.1 si mettono in prova simultaneamente n esemplari con tempi di vita indipendenti. A differenza di quanto fatto precedentemente, però, supponiamo di arrestare il processo dopo un tempo T fissato, o al guastarsi dell'n-esimo oggetto, se ciò dovesse verificarsi prima. Vogliamo stimare il valore di θ usando i dati a nostra disposizione. Denotiamo quindi con $i_1, i_2, \ldots, i_r$ i numeri identificativi degli $r \leq n$ oggetti che si sono guastati entro il tempo T, e con $x_1, x_2, \ldots, x_r$ i loro tempi di vita; resta inteso che i rimanenti $n - r$ oggetti sono sopravvissuti oltre il tempo T.

È facile verificare che la funzione di likelihood è data da

$$\begin{aligned} f(i_1, \ldots, i_r, x_1, \ldots, x_r|\theta) &= f_{X_{i_1},\ldots,X_{i_r}}(x_1, \ldots, x_r) \cdot P(X_j > T, j \notin \{i_1, \ldots, i_r\}) \\ &= \frac{1}{\theta^r} e^{-x_1/\theta} e^{-x_2/\theta} \cdots e^{-x_r/\theta} (e^{-T/\theta})^{n-r} \\ &= \frac{1}{\theta^r} \exp\left\{-\frac{1}{\theta}\sum_{i=1}^{r} x_i - \frac{(n-r)T}{\theta}\right\} \end{aligned} \tag{14.3.17}$$

Per determinare lo stimatore di massima verosimiglianza, occorre la derivata rispetto a θ del logaritmo di questa espressione,

$$\log f(i_1, \ldots, i_r, x_1, \ldots, x_r|\theta) = -r \log \theta - \frac{1}{\theta}\sum_{i=1}^{r} x_i - \frac{(n-r)T}{\theta}$$

$$\frac{\partial}{\partial\theta} \log f(i_1, \ldots, i_r, x_1, \ldots, x_r|\theta) = -\frac{r}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{r} x_i + \frac{(n-r)T}{\theta^2}$$

Uguagliando a zero l'ultima espressione e risolvendo in termini di θ si trova che la stima di massima verosimiglianza è data da

$$\frac{\sum_{i=1}^{r} x_i + (n-r)T}{r}$$

e quindi il corrispondente stimatore è dato dalla statistica seguente:

$$\hat{\theta} := \frac{\sum_{i=1}^{R} X_{(i)} + (n-R)T}{R} \tag{14.3.18}$$

dove si è denotato con R il numero di oggetti guastatisi entro il tempo T, e con $X_{(i)}$, per $i = 1, 2, \ldots, R$, i loro tempi di vita nell'ordine.

Se si denota ancora una volta con τ il total time on test associato all'esperimento, è facile convincersi che

$$\tau = \sum_{i=1}^{R} X_{(i)} + (n-R)T \tag{14.3.19}$$

infatti gli oggetti che si guastano entro l'istante T hanno tempi di vita dati da $X_{(1)}, X_{(2)}, \ldots, X_{(R)}$, mentre i restanti $n - R$ vengono mantenuti in funzione per un tempo T, fino all'interruzione dell'esperimento.

Come nelle Sezioni 14.3.1 e 14.3.2, abbiamo provato anche nel caso di prove simultanee interrotte ad un tempo prefissato, che lo stimatore di massima verosimiglianza della vita media di una popolazione esponenziale è il rapporto tra il total time on test e il numero di guasti osservati.

Osservazione 14.3.2. Come il lettore avrà ormai intuito, il fatto che negli esperimenti sui tempi di vita di componenti esponenziali, lo stimatore di massima verosimiglianza sia dato dal rapporto tra total time on test e numero di guasti osservati, è un risultato del tutto generale. Per convincerci di questo enunciato, consideriamo una *qualunque* situazione in cui siano in prova dei componenti esponenziali indipendenti, e supponiamo che alla conclusione dell'esperimento, r di essi si sono guastati, avendo avuto tempi di vita $x_1, x_2, \ldots, x_r$, mentre altri s componenti siano sopravvissuti, restando in funzione per dei tempi $y_1, y_2, \ldots, y_s$. La likelihood di θ per un tale esito è proporzionale a

$$\frac{1}{\theta^r} e^{-x_1/\theta} \cdots e^{-x_r/\theta} e^{-y_1/\theta} \cdots e^{-y_s/\theta} = \frac{1}{\theta^r} \exp\left\{-\frac{1}{\theta}\sum_{i=1}^{r} x_i - \frac{1}{\theta}\sum_{i=1}^{s} y_i\right\} \tag{14.3.20}$$

La costante di proporzionalità sottointesa dipende caso per caso dalla struttura dell'esperimento, ma non da θ. (Ad esempio può dipendere dal fatto che le durate $x_1, x_2, \ldots, x_r$ siano ordinate o no, oppure dalla scelta di interrompere la prova ad un tempo fissato o aleatorio.) È facile dedurre dall'equazione precedente che la stima di massima verosimiglianza per θ è data da

$$\frac{1}{r}\sum_{i=1}^{r} x_i + \frac{1}{r}\sum_{i=1}^{s} y_i \tag{14.3.21}$$

Se si denota con τ la statistica (aleatoria) che rappresenta il tempo complessivo di funzionamento del sistema, si vede che $\sum_{i=1}^{r} x_i + \sum_{i=1}^{s} y_i$ costituisce la sua realizzazione, quindi lo stimatore di massima verosimiglianza è anche in questo caso dato da

$$\hat{\theta} := \frac{\tau}{R} \tag{14.3.22}$$

La distribuzione di τ/R a questo livello di generalità non può essere dedotta[2], e quindi non siamo in grado di esibire intervalli di confidenza per θ.

Anziché proseguire in questa direzione ci rivolgiamo ora allo studio delle stime bayesiane.

14.3.4 Approccio bayesiano

Supponiamo di mettere in prova dei componenti con tempi di vita esponenziali e indipendenti, con media incognita θ. Come notato nell'Osservazione 14.3.2, la likelihood dei dati può essere espressa tramite

$$f(\text{dati}|\theta) = \frac{K}{\theta^r} e^{-t/\theta}$$

dove con t si è indicato il total time on test, ovvero la somma dei tempi per cui sono stati in funzione tutti i pezzi provati. Come in precedenza r denota il numero di guasti osservati.

Denotiamo con $\lambda := 1/\theta$, l'intensità della distribuzione esponenziale in esame. Nell'approccio bayesiano è più conveniente lavorare con λ che con il suo reciproco. La likelihood di questo nuovo parametro si riscrive nella forma

$$f(\text{dati}|\lambda) = K\lambda^r e^{-\lambda t} \tag{14.3.23}$$

Se si suppone prima dell'esperimento, che λ abbia densità a priori $g(\lambda)$, la relativa densità a posteriori, in funzione dei dati è

$$\begin{aligned} f(\lambda|\text{dati}) &= \frac{f(\text{dati}|\lambda)g(\lambda)}{\int f(\text{dati}|\mu)g(\mu)\,d\mu} \\ &= \frac{\lambda^r e^{-\lambda t} g(\lambda)}{\int \mu^r e^{-\mu t} g(\mu)\,d\mu} \end{aligned} \tag{14.3.24}$$

La densità a posteriori precedente assume una forma particolarmente conveniente quando g è una densità di tipo gamma. Denotiamo con b ed a i relativi parametri, in modo tale che g prende la forma seguente,

$$g(\lambda) = \frac{a^b}{\Gamma(b)} \lambda^{b-1} e^{-a\lambda}, \qquad \lambda > 0 \tag{14.3.25}$$

e l'Equazione (14.3.24) diviene

$$f(\lambda|\text{dati}) = C\lambda^{b+r-1} e^{-(a+t)\lambda}, \qquad \lambda > 0$$

[2] Una difficoltà ad esempio è costituita dal fatto che τ e R sono entrambe aleatorie e non sono indipendenti.

dove C è una costante che non dipende da λ. Siccome l'espressione precedente deve essere una densità di probabilità, vi riconosciamo una distribuzione gamma di parametri $b+r$ ed $a+t$, deduciamo che $C = (a+t)^{b+r}/\Gamma(b+r)$, e otteniamo che

$$f(\lambda|\text{dati}) = \frac{(a+t)^{b+r}}{\Gamma(b+r)} \lambda^{b+r-1} e^{-(a+t)\lambda}, \qquad \lambda > 0 \tag{14.3.26}$$

In altri termini, se la distribuzione a priori di λ è di tipo gamma con parametri b ed a, allora indipendentemente dalla struttura dell'esperimento, la distribuzione condizionale di λ, a posteriori dell'osservazione dei dati, è di tipo gamma con parametri $b+R$ e $a+\tau$, dove τ e R rappresentano come al solito il total time on test degli oggetti e il numero di guasti osservati. Poiché il valore atteso di una variabile aleatoria gamma di parametri b e a è b/a (si veda la Sezione 5.7), possiamo concludere che lo stimatore di Bayes di λ, $E[\lambda|\text{dati}]$ è dato da

$$E[\lambda|\text{dati}] = \frac{b+R}{a+\tau} \tag{14.3.27}$$

Esempio 14.3.5. Supponiamo che vengano messi in prova (in momenti diversi) 20 componenti con tempi di vita esponenziali di intensità incognita λ. Alla conclusione dell'esperimento, 10 esemplari si sono guastati, dopo essere stati in funzione per un numero di ore:

5 7 6.2 8.1 7.9 15 18 3.9 4.6 5.8

Gli altri 10 pezzi al momento dell'interruzione dell'esperimento erano stati in funzione per un numero di ore:

3 3.2 4.1 1.8 1.6 2.7 1.2 5.4 10.3 1.5

Se prima dell'esperimento la nostra convinzione era che λ avesse distribuzione gamma di parametri 2 e 20, qual è lo stimatore di Bayes per λ?

Siccome

$$\tau = 116.3 \qquad \text{e} \qquad R = 10$$

segue che la stima bayesiana di λ è

$$E[\lambda|\text{dati}] = \frac{12}{136.3} \approx 0.088 \quad \square$$

Osservazione 14.3.3. Come abbiamo visto, la scelta della gamma, come distribuzione a priori per l'intensità dei tempi di vita esponenziali, rende i calcoli piuttosto semplici. Anche se dal punto di vista delle applicazioni, questa non è una giustificazione sufficiente, tale scelta viene spesso motivata dalla flessibilità con cui si possono fissare i due parametri, che consente di approssimare ragionevolmente quasi ogni convinzione a priori si possa esprimere.

14.4 Confronto di due campioni

Una azienda possiede due stabilimenti per la produzione di valvole termoioniche. Si immagina che questi componenti abbiano tempi di vita esponenziali, e si denotano con θ_1 e θ_2 i tempi di vita medi relativi ai due impianti. Per verificare l'ipotesi che i prodotti dei due stabilimenti siano equivalenti (almeno per quanto riguarda il tempo di vita medio), si estraggono campioni indipendenti di n e m valvole rispettivamente, che vengono esaminati.

Denotiamo con $X_1, X_2, \ldots, X_n$ i tempi di vita delle n valvole campionate dal primo stabilimento, e con $Y_1, Y_2, \ldots, Y_m$ quelli delle m valvole provenienti dal secondo. Vogliamo verificare l'ipotesi $H_0 : \theta_1 = \theta_2$, supponendo che le X_i e le Y_j siano campioni aleatori indipendenti di popolazioni esponenziali con medie θ_1 e θ_2.

Per prima cosa notiamo che $\sum_{i=1}^{n} X_i$ e $\sum_{i=1}^{m} Y_i$ (essendo somme di esponenziali i.i.d.) sono variabili aleatorie gamma indipendenti con parametri rispettivamente n e $1/\theta_1$ la prima, m e $1/\theta_2$ la seconda. Dall'equivalenza tra la distribuzione gamma e la chi-quadro deduciamo che

$$\begin{aligned} \frac{2}{\theta_1}\sum_{i=1}^{n} X_i &\sim \chi^2_{2n} \\ \frac{2}{\theta_2}\sum_{i=1}^{m} Y_i &\sim \chi^2_{2m} \end{aligned} \tag{14.4.1}$$

Perciò dalla definizione di distribuzione F di Fisher otteniamo che

$$\frac{\theta_2 \overline{X}}{\theta_1 \overline{Y}} = \left(\frac{1}{2n}\frac{2}{\theta_1}\sum_{i=1}^{n} X_i\right)\left(\frac{1}{2m}\frac{2}{\theta_2}\sum_{i=1}^{m} Y_i\right)^{-1} \sim F_{n,m} \tag{14.4.2}$$

dove si sono indicate con $\overline{X}$ e $\overline{Y}$ le due medie campionarie.

Perciò se l'ipotesi $\theta_1 = \theta_2$ è soddisfatta, il rapporto $\overline{X}/\overline{Y}$ ha distribuzione F con n e m gradi di libertà. Questo fatto suggerisce di costruire il test di

$$H_0 : \theta_1 = \theta_2 \qquad \text{contro} \qquad H_1 : \theta_1 \neq \theta_2$$

come segue: **(1)** si sceglie un livello di significatività α; **(2)** si determina il valore v assunto della statistica $\overline{X}/\overline{Y}$; **(3)** si calcola $P(F \leq v)$, dove $F \sim F_{n,m}$; **(4)** se tale probabilità risulta inferiore ad $\alpha/2$ o superiore ad $1 - \alpha/2$, l'ipotesi viene rifiutata: nel primo caso perché $\overline{X}$ è sensibilmente inferiore a $\overline{Y}$, nel secondo caso perché è vero il contrario.

Cambiando punto di vista, è possibile calcolare il p-dei-dati, che è dato da

$$p\text{-dei-dati} = 2\min\{P(F \leq v),\ 1 - P(F \leq v)\} \tag{14.4.3}$$

Esempio 14.4.1. Verifichiamo al 5% di significatività l'ipotesi che i tempi di vit dei componenti provenienti dai due impianti abbiano la stessa distribuzione. Sup poniamo che i dati siano variabili aleatorie esponenziali provenienti da popolazion indipendenti. Un campione di 10 componenti del primo impianto ha totalizzato u tempo di funzionamento complessivo di 420 ore, mentre 15 pezzi provenienti da secondo impianto hanno raggiunto un totale di 510 ore.

Il valore della statistica del test è $\overline{X}/\overline{Y} = 42/34 \approx 1.235$. Per calcolare l probabilità che una F di Fisher con 10 e 15 gradi di libertà realizzi un valore inferior a questo, eseguiamo il Programma 5.8.3, ottenendo che

$$P(F_{10,15} < 1.235) \approx 0.655$$

Siccome il p-dei-dati risultante è $2 \times (1 - 0.655) = 69\%$, non si può rifiutare H_0. □

14.5 La distribuzione di Weibull

La distribuzione esponenziale che abbiamo studiato finora corrisponde ai casi in cu la funzione intensità di rotture $\lambda(t)$ si riduce ad una costante. Vi sono però molt situazioni in cui è più realistico supporre che $\lambda(t)$ sia una funzione crescente[3] o de crescente[4] del tempo.

Un esempio di funzione di rotture di questo tipo è dato da

$$\lambda(t) = \alpha\beta t^{\beta-1}, \qquad t > 0 \tag{14.5.1}$$

dove α e β sono costanti positive qualsiasi. La distribuzione che corrisponde a questa scelta di λ prende il nome di distribuzione di *Weibull* di parametri α e β. Si noti che λ è una funzione crescente se $\beta > 1$ e decrescente se $\beta < 1$, mentre se $\beta = 1$ diviene costante, e la distribuzione si riduce ad una esponenziale di intensità α.

Le funzioni di ripartizione e di densità di Weibull si ottengono a partire dall'E quazione (14.2.3) e dalla definizione di λ:

$$\begin{aligned} F(t) &= 1 - \exp\left\{-\int_0^t \lambda(s)\,ds\right\} \\ &= 1 - e^{-\alpha t^\beta}, \qquad t > 0 \end{aligned} \tag{14.5.2}$$

$$\begin{aligned} f(t) &= \frac{d}{dt}\{1 - e^{-\alpha t^\beta}\} \\ &= \alpha\beta t^{\beta-1} e^{-\alpha t^\beta}, \qquad t > 0 \end{aligned} \tag{14.5.3}$$

La Figura 14.2 rappresenta i grafici di densità di questo tipo per diversi valori di α e di β.

[3] Quando gli oggetti studiati subiscono un continuo deterioramento.

[4] Quando alcuni degli oggetti studiati hanno difetti di fabbricazione che li portano a guastarsi presto.

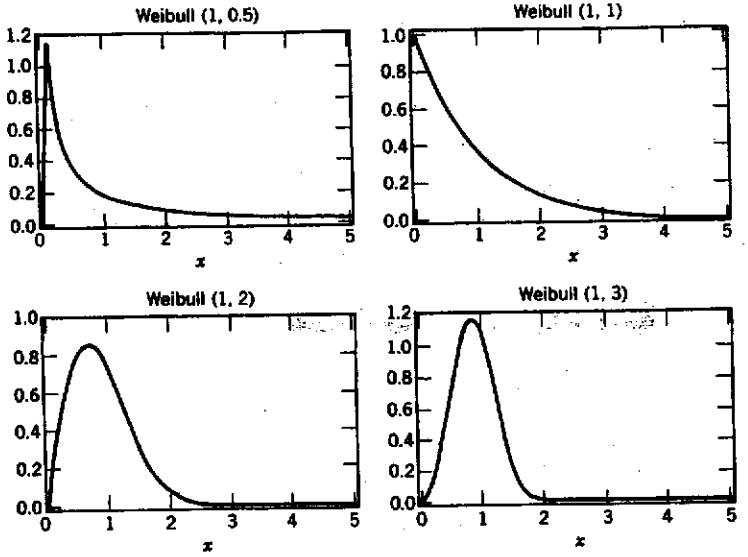


Figura 14.2 Funzioni di densità di distribuzioni Weibull.

Supponiamo ora che $X_1, X_2, \ldots, X_n$ siano variabili aleatorie di tipo Weibull indipendenti e tutte aventi i medesimi parametri α e β, che si assumono incogniti. Per stimare α e β usiamo l'approccio della massima verosimiglianza. Dall'Equazione (14.5.3) si ricava che

$$f(x_1, x_2, \ldots, x_n) = \alpha^n \beta^n x_1^{\beta-1} \cdots x_n^{\beta-1} \exp\left\{-\alpha \sum_{i=1}^{n} x_i^{\beta}\right\} \tag{14.5.4}$$

per cui il logaritmo della verosimiglianza risulta

$$\log f(x_1, x_2, \ldots, x_n) = n \log \alpha + n \log \beta + (\beta - 1) \sum_{i=1}^{n} \log x_i - \alpha \sum_{i=1}^{n} x_i^{\beta}$$

Calcoliamo le derivate parziali per cercare i punti critici della verosimiglianza,

$$\frac{\partial}{\partial \alpha} \log f(x_1, x_2, \ldots, x_n) = \frac{n}{\alpha} - \sum_{i=1}^{n} x_i^{\beta}$$

$$\frac{\partial}{\partial \beta} \log f(x_1, x_2, \ldots, x_n) = \frac{n}{\beta} + \sum_{i=1}^{n} \log x_i - \alpha \sum_{i=1}^{n} x_i^{\beta} \log x_i$$

Uguagliando a zero queste due formule si possono trovare delle relazioni per le stime di massima verosimiglianza $\widehat{\alpha}$ e $\widehat{\beta}$:

$$\begin{cases} \dfrac{n}{\widehat{\alpha}} - \displaystyle\sum_{i=1}^{n} x_i^{\widehat{\beta}} = 0 \\ \dfrac{n}{\widehat{\beta}} + \displaystyle\sum_{i=1}^{n} \log x_i - \widehat{\alpha} \sum_{i=1}^{n} x_i^{\widehat{\beta}} \log x_i = 0 \end{cases}$$

o, equivalentemente,

$$\begin{cases} \widehat{\alpha} = \left(\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i^{\widehat{\beta}}\right)^{-1} \\ \left[\dfrac{1}{\widehat{\beta}} + \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \log x_i\right] \sum_{i=1}^{n} x_i^{\widehat{\beta}} - \sum_{i=1}^{n} x_i^{\widehat{\beta}} \log x_i = 0 \end{cases} \tag{14.5.5}$$

Quest'ultimo sistema di equazioni può essere risolto, ricavando (numericamente) $\widehat{\beta}$ dalla seconda, e poi sostituendo il suo valore nella prima per ottenere $\widehat{\alpha}$.

Piuttosto che proseguire con questo approccio, preferiamo introdurre una seconda strategia, che risulta non solo computazionalmente più agevole, ma sembra anche fornire stime più accurate, come è indicato da studi di simulazione recenti.

14.5.1 Stima parametrica con il metodo dei minimi quadrati

Sia $X_1, X_2, \ldots, X_n$ un campione aleatorio di Weibull con funzione di ripartizione

$$F(x) = 1 - e^{-\alpha x^{\beta}}, \qquad x > 0$$

Possiamo linearizzare questa espressione nel modo seguente:

$$\log(1 - F(x)) = -\alpha x^{\beta}$$

$$\log\left(\frac{1}{1 - F(x)}\right) = \alpha x^{\beta}$$

$$\log\log\left(\frac{1}{1 - F(x)}\right) = \log \alpha + \beta \log x \tag{14.5.6}$$

Riordiniamo il campione dal valore minore al maggiore, denotando i dati permutati con $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$; per $i = 1, 2, \ldots, n$, indichiamo con $x_{(i)}$ il valore osservato per $X_{(i)}$.

Accettiamo per ora di essere in grado di approssimare i valori $\log\log[1/(1 - F(x_{(i)}))]$ (che sono incogniti perché non conosciamo la forma di F che dipende da α e β) con una n-upla di valori $y_1, y_2, \ldots, y_n$. Si deduce dall'Equazione (14.5.6) che

$$y_i \approx \log \alpha + \beta \log x_{(i)}, \qquad i = 1, 2, \ldots, n$$

Possiamo a questo punto scegliere α e β in modo da minimizzare la somma dei quadrati degli errori, ovvero

$$\sum_{i=1}^{n}(y_i - \beta \log x_{(i)} - \log \alpha)^2$$

In effetti, applicando la Proposizione 9.2.1 a pagina 344, si deduce subito che il minimo si ottiene quando i parametri sono $\widehat{\alpha}$ e $\widehat{\beta}$, definiti da

$$\begin{aligned} \widehat{\beta} &:= \frac{\sum_{i=1}^{n} y_i \log x_{(i)} - n\bar{y} \cdot \overline{\log x}}{\sum_{i=1}^{n} (\log x_{(i)})^2 - n(\overline{\log x})^2} \\ \widehat{\alpha} &:= \exp\{\bar{y} - \beta \overline{\log x}\} \end{aligned} \tag{14.5.7}$$

dove si è posto

$$\overline{\log x} := \frac{1}{n}\sum_{i=1}^{n} \log x_{(i)} \qquad \text{e} \qquad \bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i \tag{14.5.8}$$

Restano da determinare dei valori y_i che approssimino le quantità incognite:

$$\log\log\left(\frac{1}{1 - F(x_{(i)})}\right) = \log[-\log(1 - F(x_{(i)}))], \qquad i = 1, 2, \ldots, n$$

Presentiamo di seguito due metodi che permettono di ottenere questo tipo di approssimazioni.

Metodo 1 Si usa il fatto che

$$E[F(X_{(i)})] = \frac{i}{n+1} \tag{14.5.9}$$

e si approssima $F(x_{(i)})$ con $E[F(X_{(i)})]$, ponendo

$$\begin{aligned} y_i &:= \log(-\log(1 - E[F(X_{(i)})])) \\ &= \log\left[-\log\left(1 - \frac{i}{n+1}\right)\right] \\ &= \log\log\left(\frac{n+1}{n-i+1}\right) \end{aligned} \tag{14.5.10}$$

Metodo 2 Si usa qui il fatto che

$$E[-\log(1 - F(X_{(i)}))] = \frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{n-i+1} \tag{14.5.11}$$

ponendo di conseguenza

$$y_i := \log\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{n-i+1}\right) \tag{14.5.12}$$

Osservazione 14.5.1.

(a) Non è chiaro ad oggi quale di questi metodi fornisca le migliori stime dei parametri delle distribuzioni di Weibull.

(b) La dimostrazione delle Equazioni (14.5.9) e (14.5.11) è l'argomento dei Problemi dal 28 al 30.

Problemi

1. Una variabile aleatoria con funzione di ripartizione data da

$$F(t) = 1 - \exp\{-\alpha t^{\beta}\}, \qquad t > 0$$

si dice *di Weibull* di parametri α e β. Calcola la funzione di intensità di rotture corrispondente.

2. Siano X e Y due variabili aleatorie indipendenti con funzioni di intensità di rotture $\lambda_x(t)$ e $\lambda_y(t)$. Dimostra che la funzione intensità di rotture di $Z := \min\{X, Y\}$ è

$$\lambda_z(t) = \lambda_x(t) + \lambda_y(t)$$

3. Il rischio di contrarre un tumore ai polmoni, per un fumatore almeno quarantenne, può essere approssimato dalla funzione

$$\lambda(t) = 0.027 + 0.025\left(\frac{t-40}{40}\right)^4, \qquad t > 40$$

dove t rappresenta l'età in anni. Supponendo che un fumatore di 40 anni non muoia per altre cause, e che non smetta mai di fumare, qual è la probabilità che giunga **(a)** ai 50 anni di età, o **(b)** ai 60 anni di età, senza contrarre questa malattia?

4. Supponi che il tempo di vita di un certo prodotto abbia intensità di rotture $\lambda(t) = t^3$, per $t > 0$.

   (a) Qual è la probabilità che un esemplare funzioni per più di 2 unità di tempo?

   (b) Qual è la probabilità che si guasti tra gli istanti 0.4 e 1.4?

(c) Qual è la vita media?

(d) Qual è la probabilità che un esemplare di età 1 funzioni almeno per un'altra unità di tempo?

5. La distribuzione di un tempo di vita aleatorio si dice IFR (*increasing failure rate*) se la sua intensità di rotture è una funzione non decrescente di t.

(a) Dimostra che la seguente densità di tipo gamma è IFR:

$$f(t) = \lambda^2 t e^{-\lambda t}, \qquad t > 0$$

(b) Dimostra più in generale che una distribuzione gamma di parametri α e λ è IFR solo se $\alpha \geq 1$.

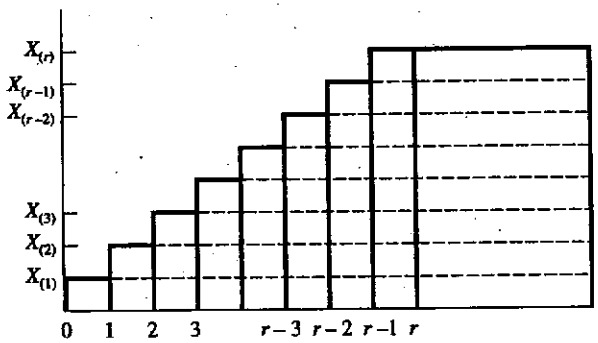
6. Dimostra che la distribuzione uniforme sull'intervallo (a, b) è IFR.

7. Per il modello della Sezione 14.3.1, spiega come si può usare la figura seguente per giustificare che

$$\tau = \sum_{j=1}^{r} Y_j$$

dove si è posto

$$Y_j := (n - j + 1)(X_{(j)} - X_{(j-1)})$$



Suggerimento: Entrambe le grandezze τ e $\sum_{j=1}^{r} Y_j$ rappresentano l'area della figura qui sopra, da due punti di vista diversi.

8. Un esperimento di prova simultanea di 30 transistor con vita esponenziale i.i.d. viene interrotto al decimo guasto. Si osservano, per i componenti che si guastano, le ore di vita seguenti:

4.1 7.3 13.2 18.8 24.5 30.8 38.1 45.5 53 62.2

(a) Qual è la stima di massima verosimiglianza per la vita media dei transistor?

(b) Calcola un intervallo di confidenza bilaterale al 95% per il tempo di vita medio.

(c) Determina un valore c che possiamo affermare essere inferiore alla media dei tempi di vita, con il 95% di confidenza.

(d) Verifica al 10% di significatività l'ipotesi che il tempo di vita medio sia di 7.5 ore, usando una alternativa bilaterale.

9. Supponi di dovere verificare l'ipotesi $H_0 : \theta = \theta_0$ in alternativa ad $H_1 : \theta \neq \theta_0$, con un esperimento strutturato secondo il modello della Sezione 14.3.1. Denota con v la realizzazione della statistica $2\tau/\theta_0$. Mostra che l'ipotesi nulla va rifiutata se il livello di significatività è superiore al valore del p-dei-dati, dato da

$$p\text{-dei-dati} = 2\min\{P(\chi^2_{2r} < v),\ 1 - P(\chi^2_{2r} < v)\}$$

Dove χ^2_{2r} rappresenta una variabile aleatoria con distribuzione chi-quadro con $2r$ gradi di libertà.

10. In un esperimento vengono messi in prova 30 componenti, e si interrompe tutto quando si verifica il guasto numero 8. I tempi in cui si hanno i guasti, in ore, sono i seguenti:

0.35 0.73 0.99 1.40 1.45 1.83 2.20 2.72

Verifica al 5% di significatività l'ipotesi che la vita media sia di 1 ora. Supponi che i tempi di vita siano esponenziali.

11. Immagina che vengano messi in prova simultanea 20 oggetti e che si sia deciso di terminare la sperimentazione in corrispondenza del decimo guasto. Calcola **(a)** valore atteso e **(b)** varianza della durata dell'esperimento.

12. Le valvole termoioniche prodotte in una certa fabbrica hanno vita esponenziale di media incognita θ. Per stimare il valore di questo parametro si vogliono mettere in prova simultaneamente n di questi componenti, fermarsi al decimo guasto, e possibilmente il tutto non dovrebbe richiedere (mediamente) più di 3 ore di sperimentazione. Se si pensa che un valore sensato per θ sia 20, quanto grande deve essere scelto il numero di componenti da esaminare n?

13. Un tipo di componenti elettronici viene sottoposto ad un esame sequenziale della durata di 300 ore. Il numero di guasti osservati è 16. Assumi che i tempi di vita (misurati in ore) siano esponenziali i.i.d. con media incognita θ.

(a) Trova la stima di massima verosimiglianza di θ.

(b) Verifica con il 5% di significatività l'ipotesi che $\theta = 20$ contro l'alternativa $\theta \neq 20$.

(c) Determina un intervallo di confidenza al 95% per θ.

***14.** Si ottiene un processo di Poisson se si conta il numero di "eventi" separati da pause esponenziali indipendenti, che si verificano in un intervallo di tempo fissato (si veda la Sezione 5.6.1). Dimostra che, se X è una variabile aleatoria di Poisson di media $x/2$, e $F_{\chi^2_{2n}}$ denota la funzione di ripartizione della distribuzione chi-quadro con $2n$ gradi di libertà,

$$P(X \geq n) = F_{\chi^2_{2n}}(x)$$

Suggerimento: Usa i risultati della Sezione 14.3.2

15. Gli oggetti di un campione estratto da una popolazione con vita esponenziale di media θ, vengono provati uno dopo l'altro, fino al momento del guasto r-esimo, o al più tardi, al raggiungimento dell'istante T.

(a) Determina la funzione di verosimiglianza.

(b) Verifica che anche in questo caso lo stimatore di massima verosimiglianza è dato dal rapporto tra il total time on test degli esemplari provati, e il numero di guasti osservati.

16. Dimostra che la stima che corrisponde alla funzione di massima verosimiglianza data dall'Equazione (14.3.20) è quella espressa dalla Equazione (14.3.21).

17. Un laboratorio ha strumentazione sufficiente a tenere in prova contemporaneamente un massimo di 5 componenti. Si devono testare 10 pezzi provenienti da una comune popolazione esponenziale, e si decide di cominciare con 5 di essi, sostituendo via via quelli guasti con altri nuovi, fino al guastarsi di tutti e dieci, o al raggiungimento delle 200 ore di prova. Se alla fine si sono contati 9 guasti, che si sono verificati agli istanti

15 28.2 46 62.2 76 86 128 153 197

Qual è la stima di massima verosimiglianza del tempo di vita medio di questi componenti?

18. Supponiamo che il tempo di remissione della leucemia dopo un tipo di trattamento chemioterapico sia (espresso in settimane) una variabile aleatoria esponenziale di media incognita θ. Si tiene sotto controllo un gruppo di 20 pazienti, e al momento attuale, i loro tempi di remissione sono di

1.2	1.8	2.2	4.1	5.6	8.4	11.8	13.4	16.2	21.7
29	41	42	42.4	49.3	60.5	61	94	98	99.2

dove si sono evidenziati con un riquadro i casi in cui la remissione non è ancora completa. Qual è la stima di massima verosimiglianza di θ?

19. Con riferimento al Problema 17, supponi che si ipotizzi che la distribuzione a priori di $\lambda := 1/\theta$, sia una gamma di parametri 1 e 100. Quanto vale lo stimatore di Bayes di λ?

20. Quale sarebbe lo stimatore di Bayes del parametro $\lambda := 1/\theta$, se nel Problema 18 fosse nota la distribuzione a priori di λ, esponenziale di intensità 30?

21. Quelli riportati qui sotto sono i minuti di funzionamento prima di rovinarsi, di due tipi di isolanti elettrici sottoposti ad una forte differenza di potenziale.

Tipo 1	212	88.5	122.3	116.4	125	132	66
Tipo 2	34.6	54	162	49	78	121	128

Verifica l'ipotesi che i due campioni di dati provengano dalla stessa distribuzione esponenziale.

22. Si suppone che due tipi di transistor abbiano tempi di vita con distribuzioni esponenziali (eventualmente) diverse. Si vuole verificare l'ipotesi che i tempi di vita medi siano identici; a questo scopo si mettono in prova n_1 trasistor del primo tipo, arrestando l'esperimento al guasto r_1-esimo, e si procede similmente con n_2 componenti del secondo tipo, interrompendo al guasto r_2-esimo.

(a) Usando i risultati della Sezione 14.3.1, mostra come l'ipotesi di uguaglianza delle medie possa essere verificata usando una statistica che, sotto l'ipotesi nulla, ha distribuzione F con $2r_1$ e $2r_2$ gradi di libertà.

(b) Supponi che i parametri concreti siano

$$n_1 = 20 \qquad r_1 = 10 \qquad n_2 = 10 \qquad r_2 = 7$$

e che gli istanti in cui si sono osservati i guasti siano stati

Tipo 1	10.4	23.2	31.4	45	61.1	69.6	81.3	95.2	112	129.4
Tipo 2	6.1	13.8	21.2	31.6	46.4	66.7	92.4			

Qual è il più piccolo livello di significatività α con il quale si rifiuta l'ipotesi che le medie siano uguali? (In altre parole: quanto vale il p-dei-dati?)

23. Sia X una variabile aleatoria di Weibull con parametri α e β. Dimostra che

$$E[X] = \alpha^{-1/\beta}\Gamma(1 + 1/\beta)$$

dove Γ denota la funzione gamma di Eulero, definita da

$$\Gamma(y) := \int_0^\infty e^{-x}x^{y-1}\,dx$$

Suggerimento: Scrivi

$$E[X] = \int_0^\infty t\alpha\beta t^{\beta-1}e^{-\alpha t^\beta}\,dt$$

quindi esegui il cambio di variabili

$$x = \alpha t^\beta, \qquad dx = \alpha\beta t^{\beta-1}dt$$

24. Mostra che la varianza di una variabilie aleatoria di Weibull di parametri α e β è data da

$$\alpha^{-2/\beta}\left[\Gamma\left(1+\frac{2}{\beta}\right) - \left(\Gamma\left(1+\frac{1}{\beta}\right)\right)^2\right]$$

25. Quelli che seguono sono dati campionati da una distribuzione di Weibull di parametri incogniti α e β. Determina le stime dei minimi quadrati dei parametri, usando ciascuno dei metodi presentati.

15.4 16.8 6.2 10.6 21.4 18.2 1.6 12.5 19.4 17

26. Mostra che se X è una variabile aleatoria di tipo Weibull di parametri α e β, allora αX^β è una variabile aleatoria esponenziale di media 1.

27. Sia U una variabile aleatoria uniforme su $(0, 1)$. Dimostra che $[-\alpha^{-1} \log U]^{1/\beta}$ è di tipo Weibull con parametri α e β.

I tre problemi seguenti riguardano le Equazioni (14.5.9) *e* (14.5.11).

28. Sia X una variabile aleatoria continua con funzione di ripartizione F. Dimostra che $F(X)$ e $1 - F(X)$ hanno entrambe distribuzione uniforme su $(0, 1)$.

29. Sia $X_{(i)}$ il valore i-esimo (in ordine crescente) di un campione di n osservazioni indipendenti di una popolazione con funzione di ripartizione F. Sia similmente $U_{(i)}$ l'i-esimo valore di n variabili aleatorie indipendenti, uniformi su $(0, 1)$.

(a) Mostra che la funzione di densità di $U_{(i)}$ è data da

$$f_{U_{(i)}} = \frac{n!}{(n-i)!(i-1)!} t^{i-1}(1-t)^{n-i}, \qquad 0 < t < 1$$

Suggerimento: Affinché l'i-esima (in ordine crescente) di n variabili aleatorie uniformi e indipendenti valga t, quante di esse devono valere meno di t, e quante di più? E quanti modi diversi ci sono per dividere un gruppo di n elementi in tre gruppi di ampiezza $i - 1$, 1 e $n - i$?

(b) Usa la parte **(a)** di questo problema per dimostrare che $E[U_{(i)}] = i/(n + 1)$.
Suggerimento: Per risolvere l'integrale risultante, usa il fatto che la densità di probabilità $f_{U_{(i)}}$ ha integrale unitario.

(c) Usa il Problema 28 per dimostrare che $E[F(X_{(i)})] = i/(n + 1)$.

30. **(a)** Dimostra che se U è uniforme su $(0, 1)$, allora $-\log U$ ha distribuzione esponenziale di media 1.

(b) Usa l'Equazione (14.3.8) e i risultati dei problemi precedenti per dimostrare la validità dell'Equazione (14.5.11).

A Tabelle

Tabella A.1 Funzione di ripartizione della distribuzione normale standard

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2}\, dy$$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Tabella A.2 Valori assunti da $\chi^2_{\alpha,n}$

	α							
n	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.00004	0.00016	0.00098	0.00393	3.841	5.024	6.635	7.879
2	0.0100	0.0201	0.0506	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672

Tabella A.3 Valori assunti da $t_{\alpha,n}$

	α				
n	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
50	1.299	1.676	2.009	2.403	2.678
70	1.294	1.667	1.994	2.381	2.648
100	1.290	1.660	1.984	2.364	2.626
∞	1.282	1.645	1.960	2.326	2.576

Tabella A.4 Valori assunti da $F_{0.05,n,m}$; n rappresenta i gradi di libertà al numeratore e m quelli al denominatore.

	n						
m	1	2	3	4	5	6	7
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01

Tabella A.5 Valori assunti da $C(m, d, \alpha)$

d		m									
		2	3	4	5	6	7	8	9	10	11
5	0.05	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17
	0.01	5.70	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48
6	0.05	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.65
	0.01	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30
7	0.05	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30
	0.01	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37	8.55
8	0.05	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05
	0.01	4.75	5.64	6.20	6.62	6.96	7.24	7.47	7.68	7.86	8.03
9	0.05	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.87
	0.01	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.33	7.49	7.64
10	0.05	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72
	0.01	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36
11	0.05	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61
	0.01	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13
12	0.05	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.51
	0.01	4.32	5.05	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94
13	0.05	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43
	0.01	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79
14	0.05	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36
	0.01	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66
15	0.05	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31
	0.01	4.17	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55
16	0.05	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26
	0.01	4.13	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46
17	0.05	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.21
	0.01	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38
18	0.05	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17
	0.01	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31
20	0.05	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11
	0.01	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19
24	0.05	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01
	0.01	3.96	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02
30	0.05	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92
	0.01	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85
40	0.05	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82
	0.01	3.82	4.37	4.70	4.93	5.11	5.26	5.39	5.50	5.60	5.69
60	0.05	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73
	0.01	3.76	4.28	4.59	4.82	4.99	5.13	5.25	5.36	5.45	5.53
120	0.05	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.64
	0.01	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.37
∞	0.05	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55
	0.01	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23

Indice analitico

Finito di stampare nel mese di marzo 2003
da Legoprint – Lavis – Trento