Volume 2
Elementi di Inferenza Statistica

Nicola Torelli
Roberta Pappadà

Università degli Studi di Trieste

Dipartimento di Scienze Economiche, Aziendali,
Matematiche e Statistiche "Bruno de Finetti" (DEAMS)

Corso di laurea in Statistica e Informatica
per l'Azienda, la Finanza e l'Assicurazione

a.a. 2019-2020

Premessa

Queste dispense sono rivolte agli studenti del corso di Inferenza Statistica del Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche (DEAMS) "Bruno de Finetti".

Gli argomenti trattati riguardano primi elementi della teoria e dei metodi dell'inferenza statistica. Nel testo si farà riferimento a numerosi risultati che sono illustrati in *Volume 1: Richiami e Complementi di Calcolo delle Probabilità* (i riferimenti incrociati al Volume 1 presenti nel testo appariranno con la sigla 'RCCP').

Si ringraziano gli studenti e tutti i lettori che vorranno manifestare suggerimenti utili al miglioramento del presente testo. In particolare, si ringrazia Vincenzo Gioia (DEAMS) per i preziosi commenti e suggerimenti forniti agli autori.

Trieste, 24 ottobre 2024

Indice

Capitolo 1

Inferenza statistica: concetti introduttivi

1.1 Introduzione

La ricerca nelle scienze empiriche, siano esse le cosiddette scienze "dure" (medicina, farmacologia, chimica, fisica, astronomia, etc.) o le scienze economico-sociali (economia, sociologia, psicologia, scienze dell'educazione, etc.), è basata sulla osservazione di fatti e dati empirici. La statistica si occupa della raccolta, del trattamento e della sintesi dei dati e dei metodi per trarre conclusioni dagli stessi e di ottenere previsioni oppure di prendere decisioni. Il metodo statistico è pertanto generalmente riconosciuto come il più razionale e completo approccio metodologico a supporto della ricerca empirica.

Si conviene di distinguere gli ambiti della statistica in relazione agli obiettivi specifici di fasi diverse dell'analisi dei dati. Le fasi di raccolta, organizzazione, pulizia e le tecniche di sintesi dei dati (anche attraverso opportuni strumenti grafici per la visualizzazione degli stessi) al fine di rendere visibili le loro caratteristiche salienti, sono tipicamente comprese nell'ambito della **statistica descrittiva** o dell'**analisi esplorativa dei dati**. Quando invece l'obiettivo dell'analisi va oltre la semplice descrizione dei dati disponibili e si propone di trarre conclusioni sulla popolazione da cui i dati sono tratti, di formulare e individuare modelli, di confermare o verificare teorie, di fornire supporto a decisioni o di ottenere previsioni, allora è necessario ricorrere ai metodi dell'**inferenza statistica**.

Sebbene la separazione fra le fasi di descrizione/esplorazione dei dati e

quella dell'inferenza statistica non sia nella pratica così netta e gli strumenti e i concetti di statistica, descrittiva o inferenziale, vadano sempre impiegati in modo integrato, sono spesso i metodi dell'inferenza statistica quelli di maggior interesse. I metodi dell'inferenza statistica consentono di trarre conclusioni che non sono infatti riferite solo allo specifico insieme di dati disponibile. La situazione più tipica, e che richiameremo più volte, fa riferimento alla possibilità di estendere a una **popolazione** molto ampia le evidenze empiriche osservate solo su un piccolo insieme di unità di quella popolazione (il **campione**).

Se ci si chiede, ad esempio, quale sia la proporzione di pazienti che avranno una completa remissione di un dato sintomo (ad esempio il mal di testa) fra coloro che hanno assunto un certo farmaco, si dovrà somministrare il farmaco a tutti coloro che hanno (o avranno) il mal di testa (in una popolazione di unità virtualmente infinita) e osservare se vi è remissione del sintomo. Nella realtà ci si limiterà a osservare se avviene la remissione del mal di testa solo su un ristretto gruppo (un campione) di pazienti cui si somministrerà il farmaco (e questi saranno i dati di cui si disporrà); impiegando i metodi dell'inferenza statistica si potranno trarre conclusioni su quale proporzione non avrà il mal di testa se il farmaco fosse stato assunto dall'intera popolazione. È immediato osservare che la conclusione che si trae utilizzando i metodi dell'inferenza statistica, essendo basata su un numero (anche molto) limitato di osservazioni relative a un sottoinsieme di unità della popolazione di interesse, è soggetta a incertezza e il risultato che si ottiene dipende dalle specificità del campione osservato. Tuttavia, l'uso dei metodi di inferenza statistica consente di misurare, controllare e, se possibile, ridurre l'inevitabile margine di incertezza.

1.2 I dati come realizzazione di un modello aleatorio

Nel contesto dell'inferenza statistica si cerca quindi di trarre conclusioni generali a partire dai dati raccolti su un campione di unità di una popolazione (molto) più ampia che in genere è troppo costoso o impossibile osservare nella sua interezza. A questo fine, si suppone che i dati (le variabili osservate sugli elementi del campione) siano ottenuti secondo uno schema che li renda assimilabili a realizzazioni di un modello aleatorio le cui caratteristiche almeno

in parte possono essere assunte come note. La variabilità che si osserva nei dati è, in questa impostazione, il risultato di un modello aleatorio.

Nel seguito si farà riferimento ad una situazione e a un modello semplice che richiameremo spesso. Si supporrà che l'interesse sia su una singola variabile aleatoria (v.a.) Y e sulla sua distribuzione nell'intera popolazione. La variabile Y è definita sul supporto R_Y e si dispone di n valori osservati $(y_1, y_2, \ldots, y_n)$ sugli elementi del campione che sono considerati realizzazioni del vettore di variabili aleatorie $(Y_1, Y_2, \ldots, Y_n)$. Per le componenti di tale vettore aleatorio si assume l'indipendenza (ovvero Y_i e Y_j sono indipendenti, per ogni coppia i, j ove $i \neq j$) e che abbiano distribuzione identica e pari a quella di Y. La legge di distribuzione di Y è specificata da una funzione di densità (o di probabilità, in relazione alle caratteristiche del supporto R_Y). Il valore n è detto **ampiezza (o dimensione) campionaria**.

Si dice in tal caso che $(Y_1, Y_2, \ldots, Y_n)$ sono variabili aleatorie *indipendenti e identicamente distribuite* (in breve, **i.i.d.**). La legge di distribuzione comune per le variabili aleatorie Y_i è fornita da $f(y)$ (che è la stessa legge che caratterizza la variabile Y nella popolazione). Le condizioni poste caratterizzano quello che viene denominato campione casuale. In questo caso, il problema di inferenza statistica equivale a individuare il modello distributivo ignoto della variabile aleatoria Y avendo osservato realizzazioni da una successione di variabili aleatorie $(Y_1, Y_2, \ldots, Y_n)$ indipendenti e identiche in distribuzione alla Y.

Definizione 1.1 (Campione casuale)**.** La successione di variabili aleatorie $(Y_1, Y_2, \ldots, Y_n)$ è un **campione casuale** tratto dalla variabile aleatoria Y definita sul supporto R_Y se $(Y_1, Y_2, \ldots, Y_n)$ sono variabili aleatorie i.i.d. e con la medesima legge di probabilità della variabile Y. I valori della n-pla campionaria appartengono all'insieme R_Y^n, detto spazio campionario.

Se $(Y_1, Y_2, \ldots, Y_n)$ è un campione casuale da Y allora la funzione di densità (probabilità) congiunta del vettore aleatorio è completamente specificata essendo semplicemente il prodotto delle funzioni di densità (probabilità) marginali.

Esempio 1.1. Si dispone di un campione casuale $(Y_1, Y_2, \ldots, Y_n)$ da Y che

è distribuito secondo la funzione di densità $f_Y(y)$, la densità del campione è

$$f_{Y_1,Y_2,\ldots,Y_n}(y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} f(y_i). \tag{1.1}$$

▲

Si osservi che nel calcolo delle probabilità si pone di solito il problema di valutare la probabilità di eventi relativi a possibili determinazioni di una variabile aleatoria Y di legge nota. Nell'inferenza statistica si ragiona invece su un problema inverso: osservata una sequenza di valori $y_1, y_2, \ldots, y_n$ che sono supposti essere determinazioni di una variabile aleatoria Y, si vuole fare inferenza sulla legge distributiva di Y che è, del tutto o in parte, ignota.

1.3 Inferenza parametrica e non-parametrica

L'inferenza sulla variabile Y equivale a fornire indicazioni sulla funzione di densità $f_Y(y)$, o di probabilità $p_Y(y)$, che caratterizza la sua distribuzione. Spesso la conoscenza del fenomeno in esame rende possibile e ragionevole restringere l'interesse ad uno specifico sottoinsieme di leggi distributive. Per brevità, nel seguito si converrà di indicare genericamente con $f(y)$ la legge di distribuzione di Y omettendo, salvo nel caso in cui questo generi confusione, la distinzione fra il caso in cui si tratti di una funzione di densità o di probabilità. In particolare, in molti casi è possibile restringere l'interesse al caso in cui la funzione $f(y)$ è nota nella sua forma funzionale ma la sua completa specificazione dipende da una o più costanti, dette **parametri**. In tal caso, si specifica il modello per Y sul quale si vuole fare inferenza con la funzione $f(y;\theta)$, ove il parametro $\theta \in \Theta$ è una costante e Θ, l'insieme dei suoi possibili valori, è denominato **spazio parametrico**.

La conoscenza di θ implica la conoscenza completa della legge di distribuzione di Y nella popolazione (a tal fine, in genere, si assume anche che a due distinti valori di θ corrispondano distinte funzioni $f(y;\theta)$).

In questo caso si parla di **inferenza parametrica** e i dati campionari sono impiegati per fare inferenza sul valore del parametro θ.

Nel caso invece in cui non si facciano assunzioni precise sulla forma di $f(y)$ si parlerà di **inferenza non-parametrica**.

Esempio 1.2. Modello parametrico bernoulliano. Si dispone di un campione casuale $(Y_1, Y_2, \ldots, Y_n)$ da Y che ha distribuzione di probabilità bernoulliana $f(y;\theta) = \theta^y(1-\theta)^{1-y}$ (il campione è allora composto da una sequenza di valori pari a zero o uno di lunghezza n, lo spazio parametrico è l'intervallo reale $\Theta = [0,1]$, lo spazio campionario è l'insieme delle sequenze di valori 0 o 1, $\{0,1\}^n$; la distribuzione di probabilità del campione è

$$f(y_1, y_2, \ldots, y_n;\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i} \tag{1.2}$$

dove $s = \sum_{i=1}^n y_i$ il numero dei valori pari a 1 (successi) osservati nel campione, allora la probabilità di osservare un dato campione con s successi è

$$f(y_1, y_2, \ldots, y_n;\theta) = \theta^s(1-\theta)^{n-s} \tag{1.3}$$

▲

Esempio 1.3. Modello parametrico gaussiano. Se si dispone di un campione casuale $(Y_1, Y_2, \ldots, Y_n)$ da Y che si suppone distribuito secondo la legge normale $Y \sim \mathcal{N}(\mu, \sigma^2)$, con σ^2 costante fissata, allora lo spazio parametrico è l'intera retta reale $\mathbb{R}$, lo spazio campionario l'insieme $\mathbb{R}^n$ e la distribuzione di probabilità del campione è

$$\begin{aligned} f(y_1, y_2, \ldots, y_n;\mu) =& \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \\ =& (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu)^2}. \end{aligned}$$

▲

Tale modello parametrico definisce una famiglia di variabili aleatorie gaussiane tutte con la stessa forma potendo variare solo la posizione, ovvero il valore medio, della variabile aleatoria.

In numerosi casi si considera anche il modello parametrico in cui il parametro è la coppia (μ, σ^2) per cui si otterranno espressioni analoghe per la densità del campione ma si scriverà $f(y_1, y_2, \ldots y_n;\mu,\sigma^2)$ e lo spazio parametrico sarà $\mathbb{R} \times \mathbb{R}^+$.

1.4 Alcuni esempi di modelli statistici

Saranno ora forniti alcuni semplici esempi di modelli statistici che definiscono prototipi di problemi inferenziali, mostrando come la formalizzazione di un modello possa essere adattato ad alcune situazioni reali. Si vuole evidenziare anche come la specificazione di un modello per un dato problema sia un elemento da analizzare criticamente caso per caso; spesso sarà in effetti necessario riformulare il modello adottato se esso si rivela poco adatto a descrivere (o spiegare) i dati.

1.4.1 Qual è la proporzione di palline bianche nell'urna?

L'esempio che segue è forse il più semplice e immediatamente comprensibile e il modello che si adotta descrive in modo del tutto adeguato il meccanismo che genera i dati; esso pertanto verrà usato anche in seguito come prototipo per introdurre alcune procedure inferenziali.

Si supponga di avere un'urna composta di palline bianche e nere e sia p la proporzione ignota di palline bianche. L'urna è chiusa e non è possibile esaminare il suo contenuto. È solo possibile estrarre con reinserimento un numero fissato di palline, diciamo n, e osservare se ogni pallina estratta è bianca o nera. Sia $y_i = 1$ se osserviamo una pallina bianca alla i-esima estrazione e 0 altrimenti. La sequenza di n valori 0 e 1 osservata, $(y_1, y_2, \dots y_n)$, può essere assimilata alla determinazione delle variabili aleatorie $(Y_1, Y_2, \dots, Y_n)$ con $Y_i \sim Be(p)$ tutte indipendenti e identicamente distribuite.

Il modello statistico che genera i dati è quindi assimilabile a un modello parametrico bernoulliano come quello definito nell'esempio 1.2. In questo specifico caso il modello statistico bernoulliano descrive in modo del tutto adeguato il problema. Nel paragrafo successivo si mostrerà che vi sono casi in cui è sensato adottare tale modello, tuttavia esso potrebbe riflettere solo in parte le caratteristiche reali dei dati che si osservano.

Per questo specifico problema è utile inoltre notare che sarebbe del tutto equivalente pensare che la sequenza osservata di palline sia un campione di dimensione n dalla popolazione costituita dalla teorica sequenza infinita di risultati dei lanci. Si può infatti ritenere che la variabile Y costituisca la descrizione della popolazione di infinite estrazione nella quale, in virtù della legge dei grandi numeri, la proporzione di palline bianche sarà pari a p. Il problema di inferenza diventa quindi quello di capire se e come, attraverso

un campione limitato di lanci, si possa inferire sulle caratteristiche della popolazione e quindi su p.

1.4.2 Modelli statistici per l'inferenza su una proporzione

Può apparire adeguato il modello dell'urna illustrato nella sezione precedente quando si voglia fare inferenza su una popolazione in cui è p la proporzione di unità che presentano una caratteristica di interesse. Vi sono numerosi esempi in cui è di interesse tale situazione: si immagini di voler conoscere la proporzione di esseri umani che hanno un particolare gruppo sanguigno o la proporzione di pezzi difettosi che vengono prodotti da un processo industriale oppure, come accade nei sondaggi di opinione o in quelli elettorali, di chiedersi quale sia la proporzione di coloro che nella popolazione di interesse intendono sostenere un dato partito politico o condividono una determinata opinione. Ovviamente l'inferenza statistica è necessaria in quanto si dispone dei dati osservati su un campione, ovvero su un sottoinsieme di elementi, tratto da tali popolazioni, spesso molto ampie.

È evidente come sia possibile fare ancora riferimento al modello statistico bernoulliano analogamente a quanto fatto nell'esempio dell'urna appena visto. In tutti questi casi, si può ipotizzare infatti che la quantità di interesse, ovvero la proporzione p, sia assimilabile al parametro di una variabile aleatoria bernoulliana, e che il valore y_i osservato per un elemento del campione sia una determinazione di una variabile $Y_i \sim Be(p)$. Infatti, se si prendesse una unità a caso da una simile popolazione si avrebbe $P(Y_i = 1) = p$.

Perché l'analogia regga occorre pensare che la selezione degli elementi che fanno parte del campione corrisponda all'estrazione di elementi da un'urna con reinserimento. È questo che consente di assumere che la sequenza $(y_1, y_2, \ldots y_n)$ possa essere vista come una realizzazione del vettore aleatorio $(Y_1, Y_2, \ldots, Y_n)$ di n variabili indipendenti e tutte distribuite secondo la legge $Be(p)$.

Tuttavia, in nessuno dei casi visti sopra, è del tutto realistico pensare che la selezione del campione corrisponda precisamente all'estrazione da un'urna. Il modello dell'estrazione dall'urna è forse ben approssimato nel caso della selezione dei pezzi difettosi (per il quale si può anche pensare che ci si riferisca a una popolazione infinita), lo è molto meno, come discusso anche più avanti,

nel caso di un sondaggio elettorale (dove si ha una popolazione finita e attuare una selezione del campione assimilabile allo schema dell'urna non è semplice).

1.4.3 Modelli statistici per l'inferenza su variabili di conteggio

Si immagini che l'interesse sia quello di individuare un modello statistico adatto a descrivere una variabile discreta Y che corrisponde a un conteggio. Ad esempio, Y potrebbe rappresentare il numero di oggetti comprati da un cliente su un portale di acquisti in rete in una sessione, oppure il numero di coloro che in un dato periodo e in una data area contraggono una certa malattia, o, ancora, il numero di sinistri denunciati dai clienti di una compagnia assicurativa. In questi casi il supporto della variabile Y è l'insieme dei numeri interi non negativi per cui $R_Y = \{0, 1, 2, 3, \dots\}$. Se si vuole adottare un modello parametrico la scelta più ovvia è quella di scegliere fra alcune famiglie di variabili aleatorie discrete.

Si potrebbe assumere, ad esempio, che Y nella popolazione sia descritta da una variabile aleatoria di Poisson di media λ. I valori di un campione $(y_1, y_2, \dots y_n)$ saranno realizzazioni di una sequenza i.i.d. di $Y_i \sim Pois(\lambda)$ con $\lambda > 0$. Se si accetta una tale assunzione distributiva allora il problema di inferenza riguarda il parametro λ (o qualche altra quantità di solito esprimibile come una funzione di λ, si potrebbe infatti voler fare inferenza su $P(Y = 0) = e^{-\lambda}$).

Un'obiezione sensata potrebbe essere che nella realtà non è ragionevole ipotizzare che i dati osservati possano assumere qualsiasi valore intero, anche molto grande, così come sarebbe per le realizzazioni di una variabile aleatoria di Poisson. Il modello statistico è in effetti sempre una rappresentazione imperfetta della realtà, una sua idealizzazione introdotta allo scopo di poter fornire una spiegazione della variabilità osservata nei dati. In effetti, nelle situazioni reali, nessun modello statistico descrive in modo perfetto la realtà ma è tuttavia uno strumento utile per valutare l'incertezza delle procedure di inferenza.

Infine, si osservi che la distribuzione di Poisson è solo una delle possibili scelte. Si potrebbe, infatti, fare ricorso a ipotesi diverse sulla legge distributiva per Y nella popolazione: ad esempio, la binomiale negativa sarebbe una buona alternativa. Essendo quest'ultima caratterizzata da due parametri si potrà rivelare più flessibile della Poisson. In generale, modelli stocastici che

coinvolgono un maggior numero di parametri possono rivelarsi più adeguati a descrivere un fenomeno. Nelle applicazioni della statistica è in effetti sempre presente il dilemma fra scegliere un modello più semplice ma meno capace di descrivere i dati osservati e uno più complesso che rischia di avere un ottimo adattamento ma solo a uno specifico insieme di dati osservato.

1.4.4 Modelli statistici per l'inferenza su una durata

In molti casi la quantità di interesse è il tempo che trascorre prima che si verifichi un dato evento. Ad esempio, il tempo che trascorre prima della rottura di un prodotto meccanico, il tempo di attesa prima della guarigione di coloro che sono stati sottoposti a una terapia o il tempo prima di decidere di riscattare una polizza pensionistica. La variabile Y che descrive tali grandezze può assumere ovviamente solo valori positivi e il tempo trascorso è quindi una variabile continua per cui il supporto è $R_Y = \mathbb{R}^+$.

Il modello esponenziale potrebbe essere adeguato in un simile caso e un campione casuale da tale popolazione potrebbe essere quindi inteso come la realizzazione di una sequenza i.i.d. di variabili $Y_i \sim \text{Esp}(\lambda)$ con il parametro $\lambda > 0$. L'inferenza riguarda quindi il parametro λ, se si ritiene l'ipotesi distributiva corretta.

Anche in questo caso i modelli alternativi che potrebbero essere proposti sono tutti quelli relativi a variabili aleatorie con supporto i reali positivi come, ad esempio, la Gamma, la Log-normale, etc.

Si noti che in questo caso, come per tutti i modelli continui, si adotta una variabile aleatoria le cui determinazioni sono numeri reali (i reali positivi in questo specifico caso). Tuttavia le osservazioni che poi si otterranno, i nostri dati, sono sempre in realtà misurati con un certo livello di approssimazione (lo strumento di misurazione ha in effetti una risoluzione finita) e saranno pertanto di fatto discreti. Vale ancora una volta l'osservazione che qualsiasi modello è un'astrazione che si ritiene fornisca un'adeguata, non perfetta quindi, descrizione della realtà.

Un'ulteriore particolarità, in questo caso, potrebbe derivare dal fatto che le osservazioni campionarie in alcuni casi sono incomplete. Ad esempio, se l'interesse fosse su Y, durata di una lampadina, si potrebbe considerare un campione di lampadine, e annotare il tempo y_i di rottura dell'i-esima lampadina. Se l'osservazione delle lampadine viene estesa fino a un dato momento alcune lampadine sono ancora in funzione e quindi il loro tempo di rottura non verrà osservato. In questo caso si osserva per alcune lampadine il tempo

completo y_i, per altre avrò solo il dato relativo al tempo trascorso fino al termine della mia rilevazione, e questo tempo non è la durata di vita y_i ma è un tempo t_i, $t_i \leq y_i$, detto tempo di censura. Nel fare inferenza si dovrà, pertanto, tenere opportunamente conto che i dati contengono informazione su durate complete (y_i) e su tempi di censura t_i.

1.4.5 Modelli gaussiani per variabili quantitative

L'assunzione che una quantità Y si distribuisce in una popolazione secondo una legge gaussiana (o normale) di media μ e varianza σ^2 è forse quella che si adotta più spesso nei modelli parametrici. Il campione $(y_1, y_2, \ldots, y_n)$ è determinazione di un vettore aleatorio gaussiano di componenti $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, la n-pla campionaria è una sequenza in $\mathbb{R}^n$ e i parametri sono definiti nello spazio parametrico $\mathbb{R} \times \mathbb{R}^+$.

Molte variabili quantitative continue hanno in effetti distribuzione che risulta ben approssimata da una legge gaussiana: di solito tale legge descrive bene grandezze biometriche (la statura, il peso dei neonati, la temperatura corporea, etc), valori di misure di grandezze fisiche (come il diametro di pezzi ottenuti da un processo produttivo, o misurazioni ripetute di una stessa quantità affette da un errore di misura), grandezze monetarie (magari dopo una trasformazione logaritmica), o valori di test standardizzati (punteggi ottenuti dalla somministrazione di molti quesiti, un esempio è il QI o il punteggio di un test tipo INVALSI). In effetti l'azione del teorema del limite centrale (sulla somma di tanti fattori accidentali aleatori) giustifica in molti casi l'assunzione gaussiana.

Si noti che, anche se la normalità implica che la variabile Y sia continua e definita sull'intero asse reale, spesso si usa il modello normale anche per grandezze che ammettono solo valori positivi. Il modello in effetti può rivelarsi adeguato anche in questi casi (ad esempio per le stature umane) in quanto valori insensati (come quelli negativi) per i valori tipici dei parametri del modello nella pratica hanno probabilità nulla di verificarsi.

Non è neppure infrequente che si adotti il modello gaussiano per variabili di natura discreta: ad esempio, nel caso di variabili di conteggio ma con media molto elevata (ad esempio, per la distribuzione del numero di assicurati con una polizza caso morte per agenzia assicurativa). Si ricorda che si è già discusso che un modello continuo comunque deve approssimare osservazioni che saranno per loro natura discrete, inoltre, in questo caso si può pensare ancora alle conseguenze del teorema del limite centrale per cui, ad esempio,

una v.a. di Poisson con parametro elevato (sopra il centinaio) risulta ben approssimata da una gaussiana.

1.4.6 Un modello statistico per valutare l'efficacia di un farmaco

Si torni ora all'esempio richiamato nel paragrafo 1.1, sull'efficacia di un farmaco per il mal di testa. Immaginiamo che l'interesse sia nel valutare una variabile quantitativa Y_F sulla quale ci si aspetta il farmaco agisca; ad esempio il medico potrebbe aspettarsi che la pressione diastolica Y_F (la cosiddetta "minima") dopo avere assunto il farmaco venga ridotta entro valori normali ritenuti non patologici. Come detto, l'interesse è sull'effetto del farmaco sulla popolazione (di chi abbia sintomi di ipertensione) ma sarà possibile misurare il valore di Y_F solo per un campione che ha assunto il farmaco (il deponente F ricorda che si tratta di soggetti che hanno assunto il farmaco). I valori $(y_{F1}, y_{F2}, \ldots, y_{Fn})$ misurati su un campione di coloro che hanno assunto il farmaco possono essere determinazioni di variabili gaussiane di media μ_F e varianza σ_F^2, cioè $Y_F \sim \mathcal{N}(\mu_F, \sigma_F^2)$.

L'assunzione di gaussianità potrebbe essere abbastanza valida in questo caso, come accade per molte grandezze biometriche. Se anche, nella realtà, il range dei valori osservabili fosse limitato ai valori positivi (considerazione che quindi vale anche per il valore della media μ_F). Si noti che il modello potrebbe essere anche uniparametrico se si supponesse noto il valore di σ_F^2 nella popolazione, cosa possibile se si avessero ampi studi precedenti che indicano che la variabilità di Y_F è nota e pari a uno specifico valore. Si è quindi nel caso illustrato dall'esempio 1.3.

Si immagini ora che l'obiettivo non sia quello di determinare la distribuzione di Y_F per coloro che hanno assunto il farmaco (che significa poter dire qualcosa sul valore di μ_F se si assume nota la varianza σ_F^2) ma quello di voler rispondere a un quesito più generale e interessante. Si consideri, quindi, il problema di voler stabilire se il farmaco sia efficace o meno. Il problema può essere affrontato cercando di capire se la distribuzione della variabile di interesse sia diversa a seconda che si assuma o meno il farmaco. In altri termini, questo equivale a confrontare la distribuzione della variabile Y_F con la variabile Y_P che descrive il valore della variabile di interesse per coloro che non assumono il farmaco. Se queste due distribuzioni sono uguali allora si potrà concludere che il farmaco non ha alcun effetto. Per poter verifica-

re l'efficacia di un farmaco occorre quindi disporre di un secondo campione, $(y_{P1}, y_{P2}, \ldots, y_{Pm})$, dalla popolazione delle misure di ipertensione per chi non assume il farmaco. A questo punto quindi il modello statistico si presenta leggermente più complesso. Vi sono due popolazioni per le quali si potrebbe assumere un analogo modello gaussiano: la prima popolazione è quella di chi assume il farmaco, in cui la pressione è distribuita secondo la legge $Y_F \sim \mathcal{N}(\mu_F, \sigma_F^2)$; la seconda è quella di chi non assume il farmaco $Y_P \sim \mathcal{N}(\mu_P, \sigma_P^2)$, e da queste vengono estratti 2 campioni separati di dimensione n e m, rispettivamente. L'inferenza riguarda quindi una coppia di variabili aleatorie indipendenti. Il modello ha complessivamente ora 4 parametri, tuttavia si potrebbe ricorrere a un modello più semplice assumendo $\sigma_F^2 = \sigma_P^2$.

Resta da osservare che l'inferenza sulla differenza fra le due popolazioni ci consente di trarre la conclusione dell'efficacia del farmaco se si riesce a escludere che la differenza fra Y_F e Y_P sia dovuta a cause diverse dall'azione del farmaco. Questo in genere si realizza attraverso un disegno di rilevazione per cui si individuano i soggetti (i malati di ipertensione in questo esempio) che devono essere trattati e quelli che non verranno trattati mediante un'assegnazione casuale. Inoltre si procede in modo che tutti i soggetti vengano apparentemente trattati, anche se alcuni (quelli per cui si osserva Y_P) prenderanno sì un farmaco ma senza principio attivo (il cosiddetto placebo). Pertanto il paziente non sa se appartiene o meno alla popolazione dei trattati con il farmaco. Infine, è buona prassi evitare che anche chi somministra la cura (il medico di solito) sappia se a ciascun soggetto sia stato somministrato il farmaco o il placebo.

1.4.7 Un modello statistico per il volume del legno degli alberi di ciliegio

Si supponga ora che l'interesse sia sulla quantità misurata con il volume (in m^3) di legno che si ottiene da un tronco di un albero di ciliegio. Anche in questo caso si potrebbe ritenere ragionevole che il volume di legno Y, se osservato su una popolazione di alberi di ciliegio, sia distribuito secondo una legge gaussiana, per cui si potrebbe assumere $Y \sim \mathcal{N}(\mu, \sigma^2)$. Tale assunzione tuttavia rischia di rivelarsi poco aderente alla realtà se, ad esempio, gli alberi di ciliegio fossero molto diversi tra loro in altezza. Allora si può rendere il modello più realistico assumendo che esistano popolazioni di al-

beri di altezza x differente. Cioè si potrebbe ipotizzare che l'assunzione di normalità sia relativa alla popolazione degli alberi di altezza comune x per cui $Y_x \sim \mathcal{N}(\mu_x, \sigma_x^2)$. Per l'inferenza su tale modello dovremmo supporre di osservare tanti campioni per quante sono le altezze x (così come nell'esempio precedente si ottenevano due campioni separatamente per i trattati col farmaco e con il placebo). Tuttavia, in questo ultimo caso, nella pratica, ciascun albero mostrerà valori di x diversi. L'inferenza sulle variabili Y_x può procedere se si aggiungono delle assunzioni semplificatrici sulle medie μ_x e sulle varianze σ_x^2).

Una prima semplificazione potrebbe essere quella di assumere che le varianze non dipendano da x per cui $Y_x \sim \mathcal{N}(\mu_x, \sigma^2)$ e quindi si suppone che la varianza della quantità di legno prodotta sia la stessa qualunque sia l'altezza dell'albero (condizione di *omoschedasticità*). Questa semplificazione però non è di grande aiuto se continua a valere che le popolazioni gaussiane hanno tante medie diverse quanti sono i valori di x. Si potrebbe ancora assumere che i valori di μ_x siano legati a x attraverso una semplice relazione per cui $\mu_x = f(x)$ ove, ad esempio, $f(x) = \beta_0 + \beta_1 x$. A questo punto il modello statistico per la variabile Y diventa $Y_x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$. Esso dipende da 3 parametri soltanto e sarà possibile fare inferenza su tale modello se si osservano per un campione casuale di unità le coppie di valori $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$.

Il modello appena visto è il più semplice esempio di modello (di regressione) lineare che costituisce una delle basi per poter pensare a modelli più complessi nei quali le ipotesi distributive sulla variabile di interesse sono legate ad alcune caratteristiche osservate (l'altezza x nel caso in questione).

1.5 Il campionamento in pratica

La selezione di un campione di unità che rappresenti la popolazione è nella pratica un'operazione tutt'altro che scontata e banale. Un campione è rappresentativo della popolazione se non discrimina alcuni elementi della popolazione o non mostra preferenze per alcuni altri elementi: si richiede quindi che ciascuna unità della popolazione abbia una probabilità non nulla di essere inclusa nel campione e che tale probabilità non sia dipendente dal valore della variabile di interesse Y_i. Campioni costruiti secondo tale regola sono campioni probabilistici. Si ricorre quindi all'uso di tecniche di selezione del campione probabilistiche ove la probabilità di inclusione è nota e control-

lata dallo statistico e in effetti buona parte dei ragionamenti inferenziali si fondano su questo presupposto.

Il caso più semplice di campione probabilistico è quello che equivale a uno schema di campionamento (di selezione quindi delle unità da inserire nel campione e su cui trarre le osservazioni) che riproduca più fedelmente possibile lo schema della selezione di elementi da un'urna con reinserimento. Si può pensare a un'urna molto grande che contenga tutte le unità della popolazione di interesse; e sarà pertanto un'urna virtuale visto che in alcuni casi la popolazione potrà essere composta da un numero di elementi infinito. Si noti che si vorrebbe che la selezione di un insieme di n unità della popolazione sia tale che, una volta osservati sulle unità del campione i valori $(y_1, y_2, \dots y_n)$, essi siano una sequenza di determinazioni di quello che abbiamo definito un campione casuale, ovvero un vettore di variabili aleatorie i.i.d. $(Y_1, Y_2, \dots, Y_n)$ tutte distribuite come la variabile Y nella popolazione.

È evidente che se si pensa a una popolazione infinita non ha ragione di essere la distinzione fra estrazione con o senza reinserimento. In entrambi i casi infatti l'estrazione casuale di elementi dalla popolazione genera eventi indipendenti. In alcuni dei casi sopra esemplificati potrebbe essere in effetti legittimo pensare che le popolazioni da cui si traggono i campioni (tutti gli umani oppure tutti i pezzi prodotti da una macchina) siano composte da infiniti elementi (ancorché virtuali). L'obiettivo è capire qual è la proporzione di pezzi difettosi in tutta la potenziale produzione di quel processo e non in quella di un particolare giorno; e si vuole inferire, ad esempio, sulla proporzione di coloro che hanno uno specifico gruppo sanguigno presente nella popolazione degli esseri umani in generale e non per uno specifico gruppo (quelli in vita adesso). Resta da assicurarsi comunque, in quest'ultimo caso, che il campione sia selezionato in modo tale da non favorire unità che hanno caratteristiche particolari. Nel caso dei gruppi sanguigni si potrebbe osservare il dato sul gruppo sanguigno 0+ scegliendo casualmente alcuni fra coloro che si recano a fare una donazione del sangue. Se in passato fossero stati incoraggiati a fare donazioni soprattutto quelli con il gruppo 0+, allora il campione non sarebbe a rigore un campione probabilistico e effetti sono sistematicamente sovra rappresentati i donatori con gruppo 0+.

Nel caso di una popolazione finita, composta cioè da un numero finito di unità N, possiamo comunque usare lo stesso modello statistico, assumendo che esso descriva, almeno approssimativamente, la variabile di interesse Y e immaginando che la selezione delle unità avvenga con reinserimento, così da mantenere l'indipendenza e l'identica distribuzione.

Si pensi ad esempio al caso dei sondaggi elettorali o, più in generale, dei sondaggi di opinione che sono a tutti gli effetti delle indagini campionarie. Qui il riferimento è a una popolazione reale composta da tutti coloro che hanno diritto di voto e si tratta quindi di una popolazione finita. Avrebbe in tal caso senso che la selezione del campione di unità avvenga in blocco (cioè senza reinserimento). Se tuttavia la dimensione della popolazione finita N è molto grande (dell'ordine di qualche milione per un sondaggio elettorale in Italia) e molto più grande di n, usare come approssimazione lo schema del campionamento senza reinserimento è ancora un'approssimazione accettabile.

Nel caso di tali indagini campionarie, è di solito molto difficile assicurare che ogni unità della popolazione abbia probabilità non nulla di entrare nel campione. Attuare una selezione che assomiglia allo schema dell'estrazione dall'urna non è spesso possibile e in effetti la realtà si discosta in modo sensibile da tale schema. Si contattano in effetti spesso solo le unità che sono reperibili in liste che talvolta nemmeno comprendono tutti gli elementi della popolazione cui si è interessati (gli elenchi telefonici ad esempio). Oppure, anche se si riuscisse a ottenere una lista che comprende tutte le unità della popolazione da cui selezionare il campione casualmente, poi accade che non sia possibile reperire molte unità oppure alcune di esse decidono di non voler rispondere (in tal caso il campione è frutto quindi di selezione non casuale o, addirittura di auto-selezione quando è l'unità stessa a decidere se partecipare all'indagine e quindi entrare o meno nel campione). I fattori di selezione non casuale concorrono a rendere il modello statistico che si adotta (in questo esempio potrebbe essere adeguato lo schema della bernoulliana) solo una approssimazione molto vaga e imprecisa della realtà. Tali scostamenti ed errori rispetto allo schema dell'urna non possono essere trattati e controllati utilizzando gli schemi del calcolo delle probabilità: essi sono infatti detti errori non campionari in quanto gli scostamenti fra quello che si osserva nel campione e quello che si osserverebbe se si potesse sondare l'intera popolazione non dipendono solo dall'aleatorietà associata al processo di selezione casuale del campione. La presenza di tali fattori di disturbo, degli errori non campionari - che non è sempre possibile tenere sotto controllo - limita il valore delle conclusioni ottenute utilizzando la teoria statistica che è in effetti costruita nelle situazioni ideali del campione casuale. Per tale motivo accade di frequente che i risultati di indagini campionarie, come i sondaggi elettorali, possano condurre a conclusioni non accurate, sebbene si usino correttamente le tecniche dell'inferenza statistica che però corrispondono a una situazione ideale dalla quale, in casi come questi, ci si discosta sensibilmente.

La validità del modello statistico dipende quindi dalla effettiva procedura di estrazione del campione dalla popolazione che deve permettere di assumere che i dati estratti siano assimilabili a un campione casuale (o comunque a campioni probabilistici). Compito dello statistico è pertanto anche quello di assicurarsi che la selezione del campione si avvicini a quanto assunto poi nella costruzione del modello statistico. Esistono a tal proposito sviluppi della teoria statistica che si occupano del problema di ottenere campioni probabilistici (e quindi rappresentativi) in situazioni molto complesse, di trarre inferenza su quantità relative a una popolazione finita (e non solo sui parametri di un modello parametrico), e anche di correggere e limitare gli effetti indesiderati legati all'uso di campioni (auto) selezionati in modo non completamente casuale o in caso vi siano mancate risposte. Il tema del campionamento da popolazione finita, quello delle tecniche per trarre inferenza da dati ottenuti con schemi di campionamento più complicati di quello casuale o anche con campioni non casuali richiedono il ricorso a tecniche e modelli statistici molto più elaborati che vanno ben oltre un testo introduttivo sui metodi dell'inferenza statistica. Nel seguito di questa trattazione si assumerà di essere nella situazione ideale del campionamento casuale tenendo però sempre a mente che nella fase di critica di un modello non va trascurato l'aspetto relativo alla effettiva selezione campionaria utilizzata.

È per i motivi fin qui illustrati che lo statistico deve sempre essere coinvolto in tutte le fasi di un'indagine scientifica che voglia utilizzare il metodo statistico: a partire dalla progettazione della ricerca e dal disegno dell'indagine, alla fase di raccolta e analisi preliminare dei dati, fino alla fase di costruzione del modello statistico, all'uso delle tecniche inferenziali di analisi dei dati, alla critica e revisione del modello e infine, al termine del processo, alla formulazione delle conclusioni e delle decisioni conseguenti.

1.6 I problemi dell'inferenza statistica

1.6.1 Il modello statistico parametrico

Come già menzionato, l'obiettivo dell'inferenza statistica è principalmente quello di estrarre informazioni dai dati osservati su un numero limitato di unità. Se si suppone che i dati osservati siano le realizzazioni di un particolare modello statistico (stocastico quindi) l'obiettivo diviene quello di ottenere informazioni sul processo generatore di quei dati.

Poiché ci si dovrà accontentare nella maggioranza dei casi di un insieme di dati limitato, l'inferenza sul processo generatore può rivelarsi complicata e si dovranno trarre conclusioni caratterizzate da un grado di incertezza che si cerca di ridurre e misurare. Si noti che se i dati sono generati da un modello che è stocastico questi mostreranno un'intrinseca variabilità. Inoltre, poiché tali dati sono ottenuti a seguito di una selezione casuale di unità della popolazione ci si attende che se l'operazione di selezione del campione fosse ripetuta più volte si otterrebbero sempre insiemi di dati diversi. Per quanto visto, obiettivo dell'inferenza statistica è quello di trarre conclusioni da un singolo insieme di dati che abbiano validità generale e non riflettano le particolarità di quello specifico insieme di dati.

Da qui in avanti l'attenzione sarà prevalentemente su modelli statistici parametrici come quelli già definiti nella sezione 1.3, per i quali vale quanto segue.

Definizione 1.2 (Modello statistico parametrico)**.** Sia Y una variabile aleatoria distribuita secondo la funzione $f_Y(y;\theta)$ su un supporto R_Y.

1. θ è un parametro che assume valori in Θ, lo spazio parametrico. Nel seguito si considera prevalentemente il caso di un parametro scalare (ovvero esso è un valore reale come, ad esempio, nel caso $Y \sim \text{Esp}(\theta)$). Si espliciterà, se necessario, il caso in cui θ è un parametro vettoriale, ad esempio se $Y \sim \mathcal{N}(\mu, \sigma^2)$ con μ e σ^2 entrambi non noti e lo spazio parametrico sarà definito opportunamente.
2. Il campione di dati osservati $(y_1, y_2, \ldots, y_n)$ è realizzazione del vettore aleatorio $(Y_1, Y_2, \ldots, Y_n)$ ove Y_i è indipendente da Y_j, $\forall i \neq j$, e tutte le Y_i hanno l'identica distribuzione di Y. L'insieme dei possibili valori del campione varia in uno spazio che in generale è composto da tutte le possibili n-ple campionarie $(y_1, y_2, \ldots, y_n) \in R_Y^n$.

In alternativa, alla luce delle considerazioni svolte nei precedenti paragrafi si può esprimere il problema di inferenza statistica (parametrica) come segue:

Sia Y una variabile statistica la cui distribuzione nella popolazione è descritta dalla funzione di densità (o di probabilità) $f_Y(y;\theta)$. Si ha completa conoscenza della distribuzione di Y nella popolazione se si conosce il valore del parametro θ.

1.6.2 Le tecniche di inferenza statistica

I problemi di inferenza statistica che si pongono una volta osservato un campione $(y_1, y_2, \ldots, y_n)$ riguardano cosa si possa dire sul valore del parametro θ e quindi sulla distribuzioni di Y e come si valuta l'incertezza di tali affermazioni.

Alcune domande che riflettono l'inferenza sul parametro θ potrebbero essere poste come segue:

1. Quale fra i possibili valori di θ è maggiormente plausibile alla luce dei dati campionari?
2. È possibile definire un insieme di valori di θ che più di altri sono plausibili avendo osservato un dato campione?
3. Se si definisce un valore (o anche più valori) di θ di particolare interesse, si può concludere che i dati osservati provengono da una popolazione Y caratterizzata da quello specifico valore?
4. Sulla base dei dati osservati, e delle conseguenti nostre congetture su θ, si è in grado di stabilire quale nuovo valore di Y sarà più plausibile osservare?

Le domande formulate sono esemplificative di alcune delle procedure tipiche dell'inferenza statistica. In particolare, le prime due attengono ai problemi di **stima**, rispettivamente **puntuale** e **intervallare**. La terza è relativa al problema della **verifica di ipotesi** mentre la quarta domanda attiene ai problemi di **previsione**.

Ciascuna di queste procedure inferenziali fornisce delle risposte incerte visto che in fondo sono conclusioni basate sui dati osservati su un (piccolo) campione mentre le affermazioni sono riferite all'intera popolazione. Le modalità con cui si valuta l'attendibilità delle risposte fornite da tali procedure e l'incertezza ad esse connessa varia a seconda della procedura scelta (e come si vedrà anche a seconda dell'approccio inferenziale scelto).

Nel seguito quindi si introdurranno elementi di teoria della stima puntuale e intervallare e della verifica di ipotesi parametriche e sarà inoltre fornito qualche cenno ai problemi di previsione.

Resta infine un problema più generale. Ovvero, il modello statistico proposto è complessivamente compatibile con i dati osservati? E fra due modelli alternativi quale si rivela più adatto? Affrontare questi problemi equivale a chiedersi se il modello statistico adottato sia adeguato o meno. Alcuni degli esempi proposti sopra già delineavano tali questioni. Se nel caso di dati di

conteggio è possibile proporre la Poisson come modello generatore dei dati allora è possibile in alternativa anche usare la binomiale negativa. Chiedersi se un modello è adeguato o quale fra due modelli parametrici alternativi sia preferibile per inferire a partire dai dati osservati ha a che vedere con problemi di **selezione** del modello e di **controllo diagnostico**. Tali temi sono molto rilevanti soprattutto nella trattazione di modelli più complessi.

Ci si limita quindi a fornire solo alcuni cenni su questi ultimi temi senza però dimenticare il noto aforisma attribuito allo statistico G.E.P. Box:

> ...all models are wrong, but some are useful...

Esso ammonisce che un modello statistico nella applicazioni concrete può e deve sempre essere oggetto di riflessioni critiche che possono condurre a specificazioni diverse (e quindi a risultati diversi).

Va inoltre ricordata quella che viene citata come una regola aurea nella costruzione di qualsiasi modello esplicativo della realtà nota come "rasoio di Occam" (che deriva da riflessioni del filosofo medioevale inglese Guglielmo da Occam). Essa consiglia di scegliere fra due modelli esplicativi in grado di descrivere (quasi) ugualmente bene la realtà quello che richiede meno assunzioni. Nel nostro caso quindi, se due modelli statistici parametrici, ad esempio, forniscono entrambi buone descrizioni dei dati o buone previsioni è buona regola scegliere quello che è meno complesso (ovvero che ha meno parametri).

1.7 Gli approcci all'inferenza statistica

Le risposte fornite dalle procedure di inferenza statistica hanno degli inevitabili margini di incertezza; in fondo si sta tentando di trarre conclusioni su qualcosa che non è osservato (la popolazione) se non parzialmente. Lo statistico quindi deve poter fornire delle misure che dicano quanto affidabili siano e che qualità abbiano i risultati inferenziali e confrontare eventualmente procedure alternative. Gli approcci con cui si deriva e si misura l'accuratezza delle conclusioni inferenziali possono tuttavia essere di diversa natura.

1.7.1 L'approccio frequentista

Per chiarire questo ultimo aspetto, si considerino i dati $(y_1, y_2, \ldots y_n)$ generati dal modello statistico proposto in 1.3. Il parametro sul quale si vuole fare

inferenza è la media μ della variabile Y che è una gaussiana di varianza nota. Si potrebbe proporre come valore plausibile per μ nella popolazione la seguente funzione dei dati campionari $\bar{y} = \sum_{1=1}^{n} y_i/n$. Poiché ci si basa su un campione finito di n dati, $\bar{y}$ non sarà uguale al parametro μ e non si è in grado di dire quanto tale valore possa essere vicino al parametro (altrimenti si conoscerebbe μ). Tuttavia, essendo il campione determinazione di un vettore aleatorio anche $\bar{y}$ sarà un valore aleatorio prodotto dalla variabile aleatoria $\overline{Y} = \sum_{1=1}^{n} Y_i/n$ che, in virtù dell'assunzione di campionamento casuale, è noto essere distribuita come una $\mathcal{N}(\mu, \sigma^2/n)$. Si può quindi dire che il valore $\overline{y}$ proviene da una v.a. $\overline{Y}$ che assume valori vicini al vero e ignoto valore μ e come misura della qualità di questa procedura si può ad esempio guardare alla varianza di $\overline{Y}$.

L'esempio illustra come la valutazione dell'incertezza in questa procedura inferenziale si basi essenzialmente sull'analisi delle proprietà della variabile aleatoria $\overline{Y}$. Ciò è rilevante in quanto la distribuzione della variabile aleatoria dipende dal parametro e in questo caso fornisce valori che sono non troppo distanti dal valore vero μ se la sua varianza è piccola.

Si noti che μ è una costante ignota per cui non è possibile fare affermazioni del tipo: "È probabile che il valore vero di μ sia compreso in un determinato intervallo". Le affermazioni che si possono fare riguardano i valori che si potrebbero ottenere a seguito del campionamento, e sono del tipo: "I valori della variabile $\overline{Y}$ saranno mediamente non troppo lontani dal valore ignoto μ" (resta da definire con precisione cosa vuol dire "lontani"). Si sa tuttavia che $\bar{y}$ è una determinazione della variabile aleatoria $\overline{Y}$.

La variabile aleatoria $\overline{Y}$ descrive quindi la distribuzione dei valori che si otterrebbe se si ripetesse più volte il campionamento ottenendo ogni volta un campione diverso. Quindi può essere interpretata come la distribuzione di frequenza dei valori $\bar{y}$ che otterrei se si generassero diversi insiemi di dati di dimensione n da quel modello statistico.

L'approccio all'inferenza basato essenzialmente sullo studio delle v.a. che descrivono quantità che sono funzioni del campione casuale è detto **frequentista** o **classico**. Per valutare la qualità delle procedure inferenziali applicate ad un singolo campione si ragiona sui valori che si potrebbero ottenere applicando la stessa procedura immaginando di ripetere il campionamento tantissime volte (**principio del campionamento ripetuto**).

1.7.2 L'approccio bayesiano

Un approccio alternativo per valutare la qualità delle informazioni su un parametro deducibili a partire da un campione di dati si basa sul fatto che si assume che il parametro oggetto dell'inferenza sia una determinazione di una variabile aleatoria. I dati, il campione osservato, sono quindi ancora prodotti da un modello generatore stocastico, come nell'esempio precedente ove i dati sono determinazioni di una gaussiana di media μ. Ma, sempre con riferimento all'esempio, se si assume che il valore μ oggetto dell'inferenza è prodotto da una variabile aleatoria allora si potrà pensare di ottenere una distribuzione di probabilità per la variabile aleatoria che descrive μ che sia aggiornata tenendo conto delle informazioni fornite dal campione osservato. Sarà pertanto necessario disporre di una distribuzione di probabilità **iniziale** su μ che riflette la conoscenza (o ignoranza) sul parametro prima di osservare il campione e dei dati campionari generati da un meccanismo generatore specificato dal modello statistico. Questi due ingredienti permettono di ricavare una nuova distribuzione di probabilità *finale* per la variabile aleatoria $\mu|y_1, y_2, \ldots, y_n$, che si avvale delle informazioni ottenute dal campionamento.

Si noti che in questo caso, a differenza dell'approccio frequentista, si possono fare affermazioni probabilistiche sul parametro. Si potrà ad esempio calcolare la probabilità che il parametro μ sia compreso in un dato intervallo e senza fare appello agli ipotetici risultati che si potrebbero ottenere se si ripetesse il campionamento più volte.

Tale approccio all'inferenza è detto approccio **bayesiano** anche (e soprattutto) perché la distribuzione di probabilità condizionata finale $\mu|y_1, y_2, \ldots, y_n$ si deriva applicando il teorema di Bayes. Un'introduzione all'inferenza bayesiana sarà discussa nel capitolo 6.

Capitolo 2

Statistiche e distribuzioni campionarie

2.1 Statistiche campionarie

Nel precedente capitolo si è osservato che il campionamento casuale da una popolazione, in cui la variabile di interesse Y è distribuita secondo una legge $f(\cdot)$, ci peremette di ottenere n valori $(y_1, y_2, \ldots, y_n)$ che sono assunti essere determinazioni del vettore aleatorio $(Y_1, Y_2, \ldots, Y_n)$. Il vettore aleatorio $(Y_1, Y_2, \ldots, Y_n)$ ha componenti indipendenti e tutte distribuite identicamente alla variabile Y. Le procedure di inferenza statistica si basano su elaborazione dei valori campionari, ovvero su funzioni del vettore campionario, dette **statistiche**. Essendo le statistiche funzioni di valori aleatori, sono esse stessa variabili aleatorie. Sono statistiche quindi semplici sintesi dei valori campionari, come la media o la devianza o anche il range $\max(y_1, y_2, \ldots, y_n) - \min(y_1, y_2, \ldots, y_n)$.

Definizione 2.1 (Statistica)**.** Dato un campione casuale $(Y_1, Y_2, \ldots, Y_n)$, si chiama **statistica** una qualsiasi funzione $T = t(Y_1, Y_2, \ldots, Y_n)$ del campione. T non dipende da alcun parametro incognito.

La distribuzione di probabilità della statistica $T = t(Y_1, Y_2, \ldots, Y_n)$, coinvolgendo il campione $(Y_1, Y_2, \ldots, Y_n)$, prende il nome di *distribuzione campionaria* e T è detta variabile aleatoria campionaria. Si consideri l'evento

$$t(Y_1, Y_2, \ldots, Y_n) \leq y \tag{2.1}$$

per un dato valore y. Allora $t(Y_1, Y_2, \ldots, Y_n)$ ha funzione di ripartizione data da

$$F(y) = P(t(Y_1, Y_2, \ldots, Y_n) \leq y) = P((Y_1, Y_2, \ldots, Y_n) \leq I_y), \tag{2.2}$$

dove I_y è l'insieme dello spazio campionario di $(Y_1, Y_2, \ldots, Y_n)$ in cui la diseguaglianza (2.1) è soddisfatta. Si evince quindi che la (2.2) dipende dalla distribuzione di probabilità del campione $(Y_1, Y_2, \ldots, Y_n)$

$$f_{Y_1,\ldots,Y_n}(y_1, y_2, \ldots, y_n) = f(y_1)f(y_2)\ldots f(y_n)$$

e quindi dal modello descrittivo della popolazione $f(y; \theta)$.

Esempio 2.1. Sia $(Y_1, \ldots, Y_n)$ un campione casuale estratto da una densità $f(\cdot; \theta)$. Allora $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ e $R_n = \frac{1}{2}[\min\{Y_1, \ldots, Y_n\} + \max\{Y_1, \ldots, Y_n\}]$ sono esempi di statistiche. Si noti che se f è la funzione di densità di una distribuzione normale di parametri θ e $\sigma^2 = 1$, e θ è incognita, allora $\overline{Y}_n - \theta$ non è una statistica perché dipenderebbe da θ. ▲

2.2 Momenti campionari

I momenti campionari sono definiti come segue.

Definizione 2.2 (Momenti Campionari)**.** Dato un campione casuale, $(Y_1, Y_2, \ldots, Y_n)$, dalla v.a. Y avente legge $f(\cdot)$, il **momento campionario (assoluto)** r-mo, o di ordine r, viene definito come

$$M'_r = \frac{1}{n}\sum_{i=1}^n Y_i^r. \tag{2.3}$$

Il **momento campionario r-mo rispetto a** $\overline{Y}_n$ viene definito come

$$M_r = \frac{1}{n}\sum_{i=1}^n (Y_i - \overline{Y}_n)^r. \tag{2.4}$$

2.2.1 Media campionaria

La (2.3) per $r = 1$ definisce *media campionaria*, indicata con $\overline{Y}$ (o $\overline{Y}_n$)

$$\overline{Y}_n := \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{2.5}$$

Il seguente importante risultato sulla *distribuzione campionaria della media* mostra che, a prescindere dal modello descrittivo della popolazione, è possibile determinare due caratteristiche della distribuzione di probabilità della media campionaria $\overline{Y}$, ovvero la sua media e la sua varianza.

Teorema 2.1. Sia $(Y_1, Y_2, \ldots, Y_n)$ un campione casuale da una v.a. Y avente legge di distribuzione $f(\cdot)$, media μ e varianza finita σ^2. Se $\overline{Y} = (1/n)\sum_{i=1}^{n} Y_i$, allora

$$E(\overline{Y}) = \mu_{\overline{Y}} = \mu \quad \text{e} \quad V(\overline{Y}) = \sigma^2_{\overline{Y}} = \frac{\sigma^2}{n}.$$

Dimostrazione. La dimostrazione segue immediatamente dal Corollario 4.1-RCCP, considerando che

$$\overline{Y} = \frac{1}{n}Y_1 + \frac{1}{n}Y_2 + \cdots + \frac{1}{n}Y_n$$

è una combinazione lineare delle variabili $Y_1, Y_2, \ldots, Y_n$, ciascuna delle quali è tale che $E(Y_i) = \mu$; si ha quindi

$$E(\overline{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(Y_i\right) = \frac{1}{n}(n\mu) = \mu$$

In modo analogo, essendo le variabili Y_i indipendenti, dalla 4.12-RCCP, ponendo $a_i = (1/n)$ e considerando che $V(Y_i) = \sigma^2$, $i = 1, \ldots, n$, si ha

$$V(\overline{Y}) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V\left(Y_i\right) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

□

Si è visto che, per qualsiasi variabile casuale Y che possiede varianza finita, la statistica media campionaria ha valore medio pari al valore atteso della popolazione e varianza pari alla varianza della popolazione diviso per n, l'ampiezza del campione. Occorre precisare che $E(\overline{Y}) = \mu$ non implica che la media $\bar{y}$ di n osservazioni è uguale alla media della popolazione μ; ci dice che se consideriamo ripetutamente campioni di ampiezza n e calcoliamo $\bar{y}$ per ciascuno di questi, in media $\bar{y}$ sarà uguale a μ. Analogamente, il risultato $V(\overline{Y}) = \sigma^2/n$ descrive la variabilità degli $\bar{y}$ osservati considerando più campioni della stessa ampiezza, e prova che la dispersione dei valori di $\overline{Y}$ intorno a μ è piccola se l'ampiezza del campione è grande. Ne deriva che per una data popolazione i valori che la media campionaria $\overline{Y}$ assume tenderanno a essere più concentrati intorno a μ rispetto alla legge di riferimento, all'aumentare dell'ampiezza n del campione.
Generalizzando questo risultato, si può affermare che il valore atteso di un momento campionario (assoluto) è uguale al corrispondente momento della popolazione, e che la sua varianza è $(1/n)$ volte una certa funzione dei momenti della popolazione.

Esempio 2.2. Sia $(Y_1, \ldots, Y_n)$ un campione casuale generato da $Y \sim Po(\lambda)$, per la quale è noto che $E(Y) = \lambda$ e $V(Y) = \lambda$. Allora:

$$E(\overline{Y}) = \lambda \quad \text{e} \quad V(\overline{Y}) = \frac{\lambda}{n}.$$

▲

Esempio 2.3. Sia $(Y_1, \ldots, Y_n)$ un campione casuale estratto da Y avente densità gamma di parametri k e λ. Si noti che per una $Y \sim Ga(k, \lambda)$ si ha $E(Y) = k/\lambda$ e $V(Y) = k/\lambda^2$, per cui dal Teorema 2.1 si ottiene

$$E(\overline{Y}) = \frac{k}{\lambda} \quad \text{e} \quad V(\overline{Y}) = \frac{k}{\lambda^2 n}.$$

▲

2.2.2 Varianza campionaria

Si consideri un campione casuale da Y, $(Y_1, Y_2, \ldots, Y_n)$, le cui componenti hanno quindi la stessa distribuzione di probabilità. Pertanto si ha

$$E(Y_i^2) = E(Y^2) = \mu^2 + \sigma^2, \quad i = 1, 2, \ldots, n,$$

dove μ e σ^2 sono, rispettivamente, la media e la varianza della popolazione. Inoltre, in virtù del Teorema 2.1, risulta

$$E(\overline{Y}^2) = \mu^2 + \frac{\sigma^2}{n}$$

da cui

$$\begin{aligned} E(Y_i^2) - E(\overline{Y}^2) &= \mu^2 + \sigma^2 - \mu^2 - \frac{\sigma^2}{n} \\ &= \frac{(n-1)\sigma^2}{n}, \quad i = 1, 2, \ldots, n \end{aligned}$$

Si osservi ora che possiamo scrivere

$$\begin{aligned} E\left(\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right) &= E\left(\sum_{i=1}^{n}(Y_i^2 + \overline{Y}^2 - 2Y_i\overline{Y})\right) \\ &= E\left(\sum_{i=1}^{n} Y_i^2 + n\overline{Y}^2 - 2(n\overline{Y})\overline{Y}\right) \\ &= E\left(\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2\right) \\ &= \sum_{i=1}^{n} E(Y_i^2) - nE(\overline{Y}^2) \\ &= n\left(\frac{(n-1)\sigma^2}{n}\right) = (n-1)\sigma^2 \end{aligned}$$

Ne segue che se definiamo la statistica S^2, detta *varianza campionaria (corretta)*, come

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{2.6}$$

allora si ottiene

$$E(S^2) = \frac{1}{n-1}E\left(\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right) = \frac{(n-1)\sigma^2}{n-1} = \sigma^2$$

provando, in parte, il risultato che segue (nei capitoli successivi verrà chiarito il motivo per il quale la (2.6) definisce la varianza campionaria cosiddetta *corretta*).

Teorema 2.2. Sia $(Y_1, Y_2, \ldots, Y_n)$ un campione casuale da una densità $f(\cdot)$, che ha media μ e varianza finita σ^2. Allora la **varianza campionaria**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

ha valore atteso e varianza date da

$$E(S^2) = \sigma^2 \quad \text{e} \quad V(S^2) = \frac{\sigma^4}{n}\left(\gamma_2 + 2\frac{n}{n-1}\right) \text{ per } n > 1,$$

dove $\gamma_2 = \frac{\overline{m}_4}{\sigma^4} - 3$ è il coefficiente di curtosi della popolazione, essendo $\overline{m}_4$ il momento quarto centrale.

La dimostrazione riguardante la varianza di S^2, più complessa, viene qui omessa, si fornisce però il seguente risultato

$$V(S^2) = \frac{1}{n}\left(\overline{m}_4 - \frac{n-3}{n-1}\sigma^4\right). \tag{2.7}$$

Per quanto riguarda il valore atteso, si osservi che un modo di procedere alternativo alla dimostrazione già vista è il seguente. Possiamo scrivere S^2 come

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \mu)^2 - \frac{n}{n-1}(\overline{Y} - \mu)^2 \tag{2.8}$$

in quanto

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \mu + \mu - \overline{Y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}((Y_i - \mu) - (\overline{Y} - \mu))^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \mu)^2 + \frac{1}{n-1}\sum_{i=1}^{n}(\overline{Y} - \mu)^2 - \frac{2}{n-1}\sum_{i=1}^{n}(Y_i - \mu)(\overline{Y} - \mu) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \mu)^2 + \frac{n}{n-1}(\overline{Y} - \mu)^2 - \frac{2}{n-1}\sum_{i=1}^{n}(Y_i - \mu)(\overline{Y} - \mu), \end{aligned}$$

da cui, essendo $\sum_i Y_i = n\overline{Y}$, si ottiene la relazione sopra riportata. Si ha allora

$$\begin{aligned} E(S^2) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\mu)^2 - \frac{n}{n-1}(\overline{Y}-\mu)^2\right) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}E(Y_i-\mu)^2 - \frac{n}{n-1}E(\overline{Y}-\mu)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\sigma^2 - \frac{\sigma^2}{n-1} = \sigma^2. \end{aligned}$$

Analogamente a quanto visto prima per la media campionaria, il risultato $E(S^2) = \sigma^2$ ci dice che se calcoliamo i valori di S^2 per campioni ripetuti della stessa ampiezza provenienti da una certa popolazione, allora la media di tali valori è uguale alla varianza della popolazione. Sembra ragionevole quindi assumere il valore osservato s^2 di S^2 come stima della varianza della popolazione sebbene, come si vedrà in seguito, questo non rappresenti l'unico criterio possibile.

Esempio 2.4. Sia $(Y_1, \ldots, Y_n)$ un campione casuale generato da una $\mathcal{N}(\mu, \sigma^2)$ per la quale è noto che $\overline{m}_4 = 3\sigma^4$. Allora, risulta

$$V(S^2) = \frac{1}{n}\left(3\sigma^4 - \frac{n-3}{n-1}\sigma^4\right) = \frac{2\sigma^4}{n-1}.$$

▲

2.3 Campionamento da alcuni modelli parametrici

Nei paragrafi che seguono si forniscono alcuni esempi di come ricavare la distribuzione esatta della media campionaria per campioni tratti da alcuni modelli parametrici comunemente adottati.

2.3.1 Dati generati da modelli di Bernoulli e di Poisson

Sia $(Y_1, Y_2, \ldots, Y_n)$ un campione casuale estratto da una distribuzione di Bernoulli

$$f(y) = p^y(1-p)^{1-y} \quad y = 0, 1$$

È noto che $\sum_{i=1}^{n} Y_i$ ha una distribuzione binomiale, cioè

$$P\left(\sum_{i=1}^{n} Y_i = k\right) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n$$

da cui si ricava la distribuzione esatta di $\overline{Y}_n$

$$P\left(\overline{Y}_n = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n \tag{2.9}$$

essendo quest'ultima la probabilità che la media campionaria assuma i valori 0, $1/n$, $2/n, \ldots 1$. Si può anche scrivere

$$P\left(\overline{Y}_n = y\right) = \binom{n}{yn} p^{yn} (1-p)^{n-yn}, \tag{2.10}$$

cioè la media campionaria nel caso di una popolazione Bernoulliana ha una distribuzione proporzionale alla binomiale

$$\overline{Y}_n \sim \frac{1}{n}\text{Bin}(n, p).$$

Sia ora $(Y_1, Y_2, \ldots, Y_n)$ un campione casuale da una distribuzione di Poisson di media λ, allora anche $\sum_{i=1}^{n} Y_i$ ha distribuzione di Poisson di media $n\lambda$ (si veda la 4.18-RCCP). Si ha che $E(\overline{Y}_n) = \lambda$ e $V(\overline{Y}_n) = \lambda/n$ e

$$P\left(\overline{Y}_n = \frac{k}{n}\right) = P\left(\sum_{i=1}^{n} Y_i = k\right) = \frac{(n\lambda)^k}{k!} e^{-n\lambda}, \quad k = 0, 1, 2, \ldots \tag{2.11}$$

In effetti, la statistica $\overline{Y}_n$ non ha distribuzione di Poisson, in quanto assume i valori $0, 1/n$, $2/n, \ldots$, e pertanto non appartiene alla stessa famiglia da cui provengono le variabili aleatorie componenti.

2.3.2 Dati da una distribuzione esponenziale

Sia $(Y_1, Y_2, \ldots, Y_n)$ un campione casuale estratto da una densità esponenziale

$$f(y) = \lambda e^{-\lambda y} \quad y \geq 0,$$

dove $\lambda > 0$.

Per quanto visto nell'esempio 4.8-RCCP, $\sum_{i=1}^{n} Y_i$ ha distribuzione gamma di parametri n e λ, pertanto

$$P\left(\sum_{i=1}^{n} Y_i \leq k\right) = \int_0^k \frac{\lambda^n z^{n-1}}{\Gamma(n)} e^{-\lambda z} \mathrm{d}z, \quad k > 0$$

Quindi possiamo scrivere

$$P\left(\overline{Y}_n \leq \frac{k}{n}\right) = \int_0^k \frac{\lambda^n z^{n-1}}{\Gamma(n)} e^{-\lambda z} \mathrm{d}z, \quad k > 0$$

oppure

$$\begin{aligned} P\left(\overline{Y}_n \leq y\right) &= \int_0^{ny} \frac{\lambda^n z^{n-1}}{\Gamma(n)} e^{-\lambda z} \mathrm{d}z \\ &= \int_0^{y} \frac{\lambda^n (nu)^{n-1}}{\Gamma(n)} e^{-n\lambda u} n \mathrm{d}u \\ &= \int_0^{y} \frac{(n\lambda)^n u^{n-1}}{\Gamma(n)} e^{-n\lambda u} \mathrm{d}u \end{aligned}$$

cioè $\overline{Y}$ ha una distribuzione gamma con parametri n e $n\lambda$.

2.3.3 Dati dalla distribuzione uniforme

Sia $(Y_1, Y_2, \ldots, Y_n)$ un campione casuale estratto da una distribuzione uniforme nell'intervallo $(0, 1]$. Utilizzando la funzione indicatrice, si può scrivere

$$f_{\overline{Y}_1}(y) = f_Y(y) = \mathbf{I}_{(0,1]}(y)$$

Per $n = 2$

$$\begin{aligned} f_{\overline{Y}_2}(y) &= 2 f_{Y_1+Y_2}(2y) \\ &= 2(2y)\mathbf{I}_{(0,1]}(2y) + 2(2-2y)\mathbf{I}_{(1,2]}(2y) \\ &= 2(2y)\mathbf{I}_{(0,1/2]}(y) + 2(2-2y)\mathbf{I}_{(1/2,1]}(y) \\ &= \begin{cases} 4y & \text{per } y \in (0, 1/2] \\ 4(1-y) & \text{per } y \in (1/2, 1] \end{cases} \end{aligned}$$

Per $n = 3$ si perviene alla seguente espressione per la densità di $\overline{Y}_3$:

$$f_{\overline{Y}_3}(y) = \frac{3}{2}(3y)^2 \mathbf{I}_{(0,1/3]}(y) + \frac{3}{2}\left[(3y)^2 - \binom{3}{1}(3y-1)^2\right]\mathbf{I}_{(1/3,2/3]}(y) + \frac{3}{2}\left[(3y)^2 - \binom{3}{1}(3y-1)^2 + \binom{3}{2}(3y-2)^2\right]\mathbf{I}_{(2/3,1]}(y)$$

In generale, la densità esatta di $\overline{Y}_n$ è data da

$$f_{\overline{Y}_n}(y) = \sum_{k=0}^{n-1} \frac{n}{(n-1)!}\Big[(ny)^{(n-1)} - \binom{n}{1}(ny-1)^{(n-1)} + \binom{n}{2}(ny-2)^{(n-1)} - \ldots (-1)^k \binom{n}{k}(ny-k)^{(n-1)}\Big]\mathbf{I}_{(k/n,(k+1)/n]}(y).$$

2.4 Campionamento dalla distribuzione normale

Quanto visto finora stabilisce che la media campionaria ha la stessa media della popolazione da cui proviene il campione e varianza più piccola di quella della popolazione, ottenuta dalla moltiplicazione del fattore $1/n$. Questo paragrafo è invece dedicata ad alcuni risultati che valgono sotto l'ipotesi che la popolazione da cui è tratto il campione sia normale.

2.4.1 Media Campionaria

Teorema 2.3. Sia $(Y_1, Y_2, \ldots, Y_n)$ un campione casuale proveniente da una popolazione normale $\mathcal{N}(\mu, \sigma^2)$. Allora la media campionaria $\overline{Y}_n$ ha una distribuzione normale con media μ e varianza σ^2/n:

$$\overline{Y}_n \sim \mathcal{N}(\mu, \sigma^2/n). \tag{2.12}$$

Dimostrazione. Per dimostrare questo risultato basta considerare che se

$$Y_i \sim \mathcal{N}(\mu, \sigma^2) \quad i = 1, \ldots, n$$

sussistono allora le condizioni del Teorema 4.6-RCCP concernente la distribuzione di probabilità di una combinazione lineare di variabili aleatorie normali indipendenti; si perviene alla (2.12) ponendo $a_i = 1/n$, $i = 1, \ldots, n$ nella 4.20-RCCP. □

Il risultato appena dimostrato mostra che la variabile aleatoria media campionaria ha una distribuzione normale se la popolazione da cui fa il campionamento è distribuita secondo una distribuzione normale, e ciò consente di calcolare, ad esempio, la probabilità (esatta) che $\overline{Y}_n$ sia maggiore/minore di un dato valore, o compreso in qualche intorno fissato del parametro incognito μ.

Esempio 2.5. In un uliveto la produzione per albero si assume descritta da una variabile aleatoria normale con media $\mu = 14.2$ kg e varianza $\sigma^2 = 9.25$. Si vuole determinare la probabilità che in un campione casuale di 10 piante la produzione media risulti minore di 12 kg. Per quanto visto la media campionaria $\overline{Y}_{10}$ ha distribuzione $\mathcal{N}(14.2, 9.25/10)$. Pertanto si ha

$$P(\overline{Y}_{10} < 12) = P\left(Z < \frac{12 - 14.2}{\sqrt{9.25/10}}\right) = \Phi\left(-2.29\right) = 0.011.$$

▲

2.4.2 Varianza Campionaria

Nell'ipotesi che il modello descrittivo della popolazione sia normale, è possibile determinare la distribuzione di probabilità di S^2 definita in (2.6). Una funzione che gioca un ruolo fondamentale nella derivazione della distribuzione di S^2 è la distribuzione *chi-quadrato*, già introdotta nel paragrafo 2.2.5-RCCP come caso particolare della distribuzione gamma.

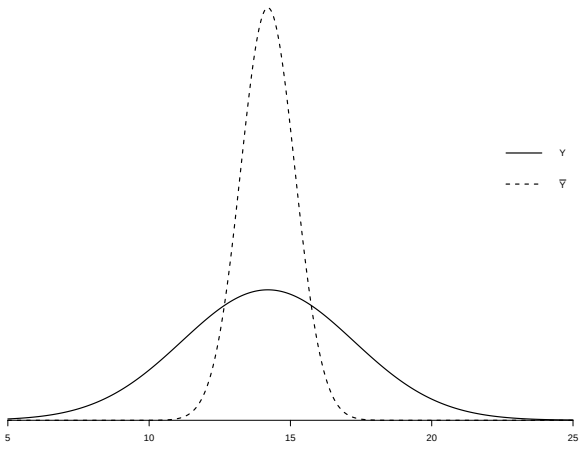


Figura 2.1: Distribuzione di Y e della media campionaria $\overline{Y}$, per $n = 10$, con parametri dati nell'Esempio 2.5.

Definizione 2.3 (Distribuzione chi-quadrato)**.** Se Y è una variabile casuale con densità

$$f(y) = \frac{1}{\Gamma(k/2)} \left(\frac{1}{2}\right)^{k/2} x^{k/2-1} e^{-y/2}, \quad y > 0 \tag{2.13}$$

allora si dice che Y ha una **distribuzione chi-quadrato con k gradi di libertà**. Il parametro k della densità chi-quadrato con k gradi di libertà definita dalla (2.13) è un intero positivo.

Si osservi che, come visto, una densità chi-quadrato è un caso particolare di una densità gamma con parametri $n = k/2$ e $\lambda = 1/2$. Se una variabile casuale Y ha una distribuzione chi-quadrato, cioè $Y \sim \chi^2_k$, si ha

$$E(X) = k, \quad V(X) = 2k$$

e

$$m(t) = \left(\frac{1/2}{(1/2) - t}\right)^{k/2} = \left(\frac{1}{1 - 2t}\right)^{k/2}, \quad t < 1/2.$$

Teorema 2.4. Se Y_i, $i = 1, \ldots, k$, sono variabili aleatorie indipendenti e normalmente distribuite con medie μ_i e varianze σ_i^2, allora

$$U = \sum_{i=1}^{k} \left(\frac{Y_i - \mu_i}{\sigma_i} \right)^2$$

ha una distribuzione chi-quadrato con k gradi di libertà.

Dimostrazione. Se si pone $Z_i = (Y_i - \mu_i)/\sigma_i$, allora $Z_i \sim \mathcal{N}(0, 1)$. Ora, usando la funzione generatrice dei momenti, si ha

$$\begin{aligned} m_U(t) = E(e^{tU}) &= E(e^{t\sum Z_i^2}) \\ &= E\left(\prod_{i=1}^{k} e^{tZ_i^2} \right) = \prod_{i=1}^{k} E(e^{tZ_i^2}) \end{aligned}$$

Si osservi che

$$\begin{aligned} E(e^{tZ^2}) &= \int_{-\infty}^{\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \mathrm{d}z \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)z^2} \mathrm{d}z \\ &= \frac{1}{\sqrt{1-2t}} \int_{-\infty}^{\infty} \frac{\sqrt{1-2t}}{\sqrt{2\pi}} e^{-\frac{1}{2}(1-2t)z^2} \mathrm{d}z \\ &= \frac{1}{\sqrt{1-2t}} \quad t < 1/2, \end{aligned}$$

essendo l'ultimo integrale uguale all'unità dato che rappresenta l'area sotto la curva normale con varianza $1/(1-2t)$. Quindi si può scrivere

$$m_U(t) = \prod_{i=1}^{k} \frac{1}{\sqrt{1-2t}} = \left(\frac{1}{1-2t} \right)^{k/2}, \quad t < 1/2.$$

ottenendo così la funzione generatrice dei momenti di una distribuzione chi-quadrato con k gradi di libertà. □

Il Teorema (2.4) afferma che la somma dei quadrati di variabili casuali normali standardizzate indipendenti ha una distribuzione chi-quadrato con

gradi di libertà pari al numero di variabili nella somma. Notiamo che al risultato appena dimostrato si poteva pervenire banalmente considerando la 4.19-RCCP.

Corollario 2.1. Se $(Y_1, Y_2, \ldots, Y_n)$ è un campione casuale estratto da una distribuzione normale con media μ e varianza σ^2, allora

$$U = \sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{\sigma^2}$$

ha una distribuzione chi-quadrato con n gradi di libertà.

Per il risultato fondamentale che segue è necessario richiamare un altro risultato che riguarda un insieme qualsiasi di variabili aleatorie incorrelate. Siano tali variabili aleatorie $Y_1, Y_2, \ldots, Y_n$, ciascuna avente media μ e varianza σ^2. Allora si può dimostrare che $\overline{Y}$ è incorrelata a ciascuna delle variabili aleatorie $Y_1 - \overline{Y}$, $Y_2 - \overline{Y}$, $\ldots Y_n - \overline{Y}$.

Sia $\overline{Y} = (1/n)\sum_{i=1}^{n} Y_i$ e si ponga $U = Y_1 - \overline{Y}$; è facile verificare che

$$U = Y_1 - \overline{Y} = \frac{n-1}{n}Y_1 - \frac{1}{n}Y_2 - \frac{1}{n}Y_3 - \cdots - -\frac{1}{n}Y_n.$$

Allora, per il Teorema 4.3-RCCP, la covarianza tra $\overline{Y}$ e U è data da

$$\begin{aligned}\mathrm{Cov}(\overline{Y}, U) &= \sum_{i=1}^{n} a_i b_i \sigma_i^2 \\ &= \sigma^2\left(\frac{1}{n}\left(\frac{n-1}{n}\right) + \frac{1}{n}\left(-\frac{1}{n}\right) + \cdots + \frac{1}{n}\left(-\frac{1}{n}\right)\right) \\ &= \sigma^2\left(\frac{n-1}{n^2} - \frac{n-1}{n^2}\right) = 0.\end{aligned}$$

Si ha quindi che $\overline{Y}$ e U sono incorrelate, e analogamente si dimostra che $\overline{Y}$ e $Y_2 - \overline{Y}$, $\overline{Y}$ e Y_3, $\ldots Y_n - \overline{Y}$ sono incorrelate.

Teorema 2.5. Sia $(Y_1, Y_2, \ldots, Y_n)$ un campione casuale estratto da una distribuzione normale con media μ e varianza σ^2. Allora

i La media campionaria $\overline{Y}$ e la varianza campionaria S^2 sono variabili aleatorie indipendenti.

ii La varianza campionaria S^2 ha distribuzione di probabilità proporzionale a quella di una chi-quadrato con $n-1$ gradi di libertà:

$$S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1} \tag{2.14}$$

Dimostrazione. Per la dimostrazione della (i) si osservi che dato un campione casuale di ampiezza n estratto da Y distribuita normalmente con media μ e varianza σ^2, le variabili

$$Z_1 = \frac{Y_1 - \mu}{\sigma},\ Z_2 = \frac{Y_2 - \mu}{\sigma},\ \ldots, Z_n = \frac{Y_n - \mu}{\sigma}$$

sono normali standardizzate indipendenti e risulta

$$\begin{aligned}\overline{Z} =& \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \mu)}{\sigma} = \frac{\overline{Y} - \mu}{\sigma},\\ Z_i - \overline{Z} =& \frac{(Y_i - \mu)}{\sigma} - \frac{(\overline{Y} - \mu)}{\sigma} = \frac{Y_i - \overline{Y}}{\sigma}\end{aligned}$$

Ora, dalla (2.8) si ha

$$\sum_{i=1}^{n}(Z_i - \overline{Z})^2 = \sum_{i=1}^{n} Z_i^2 - n\overline{Z}^2. \tag{2.15}$$

Per quanto visto in precedenza, le variabili aleatorie $\overline{Z}$ e $Z_i - \overline{Z}$ sono incorrelate, per ogni $i = 1, \ldots, n$ ed essendo distribuite normalmente sono anche indipendenti. Ne segue che $\overline{Z}$ e $\sum_{i=1}^{n}(Z_i - \overline{Z})^2$ sono anch'esse indipendenti, così come $n\overline{Z}^2$ e $\sum_{i=1}^{n}(Z_i - \overline{Z})^2$. Pertanto la (2.15) riscritta come

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n}(Z_i - \overline{Z})^2 + n\overline{Z}^2.$$

implica

$$E(e^{t\sum Z_i^2}) = E(e^{t\sum(Z_i-\overline{Z})^2})E(e^{tn\overline{Z}^2}),$$

da cui si ottiene

$$(1-2t)^{-n/2} = E(e^{t\sum(Z_i-\overline{Z})^2})(1-2t)^{-1/2}$$

essendo $\sum Z_i^2$ una variabile chi-quadrato con n gradi di libertà e $n\overline{Z}^2$ una chi-quadrato con 1 grado di libertà. Ne segue che l'ultima espressione diventa

$$E(e^{t\sum(Z_i-\overline{Z})^2}) = \frac{1}{(1-2t)^{(n-1)/2}}$$

che corrisponde alla funzione generatrice dei momenti di una chi-quadrato con $n-1$ gradi di libertà; allora, si deduce che $\sum(Z_i-\overline{Z})^2 \sim \chi^2_{n-1}$ e inoltre $\overline{Z} = (\overline{Y}-\mu)/\sigma$ è indipendente da $\sum(Z_i-\overline{Z})^2 = (n-1)S^2/\sigma$, cosicché anche $\overline{Y}$ e S^2 sono variabili aleatorie indipendenti.

Per la dimostrazione della (ii), è sufficiente considerare che

$$\sum(Z_i-\overline{Z})^2 = \sum\frac{(Y_i-\overline{Y})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

da cui si deduce che

$$S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}.$$

□

Si noti che assumendo l'indipendenza tra $\overline{Y}$ e S^2, il secondo punto poteva essere anche dimostrato come segue.

Usando la (2.8), si può scrivere

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2 - \frac{n}{\sigma^2}(\overline{Y}-\mu)^2$$

da cui

$$\sum_{i=1}^{n}\left(\frac{Y_i-\mu}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}\right)^2.$$

Posto

$$U = \sum_{i=1}^{n}\left(\frac{Y_i-\mu}{\sigma}\right)^2, \quad V = \frac{(n-1)S^2}{\sigma^2}, \quad W = \left(\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}\right)^2,$$

sussiste la seguente relazione tra le funzioni generatrici dei momenti di U, V e W:

$$m_U(t) = m_V(t)m_W(t) \tag{2.16}$$

Sapendo che $m_U(t) = (1-2t)^{-n/2}$, essendo U la somma di quadrati di n variabili aleatorie normali standardizzate, e $m_W(t) = (1-2t)^{-1/2}$, essendo W una normale standard al quadrato, si ricava

$$m_V(t) = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-(n-1)/2},$$

che è la funzione generatrice dei momenti di una variabile aleatoria chi-quadrato con $n-1$ gradi di libertà. Se ne deduce che

$$V = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}.$$

Nel paragrafo 4.7-RCCP si è visto che se Z ha distribuzione normale standardizzata, V ha una distribuzione chi-quadrato con d gradi di libertà, e se Z e V sono indipendenti, allora $Z/\sqrt{V/d}$ ha una distribuzione t *di Student* con d gradi di libertà, avente densità espressa dalla 4.34-RCCP. Il Teorema che segue mostra come si possa applicare tale risultato al campionamento da una popolazione normale.

Teorema 2.6. Se $(Y_1, Y_2, \ldots, Y_n)$ è un campione casuale di dimensione n estratto da una normale con media μ e varianza σ^2, allora il rapporto

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

dove $\overline{Y}$ e S^2 sono, rispettivamente, la media e la varianza del campione, ha una distribuzione t di Student con $n-1$ gradi di libertà.

Dimostrazione. Si noti innanzitutto che campionando da una popolazione normale $Z = (\overline{Y} - \mu)/(\sigma/\sqrt{n}) \sim \mathcal{N}(0,1)$, poiché $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$. Inoltre, in virtù del Teorema 2.5, $(n-1)S^2/\sigma^2$ ha una distribuzione chi-quadrato con $n-1$ gradi di libertà ed è indipendente da Z. Quindi

$$\frac{(\overline{Y} - \mu)}{\sigma/\sqrt{n}} \Big/ \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} = \frac{\sqrt{n}(\overline{Y} - \mu)}{S} = T \tag{2.17}$$

è il rapporto di una normale standard con la radice quadrata di una χ^2 indipendente, divisa per i suoi gradi di libertà. Ne segue che il rapporto ottenuto nella (2.17) ha distribuzione t di Student con $n-1$ gardi di libertà. □

Ciò che è importante sottolineare è che nella (2.17) il parametro incognito σ si semplifica, e pertanto la distribuzione di $\sqrt{n}(\overline{Y}-\mu)/S$ è indipendente dal valore di σ. Vedremo in seguito che la distribuzione t di Student si dimostrerà fondamentale per fare inferenza circa la media di una popolazione normale μ la cui varianza σ^2 è ignota.

Esempio 2.6. Il contenuto Y di alcool metilico di una bottiglia di vino è una variabile aleatoria $\mathcal{N}(\mu, \sigma^2)$. Considerando un campione di 9 bottiglie:

(a) calcolare $P(|\overline{Y} - \mu| > S)$ dove S^2 è la varianza campionaria.

(b) Se $\sigma^2 = 16$, calcolare la probabilità che S^2 assuma un valore compreso nell'intervallo $(5.46, 31)$.

(c) Se $\sigma^2 = 16$, determinare c tale che $P(S^2 \leq c) = 0.99$.

Per il punto (a) si può scrivere

$$\begin{aligned} P(|\overline{Y} - \mu| > S) &= 1 - P(-S < \overline{Y} - \mu < S) \\ &= 1 - P\left(-\sqrt{n} < \sqrt{n}\frac{(\overline{Y} - \mu)}{S} < \sqrt{n}\right) \\ &= 1 - P(-\sqrt{n} < T < \sqrt{n}) \end{aligned}$$

dove $T = \sqrt{n}(\overline{Y} - \mu)/S$ è distribuita come una t di Student con $n-1$ gradi di libertà. Pertanto, per $n = 9$, la probabilità cercata si ricava usando la tavola come

$$P(|\overline{Y}-\mu| > S) = 1-P(-3 < T < 3) = 2\times P(T > 3) = 2\times 0.00853 = 0.01706.$$

La probabilità al punto (b) è data da

$$\begin{aligned} P(5.46 \leq S^2 \leq 31) =& P\left(\frac{5.46 \times 8}{16} \leq \frac{8}{16}S^2 \leq \frac{31 \times 8}{16}\right) \\ =& P(2.73 \leq V \leq 15.5) \end{aligned}$$

dove $V = 8S^2/16$ ha distribuzione χ^2 con $n - 1 = 8$ gradi di libertà. Dalle tavole si trova

$$P(2.73 \leq V \leq 15.5) = P(V \leq 15.5) - P(V \leq 2.73) = 0.95 - 0.05 = 0.90.$$

Per risolvere il punto (c) si consideri ancora che $8S^2/16 \sim \chi^2_8$, pertanto si ha

$$0.99 = P(S^2 \leq c) = P(S^2/2 \leq c/2)$$

da cui si ricava

$$\frac{c}{2} = \chi^2_{8,0.99} = 20.09.$$

Quindi $c = 2 \times 20.1 = 40.18$. ▲

2.5 Campionamento da due popolazioni normali

I risultati presentati nel seguito riguardano il caso di due campioni estratti in modo indipendente; un esempio in cui si avrebbe tale situazione è quello discusso nel capitolo precedente riguardante la valutazione dell'efficacia di un farmaco.

Sia $(Y_{A1}, Y_{A2}, \ldots, Y_{An_A})$ un campione casuale di ampiezza n_A proveniente da una popolazione in cui la variabile Y_A ha media μ_A e varianza σ^2_A. Sia poi $(Y_{B1}, Y_{B2}, \ldots, Y_{Bn_B})$ un campione casuale di ampiezza n_B, indipendente dal primo, da una seconda popolazione in cui per Y_B si assume media μ_B e varianza σ^2_B. Siano

$$\overline{Y}_A = \frac{1}{n_A}\sum_{i=1}^{n_A} Y_{Ai}, \quad \overline{Y}_B = \frac{1}{n_B}\sum_{i=1}^{n_B} Y_{Bi},$$

le medie dei due campioni. Si osservi che $E(\overline{Y}_A) = \mu_A$ e $E(\overline{Y}) = \mu_B$, per cui si ha

$$E(\overline{Y}_A - \overline{Y}_B) = \mu_A - \mu_B. \tag{2.18}$$

Inoltre per la varianza si ha

$$V(\overline{Y}_A - \overline{Y}_B) = \frac{\sigma^2_A}{n_A} + \frac{\sigma^2_B}{n_B} \tag{2.19}$$

ricordando che $V(\overline{Y}_A - \overline{Y}_B) = V(\overline{Y}_A) + V(\overline{Y}_B) - 2\text{Cov}(\overline{Y}_A, \overline{Y}_B)$ e che $\text{Cov}(\overline{Y}_A, \overline{Y}_B) = 0$ per campioni indipendenti.
Si assuma ora che le popolazioni di riferimento abbiano legge normale $\mathcal{N}(\mu_A, \sigma_A^2)$ e $\mathcal{N}(\mu_B, \sigma_B^2)$, rispettivamente. In questa ipotesi, per quanto visto, si ha $\overline{Y}_A \sim \mathcal{N}(\mu_A, \sigma_A^2/n_A)$ e $\overline{Y}_B \sim \mathcal{N}(\mu_B, \sigma_B^2/n_B)$. Allora, qualunque combinazione lineare di $\overline{Y}_A$, $\overline{Y}_B$ sarà ancora una variabile aleatoria gaussiana, in particolare sarà ancora gaussiana la *differenza tra le medie campionarie* $(\overline{Y}_A - \overline{Y}_B)$. Per il Teorema 4.6-RCCP si ottiene

$$(\overline{Y}_A - \overline{Y}_B) \sim \mathcal{N}\left(\mu_A - \mu_B, \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\right), \tag{2.20}$$

mentre la corrispondente variabile aleatoria standardizzata sarà

$$Z = \frac{\overline{Y}_A - \overline{Y}_B - (\mu_A - \mu_B)}{\sqrt{(\sigma_A^2/n_A) + (\sigma_B^2/n_B)}} \sim \mathcal{N}(0,1). \tag{2.21}$$

Nel caso di due campioni casuali indipendenti, generati rispettivamente dalle popolazioni $\mathcal{N}(\mu_A, \sigma^2)$ e $\mathcal{N}(\mu_B, \sigma^2)$, per le quali la varianza coincide, la v.a. differenza tra le medie campionarie $\overline{Y}_A - \overline{Y}_B$ può essere espressa tramite una variabile casuale t di Student in base al seguente risultato.

Teorema 2.7. Siano $(Y_{A1}, Y_{A2}, \ldots, Y_{An_A})$ e $(Y_{B1}, Y_{B2}, \ldots, Y_{Bn_B})$ due campioni casuali indipendenti provenienti, rispettivamente, dalle popolazioni normali $\mathcal{N}(\mu_A, \sigma^2)$ e $\mathcal{N}(\mu_B, \sigma^2)$. Siano indicate con $S_A^2 = \sum_i^{n_A}(Y_{Ai} - \overline{Y}_A)^2/(n_A - 1)$ e $S_B^2 = \sum_i^{n_B}(Y_{Bi} - \overline{Y}_B)^2/(n_B - 1)$ le varianze dei due campioni. Allora il rapporto

$$T = \frac{(\overline{Y}_A - \overline{Y}_B) - (\mu_A - \mu_B)}{\sqrt{\dfrac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}\left(\dfrac{1}{n_A} + \dfrac{1}{n_B}\right)}} \tag{2.22}$$

ha distribuzione t di Student con $n_A + n_B - 2$ gradi di libertà.

Dimostrazione. Si osservi innanzitutto che

$$\frac{(\overline{Y}_A - \overline{Y}_B) - (\mu_A - \mu_B)}{\sigma\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}} \sim \mathcal{N}(0,1)$$

in virtù della (2.20), dove $\sigma_A^2 = \sigma_B^2 = \sigma^2$. Dalla (2.22), dividendo numeratore e denominatore per σ, si ha

$$T = \frac{[(\overline{Y}_A - \overline{Y}_B) - (\mu_A - \mu_B)]/\sigma}{\sqrt{\dfrac{(n_A-1)S_A^2/\sigma^2 + (n_B-1)S_B^2/\sigma^2}{n_A+n_B-2}\left(\dfrac{1}{n_A}+\dfrac{1}{n_B}\right)}}$$
$$= \frac{(\overline{Y}_A - \overline{Y}_B) - (\mu_A - \mu_B)}{\sigma\sqrt{1/n_A + 1/n_B}} \Big/ \sqrt{\frac{(n_A-1)S_A^2/\sigma^2 + (n_B-1)S_B^2/\sigma^2}{n_A+n_B-2}}.$$

Al numeratore si ottiene quindi una normale standardizzata per quanto osservato sopra; mentre al denominatore abbiamo la somma di $(n_A-1)S_A^2/\sigma^2$ e $(n_B-1)S_B^2/\sigma^2$, che sono variabili chi-quadrato indipendenti (si veda il Teorema 2.5) con $n_A - 1$ e $n_B - 1$ gradi di libertà, rispettivamente, cioè una chi-quadrato con $n_A + n_B - 2$ gradi di libertà, divisa per gli stessi gradi di libertà. È inoltre noto che le variabili aleatorie a rapporto sono indipendenti, essendo $\overline{Y}_A$ ed S_A^2 indipendenti, come pure $\overline{Y}_B$ ed S_B^2, in virtù della normalità delle popolazioni.

Se ne deduce che la (2.22) ha distribuzione t di Student con $n_A + n_B - 2$ gradi di libertà. □

Esempio 2.7. Si vuole valutare la differenza tra due linee di produzione che realizzano il medesimo prodotto. I tempi Y_1 e Y_2 rispettivamente impiegati dalle due linee di produzione si distribuiscono normalmente con medie $\mu_1 = 22$ e $\mu_2 = 20$ e varianze $\sigma_1^2 = 9$ e $\sigma_2^2 = 16$. Si consideri un campione di ampiezza $n_1 = 12$ da Y_1, e un secondo campione di ampiezza $n_2 = 14$ da Y_2. Si determini la probabilità $P(\overline{Y}_1 - \overline{Y}_2 \leq 3)$.

Sulla base della (2.21) otteniamo

$$P(\overline{Y}_1 - \overline{Y}_2 \leq 3) = P\left(Z \leq \frac{3-(22-20)}{\sqrt{(9/12)+(16/14)}}\right) = \Phi(0.727) = 0.766.$$

▲

È anche importante derivare la distribuzione campionaria del *rapporto tra le varianze campionarie*, sotto l'assunzione che i campioni estratti siano determinazioni di variabili gaussiane.

Nel paragrafo 4.7-RCCP, si è visto che la distribuzione del rapporto di due variabili casuali chi-quadrato indipendenti divise per i loro rispettivi gradi di libertà è la distribuzione F di Fisher. Cioè la quantità

$$Y = \frac{U/m}{V/n}$$

dove $U \sim \chi^2_m$ e $V \sim \chi^2_n$ e U e V sono indipendenti è la distribuita secondo una F di Fisher con m ed n gradi di libertà, che ha densità espressa dalla 4.35-RCCP, per $d_1 = m$ e $d_2 = n$. Nell'ambito del campionamento, vale il risultato che segue.

Teorema 2.8. Siano $(Y_{A1}, Y_{A2}, \ldots, Y_{An_A})$ e $(Y_{B1}, Y_{B2}, \ldots, Y_{Bn_B})$ due campioni casuali indipendenti provenienti, rispettivamente, dalle popolazioni normali $\mathcal{N}(\mu_A, \sigma^2_A)$ e $\mathcal{N}(\mu_B, \sigma^2_B)$. Siano indicate con S^2_A ed S^2_B le varianze dei due campioni definite come in (2.6). Il rapporto

$$\frac{S^2_A/\sigma^2_A}{S^2_B/\sigma^2_B} \tag{2.23}$$

ha distribuzione F di Fisher con $n_A - 1$ ed $n_B - 1$ gradi di libertà.

Dimostrazione. Il rapporto nell'enunciato del teorema può essere così riscritto

$$\frac{S^2_A/\sigma^2_A}{S^2_B/\sigma^2_B} = \frac{\frac{(n_A-1)S^2_A}{\sigma^2_A}/(n_A-1)}{\frac{(n_B-1)S^2_B}{\sigma^2_B}/(n_B-1)} \tag{2.24}$$

In base al Teorema 2.5, sappiamo che

$$\frac{(n_A-1)S^2_A}{\sigma^2_A} \sim \chi^2_{n_A-1}$$

e analogamente

$$\frac{(n_B-1)S^2_B}{\sigma^2_B} \sim \chi^2_{n_B-1}$$

da cui si nota che nella (2.24) al numeratore appare una variabile chi-quadrato con $n_A - 1$ gradi di libertà divisa per $n_A - 1$, e al denominatore appare una variabile chi-quadrato con $n_B - 1$ gradi di libertà divisa per $n_B - 1$. Essendo S^2_A ed S^2_B indipendenti, dall'applicazione del Teorema 4.14-RCCP discende immediatamente la tesi. □

La (2.23) è una statistica solo se si suppongono note le varianze σ_A^2 e σ_B^2. In particolare, se le varianze sono uguali $\sigma_A^2 = \sigma_B^2 = \sigma^2$, e sotto l'ipotesi di normalità, il rapporto tra le varianze campionarie

$$F = \frac{S_A^2/\sigma^2}{S_B^2/\sigma^2} = \frac{S_A^2}{S_B^2}$$

è distribuito come una variabile aleatoria F di Fisher con gradi di libertà $n_A - 1$ ed $n_B - 1$.

Esempio 2.8. Riprendendo i dati dell'esempio 2.7, si determini la costante a per cui vale $P(S_1^2 < aS_2^2) = 0.95$.

In base al Teorema(2.8), si ha

$$0.95 = P(S_1^2 < aS_2^2) = P\left(\frac{S_1^2/9}{S_2^2/16} < a\frac{16}{9}\right)$$

Poiché il valore della variabile F di Fisher con $n_1 - 1 = 11$ e $n_2 - 1 = 13$ gradi di libertà è 2.63, allora il valore di a ri ricava da $(16/9)a = 2.63$, da cui $a = 1.48$. ▲

2.6 Risultati asintotici

Come visto, la *legge debole dei grandi numeri* afferma che fissati due numeri $\epsilon > 0$ e δ, $0 < \delta < 1$, esiste un intero n tale che estraendo da $f(\cdot)$ un campione casuale di ampiezza maggiore o uguale a n, la probabilità che la media campionaria $\overline{Y}_n$ differisca da μ meno di ϵ è maggiore di $1 - \delta$ (cioè, tanto vicina a 1 quanto si vuole): per tutti gli interi $m \geq n$ vale

$$P(|\overline{Y}_m - \mu| < \epsilon) \geq 1 - \delta. \tag{2.25}$$

Il **teorema del limite centrale** (paragrafo 4.6-RCCP) afferma inoltre che la densità di $\overline{Y}$ tende a quella normale per $n \to \infty$, mentre per n finito e sufficientemente elevato, tale densità sarà ben approssimata da quella normale, indipendentemente dal modello descrittivo della popolazione.

Data l'importanza di tale risultato che fornisce la *distribuzione campionaria della media per grandi campioni*, esso verrà enunciato nuovamente.

Teorema 2.9 (Teorema del limite centrale)**.** Sia $f(\cdot)$ una densità con media μ e varianza finita σ^2, e sia $\overline{Y}_n$ la media campionaria di un campione casuale di ampiezza n estratto da $f(\cdot)$. Allora la sequenza di variabili aleatorie Z_n definita da

$$Z_n = \frac{\overline{Y}_n - E(\overline{Y}_n)}{\sqrt{V(\overline{Y}_n)}} = \frac{\sqrt{n}(\overline{Y}_n - \mu)}{\sigma} \tag{2.26}$$

converge in distribuzione alla normale standard.

Si deduce chiaramente l'importanza del Teorema 2.9 nelle applicazioni pratiche: la media $\overline{Y}_n$ di un campione casuale estratto da una qualsiasi distribuzione con varianza finita σ^2 e media μ è *approssimativamente* distribuita secondo la normale $\mathcal{N}(\mu, \sigma^2/n)$.

Si consideri ora il caso di due campioni casuali indipendenti, di ampiezza n_A ed n_B, tratti da due diverse popolazioni. Se la dimensione dei due campioni è sufficientemente elevata, il Teorema 2.9 consente di affermare che, qualunque sia il modello distributivo delle due popolazioni, la distribuzione della differenza tra le medie campionarie $(\overline{Y}_A - \overline{Y}_B)$ può essere approssimata dalla distribuzione normale $\mathcal{N}(\mu_A - \mu_B, \sigma_A^2/n_A + \sigma_B^2/n_B)$, poiché $\overline{Y}_A$ ha distribuzione limite $\mathcal{N}(\mu_A, \sigma_A^2/n_A)$ e, analogamente, $\overline{Y}_B$ ha distribuzione limite $\mathcal{N}(\mu_B, \sigma_B^2/n_B)$, assumendo che μ_A, μ_B, e σ_A^2, σ_B^2 siano, rispettivamente, le medie e le varianze delle due popolazioni da cui si generano i campioni. In particolare, se si considerano popolazioni Bernoulliane con medie p_A e p_B rispettivamente, allora $(\overline{Y}_A - \overline{Y}_B)$ ha distribuzione limite

$$\overline{Y}_A - \overline{Y}_B \sim \mathcal{N}(p_A - p_B, p_A(1-p_A)/n_A + p_B(1-p_B)/n_B).$$

Esempio 2.9. Si supponga che una distribuzione con media incognita abbia varianza $\sigma^2 = 1$. Quanto deve essere grande un campione perché si abbia una probabilità di almeno il 95% che la media campionaria $\overline{Y}_n$ disti dalla media della popolazione meno di 0.5?

Usando la (2.25), per la disuguaglianza di Chebyshev, si ha

$$P(|\overline{Y}_n - \mu| < \epsilon) \geq 1 - \frac{(1/n)\sigma^2}{\epsilon^2} \geq 1 - \delta.$$

Posto $\sigma^2 = 1$, $\epsilon = 0.5$ e $\delta = 0.05$ si ottiene

$$n > \frac{\sigma^2}{\delta\epsilon^2} = \frac{1}{0.05(0.5)^2} = 80.$$

▲

2.7 Statistiche d'ordine

Si introduce ora il concetto di statistica d'ordine e si forniscono alcune delle sue proprietà.

Definizione 2.4 (Statistiche d'ordine)**.** Dato un campione casuale di ampiezza n, $(Y_1, Y_2, \ldots, Y_n)$, estratto da una funzione di ripartizione $F(\cdot)$, si definiscono **statistiche d'ordine** corrispondenti al campione dato gli Y_i ordinati in senso non decrescente

$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$$

dove $Y_{(i)}$ è il dato (variabile aleatoria) che occupa la posizione i-ma della graduatoria.

Segue immediatamente che

$$Y_{(1)} = \min\{Y_1, Y_2, \ldots, Y_n\}, \quad Y_{(n)} = \max\{Y_1, Y_2, \ldots, Y_n\}.$$

come già visto nel paragrafo 4.4-RCCP. Gli $Y_{(i)}$ $(i = 1, \ldots, n)$ sono statistiche essendo funzioni del campione e, a differenza del campione stesso, non sono indipendenti. Infatti, se $Y_{(j)} \geq y$, allora si ha necessariamente $Y_{(j+1)} \geq y$.

Teorema 2.10. Si indichino con $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ le statistiche d'ordine estratte da una funzione di ripartizione $F(\cdot)$. La funzione di ripartizione di $Y_{(\alpha)}$ $(\alpha = 1, \ldots, n)$ è data da

$$F_{Y_{(\alpha)}}(y) = \sum_{j=\alpha}^{n} \binom{n}{j} [F(y)]^j [1 - F(y)]^{n-j} \tag{2.27}$$

Dimostrazione. Per y fissato, sia Z_i la variabile aleatoria che assume valore 1 se $Y_i \in (-\infty, y]$. Allora $Z = \sum_{i=1}^{n} Z_i$ è pari al numero di $Y_i \leq y$. Si osservi che Z ha distribuzione binomiale con parametri n e $F(y)$. Data l'equivalenza dei due eventi $\{Z \geq \alpha\}$ e $\{Y_{(\alpha)} \leq y\}$ si ha

$$F_{Y_{(\alpha)}}(y) = P(Y_{(\alpha)} \leq y) = P(Z \geq \alpha) = \sum_{j=\alpha}^{n} \binom{n}{j} [F(y)]^j [1 - F(y)]^{n-j}$$

□

Si può dimostrare che la densità congiunta delle statistiche d'ordine $Y_{(1)}$, $Y_{(2)}, \ldots, Y_{(n)}$ è data da

$$f_{Y_{(1)},\ldots,Y_{(n)}}(y_1, \ldots, y_n) = n! f(y_1) \ldots f(y_n), \quad \text{per } y_1 < y_2 < \cdots < y_n \tag{2.28}$$

e $f_{Y_{(1)},\ldots,Y_{(n)}}(y_1, \ldots, y_n) = 0$, altrimenti.

Si può ottenere un insieme di densità marginali dalla densità congiunta integrando rispetto alle variabili che non interessano. Ad esempio, per $Y_{(\alpha)}$ si ottiene

$$f_{Y_{(\alpha)}}(y) = \frac{n!}{(\alpha - 1)!(n - \alpha)!} [F(y)]^{\alpha - 1} [1 - F(y)]^{n - \alpha} f(y).$$

Sulla base delle statistiche d'ordine è possibile definire alcune funzioni delle stesse, come ad esempio la *mediana campionaria* Me_n che corrisponde alla statistica d'ordine centrale se n è dispari e alla media delle due statistiche d'ordine centrali se n è pari:

$$Me_n = \begin{cases} Y_{((n+1)/2)} & \text{se n è dispari} \\ \frac{1}{2}(Y_{(n/2)} + Y_{(n/2+1)}) & \text{se n è pari.} \end{cases}$$

Il *campo di variazione*, indicato con W, è espresso dalla differenza tra il massimo e il minimo campionario:

$$W = Y_{(n)} - Y_{(1)}.$$

Nei paragrafi precedenti si sono esaminate alcune distribuzioni asintotiche, come nel caso della media campionaria $\overline{Y}$. La domanda che ci poniamo

ora è: esiste una distribuzione asintotica per la mediana campionaria?

Al fine di ricavare risultati asintotici, si indicano con $Y^n_{(1)}$, $Y^n_{(2)}, \ldots, Y^n_{(n)}$ le statistiche d'ordine di un campione di ampiezza n. Si fornirà la distribuzione asintotica dell'(np)-esima statistica d'ordine di un campione di ampiezza n, per ogni $0 < p < 1$. Si osservi che np può non essere un intero, e pertanto definiamo p_n tale che np_n sia un intero, dove p_n è approssimativamente pari a p. In particolare, sia $p = 1/2$, e $\xi_{1/2}$ il quantile di ordine $1/2$, cioè la mediana della popolazione. Allora, si può dimostrare che la mediana campionaria Y_m è distribuita asintoticamente come una normale avente per media la mediana della popolazione e per varianza $1/(4n[f(\xi_{1/2})]^2)$ dove $f(\cdot)$ è la densità della popolazione da cui estraiamo il campione.

Se $f(\cdot)$ è una densità normale di media μ e varianza σ^2, allora la mediana campionaria ha distribuzione asintotica

$$Me_n \dot{\sim} \mathcal{N}\left(\mu, \frac{\pi\sigma^2}{2n}\right)$$

essendo $1/(4n[f(\mu)]^2) = 1/4n(2\pi\sigma^2)^{-1}$.

Un risultato più generale è il seguente.

Teorema 2.11. Sia dato un campione casuale $(Y_1, Y_2, \ldots, Y_n)$ da una popolazione con densità $f(\cdot)$ e funzione di ripartizione $F(\cdot)$, e sia $F(y)$ strettamente monotona per $0 < F(y) < 1$. Sia ξ_p l'unica soluzione di $F(y) = p$ rispetto a y per $0 < p < 1$. Sia p_n tale che np_n sia un intero e $n|p_n - p|$ limitato. Indicata con $Y^n_{(np_n)}$ l'(np_n)-esima statistica d'ordine del campione di ampiezza n, allora $Y^n_{(np_n)}$ ha distribuzione asintotica

$$Y^n_{(np_n)} \dot{\sim} \mathcal{N}\left(\xi_p, \frac{p(1-p)}{n[f(\xi_p)]^2}\right).$$

2.8 La funzione di ripartizione empirica

In questo paragrafo definiamo la *funzione di ripartizione empirica o campionaria*, che è una funzione delle statistiche d'ordine.

Definizione 2.5 (Funzione di ripartizione empirica)**.** Dato un campione casuale $(Y_1, Y_2, \ldots, Y_n)$ da una funzione di ripartizione F, la **funzione di ripartizione (f.r.) empirica** (o campionaria), indicata con $F_n(y)$, è definita da

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_{(-\infty,y]}(Y_i)$$

dove $\mathbf{I}_{(-\infty,y]}(Y_i)$ è la funzione indicatrice, uguale a 1 se $Y_i \leq y$ e uguale a 0 altrimenti.

Si noti che, indicate con $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ le statistiche d'ordine corrispondenti, la f.r. empirica può essere espressa come

$$F_n(y) = (1/n) \times \{\text{numero di } Y_{(j)} \text{ minori o uguali a } y\}.$$

Per un y fissato, $F_n(y)$ è una statistica, essendo funzione del campione. Sia $Z_i = \mathbf{I}_{(-\infty,y]}(Y_i)$; allora Z_i ha una distribuzione di Bernoulli con parametro $F(y)$. Quindi $\sum_{i=1}^{n} Z_i$ avrà una distribuzione binomiale con parametri n e $F(y)$. Ne segue che

$$P\left(F_n(y) = \frac{k}{n}\right) = \binom{n}{k} [F(y)]^k [1 - F(y)]^{n-k}, \quad k = 0, 1, \ldots, n.$$

da cui ricaviamo la media e la varianza della funzione di ripartizione campionaria:

$$E(F_n(y)) = \sum_{k=0}^{n} \frac{k}{n} \binom{n}{k} [F(y)]^k [1 - F(y)]^{n-k} = F(y),$$

$$V(F_n(y)) = \frac{1}{n} [F(y)][1 - F(y)].$$

Inoltre, poiché $F_n(y)$ è la media campionaria delle variabili Z_i, dal teorema del limite centrale sappiamo che $F_n(y)$ è asintoticamente distribuita come una normale con media $F(y)$ e varianza $(1/n)[F(y)][1 - F(y)]$.

Capitolo 3

Verosimiglianza

3.1 Ancora sulla composizione di un'urna

Nel paragrafo 1.4.1 si è introdotto un semplice problema: da un'urna composta di palline bianche e nere in cui l'ignota proporzione di palline bianche è pari a p si estraggono con reinserimento n palline e si osserva se ogni pallina estratta è bianca o nera. Il modello statistico è un modello parametrico bernoulliano e quindi il problema inferenziale riguarda il parametro p che è la media di una variabile aleatoria Bernoulliana Y per la quale $P(Y = 1) = p$ e $Y = 1$ se la pallina è bianca.

Per semplificare il problema si assuma che sia noto che le urne da cui avviene l'estrazione possono essere solo di 4 tipi, ovvero i possibili valori di p possono essere solo i seguenti: 0.1, 0.2, 0.3 e 0.5. Lo spazio parametrico P è quindi composto da quattro valori $p_j \in \{0.1, 0.2, 0.3, 0.5\}$.

Una volta osservato il campione, un obiettivo interessante è quello di riassumere l'informazione fornita dai dati osservati per ciascun valore del parametro. Per chiarire meglio la questione si consideri il seguente esempio. Sia $n = 20$ e sia risultata la seguente sequenza di palline

$$\{b, n, b, , n, n, n, b, n, b, b, n, n, n, b, n, n, b, n, b, n\}.$$

In essa contiamo 8 palline bianche (chiamiamo $s = \sum_{i=1}^{n} y_i$ il risultato di tale conteggio). Alla luce del risultato campionario si può procedere al calcolo della probabilità di osservare tale sequenza per le diverse composizioni di urna (in questo caso sono solo 4).

Si tratta pertanto di calcolare nel caso di campionamento casuale, ovvero assumendo indipendenza e identica distribuzione, le seguenti probabilità:

$$\begin{aligned} P(y_1, y_2, \ldots, y_n; p_j) &= \prod_{i=1}^{20} p_j^{y_i}(1-p_j)^{1-y_i} \\ &= p_j^{\sum_{i=1}^{20} y_i}(1-p_j)^{20-\sum_{i=1}^{20} y_i} = p_j^s(1-p_j)^{20-s}, \end{aligned}$$

per $j = 1, 2, 3, 4$. Si possono riassumere tali valori che chiameremo $L(p_j)$ nella seguente tabella (si tratta di valori molto piccoli e pertanto nella tabella sono stati moltiplicati per 10^9)

p_j	0.1	0.2	0.3	0.5
$L(p_j) \times 10^9$	2.82	175.92	908.13	953.67

Abbiamo quindi definito una funzione $L(p_j)$ che per ogni punto dello spazio parametrico (che in questo caso comprende i soli 4 valori di p_j) fornisce la probabilità che il modello statistico abbia generato lo specifico campione osservato: essa è un esempio di **funzione di verosimiglianza**.

Prima di generalizzare il ragionamento, si osservi quanto segue.

a. Per quanto i singoli valori della funzione $L(p_j)$ siano delle probabilità, essa non è una funzione di probabilità (i valori p_j non sono valori aleatori).

b. Se, nel contesto descritto sopra, si calcolasse la probabilità di una qualsiasi altra sequenza che contiene 8 palline bianche anche secondo un altro ordine, si otterrebbero gli stessi valori della verosimiglianza. Se invece si calcola la probabilità della sequenza osservata $P(S = 8; p = p_j)$ ove $S = \sum_{i=1}^{n} Y_i \sim Bin(20, p_j)$, questo ci porta a calcolare le stesse probabilità viste sopra moltiplicate per il fattore $\binom{20}{8}$ che non dipende da p_j.

c. La funzione $L(p_j)$ permette di confrontare i valori del parametro e ci indica, in particolare, quali valori dello spazio parametrico sono più plausibili alla luce dei dati osservati. Il confronto più appropriato avviene considerando il rapporto fra i valori di $L(p_j)$. Nell'esempio considerato p_3 è circa 5.16 volte più verosimile di p_2, infatti $\frac{L(p_3)}{L(p_2)} \approx 5.16$. Il rapporto fra $L(p_4)$ e $L(p_3)$ è vicino a 1 (≈ 1.05) ma indica che è più verosimile che p_4 abbia generato i dati rispetto a p_3.

d. Se si calcola la verosimiglianza con riferimento al risultato $s = 8$ invece che con riferimento alla sequenza osservata i rapporti fra le verosimiglianze restano invariati. Questo implica che conoscere la sequenza osservata o solo la la statistica campionaria S è equivalente ai fini della valutazione del rapporto fra le verosimiglianze di due valori del parametro. Quindi la funzione è definita in modo equivalente a meno di una costante di proporzionalità.

e. Se si dovesse scegliere un solo valore, quale possibile valore di p avrebbe senso scegliere p_4, poiché è quello al quale corrisponde il valore di $L(p_j)$ più elevato.

A questo punto, si può considerare una versione più generale dell'esempio introdotto: l'urna da cui avviene l'estrazione delle 20 palline ha una composizione ignota e la proporzione di palline bianche p può assumere qualsiasi valore reale compreso fra 0 e 1. Questo equivale a ipotizzare che i 20 valori osservati siano determinazione di una sequenza di 20 variabili aleatorie $Y_i \sim Be(p)$. Lo stesso ragionamento visto sopra può essere riproposto e la probabilità della generica sequenza dei valori campionari osservati $(y_1, y_2, \ldots, y_{20})$ in funzione della composizione dell'urna risulta essere la seguente funzione di p:

$$L(y_1, y_2, \ldots, y_n; p) = p^s(1-p)^{20-s} = p^8(1-p)^{12}.$$

Al fine di sottolineare che si tratta di una funzione di p in generale scriveremo $L(y_1, y_2, \ldots, y_n; p) = L(p) \quad \text{con} \quad p \in (0, 1)$. In effetti, essa è vista come funzione solo di p essendo fissata la sequenza di palline bianche e nere osservata. Avremo quindi una funzione definita sullo spazio parametrico del modello che fornisce per ogni valore di p la verosimiglianza che tale valore abbia generato la sequenza osservata.

Può essere interessante notare nella figura 3.1 che la funzione di verosimiglianza ha il suo massimo in corrispondenza del valore $\sum_{i=1}^{n} y_i/n$ che nel caso del modello Bernoulliano rappresenta la proporzione di successi. Si noti inoltre che si può esprimere la funzione di verosimiglianza anche solo conoscendo $s = \sum_{i=1}^{20} y_i$ invece che l'intera sequenza dei valori y_i.

In generale, quindi, la funzione di verosimiglianza per il parametro di un modello Bernoulliano per una variabile Y, avendo osservato un campione i.i.d. di dimensione n, è data da:

$$L(y_1, y_2, \ldots, y_n; p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{\sum_{i=1}^{n} y_i}(1-p)^{n-\sum_{i=1}^{n} y_i}.$$

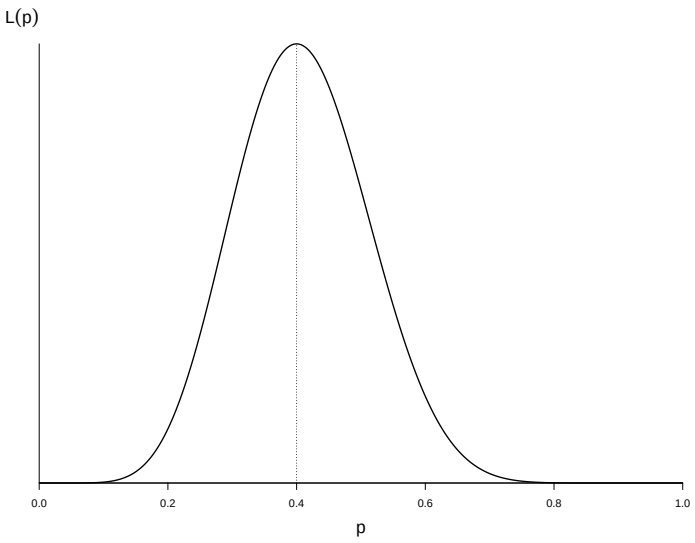


Figura 3.1: Funzione di verosimiglianza per il parametro p di $Y \sim Be(p)$, nel caso delle 4 urne con 8 successi su 20.

3.2 Funzione di verosimiglianza

I concetti illustrati nel paragrafo precedente riguardo l'individuazione di una funzione che associ ad ogni valore del parametro la probabilità che esso abbia dato luogo al campione osservato, possono essere proposti più in generale per un dato modello parametrico per Y e conducono alla seguente definizione.

Definizione 3.1 (Funzione di verosimiglianza)**.** Osservato un campione casuale $(y_1, y_2, \ldots, y_n)$ di dimensione n da Y che è distribuita secondo il modello $f(y;\theta)$, ove $\theta \in \Theta \subset \mathbb{R}$, si definisce **funzione di verosimiglianza** per θ la funzione

$$L(\theta) = f(y_1, y_2, \ldots, y_n; \theta).$$

Essa è una funzione definita su Θ che assume valori in $\mathbb{R}^+$. Nel caso di campionamento casuale, per la condizione di indipendenza delle variabili aleatorie da cui sono tratte le realizzazioni campionarie (nonché per l'identica distribuzione delle stesse), si ha:

$$L(\theta) = f(y_1, y_2, \ldots, y_n; \theta) = \prod_{1=1}^{n} f(y_i; \theta).$$

Si possono fare le seguenti considerazioni:

1. La funzione di verosimiglianza è funzione del parametro e considera i dati come costanti. Per tale motivo, come già sottolineato, nel seguito si conviene di definire tale funzione semplicemente come $L(\theta)$ omettendo, quando non sia necessario, il riferimento alla n-pla campionaria osservata.
2. Se ai fini del confronto di due valori della verosimiglianza si utilizza il rapporto, allora la verosimiglianza può essere definita a meno di una costante per cui $L(\theta) = c \cdot f(y_1, y_2, \ldots, y_n; \theta) \propto f(y_1, y_2, \ldots, y_n; \theta)$. Si noti che quello che conta è che c non dipenda da θ mentre può dipendere dai dati.
3. Tale funzione rappresenta nel caso in cui Y sia un modello per dati discreti la probabilità di osservare i dati del campione in corrispondenza di ciascun valore θ: cioè la "verosimiglianza" del campione osservato per quel valore di θ. Se per Y si utilizza un modello continuo, la probabilità di osservare y_i può essere scritta come:

   $$P(y_i - \epsilon \leq y_i \leq y_i + \epsilon) = \int_{y_i - \epsilon}^{y_i + \epsilon} f(y;\theta)dy \approx 2\epsilon \cdot f(y_i;\theta)$$

   e l'ultima approssimazione è tanto più ragionevole quanto più piccolo è ϵ. Quindi la probabilità di osservare un dato campione è approssi-

mativamente pari a $L(\theta) = 2\epsilon \prod_{1=1}^{n} f(y_i;\theta)$, dove $f(y_i;\theta)$ rappresenta la funzione di densità in y_i. Per quanto detto al punto precedente la costante può essere trascurata e pertanto la verosimiglianza può essere interpretata come probabilità di osservare un dato insieme di dati per ciascun valore del parametro anche nel caso continuo.

4. La funzione di verosimiglianza è definita per un campione $(y_1, \ldots, y_n)$ ottenuto da un opportuno modello statistico anche in assenza di indipendenza o di identica distribuzione. In effetti, è sufficiente che il modello consenta di definire la funzione di probabilità (o densità) congiunta. Una volta osservato il campione si ha che

$$L(\theta) = f(y_1, y_2, \ldots, y_n; \theta)$$

definisce la probabilità (o densità) congiunta di osservare quel campione al variare di $\theta \in \Theta$. La condizione di indipendenza permette di esprimere tale funzione come prodotto delle funzioni di densità marginali e quella di identica distribuzione di usare la stessa $f(\cdot)$ per ogni osservazione.

5. Si assume che il modello statistico sia tale che a due valori diversi di θ corrispondano distribuzioni di probabilità distinte. Tale proprietà è detta di *identificabilità*, ed è importante per giustificare la funzione di verosimiglianza come criterio per l'inferenza sul parametro di un modello statistico.

3.2.1 Funzione di log-verosimiglianza

La funzione di verosimiglianza permette quindi di esprimere, per ciascun possibile valore del parametro, una quantità che riflette quanto sia plausibile che i dati osservati siano generati da uno specifico valore di θ: possiamo ordinare i valori del parametro in base alla loro verosimiglianza per cui se $L(\theta_1) > L(\theta_2)$ allora la distribuzione con parametro θ_1 fornisce una "spiegazione" più plausibile dei dati osservati. Se consideriamo una funzione monotona di $L(\theta)$ la graduatoria fra i valori di θ resta immutata.

Nell'esempio introduttivo (3.1) l'urna 4 era la più plausibile e l'urna 3 era più plausibile della 1. Se invece di considerare $L(\theta)$ si considera il logaritmo di tale quantità ovvero $l(\theta) = \log L(\theta)$ la graduatoria resta immutata. Sempre con riferimento all'esempio dell'urna, si noti che:

1. nel caso delle 4 urne con 8 successi su 20 l'espressione del logaritmo naturale della funzione di verosimiglianza risulta agevole da esprimere

$$\begin{aligned}\log L(y_1, y_2, \ldots, y_{20}; p) &= \log\left(p^{\sum_{i=1}^{20} y_i}(1-p)^{20-\sum_{i=1}^{20} y_i}\right) \\ &= \sum_{i=1}^{20} y_i \log p + \left(20 - \sum_i y_i\right)\log(1-p) \\ &= 8\log p + 12\log(1-p)\end{aligned}$$

   (in molti casi è in effetti preferibile calcolare n termini di una sommatoria piuttosto che di una produttoria);

2. in generale, per qualsiasi valore di $p \in (0,1)$, il logaritmo si può calcolare in quanto assume valori strettamente maggiori di 0.

Studiare il logaritmo della funzione di verosimiglianza o la funzione di verosimiglianza stessa porta quindi a conclusioni equivalenti.

Definizione 3.2 (Funzione di log-verosimiglianza)**.** Si definisce la **funzione di log-verosimiglianza** come segue:

$$\ell(\theta) = \log L(\theta) = \log f(y_1, y_2, \ldots, y_n; \theta),$$

che per un campione casuale diventa

$$\ell(\theta) = \log\left(\prod_{1=1}^{n} f(y_i;\theta)\right) = \sum_{1=1}^{n} \log f(y_i;\theta)$$

Si è osservato che, in genere, si guarda ai rapporti fra valori della funzione di verosimiglianza per valutare quale fra due punti di Θ sia più verosimile per definire la popolazione Y da cui sono tratti i dati osservati. Se si considera il logaritmo del rapporto fra verosimiglianze esso equivale alle differenza fra la log-verosimiglianza in due diversi punti di Θ.

3.2.2 Invarianza della funzione di verosimiglianza

In molti casi il modello distributivo per Y può essere espresso utilizzando parametrizzazioni alternative. Ad esempio, nel caso di un modello esponenziale la cui legge di distribuzione è espressa da $f(y;\theta) = \theta e^{-\theta y}$, $\theta > 0$, si

potrebbe utilizzare un parametro $\phi = g(\theta)$ che corrisponde a una trasformazione biunivoca di θ. Ad esempio, si potrebbe porre $\phi = 1/\theta$ così che il nuovo parametro rappresenti la media della distribuzione così da agevolarne l'interpretazione.

La funzione di verosimiglianza risulta invariante rispetto a trasformazioni biunivoche del parametro; per le verosimiglianze espresse in funzione di θ e di ϕ si ha

$$L_\theta(\theta) = L_\theta(g^{-1}(\phi)) = L_\phi(\phi).$$

3.2.3 Alcuni esempi

Modello Poisson

Sia $(y_1, y_2, \ldots, y_n)$ un campione di dimensione n da Y che è distribuita secondo il modello Poisson $p(y;\lambda) = e^{-\lambda}\lambda^y/y!$, $\lambda > 0$, $R_y = \{0, 1, 2, \ldots\}$. La funzione di verosimiglianza è

$$L(y_1, y_2, \ldots, y_n; \lambda) = L(\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{y_i}}{y_i!} = \frac{e^{-\lambda n}\lambda^{\sum_{i=1}^n y_i}}{\prod y_i!}$$

Si ottiene quindi la seguente funzione di log-verosimiglianza:

$$\begin{aligned}\ell(\lambda) = \log L(\lambda) &= \log\left(e^{-\lambda n}\lambda^{\sum_{i=1}^n y_i}/\prod y_i!\right) \\ &= \log(e^{-\lambda n}) + \log\left(\lambda^{\sum_{i=1}^n y_i}\right) - \log\left(\prod y_i!\right) \\ &= -\lambda n + \log\lambda\sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\log y_i!\end{aligned} \tag{3.1}$$

Modello gaussiano

Sia dato un campione casuale $(y_1, y_2, \ldots, y_n)$ proveniente da una popolazione $\mathcal{N}(\mu, \sigma^2)$, con μ e σ^2 ignoti. Allora la funzione di verosimiglianza per il campione osservato $(y_1, y_2, \ldots, y_n)$ è

$$\begin{aligned}L(y_1, y_2, \ldots, y_n; \mu, \sigma^2) = L(\mu, \sigma^2) &= \prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-(y_i-\mu)^2/2\sigma^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\sum_{i=1}^n (y_i-\mu)^2/2\sigma^2}\end{aligned}$$

Si noti che si tratta in questo caso di valutare la verosimiglianza come funzione dei due parametri μ e σ^2. Applicando il logaritmo naturale, si ottiene la seguente funzione di log-verosimiglianza:

$$\begin{aligned}\ell(\mu, \sigma^2) = \log L(\mu, \sigma^2) &= \log\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} - \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2} \\ &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2}.\end{aligned} \tag{3.2}$$

Modello esponenziale con diversi schemi di osservazione

Sia $(y_1, y_2, \ldots, y_n)$ un campione di dimensione n da Y che è distribuita secondo il modello esponenziale di parametro λ, $f(y;\lambda) = \lambda e^{-\lambda y}$, per $\lambda > 0$ e $y \geq 0$. La funzione di verosimiglianza in corrispondenza del campione osservato è:

$$L(y_1, y_2, \ldots, y_n; \lambda) = L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} y_i}$$

Si ottiene quindi la seguente funzione di log-verosimiglianza:

$$\ell(\lambda) = \log L(\lambda) = \log\left(\lambda^n e^{-\lambda \sum_{i=1}^{n} y_i}\right) = n\log\lambda - \lambda\sum_{i=1}^{n} y_i. \tag{3.3}$$

Si immagini ora che per lo stesso campione $(y_1, y_2, \ldots, y_n)$ di dimensione n da una distribuzione univariata, discreta o continua, $f(y;\theta)$ con $\theta \in \Theta$ e $\Theta \subset \mathbb{R}$ non sia possibile osservare il valore ma si sa solo, avendo fissato una soglia t, se $y_i > t$. L'informazione è quindi parziale e si dice che dati sono soggetti a censura dicotomica. È ancora possibile fare inferenza sul parametro λ anche con questa informazione incompleta. Ovviamente occorre costruire la funzione di verosimiglianza che rifletta il tipo di informazione disponibile. In questo caso in effetti si potrebbe associare a ogni dato y_i il valore di una variabile w_i che sia funzione indicatrice dell'evento $y_i > t$ per cui

$$\begin{aligned} w_i &= 1 \quad \text{se} \quad y_i > t \\ w_1 &= 0 \quad \text{altrimenti} \end{aligned}$$

I valori di w_i sono quindi determinazioni di una variabile bernoulliana di parametro $p = P(Y_i > t)$. Ricordando l'espressione ottenuta per il parametro

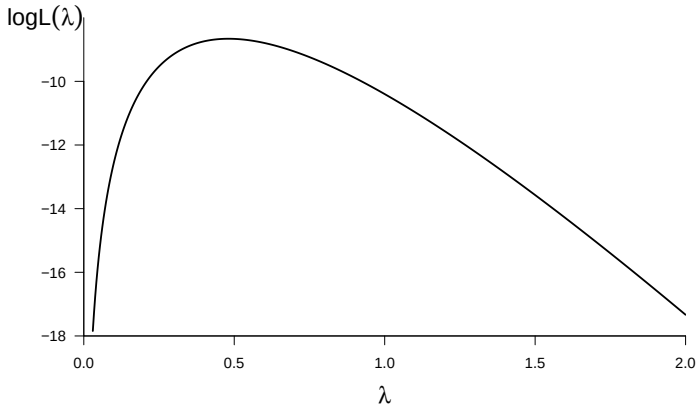


Figura 3.2: Funzione di log-verosimiglianza per il parametro λ di una v.a. $Y \sim Esp(\lambda)$.

di una bernoulliana, La funzione di verosimiglianza per p è espressa da

$$L(w_1, w_2, \ldots, w_n; p) = L(p) = p^{n_1}(1-p)^{n-n_1}$$

ove $n_1 = \sum_{i=1}^{n} w_i$ è il numero di valori y_i che sono risultati maggiori di t, da cui si ottiene la log-verosimiglianza per il modello bernoulliano data da:

$$\ell(p) = \log L(p) = n_1 \log(p) + (n - n_1) \log(1-p)$$

È interessante ora considerare il caso in cui Y ha distribuzione esponenziale di parametro λ ed è una variabile che nella popolazione misura i tempi fino al verificarsi di un dato evento. Allora $p = e^{-t\lambda}$, e

$$\begin{aligned} \ell(\lambda) = \log L(w_1, w_2, \ldots, w_n; \lambda) &= \log\left(e^{-n_1 t\lambda}(1 - e^{-t\lambda})^{n-n_1}\right) \\ &= -\lambda n_1 t + (n - n_1)\log(1 - e^{-t\lambda}). \end{aligned} \tag{3.4}$$

Si noti che questo equivale a una riparametrizzazione di un modello bernoulliano con parametro $p = e^{-t\lambda}$.

Un tipico esempio si ha considerando il tempo trascorso dall'inizio di una infezione fino alla completa guarigione per i pazienti con una data patologia curata con un certo farmaco, per il quale si può supporre una distribuzione esponenziale di parametro λ. Si supponga che il campione di osservazioni

$(y_1, y_2, \ldots, y_n)$ comprenda alcune y_i^* per le quali il dato è censurato ed è noto solo che il tempo di guarigione è superiore a y_i^*. Il campione quindi comprende alcuni dati completi y_i che contribuiranno alla funzione di verosimiglianza con $f(y_i, \lambda) = \lambda \exp(-\lambda y_i)$; mentre per i dati censurati y_i^* l'informazione è incompleta e si sa solo che il tempo sarà maggiore del tempo di censura: il contributo alla verosimiglianza per essi è $P(Y > y_i^*) = 1 - F(y_i^*) = \exp(-\lambda y_i^*)$. Se i dati censurati sono $n_1 < n$ allora la funzione di verosimiglianza risulta pari a:

$$L(\lambda) = \prod_{i=1}^{n-n1} \lambda e^{-\lambda y_i} \prod_{i=1}^{n1} (e^{-\lambda y_i^*})$$

Modello rettangolare

Sia $(y_1, y_2, \ldots, y_n)$ un campione casuale estratto da una distribuzione uniforme $U(0, \theta)$, $\theta > 0$. La funzione di verosimiglianza per θ è data da

$$L(y_1, y_2, \ldots, y_n; \theta) = L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} = \frac{1}{\theta^n}, \quad 0 \leq y_i \leq \theta, \quad i = 1, \ldots, n$$

La condizione che ciascun y_i sia compreso tra 0 e θ equivale a richiedere che il massimo di $(y_1, y_2, \ldots, y_n)$, $y_{(n)}$, sia minore o uguale a θ. Pertanto la verosimiglianza è pari a

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} = \frac{1}{\theta^n}, \qquad \text{se } y_{(n)} \leq \theta,$$

ed è pari a zero altrimenti. Si noti che l'espressione precedente può essere riscritta come

$$L(\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbf{I}_{[0,\theta]}(y_i) = \frac{1}{\theta^n} \mathbf{I}_{[y_{(n)},\infty)}(\theta).$$

La funzione presenta quindi il suo valore più elevato in corrispondenza del valore $y_{(n)}$ (si veda la figura 3.3).

3.3 Alcune proprietà della funzione di verosimiglianza

Come si è detto la funzione di verosimiglianza fornisce una sintesi del supporto del campione a ciascun valore del parametro θ in quanto esprime

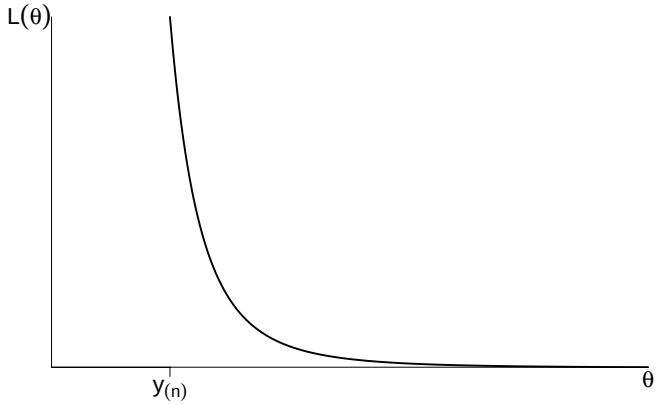


Figura 3.3: Funzione di verosimiglianza per una variabile casuale uniforme nell'intervallo $[0, \theta]$.

la probabilità di avere osservato quello specifico insieme di dati. Tuttavia se si considera uno specifico valore di θ, il comportamento della funzione $L(\theta) = f(y_1, y_2, \ldots, y_n; \theta)$, o della sua trasformata $\ell(\theta) = \log L(\theta)$, può essere studiato al variare della n-pla campionaria. $L(\theta)$ ed $\ell(\theta)$ sono variabili aleatorie campionarie e possiamo quindi valutarne la distribuzione e le proprietà.

Per evidenziare meglio che si intende qui valutare il comportamento di variabili aleatorie campionarie, sarà, in questo caso, adottata la notazione $f(y_1, y_2, \ldots, y_n; \theta)$ per indicare la funzione di verosimiglianza; inoltre qui e in futuro per casi analoghi, denoteremo con $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ l'intero vettore dei dati come realizzazione del vettore aleatorio $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$.

Ad esempio, per calcolare una quantità come $E(S)$ ove $S = g(y_1, y_2, \ldots, y_n)$ invece di scrivere

$$E(S) = \int_{R_{y_1}} \int_{R_{y_2}} \cdots \int_{R_{y_n}} g(y_1, y_2, \ldots, y_n) f(y_1, y_2, \ldots, y_n; \theta) dy_1, dy_2 \ldots dy_n$$

si scriverà semplicemente

$$E(S) = \int_{\mathbb{R}_{\boldsymbol{Y}}} g(\boldsymbol{y}) f(\boldsymbol{y}) d\boldsymbol{y},$$

ove l'integrale si intende esteso sull'intero spazio campionario $R_{\boldsymbol{Y}}$ se non diversamente specificato.

Inoltre nel confrontare il valore della funzione di verosimiglianza $L(\theta) = f(\boldsymbol{y};\theta)$ per diversi valori di $\theta \in \Theta$, indicheremo con $f(\boldsymbol{y};\theta)$ la probabilità (densità) del campione in corrispondenza del valore θ che ha generato effettivamente il campione.

3.3.1 Un'importante disuguaglianza

L'intuito suggerirebbe che punti interessanti dello spazio parametrico Θ sono quelli per cui la funzione di verosimiglianza (o di log-verosimiglianza) assume valori elevati. Ci si aspetta inoltre che la funzione di verosimiglianza sia elevata proprio in corrispondenza del valore θ che ha effettivamente generato i dati. Questo potrebbe non essere vero per un singolo campione $\boldsymbol{y}$ realizzazione della variabile aleatoria $\boldsymbol{Y}$, tuttavia se si considera il valore atteso di tale quantità allora vale la disuguaglianza presentata nel seguente teorema.

Teorema 3.1. Il valore atteso della log-verosimiglianza in θ, valore del parametro per il modello che ha generato il campione, è sempre maggiore del valore medio calcolato in $\theta' \neq \theta$ ovvero

$$E(\ell(\theta)) \geq E(\ell(\theta')) \qquad \forall \theta' \in \Theta$$

Dimostrazione. In virtù della disuguaglianza di Jensen, per la concavità della funzione logaritmo, si ha che:

$$E\left(\log \frac{f(\boldsymbol{Y};\theta')}{f(\boldsymbol{Y};\theta)}\right) \leq \log E\left(\frac{f(\boldsymbol{Y};\theta')}{f(\boldsymbol{Y};\theta)}\right).$$

Si osservi che il termine al secondo membro è nullo in quanto

$$E\left(\frac{f(\boldsymbol{Y};\theta')}{f(\boldsymbol{Y};\theta)}\right) = \int_{R_{\boldsymbol{Y}}} \frac{f(\boldsymbol{y};\theta')}{f(\boldsymbol{y};\theta)} f(\boldsymbol{y},\theta) d\boldsymbol{y} = \int_{R_{\boldsymbol{Y}}} f(\boldsymbol{y};\theta') d\boldsymbol{y} = 1$$

Da questo segue che

$$E(\log f(\boldsymbol{Y};\theta') - \log f(\boldsymbol{Y};\theta)) \leq 0$$

da cui la tesi. Si osservi anche che la condizione di identificabilità $f(\boldsymbol{y};\theta') \neq f(\boldsymbol{y};\theta)$ se $\theta \neq \theta'$ implica che $P\left(\frac{f(\boldsymbol{Y};\theta')}{f(\boldsymbol{Y};\theta)} \neq 1\right) > 0$. □

Il valore di $E\left(\log\frac{f(\boldsymbol{Y};\theta')}{f(\boldsymbol{Y};\theta)}\right)$ misura quindi la differenza fra le due distribuzioni che hanno come parametro θ' e θ. In realtà tale misura, che non è tecnicamente una distanza fra distribuzioni, è una misura di divergenza (una divergenza richiede condizioni meno stringenti di una misura di distanza). Essa è (a meno del segno) nota come la divergenza di Kullback-Leibler, ed è usata per valutare quanto diversa sia una distribuzione di probabilità presa come riferimento da un'altra definita sul medesimo supporto.

3.3.2 La funzione punteggio (score)

Nello studio della funzione di verosimiglianza, nel caso in cui Θ è un intervallo di $\mathbb{R}$, sarà importante analizzare come il valore della stessa funzione cresca o decresca in corrispondenza di specifici valori del parametro. È particolarmente interessante osservare i valori della verosimiglianza (o della log-verosimiglianza) quando ci si trovi vicino al valore θ che ha generato i dati (abbiamo già mostrato come in media la log-verosimiglianza sia in quel punto più elevata). A tal fine è di rilievo la funzione punteggio di seguito introdotta.

Definizione 3.3 (Funzione punteggio o funzione *score*)**.** La **funzione punteggio**, comunemente detta *funzione score*, dal termine in inglese, è la derivata della funzione di log-verosimiglianza rispetto a θ:

$$s(\boldsymbol{y},\theta)=\frac{d\ell(\theta)}{d\theta}=\frac{d}{d\theta}\log f(y_1,y_2,\ldots,y_n;\theta)=\sum_{i=1}^{n}\frac{d}{d\theta}\log f(y_i;\theta)$$

ove l'ultima uguaglianza vale solo nel caso di campione casuale.

Ci si attende che tale funzione assuma valori non lontani da 0 al valore di θ che ha generato i dati. E in effetti dimostreremo che tale funzione è in media pari a 0 per le famiglie parametriche per le quali valgono alcune condizioni di regolarità.

Condizioni di regolarità

Una famiglia parametrica Y caratterizzata da $f(y;\theta)$ con $\theta\in\Theta$ è regolare se valgono le seguenti condizioni:

1. Lo spazio parametrico Θ è un intervallo aperto (ovvero non include i valori estremi dell'intervallo);
2. il supporto di Y non dipende dal parametro θ;
3. è possibile scambiare il segno di derivata e quello di integrale nel definire quantità come $E(g(\boldsymbol{Y}))$;
4. la log-verosimiglianza è una funzione differenziabile due volte rispetto a θ e la quantità $E\left(s(\boldsymbol{Y},\theta)^2\right)$ esiste ed è finita.

Si noti che tali condizioni sono presenti in molte delle più comuni famiglie parametriche (ad esempio, Poisson, binomiale, gaussiana, gamma) ma che invece non sono valide per la famiglia uniforme o rettangolare, $U(0,\theta)$, che non soddisfa la seconda condizione.

Teorema 3.2. Si consideri un modello parametrico per Y, distribuita secondo la legge $f(y;\theta)$, con $\theta \in \Theta$, per il quale valgano le condizioni di regolarità sopra elencate e sia disponibile il campione casuale $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$. Allora il valore atteso della funzione punteggio è nullo:

$$E(s(\boldsymbol{Y},\theta)) = E\left(\frac{d}{d\theta}\log f(\boldsymbol{Y};\theta)\right) = 0.$$

Dimostrazione. Il valore atteso $E(s(\boldsymbol{Y},\theta))$ può essere scritto come segue:

$$E(s(\boldsymbol{Y},\theta)) = \int_{R_{\boldsymbol{Y}}} \left(\frac{d}{d\theta}\log f(\boldsymbol{y};\theta)\right) f(\boldsymbol{y};\theta)d\boldsymbol{y}.$$

Si osservi che vale

$$\frac{d}{d\theta}\log f(\boldsymbol{y};\theta) = \frac{\frac{d}{d\theta}f(\boldsymbol{y},\theta)}{f(\boldsymbol{y};\theta)}$$

e pertanto possiamo scrivere

$$\begin{aligned} E(s(\boldsymbol{Y},\theta)) &= \int_{R_{\boldsymbol{y}}} \frac{\frac{d}{d\theta}f(\boldsymbol{y},\theta)}{f(\boldsymbol{y};\theta)} f(\boldsymbol{y},\theta)d\boldsymbol{y} \\ &= \int_{R_{\boldsymbol{Y}}} \frac{d}{d\theta}f(\boldsymbol{y},\theta)d\boldsymbol{y} = \frac{d}{d\theta}\int_{R_{\boldsymbol{Y}}} f(\boldsymbol{y},\theta)d\boldsymbol{y} = \frac{d1}{d\theta} = 0, \end{aligned}$$

dove si è applicato il fatto di poter invertire il segno di integrazione e derivazione rispetto a θ.

□

3.3.3 L'informazione attesa di Fisher

La funzione punteggio per un dato campione assume valori che si aggirano attorno allo 0 in corrispondenza del valore del parametro θ che definisce la distribuzione da cui provengono i dati campionari. Un altro aspetto rilevante riguarda la velocità con cui la funzione punteggio varia in media (cioè al variare del campione) quando ci si muove attorno al valore "vero" θ. Tale aspetto può essere valutato guardando alla curvatura media della funzione di verosimiglianza in θ. Questa quantità considerata col segno negativo è detta **informazione attesa di Fisher** ed è indicata con $I(\theta)$:

$$I(\theta) = -E\left(\frac{d^2}{d\theta^2}\log f(\boldsymbol{Y};\theta)\right).$$

Si noti che la log-verosimiglianza nei pressi del valore θ avrà derivate seconde elevate se la sua curvatura è elevata (e quindi le derivate prime assumono valori elevati positivi o negativi, per cui in media la derivata prima della funzione punteggio è grande, cioè varia molto al variare del campione). Questo aspetto è espresso dal seguente risultato.

Teorema 3.3. Si consideri Y, distribuita secondo la legge $f(y;\theta)$ con $\theta \in \Theta$ per la quale valgono le condizioni di regolarità sopra elencate ed sia $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ un campione casuale da Y. Allora

$$I(\theta) = -E\left(\frac{d^2}{d\theta^2}\log f(\boldsymbol{Y};\theta)\right) = E(s(\boldsymbol{Y},\theta)^2) = V(s(\boldsymbol{Y},\theta))$$

Dimostrazione.

$$\begin{aligned} I(\theta) &= E\left(-\frac{d^2}{d\theta^2}\log f(\boldsymbol{Y};\theta)\right) = E\left\{-\frac{d}{d\theta}\left(\frac{\frac{d}{d\theta}f(\boldsymbol{Y};\theta)}{f(\boldsymbol{Y};\theta)}\right)\right\} \\ &= E\left\{-\frac{\left(\frac{d^2}{d\theta^2}f(\boldsymbol{Y};\theta)\right)f(\boldsymbol{Y};\theta) - \left(\frac{d}{d\theta}f(\boldsymbol{Y};\theta)\right)^2}{f(\boldsymbol{Y};\theta)^2}\right\} \\ &= -E\left\{\frac{\frac{d^2}{d\theta^2}f(\boldsymbol{Y};\theta)}{f(\boldsymbol{Y};\theta)}\right\} + E\left\{\frac{(\frac{d}{d\theta}f(\boldsymbol{Y};\theta))^2}{f(\boldsymbol{Y};\theta)^2}\right\} \end{aligned}$$

Possiamo allora scrivere

$$I(\theta) = -\int_{R_{\boldsymbol{Y}}} \frac{\frac{d^2}{d\theta^2} f(\boldsymbol{y};\theta)}{f(\boldsymbol{y};\theta)} f(\boldsymbol{y};\theta) d\boldsymbol{y} + \int_{R_{\boldsymbol{Y}}} \left(\frac{\frac{d}{d\theta} f(\boldsymbol{y};\theta)}{f(\boldsymbol{y};\theta)} \right)^2 f(\boldsymbol{y},\theta) d\boldsymbol{y}$$

(sfruttando le condizioni di regolarità e ricordando che il valore atteso della funzione punteggio è nulla)

$$= -\frac{d^2}{d\theta^2} \int_{R_{\boldsymbol{Y}}} f(\boldsymbol{y};\theta) d\boldsymbol{y} + \int_{R_{\boldsymbol{Y}}} s(\boldsymbol{y};\theta)^2 f(\boldsymbol{y};\theta) d\boldsymbol{y}$$

$$= -\frac{d^2 1}{d\theta^2} + \int_{R_{\boldsymbol{Y}}} s(\boldsymbol{y},\theta)^2 f(\boldsymbol{y};\theta) d\boldsymbol{y} = E(s(\boldsymbol{Y},\theta)^2) = V(s(\boldsymbol{Y},\theta)).$$

□

Quindi se una verosimiglianza è molto piatta (cioè non mostra preferenza per particolari valori di θ) la quantità di informazione data dal campione sarà poca, se invece la verosimiglianza in media decresce velocemente appena ci si allontana dal valore di θ "vero" allora ci aspettiamo per quel modello che vi sia molta informazione da parte del campione riguardo al parametro ignoto.

Inoltre se siamo nel caso di un campione i.i.d., ovvero nel caso di campionamento casuale vale quanto segue:

$$i(\theta) = \frac{I(\theta)}{n},$$

dove

$$i(\theta) = -E\left(\frac{d^2}{d\theta^2} \log f(Y_i;\theta) \right)$$

$$= E\left\{ \left(\frac{d}{d\theta} \log f(Y;\theta) \right)^2 \right\} = V\left(\frac{d}{d\theta} \log f(Y;\theta) \right)$$

è l'informazione attesa per una singola determinazione dalla variabile Y. Le variabili y_i sono tutte identicamente distribuite e sfruttando l'indipendenza si ha

$$\int_{R_{\boldsymbol{Y}}} \left[\frac{d}{d\theta} \log f(\boldsymbol{y};\theta) \right]^2 f(\boldsymbol{y};\theta) d\boldsymbol{y}$$

$$= \int_{R_{y_1}} \cdots \int_{R_{y_n}} \left[\sum_{i=1}^{n} \frac{d}{d\theta} \log f(y_i;\theta) \right]^2 f(y_1,\ldots,y_n;\theta) dy_1, \ldots dy_n$$

e poiché in caso di indipendenza la varianza di una somma è pari alla somma delle varianze, si deduce che

$$\begin{aligned} I(\theta) &= \int_{R_{\boldsymbol{Y}}} \left[\frac{d}{d\theta}\log f(\boldsymbol{y};\theta)\right]^2 f(\boldsymbol{y};\theta)d\boldsymbol{y} \\ &= n\int_{R_y} \left[\frac{d}{d\theta}\log f(y;\theta)\right]^2 f(y;\theta)dy \\ &= nV\left(\frac{d}{d\theta}\log f(y;\theta)\right) = ni(\theta). \end{aligned}$$

Capitolo 4

Stima puntuale

4.1 Stime e stimatori

Nell'introduzione si era descritto in modo informale il problema della stima puntuale come segue: dato un modello statistico parametrico per cui Y è distribuita secondo la legge $f(y;\theta)$, si vuole individuare un valore $\hat{\theta} \in \Theta$ che si ritiene sia più plausibile avendo osservato un campione casuale $(y_1, y_2, \ldots, y_n)$ da Y. Il valore $\hat{\theta}$ è ottenuto come determinazione di una statistica, che chiameremo **stimatore**, $\hat{\theta} = s(Y_1, Y_2, \ldots, Y_n)$ ed è quindi una variabile aleatoria. Il valore di tale statistica calcolata per il campione osservato $(y_1, y_2, \ldots, y_n)$ fornisce una **stima puntuale** di θ.

Chiaramente è ragionevole chiedere che le statistiche campionarie da candidare come stimatori forniscano valori nello spazio parametrico Θ. Per cui uno stimatore è una funzione che ha come dominio lo spazio campionario e codominio lo spazio parametrico.

Nel seguito, non del tutto propriamente, si denoterà con $\hat{\theta}$ sia lo stimatore (ovvero la variabile aleatoria $s(Y_1, Y_2, \ldots, Y_n)$) che il valore specifico che lo stesso assume una volta osservato un campione (ovvero la stima $\hat{\theta} = s(y_1, y_2, \ldots, y_n)$). Sarà il contesto a chiarire se si stia parlando della variabile aleatoria stimatore o della stima.

Nell'ambito della statistica classica sarà cruciale determinare la distribuzione dello stimatore $\hat{\theta}$ ovvero la distribuzione dei valori che potrei osservare se la stima fosse calcolata immaginando di ripetere il campionamento. Infatti, in genere, si osserva solo uno dei possibili valori della stima in quanto disporremo di un solo campione, ma per capire quanto tale stima sia affida-

bile e per misurare i margini di incertezza insiti nella procedura inferenziale faremo riferimento alle caratteristiche e alle proprietà della variabile aleatoria stimatore.

Il termine stima puntuale si riferisce al fatto che otteniamo un solo valore (punto) fra quelli che potrebbe generare lo stimatore (in inglese si infatti usa il termine *point estimation*).

Non esiste ovviamente una sola statistica che si può proporre come stimatore e la scelta fra diversi stimatori alternativi andrà fatta definendo alcune proprietà che ci aspettiamo abbia un "buono" stimatore, proprietà che riguardano la distribuzione degli stimatori alternativi.

Vedremo in questo capitolo, sempre restando nell'ambito della statistica classica (frequentista), quali siano i criteri da usare per giudicare alcuni stimatori più interessanti di altri e vedremo anche di come possano individuare statistiche da candidare come stimatori. Particolare attenzione verrà dedicata all'individuazione di stimatori a partire dall'esame della funzione di verosimiglianza (quantità già introdotta nel precedente capitolo).

4.1.1 Alcuni esempi elementari

Si noti che in realtà il problema di stima si pone più in generale per qualsiasi grandezza caratteristica della variabile Y. Potremmo essere quindi interessati a stimare la media Y, la sua varianza, un particolare quantile di Y o particolari valori di $F_Y(y)$. In molti casi le quantità di interesse coincidono con il parametro che caratterizza la distribuzione di Y ma in altri casi le quantità che si vogliono stimare possono essere espressi come funzioni del parametro stesso.

Stima del parametro di una bernoulliana (stima di una proporzione)

Per il modello introdotto nel paragrafo 1.4.1 si poneva il problema di fare inferenza sul valore p di una variabile bernoulliana che rappresentava la proporzione di palline bianche in un urna a partire dal risultato di una sequenza di n estrazioni con reinserimento dalla stessa urna. Si è poi osservato in 1.4.2 che lo stesso modello statistico era valido anche nel caso si voglia fare inferenza sulla proporzione p di unità che in una popolazione (molto ampia) avessero una determinata caratteristica.

I risultati del paragrafo 1.2 ci consentono di affermare che se si dispone di un campione casuale estratto da una variabile aleatoria di Bernoulli $(Y_1, Y_2, \ldots, Y_n)$ per la media $\overline{Y}$ (che nello specifico caso chiameremo anche **proporzione campionaria** e denoteremo con $\hat{p}$)

$$\overline{Y} = \hat{p} = \frac{k}{n} \tag{4.1}$$

ove k è il numero di valori per cui è $y_1 = 1$ nel campione, si ha

$$\overline{Y} \sim \frac{1}{n}\text{Bin}(n, p).$$

e pertanto $E(\overline{Y}) = E(\hat{p}) = p$ inoltre $V(\overline{Y}_n) = \frac{p(1-p)}{n}$.

Questo implica che i valori osservati di $\hat{p}$ per un dato campione si aggireranno attorno al vero valore p. Talvolta otterremo un valore più grande del vero valore p altre volte un valore più piccolo, se potessimo però ripetere il campionamento più volte i diversi valori di $\hat{p}$ sarebbero in media pari al valore vero. Si noti anche che la varianza di tali valori sarà più piccola all'aumentare della dimensione del campione n.

D'altra parte, per questo modello statistico, se cresce la dimensione del campione vale la legge debole dei grandi numeri applicata per la successione di variabili aleatorie $\overline{Y}_n$ (aggiungiamo il pedice n per sottolineare che consideriamo la media campionaria per campioni di ampiezza n crescente) che garantisce la sua convergenza in probabilità al vero valore p.

Quindi la statistica $\hat{p}$, proporzione campionaria, è senza dubbio uno stimatore interessante per la probabilità p che caratterizza la popolazione bernoulliana.

Stima della media per una gaussiana

Nel paragrafo 1.4.5 si assume che i dati siano tratti da una popolazione nella quale Y è distribuita secondo una legge gaussiana di media μ e varianza σ^2.

Abbiamo già osservato nel paragrafo 2.2.1 che la statistica $\overline{Y}_n = (1/n)\sum_i Y_i$ ha media pari a μ e varianza pari a σ^2/n (ove, ancora una volta, con il pedice n si evidenzia che si tratta di una sequenza aleatoria essendo la media calcolata su un campione di n elementi).

Quindi tale statistica fornisce valori che "ballano" attorno al vero valore del parametro del modello generatore μ. In particolare la statistica media campionaria ha distribuzione gaussiana $\overline{Y}_n \sim N(\mu, \sigma^2/n)$ e sempre la legge

dei grandi numeri garantisce che al divergere della dimensione campionaria la successione delle statistiche $\overline{Y}_n$ converge al valore μ.

Quindi la statistica media campionaria è uno stimatore ragionevole per la media μ della gaussiana da cui abbiamo tratto i dati.

Tuttavia si noti che in questo caso si potrebbe considerare anche la statistica mediana campionaria:

$$Me(Y) = \begin{cases} \frac{Y_{(\frac{n}{2})}+Y_{(\frac{n}{2}+1)}}{2}, & \text{se n è pari} \\ Y_{(n+1)/2}, & \text{se n è dispari.} \end{cases}$$

Si è già osservato (si veda il paragrafo 2.4) che la statistica $Me(Y)$ fornisce valori in media pari a μ e quindi si può considerare anche questa statistica per stimare μ.

Ma quale fra i due stimatori preferire? Ovviamente in un singolo campione non sappiamo quale dei due fornisce il valore più vicino al parametro μ (altrimenti conosceremmo μ) ma possiamo scegliere sulla base della distribuzione dei due stimatori quello che mediamente fornisce valori meno distanti da μ.

Un'idea potrebbe essere quella di guardare agli scostamenti al quadrato $(\overline{Y} - \mu)^2$ e $(Me(Y) - \mu)^2$ fra valore delle due statistiche e valore del parametro. Impiegando il quadrato dello scostamento si attribuisce peso maggiore a scostamenti elevati e stesso peso a scostamenti di pari valore assoluto. Fra i due stimatori quello che (se si ripetesse più volte il campionamento) ha in media scostamenti quadratici più piccoli è da preferire.

Quindi si dovrebbe calcolare $E[(\overline{Y} - \mu)]^2$ ed $E[Me(Y) - \mu]^2$ e scegliere la statistica per cui lo scostamento quadratico medio risulti più piccolo.

Tale calcolo potrebbe esser molto agevole nel primo caso perché esso coinvolge le distribuzione della media campionaria (che è combinazione lineare di variabili) mentre risulterebbe più laborioso nel secondo che coinvolge la distribuzione di statistiche ordinate.

Si potrebbe inoltre osservare che in entrambi i casi gli stimatori proposti hanno un buon comportamento al crescere della dimensione del campione e convergono al vero valore di μ.

4.2 Misurare la qualità di uno stimatore

Negli esempi sopra illustrati si è avuto già modo di osservare che gli stimatori interessanti sono quelli che forniscono valori che sono "vicini" al vero e ignoto

parametro θ del modello generatore dei dati. Questo dovrebbe accadere qualunque sia $\theta \in \Theta$.

Una misura della qualità di uno stimatore è quindi ragionevole che sia funzione crescente della quantità $|\hat{\theta} - \theta|$ o di tale differenza al quadrato. L'**errore quadratico medio** è definito proprio come il valore atteso della differenza fra stima e parametro al quadrato:

Definizione 4.1 (Errore quadratico medio)**.** Dato un modello statistico parametrico per cui Y è distribuita secondo la legge $f(y, \theta)$ e avendo a disposizione un campione casuale $(Y_1, Y_2, \ldots, Y_n)$ si definisce l'errore quadratico medio di uno stimatore $\hat{\theta} = g(Y_1, Y_2, \ldots, Y_n)$ per il parametro θ come

$$\mathrm{EQM}_{\hat{\theta}}(\theta) = E\{(\hat{\theta} - \theta)^2\}$$

È immediato verificare, semplicemente aggiungendo e togliendo all'espressione in parentesi $E(\hat{\theta})$, che si ha

$$\begin{aligned}\mathrm{EQM}_{\hat{\theta}}(\theta) &= E\{[(\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta)]^2\} \\ &= E\{[\hat{\theta} - E(\hat{\theta})]^2\} + E\{[E(\hat{\theta}) - \theta]^2\} + 2E[\hat{\theta} - E(\hat{\theta})][E(\hat{\theta}) - \theta)] \\ &= E\{[\hat{\theta} - E(\hat{\theta})]^2\} + [E(\hat{\theta}) - \theta]^2\end{aligned}$$

in quanto $E[\hat{\theta} - E(\hat{\theta})] = 0$. La quantità $E(\hat{\theta}) - \theta = B(\theta)$ è denominata **distorsione** per cui $\mathrm{EQM}_{\hat{\theta}}(\theta)$ è la somma di due componenti

- la distorsione al quadrato

$$B(\theta)^2 = [E(\hat{\theta}) - \theta]^2$$

- la varianza dello stimatore

$$V(\hat{\theta}) = E\{[\hat{\theta} - E(\hat{\theta})]^2\}$$

Si noti che l'EQM è funzione dell'ignoto valore di θ e un buon stimatore dovrebbe avere un EQM che è quanto più basso possibile per ogni $\theta \in \Theta$. Se, ad esempio, utilizzassimo come stimatore di θ una costante c (che non dipende dai dati quindi (ma che potrebbe rappresentare ad esempio un valore che riteniamo tipico per quel parametro) nel caso in cui effettivamente $\theta = c$

l'EQM sarebbe nullo tuttavia esso crescerebbe molto velocemente se il vero valore di θ è diverso da c poiché si tratterebbe di uno stimatore distorto $\forall \theta \neq c$.

Esempio 4.1 (Media campionaria)**.** Si consideri un modello parametrico per cui Y è distribuita secondo una legge di media μ e varianza σ^2 supposta nota. Si dispone di un campione casuale di ampiezza n e si propone quale stimatore di μ la statistica $\overline{Y} = (1/n)\sum_{i=1}^{n} Y_i$. Il suo errore quadratico medio è costante e pari a σ^2/n in quanto, come già noto, $E(\overline{Y}) = \mu$ e quindi $\mathrm{EQM}_{\overline{Y}}(\mu) = E(\overline{Y} - \mu)^2 = \sigma^2/n$. Se considerassimo lo stimatore $\hat{\theta}_1 = c$ che ignora le informazioni campionarie (ed è quindi quantomeno poco interessante) avremmo che $\mathrm{EQM}_{\hat{\theta}_1} = (c - \mu)^2$, che risulterebbe minore di $\mathrm{EQM}_{\overline{Y}}$ per $c - \sigma/\sqrt{n} < \mu < c + \sigma/\sqrt{n}$. ▲

L'esempio mostra chiaramente come in generale non sia possibile trovare uno stimatore con $\mathrm{EQM}_{\hat{\theta}}(\theta)$ che risulti più basso qualunque sia il valore di θ. Tuttavia, si noti che se si lascia tendere a infinito la dimensione del campione n, lo stimatore $\overline{Y}$ avrebbe EQM che tende a 0 per ogni μ, cosa che non accadrebbe per lo stimatore $\hat{\theta}_1 = c$, questo induce a credere che possa essere utile prendere in esame stimatori il cui EQM tende a 0 se n tende a infinito.

Il motivo che potrebbe rendere interessante lo stimatore $\overline{Y}$ è la sua caratteristica di avere distorsione pari a 0. Qualunque sia μ i valori dello stimatore avranno valore atteso che è pari al parametro da stimare e quindi forniranno valori che saranno "attorno" a μ.

Esempio 4.2 (Varianza campionaria non corretta)**.** Si consideri un campione i.i.d. da una popolazione Y che ha media μ e varianza σ^2. Si voglia usare per σ^2 il seguente stimatore

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = S^2\frac{n-1}{n}. \tag{4.2}$$

Per quanto esposto al 2.2 si ha che $E(\hat{\sigma}^2) = E(S^2\frac{n-1}{n}) = \sigma^2\frac{n-1}{n}$ e pertanto si tratta di uno stimatore distorto del parametro σ^2. La sua distorsione è pari a $B(\sigma^2) = \sigma^2\frac{n-1}{n} - \sigma^2 = -\sigma^2/n$ e sarà quindi tanto minore quanto più ampio è il campione. Tuttavia, la varianza di tale stimatore potrebbe essere inferiore a quella di S^2 per alcuni modelli distributivi di Y e quindi non si può escludere che tale stimatore abbia un EQM più basso di quello di S^2.

▲

Gli esempi illustrano che il problema a ottenere uno stimatore che abbia $\text{EQM}_{\hat{\theta}}(\theta)$ uniformemente più basso di qualsiasi altro stimatore è complicato dal *trade-off* che esiste tra distorsione e varianza: ovvero, per n finito, posso ridurre la varianza al prezzo di accettare la presenza di una distorsione oppure posso annullare la distorsione ma a quel punto devo accettare la varianza che ne consegue. Un buon stimatore sarà quindi quello per cui si realizza un buon compromesso fra distorsione e varianza qualunque sia θ.

4.2.1 Stimatori non distorti ed efficienza

Una possibile strategia nella ricerca di "buoni stimatori" è quella di controllare la distorsione e cercare stimatori che a parità di distorsione abbiano varianza più bassa possibile. Un'idea potrebbe essere quella di porre l'attenzione su quegli stimatori che non abbiano distorsione. Stimatori con distorsione nulla hanno valore atteso che è pari al parametro da stimare e per essi è talvolta usato, non del tutto propriamente, il termine "corretto". Vale la seguente definizione:

Definizione 4.2 (Stimatore non distorto)**.** Lo stimatore $\hat{\theta}$ per il parametro θ è non distorto se $E(\hat{\theta}) - \theta = 0$ ovvero se, $\forall\theta$, il valore atteso dello stimatore è pari al parametro che si vuole stimare.

Molta parte della teoria statistica classica ha enfatizzato l'attenzione sulla ricerca di stimatori non distorti. In tal caso il confronto in termini di EQM fra due stimatori entrambi non distorti si riduce al confronto fra le loro varianze.

Per uno stimatore non distorto $\hat{\theta}$ di θ la $V(\hat{\theta})$ costituisce una immediata misura sintetica della sua qualità. In realtà spesso si utilizza la radice quadrata di $V(\hat{\theta})$ che è detta **errore standard** dello stimatore.

Nel caso di stimatori non distorti si può definire $\hat{\theta}_1$ come più efficiente di $\hat{\theta}_2$ se $V(\hat{\theta}_1) < V(\hat{\theta}_2)$ per almeno un valore θ.

Il risultato che verrà presentato nel paragrafo successivo, noto come limite di Rao-Cramer, consente di giudicare l'efficienza assoluta di uno stimatore non distorto per i parametri di alcuni dei più comuni modelli parametrici confrontando la sua varianza con un valore limite che rappresenta il più basso valore cui si può pervenire.

4.2.2 Limite inferiore di Rao-Cramer per la varianza di stimatori non distorti

Il risultato che viene presentato è noto anche come disuguaglianza di Rao-Cramer e stabilisce che, sotto alcune condizioni soddisfatte per le più comuni famiglie parametriche, è possibile individuare un valore che rappresenta un limite inferiore per uno stimatore non distorto di un parametro. Se quindi si dispone di uno stimatore non distorto che abbia varianza pari a tale limite allora tale stimatore è massimamente efficiente qualunque sia θ. Il confronto della varianza di qualsiasi stimatore non distorto con il limite di Rao-Cramer è utile per valutarne la sua qualità.

Al fine di enunciare e poi dimostrare la disuguaglianza di Rao-Cramer, conviene richiamare la notazione introdotta nel paragrafo 3.3. Quindi denoteremo ancora con $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ l'intero vettore dei dati, realizzazioni del vettore aleatorio $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$. Per denotare il valore atteso di una statistica come $E(\hat{\theta})$ ove $\hat{\theta} = g(y_1, y_2, \ldots, y_n)$

$$E(\hat{\theta}) = \int_{R_{y_1}} \int_{R_{y_2}} \cdots \int_{R_{y_n}} g(y_1, y_2, \ldots, y_n) f(y_1, y_2, \ldots, y_n; \theta) dy_1, dy_2 \ldots dy_n$$

si scriverà

$$E(\hat{\theta}) = \int_{R_{\boldsymbol{Y}}} g(\boldsymbol{y}) f(\boldsymbol{y}; \theta) d\boldsymbol{y},$$

ove l'integrale si intende esteso sull'intero spazio campionario $\mathbb{R}_{\boldsymbol{Y}}$, e $f(\boldsymbol{y}; \theta)$ indica la probabilità (densità) del campione in corrispondenza del valore θ che ha generato effettivamente il campione.

Teorema 4.1 (Disuguaglianza di Rao-Cramer)**.** Sia $(Y_1, Y_2, \ldots, Y_n)$ un campione casuale tratto da una famiglia parametrica Y caratterizzata da $f(y;\theta)$, con $\theta \in \Theta$, per la quale valgono le condizioni di regolarità enunciate nel paragrafo 3.3.2, ovvero:

1. L'insieme Θ è un intervallo aperto (ovvero non include i valori estremi dell'intervallo).
2. Il supporto di Y non dipende da θ.
3. è possibile scambiare il segno di derivata e quello di integrale nel definire quantità come $E(g(\boldsymbol{Y}))$;
4. la log-verosimiglianza è una funzione differenziabile due volte rispetto a θ e la quantità $E\left(s(\boldsymbol{Y},\theta)^2\right)$ esiste ed è finita.

Sia $\hat{\theta} = g(\boldsymbol{Y})$ uno stimatore non distorto del parametro θ:

$$E(\hat{\theta}) = \theta;$$

allora

$$V(\hat{\theta}) \geq \frac{1}{I(\theta)} = \left[-E\left(\frac{d^2}{d\theta^2}\log f(\boldsymbol{y};\theta)\right)\right]^{-1}$$

la quantità $I(\theta)$ è l'informazione attesa di Fisher, introdotta in 3.3.3, e quindi nel caso di un campione i.i.d. si ha $I(\theta) = ni(\theta)$, per cui la disuguaglianza può essere espressa come segue

$$V(\hat{\theta}) \geq \frac{1}{nE\left\{\left(\frac{d}{d\theta}\log f(y;\theta)\right)^2\right\}}$$

Dimostrazione. Si consideri la funzione punteggio $s(\boldsymbol{Y},\theta) = \dfrac{d}{d\theta}\log f(\boldsymbol{Y};\theta)$ e si ricordi che per essa vale $E(s(\boldsymbol{Y},\theta)) = 0$. Si consideri ora la covarianza fra le due quantità aleatorie $U = s(\boldsymbol{Y},\theta)$ e $W = \hat{\theta} = g(\boldsymbol{Y})$ che, per la nullità del valore atteso della funzione punteggio, è semplicemente pari al valore atteso del prodotto $\mathrm{Cov}(U,W) = E(s(\boldsymbol{Y},\theta)\cdot\hat{\theta})$.

La disuguaglianza di Schwarz implica che

$$\mathrm{Cov}(U,W)^2 \leq V(U)V(W) = V(s(\boldsymbol{Y},\theta)V(\hat{\theta})) = -E\left(\frac{d^2}{d\theta^2}\log f(\boldsymbol{Y};\theta)\right)V(\hat{\theta})$$

Si osservi che

$$\begin{aligned}\mathrm{Cov}(U,W) = E(s(\boldsymbol{Y},\theta)\cdot\hat{\theta}) &= \int_{R_{\boldsymbol{Y}}}\left(\frac{d}{d\theta}\log f(\boldsymbol{y};\theta)\right)g(\boldsymbol{y})f(\boldsymbol{y};\theta)d\boldsymbol{y}\\ &= \int_{R_{\boldsymbol{Y}}} g(\boldsymbol{y})\frac{\frac{d}{d\theta}f(\boldsymbol{y},\theta)}{f(\boldsymbol{y};\theta)}f(\boldsymbol{y},\theta)d\boldsymbol{y}\\ &= \frac{d}{d\theta}\int_{R_{\boldsymbol{Y}}} g(\boldsymbol{y})f(\boldsymbol{y},\theta)d\boldsymbol{y} = \frac{d}{d\theta}E(\hat{\theta}) = \frac{d}{d\theta}\theta = 1\end{aligned}$$

dove negli ultimi passaggi si sfrutta la condizione che sia possibile scambiare derivata e integrale, e infine la condizione di non distorsione di $\hat{\theta}$. Quindi

$$1 \leq -E\left(\frac{d^2}{d\theta^2}\log f(\boldsymbol{Y};\theta)\right)V(\hat{\theta})$$

da cui segue

$$V(\hat{\theta}) \geq \frac{1}{-E\left(\frac{d^2}{d\theta^2}\log f(\boldsymbol{Y};\theta)\right)} = \frac{1}{I(\theta)}$$

□

In alcuni casi si potrà reperire uno stimatore non distorto per θ che raggiunge tale limite. Tale stimatore sarà quindi preferibile, qualunque sia θ, a qualsiasi altro stimatore che condivida la proprietà della non distorsione ed è pertanto lo stimatore MVND (stimatore a *Minima Varianza fra i Non Distorti*) quindi massimamente efficiente.

Si ricordi però che tale limite vale se sono presenti le condizioni di regolarità e che, inoltre, in molti casi si possono ottenere stimatori distorti con distorsione non eccessiva e bassa varianza. Tuttavia alcune delle famiglie parametriche di impiego più comune (Binomiale, Normale, Poisson, Gamma, etc.) sono famiglie regolari e per valutare la qualità degli stimatori dei loro parametri, se non distorti, la disuguaglianza è estremamente utile. In condizioni di regolarità esistono importanti risultati teorici che consentono di stabilire l'esistenza e di reperire lo stimatore MVND.

Negli esempi che seguono vedremo dei casi in cui possiamo definire il limite inferiore per la varianza di uno stimatore non distorto e osserveremo che

alcuni stimatori non distorti, già proposti in alcuni degli esempi, introdotti raggiungono tale limite.

Esempio 4.3 (Limite per la varianza di uno stimatore del parametro di una bernoulliana)**.** Si consideri la funzione score a partire dalla funzione di verosimiglianza già ottenuta nel paragrafo 3.1; la derivata della funzione score è

$$\frac{d^2}{dp^2}\ell(p) = \frac{d}{dp}\left[\frac{\sum_{i=1}^n y_i}{p} - \frac{(n-\sum_{i=1}^n y_i)}{(1-p)}\right] = -\left[\frac{\sum_{i=1}^n y_i}{p^2} + \frac{(n-\sum_{i=1}^n y_i)}{(1-p)^2}\right]$$

Ricordando che $E(\sum_{i=1}^n Y_i) = np$, l'informazione attesa di Fisher, cioè il valore atteso della derivata seconda di $\ell(p)$ cambiato di segno, è pari a

$$E\left[\frac{\sum_{i=1}^n Y_i}{p^2} + \frac{(n-\sum_{i=1}^n Y_i)}{(1-p)^2}\right] = \frac{np}{p^2} + \frac{n-np}{(1-p)^2} = \frac{n}{p(1-p)}$$

Qualsiasi stimatore non distorto del parametro di una bernoulliana non potrà avere varianza inferiore a $\frac{p(1-p)}{n}$. Si ricordi che questa è la varianza dello stimatore $\hat{p}$, proporzione campionaria, che pertanto risulta corretto e fra i corretti massimamente efficiente.

▲

Esempio 4.4 (Limite per la varianza di uno stimatore del parametro di una Poisson)**.** La funzione di log-verosimiglianza per un campione casuale da una Poisson di media $E(Y) = \lambda$ è

$$\ell(\lambda) = \sum_{i=1}^n [-\lambda + y_i \log\lambda - \log y_i!] = -n\lambda + \log\lambda \sum_{i=1}^n y_i - \sum_{i=1}^n \log y_i!$$

otteniamo ora la funzione score

$$\frac{d}{d\lambda}\ell(\lambda) = -n + \frac{\sum_{i=1}^n y_i}{\lambda}$$

calcoliamo ora l'informazione attesa come il valore atteso del quadrato della funzione score

$$E\left[-n + \frac{\sum_{i=1}^n Y_i}{\lambda}\right]^2 = E\left[-n + \frac{1}{\lambda}n\overline{Y}\right]^2 = E\left[\frac{n(\overline{Y}-\lambda)^2}{\lambda}\right] = \frac{n^2}{\lambda^2}E(\overline{Y}-\lambda)^2$$

Quindi si ottiene $\frac{n^2}{\lambda^2}V(\overline{Y}) = \frac{n}{\lambda}$.

▲

Esempio 4.5 (Limite per la varianza di uno stimatore della media di una Normale)**.** La log-verosimiglianza per la media μ di una gaussiana Y, dato un campione casuale $(y_1, y_2, \ldots, y_n)$ è stata già ottenuta in (3.2). Consideriamo però ora la varianza σ^2 nota

$$\ell(\mu) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}$$

Occorre calcolare il valore atteso della funzione score al quadrato e possiamo ricordare che essa, nel caso di campione casuale, è pari a $ni(\mu)$, dove $i(\mu)$ rappresenta il valore atteso del quadrato della funzione score per un singolo elemento campionario che ha densità pari a quella della popolazione $f(y, \mu)$:

$$i(\mu) = E\left[\frac{d}{d\mu}\ell(\mu)\right]^2 = E\left[\frac{(Y-\mu)}{\sigma^2}\right]^2 = \frac{1}{\sigma^4}E(Y-\mu)^2 = \frac{1}{\sigma^2}$$

L'informazione attesa di Fisher è quindi $I(\mu) = ni(\mu) = n/\sigma^2$ e il limite di Rao-Cramer per stimatori non distorti di μ è σ^2/n. Avevamo visto che anche per stimare μ si potevano considerare sia la media che la mediana campionaria entrambi stimatori non distorti. Tuttavia la media campionaria ha varianza che raggiunge il limite di Rao-Cramer e pertanto è senza dubbio uno stimatore preferibile.

▲

4.3 Proprietà asintotiche per sequenze di stimatori

Si disponga di un campione casuale $(Y_1, Y_2, \ldots, Y_n)$ da una popolazione nella quale Y è distribuita secondo la legge $f(y;\theta)$ e sia $\hat{\theta}_n = g(Y_1, Y_2, \ldots, Y_n)$ uno stimatore del parametro θ. Si noti che la funzione che definisce lo stimatore è la stessa quale che sia la dimensione campionaria n. Per evidenziare che consideriamo una sequenza di stimatori al crescere di n si è posto il suffisso n allo stimatore $\hat{\theta}$. Il comportamento aleatorio di tale sequenza di stimatori se il campione diventa sempre più grande ovvero se n tende a infinito è evidentemente di estremo rilievo per giudicare la sua qualità.

Quello che ci si aspetta è che la sequenza $\hat{\theta}_n$ al crescere della dimensione campionaria fornisca stime sempre meno incerte e prossime al vero valore del

parametro. La principale fonte di incertezza in una procedura inferenziale è infatti legata all'impiego di un campione di dimensione limitata, se fosse possibile osservare l'intera popolazione allora si dovrebbe poter dire con certezza come è distribuita Y e quindi si dovrebbe conoscere il valore del parametro che definisce la sua distribuzione. Porre quindi delle condizioni sul comportamento della successione aleatoria $\hat{\theta}_n$ al crescere della ampiezza campionaria n permette di selezionare sequenze di stimatori che si comportano coerentemente con questa aspettativa.

Le proprietà degli stimatori che riguardano il comportamento di sequenze di stimatori per ampiezza crescente del campione sono dette **proprietà asintotiche**. Mentre le proprietà dello stimatore in cui si esamina il comportamento aleatorio dello stimatore per una fissata ampiezza campionaria sono dette proprietà per campioni finiti.

4.3.1 Consistenza

La successione aleatoria di stimatori $\hat{\theta}_n = g(Y_1, Y_2, \ldots, Y_n)$ dovrebbe quindi convergere, secondo un opportuno criterio di convergenza, alla costante θ ovvero al parametro ignoto che caratterizza la distribuzione di Y nella popolazione da cui è tratto il campione. Il termine "consistenza", che è consuetudine usare per definire tale proprietà, deriva dalla (non adeguata) traduzione dal termine inglese *consistent* che in effetti sarebbe meglio reso dal termine italiano "coerente". In effetti, una sequenza di stimatori che avendo campioni sempre più grandi non tenda ad avvicinarsi sempre più al vero valore avrebbe un comportamento "incoerente".

La definizione di sequenza consistente di stimatori viene quindi qualificata dal criterio di convergenza che si sceglie di adottare. Vale quindi la seguente definizione:

Definizione 4.3 (Consistenza di una sequenza di stimatori)**.** Si disponga di un campione casuale $(Y_1, Y_2, \ldots, Y_n)$ da una variabile aleatoria Y distribuita secondo la legge $f(y;\theta)$. La sequenza di stimatori $\hat{\theta}_n = g(Y_1, Y_2, \ldots, Y_n)$ del parametro θ è consistente se $\theta_n \to \theta$ quando l'ampiezza campionaria n diverge. La natura della convergenza definisce anche il tipo di convergenza per cui:

1. se $\hat{\theta}_n \xrightarrow{qc} \theta$ allora la sequenza di stimatore è fortemente consistente,
2. se $\hat{\theta}_n \xrightarrow{p} \theta$ allora la sequenza di stimatore è debolmente consistente,
3. se $\hat{\theta}_n \xrightarrow{mq} \theta$ allora vi è consistenza in media quadratica.

Fra i diversi tipi di consistenza intercorrono quindi le medesime relazioni che intercorrono fra i tipi di convergenza per successioni di variabili aleatorie. La **consistenza forte** implica le altre mentre quella in media quadratica implica quella debole. Questo ultimo aspetto consente di ottenere facilmente una condizione necessaria per la verifica della **consistenza debole**.

Si osservi infatti che la **consistenza in media quadratica** è definita come segue:

$$\lim_{n\to\infty} E\{[\hat{\theta}_n - \theta]^2\} = \lim_{n\to\infty} \mathrm{EQM}_{\hat{\theta}_n}(\theta) = 0.$$

quindi

$$\lim_{n\to\infty} \mathrm{EQM}_{\hat{\theta}_n}(\theta) = \lim_{n\to\infty} \{E[\hat{\theta}_n - E(\hat{\theta}_n)]^2 + [E(\hat{\theta}_n) - \theta]^2\}$$

pertanto deve accadere che congiuntamente valgano i seguenti limiti

$$\lim_{n\to\infty} E\{[\hat{\theta}_n - E(\hat{\theta}_n)]^2\} = 0$$
$$\lim_{n\to\infty} [E(\hat{\theta}_n) - \theta]^2 = 0$$

la seconda condizione è equivalente a

$$\lim_{n\to\infty} E(\hat{\theta}_n) = \theta$$

che definisce la proprietà di *n*on distorsione asintotica. Pertanto una sequenza di stimatori è consistente se contemporaneamente la sequenza dei valori attesi degli stimatori tende a valore vero del parametro e la varianza tende a 0.

Se tali due condizioni sono verificate allora la sequenza di stimatori è necessariamente anche debolmente consistente.

Nel seguito, per brevità definiremo la consistenza come una proprietà dello stimatore omettendo di precisare che ci si riferisce alla sequenza degli stessi al crescere di n.

Esempio 4.6 (Varianza campionaria non corretta)**.** Si consideri il problema della stima della varianza di una popolazione Y che è gaussiana di media μ e varianza σ^2. Si era mostrato nell'esempio 4.2 che lo stimatore di σ^2

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = S^2\frac{n-1}{n}. \tag{4.3}$$

ha distorsione pari a $-\sigma^2/n$, il cui limite per n che tende a infinito è 0. Quindi si tratta di uno stimatore asintoticamente non distorto della varianza. Inoltre ricordando che $S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$, e se $W \sim \chi^2_{n-1}$ allora $V(W) = 2(n-1)$, si ha $V(S^2) = \frac{2\sigma^4}{n-1}$ e quindi

$$V(\hat{\sigma}^2) = V\left(S^2\frac{n-1}{n}\right) = \frac{2\sigma^4(n-1)}{n^2}$$

che tende a 0 se $n \to \infty$. Pertanto lo stimatore $\hat{\sigma}^2$ è consistente per σ^2.

▲

4.3.2 Approssimazione asintotiche

Si è più volte sottolineato che ai fini di giudicare la qualità della procedura inferenziale secondo il paradigma della statistica classica è essenziale individuare la distribuzione delle statistiche campionarie impiegate. E abbiamo visto come nel caso di alcuni semplici stimatori di quantità caratteristiche di modelli parametrici sia immediato ricavare la distribuzione; si pensi al caso della distribuzione della media campionaria per campioni da una variabile gaussiana. In molti altri casi ottenere la distribuzione dello stimatore è tutt'altro che agevole: sia perché lo stimatore potrebbe avere una forma più complessa sia perché il modello statistico che genera i dati è meno agevole da trattare.

Di notevole interesse è il caso in cui una sequenza di stimatori $\hat{\theta}_n$, o una sua opportuna trasformazione eventualmente dipendente da alcune costanti

note, converga in distribuzione a una variabile aleatoria nota G. Ad esempio, se accade che la sequenza aleatoria

$$b_n(\hat{\theta}_n - a_n) \xrightarrow{d} G$$

dove a_n e b_n sono costanti, diremo che $\hat{\theta}_n$ ha distribuzione asintotica G.

Uno dei casi che incontreremo più di frequente è quello in cui si dispone di un campione casuale e la distribuzione asintotica è la gaussiana. In particolare, se

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \mathcal{V})$$

allora $\sqrt{n}(\hat{\theta}_n - \theta)$ ha distribuzione asintotica gaussiana e $\mathcal{V}$ è detta **varianza asintotica**. Questo implica che, per n sufficientemente grande, lo stimatore $\hat{\theta}_n$ ha **distribuzione asintotica** $\mathcal{N}(\theta, \mathcal{V}/n)$ e scriveremo

$$\hat{\theta}_n \;\dot{\sim}\; \mathcal{N}(\theta, \mathcal{V}/n)$$

In tal caso lo stimatore $\hat{\theta}_n$ è asintoticamente non distorto. Inoltre, se la varianza asintotica è pari a $i^{-1}(\theta)$, allora $\hat{\theta}_n$ ha varianza che raggiunge il limite di Rao-Cramer e lo stimatore è detto **asintoticamente efficiente**.

Esempio 4.7 (Proporzione campionaria)**.** Si riconsideri il caso di un campione casuale da un modello bernoulliano e l'uso della proporzione campionaria $\hat{p}$ per stimare il parametro p. Si ricordi che in realtà noi conosciamo la distribuzione esatta di questo stimatore in quanto $n\hat{p} \sim \text{Bin}(n, p)$. Tuttavia sappiamo anche che, in virtù del teorema del limite centrale, la sequenza aleatoria

$$\frac{\sqrt{n}(\hat{p} - p)}{\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}(0, 1)$$

pertanto lo stimatore $\hat{p}$ ha distribuzione asintotica gaussiana e quindi

$$\hat{p} \;\dot{\sim}\; \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

L'aspetto di maggiore interesse pratico è che per n grande

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \;\dot{\sim}\; \mathcal{N}(0, 1)$$

Ricordando inoltre quanto emerso nell'esempio 4.3 possiamo anche affermare che tale stimatore è MVND $\forall n$. Si noti però che per usare concretamente l'approssimazione asintotica sarebbe necessario conoscere la varianza $p(1-p)/n$ che dipende dal parametro ignoto p. Tuttavia al denominatore al posto di p si può inserire la sua stima consistente, $\hat{p}$, e l'approssimazione asintotica sarà ancora valida; quindi per n grande possiamo scrivere

$$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \;\dot{\sim}\; \mathcal{N}(0,1).$$

▲

4.4 Stime di massima verosimiglianza

Nel capitolo 3 è stata introdotta la funzione di verosimiglianza

$$L(\theta) = f(y_1, y_2, \ldots, y_n; \theta) = \prod_{1=1}^{n} f(y_i; \theta)$$

ove l'ultima uguaglianza è lecita nel caso di campione casuale $(y_1, y_2, \ldots, y_n)$ di dimensione n da Y distribuita secondo il modello $f(y; \theta)$.

Si era osservato come essa per ogni $\theta \in \Theta$ fornisce un valore che misura quanto è plausibile (verosimile) che il campione osservato sia generato da un modello caratterizzato da quello specifico θ. I valori di θ per i quali $L(\theta)$ è più alto sono quelli più plausibili.

In 3.1 si è considerata un'urna con solo quattro possibili composizioni si era osservato che l'urna cui corrispondeva il valore più alto della verosimiglianza fosse quella che era più plausibile avesse generato il campione di palline osservato e quindi se si dovesse sceglierne una andrebbe scelta proprio quella.

Si può generalizzare tale idea e scegliere come stima di un parametro θ quel valore $\hat{\theta}_{MV}$ tale che $L(\hat{\theta}_{MV})$ sia in assoluto più elevato. Il valore $\hat{\theta}_{MV}$ è detto **stima di massima verosimiglianza**.

Definizione 4.4 (Stima di massima verosimiglianza)**.** La stima di massima verosimiglianza (SMV) $\hat{\theta}_{MV}$ di un parametro θ è tale che

$$L(\hat{\theta}_{MV}) \geq L(\theta) = f(y_1, y_2, \ldots, y_n; \theta) \qquad \forall \theta \in \Theta$$

quindi per un campione casuale (ovvero nel caso i.i.d)

$$\hat{\theta}_{MV} = \arg\max_{\theta \in \Theta} \prod_{1=1}^{n} f(y_i; \theta)$$

Si noti che se la funzione di verosimiglianza $L(\theta)$ venisse moltiplicata per una costante non dipendente da θ non cambierebbe il valore della SMV e in effetti si era osservato nel precedente capitolo che la funzione di verosimiglianza può essere definita a meno di una costante moltiplicativa $L(\theta) = c \cdot f(y_1, y_2, \ldots, y_n; \theta) \propto f(y_1, y_2, \ldots, y_n; \theta)$. Tale costante può essere eventualmente quindi ignorata nel determinare il valore $\hat{\theta}_{MV}$ per cui risulta che $L(\hat{\theta}_{MV})$ è massimo.

Inoltre per determinare la SMV è spesso conveniente lavorare sulla funzione di log-verosimiglianza $l(\theta) = \log L(\theta)$ infatti, essendo il logaritmo una trasformazione strettamente monotona si ha, nel caso di campione casuale

$$\hat{\theta}_{MV} = \arg\max_{\theta \in \Theta} l(\theta) = \arg\max_{\theta \in \Theta} \log \left(\prod_{1=1}^{n} f(y_i; \theta) \right) = \arg\max_{\theta \in \Theta} \sum_{1=1}^{n} \log f(y_i; \theta)$$

Prima di introdurre alcuni esempi di determinazione della stima di massima verosimiglianza è il caso di precisare che:

1. la SMV è definita come il massimo globale della funzione di verosimiglianza; occorrerebbe quindi essere certi che qualunque procedura si usi per individuare il massimo di $L(\theta)$ essa conduca a individuare tale punto di massimo globale. Non vi è garanzia che la funzione di verosimiglianza non presenti massimi locali: potrebbe accadere in particolare con campioni di dimensione molto limitata piccoli e per alcuni specifici modelli statistici;

2. non è garantita l'esistenza della SMV, ovvero la funzione di verosimiglianza potrebbe non essere limitata;

3. nel caso il modello statistico non sia identificabile, ovvero esistono almeno due valori del parametro θ_1 e θ_2 cui corrisponde la medesima legge distributiva per cui $f(y;\theta_1) = f(y;\theta_2)$ la SMV potrebbe non essere unica.

Tuttavia, nei casi di maggiore interesse pratico e per alcuni fra i modelli statistici più comuni si può provare l'unicità e l'esistenza della SMV. Per tali modelli la ricerca della SMV non presenta grosse criticità tuttavia in alcuni casi sarà necessario ricorrere a opportuni algoritmi numerici per reperire il massimo della funzione di verosimiglianza.

4.4.1 Determinazione della stima di massima verosimiglianza: alcuni esempi

Se la funzione di verosimiglianza è continua e differenziabile in θ il suo massimo può essere cercato calcolando la derivata rispetto a θ e uguagliandola a zero. Si tratta di risolvere cioè la seguente equazione

$$\frac{d}{d\theta}L(\theta) = \frac{d}{d\theta}f(y_1, y_2, \ldots, y_n;\theta) = \frac{d}{d\theta}\prod_{i=1}^{n} f(y_i;\theta) = 0 \tag{4.4}$$

oppure, in alternativa, si può considerare la funzione di log-verosimiglianza,

$$\frac{d}{d\theta}\log(L(\theta)) = \frac{d\ell(\theta)}{d\theta} = s(\boldsymbol{y};\theta) = \frac{d}{d\theta}\sum_{i=1}^{n}\log f(y_i;\theta) = 0 \tag{4.5}$$

dove l'ultima uguaglianza, in entrambe le equazioni, vale nel caso di campione casuale.

L'equazione (4.5) è detta **equazione di verosimiglianza** e andrebbe, a rigore, fatta la verifica che il valore che è soluzione dell'equazione sia un massimo globale.

Con riferimento al campione osservato, è poi possibile calcolare la derivata seconda della log-verosimiglianza cambiata di segno. Essa rappresenta la curvatura della funzione di verosimiglianza attorno al suo massimo e fornisce un'indicazione di quanto velocemente cala la verosimiglianza appena ci si sposta dal valore per cui essa è massima. Se tale curvatura è elevata, un punto θ_0 anche poco distante da $\hat{\theta}_{MV}$ avrebbe verosimiglianza di molto inferiore. Questo indica che il campione dà indicazioni "forti" su $\hat{\theta}_{MV}$. L'informazione

su di esso sembrerebbe quindi piuttosto precisa se tale derivata seconda è grande. Se si calcola il valore della derivata seconda in $\hat{\theta}_{MV}$ si ottiene

$$-\frac{d^2}{d\theta^2}\sum_{i=1}^{n}\log f(y_i;\hat{\theta}_{MV}) \tag{4.6}$$

Tale quantità è detta **informazione osservata di Fisher**. Si rammenti che nel capitolo precedente era stata introdotta l'informazione attesa di Fisher

$$I(\theta) = -E\left(\frac{d^2}{d\theta^2}\sum_{i=1}^{n}\log f(Y_i;\theta)\right)$$

e discusso il suo ruolo nella determinazione del limite di Rao-Cramer.

Determiniamo ora le SMV per alcuni dei modelli statistici per i quali si era introdotta la funzione di verosimiglianza nel capitolo 3.

Esempio 4.8 (SMV del parametro di una bernoulliana)**.** La funzione di verosimiglianza per il parametro p di una popolazione in cui Y è distribuita secondo una bernoulliana è (si veda 3.1)

$$L(p) = L(y_1, y_2, \ldots, y_n; p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i};$$

la funzione di log-verosimiglianza pertanto è

$$\ell(p) = \log L(y_1, y_2, \ldots, y_n; p) = \sum_{i=1}^{n} y_i \log p + \log(1-p)\left(n - \sum_{i=1}^{n} y_i\right) \tag{4.7}$$

Calcoliamo ora la funzione score

$$s(\boldsymbol{y}; p) = \frac{d}{dp}\ell(p) = \frac{\sum_{i=1}^n y_i}{p} - \frac{(n - \sum_{i=1}^n y_i)}{(1-p)}$$

e troviamo p che è soluzione dell'equazione di verosimiglianza

$$\frac{\sum_{i=1}^n y_i}{p} + \frac{(n - \sum_{i=1}^n y_i)}{(1-p)} = \frac{(1-p)\sum_{i=1}^n y_i - p(n - \sum_{i=1}^n y_i)}{p(1-p)} = 0$$

$$\sum_{i=1}^{n} y_i = pn$$

da cui otteniamo

$$\hat{p}_{MV} = \frac{\sum_{i=1}^{n} y_i}{n}$$

che è la proporzione campionaria.

▲

Esempio 4.9 (SMV per il parametro di una Poisson)**.** La (3.1) rappresenta la funzione di log-verosimiglianza per un campione casuale $(y_1, y_2, \ldots, y_n)$ da una Poisson di media $E(Y) = \lambda$

$$\ell(\lambda) = -n\lambda - \log \lambda \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \log y_i!$$

Si era inoltre ottenuta la funzione score

$$\frac{d}{d\lambda}\ell(\lambda) = -n + \frac{\sum_{i=1}^{n} y_i}{\lambda},$$

per cui l'equazione di verosimiglianza è

$$n + \frac{\sum_{i=1}^{n} y_i}{\lambda} = 0$$

risolvendo per λ si ottiene

$$\hat{\lambda}_{MV} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$$

Si noti che $\hat{\lambda}_{MV}$ è stimatore non distorto e che

$$V(\hat{\lambda}_{MV}) = V(\bar{y}) = \lambda/n$$

che è pari al limite di Rao-Cramer ottenuto nell'esempio 4.4. Pertanto si tratto di uno stimatore MVND.

▲

Esempio 4.10 (SMV per i parametri di una gaussiana)**.** La log-verosimiglianza per i parametri μ e σ^2 di una gaussiana Y, dato un campione casuale $(y_1, y_2, \ldots, y_n)$ è quella ottenuta in (3.2) che è quindi funzione di due variabili

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\sigma^2}.$$

In questo caso è possibile ottenere le stime di MV individuando la coppia (μ, σ^2) per cui la log-verosimiglianza risulta massima. Questo può essere fatto calcolando separatamente la funzione score per il parametro μ e per il parametro σ^2. Si tratta cioè di calcolare le derivate parziali rispetto ai parametri e poi eguagliarle a zero ottenendo quindi un sistema di equazioni la cui soluzione fornisce le stime di massima verosimiglianza. La log-verosimiglianza è quella ottenuta in 3.2 da cui si ottengono le funzioni score considerando prima la derivata parziale rispetto a μ

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = \frac{\sum_{i=1}^n (y_i - \mu)}{\sigma^2}.$$

e poi rispetto a σ^2

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^4}.$$

e quindi le due equazioni di verosimiglianza

$$\begin{aligned} \frac{\sum_{i=1}^n (y_i - \mu)}{\sigma^2} &= 0 \\ -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^4} &= 0 \end{aligned}$$

La prima equazione di verosimiglianza fornisce immediatamente lo stimatore di massima verosimiglianza per la media:

$$\hat{\mu}_{MV} = \frac{1}{n} \sum_{i=1}^n y_i = \overline{y}$$

e sostituendo nella seconda otteniamo lo stimatore di massima verosimiglianza per la varianza

$$\hat{\sigma}^2_{MV} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu}_{MV})^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \overline{y})^2 = \frac{S^2(n-1)}{n}$$

Si noti che lo stimatore di massima verosimiglianza per la varianza della gaussiana non ha la proprietà della non distorsione ma, come visto nell'esempio 4.6, è consistente per σ^2. ▲

Esempio 4.11 (SMV per il parametro θ di una uniforme $U(0,\theta)$)**.** Se disponiamo di un campione casuale $(y_1, y_2, \ldots, y_n)$ tratto da $Y \sim U(0,\theta)$, con $\theta > 0$ la funzione di verosimiglianza per θ è data (vedi esempio 3.2.3) da

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} = \frac{1}{\theta^n}, \qquad \text{se } y_{(n)} \leq \theta,$$

ed è pari a zero altrimenti.

Tale funzione è sempre decrescente per $\theta \geq y_{(n)}$ e la sua derivata è quindi sempre negativa. Assumerà quindi i suo valore massimo in corrispondenza del valore più piccolo del campo di definizione della funzione e pertanto la SMV è pari a $\hat{\theta}_{MV} = y_{(n)}$ ovvero al valore massimo osservato nel campione $(y_1, y_2, \ldots, y_n)$.

Si noti che tale stimatore ha evidentemente valore atteso che è inferiore a θ. Tuttavia è anche evidente che lo stimatore è consistente. Si noti che in questo caso non valgono le condizioni di regolarità (il dominio della Y dipende dal parametro).

▲

4.4.2 Stimatori di massima verosimiglianza e sue proprietà asintotiche

La funzione di verosimiglianza $L(\theta)$ può essere studiata immaginando di ripetere il campionamento, aspetto già considerato nel paragrafo 3.3 del capitolo precedente. Si considererà quindi la variabile aleatoria $\hat{\theta}_{MV}$ per valutare le proprietà dello stimatore di massima verosimiglianza. Di particolare rilievo sono le sue proprietà asintotiche; esse sono in effetti notevoli e questo giustifica il fatto che il metodo di stima della massima verosimiglianza abbia un rilievo di assoluta preminenza fra le procedure di stima.

Si noti innanzitutto che, come si è visto negli ultimi due esempi, non vi è garanzia che gli stimatori di massima verosimiglianza siano non distorti. Questo ancora una volta deve indurre a considerare la non distorsione come una proprietà non irrinunciabile mentre, come già evidenziato, è difficile giustificare uno stimatore che non sia consistente.

La sequenza di stimatori di MV al divergere della dimensione campionaria n è consistente sotto condizioni piuttosto generali e, nel caso dei parametri dei modelli di maggior interesse, la distribuzione asintotica degli stimatori di MV è gaussiana e gli stimatori di MV sono asintoticamente efficienti.

Spesso è utile poter ottenere l'errore standard dello stimatore di MV definito come $\sqrt{V(\hat{\theta}_{MV})}$, che è la principale misura sintetica della qualità dello stimatore nel caso in cui sia non distorto (o se la distorsione è molto piccola, cosa che in effetti accade per campioni casuali di dimensione ampia).

Teorema 4.2 (Proprietà asintotiche degli stimatori di massima verosimiglianza)**.** Sia $\hat{\theta}_n$ lo stimatore di massima verosimiglianza per il parametro θ, che caratterizza la legge distributiva per Y, ottenuta da un campione casuale di dimensione n. Allora, sotto condizioni piuttosto generali

1. la sequenza di stimatori $\hat{\theta}_n$ converge in probabilità a θ.
2. la sequenza
$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{i(\theta)}\right)$$
dove $i(n)$ è l'informazione attesa per un singolo valore tratto da Y
$$i(\theta) = E\left\{\left(\frac{d}{d\theta}\log f(y;\theta)\right)^2\right\}$$

Per lo stimatore di massima verosimiglianza $\hat{\theta}_{MV}$ per n grande vale quindi la seguente approssimazione asintotica (ricordando che per un campione casuale $I(\theta) = ni(\theta)$)
$$\hat{\theta}_{MV} \dot{\sim} \mathcal{N}\left(\theta, \frac{1}{I(\theta)}\right)$$

Le dimostrazioni delle proprietà asintotiche e gli aspetti formali vengono qui omessi (anche se è facile intuire che tali dimostrazioni sfrutteranno alcuni importanti risultati per sequenze aleatorie come la legge dei grandi numeri e il teorema del limite centrale). Peraltro se valgono le condizioni di regolarità introdotte in 3.3.2 allora la convergenza della sequenza al parametro è quasi certa.

È importante segnalare che lo stimatore di massima verosimiglianza, oltre a essere asintoticamente non distorto, è anche asintoticamente efficiente in quanto la varianza della distribuzione asintotica è pari al limite di Rao-Cramer.

4.4.3 Invarianza della SMV e stima di funzioni dei parametri

All'inizio del capitolo si era sottolineato come il problema di stima può riguardare altre quantità caratteristiche della legge distributiva di Y che in generale sono ignote e possono essere espresse come funzioni del parametro θ. Ad esempio, si pensi o al caso in cui si voglia stimare il rapporto di scommessa (*odds*) per una probabilità che è pari a $p/(1-p)$ per una popolazione bernoulliana $Y \sim Be(p)$, oppure la probabilità di osservare nessun evento $P(Y = 0) = e^{-\lambda}$ per una popolazione $Y \sim Po(\lambda)$. O ancora si potrebbe voler stimare la media $E(Y) = \theta$ di una popolazione esponenziale con legge $f(y; \lambda) = \lambda e^{-\lambda y}$ con $\lambda > 0$, che è quindi pari a $\theta = \frac{1}{\lambda}$.

Si noti che il problema potrebbe in realtà porsi come stima del parametro successiva a una diversa parametrizzazione della legge distributiva della variabile di interesse Y. La legge di una variabile esponenziale è talvolta espressa direttamente in termini della sua media θ, detto parametro di scala, invece che in termini del parametro λ che è il tasso di accadimento.

Anche la legge distributiva di una bernoulliana potrebbe essere espressa in termini del logaritmo del rapporto di scommessa (odds)

$$\theta = \log\left(\frac{p}{1-p}\right) = \text{logit}(p)$$

(che è noto come trasformazione *logit* di una probabilità) da cui si ricava $p = \frac{e^\theta}{1+e^\theta}$. In tal caso si noti che, riprendendo l'espressione in (4.7) essa può essere scritta come

$$\sum_{i=1}^{n} y_i \log p + (n - \sum_{i=1}^{n} y_i) \log(1-p) = \sum_{i=1}^{n} y_i \log\left(\frac{p}{1-p}\right) + n \log(1-p).$$

La funzione di log-verosimiglianza in θ risulta essere

$$\ell(\theta) = \theta \sum_{i=1}^{n} y_i - \log(1 + e^\theta).$$

La soluzione dell'equazione di verosimiglianza per θ è tuttavia meno agevole.

Nel paragrafo 3.2.2 si era introdotto la proprietà di invarianza della funzione di verosimiglianza rispetto a parametrizzazioni mediante funzioni biunivoche $\phi = h(\theta)$ e quindi il valore $\hat{\phi}_{MV}$ per cui è massima la verosimiglianza $L(\phi)$

espressa in funzione di ϕ è pari al valore $\hat{\theta}_{MV}$ per cui è massima la verosimiglianza espressa in funzione di θ trasformato tramite la funzione $g(\cdot)$, pertanto vale la seguente proprietà:

Definizione 4.5 (Proprietà di invarianza della SMV)**.** Sia $\hat{\theta}_{MV}$ la stima di massima verosimiglianza di θ e si consideri la trasformazione biunivoca $\phi = h(\theta)$. La SMV $\hat{\phi}_{MV}$ di ϕ è pari a $h(\hat{\theta}_{MV})$.

Per reperire la SMV di una trasformazione (o a seguito di una riparametrizzazione) $\phi = h(\theta)$ è sufficiente inserire il valore $\hat{\theta}_{MV}$ nell'espressione della funzione $h(\cdot)$.

Considerando quindi l'esempio fatto sopra per la trasformazione logit risulta immediato (e comodo) ottenere la stima θ_{MV} utilizzando la proprietà di invarianza da cui si ottiene

$$\hat{\theta}_{MV} = \log\left(\frac{\hat{p}_{MV}}{1-\hat{p}_{MV}}\right).$$

4.4.4 Approssimazione asintotica per SMV di funzioni dei parametri

La stima $\hat{\phi}_{MV} = h(\hat{\theta}_{MV})$ gode delle proprietà asintotiche delle SMV ed è pertanto consistente, la sua distribuzione può essere approssimata per n grande con una gaussiana di media pari al parametro ϕ e varianza $V(h(\hat{\theta}_{MV}))$. Tuttavia è difficile reperire la varianza di una trasformazione non lineare di una variabile aleatoria come $h(\hat{\theta}_{MV})$. In molti casi la varianza (eventualmente asintotica) della SMV del parametro θ è nota, come nell'esempio della funzione logit, ove la varianza della SMV di p è pari al limite di Rao-Cramer $V(h(\hat{p})) = p(1-p)/n$. Questa può essere approssimata con l'informazione osservata $\hat{p}_{MV}(1-\hat{p}_{MV})/n$ che, in virtù della consistenza di $\hat{p}_{MV}$, fornisce ancora uno stimatore consistente della varianza.

Potrebbe essere, in generale, interessante riuscire ad esprimere la varianza di $\hat{\phi}_{MV}$ in funzione della varianza di $\hat{\theta}_{MV}$. A questo scopo, sempre per n grande, vale la seguente approssimazione nota come **metodo delta**.

Definizione 4.6 (Metodo Delta). La varianza dello stimatore $\hat{\phi}_{MV}$ per il parametro $\phi = h(\theta)$, se n è sufficientemente grande, si può approssimare con la varianza dello stimatore $\hat{\theta}_{MV}$ moltiplicata per il quadrato della derivata $\frac{d}{d\theta}h(\theta)$:

$$V(h(\hat{\theta}_{MV})) = V(\hat{\theta}_{MV}) \left(\frac{d}{d\theta} h(\theta) \right)^2 .$$

Riprendendo quindi l'esempio visto in precedenza per la trasformata logit, si ha che la varianza di

$$\hat{\theta}_{MV} = \log \left(\frac{\hat{p}_{MV}}{1 - \hat{p}_{MV}} \right) = \text{logit}(\hat{p}_{MV})$$

per n grande, si può approssimare usando il metodo delta con

$$\begin{aligned} V(\hat{\theta}_{MV}) &= V(\hat{p}_{MV}) \left(\frac{d}{dp} \log \frac{p}{1-p} \right)^2 \\ &= \frac{p(1-p)}{n} \frac{1}{p(1-p)^2} \\ &= \frac{1}{np(1-p)} \end{aligned}$$

Quindi una stima consistente di tale varianza è

$$V(\hat{\theta}_{MV}) = \frac{1}{n\hat{p}_{MV}(1 - \hat{p}_{MV})}.$$

4.5 Altri metodi di stima

4.5.1 Metodo dell'inserimento

Il metodo della massima verosimiglianza è utilizzabile per ottenere la stima di un parametro della legge di distribuzione $f(y;\theta)$ di Y nella popolazione. Tuttavia non è infrequente che si voglia stimare una caratteristica di Y, come ad esempio la sua media, la sua varianza o una probabilità come $P(Y \leq k)$ per un valore k noto, anche nel caso in cui la distribuzione di Y non sia nota.

In tali casi risulta piuttosto naturale proporre come stimatori quelle funzioni del campione casuale che sono il corrispondente empirico delle caratteristiche da stimare.

Se quindi si vuole stimare la media di Y nella popolazione si può proporre come stima $\bar{y}$ ovvero la media campionaria di $(y_1, y_2, \ldots, y_n)$. Analogamente si può considerare la varianza campionaria S^2 come stima di $V(Y)$ e la proporzione di valori $y_i \leq k$ per stimare $P(Y \leq k)$. Abbiamo già trattato tali statistiche campionarie e sappiamo che in effetti forniscono stimatori non distorti delle grandezze da stimare e, ricordando alcuni risultati come la legge dei grandi numeri e il teorema del limite centrale, possiamo anche stabilire, per questi esempi, che si tratta di stimatori consistenti e con distribuzione asintotica gaussiana.

Un metodo semplice per proporre una stima di una caratteristica di Y, anche in assenza di ipotesi distributive su Y e quindi in un contesto non parametrico, è quello di proporre l'analoga caratteristica campionaria, anche nel caso in cui non abbiamo assunzioni distributive sulla Y stessa, ottenendo talvolta stimatori che hanno proprietà interessanti.

In realtà gli esempi proposti sono di fatto casi particolari dell'applicazione di un principio più generale che è noto come principio dell'inserimento (in inglese *plug-in principle*). Esso stabilisce che la caratteristica di una distribuzione Y può essere approssimata dalla stessa caratteristica della distribuzione empirica ottenuta da un campione casuale. Per esempio, il quantile di Y può essere approssimato dall'analogo quantile tratto dalla distribuzione empirica ottenuta a partire del campione. In effetti per la media $\mu = E(Y)$ si può proporre $\hat{\mu} = E_{F_n}(Y) = \sum_{j=1}^{n} y_j \hat{f}_j = \bar{y}$ ove $\hat{f}_j$ sono i valori di incremento della funzione di ripartizione empirica $F_n(y)$ nei punti distinti y_j e sono quindi pari a $\hat{f}_j = n_j/n$, ove n_j è il numero di valori nel campione pari a y_j.

4.5.2 Metodo dei momenti

L'idea di ottenere una stima utilizzando l'analogo campionario di una quantità caratteristica della popolazione può ovviamente essere impiegata anche nel caso in cui per Y abbiamo un'assunzione distributiva precisa, per cui $f(y;\theta)$ e l'inferenza riguarda quindi il parametro θ. In questo caso si possono considerare i momenti della Y che, di solito, avranno un'espressione che

dipende dal parametro ignoto. Definiti i momenti per i diversi ordini r di Y

$$\begin{aligned} m_1(\theta) =& E(Y) \\ m_2(\theta) =& E(Y^2) \\ & \vdots \\ m_r(\theta) =& E(Y^r) \end{aligned}$$

è possibile ottenere le equazioni del metodo dei momenti in cui si uguaglia il momento r-esimo di Y al momento campionario del medesimo ordine:

$$\begin{aligned} m_1(\theta) =& \frac{\sum_{i=1}^n y_i}{n} \\ m_2(\theta) =& \frac{\sum_{i=1}^n y_i^2}{n} \\ & \vdots \\ m_r(\theta) =& \frac{\sum_{i=1}^n y_i^r}{n} \end{aligned}$$

Possiamo risolvere una di queste equazioni, se abbiamo un solo parametro, per θ e il valore ottenuto fornisce la stima con il metodo dei momenti. Se abbiamo più parametri possiamo considerare tante equazioni quanti sono i parametri.

Se il parametro ignoto è il valore atteso di Y si può usare la prima equazione e, banalmente, lo stimatore con il metodo dei momenti risulta pari alla media campionaria. Consideriamo ora alcuni esempi di applicazione del metodo dei momenti.

Esempio 4.12. Sia $Y \sim U(0, \theta)$ con $\theta > 0$ e si disponga di un campione casuale $(y_1, y_2, \ldots, y_n)$. Il momento primo $m_1(\theta) = E(Y)$ è pari a $\theta/2$ pertanto si può considerare l'equazione del metodo dei momenti

$$m_1(\theta) = \theta/2 = \frac{\sum_{i=1}^n y_i}{n}$$

risolvendo per θ si ottiene la stima $\hat{\theta}_{MM} = 2\bar{y}$. Si noti che si tratta di uno stimatore non distorto. Infatti $2\bar{y} = 2\theta/2 = \theta$ ed è facile verificare la sua consistenza essendo la $V(\hat{\theta}_{MM}) = 4\theta^2/12n$ che se n tende a infinito tende a 0.

Tuttavia in un piccolo campione potrebbe accadere che $\hat{\theta}_{MM} < y_{(n)}$ ottenendo un valore della stima che non è coerente con i dati osservati poiché deve essere $y_{(n)} < \theta$. ▲

Il metodo dei momenti fornisce, sotto condizioni piuttosto generali, stimatori consistenti, e in taluni casi è agevole da calcolare. Tuttavia si noti che esso non è definito in modo univoco e, anche asintoticamente, non possiede proprietà ottimali. Per piccoli campioni potrebbe essere distorto e in alcuni casi potrebbe fornire stime al di fuori dello spazio parametrico o prive di senso come nell'esempio sopra considerato.

Capitolo 5

Stima intervallare

5.1 Esempio introduttivo

Si consideri un campione casuale di quattro osservazioni (y_1, y_2, y_3, y_4) estratto da una popolazione normale con media incognita μ e deviazione standard $\sigma = 3$. La stima di massima verosimiglianza di μ è la media campionaria $\bar{y}$ che per il campione osservato ammettiamo essere $\bar{y} = 2.8$. Per la variabile aleatoria $\overline{Y} = (1/4)\sum_{i=1}^{4} Y_i$ risulta, come è noto,

$$\overline{Y} \sim \mathcal{N}(\mu, 9/4)$$

da cui segue che la variabile aleatoria

$$Z = \frac{\overline{Y} - \mu}{\sqrt{9/4}}$$

sarà distribuita normalmente con media 0 e varianza unitaria e la sua funzione di densità che è indipendente dal valore del parametro. Se si fissa un valore $\alpha \in (0, 1)$ si può individuare una coppia di valori diciamo z_1 e z_2 con $z_1 < z_2$ tali che la probabilità che Z sia tra tali due valori arbitrari sia esattamente pari a $1 - \alpha$, ad esempio, possiamo scrivere

$$P(z_1 \leq Z \leq z_2) = \int_{z_1}^{z_2} \phi(z)dz = 1 - \alpha.$$

Ovviamente esistono infinite coppie (z_1, z_2) con tale proprietà. Possiamo nel caso in questione tuttavia considerare i valori z_1 e z_2 simmetrici rispetto

alla media ovvero $z_1 = -z_2$ che definiscono, com'è facile verificare, il più piccolo intervallo cui è associata una probabilità $1-\alpha$. Quindi in questo caso $P(Z > z_2) = P(Z \leq z_1) = \alpha/2$. Pertanto la $P(Z \leq z_2) = 1-\alpha/2$ e quindi z_2 è il quantile di ordine $1-\alpha/2$ in una normale standard cioè $\Phi^{-1}(1-\alpha/2)$. Se ad esempio fissiamo $\alpha = 0.05$ e quindi $1-\alpha = .95$ e $1-\alpha/2 = .975$ allora si ha $z_2 = \Phi^{-1}(1-\alpha/2) \approx 1.96$.

Quindi riconsiderando la variabile aleatoria $\overline{Y}$ si può scrivere

$$\begin{aligned} P(-1.96 \leq Z \leq 1.96) &= P\left(-1.96 \leq \frac{\overline{Y}-\mu}{\sqrt{9/4}} \leq 1.96\right) \\ &= P\left(\mu - 1.96\sqrt{9/4} \leq \overline{Y} \leq \mu + 1.96\sqrt{9/4}\right) \\ &= P\left(\mu - 2.94 \leq \overline{Y} \leq \mu + 2.94\right) = 0.95 \end{aligned}$$

Poiché le disequazioni $\mu - 2.94 \leq \overline{Y}$ e $\overline{Y} \leq \mu + 2.94$ sono equivalenti alle disuguaglianze $\mu \leq \overline{Y} + 2.94$ e $\mu \geq \overline{Y} - 2.94$, la relazione precedente può essere riscritta nella forma

$$P(\overline{Y} - 2.94 \leq \mu \leq \overline{Y} + 2.94) = 0.95. \tag{5.1}$$

L'espressione (5.1) definisce gli estremi di un intervallo aleatorio $(\overline{Y} - 2.94, \overline{Y} + 2.94)$ che contiene la media incognita μ con probabilità pari a 0.95. L'intervallo così definito è detto **intervallo di confidenza** (o intervallo fiduciario) con livello di confidenza 0.95. Il significato dell'intervallo sopra definito può essere chiarito dalla seguente affermazione: se si estraessero ripetutamente campioni casuali di ampiezza quattro dalla popolazione normale di varianza 9 e si calcolasse l'intervallo $(\overline{Y}-2.94, \overline{Y}+2.94)$ per ogni campione, la frequenza relativa di intervalli che contengono μ si avvicinerebbe molto a 0.95. Pertanto possiamo affermare che l'intervallo $(-0.14, 5.74)$ calcolato per l'unico campione osservato (e ottenuto sostituendo 2.8 a $\overline{Y}$ nella formula) sia, con una certa *confidenza*, uno di quelli per cui μ risulta coperto dall'intervallo. Si noti che l'intervallo espresso in termini della media campionaria $\bar{y}$ nel generico campione osservato, $(\bar{y} - 2.94, \bar{y} + 2.94)$, potrà contenere o meno il valore μ quello che possiamo affermare è solo che il generico intervallo $(\bar{y} - 2.94, \bar{y} + 2.94)$ racchiuderà μ al suo interno circa il 95% delle volte.

È questa l'interpretazione, da un punto di vista del campionamento ripetuto, della relazione

$$P(\mu - 2.94 \leq \overline{Y} \leq \mu + 2.94) = P(\overline{Y} - 2.94 \leq \mu \leq \overline{Y} + 2.94) = 0.95$$

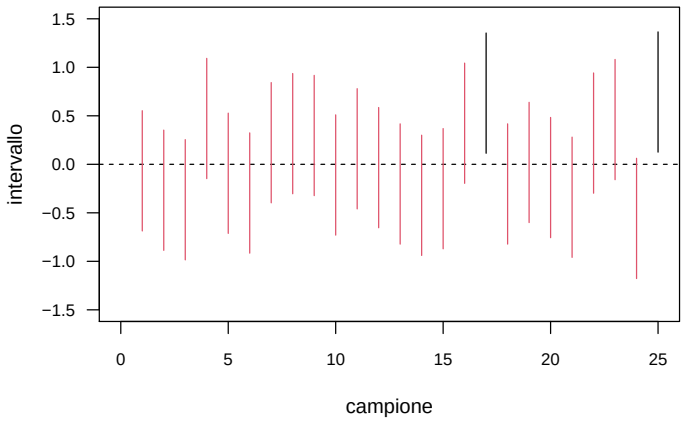


Figura 5.1: Intervalli di confidenza al 95% per μ per $N = 25$ campioni di ampiezza $n = 10$ da una distribuzione $\mathcal{N}(0, 1)$. Gli intervalli che contengono il valore $\mu = 0$ sono riportati in rosso, e si evince che $23/25 = 92\%$ degli intervalli contiene lo 0. Per $N \to \infty$ la proporzione degli intervalli che contiene il valore $\mu = 0$, ovvero il grado di copertura, tenderà a 0.95.

esaminata in precedenza (si veda la figura 5.1).

Si noti che la probabilità $1 - \alpha$ viene fissata ad un valore vicino a 1, poiché rappresenta il grado di fiducia sul fatto che l'intervallo costruito sia uno di quelli che contenga il parametro. Tale probabilità è detta **livello di confidenza** o **grado di copertura**. Si è già osservato che vi sono, in realtà, molti possibili intervalli con la stessa probabilità. Ad esempio, poiché risulta

$$P(-1.68 \leq Z \leq 2.70) = 0.95,$$

un altro intervallo di confidenza al livello del 95% per μ nell'esempio considerato è dato dall'intervallo di estremi $\bar{y} - 1.68(3/2)$ e $\bar{y} + 2.70(3/2)$, cioè dall'intervallo $(0.28, 6.85)$. Questo intervallo presenta una lunghezza data da $6.85 - 0.28 = 6.57$ che è maggiore di quella dell'intervallo $(-0.14, 5.74)$, e quindi fornisce una informazione meno precisa sulla posizione di μ. Per questo motivo, si individua l'intervallo per cui l'ampiezza $b - a$ sia minimizzata, per una certa area fissata, che per l'esempio considerato corrisponde a scegliere $b = -a$, per via della simmetria di $\phi(z)$ rispetto a $z = 0$, ottenendo così il più piccolo intervallo di confidenza al 95% per μ. Si noti che in questo caso l'idea di considerare un intervallo simmetrico ha condotto a un intervallo

che è il più corto possibile aspetto desiderabile visto che tale intervallo è più informativo se è più stretto.

Si osservi che nell'esempio proposto era cruciale la quantità Z sopra introdotta: è per mezzo di essa che è definito un intervallo di confidenza. Essa è una funzione del campione e del parametro incognito μ della popolazione ed ha una distribuzione indipendente dal parametro stesso e da ogni altro parametro. Pertanto espressioni della forma $P(a \leq Z \leq b) = 1 - \alpha$ daranno luogo a una proposizione di probabilità da cui si potrà ottenere un intervallo aleatorio che, con elevata probabilità, contiene il parametro. Questo metodo di costruzione di un intervallo di confidenza per un parametro di una popolazione sarà oggetto del paragrafo che segue.

5.2 Costruzione di intervalli di confidenza

5.2.1 Definizione

Nel paragrafo precedente è stato introdotto il concetto di intervallo di confidenza mediante un semplice esempio. Una definizione generale e più formale viene data di seguito.

Definizione 5.1 (Intervallo di confidenza)**.** Sia $(Y_1, Y_2, \ldots, Y_n)$ un campione casuale da una popolazione Y distribuita secondo la legge $f(y;\theta)$, di cui non conosciamo il parametro $\theta \in \Theta$. Siano $T_1 = t_1(Y_1, \ldots, Y_n)$ e $T_2 = t_2(Y_1, \ldots, Y_n)$ due statistiche che soddisfano $T_1 \leq T_2$, per le quali $P(T_1 \leq \theta \leq T_2) = 1 - \alpha, \alpha \in [0, 1]$. Allora l'intervallo aleatorio (T_1, T_2) è un **intervallo di confidenza** (o intervallo fiduciario) al $100(1-\alpha)\%$ per θ. Il valore $1 - \alpha$ è detto *livello di confidenza* e T_1, T_2 definiscono il limite di confidenza per θ inferiore e superiore, rispettivamente.

In questo contesto, la quantità α corrisponde alla probabilità che l'intervallo aleatorio (T_1, T_2) non contenga θ, e rappresenta quindi una misura del rischio di errore alla quale vengono solitamente assegnati valori piccoli come, ad esempio, $\alpha = 0.10, 0.05, 0.01$ che danno luogo ad intervalli di confidenza al 90%, al 95% e al 99%, rispettivamente. L'intervallo osservato dato da una coppia di valori numerici (t_1, t_2) associato ad un campione $(y_1, y_2, \ldots, y_n)$ è

il valore dell'intervallo (T_1, T_2) e rappresenta una *stima per intervallo* della quantità incognita θ (con una terminologia un po' impropria, si utilizza talvolta il termine "intervallo di confidenza" anche in riferimento a (t_1, t_2)).

Seppure sia meno consueto il suo impiego, va citato il caso in cui si pone $T_1 = -\infty$ oppure $T_2 = \infty$ ottenendo così un intervallo unilaterale.

Si osservi che una volta definito un intervallo di confidenza per θ è possibile determinare, fissato $1-\alpha$, un intervallo di confidenza al livello $1-\alpha$ per una *qualsiasi* funzione monotona di θ, $\tau = \tau(\theta)$. Se assumiamo che τ sia una funzione monotona crescente e (T_1, T_2) è un intervallo di confidenza al livello $1-\alpha$ per θ, avendo posto $T_i = t_i(Y_1, \ldots, Y_n)$, $i = 1, 2$, allora

$$P(T_1 \leq \theta \leq T_2) = P(\tau(T_1) \leq \tau(\theta) \leq \tau(T_2)) = 1 - \alpha,$$

e pertanto $(\tau(T_1), \tau(T_2))$ è un intervallo di confidenza al $100(1-\alpha)\%$ per $\tau(\theta)$.

5.2.2 Il metodo della funzione pivot

L'approccio adottato nell'esempio precedente si basava sulla individuazione della variabile aleatoria Z, funzione dei dati campionari, con distribuzione completamente nota così da poter determinare gli estremi di un intervallo con probabilità fissata e pari a $1-\alpha$ di contenere Z. La funzione Z inoltre conteneva nella sua espressione il valore del parametro di interesse ignoto ed era funzione monotona di questa. Una funzione dei dati campionari con tali caratteristiche è detta *funzione pivot*.

Definizione 5.2 (Funzione Pivot)**.** Sia $(Y_1, Y_2, \ldots, Y_n)$ un campione casuale da Y distribuita secondo la legge $f(y;\theta)$, $\theta \in \Theta$. Una funzione del campione e del parametro, $Q = q(Y_1, Y_2, \ldots, Y_n; \theta)$, la cui distribuzione sia nota e indipendente da θ viene detta **funzione pivot** (o quantità pivotale).

Si definisce inoltre una funzione pivot approssimata se la sua distribuzione asintotica non dipende da θ.

La funzione pivot è pertanto una quantità aleatoria che contiene nella sua espressione il parametro di interesse ma la cui distribuzione non dipende da esso.

Non è sempre agevole o immediato individuare una quantità pivotale da utilizzare nella determinazione di intervalli di confidenza. Di seguito riportiamo alcuni esempi notevoli per cui risulta piuttosto agevole individuare la funzione pivot e ricavare quindi l'intervallo di confidenza.

Intervallo di confidenza per la media di un'esponenziale

La variabile Y è distribuita secondo una legge esponenziale di parametro λ per cui

$$f(y, \lambda) = \lambda e^{\lambda y}.$$

Si dispone di un campione casuale di $(Y_1, Y_2, \ldots, Y_n)$ e si vuole fare inferenza su λ. A tal fine si osservi che la quantità $T = 2n\lambda\overline{Y}$ ha distribuzione nota. Infatti ricordando che somme di esponenziali indipendenti sono variabili Gamma, si ha $\sum_{i=1}^{n} Y_i \sim Ga(n, \lambda)$. Quindi, ricordando ancora che se si moltiplica una variabile aleatoria Gamma per una costante si ha ancora una gamma con parametro di forma che viene diviso per tale costante, si ha $T = 2\lambda \sum_{i=1}^{n} Y_i \sim Ga(n, \lambda/(2\lambda)) \equiv Ga(n, 1/2)$. Si tratta quindi di una funzione che ha distribuzione nota e che nella sua espressione contiene il parametro λ: ha quindi le caratteristiche di una funzione pivot.

Possiamo allora scrivere, fissata una probabilità $1-\alpha$, e considerando i due quantili $k_{\alpha/2}$ e $k_{1-\alpha/2}$ di una $Ga(n, 1/2)$

$$P\left(k_{\alpha/2} \leq 2n\lambda\overline{Y} \leq k_{1-\alpha/2}\right) = 1 - \alpha$$

da cui si ottiene

$$P\left(\frac{k_{\alpha/2}}{2n\overline{Y}} \leq \lambda \leq \frac{k_{1-\alpha/2}}{2n\overline{Y}}\right) = 1 - \alpha$$

e l'intervallo aleatorio $\left(\frac{k_{\alpha/2}}{2n\overline{Y}} \ , \ \frac{k_{1-\alpha/2}}{2n\overline{Y}}\right)$ costituisce quindi un intervallo di confidenza per λ al livello $(1-\alpha)\%$.

Si noti che la distribuzione $Ga(n, 1/2)$ è una distribuzione χ^2_{2n} per cui si potrebbero usare le tavole di tale distribuzione per individuare i due quantili. È anche il caso di segnalare che, in questo caso, essendo la distribuzione Gamma non simmetrica la scelta di lasciare a destra e a sinistra una probabilità uguale non darebbe luogo a un intervallo di ampiezza minima. Tuttavia in tali casi la scelta di avere pari probabilità sui due lati è spesso quella prevalente.

Intervallo di confidenza per la media di una gaussiana

Sia $(Y_1, Y_2, \ldots, Y_n)$ un campione casuale estratto da $Y \sim \mathcal{N}(\mu, \sigma^2)$; si vuole costruire un intervallo di confidenza per la media incognita μ. Si considera dapprima il caso in cui si assume nota la varianza della popolazione (così come nell'esempio introduttivo, successivamente si considera il caso, più realistico, di varianza ignota).

Varianza Nota Il caso in cui la varianza è nota è stato già illustrato nell'esempio introduttivo. Come visto, il rapporto

$$Z = \frac{\overline{Y} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

è una quantità pivot: infatti la sua distribuzione è una gaussiana standard ed è indipendente dal parametro μ. Inoltre la sua espressione contiene μ così che è possibile, fissata una probabilità $1 - \alpha$, cercare due valori z_1 e z_2 per cui risulti

$$P\left(z_1 \leq \frac{\overline{Y} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq z_2\right) = 1 - \alpha$$

Se in particolare si vuole che l'intervallo sia il più stretto possibile allora si può scrivere

$$P\left(-z_{1-\alpha/2} \leq \frac{\overline{Y} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

e quindi isolando μ al centro delle disequazioni si ottiene

$$P\left(\overline{Y} - z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \overline{Y} + z_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) = 1 - \alpha$$

Quindi $[\overline{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}]$ è un intervallo di confidenza al livello $(1 - \alpha)100\%$

Esempio 5.1. Si dispone dell'altezza (in centimetri) per un campione di $n = 30$ studenti. Per il suddetto campione la media è risultata essere $\bar{y} = 163.87$ cm e si assuma $\sigma^2 = 64$ noto. I limiti dell'intervallo di confidenza per μ al 95% sono $163.87 \pm 1.96(8/\sqrt{30})$, cioè 161.01 e 166.73. L'ampiezza dell'intervallo è $A = (2{\cdot}1.96{\cdot}8)/\sqrt{30} = 5.73$, mentre l'ampiezza dell'intervallo al 99% è $A = (2 \cdot 2.576 \cdot 8)/\sqrt{30} = 7.52$. ▲

Varianza ignota L'individuazione di una funzione pivot al fine di determinare un intervallo di confidenza per la media μ di una popolazione normale con varianza incognita σ^2, da un campione casuale di numerosità n, si basa su un noto risultato.

Sia $\overline{Y} = (1/n)\sum_i Y_i$ la media campionaria e $S^2 = \sum_i (Y_i - \overline{Y})^2/(n-1)$ la varianza campionaria corretta. Si cerca una quantità pivotale che contenga μ e che dipenda dal campione (non deve contenere alte quantità ignote e quindi non possiamo usare la statistica vista al punto precedente che conteneva nella sua espressione σ). Il rapporto

$$\frac{\sqrt{n}(\overline{Y} - \mu)}{\sqrt{S^2}}$$

che, come noto (si veda il paragrafo 2.4), ha una distribuzione nota essendo una t di Student con $n-1$ gradi libertà, e indipendente da μ ed è pertanto una funzione pivot. Si può scrivere

$$\frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

dove $S = +\sqrt{S^2}$. Se $t_{n-1,\alpha}$ è il quantile di ordine α della distribuzione t con $n-1$ gradi di libertà, allora

$$P\left(t_{n-1,\frac{\alpha}{2}} \leq \frac{\overline{Y} - \mu}{S/\sqrt{n}} \leq t_{n-1,1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

da cui, moltiplicando per $S/\sqrt{n}$,

$$P\left(t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \leq \overline{Y} - \mu \leq t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

quindi, moltiplicando per -1 e sommando $\overline{Y}$,

$$P\left(\overline{Y} - t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \geq \mu \geq \overline{Y} - t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1-\alpha$$

o equivalentemente, essendo $t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2}$, risulta

$$\begin{aligned} &P\left(\overline{Y} - t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{Y} - t_{n-1,\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) \\ &= P\left(\overline{Y} - t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{Y} + t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right) = 1-\alpha. \end{aligned}$$

Pertanto in intervallo di confidenza per μ al livello $1-\alpha$ è

$$\left(\overline{Y} - t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \overline{Y} + t_{n-1,1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right). \tag{5.2}$$

Si noti che, ancora una volta, si tratta di uno dei possibili intervalli che si possono costruire scegliendo due valori qualsiasi t_1 e t_2 della distribuzione t con $n-1$ gradi libertà per cui risulta soddisfatta l'uguaglianza

$$P\left(t_1 \leq \frac{\overline{Y}-\mu}{S/\sqrt{n}} \leq t_2\right) = 1-\alpha;$$

l'intervallo che ne deriva ha ampiezza $(t_2 - t_1)S/\sqrt{n}$ che, per ogni campione dato, assume valore minimo quando t_2 e t_1 sono scelti simmetricamente rispetto a 0, cioè quando $t_1 = -t_2$. L'intervallo fornito dalla (5.2) è, quindi, l'intervallo di ampiezza minima e pari a

$$A = 2t_{n-1,1-\alpha/2}S/\sqrt{n}.$$

Esempio 5.2. Si riconsideri l'esempio 5.1 ma immaginiamo ora che la varianza non sia nota per cui viene calcolata la varianza campionaria che risulta $s^2 = 36.49$ cm^2. Sulla base di questi dati si trova $s/\sqrt{30} = 1.103$ e dalle tavole della distribuzione t con 29 gradi libertà otteniamo $t_{29,0.975} = 2.0452$. Gli estremi dell'intervallo di confidenza per μ al livello del 95% sono dati da

$$163.87 \pm 2.0452(1.103)$$

cioè $(161.614, 166.126)$.

▲

Intervallo di confidenza per la varianza di una gaussiana

Si vuole ora costruire un intervallo di confidenza per la varianza σ^2 del modello gaussiano di riferimento, assumendo incognito il valore di μ. Occorre individuare una funzione *pivot* che possa essere 'trasformata' in una affermazione di probabilità su σ^2. Poiché campioniamo da una popolazione normale, sappiamo che lo stimatore S^2 di σ^2 ha distribuzione di probabilità nota e, in particolare,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

per cui possiamo considerare $Q = (n-1)S^2/\sigma^2$ come quantità pivotale e per un fissato α possiamo scrivere

$$1-\alpha = P\left(q_1 \leq Q \leq q_2\right) = P\left(\frac{1}{q_1} \geq \frac{\sigma^2}{(n-1)S^2} \geq \frac{1}{q_2}\right)$$

$$= P\left(\frac{(n-1)S^2}{q_2} \leq \sigma^2 \leq \frac{(n-1)S^2}{q_1}\right)$$

così

$$\left(\frac{(n-1)S^2}{q_2}, \frac{(n-1)S^2}{q_1}\right)$$

è un intervallo di confidenza al $(1-\alpha)100\%$ per σ^2, dove q_1 e q_2 sono tali che $P(q_1 \leq Q \leq q_2) = 1-\alpha$. Possiamo scegliere tali valori in modo che la probabilità 'nelle code' sia la stessa, cioè consideriamo i quantili della distribuzione chi-quadrato con $n-1$ gradi di libertà

$$q_2 = \chi^2_{n-1,1-\alpha/2}, \qquad q_1 = \chi^2_{n-1,\alpha/2}$$

tali che

$$P(Q \leq q_1) = P(Q \geq q_2) = \alpha/2$$

(q_1 e q_2 si ricavano dalle tavole della distribuzione chi-quadrato in corrispondenza di $n-1$ g.d.l.). Si perviene quindi all'intervallo di confidenza al $(1-\alpha)100\%$ per σ^2 dato da

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right) \tag{5.3}$$

che, diversamente dai casi precedenti, non ha la proprietà di essere di ampiezza minima. Osserviamo infine che per ogni q_1 e q_2 tali che $P(q_1 \leq Q \leq q_2) = 1-\alpha$,

$$\left(\sqrt{\frac{(n-1)S^2}{q_2}}, \sqrt{\frac{(n-1)S^2}{q_1}}\right)$$

è un intervallo di confidenza al livello $1-\alpha$ per σ.

Esempio 5.3. Si supponga di aver effettuato $n = 10$ misurazioni con un certo strumento e si assume che tali misure si distribuiscano seguendo una normale di media μ e varianza incognita σ^2. Il campione ha fornito i seguenti valori:

$$4.7,\ 5.5,\ 4.4,\ 3.3,\ 4.6,\ 5.3,\ 5.2,\ 4.8,\ 5.7,\ 5.3.$$

Sulla base dei dati osservati calcoliamo la varianza campionaria $s^2 = 0.484$. Posto $\alpha = 0.05$, dalle tavole si trova $\chi_{9,0.025} = 2.700$, $\chi_{9,0.975} = 19.023$, essendo 9 i gradi di libertà. Ne segue che il valore dell'intervallo di confidenza per σ^2 al livello $1-\alpha = 0.95$ è dato da

$$\left(\frac{9(0.484)}{19.023}, \frac{9(0.484)}{2.700}\right) = (0.229, 1.613).$$

▲

Esempio 5.4. Sia $(Y_1, Y_2, \ldots, Y_n)$ un campione casuale proveniente da una popolazione normale di media μ e varianza incognita σ^2. Si vuole costruire un intervallo di confidenza al 95% per σ^2, supponiamo però ora che sia nota la media μ. Il rapporto $\sum_{i=1}^n (Y_i-\mu)^2/\sigma^2$ ha distribuzione chi-quadrato con n gradi di libertà. Ne segue che si può scrivere

$$P\left(\chi^2_{n,\alpha/2} \leq \frac{\sum_{i=1}^n (Y_i-\mu)^2}{\sigma^2} \leq \chi^2_{n,1-\alpha/2}\right) = 1-\alpha$$

Risolvendo le disuguaglianze rispetto a σ^2, si ottiene

$$P\left(\frac{\sum_{i=1}^n (Y_i-\mu)^2}{\chi^2_{n,1-\alpha/2}} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (Y_i-\mu)^2}{\chi^2_{n,\alpha/2}}\right) = 1-\alpha$$

da cui l'intervallo cercato (ponendo $\alpha = 0.05$)

$$\left(\frac{\sum_{i=1}^n (Y_i-\mu)^2}{\chi^2_{n,0.975}}, \frac{\sum_{i=1}^n (Y_i-\mu)^2}{\chi^2_{n,0.025}}\right).$$

▲

5.3 Intervalli di confidenza approssimati

Determinare una funzione pivot esatta può risultare difficile, talvolta impossibile. In molti casi, è possibile fare ricorso a quantità pivotali approssimate, che conducono quindi a intervalli di confidenza approssimati. Di solito, si fa riferimento alla distribuzione asintotica delle quantità pivot per cui tali approssimazioni sono tanto più accettabili quanto maggiore è la dimensione campionaria. In particolare, va osservato che se si dispone di uno stimatore consistente $\hat{\theta}_n$ per il parametro di interesse θ per il quale valga

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \mathcal{V}),$$

allora

$$Z = \frac{\hat{\theta}_n - \theta}{\sqrt{\frac{\mathcal{V}}{n}}} \dot{\sim} \mathcal{N}(0,1)$$

è una quantità pivot approssimata che consente di ottenere facilmente un intervallo di confidenza per θ. Si noti che $\sqrt{\frac{\mathcal{V}}{n}}$ è l'errore standard (asintotico) di $\hat{\theta}_n$ che indicheremo con $\text{se}(\hat{\theta}_n)$.
Infatti dall'espressione

$$P\left(-z_{1-\alpha/2} \leq \frac{\hat{\theta}_n - \theta}{\text{se}(\hat{\theta}_n)} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

isolando θ al centro delle disequazioni, si ottiene

$$P\left(\hat{\theta}_n - z_{1-\alpha/2}\,\text{se}(\hat{\theta}_n) \leq \theta \leq \hat{\theta}_n + z_{1-\alpha/2}\,\text{se}(\hat{\theta}_n)\right) = 1 - \alpha$$

Quindi $[\hat{\theta}_n - z_{1-\alpha/2}\,\text{se}(\hat{\theta}_n), \hat{\theta}_n + z_{1-\alpha/2}\,\text{se}(\hat{\theta}_n)]$ è un intervallo di confidenza approssimato al livello $1 - \alpha$. Si noti che spesso il problema per l'utilizzo di questo risultato è la conoscenza dell'errore standard dello stimatore che è ignoto. Si può ricorrere in tal caso a una stima consistente dell'errore standard e il risultato asintotico resta ancora valido.

La qualità di un intervallo di confidenza approssimato dipende, come già segnalato, dalla qualità dell'approssimazione asintotica e in genere per una dimensione del campione finita la reale probabilità di copertura (livello di confidenza) differisce dalla copertura nominale.

Si presentano di seguito alcuni importanti esemplificazioni in cui si sfrutta la conoscenza della distribuzione asintotica della funzione pivot approssimata.

5.3.1 Intervallo di confidenza per una proporzione

Se $(Y_1, Y_2, \ldots, Y_n)$ è un campione casuale proveniente da una popolazione bernoulliana di parametro p e sia $\overline{Y}$ la proporzione campionaria, allora, ricordando l'esempio 4.7,

$$\frac{\sqrt{n}(\overline{Y} - p)}{\sqrt{p(1-p)}}$$

ha distribuzione limite normale $\mathcal{N}(0,1)$ ed quindi è un pivot approssimato. Uno stimatore consistente dell'errore standard $\sqrt{p(1-p)/n}$ è semplicemente $\sqrt{\overline{Y}(1-\overline{Y})/n}$ e quindi si può scrivere, per n sufficientemente grande,

$$P\left(-z_{1-\alpha/2} \leq \frac{\overline{Y} - p}{\sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}}} \leq z_{1-\alpha/2}\right) \approx 1 - \alpha$$

da cui si ottiene l'intervallo di confidenza approssimato al $100(1-\alpha)\%$ per p:

$$\left(\overline{Y} - z_{1-\alpha/2}\sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}}, \overline{Y} + z_{1-\alpha/2}\sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}}\right)$$

Ad esempio, per un campione osservato $y_1, y_2, \ldots, y_n$, da $Y \sim Be(p)$, la stima per intervallo per p al 95% (quando n è sufficientemente elevato) è

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right),$$

dove $\hat{p}$ è la proporzione campionaria.

Resta da capire quanto grande debba essere n per poter avere intervalli di confidenza con il livello di copertura effettivo vicino a quello nominale. Valgono in questo caso, in larga misura, le considerazioni fatte più volte per giudicare l'approssimazione normale-binomiale. Tuttavia il valore di p non è noto per cui una valutazione prudenziale fa ritenere di adottare l'approssimazione se n non è inferiore a qualche centinaio.

Esempio 5.5. Viene effettuato un sondaggio per stimare la percentuale di famiglie nella popolazione in una specifica regione che parteciperanno a un programma di riciclaggio proposto separando i loro rifiuti in vari componenti.

I sondaggisti hanno deciso di considerare un campione di $n = 1000$ dalla popolazione di circa 1.5 milioni di famiglie. Ogni intervistato risponderà sì o no a una domanda riguardante la propria partecipazione. Possiamo quindi presumere di campionare da un modello di Bernoulli dove $p \in [0, 1]$ è la proporzione di individui nella popolazione che risponderanno sì. Dopo aver condotto il sondaggio, si registrano 790 intervistati che hanno risposto sì e 210 che hanno risposto no. Sulla base di questi dati si ha

$$\hat{p} = \bar{y} = \frac{790}{1000} = 0.79.$$

e inoltre

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.79(1-0.79)}{1000}} = 0.01288$$

Quindi un intervallo di confidenza approssimato per p al 95% è dato da

$$(0.79 - 1.96(0.01288), 0.79 + 1.96(0.01288)) = (0.76475, 0.81525)$$

Si ribadisce che si tratta di un intervallo approssimato per p che richiede che la numerosità campionaria n sia sufficientemente elevata. L'accuratezza di tale approssimazione dipende quindi da n e, in particolare, se il vero p fosse vicino a 0 o 1, allora occorrerebbe considerare una ampiezza piuttosto elevata.

▲

5.3.2 Intervallo di confidenza per il valore atteso

Si consideri il caso in cui la variabile di interesse Y ha media $M(Y) = \mu$ e varianza $V(Y) = \sigma^2$ senza però ulteriori specificazioni riguardo la sua forma distributiva. Si immagini di voler ottener un intervallo di confidenza per μ. In tal caso, ricordando il teorema del limite centrale, la statistica $\overline{Y}$ ha distribuzione, che per n grande può essere ben approssimata da una distribuzione gaussiana e in particolare la quantità

$$Z = \frac{\sqrt{n}(\overline{Y} - \mu)}{\sigma} \dot{\sim} \mathcal{N}(0, 1).$$

Tale quantità costituisce quindi un pivot approssimato. Se σ^2 non fosse noto si potrebbe sostituire a σ^2 una sua stima consistente, ad esempio S^2. Allora

per n grande

$$Z = \frac{\sqrt{n}(\overline{Y} - \mu)}{\sqrt{S^2}} \dot{\sim} \mathcal{N}(0, 1).$$

e quindi anche questa quantità può essere impiegata come pivot approssimato al fine di costruire un intervallo di confidenza. In particolare, fissato il livello di confidenza $1 - \alpha$, si può scrivere

$$P\left(z_{\frac{\alpha}{2}} \leq \frac{\overline{Y} - \mu}{\frac{S}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}}\right) \approx 1 - \alpha$$

da cui, moltiplicando per $S/\sqrt{n}$,

$$P\left(z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \leq \overline{Y} - \mu \leq z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right) \approx 1 - \alpha$$

dove $z_{1-\frac{\alpha}{2}} = \Phi^{-1}(1 - \frac{\alpha}{2})$ e $\Phi(z)$ è la funzione di ripartizione della normale standard. Si ottiene pertanto l'intervallo di confidenza approssimato

$$\left(\overline{Y} - z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \overline{Y} + z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right).$$

Esempio 5.6. Si immagini che Y sia il reddito familiare mensile in una ampia popolazione quale potrebbe essere quella delle famiglie italiane. Si è interessati al reddito medio di tale popolazione e si ipotizzi che essa, pur essendo finita, sia sufficientemente ampia da poter considerare questo come un problema di inferenza da una popolazione infinita. Si immagini che su un campione di 200 famiglie si sia calcolato il reddito medio campionario pari a 2730 euro e che sia pari a 8053015 la media dei quadrati dei dati campionari. L'intervallo di confidenza approssimato al livello 0.99 è

$$\left(\overline{Y} - z_{0.995} \frac{S}{\sqrt{n}}, \overline{Y} + z_{0.995} \frac{S}{\sqrt{n}}\right)$$

in quanto è $\alpha/2 = .005$. Ora, $z_{0.995} \approx 2.58$ e la radice della varianza campionaria è

$$S = \sqrt{(8053015 - 2730^2)\frac{200}{199}} = \sqrt{603130.7} \approx 776.61$$

quindi si ottiene l'intervallo di confidenza approssimato

$$\left(2730 - 2.58\frac{776.61}{\sqrt{200}}, 2730 + 2.58\frac{776.61}{\sqrt{200}}\right) = (2675.1, 2784.9)$$

Si noti che la quantità pivotale è la stessa ottenuta nel caso si assumesse per Y la distribuzione gaussiana (assunzione che tuttavia sarebbe stata assai discutibile poiché la distribuzione dei redditi ha una marcata asimmetria positiva). La differenza è che in quel caso si ha la distribuzione esatta della quantità pivot e i percentili andrebbero tratti dalla t di Student con 199 gradi di libertà. Avremmo ottenuto comunque un intervallo di confidenza molto simile in quanto una t di Student con 199 gradi di libertà è quasi equivalente a una gaussiana infatti $t_{199,0.975} \approx 2,6$. Si ricordi infatti che le usuali tavole della t di Student per valori dei gradi di libertà superiori a 100 non vengono riportate potendosi usare l'approssimazione alla gaussiana standard.

La numerosità campionaria pari a 200 i questo caso è sufficiente a garantire che l'intervallo di confidenza abbia un grado di copertura effettivo molto prossimo a quello nominale. Con campioni più piccoli l'approssimazione potrebbe essere troppo grossolana.

▲

5.3.3 Intervalli di confidenza approssimati derivati da stimatori di massima verosimiglianza

È possibile fare riferimento a una funzione pivot approssimata nel caso in cui si sia ottenuto per il parametro di interesse la stima di massima verosimiglianza.

Si ricorda infatti che per lo stimatore di massima verosimiglianza $\hat{\theta}_n$ del parametro θ di una densità $f(\cdot;\theta)$ vale, nel caso di campione casuale, il seguente risultato

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{i(\theta)}\right)$$

per cui

$$\hat{\theta}_n \dot{\sim} \mathcal{N}(\theta, (ni(\theta))^{-1})$$

la varianza dello stimatore $\hat{\theta}_n$ è quindi approssimata da

$$\frac{1}{I(\theta)} = \frac{1}{ni(\theta)} = \frac{1}{-nE\left(\frac{d^2}{d\theta^2}\log f(y;\theta)\right)} = \frac{1}{nE\left\{\left(\frac{d}{d\theta}\log f(y;\theta)\right)^2\right\}}$$

dove $I(\theta)$ è l'informazione attesa di Fisher.

Ne segue che $(\hat{\theta}_n - \theta)/\sqrt{(ni(\theta))^{-1}}$ può essere utilizzata come una quantità pivotale approssimata nella determinazione di un intervallo di confidenza approssimato per θ al livello $1-\alpha$, invertendo le disuguaglianze

$$-z_{1-\alpha/2} \leq \frac{\hat{\theta}_n - \theta}{\sqrt{(ni(\theta))^{-1}}} \leq z_{1-\alpha/2}$$

dove $z_{\alpha/2} = -z_{1-\alpha/2}$ si ricava dalle tavole come il valore tale che $\Phi(z_{1-\alpha/2}) = 1-\alpha/2$.

Nella pratica alla quantità di informazione attesa $ni(\theta)$ si può sostituire la quantità di informazione osservata, introdotta in 4.4.1, cioè la derivata seconda della log-verosimiglianza col segno negativo calcolata in $\hat{\theta}$. Il valore $(ni(\hat{\theta}))^{-1}$ è una stima consistente della varianza asintotica e il grado di copertura dell'intervallo, se n è grande, non si discosta molto da $1-\alpha$, pertanto:

$$P\left(-z_{1-\alpha/2} \leq \frac{\hat{\theta}_n - \theta}{\sqrt{(ni(\hat{\theta}_n))^{-1}}} \leq z_{1-\alpha/2}\right) \approx 1-\alpha$$

e invertendo l'espressione si ottiene

$$P\left(\hat{\theta}_n - z_{1-\frac{\alpha}{2}}\sqrt{(ni(\hat{\theta}_n))^{-1}} \leq \theta \leq \hat{\theta}_n + z_{1-\frac{\alpha}{2}}\sqrt{(ni(\hat{\theta}_n))^{-1}}\right) \approx 1-\alpha.$$

Quindi, fissato un livello di confidenza pari a $\gamma = 1-\alpha$, si arriva facilmente ad ottenere un intervallo approssimato per θ:

$$\left(\hat{\theta}_n - z_{1-\frac{\alpha}{2}}\sqrt{(ni(\hat{\theta}_n))^{-1}},\ \hat{\theta}_n + z_{1-\frac{\alpha}{2}}\sqrt{(ni(\hat{\theta}_n))^{-1}}\right). \tag{5.4}$$

Esempio 5.7. Sia $(Y_1, Y_2, \ldots, Y_n)$ un campione casuale estratto dalla densità $f(y;\lambda) = \lambda e^{-\lambda y}$, $x > 0$. È agevole verificare che lo stimatore di massima verosimiglianza $\hat{\lambda}_{MV}$ di λ è $1/\overline{Y}$ ed può essere approssimato con una distribuzione asintotica normale con media λ e varianza

$$\frac{1}{nE\left\{\left(\frac{d}{d\lambda}\log(\lambda e^{-\lambda Y})\right)^2\right\}} = \frac{1}{nE\left\{\left(\frac{1}{\lambda} - Y\right)^2\right\}} = \frac{1}{nV(Y)} = \frac{\lambda^2}{n}$$

Perciò, fissato $\gamma = 0.98$, possiamo scrivere

$$\begin{aligned}
0.98 &\approx P\left(-z_{0.99} \leq \frac{1/\overline{Y} - \lambda}{\sqrt{\lambda^2/n}} \leq z_{0.99}\right) \\
&= P\left(\frac{-z_{0.99}\lambda}{\sqrt{n}} \leq \frac{1}{\overline{Y}} - \lambda \leq \frac{z_{0.99}\lambda}{\sqrt{n}}\right) \\
&= P\left(\frac{-z_{0.99}}{\sqrt{n}} \leq \frac{1}{\overline{Y}\lambda} - 1 \leq \frac{z_{0.99}}{\sqrt{n}}\right) \\
&= P\left(\frac{1}{1 + z_{0.99}/\sqrt{n}} \leq \overline{Y}\lambda \leq \frac{1}{1 - z_{0.99}/\sqrt{n}}\right) \\
&= P\left(\frac{1/\overline{Y}}{1 + z_{0.99}/\sqrt{n}} \leq \lambda \leq \frac{1/\overline{Y}}{1 - z_{0.99}/\sqrt{n}}\right)
\end{aligned}$$

e quindi

$$\left(\frac{1/\overline{Y}}{1 + z_{0.99}/\sqrt{n}}, \frac{1/\overline{Y}}{1 - z_{0.99}/\sqrt{n}}\right)$$

è un intervallo di confidenza per λ per grandi campioni con un livello di confidenza approssimativamente uguale a 0.98, dove $z_{0.99} \approx 2.326$.

▲

5.3.4 Intervallo di confidenza approssimato per una funzione del parametro

Si ricordi quanto già introdotto al paragrafo 4.4.4 riguardo la distribuzione asintotica di funzioni degli stimatori di massima verosimiglianza. Si supponga di disporre di un campione casuale piuttosto ampio di valori tratti da Y con distribuzione $f(y;\theta)$, e che l'interesse sia rivolto alla funzione biettiva $\phi = h(\theta)$ del parametro. Sia $\hat{\phi}_{MV} = h(\hat{\theta}_{MV})$ lo stimatore di massima verosimiglianza di ϕ. In virtù della proprietà di invarianza e utilizzando il metodo delta per approssimare la varianza asintotica, si deduce che la distribuzione approssimata di $\hat{\phi}_{MV}$ è una gaussiana di media ϕ e varianza pari a $V(h(\hat{\theta}_{MV})) = V(\hat{\theta}_{MV})\left(\frac{d}{d\theta}h(\theta)\right)^2$. Allora la quantità

$$Z = \frac{h(\hat{\theta}_{MV}) - h(\theta)}{\sqrt{V(\hat{\theta}_{MV})\left(\frac{d}{d\theta}h(\theta)\right)^2}} \dot{\sim} \mathcal{N}(0,1)$$

La quantità al denominatore può essere riscritta come

$$V(\hat{\theta}_{MV})\left(\frac{d}{d\theta}h(\theta)\right)^2 = I(\theta)^{-1}\left(\frac{d}{d\theta}h(\theta)\right)^2$$

ed essere stimata consistentemente sostituendo a θ la sua stima di massima verosimiglianza $\hat{\theta}_{MV}$.

Esempio 5.8. Si consideri il caso in cui si disponga di un campione casuale di dimensione n da una $Y \sim Be(p)$ e si sia interessati a costruire un intervallo di confidenza per il logit di p, cioè $\phi = \log\left(\frac{p}{1-p}\right) = \text{logit}(p)$. Riprendendo i risultati del paragrafo 4.4.4, si ha

$$\hat{\phi}_{MV} = \log\left(\frac{\hat{p}_{MV}}{1-\hat{p}_{MV}}\right)$$

ove $\hat{p}_{MV}$ è la proporzione campionaria $\bar{y}$. La varianza, per n grande puèssere approssimata con

$$\begin{aligned} V(\hat{\phi}_{MV}) &= V(\hat{p}_{MV})\left(\frac{d}{dp}\log\left(\frac{p}{1-p}\right)\right)^2 \\ &= \frac{p(1-p)}{n}\frac{1}{p(1-p)^2} \\ &= \frac{1}{np(1-p)} \end{aligned} \tag{5.5}$$

quindi si ha il seguente pivot approssimato

$$Z = \frac{\hat{\phi}_{MV} - \phi}{\sqrt{V(\hat{\phi}_{MV})}} = \frac{\log\left(\frac{\hat{p}_{MV}}{1-\hat{p}_{MV}}\right) - \log\left(\frac{p}{1-p}\right)}{\sqrt{\frac{1}{np(1-p)}}} \dot{\sim} \mathcal{N}(0,1)$$

La stima di massima verosimiglianza $\hat{p}_{MV}$ di p è la proporzione campionaria $\bar{y} = \hat{p}$ per cui possiamo sostituire questa nella espressione della 5.5, ottenendo il seguente intervallo di confidenza approssimato per ϕ al livello $(1-\alpha)100\%$

$$\left(\log\left(\frac{\hat{p}}{1-\hat{p}}\right) - z_{1-\alpha/2}\sqrt{\frac{1}{n\hat{p}(1-\hat{p})}}\ ,\ \log\left(\frac{\hat{p}}{1-\hat{p}}\right) + z_{1-\alpha/2}\sqrt{\frac{1}{n\hat{p}(1-\hat{p})}}\right).$$

▲

5.4 Intervalli di confidenza e ampiezza del campione

Molto spesso ci si pone il problema di determinare la dimensione del campione n per garantire che con una probabilità molto alta i risultati di un'analisi statistica forniscano stime accurate. Ad esempio, supponiamo di prendere un campione di dimensione n da una popolazione e di voler stimare la media della popolazione in modo che la stima sia entro 0.5 della media reale con probabilità almeno 0.95. Ciò significa che vogliamo che la semi lunghezza, o margine di errore, dell'intervallo di confidenza al 95% per la media sia inferiore a 0.5. Consideriamo tali problemi nei seguenti esempi. Innanzitutto, si considera il problema di selezionare la dimensione del campione per garantire che un intervallo di confidenza sia più breve di un valore prescritto.

Esempio 5.9. Si consideri la situazione illustrata in precedenza in cui il campionamento è fatto da una popolazione normale $\mathcal{N}(\mu, \sigma^2)$, dove $\mu \in \mathbb{R}$ è incognita e $\sigma^2 > 0$ è nota. L'ampiezza dell'intervallo di confidenza per μ al livello $100(1-\alpha)\%$ è

$$A = 2z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},$$

perciò volendo determinare n tale che il *margine di errore* non sia maggiore di un certo valore $\delta > 0$, occorre porre

$$\frac{A}{2} = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \delta,$$

o equivalentemente

$$n \geq \sigma^2 \left(\frac{z_{1-\frac{\alpha}{2}}}{\delta} \right)^2.$$

Supponiamo, ad esempio, che nella popolazione $\sigma^2 = 10$, e fissiamo $\gamma = 1 - \alpha = 0.95$, $\delta = 0.5$, allora

$$n \geq 10 \left(\frac{1.96}{0.5} \right)^2 = 153.664$$

e dunque il più piccolo valore per n è $n = 154$.

Se invece non si conosce σ^2, in tal caso la disuguaglianza di interesse è

$$t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \delta,$$

che implica

$$n \geq s^2 \left(\frac{t_{n-1,1-\frac{\alpha}{2}}}{\delta} \right)^2,$$

da cui si evince che il calcolo richiederebbe il valore s, che però non conosciamo. Tuttavia, in molte circostanze possiamo almeno individuare un limite superiore a σ, cioè un valore b tale che $\sigma \leq b$. Ad esempio, se consideriamo un campione di altezze di individui espresse in centimetri, sotto l'assunzione di normalità della popolazione, l'intervallo $(\mu - 3\sigma, \mu + 3\sigma)$ sarà contenuto tra due valori che possiamo individuare come i limiti inferiore e superiore per i valori dell'altezza. Dividendo l'ampiezza per 6, possiamo ottenere un valore plausibile del limite superiore b per il valore di σ, e ci aspettiamo che $s \leq b$. Quindi determiniamo n tale che

$$n \geq b^2 \left(\frac{t_{n-1,1-\frac{\alpha}{2}}}{\delta} \right)^2,$$

dove si osservi che occorre valutare anche $t_{n-1,1-\frac{\alpha}{2}}$ per ogni n, e sarà quindi opportuno considerare scelte conservative dei possibili valori per n, cioè non scegliere il valore più piccolo possibile.

▲

Esempio 5.10. Si consideri il caso di un campione casuale $(Y_1, Y_2, \ldots, Y_n)$ tratto da una popolazione bernoulliana di parametro incognito p. Per quanto visto, l'intervallo nella (5.3.1) rappresenta un intervallo approssimato per p al livello $1-\alpha$. Si vuole determinare la dimensione del campione n tale che il margine di errore (la semi ampiezza dell'intervallo) sia non superiore a $\delta > 0$, dunque si richiede che

$$z_{1-\alpha/2} \sqrt{\frac{\bar{y}(1-\bar{y})}{n}} \leq \delta, \tag{5.6}$$

cioè

$$n \geq \bar{y}(1-\bar{y}) \left(\frac{z_{1-\alpha/2}}{\delta} \right)^2.$$

Per ovviare al fatto che la disuguaglianza dipende da $\bar{y}$, si osservi che $0 \leq \bar{y}(1-\bar{y}) \leq 1/4$ per ogni $\bar{y}$, dove si raggiunge il valore $1/4$ quando $\bar{y} = \hat{p} = 0.5$. Pertanto se determiniamo n tale che

$$n \geq \frac{1}{4} \left(\frac{z_{1-\alpha/2}}{\delta} \right)^2,$$

allora (5.6) sarà soddisfatta. Ad esempio, se $\gamma = 0.95$, allora $\alpha = 0.05$, e fissando $\delta = 0.1$ si pone

$$n \geq \frac{1}{4}\left(\frac{1.96}{0.1}\right)^2 = 96.04$$

da cui deriva che n deve essere almeno pari a 97.

▲

5.5 Ulteriori esempi: intervalli di confidenza per due popolazioni

Si riportano di seguito alcuni risultati classici che riguardano il caso in cui la quantità di interesse riguardi i parametri di due popolazioni. Ad esempio, spesso si vuole fare inferenza sulla differenza fra i valori medi di due popolazioni come nell'esempio introdotto in 1.4.6 ove la quantità che interessa è la differenza fra la media per i due gruppi (quelli del gruppo trattato con il farmaco e quello di controllo).

Stima della differenza fra le medie di popolazioni normali con varianze note

Se assumiamo che le varianze delle due popolazioni σ_1^2 e σ_2^2 siano note, è ragionevole basare la costruzione dell'intervallo di confidenza per $\mu_1 - \mu_2$ sulla statistica $\overline{X} - \overline{Y}$ di media $\mu_1 - \mu_2$ e varianza $\sigma_1^2/n + \sigma_2^2/m$, assumendo come prima di avere due campioni casuali indipendenti. Allora si ha

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \sim \mathcal{N}(0,1) \tag{5.7}$$

e posto $\sigma^2_{\bar{X}-\bar{Y}} = \sigma_1^2/n + \sigma_2^2/m$, si può scrivere

$$P\left(z_{\alpha/2} \leq \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma_{\bar{X}-\bar{Y}}} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

da cui, con il solito procedimento, si ricava l'intervallo di confidenza per $\mu_1 - \mu_2$ al $(1-\alpha)100\%$

$$\left(\overline{X} - \overline{Y} - z_{1-\frac{\alpha}{2}}\sigma_{\bar{X}-\bar{Y}},\ \overline{X} - \overline{Y} + z_{1-\frac{\alpha}{2}}\sigma_{\bar{X}-\bar{Y}}\right). \tag{5.8}$$

In questo paragrafo consideriamo due campioni indipendenti: $(X_1, \ldots, X_n)$ estratto da una distribuzione $\mathcal{N}(\mu_1, \sigma_1^2)$ e $(Y_1, \ldots, Y_m)$ estratto da una distribuzione $\mathcal{N}(\mu_2, \sigma_2^2)$. Nel seguito costruiamo un intervallo di confidenza per la **differenza tra le medie** delle due popolazioni $\mu_1 - \mu_2$, supponendo che le varianze σ_1^2 e σ_2^2 non siano note.

Popolazioni normali con varianze ignote e uguali

Ipotizziamo che le varianze delle popolazioni siano uguali, cioè $\sigma_1^2 = \sigma_2^2$ e denotiamo con σ^2 la varianza comune. La variabile aleatoria $\overline{X} - \overline{Y}$ ha quindi distribuzione normale con media $\mu_1 - \mu_2$ e varianza $\sigma^2/n + \sigma^2/m$. Inoltre $\sum_i (X_i - \overline{X})^2/\sigma^2$ ha distribuzione chi-quadrato con $n-1$ gradi di libertà e $\sum_i (Y_i - \overline{Y})^2/\sigma^2$ ha distribuzione chi-quadrato con $m-1$ gradi di libertà. Quindi

$$\frac{\sum_i (X_i - \overline{X})^2}{\sigma^2} + \frac{\sum_i (Y_i - \overline{Y})^2}{\sigma^2}$$

ha distribuzione chi-quadrato con $m + n - 2$ gradi di libertà. Infine,

$$\begin{aligned} Q =& \frac{(\overline{X} - \overline{Y} - (\mu_1 - \mu_2))/(\sqrt{\sigma^2/n + \sigma^2/m})}{\sqrt{(\sum_i (X_i - \overline{X})^2/\sigma^2 + \sum_i (Y_i - \overline{Y})^2/\sigma^2)/(m+n-2)}} \\ =& \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{(1/n + 1/m)\left(\sum_i (X_i - \overline{X})^2 + \sum_i (Y_i - \overline{Y})^2\right)/(m+n-2)}} \\ =& \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{(1/n + 1/m) S_p^2}} \end{aligned}$$

ha una distribuzione t con $m+n-2$ gradi di libertà (si veda il Teorema 2.7), essendo il rapporto tra una normale standard e una chi-quadrato divisa per i suoi gradi di libertà (che sono $m + n - 2$), dove si è posto

$$S_p^2 = \frac{1}{n+m-2}\left(\sum_i (X_i - \overline{X})^2 + \sum_i (Y_i - \overline{Y})^2\right).$$

S_p^2 è uno stimatore non distorto della varianza comune σ^2 (anche detto stimatore *pooled* o combinato della varianza), definito dalla media ponderata degli stimatori varianza campionaria S_1^2 e S_2^2 di σ^2 per le due popolazioni che utilizza tutti gli $n+m$ dati (nell'ipotesi di omoschedasticità):

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{m+n-2}.$$

Adottando la quantità

$$Q = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{(1/n + 1/m)}} \sim t_{n+m-2}$$

come funzione *pivot* e fissato un livello $1-\alpha = \gamma \in (0,1)$ possiamo scrivere

$$P(-t_{1-\alpha/2} \leq Q \leq t_{1-\alpha/2}) = 1-\alpha,$$

dove $t_{1-\alpha/2}$ è il quantile $(1-\alpha/2)$-esimo della distribuzione t con $m+n-2$ gradi di libertà (si ricordi che vale $-t_{1-\alpha/2} = t_{\alpha/2}$). Dall'espressione precedente si ricava che la probabilità dell'evento

$$(\overline{X} - \overline{Y}) - t_{1-\frac{\alpha}{2}} S_p\sqrt{\frac{1}{n} + \frac{1}{m}} \leq \mu_1 - \mu_2 \leq (\overline{X} - \overline{Y}) + t_{1-\frac{\alpha}{2}} S_p\sqrt{\frac{1}{n} + \frac{1}{m}}$$

è $1-\alpha$. Allora, l'intervallo di confidenza per $\mu_1 - \mu_2$ con grado di copertura $1-\alpha$ ha estremi

$$\left(\overline{X} - \overline{Y} - t_{1-\frac{\alpha}{2}}\, S_p\sqrt{\frac{1}{n} + \frac{1}{m}},\ \overline{X} - \overline{Y} + t_{1-\frac{\alpha}{2}}\, S_p\sqrt{\frac{1}{n} + \frac{1}{m}}\right). \tag{5.9}$$

Intervalli di confidenza per la differenza fra le media per grandi campioni

Nel caso le dimensioni campionarie siano elevate è possibile determinare gli stimatori intervallari validi a prescindere dal modello distributivo delle popolazioni e dalle proprietà delle varianze, in virtù dell'approssimazione fornita dalla distribuzione normale.

Se i campioni $(X_1, \ldots, X_n)$ e $(Y_1, \ldots, Y_m)$ sono indipendenti e hanno entrambi dimensione sufficientemente grande, il rapporto

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n + S_2^2/m}}$$

ottenuto dalla (5.7) sostituendo a σ_1^2 e σ_2^2 i rispettivi stimatori S_1^2 ed S_2^2, ha distribuzione che può essere approssimata con una normale $\mathcal{N}(0,1)$. Ne segue che vale la relazione

$$P\left(z_{\alpha/2} \leq \frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{S_1^2/n+S_2^2/m}} \leq z_{1-\alpha/2}\right) = 1-\alpha$$

da cui è facile pervenire all'intervallo con livello di confidenza $1-\alpha$ per grandi campioni per $\mu_1-\mu_2$:

$$\left((\overline{X}-\overline{Y}) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}},\ (\overline{X}-\overline{Y}) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}}\right). \tag{5.10}$$

5.5.1 Dati appaiati

Abbiamo assunto prima di avere due campioni casuali indipendenti. Supponiamo ora che non sia valida l'assunzione di indipendenza e di disporre di un campione casuale $(X_1,Y_1),\ldots,(X_n,Y_n)$ estratto dalla distribuzione normale bivariata di parametri $\mu_1 = E(X)$, $\mu_2 = E(Y)$, $\sigma_1^2 = V(X)$, $\sigma_2^2 = V(Y)$, e $\rho = \mathrm{Cov}(X,Y)/\sigma_1\sigma_2$.

Sia $D_i = Y_i - X_i$, $i = 1,\ldots,n$, allora $D_1,\ldots,D_n$, sono variabili casuali indipendenti e identicamente distribuite e inoltre

$$D_i \sim \mathcal{N}(\mu_2-\mu_1, \sigma_1^2+\sigma_1^2-2\mathrm{Cov}(X,Y))$$

da cui si evince che

$$\frac{\overline{D}-\mu_D}{S_D/\sqrt{n}} \sim t_{n-1},$$

dove $\mu_D = \mu_2-\mu_1$, $\overline{D} = (1/n)\sum_{i=1}^n D_i$ e $S_D^2 = (1/(n-1))\sum_i (D_i-\overline{D})^2$.

Otteniamo il seguente intervallo di confidenza per μ_D al livello $(1-\alpha)100\%$:

$$\left(\overline{D} - t_{n-1,1-\frac{\alpha}{2}}\frac{S_D}{\sqrt{n}},\ \overline{D} + t_{n-1,1-\frac{\alpha}{2}}\frac{S_D}{\sqrt{n}}\right), \tag{5.11}$$

dove $t_{n-1,1-\frac{\alpha}{2}}$ è, come prima, il valore della distribuzione t con $n-1$ gradi di libertà che lascia una probabilità nella code uguale e pari a $\alpha/2$.

Si fa riferimento all'intervallo appena costruito come all'intervallo di confidenza per la differenza delle medie di osservazioni *appaiate*, cioè tali che l'osservazione X_i è accoppiata all'osservazione Y_i.

Esempio 5.11. Per due fondi di investimento si sono rilevati i rendimenti giornalieri: il rendimento, X, per il primo fondo a 30 giorni ha riportato un valore medio pari a $\bar{x} = 1.46$ (per 10000) con media dei quadrati $(1/30)\sum_{i=1} x_i^2 = 2.80$, mentre per il secondo fondo, osservato negli stessi 30 giorni, si è calcolato un rendimento medio $\bar{y} = 1.16$ con media dei quadrati $(1/30)\sum_{i=1} y_i^2 = 2.42$. Inoltre è noto che $\sum_{i=1}^{30} x_i y_i = 1.93$. Assumiamo che i rendimenti giornalieri dei due fondi siano distribuiti indipendentemente nei diversi giorni e che seguano una distribuzione normale. Si supponga inoltre che i rendimenti dei due fondi nello stesso giorno siano dipendenti e sia $\mathrm{Cov}(X, Y)$ la covarianza tra X e Y. Supponendo che $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ e $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$, possiamo costruire l'intervallo di confidenza al 95% per la media di $D = X - Y$, indicata con $\mu_D = \mu_x - \mu_y$, considerando che $D \sim \mathcal{N}(\mu, \sigma^2)$, dove $\sigma^2 = \sigma_x^2 + \sigma_y^2 - 2\mathrm{Cov}(X, Y)$.

Si tratta quindi di ricavare l'intervallo per la media di una popolazione normale con varianza incognita, pertanto l'intervallo cercato ha estremi

$$\overline{D} \pm t_{n-1,0.975}\sqrt{S^2/n}$$

dove con $\overline{D}$ e S^2 indichiamo, rispettivamente, la media campionaria e la varianza campionaria corretta del campione $(X_i - Y_i)$, $\quad i = 1, \ldots, 30$. Sulla base dei dati a disposizione, la stima di $\overline{D}$ è $\hat{d} = 1.46 - 1.16 = 0.3$ e la stima di S^2 si ottiene come

$$\begin{aligned} s^2 &= \frac{n}{n-1}\left(\frac{\sum_{i=1} x_i^2}{n} - \bar{x}^2 + \frac{\sum_{i=1} y_i^2}{n} - \bar{y}^2 - 2\left((1/n)\sum_i x_i y_i - \bar{x}\bar{y}\right)\right) \\ &= \frac{30}{29}\left(2.8 - 1.46^2 + 2.42 - 1.16^2 - 2\left(\frac{1.93}{30} - 1.46 \cdot 1.16\right)\right) = 5.174. \end{aligned}$$

Quindi si ottiene l'intervallo al livello 95% di estremi $0.3 \pm 2.045\sqrt{5.174/30}$, cioè risulta $(-0.55, 0.85)$.

▲

5.5.2 Differenza tra le medie di due Bernoulliane

Un caso peculiare che merita particolare attenzione riguarda la costruzione di un intervallo di confidenza per la differenza tra le medie di due popolazioni Bernoulliane. Si considerino due campioni indipendenti $(X_1, \ldots, X_{n_1})$ e

$(Y_1, \ldots, Y_{n_2})$ estratti, rispettivamente, da una distribuzione $Be(p_1)$ e $Be(p_2)$. Per ottenere un intervallo di confidenza per la differenza tra le medie $p_1 - p_2$ consideriamo valida l'approssimazione normale per grandi campioni delle distribuzioni degli stimatori $\hat{p}_1$ e $\hat{p}_2$ che denotano le proporzioni campionarie nei due campioni. Ne segue che la distribuzione della variabile aleatoria $\hat{p}_1 - \hat{p}_2$ è approssimativamente

$$\hat{p}_1 - \hat{p}_2 \dot{\sim} \mathcal{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

da cui segue che il rapporto

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

ha distribuzione limite $\mathcal{N}(0,1)$. Quindi, fissato $\alpha \in (0,1)$, possiamo scrivere

$$P((\hat{p}_1 - \hat{p}_2) - z_{1-\alpha/2}\sqrt{S^2} \leq (p_1 - p_2) \leq (\hat{p}_1 - \hat{p}_2) + z_{1-\alpha/2}\sqrt{S^2}) = 1 - \alpha,$$

dove $S^2 = \hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2$. Si perviene quindi all'intervallo di confidenza al livello $(1-\alpha)$ per la differenza tra due proporzioni:

$$\left(\hat{p}_1 - \hat{p}_2 - z_{1-\frac{\alpha}{2}}S,\, \hat{p}_1 - \hat{p}_2 + z_{1-\frac{\alpha}{2}}S\right). \tag{5.12}$$

Esempio 5.12. Due campioni indipendenti di 1000 e 800 votanti in due città differenti indicano che nella città A 132 votanti sono a favore di una riforma elettorale, mentre nella città B sono 194 i cittadini che si sono espressi a favore della riforma. Vogliamo ottenere un intervallo fiduciario al livello del 99% per la differenza tra le proporzioni dei votanti a favore della riforma nelle due città. Le proporzioni osservate nei due campioni di numerosità $n_1 = 1000$ e $n_2 = 800$ sono $\hat{p}_1 = 132/1000 = 0.132$ e $\hat{p}_2 = 194/800 = 0.2425$. Otteniamo una stima della varianza della differenza come

$$s^2 = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2} = \frac{0.1146}{1000} + \frac{0.1837}{800} = 0.00034$$

Considerato che $\alpha = 1 - 0.99 = 0.01$ e $z_{0.995} = 2.576$, l'intervallo cercato ha estremi

$$(0.132 - 0.2425) \pm 2.576\sqrt{0.00034}$$

cioè $(-0.158, -0.063)$.

▲

Capitolo 6

Introduzione all'inferenza bayesiana

6.1 Ancora sulla scelta dell'urna

Si riconsideri l'esempio dell'urna introdotta nel paragrafo 3.1. L'urna è composta da palline bianche e nere in proporzione ignota, è chiusa e non possiamo esaminare il suo contenuto. Si estraggono con reinserimento n palline e si osserva se ogni pallina estratta è bianca o nera. Inoltre è noto che ci sono solo 4 tipi di urne disponibili. La prima urna, U_1, ha una proporzione di palline bianche pari a 0.1; la seconda, U_2, ha il 20% di palline bianche; la terza, U_3, ha una proporzione pari a 0.3, e la quarta, U_4, ha tante palline bianche quante nere. Il problema è quello di ricercare una strategia per "individuare" l'urna dalla quale sono state estratte le palline.

Si è già osservato che è ragionevole la strategia di privilegiare quell'urna che è più verosimile abbia permesso di ottenere la sequenza di palline bianche e nere osservata nelle n estrazioni. Sia y ($y = 0, 1, \ldots, n$) il numero di palline bianche ottenuto nella sequenza di estrazioni con reinserimento e si ricordi che quello che importa ai fini della decisione è esclusivamente il numero di palline estratto e non l'ordine con cui queste appaiono nella sequenza.

Detta Y la variabile aleatoria che riguarda il conteggio di palline bianche si può calcolare la probabilità di osservare y palline bianche per ciascun tipo di urna:

$$L(p_i) = P(Y = y; p_i) = \binom{n}{y} p_i^y (1 - p_i)^{1-y} \tag{6.1}$$

dove p_i per $i = 1, 2, 3, 4$ può assumere i valori 0.1, 0.2, 0.3, 0.5.

Una volta osservato il risultato y, otteniamo quindi i 4 valori $L(p_1), L(p_2), L(p_3), L(p_4)$ delle probabilità definite sopra ed è ragionevole scegliere l'urna per la quale il valore $L(p_i)$ è più elevato. Si tratta infatti dell'urna che è più verosimile abbia prodotto il risultato osservato. Si noti che l'ordinamento dei valori $L(p_i)$ non dipende dal fattore moltiplicativo $\binom{n}{y}$ che è lo stesso qualunque sia l'urna.

La strategia illustrata corrisponde all'approccio che conduce alla stima di massima verosimiglianza in un caso in cui lo spazio parametrico è discreto e $L(p_i)$ è la funzione di verosimiglianza.

Un esempio numerico può essere utile. Si supponga di avere compiuto $n = 20$ estrazioni e di avere ottenuto $y = 8$ palline bianche, i valori di $L(p_i)$ risultano proporzionali (omettiamo il fattore $\binom{n}{y}$ e moltiplichiamo i valori per 10^9) a

$$\begin{aligned} L(0.1) = &\quad 0.1^8 * 0.9^{12} = 2.824295 \\ L(0.2) = &\quad 0.2^8 * 0.8^{12} = 175.9219 \\ L(0.3) = &\quad 0.3^8 * 0.7^{12} = 908.1269 \\ L(0.5) = &\quad\quad\quad 0.5^{20} = 953.6743 \end{aligned}$$

Il criterio della massima verosimiglianza porta quindi a scegliere l'urna 4 e a determinare come stima della proporzione di palline bianche il valore $p_4 = 0.5$.

6.1.1 E se avessimo informazioni sull'urna da cui estraiamo?

Nell'esempio precedente sapevamo solo che vi erano 4 possibili urne ciascuna con le proporzioni p_i $(i = 1, 2, 3, 4)$.

Si immagini ora un diverso contesto per l'esperimento. Si supponga sia noto che le urne U_i sono collocate al piano i-mo di un palazzo e che si chieda all'usciere, che risiede al primo piano, di condurre allo stesso piano un'urna dalla quale si effettueranno le estrazioni.

In questo caso, si potrebbe valutare il fatto che l'usciere possa scegliere con maggiore probabilità un'urna collocata in posizione più comoda per cui, prima di fare l'esperimento, si potrebbe formulare una valutazione della probabilità che l'usciere porti una della quattro urne. Per esempio, si potrebbe

sintetizzare tale valutazione attraverso una distribuzione di probabilità sui valori delle proporzioni p_i:

p_i	$P(p_i)$
0.1	0.4
0.2	0.3
0.3	0.2
0.5	0.1

Si ritiene cioè che sia più probabile che l'usciere conduca al I piano l'urna che è per lui meno faticosa da trasportare. Questa informazione modifica in modo sostanziale il problema: ora si ha una distribuzione di probabilità sulle possibili proporzioni p_i di palline bianche nelle urne.

Quindi avendo davanti l'urna e se non si osserva alcun dato, ovvero se non si effettuano estrazioni, la tabella sopra è *a-priori* la distribuzione di probabilità che fornisce la valutazione di incertezza sulla composizione dell'urna e di conseguenza una valutazione su quale è la proporzione di palline bianche in essa contenuta.

Ci si può chiedere come si modifica questa valutazione, espressa da questa distribuzione iniziale di probabilità sulla variabile aleatoria P che assume valori nello spazio delle possibili proporzioni p_i, se si acquisiscono nuove informazioni, ad esempio se si osservano i risultati di n estrazioni con reinserimento dall'urna.

A questo punto si potrebbe calcolare la distribuzione di probabilità sulle possibili urne condizionatamente al risultato dell'estrazione, in altri termini sarebbe interessante ottenere la distribuzione di probabilità $P(P = p_i|y)$ ove y è il numero di palline bianche ottenuto in n estrazioni.

Ora come abbiamo visto, si può di valutare $P(y|P = p_i)$ ovvero la probabilità di ottenere il risultato y condizionatamente al fatto che l'estrazione avvenga dall'urna i-esima. Tale probabilità è in effetti pari alla verosimiglianza $L(p_i)$ già calcolata in (6.1) (si noti però che poiché ora p_i è la realizzazione di una variabile aleatoria questa quantità va espressa come una probabilità condizionata) e poiché si è espressa già la distribuzione di probabilità *a-priori* $P(P = p_i)$ è possibile utilizzare la formula di Bayes e ottenere:

$$P(P = p_i|y) = \frac{P(y|P = p_i) \cdot P(P = p_i)}{\sum_{i=1}^{4} P(y|P = p_i) \cdot P(P = p_i)}$$

Si noti che quindi che si può anche scrivere anche

$$P(P = p_i|y) \propto P(y|P = p_i) \cdot P(P = p_i) = \text{verosimiglianza} \times \textit{a-priori}$$

Si considerino quindi i dati dell'esempio precedente, ovvero y=8 in $n = 20$ estrazioni, e calcolare la distribuzione di probabilità su P che combina la distribuzione a priori e la verosimiglianza attraverso la formula di Bayes. Tale distribuzione di probabilità è detta *a-posteriori* ed è ottenuta aggiornando l'incertezza *a-priori* con le informazioni ottenute dall'estrazione.

Si calcoli dapprima il denominatore (per comodità moltiplichiamo sempre per 10^9 e notiamo che il fattore $\binom{n}{y}$ è sia al numeratore che al denominatore per cui si semplifica)

$$\sum_{i=1}^{4} P(Y|P = p_i) \cdot P(P = p_i) = 2.824 \cdot 0.4 + 175.921 \cdot 0.3 + 908.126 \cdot 0.2 + \\ + 953.674 \cdot 0.1 = 330.8991.$$

A questo punto si ottiene la distribuzione *a posteriori* di probabilità $P|y$

$P\|y$	$P(P = p_i\|y)$		
p_1	$0.4 \cdot 2.824295/330.8991$	=	0.003414
p_2	$0.3 \cdot 175.9219/330.8991$	=	0.159494
p_3	$0.2 \cdot 908.1269/330.8991$	=	0.548884
p_4	$0.1 \cdot 953.6743/330.8991$	=	0.288207

Questa distribuzione di probabilità è quindi la sintesi di tutte le informazioni disponibili e misura l'incertezza su quale urna possa essere quella da cui sono state estratto le palline.

Si potrebbe decidere anche in questo caso, se si dovesse scegliere una urna specifica, di prendere quella cui corrisponde la massima probabilità. Si noti che se si adottasse questa scelta, con gli stessi dati si sarebbe ora portati a scegliere l'urna 3 (e non nella 4).

6.2 L'approccio bayesiano all'inferenza

6.2.1 Estensione dell'esempio sull'estrazione da un'urna

L'esempio appena proposto illustra l'approccio bayesiano all'inferenza.

Le sue principali caratteristiche possono essere illustrate con riferimento al semplice esempio introduttivo.

- Le conoscenze sul parametro, che in questo caso è l'ignota proporzione di palline bianche nell'urna, sono espresse ora tramite una distribuzione di probabilità . Nel caso di un problema di inferenza parametrica questo equivale a dire che spostiamo il problema da quello di determinare il parametro "vero" che si suppone abbia generato i dati a quello di sintetizzare tutte le conoscenze (quelle *a-priori* e quelle che derivano dalle estrazioni dall'urna) fornendo una distribuzione di probabilità aggiornata (*a-posteriori*) sullo spazio parametrico.

- Si suppone sia possibile fornire una distribuzione di probabilità iniziale, la distribuzione *a-priori*, che sintetizza le conoscenze sul parametro prima di ottenere i dati del campione. Questa operazione è coerente con la definizione soggettivista della probabilità .

- Si noti che, in questo caso, si potrebbe fissare una distribuzione di probabilità che non suppone alcuna preferenza fra le urne, nell'esempio potrebbe essere una distribuzione discreta uniforme, e si otterrebbe quindi un ordine di preferenza per le urne analogo a quello ottenuto senza informazioni *a-priori*. Tuttavia in questa impostazione si otterrebbe comunque una distribuzione di probabilità *a-posteriori* sulle possibili urne. Il problema inferenziale risulterebbe quindi comunque formulato in modo logicamente diverso.

La distribuzione di probabilità *a-posteriori* quindi sintetizza in modo coerente lo stato di informazione sul parametro dopo avere osservato i dati campionari (le estrazioni nell'esempio). E sarà a questa distribuzione che ci si affiderà per fare inferenza sul parametro ignoto (ovvero per scegliere una fra le urne o, se si vuole, per scommettere in modo coerente su quale sia l'urna da cui abbiamo estratto).

6.2.2 Inferenza su una proporzione

Si potrebbe quindi ora estendere l'esempio e pensare, più in generale, che vi siano infinite urne ciascuna caratterizzata da una proporzione p di palline bianche. In questo caso quindi il valore p è la quantità (il parametro) sul quale si vuole fare inferenza e lo spazio parametrico sarebbe costituito dai valori reali dell'intervallo $(0, 1)$.

Si supporrà di avere una distribuzione di probabilità *a-priori* per p, espressa dalla funzione di densità $\pi(p)$ definita su (0,1) che riassume le cono-

scenze (o meglio l'incertezza) sulla composizione dell'urna prima di estrarre informazioni.

Ora si supponga di estrarre n palline con reinserimento dall'urna che ha una proporzione di palline bianche p. Condizionatamente a p se si estrae un numero $Y = y$ di palline bianche, si potrà calcolare la probabilità $P(Y = y|p)$ che sarà quindi una binomiale $Bin(n, p)$.

Ripercorrendo lo schema precedente si può ottenere la distribuzione di probabilità *a-posteriori* per $p|Y = y$ usando la formula di Bayes

$$g(p|Y = y) = \frac{P(Y = y|p) \cdot \pi(p)}{\int_0^1 P(Y = y|p) \cdot \pi(p)dp}. \tag{6.2}$$

In altri termini, in generale, si potrebbe dire che data una distribuzione di probabilità iniziale sul parametro p e calcolata la verosimiglianza dei dati osservati in corrispondenza di ciascun possibile valore p, $f(y|p)$, l'informazione *a-posteriori* per p è

$$g(p|y) \propto \pi(p)f(y|p) \tag{6.3}$$

Se si dovesse scegliere una distribuzione di probabilità che riassume le informazioni su p prima di effettuare le estrazioni questa dovrebbe avere come supporto l'insieme dei possibili valori p. Quindi si tratterebbe di una distribuzione di probabilità sull'intervallo reale (0,1). Per esempio sarebbe comodo scegliere una distribuzione di probabilità della famiglia delle Beta. Come è noto una variabile aleatoria distribuita come una $Beta(\alpha, \beta)$ (con $\alpha, \beta > 0$) ha funzione di densità

$$\pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1}, \qquad 0 < p < 1$$

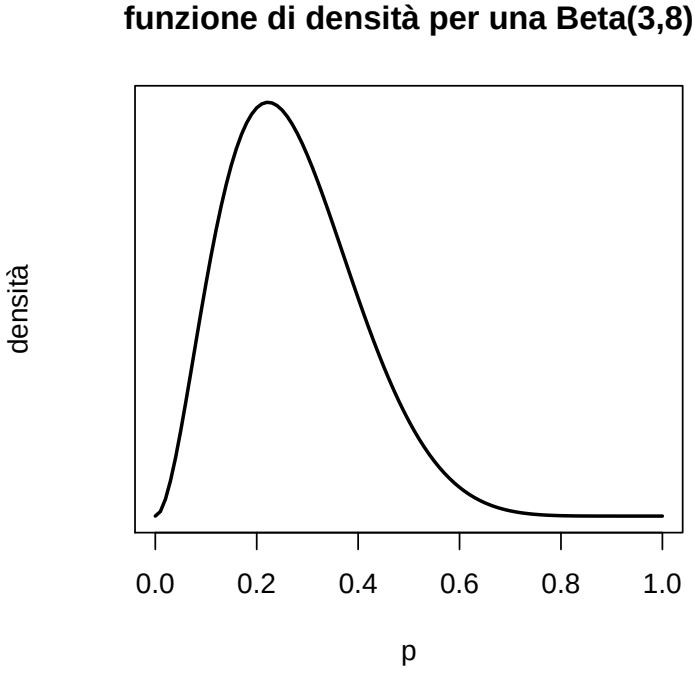
I due parametri α e β potrebbero esser scelti così da riflettere le preferenze *a-priori* sui possibili valori di p. Si ricorda che la media di una $Beta(\alpha, \beta)$ risulta pari a $M = \alpha/(\alpha + \beta)$ e la varianza è $V = M(1 - M)/(\alpha + \beta + 1)$.

Se ad esempio si ritenesse che sono più plausibili i valori inferiori a 0.5 allora si potrebbe usare come *a-priori* una $Beta(3, 8)$ che ha media $3/11$ e la cui funzione di densità assegna maggiore probabilità ai valori minori di 0.5 come si vede dal grafico che segue.

La densità *a-posteriori* per $P|y$ sarà quindi:

$$g(p|Y = y) = \frac{p^y(1-p)^{n-y}p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 p^y(1-p)^{n-y}p^{\alpha-1}(1-p)^{\beta-1}dp}, \qquad 0 < p < 1$$

il fattore $\binom{n}{y}$ appare sia al numeratore che al denominatore per cui si semplifica.



Si verifica agevolmente che

$$g(p|y) \propto p^{\alpha+y-1}(1-p)^{\beta+n-y-1}, \qquad 0 < p < 1$$

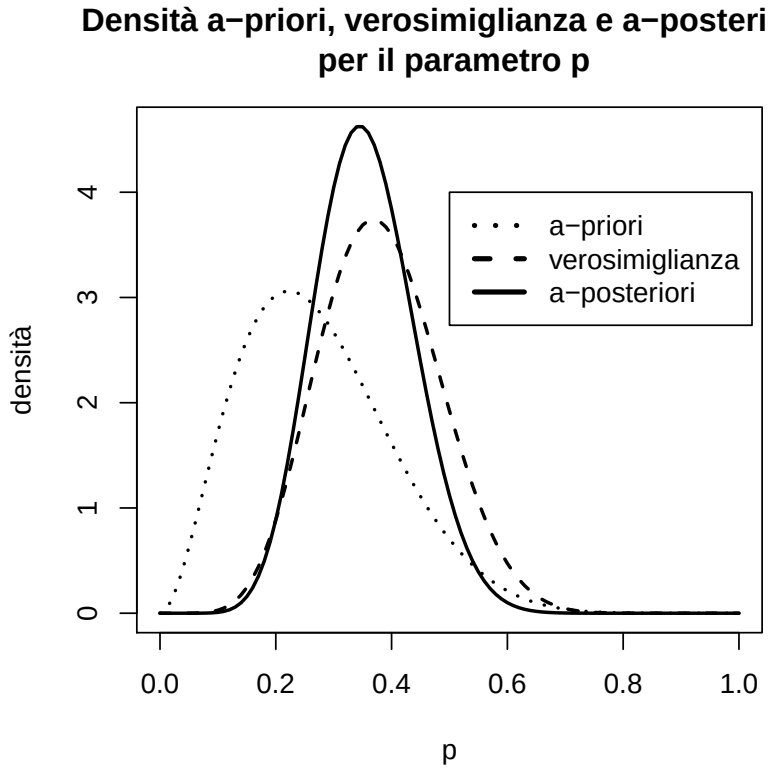
La densità *a-posteriori* risulta quindi essere ancora una distribuzione della famiglia Beta con parametri $\alpha + y$ e $\beta + n - y$. Se si vuole quindi fornire un unico valore a riassumere la distribuzione di probabilità del parametro si può usare una qualsiasi misura di sintesi per la distribuzione. Nell'esempio esposto, si immaginava che una sintesi efficace (quella che potremmo considerare una stima puntuale) sia la moda della distribuzione *a-posteriori*. Un'eccellente sintesi di una distribuzione potrebbe essere la media. Allo stesso fine si potrebbe anche considerare la mediana quale misura di sintesi. Una misura riassuntiva dell'incertezza potrebbe essere inoltre fornita dalla varianza della distribuzione *a-posteriori*.

Se si decidesse nell'esempio visto prima di riassumere le informazioni sul parametro attraverso la media della *a-posteriori*, essa sarebbe pari a

$$\frac{\alpha + y}{\alpha + \beta + n} = \frac{\alpha + \beta}{\alpha + \beta + n} \cdot \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \cdot \frac{y}{n} \tag{6.4}$$

La media dell'*a-posteriori* è quindi una media ponderata della media dell'*a-priori*, pari ad $\alpha/(\alpha+\beta)$, e il valore y/n, la proporzione campionaria, per cui è massima la verosimiglianza. Pertanto la media dell'*a-posteriori* è compresa fra le medie di *a-priori* e il massimo della verosimiglianza.

Si noti inoltre che il peso della seconda componente cresce con la dimensione campionaria per cui al crescere del campione la media della *a-posteriori* si avvicinerà sempre di più a coincidere con la stima di massima verosimiglianza. Si noti che anche la moda dell'*a-posteriori* è compresa fra le mode di *a-priori* e funzione di verosimiglianza, come si evince dalla figura dove sono riportate le distribuzioni *a-priori*, *a-posteriori* e la verosimiglianza (normalizzata così che l'area sotto di essa sia pari a 1) per i dati dell'esempio utilizzato. Se si estraesse quindi un numero crescente di palline la media dell'*a-posteriori* tenderà a essere pari al valore che è la stima di massima verosimiglianza y/n. Quindi il valore dell'informazione campionaria sarà dominante rispetto all'informazione *a-priori* che diventerà sempre meno rilevante al crescere della dimensione del campione. Si noti poi che in questo caso la varianza della *a-posteriori* si ridurrà e tenderà a zero al crescere del numero di palline estratte.



La media della *a-priori* è $3/11{=}0.2727$, mentre la media della *a-posteriori* è $11/31{=}0.3548$.

Nei casi intermedi si può pensare che l'informazione che deriva del campione permette di aggiornare coerentemente quanto si sapeva sul parametro prima di conoscere l'informazione campionaria (di estrarre cioè le palline dall'urna).

6.2.3 Intervalli di confidenza e intervalli di credibilità

In questo caso tutta l'incertezza sarà contenuta nella distribuzione *a-posteriori*. Sarà questo lo strumento per fare affermazioni inferenziali.

Si potrebbe quindi decidere di fare affermazioni probabilistiche sul valore di p una volta osservato un risultato campionario. Ad esempio calcolando la $P_{P|y}(p \leq 0.5)$ o la probabilità nell'*a-posteriori* che p caschi in un intervallo definito. O ancora si potrebbe calcolare un intervallo in cui p ricade con probabilità elevata fissata pari a $1 - \alpha$. In questo caso si otterrebbe un intervallo che è detto di credibilità (non di confidenza).

Nel caso dell'*intervallo di confidenza classico*, è l'intervallo a essere casuale e il valore della costante p sarà contenuto nell'intervallo in una proporzione $(1-\alpha)$ dei campioni casuali di pari dimensione che venissero tratti dalla stessa popolazione.

Per l'*intervallo di credibilità* bayesiano p è determinazione di una variabile aleatoria e l'interpretazione è quindi diversa e forse più agevole e naturale: in questo caso si può legittimamente affermare che la probabilità che p sia interna al suddetto intervallo è pari a quella da noi fissata e pari a $1-\alpha$ senza ricorrere all'espediente di chiedersi cosa accadrebbe in una ipotetica ripetizione del campionamento.

6.3 Verso la statistica bayesiana

6.3.1 Impostazione generale

L'esempio illustrato è in grado di chiarire alcuni passaggi chiave dell'approccio all'inferenza bayesiana parametrica che deriva quindi da un'impostazione molto diversa del problema inferenziale. Non si fanno affermazioni con riferimento all'ipotetica distribuzione delle statistiche che si otterrebbero se si immaginasse di ripetere più volte il campionamento. Si ricava invece una distribuzione di probabilità per il parametro di interesse combinando tutta l'informazione disponibile.

L'esempio visto è uno dei più semplici e in realtà per altri casi sarà necessari ripetere i passaggi logici visti per ricavare l'*a-posteriori* sulla base della quale fondare le affermazioni inferenziali.

Sia quindi θ il parametro che caratterizza la distribuzione di Y nella popolazione. E si immagini che il valore $\theta \in \Theta \subset \Re$ sia determinazione di una variabile aleatoria la cui funzione di densità è $\pi(\theta)$ definita su Θ. La densità $\pi(\theta)$ descrive quindi l'incertezza su θ prima di osservare un campione di determinazioni da Y. Osservati i dati campionari $y_1, y_2, \ldots, y_n$ ed essendo nota la legge di distribuzione di $f(y|\theta)$, si può definire la verosimiglianza del campione osservato come $L(y_1, y_2, \ldots, y_n|\theta)$. A questo punto l'applicazione della formula di Bayes permette di ottenere la **distribuzione a-posteriori** $g(\theta|y_1, y_2, \ldots, y_n)$. Tale distribuzione finale aggiorna quindi le conoscenze *a-priori* dopo avere osservato il campione.

$$g(\theta|y_1, y_2, \ldots, y_n) = \frac{L(y_1, y_2, \ldots, y_n|\theta) \cdot \pi(\theta)}{\int_{\theta\in\Theta} L(y_1, y_2, \ldots, y_n|\theta) \cdot \pi(\theta) d\theta}, \qquad \theta \in \Theta \qquad (6.5)$$

6.3.2 Aspetti ulteriori

Come si vede la *a-posteriori* è pari al prodotto fra verosimiglianza e *a-priori* diviso una costante di normalizzazione (l'integrale al denominatore).

Nell'esempio illustrato relativo al parametro p di una Bernoulliana, la verosimiglianza $f(y|p)$ e la *a-priori* $\pi(p)$ si combinavano in modo agevole ricavando una densità *a-posteriori* che ha forma nota. In particolare, accadeva che *a-priori* e *a-posteriori* appartenessero alla stessa famiglia (quella delle Beta).

Questo accade con altre distribuzioni (ad esempio, nel caso di una *a-priori* gaussiana per il parametro μ di una gaussiana, l'*a-posteriori* è ancora gaussiana, oppure nel caso di una *a-priori* gamma per il parametro λ di una Poisson si ricava l'*a-posteriori* che è ancora una gamma).

In casi come questi si dice che la *a-priori* è **coniugata** con la distribuzione che si suppone descriva la popolazione da cui si sono tratto i dati.

Tuttavia le cose non vanno sempre così bene e ricavare la *a-posteriori* non è sempre agevole. E in numerosi casi non è facile ricavare la costante di normalizzazione al denominatore nella formula.

Riassumere poi le informazioni *a-priori* in una distribuzione non è sempre operazione semplice. Spesso le critiche all'impostazione bayesiana si concentrano proprio sul fatto che l'inferenza potrebbe dipendere dalla formulazione della distribuzione *a-priori*, che potrebbe essere differente per diversi ricercatori, introducendo margini di arbitrarietà. Per tale motivo, sono stati sviluppati studi che mostrano la robustezza dei risultati che si ottengono in presenza di differenti informazioni *a-priori*. Si ricorda peraltro che se vi è tanta informazione campionaria, ovvero in presenza di campioni numerosi, il peso della informazione *a-priori* diminuisce di importanza.

Anche per evitare tali critiche, in alcuni casi alle *a-priori* che danno preferenza ad alcuni valori del parametro (dette *a-priori* informative) si preferiscono distribuzioni *a-priori* sullo spazio dei parametri che sono vaghe e diffuse (quindi poco informative o non informative) così da enfatizzare il ruolo dell'informazione campionaria. Sono state proposte varie soluzioni per individuare distribuzioni *a-priori* non informative tuttavia la loro trattazione va ben oltre gli scopi di questa breve introduzione.

Recentemente molti dei problemi legati alla derivazione della *a-posteriori* sono stati superati in quanto sono stati sviluppati efficienti metodi di simulazione Monte Carlo per ottenere campioni dalla distribuzione *a-posteriori* anche quando la forma della stessa non può essere ottenuta analiticamente: si tratta di simulazioni che sfruttano la teoria dei processi stocastici detti catene di Markov e per tale motivo si parla di metodi MCMC (Monte Carlo Markov Chain). Questo ha consentito di estendere l'uso del ragionamento bayesiano a problemi di inferenza anche molto complessi.

L'approccio all'inferenza bayesiana è quindi per molti versi da ritenersi un'eccellente approccio per affrontare i problemi di inferenza statistica e la valutazione dell'incertezza si basa su fondamenti, in buona parte, diversi da quelli tipici dell'approccio classico.

La valutazione dell'incertezza, inevitabilmente presente in qualsiasi conclusione inferenziale, è caratterizzata nell'approccio bayesiano da una maggiore chiarezza interpretativa non essendo necessario fare ricorso a quanto potrebbe accadere in ipotetiche replicazioni della rilevazione campionaria.

Capitolo 7

Verifica di ipotesi

7.1 Introduzione

Gregor Mendel, nei suoi esperimenti sulla fecondazione del *Pisum*, osservò la variazione di una determinata caratteristica da una generazione all'altra. Il suo intento era quello di formulare una legge per la trasmissione dei caratteri. Le varietà di piante scelte per gli esperimenti differivano su diverse caratteristiche come il colore, la forma e le dimensioni dei baccelli. In particolare, egli condusse un famoso esperimento incrociando piante con baccelli verdi (caratteristica dominante) e baccelli gialli (caratteristica recessiva) ottenendo un campione di nuova generazione formato da 580 piante di piselli, le cui caratteristiche si possono riassumere come segue:

baccelli verdi	428
baccelli gialli	152
	580

Mendel ipotizzava che un quarto delle piante del campione di nuova generazione avrebbe presentato il carattere recessivo, ovvero il baccello di colore giallo. Tuttavia, dai dati ottenuti egli osserva una proporzione di piante con baccelli gialli pari a $152/580 = 0.262$. Quindi la percentuale osservata fu pari a 26.2%. Si può dire in questo caso che i risultati dell'esperimento non confermano la teoria sviluppata da Mendel? Cosa ci si aspetta di osservare se si ripetesse l'esperimento e fosse vera l'ipotesi che la probabilità di riproduzione del carattere recessivo è pari al 25%?

In altri termini, formulata una **ipotesi**, quindi una affermazione su una proprietà di una popolazione (la popolazione delle piante di pisello nell'esempio), come si decide se i dati campionari siano conformi all'ipotesi?

Per rispondere a tali domande è possibile ricorrere a procedure statistiche per decidere se i risultati siano effettivamente conformi a quello che ci si aspetterebbe se il vero valore fosse 25%. Tali procedure sono riconducibili alle procedure di verifica di ipotesi statistiche. Tali ipotesi possono riguardare la distribuzione di una variabile Y e quindi possono essere relative a un parametro, e vengono dette ipotesi parametriche, o ad altre caratteristiche della distribuzione. Di seguito si introducono i concetti alla base della verifica di ipotesi sul parametro di una popolazione.

7.2 Test di significatività

Nell'introduzione abbiamo fornito un esempio in cui si formulava un'ipotesi sul valore di una proporzione. Generalizzando l'esempio, si consideri una popolazione nella quale la variabile Y e distribuita secondo una legge nota ma alla quale è associato un parametro θ ignoto. Si supponga che sia possibile formulare, sulla base di una teoria o una congettura, una ipotesi sul valore del parametro. Tale ipotesi si indica con H_0 ed è detta **ipotesi nulla**:

$$H_0 : \theta = \theta_0$$

Come si vedrà in seguito, tale ipotesi può essere di diversi tipi. Nell'esempio considerato, l'ipotesi nulla è espressa dall'affermazione

La proporzione di piselli con baccelli gialli è uguale a 0.25

essa riguarda in realtà una variabile $Y \sim Be(p)$ dove p rappresenta la proporzione di piante con baccelli gialli nella popolazione e si traduce quindi in

$$H_0 : p = 0.25$$

Il punto di partenza è quindi l'identificazione dell'ipotesi da sottoporre a verifica che esprime, ad esempio, if fatto che una terapia non sia stata efficace, o che l'utilizzo di un particolare trattamento non abbia portato ad un miglioramento. Come sempre si immagina di osservare un campione $Y_1, Y_2, \ldots, Y_n$ e il nostro obiettivo è valutare se i dati osservati è ragionevole pensare che si siano ottenuti essendo vera l'ipotesi nulla. Una procedura che si può adottare

è quella dei cosiddetti **test di significatività** che implica la misurazione di quanto estremo sia il dato ottenuto dal campione quando si assume che H_0 sia vera. Si tratta di valutare se e quanto il dato campionario sia sorprendente se si ipotizzasse che H_0 è vera.

Questa valutazione viene effettuata sulla base di una probabilità, detta **valore-p** (*p-value*), che fornisce una misura di quanto sia plausibile osservare dati ancora più estremi di quelli osservati se si assume vera l'ipotesi nulla. Un valore molto piccolo di tale probabilità è una misura informale di quanta evidenza vi sia contro l'ipotesi nulla fatta. Il valore-p è una grandezza continua compresa tra 0 e 1. Proseguendo il ragionamento informale senza considerare in modo rigido le soglie fornite, si riassume l'evidenza contro H_0 nella tabella che segue

valore di p	evidenza contro H_0
$p > 0.1$	nessuna o poca evidenza
$0.01 < p \leq 0.1$	evidenza debole
$0.001 < p \leq 0.01$	evidenza sostanziale
$p \leq 0.001$	forte evidenza

Si osservi tuttavia che un valore-p elevato non è la prova che l'ipotesi nulla sia vera, né rappresenta la probabilità che l'ipotesi nulla sia vera. Il valore-p può essere pensato come un 'fattore sorpresa', che risponde alla domanda su quanto è sorprendente avere osservato quel campione se è vera l'ipotesi nulla. Nella sezione che seguono sono illustrati alcuni esempi per chiarire ulteriormente il ruolo del valore-p nella verifica di ipotesi.

7.2.1 Verifica d'ipotesi su una proporzione

Partendo dall'esperimento di Mendel si formula l'ipotesi H_0 espressa dalla seguente affermazione: nella popolazione delle piante di pisello, la probabilità che una pianta di nuova generazione abbia il baccello giallo è pari a $1/4$. Si dispone dei dati dell'esperimento sugli incroci da cui risultavano 152 baccelli gialli e 428 baccelli verdi. La proporzione di baccelli gialli nel campione, $\hat{p}$, ha, approssimativamente, distribuzione normale con media p (probabilità di una pianta con baccello giallo) e varianza $p(1-p)/n$. Se l'ipotesi nulla è vera, il numero di baccelli gialli, che corrisponde a n volte la proporzione campionaria, è assimilabile a un valore tratto da una distribuzione $n\hat{p} \sim Bin(n, p)$. Si potrebbe chiedersi se il valore osservato $\hat{p}_{oss} = \frac{152}{580} = 0.262$ è un valore che è plausibile ottenere da questa distribuzione. In effetti essendo la

media della distribuzione binomiale pari a $n \cdot p = 580 \cdot 0.25 = 145$ e l'errore standard pari a circa 10.4, il valore osservato pari a 152 dista dalla media meno di uno scarto quadratico medio ed è quindi un valore che non si può ritenere poco plausibile. Tuttavia, per comprendere meglio se 152 sia un valore "estremo" quando è vera l'ipotesi di Mendel, potremmo considerare la probabilità di ottenere valori ancora più estremi calcolando ad esempio la probabilità che nella binomiale suddetta si osservino più di 152 baccelli gialli. Se tale probabilità fosse molto piccola allora potremmo pensare che i dati osservati possano provenire da una variabile aleatoria diversa da quella descritta da H_0. Si noti che per svolgere tale calcolo, data l'ampia dimensione campionaria, si può ricorrere all'approssimazione gaussiana, per cui

$$\hat{p} \sim \mathcal{N}(0.25, 0.1875/n)$$

e valutare quanto il valore 0.262 osservato per $\hat{p}$ sia compatibile con l'ipotesi nulla basandoci sulla probabilità di osservare uno scostamento da 0.25 ancora più grande, se l'ipotesi è vera. Tenendo conto che $n = 580$, si calcola la seguente probabilità approssimata

$$\begin{aligned} P(|\hat{p} - 0.25| \geq |0.25 - 0.262|) &= P\left(\left|\frac{\hat{p} - 0.25}{\sqrt{0.1875/n}}\right| \geq \frac{0.012}{\sqrt{0.1875/n}}\right) \\ &= 2(1 - \Phi(0.667)) \\ &= 0.505 \end{aligned}$$

Si valuta in questo caso la probabilità di scostamenti superiori allo scarto, preso in valore assoluto, tra 0.262 e 0.25. In effetti, sarebbero indicativi di scarsa aderenza dei dati osservati all'ipotesi nulla sia valori molto più piccoli di 0.25 che molto più grandi di 0.25, non essendovi indicazioni di sorta su cosa dovrebbe (potrebbe) accadere se H_0 non fosse vera. Cioè non si hanno indicazioni precise sul fatto che dati più sfavorevoli ad H_0 si ottengano solo con proporzioni più elevate (o meno elevate) di quella specificata da H_0.

La probabilità che si ottiene in questo esempio non è trascurabile, e questo porta a sostenere che i dati campionari sono compatibili con l'ipotesi fatta. Si può generalizzare quanto visto come segue.

Modello bernoulliano: test bilaterale

Si dispone di un campione $(Y_1, Y_2, \ldots, Y_n)$ da Y che ha distribuzione di probabilità bernoulliana di parametro incognito $p \in [0, 1]$. Si supponga di for-

mulare l'ipotesi $H_0 : p = p_0$. Sotto H_0, e se n è elevato tanto da assicurare che si possa ricorrere all'approssimazione normale, vale

$$Z = \frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1-p_0)}} \dot{\sim} \mathcal{N}(0,1). \tag{7.1}$$

Come si è discusso nell'esempio, per calcolare la probabilità di ottenere valori ancora più estremi di quello osservato occorrerebbe decidere cosa vuol dire "più estremi". Nell'esempio non vi era ragione per stabilire che fossero estreme proporzioni osservate superiori oppure inferiori a $p = 0.25$. Un modo sintetico per fornire tale indicazione è quello di esplicitare un'ipotesi alternativa. Spesso si rappresenta tale alternativa specificando che essa comprende tutti valori del parametro eccetto quello specificato da H_0. In questo caso, si potrebbe scrivere l'alternativa come $p \neq p_0$ (l'ipotesi alternativa viene comunemente indicata con H_1). Si noti che in questo contesto l'**ipotesi alternativa** H_1 ha solo il ruolo di guida per valutare quali siano i risultati campionari meno favorevoli ad H_0 di quello effettivamente osservato.

Il valore-p approssimato (per n grande) per verificare l'ipotesi H_0 a fronte dell'alternativa specificata sopra risulta essere

$$P\left(|Z| \geq \left|\frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1-p_0)}}\right|\right) \approx 2\left[1 - \Phi\left(\left|\frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1-p_0)}}\right|\right)\right], \tag{7.2}$$

dove $\hat{p}$ è il valore osservato della proporzione campionaria.

Come si è visto il valore-p fornisce la probabilità di ottenere un valore della quantità in (7.1) che sia estremo almeno come quello ottenuto dai dati campionari, assumendo che l'ipotesi nulla sia vera. È chiaro che l'ipotesi nulla è tanto meno "plausibile" quanto minore è il valore-p.

Modello bernoulliano: test unilaterale

Si è già avuto modo di osservare che è necessario formulare un criterio che ci consenta di dire quali dati siano ancor meno favorevoli ad H_0, più estremi quindi, di quelli effettivamente osservati. Nel caso sopra si era proposta l'introduzione di un'ipotesi alternativa ad H_0 che definiva valori alternativi del parametro e che indicherebbero in quale direzione cercare i dati "più estremi" se vera H_0. Il contesto applicativo a volte suggerisce di guardare ai dati ancor meno favorevoli ad H_0 solo in una specifica direzione. Se si considera ad esempio il problema di stabilire quale sia la proporzione di coloro

che hanno remissione di un sintomo dopo aver assunto un farmaco e fosse noto che in condizioni normali il 30% dei pazienti guarisce spontaneamente, allora H_0 specificherebbe che la probabilità di guarire è pari a 0.3 e si valuterà se i dati supportano tale ipotesi oppure se essi forniscano valori molto inusuali sotto H_0, cioè una proporzione osservata che è molto più elevata di 0.3. Valori ancora più elevati di 0.3 sono quindi indicativi del fatto che è improbabile che la differenza nella proporzione dei guariti sia imputabile al caso, dando scarso supporto alla ipotesi di nessuna efficacia del farmaco. Per rendere esplicito che sono solo i valori elevati, maggiori di 0.3, quelli che si considerano per ottenere il valore-p, si indicherà con $p > 0.3$ l'*ipotesi alternativa unilaterale*. Ovviamente vi saranno altri casi in cui si guardarà solo a valori a sinistra di 0.3 e l'ipotesi alternativa sarà specificata di conseguenza.

Si supponga di osservare un campione $(y_1, y_2, \ldots, y_n)$ da Y che ha distribuzione di probabilità bernoulliana di parametro incognito $p \in [0, 1]$.

(a.) Si formuli l'ipotesi

$$H_0 : p = p_0$$

Sotto H_0 vale la (7.1) e il valore-p approssimato (per n grande) per verificare l'ipotesi H_0 a fronte dell'alternativa $p > p_0$ è espresso da

$$P\left(Z > \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right) \approx 1 - \Phi\left(\frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1-p_0)}}\right), \tag{7.3}$$

dove $\hat{p}$ è il valore osservato della proporzione campionaria.

(b.) Si formuli l'ipotesi

$$H_0 : p = p_0$$

Sotto H_0 vale la (7.1) e il valore-p approssimato (per n grande) per verificare l'ipotesi H_0 a fronte dell'alternativa $p < p_0$ è

$$P\left(Z < \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right) \approx \Phi\left(\frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1-p_0)}}\right) \tag{7.4}$$

Quanto illustrato fornisce un metodo per la verifica di una ipotesi unilaterale sul parametro di una popolazione bernoulliana sulla base della probabilità di osservare, sotto la validità di H_0, uno scostamento ancora maggiore di quello del dato osservato dal valore ipotizzato p_0.

Esempio 7.1. Si immagini di giocare a 'testa' o 'croce' e di scommettere su 'croce'; si vince 1 euro se nel lancio esce croce e si perde 1 euro se esce testa. Si tratta di un gioco equo se si conta sul fatto che la moneta è equilibrata ovvero che sia $p = 0.5$ la probabilità di osservare una croce. Dopo 200 lanci si ottiene 72 volte testa si osserva una perdita di 56 euro e sorge il sospetto che la moneta sia truccata (cioè p potrebbe essere in realtà inferiore a 0.5). Ci si può chiedere se perdere 56 euro dopo 200 lanci sia compatibile con l'ipotesi $H_0 : p = 0.5$ o non sia un dato estremo che farebbe sospettare che tale ipotesi non sia vera. In tal caso, interessa valutare come "estremi" solo i casi in cui la perdita sia ancora superiore, ovvero $H_1 : p < 0.5$. In questo caso $\hat{p} = 72/200 = 0.36$ il valore-p è rappresentato dalla probabilità di ottenere ancor meno di 72. Utilizzando l'approssimazione binomiale-normale risulta

$$\Phi\left(\frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1-p_0)}}\right) = \Phi\left(\frac{\sqrt{200}(0.36 - 0.5)}{\sqrt{0.36(0.64)}}\right) = \Phi(-4.125) = 0.000018$$

Si tratta di una probabilità estremamente bassa e vi è quindi una forte evidenza contro l'ipotesi che si tratti di una moneta equilibrata. D'altra parte, il valore -4.125 è un valore molto piccolo nella coda sinistra di una gaussiana standard e quindi si può ritenere molto implausibile, per quanto non impossibile, che provenga da tale distribuzione. Converrebbe comunque smettere di giocare a quelle condizioni. ▲

Esempio 7.2. Alle ultime elezioni un partito ha ricevuto il 29% dei voti. Due anni dopo, da un sondaggio di opinione basato su 300 interviste si è trovato che il 32% degli intervistati ha dichiarato di essere disposto a votare per lo stesso partito. L'ipotesi che si può formulare è che la proporzione dei cittadini che voterebbe per il partito considerato sia uguale a $p = 0.29$, a fronte dell'alternativa che l'appoggio al partito sia diminuito, cioè $p < 0.29$. Assumendo valida l'approssimazione normale, si calcola il p-value approssimato come

$$\Phi\left(\frac{0.32 - 0.29}{\sqrt{\frac{0.29(1-0.29)}{300}}}\right) = \Phi(1.145) \approx 0.87$$

La probabilità ottenuta è piuttosto elevata, tale da non mettere in dubbio l'ipotesi nulla, per cui concludiamo che non ci sono evidenze campionarie sufficienti per smentire l'ipotesi formulata. ▲

7.2.2 Test di significatività per la media di una normale

Si considera ora un procedimento per valutare quanto supporto l'osservazione di un campione da una popolazione normale $\mathcal{N}(\mu, \sigma^2)$ possa fornire a un'ipotesi sulla media incognita μ del tipo $H_0 : \mu = \mu_0$, dove μ_0 è un valore fissato.

Popolazione normale con varianza nota

Dato un campione casuale $(Y_1, \ldots, Y_n)$ da $Y \sim \mathcal{N}(\mu, \sigma^2)$, dove si assume σ nota, se l'ipotesi nulla è

$$H_0 : \mu = \mu_0$$

dove μ_0 è un valore dato, la distribuzione della media campionaria nell'ipotesi nulla è una normale e si ha $\overline{Y} \sim \mathcal{N}(\mu_0, \sigma^2/n)$. Allora un metodo per valutare se l'ipotesi formulata è plausibile è quello di confrontare il valore $\bar{y}$ ottenuto dai dati campionaria con i valori che potrebbero esser prodotti da questa distribuzione. Poiché $\overline{Y}$ segue una distribuzione normale, ciò implica che valuteremo valori di $\bar{y}$ che sono più distanti da μ_0 nelle code della distribuzione come realizzazioni meno plausibili dalla distribuzione di $\overline{Y}$ sotto H_0. Sapendo che quando è vera H_0 risulta

$$Z = \frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \tag{7.5}$$

il valore-p del **test bilaterale** si calcola come segue

$$\begin{aligned} P(|\overline{Y} - \mu_0| \geq |\bar{y} - \mu_0|) =& P\left(|Z| \geq \left|\frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}\right|\right) \\ =& 2\left[1 - \Phi\left(\left|\frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}\right|\right)\right]. \end{aligned}$$

Se si ottenesse un valore molto piccolo del valore-p allora vorrebbe dire che $\bar{y}$ osservato è un valore molto distante da μ_0 e, se H_0 fosse vera, questo sarebbe un valore poco probabile. In tal caso, i dati campionari non supporterebbero l'ipotesi formulata.

Esempio 7.3. Si supponga di aver generato $n = 10$ valori da una $\mathcal{N}(25, 3)$. Si assuma, ad esempio, di non conoscere il valore della media della popolazione da cui i dati sono stati generati, e formuliamo l'ipotesi $H_0 : \mu = 24$.

23.4465	27.7502	24.8610	26.2262	28.4372
25.3202	23.0421	25.2294	24.5848	24.7596

Il valore-p associato è dato da (si utilizza il software `R` per ottenere il valore cercato di Φ):

$$2\left[1-\Phi\left(\frac{|\bar{y}-\mu_0|}{\sigma/\sqrt{n}}\right)\right] = 2\left[1-\Phi\left(\frac{|25.3657-24|}{\sqrt{3/10}}\right)\right]$$
$$= 2(1-\Phi(2.4934)) = 0.0127,$$

che rappresenta una probabilità abbastanza piccola. Ciò implica che se l'ipotesi nulla fosse vera, osserveremmo un valore di $\bar{y}$ distante dall'ipotesi almeno quanto quello ottenuto ($\bar{y} = 25.3657$) solo nel 1.27% dei casi, che ci porta a avanzare qualche dubbio dell'ipotesi nulla (e avendo generato i dati sappiamo che i nostri dubbi sono fondati). ▲

Popolazione normale con varianza incognita

Si procedere in modo simile a quanto fatto in precedenza se, mantenendo l'assunzione che la popolazione sia normale, si assume di non conoscere la varianza σ_0^2. La verifica dell'ipotesi $H_0 : \mu = \mu_0$ si baserà sul rapporto

$$T = \frac{\overline{Y}-\mu_0}{S/\sqrt{n}} \tag{7.6}$$

che ha distribuzione t di Student con $n-1$ gradi di libertà. Quindi verifichiamo l'ipotesi formulata mediante il calcolo del valore-p

$$P\left(|T| \geq \left|\frac{\bar{y}-\mu_0}{s/\sqrt{n}}\right|\right) = 2\left[1-F_{t_{n-1}}\left(\left|\frac{\bar{y}-\mu_0}{s/\sqrt{n}}\right|\right)\right], \tag{7.7}$$

dove $F_{t_{n-1}}$ è la funzione di ripartizione della t di Student con $n-1$ gradi di libertà.

Esempio 7.4. Sono state osservate le pulsazioni cardiache (in battiti per minuto) di 100 studenti. La media e la varianza ottenute dal campione sono state $\bar{y} = 68.7$ e $s^2 = 75.12$. Ci si propone di verificare l'ipotesi che la media dei battiti al minuto sia uguale al valore $\mu_0 = 70$. Anche se

non abbiamo l'informazione che il campione provenga da una normale, data l'ampiezza sufficientemente elevata, possiamo assumere che $\overline{Y}$ si distribuisca approssimativamente come una normale. Allora, si calcola la quantità

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{68.7 - 70}{8.667/10} = -1.4999$$

Il riferimento è alla t di Student con 99 gradi di libertà, quindi si ottiene il valore-p come

$$2\left[1 - F_{t_{99}}\left(\left|\frac{\bar{y} - \mu_0}{s/\sqrt{n}}\right|\right)\right] = 2(1 - F_{t_{99}}(1.4999))$$

Usando il software `R` si trova $2(1 - 0.9315887) = 0.137$, ma considerando che la t di Student con 99 gradi di libertà è approssimativamente uguale alla normale, possiamo ricavare la probabilità cercata dalle tavole della distribuzione normale standard:

$$2\left[1 - \Phi\left(\left|\frac{68.7 - 70}{8.667/10}\right|\right)\right] = 2(1 - \Phi(1.4999)) = 0.134.$$

Si deduce che l'osservazione fatta non permette di escludere la validità dell'ipotesi. ▲

7.2.3 Test unilaterale per la media di una normale

Nella sezione 7.2.2 si è formulata l'ipotesi che il parametro di interesse sia uguale ad uno valore dato μ_0. Altre ipotesi di interesse considerano ancora $H_0 : \mu = \mu_0$ ma i valori non plausibili sotto H_0 sono solo a destra o a sinistra di μ. La procedura di verifica d'ipotesi si conduce, in tal caso, in modo simile a quanto già visto ma il calcolo del valore-p dovrà riflettere la natura unilaterale dell'ipotesi.

Varianza nota

Sia $(Y_1, \ldots, Y_n)$ un campione casuale da $Y \sim \mathcal{N}(\mu, \sigma^2)$, dove si suppone $\sigma > 0$ nota. Si formula l'ipotesi

$$H_0 : \mu = \mu_0$$

che vogliamo verificare sulla base dei dati osservati. Il test si basa ancora sulla quantità

$$Z = \frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}} = \frac{\overline{Y} - \mu + \mu - \mu_0}{\sigma/\sqrt{n}} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} + \frac{\mu - \mu_0}{\sigma/\sqrt{n}} \tag{7.8}$$

da cui deriva

$$Z \sim \mathcal{N}\left(\frac{\mu - \mu_0}{\sigma/\sqrt{n}}, 1\right).$$

Quindi $Z \sim \mathcal{N}(0,1)$ quando H_0 è vera. Si osservi che se $\mu > \mu_0$, allora $\sqrt{n}(\mu - \mu_0)/\sigma > 0$. Ciò implica che, se si ritiene estremi i valori di μ che sono più grandi di μ_0, in corrispondenza di tali valori si avranno valori di Z nella coda destra della distribuzione $\mathcal{N}(0,1)$, mentre quando l'ipotesi nulla è vera, si osserveranno valori di Z ragionevoli per la distribuzione $\mathcal{N}(0,1)$. Ne segue che

(a.) per la verifica di $H_0 : \mu = \mu_0$ a fronte dell'alternativa unilaterale destra $H_1 : \mu > \mu_0$ il valore-p è

$$P\left(Z \geq \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}\right), \tag{7.9}$$

dove $Z \sim \mathcal{N}(0,1)$ e $\bar{y}$ è la media del campione osservato;

(b.) allo stesso modo, se si ipotizza $H_0 : \mu = \mu_0$, a fronte dell'alternativa $H_1 : \mu < \mu_0$, il valore-p si ottiene come

$$P\left(Z \leq \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}\right). \tag{7.10}$$

A seconda del caso considerato, si conclude che i dati sono in disaccordo con l'ipotesi formulata se si ottiene una valore molto piccolo della (7.9) o della (7.10).

Varianza incognita

Con le stesse considerazioni svolte in precedenza, si ammetta ora di non conoscere la varianza σ_0^2. La verifica dell'ipotesi $H_0 : \mu = \mu_0$ contro una alternativa del tipo $\mu < \mu_0$ o $\mu > \mu_0$ si baserà ancora sulla quantità

$$t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}} \tag{7.11}$$

dove $\overline{y}$ e s sono la media e la deviazione standard del campione osservato. Se l'ipotesi alternativa è $H_1 : \mu > \mu_0$ allora il *p-value* è espresso da

$$P(T \geq t) = P\left(T \geq \frac{\bar{y} - \mu_0}{s/\sqrt{n}}\right) = 1 - F_{t_{n-1}}\left(\frac{\bar{y} - \mu_0}{s/\sqrt{n}}\right), \tag{7.12}$$

dove $F_{t_{n-1}}$ è ancora la funzione di ripartizione della t di Student con $n-1$ gradi di libertà.

Analogamente, per la verifica della stessa ipotesi nulla ma con $H_1 : \mu < \mu_0$ si calcolerà la probabilità di osservare, quando l'ipotesi nulla è vera, un valore più piccolo di t:

$$P(T \leq t) = P\left(T \leq \frac{\bar{y} - \mu_0}{s/\sqrt{n}}\right) = F_{t_{n-1}}\left(\frac{\bar{y} - \mu_0}{s/\sqrt{n}}\right), \tag{7.13}$$

essendo T distribuita secondo una t di Student con $n-1$ gradi di libertà.

Esempio 7.5. Una nuova dieta si propone di far perdere almeno 4.5 kg a chi la segue per due mesi. Si sono rilevati i dati per 13 persone, selezionate casualmente, che hanno seguito la dieta per due mesi e si è rilevato che hanno perso

$$3.9, 4.3, 5.6, 5.6, 4.1, 6.5, 3.7, 5.9, 4.3, 3.7, 4.4, 5.9, 5.0\, kg.$$

Assumendo che nella popolazione il peso perso si distribuisca come una gaussiana di media μ, si vuole sottoporre a verifica l'ipotesi $H_0 : \mu = 4.5$, a fronte dell'alternativa che il valore di μ sia inferiore. Si tratta quindi di condurre un test di significatività per la media di una popolazione normale con varianza non nota, stimata mediante $s^2 = 0.916$ calcolata per il campione dato. Dai dati si trova anche $\bar{y} = 4.838$. Il livello di significativit'a osservato è

$$P\left(\frac{\overline{Y} - \mu_0}{S/\sqrt{n}} < \frac{4.838 - 4.5}{0.957/\sqrt{13}}\right) = F_{t_{12}}(1.273)$$

dove $F_{t_{12}}$ indica la funzione di ripartizione della t con 12 gradi di libertà. Dalle tavole si trova che $F_{t_{12}}(1.273)$ è un valore compreso tra 0.85 e 0.9, a supporto dell'ipotesi H_0. ▲

7.3 Verifica di ipotesi e stima intervallare

Come visto nel Capitolo 5, un intervallo di confidenza è un modo di esprimere stima e incertezza insieme. Gli estremi dell'intervallo sono costruiti in modo che la probabilità che l'intervallo contenga il vero valore del parametro sia uguale a un livello prefissato, detto livello di confidenza. Un altro approccio alla verifica di ipotesi è fornito dagli intervalli di confidenza. Si consideri un intervallo di confidenza (T_1, T_2) al livello $1-\alpha$ per μ, con $\alpha \in (0,1)$. Allora se $\mu_0 \notin (T_1, T_2)$ possiamo dedurre che l'evidenza campionaria è contro H_0.

Più in generale, la verifica di una ipotesi tramite intervalli di confidenza risulta equivalente, sotto alcune condizioni, alla verifica di una ipotesi si μ mediante un test di significatività come quelli presentati nelle sezioni precedenti. Vediamo tale equivalenza più in dettaglio con alcuni esempi.

Esempio 7.6 (Modello normale con varianza nota)**.** Si vuole mostrare l'equivalenza tra ottenere un valore-p inferiore a α per $H_0 : \mu = \mu_0$ e verificare che l'intervallo di livello $1-\alpha$ per μ non contiene μ_0. Osserviamo che la disuguaglianza

$$\alpha \leq 2\left[1 - \Phi\left(\left|\frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}\right|\right)\right]$$

è soddisfatta se e solo se

$$\Phi\left(\left|\frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}\right|\right) \leq 1 - \frac{\alpha}{2},$$

e cioè, se e solo se

$$\frac{|\bar{y} - \mu_0|}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2}$$

che implica

$$\mu_0 \in \left[\bar{y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Ciò implica che l'intervallo di confidenza per μ comprende quei valori μ_0 per i quali il valore-p per l'ipotesi $H_0 : \mu = \mu_0$ è maggiore di α. Pertanto, il valore-p basato sulla statistica (7.5) per il test su H_0 sarà inferiore a α se e solo se μ_0 non è nell'intervallo di confidenza per μ. Ad esempio, se decidiamo che per qualsiasi valore-p minore di 0.05 concluderemo che i dati non supportano l'ipotesi nulla, allora potremo trarre la stessa conclusione ogni qual volta l'intervallo di confidenza al 95% per μ non contiene μ_0.

Riprendendo i dati dell'esempio 7.3 l'intervallo di confidenza per μ al 95% ha estremi

$$\bar{y} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 25.3657 \pm 1.96\sqrt{3/10} = [24.29, 26.44].$$

L'intervallo ottenuto non contiene $\mu_0 = 24$ e questo implica che i dati non supportano l'ipotesi H_0. ▲

Esempio 7.7 (Modello bernoulliano)**.** Utilizzando lo stesso ragionamento del test sulla media illustrato nell'esempio precedente possiamo verificare l'ipotesi formulata sul parametro di una distribuzione bernoulliana, disponendo di un campione casuale di ampiezza elevata da $Y \sim Be(p)$, mediante la costruzione di un intervallo per p al livello $1 - \alpha = 0.95$:

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

dove $\hat{p}$ denota la proporzione campionaria.

Se consideriamo i dati dell'esperimento di Mendel illustrati nella sezione introduttiva, si ottiene il seguente intervallo al 95% per p

$$0.262 \pm 1.96\sqrt{\frac{0.262(1-0.262)}{580}} = [0.2262, 0.2978]$$

che contiene $p_0 = 0.25$, e quindi non fornisce alcuna ragione per dubitare dell'ipotesi nulla $H_0 : p = p_0 = 0.25$.

▲

7.3.1 Test sulla varianza di una popolazione normale

Abbiamo finora analizzato il problema di verifica di una ipotesi riguardante la media di una distribuzione considerando il campionamento da un modello Bernoulliano di parametro p o da una popolazione $\mathcal{N}(\mu, \sigma^2)$, dove l'interesse era su p e μ rispettivamente.

In questa sezione consideriamo invece il caso in cui il parametro incognito di interesse su cui fare inferenza è σ^2, varianza della distribuzione $\mathcal{N}(\mu, \sigma^2)$. Si supponga, ad esempio, che da studi precedenti il valore incognito della varianza è molto vicino ad un valore σ_0^2. Disponendo allora di un nuovo

campione, il quesito di interesse è se la variabilità del processo studiato sia cambiata o no, per cui si formula l'ipotesi nulla $H_0 : \sigma^2 = \sigma_0^2$.

Consideriamo un campione casuale $(Y_1, \ldots, Y_n)$ da $Y \sim \mathcal{N}(\mu, \sigma^2)$ dove $\mu \in \mathbb{R}$ e $\sigma > 0$ sono incogniti. Uno stimatore non distorto per σ^2 è la varianza campionaria S^2 e inoltre dalla normalità della popolazione segue che $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Possiamo allora considerare un intervallo al livello $1-\alpha$ per σ^2, espresso dalla (5.3): per ogni $(\mu, \sigma) \in \mathbb{R} \times (0, \infty)$ si ha

$$1 - \alpha = P\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right)$$

dove $\chi^2_{m,q}$ indica il quantile q-esimo della distribuzione chi-quadrato con m gradi di libertà. Allora per verificare l'ipotesi

$$H_0 : \sigma^2 = \sigma_0^2$$

al livello α verificheremo se σ_0^2 si trova nell'intervallo

$$\left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right].$$

Per il calcolo del valore-p per la verifica di ipotesi sulla varianza di una popolazione normale ci si baserà sempre sulla quantità

$$T = \frac{(n-1)S^2}{\sigma_0^2}$$

che, nell'ipotesi nulla, è distribuita secondo una χ^2 con $n-1$ gradi di libertà. Nel caso del test bilaterale, un valore molto piccolo o molto grande della statistica T rappresenta una evidenza campionaria contro $H_0 : \sigma^2 = \sigma_0^2$. Pertanto, se $t = (n-1)s^2/\sigma_0^2$ è il valore osservato della statistica, il valore-p sarà pari a due volte l'area nella coda individuata da t.

Se invece l'ipotesi alternativa è di tipo unilaterale, ad esempio $H_1 : \sigma^2 > \sigma_0^2$, allora se questa è vera, il valore t tenderà ad assumere valori "grandi", poiché s^2 sarà molto più grande di σ_0^2. Ne segue che l'ipotesi nulla viene smentita se il valore-p dato dalla probabilità

$$P(T \geq t) = P\left(T \geq \frac{(n-1)s^2}{\sigma_0^2}\right)$$

risulta essere un valore molto piccolo, essendo $T \sim \chi^2_{n-1}$. In modo analogo, se si verifica l'ipotesi nulla $H_0 : \sigma^2 = \sigma_0^2$ a fronte dell'alternativa $H_1 : \sigma^2 < \sigma_0^2$, allora si rifiuterà H_0 per valori piccoli della statistica T, e quindi il p-value si calcola come la probabilità di ottenere un valore pari a t o un valore ancora più 'piccolo':

$$P(T \leq t) = P\left(T \leq \frac{(n-1)s^2}{\sigma_0^2}\right).$$

Esempio 7.8. Si considerino nuovamente i dati dell'esempio 7.3 e si assuma ora di non sapere che $\sigma^2 = 3$. Per il campione dato con $n = 10$ si ha $s^2 = 2.9106$; inoltre $\chi^2_{9,0.025} = 2.700$ e $\chi^2_{9,0.975} = 19.023$ sono i valori della chi-quadrato con 9 gradi di libertà che lasciano in ciascuna coda una probabilità pari a 0.025. Si ricava perció l'intervallo di confidenza per σ^2 al livello 0.95

$$\left[\frac{9s^2}{\chi^2_{9,0.975}}, \frac{9s^2}{\chi^2_{9,0.025}}\right] = [1.377, 9.702].$$

L'intervallo è piuttosto ampio e indica un grado di incertezza sul parametro σ^2 non trascurabile. Tuttavia, l'intervallo al livello del 95% contiene $\sigma^2 = 3$ e quindi concludiamo che non possiamo rifiutare l'ipotesi fatta.

▲

Esempio 7.9. Un metodo comunemente usato per determinare il calore specifico del bronzo è noto produca misure caratterizzate da uno scarto quadratico medio pari a 0.01. Un nuovo metodo viene verificato su un campione 9 volte e per le misure ottenute si calcola lo scarto quadratico medio campionario che risulta pari a 0.0086. Si supponga che le misure di calore specifico del bronzo nella popolazione abbiano distribuzione normale di media μ non nota. Si può affermare che il nuovo metodo riduce lo scarto quadratico medio?

Consideriamo la verifica dell'ipotesi $H_0 : \sigma^2 = 0.01^2$, a fronte dell'alternativa $\sigma^2 < 0.01^2$. Ponendo $\sigma_0^2 = 0.01^2$ e $n = 9$, sotto l'ipotesi nulla, data la normalità della popolazione, la varianza campionaria ha distribuzione $S^2 \sim \chi^2_{n-1}\sigma_0^2/(n-1)$. Dal campione si trova

$$t = \frac{8}{0.0001} 0.0086^2 = 5.92$$

e quindi il valore-p è $P(T < 5.92)$ essendo $T \sim \chi^2_8$. Dalle tavole si trova che tale probabilità è superiore a 0.25 poiché $P(T < 5.0706) = 0.25$ (usando `R`

si trova $P(T < 5.92) \approx 0.344$). Si conclude che il livello di significatività trovato non porta a dubitare dell'ipotesi nulla, che si può quindi ritenere in accordo con i dati. Non abbiamo quindi gli elementi per affermare che il nuovo metodo riduce lo scarto quadratico medio.

▲

Media nota

Se si suppone di disporre di un campione casuale $(Y_1, \ldots, Y_n)$ da $Y \sim \mathcal{N}(\mu, \sigma^2)$ dove μ è nota e σ^2 è il parametro su cui facciamo inferenza, allora quando l'ipotesi nulla

$$H_0 : \sigma^2 = \sigma_0^2$$

è vera, la quantità

$$\frac{\sum_{i=1}^{n}(Y_i - \mu)^2}{\sigma_0^2}$$

ha distribuzione chi-quadrato con n gradi di libertà (si veda l'esempio 5.4). Possiamo allora costruire un intervallo al livello $1 - \alpha$ per σ^2, espresso da

$$\left[\frac{n\hat{\sigma}^2}{\chi^2_{n,1-\alpha/2}}, \frac{n\hat{\sigma}^2}{\chi^2_{n,\alpha/2}}\right]$$

dove $\hat{\sigma}^2 = \sum_{i=1}^{n}(Y_i - \mu)^2/n$, e condurre la verifica dell'ipotesi nulla verificando se σ_0^2 si trova oppure no nell'intervallo sopra riportato.

7.4 Approccio di Neyman-Pearson

L'approccio di Neyman-Pearson fa riferimento a un sistema d'ipotesi che consiste nell'*ipotesi nulla*, H_0, e nell'*ipotesi alternativa*, H_1: la prima è espressa tramite la specificazione di un insieme di valori, ad esempio $\theta = \theta_0$ (ipotesi semplice), $\theta \geq \theta_0$, $\theta < \theta_0$ (ipotesi composte), essendo θ il parametro che caratterizza il modello descrittivo della popolazione, e θ_0 uno specifico valore del parametro; l'ipotesi H_1 è l'affermazione o la congettura contrapposta.

L'approccio illustrato consiste nel formulare una regola per cui, a seconda del valore assunto dal campione, si decide di accettare o rifiutare l'ipotesi nulla. Detto Ω lo spazio campionario, un test individua una partizione di Ω in due sottoinsiemi: la regione contenente i campioni per i quali la regola impone

di rifiutare H_0, detta **regione di rifiuto** e indicata con R, e il complementare della regione di rifiuto, detta anche **regione di accettazione**, A, con $A \cup R = \Omega$, $A \cap R = \varnothing$. La decisione presa può essere corretta o sbagliata e in particolare distinguiamo due possibili errori (si veda la tabella 7.1):

1. I tipo: rifiutare H_0 quando è vera
2. II tipo: accettare H_0 quando è falsa

Per verificare un sistema d'ipotesi col criterio di Neyman-Pearson si procede come segue:

- Si formulano l'ipotesi nulla e l'ipotesi alternativa (ad esempio, $H_0 : \mu = \mu_0$ e $H_1 : \mu = \mu_1 > \mu_0$);
- Si determina la forma della regione di rifiuto (ad esempio, nel caso di verifica dell'ipotesi al punto precedente, considereremo una regione di rifiuto del tipo $\overline{Y} > k$);
- Si sceglie una particolare regione di rifiuto in modo che la probabilità di errore di I tipo sia al più pari ad α (regione di livello α), con α fissato e piccolo (ad esempio, 0.05, 0.01) per tenere sotto controllo l'errore che si commette rifiutando H_0 quando in realtà è vera.

La probabilità di errore di I tipo, α, è il **livello di significatività** del test. Inoltre, l'errore di II tipo è considerato in un certo senso "meno grave" e la probabilità di commetterlo, β, è determinata dalla scelta di α. Il complemento a 1 della probabilità di errore di II tipo è la probabilià di rifiutare H_0 quando questa è falsa ed è la **potenza del test**.

	H_0 vera	H_0 falsa
Accettazione di H_0	decisione corretta	Errore di secondo tipo
Rifiuto di H_0	Errore di primo tipo	decisione corretta

Tabella 7.1: Errori di I e II tipo nell'approccio di Neyman-Pearson.

Si consideri il sistema di ipotesi semplici per $\theta \in \Theta$ (il modello è completamente specificato sotto H_0 o H_1)

$$H_0 : \theta = \theta_0; \qquad H_1 : \theta = \theta_1$$

che genera un test unilaterale (o ad una coda) nei due casi $\theta_1 > \theta_0$ e $\theta_1 < \theta_0$. Come visto, il test consiste in una procedura con cui decidiamo, alla luce dei dati del campione, se accettare o rifiutare l'ipotesi formulata. Non è quindi possibile commettere entrambi gli errori di I e II tipo contemporaneamente e si ha

$$\begin{aligned}\alpha &= P(\text{errore I tipo}) \\ &= P(\text{rifiutare } H_0, \text{ dato che } H_0 \text{ è vera}) \\ &= P(\text{rifiutare } H_0 | \, \theta = \theta_0)\end{aligned}$$

$$\begin{aligned}\beta &= P(\text{errore II tipo}) \\ &= P(\text{accettare } H_0, \text{ dato che } H_1 \text{ è vera}) \\ &= P(\text{accettare } H_0 | \, \theta = \theta_1)\end{aligned}$$

Ad ogni test possiamo associare una coppia di valori (α, β). Si osservi che non è possibile per un test minimizzare simultaneamente α e β. Innanzitutto, il valore minimo che α e β possono assumere è 0, essendo queste due probabilità. Si supponga ora di adottare la seguente regola di decisione: si accetti l'ipotesi nulla, qualunque sia il dato campionario che si osserva. Per questo test la probabilità dell'errore di I tipo sarà $\alpha = 0$, dal momento che non si commette mai l'errore di rifiutare H_0 quando questa è vera. D'altra parte, a prescindere dal campione osservato, la probabilità dell'errore di II tipo sarà $\beta = 1$, il massimo valore possibile. Se invece si considerasse un test in cui si rifiuta H_0, qualunque sia il campione osservato, allora avremmo $\alpha = 1$ e $\beta = 0$. Nessuno di questi due test rappresenta un buon test poiché, nel minimizzare la probabilità di uno dei due possibili errori, massimizzano la probabilità dell'altro.

La *potenza* del test per un valore $\theta_1 \neq \theta_0$ varia di conseguenza, essendo pari a

$$\pi = 1 - \beta = P(\text{rifiutare } H_0 | \, \theta = \theta_1).$$

In particolare si vedrà che la potenza di un test risulta influenzata dal livello di significatività prescelto α, dalla specificazione dell'ipotesi alternativa e dalla numerosità del campione.

Regioni di rifiuto: media della normale con varianza nota

Sia $(Y_1, Y_2, \ldots, Y_n)$ un campione proveniente da Y distribuita normalmente, con varianza nota e pari a σ^2. Si consideri il sistema di ipotesi semplici

$$\begin{cases} H_0 & : \mu = \mu_0 \\ H_1 & : \mu = \mu_1 (> \mu_0) \end{cases} \tag{7.14}$$

Nella sezione 7.2.2 si è visto che possiamo basare il problema di verifica di $H_0 : \mu = \mu_0$ sulla statistica $\bar{Y}$ che, sotto H_0, ha distribuzione $\overline{Y} \sim \mathcal{N}(\mu_0, \sigma^2/n)$.

Definiamo la regione di rifiuto

$$R = \{\overline{Y} > k\}$$

dove k denota una soglia critica superata la quale l'ipotesi nulla viene rifiutata. Per determinare un k fissiamo la probabilità dell'errore di I tipo:

$$\begin{aligned} \alpha &= P(\text{rifiutare } H_0 | \, \mu = \mu_0) \\ &= P(\overline{Y} > k | \, \mu = \mu_0) \\ &= 1 - \Phi\left(\frac{k - \mu_0}{\sigma/\sqrt{n}}\right) \end{aligned}$$

essendo $\overline{Y} \sim \mathcal{N}(\mu_0, \sigma^2/n)$ sotto H_0. Ne deriva che

$$k = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

e quindi, in definitiva,

$$R = \{\overline{Y} > \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}\}. \tag{7.15}$$

La regione di rifiuto con probabilità di errore di I tipo al più pari ad α è detta regione di livello α. Si noti che per ottenere la regione di rifiuto si fa riferimento alla distribuzione campionaria nell'ipotesi nulla, cioè $\overline{Y} \sim \mathcal{N}(\mu_0, \sigma^2/n)$. Sotto H_1 si ha invece $\overline{Y} \sim \mathcal{N}(\mu_1, \sigma^2/n)$ e la probabilità di errore di II tipo è

$$\begin{aligned} \beta &= P(\text{accettare } H_0 | \, \mu = \mu_1) \\ &= P(\overline{Y} \leq k | \, \mu = \mu_1) \\ &= \Phi\left(\frac{k - \mu_1}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{1-\alpha}\right) \end{aligned}$$

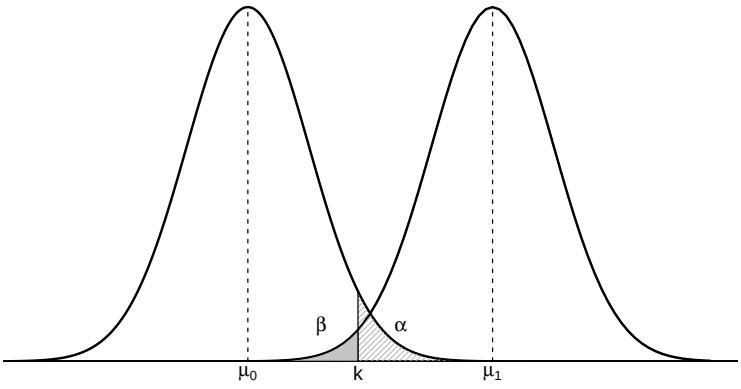


Figura 7.1: Probabilità α e β per la verifica dell'ipotesi $H_0 : \mu = \mu_0$ contro $H_1 : \mu = \mu_1 > \mu_0$.

La potenza è data dal complemento a 1 della probabilità di errore di II tipo

$$\pi = 1 - \beta = 1 - \Phi\left(\frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} + z_{1-\alpha}\right)$$

Nella figura 7.1 dove entrambe le ipotesi sono semplici è stata evidenziata graficamente la regione di rifiuto dell'ipotesi H_0 e l'area (in grigio chiaro) corrispondente alla probabilità dell'errore di I tipo; l'area in grigio scuro corrisponde alla probabilità dell'errore di II tipo, infine la potenza risulta graficamente espressa dall'area sottesa alla curva di destra relativa all'intervallo $(k, +\infty)$. Si osservi che al crescere di k diminuisce la probabilità di errore di I tipo e aumenta la probabilità di errore di II tipo. Viceversa, si vede chiaramente come l'incremento di α comporta l'allargamento della regione di rifiuto, determinando una riduzione della probabilità dell'errore di II tipo e di conseguenza un aumento della potenza del test. Il *trade-off* tra α e β viene illustrato nell'esempio che segue.

Esempio 7.10. Sia $(Y_1, Y_2, \ldots, Y_9)$ un campione tratto da una normale di varianza nota e pari a $\sigma^2 = 1$. Vogliamo verificare l'ipotesi semplice $H_0 : \mu = 2$ contro l'alternativa $H_1 : \mu = 3$. Per decidere sull'ipotesi nulla guarderemo al valore osservato al valore osservato della media campionaria $\overline{Y} = (1/9)\sum_{i=1} Y_i$. Infatti, se H_0 è vera, ci aspettiamo un valore di $\bar{y}$ più

vicino a 2 che a 3, mentre se H_0 è falsa, sarà l'opposto. Consideriamo quindi il test che accetta H_0 se $\bar{y} \leq k$ e rifiuta H_0 se $\bar{y} > k$, con k tale che la probabilità dell'errore di I tipo sia pari ad un α fissato:

$$\alpha = P(\overline{Y} > k|H_0) = 1 - \Phi\left(\frac{k-2}{1/3}\right) = \Phi(3(2-k))$$

essendo $\overline{Y}$ normalmente distribuita con media $\mu = \mu_0 = 2$ e varianza $\sigma^2 = 1/9$ se H_0 è vera. Inoltre, si ha

$$\beta = P(\overline{Y} \leq k|H_1) = \Phi\left(\frac{k-3}{1/3}\right) = \Phi(3(k-3))$$

poiché, sotto H_1, $\overline{Y}$ è normalmente distribuita con media $\mu = \mu_1 = 3$ e varianza $\sigma^2 = 1/9$. È facile vedere come la scelta di k influenzi i valori di α e β per questo test. Si riportano, a titolo di esempio quattro possibili valori di k in ordine crescente a cui corrispondono valori di α e β decrescenti e crescenti, rispettivamente:

k	α	β
2.2	0.274	0.008
2.4	0.115	0.036
2.6	0.036	0.115
2.8	0.008	0.274

Viceversa, fissato $\alpha = 0.05$, si ottiene $k = 2 - z_{0.05}/3 = 2.548$ e la potenza del test al livello del 5% è

$$\pi = 1 - \beta = 1 - \Phi(3(2.548 - 3)) = 0.912.$$

▲

Si consideri il sistema di ipotesi bilaterale

$$\begin{cases} H_0 & : \mu = \mu_0 \\ H_1 & : \mu \neq \mu_0 \end{cases} \tag{7.16}$$

Per uno specifico valore di α, si rifiuta H_0 per valori di $\bar{y}$ molto distanti da μ_0 nell'una o nell'altra direzione. Pertanto, è ragionevole considerare una regione di rifiuto del tipo

$$R = \{\overline{Y} < k_1 \vee \overline{Y} > k_2\}$$

dove k_1 e k_2 sono tali che

$$\alpha/2 = P(\overline{Y} < k_1 | \mu = \mu_0) = P(\overline{Y} > k_2 | \mu = \mu_0)$$

da cui

$$k_1 = \mu_0 + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \quad k_2 = \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}$$

Pertanto la regione di rifiuto di livello α per la verifica dell'ipotesi $H_0 : \mu = \mu_0$ contro l'ipotesi $H_1 : \mu \neq \mu_0$ può essere scritta come $R = \{\overline{Y} < \mu_0 - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \vee \overline{Y} > \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\}$ o, equivalentemente, come

$$R = \left\{ \bar{y} : \left| \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{1-\alpha/2} \right\}.$$

Si noti che nel caso in esame, sotto l'ipotesi alternativa, μ può assumere qualsiasi valore più piccolo o più grande di μ_0, e quindi anche la probabilità β e la potenza non sono costanti ma variano con μ. Si introduce quindi la **funzione di potenza** del test, che corrisponde alla probabilità che la $n-$pla campionaria appartenga alla regione di rifiuto R sotto l'ipotesi alternativa $H_1 : \mu \neq \mu_0$. Essa può essere calcolata come segue:

$$\begin{aligned}
\pi(\mu) &= P(R | \mu \neq \mu_0) \\
&= P\left(\overline{Y} < \mu_0 - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \middle| \mu \neq \mu_0 \right) + P\left(\overline{Y} > \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \middle| \mu \neq \mu_0 \right) \\
&= P\left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} < \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{1-\alpha/2} \right) + P\left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} > \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha/2} \right) \\
&= \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{1-\alpha/2} \right) + 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha/2} \right) \qquad (7.17)
\end{aligned}$$

dove si è utilizzato il fatto che, sotto H_1, $Y \sim \mathcal{N}(\mu, \sigma^2)$, con σ nota e $\mu \neq \mu_0$. La funzione di potenza per la verifica d'ipotesi (7.16) è illustrata con un esempio in figura 7.2. Si osservi che $\pi(\mu)$ è crescente se $\mu > \mu_0$, e decrescente se $\mu < \mu_0$, ha il minimo in $\mu = \mu_0$ e $\pi(\mu) \to 1$ per $\mu \to \pm\infty$.

Ipotesi composte

La procedura di verifica d'ipotesi con approccio del tipo Neyman-Pearson è stata finora limitata al caso di una ipotesi nulla semplice, che quindi specifica

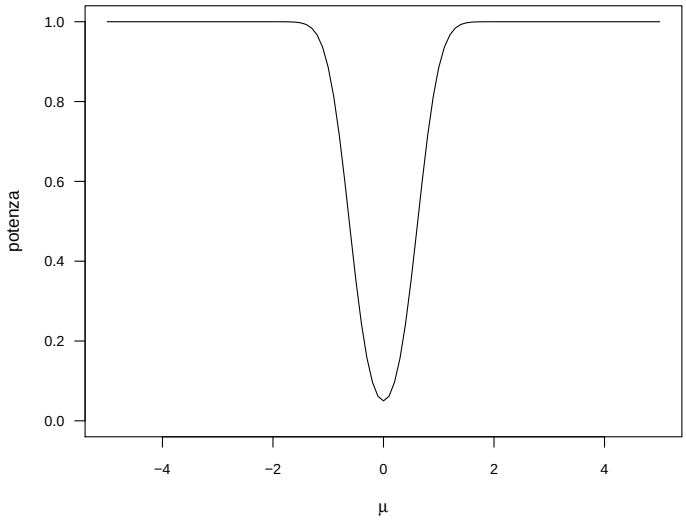


Figura 7.2: Funzione di potenza $\pi(\mu)$ del test sulla media con ipotesi nulla $H_0 : \mu = 0$ contro $H_1 : \mu \neq 0$, $n = 10$, $\sigma = 1$, $\alpha = 0.05$.

completamente la distribuzione della variabile Y sotto H_0. Si consideri ora il caso più generale in cui sia l'ipotesi nulla che l'alternativa sono *composte*.

Per il modello normale con varianza nota, si vuole verificare

$$\begin{cases} H_0 & : \mu \leq \mu_0 \\ H_1 & : \mu > \mu_0 \end{cases} \tag{7.18}$$

dove l'ipotesi nulla è del tipo $H_0 : \theta \in \Theta_0$, con Θ_0 sottoinsieme dello spazio parametrico Θ che contiene almeno una coppia di valori del parametro θ. Qui il parametro è la media della popolazione normale. Definiamo allora una regione di rifiuto R di livello α, cioè un sottoinsieme dello spazio campionario $R \subset \Omega$ che soddisfa

$$P((Y_1, Y_2, \ldots, Y_n) \in R | H_0) \leq \alpha$$

dove α è il livello di significatività del test. La zona di rifiuto R si basa ancora

sulla media campionaria $\overline{Y}$, in particolare R è data da

$$R = \left\{ \overline{Y} > \mu_0 + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right\} = \{ Z > z_{1-\alpha} \}$$

(che è equivalentemente alla (7.15)) e la probabilità di errore di I tipo è

$$\begin{aligned} \alpha(\mu) =& P(\overline{Y} \in R | H_0) \\ =& P(\overline{Y} > \mu_0 + z_{1-\alpha}\sigma/\sqrt{n} | \mu \leq \mu_0) \\ =& 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha} \right) \end{aligned} \tag{7.19}$$

La (7.19) è una funzione crescente di μ che assume il suo valore massimo per $\mu = \mu_0$:

$$1 - \Phi\left(\frac{\mu_0 - \mu_0}{\sigma/\sqrt{n}} + z_{1-\alpha} \right) = 1 - \Phi(z_{1-\alpha}) = \alpha.$$

Ne segue che l'ipotesi nulla composta $H_0 : \mu \leq \mu_0$ è riconducibile all'ipotesi $H_0 : \mu = \mu_0$, e che il rifiuto di $H_0 : \mu = \mu_0$ implica anche il rifiuto dell'ipotesi $H_0 : \mu \leq \mu_0$. Infatti, per quanto appena visto si ha

$$\begin{aligned} P(\overline{Y} > k | \mu = \mu_0) = P\left(Z > \frac{k - \mu_0}{\sigma/\sqrt{n}} \right) &> P(\overline{Y} > k | \mu = \mu' < \mu_0) \\ &= P\left(Z > \frac{k - \mu'}{\sigma/\sqrt{n}} \right) \end{aligned}$$

in quanto

$$\frac{k - \mu_0}{\sigma/\sqrt{n}} < \frac{k - \mu'}{\sigma/\sqrt{n}}$$

per qualsiasi valore $\mu' < \mu_0$.

Si procede in modo analogo per la verifica del sistema di ipotesi

$$\begin{cases} H_0 & : \mu \geq \mu_0 \\ H_1 & : \mu < \mu_0 \end{cases} \tag{7.20}$$

dove si dimostra che

$$R = \left\{ \overline{Y} < \mu_0 - z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right\} = \{ Z < -z_{1-\alpha} \} = \{ Z < z_\alpha \}$$

è una regione di rifiuto di livello α. Una sintesi dei casi relativi alla normale con varianza nota è fornita nella Tabella 7.2.

Se $H_0 : \mu = \mu_0$ e l'ipotesi alternativa è composta del tipo $H_1 : \mu > \mu_0$, allora la potenza del test è funzione di μ, che può assumere qualsiasi valore maggiore di μ_0. La funzione di potenza del test si ottiene come segue

$$\begin{aligned}\pi(\mu) &= 1 - \beta(\mu) = 1 - P\left(\overline{Y} \leq \mu_0 + z_{1-\alpha}\sigma/\sqrt{n} | \mu > \mu_0\right) \\ &= 1 - P\left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) \\ &= 1 - \Phi\left(z_{1-\alpha} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right). \end{aligned} \tag{7.21}$$

Si osservi che il minimo della potenza si ottiene per $\mu = \mu_0$ e corrisponde ad α, inoltre la funzione $\pi(\mu)$ è crescente e tende ad 1 se μ tende a infinito.

Analogamente, quando l'ipotesi alternativa è del tipo $H_1 : \mu < \mu_0$, la funzione di potenza del test con regione di rifiuto di livello α è data da

$$\begin{aligned}\pi(\mu) &= 1 - \beta(\mu) = 1 - P\left(\overline{Y} \geq \mu_0 - z_{1-\alpha}\sigma/\sqrt{n} | \mu < \mu_0\right) \\ &= P\left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \leq -z_{1-\alpha} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}\right) \\ &= \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{1-\alpha}\right). \end{aligned} \tag{7.22}$$

Tale funzione cresce al diminuire di μ ed ha un andamento simmetrico rispetto alla funzione di potenza del test con alternativa $H_1 : \mu > \mu_0$.

Regioni di rifiuto: media della normale con varianza incognita

Nel caso di una verifica su una media con varianza non nota, si usa il risultato già discusso, in base al quale se $\mu = \mu_0$, allora la distribuzione di $\sqrt{n}(\overline{Y} - \mu_0)/S$ è

$$T = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

da cui le regioni di rifiuto illustrate in Tabella 7.3.

H_0	H_1	R
$\mu \leq \mu_0$	$\mu > \mu_0$	$\bar{y} > \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\bar{y} < \mu_0 - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\left\lvert \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \right\rvert > z_{1-\alpha/2}$

Tabella 7.2: Regioni di rifiuto per la verifica delle ipotesi sulla media di una normale con varianza nota.

H_0	H_1	R
$\mu \leq \mu_0$	$\mu > \mu_0$	$\bar{y} > \mu_0 + \frac{s}{\sqrt{n}} t_{n-1,1-\alpha}$
$\mu \geq \mu_0$	$\mu < \mu_0$	$\bar{y} < \mu_0 - \frac{s}{\sqrt{n}} t_{n-1,1-\alpha}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\left\lvert \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \right\rvert > t_{n-1,1-\alpha/2}$

Tabella 7.3: Regioni di rifiuto per la verifica delle ipotesi sulla media di una normale con varianza incognita.

Esempio 7.11. Per controllare se il peso dichiarato su una confezione risponda al vero, un pastificio analizza il peso di un campione $n = 10$ confezioni di pasta. Le confezioni hanno peso dichiarato 500 g, ed è ragionevole assumere che il peso di una confezione sia una variabile aleatoria normale di varianza nota pari a $\sigma^2 = 42.5$. Per controllare che la macchina non generi confezioni con peso diverso, si verifica se il peso medio del campione si discosta troppo da 500, nel qual caso si sospende la produzione.

Per decidere quando il peso medio è troppo diverso da 500 si impiega una regione di rifiuto al 5%. Il sistema di ipotesi è

$$H_0 : \mu = 500 \qquad H_1 : \mu \neq 500$$

quindi la regione di rifiuto è

$$\left|\frac{\bar{y}-500}{\sqrt{42.5/10}}\right| > z_{0.975} = 1.96.$$

Si supponga che per il campione da 10 unità l'ufficio qualità rilevi $\bar{y} = 503.7$. Allora, si ha

$$\frac{503.7-500}{\sqrt{42.5/10}} = 1.79$$

e quindi l'ipotesi nulla non viene rifiutata al livello del 5%.

Il livello di significatività osservato per il test (valore-p) è

$$2\left[1-\Phi\left(\left|\frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}\right|\right)\right] = 2(1-\Phi(1.795)) = 0.073$$

Se la macchina produce confezioni con peso medio 495, qual è la probabilità che la produzione venga interrotta? Si tratta della probabilità di rifiutare supponendo pari a 495 la media del processo:

$$\begin{aligned} P(\overline{Y} \in R|\mu = 495) &= 1 - P\left(-1.96 < \frac{\overline{Y}-\mu_0}{\sigma/\sqrt{n}} < 1.96|\mu = 495\right) \\ &= 1 + \Phi\left(\frac{500-495}{\sqrt{42.5/10}} - 1.96\right) - \Phi\left(\frac{500-495}{\sqrt{42.5/10}} + 1.96\right) \\ &\approx 0.68 \end{aligned}$$

Si osservi che la probabilità ottenuta rappresenta la potenza del test, al livello del 5%, quando $\mu = 495$.

Fissata la probabilità di errore di I tipo $\alpha = 0.05$, la probabilità di errore del II tipo è dunque

$$\beta = 1 - 0.68 = 0.32.$$

Si calcolino nuovamente le quantità ottenute variando la dimensione campionaria, in particolare sia $n = 30$. La regione di accettazione al 5% diventa

$$500 \pm 1.96(\sqrt{42.5/30}) = 500 \pm 1.96(1.19) \rightarrow [497.67, 502.33]$$

Se $\mu = 495$, la potenza del test è

$$1 + \Phi\left(\frac{500-495}{\sqrt{42.5/30}} - 1.96\right) - \Phi\left(\frac{500-495}{\sqrt{42.5/30}} + 1.96\right) \approx 0.99$$

da cui si ottiene un valore più piccolo di quello ottenuto in precedenza per la probabilità dell'errore di II tipo, $\beta = 1 - 0.99 = 0.01$.

Si evince che fissata la probabilità di errore di I tipo, aumentando la numerosità campionaria diminuisce la probabilità di errore di II tipo.

Infine, si consideri lo stesso sistema di ipotesi nel caso in cui non sia nota la varianza della popolazione. Se $n = 10$, la regione di rifiuto diventa

$$\left| \frac{\bar{y} - 500}{s/\sqrt{10}} \right| > t_{9,0.975} = 2.26$$

Se, ad esempio, si osservasse un campione con $s^2 = 42.5$ (pari alla varianza nota nella prima trattazione), si rifiuterebbe H_0 se il peso medio delle 10 confezioni cadesse fuori dall'intervallo

$$500 \pm 2.26(2.062)$$

che rappresenta un intervallo più ampio di quello ottenuto supponendo $\sigma^2 = 42.5$ nota.

▲

Regioni di rifiuto: varianza della normale

Sia ancora $Y \sim \mathcal{N}(\mu, \sigma^2)$, con σ^2 incognita; la quantità μ può essere o non essere conosciuta. Si supponga di disporre di un campione casuale di ampiezza n da Y. Vogliamo verificare ipotesi su σ^2. Consideriamo il sistema di ipotesi

$$\begin{cases} H_0 & : \sigma^2 \leq \sigma_0^2 \\ H_1 & : \sigma^2 > \sigma_0^2 \end{cases} \tag{7.23}$$

Se μ è incognita, possiamo determinare un test usando la statistica già vista in sezione 7.3.1, $V = \sum(Y_i - \overline{Y})^2/\sigma_0^2 = (n-1)S^2/\sigma_0^2$. V sarà per $\sigma^2 > \sigma_0^2$ più grande che per $\sigma^2 \leq \sigma_0^2$; quindi un test ragionevole è quello che consiste nel rifiutare H_0 per V grande. Se $\sigma^2 = \sigma_0^2$ allora V ha una distribuzione chi-quadrato con $n-1$ gradi di libertà e quindi $P(V > \chi^2_{n-1,1-\alpha}) = \alpha$, dove $\chi^2_{n-1,1-\alpha}$ è il quantile di ordine $1-\alpha$ di una distribuzione chi-quadrato con $n-1$ gradi di libertà. Pertanto definiamo la seguente regione di rifiuto

$$R = \left\{ V > \chi^2_{n-1,1-\alpha} \right\} = \left\{ \sum (Y_i - \overline{Y})^2 > \sigma_0^2 \chi^2_{n-1,1-\alpha} \right\}$$

Osservato un campione $(y_1, y_2, \ldots, y_n)$ e fissato un livello $\alpha \in (0,1)$, se risulta $\sum(y_i - \bar{y})^2 > \sigma_0^2 \chi^2_{n-1,1-\alpha}$ per il campione, allora si deciderà di rifiutare $H_0 : \sigma^2 \leq \sigma_0^2$ al livello $\alpha 100\%$.

Con ragionamenti analoghi si perviene alle regioni di rifiuto illustrate in Tabella 7.4.

H_0	H_1	R
$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\sum (y_i - \bar{y})^2 > \sigma_0^2 \chi^2_{n-1,1-\alpha}$
$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\sum (y_i - \bar{y})^2 < \sigma_0^2 \chi^2_{n-1,\alpha}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\sum (y_i - \bar{y})^2 < \sigma_0^2 \chi^2_{n-1,\alpha/2}$ o $\sum (y_i - \bar{y})^2 > \sigma_0^2 \chi^2_{n-1,1-\alpha/2}$

Tabella 7.4: Regioni di rifiuto per la verifica delle ipotesi sulla varianza di una normale.

Regioni di rifiuto per la verifica di ipotesi su una proporzione

Consideriamo una popolazione Bernoulliana di parametro p. La verifica di ipotesi sulla media della popolazione si basa, come già discusso in 7.2.1, sulla proporzione campionaria $\hat{p} = \overline{Y}$. Sappiamo che se $p = p_0$, allora è approssimativamente vero che

$$Z = \frac{\overline{Y} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim \mathcal{N}(0,1) \tag{7.24}$$

(se l'ampiezza del campione è sufficientemente elevata). Si consideri allora il sistema di ipotesi

$$\begin{cases} H_0 & : p \geq p_0 \\ H_1 & : p < p_0 \end{cases} \tag{7.25}$$

H_0	H_1	R
$p \leq p_0$	$p > p_0$	$\hat{p} > p_0 + z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}$
$p \geq p_0$	$p < p_0$	$\hat{p} < p_0 - z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}$
$p = p_0$	$p \neq p_0$	$\left\lvert \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \right\rvert > z_{1-\alpha/2}$

Tabella 7.5: Regioni di rifiuto per la verifica delle ipotesi sul parametro p di una Bernoulliana.

Fissato un livello di significatività, α, la zona di rifiuto per il test sarà del tipo $\{\hat{p} < k\}$, dove k deve essere tale che

$$\alpha \geq P(\overline{Y} < k | H_0) = \Phi\left(\frac{k - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right)$$

da cui, risolvendo l'uguaglianza in k, si trova

$$k = p_0 - z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}$$

Se ne deduce che

$$R = \{Z < -z_{1-\alpha}\} = \left\{\overline{Y} < p_0 - z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}\right\}$$

è una regione di rifiuto di livello α.

Le regioni di rifiuto per il caso di una proporzione sono riassunte in tabella 7.5. Un esempio della regione di rifiuto per il test unilaterale destro è illustrato in figura 7.3.

7.4.1 Lemma di Neyman-Pearson

Si è già accennato il problema del *trade-off* che esiste tra la probabilità dell'errore di tipo I e la probabilità dell'errore di tipo II per la verifica di

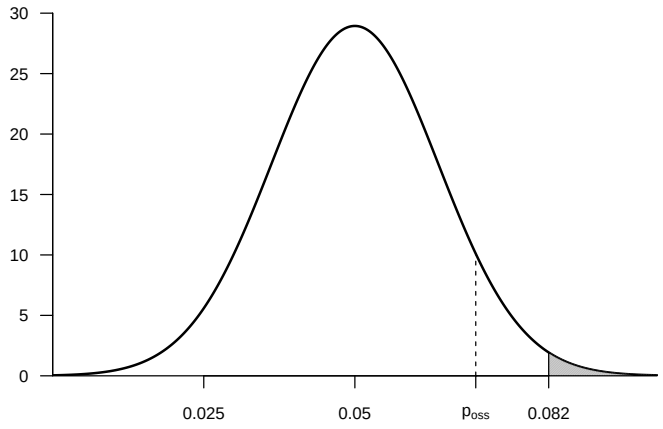


Figura 7.3: Valore osservato e regione di rifiuto al livello 1% per la verifica d'ipotesi $H_0 : p = 0.05$ contro $H_1 : p > p_0 = 0.05$ basata su un campione di ampiezza $n = 250$. L'area evidenziata è pari al livello di significatività $\alpha = 0.01$.

un'ipotesi semplice $H_0 : \theta = \theta_0$ contro un'alternativa pure essa semplice sul generico parametro θ, $H_1 : \theta = \theta_1$. Il problema che si affronta in questa sezione è quello di individuare un 'buon' criterio per scegliere una regione di rifiuto che possa rappresentare un compromesso tra le variazioni di α e quelle di β.

Una strategia classica consiste nel cercare un test che presenti una probabilità dell'errore di II tipo, β, più piccola rispetto a tutti gli altri test con livello di significatività α fissato. Il altri termini, si potrebbe scegliere un livello α piccolo (ad esempio 0.05, 0.01, 0.001) per controllare il rischio dell'errato rifiuto di H_0 (errore I tipo), e poi individuare tra i test che soddisfano tale requisito quello che ha maggiore potenza. È opportuno introdurre la seguente definizione.

Definizione 7.1 (Test più potenti per ipotesi semplici)**.** Si definisce test più potente al livello di significatività α $(0 < \alpha < 1)$ un test per la verifica di $H_0 : \theta = \theta_0$ contro $H_1 : \theta = \theta_1$ con regione di rifiuto R che soddisfa

(i) $P((Y_1, Y_2, \ldots, Y_n) \in R | H_0) = \alpha$

(ii) $\pi_R(\theta_1) \geq \pi_{R'}(\theta_1)$ per qualsiasi altra R' di livello α prefissato

Sia Ω lo spazio campionario e $\tilde{y} = (y_1, y_2, \ldots, y_n)$ il campione osservato di n elementi. La funzione di verosimiglianza sotto l'ipotesi nulla H_0, $L(\tilde{y}; \theta_0) = L(\theta_0)$, corrisponde alla probabilità di osservare $\tilde{y}$ quando $\theta = \theta_0$, cioè quando è vera H_0; analogamente, la funzione di verosimiglianza sotto l'alternativa H_1, $L(\tilde{y}; \theta_1) = L(\theta_1)$, è la probabilità di osservare $\tilde{y}$ assumendo che il vero valore del parametro sia θ_1. Sembra perciò intuitivamente ragionevole costruire il test sulla base del confronto tra $L(\theta_0)$ ed $L(\theta_1)$.

Il teorema che segue fornisce un metodo per individuare una partizione di Ω in regione di rifiuto e regione di accettazione, tale che il test così individuato rappresenta il test più potente al livello di significatività α, per ogni α fissato.

Teorema 7.1 (Lemma di Neyman-Pearson)**.** Sia $(y_1, y_2, \ldots, y_n)$ un campione casuale osservato da Y che è distribuita secondo il modello $f(y; \theta)$, $\theta \in \Theta$, e si voglia verificare $H_0 : \theta = \theta_0$ contro $H_1 : \theta = \theta_1$. Se $L(\theta)$ è la funzione di verosimiglianza di θ, allora il test più potente al livello di significatività α per H_0 contro H_1 ha regione di rifiuto R^* che soddisfa

$$L(\theta_1) \geq kL(\theta_0)$$

dove k è una costante positiva tale che $P((Y_1, Y_2, \ldots, Y_n) \in R^* | H_0) = \alpha$.

Dimostrazione. Sia R^* la regione di rifiuto di un test con livello di significatività α. Se, per un fissato α, è unica la regione di rifiuto R^* di livello α, allora R^* è anche la regione di rifiuto del test più potente al livello di significatività α per H_0 contro H_1, poiché essa è l'unica che fornisce un certo valore della potenza π.

Si assuma l'esistenza di almeno un'altra regione R corrispondente ad un altro test tale che $P((Y_1, Y_2, \ldots, Y_n) \in R|H_0) \leq \alpha$. Dato un generico sottoinsieme $I \subset \Omega$, si porrà sinteticamente $P(I|H_0) = P((Y_1, Y_2, \ldots, Y_n) \in I|H_0)$ e $P(I|H_1) = P((Y_1, Y_2, \ldots, Y_n) \in I|H_1)$.

Si illustra la dimostrazione nel caso continuo; nel caso di distribuzioni discrete si segue un procedimento analogo. Occorre dimostrare che la potenza del test con regione di rifiuto R^* è almeno pari alla potenza del test con regione di rifiuto R, ovvero $\pi_{R^*}(\theta_1) \geq \pi_R(\theta_1)$. Per un campione casuale $\tilde{y} = (y_1, y_2, \ldots, y_n)$, le funzioni di verosimiglianza calcolate sotto le due ipotesi sono

$$L(\theta_0) = L(\tilde{y}; \theta_0) = \prod_{i=1}^{n} f(y_i; \theta_0)$$
$$L(\theta_1) = L(\tilde{y}; \theta_1) = \prod_{i=1}^{n} f(y_i; \theta_1)$$

e la regione di rifiuto individuata dal Lemma è tale che

$$L(\theta_1) \geq kL(\theta_0), \text{ per } \tilde{y} \in R^*; \qquad L(\theta_1) < kL(\theta_0), \text{ per } \tilde{y} \in \bar{R}^*$$

dove $\bar{R}^*$ è il complementare di R^*. Inoltre possiamo scrivere

$$\pi_{R^*}(\theta_1) = P(R^*|H_1) = \int_{R^*} \cdots \int L(\theta_1)\, dy_1 dy_2 \ldots dy_n = \int_{R^*} L(\theta_1) d\tilde{y}$$
$$\pi_R(\theta_1) = P(R|H_1) = \int_{R} \cdots \int L(\theta_1)\, dy_1 dy_2 \ldots dy_n = \int_{R} L(\theta_1) d\tilde{y}$$

Osserviamo che per due qualsiasi regioni di rifiuto R^* e R incluse nello spazio campionario valgono le relazioni seguenti:

$$R^* = (R^* \cap R) \cup (R^* \cap \bar{R}); \qquad R = (R \cap R^*) \cup (R \cap \bar{R}^*)$$

da cui si evince la (eventuale) regione comune alle due regioni di rifiuto, $R^* \cap R$ (si veda la figura 7.4). Poiché $(R^* \cap \bar{R}) \subset R^*$, la disuguaglianza $L(\theta_1) \geq kL(\theta_0)$ è valida anche per $\tilde{y} \in (R^* \cap \bar{R})$. Analogamente, poiché $(R \cap \bar{R}^*) \subset \bar{R}^*$, la disuguaglianza $L(\theta_1) < kL(\theta_0)$ e la sua opposta $-L(\theta_1) > -kL(\theta_0)$ sono valide anche per $\tilde{y} \in (R \cap \bar{R}^*)$. Allora la differenza tra le potenze calcolate sulle due regioni di rifiuto che presentano $(R^* \cap R)$ in comune è espressa da

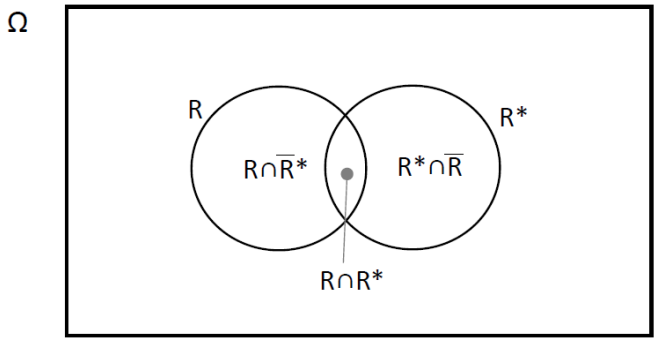


Figura 7.4: Regioni di rifiuto R ed R^* in Ω.

$$\begin{aligned}
\pi_{R^*} - \pi_R &= \int_{R^*} L(\theta_1)d\tilde{y} - \int_R L(\theta_1)d\tilde{y} = \int_{R^*\cap\bar{R}} L(\theta_1)d\tilde{y} - \int_{R\cap\bar{R}^*} L(\theta_1)d\tilde{y} \\
&\geq k\int_{R^*\cap\bar{R}} L(\theta_0)d\tilde{y} - k\int_{R\cap\bar{R}^*} L(\theta_0)d\tilde{y} \\
&= k\left[\int_{R^*\cap\bar{R}} L(\theta_0)d\tilde{y} + \int_{R^*\cap R} L(\theta_0)d\tilde{y}\right] \\
&\qquad - k\left[\int_{R\cap\bar{R}^*} L(\theta_0)d\tilde{y} + \int_{R^*\cap R} L(\theta_0)d\tilde{y}\right] \\
&= k\int_{R^*} L(\theta_0)d\tilde{y} - k\int_R L(\theta_0)d\tilde{y} = kP(R^*|H_0) - kP(R|H_0)
\end{aligned}$$

Quindi risulta

$$\pi_{R^*}(\theta_1) - \pi_R(\theta_1) \geq kP(R^*|H_0) - kP(R|H_0) \geq 0 \tag{7.26}$$

essendo $P(R^*|H_0) = \alpha$ e $P(R|H_0) \leq \alpha$. La diseguaglianza (7.26) indica che la potenza del test avente zona di rifiuto R^* è almeno uguale a quella del test con zona di rifiuto R, il che dimostra il teorema.

□

Il principio del Lemma appena dimostrato è il confronto basato sulla verosimiglianza: si preferisce quell'ipotesi che risulta k volte più plausibile il

termini di verosimiglianza, dove la costante k è determinata in modo tale che il rischio di commettere l'errore di I tipo, ritenuto più grave, sia pari ad un livello α prefissato. È importante osservare che

- non sempre esistono k e R^* che soddisfano le condizioni del Lemma, in tal caso il Teorema 7.1 non fornisce un test più potente al livello di significatività α;
- spesso, in pratica, non è necessario calcolare k e R^*, la disuguaglianza $(L(\theta_0)/L(\theta_1)) \leq k$ può essere trasformata in una disuguaglianza equivalente che è più facile da trattare;
- il valore numerico della costante k viene determinato dal livello α e il vincolo $(L(\theta_0)/L(\theta_1)) \leq k$ determina la forma della regione di rifiuto.

Esempio 7.12. Sia $(Y_1, \ldots, Y_n)$ un campione casuale tratto da $Y \sim \mathcal{N}(\mu, \sigma^2)$, dove si considera σ^2 nota. Si vuole determinare il test più potente per la verifica di ipotesi $H_0 : \mu = \mu_0$ contro l'alternativa $H_1 : \mu = \mu_1$, $\mu_1 > \mu_0$. Valutiamo il *rapporto di verosimiglianza* sotto le due ipotesi

$$\begin{aligned}
\frac{L(\mu_1)}{L(\mu_0)} &= \frac{\prod_{i=1}^n (\sigma\sqrt{2\pi})^{-1} e^{-(y_i-\mu_1)^2/2\sigma^2}}{\prod_{i=1}^n (\sigma\sqrt{2\pi})^{-1} e^{-(y_i-\mu_0)^2/2\sigma^2}} \\
&= \frac{(\sigma\sqrt{2\pi})^{-n} e^{-\sum_i (y_i-\mu_1)^2/2\sigma^2}}{(\sigma\sqrt{2\pi})^{-n} e^{-\sum_i (y_i-\mu_0)^2/2\sigma^2}} \\
&= \frac{(\sigma\sqrt{2\pi})^{-n} e^{-\sum_i y_i^2/2\sigma^2} e^{-(-2\mu_1 \sum_i y_i + n\mu_1^2)/2\sigma^2}}{(\sigma\sqrt{2\pi})^{-n} e^{-\sum_i y_i^2/2\sigma^2} e^{-(-2\mu_0 \sum_i y_i + n\mu_0^2)/2\sigma^2}} \\
&= e^{(\mu_1-\mu_0)\sum_i y_i/\sigma^2 - n(\mu_1^2-\mu_0^2)/2\sigma^2}
\end{aligned}$$

Quindi la disuguaglianza $L(\mu_1)/L(\mu_0)) \geq k$ si realizza se e solo se

$$e^{(\mu_1-\mu_0)\sum_i y_i/\sigma^2 - n(\mu_1^2-\mu_0^2)/2\sigma^2} \geq k$$

cioè se e solo se

$$\frac{(\mu_1-\mu_0)\sum_i y_i}{\sigma^2} - \frac{n(\mu_1^2-\mu_0^2)}{2\sigma^2} \geq \log k$$

da cui

$$\sum_i y_i \geq \frac{2\sigma^2 \log k + n(\mu_1^2-\mu_0^2)}{2(\mu_1-\mu_0)}$$

o equivalentemente,

$$\bar{y} \geq \frac{2\sigma^2 \log k + n(\mu_1^2 - \mu_0^2)}{2n(\mu_1 - \mu_0)} = k'.$$

Dall'ultima disuguaglianza si individua la zona di rifiuto R^* ponendo

$$\alpha = P(R^*|H_0) = P(\overline{Y} \geq k'|H_0) = P\left(\frac{\overline{Y} - \mu_0}{\sigma/\sqrt{n}} \geq \frac{k' - \mu_0}{\sigma/\sqrt{n}}\right)$$

Pertanto si trova

$$\frac{k' - \mu_0}{\sigma/\sqrt{n}} = z_{1-\alpha} \rightarrow k' = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

determinando la regione di rifiuto

$$R^* = \left\{(y_1, \ldots, y_n) : \bar{y} \geq \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right\}$$

del test più potente di livello α per H_0 contro H_1.

Se il test ha ipotesi alternativa $H_1 : \mu = \mu_1$, $\mu_1 < \mu_0$, con passaggi analoghi, si trova che il test più potente di livello α ha regione di rifiuto del tipo

$$R^* = \left\{(y_1, \ldots, y_n) : \bar{y} \leq \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right\}$$

Si osservi che in entrambi i casi la regione di rifiuto individuata non dipende dal particolare valore di μ_1. Si noti inoltre i test ottenuti coincidono con quelli già discussi nella sezione 7.4 per la media di una normale con varianza nota.

▲

Esempio 7.13. Sia $(Y_1, \ldots, Y_n)$ un campione casuale tratto da Y distribuita secondo il modello $f(\lambda) = \lambda e^{-\lambda y}$, $y > 0$, dove $\lambda = \lambda_0$ o $\lambda = \lambda_1 > \lambda_0$. Consideriamo la verifica di ipotesi $H_0 : \lambda = \lambda_0$ contro l'alternativa $H_1 : \lambda = \lambda_1$. Il rapporto di verosimiglianza sotto le due ipotesi è

$$\frac{L(\lambda_1)}{L(\lambda_0)} = \frac{\lambda_1^n e^{-\lambda_1 \sum_{i=1}^n y_i}}{\lambda_0^n e^{-\lambda_0 \sum_{i=1}^n y_i}} = \left(\frac{\lambda_1}{\lambda_0}\right)^n e^{-(\lambda_1 - \lambda_0)\sum_{i=1}^n y_i}$$

Secondo il Lemma di Neyman-Pearson, il test più potente sarà del tipo: si rifiuti H_0 se $L(\lambda_1)/L(\lambda_0) \geq k$, cioè

$$\left(\frac{\lambda_1}{\lambda_0}\right)^n e^{-(\lambda_1-\lambda_0)\sum_{i=1}^n y_i} \geq k$$

da cui si ottiene

$$\sum_{i=1}^n y_i \leq \frac{1}{\lambda_1 - \lambda_0} \log\left\{\left(\frac{\lambda_1}{\lambda_0}\right)^n k^{-1}\right\} = k'$$

dove k' è una costante. La condizione $P((Y_1, Y_2, \ldots, Y_n) \in R^*|H_0) = \alpha$ è $P(\sum_i Y_i \leq k'|\lambda = \lambda_0) = \alpha$, dove $\sum_i Y_i \sim Ga(n, \lambda_0)$, essendo le Y_i esponenziali indipendenti di parametro λ_0 quando H_0 è vera. Dunque si trova che il test più potente al livello di significatività α ha regione di rifiuto del tipo $R^* = \{\sum_i y_i \leq k'\}$, dove k' è il quantile di ordine α della distribuzione gamma di parametri n e λ_0.

▲

7.4.2 Test uniformemente più potenti

Nella sezione 7.4.1 si è presa in esame la verifica di ipotesi semplici rispetto ad un'alternativa semplice. Considerando il problema più generale di verifica dell'ipotesi $H_0 : \theta = \theta_0$ in alternativa ad una ipotesi composta, ad esempio $H_1 : \theta > \theta_0$ o $H_1 : \theta < \theta_0$.

Si è già osservato in precedenza che, se l'ipotesi alternativa è composta, la probabilità β e quindi anche la potenza π sono funzioni di θ, che varia nell'intervallo che definisce l'ipotesi alternativa. Ad esempio, nel caso del test per la media di una normale con varianza nota, se $H_0 : \mu = \mu_0$ e $H_1 : \mu > \mu_0$ la funzione di potenza del test con livello di significatività α è

$$\pi(\mu) = 1 - \beta(\mu) = 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha}\right)$$

che, come si è osservato in precedenza, cresce all'aumentare di μ e tende ad 1 al tendere di μ ad infinito. Se l'ipotesi alternativa è $H_1 : \mu < \mu_0$, allora si ha

$$\pi(\mu) = 1 - \beta(\mu) = \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{1-\alpha}\right).$$

Un criterio per la ricerca di un test è quello di individuare, ove possibile, il test per cui la funzione di potenza è massima per ogni μ contemplato dall'ipotesi alternativa, che corrisponde alla definizione di test *uniformemente più potente*. L'avverbio "uniformemente" si riferisce, chiaramente, a tutti i possibili valori che θ può assumere sotto l'ipotesi alternativa e pertanto, quando esiste, il test presenta tra tutti quelli di livello α, la maggiore probabilità di rifiutare correttamente l'ipotesi nulla.

Definizione 7.2 (Test uniformemente più potente)**.** Sia $(Y_1, \ldots, Y_n)$ un campione casuale tratto da Y che è distribuita secondo il modello $f(y;\theta)$, $\theta \in \Theta$, e si voglia verificare $H_0 : \theta = \theta_0$ contro $H_1 : \theta < \theta_0$, oppure $H_1 : \theta > \theta_0$. Si dice **test uniformemente più potente**, al livello di significatività α, il test che ha regione di rifiuto R^* per la quale vale la relazione

$$\pi_{R^*}(\theta) \geq \pi_R(\theta),$$

per ogni θ nell'ipotesi alternativa, essendo $\pi_R(\theta)$ la funzione di potenza di un qualsiasi altro test con zona di rifiuto $R \neq R^*$.

Un caso in cui è possibile individuare un test uniformemente più potente si presenta quando il test individuato dal Lemma di Neyman-Pearson non dipende da particolare valore di θ se non per la condizione $\theta > \theta_0$ o $\theta < \theta_0$, in tal caso si perviene allo stesso test più potente per *ogni* θ contemplato dall'ipotesi alternativa, che quindi si può affermare essere *uniformemente* più potente.

Esempio 7.14 (Test uniformemente più potente per la media di una normale)**.** Si riprenda l'esempio 7.12 dove $H_0 : \mu = \mu_0$ è l'ipotesi nulla, e si riformuli l'ipotesi alternativa, rendendola composta: $H_1 : \mu > \mu_0$. Per un particolare valore di μ tale che $\mu_1 > \mu_0$ il test per la verifica $H_0 : \mu = \mu_0$ contro $H_1 : \mu = \mu_1 > \mu_0$ trovato mediante il Lemma di Neyman-Pearson ha regione di rifiuto $R^* = \{\bar{y} \geq \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}\}$. Allora il test con regione di rifiuto R^* è il più potente al livello di significatività α e poiché R^* non dipende da μ_1, esso è anche il test uniformemente più potente nel senso della definizione 7.2.

▲

La ricerca del test uniformemente più potente è resa più agevole se il modello distributivo di riferimento $f(y;\theta)$, $\theta \in \Theta$, ha *rapporto di verosimiglianza*

monotono, cioè se esiste una statistica $t(Y_1, Y_2, \ldots, Y_n)$ tale che, per ogni $\theta', \theta'' \in \Theta$, con $\theta' < \theta''$ il rapporto $\frac{L(\theta')}{L(\theta'')}$ sia una funzione non crescente o non decrescente di $t(y_1, \ldots, y_n)$.

Nel caso di famiglie di densità con rapporto di verosimiglianza monotono vale il seguente risultato per la verifica d'ipotesi nulla e alternativa composte unilaterali.

Teorema 7.2. Siano $H_0 : \theta \leq \theta_0$ l'ipotesi nulla e $H_1 : \theta > \theta_0$ l'ipotesi alternativa concernenti il parametro θ di una popolazione descritta dal modello $f(y;\theta)$. Assumendo che $f(y;\theta)$ abbia un rapporto di verosimiglianza monotono nella statistica $T = t(Y_1, Y_2, \ldots, Y_n)$:

(i) se il rapporto è *non decrescente* in $t(y_1, \ldots, y_n)$ ed esiste k^* tale che $P(t(Y_1, \ldots, Y_n) < k^*) = \alpha$, allora il test con regione di rifiuto

$$R = \{(y_1, \ldots, y_n) : t(y_1, \ldots, y_n) < k^*\}$$

è un test uniformemente più potente al livello α per la verifica delle ipotesi suddette;

(ii) se il rapporto è *non crescente* in $t(y_1, \ldots, y_n)$ ed esiste k^* tale che $P(t(Y_1, \ldots, Y_n) > k^*) = \alpha$, allora il test con regione di rifiuto

$$R = \{(y_1, \ldots, y_n) : t(y_1, \ldots, y_n) > k^*\}$$

è un test uniformemente più potente al livello α per la verifica delle ipotesi suddette.

Il risultato è valido anche nel caso in cui si invertano le ipotesi H_0 ed H_1, a patto che si invertano le disuguaglianze che definiscono le regioni di rifiuto.

Esempio 7.15. Si consideri un campione casuale proveniente da una popolazione esponenziale con parametro λ e si vuole verificare l'ipotesi nulla $H_0 : \lambda \leq \lambda_0$. Se λ_1 è un valore di λ tale che $\lambda_1 > \lambda_0$ il rapporto

$$\frac{L(\lambda_0)}{L(\lambda_1)} = \left(\frac{\lambda_0}{\lambda_1}\right)^n e^{-(\lambda_0 - \lambda_1)\sum_{i=1}^n y_i}$$

è funzione crescente della statistica $t(y_1, \ldots, y_n) = \sum_{i=1}^n y_i$. Allora applicando la (i) del Teorema 7.2, un test uniformemente più potente è per la verifica

di $H_0 : \lambda \leq \lambda_0$ contro $H_1 : \lambda > \lambda_0$ è del tipo: si rifiuti H_0 se e solo se $\sum_{i=1}^{n} y_i < k^*$, dove k^* si ottiene da

$$\alpha = P\left(\sum_i Y_i < k^* | H_0\right) = \int_0^{k^*} \frac{1}{\Gamma(n)} \lambda_0^n u^{n-1} e^{-\lambda_0 u} du$$

quindi il test più potente trovato nell'esempio 7.13, è anche uniformemente più potente al livello di significatività α.

▲

I risultati visti finora per la ricerca di un test uniformemente più potente hanno riguardato esclusivamente il caso di ipotesi unilaterali. Si consideri allora il caso di un test bilaterale per la media di una popolazione normale con varianza nota. La regione di rifiuto con livello di significatività α per la verifica $H_0 : \mu = \mu_0$ in alternativa a $H_1 : \mu \neq \mu_1$ è

$$|\bar{y} - \mu_0| > z_{1-\alpha/2}\sigma/\sqrt{n}$$

e, in tal caso, non fornisce un test uniformemente più potente; infatti, se il vero valore di μ fosse superiore a μ_0, il test più potente sarebbe quello per l'alternativa $H_1 : \mu = \mu_1 > \mu_0$ con regione di rifiuto

$$R = \{(y_1, \ldots, y_n) : \bar{y} > \mu_0 + z_{1-\alpha}\sigma/\sqrt{n}\},$$

che ha funzione di potenza data dalla (7.21). Se invece il vero valore di μ fosse inferiore a μ_0 il test più potente sarebbe quello per l'alternativa $H_1 : \mu = \mu_1 < \mu_0$ con regione di rifiuto

$$R = \{(y_1, \ldots, y_n) : \bar{y} < \mu_0 - z_{1-\alpha}\sigma/\sqrt{n}\},$$

la cui funzione di potenza è data dalla (7.22).

7.5 Ampiezza campionaria e potenza di un test

Quanto discusso finora ha evidenziato che la potenza di un test rappresenta una misura della capacità del test di riconoscere che l'ipotesi nulla sia falsa. Fissato un α piccolo (ad esempio $\alpha = 0.05, 0.01$, ecc.), allora se $\theta \neq \theta_0$, $\pi(\theta)$ è la probabilità di prendere la decisione corretta, riconoscendo che H_0 è falsa.

Se un test di ipotesi è condotto con dati campionari consistenti in poche osservazioni, la potenza sarà bassa, e aumenterà al crescere della dimensione campionaria n. È interessante chiedersi che valore debba assumere n affinché sia sufficientemente alta la probabilità di rifiutare l'ipotesi H_0 quando $\theta \neq \theta_0$. In particolare, fissato un livello di probabilità π_0, si vuole individuare n tale che $\pi(\theta_1) \geq \pi_0$. Nel seguito si considereranno alcuni esempi per chiarire il legame tra potenza di un test e numerosità del campione.

Esempio 7.16. Si consideri la verifica dell'ipotesi $H_0 : \mu = \mu_0$, contro l'alternativa $H_1 : \mu \neq \mu_0$, con μ media della distribuzione $\mathcal{N}(\mu, \sigma^2)$ (σ^2 nota) illustrata nella sezione 7.2.2. Per il test suddetto la funzione di potenza è data dalla (7.17). La funzione π è simmetrica rispetto al valore μ_0, ed è crescente in n. Sia $\sigma^2 = 1$ e $\alpha = 0.05$. Si consideri inoltre un valore μ_1 che si discosta di 0.1 da μ_0, $\mu_1 = \mu_0 + 0.1$. Allora la potenza è espressa da

$$\Phi(-0.1\sqrt{n} - 1.96) + 1 - \Phi(-0.1\sqrt{n} + 1.96).$$

(lo stesso risultato si ottiene per $\mu_1 = \mu_0 - 0.1$ in virtù della simmetria).

Se ne deduce che una volta determinato n tale che $\pi(\mu_1) \geq \pi_0$, si è in grado di dire che la potenza del test sarà almeno pari a π_0 per tutti i valori di μ che soddisfano $|\mu_0 - \mu| \geq |\mu_0 - \mu_1|$. Ad esempio, sia $\mu_0 = 0$, $n = 10$ e $\alpha = 0.05$ come nella figura 7.2. Per $\mu = 1$ la potenza del test è data da $\pi(\mu) = 0.885$. Quindi, per qualsiasi valore μ tale che $\mu < -1$ o $\mu > 1$ la potenza sarà maggiore del valore trovato per $\pi(1)$. ▲

Esempio 7.17. Si supponga di voler verificare l'ipotesi $H_0 : \mu = 2$ a fronte dell'alternativa $H_1 : \mu > 2$, dove μ è la media di una popolazione $\mathcal{N}(\mu, \sigma^2)$ (con σ nota). Si assuma $\sigma = 2$ e si consideri un campione casuale di ampiezza n da $Y \sim \mathcal{N}(\mu, \sigma^2)$. Fissato $\alpha = 0.01$ a funzione di potenza è data dalla (7.21). Sia $\mu_1 = \mu_0 + 3$ e si voglia determinare la numerosità n affinché la potenza sia almeno pari a 0.95, quando $\mu = \mu_1$. Poiché si richiede che

$$\pi(\mu_1) \geq 0.95,$$

risolviamo l'espressione in n

$$0.95 = \pi(\mu_1) = 1 - \Phi\left(z_{0.99} + \frac{2-5}{2/\sqrt{n}}\right)$$

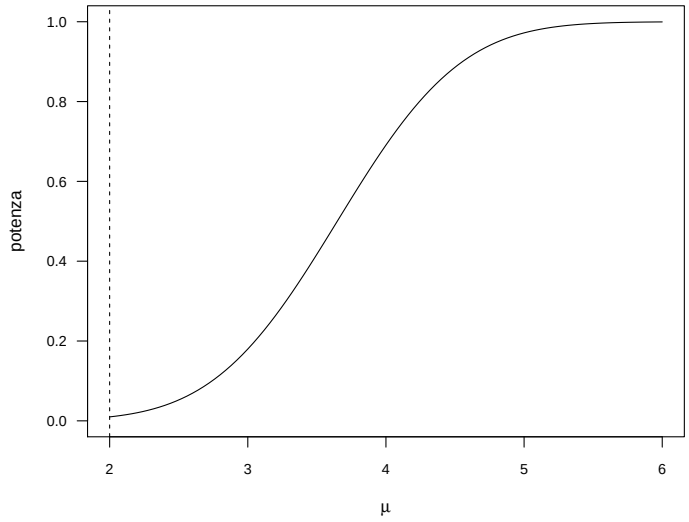


Figura 7.5: Funzione di potenza $\pi(\mu)$ del test sulla media con ipotesi nulla $H_0 : \mu = 2$, e ipotesi alternativa $H_1 : \mu > 2$, $n = 8$, $\sigma = 2$, $\alpha = 0.01$ (esempio 7.17).

Sostituendo $z_{0.99} = 2.33$, si ottiene

$$n = \frac{2^2(2.33 + 1.64)^2}{3^2} = 7.005$$

Quindi sarà necessario considerare $n \geq 7.005$ per soddisfare la condizione richiesta sulla potenza. Si arrotondi per eccesso all'intero $n = 8$, che determina un valore della potenza pari a $\pi(\mu_1) = 1 - \Phi(-1.913) = 0.97$, superiore al limite fissato di 0.95, se si sceglie $n = 7$ si ottiene un valore molto vicino ma inferiore a 0.95. ▲

7.6 Test del rapporto di verosimiglianza

I risultati visti nella sezione precedente consentono di derivare la migliore regione di rifiuto anche in alcuni casi in cui H_0 e H_1 non sono ipotesi semplici,

sebbene il teorema di Neyman-Pearson riguardi esclusivamente il caso in cui sia l'ipotesi nulla che quella alternativa sono semplici. Tuttavia, come si è visto, esso non consente sempre l'individuazione del test più potente o non sempre è agevole dedurre dai risultati visti la forma del test per il singolo specifico problema.

Una procedura generale per l'individuazione della regione critica che dà usualmente buoni risultati si basa sul rapporto delle funzioni di verosimiglianza già introdotto proprio nell'ambito del Lemma di Neyman-Pearson. Si è interessati a verificare l'ipotesi $H_0 : \theta = \theta_0$ rispetto all'ipotesi alternativa $H_1 : \theta \neq \theta_0$. L'idea alla base del **test del rapporto di verosimiglianza** è quella di valutare come il campione "è spiegato" dall'ipotesi nulla rispetto a come "è spiegato" senza imporre alcun vincolo. Sia $\hat{\theta}_{MV}$ la stima di massima verosimiglianza per θ. Possiamo considerare il rapporto tra il valore della funzione di verosimiglianza in corrispondenza di θ_0 e quello in corrispondenza della stima di massima verosimiglianza, $\hat{\theta}_{MV}$, e se tale rapporto è piccolo significa che nei dati è presente evidenza empirica contraria all'ipotesi nulla.

La regione di rifiuto del test del rapporto di verosimiglianza è formata da tutti i punti campionari tali per cui risulta piccolo il rapporto

$$\lambda = \frac{L(y_1, y_2, \ldots, y_n; \theta_0)}{\max_{\theta \in \Theta} L(y_1, y_2, \ldots, y_n; \theta)} = \frac{L(\theta_0)}{L(\hat{\theta}_{MV})} \tag{7.27}$$

dove $0 \leq \lambda \leq 1$. Se H_0 è vera, allora λ è vicino a 1. La regione di rifiuto quindi viene espressa in termini di una soglia K al di sotto della quale l'ipotesi nulla viene rifiutata, cioè la regione di rifiuto è del tipo $\{\lambda < K\}$, con $0 \leq K \leq 1$, dove K è tale che la probabilità di commettere un errore di I tipo sia uguale ad α:

$$P(\lambda < K | H_0) = \alpha.$$

Esempio 7.18. Si consideri la verifica dell'ipotesi sulla media di una popolazione normale con varianza nota. Sia $H_0 : \mu = \mu_0$ l'ipotesi nulla e $H_1 : \mu \neq \mu_0$ l'ipotesi alternativa composta bidirezionale. Ricorriamo al test del rapporto di verosimiglianza per individuare una regione di rifiuto. Sotto H_0 si ha

$$L(\mu_0) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\sum_i (y_i - \mu_0)^2 / 2\sigma^2}$$

mentre sotto H_1

$$L(\hat{\mu}) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\sum_i (y_i - \bar{y})^2 / 2\sigma^2}$$

essendo $\bar{y}$ la stima di massima verosimiglianza di μ. Allora la regione di rifiuto del test è definita dalla disuguaglianza

$$\lambda = \frac{e^{-\sum_i (y_i-\mu_0)^2/2\sigma^2}}{e^{-\sum_i (y_i-\bar{y})^2/2\sigma^2}} = e^{-\frac{n}{2\sigma^2}(\bar{y}-\mu_0)^2} < K$$

o, equivalentemente

$$\log(\lambda) = -\frac{n}{2\sigma^2}(\bar{y}-\mu_0)^2 < K', \;\; K' < 0$$

Quest'ultima equivale a rifiutare se

$$\left| \frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}} \right| > \sqrt{-2K'} = K^*$$

dove K' si trova imponendo

$$P\left(\left| \frac{\overline{Y}-\mu_0}{\sigma/\sqrt{n}} \right| > K^* | H_0 \right) = \alpha.$$

Essendo $(\overline{Y}-\mu_0)/(\sigma/\sqrt{n}) \sim \mathcal{N}(0,1)$ quando $\mu = \mu_0$, si trova

$$K^* = z_{1-\alpha/2}.$$

Si noti che la regione di rifiuto trovata coincide con quella già vista in precedenza (si veda la (7.16)).

▲

Una prima osservazione rilevante è che se $\Theta = \{\theta_0, \theta_1\}$ allora la (7.27) definisce la stessa zona di rifiuto del Lemma di Neyman-Pearson e quindi individua il test più potente al livello di significatività α. Una seconda osservazione riguarda la costante K per determinare la quale occorre conoscere la distribuzione del rapporto stesso in corrispondenza del livello di significatività prefissato. In generale, però, tale distribuzione non è di semplice derivazione. In tal caso si procede, come visto nel caso delle applicazioni del Lemma di Neyman-Pearson, cercando di semplificare la disuguaglianza $L(\theta_0)/L(\hat{\theta}) < K$ o l'equivalente in termini di log-verosimiglianza, in modo da ricondurla ad una statistica la cui distribuzione è nota.

Inoltre, sotto condizioni di regolarità abbastanza generali, si dimostra che qualora l'ipotesi H_0 sia semplice e riguardi un solo parametro (come assunto finora), allora la distribuzione asintotica, sotto H_0, della (7.27) è

$$RV = -2\log\lambda_n \xrightarrow{d} \chi^2_1$$

per $n \to \infty$. Se l'ipotesi nulla è falsa, la statistica RV assume valori elevati; di conseguenza si rifiuterà H_0, al livello α, quando

$$RV = -2 \log \lambda_n \geq \chi^2_{1,1-\alpha}$$

dove $\chi^2_{1,1-\alpha}$ è il quantile al livello $(1-\alpha)$ della chi-quadrato con 1 grado di libertà.

Esempio 7.19. Per un modello normale $\mathcal{N}(\mu, \sigma^2)$ si vuole verificare $H_0 : \mu = \mu_0$. Allora, fissato il livello di significatività $\alpha = 0.05$, si rifiuterà H_0 ogni qual volta $-2 \log \lambda > \chi^2_{1,0.95} = 3.84$. Si osservi che la statistica del rapporto di verosimiglianza $-2 \log \lambda_n$ può essere espressa come

$$RV = -2 \log \left(\frac{L(\mu_0)}{L(\hat{\mu}_{MV})} \right) = 2(\log L(\hat{\mu}_{MV}) - \log L(\mu_0))$$

cioè è pari a due volte la differenza tra il valore della log-verosimiglianza in $\hat{\mu}_{MV}$ e in μ_0.

▲

7.7 Procedure di test asintotiche

Quando si usa la stima di massima verosimiglianza si può ricorrere a procedure di test che sfruttano le proprietà asintotiche di tali stimatori che sono state illustrate precedentemente. Nel paragrafo precedente è stato introdotto il test del rapporto di verosimiglianza che si basa proprio sul confronto tra le funzioni di verosimiglianza sotto l'ipotesi nulla e alternativa e le proprietà asintotiche della statistica test. Anche se talvolta conduce a distribuzioni esatte, il criterio del rapporto di verosimiglianza può essere considerato un test asintotico poiché, quasi sempre, si ricorre alla distribuzione della statistica test sotto H_0 per $n \to \infty$.

Una procedura alternativa basata sulla stima di massima verosimiglianza prende il nome di *Test di Wald* e viene brevemente illustrata nel seguito.

7.7.1 Test di Wald

Sia ancora $\ell(\theta)$ la funzione di log-verosimiglianza di θ, $\hat{\theta}_{MV}$ lo stimatore di massima verosimiglianza per θ. Si consideri innanzitutto la verifica dell'ipotesi

$$H_0 : \theta = \theta_0$$

Il test di Wald si basa sull'idea che la verifica dell'ipotesi nulla possa essere basata su una misura della distanza tra $\hat{\theta}_{MV}$ e θ_0, in particolare si definisce la statistica di Wald come

$$W = (\hat{\theta}_{MV} - \theta_0)^2 I(\theta)$$

dove $I(\theta)$ denota l'informazione attesa di Fisher. Si noti che la differenza $\hat{\theta}_{MV} - \theta_0$ al quadrato viene pesata con una misura della curvatura della log-verosimiglianza, la sua derivata seconda cambiata di segno. Infatti, per quanto visto

$$I(\theta) = -E\left(\frac{d^2 \log L(\theta)}{d\theta^2}\right).$$

Ricordando il teorema 4.2, per grandi campioni è possibile riformulare la statistica di Wald come segue

$$W = (\hat{\theta}_{MV} - \theta_0)^2 V(\hat{\theta}_{MV})^{-1} = \left(\frac{\hat{\theta}_{MV} - \theta_0}{\sqrt{V(\hat{\theta}_{MV})}}\right)^2 \tag{7.28}$$

dove, per implementare il test, la varianza $V(\hat{\theta}_{MV})$ sarà sostituita dall'informazione osservata $1/I(\hat{\theta}_{MV})$.

Sotto H_0, la distribuzione della statistica di Wald è

$$W \xrightarrow{d} \chi^2_1$$

pertanto si rifiuta H_0 al livello $\alpha(100)\%$ se $W \geq \chi^2_{1,1-\alpha}$, dove $\chi^2_{1,1-\alpha}$ è il quantile di ordine α della chi-quadrato con 1 grado di libertà.

Calcolando ora la radice quadrata si ottiene esattamente la statistica t

$$t = \sqrt{W} = \frac{\hat{\theta}_{MV} - \theta_0}{\sqrt{V(\hat{\theta}_{MV})}} \tag{7.29}$$

si osservi che il test t è anche un test di Wald e la distribuzione di t sotto H_0 è asintoticamente $\mathcal{N}(0,1)$.

Infine si noti che il test di Wald non richiede la specificazione di una ipotesi alternativa e può essere applicato a tutti gli stimatori consistenti e con distribuzione asintotica normale.

Se $g\colon \mathbb{R} \to \mathbb{R}$ è una funzione continua, allora per la verifica dell'ipotesi

$$H_0 : g(\theta) = q$$

si potrà adottare la statistica test

$$W = \frac{(g(\hat{\theta}_{MV}) - q)^2}{V(g(\hat{\theta}_{MV}))}$$

che, sotto H_0, ha distribuzione asintotica chi-quadrato con 1 grado di libertà. Una stima consistente della varianza al denominatore può essere ottenuta utilizzando il metodo delta illustrato in sezione 4.4.4.

Esempio 7.20. Si è osservato un campione $(y_1, \ldots, y_{400})$ di 400 valori da una variabile aleatoria esponenziale di parametro λ. Dal campione risulta una media pari a $\bar{y} = 0.8$. Si vuole costruire un test asintotico per la verifica dell'ipotesi $H_0 : \lambda = \lambda_0$, $\lambda_0 = 1$.

È noto che lo stimatore di massima verosimiglianza del parametro λ di una esponenziale è $\hat{\lambda}_{MV} = 1/\bar{y} = 1.25$ ed essendo n elevato vale l'approssimazione

$$\hat{\lambda}_{MV} \sim \mathcal{N}(\lambda, I(\lambda)^{-1})$$

dove $I(\lambda)^{-1} = \lambda^2/n$ è il reciproco dell'informazione attesa di Fisher calcolata come nell'esempio 5.7. La verifica dell'ipotesi $H_0 : \lambda = 1$ si basa sulla statistica di Wald

$$W = \frac{(\hat{\lambda}_{MV} - 1)^2}{I(\lambda)^{-1}}$$

che ha, sotto H_0, distribuzione chi-quadrato con 1 grado di libertà. La varianza asintotica viene stimata dai dati campionari mediante $\hat{\lambda}^2_{MV}/n$. Pertanto, si trova $w = 400(1.25 - 1)^2/1.25^2 = 16$ ed essendo $\chi^2_{1,0.95} = 3.84$ il quantile di ordine 0.95 della chi-quadrato con 1 grado di libertà, l'ipotesi nulla viene rifiutata al livello del 5%. ▲

7.8 Confronto di popolazioni

Sin qui si è considerata l'inferenza per una singola popolazione. Vi sono però problemi di verifica di ipotesi di grande interesse che prevedono dei confronti tra gruppi diversi. Ad esempio, si supponga di voler verificare se i redditi medi di due regioni siano diversi e in che misura, o ancora si vuole testare l'efficacia di un nuovo farmaco rispetto a quello esistente. La sezione riguarda la verifica di ipotesi nel contesto del confronto di popolazioni, partendo dal caso elementare della verifica di ipotesi sulla differenza tra le medie di due popolazioni normali.

7.8.1 Inferenza per la differenza di medie di popolazioni normali

Il problema del confronto tra le medie di due popolazioni normali può essere formulato come segue: consideriamo due variabili aleatorie Y_1 e Y_2 che si distribuiscono normalmente, con medie μ_1, μ_2 e varianze σ_1^2 e σ_2^2, rispettivamente. Si osservano due campioni indipendenti, sulla base due quali vogliamo verificare una ipotesi relativa alla differenza tra le medie $\mu_1 - \mu_2$.

Un esempio tipico in questo contesto è quello della verifica dell'efficacia di un farmaco. Si considerano n pazienti che soffrono di una certa patologia e si dividono casualmente in due gruppi. Ad uno di questi si somministra il farmaco, all'altro un placebo. Se si suppone, ad esempio, che il farmaco agisca sui livelli di colesterolo, allora si misurerà l'efficacia del trattamento sulla base del confronto tra i livelli medi nei due gruppi (si veda l'esempio illustrato nel paragrafo 1.4.6). Si possono allora considerare due popolazioni, quella di chi assume il placebo, che si supporrà distribuita secondo la legge $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ e quella di chi assume il farmaco $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, e da queste vengono estratti due campioni di dimensione n_1 e n_2, rispettivamente. L'inferenza riguarda quindi una coppia di variabili aleatorie indipendenti.

Osservando due campioni, si otterranno le medie campionarie $\bar{y}_1$ e $\bar{y}_2$, dove si ricordi che le variabili $\overline{Y}_1$ e $\overline{Y}_2$ hanno distribuzione $\overline{Y}_1 \sim \mathcal{N}(\mu_1, \sigma_1^2/n_1)$ e $\overline{Y}_2 \sim \mathcal{N}(\mu_2, \sigma_2^2/n_2)$, in virtù dell'assunzione di gaussianità delle popolazioni. L'oggetto di interesse è ora la differenza tra le medie $\mu_1 - \mu_2$, e in particolare interessa verificare l'ipotesi che il farmaco sia inefficace, cioè

$$\mu_1 - \mu_2 = 0.$$

Se infatti fosse $\mu_1 - \mu_2 = 0$, allora ciò implicherebbe che la media nei due gruppi è la stessa e quindi il farmaco non ha sortito l'effetto desiderato. La differenza $\mu_1 - \mu_2$ si può ragionevolmente stimare con $\overline{Y}_1 - \overline{Y}_2$, dove si noti che, supponendo siano indipendenti i due campioni,

$$\overline{Y}_1 - \overline{Y}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

In analogia a quanto fatto nel caso della verifica sulla media di una popolazione normale, per verificare l'ipotesi $\mu_D = \mu_1 - \mu_2 = 0$ si userà il fatto che

$$D = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

ha distribuzione normale per ricavare test di significatività e regioni di rifiuto per $\mu_1 - \mu_2$.

Si è mostrato nel paragrafo 5.5 come, partendo da

$$P\left(-z_{1-\alpha/2} \leq \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

per un α fissato, si ricava

$$P\left(\overline{Y}_1 - \overline{Y}_2 - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \overline{Y}_1 - \overline{Y}_2 + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1-\alpha$$

da cui si ottiene un intervallo di livello $1 - \alpha$ per $\mu_1 - \mu_2$ di estremi

$$(\overline{Y}_1 - \overline{Y}_2) \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{7.30}$$

dove le varianze delle popolazioni σ_1^2 e σ_2^2 si erano supposte note. Di conseguenza l'ipotesi nulla $\mu_1 = \mu_2$ (o $\mu_D = 0$) può essere verificata determinando se l'intervallo di confidenza include o meno lo 0. Dunque fissato α (ad esempio, $\alpha = 0.05$), se gli estremi dell'intervallo che si ottiene dall'ultima espressione contengono lo zero, allora si potrà concludere che il valore reale di μ_D non è significativamente diverso da zero.

In alternativa, per condurre il **test bilaterale per la differenza tra le medie** si consideri che sotto l'ipotesi nulla

$$H_0 : \mu_1 - \mu_2 = 0$$

si ha, quando H_0 è vera,

$$D_0 = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1)$$

che è tanto più grande in valore assoluto quanto più il campione si discosta dall'ipotesi formulata. Il valore-p è quindi dato da

$$P(|D_0| > |d_0|) = P\left(\left|\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right| > \left|\frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right|\right)$$
$$= 2\left(1 - \Phi\left(\left|\frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right|\right)\right). \tag{7.31}$$

Volendo considerare il problema della verifica di

$$H_0 : \mu_1 - \mu_2 = 0; \qquad H_1 : \mu_1 - \mu_2 \neq 0 \tag{7.32}$$

che equivale a verificare $H_0 : \mu_1 = \mu_2$ contro l'alternativa $H_1 : \mu_1 \neq \mu_2$, fissato α, si pone

$$P(|D_0| > z_{1-\alpha/2}|H_0) = \alpha$$

dove $z_{1-\alpha/2}$ è il quantile di ordine $1 - \alpha/2$ della normale standard e $D_0 \sim \mathcal{N}(0, 1)$ sotto H_0. Si trova quindi la regione di rifiuto di livello α:

$$R = \left\{|D_0| > z_{1-\alpha/2}\right\} = \left\{\left|\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}\right| > z_{1-\alpha/2}\right\}.$$

Per il test **unilaterale (destro)** per la verifica dell'ipotesi nulla sulla differenza tra le medie

$$H_0 : \mu_1 - \mu_2 \leq 0$$

che equivale a $H_0 : \mu_1 \leq \mu_2$, ci si riferisce sempre alla quantità D_0 ma lo scostamento è indicato da valori grandi (positivi), quindi il valore-p del test è

$$P(D_0 > d_0) = P\left(\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right)$$
$$= 1 - \Phi\left(\frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right), \tag{7.33}$$

dove si è indicato con d_0 il valore osservato di D_0. La regione di rifiuto di livello α per la verifica del sistema d'ipotesi

$$H_0 : \mu_1 - \mu_2 \leq 0; \qquad H_1 : \mu_1 - \mu_2 > 0 \tag{7.34}$$

è espressa da

$$R = \{D_0 > z_{1-\alpha}\} = \left\{ \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} > z_{1-\alpha} \right\}$$

cioè si deciderà di rifiutare l'ipotesi nulla se nel campione si è osservato $\bar{y}_1 - \bar{y}_2 > z_{1-\alpha}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$.

Infine se l'ipotesi nulla è (**test unilaterale sinistro**)

$$H_0 : \mu_1 - \mu_2 \geq 0$$

allora si guarderà alla probabilità di ottenere un valore più piccolo di quello osservato d_0, che fornisce il livello di significatività osservato per la verifica di H_0:

$$\begin{aligned} P(D_0 < d_0) =& P\left(\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right) \\ =& \Phi\left(\frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right). \end{aligned} \tag{7.35}$$

Per l'individuazione della regione di rifiuto di livello α nella formulazione

$$H_0 : \mu_1 - \mu_2 \geq 0; \qquad H_1 : \mu_1 - \mu_2 < 0$$

si procede in modo analogo a quanto visto prima, pervenendo alla regione

$$R = \{D_0 < -z_{1-\alpha}\} = \left\{ \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < -z_{1-\alpha} \right\}.$$

Le regioni di rifiuto sono sintetizzate in tabella 7.6.

Si è assunto finora di conoscere le varianze. Se ciò non si verifica, come tipicamente avviene, occorre estendere quanto visto sopra al caso in cui la varianza debba essere stimata. Più precisamente, si procede in maniera analoga a quanto fatto nel caso di un campione, salvo un'ipotesi aggiuntiva che riguarda l'eguaglianza delle varianze nei due gruppi.

H_0	H_1	R
$\mu_1 - \mu_2 \leq 0$	$\mu_1 - \mu_2 > 0$	$\bar{y}_1 - \bar{y}_2 > z_{1-\alpha}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$
$\mu_1 - \mu_2 \geq 0$	$\mu_1 - \mu_2 < 0$	$\bar{y}_1 - \bar{y}_2 < -z_{1-\alpha}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$
$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	$\lvert\bar{y}_1 - \bar{y}_2\rvert > z_{1-\frac{\alpha}{2}}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$

Tabella 7.6: Regioni di rifiuto per la verifica delle ipotesi sulla differenza tra le medie di due popolazioni normali con varianze note.

Caso delle varianze non note

Rispetto alla situazione precedente, occorre supporre che la varianza sia la medesima nelle due popolazioni, cioè $\sigma_1^2 = \sigma_2^2$, analogamente a quanto fatto in sezione 5.5 per la costruzione di intervalli di confidenza. Per stimare la varianza comune σ^2 si ricorre a

$$S_p^2 = \frac{1}{n_1 + n_2 - 2}((n_1 - 1)S_1^2 + (n_2 - 1)S_2^2)$$

Poiché, come visto,

$$D = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

possiamo scrivere

$$P(-t_{n_1+n_2-2,1-\alpha/2} \leq D \leq t_{n_1+n_2-2,1-\alpha/2}) = 1 - \alpha,$$

da cui si ricava un intervallo di confidenza di livello $1 - \alpha$ per $\mu_1 - \mu_2$

$$\overline{Y}_1 - \overline{Y}_2 \pm t_{n_1+n_2-2,1-\frac{\alpha}{2}} S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

In analogia al caso precedente, si può dire che si accetta l'ipotesi $\mu_D = \mu_1 - \mu_2 = 0$ se l'intervallo di confidenza per $\mu_1 - \mu_2$ include il valore ipotizzato per la differenza, ovvero lo zero.

Equivalentemente, fissato α, si può determinare la soglia critica k della regione di rifiuto per il test $H_0 : \mu_D = \mu_1 - \mu_2 = 0$, $H_1 : \mu_D = \mu_1 - \mu_2 \neq 0$, tale che

$$P(|\overline{Y}_1 - \overline{Y}_2| > k|H_0) = \alpha$$

da cui si ottiene

$$k = t_{n_1+n_2-2,1-\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Quindi, se si indica con d_0 il valore osservato di

$$D_0 = \frac{\overline{Y}_1 - \overline{Y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

la regione di rifiuto di livello α è scritta equivalentemente come

$$|d_0| > t_{n_1+n_2-2,1-\alpha/2},$$

essendo la distribuzione di D_0 sotto l'ipotesi nulla una t di Student con $n_1 + n_2 - 2$ gradi di libertà.

D'altra parte, per giudicare se la differenza osservata $(\bar{y}_1 - \bar{y}_2)$ devia fortemente da 0 nell'ipotesi $H_0 : \mu_D = \mu_1 - \mu_2 = 0$, si calcola la probabilità (valore-p)

$$\begin{aligned} P\left(|\overline{Y}_1 - \overline{Y}_2| > |\bar{y}_1 - \bar{y}_2|\right) &= P\left(|D_0| > \left|\frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{(1/n_1) + (1/n_2)}}\right|\right) \\ &= 2\left[1 - F_{t_{n_1+n_2-2}}\left(\left|\frac{\bar{y}_1 - \bar{y}_2}{s_p\sqrt{(1/n_1) + (1/n_2)}}\right|\right)\right], \end{aligned}$$

dove si è indicato con $F_{t_{n_1+n_2-2}}$ la funzione di ripartizione della t di Student con $n_1 + n_2 - 2$ gradi di libertà.

Si procede analogamente per ricavare le regioni di rifiuto al livello α per le altre due formulazioni dell'ipotesi nulla $H_0 : \mu_1 - \mu_2 \leq 0$ e $H_0 : \mu_1 - \mu_2 \geq 0$ (tabella 7.7).

Se l'ipotesi nulla è

$$H_0 : \mu_1 - \mu_2 \leq 0$$

allora il valore-p si ottiene calcolando

$$P(D_0 > d_0) = 1 - F_{t_{n_1+n_2-2}}(d_0)$$

dove $d_0 = (\bar{y}_1 - \bar{y}_2)/(s_p\sqrt{(1/n_1) + (1/n_2)})$.

Infine se si vuole verificare

$$H_0 : \mu_1 - \mu_2 \geq 0$$

allora si valuterà la probabilità

$$P(D_0 < d_0) = F_{t_{n_1+n_2-2}}(d_0)$$

che fornisce evidenza contro l'ipotesi nulla quando è molto piccola.

Esempio 7.21. Un produttore di lampadine afferma di aver realizzato un nuovo sistema che prolunga la durata delle lampadine. Per testare la validità delle sue affermazioni, si formula l'ipotesi che il nuovo dispositivo non abbia effetto sulla durata della lampadina, cioè che la durata media, μ_1, delle lampadine dotate del nuovo sistema sia minore o uguale alla durata media, μ_2, delle lampadine del vecchio tipo). Allora si ha l'ipotesi nulla $H_0 : \mu_1 \leq \mu_2$, e l'alternativa è che il nuovo sistema sia migliore di quello già in uso, $H_1 : \mu_1 > \mu_2$. Si osservano due campioni da ciascun tipo di lampadina, entrambi di numerosità $n_1 = n_2 = 31$. Dal campione delle lampadine dotate del nuovo sistema si ottiene $\bar{y}_1 = 1195.16$ e $s_1^2 = 118.13$, mentre il campione di lampadine del vecchio tipo ha fornito $\bar{y}_2 = 1180.05$ e $s_2^2 = 124.34$.

Per entrambe le popolazioni si assuma la normalità della variabile durata di vita. Inoltre, nell'ipotesi che le varianze delle due popolazioni siano uguali, la varianza comune è stimata mediante la media ponderata delle varianze campionarie corrette

$$s_p^2 = \frac{30(118.13) + 30(124.34)}{60} = 121.23.$$

Fissato un livello di significatività, ad esempio, pari ad $\alpha = 0.01$, si rifiuta l'ipotesi nulla se

$$\bar{y}_1 - \bar{y}_2 > t_{n_1+n_2-2,1-\alpha}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Equivalentemente, si confronta il valore osservato della statistica test

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{1195.16 - 1180.05}{121.23(1/31 + 1/31)} = 5.40$$

H_0	H_1	R
$\mu_1 - \mu_2 \leq 0$	$\mu_1 - \mu_2 > 0$	$\bar{y}_1 - \bar{y}_2 > t_{n_1+n_2-2,1-\alpha}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$
$\mu_1 - \mu_2 \geq 0$	$\mu_1 - \mu_2 < 0$	$\bar{y}_1 - \bar{y}_2 < -t_{n_1+n_2-2,1-\alpha}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$
$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	$\lvert\bar{y}_1 - \bar{y}_2\rvert > t_{n_1+n_2-2,1-\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

Tabella 7.7: Regioni di rifiuto per la verifica delle ipotesi sulla differenza tra le medie di due popolazioni normali con varianze incognite uguali.

con il quantile di ordine $1 - \alpha = 0.99$ della t con 60 gradi di libertà, $t_{60,0.99} = 2.39$. Essendo $5.40 > 2.39$ si decide rifiutare l'ipotesi nulla al livello dell'1%, cioè si riconosce al nuovo tipo di lampadina migliori qualità rispetto al vecchio.

Ricordando che il livello di significatività osservato è la probabilità di osservare valori della funzione test meno favorevoli ad H_0 del valore effettivamente ottenuto, si ottiene un valore-p dato da

$$P(T > 5.40) < 0.0005$$

poiché dalle tavole risulta $P(T > 3.4602) = 0.0005$, per $T \sim t_{60}$. Possiamo dunque concludere che sembra sussistere una aumento significativo della durata nelle lampadine con il nuovo sistema, a sostegno dell'ipotesi alternativa.

▲

7.8.2 Inferenza per coppie appaiate

Nel paragrafo 7.8.1 si è trattato l'inferenza circa le medie di due popolazioni normali indipendenti. In questo paragrafo si considera invece il caso in cui i due campioni estratti siano dipendenti, ed in tal caso la verifica di ipotesi sulla differenza tra le medie delle due popolazioni si basa sull'osservazione di *coppie appaiate*.

Un tipico esempio è quello in cui un gruppo di individui si sottopone ad un trattamento dietetico, e si verifica l'efficacia della dieta usando il peso di

H_0	H_1	R
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	$\bar{d} > \frac{s_D}{\sqrt{n}} t_{n-1,1-\alpha}$
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$	$\bar{d} > \frac{s_D}{\sqrt{n}} t_{n-1,\alpha}$
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$\lvert\bar{d}\rvert > \frac{s_D}{\sqrt{n}} t_{n-1,1-\alpha/2}$

Tabella 7.8: Regioni di rifiuto per la verifica delle ipotesi sulla media per coppie appaiate.

ciascun soggetto "prima" e "dopo" il trattamento. In questa situazione ogni valore precedente alla dieta è associato al valore che si osserva al termine della dieta e ogni coppia di misure si riferisce alla stessa persona.

Questa stessa situazione è già stata descritta nella procedura di costruzione di intervalli di confidenza. Sia $D_i = X_i - Y_i$, $i = 1, \ldots, n$, la differenza tra le variabili della coppia associata all'elemento i-mo del campione, allora le D_i sono variabili indipendenti, e assumendo $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ e $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, si può scrivere $D_i \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)$. Si osservi che $\mu_D = \mu_1 - \mu_2 = 0$ equivale a $\mu_1 = \mu_2$, e pertanto si può verificare una ipotesi su μ_D ricorrendo al test per la media di una normale con varianza incognita basato sul rapporto

$$T = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}}$$

che ha, sotto $H_0 : \mu_D = 0$, distribuzione t di Student con $n-1$ gradi di libertà, essendo $\overline{D} = (1/n)\sum_{1=1}^{n} D_i$ e $S_D^2 = \sum_{1=1}^{n}(D_i - \overline{D})^2/(n-1)$ la varianza campionaria delle differenze D_i. Le regioni di rifiuto di livello α per le diverse formulazioni delle ipotesi sono riportate in tabella 7.8.

7.8.3 Grandi campioni

Se l'obiettivo è l'inferenza sulla media o sulla differenza tra medie e il/i campione/i sono molto grandi, allora si possono usare i test ricavati sopra con l'assunzione di normalità anche se la popolazione non è normale. Infatti, se l'ampiezza campionaria è sufficientemente elevata allora, come visto, le medie campionarie sono distribuite comunque approssimativamente secondo

una normale in virtù del teorema del limite centrale. Inoltre si possono usare i test sviluppati per varianze note anche se le varianze sono stimate con le varianze campionarie. Se l'inferenza è su una singola popolazione, questo non comporta grandi cambiamenti. Se invece l'obiettivo riguarda l'inferenza sul confronto tra medie, allora tale semplificazione consente di evitare l'assunzione di eguaglianza tra le varianze.

Esempio 7.22. Si hanno due popolazioni Y_1 e Y_2 dalle quali si osservano due campioni di ampiezza n_1 e n_2, entrambi elevate. Le medie campionarie hanno approssimativamente distribuzione

$$\overline{Y_1} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \qquad \overline{Y_2} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

da cui, tenendo conto che, approssimativamente,

$$\frac{\overline{Y_1} - \overline{Y_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim \mathcal{N}(0,1)$$

si ricava il valore-p per il test di significatività per la verifica dell'ipotesi $H_0 : \mu_1 - \mu_2 = 0$

$$P\left(\left|\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}\right| > \left|\frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}\right|\right) = 2\left(1 - \Phi\left(\left|\frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}\right|\right)\right).$$

La regione di rifiuto di livello α per il sistema di ipotesi

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 - \mu_2 \neq 0$$

è data da

$$R = \left\{\left|\frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}\right| > z_{1-\alpha/2}\right\}.$$

▲

7.8.4 Inferenza per la differenza tra proporzioni

Si supponga di osservare due campioni indipendenti da due popolazioni bernoulliane con parametro p_1 e p_2 rispettivamente, dai quali si ottengono le proporzioni campionarie $\hat{p}_1$ e $\hat{p}_2$. Nel capitolo 5 si è visto come la costruzione di un intervallo di confidenza per la differenza tra p_1 e p_2 si può basare sulla quantità, approssimativamente normale,

$$D = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

determinando l'intervallo di livello α per $p_1 - p_2$ di estremi

$$\hat{p}_1 - \hat{p}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Per i test sulla differenza tra proporzioni è quindi ragionevole impiegare la quantità

$$D_0 = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{7.36}$$

dove

$$\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

à la proporzione campionaria complessiva per i due campioni. Nell'ipotesi $H_0 : p_1 = p_2$ la quantità D_0 è approssimativamente distribuita secondo una $\mathcal{N}(0,1)$, per cui si ricavano agevolmente i valori-P per i test bilaterale e unilaterali. Se d_0 è il valore osservato della (7.36) si ha:

- per la verifica dell'ipotesi $H_0 : p_1 = p_2$, il valore-p è espresso dalla probabilità $2(1 - \Phi(|d_0|))$;
- per il test unilaterale destro con $H_0 : p_1 \leq p_2$, il valore-p è espresso dalla probabilità $1 - \Phi(d_0)$;
- per il test unilaterale sinistro con $H_0 : p_1 \geq p_2$, il valore-p è espresso dalla probabilità $\Phi(d_0)$.

Le regioni di rifiuto di livello α sono riportate in tabella 7.9.

H_0	H_1	R
$p_1 \leq p_2$	$p_1 > p_2$	$\hat{p}_1 - \hat{p}_2 > z_{1-\alpha}\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$
$p_1 \geq p_2$	$p_1 < p_2$	$\hat{p}_1 - \hat{p}_2 < -z_{1-\alpha}\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$
$p_1 = p_2$	$p_1 \neq p_2$	$\lvert\hat{p}_1 - \hat{p}_2\rvert > z_{1-\alpha/2}\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

Tabella 7.9: Regioni di rifiuto per la verifica delle ipotesi sulla differenza tra le proporzioni.

Esempio 7.23. Un partito politico ha avviato una indagine sull'orientamento della popolazione in merito ad un prossimo referendum. L'interesse si concentra sull'opinione dei votanti nelle regioni che denotiamo con A e B. Il sondaggio fornisce i seguenti dati: su 500 intervistati dalla regione A, 300 intendono votare per il sì; nella regione B sono 340 su 600 coloro i quali hanno dichiarato di propendere per il sì. Si è interessati a confrontare le due regioni sulla base dei dati raccolti, quindi si formula l'ipotesi $H_0 : p_1 - p_2 = 0$, contro l'alternativa $H_1 : p_1 - p_2 \neq 0$. Fissato il livello di significatività $\alpha = 0.05$, si rifiuterà H_0 al livello del 5% se

$$\frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} > 1.96$$

Poiché $\hat{p}_c = (300+340)/(500+600) = 0.58$, si ha

$$\frac{(300/500) - (340/600)}{\sqrt{0.58(1-0.58)\left(\frac{1}{500}+\frac{1}{600}\right)}} = 1.115$$

Pertanto l'ipotesi nulla non viene rifiutata al livello $\alpha = 0.05$. D'altra parte il livello di significatività osservato è dato da

$$2(1 - \Phi(1.115)) = 0.265$$

a riprova dell'evidenza campionaria a favore dell'ipotesi nulla.

▲

7.8.5 Inferenza sulle varianze di due popolazioni normali

Il problema del confronto delle varianze di due popolazioni assume rilievo in diverse circostanze, ad esempio, nella comparazione della variabilità di due processi diversi per la produzione di uno stesso prodotto. Il confronto tra le varianze è inoltre una verifica preliminare necessaria per assumere l'uguaglianza delle varianze nella verifica di ipotesi sulla differenza tra le medie di due popolazioni normali con varianze incognite.

Si considerino due campioni $(Y_{A1}, \ldots, Y_{An_A})$ e $(Y_{B1}, \ldots, Y_{Bn_B})$ indipendenti da $Y_A \sim \mathcal{N}(\mu_A, \sigma_A^2)$ e $Y_B \sim \mathcal{N}(\mu_B, \sigma_B^2)$, rispettivamente. Si consideri l'ipotesi nulla

$$H_0 : \sigma_A^2/\sigma_B^2 = 1$$

Se l'ipotesi nulla è vera, allora le varianze delle popolazioni sono uguali. La statistica test viene derivata dalle due proprietà $(n_A - 1)S_A^2/\sigma_A^2 \sim \chi^2_{n_A-1}$ e $(n_B - 1)S_B^2/\sigma_B^2 \sim \chi^2_{n_B-1}$, essendo S_A^2 ed S_B^2 le varianze campionarie corrette per i due campioni. Come si è visto, la variabile casuale F si definisce come il rapporto di due variabili casuali chi-quadrato indipendenti divise per i rispettivi gradi di libertà. In questo caso il rapporto da considerare è dato da

$$F = \frac{\frac{(n_A-1)S_A^2/\sigma_A^2}{n_A-1}}{\frac{(n_B-1)S_B^2/\sigma_B^2}{n_B-1}} = \frac{S_A^2/\sigma_A^2}{S_B^2/\sigma_B^2} \sim F_{n_A-1,n_B-1}$$

Se l'ipotesi nulla è $H_0 : \sigma_A^2/\sigma_B^2 = 1$ è vera, la statistica test $F = S_A^2/S_B^2$ ha distribuzione F con $n_A - 1$ gradi di libertà al numeratore e $n_B - 1$ al denominatore. Se l'ipotesi alternativa è $H_1 : \sigma_A^2/\sigma_B^2 \neq 1$, allora il test è a due code e, fissato il livello di significatività α, si rifiuterà H_0 se il valore osservato $f = s_A^2/s_B^2$ è $f \leq f_{n_A-1,n_B-1;\alpha/2}$ o se $f \geq f_{n_A-1,n_B-1;1-\alpha/2}$, dove $f_{n_A-1,n_B-1;\gamma}$ è il quantile di ordine γ della distribuzione F di Fisher con i gradi di libertà indicati.

Se il test è a una coda, con $H_0 : \sigma_A^2/\sigma_B^2 \leq 1$ (cioè $H_0 : \sigma_A^2 \leq \sigma_B^2$) e $H_1 : \sigma_A^2/\sigma_B^2 > 1$, si rifiuterà l'ipotesi nulla se $f = s_A^2/s_B^2 > f_{n_A-1,n_B-1;1-\alpha}$. Per la verifica d'ipotesi $H_0 : \sigma_A^2/\sigma_B^2 \geq 1$, $H_1 : \sigma_A^2/\sigma_B^2 < 1$ la determinazione della regione di rifiuto al livello α è lasciata per esercizio.

Esempio 7.24. Si vuole verificare l'ipotesi $H_0 : \sigma_1^2 = \sigma_2^2$ contro l'alternativa $H_1 : \sigma_1^2 \neq \sigma_2^2$, dove σ_1^2, σ_2^2 sono le varianze incognite di due popolazioni

normali $\mathcal{N}(\mu_1, \sigma_1^2)$ e $\mathcal{N}(\mu_2, \sigma_2^2)$. Si estraggono due campioni indipendenti di ampiezza 20 dai quali si osserva $s_1 = 3.6$ e $s_2 = 1.4$. Fissato $\alpha = 0.05$, la regione di rifiuto è del tipo $R = \{s_1^2/s_2^2 \notin (f_{0.0255}, f_{0.975})\}$, dove con f_α si è indicato il quantile di ordine α della distribuzione $F \sim F_{19,19}$. Dalle tavole si trova $f_{0.025} = 0.396$ e $f_{0.975} = 2.526$. Poiché dal campione risulta

$$\frac{s_1^2}{s_2^2} = \frac{3.6^2}{1.4^2} = 6.61$$

il valore ottenuto è nella regione di rifiuto di livello 0.05 e pertanto l'ipotesi nulla viene rifiutata.

▲

7.9 Alcuni test non parametrici

Nei paragrafi precedenti si è discusso delle procedure per verificare ipotesi sui parametri caratterizzanti la distribuzione di Y, della quale abbiamo sempre assunto che la forma della sua funzione di massa o di densità fosse nota a meno del valore di tali parametri. Può però capitare di essere interessati a verificare se il modello scelto per rappresentare la distribuzione di Y è effettivamente ben scelto, cioè se i dati sono ben rappresentati dalla distribuzione assunta. I test che verranno trattati nel seguito servono a sottoporre a verifica ipotesi concernenti il modello stesso che caratterizza una popolazione.

Esistono diversi metodi per sottoporre a test ipotesi sulla forma della distribuzione: nel seguito si considereranno il test χ^2 per la verifica di ipotesi sul modello distributivo di una popolazione e per la verifica dell'ipotesi di indipendenza. In ambito nonparametrico, un altro test di notevole rilevanza è il test di Kolmogorov-Smirnov, basato sul confronto tra la funzione di ripartizione empirica e quella teorica.

7.9.1 Test di conformità (o adattamento) del χ^2

Si consideri la situazione in cui si dispone delle frequenze associate a k modalità di un certo carattere, $y_1, y_2, \ldots, y_k$. Le modalità y_i possono essere le categorie di una caratteristica qualitativa, i valori numerici di una variabile quantitativa o intervalli di una caratteristica quantitativa continua. La

verifica in questo contesto riguarda una teoria/modello su come potrebbe essere fatta la distribuzione, cioè una distribuzione di probabilità $\pi_1, \pi_2, \ldots, \pi_k$, $\pi_i \geq 0$, $\sum_{i=1}^{k} \pi_i = 1$. I dati sono riassunti nella tabella seguente.

modalità	frequenze ass.	frequenze rel.	frequenze teoriche
y_1	n_1	f_1	π_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_i	n_i	f_i	π_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_k	n_k	f_k	π_k

L'ipotesi da verificare è

$$H_0 : P(Y = y_i) = \pi_i, \quad i = 1, \ldots, k$$

Dato un campione di n unità si pone quindi il problema di confrontare la distribuzione di frequenza (osservata), $n_1, n_2, \ldots, n_k$, con le frequenze (probabilità) teoriche (ipotizzate). Sotto l'ipotesi nulla, si ha

$$E(n_i|H_0) = n\pi_i$$

cioè la frequenza attesa della modalità y_i sotto l'ipotesi nulla è $n\pi_i$. Il test si basa allora sulla statistica

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n\pi_i)^2}{n\pi_i} \tag{7.37}$$

che, per n grande, si distribuisce secondo una chi-quadrato con $k-1$ gradi di libertà.

Considerando che la (7.37) è una misura della distanza tra dati effettivi e dati attesi, l'ipotesi nulla sarà da rifiutare per valori molto grandi di χ^2, cioè per valori maggiori del quantile di ordine $1-\alpha$, $\chi^2_{k-1,1-\alpha}$, tale che

$$P(\chi^2 \geq \chi^2_{k-1,1-\alpha}|H_0) = \alpha.$$

La zona di rifiuto di livello α è

$$R = \{\chi^2 \geq \chi^2_{k-1,1-\alpha}\}.$$

Inoltre il valore-p è

$$P(\chi^2_{k-1} > \chi^2) = 1 - F_{\chi^2_{k-1}}(\chi^2).$$

Esempio 7.25. Una linea metropolitana ha quattro ingressi e si vuole verificare l'ipotesi che gli ingressi siano utilizzati con la stessa intensità. Si osserva un campione di 100 unità, e si schematizzano i dati ottenuti come segue

Ingresso	A	B	C	D
Frequenza	29	28	25	18

L'ipotesi nulla è pertanto $H_0 : \pi_i = 1/4$, $i = 1, \ldots, 4$. Sulla base del modello, la frequenza attesa dell'i-ma modalità è

$$n\pi_i = 100(1/4) = 25$$

e quindi la (7.37) vale

$$\chi^2 = \frac{(29-25)^2}{25} + \frac{(28-25)^2}{25} + \frac{(25-25)^2}{25} + \frac{(18-25)^2}{25} = 2.96$$

La regione di rifiuto al 5% è $\{\chi^2 > \chi^2_{3,0.95}\} = \{\chi^2 > 7.185\}$. Ne segue che l'ipotesi non viene rifiutata.

▲

Esempio 7.26. Si ipotizza che il numero di incidenti che si verifica ogni settimana in un tratto di strada sia distribuito secondo una Poisson di parametro $\lambda = 0.5$. Si osserva il numero di incidenti per 50 settimane, e si vuole verificare se le osservazioni provengono dalla distribuzione ipotizzata $P(0.5)$.

y_i	n_i
0	20
1	17
2	9
3+	4
	$n = 50$

Si calcolano allora le probabilità teoriche

$$\pi_i = \begin{cases} \frac{(0.5)^i}{i!} e^{-0.5} & i = 0, 1, 2 \\ 1 - \sum_{i=0}^{2} \frac{(0.5)^i}{i!} e^{-0.5} & i = 3 \end{cases}$$

e le frequenze attese $n\pi_i$:

i	π_i	$n\pi_i$
0	0.61	30.5
1	0.30	15.0
2	0.08	4.0
3	0.014	0.7

Quindi si ottiene

$$\chi^2 = \frac{(20-30.5)^2}{30.5} + \frac{(17-15.0)^2}{15} + \frac{(9-4)^2}{4} + \frac{(4-0.7)^2}{0.7} = 25.69$$

mentre $\chi^2_{3,0.95} = 7.81$, pertanto si rifiuta l'ipotesi.

▲

Si osservi che talora le modalità fossero classi di frequenza, il loro numero e la loro ampiezza possono incidere sul risultato del test: il test χ^2 dipende infatti anche dal tipo di classificazione adottata ed è inevitabile un certo grado di arbitrarietà. Un'indicazione generale è quella di considerare classi la cui frequenza non sia troppo piccola rispetto alle altre.

Nell'esempio precedente il test di conformità porta a rifiutare l'ipotesi nulla $H_0 : p(y) = (0.5)^i e^{-0.5}/i!$, $i = 0, 1, 2, \ldots$. Pertanto la distribuzione potrebbe non essere Poisson, oppure la distribuzione potrebbe essere Poisson ma con $\lambda \neq 0.5$. In particolare, si ricordi che la stima di massima verosimiglianza della media di $Y \sim Po(\lambda)$ è la media campionaria; quindi si procede a sostituire λ con il valore della media osservata nel campione e si calcolano con tale stima le frequenze attese. Il test usa sempre la (7.37) ma, avendo stimato un parametro, il quantile a cui riferirsi per la verifica dell'ipotesi nulla non è quello di un χ^2_{k-1} ma quello di un χ^2_{k-2}.

Esempio 7.27. Si consideri nuovamente l'esempio 7.26. La media campionaria non è calcolabile con i dati a disposizione (per via del fatto che sono messe assieme le modalità maggiori di 3), però sarà certamente maggiore di

$$\frac{1}{50}(17 + 18 + 12) = 0.94$$

che è comunque maggiore del valore ipotizzato. Ha senso allora confrontare i valori osservati con una $Po(0.94)$

y	n	π_i	$n\pi_i$
0	20	0.39	19.5
1	17	0.37	18.5
2	9	0.17	8.5
3+	4	0.07	3.5

il valore della statistica è

$$\chi^2 = \frac{(20-19.5)^2}{19.5} + \frac{(17-18.5)^2}{18.5} + \frac{(9-8.5)^2}{8.5} + \frac{(4-3.5)^2}{3.5} = 0.235$$

ed essendo $\chi^2_{2,0.95} = 5.99$, non si rifiuta l'ipotesi al livello del 5%.

▲

7.9.2 Test di adattamento di Kolmogorov-Smirnov

Sia $(Y_1, \ldots, Y_n)$ un campione casuale proveniente da una popolazione descritta dalla funzione di ripartizione continua $F(y)$. Si è accennato in precedenza che la funzione di ripartizione empirica

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_{(-\infty,y]}(Y_i)$$

è, per n grande, prossima alla funzione di ripartizione $F(y)$ della popolazione di provenienza del campione.

L'ipotesi nulla che si vuole sottoporre a verifica è che la distribuzione dalla quale proviene il campione sia una determinata distribuzione continua (normale, esponenziale, rettangolare, ecc.)

$$H_0 : F(y) = F_0(y)$$

in alternativa a

$$H_1 : F(y) \neq F_0(y)$$

dove F_0 definisce una legge continua su $\mathbb{R}$. L'ipotesi nulla specifica in modo completo la distribuzione della popolazione e pertanto la validità dell'ipotesi nulla può essere verificata utilizzando una misura di distanza tra F_n e F_0. In particolare, il test di Kolmogorov-Smirnov si basa sulla statistica

$$D_n = \sup_{-\infty<y<\infty} |F_n(y) - F_0(y)| \tag{7.38}$$

che è la massima differenza tra la funzione di ripartizione empirica e quella teorica (ipotizzata dall'ipotesi nulla). Se H_0 è falsa, $F_n(y)$ tenderà ad essere prossima alla vera $F(\cdot)$ e lontana dalla $F_0(\cdot)$ e, di conseguenza, valori grandi della (7.38) inducono a rifiutare l'ipotesi nulla.

Fissato α, un criterio di verifica dell'ipotesi nulla è quindi quello di rifiutare H_0 se $\{D_n > d_{1-\alpha}\}$, dove $d_{1-\alpha}$ è la soglia critica tale che $P(D_n \geq d_{1-\alpha}|H_0) = \alpha$. Quindi la regione

$$R = \left\{(y_1, \ldots, y_n) : \sup_{-\infty<y<\infty} |F_n(y) - F_0(y)| > d_{1-\alpha}\right\}$$

definisce un test che ha approssimativamente livello α. Per il calcolo di D_n si può procedere come segue:

- Si ordinano in modo crescente le osservazioni campionarie

$$y_{(1)} < y_{(2)} < \cdots < y_{(n)}$$

- si determinano i valori della funzione di ripartizione teorica F_0 in $y_{(i)}$, $i = 1, \ldots, n$ e le differenze

$$\frac{i}{n} - F_0(y_{(i)}); \quad F_0(y_{(i)}) - \frac{i-1}{n}, \ i = 1, \ldots, n$$

- si calcolano i valori

$$D_n = \max_{1\leq i\leq n} \left\{\max\left\{\frac{i}{n} - F_0(y_{(i)}), F_0(y_{(i)}) - \frac{i-1}{n}\right\}\right\} \tag{7.39}$$

Per dimensioni campionarie non elevate, si può individuare i centili $d_{1-\alpha}$ della distribuzione di D_n (si veda la tabella 7.11). Per campioni di dimensioni più elevate si può ricorrere alla distribuzione limite della statistica

$$\lim_{n\to\infty} P(\sqrt{n}D_n \leq t) = \lim_{n\to\infty} F_{\sqrt{n}D_n}(t) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2t^2} = H(t), t > 0.$$

Utilizzando solo il primo termine della serie, per approssimare il quantile di ordine $1-\alpha$ basta risolvere rispetto a t l'equazione:

$$1 - \alpha = P(\sqrt{n}D_n \leq t) = P(D_n \leq t/\sqrt{n}) \approx 1 - 2e^{-2t^2},$$

e porre $d_{1-\alpha} = t/\sqrt{n}$.

Si osservi che la distribuzione campionaria di D_n, quando le Y_i hanno funzione di ripartizione F_0, non dipende da $F_0(y)$, ma solo da n. Infatti le variabili aleatorie

$$U_i = F_0(Y_i)$$

hanno distribuzione uniforme $U_i \sim U(0,1)$ e sono indipendenti.

Esempio 7.28. Un computer ha generato $n = 20$ numeri casuali y_i che sono riportati, in ordine crescente, nella prima colonna della tabella 7.10. Applichiamo il test di Kolmogorov-Smirnov per verificare l'ipotesi che essi siano determinazioni della distribuzione normale di media 2 e varianza 1. Nella seconda colonna viene riportata la funzione di ripartizione teorica per ogni osservazione, e i valori per calcolare la statistica D_n con la (7.39).

$y_{(i)}$	$F_0(y_{(i)})$	$\frac{i}{n} - F_0(y_{(i)})$	$F_0(y_{(i)}) - \frac{i-1}{n}$
0.36	0.0505	-0.0005	0.0505
0.82	0.1190	-0.0190	0.0690
0.86	0.1271	0.0229	0.0271
1.04	0.1685	0.0315	0.0185
1.06	0.1736	0.0764	-0.0264
1.12	0.1894	0.1106	-0.0606
1.14	0.1949	0.1551	-0.1051
1.44	0.2877	0.1123	-0.0623
1.47	0.2981	0.1519	-0.1019
1.78	0.4129	0.0871	-0.0371
1.80	0.4207	0.1293	-0.0793
1.90	0.4602	0.1398	-0.0898
2.29	0.6141	0.0359	0.0141
2.67	0.7486	-0.0486	0.0986
2.73	0.7673	-0.0173	0.0673
2.75	0.7734	0.0266	0.0234
3.11	0.8665	-0.0165	0.0665
3.18	0.8810	0.0190	0.0310
3.24	0.8925	0.0575	-0.0075
3.56	0.9406	0.0594	-0.0094

Tabella 7.10: Dati per l'esempio 7.10.

Si trova quindi $D_n = 0.1551$, cercando il massimo valore delle differenze riportate nella terza e quarta colonna. Dalla tavola dei valori $d_{1-\alpha}$ si trova $d_{1-\alpha} = 0.352$ per $n = 20$ e $\alpha = 0.01$. Dunque per $\alpha = 0.01$ la regione di rifiuto è $D_n \in [0.352, 1]$. Poiché il dato empirico non appartiene a tale intervallo si conclude che l'ipotesi nulla che i valori generati provengano da una distribuzione normale con media 2 e varianza 1 non viene rifiutata al livello di significatività 0.01. ▲

7.9.3 Test χ^2 per la verifica dell'ipotesi di indipendenza

Il test del χ^2 sarà illustrato in questa sezione con riferimento alla verifica dell'ipotesi di indipendenza tra due caratteristiche di una popolazione. In particolare, si supponga che le due variabili X e Y possano assumere solo un numero finito di valori (in genere qualitativi), rispettivamente $x_1, \ldots, x_t$ e $y_1, \ldots, y_s$ (si può anche pensare che i possibili valori di X e Y siano classificati in un certo numero di classi). Si consideri quindi un campione casuale di ampiezza N dalla popolazione in oggetto, le cui unità vengono classificate secondo le modalità dei caratteri di X e Y. Sia n_{ij} il numero di unità che nel campione presentano le modalità i e j. I dati sono organizzati in una tabella di contingenza come quella riportata di seguito:

	X						
Y	x_1	x_2	$\cdots$	x_j	$\cdots$	x_t	
y_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1t}	$n_{1\cdot}$
y_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2t}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{it}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
y_s	n_{s1}	n_{s2}	$\cdots$	n_{sj}	$\cdots$	n_{st}	$n_{s\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$		$n_{\cdot j}$		$n_{\cdot t}$	N

dove

$$n_{i\cdot} = \sum_{j=1}^{t} n_{ij}; \qquad n_{\cdot j} = \sum_{i=1}^{s} n_{ij}$$

Le probabilità teoriche congiunte p_{ij}, e marginali $p_{\cdot j}$ e $p_{i\cdot}$ con le quali si presentano i valori delle v.a. non sono note, ma possono essere stimate tramite i loro stimatori MV, cioè tramite le frequenze empiriche relative congiunte e marginali:

$$\hat{p}_{ij} = \frac{n_{ij}}{N} \qquad \hat{p}_{i\cdot} = \frac{n_{i\cdot}}{N} \qquad \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{N}, \quad i = 1, \ldots, s; j = 1, \ldots, t.$$

L'indipendenza tra X e Y si verifica se e solo se sussiste la relazione $p_{ij} = p_{i\cdot}p_{\cdot j}$ e pertanto la verifica à espressa da

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j} \quad \forall (i, j)$$

Si tratta dunque di misurare la distanza tra le frequenze effettive e quelle attese sotto l'ipotesi nulla; in particolare, la *frequenza (assoluta) attesa* sotto l'ipotesi di indipendenza per le modalità i e j è

$$\hat{n}_{ij} = N\hat{p}_{i\cdot}\hat{p}_{\cdot j} = N\frac{n_{i\cdot}}{N}\frac{n_{\cdot j}}{N} = \frac{n_{i\cdot}n_{\cdot j}}{N}$$

Per confrontare le frequenze osservate con $\hat{n}_{ij}$ si utilizza quindi la statistica

$$\chi^2 = \sum_{i=1}^{s}\sum_{j=1}^{t}\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \tag{7.40}$$

che può essere ricondotta alla forma alternativa in termini delle frequenze relative osservate e teoriche:

$$\chi^2 = N\sum_{i=1}^{s}\sum_{j=1}^{t}\frac{(\hat{p}_{ij} - \hat{p}_{i\cdot}\hat{p}_{\cdot j})^2}{\hat{p}_{i\cdot}\hat{p}_{\cdot j}}$$

Si dimostra che, sotto H_0, la distribuzione limite della (7.40) è una chi-quadrato con $(s-1)(t-1)$ gradi di libertà. Le osservazioni che sono "distanti" dall'ipotesi nulla di indipendenza determinano un valore elevato della statistica considerata, pertanto il valore-p si calcola come

$$P(Q > \chi_0^2) = 1 - F_{\chi^2_{(s-1)(t-1)}}(\chi_0^2)$$

dove $Q \sim \chi^2_{(s-1)(t-1)}$ ha distribuzione chi-quadrato con $(s-1)(t-1)$ gradi di libertà e χ_0^2 è il valore osservato della statistica.

La regione di rifiuto del test di livello α è

$$\{\chi_0^2 > \chi^2_{(s-1)(t-1);1-\alpha}\}$$

cioè si rifiuterà l'ipotesi nulla per valori superiori al quantile di ordine $1-\alpha$ di una chi-quadrato con i gradi di libertà indicati.

Esempio 7.29. Si consideri un campione di $N = 232$ individui classificati, mediante una indagine tra le forze di lavoro, secondo il genere e la condizione occupazionale:

	M	F	
occupati	0.24	0.51	0.75
inattivi	0.055	0.095	0.15
disoccupati	0.005	0.095	0.10
	0.3	0.7	1

Si vuole verificare l'ipotesi di indipendenza tra sesso e condizione occupazionale usando un livello di significatività del 1%.

Passando alle frequenze assolute si ricava la tabella (effettuando gli opportuni arrotondamenti)

	M	F	
occupati	56	118	174
inattivi	13	22	35
disoccupati	1	22	23
	70	162	232

Queste vanno confrontate con le frequenze attese (teoriche) $\hat{n}_{ij} = (n_{i.}n_{.j})/232$ riportate in tabella

	M	F	
occupati	52.5	121.5	174
inattivi	10.56	24.44	35
disoccupati	6.94	16.06	23
	70	162	232

Il confronto fra le 6 frequenze osservate e teoriche si basa allora sulla somma della quantità $(n_{ij} - \hat{n}_{ij})^/\hat{n}_{ij}$, calcolata per ogni cella:

$$\chi_0^2 = \frac{(56-52.5)^2}{52.5} + \frac{(118-121.5)^2}{121.5} + \frac{(13-10.56)^2}{10.56} + \frac{(22-24.44)^2}{24.44} + \frac{(1-6.94)^2}{6.94} + \frac{(22-16.06)^2}{16.06} = 8.42$$

Il valore osservato della statistica è $\chi_0^2 = 8.42$, e la regione di rifiuto, fissato α è $\{\chi_0^2 > \chi_{2,1-\alpha}^2\}$, dove $\chi_{2,1-\alpha}^2$ è il quantile di ordine $1-\alpha$ della distribuzione chi-quadrato con 2 gradi di libertà. Dalle tavole si trova $\chi_{2,0.99}^2 = 9.21$. Pertanto con $\alpha = 0.01$ non si rifiuta l'ipotesi di indipendenza tra il sesso e lo stato occupazionale. Si osservi che se avessimo fissato il livello di significatività $\alpha = 0.05$ avremmo rifiutato l'ipotesi nulla essendo $\chi_{2,0.95}^2 = 5.99$.

▲

	α				
n	0.2	0.1	0.05	0.02	0.01
1	0.900	0.950	0.975	0.990	0.995
2	0.684	0.776	0.842	0.900	0.929
3	0.565	0.636	0.708	0.785	0.829
4	0.493	0.565	0.624	0.689	0.734
5	0.447	0.509	0.563	0.627	0.669
6	0.410	0.468	0.519	0.577	0.617
7	0.381	0.436	0.483	0.538	0.576
8	0.358	0.410	0.454	0.507	0.542
9	0.339	0.387	0.430	0.480	0.513
10	0.323	0.369	0.409	0.457	0.489
11	0.308	0.352	0.391	0.437	0.468
12	0.296	0.338	0.375	0.419	0.449
13	0.285	0.325	0.361	0.404	0.432
14	0.275	0.314	0.349	0.390	0.418
15	0.266	0.304	0.338	0.377	0.404
16	0.258	0.295	0.327	0.366	0.392
17	0.250	0.286	0.318	0.355	0.381
18	0.244	0.028	0.309	0.346	0.371
19	0.237	0.271	0.301	0.337	0.361
20	0.232	0.265	0.294	0.329	0.352
21	0.226	0.259	0.287	0.321	0.344
22	0.221	0.253	0.281	0.314	0.337
23	0.216	0.247	0.275	0.307	0.330
24	0.212	0.242	0.264	0.301	0.323
25	0.208	0.238	0.264	0.295	0.317
26	0.204	0.233	0.259	0.290	0.311
27	0.200	0.229	0.254	0.284	0.305
28	0.197	0.225	0.250	0.279	0.300
29	0.193	0.221	0.246	0.275	0.295
30	0.190	0.218	0.242	0.270	0.281

Tabella 7.11: Valore critico $d_{1-\alpha}$ per $n = 1, \ldots, 30$, per il test di Kolmogorov-Smirnov; $d_{1-\alpha}$ è tale che $P(D_n \geq d_{1-\alpha}) = \alpha$.