

LEGI
Via E.
Tel

Antonio Di Crescenzo Luigi M. Ricciardi

519.5

DIC

4.6

Elementi
di statistica

ISPS: 144846

CDD: 519.5



CONSULTAZIONE

Liguori Editore

Tutti i diritti sono riservati. Nessuna parte di questa pubblicazione può essere tradotta, riprodotta, copiata o trasmessa senza l'autorizzazione dell'editore. L'AIDRO (Associazione Italiana per i Diritti di Riproduzione delle Opere dell'Ingegno), via delle Erbe 2, 20121 Milano, potrà concedere una licenza di riproduzione a pagamento per una porzione non superiore a un decimo del presente volume.

Prima edizione italiana Aprile 2000

Liguori Editore, S.r.l.
via Posillipo 394
I 80123 Napoli
<http://www.liuorit.it>

Copyright © Liguori Editore, S.r.l. 2000

Di Crescenzo, Antonio :
Elementi di statistica Antonio Di Crescenzo, Luigi M. Ricciardi
Napoli : Liguori, 2000
ISBN 88 - 207 - 3052 - 9

1. Statistica matematica 2. Campionamento statistico I. Titolo

Ristampe:

9 8 7 6 5 4 3 2 1 0 2005 2004 2003 2002 2001 2000

Questo volume è stampato in Italia dalle Officine Grafiche Liguori - Napoli su carta inalterabile, priva di acidi, a pH neutro, conforme alle norme Iso 9706 ∞.

Indice

Premessa	ix
1 Generalità sul campionamento	1
1.1 Preliminari	3
1.2 Campionamento	7
1.3 Media e varianza campionarie	14
1.4 Campioni di popolazioni normali	14
2 Distribuzioni speciali	23
2.1 Distribuzione chi-quadrato	32
2.2 Distribuzione di Student	36
2.3 Distribuzione di Fisher	42
2.4 Distribuzione normale multivariata	48
2.5 Distribuzione normale inversa	48
3 Stima puntuale	55
3.1 Statistiche d'ordine	59
3.2 Stimatori corretti	68
3.3 Stimatori a varianza minima	83
3.4 Proprietà asintotiche degli stimatori	90
3.5 Statistiche sufficienti	102
3.6 Statistiche complete	107
3.7 Metodo della massima verosimiglianza	124
3.8 Metodo dei momenti	127
3.9 Stimatori di Bayes	127
4 Stima intervallare	137
4.1 Intervalli fiduciari	141
4.2 Metodo del cardine	144
4.3 Intervalli fiduciari per medie	144
4.3.1 Varianza nota	144
4.3.2 Varianza incognita	150
4.4 Differenze tra medie	151
4.4.1 Varianze note	151

4.4.2 Varianze incognite	153
4.5 Intervalli fiduciari per varianze	155
4.5.1 Media nota	155
4.5.2 Media incognita	157
4.6 Rapporti di varianze	158
4.7 Popolazioni di Bernoulli	160
4.8 Popolazioni esponenziali	162
4.9 Stime per quantili	165
5 Ipotesi statistiche	
5.1 Verifica delle ipotesi	171
5.2 Lemma di Neyman-Pearson	181
5.3 Rapporto di verosimiglianze	191
5.4 Test chi-quadrato	205
5.5 Differenze tra proporzioni	214
5.6 Tabelle di contingenza	218
6 Regressione e correlazione	
6.1 Regressione	225
6.2 Approssimazione ai minimi quadrati	229
6.3 Regressione non lineare	238
6.4 Stime puntuali	242
6.5 Regressione normale	248
6.6 Stima intervallare	250
6.7 Verifica di ipotesi	257
6.8 Adeguatezza del modello	263
6.9 Minimi quadrati pesati	266
6.10 Regressione polinomiale	272
6.11 Regressione lineare multivariata	276
6.12 Correlazione normale	279
7 Analisi della varianza	
7.1 Introduzione	285
7.2 Esperimenti ad un fattore	285
7.2.1 Stima puntuale dei parametri	289
7.2.2 Verifica di ipotesi	291
7.3 Il piano degli esperimenti	297
7.4 Esperimenti a due fattori	299
7.4.1 Stima puntuale dei parametri	300
7.4.2 Verifica di ipotesi	302
8 Rappresentazione dei dati	
8.1 Diagramma delle frequenze	307
8.2 Istogramma	312
8.3 Istogramma cumulativo	318

8.4 Teorema di Glivenko-Cantelli	325
Appendice A: Principali variabili casuali	329
A.1 Variabili casuali discrete	329
A.1.1 Variabile uniforme	329
A.1.2 Variabile di Bernoulli	330
A.1.3 Variabile binomiale	331
A.1.4 Variabile di Poisson	332
A.1.5 Variabile geometrica	333
A.1.6 Variabile di Pascal	334
A.1.7 Variabile binomiale negativa	335
A.1.8 Variabile ipergeometrica	336
A.1.9 Variabile multinomiale	337
A.2 Variabili casuali continue	338
A.2.1 Variabile uniforme	338
A.2.2 Variabile normale	339
A.2.3 Variabile esponenziale	340
A.2.4 Variabile gamma	341
A.2.5 Variabile beta	342
A.2.6 Variabile chi-quadrato	343
A.2.7 Variabile di Student	344
A.2.8 Variabile di Fisher	344
A.2.9 Variabile di Laplace	345
A.2.10 Variabile normale inversa	346
A.2.11 Variabile lognormale	347
Appendice B: Tabelle	349
Indice analitico	357

Premessa

L'insegnamento della Statistica dovrebbe, com'è consuetudine nei Paesi a più lunga tradizione nella didattica di questa materia, svilupparsi attraverso due successivi stadi: il primo da dedicarsi agli elementi di base, con enfasi sulle finalità e sugli aspetti applicativi; il secondo incentrato sulla formalizzazione e sulle estensioni attraverso l'uso di strumenti matematici più avanzati. Il troppo modesto spazio, che anche nell'ambito delle avvenute modificazioni statutarie viene di norma reso disponibile all'insegnamento della Statistica e delle sue applicazioni, costringe pertanto a individuare percorsi formativi che nell'ottemperare ad irrinunciabili esigenze di rigore nella formalizzazione matematica di questa materia, ne sottolineino al contempo la concretezza nella genesi e nelle applicazioni attraverso un esplicito, costante riferimento ad istanze specifiche, sia pur opportunamente schematizzate.

Questi Elementi, nati da una raccolta di lezioni ed esercitazioni rivolte a studenti del corso di laurea in matematica, e pertanto condizionati dai limiti sopra indicati, sono stati redatti in una forma e con un linguaggio atti a consentirne larga utilizzazione nei corsi di laurea della Facoltà di Scienze nonché nell'ambito di vari altri curricula, ad esempio dell'Ingegneria e dell'Economia.

Sia per limiti di spazio, che in quanto già ampiamente disponibili altrove, in questo volume non trovano posto esempi di utilizzazione di pacchetti di software statistico che pure nel corso dell'insegnamento vanno forniti con dovizia a integrazione delle lezioni di teoria e delle esercitazioni a carattere numerico. Al Lettore che intenda avvicinarsi alla Statistica attraverso questi Elementi si è poi voluto richiedere soltanto la conoscenza dei fondamenti del calcolo differenziale e integrale nonché dei tratti essenziali del calcolo delle probabilità. Si è così, ad esempio, deciso di non ricorrere all'uso delle funzioni di variabile complessa utilizzando quindi la funzione generatrice dei momenti in luogo della funzione caratteristica.

Per agevolare la lettura, contestuale richiamo viene effettuato di ogni specifico teorema o particolare formula della teoria della probabilità allorché per la prima volta ne viene fatto ricorso. È stata inoltre inclusa in un'appendice una sintesi delle principali proprietà delle variabili casuali discrete e assolutamente continue con l'indicazione dei rispettivi parametri descrittivi, tra cui i momenti (centrali o intorno all'origine sulla base di criteri di semplicità), e delle funzioni generatrici, quando non eccessivamente complicate. Ancora in un'appendice sono altresì riportate tabelle relative alle principali distribuzioni utilizzate nel corso del volume.

La presentazione della materia, conformemente con i canoni della matematica, avviene attraverso la sua articolazione in definizioni, lemmi, teoremi, proposizioni, corollari e osservazioni, con le relative dimostrazioni quando non ovvie. Tuttavia, occasionalmente, ad

evitare eccessivi appesantimenti qualche dimostrazione o è stata omessa ovvero ci si è limitati soltanto ad accennarla. Al fine di sottolineare l'applicabilità immediata dei risultati via via presentati, e per meglio fissare i concetti, viene effettuato frequente uso di esempi (se ne contano oltre cento nel corso del volume); ad essi, in contrasto con una diffusa consuetudine, non si è destinato corpo minore di stampa a sottolinearne la rilevanza didattica. I risultati numerici, approssimati alla quarta cifra decimale, sono stati ottenuti mediante una normale calcolatrice tascabile ad uso scientifico.

Nella scelta degli argomenti da includere è stato necessario tener conto di prefissati limiti di spazio. Il criterio di massima adottato è consistito nell'escludere argomenti non indispensabili in un primo avvicinamento alla materia, quali test non parametrici e analisi dei clusters, e nel rinunciare alla discussione di tematiche, come il trattamento delle serie storiche, che richiedono conoscenze non superficiali della teoria dei processi stocastici. Il carattere introduttivo di questi Elementi ha pertanto suggerito di enfatizzare la considerazione di problemi monoparametrici e la descrizione di metodi e strumenti di statistica univariata nella certezza che di estensioni o generalizzazioni potrà più efficacemente impadronirsi il Lettore in successivi momenti, quando abbia già fatto propria la conoscenza dei fondamenti del metodo statistico e dei principi matematici sui quali esso poggia.

Nel corso della stesura gli Autori hanno beneficiato di numerosi suggerimenti da parte della Prof.ssa Amelia G. Nobile. A lei, ed alle Dott.sse Elvira Di Nardo, Maria Longobardi ed Enrica Pirozzi che si sono prestate ad una paziente opera di lettura critica di una precedente versione correggendo imprecisioni e refusi tipografici, va il loro ringraziamento.

Gli Autori:
Napoli, 3 aprile 2000

Capitolo 1 Generalità sul campionamento

1.1 Preliminari

Il calcolo delle probabilità trova larga applicazione in statistica sia allo scopo di interpretare e di prevedere fenomeni a carattere aleatorio, sia come mezzo di verifica della validità dei risultati cui la teoria della probabilità conduce. La statistica può essenzialmente riguardarsi come un insieme di metodi di natura logica e matematica atti a raccogliere, elaborare, analizzare e interpretare dati con la finalità di descrivere fenomeni collettivi o di estendere la descrizione di taluni fenomeni osservati ad altri fenomeni dello stesso tipo non ancora osservati, prevedendone l'evoluzione.

Uno dei problemi fondamentali della statistica consiste nell'analisi di dati sperimentali e nella ricerca della funzione di distribuzione da associare ad una o più variabili casuali che si assume abbiano generato i dati in esame. La ricerca della soluzione di un tale problema si articola quindi in varie fasi, tra cui le seguenti:

- (a) costruzione di un modello ipotizzando che variabili casuali di un certo tipo siano responsabili della generazione dei dati osservati e, successivamente, determinazione della funzione di distribuzione di tali variabili;
- (b) specificazione, sulla base dei dati osservati, dei parametri che compaiono nella funzione di distribuzione;
- (c) valutazione della validità del modello tramite confronto con dati sperimentali.

Per quanto riguarda la costruzione del modello (punto (a)) e, quindi, per pervenire a delle ipotesi sul tipo di distribuzione da utilizzare, è necessario disporre di un certo numero di indicazioni come illustrato, a titolo esemplificativo, dalle seguenti semplici considerazioni. Supponiamo dunque che i valori assunti da una successione X_0, X_1, \dots di variabili casuali siano determinati dall'azione di un'altra successione di variabili casuali, Z_1, Z_2, \dots , indipendenti e identicamente distribuite, che chiameremo "cause". Precisamente, supponiamo che

X_0 sia arbitraria, e che i valori di X_1, X_2, \dots siano regolati dalle equazioni

$$X_{i+1} = X_i + h(X_i) Z_{i+1} \quad (i = 0, 1, \dots), \quad (1.1)$$

dove h è un'opportuna funzione positiva la cui specificazione caratterizza il modello in considerazione. La (1.1) descrive dunque una situazione in cui il valore della generica variabile X_{i+1} dipende in modo diretto dalla precedente variabile X_i e dalla causa Z_{i+1} . Dalla (1.1) si trae:

$$Z_{i+1} = \frac{X_{i+1} - X_i}{h(X_i)} \quad (i = 0, 1, \dots),$$

da cui segue:

$$Z_1 + Z_2 + \dots + Z_k = \sum_{i=0}^{k-1} \frac{X_{i+1} - X_i}{h(X_i)}. \quad (1.2)$$

Notiamo che se le variabili casuali Z_1, Z_2, \dots sono dotate di media α e varianza β finite e se k è molto grande, per il teorema centrale del limite¹ il primo membro della (1.2) è approssimativamente una variabile casuale normale di media $\mu = k\alpha$ e varianza $\sigma^2 = k\beta$. Dalla (1.2) segue allora che anche

$$Y \stackrel{\text{def}}{=} \sum_{i=0}^{k-1} \frac{X_{i+1} - X_i}{h(X_i)}$$

è una variabile casuale normale di media μ e varianza σ^2 . Se indichiamo con x_i il generico valore assunto da X_i , il generico valore della variabile Y può approssimativamente scriversi al seguente modo:

$$y \stackrel{\text{def}}{=} \sum_{i=0}^{k-1} \frac{x_{i+1} - x_i}{h(x_i)} \approx \int_{x_0}^x \frac{1}{h(z)} dz \stackrel{\text{def}}{=} g(x), \quad (1.3)$$

purché k sia molto grande e ciascun termine $x_{i+1} - x_i$ sia molto piccolo. In conclusione, nelle sopramenzionate ipotesi la (1.3) conduce alla relazione $Y = g(X)$ tra variabili casuali, con Y variabile normale di media μ e varianza σ^2 , dotata quindi di densità di probabilità:

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(y-\mu)^2}{2\sigma^2} \right] \quad (y \in \mathbb{R}).$$

Nel caso in cui g è una funzione strettamente monotona, dotata quindi di funzione inversa g^{-1} , facendo uso delle note leggi di trasformazione di variabili casuali la densità di probabilità $f_X(x)$ di $X = g^{-1}(Y)$ si ottiene da $f_Y(y)$ al seguente modo:

$$f_X(x) = \left| \frac{d}{dx} g(x) \right| f_Y(g(x)) = \left| \frac{d}{dx} g(x) \right| \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{|g(x)-\mu|^2}{2\sigma^2} \right\}, \quad (1.4)$$

¹Ricordiamo che, nella sua enunciazione più semplice, il teorema centrale del limite afferma la convergenza in distribuzione alla variabile casuale normale standard della successione $\{(\sum_{i=1}^n X_i - n\mu)/\sigma\sqrt{n}\}_n$ delle somme standardizzate di variabili casuali indipendenti $\{X_i\}_n$, identicamente distribuite a media μ e varianza σ^2 finite:

$$\lim_{n \rightarrow \infty} P \left(\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \leq x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy \quad (x \in \mathbb{R}).$$

1.2. CAMPIONAMENTO

dove la (1.4) è valida per $x \in (g^{-1}(-\infty), g^{-1}(\infty))$ se g è crescente, per $x \in (g^{-1}(\infty), g^{-1}(-\infty))$ se g è decrescente, mentre $f_X(x)$ è nulla altrimenti. Esaminiamo ora il caso $h(z) = z$ che, come mostra la (1.1), si riferisce ad una situazione in cui l'effetto di una data causa Z_{i+1} è proporzionale al valore di X_i . Dalla (1.3), scegliendo $x_0 = 1$, segue $g(x) = \ln x$; dalla (1.4) si ricava poi che la variabile casuale $Y = g(X) \equiv \ln X$ è approssimativamente normale, e che quindi $X = g^{-1}(Y) \equiv e^Y$ ha densità di probabilità $f_X(x)$ approssimativamente lognormale:

$$f_X(x) \simeq \frac{1}{x\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(\ln x - \mu)^2}{2\sigma^2} \right], \quad x > 0.$$

Le considerazioni appena svolte costituiscono un semplice esempio di costruzione di un modello probabilistico, quello espresso dalla densità di probabilità (1.4). Una volta individuato il tipo di modello, la determinazione dei parametri coinvolti (punto (b)) e la valutazione della sua validità (punto (c)) possono essere effettuate utilizzando i dati disponibili ed i metodi della cosiddetta inferenza statistica dei quali si dirà nel seguito (cfr. § 1.3).

1.2 Campionamento

Fondamentale, in statistica, è il concetto di *campionamento*. Il campionamento fa riferimento ad una assegnata *popolazione*, ossia ad un assegnato insieme di oggetti o di individui (non necessariamente persone). Ad esempio, ci si può riferire alla popolazione consistente nelle automobili prodotte in una giornata lavorativa in una certa fabbrica (popolazione discreta) oppure alla popolazione costituita dagli intervalli di tempo intercorrenti tra coppie di successive chiamate telefoniche ricevute da un certo centralino (popolazione continua). Introdurremo ora due concetti fondamentali in statistica attraverso le definizioni di campione casuale e di variabile casuale genitrice; questi troveranno poi subito una più rigorosa giustificazione — sia pure non indispensabile per la comprensione del seguito — che passeremo ad illustrare approfondendo la nozione di campionamento secondo il punto di vista di H.G. Tucker.

Definizione 1.2.1 Siano X_1, X_2, \dots, X_n variabili casuali osservabili² indipendentemente distribuite. Il vettore (X_1, X_2, \dots, X_n) si dice campione casuale di taglia n estratto da una popolazione infinita o da una popolazione finita con rimpiazzamento.

La funzione di distribuzione del campione, ossia la funzione di distribuzione congiunta di (X_1, X_2, \dots, X_n) è quindi:

$$F(x_1, x_2, \dots, x_n) = \prod_{i=1}^n F_X(x_i),$$

dove $F_X(x_i)$ denota la comune funzione di distribuzione delle variabili casuali X_i ($i = 1, 2, \dots, n$).

Definizione 1.2.2 Si dice variabile casuale genitrice di un campione casuale (X_1, X_2, \dots, X_n) una variabile X avente la medesima funzione di distribuzione delle variabili che costituiscono il campione casuale.

²Si dicono "osservabili" le variabili casuali i cui valori possono essere realmente osservati.

Definizione 1.2.3 Dato un campione casuale (X_1, X_2, \dots, X_n) di variabile genitrice X , le variabili casuali X_1, X_2, \dots, X_n si dicono costituire n osservazioni indipendenti di X , e X_k ($k = 1, 2, \dots, n$) prende il nome di k -esima osservazione.

Definizione 1.2.4 Dato un campione casuale (X_1, X_2, \dots, X_n) , ogni n -upla (x_1, x_2, \dots, x_n) di reali rappresentanti valori effettivamente assunti dalle variabili casuali X_1, X_2, \dots, X_n viene detta realizzazione del campione casuale.

Veniamo ora al preannunciato approfondimento e, a tal fine, riferiamoci ad una certa popolazione, la cui natura non occorre qui specificare. Denoteremo la popolazione con Ω , e con $\omega \in \Omega$ indicheremo il suo generico elemento. Ad ogni elemento della popolazione associeremo un reale che ne indica una qualche caratteristica. Così, se Ω denota l'insieme delle automobili dell'esempio di cui sopra, il reale associato potrebbe indicare il numero di difetti riscontrati in ciascuna auto. In altri termini, si considera una funzione $X: \Omega \rightarrow \mathbb{R}$ che ad ogni elemento, o "individuo", $\omega \in \Omega$ associa un reale $X(\omega)$. Pertanto Ω sarà riguardato come lo spazio campione di uno spazio di probabilità (Ω, \mathcal{F}, P) e si assumerà che X sia una variabile casuale ivi definita. D'ora innanzi ci riferiremo ad X come alla *variabile casuale genitrice*³ (cfr. Definizione 1.2.2).

Nel seguito, a meno di esplicito avviso contrario, considereremo sempre il caso di campionamento con rimpiazzamento. Supponiamo dunque di effettuare un *campionamento n volte con rimpiazzamento*. Ciò significa che si sceglie un elemento ω_1 a caso da Ω , lo si rimpiazza, si sceglie un secondo elemento ω_2 a caso da Ω , lo si rimpiazza, e così via fino a scegliere un elemento ω_n a caso da Ω . In tal modo, mediante estrazioni ripetute con rimpiazzamento, ossia mediante prove ripetute, si è costruita la n -upla $(\omega_1, \omega_2, \dots, \omega_n)$ di elementi, o individui, di Ω . Questa può essere riguardata come elemento di un nuovo spazio campione $\Omega^{(n)}$ i cui elementi sono tutte le n -uple ordinate di elementi, anche ripetuti, di Ω . Con $\omega^{(n)} \stackrel{\text{def}}{=} (\omega_1, \omega_2, \dots, \omega_n)$ indicheremo il generico elemento di $\Omega^{(n)}$. Definiamo ora una σ -algebra $\mathcal{F}^{(n)}$ di eventi costruita a partire da eventi elementari di questo nuovo spazio campione $\Omega^{(n)}$. Per ogni n -upla (E_1, E_2, \dots, E_n) , tale che $E_i \in \mathcal{F}$ per $i = 1, 2, \dots, n$, definiamo il seguente evento:

$$E_1 \times E_2 \times \dots \times E_n = \{\omega^{(n)} \in \Omega^{(n)} : \omega_1 \in E_1, \omega_2 \in E_2, \dots, \omega_n \in E_n\}.$$

Si tratta dunque dell'evento composto " E_1 si verifica alla prima prova, E_2 si verifica alla seconda prova, ..., E_n si verifica all' n -esima prova". Un tale evento verrà detto *rettangolare*. Noi richiederemo che $\mathcal{F}^{(n)}$ contenga tutti questi eventi rettangolari. Precisamente, detto $\mathcal{R}^{(n)}$ l'insieme di tutti questi eventi rettangolari, definiamo $\mathcal{F}^{(n)}$ come la *minima* σ -algebra di sottoinsiemi di $\Omega^{(n)}$ contenente $\mathcal{R}^{(n)}$, ossia identifichiamo $\mathcal{F}^{(n)}$ con l'intersezione di tutte le σ -algebre di sottoinsiemi di $\Omega^{(n)}$ che contengono $\mathcal{R}^{(n)}$. Si noti che affinché questa definizione abbia senso occorre dimostrare (i) che l'intersezione di qualsiasi collezione di σ -algebre di sottoinsiemi di $\Omega^{(n)}$ è una σ -algebra ed inoltre (ii) che la collezione di σ -algebre contenenti $\mathcal{R}^{(n)}$ non è vuota. Il punto (i) è dimostrato dal seguente lemma.

Lemma 1.2.1 Sia $\{\mathcal{A}_\lambda, \lambda \in \Lambda\}$ una collezione non vuota di σ -algebre di sottoinsiemi di $\Omega^{(n)}$. Allora $\bigcap_{\lambda \in \Lambda} \mathcal{A}_\lambda$ è una σ -algebra.

³Va menzionato che talora per semplicità, ma impropriamente, i termini "variabile casuale genitrice" e "popolazione" vengono usati in maniera intercambiabile.

1.2. CAMPIONAMENTO

Dim. Osserviamo anzitutto che poiché $\Omega^{(n)} \in \mathcal{A}_\lambda$ per ogni $\lambda \in \Lambda$, $\Omega^{(n)}$ appartiene anche a $\bigcap_{\lambda \in \Lambda} \mathcal{A}_\lambda$. Sia $B \in \bigcap_{\lambda \in \Lambda} \mathcal{A}_\lambda$, così che $B \in \mathcal{A}_\lambda$ per ogni $\lambda \in \Lambda$. Poiché ogni \mathcal{A}_λ è una σ -algebra, $B \in \mathcal{A}_\lambda$ implica che $\bar{B} \in \mathcal{A}_\lambda$ per ogni $\lambda \in \Lambda$. Quindi $\bar{B} \in \bigcap_{\lambda \in \Lambda} \mathcal{A}_\lambda$, e pertanto $\bigcap_{\lambda \in \Lambda} \mathcal{A}_\lambda$ è chiusa rispetto all'operazione di complementazione. Mostriamo che $\bigcap_{\lambda \in \Lambda} \mathcal{A}_\lambda$ è chiusa rispetto all'operazione di unione numerabile. Sia $B_k \in \bigcap_{\lambda \in \Lambda} \mathcal{A}_\lambda$ per $k = 1, 2, \dots$. Allora per ogni $\lambda \in \Lambda$ si ha $B_k \in \mathcal{A}_\lambda$ per ogni k . Poiché \mathcal{A}_λ è una σ -algebra per ogni $\lambda \in \Lambda$, $B_k \in \mathcal{A}_\lambda$ per $k = 1, 2, \dots$ implica $\bigcup_{k=1}^{\infty} B_k \in \mathcal{A}_\lambda$. Pertanto, $\bigcup_{k=1}^{\infty} B_k \in \bigcap_{\lambda \in \Lambda} \mathcal{A}_\lambda$, e quindi $\bigcap_{\lambda \in \Lambda} \mathcal{A}_\lambda$ è chiusa rispetto all'unione numerabile. Dunque, $\bigcap_{\lambda \in \Lambda} \mathcal{A}_\lambda$ contiene $\Omega^{(n)}$, contiene il complementare di ogni suo elemento ed è chiusa rispetto all'unione numerabile; quindi è una σ -algebra. ■

Per dimostrare il punto (ii) occorre mostrare che la collezione delle σ -algebre contenenti $\mathcal{R}^{(n)}$ è non vuota. Ciò è evidente dal momento che l'insieme delle parti di $\Omega^{(n)}$ è una σ -algebra e contiene $\mathcal{R}^{(n)}$.

A questo stadio siamo passati dallo spazio di probabilità (Ω, \mathcal{F}, P) allo spazio probabilizzabile $(\Omega^{(n)}, \mathcal{F}^{(n)})$. Ci prefiggiamo ora di definire una misura di probabilità $P^{(n)}$ in $\mathcal{F}^{(n)}$. A tal fine, per ogni evento rettangolare $E_1 \times E_2 \times \dots \times E_n$ in $\mathcal{F}^{(n)}$ poniamo:

$$P^{(n)}(E_1 \times E_2 \times \dots \times E_n) = \prod_{k=1}^n P(E_k). \quad (1.5)$$

Il motivo è chiarito dalle considerazioni che seguono. L'evento $E_1 \times E_2 \times \dots \times E_n$ è un elemento di $\Omega^{(n)}$ così definito:

$$E_1 \times E_2 \times \dots \times E_n = \bigcap_{k=1}^n E_{k,k}^{(n)},$$

dove si è posto:

$$E_{1,1}^{(n)} = E_1 \times \Omega \times \Omega \times \dots \times \Omega \quad (E_1 \text{ si verifica alla prima prova})$$

$$E_{2,2}^{(n)} = \Omega \times E_2 \times \Omega \times \dots \times \Omega \quad (E_2 \text{ si verifica alla seconda prova})$$

⋮

$$E_{n,n}^{(n)} = \Omega \times \Omega \times \dots \times \Omega \times E_n \quad (E_n \text{ si verifica alla } n\text{-esima prova}).$$

Si noti che gli $E_{k,k}^{(n)}$ sono elementi di $\Omega^{(n)}$. Poiché nel campionamento con rimpiazzamento i risultati di prove distinte vanno riguardati come indipendenti, è ragionevole richiedere che risult

$$P^{(n)}(E_{k,k}^{(n)}) = P(E_k) \quad (k = 1, 2, \dots, n).$$

Ne segue:

$$P^{(n)}(E_1 \times E_2 \times \dots \times E_n) = P^{(n)}\left(\bigcap_{k=1}^n E_{k,k}^{(n)}\right) = \prod_{k=1}^n P^{(n)}(E_{k,k}^{(n)}) = \prod_{k=1}^n P(E_k),$$

che coincide con la (1.5). Abbiamo dunque definito $P^{(n)}$ su $\mathcal{R}^{(n)}$; quest'ultimo — se, come supponiamo, \mathcal{F} non si riduce alla σ -algebra banale $\{\emptyset, \Omega\}$ — è un sottoinsieme proprio di

$\mathcal{F}^{(n)}$. La misura $P^{(n)}$ può essere allora estesa ad $\mathcal{F}^{(n)}$ in maniera unica,⁴ nel senso che esiste una e una sola misura di probabilità $\Pi^{(n)}$ definita su $\mathcal{F}^{(n)}$ che si riduce a $P^{(n)}$ su $\mathcal{R}^{(n)}$. In conclusione, campionando n volte con rimpiazzamento siamo passati dallo spazio di probabilità (Ω, \mathcal{F}, P) allo spazio di probabilità $(\Omega^{(n)}, \mathcal{F}^{(n)}, \Pi^{(n)})$, dove la misura di probabilità $\Pi^{(n)}$ si riduce alla (1.5) su ogni evento rettangolare.

Nel seguito per semplificare la notazione denoteremo $\Pi^{(n)}$ ancora con P , riferendoci così allo spazio di probabilità $(\Omega^{(n)}, \mathcal{F}^{(n)}, P)$, con P estensione di $P^{(n)}$ da $\mathcal{R}^{(n)}$ a $\mathcal{F}^{(n)}$.

Con riferimento al campionamento n volte con rimpiazzamento, indichiamo con X_1, X_2, \dots, X_n i reali associati rispettivamente alla prima, alla seconda, ..., alla n -esima prova (o osservazione). In altri termini, per ogni

$$\omega^{(n)} = (\omega_1, \omega_2, \dots, \omega_n) \in \Omega^{(n)}$$

poniamo:

$$X_k(\omega^{(n)}) = \tilde{X}(\omega_k) \quad (k = 1, 2, \dots, n),$$

dove $X: \Omega \rightarrow \mathbb{R}$ denota la variabile casuale genitrice prima introdotta. Sussiste il seguente fondamentale teorema:

Teorema 1.2.1 *Le funzioni $X_k: \Omega^{(n)} \rightarrow \mathbb{R}$ ($k = 1, 2, \dots, n$) sono variabili casuali indipendenti, ciascuna avente funzione di distribuzione coincidente con quella di X .*

Dim. Poiché per ogni k ($k = 1, 2, \dots, n$) e per ogni $x_k \in \mathbb{R}$ risulta

$$\begin{aligned} \{X_k \leq x_k\} &\equiv \{\omega^{(n)} \in \Omega^{(n)} : X_k(\omega^{(n)}) \leq x_k\} \\ &= \Omega \times \dots \times \Omega \times \{\omega_k \in \Omega : X(\omega_k) \leq x_k\} \times \Omega \times \dots \times \Omega \in \mathcal{F}^{(n)}, \end{aligned} \quad (1.6)$$

dove $\{\omega_k \in \Omega : X(\omega_k) \leq x_k\}$ è situato al k -esimo posto, gli insiemi a secondo membro sono misurabili, e quindi $X_k: \Omega^{(n)} \rightarrow \mathbb{R}$ è una variabile casuale. Inoltre, per $k = 1, 2, \dots, n$ e per ogni $x \in \mathbb{R}$, dalle (1.5) e (1.6) si trae:

$$\begin{aligned} P^{(n)}\{X_k \leq x\} &= P(\Omega) \cdot \dots \cdot P(\Omega) \cdot P\{X \leq x\} \cdot P(\Omega) \cdot \dots \cdot P(\Omega) \\ &= P\{X \leq x\}, \end{aligned} \quad (1.7)$$

così che X_k ha la stessa distribuzione di X . Infine, poiché dalla (1.6) si ha

$$\bigcap_{k=1}^n \{X_k \leq x_k\} = \{X \leq x_1\} \times \dots \times \{X \leq x_n\},$$

facendo uso delle (1.5) e (1.7) si ricava:

$$\begin{aligned} P^{(n)}\left(\bigcap_{k=1}^n \{X_k \leq x_k\}\right) &= P^{(n)}(\{X \leq x_1\} \times \dots \times \{X \leq x_n\}) \\ &= \prod_{k=1}^n P\{X \leq x_k\} = \prod_{k=1}^n P^{(n)}\{X_k \leq x_k\}, \end{aligned}$$

il che prova l'indipendenza di X_1, X_2, \dots, X_n . ■

⁴Ciò è conseguenza del teorema di Carathéodory, ben noto in teoria della probabilità.

Sulla base di queste considerazioni appare ora ben giustificato l'aver chiamate "osservazioni indipendenti" della variabile genitrice X le variabili casuali X_1, X_2, \dots, X_n . Il vettore (X_1, X_2, \dots, X_n) costituisce così un campione casuale⁵ di taglia n di variabile casuale genitrice X tratto dalla popolazione Ω . Ogni sua realizzazione, ossia ogni n -upla (x_1, x_2, \dots, x_n) di reali rappresentanti valori effettivamente assunti rispettivamente dalle variabili casuali X_1, X_2, \dots, X_n , viene talora detta anche " n -upla di misure," o " n -upla di osservazioni," con rimpiazzamento della variabile casuale X .

Si noti che le n -uple ordinate $\omega^{(n)} = (\omega_1, \omega_2, \dots, \omega_n) \in \Omega^{(n)}$ possono contenere elementi ripetuti di Ω . Ciò non accade se il campionamento avviene senza rimpiazzamento, nel qual caso $\omega^{(n)}$ consisterebbe di n elementi, o individui, *distinti* della popolazione Ω , ed $\Omega^{(n)}$ sarebbe costituita dall'insieme di tutte le n -uple siffatte.]

1.3 Media e varianza campionarie

Di particolare importanza è la cosiddetta *inferenza statistica*. Questa ha lo scopo di estendere conclusioni quantitative ricavate dall'esame di un campione alla popolazione da cui il campione è stato estratto. Esistono due metodi fondamentali di inferenza: *stima dei parametri e verifica delle ipotesi*. Va inoltre menzionato che problemi in cui la forma della distribuzione della variabile casuale genitrice è specificata a meno di un insieme di parametri incogniti sono detti problemi di *inferenza parametrica*, mentre quelli in cui non si hanno informazioni sulla forma di tale distribuzione sono detti problemi di *inferenza non parametrica*.

La stima dei parametri si prefigge la determinazione dei valori incogniti dei parametri di una popolazione (media, varianza, ecc.) per tramite di corrispondenti quantità (media campionaria, varianza campionaria, ecc.) ricavate dal campione estratto dalla popolazione. Si possono usare stime puntuale o stime intervallari. Si parla di *stima puntuale* quando si stima un parametro incognito della popolazione usando un singolo valore reale. Alla stima puntuale di un parametro di una popolazione (costituita, come si è detto, da un solo valore) si preferisce sovente sostituire un intervallo di valori, detto *intervallo fiduciario*; in tal caso si cerca di determinare, in base ai dati del campione, due limiti entro i quali sia compreso il valore incognito del parametro con una prefissata probabilità detta *coefficiente di fiducia*.

La verifica delle ipotesi tratta del problema di stabilire se un campione, di cui sia stata osservata una realizzazione, possa essere stato generato da una ipotizzata popolazione.

Va subito detto che affinché siano valide le conclusioni tratte mediante i metodi dell'inferenza statistica occorre che i campioni siano scelti in modo da essere rappresentativi della popolazione da cui vengono estratti. Va da sé che essenziale ai fini dell'attendibilità delle conclusioni cui la statistica conduce è proprio l'adeguatezza del campione a caratterizzare la popolazione in esame. Particolarmente delicato è pertanto il compito di valutare la significatività del campione per la finalità che ci si prefigge.

Nel seguito, secondo la vigente consuetudine, utilizzeremo il termine "statistica" anche per denotare variabili casuali costruite a partire da campioni casuali. Termini alternativi talora ricorrenti nella letteratura sono anche quelli di "riassunti campionari" o, più semplicemente, "riassunti". Diamo dunque la seguente definizione:

⁵Nel seguito useremo talvolta il termine "campione" in luogo di "campione casuale".

CAPITOLO 1. GENERALITÀ SUL CAMPIONAMENTO

Definizione 1.3.1 Una statistica $g(X_1, X_2, \dots, X_n)$ è una variabile casuale osservabile che non dipende da parametri incogniti e che è funzione delle variabili casuali osservabili X_1, X_2, \dots, X_n costituenti un campione casuale.

Pertanto, se (X_1, X_2, \dots, X_n) è un campione casuale, le variabili casuali

$$\frac{1}{n} \sum_{i=1}^n X_i, \quad \sum_{i=1}^n X_i^2, \quad \prod_{i=1}^n X_i$$

costituiscono esempi di statistiche. Inoltre, se la variabile casuale genitrice X è dotata di media μ e varianza σ^2 finite, le variabili casuali

$$\frac{X_1^2}{\sigma^2}, \quad X_2 - \mu, \quad \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2$$

non sono ad esempio statistiche se μ e σ^2 sono incognite, mentre vanno riguardate come delle statistiche se i valori di μ e σ^2 sono noti.

Dato che le statistiche sono variabili casuali dipendenti dal campione, i loro valori cambiano al variare della realizzazione osservata. È pertanto consuetudine riferirsi alle distribuzioni delle statistiche come alle *distribuzioni campionarie*.

Statistiche tipiche sono la media campionaria e la varianza campionaria qui di seguito definite.

Definizione 1.3.2 Se (X_1, X_2, \dots, X_n) è un campione casuale, la statistica

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

è detta media campionaria, mentre la statistica

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (n > 1)$$

è detta varianza campionaria.

Ulteriori statistiche sono i cosiddetti *momenti campionari* che passiamo a definire.

Definizione 1.3.3 Se (X_1, X_2, \dots, X_n) è un campione casuale e k è un intero positivo la statistica

$$\bar{X}^{(k)} = \frac{1}{n} \sum_{i=1}^n X_i^k$$

è detta momento campionario di ordine k .

Con \bar{x} , s^2 e $\bar{x}^{(k)}$ si denotano i valori assunti rispettivamente dalla media campionaria, dalla varianza campionaria e dal momento campionario di ordine k di un campione casuale in corrispondenza di una realizzazione osservata (x_1, x_2, \dots, x_n) . Risulta dunque:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \bar{x}^{(k)} = \frac{1}{n} \sum_{i=1}^n x_i^k. \quad (1.8)$$

La quantità $\bar{x}^{(k)}$ è anche detta *momento empirico intorno all'origine di ordine k* .

1.3. MEDIA E VARIANZA CAMPIONARIE

Esempio 1.3.1 Un'industria produce lampadine che è ragionevole ritenere identiche. Per un test di affidabilità si preleva un lotto di 14 lampadine che vengono accese simultaneamente. Le durate di ciascuna delle lampadine possono riguardarsi come variabili casuali indipendenti costituenti un campione casuale (X_1, X_2, \dots, X_n) di taglia $n = 14$. Le durate, in ore, misurate risultano essere le seguenti:

$$(220, 195, 235, 192, 215, 201, 241, 187, 197, 213, 232, 192, 211, 206).$$

Esse costituiscono la realizzazione osservata del campione casuale considerato. Essendo

$$n = 14, \quad \sum_{i=1}^n x_i = 2937, \quad \sum_{i=1}^n x_i^2 = 620013$$

dalle (1.8) si conclude che la media campionaria e la varianza campionaria del campione casuale considerato assumono i seguenti valori

$$\bar{x} = 209.78, \quad s^2 = 297.87$$

in corrispondenza della realizzazione osservata. ◆

Dalle Definizioni 1.3.2 e 1.3.3 appare evidente che è possibile esprimere la media e la varianza campionaria attraverso i momenti campionari del primo e del secondo ordine. Sussiste infatti il seguente risultato:

Proposizione 1.3.1 Dato un campione casuale (X_1, X_2, \dots, X_n) la media campionaria e la varianza campionaria sono così esprimibili:

$$\bar{X} = \bar{X}^{(1)}, \quad S^2 = \frac{n}{n-1} (\bar{X}^{(2)} - \bar{X}^2).$$

Dim. L'uguaglianza $\bar{X} = \bar{X}^{(1)}$ segue direttamente dalla definizione di media campionaria e di momento campionario del primo ordine. Inoltre, dalla definizione di varianza campionaria e di momento campionario del secondo ordine si ha:

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i^2 - 2X_i\bar{X} + \bar{X}^2) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + n\bar{X}^2 \right) \\ &= \frac{1}{n-1} \left[\sum_{i=1}^n X_i^2 - 2\bar{X}(n\bar{X}) + n\bar{X}^2 \right] \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right) = \frac{n}{n-1} (\bar{X}^{(2)} - \bar{X}^2), \end{aligned}$$

il che completa la dimostrazione. ◆

Determiniamo ora valore medio e varianza delle statistiche media campionaria e varianza campionaria.

Teorema 1.3.1 Se (X_1, X_2, \dots, X_n) è un campione casuale estratto da una popolazione dotata di valore medio μ , varianza σ^2 e momento centrale del quart'ordine μ_4 , risulta:

$$E(\bar{X}) = \mu, \quad D^2(\bar{X}) = \frac{\sigma^2}{n}, \quad (1.9)$$

$$E(S^2) = \sigma^2, \quad D^2(S^2) = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right) \quad (n > 1). \quad (1.10)$$

Dim. Per dimostrare le (1.9) osserviamo anzitutto che dalla definizione di media campionaria si ha:

$$E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right).$$

Poiché il valore medio della somma di variabili casuali è uguale alla somma dei valori medi delle variabili stesse, segue:

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu,$$

avendo fatto uso dell'ipotesi $E(X_i) = \mu$ per $i = 1, 2, \dots, n$. Ricordiamo ora che se X_1, X_2, \dots, X_n sono variabili casuali indipendenti e se c_1, c_2, \dots, c_n sono costanti arbitrarie, risulta:

$$\begin{aligned} D^2\left(\sum_{i=1}^n c_i X_i\right) &= E\left\{\left[\sum_{i=1}^n c_i (X_i - \mu)\right]^2\right\} \\ &= \sum_{i,j=1}^n c_i c_j E[(X_i - \mu)(X_j - \mu)] \\ &= \sum_{\substack{i,j=1 \\ i=j}}^n c_i c_j \text{cov}(X_i, X_j) + \sum_{\substack{i,j=1 \\ i \neq j}}^n c_i c_j \text{cov}(X_i, X_j). \end{aligned}$$

Notiamo che $\text{cov}(X_i, X_i) = D^2(X_i)$ per $i = 1, 2, \dots, n$. Inoltre, poiché le variabili che costituiscono un campione casuale sono indipendenti, si ha $\text{cov}(X_i, X_j) = 0$ per $i, j = 1, 2, \dots, n, i \neq j$; così che

$$D^2\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i^2 D^2(X_i).$$

Per $c_1 = c_2 = \dots = c_n = 1/n$ segue allora:

$$D^2(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n D^2(X_i) = \frac{\sigma^2}{n},$$

dato che $D^2(X_i) = \sigma^2$ per $i = 1, 2, \dots, n$. La (1.9) è così dimostrata. Dimostriamo ora la prima delle (1.10). Dalla definizione di varianza campionaria si trae:

$$E(S^2) = \frac{1}{n-1} E\left\{\sum_{i=1}^n [(X_i - \mu) - (\bar{X} - \mu)]^2\right\}$$

1.3. MEDIA E VARIANZA CAMPIONARIE

$$\begin{aligned} &= \frac{1}{n-1} \sum_{i=1}^n E[(X_i - \mu)^2 + (\bar{X} - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu)] \\ &= \frac{1}{n-1} \left\{ \sum_{i=1}^n D^2(X_i) + n D^2(\bar{X}) - 2 \sum_{i=1}^n E[(X_i - \mu)(\bar{X} - \mu)] \right\} \\ &= \frac{1}{n-1} \left(n \sigma^2 + n \frac{\sigma^2}{n} - 2n \frac{\sigma^2}{n} \right) = \sigma^2, \end{aligned}$$

avendo fatto uso della relazione:

$$\begin{aligned} \sum_{i=1}^n E[(X_i - \mu)(\bar{X} - \mu)] &= E\left[\sum_{i=1}^n (X_i - \mu)(\bar{X} - \mu)\right] \\ &= E\left[(\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu)\right] = E[(\bar{X} - \mu)n(\bar{X} - \mu)] \\ &= n D^2(\bar{X}). \end{aligned}$$

Dimostriamo la seconda delle (1.10). Osserviamo preliminarmente che, se si pone $Y_i = X_i - \mu$ per $i = 1, 2, \dots, n$, dalla definizione di S^2 si ha:

$$\begin{aligned} S^2 &\equiv \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \sum_{i=1}^n \left[X_i - \mu - \frac{1}{n} \sum_{j=1}^n (X_j - \mu) \right]^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n \left(Y_i - \frac{1}{n} \sum_{j=1}^n Y_j \right)^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2, \end{aligned}$$

dove si è posto $\bar{Y} = \sum_{j=1}^n Y_j / n$. Osserviamo poi che, procedendo analogamente a quanto effettuato nella dimostrazione della Proposizione 1.3.1, si ottiene:

$$\begin{aligned} S^2 &= \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n Y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n Y_i \right)^2 \right] \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n Y_i^2 - \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n Y_i Y_j \right) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n Y_i^2 - \frac{1}{n} \sum_{i=1}^n Y_i^2 - \frac{2}{n} \sum_{i < j} Y_i Y_j \right) \\ &= \frac{1}{n} \sum_{i=1}^n Y_i^2 - \frac{2}{n(n-1)} \sum_{i < j} Y_i Y_j. \end{aligned} \quad (1.11)$$

Utilizziamo ora la (1.11) per determinare il momento del secondo ordine di S^2 :

$$E[(S^2)^2] = E\left\{\left[\frac{1}{n} \sum_{i=1}^n Y_i^2 - \frac{2}{n(n-1)} \sum_{i < j} Y_i Y_j\right]^2\right\}$$

$$= E\left[\frac{1}{n^2}\left(\sum_{i=1}^n Y_i^2\right)^2 - \frac{4}{n^2(n-1)} \sum_{r=1}^n Y_r^2 \sum_{i < j} Y_i Y_j + \frac{4}{n^2(n-1)^2} \left(\sum_{i < j} Y_i Y_j\right)^2\right]. \quad (1.12)$$

Per l'indipendenza delle variabili Y_i , ed essendo $E(Y_i) = E(X_i - \mu) = 0$, risulta:

$$\begin{aligned} E\left[\frac{1}{n^2}\left(\sum_{i=1}^n Y_i^2\right)^2\right] &= \frac{1}{n^2} E\left(\sum_{i=1}^n Y_i^2 \sum_{j=1}^n Y_j^2\right) = \frac{1}{n^2} E\left(\sum_{i=1}^n Y_i^4 + \sum_{i \neq j} Y_i^2 Y_j^2\right) \\ &= \frac{1}{n^2} E\left(\sum_{i=1}^n Y_i^4\right) + \frac{2}{n^2} E\left(\sum_{i < j} Y_i^2 Y_j^2\right), \end{aligned} \quad (1.13)$$

$$E\left[\frac{4}{n^2(n-1)} \sum_{r=1}^n Y_r^2 \sum_{i < j} Y_i Y_j\right] = 0, \quad (1.14)$$

$$\begin{aligned} E\left[\frac{4}{n^2(n-1)^2} \left(\sum_{i < j} Y_i Y_j\right)^2\right] &= \frac{4}{n^2(n-1)^2} E\left(\sum_{h < k < r < s} Y_h Y_k Y_r Y_s\right) \\ &= \frac{4}{n^2(n-1)^2} E\left(\sum_{i < j} Y_i^2 Y_j^2\right). \end{aligned} \quad (1.15)$$

Facendo uso delle (1.13), (1.14) e (1.15) nella (1.12), si ricava:

$$E[(S^2)^2] = E\left[\frac{1}{n^2} \sum_{i=1}^n Y_i^4 + \frac{2(n-1)^2 + 4}{n^2(n-1)^2} \sum_{i < j} Y_i^2 Y_j^2\right].$$

Da questa, essendo $E(Y_i^4) = E[(X_i - \mu)^4] = \mu_4$ e $E(Y_i^2) = E[(X_i - \mu)^2] = \sigma^2$, segue:

$$E[(S^2)^2] = \frac{1}{n} \mu_4 + \frac{(n-1)^2 + 2}{n(n-1)} \sigma^4. \quad (1.16)$$

In virtù della prima delle (1.10) e della (1.16), si ottiene infine:

$$D^2(S^2) = E[(S^2)^2] - [E(S^2)]^2 = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right),$$

che coincide con la seconda delle (1.10).

Un'ulteriore statistica che incontreremo nel seguito è la *deviazione standard campionaria* S , così definita:

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} \quad (n > 1). \quad (1.17)$$

Si noti che media campionaria, varianza campionaria, momenti campionari e deviazione standard campionaria dipendono dalla taglia del campione. Se questa viene considerata fissa,

le notazioni adottate \bar{X} , S^2 , $\bar{X}^{(k)}$ e S sono opportune in quanto semplici. Allorché, tuttavia, tali statistiche intervengono in procedimenti in cui la taglia del campione viene fatta variare, occorre far ricorso a notazioni più idonee in cui si evidenzia la loro dipendenza dalla taglia. In generale una statistica $T = g(X_1, X_2, \dots, X_n)$ sarà dunque denotata con T_n quando è opportuno indicare che essa si riferisce ad un campione casuale di taglia n .

In numerosi contesti accade che si debba fare riferimento a più di un campione casuale; sorge quindi la necessità di confrontare le informazioni fornite da ciascuno di questi. Si considera a tale scopo una variabile casuale costruita a partire da statistiche relative ai diversi campioni. Un semplice esempio nasce quando, disponendo di due campioni indipendenti, si considera la variabile casuale definita come differenza delle due medie campionarie. Nel teorema che segue vengono specificate media e varianza di tale variabile casuale.

Teorema 1.3.2 *Siano $X_{11}, X_{12}, \dots, X_{1n}$ e $X_{21}, X_{22}, \dots, X_{2m}$ variabili casuali indipendenti. Se le prime n costituiscono un campione casuale estratto da una popolazione avente valore medio μ_1 e varianza σ_1^2 e se le rimanenti m costituiscono un campione casuale estratto da una popolazione avente valore medio μ_2 e varianza σ_2^2 , si ha:*

$$E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2, \quad D^2(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m},$$

dove \bar{X}_1 e \bar{X}_2 denotano le medie campionarie dei due campioni casuali.

Dim. Il valore medio di $\bar{X}_1 - \bar{X}_2$ segue dalla prima delle (1.9) e dalla linearità dell'operazione di media. Dall'indipendenza delle variabili casuali che costituiscono i due campioni segue poi:

$$\begin{aligned} D^2(\bar{X}_1 - \bar{X}_2) &= D^2\left(\frac{1}{n} \sum_{k=1}^n X_{1k} - \frac{1}{m} \sum_{k=1}^m X_{2k}\right) \\ &= \frac{1}{n^2} \sum_{k=1}^n D^2(X_{1k}) + \frac{1}{m^2} \sum_{k=1}^m D^2(X_{2k}) \\ &= \frac{1}{n^2} \sum_{k=1}^n \sigma_1^2 + \frac{1}{m^2} \sum_{k=1}^m \sigma_2^2 = \frac{1}{n^2} n \sigma_1^2 + \frac{1}{m^2} m \sigma_2^2 \\ &= \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}, \end{aligned}$$

avendo fatto uso dell'ipotesi $D^2(X_{jk}) = \sigma_j^2$ ($j = 1, 2$). ■

Solitamente i valori assunti dalle medie campionarie \bar{X}_1 e \bar{X}_2 in corrispondenza delle realizzazioni $(x_{11}, x_{12}, \dots, x_{1n})$ e $(x_{21}, x_{22}, \dots, x_{2m})$ osservate vengono denotati rispettivamente con \bar{x}_1 e con \bar{x}_2 :

$$\bar{x}_1 = \frac{1}{n} \sum_{i=1}^n x_{1i}, \quad \bar{x}_2 = \frac{1}{m} \sum_{i=1}^m x_{2i}. \quad (1.18)$$

Esempio 1.3.2 In un ospedale vengono curati 14 pazienti seguendo due diverse terapie. Precisamente, a 6 pazienti viene somministrato un farmaco di tipo A ed ai rimanenti 8 un farmaco

di tipo *B*. Per ciascun paziente si registra quindi il numero di giorni intercorrenti tra somministrazione dei farmaci e guarigione. Questo numero, per ogni paziente, non è prevedibile con certezza, ed è quindi ragionevole riguardarlo come una variabile casuale; è inoltre anche legittimo supporre che i tempi di guarigione siano tra loro indipendenti. Infine, su base intuitiva si comprende come i tempi di guarigione dei pazienti sottoposti a medesima terapia possano riguardarsi come identicamente distribuiti, mentre non è consentito escludere che le due distinte terapie possano dar luogo a tempi diversi di guarigione. In termini statistici, è ragionevole affermare che le durate dei tempi di guarigione dei pazienti trattati con farmaci di tipo *A* costituiscono un campione casuale $(X_{11}, X_{12}, \dots, X_{1n})$ di taglia $n = 6$ e che le variabili casuali relative ai pazienti trattati con farmaci di tipo *B* formano un campione casuale $(X_{21}, X_{22}, \dots, X_{2m})$ di taglia $m = 8$. Riportiamo qui di seguito le durate, in giorni, dei tempi di guarigione dei 16 pazienti:

Farmaco di tipo *A*: (15, 24, 20, 24, 18, 28)

Farmaco di tipo *B*: (16, 32, 24, 26, 18, 26, 30, 28).

Esse costituiscono le realizzazioni osservate dei due campioni casuali, in corrispondenza delle quali dalle (1.18) segue che le medie campionarie dei due campioni casuali assumono i seguenti valori:

$$\bar{x}_1 = \frac{1}{n} \sum_{i=1}^n x_{1i} = 21.5, \quad \bar{x}_2 = \frac{1}{m} \sum_{i=1}^m x_{2i} = 25.$$

◆

1.4 Campioni di popolazioni normali

In statistica di particolare rilevanza è la popolazione normale, così che ruolo importante rivestono i campioni casuali di variabile genitrice a distribuzione normale. Il motivo principale risiede nella circostanza che proprietà relative a popolazioni normali possono poi spesso essere estese a popolazioni di diversa natura facendo ricorso al teorema centrale del limite.

Ricordiamo che una variabile casuale continua⁶ Y ha distribuzione normale di media μ e varianza σ^2 se ha densità di probabilità

$$f_Y(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad (x \in \mathbb{R}),$$

con $\mu \in \mathbb{R}$ e $\sigma > 0$. Notiamo che se Y ha distribuzione normale di media μ e varianza σ^2 la trasformazione lineare

$$X = \frac{Y - \mu}{\sigma} \quad (1.19)$$

definisce la variabile casuale normale standard. Nello studio di variabili casuali normali è possibile dunque ricondursi sempre a variabili normali standard mediante la trasformazione (1.19). Nella Figura 1.1 è mostrato il grafico della densità normale standard, ossia della funzione

⁶Per brevità, talora parleremo di variabili casuali "continue" in luogo di variabili casuali "assolutamente continue".

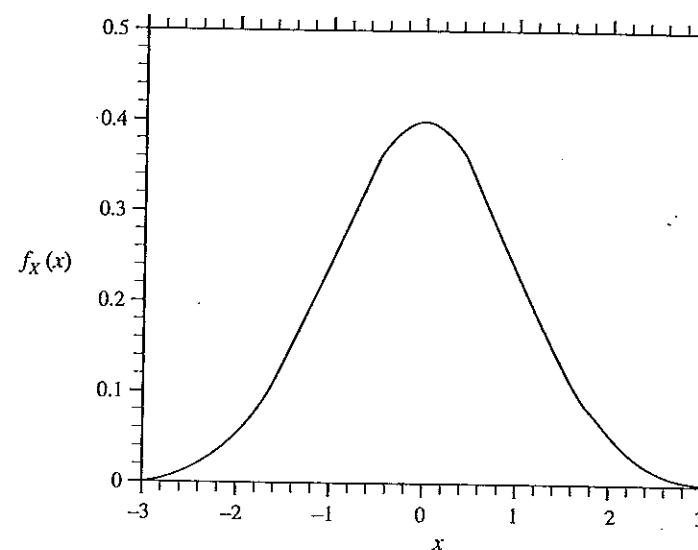


Figura 1.1: Densità di probabilità della variabile normale standard.

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad (x \in \mathbb{R}). \quad (1.20)$$

La prima difficoltà che sorge nello studio di una variabile casuale normale è che non esiste un'espressione semplice per la sua funzione di distribuzione. Per una variabile casuale X normale standard la funzione di distribuzione $\Phi(x)$ è rappresentata solitamente nella forma integrale

$$\Phi(x) \equiv P(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy \quad (x \in \mathbb{R}) \quad (1.21)$$

il cui grafico è mostrato nella Figura 1.2.

Per il frequente e fondamentale uso della distribuzione normale in statistica sorge comunque l'esigenza di poter agevolmente calcolare la probabilità che una variabile casuale normale assuma valori in un qualunque prefissato intervallo. A tale scopo sono state approntate tabelle numeriche concernenti integrali della densità normale standard. In particolare, nella Tabella 1 dell'Appendice B sono riportati i valori della funzione

$$\Psi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-y^2/2} dy \quad (1.22)$$

in corrispondenza di alcuni valori di x compresi tra 0 e 3.5. Risulta così possibile calcolare la funzione di distribuzione (1.21) nei corrispondenti punti; infatti, essendo $\Phi(0) = 1/2$ ed

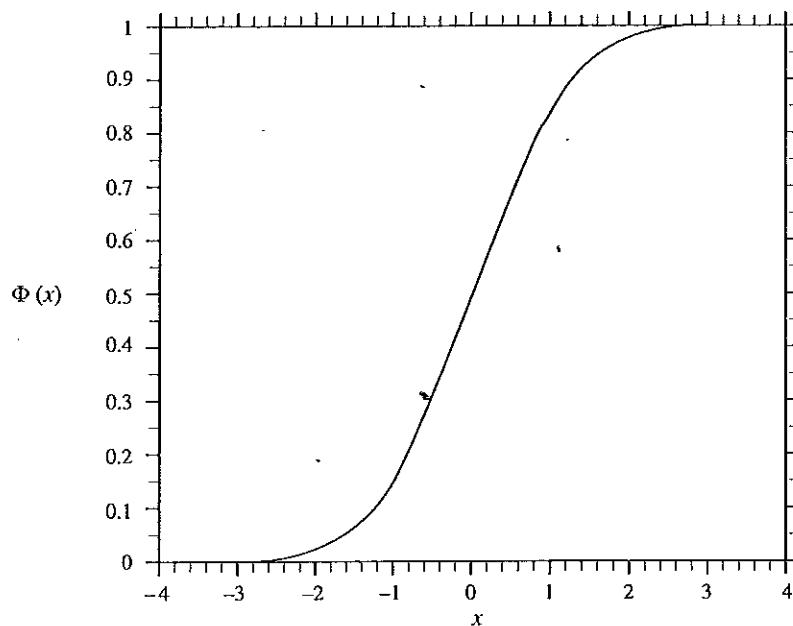


Figura 1.2: Funzione di distribuzione della variabile normale standard.

essendo $e^{-y^2/2}$ una funzione pari, si ha:

$$\Phi(x) = \begin{cases} \frac{1}{2} + \Psi(x) & \text{per } x \geq 0, \\ \frac{1}{2} - \Psi(-x) & \text{per } x < 0. \end{cases} \quad (1.23)$$

Utilizzando la (1.23) è possibile, ad esempio, valutare la probabilità che una variabile casuale normale Y di media μ e varianza σ^2 assuma valori in prefissati intorni della media. Infatti dalla (1.23), ricordando la (1.19), per $\delta > 0$ si ha

$$\begin{aligned} P(\mu - \delta < Y < \mu + \delta) &= P\left(-\frac{\delta}{\sigma} < X < \frac{\delta}{\sigma}\right) = \Phi\left(\frac{\delta}{\sigma}\right) - \Phi\left(-\frac{\delta}{\sigma}\right) \\ &= 2\Psi\left(\frac{\delta}{\sigma}\right). \end{aligned} \quad (1.24)$$

Per $\delta = 3\sigma$, poiché risulta $\Psi(3) \approx 0.4987$ (cfr. Tabella 1 dell'Appendice B), dalla (1.24) segue

$$P(\mu - 3\sigma < Y < \mu + 3\sigma) \approx 0.9974. \quad (1.25)$$

1.4. CAMPIONI DI POPOLAZIONI NORMALI

La (1.25), nota col nome di *legge del 3σ*, mostra che una variabile casuale normale assume valori nell'intervallo di centro μ e semiampiezza 3σ con probabilità davvero molto prossima all'unità.

Esempio 1.4.1 Un distributore automatico di caffè è programmato in modo tale da includere 2 grammi di zucchero in ogni razione. In realtà, a causa di imperfezioni meccaniche, la quantità, in grammi, di zucchero distribuita è riguardabile come una variabile casuale di media 2 e deviazione standard 0.4. Conseguentemente, per stabilire se la macchina distribuisce una quantità di zucchero che si discosta eccessivamente da quella programmata (2 grammi), si decide di calcolare la probabilità che in un campione di 48 razioni di caffè la quantità media di zucchero distribuita sia o inferiore a 1.9 grammi o superiore a 2.1 grammi, identificando in ciò una situazione di non adeguata dolcificazione. Ricordando che la media campionaria \bar{X} descrive qui la quantità media di zucchero distribuita nelle 48 razioni di caffè, in virtù del teorema centrale del limite possiamo assumere che la variabile

$$Z = \frac{X_1 + \dots + X_n - n\mu}{\sqrt{n}\sigma} \doteq \frac{\bar{X} - 2}{0.4/\sqrt{48}}$$

sia approssimativamente normale standard, dove l'ultima uguaglianza discende dall'essere $\mu = 2$, $\sigma = 0.4$ ed $n = 48$. Pertanto, la probabilità richiesta vale

$$\begin{aligned} P(\bar{X} < 1.9) + P(\bar{X} > 2.1) &= 1 - P(1.9 \leq \bar{X} \leq 2.1) \\ &= 1 - P\left(\frac{1.9 - 2}{0.0577} \leq \frac{\bar{X} - 2}{0.0577} \leq \frac{2.1 - 2}{0.0577}\right) \\ &= 1 - P(-1.77 \leq Z \leq 1.77) \\ &= 1 - 2P(0 \leq Z \leq 1.77). \end{aligned}$$

Essendo la variabile Z approssimativamente normale standard, dalla Tabella 1 dell'Appendice B segue $P(0 \leq Z \leq 1.77) = \Psi(1.77) \approx 0.4616$, così che la probabilità richiesta è data da

$$P(\bar{X} < 1.9) + P(\bar{X} > 2.1) \approx 1 - 2 \cdot 0.4582 = 0.0836.$$

Si conclude che, approssimativamente, in media 4 caffè su 48 non sono adeguatamente zuccherati. ◆

Talora è conveniente riferirsi ad un tipo particolare di integrale della densità normale standard introducendo una quantità, z_α , che chiameremo *quantile superiore*⁷. Data dunque una variabile casuale X avente distribuzione normale standard, il quantile superiore z_α è il reale tale da aversi:

$$P(X \geq z_\alpha) \equiv \int_{z_\alpha}^{\infty} f_X(x) dx = \alpha, \quad (1.26)$$

dove $0 < \alpha < 1$ e dove $f_X(x)$ denota la densità di probabilità (1.20) della variabile casuale normale standard. Si noti che la (1.26) può equivalentemente esprimersi come

$$\Phi(z_\alpha) = 1 - \alpha, \quad (1.27)$$

dove Φ è data dalla (1.21).

⁷L'origine di tale denominazione risulterà evidente nel § 4.9 in cui vengono definiti i quantili.

Proposizione 1.4.1 Per ogni α in $(0, 1)$ il reale z_α definito dalla (1.26) soddisfa la seguente relazione:

$$z_{1-\alpha} = -z_\alpha. \quad (1.28)$$

Dim. Dalla (1.26) segue:

$$1 - \alpha = P(X \geq z_{1-\alpha}) = \int_{z_{1-\alpha}}^{\infty} f_X(x) dx.$$

Ponendo $x = -y$ nell'integrale e ricordando che $f_X(x)$ è una funzione pari, si ottiene:

$$1 - \alpha = \int_{-\infty}^{-z_{1-\alpha}} f_X(y) dy = P(X \leq -z_{1-\alpha}),$$

da cui si trae

$$P(X \geq -z_{1-\alpha}) = \alpha,$$

ovvero

$$1 - \Phi(-z_{1-\alpha}) = \alpha. \quad (1.29)$$

Poiché Φ è strettamente crescente, dal confronto della (1.27) con la (1.29) segue infine la (1.28). ■

I valori z_α di frequente utilizzazione sono riportati nella Tabella 2 dell'Appendice B. Essi si possono peraltro in generale calcolare facendo ricorso alla Tabella 1 dell'Appendice B. Come già detto, in essa vengono indicati i valori $\Psi(x) \equiv \Phi(x) - 1/2$ in corrispondenza di alcuni valori positivi di x . Poiché, come si è visto, risulta $\Phi(z_\alpha) = 1 - \alpha$, si ha:

$$\Psi(z_\alpha) \equiv \Phi(z_\alpha) - \frac{1}{2} = \frac{1}{2} - \alpha.$$

Pertanto per $0 < \alpha < 1/2$ in corrispondenza di z_α nella tabella compare $\Phi(z_\alpha) - 1/2$, ossia la quantità $1/2 - \alpha$. Per determinare z_α occorre quindi individuare il valore $1/2 - \alpha$ tra quelli presenti nella tabella. La sua controimmagine fornisce allora una buona approssimazione di z_α , salvo questioni di approssimazioni numeriche sulle quali non possiamo qui addentrarci. A titolo di esempio, si supponga di aver fissato il valore $\alpha = 0.05$, così che risulta $1/2 - \alpha = 0.45$. Con riferimento alla Tabella 1 dell'Appendice B si vede che esistono due valori che ugualmente approssimano 0.45, ossia i valori 0.4495 e 0.4505 ai quali corrispondono rispettivamente le controimmagini 1.64 e 1.65. Sarà pertanto ragionevole assumere per z_α la media di queste ultime, ossia porre $z_{0.05} = 1.645$. Si noti che a mezzo della Tabella 1 è possibile determinare i valori di z_α con $0 < \alpha < 1/2$; mediante la (1.28) è poi possibile ottenere i valori di z_α con $1/2 < \alpha < 1$.

Vale qui la pena di osservare che è anche possibile fare uso di un risultato limite che consente di ottenere una valutazione asintotica di talune probabilità per variabili normali. Nella proposizione che segue si ricava infatti un'approssimazione per la probabilità $P(X > x)$ di una variabile casuale X normale standard, valida per grandi valori di x .

Proposizione 1.4.2 Se X è una variabile casuale normale standard risulta:

$$P(X > x) \sim \frac{f_X(x)}{x} \quad \text{per } x \rightarrow \infty,$$

ossia:

$$\lim_{x \rightarrow \infty} \frac{P(X > x)}{f_X(x)/x} = 1.$$

Dim. Osserviamo anzitutto che per $x > 0$ risulta:

$$P(X > x) \leq \frac{f_X(x)}{x}. \quad (1.30)$$

Infatti, ricordando la (1.20), per $x > 0$ si ha:

$$\begin{aligned} P(X > x) &= \int_x^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\ &\leq \frac{1}{x} \int_x^{\infty} \frac{1}{\sqrt{2\pi}} y e^{-y^2/2} dy \equiv \frac{1}{x} f_X(x). \end{aligned}$$

Dal rapporto

$$\frac{P(X > x)}{f_X(x)/x} = \frac{x \int_x^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy}{\frac{1}{\sqrt{2\pi}} e^{-x^2/2}},$$

applicando la regola di L'Hospital si trae:

$$\lim_{x \rightarrow \infty} \frac{P(X > x)}{f_X(x)/x} = 1 - \lim_{x \rightarrow \infty} \frac{\int_x^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy}{x \frac{1}{\sqrt{2\pi}} e^{-x^2/2}} = 1 - \lim_{x \rightarrow \infty} \frac{P(X > x)}{x f_X(x)}.$$

In virtù della (1.30) si ottiene così:

$$\lim_{x \rightarrow \infty} \frac{P(X > x)}{f_X(x)/x} \geq 1 - \lim_{x \rightarrow \infty} \frac{1}{x^2} = 1. \quad (1.31)$$

Osservando infine che per la (1.30) risulta

$$\lim_{x \rightarrow \infty} \frac{P(X > x)}{f_X(x)/x} \leq 1, \quad (1.32)$$

dalle (1.31) e (1.32) segue la tesi. ■

A titolo di esempio nella Tabella 1.1 sono riportati i valori della probabilità $P(X > x)$ e dell'approssimazione $f_X(x)/x$ in corrispondenza di alcune scelte di x . È evidente come l'approssimazione migliori al crescere di x .

Determiniamo ora la distribuzione campionaria della media campionaria relativa ad un campione casuale estratto da popolazione normale.

Tabella 1.1: Probabilità $P(X > x)$ e corrispondenti approssimazioni $f_X(x)/x$ per alcune scelte di x .

x	$P(X > x)$	$f_X(x)/x$
1	0.1587	0.2420
1.5	0.0668	0.0863
2	0.0228	0.0270
2.5	0.0062	0.0070
3	0.0013	0.0014

Teorema 1.4.1 La media campionaria di un campione casuale di taglia n estratto da una popolazione normale di valore medio μ e varianza σ^2 ha distribuzione normale di media μ e varianza σ^2/n .

Dim. Sia (X_1, X_2, \dots, X_n) il campione casuale. Poiché le variabili che lo costituiscono sono indipendenti e identicamente distribuite, per la funzione generatrice $M_{\bar{X}}(t)$ della media campionaria si ha:

$$\begin{aligned} M_{\bar{X}}(t) &= E\left(e^{\bar{X}}\right) = E\left\{\exp\left[\frac{t}{n}(X_1 + \dots + X_n)\right]\right\} \\ &= E\left[\exp\left(\frac{t}{n}X_1\right)\right] \cdots E\left[\exp\left(\frac{t}{n}X_n\right)\right] \\ &= \left[M_X\left(\frac{t}{n}\right)\right]^n, \end{aligned}$$

dove $M_X(t)$ denota la funzione generatrice delle variabili che costituiscono il campione casuale. Poiché, per ipotesi, queste hanno distribuzione normale di media μ e varianza σ^2 , la funzione generatrice $M_X(t)$ ha la seguente espressione (cfr. § A.2.2):

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right).$$

Si ha dunque:

$$M_{\bar{X}}(t) = \left[M_X\left(\frac{t}{n}\right)\right]^n = \left[\exp\left(\mu \frac{t}{n} + \frac{1}{2}\sigma^2 \frac{t^2}{n^2}\right)\right]^n = \exp\left(\mu t + \frac{1}{2}\frac{\sigma^2}{n}t^2\right),$$

che è la funzione generatrice dei momenti di una variabile casuale normale di media μ e varianza σ^2/n . Per la corrispondenza biunivoca tra funzioni generatrici dei momenti e densità di probabilità segue infine che \bar{X} ha distribuzione normale di valore medio μ e varianza σ^2/n . ■

Esempio 1.4.2 Riprendendo in esame l'Esempio 1.3.1 supponiamo che le durate di ciascuna delle lampadine prodotte possa essere riguardata, a patto di una evidente approssimazione, come una variabile casuale normale di media $\mu = 215$ e deviazione standard $\sigma = 20$. Il lotto delle 14 lampadine esaminate costituisce allora un campione casuale estratto da una popolazione normale. Assumiamo che il test di affidabilità viene ritenuto superato se è maggiore di

0.05 la probabilità che la media campionaria \bar{X} sia minore del valore osservato $\bar{x} = 209.78$. Se si pone $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$, con $n = 14$, $\mu = 215$ e $\sigma = 20$, la probabilità da calcolare è la seguente:

$$P(\bar{X} < 209.78) = P\left(Z < \frac{209.78 - 215}{5.35}\right) = P(Z < -0.98).$$

In virtù del Teorema 1.4.1 la media campionaria \bar{X} ha distribuzione normale, cosicché Z è una variabile casuale normale standard. Facendo allora uso della Tabella 1 dell'Appendice B si ha:

$$\begin{aligned} P(Z < -0.98) &= P(Z > 0.98) = P(Z > 0) - P(0 < Z \leq 0.98) \\ &= 0.5 - 0.3365 = 0.1635, \end{aligned}$$

così che la probabilità $P(\bar{X} < 209.78)$ risulta maggiore di 0.05. Ne segue che, nell'ipotesi di validità della distribuzione normale per le durate di funzionamento delle lampadine, il test di affidabilità si deve ritenere superato in base alla realizzazione osservata. ♦

Determiniamo ora la distribuzione campionaria della differenza delle medie campionarie relative a due campioni casuali estratti da differenti popolazioni normali.

Teorema 1.4.2 Siano $X_{11}, X_{12}, \dots, X_{1n}$ e $X_{21}, X_{22}, \dots, X_{2m}$ variabili casuali indipendenti. Se le prime n di esse costituiscono un campione casuale estratto da una popolazione normale di valore medio μ_1 e varianza σ_1^2 e se le rimanenti m costituiscono un campione casuale estratto da una popolazione normale di valore medio μ_2 e varianza σ_2^2 , la differenza $\bar{X}_1 - \bar{X}_2$ tra le medie campionarie ha distribuzione normale di valore medio $\mu_1 - \mu_2$ e varianza $\sigma_1^2/n + \sigma_2^2/m$.

Dim. Procedendo in analogia con la dimostrazione del Teorema 1.4.1 osserviamo che la funzione generatrice dei momenti della differenza tra le medie campionarie è esprimibile al seguente modo:

$$\begin{aligned} M_{\bar{X}_1 - \bar{X}_2}(t) &= E\left[e^{t(\bar{X}_1 - \bar{X}_2)}\right] = E\left[\exp\left(\frac{t}{n}\sum_{k=1}^n X_{1k} - \frac{t}{m}\sum_{k=1}^m X_{2k}\right)\right] \\ &= E\left[\prod_{k=1}^n \exp\left(\frac{t}{n}X_{1k}\right) \prod_{k=1}^m \exp\left(-\frac{t}{m}X_{2k}\right)\right] \\ &= \prod_{k=1}^n M_{1k}\left(\frac{t}{n}\right) \prod_{k=1}^m M_{2k}\left(-\frac{t}{m}\right), \end{aligned}$$

dove $M_{jk}(t)$ denota la funzione generatrice dei momenti della variabile X_{jk} . Quest'ultima è una variabile casuale normale di valore medio μ_j e varianza σ_j^2 , così che risulta:

$$M_{jk}(t) = \exp\left(\mu_j t + \frac{1}{2}\sigma_j^2 t^2\right).$$

Si ottiene allora:

$$\begin{aligned} M_{\bar{X}_1 - \bar{X}_2}(t) &= \prod_{k=1}^n \exp\left(\mu_1 \frac{t}{n} + \frac{\sigma_1^2 t^2}{2 n^2}\right) \prod_{k=1}^m \exp\left(-\mu_2 \frac{t}{m} + \frac{\sigma_2^2 t^2}{2 m^2}\right) \\ &= \exp\left[(\mu_1 - \mu_2)t + \frac{1}{2}\left(\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)t^2\right], \end{aligned}$$

che è la funzione generatrice dei momenti di una variabile casuale normale di valore medio $\mu_1 - \mu_2$ e varianza $\sigma_1^2/n + \sigma_2^2/m$. Per la corrispondenza bivoca tra funzioni generatrici dei momenti e densità di probabilità, si conclude che $\bar{X}_1 - \bar{X}_2$ ha distribuzione normale di valore medio $\mu_1 - \mu_2$ e varianza $\sigma_1^2/n + \sigma_2^2/m$. ■

Esempio 1.4.3 Con riferimento all'Esempio 1.3.2 si assuma che i tempi di guarigione di pazienti trattati con farmaci di tipo A o di tipo B possano essere riguardati, approssimativamente, come variabili casuali aventi distribuzioni normali di varianza $\sigma^2 = 4$. A partire dalle realizzazioni osservate dei due campioni casuali si desidera allora ottenere delle informazioni per stabilire se è lecito ritenere che la media delle variabili casuali costituenti il campione casuale $(X_{11}, X_{12}, \dots, X_{1n})$ è uguale a quella delle variabili del campione $(X_{21}, X_{22}, \dots, X_{2m})$. Supponendo ora che le medie dei due campioni coincidano, essendo $n = 6, m = 8$ e $\sigma^2 = 4$ dal Teorema 1.4.2 segue che la variabile $\bar{X}_1 - \bar{X}_2$ ha distribuzione normale di media zero e varianza $\sigma^2(1/n + 1/m) = 1.1667$. La variabile

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{1.1667}} \quad (1.33)$$

ha dunque distribuzione normale standard. Ricordando che le realizzazioni osservate forniscono $\bar{x}_1 = 21.5$ e $\bar{x}_2 = 25$, la variabile (1.33) assume il valore

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{1.1667}} = \frac{21.5 - 25}{\sqrt{1.1667}} = -3.24. \quad (1.34)$$

Dalla Tabella 1 dell'Appendice B risulta poi $P(0 < Z < 3.24) = 0.4994$ di modo che la probabilità che Z assuma un valore non superiore a -3.24 (cfr. (1.34)) vale

$$\begin{aligned} P(Z \leq -3.24) &= P(Z \geq 3.24) = P(Z > 0) - P(0 < Z < 3.24) \\ &= 0.5 - 0.4994 = 0.0006, \end{aligned}$$

che è certamente molto piccola in questo contesto. Ne segue che, in base ai dati osservati, è preferibile rifiutare l'ipotesi che le medie dei due campioni casuali siano identiche e ritenere, pertanto, che terapie basate su trattamenti con medicinali di tipo A o di tipo B debbano essere riguardate come significativamente differenti. ♦

Capitolo 2 Distribuzioni speciali

2.1 Distribuzione chi-quadrato

In questo paragrafo analizzeremo le principali proprietà di una particolare variabile casuale, la cosiddetta variabile "chi-quadrato", indicata in letteratura anche con il simbolo χ^2 , che riveste grande rilevanza per le sue molteplici applicazioni statistiche. A tal fine è utile richiamare la definizione ed alcune proprietà della funzione gamma di Eulero, alle quali dovremo fare frequente ricorso.

Definizione 2.1.1 Dicesi funzione gamma di Eulero, o semplicemente funzione gamma, la funzione

$$\Gamma(t) = \int_0^\infty z^{t-1} e^{-z} dz \quad (t > 0). \quad (2.1)$$

$\Gamma(t)$ non è dunque altro che una notazione per indicare l'integrale definito (esistente per $t > 0$) che appare a secondo membro della (2.1).

Osservazione 2.1.1 La funzione gamma gode delle seguenti proprietà:

(i) Per ogni $t > 0$ risulta:

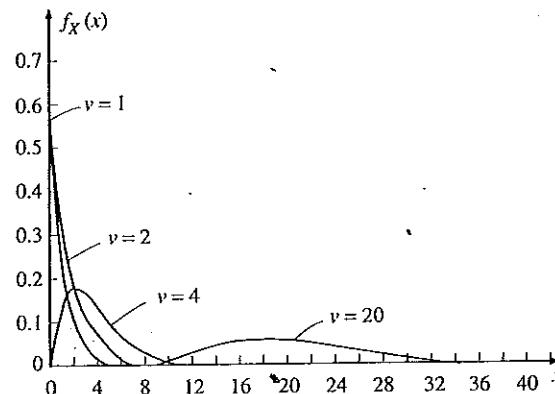
$$\Gamma(t+1) = t \Gamma(t); \quad (2.2)$$

(ii) si ha:

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}. \quad (2.3)$$

Dim. La (i) segue immediatamente dalla (2.1) mediante integrazione per parti. Si noti in particolare che $\Gamma(t+1) = t!$ se t è un intero non negativo. Per dimostrare la (ii) osserviamo che dalla (2.1) segue:

$$\begin{aligned} \Gamma\left(\frac{1}{2}\right) &= \int_0^\infty \frac{1}{\sqrt{z}} e^{-z} dz = \sqrt{2} \int_0^\infty e^{-y^2/2} dy \\ &= \frac{1}{\sqrt{2}} \int_{-\infty}^\infty e^{-y^2/2} dy = \sqrt{\pi}, \end{aligned}$$

Figura 2.1: Densità chi-quadrato con $v = 1, 2, 4, 20$ gradi di libertà.

avendo effettuato la sostituzione $z = y^2/2$.

Definizione 2.1.2 Una variabile casuale X continua si dice avere distribuzione chi-quadrato, o distribuzione χ^2 , con v gradi di libertà, dove v è un intero positivo, se essa ha la seguente densità di probabilità:

$$f_X(x) = \begin{cases} \frac{x^{v/2-1} e^{-x/2}}{2^{v/2} \Gamma(\frac{v}{2})} & \text{per } x \geq 0, \\ 0 & \text{altrimenti.} \end{cases} \quad (2.4)$$

I grafici della densità chi-quadrato sono riportati in Figura 2.1 per alcuni valori del numero v di gradi di libertà. Tali grafici indicano una progressiva diminuzione dell'asimmetria della densità all'aumentare del numero di gradi di libertà. Si noti che già per $v = 20$ la densità ha un andamento molto simile a quello della densità di una variabile casuale normale.

Ricaviamo ora la funzione generatrice dei momenti della variabile casuale chi-quadrato.

Proposizione 2.1.1 La funzione generatrice dei momenti della variabile casuale X a distribuzione χ^2 con v gradi di libertà è data da

$$M_X(t) = (1 - 2t)^{-\frac{v}{2}} \quad (t < 1/2). \quad (2.5)$$

Dim. Facendo uso della (2.4) si ha infatti:

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} f_X(x) dx = \int_0^{\infty} e^{tx} \frac{x^{v/2-1} e^{-x/2}}{2^{v/2} \Gamma(v/2)} dx \\ &= \frac{1}{\Gamma(v/2)} \int_0^{\infty} \left(\frac{x}{2}\right)^{v/2-1} e^{-x(1/2-t)} \frac{1}{2} dx \\ &= \frac{(1-2t)^{-v/2}}{\Gamma(v/2)} \int_0^{\infty} \left[\frac{x}{2}(1-2t)\right]^{v/2-1} e^{-x(1/2-t)} \frac{1-2t}{2} dx. \end{aligned}$$

2.1. DISTRIBUZIONE CHI-QUADRATO

Effettuando la sostituzione $z = x(1/2-t)$, con $t < 1/2$, si ottiene:

$$M_X(t) = \frac{(1-2t)^{-v/2}}{\Gamma(v/2)} \int_0^{\infty} z^{v/2-1} e^{-z} dz.$$

Infine, ricordando la (2.1), per $t < 1/2$ segue la (2.5).

Osservazione 2.1.2 La variabile casuale chi-quadrato con $v = 2$ gradi di libertà ha distribuzione coincidente con quella di una variabile casuale esponenziale di media 2.

Dim. Segue immediatamente dall'espressione (2.5) della funzione generatrice. Infatti, ponendovi $v = 2$ si ricava $M_X(t) = (1-2t)^{-1}$ (con $t < 1/2$) che coincide con la funzione generatrice di una variabile casuale esponenziale di media 2. Più direttamente, basta porre $v = 2$ nella (2.4) per ottenere la densità di probabilità esponenziale di media 2.

Siamo ora in grado di ricavare l'espressione dei momenti intorno all'origine della variabile casuale chi-quadrato.

Proposizione 2.1.2 Se X è una variabile casuale a distribuzione χ^2 con v gradi di libertà, si ha:

$$E(X^k) = 2^k \frac{\Gamma(\frac{v}{2} + k)}{\Gamma(\frac{v}{2})} \quad (k = 1, 2, \dots). \quad (2.6)$$

Dim. Dalla (2.5) si ottiene facilmente:

$$\begin{aligned} E(X^k) &= \frac{d^k}{dt^k} M_X(t) \Big|_{t=0} \\ &= \left(-\frac{v}{2}\right) \left(-\frac{v}{2}-1\right) \cdots \left(-\frac{v}{2}-k+1\right) (-2)^k \\ &= 2^k \frac{\Gamma(v/2+k)}{\Gamma(v/2)} \quad (k = 1, 2, \dots), \end{aligned}$$

dove si è fatto uso della relazione

$$\left(\frac{v}{2}+k-1\right) \left(\frac{v}{2}+k-2\right) \cdots \frac{v}{2} \Gamma\left(\frac{v}{2}\right) = \Gamma\left(\frac{v}{2}+k\right),$$

conseguenza della (2.2).

Osservazione 2.1.3 Media e varianza di una variabile casuale X a distribuzione χ^2 con v gradi di libertà, sono rispettivamente

$$\overline{E(X)} = v \quad (2.7)$$

$$D^2(X) = 2v. \quad (2.8)$$

Dim. Dalla (2.6) per $k = 1$ si ottiene immediatamente la (2.7) quando si faccia uso della (2.2). In maniera analoga si ricava

$$E(X^2) = v^2 + 2v. \quad (2.9)$$

Dalla (2.7), dalla (2.9) e dalla relazione

$$D^2(X) = E(X^2) - [E(X)]^2$$

segue infine la (2.8). \square

Le variabili chi-quadrato, come vedremo nei teoremi che seguono, sono collegabili alle variabili casuali normali.

Teorema 2.1.1 Se X è una variabile casuale normale standard, la variabile casuale $Y = X^2$ ha distribuzione chi-quadrato con 1 grado di libertà.

Dim. Calcoliamo la funzione di distribuzione $F_Y(y)$ di $Y = X^2$. Per ogni y reale si ha:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X^2 \leq y) \\ &= \begin{cases} P(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \text{per } y > 0, \\ 0 & \text{altrimenti,} \end{cases} \end{aligned}$$

dove $F_X(x)$ è la funzione di distribuzione della variabile casuale normale standard X . Derivando ambo i membri rispetto a y si ricava $f_Y(y) = 0$ per $y \leq 0$, mentre per $y > 0$ si ottiene:

$$f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}} = \frac{f_X(\sqrt{y})}{\sqrt{y}},$$

dove l'ultima uguaglianza segue dall'osservazione che la densità di probabilità

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad (x \in \mathbb{R})$$

è una funzione pari. Si è dunque ottenuto:

$$f_Y(y) = \begin{cases} \frac{f_X(\sqrt{y})}{\sqrt{y}} = \frac{1}{\sqrt{2\pi y}} e^{-y/2} & \text{per } y > 0, \\ 0 & \text{altrimenti.} \end{cases}$$

In virtù dell'identità (2.3) e dell'espressione (2.4) si riconosce che la funzione $f_Y(y)$ è proprio la densità di probabilità di una variabile casuale chi-quadrato con $v = 1$, ossia con 1 grado di libertà. \square

Teorema 2.1.2 Se X_1, X_2, \dots, X_n sono variabili casuali indipendenti aventi distribuzioni chi-quadrato con rispettivamente v_1, v_2, \dots, v_n gradi di libertà, la variabile casuale $Y = X_1 + X_2 + \dots + X_n$ ha distribuzione chi-quadrato con $v_1 + v_2 + \dots + v_n$ gradi di libertà.

2.1. DISTRIBUZIONE CHI-QUADRATO

Dim. Per l'indipendenza di X_1, X_2, \dots, X_n , la funzione generatrice dei momenti di $Y = X_1 + X_2 + \dots + X_n$ si fattorizza nel prodotto delle funzioni generatrici $M_i(t)$ delle variabili X_i . Poiché X_i ha distribuzione chi-quadrato con v_i gradi di libertà, facendo uso della (2.5) si ottiene:

$$M_Y(t) = \prod_{i=1}^n (1 - 2t)^{-v_i/2} = (1 - 2t)^{-\frac{v_1+v_2+\dots+v_n}{2}} \quad (t < 1/2),$$

che riconosciamo essere proprio la funzione generatrice dei momenti di una variabile casuale chi-quadrato con $v_1 + v_2 + \dots + v_n$ gradi di libertà. Per la biequivocità della corrispondenza tra funzioni generatrici dei momenti e densità di probabilità segue infine che $Y = X_1 + X_2 + \dots + X_n$ ha distribuzione chi-quadrato con $v_1 + v_2 + \dots + v_n$ gradi di libertà. \blacksquare

Dai Teoremi 2.1.1 e 2.1.2 discende immediatamente il seguente corollario.

Corollario 2.1.1 Se X_1, X_2, \dots, X_n sono variabili casuali indipendenti a distribuzioni normali standard, la variabile $Y = X_1^2 + X_2^2 + \dots + X_n^2$ ha distribuzione chi-quadrato con n gradi di libertà.

Teorema 2.1.3 Siano X_1 e X_2 variabili casuali indipendenti. Se X_1 ha distribuzione chi-quadrato con v_1 gradi di libertà e se $X_1 + X_2$ ha distribuzione chi-quadrato con $v_1 + v_2$ gradi di libertà, allora X_2 ha distribuzione chi-quadrato con v_2 gradi di libertà.

Dim. Per l'indipendenza di X_1 e X_2 sussiste la seguente relazione tra funzioni generatrici dei momenti:

$$M_{X_1+X_2}(t) = M_{X_1}(t) M_{X_2}(t)$$

da cui, per $t < 1/2$, si trae:

$$M_{X_2}(t) = \frac{M_{X_1+X_2}(t)}{M_{X_1}(t)} = \frac{(1 - 2t)^{-\frac{v_1+v_2}{2}}}{(1 - 2t)^{-\frac{v_1}{2}}} = (1 - 2t)^{-\frac{v_2}{2}}.$$

Questa è la funzione generatrice dei momenti di una variabile casuale chi-quadrato con v_2 gradi di libertà, così che l'asserto resta provato. \blacksquare

Nel teorema che segue sono dimostrati due significativi risultati concernenti campioni casuali estratti da popolazioni normali. Anzitutto, viene evidenziata l'importante proprietà che per siffatti campioni media e varianza campionarie sono statistiche indipendenti;¹ si determina, poi, la distribuzione campionaria della variabile casuale $(n-1)S^2/\sigma^2$ che, val la pena ricordare, risulta essere una statistica se e solo se la varianza σ^2 è nota (cfr. Definizione 1.3.1).

Teorema 2.1.4 Siano \bar{X} e S^2 rispettivamente la media campionaria e la varianza-campionaria di un campione casuale di taglia n estratto da una popolazione normale di varianza σ^2 . Sussistono le seguenti proprietà:

(i) \bar{X} e S^2 sono indipendenti;

(ii) la variabile casuale $(n-1)S^2/\sigma^2$ ha distribuzione χ^2 con $n-1$ gradi di libertà. \blacksquare

¹ Si sottolinea che tale proprietà non sussiste in generale per campioni estratti da popolazioni qualsiasi.

Dim. Denotiamo con (X_1, X_2, \dots, X_n) il campione casuale, con le variabili X_i normali, indipendenti, di media μ e varianza σ^2 . La sua densità di probabilità è pertanto:

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right].$$

Effettuiamo la trasformazione

$$\begin{aligned} \bar{X} &= \frac{1}{n} (X_1 + X_2 + \dots + X_n) \\ X_2 &= X_2 \\ &\vdots \\ X_n &= X_n. \end{aligned}$$

Osservando che il determinante Jacobiano vale $1/n$ e che, inoltre, può scriversi $X_1 = n\bar{X} - X_2 - \dots - X_n$, si ha:

$$\begin{aligned} f_{\bar{X}, X_2, \dots, X_n}(\bar{x}, x_2, \dots, x_n) &= \\ &= n f_{X_1, \dots, X_n}(n\bar{x} - x_2 - \dots - x_n, x_2, \dots, x_n) \\ &= \frac{n}{(2\pi)^{n/2} \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \left[(n\bar{x} - x_2 - \dots - x_n - \mu)^2 + \sum_{i=2}^n (x_i - \mu)^2 \right] \right\}. \end{aligned}$$

Facendo poi uso dell'identità

$$\begin{aligned} (n\bar{x} - x_2 - \dots - x_n - \mu)^2 + \sum_{i=2}^n (x_i - \mu)^2 \\ = \sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2, \end{aligned}$$

si trae:

$$\begin{aligned} f_{\bar{X}, X_2, \dots, X_n}(\bar{x}, x_2, \dots, x_n) &= \\ &= \frac{n}{(2\pi)^{n/2} \sigma^n} \exp \left\{ -\frac{1}{2\sigma^2} \left[\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \right] \right\}. \end{aligned}$$

Poiché la media campionaria \bar{X} ha distribuzione normale di media μ e varianza σ^2/n , la densità della $(n-1)$ -upla (X_2, \dots, X_n) condizionata da \bar{X} è la seguente:

$$\begin{aligned} f_{X_2, \dots, X_n|\bar{X}}(x_2, \dots, x_n|\bar{x}) &= f_{\bar{X}, X_2, \dots, X_n}(\bar{x}, x_2, \dots, x_n) / f_{\bar{X}}(\bar{x}) \\ &= \frac{\sqrt{n}}{(\sqrt{2\pi}\sigma)^{n-1}} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right]. \end{aligned} \quad (2.10)$$

Facciamo uso della (2.10) per calcolare la funzione generatrice dei momenti della variabile casuale

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (2.11)$$

2.1. DISTRIBUZIONE CHI-QUADRATO

condizionata da $\bar{X} = \bar{x}$. Si ha:

$$\begin{aligned} E \left\{ \exp \left[\frac{t}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \right] \middle| \bar{X} = \bar{x} \right\} \\ = \frac{\sqrt{n}}{(\sqrt{2\pi}\sigma)^{n-1}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left[\frac{t}{\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right] \\ \times \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right] dx_2 \cdots dx_n \\ = \frac{\sqrt{n}}{(\sqrt{2\pi}\sigma)^{n-1}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left[-\frac{1}{2\bar{\sigma}^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right] dx_2 \cdots dx_n, \end{aligned} \quad (2.12)$$

dove si è posto $\bar{\sigma}^2 = \sigma^2/(1-2t)$. Dalla condizione di normalizzazione della densità (2.10) segue poi:

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left[-\frac{1}{2\bar{\sigma}^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right] dx_2 \cdots dx_n = \frac{(\sqrt{2\pi}\sigma)^{n-1}}{\sqrt{n}}.$$

Facendo uso della (2.1), dalla (2.12) per $t < 1/2$ si trae:

$$\begin{aligned} E \left\{ \exp \left[\frac{t}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \right] \middle| \bar{X} = \bar{x} \right\} &= \frac{\sqrt{n}}{(\sqrt{2\pi}\sigma)^{n-1}} \left(\frac{\sqrt{2\pi}\sigma}{\sqrt{1-2t}} \right)^{n-1} \frac{1}{\sqrt{n}} \\ &= (1-2t)^{-(n-1)/2}. \end{aligned} \quad (2.13)$$

La funzione generatrice dei momenti della variabile casuale (2.11) condizionata da $\bar{X} = \bar{x}$ non dipende dunque da \bar{x} . Ciò implica, com'è immediato mostrare, che anche la funzione generatrice dei momenti della variabile casuale $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)$, condizionata da $\bar{X} = \bar{x}$, non dipende da \bar{x} . Se ne conclude che \bar{X} e S^2 sono indipendenti. La (i) è dunque provata. La dimostrazione della (ii) segue poi dalla (2.13) in quanto questa è proprio la funzione generatrice dei momenti di una variabile casuale χ^2 con $n-1$ gradi di libertà. Alternativamente, la (ii) si può dimostrare osservando che in base alle definizioni di \bar{X} e S^2 si ha:

$$\begin{aligned} \frac{(n-1)S^2}{\sigma^2} &= \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} - \frac{\bar{X} - \mu}{\sigma} \right)^2 \\ &= \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 - 2 \frac{\bar{X} - \mu}{\sigma} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} + n \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 \\ &= \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 - \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2. \end{aligned} \quad (2.14)$$

Poiché $(X_i - \mu)/\sigma$ ha distribuzione normale standard, dal Teorema 2.1.1 segue che $(X_i - \mu)^2/\sigma^2$ ha distribuzione chi-quadrato con 1 grado di libertà. Per l'indipendenza delle variabili X_1, X_2, \dots, X_n , in virtù del Teorema 2.1.2 segue quindi che $\sum_{i=1}^n (X_i - \mu)^2/\sigma^2$ ha distribuzione chi-quadrato con n gradi di libertà. Inoltre, ricordando il Teorema 1.4.1, $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ ha distribuzione normale standard, così che il suo quadrato ha distribuzione chi-quadrato con

1 grado di libertà. Dalla proprietà (i) segue, ancora, che $(\bar{X} - \mu)^2 / (\sigma / \sqrt{n})^2$ e $(n - 1)S^2 / \sigma^2$ sono variabili casuali indipendenti.² Dal Teorema 2.1.3 e dalla (2.14) si conclude allora che $(n - 1)S^2 / \sigma^2$ ha distribuzione chi-quadrato con $n - 1$ gradi di libertà. ■

Come conseguenza immediata del Teorema 2.1.4 è possibile calcolare la varianza della varianza campionaria di una popolazione normale.

Corollario 2.1.2 Per la varianza campionaria S^2 di un campione casuale di taglia n estratto da una popolazione normale di varianza σ^2 si ha:

$$D^2(S^2) = \frac{2\sigma^4}{n-1}.$$

Dim. Nel Teorema 2.1.4 si è mostrato che $(n - 1)S^2 / \sigma^2$ ha distribuzione chi-quadrato con $n - 1$ gradi di libertà. Pertanto, ricordando la (2.8) si ottiene:

$$D^2\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1). \quad (2.15)$$

Poiché risulta

$$D^2\left[\frac{(n-1)S^2}{\sigma^2}\right] = \frac{(n-1)^2}{\sigma^4} D^2(S^2), \quad (2.16)$$

dalle (2.15) e (2.16) segue immediatamente la tesi. ■

La distribuzione chi-quadrato riveste un ruolo notevole in statistica non solo sotto il profilo teorico ma anche per le applicazioni cui conduce. Per questo motivo, poiché la funzione di distribuzione della variabile chi-quadrato non è in generale esprimibile in forma chiusa, sono state costruite tabelle di vario tipo coinvolgenti integrali della corrispondente densità di probabilità. In particolare, data una variabile casuale X avente distribuzione chi-quadrato con v gradi di libertà si indica con $\chi_{\alpha,v}^2$ il quantile superiore, ossia quel reale tale da aversi:

$$P(X \geq \chi_{\alpha,v}^2) \equiv \int_{\chi_{\alpha,v}^2}^{\infty} f_X(x) dx = \alpha,$$

con $0 < \alpha < 1$. La Tabella 3 dell'Appendice B riporta i valori di $\chi_{\alpha,v}^2$ per alcune scelte di α e di v .

Esempio 2.1.1 Si supponga che lo spessore del filamento di una lampadina sia una caratteristica critica e che il suo processo di realizzazione, in una fabbrica di componenti elettrici, sia da ritenersi sotto controllo se la variazione di spessore dei filamenti prodotti è caratterizzata, in millimetri, da una deviazione standard non superiore a $\sigma = 0.05$. Per un controllo di qualità vengono esaminati periodicamente lotti di 20 filamenti appena realizzati e per ciascun lotto vengono misurati gli spessori $(x_1, x_2, \dots, x_{20})$ dei filamenti. Si adotta poi il seguente

²Ricordiamo che se U_1, U_2, \dots, U_n sono variabili casuali indipendenti, tali risultano anche le variabili casuali $W_1 = \varphi_1(U_1), W_2 = \varphi_2(U_2), \dots, W_n = \varphi_n(U_n)$, dove $\varphi_1, \varphi_2, \dots, \varphi_n$ sono arbitrarie funzioni Borel-misurabili.

criterio di valutazione: il processo realizzativo viene ritenuto accettabile se la probabilità che la varianza campionaria S^2 assuma un valore maggiore o uguale a

$$s^2 = \frac{1}{19} \sum_{i=1}^{20} (x_i - \bar{x})^2 \quad (2.17)$$

è inferiore a 0.025. Supponiamo che una realizzazione osservata sia la seguente:

$$(1.05, 1.07, 1.12, 0.92, 0.98, 1.07, 1.11, 1.02, 0.98, 0.99, \\ 1.02, 1.12, 1.05, 0.99, 0.94, 1.05, 0.99, 1.11, 0.97, 1.12)$$

e che questa possa essere riguardata come estratta da una popolazione normale. Dalla (2.17) si ricava allora che il valore assunto dalla varianza campionaria è $s^2 = 3.95 \cdot 10^{-3}$. Pertanto, essendo $n = 20$ e $\sigma^2 = 2.5 \cdot 10^{-3}$, la probabilità da calcolare è

$$P(S^2 \geq 3.95 \cdot 10^{-3}) = P\left[\frac{(n-1)S^2}{\sigma^2} \geq 30.02\right]. \quad (2.18)$$

Poiché la popolazione è per ipotesi normale, in virtù del Teorema 2.1.4 la statistica $(n - 1)S^2 / \sigma^2$ ha distribuzione chi-quadrato con 19 gradi di libertà. Dalla Tabella 3 dell'Appendice B risulta $\chi_{0.025;19}^2 = 32.852$, di modo che

$$P\left[\frac{(n-1)S^2}{\sigma^2} \geq 32.852\right] = 0.025. \quad (2.19)$$

In definitiva, dalle (2.18) e (2.19) segue:

$$P(S^2 \geq 3.95 \cdot 10^{-3}) > 0.025,$$

da cui si conclude che, sulla base della realizzazione considerata, il processo di costruzione dei filamenti non può ritenersi accettabile. ♦

È opportuno qui mostrare come la distribuzione normale standard nasca anche come limite della distribuzione chi-quadrato al divergere del numero di gradi di libertà, come indicato dalla proposizione che segue.

Proposizione 2.1.3 Sia X una variabile chi-quadrato con v gradi di libertà. Al divergere di v , ciascuna delle seguenti variabili converge in distribuzione³ ad una variabile casuale normale standard:

$$(i) \frac{X - v}{\sqrt{2v}},$$

$$(ii) \sqrt{2X} - \sqrt{2v}.$$

³Si dice che una successione $\{X_n\}$ di variabili casuali converge in distribuzione ad una variabile casuale X se risulta $\lim_{n \rightarrow \infty} P(X_n \leq x) = P(X \leq x) \equiv F_X(x)$ in ogni $x \in \mathbb{R}$ che sia punto di continuità della funzione di distribuzione F_X di X .

Dim. Con riferimento alla (i), notiamo che, in virtù del Teorema 2.1.2, la variabile casuale X può essere riguardata come somma di v variabili casuali indipendenti ed identicamente distribuite, aventi distribuzione chi-quadrato con 1 grado di libertà. Conseguentemente la variabile $(X - v)/\sqrt{2v}$ può essere interpretata come una somma standardizzata di v variabili casuali. Per il teorema centrale del limite, per $v \rightarrow \infty$ essa tende allora ad una variabile casuale normale standard. Riguardo la (ii), osserviamo che per ogni $x \in \mathbb{R}$ si ha:

$$\begin{aligned} P(\sqrt{2X} - \sqrt{2v} \leq x) &= P\left(\sqrt{X} \leq \frac{x}{\sqrt{2}} + \sqrt{v}\right) \\ &= P\left(X \leq \frac{x^2}{2} + x\sqrt{2v} + v\right) \\ &= P\left(\frac{X - v}{\sqrt{2v}} \leq x + \frac{x^2}{2\sqrt{2v}}\right). \end{aligned}$$

Procedendo al limite per $v \rightarrow \infty$, in virtù di quanto già mostrato per la (i) si ricava:

$$\lim_{v \rightarrow \infty} P(\sqrt{2X} - \sqrt{2v} \leq x) = P(Z \leq x),$$

con Z variabile casuale normale standard. ■

2.2 Distribuzione di Student

In questo paragrafo prenderemo in esame un'altra variabile casuale, particolarmente utile in applicazioni statistiche, la cui funzione di distribuzione prende il nome dello pseudonimo "Student" con cui lo statistico W.S. Gosset firmò l'articolo in cui la introdusse.

Definizione 2.2.1 Una variabile casuale T continua si dice avere distribuzione di Student con v gradi di libertà, dove v è un intero positivo, se ha densità di probabilità

$$f_T(x) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\pi v} \Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}} \quad (x \in \mathbb{R}). \quad (2.20)$$

Si noti che per $v = 1$ la densità di Student si identifica con la densità di Cauchy.

Il grafico della densità di Student è riportato in Figura 2.2 per $v = 1$ e $v = 9$ gradi di libertà. Si noti come il suo andamento, soprattutto nel caso $v = 9$, sia reminiscente di quello relativo alla densità normale standard. In effetti l'approssimazione della densità di Student alla normale standard migliora al crescere del numero di gradi di libertà, come mostrato nella proposizione seguente.

Proposizione 2.2.1 Al divergere del numero v di gradi di libertà la densità di Student tende alla densità normale standard.

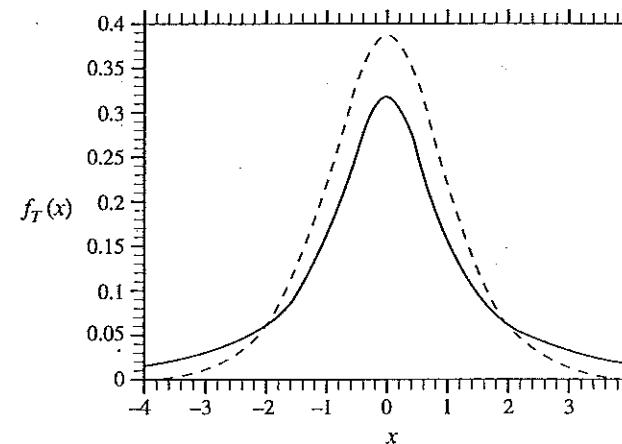


Figura 2.2: Densità di Student con $v = 1$ (curva continua) e $v = 9$ (curva tratteggiata) gradi di libertà.

Dim. Applicando la formula di Stirling⁴ alla densità (2.20), per $v \rightarrow \infty$ e per ogni $x \in \mathbb{R}$ si ha:

$$\begin{aligned} f_T(x) &= \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\pi v} \Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}} \\ &\sim \frac{1}{\sqrt{\pi v}} \frac{\left(\frac{v-1}{2}\right)^{(v-1)/2} e^{-(v-1)/2} \sqrt{2\pi \left(\frac{v-1}{2}\right)}}{\left(\frac{v}{2}-1\right)^{(v/2)-1} e^{-(v/2)+1} \sqrt{2\pi \left(\frac{v}{2}-1\right)}} \left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}} \\ &= \frac{1}{\sqrt{2\pi v}} \left(\frac{v-1}{v-2}\right)^{v/2} \frac{(v-1)^{-1/2}}{(v-2)^{-1}} e^{-1/2} \sqrt{\frac{v-1}{v-2}} \left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}}. \end{aligned}$$

Si può poi mostrare che sussistono i seguenti tre limiti:

$$\begin{aligned} \lim_{v \rightarrow \infty} \left(\frac{v-1}{v-2}\right)^{v/2} &= e^{1/2}, & \lim_{v \rightarrow \infty} \frac{(v-1)^{-1/2}}{(v-2)^{-1}} &= \sqrt{v}, \\ \lim_{v \rightarrow \infty} \sqrt{\frac{v-1}{v-2}} \left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}} &= e^{-x^2/2}; \end{aligned}$$

⁴La formula di Stirling afferma che risulta $\Gamma(\alpha+1) \sim \alpha^\alpha e^{-\alpha} \sqrt{2\pi\alpha}$, dove il simbolo " \sim " sta ad indicare che il rapporto delle due funzioni fra cui è posto tende ad 1 per $\alpha \rightarrow \infty$.

segue pertanto:

$$\lim_{v \rightarrow \infty} f_T(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad (x \in \mathbb{R}),$$

cosicché la tesi è dimostrata. ■

Come conseguenza pratica della Proposizione 2.2.1, si è soliti ritenere che l'approssimazione di una distribuzione di Student con una distribuzione normale standard sia soddisfacente per $v > 30$.

Nel seguente teorema è mostrato come si ricava una variabile di Student a partire da una variabile normale e da una variabile chi-quadrato.

Teorema 2.2.1 Siano X e Y variabili casuali indipendenti aventi rispettivamente distribuzione normale standard e distribuzione chi-quadrato con v gradi di libertà. La variabile casuale $T = X/\sqrt{Y/v}$ ha allora distribuzione di Student con v gradi di libertà.

Dim. Essendo X e Y indipendenti, la loro densità di probabilità congiunta è

$$f_{X,Y}(x,y) = \begin{cases} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \frac{y^{v/2-1} e^{-y/2}}{\Gamma(v/2) 2^{v/2}} & \text{per } x \in \mathbb{R}, y > 0, \\ 0 & \text{altrimenti.} \end{cases}$$

Mediane la trasformazione

$$T = g(X) = \frac{X}{\sqrt{Y/v}},$$

ovvero

$$X = g^{-1}(T) = T \sqrt{Y/v},$$

si ha:

$$f_{T,Y}(t,y) = f_{X,Y}[g^{-1}(t), y] \left| \frac{d}{dt} g^{-1}(t) \right|,$$

con $dg^{-1}(t)/dt = \sqrt{y/v}$. Per $t \in \mathbb{R}$ e $y > 0$ segue allora:

$$\begin{aligned} f_{T,Y}(t,y) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \frac{1}{\Gamma(v/2) 2^{v/2}} y^{v/2-1} e^{-y/2} \sqrt{\frac{y}{v}} \\ &= \frac{1}{\sqrt{\pi v} \Gamma(v/2)} 2^{-(v+1)/2} y^{(v-1)/2} \exp\left[-\frac{y}{2} \left(1 + \frac{t^2}{v}\right)\right]. \end{aligned}$$

Integriamo la densità bidimensionale $f_{T,Y}(t,y)$ rispetto ad y nell'intervallo $(0, \infty)$ per ottenere la densità marginale di T . Facendo uso del cambiamento di variabile $w = (1 + t^2/v)y/2$ nell'integrale, e ricordando la definizione (2.1) della funzione gamma, si ottiene:

$$\begin{aligned} f_T(t) &= \int_0^\infty \frac{1}{\sqrt{\pi v} \Gamma(v/2)} 2^{-(v+1)/2} y^{(v-1)/2} \exp\left[-\frac{y}{2} \left(1 + \frac{t^2}{v}\right)\right] dy \\ &= \int_0^\infty \frac{1}{\sqrt{\pi v} \Gamma(v/2)} 2^{-(v+1)/2} \left(\frac{2w}{1+t^2/v}\right)^{(v-1)/2} e^{-w} \frac{2}{1+t^2/v} dw \end{aligned}$$

2.2. DISTRIBUZIONE DI STUDENT

$$\begin{aligned} &= \frac{1}{\sqrt{\pi v} \Gamma(v/2)} \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2} \int_0^\infty w^{(v-1)/2} e^{-w} dw \\ &= \frac{1}{\sqrt{\pi v} \Gamma(v/2)} \Gamma\left(\frac{v+1}{2}\right) \left(1 + \frac{t^2}{v}\right)^{-(v+1)/2} \quad (t \in \mathbb{R}). \end{aligned}$$

La densità di probabilità di T è dunque quella di una variabile casuale avente distribuzione di Student con v gradi di libertà. ■

Come affermato nel Teorema 1.4.1, la media campionaria di un campione casuale di taglia n estratto da una popolazione normale di media μ e varianza σ^2 ha distribuzione normale di media μ e varianza σ^2/n . La difficoltà maggiore nell'applicazione di tale risultato risiede nella circostanza che solitamente la varianza del campione è incognita. Conviene pertanto sostituire questa con il valore della varianza campionaria che, come espresso dalla Proposizione 3.2.1, costituisce uno stimatore corretto della varianza σ^2 . Pertanto, quando la media μ è nota, diventa importante poter determinare la distribuzione campionaria della statistica

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}. \quad (2.21)$$

Nel teorema che segue vedremo che nel caso di campioni casuali normali la variabile (2.21) risulta avere distribuzione di Student.

Teorema 2.2.2 Se \bar{X} e S^2 sono la media campionaria e la varianza campionaria di un campione casuale di taglia n estratto da una popolazione normale di media μ , la variabile casuale $T = (\bar{X} - \mu)/(S/\sqrt{n})$ ha distribuzione di Student con $n - 1$ gradi di libertà.

Dim. Detta σ^2 la varianza della popolazione, dal Teorema 2.1.4 segue che le variabili casuali

$$X = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}, \quad Y = \frac{n-1}{\sigma^2} S^2$$

sono indipendenti. Esse hanno rispettivamente distribuzione normale standard e distribuzione chi-quadrato con $n - 1$ gradi di libertà. Facendo allora uso del Teorema 2.2.1 si conclude che la variabile casuale

$$T = \frac{X}{\sqrt{Y/(n-1)}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sqrt{\frac{\sigma^2}{S^2}} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

ha distribuzione di Student con $n - 1$ gradi di libertà. ■

Data l'importanza della distribuzione di Student in svariate applicazioni statistiche, sono stati tabulati integrali della sua densità di probabilità non essendo esprimibile in forma chiusa la sua funzione di distribuzione. In particolare, data una variabile casuale T avente distribuzione di Student con v gradi di libertà si denota con $t_{\alpha,v}$ il quantile superiore, definito dalla relazione

$$P(T \geq t_{\alpha,v}) \equiv \int_{t_{\alpha,v}}^\infty f_T(x) dx = \alpha, \quad (2.22)$$

con $0 < \alpha < 1$. Si noti che sussiste la relazione:

$$t_{1-\alpha;v} = -t_{\alpha;v}. \quad (2.23)$$

Alla dimostrazione della (2.23) si perviene procedendo in modo analogo al caso della Proposizione 1.4.1 poiché anche la densità di probabilità (2.20) di una variabile casuale di Student è una funzione pari. Nell'Appendice B è riportata la Tabella 4 in cui vengono elencati i valori di $t_{\alpha;v}$ corrispondenti a talune scelte di α e di v .

Esempio 2.2.1 Una società che effettua riparazioni di impianti termici afferma che i suoi operai sono in grado di portare a termine una generica riparazione nell'arco di 30 minuti primi (min). Si desidera verificare se tale affermazione è giustificata sapendo che nel corso di 25 interventi di riparazione gli operai hanno impiegato mediamente 40 min, con una varianza campionaria osservata pari a 14 min^2 . Riguardando questi dati come estratti da una popolazione normale, facendo ricorso al Teorema 2.2.2 si può assumere che la statistica

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

abbia distribuzione di Student con 24 gradi di libertà. Il valore di T in corrispondenza della realizzazione osservata è il seguente:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{40 - 30}{\sqrt{14/25}} = 13.36.$$

Poiché dalla Tabella 4 dell'Appendice B risulta $t_{0.005;24} = 2.797$, si ha

$$P(T > 2.797) = 0.005,$$

e quindi risulta trascurabile la probabilità che T assuma valori molto maggiori di qualche unità, quale risulta essere il valore $t = 13.36$ corrispondente alla realizzazione osservata. Ne consegue che il tempo impiegato per una riparazione deve essere ritenuto sensibilmente superiore a 30 minuti. ♦

2.3 Distribuzione di Fisher

Proseguiamo lo studio delle variabili casuali di interesse statistico introducendo la distribuzione che prende il nome dello statistico R.A. Fisher.

Definizione 2.3.1 Una variabile casuale X continua si dice avere distribuzione di Fisher con v_1 e v_2 gradi di libertà, dove v_1 e v_2 sono interi positivi, se essa ha densità di probabilità

$$f_X(x) = \begin{cases} \frac{\Gamma(\frac{v_1+v_2}{2})}{\Gamma(\frac{v_1}{2})\Gamma(\frac{v_2}{2})} \left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}} x^{\frac{v_1}{2}-1} \left(1 + \frac{v_1}{v_2}x\right)^{-\frac{v_1+v_2}{2}} & \text{per } x > 0, \\ 0 & \text{altrimenti.} \end{cases}$$

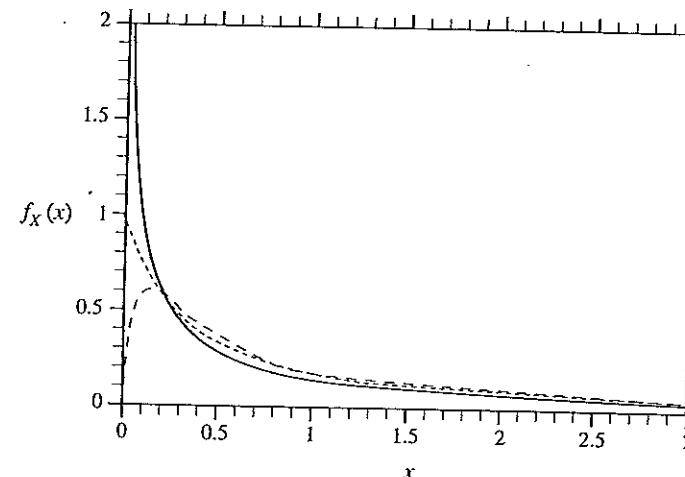


Figura 2.3: Densità di Fisher con gradi di libertà $v_1 = 1$ (curva continua), $v_1 = 2$ (curva tratteggiata) e $v_1 = 3$ (curva a tratti e punti); il secondo grado di libertà è $v_2 = 1$ in tutti e tre i casi.

Nella Figura 2.3 è mostrato il grafico della densità di Fisher quando il numero di gradi di libertà è $v_1 = 1$, $v_1 = 2$ e $v_1 = 3$, con $v_2 = 1$.

Nel teorema che segue si dimostra come sia possibile ricavare una variabile di Fisher a partire da due variabili chi-quadrato.

Teorema 2.3.1 Se Y_1 e Y_2 sono variabili casuali indipendenti aventi distribuzioni chi-quadrato con rispettivamente v_1 e v_2 gradi di libertà, la variabile casuale

$$X = \frac{Y_1/v_1}{Y_2/v_2}$$

ha distribuzione di Fisher con v_1 e v_2 gradi di libertà.

Dim. Poiché Y_1 e Y_2 sono indipendenti, la loro densità di probabilità congiunta è data da

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} \frac{y_1^{v_1/2-1} e^{-y_1/2}}{2^{v_1/2} \Gamma(v_1/2)} \frac{y_2^{v_2/2-1} e^{-y_2/2}}{2^{v_2/2} \Gamma(v_2/2)} & \text{per } y_1 > 0, y_2 > 0, \\ 0 & \text{altrimenti.} \end{cases}$$

Effettuando la trasformazione $X = g(Y_1) = (Y_1/v_1)/(Y_2/v_2)$, ovvero ponendo $Y_1 = g^{-1}(X) =$

$(v_1/v_2)Y_2 X$, si ricava:

$$f_{X,Y_2}(x,y_2) = f_{Y_1,Y_2}[g^{-1}(x),y_2] \left| \frac{d}{dx} g^{-1}(x) \right|,$$

con $dg^{-1}(x)/dx = (v_1/v_2)y_2$. Per $x > 0$ e $y_2 > 0$ si ha allora:

$$\begin{aligned} f_{X,Y_2}(x,y_2) &= \frac{1}{2^{v_1/2}\Gamma(v_1/2)} \left(\frac{v_1}{v_2} xy_2 \right)^{v_1/2-1} \exp\left(-\frac{v_1}{v_2} \frac{xy_2}{2}\right) \\ &\quad \times \frac{1}{2^{v_2/2}\Gamma(v_2/2)} y_2^{v_2/2-1} e^{-y_2/2} \frac{v_1}{v_2} y_2 \\ &= \frac{(v_1/v_2)^{v_1/2}}{2^{(v_1+v_2)/2}\Gamma(v_1/2)\Gamma(v_2/2)} x^{v_1/2-1} y_2^{(v_1+v_2)/2-1} \\ &\quad \times \exp\left[-\frac{y_2}{2} \left(\frac{v_1}{v_2} x + 1\right)\right]. \end{aligned}$$

Integriamo la densità $f_{X,Y_2}(x,y_2)$ rispetto ad y_2 nell'intervallo $(0, \infty)$ per determinare la densità marginale di X . Ponendo $z = (xv_1/v_2 + 1)y_2/2$ nell'integrale che segue, per $x > 0$ otteniamo:

$$\begin{aligned} f_X(x) &= \int_0^\infty \frac{(v_1/v_2)^{v_1/2}}{2^{(v_1+v_2)/2}\Gamma(v_1/2)\Gamma(v_2/2)} x^{v_1/2-1} y_2^{(v_1+v_2)/2-1} \\ &\quad \times \exp\left[-\frac{y_2}{2} \left(\frac{v_1}{v_2} x + 1\right)\right] dy_2 \\ &= \frac{(v_1/v_2)^{v_1/2}}{2^{(v_1+v_2)/2}\Gamma(v_1/2)\Gamma(v_2/2)} x^{v_1/2-1} \\ &\quad \times \int_0^\infty \frac{2}{xv_1/v_2 + 1} \left(\frac{2z}{xv_1/v_2 + 1}\right)^{(v_1+v_2)/2-1} e^{-z} dz \\ &= \frac{1}{\Gamma(v_1/2)\Gamma(v_2/2)} \left(\frac{v_1}{v_2}\right)^{v_1/2} \left(\frac{v_1}{v_2} x + 1\right)^{-(v_1+v_2)/2} x^{v_1/2-1} \\ &\quad \times \int_0^\infty z^{(v_1+v_2)/2-1} e^{-z} dz \\ &= \frac{\Gamma[(v_1+v_2)/2]}{\Gamma(v_1/2)\Gamma(v_2/2)} \left(\frac{v_1}{v_2}\right)^{v_1/2} x^{v_1/2-1} \left(\frac{v_1}{v_2} x + 1\right)^{-(v_1+v_2)/2}, \end{aligned}$$

dove si è fatto uso della definizione (2.1). La variabile casuale X ha dunque distribuzione di Fisher con v_1 e v_2 gradi di libertà. ■

Il Teorema 2.3.1 risulta particolarmente utile in problemi in cui si confrontano le varianze σ_1^2 e σ_2^2 di due popolazioni normali, ad esempio quando si vuole stimare il rapporto σ_1^2/σ_2^2 o verificare se risulta $\sigma_1^2 = \sigma_2^2$. Consideriamo due popolazioni normali e da ciascuna estraiamo un campione casuale. Denotate con n_1 e n_2 le relative taglie, per il Teorema 2.1.4 le variabili casuali

$$X_1 = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \quad (2.24)$$

$$X_2 = \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \quad (2.25)$$

hanno distribuzione chi-quadrato con $n_1 - 1$ e $n_2 - 1$ gradi di libertà rispettivamente, sotto l'ipotesi che le $n_1 + n_2$ variabili casuali che costituiscono i due campioni sono collettivamente indipendenti. Le variabili (2.24) e (2.25) sono indipendenti, così che dal Teorema 2.3.1 scaturisce il seguente teorema di dimostrazione immediata.

Teorema 2.3.2 *Siano S_1^2 e S_2^2 le varianze campionarie di due campioni casuali indipendenti di taglie n_1 e n_2 estratti da popolazioni normali di varianze σ_1^2 e σ_2^2 . La variabile casuale*

$$X = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

ha allora distribuzione di Fisher con $n_1 - 1$ e $n_2 - 1$ gradi di libertà.

Come per le altre distribuzioni finora esaminate, e per la stessa ragione, anche per la distribuzione di Fisher sono state costruite tabelle di integrali della densità di probabilità. Così, se X è una variabile casuale di Fisher con v_1 e v_2 gradi di libertà si denota con $F_{\alpha;v_1,v_2}$ il quantile superiore:

$$P(X \geq F_{\alpha;v_1,v_2}) \equiv \int_{F_{\alpha;v_1,v_2}}^\infty f_X(x) dx = \alpha, \quad (2.26)$$

con $0 < \alpha < 1$. Nell'Appendice B sono presenti le Tabelle 5 e 6; per alcune scelte di v_1 e v_2 la prima riporta i valori di $F_{0.01;v_1,v_2}$ e la seconda i valori di $F_{0.05;v_1,v_2}$.

Una variabile casuale di Fisher gode della peculiare proprietà che anche la sua reciproca ha distribuzione di Fisher. Sussiste invero la seguente Proposizione:

Proposizione 2.3.1 *Se X è una variabile casuale avente distribuzione di Fisher con v_1 e v_2 gradi di libertà, la variabile $1/X$ ha anch'essa distribuzione di Fisher con v_2 e v_1 gradi di libertà.*

Dim. Se X ha distribuzione di Fisher con v_1 e v_2 gradi di libertà esistono due variabili casuali X_1 e X_2 indipendenti, aventi distribuzione chi-quadrato con rispettivamente v_1 e v_2 gradi di libertà, tali da risultare:

$$X = \frac{X_1/v_1}{X_2/v_2},$$

ovvero:

$$\frac{1}{X} = \frac{X_2/v_2}{X_1/v_1}.$$

Però tanto, per il Teorema 2.3.2, $1/X$ ha distribuzione di Fisher con v_2 e v_1 gradi di libertà. ■

Proposizione 2.3.2 *Sussiste la seguente relazione:*

$$F_{1-\alpha;v_1,v_2} = \frac{1}{F_{\alpha;v_2,v_1}}. \quad (2.27)$$

Dim. Infatti, se X ha distribuzione di Fisher con v_1 e v_2 gradi di libertà, posto $Y = 1/X$ dalla (2.26) si trae:

$$\begin{aligned} 1 - \alpha &= P(X \geq F_{1-\alpha; v_1, v_2}) = P\left(\frac{1}{Y} \geq F_{1-\alpha; v_1, v_2}\right) \\ &= P\left(Y \leq \frac{1}{F_{1-\alpha; v_1, v_2}}\right), \end{aligned}$$

ovvero:

$$\alpha = 1 - P\left(Y \leq \frac{1}{F_{1-\alpha; v_1, v_2}}\right) = P\left(Y > \frac{1}{F_{1-\alpha; v_1, v_2}}\right). \quad (2.28)$$

Ricordando che, per la Proposizione 2.3.1, Y ha distribuzione di Fisher con v_2 e v_1 gradi di libertà si ha poi:

$$P(Y \geq F_{\alpha; v_2, v_1}) = \alpha. \quad (2.29)$$

Dalle (2.28) e (2.29) si ricava:

$$P\left(Y > \frac{1}{F_{1-\alpha; v_1, v_2}}\right) = P(Y \geq F_{\alpha; v_2, v_1}),$$

ossia:

$$\int_{1/F_{1-\alpha; v_1, v_2}}^{\infty} f_Y(x) dx = \int_{F_{\alpha; v_2, v_1}}^{\infty} f_Y(x) dx. \quad (2.30)$$

Dalla (2.30) segue infine la (2.27). ■

Esempio 2.3.1 Una casa automobilistica ha prodotto due diversi prototipi di automobili le cui prestazioni vengono valutate mediante percorrenze prestabilite in un autodromo. Il primo prototipo effettua $n = 8$ giri impiegando, in minuti, i seguenti tempi:

$$(5.42, 5.53, 6.05, 5.92, 5.31, 6.16, 5.92, 5.26).$$

Il secondo, invece, effettua $m = 10$ giri con i seguenti tempi:

$$(4.83, 4.96, 5.22, 5.36, 4.87, 5.01, 4.87, 5.13, 5.22, 4.98).$$

Supponiamo che i dati osservati possano essere riguardati come realizzazioni estratte da due popolazioni normali di varianze σ_1^2 e σ_2^2 uguali; ciò equivale ad ipotizzare che le prestazioni dei due prototipi siano soggette alla stessa variabilità. In virtù delle ipotesi fatte, dal Teorema 2.3.2 si ricava che la statistica

$$F = \frac{s_1^2}{s_2^2}$$

è una variabile di Fisher con 7 e 9 gradi di libertà. Dai dati osservati segue $s_1 = 0.3553$ e $s_2 = 0.1788$, così che

$$\left(\frac{s_1}{s_2}\right)^2 = \left(\frac{0.3553}{0.1788}\right)^2 = 3.9487$$

è il valore che assume F in corrispondenza delle due realizzazioni. Poiché dalla Tabella 6 Nell'Appendice B si ha $F_{0.005; 7, 9} = 3.29$, risulta:

$$P(F \geq 3.9487) < P(F \geq 3.29) = 0.05.$$

Si è quindi ricavato che è inferiore a cinque centesimi la probabilità che la statistica $F = S_1^2/S_2^2$ assuma un valore maggiore o uguale al valore s_1^2/s_2^2 corrispondente alle realizzazioni osservate. Essendo tale probabilità molto piccola, è lecito concludere che le variabilità delle prestazioni dei due prototipi non sono uguali. ◆

Concludiamo con due significative proprietà della distribuzione di Fisher, espresse dalle proposizioni che seguono.

Proposizione 2.3.3 *Sia X una variabile casuale di Fisher con v_1 e v_2 gradi di libertà. Al divergere di v_2 , la densità della variabile $v_1 X$ tende alla densità chi-quadrato con v_1 gradi di libertà.*

Dim. Le densità delle variabili casuali X e $v_1 X$ sono così legate:

$$f_{v_1 X}(x) = \frac{1}{v_1} f_X\left(\frac{x}{v_1}\right).$$

Pertanto, ricordando la Definizione 2.3.1, per $x > 0$ si ha:

$$f_{v_1 X}(x) = \frac{1}{v_1} \frac{\Gamma\left(\frac{v_1 + v_2}{2}\right)}{\Gamma\left(\frac{v_1}{2}\right) \Gamma\left(\frac{v_2}{2}\right)} \left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}} \left(\frac{x}{v_1}\right)^{\frac{v_1}{2}-1} \left(1 + \frac{x}{v_2}\right)^{-\frac{v_1+v_2}{2}}. \quad (2.31)$$

Osserviamo che, per la formula di Stirling, quando $v_2 \rightarrow \infty$ risulta:

$$\begin{aligned} \frac{\Gamma\left(\frac{v_1 + v_2}{2}\right)}{\Gamma\left(\frac{v_2}{2}\right)} &\sim \frac{\left(\frac{v_1 + v_2}{2} - 1\right)^{\frac{v_1+v_2}{2}-1} e^{-\frac{v_1+v_2}{2}+1} \sqrt{2\pi \left(\frac{v_1 + v_2}{2} - 1\right)}}{\left(\frac{v_2}{2} - 1\right)^{\frac{v_2}{2}-1} e^{-\frac{v_2}{2}+1} \sqrt{2\pi \left(\frac{v_2}{2} - 1\right)}} \\ &= \left(\frac{v_1 + v_2 - 2}{v_2 - 2}\right)^{\frac{v_2}{2}-1} \left(\frac{v_1 + v_2}{2} - 1\right)^{\frac{v_1}{2}} e^{-v_1/2} \sqrt{\frac{v_1 + v_2 - 2}{v_2 - 2}}. \end{aligned}$$

Dalla (2.31) si trae allora:

$$\begin{aligned} \lim_{v_2 \rightarrow \infty} f_{v_1 X}(x) &= \frac{x^{(v_1/2)-1}}{\Gamma(v_1/2)} e^{-v_1/2} \lim_{v_2 \rightarrow \infty} \left(1 + \frac{x}{v_2}\right)^{-\frac{v_1+v_2}{2}} \sqrt{\frac{v_1 + v_2 - 2}{v_2 - 2}} \\ &\times \lim_{v_2 \rightarrow \infty} \left(\frac{v_1 + v_2 - 2}{v_2 - 2}\right)^{\frac{v_2}{2}-1} \left(\frac{v_1 + v_2 - 2}{2v_2}\right)^{\frac{v_1}{2}}. \end{aligned}$$

Potendosi inoltre dimostrare che

$$\lim_{v_2 \rightarrow \infty} \left(1 + \frac{x}{v_2}\right)^{-\frac{v_1+v_2}{2}} \sqrt{\frac{v_1 + v_2 - 2}{v_2 - 2}} = e^{-x/2}$$

e che

$$\lim_{v_2 \rightarrow \infty} \left(\frac{v_1 + v_2 - 2}{v_2 - 2}\right)^{\frac{v_2}{2}-1} \left(\frac{v_1 + v_2 - 2}{2v_2}\right)^{\frac{v_1}{2}} = e^{v_1/2} \left(\frac{1}{2}\right)^{v_1/2},$$

si ricava immediatamente:

$$\lim_{v_2 \rightarrow \infty} f_{V_1 X}(x) = \frac{x^{(v_1/2)-1} e^{-x/2}}{2^{v_1/2} \Gamma(v_1/2)} \quad (x > 0),$$

dove il secondo membro è la densità chi-quadrato caratterizzata da v_1 gradi di libertà. ■

Proposizione 2.3.4 *Sia X una variabile casuale di Student con v gradi di libertà. La variabile $Y = X^2$ ha allora distribuzione di Fisher con 1 e v gradi di libertà.*

Dim. Come visto nel corso della dimostrazione del Teorema 2.1.1, le densità di Y e di X sono così legate:

$$f_Y(y) = \begin{cases} \frac{f_X(\sqrt{y})}{\sqrt{y}} & \text{per } y > 0, \\ 0 & \text{altrimenti.} \end{cases}$$

Ricordando l'espressione (2.20) di $f_X(x)$, per $y > 0$ si ha allora:

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{y}} \frac{\Gamma(\frac{v+1}{2})}{\sqrt{\pi v} \Gamma(\frac{v}{2})} \left(1 + \frac{y}{v}\right)^{-\frac{v+1}{2}} \\ &= \frac{\Gamma(\frac{1+v}{2})}{\Gamma(\frac{1}{2}) \Gamma(\frac{v}{2})} \left(\frac{1}{v}\right)^{1/2} y^{-1/2} \left(1 + \frac{y}{v}\right)^{-\frac{1+v}{2}}. \end{aligned}$$

Confrontando, infine, la densità appena ottenuta con quella di Fisher si ricava che la variabile Y ha distribuzione di Fisher con 1 e v gradi di libertà. ■

2.4 Distribuzione normale multivariata

La densità di probabilità congiunta di un campione casuale estratto da una popolazione normale è esplicitabile agevolmente essendo il prodotto delle densità di probabilità delle singole variabili in virtù della loro indipendenza. Talora, però, ci si imbatte in situazioni in cui compaiono n variabili casuali congiuntamente normali ma non indipendenti. La densità di probabilità congiunta di tali variabili non è pertanto esprimibile come prodotto delle densità marginali. Invero, si dà la seguente definizione di variabili casuali congiuntamente normali.

Definizione 2.4.1 *Le variabili casuali X_1, X_2, \dots, X_n sono dette congiuntamente normali, ovvero dotate di distribuzione normale multivariata, se esistono n variabili normali standard Z_1, Z_2, \dots, Z_n indipendenti, n costanti reali m_1, m_2, \dots, m_n ed una matrice $A \equiv \|a_{i,j}\|$ non singolare $n \times n$ tali da aversi*

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix} + \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{pmatrix}.$$

2.4. DISTRIBUZIONE NORMALE MULTIVARIATA

In termini matriciali questa relazione, con ovvia notazione, può riscriversi come

$$\mathbf{X} = A\mathbf{Z} + \mathbf{m}.$$

Determiniamo la densità di probabilità congiunta $f_{\mathbf{X}}(x_1, \dots, x_n)$ del vettore casuale $\mathbf{X} = (X_1, \dots, X_n)^T$. Per far ciò cominciamo col ricordare che la densità di probabilità di $\mathbf{Z} = (Z_1, \dots, Z_n)^T$ è

$$\begin{aligned} f_{\mathbf{Z}}(z_1, \dots, z_n) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2} \sum_{i=1}^n z_i^2\right) \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \mathbf{z}^T \mathbf{z}\right). \end{aligned}$$

Consideriamo poi la trasformazione

$$\mathbf{z} = A^{-1}(\mathbf{x} - \mathbf{m}),$$

dove $A^{-1} = \|b_{ij}\|$ denota la matrice inversa di A . È facile verificare che risulta $\partial z_i / \partial x_j = b_{ij}$, così che lo Jacobiano della trasformazione coincide con il determinante di A^{-1} , che denotremo con $|A^{-1}|$. Si ha dunque:

$$\begin{aligned} f_{\mathbf{X}}(x_1, \dots, x_n) &= \left|\frac{\partial \mathbf{z}}{\partial \mathbf{x}}\right| f_{\mathbf{Z}}(z_1, \dots, z_n) \\ &= \frac{|A^{-1}|}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \mathbf{m})^T (A^{-1})^T A^{-1} (\mathbf{x} - \mathbf{m})\right]. \end{aligned} \quad (2.32)$$

Introduciamo la matrice C tale che $C = AA^T$. Poiché $(A^T)^{-1} = (A^{-1})^T$ si ha:

$$(A^{-1})^T A^{-1} = (A^T)^{-1} A^{-1} = (AA^T)^{-1} = C^{-1}.$$

Pertanto $|A^{-1}|^2 = |C^{-1}|$ e quindi:

$$|A^{-1}| = \sqrt{|C^{-1}|}.$$

Dalla (2.32) si trae allora:

$$f_{\mathbf{X}}(x_1, \dots, x_n) = \frac{\sqrt{|C^{-1}|}}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \mathbf{m})^T C^{-1} (\mathbf{x} - \mathbf{m})\right]. \quad (2.33)$$

Rimangono da determinare la matrice C e il vettore \mathbf{m} . Dimostriamo innanzitutto il seguente lemma.

Lemma 2.4.1 *La matrice C^{-1} è simmetrica e definita positiva.*

Dim. C^{-1} è simmetrica, poiché

$$(C^{-1})^T = [(AA^T)^{-1}]^T = [(AA^T)]^{-1} = (AA^T)^{-1} = C^{-1}.$$

Essa inoltre è definita positiva; infatti per $\mathbf{x} \neq \mathbf{0}$ si ha:

$$\mathbf{x}^T C^{-1} \mathbf{x} = \mathbf{x}^T (A^{-1})^T A^{-1} \mathbf{x} = (A^{-1} \mathbf{x})^T (A^{-1} \mathbf{x}) > 0.$$

Definizione 2.4.2 Se $H = \{H_{ij}\}$ è una matrice di variabili casuali, $E(H)$ denota la matrice $\{E(H_{ij})\}$ dei valori medi. Inoltre, se $G(x) = \{g_{ij}(x)\}$ è una matrice di funzioni definite in un intervallo (a, b) , l'integrale $\int_a^b G(x) dx$ denota la matrice di elementi $\left| \int_a^b g_{ij}(x) dx \right|$. Infine, dx^T denota il prodotto $dx_1 dx_2 \dots dx_n$ presente negli integrali n -dimensionali.

Siamo ora in grado di determinare la densità di probabilità congiunta delle n variabili congiuntamente normali X_1, X_2, \dots, X_n e di fornire un'interpretazione del vettore \mathbf{m} e della matrice C .

Teorema 2.4.1 Se X_1, X_2, \dots, X_n sono variabili casuali congiuntamente normali, la loro densità di probabilità congiunta è

$$f_X(x_1, \dots, x_n) = \frac{\sqrt{|C^{-1}|}}{(2\pi)^{n/2}} \exp \left[-\frac{1}{2} (\mathbf{x} - \mathbf{m})^T C^{-1} (\mathbf{x} - \mathbf{m}) \right], \quad (2.34)$$

dove $\mathbf{m} = E(\mathbf{X})$ è il vettore dei valori medi e dove $C = \{ \text{cov}(X_i, X_j) \}$ è la matrice di covarianza di \mathbf{X} .

Dim. Dalla relazione $\mathbf{X} = A\mathbf{Z} + \mathbf{m}$ e da $E(\mathbf{Z}) = \mathbf{0}$ segue immediatamente $E(\mathbf{X}) = \mathbf{m}$. Il vettore \mathbf{m} è dunque il vettore dei valori medi di \mathbf{X} . Se poi denotiamo con $C_0 = \{ \text{cov}(X_i, X_j) \}$ la matrice di covarianza di \mathbf{X} , si ha:

$$C_0 = E[(\mathbf{X} - \mathbf{m})(\mathbf{X} - \mathbf{m})^T] = \int_{\mathbb{R}^n} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T f_X(\mathbf{x}^T) d\mathbf{x}^T.$$

Dalla (2.33) segue allora:

$$C_0 = \frac{\sqrt{|C^{-1}|}}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T \exp \left[-\frac{1}{2} (\mathbf{x} - \mathbf{m})^T C^{-1} (\mathbf{x} - \mathbf{m}) \right] d\mathbf{x}^T$$

da cui, applicando la trasformazione $\mathbf{x} = A\mathbf{z} + \mathbf{m}$ di Jacobiano $|A| = \sqrt{|C|}$, si ottiene:

$$\begin{aligned} C_0 &= \frac{\sqrt{|C^{-1}|}}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} A \mathbf{z} \mathbf{z}^T A^T \exp \left(-\frac{1}{2} \mathbf{z}^T \mathbf{z} \right) \sqrt{|C|} d\mathbf{z}^T \\ &= A \left[(2\pi)^{-n/2} \int_{\mathbb{R}^n} \mathbf{z} \mathbf{z}^T \exp \left(-\frac{1}{2} \mathbf{z}^T \mathbf{z} \right) d\mathbf{z}^T \right] A^T. \end{aligned} \quad (2.35)$$

Poiché risulta:

$$(2\pi)^{-n/2} \int_{\mathbb{R}^n} \mathbf{z} \mathbf{z}^T \exp \left(-\frac{1}{2} \mathbf{z}^T \mathbf{z} \right) d\mathbf{z}^T = I,$$

dove I denota la matrice identità $n \times n$, dalla (2.35) segue:

$$C_0 = A I A^T = A A^T = C.$$

In conclusione, sussiste la (2.34), in cui \mathbf{m} è il vettore dei valori medi e C è la matrice di covarianza. ■

Un importante teorema, che per brevità ci limitiamo solo ad enunciare, è il seguente.

Teorema 2.4.2 Se X_1, X_2, \dots, X_n sono variabili casuali congiuntamente normali, ciascun gruppo di esse ha distribuzione normale multivariata.

È opportuno menzionare che può accadere che ciascuna delle n variabili casuali X_1, X_2, \dots, X_n abbia distribuzione normale, mentre la loro distribuzione congiunta non è normale, come mostra l'esempio che segue.

Esempio 2.4.1 Consideriamo le variabili casuali X_1 e X_2 di densità di probabilità congiunta:

$$f_{X_1, X_2}(x_1, x_2) = \begin{cases} \frac{1}{\pi} \exp \left(-\frac{x_1^2 + x_2^2}{2} \right) & \text{per } x_1 < 0, x_2 \geq 0 \\ & \text{e per } x_1 \geq 0, x_2 < 0, \\ 0 & \text{altrimenti} \end{cases}$$

che, evidentemente, non sono congiuntamente normali. È facile verificare che X_1 e X_2 non sono indipendenti in quanto ad esempio nei punti in cui risulta $f_{X_1, X_2}(x_1, x_2) = 0$ si ha $f_{X_1, X_2}(x_1, x_2) \neq f_{X_1}(x_1) f_{X_2}(x_2)$, dove con $f_{X_i}(x_i)$ si sono denotate le densità marginali delle X_i . Queste ultime, come si ricava per integrazione della densità congiunta, sono densità normali standard. ♦

Analizzeremo ora in dettaglio la distribuzione normale multivariata nel caso particolare $n = 2$ denotando la matrice di covarianza C nel seguente modo:

$$C = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

dove

$$\sigma_i^2 = D^2(X_i) \quad (i = 1, 2), \quad \sigma_{12} = \text{cov}(X_1, X_2).$$

Indichiamo poi con

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

il coefficiente di correlazione di X_1 e X_2 . Poiché in virtù del Lemma 2.4.1 la matrice C^{-1} è simmetrica e definita positiva, anche la sua inversa, ossia C , è tale. Pertanto il suo determinante è positivo:

$$|C| = \begin{vmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{vmatrix} = \sigma_1^2 \sigma_2^2 - \sigma_{12}^2 = \sigma_1^2 \sigma_2^2 (1 - \rho^2) > 0.$$

Inoltre, la matrice C^{-1} è data da

$$C^{-1} = \frac{1}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)} \begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix}.$$

Da quanto visto finora segue che l'espressione in parentesi quadra a secondo membro della (2.34) assume la seguente forma:

$$\begin{aligned} -\frac{1}{2} (\mathbf{x} - \mathbf{m})^T C^{-1} (\mathbf{x} - \mathbf{m}) &= -\frac{1}{2(1 - \rho^2)} \\ &\times \left[\left(\frac{x_1 - m_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_1 - m_1}{\sigma_1} \right) \left(\frac{x_2 - m_2}{\sigma_2} \right) + \left(\frac{x_2 - m_2}{\sigma_2} \right)^2 \right]. \end{aligned} \quad (2.36)$$

Facendo uso della (2.36) e osservando che

$$|C^{-1}| = \frac{1}{|C|} = \frac{1}{\sigma_1^2 \sigma_2^2 (1 - \rho^2)},$$

dalla (2.34) segue l'espressione della densità normale bivariata:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} \times \left[\left(\frac{x_1-m_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-m_1}{\sigma_1}\right)\left(\frac{x_2-m_2}{\sigma_2}\right) + \left(\frac{x_2-m_2}{\sigma_2}\right)^2 \right]\right\}. \quad (2.37)$$

Osservazione 2.4.1 Si noti che la densità (2.37) è normalizzata, ossia che risulta:

$$I \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_2 dx_1 = 1.$$

Dim. Effettuando la sostituzione

$$y_1 = \frac{x_1-m_1}{\sigma_1}, \quad y_2 = \frac{x_2-m_2}{\sigma_2}$$

si ha:

$$\begin{aligned} I &= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left[-\frac{y_1^2 - 2\rho y_1 y_2 + y_2^2}{2(1-\rho^2)}\right] dy_2 dy_1 \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\left[-\frac{(y_1 - \rho y_2)^2 + (1-\rho^2)y_2^2}{2(1-\rho^2)}\right] dy_2 dy_1. \end{aligned}$$

Ponendo poi

$$z = \frac{y_1 - \rho y_2}{\sqrt{1-\rho^2}},$$

si ricava:

$$I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2} dz \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y_2^2/2} dy_2 = 1,$$

dove l'ultima uguaglianza segue dalla condizione di normalizzazione della densità normale standard. ■

L'osservazione che segue è una verifica diretta di quanto espresso dal Teorema 2.4.2.

Osservazione 2.4.2 Se X_1, X_2 sono variabili congiuntamente normali dotate di densità congiunta (2.37), allora X_1 e X_2 sono variabili normali aventi densità marginali

$$f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{x_1-m_1}{\sigma_1}\right)^2\right] \quad (2.38)$$

e

$$f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{1}{2}\left(\frac{x_2-m_2}{\sigma_2}\right)^2\right], \quad (2.39)$$

rispettivamente.

2.4. DISTRIBUZIONE NORMALE MULTIVARIATA

Dim. Osserviamo anzitutto che per definizione di densità marginale si ha:

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2,$$

dove $f(x_1, x_2)$ è data dalla (2.37). Allo scopo di ottenere un quadrato perfetto, aggiungiamo e sottraiamo la quantità $\rho^2(x_1 - m_1)^2/\sigma_1^2$ nell'espressione in parentesi quadra della (2.37). Si ricava così:

$$\begin{aligned} f_{X_1}(x_1) &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2}\left(\frac{x_1-m_1}{\sigma_1}\right)^2 - \frac{1}{2(1-\rho^2)} \left[\left(\frac{x_2-m_2}{\sigma_2}\right) - \rho\left(\frac{x_1-m_1}{\sigma_1}\right)\right]^2\right\} dx_2. \end{aligned}$$

Ponendo

$$z = \frac{1}{\sqrt{2(1-\rho^2)}} \left[\left(\frac{x_2-m_2}{\sigma_2}\right) - \rho\left(\frac{x_1-m_1}{\sigma_1}\right)\right],$$

si ottiene poi:

$$f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{x_1-m_1}{\sigma_1}\right)^2\right] \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

Poiché l'integrale a secondo membro vale 1, segue che la densità marginale $f_{X_1}(x_1)$ è data dalla (2.38), che coincide con la densità di probabilità di una variabile normale di valore medio m_1 e varianza σ_1^2 . Procedendo in maniera analoga si ricava che la densità marginale $f_{X_2}(x_2)$ è data dalla (2.39), che denota la densità di probabilità di una variabile normale di media m_2 e varianza σ_2^2 . ■

Dalle espressioni della densità congiunta (2.37) e delle densità marginali (2.38) e (2.39) si ricava che le variabili X_1, X_2 congiuntamente normali sono indipendenti se e solo se X_1 e X_2 sono non correlate. Infatti, risulta $f(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$ se e solo se è $\rho = 0$.

Notevole importanza rivestono le densità condizionate di variabili casuali normali; le relative espressioni sono fornite dalla Osservazione 2.4.3.

Osservazione 2.4.3 Se X_1, X_2 sono variabili congiuntamente normali dotate di densità congiunta (2.37), allora la densità di X_1 condizionata da X_2 e la densità di X_2 condizionata da X_1 sono date da

$$\begin{aligned} f_{X_1|X_2}(x_1|x_2) &= \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_1} \\ &\times \exp\left\{-\frac{1}{2\sigma_1(1-\rho^2)} \left[x_1 - m_1 - \rho\frac{\sigma_1}{\sigma_2}(x_2 - m_2)\right]^2\right\} \quad (2.40) \end{aligned}$$

e

$$\begin{aligned} f_{X_2|X_1}(x_2|x_1) &= \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma_2} \\ &\times \exp\left\{-\frac{1}{2\sigma_2(1-\rho^2)} \left[x_2 - m_2 - \rho\frac{\sigma_2}{\sigma_1}(x_1 - m_1)\right]^2\right\}, \quad (2.41) \end{aligned}$$

rispettivamente.

Dim. Le densità condizionate delle variabili X_1 e X_2 si ottengono dalla densità congiunta (2.37) e dalle densità marginali (2.38) e (2.39) mediante le note relazioni:

$$f_{X_1|X_2}(x_1|x_2) = \frac{f(x_1, x_2)}{f_{X_2}(x_2)}, \quad f_{X_2|X_1}(x_2|x_1) = \frac{f(x_1, x_2)}{f_{X_1}(x_1)}.$$

Con facili calcoli si ricava così che la densità condizionata di X_1 dato $X_2 = x_2$ è normale di valore medio

$$E(X_1|X_2 = x_2) = m_1 + p \frac{\sigma_1}{\sigma_2} (x_2 - m_2) \quad (2.42)$$

e varianza

$$D^2(X_1|X_2 = x_2) = \sigma_1^2(1 - p^2).$$

Si ha quindi la (2.40). In modo analogo si dimostra che la densità condizionata di X_2 dato $X_1 = x_1$ è normale di valore medio

$$E(X_2|X_1 = x_1) = m_2 + p \frac{\sigma_2}{\sigma_1} (x_1 - m_1)$$

e varianza

$$D^2(X_2|X_1 = x_1) = \sigma_2^2(1 - p^2). \quad (2.43)$$

Segue quindi la (2.41). \blacksquare

Una importante proprietà di cui godono le variabili casuali normali è espressa dalla proposizione che segue.

Proposizione 2.4.1 Se X_1, X_2, \dots, X_n sono variabili normali indipendenti di valori medi $\mu_1, \mu_2, \dots, \mu_n$ e varianze $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$, allora la variabile $Y = \sum_{i=1}^n c_i X_i$ è normale di media $\sum_{i=1}^n c_i \mu_i$ e varianza $\sum_{i=1}^n c_i^2 \sigma_i^2$.

Dim. Segue facilmente utilizzando un procedimento analogo a quello seguito per la dimostrazione dei Teoremi 1.4.1 e 1.4.2. \blacksquare

2.5 Distribuzione normale inversa

Di interesse in questioni applicative, tra cui problemi di assorbimento in processi stocastici di tipo diffusivo, è anche la cosiddetta distribuzione normale inversa.

Definizione 2.5.1 Una variabile casuale X continua si dice avere distribuzione normale inversa di parametri μ e λ reali positivi se ha densità di probabilità

$$f_X(x) = \begin{cases} \left(\frac{\lambda}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right] & \text{per } x > 0, \\ 0 & \text{altrimenti.} \end{cases} \quad (2.44)$$

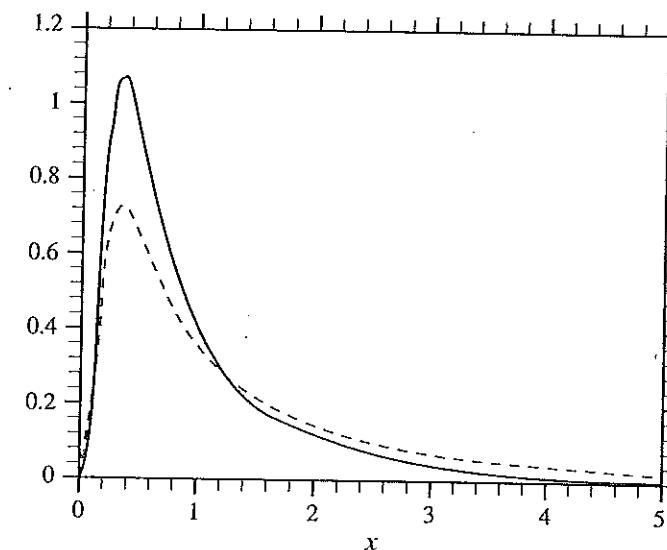


Figura 2.4: Densità normale inversa con $\lambda = 1$ e con $\mu = 1$ (curva continua) e $\mu = 2$ (curva tratteggiata).

Nella Figura 2.4 sono riportati i grafici delle densità normali inverse per $\lambda = 1$ e per $\mu = 1, \mu = 2$. Si noti che essi sono unimodali. In realtà si può dimostrare che la densità di una variabile normale inversa è sempre unimodale.

Val la pena menzionare che ponendo

$$\mu = \frac{S - x_0}{\eta}, \quad \lambda = \frac{(S - x_0)^2}{\sigma^2}$$

la (2.44) fornisce la densità di probabilità del tempo di primo passaggio attraverso lo stato S di un processo di Wiener $\{W(x), x \geq 0\}$ di deriva η e varianza infinitesimale σ^2 avente origine in x_0 ($x_0 < S$), ossia tale da risultare $P[W(0) = x_0] = 1$.

Si noti che esiste il limite della (2.44) per $\mu \rightarrow \infty$, che è ancora una densità di probabilità. Questa è interpretabile come la densità del tempo di primo passaggio (questa volta a valore medio infinito) attraverso lo stato S di un processo di Wiener $\{W(x), x \geq 0\}$ di deriva nulla e varianza infinitesimale σ^2 avente origine in x_0 ($x_0 < S$).

Nella seguente proposizione si ricavano funzione generatrice dei momenti, media e varianza della variabile casuale normale inversa.

Proposizione 2.5.1 La funzione generatrice dei momenti di una variabile casuale X di di-

stribuzione normale inversa è

$$M_X(t) = \exp \left\{ \frac{\lambda}{\mu} \left[1 - \left(1 - \frac{2\mu^2 t}{\lambda} \right)^{1/2} \right] \right\} \quad \left(t < \frac{\lambda}{2\mu^2} \right). \quad (2.45)$$

Inoltre, valore medio e varianza di X sono:

$$E(X) = \mu, \quad D^2(X) = \frac{\mu^3}{\lambda}.$$

Dim. Per definizione di funzione generatrice dei momenti si ha:

$$\begin{aligned} M_X(t) &= \int_0^\infty e^{tx} \left(\frac{\lambda}{2\pi x^3} \right)^{1/2} \exp \left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x} \right] dx \\ &= \left(\frac{\lambda}{2\pi} \right)^{1/2} e^{\lambda/\mu} \int_0^\infty x^{-3/2} \exp \left[-\frac{\lambda}{2x} - x \left(\frac{\lambda}{2\mu^2} - t \right) \right] dx. \end{aligned}$$

Di qui segue immediatamente la (2.45) facendo uso del seguente risultato⁵

$$\int_0^\infty x^{-3/2} \exp \left(-\frac{\beta}{x} - \gamma x \right) dx = \left(\frac{\pi}{\beta} \right)^{1/2} \exp(-2\sqrt{\beta}\gamma),$$

valido per ogni β e γ positivi. Dalla (2.45) si ricava poi il valore medio:

$$E(X) = \frac{\partial}{\partial t} M_X(t) \Big|_{t=0} = M_X(0) \mu \left(1 - \frac{2\mu^2 t}{\lambda} \right)^{-1/2} \Big|_{t=0} = \mu.$$

Il momento del secondo ordine è invece:

$$\begin{aligned} E(X^2) &= \frac{\partial^2}{\partial t^2} M_X(t) \Big|_{t=0} \\ &= M_X(0) \left[\mu^2 \left(1 - \frac{2\mu^2 t}{\lambda} \right)^{-1} + \frac{\mu^3}{\lambda} \left(1 - \frac{2\mu^2 t}{\lambda} \right)^{-3/2} \right] \Big|_{t=0} = \mu^2 + \frac{\mu^3}{\lambda}. \end{aligned}$$

Si ottiene così la varianza di X :

$$D^2(X) = E(X^2) - [E(X)]^2 = \frac{\mu^3}{\lambda},$$

il che completa la dimostrazione. ■

È interessante notare che se X ha distribuzione normale inversa di parametri μ e λ , per ogni $c > 0$ la variabile $Y = cX$ risulta avere distribuzione normale inversa di parametri $c\mu$ e $c\lambda$. Per analogia con la distribuzione normale, si potrebbe essere indotti a ritenere che combinazioni lineari di variabili normali inverse abbiano ancora distribuzione normale inversa. Esiste invece una diversa proprietà additiva, espressa dalla proposizione che segue.

⁵Cfr., ad esempio, formula (27) pag. 146 del § 4.5 in A. Erdélyi, W. Magnus, F. Oberhettinger and F.G. Tricomi (1954), *Tables of Integral Transforms*, Volume I, McGraw-Hill, N.Y., nella quale si ponga $a = \beta/4$ e $p = \gamma$.

Proposizione 2.5.2 Siano X_1, X_2, \dots, X_n variabili casuali indipendenti a distribuzioni normali inverse di parametri μ_i e λ_i ($i = 1, 2, \dots, n$) tali da risultare

$$\frac{\lambda_1}{c_1 \mu_1^2} = \frac{\lambda_2}{c_2 \mu_2^2} = \dots = \frac{\lambda_n}{c_n \mu_n^2} = \xi, \quad (2.46)$$

dove c_1, c_2, \dots, c_n e ξ sono costanti positive. La variabile casuale $Y = \sum_{i=1}^n c_i X_i$ ha allora distribuzione normale inversa di parametri

$$\mu = \sum_{i=1}^n c_i \mu_i, \quad \lambda = \xi \left(\sum_{i=1}^n c_i \mu_i \right)^2. \quad (2.47)$$

Dim. In virtù dell'ipotesi (2.46) e dell'espressione (2.45), la funzione generatrice dei momenti di Y è

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^n M_{X_i}(c_i t) = \prod_{i=1}^n \exp \left\{ \frac{\lambda_i}{\mu_i} \left[1 - \left(1 - \frac{2\mu_i^2 c_i t}{\lambda_i} \right)^{1/2} \right] \right\} \\ &= \exp \left\{ \left(\frac{\lambda_1}{\mu_1} + \dots + \frac{\lambda_n}{\mu_n} \right) \left[1 - \left(1 - \frac{2t}{\xi} \right)^{1/2} \right] \right\}, \end{aligned}$$

per

$$t < \frac{\xi}{2} \min \{c_1, \dots, c_n\}.$$

Confrontando $M_Y(t)$ con la funzione generatrice (2.45) segue immediatamente che Y è una variabile casuale normale inversa i cui parametri μ e λ soddisfano le seguenti relazioni:

$$\frac{\lambda}{\mu^2} = \xi, \quad \frac{\lambda}{\mu} = \frac{\lambda_1}{\mu_1} + \dots + \frac{\lambda_n}{\mu_n} \equiv \xi(c_1 \mu_1 + \dots + c_n \mu_n).$$

Da queste si ricava infine che i valori di μ e λ sono proprio quelli specificati nella (2.47). ■

È ora opportuno indicare il motivo per il quale la variabile casuale normale inversa è così denominata. A tal fine, data una variabile casuale Z , consideriamo la funzione

$$L_Z(t) = \ln E(e^{-tZ}),$$

che qui chiameremo, con un abuso di terminologia, *funzione generatrice dei cumulanti* di Z .

Definizione 2.5.2 Due variabili casuali X e Y si dicono inverse se esiste una funzione $L(t)$ tale che le rispettive funzioni generatrici dei cumulanti $L_X(t)$ e $L_Y(t)$ soddisfano le condizioni

$$L_X(t) = \alpha L(t), \quad L_Y(t) = \beta L^{-1}(t) \quad (2.48)$$

per ogni reale t comune ai domini di $L_X(t)$ e $L_Y(t)$, con α e β costanti reali. Coppie di funzioni di distribuzione le cui funzioni generatrici dei cumulanti soddisfano le relazioni (2.48) sono poi dette funzioni di distribuzione inverse.

Verifichiamo che la funzione di distribuzione normale e la funzione di distribuzione normale inversa sono distribuzioni inverse secondo la Definizione 2.5.2. A tal fine, sia X una variabile casuale normale di media m e varianza σ^2 :

$$f_X(x) = \left(\frac{1}{2\pi\sigma^2} \right)^{1/2} \exp \left[-\frac{(x-m)^2}{2\sigma^2} \right] \quad (x \in \mathbb{R}).$$

È facile dimostrare che risulta:

$$L_X(t) = -mt + \frac{1}{2}\sigma^2 t^2 \quad (t \in \mathbb{R}).$$

Se poniamo

$$L(t) = -t + \frac{\sigma^2}{2m} t^2 \quad (m \neq 0),$$

si ha:

$$L_X(t) = m L(t).$$

Se si restringe $L(t)$ all'intervallo $t < 2m/\sigma^2$, la funzione inversa di $L(t)$ è

$$L^{-1}(t) = \frac{m}{\sigma^2} \left[1 - \left(1 + \frac{2\sigma^2}{m} t \right)^{1/2} \right]. \quad (2.49)$$

Osserviamo che se Y è una variabile casuale normale inversa di parametri μ e λ , dalla (2.45) si ricava facilmente che la funzione generatrice dei cumulantini è

$$L_Y(t) = \frac{\lambda}{\mu} \left[1 - \left(1 + \frac{2\mu^2}{\lambda} t \right)^{1/2} \right]. \quad (2.50)$$

Confrontando la (2.49) con la (2.50) si trae immediatamente che ponendo $\lambda/\mu^2 = m/\sigma^2$ risulta:

$$L_Y(t) = \mu L^{-1}(t).$$

Pertanto, in accordo con la Definizione 2.5.2, X e Y sono variabili casuali inverse se si ha $\lambda/\mu^2 = m/\sigma^2$.

Proposizione 2.5.3 Siano X e Y variabili casuali inverse soddisfacenti la relazione (2.48). Se si denotano con K_1 e K_2 primo e secondo cumulante della variabile casuale di funzione generatrice dei cumulantini $L(t)$, medie e varianze di X e Y sono date da

$$E(X) = \alpha K_1, \quad D^2(X) = \alpha K_2, \quad (2.51)$$

$$E(Y) = \frac{\beta}{K_1}, \quad D^2(Y) = \frac{\beta K_2}{(K_1)^3}. \quad (2.52)$$

Inoltre, i coefficienti di variazione di X e Y coincidono se e solo se è $\beta = \alpha K_1$; in tal caso risulta:

$$E(X) = \beta, \quad E(Y) = \alpha.$$

Dim. Essendo $L_X(t) = \alpha L(t)$, e poiché K_1 e K_2 sono rispettivamente media e varianza della variabile casuale avente funzione generatrice dei cumulantini $L(t)$, le (2.51) sono ovvie. Le (2.52) si ottengono invece osservando che risulta:

$$\frac{d}{dt} L^{-1}(t) = \left[\frac{d}{dt} L(t) \right]^{-1}, \quad \frac{d^2}{dt^2} L^{-1}(t) = \frac{d^2}{dt^2} L(t) \left[\frac{d}{dt} L(t) \right]^{-3}.$$

I coefficienti di variazione di X e Y si ricavano poi facilmente ricordando che il coefficiente di variazione $CV(Z)$ di una variabile casuale Z è così definito:

$$CV(Z) = \frac{D(Z)}{E(Z)}. \quad (2.53)$$

Facendo uso delle (2.51) e (2.52), dalla (2.53) si ottiene pertanto:

$$CV(X) = \frac{D(X)}{E(X)} = \frac{\sqrt{K_2}}{\sqrt{\alpha K_1}}, \quad CV(Y) = \frac{D(Y)}{E(Y)} = \sqrt{\frac{K_2}{\beta K_1}}.$$

Quindi $CV(X) = CV(Y)$ se e solo se $\alpha K_1 = \beta$, da cui segue $E(X) = \beta$ e $E(Y) = \alpha$. ■

Concludiamo queste considerazioni menzionando che altre due coppie di funzioni di distribuzione inverse sono le seguenti: a) binomiale e binomiale negativa; b) Poisson e gamma.

Capitolo 3

Stima puntuale

3.1 Statistiche d'ordine

Ci poniamo ora il problema dell'ordinamento dei valori assunti dalle variabili casuali costituenti un campione (X_1, X_2, \dots, X_n) estratto da una popolazione continua. Indichiamo con $X_{(1)}$ la statistica che descrive il minimo dei valori del campione, con $X_{(2)}$ quella che descrive il successivo in ordine di grandezza crescente e così via, fino a $X_{(n)}$ che descrive il valore massimo; le statistiche $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ vengono dette *statistiche d'ordine*¹. In particolare, $X_{(1)}$ è detta prima statistica d'ordine, $X_{(2)}$ seconda statistica d'ordine, ecc. Denotiamo poi con $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ i valori assunti dalle statistiche d'ordine in corrispondenza di una realizzazione (x_1, x_2, \dots, x_n) . Osserviamo che, poiché il campione è costituito da variabili casuali continue, è nulla la probabilità che due o più valori del campione siano uguali, ed è pertanto nulla la probabilità che due statistiche d'ordine assumano lo stesso valore.

Determineremo ora la funzione di distribuzione e la densità di probabilità di una generica statistica d'ordine.

Teorema 3.1.1 *La funzione di distribuzione $F_{(k)}(x)$ e la densità di probabilità $f_{(k)}(x)$ della k-esima statistica d'ordine $X_{(k)}$ di un campione casuale di taglia n estratto da una popolazione infinita continua di funzione di distribuzione $F(x)$ e densità di probabilità $f(x)$ sono*

$$F_{(k)}(x) = \sum_{j=k}^n \binom{n}{j} [F(x)]^j [1 - F(x)]^{n-j} \quad (k = 1, 2, \dots, n) \quad (3.1)$$

$$f_{(k)}(x) = \binom{n}{k} k [F(x)]^{k-1} [1 - F(x)]^{n-k} f(x) \quad (k = 1, 2, \dots, n). \quad (3.2)$$

Dim. Cominciamo col calcolare la funzione di distribuzione $F_{(k)}(x) = P(X_{(k)} \leq x)$ di $X_{(k)}$ osservando che l'evento $\{X_{(k)} \leq x\}$ equivale all'evento "almeno k delle variabili del campione casuale sono minori o uguali a x "; infatti l'evento $\{X_{(k)} \leq x\}$ si verifica se e solo se

¹Talora, quando è opportuno evidenziare la dipendenza dalla taglia del campione, le statistiche d'ordine vengono denotate con $X_{(1:n)}, X_{(2:n)}, \dots, X_{(n:n)}$.

si verificano almeno k degli eventi $\{X_1 \leq x\}, \{X_2 \leq x\}, \dots, \{X_n \leq x\}$. Questi sono collettivamente indipendenti, ciascuno con probabilità $F(x)$. Ricordando che la probabilità che si verifichino esattamente j di n eventi indipendenti aventi uguali probabilità di occorrenza è espressa dalla distribuzione binomiale, si ricava immediatamente la (3.1). Si noti che poiché X è continua, tale è anche $X_{(k)}$. Per calcolare la densità di probabilità $f_{(k)}(x)$ deriviamo la funzione di distribuzione $F_{(k)}(x)$. Si ottiene:

$$\begin{aligned} f_{(k)}(x) = \frac{d}{dx} F_{(k)}(x) &= \sum_{j=k}^n \binom{n}{j} j [F(x)]^{j-1} [1 - F(x)]^{n-j} f(x) \\ &\quad - \sum_{j=k+1}^{n-1} \binom{n}{j} (n-j) [F(x)]^j [1 - F(x)]^{n-j-1} f(x) \end{aligned}$$

ovvero, ponendo $s = j+1$ nella seconda somma a secondo membro:

$$\begin{aligned} f_{(k)}(x) &= \sum_{j=k}^n \binom{n}{j} j [F(x)]^{j-1} [1 - F(x)]^{n-j} f(x) \\ &\quad - \sum_{s=k+1}^n \binom{n}{s-1} (n-s+1) [F(x)]^{s-1} [1 - F(x)]^{n-s} f(x). \end{aligned}$$

Raccogliendo i termini si ha poi:

$$\begin{aligned} f_{(k)}(x) &= \binom{n}{k} k [F(x)]^{k-1} [1 - F(x)]^{n-k} f(x) \\ &\quad + \sum_{j=k+1}^n \left[\binom{n}{j} j - \binom{n}{j-1} (n-j+1) \right] \\ &\quad \times [F(x)]^{j-1} [1 - F(x)]^{n-j} f(x). \end{aligned} \quad (3.3)$$

Osservando che risulta

$$\begin{aligned} \binom{n}{j} j - \binom{n}{j-1} (n-j+1) &= \frac{n! j}{j!(n-j)!} - \frac{n!(n-j+1)}{(j-1)!(n-j+1)!} \\ &= \frac{n!}{(j-1)!(n-j)!} - \frac{n!}{(j-1)!(n-j)!} = 0, \end{aligned}$$

dalla (3.3) segue infine la (3.2).

Casi particolari di statistiche d'ordine sono il minimo e il massimo:

$$X_{(1)} = \min\{X_1, X_2, \dots, X_n\}, \quad X_{(n)} = \max\{X_1, X_2, \dots, X_n\}.$$

Per ogni $x \in \mathbb{R}$, le rispettive funzioni di distribuzione sono:

$$F_{(1)}(x) = 1 - [1 - F(x)]^n, \quad F_{(n)}(x) = [F(x)]^n, \quad (3.4)$$

mentre le densità di probabilità sono rispettivamente:

$$f_{(1)}(x) = n [1 - F(x)]^{n-1} f(x), \quad f_{(n)}(x) = n [F(x)]^{n-1} f(x), \quad (3.5)$$

3.1. STATISTICHE D'ORDINE

dove $F(x)$ e $f(x)$ denotano funzione di distribuzione e densità di probabilità della variabile casuale genitrice.

Introduciamo ora una grandezza, detta *mediana*, che esprime un'importante caratteristica della funzione di distribuzione di una variabile casuale.

Definizione 3.1.1 Si dice *mediana* di una variabile casuale X ogni reale m tale da aversi:

$$P(X \leq m) \geq \frac{1}{2}, \quad P(X \geq m) \geq \frac{1}{2}.$$

Quindi, denotata con $F(x)$ la funzione di distribuzione di X , la mediana m è tale che

$$\frac{1}{2} \leq F(m) \leq \frac{1}{2} + P(X = m). \quad (3.6)$$

Se X è una variabile casuale continua, essendo $P(X = x) = 0$ per ogni $x \in \mathbb{R}$, dalla (3.6) segue che la mediana soddisfa l'equazione

$$F(m) = \frac{1}{2}.$$

È allora evidente che la mediana è unica quando la funzione di distribuzione $F(x)$ è strettamente crescente.

La mediana costituisce il punto intermedio di una distribuzione nel senso che ogni variabile casuale continua assume valori inferiori o superiori alla sua mediana con uguali probabilità.

Per illustrare ulteriormente il significato di mediana, accenniamo al ruolo che essa riveste in problemi di predizione dei valori assunti da una variabile casuale. A tal fine supponiamo di essere indotti a prevedere che sia c il valore che sarà assunto da una data variabile casuale continua X . Conseguentemente, il valore assoluto dell'errore associato a tale previsione è $|X - c|$. Calcoliamone il valore medio indicando con $f(x)$ e $F(x)$ densità e funzione di distribuzione di X . Si ha:

$$\begin{aligned} E(|X - c|) &= \int_{-\infty}^{\infty} |x - c| f(x) dx \\ &= \int_{-\infty}^c (c - x) f(x) dx + \int_c^{\infty} (x - c) f(x) dx \\ &= c F(c) - \int_{-\infty}^c x f(x) dx + \int_c^{\infty} x f(x) dx - c [1 - F(c)]. \end{aligned}$$

Volendo determinare il minimo di $E(|X - c|)$ al variare di c , notiamo che risulta:

$$\begin{aligned} \frac{d}{dc} E(|X - c|) &= F(c) + c f(c) - c f(c) - c f(c) - 1 + F(c) + c f(c) \\ &= 2F(c) - 1, \end{aligned}$$

così che la derivata di $E(|X - c|)$ è nulla se e solo se risulta $F(c) = 1/2$, ossia se e solo se c coincide con la mediana di X . Osservando poi che si ha

$$\frac{d^2}{dc^2} E(|X - c|) = 2f(c) \geq 0,$$

si conclude che la media $E(|X - c|)$ dell'errore assoluto di previsione è minima quando c è uguale alla mediana m di X . La mediana è dunque il migliore preditore di una variabile casuale sotto il vincolo che sia minimo l'errore assoluto medio della predizione.

Dato un campione casuale di taglia n , una statistica definita in analogia con la mediana è la *mediana campionaria* \tilde{X} :

$$\tilde{X} = \begin{cases} X_{(k+1)} & \text{per } n = 2k + 1, \\ \frac{X_{(k)} + X_{(k+1)}}{2} & \text{per } n = 2k. \end{cases} \quad (3.7)$$

Se la taglia del campione è dispari, la mediana campionaria si identifica così con la *statistica d'ordine centrale*; se essa è invece pari, la mediana campionaria viene definita come media aritmetica delle due statistiche d'ordine centrali.

Di interesse è anche la statistica costituita dalla differenza

$$R = X_{(n)} - X_{(1)} \quad (3.8)$$

tra il massimo e il minimo delle variabili del campione casuale. Infatti, R costituisce il *campo di variazione campionario* in quanto misura l'ampiezza dell'intervallo che contiene il campione, e fornisce quindi informazioni sulla dispersione della variabile casuale genitrice.

Esempio 3.1.1 Nel corso di un'ispezione condotta presso un'azienda sono state esaminate 10 confezioni di mangime. In queste si sono rilevate le seguenti percentuali di impurità:

$$(2.23, 1.95, 2.15, 1.90, 2.24, 2.01, 1.87, 1.97, 2.13, 2.07).$$

Riguardando questi dieci numeri come costituenti la realizzazione di un campione casuale di taglia 10, le statistiche d'ordine corrispondenti assumono i seguenti valori:

$$\begin{aligned} x_{(1)} &= 1.87 & x_{(2)} &= 1.90 & x_{(3)} &= 1.95 & x_{(4)} &= 1.97 & x_{(5)} &= 2.01 \\ x_{(6)} &= 2.07 & x_{(7)} &= 2.13 & x_{(8)} &= 2.15 & x_{(9)} &= 2.23 & x_{(10)} &= 2.24. \end{aligned}$$

La mediana campionaria assume pertanto il valore

$$\frac{x_{(5)} + x_{(6)}}{2} = \frac{2.01 + 2.07}{2} = 2.04,$$

mentre per il campo di variazione campionario si ha

$$x_{(10)} - x_{(1)} = 2.24 - 1.87 = 0.37.$$

È ragionevole ritenere che il processo di produzione sia insoddisfacente, in termini di contenuto di impurità del prodotto, se è piccola la probabilità che l'impurità massima a priori registrabile sia maggiore della massima impurità presente nella realizzazione osservata del campione. Supponiamo ora che il campione casuale in esame si possa ritenere estratto da una popolazione normale di media $\mu = 1.95$ e deviazione standard $\sigma = 0.1$. Per concretezza, assumiamo poi che il processo di produzione delle confezioni non sia da ritenersi soddisfacente

3.2. STIMATORI CORRETTI

se è minore di 0.04 la probabilità che il massimo $X_{(10)}$ del campione sia superiore al valore $x_{(10)} = 2.24$ osservato, ossia assumiamo che il processo di produzione sia insoddisfacente se risulta $P(X_{(10)} > x_{(10)}) < 0.04$. Ricordando la seconda delle (3.4), notiamo che risulta:

$$P(X_{(10)} > 2.24) = 1 - F_{(10)}(2.24) = 1 - [F(2.24)]^{10},$$

dove $F(x)$ è la funzione di distribuzione della variabile casuale genitrice X che, per ipotesi, è normale di media $\mu = 1.95$ e deviazione standard $\sigma = 0.1$. Si ha quindi:

$$\begin{aligned} F(2.24) &= P(X \leq 2.24) = P\left(\frac{X - 1.95}{0.1} \leq \frac{2.24 - 1.95}{0.1}\right) \\ &= P(Z \leq 2.9), \end{aligned}$$

dove $Z = (X - 1.95)/0.1$ è la variabile casuale normale standard. Dalla Tabella 1 dell'Appendice B risulta

$$P(Z \leq 2.9) = P(Z < 0) + P(0 \leq Z \leq 2.9) = \frac{1}{2} + 0.4981 = 0.9981$$

e pertanto

$$P(X_{(10)} > 2.24) = 1 - (0.9981)^{10} = 0.0188.$$

Poiché questa probabilità è minore di 0.04, il processo di produzione è da ritenersi non soddisfacente. ◆

3.2 Stimatori corretti

Uno dei problemi centrali dell'inferenza statistica consiste nello studiare una popolazione, la cui legge di probabilità è di forma nota, ma che è caratterizzata da uno o più parametri incogniti. Va sottolineato che se questi fossero noti, la legge di probabilità sarebbe completamente specificata. Un fondamentale problema consiste dunque nell'individuazione dei valori incogniti dei parametri effettuandone la stima a partire da realizzazioni di campioni casuali. La *stima dei parametri* costituisce il procedimento con cui da una realizzazione osservata (x_1, x_2, \dots, x_n) si traggono informazioni che consentano di assegnare a ciascuno dei parametri incogniti un singolo valore (nel qual caso si parla di *stima puntuale*) o un intervallo di valori (*stima intervallare*). Tratteremo la stima puntuale nella parte rimanente del presente capitolo, mentre la stima intervallare sarà oggetto del Cap. 4.

Definizione 3.2.1 Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione di cui θ è un parametro incognito. Si dice *stimatore* di θ una statistica $\hat{\theta} = g(X_1, X_2, \dots, X_n)$ i cui valori sono usati per stimare θ . Il valore $\hat{\theta} = g(x_1, x_2, \dots, x_n)$ assunto da $\hat{\theta}$ in corrispondenza della realizzazione (x_1, x_2, \dots, x_n) osservata è detto *stima puntuale di θ* .

Talora, ricorrendo ad un linguaggio più formale, si dice che una statistica $\hat{\theta}_n = g(X_1, X_2, \dots, X_n)$ costituisce uno stimatore del parametro θ se la successione di variabili casuali $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_n, \dots$ converge a θ in accordo con qualche prestabilito criterio di convergenza.

È opportuno chiarire anzitutto che gli stimatori, essendo essi stessi variabili casuali, sono soggetti a variabilità nel senso che le stime puntuali $\hat{\theta} = g(x_1, x_2, \dots, x_n)$ possono cambiare in

maniera rilevante in corrispondenza della realizzazione (x_1, x_2, \dots, x_n) osservata. Per valutare la bontà di uno stimatore è allora necessario studiarne la distribuzione campionaria, ponendosi come obiettivo fondamentale quello di determinare uno stimatore la cui distribuzione campionaria sia quanto più possibile "concentrata" in prossimità di θ , in modo che le stime puntuale siano quanto più vicine al parametro incognito. D'altro canto, poiché non esiste uno stimatore perfetto, ossia tale da fornire sempre la stima giusta, in generale accade che si faccia riferimento a più stimatori, tutti in qualche modo idonei alla stima di un parametro incognito di una popolazione. Per pervenire alla migliore stima possibile occorre allora individuare talune proprietà degli stimatori che consentano di stabilire quando uno stimatore sia da preferirsi ad altri. Anzitutto, sembra ragionevole richiedere che uno stimatore fornisca la stima giusta almeno in media: è quindi desiderabile che il valore atteso di uno stimatore uguagli il parametro da stimare. Nasce così la seguente definizione, che esprime la proprietà più immediata da richiedersi agli stimatori:

Definizione 3.2.2 Uno stimatore $\hat{\theta} = g(X_1, X_2, \dots, X_n)$ del parametro incognito θ è detto corretto, o non distorto, se risulta:

$$E(\hat{\theta}) = E[g(X_1, X_2, \dots, X_n)] \stackrel{!}{=} \theta$$

per ogni valore di θ .

La differenza $E(\hat{\theta}) - \theta$, detta *distorsione o errore sistematico*, è dunque nulla se $\hat{\theta}$ è uno stimatore corretto di θ .

Per chiarire meglio il significato dell'operazione di media che interviene nella Definizione 3.2.2, mostriamo come anche la media, così come la mediana, rivesta un ruolo significativo nella predizione del valore che una variabile casuale assume. Supponiamo a tal fine di prevedere che c sia il valore che sarà assunto da una certa variabile casuale X dotata di media finita μ ; il quadrato dell'errore di tale previsione è quindi $(X - c)^2$, il cui valore medio soddisfa la seguente relazione:

$$\begin{aligned} E[(X - c)^2] &= E[(X - \mu + \mu - c)^2] \\ &= E[(X - \mu)^2] + 2(\mu - c)E(X - \mu) + (\mu - c)^2 \\ &= E[(X - \mu)^2] + (\mu - c)^2 \\ &\geq E[(X - \mu)^2]. \end{aligned}$$

L'errore quadratico medio $E[(X - c)^2]$ è pertanto minimo quando c è pari alla media μ di X . La media è dunque il migliore preditore di una variabile casuale quando si desideri minimizzare l'errore quadratico medio della predizione.

Da queste considerazioni e dalla Definizione 3.2.2 si conclude che se $\hat{\theta}$ è uno stimatore corretto di un parametro θ , è legittimo attendersi che sia proprio θ il valore da esso assunto in corrispondenza di una realizzazione del campione casuale.

Mostriremo ora che talune delle principali statistiche fin qui esaminate sono stimatori corretti.

Proposizione 3.2.1 La media campionaria \bar{X} , la varianza campionaria S^2 e il momento campionario $\bar{X}^{(k)}$ di ordine k di un campione casuale (X_1, X_2, \dots, X_n) sono stimatori corretti

3.2. STIMATORI CORRETTI

rispettivamente del valore medio, della varianza e del momento intorno all'origine di ordine k della popolazione.

Dim. In accordo con la Definizione 3.2.2, la media campionaria e la varianza campionaria sono stimatori corretti in virtù del Teorema 1.3.1. Per mostrare che il momento campionario $\bar{X}^{(k)}$ è uno stimatore corretto del momento intorno all'origine $\mu'_k = E(X^k)$, osserviamo che dalla Definizione 1.3.3 segue:

$$E\left(\bar{X}^{(k)}\right) = E\left(\frac{1}{n} \sum_{i=1}^n X_i^k\right) = \frac{1}{n} \sum_{i=1}^n E(X_i^k) = \frac{1}{n} \sum_{i=1}^n \mu'_k = \mu'_k.$$

La Proposizione 3.2.1 chiarisce il motivo per il quale nella definizione di varianza campionaria S^2 appare $n - 1$ a denominatore, anziché n come sarebbe naturale attendersi: invero in tal modo, come mostra la prima delle (1.10), S^2 risulta essere uno stimatore corretto della varianza della popolazione. È interessante osservare che invece la mediana campionaria non è in generale uno stimatore corretto della mediana, come mostra l'esempio che segue.

Esempio 3.2.1 Sia \tilde{X} la mediana campionaria di un campione casuale di taglia $n = 3$ estratto da una popolazione esponenziale di media θ incognita. Dalla (3.7) segue allora $\tilde{X} = X_{(2)}$. Pertanto, ricordando che densità e funzione di distribuzione di una variabile esponenziale di media θ sono rispettivamente

$$f(x) = \frac{1}{\theta} e^{-x/\theta} \quad (x > 0), \quad F(x) = 1 - e^{-x/\theta} \quad (x > 0),$$

e ricordando l'espressione della densità $f_{(2)}$ fornita dal Teorema 3.1.1, si ricava la densità di \tilde{X} :

$$\begin{aligned} f_{(2)}(x) &= \binom{3}{2} 2F(x)[1 - F(x)]f(x) \\ &= \frac{6}{\theta} e^{-2x/\theta} (1 - e^{-x/\theta}) \quad (x > 0). \end{aligned}$$

Ne segue:

$$\begin{aligned} E(\tilde{X}) &= \int_0^\infty x f_{(2)}(x) dx = \frac{6}{\theta} \int_0^\infty x e^{-2x/\theta} (1 - e^{-x/\theta}) dx \\ &= 3 \int_0^\infty x \frac{2}{\theta} e^{-2x/\theta} dx - 2 \int_0^\infty x \frac{3}{\theta} e^{-3x/\theta} dx \\ &= \frac{5}{6} \theta = 0.8333 \theta. \end{aligned}$$

D'altro canto

$$m = \theta \ln 2 = 0.6931 \theta$$

è la mediana di una variabile esponenziale di media θ , come è immediato ricavare dalla condizione

$$F(m) \equiv 1 - e^{-m/\theta} = \frac{1}{2}.$$

La mediana campionaria \tilde{X} non è dunque uno stimatore corretto della mediana m , avendosi la distorsione

$$E(\tilde{X}) - m = 0.8333\theta - 0.6931\theta = 0.1402\theta.$$

Inoltre, essendo

$$E(\tilde{X}) - \theta = -0.1667\theta,$$

\tilde{X} non è uno stimatore corretto neanche del parametro θ . \diamond

La Proposizione 3.2.1 afferma, tra l'altro, che la varianza campionaria è uno stimatore corretto della varianza σ^2 della popolazione. Nel caso particolare in cui il campione casuale è estratto da una popolazione di media μ nota, è possibile costruire un ulteriore stimatore corretto della varianza σ^2 , come indicato dalla proposizione che segue.

Proposizione 3.2.2 *Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione di media μ nota, varianza σ^2 e momento centrale del quart'ordine μ_4 . Per la statistica*

$$\hat{\Theta} = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 \quad (3.9)$$

risulta:

$$E(\hat{\Theta}) = \sigma^2, \quad D^2(\hat{\Theta}) = \frac{\mu_4 - \sigma^4}{n}. \quad (3.10)$$

Dim. Dalla (3.9) si ha:

$$\begin{aligned} E(\hat{\Theta}) &= E\left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2\right] = \frac{1}{n} \sum_{i=1}^n E[(X_i - \mu)^2] \\ &= \frac{1}{n} \sum_{i=1}^n \sigma^2 = \sigma^2, \end{aligned} \quad (3.11)$$

ossia la prima delle (3.10). Dalla (3.11) e dalla (3.9) risulta poi:

$$\begin{aligned} D^2(\hat{\Theta}) &= E(\hat{\Theta}^2) - [E(\hat{\Theta})]^2 \\ &= E\left\{\left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2\right]^2\right\} - \sigma^4 \\ &= \frac{1}{n^2} \sum_{i,j=1}^n E[(X_i - \mu)^2(X_j - \mu)^2] - \sigma^4 \\ &= \frac{1}{n^2} \sum_{i \neq j} E[(X_i - \mu)^2]E[(X_j - \mu)^2] + \frac{1}{n^2} \sum_{i=j} E[(X_i - \mu)^4] - \sigma^4 \\ &= \frac{1}{n^2} n(n-1)\sigma^4 + \frac{1}{n}\mu_4 - \sigma^4 \\ &= \frac{\mu_4 - \sigma^4}{n}. \end{aligned}$$

3.2. STIMATORI CORRETTI

Verranno ora ricavati due stimatori corretti della deviazione standard σ per una popolazione normale di media μ nota e varianza σ^2 incognita.

Proposizione 3.2.3 *Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione normale di media μ e varianza σ^2 . Se μ è nota, sono corretti gli stimatori*

$$\hat{\Theta} = \frac{1}{\sqrt{2}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} \sqrt{\sum_{i=1}^n (X_i - \mu)^2} \quad (3.12)$$

$$\hat{\Theta}' = \frac{1}{n} \sqrt{\frac{\pi}{2}} \sum_{i=1}^n |X_i - \mu| \quad (3.13)$$

per la stima della deviazione standard σ .

Dim. Per mostrare che lo stimatore (3.12) è corretto, osserviamo che $\hat{\Theta}$ può essere così riscritto:

$$\hat{\Theta} = \frac{\sigma}{\sqrt{2}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} T, \quad (3.14)$$

dove si è posto

$$T = \sqrt{\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2}.$$

La variabile casuale $Y = T^2$ è quindi somma di n variabili normali standard indipendenti così che, per il Corollario 2.1.1, essa ha distribuzione chi-quadrato con n gradi di libertà. Ricordando allora la densità (2.4), possiamo scrivere:

$$\begin{aligned} E(T) &= E(\sqrt{Y}) = \int_0^\infty \sqrt{x} \frac{x^{(n/2)-1} e^{-x/2}}{2^{n/2} \Gamma(n/2)} dx \\ &= \sqrt{2} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \int_0^\infty \frac{x^{(n-1)/2} e^{-x/2}}{2^{(n+1)/2} \Gamma((n+1)/2)} dx. \end{aligned}$$

Osserviamo che quest'ultimo integrale vale 1 in quanto la funzione integranda è la densità di una variabile casuale chi-quadrato con $n+1$ gradi di libertà; ne segue:

$$E(T) = \sqrt{2} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}. \quad (3.15)$$

Facendo uso della (3.15) nella (3.14) si ricava infine $E(\hat{\Theta}) = \sigma$, da cui si conclude che $\hat{\Theta}$ è uno stimatore corretto della deviazione standard σ . Per quanto riguarda la statistica (3.13), osserviamo anzitutto che risulta:

$$E(|X_i - \mu|) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^\infty |x - \mu| e^{-(x-\mu)^2/(2\sigma^2)} dx$$

da cui, effettuando la sostituzione $y = (x - \mu)/\sigma$, si ricava:

$$E(|X_i - \mu|) = \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |y| e^{-y^2/2} dy = \sqrt{\frac{2}{\pi}} \sigma \int_0^{\infty} y e^{-y^2/2} dy = \sqrt{\frac{2}{\pi}} \sigma. \quad (3.16)$$

Dalle (3.13) e (3.16) si trae poi:

$$E(\hat{\Theta}') = \frac{1}{n} \sqrt{\frac{\pi}{2}} \sum_{i=1}^n E(|X_i - \mu|) = \frac{1}{n} \sqrt{\frac{\pi}{2}} \sum_{i=1}^n \sqrt{\frac{2}{\pi}} \sigma = \sigma,$$

così che anche $\hat{\Theta}'$ è uno stimatore corretto di σ . ■

È opportuno rilevare che se $\hat{\Theta}$ è uno stimatore corretto di un parametro θ , se g denota una funzione arbitraria, $g(\hat{\Theta})$ non è in generale uno stimatore corretto di $g(\theta)$. Inoltre, se ad esempio $\hat{\Theta}^2$ è uno stimatore corretto di θ^2 , non è detto che $\hat{\Theta}$ sia uno stimatore corretto di θ . Una riprova di tale affermazione la si ritrova nella seguente proposizione nella quale si mostra che nel caso di popolazione normale la deviazione standard campionaria S non è uno stimatore corretto di σ anche se, come mostrato della prima delle (1.10), la varianza campionaria S^2 è uno stimatore corretto di σ^2 .

Proposizione 3.2.4 *La deviazione standard campionaria S di un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione normale di varianza σ^2 non è uno stimatore corretto della deviazione standard σ avendosi:*

$$E(S) = \sigma \sqrt{\frac{2}{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}. \quad (3.17)$$

Dim. Come mostrato nel Teorema 2.1.4, se S^2 è la varianza campionaria di un campione casuale di taglia n estratto da una popolazione normale di varianza σ^2 , la variabile casuale $Y = (n-1)S^2/\sigma^2$ ha distribuzione chi-quadrato con $n-1$ gradi di libertà. Posto

$$T = \sqrt{Y} = \frac{\sqrt{n-1}}{\sigma} S,$$

applicando un procedimento del tutto analogo a quello usato nella dimostrazione della Proposizione 3.2.3, si ottiene:

$$E(T) = \sqrt{2} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}$$

e quindi:

$$E(S) = \frac{\sigma}{\sqrt{n-1}} E(T) = \sigma \sqrt{\frac{2}{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)},$$

cosicché la deviazione standard campionaria S non è uno stimatore corretto di σ . ■

Sebbene la proprietà di correttezza sia molto significativa, essa da sola non è sufficiente a garantire la bontà di uno stimatore. Infatti se $\hat{\Theta}$ è uno stimatore corretto, può comunque accadere che i valori da esso assunti si discostino molto dal suo valore medio θ . Diventa allora necessario richiedere che la dispersione di uno stimatore attorno al parametro da stimare sia la più piccola possibile. A questo proposito conviene dare la seguente definizione:

Definizione 3.2.3 *Dato un campione casuale (X_1, X_2, \dots, X_n) , sia $\hat{\Theta} = g(X_1, X_2, \dots, X_n)$ uno stimatore del parametro incognito θ . Si definisce errore quadratico medio di $\hat{\Theta}$ la quantità*

$$\text{mse}(\hat{\Theta}) = E[(\hat{\Theta} - \theta)^2],$$

sempre che il valore medio a secondo membro esista.²

È dunque evidente che tra più stimatori dello stesso parametro conviene preferire quello il cui errore quadratico medio è minimo per ogni valore di θ . Così, se $\hat{\Theta}$ e $\hat{\Theta}'$ sono stimatori di θ , si preferisce $\hat{\Theta}$ se risulta

$$\text{mse}(\hat{\Theta}) \leq \text{mse}(\hat{\Theta}') \quad \text{per ogni } \theta. \quad (3.18)$$

Si sottolinea che la diseguaglianza (3.18) deve sussistere per ogni valore di θ . Ciò comporta che non è sempre possibile confrontare stimatori in base ai rispettivi errori quadratici medi in quanto tale diseguaglianza potrebbe non essere soddisfatta per tutti i valori di θ ma solo per alcuni di essi.

È interessante notare che l'errore quadratico medio di uno stimatore $\hat{\Theta}$ può esprimersi come somma della sua varianza e del quadrato della distorsione:

$$\text{mse}(\hat{\Theta}) = D^2(\hat{\Theta}) + [E(\hat{\Theta}) - \theta]^2. \quad (3.19)$$

Infatti, posto $\mu = E(\hat{\Theta})$ risulta:

$$\begin{aligned} \text{mse}(\hat{\Theta}) &= E[(\hat{\Theta} - \theta)^2] = E[(\hat{\Theta} - \mu + \mu - \theta)^2] \\ &= E[(\hat{\Theta} - \mu)^2] + 2(\mu - \theta)E(\hat{\Theta} - \mu) + (\mu - \theta)^2 \\ &= D^2(\hat{\Theta}) + [E(\hat{\Theta}) - \theta]^2. \end{aligned}$$

Dalla (3.19) discende che uno stimatore $\hat{\Theta}$ è corretto se e solo se il suo errore quadratico medio è minimo ed uguale alla varianza $D^2(\hat{\Theta})$ per ogni θ .

È opportuno osservare che non è detto che uno stimatore corretto di un parametro possa segnare un errore quadratico medio inferiore a quello di ogni altro stimatore non corretto dello stesso parametro, come mostrato nell'esempio che segue.

Esempio 3.2.2 Sia (X_1, X_2, \dots, X_n) un campione estratto da una popolazione di variabile casuale genitrice di Bernoulli di parametro θ incognito. La variabile $X = X_1 + X_2 + \dots + X_n$ ha allora distribuzione binomiale di parametro θ . Consideriamo i seguenti stimatori di θ :

$$\hat{\Theta} = \frac{X}{n}, \quad \hat{\Theta}' = \frac{X+1}{n+2}.$$

²La notazione "mse" è l'abbreviazione di "mean square error" che in Inglese significa proprio errore quadratico medio.

Poiché $E(X) = n\theta$, si ha $E(\hat{\Theta}) = \theta$ e $E(\hat{\Theta}') = (n\theta + 1)/(n+2)$. Pertanto $\hat{\Theta}$, a differenza di $\hat{\Theta}'$, è uno stimatore corretto di θ . Calcoliamo ora l'errore quadratico medio per entrambi gli stimatori. Ricordando che la varianza di X è $D^2(X) = n\theta(1-\theta)$, si ha:

$$\begin{aligned} \text{mse}(\hat{\Theta}) &= E[(\hat{\Theta} - \theta)^2] = \frac{E[(X - n\theta)^2]}{n^2} = \frac{\theta(1-\theta)}{n}, \\ \text{mse}(\hat{\Theta}') &= E[(\hat{\Theta}' - \theta)^2] = E\left[\left(\frac{X+1}{n+2} - \theta\right)^2\right] \\ &= \frac{E[(X - n\theta + 1 - 2\theta)^2]}{(n+2)^2} = \frac{E[(X - n\theta)^2] + (1-2\theta)^2}{(n+2)^2} \\ &= \frac{n\theta(1-\theta) + (1-2\theta)^2}{(n+2)^2}. \end{aligned}$$

Al fine di confrontare gli errori quadratici medi dei due stimatori, osserviamo che la condizione $\text{mse}(\hat{\Theta}) < \text{mse}(\hat{\Theta}')$ equivale alla seguente:

$$\frac{\theta(1-\theta)}{n} < \frac{n\theta(1-\theta) + (1-2\theta)^2}{(n+2)^2}$$

che è soddisfatta se e solo se risulta:

$$\theta^2 - \theta + \frac{n}{4(2n+1)} > 0.$$

Le radici dell'equazione

$$\theta^2 - \theta + \frac{n}{4(2n+1)} = 0$$

sono

$$\theta_1 = \frac{1}{2} \left(1 - \sqrt{1 - \frac{n}{2n+1}} \right), \quad \theta_2 = \frac{1}{2} \left(1 + \sqrt{1 - \frac{n}{2n+1}} \right),$$

così che l'errore quadratico medio dello stimatore corretto $\hat{\Theta}$ è minore di quello di $\hat{\Theta}'$ solo per $0 < \theta < \theta_1$ e per $\theta_2 < \theta < 1$. Di conseguenza, non essendo la diseguaglianza (3.18) soddisfatta per ogni $\theta \in (0, 1)$, non si può a priori ritenere che lo stimatore $\hat{\Theta}$ sia da preferire a $\hat{\Theta}'$, pur essendo il primo corretto. ◆

Dato un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione caratterizzata da un parametro θ incognito, siano T_1, T_2, \dots, T_k stimatori indipendenti e corretti di θ . Possiamo definire un ulteriore stimatore $\hat{\Theta}$ del parametro θ mediante la combinazione lineare

$$\hat{\Theta} = \sum_{i=1}^k c_i T_i, \quad (3.20)$$

dove c_1, c_2, \dots, c_k sono costanti reali tali da aversi

$$\sum_{i=1}^k c_i = 1. \quad (3.21)$$

Poiché

$$E(\hat{\Theta}) = \sum_{i=1}^k c_i E(T_i) = \sum_{i=1}^k c_i \theta = \theta,$$

lo stimatore $\hat{\Theta}$ è corretto. Ogni stimatore definito come nella (3.20) viene detto stimatore lineare corretto. Ricordando che l'errore quadratico medio di uno stimatore corretto coincide con la sua varianza, il problema della ricerca del migliore stimatore lineare corretto è ricondotto alla ricerca dello stimatore lineare corretto a varianza minima. La soluzione è fornita dal teorema che segue.

Teorema 3.2.1 Siano T_1, T_2, \dots, T_k stimatori indipendenti e corretti di un parametro incognito θ dotati di varianze finite $\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2$. Tra gli stimatori lineari corretti definiti come nella (3.20) ha varianza minima quello per il quale risulta

$$c_i = \frac{1/\sigma_i^2}{\sum_{i=1}^k 1/\sigma_i^2} \quad (i = 1, 2, \dots, k). \quad \begin{array}{l} \text{Caso Correttivo:} \\ \text{D'uno media} \\ \text{quadrat.} \\ \text{Perciò cor.} \\ \text{e var. min.} \end{array} \quad (3.22)$$

Dim. Dalla (3.20) segue la varianza di $\hat{\Theta}$:

$$D^2(\hat{\Theta}) = \sum_{i=1}^k c_i^2 D^2(T_i) = \sum_{i=1}^k c_i^2 \sigma_i^2.$$

Poiché siamo interessati a determinare i valori c_i che la rendono minima, facciamo uso della diseguaglianza³

$$\left(\sum_{i=1}^k a_i b_i \right)^2 \leq \left(\sum_{i=1}^k a_i^2 \right) \left(\sum_{i=1}^k b_i^2 \right). \quad (3.23)$$

Notiamo che nella (3.23) sussiste il segno di uguaglianza se e solo se esiste una costante λ tale che $a_i = \lambda b_i$ per $i = 1, 2, \dots, k$. Se si pone ora $a_i = c_i \sigma_i$, $b_i = 1/\sigma_i$, la (3.23) si legge:

$$\left[\sum_{i=1}^k (c_i \sigma_i)^2 \right] \left(\sum_{i=1}^k \frac{1}{\sigma_i^2} \right) \geq \left(\sum_{i=1}^k c_i \right)^2 = 1,$$

ossia:

$$\sum_{i=1}^k c_i^2 \sigma_i^2 \geq \left(\sum_{i=1}^k \frac{1}{\sigma_i^2} \right)^{-1}.$$

Questa diventa un'uguaglianza se e solo se $c_i \sigma_i = \lambda/\sigma_i$, ossia per

$$c_i = \frac{\lambda}{\sigma_i^2}. \quad (3.24)$$

Sommendo ambo i membri della (3.24) sull'indice i , ricaviamo la costante λ . In virtù della (3.21) si ha infatti:

$$1 = \sum_{i=1}^k c_i = \lambda \sum_{i=1}^k \frac{1}{\sigma_i^2}. \quad (3.25)$$

³Questa diseguaglianza segue immediatamente osservando che la forma quadratica $\sum_{i=1}^k (a_i x - b_i)^2$ è non negativa per ogni x reale, il che implica che il suo discriminante è non negativo.

Sostituendo nella (3.24) il valore di λ ottenuto dalla (3.25) segue infine la (3.22). \blacksquare

È opportuno notare che il coefficiente c_i presente nella (3.22) è inversamente proporzionale alla varianza σ_i^2 . Ciò comporta che nella combinazione lineare (3.20) hanno maggiore peso gli stimatori corretti T_i a varianza piccola.

Una conseguenza immediata del Teorema 3.2.1 la si ritrova a proposito degli stimatori lineari corretti della media di una popolazione, come qui appresso mostrato.

Corollario 3.2.1 *Dato un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione di valore medio μ e varianza σ^2 , tra gli stimatori lineari corretti di μ aventi forma*

$$\hat{\Theta} = \sum_{i=1}^n c_i X_i,$$

lo stimatore a varianza minima è la media campionaria \bar{X} .

Dim. Segue dal Teorema 3.2.1 scegliendo come stimatori T_i del parametro incognito μ ciascuna delle variabili X_i costituenti il campione. Si noti che tali stimatori sono certamente corretti avendosi $E(X_i) = \mu$. In accordo con la (3.22) poniamo allora:

$$c_i = \frac{1/\sigma^2}{\sum_{i=1}^n 1/\sigma^2} = \frac{1}{n} \quad (i = 1, 2, \dots, n).$$

Segue pertanto:

$$\hat{\Theta} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X},$$

la cui varianza, che come sappiamo è σ^2/n , è pertanto minima. \blacksquare

3.3 Stimatori a varianza minima

Allorché occorre effettuare la stima di un parametro incognito di una popolazione, e sono quindi stati presi in considerazione degli stimatori idonei, nasce il problema di individuare tra essi il "migliore". Per quanto visto nel § 3.2, è opportuno scegliere lo stimatore a errore quadratico medio minimo. Per stimatori corretti l'errore quadratico medio coincide con la varianza; appare così naturale paragonare stimatori corretti di uno stesso parametro mediante confronto delle rispettive varianze.

Definizione 3.3.1 *Se $\hat{\Theta}_1$ e $\hat{\Theta}_2$ sono stimatori corretti di un parametro, il rapporto $D^2(\hat{\Theta}_2)/D^2(\hat{\Theta}_1)$ viene chiamato efficienza di $\hat{\Theta}_1$ relativa a $\hat{\Theta}_2$. Si dice, inoltre, relativamente più efficiente dell'altro lo stimatore che, tra i due, ha varianza minima.*

In base a questa definizione si conviene che se $\hat{\Theta}_1$ e $\hat{\Theta}_2$ sono stimatori corretti di un parametro, è da preferirsi lo stimatore relativamente più efficiente. Esamineremo ora due esempi in cui si fa uso della Definizione 3.3.1 per individuare degli stimatori corretti relativamente più efficienti.

Esempio 3.3.1 Si consideri un campione casuale (X_1, X_2) estratto da una popolazione normale di media 0 e varianza σ^2 . Poiché $E(X_i) = 0$, risulta $E(X_1^2) = E(X_2^2) = \sigma^2$, così che la statistica

$$T = \frac{X_1^2 + X_2^2}{2}$$

è uno stimatore corretto del parametro σ^2 . La varianza di T si riconosce poi essere

$$\begin{aligned} D^2(T) &= \frac{1}{4} [D^2(X_1^2) + D^2(X_2^2)] = \frac{1}{2} D^2(X^2) \\ &= \frac{1}{2} \{E(X^4) - [E(X^2)]^2\} = \frac{1}{2} [3\sigma^4 - (\sigma^2)^2] = \sigma^4, \end{aligned}$$

avendo fatto uso della relazione $E(X^4) = 3[D^2(X)]^2$, valida per ogni variabile normale X a media nulla.⁴ Come mostrato nella Proposizione 3.2.1, un altro stimatore corretto di σ^2 è la varianza campionaria S^2 . Poiché il campione ha taglia 2, dal Corollario 2.1.2 segue:

$$D^2(S^2) = 2\sigma^4;$$

avendosi inoltre $D^2(S^2) = 2D^2(T)$, l'efficienza di S^2 relativa a T è data da $D^2(T)/D^2(S^2) = 0.5$. In conclusione, per il campione casuale analizzato T è relativamente più efficiente di S^2 . \diamond

Esempio 3.3.2 Consideriamo un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione normale di media μ nota e varianza σ^2 incognita. Come mostrato nella Proposizione 3.2.3, gli stimatori (3.12) e (3.13) della deviazione standard σ sono entrambi corretti. Calcoliamone la varianza. Dalla (3.12) segue:

$$E(\hat{\Theta}^2) = \frac{1}{2} \left[\frac{\Gamma(n/2)}{\Gamma(n+1/2)} \right]^2 \sum_{i=1}^n E[(X_i - \mu)^2] = \frac{n}{2} \left[\frac{\Gamma(n/2)}{\Gamma(n+1/2)} \right]^2 \sigma^2;$$

pertanto, ricordando che $E(\hat{\Theta}) = \sigma$, risulta:

$$D^2(\hat{\Theta}) = E(\hat{\Theta}^2) - [E(\hat{\Theta})]^2 = \left\{ \frac{n}{2} \left[\frac{\Gamma(n/2)}{\Gamma(n+1/2)} \right]^2 - 1 \right\} \sigma^2. \quad (3.26)$$

Dalla (3.13) si ha invece:

$$D^2(\hat{\Theta}') = \frac{1}{n^2} \frac{\pi}{2} D^2 \left(\sum_{i=1}^n |X_i - \mu| \right) = \frac{1}{n} \frac{\pi}{2} D^2(|X_i - \mu|) = \frac{\pi - 2}{2n} \sigma^2, \quad (3.27)$$

⁴Per una variabile X normale a media nulla e varianza σ^2 si ha infatti:

$$\begin{aligned} E(X^4) &\equiv \int_{-\infty}^{\infty} \frac{x^4}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} dx = 3\sigma^2 \int_{-\infty}^{\infty} \frac{x^2}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} dx \\ &= 3\sigma^2 D^2(X) \equiv 3[D^2(X)]^2. \end{aligned}$$

dove si è fatto uso della relazione

$$D^2(|X_i - \mu|) = E(|X_i - \mu|^2) - [E(|X_i - \mu|)]^2 = \sigma^2 - \frac{2}{\pi} \sigma^2.$$

Facciamo ora ricorso alle varianze (3.26) e (3.27) per determinare l'efficienza di $\hat{\Theta}$ relativa a $\hat{\Theta}'$. Per $n = 1$ i due stimatori coincidono, e quindi le loro varianze sono uguali. Per $n = 2$, ricordando l'Osservazione 2.1.1, si ha:

$$\begin{aligned} D^2(\hat{\Theta}) &= \left\{ \frac{1}{[\Gamma(3/2)]^2} - 1 \right\} \sigma^2 = \left\{ \frac{4}{[\Gamma(1/2)]^2} - 1 \right\} \sigma^2 \\ &= \left(\frac{4}{\pi} - 1 \right) \sigma^2 = 0.2732 \sigma^2, \\ D^2(\hat{\Theta}') &= \frac{\pi - 2}{4} \sigma^2 = 0.2854 \sigma^2. \end{aligned}$$

Si trae quindi:

$$\frac{D^2(\hat{\Theta}')}{D^2(\hat{\Theta})} = \frac{0.2854 \sigma^2}{0.2732 \sigma^2} = 1.0447,$$

il che, per la Definizione 3.3.1, implica che in questo caso $\hat{\Theta}$ è relativamente più efficiente di $\hat{\Theta}'$. Si potrebbe poi mostrare che lo stimatore $\hat{\Theta}$ è da preferire a $\hat{\Theta}'$ per ogni valore di n . ◆

È opportuno notare che il concetto di efficienza relativa può essere introdotto anche in situazioni in cui la taglia del campione casuale viene fatta divergere. Precisamente, si dà la seguente definizione.

Definizione 3.3.2 Considerato un campione casuale di taglia n , se $\hat{\Theta}_{1,n}$ e $\hat{\Theta}_{2,n}$ sono stimatori corretti di un parametro incognito, il limite

$$\lim_{n \rightarrow \infty} \frac{D^2(\hat{\Theta}_{2,n})}{D^2(\hat{\Theta}_{1,n})}$$

viene detto efficienza asintotica di $\hat{\Theta}_{1,n}$ relativa a $\hat{\Theta}_{2,n}$.

Per illustrare in maniera intuitiva il significato di questa definizione, menzioniamo che si può affermare che se l'efficienza asintotica di $\hat{\Theta}_{1,n}$ relativa a $\hat{\Theta}_{2,n}$ è pari a r/k , con r e k interi, gli stimatori $\hat{\Theta}_{1,kn}$ e $\hat{\Theta}_{2,mn}$ forniscono approssimativamente stesse "affidabilità" di stima, e che queste tendono a coincidere per $n \rightarrow \infty$.

Esempio 3.3.3 Si consideri un campione casuale (X_1, X_2, \dots, X_n) (con $n > 1$) estratto da una popolazione di media θ nota, varianza σ^2 e momento centrale del quart'ordine μ_4 . Nelle Proposizioni 3.2.1 e 3.2.2 si è visto che la varianza campionaria e la statistica (3.9) sono entrambe stimatori corretti della varianza σ^2 . Dalle seconde delle (1.10) e (3.10) segue:

$$D^2(S^2) - D^2(\hat{\Theta}) = \frac{1}{n} \left(\mu_4 - \frac{n-3}{n-1} \sigma^4 \right) - \frac{\mu_4 - \sigma^4}{n} = \frac{2\sigma^4}{n(n-1)}, \quad (3.28)$$

3.3. STIMATORI A VARIANZA MINIMA

da cui, in virtù della Definizione 3.3.1, si ricava che $\hat{\Theta}$ è relativamente più efficiente di S^2 . Ciò mostra che per la stima di σ^2 la conoscenza di μ è significativa per ottenere uno stimatore, $\hat{\Theta}$, che sia relativamente più efficiente della varianza campionaria. È bene inoltre notare che risulta

$$\lim_{n \rightarrow \infty} \frac{D^2(\hat{\Theta})}{D^2(S^2)} = \lim_{n \rightarrow \infty} \frac{\mu_4 - \sigma^4}{\mu_4 - \frac{n-3}{n-1} \sigma^4} = 1;$$

dalla Definizione 3.3.2 segue allora che l'efficienza asintotica di S^2 relativa a $\hat{\Theta}$ è pari ad 1. Si conclude che se la taglia n del campione è elevata, la varianza campionaria e la statistica (3.9) forniscono approssimativamente la stessa affidabilità di stima. La maggiore efficienza relativa di $\hat{\Theta}$ è dunque significativa solo quando la taglia del campione non è molto grande. Invece, come mostra la (3.28), la differenza $D^2(S^2) - D^2(\hat{\Theta})$ tende a zero rapidamente (come $1/n^2$) al crescere di n . ◆

Esempio 3.3.4 Sia (X_1, X_2, \dots, X_n) (con $n > 1$) un campione casuale estratto da una popolazione di Poisson di parametro θ . Poiché media, varianza e momento centrale del quart'ordine della variabile casuale genitrice X sono rispettivamente $E(X) = \theta$, $D^2(X) = \theta$ e $\mu_4 = \theta + 3\theta^2$, dal Teorema 1.3.1 discende che medie e varianze della media campionaria e della varianza campionaria sono rispettivamente:

$$E(\bar{X}) = \theta, \quad D^2(\bar{X}) = \frac{\theta}{n},$$

$$E(S^2) = \theta, \quad D^2(S^2) = \frac{\theta}{n} + \frac{2\theta^2}{n-1}.$$

Pertanto sia la media campionaria che la varianza campionaria sono stimatori corretti di θ . Valutiamone l'efficienza relativa: essendo

$$D^2(\bar{X}) \equiv \frac{\theta}{n} < \frac{\theta}{n} + \frac{2\theta^2}{n-1} \equiv D^2(S^2),$$

dalla Definizione 3.3.1 si ricava che \bar{X} è relativamente più efficiente di S^2 , così che nella stima di θ la media campionaria è da preferirsi alla varianza campionaria. Dalla Definizione 3.3.2 si ha poi che l'efficienza asintotica di \bar{X} relativa a S^2 è

$$\lim_{n \rightarrow \infty} \frac{D^2(S^2)}{D^2(\bar{X})} = \lim_{n \rightarrow \infty} \left[1 + 2\theta \frac{n}{n-1} \right] = 1 + 2\theta. \quad (3.29)$$

Pertanto anche quando la taglia n del campione è elevata la media campionaria fornisce maggiore affidabilità di stima rispetto alla varianza campionaria, e ciò risulta tanto più significativo quanto maggiore è θ , essendo il limite nella (3.29) funzione crescente di θ . ◆

Esempio 3.3.5 Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione di media θ e varianza σ^2 . Uno stimatore corretto di θ è il seguente:

$$\hat{\Theta} = \frac{2}{n(n+1)} \sum_{i=1}^n iX_i. \quad (3.30)$$

Infatti risulta:

$$E(\hat{\theta}) = \frac{2}{n(n+1)} \sum_{i=1}^n i E(X_i) = \frac{2\theta}{n(n+1)} \sum_{i=1}^n i = \theta.$$

Facendo poi uso dell'identità

$$\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6},$$

si ricava la varianza di $\hat{\theta}$:

$$\begin{aligned} D^2(\hat{\theta}) &= \frac{4}{n^2(n+1)^2} \sum_{i=1}^n i^2 D^2(X_i) = \frac{4\sigma^2}{n^2(n+1)^2} \sum_{i=1}^n i^2 \\ &= \frac{2\sigma^2(2n+1)}{3n(n+1)}. \end{aligned} \quad (3.31)$$

Confrontiamo lo stimatore corretto $\hat{\theta}$ con la media campionaria \bar{X} che, come visto nella Proposizione 3.2.1, è anch'essa uno stimatore corretto della media. Essendo $D^2(\bar{X}) = \sigma^2/n$, dalla Definizione 3.3.1 e dalla (3.31) si ottiene l'efficienza di \bar{X} relativa a $\hat{\theta}$:

$$\frac{D^2(\hat{\theta})}{D^2(\bar{X})} = \frac{2(2n+1)}{3(n+1)}. \quad (3.32)$$

Si noti che il rapporto (3.32) è maggiore di 1 per ogni $n > 1$, mentre è uguale a 1 nel caso banale $n = 1$ in cui i due stimatori coincidono. Ciò comporta che \bar{X} è relativamente più efficiente di $\hat{\theta}$. Dalla (3.32) segue infine l'efficienza asintotica di \bar{X} relativa a $\hat{\theta}$:

$$\lim_{n \rightarrow \infty} \frac{D^2(\hat{\theta})}{D^2(\bar{X})} = \frac{4}{3}.$$

Così, la media campionaria in corrispondenza di un campione casuale di taglia 300 fornisce approssimativamente la stessa affidabilità di stima che fornisce $\hat{\theta}$ in corrispondenza di un campione di taglia 400. ♦

Osserviamo che, per come è definito, lo stimatore (3.30) attribuisce pesi crescenti con il numero d'ordine delle variabili costituenti il campione. Ciò ne rende ingiustificato l'uso in quanto tali variabili sono indipendenti e identicamente distribuite, così che esse dovrebbero intervenire tutte in pari misura nella stima della media θ . È questa una indiretta giustificazione della maggiore efficienza della media campionaria rispetto allo stimatore (3.30).

Si è visto che tra più stimatori corretti di uno stesso parametro si preferisce quello la cui distribuzione campionaria ha varianza minima. Questo criterio suggerisce di introdurre la seguente definizione:

Definizione 3.3.3 Uno stimatore corretto $\hat{\theta}$ di un parametro θ si dice *efficiente*, o a varianza uniformemente minima, se per ogni altro stimatore corretto $\hat{\theta}'$ di θ risulta $D^2(\hat{\theta}) \leq D^2(\hat{\theta}')$ per ogni valore di θ .

3.3. STIMATORI A VARIANZA MINIMA

È bene osservare esplicitamente che per stimatori corretti $\hat{\theta}$ e $\hat{\theta}'$ può accadere che la diseguaglianza $D^2(\hat{\theta}) \leq D^2(\hat{\theta}')$ sussista soltanto per taluni valori del parametro θ da stimare, e non per tutti. In tal caso non si può affermare che $\hat{\theta}$ ha varianza *uniformemente* minima.

Nel teorema che segue viene introdotta una diseguaglianza, detta di Cramér-Rao, che fornisce un limite inferiore alla varianza di stimatori corretti, e che quindi è di notevole ausilio nella ricerca di stimatori corretti efficienti.

Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione continua di variabile casuale genitrice X caratterizzata da un parametro incognito θ . Qui denoteremo con $f(x; \theta)$ la densità di probabilità di X e con

$$f(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta) \quad (3.33)$$

la densità di probabilità del campione.

Teorema 3.3.1 (Cramér-Rao) Nelle ipotesi che

- (i) $\frac{\partial}{\partial \theta} \ln f(x; \theta)$ esiste per ogni x e per ogni θ
- (ii) $\int_{\mathbb{R}^n} g(x_1, x_2, \dots, x_n) \frac{\partial}{\partial \theta} f(x_1, x_2, \dots, x_n; \theta) dx_1 dx_2 \dots dx_n$
 $= \frac{\partial}{\partial \theta} \int_{\mathbb{R}^n} g(x_1, x_2, \dots, x_n) f(x_1, x_2, \dots, x_n; \theta) dx_1 dx_2 \dots dx_n$
- (iii) $\int_{\mathbb{R}} \frac{\partial}{\partial \theta} f(x; \theta) dx = \frac{\partial}{\partial \theta} \int_{\mathbb{R}} f(x; \theta) dx$
- (iv) $E\left\{\left[\frac{\partial}{\partial \theta} \ln f(X; \theta)\right]^2\right\}$ esiste finito per ogni θ ,

per ogni stimatore corretto $\hat{\theta} = g(X_1, X_2, \dots, X_n)$ del parametro θ che sia dotato di varianza finita, risulta:

$$D^2(\hat{\theta}) \geq \frac{1}{n E\left\{\left[\frac{\partial}{\partial \theta} \ln f(X; \theta)\right]^2\right\}}, \quad (3.34)$$

sussistendo l'uguaglianza se e solo se esiste una funzione $\alpha(\theta)$ tale da aversi per ogni θ :

$$g(X_1, X_2, \dots, X_n) - \theta = \alpha(\theta) \sum_{j=1}^n \frac{\partial}{\partial \theta} \ln f(X_j; \theta). \quad (3.35)$$

Dim. Mediante applicazione della diseguaglianza di Schwarz⁵ alla coppia di variabili casuali $\hat{\theta} - \theta$ e $\partial \ln f(X_1, X_2, \dots, X_n; \theta) / \partial \theta$ si ottiene:

$$\left\{E\left[(\hat{\theta} - \theta) \frac{\partial}{\partial \theta} \ln f(X_1, X_2, \dots, X_n; \theta)\right]\right\}^2$$

⁵Ricordiamo che se le variabili casuali X e Y possiedono momenti del secondo ordine finiti, la diseguaglianza di Schwarz si scrive $[E(XY)]^2 \leq E(X^2)E(Y^2)$, dove l'uguaglianza sussiste se e solo se esiste un $\lambda \in \mathbb{R}$ tale che $P(Y = \lambda X) = 1$.

$$\leq E[(\hat{\theta} - \theta)^2] E\left\{ \left[\frac{\partial}{\partial \theta} \ln f(X_1, X_2, \dots, X_n; \theta) \right]^2 \right\}. \quad (3.36)$$

Dimostriamo innanzitutto che il primo membro della (3.36) è pari ad 1. Osserviamo a tal fine che dalla (3.33) segue l'identità

$$\frac{\partial}{\partial \theta} \ln f(X_1, X_2, \dots, X_n; \theta) = \sum_{j=1}^n \frac{\partial}{\partial \theta} \ln f(X_j; \theta), \quad (3.37)$$

così che:

$$\begin{aligned} & E\left[(\hat{\theta} - \theta) \frac{\partial}{\partial \theta} \ln f(X_1, X_2, \dots, X_n; \theta) \right] \\ &= E\left[\hat{\theta} \frac{\partial}{\partial \theta} \ln f(X_1, X_2, \dots, X_n; \theta) \right] - \theta \sum_{j=1}^n E\left[\frac{\partial}{\partial \theta} \ln f(X_j; \theta) \right]. \end{aligned} \quad (3.38)$$

Facendo uso dell'ipotesi (ii) si ha poi:

$$\begin{aligned} & E\left[\hat{\theta} \frac{\partial}{\partial \theta} \ln f(X_1, X_2, \dots, X_n; \theta) \right] \\ &= \int_{\mathbf{R}^n} g(x_1, x_2, \dots, x_n) \frac{\partial}{\partial \theta} \ln f(x_1, x_2, \dots, x_n; \theta) \\ &\quad \times f(x_1, x_2, \dots, x_n; \theta) dx_1 dx_2 \dots dx_n \\ &= \int_{\mathbf{R}^n} g(x_1, x_2, \dots, x_n) \frac{\partial}{\partial \theta} f(x_1, x_2, \dots, x_n; \theta) dx_1 dx_2 \dots dx_n \\ &= \frac{\partial}{\partial \theta} \int_{\mathbf{R}^n} g(x_1, x_2, \dots, x_n) f(x_1, x_2, \dots, x_n; \theta) dx_1 dx_2 \dots dx_n \\ &= \frac{\partial}{\partial \theta} E(\hat{\theta}) = \frac{\partial}{\partial \theta} \theta = 1, \end{aligned}$$

essendo, per ipotesi, $\hat{\theta} = g(X_1, X_2, \dots, X_n)$ uno stimatore corretto di θ . Inoltre, per l'ipotesi (iii), risulta:

$$\begin{aligned} E\left[\frac{\partial}{\partial \theta} \ln f(X; \theta) \right] &= \int_{\mathbf{R}} f(x; \theta) \frac{\partial}{\partial \theta} \ln f(x; \theta) dx = \int_{\mathbf{R}} \frac{\partial}{\partial \theta} f(x; \theta) dx \\ &= \frac{\partial}{\partial \theta} \int_{\mathbf{R}} f(x; \theta) dx = \frac{\partial}{\partial \theta} 1 = 0. \end{aligned} \quad (3.39)$$

La (3.38) diventa allora:

$$E\left[(\hat{\theta} - \theta) \frac{\partial}{\partial \theta} \ln f(X_1, X_2, \dots, X_n; \theta) \right] = 1, \quad (3.40)$$

restando così dimostrato che è unitario il primo membro della (3.36). Sostituendo la (3.40) nella (3.36), e ricordando che $D^2(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$ per la supposta correttezza dello stimatore

$\hat{\theta}$, si trae inoltre:

$$D^2(\hat{\theta}) \geq \frac{1}{E\left\{ \left[\frac{\partial}{\partial \theta} \ln f(X_1, X_2, \dots, X_n; \theta) \right]^2 \right\}}. \quad (3.41)$$

Si consideri il valore medio che compare a secondo membro della (3.41). Facendo uso della (3.37) si ha:

$$\begin{aligned} E\left\{ \left[\frac{\partial}{\partial \theta} \ln f(X_1, X_2, \dots, X_n; \theta) \right]^2 \right\} &= E\left\{ \left[\sum_{j=1}^n \frac{\partial}{\partial \theta} \ln f(X_j; \theta) \right]^2 \right\} \\ &= E\left[\sum_{j,k=1}^n \frac{\partial}{\partial \theta} \ln f(X_j; \theta) \frac{\partial}{\partial \theta} \ln f(X_k; \theta) \right]. \end{aligned} \quad (3.42)$$

Per $j \neq k$, in virtù dell'indipendenza di X_j e X_k , e facendo uso della (3.39), si ottiene:

$$\begin{aligned} & E\left[\frac{\partial}{\partial \theta} \ln f(X_j; \theta) \frac{\partial}{\partial \theta} \ln f(X_k; \theta) \right] \\ &= E\left[\frac{\partial}{\partial \theta} \ln f(X_j; \theta) \right] E\left[\frac{\partial}{\partial \theta} \ln f(X_k; \theta) \right] = 0. \end{aligned}$$

La (3.42) pertanto diventa:

$$\begin{aligned} E\left\{ \left[\frac{\partial}{\partial \theta} \ln f(X_1, X_2, \dots, X_n; \theta) \right]^2 \right\} &= \sum_{r=1}^n E\left\{ \left[\frac{\partial}{\partial \theta} \ln f(X_r; \theta) \right]^2 \right\} \\ &= n E\left\{ \left[\frac{\partial}{\partial \theta} \ln f(X; \theta) \right]^2 \right\}. \end{aligned} \quad (3.43)$$

Sostituendo la (3.43) nella (3.41) si perviene alla (3.34). In essa, per la diseguaglianza di Schwarz, sussiste il segno di uguaglianza se e solo se per ogni valore di θ le variabili casuali $\hat{\theta} - \theta$ e $\partial \ln f(X_1, X_2, \dots, X_n; \theta) / \partial \theta$ con probabilità unitaria sono tra loro proporzionali, ossia se e solo se

$$\hat{\theta} - \theta = \alpha(\theta) \frac{\partial}{\partial \theta} \ln f(X_1, X_2, \dots, X_n; \theta). \quad (3.44)$$

Per la (3.37), dalla (3.44) segue infine la condizione (3.35).

Il Teorema 3.3.1 solo per semplicità è stato enunciato e dimostrato per popolazioni continue. Esso è tuttavia valido, sotto ipotesi equivalenti, anche per popolazioni discrete; in tal caso la funzione $f(x; \theta)$ va interpretata come distribuzione di probabilità della variabile casuale genitrice.

D'ora innanzi nel fare riferimento al Teorema 3.3.1 ed alle sue implicazioni assumeremo tacitamente che siano sempre soddisfatte le ipotesi che ne assicurano la validità.

Vale la pena sottolineare che la diseguaglianza (3.34) fornisce un limite inferiore alla varianza degli stimatori corretti, ma non afferma che esiste uno stimatore corretto di varianza uguale a tale limite inferiore.



Definizione 3.3.4 Sia $\hat{\theta} = g(X_1, X_2, \dots, X_n)$ uno stimatore corretto di un parametro θ . Si definisce efficienza di $\hat{\theta}$ il numero reale

$$E_{\hat{\theta}} = \frac{1}{n D^2(\hat{\theta}) E\left\{\left[\frac{\partial}{\partial \theta} \ln f(X; \theta)\right]^2\right\}},$$

sempre che esistano finite e non nulle le quantità che compaiono al denominatore. Se $E_{\hat{\theta}} = 1$, $\hat{\theta}$ si dice pienamente efficiente.

Si noti che è $0 < E_{\hat{\theta}} \leq 1$ in virtù della disegualanza di Cramér-Rao (3.34). Inoltre, dal Teorema 3.3.1 discende che la (3.35) costituisce condizione necessaria e sufficiente perché lo stimatore $\hat{\theta}$ sia pienamente efficiente. Dalle Definizioni 3.3.3 e 3.3.4 segue poi che uno stimatore corretto e pienamente efficiente è efficiente, mentre non sussiste necessariamente l'implicazione inversa; uno stimatore corretto efficiente può non essere pienamente efficiente.

Negli esempi che seguono si fa riferimento alla media campionaria \bar{X} che, come mostrato nella Proposizione 3.2.1, è uno stimatore corretto del valore medio di una popolazione. Come si vedrà, per campioni casuali tratti da popolazioni normali, esponenziali e di Poisson lo stimatore \bar{X} risulta essere anche pienamente efficiente.

È infine appena il caso di sottolineare come la piena efficienza di uno stimatore sia una caratteristica altamente desiderabile risultando in tal caso minima la sua varianza proprio in virtù della disegualanza di Cramér-Rao.

Esempio 3.3.6 Un campione casuale sia estratto da una popolazione normale di media μ e varianza σ^2 nota; la densità di probabilità della variabile casuale genitrice è dunque:

$$f(x; \mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad (x \in \mathbb{R}),$$

così che

$$\ln f(x; \mu) = -\ln(\sqrt{2\pi}\sigma) - \frac{(x-\mu)^2}{2\sigma^2}$$

$$\frac{\partial}{\partial \mu} \ln f(x; \mu) = \frac{x-\mu}{\sigma^2},$$

e quindi:

$$E\left\{\left[\frac{\partial}{\partial \mu} \ln f(X; \mu)\right]^2\right\} = E\left[\left(\frac{X-\mu}{\sigma^2}\right)^2\right] = \frac{D^2(X)}{\sigma^4} = \frac{1}{\sigma^2}.$$

Di qui si trae infine:

$$E_{\bar{X}} = \frac{1}{n D^2(\bar{X}) E\left\{\left[\frac{\partial}{\partial \mu} \ln f(X; \mu)\right]^2\right\}} = \frac{1}{n \frac{\sigma^2}{n} \frac{1}{\sigma^2}} = 1.$$

Per la Definizione 3.3.4 la media campionaria \bar{X} è dunque uno stimatore pienamente efficiente della media μ .

Esempio 3.3.7 Esaminiamo il caso di un campione estratto da una popolazione esponenziale di valore medio θ . La densità di probabilità della variabile casuale genitrice è dunque:

$$f(x; \theta) = \begin{cases} 0 & \text{per } x \leq 0, \\ \frac{1}{\theta} e^{-x/\theta} & \text{per } x > 0. \end{cases}$$

Dalla relazione

$$\frac{\partial}{\partial \theta} \ln f(x; \theta) = \frac{x-\theta}{\theta^2} \quad (x > 0),$$

ed essendo $D^2(X) = \theta^2$, segue:

$$E\left\{\left[\frac{\partial}{\partial \theta} \ln f(X; \theta)\right]^2\right\} = E\left[\frac{(X-\theta)^2}{\theta^4}\right] = \frac{D^2(X)}{\theta^4} = \frac{1}{\theta^2}.$$

Si ottiene così:

$$E_{\bar{X}} = \frac{1}{n D^2(\bar{X}) E\left\{\left[\frac{\partial}{\partial \theta} \ln f(X; \theta)\right]^2\right\}} = \frac{1}{n \frac{\theta^2}{n} \frac{1}{\theta^2}} = 1,$$

che mostra che \bar{X} è uno stimatore pienamente efficiente di θ .

Esempio 3.3.8 Riferiamoci ad un campione casuale estratto da una popolazione di variabile genitrice X poissoniana di parametro λ , e quindi di distribuzione di probabilità

$$f(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \quad (x = 0, 1, \dots),$$

con

$$\lambda = E(X) = D^2(X). \quad (3.45)$$

Pertanto

$$E\left\{\left[\frac{\partial}{\partial \lambda} \ln f(X; \lambda)\right]^2\right\} = E\left[\frac{(X-\lambda)^2}{\lambda^2}\right] = \frac{D^2(X)}{\lambda^2} = \frac{1}{\lambda},$$

dove si è fatto uso della (3.45). Si ottiene quindi:

$$E_{\bar{X}} = \frac{1}{n D^2(\bar{X}) E\left\{\left[\frac{\partial}{\partial \lambda} \ln f(X; \lambda)\right]^2\right\}} = \frac{1}{n \frac{\lambda}{n} \frac{1}{\lambda}} = 1,$$

che mostra come la media campionaria \bar{X} sia uno stimatore pienamente efficiente di λ .

Gli esempi considerati non inducono a ritenere che la proprietà di piena efficienza sia di validità generale. Vi sono infatti casi in cui la media campionaria non è uno stimatore pienamente efficiente della media, pur essendone sempre uno stimatore corretto.

Nel teorema che segue è mostrato che gli stimatori pienamente efficienti sono caratterizzati da una particolare struttura.

Teorema 3.3.2 Se $\hat{\Theta} = g(X_1, X_2, \dots, X_n)$ è uno stimatore corretto pienamente efficiente del parametro θ , si ha:

$$\hat{\Theta} = \frac{1}{n} \sum_{j=1}^n h(X_j),$$

dove $h(x) = g(x, x, \dots, x)$.

Dim. Poiché, per ipotesi, $\hat{\Theta}$ è uno stimatore pienamente efficiente, dal Teorema 3.3.1 per ogni θ e per qualche funzione $\alpha(\theta)$ di θ segue

$$g(x_1, x_2, \dots, x_n) - \theta = \alpha(\theta) \sum_{j=1}^n \frac{\partial}{\partial \theta} \ln f(x_j; \theta). \quad (3.46)$$

La (3.46) sussiste per ogni n -upla (x_1, x_2, \dots, x_n) e quindi, in particolare, per $x_1 = x_2 = \dots = x_n = x$. Posto allora $h(x) = g(x, x, \dots, x)$, dalla (3.46) si ottiene:

$$h(x) - \theta = n \alpha(\theta) \frac{\partial}{\partial \theta} \ln f(x; \theta). \quad (3.47)$$

Ponendo $x = x_j$ ($j = 1, 2, \dots, n$) nella (3.47) e sommando le n equazioni risultanti, si ricava:

$$\sum_{j=1}^n h(x_j) - n\theta = n \alpha(\theta) \sum_{j=1}^n \frac{\partial}{\partial \theta} \ln f(x_j; \theta). \quad (3.48)$$

Facendo uso della (3.46), dalla (3.48) si ottiene poi:

$$\sum_{j=1}^n h(x_j) - n\theta = n [g(x_1, x_2, \dots, x_n) - \theta],$$

da cui la tesi. ■

Allorché ci si confronta con problemi che coinvolgono l'osservazione di più campioni casuali, sorge talora l'esigenza di paragonare o di raggruppare risultati forniti dalle diverse realizzazioni osservate. È opportuno allora individuare condizioni che consentano di non perdere talune significative proprietà degli stimatori relativi a campioni casuali differenti. Il seguente teorema fornisce un risultato di questo tipo nel caso di due campioni casuali.

Teorema 3.3.3 Siano $(X_{11}, X_{12}, \dots, X_{1n})$ e $(X_{21}, X_{22}, \dots, X_{2m})$ campioni casuali estratti da due popolazioni indipendenti caratterizzate da un parametro comune θ . Dati due stimatori

$$\hat{\Theta}_1 = g_1(X_{11}, X_{12}, \dots, X_{1n}), \quad \hat{\Theta}_2 = g_2(X_{21}, X_{22}, \dots, X_{2m})$$

di θ , posto

$$\hat{\Theta}^{(\alpha)} = \alpha \hat{\Theta}_1 + (1 - \alpha) \hat{\Theta}_2 \quad (0 \leq \alpha \leq 1) \quad (3.49)$$

sussiste quanto segue:

(i) se $\hat{\Theta}_1$ e $\hat{\Theta}_2$ sono stimatori corretti di θ , tale risulta $\hat{\Theta}^{(\alpha)}$;

(ii) se $\hat{\Theta}_1$ e $\hat{\Theta}_2$ sono dotati di varianza, la varianza di $\hat{\Theta}^{(\alpha)}$ è minima per $\alpha = \alpha_0$, dove

$$\alpha_0 = \frac{D^2(\hat{\Theta}_2)}{D^2(\hat{\Theta}_1) + D^2(\hat{\Theta}_2)}.$$

Dim. Poiché dalla (3.49) risulta che $\hat{\Theta}^{(\alpha)}$ è una combinazione lineare degli stimatori corretti $\hat{\Theta}_1$ e $\hat{\Theta}_2$, il punto (i) del teorema è di immediata dimostrazione avendosi:

$$E(\hat{\Theta}^{(\alpha)}) = \alpha E(\hat{\Theta}_1) + (1 - \alpha) E(\hat{\Theta}_2) = \alpha \theta + (1 - \alpha) \theta = \theta.$$

Per dimostrare il punto (ii) notiamo che dalla definizione (3.49) la varianza di $\hat{\Theta}^{(\alpha)}$ è data da

$$D^2(\hat{\Theta}^{(\alpha)}) = \alpha^2 D^2(\hat{\Theta}_1) + (1 - \alpha)^2 D^2(\hat{\Theta}_2), \quad (3.50)$$

la cui derivata

$$\frac{d}{d\alpha} D^2(\hat{\Theta}^{(\alpha)}) = 2 [\alpha D^2(\hat{\Theta}_1) - (1 - \alpha) D^2(\hat{\Theta}_2)]$$

si annulla per

$$\alpha = \frac{D^2(\hat{\Theta}_2)}{D^2(\hat{\Theta}_1) + D^2(\hat{\Theta}_2)} \equiv \alpha_0.$$

Dalla forma della funzione (3.50) si conclude che questo è un punto di minimo assoluto per $D^2(\hat{\Theta}^{(\alpha)})$, così che risulta $D^2(\hat{\Theta}^{(\alpha_0)}) \leq D^2(\hat{\Theta}^{(\alpha)})$ per ogni $\alpha \in [0, 1]$ e per ogni θ . ■

Esaminiamo due esempi concernenti campioni estratti da popolazioni normali in cui si fa uso del Teorema 3.3.3.

Esempio 3.3.9 Siano \bar{X}_1 la media campionaria di un campione casuale di taglia n estratto da una popolazione normale di media μ e varianza σ_1^2 nota e \bar{X}_2 la media campionaria di un campione casuale anch'esso di taglia n estratto però da una differente popolazione normale, caratterizzata da media μ e da varianza σ_2^2 nota. Essendo \bar{X}_1 e \bar{X}_2 stimatori corretti del parametro μ , dalla (i) del Teorema 3.3.3 segue che

$$X^{(\alpha)} = \alpha \bar{X}_1 + (1 - \alpha) \bar{X}_2 \quad (0 \leq \alpha \leq 1)$$

è anch'esso uno stimatore corretto di μ . Inoltre, poiché $D^2(\bar{X}_1) = \sigma_1^2/n$ e $D^2(\bar{X}_2) = \sigma_2^2/n$, risulta:

$$\frac{D^2(\bar{X}_2)}{D^2(\bar{X}_1) + D^2(\bar{X}_2)} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

In virtù della (ii) del Teorema 3.3.3 la varianza dello stimatore $X^{(\alpha)}$ è minima se si sceglie

$$\alpha = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

Esempio 3.3.10 Sia S_1^2 la varianza campionaria di un campione casuale di taglia n estratto da una popolazione normale di media μ_1 nota e varianza σ^2 incognita. Sia poi S_2^2 la varianza campionaria di un campione casuale di taglia m estratto da una differente popolazione normale avente media μ_2 nota e varianza σ^2 incognita. Come affermato dalla Proposizione 3.2.1, S_1^2 e S_2^2 sono stimatori corretti di σ^2 . In virtù della (i) del Teorema 3.3.3 anche

$$S_{(\alpha)}^2 = \alpha S_1^2 + (1 - \alpha) S_2^2 \quad (0 \leq \alpha \leq 1) \quad (3.51)$$

è uno stimatore corretto di σ^2 . Ricordiamo poi che, per il Corollario 2.1.2, le varianze di S_1^2 e di S_2^2 sono

$$D^2(S_1^2) = \frac{2\sigma^4}{n-1}, \quad D^2(S_2^2) = \frac{2\sigma^4}{m-1},$$

così che

$$\frac{D^2(S_2^2)}{D^2(S_1^2) + D^2(S_2^2)} = \frac{n-1}{n+m-2}.$$

Dalla (ii) del Teorema 3.3.3 segue che la scelta

$$\alpha = \frac{n-1}{n+m-2}$$

rende minima la varianza dello stimatore $S_{(\alpha)}^2$ definito dalla (3.51). \diamond

Concludiamo questo paragrafo introducendo un'importante nozione che consente un'ulteriore caratterizzazione degli stimatori.

Definizione 3.3.5 Sia $\hat{\theta}$ uno stimatore di un parametro incognito θ . Fissata una costante positiva c , la probabilità

$$P(|\hat{\theta} - \theta| < c)$$

viene detta concentrazione di $\hat{\theta}$ intorno a θ .

Dalla Definizione 3.3.5 appare evidente che uno stimatore $\hat{\theta}$ è tanto migliore quanto maggiore è la sua concentrazione intorno a θ . Si noti che la diseguaglianza di Chebyshev⁶ fornisce un limite inferiore alla concentrazione di stimatori corretti dotati di varianza finita. Infatti, se $E(\hat{\theta}) = \theta$ e $D^2(\hat{\theta}) = \sigma^2 < \infty$, si ha:

$$P(|\hat{\theta} - \theta| < c) \geq 1 - \frac{\sigma^2}{c^2}.$$

Da quest'ultima relazione segue quindi che uno stimatore corretto a varianza molto piccola ha concentrazione molto elevata.

⁶La diseguaglianza di Chebyshev afferma che se X è una variabile casuale dotata di media μ e di varianza σ^2 finite, risulta $P(|X - \mu| < \epsilon) \geq 1 - \sigma^2/\epsilon^2$ per ogni arbitrariamente fissato $\epsilon > 0$.

È bene sottolineare che la concentrazione di uno stimatore in generale dipende dal parametro θ incognito. Ad esempio, nell'ambito della stima della varianza σ^2 di una popolazione normale la concentrazione dello stimatore corretto S^2 intorno a σ^2 è data da

$$\begin{aligned} P(|S^2 - \sigma^2| < c) &= P\left[\left|\frac{(n-1)S^2}{\sigma^2} - (n-1)\right| < \frac{(n-1)c}{\sigma^2}\right] \\ &= P\left[|Y - (n-1)| < \frac{(n-1)c}{\sigma^2}\right], \end{aligned} \quad (3.52)$$

dove $Y = (n-1)S^2/\sigma^2$, in virtù del Teorema 2.1.4, è una variabile casuale a distribuzione chi-quadrato con $n-1$ gradi di libertà. Dalla (3.52) è evidente che la concentrazione di S^2 dipende dal parametro σ^2 , essendo in particolare decrescente con esso. Non mancano comunque casi in cui la concentrazione non dipende dal parametro incognito. Se ad esempio si considera il problema della stima della media μ di una popolazione normale a varianza σ^2 nota, la concentrazione della media campionaria \bar{X} intorno a μ è

$$P(|\bar{X} - \mu| < c) = P\left(\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| < \frac{c}{\sigma/\sqrt{n}}\right) = P\left(|Z| < \frac{c}{\sigma/\sqrt{n}}\right), \quad (3.53)$$

con $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ variabile casuale normale standard. La (3.53) mostra che la concentrazione di \bar{X} intorno a μ non dipende da μ , ma dipende esclusivamente da c , da σ e da n .

Come mostrato dalla definizione che segue, il concetto di concentrazione consente di introdurre un ulteriore criterio per il confronto di stimatori.

Definizione 3.3.6 Siano $\hat{\theta}_1$ e $\hat{\theta}_2$ stimatori di un parametro incognito θ . Se per ogni $c > 0$ e per ogni valore di θ risulta

$$P(|\hat{\theta}_1 - \theta| < c) \geq P(|\hat{\theta}_2 - \theta| < c), \quad (3.54)$$

$\hat{\theta}_1$ è detto più concentrato di $\hat{\theta}_2$ intorno a θ .

Dalla Definizione 3.3.6 segue immediatamente che se uno stimatore $\hat{\theta}_1$ è più concentrato di $\hat{\theta}_2$ intorno a un parametro θ , allora $\hat{\theta}_1$ è da preferire a $\hat{\theta}_2$ per la stima di θ . Un inconveniente che sorge nell'uso di tale criterio di confronto è però costituito dalla circostanza che la diseguaglianza (3.54) deve essere soddisfatta per ogni valore di θ , il che accade soltanto in rari casi, uno dei quali è indicato nella proposizione che segue.

Proposizione 3.3.1 Siano $\hat{\theta}_1$ e $\hat{\theta}_2$ stimatori corretti di un parametro incognito θ dotati rispettivamente di varianze σ_1^2 e σ_2^2 . Se le variabili standardizzate

$$\frac{\hat{\theta}_1 - \theta}{\sigma_1}, \quad \frac{\hat{\theta}_2 - \theta}{\sigma_2}$$

hanno la medesima distribuzione, allora $\hat{\theta}_1$ è più concentrato di $\hat{\theta}_2$ intorno a θ se e solo se risulta $\sigma_1^2 \leq \sigma_2^2$.

Dim. Nelle ipotesi fatte, la diseguaglianza (3.54) è soddisfatta se e solo se

$$P\left(\frac{|\hat{\theta}_1 - \theta|}{\sigma_1} < \frac{c}{\sigma_1}\right) \geq P\left(\frac{|\hat{\theta}_2 - \theta|}{\sigma_2} < \frac{c}{\sigma_2}\right),$$

ossia se e solo se

$$P\left(-\frac{c}{\sigma_1} < \frac{\hat{\theta}_1 - \theta}{\sigma_1} < \frac{c}{\sigma_1}\right) \geq P\left(-\frac{c}{\sigma_2} < \frac{\hat{\theta}_2 - \theta}{\sigma_2} < \frac{c}{\sigma_2}\right). \quad (3.55)$$

Poiché, per ipotesi, le variabili $(\hat{\theta}_1 - \theta)/\sigma_1$ e $(\hat{\theta}_2 - \theta)/\sigma_2$ sono identicamente distribuite, la diseguaglianza (3.55) è soddisfatta per ogni $c > 0$ e per ogni θ se e solo se è $\sigma_1^2 \leq \sigma_2^2$. ■

L'esempio che segue mostra l'utilizzazione della Proposizione 3.3.1 nel caso di campioni casuali normali.

Esempio 3.3.11 Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione normale di valore medio θ e varianza σ^2 . Consideriamo due diversi stimatori corretti della media θ , il primo dato dalla media campionaria \bar{X} ed il secondo dallo stimatore lineare

$$\hat{\theta} = \sum_{i=1}^n c_i X_i,$$

dove c_1, c_2, \dots, c_n sono costanti reali tali da aversi

$$\sum_{i=1}^n c_i = 1.$$

Abbiamo già visto nel Corollario 3.2.1 che \bar{X} ha varianza inferiore a quella di $\hat{\theta}$. Confrontiamo tali stimatori anche sotto il profilo delle rispettive concentrazioni. Ricordando il Teorema 1.4.1 e la Proposizione 2.4.1, notiamo che sia \bar{X} che $\hat{\theta}$ hanno distribuzione normale di media θ ; il primo stimatore ha però varianza σ^2/n , mentre per il secondo risulta

$$D^2(\hat{\theta}) = \sum_{i=1}^n c_i^2 D^2(X_i) = \sigma^2 \sum_{i=1}^n c_i^2.$$

Poiché le variabili $(\bar{X} - \theta)/D(\bar{X})$ e $(\hat{\theta} - \theta)/D(\hat{\theta})$ hanno entrambe distribuzione normale standard, per la Proposizione 3.3.1 \bar{X} è più concentrato di $\hat{\theta}$ intorno a θ se e solo se è $D^2(\bar{X}) \leq D^2(\hat{\theta})$. Osservando poi che se si pone $k = n$ nella (3.23) e si sceglie $a_i = c_i$ e $b_i = 1/n$ per $i = 1, 2, \dots, n$ risulta

$$D^2(\bar{X}) = \frac{\sigma^2}{n} \leq \sigma^2 \sum_{i=1}^n c_i^2 = D^2(\hat{\theta}),$$

si ricava che la media campionaria \bar{X} è sempre più concentrata di $\hat{\theta}$ intorno alla media θ della popolazione normale considerata. ◆

3.4 Proprietà asintotiche degli stimatori

In questo paragrafo si esamineranno alcune proprietà degli stimatori in relazione al divergere della taglia n del campione casuale cui essi si riferiscono. A tal fine si consideri una popolazione da cui si estrae un campione casuale (X_1, X_2, \dots, X_n) e sia $\hat{\theta}_n = g_n(X_1, X_2, \dots, X_n)$ uno stimatore di un parametro θ incognito. Al variare di n resta allora definita la successione

$$\hat{\theta}_1 = g_1(X_1), \quad \hat{\theta}_2 = g_2(X_1, X_2), \dots, \hat{\theta}_n = g_n(X_1, X_2, \dots, X_n), \dots$$

Si desidera che al crescere di n la successione $\{\hat{\theta}_n\}$ di stimatori fornisca una stima sempre più accurata del parametro θ . Proprietà esibite da stimatori al divergere della taglia del campione sono dette *proprietà asintotiche*.

Se lo stimatore $\hat{\theta}_n$ non è corretto, è desiderabile richiedere che la sua distorsione $E(\hat{\theta}_n) - \theta$ tenda a zero al divergere della taglia del campione. Conviene pertanto dare la seguente definizione:

Definizione 3.4.1 Uno stimatore $\hat{\theta}_n = g_n(X_1, X_2, \dots, X_n)$ del parametro incognito θ è detto asintoticamente corretto, o asintoticamente non distorto, se

$$\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \lim_{n \rightarrow \infty} E[g_n(X_1, X_2, \dots, X_n)] = \theta.$$

La correttezza asintotica richiede dunque che la media di $\hat{\theta}_n$ tenda al parametro θ da stimare al divergere della taglia del campione.

È evidente che uno stimatore corretto è anche asintoticamente corretto. Forniremo ora due esempi di stimatori che non sono corretti, ma che lo sono asintoticamente.

Esempio 3.4.1 Sia \bar{X}_n la media campionaria di un campione casuale di taglia n estratto da una popolazione di media μ e varianza σ^2 . Dal Teorema 1.3.1 segue:

$$E(\bar{X}_n^2) = D^2(\bar{X}_n) + [E(\bar{X}_n)]^2 = \frac{\sigma^2}{n} + \mu^2,$$

così che \bar{X}_n^2 non è uno stimatore corretto di μ^2 . Ma poiché

$$\lim_{n \rightarrow \infty} E(\bar{X}_n^2) = \mu^2,$$

si conclude che \bar{X}_n^2 è uno stimatore asintoticamente corretto di μ^2 . ◆

Esempio 3.4.2 Sia S la deviazione standard campionaria di un campione casuale di taglia n estratto da una popolazione normale di varianza σ^2 . Come mostrato nella Proposizione 3.2.4, S non è uno stimatore corretto di σ . Tuttavia, come ora mostreremo, esso è asintoticamente corretto. Invero, notiamo anzitutto che dalla (3.17) si ha:

$$\lim_{n \rightarrow \infty} E(S) = \sigma \lim_{n \rightarrow \infty} \sqrt{\frac{2}{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}. \quad (3.56)$$

Dalla formula di Stirling, per $n \rightarrow \infty$ si ottiene

$$\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sim \frac{\left(\frac{n}{2}-1\right)^{\frac{n}{2}-1} e^{-\frac{n}{2}+1} \sqrt{2\pi\left(\frac{n}{2}-1\right)}}{\left(\frac{n-3}{2}\right)^{\frac{n-3}{2}} e^{-\frac{n}{2}+\frac{3}{2}} \sqrt{2\pi\left(\frac{n-3}{2}\right)}},$$

così che

$$\lim_{n \rightarrow \infty} \sqrt{\frac{2}{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} = \lim_{n \rightarrow \infty} \sqrt{\frac{n-2}{n-1}} \left(\frac{n-2}{n-3}\right)^{\frac{n-2}{2}-1} e^{-1/2} = 1.$$

Pertanto, dalla (3.56) segue $\lim_{n \rightarrow \infty} E(S) = \sigma$, da cui si conclude che S è uno stimatore asintoticamente corretto di σ . \diamond

Come visto nel § 3.3, l'idea di ricercare stimatori corretti efficienti è suggerita dall'esigenza di minimizzare il discostamento tra i valori di uno stimatore e il parametro da stimare. Non è però sempre possibile individuare uno stimatore caratterizzato da tale proprietà; esso potrebbe, ad esempio, non possedere varianza finita, nel qual caso l'obiettivo di minimizzazione della varianza perderebbe di significato. Anche quando non è possibile ottenere stimatori efficienti si è comunque interessati a costruire stimatori che forniscano stime puntuali che al divergere della taglia del campione siano "prossime il più possibile" al parametro da stimare. Questo concetto viene formalizzato nella definizione di stimatore *consistente*.

Definizione 3.4.2 *Uno stimatore $\hat{\Theta}_n = g_n(X_1, X_2, \dots, X_n)$ del parametro θ è detto consistente se per ogni $\epsilon > 0$ si ha:*

$$\lim_{n \rightarrow \infty} P(|\hat{\Theta}_n - \theta| < \epsilon) = 1.$$

È bene rilevare che il termine "consistente" è l'italianizzazione incorretta della parola inglese "consistent" che in Italiano significa "coerente". L'uso ormai universalmente accettato dell'aggettivo "consistente" nel contesto della teoria degli stimatori ne suggerisce, peraltro, anche qui l'uso.

La Definizione 3.4.2 di stimatore consistente si può anche riformulare affermando che $\hat{\Theta}_n$ tende in probabilità⁷ al parametro θ da stimare.

Si noti che la consistenza, così come la correttezza asintotica, è una proprietà di tipo asintotico. In maniera informale si può affermare che se $\hat{\Theta}_n$ è uno stimatore consistente di θ , al divergere della taglia n del campione tende all'unità la probabilità che l'errore assoluto $|\hat{\Theta}_n - \theta|$ sia inferiore ad ogni preassegnata costante positiva.

Nel teorema seguente sono fornite delle condizioni sufficienti affinché uno stimatore sia consistente.

Teorema 3.4.1 *Sia $\hat{\Theta}_n = g_n(X_1, X_2, \dots, X_n)$ uno stimatore asintoticamente corretto del parametro θ dotato di varianza finita. Se risulta*

$$\lim_{n \rightarrow \infty} D^2(\hat{\Theta}_n) = 0, \quad (3.57)$$

⁷Ricordiamo che una successione $\{X_n\}$ di variabili casuali si dice tendere in probabilità ad una variabile casuale X se risulta $\lim_{n \rightarrow \infty} P(|X_n - X| < \epsilon) = 1$ per ogni prefissato $\epsilon > 0$.

3.4. PROPRIETÀ ASINTOTICHE DEGLI STIMATORI

allora $\hat{\Theta}_n$ è consistente. \square $E_{\theta}[\hat{\Theta}_n - \theta] / \epsilon^2 = 1$

Dim. Cominciamo col dimostrare che per ogni $\epsilon > 0$ sussiste la seguente diseguaglianza:

$$P(|\hat{\Theta}_n - \theta| \geq \epsilon) \leq \frac{E[(\hat{\Theta}_n - \theta)^2]}{\epsilon^2} = \frac{D^2(\hat{\Theta}_n) + [E(\hat{\Theta}_n - \theta)]^2}{\epsilon^2}. \quad (3.58)$$

A tal fine osserviamo che, denotando con F_n la funzione di distribuzione di $\hat{\Theta}_n$, per ogni $\epsilon > 0$ risulta

$$\begin{aligned} E[(\hat{\Theta}_n - \theta)^2] &= \int_{\mathbf{R}} (x - \theta)^2 dF_n(x) \geq \int_{\{x: |x-\theta| \geq \epsilon\}} (x - \theta)^2 dF_n(x) \\ &\geq \epsilon^2 \int_{\{x: |x-\theta| \geq \epsilon\}} dF_n(x) \equiv \epsilon^2 P(|\hat{\Theta}_n - \theta| \geq \epsilon), \end{aligned}$$

che fornisce la (3.58). La dimostrazione del teorema segue ora facilmente. Invero, per la (3.57) e per la postulata asintotica correttezza di $\hat{\Theta}_n$, il secondo membro della (3.58) tende a zero al divergere di n , così che per ogni $\epsilon > 0$ si ha:

$$\lim_{n \rightarrow \infty} P(|\hat{\Theta}_n - \theta| < \epsilon) = 1 - \lim_{n \rightarrow \infty} P(|\hat{\Theta}_n - \theta| \geq \epsilon) = 1.$$

In virtù della Definizione 3.4.2, $\hat{\Theta}_n$ è dunque uno stimatore consistente di θ . \blacksquare

Si noti che se $\hat{\Theta}_n$ è uno stimatore corretto, la condizione (3.57) ne assicura a fortiori la consistenza. Vä inoltre sottolineato che uno stimatore non corretto può essere consistente solo se è asintoticamente corretto.

Ricorrendo al Teorema 3.4.1 forniremo ora alcuni esempi di stimatori consistenti.

Esempio 3.4.3 Consideriamo un campione casuale estratto da una popolazione di valore medio μ e varianza σ^2 finiti e dimostriamo che la media campionaria \bar{X} è uno stimatore consistente di μ . Invero, dal Teorema 1.3.1 si ha $E(\bar{X}) = \mu$, così che la media campionaria è uno stimatore corretto di μ . Inoltre risulta $D^2(\bar{X}) = \sigma^2/n$, il che rende soddisfatta l'ipotesi (3.57) per \bar{X} . Dal Teorema 3.4.1 segue dunque che \bar{X} è uno stimatore consistente di μ . \diamond

Esempio 3.4.4 Nel caso di un campione casuale estratto da una popolazione normale di varianza σ^2 finita, la varianza campionaria S^2 è uno stimatore consistente di σ^2 . Invero, S^2 è uno stimatore corretto di σ^2 essendo $E(S^2) = \sigma^2$. Inoltre, poiché nel Corollario 2.1.2 è mostrato che $D^2(S^2) = 2\sigma^4/(n-1)$, l'ipotesi (3.57) del Teorema 3.4.1 è soddisfatta. Quindi S^2 è uno stimatore consistente di σ^2 . \diamond

Esempio 3.4.5 Dato un campione (X_1, X_2, \dots, X_n) estratto da una popolazione di variabile casuale genitrice X , mostriamo che i momenti campionari $\bar{X}^{(k)} = (n)^{-1} \sum_{i=1}^n X_i^k$ sono stimatori consistenti dei momenti $\mu'_k = E(X^k)$. Infatti, come visto nella Proposizione 3.2.1, si ha $E(\bar{X}^{(k)}) = \mu'_k$, così che il momento campionario di ordine k è uno stimatore corretto del

momento intorno all'origine di pari ordine. Inoltre, per la varianza di $\bar{X}^{(k)}$ si ha:

$$\begin{aligned} D^2(\bar{X}^{(k)}) &= E[(\bar{X}^{(k)})^2] - [E(\bar{X}^{(k)})]^2 \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n E(X_i^k X_j^k) - (\mu'_k)^2 \\ &= \frac{1}{n^2} \sum_{i=1}^n E(X_i^{2k}) + \frac{1}{n^2} \sum_{i \neq j} E(X_i^k) E(X_j^k) - (\mu'_k)^2 \\ &= \frac{1}{n} \mu_{2k} + \frac{1}{n^2} n(n-1) (\mu'_k)^2 - (\mu'_k)^2 \\ &= \frac{\mu'_{2k} - (\mu'_k)^2}{n}, \end{aligned} \quad (3.59)$$

così che questa esiste finita se il momento μ'_{2k} è finito. Dalla (3.59) segue l'annullarsi di $D^2(\bar{X}^{(k)})$ al limite in cui $n \rightarrow \infty$; per il Teorema 3.4.1 il momento campionario di ordine k è dunque uno stimatore consistente di μ'_k . \diamond

Esempio 3.4.6 Sia (X_1, X_2, \dots, X_n) un campione casuale costituito da variabili di Bernoulli. La variabile $X = X_1 + X_2 + \dots + X_n$ ha dunque distribuzione binomiale di parametro θ . Nell'Esempio 3.2.2 abbiamo preso in esame i seguenti stimatori di θ :

$$\hat{\Theta} = \frac{X}{n}, \quad \hat{\Theta}' = \frac{X+1}{n+2}.$$

Essendo

$$E(\hat{\Theta}) = \theta, \quad E(\hat{\Theta}') = \frac{n\theta+1}{n+2},$$

$\hat{\Theta}$ è uno stimatore corretto mentre $\hat{\Theta}'$ è uno stimatore asintoticamente corretto. Inoltre, poiché $D^2(X) = n\theta(1-\theta)$, si ha:

$$\begin{aligned} D^2(\hat{\Theta}) &= D^2\left(\frac{X}{n}\right) = \frac{D^2(X)}{n^2} = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n} \\ D^2(\hat{\Theta}') &= D^2\left(\frac{X+1}{n+2}\right) = \frac{D^2(X)}{(n+2)^2} = \frac{n\theta(1-\theta)}{(n+2)^2}. \end{aligned}$$

L'ipotesi (3.57) è dunque soddisfatta sia per $\hat{\Theta}$ che per $\hat{\Theta}'$. Entrambi sono dunque stimatori consistenti di θ . \diamond

Esempio 3.4.7 Sia (X_1, X_2, \dots, X_n) un campione casuale tratto da una popolazione uniforme in $(0, \theta)$, dove θ è un parametro positivo incognito. La densità di probabilità della variabile casuale genitrice è quindi:

$$f(x) = \begin{cases} \frac{1}{\theta} & \text{per } 0 < x < \theta, \\ 0 & \text{altrimenti.} \end{cases}$$

Consideriamo le seguenti statistiche, definite in termini del massimo $X_{(n)}$ e dal minimo $X_{(1)}$ del campione:

$$\hat{\Theta}_n = \frac{n+1}{n} X_{(n)}, \quad \hat{\Theta}'_n = (n+1) X_{(1)}.$$

Entrambe sono stimatori corretti di θ . Infatti, ricordando le (3.5), si ha:

$$E(X_{(n)}) = \int_0^\theta x f_{(n)}(x) dx = \frac{n}{\theta^n} \int_0^\theta x^n dx = \frac{n}{n+1} \theta \quad (3.60)$$

$$\begin{aligned} E(X_{(1)}) &= \int_0^\theta x f_{(1)}(x) dx = n \int_0^\theta \frac{x}{\theta} \left(1 - \frac{x}{\theta}\right)^{n-1} dx \\ &= \left[-x \left(1 - \frac{x}{\theta}\right)^n\right]_0^\theta + \int_0^\theta \left(1 - \frac{x}{\theta}\right)^n dx = \frac{\theta}{n+1}, \end{aligned} \quad (3.61)$$

da cui segue:

$$E(\hat{\Theta}_n) = \frac{n+1}{n} E(X_{(n)}) = \theta, \quad E(\hat{\Theta}'_n) = (n+1) E(X_{(1)}) = \theta.$$

Si noti che, come c'era da attendersi, i valori medi (3.60) e (3.61) sono tali da aversi

$$E(X_{(n)}) - \frac{\theta}{2} = \frac{\theta}{2} - E(X_{(1)}),$$

ossia sono simmetrici rispetto alla media $\theta/2$ della popolazione. Volendo stabilire quale dei due stimatori sia dà preferire, calcoliamone le varianze. Dalle (3.5) si ricavano anzitutto i momenti del secondo ordine di $X_{(n)}$ e $X_{(1)}$:

$$\begin{aligned} E(X_{(n)}^2) &= \int_0^\theta x^2 f_{(n)}(x) dx = \frac{n}{\theta^n} \int_0^\theta x^{n+1} dx = \frac{n}{n+2} \theta^2 \\ E(X_{(1)}^2) &= \int_0^\theta x^2 f_{(1)}(x) dx = n \int_0^\theta \frac{x^2}{\theta} \left(1 - \frac{x}{\theta}\right)^{n-1} dx \\ &= \left[-x^2 \left(1 - \frac{x}{\theta}\right)^n\right]_0^\theta + 2 \int_0^\theta x \left(1 - \frac{x}{\theta}\right)^n dx = \frac{2\theta^2}{(n+1)(n+2)}. \end{aligned}$$

Da questi e dalle (3.60) e (3.61) si ottiene:

$$\begin{aligned} D^2(X_{(n)}) &= E(X_{(n)}^2) - [E(X_{(n)})]^2 = \frac{n}{n+2} \theta^2 - \frac{n^2}{(n+1)^2} \theta^2 \\ &= \frac{n}{(n+2)(n+1)^2} \theta^2 \end{aligned} \quad (3.62)$$

$$\begin{aligned} D^2(X_{(1)}) &= E(X_{(1)}^2) - [E(X_{(1)})]^2 = \frac{2\theta^2}{(n+1)(n+2)} - \frac{\theta^2}{(n+1)^2} \\ &= \frac{n}{(n+2)(n+1)^2} \theta^2, \end{aligned} \quad (3.63)$$

cosicché le varianze delle statistiche $X_{(n)}$ e $X_{(1)}$ coincidono, come d'altronde poteva prevedersi in base a considerazioni sulla simmetria di $X_{(1)}$ e $X_{(n)}$ rispetto alla media della popolazione. Dalle (3.62) e (3.63) si ricavano poi le varianze degli stimatori $\hat{\theta}_n$ e $\hat{\theta}'_n$:

$$\begin{aligned} D^2(\hat{\theta}_n) &= \frac{(n+1)^2}{n^2} D^2(X_{(n)}) = \frac{1}{n(n+2)} \theta^2, \\ D^2(\hat{\theta}'_n) &= (n+1)^2 D^2(X_{(1)}) = \frac{n}{n+2} \theta^2. \end{aligned}$$

Queste per $n=1$ coincidono, come c'era da attendersi in quanto in questo caso i due stimatori si identificano. Per $n>1$ si ha invece:

$$D^2(\hat{\theta}_n) = \frac{1}{n(n+2)} \theta^2 < \frac{n}{n+2} \theta^2 = D^2(\hat{\theta}'_n).$$

Se ne trae che $\hat{\theta}_n$ è da preferire a $\hat{\theta}'_n$ nella stima di θ . Si osservi inoltre che lo stimatore $\hat{\theta}_n$ è consistente dal momento che esso soddisfa le ipotesi del Teorema 3.4.1. Invece lo stimatore $\hat{\theta}'_n$ non è consistente. Ricordando infatti che

$$P(X_{(1)} > x) = \left(1 - \frac{x}{\theta}\right)^n \quad \text{per } 0 < x < \theta,$$

per ogni ε positivo (e minore di θ) risulta:

$$\begin{aligned} P(|\hat{\theta}'_n - \theta| < \varepsilon) &= P\left(\frac{\theta - \varepsilon}{n+1} < X_{(1)} < \frac{\theta + \varepsilon}{n+1}\right) \\ &= P\left(X_{(1)} > \frac{\theta - \varepsilon}{n+1}\right) - P\left(X_{(1)} > \frac{\theta + \varepsilon}{n+1}\right) \\ &= \left[1 - \frac{\theta - \varepsilon}{\theta(n+1)}\right]^n - \left[1 - \frac{\theta + \varepsilon}{\theta(n+1)}\right]^n. \end{aligned}$$

Da questa, passando al limite per $n \rightarrow \infty$, si ottiene:

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}'_n - \theta| < \varepsilon) = e^{-(\theta-\varepsilon)/\theta} - e^{-(\theta+\varepsilon)/\theta} < 1;$$

quindi $\hat{\theta}'_n$ non è uno stimatore consistente di θ . \diamond

È opportuno sottolineare che il Teorema 3.4.1 fornisce delle condizioni sufficienti, ma non necessarie, per la consistenza di uno stimatore. È invero possibile mostrare che esistono stimatori che sono consistenti ma che non posseggono varianza. A tal proposito, ricordiamo che nell'Esempio 3.4.3 si è mostrato che la media campionaria è uno stimatore consistente della media nell'ipotesi di finitezza di media e varianza; osserviamo tuttavia che per la legge debole dei grandi numeri⁸ la media campionaria risulta essere uno stimatore consistente della media anche se si rilassa l'ipotesi di finitezza della varianza.

Ricaveremo ora un'interessante proprietà degli stimatori consistenti che non trova l'analogo nel caso degli stimatori corretti. Nel seguente teorema si mostra invero come uno stimatore che sia funzione di uno stimatore consistente è anch'esso consistente.

⁸Ricordiamo che la legge debole dei grandi numeri afferma che se X_1, X_2, \dots sono variabili casuali indipendenti, identicamente distribuite e dotate di media finita μ , al divergere di n la variabile casuale $(X_1 + \dots + X_n)/n$ tende in probabilità a μ .

3.4. PROPRIETÀ ASINTOTICHE DEGLI STIMATORI

Teorema 3.4.2 *Sia $\hat{\theta}_n$ uno stimatore consistente di un parametro θ . Se $h(\hat{\theta}_n)$ è uno stimatore di $h(\theta)$, esso è consistente se h è una funzione continua.*

Dim. Essendo $h(\theta)$ continua, per ogni $\varepsilon > 0$ esiste un reale positivo δ_ε tale che $|x - \theta| < \delta_\varepsilon$ implica $|h(x) - h(\theta)| < \varepsilon$. Pertanto risulta:

$$\{|\hat{\theta}_n - \theta| < \delta_\varepsilon\} \subset \{|h(\hat{\theta}_n) - h(\theta)| < \varepsilon\},$$

da cui segue:

$$P(|\hat{\theta}_n - \theta| < \delta_\varepsilon) \leq P(|h(\hat{\theta}_n) - h(\theta)| < \varepsilon). \quad (3.64)$$

Poiché $\hat{\theta}_n$ è consistente, si ha:

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| < \delta_\varepsilon) = 1. \quad (3.65)$$

Al limite per $n \rightarrow \infty$, dalla (3.64) e dalla (3.65) segue

$$\lim_{n \rightarrow \infty} P(|h(\hat{\theta}_n) - h(\theta)| < \varepsilon) = 1,$$

così che $h(\hat{\theta}_n)$ è uno stimatore consistente di $h(\theta)$. \blacksquare

Il risultato appena dimostrato si può estendere facilmente al caso di più parametri. Sussiste infatti la seguente proposizione la cui dimostrazione procede analogamente a quella del Teorema 3.4.2, ed è pertanto omessa.

Proposizione 3.4.1 *Siano $\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}, \dots, \hat{\theta}_n^{(k)}$ stimatori consistenti di k parametri $\theta_1, \theta_2, \dots, \theta_k$. Se uno stimatore di $h(\theta_1, \theta_2, \dots, \theta_k)$ è dato da $h(\hat{\theta}_n^{(1)}, \hat{\theta}_n^{(2)}, \dots, \hat{\theta}_n^{(k)})$ con h funzione continua, allora tale stimatore è consistente.*

Esempio 3.4.8 Nell'Esempio 3.4.1 si è mostrato che se \bar{X}_n è la media campionaria di un campione casuale di taglia n estratto da una popolazione di media μ e varianza σ^2 , allora \bar{X}_n^2 è uno stimatore asintoticamente corretto di μ^2 . In aggiunta si può ora mostrare che \bar{X}_n^2 è uno stimatore consistente. Nell'Esempio 3.4.3, infatti, si è visto che \bar{X}_n è uno stimatore consistente di μ . Pertanto, poiché $h(\mu) = \mu^2$ è una funzione continua, dal Teorema 3.4.2 discende che \bar{X}_n^2 è uno stimatore consistente di μ^2 . \diamond

Fin qui si è parlato di stimatori consistenti; va tuttavia segnalato che alcuni autori preferiscono la terminologia "debolmente consistente" in luogo di "consistente"; ciò in quanto è possibile definire anche stimatori "fortemente consistenti", come qui appresso indicato.

Definizione 3.4.3 *Uno stimatore $\hat{\theta}_n = g_n(X_1, X_2, \dots, X_n)$ del parametro θ dotato di momento del secondo ordine finito è detto fortemente consistente, o anche consistente in media quadratica, se si ha:*

$$\lim_{n \rightarrow \infty} \text{mse}(\hat{\theta}_n) = 0.$$

close mso E[(\hat{\theta}_n - \theta)^2]

La proprietà di forte consistenza implica quella di consistenza debole, come espresso dalla proposizione che segue.

Proposizione 3.4.2 Se $\hat{\Theta}_n = g_n(X_1, X_2, \dots, X_n)$ è uno stimatore fortemente consistente del parametro θ , esso è anche consistente.

Dim. Dalla diseguaglianza (3.58), per ogni $\epsilon > 0$ si ricava:

$$P(|\hat{\Theta}_n - \theta| < \epsilon) \geq 1 - \frac{E[(\hat{\Theta}_n - \theta)^2]}{\epsilon^2}.$$

Essendo $\hat{\Theta}_n$ uno stimatore fortemente consistente di θ , l'errore quadratico medio $\text{mse}(\hat{\Theta}_n) = E[(\hat{\Theta}_n - \theta)^2]$ tende a zero al divergere di n , così che risulta $\lim_{n \rightarrow \infty} P(|\hat{\Theta}_n - \theta| < \epsilon) = 1$ per ogni $\epsilon > 0$. ■

La terminologia "debolmente consistente" e "fortemente consistente" trova le proprie origini nelle leggi debole e forte dei grandi numeri. Ad esempio, la media campionaria è sempre uno stimatore debolmente consistente proprio in virtù della legge debole dei grandi numeri, laddove diventa fortemente consistente nell'ipotesi aggiuntiva di varianza finita della variabile genitrice.

A conclusione di questo paragrafo va menzionato che le proprietà asintotiche degli stimatori, quali correttezza asintotica e consistenza, perdono d'interesse se i campioni di cui si dispone sono, come spesso accade, di piccola taglia, laddove risultano essere significative nel caso di campioni di taglia elevata. Va anche detto che a partire da stimatori caratterizzati da assegnate proprietà è talora possibile costruirne altri che godono delle medesime proprietà. Se, ad esempio, $\hat{\Theta}_n$ è uno stimatore asintoticamente corretto di un parametro θ , è possibile ottenere altri stimatori caratterizzati dalla stessa proprietà asintotica: basterà infatti porre $\hat{\Theta}'_n = \lambda_n \hat{\Theta}_n$, dove $\{\lambda_n\}$ è una successione convergente ad 1 i cui elementi non dipendono da θ . Tale ambiguità si può talvolta eliminare mediante un'operazione di "rinormalizzazione": invero, se $\hat{\Theta}_n$ è uno stimatore asintoticamente corretto del parametro θ e se risulta $E(\hat{\Theta}_n) = \theta/\lambda_n$, basterà porre $\hat{\Theta}'_n = \lambda_n \hat{\Theta}_n$ perché lo stimatore $\hat{\Theta}'_n$ sia corretto.

3.5 Statistiche sufficienti

In problemi statistici coinvolgenti grandi moli di dati si ottengono talora rilevanti semplificazioni ricorrendo a delle opportune quantità che sintetizzano informazioni contenute nei dati in esame. Un'analisi statistica imperniata su siffatte quantità può invero alle volte essere altrettanto efficace quanto quella effettuata sulla totalità dei dati; queste quantità vengono pertanto denominate "statistiche sufficienti". Più precisamente, dato un campione casuale (X_1, X_2, \dots, X_n) caratterizzato dal parametro incognito θ , una statistica $T = g(X_1, X_2, \dots, X_n)$ si dice "sufficiente" se utilizza ciò che del campione casuale è rilevante per la stima di θ , ossia se tutte le informazioni ottenibili su θ a partire dalla realizzazione del campione sono ricavabili anche dai soli valori $t = g(x_1, x_2, \dots, x_n)$ assunti dalla statistica T .

Per chiarire formalmente questo concetto, consideriamo un campione (X_1, X_2, \dots, X_n) estratto da una popolazione caratterizzata dal parametro incognito θ e sia $T = g(X_1, X_2, \dots, X_n)$ una statistica. Denotiamo con $f(x_1, x_2, \dots, x_n | t)$ la densità [distribuzione] di probabilità del campione casuale condizionata da $T = t$ nell'ipotesi in cui la popolazione che genera il campione è continua [discreta]. Sia Ω l'insieme delle realizzazioni osservabili, e Ω^t l'insieme

3.5. STATISTICHE SUFFICIENTI

costituito dalle realizzazioni del campione in corrispondenza delle quali la statistica T assume il valore t :

$$\Omega^t = \{(x_1, x_2, \dots, x_n) \in \Omega : g(x_1, x_2, \dots, x_n) = t\}.$$

La collezione di insiemi Ω^t costituisce una partizione di Ω , essendo

$$\Omega = \bigcup_t \Omega^t, \quad \Omega^t \cap \Omega^{\tau} = \emptyset \quad \text{per } t \neq \tau.$$

Se $f(x_1, x_2, \dots, x_n | t)$ dipende dal parametro θ , al variare di questo tra le realizzazioni di Ω^t ve ne saranno alcune caratterizzate da densità [probabilità] maggiore di quella di altre. Sapere quale realizzazione è stata osservata può quindi essere di ausilio per la stima di θ giacché l'osservazione effettuata fornisce in generale delle informazioni aggiuntive rispetto alla mera conoscenza che T assume il valore t . Ad esempio, siano $(x'_1, x'_2, \dots, x'_n)$ e $(x''_1, x''_2, \dots, x''_n)$ realizzazioni di un campione casuale, entrambe appartenenti all'insieme Ω^t . Supponiamo che $f(x_1, x_2, \dots, x_n | t)$ dipenda da θ così che al variare di θ la densità [distribuzione] di probabilità del campione casuale condizionata da $T = t$ assume, in generale, valori differenti. Supponiamo che si abbia

$$\begin{aligned} f(x'_1, x'_2, \dots, x'_n | t) &> f(x''_1, x''_2, \dots, x''_n | t) && \text{per } \theta \in \Theta_0, \\ f(x'_1, x'_2, \dots, x'_n | t) &< f(x''_1, x''_2, \dots, x''_n | t) && \text{per } \theta \in \Theta_1, \end{aligned}$$

con $\Theta_0 \cap \Theta_1 = \emptyset$. La conoscenza di quale delle due realizzazioni sia stata osservata fornisce quindi delle informazioni per la stima di θ che sono aggiuntive rispetto al sapere solo che la realizzazione osservata appartiene a Ω^t , ossia che la statistica T assume il valore t . Infatti, se la realizzazione osservata è $(x'_1, x'_2, \dots, x'_n)$ si è maggiormente inclini a ritenerne che il valore del parametro θ è incluso in Θ_0 ; al contrario, l'osservazione di $(x''_1, x''_2, \dots, x''_n)$ fa ritenere che sia $\theta \in \Theta_1$. Si noti che se $f(x_1, x_2, \dots, x_n | t)$ non dipendesse da θ , ossia se la densità [distribuzione] di probabilità delle realizzazioni del campione appartenenti a Ω^t rimanesse immutata al variare di θ , il sapere quale delle due realizzazioni sia stata osservata non contribuirebbe nel senso prima specificato alla stima di θ . Chiariamo questi concetti con un esempio.

Esempio 3.5.1 Sia (X_1, X_2, \dots, X_n) un campione costituito da variabili casuali di Bernoulli di parametro $\theta \in (0, 1)$. L'insieme Ω consiste quindi di tutte le sequenze di n elementi ciascuno dei quali è uguale a 0 oppure a 1:

$$\Omega = \{(x_1, x_2, \dots, x_n) : x_i \in \{0, 1\}, i = 1, 2, \dots, n\}.$$

Analizzeremo qui due differenti scelte per la statistica T .

(i) Sia $T = X_{(n)} = \max\{X_1, X_2, \dots, X_n\}$. Poiché risulta

$$\begin{aligned} P(T = t | X_1 = x_1, \dots, X_n = x_n) &= \begin{cases} 1 & \text{per } t = \max\{x_1, x_2, \dots, x_n\}, \\ 0 & \text{altrimenti,} \end{cases} \\ P(X_1 = x_1, \dots, X_n = x_n) &= \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} \\ &= \theta^{x_1 + \dots + x_n} (1-\theta)^{n-(x_1 + \dots + x_n)} \end{aligned}$$

e inoltre

$$P(T=t) = [1 - (1-\theta)^n]^t (1-\theta)^{n(1-t)} \quad (t=0,1),$$

la probabilità condizionata $f(x_1, x_2, \dots, x_n | t)$ è data da

$$\begin{aligned} f(x_1, x_2, \dots, x_n | t) &\equiv P(X_1 = x_1, \dots, X_n = x_n | T = t) \\ &= P(T = t | X_1 = x_1, \dots, X_n = x_n) \frac{P(X_1 = x_1, \dots, X_n = x_n)}{P(T = t)} \\ &= \frac{\theta^{x_1 + \dots + x_n} (1-\theta)^{n-(x_1 + \dots + x_n)}}{[1 - (1-\theta)^n]^t (1-\theta)^{n(1-t)}} \Big|_{\max\{x_1, x_2, \dots, x_n\}=t} \\ &= \frac{1}{[(1-\theta)^{-n}-1]^t} \left(\frac{\theta}{1-\theta} \right)^{x_1 + \dots + x_n} \Big|_{\max\{x_1, x_2, \dots, x_n\}=t}. \end{aligned} \quad (3.66)$$

Per ogni $t \in \{0,1\}$ la probabilità condizionata (3.66) dipende dal parametro θ . Verifichiamo quanto affermato in precedenza, ossia che per la stima di θ la conoscenza della realizzazione osservata fornisce informazioni ulteriori rispetto al sapere che la variabile casuale T assume il valore t . Infatti, supponiamo ad esempio che non sia nota la realizzazione osservata, ma che si sappia che è una delle seguenti: $(1, 0, \dots, 0)$ oppure $(1, 1, \dots, 1)$. In entrambi i casi la statistica T assume valore 1. Inoltre, dalla (3.66) si trae:

$$\begin{aligned} f(1, 0, \dots, 0 | 1) &= \frac{1}{(1-\theta)^{-n}-1} \left(\frac{\theta}{1-\theta} \right) \\ f(1, 1, \dots, 1 | 1) &= \frac{1}{(1-\theta)^{-n}-1} \left(\frac{\theta}{1-\theta} \right)^n, \end{aligned}$$

da cui segue:

$$\begin{aligned} f(1, 0, \dots, 0 | 1) &> f(1, 1, \dots, 1 | 1) && \text{per } \theta < \frac{1}{2}, \\ f(1, 0, \dots, 0 | 1) &< f(1, 1, \dots, 1 | 1) && \text{per } \theta > \frac{1}{2}. \end{aligned}$$

Queste diseguaglianze conducono alle seguenti conclusioni: se la realizzazione osservata è $(1, 0, \dots, 0)$ si è maggiormente inclini ad affermare che sia $\theta < 1/2$, mentre se si osserva $(1, 1, \dots, 1)$ si è portati a ritenere che sia $\theta > 1/2$. Il sapere che T assume il valore 1 non è quindi sufficiente per la stima di θ , visto che la conoscenza della realizzazione osservata fornisce informazioni ulteriori.

(ii) Sia $T = X_1 + X_2 + \dots + X_n$. In questo caso risulta

$$P(T=t | X_1 = x_1, \dots, X_n = x_n) = \begin{cases} 1 & \text{per } t = x_1 + x_2 + \dots + x_n, \\ 0 & \text{altrimenti} \end{cases}$$

e

$$P(T=t) = \binom{n}{t} \theta^t (1-\theta)^{n-t},$$

essendo T una variabile binomiale di parametro θ . Si ricava allora facilmente la probabilità condizionata $f(x_1, x_2, \dots, x_n | t)$:

$$\begin{aligned} f(x_1, x_2, \dots, x_n | t) &= \frac{\theta^{x_1 + \dots + x_n} (1-\theta)^{n-(x_1 + \dots + x_n)}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} \Big|_{x_1 + \dots + x_n=t} \\ &= \frac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} = \binom{n}{t}^{-1}. \end{aligned} \quad (3.67)$$

Stavolta $f(x_1, x_2, \dots, x_n | t)$ per ogni $t = 0, 1, \dots, n$ non dipende dal parametro θ . A differenza del caso (i), la conoscenza della realizzazione osservata ora non fornisce informazioni ulteriori per la stima di θ rispetto al sapere che T assume il valore t . Infatti, se ad esempio è noto che la realizzazione osservata è tale che la statistica T assume valore k , la conoscenza dell'intera realizzazione non aggiunge nulla di utile per la stima di θ ; per ogni realizzazione tale che $x_1 + x_2 + \dots + x_n = k$ la (3.67) fornisce infatti

$$f(x_1, x_2, \dots, x_n | k) = \binom{n}{k}^{-1},$$

e questa non dipende da θ . La conoscenza del valore assunto da T è quindi sufficiente per la stima di θ . ◆

In base a quanto visto finora siamo dunque condotti a dare la seguente definizione:

Definizione 3.5.1 Una statistica $T = g(X_1, X_2, \dots, X_n)$ è detta *sufficiente per la stima del parametro θ* se per ogni valore t assunto da T la densità [distribuzione] di probabilità condizionata $f(x_1, x_2, \dots, x_n | t)$ del campione (X_1, X_2, \dots, X_n) , dato che $T = t$, non dipende da θ .

Con riferimento al precedente esempio, la Definizione 3.5.1 porta così alla conclusione che, al contrario della statistica $T = X_1 + X_2 + \dots + X_n$, la statistica $X_{(n)} = \max\{X_1, X_2, \dots, X_n\}$ non è sufficiente per la stima del parametro θ di una popolazione di Bernoulli.

Esempio 3.5.2 Se (X_1, X_2, \dots, X_n) è un campione casuale costituito da variabili di Poisson di parametro θ , la statistica $T = X_1 + X_2 + \dots + X_n$ è sufficiente per la stima di θ . Si ha infatti:

$$\begin{aligned} f(x_1, x_2, \dots, x_n | t) &= P(X_1 = x_1, \dots, X_n = x_n | T = t) \\ &= \frac{1}{P(T=t)} \left(\prod_{i=1}^n e^{-\theta} \frac{\theta^{x_i}}{x_i!} \right) \Big|_{x_1 + \dots + x_n=t} \\ &= \frac{1}{P(T=t)} \frac{e^{-n\theta} \theta^t}{(x_1! \dots x_n!) \Big|_{x_1 + \dots + x_n=t}}. \end{aligned}$$

Poiché $T = X_1 + X_2 + \dots + X_n$ è una variabile casuale di Poisson di parametro $n\theta$, risulta:

$$P(T=t) = e^{-n\theta} \frac{(n\theta)^t}{t!} \quad (t=0, 1, \dots),$$

e quindi:

$$\begin{aligned} f(x_1, x_2, \dots, x_n | t) &= e^{n\theta} \frac{t!}{(n\theta)^t} \frac{e^{-n\theta} \theta^t}{(x_1! \cdots x_n!)_{x_1+\cdots+x_n=t}} \\ &= \frac{t!}{n! (x_1! \cdots x_n!)_{x_1+\cdots+x_n=t}}. \end{aligned}$$

Per ogni valore t di T la probabilità condizionata $f(x_1, x_2, \dots, x_n | t)$ non dipende dunque dal parametro θ . Dalla Definizione 3.5.1 si conclude che T è una statistica sufficiente per la stima di θ . \diamond

Dimostrare che una statistica è sufficiente facendo uso della Definizione 3.5.1 è in genere complicato. Una maniera più agevole consiste nel ricorrere al seguente *teorema di fattorizzazione*.

Teorema 3.5.1 *Dato un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione caratterizzata da un parametro incognito θ , una statistica $T = g(X_1, X_2, \dots, X_n)$ è detta statistica sufficiente per stimare θ se e solo se la densità [distribuzione] di probabilità congiunta $f(x_1, x_2, \dots, x_n; \theta)$ del campione casuale può essere così fattorizzata:*

$$f(x_1, x_2, \dots, x_n; \theta) = k(t, \theta) h(x_1, x_2, \dots, x_n), \quad (3.68)$$

dove $k(t, \theta)$ è una funzione dipendente solo da $t \equiv g(x_1, x_2, \dots, x_n)$ e da θ , mentre $h(x_1, x_2, \dots, x_n)$ è una funzione che non dipende da θ .

Dim. Limitiamo la dimostrazione al caso di un campione estratto da una popolazione discreta. Supponiamo che la distribuzione di probabilità $f(x_1, x_2, \dots, x_n; \theta)$ di (X_1, X_2, \dots, X_n) si fattorizzi come indicato nella (3.68). Se indichiamo con $t \equiv g(x_1, x_2, \dots, x_n)$ un valore assunto dalla statistica T , dalla (3.68) segue:

$$\begin{aligned} P(T = t) &= P[g(X_1, X_2, \dots, X_n) = t] = \sum_{\Omega'} f(x_1, x_2, \dots, x_n; \theta) \\ &= \sum_{\Omega'} k[g(x_1, x_2, \dots, x_n), \theta] h(x_1, x_2, \dots, x_n), \end{aligned} \quad (3.69)$$

dove le somme si intendono estese a tutte le n -uple (x_1, x_2, \dots, x_n) appartenenti all'insieme

$$\Omega' = \{(x_1, x_2, \dots, x_n); g(x_1, x_2, \dots, x_n) = t\}.$$

Dalla (3.69) segue:

$$P(T = t) = k(t, \theta) \sum_{\Omega'} h(x_1, x_2, \dots, x_n).$$

Pertanto, in virtù della (3.68), la distribuzione di X_1, X_2, \dots, X_n condizionata da $T = t$ è data da

$$f(x_1, x_2, \dots, x_n | t) = \frac{P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n, T = t)}{P(T = t)}$$

3.5. STATISTICHE SUFFICIENTI

$$\begin{aligned} &= \frac{f(x_1, x_2, \dots, x_n; \theta)|_{g(x_1, x_2, \dots, x_n)=t}}{k(t, \theta) \sum_{\Omega'} h(x_1, x_2, \dots, x_n)} \\ &= \frac{k(t, \theta) h(x_1, x_2, \dots, x_n)|_{g(x_1, x_2, \dots, x_n)=t}}{k(t, \theta) \sum_{\Omega'} h(x_1, x_2, \dots, x_n)} \\ &= \frac{h(x_1, x_2, \dots, x_n)|_{g(x_1, x_2, \dots, x_n)=t}}{\sum_{\Omega'} h(x_1, x_2, \dots, x_n)} \end{aligned}$$

che per ogni t non dipende da θ . In virtù della Definizione 3.5.1 si conclude che T è una statistica sufficiente per la stima di θ . Per dimostrare la parte necessaria del teorema, supponiamo che T sia una statistica sufficiente. Poiché

$$P(T = t | X_1 = x_1, \dots, X_n = x_n) = \begin{cases} 1 & \text{per } g(x_1, x_2, \dots, x_n) = t, \\ 0 & \text{altrimenti,} \end{cases}$$

ponendo $t = g(x_1, x_2, \dots, x_n)$ la distribuzione di probabilità del campione casuale si esprime nel seguente modo:

$$\begin{aligned} &f(x_1, x_2, \dots, x_n; \theta) \\ &= P(X_1 = x_1, \dots, X_n = x_n) P(T = t | X_1 = x_1, \dots, X_n = x_n) \\ &= P(X_1 = x_1, \dots, X_n = x_n, T = t) \\ &= P(T = t) P(X_1 = x_1, \dots, X_n = x_n | T = t). \end{aligned}$$

Posto poi

$$\begin{aligned} k(t, \theta) &= P(T = t) \\ h(x_1, x_2, \dots, x_n) &= P(X_1 = x_1, \dots, X_n = x_n | T = t), \end{aligned}$$

e ricordando che $h(x_1, x_2, \dots, x_n)$ non dipende da θ perché per ipotesi T è una statistica sufficiente per la stima di θ , segue immediatamente:

$$f(x_1, x_2, \dots, x_n; \theta) = k(t, \theta) h(x_1, x_2, \dots, x_n),$$

ossia la (3.68). \blacksquare

Esempio 3.5.3 Sia (X_1, X_2, \dots, X_n) un campione estratto da una popolazione binomiale di parametri k noto e θ incognito. Facendo ricorso al Teorema 3.5.1 mostriamo che $T = X_1 + X_2 + \dots + X_n$ è una statistica sufficiente per la stima di θ . A tal fine consideriamo la probabilità congiunta

$$\begin{aligned} f(x_1, x_2, \dots, x_n; \theta) &\equiv P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n; \theta) \\ &= \theta^{x_1+ \dots + x_n} (1-\theta)^{nk-(x_1+ \dots + x_n)} \prod_{i=1}^n \binom{k}{x_i}, \end{aligned}$$

dove $x_1, x_2, \dots, x_n \in \{0, 1, \dots, k\}$. Per $t = x_1 + x_2 + \dots + x_n$ si ha:

$$f(x_1, x_2, \dots, x_n; \theta) = \theta^t (1-\theta)^{nk-t} \prod_{i=1}^n \binom{k}{x_i}.$$

Se si sceglie

$$k(t, \theta) = \theta'(1-\theta)^{n-k-t}, \quad h(x_1, x_2, \dots, x_n) = \prod_{i=1}^n \binom{k}{x_i},$$

la relazione (3.68) è soddisfatta, e pertanto $T = X_1 + X_2 + \dots + X_n$ è una statistica sufficiente per la stima di θ . \diamond

Esempio 3.5.4 Consideriamo un campione (X_1, X_2, \dots, X_n) estratto da una popolazione normale di media μ e varianza σ^2 nota, e mostriamo che la statistica $T = X_1 + X_2 + \dots + X_n$ è sufficiente per la stima di μ . La densità congiunta del campione è infatti

$$\begin{aligned} f(x_1, x_2, \dots, x_n; \mu) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right] \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2 \right)\right]; \end{aligned}$$

ponendovi $x_1 + x_2 + \dots + x_n = t$ e scegliendo

$$\begin{aligned} \exp\left[-\frac{1}{2\sigma^2} (-2\mu t + n\mu^2)\right] &= k(t, \mu) \\ \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2\right) &= h(x_1, x_2, \dots, x_n), \end{aligned}$$

la densità congiunta del campione si fattorizza come indicato nella (3.68). Se ne conclude che $T = X_1 + X_2 + \dots + X_n$ è una statistica sufficiente per la stima della media di una popolazione normale. \diamond

I due corollari che seguono esprimono significative proprietà di cui godono le statistiche sufficienti.

Corollario 3.5.1 Sia $T = g(X_1, X_2, \dots, X_n)$ una statistica sufficiente per la stima di un parametro incognito θ e sia $u(\cdot)$ una funzione invertibile che non coinvolge θ . La statistica $Z = u(T)$ è allora anch'essa sufficiente per la stima di θ .

Dim. Questo risultato segue dal Teorema 3.5.1; invero, la formula (3.68) si può riscrivere come

$$f(x_1, x_2, \dots, x_n; \theta) = k[u^{-1}(z), \theta] h(x_1, x_2, \dots, x_n),$$

con $u^{-1}(z) = g(x_1, x_2, \dots, x_n)$ e dove la funzione $k[u^{-1}(z), \theta]$ dipende solo da z e θ . \blacksquare

Corollario 3.5.2 La statistica $Z = u(T)$ di cui al Corollario 3.5.1 è una statistica sufficiente per la stima del parametro $\lambda = u(\theta)$.

Dim. Anche questo risultato segue dal Teorema 3.5.1 dato che la formula (3.68) si può riscrivere nel seguente modo:

$$f[x_1, x_2, \dots, x_n; u^{-1}(\lambda)] = k[u^{-1}(z), u^{-1}(\lambda)] h(x_1, x_2, \dots, x_n).$$

Sia ora (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione di media μ . Se $T = X_1 + X_2 + \dots + X_n$ è una statistica sufficiente per la stima del parametro μ , dal Corollario 3.5.1 discende che la media campionaria $\bar{X} = T/n \equiv (X_1 + X_2 + \dots + X_n)/n$ è anch'essa una statistica sufficiente per la stima di μ ; in aggiunta, ne è anche uno stimatore corretto.

Una definizione analoga alla 3.5.1 viene data nel caso di stima di più parametri.

Definizione 3.5.2 Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione caratterizzata da k parametri incogniti $\theta_1, \theta_2, \dots, \theta_k$ e siano $T_i = g_i(X_1, X_2, \dots, X_n)$ ($i = 1, 2, \dots, k$) delle statistiche. Il vettore casuale (T_1, T_2, \dots, T_k) è detto statistica sufficiente per la stima dei parametri $\theta_1, \theta_2, \dots, \theta_k$ se la densità [distribuzione] condizionata $f(x_1, x_2, \dots, x_n | t_1, t_2, \dots, t_k)$, dato che $T_1 = t_1, T_2 = t_2, \dots, T_k = t_k$, non dipende da $\theta_1, \theta_2, \dots, \theta_k$ per ogni k -upla (t_1, t_2, \dots, t_k) di valori assunti da (T_1, T_2, \dots, T_k) .

Il Teorema 3.5.1, come subito vedremo omettendone per brevità la dimostrazione, si può estendere al caso di stima di più parametri. Nel seguito, quando verrà fatto riferimento ad una popolazione caratterizzata da k parametri incogniti si indicherà con θ il vettore dei parametri $(\theta_1, \theta_2, \dots, \theta_k)$.

Teorema 3.5.2 Dato un campione (X_1, X_2, \dots, X_n) estratto da una popolazione caratterizzata da k parametri incogniti $\theta = (\theta_1, \theta_2, \dots, \theta_k)$, la statistica k -dimensionale (T_1, T_2, \dots, T_k) , con $T_i = g_i(X_1, X_2, \dots, X_n)$, è detta statistica sufficiente per θ se e solo se la densità [distribuzione] di probabilità congiunta $f(x_1, x_2, \dots, x_n; \theta)$ del campione può essere così fattorizzata:

$$f(x_1, x_2, \dots, x_n; \theta) = k(t_1, t_2, \dots, t_k, \theta) h(x_1, x_2, \dots, x_n), \quad (3.70)$$

dove $k(t_1, t_2, \dots, t_k, \theta)$ è una funzione che dipende solo da θ e dai valori $t_i \equiv g_i(x_1, x_2, \dots, x_n)$, mentre $h(x_1, x_2, \dots, x_n)$ è una funzione che non dipende da θ .

Introdurremo ora una particolare famiglia di densità [distribuzioni] di probabilità caratterizzate da una particolare struttura grazie alla quale è possibile ricavare facilmente statistiche sufficienti.

Definizione 3.5.3 Si dice famiglia esponenziale una famiglia di densità [distribuzioni] di probabilità della forma

$$f(x; \theta) = c(\theta) \varphi(x) \exp\left[\sum_{i=1}^k \pi_i(\theta) d_i(x)\right], \quad (3.71)$$

dove $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ è un vettore di parametri mentre c , φ , π_i e d_i ($i = 1, 2, \dots, k$) sono funzioni opportune.

Si noti che la funzione $c(\theta)$, che appare nella (3.71), si può determinare dalla condizione di normalizzazione di f .

Teorema 3.5.3 Se (X_1, X_2, \dots, X_n) è un campione casuale estratto da una famiglia esponenziale, la k -upla

$$\left[T_1 \stackrel{\text{def}}{=} \sum_{j=1}^n d_1(X_j), T_2 \stackrel{\text{def}}{=} \sum_{j=1}^n d_2(X_j), \dots, T_k \stackrel{\text{def}}{=} \sum_{j=1}^n d_k(X_j) \right] \quad (3.72)$$

è una statistica sufficiente per la stima di $\theta = (\theta_1, \theta_2, \dots, \theta_k)$.

Dim. Indichiamo con $f(x_1, x_2, \dots, x_n; \theta)$ la densità [distribuzione] di probabilità congiunta del campione casuale. Facendo uso della (3.71) si ottiene:

$$f(x_1, x_2, \dots, x_n; \theta) = [c(\theta)]^n \exp \left[\sum_{i=1}^k \pi_i(\theta) \sum_{j=1}^n d_i(x_j) \right] \prod_{j=1}^n \varphi(x_j).$$

Pertanto, posto

$$t_i = \sum_{j=1}^n d_i(x_j) \quad (i = 1, 2, \dots, k),$$

$$k(t_1, t_2, \dots, t_k; \theta) = [c(\theta)]^n \exp \left[\sum_{i=1}^k \pi_i(\theta) t_i \right]$$

$$h(x_1, x_2, \dots, x_n) = \prod_{j=1}^n \varphi(x_j),$$

in virtù del Teorema 3.5.2 si conclude che la statistica (3.72) è sufficiente per la stima dei parametri $\theta_1, \theta_2, \dots, \theta_k$. \blacksquare

Negli esempi che seguono si fa uso del Teorema 3.5.3 per costruire statistiche sufficienti.

Esempio 3.5.5 Sia (X_1, X_2, \dots, X_n) un campione estratto da una popolazione di variabile generatrice a distribuzione gamma di parametri α e β , ossia di densità di probabilità

$$f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \quad (x > 0).$$

Notando che può equivalentemente scriversi

$$f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \exp \left[(\alpha-1) \ln x - \frac{x}{\beta} \right] \quad (x > 0),$$

si constata che la densità $f(x; \alpha, \beta)$ si può rappresentare nella forma (3.71) pur di porre

$$c(\alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)}, \quad \varphi(x) = 1,$$

$$d_1(x) = \ln x, \quad d_2(x) = x,$$

$$\pi_1(\alpha, \beta) = \alpha - 1, \quad \pi_2(\alpha, \beta) = -\frac{1}{\beta}.$$

3.5. STATISTICHE SUFFICIENTI

In virtù del Teorema 3.5.3 si conclude che il vettore

$$\left(\sum_{i=1}^n \ln X_i, \sum_{i=1}^n X_i \right)$$

costituisce una statistica bidimensionale sufficiente per la stima dei parametri α e β . Per il Corollario 3.5.1, segue inoltre che anche la coppia

$$\left(\prod_{i=1}^n X_i, \sum_{i=1}^n X_i \right)$$

costituisce una statistica bidimensionale sufficiente per la stima di α e β . \diamond

Esempio 3.5.6 Consideriamo un campione (X_1, X_2, \dots, X_n) estratto da una popolazione normale di media μ e varianza σ^2 incognite. La densità di probabilità della variabile casuale genitrice è dunque:

$$\begin{aligned} f(x; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{\mu^2}{2\sigma^2} \right) \exp \left(\frac{\mu}{\sigma^2} x - \frac{x^2}{2\sigma^2} \right) \quad (x \in \mathbb{R}). \end{aligned}$$

Notiamo che questa si può esprimere nella forma (3.71) che definisce la famiglia esponenziale, pur di scegliere

$$c(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{\mu^2}{2\sigma^2} \right), \quad \varphi(x) = 1,$$

$$\begin{aligned} \pi_1(\mu, \sigma^2) &= \frac{\mu}{\sigma^2}, & \pi_2(\mu, \sigma^2) &= -\frac{1}{2\sigma^2}, \\ d_1(x) &= x, & d_2(x) &= x^2. \end{aligned}$$

Dal Teorema 3.5.3 segue allora che la coppia

$$\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right) \quad (3.73)$$

è una statistica sufficiente per la stima di (μ, σ^2) . Si noti, infine, che a partire dalla coppia di statistiche sufficienti (3.73) si costruiscono gli stimatori

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad S^2 = \frac{1}{n-1} \left[\sum_{i=1}^n X_i^2 - \frac{1}{n} \left(\sum_{i=1}^n X_i \right)^2 \right]$$

della media μ e della varianza σ^2 . \diamond

Esempio 3.5.7 Si consideri un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione normale inversa di parametri μ e λ , ossia di densità di probabilità (cfr. Definizione 2.5.1)

$$\begin{aligned} f(x; \mu, \lambda) &= \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right] \\ &= \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} e^{\lambda/\mu} \exp\left(-\frac{\lambda}{2\mu^2}x - \frac{\lambda}{2} \frac{1}{x}\right) \quad (x > 0). \end{aligned} \quad (3.74)$$

Se si pone

$$c(\mu, \lambda) = \left(\frac{\lambda}{2\pi}\right)^{1/2} e^{\lambda/\mu}, \quad \varphi(x) = \left(\frac{1}{x}\right)^{3/2},$$

$$\pi_1(\mu, \lambda) = -\frac{\lambda}{2\mu^2}, \quad \pi_2(\mu, \lambda) = -\frac{\lambda}{2},$$

$$d_1(x) = x, \quad d_2(x) = \frac{1}{x},$$

la densità (3.74) assume la forma (3.71). Dal Teorema 3.5.3 segue allora che la coppia

$$\left(\sum_{i=1}^n X_i, \sum_{i=1}^n \frac{1}{X_i} \right)$$

è una statistica sufficiente per la stima dei parametri μ e λ . \diamond

Vedremo ora come sia possibile considerare un'altra famiglia di distribuzioni per la quale, come per quella esponenziale, si riescono a determinare delle statistiche sufficienti. Nel seguito con $I_A(z)$ denoteremo la funzione indicatrice così definita:

$$I_A(z) = \begin{cases} 1 & \text{per } z \in A, \\ 0 & \text{altrimenti.} \end{cases} \quad (3.75)$$

Teorema 3.5.4 Sia (X_1, X_2, \dots, X_n) un campione casuale di taglia n estratto da una popolazione avente densità [distribuzione] di probabilità della forma

$$f(x; \theta) = c(\theta) \varphi(x) I_{(\gamma, \pi_1(\theta))}(x) \exp\left[\sum_{i=2}^k \pi_i(\theta) d_i(x)\right], \quad (3.76)$$

dove c , φ , π_i ($i = 1, 2, \dots, k$) e d_i ($i = 2, 3, \dots, k$) sono funzioni opportune, θ denota la n -upla $(\theta_1, \theta_2, \dots, \theta_k)$ dei parametri e γ è una costante tale che $\pi_1(\theta) > \gamma$ per ogni θ . La k -upla

$$\left[T_1 \stackrel{\text{def}}{=} \max\{X_1, X_2, \dots, X_n\}, T_2 \stackrel{\text{def}}{=} \sum_{j=1}^n d_2(X_j), \dots, T_k \stackrel{\text{def}}{=} \sum_{j=1}^n d_k(X_j) \right] \quad (3.77)$$

è allora una statistica sufficiente per la stima di θ .

3.5. STATISTICHE SUFFICIENTI

Dim. Per la (3.76) la densità [distribuzione] di probabilità congiunta del campione può così esprimersi:

$$\begin{aligned} f(x_1, x_2, \dots, x_n; \theta) &= [c(\theta)]^n \exp\left[\sum_{i=2}^k \pi_i(\theta) \sum_{j=1}^n d_i(x_j)\right] \\ &\times \prod_{j=1}^n \left[\varphi(x_j) I_{(\gamma, \pi_1(\theta))}(x_j) \right]. \end{aligned}$$

Poiché risulta

$$\prod_{j=1}^n I_{(\gamma, \pi_1(\theta))}(x_j) = 1$$

se e solo se

$$I_{(\gamma, \pi_1(\theta))}(\max\{x_1, x_2, \dots, x_n\}) = 1,$$

si può scrivere:

$$\begin{aligned} f(x_1, x_2, \dots, x_n; \theta) &= [c(\theta)]^n \exp\left[\sum_{i=2}^k \pi_i(\theta) \sum_{j=1}^n d_i(x_j)\right] \\ &\times I_{(\gamma, \pi_1(\theta))}(\max\{x_1, x_2, \dots, x_n\}) \prod_{j=1}^n \varphi(x_j). \end{aligned}$$

Posto quindi

$$\begin{aligned} k(t_1, t_2, \dots, t_k; \theta) &= [c(\theta)]^n \exp\left[\sum_{i=2}^k \pi_i(\theta) \sum_{j=1}^n d_i(x_j)\right] \\ &\times I_{(\gamma, \pi_1(\theta))}(\max\{x_1, x_2, \dots, x_n\}) \\ h(x_1, x_2, \dots, x_n) &= \prod_{j=1}^n \varphi(x_j), \end{aligned}$$

dal Teorema 3.5.2 di fattorizzazione segue che la (3.77) è una statistica sufficiente per la stima dei parametri $\theta_1, \theta_2, \dots, \theta_k$. \blacksquare

Esempio 3.5.8 Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione uniforme nell'intervallo $(0, \theta)$, ossia di variabile genitrice di densità di probabilità

$$f(x; \theta) = \frac{1}{\theta} I_{(0, \theta)}(x). \quad (3.78)$$

Ricaviamo una statistica sufficiente per la stima del parametro θ . A tal fine osserviamo che alla (3.78) può darsi la forma (3.76) ponendo $c(\theta) = 1/\theta$, $\varphi(x) = 1$, $\pi_1(\theta) = \theta$. In virtù del Teorema 3.5.4 concludiamo che la statistica $T_1 \equiv X_{(n)} = \max\{X_1, X_2, \dots, X_n\}$ è sufficiente per la stima di θ . \diamond



3.6 Statistiche complete

Vedremo in questo paragrafo come le statistiche sufficienti possano essere utilizzate per ricercare estimatori corretti ed efficienti, ossia a varianza uniformemente minima, per la stima di un parametro incognito. A tale scopo consideriamo il seguente teorema:

Teorema 3.6.1 (Blackwell-Rao) *Siano X e Y variabili casuali con Y dotata di media e varianza finite. Posto $\varphi(x) = E(Y|X=x)$, risulta:*

$$E[\varphi(X)] = E(Y), \quad D^2[\varphi(X)] \leq D^2(Y), \quad (3.79)$$

sussistendo l'uguaglianza se e solo se $P[Y = \varphi(X)] = 1$.

Dim. Effettuiamo la dimostrazione nel caso in cui X e Y sono variabili continue. Poiché risulta

$$\begin{aligned} E[E(Y^k|X)] &= \int_{\mathbb{R}} E(Y^k|X=x) f_X(x) dx \\ &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} y^k f_{Y|X}(y|x) dy \right] f_X(x) dx \\ &= \int_{\mathbb{R}} y^k \left[\int_{\mathbb{R}} f_{Y|X}(y|x) f_X(x) dx \right] dy \\ &= \int_{\mathbb{R}} y^k f_Y(y) dy \\ &= E(Y^k), \end{aligned} \quad (3.80)$$

la prima delle (3.79) discende immediatamente dalla (3.80) per $k = 1$. Osserviamo ora che

$$D^2(Y) = E[D^2(Y|X)] + D^2[E(Y|X)]. \quad (3.81)$$

Infatti, facendo uso della (3.80) per $k = 1$ e per $k = 2$, si ha:

$$\begin{aligned} D^2(Y) &= E(Y^2) - [E(Y)]^2 \\ &= E[E(Y^2|X)] - \{E[E(Y|X)]\}^2 \\ &= E[E(Y^2|X)] - E\{[E(Y|X)]^2\} \\ &\quad + E\{[E(Y|X)]^2\} - \{E[E(Y|X)]\}^2 \\ &= E[D^2(Y|X)] + D^2[E(Y|X)]. \end{aligned}$$

Dalla (3.81) si ottiene poi:

$$D^2[\varphi(X)] = D^2[E(Y|X)] = D^2(Y) - E[D^2(Y|X)],$$

da cui, osservando che $E[D^2(Y|X)]$ è una quantità non negativa, discende la seconda delle (3.79). Sussiste inoltre l'uguaglianza $D^2[\varphi(X)] = D^2(Y)$ se e solo se $E[D^2(Y|X)] = 0$. Pertanto, avendosi

$$E[D^2(Y|X)] = \int_{\mathbb{R}} D^2(Y|X=x) f_X(x) dx$$

$$\begin{aligned} &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} [y - E(Y|X=x)]^2 f_{Y|X}(y|x) dy \right] f_X(x) dx \\ &= \int_{\mathbb{R}} dx \int_{\mathbb{R}} [y - E(Y|X=x)]^2 f_{X,Y}(x,y) dy \\ &= E\{[Y - E(Y|X)]^2\}, \end{aligned}$$

L'uguaglianza $D^2[\varphi(X)] = D^2(Y)$ risulta soddisfatta se e solo se risulta $E\{[Y - E(Y|X)]^2\} = 0$, il che accade se e solo se $P[Y = \varphi(X)] = 1$ giacché $[Y - E(Y|X)]^2$ è una variabile casuale non negativa. La dimostrazione nel caso discreto si effettua in modo analogo. ■

Come immediata conseguenza del Teorema 3.6.1 discende il teorema seguente:

Teorema 3.6.2 *Siano (X_1, X_2, \dots, X_n) un campione estratto da una popolazione caratterizzata dal parametro incognito θ , $T = g(X_1, X_2, \dots, X_n)$ una statistica sufficiente per la stima di θ e $\hat{\Theta}' = \gamma(X_1, X_2, \dots, X_n)$ uno stimatore corretto di θ a varianza finita; la variabile casuale $\hat{\Theta} = E(\hat{\Theta}'|T)$ è allora tale da avere*

$$E(\hat{\Theta}) = \theta, \quad D^2(\hat{\Theta}) \leq D^2(\hat{\Theta}'), \quad (3.82)$$

dove l'uguaglianza sussiste se e solo se $P(\hat{\Theta} = \hat{\Theta}') = 1$. Inoltre, la variabile casuale $\hat{\Theta}$ non dipende⁹ da θ .

Dim. Ponendo $X = T$ e $Y = \hat{\Theta}'$, dal Teorema 3.6.1 segue

$$E[E(\hat{\Theta}'|T)] = E(\hat{\Theta}'), \quad D^2[E(\hat{\Theta}'|T)] \leq D^2(\hat{\Theta}'),$$

dove l'uguaglianza sussiste se e solo se risulta $P[\hat{\Theta}' = E(\hat{\Theta}'|T)] = 1$. Poiché per ipotesi $\hat{\Theta} = E(\hat{\Theta}'|T)$ e $\hat{\Theta}'$ è uno stimatore corretto di θ , si ricava subito la (3.82) nella quale si ha l'uguaglianza se e solo se $P(\hat{\Theta} = \hat{\Theta}') = 1$. Dimostriamo ora nel caso continuo che la variabile casuale $\hat{\Theta}$ non dipende da θ . Notiamo anzitutto che sussiste la relazione

$$E(\hat{\Theta}'|T=t) = E[\gamma(X_1, X_2, \dots, X_n)|T=t] = \int_{\mathbb{R}^n} \gamma(x) f_{X|T}(x|t) dx, \quad (3.83)$$

dove $f_{X|T}(x|t)$ denota la densità del campione casuale condizionata da $T = t$. Essendo T una statistica sufficiente per la stima di θ , per la Definizione 3.5.1 $f_{X|T}(x|t)$ non dipende da θ ; conseguentemente, poiché nemmeno $\gamma(x)$ dipende da θ , anche il valore medio condizionato $E(\hat{\Theta}'|T=t)$, dato dalla (3.83), non dipende da θ . Da ciò, osservando che al variare di t la variabile casuale $\hat{\Theta} = E(\hat{\Theta}'|T)$ assume i valori $E(\hat{\Theta}'|T=t)$ con densità $f_T(t)$, segue che $\hat{\Theta}$ non dipende da θ in quanto anche la statistica T non dipende da θ . Il caso discreto può essere trattato in modo analogo. ■

Il teorema appena dimostrato riveste notevole importanza. Esso infatti afferma che nelle ipotesi in cui sono noti uno stimatore corretto di un parametro incognito ed una statistica sufficiente per la sua stima è possibile costruire un secondo stimatore corretto la cui varianza è non maggiore di quella del primo stimatore. Ricordando la Definizione 3.3.1, si conclude

⁹Per questo motivo $\hat{\Theta}$ può essere usata come stimatore di θ .

che il secondo stimatore è relativamente più efficiente del primo qualora le rispettive varianze non coincidano.

Esaminiamo un esempio in cui partendo da uno stimatore corretto e da una statistica sufficiente si fa uso del Teorema 3.6.2 per ricavare uno stimatore corretto relativamente più efficiente del primo.

Esempio 3.6.1 Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione di Bernoulli di parametro θ . Come si è mostrato nell'Esempio 3.5.1, $T = X_1 + X_2 + \dots + X_n$ è una statistica sufficiente per la stima di θ . Inoltre la variabile casuale X_1 è uno stimatore corretto di θ a varianza finita. Dal Teorema 3.6.2 segue allora che $\hat{\theta} = E(X_1|T)$ è uno stimatore corretto di θ ed è tale che $D^2(\hat{\theta}) \leq D^2(X_1)$. Per determinarlo osserviamo che si ha:

$$\begin{aligned} E(X_1|T=t) &= \sum_{k=0}^1 kP(X_1=k|T=t) = P(X_1=1|T=t) \\ &= \frac{P(X_1=1)P(T=t|X_1=1)}{P(T=t)}. \end{aligned} \quad (3.84)$$

Poiché somme di variabili di Bernoulli indipendenti e identicamente distribuite forniscono variabili binomiali, si ricavano facilmente le seguenti relazioni:

$$\begin{aligned} P(T=t) &= P(X_1+X_2+\dots+X_n=t) \\ &= \binom{n}{t} \theta^t (1-\theta)^{n-t}, \end{aligned} \quad (3.85)$$

$$\begin{aligned} P(T=t|X_1=1) &= P(X_1+X_2+\dots+X_n=t|X_1=1) \\ &= P(X_2+\dots+X_n=t-1) \\ &= \binom{n-1}{t-1} \theta^{t-1} (1-\theta)^{n-t}. \end{aligned} \quad (3.86)$$

Sostituendo le (3.85) e (3.86) nella (3.84), e ricordando che $P(X_1=1)=\theta$, si ha:

$$E(X_1|T=t) = \frac{\theta \binom{n-1}{t-1} \theta^{t-1} (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} = \frac{t}{n}.$$

Pertanto

$$\hat{\theta} = E(X_1|T) = \frac{T}{n} = \frac{1}{n} \sum_{i=1}^n X_i \equiv \bar{X},$$

ossia lo stimatore $\hat{\theta}$ coincide con la media campionaria. Poiché inoltre per $n > 1$ risulta

$$D^2(X_1) = \theta(1-\theta) > \frac{\theta(1-\theta)}{n} = D^2(\hat{\theta}),$$

$\hat{\theta}$ è relativamente più efficiente di X_1 .

Abbiamo visto che il Teorema 3.6.2 può essere usato per "migliorare" l'efficienza di stimatori corretti nel senso che, partendo da uno stimatore corretto e da una statistica sufficiente T , è possibile ricavare un secondo stimatore corretto di varianza non maggiore di quella del primo. È allora naturale chiedersi se la varianza dello stimatore così ottenuto possa essere ulteriormente ridotta, applicando ad esempio nuovamente la procedura suggerita dal Teorema 3.6.2 e facendo uso di una diversa statistica sufficiente. In effetti, la risposta a tale quesito è negativa se la densità [distribuzione] di probabilità di tale statistica appartiene ad una classe, detta *famiglia completa*, qui appresso definita.

Definizione 3.6.1 Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione caratterizzata da un parametro θ incognito appartenente ad un insieme Θ e sia $f(t; \theta)$ la densità [distribuzione] di probabilità di una statistica $T = g(X_1, X_2, \dots, X_n)$. La famiglia di funzioni $\{f(t; \theta), \theta \in \Theta\}$ si dice completa se per ogni funzione $h(t)$ la relazione

$$E[h(T)] = 0 \quad \text{per ogni } \theta \in \Theta$$

implica $h(t) = 0$ per ogni reale t per il quale risulta $f(t; \theta) > 0$.

Va menzionato che talora la proprietà di completezza viene riferita alla statistica T invece che alla famiglia di funzioni $f(t; \theta)$; in tal caso si dice che T è una statistica completa per la stima di θ .

Esaminiamo alcuni esempi in cui si determinano delle statistiche complete.

Esempio 3.6.2 Sia (X_1, X_2, \dots, X_n) un campione estratto da una popolazione binomiale di parametri k noto e θ incognito. Nell'Esempio 3.5.3 abbiamo visto che la statistica $T = X_1 + X_2 + \dots + X_n$ è sufficiente per la stima di θ ; dimostriamo ora che essa è completa. Cominciamo con l'osservare che T , essendo somma di variabili casuali binomiali indipendenti e identicamente distribuite, è anch'essa una variabile binomiale, caratterizzata dai parametri nk e θ . Pertanto la sua distribuzione di probabilità è

$$f(t; \theta) = \binom{nk}{t} \theta^t (1-\theta)^{nk-t} \quad (t = 0, 1, \dots, nk). \quad (3.87)$$

Consideriamo ora una funzione arbitraria $h(t)$ e imponiamo che risulti $E[h(T)] = 0$ per ogni $\theta \in (0, 1)$. Deve allora essere

$$E[h(T)] = \sum_{t=0}^{nk} h(t) f(t; \theta) = 0$$

ossia, ricordando la (3.87),

$$\sum_{t=0}^{nk} h(t) \binom{nk}{t} \left(\frac{\theta}{1-\theta} \right)^t = 0 \quad (3.88)$$

per ogni $\theta \in (0, 1)$. È facile rendersi conto che la (3.88) implica che la funzione $h(t)$ è nulla per $t = 0, 1, \dots, nk$. Pertanto, per la Definizione 3.6.1, la (3.87) costituisce una famiglia completa, così che T è una statistica completa per la stima di θ .

Esempio 3.6.3 Sia (X_1, X_2, \dots, X_n) un campione estratto da una popolazione di Poisson di parametro θ . Come mostrato nell'Esempio 3.5.2, la statistica $T = X_1 + X_2 + \dots + X_n$ è sufficiente per la stima di θ . Allo scopo di dimostrare che è anche completa, osserviamo che essa è la somma di n variabili casuali di Poisson indipendenti e identicamente distribuite, e pertanto T è una variabile casuale di Poisson di parametro $n\theta$, ossia di distribuzione di probabilità

$$f(t; \theta) = \frac{(n\theta)^t}{t!} e^{-n\theta} \quad (t = 0, 1, \dots). \quad (3.89)$$

Indichiamo ora con $h(t)$ una funzione qualsiasi ed imponiamo che risulti $E[h(T)] = 0$ per ogni $\theta > 0$:

$$E[h(T)] = \sum_{t=0}^{\infty} h(t) f(t; \theta) = 0. \quad (3.90)$$

Dalle (3.89) e (3.90) si trae che per ogni $\theta > 0$ deve avversi:

$$\sum_{t=0}^{\infty} h(t) \frac{(n\theta)^t}{t!} = 0. \quad (3.91)$$

La (3.91) implica che $h(t)$ è nulla per $t = 0, 1, \dots$. Da ciò, in virtù della Definizione 3.6.1, segue che la (3.89) è una famiglia completa, ossia che T è una statistica completa per la stima di θ . \diamond

Siamo ora in grado di dimostrare il seguente teorema che consente di ricavare stimatori corretti ed efficienti.

Teorema 3.6.3 (Lehmann-Scheffé) *Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione caratterizzata da un parametro incognito θ e sia $T = g(X_1, X_2, \dots, X_n)$ una statistica sufficiente e completa per la stima di θ . Se $\hat{\Theta} = \varphi(T)$ è uno stimatore corretto di θ dipendente da T , allora $\hat{\Theta}$ è unico ed efficiente.*

Dim. Supponiamo che oltre a $\hat{\Theta} = \varphi(T)$ esista un altro stimatore corretto di θ dipendente da T , che denotiamo con $\tilde{\Theta} = \psi(T)$, così che per ogni $\theta \in \Theta$ risulta $E(\hat{\Theta}) = E(\tilde{\Theta}) = \theta$. Posto $h(t) = \varphi(t) - \psi(t)$, si ha quindi:

$$E[h(T)] = E[\varphi(T) - \psi(T)] = E(\hat{\Theta} - \tilde{\Theta}) = 0$$

per ogni $\theta \in \Theta$. Essendo T una statistica completa, $h(t) = 0$ per la Definizione 3.6.1, e quindi $\varphi(t) = \psi(t)$ per ogni valore t per cui $f(t; \theta) > 0$. Esiste dunque un unico stimatore corretto $\hat{\Theta} = \varphi(T)$ dipendente da T . Prendiamo ora in esame uno stimatore qualsiasi $\hat{\Theta}'$ per la stima di θ che sia corretto e dotato di varianza finita. In virtù del Teorema 3.6.2 la variabile casuale $E(\hat{\Theta}'|T)$ è uno stimatore corretto di θ e dipende dalla statistica sufficiente T . Inoltre essa è tale da avversi $D^2[E(\hat{\Theta}'|T)] \leq D^2(\hat{\Theta}')$, dove l'uguaglianza sussiste se e solo se $P[E(\hat{\Theta}'|T) = \hat{\Theta}'] = 1$. Avendo già dimostrato che $\hat{\Theta} = \varphi(T)$ è l'unico stimatore corretto di θ dipendente da T , deve necessariamente avversi $\hat{\Theta} = E(\hat{\Theta}'|T)$, dove tale stimatore è corretto ed efficiente. \blacksquare

3.7. METODO DELLA MASSIMA VEROSIMIGLIANZA

Nell'esempio che segue faremo uso del Teorema 3.6.3 per ricavare uno stimatore corretto ed efficiente per la stima della media di popolazioni normali.

Esempio 3.6.4 Consideriamo un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione normale di media μ e varianza σ^2 nota. Come mostrato nell'Esempio 3.5.4, la statistica $T = X_1 + X_2 + \dots + X_n$ è sufficiente per la stima di μ . Essendo somma di n variabili casuali normali di media μ e varianza σ^2 , essa ha distribuzione normale di media $n\mu$ e varianza $n\sigma^2$, così che la sua densità di probabilità è

$$f_T(t; \mu) = \frac{1}{\sqrt{2\pi n\sigma^2}} \exp \left[-\frac{(t - n\mu)^2}{2n\sigma^2} \right] \quad (t \in \mathbb{R}). \quad (3.92)$$

Per dimostrare che T è anche una statistica completa, consideriamo una funzione arbitraria $h(t)$ e richiediamo che sia $E[h(T)] = 0$ per ogni $\theta > 0$. Deve allora avversi

$$E[h(T)] = \int_{\mathbb{R}} h(t) f_T(t; \mu) dt = 0$$

e quindi, per la (3.92), per ogni $\mu \in \mathbb{R}$ deve risultare:

$$\int_{\mathbb{R}} h(t) \exp \left[-\frac{(t - n\mu)^2}{2n\sigma^2} \right] dt = 0. \quad (3.93)$$

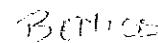
Dalla (3.93) si ricava che $h(t)$ è nulla per ogni $t \in \mathbb{R}$. In virtù della Definizione 3.6.1, la (3.92) è dunque una famiglia completa, ossia T è una statistica completa per la stima di μ . La statistica

$$\hat{\Theta} = \frac{T}{n} = \frac{1}{n} \sum_{i=1}^n X_i \equiv \bar{X}$$

risulta inoltre essere uno stimatore corretto del parametro μ dal momento che essa coincide con \bar{X} che, come sappiamo, è uno stimatore corretto di μ . Di conseguenza, $\hat{\Theta} \equiv \bar{X} = T/n$ è uno stimatore corretto di μ ed è funzione di una statistica T sufficiente e completa; in virtù del Teorema 3.6.3 segue così che la media campionaria è l'unico stimatore efficiente della media μ . \diamond

3.7 Metodo della massima verosimiglianza

Si è visto che gli stimatori preposti a fornire una stima puntuale dei parametri incogniti di una popolazione sulla base di un suo campione casuale (X_1, X_2, \dots, X_n) possono godere di alcune significative proprietà, quali correttezza, efficienza e consistenza. Si è anche notato che non sempre queste proprietà convivono in uno stesso stimatore. La decisione di quale proprietà sia da preferire, e quindi dello stimatore da adottare, è solitamente legata alle finalità che ci si prefissano. Abbiamo ad esempio fatto rilevare come la proprietà di consistenza sia significativa solo nel caso di campioni di taglia elevata. Rimane comunque da chiarire in che modo risulti possibile individuare gli stimatori che possano considerarsi potenziali buoni candidati per la stima dei parametri incogniti. A tale scopo verrà ora esposto un metodo dovuto allo statistico inglese R.A. Fisher, detto *metodo della massima verosimiglianza*, che consente di determinare degli stimatori $\hat{\Theta}_j = g_j(X_1, X_2, \dots, X_n)$ ($j = 1, 2, \dots, k$) dei parametri incogniti $\theta_1, \theta_2, \dots, \theta_k$.



Definizione 3.7.1 Dicesi funzione di verosimiglianza $L(\theta_1, \theta_2, \dots, \theta_k)$ di un campione casuale (X_1, X_2, \dots, X_n) la densità [distribuzione] di probabilità congiunta del campione:

$$\begin{aligned} L(\theta_1, \theta_2, \dots, \theta_k) &= f(x_1, x_2, \dots, x_n; \theta_1, \theta_2, \dots, \theta_k) \\ &= \prod_{i=1}^n f(x_i; \theta_1, \theta_2, \dots, \theta_k), \end{aligned} \quad (3.94)$$

riguardata come funzione dei parametri $\theta_1, \theta_2, \dots, \theta_k$.

La funzione di verosimiglianza può essere utilizzata per ottenere una stima puntuale degli incogniti parametri facendo ricorso a quello che qui chiameremo *postulato zero della statistica*. Questo può enunciarsi affermando che in assenza di specifiche indicazioni, tra tutte le possibili stime dei parametri va effettuata quella che rende massima la probabilità di occorrenza, o la densità di probabilità, della realizzazione che si sta osservando. In base a questo postulato appare dunque ragionevole scegliere come stima puntuale dei parametri incogniti i valori di $\theta_1, \theta_2, \dots, \theta_k$ che rendono massima la funzione di verosimiglianza $L(\theta_1, \theta_2, \dots, \theta_k)$ in corrispondenza della realizzazione (x_1, x_2, \dots, x_n) osservata, ossia che massimizzano ivi la densità [distribuzione] di probabilità congiunta del campione. Il metodo della massima verosimiglianza consiste dunque nel ricercare il massimo assoluto della funzione di verosimiglianza $L(\theta_1, \theta_2, \dots, \theta_k)$ riguardata come funzione dei soli parametri in quanto l' n -upla (x_1, x_2, \dots, x_n) viene considerata fissata. I valori dei parametri $\theta_1, \theta_2, \dots, \theta_k$ che la massimizzano sono detti *stime di massima verosimiglianza*; essi sono espressi come funzioni della realizzazione (x_1, x_2, \dots, x_n) e denotati con $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$. Le corrispondenti statistiche $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ sono dette *stimatori di massima verosimiglianza*. Più formalmente, si dà la seguente definizione:

Definizione 3.7.2 Siano $\hat{\theta}_j = g_j(x_1, x_2, \dots, x_n)$ ($j = 1, 2, \dots, k$) assegnate funzioni della generica realizzazione (x_1, x_2, \dots, x_n) del campione casuale (X_1, X_2, \dots, X_n) . I valori $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ corrispondenti alla realizzazione effettivamente osservata sono detti *stime di massima verosimiglianza dei parametri* $\theta_1, \theta_2, \dots, \theta_k$ se per ogni altra scelta $\theta'_1, \theta'_2, \dots, \theta'_k$ di questi risulta:

$$L(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k) \geq L(\theta'_1, \theta'_2, \dots, \theta'_k).$$

Le corrispondenti statistiche $\hat{\theta}_j = g_j(X_1, X_2, \dots, X_n)$ ($j = 1, 2, \dots, k$) sono dette *stimatori di massima verosimiglianza*.

Con un semplice esempio chiariamo il principio su cui poggia il metodo della massima verosimiglianza.

Esempio 3.7.1 Si vuole stabilire se una certa moneta è truccata. Supponiamo di non conoscere il valore esatto della probabilità θ che lanciando la moneta esca testa, ma di sapere che esso può essere $\theta_1 = 0.5$ oppure $\theta_2 = 0.4$. Effettuiamo n lanci indipendenti ed indichiamo con x_i l'esito del lancio i -esimo, ponendo $x_i = 1$ se esce testa e $x_i = 0$ se esce croce. Il campione (X_1, X_2, \dots, X_n) cui questo esperimento casuale si riferisce è dunque costituito da variabili casuali di Bernoulli; la funzione di verosimiglianza è pertanto:

$$L(\theta) = f(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i}$$

3.7. METODO DELLA MASSIMA VEROVEROSIMIGLIANZA

$$= \theta^{x_1 + \dots + x_n} (1 - \theta)^{n - (x_1 + \dots + x_n)}. \quad (3.95)$$

Supponiamo che su 100 lanci effettuati si siano ottenute 46 teste e 54 croci; la (3.95) assume allora i seguenti valori:

$$\begin{aligned} L(\theta_1) &= L(0.5) = (0.5)^{46} (0.5)^{54} = (0.5)^{100} = 7.88 \cdot 10^{-31} \\ L(\theta_2) &= L(0.4) = (0.4)^{46} (0.6)^{54} = 5.18 \cdot 10^{-31}. \end{aligned}$$

Essendo $L(\theta_1) > L(\theta_2)$, il metodo della massima verosimiglianza suggerisce di scegliere come stima di θ il valore θ_1 , ossia quello che massimizza la funzione di verosimiglianza $L(\theta)$. Si noti che tra i valori possibili di θ , ossia tra θ_1 e θ_2 , si sceglie θ_1 in quanto esso è il più plausibile nel senso che fornisce la "migliore" spiegazione dei dati osservati. Se θ fosse uguale a θ_2 la realizzazione osservata sarebbe invero meno probabile, e quindi la sua occorrenza sarebbe a priori meno giustificata. ♦

L'applicazione del metodo della massima verosimiglianza richiede di determinare il massimo della funzione di verosimiglianza (3.94); di solito, però, in luogo di massimizzare tale funzione è più agevole massimizzarne il logaritmo, ossia la funzione

$$\ln L(\theta_1, \theta_2, \dots, \theta_k) = \sum_{i=1}^n \ln f(x_i; \theta_1, \theta_2, \dots, \theta_k).$$

Ciò è lecito poiché il massimo assoluto di $L(\theta_1, \theta_2, \dots, \theta_k)$ coincide con il massimo assoluto di $\ln L(\theta_1, \theta_2, \dots, \theta_k)$.

Quando i parametri $\theta_1, \theta_2, \dots, \theta_k$ assumono valori in un continuo, il metodo della massima verosimiglianza richiede anzitutto la determinazione dei valori di $\theta_1, \theta_2, \dots, \theta_k$ tali da

$$\frac{\partial}{\partial \theta_i} \ln L(\theta_1, \theta_2, \dots, \theta_k) = 0 \quad (i = 1, 2, \dots, k);$$

si stabilisce poi se questi forniscano effettivamente un massimo relativo per $L(\theta_1, \theta_2, \dots, \theta_k)$. La ricerca del massimo si conclude esaminando i valori che assume $\ln L(\theta_1, \theta_2, \dots, \theta_k)$ sulla frontiera della regione di variabilità di $\theta_1, \theta_2, \dots, \theta_k$ ove il punto di massimo potrebbe giacere.

Il metodo della massima verosimiglianza testé descritto è largamente usato in problemi di stima puntuale dato che, sotto condizioni molto generali, esso conduce a estimatori sufficienti, asintoticamente corretti ed efficienti.

Esaminiamo alcuni esempi di applicazione del metodo della massima verosimiglianza.

Esempio 3.7.2 Si supponga che il campione casuale (X_1, X_2, \dots, X_n) sia stato estratto da una popolazione avente distribuzione gamma di parametri α noto e β incognito, e che di quest'ultimo si desideri determinare lo stivatore di massima verosimiglianza sulla base della realizzazione (x_1, x_2, \dots, x_n) osservata. Per $\beta > 0$ la funzione di verosimiglianza del campione è

$$\begin{aligned} L(\beta) &= \prod_{i=1}^n \frac{1}{\beta^\alpha \Gamma(\alpha)} x_i^{\alpha-1} e^{-x_i/\beta} \\ &= \left[\frac{1}{\beta^\alpha \Gamma(\alpha)} \right]^n (x_1 \cdots x_n)^{\alpha-1} e^{-(x_1 + \dots + x_n)/\beta}, \end{aligned}$$

per $x_1, x_2, \dots, x_n > 0$, mentre è nulla altrove. Si ha quindi:

$$\ln L(\beta) = -n[\alpha \ln \beta + \ln \Gamma(\alpha)] + (\alpha - 1) \ln(x_1 \cdots x_n) - \frac{x_1 + \cdots + x_n}{\beta},$$

da cui segue:

$$\frac{L'(\beta)}{L(\beta)} \equiv \frac{d}{d\beta} \ln L(\beta) = -n \frac{\alpha}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^n x_i.$$

Imponendo che la derivata si annulli, si ricava:

$$\beta = \frac{1}{n\alpha} \sum_{i=1}^n x_i \equiv \frac{\bar{x}}{\alpha}.$$

Questo è un punto di massimo relativo per $L(\beta)$ essendo $L'(\beta) > 0$ per $\beta < \bar{x}/\alpha$ e $L'(\beta) < 0$ per $\beta > \bar{x}/\alpha$. Ma poiché risulta

$$\lim_{\beta \rightarrow 0} \ln L(\beta) = \lim_{\beta \rightarrow \infty} \ln L(\beta) = -\infty,$$

si conclude che il massimo relativo ottenuto è in realtà un punto di massimo assoluto, che pertanto fornisce la stima di massima verosimiglianza di β . Lo stimatore di massima verosimiglianza del parametro β di una popolazione gamma, qualora α sia noto, è quindi:

$$\hat{\beta} = \frac{1}{n\alpha} \sum_{i=1}^n X_i \equiv \frac{\bar{X}}{\alpha}.$$

Si osservi che nel caso particolare $\alpha = 1$ la popolazione in esame diventa esponenziale di media β . Se ne conclude che la media campionaria è lo stimatore di massima verosimiglianza della media di una popolazione esponenziale. ♦

Esempio 3.7.3 In un esperimento consistente in k prove ripetute sia θ la probabilità di occorrenza di un dato evento A in ogni prova. Si supponga che l'esperimento venga ripetuto n volte e si denoti con x_i ($i = 1, 2, \dots, n$) il numero di occorrenze dell'evento A nell'esperimento i -esimo. Si costruisce così un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione avente distribuzione binomiale di parametri k e θ . Determiniamo lo stimatore di massima verosimiglianza del parametro θ sapendo che è stata osservata la realizzazione (x_1, x_2, \dots, x_n) . Per $0 \leq \theta \leq 1$ la funzione di verosimiglianza del campione è

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n \binom{k}{x_i} \theta^{x_i} (1-\theta)^{k-x_i} \\ &= \binom{k}{x_1} \cdots \binom{k}{x_n} \theta^{x_1+x_2+\cdots+x_n} (1-\theta)^{nk-(x_1+x_2+\cdots+x_n)}, \end{aligned} \quad (3.96)$$

dove ciascuna delle x_i può assumere i valori $0, 1, \dots, k$. Indicando per brevità con C il prodotto dei coefficienti binomiali che compare nella (3.96), si ha:

$$\begin{aligned} \ln L(\theta) &= \ln C + (x_1 + x_2 + \cdots + x_n) \ln \theta \\ &\quad + [nk - (x_1 + x_2 + \cdots + x_n)] \ln(1-\theta), \end{aligned}$$

da cui segue:

$$\begin{aligned} \frac{L'(\theta)}{L(\theta)} &\equiv \frac{d}{d\theta} \ln L(\theta) = \frac{x_1 + x_2 + \cdots + x_n}{\theta} - \frac{nk - (x_1 + x_2 + \cdots + x_n)}{1-\theta} \\ &= \frac{(1-\theta)(x_1 + x_2 + \cdots + x_n) - nk\theta + \theta(x_1 + x_2 + \cdots + x_n)}{\theta(1-\theta)} \\ &= \frac{(x_1 + x_2 + \cdots + x_n) - nk\theta}{\theta(1-\theta)}. \end{aligned} \quad (3.97)$$

Imponendo che tale derivata si annulli, si ottiene:

$$\theta = \frac{1}{nk} \sum_{i=1}^n x_i. \quad (3.98)$$

Osserviamo inoltre che risulta:

$$\begin{aligned} L'(\theta) &> 0 \quad \text{per } \theta < \frac{1}{nk} \sum_{i=1}^n x_i, \\ L'(\theta) &< 0 \quad \text{per } \theta > \frac{1}{nk} \sum_{i=1}^n x_i. \end{aligned}$$

La (3.98) rappresenta quindi un punto di massimo relativo per $L(\theta)$. Poiché

$$\lim_{\theta \rightarrow 0} \ln L(\theta) = \lim_{\theta \rightarrow 1} \ln L(\theta) = -\infty,$$

questo è in realtà un massimo assoluto, così che la (3.98) è la stima di massima verosimiglianza di θ . Il corrispondente stimatore di massima verosimiglianza è quindi:

$$\hat{\theta} = \frac{1}{nk} \sum_{i=1}^n X_i \equiv \frac{\bar{x}}{k}. \quad (3.99)$$

Poiché $\sum_{i=1}^n x_i$ rappresenta la frequenza assoluta dell'evento A negli n esperimenti ed essendo nk il numero totale di prove, la (3.98) può essere interpretata come la frequenza relativa dell'evento A riscontrata nella serie di prove ripetute. D'altra parte x_i/k rappresenta la frequenza relativa di A nell'esperimento i -esimo, così che la (3.98) può essere riguardata anche come la frequenza media dell'evento A negli n esperimenti.

Si noti che nel caso particolare $k = 1$ la popolazione in esame è costituita da variabili di Bernoulli di parametro θ . In tal caso dalla (3.99) discende che la media campionaria è lo stimatore di massima verosimiglianza del parametro θ di una popolazione di Bernoulli, ed anche del parametro θ di una distribuzione binomiale.

È opportuno infine notare che quanto qui discusso è suscettibile di un utile generalizzazione che passiamo ad illustrare. Supponiamo di effettuare ancora n esperimenti, ciascuno consistente in prove ripetute, e sia θ la probabilità di occorrenza di un dato evento A in ogni prova. A differenza del caso precedente supponiamo, però, che il numero di prove in ogni esperimento sia variabile; indichiamo quindi con k_i il numero di prove relative all'esperimento i -esimo, mentre con x_i denoteremo ancora il numero di occorrenze

dell'evento A nell'esperimento i -esimo ($i = 1, 2, \dots, n$). Le variabili casuali indipendenti X_1, X_2, \dots, X_n stavolta non costituiscono un campione casuale in quanto non sono identicamente distribuite. Infatti la generica variabile X_i ha distribuzione binomiale di parametri k_i e θ ($i = 1, 2, \dots, n$). Ciononostante è possibile fare ricorso al principio che sottende il metodo della massima verosimiglianza per ricavare uno stimatore del parametro θ poiché $L(\theta) \equiv P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n; \theta)$ è ancora interpretabile alla stregua di una funzione di verosimiglianza. Per $0 \leq \theta \leq 1$ risulta:

$$\begin{aligned} L(\theta) &= P(X_i = x_i; \theta) = \prod_{i=1}^n \binom{k_i}{x_i} \theta^{x_i} (1-\theta)^{k_i-x_i} \\ &= \binom{k_1}{x_1} \cdots \binom{k_n}{x_n} \theta^{x_1+x_2+\cdots+x_n} (1-\theta)^{K-(x_1+x_2+\cdots+x_n)}, \end{aligned}$$

dove ciascuna delle x_i può assumere i valori $0, 1, \dots, k_i$, e dove si è posto $K = k_1 + k_2 + \cdots + k_n$. Come si vede facilmente, la derivata rispetto a θ della funzione $\ln L(\theta)$ coincide con quella del caso precedente, ossia con la (3.97) a patto di porvi K al posto di nk . Procedendo in modo del tutto analogo si ricava infine lo stimatore di massima verosimiglianza di θ :

$$\hat{\theta} = \frac{1}{K} \sum_{i=1}^n X_i \equiv \frac{X_1 + X_2 + \cdots + X_n}{k_1 + k_2 + \cdots + k_n}.$$

Esempio 3.7.4 Sia (X_1, X_2, \dots, X_n) un campione estratto da una popolazione distribuita uniformemente in $[0, \theta]$ e sia (x_1, x_2, \dots, x_n) una sua fissata realizzazione, avendosi evidentemente $x_i \in [0, \theta]$ per $i = 1, 2, \dots, n$. Si desidera determinare lo stimatore di massima verosimiglianza del parametro positivo θ . La funzione di verosimiglianza del campione è la seguente:

$$L(\theta) = \prod_{i=1}^n \frac{1}{\theta} = \frac{1}{\theta^n}.$$

Per determinarne il massimo, osserviamo che essa è una funzione decrescente per $\theta > 0$. Inoltre, per ipotesi si ha $\theta \geq x_1, x_2, \dots, x_n$, non potendo θ essere inferiore ad alcuno degli elementi del campione. La funzione $L(\theta)$ assume pertanto il suo massimo (assoluto) quando θ è pari al massimo dei valori del campione casuale. Se poniamo $x_{(n)} = \max\{x_1, x_2, \dots, x_n\}$, la funzione di verosimiglianza $L(\theta)$, sotto il vincolo $\theta \geq x_1, x_2, \dots, x_n$, è pertanto massima per $\theta = x_{(n)}$. Si ricava così che lo stimatore desiderato si identifica con la n -esima statistica d'ordine del campione casuale:

$$\hat{\theta} \equiv X_{(n)} = \max\{X_1, X_2, \dots, X_n\}. \quad (3.100)$$

Ricordando la seconda delle (3.5), si ricava facilmente la densità di $\hat{\theta}$:

$$f_{\hat{\theta}}(x) = \begin{cases} \frac{n}{\theta} \left(\frac{x}{\theta}\right)^{n-1} & \text{per } 0 < x < \theta, \\ 0 & \text{altrimenti.} \end{cases} \quad (3.101)$$

Il valore medio di $\hat{\theta}$ è quindi:

$$E(\hat{\theta}) = \int_0^\theta x f_{\hat{\theta}}(x) dx = \int_0^\theta x \frac{n}{\theta} \left(\frac{x}{\theta}\right)^{n-1} dx = \frac{n}{n+1} \theta. \quad (3.102)$$

Lo stimatore di massima verosimiglianza $\hat{\theta}$, che nell'Esempio 3.5.8 si è mostrato essere una statistica sufficiente per la stima di θ , non è dunque corretto, ma è asintoticamente corretto, così da risultare idoneo per la stima di θ solo nel caso di campioni di taglia elevata. ♦

Con riferimento all'Esempio 3.7.4 va rilevato come a partire da uno stimatore di massima verosimiglianza non corretto sia possibile pervenire a un nuovo stimatore che, oltre alla correttezza, esibisce anche altre caratteristiche tali da renderlo preferibile al precedente. A chiarimento di questa affermazione forniamo l'esempio che segue.

Esempio 3.7.5 Sia (X_1, X_2, \dots, X_n) il campione casuale di cui all'Esempio 3.7.4. Osserviamo che se si pone

$$\hat{\theta}' = \frac{n+1}{n} \hat{\theta} \equiv \frac{n+1}{n} \max\{X_1, X_2, \dots, X_n\}, \quad (3.103)$$

dalla (3.102) si ha:

$$E(\hat{\theta}') \equiv \frac{n+1}{n} E(\hat{\theta}) = \theta,$$

così che $\hat{\theta}'$ è uno stimatore corretto del parametro θ . Mostriamo che lo stimatore $\hat{\theta}'$ è da preferirsi allo stimatore asintoticamente corretto $\hat{\theta}$ dato dalla (3.100) mediante confronto dei rispettivi errori quadratici medi. Facendo uso della (3.101) si ottiene l'errore quadratico medio di $\hat{\theta}$:

$$\begin{aligned} \text{mse}(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] = \int_0^\theta (x - \theta)^2 \frac{n}{\theta} \left(\frac{x}{\theta}\right)^{n-1} dx \\ &= \frac{n}{\theta^n} \int_0^\theta (x^{n+1} - 2\theta x^n + \theta^2 x^{n-1}) dx \\ &= n\theta^2 \left(\frac{1}{n+2} - \frac{2}{n+1} + \frac{1}{n} \right) \\ &= \frac{2\theta^2}{(n+1)(n+2)}. \end{aligned} \quad (3.104)$$

Ricordando la (3.103) si ricava poi l'errore quadratico medio di $\hat{\theta}'$:

$$\begin{aligned} \text{mse}(\hat{\theta}') &= E \left[\left(\frac{n+1}{n} \hat{\theta} - \theta \right)^2 \right] = \int_0^\theta \left(\frac{n+1}{n} x - \theta \right)^2 \frac{n}{\theta} \left(\frac{x}{\theta}\right)^{n-1} dx \\ &= \int_0^\theta \left[\frac{(n+1)^2}{n^2} x^2 - 2\theta \frac{n+1}{n} x + \theta^2 \right] \frac{n}{\theta} \left(\frac{x}{\theta}\right)^{n-1} dx \\ &= \theta^2 \left[\frac{(n+1)^2}{n(n+2)} - 1 \right] = \frac{\theta^2}{n(n+2)}. \end{aligned} \quad (3.105)$$

Dalle (3.104) e (3.105) segue che $\text{mse}(\hat{\Theta}') \leq \text{mse}(\hat{\Theta})$ se e solo se

$$\frac{\theta^2}{n(n+2)} \leq \frac{2\theta^2}{(n+1)(n+2)}. \quad (3.106)$$

Poiché la (3.106) è soddisfatta per ogni $\theta > 0$ e per ogni $n \geq 1$, lo stimatore corretto $\hat{\Theta}'$ è sempre da preferirsi allo stimatore di massima verosimiglianza $\hat{\Theta}$.

Allo scopo di ottenere una conferma di tale conclusione, confrontiamo i valori assoluti degli errori relativi degli stimatori $\hat{\Theta}$ e $\hat{\Theta}'$. Definiamo pertanto le statistiche

$$U_n = \frac{|\theta - \hat{\Theta}|}{\theta} \equiv \frac{\theta - \hat{\Theta}}{\theta}, \quad U'_n = \frac{|\theta - \hat{\Theta}'|}{\theta},$$

rappresentanti i valori assoluti degli errori relativi che si commettono quando si utilizzano $\hat{\Theta}$ e $\hat{\Theta}'$ per stimare θ . Si noti che il segno di identità nella definizione di U_n è conseguenza della (3.100) e dall'essere $P(X_i \leq \theta) = 1$ ($i = 1, 2, \dots, n$). La funzione di distribuzione di U_n è data da

$$F_{U_n}(x) = P\left(\frac{\theta - \hat{\Theta}}{\theta} \leq x\right) = P[\hat{\Theta} \geq \theta(1-x)]$$

$$= 1 - F_{\hat{\Theta}}[\theta(1-x)] = \begin{cases} 0 & \text{per } x < 0, \\ 1 - (1-x)^n & \text{per } 0 \leq x < 1, \\ 1 & \text{per } x \geq 1. \end{cases}$$

Ricordando l'espressione (3.102) della media di $\hat{\Theta}$, si ricava poi:

$$E(U_n) = \frac{\theta - E(\hat{\Theta})}{\theta} = \frac{1}{n+1}. \quad (3.107)$$

Se si utilizza lo stimatore $\hat{\Theta}$ si commette dunque un errore relativo il cui valore assoluto è in media $1/(n+1)$. Ciò conferma che conviene fare uso di questo stimatore solo se la taglia n del campione è elevata. Tale affermazione viene confortata dall'analisi della funzione di distribuzione di U_n , come evidenziato dalla Tabella 3.1 in cui, a titolo di esempio, sono riportati i valori di $F_{U_n}(0.01) \equiv P(U_n \leq 0.01)$ e di $F_{U_n}(0.05) \equiv P(U_n \leq 0.05)$ in corrispondenza di talune scelte di n : è evidente come questi valori si avvicinino all'unità al crescere di n .

Esaminiamo ora U'_n ; la sua funzione di distribuzione è data da

$$F_{U'_n}(x) = P\left(\frac{|\theta - \hat{\Theta}'|}{\theta} \leq x\right) = P[\theta(1-x) \leq \hat{\Theta}' \leq \theta(1+x)]$$

$$= P\left[\theta(1-x)\frac{n}{n+1} \leq \hat{\Theta} \leq \theta(1+x)\frac{n}{n+1}\right]$$

$$= F_{\hat{\Theta}}\left[\theta(1+x)\frac{n}{n+1}\right] - F_{\hat{\Theta}}\left[\theta(1-x)\frac{n}{n+1}\right],$$

3.7. METODO DELLA MASSIMA VEROVEROSIMIGLIANZA

Tabella 3.1: Alcuni valori delle funzioni di distribuzione degli errori U_n e U'_n .

n	$F_{U_n}(0.01)$	$F_{U_n}(0.05)$	$F_{U'_n}(0.01)$	$F_{U'_n}(0.05)$
10	0.0956	0.4013	0.0772	0.3971
20	0.1821	0.6415	0.1515	0.8649
30	0.2603	0.7854	0.2274	0.9197
40	0.3310	0.8715	0.3055	0.9521
50	0.3949	0.9230	0.3858	0.9714
60	0.4528	0.9539	0.4712	0.9829
70	0.5051	0.9724	0.5604	0.9898
80	0.5525	0.9835	0.6549	0.9939
90	0.5953	0.9901	0.7560	0.9963
100	0.6339	0.9941	0.8646	0.9978

ossia:

$$F_{U'_n}(x) = \begin{cases} 0 & \text{per } x < 0, \\ \left[(1+x)\frac{n}{n+1}\right]^n - \left[(1-x)\frac{n}{n+1}\right]^n & \text{per } 0 \leq x < \frac{1}{n}, \\ 1 - \left[(1-x)\frac{n}{n+1}\right]^n & \text{per } \frac{1}{n} \leq x < 1, \\ 1 & \text{per } x \geq 1, \end{cases}$$

dalla quale si ottiene il valore medio¹⁰

$$\begin{aligned} E(U'_n) &= \int_0^1 [1 - F_{U'_n}(x)] dx \\ &= \int_0^{1/n} \left\{1 - \left[(1+x)\frac{n}{n+1}\right]^n\right\} dx + \int_0^1 \left[(1-x)\frac{n}{n+1}\right]^n dx \\ &= \frac{2}{n} \left(\frac{n}{n+1}\right)^{n+1}. \end{aligned} \quad (3.108)$$

Osserviamo che U'_n è in media minore di U_n ; invero, dalla (3.107) e dalla (3.108) segue $E(U'_n) \leq E(U_n)$ se e solo se

$$\left(1 + \frac{1}{n}\right) \left(1 - \frac{1}{n+1}\right)^{n+1} \leq \frac{1}{2};$$

¹⁰Si ricorda che se X è una variabile casuale di funzione di distribuzione $F(x)$, dotata di valore medio, si ha:

$$E(X) \equiv \int_{-\infty}^{\infty} x dF(x) = - \int_{-\infty}^0 F(x) dx + \int_0^{\infty} [1 - F(x)] dx.$$

ma questa relazione è sempre verificata poiché il primo membro è strettamente decrescente in n , vale $1/2$ per $n = 1$ e tende a $1/e$ per $n \rightarrow \infty$. Per un più completo confronto degli stimatori $\hat{\theta}$ e $\hat{\theta}'$ di θ è opportuno paragonare anche le funzioni di distribuzione delle statistiche U_n e U'_n corrispondenti. A tal fine, nella Tabella 3.1 sono stati riportati anche $F_{U'_n}(0.01) \equiv P(U'_n \leq 0.01)$ e $F_{U'_n}(0.05) \equiv P(U'_n \leq 0.05)$ in corrispondenza dei medesimi valori di n . È evidente che al crescere di n queste probabilità tendono ad 1 più rapidamente delle corrispondenti probabilità relative a U_n . Ciò conferma ulteriormente che per la stima di θ lo stimatore $\hat{\theta}'$ è da preferirsi a $\hat{\theta}$. ♦

Esempio 3.7.6 Consideriamo un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione distribuita alla Poisson con parametru θ e determiniamone lo stimatore di massima verosimiglianza. Per $\theta > 0$ la funzione di verosimiglianza è data da

$$L(\theta) = \prod_{i=1}^n e^{-\theta} \frac{\theta^{x_i}}{x_i!} \triangleq \frac{1}{A} e^{-n\theta} \theta^{x_1+x_2+\dots+x_n},$$

dove ciascuna delle x_i può assumere i valori $0, 1, 2, \dots$ e dove A denota il prodotto $x_1! \cdots x_n!$. Passando ai logaritmi si ha:

$$\ln L(\theta) = -\ln A - n\theta + \left(\sum_{i=1}^n x_i \right) \ln \theta,$$

da cui segue:

$$\frac{L'(\theta)}{L(\theta)} \equiv \frac{d}{d\theta} \ln L(\theta) = -n + \frac{1}{\theta} \sum_{i=1}^n x_i.$$

Imponendo l'annullarsi di $L'(\theta)$, si ricava:

$$\theta = \frac{1}{n} \sum_{i=1}^n x_i. \quad (3.109)$$

Osservando poi che risulta

$$\frac{d^2}{d\theta^2} \ln L(\theta) = -\frac{1}{\theta^2} \sum_{i=1}^n x_i \leq 0,$$

si conclude che la (3.109) rappresenta un punto di massimo relativo per $L(\theta)$. Poiché

$$\lim_{\theta \rightarrow 0} \ln L(\theta) = \lim_{\theta \rightarrow \infty} \ln L(\theta) = -\infty,$$

la (3.109) fornisce in realtà un punto di massimo assoluto per $L(\theta)$. Si conclude così che la media campionaria

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i \equiv \bar{X}$$

è lo stimatore di massima verosimiglianza della media θ della distribuzione di Poisson. ♦

3.7. METODO DELLA MASSIMA VERO SIMIGLIANZA

Esempio 3.7.7 Sia dato un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione normale di valore medio μ e varianza σ^2 . Si desidera determinare gli stimatori di massima verosimiglianza dei parametri μ e σ^2 . La funzione di verosimiglianza per $\mu \in \mathbb{R}$ e $\sigma^2 > 0$ è data da

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(x_i - \mu)^2}{2\sigma^2} \right] \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right], \end{aligned}$$

dove $x_1, x_2, \dots, x_n \in \mathbb{R}$. Passando ai logaritmi si ottiene:

$$\ln L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

Allo scopo di determinare il massimo di $L(\mu, \sigma^2)$ calcoliamo le derivate parziali di $\ln L(\mu, \sigma^2)$ rispetto a μ e rispetto a σ^2 . Si ottiene così:

$$\frac{\partial}{\partial \mu} \ln L(\mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu), \quad (3.110)$$

$$\frac{\partial}{\partial \sigma^2} \ln L(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2. \quad (3.111)$$

Imponendo che la derivata (3.110) sia nulla si ricava:

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i \equiv \bar{x}, \quad (3.112)$$

dove \bar{x} è la media aritmetica dei valori x_1, x_2, \dots, x_n costituenti la realizzazione del campione. Facendo poi uso della (3.112) nella (3.111) e imponendo che la derivata ivi presente si annulli si trae:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad (3.113)$$

che fornisce il cosiddetto *scarto quadratico medio* della realizzazione del campione. Dalle (3.110) e (3.111) si ha poi:

$$\begin{aligned} \frac{\partial^2}{\partial \mu^2} \ln L(\mu, \sigma^2) &= -\frac{n}{\sigma^2} \\ \frac{\partial^2}{\partial \mu \partial \sigma^2} \ln L(\mu, \sigma^2) &= -\frac{1}{\sigma^4} \sum_{i=1}^n (x_i - \mu) \\ \frac{\partial^2}{\partial (\sigma^2)^2} \ln L(\mu, \sigma^2) &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n (x_i - \mu)^2. \end{aligned}$$

Indicando con $\hat{\mu}$ e $\hat{\sigma}^2$ le stime (3.112) e (3.113), ossia ponendo

$$\hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

si ottiene così:

$$\begin{aligned}\frac{\partial^2}{\partial \mu^2} \ln L(\mu, \sigma^2) \Big|_{\mu=\hat{\mu}, \sigma^2=\hat{\sigma}^2} &= -\frac{n}{\hat{\sigma}^2} < 0 \\ \frac{\partial^2}{\partial \mu \partial \sigma^2} \ln L(\mu, \sigma^2) \Big|_{\mu=\hat{\mu}, \sigma^2=\hat{\sigma}^2} &= 0 \\ \frac{\partial^2}{\partial (\sigma^2)^2} \ln L(\mu, \sigma^2) \Big|_{\mu=\hat{\mu}, \sigma^2=\hat{\sigma}^2} &= -\frac{n}{2\hat{\sigma}^4} < 0.\end{aligned}$$

La matrice hessiana di $\ln L(\mu, \sigma^2)$ ha autovalori entrambi negativi, cosicché il punto determinato è un massimo relativo. In realtà si tratta di un massimo assoluto in quanto la funzione $\ln L(\mu, \sigma^2)$ diverge negativamente sulla frontiera della regione $\{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$ di variabilità dei parametri. Pertanto lo stimatore di massima verosimiglianza del valore medio μ di una popolazione normale è la media campionaria \bar{X} . Lo stimatore di massima verosimiglianza della varianza σ^2 è dato da

$$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \equiv \frac{n-1}{n} S^2;$$

a quest'ultimo si dà il nome di *scarto quadratico medio campionario*. Si noti che qualunque sia la popolazione in esame, a differenza della media campionaria che è uno stimatore corretto della media, lo scarto quadratico medio campionario non è uno stimatore corretto di σ^2 , ma asintoticamente corretto; invero, essendo $E(S^2) = \sigma^2$, si ha

$$E\left(\frac{n-1}{n} S^2\right) = \frac{n-1}{n} \sigma^2,$$

che tende a σ^2 per $n \rightarrow \infty$.

Esempio 3.7.8 Consideriamo un campione (X_1, X_2, \dots, X_n) estratto da una popolazione normale inversa di parametri μ e λ . Allo scopo di determinare gli stimatori di massima verosimiglianza di μ e λ , esplicitiamo la funzione di verosimiglianza $L(\mu, \lambda)$. Poiché (cfr. § 2.5) la densità di probabilità di una variabile normale inversa è data da

$$f(x) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right] \quad (x > 0),$$

per $\mu > 0$ e $\lambda > 0$ risulta:

$$L(\mu, \lambda) = \exp\left[-\frac{\lambda}{2\mu^2} \sum_{i=1}^n \frac{(x_i - \mu)^2}{x_i}\right] \prod_{i=1}^n \left(\frac{\lambda}{2\pi x_i^3}\right)^{1/2},$$

con $x_1, x_2, \dots, x_n > 0$. Per determinare il massimo di $L(\mu, \lambda)$ calcoliamo le derivate parziali della funzione

$$\ln L(\mu, \lambda) = -\frac{\lambda}{2} \sum_{i=1}^n \frac{1}{x_i} \left(\frac{x_i}{\mu} - 1\right)^2 + \frac{1}{2} \sum_{i=1}^n \ln\left(\frac{\lambda}{2\pi x_i^3}\right)$$

3.7. METODO DELLA MASSIMA VERO SIMIGLIANZA

rispetto a μ e rispetto a λ ed imponiamone l'annullarsi. Si ha dunque anzitutto:

$$\frac{\partial}{\partial \mu} \ln L(\mu, \lambda) = \frac{\lambda}{\mu^3} \sum_{i=1}^n (x_i - \mu) = 0,$$

così che

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i \equiv \bar{x}. \quad (3.114)$$

Imponendo inoltre

$$\frac{\partial}{\partial \lambda} \ln L(\mu, \lambda) = -\frac{1}{2} \sum_{i=1}^n \frac{1}{x_i} \left(\frac{x_i}{\mu} - 1\right)^2 + \frac{n}{2\lambda} = 0, \quad (3.115)$$

e facendo uso della (3.114), si ottiene:

$$\lambda = n \left[\sum_{i=1}^n \frac{1}{x_i} \left(\frac{x_i}{\mu} - 1\right)^2 \right]_{\mu=\bar{x}}^{-1}.$$

Osservando poi che

$$\left[\sum_{i=1}^n \frac{1}{x_i} \left(\frac{x_i}{\mu} - 1\right)^2 \right]_{\mu=\bar{x}} = \sum_{i=1}^n \frac{1}{x_i} + \frac{1}{\bar{x}^2} \sum_{i=1}^n x_i - 2 \frac{n}{\bar{x}} = \sum_{i=1}^n \left(\frac{1}{x_i} - \frac{1}{\bar{x}}\right),$$

dalla (3.115) si trae:

$$\lambda = n \left[\sum_{i=1}^n \left(\frac{1}{x_i} - \frac{1}{\bar{x}}\right) \right]^{-1}. \quad (3.116)$$

Calcoliamo le derivate seconde:

$$\begin{aligned}\frac{\partial^2}{\partial \mu^2} \ln L(\mu, \lambda) &= -\frac{\lambda}{\mu^3} \left[n + \frac{3}{\mu} \sum_{i=1}^n (x_i - \mu)\right] \\ \frac{\partial^2}{\partial \mu \partial \lambda} \ln L(\mu, \lambda) &= \frac{1}{\mu^3} \sum_{i=1}^n (x_i - \mu) \\ \frac{\partial^2}{\partial \lambda^2} \ln L(\mu, \lambda) &= -\frac{n}{2\lambda^2}.\end{aligned}$$

Indicando con $\hat{\mu}$ e $\hat{\lambda}$ le stime (3.114) e (3.116), ossia ponendo

$$\hat{\mu} = \bar{x}, \quad \hat{\lambda} = n \left[\sum_{i=1}^n \left(\frac{1}{x_i} - \frac{1}{\bar{x}}\right) \right]^{-1},$$

si ottiene:

$$\frac{\partial^2}{\partial \mu^2} \ln L(\mu, \lambda) \Big|_{\mu=\hat{\mu}, \lambda=\hat{\lambda}} = -\frac{\hat{\lambda}}{\hat{\mu}^3} n < 0$$

$$\begin{aligned}\frac{\partial^2}{\partial \mu \partial \lambda} \ln L(\mu, \lambda) \Big|_{\mu=\hat{\mu}, \lambda=\hat{\lambda}} &= 0 \\ \frac{\partial^2}{\partial \lambda^2} \ln L(\mu, \lambda) \Big|_{\mu=\hat{\mu}, \lambda=\hat{\lambda}} &= -\frac{n}{2\hat{\lambda}^2} < 0.\end{aligned}$$

La matrice hessiana di $\ln L(\mu, \lambda)$ ha autovalori entrambi negativi, così che il punto di coordinate $(\hat{\mu}, \hat{\lambda})$ così determinato è un massimo relativo. In realtà trattasi di un massimo assoluto dal momento che la funzione $\ln L(\mu, \lambda)$ diverge negativamente sulla frontiera della regione $\{(\mu, \lambda) : \mu > 0, \lambda > 0\}$ di variabilità dei parametri. Se ne deduce che la media campionaria \bar{X} è lo stimatore di massima verosimiglianza del parametro μ di una popolazione normale inversa, mentre la statistica

$$n \left[\sum_{i=1}^n \left(\frac{1}{X_i} - \frac{1}{\bar{X}} \right) \right]^{-1}$$

è lo stimatore di massima verosimiglianza del parametro λ . ◆

È opportuno segnalare a questo punto che non mancano casi in cui il metodo della massima verosimiglianza fallisce, come indica il seguente esempio.

Esempio 3.7.9 Sia Z una variabile casuale normale standard e siano $f_Z(x)$ e $F_Z(x)$ le corrispondenti densità di probabilità e funzione di distribuzione. Si consideri ora la cosiddetta variabile casuale *normale asimmetrica* X , la cui densità di probabilità è così definita:

$$f_X(x; \theta) = 2 f_Z(x) F_Z(\theta x), \quad x \in \mathbb{R}, \quad (3.117)$$

dove θ è un parametro reale. Riguardando X come campione casuale di taglia unitaria, e denotando con x una sua realizzazione, ricerchiamo lo stimatore di massima verosimiglianza di θ . Ricordando le (3.94) e (3.117), la funzione di verosimiglianza è data da

$$L(\theta) = 2 f_Z(x) F_Z(\theta x). \quad (3.118)$$

Dalla (3.118) si evince che se $x = 0$ si ha $L(\theta) = 1/\sqrt{2\pi}$, così che la funzione di verosimiglianza non dipende da θ . Se, invece, x è positivo [negativo] $L(\theta)$ è strettamente crescente [decrescente] in θ e risulta

$$\lim_{\theta \rightarrow \infty} L(\theta) = 2 f_Z(x) \quad \left[\lim_{\theta \rightarrow -\infty} L(\theta) = 2 f_Z(x) \right].$$

In base alla Definizione 3.7.2 si conclude che non è possibile attribuire una stima di massima verosimiglianza a θ in quanto non è possibile individuare una funzione reale $\hat{\theta} = g(x)$, dipendente dal generico valore x osservato, tale da aversi $L(\hat{\theta}) \geq L(\theta')$ per ogni scelta di θ' . ◆

Esamineremo ora alcuni aspetti particolarmente significativi del metodo della massima verosimiglianza, evidenziando le principali proprietà degli stimatori che ne discendono.

3.7. METODO DELLA MASSIMA VEROsimiglianza

Proposizione 3.7.1 Sia $\hat{\theta}$ lo stimatore di massima verosimiglianza di un parametro θ . Se $\lambda = u(\theta)$ è una funzione iniettiva,¹¹ $u(\hat{\theta})$ risulta essere lo stimatore di massima verosimiglianza di $u(\theta)$.

Dim. Per ipotesi la funzione di verosimiglianza $L(\theta)$ è massima-in- $\hat{\theta}$, dove $\hat{\theta}$ è la stima di massima verosimiglianza di θ ottenuta, per ipotesi, mediante lo stimatore $\hat{\theta}$. Per l'ipotesi di iniettività di u , la funzione di verosimiglianza $L^*(\lambda) \stackrel{\text{def}}{=} L[u^{-1}(\lambda)]$, relativa al campione estratto dalla popolazione di parametro incognito $\lambda = u(\theta)$, è massima in quell'unico punto $\hat{\lambda}$ tale da aversi $\hat{\lambda} = u(\hat{\theta})$. Quindi, lo stimatore di massima verosimiglianza di $u(\theta)$ è dato da $u(\hat{\theta})$. ■

Esempio 3.7.10 Sia dato un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione normale di valore medio μ noto e varianza σ^2 incognita. Con un procedimento perfettamente analogo a quello adottato nell'Esempio 3.7.7, è facile ricavare che lo stimatore di massima verosimiglianza di σ^2 è dato da

$$\hat{\Sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2.$$

Facendo uso della Proposizione 3.7.1 si ricava poi lo stimatore di massima verosimiglianza della deviazione standard σ :

$$\hat{\Sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2}. \quad \diamond$$

Proposizione 3.7.2 Se $T = g(X_1, X_2, \dots, X_n)$ è una statistica sufficiente per la stima di un parametro θ , lo stimatore di massima verosimiglianza di θ , se unico, è una funzione di T .

Dim. Essendo T una statistica sufficiente, dal Teorema 3.5.1 segue:

$$L(\theta) \equiv f(x_1, x_2, \dots, x_n; \theta) = k[g(x_1, x_2, \dots, x_n), \theta] h(x_1, x_2, \dots, x_n).$$

Massimizzare $L(\theta)$ equivale quindi a massimizzare $k[g(x_1, x_2, \dots, x_n), \theta]$. Il valore di θ che massimizza $k[g(x_1, x_2, \dots, x_n), \theta]$ è pertanto una funzione di $g(x_1, x_2, \dots, x_n)$. Lo stimatore di massima verosimiglianza di θ è quindi una funzione della statistica $T = g(X_1, X_2, \dots, X_n)$. ■

Proposizione 3.7.3 Se esistente, uno stimatore corretto e pienamente efficiente di un parametro θ è uno stimatore di massima verosimiglianza.

Dim. Sia $\hat{\theta} = g(X_1, X_2, \dots, X_n)$ uno stimatore corretto e pienamente efficiente di θ . Per il Teorema 3.3.1 esiste allora una funzione $\alpha(\theta)$ tale che per ogni θ risulta:

$$g(X_1, X_2, \dots, X_n) - \theta = \alpha(\theta) \sum_{j=1}^n \frac{\partial}{\partial \theta} \ln f(X_j; \theta),$$

¹¹Ricordiamo che una funzione $u(\cdot)$ si dice iniettiva se $\theta_1 \neq \theta_2$ implica $u(\theta_1) \neq u(\theta_2)$.

così che la derivata logaritmica della funzione di verosimiglianza fornisce:

$$\frac{\partial}{\partial \theta} \ln L(\theta) \equiv \sum_{j=1}^n \frac{\partial}{\partial \theta} \ln f(x_j; \theta) = \frac{1}{\alpha(\theta)} [g(x_1, x_2, \dots, x_n) - \theta].$$

Imponendone l'annullarsi, si ricava $\hat{\theta} = g(x_1, x_2, \dots, x_n)$. Lo stimatore di massima verosimiglianza, se esiste, coincide quindi con lo stimatore corretto e pienamente efficiente $\hat{\theta} = g(X_1, X_2, \dots, X_n)$. ■

Ricaveremo ora un'interessante caratterizzazione della distribuzione normale facendo uso del principio della massima verosimiglianza.

Proposizione 3.7.4 *Sia (X_1, X_2, \dots, X_n) un campione estratto da una popolazione continua di valore medio μ , e sia $L(\mu) = \prod_{i=1}^n f(x_i - \mu)$ la sua funzione di verosimiglianza. Supponiamo che (i) f è una funzione pari, (ii) f è continua e derivabile, (iii) f ha per supporto l'insieme dei numeri reali.¹² Se la media campionaria è lo stimatore di massima verosimiglianza di μ , allora la popolazione ha distribuzione normale.*

Dim. Per ipotesi la media campionaria è lo stimatore di massima verosimiglianza del valore medio μ , cosicché la funzione di verosimiglianza $L(\mu) = \prod_{i=1}^n f(x_i - \mu)$ è dotata di massimo per $\mu = \bar{x}$. Di conseguenza anche la funzione

$$\ln L(\mu) = \ln \prod_{i=1}^n f(x_i - \mu) = \sum_{i=1}^n \ln f(x_i - \mu)$$

è dotata di massimo per $\mu = \bar{x}$. Ponendo $\varphi = \ln f$ e derivando rispetto a μ si ha allora:

$$\sum_{i=1}^n \varphi'(x_i - \mu) \Big|_{\mu=\bar{x}} = 0. \quad (3.119)$$

In virtù dell'arbitrarietà della realizzazione (x_1, x_2, \dots, x_n) , possiamo scegliere $x_1 = x_2 = \dots = x_{n-1} = y$ e lasciare x_n arbitrario. Così, risulta

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{n-1}{n} y + \frac{1}{n} x_n,$$

e quindi:

$$x_n - \bar{x} = (n-1)(\bar{x} - y),$$

di modo che la (3.119) diventa:

$$\sum_{i=1}^{n-1} \varphi'(x_i - \bar{x}) + \varphi'(x_n - \bar{x}) = \sum_{i=1}^{n-1} \varphi'(x_i - \bar{x}) + \varphi'[(n-1)(\bar{x} - y)] = 0,$$

ossia:

$$(n-1) \varphi'(\bar{x} - y) = -\varphi'[(n-1)(\bar{x} - y)].$$

¹²Ciò significa che risulta $f(x) > 0$ per ogni $x \in \mathbb{R}$.

3.7. METODO DELLA MASSIMA VEROsimiglianza

Poiché per ipotesi f è una funzione pari, tale è anche φ , e quindi φ' è una funzione dispari, così che $\varphi'(u) = -\varphi'(-u)$. Ne segue:

$$(n-1) \varphi'(\bar{x} - y) = \varphi'[(n-1)(\bar{x} - y)].$$

Ponendo $z = \bar{x} - y$ e $m = n-1$ si ha allora:

$$m \varphi'(z) = \varphi'(mz). \quad (3.120)$$

L'equazione ottenuta, in cui m è un intero, esprime una relazione di linearità per la funzione φ' . Mostriamo, invero, che risulta $\varphi'(x) = ax$. A tal fine osserviamo che ponendo $z = u/k$, con k razionale, la (3.120) diventa:

$$m \varphi'\left(\frac{u}{k}\right) = \varphi'\left(\frac{m}{k} u\right), \quad (3.121)$$

mentre per $m = 1/k$ si ricava:

$$\frac{1}{k} \varphi'(z) = \varphi'\left(\frac{z}{k}\right). \quad (3.122)$$

Dalle (3.121) e (3.122) segue:

$$\frac{m}{k} \varphi'(z) = \varphi'\left(\frac{m}{k} z\right).$$

Ponendo $m/k = x$ (si noti che x è un numero razionale), $z = 1$ e $\varphi'(1) = a$ si ha $\varphi'(x) = ax$, da cui segue:

$$\varphi(x) = \frac{a}{2} x^2 + b. \quad (3.123)$$

La relazione (3.123), dimostrata per qualsiasi x razionale, sussiste anche per qualsiasi x reale per la postulata continuità della funzione $\varphi(x)$. Ricordando che $\varphi = \ln f$, si ricava poi:

$$f(x) = c \exp\left(\frac{a}{2} x^2 + b\right).$$

Infine, applicando la condizione di normalizzazione $\int_{-\infty}^{\infty} f(x) dx = 1$, e ricordando che per ipotesi è $f(x) > 0$ per ogni $x \in \mathbb{R}$, segue che f è una densità normale. ■

La proposizione che segue, che qui ci limitiamo ad enunciare¹³, fornisce un'interessante estensione della Proposizione 3.7.4.

Proposizione 3.7.5 *Per $n \geq 3$, sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione continua di densità di probabilità $f(x - \theta)$, e siano $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ le corrispondenti statistiche d'ordine. Se*

$$\hat{\theta} = \sum_{i=1}^n a_i X_{(i)} \quad (a_i \geq 0, \sum_{i=1}^n a_i = 1)$$

è uno stimatore di massima verosimiglianza del parametro θ , si verifica una delle seguenti eventualità:

¹³Cfr. Z. Buczolich e G.J. Székely (1992) *Adv. Appl. Math.* **10**, 439–456.

- (i) $a_1 = a_2 = \dots = a_n = 1/n$;
- (ii) $a_1 + a_n = 1$, $a_1 a_n > 0$;
- (iii) $a_j + a_{j+1} = 1$, $a_j a_{j+1} > 0$, con $j \in \{1, 2, \dots, n-1\}$;
- (iv) $a_j = 1$, con $j \in \{1, 2, \dots, n\}$.

Nel caso (i) f è una densità normale; nel caso (ii) f è una densità uniforme; nel caso (iii) f è una densità di Laplace.

3.8 Metodo dei momenti

In questo paragrafo esporremo uno dei primi metodi ideati per la ricerca di stimatori di parametri incogniti, il cosiddetto *metodo dei momenti* introdotto da Karl Pearson nel 1894.

Consideriamo un campione (X_1, X_2, \dots, X_n) tratto da una popolazione di variabile genetrice X caratterizzata da k parametri incogniti $\theta_1, \theta_2, \dots, \theta_k$, e supponiamo che questi si possano esprimere come funzioni dei momenti intorno all'origine $\mu'_j = E(X^j)$ ($j = 1, 2, \dots, r$), che si assume esistano finiti:

$$\theta_i = g_i(\mu'_1, \mu'_2, \dots, \mu'_r) \quad (i = 1, 2, \dots, k). \quad (3.124)$$

Il metodo dei momenti consiste nello stimare i parametri incogniti θ_i facendo uso degli stimatori

$$\hat{\theta}_i = g_i(\bar{X}^{(1)}, \bar{X}^{(2)}, \dots, \bar{X}^{(r)}) \quad (i = 1, 2, \dots, k), \quad (3.125)$$

dove

$$\bar{X}^{(j)} = \frac{1}{n} \sum_{i=1}^n X_i^j \quad (j = 1, 2, \dots, r)$$

denota il momento campionario di ordine j (cfr. la Definizione 1.3.3). La scelta degli stimatori (3.125) è suggerita dalla circostanza che, come abbiamo visto nell'Esempio 3.4.5, $\bar{X}^{(j)}$ è uno stimatore consistente di μ'_j . Pertanto, se g_i è una funzione continua, $\hat{\theta}_i$ è uno stimatore consistente del parametro θ_i in virtù della Proposizione 3.4.1.

Per poter esprimere i parametri incogniti θ_i come funzioni dei momenti intorno all'origine μ'_j , conviene generalmente procedere al seguente modo: dall'espressione dei momenti μ'_j si individua la dipendenza di questi dai parametri incogniti θ_i :

$$\mu'_j = \varphi_j(\theta_1, \theta_2, \dots, \theta_k) \quad (j = 1, 2, \dots, r); \quad (3.126)$$

risolvendo il sistema (3.126) rispetto ai parametri $\theta_1, \theta_2, \dots, \theta_k$ incogniti, si ricavano poi le equazioni (3.124) che conducono immediatamente agli stimatori (3.125). È bene notare che il numero r di equazioni che costituiscono il sistema (3.126), ossia il numero di momenti intorno all'origine μ'_j coinvolti nella (3.124), deve essere non inferiore al numero k di parametri incogniti in modo che il sistema (3.126) possa ammettere soluzione.

Esaminiamo alcuni esempi di applicazione del metodo dei momenti.

3.8. METODO DEI MOMENTI

Esempio 3.8.1 Dato un campione (X_1, X_2, \dots, X_n) estratto da una popolazione per la quale la media μ e la varianza σ^2 sono parametri incogniti, si desidera determinarne degli stimatori mediante il metodo dei momenti. Il sistema (3.126) in questo caso diventa:

$$\begin{cases} \mu'_1 = \mu, \\ \mu'_2 = \mu^2 + \sigma^2, \end{cases}$$

la cui soluzione

$$\mu = \mu'_1, \quad \sigma^2 = \mu'_2 - (\mu'_1)^2,$$

in accordo con la (3.124), esprime i parametri incogniti μ e σ^2 in funzione dei momenti intorno all'origine μ'_1 e μ'_2 . Come indicato dalla (3.125), gli stimatori dei parametri μ e σ^2 sono quindi:

$$\begin{aligned} \hat{M} &= \bar{X}^{(1)} \equiv \bar{X}, \\ \hat{\Sigma}^2 &= \bar{X}^{(2)} - (\bar{X}^{(1)})^2 \equiv \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i^2 - X_i \bar{X}) = \frac{1}{n} \sum_{i=1}^n (X_i^2 - 2X_i \bar{X} + \bar{X}^2) \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n} S^2 \quad (n > 1). \end{aligned} \quad (3.127)$$

Per una popolazione qualsiasi gli stimatori di media e varianza suggeriti dal metodo dei momenti sono dunque rispettivamente la media campionaria e lo scarto quadratico medio campionario. Si noti che il primo dei due è uno stimatore corretto, mentre il secondo lo è solo asintoticamente. Osserviamo, infine, che nel caso di campioni casuali normali gli stimatori testé ottenuti coincidono con quelli forniti dal metodo della massima verosimiglianza (cfr. Esempio 3.7.7). ◆

Esempio 3.8.2 Sia (X_1, X_2, \dots, X_n) (con $n > 1$) un campione estratto da una popolazione avente distribuzione gamma di parametri α e β . Determiniamo degli stimatori di questi parametri facendo uso del metodo dei momenti. Ricordando che la media e la varianza di una variabile gamma sono rispettivamente $\alpha\beta$ e $\alpha\beta^2$ (cfr. § A.2.4), il sistema (3.126) diventa:

$$\begin{cases} \mu'_1 = \alpha\beta, \\ \mu'_2 = \alpha(\alpha+1)\beta^2, \end{cases}$$

la cui soluzione, corrispondente alla (3.124), è data da

$$\alpha = \frac{(\mu'_1)^2}{\mu'_2 - (\mu'_1)^2}, \quad \beta = \frac{\mu'_2 - (\mu'_1)^2}{\mu'_1}.$$

Pertanto, come indicato dalla (3.125), gli stimatori di α e β sono rispettivamente:

$$\hat{\alpha} = \frac{(\bar{X}^{(1)})^2}{\bar{X}^{(2)} - (\bar{X}^{(1)})^2}, \quad \hat{\beta} = \frac{\bar{X}^{(2)} - (\bar{X}^{(1)})^2}{\bar{X}^{(1)}}. \quad (3.128)$$

Poiché, come mostrato in (3.127), risulta

$$\bar{X}^{(2)} - (\bar{X}^{(1)})^2 = \frac{n-1}{n} S^2,$$

gli stimatori (3.128) si esprimono infine al seguente modo:

$$\hat{A} = \frac{n}{n-1} \frac{\bar{X}^2}{S^2}, \quad \hat{B} = \frac{n-1}{n} \frac{S^2}{\bar{X}}.$$

Si noti che entrambi sono consistenti, essendo funzioni continue di stimatori consistenti (cfr. Proposizione 3.4.1).

Nell'esempio che segue, nell'indicare un'applicazione del metodo dei momenti atta alla determinazione di uno stimatore del parametro di una distribuzione uniforme verrà anche esplicitamente sottolineato come tale metodo possa risultare talvolta insoddisfacente in quanto la stima così ottenuta può essere in contrasto con i dati di cui si dispone. Verrà inoltre mostrato che nel caso di popolazioni uniformi, così come accade utilizzando il metodo della massima verosimiglianza, il metodo dei momenti non conduce a stimatori ad errore quadratico medio minimo.

Esempio 3.8.3 Dato un campione (X_1, X_2, \dots, X_n) estratto da una popolazione distribuita uniformemente nell'intervallo $[0, \theta]$, si vuole determinare uno stimatore di θ facendo uso del metodo dei momenti. Essendo $\mu'_1 = \theta/2$, il metodo dei momenti suggerisce di stimare θ mediante lo stimatore $\hat{\theta} = 2\bar{X}$. Questo è corretto, avendosi:

$$E(\hat{\theta}) = 2E(\bar{X}) = 2\mu'_1 = \theta.$$

Inoltre, la sua varianza è data da

$$D^2(\hat{\theta}) = 4D^2(\bar{X}) = 4 \frac{1}{n} \frac{\theta^2}{12} = \frac{\theta^2}{3n}, \quad (3.129)$$

essendo $\theta^2/12$ la varianza di una variabile distribuita uniformemente in $[0, \theta]$, come specificato nell'Appendice A.2.1. Dal Teorema 3.4.1 segue allora che $\hat{\theta}$ è uno stimatore consistente di θ . Si noti che sebbene $\hat{\theta}$ sia uno stimatore corretto e consistente di θ , esso può condurre in taluni casi a risultati banalmente erronei in quanto può fornire stime di θ che risultano minori di elementi della realizzazione osservata. Così, se in un campione di taglia 3 si è presentata la realizzazione $(1.3, 2.1, 8.6)$, la stima di θ fornita da $\hat{\theta}$ è

$$\hat{\theta} = \frac{2}{3}(1.3 + 2.1 + 8.6) = 8.$$

Essa è però ovviamente erronea in quanto minore del terzo valore osservato. Un inconveniente di questo tipo non si incontra, invece, quando si fa ricorso ai due stimatori di cui agli Esempi 3.7.4 e 3.7.5. In quel caso abbiamo infatti visto che lo stimatore di massima verosimiglianza del parametro θ , dato dalla statistica d'ordine $X_{(n)} \equiv \max\{X_1, X_2, \dots, X_n\}$, non

3.9. STIMATORI DI BAYES

è corretto, ma lo è asintoticamente. Abbiamo poi visto che, in alternativa a tale stimatore è preferibile usare lo stimatore corretto di θ dato da $(n+1)X_{(n)}/n$, avendosi

$$\text{mse}(X_{(n)}) = \frac{2\theta^2}{(n+1)(n+2)} \geq \frac{\theta^2}{n(n+2)} = \text{mse}\left(\frac{n+1}{n} X_{(n)}\right). \quad (3.130)$$

Poiché lo stimatore $\hat{\theta} = 2\bar{X}$ fornito dal metodo dei momenti ha errore quadratico medio $\text{mse}(\hat{\theta}) = D^2(\hat{\theta})$, facendo uso della (3.129) si ricava facilmente che per ogni $\theta > 0$ e per ogni intero positivo n risulta:

$$\text{mse}(\hat{\theta}) = \frac{\theta^2}{3n} \geq \frac{2\theta^2}{(n+1)(n+2)} = \text{mse}(X_{(n)}),$$

ossia che $\hat{\theta}$ ha errore quadratico medio non inferiore a quello di $X_{(n)}$. Pertanto, per la (3.130), lo stimatore ad errore quadratico medio minimo tra i tre stimatori considerati è $(n+1)X_{(n)}/n$. Questo è quindi da preferirsi per la stima di θ , sempre che il criterio di valutazione consista nel preferire lo stimatore a minimo errore quadratico medio.

Esempio 3.8.4 Consideriamo un campione (X_1, X_2, \dots, X_n) estratto da una popolazione distribuita uniformemente in $[-\theta, \theta]$, dove θ è un parametro positivo che intendiamo stimare mediante il metodo dei momenti. Osserviamo innanzitutto che risulta $\mu'_1 = 0$, e che quindi questa equazione non è utile per applicare il metodo dei momenti in quanto non coinvolge il parametro incognito. Occorre dunque ricorrere al momento intorno all'origine di ordine 2 che è dato da

$$\mu'_2 = \frac{\theta^2}{3}.$$

Lo stimatore di θ suggerito dal metodo dei momenti è dunque:

$$\hat{\theta} = \sqrt{3\bar{X}^{(2)}}.$$

Questo è consistente, in quanto funzione continua di stimatore consistente (cfr. Teorema 3.4.2).

3.9. Stimatori di Bayes

Sia dato un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione caratterizzata da un parametro incognito θ . In talune situazioni appare plausibile riguardare θ come valore assunto da una variabile casuale Θ . Ciò si verifica solitamente quando, indipendentemente dall'osservazione di una realizzazione (x_1, x_2, \dots, x_n) del campione casuale in esame, sono disponibili informazioni sul valore di tale parametro e, specificamente, quando queste informazioni sono espresse a mezzo della densità [distribuzione] di probabilità $f_\theta(\theta)$ di Θ , detta anche densità [distribuzione] di probabilità *a priori*. Ad esempio, supponiamo che da una precedente esperienza risulta che un parametro θ può assumere indifferentemente un valore qualsiasi dell'intervallo $(0, 1)$: è allora ragionevole riguardare il valore di θ come il valore assunto da una variabile casuale Θ uniformemente distribuita in tale intervallo.

Supponiamo, più in generale, che le informazioni disponibili su un parametro incognito θ suggeriscano di ritenere che esso sia generato da una variabile Θ avente densità [distribuzione] a priori $f_\Theta(\theta)$, e immaginiamo di essere sul punto di osservare una realizzazione (x_1, x_2, \dots, x_n) di un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione caratterizzata dal parametro θ . Indichiamo con $f(x_1, \dots, x_n | \theta)$ la densità [distribuzione] di probabilità del campione condizionata da $\Theta = \theta$. Siano (x_1, x_2, \dots, x_n) la realizzazione osservata e $f(\theta | x_1, \dots, x_n)$ la densità [distribuzione] di probabilità della variabile casuale Θ condizionata da $X_i = x_i$ ($i = 1, 2, \dots, n$), ossia condizionata dai valori assunti dal campione. Questa è anche detta densità [distribuzione] di probabilità *a posteriori*. Nel caso in cui Θ è continua, in virtù di ben note formule, si ha:

$$f(\theta | x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n | \theta) f_\Theta(\theta)}{\int_{\mathbb{R}} f(x_1, \dots, x_n | \theta) f_\Theta(\theta) d\theta}. \quad (3.131)$$

Il significato di "densità a priori" e "densità a posteriori" va ricercato nella circostanza che la densità $f_\Theta(\theta)$ fornisce informazioni sul parametro *prima* che la realizzazione (x_1, x_2, \dots, x_n) sia disponibile, mentre la densità $f(\theta | x_1, \dots, x_n)$ esprime le informazioni su θ *dopo* che la realizzazione (x_1, x_2, \dots, x_n) è stata osservata.

Nel § 3.2 si è visto che la migliore stima di previsione per il valore che assume una generica variabile casuale è la sua media, nel senso che questa minimizza l'errore quadratico medio. Appare quindi ragionevole stimare θ sulla base della realizzazione (x_1, x_2, \dots, x_n) osservata mediante il cosiddetto *stimatore di Bayes* così definito:

$$\hat{\theta} \stackrel{\text{def}}{=} E(\Theta | X_1, \dots, X_n).$$

In tal modo il parametro θ viene stimato valutando lo stimatore di Bayes in corrispondenza della realizzazione osservata. Per la stima $\hat{\theta}$ di θ si ha pertanto:

$$\hat{\theta} \equiv E(\Theta | X_1 = x_1, \dots, X_n = x_n) = \int_{\mathbb{R}} \theta f(\theta | x_1, \dots, x_n) d\theta.$$

In considerazione dell'importanza che la distribuzione normale riveste, esamineremo ora in dettaglio il caso in cui il parametro incognito di una variabile casuale normale è la media, che supporremo essere caratterizzata a sua volta da una distribuzione a priori normale.

Teorema 3.9.1 *Sia \bar{x} il valore assunto dalla media campionaria \bar{X} di un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione normale di media θ incognita e di varianza σ^2 nota. Se θ viene riguardato come valore assunto da una variabile casuale normale Θ di media μ_0 e varianza σ_0^2 note, la densità a posteriori di Θ è normale di media*

$$E(\Theta | x_1, \dots, x_n) = \frac{n\sigma_0^2 \bar{x} + \sigma^2 \mu_0}{n\sigma_0^2 + \sigma^2} \quad (3.132)$$

e varianza

$$D^2(\Theta | x_1, \dots, x_n) = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}. \quad (3.133)$$

3.9. STIMATORI DI BAYES

Dim. Per ipotesi la densità di probabilità del campione è data da

$$f(x_1, \dots, x_n | \theta) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left[-\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \theta}{\sigma} \right)^2 \right],$$

mentre la densità di probabilità a priori di Θ è

$$f_\Theta(\theta) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp \left[-\frac{1}{2} \left(\frac{\theta - \mu_0}{\sigma_0} \right)^2 \right] \quad (\theta \in \mathbb{R}).$$

Ne segue che, per la (3.131), la densità a posteriori è data da

$$f(\theta | x_1, \dots, x_n) = C \exp \left[-\frac{1}{2} \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \theta^2 + \left(\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) \theta - \frac{1}{2} \left(\frac{1}{\sigma^2} \sum_{i=1}^n x_i^2 + \frac{\mu_0^2}{\sigma_0^2} \right) \right], \quad (3.134)$$

con $\bar{x} = (x_1 + x_2 + \dots + x_n)/n$ e con C costante che non dipende da θ . Se si pone

$$\mu_1 = \frac{n\sigma_0^2 \bar{x} + \sigma^2 \mu_0}{n\sigma_0^2 + \sigma^2} \equiv \left(\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) \sigma_1^2, \quad (3.135)$$

$$\sigma_1^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right)^{-1} \equiv \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}, \quad (3.136)$$

è facile verificare che l'espressione in parentesi quadra nella (3.134) diventa:

$$-\frac{1}{2\sigma_1^2} \theta^2 + \frac{\mu_1}{\sigma_1^2} \theta - \frac{1}{2} \left(\frac{1}{\sigma^2} \sum_{i=1}^n x_i^2 + \frac{\mu_0^2}{\sigma_0^2} \right) = -\frac{(\theta - \mu_1)^2}{2\sigma_1^2} + K, \quad (3.137)$$

con K costante che non dipende da θ . Pertanto, applicando la condizione di normalizzazione alla (3.134) e facendo uso della (3.137), si ricava facilmente la desiderata densità a posteriori:

$$f(\theta | x_1, \dots, x_n) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp \left[-\frac{1}{2} \left(\frac{\theta - \mu_1}{\sigma_1} \right)^2 \right] \quad (\theta \in \mathbb{R}),$$

dove μ_1 e σ_1^2 sono dati dalle (3.135) e (3.136). ■

Dalla (3.132) segue che lo stimatore di Bayes per il parametro θ è il seguente:

$$\hat{\theta} \equiv E(\Theta | X_1, \dots, X_n) = \frac{n\sigma_0^2 \bar{X} + \sigma^2 \mu_0}{n\sigma_0^2 + \sigma^2}. \quad (3.138)$$

È opportuno notare che $\hat{\theta}$ è espresso come una media pesata tra la media campionaria \bar{X} e il valore medio μ_0 della densità a priori. Infatti dalla (3.138) si ricava:

$$\hat{\theta} = \alpha \bar{X} + (1 - \alpha) \mu_0, \quad (3.139)$$

dove

$$\alpha \equiv \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}$$

è compreso tra 0 ed 1. Si noti che α diventa sempre più prossimo ad 1 al crescere della taglia n del campione. Ciò comporta che al divergere di n lo stimatore di Bayes $\hat{\Theta}$ tende alla media campionaria \bar{X} che, come sappiamo, risulta essere lo stimatore di θ suggerito sia dal metodo della massima verosimiglianza che dal metodo dei momenti (cfr. Esempi 3.7.7 e 3.8.1).

Esempio 3.9.1 Una società petrolifera che gestisce distributori di carburante in base ad un proprio modello ritiene che ciascun distributore vende in media settimanalmente $\mu_0 = 537$ confezioni di olio per motori, con una fluttuazione da distributore a distributore caratterizzata da deviazione standard $\sigma_0 = 11.6$. Naturalmente, il numero di confezioni che ogni distributore vende varia da settimana a settimana. Si supponga che tale variazione venga misurata mediante la deviazione standard $\sigma = 36.4$. Rilevato che un distributore di nuova installazione vende 6460 confezioni di olio durante le sue prime 12 settimane di attività, si decide di stimare il numero medio θ di confezioni che esso vende settimanalmente, prefissandosi inoltre di determinare la probabilità con cui tale numero medio è compreso tra 520 e 550. Per risolvere i due quesiti si denoti con $(X_1, X_2, \dots, X_{12})$ il campione casuale, estratto da una popolazione supposta approssimativamente normale, che descrive il numero di confezioni vendute dal nuovo distributore nelle prime 12 settimane di attività, e con $(x_1, x_2, \dots, x_{12})$ i numeri effettivi (non noti) di confezioni vendute in ciascuna delle suddette 12 settimane. Per ipotesi, risulta:

$$\bar{x} = \frac{1}{12} \sum_{i=1}^{12} x_i = \frac{6460}{12} = 538.33. \quad (3.140)$$

Si assuma, inoltre, che la distribuzione a priori di θ sia anch'essa normale, di valore medio $\mu_0 = 537$ e deviazione standard $\sigma_0 = 11.6$. La (3.138) fornisce lo stimatore di Bayes

$$\hat{\Theta} \equiv \frac{n\sigma_0^2 \bar{X} + \sigma^2 \mu_0}{n\sigma_0^2 + \sigma^2},$$

mediante il quale si ricava che, tenendo conto del punto di vista della società petrolifera (ossia sulla base dei valori μ_0 e σ_0 da essa ritenuti validi), la stima $\hat{\theta}$ del numero medio θ di confezioni che vengono vendute in una settimana dal nuovo distributore è la seguente:

$$\hat{\theta} \equiv \frac{n\sigma_0^2 \bar{x} + \sigma^2 \mu_0}{n\sigma_0^2 + \sigma^2} = \frac{12(11.6)^2(6460/12) + (36.4)^2 537}{12(11.6)^2 + (36.4)^2} = 537.73,$$

dove si è fatto uso della (3.140). Non molto dissimile è la stima fornita dalla media campionaria, la quale non tiene conto del punto di vista della Società. Infatti, la (3.140) mostra che risulta $\bar{x} = 538.33$.

Per risolvere il secondo quesito, si osservi che per il Teorema 3.9.1, sotto l'ipotesi adottata di validità dell'approssimazione normale, la densità a posteriori è normale di media

$$E(\Theta|X_1 = x_1, \dots, X_{12} = x_{12}) = 537.73$$

3.9. STIMATORI DI BAYES

e varianza

$$D^2(\Theta|X_1 = x_1, \dots, X_{12} = x_{12}) = 60.65.$$

La probabilità richiesta è allora

$$P(520 \leq \Theta \leq 550|X_1 = x_1, \dots, X_{12} = x_{12}) = P(-2.28 \leq Z \leq 1.57),$$

dove $Z \stackrel{\text{def}}{=} (\Theta - 537.73)/\sqrt{60.65}$ è la variabile casuale normale standard. Facendo uso della Tabella 1 dell'Appendice B segue:

$$\begin{aligned} P(-2.28 \leq Z \leq 1.57) &= P(0 \leq Z \leq 1.57) + P(0 \leq Z \leq 2.28) \\ &= 0.4418 + 0.4887 = 0.9305. \end{aligned}$$

Nell'approssimazione normale, e con il criterio seguito dalla Società, pari a 0.9305 è dunque la probabilità che il numero medio di confezioni vendute settimanalmente dal nuovo distributore sia compreso tra 520 e 550. ◆

Esempio 3.9.2 Consideriamo un campione casuale (X_1, X_2, \dots, X_n) di densità di probabilità

$$f(x_1, \dots, x_n|\theta) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \left\{ \theta \exp \left[-\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu_1)^2}{\sigma^2} \right] + (1-\theta) \exp \left[-\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu_0)^2}{\sigma^2} \right] \right\}, \quad (3.141)$$

con $\mu_0, \mu_1 \in \mathbb{R}$, $\mu_0 < \mu_1$ e $\sigma > 0$. Assumiamo che queste costanti siano note e che θ sia un parametro incognito che può assumere i valori 0 oppure 1. Notiamo che è possibile dare una particolare interpretazione alla popolazione in esame: dalla densità (3.141) si deduce infatti che la variabile casuale genitrice è normale con varianza σ^2 e con media uguale a μ_0 se $\theta = 0$, uguale a μ_1 se $\theta = 1$. Supponiamo che il valore assunto da θ sia rappresentabile mediante una variabile casuale di Bernoulli Θ di distribuzione di probabilità a priori

$$f_\Theta(\theta) = \frac{1}{2} \quad (\theta = 0, 1). \quad (3.142)$$

Lo stimatore di Bayes è immediatamente ottenuto:

$$\hat{\Theta} \equiv E(\Theta|X_1, \dots, X_n) = P(\Theta = 1|X_1, \dots, X_n). \quad (3.143)$$

Facendo uso dell'espressione generale (3.131) della distribuzione di probabilità a posteriori e della (3.142) segue:

$$f(1|x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n|1)f_\Theta(1)}{\sum_{\theta=0}^1 f(x_1, \dots, x_n|\theta)f_\Theta(\theta)} = \frac{f(x_1, \dots, x_n|1)}{\sum_{\theta=0}^1 f(x_1, \dots, x_n|\theta)},$$

da cui, in virtù della (3.141), si ricava:

$$f(1|x_1, \dots, x_n) = \frac{\exp \left[-\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu_1)^2}{\sigma^2} \right]}{\exp \left[-\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu_1)^2}{\sigma^2} \right] + \exp \left[-\frac{1}{2} \sum_{i=1}^n \frac{(x_i - \mu_0)^2}{\sigma^2} \right]}. \quad (3.144)$$

Osserviamo poi che si ha:

$$\begin{aligned} \frac{\exp\left[-\frac{1}{2}\sum_{i=1}^n \frac{(x_i - \mu_1)^2}{\sigma^2}\right]}{\exp\left[-\frac{1}{2}\sum_{i=1}^n \frac{(x_i - \mu_0)^2}{\sigma^2}\right]} &= \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n [(x_i - \mu_1)^2 - (x_i - \mu_0)^2]\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n [(\mu_0 - \mu_1)2x_i + \mu_1^2 - \mu_0^2]\right\} \\ &= \exp\left\{n \frac{\mu_1 - \mu_0}{\sigma^2} \left(\bar{x} - \frac{\mu_0 + \mu_1}{2}\right)\right\}, \end{aligned} \quad (3.145)$$

dove \bar{x} denota il valore assunto dalla media campionaria \bar{X} . Così, ponendo

$$a = -n \frac{\mu_1 - \mu_0}{\sigma^2} \frac{\mu_0 + \mu_1}{2}, \quad b = n \frac{\mu_1 - \mu_0}{\sigma^2}, \quad (3.146)$$

e facendo uso della (3.145), la (3.144) diviene:

$$f(1|x_1, \dots, x_n) \equiv \hat{\theta} = \frac{e^{a+b\bar{x}}}{1 + e^{a+b\bar{x}}}. \quad (3.147)$$

Dalle (3.143) e (3.147) si ricava infine lo stimatore di Bayes:

$$\hat{\Theta} \equiv P(\Theta = 1|X_1, \dots, X_n) = \frac{e^{a+b\bar{X}}}{1 + e^{a+b\bar{X}}}. \quad (3.148)$$

Si noti che $\hat{\Theta}$ è uno stimatore della probabilità che la variabile casuale di Bernoulli Θ assuma il valore 1. La stima (3.147) di tale probabilità risulta maggiore di 1/2 quando $e^{a+b\bar{x}}$ è maggiore di 1; cioè, per le posizioni (3.146) e per l'ipotesi che $\mu_0 < \mu_1$, si verifica allorché \bar{x} è maggiore della media aritmetica di μ_0 e μ_1 .

Lo stimatore (3.148) viene detto *logistico* in analogia con la denominazione di "logistica" attribuita alla curva di equazione

$$y = \frac{e^{a+b\bar{x}}}{1 + e^{a+b\bar{x}}}.$$



Esempio 3.9.3 Consideriamo un campione casuale (X_1, X_2, \dots, X_n) costituito da variabili di Poisson di parametro θ incognito. Supponendo che questo sia rappresentabile mediante una variabile casuale esponenziale a media unitaria, determiniamo lo stimatore di Bayes $\hat{\Theta} = E(\Theta|X_1, \dots, X_n)$. Osserviamo anzitutto che la densità del campione e la densità a priori sono rispettivamente

$$\begin{aligned} f(x_1, \dots, x_n|\theta) &= \prod_{i=1}^n \frac{\theta^{x_i}}{x_i!} e^{-\theta} \\ &= \frac{\theta^{x_1+\dots+x_n} e^{-n\theta}}{x_1! \dots x_n!} \quad (x_i > 0; i = 1, 2, \dots, n) \end{aligned} \quad (3.149)$$

$$f_\theta(\theta) = e^{-\theta} \quad (\theta > 0). \quad (3.150)$$

3.9. STIMATORI DI BAYES

Si ricava dunque:

$$\int_0^\infty f(x_1, \dots, x_n|\theta) f_\theta(\theta) d\theta = \int_0^\infty \frac{\theta^x e^{-\theta(n+1)}}{x_1! \dots x_n!} d\theta, \quad (3.151)$$

dove si è posto $x = x_1 + \dots + x_n$. Il secondo membro della (3.151) può esprimersi come il prodotto della costante $[(n+1)x_1! \dots x_n!]^{-1}$ per il momento di ordine x intorno all'origine di una variabile casuale esponenziale di media $(n+1)^{-1}$ che, come è facile verificare, è dato da $x! (n+1)^{-x}$. Pertanto la (3.151) diventa:

$$\int_0^\infty f(x_1, \dots, x_n|\theta) f_\theta(\theta) d\theta = \frac{x!}{x_1! \dots x_n!} \frac{1}{(n+1)^{x+1}}. \quad (3.152)$$

Facendo uso delle (3.149), (3.150) e dell'integrale (3.152), dalla (3.131) si ottiene immediatamente la densità a posteriori:

$$f(\theta|x_1, \dots, x_n) = \frac{(n+1)^{x+1}}{x!} \theta^x e^{-(n+1)\theta} \quad (\theta > 0).$$

Questa è la densità di una variabile casuale di tipo gamma di parametri $\alpha = x+1$ e $\beta = (n+1)^{-1}$, dotata quindi di valore medio $\alpha\beta$, così che

$$E(\Theta|X_1 = x_1, \dots, X_n = x_n) \equiv \alpha\beta = \frac{x+1}{n+1}.$$

In definitiva,

$$\hat{\Theta} = \frac{1}{n+1} \left(\sum_{i=1}^n x_i + 1 \right)$$

è lo stimatore di Bayes richiesto. Si noti che questo stimatore differisce dallo stimatore di massima verosimiglianza; nell'Esempio 3.7.6 si è infatti visto che lo stimatore di massima verosimiglianza per il parametro di una distribuzione di Poisson è dato dalla media campionaria \bar{X} . Riferendoci ad un caso specifico, supponiamo che il numero di componenti imperfetti prodotti giornalmente da un dato impianto industriale sia distribuito secondo una variabile di Poisson di media θ incognita e assumiamo che il valore di θ sia distribuito secondo una variabile esponenziale di media unitaria. Se nell'arco di 15 giorni vengono prodotti 74 componenti imperfetti, la stima di Bayes di θ è data da

$$\hat{\theta} = \frac{x+1}{n+1} = \frac{75}{16} = 4.6875,$$

mentre la stima di massima verosimiglianza di θ , per la quale non si fa uso dell'ipotesi che il valore di θ sia distribuito esponenzialmente, è

$$\bar{x} = \frac{x}{n} = \frac{74}{15} = 4.9333.$$



Esempio 3.9.4 Consideriamo un campione casuale (X_1, X_2, \dots, X_n) costituito da variabili di Bernoulli di parametro θ incognito che si suppone uniformemente distribuito nell'intervallo $(0, 1)$. Per determinare lo stimatore di Bayes di θ , notiamo che risulta:

$$f(x_1, \dots, x_n | \theta) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} = \theta^x (1-\theta)^{n-x} \quad (x = 0, 1, \dots, n), \quad (3.153)$$

con $x = x_1 + x_2 + \dots + x_n$, e che invece è

$$f_\Theta(\theta) = 1 \quad (0 < \theta < 1). \quad (3.154)$$

Dalle (3.153) e (3.154) segue:

$$\int_0^1 f(x_1, \dots, x_n | \theta) f_\Theta(\theta) d\theta = \int_0^1 \theta^x (1-\theta)^{n-x} d\theta. \quad (3.155)$$

Poiché, come è facile dimostrare mediante ripetute integrazioni per parti, per k e r interi positivi risulta

$$\int_0^1 \theta^k (1-\theta)^r d\theta = \frac{k! r!}{(k+r+1)!}, \quad (3.156)$$

dalla (3.155) si ha:

$$\int_0^1 f(x_1, \dots, x_n | \theta) f_\Theta(\theta) d\theta = \frac{x!(n-x)!}{(n+1)!}. \quad (3.157)$$

Facendo poi uso delle (3.153), (3.154) e dell'integrale (3.157) nella (3.131), si giunge infine all'espressione della densità a posteriori:

$$f(\theta | x_1, \dots, x_n) = \frac{(n+1)!}{x!(n-x)!} \theta^x (1-\theta)^{n-x} \quad (0 < \theta < 1).$$

Facendo ancora ricorso alla (3.156) si ottiene poi:

$$\begin{aligned} E(\theta | X_1 = x_1, \dots, X_n = x_n) &= \frac{(n+1)!}{x!(n-x)!} \int_0^1 \theta^{x+1} (1-\theta)^{n-x} d\theta \\ &= \frac{(n+1)!}{x!(n-x)!} \frac{(x+1)!(n-x)!}{(n+2)!} \\ &= \frac{x+1}{n+2}. \end{aligned}$$

Lo stimatore di Bayes richiesto è dunque

$$\hat{\theta} = \frac{1}{n+2} \left(\sum_{i=1}^n X_i + 1 \right).$$

Osserviamo che questo differisce dallo stimatore di massima verosimiglianza dato dalla media campionaria \bar{X} (si veda l'Esempio 3.7.3 nel caso $k = 1$).

Si consideri, come caso particolare, un macchinario che produce dei dispositivi, ciascuno dei quali è difettoso con probabilità θ , e supponiamo che θ sia distribuito uniformemente

nell'intervallo $(0, 1)$. Se vengono prodotti 235 dispositivi, 23 dei quali sono difettosi, la stima di Bayes di θ è

$$\hat{\theta} = \frac{x+1}{n+2} = \frac{24}{237} = 0.1013,$$

mentre per la stima di massima verosimiglianza di θ , la quale non tiene conto dell'ipotesi che θ venga scelto uniformemente in $(0, 1)$, si ha:

$$\bar{x} = \frac{x}{n} = \frac{23}{235} = 0.0979.$$

Consideriamo ora il caso in cui un parametro incognito θ è uniformemente distribuito nell'intervallo (a, b) . La densità a priori è dunque:

$$f_\Theta(\theta) = \begin{cases} \frac{1}{b-a} & \text{per } a < \theta < b, \\ 0 & \text{altrimenti.} \end{cases}$$

Sotto questa ipotesi la densità a posteriori, in virtù della (3.131), è data da

$$\begin{aligned} f(\theta | x_1, \dots, x_n) &= \frac{f(x_1, \dots, x_n | \theta) f_\Theta(\theta)}{\int_a^b f(x_1, \dots, x_n | \theta) f_\Theta(\theta) d\theta} \\ &= \frac{f(x_1, \dots, x_n | \theta)}{\int_a^b f(x_1, \dots, x_n | \theta) d\theta} \quad (a < \theta < b). \end{aligned} \quad (3.158)$$

Dalla (3.158) si deduce subito che una moda¹⁴ della densità a posteriori $f(\theta | x_1, \dots, x_n)$ è il valore di θ che massimizza la funzione $f(x_1, \dots, x_n | \theta)$. Notiamo che, considerando θ come parametro e non come valore assunto dalla variabile casuale Θ , la funzione $f(x_1, \dots, x_n | \theta)$ si identifica con la funzione di verosimiglianza $L(\theta)$. Come è noto dal § 3.7, il valore di θ che massimizza la funzione $L(\theta) \equiv f(x_1, \dots, x_n | \theta)$ costituisce la stima di massima verosimiglianza del parametro θ . Se ne conclude che la stima di massima verosimiglianza di θ , ristretta all'intervallo (a, b) , è uguale alla moda della densità a posteriori sotto l'ipotesi che la densità a priori sia uniforme nell'intervallo (a, b) .

¹⁴Una moda di una densità di probabilità di una variabile casuale è un valore per il quale la densità di probabilità presenta un massimo relativo.

Capitolo 4

Stima intervallare

4.1 Intervalli fiduciari

La stima puntuale, di cui si è parlato nel Cap. 3, costituisce certo uno strumento efficace per l'individuazione dei parametri incogniti di una popolazione, ma non è esente da difetti. Si consideri, ad esempio, un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione di parametro incognito θ che viene stimato a mezzo di uno stimatore corretto $\hat{\theta} = g(X_1, X_2, \dots, X_n)$. Così, a θ viene attribuito il valore $\hat{\theta} = g(x_1, x_2, \dots, x_n)$ che $\hat{\theta}$ assume in corrispondenza di una realizzazione (x_1, x_2, \dots, x_n) osservata. Se da un lato così facendo tale valore resta individuato con precisione, dall'altro esso non è sempre significativo; soprattutto se la taglia n del campione è piccola, anche quando $\hat{\theta}$ è consistente. Invero, solo al limite in cui $n \rightarrow \infty$ il valore $\theta = E(\hat{\theta})$ viene assunto con certezza da $\hat{\theta}$, avendosi

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| < \varepsilon) = 1$$

per ogni $\varepsilon > 0$ in virtù dell'ipotesi di consistenza. Ma se, come necessariamente accade, la taglia del campione è finita, $\hat{\theta}$ è una variabile casuale che assume il suo valore medio θ con un'incertezza valutabile, ad esempio, mediante la deviazione standard $D(\hat{\theta})$, se esistente. In particolare, se il campione è estratto da una popolazione continua si ha $P(\hat{\theta} = \theta) = 0$ essendo nulla la probabilità che una variabile casuale continua assuma un singolo valore.

Per meglio chiarire questo concetto consideriamo a titolo esemplificativo un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione normale di media θ incognita e varianza σ^2 nota. Per stimare il parametro incognito θ a mezzo di una realizzazione del campione conviene fare uso della media campionaria \bar{X} che, come sappiamo, è uno stimatore corretto, consistente e pienamente efficiente. È però facile, ed istruttivo, dimostrare che risulta $P(\bar{X} = \theta) = 0$. Infatti, per ogni $\varepsilon > 0$ sussiste la seguente relazione di inclusione tra eventi:

$$\{\bar{X} = \theta\} \subset \left\{ \theta - \frac{\varepsilon\sigma}{2\sqrt{n}} \leq \bar{X} \leq \theta + \frac{\varepsilon\sigma}{2\sqrt{n}} \right\}.$$

Passando alle probabilità, ed osservando che $(\bar{X} - \theta)/\sqrt{n}/\sigma$ è una variabile casuale normale

standard, si ottiene:

$$\begin{aligned} P(\bar{X} = \theta) &\leq P\left(\theta - \frac{\varepsilon\sigma}{2\sqrt{n}} \leq \bar{X} \leq \theta + \frac{\varepsilon\sigma}{2\sqrt{n}}\right) \\ &= P\left(-\frac{\varepsilon}{2} \leq \frac{\bar{X} - \theta}{\sigma/\sqrt{n}} \leq \frac{\varepsilon}{2}\right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\varepsilon/2}^{\varepsilon/2} e^{-t^2/2} dt = \frac{\varepsilon}{\sqrt{2\pi}} e^{-\eta^2/2}, \end{aligned} \quad (4.1)$$

dove, avendo fatto uso del teorema della media, η denota un punto interno all'intervallo $(-\varepsilon/2, \varepsilon/2)$. Maggiorando $e^{-\eta^2/2}$ con l'unità, dalla (4.1) segue in definitiva:

$$P(\bar{X} = \theta) \leq \frac{\varepsilon}{\sqrt{2\pi}},$$

così che $P(\bar{X} = \theta) = 0$ in virtù dell'arbitrarietà di ε .

È opportuno notare che anche nel caso discreto accade che la stima fornita da stimatori corretti e consistenti possa essere imprecisa. Basti osservare che nel caso di un campione casuale di taglia n estratto da una popolazione di Bernoulli di parametro $\theta \equiv 1/2$ la media campionaria \bar{X} , sebbene sia uno stimatore corretto e consistente di θ , non assume mai valore $1/2$ se n è dispari, per elevata che sia la taglia del campione.

Sorge in conclusione la necessità di disporre di un ulteriore metodo di stima: in luogo di uno stimatore che fornisce quale stima un unico valore esattamente specificato ma che può avere bassa probabilità di occorrenza si ricerca un intervallo aleatorio, i cui estremi sono variabili casuali, caratterizzato da alta probabilità di contenere il valore incognito del parametro. Si passa quindi da uno stimatore $\hat{\theta}$ che fornisce una stima puntuale del parametro θ ad una coppia di variabili casuali rappresentanti gli estremi dell'intervallo che si vuole contenga il valore incognito di θ con un prefissato piccolo margine di incertezza. Nasce così la cosiddetta *stima intervallare* dei parametri incogniti.

Ritornando all'esempio sulla popolazione normale di media θ incognita va comunque detto che se è certo vero che la media campionaria \bar{X} non uguaglia la media con probabilità unitaria, i valori che assume sono comunque "prossimi" ad essa. Per meglio rendere l'idea di quanto siano prossimi a θ è opportuno rendere meno rigida la stima di questo parametro considerando, ad esempio, in luogo di \bar{X} le statistiche

$$\begin{aligned} C^- &= g^-(X_1, X_2, \dots, X_n) \stackrel{\text{def}}{=} \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} \\ C^+ &= g^+(X_1, X_2, \dots, X_n) \stackrel{\text{def}}{=} \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}. \end{aligned}$$

Calcoliamo la probabilità dell'evento

$$\{C^- < \theta < C^+\} = \{C^- < \theta\} \cap \{C^+ > \theta\}.$$

Si ha evidentemente:

$$\begin{aligned} P(C^- < \theta < C^+) &= P\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \theta < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) \\ &= P\left(-1.96 < \frac{\bar{X} - \theta}{\sigma/\sqrt{n}} < 1.9\right). \end{aligned}$$

4.1. INTERVALLI FIDUCIARI

Poiché $Z \stackrel{\text{def}}{=} (\bar{X} - \theta)/\sigma/\sqrt{n}$ è una variabile casuale normale standard, risulta:

$$P(C^- < \theta < C^+) = 2P(0 < Z < 1.96) = 0.95, \quad (4.2)$$

avendo fatto uso della Tabella 1 dell'Appendice B dalla quale si evince che $P(0 < Z < 1.96) = 0.475$. Il risultato al quale siamo giunti suggerisce di procedere nel seguente modo. Osservata una realizzazione (x_1, x_2, \dots, x_n) del campione casuale considerato, calcoliamo le quantità

$$c^- = \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \quad c^+ = \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}},$$

dove $\bar{x} = (x_1 + x_2 + \dots + x_n)/n$. In virtù della (4.2), l'affermazione che risulta $c^- < \theta < c^+$ è alla lunga corretta nel 95% dei casi o, equivalentemente, essa risulta inattendibile soltanto nel 5% dei casi quando si ripeta un gran numero di volte l'osservazione del campione casuale estratto dalla popolazione in esame.

In maniera analoga si procede nel caso di popolazioni di natura qualsiasi sostituendo alla stima puntuale di un parametro della popolazione in considerazione un intervallo di valori designato quale *intervallo fiduciario*. Si determinano, invero, due limiti entro i quali sia compreso il valore incognito del parametro con una assegnata probabilità denominata *coefficiente o grado di fiducia*. Il nuovo tipo di stima, che si propone come alternativa alla stima puntuale, è detta appunto *stima intervallare*.

Definizione 4.1.1 Dato un campione casuale (X_1, X_2, \dots, X_n) di taglia n estratto da una generica popolazione di parametro incognito θ , denotiamo con $C^- = g^-(X_1, X_2, \dots, X_n)$ e $C^+ = g^+(X_1, X_2, \dots, X_n)$ due statistiche indipendenti da θ soddisfacenti la condizione¹ $C^- < C^+$. Se risulta

$$P(C^- < \theta < C^+) = 1 - \alpha, \quad (4.3)$$

dove α è un reale compreso tra 0 e 1 che non dipende da θ , l'intervallo (C^-, C^+) viene detto *intervallo fiduciario per θ di coefficiente $1 - \alpha$* .

Le statistiche C^- e C^+ sono note come *estremo sinistro* e *estremo destro* dell'intervallo fiduciario. Denotati poi con $c^- = g^-(x_1, x_2, \dots, x_n)$ e $c^+ = g^+(x_1, x_2, \dots, x_n)$ i valori assunti dalle statistiche C^- e C^+ in corrispondenza di una realizzazione (x_1, x_2, \dots, x_n) del campione casuale considerato, l'intervallo (c^-, c^+) costituisce una *stima dell'intervallo fiduciario* di coefficiente $1 - \alpha$ per il parametro θ .

Si noti che la (4.3) afferma che l'intervallo casuale (C^-, C^+) contiene il parametro incognito θ con probabilità $1 - \alpha$.

Riassumendo, per effettuare una stima intervallare del parametro θ di una popolazione si assegna innanzitutto un coefficiente di fiducia $1 - \alpha$; si ricercano successivamente due statistiche, $C^- = g^-(X_1, X_2, \dots, X_n)$ e $C^+ = g^+(X_1, X_2, \dots, X_n)$ dipendenti dal campione casuale e dal fissato α , tali che risulti $P(C^- < \theta < C^+) = 1 - \alpha$. I valori $c^- = g^-(x_1, x_2, \dots, x_n)$ e $c^+ = g^+(x_1, x_2, \dots, x_n)$ assunti dalle statistiche C^- e C^+ in corrispondenza della realizzazione (x_1, x_2, \dots, x_n) forniscono la stima (c^-, c^+) di grado $1 - \alpha$ per il parametro θ .

¹La condizione $C^- < C^+$ significa che $g^-(x_1, x_2, \dots, x_n)$ risulta sempre minore di $g^+(x_1, x_2, \dots, x_n)$ per ogni realizzazione (x_1, x_2, \dots, x_n) del campione.

Un'applicazione immediata della definizione di coefficiente fiduciario la si trova in problemi in cui ci si chiede se θ assume uno specificato valore. Ad esempio, supponiamo che si desidera stabilire se il valore incognito di un parametro θ è uguale ad un reale θ_0 prefissato. Se (c^-, c^+) è la stima dell'intervallo fiduciario per θ di coefficiente $1 - \alpha$, con α prossimo a zero, e se risulta $\theta_0 \notin (c^-, c^+)$, si è maggiormente inclini a ritenere che θ non sia uguale a θ_0 dal momento che la stima intervallare di θ non include tale valore.

Per chiarire ulteriormente il significato di intervallo fiduciario consideriamo il caso di stima di un parametro θ e supponiamo che, ad esempio, (c^-, c^+) costituisca la stima del relativo intervallo fiduciario di coefficiente 0,95. Ciò non va interpretato affermando che la probabilità che θ appartenga a (c^-, c^+) è 0,95, perché tale affermazione non coinvolge alcuna variabile casuale, così che non ha senso parlare di probabilità. Si deve invece affermare che la tecnica usata per ottenere la stima dell'intervallo fiduciario è tale che nel 95% dei casi in cui viene usata essa conduce ad un intervallo (c^-, c^+) che contiene il parametro θ . In altri termini, prima di osservare la realizzazione del campione possiamo assicurare che con probabilità 0,95 l'intervallo che si otterrà conterrà θ , mentre dopo aver osservato i dati si può solo assicurare che l'intervallo (c^-, c^+) contiene θ con "grado di fiducia pari a 0,95".

Dato un intervallo fiduciario (C^-, C^+) di coefficiente $1 - \alpha$, la statistica

$$\begin{aligned} A(X_1, X_2, \dots, X_n) &= C^+ - C^- \\ &\equiv g^+(X_1, X_2, \dots, X_n) - g^-(X_1, X_2, \dots, X_n) \end{aligned} \quad (4.4)$$

costituisce l'ampiezza dell'intervallo fiduciario. Essa determina il grado di precisione della stima. Fissata infatti α e considerata una data realizzazione (x_1, x_2, \dots, x_n) del campione casuale, quanto minore è $A(x_1, x_2, \dots, x_n)$ tanto maggiore è la precisione della stima.

Esempio 4.1.1 Si consideri un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione la cui variabile casuale genitrice ha densità di probabilità

$$f(x; \theta) = \begin{cases} e^{-(x-\theta)} & \text{per } x > \theta, \\ 0 & \text{altrimenti,} \end{cases}$$

dove θ è un parametro positivo. Mostriamo che l'intervallo

$$(X_{(1)} + \frac{\ln \alpha}{n}, X_{(1)}), \quad (4.5)$$

dove $X_{(1)} \equiv \min\{X_1, X_2, \dots, X_n\}$ denota la prima statistica d'ordine, costituisce un intervallo fiduciario per θ di coefficiente $1 - \alpha$. Infatti si ha:

$$\begin{aligned} P\left(X_{(1)} + \frac{\ln \alpha}{n} < \theta < X_{(1)}\right) &= P\left(\theta < X_{(1)} < \theta - \frac{\ln \alpha}{n}\right) \\ &= F_{(1)}\left(\theta - \frac{\ln \alpha}{n}\right) - F_{(1)}(\theta), \end{aligned} \quad (4.6)$$

dove $F_{(1)}(x)$ denota la funzione di distribuzione di $X_{(1)}$. Ricordando la (3.4), si ottiene:

$$F_{(1)}(x) = \begin{cases} 0 & \text{per } x < \theta, \\ 1 - e^{-n(x-\theta)} & \text{per } x \geq \theta. \end{cases} \quad (4.7)$$

4.2. METODO DEL CARDINE

Dalla (4.6) e (4.7) si trae poi:

$$P\left(X_{(1)} + \frac{\ln \alpha}{n} < \theta < X_{(1)}\right) = 1 - \alpha. \quad (4.8)$$

La (4.8) mostra che la (4.5) è un intervallo fiduciario per θ di coefficiente $1 - \alpha$. Notiamo che, per la (4.4), l'ampiezza di tale intervallo è

$$A(X_1, X_2, \dots, X_n) = -\frac{\ln \alpha}{n},$$

che risulta essere decrescente in n . La precisione della stima aumenta dunque al crescere della taglia del campione. ♦

Osserviamo che nel caso di campioni casuali estratti da popolazione discreta può non risultare possibile determinare C^- e C^+ in modo che la probabilità $P(C^- < \theta < C^+)$ sia esattamente uguale al coefficiente di fiducia $1 - \alpha$ prefissato. In tal caso si adotta come intervallo fiduciario di coefficiente $1 - \alpha$ il più piccolo intervallo (C^-, C^+) per il quale, in luogo della (4.3), risulta:

$$P(C^- < \theta < C^+) \geq 1 - \alpha.$$

È anche opportuno menzionare che gli intervalli fiduciari (C^-, C^+) , introdotti nella Definizione 4.1.1, sono detti *bidirezionali*. In realtà, in applicazioni che richiedono la presenza di limiti solo superiori o solo inferiori è anche possibile introdurre intervalli fiduciari *unidirezionali*, ossia del tipo $(-\infty, C^+)$ o (C^-, ∞) .

Nel paragrafo che segue esporremo un metodo per la costruzione di intervalli fiduciari nel caso di popolazione continua.

4.2 Metodo del cardine

Efficace per la costruzione di intervalli fiduciari per la stima intervallare di parametri di popolazioni continue è il cosiddetto *metodo del pivot*, che qui preferiamo denotare quale *metodo del cardine* o dell'*inversione analitica*. Questo consiste nel determinare una variabile casuale

$$C = \gamma(X_1, X_2, \dots, X_n; \theta) \quad (4.9)$$

che gode delle seguenti proprietà:

- (i) i suoi valori dipendono dalla realizzazione (x_1, x_2, \dots, x_n) del campione (X_1, X_2, \dots, X_n) e dal parametro θ da stimare;
- (ii) la sua funzione di distribuzione non dipende dal parametro θ ;
- (iii) per ogni $\alpha \in (0, 1)$ esistono due reali $\alpha_1 = h_1(\alpha)$ e $\alpha_2 = h_2(\alpha)$ (con $\alpha_1 < \alpha_2$) dipendenti da α tali che per ogni valore di θ risulta

$$P(\alpha_1 < C < \alpha_2) = 1 - \alpha; \quad (4.10)$$

(iv) per ogni realizzazione (x_1, x_2, \dots, x_n) e per ogni θ sussiste la seguente equivalenza:

$$\alpha_1 < \gamma(x_1, x_2, \dots, x_n; \theta) < \alpha_2 \iff g^-(x_1, x_2, \dots, x_n) < \theta < g^+(x_1, x_2, \dots, x_n), \quad (4.11)$$

con $g^-(x_1, x_2, \dots, x_n)$ e $g^+(x_1, x_2, \dots, x_n)$ funzioni dipendenti soltanto da (x_1, x_2, \dots, x_n) .

Dalla (4.11), in virtù della (4.9), segue che la (4.10) può essere così espressa:

$$P[g^-(X_1, X_2, \dots, X_n) < \theta < g^+(X_1, X_2, \dots, X_n)] = 1 - \alpha.$$

Per la Definizione 4.1.1

$$(g^-(X_1, X_2, \dots, X_n), g^+(X_1, X_2, \dots, X_n))$$

costituisce allora un intervallo fiduciario per θ di coefficiente $1 - \alpha$. La variabile casuale (4.9) è detta *cardine* e $c = \gamma(x_1, x_2, \dots, x_n; \theta)$ è il valore da essa assunto in corrispondenza della realizzazione (x_1, x_2, \dots, x_n) .

La condizione (iv) è sicuramente soddisfatta se la funzione $c = \gamma(x_1, x_2, \dots, x_n; \theta)$ è invertibile rispetto a θ . Per rendercene conto, indichiamo con $\gamma^{-1}(x_1, x_2, \dots, x_n; c)$ la funzione inversa di $\gamma(x_1, x_2, \dots, x_n; \theta)$ e poniamo

$$g^-(x_1, x_2, \dots, x_n) = \begin{cases} \gamma^{-1}(x_1, x_2, \dots, x_n; \alpha_1) & \text{se } \gamma \text{ è strettamente crescente in } \theta, \\ \gamma^{-1}(x_1, x_2, \dots, x_n; \alpha_2) & \text{se } \gamma \text{ è strettamente decrescente in } \theta \end{cases}$$

$$g^+(x_1, x_2, \dots, x_n) = \begin{cases} \gamma^{-1}(x_1, x_2, \dots, x_n; \alpha_2) & \text{se } \gamma \text{ è strettamente crescente in } \theta, \\ \gamma^{-1}(x_1, x_2, \dots, x_n; \alpha_1) & \text{se } \gamma \text{ è strettamente decrescente in } \theta. \end{cases}$$

Le relazioni

$$\alpha_1 < \gamma(x_1, x_2, \dots, x_n; \theta) < \alpha_2$$

$$g^-(x_1, x_2, \dots, x_n) < \theta < g^+(x_1, x_2, \dots, x_n)$$

sono allora equivalenti, di modo che la (4.11) è soddisfatta.

In conclusione per determinare un intervallo fiduciario per θ mediante il metodo del cardine, una volta fissato il coefficiente $1 - \alpha$ si effettua la scelta di due valori $\alpha_1 = h_1(\alpha)$ e $\alpha_2 = h_2(\alpha)$ dipendenti da α tali che $P(\alpha_1 < C < \alpha_2) = 1 - \alpha$, dove C è una variabile cardine. Per la (4.11), l'evento $\{\alpha_1 < C < \alpha_2\}$ coincide con l'evento $\{C^- < \theta < C^+\}$, dove $C^- = g^-(X_1, X_2, \dots, X_n)$ e $C^+ = g^+(X_1, X_2, \dots, X_n)$. Si ottiene così l'intervallo fiduciario (C^-, C^+) desiderato di coefficiente $1 - \alpha$.

L'utilizzazione del metodo appena descritto richiede preliminarmente la determinazione di una variabile cardine. Dopo aver dimostrato la proposizione che segue, vedremo che, sotto opportune ipotesi, ciò risulta sempre possibile.

4.2. METODO DEL CARDINE

Proposizione 4.2.1 Se X è una variabile casuale continua di funzione di distribuzione $F(x)$ continua e strettamente crescente, la variabile casuale $U = F(X)$ ha distribuzione uniforme nell'intervalllo $(0, 1)$.

Dim. Se indichiamo con F_X^{-1} la funzione inversa di F_X , per $0 < u < 1$ risulta:

$$\begin{aligned} F_U(u) &= P(U \leq u) = P[F_X(X) \leq u] = P[X \leq F_X^{-1}(u)] \\ &= F_X[F_X^{-1}(u)] = u \end{aligned}$$

così che U è una variabile casuale uniforme nell'intervalllo $(0, 1)$. ■

Siamo ora in grado di dimostrare il seguente teorema:

Teorema 4.2.1 Se (X_1, X_2, \dots, X_n) è un campione casuale estratto da una popolazione continua, caratterizzata dal parametro θ , di variabile casuale genitrice dotata di funzione di distribuzione $F(x; \theta)$ continua e strettamente crescente in x e continua e strettamente monotona in θ , allora

$$C = \gamma(X_1, X_2, \dots, X_n; \theta) = -\sum_{i=1}^n \ln F(X_i; \theta) \quad (4.12)$$

è una variabile cardine per la stima intervallare di θ .

Dim. Essendo $F(x; \theta)$ continua e strettamente monotona in θ per ogni $i = 1, 2, \dots, n$, la funzione $\gamma(x_1, x_2, \dots, x_n; \theta)$, definita nella (4.12), è anch'essa continua e strettamente monotona in θ . Inoltre, per la Proposizione 4.2.1 le variabili casuali $F(X_i; \theta)$ hanno distribuzione uniforme 0 nell'intervalllo $(0, 1)$ e quindi:

$$P[F(X_i; \theta) \geq u] = 1 - u \quad (0 \leq u \leq 1).$$

Di conseguenza risulta:

$$P[-\ln F(X_i; \theta) \leq x] = P[F(X_i; \theta) \geq e^{-x}] = 1 - e^{-x} \quad (x > 0);$$

le variabili casuali $-\ln F(X_i; \theta)$ hanno dunque distribuzione esponenziale di media unitaria. È poi immediato dimostrare che la variabile casuale C , essendo somma di n variabili casuali indipendenti esponenziali di media unitaria, ha distribuzione gamma di parametri $\alpha = n$ e $\beta = 1$. Infatti risulta:

$$\begin{aligned} M_C(t) &\stackrel{\text{def}}{=} E(e^{tC}) = E\left\{\exp\left[-t\sum_{i=1}^n \ln F(X_i; \theta)\right]\right\} \\ &= \prod_{i=1}^n E\left[e^{-t \ln F(X_i; \theta)}\right] = \prod_{i=1}^n \left(\frac{1}{1-t}\right) \\ &= \left(\frac{1}{1-t}\right)^n \quad (t < 1), \end{aligned}$$

che è immediato riconoscere essere la funzione generatrice dei momenti di una variabile casuale gamma. Ne segue che la distribuzione di C non dipende da θ , così che la (4.12) è una variabile cardine. ■

Nella proposizione seguente verrà mostrato come sia possibile costruire intervalli fiduciari per parametri espressi come funzioni monotone di altri parametri.

Proposizione 4.2.2 *Sia (C^-, C^+) un intervallo fiduciario di coefficiente $1 - \alpha$ per un parametro θ e sia $\lambda = h(\theta)$, con h funzione strettamente monotona. Se h è strettamente crescente [decrecente], allora $(h(C^-), h(C^+))$ [$(h(C^+), h(C^-))$] costituisce un intervallo fiduciario di coefficiente $1 - \alpha$ per λ .*

Dim. Poiché (C^-, C^+) è un intervallo fiduciario di coefficiente $1 - \alpha$ per θ , sussiste la (4.3). Essendo $\lambda = h(\theta)$, segue allora:

$$1 - \alpha = \begin{cases} P[h(C^-) < \lambda < h(C^+)] & \text{se } h \text{ è strettamente crescente,} \\ P[h(C^+) < \lambda < h(C^-)] & \text{se } h \text{ è strettamente decrecente,} \end{cases}$$

e quindi la tesi. ■

Nei paragrafi seguenti vedremo in dettaglio come costruire intervalli fiduciari per la stima di parametri relativi a campioni casuali estratti da popolazioni particolari.

4.3 Intervalli fiduciari per medie

Dato campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione normale di media μ e varianza σ^2 , determiniamo un intervallo fiduciario di coefficiente $1 - \alpha$ per la media. Occorre trattare separatamente i casi di varianza nota e di varianza incognita.

4.3.1 Varianza nota

Teorema 4.3.1 *Se \bar{X} è la media campionaria di un campione casuale di taglia n estratto da una popolazione normale di varianza σ^2 nota, un intervallo fiduciario di coefficiente $1 - \alpha$ per la media μ della popolazione è il seguente:*

$$\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right), \quad (4.13)$$

con $z_{\alpha/2}$ ricavabile dalla (1.26).

Dim. Consideriamo la variabile casuale

$$C = \gamma(X_1, X_2, \dots, X_n; \mu) = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}. \quad (4.14)$$

Osserviamo che essa è una variabile cardine poiché dipende dal campione (X_1, X_2, \dots, X_n) e dal parametro μ da stimare (avendo supposto che la varianza σ^2 è nota). Inoltre, per il Teorema 1.4.1, la media campionaria \bar{X} ha distribuzione normale di media μ e varianza σ^2/n . Ne segue che la variabile casuale C , definita dalla (4.14), ha distribuzione normale standard, così che la sua densità di probabilità non dipende dal parametro μ . La funzione γ è infine

4.3. INTERVALLI FIDUCIARI PER MEDIE

strettamente decrescente rispetto a μ . Allo scopo di determinare un intervallo fiduciario per μ possiamo far uso del metodo del cardine. Questo, come si è detto, richiede che si scelgano due valori α_1 e α_2 dipendenti da α e tali da aversi

$$P(\alpha_1 < C < \alpha_2) = 1 - \alpha.$$

Come espresso dalla (1.26), se Z è una variabile casuale normale standard con z_α si denota il reale tale che $P(Z > z_\alpha) = P(Z \geq z_\alpha) = \alpha$. Poiché la variabile C definita nella (4.14) ha proprio distribuzione normale standard, scegliamo $\alpha_1 = -z_{\alpha/2}$ e $\alpha_2 = z_{\alpha/2}$. In tal modo, essendo $-z_{\alpha/2} = z_{1-\alpha/2}$ (cfr. Proposizione 1.4.1), si ha:

$$\begin{aligned} P(\alpha_1 < C < \alpha_2) &= P(-z_{\alpha/2} < C < z_{\alpha/2}) \\ &= P(C > -z_{\alpha/2}) - P(C \geq z_{\alpha/2}) \\ &= P(C > z_{1-\alpha/2}) - P(C \geq z_{\alpha/2}) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha. \end{aligned}$$

Dalla (4.14) si ottiene poi:

$$P\left(-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

ossia:

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Le variabili casuali

$$C^- = \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad C^+ = \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

sono pertanto rispettivamente estremi sinistro e destro dell'intervallo fiduciario per la media μ , così che tale intervallo coincide con la (4.13). ■

È opportuno notare che l'ampiezza dell'intervallo fiduciario fornito dal Teorema 4.3.1 è data da

$$A = C^+ - C^- = z_{\alpha/2} \frac{2\sigma}{\sqrt{n}}. \quad (4.15)$$

Essa non dipende dal campione casuale (X_1, X_2, \dots, X_n) ed è inoltre strettamente decrescente in n , così che il grado di precisione della stima aumenta con la taglia del campione. L'espressione (4.15), oltre a evidenziare che A è direttamente proporzionale alla deviazione standard σ , suggerisce che se si considerano due campioni casuali di taglie n ed m estratti da popolazioni normali aventi stessa media μ ma varianze σ_1^2 e σ_2^2 differenti, per far sì che le ampiezze degli intervalli fiduciari per μ forniti dal Teorema 4.3.1 siano uguali deve risultare:

$$\frac{\sigma_1^2}{n} = \frac{\sigma_2^2}{m},$$

indicante che a una varianza maggiore deve corrispondere una maggiore taglia del campione.

Tabella 4.1: Dati, medie campionarie e intervalli fiduciari per l'Esempio 4.31.

i	x_i	\bar{x}_i	c_i^-	c_i^+
1	69.3	69.30	38.94	99.66
2	51.8	60.55	39.08	82.02
3	54.1	58.40	40.87	75.93
4	34.3	52.37	37.19	67.55
5	81.8	58.26	44.68	71.84
6	65.6	59.48	47.08	71.88
7	61.7	59.80	48.32	71.28
8	56.2	59.35	48.61	70.09
9	25.7	55.61	45.49	65.73
10	45.2	54.57	44.97	64.17
11	72.1	56.16	47.00	65.32
12	59.4	56.43	47.66	65.20
13	30.8	54.46	46.04	62.88
14	53.7	54.41	46.29	62.53
15	67.8	55.30	47.46	63.14
16	41.6	54.44	46.85	62.03
17	76.3	55.73	48.37	63.09
18	48.6	55.33	48.17	62.49
19	38.5	54.45	47.48	61.42
20	63.5	54.90	48.11	61.69

Esempio 4.3.1 Consideriamo la seguente realizzazione di un campione casuale di taglia $n = 20$ estratto da una popolazione normale avente varianza $\sigma^2 = 240$:

$$(69.3 \quad 51.8 \quad 54.1 \quad 34.3 \quad 81.8 \quad 65.6 \quad 61.7 \quad 56.2 \quad 25.7 \quad 45.2 \\ 72.1 \quad 59.4 \quad 30.8 \quad 53.7 \quad 67.8 \quad 41.6 \quad 76.3 \quad 48.6 \quad 38.5 \quad 63.5)$$

Vogliamo costruire un intervallo fiduciario di coefficiente $1 - \alpha = 0.95$ per l'incognito valore medio μ della popolazione. Dalla Tabella 2 dell'Appendice B si ricava $z_{\alpha/2} = z_{0.025} = 1.96$. Osservando che la media campionaria assume il valore $\bar{x} = 54.9$ si ottiene:

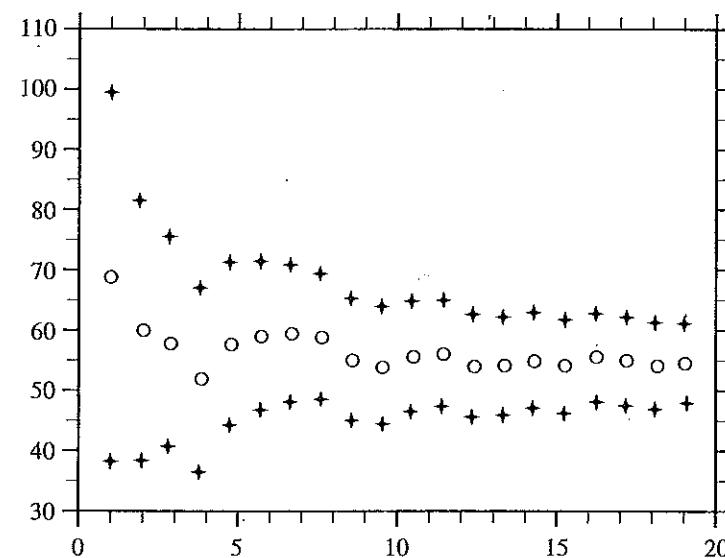
$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 54.9 - 1.96 \sqrt{\frac{240}{20}} = 48.11$$

$$\bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 54.9 + 1.96 \sqrt{\frac{240}{20}} = 61.69$$

da cui, per il Teorema 4.3.1, segue che la stima dell'intervallo fiduciario è $(48.11, 61.69)$.

È ora utile esaminare come cambia tale stima al variare del numero di dati osservati. A tal fine, per $i = 1, 2, \dots, 20$ nella Tabella 4.1 sono riportati i dati osservati x_i , le medie

4.3. INTERVALLI FIDUCIARI PER MEDIE

Figura 4.1: In relazione alla Tabella 4.1, al variare di i in neretto compaiono gli intervalli fiduciari mentre i cerchietti indicano le medie campionarie.

campionarie

$$\bar{x}_i = \frac{1}{i} \sum_{k=1}^i x_k$$

e gli estremi delle stime degli intervalli fiduciari corrispondenti ai dati (x_1, x_2, \dots, x_i) . Nella Figura 4.1 è rappresentato l'andamento delle medie campionarie e degli estremi delle stime degli intervalli fiduciari.

Si noti che in virtù del teorema centrale del limite quanto affermato dal Teorema 4.3.1 sussiste approssimativamente anche per campioni casuali di grandi dimensioni estratti da popolazioni (di varianza σ^2 nota) che non abbiano distribuzione normale. Nella pratica si fa uso del Teorema 4.3.1 per campioni casuali di taglia $n \geq 30$.

È opportuno notare che si possono costruire più intervalli fiduciari di coefficiente $1 - \alpha$ per le medie. Nel teorema che segue ne viene indicata un'intera famiglia, uno dei cui membri è caratterizzato da massima precisione di stima, ossia da ampiezza minima.

Teorema 4.3.2 Se \bar{X} è la media campionaria di un campione casuale di taglia n estratto da una popolazione normale di media μ incognita e varianza σ^2 nota, fissato un arbitrario reale

α , con $0 < \alpha < 1$, ogni intervallo del tipo

$$\left(\bar{X} - z_v \frac{\sigma}{\sqrt{n}}, \bar{X} - z_{v+1-\alpha} \frac{\sigma}{\sqrt{n}} \right) \quad (4.16)$$

è un intervallo fiduciario di coefficiente $1 - \alpha$ per μ se $v \in (0, 1)$. Di questi, l'intervallo

$$\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right), \quad (4.17)$$

corrispondente a $v = \alpha/2$, è quello avente ampiezza minima.

Dim. Comunque si fissi un intervallo $(z_{v+1-\alpha}, z_v)$, con $v \in (0, 1)$, facendo uso della definizione (1.26) di quantile superiore si ha:

$$\begin{aligned} P\left(z_{v+1-\alpha} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_v\right) &= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > z_{v+1-\alpha}\right) - P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq z_v\right) \\ &= v + 1 - \alpha - v = 1 - \alpha, \end{aligned}$$

da cui si ricava:

$$P\left(\bar{X} - z_v \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - z_{v+1-\alpha} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

La (4.16) individua dunque una famiglia di intervalli fiduciari di coefficiente $1 - \alpha$ per μ . L'ampiezza di tali intervalli, in virtù della definizione (4.4), è data da

$$A(v) = (z_v - z_{v+1-\alpha}) \frac{\sigma}{\sqrt{n}}.$$

Per minimizzare $A(v)$ imponiamo l'annullarsi della derivata

$$\frac{d}{dv} A(v) = \frac{d}{dv} (z_v - z_{v+1-\alpha}) \frac{\sigma}{\sqrt{n}}. \quad (4.18)$$

Consideriamo a tal fine l'identità

$$\int_{z_{v+1-\alpha}}^{z_v} f(x) dx = 1 - \alpha, \quad (4.19)$$

dove $f(x)$ è la densità normale standard. Derivando ambo i membri della (4.19) rispetto a v si ottiene:

$$f(z_v) \frac{d}{dv} z_v - f(z_{v+1-\alpha}) \frac{d}{dv} z_{v+1-\alpha} = 0,$$

così che la (4.18) diventa:

$$\frac{d}{dv} A(v) = \frac{d}{dv} z_v \left[1 - \frac{f(z_v)}{f(z_{v+1-\alpha})} \right] \frac{\sigma}{\sqrt{n}}. \quad (4.20)$$

Dimostriamo ora che la derivata di z_v rispetto a v è negativa. Essendo $\Phi(z_v) = 1 - v$, risulta $z_v = \Phi^{-1}(1 - v)$, dove Φ^{-1} denota l'inversa della funzione di distribuzione di una variabile

4.3. INTERVALLI FIDUCIARI PER MEDIE

casuale normale standard. Poiché Φ^{-1} è una funzione crescente, $z_v \equiv \Phi^{-1}(1 - v)$ è decrescente in v , così che

$$\frac{d}{dv} z_v < 0. \quad (4.21)$$

Dalle (4.20) e (4.21) segue che la derivata di $A(v)$ si annulla per $f(z_v) = f(z_{v+1-\alpha})$. Essendo $f(x)$ una funzione pari, affinché tale uguaglianza sia soddisfatta dovrebbe risultare $z_v = z_{v+1-\alpha}$ oppure $z_v = -z_{v+1-\alpha}$. Ma in virtù della (4.19) e dell'ipotesi che α è interno all'intervallo $(0, 1)$, non può aversi $z_v = z_{v+1-\alpha}$; l'equazione $z_v = -z_{v+1-\alpha}$ è invece soddisfatta per $v = \alpha/2$, avendosi $z_v = -z_{1-\alpha}$ in conseguenza della (1.28). La derivata di $A(v)$ è pertanto nulla in $v = \alpha/2$. Per mostrare che questo è un punto di massimo esaminiamo la derivata di $A(v)$ in un suo intorno. Dalla (4.20) per ogni fissato reale ϵ segue

$$\frac{d}{dv} A(v) \Big|_{v=\frac{\alpha}{2}+\epsilon} = \frac{d}{dv} z_v \Big|_{v=\frac{\alpha}{2}+\epsilon} \left[1 - \frac{f(z_{\frac{\alpha}{2}+\epsilon})}{f(z_{1-\frac{\alpha}{2}-\epsilon})} \right] \frac{\sigma}{\sqrt{n}}. \quad (4.22)$$

Essendo

$$f(z_{\frac{\alpha}{2}+\epsilon}) = f(-z_{\frac{\alpha}{2}-\epsilon}) = f(z_{\frac{\alpha}{2}-\epsilon}),$$

ed osservando che dall'espressione (1.20) della densità normale standard si ricava facilmente l'identità

$$\frac{f(z_{\frac{\alpha}{2}+\epsilon})}{f(z_{\frac{\alpha}{2}-\epsilon})} = \exp\left(-2\epsilon z_{\frac{\alpha}{2}}\right), \quad f(z_{\frac{\alpha}{2}-\epsilon}) = \frac{1}{\sqrt{2\pi}} e^{-z_{\frac{\alpha}{2}}^2}$$

la (4.22) diventa

$$\frac{d}{dv} A(v) \Big|_{v=\frac{\alpha}{2}+\epsilon} = \frac{d}{dv} z_v \Big|_{v=\frac{\alpha}{2}+\epsilon} \left[1 - \exp\left(-2\epsilon z_{\frac{\alpha}{2}}\right) \right] \frac{\sigma}{\sqrt{n}}. \quad (4.23)$$

Osserviamo che è $z_{\frac{\alpha}{2}} > 0$ in quanto è $0 < \alpha < 1$; in virtù della (4.21) si ricava che la (4.23) è negativa [positiva] per ϵ negativo [positivo]. Da ciò si trae che $v = \alpha/2$ è un punto di minimo per $A(v)$. Ponendo $v = \alpha/2$ nella (4.16) segue quindi che l'intervallo fiduciario (4.17) è quello che, per il fissato α , ha ampiezza minima, il che implica massima precisione della stima. ■

Si noti che tra gli intervalli fiduciari della media μ aventi forma (4.16) l'intervallo (4.17) risulta essere l'unico il cui centro coincide con la media campionaria \bar{X} , che sappiamo essere uno stimatore corretto di μ . Esso, inoltre, è l'unico intervallo per il quale risulta

$$P(C^- < \mu < \bar{X}) = P(\bar{X} < \mu < C^+);$$

per questo motivo viene detto *intervallo centrale*.

Consideriamo ora nuovamente l'intervallo fiduciario di coefficiente $1 - \alpha$ per la media μ dato dalla (4.16). Facendo uso delle (1.26) e (1.28), si ricava che l'intervallo (4.16) per $v = 0$ diventa

$$\left(\bar{X} - z_0 \frac{\sigma}{\sqrt{n}}, \bar{X} - z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right) \equiv \left(-\infty, \bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}} \right),$$

mentre per $v = \alpha$ fornisce

$$\left(\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}, \bar{X} - z_1 \frac{\sigma}{\sqrt{n}} \right) \equiv \left(\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty \right).$$

In entrambi i casi si è dunque ottenuto un intervallo fiduciario unidirezionale.

Val la pena sottolineare come l'intervallo (4.17) del Teorema 4.3.2 coincida con l'intervallo (4.13) del Teorema 4.3.1.

4.3.2 Varianza incognita

Teorema 4.3.3 *Se \bar{X} e S sono la media campionaria e la deviazione standard campionaria di un campione casuale di taglia n estratto da una popolazione normale di varianza incognita, un intervallo fiduciario di coefficiente $1 - \alpha$ per l'incognito valore medio μ della popolazione risulta essere:*

$$\left(\bar{X} - t_{\alpha/2; n-1} \frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2; n-1} \frac{S}{\sqrt{n}} \right),$$

dove $t_{\alpha; v}$ è definito nella (2.22).

Dim. Consideriamo la variabile casuale

$$C = \gamma(X_1, X_2, \dots, X_n; \mu) = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (4.24)$$

dove μ è il valore medio incognito da stimare e S è la deviazione standard campionaria. Essa è una variabile cardine poiché C dipende dal campione casuale (X_1, X_2, \dots, X_n) e dal parametro μ da stimare. Per il Teorema 2.2.2 C ha distribuzione di Student con $n - 1$ gradi di libertà, così che la sua densità di probabilità non dipende dal parametro μ . La funzione γ che compare nella (4.24) è poi strettamente decrescente rispetto a μ . Come prescritto dal metodo del cardine, scegliamo due valori α_1 e α_2 dipendenti da α e tali da aversi

$$P(\alpha_1 < C < \alpha_2) = 1 - \alpha.$$

Ricordiamo che se T è una variabile casuale di Student con v gradi di libertà, con $t_{\alpha; v}$ si denota il reale tale da risultare $P(T > t_{\alpha; v}) = \alpha$. Poiché la variabile C definita dalla (4.24) ha distribuzione di Student con $n - 1$ gradi di libertà, scegliamo $\alpha_1 = -t_{\alpha/2; n-1}$ e $\alpha_2 = t_{\alpha/2; n-1}$. Essendo

$$P(C > t_{\alpha/2; n-1}) = \frac{\alpha}{2}, \quad P(C < -t_{\alpha/2; n-1}) = \frac{\alpha}{2},$$

ricordando la (4.24) si trae:

$$P\left(-t_{\alpha/2; n-1} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2; n-1}\right) = 1 - \alpha,$$

ossia:

$$P\left(\bar{X} - t_{\alpha/2; n-1} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2; n-1} \frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Pertanto

$$C^- = \bar{X} - t_{\alpha/2; n-1} \frac{S}{\sqrt{n}}, \quad C^+ = \bar{X} + t_{\alpha/2; n-1} \frac{S}{\sqrt{n}}$$

sono rispettivamente estremi sinistro e destro dell'intervallo fiduciario. ■

4.4 DIFFERENZE TRA MEDIE

Esempio 4.3.2 Consideriamo la seguente realizzazione di un campione casuale di taglia $n = 16$ estratto da una popolazione normale a varianza incognita:

$$(35.7 \ 40.8 \ 43.3 \ 48.5 \ 51.3 \ 54.7 \ 56.2 \ 58.3 \\ 59.7 \ 61.4 \ 63.1 \ 66.2 \ 69.6 \ 72.5 \ 74.6 \ 80.1).$$

Media e varianza campionarie assumono rispettivamente i valori $\bar{x} = 58.5$ e $s^2 = 158.07$. Vogliamo costruire un intervallo fiduciario di coefficiente $1 - \alpha = 0.95$ per il valore medio μ della popolazione. Consultando la Tabella 4 dell'Appendice B si ottiene $t_{\alpha/2; n-1} = t_{0.025; 15} = 2.131$. Poiché risulta

$$\bar{x} - t_{\alpha/2; n-1} \frac{s}{\sqrt{n}} = 58.5 - 2.131 \sqrt{\frac{158.07}{16}} = 51.8 \\ \bar{x} + t_{\alpha/2; n-1} \frac{s}{\sqrt{n}} = 58.5 + 2.131 \sqrt{\frac{158.07}{16}} = 65.2,$$

dal Teorema 4.3.3 segue che (51.8, 65.2) costituisce la stima dell'intervallo fiduciario per la media μ . ♦

4.4 Differenze tra medie

Siano $(X_{11}, X_{12}, \dots, X_{1n})$ un campione estratto da una popolazione normale di media μ_1 e varianza σ_1^2 e $(X_{21}, X_{22}, \dots, X_{2m})$ un campione estratto da una popolazione normale avente valore medio μ_2 e varianza σ_2^2 . Supponiamo che le $n + m$ variabili casuali che costituiscono i due campioni siano collettivamente indipendenti. Ricordiamo che, per il Teorema 1.4.2, la differenza tra le medie campionarie $\bar{X}_1 - \bar{X}_2$ è una variabile casuale normale di media $\mu_1 - \mu_2$ e varianza $\sigma_1^2/n + \sigma_2^2/m$. Vogliamo qui determinare un intervallo fiduciario di coefficiente $1 - \alpha$ per la differenza $\mu_1 - \mu_2$. Occorre trattare separatamente i casi di varianze note e di varianze incognite.

4.4.1 Varianze note

Teorema 4.4.1 *Se \bar{X}_1 e \bar{X}_2 sono le medie campionarie di due campioni casuali indipendenti di taglie n e m estratti rispettivamente da popolazioni normali di varianze σ_1^2 e σ_2^2 note, un intervallo fiduciario di coefficiente $1 - \alpha$ per la differenza $\mu_1 - \mu_2$ dei valori medi delle popolazioni è dato da*

$$\left(\bar{X}_1 - \bar{X}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}, \bar{X}_1 - \bar{X}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \right).$$

Dim. Consideriamo la variabile casuale

$$C = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\delta}, \quad (4.25)$$

dove

$$\delta = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

denota la deviazione standard di $\bar{X}_1 - \bar{X}_2$. Per ipotesi, le varianze σ_1^2 e σ_2^2 sono note; la (4.25) dipende quindi dai campioni $(X_{11}, X_{12}, \dots, X_{1n})$ e $(X_{21}, X_{22}, \dots, X_{2m})$ e dalla differenza $\mu_1 - \mu_2$ da stimare, ed è pertanto una variabile cardine. Inoltre, poiché $\bar{X}_1 - \bar{X}_2$ è una variabile casuale normale di media $\mu_1 - \mu_2$ e varianza $\delta^2 = \sigma_1^2/n + \sigma_2^2/m$, C ha distribuzione normale standard, così che la sua densità di probabilità non dipende dalla differenza $\mu_1 - \mu_2$. Infine C , riguardata come funzione della differenza $\mu_1 - \mu_2$, è strettamente decrescente. Per applicare il metodo del cardine, sceglieremo due valori α_1 e α_2 dipendenti da α e tali che

$$P(\alpha_1 < C < \alpha_2) = 1 - \alpha.$$

In maniera analoga a quanto indicato per il Teorema 4.3.1, poiché C ha distribuzione normale standard sceglieremo $\alpha_1 = -z_{\alpha/2}$ e $\alpha_2 = z_{\alpha/2}$. Ricordando la (4.25) si ha:

$$P\left(-z_{\alpha/2} < \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\delta} < z_{\alpha/2}\right) = 1 - \alpha,$$

ossia:

$$P\left(\bar{X}_1 - \bar{X}_2 - z_{\alpha/2} \delta < \mu_1 - \mu_2 < \bar{X}_1 - \bar{X}_2 + z_{\alpha/2} \delta\right) = 1 - \alpha.$$

Le statistiche C^- e C^+ , rispettivamente estremi sinistro e destro dell'intervallo fiduciario, sono quindi le seguenti:

$$C^- = \bar{X}_1 - \bar{X}_2 - z_{\alpha/2} \delta, \quad C^+ = \bar{X}_1 - \bar{X}_2 + z_{\alpha/2} \delta.$$

In virtù del teorema centrale del limite l'intervallo fiduciario di cui al Teorema 4.4.1 è utilizzabile in via approssimata anche per campioni casuali di taglia elevata estratti da popolazioni che non abbiano distribuzione normale. Nella pratica si fa uso del Teorema 4.4.1 per $n, m \geq 30$.

Esempio 4.4.1 Una ditta produce due tipi di lampadine che vengono sottoposte a test. Un campione di 35 lampadine del primo tipo ha funzionato in media per 408 ore mentre un campione di 45 lampadine del secondo tipo è durato 396 ore in media. Supponiamo che le deviazioni standard σ_1 e σ_2 delle durate per i due tipi di lampadine siano note, e che risultino $\sigma_1 = 25$ ore, $\sigma_2 = 20$ ore. Vogliamo costruire un intervallo fiduciario di coefficiente $1 - \alpha = 0.94$ per la differenza $\mu_1 - \mu_2$ tra i tempi medi di durata dei due tipi di lampadine. Anzitutto osserviamo che dalla Tabella 1 dell'Appendice B si deduce $z_{\alpha/2} = z_{0.03} = 1.88$. Anche se ci è ignota la distribuzione della durata dei due tipi di lampadine, poiché i campioni sono di taglia elevata possiamo ricorrere al Teorema 4.4.1 ed ottenere così la stima desiderata. Risultando

$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \delta = 408 - 396 - 1.88 \sqrt{\frac{625}{35} + \frac{400}{45}} = 2.28$$

$$\bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \delta = 408 - 396 + 1.88 \sqrt{\frac{625}{35} + \frac{400}{45}} = 21.72,$$

la stima dell'intervallo fiduciario, in ore, è $(2.28, 21.72)$. Poiché essa non include il valore zero, con "fiducia" $1 - \alpha = 0.94$, quindi molto elevata, le durate medie dei due tipi di lampadine differiscono in modo significativo. ♦

4.4.2 Varianze incognite

Consideriamo ora il caso di due popolazioni normali aventi varianze incognite che sono uguali:²

$$\sigma_1^2 = \sigma_2^2 = \sigma^2.$$

Teorema 4.4.2 Siano \bar{X}_1 e \bar{X}_2 le medie campionarie di due campioni casuali indipendenti di taglie n ed m estratti da popolazioni normali di varianze uguali e incognite, e sia S_p la radice quadrata della varianza campionaria pesata

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}. \quad (4.26)$$

Un intervallo fiduciario di coefficiente $1 - \alpha$ per la differenza $\mu_1 - \mu_2$ degli incogniti valori medi delle popolazioni è dato da

$$(\bar{X}_1 - \bar{X}_2 - t_{\alpha/2;n+m-2} S_p \delta, \bar{X}_1 - \bar{X}_2 + t_{\alpha/2;n+m-2} S_p \delta), \quad (4.27)$$

dove

$$\delta = \sqrt{\frac{1}{n} + \frac{1}{m}}.$$

Dim. La variabile casuale

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma \delta} \quad (4.28)$$

ha distribuzione normale standard. Inoltre, per il Teorema 2.1.4 le variabili casuali indipendenti

$$\frac{n-1}{\sigma^2} S_1^2, \quad \frac{m-1}{\sigma^2} S_2^2$$

hanno distribuzione chi-quadrato con rispettivamente $n-1$ e $m-1$ gradi di libertà. La loro somma

$$Y = \frac{n-1}{\sigma^2} S_1^2 + \frac{m-1}{\sigma^2} S_2^2 = \frac{n+m-2}{\sigma^2} S_p^2 \quad (4.29)$$

ha allora distribuzione chi-quadrato con $n+m-2$ gradi di libertà. Si potrebbe poi dimostrare che le variabili casuali Z e Y , definite dalle (4.28) e (4.29), sono indipendenti. Per il Teorema 2.2.1 la variabile casuale

$$C = \frac{Z}{\sqrt{Y/(n+m-2)}} = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p \delta} \quad (4.30)$$

²A questo caso ci si può sempre ricondurre con un'opportuna trasformazione di una delle variabili casuali genetiche nonché del corrispondente campione e della sua realizzazione osservata.

ha dunque distribuzione di Student con $n+m-2$ gradi di libertà, così che la sua densità di probabilità non dipende da $\mu_1 - \mu_2$. La (4.30) mostra che C è una funzione strettamente decrescente rispetto alla differenza $\mu_1 - \mu_2$; pertanto essa è una variabile cardine giacché dipende dai campioni casuali $(X_{11}, X_{12}, \dots, X_{1n})$ e $(X_{21}, X_{22}, \dots, X_{2m})$ e dalla differenza $\mu_1 - \mu_2$ da stimare. Possiamo così applicare il metodo del cardine per determinare un intervallo fiduciario per $\mu_1 - \mu_2$. Dobbiamo dunque scegliere due valori α_1 e α_2 dipendenti da α e tali da avversi

$$P(\alpha_1 < C < \alpha_2) = 1 - \alpha.$$

In modo analogo a quanto effettuato nella dimostrazione del Teorema 4.3.3, poiché C ha distribuzione di Student con $n+m-2$ gradi di libertà poniamo $\alpha_1 = -t_{\alpha/2;n+m-2}$ e $\alpha_2 = t_{\alpha/2;n+m-2}$. Otteniamo così:

$$P\left(-t_{\alpha/2;n+m-2} < \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{s_p \delta} < t_{\alpha/2;n+m-2}\right) = 1 - \alpha,$$

che implica la (4.27). ■

Si noti che la varianza campionaria pesata (4.26) coincide con lo stimatore corretto $S_{(a)}^2$ di σ^2 a varianza minima, ossia con $\alpha = (n-1)/(n+m-2)$, analizzato nell'Esempio 3.3.10.

Esempio 4.4.2 Confrontando il contenuto di nicotina di due tipi di sigarette si è riscontrato che dieci sigarette di tipo A hanno in media un contenuto di 3.1 mg di nicotina con una deviazione standard campionaria di 0.5 mg, mentre otto sigarette di tipo B hanno in media un contenuto di 2.7 mg di nicotina con una deviazione standard campionaria di 0.7 mg. Supponendo che i dati costituiscono campioni casuali di due popolazioni normali di uguali varianze, si vuole costruire un intervallo fiduciario di coefficiente $1 - \alpha = 0.95$ per la differenza del contenuto medio di nicotina dei due tipi di sigarette. Dai dati forniti si ricava:

$$\begin{aligned} n &= 10; \quad \bar{x}_1 = 3.1; \quad s_1 = 0.5; \\ m &= 8; \quad \bar{x}_2 = 2.7; \quad s_2 = 0.7. \end{aligned}$$

Essendo $\alpha = 0.05$ e $n+m-2 = 16$, dalla Tabella 4 dell'Appendice B si ottiene $t_{\alpha/2;n+m-2} = t_{0.025;16} = 2.12$. Il valore assunto dalla statistica s_p è poi:

$$s_p = \sqrt{\frac{9 \cdot 0.25 + 7 \cdot 0.49}{16}} = 0.596.$$

Poiché risulta

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2;n+m-2} s_p \delta = 3.1 - 2.7 - 2.12 \cdot 0.596 \sqrt{\frac{1}{10} + \frac{1}{8}} = -0.2$$

$$\bar{x}_1 - \bar{x}_2 + t_{\alpha/2;n+m-2} s_p \delta = 3.1 - 2.7 + 2.12 \cdot 0.596 \sqrt{\frac{1}{10} + \frac{1}{8}} = 1$$

attraverso il Teorema 4.4.2 si giunge alla stima dell'intervallo fiduciario $(-0.2, 1)$. Quindi con un grado di fiducia pari a 0.95 la differenza dei valori medi di nicotina deve ritenersi compresa nell'intervallo $(-0.2, 1)$. Poiché in questo intervallo è compreso il valore zero, non è possibile concludere che sono significativamente diversi i contenuti di nicotina dei due tipi di sigarette. ♦

4.5 Intervalli fiduciari per varianze

Dato un campione casuale estratto da una popolazione normale di media μ e varianza σ^2 , esaminiamo come sia possibile determinare un intervallo fiduciario di coefficiente $1 - \alpha$ per stimare la varianza σ^2 . Distinguiamo i casi di media nota e media incognita.

4.5.1 Media nota

Sussiste il seguente teorema:

Teorema 4.5.1 *Dato un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione normale di media μ nota, un intervallo fiduciario di coefficiente $1 - \alpha$ per la varianza σ^2 della popolazione è il seguente:*

$$\left(\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{\alpha/2;n}^2}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{1-\alpha/2;n}^2} \right).$$

Dim. Consideriamo la variabile casuale

$$C = \gamma(X_1, X_2, \dots, X_n; \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2. \quad (4.31)$$

Questa è una variabile cardine giacché dipende dal campione casuale e dal parametro σ^2 da stimare. Inoltre, essendo somma dei quadrati di n variabili indipendenti normali standard (cfr. Corollario 2.1.1), essa ha distribuzione chi-quadrato con n gradi di libertà, così che la sua funzione di distribuzione non dipende dal parametro σ^2 . La funzione γ , espressa dalla (4.31), è inoltre strettamente decrescente rispetto a σ^2 . È allora possibile fare uso del metodo del cardine per determinare un intervallo fiduciario di coefficiente $1 - \alpha$ per σ^2 . Scegliamo a tale scopo due valori α_1 e α_2 dipendenti da α e tali che

$$P(\alpha_1 < C < \alpha_2) = 1 - \alpha.$$

Ricordiamo che se X è una variabile casuale a distribuzione chi-quadrato con n gradi di libertà, con $\chi_{\alpha;n}^2$ si denota quel reale positivo tale da avversi $P(X > \chi_{\alpha;n}^2) = \alpha$. Poiché C ha distribuzione chi-quadrato con n gradi di libertà, scegliamo $\alpha_1 = \chi_{1-\alpha/2;n}^2$ e $\alpha_2 = \chi_{\alpha/2;n}^2$. In tal modo risulta:

$$P(C > \chi_{\alpha/2;n}^2) = P(0 < C < \chi_{1-\alpha/2;n}^2) = \frac{\alpha}{2},$$

da cui si ottiene:

$$P(\chi_{1-\alpha/2;n}^2 < C < \chi_{\alpha/2;n}^2) = P(C > \chi_{1-\alpha/2;n}^2) - P(C > \chi_{\alpha/2;n}^2) = 1 - \alpha,$$

ovvero:

$$P\left[\chi_{1-\alpha/2;n}^2 < \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 < \chi_{\alpha/2;n}^2\right] = 1 - \alpha.$$

Di qui segue:

$$P\left[\frac{\sum_{i=1}^n(X_i-\mu)^2}{\chi_{\alpha/2;n}^2} < \sigma^2 < \frac{\sum_{i=1}^n(X_i-\mu)^2}{\chi_{1-\alpha/2;n}^2}\right] = 1 - \alpha.$$

Pertanto le statistiche C^- e C^+ , rispettivamente estremi sinistro e destro dell'intervallo fiduciario per σ^2 , hanno le seguenti espressioni:

$$C^- = \frac{\sum_{i=1}^n(X_i-\mu)^2}{\chi_{\alpha/2;n}^2}, \quad C^+ = \frac{\sum_{i=1}^n(X_i-\mu)^2}{\chi_{1-\alpha/2;n}^2}.$$

Mostreremo ora che, nel caso di media nota, è possibile ricavare anche un intervallo fiduciario per la deviazione standard σ a partire da un intervallo fiduciario per la varianza σ^2 .

Corollario 4.5.1 *Dato un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione normale di media μ nota, un intervallo fiduciario di coefficiente $1 - \alpha$ per la deviazione standard σ della popolazione risulta essere il seguente:*

$$\left(\sqrt{\frac{\sum_{i=1}^n(X_i-\mu)^2}{\chi_{\alpha/2;n}^2}}, \sqrt{\frac{\sum_{i=1}^n(X_i-\mu)^2}{\chi_{1-\alpha/2;n}^2}} \right).$$

Dim. Segue direttamente dalla Proposizione 4.2.2 e dal Teorema 4.5.1, osservando che $h(x) = \sqrt{x}$ è una funzione strettamente crescente per $x \geq 0$. ■

Esempio 4.5.1 Con riferimento all'Esempio 2.1.1, riconsideriamo la realizzazione osservata

$$(1.05, 1.07, 1.12, 0.92, 0.98, 1.07, 1.11, 1.02, 0.98, 0.99, 1.02, 1.12, 1.05, 0.99, 0.94, 1.05, 0.99, 1.11, 0.97, 1.12),$$

relativa ad un campione casuale $(X_1, X_2, \dots, X_{20})$ estratto da una popolazione normale di media $\mu = 1$ e varianza σ^2 incognita descrivente lo spessore dei filamenti di 20 lampadine prodotte in una fabbrica di componenti elettrici. Vogliamo determinare un intervallo fiduciario di coefficiente $1 - \alpha = 0.95$ per la deviazione standard σ del campione. Notiamo anzitutto che si ha

$$\sum_{i=1}^{20}(x_i - 1)^2 = 0.0975.$$

Inoltre, dalla consultazione della Tabella 3 dell'Appendice B constatiamo che risulta $\chi_{\alpha/2;n}^2 = \chi_{0.025;20}^2 = 34.170$ e $\chi_{1-\alpha/2;n}^2 = \chi_{0.975;20}^2 = 9.591$. Quindi, in virtù del Teorema 4.5.1, la stima dell'intervallo fiduciario per la varianza σ^2 è

$$\begin{aligned} \left(\frac{\sum_{i=1}^{20}(x_i - 1)^2}{\chi_{0.025;20}^2}, \frac{\sum_{i=1}^{20}(x_i - 1)^2}{\chi_{0.975;20}^2} \right) &= \left(\frac{0.0975}{34.170}, \frac{0.0975}{9.591} \right) \\ &= (2.85 \cdot 10^{-3}, 0.0102). \end{aligned}$$

4.5. INTERVALLI FIDUCIARI PER VARIANZE

Per il Corollario 4.5.1 la stima dell'intervallo fiduciario per la deviazione standard σ è quindi $(0.0534, 0.1008)$. Si noti che la stima ricavata induce a ritenere che con attendibilità o "fiducia" $1 - \alpha = 0.95$, la deviazione standard σ sia superiore al valore critico 0.05. Pertanto nell'ipotesi $\mu = 1$ la stima dell'intervallo fiduciario per σ conduce alla conclusione che il processo di realizzazione dei filamenti non può ritenersi accettabile. Tale risultato è in accordo con quello cui si è pervenuti adottando il criterio dell'Esempio 2.1.1. ♦

4.5.2 Media incognita

Il risultato fornito dal Teorema 4.5.1 non è applicabile nei casi, particolarmente realistici, in cui la media della popolazione è incognita. Sorge allora l'esigenza di determinare un intervallo fiduciario per la varianza σ^2 quando anche la media μ è incognita. Al fine di fare ancora uso del metodo del cardine, riconsideriamo la variabile cardine (4.31) introdotta nel caso di media μ nota, sostituendo però in essa μ con il suo stimatore dato dalla media campionaria. Si ottiene così la variabile

$$\frac{1}{\sigma^2} \sum_{i=1}^n(X_i - \bar{X})^2 \equiv \frac{n-1}{\sigma^2} S^2$$

che interviene nel teorema che segue.

Teorema 4.5.2 *Se S^2 è la varianza campionaria di un campione casuale di taglia n estratto da una popolazione normale di media incognita, un intervallo fiduciario di coefficiente $1 - \alpha$ per la varianza σ^2 della popolazione è il seguente:*

$$\left(\frac{n-1}{\chi_{\alpha/2;n-1}^2} S^2, \frac{n-1}{\chi_{1-\alpha/2;n-1}^2} S^2 \right).$$

Dim. La variabile casuale

$$C = \gamma(X_1, X_2, \dots, X_n; \sigma^2) = \frac{n-1}{\sigma^2} S^2 \quad (4.32)$$

è una variabile cardine poiché dipende dal campione casuale e dal parametro σ^2 da stimare. Inoltre, come mostrato nel Teorema 2.1.4, C ha distribuzione chi-quadrato con $n-1$ gradi di libertà, così che la sua funzione di distribuzione non dipende dal parametro σ^2 . La funzione γ definita nella (4.32) è poi strettamente decrescente rispetto a σ^2 . Possiamo allora far uso del metodo del cardine per determinare un intervallo fiduciario di coefficiente $1 - \alpha$ per la varianza σ^2 . Sceglieremo a tal fine due valori α_1 e α_2 dipendenti da α e tali da aversi

$$P(\alpha_1 < C < \alpha_2) = 1 - \alpha.$$

Poiché, come si è già notato, C ha distribuzione chi-quadrato con $n-1$ gradi di libertà, in analogia col Teorema 4.5.1 sceglieremo $\alpha_1 = \chi_{1-\alpha/2;n-1}^2$ e $\alpha_2 = \chi_{\alpha/2;n-1}^2$. In tal modo si ha:

$$P\left(\chi_{1-\alpha/2;n-1}^2 < \frac{n-1}{\sigma^2} S^2 < \chi_{\alpha/2;n-1}^2\right) = 1 - \alpha.$$

Di qui segue:

$$P\left(\frac{n-1}{\chi_{\alpha/2;n-1}^2} S^2 < \sigma^2 < \frac{n-1}{\chi_{1-\alpha/2;n-1}^2} S^2\right) = 1 - \alpha.$$

Le statistiche C^- e C^+ , rispettivamente estremi sinistro e destro dell'intervallo fiduciario per σ^2 , sono allora così specificate:

$$C^- = \frac{n-1}{\chi_{\alpha/2;n-1}^2} S^2, \quad C^+ = \frac{n-1}{\chi_{1-\alpha/2;n-1}^2} S^2.$$

Analogamente al caso di media μ nota, dal Teorema 4.5.2 si ricava un intervallo fiduciario per la deviazione standard σ quando la media è incognita.

Corollario 4.5.2 Se S^2 è la varianza campionaria di un campione casuale di taglia n estratto da una popolazione normale di media incognita, un intervallo fiduciario di coefficiente $1 - \alpha$ per la deviazione standard σ della popolazione è il seguente:

$$\left(\sqrt{\frac{n-1}{\chi_{\alpha/2;n-1}^2} S^2}, \sqrt{\frac{n-1}{\chi_{1-\alpha/2;n-1}^2} S^2} \right).$$

Esempio 4.5.2 Dato un campione di taglia $n = 16$ estratto da una popolazione normale di media e varianza incognite, si voglia determinare un intervallo fiduciario di coefficiente $1 - \alpha = 0.99$ per la varianza σ^2 della popolazione sapendo che in base ad una realizzazione del campione risulta $s = 2.2$. Consultando la Tabella 3 dell'Appendice B osserviamo che si ha $\chi_{\alpha/2;n-1}^2 = \chi_{0.005;15}^2 = 32.801$ e $\chi_{1-\alpha/2;n-1}^2 = \chi_{0.995;15}^2 = 4.601$, così che

$$\frac{n-1}{\chi_{\alpha/2;n-1}^2} s^2 = \frac{15(2.2)^2}{32.801} = 2.21$$

$$\frac{n-1}{\chi_{1-\alpha/2;n-1}^2} s^2 = \frac{15(2.2)^2}{4.601} = 15.78.$$

Dal Teorema 4.5.2 segue dunque che $(2.21, 15.78)$ è la stima dell'intervallo fiduciario. ◆

4.6 Rapporti di varianze

Consideriamo due campioni casuali indipendenti estratti da popolazioni normali di varianze σ_1^2 e σ_2^2 con la finalità di determinare un intervallo fiduciario di coefficiente $1 - \alpha$ per il rapporto σ_1^2/σ_2^2 .

Teorema 4.6.1 Se S_1^2 e S_2^2 denotano le varianze campionarie di due campioni casuali indipendenti di taglie n e m estratti da popolazioni normali, un intervallo fiduciario di coefficiente $1 - \alpha$ per il rapporto σ_1^2/σ_2^2 tra le incognite varianze è il seguente:

$$\left(\frac{1}{F_{\alpha/2;n-1,m-1}} \frac{S_1^2}{S_2^2}, F_{\alpha/2;m-1,n-1} \frac{S_1^2}{S_2^2} \right).$$

4.6. RAPPORTI DI VARIANZE

Dim. Nel Teorema 2.3.2 si è visto che se S_1^2 e S_2^2 sono le varianze campionarie di due campioni indipendenti di taglie n e m estratti da popolazioni normali di varianze σ_1^2 e σ_2^2 , la variabile

$$C = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \quad (4.33)$$

ha distribuzione di Fisher con $n - 1$ e $m - 1$ gradi di libertà, così che la sua funzione di distribuzione non dipende dal rapporto σ_1^2/σ_2^2 . Poiché, inoltre, C dipende dai due campioni casuali e dal rapporto σ_1^2/σ_2^2 da stimare ed è strettamente decrescente rispetto a σ_1^2/σ_2^2 , concludiamo che essa è una variabile cardine. Possiamo allora far uso del metodo del cardine per determinare un intervallo fiduciario di coefficiente $1 - \alpha$ per il rapporto di varianze σ_1^2/σ_2^2 . Sceglieremo a tal fine due valori α_1 e α_2 dipendenti da α e tali che risultino

$$P(\alpha_1 < C < \alpha_2) = 1 - \alpha.$$

Ricordiamo che se X è una variabile casuale avente distribuzione di Fisher con v_1 e v_2 gradi di libertà, con $F_{\alpha;v_1,v_2}$ si denota quel reale tale da aversi $P(X \geq F_{\alpha;v_1,v_2}) = \alpha$. Poiché C ha distribuzione di Fisher con $n - 1$ e $m - 1$ gradi di libertà, poniamo $\alpha_1 = F_{1-\alpha/2;n-1,m-1}$ e $\alpha_2 = F_{\alpha/2;n-1,m-1}$. Essendo C a valori positivi, si ha:

$$P(0 < C < F_{1-\alpha/2;n-1,m-1}) = P(C > F_{\alpha/2;n-1,m-1}) = \frac{\alpha}{2},$$

da cui segue:

$$\begin{aligned} P(F_{1-\alpha/2;n-1,m-1} < C < F_{\alpha/2;n-1,m-1}) \\ &= 1 - P(0 < C < F_{1-\alpha/2;n-1,m-1}) - P(C > F_{\alpha/2;n-1,m-1}) \\ &= 1 - \alpha. \end{aligned}$$

Si ricava pertanto:

$$P\left(F_{1-\alpha/2;n-1,m-1} < \frac{\sigma_1^2 S_1^2}{\sigma_2^2 S_2^2} < F_{\alpha/2;n-1,m-1}\right) = 1 - \alpha,$$

ossia:

$$P\left(\frac{1}{F_{\alpha/2;n-1,m-1}} \frac{S_1^2}{S_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{F_{1-\alpha/2;n-1,m-1}} \frac{S_1^2}{S_2^2}\right) = 1 - \alpha. \quad (4.34)$$

Poiché per la Proposizione 2.3.2 risulta

$$F_{1-\alpha/2;n-1,m-1} = \frac{1}{F_{\alpha/2;n-1,m-1}},$$

dalla (4.34) si trae:

$$P\left(\frac{1}{F_{\alpha/2;n-1,m-1}} \frac{S_1^2}{S_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < F_{\alpha/2;n-1,m-1} \frac{S_1^2}{S_2^2}\right) = 1 - \alpha.$$

Le statistiche C^- e C^+ , rispettivamente estremi sinistro e destro dell'intervallo fiduciario per il rapporto σ_1^2/σ_2^2 sono pertanto le seguenti:

$$C^- = \frac{1}{F_{\alpha/2; n-1, m-1}} \frac{s_1^2}{s_2^2}, \quad C^+ = F_{\alpha/2; n-1, m-1} \frac{s_1^2}{s_2^2}.$$

Esempio 4.6.1 Si desidera costruire un intervallo fiduciario di coefficiente $1 - \alpha = 0.98$ per il rapporto σ_1^2/σ_2^2 nel caso dei campioni casuali considerati nell'Esempio 4.4.2. Essendo $n = 10$, $m = 8$, $s_1 = 0.5$, $s_2 = 0.7$, dalla consultazione della Tabella 5 dell'Appendice B si ricava $F_{0.01; 9,7} = 6.72$ e $F_{0.01; 7,9} = 5.61$, così che

$$\begin{aligned} \frac{1}{F_{\alpha/2; n-1, m-1}} \frac{s_1^2}{s_2^2} &= \frac{1}{6.72} \frac{0.25}{0.49} = 0.076 \\ F_{\alpha/2; n-1, m-1} \frac{s_1^2}{s_2^2} &= 5.61 \frac{0.25}{0.49} = 2.862 \end{aligned}$$

da cui, in virtù del Teorema 4.6.1, segue la stima $(0.076, 2.862)$ dell'intervallo fiduciario. Si noti che questo contiene il valore 1; non è dunque lecito concludere che la congettura $\sigma_1^2 = \sigma_2^2$ sia errata. ♦

Esempio 4.6.2 Con riferimento all'Esempio 2.3.1 costruiamo un intervallo fiduciario di coefficiente $1 - \alpha = 0.9$ per il rapporto σ_1^2/σ_2^2 . Dai dati osservati risulta $(s_1/s_2)^2 = 3.9487$; inoltre dalla Tabella 6 dell'Appendice B si trae $F_{0.05; 7,9} = 3.29$ e $F_{0.05; 9,7} = 3.68$. Essendo $n = 8$ e $m = 10$ si ha dunque:

$$\begin{aligned} \frac{1}{F_{\alpha/2; n-1, m-1}} \frac{s_1^2}{s_2^2} &= \frac{1}{3.29} 3.9487 = 1.2 \\ F_{\alpha/2; n-1, m-1} \frac{s_1^2}{s_2^2} &= 3.68 \times 3.9487 = 14.5 \end{aligned}$$

di modo che $(1.2, 14.5)$ costituisce la stima dell'intervallo fiduciario per il rapporto σ_1^2/σ_2^2 in virtù del Teorema 4.6.1. Poiché questo intervallo non contiene il valore 1 è lecito ritenere che le varianze σ_1^2 e σ_2^2 non siano uguali. ♦

4.7 Popolazioni di Bernoulli

Consideriamo un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione avente distribuzione di Bernoulli di parametro θ ($0 < \theta < 1$):

$$P(X_i = 1) = \theta, \quad P(X_i = 0) = 1 - \theta \quad (i = 1, 2, \dots, n);$$

cioè allo scopo di determinare un intervallo fiduciario per θ . Sussiste il seguente teorema:

4.7. POPOLAZIONI DI BERNOULLI

Teorema 4.7.1 Se \bar{X} è la media campionaria di un campione casuale di taglia n estratto da una popolazione di Bernoulli di parametro θ , per n grande un intervallo fiduciario per θ di coefficiente approssimativamente $1 - \alpha$ è il seguente:

$$\left(\bar{X} - z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \bar{X} + z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \right).$$

Dim. Poiché (cfr. § A.1.2) $E(X_i) = \theta$ e $D^2(X_i) = \theta(1-\theta)$ ($i = 1, 2, \dots, n$), dal teorema centrale del limite segue che per $n \rightarrow \infty$ la variabile casuale

$$\begin{aligned} C &= \gamma(X_1, X_2, \dots, X_n; \theta) = \frac{X_1 + X_2 + \dots + X_n - n\theta}{\sqrt{n\theta(1-\theta)}} \\ &= \frac{\bar{X} - \theta}{\sqrt{\theta(1-\theta)/n}} \end{aligned} \quad (4.35)$$

converge in distribuzione ad una variabile casuale normale standard. Faremo ora uso di tale risultato per determinare un intervallo fiduciario per il parametro θ . Notiamo anzitutto che C dipende dal campione casuale (X_1, X_2, \dots, X_n) e dal parametro θ da stimare. Inoltre al crescere di n essa tende ad assumere distribuzione normale standard, così che nella sua funzione di distribuzione la dipendenza da θ tende a scomparire. Anche se C non può essere riguardata come variabile cardine in quanto la sua distribuzione dipende per ogni fissato n dal parametro θ da stimare, per $n \rightarrow \infty$ è possibile ugualmente applicare il metodo del cardine per determinare un intervallo fiduciario per θ di coefficiente all'incirca $1 - \alpha$. Infatti, come subito vedremo, è possibile mostrare che per n grande la relazione $\alpha_1 < \epsilon < \alpha_2$ è approssimativamente equivalente a $C^- < \theta < C^+$, con C^- e C^+ statistiche opportune. Ricordando la definizione di z_α (cfr. § 1.4), e utilizzando la circostanza che al crescere di n la variabile C tende ad assumere distribuzione normale standard, dalla (4.35) si ricava:

$$P\left[-z_{\alpha/2} < \frac{\bar{X} - \theta}{\sqrt{\theta(1-\theta)/n}} < z_{\alpha/2}\right] \approx 1 - \alpha, \quad (4.36)$$

dove l'approssimazione migliora al crescere di n . Risolviamo rispetto a θ le diseguaglianze presenti nella (4.36), che sintetizziamo nella forma

$$\left| \frac{\bar{X} - \theta}{\sqrt{\theta(1-\theta)/n}} \right| < z_{\alpha/2}.$$

Si ha allora:

$$\left[\frac{\bar{X} - \theta}{\sqrt{\theta(1-\theta)/n}} \right]^2 < z_{\alpha/2}^2,$$

ossia:

$$\bar{X}^2 - 2\theta\bar{X} + \theta^2 < z_{\alpha/2}^2 \frac{\theta(1-\theta)}{n},$$

da cui segue:

$$\theta^2(n + z_{\alpha/2}^2) - \theta(2n\bar{X} + z_{\alpha/2}^2) + n\bar{X}^2 < 0. \quad (4.37)$$

Posto

$$C^- = \frac{2n\bar{X} + z_{\alpha/2}^2 - z_{\alpha/2}\sqrt{z_{\alpha/2}^2 + 4n\bar{X} - 4n\bar{X}^2}}{2(n+z_{\alpha/2}^2)} \quad (4.38)$$

$$C^+ = \frac{2n\bar{X} + z_{\alpha/2}^2 + z_{\alpha/2}\sqrt{z_{\alpha/2}^2 + 4n\bar{X} - 4n\bar{X}^2}}{2(n+z_{\alpha/2}^2)}, \quad (4.39)$$

la disequazione (4.37) è verificata se e solo se risulta

$$C^- < \theta < C^+.$$

Trascurando nelle (4.38) e (4.39) i termini che per $n \rightarrow \infty$ tendono a zero più rapidamente di $1/\sqrt{n}$, si ricava che per n grande le statistiche C^- e C^+ sono approssimativamente le seguenti:

$$C^- \approx \bar{X} - z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \quad C^+ \approx \bar{X} + z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}.$$

Resta così individuato l'intervallo fiduciario approssimato (C^-, C^+) , con

$$P\left[\bar{X} - z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} < \theta < \bar{X} + z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right] \approx 1 - \alpha.$$

Esempio 4.7.1 Una ditta è interessata a stabilire se un nuovo prodotto risulta gradito ai suoi potenziali acquirenti. Da un'indagine condotta su 400 persone emerge che 140 di esse hanno espresso giudizio positivo. Si voglia determinare un intervallo fiduciario di coefficiente approssimativamente $1 - \alpha = 0.95$ per la frequenza θ dei casi in cui il prodotto risulta gradito. Si tratta di una popolazione di Bernoulli tale che su $n = 400$ prove l'evento "il prodotto risulta gradito" si è verificato 140 volte. Quindi la media campionaria assume il valore $\bar{x} = 0.35$. Essendo $n > 30$, possiamo fare uso del Teorema 4.7.1 per stimare θ . Dalla Tabella 2 dell'Appendice B si ottiene $z_{\alpha/2} = z_{0.025} = 1.96$. Inoltre si ha:

$$\begin{aligned} \bar{x} - z_{\alpha/2} \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} 20 &= 0.35 - 1.96 \sqrt{\frac{0.35(1-0.35)}{400}} = 0.303 \\ \bar{x} + z_{\alpha/2} \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} 20 &= 0.35 + 1.96 \sqrt{\frac{0.35(1-0.35)}{400}} = 0.397. \end{aligned}$$

Per il Teorema 4.7.1, la stima approssimata dell'intervallo fiduciario è dunque $(0.303, 0.397)$.

4.8 Popolazioni esponenziali

Consideriamo un campione casuale estratto da una popolazione esponenziale di valore medio θ (con $\theta > 0$). Si vuole determinare un intervallo fiduciario per il parametro θ .

Teorema 4.8.1 Se (X_1, X_2, \dots, X_n) è un campione casuale estratto da una popolazione esponenziale, un intervallo fiduciario di coefficiente $1 - \alpha$ per il valore medio θ è il seguente:

$$\left(\frac{2}{\chi_{\alpha/2;2n}^2} \sum_{i=1}^n X_i, \frac{2}{\chi_{1-\alpha/2;2n}^2} \sum_{i=1}^n X_i \right).$$

Dim. Se X_i è una variabile casuale esponenziale di media θ , $2X_i/\theta$ è una variabile casuale esponenziale di media 2 e, quindi (cfr. Osservazione 2.1.2), è una variabile chi-quadrato con 2 gradi di libertà. Pertanto in virtù del Teorema 2.1.2 la variabile casuale

$$C = \gamma(X_1, X_2, \dots, X_n; \theta) = \frac{2}{\theta} \sum_{i=1}^n X_i \quad (4.40)$$

ha distribuzione chi-quadrato con $2n$ gradi di libertà in quanto espressa come somma di n variabili chi-quadrato indipendenti ciascuna avente 2 gradi di libertà. La (4.40) individua dunque una variabile cardine poiché C dipende dal campione casuale (X_1, X_2, \dots, X_n) e dal parametro θ , mentre la sua distribuzione, come si è visto, non dipende da θ . La funzione γ è infine strettamente decrescente in θ , e quindi invertibile. Possiamo allora fare uso del metodo del cardine per ricavare un intervallo fiduciario di coefficiente $1 - \alpha$ per il parametro θ . Ricordiamo che, poiché la variabile (4.40) ha distribuzione chi-quadrato con $2n$ gradi di libertà, i valori $\chi_{\alpha/2;2n}^2$ e $\chi_{1-\alpha/2;2n}^2$ sono tali da aversi:

$$P\left(\chi_{1-\alpha/2;2n}^2 < \frac{2}{\theta} \sum_{i=1}^n X_i < \chi_{\alpha/2;2n}^2\right) = 1 - \alpha,$$

ovvero:

$$P\left(\frac{2}{\chi_{\alpha/2;2n}^2} \sum_{i=1}^n X_i < \theta < \frac{2}{\chi_{1-\alpha/2;2n}^2} \sum_{i=1}^n X_i\right) = 1 - \alpha.$$

Pertanto le statistiche

$$C^- = \frac{2}{\chi_{\alpha/2;2n}^2} \sum_{i=1}^n X_i, \quad C^+ = \frac{2}{\chi_{1-\alpha/2;2n}^2} \sum_{i=1}^n X_i$$

costituiscono gli estremi sinistro e destro dell'intervallo fiduciario per la media θ .

Esempio 4.8.1 Si supponga che una variabile casuale esponenziale di media incognita θ sia idonea a descrivere l'intervallo di tempo che intercorre tra due consecutivi terremoti di prefissata magnitudo che si verificano in una data regione, e che si desideri ricavare un intervallo fiduciario di coefficiente $1 - \alpha = 0.9$ per θ sulla base dei seguenti 10 dati, ciascuno dei quali è un intervallo di tempo espresso in mesi:

$$(31.5, 22.2, 25.6, 17.8, 39.8, 20.2, 27.5, 24.1, 19.2, 47.9).$$

La somma di questi dati è pari a 275.8; inoltre dalla Tabella 3 dell'Appendice B risulta $\chi_{\alpha/2;2n}^2 = \chi_{0.05;20}^2 = 31.410$ e $\chi_{1-\alpha/2;2n}^2 = \chi_{0.95;20}^2 = 10.851$. Pertanto si ha:

$$\frac{2}{\chi_{\alpha/2;2n}^2} \sum_{i=1}^n x_i = \frac{2 \cdot 275.8}{31.410} = 17.56, \quad \frac{2}{\chi_{1-\alpha/2;2n}^2} \sum_{i=1}^n x_i = \frac{2 \cdot 275.8}{10.851} = 50.83.$$

Dal Teorema 4.8.1 si ricava dunque che la stima dell'intervallo fiduciario è $(17.56, 50.83)$.

Esporremo qui di seguito un ulteriore procedimento che consente di ricavare intervalli fiduciari approssimati per la media θ di una popolazione esponenziale nel caso in cui si disponga di campioni casuali di taglia elevata.

Teorema 4.8.2 Sia \bar{X} la media campionaria di un campione casuale di taglia n estratto da una popolazione esponenziale di valore medio θ . Se n è grande, un intervallo fiduciario di coefficiente approssimativamente $1 - \alpha$ per il parametro θ risulta essere:

$$\left(\frac{\bar{X}}{1 + \frac{z_{\alpha/2}}{\sqrt{n}}}, \frac{\bar{X}}{1 - \frac{z_{\alpha/2}}{\sqrt{n}}} \right).$$

Dim. Poiché (cfr. § A.2.3) $E(X_i) = \theta$ e $D^2(X_i) = \theta^2$ ($i = 1, 2, \dots, n$), per il teorema centrale del limite la variabile

$$C = \frac{X_1 + X_2 + \dots + X_n - nE(X_i)}{\sqrt{nD^2(X_i)}} = \sqrt{n} \left(\frac{\bar{X}}{\theta} - 1 \right) \quad (4.41)$$

al divergere di n converge in distribuzione ad una variabile casuale normale standard. Ciò consente di determinare un intervallo fiduciario per θ . Notiamo che la variabile C data dalla (4.41) dipende dal campione casuale (X_1, X_2, \dots, X_n) e dal parametro da stimare θ . Inoltre, al crescere di n essa tende ad assumere distribuzione normale standard, così che la dipendenza della sua funzione di distribuzione dal parametro θ tende a scomparire. La funzione (4.41) è, infine, strettamente decrescente in θ , e quindi dotata di inversa. Per n grande C può dunque essere riguardata approssimativamente come una variabile cardinale, così che per n grande si può applicare il metodo del cardine al fine di determinare un intervallo fiduciario di coefficiente all'incirca $1 - \alpha$ per il parametro θ . Ricordando la definizione di z_α (cfr. § 1.4), e facendo uso della circostanza che al crescere di n la variabile C tende ad assumere distribuzione normale standard, dalla (4.41) si ottiene:

$$P \left[-z_{\alpha/2} < \sqrt{n} \left(\frac{\bar{X}}{\theta} - 1 \right) < z_{\alpha/2} \right] \approx 1 - \alpha,$$

con un'approssimazione che migliora al crescere di n . Ne segue:

$$P \left[\bar{X} \left(1 + \frac{z_{\alpha/2}}{\sqrt{n}} \right)^{-1} < \theta < \bar{X} \left(1 - \frac{z_{\alpha/2}}{\sqrt{n}} \right)^{-1} \right] \approx 1 - \alpha.$$

Le statistiche C^- e C^+ , rispettivamente estremi sinistro e destro dell'intervallo fiduciario, sono dunque

$$C^- = \bar{X} \left(1 + \frac{z_{\alpha/2}}{\sqrt{n}} \right)^{-1}, \quad C^+ = \bar{X} \left(1 - \frac{z_{\alpha/2}}{\sqrt{n}} \right)^{-1}.$$

Esempio 4.8.2 Si consideri un campione casuale di taglia $n = 400$ estratto da una popolazione esponenziale di media θ . Si vuole determinare un intervallo fiduciario di coefficiente

$1 - \alpha = 0.98$ per il parametro θ sapendo che la media campionaria assume il valore $\bar{x} = 3.2$. Dalla Tabella 2 dell'Appendice B si trae $z_{\alpha/2} = z_{0.01} = 2.326$, di modo che risulta:

$$\begin{aligned} \bar{x} \left(1 + \frac{z_{\alpha/2}}{\sqrt{n}} \right)^{-1} &= 3.2 \left(1 + \frac{2.326}{20} \right)^{-1} = 2.8666 \\ \bar{x} \left(1 - \frac{z_{\alpha/2}}{\sqrt{n}} \right)^{-1} &= 3.2 \left(1 - \frac{2.326}{20} \right)^{-1} = 3.6209. \end{aligned}$$

Essendo $n > 30$, possiamo fare uso del Teorema 4.8.2 ricavando così che la stima approssimata dell'intervallo fiduciario è $(2.8667, 3.6209)$. ◇

4.9 Stime per quantili

Verranno ora definite delle quantità, dette quantili, che forniscono indicazioni sulla forma della funzione di distribuzione di una variabile casuale.

Definizione 4.9.1 Data una variabile casuale X di funzione di distribuzione $F_X(x)$, si dice quantile p -esimo di X ogni reale ξ_p che soddisfa entrambe le relazioni

$$P(X \leq \xi_p) \geq p, \quad P(X \geq \xi_p) \geq 1 - p,$$

per p fissato, con $0 < p < 1$.

Si noti che ξ_p è tale da aversi:

$$p \leq F_X(\xi_p) \leq p + P(X = \xi_p). \quad (4.42)$$

In particolare, ricordando la Definizione 3.1.1, si verifica che il quantile $\xi_{1/2}$ coincide con la mediana.

Nei casi in cui la funzione di distribuzione della popolazione in esame è ignota la conoscenza dei quantili fornisce informazioni utili per stimarne l'andamento. In questo paragrafo considereremo campioni casuali estratti da una popolazione continua dotata di funzione di distribuzione $F_X(x)$ che sia continua e strettamente crescente. Pertanto, poiché per variabili casuali continue si ha $P(X = x) = 0$, le diseguaglianze (4.42) conducono all'equazione

$$F_X(\xi_p) = p \quad (4.43)$$

che, in virtù dell'ipotesi che $F_X(x)$ è strettamente crescente, ammette un'unica soluzione. Nel seguito considereremo solo casi in cui il quantile ξ_p esiste ed è unico.

Si noti che, in accordo con la terminologia introdotta nel § 1.4, il quantile superiore z_α di una variabile continua X è definito dalla relazione

$$P(X \geq z_\alpha) = \alpha. \quad (4.44)$$

In virtù della (4.43), risulta poi:

$$P(X \leq \xi_p) = p. \quad (4.45)$$

Dal confronto delle (4.44) e (4.45) si comprende l'origine del termine "quantile superiore" già attribuito a z_α . Per brevità, ξ_p è stato denominato semplicemente "quantile" in luogo di "quantile inferiore".

Affronteremo ora il problema della stima puntuale di un quantile ξ_p , con p fissato, di una popolazione continua.

Denotate con $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ le statistiche d'ordine relative ad un campione casuale (X_1, X_2, \dots, X_n) di variabile casuale genitrice di funzione di distribuzione F_X , introduciamo le variabili casuali V_1, V_2, \dots, V_n con $V_k = F_X(X_{(k)})$ per $k = 1, 2, \dots, n$.

Proposizione 4.9.1 La variabile casuale V_k ha funzione di distribuzione coincidente con quella della k -esima statistica d'ordine d'un campione casuale estratto da una popolazione distribuita uniformemente in $(0, 1)$ e possiede valore medio

$$E(V_k) = \frac{k}{n+1} \quad (k = 1, 2, \dots, n). \quad (4.46)$$

Dim. Per $k = 1, 2, \dots, n$ si ha invero:

$$P(V_k \leq u) = P[F_X(X_{(k)}) \leq u] = P[X_{(k)} \leq F_X^{-1}(u)] = F_{(k)}[F_X^{-1}(u)],$$

dove (cfr. § 3.1)

$$F_{(k)}(x) = \sum_{j=k}^n \binom{n}{j} [F_X(x)]^j [1 - F_X(x)]^{n-j} \quad (4.47)$$

denota la funzione di distribuzione della statistica d'ordine $X_{(k)}$. Pertanto, per $0 < u < 1$, risulta:

$$\begin{aligned} P(V_k \leq u) &= F_{(k)}[F_X^{-1}(u)] \\ &= \sum_{j=k}^n \binom{n}{j} \{F_X[F_X^{-1}(u)]\}^j \{1 - F_X[F_X^{-1}(u)]\}^{n-j} \\ &= \sum_{j=k}^n \binom{n}{j} u^j (1-u)^{n-j}. \end{aligned} \quad (4.48)$$

Ricordando che $F_U(u) = u$ per $0 < u < 1$ è la funzione di distribuzione della variabile casuale U uniforme in $(0, 1)$, la (4.48) diventa:

$$P(V_k \leq u) = \sum_{j=k}^n \binom{n}{j} [F_U(u)]^j [1 - F_U(u)]^{n-j} \quad (0 < u < 1).$$

La variabile casuale V_k ha dunque la medesima funzione di distribuzione della k -esima statistica d'ordine d'un campione casuale estratto da una popolazione distribuita uniformemente in $(0, 1)$. Osservando che dalla (4.48) si trae

$$P(V_k > u) = \sum_{j=0}^{k-1} \binom{n}{j} u^j (1-u)^{n-j} \quad (0 < u < 1),$$

4.9. STIME PER QUANTILI

il valore medio di V_k è (v. nota a piè pagina nell'Esempio 3.7.5)

$$E(V_k) = \int_0^1 P(V_k > u) du = \sum_{j=0}^{k-1} \binom{n}{j} \int_0^1 u^j (1-u)^{n-j} du. \quad (4.49)$$

Ricordando la (3.156), la (4.49) diventa

$$E(V_k) = \sum_{j=0}^{k-1} \binom{n}{j} \frac{j!(n-j)!}{(n+1)!},$$

ossia la (4.46). ■

L'espressione della media di V_k consente di costruire degli stimatori per i quantili ξ_p . Infatti, se $p = k/(n+1)$, dalle (4.43) e (4.46) risulta

$$E(V_k) \equiv E[F_X(X_{(k)})] = \frac{k}{n+1} = F_X[\xi_{k/(n+1)}].$$

Questa uguaglianza suggerisce di stimare il quantile $\xi_{k/(n+1)}$ mediante la statistica d'ordine $X_{(k)}$. Invero, se fosse lecito scambiare l'operatore di media E con la funzione F_X , si otterrebbe $F_X[E(X_{(k)})] = F_X[\xi_{k/(n+1)}]$. Ciò, in virtù dell'ipotesi che $F_X(x)$ è strettamente crescente, condurrebbe all'identità $E(X_{(k)}) = \xi_{k/(n+1)}$, così che $X_{(k)}$ sarebbe uno stimatore corretto di $\xi_{k/(n+1)}$. Anche se la summenzionata operazione di scambio non è in generale lecita, rimane il suggerimento di stimare il quantile $\xi_{k/(n+1)}$ attraverso la statistica d'ordine $X_{(k)}$. Per $k/(n+1) < p < (k+1)/(n+1)$ si può poi stimare ξ_p mediante la statistica

$$\hat{Q}_p = [k+1 - (n+1)p]X_{(k)} + [(n+1)p - k]X_{(k+1)},$$

costituente una combinazione lineare delle statistiche d'ordine $X_{(k)}$ e $X_{(k+1)}$. È opportuno notare, ad esempio, che $\hat{Q}_{1/2}$ coincide con la mediana campionaria \tilde{X} definita nella (3.7).

Vale la pena di sottolineare che gli stimatori dei quantili qui suggeriti possono essere ricavati a prescindere dalla conoscenza della forma della funzione di distribuzione $F_X(x)$. Osserviamo però che, in generale, tali stimatori non possiedono caratteristiche tali da renderli particolarmente "efficaci". Si può infatti verificare, come vedremo nel prossimo esempio, che la conoscenza della forma della distribuzione della popolazione consenta di ricavare degli stimatori dei quantili che sono preferibili alle statistiche d'ordine.

Esempio 4.9.1 Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione distribuita uniformemente in $(0, \theta)$, con $\theta > 0$. Supponiamo che si desideri stimare il quantile p -esimo, con $p = k/n$. Se non si conoscesse la distribuzione della popolazione, per stimare $\xi_p \equiv \theta p$ si utilizzerebbe la statistica d'ordine $X_{(k)}$. Analizziamo le proprietà di tale stimatore osservando che risulta $X_{(k)} = \theta V_k$, dove V_k è la k -esima statistica d'ordine d'un campione di taglia n estratto da una popolazione distribuita uniformemente in $(0, 1)$. Dalla Proposizione 4.9.1 si ha:

$$E(X_{(k)}) = \theta E(V_k) = \theta \frac{k}{n+1} \equiv \xi_p \frac{n}{n+1}, \quad (4.50)$$

così che $X_{(k)}$ non è uno stimatore corretto. Per il Teorema 3.1.1 si ha poi che la densità di probabilità di $X_{(k)}$ è data da

$$f_{(k)}(x) = \binom{n}{k} k \left(\frac{x}{\theta}\right)^{k-1} \left(1 - \frac{x}{\theta}\right)^{n-k} \frac{1}{\theta} \quad (0 < x < \theta).$$

Pertanto

$$\begin{aligned} E(X_{(k)}^2) &= \binom{n}{k} \frac{k}{\theta^k} \int_0^\theta x^{k+1} \left(1 - \frac{x}{\theta}\right)^{n-k} dx \\ &= \binom{n}{k} k \theta^2 \int_0^1 z^{k+1} (1-z)^{n-k} dz, \end{aligned} \quad (4.51)$$

avendo posto $z = x/\theta$. Poiché dalla (3.156) si trae

$$\int_0^1 z^{k+1} (1-z)^{n-k} dz = \frac{(k+1)! (n-k)!}{(n+2)!} = \frac{k+1}{(n+1)(n+2)} \binom{n}{k}^{-1},$$

la (4.51) diventa:

$$E(X_{(k)}^2) = \frac{k(k+1)}{(n+1)(n+2)} \theta^2. \quad (4.52)$$

Da questa, e dalla media (4.50), si ricava l'errore quadratico medio di $X_{(k)}$:

$$\begin{aligned} \text{mse}(X_{(k)}) &= E[(X_{(k)} - \xi_p)^2] = E(X_{(k)}^2) - 2\xi_p E(X_{(k)}) + \xi_p^2 \\ &= \left[\frac{k(k+1)}{(n+1)(n+2)} - 2p \frac{k}{n+1} + p^2 \right] \theta^2. \end{aligned} \quad (4.53)$$

Si ricorderà che nell'Esempio 3.8.3 si era visto essere preferibile adottare $(n+1)X_{(n)}/n$ come stimatore di θ . Per stimare il quantile $\xi_p \equiv \theta p$ sembra allora opportuno usare lo stimatore corretto

$$\hat{Q} = \frac{n+1}{n} p X_{(n)}.$$

Dalla Proposizione 4.9.1 segue infatti:

$$E(\hat{Q}) = \frac{n+1}{n} p E(X_{(n)}) = \frac{n+1}{n} p \theta E V_n = p \theta \equiv \xi_p.$$

Notando che, in virtù delle (4.52) e (4.50), risulta

$$D^2(X_{(n)}) = E(X_{(n)}^2) - [E(X_{(n)})]^2 = \frac{n \theta^2}{(n+1)^2 (n+2)},$$

si ha poi:

$$\text{mse}(\hat{Q}) = D^2(\hat{Q}) = \frac{(n+1)^2}{n^2} p^2 D^2(X_{(n)}) = \frac{p^2 \theta^2}{n(n+2)}. \quad (4.54)$$

4.9. STIME PER QUANTILI

Infine, si ricava facilmente che risulta $\text{mse}(\hat{Q}) \leq \text{mse}(X_{(k)})$. Ricordando che $p = k/n$, dagli errori quadratici medi (4.53) e (4.54) si ricava infatti:

$$\begin{aligned} \text{mse}(X_{(k)}) - \text{mse}(\hat{Q}) &= k \theta^2 \left[\frac{k+1}{(n+1)(n+2)} - \frac{2k}{n(n+1)} + \frac{k}{n^2} - \frac{k}{n^3(n+2)} \right] \\ &= k \theta^2 \frac{n^2(n-k)+k(n-1)}{n^3(n+1)(n+2)} \geq 0, \end{aligned}$$

dove la diseguaglianza è conseguenza dell'essere $1 \leq k \leq n$. La conoscenza della forma della distribuzione della popolazione ha dunque consentito di ricavare per i quantili uno stimatore con errore quadratico medio non superiore a quello della statistica d'ordine $X_{(k)}$. \diamond

Affrontiamo ora il problema della stima intervallare di un quantile ξ_p , con p fissato, di una popolazione continua.

Un intervallo fiduciario per la stima dei quantili ξ_p può essere ottenuto usando due statistiche d'ordine $X_{(r)}$ e $X_{(k)}$ (con $r < k$) come estremi dell'intervallo fiduciario. Per determinare il coefficiente di fiducia dell'intervallo $(X_{(r)}, X_{(k)})$ osserviamo che, essendo $r < k$, per l'ipotizzata continuità delle variabili $X_{(i)}$, risulta:

$$\begin{aligned} P(X_{(r)} < \xi_p < X_{(k)}) &= P(X_{(r)} < \xi_p, X_{(k)} > \xi_p) \\ &= P(X_{(r)} < \xi_p) - P(X_{(r)} < \xi_p, X_{(k)} \leq \xi_p) \\ &= P(X_{(r)} < \xi_p) - P(X_{(k)} \leq \xi_p) \\ &= F_{(r)}(\xi_p) - F_{(k)}(\xi_p). \end{aligned}$$

Facendo uso della (4.47), e ricordando che $F_X(\xi_p) = p$ si ha poi:

$$\begin{aligned} F_{(k)}(\xi_p) &= \sum_{j=k}^n \binom{n}{j} [F_X(\xi_p)]^j [1 - F_X(\xi_p)]^{n-j} \\ &= \sum_{j=k}^n \binom{n}{j} p^j (1-p)^{n-j}. \end{aligned}$$

Il coefficiente di fiducia dell'intervallo $(X_{(r)}, X_{(k)})$ è dunque il seguente:

$$\begin{aligned} P(X_{(r)} < \xi_p < X_{(k)}) &= \sum_{j=r}^n \binom{n}{j} p^j (1-p)^{n-j} - \sum_{j=k}^n \binom{n}{j} p^j (1-p)^{n-j} \\ &= \sum_{j=r}^{k-1} \binom{n}{j} p^j (1-p)^{n-j}. \end{aligned}$$

Ad esempio, per un campione casuale di taglia $n = 12$ il coefficiente di fiducia dell'intervallo $(X_{(3)}, X_{(10)})$ per la stima della mediana della popolazione è dato da

$$P(X_{(3)} < \xi_{1/2} < X_{(10)}) = \sum_{j=3}^9 \binom{12}{j} \left(\frac{1}{2}\right)^{12} = 0.9614.$$

Il metodo di stima intervallare ora illustrato è di ampia utilizzazione potendo essere applicato a campioni casuali estratti da popolazioni qualsiasi purché dotate di funzione di distribuzione continua e strettamente crescente. In aggiunta, esso è particolarmente semplice, richiedendo esclusivamente il calcolo di probabilità di tipo binomiale. Diversamente da come si procede nell'utilizzazione del metodo del cardine vi è ora, però, l'inconveniente di dover prima fissare gli estremi dell'intervallo fiduciario e poi determinarne il coefficiente di fiducia. Ciò in taluni casi può portare ad intervalli fiduciari insoddisfacenti. Comunque, se si desidera un intervallo fiduciario per ξ_p che sia almeno di coefficiente fiduciario $1 - \alpha$, occorre scegliere r e k in modo tale che sussista la diseguaglianza

$$P(X_{(r)} < \xi_p < X_{(k)}) \equiv \sum_{j=r}^{k-1} \binom{n}{j} p^j (1-p)^{n-j} \geq 1 - \alpha. \quad (4.55)$$

Nella pratica si scelgono r e k in guisa da minimizzare la differenza $k - r$ sotto il vincolo (4.55).

Va infine notato che, poiché il coefficiente di fiducia presente nella (4.55) è espresso come somma di probabilità binomiali (assumenti quindi un insieme discreto di valori), talora accade che non sia possibile ottenere *esattamente* il valore $1 - \alpha$ desiderato, ossia di pervenire *esattamente* al preassegnato coefficiente di fiducia $1 - \alpha$.

Capitolo 5

Ipotesi statistiche

5.1 Verifica delle ipotesi

Nei capitoli precedenti ci siamo interessati alla stima puntuale e intervallare dei parametri incogniti di una popolazione. Ci occuperemo ora di un altro problema tipico dell'inferenza statistica: la verifica delle ipotesi statistiche. Si tratta, in sostanza, di stabilire se una data realizzazione (x_1, x_2, \dots, x_n) di un campione casuale può essere riguardata o meno come estratta da una specificata popolazione. Per rispondere a tale quesito è consuetudine calcolare anzitutto una prima quantità che è funzione della realizzazione, e successivamente una seconda quantità funzione della popolazione che si suppone generi il campione. Si valuta, poi, se la differenza tra queste non è significativa, nel qual caso l'ipotesi viene accettata, oppure se è causata da una scelta erronea del modello di popolazione ipotizzato, ed allora l'ipotesi viene respinta.

Precisiamo anzitutto che cosa debba intendersi per ipotesi statistica.

Definizione 5.1.1 *Un'ipotesi statistica è una congettura relativa alla distribuzione di una o più variabili casuali. Se l'ipotesi statistica individua completamente la distribuzione, nel senso che specifica la forma funzionale della distribuzione e i valori dei parametri che vi coimpaiono, essa è detta ipotesi semplice; altrimenti, ossia se essa non precisa il tipo di distribuzione o i valori di tutti i parametri che la caratterizzano, essa viene detta ipotesi composta.*

Solitamente un'ipotesi statistica viene formulata in guisa da valutarne l'opportunità di rifiuto. Se, ad esempio, si vuole stabilire se una moneta è truccata, si formula dapprima l'ipotesi che la moneta sia onesta (ossia che la probabilità di avere testa nel singolo lancio sia pari a $1/2$) e si valuta poi se tale ipotesi debba rigettarsi. Analogamente, quando si vuole stabilire se un certo procedimento è migliore di un altro, si formula l'ipotesi che i due procedimenti si equivalgono e si valuta poi se ciò debba considerarsi inaccettabile.

Quando le ipotesi sottoposte a verifica coinvolgono uno o più parametri incogniti della popolazione in esame, esse sono dette *ipotesi parametriche*. Nel caso di ipotesi parametriche, il problema della verifica di ipotesi viene formulato in accordo con le considerazioni che

seguono. Sia Θ lo spazio dei parametri, ossia l'insieme dei valori che possono assumere i parametri incogniti coinvolti nelle ipotesi statistiche in questione. Qui, solitamente, ci limiteremo a trattare ipotesi concernenti un unico parametro che indicheremo con θ . Siano poi Θ_0 e Θ_1 sottoinsiemi disgiunti di Θ . Si considerano due ipotesi statistiche: la prima, quella soggetta a verifica, si dice *ipotesi nulla* e viene denotata con H_0 ; essa si realizza quando $\theta \in \Theta_0$. La seconda è detta *ipotesi alternativa* e si denota con H_1 ; essa si realizza quando $\theta \in \Theta_1$. Tale situazione si rappresenta schematicamente nel seguente modo:

$$H_0: \theta \in \Theta_0$$

$$H_1: \theta \in \Theta_1.$$

Esempio 5.1.1 Supponiamo che una casa farmaceutica debba decidere, sulla base di dati osservati, se almeno il 90 per cento dei pazienti trattati con un particolare farmaco guarisce da una certa sindrome. Tale problema può essere interpretato intendendo che la casa farmaceutica vuole stabilire se θ , il parametro di una popolazione di Bernoulli, è maggiore di 0.9. In questo caso lo spazio dei parametri è l'intervallo $\Theta = [0, 1]$, con $\Theta_0 = [0.9, 1]$ e $\Theta_1 = [0, 0.9]$. Le ipotesi statistiche si possono allora indicare al seguente modo:

$$H_0: \theta \geq 0.9$$

$$H_1: \theta < 0.9.$$

Si noti che in questo esempio entrambe le ipotesi sono composte giacché esse non coinvolgono un unico valore del parametro θ e, pertanto, non specificano completamente la distribuzione del campione casuale. Per ipotesi del tipo

$$H_0: \theta = 0.9$$

$$H_1: \theta \neq 0.9,$$

l'ipotesi nulla H_0 è invece semplice, mentre l'ipotesi alternativa H_1 è ancora composta.

Diamo ora la definizione di *test di ipotesi*.

Definizione 5.1.2 Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione caratterizzata da un parametro incognito θ e siano H_0 e H_1 ipotesi statistiche su θ . Dicesi *test* ogni procedimento a carattere algoritmico che fa uso di una realizzazione osservata del campione per decidere se accettare l'ipotesi H_0 o se rifiutarla a favore dell'ipotesi alternativa H_1 .

Allo scopo di decidere se accettare o rifiutare l'ipotesi nulla H_0 , si costruisce anzitutto una regione C , detta *regione critica del test* o *regione di rifiuto*; questa viene determinata in modo da contenere tutte e sole le realizzazioni del campione casuale in base alle quali l'ipotesi nulla H_0 non è statisticamente plausibile, ed è quindi da rifiutare. Pertanto, una volta osservata una realizzazione (x_1, x_2, \dots, x_n) del campione casuale in esame, si verifica se essa cade o meno nella regione critica, operando poi al seguente modo:

- se $(x_1, x_2, \dots, x_n) \notin C$ si accetta l'ipotesi nulla H_0 ;

- se $(x_1, x_2, \dots, x_n) \in C$ si rifiuta l'ipotesi nulla H_0 (e quindi si accetta l'ipotesi alternativa H_1).

Riassumendo, facendosi guidare da considerazioni varie (natura del modello statistico, agenti responsabili della generazione del campione, informazioni eventuali sulla significatività di possibili valori dei parametri, ecc.) si costruisce anzitutto la regione critica C ; successivamente si controlla se la realizzazione osservata appartiene alla regione critica: in caso affermativo l'ipotesi nulla H_0 viene rifiutata, altrimenti viene accettata.

Esempio 5.1.2 Si desidera stabilire se una moneta è truccata, ossia se la probabilità θ di uscita di testa in un singolo lancio è diversa da $1/2$, effettuandone 10 lanci. Si ottiene così il campione casuale $(X_1, X_2, \dots, X_{10})$, dove $X_i = 1$ se nel lancio i -esimo si ha testa, $X_i = 0$ se si ha croce. Si formulano poi le seguenti ipotesi:

$$H_0: \theta = 0.5$$

$$H_1: \theta \neq 0.5.$$

La regione critica del test andrà identificata con l'insieme delle realizzazioni $(x_1, x_2, \dots, x_{10})$ che non giustificano l'ipotesi nulla $\theta = 0.5$, ossia l'ipotesi di correttezza della moneta. Ad esempio se si pone

$$C = \left\{ (x_1, x_2, \dots, x_{10}): \sum_{i=1}^{10} x_i \leq 1 \text{ oppure } \sum_{i=1}^{10} x_i \geq 9 \right\}, \quad (5.1)$$

l'ipotesi che la moneta non sia truccata viene rifiutata se su 10 lanci indipendenti esce un numero di teste inferiore a 2 o superiore a 8.

Criteri rigorosi per la costruzione di regioni critiche di test di ipotesi verranno introdotti nei paragrafi seguenti.

Il criterio sopra indicato per l'accettazione o il rifiuto dell'ipotesi nulla non è, però, in generale scevro da errori; questi si vogliono classificare al seguente modo:

- *errore di I tipo*: lo si commette quando si rifiuta l'ipotesi nulla H_0 nel caso in cui questa è vera;

- *errore di II tipo*: consiste nell'accettare l'ipotesi nulla H_0 nel caso in cui questa è falsa.

Solitamente si denota con $\alpha(\theta)$ la probabilità di commettere un errore di I tipo e con $\beta(\theta)$ la probabilità di commettere un errore di II tipo. Più precisamente, dato un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione caratterizzata da un parametro incognito θ , siano $H_0: \theta \in \Theta_0$ e $H_1: \theta \in \Theta_1$ ipotesi statistiche su θ che si desidera sottoporre a test; se C è la regione critica del test, le probabilità d'errore di I e di II tipo si esprimono nel seguente modo:

$$\alpha(\theta) = P[(X_1, X_2, \dots, X_n) \in C | \theta \in \Theta_0] \quad (5.2)$$

$$\beta(\theta) = P[(X_1, X_2, \dots, X_n) \notin C | \theta \in \Theta_1]. \quad (5.3)$$

La scrittura $\alpha(\theta)$ e $\beta(\theta)$ sta ad indicare che le probabilità d'errore variano al variare del valore del parametro θ sottoposto a verifica. Va inoltre segnalato che la notazione usata nelle (5.2) e (5.3), che verrà adottata anche nel seguito, non coinvolge probabilità condizionate in senso stretto in quanto ciò che segue la barra verticale non è un evento, ma rappresenta una circostanza che viene assunta come vera. Così, l'espressione $P[(X_1, X_2, \dots, X_n) \in C | \theta \in \Theta_0]$ rappresenta la probabilità che il campione casuale (X_1, X_2, \dots, X_n) appartenga alla regione critica qualora il parametro θ assuma un valore dell'insieme Θ_0 .

Esempio 5.1.3 Effettuate n osservazioni indipendenti (x_1, x_2, \dots, x_n) di una variabile casuale X di funzione di distribuzione incognita si vuole stabilire mediante un test se X può assumere valori negativi. Posto a tal fine $\theta \stackrel{\text{def}}{=} P(X < 0)$, formuliamo l'ipotesi $H_0: \theta = 0$ contro l'ipotesi $H_1: \theta = 0.02$. Dovendo identificare la regione critica come quella contenente tutti e soli i valori che non giustificano l'ipotesi nulla $H_0: \theta \leq P(X < 0) = 0$, conviene porre...

$$C = \{(x_1, x_2, \dots, x_n) : \text{almeno una delle } x_i \text{ è negativa}\}.$$

Poiché $\Theta_0 = \{0\}$, per la (5.2) la probabilità d'errore di I tipo è nulla; infatti risulta:

$$\begin{aligned} \alpha(\theta) &= P[(X_1, X_2, \dots, X_n) \in C | \theta \in \Theta_0] \\ &= P\left(\bigcup_{i=1}^n \{X_i < 0\} \mid \theta = 0\right) = 0. \end{aligned}$$

Se la realizzazione (x_1, x_2, \dots, x_n) contiene almeno un valore negativo, si rifiuta, com'è ovvio, l'ipotesi nulla. Si noti che in tal caso non sussiste il rischio di commettere un errore di I tipo essendo nulla la probabilità di quest'ultimo. Poiché $\Theta_1 = \{0.02\}$, dalla (5.3) si ricava poi la probabilità d'errore di II tipo:

$$\begin{aligned} \beta(\theta) &= P[(X_1, X_2, \dots, X_n) \notin C | \theta \in \Theta_1] \\ &= P\left(\bigcap_{i=1}^n \{X_i \geq 0\} \mid \theta = 0.02\right) \\ &= \prod_{i=1}^n P(X_i \geq 0 | \theta = 0.02) = (1 - 0.02)^n = (0.98)^n. \end{aligned}$$

Se i valori della realizzazione (x_1, x_2, \dots, x_n) sono tutti positivi, si accetta dunque l'ipotesi nulla commettendo un errore di II tipo con probabilità $(0.98)^n$. Perché, ad esempio, tale probabilità sia non superiore a 0.01, deve avversi

$$(0.98)^n \leq 0.01,$$

ossia

$$n \geq \frac{\ln 0.01}{\ln 0.98} = \frac{-4.6052}{-0.0202} = 227.98.$$

Quindi, affinché la probabilità d'errore di II tipo sia non superiore all'1%, la taglia del campione deve essere maggiore o uguale a 228. ♦

Esempio 5.1.4 Dato un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione normale di valore medio μ e varianza $\sigma^2 = 1$ si desidera sottoporre a test l'ipotesi nulla semplice $H_0: \mu = \mu_0$ contro l'ipotesi alternativa semplice $H_1: \mu = \mu_1$, con $\mu_1 > \mu_0$. Si supponga che la regione di rifiuto per l'ipotesi nulla H_0 sia del tipo

$$C = \{(x_1, x_2, \dots, x_n) : \bar{x} > k\}, \quad (5.4)$$

con k costante. Si rifiuta dunque l'ipotesi nulla quando la media campionaria \bar{X} assume un valore maggiore della costante k , che verrà ora determinata in modo che la probabilità d'errore di I tipo, $\alpha(\mu_0)$, sia pari a 0.05. Dalla (5.2), essendo $\Theta_0 = \{\mu_0\}$, si ricava:

$$\begin{aligned} \alpha(\mu_0) &= P(\bar{X} > k | \mu = \mu_0) = P\left(\frac{\bar{X} - \mu_0}{1/\sqrt{n}} > \frac{k - \mu_0}{1/\sqrt{n}} \mid \mu = \mu_0\right) \\ &= P\left(Z > \frac{k - \mu_0}{1/\sqrt{n}}\right) \end{aligned}$$

dove, sotto l'ipotesi nulla $\mu = \mu_0$, la variabile $Z = (\bar{X} - \mu_0)/(1/\sqrt{n})$ ha distribuzione normale standard. Poiché dalla Tabella 2 dell'Appendice B risulta

$$P(Z > z_{0.05}) = P(Z > 1.645) = 0.05,$$

affinché sia $\alpha(\mu_0) = 0.05$ deve risultare $(k - \mu_0)\sqrt{n} = 1.645$, ossia

$$k = \mu_0 + \frac{1.645}{\sqrt{n}}. \quad (5.5)$$

Dalle (5.4) e (5.5) segue allora la regione di rifiuto per la quale si ha $\alpha(\mu_0) = 0.05$:

$$C = \left\{ (x_1, x_2, \dots, x_n) : \bar{x} > \mu_0 + \frac{1.645}{\sqrt{n}} \right\}. \quad (5.6)$$

A questo punto è possibile determinare i valori di n che rendono prefissatamente piccola la probabilità d'errore di II tipo. Si desideri, ad esempio, stabilire per quali valori di n risulta $\beta(\mu_1) \leq 0.05$. Essendo $\Theta_1 = \{\mu_1\}$, dalla (5.3) e dalla (5.5) segue:

$$\begin{aligned} \beta(\mu_1) &= P(\bar{X} \leq k | \mu = \mu_1) = P\left(\frac{\bar{X} - \mu_1}{1/\sqrt{n}} \leq \frac{k - \mu_1}{1/\sqrt{n}} \mid \mu = \mu_1\right) \\ &= P[Z \leq (\mu_0 - \mu_1)\sqrt{n} + 1.645] \end{aligned}$$

dove, sotto l'ipotesi alternativa $\mu = \mu_1$, la variabile $Z = (\bar{X} - \mu_1)/(1/\sqrt{n})$ ha distribuzione normale standard. Sfruttando la simmetria di Z si ha poi:

$$\beta(\mu_1) = P[Z \geq (\mu_1 - \mu_0)\sqrt{n} - 1.645].$$

Risulta dunque $\beta(\mu_1) \leq 0.05$ quando

$$(\mu_1 - \mu_0)\sqrt{n} - 1.645 \geq z_{0.05} \equiv 1.645,$$

ossia se la taglia n del campione è tale da aversi:

$$n \geq \frac{(3.29)^2}{(\mu_1 - \mu_0)^2} = \frac{10.8241}{(\mu_1 - \mu_0)^2}.$$

Una volta fissata la regione critica (5.6), perché sia $\beta(\mu_1) \leq 0.05$ quando $\mu_1 > \mu_0 + 3.29$ è allora sufficiente disporre di un campione casuale di taglia unitaria. Se, invece, si ha $\mu_1 = \mu_0 + 1$, occorre un campione di taglia non inferiore a 11. ◆

Di solito si considerano test statisticci di due tipi: test unilaterali e test bilaterali. Se le ipotesi sono del tipo

$$H_0: \theta \leq \theta_0$$

$$H_1: \theta > \theta_0,$$

oppure

$$H_0: \theta = \theta_0$$

$$H_1: \theta > \theta_0,$$

il test è detto unilaterale. Se invece le ipotesi si presentano nella forma

$$H_0: \theta = \theta_0$$

$$H_1: \theta \neq \theta_0,$$

oppure

$$H_0: \theta_0 \leq \theta \leq \theta_1$$

$$H_1: \theta < \theta_0 \text{ oppure } \theta > \theta_1,$$

il test viene detto bilaterale.

Osserviamo che dalla (5.2) è evidente che la probabilità d'errore di I tipo varia al variare di θ in Θ_0 . Poiché si desidera che tale probabilità sia piccola per ogni $\theta \in \Theta_0$, ruolo fondamentale gioca l'estremo superiore di $\alpha(\theta)$, il che giustifica la definizione che segue.

Definizione 5.1.3 L'ampiezza di un test dell'ipotesi nulla $H_0: \theta \in \Theta_0$ contro l'ipotesi alternativa $H_1: \theta \in \Theta_1$ è data da

$$\alpha = \sup_{\theta \in \Theta_0} \alpha(\theta).$$

L'ampiezza del test viene anche detta *ampiezza della regione critica*; essa fornisce una valutazione della probabilità massima di occorrenza di un errore di I tipo al variare di θ in Θ_0 .

Esempio 5.1.5 In base alla Definizione 5.1.3 l'ampiezza della regione critica (5.1) del test bilaterale trattato nell'Esempio 5.1.2 vale

$$\begin{aligned} \alpha &= P[(X_1, X_2, \dots, X_{10}) \in C | \theta \in \Theta_0] \\ &= P\left(\sum_{i=1}^{10} X_i \leq 1 \mid \theta = \frac{1}{2}\right) + P\left(\sum_{i=1}^{10} X_i \geq 9 \mid \theta = \frac{1}{2}\right) \\ &= P(Y \leq 1) + P(Y \geq 9), \end{aligned}$$

5.1. VERIFICA DELLE IPOTESI

dove Y è una variabile casuale binomiale di parametri 10 e 1/2. Si ha quindi:

$$\alpha = \left(\frac{1}{2}\right)^{10} \left[\sum_{j=0}^1 \binom{10}{j} + \sum_{j=9}^{10} \binom{10}{j} \right] = \frac{11}{2^9} = 0.0215.$$

Se ne conclude che adottando la (5.1) quale regione critica si commette un errore di I tipo con probabilità di circa il 2%; se, invece, si fosse utilizzata la regione critica

$$C' = \left\{ (x_1, x_2, \dots, x_{10}) : \sum_{i=1}^{10} x_i \leq 2 \text{ oppure } \sum_{i=1}^{10} x_i \geq 8 \right\}$$

si sarebbe avuta ampiezza

$$\alpha' = \left(\frac{1}{2}\right)^{10} \left[\sum_{j=0}^2 \binom{10}{j} + \sum_{j=8}^{10} \binom{10}{j} \right] = \frac{56}{2^9} = 0.1094,$$

che è di circa cinque volte maggiore della precedente. ◆

Nei test coinvolgenti un parametro incognito θ in cui sia l'ipotesi nulla che l'ipotesi alternativa sono composte, risulta in genere complicato individuare i criteri da adottare per costruire la regione critica del test e per fissarne l'ampiezza in quanto le probabilità di entrambi i tipi di errori dipendono da θ . Per esprimere sinteticamente la dipendenza di $\alpha(\theta)$ e di $\beta(\theta)$ da θ si definisce una funzione particolare, detta *funzione potenza* o più brevemente *potenza*, come indicato nella definizione che segue.

Definizione 5.1.4 La potenza $\pi(\theta)$ di un test dell'ipotesi nulla $H_0: \theta \in \Theta_0$ contro l'ipotesi alternativa $H_1: \theta \in \Theta_1$ è così definita:

$$\pi(\theta) = \begin{cases} \alpha(\theta) & \text{per } \theta \in \Theta_0, \\ 1 - \beta(\theta) & \text{per } \theta \in \Theta_1. \end{cases}$$

Ricordando le definizioni (5.2) e (5.3), si riconosce che i valori della potenza coincidono con le probabilità di rifiutare l'ipotesi nulla H_0 al variare di θ nello spazio dei parametri Θ . Notiamo che per $\theta \in \Theta_0$ la potenza uguaglia la probabilità di commettere un errore del I tipo, mentre per $\theta \in \Theta_1$ essa è uguale alla probabilità di non commettere un errore del II tipo. Un test ideale è dunque quello la cui potenza è nulla per $\theta \in \Theta_0$ e unitaria per $\theta \in \Theta_1$; è infatti in tal caso nulla la probabilità di errori sia di I che di II tipo, di modo che si accetta l'ipotesi nulla H_0 quando essa è vera e la si rifiuta quando è falsa. La potenza risulta quindi utile per valutare la bontà di un test.

Esempio 5.1.6 Consideriamo una popolazione normale di media μ incognita e varianza $\sigma^2 = 1$ e analizziamo la potenza del test consistente nel sottoporre a verifica l'ipotesi nulla composta $H_0: \mu \leq 1$ contro l'ipotesi alternativa composta $H_1: \mu > 1$. Indicando con (x_1, x_2, \dots, x_n) la generica realizzazione di un campione casuale di taglia n estratto dalla popolazione in questione, scegliamo la regione critica del test nel modo seguente:

$$C = \left\{ (x_1, x_2, \dots, x_n) : \bar{x} \geq 1 + \frac{z_\alpha}{\sqrt{n}} \right\}.$$

Dalle (5.2) e (5.3) segue che le probabilità di commettere errori di I e II tipo sono rispettivamente:

$$\alpha(\mu) = P\left(\bar{X} \geq 1 + \frac{z_\alpha}{\sqrt{n}} \mid \mu \leq 1\right) \quad (\mu \leq 1)$$

$$\beta(\mu) = P\left(\bar{X} < 1 + \frac{z_\alpha}{\sqrt{n}} \mid \mu > 1\right) \quad (\mu > 1).$$

Posto poi $Z = (\bar{X} - \mu)\sqrt{n}$, le probabilità $\alpha(\mu)$ e $\beta(\mu)$ assumono le seguenti forme:

$$\alpha(\mu) = P[Z \geq (1-\mu)\sqrt{n} + z_\alpha] \quad (\mu \leq 1)$$

$$\beta(\mu) = P[Z < (1-\mu)\sqrt{n} + z_\alpha] \quad (\mu > 1),$$

dove Z ha distribuzione normale standard. Supponiamo, per concretezza, che il test si riferisca ad un campione di taglia 4 e poniamo $\alpha = 0.123$ di modo che dalla Tabella 1 dell'Appendice B si ottiene $z_\alpha = z_{0.123} = 1.16$. Segue allora:

$$\alpha(\mu) = P[Z \geq 2(1-\mu) + 1.16] \quad (\mu \leq 1)$$

$$\beta(\mu) = P[Z < 2(1-\mu) + 1.16] \quad (\mu > 1).$$

In virtù della Definizione 5.1.4 la potenza del test è allora:

$$\pi(\mu) = P[Z \geq 2(1-\mu) + 1.16] \quad (\mu \in \mathbb{R}).$$

Nella Tabella 5.1 sono riportate le probabilità di errore $\alpha(\mu)$ e $\beta(\mu)$, corrispondenti a valori di μ in $(0, 1]$ per $\alpha(\mu)$ e in $[1, 2.5]$ per $\beta(\mu)$, ricavate mediante la Tabella 1 dell'Appendice B. Sono anche indicati i valori della potenza $\pi(\mu)$ del test in corrispondenza di alcune scelte di μ . La Figura 5.1 mostra il grafico della funzione $\pi(\mu)$ al variare di μ . ◆

Esempio 5.1.7 Dato un campione casuale $(X_1, X_2, \dots, X_{20})$ costituito da variabili di Bernoulli di parametro θ , verifichiamo l'ipotesi nulla semplice $H_0: \theta = 0.9$ contro l'ipotesi alternativa semplice $H_1: \theta = 0.6$. Assumiamo che la regione di rifiuto per l'ipotesi nulla H_0 , ossia la regione critica, sia così specificata:

$$C = \left\{ (x_1, x_2, \dots, x_{20}): \sum_{i=1}^{20} x_i \leq 14 \right\}.$$

Posto $x = x_1 + x_2 + \dots + x_{20}$, le alternative possibili sono dunque le seguenti:

- se x è maggiore di 14 si accetta l'ipotesi nulla H_0 ;
- se x risulta minore di 15 si rifiuta l'ipotesi nulla H_0 (e quindi si accetta l'ipotesi alternativa H_1).

Facciamo uso delle (5.2) e (5.3) per valutare le probabilità di incorrere in errori di I tipo e di II tipo. Ricordando che la variabile casuale $X = X_1 + X_2 + \dots + X_{20}$ ha distribuzione

Tabella 5.1: Probabilità di errori di I e II tipo e potenza del test per l'Esempio 5.1.5.

μ	$\alpha(\mu)$	$\beta(\mu)$	$\pi(\mu)$
0.5	0.0154		0.0154
0.6	0.0250		0.0250
0.7	0.0392		0.0392
0.8	0.0594		0.0594
0.9	0.0869		0.0869
1	0.1230		0.1230
1.1		0.8315	0.1685
1.2		0.7764	0.2236
1.3		0.7123	0.2877
1.4		0.6406	0.3594
1.5		0.5636	0.4364
1.6		0.4840	0.5160
1.7		0.4052	0.5948
1.8		0.3300	0.6700
1.9		0.2611	0.7389
2		0.2005	0.7995
2.1		0.1492	0.8508
2.2		0.1075	0.8925
2.3		0.0749	0.9251
2.4		0.0505	0.9495
2.5		0.0329	0.9671

binomiale di parametro θ , poiché si ha $\Theta_0 = \{0.9\}$ e $\Theta_1 = \{0.6\}$, risulta:

$$\alpha(0.9) = P(X \leq 14 \mid \theta = 0.9) = \sum_{x=0}^{14} \binom{20}{x} (0.9)^x (0.1)^{20-x} = 0.0114$$

$$\beta(0.6) = P(X \geq 15 \mid \theta = 0.6) = \sum_{x=15}^{20} \binom{20}{x} (0.6)^x (0.4)^{20-x} = 0.1255.$$

Dalla Definizione 5.1.3 segue l'ampiezza del test:

$$\alpha = \sup_{\theta \in \{0.9\}} \alpha(\theta) = \alpha(0.9) = 0.0114.$$

Esaminiamo che cosa accade allorché si modifica la regione critica. Supponiamo, ad esempio, che la regione di rifiuto per l'ipotesi nulla H_0 sia

$$C = \left\{ (x_1, x_2, \dots, x_{20}): \sum_{i=1}^{20} x_i \leq 15 \right\}.$$

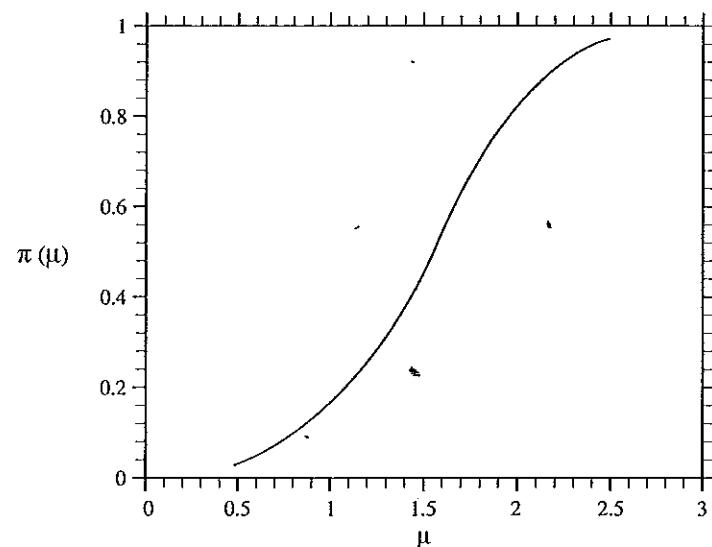


Figura 5.1: Potenza del test di cui all'Esempio 5.1.4.

In tal caso risulta:

$$\alpha(0.9) = P(X \leq 15 | \theta = 0.9) = \sum_{x=0}^{15} \binom{20}{x} (0.9)^x (0.1)^{20-x} = 0.0433$$

$$\beta(0.6) = P(X \geq 16 | \theta = 0.6) = \sum_{x=16}^{20} \binom{20}{x} (0.6)^x (0.4)^{20-x} = 0.0509$$

mentre l'ampiezza del test diventa:

$$\alpha = \sup_{\theta \in \{\theta_0\}} \alpha(\theta) = \alpha(0.9) = 0.0433.$$

◆

Nell'esempio appena discusso è accaduto che ampliando la regione critica del test la probabilità d'errore di II tipo è diminuita mentre la probabilità d'errore di I tipo è aumentata. Tale situazione è tipica dei problemi di decisione statistica: per campioni casuali di taglia fissata, se al variare della regione critica diminuisce la probabilità di un tipo d'errore concomitamente aumenta la probabilità di commettere un errore dell'altro tipo; ciò è d'altronde anche evidente dalle definizioni (5.2) e (5.3). Una maniera per tentare di ridurre entrambe le probabilità $\alpha(\theta)$ e $\beta(\theta)$ consiste nell'aumentare la taglia del campione, il che non è però sempre possibile.

5.2. LEMMA DI NEYMAN-PEARSON

Le considerazioni svolte conducono a scegliere dei test per i quali sia fissata la probabilità d'errore di I tipo ed a ricercare, tra questi, un test per il quale sia minima la probabilità d'errore di II tipo. Si giunge così alla definizione che segue.

Definizione 5.1.5 Un test ϕ di ampiezza α per verificare l'ipotesi nulla $H_0: \theta \in \Theta_0$ contro l'ipotesi alternativa $H_1: \theta \in \Theta_1$ si dice uniformemente più potente se per ogni altro test ϕ' di ampiezza $\alpha' \leq \alpha$ risulta $\beta'(\theta) \geq \beta(\theta)$ per ogni $\theta \in \Theta_1$.

Un test uniformemente più potente di ampiezza α è quindi tale da rendere minima la probabilità d'errore di II tipo.

5.2 Lemma di Neyman-Pearson

In questo paragrafo analizzeremo un metodo che consente di costruire test uniformemente più potenti di ampiezza α nel caso della verifica di ipotesi sulle semplici contro ipotesi alternative semplici. In questo caso particolare, quando cioè entrambe le ipotesi sono semplici, un test uniformemente più potente viene detto *semplicemente più potente*.

Dato un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione caratterizzata da un parametro θ , sia $\Theta = \{\theta_0, \theta_1\}$ lo spazio dei parametri per θ . Si desidera verificare l'ipotesi nulla semplice $H_0: \theta = \theta_0$ contro l'ipotesi alternativa semplice $H_1: \theta = \theta_1$. Indicando con C la regione critica del test, dalla Definizione 5.1.3 segue che, essendo $\Theta_0 = \{\theta_0\}$, l'ampiezza del test è data da

$$\alpha = \sup_{\theta \in \Theta_0} \alpha(\theta) = \alpha(\theta_0) = P[(X_1, X_2, \dots, X_n) \in C | \theta = \theta_0].$$

Dalla Definizione 5.1.4 segue poi la potenza del test:

$$\pi(\theta) = \begin{cases} \alpha(\theta_0) & \text{per } \theta = \theta_0, \\ 1 - \beta(\theta_1) & \text{per } \theta = \theta_1. \end{cases}$$

Esamineremo ora un teorema dovuto a J. Neyman e E.S. Pearson, noto anche quale *Lemma di Neyman-Pearson* per il ruolo fondamentale che esso riveste in inferenza statistica, che fornisce un metodo per costruire test semplicemente più potenti.

Teorema 5.2.1 (Neyman-Pearson) Siano (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione caratterizzata da un parametro incognito θ , (x_1, x_2, \dots, x_n) una sua realizzazione e $L(\theta; x_1, x_2, \dots, x_n)$ la corrispondente funzione di verosimiglianza. Il test semplicemente più potente di ampiezza α per verificare l'ipotesi nulla $H_0: \theta = \theta_0$ contro l'ipotesi alternativa $H_1: \theta = \theta_1$ è quello di regione critica

$$C = \left\{ (x_1, x_2, \dots, x_n) : \frac{L(\theta_0; x_1, x_2, \dots, x_n)}{L(\theta_1; x_1, x_2, \dots, x_n)} \leq k \right\}, \quad (5.7)$$

dove k è una costante tale da rendere uguale ad α l'ampiezza di C .

Dim. Ci limitiamo a dimostrare il teorema nel caso di un campione casuale estratto da popolazione continua; il caso di popolazione discreta si tratta in modo analogo. Sia C una regione critica di ampiezza α che soddisfa le ipotesi del teorema e sia \mathcal{D} un'altra regione critica anch'essa di ampiezza α . Si ha:

$$P[(X_1, X_2, \dots, X_n) \in C | \theta = \theta_0] = P[(X_1, X_2, \dots, X_n) \in \mathcal{D} | \theta = \theta_0] = \alpha,$$

ossia:

$$\int_C L(\theta_0; \mathbf{x}) d\mathbf{x} = \int_{\mathcal{D}} L(\theta_0; \mathbf{x}) d\mathbf{x} = \alpha, \quad (5.8)$$

dove \mathbf{x} denota la n -upla (x_1, x_2, \dots, x_n) e $d\mathbf{x}$ il prodotto $dx_1 dx_2 \cdots dx_n$. Osserviamo che C è l'unione degli insiemi disgiunti $C \cap \mathcal{D}$ e $C \cap \bar{\mathcal{D}}$, mentre \mathcal{D} è l'unione degli insiemi disgiunti $C \cap \mathcal{D}$ e $\bar{C} \cap \mathcal{D}$; pertanto risulta:

$$\begin{aligned} \int_C L(\theta_0; \mathbf{x}) d\mathbf{x} &= \int_{C \cap \mathcal{D}} L(\theta_0; \mathbf{x}) d\mathbf{x} + \int_{C \cap \bar{\mathcal{D}}} L(\theta_0; \mathbf{x}) d\mathbf{x}, \\ \int_{\mathcal{D}} L(\theta_0; \mathbf{x}) d\mathbf{x} &= \int_{C \cap \mathcal{D}} L(\theta_0; \mathbf{x}) d\mathbf{x} + \int_{\bar{C} \cap \mathcal{D}} L(\theta_0; \mathbf{x}) d\mathbf{x}. \end{aligned}$$

Facendo uso di queste relazioni, la (5.8) diventa:

$$\int_{C \cap \bar{\mathcal{D}}} L(\theta_0; \mathbf{x}) d\mathbf{x} = \int_{\bar{C} \cap \mathcal{D}} L(\theta_0; \mathbf{x}) d\mathbf{x}. \quad (5.9)$$

D'altra canto, dalla (5.7) si ha $L(\theta_1; \mathbf{x}) \geq L(\theta_0; \mathbf{x})/k$ per $\mathbf{x} \in C$. Inoltre, essendo

$$\bar{C} = \left\{ (x_1, x_2, \dots, x_n) : \frac{L(\theta_0; x_1, x_2, \dots, x_n)}{L(\theta_1; x_1, x_2, \dots, x_n)} > k \right\},$$

si trae $L(\theta_1; \mathbf{x}) < L(\theta_0; \mathbf{x})/k$ per $\mathbf{x} \in \bar{C}$. Dall'identità (5.9) segue allora:

$$\begin{aligned} \int_{C \cap \bar{\mathcal{D}}} L(\theta_1; \mathbf{x}) d\mathbf{x} &\geq \int_{C \cap \bar{\mathcal{D}}} \frac{L(\theta_0; \mathbf{x})}{k} d\mathbf{x} = \int_{\bar{C} \cap \mathcal{D}} \frac{L(\theta_0; \mathbf{x})}{k} d\mathbf{x} \\ &> \int_{\bar{C} \cap \mathcal{D}} L(\theta_1; \mathbf{x}) d\mathbf{x}, \end{aligned}$$

e quindi:

$$\int_{C \cap \bar{\mathcal{D}}} L(\theta_1; \mathbf{x}) d\mathbf{x} > \int_{\bar{C} \cap \mathcal{D}} L(\theta_1; \mathbf{x}) d\mathbf{x}. \quad (5.10)$$

Facendo poi uso della (5.10) si ottiene:

$$\begin{aligned} \int_C L(\theta_1; \mathbf{x}) d\mathbf{x} &= \int_{C \cap \mathcal{D}} L(\theta_1; \mathbf{x}) d\mathbf{x} + \int_{C \cap \bar{\mathcal{D}}} L(\theta_1; \mathbf{x}) d\mathbf{x} \\ &> \int_{C \cap \mathcal{D}} L(\theta_1; \mathbf{x}) d\mathbf{x} + \int_{\bar{C} \cap \mathcal{D}} L(\theta_1; \mathbf{x}) d\mathbf{x} \\ &= \int_{\mathcal{D}} L(\theta_1; \mathbf{x}) d\mathbf{x}. \end{aligned} \quad (5.11)$$

5.2. LEMMA DI NEYMAN-PEARSON

La diseguaglianza

$$\int_C L(\theta_1; \mathbf{x}) d\mathbf{x} > \int_{\mathcal{D}} L(\theta_1; \mathbf{x}) d\mathbf{x},$$

espressa dalla (5.11) equivale ad affermare che risulta

$$\begin{aligned} 1 - \beta_C(\theta) &\equiv P[(X_1, X_2, \dots, X_n) \in C | \theta = \theta_1] \\ &> P[(X_1, X_2, \dots, X_n) \in \mathcal{D} | \theta = \theta_1] \\ &\equiv 1 - \beta_{\mathcal{D}}(\theta), \end{aligned}$$

da cui segue:

$$\beta_C(\theta) < \beta_{\mathcal{D}}(\theta).$$

Per la regione critica C , soddisfacente le ipotesi (5.7), la probabilità di commettere un errore di II tipo è quindi minore di quella corrispondente ad ogni altra regione critica \mathcal{D} avente ampiezza α . In virtù della Definizione 5.1.5 il test avente regione critica C è pertanto il più potente per verificare l'ipotesi nulla semplice $H_0: \theta = \theta_0$ contro l'ipotesi alternativa semplice $H_1: \theta = \theta_1$. ■

Esaminiamo alcuni risultati indicanti come si possa utilizzare il Lemma di Neyman-Pearson per determinare test semplicemente più potenti. Cominciamo con l'esaminare ipotesi riguardanti media e varianza di una popolazione normale.

Proposizione 5.2.1 *Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione normale di media μ incognita e di varianza σ^2 nota. Il test semplicemente più potente di ampiezza α per verificare l'ipotesi nulla semplice $H_0: \mu = \mu_0$ contro l'ipotesi alternativa semplice $H_1: \mu = \mu_1$ (con $\mu_1 > \mu_0$) è quello di regione critica*

$$C = \left\{ (x_1, x_2, \dots, x_n) : \bar{x} \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} \right\}. \quad (5.12)$$

Dim. In virtù del Teorema 5.2.1 la regione critica del test deve avere la forma (5.7). Essendo

$$L(\mu) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right],$$

si ha:

$$\begin{aligned} \frac{L(\mu_0)}{L(\mu_1)} &= \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n [(x_i - \mu_0)^2 - (x_i - \mu_1)^2] \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n [\mu_0^2 - \mu_1^2 - 2x_i(\mu_0 - \mu_1)] \right\} \\ &= \exp \left[\frac{n}{2\sigma^2} (\mu_1^2 - \mu_0^2) + \frac{\mu_0 - \mu_1}{\sigma^2} \sum_{i=1}^n x_i \right] \\ &= \exp \left[\frac{n}{2\sigma^2} (\mu_1^2 - \mu_0^2) - \frac{n}{\sigma^2} (\mu_1 - \mu_0) \bar{x} \right] \\ &= \exp \left[\frac{n}{\sigma^2} (\mu_1 - \mu_0) \left(\frac{\mu_1 + \mu_0}{2} - \bar{x} \right) \right]. \end{aligned}$$

È allora $L(\mu_0)/L(\mu_1) \leq k$ se e solo se è

$$\exp \left[\frac{n}{\sigma^2} (\mu_1 - \mu_0) \left(\frac{\mu_1 + \mu_0}{2} - \bar{x} \right) \right] \leq k, \quad (5.13)$$

ossia, essendo $\mu_1 > \mu_0$, se e solo se risulta

$$\bar{x} \geq \frac{\mu_1 + \mu_0}{2} - \frac{\sigma^2 \ln k}{n(\mu_1 - \mu_0)}. \quad (5.14)$$

Per le (5.7) e (5.14) e per l'arbitrarietà di k segue che la regione critica C deve essere del tipo

$$C = \{(x_1, x_2, \dots, x_n) : \bar{x} \geq A\},$$

dove A è una costante da determinarsi in modo che C abbia ampiezza α . Per la Definizione 5.1.3 di ampiezza di un test si vede che per determinare la costante A occorre risolvere l'equazione:

$$P(\bar{X} \geq A | \mu = \mu_0) = P \left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{A - \mu_0}{\sigma/\sqrt{n}} \mid \mu = \mu_0 \right) = \alpha. \quad (5.15)$$

Osserviamo a tal fine che sotto l'ipotesi nulla H_0 la variabile \bar{X} ha distribuzione normale di media μ_0 e varianza σ^2/n , così che la variabile casuale $Z \stackrel{\text{def}}{=} (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ è normale standard. Ricordiamo poi (cfr. § 1.3) che se Z è una variabile normale standard, con z_α si denota il reale tale che $P(Z \geq z_\alpha) = \alpha$. Se allora si pone $(A - \mu_0)/(\sigma/\sqrt{n}) = z_\alpha$, la (5.15) è soddisfatta. Si ha quindi $A = \mu_0 + z_\alpha \sigma/\sqrt{n}$, da cui segue che il test corrispondente alla regione critica (5.12) risulta essere il test più potente per verificare l'ipotesi nulla $H_0: \mu = \mu_0$ contro l'ipotesi alternativa $H_1: \mu = \mu_1$. ■

È interessante notare che la regione critica (5.12) cambia al variare di α e σ . È infatti evidente che $A = \mu_0 + z_\alpha \sigma/\sqrt{n}$ diverge positivamente al tendere di α a zero o di σ a ∞ , di modo che corrispondentemente la regione critica tende all'insieme vuoto.

Nel caso in cui risulta $\mu_0 > \mu_1$, procedendo in modo analogo al caso trattato nella Proposizione 5.2.1 è facile ricavare che alla regione critica

$$C = \{(x_1, x_2, \dots, x_n) : \bar{x} \leq \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}\}.$$

di ampiezza α corrisponde il test semplicemente più potente per verificare l'ipotesi nulla $H_0: \mu = \mu_0$ contro l'ipotesi alternativa $H_1: \mu = \mu_1$.

Esempio 5.2.1 La temperatura in gradi centigradi registrata periodicamente in una regione sottoposta a controllo termico è riguardata come una variabile casuale normale di media $\mu_0 = 18$ e deviazione standard $\sigma = 2$. Sulla base dei seguenti dati:

$$(18.3, 19.5, 20.8, 17.5, 19.8, 21.3, 16.8, 20.6, 18.7, 21.5)$$

registrati nel corso delle ultime 10 misure si ritiene che si sia verificata una variazione significativa della temperatura da quantificarsi in un aumento medio di 2 unità. Ci si pone il problema di verificare se la media μ incognita sia pari a $\mu_0 = 18$ oppure a $\mu_1 = 20$. Riguardando

5.2. LEMMA DI NEYMAN-PEARSON

le temperature misurate come realizzazione $(x_1, x_2, \dots, x_{10})$ di un campione casuale di taglia 10 estratto da una popolazione normale di media μ incognita e di varianza $\sigma^2 = 4$, si formano pertanto le ipotesi $H_0: \mu = 18$ contro $H_1: \mu = 20$ da sottoporre a verifica. Si costruisce quindi un test uniformemente più potente per μ in accordo con la Proposizione 5.2.1, ossia scegliendo la regione critica nel seguente modo:

$$C = \{(x_1, x_2, \dots, x_{10}) : \bar{x} \geq 18 + z_\alpha \frac{2}{\sqrt{10}}\}.$$

Se si pone $\alpha = 0.025$ come ampiezza della regione critica, dalla Tabella 2 dell'Appendice B segue $z_\alpha = z_{0.025} = 1.96$, e quindi

$$C = \{(x_1, x_2, \dots, x_{10}) : \bar{x} \geq 19.2396\}.$$

Poiché la media campionaria assume il valore $\bar{x} = 19.48$, la realizzazione osservata appartiene alla regione critica C di ampiezza $\alpha = 0.025$. Il test effettuato sulla base dei dati disponibili suggerisce allora di rifiutare l'ipotesi nulla $\mu = 18$ a favore dell'ipotesi alternativa $\mu = 20$. Si accetta dunque l'ipotesi che si sia verificato un aumento medio di 2 gradi nella temperatura della regione esaminata. ♦

Proposizione 5.2.2 Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione normale di media μ nota e di varianza σ^2 incognita. Il test semplicemente più potente di ampiezza α per verificare l'ipotesi nulla semplice $H_0: \sigma^2 = \sigma_0^2$ contro l'ipotesi alternativa semplice $H_1: \sigma^2 = \sigma_1^2$ (con $\sigma_1^2 > \sigma_0^2$) è quello che ha come regione critica

$$C = \{(x_1, x_2, \dots, x_n) : \sum_{i=1}^n (x_i - \mu)^2 \geq \sigma_0^2 \chi_{\alpha, n}^2\}. \quad (5.16)$$

Dim. Per il Teorema 5.2.1 la regione critica del test deve essere scelta come specificato dalla (5.7). Dalla funzione di verosimiglianza

$$L(\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right]$$

segue:

$$\begin{aligned} \frac{L(\sigma_0^2)}{L(\sigma_1^2)} &= \left(\frac{\sigma_1^2}{\sigma_0^2} \right)^{n/2} \exp \left[-\frac{1}{2} \left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2} \right) \sum_{i=1}^n (x_i - \mu)^2 \right] \\ &= \left(\frac{\sigma_1^2}{\sigma_0^2} \right)^{n/2} \exp \left[\frac{\sigma_0^2 - \sigma_1^2}{2\sigma_1^2} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma_0} \right)^2 \right]. \end{aligned}$$

Pertanto risulta $L(\sigma_0^2)/L(\sigma_1^2) \leq k$ se e solo se è

$$\left(\frac{\sigma_1^2}{\sigma_0^2} \right)^{n/2} \exp \left[\frac{\sigma_0^2 - \sigma_1^2}{2\sigma_1^2} \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma_0} \right)^2 \right] \leq k. \quad (5.17)$$

Essendo $\sigma_1^2 > \sigma_0^2$, la (5.17) diventa:

$$\sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma_0} \right)^2 \geq \frac{2\sigma_1^2}{\sigma_1^2 - \sigma_0^2} \left[\ln \frac{1}{k} + \frac{n}{2} \ln \left(\frac{\sigma_1^2}{\sigma_0^2} \right) \right].$$

Ricordando la (5.7), per l'arbitrarietà di k segue che la regione critica C deve essere del seguente tipo:

$$C = \left\{ (x_1, x_2, \dots, x_n) : \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma_0} \right)^2 \geq A \right\}.$$

La costante A è da determinarsi in modo che C abbia ampiezza α ; occorre quindi risolvere la seguente equazione:

$$\alpha = P \left[\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma_0} \right)^2 \geq A \mid \sigma^2 = \sigma_0^2 \right]. \quad (5.18)$$

Sotto l'ipotesi nulla H_0 le variabili X_i sono indipendenti ed hanno distribuzione normale di valore medio μ e varianza σ_0^2 . Le variabili casuali $(X_i - \mu)^2 / \sigma_0^2$ sono allora indipendenti ed hanno distribuzione chi-quadrato con 1 grado di libertà; la loro somma ha pertanto distribuzione chi-quadrato con n gradi di libertà. Ricordiamo (cfr. § 2.1), che se Y è una variabile casuale chi-quadrato con n gradi di libertà, con $\chi_{\alpha;n}^2$ si denota il reale tale che $P(Y \geq \chi_{\alpha;n}^2) = \alpha$; dalla (5.18) segue dunque $A = \chi_{\alpha;n}^2$. Il test corrispondente alla regione critica di ampiezza α (5.16) è quindi il test semplicemente più potente per verificare l'ipotesi nulla $H_0: \sigma^2 = \sigma_0^2$ contro l'ipotesi alternativa $H_1: \sigma^2 = \sigma_1^2$ nel caso in cui è $\sigma_1^2 > \sigma_0^2$. ■

Procedendo analogamente alla dimostrazione della Proposizione 5.2.2, si può ricavare facilmente che la regione critica

$$C = \left\{ (x_1, x_2, \dots, x_n) : \sum_{i=1}^n (x_i - \mu)^2 \leq \sigma_0^2 \chi_{1-\alpha;n}^2 \right\}$$

di ampiezza α risulta corrispondere al test semplicemente più potente per verificare l'ipotesi nulla $H_0: \sigma^2 = \sigma_0^2$ contro l'ipotesi alternativa $H_1: \sigma^2 = \sigma_1^2$ quando risulta $\sigma_0^2 > \sigma_1^2$.

Esempio 5.2.2 Ad un campione di 16 studenti vengono sottoposti dei quesiti da risolversi il più rapidamente possibile. I tempi, in minuti, che essi impiegano possono riguardarsi come realizzazione $(x_1, x_2, \dots, x_{16})$ di un campione casuale estratto da una popolazione normale di media $\mu = 20$. Tali tempi risultano essere i seguenti:

$$(20.4, 21.2, 21.0, 22.1, 20.4, 19.5, 18.8, 21.9, \\ 20.2, 18.3, 21.5, 18.6, 19.8, 18.2, 19.5, 20.3).$$

Si formuli sulla varianza σ^2 l'ipotesi $H_0: \sigma^2 = 1.5$ contro l'ipotesi $H_1: \sigma^2 = 2$. Ricordando la Proposizione 5.2.2 si costruisce un test uniformemente più potente per σ^2 scegliendo la regione critica secondo la (5.16). Si ha quindi:

$$C = \left\{ (x_1, x_2, \dots, x_{16}) : \sum_{i=1}^n (x_i - 20)^2 \geq 1.5 \chi_{\alpha;16}^2 \right\}.$$

5.2. LEMMA DI NEYMAN-PEARSON

Scegliendo $\alpha = 0.05$ come ampiezza della regione critica, dalla Tabella 3 dell'Appendice B si ha $\chi_{\alpha;16}^2 = \chi_{0.05;16}^2 = 26.296$. Pertanto la regione critica diventa:

$$C = \left\{ (x_1, x_2, \dots, x_{16}) : \sum_{i=1}^n (x_i - 20)^2 \geq 39.444 \right\}.$$

Poiché nell'esperimento effettuato risulta

$$\sum_{i=1}^n (x_i - 20)^2 = 23.23,$$

la realizzazione osservata non appartiene alla regione critica. Non c'è dunque motivo per rifiutare l'ipotesi nulla $\sigma^2 = 1.5$. ♦

Saranno ora indicati altri risultati in cui si mostra come il Lemma di Neyman-Pearson possa essere utilizzato per ricavare test semplicemente più potenti nel caso di ipotesi concernenti i parametri di popolazioni di Bernoulli, esponenziale e geometrica.

Proposizione 5.2.3 *Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione di Bernoulli di parametro θ incognito. Il test semplicemente più potente di ampiezza α per verificare l'ipotesi nulla semplice $H_0: \theta = \theta_0$ contro l'ipotesi alternativa semplice $H_1: \theta = \theta_1$ (con $\theta_1 > \theta_0$) è quello di regione critica*

$$C = \left\{ (x_1, x_2, \dots, x_n) : \sum_{i=1}^n x_i \geq k_\alpha \right\}, \quad (5.19)$$

dove k_α è il minimo intero tale da aversi

$$\sum_{r=k_\alpha}^n \binom{n}{r} \theta_0^r (1-\theta_0)^{n-r} \leq \alpha.$$

Dim. In virtù del Teorema 5.2.1, la regione critica del test deve essere scelta in accordo con la (5.7). Dalla funzione di verosimiglianza

$$L(\theta) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} = \theta^x (1-\theta)^{n-x},$$

in cui si è posto $x = x_1 + x_2 + \dots + x_n$, segue:

$$\frac{L(\theta_0)}{L(\theta_1)} = \left(\frac{\theta_0}{\theta_1} \right)^x \left(\frac{1-\theta_0}{1-\theta_1} \right)^{n-x}.$$

Si ha allora $L(\theta_0)/L(\theta_1) \leq k$ se e solo se risulta

$$\left(\frac{\theta_0}{\theta_1} \right)^x \left(\frac{1-\theta_0}{1-\theta_1} \right)^{n-x} \leq k. \quad (5.20)$$

Essendo $\theta_1 > \theta_0$, si ha:

$$x \geq \left(\ln \frac{\theta_1}{\theta_0} + \ln \frac{1-\theta_0}{1-\theta_1} \right)^{-1} \left(\ln \frac{1}{k} + n \ln \frac{1-\theta_0}{1-\theta_1} \right).$$

Per l'arbitrarietà di k , la regione critica si può esprimere nel seguente modo:

$$C = \left\{ (x_1, x_2, \dots, x_n) : \sum_{i=1}^n x_i \geq A \right\},$$

dove A va determinato in modo che C abbia ampiezza α . Ricordando la Definizione 5.1.3 di ampiezza di un test, occorre dunque risolvere l'equazione

$$P(X \geq A | \theta = \theta_0) = \alpha,$$

dove $X = X_1 + X_2 + \dots + X_n$ è una variabile casuale binomiale in quanto somma di variabili casuali di Bernoulli. In particolare, sotto l'ipotesi nulla H_0 la variabile X ha distribuzione binomiale di parametro θ_0 . Si sceglie quindi $A = k_\alpha$, dove k_α è il minimo intero tale da risultare

$$P(X \geq k_\alpha | \theta = \theta_0) = \sum_{r=k_\alpha}^n \binom{n}{r} \theta_0^r (1-\theta_0)^{n-r} \leq \alpha.$$

In tal modo l'ampiezza della regione critica è il più possibile prossima ad α , pur senza eccedere questo valore. Quando risulta $\theta_1 > \theta_0$, il test corrispondente alla regione critica (5.19) di ampiezza α è dunque il test semplicemente più potente per verificare l'ipotesi nulla $H_0: \theta = \theta_0$ contro l'ipotesi alternativa $H_1: \theta = \theta_1$. ■

È facile ricavare che nel caso $\theta_0 > \theta_1$ il test semplicemente più potente di ampiezza α per verificare l'ipotesi nulla semplice $H_0: \theta = \theta_0$ contro l'ipotesi alternativa semplice $H_1: \theta = \theta_1$ è quello di regione critica

$$C = \left\{ (x_1, x_2, \dots, x_n) : \sum_{i=1}^n x_i \leq k'_\alpha \right\},$$

dove k'_α è il massimo intero tale da aversi

$$\sum_{r=0}^{k'_\alpha} \binom{n}{r} \theta_0^r (1-\theta_0)^{n-r} \leq \alpha.$$

Esempio 5.2.3 In un'industria automobilistica si desidera sottoporre a verifica la probabilità θ che una generica autovettura prodotta superi il collaudo sapendo che in un campione di 25 autovetture 21 di esse hanno superato il collaudo. Il campione casuale può ritenersi estratto da una popolazione di Bernoulli di parametro θ in cui la variabile i -esima assume valore 1 se la vettura supera il collaudo, 0 altrimenti. Assumiamo che su θ venga formulata l'ipotesi $H_0: \theta = 0.8$ contro l'ipotesi $H_1: \theta = 0.9$. Facendo uso della Proposizione 5.2.3 si costruisce la regione critica di ampiezza 0.05 ponendo

$$C = \left\{ (x_1, x_2, \dots, x_{25}) : \sum_{i=1}^{25} x_i \geq k_{0.05} \right\},$$

dove $k_{0.05}$ è il minimo intero tale da aversi

$$\sum_{r=k_{0.05}}^{25} \binom{25}{r} (0.8)^r (0.2)^{25-r} \leq 0.05.$$

Essendo

$$\binom{25}{r} (0.8)^r (0.2)^{25-r} = \begin{cases} 0.0038 & \text{per } r = 25, \\ 0.0236 & \text{per } r = 24, \\ 0.0708 & \text{per } r = 23, \end{cases}$$

si ha $k_{0.05} = 24$. Poiché risulta $\sum_{i=1}^{25} x_i = 21$, la realizzazione osservata non appartiene alla regione critica, così che non vi sono motivi per rifiutare l'ipotesi nulla $H_0: \theta = 0.8$. ♦

Proposizione 5.2.4 Sia X una variabile esponenziale di media θ incognita, riguardata come campione casuale di taglia unitaria estratto da una popolazione esponenziale. Il test semplicemente più potente di ampiezza α per verificare l'ipotesi nulla semplice $H_0: \theta = \theta_0$ contro l'ipotesi alternativa semplice $H_1: \theta = \theta_1$, con $\theta_1 > \theta_0$, è quello di regione critica

$$C = \left\{ x : x \geq \theta_0 \ln \frac{1}{\alpha} \right\}. \quad (5.21)$$

Dim. Per il Teorema 5.2.1 la regione critica del test deve essere scelta come indicato dalla (5.7). La funzione di verosimiglianza è ora

$$L(\theta) = \frac{1}{\theta} e^{-x/\theta} \quad (x > 0),$$

così che

$$\frac{L(\theta_0)}{L(\theta_1)} = \frac{\theta_1}{\theta_0} \exp \left[-x \left(\frac{1}{\theta_0} - \frac{1}{\theta_1} \right) \right].$$

Si ha allora $L(\theta_0)/L(\theta_1) \leq k$ se e solo se risulta

$$\frac{\theta_1}{\theta_0} \exp \left[-x \left(\frac{1}{\theta_0} - \frac{1}{\theta_1} \right) \right] \leq k. \quad (5.22)$$

Essendo $\theta_1 > \theta_0$, si ha:

$$x \geq \frac{\theta_0 \theta_1}{\theta_0 - \theta_1} \ln \left(k \frac{\theta_0}{\theta_1} \right),$$

così che la regione critica deve essere del tipo

$$C = \{x : x \geq A\}$$

dove A è da determinarsi come soluzione dell'equazione

$$P(X \geq A | \theta = \theta_0) = \alpha,$$

in modo che C abbia ampiezza α . Sotto l'ipotesi nulla H_0 la variabile X ha distribuzione esponenziale di valore medio θ_0 ; pertanto risulta

$$P(X \geq A | \theta = \theta_0) = e^{-A/\theta_0}.$$

Si giunge dunque all'equazione $e^{-A/\theta_0} = \alpha$, da cui si trae:

$$A = \theta_0 \ln \frac{1}{\alpha}.$$

La regione critica di ampiezza α è quindi data dalla (5.21).

Si noti che quando, ad esempio, risulta $\alpha = 0.05$ si ha $\ln(1/\alpha) = 2.9957$, di modo che la regione critica (5.21) di ampiezza α diventa

$$C = \{x: x \geq 2.9957\theta_0\}.$$

Ne segue che se si rifiuta l'ipotesi nulla $\theta = \theta_0$ ogni volta che si osserva un valore x maggiore di circa tre volte θ_0 , si commette un errore di I tipo con probabilità $\alpha = 0.05$. Osserviamo poi che, per la (5.3), la probabilità di commettere un errore di II tipo è

$$\beta(\theta_1) = P(X \notin C | \theta \in \Theta_1) = P\left(X < \theta_0 \ln \frac{1}{\alpha} \mid \theta = \theta_1\right) = 1 - \alpha^{\theta_0/\theta_1}.$$

Pertanto, se si pone $\alpha = 0.05$ si ottiene ad esempio:

$$\beta(\theta_1) = 1 - (0.05)^{\theta_0/\theta_1} = \begin{cases} 0.1391 & \text{per } \theta_0 = 0.05\theta_1, \\ 0.2589 & \text{per } \theta_0 = 0.1\theta_1, \\ 0.5929 & \text{per } \theta_0 = 0.3\theta_1, \\ 0.7764 & \text{per } \theta_0 = 0.5\theta_1. \end{cases}$$

Se ne conclude che solo per $\theta_1 \gg \theta_0$ la probabilità d'errore di II tipo è prossima a zero.

Nel caso $\theta_0 > \theta_1$ si ricava in modo analogo che il test semplicemente più potente di ampiezza α per verificare l'ipotesi nulla semplice $H_0: \theta = \theta_0$ contro l'ipotesi alternativa semplice $H_1: \theta = \theta_1$ ha regione critica

$$C = \left\{x: x \leq \theta_0 \ln \frac{1}{1-\alpha}\right\}.$$

Proposizione 5.2.5 Sia X una variabile geometrica di parametro θ incognito, con $0 < \theta < 1$, riguardata come campione casuale di taglia unitaria estratto da una popolazione geometrica. Il test semplicemente più potente di ampiezza α per verificare l'ipotesi nulla semplice $H_0: \theta = \theta_0$ contro l'ipotesi alternativa semplice $H_1: \theta = \theta_1$ (con $\theta_1 > \theta_0$) è quello di regione critica¹

$$C = \{x: x \geq [\log_{\theta_0} \alpha]\}. \quad (5.23)$$

Dim. Per il Lemma di Neyman-Pearson la regione critica del test deve essere scelta secondo la (5.7). Dalla funzione di verosimiglianza

$$L(\theta) = \theta^x (1-\theta)^{1-x} \quad (x = 0, 1, \dots)$$

¹ Si ricorda che con $[x]$ si intende il minimo intero non minore di x , mentre $\lfloor x \rfloor$ denota il massimo intero non maggiore di x .

si trae

$$\frac{L(\theta_0)}{L(\theta_1)} = \left(\frac{\theta_0}{\theta_1}\right)^x \frac{1-\theta_0}{1-\theta_1}.$$

Ne segue che risulta $L(\theta_0)/L(\theta_1) \leq k$ se e solo se

$$\left(\frac{\theta_0}{\theta_1}\right)^x \frac{1-\theta_0}{1-\theta_1} \leq k.$$

Per $\theta_1 > \theta_0$ questa diventa:

$$x \geq \left(\ln \frac{\theta_0}{\theta_1}\right)^{-1} \ln \left(k \frac{1-\theta_1}{1-\theta_0}\right).$$

La regione critica C è quindi del tipo

$$C = \{x: x \geq A\},$$

dove A è il minimo intero soddisfacente la relazione

$$P(X \geq A | \theta = \theta_0) \leq \alpha. \quad (5.24)$$

Invero, in tal modo si determina quel valore intero di A per il quale la regione critica C ha ampiezza α . Sotto l'ipotesi nulla H_0 la variabile X ha distribuzione geometrica di parametro θ_0 , e quindi

$$P(X \geq A | \theta = \theta_0) = \sum_{x=A}^{\infty} \theta_0^x (1-\theta_0)^{1-x} = \theta_0^A. \quad (5.25)$$

Dalle (5.24) e (5.25) segue la disequazione $\theta_0^A \leq \alpha$, da cui si trae

$$A = \lceil \log_{\theta_0} \alpha \rceil.$$

La regione critica di ampiezza α è dunque la (5.23).

Nel caso $\theta_0 > \theta_1$ è facile mostrare che la regione critica di ampiezza α del test semplicemente più potente per verificare l'ipotesi nulla semplice $H_0: \theta = \theta_0$ contro l'ipotesi alternativa semplice $H_1: \theta = \theta_1$ è la seguente:

$$C = \{x: x \leq \lfloor \log_{\theta_0} (1-\alpha) - 1 \rfloor\}.$$

5.3. Rapporto di verosimiglianze

Nei casi concreti accade raramente di dover verificare ipotesi sulle semplici contro ipotesi alternative semplici. In generale almeno una delle due ipotesi è composta, ed i test uniformemente più potenti si presentano soltanto in casi particolari.

Esporremo ora un metodo generale che si applica quando almeno una delle ipotesi da verificare è composta. Questo metodo è basato sul *test del rapporto di verosimiglianze*; esso

risulta soddisfacente soprattutto per campioni di taglia elevata ed è di carattere empirico, così che non mancano casi in cui esso fallisce. Tuttavia, sebbene non fornisca necessariamente test uniformemente più potenti, esso possiede buone qualità asintotiche e, soprattutto, è di applicazione agevole e generale.

Sia (X_1, X_2, \dots, X_n) un campione casuale di taglia n estratto da una popolazione caratterizzata da un parametro incognito θ . Indichiamo con Θ lo spazio dei parametri ripartito in due insiemi disgiunti Θ_0 e Θ_1 , con $\Theta = \Theta_0 \cup \Theta_1$. Siano $H_0: \theta \in \Theta_0$ l'ipotesi nulla che vogliamo verificare e $H_1: \theta \in \Theta_1$ l'ipotesi alternativa, dove ora assumiamo che almeno una delle due ipotesi è composta. Se, come al solito, denotiamo con $L(\theta)$ la funzione di verosimiglianza del campione e con $f(x; \theta)$ la densità [distribuzione] di probabilità della variabile casuale genitrice, risulta:

$$L(\hat{\theta}_0) = \max_{\theta \in \Theta_0} L(\theta) = \prod_{i=1}^n f(x_i; \hat{\theta}_0), \quad L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta) = \prod_{i=1}^n f(x_i; \hat{\theta}),$$

dove $\hat{\theta}_0$ costituisce la stima di massima verosimiglianza del parametro θ quando $\theta \in \Theta_0$, e $\hat{\theta}$ la stima di massima verosimiglianza di θ quando $\theta \in \Theta$. La quantità

$$\lambda \stackrel{\text{def}}{=} \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} = \frac{\max_{\theta \in \Theta_0} L(\theta)}{\max_{\theta \in \Theta} L(\theta)} = \frac{\prod_{i=1}^n f(x_i; \hat{\theta}_0)}{\prod_{i=1}^n f(x_i; \hat{\theta})}, \quad (5.26)$$

è detta *rapporto di verosimiglianze*. Trattandosi di un rapporto di valori assunti da funzioni di verosimiglianza, si ha $\lambda \geq 0$; inoltre, poiché Θ_0 è un sottoinsieme dello spazio dei parametri Θ , risulta

$$\max_{\theta \in \Theta_0} L(\theta) \leq \max_{\theta \in \Theta} L(\theta)$$

e quindi $\lambda \leq 1$. Nel caso in cui l'ipotesi nulla è falsa, ci si attende che $\max_{\theta \in \Theta_0} L(\theta)$ sia piccolo rispetto a $\max_{\theta \in \Theta} L(\theta)$, e quindi che il rapporto di verosimiglianze (5.26) sia prossimo a 0; d'altra parte, nel caso in cui l'ipotesi nulla è vera, ossia quando $\theta \in \Theta_0$, ci si attende che $\max_{\theta \in \Theta_0} L(\theta)$ sia molto vicino a $\max_{\theta \in \Theta} L(\theta)$, e quindi che il rapporto di verosimiglianze (5.26) sia prossimo ad 1. Da tali considerazioni scaturisce un criterio per costruire la regione critica del test: è infatti plausibile identificarla con l'insieme delle realizzazioni (x_1, x_2, \dots, x_n) per le quali il rapporto di verosimiglianze λ è prossimo a 0. Il metodo del rapporto di verosimiglianze afferma dunque che l'ipotesi nulla H_0 viene rifiutata se e solo se la realizzazione osservata (x_1, x_2, \dots, x_n) è tale che il rapporto di verosimiglianze λ , dato dalla (5.26), appartiene alla regione critica C così definita:

$$C \stackrel{\text{def}}{=} \{(x_1, x_2, \dots, x_n) : \lambda \leq k\}, \quad (5.27)$$

dove k è una costante compresa tra 0 e 1 scelta in maniera opportuna, ossia in modo che la regione critica (5.27) abbia ampiezza α . Il test corrispondente alla regione critica (5.27) è anche denominato "test del rapporto di verosimiglianze di ampiezza α per verificare l'ipotesi nulla $H_0: \theta \in \Theta_0$ contro l'ipotesi alternativa $H_1: \theta \in \Theta_1$ ".

Supponiamo anzitutto che l'ipotesi nulla H_0 sia semplice, con $\Theta_0 = \{\theta_0\}$. Posto

$$Y = \frac{\prod_{i=1}^n f(X_i; \theta_0)}{\prod_{i=1}^n f(X_i; \hat{\theta})},$$

5.3. RAPPORTO DI VERO SIMIGLIANZE

si sceglie dunque k in modo tale da aversi

$$P(Y \leq k | \theta = \theta_0) \equiv P\left[\frac{\prod_{i=1}^n f(X_i; \theta_0)}{\prod_{i=1}^n f(X_i; \hat{\theta})} \leq k \mid \theta = \theta_0\right] = \alpha.$$

Va osservato che, a differenza delle variabili casuali X_i , le densità [distribuzioni] di probabilità $f(\cdot; \theta_0)$ e $f(\cdot; \hat{\theta})$ non dipendono da θ . Nel caso continuo, denotando con $f_Y(y)$ la densità di probabilità della statistica Y , si sceglie k in modo che risulti:

$$\int_0^k f_Y(y) dy = \alpha.$$

Nel caso discreto si ha invece:

$$\sum_{y \leq k} P(Y = y) \leq \alpha, \quad (5.28)$$

dove k viene identificato con il massimo valore per il quale la somma (5.28) è non superiore ad α . In entrambi i casi, per la Definizione 5.1.3 la regione critica (5.27) ha ampiezza α .

Supponiamo ora che l'ipotesi nulla H_0 sia composta. Se si pone

$$Y = \frac{\prod_{i=1}^n f(X_i; \hat{\theta}_0)}{\prod_{i=1}^n f(X_i; \hat{\theta})},$$

k va scelto in modo che sia

$$\sup_{\theta \in \Theta_0} P(Y \leq k | \theta \in \Theta_0) \equiv \sup_{\theta \in \Theta_0} P\left[\frac{\prod_{i=1}^n f(X_i; \hat{\theta}_0)}{\prod_{i=1}^n f(X_i; \hat{\theta})} \leq k \mid \theta \in \Theta_0\right] = \alpha.$$

Ciò, ricordando la (5.2), equivale a richiedere che la probabilità $\alpha(\theta)$ di errore di I tipo sia minore o uguale ad α per ogni $\theta \in \Theta_0$ e sia uguale ad α , se possibile, per almeno un valore di θ in Θ_0 . In tal modo la scelta di k , per la Definizione 5.1.3, assicura che la regione critica (5.27) abbia ampiezza α .

Nel seguito vengono presentati alcuni risultati che evidenziano come si possa fare uso del metodo del rapporto di verosimiglianze per ricavare test concernenti parametri di una popolazione normale. Cominciamo con due test riguardanti la media μ .

Proposizione 5.3.1 Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione normale di media μ incognita e di varianza σ^2 nota. Il test del rapporto di verosimiglianze di ampiezza α per verificare l'ipotesi nulla $H_0: \mu \leq \mu_0$ contro l'ipotesi alternativa $H_1: \mu > \mu_0$ è quello che ha regione critica

$$C = \{(x_1, x_2, \dots, x_n) : \bar{x} \geq \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}\}. \quad (5.29)$$

Dim. Ricaviamo anzitutto la funzione di verosimiglianza del campione casuale:

$$L(\mu) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right]. \quad (5.30)$$

Notando che lo spazio dei parametri $\Theta \equiv \mathbb{R}$ è, in questo caso, ripartito negli insiemi $\Theta_0 = (-\infty, \mu_0]$ e $\Theta_1 = (\mu_0, \infty)$, determiniamo poi il rapporto di verosimiglianze. Poiché lo stimatore di massima verosimiglianza di μ è la media campionaria (cfr. Esempio 3.7.7), si ha:

$$\max_{\mu \in \mathbb{R}} L(\mu) = L(\bar{x}) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right].$$

Per calcolare $\max_{\mu \leq \mu_0} L(\mu)$ poniamoci nel caso $\bar{x} > \mu_0$, dato che il caso $\bar{x} \leq \mu_0$ individua una situazione in cui l'ipotesi nulla $H_0: \mu \leq \mu_0$ non può essere rifiutata. Osserviamo che dalla (5.30) segue:

$$\frac{d}{d\mu} L(\mu) = \frac{n}{\sigma^2} (\bar{x} - \mu) L(\mu),$$

così che $L(\mu)$ è strettamente crescente per $\mu < \bar{x}$ e quindi, in particolare, anche per $\mu \leq \mu_0$. Si ricava pertanto:

$$\max_{\mu \leq \mu_0} L(\mu) = L(\mu_0) = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_0)^2 \right].$$

Il rapporto di verosimiglianze per $\bar{x} > \mu_0$ è dunque:

$$\lambda = \frac{L(\mu_0)}{L(\bar{x})} = \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n [(x_i - \mu_0)^2 - (\bar{x} - \mu_0)^2] \right\}. \quad (5.31)$$

Osservando che risulta

$$\begin{aligned} \sum_{i=1}^n [(x_i - \mu_0)^2 - (\bar{x} - \mu_0)^2] &= n(\mu_0^2 - \bar{x}^2) + 2(\bar{x} - \mu_0) \sum_{i=1}^n x_i \\ &= n(\mu_0 - \bar{x})(\mu_0 + \bar{x} - 2\bar{x}) = n(\bar{x} - \mu_0)^2, \end{aligned}$$

la (5.31) diventa:

$$\lambda = \exp \left[-\frac{n}{2\sigma^2} (\bar{x} - \mu_0)^2 \right].$$

La diseguaglianza $\lambda \leq k$ si esprime allora al seguente modo:

$$\exp \left[-\frac{n}{2\sigma^2} (\bar{x} - \mu_0)^2 \right] \leq k,$$

ovvero,

$$|\bar{x} - \mu_0| \geq \sqrt{\frac{2\sigma^2}{n} \ln \frac{1}{k}}. \quad (5.32)$$

Poiché k è una costante arbitraria, ricordando la (5.27) e tenendo presente che è $\bar{x} > \mu_0$ dalla (5.32) segue che la regione critica \mathcal{C} è costituita dalle n -uple (x_1, x_2, \dots, x_n) per le quali risulta:

$$\bar{x} \geq A,$$

dove la costante A deve essere determinata in modo che \mathcal{C} abbia ampiezza α , ossia in modo da aversi

$$\sup_{\mu \leq \mu_0} P(\bar{X} \geq A | \mu \leq \mu_0) = \alpha.$$

Poiché la media campionaria \bar{X} sotto l'ipotesi nulla H_0 ha distribuzione normale di media $\mu \leq \mu_0$ e varianza σ^2/n , scegliamo A in modo che risulti

$$\sup_{\mu \leq \mu_0} P \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{A - \mu}{\sigma/\sqrt{n}} \mid \mu \leq \mu_0 \right) = \sup_{\mu \leq \mu_0} P \left(Z \geq \frac{A - \mu}{\sigma/\sqrt{n}} \right) = \alpha,$$

dove $Z \stackrel{\text{def}}{=} (\bar{X} - \mu)/(\sigma/\sqrt{n})$ è una variabile casuale normale standard. Essendo

$$\sup_{\mu \leq \mu_0} P \left(Z \geq \frac{A - \mu}{\sigma/\sqrt{n}} \right) = P \left(Z \geq \frac{A - \mu_0}{\sigma/\sqrt{n}} \right),$$

e ricordando che z_α è il reale tale da aversi $P(Z \geq z_\alpha) = \alpha$, poniamo

$$\frac{A - \mu_0}{\sigma/\sqrt{n}} = z_\alpha,$$

così che

$$A = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}.$$

Se ne conclude che la (5.29) è la regione critica di ampiezza α del test del rapporto di verosimiglianze. ■

Dalla Proposizione 5.3.1 discende che l'ipotesi nulla $H_0: \mu \leq \mu_0$ viene rifiutata se si osserva una realizzazione (x_1, x_2, \dots, x_n) tale da aversi $\bar{x} \geq \mu_0 + z_\alpha \sigma/\sqrt{n}$, mentre essa viene accettata se si ha $\bar{x} < \mu_0 + z_\alpha \sigma/\sqrt{n}$. Ciò giustifica l'ipotesi $\bar{x} > \mu_0$ considerata nel corso della dimostrazione. È inoltre interessante notare che quello fornito dalla Proposizione 5.3.1 è un test uniformemente più potente di ampiezza α giacché coincide con quello ottenuto nella Proposizione 5.2.1.

Proposizione 5.3.2 *Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione normale di media μ incognita e di varianza σ^2 nota. Il test del rapporto di verosimiglianze di ampiezza α per verificare l'ipotesi nulla $H_0: \mu = \mu_0$ contro l'ipotesi alternativa $H_1: \mu \neq \mu_0$ è quello avente regione critica*

$$\mathcal{C} = \left\{ (x_1, x_2, \dots, x_n) : |\bar{x} - \mu_0| \geq z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\}. \quad (5.33)$$

Dim. Osserviamo anzitutto che la funzione di verosimiglianza del campione casuale è la stessa della Proposizione 5.3.1. Lo spazio dei parametri $\Theta \equiv \mathbb{R}$ si ripartisce ora negli insiemi $\Theta_0 = \{\mu_0\}$ e $\Theta_1 = (-\infty, \mu_0] \cup (\mu_0, \infty)$. Conseguentemente risulta:

$$\max_{\mu \in \{\mu_0\}} L(\mu) = L(\mu_0), \quad \max_{\mu \in \mathbb{R}} L(\mu) = L(\bar{x}),$$

così che il rapporto di verosimiglianze

$$\lambda = \frac{L(\mu_0)}{L(\bar{x})} = \exp \left[-\frac{n}{2\sigma^2} (\bar{x} - \mu_0)^2 \right],$$

coincide con quello ricavato nel corso della dimostrazione della Proposizione 5.3.1. Dalla (5.27) segue che ora la regione critica C è costituita dalle realizzazioni (x_1, x_2, \dots, x_n) tali da avversi $\lambda \leq k$, ossia:

$$(\bar{x} - \mu_0)^2 \geq -\frac{2\sigma^2}{n} \ln k.$$

Tale relazione, per l'arbitrarietà di k , equivale alla seguente:

$$|\bar{x} - \mu_0| \geq A,$$

dove la costante A va scelta in modo che la regione critica C abbia ampiezza α , ossia in modo che risulti

$$P(|\bar{X} - \mu_0| \geq A | \mu = \mu_0) = \alpha.$$

Poiché sotto l'ipotesi nulla H_0 la media campionaria \bar{X} ha distribuzione normale di media μ_0 e varianza σ^2/n , scegliamo A in modo da avversi

$$P \left(\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| \geq A \frac{\sqrt{n}}{\sigma} | \mu = \mu_0 \right) = P \left(|Z| \geq A \frac{\sqrt{n}}{\sigma} \right) = \alpha,$$

dove $Z \stackrel{\text{def}}{=} (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ è una variabile casuale normale standard. Ricordando che $z_{\alpha/2}$ è il reale tale che $P(|Z| \geq z_{\alpha/2}) = \alpha$, poniamo

$$A \frac{\sqrt{n}}{\sigma} = z_{\alpha/2},$$

così che

$$A = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Se ne trae che la regione critica C di ampiezza α del test del rapporto di verosimiglianze è data dalla (5.33). ■

Dalla Proposizione 5.3.2 segue dunque che l'ipotesi nulla $H_0: \mu = \mu_0$ viene rifiutata se la realizzazione osservata (x_1, x_2, \dots, x_n) è tale da avversi $|\bar{x} - \mu_0| \geq z_{\alpha/2} \sigma/\sqrt{n}$, mentre viene accettata se risulta $|\bar{x} - \mu_0| < z_{\alpha/2} \sigma/\sqrt{n}$.

Esempio 5.3.1 Riprendiamo in esame l'Esempio 1.3.1 nel quale si è fatto riferimento alla realizzazione

$$(220, 195, 235, 192, 215, 201, 241, 187, 197, 213, 232, 192, 211, 206)$$

di un campione casuale di taglia 14 che, come precisato nell'Esempio 1.4.2, si suppone estratto da una popolazione normale di media di media $\mu = 215$ e deviazione standard $\sigma = 20$. Si desidera ora verificare se è plausibile l'ipotesi che la media sia $\mu = 215$. Formuliamo a tal

fine l'ipotesi $H_0: \mu = 215$ contro l'ipotesi alternativa $H_1: \mu \neq 215$, da verificarsi mediante il test di cui alla Proposizione 5.3.2. La (5.33) fornisce la regione critica di ampiezza $\alpha = 0.05$:

$$C = \left\{ (x_1, x_2, \dots, x_{14}): |\bar{x} - 215| \geq z_{0.025} \frac{20}{\sqrt{14}} \right\}.$$

Dalla Tabella 2 dell'Appendice B si trae $z_{0.025} = 1.96$, così che risulta:

$$C = \{(x_1, x_2, \dots, x_{14}): |\bar{x} - 215| \geq 10.48\}.$$

Poiché la media campionaria assume valore $\bar{x} = 209.78$ in corrispondenza della realizzazione osservata, si ha $|\bar{x} - 215| = 5.22 < 10.48$ così che i valori osservati non appartengono alla regione critica; di conseguenza non si può rifiutare l'ipotesi nulla $H_0: \mu = 215$. ◆

Nelle proposizioni che seguono si fa riferimento ad altri due casi in cui si fa uso del metodo del rapporto di verosimiglianze. Si discutono, invero, dei test atti alla verifica di ipotesi concernenti media e varianza di popolazione normale quando entrambe sono incognite. Lo spazio dei parametri è ora, evidentemente, bidimensionale; il metodo del rapporto di verosimiglianze si applica tuttavia in modo analogo a quanto visto finora per il caso unidimensionale.

Proposizione 5.3.3 Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione normale di media μ e di varianza σ^2 entrambe incognite. Il test del rapporto di verosimiglianze di ampiezza α per verificare l'ipotesi nulla $H_0: \mu = \mu_0$ contro l'ipotesi alternativa $H_1: \mu \neq \mu_0$ è quello avente regione critica

$$C = \left\{ (x_1, x_2, \dots, x_n): \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| \geq t_{\alpha/2; n-1} \right\}. \quad (5.34)$$

Dim. Lo spazio dei parametri corrispondente alla coppia (μ, σ^2) è $\Theta \equiv \mathbb{R} \times \mathbb{R}^+$. Quest'ultimo è ripartito negli insiem $\Theta_0 = \{\mu_0\} \times \mathbb{R}^+$ e $\Theta_1 = (-\infty, \mu_0) \cup (\mu_0, \infty) \times \mathbb{R}^+$. Poiché la funzione di verosimiglianza è

$$L(\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right],$$

risulta:

$$\begin{aligned} \max_{\mu \in \{\mu_0\}, \sigma^2 \in \mathbb{R}^+} L(\mu, \sigma^2) &= L(\mu_0, \sigma_0^2) \\ &= \left(\frac{1}{2\pi\sigma_0^2} \right)^{n/2} \exp \left[-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu_0)^2 \right], \end{aligned}$$

dove

$$\sigma_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_0)^2$$

denota la stima di massima verosimiglianza della varianza σ^2 sotto l'ipotesi nulla H_0 . Si ha quindi:

$$L(\mu_0, \sigma_0^2) = \left(\frac{1}{2\pi\sigma_0^2} \right)^{n/2} e^{-n/2}.$$

Ricordando poi che gli stimatori di massima verosimiglianza di μ e σ^2 sono rispettivamente \bar{X} e $(n-1)S^2/n$ (cfr. Esempio 3.7.7), segue:

$$\begin{aligned} \max_{\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+} L(\mu, \sigma^2) &= L(\bar{x}, (n-1)s^2/n) \\ &= \left[\frac{n}{2\pi(n-1)s^2} \right]^{n/2} \exp \left[-\frac{n}{2(n-1)s^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right] \\ &= \left[\frac{n}{2\pi(n-1)s^2} \right]^{n/2} e^{-n/2}. \end{aligned}$$

Si ottiene dunque il rapporto di verosimiglianze:

$$\lambda = \frac{L(\mu_0, \sigma_0^2)}{L(\bar{x}, (n-1)s^2/n)} = \left[\frac{(n-1)s^2}{n\sigma_0^2} \right]^{n/2}. \quad (5.35)$$

Dimostriamo ora che risulta:

$$\frac{n\sigma_0^2}{(n-1)s^2} = 1 + \frac{t^2}{n-1}, \quad (5.36)$$

con

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}.$$

Ricordando la definizione di σ_0^2 , si ha infatti:

$$\begin{aligned} 1 + \frac{t^2}{n-1} &= 1 + \frac{n}{n-1} \frac{(\bar{x} - \mu_0)^2}{s^2} \\ &= \frac{(n-1)s^2 + n(\bar{x} - \mu_0)^2}{(n-1)s^2} \\ &= \frac{1}{(n-1)s^2} \left[\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2 \right]. \end{aligned} \quad (5.37)$$

Inoltre,

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2 &= \sum_{i=1}^n (x_i^2 - 2\bar{x}\mu_0 + \mu_0^2) \\ &= \sum_{i=1}^n (x_i^2 - 2x_i\mu_0 + \mu_0^2) \\ &= \sum_{i=1}^n (x_i - \mu_0)^2 = n\sigma_0^2. \end{aligned} \quad (5.38)$$

Dalle (5.37) e (5.38) segue la (5.36). Facendo uso di quest'ultima, si ricava che il rapporto di verosimiglianze (5.35) assume la forma

$$\lambda = \left(1 + \frac{t^2}{n-1} \right)^{-n/2},$$

così che la diseguaglianza $\lambda \leq k$ diventa:

$$t^2 \geq (n-1)(k^{-2/n} - 1).$$

Poiché k è una costante arbitraria, ricordando la (5.27) si ricava che la regione critica C è costituita dalle n -uple (x_1, x_2, \dots, x_n) per le quali risulta

$$|t| \equiv \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| \geq A,$$

con A scelto in modo che la regione critica C abbia ampiezza α , ossia tale da aversi

$$P \left(\left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| \geq A \mid \mu = \mu_0 \right) = \alpha.$$

Sotto l'ipotesi nulla H_0 , la statistica $T = (\bar{X} - \mu_0)/(S/\sqrt{n})$ ha distribuzione di Student con $n-1$ gradi di libertà. Si deve allora scegliere A in maniera che risulti

$$P(|T| \geq A) = \alpha.$$

Ricordando che $t_{\alpha/2;n-1}$ è il reale tale che $P(T \geq t_{\alpha/2;n-1}) = \alpha$ (cfr. § 2.2), possiamo porre

$$A = t_{\alpha/2;n-1}.$$

Ne discende che la regione critica C di ampiezza α del test del rapporto di verosimiglianze è data dalla (5.34). ■

Esempio 5.3.2 Un'azienda chimica produce un nuovo tipo di vernice che si desidera sottoporre a controllo. Questa viene usata per verniciare 10 pannelli di uguale dimensione. Si misura, in secondi, il tempo che ogni pannello impiega ad asciugarsi. Si costruisce così un campione casuale $(X_1, X_2, \dots, X_{10})$ del quale si osserva la realizzazione seguente:

$$(238, 152, 215, 198, 167, 228, 172, 182, 148, 218).$$

Supponendo che il tempo di asciugatura possa essere riguardato come una variabile casuale normale di media μ e varianza σ^2 entrambe incognite, si desidera sottoporre a verifica l'ipotesi $H_0: \mu = 200$ contro l'ipotesi alternativa $H_1: \mu \neq 200$. Facendo appello alla Proposizione 5.3.3, introduciamo la regione critica di ampiezza $\alpha = 0.05$ del test del rapporto di verosimiglianze:

$$C = \left\{ (x_1, x_2, \dots, x_{10}) : \left| \frac{\bar{x} - 200}{S/\sqrt{10}} \right| \geq 2.262 \right\},$$

dove si è fatto uso della relazione $t_{0.025;9} = 2.262$ (cfr. Tabella 4 dell'Appendice B). Dalla realizzazione osservata si trae:

$$\bar{x} = 191.8 \quad s = 32.169,$$

così che risulta:

$$\left| \frac{\bar{x} - 200}{s/\sqrt{10}} \right| = \left| \frac{191.8 - 200}{32.169/\sqrt{10}} \right| = 0.8 < 2.262.$$

Poiché la realizzazione osservata non appartiene alla regione critica, non si può rifiutare l'ipotesi nulla $H_0: \mu = 200$. \diamond

Proposizione 5.3.4 Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione normale di media μ e di varianza σ^2 entrambe incognite. Il test del rapporto di verosimiglianze di ampiezza α per verificare l'ipotesi nulla $H_0: \sigma^2 = \sigma_0^2$ contro l'ipotesi alternativa $H_1: \sigma^2 \neq \sigma_0^2$ è quello di regione critica

$$C = \left\{ (x_1, x_2, \dots, x_n) : 0 \leq \frac{(n-1)s^2}{\sigma_0^2} \leq \chi_{1-\alpha/2; n-1}^2 \text{ oppure } \frac{(n-1)s^2}{\sigma_0^2} \geq \chi_{\alpha/2; n-1}^2 \right\}.$$

Dim. $\Theta = \mathbb{R} \times \mathbb{R}^+$ è lo spazio dei parametri corrispondente alla coppia (μ, σ^2) . Tale spazio è ripartito negli insiemi $\Theta_0 = \mathbb{R} \times \{\sigma_0^2\}$ e $\Theta_1 = \mathbb{R} \times (-\infty, \sigma_0^2) \cup (\sigma_0^2, \infty)$. La funzione di verosimiglianza del campione casuale è la seguente:

$$L(\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right].$$

Poiché il metodo della massima verosimiglianza fornisce \bar{X} e $(n-1)S^2/n$ come stimatori di μ e σ^2 (cfr. Esempio 3.7.7), si ha:

$$\begin{aligned} \max_{\mu \in \mathbb{R}, \sigma^2 \in \{\sigma_0^2\}} L(\mu, \sigma^2) &= L(\bar{x}, \sigma_0^2) \\ &= \left(\frac{1}{2\pi\sigma_0^2} \right)^{n/2} \exp \left[-\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right] \\ &= \left(\frac{1}{2\pi\sigma_0^2} \right)^{n/2} \exp \left[-\frac{(n-1)s^2}{2\sigma_0^2} \right] \\ \max_{\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+} L(\mu, \sigma^2) &= L\left(\bar{x}, (n-1)\frac{s^2}{n}\right) \\ &= \left[\frac{n}{2\pi(n-1)s^2} \right]^{n/2} \exp \left[-\frac{n}{2(n-1)s^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right] \\ &= \left[\frac{n}{2\pi(n-1)s^2} \right]^{n/2} \exp\left(-\frac{n}{2}\right). \end{aligned}$$

5.3. RAPPORTO DI VERO SIMIGLIANZE

Il rapporto di verosimiglianze λ è dunque il seguente:

$$\begin{aligned} \lambda &= \frac{L(\bar{x}, \sigma_0^2)}{L(\bar{x}, (n-1)s^2/n)} \\ &= \left[\frac{(n-1)s^2}{n\sigma_0^2} \right]^{n/2} \exp \left\{ -\frac{n}{2} \left[\frac{(n-1)s^2}{n\sigma_0^2} - 1 \right] \right\}. \end{aligned}$$

La diseguaglianza $\lambda \leq k$ quindi diventa:

$$\left[\frac{(n-1)s^2}{n\sigma_0^2} \right]^{n/2} \exp \left\{ -\frac{n}{2} \left[\frac{(n-1)s^2}{n\sigma_0^2} - 1 \right] \right\} \leq k;$$

di qui, per l'arbitrarietà di k , si trae:

$$\frac{(n-1)s^2}{\sigma_0^2} B^{(n-1)s^2/\sigma_0^2} \leq A, \quad (5.39)$$

dove $A \stackrel{\text{def}}{=} n e^{-1} k^{2/n}$ è una costante positiva arbitraria e dove $B \stackrel{\text{def}}{=} e^{-1/n}$ è evidentemente tale da aversi $0 < B < 1$. Se A è sufficientemente prossimo a zero, la diseguaglianza (5.39) è verificata per

$$0 \leq \frac{(n-1)s^2}{\sigma_0^2} \leq A_1, \quad \frac{(n-1)s^2}{\sigma_0^2} \geq A_2,$$

con A_1 e A_2 ($A_1 < A_2$) radici dell'equazione

$$\frac{(n-1)s^2}{\sigma_0^2} B^{(n-1)s^2/\sigma_0^2} = A.$$

La regione critica C del test del rapporto di verosimiglianze assume dunque la seguente forma:

$$C = \left\{ (x_1, x_2, \dots, x_n) : 0 \leq \frac{(n-1)s^2}{\sigma_0^2} \leq A_1 \text{ o } \frac{(n-1)s^2}{\sigma_0^2} \geq A_2 \right\},$$

dove le costanti A_1 e A_2 vengono determinate in modo che C abbia ampiezza α . Sotto l'ipotesi nulla $H_0: \sigma^2 = \sigma_0^2$ deve quindi risultare:

$$P \left[0 \leq \frac{(n-1)s^2}{\sigma_0^2} \leq A_1 \mid \sigma^2 = \sigma_0^2 \right] + P \left[\frac{(n-1)s^2}{\sigma_0^2} \geq A_2 \mid \sigma^2 = \sigma_0^2 \right] = \alpha. \quad (5.40)$$

Poiché sotto l'ipotesi nulla la statistica $(n-1)s^2/\sigma_0^2$ ha distribuzione chi-quadrato con $n-1$ gradi di libertà, scegliamo

$$A_1 = \chi_{1-\alpha/2; n-1}^2, \quad A_2 = \chi_{\alpha/2; n-1}^2.$$

In tal caso, infatti, l'uguaglianza (5.40) è verificata. Si ricava così che la regione critica C di ampiezza α del test del rapporto di verosimiglianze è costituita dalle n -uple (x_1, x_2, \dots, x_n) tali da aversi

$$0 \leq \frac{(n-1)s^2}{\sigma_0^2} \leq \chi_{1-\alpha/2; n-1}^2 \quad \text{oppure} \quad \frac{(n-1)s^2}{\sigma_0^2} \geq \chi_{\alpha/2; n-1}^2.$$

Esempio 5.3.3 Riprendiamo in esame l'Esempio 5.3.2. Si ha dunque un campione casuale $(X_1, X_2, \dots, X_{10})$ che si suppone estratto da una popolazione normale di media μ e varianza σ^2 entrambe incognite. Sulla base della realizzazione osservata

$$(238, 152, 215, 198, 167, 228, 172, 182, 148, 218),$$

si desidera sottoporre a verifica l'ipotesi $H_0: \sigma^2 = 800$ contro l'ipotesi alternativa $H_1: \sigma^2 \neq 800$. Dalla Proposizione 5.3.4 segue che la regione critica di ampiezza $\alpha = 0.05$ del test del rapporto di verosimiglianze è la seguente:

$$C = \left\{ (x_1, x_2, \dots, x_{10}): 0 \leq \frac{9s^2}{800} \leq 2.7 \text{ oppure } \frac{9s^2}{800} \geq 19.023 \right\},$$

avendo sfruttato le relazioni $\chi^2_{0.975,9} = 2.7$ e $\chi^2_{0.025,9} = 19.023$ (cfr. Tabella 3 dell'Appendice B). Notiamo che dalla realizzazione osservata risulta $s = 32.169$, e quindi

$$\frac{9s^2}{800} = 11.642 \in (2.7, 19.023).$$

Poiché la realizzazione osservata non appartiene alla regione critica C , non si può rifiutare l'ipotesi nulla $H_0: \sigma^2 = 800$. \diamond

A conclusione di questo paragrafo esaminiamo l'uso del metodo del rapporto di verosimiglianze per verificare ipotesi concernenti differenze tra medie di due popolazioni. In analogia con quanto visto nel § 4.4, in cui si è effettuata la stima intervallare di differenze tra medie, tratteremo qui i casi di varianze note e di varianze incognite.

Proposizione 5.3.5 *Dati due campioni indipendenti $(X_{11}, X_{12}, \dots, X_{1n})$ e $(X_{21}, X_{22}, \dots, X_{2m})$ estratti da popolazioni normali di medie μ_1 e μ_2 incognite e di varianze σ_1^2 e σ_2^2 note, e denotato con ω un fissato reale, il test del rapporto di verosimiglianze di ampiezza α per verificare l'ipotesi nulla $H_0: \mu_1 - \mu_2 = \omega$ contro l'ipotesi alternativa $H_1: \mu_1 - \mu_2 \neq \omega$ è quello avente regione critica*

$$C = \left\{ (x_{11}, \dots, x_{1n}), (x_{21}, \dots, x_{2m}): \frac{|\bar{x}_1 - \bar{x}_2 - \omega|}{\delta} \geq z_{\alpha/2} \right\}, \quad (5.41)$$

dove \bar{x}_1 e \bar{x}_2 denotano i valori assunti dalle medie campionarie delle popolazioni, e dove si è posto

$$\delta = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}. \quad (5.42)$$

Dim. Per brevità accenneremo soltanto alla dimostrazione. Applicando il metodo del rapporto di verosimiglianze, la diseguaglianza $\lambda \leq k$ di cui alla (5.27) si esprime come segue:

$$\frac{|\bar{x}_1 - \bar{x}_2 - \omega|}{\delta} \geq A. \quad (5.43)$$

Sceglieremo la costante A in modo che la regione critica C abbia ampiezza α , ossia in modo da aversi

$$P\left(\frac{|\bar{X}_1 - \bar{X}_2 - \omega|}{\delta} \geq A \mid \mu_1 - \mu_2 = \omega\right) = \alpha.$$

Poiché sotto l'ipotesi nulla $H_0: \mu_1 - \mu_2 = \omega$ la differenza $\bar{X}_1 - \bar{X}_2$ ha distribuzione normale di media ω e varianza δ^2 (cfr. Teorema 1.4.2), la variabile casuale $Z \stackrel{\text{def}}{=} (\bar{X}_1 - \bar{X}_2 - \omega)/\delta$ è normale standard. Poniamo allora $A = z_{\alpha/2}$, così che risulta

$$P\left(\frac{|\bar{X}_1 - \bar{X}_2 - \omega|}{\delta} \geq z_{\alpha/2} \mid \mu_1 - \mu_2 = \omega\right) = P(|Z| \geq z_{\alpha/2}) = \alpha.$$

Se ne trae che la (5.41) costituisce la regione critica di ampiezza α del test del rapporto di verosimiglianze. \blacksquare

Esempio 5.3.4 I numeri di impianti termici prodotti giornalmente da due stabilimenti sono caratterizzati rispettivamente da varianze $\sigma_1^2 = 9$ e $\sigma_2^2 = 10$. Sulla base dei numeri medi osservati $\bar{x}_1 = 85$ e $\bar{x}_2 = 76$ di impianti prodotti rispettivamente nell'arco di $n = 40$ e $m = 55$ giornate lavorative, si desidera verificare se la differenza tra i numeri medi μ_1 e μ_2 di impianti prodotti quotidianamente dai due stabilimenti sia pari ad 8. Si desidera quindi sottoporre a verifica l'ipotesi nulla $H_0: \mu_1 - \mu_2 = 8$ contro l'ipotesi alternativa $H_1: \mu_1 - \mu_2 \neq 8$. Poiché n ed m sono sufficientemente grandi, è possibile sfruttare l'approssimazione normale fornita dal teorema centrale del limite. Facciamo allora riferimento alla regione critica (5.41) di ampiezza $\alpha = 0.05$ del test del rapporto di verosimiglianze ricavata nella Proposizione 5.3.5. Poiché la (5.42) fornisce

$$\delta = \sqrt{\frac{9}{40} + \frac{10}{55}} = 0.6378,$$

il primo membro della (5.43),

$$\frac{|\bar{x}_1 - \bar{x}_2 - \omega|}{\delta} = \frac{|85 - 76 - 8|}{0.6378} = 1.57,$$

è minore di $z_{\alpha/2} = z_{0.025} = 1.96$ (cfr. Tabella 2 dell'Appendice B), così che le realizzazioni osservate non appartengono alla regione critica. La differenza tra i valori medi osservati non è quindi tale da escludere che $\mu_1 - \mu_2$ sia pari ad 8. \diamond

Nel caso in cui le varianze delle due popolazioni sono incognite non è possibile fare ricorso alla Proposizione 5.3.5; si adotta invece il criterio di cui alla proposizione seguente, la cui dimostrazione sarà per brevità solo accennata.

Proposizione 5.3.6 *Dati due campioni indipendenti $(X_{11}, X_{12}, \dots, X_{1n})$ e $(X_{21}, X_{22}, \dots, X_{2m})$ estratti da popolazioni normali di medie μ_1 e μ_2 incognite e di varianze incognite, e denotato con ω un fissato reale, il test del rapporto di verosimiglianze di ampiezza α per verificare l'ipotesi nulla $H_0: \mu_1 - \mu_2 = \omega$ contro l'ipotesi alternativa $H_1: \mu_1 - \mu_2 \neq \omega$ è quello avente regione critica*

$$C = \left\{ (x_{11}, \dots, x_{1n}), (x_{21}, \dots, x_{2m}): \frac{|\bar{x}_1 - \bar{x}_2 - \omega|}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \geq t_{\alpha/2; n+m-2} \right\}, \quad (5.44)$$

con \bar{x}_1 e \bar{x}_2 valori assunti dalle medie campionarie delle popolazioni e dove si è posto

$$s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}, \quad (5.45)$$

con s_1^2 e s_2^2 valori assunti dalle varianze campionarie delle popolazioni.

Dim. Facendo uso del metodo del rapporto di verosimiglianze, la diseguaglianza $\lambda \leq k$ è così espressa:

$$\frac{|\bar{x}_1 - \bar{x}_2 - \omega|}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \geq A. \quad (5.46)$$

Si determina ora la costante A in modo che la regione critica C abbia ampiezza α , ossia in modo che risulti

$$P\left(\frac{|\bar{X}_1 - \bar{X}_2 - \omega|}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \geq A \mid \mu_1 - \mu_2 = \omega\right) = \alpha.$$

Tale uguaglianza è soddisfatta se si pone $A = t_{\alpha/2; n+m-2}$. Infatti, come abbiamo visto nel corso della dimostrazione del Teorema 4.4.2, la variabile casuale

$$T \stackrel{\text{def}}{=} \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$$

ha distribuzione di Student con $n+m-2$ gradi di libertà, così che

$$P\left(\frac{|\bar{X}_1 - \bar{X}_2 - \omega|}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \geq t_{\alpha/2; n+m-2} \mid \mu_1 - \mu_2 = \omega\right) = P(|T| \geq t_{\alpha/2; n+m-2}) = \alpha.$$

La (5.44) costituisce pertanto la regione critica di ampiezza α del test del rapporto di verosimiglianze. ■

Esempio 5.3.5 Ad una prova di dattilografia due candidati completano rispettivamente $n = 15$ e $m = 10$ pagine. I numeri medi di errori commessi per pagina sono rispettivamente $\bar{x}_1 = 3.2$ e $\bar{x}_2 = 4.7$, con varianze osservate $s_1^2 = 2.1$ e $s_2^2 = 2.5$. Si desidera verificare se i numeri medi μ_1 e μ_2 di errori per pagina siano uguali. Pertanto, assumendo che il numero di errori per pagina sia distribuito secondo variabili casuali normali di medie μ_1 e μ_2 , si sottopone a verifica l'ipotesi nulla $H_0: \mu_1 - \mu_2 = 0$ contro l'ipotesi alternativa $H_1: \mu_1 - \mu_2 \neq 0$. Utilizziamo la regione critica (5.44) di ampiezza $\alpha = 0.05$ del test del rapporto di verosimiglianze fornita dalla Proposizione 5.3.6. Dalla (5.45) si ha

$$s_p^2 = \frac{(15-1)2.1 + (10-1)2.5}{15+10-2} = 2.2565,$$

così che il primo membro della (5.46) assume il valore

$$\frac{|3.2 - 4.7|}{\sqrt{2.2565} \sqrt{\frac{1}{15} + \frac{1}{10}}} = 2.446.$$

5.4. TEST CHI-QUADRATO

Questo è maggiore di $t_{\alpha/2; n+m-2} = t_{0.025; 23} = 2.069$ (cfr. Tabella 4 dell'Appendice B), e quindi le realizzazioni osservate appartengono alla regione critica. L'ipotesi che i numeri medi di errori commessi dai due impiegati coincidano va dunque rifiutata.

Se invece si sottopone a verifica l'ipotesi nulla $H'_0: \mu_1 - \mu_2 = -1$ contro l'ipotesi alternativa $H'_1: \mu_1 - \mu_2 \neq -1$, si ha

$$\frac{|\bar{x}_1 - \bar{x}_2 - \omega|}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{|3.2 - 4.7 + 1|}{\sqrt{2.2565} \sqrt{\frac{1}{15} + \frac{1}{10}}} = 0.815,$$

che risulta minore di $t_{\alpha/2; n+m-2} = t_{0.025; 23} = 2.069$. Le realizzazioni osservate appartengono quindi alla regione critica (5.44) di ampiezza $\alpha = 0.05$ e pertanto non va rifiutata l'ipotesi che il primo candidato commetta un numero medio di errori inferiore di un'unità a quello commesso dall'altro candidato. ◆

5.4 Test chi-quadrato

In numerose applicazioni statistiche si presentano situazioni in cui si desidera stabilire se un insieme di dati possa essere riguardato come realizzazione di un campione casuale estratto da una popolazione avente una preassegnata funzione di distribuzione. Analizziamo il caso di una popolazione che è suddivisa in k classi, nel senso che i suoi possibili valori appartengono ad un insieme che è ripartito nelle k classi C_1, C_2, \dots, C_k , dove ciascuna di queste rappresenta una o più categorie di una caratteristica qualitativa o quantitativa della popolazione. Indichiamo poi con X la variabile casuale genitrice e con p_i la probabilità della classe C_i ($i = 1, 2, \dots, k$). Se la popolazione è discreta, si ha:

$$p_i = \sum_{\{x_i \in C_i\}} P(X = x_i) \quad (i = 1, 2, \dots, k).$$

Se, invece, X è continua, denotata con $f_X(x)$ la sua densità di probabilità, risulta:

$$p_i = \int_{C_i} f_X(x) dx \quad (i = 1, 2, \dots, k).$$

Se, quindi, si sceglie a caso un elemento della popolazione, questo appartiene alla classe C_i con probabilità p_i , qualunque sia la natura, discreta o continua, della popolazione. Notiamo che, essendo stato supposto finito il numero delle classi, quando — come accade in numerose applicazioni concrete — la variabile genitrice assume valori in un insieme infinito, alcune delle classi C_i possono avere ampiezza infinita. Se, ad esempio, si considera una popolazione normale, questa può essere suddivisa in k classi fissando $k-1$ numeri reali $t_1 < t_2 < \dots < t_{k-1}$ e ponendo:

$$C_1 = (-\infty, t_1], \quad C_i = (t_{i-1}, t_i] \quad (i = 2, 3, \dots, k-1), \quad C_k = (t_{k-1}, \infty).$$

In generale la distribuzione di probabilità (p_1, p_2, \dots, p_k) è incognita, così che un problema fondamentale consiste nell'assegnare una k -upla di reali $(\pi_1, \pi_2, \dots, \pi_k)$, ciascuno appartenente all'intervallo $[0, 1]$, e nel verificare se questi sono interpretabili come probabilità

delle k classi della popolazione. Fissata la k -upla $(\pi_1, \pi_2, \dots, \pi_k)$ si desidera quindi verificare l'ipotesi nulla

$$H_0: \{p_i\} = \{\pi_i\} \quad (5.47)$$

contro l'ipotesi alternativa

$$H_1: \{p_i\} \neq \{\pi_i\}.$$

Si effettua così un *test di conformità*, nel senso che mediante esso si verifica se i valori assegnati $(\pi_1, \pi_2, \dots, \pi_k)$ sono conformi alle probabilità teoriche (p_1, p_2, \dots, p_k) delle k classi nel senso che risulta $p_i = \pi_i$ ($i = 1, 2, \dots, k$).

Si noti che la (5.47) è un'ipotesi semplice dal momento che i valori π_i ipotizzati specificano completamente la distribuzione di probabilità delle classi della popolazione.

Dato un campione casuale (X_1, X_2, \dots, X_n) estratto dalla popolazione in questione, vediamo ora come si può procedere alla costruzione della regione critica del test. Definiamo a tal fine le variabili casuali N_1, N_2, \dots, N_k , dette *frequenze assolute* delle classi C_i , che rappresentano i numeri di elementi del campione appartenenti rispettivamente alle classi C_1, C_2, \dots, C_k . Notiamo anzitutto che tali variabili sono legate dalla relazione

$$N_1 + N_2 + \dots + N_k = n, \quad (5.48)$$

e che il vettore casuale (N_1, N_2, \dots, N_k) ha distribuzione multinomiale. Invero, sotto l'ipotesi nulla (5.47), si ha:²

$$P(N_1 = n_1, N_2 = n_2, \dots, N_k = n_k | p_i = \pi_i) = \frac{n!}{n_1! n_2! \dots n_k!} \pi_1^{n_1} \pi_2^{n_2} \dots \pi_k^{n_k},$$

con $n_1, n_2, \dots, n_k \geq 0$ e $n_1 + n_2 + \dots + n_k = n$. Osserviamo ora che le variabili casuali N_i possono esprimersi al seguente modo:

$$N_i = \left| \{X_j : X_j \in C_i; j = 1, 2, \dots, n\} \right| = \sum_{j=1}^n I_{C_i}(X_j) \quad (i = 1, 2, \dots, k) \quad (5.49)$$

dove I denota la funzione indicatrice definita nella (3.75). Poiché la variabile $I_{C_i}(X_j)$ per $j = 1, 2, \dots, n$ ha distribuzione di Bernoulli di parametro p_i , dalla (5.49) discende che N_i ha distribuzione binomiale di parametri n e p_i . Sotto l'ipotesi nulla (5.47), per $i = 1, 2, \dots, k$ la distribuzione di probabilità, la media e la varianza di N_i sono pertanto le seguenti:

$$\begin{aligned} P(N_i = x | p_i = \pi_i) &= \binom{n}{x} \pi_i^x (1 - \pi_i)^{n-x} \quad (x = 0, 1, \dots, n), \\ E(N_i | p_i = \pi_i) &= n \pi_i, \quad D^2(N_i | p_i = \pi_i) = n \pi_i (1 - \pi_i). \end{aligned} \quad (5.50)$$

Quindi, quando l'ipotesi nulla è vera in media $n \pi_i$ elementi di una realizzazione del campione casuale considerato appartengono alla classe C_i .

²Per semplicità di scrittura nelle formule che seguono con $p_i = \pi_i$ si intende che tale uguaglianza sussiste per $i = 1, 2, \dots, k$.

Allo scopo di costruire una regione di rifiuto per l'ipotesi nulla, introduciamo la statistica

$$\Xi = \sum_{i=1}^k \frac{(N_i - n \pi_i)^2}{n \pi_i}; \quad (5.51)$$

Proposizione 5.4.1 *Sotto la condizione che l'ipotesi nulla $H_0: \{p_i\} = \{\pi_i\}$ sia vera, al divergere di n la statistica (5.51) assume distribuzione chi-quadrato con $k-1$ gradi di libertà.*

Dim. Ci limiteremo qui a dimostrare la proposizione nel caso di $k = 2$ classi poiché se k è un intero positivo arbitrario la dimostrazione, dovuta a E.S. Pearson, è eccessivamente laboriosa. Per $k = 2$, si ha $N_1 + N_2 = n$ e $\pi_2 = 1 - \pi_1$ così che la (5.51) diventa

$$\Xi = \sum_{i=1}^2 \frac{(N_i - n \pi_i)^2}{n \pi_i} = \frac{(N_1 - n \pi_1)^2}{n \pi_1 (1 - \pi_1)} = \left[\frac{N_1 - E(N_1 | p_i = \pi_i)}{D(N_1 | p_i = \pi_i)} \right]^2, \quad (5.52)$$

dove l'ultima uguaglianza segue dalle (5.50). Poiché la variabile binomiale standardizzata $[N_1 - E(N_1 | p_i = \pi_i)]/D(N_1 | p_i = \pi_i)$ per $n \rightarrow \infty$ tende ad una variabile normale standard, per il Teorema 2.1.1 il suo quadrato (5.52) tende ad una variabile chi-quadrato con 1 grado di libertà. ■

Per fornire una sia pur intuitiva giustificazione dell'affermazione generale espressa dalla Proposizione 5.4.1, osserviamo che se si pone

$$Y_i = \frac{N_i - E(N_i | p_i = \pi_i)}{D(N_i | p_i = \pi_i)} \quad (i = 1, 2, \dots, k), \quad (5.53)$$

dalle (5.50) e (5.51) si ricava:

$$\Xi = \sum_{i=1}^k \frac{(N_i - n \pi_i)^2}{n \pi_i} \approx \sum_{i=1}^k Y_i^2,$$

dove l'approssimazione nasce dall'avere supposto che tutti i valori π_i siano molto minori dell'unità. In tal caso, infatti, dalla seconda delle (5.50) segue $D^2(N_i | p_i = \pi_i) \approx n \pi_i$. Per n grande, inoltre, le variabili Y_i hanno approssimativamente distribuzione normale standard, e quindi le variabili Y_i^2 hanno approssimativamente distribuzione chi-quadrato con 1 grado di libertà. La statistica Ξ è quindi approssimativamente la somma di k variabili chi-quadrato che, però, non sono indipendenti; invero, in virtù delle (5.48) e (5.53) tra esse sussiste la seguente relazione lineare:

$$\sum_{i=1}^k D(N_i | p_i = \pi_i) Y_i = \sum_{i=1}^k [N_i - E(N_i | p_i = \pi_i)] = \sum_{i=1}^k N_i - n \sum_{i=1}^k \pi_i = 0.$$

Ciascuna delle k variabili Y_i è pertanto dipendente dalle rimanenti $k-1$ variabili; ciò comporta che per n grande la statistica Ξ ha approssimativamente distribuzione chi-quadrato con $k-1$ gradi di libertà. Va comunque ribadito che il procedimento appena delineato non è rigoroso poiché poggia sull'approssimazione $D^2(N_i | p_i = \pi_i) \approx n \pi_i$, non sempre legittima.

In virtù della Proposizione 5.4.1 si può dunque affermare che se la taglia n del campione è elevata, la distribuzione della statistica Ξ è approssimabile con quella di una variabile chi-quadrato con $k - 1$ gradi di libertà. Nei casi concreti l'approssimazione viene considerata soddisfacente se risulta $n\pi_i \geq 5$ per $i = 1, 2, \dots, k$. Una conseguenza immediata di tale approssimazione è che è possibile costruire una regione critica per verificare l'ipotesi nulla (5.47). Infatti, dalla Proposizione 5.4.1 per n elevato segue:

$$P(\Xi \geq \chi^2_{\alpha;k-1} | p_i = \pi_i) \approx 1 - \alpha, \quad (5.54)$$

dove l'approssimazione migliora al crescere di n . Per costruire la regione critica del test denotiamo allora con (x_1, x_2, \dots, x_n) la realizzazione osservata di un campione casuale estratto dalla popolazione considerata, e indichiamo con

$$n_i = \left| \{x_j : x_j \in C_i; j = 1, 2, \dots, n\} \right| = \sum_{j=1}^n I_{C_i}(x_j)$$

la frequenza assoluta della classe C_i , ossia il numero di elementi della realizzazione appartenenti a C_i ($i = 1, 2, \dots, k$). Indichiamo inoltre con

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n\pi_i)^2}{n\pi_i} \quad (5.55)$$

il valore assunto dalla statistica (5.51) in corrispondenza della realizzazione (x_1, x_2, \dots, x_n) . Osserviamo che la (5.55) fornisce una qualche misura del discostamento tra le frequenze assolute osservate n_i e le corrispondenti grandezze teoriche $n\pi_i$: pertanto, quanto maggiore è il valore χ^2 tanto più ci si attende che l'ipotesi nulla sia falsa. Inoltre, la circostanza che nella (5.55) i termini $(n_i - n\pi_i)^2$ sono divisi per $n\pi_i$ sta ad indicare che il quadrato del discostamento tra la frequenza assoluta n_i e il corrispondente valore ipotizzato $n\pi_i$ ha maggiore peso quando la frequenza attesa $n\pi_i$ è piccola. In virtù della (5.54) e della definizione (5.55) si trae che, per la verifica dell'ipotesi nulla, l'insieme

$$\mathcal{C} = \{(x_1, x_2, \dots, x_n) : \chi^2 \geq \chi^2_{\alpha;k-1}\} \quad (5.56)$$

costituisce una regione critica che ha approssimativamente ampiezza α , dove l'approssimazione migliora al crescere della taglia n del campione. Pertanto, adottando la (5.56) come regione critica del test, l'ipotesi nulla (5.47) viene rifiutata se la realizzazione (x_1, x_2, \dots, x_n) osservata appartiene a \mathcal{C} ; in caso contrario essa viene accettata.

Si noti che la condizione $\chi^2 \geq \chi^2_{\alpha;k-1}$ che compare nella definizione (5.56) della regione critica concorda pienamente con la considerazione, effettuata in precedenza, in base alla quale ci si attende che per valori grandi di χ^2 l'ipotesi nulla sia da rifiutare.

Quello ora descritto prende il nome di *test chi-quadrato* per via della approssimazione della statistica Ξ con una variabile chi-quadrato.

È importante sottolineare che la scelta delle classi gioca un ruolo fondamentale nell'applicazione del test chi-quadrato. Numero e ampiezza delle classi possono infatti essere determinanti per l'esito dell'analisi statistica: tipi differenti di classificazione dei medesimi dati possono invero condurre a conclusioni opposte circa la validità dell'ipotesi nulla. Sebbene

5.4. TEST CHI-QUADRATO

non sia in generale possibile fornire indicazioni rigorose e sempre applicabili per la costruzione delle classi, un utile criterio empirico consiste nello scegliere le classi C_i in modo che ciascuna frequenza attesa $n\pi_i$ superi una certa prefissata soglia. Abbiamo ad esempio menzionato che nei casi pratici si ha una buona approssimazione della distribuzione di Ξ con una distribuzione chi-quadrato quando n è grande e quando risulta $n\pi_i \geq 5$ per $i = 1, 2, \dots, n$.

Taluni autori suggeriscono di costruire le classi C_i facendo sì, quando possibile, che queste siano equiampie, ossia costituite da intervalli di uguali ampiezze. Esamineremo ora un esempio di utilizzazione del test chi-quadrato in cui si fa uso proprio di quest'ultimo criterio.

Esempio 5.4.1 Un'agenzia incaricata di compiere un sondaggio intervista 80 telespettatori registrando per quanto tempo essi hanno seguito una certa trasmissione televisiva della durata di 4 ore. Suddividendo la durata della trasmissione in 4 fasce orarie, si costruiscono le seguenti classi:

$$C_1 = [0, 1], \quad C_2 = (1, 2], \quad C_3 = (2, 3], \quad C_4 = (3, 4].$$

Sulla base dei dati $(x_1, x_2, \dots, x_{80})$ raccolti risulta che le frequenze assolute delle 4 classi C_i sono le seguenti:

$$n_1 = 38, \quad n_2 = 23, \quad n_3 = 11, \quad n_4 = 8.$$

Si desidera verificare se le probabilità p_i che un generico telespettatore ha seguito la trasmissione per un periodo di tempo appartenente alla classe C_i sia del tipo

$$\pi_i = \frac{16}{15} \frac{1}{2^i} \equiv \frac{2^{4-i}}{15} \quad (i = 1, 2, 3, 4).$$

Formuliamo a tal fine le ipotesi $H_0: \{p_i\} = \{\pi_i\}$ e $H_1: \{p_i\} \neq \{\pi_i\}$, che andremo ora a verificare mediante il test chi-quadrato. In accordo con la (5.56) costruiamo una regione critica che abbia approssimativamente ampiezza $\alpha = 0.05$. Poiché dalla Tabella 3 dell'Appendice B si ricava $\chi^2_{\alpha;k-1} = \chi^2_{0.05;3} = 7.815$, si ha:

$$\mathcal{C} = \{(x_1, x_2, \dots, x_{80}) : \chi^2 \geq 7.815\}.$$

Notiamo che risulta

$$n\pi_1 = 42.\bar{6}, \quad n\pi_2 = 21.\bar{3}, \quad n\pi_3 = 10.\bar{6}, \quad n\pi_4 = 5.\bar{3};$$

pertanto, essendo $n\pi_i \geq 5$ per $i = 1, 2, 3, 4$, le classi C_i e i valori π_i soddisfano la condizione che il numero atteso di elementi appartenenti ad ogni classe sia non inferiore a 5. Inoltre, $n = 80$ è sufficientemente elevato da poter ritenere che l'approssimazione che conduce al test chi-quadrato sia soddisfacente. Dalla (5.55) si ricava infine:

$$\begin{aligned} \chi^2 &= \sum_{i=1}^k \frac{(n_i - n\pi_i)^2}{n\pi_i} \\ &= \frac{(38 - 42.\bar{6})^2}{42.\bar{6}} + \frac{(23 - 21.\bar{3})^2}{21.\bar{3}} + \frac{(11 - 10.\bar{6})^2}{10.\bar{6}} + \frac{(8 - 5.\bar{3})^2}{5.\bar{3}} = 1.9843. \end{aligned}$$

Il valore ottenuto $\chi^2 = 1.9843$ è minore di 7.815, così che la realizzazione osservata non appartiene alla regione critica \mathcal{C} . Si accetta dunque l'ipotesi nulla $H_0: \{p_i\} = \{2^{4-i}/15\}$. ♦

Tabella 5.2: Matrimoni celebrati in ciascun mese nell'anno 1985.

i	mesi	matrimoni (n_i)
1	gennaio	11420
2	febbraio	10241
3	marzo	11411
4	aprile	29149
5	maggio	24138
6	giugno	40768
7	luglio	32769
8	agosto	29813
9	settembre	49322
10	ottobre	30910
11	novembre	6919
12	dicembre	19130

Sovente nell'applicazione del test chi-quadrato le classi C_i vengono scelte equiprobabili, ossia tali che la distribuzione di probabilità $\{\pi_i\}$ è uniforme. In questo caso la determinazione della regione critica (5.56) si semplifica. Infatti, se $\pi_i = 1/k$ per $i = 1, 2, \dots, k$, per la condizione $n_1 + n_2 + \dots + n_k = n$ la (5.55) diventa:

$$\chi^2 = \frac{k}{n} \left(\sum_{i=1}^k n_i^2 \right) - n. \quad (5.57)$$

Notiamo, inoltre, che se per $i = 1, 2, \dots, k$ risulta $\pi_i = 1/k$, la relazione $n \pi_i \geq 5$ è soddisfatta quando il numero k di classi è non maggiore di $n/5$.

Nell'esempio che segue si utilizza il test chi-quadrato per verificare un'ipotesi nel caso di classi equiprobabili.

Esempio 5.4.2 Nella Tabella 5.2 è riportato il numero di matrimoni celebrati in Italia in ciascun mese dell'anno 1985. Vogliamo verificare, utilizzando il test chi-quadrato, se è plausibile ipotizzare — come peraltro certo non indicano, ad un sia pur qualitativo esame, i dati della tabella — se la data dei matrimoni possa considerarsi distribuita uniformemente nei 12 mesi. A tal fine formuliamo le ipotesi $H_0: \{p_i\} = \{\pi_i\}$ e $H_1: \{p_i\} \neq \{\pi_i\}$, dove p_i ($i = 1, 2, \dots, 12$) è la probabilità che un generico matrimonio si celebri nel mese i -esimo, e dove $\pi_i = 1/12$ per $i = 1, 2, \dots, 12$. La distribuzione che si desidera sottoporre a verifica è quindi quella uniforme. Denotiamo con $n = 295990$ il numero totale di matrimoni celebrati nell'anno 1985, con n_i il numero di matrimoni celebrati nel mese i -esimo ($i = 1, 2, \dots, 12$) e con $k = 12$ il numero di classi. Dalla (5.57) segue:

$$\chi^2 = \frac{k}{n} \left(\sum_{i=1}^k n_i^2 \right) - n = 77984.069.$$

Se si sceglie $\alpha = 0.01$, dalla Tabella 3 dell'Appendice B risulta $\chi^2_{\alpha;k-1} = \chi^2_{0.01;11} = 24.725$,

5.4. TEST CHI-QUADRATO

così che la regione critica (5.56) diventa:

$$C = \{(x_1, x_2, \dots, x_n) : \chi^2 \geq 24.725\}.$$

Poiché il valore calcolato $\chi^2 = 77984.069$ è maggiore di 24.725, la realizzazione osservata appartiene alla regione critica. Siamo dunque condotti a rifiutare l'ipotesi nulla e ad accettare l'ipotesi alternativa, ossia che le date dei matrimoni non sono uniformemente distribuite nei 12 mesi. \diamond

Il test chi-quadrato viene frequentemente utilizzato quando la popolazione in esame è distribuita secondo una legge dedotta, ad esempio, attraverso la formulazione di un modello matematico, così che è essenziale verificare anzitutto che il modello sia compatibile con i dati osservati. La distribuzione di probabilità $(\pi_1, \pi_2, \dots, \pi_k)$ da assegnare alle classi è dunque suggerita proprio dal modello matematico cui si è fatto ricorso. Queste probabilità, in generale, dipendono da taluni parametri $\theta_1, \theta_2, \dots, \theta_r$. La distribuzione di probabilità si assume quindi nota a meno di eventuali parametri, così che scriveremo:

$$\pi_i = \varphi_i(\theta_1, \theta_2, \dots, \theta_r) \quad (i = 1, 2, \dots, k).$$

Chiaramente, se si denota con $f(x; \theta_1, \theta_2, \dots, \theta_r)$ la densità [distribuzione] di probabilità della variabile casuale genitrice X , risulta

$$\varphi_i(\theta_1, \theta_2, \dots, \theta_r) = \begin{cases} \int_{C_i} f(x; \theta_1, \theta_2, \dots, \theta_r) dx & \text{se } X \text{ è continua,} \\ \sum_{x \in C_i} f(x; \theta_1, \theta_2, \dots, \theta_r) & \text{se } X \text{ è discreta.} \end{cases}$$

I parametri $\theta_1, \theta_2, \dots, \theta_r$ vanno stimati facendo uso della realizzazione osservata del campione. Applicando i noti metodi di stima (ad esempio il metodo della massima verosimiglianza o il metodo dei momenti) si ricavano dunque le stime puntuali $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_r$, da cui si ottengono le probabilità

$$\hat{\pi}_i = \varphi_i(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_r) \quad (i = 1, 2, \dots, k).$$

A questo punto, per verificare l'ipotesi nulla $H_0: \{p_i\} = \{\hat{\pi}_i\}$ contro l'ipotesi alternativa $H_1: \{p_i\} \neq \{\hat{\pi}_i\}$, si costruisce la statistica

$$\Xi = \sum_{i=1}^k \frac{(N_i - n \hat{\pi}_i)^2}{n \hat{\pi}_i}. \quad (5.58)$$

È possibile dimostrare che, sotto l'ipotesi nulla e sotto ipotesi di regolarità — tra cui l'ipotesi che le funzioni $\varphi_i(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_r)$ posseggono derivate parziali continue rispetto a $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_r$ —, la statistica (5.58) possiede asintoticamente distribuzione chi-quadrato con $k - r - 1$ gradi di libertà. Si noti che, a differenza di quanto visto nella Proposizione 5.4.1, il numero $k - 1$ di gradi di libertà va ridotto ulteriormente di tante unità per quanti sono i parametri incogniti che vengono stimati a partire dalla realizzazione osservata del campione casuale. Ne segue che quest'ultimo procedimento può essere adottato solo quando il numero k di classi è maggiore di 2.

Una volta osservata una realizzazione (x_1, x_2, \dots, x_n) , si rifiuta l'ipotesi nulla se essa appartiene alla regione critica del test chi-quadrato

$$C = \{(x_1, x_2, \dots, x_n) : \chi^2 \geq \chi^2_{\alpha k - r - 1}\}, \quad (5.59)$$

dove

$$\chi^2 = \sum_{i=1}^k \frac{(n_i - n\hat{\pi}_i)^2}{n\hat{\pi}_i}. \quad (5.60)$$

Come visto in precedenza, l'ampiezza della regione critica (5.59) è approssimativamente α , e l'approssimazione migliora al crescere di n .

Esempio 5.4.3 Un centralino telefonico viene posto sotto controllo con l'obiettivo di analizzare con che frequenza riceve chiamate. Nel corso di un certo periodo di attività si registrano i seguenti 40 tempi di interrivo, in secondi, tra ogni telefonata e la successiva:

$$\begin{array}{cccccccccc} (5.6 & 20.5 & 31.2 & 2.1 & 42.6 & 11.6 & 58.3 & 24.5 & 7.8 & 18.6 \\ 48.5 & 3.5 & 15.3 & 51.1 & 4.7 & 26.3 & 1.9 & 33.4 & 17.6 & 6.7 \\ 10.1 & 36.1 & 9.2 & 22.6 & 46.7 & 3.8 & 12.8 & 21.8 & 8.1 & 38.4 \\ 5.2 & 27.2 & 13.5 & 45.1 & 7.1 & 14.3 & 28.3 & 2.7 & 16.2 & 0.5 \end{array}).$$

Riguardiamo questi dati come la realizzazione $(x_1, x_2, \dots, x_{40})$ di un campione casuale estratto da una popolazione di distribuzione incognita. Si desidera stabilire se, sulla base dei dati osservati, tale distribuzione possa essere riguardata come esponenziale. Suddividiamo a tal fine l'asse dei tempi nelle seguenti classi:

$$C_1 = (0, 10], C_2 = (10, 20], C_3 = (20, 30], C_4 = (30, 40], C_5 = (40, \infty).$$

Le probabilità da associare alle varie classi hanno la seguente forma:

$$\pi_i = \int_{C_i} \frac{1}{\theta} e^{-x/\theta} dx \quad (i = 1, 2, \dots, 5),$$

dove con θ indichiamo la media della variabile genitrice esponenziale. Come visto in precedenza (cfr. Esempio 3.7.2), il parametro incognito θ può essere stimato facendo ricorso allo stimatore di massima verosimiglianza costituito dalla media campionaria. Una stima puntuale di θ è quindi data da

$$\hat{\theta} = \bar{x} = \frac{1}{40} \sum_{i=1}^{40} x_i = 20.0375.$$

Le probabilità delle classi C_i da sottoporre a verifica sono dunque le seguenti:

$$\hat{\pi}_i = \int_{C_i} \frac{1}{\bar{x}} e^{-x/\bar{x}} dx = \begin{cases} 0.3929 & \text{per } i = 1, \\ 0.2385 & \text{per } i = 2, \\ 0.1448 & \text{per } i = 3, \\ 0.0879 & \text{per } i = 4, \\ 0.1359 & \text{per } i = 5. \end{cases} \quad (5.61)$$

5.4. TEST CHI-QUADRATO

Ricorriamo al test chi-quadrato per verificare l'ipotesi nulla $H_0: \{p_i\} = \{\hat{\pi}_i\}$ contro l'ipotesi alternativa $H_1: \{p_i\} \neq \{\hat{\pi}_i\}$. Dalla realizzazione osservata si ricavano le frequenze assolute per le 5 classi:

$$n_1 = 14, \quad n_2 = 9, \quad n_3 = 7, \quad n_4 = 4, \quad n_5 = 6.$$

Di conseguenza, per $n = 40$ e $k = 5$ la (5.60) diventa:

$$\begin{aligned} \chi^2 &= \sum_{i=1}^k \frac{(n_i - n\hat{\pi}_i)^2}{n\hat{\pi}_i} \\ &= \frac{(14 - 15.716)^2}{15.716} + \frac{(9 - 9.54)^2}{9.54} + \frac{(7 - 5.792)^2}{5.792} \\ &\quad + \frac{(4 - 3.516)^2}{3.516} + \frac{(6 - 5.436)^2}{5.436} \\ &= 3.2446 + 0.0306 + 0.2519 + 0.0666 + 0.0585 \\ &= 3.6522. \end{aligned}$$

Essendo $k - r - 1 = 3$, la regione critica (5.59) del test chi-quadrato avente approssimativamente ampiezza $\alpha = 0.025$ diventa:

$$C = \{(x_1, x_2, \dots, x_{40}) : \chi^2 \geq \chi^2_{0.025; 3}\}.$$

Dalla Tabella 3 dell'Appendice B risulta $\chi^2_{0.025; 3} = 9.348$; pertanto, poiché il valore $\chi^2 = 3.6522$ calcolato è minore di 9.348 la realizzazione osservata non appartiene alla regione critica, così che si può accettare l'ipotesi nulla che i tempi che intercorrono tra chiamate successive siano distribuiti esponenzialmente con media $\theta = 20.0375$. Osserviamo, però, che risulta $n\hat{\pi}_4 = 40 \cdot 0.0876 = 3.516$. In tal caso la condizione $n\hat{\pi}_i \geq 5$ non è soddisfatta per ogni valore i , e ciò potrebbe invalidare il test appena effettuato. Si conviene allora di unire la classe C_4 con la classe C_5 in modo da costruire una classe più ampia $C'_4 = C_4 \cup C_5$ la cui probabilità $\hat{\pi}'_4 \equiv \hat{\pi}_4 + \hat{\pi}_5$ risulti maggiore di $5/n$. In questo caso si ha $\hat{\pi}'_4 = 0.2238$ e $n'_4 = 10$. Pertanto, per $n = 40$ e $k = 4$ la (5.60) fornisce:

$$\chi^2 = 3.2446 + 0.0306 + 0.2519 + \frac{(10 - 8.952)^2}{8.952} = 3.6498.$$

Per $k - r - 1 = 2$, la regione critica (5.59) con ampiezza approssimativamente $\alpha = 0.025$ diventa allora:

$$C = \{(x_1, x_2, \dots, x_{40}) : \chi^2 \geq \chi^2_{0.025; 2}\}.$$

La Tabella 3 dell'Appendice B fornisce $\chi^2_{0.025; 2} = 7.378$; quindi, poiché il valore $\chi^2 = 3.6498$ calcolato è minore di 7.378, la realizzazione osservata non appartiene neanche alla nuova regione critica. Si può dunque nuovamente accettare l'ipotesi nulla. ◆

5.5 Differenze tra proporzioni

Un'importante applicazione del test chi-quadrato si presenta nelle situazioni in cui si desidera stabilire se differenze tra proporzioni o percentuali osservate sono statisticamente significative. Ad esempio, supponiamo che 12 di 90 confezioni di prodotti alimentari esaminate in un deposito non soddisfano i requisiti di corretta conservazione, mentre 9 di 50 confezioni in un differente deposito non sono soddisfacenti. Ci si può allora chiedere se le differenze tra le percentuali $12/90 = 0.13$ e $9/50 = 0.18$ siano da ritenersi significative.

Al fine di indicare un metodo generaleatto alla risoluzione di problemi di questo tipo, consideriamo k variabili casuali X_1, X_2, \dots, X_k indipendenti, dove ogni variabile X_i ha distribuzione binomiale di parametri θ_i incognito e n_i noto. Si desidera costruire un test per verificare l'ipotesi nulla

$$H_0: \theta_1 = \theta_2 = \dots = \theta_k = \theta_0, \quad (5.62)$$

contro l'ipotesi alternativa che almeno uno dei parametri θ_i non sia uguale a θ_0 .

Proposizione 5.5.1 *Sotto la condizione che l'ipotesi nulla (5.62) sia vera, con θ_0 noto, al divergere di n_1, n_2, \dots, n_k la statistica*

$$\Xi = \sum_{i=1}^k \frac{(X_i - n_i \theta_0)^2}{n_i \theta_0 (1 - \theta_0)} \quad (5.63)$$

assume distribuzione chi-quadrato con k gradi di libertà.

Dim. Sotto l'ipotesi nulla, per $i = 1, 2, \dots, k$ la variabile casuale

$$Z_i = \frac{X_i - n_i \theta_0}{\sqrt{n_i \theta_0 (1 - \theta_0)}}$$

al divergere di n_i assume distribuzione normale standard. Pertanto, in virtù del Teorema 2.1.1, la statistica (5.63) è espressa come somma di k variabili casuali indipendenti ciascuna avente asintoticamente distribuzione chi-quadrato con 1 grado di libertà. Dal Teorema 2.1.2 segue allora che Ξ assume distribuzione chi-quadrato con k gradi di libertà al divergere di n_1, n_2, \dots, n_k . ■

Esaminiamo ora come si procede per costruire una regione di rifiuto dell'ipotesi nulla (5.62). In virtù della Proposizione 5.5.1, al divergere di n_1, n_2, \dots, n_k la variabile (5.63) ha distribuzione chi-quadrato con k gradi di libertà. Pertanto, se indichiamo con x_1, x_2, \dots, x_k i valori osservati delle variabili casuali binomiali X_1, X_2, \dots, X_k e con

$$\chi^2 = \sum_{i=1}^k \frac{(x_i - n_i \theta_0)^2}{n_i \theta_0 (1 - \theta_0)}$$

il valore assunto da Ξ in corrispondenza di x_1, x_2, \dots, x_k , l'insieme

$$C = \{(x_1, x_2, \dots, x_k) : \chi^2 \geq \chi^2_{\alpha; k}\} \quad (5.64)$$

costituisce una regione critica di ampiezza all'incirca α , dove l'approssimazione migliora al crescere di ciascuno dei parametri n_1, n_2, \dots, n_k . L'ipotesi nulla (5.62) viene dunque rifiutata se i valori osservati x_1, x_2, \dots, x_k appartengono alla regione critica (5.64); in caso contrario essa viene accettata.

Esempio 5.5.1 Una società viene incaricata di compiere un'indagine demoscopica per valutare il gradimento di una certa propaganda pubblicitaria per fasce d'età prefissate. Dei 120 giovani intervistati, 98 sono favorevoli; dei 180 adulti intervistati, ne sono favorevoli 137; infine, su 140 anziani intervistati si sono riscontrati 103 pareri favorevoli. Assumiamo che le decisioni prese da ogni individuo intervistato siano tra loro indipendenti, e supponiamo inoltre che ogni giovane, ogni adulto, ogni anziano intervistato sia favorevole alla propaganda pubblicitaria con probabilità $\theta_1, \theta_2, \theta_3$, rispettivamente. Sotto tali ipotesi le variabili casuali X_1, X_2, X_3 , descriventi il numero di giovani, adulti, anziani favorevoli al messaggio pubblicitario posseggono distribuzioni binomiali di parametri $\theta_1, \theta_2, \theta_3$ e $n_1 = 120, n_2 = 180, n_3 = 140$, rispettivamente. Si desidera sottoporre a verifica l'ipotesi nulla

$$H_0: \theta_1 = \theta_2 = \theta_3 = 0.75$$

contro l'ipotesi alternativa che almeno uno dei parametri θ_i non sia uguale a 0.75. Sotto l'ipotesi nulla H_0 , la statistica (5.63), che ha qui approssimativamente distribuzione chi-quadrato con $k = 3$ gradi di libertà, assume il valore

$$\begin{aligned} \chi^2 &= \sum_{i=1}^3 \frac{(x_i - 0.75 n_i)^2}{0.1875 n_i} \\ &= \frac{(98 - 120 \cdot 0.75)^2}{120 \cdot 0.1875} + \frac{(137 - 180 \cdot 0.75)^2}{180 \cdot 0.1875} + \frac{(103 - 140 \cdot 0.75)^2}{140 \cdot 0.1875} \\ &= 3.1153, \end{aligned}$$

avendo indicato con $x_1 = 98, x_2 = 137$ e $x_3 = 103$ i valori assunti dalle variabili casuali binomiali. Dalla (5.64) segue che per $\alpha = 0.05$ la regione di rifiuto dell'ipotesi nulla è la seguente

$$C = \{(x_1, x_2, x_3) : \chi^2 \geq \chi^2_{0.05; 3}\}.$$

Dalla Tabella 3 dell'Appendice B si ricava $\chi^2_{0.05; 3} = 7.815$; pertanto, poiché il valore $\chi^2 = 3.1153$ calcolato è minore di 7.815, i valori x_1, x_2, x_3 osservati non appartengono a C . Si conclude che è plausibile accettare l'ipotesi nulla $H_0: \theta_1 = \theta_2 = \theta_3 = 0.75$. ◆

Finora è stato esaminato il caso in cui nell'ipotesi nulla (5.62) il parametro θ_0 è supposto noto. Spesso, però, accade che il valore di θ_0 sia incognito, e che debba quindi essere stimato a partire dai dati di cui si dispone. In questo caso in luogo della (5.62) si sottopone a verifica l'ipotesi nulla

$$H_0: \theta_1 = \theta_2 = \dots = \theta_k \quad (5.65)$$

contro l'ipotesi alternativa che i parametri θ_i non siano tutti uguali tra loro. Per costruire la regione critica del test, il valore θ_0 che appare nella statistica (5.63) va sostituito con la sua stima di massima verosimiglianza (cfr. Esempio 3.7.3)

$$\hat{\theta} = \frac{x_1 + x_2 + \dots + x_k}{n_1 + n_2 + \dots + n_k}. \quad (5.66)$$

Tabella 5.3: Polizze contro incendi (x_i) e in totale (n_i).

i	x_i	n_i
1	32	125
2	17	130
3	22	115
4	12	80
5	16	175
6	56	150

Proposizione 5.5.2 Sotto la condizione che l'ipotesi nulla (5.65) sia vera, al divergere di n_1, n_2, \dots, n_k la statistica

$$\Xi = \sum_{i=1}^k \frac{(x_i - n_i \hat{\theta})^2}{n_i \hat{\theta}(1 - \hat{\theta})} \quad (5.67)$$

assume distribuzione chi-quadrato con $k - 1$ gradi di libertà.

Mentre nella Proposizione 5.5.1 la variabile Ξ assume asintoticamente distribuzione chi-quadrato con k gradi di libertà, ora la statistica (5.67) di gradi di libertà ne ha $k - 1$. Il numero di gradi di libertà si decrementa di un'unità in quanto nella stima (5.66), utilizzata nella (5.67), si fa uso dei dati x_1, x_2, \dots, x_k .

Per costruire la regione di rifiuto dell'ipotesi nulla (5.65) si procede come nel caso precedente. Pertanto, se denotiamo con

$$\chi^2 = \sum_{i=1}^k \frac{(x_i - n_i \hat{\theta})^2}{n_i \hat{\theta}(1 - \hat{\theta})} \quad (5.68)$$

il valore assunto dalla statistica (5.67) in corrispondenza dei dati osservati x_1, x_2, \dots, x_k , l'insieme

$$C = \{(x_1, x_2, \dots, x_k) : \chi^2 \geq \chi^2_{\alpha; k-1}\} \quad (5.69)$$

costituisce una regione critica di ampiezza all'incirca α . L'approssimazione, al solito, migliora al crescere di ognuno dei parametri n_1, n_2, \dots, n_k . Si conclude che l'ipotesi nulla (5.65) è da rifiutarsi se i valori x_1, x_2, \dots, x_k osservati appartengono alla regione critica (5.69); in caso contrario essa va accettata.

Esempio 5.5.2 Una compagnia assicuratrice è proprietaria di 6 agenzie in altrettante città. Nella Tabella 5.3 sono riportati il numero x_i di polizze contro incendi e il numero totale n_i di polizze stipulate in un mese dall'agenzia i -esima ($i = 1, 2, \dots, 6$). Supponiamo che la variabile casuale X_i , descrivente il numero di polizze contro incendi stipulate nel mese in esame dall'agenzia i -esima possa essere riguardata come variabile binomiale di parametri θ_i e n_i ($i = 1, 2, \dots, 6$). Assumendo che X_1, X_2, \dots, X_6 siano indipendenti, si desidera costruire un test per verificare l'ipotesi nulla

$$H_0: \theta_1 = \theta_2 = \dots = \theta_6$$

5.5. DIFFERENZE TRA PROPORZIONI

contro l'ipotesi alternativa che i parametri θ_i non siano tutti uguali tra loro. Dalla (5.66) segue la stima

$$\hat{\theta} = \frac{x_1 + x_2 + \dots + x_6}{n_1 + n_2 + \dots + n_6} = \frac{155}{775} = 0.2.$$

Sotto l'ipotesi nulla H_0 la statistica (5.67), avente approssimativamente distribuzione chi-quadrato con $k - 1 = 5$ gradi di libertà, assume allora il valore

$$\begin{aligned} \chi^2 &= \sum_{i=1}^6 \frac{(x_i - n_i \cdot 0.2)^2}{n_i \cdot 0.16} \\ &= \frac{(32 - 125 \cdot 0.2)^2}{125 \cdot 0.16} + \frac{(17 - 130 \cdot 0.2)^2}{130 \cdot 0.16} + \frac{(22 - 115 \cdot 0.2)^2}{115 \cdot 0.16} \\ &\quad + \frac{(12 - 80 \cdot 0.2)^2}{80 \cdot 0.16} + \frac{(16 - 175 \cdot 0.2)^2}{175 \cdot 0.16} + \frac{(56 - 150 \cdot 0.2)^2}{150 \cdot 0.16} = 48.7. \end{aligned}$$

Facendo uso della (5.69) si deduce che per $\alpha = 0.025$ la regione di rifiuto dell'ipotesi nulla è la seguente:

$$C = \{(x_1, x_2, \dots, x_6) : \chi^2 \geq \chi^2_{0.025; 5}\}.$$

Dalla Tabella 3 dell'Appendice B si ricava $\chi^2_{0.025; 5} = 12.832$; poiché $\chi^2 = 48.7$ è maggiore di 12.832, i valori x_1, x_2, \dots, x_6 osservati appartengono a C , così che l'ipotesi nulla $H_0: \theta_1 = \theta_2 = \dots = \theta_6$ non è accettabile.

Notiamo che dalla Tabella 5.3 si possono ricavare le percentuali osservate $f_i \equiv x_i/n_i$:

$$f_1 = 0.256, f_2 = 0.131, f_3 = 0.191, f_4 = 0.150, f_5 = 0.091, f_6 = 0.373.$$

Da queste si ricava che le percentuali f_5 e f_6 sono quelle che si discostano maggiormente dalla stima $\hat{\theta} = 0.2$. Ciò suggerisce di verificare l'ipotesi che i parametri θ_i siano uguali tra loro, restringendo l'ipotesi ai soli casi $i = 1, 2, 3, 4$. Verifichiamo dunque l'ipotesi nulla

$$H_0: \theta_1 = \theta_2 = \theta_3 = \theta_4$$

contro l'ipotesi alternativa che tali parametri non siano tutti uguali tra loro. Per $k = 4$, dalla (5.66) segue ora la stima

$$\hat{\theta} = \frac{x_1 + x_2 + x_3 + x_4}{n_1 + n_2 + n_3 + n_4} = \frac{83}{450} = 0.18\bar{4}.$$

Conseguentemente sotto la nuova ipotesi nulla la statistica (5.67), che ha ora approssimativamente distribuzione chi-quadrato con $k - 1 = 3$ gradi di libertà, assume il valore

$$\begin{aligned} \chi^2 &= \sum_{i=1}^4 \frac{(x_i - n_i \cdot 0.18\bar{4})^2}{n_i \cdot 0.1504} \\ &= \frac{(32 - 125 \cdot 0.18\bar{4})^2}{125 \cdot 0.1504} + \frac{(17 - 130 \cdot 0.18\bar{4})^2}{130 \cdot 0.1504} \\ &\quad + \frac{(22 - 115 \cdot 0.18\bar{4})^2}{115 \cdot 0.1504} + \frac{(12 - 80 \cdot 0.18\bar{4})^2}{80 \cdot 0.1504} = 7.42. \end{aligned}$$

Per $\alpha = 0.025$ la regione critica (5.69) diventa:

$$C = \{(x_1, x_2, x_3, x_4) : \chi^2 \geq \chi^2_{0.025;3}\},$$

e dalla Tabella 3 dell'Appendice B si ottiene $\chi^2_{0.025;3} = 9.348$. Stavolta il valore $\chi^2 = 7.42$ osservato è minore di 9.348, così che i valori x_1, x_2, x_3, x_4 osservati non appartengono alla regione critica C : di conseguenza l'ipotesi nulla $H_0: \theta_1 = \theta_2 = \theta_3 = \theta_4$ è da accettare.

Dai dati riportati nella Tabella 5.3 si traggono pertanto le seguenti conclusioni: le percentuali teoriche di polizze contro incendi stipulate nelle agenzie delle prime 4 città si possono assumere identiche; ciò non è invece legittimo per la totalità delle 6 agenzie. ♦

5.6 Tabelle di contingenza

In numerose applicazioni il test chi-quadrato costituisce uno strumento utile per stabilire se due o più caratteristiche di una popolazione sono indipendenti, come ora indicheremo.

Supponiamo di avere in esame una popolazione contraddistinta da due caratteristiche qualitative o quantitative, A e B , ciascuna classificabile in un prefissato numero di classi. Se indichiamo tali classi rispettivamente con A_1, A_2, \dots, A_k e B_1, B_2, \dots, B_l , la popolazione risulta suddivisa in $k \cdot l$ classi. Denotiamo queste con $C_{ij} = A_i \cap B_j$, e le probabilità corrispondenti con p_{ij} ($i = 1, 2, \dots, k; j = 1, 2, \dots, l$). Ciò significa che un elemento della popolazione scelto a caso appartiene alla classe C_{ij} con probabilità p_{ij} .

Consideriamo un campione casuale (X_1, X_2, \dots, X_n) estratto dalla popolazione e, in analogia con la (5.49), definiamo la frequenza assoluta

$$N_{ij} = \left| \{X_r : X_r \in C_{ij}; r = 1, 2, \dots, n\} \right| = \sum_{r=1}^n I_{C_{ij}}(X_r) \quad (i = 1, 2, \dots, k; j = 1, 2, \dots, l)$$

che possiede di distribuzione binomiale di parametri n e p_{ij} . Data la realizzazione (x_1, x_2, \dots, x_n) di un campione estratto dalla popolazione in esame, indichiamo con n_{ij} il valore assunto dalla frequenza assoluta N_{ij} , ossia il numero di elementi della realizzazione che appartengono alla classe C_{ij} . Si ha:

$$\sum_{i=1}^k \sum_{j=1}^l n_{ij} = n.$$

Si possono inoltre definire le seguenti *frequenze marginali*:

$$N_{i \cdot} = \sum_{j=1}^l N_{ij} \quad (i = 1, 2, \dots, k)$$

$$N_{\cdot j} = \sum_{i=1}^k N_{ij} \quad (j = 1, 2, \dots, l),$$

cui corrispondono i seguenti valori

$$n_{i \cdot} = \sum_{j=1}^l n_{ij} \quad n_{\cdot j} = \sum_{i=1}^k n_{ij}$$

5.6. TABELLE DI CONTINGENZA

in base alla realizzazione (x_1, x_2, \dots, x_n) . Pertanto $n_{i \cdot}$ e $n_{\cdot j}$ denotano rispettivamente la frequenza assoluta della classe A_i e quella della classe B_j . È conveniente riportare le frequenze assolute di ogni classe C_{ij} e le frequenze marginali corrispondenti in una tabella del seguente tipo, detta *tavola di contingenza*:

	B_1	B_2	\dots	B_j	\dots	B_l	
A_1	n_{11}	n_{12}	\dots	n_{1j}	\dots	n_{1l}	$n_{1 \cdot}$
A_2	n_{21}	n_{22}	\dots	n_{2j}	\dots	n_{2l}	$n_{2 \cdot}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
A_l	n_{l1}	n_{l2}	\dots	n_{lj}	\dots	n_{ll}	$n_{l \cdot}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
A_k	n_{k1}	n_{k2}	\dots	n_{kj}	\dots	n_{kl}	$n_{k \cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	\dots	$n_{\cdot j}$	\dots	$n_{\cdot l}$	n

Il problema che si vuole ora affrontare consiste nello stabilire se le caratteristiche A e B della popolazione possano riguardarsi, o meno, come indipendenti. Se si indicano rispettivamente con $p_{i \cdot}$ e con $p_{\cdot j}$ le probabilità marginali relative alle classi A_i e B_j , si ha:

$$p_{i \cdot} = \sum_{j=1}^l p_{ij} \quad (i = 1, 2, \dots, k)$$

$$p_{\cdot j} = \sum_{i=1}^k p_{ij} \quad (j = 1, 2, \dots, l).$$

Si noti che le variabili casuali $N_{i \cdot}$ hanno distribuzione binomiale di parametri n e $p_{i \cdot}$, e le $N_{\cdot j}$ distribuzione binomiale di parametri n e $p_{\cdot j}$. Le caratteristiche A e B della popolazione sono dunque indipendenti quando risulta

$$p_{ij} = p_{i \cdot} p_{\cdot j}$$

per ogni coppia i, j con $i = 1, 2, \dots, k$ e $j = 1, 2, \dots, l$. L'ipotesi nulla da verificare è allora:

$$H_0: \{p_{ij}\} = \{p_{i \cdot} p_{\cdot j}\} \quad (5.70)$$

contro l'ipotesi alternativa

$$H_1: \{p_{ij}\} \neq \{p_{i \cdot} p_{\cdot j}\}.$$

A questo punto si possono presentare due situazioni.

(i) Se le probabilità marginali $p_{i \cdot}$ e $p_{\cdot j}$ sono note per ogni i, j , la verifica dell'ipotesi nulla (5.70) si riconduce al test di conformità che abbiamo visto in precedenza. In questo caso si fa ricorso alla statistica

$$\Xi = \sum_{i=1}^k \sum_{j=1}^l \frac{(N_{ij} - n p_{i \cdot} p_{\cdot j})^2}{n p_{i \cdot} p_{\cdot j}}. \quad (5.71)$$

Questa, in analogia con la (5.51) e con quanto mostrato nella Proposizione 5.4.1, ha asintoticamente distribuzione chi-quadrato con $k l - 1$ gradi di libertà. Posto

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij} - n p_i, p_j)^2}{n p_i, p_j},$$

l'ipotesi nulla di indipendenza (5.70) viene rifiutata se la realizzazione osservata appartiene alla regione critica

$$C = \{(x_1, x_2, \dots, x_n) : \chi^2 \geq \chi^2_{\alpha; k l - 1}\}$$

di ampiezza approssimativamente α ; in caso contrario l'ipotesi nulla H_0 viene accettata. Anche in questo caso si fa ricorso ad una approssimazione chi-quadrato; si assume, allora, che il criterio adottato sia soddisfacente per n grande e quando risulta $n p_i, p_j \geq 5$ per ogni i, j .

(ii) Se le probabilità marginali p_i e p_j sono incognite — situazione che si incontra con maggiore frequenza nelle applicazioni concrete —, è necessario stimarle a partire dalla realizzazione osservata del campione. Poiché le frequenze marginali N_i e N_j hanno distribuzioni binomiali, la stima di p_i e p_j si riduce alla stima del parametro di una distribuzione binomiale. Come visto nell'Esempio 3.7.3, il metodo della massima verosimiglianza suggerisce di stimare il parametro di una distribuzione binomiale mediante la media campionaria. In questo caso tale stima si esprime attraverso le frequenze relative corrispondenti:

$$\begin{aligned} \hat{p}_{i \cdot} &= \frac{n_{i \cdot}}{n} = \frac{n_{1 \cdot} + n_{2 \cdot} + \dots + n_{k \cdot}}{n} \quad (i = 1, 2, \dots, k) \\ \hat{p}_{\cdot j} &= \frac{n_{\cdot j}}{n} = \frac{n_{1 j} + n_{2 j} + \dots + n_{k j}}{n} \quad (j = 1, 2, \dots, l). \end{aligned}$$

Sostituendo tali stime nella (5.71) si ricava la statistica

$$\Xi = \sum_{i=1}^k \sum_{j=1}^l \frac{(N_{ij} - n \hat{p}_{i \cdot} \hat{p}_{\cdot j})^2}{n \hat{p}_{i \cdot} \hat{p}_{\cdot j}} = \sum_{i=1}^k \sum_{j=1}^l \frac{(N_{ij} - n_{i \cdot} n_{\cdot j} / n)^2}{n_{i \cdot} n_{\cdot j} / n}. \quad (5.72)$$

Poiché

$$\sum_{i=1}^k p_{i \cdot} = \sum_{j=1}^l p_{\cdot j} = 1,$$

i parametri da stimare sono in realtà $k - 1 + l - 1$. Pertanto la statistica (5.72) ha asintoticamente distribuzione chi-quadrato con un numero di gradi libertà pari a

$$k l - 1 - (k - 1 + l - 1) = (k - 1)(l - 1).$$

Posto

$$\chi^2 = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij} - n_{i \cdot} n_{\cdot j} / n)^2}{n_{i \cdot} n_{\cdot j} / n}, \quad (5.73)$$

l'ipotesi di indipendenza $H_0: \{p_{ij}\} = \{\hat{p}_{i \cdot}, \hat{p}_{\cdot j}\}$ va dunque rifiutata se la realizzazione osservata appartiene alla regione critica

$$C = \{(x_1, x_2, \dots, x_n) : \chi^2 \geq \chi^2_{\alpha; (k-1)(l-1)}\} \quad (5.74)$$

che ha approssimativamente ampiezza α ; in caso contrario l'ipotesi nulla H_0 viene accettata. Analogamente al caso (i), il criterio si ritiene soddisfacente per n grande e allorché risulta $n \hat{p}_{i \cdot} \hat{p}_{\cdot j} \equiv n_{i \cdot} n_{\cdot j} / n \geq 5$.

Esempio 5.6.1 Un commerciante di elettrodomestici si rifornisce di pezzi di ricambio acquistandone rispettivamente $n_{1 \cdot} = 100$, $n_{2 \cdot} = 150$, e $n_{3 \cdot} = 200$ presso tre diversi centri di produzione. Nel totale degli $n = 450$ pezzi acquistati, $n_{\cdot 1} = 270$ risultano essere di qualità eccellente, $n_{\cdot 2} = 105$ di buona qualità e $n_{\cdot 3} = 75$ di qualità scadente. Nella seguente tabella di contingenza sono indicate in dettaglio le frequenze assolute della qualità dei pezzi di ricambio acquistati presso i tre centri di produzione:

	B_1	B_2	B_3	
A_1	61	25	14	100
A_2	88	37	25	150
A_3	121	43	36	200
	270	105	75	450

dove A_1, A_2 e A_3 indicano i centri di produzione e B_1, B_2 e B_3 stanno per qualità eccellente, buona e scadente. Si desidera verificare se esiste indipendenza tra i centri di produzione dei pezzi di ricambio ed i corrispondenti livelli di qualità. Poiché le distribuzioni marginali non sono note, si ricorre alla loro stima mediante le frequenze relative $\hat{p}_{i \cdot}$ e $\hat{p}_{\cdot j}$, così che l'ipotesi nulla da verificare è $H_0: \{p_{ij}\} = \{\hat{p}_{i \cdot}, \hat{p}_{\cdot j}\}$. Essendo $k = l = 3$ e $n = 450$, la (5.73) ora fornisce:

$$\begin{aligned} \chi^2 &= \sum_{i=1}^3 \sum_{j=1}^3 \frac{(n_{ij} - n_{i \cdot} n_{\cdot j} / 450)^2}{n_{i \cdot} n_{\cdot j} / 450} \\ &= \frac{(61 - 60)^2}{60} + \frac{(25 - 23.3)^2}{23.3} + \frac{(14 - 16.6)^2}{16.6} \\ &\quad + \frac{(88 - 90)^2}{90} + \frac{(37 - 35)^2}{35} + \frac{(25 - 25)^2}{25} \\ &\quad + \frac{(121 - 120)^2}{120} + \frac{(43 - 46.6)^2}{46.6} + \frac{(36 - 33.3)^2}{33.3} = 1.229. \end{aligned}$$

Se fissiamo $\alpha = 0.05$, in virtù della (5.74) la regione di rifiuto dell'ipotesi nulla è costituita dalle realizzazioni per le quali risulta $\chi^2 \geq \chi^2_{0.05; 4}$, essendo $(k-1)(l-1) = 4$. Poiché dalla Tabella 3 dell'Appendice B si ricava $\chi^2_{0.05; 4} \equiv 9.488$, il valore calcolato $\chi^2 = 1.229$ è tale che la realizzazione non appartiene alla regione critica. Si può dunque accettare l'ipotesi di indipendenza tra i centri di produzione dei pezzi di ricambio ed i corrispondenti livelli di qualità. ◆

È opportuno sottolineare che il procedimento sopra descritto non è altro che un'estensione di quello presentato nel § 5.5. Invero, vedremo ora che nel caso particolare $l = 2$ la (5.73) si identifica con la (5.68). Supponiamo dunque che la caratteristica B della popolazione sia classificabile in $l = 2$ classi distinte. Effettuiamo inoltre le seguenti posizioni:

$$n_{ij} = \begin{cases} x_i & \text{per } j = 1, \\ n_i - x_i & \text{per } j = 2, \end{cases} \quad n_{i \cdot} = n_i \quad (i = 1, 2, \dots, k). \quad (5.75)$$

Essendo $n_1 + n_2 + \dots + n_k = n$, la (5.66) fornisce

$$\hat{\theta} = \frac{x_1 + x_2 + \dots + x_k}{n}$$

e quindi risulta:

$$n_{i,j} = \begin{cases} x_1 + x_2 + \dots + x_k \equiv n\hat{\theta} & \text{per } j = 1, \\ n - (x_1 + x_2 + \dots + x_k) \equiv n(1 - \hat{\theta}) & \text{per } j = 2. \end{cases}$$

La tabella di contingenza relativa al caso $l = 2$ si può dunque rappresentare nel seguente modo:

	B_1	B_2	
A_1	x_1	$n_1 - x_1$	n_1
A_2	x_2	$n_2 - x_2$	n_2
⋮	⋮	⋮	⋮
A_i	x_i	$n_i - x_i$	n_i
⋮	⋮	⋮	⋮
A_k	x_k	$n_k - x_k$	n_k
	$x_1 + \dots + x_k$	$n - (x_1 + \dots + x_k)$	n

Per le posizioni (5.75) effettuate, la (5.73) diventa:

$$\begin{aligned} \chi^2 &= \sum_{i=1}^k \sum_{j=1}^2 \frac{(n_{ij} - n_i n_{.j}/n)^2}{n_i n_{.j}/n} \\ &= \sum_{i=1}^k \left\{ \frac{(x_i - n_i \hat{\theta})^2}{n_i \hat{\theta}} + \frac{[n_i - x_i - n_i(1 - \hat{\theta})]^2}{n_i(1 - \hat{\theta})} \right\} \\ &= \sum_{i=1}^k \frac{(x_i - n_i \hat{\theta})^2}{n_i} \left[\frac{1}{\hat{\theta}} + \frac{1}{(1 - \hat{\theta})} \right] \\ &= \sum_{i=1}^k \frac{(x_i - n_i \hat{\theta})^2}{n_i \hat{\theta}(1 - \hat{\theta})}, \end{aligned}$$

che, come preannunciato, coincide con la (5.68). Si noti, inoltre, che nel caso particolare $k = l = 2$ risulta:

$$(n_{ij} - n_i n_{.j}/n)^2 = \frac{n_{11} n_{22} - n_{12} n_{21}}{n^2} \quad (i, j = 1, 2),$$

così che la (5.73) assume la seguente più semplice forma:

$$\chi^2 = \frac{(n_{11} n_{22} - n_{12} n_{21})^2}{n} \sum_{i,j=1}^2 \frac{1}{n_i n_{.j}}. \quad (5.76)$$

L'ipotesi di indipendenza $H_0: \{p_{ij}\} = \{\hat{p}_i, \hat{p}_{.j}\}$ va quindi rifiutata se la realizzazione osservata appartiene alla regione critica

$$C = \{(x_1, x_2, \dots, x_n) : \chi^2 \geq \chi^2_{\alpha/2}\} \quad (5.77)$$

la cui ampiezza è all'incirca α ; in caso contrario l'ipotesi nulla H_0 viene accettata.

Esempio 5.6.2 Un corso universitario è seguito da $n = 150$ studenti dei quali 36 sono maschi e 114 femmine. Dei 36 maschi, 31 seguono il piano di studi individuale e 5 quello statutario, mentre delle 114 femmine 95 seguono il piano di studi individuale e 19 quello statutario. Vogliamo verificare se esiste indipendenza tra il sesso degli studenti e il tipo di indirizzo scelto assumendo $\alpha = 0.05$ come ampiezza della regione critica. A tal fine denotiamo con M e F le categorie dei maschi e delle femmine, rispettivamente, e con I e S le categorie degli studenti con piano di studi individuale e statutario, rispettivamente. Possiamo allora costruire la seguente tabella di contingenza:

	I	S	
M	31	5	36
F	95	19	114
	126	24	150

Poiché siamo nel caso $k = l = 2$, facciamo uso della formula (5.76) ottenendo:

$$\begin{aligned} \chi^2 &= \frac{(n_{MI} n_{FS} - n_{MS} n_{FI})^2}{n} \left(\frac{1}{n_M n_I} + \frac{1}{n_M n_S} + \frac{1}{n_F n_I} + \frac{1}{n_F n_S} \right) \\ &= \frac{12996}{150} \left(\frac{1}{4536} + \frac{1}{864} + \frac{1}{14364} + \frac{1}{2736} \right) = 0.157. \end{aligned}$$

In corrispondenza del valore $\alpha = 0.05$ si ha $\chi^2_{0.05/2} = 3.841$ (cfr. Tabella 3 in Appendice B), così che la regione critica (5.77) diventa

$$C = \{(x_1, x_2, \dots, x_n) : \chi^2 \geq 3.841\}.$$

Poiché il valore calcolato $\chi^2 = 0.157$ è minore di 3.841, la realizzazione osservata non appartiene alla regione critica: l'ipotesi di indipendenza tra sesso degli studenti e tipo di indirizzo scelto va dunque accettata. ◆

Capitolo 6

Regressione e correlazione

6.1 Regressione

Tra le finalità delle indagini statistiche rientra l'individuazione di relazioni tra grandezze casuali atte a far sì che talune di queste siano predicibili a partire da altre. Si pensi, ad esempio, alla possibilità di prevedere il consumo di un'automobile in base alla lunghezza di un percorso da effettuarsi, il peso che un paziente raggiungerà al termine di un prefissato numero di giorni di dieta, il rendimento di uno studente ad un esame in conseguenza del numero di ore di studio, e così via.

Per quanto auspicabile sia il predire correttamente una variabile mediante altre, occorre tener conto che si ha a che fare con grandezze casuali, e che quindi solitamente si possono al più ricavare informazioni circa i comportamenti medi delle variabili in gioco. Così, ad esempio, non si è in grado di predire esattamente il consumo d'olio di una particolare auto dopo un prefissato chilometraggio; a partire, però, da opportuni dati può risultare possibile predire il consumo medio d'olio di un'auto di un certo tipo in funzione del numero di chilometri percorsi.

Nel seguito ci riferiremo a problemi di previsione coinvolgenti variabili casuali che supporremo definite in uno stesso spazio di probabilità.

Date due variabili casuali X e Y , il problema principale della *regressione bivariata* consiste nel determinare la media condizionata $E(Y|x)$, ossia il valore medio di Y dato che $X = x$. Se la densità [distribuzione] di probabilità congiunta di X e Y è nota, il problema della regressione bivariata si risolve agevolmente. Infatti se X e Y sono continue con densità di probabilità congiunta $f_{X,Y}(x,y)$, la media condizionata $E(Y|x)$ è la seguente:

$$E(Y|x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy, \quad (6.1)$$

dove la densità di Y condizionata da $X = x$ si ottiene dalla relazione

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \quad (6.2)$$

con

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy. \quad (6.3)$$

Nel caso in cui X e Y sono variabili casuali discrete si presentano distribuzioni di probabilità invece che densità di probabilità; alla (6.1) va pertanto sostituita la relazione seguente:

$$E(Y|x) = \sum_y y P(Y=y|X=x), \quad (6.4)$$

dove

$$P(Y=y|X=x) = \frac{P(X=x, Y=y)}{P(X=x)}$$

con

$$P(X=x) = \sum_y P(X=x, Y=y).$$

Le relazioni (6.1) e (6.4) sono dette *equazioni di regressione* di Y data X ; si noti che i primi membri sono funzioni della variabile indipendente x .

Esamineremo ora due esempi in cui si ricavano le equazioni di regressione per variabili casuali continue.

Esempio 6.1.1 Sia

$$f_{X,Y}(x,y) = \begin{cases} \frac{3y}{1-x^2} & \text{per } 0 < x < 1, 0 < y < 1-x^2, \\ 0 & \text{altrimenti} \end{cases} \quad (6.5)$$

la densità di probabilità congiunta delle variabili casuali X e Y . Dalle (6.3) e (6.5) segue:

$$f_X(x) = \frac{3}{1-x^2} \int_0^{1-x^2} y dy = \frac{3}{2} (1-x^2) \quad (0 < x < 1). \quad (6.6)$$

Utilizzando le (6.2) e (6.6) ricaviamo la densità di probabilità condizionata di Y data X :

$$f_{Y|X}(y|x) = \frac{2y}{(1-x^2)^2} \quad (0 < x < 1, 0 < y < 1-x^2). \quad (6.7)$$

Dalle (6.1) e (6.7) segue infine il valore medio di Y condizionato da $X=x$:

$$\begin{aligned} E(Y|x) &= \int_0^{1-x^2} y f_{Y|X}(y|x) dy = \int_0^{1-x^2} \frac{2y^2}{(1-x^2)^2} dy \\ &= \frac{2}{3} (1-x^2) \quad (0 < x < 1). \end{aligned} \quad (6.8)$$

L'equazione di regressione (6.8) ottenuta è dunque quadratica in x .

6.1. REGRESSIONE

Esempio 6.1.2 Supponiamo che le variabili casuali continue X e Y abbiano densità di probabilità congiunta

$$f_{X,Y}(x,y) = \begin{cases} \frac{2a}{(x+y+a)^3} & \text{per } x, y > 0, \\ 0 & \text{altrimenti,} \end{cases} \quad (6.9)$$

dove a è una costante positiva arbitraria. Dalle (6.3) e (6.9) si ricava la densità di probabilità di X :

$$f_X(x) = \int_0^{\infty} \frac{2a}{(x+y+a)^3} dy = \frac{a}{(x+a)^2} \quad (x > 0). \quad (6.10)$$

Utilizzando le (6.2) e (6.10) si ottiene la densità di probabilità di Y condizionata da $X=x$:

$$f_{Y|X}(y|x) = \frac{2(x+a)^2}{(x+y+a)^3} \quad (x, y > 0). \quad (6.11)$$

Dall'equazione di regressione (6.1), e facendo uso della (6.11), si ha poi:

$$\begin{aligned} E(Y|x) &= \int_0^{\infty} \frac{2y(x+a)^2}{(x+y+a)^3} dy \\ &= \left[\frac{y(x+a)^2}{(x+y+a)^2} \right]_0^{\infty} + \int_0^{\infty} \left(\frac{x+a}{x+y+a} \right)^2 dy \\ &= x+a \quad (x > 0). \end{aligned} \quad (6.12)$$

Come mostra la (6.12), in questo caso si ottiene un'equazione di regressione lineare in x . ◆

Particolare importanza riveste il caso della *regressione lineare*, ossia quando risulta

$$E(Y|x) = a + bx, \quad (6.13)$$

dove a e b sono costanti reali dette *coefficienti di regressione* dell'equazione di regressione lineare. Vi sono svariati motivi per i quali l'equazione di regressione lineare riveste particolare interesse. Innanzitutto, in virtù della sua linearità l'equazione consente d'essere trattata matematicamente in maniera agevole; essa, poi, costituisce talora una buona approssimazione di equazioni non lineari. Si noti, inoltre, che nel caso di distribuzione normale bivariata l'equazione di regressione risulta essere proprio lineare, come si evince dalla (2.42).

Quando l'equazione di regressione è lineare i coefficienti di regressione a e b si possono esprimere mediante i valori medi $\mu_X = E(X)$, $\mu_Y = E(Y)$, le deviazioni standard $\sigma_X = D(X)$, $\sigma_Y = D(Y)$ ed il coefficiente di correlazione

$$\rho = \frac{\text{cov}(X, Y)}{D(X)D(Y)}.$$

Sussiste infatti il seguente teorema:

Teorema 6.1.1 Se l'equazione di regressione di Y data X è lineare si ha:

$$E(Y|x) = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X). \quad (6.14)$$

Dim. Ci limitiamo ad effettuare la dimostrazione nel caso continuo; il caso discreto si tratta in maniera perfettamente analoga. Osserviamo anzitutto che se l'equazione di regressione è lineare sussiste la (6.13), e quindi dalla (6.1) si ha:

$$\int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy = a + b x. \quad (6.15)$$

Moltiplicando ambo i membri della (6.15) per la densità di probabilità $f_X(x)$ e integrando rispetto ad x si ottiene:

$$\begin{aligned} & \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} y f_{Y|X}(y|x) f_X(x) dy \\ &= a \int_{-\infty}^{\infty} f_X(x) dx + b \int_{-\infty}^{\infty} x f_X(x) dx. \end{aligned}$$

Esprimendo il prodotto $f_{Y|X}(y|x) f_X(x)$ tramite la (6.2), si ha poi:

$$E(Y) = a + b E(X). \quad (6.16)$$

Se, invece, si moltiplicano ambo i membri della (6.15) per $x f_X(x)$ e si integra rispetto ad x , si perviene all'equazione

$$\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} xy f_{X,Y}(x,y) dy = a \int_{-\infty}^{\infty} x f_X(x) dx + b \int_{-\infty}^{\infty} x^2 f_X(x) dx,$$

ossia:

$$E(XY) = aE(X) + bE(X^2). \quad (6.17)$$

Risolvendo il sistema costituito dalle equazioni (6.16) e (6.17) si determinano poi i coefficienti di regressione a e b :

$$a = E(Y) - bE(X) = E(Y) - \rho \frac{D(Y)}{D(X)} E(X), \quad (6.18)$$

$$b = \frac{E(XY) - E(X)E(Y)}{D^2(X)} = \frac{\text{cov}(X,Y)}{D^2(X)} = \rho \frac{D(Y)}{D(X)} \quad (6.19)$$

Con la sostituzione delle espressioni (6.18) e (6.19) di a e b nell'equazione di regressione (6.13) segue infine la (6.14).

È interessante osservare che se l'equazione di regressione è lineare e se le variabili X e Y sono scorrelate, la retta di regressione è parallela all'asse delle ascisse. Se X e Y sono scorrelate si ha infatti $\text{cov}(X,Y) = 0$, così che il coefficiente di correlazione ρ è nullo. In tal caso dall'equazione (6.14) si trae $E(Y|x) = \mu_Y$.

Si noti che il coefficiente di correlazione riveste un ruolo importante nell'ambito dell'analisi statistica di variabili casuali congiunte. Ricordando che è sempre $-1 \leq \rho \leq 1$, dal Teorema 6.1.1 segue che se l'equazione di regressione è lineare il segno di ρ determina il segno del coefficiente angolare della retta di regressione. Al variare di ρ la (6.14) rappresenta dunque l'equazione di un fascio di rette passanti per il punto di coordinate (μ_X, μ_Y) .

In problemi di regressione che coinvolgono più di due variabili casuali si parla di *regressione multivariata*. Nel caso di una variabile casuale Y condizionata da k variabili casuali, ad esempio, si è interessati al calcolo della media condizionata $E(Y|x_1, x_2, \dots, x_k)$, ossia al valore medio di Y dato che $X_1 = x_1, X_2 = x_2, \dots, X_k = x_k$.

Tabella 6.1: Età e pressioni sistoliche per un campione di 20 individui.

età	pressione	età	pressione
17	114	51	146
21	120	53	155
24	121	58	152
29	135	59	158
34	132	60	165
36	126	65	160
38	145	67	162
41	155	69	158
44	146	70	170
49	148	72	175

6.2 Approssimazione ai minimi quadrati

Nel paragrafo precedente abbiamo analizzato in particolare il problema della regressione bivariata nel caso di variabili casuali la cui distribuzione di probabilità congiunta è nota. Nelle applicazioni, però, ci si confronta di solito con problemi in cui questa è incognita: vengono, cioè, osservate sperimentalmente coppie di dati dalle quali si evince che l'equazione di regressione può essere di tipo lineare, pur ignorando la distribuzione congiunta delle variabili casuali che generano i dati osservati.

Esempio 6.2.1 Riferiamoci alla Tabella 6.1, in cui sono riportate le età x in anni e le pressioni sistoliche y in mmHg (millimetri di mercurio) di un campione composto da 20 individui.

Se si riportano su di un grafico bidimensionale le 20 coppie di dati che costituiscono la realizzazione del campione osservato, ci si rende conto che una linea retta ne può fornire un buon attagliamento (cfr. Figura 6.2). Sebbene, invero, i punti non siano rigorosamente allineati, la loro collocazione suggerisce che la pressione sistolica media di un individuo possa essere legata alla sua età mediante un'equazione del tipo $E(Y|x) = a + bx$, con un'opportuna scelta dei parametri a e b .

Il problema della stima dei coefficienti incogniti di un'equazione di regressione lineare può essere affrontato ricorrendo al *metodo o principio dei minimi quadrati*, proposto da A. Legendre nel secolo scorso, che passiamo a descrivere.

Supponiamo di aver osservato n coppie di dati (x_i, y_i) che costituiscono la realizzazione di un campione casuale estratto da una popolazione bidimensionale, avente cioè come variabile casuale genitrice un vettore casuale bidimensionale (X, Y) . Una volta stabilito che è ragionevole ritenere che l'equazione di regressione di Y dato $X = x$ sia lineare, ossia del tipo $E(Y|x) = a + bx$, si pone il problema di stimare i coefficienti a e b incogniti. Si devono pertanto determinare le stime \hat{a} e \hat{b} in modo che la retta di regressione $\hat{y} = \hat{a} + \hat{b}x$ stimata fornisca un attagliamento dei dati (x_i, y_i) che sia in qualche senso il migliore possibile. Affinché il problema abbia senso occorre evidentemente supporre che sia $n \geq 2$. Si noti che $n = 2$ è comunque un caso banale dato che per due punti distinti $(x_1, y_1), (x_2, y_2)$ passa un'unica

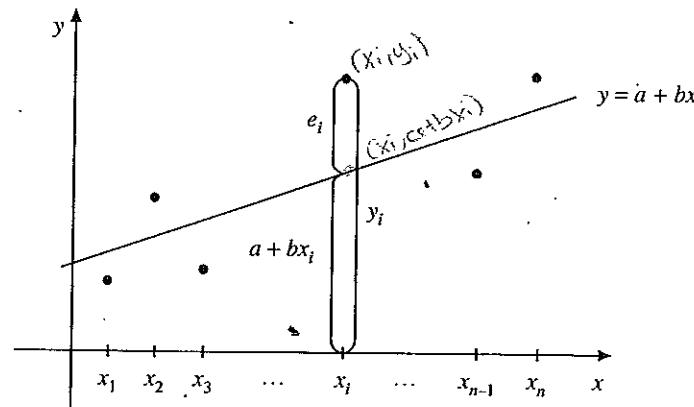


Figura 6.1: Retta di regressione.

retta; il problema è dunque significativo per $n \geq 3$ e quando le coppie (x_i, y_i) di dati non sono allineate. Indichiamo con

$$e_i = y_i - (a + bx_i) \quad (i = 1, 2, \dots, n).$$

la differenza tra il generico valore y_i osservato e l'ordinata $a + bx_i$ del punto corrispondente sulla retta di regressione $y = a + bx$, ossia il discostamento in verticale tra il generico punto di coordinate (x_i, y_i) e il punto corrispondente sulla retta di regressione, avente coordinate $(x_i, a + bx_i)$ (cfr. Figura 6.1). Il metodo dei minimi quadrati suggerisce di determinare le stime \hat{a} e \hat{b} minimizzando la somma dei quadrati dei discostamenti e_i . Si scelgono dunque come stime dei coefficienti di regressione i valori di a e b per i quali la funzione

$$Q(a, b) \stackrel{\text{def}}{=} \sum_{i=1}^n e_i^2 = \sum_{i=1}^n [y_i - (a + bx_i)]^2 \quad (6.20)$$

assume valore minimo, così che si abbia

$$Q(\hat{a}, \hat{b}) \leq Q(a, b) \quad \forall a, b \in \mathbb{R}.$$

Osserviamo che, per la presenza dei quadrati, la funzione $Q(a, b)$ tiene conto in egual misura sia dei discostamenti e_i positivi che di quelli negativi. Si noti poi che essa, per come è definita, possiede certamente un punto di minimo. Per determinarlo calcoliamo le derivate parziali della funzione (6.20) rispetto a a e b :

$$\frac{\partial}{\partial a} Q(a, b) = -2 \sum_{i=1}^n [y_i - (a + bx_i)]$$

$$\frac{\partial}{\partial b} Q(a, b) = -2 \sum_{i=1}^n x_i [y_i - (a + bx_i)].$$

Imponendone l'annullarsi si ricava il seguente sistema di equazioni algebriche nelle incognite a e b , dette *equazioni normali*:

$$an + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \quad (6.21)$$

$$a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \quad (6.22)$$

la cui soluzione fornisce i valori di a e b per i quali $Q(a, b)$ è minima. Si ottiene così la seguente stima del coefficiente di regressione a :

$$\hat{a} = \frac{1}{n} \left(\sum_{i=1}^n y_i - \hat{b} \sum_{i=1}^n x_i \right); \quad (6.23)$$

per il coefficiente b si ottiene invece la seguente stima:

$$\hat{b} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{j=1}^n y_j}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}. \quad (6.24)$$

Denotando con \bar{x} la media aritmetica dei valori x_1, x_2, \dots, x_n e con \bar{y} la media aritmetica dei valori y_1, y_2, \dots, y_n rispettivamente, le stime (6.23) e (6.24) possono più sinteticamente esprimersi al seguente modo:

$$\hat{a} = \bar{y} - \hat{b} \bar{x} \quad (6.25)$$

$$\hat{b} = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}. \quad (6.26)$$

Notiamo che, ad eccezione del caso banale $x_1 = x_2 = \dots = x_n$, la retta di regressione $\hat{y} = \hat{a} + \hat{b}x$ stimata esiste ed è unica in quanto il denominatore a secondo membro della (6.26) è positivo, avendosi

$$\begin{aligned} \sum_{i=1}^n x_i^2 - n \bar{x}^2 &= \sum_{i=1}^n x_i^2 - 2n \bar{x}^2 + n \bar{x}^2 \\ &= \sum_{i=1}^n (x_i^2 - 2x_i \bar{x} + \bar{x}^2) \\ &= \sum_{i=1}^n (x_i - \bar{x})^2. \end{aligned} \quad (6.27)$$

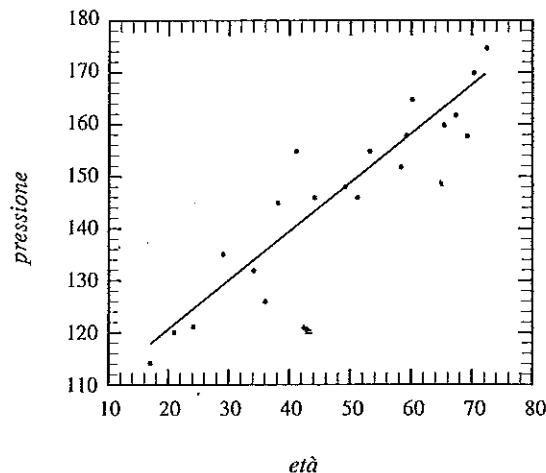


Figura 6.2: Una interpolazione lineare dei dati di cui alla Tabella 6.1.

Esempio 6.2.2 Calcoliamo le stime \hat{a} e \hat{b} relative all'Esempio 6.2.1. Facendo uso della Tabella 6.1, si ottiene:

$$\sum_i x_i = 957, \quad \sum_i x_i^2 = 51475, \quad \sum_i y_i = 2943, \quad \sum_i x_i y_i = 146197.$$

Sostituendo tali valori nelle (6.23) e (6.24) si ricavano le stime \hat{a} e \hat{b} :

$$\begin{aligned} \hat{a} &= \frac{1}{20} (2943 - 0.9457 \cdot 957) = 101.8943 \\ \hat{b} &= \frac{20 \cdot 146197 - 957 \cdot 2943}{20 \cdot 51475 - (957)^2} = 0.9457. \end{aligned}$$

La retta di regressione stimata, mostrata in Figura 6.2, è quindi

$$\hat{y} = 101.8943 + 0.9457x.$$

Ne segue, ad esempio, che la stima della pressione media per individui di età $x = 50$ (anni) è $\hat{y} = 101.8943 + 0.9457 \cdot 50 \approx 149$ (mmHg). ◆

Esaminiamo un altro esempio di applicazione del metodo dei minimi quadrati.

Esempio 6.2.3 Consideriamo la Tabella 6.2 relativa ad un campione di 16 studenti nella quale sono riportati i numeri x dei giorni di studio e i voti y ottenuti in sede d'esame. Dal grafico del campione (cfr. Figura 6.3) si evince che anche in questo caso una retta può fornire

Tabella 6.2: Giorni di studio e voti in un campione di 16 studenti.

giorni x	voti y	giorni x	voti y
24	20	37	25
26	23	39	27
28	22	41	27
29	24	43	28
31	26	45	30
32	25	48	28
35	27	53	30
36	26	56	30

un buon attagliamento dei dati. I punti indicati nel grafico non sono allineati, ma si presentano in una maniera che suggerisce che la votazione media di uno studente può dipendere linearmente dal numero di giorni di studio, ossia che il legame tra le due variabili casuali, voti Y e numero di giorni di studio X , può essere espresso da un'equazione di regressione lineare, ossia da una relazione del tipo $E(Y|x) = a + bx$. Per calcolare le stime dei coefficienti di regressione osserviamo che si ha:

$$\sum_i x_i = 603, \quad \sum_i x_i^2 = 24077, \quad \sum_i y_i = 418, \quad \sum_i x_i y_i = 16130.$$

Sostituendo questi valori nelle (6.23) e (6.24) si ricava:

$$\begin{aligned} \hat{a} &= \frac{1}{16} (418 - 0.2786 \cdot 603) = 15.622 \\ \hat{b} &= \frac{16 \cdot 16130 - 603 \cdot 418}{16 \cdot 24077 - (603)^2} = 0.2786, \end{aligned}$$

da cui segue la retta di regressione stimata:

$$\hat{y} = 15.622 + 0.2786x.$$

Dal risultato ottenuto si ottiene, ad esempio, che è $\hat{y} \geq 28$ se e solo se si ha

$$x \geq \frac{28 - 15.622}{0.2786} = 44.4293;$$

conseguentemente, sulla base del campione considerato il numero x di giorni di studio necessari rhediamente per ottenere un voto non inferiore a 28 è pari a 45. ◆

Notiamo ora che se si pone

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2, \quad (6.28)$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}, \quad (6.29)$$

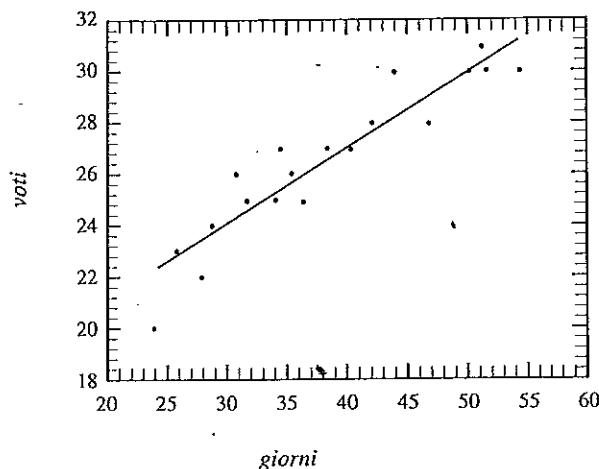


Figura 6.3: Una interpolazione lineare dei dati della Tabella 6.2.

le stime (6.23) e (6.24) si possono esprimere nel seguente modo:

$$\hat{a} = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}, \quad \hat{b} = \frac{S_{yy}}{S_{xx}} \quad (6.30)$$

dove, come al solito, \bar{x} e \bar{y} denotano le medie aritmetiche dei valori x_i ed y_i . La retta di regressione $\hat{y} = \hat{a} + \hat{b}x$ stimata assume allora la seguente forma:

$$\hat{y} = \bar{y} + \frac{S_{xy}}{S_{xx}} (x - \bar{x}). \quad (6.31)$$

Confrontando le equazioni (6.14) e (6.31) della retta di regressione e della sua stima appare evidente l'analogia tra i termini coinvolti. Infatti, \bar{y} e \bar{x} sono stime delle medie μ_Y e μ_X , mentre il rapporto S_{xy}/S_{xx} costituisce la stima di

$$\rho \frac{\sigma_Y}{\sigma_X} \equiv \frac{\text{cov}(X, Y)}{D^2(X)}.$$

Esporremo ora un'interessante interpretazione geometrica del metodo dei minimi quadrati. Una volta osservate le coppie di dati (x_i, y_i) , ($i = 1, 2, \dots, n$), con $n \geq 3$, come si è detto il problema consiste nell'individuare i valori di a e b per i quali risulti

$$a + b x_i = y_i \quad (i = 1, 2, \dots, n).$$

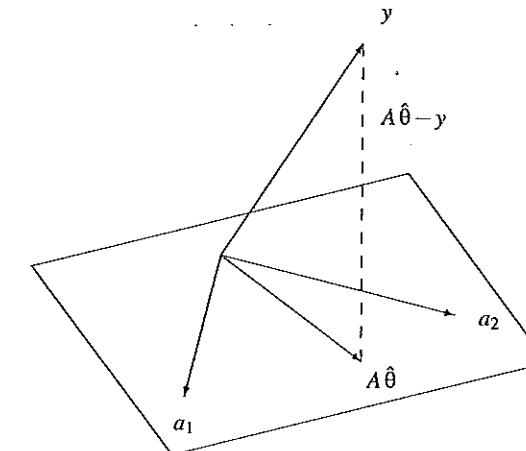


Figura 6.4: Interpretazione geometrica del metodo dei minimi quadrati.

Questo sistema di equazioni può essere rappresentato in forma matriciale come $A \theta = y$, dove

$$A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \theta = \begin{bmatrix} a \\ b \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}. \quad (6.32)$$

Consideriamo lo spazio vettoriale generato dai vettori colonna costituenti la matrice A , che indicheremo con a_1 e a_2 :

$$a_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad a_2 = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Con l'eccezione del caso banale $x_1 = x_2 = \dots = x_n$, tale spazio vettoriale ha dimensione 2. Come è noto, il sistema $A \theta = y$ ammette soluzione se il vettore y è esprimibile come combinazione lineare di a_1 e a_2 ; ciò comporta che y appartiene allo spazio delle colonne di A , ossia che i punti di coordinate (x_i, y_i) sono allineati, il che, tuttavia, di norma non accade quando si ha a che fare con dati effettivi. Ci si pone allora il problema di determinare una soluzione approssimata θ che renda minimo l'errore $\|A \theta - y\|$ rappresentante la distanza tra y e il punto $A \theta$ nello spazio delle colonne di A . Occorre quindi ricercare il punto $A \hat{\theta}$ situato a distanza minima da y rispetto a qualunque altro punto nello spazio delle colonne. Il vettore errore $A \hat{\theta} - y$ deve dunque identificarsi con la proiezione di y sullo spazio delle colonne, e il vettore errore $A \hat{\theta} - y$ deve essere perpendicolare a tale spazio, come mostrato nella Figura 6.4. Questa

condizione di perpendicolarità si esprime imponendo che per ogni scelta di

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix},$$

il vettore $A\mathbf{z}$, che si trova nello spazio delle colonne, deve essere perpendicolare al vettore $A\hat{\theta} - \mathbf{y}$. Quindi deve essere:

$$(\mathbf{A}\mathbf{z})^T (\mathbf{A}\hat{\theta} - \mathbf{y}) = 0,$$

ovvero:

$$\mathbf{z}^T (\mathbf{A}^T \mathbf{A} \hat{\theta} - \mathbf{A}^T \mathbf{y}) = 0. \quad (6.33)$$

Dovendo la (6.33) sussistere per ogni \mathbf{z} , il vettore $\mathbf{A}^T \mathbf{A} \hat{\theta} - \mathbf{A}^T \mathbf{y}$ deve essere nullo. Si ricava così il sistema

$$\mathbf{A}^T \mathbf{A} \hat{\theta} = \mathbf{A}^T \mathbf{y}, \quad (6.34)$$

cui $\hat{\theta}$ deve soddisfare per essere la migliore soluzione approssimata del sistema $\mathbf{A} \hat{\theta} = \mathbf{y}$. Se le colonne di \mathbf{A} sono indipendenti, ossia se non si verifica $x_1 = x_2 = \dots = x_n$, la matrice quadrata $\mathbf{A}^T \mathbf{A}$ è invertibile, e quindi il sistema (6.34) ammette un'unica soluzione. Notando che risulta

$$\mathbf{A}^T \mathbf{A} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \quad \mathbf{A}^T \mathbf{y} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix},$$

si conclude che il sistema (6.34) coincide con quello delle equazioni normali del metodo dei minimi quadrati. La soluzione

$$\hat{\theta} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix},$$

ottenuta con considerazioni geometriche è quindi la stessa cui si perviene usando il metodo dei minimi quadrati (cfr. le (6.23) e (6.24)).

Una volta osservate le coppie di dati (x_i, y_i) e stabilito che l'equazione di regressione di Y dato $X = x$ è da assumersi lineare, può talora accadere che in luogo di $E(Y|x) = a + bx$ sia preferibile fare ricorso all'equazione $E(Y|x) = bx$ per la quale va stimato soltanto il parametro b incognito. Il procedimento che si segue è analogo a quello adottato nel caso lineare a due parametri. Così, applicando il metodo dei minimi quadrati, si determina la stima \hat{b} minimizzando la somma dei quadrati dei discostamenti $e_i = y_i - bx_i$, ossia scegliendo come stima di b i valori per i quali la funzione

$$Q(b) \stackrel{\text{def}}{=} \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - bx_i)^2$$

assume valore minimo. Imponendo che la derivata

$$\frac{d}{db} Q(b) = -2 \sum_{i=1}^n x_i (y_i - bx_i)$$

Tabella 6.3: Distanze percorse e relativi costi per un campione di 18 automobili.

Km x	costo y	Km x	costo y
393.8	53	442.0	62
484.4	64	507.7	70
403.3	60	433.6	58
526.5	67	387.1	55
446.4	55	433.5	62
397.1	52	478.4	60
434.2	60	435.2	63
458.3	65	351.4	46
477.9	63	399.7	55

si annulli, si ricava il valore di b per il quale $Q(b)$ è minima. Si ottiene così la seguente stima di b :

$$\hat{b} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}. \quad (6.35)$$

Si noti che, ad eccezione del caso banale $x_1 = x_2 = \dots = x_n = 0$, la retta di regressione $\hat{y} = \hat{b}x$ stimata esiste ed è unica in quanto il denominatore a secondo membro della (6.35) è positivo.

Esaminiamo un esempio di applicazione del metodo dei minimi quadrati al caso di retta di regressione stimata $\hat{y} = \hat{b}x$.

Esempio 6.2.4 Prendiamo in esame la Tabella 6.3 relativa ad un campione costituito da 18 automobili dello stesso tipo; in essa è riportata la distanza x percorsa da ciascuna di esse espressa in chilometri in un test di consumo e il costo y corrispondente in euro. I dati riscontrati, il cui grafico è mostrato nella Figura 6.5, possono verosimilmente essere ben approssimati da una retta. È però evidente che nell'ipotizzare una retta di regressione di equazione $y = a + bx$ occorre fissare $a = 0$ in quanto ad un percorso x nullo deve necessariamente corrispondere un costo y nullo. La retta di regressione stimata è dunque del tipo $\hat{y} = \hat{b}x$. Facendo uso della (6.35), dai dati disponibili si ricava facilmente la stima $\hat{b} = 0.13539$, così che risulta

$$\hat{y} = 0.13539x.$$

Tale equazione, il cui grafico è tracciato nella Figura 6.5, può essere dunque utilizzata per effettuare previsioni del consumo di automobili del tipo esaminato in relazione alla distanza percorsa. ♦

È appena il caso di rilevare che nei casi in cui si adopera l'equazione di regressione $E(Y|x) = a$ il metodo dei minimi quadrati fornisce come stima di a la media aritmetica \bar{y} dei

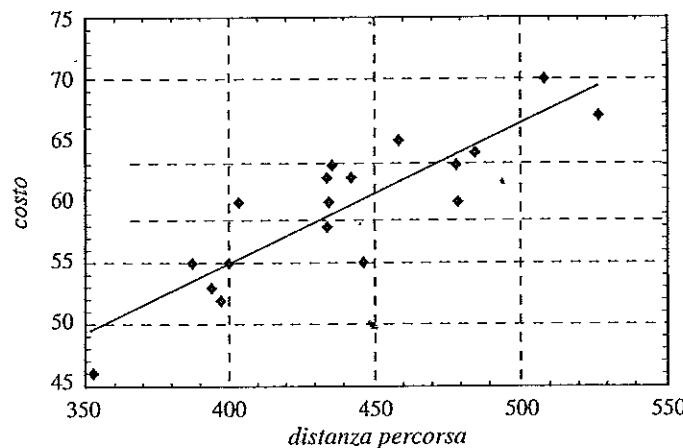


Figura 6.5: Interpolazione lineare dei costi in relazione alle distanze percorse.

valori y_i . Infatti, in questo caso la funzione

$$Q(a) \stackrel{\text{def}}{=} \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a)^2$$

attiene il suo minimo nel punto

$$\hat{a} = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}. \quad (6.36)$$

6.3 Regressione non lineare

Nel paragrafo precedente abbiamo visto che il problema della regressione bivariata può essere affrontato facendo ricorso al metodo dei minimi quadrati che consente di pervenire a stime puntuale dei parametri incogniti di un'equazione di regressione di tipo lineare. Mostriremo ora come tale metodo possa essere utilizzato anche nel caso di talune equazioni di regressione di tipo non lineare.

Supponiamo che (x_i, v_i) , $(i = 1, 2, \dots, n)$, siano coppie di dati osservati, ed assumiamo che l'equazione di regressione di V dato $X = x$ sia di tipo esponenziale:

$$E(V|x) = cd^x \quad (6.37)$$

con c e d costanti positive. Come nel caso lineare, ci poniamo il problema di ricavare una stima puntuale dei parametri c e d incogniti. Si devono dunque determinare le stime \hat{c} e \hat{d} in

modo che la curva di regressione stimata,

$$\hat{v} = \hat{c}d^{\hat{x}}, \quad (6.38)$$

fornisca un buon attagliamento dei dati (x_i, v_i) . Osserviamo che dalla (6.38) si ottiene l'equazione

$$\ln \hat{v} = \ln \hat{c} + x \ln \hat{d} \quad (6.39)$$

che, posto $\hat{y} = \ln \hat{v}$, $\hat{a} = \ln \hat{c}$ e $\hat{b} = \ln \hat{d}$, coincide con la retta di regressione stimata $\hat{y} = \hat{a} + \hat{b}x$ della regressione lineare. Il problema dell'individuazione delle stime di c e d nel modello esponenziale è stato dunque ricondotto a quello dell'individuazione delle stime di $a = \ln c$ e di $b = \ln d$ in un modello lineare. Si possono pertanto usare direttamente i risultati (6.23) e (6.24) dopo avervi posto $y_i = \ln v_i$ per $i = 1, 2, \dots, n$. Una volta individuati i valori \hat{a} e \hat{b} , applicando le trasformazioni inverse $\hat{c} = e^{\hat{a}}$ e $\hat{d} = e^{\hat{b}}$ si perviene infine ai coefficienti della curva di regressione stimata (6.38).

Esaminiamo un esempio di applicazione del procedimento appena illustrato.

Esempio 6.3.1 Un'industria lancia sul mercato un nuovo tipo di integratore alimentare. Nel corso delle prime 6 settimane di vendita i numeri (in migliaia) v_1, v_2, \dots, v_6 di confezioni vendute sono i seguenti:

$$v_1 = 8, \quad v_2 = 13, \quad v_3 = 17, \quad v_4 = 25, \quad v_5 = 41, \quad v_6 = 59.$$

Assumendo che l'equazione di regressione per il numero V di confezioni vendute nella settimana x sia esponenziale, ossia data dalla (6.37), si desidera determinare la stima ai minimi quadrati dei coefficienti incogniti c e d . Procedendo come visto poc'anzi, trasformiamo i parametri ponendo $a = \ln c$, $b = \ln d$ e $y_i = \ln v_i$ per $i = 1, 2, \dots, 6$. Si ottiene così:

$$\begin{aligned} y_1 &= 2.079, & y_2 &= 2.565, & y_3 &= 2.833, \\ y_4 &= 3.219, & y_5 &= 3.714, & y_6 &= 4.078. \end{aligned}$$

Per $x_i = i$ ($i = 1, 2, \dots, 6$), si ricava:

$$\sum_i x_i = 21, \quad \sum_i x_i^2 = 91, \quad \sum_i y_i = 18.488, \quad \sum_i x_i y_i = 71.622.$$

Il modello esponenziale è stato trasformato in un modello di regressione lineare per la cui analisi possiamo utilizzare il metodo dei minimi quadrati. Facendo uso delle (6.23) e (6.24), con $x_i = i$ ($i = 1, 2, \dots, 6$), si ottengono le seguenti stime di a e b :

$$\begin{aligned} \hat{a} &= \ln \hat{c} = \frac{1}{6} (18.488 - 0.395 \cdot 21) = 1.699 \\ \hat{b} &= \ln \hat{d} = \frac{6 \cdot 71.622 - 21 \cdot 18.488}{6 \cdot 91 - (21)^2} = 0.395, \end{aligned}$$

da cui segue

$$\hat{c} = e^{\hat{a}} = e^{1.699} = 5.468, \quad \hat{d} = e^{\hat{b}} = e^{0.395} = 1.484.$$

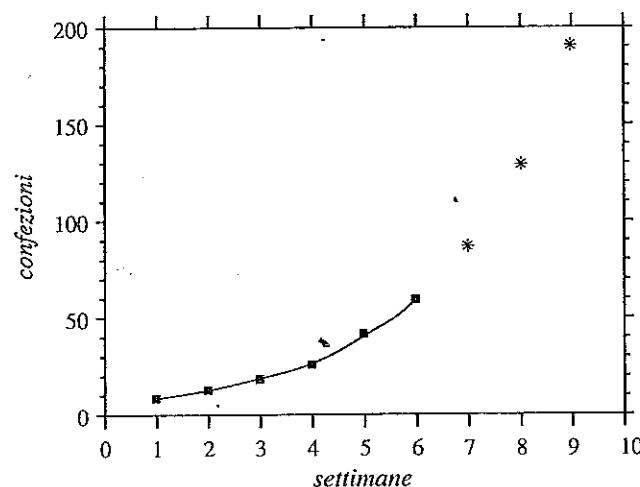


Figura 6.6: Interpolazione dei dati dell'Esempio 6.3.1. I quadrati pieni indicano valori osservati e gli asterischi valori previsti.

La curva di regressione stimata è dunque:

$$\hat{v} = 5.468(1.484)^x.$$

Il risultato ottenuto può essere utilizzato per effettuare previsioni sul numero di confezioni che saranno vendute nelle settimane successive (v. Figura 6.6). Così, se non vi sono elementi che facciano ritenere che esistano delle significative variazioni nell'andamento delle vendite, i numeri di migliaia di confezioni che si stima verranno vendute nelle successive settimane sono ad esempio

$$\hat{v}_7 = 5.468(1.484)^7 = 86.669$$

$$\hat{v}_8 = 5.468(1.484)^8 = 128.617$$

$$\hat{v}_9 = 5.468(1.484)^9 = 190.868.$$

Esamineremo ora un altro tipo di equazione di regressione non lineare che può essere studiato mediante il metodo dei minimi quadrati. Consideriamo a tal fine delle coppie di dati (u_i, v_i) , $(i = 1, 2, \dots, n)$, per le quali l'equazione di regressione di V dato $U = u$ è del tipo

$$E(V|u) = cu^b, \quad (6.40)$$

dove c è una costante positiva e b è una costante qualsiasi, entrambe da determinarsi. Volendo fornire una stima, occorre determinare i valori \hat{c} e \hat{b} in corrispondenza dei quali la curva di

regressione stimata

$$\hat{v} = \hat{c}u^{\hat{b}} \quad (6.41)$$

fornisce un buon attagliamento dei dati. Passando ai logaritmi l'equazione (6.41) diviene:

$$\ln \hat{v} = \ln \hat{c} + \hat{b} \ln u, \quad (6.42)$$

Notiamo che se si pone $\hat{y} = \ln \hat{v}$, $x = \ln u$ e $\hat{a} = \ln \hat{c}$, la (6.42) coincide con l'espressione $\hat{y} = \hat{a} + \hat{b}x$ relativa al caso di regressione lineare. Per individuare le stime di c e b nel modello (6.40) è dunque sufficiente individuare delle stime di $a = \ln c$ e di b in un modello lineare. A tal fine si fa ricorso alle (6.23) e (6.24) dopo avervi posto $y_i = \ln v_i$ e $x_i = \ln u_i$ per $i = 1, 2, \dots, n$. In tal modo, una volta individuati i valori \hat{a} e \hat{b} , essendo $\hat{c} = e^{\hat{a}}$ si ottengono i coefficienti della curva di regressione stimata (6.41). Il procedimento qui descritto verrà ora utilizzato nell'esempio che segue.

Esempio 6.3.2 In una fabbrica vengono prodotti microprocessori il cui costo di produzione varia in funzione del numero di unità prodotte. Se v_i rappresenta il costo unitario in euro dei microprocessori in corrispondenza di lotti costituiti da u_i unità, supponiamo che risulti:

$$u_1 = 10, \quad u_2 = 20, \quad u_3 = 50, \quad u_4 = 100, \quad u_5 = 200, \quad u_6 = 500,$$

$$v_1 = 1200, \quad v_2 = 550, \quad v_3 = 300, \quad v_4 = 125, \quad v_5 = 80, \quad v_6 = 45.$$

Assumiamo inoltre che la (6.40) sia l'equazione di regressione relativa al costo unitario V di microprocessori appartenenti a lotti costituiti da x unità. Per determinare la stima ai minimi quadrati dei coefficienti incogniti c e b , trasformiamo il parametro c ponendo $a = \ln c$. Posto $x_i = \ln u_i$ per $i = 1, 2, \dots, 6$ otteniamo poi

$$x_1 = 2.3026, \quad x_2 = 2.9957, \quad x_3 = 3.9120,$$

$$x_4 = 4.6052, \quad x_5 = 5.2983, \quad x_6 = 6.2146,$$

mentre posto $y_i = \ln v_i$ per $i = 1, 2, \dots, 6$ abbiamo

$$y_1 = 7.0901, \quad y_2 = 6.3099, \quad y_3 = 5.7038,$$

$$y_4 = 4.8283, \quad y_5 = 4.3820, \quad y_6 = 3.8067.$$

Ricaviamo pertanto:

$$\sum_i x_i = 25.328, \quad \sum_i x_i^2 = 117.481,$$

$$\sum_i y_i = 32.121, \quad \sum_i x_i y_i = 126.651.$$

L'equazione di regressione $v = cu^b$ è stata trasformata nell'equazione lineare $y = a + bx$ la cui specificazione può essere effettuata ricorrendo al metodo dei minimi quadrati. Invero, usando la relazione (6.24) si ottiene

$$\hat{b} = \frac{6 \cdot 126.651 - 25.328 \cdot 32.121}{6 \cdot 117.481 - (25.328)^2} = -0.8466,$$

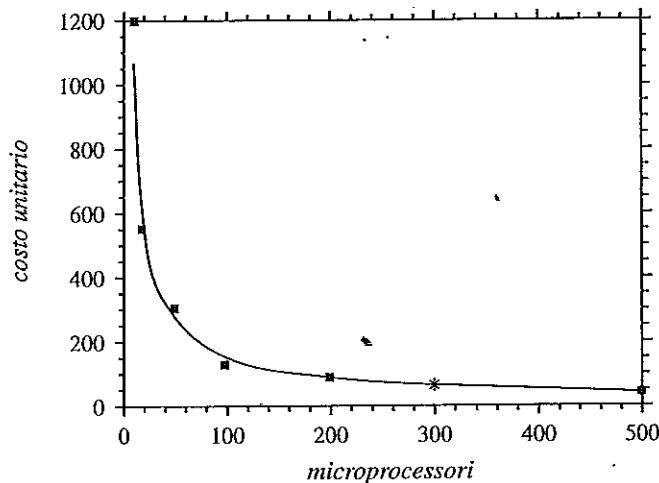


Figura 6.7: Interpolazione dei dati dell'Esempio 6.3.2. I quadrati pieni indicano i valori osservati e l'asterisco il valore previsto.

mentre dalla (6.23) segue

$$\hat{a} = \ln \hat{c} = \frac{1}{6} (32.121 + 0.8466 \cdot 25.328) = 8.9273,$$

da cui si ricava

$$\hat{c} = e^{\hat{a}} = e^{8.9273} = 7534.89.$$

La curva di regressione stimata (v. Figura 6.7) è dunque:

$$\hat{y} = 7534.89 u^{-0.8466}.$$

Questo risultato può essere utilizzato per stimare i costi unitari per lotti di preassegnate dimensioni. Ad esempio per il costo unitario (in euro) di microprocessori costituenti un lotto di $u = 300$ unità si ottiene la seguente stima:

$$\hat{v} = 7534.89 (300)^{-0.8466} = 60.248.$$

6.4 Stime puntuale

Nel contesto della regressione lineare le coppie (x_i, y_i) , $(i = 1, 2, \dots, n)$, osservate possono essere riguardate come la realizzazione di un campione casuale la cui variabile genitrice

bidimensionale (X, Y) possiede componenti X e Y caratterizzate dalla relazione

$$Y = a + bX + Z, \quad (6.43)$$

dove a e b sono parametri mentre Z è una variabile casuale di media nulla e varianza σ^2 . In tal caso l'equazione di regressione di Y dato $X = x$ è lineare, avendosi $E(Y|x) = a + bx$. Il modello (6.43) si presta ad una interessante interpretazione. Per le coppie osservate (x_i, y_i) si può ad esempio supporre che esista una relazione di tipo causa-effetto tra i valori x_i ed i corrispondenti y_i . In tal caso si assume che x_1, x_2, \dots, x_n siano delle costanti mentre l' n -upla (y_1, y_2, \dots, y_n) viene riguardata come la realizzazione di un vettore casuale (Y_1, Y_2, \dots, Y_n) in cui per ogni i il valore y_i assunto dalla variabile Y_i dipende da x_i . In altri termini, si suppone che sussistano le seguenti relazioni:

$$Y_i = a + b x_i + Z_i \quad (i = 1, 2, \dots, n), \quad (6.44)$$

dove le variabili casuali Z_1, Z_2, \dots, Z_n sono indipendenti e identicamente distribuite con media nulla e varianza σ^2 . Le variabili casuali Y_1, Y_2, \dots, Y_n sono pertanto indipendenti ed hanno medie e varianze

$$E(Y_i) = a + b x_i, \quad D^2(Y_i) = \sigma^2 \quad (i = 1, 2, \dots, n). \quad (6.45)$$

Vedremo ora come risulti possibile ricavare degli stimatori dei parametri a , b e σ^2 .

Anzitutto osserviamo che, come mostrato dalla prima delle (6.45), per ogni i la media di Y_i è lineare in x_i . Pertanto, si può applicare il metodo dei minimi quadrati per stimare a e b . Così facendo, in analogia con le stime (6.25) e (6.26) si ottengono i seguenti stimatori dei parametri a e b :

$$\hat{A} = \bar{Y} - \hat{B} \bar{x} \quad (6.46)$$

$$\hat{B} = \frac{\sum_{i=1}^n x_i Y_i - n \bar{x} \bar{Y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}, \quad (6.47)$$

dove si è posto $\bar{Y} = (\sum_{i=1}^n Y_i)/n$ e dove \bar{x} denota la media aritmetica dei valori x_1, x_2, \dots, x_n .

Si noti che gli stimatori (6.46) e (6.47) sono entrambi funzioni lineari delle variabili casuali Y_1, Y_2, \dots, Y_n . Calcoliamone valore medio, varianza e covarianza con l'intento di mostrare, in particolare, che \hat{A} e \hat{B} sono stimatori corretti dei parametri a e b .

Proposizione 6.4.1 *Valore medio, varianza e covarianza degli stimatori \hat{A} e \hat{B} sono rispettivamente dati da*

$$E(\hat{A}) = a \quad D^2(\hat{A}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \quad (6.48)$$

$$E(\hat{B}) = b \quad D^2(\hat{B}) = \frac{\sigma^2}{S_{xx}} \quad (6.49)$$

$$\text{cov}(\hat{A}, \hat{B}) = -\bar{x} \frac{\sigma^2}{S_{xx}}.$$

Dim. Calcoliamo anzitutto valore medio e varianza di \hat{B} . Dalle (6.28) e (6.47) si trae:

$$\hat{B} = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) Y_i. \quad (6.50)$$

Pertanto, ricordando la (6.28) e la (6.45), dalla (6.50) segue:

$$\begin{aligned} E(\hat{B}) &= \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) E(Y_i) = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x})(a + b x_i) \\ &= \frac{a}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) + \frac{b}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) x_i = b, \end{aligned}$$

mentre per la varianza di \hat{B} si ottiene

$$\begin{aligned} D^2(\hat{B}) &= \frac{1}{S_{xx}^2} \sum_{i=1}^n (x_i - \bar{x})^2 D^2(Y_i) \\ &= \frac{1}{S_{xx}^2} \sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2 = \frac{\sigma^2}{S_{xx}}. \end{aligned}$$

Calcoliamo ora valore medio e varianza di \hat{A} . Dalla (6.46) e dalla prima delle (6.49) si ricava:

$$E(\hat{A}) = E(\bar{Y} - \hat{B}\bar{x}) = E(\bar{Y}) - b\bar{x}.$$

Da questa, poiché dalla prima delle (6.45) risulta

$$E(\bar{Y}) = \frac{1}{n} \sum_{i=1}^n E(Y_i) = \frac{1}{n} \sum_{i=1}^n (a + b x_i) = a + b\bar{x},$$

si ottiene in definitiva:

$$E(\hat{A}) = a + b\bar{x} - b\bar{x} = a,$$

ossia la prima delle (6.48). Per ricavare la seconda delle (6.48) osserviamo che risulta:¹

$$D^2(\hat{A}) = D^2(\bar{Y} - \hat{B}\bar{x}) = D^2(\bar{Y}) + \bar{x}^2 D^2(\hat{B}) - 2\bar{x} \operatorname{cov}(\bar{Y}, \hat{B}).$$

Pertanto facendo uso delle relazioni

$$\begin{aligned} D^2(\bar{Y}) &= \frac{1}{n^2} \sum_{i=1}^n D^2(Y_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n} \\ \operatorname{cov}(\bar{Y}, \hat{B}) &= \operatorname{cov}\left[\frac{1}{n} \sum_{j=1}^n Y_j, \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) Y_i\right] \\ &= \frac{1}{n S_{xx}} \sum_{i,j=1}^n (x_i - \bar{x}) \operatorname{cov}(Y_i, Y_j) \\ &= \frac{\sigma^2}{n S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) = 0 \end{aligned} \quad (6.51)$$

¹Si ricordi che la varianza di una combinazione lineare $aX + bY$ di due variabili casuali X e Y ha la seguente espressione: $D^2(aX + bY) = a^2 D^2(X) + b^2 D^2(Y) + 2ab \operatorname{cov}(X, Y)$.

6.4. STIME PUNTUALI

e della seconda delle (6.49) segue:

$$D^2(\hat{A}) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right).$$

Resta così dimostrata la seconda delle (6.48). Infine, notando che in generale risulta

$$\operatorname{cov}(aX + bY, Z) = a \operatorname{cov}(X, Z) + b \operatorname{cov}(Y, Z),$$

dalla (6.46) si trae

$$\operatorname{cov}(\hat{A}, \hat{B}) = \operatorname{cov}(\bar{Y} - \hat{B}\bar{x}, \hat{B}) = \operatorname{cov}(\bar{Y}, \hat{B}) - \bar{x} D^2(\hat{B}). \quad (6.52)$$

Ricordando la (6.51) e la seconda delle (6.49), dalla (6.52) si ricava infine:

$$\operatorname{cov}(\hat{A}, \hat{B}) = -\bar{x} \frac{\sigma^2}{S_{xx}}.$$

La dimostrazione è così completata. ■

Nella Proposizione 6.4.1 è mostrato, tra l'altro, che le varianze degli stimatori \hat{A} e \hat{B} dipendono in maniera inversa da S_{xx} che, a sua volta, dipende dalla taglia n del campione. Si vede facilmente che se S_{xx} risultasse divergente con n , gli stimatori \hat{A} e \hat{B} sarebbero consistenti. Ciò, tuttavia, può non accadere, ossia può verificarsi che S_{xx} non diventi infinitamente grande con la taglia del campione. La proposizione che segue afferma comunque che S_{xx} non risulta mai decrescente in n , e fornisce delle condizioni sufficienti affinché S_{xx} diverga con n .

Proposizione 6.4.2 *Sussistono i seguenti risultati:*

- (i) S_{xx} è non decrescente in n ;
- (ii) se risulta $x_1 \leq x_2 \leq x_3 \leq \dots$ e se esistono una successione estratta $\{x_{j_k}\}_k$ e un reale $\varepsilon > 0$ tale da aversi $x_{j_{k+1}} - x_{j_k} > \varepsilon$ per ogni $k = 1, 2, \dots$, allora S_{xx} diverge per $n \rightarrow \infty$.

Dim. Per evidenziare che \bar{x} e S_{xx} dipendono da n , nel corso della dimostrazione verrà adottata la seguente notazione:

$$\bar{x}^{(n)} = \frac{1}{n} \sum_{i=1}^n x_i, \quad S_{xx}^{(n)} = \sum_{i=1}^n [x_i - \bar{x}^{(n)}]^2. \quad (6.53)$$

Per dimostrare la (i) osserviamo che dalla seconda delle (6.53) si trae:

$$\begin{aligned} S_{xx}^{(n+1)} - S_{xx}^{(n)} &= [x_{n+1} - \bar{x}^{(n+1)}]^2 \\ &\quad + \sum_{i=1}^n [x_i - \bar{x}^{(n+1)}]^2 - \sum_{i=1}^n [x_i - \bar{x}^{(n)}]^2. \end{aligned} \quad (6.54)$$

Notiamo che la differenza tra le due somme al secondo membro della (6.54) è non negativa in virtù della nota proprietà che la funzione

$$g(\alpha) = \sum_{i=1}^n (x_i - \alpha)^2$$

è minima per $\alpha = \sum_{i=1}^n x_i/n \equiv \bar{x}^{(n)}$. Si ha dunque $S_{xx}^{(n+1)} - S_{xx}^{(n)} \geq 0$, ossia $S_{xx}^{(n)}$ è non decrecente in n . La (ii) segue poi osservando anzitutto che risulta

$$\begin{aligned} S_{xx}^{(j_{k+1})} - S_{xx}^{(j_k)} &\equiv \sum_{i=1}^{j_{k+1}} [x_i - \bar{x}^{(j_{k+1})}]^2 - \sum_{i=1}^{j_k} [x_i - \bar{x}^{(j_k)}]^2 \\ &\geq \sum_{i=1}^{j_k} [x_i - \bar{x}^{(j_{k+1})}]^2 - \sum_{i=1}^{j_k} [x_i - \bar{x}^{(j_k)}]^2 \\ &= -2 [\bar{x}^{(j_{k+1})} - \bar{x}^{(j_k)}] \sum_{i=1}^{j_k} x_i + j_k \left\{ [\bar{x}^{(j_{k+1})}]^2 - [\bar{x}^{(j_k)}]^2 \right\} \\ &= -2 j_k \bar{x}^{(j_{k+1})} \bar{x}^{(j_k)} + 2 j_k [\bar{x}^{(j_k)}]^2 + j_k [\bar{x}^{(j_{k+1})}]^2 - j_k [\bar{x}^{(j_k)}]^2 \\ &= j_k [\bar{x}^{(j_{k+1})} - \bar{x}^{(j_k)}]^2. \end{aligned} \quad (6.55)$$

La successione $\{\bar{x}^{(j_k)}\}_k$ è crescente in k . Invero, dalla prima delle (6.53) si ha

$$\begin{aligned} \bar{x}^{(j_{k+1})} - \bar{x}^{(j_k)} &\equiv \frac{1}{j_{k+1}} \sum_{i=1}^{j_{k+1}} x_i - \frac{1}{j_k} \sum_{r=1}^{j_k} x_r \\ &= \frac{1}{j_{k+1} j_k} \left(\sum_{r=1}^{j_k} \sum_{i=1}^{j_{k+1}} x_i - \sum_{r=1}^{j_k} \sum_{i=1}^{j_{k+1}} x_r \right) \\ &= \frac{1}{j_{k+1} j_k} \left(\sum_{r=1}^{j_k} \sum_{i=j_k+1}^{j_{k+1}} x_i - \sum_{r=1}^{j_k} \sum_{i=j_k+1}^{j_{k+1}} x_r \right) \\ &= \frac{1}{j_{k+1} j_k} \sum_{r=1}^{j_k} \sum_{i=j_k+1}^{j_{k+1}} (x_i - x_r) \\ &> \frac{1}{j_{k+1} j_k} \varepsilon, \end{aligned} \quad (6.56)$$

dove la disegualanza è conseguenza dell'ipotesi $x_i - x_r \geq 0$ per $r < i$ e $x_i - x_r > \varepsilon$ per $r = j_k$ e $i = j_{k+1}$. Facendo uso della (6.56) nella (6.55), si ricava poi

$$S_{xx}^{(j_{k+1})} - S_{xx}^{(j_k)} > j_k \left(\frac{1}{j_{k+1} j_k} \varepsilon \right)^2 \stackrel{\text{def}}{=} h > 0, \quad (6.57)$$

da cui segue che $S_{xx}^{(j_k)}$ è crescente in k . Per iterazione sull'indice k , dalla (6.57) si ottiene

$$S_{xx}^{(j_{k+1})} > S_{xx}^{(j_1)} + kh \geq kh. \quad (6.58)$$

Dalla (6.58) segue infine $\lim_{k \rightarrow \infty} S_{xx}^{(j_k)} = \infty$, e quindi anche $\lim_{n \rightarrow \infty} S_{xx}^{(n)} = \infty$. \blacksquare

La (ii) della Proposizione 6.4.2 fornisce una condizione sufficiente affinché S_{xx} risulti divergente in n . Tale proprietà è molto significativa in quanto, in virtù della Proposizione 6.4.1,

6.4. STIME PUNTUALI

essa comporta che, purché \bar{x}^2 si mantenga limitato al crescere di n , sia \hat{A} che \hat{B} sono stimatori consistenti, oltre che corretti, dei parametri a e b . Vale infine la pena di sottolineare che l'ipotesi $x_1 \leq x_2 \leq x_3 \leq \dots$ nella (ii) della Proposizione 6.4.2 non è restrittiva potendo essere sempre soddisfatta mediante permutazioni degli elementi delle realizzazioni considerate.

Si noti che la stima di b assume un ruolo cruciale nella regressione lineare. Invero, l'essere ad esempio prossimo a 0 il valore stimato \hat{b} potrebbe essere indicazione che il valore esatto di b è proprio 0, e che quindi nell'equazione di regressione $E(Y|x) = a + bx$ manca la dipendenza da x . È allora auspicabile che lo stimatore \hat{B} fornisca una stima quanto più "affidabile" possibile, onde evitare interpretazioni erronee dell'equazione di regressione del modello considerato. Va infine osservato che le condizioni presenti nel punto (ii) della Proposizione 6.4.2 acquistano naturale rilievo soprattutto in applicazioni in cui x_1, x_2, \dots, x_n sono dei tempi, così che l'ordinamento crescente è automaticamente realizzato.

Passiamo ora ad esaminare uno stimatore della varianza σ^2 . Con riferimento all'equazione (6.44), ricordiamo che σ^2 è la varianza delle variabili $Z_i = Y_i - (a + bx_i)$, per $i = 1, 2, \dots, n$. Pertanto essa costituisce una misura dei discostamenti tra le variabili Y_i e i valori $a + bx_i$ corrispondenti sulla retta di regressione. Consideriamo la differenza

$$Y_i - \hat{Y}_i = Y_i - (\hat{A} + \hat{B}x_i) \quad (6.59)$$

tra la variabile Y_i che descrive il valore osservato in corrispondenza di x_i e la variabile $\hat{Y}_i = \hat{A} + \hat{B}x_i$ che descrive il corrispondente valore stimato. Ad essa si dà il nome di *residuo*. Mostreremo ora che la somma dei quadrati dei residui²

$$SS_R \stackrel{\text{def}}{=} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{A} - \hat{B}x_i)^2 \quad (6.60)$$

può essere utilizzata per costruire uno stimatore della varianza σ^2 . Sussiste, invero, la proposizione seguente:

Proposizione 6.4.3 *La statistica $SS_R/(n-2)$ è uno stimatore corretto della varianza σ^2 .*

Dim. Osserviamo che dalla (6.46) segue

$$\begin{aligned} SS_R &= \sum_{i=1}^n (Y_i - \hat{A} - \hat{B}x_i)^2 = \sum_{i=1}^n [Y_i - \bar{Y} - \hat{B}(x_i - \bar{x})]^2 \\ &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - 2\hat{B} \sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x}) + \hat{B}^2 \sum_{i=1}^n (x_i - \bar{x})^2. \end{aligned} \quad (6.61)$$

In virtù delle (6.28) e (6.50), dalla (6.61) si ottiene poi

$$SS_R = \sum_{i=1}^n Y_i^2 - n\bar{Y}^2 - S_{xx}\hat{B}^2. \quad (6.62)$$

²La notazione SS_R è reminiscente della terminologia inglese "sum of squares of residuals".

Facendo uso delle (6.45) e (6.49), dalla (6.50) si ricava la media della somma dei quadrati dei residui:

$$\begin{aligned} E(\text{SS}_R) &= E\left(\sum_{i=1}^n Y_i^2 - n\bar{Y}^2 - S_{xx}\hat{B}^2\right) \\ &= \sum_{i=1}^n \{D^2(Y_i) + [E(Y_i)]^2\} - n\{D^2(\bar{Y}) + [E(\bar{Y})]^2\} \\ &\quad - S_{xx}\{D^2(\hat{B}) + [E(\hat{B})]^2\} \\ &= n\sigma^2 + \sum_{i=1}^n (a + b x_i)^2 - n\left[\frac{\sigma^2}{n} + (a + b \bar{x})^2\right] \\ &\quad - S_{xx}\left(\frac{\sigma^2}{S_{xx}} + b^2\right) \\ &= (n-2)\sigma^2 - S_{xx}b^2 + \sum_{i=1}^n [(a + b x_i)^2 - (a + b \bar{x})^2] \\ &= (n-2)\sigma^2 - S_{xx}b^2 + b^2 \sum_{i=1}^n (x_i^2 - \bar{x}^2) = (n-2)\sigma^2, \end{aligned}$$

da cui si ricava

$$E\left(\frac{\text{SS}_R}{n-2}\right) = \sigma^2,$$

e quindi l'asserto. ■

Notiamo che nel caso in cui si utilizza $E(Y|x) = a$ quale modello di regressione, la retta di regressione stimata è $\hat{y} = \hat{a}$ che, in virtù della (6.36), diventa $\hat{y} = \bar{y}$. Pertanto, \bar{y} rappresenta la stima della costante a . Per questo modello i residui (6.59) diventano

$$Y_i - \hat{Y}_i = Y_i - \bar{Y},$$

così che

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 \stackrel{\text{def}}{=} S_{YY} \quad (6.63)$$

si identifica con la somma (6.60) dei quadrati dei residui.

6.5 Regressione normale

Come visto nel § 6.4, allorché si effettua l'analisi di regressione su coppie di dati (x_i, y_i) , ($i = 1, 2, \dots, n$) solitamente si assume che le x_i siano costanti mentre le y_i vengono interpretate come i valori assunti dalle variabili casuali Y_i che si è visto essere tra loro indipendenti.

In questo paragrafo verrà analizzato un caso particolare di regressione, la *regressione normale*, che si riferisce al caso in cui la variabile casuale Y_i condizionata da x_i è di tipo normale. Assumeremo, dunque, che per ogni valore x_i fissato la densità condizionata della variabile casuale corrispondente Y_i ha distribuzione normale di valore medio $a + b x_i$ e varianza σ^2 .

6.5. REGRESSIONE NORMALE

Nell'ambito del modello (6.44) ciò equivale a supporre che le variabili casuali Z_1, Z_2, \dots, Z_n hanno distribuzione normale di media nulla e varianza σ^2 . In queste ipotesi per $i = 1, 2, \dots, n$ la variabile Y_i condizionata da x_i ha densità di probabilità

$$f(y_i | x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{[y_i - (a + b x_i)]^2}{2\sigma^2}\right\}. \quad (6.64)$$

L'analisi della regressione normale consiste anzitutto nella stima puntuale dei parametri a , b e σ^2 a partire dalle coppie di dati osservati. Altro settore d'indagine è poi costituito dalla predizione basata sull'equazione di regressione $\hat{y} = \hat{a} + \hat{b}x$ stimata, dove \hat{a} e \hat{b} denotano le stime di a e b .

Per ottenere le stime di massima verosimiglianza dei parametri a , b e σ^2 osserviamo che alla (6.64) corrisponde la funzione di verosimiglianza

$$L(a, b, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n [y_i - (a + b x_i)]^2\right\}.$$

Per semplificare la ricerca del massimo di $L(a, b, \sigma^2)$, passiamo ai logaritmi naturali:

$$\ln L(a, b, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n [y_i - (a + b x_i)]^2. \quad (6.65)$$

Imponendo che si annullino le derivate parziali della (6.65) rispetto ad a , b e σ^2 si ricavano le equazioni normali:

$$\frac{\partial \ln L}{\partial a} = \frac{1}{\sigma^2} \sum_{i=1}^n [y_i - (a + b x_i)] = 0 \quad (6.66)$$

$$\frac{\partial \ln L}{\partial b} = \frac{1}{\sigma^2} \sum_{i=1}^n x_i [y_i - (a + b x_i)] = 0 \quad (6.67)$$

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n [y_i - (a + b x_i)]^2 = 0. \quad (6.68)$$

È interessante osservare che le (6.66) e (6.67) coincidono con le equazioni normali (6.21) e (6.22). Le stime di massima verosimiglianza \hat{a} e \hat{b} coincidono dunque con le stime ai minimi quadrati di a e b . Ricordando le (6.30) si ha quindi

$$\hat{a} = \bar{y} - \hat{b} \bar{x} \quad \hat{b} = \frac{S_{xy}}{S_{xx}}, \quad (6.69)$$

dove S_{xx} e S_{xy} sono definite nelle (6.28) e (6.29). È possibile ora determinare la stima $\hat{\sigma}^2$ della varianza σ^2 sostituendo a e b nell'equazione (6.68) con \hat{a} e \hat{b} . Si ottiene così:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n [y_i - (\hat{a} + \hat{b} x_i)]^2. \quad (6.70)$$

Nella pratica è però più conveniente utilizzare una formula alternativa per la stima $\hat{\sigma}^2$. Per ricavarla osserviamo che se si pone

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2, \quad (6.71)$$

facendo uso delle (6.69) si ottiene:

$$\begin{aligned} \sum_{i=1}^n [y_i - (\hat{a} + \hat{b}x_i)]^2 &= \sum_{i=1}^n (y_i - \bar{y} + \hat{b}\bar{x} - \hat{b}x_i)^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2\hat{b} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + \hat{b}^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= S_{yy} - 2\hat{b}S_{xy} + \hat{b}^2 S_{xx} \\ &= S_{yy} - \hat{b}S_{xy}. \end{aligned} \quad (6.72)$$

Dalle (6.70) e (6.72) segue infine la preannunciata espressione alternativa per la stima di σ^2 :

$$\hat{\sigma}^2 = \frac{S_{yy} - \hat{b}S_{xy}}{n}. \quad (6.73)$$

Confrontando la (6.70) con la (6.60) si conclude che lo stimatore di massima verosimiglianza di σ^2 è dato dalla statistica SS_R/n . Poiché nella Proposizione 6.4.3 è mostrato che la statistica $SS_R/(n-2)$ è uno stimatore corretto di σ^2 , si conclude che lo stimatore di massima verosimiglianza SS_R/n per la stima di σ^2 non è corretto, ma lo è soltanto asintoticamente.

Osserviamo che il valore assunto dalla somma dei quadrati dei residui (6.60) è dato dalla (6.72); quindi, in virtù della (6.73), esso coincide con

$$n\hat{\sigma}^2 = S_{yy} - \hat{b}S_{xy} = \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}}, \quad (6.74)$$

dove l'ultima uguaglianza segue dalla seconda delle (6.69). Usando la notazione

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad (6.75)$$

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad (6.76)$$

si può pertanto anche scrivere

$$SS_R = \frac{S_{xx}S_{yy} - S_{xy}^2}{S_{xx}}. \quad (6.77)$$

6.6 Stima intervallare

Nel paragrafo precedente abbiamo ricavato delle stime per i parametri a e b nel caso di regressione normale. Passiamo ora a indicarne l'uso nel contesto della stima intervallare.

Per determinare degli intervalli fiduciari per a e b faremo uso del seguente lemma, che per brevità ci limitiamo ad enunciare:

6.6. STIMA INTERVALLARE

Lemma 6.6.1 Nella regressione normale la variabile casuale SS_R/σ^2 ha distribuzione chi-quadrato con $n-2$ gradi di libertà ed è indipendente da \hat{A} e \hat{B} .

Una giustificazione euristica dell'origine della distribuzione chi-quadrato con $n-2$ gradi di libertà per SS_R/σ^2 discende dalle considerazioni seguenti. Poiché le variabili casuali Y_i sono indipendenti e dotate di distribuzione normale di media $E(Y_i) = a + b x_i$ e varianza $D^2(Y_i) = \sigma^2$, la variabile casuale

$$\sum_{i=1}^n \frac{[Y_i - E(Y_i)]^2}{D^2(Y_i)} = \sum_{i=1}^n \frac{[Y_i - (a + b x_i)]^2}{\sigma^2} \quad (6.78)$$

ha distribuzione chi-quadrato con n gradi di libertà. Accade poi che se nel secondo membro della (6.78) si sostituiscono a e b con i rispettivi stimatori \hat{A} e \hat{B} dati dalle (6.46) e (6.47) si perdono 2 gradi di libertà, in accordo con la circostanza che si perde un grado di libertà ogni volta che si impone una relazione tra le variabili casuali Y_i . Ciò, ricordando la (6.60), suggerisce che la variabile SS_R/σ^2 ha distribuzione chi-quadrato con $n-2$ gradi di libertà. Rileviamo poi che l'indipendenza tra le variabili SS_R/σ^2 , \hat{A} e \hat{B} è reminiscente dell'indipendenza tra media campionaria e varianza campionaria di campioni casuali tratti da popolazione normale (cfr. Teorema 2.1.4).

Per la (6.50), \hat{B} è una funzione lineare delle variabili Y_1, Y_2, \dots, Y_n ; dalle ipotesi di indipendenza e di distribuzione normale di queste discende che anche \hat{B} ha distribuzione normale, della quale la (6.49) fornisce media e varianza. In conclusione, \hat{B} è una variabile casuale normale di valore medio b e varianza σ^2/S_{xx} . Dalla (6.46) appare poi evidente che anche lo stimatore \hat{A} è una funzione lineare delle variabili Y_1, Y_2, \dots, Y_n . Procedendo in modo analogo a quanto visto per \hat{B} , si ricava facilmente che \hat{A} è una variabile casuale normale di valore medio a e varianza

$$\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right).$$

Teorema 6.6.1 Nella regressione normale un intervallo fiduciario di coefficiente $1-\alpha$ per il parametro b è il seguente:

$$\left(\hat{B} - t_{\alpha/2, n-2} \sqrt{\frac{SS_R}{(n-2)S_{xx}}}, \hat{B} + t_{\alpha/2, n-2} \sqrt{\frac{SS_R}{(n-2)S_{xx}}} \right),$$

mentre un intervallo fiduciario di coefficiente $1-\alpha$ per il parametro a è dato da

$$(\hat{A} - t_{\alpha/2, n-2} \Delta_1, \hat{A} + t_{\alpha/2, n-2} \Delta_1),$$

dove

$$\Delta_1 = \sqrt{\frac{SS_R}{(n-2)} \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}. \quad (6.79)$$

Dim. Si è visto poc'anzi che \hat{B} ha distribuzione normale di valore medio b e varianza σ^2/S_{xx} ; inoltre, in virtù del Lemma 6.6.1, SS_R/σ^2 ha distribuzione chi-quadrato con $n-2$ gradi di

libertà ed è indipendente da \hat{B} . In virtù del Teorema 2.2.1 la variabile casuale

$$T_b = \frac{\hat{B} - b}{\sqrt{\frac{SS_R}{\sigma^2/(n-2)}}} = (\hat{B} - b) \sqrt{\frac{(n-2)S_{xx}}{SS_R}}$$

ha allora distribuzione di Student con $n-2$ gradi di libertà. Quindi si ha:

$$P \left[-t_{\alpha/2, n-2} < (\hat{B} - b) \sqrt{\frac{(n-2)S_{xx}}{SS_R}} < t_{\alpha/2, n-2} \right] = 1 - \alpha,$$

ovvero:

$$P \left[\hat{B} - t_{\alpha/2, n-2} \sqrt{\frac{SS_R}{(n-2)S_{xx}}} < b < \hat{B} + t_{\alpha/2, n-2} \sqrt{\frac{SS_R}{(n-2)S_{xx}}} \right] = 1 - \alpha,$$

da cui resta specificato l'intervallo fiduciario di coefficiente $1 - \alpha$ per il parametro b . Per la stima intervallare di a si può procedere in modo analogo. Invece, facendo uso del Teorema 2.2.1 si ricava facilmente che la variabile casuale

$$T_a = \frac{\hat{A} - a}{\sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \frac{1}{\sqrt{\frac{SS_R}{\sigma^2/(n-2)}}} = \frac{\hat{A} - a}{\sqrt{\frac{SS_R}{(n-2)} \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}} = \frac{\hat{A} - a}{\Delta_1}$$

ha distribuzione di Student con $n-2$ gradi di libertà. Da ciò segue:

$$P \left(-t_{\alpha/2, n-2} < \frac{\hat{A} - a}{\Delta_1} < t_{\alpha/2, n-2} \right) = 1 - \alpha,$$

ovvero:

$$P(\hat{A} - t_{\alpha/2, n-2} \Delta_1 < a < \hat{A} + t_{\alpha/2, n-2} \Delta_1) = 1 - \alpha,$$

da cui si ricava l'intervallo fiduciario di coefficiente $1 - \alpha$ per a .

Talvolta, data l'equazione di regressione lineare $E(Y|x) = a + bx$ ed uno specificato valore x_0 , risulta anche essere utile la stima di

$$E[Y(x_0)] \stackrel{\text{def}}{=} E(Y|x_0) = a + bx_0, \quad (6.80)$$

ossia la stima del valore medio di Y condizionato dal prefissato valore di X . Se si desidera una stima puntuale è naturale usare $\hat{A} + \hat{B}x_0$, che costituisce uno stimatore corretto di $a + bx_0$ avendosi

$$E(\hat{A} + \hat{B}x_0) = E(\hat{A}) + E(\hat{B})x_0 = a + bx_0. \quad (6.81)$$

Se, invece, si desidera ricorrere ad una stima intervallare per $a + bx_0$, nel caso di regressione normale si può usare l'intervallo fiduciario specificato nel teorema che segue.

6.6. STIMA INTERVALLARE

Teorema 6.6.2 Nella regressione normale un intervallo fiduciario di coefficiente $1 - \alpha$ per $a + bx_0$ è dato da

$$(\hat{A} + \hat{B}x_0 - t_{\alpha/2, n-2} \Delta_2, \hat{A} + \hat{B}x_0 + t_{\alpha/2, n-2} \Delta_2),$$

dove

$$\Delta_2 = \sqrt{\frac{SS_R}{(n-2)} \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}. \quad (6.82)$$

Dimm. Per determinare la distribuzione di $\hat{A} + \hat{B}x_0$ osserviamo che dalle relazioni (6.46) e (6.50) si ricava:

$$\hat{A} + \hat{B}x_0 = \bar{Y} - \hat{B}(\bar{x} - x_0) = \sum_{i=1}^n Y_i \left[\frac{1}{n} - \frac{(x_i - \bar{x})(\bar{x} - x_0)}{S_{xx}} \right]. \quad (6.83)$$

La (6.83) mostra che la variabile casuale $\hat{A} + \hat{B}x_0$ è esprimibile come combinazione lineare delle variabili Y_i , le quali sono indipendenti e a distribuzione normale. Ciò comporta che anche $\hat{A} + \hat{B}x_0$ ha distribuzione normale, di media (6.81). Per determinarne la varianza, osserviamo che dalla (6.83) si ottiene

$$\begin{aligned} D^2(\hat{A} + \hat{B}x_0) &= \sum_{i=1}^n D^2(Y_i) \left[\frac{1}{n} - \frac{(x_i - \bar{x})(\bar{x} - x_0)}{S_{xx}} \right]^2 \\ &= \sigma^2 \sum_{i=1}^n \left[\frac{1}{n^2} - 2 \frac{(x_i - \bar{x})(\bar{x} - x_0)}{n S_{xx}} + \frac{(x_i - \bar{x})^2(\bar{x} - x_0)^2}{S_{xx}^2} \right] \\ &= \sigma^2 \left[\frac{1}{n} - 2 \frac{(\bar{x} - x_0)}{n S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) + \frac{(\bar{x} - x_0)^2}{S_{xx}^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right] \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}} \right], \end{aligned}$$

avendo fatto uso della (6.28) e delle relazioni

$$D^2(Y_i) = \sigma^2, \quad \sum_{i=1}^n (x_i - \bar{x}) = 0.$$

Per il Lemma 6.6.1 lo stimatore $\hat{A} + \hat{B}x_0$ è indipendente da SS_R/σ^2 , che ha distribuzione chi-quadrato con $n-2$ gradi di libertà. In virtù del Teorema 2.2.1 la variabile casuale

$$T = \frac{\hat{A} + \hat{B}x_0 - (a + bx_0)}{\sqrt{\sigma^2 \left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}} \right]}} \frac{1}{\sqrt{\frac{SS_R}{\sigma^2/(n-2)}}} = \frac{\hat{A} + \hat{B}x_0 - (a + bx_0)}{\Delta_2}$$

ha pertanto distribuzione di Student con $n-2$ gradi di libertà. Quindi risulta:

$$P \left[-t_{\alpha/2, n-2} < \frac{\hat{A} + \hat{B}x_0 - (a + bx_0)}{\Delta_2} < t_{\alpha/2, n-2} \right] = 1 - \alpha,$$

da cui si ricava:

$$P(\hat{A} + \hat{B}x_0 - t_{\alpha/2,n-2}\Delta_2 < a + bx_0 < \hat{A} + \hat{B}x_0 + t_{\alpha/2,n-2}\Delta_2) = 1 - \alpha.$$

Resta così individuato l'intervallo fiduciario di coefficiente $1 - \alpha$ per $a + bx_0$. ■

Il Teorema 6.6.2 fornisce un intervallo fiduciario per $a + bx_0$. Ciò è significativo in quanto, come mostra l'equazione (6.80), $a + bx_0$ rappresenta il valore mediamente assunto dalla variabile casuale $Y(x_0)$ che, nell'ambito del modello (6.43), denota la variabile Y condizionata dal prefissato valore $X = x_0$. In numerose situazioni risulta però di maggiore interesse ottenere informazioni direttamente sulla variabile casuale $Y(x_0)$ piuttosto che sul suo valore medio $E[Y(x_0)]$. Ad esempio, considerato un certo esperimento effettuato alla pressione atmosferica x_0 potrebbe essere più significativo prevederne l'esito $Y(x_0)$ piuttosto che stimarne il valore medio $E[Y(x_0)] \equiv a + bx_0$. Se invece si effettuasse una serie di esperimenti identici a pressione atmosferica x_0 , la media $a + bx_0$ potrebbe rivestire maggiore interesse.

Il problema della previsione del valore assunto da una variabile casuale $Y(x_0)$ può essere affrontato in vari modi: si può ad esempio utilizzare la media, che minimizza l'errore quadratico medio di previsione; oppure la mediana, che minimizza l'errore assoluto medio di previsione; o, ancora, la moda, che è il valore più probabile. Poiché per una variabile casuale normale media, mediana e moda coincidono, il problema della scelta del predittore di una siffatta variabile non presenta ambiguità. Nell'ipotesi di regressione normale è dunque naturale ricorrere alla media $a + bx_0$ per prevedere il valore assunto dalla variabile casuale $Y(x_0)$. Quando i parametri a e b sono incogniti è ragionevole sostituirli con i rispettivi stimatori \hat{A} e \hat{B} , e quindi usare $\hat{A} + \hat{B}x_0$ come predittore di $Y(x_0)$.

Talvolta, piuttosto che utilizzare un singolo numero per prevedere il valore della variabile casuale $Y(x_0)$, può essere preferibile utilizzare un intervallo. In tal caso si fa ricorso all'*intervallo di previsione* di coefficiente $1 - \alpha$, che consiste in un intervallo (C_p^-, C_p^+) per il quale risulta

$$P[C_p^- < Y(x_0) < C_p^+] = 1 - \alpha, \quad (6.84)$$

dove α è un reale compreso tra 0 e 1, mentre C_p^- e C_p^+ sono due statistiche che soddisfano la condizione $C_p^- < C_p^+$. L'intervallo di previsione non va confuso con l'intervallo fiduciario. Infatti, mentre quest'ultimo è un intervallo casuale che contiene un certo parametro incognito con probabilità $1 - \alpha$, l'intervallo di previsione è un intervallo casuale all'interno del quale la variabile casuale $Y(x_0)$ assume valore con probabilità $1 - \alpha$.

Mostriremo ora come possa ottersi un intervallo di previsione per $Y(x_0)$ nel caso di regressione normale.

Teorema 6.6.3 Nella regressione normale un intervallo di previsione di coefficiente $1 - \alpha$ per $Y(x_0)$ è dato da

$$(\hat{A} + \hat{B}x_0 - t_{\alpha/2,n-2}\Delta_3, \hat{A} + \hat{B}x_0 + t_{\alpha/2,n-2}\Delta_3),$$

dove

$$\Delta_3 = \sqrt{\frac{SS_R}{(n-2)} \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right]}. \quad (6.85)$$

6.6. STIMA INTERVALLARE

Dim. Ricaviamo anzitutto la distribuzione di $Y(x_0) - (\hat{A} + \hat{B}x_0)$. Ricordiamo che, nella regressione normale, $Y(x_0)$ è una variabile casuale normale di media $a + bx_0$ e varianza σ^2 mentre, come visto nel corso della dimostrazione del Teorema 6.6.2, $\hat{A} + \hat{B}x_0$ è una variabile casuale normale di media $a + bx_0$ e varianza

$$\sigma^2 \left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}} \right].$$

Ne segue che $Y(x_0) - (\hat{A} + \hat{B}x_0)$ ha distribuzione normale di media nulla e varianza

$$\sigma^2 \left[1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}} \right].$$

Per il Lemma 6.6.1, SS_R/σ^2 ha distribuzione chi-quadrato con $n - 2$ gradi di libertà. Pertanto, sfruttando l'indipendenza di SS_R da \hat{A} e \hat{B} , oltre che da $Y(x_0)$, dal Teorema 2.2.1 segue che la variabile casuale

$$T = \frac{Y(x_0) - (\hat{A} + \hat{B}x_0)}{\sqrt{\sigma^2 \left[1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}} \right]}} \frac{1}{\sqrt{\frac{SS_R}{\sigma^2} / (n-2)}} = \frac{Y(x_0) - (\hat{A} + \hat{B}x_0)}{\Delta_3}$$

ha distribuzione di Student con $n - 2$ gradi di libertà. Si ricava dunque:

$$P \left[-t_{\alpha/2,n-2} < \frac{Y(x_0) - (\hat{A} + \hat{B}x_0)}{\Delta_3} < t_{\alpha/2,n-2} \right] = 1 - \alpha,$$

da cui si ottiene:

$$P(\hat{A} + \hat{B}x_0 - t_{\alpha/2,n-2}\Delta_3 < Y(x_0) < \hat{A} + \hat{B}x_0 + t_{\alpha/2,n-2}\Delta_3) = 1 - \alpha,$$

e quindi la tesi. ■

Notiamo che l'intervallo fiduciario per $a + bx_0$ di cui al Teorema 6.6.2 differisce dall'intervallo di previsione per $Y(x_0)$ fornito dal Teorema 6.6.3 soltanto per la presenza nel secondo della variabile casuale Δ_3 che per ogni fissato valore di SS_R assume valori certamente maggiori di quelli corrispondentemente assunti da Δ_2 . Da ciò discende che l'intervallo di previsione per il generico valore assunto da $Y(x_0)$ contiene l'intervallo fiduciario per la media $a + bx_0$.

Forniremo ora un esempio in cui si fa uso dei precedenti teoremi per ricavare degli intervalli fiduciari.

Esempio 6.6.1 Riconsideriamo l'Esempio 6.2.3 ed assumiamo che il voto Y ottenuto da uno studente dipenda dal numero X di giorni di studio attraverso una variabile che, con una ragionevole oltre che abituale approssimazione, supporremo normale. Assumiamo pertanto che la variabile Y , condizionata da $X = x$, sia normale di media $a + bx$ e varianza σ^2 . Dai dati presenti nella Tabella 6.2 risulta $\sum_i x_i = 603$, $\sum_i x_i^2 = 24077$, $\sum_i y_i = 418$, $\sum_i y_i^2 = 11046$ e

$\sum_i x_i y_i = 16130$, così che si ha

$$S_{xx} = \sum_i x_i^2 - \frac{1}{n} \left(\sum_i x_i \right)^2 = 24077 - \frac{(603)^2}{16} = 1351.44$$

$$S_{yy} = \sum_i y_i^2 - \frac{1}{n} \left(\sum_i y_i \right)^2 = 11046 - \frac{(418)^2}{16} = 125.75$$

$$S_{xy} = \sum_i x_i y_i - \frac{1}{n} \sum_i x_i \sum_i y_i = 16130 - \frac{603 \cdot 418}{16} = 376.62.$$

Facendo uso della (6.74) si ricava il valore assunto dalla somma dei quadrati dei residui:

$$n \hat{\sigma}^2 = \frac{S_{xx} S_{yy} - S_{xy}^2}{S_{xx}} = \frac{1351.44 \cdot 125.75 - (376.62)^2}{1351.44} = 20.79.$$

Pertanto, il valore assunto dalla statistica

$$\sqrt{\frac{SS_R}{(n-2)S_{xx}}}$$

è dato da

$$\sqrt{\frac{20.79}{14 \cdot 1351.44}} = 0.033. \quad (6.86)$$

Per stimare un intervallo fiduciario di coefficiente $1 - \alpha$ per il parametro b sceglieremo ad esempio $\alpha = 0.05$. Dalla Tabella 4 dell'Appendice B si ricava $t_{\alpha/2; n-2} = t_{0.025; 14} = 2.145$; facendo uso della (6.86) e ricordando che $\hat{b} = 0.2786$, si ottiene pertanto:

$$\hat{b} - t_{\alpha/2; n-2} \hat{\sigma} \sqrt{\frac{n}{(n-2)S_{xx}}} = 0.2786 - 2.145 \cdot 0.033 = 0.2078$$

$$\hat{b} + t_{\alpha/2; n-2} \hat{\sigma} \sqrt{\frac{n}{(n-2)S_{xx}}} = 0.2786 + 2.145 \cdot 0.033 = 0.3494.$$

In virtù del Teorema 6.6.1 concludiamo che $(0.2078, 0.3494)$ è l'intervallo fiduciario stimato per b . Poiché questo non contiene il valore 0, in base ai dati osservati appare legittimo ritenere che vi sia una relazione lineare tra il voto ottenuto ed il numero di giorni di studio.

Procediamo in modo analogo per ricavare un intervallo fiduciario stimato di coefficiente $1 - \alpha = 0.9$ per il parametro a . Osserviamo che

$$\delta_1 = \sqrt{\frac{20.79}{14} \left[\frac{1}{16} + \frac{(603/16)^2}{1351.44} \right]} = 1.286 \quad (6.87)$$

è il valore assunto dalla variabile casuale (6.79). Dalla Tabella 4 dell'Appendice B si trae poi $t_{\alpha/2; n-2} = t_{0.05; 14} = 1.761$. Pertanto, facendo uso della (6.87) e ricordando che $\hat{a} = 15.622$, si ottiene:

$$\hat{a} - t_{\alpha/2; n-2} \delta_1 = 15.622 - 1.761 \cdot 1.286 = 13.357$$

$$\hat{a} + t_{\alpha/2; n-2} \delta_1 = 15.622 + 1.761 \cdot 1.286 = 17.887.$$

6.7. VERIFICA DI IPOTESI

Dal Teorema 6.6.1 segue allora che l'intervallo fiduciario stimato per a è $(13.357, 17.887)$. Osserviamo che a costituisce il valore fornito dalla retta di regressione $y = a + bx$ in corrispondenza di $x = 0$. Pertanto, poiché l'intervallo fiduciario stimato per a esclude il valore 18 (ossia la sufficienza), i dati disponibili mostrano che, con una "fiducia" pari al 90%, in corrispondenza di 0 giorni di studio ci si attende un voto inferiore alla sufficienza.

Supponendo ora che degli studenti sostengono l'esame dopo 40 giorni di studio, determiniamo una stima intervallare del voto medio $a + 40b$. Ora la variabile casuale (6.82) assume il valore

$$\delta_2 = \sqrt{\frac{20.79}{14} \left[\frac{1}{16} + \frac{(40 - 603/16)^2}{1351.44} \right]} = 0.314. \quad (6.88)$$

Scegliendo ad esempio $\alpha = 0.02$, la Tabella 4 dell'Appendice B fornisce $t_{\alpha/2; n-2} = t_{0.01; 14} = 2.624$. Così, usando la (6.88) e ricordando le stime $\hat{a} = 15.622$ e $\hat{b} = 0.2786$, si ricava:

$$\hat{a} + 40\hat{b} - t_{\alpha/2; n-2} \delta_2 = 15.622 + 40 \cdot 0.2786 - 2.624 \cdot 0.314 = 25.94$$

$$\hat{a} + 40\hat{b} + t_{\alpha/2; n-2} \delta_2 = 15.622 + 40 \cdot 0.2786 + 2.624 \cdot 0.314 = 27.59.$$

In virtù del Teorema 6.6.2 l'intervallo fiduciario stimato per $a + 40b$ è pertanto $(25.94, 27.59)$. Da ciò si conclude che con una "fiducia" pari al 98% in corrispondenza di 40 giorni di studio ci si attende un voto medio compreso nell'intervallo $(25.94, 27.59)$.

Dopo aver ricavato una stima intervallare del voto medio $a + 40b$, ci prefiggiamo ora di determinare un intervallo di previsione per il voto che otterrà un singolo studente che sostenga l'esame dopo 40 giorni di studio. A tale scopo calcoliamo il valore assunto dalla variabile casuale (6.85):

$$\delta_3 = \sqrt{\frac{20.79}{14} \left[1 + \frac{1}{16} + \frac{(40 - 603/16)^2}{1351.44} \right]} = 1.258. \quad (6.89)$$

Ponendo ancora $\alpha = 0.02$, con $t_{\alpha/2; n-2} = t_{0.01; 14} = 2.624$, usando la (6.89) e ricorrendo alle stime $\hat{a} = 15.622$ e $\hat{b} = 0.2786$, si ricava:

$$\hat{a} + 40\hat{b} - t_{\alpha/2; n-2} \delta_3 = 15.622 + 40 \cdot 0.2786 - 2.624 \cdot 1.258 = 23.46$$

$$\hat{a} + 40\hat{b} + t_{\alpha/2; n-2} \delta_3 = 15.622 + 40 \cdot 0.2786 + 2.624 \cdot 1.258 = 30.07.$$

In virtù del Teorema 6.6.3, l'intervallo di previsione per $Y(40)$ è pertanto $(23.46, 30.07)$, così che con una "fiducia" pari al 98% con 40 giorni di studio lo studente otterrà un voto superiore a 23. ♦

6.7 Verifica di ipotesi

Come si è visto in precedenza, il problema della regressione lineare nel caso normale può essere trattato riguardando le coppie osservate (x_i, y_i) , $(i = 1, 2, \dots, n)$, come se i valori x_1, x_2, \dots, x_n fossero delle costanti e l' n -upla (y_1, y_2, \dots, y_n) fosse la realizzazione di un vettore casuale (Y_1, Y_2, \dots, Y_n) in cui ciascuna variabile Y_i ha distribuzione normale di media $a + bx_i$ e varianza σ^2 . Finora abbiamo affrontato il problema delle stime puntuale e intervallare dei parametri che caratterizzano tali variabili; unitamente alla loro stima, in talune

situazioni può però anche essere di ausilio la verifica di ipotesi formulate su detti parametri. Con riferimento alla media $E(Y_i) = a + b x_i$, vedremo qui come sia possibile costruire dei test per verificare ipotesi sui parametri b ed a considerati entrambi incogniti.

Proposizione 6.7.1 Nella regressione normale un test di ampiezza α per verificare l'ipotesi nulla $H_0: b = b_0$ contro l'ipotesi alternativa $H_1: b \neq b_0$ è quello che ha regione critica

$$C = \{(x_i, y_i); i = 1, 2, \dots, n; |t| \geq t_{\alpha/2, n-2}\}, \quad (6.90)$$

dove t è il valore assunto dalla variabile casuale

$$T_{b_0} = (\hat{B} - b_0) \sqrt{\frac{(n-2)S_{xx}}{SS_R}}. \quad (6.91)$$

Dim. Analogamente a quanto visto nel corso del Teorema 6.6.1, sotto l'ipotesi nulla $H_0: b = b_0$ la variabile casuale (6.91) ha distribuzione di Student con $n - 2$ gradi di libertà. Risulta quindi:

$$P(|T_{b_0}| \geq t_{\alpha/2, n-2} | b = b_0) = \alpha.$$

Da ciò si ricava che la regione critica (6.90) ha ampiezza α . ■

Come mostrano le (6.74) e (6.30), la somma dei quadrati dei residui SS_R assume valore $(S_{xx} S_{yy} - S_{xy}^2)/S_{xx}$, mentre il valore assunto dallo stimatore $\hat{B} = \hat{b} = S_{xy}/S_{xx}$. Il valore t che appare in (6.90) si può dunque riesprimere al seguente modo:

$$t = (S_{xy} - b_0 S_{xx}) \sqrt{\frac{n-2}{S_{xx} S_{yy} - S_{xy}^2}}. \quad (6.92)$$

Si noti che di particolare rilievo è il caso in cui si desidera verificare l'ipotesi nulla $H_0: b = 0$ contro l'ipotesi alternativa $H_1: b \neq 0$. Infatti l'ipotesi $b = 0$ equivale all'ipotesi che l'equazione di regressione lineare si riduca a $E(Y|x) = a$, ossia che in essa non appaia la dipendenza da x . Esamineremo ora un esempio in cui si fa uso della Proposizione 6.7.1 per verificare un'ipotesi di questo tipo.

Esempio 6.7.1 In un laboratorio viene effettuato un esperimento che consiste nel misurare il peso y in milligrammi di una certa sostanza al variare della temperatura x in gradi centigradi. Nella Tabella 6.4 sono riportati i valori osservati nel corso di 20 ripetizioni dell'esperimento. Supponiamo che il peso corrispondente alla misura effettuata a temperatura x_i sia descritto da una variabile casuale normale Y_i di media $E(Y_i) = a + b x_i$, così che siano valide le ipotesi della regressione normale. L'aumento dei dati registrati, riportati nella Figura 6.8, suggerisce che una retta di regressione stimata $\hat{y} = \hat{a} + \hat{b} x$ potrebbe fornire un buon attagliamento delle coppie (x_i, y_i) se il coefficiente stimato \hat{b} fosse positivo. Ciò mostrerebbe che esiste un dipendenza del peso della sostanza dalla temperatura. Ricaviamo allora la retta di regressione stimata $\hat{y} = \hat{a} + \hat{b} x$. Dai dati elencati nella Tabella 6.4 segue $\sum_i x_i = 580$, $\sum_i x_i^2 = 19480$, $\sum_i y_i = 30.6$, $\sum_i y_i^2 = 47.3$ e $\sum_i x_i y_i = 912.8$. Da questi, essendo $n = 20$, si trae:

$$S_{xx} = \sum_i x_i^2 - \frac{1}{n} \left(\sum_i x_i \right)^2 = 19480 - \frac{(580)^2}{20} = 2660$$

6.7. VERIFICA DI IPOTESI

Tabella 6.4: Temperatura e peso per un campione di 20 osservazioni.

temperatura x	peso y	temperatura x	peso y
10	1.2	30	1.5
12	1.4	32	1.6
14	1.3	34	1.7
16	1.5	36	1.5
18	1.4	38	1.6
20	1.3	40	1.6
22	1.5	42	1.7
24	1.6	44	1.4
26	1.7	46	1.8
28	1.6	48	1.7

$$\begin{aligned} S_{yy} &= \sum_i y_i^2 - \frac{1}{n} \left(\sum_i y_i \right)^2 = 47.3 - \frac{(30.6)^2}{20} = 0.482 \\ S_{xy} &= \sum_i x_i y_i - \frac{1}{n} \sum_i x_i \sum_i y_i = 912.8 - \frac{580 \cdot 30.6}{20} = 25.4. \end{aligned}$$

Le stime (6.30) forniscono quindi:

$$\begin{aligned} \hat{a} &= \frac{1}{n} \left(\sum_i y_i - \frac{S_{xy}}{S_{xx}} \sum_i x_i \right) = \frac{1}{20} \left(30.6 - \frac{25.4}{2660} \cdot 580 \right) = 1.254, \\ \hat{b} &= \frac{S_{xy}}{S_{xx}} = \frac{25.4}{2660} = 0.0095. \end{aligned}$$

La retta di regressione stimata, rappresentata in Figura 6.8, ha dunque equazione $\hat{y} = 1.254 + 0.0095x$. La circostanza che il suo coefficiente angolare è molto prossimo a zero potrebbe far sorgere il dubbio che in realtà il valore di b possa essere proprio nullo. Si potrebbe quindi ipotizzare che non vi sia dipendenza del peso della sostanza dalla temperatura. Occorre allora sottoporre a verifica l'ipotesi nulla $H_0: b = 0$ contro l'ipotesi alternativa $H_1: b \neq 0$. In altri termini, avendo notato che i pesi osservati y_i tendono mediamente a crescere con le temperature x_i , si è ricorso ad un modello di regressione normale $E(Y_i) = a + b x_i$. Si desidera ora verificare se in realtà $E(Y_i)$ sia indipendente da x_i , ossia se risulti $b = 0$, nel qual caso l'aumento dei pesi y_i registrati può immaginarsi dovuto esclusivamente al caso. Facciamo a tal fine ricorso alla Proposizione 6.7.1 per costruire la regione critica del test. Poiché è $b_0 = 0$, la (6.92) fornisce:

$$t = 25.4 \sqrt{\frac{18}{2660 \cdot 0.482 - (25.4)^2}} = 4.27.$$

Scegliendo ad esempio $\alpha = 0.01$, dalla Tabella 4 dell'Appendice B si ricava $t_{\alpha/2, n-2} =$

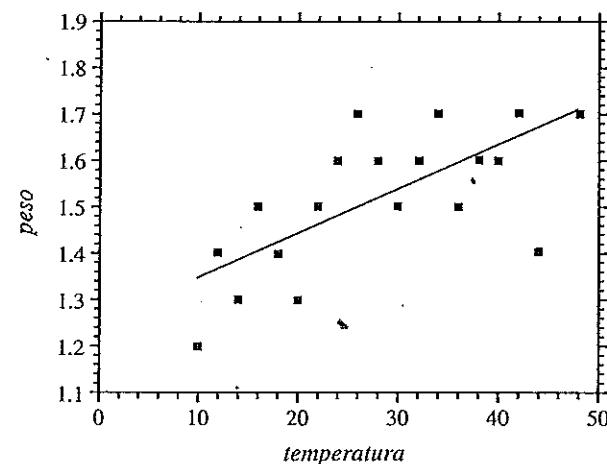


Figura 6.8: Interpolazione dei dati dell'Esempio 6.7.1.

$t_{0.005;18} = 2.861$. Dalla (6.90) segue allora che la regione critica del test suggerita nella Proposizione 6.7.1 è la seguente:

$$\mathcal{C} = \{(x_i, y_i); i = 1, 2, \dots, 20; |t| \geq 2.861\},$$

così che il valore osservato $t = 4.27$ appartiene a \mathcal{C} . Di conseguenza, in base ai dati ottenuti, l'ipotesi nulla $H_0: b = 0$ è da rifiutare. Si conclude che nel modello di regressione normale $E(Y_i) = a + bx_i$ ipotizzato la possibilità che il parametro b sia nullo va scartata. L'aumento dei pesi y_i non è dunque da attribuirsi al caso, ma alla circostanza che esiste una dipendenza lineare crescente tra il peso medio della sostanza in esame e la temperatura. ♦

Vedremo ora come sia possibile costruire un test per verificare ipotesi sul parametro a di un'equazione di regressione normale $E(Y_i) = a + bx_i$, supponendo b incognito.

Proposizione 6.7.2 Nella regressione normale un test di ampiezza α per verificare l'ipotesi nulla $H_0: a = a_0$ contro l'ipotesi alternativa $H_1: a \neq a_0$ è quello che ha come regione critica

$$\mathcal{C} = \{(x_i, y_i); i = 1, 2, \dots, n; |t| \geq t_{\alpha/2;n-2}\}, \quad (6.93)$$

dove t è il valore assunto dalla variabile casuale

$$T_{a_0} = \frac{\hat{a} - a_0}{\sqrt{\frac{SS_R}{(n-2)} \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}}. \quad (6.94)$$

6.7. VERIFICA DI IPOTESI

Tabella 6.5: Tempi di lavoro e carburante per un campione di 14 osservazioni.

carburante x	tempo y	carburante x	tempo y
1.0	0.22	4.5	1.01
1.5	0.36	5.0	1.12
2.0	0.42	5.5	1.14
2.5	0.58	6.0	1.30
3.0	0.70	6.5	1.34
3.5	0.78	7.0	1.49
4.0	0.84	7.5	1.55

Dim. Seguendo lo stesso procedimento adottato per la dimostrazione del Teorema 6.6.1, sotto l'ipotesi nulla $H_0: a = a_0$ la variabile casuale (6.94) ha distribuzione di Student con $n - 2$ gradi di libertà. Sussiste dunque la seguente relazione:

$$P(|T_{a_0}| \geq t_{\alpha/2;n-2} | a = a_0) = \alpha.$$

La regione critica (6.93) ha pertanto ampiezza α . ■

Nell'esempio che segue si fa uso della Proposizione 6.7.2 per verificare un'ipotesi sul parametro a .

Esempio 6.7.2 In un'officina vengono effettuate prove di funzionamento di motori consistenti nel sottoporre 14 esemplari dello stesso tipo a diversi carichi di lavoro. Nella Tabella 6.5 è riportato per ognuno dei 14 motori il tempo di lavoro y (in ore) in corrispondenza all'ammonitare x (in litri) di carburante adoperato. Si suppone che sia utilizzabile l'ipotesi di regressione normale, ossia che il tempo di funzionamento corrispondente al motore rifornito di una quantità x_i di carburante sia rappresentabile da una variabile casuale normale Y_i di media $E(Y_i) = a + bx_i$. Determiniamo la retta di regressione stimata $\hat{y} = \hat{a} + \hat{b}x$. Dai dati riportati nella Tabella 6.5 si ricava $\sum x_i = 59.50$, $\sum x_i^2 = 309.75$, $\sum y_i = 12.85$, $\sum y_i^2 = 14.18$, $\sum x_i y_i = 66.22$. Da questi, essendo $n = 14$, si ottiene:

$$S_{xx} = \sum_i x_i^2 - \frac{1}{n} \left(\sum_i x_i \right)^2 = 309.75 - \frac{(59.50)^2}{14} = 56.87$$

$$S_{yy} = \sum_i y_i^2 - \frac{1}{n} \left(\sum_i y_i \right)^2 = 14.18 - \frac{(12.85)^2}{14} = 2.39$$

$$S_{xy} = \sum_i x_i y_i - \frac{1}{n} \sum_i x_i \sum_i y_i = 66.22 - \frac{59.50 \cdot 12.85}{14} = 11.61.$$

Le stime (6.30) forniscono quindi:

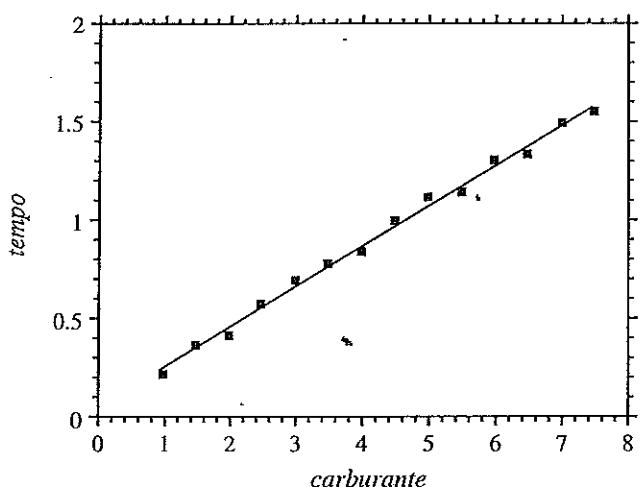


Figura 6.9: Interpolazione dei dati dell'Esempio 6.7.2.

$$\begin{aligned} \hat{a} &= \frac{1}{n} \left(\sum_i y_i - \frac{S_{xy}}{S_{xx}} \sum_i x_i \right) = \frac{1}{14} \left(12.85 - \frac{11.61}{56.87} \cdot 59.50 \right) = 0.05 \\ \hat{b} &= \frac{S_{xy}}{S_{xx}} = \frac{11.61}{56.87} = 0.20. \end{aligned}$$

La retta di regressione stimata, indicata nella Figura 6.9, ha dunque equazione $\hat{y} = 0.05 + 0.20x$. Notiamo che il termine noto in tale equazione è molto prossimo a zero. Ciò, unitamente alla considerazione che in assenza di carburante ci si attende durata nulla di funzionamento, indica che il valore reale di a dovrebbe essere 0. Conviene allora sottoporre a verifica l'ipotesi nulla $H_0: a = 0$ contro l'ipotesi alternativa $H_1: a \neq 0$. In altri termini, dopo aver inizialmente supposto valido il modello di regressione normale del tipo $E(Y_i) = a + bx_i$, si desidera ora verificare se in realtà non sia più realistico il modello del tipo $E(Y_i) = bx_i$. La regione critica del test può essere ricavata facendo uso della Proposizione 6.7.2. A tal fine notiamo che la somma dei quadrati dei residui SS_R assume valore

$$\frac{S_{xx} S_{yy} - S_{xy}^2}{S_{xx}} = \frac{56.87 \cdot 2.39 - (11.61)^2}{56.87} = 0.02;$$

pertanto, essendo $a_0 = 0$ e $\hat{a} = 0.05$, la variabile casuale (6.94) assume il valore

$$t = \frac{0.05}{\sqrt{\frac{0.02}{12} \left[\frac{1}{14} + \frac{1}{56.87} \left(\frac{59.50}{14} \right)^2 \right]}} = 1.96.$$

6.8. ADEGUATEZZA DEL MODELLO

Se ad esempio si sceglie $\alpha = 0.01$, la Tabella 4 dell'Appendice B fornisce $t_{\alpha/2, n-2} = t_{0.005, 12} = 3.055$. Dalla (6.93) si ricava così la regione critica del test fornita dalla Proposizione 6.7.2:

$$C = \{(x_i, y_i); i = 1, 2, \dots, 20; |t| \geq 3.055\}.$$

Poiché il valore osservato $t = 1.96$ non appartiene a C , in base ai dati disponibili l'ipotesi nulla $H_0: a = 0$ è da accettare. Si può dunque concludere che il modello di regressione $E(Y_i) = a + bx_i$ va sostituito con il modello $E(Y_i) = bx_i$. È allora opportuno determinare la nuova retta di regressione stimata $\hat{y} = \hat{b}x$. Facendo uso della (6.35) si ricava la stima del coefficiente di regressione b quando è $a = 0$:

$$\hat{b} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{66.22}{309.75} = 0.21.$$

La nuova retta di regressione stimata è quindi $\hat{y} = 0.214x$.

6.8 Adeguatezza del modello

È qui opportuno svolgere alcune considerazioni che consentano di chiarire in quali casi sia plausibile ritenere valido un ipotizzato modello di regressione. Si è precedentemente visto che, fissato il modello di regressione, si perviene agevolmente alla curva di regressione stimata sia nel caso lineare che in casi a questo direttamente riconducibili. Occorre però prestare attenzione alla circostanza che anche se si riesce a determinare una curva in qualche modo interpolante i dati disponibili, non è detto che il modello di regressione ipotizzato sia da ritenersi idoneo a giustificare la relazione individuata tra le variabili coinvolte. Si può ad esempio verificare che una curva di regressione stimata si discosti eccessivamente dai dati osservati, il che può far sospettare che il modello ipotizzato non sia quello maggiormente idoneo a descrivere i dati in considerazione.

In un modello di regressione lineare spesso si perviene ad una valutazione della bontà dell'approssimazione basata sul metodo dei minimi quadrati facendo ricorso a talune opportune variabili casuali. Una di queste è il *coefficiente di determinazione*

$$R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = 1 - \frac{SS_R}{S_{yy}}, \quad (6.95)$$

con S_{xy} , S_{xx} , S_{yy} e SS_R definite rispettivamente dalle (6.75), (6.28), (6.76) e (6.60) e dove la seconda uguaglianza nella (6.95) discende dalla (6.77). Osserviamo che poiché risulta

$$\frac{S_{xy}^2}{S_{xx} S_{yy}} \geq 0, \quad \frac{SS_R}{S_{yy}} \geq 0,$$

dalla (6.95) segue $0 \leq R^2 \leq 1$. Si noti che R^2 assume il valore 0 quando è nullo il valore S_{xy} assunto dalla variabile casuale S_{xy} . Come indicato dalla (6.30), tale condizione si verifica quando è nulla la stima del coefficiente b in un modello di regressione lineare $E(Y|x) =$

$a + bx$, ossia quando la retta di regressione stimata è $\hat{y} = \hat{a}$. D'altro canto, R^2 assume il valore 1 quando la somma dei quadrati dei residui $\sum R_i^2$ assume il valore 0. La (6.60) mostra che ciò accade allorché risulta

$$\sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i)^2 = 0,$$

ossia quando i dati y_i sono perfettamente allineati lungo la retta di regressione stimata $\hat{y} = \hat{a} + \hat{b}x$.

Dalla (6.95) scaturisce un'interessante osservazione: in virtù delle (6.60) e (6.71) la condizione $R^2 \geq 0$ equivale alla seguente:

$$\sum_{i=1}^n (Y_i - \hat{A} - \hat{B}x_i)^2 \leq \sum_{i=1}^n (Y_i - \bar{Y})^2 \equiv S_{YY}. \quad (6.96)$$

La diseguaglianza (6.96) esprime la circostanza che la somma dei quadrati dei residui (6.60) nel modello $E(Y|x) = a + bx$ è sempre non maggiore di quella corrispondente (6.63) per il modello $E(Y|x) = a$, il che è una conferma della ovvia superiorità del modello a due parametri.

Unitamente al coefficiente di determinazione R^2 , largamente usato è l'*indice di "fit"*, o *indice di attaglimento*

$$R = \sqrt{R^2} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}. \quad (6.97)$$

Il valore r che esso assume in corrispondenza delle coppie di dati considerate viene spesso interpretato come indicatore della bontà dell'approssimazione ottenuta mediante la retta ai minimi quadrati. Un valore molto prossimo all'unità evidenzia una situazione in cui la somma dei quadrati dei residui $\sum R_i^2$ assume valore prossimo a zero, così che i dati sono "ben approssimati" dalla retta di regressione stimata ai minimi quadrati. Va però detto che un valore elevato di r non implica che il modello di regressione lineare sia appropriato, così come un valore di r prossimo a 0 non indica necessariamente che esso sia inadeguato. Invero, il valore assunto da R fornisce solo un'indicazione di quanto l'approssimazione costituita dalla retta $\hat{y} = \hat{a} + \hat{b}x$ sia migliore di quella fornita da $\hat{y} = \hat{a}$. Così, nell'Esempio 6.7.1 risulta

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{25.4}{\sqrt{2660 \cdot 0.482}} = 0.709,$$

mentre per l'Esempio 6.7.2 si ha

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{11.6075}{\sqrt{56.875 \cdot 2.3825}} = 0.997.$$

Utilizzando la retta $\hat{y} = \hat{a} + \hat{b}x$ in luogo di $\hat{y} = \hat{a}$ si ottiene dunque un miglioramento dell'approssimazione, maggiore nel caso dell'Esempio 6.7.2.

Un ulteriore strumento che consente di verificare la validità del modello di regressione lineare $E(Y|x) = a + bx$ è l'analisi dei residui $R_i - \hat{Y}_i = Y_i - (\hat{A} + \hat{B}x_i)$. Come mostrato nella Proposizione 6.4.3, la statistica $\sum R_i^2/(n-2)$ è uno stimatore corretto della varianza $\sigma^2 \equiv D^2(Y_i)$. È allora possibile definire i *residui standardizzati* nel seguente modo:

$$\frac{Y_i - (\hat{A} + \hat{B}x_i)}{\sqrt{\sum R_i^2/(n-2)}} \quad (i = 1, 2, \dots, n). \quad (6.98)$$

6.8. ADEGUATEZZA DEL MODELLO

Tabella 6.6: Valori assunti dai residui standardizzati.

x_i	residui stand.	x_i	residui stand.
24	-1.8943	37	-0.7633
26	0.1103	39	0.4206
28	-1.1676	41	-0.0366
29	0.2450	43	0.3268
31	1.4290	45	1.5107
32	0.3798	48	-0.8163
35	1.3351	53	-0.3182
36	0.2859	56	-1.0041

Quando il modello di regressione lineare è corretto i residui standardizzati (6.98) sono approssimativamente indipendenti, con distribuzione normale standard. Ciò comporta che all'incirca il 95% dei valori da questi assunti sono compresi tra -2 e 2, avendosi $P(-1.96 < Z < 1.96) = 0.95$ per la variabile casuale normale standard Z . Così, per valutare anche visivamente la ragionevolezza del modello di regressione lineare utilizzato, è opportuno riportare in un grafico i valori assunti dai residui standardizzati: se questi si dispongono in guisa da costituire una figura che non presenta asimmetrie significative rispetto all'asse delle ascisse, o se più del 5% di essi fuoriesce dall'intervallo (-2, 2), c'è da sospettare che il modello di regressione lineare non sia adeguato per la descrizione dei dati.

Esempio 6.8.1 Prendiamo in esame i dati dell'Esempio 6.2.3. Ricordando anche quanto visto nell'Esempio 6.6.1, risulta $S_{xx} = 1351.44$, $S_{yy} = 125.75$ e $S_{xy} = 376.62$. Il valore dell'indice di fit è dunque

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{376.62}{\sqrt{1351.44 \cdot 125.75}} = 0.92$$

che è prossimo all'unità. Ciò indica che i dati sono ben approssimati dalla retta di regressione stimata ai minimi quadrati $\hat{y} = 15.622 + 0.2786x$. Poiché la statistica $\sqrt{\sum R_i^2/(n-2)}$ assume il valore $\sqrt{20.79/14} = 1.2186$, i residui standardizzati assumono i valori

$$\frac{y_i - (15.622 + 0.2786x_i)}{1.2186} \quad (i = 1, 2, \dots, 16).$$

Questi sono stati riportati nella Tabella 6.6 e nella Figura 6.10. Come si vede, essi cadono tutti nell'intervallo (-2, 2) e non esibiscono alcuna particolare tendenza o struttura. Eppure, poiché gli ultimi 3 valori sono negativi, è giustificato il sospetto che il modello lineare non sia del tutto adeguato nel senso che per grandi valori di x la retta di approssimazione $\hat{y} = 15.622 + 0.2786x$ potrebbe fornire una stima distorta per eccesso. Ciò è anche confermato dalla circostanza che le y_i sono dei voti, che quindi non possono essere maggiori di 30. Di conseguenza, la retta ai minimi quadrati $\hat{y} = 15.622 + 0.2786x$ può fornire una stima attendibile solo quando x è tale da aversi $\hat{y} \leq 30$, ossia per $x \leq (30 - 15.622)/0.2786 = 51.608$. ♦

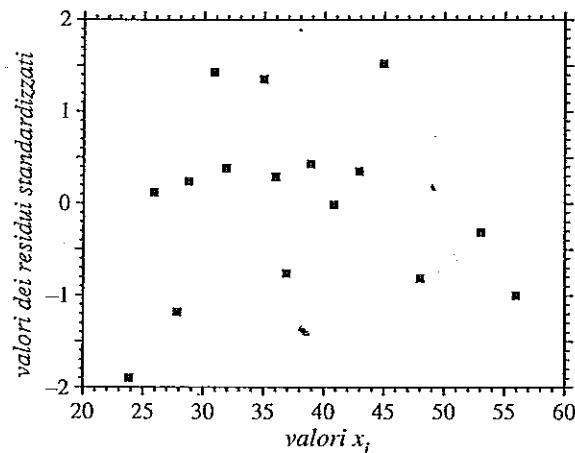


Figura 6.10: Rappresentazione dei residui standardizzati di cui alla Tabella 6.5.

6.9 Minimi quadrati pesati

Nel § 6.4 si è visto che in diversi contesti le coppie osservate (x_i, y_i) , $(i = 1, 2, \dots, n)$, possono essere riguardate come se i valori x_1, x_2, \dots, x_n fossero delle costanti e l' n -upla (y_1, y_2, \dots, y_n) fosse la realizzazione di un vettore casuale (Y_1, Y_2, \dots, Y_n) , dove le variabili casuali Y_i sono indipendenti con medie $E(Y_i) = a + b x_i$ e varianze $D^2(Y_i) = \sigma^2$ ($i = 1, 2, \dots, n$). Talvolta l'ipotesi che le varianze $D^2(Y_i)$ siano costanti non è adeguata per descrivere il fenomeno in esame. Sorge allora la necessità di analizzare anche il caso più generale in cui $D^2(Y_i)$ dipende dall'indice i . A tale scopo rivolgeremo qui l'attenzione al caso in cui le variabili casuali Y_i sono indipendenti con medie e varianze

$$E(Y_i) = a + b x_i \quad D^2(Y_i) = \frac{\sigma^2}{w_i} \quad (i = 1, 2, \dots, n), \quad (6.99)$$

dove a , b e σ^2 sono parametri incogniti e w_1, w_2, \dots, w_n sono reali positivi supposti noti. Utilizzando il principio dei minimi quadrati per il presente modello perveniamo ad una stima dei coefficienti di regressione a e b . Stavolta converrà scegliere come stime i valori di a e b per i quali, in luogo della (6.20), assume valore minimo la funzione $Q(a, b)$ così definita:

$$Q(a, b) \stackrel{\text{def}}{=} \sum_{i=1}^n \frac{[y_i - E(Y_i)]^2}{D^2(Y_i)} = \frac{1}{\sigma^2} \sum_{i=1}^n w_i [y_i - (a + b x_i)]^2. \quad (6.100)$$

È questo il cosiddetto *metodo dei minimi quadrati pesati*, in quanto $Q(a, b)$ rappresenta in questo caso una somma pesata dei quadrati dei discostamenti $e_i = y_i - (a + b x_i)$. Per determinare il minimo della funzione (6.100) imponiamo l'annullarsi delle sue derivate parziali

6.9. MINIMI QUADRATI PESATI

rispetto ad a e a b :

$$\frac{\partial}{\partial a} Q(a, b) = -\frac{2}{\sigma^2} \sum_{i=1}^n w_i [y_i - (a + b x_i)] = 0$$

$$\frac{\partial}{\partial b} Q(a, b) = -\frac{2}{\sigma^2} \sum_{i=1}^n w_i x_i [y_i - (a + b x_i)] = 0;$$

perveniamo così al sistema delle equazioni normali nelle incognite a e b

$$a \sum_{i=1}^n w_i + b \sum_{i=1}^n w_i x_i = \sum_{i=1}^n w_i y_i$$

$$a \sum_{i=1}^n w_i x_i + b \sum_{i=1}^n w_i x_i^2 = \sum_{i=1}^n w_i x_i y_i$$

la cui soluzione fornisce i valori \hat{a} e \hat{b} per i quali $Q(a, b)$ è minima:

$$\hat{a} = \frac{\sum_{i=1}^n w_i y_i - \hat{b} \sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \quad (6.101)$$

$$\hat{b} = \frac{\sum_{j=1}^n w_j \sum_{i=1}^n w_i x_i y_i - \sum_{i=1}^n w_i x_i \sum_{j=1}^n w_j y_j}{\sum_{j=1}^n w_j \sum_{i=1}^n w_i x_i^2 - \left(\sum_{i=1}^n w_i x_i \right)^2}. \quad (6.102)$$

È immediato verificare che per $w_1 = w_2 = \dots = w_n$ le stime ai minimi quadrati pesati (6.101) e (6.102) coincidono rispettivamente con le (6.23) e (6.24). Ad eccezione del caso banale $x_1 = x_2 = \dots = x_n$, anche ora la retta di regressione $\hat{y} = \hat{a} + \hat{b} x$ stimata esiste ed è unica; invero, il denominatore a secondo membro della (6.102) è positivo avendosi:³

$$\sum_{j=1}^n w_j \sum_{i=1}^n w_i x_i^2 - \left(\sum_{i=1}^n w_i x_i \right)^2 = \sum_{\substack{i,j=1 \\ i < j}}^n w_i w_j (x_i - x_j)^2.$$

È anche opportuno sottolineare che nella somma a secondo membro della (6.100) viene dato maggiore peso ai termini contenenti grandi valori di w_i , ossia piccoli valori della varianza $D^2(Y_i)$. In altri termini, nel metodo dei minimi quadrati pesati i valori osservati y_i più attendibili danno un contributo maggiore nella stima dei coefficienti di regressione.

Nell'esempio che segue si illustra una situazione che evidenzia l'utilità del metodo dei minimi quadrati pesati.

³La dimostrazione è analoga a quella relativa alla (6.27).

Esempio 6.9.1 Sia (X_1, X_2, \dots, X_n) un campione casuale estratto da una popolazione caratterizzata da media θ e varianza σ^2 . Supponiamo che le variabili costituenti il campione non siano direttamente osservabili, ma che lo siano le variabili Y_1 e Y_2 , così definite:

$$Y_1 = X_1 + X_2 + \dots + X_r, \quad Y_2 = X_{r+1} + X_{r+2} + \dots + X_n,$$

con $1 \leq r < n$. Si desidera stimare la media θ ricorrendo alle statistiche Y_1 e Y_2 . Poiché risulta

$$E(Y_1) = r\theta, \quad E(Y_2) = (n-r)\theta, \quad (6.103)$$

il metodo dei minimi quadrati suggerisce di usare come stima di θ il minimo della funzione

$$Q(\theta) = \sum_{i=1}^2 [y_i - E(Y_i)]^2 = (y_1 - r\theta)^2 + [y_2 - (n-r)\theta]^2.$$

Imponendo che la derivata di $Q(\theta)$ si annulli, otteniamo l'equazione

$$-2r(y_1 - r\theta) - 2(n-r)[y_2 - (n-r)\theta] = 0,$$

la cui soluzione conduce al seguente stimatore di θ :

$$\hat{\theta} = \frac{rY_1 + (n-r)Y_2}{r^2 + (n-r)^2}. \quad (6.104)$$

Questo è uno stimatore lineare, ed è corretto poiché in virtù delle (6.103) risulta

$$E(\hat{\theta}) = \frac{rE(Y_1) + (n-r)E(Y_2)}{r^2 + (n-r)^2} = \frac{r^2\theta + (n-r)^2\theta}{r^2 + (n-r)^2} = \theta.$$

Tuttavia esso non è il migliore stimatore della media; invero, tra gli stimatori lineari corretti della media di una popolazione quello a varianza minima è la media campionaria \bar{X} (cfr. Corollario 3.2.1).

Per individuare lo stimatore della media θ al quale il criterio dei minimi quadrati pesati conduce, consideriamo la funzione

$$Q(\theta) = \sum_{i=1}^2 \frac{[y_i - E(Y_i)]^2}{D^2(Y_i)} = \frac{(y_1 - r\theta)^2}{r\sigma^2} + \frac{[y_2 - (n-r)\theta]^2}{(n-r)\sigma^2},$$

imponendo che risulti

$$\frac{d}{d\theta} Q(\theta) = -\frac{2}{\sigma^2}(y_1 + y_2 - n\theta) = 0,$$

si perviene al seguente stimatore:

$$\hat{\theta}' = \frac{Y_1 + Y_2}{n} = \frac{1}{n} \sum_{i=1}^n X_i \equiv \bar{X}.$$

Questo è preferibile allo stimatore (6.104) in quanto coincide con la media campionaria che è uno stimatore corretto a varianza minore di quella di $\hat{\theta}$, avendosi

$$D^2(\hat{\theta}') = \frac{r^3 + (n-r)^3}{[r^2 + (n-r)^2]^2} \sigma^2 > \frac{\sigma^2}{n} \equiv D^2(\bar{X}) \quad (n > r \geq 1).$$

6.9. MINIMI QUADRATI PESATI

Mostriremo ora come il principio dei minimi quadrati sia in relazione con quello dei minimi quadrati pesati. A tal fine osserviamo che le coppie (x_i, y_i) possono anche essere riguardate assumendo x_1, x_2, \dots, x_n come costanti e y_1, y_2, \dots, y_n come valori assunti dalle variabili casuali

$$Y_i = a + b x_i + V_i \quad (i = 1, 2, \dots, n), \quad (6.105)$$

con V_1, V_2, \dots, V_n variabili casuali indipendenti con

$$E(V_i) = 0, \quad D^2(V_i) = \frac{\sigma^2}{w_i}. \quad (6.106)$$

Invero, dalle (6.105) e (6.106) seguono le (6.99). Moltiplicando ambo i membri per $\sqrt{w_i}$, le equazioni (6.105) si trasformano nelle seguenti:

$$Y_i \sqrt{w_i} = a \sqrt{w_i} + b x_i \sqrt{w_i} + V_i \sqrt{w_i} \quad (i = 1, 2, \dots, n),$$

dove le variabili casuali $V_i \sqrt{w_i}$ sono indipendenti a media nulla e varianza σ^2 . Si è così condotti ad un sistema che può essere affrontato mediante il principio dei minimi quadrati. Quest'ultimo richiede la minimizzazione della funzione

$$\sum_{i=1}^n [y_i \sqrt{w_i} - (a \sqrt{w_i} + b x_i \sqrt{w_i})]^2 = \sum_{i=1}^n w_i [y_i - (a + b x_i)]^2,$$

il cui minimo coincide con quello della funzione (6.100) relativa ai minimi quadrati pesati.

L'interpretazione geometrica del metodo dei minimi quadrati (cfr. § 6.2) può essere ora estesa al metodo dei minimi quadrati pesati. In questo caso una volta acquisite le coppie di dati (x_i, y_i) , $(i = 1, 2, \dots, n)$, con $n \geq 3$, occorre individuare i valori di a e b per i quali

$$a \sqrt{w_i} + b x_i \sqrt{w_i} = y_i \sqrt{w_i} \quad (i = 1, 2, \dots, n).$$

In forma matriciale questo sistema diventa

$$WA\theta = Wy, \quad (6.107)$$

dove

$$W = \begin{bmatrix} \sqrt{w_1} & 0 & \cdots & 0 \\ 0 & \sqrt{w_2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sqrt{w_n} \end{bmatrix}.$$

e dove A , θ e y sono definite nella (6.32). Il sistema (6.107) ammette soluzione quando il vettore Wy appartiene allo spazio delle colonne di WA ; se ciò non si verifica occorre determinare una soluzione approssimata $\hat{\theta}$ che minimizzi l'errore $\|WA\hat{\theta} - Wy\|$, ossia occorre ricercare il punto $WA\hat{\theta}$ a distanza minima da Wy . Procedendo come per i minimi quadrati non pesati, si impone che il vettore errore $WA\hat{\theta} - Wy$ sia perpendicolare a WAz per ogni scelta di z . Deve allora aversi:

$$(WAz)^T (WA\hat{\theta} - Wy) = 0,$$

ossia:

$$z^T (A^T W^T WA \hat{\theta} - A^T W^T Wy) = 0. \quad (6.108)$$

Poiché la (6.108) deve sussistere per ogni scelta di z , il vettore in parentesi si deve annullare; si giunge così al sistema

$$A^T W^T WA \hat{\theta} = A^T W^T Wy. \quad (6.109)$$

La soluzione $\hat{\theta}$ della (6.109) è anche la migliore soluzione approssimata del sistema (6.107). Se le colonne di WA sono indipendenti, ossia se non risulta $x_1 = x_2 = \dots = x_n$, la matrice quadrata $A^T W^T WA$ è invertibile, ed allora il sistema (6.109) ammette soluzione unica. Osservando che risulta

$$A^T W^T WA = \begin{bmatrix} \sum_{i=1}^n w_i & \sum_{i=1}^n w_i x_i \\ \sum_{i=1}^n w_i x_i & \sum_{i=1}^n w_i x_i^2 \end{bmatrix} \quad A^T W^T Wy = \begin{bmatrix} \sum_{i=1}^n w_i y_i \\ \sum_{i=1}^n w_i x_i y_i \end{bmatrix},$$

si riconosce che il sistema (6.109) coincide con quello costituito delle equazioni normali del metodo dei minimi quadrati pesati, così che le rispettive soluzioni sono identiche.

Il metodo dei minimi quadrati pesati può apparire poco utile dal momento che per la sua utilizzazione si richiede la conoscenza delle varianze $D^2(Y_i)$ a meno di una costante moltiplicativa. In realtà esso può essere applicato laddove il modello in esame consenta di valutare tali varianze a partire dai valori osservati x_i . Ad esempio, in molteplici situazioni la varianza delle variabili Y_i risulta dipendere linearmente da x_i ; in tali casi in luogo delle (6.99) si ha:

$$E(Y_i) = a + b x_i \quad D^2(Y_i) = \sigma^2 x_i \quad (i = 1, 2, \dots, n),$$

con a , b e σ^2 parametri incogniti e con x_1, x_2, \dots, x_n reali positivi. Si è così ricondotti al caso dei minimi quadrati pesati a patto di porre $w_i = 1/x_i$ nella seconda delle (6.99). Occorre allora determinare il minimo della funzione

$$Q(a, b) \equiv \sum_{i=1}^n \frac{[y_i - E(Y_i)]^2}{D^2(Y_i)} = \frac{1}{\sigma^2} \sum_{i=1}^n \frac{[y_i - (a + b x_i)]^2}{x_i},$$

che coincide con la (6.100) per $w_i = 1/x_i$. Le stime di a e b in questo caso si ricavano ponendo $w_i = 1/x_i$ nelle (6.101) e (6.102):

$$\hat{a} = \frac{\sum_{i=1}^n \frac{y_i}{x_i} - n \hat{b}}{\sum_{i=1}^n \frac{1}{x_i}}, \quad \hat{b} = \frac{\sum_{j=1}^n \frac{1}{x_j} \sum_{i=1}^n y_i - n \sum_{j=1}^n \frac{y_j}{x_j}}{\sum_{j=1}^n \frac{1}{x_j} \sum_{i=1}^n x_i - n^2}. \quad (6.110)$$

Esempio 6.9.2 Siano $\{Y_i = Y_i(t), t \geq 0\}$ ($i = 1, 2, \dots, n$) processi di Wiener di deriva η e varianza infinitesimale σ^2 aventi origine nello stato $Y_i(0) = a$, $a \in \mathbb{R}$. Pertanto, per $i = 1, 2, \dots, n$ e per ogni $t > 0$ sussiste quanto segue:

Tabella 6.7: Osservazione di 10 processi di Wiener.

i	1	2	3	4	5	6	7	8	9	10
t_i	0.7	0.3	0.8	0.2	0.4	1.3	0.1	1.1	0.6	1.2
y_i	0.6	1.8	2.2	1.9	1.1	-0.8	1.8	0.3	2.1	3.2

- la variabile casuale $Y_i(t)$ ha distribuzione normale;
- $Y_i(t)$ ha incrementi stazionari indipendenti;
- $P[Y_i(0) = a] = 1$;
- $E[Y_i(t)] = a + bt$, $D^2[Y_i(t)] = \sigma^2 t$.

Supponiamo che vengano fissati n istanti t_1, t_2, \dots, t_n e corrispondentemente registrati i valori y_1, y_2, \dots, y_n rispettivamente assunti dalle variabili casuali $Y_1(t_1), Y_2(t_2), \dots, Y_n(t_n)$. Si chiede di stimare lo stato iniziale a e la deriva η a partire dalle coppie (t_i, y_i) di dati registrati. In virtù delle ipotesi di cui sopra risulta possibile fare a tal fine uso del metodo dei minimi quadrati pesati, e quindi utilizzare le (6.110) sostituendo \hat{b} con la stima $\hat{\eta}$ di η e x_i con t_i :

$$\hat{a} = \frac{\sum_{i=1}^n \frac{y_i}{t_i} - n \hat{\eta}}{\sum_{i=1}^n \frac{1}{t_i}}, \quad \hat{\eta} = \frac{\sum_{j=1}^n \frac{1}{t_j} \sum_{i=1}^n y_i - n \sum_{j=1}^n \frac{y_j}{t_j}}{\sum_{j=1}^n \frac{1}{t_j} \sum_{i=1}^n t_i - n^2}. \quad (6.111)$$

Nella Tabella 6.7 sono riportati i dati registrati in un esperimento consistente nell'osservazione di 10 processi di Wiener del tipo suddetto, per i quali risulta

$$\sum_i t_i = 6.7 \quad \sum_i y_i = 14.2 \quad \sum_i \frac{1}{t_i} = 27.69 \quad \sum_i \frac{y_i}{t_i} = 45.681.$$

Facendo uso delle (6.111) si ottengono le stime della deriva $\hat{\eta}$ e del valore iniziale \hat{a} :

$$\hat{\eta} = \frac{27.69 \cdot 14.2 - 10 \cdot 45.681}{27.69 \cdot 6.7 - 10^2} = -0.7438,$$

$$\hat{a} = \frac{45.681 + 10 \cdot 0.7438}{27.69} = 1.9183.$$

La retta di regressione stimata ai minimi quadrati pesati è dunque

$$\hat{y} = 1.9183 - 0.7438t. \quad (6.112)$$

In Figura 6.11 tale retta è rappresentata insieme con i dati di cui alla Tabella 6.7. Si noti che se, erroneamente, si facesse uso del metodo dei minimi quadrati non pesati la retta di regressione stimata sarebbe

$$\hat{y} = 1.9932 - 0.8556x$$

che differisce sensibilmente dalla (6.112).

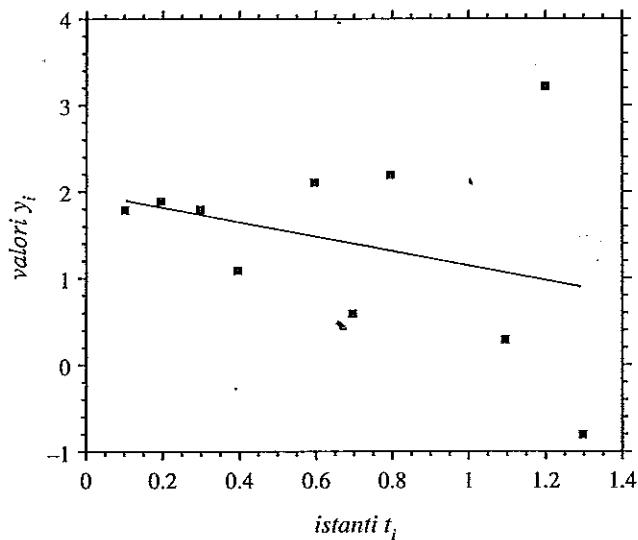


Figura 6.11: I dati di cui alla Tabella 6.5 e la relativa retta interpolante.

6.10 Regressione polinomiale

Finora si è visto che in numerose situazioni è possibile applicare il metodo dei minimi quadrati per determinare la retta $\hat{y} = \hat{a} + \hat{b}x$ che meglio approssima le coppie osservate (x_i, y_i) , ($i = 1, 2, \dots, n$). Analizzando poi l'indice di fit e i residui standardizzati si cerca di stabilire se la retta ai minimi quadrati fornisce un'approssimazione soddisfacente. Nei casi in cui ciò non accade è necessario valutare se sia preferibile adoperare un diverso modello di regressione, che sia dunque non lineare. Ad esempio, come visto nel § 6.3, talora occorre adottare equazioni di regressione del tipo $E(Y|x) = c d^x$ oppure $E(Y|u) = c u^b$. In altri contesti si può ricorrere ad equazioni di regressione di tipo polinomiale, ossia del tipo

$$E(Y|x) = a + b_1 x + b_2 x^2 + \dots + b_k x^k.$$

In questo caso vanno stimati i coefficienti di regressione a, b_1, b_2, \dots, b_k al fine di individuare la curva di regressione stimata

$$\hat{y} = \hat{a} + \hat{b}_1 x + \hat{b}_2 x^2 + \dots + \hat{b}_k x^k \quad (6.113)$$

che meglio approssima i dati (x_i, y_i) . Anche nel caso di questo modello è possibile adoperare il principio dei minimi quadrati per ricavare le stime dei coefficienti di regressione; queste

6.10. REGRESSIONE POLINOMIALE

sono costituite dai valori per i quali è minima la funzione

$$Q(a, b_1, b_2, \dots, b_k) \stackrel{\text{def}}{=} \sum_{i=1}^n [y_i - (a + b_1 x_i + b_2 x_i^2 + \dots + b_k x_i^k)]^2. \quad (6.114)$$

Per determinare il minimo della (6.114) imponiamo che si annullino le sue derivate parziali rispetto ai coefficienti di regressione; si ottiene così il seguente sistema di equazioni normali

$$\begin{cases} \sum_{i=1}^n y_i = a n + b_1 \sum_{i=1}^n x_i + b_2 \sum_{i=1}^n x_i^2 + \dots + b_k \sum_{i=1}^n x_i^k \\ \sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 + b_2 \sum_{i=1}^n x_i^3 + \dots + b_k \sum_{i=1}^n x_i^{k+1} \\ \sum_{i=1}^n x_i^2 y_i = a \sum_{i=1}^n x_i^2 + b_1 \sum_{i=1}^n x_i^3 + b_2 \sum_{i=1}^n x_i^4 + \dots + b_k \sum_{i=1}^n x_i^{k+2} \\ \vdots \quad \vdots \\ \sum_{i=1}^n x_i^k y_i = a \sum_{i=1}^n x_i^k + b_1 \sum_{i=1}^n x_i^{k+1} + b_2 \sum_{i=1}^n x_i^{k+2} + \dots + b_k \sum_{i=1}^n x_i^{2k}, \end{cases} \quad (6.115)$$

la cui soluzione conduce alle stime $\hat{a}, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_k$ e agli stimatori $\hat{A}, \hat{B}_1, \hat{B}_2, \dots, \hat{B}_k$ corrispondenti.

Nell'approssimazione dei dati mediante equazioni di regressione di tipo polinomiale il problema cruciale consiste nella determinazione del grado k del polinomio (6.113). Solitamente conviene identificare k con il minimo intero che consenta di ottenere una buona approssimazione. Notiamo che mentre è sempre possibile individuare un polinomio di grado n che interpoli n coppie (x_i, y_i) , è inverosimile riporre fiducia in una tale approssimazione perché polinomi di grado elevato esibiscono forti oscillazioni. È comunque consigliabile fare uso delle approssimazioni polinomiali del tipo (6.113) per effettuare previsioni soltanto in corrispondenza di valori di x prossimi alle ascisse osservate x_i .

Esempio 6.10.1 Al termine della diciottesima settimana di attività di un'azienda si desidera prevedere quale sarà il profitto che verrà conseguito nel corso della settimana successiva. Si effettua tale previsione sulla base dei dati riportati nella Tabella 6.8 relativi ai profitti y_i (in migliaia di euro) conseguiti dall'azienda in ciascuna settimana x_i . Facendo riferimento ad un modello di regressione lineare $E(Y|x) = a + bx$ occorre ricavare la retta di regressione stimata ai minimi quadrati. Per calcolare le stime dei coefficienti di regressione notiamo che risulta

$$\sum_i x_i = 171, \quad \sum_i x_i^2 = 2109, \quad \sum_i y_i = 234.3, \quad \sum_i x_i y_i = 2323.8. \quad (6.116)$$

Sostituendo questi valori nelle (6.23) e (6.24) si ricava

$$\hat{a} = \frac{1}{18} (234.3 - 0.2022 \cdot 171) = 11.0961$$

$$\hat{b} = \frac{18 \cdot 2323.8 - 171 \cdot 234.3}{18 \cdot 2109 - (171)^2} = 0.2022.$$

Tabella 6.8: Profitti conseguiti dall'azienda in 18 settimane.

settimane	profitti	settimane	profitti
1	12.2	10	12.4
2	11.6	11	13.1
3	12.4	12	12.5
4	12.2	13	13.4
5	11.6	14	14.2
6	12.1	15	14.9
7	11.9	16	14.3
8	12.3	17	14.8
9	12.9	18	15.5

La retta di regressione stimata è pertanto:

$$\hat{y} = 11.0961 + 0.2022x, \quad (6.117)$$

in base alla quale

$$\hat{y}_{19} = 11.0961 + 0.2022 \cdot 19 = 14.9379 \quad (6.118)$$

è il profitto stimato per la diciannovesima settimana. Volendo verificare la validità del modello lineare calcoliamo il valore che assume l'indice di fit. Facendo uso delle (6.116), e osservando che risulta $\sum_i y_i^2 = 3074.89$, si ha:

$$S_{xx} = \sum_i x_i^2 - \frac{1}{n} \left(\sum_i x_i \right)^2 = 2109 - \frac{(171)^2}{18} = 484.5$$

$$S_{yy} = \sum_i y_i^2 - \frac{1}{n} \left(\sum_i y_i \right)^2 = 3074.89 - \frac{(234.3)^2}{18} = 25.085$$

$$S_{xy} = \sum_i x_i y_i - \frac{1}{n} \sum_i x_i \sum_i y_i = 2323.8 - \frac{171 \cdot 234.3}{18} = 97.95.$$

L'indice di fit assume dunque il valore

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{97.95}{\sqrt{484.5 \cdot 25.085}} = 0.8885.$$

Poiché questo è abbastanza prossimo all'unità, non vi sono elementi per ritenere che i dati non siano ben approssimati dalla retta di regressione stimata (6.117). I valori dei residui standardizzati, indicati nella Figura 6.12, presentano però un andamento palesemente asimmetrico rispetto all'asse delle ascisse: si riscontrano infatti valori positivi solo per valori piccoli o per valori grandi di x_i . Tale asimmetria fa sospettare che il modello di regressione lineare non sia il più appropriato per l'approssimazione dei dati osservati; è quindi opportuno verificare se un modello polinomiale lo sia maggiormente. Analizziamo a tal fine il

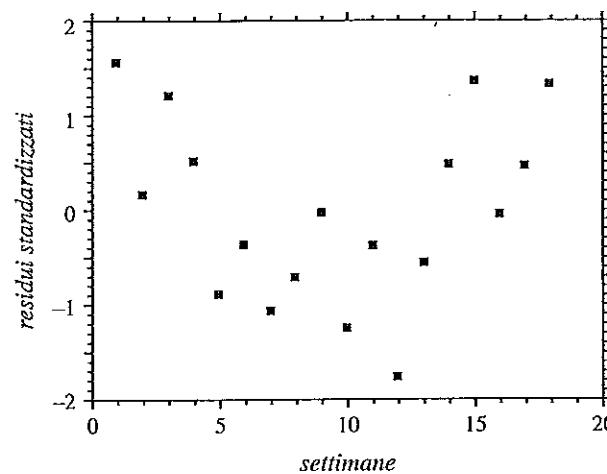


Figura 6.12: Valori dei residui standardizzati nel modello lineare.

modello di regressione polinomiale di grado $k=2$, ossia il modello di regressione quadratico, $E(Y|x) = a + b_1 x + b_2 x^2$. Per ricavare le stime ai minimi quadrati \hat{a} , \hat{b}_1 e \hat{b}_2 occorre risolvere il sistema di equazioni normali (6.115) che nel caso in esame diventa

$$\begin{bmatrix} 18 & 171 & 2109 \\ 171 & 2109 & 29241 \\ 2109 & 29241 & 432345 \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 234.3 \\ 2323.8 \\ 29491.6 \end{bmatrix}.$$

La sua soluzione conduce alle seguenti stime dei coefficienti di regressione:

$$\hat{a} = 12.1892 \quad \hat{b}_1 = -0.1258 \quad \hat{b}_2 = 0.0173,$$

così che

$$\hat{y} = 12.1892 - 0.1258x + 0.0173x^2 \quad (6.119)$$

è la curva di regressione stimata. Usando la (6.119) si deduce il profitto stimato per la diciannovesima settimana:

$$\hat{y}_{19} = 12.1892 - 0.1258 \cdot 19 + 0.0173(19)^2 = 16.0443,$$

che differisce sensibilmente dalla stima (6.118) basata sul modello lineare. Per i dati di cui alla Tabella 6.8 la curva di approssimazione quadratica (6.119) è rappresentata in Figura 6.13 insieme con la retta di regressione stimata (6.117). Come è evidente, il polinomio di secondo grado fornisce un'approssimazione migliore di quella ottenuta mediante la retta ai minimi quadrati. ♦

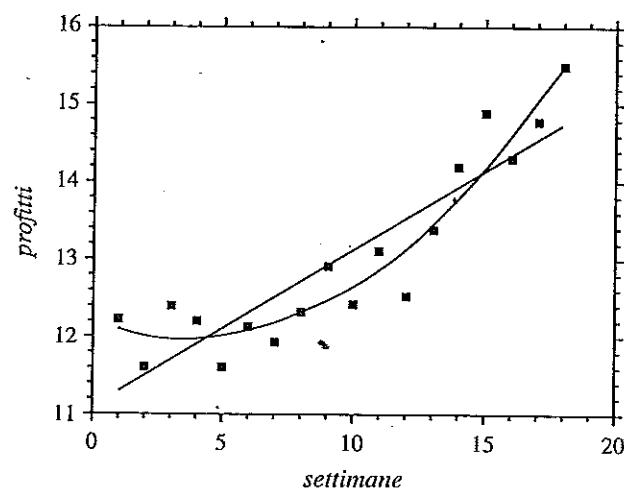


Figura 6.13: Dati di cui alla Tabella 6.7 e relative curve interpolanti.

6.11 Regressione lineare multivariata

Nei paragrafi precedenti sono stati analizzati vari aspetti concernenti il problema della regressione bivariata. Come accennato nel § 6.1, esistono tuttavia numerose situazioni in cui i problemi di regressione coinvolgono più di due variabili casuali. Si ricorre allora alla regressione multivariata che coinvolge la media condizionata $E(Y|x_1, x_2, \dots, x_k)$, ossia il valore medio di Y dato che $X_1 = x_1, X_2 = x_2, \dots, X_k = x_k$, ($k \geq 2$). Esamineremo qui la regressione lineare multivariata in base alla quale i dati osservati $(x_{i1}, x_{i2}, \dots, x_{ik}, y_i)$, ($i = 1, 2, \dots, n$), possono essere riguardati come la realizzazione di un campione casuale la cui variabile genitrice $(k+1)$ -dimensionale, $(X_1, X_2, \dots, X_k, Y)$, possiede componenti caratterizzate dalla relazione

$$Y = a + b_1 X_1 + b_2 X_2 + \dots + b_k X_k + Z, \quad (6.120)$$

dove a, b_1, b_2, \dots, b_k sono parametri reali mentre Z è una variabile casuale a media nulla. L'equazione di regressione di Y dati $X_1 = x_1, X_2 = x_2, \dots, X_k = x_k$ è dunque lineare, avendosi

$$E(Y|x_1, x_2, \dots, x_k) = a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k.$$

Come visto nel caso della regressione bivariata, la (6.120) si può interpretare anche così: supponendo che vi sia una relazione di tipo causa-effetto tra i valori $x_{i1}, x_{i2}, \dots, x_{ik}$ ed i corrispondenti valori y_i , si assume che $x_{i1}, x_{i2}, \dots, x_{ik}$ siano delle costanti mentre l' n -upla (y_1, y_2, \dots, y_n) viene riguardata come realizzazione di un vettore casuale (Y_1, Y_2, \dots, Y_n) così che, per ogni i , il valore y_i assunto dalla variabile Y_i dipende da $x_{i1}, x_{i2}, \dots, x_{ik}$. Si suppone, dunque, che

6.11. REGRESSIONE LINEARE MULTIVARIATA

sussistano le seguenti relazioni:

$$Y_i = a + b_1 x_{i1} + b_2 x_{i2} + \dots + b_k x_{ik} + Z_i \quad (i = 1, 2, \dots, n),$$

dove le variabili casuali Z_1, Z_2, \dots, Z_n sono indipendenti, identicamente distribuite e a media nulla. Le variabili casuali Y_1, Y_2, \dots, Y_n sono pertanto indipendenti ed hanno medie

$$E(Y_i) = a + b_1 x_{i1} + b_2 x_{i2} + \dots + b_k x_{ik} \quad (i = 1, 2, \dots, n).$$

Faremo ora ricorso al principio dei minimi quadrati per ricavare le stime dei coefficienti di regressione $\hat{a}, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_k$, ossia i valori per i quali è minima la funzione

$$Q(a, b_1, b_2, \dots, b_k) \stackrel{\text{def}}{=} \sum_{i=1}^n [y_i - (a + b_1 x_{i1} + b_2 x_{i2} + \dots + b_k x_{ik})]^2. \quad (6.121)$$

Imponendo che si annullino le derivate parziali della (6.121) rispetto ai coefficienti di regressione, si ottiene il sistema di equazioni normali

$$\left\{ \begin{array}{l} \sum_{i=1}^n y_i = a n + b_1 \sum_{i=1}^n x_{i1} + b_2 \sum_{i=1}^n x_{i2} + \dots + b_k \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} y_i = a \sum_{i=1}^n x_{i1} + b_1 \sum_{i=1}^n x_{i1}^2 + b_2 \sum_{i=1}^n x_{i1} x_{i2} + \dots + b_k \sum_{i=1}^n x_{i1} x_{ik} \\ \sum_{i=1}^n x_{i2} y_i = a \sum_{i=1}^n x_{i2} + b_1 \sum_{i=1}^n x_{i2} x_{i1} + b_2 \sum_{i=1}^n x_{i2}^2 + \dots + b_k \sum_{i=1}^n x_{i2} x_{ik} \\ \vdots \quad \vdots \\ \sum_{i=1}^n x_{ik} y_i = a \sum_{i=1}^n x_{ik} + b_1 \sum_{i=1}^n x_{ik} x_{i1} + b_2 \sum_{i=1}^n x_{ik} x_{i2} + \dots + b_k \sum_{i=1}^n x_{ik}^2, \end{array} \right. \quad (6.122)$$

del quale le stime $\hat{a}, \hat{b}_1, \hat{b}_2, \dots, \hat{b}_k$ costituiscono la soluzione; gli stimatori corrispondenti $\hat{A}, \hat{B}_1, \hat{B}_2, \dots, \hat{B}_k$ si ricavano di conseguenza.

Esempio 6.11.1 Un semplice esempio di problema idoneo ad un'analisi di regressione multivariata è offerto dallo studio dell'età e del prezzo di automobili usate. Si possono infatti trattare le età x_{i1} e le cilindrate x_{i2} delle automobili come costanti note, e i prezzi y_i corrispondenti come i valori assunti dalle variabili casuali Y_i . Adottando il modello di regressione lineare multivariata con $k = 2$, si assume allora che le variabili casuali Y_i siano caratterizzate da medie

$$E(Y_i) = a + b_1 x_{i1} + b_2 x_{i2} \quad (i = 1, 2, \dots, n).$$

Consideriamo un caso concreto in cui si fa riferimento a 5 automobili di cui si conoscono le cilindrate ed i prezzi per ciascuno dei primi 4 anni di età. I dati corrispondenti ($n = 20$) sono riportati nella Tabella 6.9 nella quale i prezzi sono espressi in migliaia di euro e le cilindrate in centimetri cubi. Per determinare le stime dei coefficienti di regressione a, b_1, b_2 occorre

Tabella 6.9: Prezzi di 5 automobili in relazione ad età e cilindrata.

età (x_{i1})	cilindrata (x_{i2})				
	903	999	1372	1581	1773
1	10.000	10.900	14.000	20.900	21.250
2	7.825	8.875	11.725	17.475	19.225
3	6.075	7.225	8.575	12.825	16.675
4	5.500	6.700	7.650	11.450	15.500

risolvere il sistema (6.122), che in questo caso diventa

$$\begin{cases} \sum_{i=1}^{20} y_i = a \cdot 20 + b_1 \sum_{i=1}^{20} x_{i1} + b_2 \sum_{i=1}^{20} x_{i2} \\ \sum_{i=1}^{20} x_{i1} y_i = a \sum_{i=1}^{20} x_{i1} + b_1 \sum_{i=1}^{20} x_{i1}^2 + b_2 \sum_{i=1}^{20} x_{i1} x_{i2} \\ \sum_{i=1}^{20} x_{i2} y_i = a \sum_{i=1}^{20} x_{i2} + b_1 \sum_{i=1}^{20} x_{i2} x_{i1} + b_2 \sum_{i=1}^{20} x_{i2}^2. \end{cases} \quad (6.123)$$

Poiché

$$x_{i1} = \begin{cases} 1 & \text{per } i = 1, \dots, 5 \\ 2 & \text{per } i = 6, \dots, 10 \\ 3 & \text{per } i = 11, \dots, 15 \\ 4 & \text{per } i = 16, \dots, 20 \end{cases} \quad x_{i2} = \begin{cases} 903 & \text{per } i = 1, 6, 11, 16 \\ 999 & \text{per } i = 2, 7, 12, 17 \\ 1372 & \text{per } i = 3, 8, 13, 18 \\ 1581 & \text{per } i = 4, 9, 14, 19 \\ 1773 & \text{per } i = 5, 10, 15, 20, \end{cases}$$

ricavando i valori y_i dalla Tabella 6.9 il sistema (6.123) diventa:

$$\begin{bmatrix} 20 & 50 & 26512 \\ 50 & 150 & 66280 \\ 26512 & 66280 & 37355536 \end{bmatrix} \begin{bmatrix} a \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 240.350 \\ 548.625 \\ 345628 \end{bmatrix},$$

la cui soluzione conduce alle stime desiderate:

$$\hat{a} = 1.04437 \quad \hat{b}_1 = -2.09 \quad \hat{b}_2 = 0.01222$$

e quindi alla retta di regressione stimata

$$\hat{y} = 1.04437 - 2.09 x_1 + 0.01222 x_2.$$

Questa può essere utilizzata per stimare il costo di un'automobile di età e cilindrata prefissate. Ad esempio,

$$\hat{y} = 1.04437 - 2.09 \cdot 3 + 0.01222 \cdot 1100 = 8.216$$

è la stima del costo di un'automobile di $x_1 = 3$ anni e cilindrata $x_2 = 1100$. ♦

6.12. CORRELAZIONE NORMALE

6.12 Correlazione normale

Nel § 6.4 si è mostrato che nel contesto della regressione lineare si può procedere in due modi equivalenti: o si riguardano le coppie osservate (x_i, y_i) , ($i = 1, 2, \dots, n$), come la realizzazione di un campione casuale di variabile genitrice bidimensionale (X, Y) , oppure si assume che x_1, x_2, \dots, x_n siano delle costanti e che l' n -upla (y_1, y_2, \dots, y_n) sia la realizzazione di un vettore casuale (Y_1, Y_2, \dots, Y_n) per il quale per ogni i il valore y_i assunto dalla variabile Y_i dipende da x_i . È conveniente adottare il secondo approccio allorché le costanti x_1, x_2, \dots, x_n , fissate dallo sperimentatore, si riferiscono ad una grandezza deterministica; il primo approccio è invece seguito quando i valori x_1, x_2, \dots, x_n sono generati da una variabile casuale X . In quest'ultimo caso all'analisi di regressione va accompagnata la cosiddetta *analisi della correlazione*: proprio a questa è dedicato il presente paragrafo.

Ci si prefigge dunque di rivolgere l'attenzione all'analisi della correlazione normale; questa viene effettuata su coppie di dati (x_i, y_i) ($i = 1, 2, \dots, n$) che si assume costituiscano la realizzazione di un campione casuale estratto da una popolazione normale bivariata di valori medi μ_1 e μ_2 , varianze σ_1^2 e σ_2^2 e coefficiente di correlazione ρ . Si assume, pertanto, che la densità congiunta della variabile genitrice bidimensionale (X, Y) sia

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho)^2} \right. \\ \times \left[\left(\frac{x-\mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x-\mu_1}{\sigma_1} \right) \left(\frac{y-\mu_2}{\sigma_2} \right) + \left(\frac{y-\mu_2}{\sigma_2} \right)^2 \right] \}. \quad (6.124)$$

Per determinare le stime di massima verosimiglianza dei parametri presenti nella (6.124) costruiamo la funzione di verosimiglianza:

$$L(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \prod_{i=1}^n f(x_i, y_i) \\ = \left(\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \right)^n \exp \left\{ -\frac{1}{2(1-\rho)^2} \right. \\ \times \sum_{i=1}^n \left[\left(\frac{x_i-\mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_i-\mu_1}{\sigma_1} \right) \left(\frac{y_i-\mu_2}{\sigma_2} \right) + \left(\frac{y_i-\mu_2}{\sigma_2} \right)^2 \right] \}.$$

Per renderne più agevole la ricerca del massimo, come al solito passiamo ai logaritmi naturali:

$$\ln L(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = -n \ln \left(2\pi\sigma_1\sigma_2\sqrt{1-\rho^2} \right) - \frac{1}{2(1-\rho)^2} \\ \times \sum_{i=1}^n \left[\left(\frac{x_i-\mu_1}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_i-\mu_1}{\sigma_1} \right) \left(\frac{y_i-\mu_2}{\sigma_2} \right) + \left(\frac{y_i-\mu_2}{\sigma_2} \right)^2 \right]. \quad (6.125)$$

Derivando la (6.125) rispetto a μ_1 e a μ_2 e imponendo che le derivate si annullino otteniamo le seguenti equazioni:

$$\frac{\partial}{\partial \mu_1} \ln L = -\frac{1}{(1-\rho)^2 \sigma_1^2} \sum_{i=1}^n \left[-\left(\frac{x_i-\mu_1}{\sigma_1} \right) + \rho \left(\frac{y_i-\mu_2}{\sigma_2} \right) \right] = 0$$

$$\frac{\partial}{\partial \mu_2} \ln L = -\frac{1}{(1-\rho)^2 \sigma_2} \sum_{i=1}^n \left[\rho \left(\frac{x_i - \bar{x}}{\sigma_1} \right) - \left(\frac{y_i - \bar{y}}{\sigma_2} \right) \right] = 0.$$

Da queste segue che le stime di massima verosimiglianza per μ_1 e μ_2 si identificano con le medie campionarie:

$$\hat{\mu}_1 = \bar{x}, \quad \hat{\mu}_2 = \bar{y}.$$

Imponendo poi che le derivate $\partial \ln L / \partial \sigma_1$, $\partial \ln L / \partial \sigma_2$ e $\partial \ln L / \partial \rho$ siano nulle e sostituendo nelle equazioni così ottenute μ_1 e μ_2 con le rispettive stime \bar{x} e \bar{y} , si ricava il sistema di equazioni

$$\begin{aligned} -\frac{n}{\sigma_1} - \frac{1}{(1-\rho)^2 \sigma_1} \sum_{i=1}^n \left[-\left(\frac{x_i - \bar{x}}{\sigma_1} \right)^2 + \rho \left(\frac{x_i - \bar{x}}{\sigma_1} \right) \left(\frac{y_i - \bar{y}}{\sigma_2} \right) \right] &= 0, \\ -\frac{n}{\sigma_2} - \frac{1}{(1-\rho)^2 \sigma_2} \sum_{i=1}^n \left[\rho \left(\frac{x_i - \bar{x}}{\sigma_1} \right) \left(\frac{y_i - \bar{y}}{\sigma_2} \right) - \left(\frac{y_i - \bar{y}}{\sigma_2} \right)^2 \right] &= 0, \\ \frac{n\rho}{1-\rho^2} + \frac{1}{(1-\rho)^2} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma_1} \right) \left(\frac{y_i - \bar{y}}{\sigma_2} \right) - \frac{1}{(1-\rho)^3} \times \sum_{i=1}^n \left[\left(\frac{x_i - \bar{x}}{\sigma_1} \right)^2 - 2\rho \left(\frac{x_i - \bar{x}}{\sigma_1} \right) \left(\frac{y_i - \bar{y}}{\sigma_2} \right) + \left(\frac{y_i - \bar{y}}{\sigma_2} \right)^2 \right] &= 0 \end{aligned}$$

la cui risoluzione conduce alle stime di massima verosimiglianza di σ_1 , σ_2 e ρ :

$$\hat{\sigma}_1 = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\sigma}_2 = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}, \quad (6.126)$$

$$\hat{\rho} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}. \quad (6.127)$$

È interessante osservare che ciascuna delle stime $\hat{\sigma}_1$ e $\hat{\sigma}_2$ date dalla (6.126) coincide con la stima di massima verosimiglianza della deviazione standard σ di una popolazione normale unidimensionale (cfr. Esempio 3.7.10). Le stime ottenute differiscono quindi dalle rispettive deviazioni standard campionarie soltanto per il fattore $\sqrt{(n-1)/n}$, che peraltro tende all'unità al crescere della taglia del campione.

Osserviamo che in virtù delle (6.28), (6.29) e (6.71) le stime (6.126) e (6.127) possono riscriversi nel seguente modo:

$$\hat{\sigma}_1 = \sqrt{\frac{S_{xx}}{n}}, \quad \hat{\sigma}_2 = \sqrt{\frac{S_{yy}}{n}}, \quad \hat{\rho} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}. \quad (6.128)$$

È evidente che la variabile casuale cui corrisponde la stima $\hat{\rho}$ coincide con l'indice di fit R introdotto nella (6.97). Tale indice, nel contesto dell'analisi della correlazione, è detto *coefficiente di correlazione campionario*; il valore che esso assume viene solitamente denotato con

6.12. CORRELAZIONE NORMALE

la lettera r in luogo di $\hat{\rho}$. Il calcolo di r nella pratica si semplifica usando la seguente formula alternativa

$$r = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2}} \quad (6.129)$$

che, come è facile verificare, coincide con la (6.127).

Il coefficiente di correlazione ρ gioca un ruolo fondamentale nell'analisi delle relazioni lineari tra coppie (X, Y) di variabili casuali in quanto numerosi sono i problemi in cui stima di ρ e relativi test di ipotesi rivestono grande interesse. Va qui ricordato che se è $\rho = 0$ le due variabili casuali si dicono scorrelate; se, inoltre, due variabili scorrelate sono congiuntamente normali, esse sono anche indipendenti; se, infine, risulta $|\rho| = 1$, con probabilità unitaria tra le variabili casuali X e Y sussiste una relazione lineare del tipo $Y = a + bX$.

Come mostra la (2.43), se X e Y sono congiuntamente normali la varianza σ^2 di Y dato $X = x$ vale

$$\sigma^2 \stackrel{\text{def}}{=} D^2(Y | X = x) = \sigma_2^2(1 - \rho^2),$$

dove $\sigma_2^2 = D^2(Y)$ e ρ è il coefficiente di correlazione di X e Y . Una relazione analoga sussiste tra le corrispondenti stime di massima verosimiglianza; ricordando, invero, che $\hat{\sigma}^2$, $\hat{\sigma}_2^2$ e r denotano rispettivamente le stime di σ^2 , σ_2^2 e ρ , dalla (6.73), dalla seconda delle (6.30) e dalle (6.128) si trae:

$$\hat{\sigma}^2 = \frac{S_{xx} S_{yy} - S_{xy}^2}{n S_{xx}} = \hat{\sigma}_2^2(1 - r^2). \quad (6.130)$$

Si noti che la formula (6.130) consente di evidenziare un legame esistente tra i concetti di regressione e di correlazione: è infatti evidente che risulta $\hat{\sigma}^2 = 0$ se e solo se è $|r| = 1$, ossia se e solo se i dati (x_i, y_i) giacciono su di una retta; questa ha coefficiente angolare positivo [negativo] quando risulta $r = 1$ [$r = -1$].

Per porre in maggior risalto il significato di r , osserviamo che dalla (6.130) segue

$$r^2 = \frac{\hat{\sigma}_2^2 - \hat{\sigma}^2}{\hat{\sigma}_2^2}.$$

Poiché $\hat{\sigma}_2^2$ costituisce una misura della fluttuazione dei dati y_i mentre $\hat{\sigma}^2$ esprime la fluttuazione condizionata dei dati y_i per valori x_1, x_2, \dots, x_n fissati, la differenza $\hat{\sigma}_2^2 - \hat{\sigma}^2$ rappresenta quella parte di fluttuazione totale dei dati y_i che è attribuibile alla relazione di Y con X . Pertanto r^2 rappresenta la frazione di fluttuazione totale dei dati y_i che è dovuta alla relazione sussistente tra X e Y . Ad esempio per $r = 0.5$ una percentuale pari a un quarto della fluttuazione dei dati y_i è giustificata dalla relazione di Y con X ; quando è invece $r = 0.7$, tale percentuale diventa 0.49. Una correlazione corrispondente ad $r = 0.7$ è quindi di "intensità" quasi doppia rispetto ad una correlazione relativa a $r = 0.5$; analogamente, una correlazione relativa a $r = 0.6$ ha "intensità" nove volte maggiore rispetto ad una correlazione corrispondente a $r = 0.2$, avendosi $(0.6)^2 = 9(0.2)^2$.

Per la rilevanza del ruolo che il coefficiente di correlazione campionario R svolge, appare utile determinarne la distribuzione che, peraltro, risulta essere piuttosto complicata; è allora

conveniente ricorrere ad una approssimazione. A tal fine facciamo uso della circostanza (qui affermata, per brevità, senza dimostrazione) che per n grande la statistica

$$U = \frac{1}{2} \ln \frac{1+R}{1-R}$$

è approssimativamente normale con valore medio e varianza⁴

$$E(U) = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \quad D^2(U) = \frac{1}{n-3}.$$

Pertanto la variabile casuale

$$Z = \frac{\frac{1}{2} \ln \frac{1+R}{1-R} - \frac{1}{2} \ln \frac{1+\rho}{1-\rho}}{\sqrt{\frac{1}{n-3}}} = \frac{\sqrt{n-3}}{2} \ln \frac{(1+R)(1-\rho)}{(1-R)(1+\rho)} \quad (6.131)$$

ha approssimativamente distribuzione normale standard. Tale approssimazione risulta quindi utile per costruire test atti alla verifica di ipotesi del tipo $\rho = \rho_0$, oppure per determinare intervalli fiduciali per il parametro ρ . Vedremo ora come sia possibile costruire un test per verificare l'ipotesi che il coefficiente di correlazione ρ sia nullo.

Proposizione 6.12.1 Nella correlazione normale un test di ampiezza approssimativamente α per verificare l'ipotesi nulla $H_0: \rho = 0$ contro l'ipotesi alternativa $H_1: \rho \neq 0$ è quello che ha regione critica

$$\mathcal{C} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) : |z| \geq z_{\alpha/2}\}, \quad (6.132)$$

dove z è il valore assunto dalla variabile casuale

$$\frac{\sqrt{n-3}}{2} \ln \frac{1+R}{1-R}. \quad (6.133)$$

Dim. Come visto in precedenza, la variabile casuale (6.131) ha approssimativamente distribuzione normale standard. Pertanto sotto l'ipotesi nulla $H_0: \rho = 0$ la variabile casuale (6.133) ha approssimativamente distribuzione normale standard, così che sussiste la seguente relazione:

$$P(|Z| \geq z_{\alpha/2} | \rho = 0) = P\left(\left|\frac{\sqrt{n-3}}{2} \ln \frac{1+R}{1-R}\right| \geq z_{\alpha/2}\right) \approx \alpha.$$

La regione critica (6.132) ha quindi all'incirca ampiezza α .

Esaminiamo un esempio di applicazione della Proposizione 6.12.1.

Esempio 6.12.1 Si voglia stabilire se esiste una relazione tra il tempo, in minuti, che un autista impiega per percorrere un certo tragitto in mattinata (X) e nel tardo pomeriggio (Y), sulla base dei seguenti dati:

$$\begin{aligned} x &= (35, 40, 32, 37, 35, 41, 34, 38) \\ y &= (38, 42, 30, 38, 36, 40, 36, 42). \end{aligned}$$

⁴Cfr. A. Stuart and J.K. Ord (1987), *Kendall's Advanced Theory of Statistics*. Volume I. §16.33. Charles Griffin & Company Limited, London.

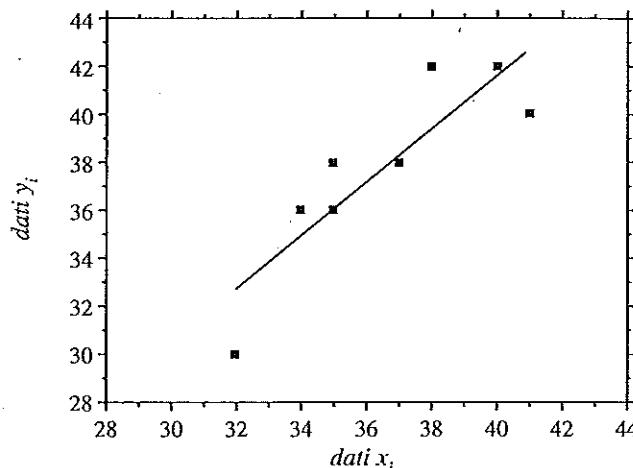


Figura 6.14: Dati dell'Esempio 6.12.1.

Si può anzitutto effettuare un test per verificare se esiste correlazione tra i dati x_i e y_i . Si assume come ipotesi nulla $H_0: \rho = 0$ contro l'ipotesi alternativa $H_1: \rho \neq 0$; poiché $n = 8$, $\sum x_i = 292$, $\sum x_i^2 = 10724$, $\sum y_i = 302$, $\sum y_i^2 = 11508$ e $\sum x_i y_i = 11096$, la (6.129) dà

$$r = \frac{8 \cdot 11096 - 292 \cdot 302}{\sqrt{8 \cdot 10724 - (292)^2} \sqrt{8 \cdot 11508 - (302)^2}} = 0.8666.$$

Questo risultato sta ad indicare che esiste una correlazione positiva molto marcata tra i tempi impiegati dall'autista in mattinata e nel tardo pomeriggio. Ciò appare anche evidente dalla Figura 6.14 nella quale sono riportati i dati osservati e la retta interpolante

$$\hat{y} = -2.6212 + 1.1061x.$$

Poiché risulta $r^2 = 0.7510$, possiamo affermare che all'incirca il 75% della fluttuazione dei dati y_i è dovuta ad una relazione lineare tra le variabili X e Y che li hanno generati. Onde effettuare poi il test per verificare se esiste correlazione tra i dati x_i e y_i , sceglieremo la regione critica \mathcal{C} come specificato dalla (6.132) e fissiamo un ragionevole valore dell'ampiezza, ponendo ad esempio $\alpha = 0.04$. La regione critica è pertanto l'insieme costituito dalle coppie (x_i, y_i) tali che $|z| \geq z_{0.02}$. Poiché si ha $n = 8$ e $r = 0.8666$, la (6.133) assume il valore

$$\frac{\sqrt{5}}{2} \ln \frac{1.8666}{0.1334} = 2.9499.$$

Questo eccede, dunque, il valore $z_{0.02} = 2.055$ (cfr. Tabella 1 dell'Appendice B), così che l'ipotesi nulla di non correlazione deve essere rifiutata, concludendosi che esiste una relazione

tra il tempo impiegato dall'autista in mattinata e quello impiegato nel tardo pomeriggio. Va comunque osservato che questo test ha solo valore indicativo a causa dell'approssimazione normale su cui poggia.

È opportuno rilevare la seguente significativa proprietà di cui gode il coefficiente di correlazione campionario, che discende direttamente dalla (6.129). Denotiamo con $r_{x,y}$ il valore assunto da tale coefficiente in corrispondenza di n coppie di dati (x_i, y_i) , $(i = 1, 2, \dots, n)$; se si pone $u_i = ax_i + b$ e $v_i = cy_i + d$ con a e c reali positivi e con b e d reali qualsiasi, non è difficile mostrare che il valore $r_{u,v}$ assunto dal coefficiente di correlazione campionario in corrispondenza delle coppie (u_i, v_i) , $(i = 1, 2, \dots, n)$, coincide con $r_{x,y}$. Il coefficiente di correlazione campionario è dunque invariante rispetto a trasformazioni lineari, in analogia con la proprietà che anche il coefficiente di correlazione è invariante rispetto a siffatte trasformazioni avendosi $\rho(X, Y) = \rho(ax + b, cy + d)$.

Da quanto esposto finora appare evidente che i limiti dell'analisi della correlazione normale risiedono nella circostanza che si tratta di una tecnica idonea soltanto nei casi in cui le coppie osservate di dati (x_i, y_i) vanno riguardate come realizzazioni di un campione casuale di variabile genitrice bidimensionale normale. Ciò nonostante i risultati cui tale tecnica conduce possono essere talora estesi, effettuando delle approssimazioni opportune, ai casi in cui la variabile genitrice bidimensionale è solo approssimativamente normale. Recentemente sono stati comunque sviluppati dei metodi per l'analisi della correlazione che non poggiano su ipotesi semplificatrici sulla distribuzione del campione. Tali metodi, basati sull'uso massiccio dei moderni sistemi di calcolo elettronico, risultano spesso efficaci nel trattare situazioni in cui non è sostenibile l'ipotesi normale o in cui non è facile verificare se tale ipotesi è accettabile a causa, ad esempio, della scarsità dei dati disponibili.

È bene però precisare che, qualunque sia la tecnica che si adopera, può risultare possibile individuare la presenza di correlazione tra due variabili casuali ma non certo individuare la causa di tale legame. Invero, in generale, le tecniche statistiche basate sull'osservazione di dati sono molto limitate nell'evidenziare le relazioni causali tra le grandezze esaminate. Così, se appare che X e Y sono variabili correlate, le tecniche di analisi della correlazione non consentono di stabilire se la causa di tale correlazione sia intrinseca nella natura delle variabili X e Y o se, ad esempio, sia conseguenza dell'effetto di una terza variabile che le influenza entrambe. Per chiarire questa affermazione, supponiamo di aver riscontrato la presenza di correlazione positiva tra il numero di visite mediche effettuate in una città e lo status socioeconomico degli individui che vi risiedono. Possono individuarsi tre spiegazioni plausibili: la prima è che la presenza di un numero elevato di visite mediche fa sì che gli individui siano maggiormente sani, e quindi più efficienti nel loro lavoro, risultandone conseguentemente innalzati i guadagni e quindi il loro status socioeconomico; la seconda è che proprio gli individui caratterizzati da status elevato hanno possibilità di sottoporsi a frequenti visite mediche; la terza è che un'ulteriore variabile condiziona indipendentemente il numero di visite mediche e lo status socioeconomico: ad esempio città di grandi dimensioni presumibilmente offrono un numero elevato di centri medici e opportunità lavorative meglio remunerate.

Capitolo 7

Analisi della varianza

7.1 Introduzione

Nelle Proposizioni 5.3.5 e 5.3.6 è stato trattato il problema della verifica di ipotesi riguardanti medie di due popolazioni. Affronteremo ora un problema a questo collegato consistente nel fornire un metodo per stabilire se le eventuali differenze sussistenti tra i valori osservati delle medie campionarie di più popolazioni possono essere attribuite al caso o se siano, invece, imputabili ad effettive significative differenze tra le medie delle popolazioni esaminate. Ad esempio si potrebbe voler stabilire, sulla base di realizzazioni osservate, se differenze reali caratterizzino le durate di funzionamento di alcuni componenti elettronici o le quantità di merci prodotte da aziende in competizione. Naturalmente le differenze osservate potrebbero trarre origine da fattori diversificati e non presi in considerazione. Così, con riferimento ai precedenti esempi, le diversità delle durate di funzionamento di taluni componenti elettronici potrebbero imputarsi a disparità dei carichi di lavoro cui essi sono stati sottoposti, mentre le differenze nelle quantità delle merci prodotte potrebbe trovare giustificazione nelle differenti condizioni aziendali e socio-economiche delle rispettive aree geografiche.

Prendendo spunto da quanto testé menzionato, saranno qui ora esposte delle considerazioni concernenti il cosiddetto *piano degli esperimenti* con la finalità di fornire indicazioni e strumenti utili alla individuazione delle reali cause di origine delle eventuali diversità dei risultati registrati nell'osservazione di dati.

7.2 Esperimenti ad un fattore

Iniziamo col riferirci ai cosiddetti *esperimenti ad 1 fattore*. In generale, un problema che coinvolge esperimenti ad un fattore può essere schematizzato mediante k campioni casuali indipendenti di identiche taglie estratti da altrettante popolazioni. Indichiamo con X_{ij} ($i = 1, 2, \dots, k$; $j = 1, 2, \dots, n$) la j -esima componente del campione casuale i -esimo e denotiamo con x_{ij} la realizzazione osservata della variabile casuale X_{ij} . Le k realizzazioni osservate

possono dunque essere così schematizzate:

$$\begin{aligned} \text{popolazione 1: } & x_{11}, x_{12}, \dots, x_{1j}, \dots, x_{1n}; \\ \text{popolazione 2: } & x_{21}, x_{22}, \dots, x_{2j}, \dots, x_{2n}; \\ & \vdots \\ \text{popolazione } i: & x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{in}; \\ & \vdots \\ \text{popolazione } k: & x_{k1}, x_{k2}, \dots, x_{kj}, \dots, x_{kn}. \end{aligned}$$

Assumeremo d'ora innanzi che le $n \cdot k$ variabili casuali caratterizzanti i k campioni sono indipendenti e dotate di distribuzioni normali con medie in generale differenti ma con varianze uguali. Indicheremo quindi con μ_i la media di X_{ij} e con σ^2 la sua varianza. Il modello che descrive le variabili del campione è allora il seguente:

$$X_{ij} = \mu_i + E_{ij} \quad (i = 1, 2, \dots, k; j = 1, 2, \dots, n), \quad (7.1)$$

dove le E_{ij} sono $n \cdot k$ variabili casuali normali indipendenti a medie nulle e varianze σ^2 . Passando ai dati, ossia alle osservazioni x_{ij} , dalle (7.1) si ricava:

$$x_{ij} = \mu_i + e_{ij} \quad (i = 1, 2, \dots, k; j = 1, 2, \dots, n), \quad (7.2)$$

dove con e_{ij} sono stati denotati i valori assunti dalle variabili casuali normali E_{ij} . Il modello (7.2) si riferisce quindi ad una situazione in cui ciascun dato x_{ij} osservato è interpretabile come somma di una componente sistematica μ_i , comune ad ogni elemento del campione i -esimo, e di una componente aleatoria e_{ij} . Notiamo che se si pone

$$\mu = \frac{1}{k} \sum_{i=1}^k \mu_i, \quad (7.3)$$

$$\alpha_i = \mu_i - \mu \quad (i = 1, 2, \dots, k), \quad (7.4)$$

le relazioni (7.1) e (7.2) possono essere riscritte al seguente modo:

$$X_{ij} = \mu + \alpha_i + E_{ij} \quad (i = 1, 2, \dots, k; j = 1, 2, \dots, n) \quad (7.5)$$

$$x_{ij} = \mu + \alpha_i + e_{ij} \quad (i = 1, 2, \dots, k; j = 1, 2, \dots, n). \quad (7.6)$$

Attraverso la (7.3) μ è definita come media aritmetica delle medie μ_i delle k popolazioni, ed è pertanto denominata *media generale* delle k popolazioni; le (7.4) introducono invece per ogni popolazione i lo scarto α_i tra la sua media e la media generale. Notiamo che dalle (7.3) e (7.4) segue

$$\sum_{i=1}^k \alpha_i = 0, \quad (7.7)$$

esprimente l'annullarsi della variazione totale tra le medie.

Il modello (7.6) costituisce uno schema più dettagliato di quello relativo al modello (7.2) in quanto i dati x_{ij} osservati sono ora decomposti nella somma di tre termini: una componente

7.2. ESPERIMENTI AD UN FATTORE

sistematica comune a tutte le osservazioni, una componente sistematica comune a tutti gli elementi di ciascuna realizzazione ed, infine, una componente aleatoria.

Per approfondire il significato della (7.5) sarà qui opportuno svolgere alcune considerazioni. Anzitutto, ricordiamo che le variabili che intervengono in un campione casuale possono riguardarsi come descriventi l'esito di esperimenti casuali indipendenti ripetuti nelle stesse condizioni. I campioni casuali relativi alle k popolazioni sopra introdotte possono allora essere reinterpretati come descriventi k serie ciascuna di n esperimenti effettuati sulla medesima popolazione. Ogni siffatta serie di esperimenti è dunque costituita da n esperimenti indipendenti ripetuti in condizioni identiche, mentre può immaginarsi che gli esperimenti relativi a serie differenti siano effettuati in condizioni diverse. Tale situazione può essere descritta affermando che l'esito degli esperimenti dipende da più fattori uno dei quali, il cosiddetto *fattore sperimentale*, viene fatto variare. In altri termini, ciascuno dei k campioni casuali fa riferimento a esperimenti compiuti in corrispondenza di uno dei k differenti *livelli* relativi al fattore sperimentale. Ad esempio, consideriamo una sperimentazione in cui si esamina la produzione per unità di superficie coltivata di grano in relazione a k diversi tipi di fertilizzante adoperato. Ogni tipo di fertilizzante viene utilizzato in n differenti superfici coltivate, in modo da ottenere in totale $n \cdot k$ sperimentazioni. In questo caso il fattore sperimentale, ossia quello che viene fatto variare, è il fertilizzante, ed i livelli sono identificabili nei k diversi tipi di fertilizzante esaminato. I fattori diversi dal fattore sperimentale, detti *fattori subsperimentali*, vengono lasciati invariati nel corso dell'effettuazione degli esperimenti. Così, nell'esempio di cui sopra l'umidità, la composizione, la posizione delle superfici coltivate, ecc. sono fattori che rimangono invariati nel corso dell'esperimento e quindi costituiscono i fattori subsperimentali. La variabile dipendente che si esamina nell'effettuare gli esperimenti prende il nome di *risposta*, mentre le combinazioni dei livelli dei fattori coinvolti negli esperimenti vengono dette *trattamenti*. Nell'esempio della produzione agraria, la risposta è la produzione di grano; se ad esempio come fattore sperimentale si assumesse anche la temperatura, i trattamenti sarebbero tutti gli abbinamenti possibili tra tipo di fertilizzante e temperatura. La problematica legata alla progettazione di un esperimento viene infine indicata come *piano degli esperimenti*. Questo piano fa riferimento alla scelta delle varie grandezze in gioco quali la risposta, i fattori, i livelli, il numero di osservazioni per ogni trattamento. Il problema cruciale che ci si pone consiste nello stabilire se esiste una relazione tra il fattore sperimentale (o i fattori sperimentali, se sono più d'uno) e la variabile risposta. La metodologia che solitamente si utilizza per affrontare tale problema costituisce la cosiddetta *analisi della varianza*, o ANOVA, acronimo della terminologia inglese *ANALYSIS OF VARIANCE*.

Passiamo ora ad esaminare l'analisi della varianza per esperimenti caratterizzati da un solo fattore sperimentale.

Alla luce delle considerazioni su poste è possibile interpretare la (7.6) in maniera più esplicita. Ciascuna osservazione x_{ij} può infatti essere espressa come somma di tre termini: il primo coincide con la media generale μ , che non è altro che il valore atteso della *media generale delle osservazioni*

$$\bar{X}_{..} = \frac{1}{nk} \sum_{i=1}^k \sum_{j=1}^n X_{ij}, \quad (7.8)$$

avendosi

$$E(\bar{X}_{..}) = \frac{1}{k} \sum_{i=1}^k E\left(\frac{1}{n} \sum_{j=1}^n X_{ij}\right) = \frac{1}{k} \sum_{i=1}^k \mu_i \equiv \mu, \quad (7.9)$$

il secondo termine, α_i , è associato alle osservazioni relative al livello i del fattore sperimentale; il terzo termine, infine, tiene conto delle variazioni dovute al caso. Notiamo, inoltre, che dalla (7.5) segue:

$$E(X_{ij}) = \mu + \alpha_i \quad (i = 1, 2, \dots, k; j = 1, 2, \dots, n), \quad (7.10)$$

dalla quale è evidente che α_i fornisce una misura dell'*effetto* del livello i del fattore sperimentale sul valore atteso di X_{ij} .

Uno dei problemi maggiormente significativi da affrontare nell'ambito dell'analisi della varianza consiste nello stabilire se l'effetto dei vari trattamenti conduce a differenze significative tra le medie μ_i . Si può pertanto sottoporre a verifica l'ipotesi che le medie μ_i siano uguali tra loro contro l'ipotesi che esse non siano tutte coincidenti. In virtù delle (7.3) e (7.4) si tratta di verificare le seguenti ipotesi:

$$H_0: \alpha_i = 0 \quad \text{per } i = 1, 2, \dots, k,$$

$$H_1: \alpha_i \neq 0 \quad \text{per almeno un valore di } i.$$

Notiamo che l'ipotesi nulla equivale ad affermare che le $n \cdot k$ variabili casuali X_{ij} sono caratterizzate da distribuzioni normali di medie μ e varianze σ^2 . È ragionevole assumere che la verifica dell'ipotesi nulla si effettui mediante una valutazione di quanto nel loro complesso tali variabili si discostano dalla media generale delle osservazioni effettuate. Si ricorre invero alla seguente statistica:

$$SS_T = \sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_{..})^2, \quad (7.11)$$

detta *somma totale dei quadrati*, dove $\bar{X}_{..}$ denota la media generale delle osservazioni introdotta nella (7.9). Se l'ipotesi nulla è vera, la variabilità presente nei dati osservati è dovuta esclusivamente al caso; se invece essa è falsa, parte della variabilità è imputabile alle differenze tra le medie μ_i .

La proposizione che segue mostra come sia possibile individuare due significativi termini nella (7.11). A tal fine notiamo preliminarmente che, posto

$$\bar{X}_{i..} = \frac{1}{n} \sum_{j=1}^n X_{ij} \quad (i = 1, 2, \dots, k), \quad (7.12)$$

la (7.8) può riscriversi nella forma

$$\bar{X}_{..} = \frac{1}{k} \sum_{i=1}^k \bar{X}_{i..}. \quad (7.13)$$

Proposizione 7.2.1 *Si ha:*

$$SS_T = SS_F + SS_E,$$

7.2. ESPERIMENTI AD UN FATTORE

con

$$SS_F = n \sum_{i=1}^k (\bar{X}_{i..} - \bar{X}_{..})^2 \quad (7.14)$$

$$SS_E = \sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_{i..})^2. \quad (7.15)$$

Dim. Dalla (7.11) segue:

$$\begin{aligned} SS_T &= \sum_{i=1}^k \sum_{j=1}^n [(\bar{X}_{i..} - \bar{X}_{..}) - (X_{ij} - \bar{X}_{i..})]^2 \\ &= n \sum_{i=1}^k (\bar{X}_{i..} - \bar{X}_{..})^2 - 2 \sum_{i=1}^k (\bar{X}_{i..} - \bar{X}_{..}) \sum_{j=1}^n (X_{ij} - \bar{X}_{i..}) \\ &\quad + \sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_{i..})^2 \\ &= SS_F + SS_E, \end{aligned}$$

dove l'ultima uguaglianza segue in virtù delle definizioni (7.14) e (7.15) e della identità $\sum_{j=1}^n (\bar{X}_{i..} - X_{ij}) = 0$ valida per ogni i , conseguenza della (7.12).

Notiamo che la variabile casuale SS_F è una somma di quadrati esprimente la variabilità tra i k gruppi di osservazioni, mentre SS_E è una somma di quadrati che traduce la variabilità esistente all'interno dei k gruppi. Da qui discende il significato dei pedici F ed E , rispettivamente: "F" sta per "fattore sperimentale", mentre "E" sta per "errore all'interno dei gruppi".

7.2.1 Stima puntuale dei parametri

Verrà ora esaminato il problema della stima puntuale dei parametri coinvolti nell'analisi della varianza per esperimenti ad un fattore.

Con riferimento al modello (7.10) facciamo ricorso al metodo dei minimi quadrati per ricavare le stime dei parametri $\mu, \alpha_1, \alpha_2, \dots, \alpha_k$. Come illustrato nel § 6.2, tale metodo richiede la minimizzazione della funzione

$$Q(\mu, \alpha) = \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \mu - \alpha_i)^2,$$

dove $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$. Imponendo l'annullarsi delle derivate di Q rispetto alle $k+1$

variabili $\mu, \alpha_1, \alpha_2, \dots, \alpha_k$ si ricava il seguente sistema di equazioni normali:

$$\left\{ \begin{array}{l} \sum_{i=1}^k \sum_{j=1}^n (\bar{x}_{ij} - \mu - \alpha_i) = 0 \\ \sum_{j=1}^n (\bar{x}_{1j} - \mu - \alpha_1) = 0 \\ \vdots \\ \sum_{j=1}^n (\bar{x}_{kj} - \mu - \alpha_k) = 0, \end{array} \right.$$

che in virtù della (7.7) diventa:

$$\left\{ \begin{array}{l} \mu = \frac{1}{nk} \sum_{i=1}^k \sum_{j=1}^n x_{ij} \\ \alpha_1 = \frac{1}{n} \sum_{j=1}^n x_{1j} - \mu \\ \vdots \\ \alpha_k = \frac{1}{n} \sum_{j=1}^n x_{kj} - \mu. \end{array} \right.$$

Lo stimatore ai minimi quadrati della media μ è dunque

$$\hat{\mu} = \bar{X}_{..}, \quad (7.16)$$

ossia coincide con la media generale delle osservazioni, mentre lo stimatore di α_i è dato da

$$\hat{\alpha}_i = \bar{X}_{i..} - \bar{X}_{..} \quad (i = 1, 2, \dots, k). \quad (7.17)$$

Si osservi che la correttezza dello stimatore $\hat{\mu}$ segue direttamente dalla (7.9), mentre la correttezza di $\hat{\alpha}_i$ si ricava notando che per la (7.10) risulta

$$E(\bar{X}_{i..}) - E(\bar{X}_{..}) = \frac{1}{n} \sum_{j=1}^n (\mu + \alpha_i) - \mu = \alpha_i \quad (i = 1, 2, \dots, k).$$

Resta ora da determinare uno stimatore della varianza σ^2 . A tal fine notiamo che le variabili X_{11}, \dots, X_{in} costituiscono un campione casuale tratto da una popolazione normale di varianza σ^2 ; pertanto, ricordando il Teorema 2.1.4, per $i = 1, 2, \dots, k$ la variabile casuale

$$\frac{1}{\sigma^2} \sum_{j=1}^n (\bar{x}_{ij} - \bar{X}_{i..})^2$$

ha distribuzione chi-quadrato con $n - 1$ gradi di libertà cosicché, per l'indipendenza dei k campioni casuali, grazie al Teorema 2.1.2 si conclude che

$$\frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^n (\bar{x}_{ij} - \bar{X}_{i..})^2 \equiv \frac{SS_E}{\sigma^2} \quad (7.18)$$

7.2. ESPERIMENTI AD UN FATTORE

è una variabile chi-quadrato avente $k(n - 1)$ gradi di libertà. Poiché la media di una variabile casuale chi-quadrato uguaglia il numero di gradi di libertà (cfr. (2.7)), la variabile (7.18) ha media $k(n - 1)$, e dunque

$$\frac{SS_E}{k(n - 1)} \equiv \frac{1}{k(n - 1)} \sum_{i=1}^k \sum_{j=1}^n (\bar{x}_{ij} - \bar{X}_{i..})^2 \quad (7.19)$$

è uno stimatore corretto di σ^2 .

È opportuno sottolineare che gli stimatori (7.16) e (7.17) ottenuti mediante il metodo dei minimi quadrati sono validi anche nel caso in cui i k campioni casuali sono tratti da popolazioni che non sono normali; ciò in quanto il procedimento sopra adottato per ottenerli non richiede la specificazione della distribuzione del campione. L'ipotesi di normalità è invece fondamentale per la determinazione dello stimatore (7.19) di σ^2 .

7.2.2 Verifica di ipotesi

Come già accennato, un problema fondamentale nell'analisi della varianza consiste nel sottoporre a verifica l'ipotesi che le medie delle k popolazioni sono uguali, il che corrisponde ad ipotizzare che si abbia $\alpha_i = 0$ per ogni i . Assumendo questa come ipotesi nulla, esaminiamo come sia possibile sottoporla a verifica.

Osserviamo anzitutto che in generale la variabile $\bar{X}_{i..}$ ($i = 1, 2, \dots, k$) ha distribuzione normale di media μ_i e varianza σ^2/n . Pertanto, allorché l'ipotesi nulla è vera le k variabili $\bar{X}_{i..}$ sono indipendenti e identicamente distribuite, così che per il Teorema 2.1.4

$$\frac{n}{\sigma^2} \sum_{i=1}^k (\bar{X}_{i..} - \bar{X}_{..})^2 \equiv \frac{SS_F}{\sigma^2} \quad (7.20)$$

è una variabile casuale chi-quadrato con $k - 1$ gradi di libertà. Da ciò segue che, sotto l'ipotesi nulla, la variabile

$$\frac{SS_F}{k-1} \equiv \frac{n}{k-1} \sum_{i=1}^k (\bar{X}_{i..} - \bar{X}_{..})^2 \quad (7.21)$$

è uno stimatore corretto di σ^2 . Ciò non è però vero se l'ipotesi nulla è falsa. È allora opportuno determinare il valore medio della variabile (7.21) nel caso più generale.

Proposizione 7.2.2 Risulta:

$$E\left(\frac{SS_F}{k-1}\right) = \sigma^2 + \frac{n}{k-1} \sum_{i=1}^k \alpha_i^2. \quad (7.22)$$

Dim. Analizziamo la variabile SS_F/n . Dalla (7.14) segue

$$\frac{SS_F}{n} = \sum_{i=1}^k (\bar{X}_{i..} - \bar{X}_{..})^2 = \sum_{i=1}^k (\bar{X}_{i..}^2 - 2\bar{X}_{i..}\bar{X}_{..} + \bar{X}_{..}^2) = \sum_{i=1}^k \bar{X}_{i..}^2 - k\bar{X}_{..}^2. \quad (7.23)$$

Notiamo ora che $\bar{X}_{i..}$ ha media μ_i e varianza σ^2/n , così che

$$E(\bar{X}_{i..}^2) = [E(\bar{X}_{i..})]^2 + D^2(\bar{X}_{i..}) = \mu_i^2 + \frac{\sigma^2}{n}. \quad (7.24)$$

Ricordando la (7.9) e osservando che

$$D^2(\bar{X}_{..}) = \frac{1}{k^2} \sum_{i=1}^k D^2(\bar{X}_{i..}) = \frac{\sigma^2}{kn},$$

si ha:

$$E(\bar{X}_{..}^2) = [E(\bar{X}_{..})]^2 + D^2(\bar{X}_{..}) = \mu^2 + \frac{\sigma^2}{kn}. \quad (7.25)$$

Pertanto, facendo uso delle relazioni (7.23)–(7.25), si ottiene:

$$\begin{aligned} E\left(\frac{SS_F}{n}\right) &= \sum_{i=1}^k E(\bar{X}_{i..}^2) - kE(\bar{X}_{..}^2) \\ &= \sum_{i=1}^k \left(\mu_i^2 + \frac{\sigma^2}{n}\right) - k\left(\mu^2 + \frac{\sigma^2}{kn}\right) \\ &= (k-1)\frac{\sigma^2}{n} + \sum_{i=1}^k \mu_i^2 - k\mu^2 \\ &= (k-1)\frac{\sigma^2}{n} + \sum_{i=1}^k (\mu_i - \mu)^2. \end{aligned}$$

Da questa, ricordando la (7.4), si ricava immediatamente la (7.22). ■

La (7.22) indica che il valore medio di $SS_F/(k-1)$ è maggiore di σ^2 quando l'ipotesi nulla è falsa, mentre è pari a σ^2 quando l'ipotesi nulla è vera. Invece, come visto nel § 7.2.1, la media di $SS_E/[k(n-1)]$ è sempre pari a σ^2 . Ciò suggerisce di rifiutare l'ipotesi nulla $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$ quando il valore assunto da $SS_F/(k-1)$ è apprezzabilmente maggiore di quello assunto da $SS_E/[k(n-1)]$. Tale criterio viene formalizzato nel teorema che segue.

Teorema 7.2.1 *Nell'analisi della varianza di popolazioni normali una regione critica di ampiezza α per rifiutare l'ipotesi nulla $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$ è costituita dalle realizzazioni tali che*

$$F \stackrel{\text{def}}{=} \frac{k(n-1)SS_F}{(k-1)SS_E} \quad (7.26)$$

assume un valore maggiore di $F_{\alpha;k-1,k(n-1)}$, con $F_{\alpha;n_1,n_2}$ definita nella (2.26).

Dim. Dal Teorema 2.1.4 segue che per $i = 1, 2, \dots, k$ le variabili casuali $\bar{X}_{i..}$ e

$$S_i^2 \stackrel{\text{def}}{=} \sum_{j=1}^n \frac{(X_{ij} - \bar{X}_{i..})^2}{n-1} \quad (7.27)$$

7.2. ESPERIMENTI AD UN FATTORE

sono indipendenti in quanto media campionaria e varianza campionaria, rispettivamente, del campione casuale normale (X_{i1}, \dots, X_{in}) . Le variabili

$$\frac{n-1}{\sigma^2} \sum_{i=1}^k S_i^2 \equiv \frac{SS_E}{\sigma^2}$$

e

$$\frac{n}{\sigma^2} \sum_{i=1}^k (\bar{X}_{i..} - \bar{X}_{..})^2 \equiv \frac{SS_F}{\sigma^2},$$

che coincidono rispettivamente con le (7.18) e (7.20), sono pertanto indipendenti ed hanno distribuzioni chi-quadrato con gradi di libertà $k(n-1)$ e $k-1$, rispettivamente, quando è vera l'ipotesi nulla. Pertanto, per il Teorema 2.3.1 la variabile casuale

$$\frac{SS_F/[\sigma^2(k-1)]}{SS_E/[\sigma^2 k(n-1)]}$$

ha distribuzione di Fisher con $k-1$ e $k(n-1)$ gradi di libertà. Ne segue che, in virtù della (2.26), la diseguaglianza

$$\frac{k(n-1)SS_F}{(k-1)SS_E} > F_{\alpha;k-1,k(n-1)}$$

è soddisfatta con probabilità α quando è vera l'ipotesi nulla; pertanto essa definisce una regione critica del test d'ampiezza α .

Gli elementi significativi dell'analisi della varianza per esperimenti ad un fattore sono riassunti nella Tabella 7.1, che prende talora il nome di Tabella ANOVA; in essa le medie dei quadrati denotano semplicemente le variabili casuali somme dei quadrati (che, come si è detto, quando è vera l'ipotesi nulla hanno distribuzione chi-quadrato se divise per σ^2) divise per il corrispondente numero di gradi di libertà.

Sarà ora opportuno evidenziare come sia possibile esprimere le somme dei quadrati SS_F e SS_E in una forma computazionalmente più conveniente.

Tabella 7.1: Tabella ANOVA per esperimenti ad un fattore.

Fonti di variabilità	Somme dei quadrati	Gradi di libertà	Medie dei quadrati	F
Fattore	SS_F	$k-1$	$\frac{SS_F}{k-1}$	$\frac{k(n-1)SS_F}{(k-1)SS_E}$
Errore	SS_E	$k(n-1)$	$\frac{SS_E}{k(n-1)}$	
Totale	SS_T	$kn-1$		

Proposizione 7.2.3 Le somme dei quadrati SS_F e SS_E possono essere così espresse:

$$SS_F = n \sum_{i=1}^k \bar{X}_{i..}^2 - nk \bar{X}_{..}^2, \quad (7.28)$$

$$SS_E = \sum_{i=1}^k \sum_{j=1}^n X_{ij}^2 - n \sum_{i=1}^k \bar{X}_{i..}^2. \quad (7.29)$$

Dim. Dalla (7.14) segue:

$$\begin{aligned} SS_F &= n \sum_{i=1}^k (\bar{X}_{i..} - \bar{X}_{..})^2 \\ &= n \sum_{i=1}^k \bar{X}_{i..}^2 - 2n\bar{X}_{..} \sum_{i=1}^k \bar{X}_{i..} + nk\bar{X}_{..}^2 \\ &= n \sum_{i=1}^k \bar{X}_{i..}^2 - nk\bar{X}_{..}^2, \end{aligned}$$

ossia la (7.28). La (7.15) fornisce poi la (7.29) avendosi:

$$\begin{aligned} SS_E &= \sum_{i=1}^k \sum_{j=1}^n (\bar{X}_{ij} - \bar{X}_{i..})^2 \\ &= \sum_{i=1}^k \sum_{j=1}^n X_{ij}^2 - 2 \cdot \sum_{i=1}^k \bar{X}_{i..} \sum_{j=1}^n \bar{X}_{ij} + n \sum_{i=1}^k \bar{X}_{i..}^2 \\ &= \sum_{i=1}^k \sum_{j=1}^n X_{ij}^2 - n \sum_{i=1}^k \bar{X}_{i..}^2. \end{aligned}$$

Notiamo che le (7.28) e (7.29) conducono alla seguente espressione della somma dei quadrati SS_T :

$$SS_T = SS_F + SS_E = \sum_{i=1}^k \sum_{j=1}^n X_{ij}^2 - nk\bar{X}_{..}^2. \quad (7.30)$$

Forniremo ora un esempio che illustrerà come applicare i risultati dell'analisi della varianza finora considerati.

Esempio 7.2.1 Tre tipi di condensatori vengono esaminati al fine di determinare le intensità delle forze elettromotrici che li danneggiano. Qui di seguito sono elencate le intensità, espresse in preassegnate unità, che hanno causato il danneggiamento in 5 condensatori di ciascuno dei tre diversi tipi:

- tipo A: 42, 36, 41, 32, 39;
- tipo B: 30, 34, 27, 33, 31;
- tipo C: 40, 39, 35, 42, 41.

7.2. ESPERIMENTI AD UN FATTORE

Si desidera verificare se vanno considerati coincidenti i valori medi μ_1, μ_2, μ_3 delle intensità delle forze elettromotrici che causano il danneggiamento dei condensatori di tipo A, B e C rispettivamente. Ricordando le definizioni (7.3) e (7.4) si sottopongono allora a verifica le seguenti ipotesi:

$$H_0: \alpha_1 = \alpha_2 = \alpha_3 = 0$$

$H_1: \alpha_i \neq 0$ per almeno un valore di i .

Assumiamo valida l'approssimazione normale e ricorriamo al Teorema 7.2.1 per verificare tali ipotesi. Scegliendo $\alpha = 0.01$ come ampiezza della regione critica, essendo $k = 3$ e $n = 5$ detta regione critica è costituita dalle realizzazioni tali che la variabile casuale

$$F \equiv 6 \frac{SS_F}{SS_E} \quad (7.31)$$

assume un valore maggiore di $F_{0.01,2,12}$ che, come indicato dalla Tabella 5 dell'Appendice B, vale 6.93. Per stabilire se le realizzazioni osservate appartengono alla regione critica calcoliamo i valori che assumono le somme dei quadrati SS_F e SS_E . Dai dati in nostro possesso risulta che le medie relative ai tre livelli del fattore sperimentale assumono i valori

$$\bar{x}_{1..} = 38 \quad \bar{x}_{2..} = 31 \quad \bar{x}_{3..} = 39, \quad (7.32)$$

così che $\bar{x}_{..} = 36$ è il valore assunto dalla media generale delle osservazioni. Ricordando le (7.14) e (7.15) si ricavano poi i valori delle somme dei quadrati SS_F e SS_E :

$$\begin{aligned} n \sum_{i=1}^k (\bar{x}_{i..} - \bar{x}_{..})^2 &= 180 \\ \sum_{i=1}^k \sum_{j=1}^n (\bar{x}_{ij} - \bar{x}_{i..})^2 &= 126. \end{aligned}$$

La variabile (7.31) assume pertanto il valore 8.57, così che le realizzazioni osservate appartengono alla regione critica. L'ipotesi che i valori medi delle intensità delle forze elettromotrici critiche per i tre tipi di condensatori coincidono va dunque rifiutata. I calcoli relativi all'analisi effettuata sono riassunti nella seguente tabella:

Fonte di variabilità	Somma dei quadrati	Gradi di libertà	Media dei quadrati	F
Fattore	180	2	180/2 = 90	90/10.5 = 8.57
Errore	126	12	126/12 = 10.5	
Totale	306	14		

Come preannunciato all'inizio del § 7.1, il problema dell'analisi della varianza per esperimenti ad un fattore può essere riguardato come un'estensione del problema della verifica di ipotesi riguardanti medie di due differenti popolazioni di varianze incognite, già trattato nel § 5.3. Infatti, sussiste il seguente risultato:

Osservazione 7.2.1 Per $k = 2$ il test dell'analisi della varianza fornito dal Teorema 7.2.1 coincide con quello fornito dalla Proposizione 5.3.6 per $n = m$ e $\omega = 0$.

Dim. Essendo $k = 2$, risulta

$$\bar{X}_{..} = \frac{\bar{X}_1 + \bar{X}_2}{2},$$

e pertanto dalla (7.14) si trae

$$2SS_F = 2n[(\bar{X}_1 - \bar{X}_{..})^2 + (\bar{X}_2 - \bar{X}_{..})^2] = n(\bar{X}_1 - \bar{X}_2)^2. \quad (7.33)$$

Inoltre, dalla (7.15) e dalla (7.27) segue

$$\frac{SS_E}{n-1} = \sum_{i=1}^2 \sum_{j=1}^n \frac{(X_{ij} - \bar{X}_{i..})^2}{n-1} = S_1^2 + S_2^2. \quad (7.34)$$

Facendo poi uso delle (7.33) e (7.34) si ricava che la variabile (7.26) si esprime come

$$F = \frac{2(n-1)SS_F}{SS_E} = \frac{n(\bar{X}_1 - \bar{X}_2)^2}{S_1^2 + S_2^2}. \quad (7.35)$$

Così, dal Teorema 7.2.1 si ha che nell'analisi della varianza di due popolazioni normali una regione critica di ampiezza α per rifiutare l'ipotesi nulla $H_0: \mu_1 = \mu_2$ è costituita dalle realizzazioni tali che la (7.35) assume un valore maggiore di $F_{\alpha;1,2(n-1)}$. Dalla Proposizione 5.3.6 segue inoltre che una regione critica per rifiutare la medesima ipotesi è costituita dalle realizzazioni per cui

$$T \stackrel{\text{def}}{=} \sqrt{n} \frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{S_1^2 + S_2^2}} \quad (7.36)$$

assume un valore maggiore di $t_{\alpha/2,2(n-1)}$. Poiché dalle (7.35) e (7.36) segue che $F = T^2$, per mostrare che le due regioni critiche coincidono occorre mostrare che risulta

$$F_{\alpha;1,2(n-1)} = t_{\alpha/2,2(n-1)}^2. \quad (7.37)$$

Ricordiamo che, come visto nelle dimostrazioni del Teorema 7.2.1 e della Proposizione 5.3.6, nelle ipotesi di popolazioni normali le variabili (7.35) e (7.36) hanno rispettivamente distribuzione di Fisher con 1 e $2(n-1)$ gradi di libertà e di Student con $2(n-1)$ gradi di libertà. La relazione (7.37) è allora un'immediata conseguenza della Proposizione 2.3.4, dove si è visto che se X è una variabile casuale di Student con v gradi di libertà, allora la variabile $Y = X^2$ ha distribuzione di Fisher con 1 e v gradi di libertà, così che in generale risulta $F_{\alpha;1,v} = t_{\alpha/2,v}^2$. ■

Per concludere è opportuno osservare che i risultati qui illustrati sono facilmente estendibili al caso di k campioni casuali di taglie n_1, n_2, \dots, n_k arbitrarie. In tal caso è invece facile verificare che risulta:

$$\begin{aligned} SS_T &= \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i..})^2 = \sum_{i=1}^k n_i (\bar{X}_i - \bar{X}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{X}_{ij} - \bar{X}_{i..})^2 \\ &\equiv SS_F + SS_E, \end{aligned}$$

7.3. IL PIANO DEGLI ESPERIMENTI

dove ora la (7.12) va sostituita con

$$\bar{X}_{i..} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \quad (i = 1, 2, \dots, k),$$

mentre $\bar{X}_{..}$ è sempre espresso dalla (7.13). Inoltre, analogamente a quanto visto nel Teorema 7.2.1, nel caso di popolazioni normali una regione critica di ampiezza α per rifiutare l'ipotesi nulla che le medie delle k popolazioni coincidono è costituita dalle realizzazioni tali che

$$F \equiv \frac{(N-k)SS_F}{(k-1)SS_E}$$

assume valore maggiore di $F_{\alpha;k-1,N-k}$, avendo posto $N = n_1 + n_2 + \dots + n_k$.

7.3 Il piano degli esperimenti

Le tecniche dell'analisi della varianza, come visto nel § 7.2, consentono di stabilire se le differenze tra le medie campionarie di vari campioni casuali sono troppo ingenti per poter essere attribuite al caso. È importante però sottolineare che queste tecniche non chiariscono l'origine di tali differenze. A titolo esemplificativo, notiamo che nell'Esempio 7.2.1 si è giunti alla conclusione che i valori medi delle intensità delle forze elettromotrici critiche dei tre tipi di condensatori non coincidono, ma non ne è stato spiegato il motivo. Osservando i valori medi sperimentali (7.32) si può essere indotti a ritenere che l'intensità della forza elettromotrice critica dei condensatori di tipo B sia sensibilmente inferiore a quella dei condensatori di tipo A e C; eppure, potrebbe accaduto che gli esperimenti condotti sui tre tipi di condensatori siano stati effettuati in condizioni differenti, ad esempio in diverse condizioni di temperatura ambientale o di umidità, e che proprio la diversità di queste condizioni sia all'origine del rifiuto dell'ipotesi nulla.

In generale, se si desidera mostrare che un certo fattore tra gli altri può essere considerato causa di un certo fenomeno osservato, occorre in qualche modo sincerarsi che nessuno degli altri fattori possa ritenersi responsabile dei risultati sperimentali. Ciò può essere ottenuto in vari modi. Ad esempio si può effettuare un esperimento rigorosamente controllato in cui tutte le variabili sono tenute fisse tranne quelle di interesse. In tal caso il risultato cui si perviene è rigoroso, ma non è detto che esso fornisca proprio l'informazione desiderata; invero, potrebbe verosimilmente accadere che durante l'osservazione del fenomeno nella sua evoluzione naturale le variabili estranee non si mantengano spontaneamente fissate. Può essere quindi talvolta preferibile ottenere un risultato che tenga anche conto delle condizioni naturali di svolgimento dei fenomeni. In alternativa ad un esperimento controllato rigorosamente si può condurre un esperimento in cui nessuno dei fattori estranei viene controllato, ma il cui eventuale effetto viene eliminato attraverso una tecnica di casualizzazione, detta anche "randomizzazione" in virtù di un evidente anglicismo. Questa tecnica consiste nel pianificare gli esperimenti in modo che le variazioni causate dai fattori estranei possano essere combinate in modo casuale. Nell'Esempio 7.2.1 ciò corrisponde a scegliere a caso i 5 condensatori di ciascun tipo da testare, e poi determinare a caso l'ordine con cui essi vengono sottoposti all'esame. Quando le variazioni dovute ai fattori estranei non controllati possono essere incluse sotto l'effetto di variazioni casuali, il piano degli esperimenti viene considerato come

completamente randomizzato. È evidente che la randomizzazione protegge dagli effetti dei fattori estranei soltanto in un senso probabilistico, e quindi essa non costituisce una tecnica infallibile. Se ad esempio i 15 condensatori dell'Esempio 7.2.1 vengono testati a mezzo di dispositivi di controllo scelti a caso, potrebbe accadere che i 5 condensatori di uno stesso tipo vengano assegnati ai 5 dispositivi meno precisi, il che evidentemente falserebbe il risultato per quel tipo specifico di condensatori. Per evitare questo genere di effetti si decide spesso di controllare rigorosamente alcuni fattori e di randomizzarne altri, cercando comunque di usare tecniche che siano una via di mezzo fra queste due.

Per introdurre un altro importante concetto concernente il piano degli esperimenti, consideriamo i seguenti dati riferintisi al profitto, in migliaia di euro, conseguito da 4 aziende nel corso di 4 settimane di attività:

azienda 1; 12, 18, 15, 19;

azienda 2: 11, 13, 13, 15;

azienda 3: 11, 18, 17, 22;

azienda 4: 15, 16, 19, 22.

Le medie osservate dei 4 campioni casuali sono

$$\bar{x}_1 = 16, \quad \bar{x}_2 = 13, \quad \bar{x}_3 = 17, \quad \bar{x}_4 = 18, \quad (7.38)$$

da cui segue che la media generale delle osservazioni assume il valore

- 16 -

Effettuando l'analisi della varianza si ottengono i risultati riassunti nella tabella seguente, per $n = 4$ e $k = 4$:

Fonti di variabilità	Somme dei quadrati	Gradi di libertà	Medie dei quadrati	<i>F</i>
Fattore	56	3	18.67	18.67/10.83 = 1.72
Errore	130	12	10.83	
Totale	186	15		

Essendo $F_{0.05;3,12} = 3.49$ (cfr. Tabella 6 dell'Appendice B) ed essendo $F = 1.72$, le realizzazioni osservate non appartengono alla regione critica di ampiezza $\alpha = 0.05$ suggerita dal Teorema 7.2.1. Pertanto l'ipotesi che i profitti medi delle 4 aziende coincidono non può essere rifiutata.

Sebbene l'analisi effettuata suggerisca la validità dell'ipotesi nulla nell'esempio testé considerato, sono comunque presenti elementi in tal senso discordanti. Innanzitutto si rilevano differenze significative tra le medie osservate (7.38), ma sono anche presenti differenze non trascurabili tra i valori osservati nei 4 campioni: nel primo caso questi variano da 12 a 19, negli altri casi da 11 a 15, da 11 a 22, e da 15 a 22. Si può inoltre notare che, in ciascuna delle 4 realizzazioni, dei valori osservati il primo è quello minimo e l'ultimo è quello massimo. Tali considerazioni suggeriscono che le variazioni riscontrate all'interno delle realizzazioni potrebbero essere causate da differenze nelle condizioni operative delle aziende nell'arco

Tabella 7.2: Osservazioni effettuate in un esperimento dipendente da due fattori.

Livello del fattore A	Livello del fattore B						Totale
	1	2	...	j	...	n	
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1n}	$x_{1..}$
2	x_{21}	x_{22}	...	x_{2j}	...	x_{2n}	$x_{2..}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{in}	$x_{i..}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	x_{k1}	x_{k2}	...	x_{kj}	...	x_{kn}	$x_{k..}$
Totale	$x_{..1}$	$x_{..2}$...	$x_{..j}$...	$x_{..n}$	$x_{...}$

delle varie settimane. Se ciò fosse vero, le variazioni dovute a questa causa verrebbero incluse nella somma dei quadrati $\sum S_E$ nell'analisi della varianza; pertanto il denominatore del valore F osservato risulterebbe essere "inflazionato". Questa potrebbe essere la causa del risultato di accettazione dell'ipotesi nulla cui ha condotto l'analisi della varianza dell'esperimento in questione.

Per evitare questo genere di inconvenienti si potrebbe pensare di mantenere fissati i fattori esterni, anche se non sempre ciò conduce ai risultati desiderati. Se, ancora, nell'esempio considerato si limitasse l'analisi dei profitti ad un'unica settimana, non vi sarebbe certezza che la validità delle conclusioni ottenute sussisterebbe anche per altre settimane d'attività. Un'altra possibilità sarebbe quella di variare i fattori estranei deliberatamente su di un ampio spettro di valori di modo che la variazione da essi causata possa essere misurata e quindi eliminata dalla somma dei quadrati SSE . Questo vuol dire che l'esperimento va pianificato in modo che sia possibile effettuare un'*analisi della varianza per esperimenti a due fattori*, in cui la variazione totale dei dati osservati è suddivisa in tre componenti, due delle quali sono dovute ai due fattori sperimentali e la terza all'errore sperimentale o al caso. Nel seguente paragrafo sarà esaminata in dettaglio tale tecnica d'indagine.

7.4 Esperimenti a due fattori

Nel trattare esperimenti dipendenti da due fattori è necessario tenere conto del fatto che questi possono essere indipendenti oppure in mutua interazione. Nel seguito esamineremo in dettaglio il solo caso di fattori indipendenti.

Con riferimento alla terminologia introdotta nel § 7.2, indichiamo con *fattore A* e *fattore B* i due fattori da cui dipendono le osservazioni di un certo esperimento. Rappresentiamo nella Tabella 7.2 le osservazioni effettuate, assumendo d'ora innanzi che il fattore *A* sia caratterizzato da k livelli, il fattore *B* da n livelli e che vi sia una singola osservazione per ogni combinazione dei livelli dei due fattori.¹ Le variabili casuali X_{ij} descriventi le osservazioni vengono

¹Risultati analoghi a quelli esposti nel seguito si ottengono anche nel caso in cui vi siano più osservazioni per ogni combinazione dei livelli.

supposte a distribuzione normale di medie μ_{ij} e varianza σ^2 . Per il modello dell'analisi della varianza a due fattori (senza interazione), in analogia con la (7.5) si postula ora la seguente relazione:

$$X_{ij} = \mu + \alpha_i + \beta_j + E_{ij} \quad (i = 1, 2, \dots, k; j = 1, 2, \dots, n). \quad (7.39)$$

Qui

$$\mu \equiv \frac{1}{nk} \sum_{i=1}^k \sum_{j=1}^n \mu_{ij}$$

è la *media generale*, mentre α_i e β_j costituiscono rispettivamente gli *effetti* sulla variabile risposta del livello i del fattore A , e del livello j del fattore B ; come per la (7.5), le E_{ij} sono infine delle variabili casuali normali indipendenti di media nulla e varianza σ^2 . Dalla (7.39) segue poi:

$$E(X_{ij}) \equiv \mu_{ij} = \mu + \alpha_i + \beta_j \quad (i = 1, 2, \dots, k; j = 1, 2, \dots, n). \quad (7.40)$$

Il valore medio di X_{ij} è quindi dipendente da i e da j in una forma tale da far sussistere alcune particolari proprietà che passiamo ad illustrare. Siano dunque j e r due generici livelli del fattore B ; dalla (7.40) si trae allora:

$$\mu_{ij} - \mu_{ir} = \beta_j - \beta_r, \quad (7.41)$$

così che l'effetto di un generico livello i del fattore A sul valore atteso μ_{ij} è costante al variare del livello j del fattore B . Analoga proprietà sussiste per un generico livello j del fattore B : il suo effetto su μ_{ij} rimane costante se il livello i del fattore A viene fatto variare, avendosi

$$\mu_{ij} - \mu_{sj} = \alpha_i - \alpha_s, \quad (7.42)$$

per ogni coppia i, s di livelli del fattore A .

Notiamo che, analogamente alla (7.7) relativa al caso di esperimenti ad un fattore, le quantità α_i e β_j soddisfano ora le relazioni

$$\sum_{i=1}^k \alpha_i = 0, \quad \sum_{j=1}^n \beta_j = 0. \quad (7.43)$$

7.4.1 Stima puntuale dei parametri

In analogia con quanto visto nel § 7.2.1 sarà ora analizzato il problema della stima puntuale dei parametri presenti nell'analisi della varianza per esperimenti a due fattori.

Riferendoci al modello (7.40) utilizziamo il metodo dei minimi quadrati per stimare i parametri $\mu, \alpha_1, \alpha_2, \dots, \alpha_k$ e $\beta_1, \beta_2, \dots, \beta_n$. Di nuovo non sarà necessario far uso dell'ipotesi di normalità delle variabili X_{ij} . Ricordando quanto esposto nel § 6.2, occorre ora minimizzare la funzione

$$Q(\mu, \alpha, \beta) = \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \mu - \alpha_i - \beta_j)^2,$$

dove $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$ e $\beta = (\beta_1, \beta_2, \dots, \beta_n)$. Imponendo che le $n+k+1$ derivate prime di Q si annullino, otteniamo il seguente sistema di equazioni normali:

$$\begin{cases} \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \mu - \alpha_i - \beta_j) = 0 \\ \sum_{j=1}^n (x_{ij} - \mu - \alpha_i - \beta_j) = 0 \quad (i = 1, 2, \dots, k) \\ \sum_{i=1}^k (x_{ij} - \mu - \alpha_i - \beta_j) = 0 \quad (j = 1, 2, \dots, n) \end{cases}$$

che, in vista delle relazioni (7.43), diventa:

$$\begin{cases} \mu = \frac{1}{nk} \sum_{i=1}^k \sum_{j=1}^n x_{ij} \\ \alpha_i = \frac{1}{n} \sum_{j=1}^n x_{ij} - \mu \quad (i = 1, 2, \dots, k) \\ \beta_j = \frac{1}{k} \sum_{i=1}^k x_{ij} - \mu \quad (j = 1, 2, \dots, n). \end{cases}$$

Si ottengono pertanto i seguenti stimatori ai minimi quadrati:

$$\begin{aligned} \hat{M} &= \bar{X}_{..}, \\ \hat{\alpha}_i &= \bar{X}_{i..} - \bar{X}_{..} \quad (i = 1, 2, \dots, k), \\ \hat{\beta}_j &= \bar{X}_{..j} - \bar{X}_{..} \quad (j = 1, 2, \dots, n), \end{aligned}$$

dove si è posto

$$\bar{X}_{..} = \frac{1}{nk} \sum_{i=1}^k \sum_{j=1}^n X_{ij}, \quad \bar{X}_{i..} = \frac{1}{n} \sum_{j=1}^n X_{ij}, \quad \bar{X}_{..j} = \frac{1}{k} \sum_{i=1}^k X_{ij}.$$

È facile rendersi conto che gli stimatori \hat{M} e $\hat{\alpha}_i$ coincidono con quelli forniti dalle (7.16) e (7.17) relativi ad esperimenti a un fattore, la cui correttezza è già stata dimostrata. La correttezza di $\hat{\beta}_j$ segue invece osservando che dalla (7.40) si trae:

$$E(\bar{X}_{..j}) - E(\bar{X}_{..}) = \frac{1}{k} \sum_{i=1}^k (\mu + \alpha_i + \beta_j) - \mu = \beta_j \quad (j = 1, 2, \dots, n).$$

Procedendo in modo analogo al caso di esperimenti ad un fattore si ricava poi che, nell'ipotesi di normalità dei campioni, la (7.19) fornisce uno stimatore corretto di σ^2 anche nel caso di esperimenti a due fattori.

7.4.2 Verifica di ipotesi

Nel caso di esperimenti a due fattori le due ipotesi nulle che si desidera sottoporre a verifica consistono nel supporre che tutti gli effetti α_i del fattore A e tutti gli effetti β_j del fattore B siano nulli. Formalmente si pone

$$\begin{aligned} H_0: \alpha_i &= 0 \quad \text{per } i = 1, 2, \dots, k \\ H'_0: \beta_j &= 0 \quad \text{per } j = 1, 2, \dots, n. \end{aligned}$$

Come alternative si assume invece che gli effetti non sono tutti nulli; si ha quindi

$$\begin{aligned} H_1: \alpha_i &\neq 0 \quad \text{per almeno un valore di } i \\ H'_1: \beta_j &\neq 0 \quad \text{per almeno un valore di } j. \end{aligned}$$

L'analisi della varianza in questo caso è basata sulla decomposizione della somma totale dei quadrati SS_T nella somma di 3 termini, come vedremo nella proposizione seguente, evidente estensione della Proposizione 7.2.1.

Proposizione 7.4.1 Si ha

$$SS_T = SS_A + SS_B + SS_E,$$

dove

$$SS_A = n \sum_{i=1}^k (\bar{X}_{i\cdot} - \bar{X}_{..})^2,$$

$$SS_B = k \sum_{j=1}^n (\bar{X}_{\cdot j} - \bar{X}_{..})^2$$

$$SS_E = \sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{..})^2. \quad (7.44)$$

Dim. Risulta utile decomporre ciascuna variabile X_{ij} nella somma delle 4 variabili i cui rispettivi significati sono sotto ciascuna indicati:

$$\begin{aligned} X_{ij} &= \underbrace{\bar{X}_{..}}_{\substack{\text{media} \\ \text{generale}}} + \underbrace{\bar{X}_{i\cdot} - \bar{X}_{..}}_{\substack{\text{deviazione per} \\ \text{effetto del fattore A}}} + \underbrace{\bar{X}_{\cdot j} - \bar{X}_{..}}_{\substack{\text{deviazione per} \\ \text{effetto del fattore B}}} \\ &\quad + \underbrace{X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{..}}_{\substack{\text{termine residuo} \\ \text{o errore}}} \end{aligned}$$

Dalla (7.11) si ha così:

$$\begin{aligned} SS_T &= \sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_{..})^2 \\ &= \sum_{i=1}^k \sum_{j=1}^n [(\bar{X}_{i\cdot} - \bar{X}_{..}) + (\bar{X}_{\cdot j} - \bar{X}_{..}) + (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{..})]^2. \end{aligned}$$

Sviluppando il quadrato a secondo membro è facile rendersi conto che le somme corrispondenti ai termini prodotto sono nulle, avendosi ad esempio

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^n (\bar{X}_{i\cdot} - \bar{X}_{..})(\bar{X}_{\cdot j} - \bar{X}_{..}) &= \sum_{i=1}^k (\bar{X}_{i\cdot} - \bar{X}_{..}) \sum_{j=1}^n (\bar{X}_{\cdot j} - \bar{X}_{..}) \\ &= (k\bar{X}_{..} - k\bar{X}_{..})(n\bar{X}_{..} - n\bar{X}_{..}) = 0. \end{aligned}$$

Si ha quindi:

$$SS_T = n \sum_{i=1}^k (\bar{X}_{i\cdot} - \bar{X}_{..})^2 + k \sum_{j=1}^n (\bar{X}_{\cdot j} - \bar{X}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{..})^2,$$

da cui segue immediatamente la tesi. ■

La Proposizione 7.4.1 mostra come nell'analisi della varianza per esperimenti a due fattori la somma totale dei quadrati si possa rappresentare come somma di 3 addendi dei quali SS_A esprime la variabilità tra i k livelli del fattore A, SS_B la variabilità tra gli n livelli del fattore B ed, infine, SS_E è il termine legato all'errore.

Notiamo che quando l'ipotesi nulla H_0 è vera le variabili X_{ij} hanno distribuzione normale di media $\mu + \beta_j$ e varianza σ^2 , così che le variabili $\bar{X}_{i\cdot}$ sono indipendenti e identicamente normalmente distribuite con media μ e varianza σ^2/n . Ne segue che, come per SS_F/σ^2 nel caso di un solo fattore, la variabile SS_A/σ^2 ha distribuzione chi-quadrato con $k-1$ gradi di libertà. Consideriamo ora la variabile $X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{..}$; quando l'ipotesi nulla H_0 è vera essa ha valore medio

$$E(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{..}) = \mu + \beta_j - \mu - (\mu + \beta_j) + \mu = 0;$$

inoltre X_{ij} ha varianza σ^2 . La variabile (7.44) è allora espressa come somma di $n \cdot k$ quadrati delle variabili X_{ij} alle quali sono imposti $n+k-1$ vincoli, di cui $k-1$ dovuti a $\bar{X}_{i\cdot}$, $n-1$ dovuti a $\bar{X}_{\cdot j}$ e 1 dovuto a $\bar{X}_{..}$. Conseguentemente,

$$\frac{1}{\sigma^2} \sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{..})^2 \equiv \frac{SS_E}{\sigma^2} \quad (7.45)$$

ha distribuzione chi-quadrato con $nk - (n+k-1) = (n-1)(k-1)$ gradi di libertà.

Quando è vera l'ipotesi nulla H'_0 , le variabili $\bar{X}_{\cdot j}$ sono indipendenti e identicamente distribuite con distribuzione normale di media μ e varianza σ^2/k . Conseguentemente la variabile $k(\bar{X}_{\cdot j} - \bar{X}_{..})^2/\sigma^2$ ha distribuzione chi-quadrato con 1 grado di libertà, e quindi

$$\frac{k}{\sigma^2} \sum_{j=1}^n (\bar{X}_{\cdot j} - \bar{X}_{..})^2 \equiv \frac{SS_B}{\sigma^2}$$

ha distribuzione chi-quadrato con $n-1$ gradi di libertà. In modo analogo al caso di H_0 , quando è vera l'ipotesi nulla H'_0 si può mostrare che la (7.45) ha ancora distribuzione chi-quadrato con $(n-1)(k-1)$ gradi di libertà. I risultati qui ottenuti consentiranno poi di

ricavare le regioni critiche dei test relative alle ipotesi formulate in precedenza per l'analisi della varianza nel caso di due fattori.

In analogia con quanto mostrato nella Proposizione 7.2.2 è ora opportuno ricavare i valori medi delle variabili $\text{SS}_A/(k-1)$ e $\text{SS}_B/(n-1)$ nel caso più generale, ossia quando le ipotesi nulle H_0 e H'_0 non sono necessariamente vere.

Proposizione 7.4.2 Risulta:

$$E\left(\frac{\text{SS}_A}{k-1}\right) = \sigma^2 + \frac{n}{k-1} \sum_{i=1}^k \alpha_i^2,$$

$$E\left(\frac{\text{SS}_B}{n-1}\right) = \sigma^2 + \frac{k}{n-1} \sum_{j=1}^n \beta_j^2.$$

Dim. Procede in modo affatto analogo alla dimostrazione della Proposizione 7.2.2. ■

La proposizione appena enunciata mostra che $\text{SS}_A/(k-1)$ e $\text{SS}_B/(n-1)$ sono stimatori corretti di σ^2 solo se valgono rispettivamente le ipotesi nulle H_0 e H'_0 .

In modo analogo a quanto esposto nel Teorema 7.2.1 siamo ora in grado di formulare il risultato che consente di verificare le ipotesi nulle H_0 e H'_0 per esperimenti a due fattori.

Teorema 7.4.1 Nell'analisi della varianza di popolazioni normali in esperimenti dipendenti da due fattori

(i) una regione critica di ampiezza α per rifiutare l'ipotesi nulla $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$ è costituita dalle realizzazioni tali che

$$F_A \stackrel{\text{def}}{=} \frac{\text{SS}_A/(k-1)}{\text{SS}_E/[(n-1)(k-1)]} = (n-1) \frac{\text{SS}_A}{\text{SS}_E} \quad (7.46)$$

assume un valore maggiore di $F_{\alpha; k-1, (n-1)(k-1)}$;

(ii) una regione critica di ampiezza α per rifiutare l'ipotesi nulla $H'_0: \beta_1 = \beta_2 = \dots = \beta_n = 0$ è costituita dalle realizzazioni tali che

$$F_B \stackrel{\text{def}}{=} \frac{\text{SS}_B/(n-1)}{\text{SS}_E/[(n-1)(k-1)]} = (k-1) \frac{\text{SS}_B}{\text{SS}_E} \quad (7.47)$$

assume un valore maggiore di $F_{\alpha; n-1, (n-1)(k-1)}$.

Dim. La dimostrazione del teorema procede in modo analogo a quella del Teorema 7.2.1, ed è basata essenzialmente sull'osservazione che le variabili casuali (7.46) e (7.47) sotto le ipotesi nulle H_0 e H'_0 hanno distribuzione di Fisher in quanto espresse come rapporti tra variabili casuali chi-quadrato ed i corrispondenti numeri di gradi di libertà. ■

Possiamo rappresentare sinteticamente le quantità significative dell'analisi della varianza per esperimenti a due fattori mediante la Tabella 7.3.

Come nel caso di un solo fattore, per esperimenti a due fattori è possibile valutare le somme dei quadrati in una forma computazionalmente più conveniente, come indicato dalla proposizione che segue.

7.4. ESPERIMENTI A DUE FATTORI

Tabella 7.3: Tabella ANOVA per esperimenti a due fattori.

Fonti di variabilità	Somme dei quadrati	Gradi di libertà	Medie dei quadrati	F
Fattore A	SS_A	$k-1$	$\frac{\text{SS}_A}{k-1}$	$F_A = (n-1) \frac{\text{SS}_A}{\text{SS}_E}$
Fattore B	SS_B	$n-1$	$\frac{\text{SS}_B}{n-1}$	$F_B = (k-1) \frac{\text{SS}_B}{\text{SS}_E}$
Errore	SS_E	$(k-1)(n-1)$	$\frac{\text{SS}_E}{(k-1)(n-1)}$	
Totale	SS_T	$kn-1$		

Proposizione 7.4.3 Le somme dei quadrati per esperimenti a due fattori possono essere così espresse:

$$\text{SS}_A = n \sum_{i=1}^k \bar{X}_{i..}^2 - nk \bar{X}_{..}^2,$$

$$\text{SS}_B = k \sum_{j=1}^n \bar{X}_{.j}^2 - nk \bar{X}_{..}^2,$$

$$\text{SS}_E = \sum_{i=1}^k \sum_{j=1}^n X_{ij}^2 - n \sum_{i=1}^k \bar{X}_{i..}^2 - k \sum_{j=1}^n \bar{X}_{.j}^2 + nk \bar{X}_{..}^2.$$

Dim. Il termine SS_A coincide con il termine SS_F valutato nella (7.28). Il termine SS_B si ricava usando un procedimento analogo a quanto visto per SS_F nella dimostrazione della Proposizione 7.2.3. Per quanto concerne SS_E , notiamo che dalla (7.44) segue:

$$\begin{aligned} \text{SS}_E &= \sum_{i=1}^k \sum_{j=1}^n (X_{ij} - \bar{X}_{i..} - \bar{X}_{.j} + \bar{X}_{..})^2 \\ &= \sum_{i=1}^k \sum_{j=1}^n X_{ij}^2 + n \sum_{i=1}^k \bar{X}_{i..}^2 + k \sum_{j=1}^n \bar{X}_{.j}^2 + nk \bar{X}_{..}^2 \\ &\quad - 2n \sum_{i=1}^k \bar{X}_{i..}^2 - 2k \sum_{j=1}^n \bar{X}_{.j}^2 + 2nk \bar{X}_{..}^2 \\ &\quad + 2nk \bar{X}_{..}^2 - 2nk \bar{X}_{..}^2 - 2nk \bar{X}_{..}^2, \end{aligned}$$

da cui la tesi. ■

Osserviamo che dalle espressioni ricavate nella Proposizione 7.4.3 si trae la relazione

$$\text{SS}_T = \text{SS}_A + \text{SS}_B + \text{SS}_E = \sum_{i=1}^k \sum_{j=1}^n X_{ij}^2 - nk \bar{X}_{..}^2,$$

che coincide con la (7.30) ottenuta nel caso di esperimenti ad un fattore.

Nell'esempio che segue esamineremo nuovamente l'esperimento di cui al § 7.3 discriminando però su due fattori anziché su di uno solo.

Esempio 7.4.1 Con riferimento all'esempio trattato nel § 7.3, prendiamo in esame i dati osservati analizzando la varianza nel caso di due fattori, assumendo come fattore A l'azienda e come fattore B la settimana. Si ha:

$$\bar{x}_{.1} = 12.25, \quad \bar{x}_{.2} = 16.25, \quad \bar{x}_{.3} = 16, \quad \bar{x}_{.4} = 19.5. \quad (7.48)$$

I risultati sono riassunti nella seguente tabella, con $k = 4$ e $n = 4$:

Fonti di variabilità	Somme dei quadrati	Gradi di libertà	Medie dei quadrati	F
Fattore A	56	3	18.67	18.67/2.74 = 6.80
Fattore B	105.31	3	35.10	35.10/2.74 = 12.80
Errore	24.69	9	2.74	
Totalle	186	15		

I valori osservati di F_A e F_B sono dunque 6.80 e 12.80. Per $\alpha = 0.05$, entrambi sono maggiori sia di $F_{\alpha/2, k-1, (n-1)(k-1)}$ che di $F_{\alpha, n-1, (n-1)(k-1)}$ in quanto dalla Tabella 6 dell'Appendice B risulta $F_{0.05; 3, 9} = 3.86$. Le ipotesi sulle $H_0: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$ e $H_0': \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ sono dunque entrambe da rifiutarsi. Ciò vuol dire che le differenze tra i valori medi osservati (7.38) per le 4 aziende sono significative, così come significative sono le differenze tra i valori medi osservati (7.48) per le 4 settimane esaminate. Si noti, tuttavia, che non è possibile trarre la conclusione che i profitti della prima settimana sono necessariamente i peggiori e che quelli della quarta sono sempre i migliori: ciò che l'analisi effettuata consente di affermare è che esistono delle differenze. Per uno studio più dettagliato sulla natura di tali differenze si dovrebbe ricorrere a dei test basati su confronti multipli. ♦

Capitolo 8

Rappresentazione dei dati

8.1 Diagramma delle frequenze

Oggetto di questo capitolo è la descrizione di alcuni metodi comunemente utilizzati per rappresentare dati sperimentali che, come al solito, riguarderemo come realizzazioni di campioni casuali. Parleremo pertanto semplicemente di "realizzazioni", assumendo che si tratti di n -uple di dati generati in maniera indipendente da una variabile casuale (variabile genitrice) oppure da un campione casuale di taglia n .

In questo paragrafo esamineremo il caso in cui la variabile casuale-genitrice X è discreta ed assume valori nell'insieme $\mathcal{D} = \{d_1, d_2, \dots, d_k\}$ costituito da k reali ordinati $d_1 < d_2 < \dots < d_k$. Sia poi (x_1, x_2, \dots, x_n) una particolare realizzazione di un campione casuale (X_1, X_2, \dots, X_n) di taglia n di variabile genitrice X . Adottando una notazione già utilizzata nel § 5.4, introduciamo le variabili casuali N_1, N_2, \dots, N_k rappresentanti rispettivamente le frequenze assolute dei valori d_1, d_2, \dots, d_k presenti nel campione. Se si denota con $p_i = P(X = d_i)$ la probabilità di occorrenza del dato d_i ($i = 1, 2, \dots, k$), il vettore casuale (N_1, N_2, \dots, N_k) ha distribuzione multinomiale:

$$P(N_1 = n_1, N_2 = n_2, \dots, N_k = n_k) = \frac{n!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k},$$

con $n_1, n_2, \dots, n_k \geq 0$ e $n_1 + n_2 + \dots + n_k = n$. Le frequenze assolute N_i sono invero legate dalla relazione $N_1 + N_2 + \dots + N_k = n$. Osserviamo inoltre che, per $i = 1, 2, \dots, k$, risulta:

$$N_i = \left| \{X_j : X_j = d_i; j = 1, 2, \dots, n\} \right| = \sum_{j=1}^n I_{\{d_i\}}(X_j), \quad (8.1)$$

dove I denota la funzione indicatrice definita nella (3.75). La variabile $I_{\{d_i\}}(X_j)$, $j = 1, 2, \dots, n$, ha distribuzione di Bernoulli di parametro p_i ; dalla (8.1) discende quindi che N_i ha distribuzione binomiale di parametri n e p_i , con media e varianza

$$E(N_i) = np_i, \quad D^2(N_i) = np_i(1 - p_i). \quad (8.2)$$

D'altro canto l'estrazione (con rimpiazzamento) di una particolare n -upla (x_1, x_2, \dots, x_n) di dati dall'insieme finito $\mathcal{D} = \{d_1, d_2, \dots, d_k\}$ costituisce un esperimento di prove ripetute e indipendenti in ciascuna delle quali l'estrazione del dato d_i si verifica con probabilità costante $p_i = P(X = d_i)$. Pertanto la frequenza assoluta N_i , esprimente il numero di volte in cui il dato d_i si presenta nel campione casuale (X_1, X_2, \dots, X_n) , è una variabile casuale binomiale. È quindi immediato calcolare valore medio e varianza della variabile casuale¹

$$F_i^{(n)} = \frac{N_i}{n} \quad (i = 1, 2, \dots, k), \quad (8.3)$$

rappresentante la frequenza relativa del dato d_i nel campione; invero, dalle (8.2) e (8.3) segue:

$$E[F_i^{(n)}] = \frac{E(N_i)}{n} = p_i, \quad (8.4)$$

$$D^2[F_i^{(n)}] = \frac{D^2(N_i)}{n^2} = \frac{p_i(1-p_i)}{n}. \quad (8.5)$$

Indichiamo ora con

$$n_i = \left| \{x_j : x_j = d_i; j = 1, 2, \dots, n\} \right| = \sum_{j=1}^n I_{\{d_i\}}(x_j)$$

la frequenza assoluta del dato d_i nella realizzazione, ossia il numero di volte in cui il dato d_i ($i = 1, 2, \dots, k$) compare nella realizzazione (x_1, x_2, \dots, x_n) osservata. Il rapporto

$$f_i^{(n)} = \frac{n_i}{n} \quad (i = 1, 2, \dots, k)$$

costituisce pertanto la frequenza relativa del dato d_i , ossia il rapporto tra il numero di occorrenze di d_i nella realizzazione ed il numero complessivo di dati osservati. La frequenza assoluta n_i e la frequenza relativa $f_i^{(n)}$ denotano dunque, rispettivamente, i generici valori delle variabili casuali N_i e $F_i^{(n)}$.

Per disporre di una rappresentazione grafica dei dati (x_1, x_2, \dots, x_n) osservati si ricorre sovente al cosiddetto *diagramma delle frequenze*. Questo, (cfr. Figura 8.1) consiste in un diagramma cartesiano nel quale sono riportate le frequenze relative $f_i^{(n)}$ in corrispondenza delle ascisse d_i ($i = 1, 2, \dots, k$). Si noti che l'uso delle frequenze relative in luogo di quelle assolute consente di ricondursi sempre all'intervallo $[0, 1]$ qualunque sia il tipo di esperimento che genera i dati e qualunque sia il numero n di questi.

Avendo supposto che i dati sono generati da una variabile casuale X , per la legge empirica del caso² al crescere di n i valori assunti da ciascuna delle frequenze relative $F_i^{(n)}$ si stabilizzano intorno alla probabilità $p_i = P(X = d_i)$, ossia intorno alla probabilità di occorrenza del dato d_i ($i = 1, 2, \dots, k$). Nella proposizione che segue di questa affermazione a carattere empirico viene fornita una precisazione quantitativa. Si dimostra infatti che al crescere del

¹La notazione utilizzata vuole sottolineare la dipendenza dalla taglia n del campione.

²La legge empirica del caso asserisce che in esperimenti casuali consistenti in n prove ripetute la frequenza relativa di un fissato evento al crescere di n si stabilizza intorno ad un valore che è proprio la probabilità di occorrenza di tale evento.

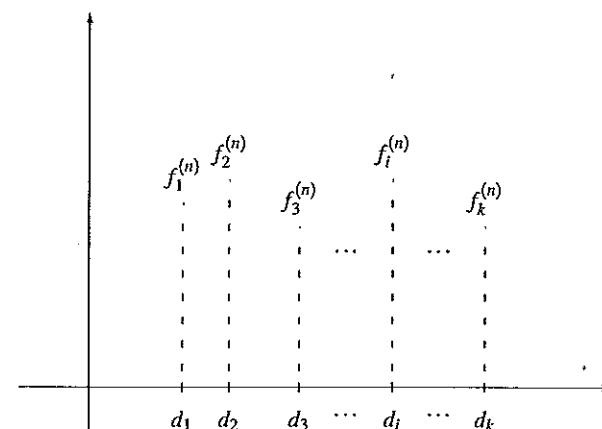


Figura 8.1: Diagramma delle frequenze.

numero n di dati osservati le frequenze relative tendono alle probabilità della variabile casuale genitrice nel senso che discostamenti tra le frequenze relative $F_i^{(n)}$ e le corrispondenti probabilità p_i diventano sempre più improbabili al crescere di n .

Proposizione 8.1.1 Per ogni $\varepsilon > 0$ risulta:

$$\lim_{n \rightarrow \infty} P\left[\left| F_i^{(n)} - p_i \right| < \varepsilon \right] = 1 \quad (i = 1, 2, \dots, k). \quad (8.6)$$

Dim. Dalla diseguaglianza di Chebyshev per ogni $\varepsilon > 0$ si ha:

$$P\left\{ \left| F_i^{(n)} - E\left[F_i^{(n)}\right] \right| < \varepsilon \right\} \geq 1 - \frac{D^2\left[F_i^{(n)}\right]}{\varepsilon^2}.$$

In virtù delle (8.4) e (8.5) si ottiene quindi:

$$P\left[\left| F_i^{(n)} - p_i \right| < \varepsilon \right] \geq 1 - \frac{p_i(1-p_i)}{n\varepsilon^2}. \quad (8.7)$$

Passando al limite per $n \rightarrow \infty$, dalla (8.7) segue la (8.6). ■

Conseguenza immediata delle (8.4) e (8.6) è che la frequenza relativa $F_i^{(n)}$ risulta essere uno stimatore corretto e consistente della probabilità p_i . Pertanto ogni volta che si osserva la realizzazione di un campione casuale di taglia elevata tratto da una popolazione discreta, le frequenze relative osservate $f_i^{(n)}$ costituiscono una buona approssimazione delle probabilità

Tabella 8.1: Frequenze assolute empiriche e teoriche di cui all'Esempio 7.1.1.

i	n_i	$n\hat{\pi}_i$	i	n_i	$n\hat{\pi}_i$
0	57	54.40	8	45	67.88
1	203	210.52	9	27	29.19
2	383	407.36	10	10	11.30
3	525	525.50	11	4	3.97
4	532	508.42	12	0	1.28
5	408	394.52	13	1	0.38
6	273	253.82	14	1	0.11
7	139	140.32	≥ 15	0	0.03

p_i , e quindi il diagramma delle frequenze fornisce una buona approssimazione del grafico della distribuzione di probabilità della variabile genitrice X .

Nel seguito è mostrato un esempio di costruzione di diagramma delle frequenze con riferimento ad un particolare insieme di dati ottenuti in un esperimento realmente effettuato.³

Esempio 8.1.1 Nell'osservazione di un fenomeno casuale consistente nel decadimento di atomi radioattivi di una certa sostanza, nell'arco di un intervallo di tempo pari a 326 minuti si sono verificati in totale 10097 decadimenti. Dividiamo questo intervallo di tempo in $n = 2608$ subintervalli ciascuno di durata 7.5 secondi e indichiamo con n_i il numero di subintervalli in cui è riscontrato un numero di decadimenti pari ad i ($i = 0, 1, \dots$). I dati sperimentalmente ottenuti sono riportati nella Tabella 8.1 nella quale la seconda e la quinta colonna indicano le frequenze assolute n_i osservate.

Supponiamo che il fenomeno del decadimento di atomi radioattivi sia descritto da una variabile casuale genitrice X che in base a considerazioni fisiche è ragionevole assumere abbia distribuzione di Poisson di media $\mu > 0$ incognita. La probabilità che in ciascun subintervallo considerato si verifichi un numero di decadimenti pari ad i è allora esprimibile al seguente modo:

$$p_i = P(X = i) = \frac{\mu^i}{i!} e^{-\mu} \quad (i = 0, 1, \dots).$$

Il parametro incognito μ va stimato usando i dati disponibili. Ricordando che lo stimatore di massima verosimiglianza della media di una popolazione di Poisson è la media campionaria (cfr. Esempio 3.7.6), si può usare come stima di μ il rapporto tra il numero totale di decadimenti osservati ed il numero n di subintervalli:

$$\hat{\mu} = \frac{10097}{2608} = \frac{1}{2608} \sum_{i=0}^{14} i n_i = 3.87.$$

Ciò comporta la seguente stima $\hat{\pi}_i$ della probabilità p_i :

$$\hat{\pi}_i = \frac{\hat{\mu}^i}{i!} e^{-\hat{\mu}} = \frac{(3.87)^i}{i!} e^{-3.87} \quad (i = 0, 1, \dots). \quad (8.8)$$

³Dati e figure del presente capitolo sono rielaborazioni tratte da classici esempi.

8.1. DIAGRAMMA DELLE FREQUENZE

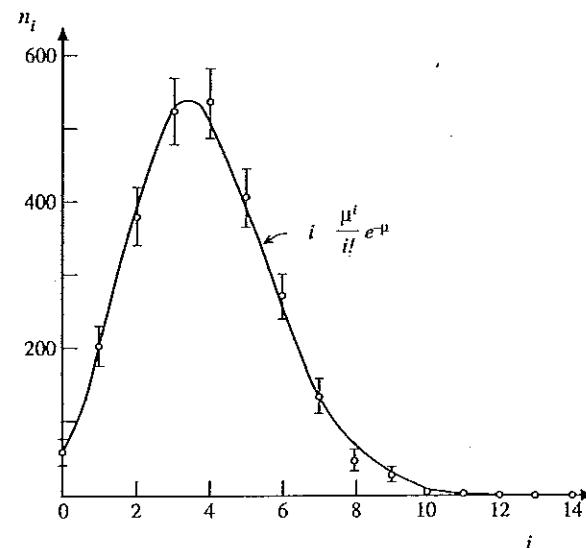


Figura 8.2: Diagramma delle frequenze per i dati di cui alla Tabella 8.1.

Nell'esempio considerato l'insieme \mathcal{D} dei dati osservabili è costituito in teoria dall'insieme degli interi non negativi, ma in realtà è inimmaginabile che in un intervallo di tempo di ampiezza finita (pari a 7.5 secondi) possa verificarsi un numero arbitrariamente elevato di decadimenti. La legge di probabilità di Poisson ipotizzata non appare tuttavia in contrasto con tale osservazione quando si riflette che le probabilità p_i possono considerarsi praticamente nulle per i molto grande. Nella Tabella 8.1 accanto ai valori n_i delle frequenze misurate sono riportati i valori

$$n\hat{\pi}_i = 2608 \frac{(3.87)^i}{i!} e^{-3.87} \quad (i = 0, 1, \dots) \quad (8.9)$$

delle corrispondenti frequenze teoriche. I dati presenti nell'ultima riga della tabella stanno ad indicare che in nessuno dei subintervalli si è registrato un numero di decadimenti maggiore o uguale a 15, e che la frequenza teorica corrispondente a tale evento è data da

$$n \sum_{i=15}^{\infty} \hat{\pi}_i = 2608 e^{-3.87} \sum_{i=15}^{\infty} \frac{(3.87)^i}{i!} = 0.03.$$

Riportiamo ora in un diagramma le frequenze assolute n_i ottenute sperimentalmente e le stime (8.9) delle quantità teoriche corrispondenti. Come indica la Figura 8.2, i discostamenti non sono elevati, così che l'ipotesi che il numero di decadimenti di atomi negli intervalli temporali esaminati sia regolato da una legge di probabilità di Poisson di media $\mu = 3.87$ sembra accettabile, quanto meno ad una ispezione visiva.

Si noti che in Figura 8.2 la curva continua interpolante le frequenze teoriche (8.9) è stata tracciata per mera comodità al fine di evidenziare meglio l'accordo dei dati previsti con quelli registrati.

Onde ottenere una conferma quantitativa dell'accettabilità della suddetta conclusione è conveniente ricorrere al test chi-quadrato per verificare l'ipotesi nulla $H_0: \{p_i\} = \{\hat{p}_i\}$ contro l'ipotesi alternativa $H_1: \{p_i\} \neq \{\hat{p}_i\}$, con le \hat{p}_i date dalle (8.8). È evidente che conviene scegliere come classi del test chi-quadrato proprio i valori d_i osservabili; ricordando inoltre la regola empirica secondo cui per ogni classe deve avversi $n\hat{p}_i \geq 5$, conviene considerare 12 classi e porre:

$$C_i = \{i-1\} \quad (i=1, 2, \dots, 11), \quad C_{12} = \{11, 12, \dots\}.$$

In accordo con quanto visto nel § 5.4, l'ipotesi nulla va rifiutata se la realizzazione osservata appartiene alla regione critica C , che sceglieremo in accordo con la (5.59). Se si pone $\alpha = 0.05$, essendo $k = 12$ e $r = 1$ dalla Tabella 3 dell'Appendice B si trae $\chi^2_{\alpha; k-r-1} = \chi^2_{0.05; 10} = 18.307$; la regione critica di ampiezza approssimativamente 0.05 è dunque

$$C = \left\{ (x_1, x_2, \dots, x_n) : \sum_{i=1}^k \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} \geq 18.307 \right\}.$$

Poiché in base ai dati risulta

$$\sum_{i=1}^k \frac{(n_i - n\hat{p}_i)^2}{n\hat{p}_i} = 12.973,$$

si conclude che la realizzazione osservata non appartiene alla regione critica, cosicché l'ipotesi nulla $H_0: \{p_i\} = \{\hat{p}_i\}$ non va rifiutata.

Concludendo, le differenze tra frequenze assolute osservate e quantità teoriche corrispondenti non sono tanto significative da indurci a ritenere che il modello poissoniano scelto debba considerarsi inaccettabile.

Va infine fatto rilevare che nella Figura 8.2 in corrispondenza di ciascuna delle frequenze assolute n_i registrate è riportato l'intervallo, o *errore*, $n_i \pm 2\sqrt{n_i}$ che troverà una giustificazione nelle considerazioni che svolgeremo a conclusione del § 8.3. ◆

8.2 Istogramma

Nel presente paragrafo si affronta il problema della rappresentazione di dati generati da una variabile casuale continua.

Esaminiamo il caso in cui si dispone di una n -upla (x_1, x_2, \dots, x_n) di dati appartenenti ad un insieme continuo che per semplicità supponiamo essere l'intervallo (a, b) . Più precisamente, ipotizziamo che i dati osservati siano la realizzazione di un campione casuale (X_1, X_2, \dots, X_n) la cui variabile genitrice X è continua e possiede densità di probabilità $f(x)$ continua nel suo supporto⁴ (a, b) . Allo scopo di categorizzare i dati ottenuti, è opportuno suddividere l'intervallo (a, b) in k subintervalli, o *classi*, C_1, C_2, \dots, C_k rispettivamente di ampiezze $\Delta t_1, \Delta t_2, \dots, \Delta t_k$. Si fissano, dunque, $k+1$ reali t_0, t_1, \dots, t_k con

$$a \equiv t_0 < t_1 < t_2 < \dots < t_{k-1} < t_k \equiv b$$

⁴Ciò significa che risulta $f(x) > 0$ per ogni $x \in (a, b)$ e $f(x) = 0$ altrimenti.

8.2. ISTOGRAMMA

e si costruiscono poi le classi $C_i = (t_{i-1}, t_i)$ ($i = 1, 2, \dots, k$), di ampiezze $\Delta t_i = t_i - t_{i-1}$. Avendo così suddiviso l'intervallo (a, b) , è evidente che ciascuno dei dati osservati (x_1, x_2, \dots, x_n) cade certamente all'interno di una delle classi C_i . In analogia con la (8.1) indichiamo con N_1, N_2, \dots, N_k le variabili casuali rappresentanti rispettivamente le frequenze assolute dei valori appartenenti alle classi C_1, C_2, \dots, C_k . In perfetta analogia con il caso discreto, ciascuna variabile casuale N_i risulta avere distribuzione binomiale; infatti ciascuna delle variabili del campione (X_1, X_2, \dots, X_n) assume valori nella classe C_i con probabilità costante

$$p_i \stackrel{\text{def}}{=} P(t_{i-1} < X \leq t_i) = \int_{t_{i-1}}^{t_i} f(x) dx \quad (i = 1, 2, \dots, k) \quad (8.10)$$

dove, come si è detto, $f(x)$ è la densità di probabilità della variabile casuale genitrice X che si è supposto generi i dati. L'osservazione di una n -upla di variabili (X_1, X_2, \dots, X_n) , indipendenti e identicamente distribuite, può dunque riguardarsi come il risultato di un esperimento di n prove ripetute e indipendenti in ciascuna delle quali l'estrazione di un valore che appartenga alla classe $C_i = (t_{i-1}, t_i)$ si verifica con probabilità costante p_i . La frequenza N_i , che esprime il numero di volte in cui un elemento del campione causale assume valore in C_i , è quindi una variabile casuale binomiale N_i di parametri n e p_i , dotata pertanto di valore medio e varianza (8.2) dove ora le p_i sono date dalle (8.10).

Denotiamo con $H_n(t)$ la variabile casuale (dipendente dal parametro t) definita nel seguente modo:

$$H_n(t) = \begin{cases} \frac{1}{\Delta t_i} \frac{N_i}{n} & \text{per } t \in C_i \quad (i = 1, 2, \dots, k), \\ 0 & \text{altrimenti.} \end{cases} \quad (8.11)$$

Ad essa attribuiamo la denominazione di *variabile casuale istogramma di classi* C_1, C_2, \dots, C_k . Se indichiamo con n_i il numero di elementi della realizzazione osservata (x_1, x_2, \dots, x_n) che appartengono alla classe C_i ($i = 1, 2, \dots, k$), ossia la frequenza assoluta dei valori della realizzazione del campione che cadono in C_i , la funzione costante a tratti

$$h_n(t) = \begin{cases} \frac{1}{\Delta t_i} \frac{n_i}{n} & \text{per } t \in C_i \quad (i = 1, 2, \dots, k), \\ 0 & \text{altrimenti,} \end{cases} \quad (8.12)$$

detta *istogramma di classi* C_1, C_2, \dots, C_k , denota i valori assunti dalla variabile casuale (8.11) al variare di t . In maniera analoga al diagramma delle frequenze del caso discreto, $h_n(t)$ fornisce una descrizione grafica della realizzazione (x_1, x_2, \dots, x_n) , come è ad esempio mostrato nella Figura 8.3.

Notiamo esplicitamente che l'area di ciascun rettangolo di base Δt_i e altezza $h_n(t)$ egualia la frequenza relativa n_i/n ; la somma delle aree dei rettangoli che compongono l'istogramma (8.12) è dunque

$$\sum_{i=1}^k \frac{n_i}{n} = 1.$$

Le proprietà

$$h_n(t) \geq 0, \quad \int_a^b h_n(t) dt = \sum_{i=1}^k \frac{n_i}{n} = 1.$$

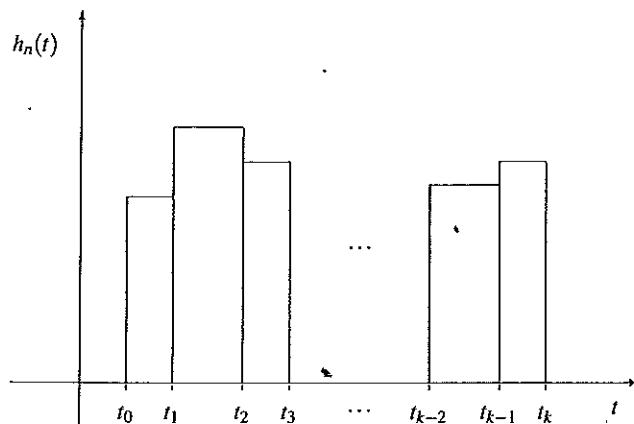


Figura 8.3: Istogramma.

dell'istogramma sono reminiscenti delle analoghe proprietà di cui godono le densità di probabilità. Tale analogia è in realtà la manifestazione di uno stretto legame esistente tra queste due funzioni; invero, al crescere del numero n di dati osservati il grafico dell'istogramma si avvicina al grafico della densità di probabilità della variabile casuale che si è supposto generi i dati, come specificato dalla seguente proposizione:

Proposizione 8.2.1 *Sia $H_n(t)$ la variabile casuale istogramma di classi C_1, C_2, \dots, C_k per il campione casuale (X_1, X_2, \dots, X_n) di variabile casuale genitrice X avente densità di probabilità $f(x)$ continua nel suo supporto (a, b) . Per ogni $t \in C_i$ esiste un reale $\xi_i \in (t_{i-1}, t_i)$ dipendente da C_i ma non da t tale da aversi*

$$\lim_{n \rightarrow \infty} P[|H_n(t) - f(\xi_i)| < \varepsilon] = 1 \quad (i = 1, 2, \dots, k) \quad (8.13)$$

per ogni $\varepsilon > 0$.

Dim. Dalla (8.11), in virtù della prima delle (8.2), si ha:

$$E[H_n(t)] = \begin{cases} \frac{p_i}{\Delta t_i} & \text{per } t \in C_i, \quad (i = 1, 2, \dots, k) \\ 0 & \text{altrimenti,} \end{cases} \quad (8.14)$$

mentre per la seconda delle (8.2) si ottiene:

$$D^2[H_n(t)] = \begin{cases} \frac{p_i(1-p_i)}{n(\Delta t_i)^2} & \text{per } t \in C_i, \quad (i = 1, 2, \dots, k) \\ 0 & \text{altrimenti,} \end{cases} \quad (8.15)$$

8.2. ISTOGRAMMA

con le p_i definite dalle (8.10). Per la supposta continuità della $f(x)$ in (a, b) , il teorema della media assicura per ogni C_i l'esistenza di un punto ξ_i interno a C_i tale da aversi

$$p_i = f(\xi_i)(t_i - t_{i-1}) = f(\xi_i)\Delta t_i \quad (i = 1, 2, \dots, k), \quad (8.16)$$

così che le (8.14) e (8.15) diventano

$$E[H_n(t)] = \begin{cases} f(\xi_i) & \text{per } t \in C_i \quad (i = 1, 2, \dots, k), \\ 0 & \text{altrimenti,} \end{cases} \quad (8.17)$$

$$D^2[H_n(t)] = \begin{cases} \frac{f(\xi_i)[1-f(\xi_i)\Delta t_i]}{n\Delta t_i} & \text{per } t \in C_i \quad (i = 1, 2, \dots, k), \\ 0 & \text{altrimenti.} \end{cases} \quad (8.18)$$

Utilizzando la diseguaglianza di Chebyshev per $t \in C_i$ ($i = 1, 2, \dots, k$), in cui è $D^2[H_n(t)] > 0$, per le (8.17) e (8.18) per ogni $\varepsilon > 0$ si ottiene:

$$\begin{aligned} P\{|H_n(t) - E[H_n(t)]| < \varepsilon\} &\equiv P[|H_n(t) - f(\xi_i)| < \varepsilon] \\ &\geq 1 - \frac{f(\xi_i)[1-f(\xi_i)\Delta t_i]}{n\Delta t_i\varepsilon^2}. \end{aligned} \quad (8.19)$$

Dalla (8.19), al limite per $n \rightarrow \infty$ si perviene infine alla tesi (8.13). ■

La Proposizione 8.2.1 mostra che le probabilità di piccoli discostamenti tra la variabile casuale-istogramma $H_n(t)$ e i valori $f(\xi_i)$ della densità della variabile genitrice X diventano sempre più prossime all'unità al crescere della taglia n del campione casuale. Pertanto quanto maggiore è il numero di dati osservati tanto maggiore è la bontà dell'approssimazione della densità di probabilità $f(\xi_i)$ da parte dell'istogramma $h_n(t)$.

Un ruolo cruciale è svolto dalle ampiezze Δt_i delle classi C_i . Per effettuare un confronto dettagliato tra modello teorico e dati sperimentali le ampiezze Δt_i devono, infatti, essere piccole; d'altra parte, come mostra la (8.18), se Δt_i tende a zero la varianza di $H_n(t)$ diverge, ed in tal caso la media $E[H_n(t)] = f(\xi_i)$ non è significativa. Inoltre, se le ampiezze delle classi C_i sono molto piccole si ha $f(\xi_i) \approx f(t_{i-1}) \approx f(t_i) \approx f(t)$, così che l'istogramma $h_n(t)$ fornisce proprio un'approssimazione di $f(t)$; viceversa, se le ampiezze Δt_i sono grandi può accadere che la suddetta approssimazione non sia buona, e che quindi non sia adeguata l'approssimazione di $f(t)$ con l'istogramma $h_n(t)$. Occorre quindi individuare dei criteri opportuni per la scelta delle ampiezze delle classi.

È importante notare che anche se il campione casuale è molto complesso non è solitamente necessario studiare in dettaglio con la stessa accuratezza tutte le zone dell'istogramma. Può pertanto essere utile, nella rappresentazione dei dati, considerare una suddivisione più fine per la zona di interesse maggiore ed, invece, una suddivisione meno fine-laddove non ha molto interesse scendere in grandi dettagli. Ad esempio è possibile porre

$$t_i = \xi_{i/k} \quad (i = 0, 1, \dots, k), \quad (8.20)$$

dove ξ_p denota il quantile p -esimo (cfr. Definizione 4.9.1) della variabile genitrice, con $\xi_0 \equiv a$ e $\xi_1 \equiv b$. In tal modo le classi C_i risultano essere equiprobabili. Invero, per la (8.20) dalla

(8.10) si ottiene:

$$p_i = \int_{\xi_{(i-1)/k}}^{\xi_{i/k}} f(x) dx = \frac{i}{k} - \frac{i-1}{k} = \frac{1}{k} \quad (i = 1, 2, \dots, k).$$

Se i quantili non sono noti se ne può utilizzare una stima in accordo con il criterio di cui al § 4.9. Infatti, se il numero k delle classi è un divisore di $n+1$, ossia se risulta $k = (n+1)/r$ con r intero positivo, il quantile $\xi_{i/k}$ può essere stimato col valore assunto dalla statistica d'ordine $X_{(ri)}$, per $i = 1, 2, \dots, k-1$. Così facendo, la (8.20) va sostituita con

$$t_0 = a, \quad t_i = x_{(ri)} \quad (i = 1, 2, \dots, k-1), \quad t_k = b.$$

Se poi non vi sono motivi preferenziali, può essere conveniente scegliere classi C_i di uguali ampiezze $\Delta t_i = \Delta t$ ($i = 1, 2, \dots, k$); ciò equivale allo scegliere come estremi delle classi C_i i reali

$$t_i = a + \frac{i}{k} (b-a) \quad (i = 0, 1, \dots, k),$$

così che risulti $\Delta t_i = (b-a)/k$ per ogni i .

Talvolta nella scelta delle ampiezze delle classi C_i si segue una regola empirica consistente nel richiedere, quando possibile, che in ogni classe siano contenute almeno cinque osservazioni.

In taluni casi, allorché ad esempio i valori dell'istogramma $h_n(t)$ sono molto piccoli, può essere preferibile riportare nel grafico i corrispondenti valori "assoluti" $n h_n(t)$. In questo caso, ovviamente, il confronto va effettuato con i valori teorici corrispondenti $n f(\xi_i)$.

Esempio 8.2.1 Ad illustrazione delle considerazioni svolte consideriamo un esempio di costruzione di istogrammi con riferimento ad un esperimento consistito nello scoccare $n = 96$ frecce contro un bersaglio posto a 30 m di distanza e nel misurare le deviazioni dei dardi in orizzontale (azimut) e in verticale (altezza) rispetto al centro del bersaglio. I dati ottenuti sono riportati nelle Tabelle 8.2 e 8.3. Costruiamo ora i due istogrammi usando classi della stessa ampiezza $\Delta t = 10$ cm sia per la deviazione in azimut (Figura 8.4) che per quella in altezza (Figura 8.5).⁵ Poiché il numero di dati non è molto elevato, tracciamo questi grafici in termini di frequenze assolute anziché relative. Per il loro aspetto, si è tentati di interpolare gli istogrammi costruiti con una curva simmetrica, nella fattispecie con una densità normale avente come valore medio μ e come varianza σ^2 le stime suggerite dal metodo della massima verosimiglianza, ossia (cfr. Esempio 3.7.7) la media campionaria e lo scarto quadratico medio dei dati del campione:

$$\mu = \frac{1}{n} \sum_{j=1}^n x_j, \quad \sigma^2 = \frac{1}{n} \sum_{j=1}^n (x_j - \mu)^2. \quad (8.21)$$

Denotando con $\psi(t)$ la densità di probabilità di una variabile casuale normale standard, onde effettuare il confronto riportiamo dunque sui grafici delle Figure 8.4 e 8.5 i valori della densità

$$\frac{1}{\sigma} \psi\left(\frac{t-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right]$$

⁵Poiché l'esempio si riferisce a particolari campioni di taglia n fissata, la dipendenza degli istogrammi dall'indice n è stata omessa nelle figure.

Tabella 8.2: Deviazioni dei dardi in azimut.

i	C_i	n_i	$n p_i$
1	(-35, -25]	0	0.17
2	(-25, -15]	2	1.80
3	(-15, -5]	9	9.73
4	(-5, 5]	28	25.43
5	(5, 15]	30	32.22
6	(15, 25]	21	19.83
7	(25, 35]	5	5.91
8	(35, 45]	1	0.85
9	(45, 55]	0	0.06

Tabella 8.3: Deviazioni dei dardi in altezza.

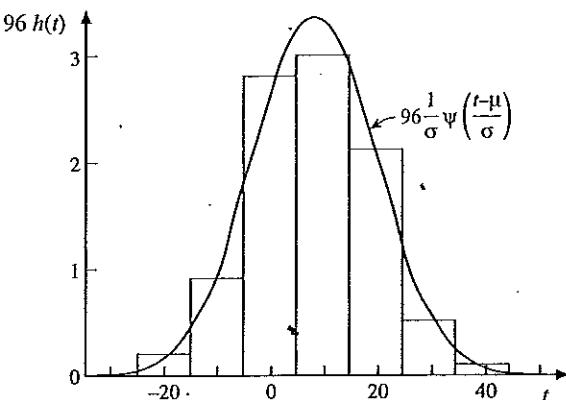
i	C_i	n_i	$n p_i$
1	(-65, -55]	0	0.54
2	(-55, -45]	3	1.95
3	(-45, -35]	5	5.92
4	(-35, -25]	13	12.86
5	(-25, -15]	18	19.94
6	(-15, -5]	21	22.07
7	(-5, 5]	21	17.46
8	(5, 15]	10	9.86
9	(15, 25]	5	3.98
10	(25, 35]	0	1.42

della variabile casuale normale di media μ e varianza σ^2 moltiplicati per la taglia del campione ($n = 96$), tenendo conto che le stime (8.21) forniscono $\mu = 8.3$ e $\sigma^2 = 129.96$ per le deviazioni in azimut, e $\mu = -12.0$ e $\sigma^2 = 289.0$ per le deviazioni in altezza. Da un esame visivo dei grafici l'ipotesi normale appare ragionevole. Esaminiamo inoltre le frequenze assolute n_i e le quantità teoriche $n p_i$ corrispondenti; queste, ricordando la (8.10), sono valutate mediante l'espressione

$$n p_i = n \int_{t_{i-1}}^{t_i} f(x) dx = n \int_{t_{i-1}}^{t_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx$$

facendo ricorso alla Tabella 1 dell'Appendice B relativa alla distribuzione normale. Le frequenze n_i e le corrispondenti quantità teoriche $n p_i$, elencate rispettivamente nella terza e nella quarta colonna delle Tabelle 8.2 e 8.3, sono molto prossime, il che indica ulteriormente la bontà dell'approssimazione normale.

Riduciamo ora le ampiezze delle classi di un fattore 5 in modo da aversi $\Delta t = 2$ cm. Come c'è da attendersi, compaiono fluttuazioni molto maggiori rispetto alla densità normale

Figura 8.4: Deviazione in azimut. $\Delta t = 10$ cm.

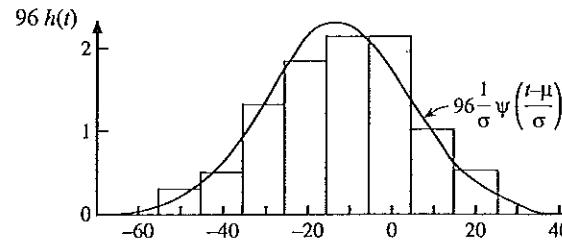
costruita teoricamente usando gli stessi parametri. Ciò appare evidente dagli istogrammi così ottenuti per le deviazioni in azimut (Figura 8.6) e in altezza (Figura 8.7).

In questo caso, a prima vista sembrerebbe che l'approssimazione normale non sia adatta a rappresentare gli istogrammi; è però possibile notare che l'area in cui sussiste un buon accordo è superiore all'80% dell'area totale, così che anche con questa suddivisione più fitta l'approssimazione può ritenersi accettabile. Va peraltro rilevato che la nuova ampiezza scelta per le classi è inadeguata a soddisfare la regola empirica prima ricordata sul numero di dati che dovrebbe cadere in ciascuna classe. Un esame della densità normale nei due grafici mostra anche che il valore medio è situato nel punto di ascissa 8.3 cm per la deviazione in azimut e di ascissa -12.0 cm per la deviazione in altezza; ciò fa pensare che il campione di dati qui considerato sia indicativo di una deviazione sistematica del puntamento dell'arciere sia verso destra che verso il basso. ♦

8.3 Iстограмма cumulativo

Un altro metodo largamente utilizzato per rappresentare dati sperimentali consiste nel costruire una funzione empirica che sia l'analogo di una funzione di distribuzione. Dato un campione casuale (X_1, X_2, \dots, X_n) , per ogni $t \in \mathbb{R}$ denotiamo con $G_n(t)$ la variabile casuale così definita:

$$G_n(t) = \frac{1}{n} \sum_{j=1}^n I_{(-\infty, t]}(X_j) \quad (t \in \mathbb{R}), \quad (8.22)$$

Figura 8.5: Deviazione in altezza. $\Delta t = 10$ cm.

dove $I_A(z)$ denota la funzione indicatrice definita nella (3.75). Alla (8.22) diamo il nome di *variabile casuale istogramma cumulativo*. Il valore

$$g_n(t) = \frac{1}{n} \sum_{j=1}^n I_{(-\infty, t]}(x_j) = \frac{1}{n} \left| \{x_1, x_2, \dots, x_n; x_j \leq t\} \right| \quad (t \in \mathbb{R}) \quad (8.23)$$

che essa assume in corrispondenza di una realizzazione (x_1, x_2, \dots, x_n) del campione casuale si dice *istogramma cumulativo* o *funzione di distribuzione empirica* del campione casuale. La funzione $g_n(t)$, che è non decrescente e costante a tratti, può essere anche rappresentata in termini dei valori $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ delle statistiche d'ordine del campione. Infatti, se in aggiunta si pone $x_{(0)} = -\infty$ e $x_{(n+1)} = \infty$, è facile rendersi conto che risulta

$$g_n(t) = \frac{r}{n} \quad \text{per } x_{(r)} \leq t < x_{(r+1)} \quad (8.24)$$

in quanto la condizione $x_{(r)} \leq t < x_{(r+1)}$ equivale ad affermare che esattamente r valori della realizzazione (x_1, x_2, \dots, x_n) sono non maggiori di t , ossia che risulta

$$\sum_{j=1}^n I_{(-\infty, t]}(x_j) = r.$$

Nella Figura 8.8 è mostrato il grafico di un istogramma cumulativo $g_n(t)$, in cui appare evidente la proprietà (8.24).

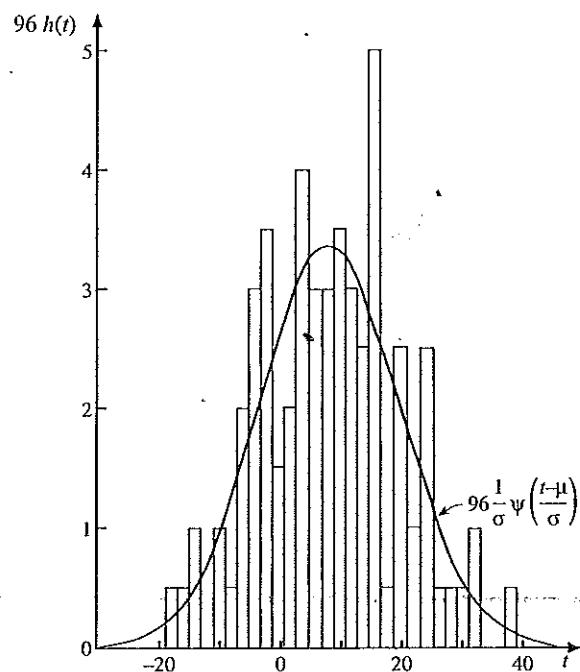
È interessante osservare che sussiste una relazione che lega l'istogramma $h_n(t)$ all'istogramma cumulativo $g_n(t)$.

Osservazione 8.3.1 Denotati con t_0, t_1, \dots, t_k gli estremi delle classi C_i di un istogramma $h_n(t)$, si ha:

$$g_n(t_i) = \int_{-\infty}^{t_i} h_n(t) dt \quad (i = 1, 2, \dots, k). \quad (8.25)$$

Dim. In virtù della (8.23), indicata con n_r la frequenza assoluta della generica classe C_r segue:

$$g_n(t_i) = \frac{1}{n} \left| \{x_1, x_2, \dots, x_n; x_j \leq t_i\} \right|$$

Figura 8.6: Deviazione in azimut. $\Delta t = 2$ cm.

$$\begin{aligned}
 &= \frac{1}{n} \left| \{x_1, x_2, \dots, x_n; x_j \in C_1 \cup C_2 \cup \dots \cup C_i\} \right| \\
 &= \frac{1}{n} (n_1 + n_2 + \dots + n_i).
 \end{aligned} \tag{8.26}$$

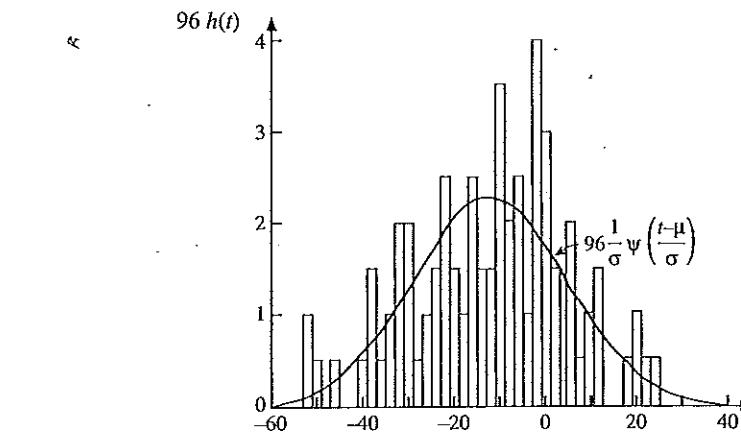
Dalla definizione (8.12) dell'istogramma $h_n(t)$ si ha poi:

$$\int_{-\infty}^{t_i} h_n(t) dt = \sum_{r=1}^i \int_{t_{r-1}}^{t_r} \frac{1}{\Delta t_r} \frac{n_r}{n} dt = \sum_{r=1}^i \frac{n_r}{n}. \tag{8.27}$$

Confrontando le (8.26) e (8.27) segue la (8.25). ■

Si noti che è possibile ricavare un'espressione analoga alla (8.25) che lega l'istogramma cumulativo $g_n(t)$ al diagramma delle frequenze introdotto nel § 8.1. Nel caso di una realizzazione (x_1, x_2, \dots, x_n) estratta da una popolazione discreta che assume valori nell'insieme

8.3. ISTOGRAMMA CUMULATIVO

Figura 8.7: Deviazione in altezza. $\Delta t = 2$ cm.

ordinato $\mathcal{D} = \{d_1, d_2, \dots, d_k\}$ è facile rendersi conto che risulta

$$g_n(d_i) = \sum_{r=1}^i f_r^{(n)} \quad (i = 1, 2, \dots, k),$$

dove $f_r^{(n)}$ è la frequenza relativa del dato d_r .

Proposizione 8.3.1 Sia $G_n(t)$ la variabile casuale istogramma cumulativo relativa ad un campione casuale (X_1, X_2, \dots, X_n) di variabile casuale genitrice X avente funzione di distribuzione $F(t)$. Per ogni $t \in \mathbb{R}$ e per ogni $\varepsilon > 0$ si ha

$$\lim_{n \rightarrow \infty} P[|G_n(t) - F(t)| < \varepsilon] = 1. \tag{8.28}$$

Dim. Osserviamo che per ogni $t \in \mathbb{R}$ la variabile casuale

$$N(t) \stackrel{\text{def}}{=} \sum_{j=1}^n I_{(-\infty, t]}(X_j) \tag{8.29}$$

ha distribuzione binomiale di parametri n e $F(t)$ in quanto somma di n variabili casuali di Bernoulli indipendenti ciascuna delle quali assume valore 1 con probabilità $P(X_j \leq t) \equiv F(t)$. Il valore medio e la varianza di $N(t)$ sono

$$E[N(t)] = nF(t), \quad D^2[N(t)] = nF(t)[1 - F(t)]. \tag{8.30}$$

Poiché dalle (8.22) e (8.29) segue $G_n(t) = N(t)/n$, facendo uso delle (8.30) possiamo calcolare valore medio e varianza della variabile casuale istogramma cumulativo:

$$E[G_n(t)] = \frac{E[N(t)]}{n} = F(t),$$

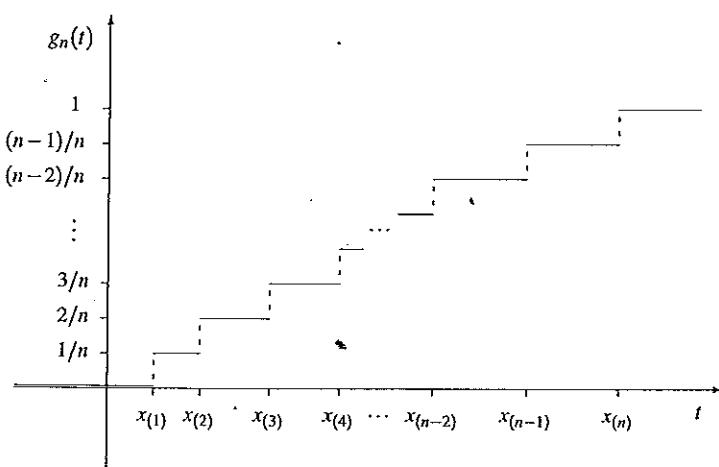


Figura 8.8: Istogramma cumulativo.

$$D^2[G_n(t)] = \frac{D^2[N(t)]}{n^2} = \frac{F(t)[1-F(t)]}{n} \quad (8.31)$$

Facendo uso della diseguaglianza di Chebyshev, per ogni $\epsilon > 0$ ricaviamo poi:

$$\begin{aligned} P\{|G_n(t) - E[G_n(t)]| < \epsilon\} &\equiv P\{|G_n(t) - F(t)| < \epsilon\} \\ &\geq 1 - \frac{F(t)[1-F(t)]}{n\epsilon^2}. \end{aligned} \quad (8.32)$$

Procedendo al limite per $n \rightarrow \infty$, dalla (8.32) si giunge infine alla tesi (8.28). ■

La Proposizione 8.3.1 implica che per ogni $t \in \mathbb{R}$ la variabile casuale istogramma cumulativo $G_n(t)$ è uno stimatore corretto e consistente della funzione di distribuzione $F(t)$ della variabile casuale genitrice. Ciò comporta che una funzione di distribuzione $F(t)$ incognita può essere stimata con l'istogramma cumulativo $g_n(t)$ qualora la taglia n del campione casuale sia molto elevata.

Si noti che talora in luogo di $g_n(t)$ si preferisce considerare la funzione $ng_n(t)$ che rappresenta l'istogramma cumulativo "assoluto", e che quindi va confrontata con il prodotto $nF(t)$.

Gli istogrammi cumulativi relativi all'Esempio 8.2.1 sono riportati nelle Figure 8.9 e 8.10 nelle quali l'istogramma cumulativo è indicato più semplicemente con $g(t)$. In esse sono graficate le funzioni $ng(t)$ e il prodotto della taglia del campione, pari a 96, per la funzione

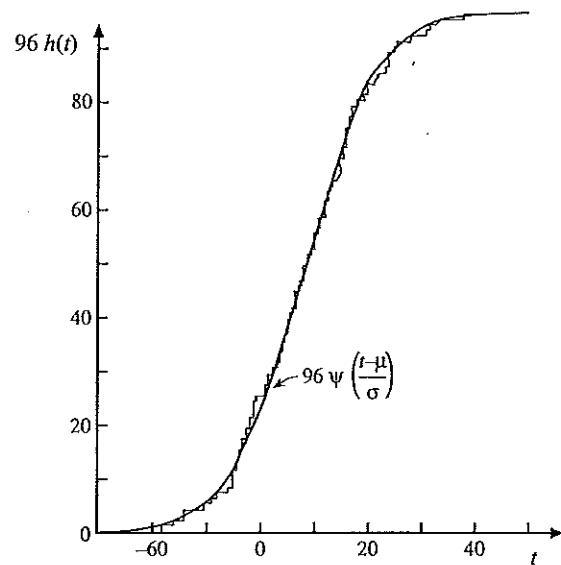


Figura 8.9: Istogramma cumulativo (deviazioni in azimut).

di distribuzione

$$\Psi\left(\frac{t-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^t \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx$$

di una variabile casuale normale di media μ e varianza σ^2 , con $\mu = 8.3$ e $\sigma^2 = 129.96$ per le deviazioni in azimut, e $\mu = -12.0$ e $\sigma^2 = 289.0$ per le deviazioni in altezza. Si può notare che l'approssimazione della funzione teorica $96\Psi[(t-\mu)/\sigma]$ mediante la funzione $96g(t)$ appare migliore di quella che si è ottenuta approssimando la densità di probabilità mediante l'istogramma (cfr. Figure 8.4 e 8.5). Nell'istogramma cumulativo risultano infatti in generale ridotte le fluttuazioni in quanto discostamenti di segno opposto dalla curva teorica vengono sommati, e quindi tendono a compensarsi. Va poi sottolineato che la presenza di una suddivisione in classi fa sì che l'istogramma fornisca informazioni che dipendono in maniera significativa dalla scelta delle classi, mentre per l'istogramma cumulativo un tale inconveniente non sussiste. Così, mentre l'istogramma $h_n(t)$ fornisce un'approssimazione per la densità $f(t)$ che dipende dalla classe C_i contenente il valore t , l'istogramma cumulativo $g_n(t)$ costituisce una stima di $F(t)$ che non dipende da quantità estranee ai dati costituenti la realizzazione.

Qualunque sia il tipo di rappresentazione dei dati (diagramma delle frequenze, istogramma o istogramma cumulativo) è importante determinare l'ampiezza dell'intervallo di fluttuazione delle grandezze in esame. Esaminiamo le espressioni delle varianze ottenute rispettivamente nei tre casi considerati di rappresentazione dei dati. Cominciamo con l'osser-

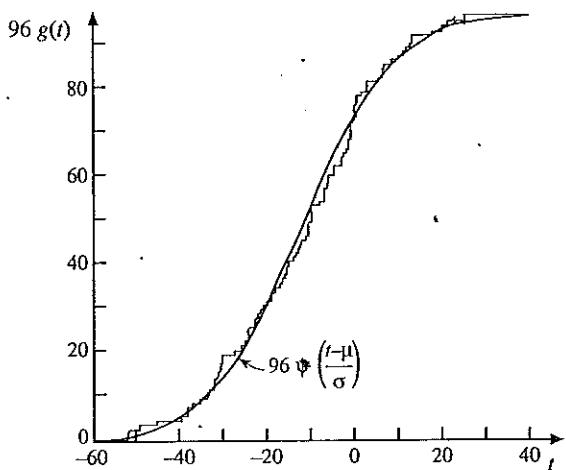


Figura 8.10: Istogramma cumulativo (deviazione in altezza).

vare che nel caso del diagramma delle frequenze dalla seconda delle (8.2) segue che la deviazione standard $D(N_i)$ della frequenza assoluta N_i è approssimabile al seguente modo:

$$D(N_i) \equiv \sqrt{n p_i (1 - p_i)} \approx \sqrt{n p_i} = \sqrt{E(N_i)} \approx \sqrt{n_i} \quad (8.33)$$

nell'ipotesi in cui p_i è molto minore dell'unità (così che $\sqrt{1 - p_i}$ è molto prossimo ad 1) e se si assume che il valore osservato n_i differisce solo di poco dal valore medio $E(N_i)$. Riferiamoci ora all'istogramma per il quale, facendo uso della seconda delle (8.2) con p_i data dalla (8.16), la deviazione standard della frequenza assoluta del numero di elementi del campione appartenenti alla classe C_i è così approssimabile:

$$\begin{aligned} D(N_i) &\equiv \sqrt{n p_i (1 - p_i)} = \sqrt{n f(\xi_i) \Delta_i [1 - f(\xi_i) \Delta_i]} \\ &\approx \sqrt{n f(\xi_i) \Delta_i} = \sqrt{n p_i} = \sqrt{E(N_i)} \approx \sqrt{n_i}, \end{aligned} \quad (8.34)$$

dove n_i denota il numero di elementi della realizzazione che cadono nella classe C_i , avendo supposto che $f(\xi_i) \Delta_i$ sia trascurabile rispetto all'unità e che il valore osservato n_i differisca di poco dal valore medio $E(N_i)$. Infine, nel caso dell'istogramma cumulativo considerazioni analoghe conducono all'approssimazione

$$D[N(t)] \equiv \sqrt{n F(t) [1 - F(t)]} \approx \sqrt{n F(t)} \approx \sqrt{n(t)} \quad (8.35)$$

dove $n(t)$ denota il numero di valori nella realizzazione che sono minori o uguali a t .

Le espressioni (8.33), (8.34) e (8.35) delle deviazioni standard spiegano perché i risultati di misure o osservazioni sperimentali siano spesso rappresentati nelle forme

$$n_i \pm \ell \sqrt{n_i}, \quad n(t) \pm \ell \sqrt{n(t)}, \quad (8.36)$$

8.4. TEOREMA DI GLIVENKO-CANTELLI

dove ℓ è un intero che solitamente assume il valore 1 o, come nel caso dei dati riportati in Figura 8.2, il valore 2. Invero, le (8.36) individuano degli intervalli all'interno dei quali con alta probabilità assumono valori le variabili casuali N_i e $N(t)$.

8.4 Teorema di Glivenko-Cantelli

Nel paragrafo precedente abbiamo visto che per rappresentare dei dati sperimentali sovente si fa uso dell'istogramma cumulativo il cui grafico è l'analogo di quello di una funzione di distribuzione. In questo paragrafo verrà dimostrato che l'istogramma cumulativo, ossia la funzione di distribuzione empirica di un campione casuale, converge alla funzione di distribuzione delle variabili del campione secondo il criterio di convergenza quasi certa. Questo risultato fornisce una giustificazione quantitativa della comune affermazione qualitativa e imprecisa che "un campione casuale di grande taglia descrive una popolazione".

Consideriamo un campione casuale (X_1, X_2, \dots, X_n) estratto da una popolazione avente variabile casuale genitrice X . Un problema di fondamentale interesse in statistica consiste nel ricostruire la funzione $F(x)$ partendo dalle variabili osservabili costituenti il campione considerato. Il teorema di Glivenko-Cantelli, che andremo subito ad introdurre, afferma che ciò è possibile, ed in modo uniforme.

Cominciamo col ricordare la definizione (8.22) della variabile casuale funzione di distribuzione empirica n -esima $G_n(x)$ della funzione di distribuzione $F(x)$:

$$G_n(x) = \frac{1}{n} \sum_{k=1}^n I_{(-\infty, x]}(X_k).$$

Lemma 8.4.1 La successione $\{G_n(x)\}$ converge quasi certamente⁶ alla funzione di distribuzione $F(x)$, ossia per ogni $x \in \mathbb{R}$ risulta

$$G_n(x) \xrightarrow{q.c.} F(x) \quad (8.37)$$

$$G_n(x^-) \xrightarrow{q.c.} F(x^-), \quad (8.38)$$

dove $F(x^-)$ denota $\lim_{y \rightarrow x^-} F(y)$.

Dim. Per dimostrare la (8.37) consideriamo la successione di variabili casuali indipendenti e identicamente distribuite $\{I_{(-\infty, x]}(X_n)\}_n$. Chiaramente risulta $E[I_{(-\infty, x]}(X_n)] = P(X_n \leq x) = F(x)$. Dalla legge forte dei grandi numeri⁷ si ha dunque:

$$G_n(x) = \frac{1}{n} \sum_{k=1}^n I_{(-\infty, x]}(X_k) \xrightarrow{q.c.} F(x),$$

ossia la (8.37). Per dimostrare la (8.38) si applica un procedimento analogo, con la sola differenza che in questo caso occorre considerare la successione di variabili $\{I_{(-\infty, x]}(X_n)\}_n$.

⁶Una successione $\{X_n\}$ di variabili casuali si dice convergere quasi certamente ad una variabile casuale X , e lo si denota con $X_n \xrightarrow{q.c.} X$, se $\{\omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\}$ è un evento quasi certo, ossia se risulta $P(\lim_{n \rightarrow \infty} X_n = X) = 1$.

⁷La legge forte dei grandi numeri afferma che se X_1, X_2, \dots è una successione di variabili casuali indipendenti, identicamente distribuite e dotate di media finita μ , allora $\frac{X_1 + X_2 + \dots + X_n}{n} \xrightarrow{q.c.} \mu$.

Va sottolineato che le affermazioni (8.37) e (8.38) sussistono per ogni fissato x reale. A questo punto nasce il seguente quesito:

(i) è vero che

$$P[G_n(x) \rightarrow F(x) \text{ per ogni } x \in \mathbb{R}] = 1? \quad (8.39)$$

Si noti che in parentesi quadre compare una intersezione *non numerabile* di eventi, che quindi non è detto sia essa stessa un evento. Comunque, supponendo che tale intersezione sia un evento, e supponendo che la (8.39) sussista, è possibile considerare l'ulteriore quesito:

(ii) è vero che

$$P[G_n(x) \rightarrow F(x) \text{ uniformemente in } x] = 1?$$

Il seguente teorema, noto anche quale Teorema di Glivenko-Cantelli, stabilisce che la risposta al quesito (ii) è affermativa.

Teorema 8.4.1 (Glivenko-Cantelli) Se $\{X_n\}$ è una successione di variabili casuali indipendenti e identicamente distribuite, con funzione di distribuzione $F(x)$, si ha:

$$P\left[\sup_{x \in \mathbb{R}} |G_n(x) - F(x)| \rightarrow 0\right] = 1.$$

Dim. Per ogni intero $r \geq 2$ e per $k = 1, 2, \dots, r-1$ poniamo

$$x_{r,k} = \min\left\{x : \frac{k}{r} \leq F(x)\right\}. \quad (8.40)$$

Sia inoltre $x_{r,0} = -\infty$ e $x_{r,r} = \infty$. Consideriamo gli intervalli non vuoti del tipo $[x_{r,k}, x_{r,k+1})$. Se

$$x \in [x_{r,k}, x_{r,k+1}),$$

risulta:

$$G_n(x_{r,k}) \leq G_n(x) \leq G_n(x_{r,k+1}), \quad F(x_{r,k}) \leq F(x) \leq F(x_{r,k+1}).$$

Di qui segue:

$$\begin{aligned} G_n(x) - F(x) &\leq G_n(x_{r,k+1}) - F(x_{r,k}) \\ &= G_n(x_{r,k+1}) - F(x_{r,k+1}) + F(x_{r,k+1}) - F(x_{r,k}) \\ &\leq G_n(x_{r,k+1}) - F(x_{r,k+1}) + \frac{1}{r}, \end{aligned} \quad (8.41)$$

$$\begin{aligned} G_n(x) - F(x) &\geq G_n(x_{r,k}) - F(x_{r,k+1}) \\ &= G_n(x_{r,k}) - F(x_{r,k}) - [F(x_{r,k+1}) - F(x_{r,k})] \\ &\geq G_n(x_{r,k}) - F(x_{r,k}) - \frac{1}{r}. \end{aligned} \quad (8.42)$$

Le ultime diseguaglianze nelle (8.41) e (8.42) scaturiscono dalla circostanza che dalla (8.40) si trae:

$$F(x_{r,k+1}) - F(x_{r,k}) \leq \frac{1}{r}.$$

Si noti poi che le (8.41) e (8.42) sussistono per ogni $\omega \in \Omega$. Pertanto, per ogni x reale si ha:

$$\begin{aligned} &|G_n(x) - F(x)| \\ &\leq \max_{1 \leq k, j \leq r-1} \left\{ |G_n(x_{r,k}) - F(x_{r,k})|, |G_n(x_{r,j}) - F(x_{r,j})| \right\} + \frac{1}{r}. \end{aligned} \quad (8.43)$$

La diseguagliaza (8.43) vale anche per gli estremi superiori di ambo i membri così che, facendo uso del Lemma 8.4.1, si ottiene:

$$\limsup_{x \in \mathbb{R}} |G_n(x) - F(x)| \leq \frac{1}{r}, \quad \text{q.c.}$$

Dall'arbitrarietà di r segue infine la tesi. \square

Si noti che nella dimostrazione effettuata si richiede che

$$\sup_{x \in \mathbb{R}} |G_n(x) - F(x)|$$

sia una variabile casuale. Ciò, peraltro, può facilmente dimostrarsi.

Appendice A

Principali variabili casuali

A.1 Variabili casuali discrete

Vengono qui richiamate le definizioni e gli elementi descrittivi essenziali delle principali variabili casuali discrete. Per semplicità di notazione la distribuzione di probabilità di una variabile casuale discreta X sarà denotata con

$$P_X(k) = P(X = k) \quad (k \in \mathcal{S}),$$

dove \mathcal{S} indica l'insieme, finito o numerabile, dei valori assunti da X .

A.1.1 Variabile uniforme

La variabile casuale X uniforme descrive l'esito di un esperimento casuale caratterizzato da n possibili esiti equiprobabili. La sua distribuzione di probabilità, detta *distribuzione uniforme discreta*, è la seguente:

$$P_X(k) = \frac{1}{n} \quad (k = 1, 2, \dots, n),$$

dove n è un intero positivo.

(i) *Funzione generatrice dei momenti*

$$M_X(t) \stackrel{\text{def}}{=} E(e^{tX}) = \sum_{k=1}^n \frac{1}{n} e^{kt} = \frac{e^t}{n} \frac{1 - e^{nt}}{1 - e^t} \quad (t \in \mathbb{R}).$$

(ii) *Valore medio*

$$E(X) = \frac{n+1}{2}.$$

(iii) *Varianza*

$$D^2(X) = \frac{n^2 - 1}{12}.$$

(iv) Coefficiente di variazione

$$\text{CV}(X) \stackrel{\text{def}}{=} \frac{D(X)}{E(X)} = \sqrt{\frac{n-1}{3(n+1)}}.$$

(v) Momento del terzo ordine

$$E(X^3) = \frac{n(n+1)^2}{4}.$$

(vi) Momento del quarto ordine

$$E(X^4) = \frac{(n+1)(2n+1)(3n^2+3n-1)}{30}.$$

A.1.2 Variabile di Bernoulli

In un esperimento casuale sia θ ($0 \leq \theta \leq 1$) la probabilità di occorrenza di un evento E , e quindi $1-\theta$ la probabilità dell'evento complementare \bar{E} . Il risultato dell'esperimento è descritto mediante una variabile casuale X che assume valore 1 se si verifica E e valore 0 se si verifica \bar{E} , ossia se E non si verifica. Ad X si dà il nome di variabile casuale di Bernoulli. La sua distribuzione di probabilità, detta *distribuzione di Bernoulli*, è la seguente:

$$P_X(k) = \theta^k (1-\theta)^{1-k} \quad (k = 0, 1).$$

(i) Funzione generatrice dei momenti

$$M_X(t) \stackrel{\text{def}}{=} E(e^{tX}) = 1 - \theta + \theta e^t \quad (t \in \mathbb{R}).$$

(ii) Valore medio

$$E(X) = \theta.$$

(iii) Varianza

$$D^2(X) = \theta(1-\theta).$$

(iv) Coefficiente di variazione

$$\text{CV}(X) \stackrel{\text{def}}{=} \frac{D(X)}{E(X)} = \sqrt{\frac{1-\theta}{\theta}}.$$

(v) Momento centrale del terzo ordine

$$E\{[X - E(X)]^3\} = \theta(1-\theta)(1-2\theta).$$

(vi) Momento centrale del quarto ordine

$$E\{[X - E(X)]^4\} = 3\theta^2(1-\theta)^2 + \theta(1-\theta)[1-6\theta(1-\theta)].$$

A.1. VARIABILI CASUALI DISCRETE

(vii) Coefficiente di asimmetria

$$\frac{E\{[X - E(X)]^3\}}{D^3(X)} = \frac{1-2\theta}{\sqrt{\theta(1-\theta)}}.$$

(viii) Coefficiente di piccatezza

$$\frac{E\{[X - E(X)]^4\}}{D^4(X)} = \frac{1}{\theta(1-\theta)} - 3.$$

A.1.3 Variabile binomiale

In un esperimento casuale consistente nella ripetizione di n prove identiche indipendenti sia θ la probabilità di occorrenza di un dato evento E in ogni prova. Si consideri la variabile casuale Y_n che indica il numero di prove dell'esperimento in cui E si verifica. Se si denota con X_i la variabile casuale che descrive l'esito della prova i -esima ($i = 1, 2, \dots, n$), si ha $Y_n = X_1 + X_2 + \dots + X_n$, dove X_1, X_2, \dots, X_n sono variabili di Bernoulli indipendenti e identicamente distribuite. La variabile casuale Y_n è detta variabile binomiale. La sua distribuzione di probabilità, detta *distribuzione binomiale*, ha la seguente espressione:

$$P_{Y_n}(k) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \quad (k = 0, 1, \dots, n)$$

ed è quindi caratterizzata dai due parametri θ ed n , con $0 \leq \theta \leq 1$ ed n intero positivo.

(i) Funzione generatrice dei momenti

$$M_{Y_n}(t) \stackrel{\text{def}}{=} E(e^{tY_n}) = (1 - \theta + \theta e^t)^n \quad (t \in \mathbb{R}).$$

(ii) Valore medio

$$E(Y_n) = n\theta.$$

(iii) Varianza

$$D^2(Y_n) = n\theta(1-\theta).$$

(iv) Coefficiente di variazione

$$\text{CV}(Y_n) \stackrel{\text{def}}{=} \frac{D(Y_n)}{E(Y_n)} = \sqrt{\frac{1-\theta}{n\theta}}.$$

(v) Momento centrale del terzo ordine

$$E\{[Y_n - E(Y_n)]^3\} = n\theta(1-\theta)(1-2\theta).$$

(vi) Momento centrale del quarto ordine

$$E\{[Y_n - E(Y_n)]^4\} = 3n^2\theta^2(1-\theta)^2 + n\theta(1-\theta)[1-6\theta(1-\theta)].$$

(vii) Coefficiente di asimmetria

$$\frac{E\{[Y_n - E(Y_n)]^3\}}{D^3(Y_n)} = \frac{1 - 2\theta}{\sqrt{n\theta(1-\theta)}}.$$

(viii) Coefficiente di piccatezza

$$\frac{E\{[Y_n - E(Y_n)]^4\}}{D^4(Y_n)} = \frac{1}{n\theta(1-\theta)} + 3 - \frac{6}{n}.$$

Si noti che il coefficiente di asimmetria è nullo per $\theta = 1/2$ e che il coefficiente di piccatazza tende a 3 per $n \rightarrow \infty$.

A.1.4 Variabile di Poisson

Si assuma che la probabilità θ di occorrenza nella singola prova di un evento E in un esperimento di n prove ripetute dipenda inversamente dal numero di prove:

$$\theta = \frac{\lambda}{n} \quad (\lambda \text{ costante positiva fissata}).$$

Detta Y_n la variabile casuale binomiale corrispondente, al limite per $n \rightarrow \infty$ risulta:

$$\begin{aligned} \lim_{n \rightarrow \infty} P_{Y_n}(k) &= \lim_{n \rightarrow \infty} \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \lim_{n \rightarrow \infty} \frac{n}{k!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \\ &= \frac{\lambda^k}{k!} e^{-\lambda}. \end{aligned}$$

La distribuzione di probabilità limite

$$P_X(k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad (k = 0, 1, \dots),$$

con $\lambda > 0$, viene detta *distribuzione di Poisson*; la variabile casuale X corrispondente è denominata variabile casuale di Poisson.

(i) Funzione generatrice dei momenti

$$M_X(t) \stackrel{\text{def}}{=} E(e^{tX}) = \exp[\lambda(e^t - 1)] \quad (t \in \mathbb{R}).$$

(ii) Valore medio

$$E(X) = \lambda.$$

A.1. VARIABILI CASUALI DISCRETE(iii) Varianza¹

$$D^2(X) = \lambda.$$

(iv) Coefficiente di variazione

$$CV(X) \stackrel{\text{def}}{=} \frac{D(X)}{E(X)} = \frac{1}{\sqrt{\lambda}}.$$

(v) Momento centrale del terzo ordine

$$E\{[X - E(X)]^3\} = \lambda.$$

(vi) Momento centrale del quarto ordine

$$E\{[X - E(X)]^4\} = \lambda + 3\lambda^2.$$

(vii) Coefficiente di asimmetria

$$\frac{E\{[X - E(X)]^3\}}{D^3(X)} = \frac{1}{\sqrt{\lambda}}.$$

(viii) Coefficiente di piccatazza

$$\frac{E\{[X - E(X)]^4\}}{D^4(X)} = \frac{1}{\lambda} + 3.$$

A.1.5 Variabile geometrica

Si consideri un esperimento consistente nella ripetizione indefinita di prove indipendenti in ciascuna delle quali un evento E si verifica con probabilità θ . Sia X la variabile casuale che denota il numero di prove dopo il quale E si verifica per la prima volta. Questa descrive dunque il "tempo di attesa" della prima occorrenza di E , ed è denominata variabile casuale geometrica. La *distribuzione geometrica*, definita come la distribuzione di probabilità di X , è data da

$$P_X(k) = \theta(1 - \theta)^{k-1} \quad (k = 1, 2, \dots),$$

con $0 \leq \theta \leq 1$.

(i) Funzione generatrice dei momenti

$$M_X(t) \stackrel{\text{def}}{=} E(e^{tX}) = \frac{\theta e^t}{1 - (1 - \theta)e^t} \quad \left(t < \ln \frac{1}{1 - \theta}\right).$$

(ii) Valore medio

$$E(X) = \frac{1}{\theta}.$$

(iii) Varianza

$$D^2(X) = \frac{1-\theta}{\theta^2}.$$

(iv) Coefficiente di variazione

$$\text{CV}(X) \stackrel{\text{def}}{=} \frac{D(X)}{E(X)} = \sqrt{1-\theta}.$$

(v) Momento centrale del terzo ordine

$$E\{[X-E(X)]^3\} = \frac{1-\theta}{\theta^2} \left(1+2\frac{1-\theta}{\theta}\right).$$

(vi) Momento centrale del quarto ordine

$$E\{[X-E(X)]^4\} = \frac{1-\theta}{\theta^2} \left(1+9\frac{1-\theta}{\theta^2}\right).$$

(vii) Coefficiente di asimmetria

$$\frac{E\{[X-E(X)]^3\}}{D^3(X)} = \frac{\theta}{\sqrt{1-\theta}} \left(1+2\frac{1-\theta}{\theta}\right).$$

(viii) Coefficiente di piccatesza

$$\frac{E\{[X-E(X)]^4\}}{D^4(X)} = \frac{\theta^2}{1-\theta} \left(1+9\frac{1-\theta}{\theta^2}\right).$$

A.1.6 Variabile di Pascal

Con riferimento ad un esperimento consistente nella ripetizione indefinita di prove indipendenti in ciascuna delle quali un evento E si verifica con probabilità θ , sia Y_n la variabile casuale che denota il numero di prove dopo il quale E si verifica per la n -esima volta. Essa rappresenta dunque il "tempo di attesa" della n -esima occorrenza di E , ed è denominata variabile casuale di Pascal. Se si denota con X_i la variabile casuale che descrive il tempo che intercorre tra la $(i-1)$ -esima e la i -esima occorrenza dell'evento E , si ha $Y_n = X_1 + X_2 + \dots + X_n$, dove X_1, X_2, \dots, X_n sono variabili casuali geometriche indipendenti ed identicamente distribuite. La *distribuzione di Pascal*, definita come la distribuzione di probabilità di Y_n , è la seguente:

$$P_{Y_n}(k) = \binom{k-1}{n-1} \theta^n (1-\theta)^{k-n} \quad (k=n, n+1, \dots),$$

con $0 \leq \theta \leq 1$ ed n intero positivo.

(i) Funzione generatrice dei momenti

$$M_{Y_n}(t) \stackrel{\text{def}}{=} E(e^{tY_n}) = \left[\frac{\theta e^t}{1-(1-\theta)e^t} \right]^n \quad \left(t < \ln \frac{1}{1-\theta} \right).$$

(ii) Valore medio

$$E(Y_n) = \frac{n}{\theta}.$$

(iii) Varianza

$$D^2(Y_n) = n \frac{1-\theta}{\theta^2}.$$

(iv) Coefficiente di variazione

$$\text{CV}(Y_n) \stackrel{\text{def}}{=} \frac{D(Y_n)}{E(Y_n)} = \sqrt{\frac{1-\theta}{n}}.$$

(v) Momento centrale del terzo ordine

$$E\{[Y_n-E(Y_n)]^3\} = n \frac{1-\theta}{\theta^2} \left(1+2\frac{1-\theta}{\theta}\right).$$

(vi) Momento centrale del quarto ordine

$$E\{[Y_n-E(Y_n)]^4\} = n \frac{1-\theta}{\theta^2} \left[1+(6+3n)\frac{1-\theta}{\theta^2}\right].$$

(vii) Coefficiente di asimmetria

$$\frac{E\{[Y_n-E(Y_n)]^3\}}{D^3(Y_n)} = \frac{2-\theta}{\sqrt{n(1-\theta)}}.$$

(viii) Coefficiente di piccatesza

$$\frac{E\{[Y_n-E(Y_n)]^4\}}{D^4(Y_n)} = \frac{\theta^2 + (6+3n)(1-\theta)}{n(1-\theta)}.$$

A.1.7 Variabile binomiale negativa

La *distribuzione binomiale negativa* è la distribuzione di probabilità di una variabile casuale X tale che

$$P_X(k) = \binom{r+k-1}{k} \theta^r (1-\theta)^k \quad (k=0,1,\dots),$$

con $0 \leq \theta \leq 1$, $r > 0$ e $\binom{\alpha}{k}$, con α reale, così definito:

$$\binom{\alpha}{k} = \frac{\alpha(\alpha-1)\cdots(\alpha-k+1)}{k!}.$$

(i) Funzione generatrice dei momenti

$$M_X(t) \stackrel{\text{def}}{=} E(e^{tX}) = \left[\frac{\theta}{1-(1-\theta)e^t} \right]^r \quad (t < \ln \frac{1}{1-\theta}).$$

(ii) Valore medio

$$E(X) = r \frac{1-\theta}{\theta}.$$

(iii) Varianza

$$D^2(X) = r \frac{1-\theta}{\theta^2}.$$

(iv) Coefficiente di variazione

$$\text{CV}(X) \stackrel{\text{def}}{=} \frac{D(X)}{E(X)} = \frac{1}{\sqrt{r(1-\theta)}}.$$

(v) Momento centrale del terzo ordine

$$E\{[X - E(X)]^3\} = r \frac{1-\theta}{\theta^2} \left(1 + 2 \frac{1-\theta}{\theta} \right).$$

(vi) Momento centrale del quarto ordine

$$E\{[X - E(X)]^4\} = r \frac{1-\theta}{\theta^2} \left[1 + (6+3r) \frac{1-\theta}{\theta^2} \right].$$

(vii) Coefficiente di asimmetria

$$\frac{E\{[X - E(X)]^3\}}{D^3(X)} = \frac{2-\theta}{\sqrt{r(1-\theta)}}.$$

(viii) Coefficiente di piccatesza

$$\frac{E\{[X - E(X)]^4\}}{D^4(X)} = \frac{\theta^2 + (6+3r)(1-\theta)}{r(1-\theta)}.$$

Non è difficile mostrare che se Y_n è una variabile casuale di Pascal di parametri θ ed n , la variabile $X = Y_n - n$ ha distribuzione binomiale negativa di parametri θ ed n .

A.1.8 Variabile ipergeometrica

Questa variabile casuale nasce in problemi di estrazione senza rimpiazzamento. Un modello esemplificativo fa riferimento ad un'urna contenente n biglie, m delle quali sono bianche. Si estraiano a caso r biglie dall'urna ($0 \leq r \leq n$), l'una dopo l'altra senza rimpiazzamento. La variabile casuale X che rappresenta il numero di biglie bianche presenti tra le r biglie estratte

viene detta variabile casuale ipergeometrica. La sua distribuzione, che prende il nome di *distribuzione ipergeometrica*, è così definita:

$$P_X(k) = \frac{\binom{m}{k} \binom{n-m}{r-k}}{\binom{n}{r}} \quad (k = 0, 1, \dots, r; r-n+m \leq k \leq m),$$

con n intero positivo ed $m, r = 0, 1, \dots, n$.

(i) Valore medio

$$E(X) = r \frac{m}{n}.$$

(ii) Varianza

$$D^2(X) = \frac{rm(n-m)(n-r)}{n^2(n-1)}.$$

(iii) Coefficiente di variazione

$$\text{CV}(X) \stackrel{\text{def}}{=} \frac{D(X)}{E(X)} = \sqrt{\frac{(n-m)(n-r)}{rm(n-1)}}.$$

(iv) Momenti fattoriali discendenti

$$E[X(X-1)\cdots(X-s+1)] = s! \frac{\binom{m}{s} \binom{r}{s}}{\binom{n}{s}}.$$

A.1.9 Variabile multinomiale

Si consideri un esperimento casuale consistente nella ripetizione di n prove identiche indipendenti. Siano p_1, p_2, \dots, p_k le probabilità di occorrenza in ogni prova di k eventi E_1, E_2, \dots, E_k incompatibili ed esaustivi. Si denoti con X_i la variabile casuale rappresentante il numero di prove dell'esperimento in cui si verifica l'evento E_i . La variabile casuale k -dimensionale (X_1, X_2, \dots, X_k) è detta variabile casuale multinomiale. La sua distribuzione di probabilità, denominata *distribuzione multinomiale*, è la seguente:

$$P_{X_1, X_2, \dots, X_k}(n_1, n_2, \dots, n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k},$$

$$(n_1, n_2, \dots, n_k \geq 0, n_1 + n_2 + \cdots + n_k = n),$$

con $p_1, p_2, \dots, p_k \geq 0$ e $p_1 + p_2 + \cdots + p_k = 1$.

(i) Funzione generatrice dei momenti

$$\begin{aligned} M_{X_1, X_2, \dots, X_k}(t_1, t_2, \dots, t_k) &\stackrel{\text{def}}{=} E(e^{t_1 X_1 + t_2 X_2 + \dots + t_k X_k}) \\ &= (p_1 e^{t_1} + p_2 e^{t_2} + \dots + p_k e^{t_k})^n \\ &\quad (t_1, t_2, \dots, t_k \in \mathbb{R}). \end{aligned}$$

(ii) Momento prodotto di ordine k

$$E(X_1 X_2 \cdots X_k) = n(n-1) \cdots (n-k+1) p_1 p_2 \cdots p_k.$$

(iii) Momento prodotto di ordine 2

$$E(X_i X_j) = n(n-1) p_i p_j.$$

(iv) Covarianza

$$\text{cov}(X_i, X_j) = -n p_i p_j.$$

Si noti che ciascuna delle variabili casuali X_i ha distribuzione binomiale di parametri p_i ed n .

A.2 Variabili casuali continue

Viene qui fornita una rassegna delle principali variabili casuali continue richiamandone sinteticamente le definizioni e gli elementi descrittivi di maggior interesse.

A.2.1 Variabile uniforme

Una variabile casuale X si dice uniforme nell'intervallo (a, b) se ha densità di probabilità

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{per } a < x < b, \\ 0 & \text{altrimenti.} \end{cases}$$

(i) Funzione generatrice dei momenti

$$M_X(t) \stackrel{\text{def}}{=} E(e^{tX}) = \frac{e^{bt} - e^{at}}{(b-a)t} \quad (t \in \mathbb{R}).$$

(ii) Valore medio

$$E(X) = \frac{a+b}{2}.$$

(iii) Varianza

$$D^2(X) = \frac{(b-a)^2}{12}.$$

A.2.2 Variabili casuali continue

(iv) Coefficiente di variazione

$$\text{CV}(X) \stackrel{\text{def}}{=} \frac{D(X)}{E(X)} = \sqrt{\frac{b-a}{3(a+b)}}.$$

(v) Momenti centrali di ordine dispari

$$E\{[X - E(X)]^{2n+1}\} = 0 \quad (n = 0, 1, \dots).$$

(vi) Momento centrale del quarto ordine

$$E\{[X - E(X)]^4\} = \frac{(b-a)^4}{80}.$$

(vii) Coefficiente di asimmetria

$$\frac{E\{[X - E(X)]^3\}}{D^3(X)} = 0.$$

(viii) Coefficiente di piccatesza

$$\frac{E\{[X - E(X)]^4\}}{D^4(X)} = \frac{9}{5}.$$

A.2.2 Variabile normale

Una variabile casuale X si dice normale di parametri μ e σ^2 , con $\mu \in \mathbb{R}$ e $\sigma \in \mathbb{R}^+$, se ha densità di probabilità

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad (x \in \mathbb{R}).$$

Questa è nota anche come densità di Gauss, o gaussiana.

(i) Funzione generatrice dei momenti

$$M_X(t) \stackrel{\text{def}}{=} E(e^{tX}) = \exp\left(\mu t + \frac{\sigma^2}{2} t^2\right) \quad (t \in \mathbb{R}).$$

(ii) Valore medio

$$E(X) = \mu.$$

(iii) Varianza

$$D^2(X) = \sigma^2.$$

(iv) Coefficiente di variazione

$$\text{CV}(X) \stackrel{\text{def}}{=} \frac{D(X)}{E(X)} = \frac{\sigma}{\mu}.$$

(v) Momenti centrali di ordine dispari

$$E\{[X - E(X)]^{2n+1}\} = 0 \quad (n = 0, 1, \dots).$$

(vi) Momento centrale del quarto ordine

$$E\{[X - E(X)]^4\} = 3\sigma^4.$$

(vii) Coefficiente di asimmetria

$$\frac{E\{[X - E(X)]^3\}}{D^3(X)} = 0.$$

(viii) Coefficiente di piccatezza

$$\frac{E\{[X - E(X)]^4\}}{D^4(X)} = 3.$$

La densità di Gauss è un caso particolare della cosiddetta densità gaussiana generalizzata di ordine α data da

$$f(x) = \frac{\alpha^{1-\frac{1}{\alpha}}}{2\sigma\Gamma(\frac{1}{\alpha})} \exp\left(-\frac{|x-\mu|^\alpha}{\alpha\sigma^\alpha}\right) \quad (x \in \mathbb{R}),$$

con $\mu \in \mathbb{R}$ e $\sigma, \alpha \in \mathbb{R}^+$.

A.2.3 Variabile esponenziale

Si dice che X è una variabile casuale esponenziale se essa ha densità di probabilità

$$f_X(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & \text{per } x > 0, \\ 0 & \text{altrimenti,} \end{cases}$$

con $\theta > 0$.

(i) Funzione generatrice dei momenti

$$M_X(t) \stackrel{\text{def}}{=} E(e^{tX}) = \frac{1}{1-\theta t} \quad \left(t < \frac{1}{\theta}\right).$$

(ii) Valore medio

$$E(X) = \theta.$$

(iii) Varianza

$$D^2(X) = \theta^2.$$

(iv) Coefficiente di variazione

$$CV(X) \stackrel{\text{def}}{=} \frac{D(X)}{E(X)} = 1.$$

(v) Momento centrale del terzo ordine

$$E\{[X - E(X)]^3\} = 2\theta^3.$$

(vi) Momento centrale del quarto ordine

$$E\{[X - E(X)]^4\} = 9\theta^4.$$

(vii) Coefficiente di asimmetria

$$\frac{E\{[X - E(X)]^3\}}{D^3(X)} = 2.$$

(viii) Coefficiente di piccatezza

$$\frac{E\{[X - E(X)]^4\}}{D^4(X)} = 9.$$

(ix) Momenti intorno all'origine

$$E(X^k) = k! \theta^k \quad (k = 1, 2, \dots).$$

A.2.4 Variabile gamma

Una variabile casuale X si dice essere di tipo gamma di parametri $\alpha > 0$ e $\beta > 0$ se ha densità di probabilità

$$f_X(x) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}} & \text{per } x > 0, \\ 0 & \text{altrimenti,} \end{cases} \quad (\text{A.1})$$

dove $\Gamma(\alpha)$ denota la funzione gamma di Eulero:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \quad (\alpha > 0). \quad (\text{A.2})$$

(i) Funzione generatrice dei momenti

$$M_X(t) \stackrel{\text{def}}{=} E(e^{tX}) = \left(\frac{1}{1-\beta t}\right)^\alpha \quad \left(t < \frac{1}{\beta}\right).$$

(ii) Valore medio

$$E(X) = \alpha\beta.$$

(iii) Varianza

$$D^2(X) = \alpha\beta^2.$$

(iv) Coefficiente di variazione

$$CV(X) \stackrel{\text{def}}{=} \frac{D(X)}{E(X)} = \frac{1}{\sqrt{\alpha}}.$$

(v) *Momento centrale del terzo ordine*

$$E\{[X - E(X)]^3\} = 2\alpha\beta^3.$$

(vi) *Momento centrale del quarto ordine*

$$E\{[X - E(X)]^4\} = (6\alpha + 3\alpha^2)\beta^4.$$

(vii) *Coefficiente di asimmetria*

$$\frac{E\{[X - E(X)]^3\}}{D^3(X)} = \frac{2}{\sqrt{\alpha}}.$$

(viii) *Coefficiente di piccatesza*

$$\frac{E\{[X - E(X)]^4\}}{D^4(X)} = 3 + \frac{6}{\alpha}.$$

Si noti che per $\alpha = 1$ la densità gamma si riduce ad una densità esponenziale di parametro β ; se, inoltre, X_i ($i = 1, 2, \dots, n$) sono variabili casuali esponenziali indipendenti di valore medio β , la variabile casuale $X_1 + X_2 + \dots + X_n$ è di tipo gamma con parametri $\alpha \equiv n$ e β .

La (A.1) può riguardarsi come caso particolare della cosiddetta densità gamma generalizzata

$$g(x; \alpha, \beta, \gamma) = \begin{cases} \frac{\gamma}{\beta^{\alpha\gamma}\Gamma(\alpha)} x^{\alpha\gamma-1} e^{-\left(\frac{x}{\beta}\right)^{\gamma}} & \text{per } x > 0, \\ 0 & \text{altrimenti,} \end{cases}$$

con α, β e γ parametri positivi.

A.2.5 Variabile beta

Una variabile casuale X è di tipo beta di parametri a e b ($a, b \in \mathbb{R}^+$) se ha densità di probabilità

$$f_X(x) = \begin{cases} \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1} & \text{per } 0 < x < 1, \\ 0 & \text{altrimenti,} \end{cases}$$

dove $B(a, b)$ denota la funzione beta:

$$B(a, b) = \int_0^1 x^{a-1} (1-x)^{b-1} dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

(i) *Funzione generatrice dei momenti*

$$M_X(t) \stackrel{\text{def}}{=} E(e^{tX}) = \sum_{k=0}^{\infty} \frac{B(a+k, b)}{B(a, b)} \frac{t^k}{k!} \quad (t \in \mathbb{R}).$$

A.2. VARIABILI CASUALI CONTINUE

(ii) *Valore medio*

$$E(X) = \frac{a}{a+b}.$$

(iii) *Varianza*

$$D^2(X) = \frac{ab}{(a+b+1)(a+b)^2}.$$

(iv) *Coefficiente di variazione*

$$CV(X) \stackrel{\text{def}}{=} \frac{D(X)}{E(X)} = \sqrt{\frac{b}{a(a+b+1)}}.$$

(v) *Momenti intorno all'origine*

$$E(X') = \frac{B(a+r, b)}{B(a, b)}.$$

A.2.6 Variabile chi-quadrato

Una variabile casuale X si dice essere chi-quadrato con v gradi di libertà se ha densità di probabilità

$$f_X(x) = \begin{cases} \frac{x^{\frac{v}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{v}{2}} \Gamma\left(\frac{v}{2}\right)} & \text{per } x > 0, \\ 0 & \text{altrimenti,} \end{cases}$$

dove v è un intero positivo e $\Gamma(\cdot)$ è la funzione gamma di Eulero definita nella (A.2).(i) *Funzione generatrice dei momenti*

$$M_X(t) \stackrel{\text{def}}{=} E(e^{tX}) = \left(\frac{1}{1-2t}\right)^{\frac{v}{2}} \quad \left(t < \frac{1}{2}\right).$$

(ii) *Valore medio*

$$E(X) = v.$$

(iii) *Varianza*

$$D^2(X) = 2v.$$

(iv) *Coefficiente di variazione*

$$CV(X) \stackrel{\text{def}}{=} \frac{D(X)}{E(X)} = \sqrt{\frac{2}{v}}.$$

(v) *Momento centrale del terzo ordine*

$$E\{[X - E(X)]^3\} = 8v.$$

(vi) *Momento centrale del quarto ordine*

$$E\{[X - E(X)]^4\} = 12v(4+v).$$

(vii) *Coefficiente di asimmetria*

$$\frac{E\{[X - E(X)]^3\}}{D^3(X)} = 2\sqrt{\frac{2}{v}}.$$

(viii) *Coefficiente di piccatezza*

$$\frac{E\{[X - E(X)]^4\}}{D^4(X)} = 3 + \frac{12}{v}.$$

Si noti che la densità chi-quadrato con v gradi di libertà si identifica con la densità gamma di parametri $\beta = 2$ e $\alpha = v/2$.

A.2.7 Variabile di Student

Una variabile casuale X si dice essere di Student di parametro v , con $v > 0$, se ha densità di probabilità

$$f_X(x) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{\pi v} \Gamma(\frac{v}{2})} \left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}} \quad (x \in \mathbb{R}),$$

dove $\Gamma(\cdot)$ è la funzione gamma di Eulero definita nella (A.2).

(i) *Valore medio*

$$E(X) = 0 \quad (\text{per } v > 1).$$

(ii) *Varianza*

$$D^2(X) = \frac{v}{v-2} \quad (\text{per } v > 2).$$

A.2.8 Variabile di Fisher

Una variabile casuale X si dice essere di Fisher di parametri v_1 e v_2 , con v_1 e v_2 interi positivi, se ha densità di probabilità

$$f_X(x) = \begin{cases} \frac{\Gamma(\frac{v_1+v_2}{2})}{\Gamma(\frac{v_1}{2}) \Gamma(\frac{v_2}{2})} \left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}} x^{\frac{v_1}{2}-1} \left(1 + \frac{v_1}{v_2}x\right)^{-\frac{v_1+v_2}{2}} & \text{per } x > 0, \\ 0 & \text{altrimenti,} \end{cases}$$

dove $\Gamma(\cdot)$ denota la funzione gamma di Eulero definita nella (A.2).

A.2. VARIABILI CASUALI CONTINUE

(i) *Valore medio*

$$E(X) = \frac{v_2}{v_2 - 2} \quad (\text{per } v_2 > 2).$$

(ii) *Varianza*

$$D^2(X) = \frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)} \quad (\text{per } v_2 > 4).$$

(iii) *Coefficiente di variazione*

$$CV(X) \stackrel{\text{def}}{=} \frac{D(X)}{E(X)} = \sqrt{\frac{2(v_1 + v_2 - 2)}{v_1(v_2 - 4)}} \quad (\text{per } v_2 > 4).$$

A.2.9 Variabile di Laplace

Si dice che X è una variabile casuale di Laplace (o variabile casuale esponenziale doppia, o variabile casuale esponenziale bilatera) se ha densità di probabilità

$$f_X(x) = \frac{1}{2\beta} \exp\left(-\frac{|x-\alpha|}{\beta}\right) \quad (x \in \mathbb{R}),$$

con $\beta > 0$ e $\alpha \in \mathbb{R}$.

(i) *Funzione generatrice dei momenti*

$$M_X(t) \stackrel{\text{def}}{=} E(e^{tX}) = \frac{e^{\alpha t}}{1 - (\theta t)^2} \quad \left(-\frac{1}{\beta} < t < \frac{1}{\beta}\right).$$

(ii) *Valore medio*

$$E(X) = \alpha.$$

(iii) *Varianza*

$$D^2(X) = 2\beta^2.$$

(iv) *Coefficiente di variazione*

$$CV(X) \stackrel{\text{def}}{=} \frac{D(X)}{E(X)} = \sqrt{2} \frac{\beta}{\alpha}.$$

(v) *Momenti centrali di ordine dispari*

$$E\{[X - E(X)]^{2n+1}\} = 0 \quad (n = 0, 1, \dots).$$

(vi) *Momento centrale del quarto ordine*

$$E\{[X - E(X)]^4\} = 24\beta^4.$$

(vii) Coefficiente di asimmetria

$$\frac{E\{[X - E(X)]^3\}}{D^3(X)} = 0.$$

(viii) Coefficiente di piccatezza

$$\frac{E\{[X - E(X)]^4\}}{D^4(X)} = 6.$$

A.2.10 Variabile normale inversa

Una variabile casuale X si dice normale inversa di parametri μ e λ ($\mu, \lambda \in \mathbb{R}^+$) se ha densità di probabilità

$$f_X(x) = \begin{cases} \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right] & \text{per } x > 0, \\ 0 & \text{altrimenti.} \end{cases}$$

(i) Funzione generatrice dei momenti

$$M_X(t) \stackrel{\text{def}}{=} E(e^{tX}) = \exp\left\{\frac{\lambda}{\mu}\left[1 - \left(1 - \frac{2\mu^2 t}{\lambda}\right)^{1/2}\right]\right\} \quad \left(t < \frac{\lambda}{2\mu^2}\right).$$

(ii) Valore medio

$$E(X) = \mu.$$

(iii) Varianza

$$D^2(X) = \frac{\mu^3}{\lambda}.$$

(iv) Coefficiente di variazione

$$CV(X) \stackrel{\text{def}}{=} \frac{D(X)}{E(X)} = \sqrt{\frac{\mu}{\lambda}}.$$

(v) Momento centrale del terzo ordine

$$E\{[X - E(X)]^3\} = 3 \frac{\mu^5}{\lambda^2}.$$

(vi) Momento centrale del quarto ordine

$$E\{[X - E(X)]^4\} = 3 \frac{\mu^6}{\lambda^2} \left(1 + 5 \frac{\mu}{\lambda}\right).$$

(vii) Coefficiente di asimmetria

$$\frac{E\{[X - E(X)]^3\}}{D^3(X)} = 3 \sqrt{\frac{\mu}{\lambda}}.$$

(viii) Coefficiente di piccatezza

$$\frac{E\{[X - E(X)]^4\}}{D^4(X)} = 3 \left(1 + 5 \frac{\mu}{\lambda}\right).$$

A.2.11 Variabile lognormale

Se Y è una variabile casuale normale di media μ e varianza σ^2 , la variabile casuale $X = e^Y$ si dice lognormale di parametri μ e σ^2 , con $\mu \in \mathbb{R}$ e $\sigma \in \mathbb{R}^+$. La sua densità di probabilità è pertanto

$$f_X(x) = \begin{cases} \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right] & \text{per } x > 0, \\ 0 & \text{altrimenti.} \end{cases}$$

(i) Momenti intorno all'origine

$$E(X^r) = \exp\left(\mu r + \frac{\sigma^2}{2} r^2\right) \quad (r \in \mathbb{R}).$$

(ii) Valore medio

$$E(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right).$$

(iii) Varianza

$$D^2(X) = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}.$$

(iv) Coefficiente di variazione

$$CV(X) \stackrel{\text{def}}{=} \frac{D(X)}{E(X)} = \sqrt{e^{\sigma^2} - 1}.$$

(v) Momento centrale del terzo ordine

$$E\{[X - E(X)]^3\} = \exp\left(3\mu + \frac{3}{2}\sigma^2\right) (e^{3\sigma^2} - 3e^{\sigma^2} + 2).$$

(vi) Momento centrale del quarto ordine

$$E\{[X - E(X)]^4\} = \exp(4\mu + 2\sigma^2) (e^{6\sigma^2} - 4e^{3\sigma^2} + 6e^{\sigma^2} - 3).$$

(vii) Coefficiente di asimmetria

$$\frac{E\{[X - E(X)]^3\}}{D^3(X)} = \frac{e^{3\sigma^2} - 3e^{\sigma^2} + 2}{(e^{\sigma^2} - 1)^{3/2}}.$$

(viii) Coefficiente di piccatezza

$$\frac{E\{[X - E(X)]^4\}}{D^4(X)} = \frac{e^{6\sigma^2} - 4e^{3\sigma^2} + 6e^{\sigma^2} - 3}{(e^{\sigma^2} - 1)^2}.$$

Appendice B
Tabelle

Tabella 1
Distribuzione normale standard

I valori di $\Psi(x) = (2\pi)^{-1/2} \int_0^x e^{-z^2/2} dz$ sono riportati per alcune scelte di x . Ad esempio a $x = 2.54$ (ottenuto come $2.5 + .04$) corrisponde $\Psi(x) = 0.4945$.

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0474	.0517	.0557	.0595	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998

Tabella 2
Distribuzione normale standard: valori di z_α

α	0.10	0.05	0.025	0.01	0.005	0.001	0.0005	0.00005	0.000005
z_α	1.282	1.645	1.960	2.326	2.576	3.090	3.291	3.891	4.417

Tabella 3
Distribuzione chi-quadrato: valori di $\chi_{\alpha,v}^2$

$v \setminus \alpha$.995	.99	.975	.95	.05	.025	.01	.005
1	0.0000393	0.000157	0.000982	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	0.0717	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	12.832	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	45.558
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	43.194	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 4
Distribuzione di Student: valori di $t_{\alpha,v}$

$v \setminus \alpha$.10	.05	.025	.01	.005
1	3.078	6.314	12.706	41.821	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
∞	1.282	1.645	1.960	2.326	2.576

Tabella 5 (prima parte)
Distribuzione di Fisher: valori di $F_{0.01,v_1,v_2}$

$v_2 \setminus v_1$	1	2	3	4	5	6	7	8	9
1	4052	5.000	5.403	5.625	5.764	5.859	5.928	5.982	6.023
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

Tabella 5 (seconda parte)

Distribuzione di Fisher: valori di $F_{0.01;v_1,v_2}$

$v_2 \setminus v_1$	10	12	15	20	24	30	40	60	120	∞
1	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	99.5
3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
4	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
5	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	4.91
11	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17
30	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Tabella 6 (prima parte)

Distribuzione di Fisher: valori di $F_{0.05;v_1,v_2}$

$v_2 \setminus v_1$	1	2	3	4	5	6	7	8	9
1	161	200	216	225	230	234	237	239	241
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

Tabella 6 (seconda parte)
Distribuzione di Fisher: valori di $F_{0.05;v_1,v_2}$

$v_2 \setminus v_1$	10	12	15	20	24	30	40	60	120	∞
1	242	244	246	248	249	250	251	252	253	254
2	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Indice analitico

- ampiezza di un test, *v.* test
- analisi della varianza, 335–360
 - a due fattori, 352–360
 - ad un fattore, 336–349
- media generale, 338
- piano degli esperimenti, 338, 349–352
- somma dei quadrati, 339
- stima puntuale, 340–342, 353–355
- verifica di ipotesi, 342–349, 355–360
- Blackwell-Rao**
 - teorema di, 118
- campionamento, 3
- campione casuale, 4
 - realizzazione di un, 4
- campo di variazione, 66
- cardine
 - metodo del, 164–166
- chi-quadrato (χ^2), *v.* distribuzione
- classi, 239
- equiampie, 243
- equiprobabili, 244
- coefficiente
 - di correlazione campionario, 329
 - di determinazione, 308
 - di fiducia, 161, 162
 - di variazione, 61
- concentrazione, *v.* stimatore
- correlazione normale, 327
- Cramér-Rao
 - disuguaglianza di, 85
- densità
 - a posteriori, 149
 - a priori, 149
- deviazione standard campionaria, 15, 74, 97
- diagramma delle frequenze, 361–364
- distorsione, *v.* stimatore
- distribuzione
 - a posteriori, 149
 - a priori, 149
 - beta, 402–403
 - binomiale, 389–390
 - binomiale negativa, 394–395
 - campionaria, 10
 - chi-quadrato (χ^2), 27–37, 403–404, 411
 - di Bernoulli, 388–389
 - di Fisher, 42–49, 404–405, 413–416
 - di Laplace, 405–406
 - di Pascal, 393–394
 - di Poisson, 390–391
 - di Student, 37–42, 404, 412
 - esponenziale, 399–400
 - gamma, 401–402
 - geometrica, 392–393
 - ipergeometrica, 395–396
 - lognormale, 407–408
 - multinomiale, 396–397
 - normale, 17–26, 398–399, 410
 - normale bivariata, 53
 - normale inversa, 57–62, 406–407
 - normale multivariata, 49–52
 - uniforme continua, 397–398
 - uniforme discreta, 387–388
 - equazioni normali, 270
 - errore
 - di I tipo, 202.
 - di II tipo, 202

quadratico medio, 75
famiglia
completa, 122–125
esponenziale, 114
fattorizzazione
teorema di, 109, 113
Fisher, v. distribuzione
Fisher, R.A., 125
fit, v. indice di fit
funzione
di distribuzione empirica, 375
di verosimiglianza, 126
gamma, 27
Glivenko-Cantelli
teorema di, 384
grado di fiducia, 161
indice di fit, 309
inferenza
non parametrica, 8
parametrica, 8
statistica, 8
intervallo di previsione, 297
intervallo fiduciario, 159–164
ampiezza di, 163
centrale, 174
per deviazioni standard, 184
per differenza tra medie, 176, 178
per medie, 168, 172, 174
per popolazioni di Bernoulli, 187
per popolazioni esponenziali, 189, 191
per rapporti di varianze, 185
per varianze, 180, 183
ipotesi
alternativa, 200
composta, 199
nulla, 200
parametrica, 200
semplice, 199
statistica, 199
test di, 200, 201
verifica delle, 199
istogramma, 367–375
cumulativo, 375–381

legge empirica del caso, 363
Lehmann-Scheffé
teorema di, 124
massima verosimiglianza
metodo della, 125–128
stimatore di, v. estimatore
media campionaria, 10–12, 16, 23, 25, 32, 70, 79, 88–90, 92, 95, 97, 99, 103, 113, 125, 129, 130, 136, 138, 140, 142, 146, 150, 151, 157, 160, 168, 174, 175, 203
mediana, 65, 193, 198
campionaria, 66, 70, 195
di variabile esponenziale, 71
minimi quadrati, 267
pesati, 311
momenti
metodo dei, 145
momento
campionario, 10, 70, 99–100, 145
empirico, 10
Neyman-Pearson
lemma di, 211
teorema di, 211
Pearson, E.S., 241
Pearson, K., 145
pivot, v. cardine
metodo del, 164–166
popolazione, 3
binomiale, 111, 122, 129
di Bernoulli, 76, 100, 106, 108, 121, 127, 130, 156, 187, 218
di Poisson, 89, 109, 123, 135, 155
esponenziale, 89, 129, 189, 191, 220
gamma, 115, 128, 146
geometrica, 222
normale, 17, 23, 25, 32, 35, 41, 72, 74, 80, 88, 92, 95, 99, 112, 115, 124, 136, 141, 142, 150, 159, 160, 168, 169, 171, 174–176, 178, 180, 183–185, 203, 207, 213, 215, 225, 228, 229, 233, 235, 237
normale inversa, 116, 138

uniforme, 101, 118, 131, 147, 148, 193, 195
potenza, v. test
proporzioni
differenze tra, 249
quantile, 192
superiore, 21
rapporto di verosimiglianze, 224
test del, 223–239
rappresentazione dei dati, 361
regione
critica, 201
ampiezza di una, 205
di rifiuto, 201
regressione, 263
bivariata, 263
coeffienti di, 266
equazioni di, 264
lineare, 266
lineare multivariata, 323
multivariata, 267
non lineare, 278
normale, 290
polinomiale, 319
stima puntuale nella, 283
verifica di ipotesi nella, 301
residuo, 289
standardizzato, 310
riassunti campionari, 9
scarto quadratico medio, 137
campionario, 138, 146
spazio dei parametri, 200
statistica
completa, 118, 122–125
sufficiente, 105–125, 132, 141
statistiche d'ordine, 63, 144, 193, 195
massimo di, 65, 131, 148
minimo di, 65, 163
stima
dei parametri, 68
dell'intervallo fiduciario, 162
di massima verosimiglianza, 126, 158
intervallare, 68, 159–161

UNIVERSITÀ FEDERICO DI NAPOLI
BIBLIOTECA CENTRALE
FACOLTÀ DI SCIENZE MATEMATICHE
INV. 891

BIBLIOTECA FACOLTA'
SC.MM.FF.NN.

Serie di matematica e fisica
(Collana fondata da G. Stampacchia e diretta da G. Vidossich)

1. J.P. Cecconi e G. Stampacchia, *Analisi matematica*. Vol. I: *Funzioni di una variabile*
2. J.P. Cecconi e G. Stampacchia, *Analisi matematica*. Vol. II: *Funzioni di più variabili*
4. G. Strang, *Algebra lineare e sue applicazioni*
5. L.C. Piccinini, G. Stampacchia e G. Vidossich, *Equazioni differenziali ordinarie in R^n (problemi e metodi)*
6. J.P. Cecconi, L.C. Piccinini, G. Stampacchia, *Esercizi e problemi di analisi matematica*. Vol. I: *Funzioni di una variabile*; Vol. II: *Funzioni di più variabili*
7. G. Andreatta, W. Rungaldier, *Statistica matematica - Problemi ed esercizi risolti*
8. E. Acerbi, L. Modica, S. Spagnolo, *Problemi scelti di Analisi Matematica I*
9. H. Brezis, *Analisi funzionale. Teoria e applicazioni*
10. E. Acerbi, L. Modica, S. Spagnolo, *Problemi scelti di Analisi Matematica II*
11. A. Ambrosetti, I. Musu, *Matematica generale e applicazioni all'economia*
12. A. Bacciotti, F. Ricci, *Analisi matematica I*
13. G. Dell'Antonio, *Elementi di meccanica. I: Meccanica classica*
14. M. Nacinovich, *Elementi di geometria analitica*
15. A. Di Crescenzo, L.M. Ricciardi, *Elementi di statistica*

ERRATA CORRIGE

Il presente fascicolo sostituisce a tutti gli effetti l'indice analitico già inserito nel volume A. Di Crescenzo - L. M. Ricciardi, *Elementi di statistica*, ISBN 88-207-3052-9, alle pp. 357-360.

Indice analitico

NI 1489200
ELEMENTI DI
STATISTICA
A. DICRESCENZO
L.M. RICCIARDI
LIGUORI
EDITORE N 00000705

- ampiezza di un test, *v. test*
- analisi della varianza, 285-306
 - a due fattori, 299-306
 - ad un fattore, 285-297
- media generale, 287
- piano degli esperimenti, 287, 297-299
- somma dei quadrati, 288
- stima puntuale, 289-291, 300-301
- verifica di ipotesi, 291-297, 302-306
- Blackwell-Rao
 - teorema di, 102
- campionamento, 3
- campione casuale, 3
 - realizzazione di un, 4
- campo di variazione, 58
- cardine
 - metodo del, 141-142
- chi-quadrato (χ^2), *v. distribuzione classi*, 205
 - equiampie, 209
 - equiprobabili, 210
- coefficiente
 - di correlazione campionario, 280
 - di determinazione, 263
 - di fiducia, 139
 - di variazione, 53
- concentrazione, *v. stimatore*
- correlazione normale, 279
- Cramér-Rao
 - disuguaglianza di, 73
- densità
 - a posteriori, 128
 - a priori, 127
- deviazione standard campionaria, 12, 64, 83
- diagramma delle frequenze, 307-310
- distorsione, *v. stimatore*
- distribuzione
 - a posteriori, 128
 - a priori, 127
 - beta, 342-343
 - binomiale, 331-332
 - binomiale negativa, 335-336
 - campionaria, 8
 - chi-quadrato (χ^2), 23-32, 343-344, 351
 - di Bernoulli, 330-331
 - di Fisher, 36-42, 344-345, 353-356
 - di Laplace, 345-346
 - di Pascal, 334-335
 - di Poisson, 332-333
 - di Student, 32-36, 344, 352
 - esponenziale, 340-341
 - gamma, 341-342
 - geometrica, 333-334
 - ipergeometrica, 336-337
 - lognormale, 347

- multinomiale, 337–338
 normale, 14–22, 339–340, 350
 normale bivariata, 46
 normale inversa, 48–52, 346
 normale multivariata, 42–45
 uniforme continua, 338–339
 uniforme discreta, 329–330
- equazioni normali, 231
 errore
 di I tipo, 173
 di II tipo, 173
 quadratico medio, 65
- famiglia
 completa, 105–107
 esponenziale, 97
- fattorizzazione
 teorema di, 94, 97
- Fisher, *v.* distribuzione
 Fisher, R.A., 107
 fit, *v.* indice di fit
 funzione
 di distribuzione empirica, 319
 di verosimiglianza, 108
 gamma, 23
- Glivenko-Cantelli
 teorema di, 326
- grado di fiducia, 139
- indice di fit, 264
- inferenza
 non parametrica, 7
 parametrica, 7
 statistica, 7
- intervallo di previsione, 254
- intervallo fiduciario, 137–141
 ampiezza di, 140
 centrale, 149
 per deviazioni standard, 158
 per differenza tra medie, 151, 153
- per medie, 144, 148, 150
 per popolazioni di Bernoulli, 161
 per popolazioni esponenziali, 163, 164
 per rapporti di varianze, 158
 per varianze, 155, 157
- ipotesi
 alterativa, 172
 composta, 171
 nulla, 172
 parametrica, 171
 semplificata, 171
 statistica, 171
 test di, 172
 verifica delle, 171
- istogramma, 312–318
 cumulativo, 318–323
- legge empirica del caso, 308
- Lehmann-Scheffé
 teorema di, 106
- massima verosimiglianza
 metodo della, 107–109
 stimatori di, *v.* stimatore
- media campionaria, 8–10, 13, 20, 21, 27, 60, 68, 76, 77, 79, 82, 83, 85, 88, 97, 107, 110, 111, 116, 118, 120, 122, 125, 128, 129, 134, 137, 138, 144, 150, 175
- mediana, 57, 165, 169
 campionaria, 58, 61, 167
 di variabile esponenziale, 61
- minimi quadrati, 229
 pesati, 299
- momenti
 metodo dei, 124
- momento
 campionario, 8, 60, 85–86, 124
 empirico, 8

- Neyman-Pearson
 lemma di, 181
 teorema di, 181
- Pearson, E.S., 207
 Pearson, K., 124
 pivot, *v.* cardine
 metodo del, 141–142
- popolazione, 3
 binomiale, 95, 105, 110
 di Bernoulli, 65, 86, 91, 93, 104, 108, 111, 134, 160, 187
 di Poisson, 77, 93, 106, 116, 132
 esponenziale, 77, 110, 162, 164, 189
 gamma, 98, 109, 125
 geometrica, 190
 normale, 14, 20, 21, 27, 30, 35, 63, 64, 69, 76, 79, 80, 82, 83, 85, 96, 99, 107, 117, 121, 122, 128, 137, 138, 144, 146, 147, 150, 151, 153, 155, 157, 158, 175, 177, 183, 185, 193, 195, 197, 200, 202, 203
 normale inversa, 100, 118
 uniforme, 86, 101, 112, 126, 127, 166, 167
- potenza, *v.* test
 proporzioni
 differenze tra, 214
- quantile, 165
 superiore, 17
- rapporto di verosimiglianze, 192
 test del, 191–196
- rappresentazione dei dati, 307
- regione
 critica, 172
- ampiezza di una, 176
 di rifiuto, 172
- regressione, 225
 bivariata, 225
 coefficienti di, 227
 equazioni di, 226
 lineare, 227
 lineare multivariata, 276
 multivariata, 228
 non lineare, 238
 normale, 248
 polinomiale, 272
 stima puntuale nella, 242
 verifica di ipotesi nella, 257
- residuo, 247
 standardizzato, 264
- riassunti campionari, 7
- scarto quadratico medio, 117
 campionario, 118, 125
- spazio dei parametri, 172
- statistica
 completa, 102, 105–107
 sufficiente, 90–107, 113, 121
- statistiche d'ordine, 55, 123, 166, 167
- massimo di, 56, 112, 126
 minimo di, 56, 140
- stima
 dei parametri, 59
 dell'intervallo fiduciario, 139
 di massima verosimiglianza, 108, 135
- intervallare, 59, 137–139, 169, 250
- puntuale, 7, 59, 107, 124, 128, 167
- stimatore, 59
 a varianza uniformemente minima, 72
- asintoticamente corretto, 83–86, 90, 113

- concentrazione di uno, 80
consistente, 84–90, 124, 126,
127, 309, 322
corretto, 59–82, 85–87, 103,
106, 107, 113, 121, 127,
309, 322
debolmente consistente, 89
di Bayes, 127–135
di massima verosimiglianza,
108–124, 126, 133, 310
distorsione di uno, 60, 83
efficiente, 72, 76, 84, 106
efficienza asintotica di uno, 70
efficienza di uno, 68–70, 76
errore sistematico di uno, 60
fortemente consistente, 89–90
lineare corretto, 67, 68, 82
non distorto, 60
per quantili, 167
pienamente efficiente, 76–78,
121
proprietà asintotiche di uno,
83–90
relativamente più efficiente, 68,
104
Student, *v.* distribuzione

tabella di contingenza, 219
test
 ampiezza di un, 176
 bilaterale, 176
 chi-quadrato, 205–213
 di conformità, 206, 219
 potenza di un, 177
 semplicemente più potente, 181–
 191
 uniformemente più potente,
 181
 unilaterale, 176

varianza campionaria, 8–10, 27,
30, 60, 69, 80, 85, 153
verifica delle ipotesi, *v.* ipotesi
verosimiglianza, *v.* funzione