


La conoscenza è la vera protagonista nell'economia contemporanea e assume il ruolo di fattore primario nei più svariati contesti. Sono gli individui a detenere il potere della conoscenza, che è necessaria per gestire e anticipare i continui cambiamenti e le condizioni di incertezza che caratterizzano la nostra epoca. Questo volume rappresenta una guida, uno strumento concreto per approcciarsi alla conoscenza della realtà con rigore scientifico.

L'approccio statistico alla conoscenza dei fenomeni reali consente di raccogliere informazioni, elaborarle e analizzarle, guidando appunto l'individuo alla presa di decisioni. Un sapere puramente teorico può facilmente divenire obsoleto, ma l'applicabilità promossa dal testo consente di acquisire strumenti atti a gestire il cambiamento e a prevedere l'imprevedibile, vivendo così un processo di crescita continua. Il percorso proposto da questo libro non sarà, quindi, quello tradizionale di un qualsiasi manuale di statistica, ma guiderà in modo graduale all'utilizzo degli strumenti specifici di questa disciplina, attraverso una presentazione e articolazione dei metodi che ne sono alla base, prediligendo maggiormente un approccio di tipo deduttivo ai problemi reali.

Lo scopo del volume è quello di mettere il lettore in condizione di poter facilmente trasferire le conoscenze statistiche acquisite in campo socio-economico, nel settore di istruzione e formazione e più in generale nei diversi indirizzi della ricerca scientifica.

Tonio Di Battista è professore ordinario di statistica. Dal 2004 è Presidente dei Corsi di Laurea di Scienze dell'Educazione e della Formazione di Scienze Pedagogiche della Facoltà di Scienze della Formazione nell'Università degli Studi "G. d'Annunzio" di Chieti-Pescara, dove detiene i corsi di metodi e tecniche della valutazione e analisi e valutazione dei processi formativi. È Direttore del Centro di Ricerca Universitario per la Valutazione e lo Sviluppo (CERVAS). È membro del Consiglio Direttivo della Società Italiana di Statistica dal 2010. I principali temi di ricerca riguardano lo studio della biodiversità, del campionamento statistico e dei metodi e delle tecniche di valutazione dei servizi pubblici.

 **FrancoAngeli**
La passione per le conoscenze

€ 39,00 (U)

ISBN 978-88-204-0397-3



9 788820 403973

367.71 T. DI BATTISTA

METODI E TECNICHE PER LA VALUTAZIONE

Economia

Tonio Di Battista

**Metodi e tecniche
per la valutazione**

Un approccio statistico

FrancoAngeli

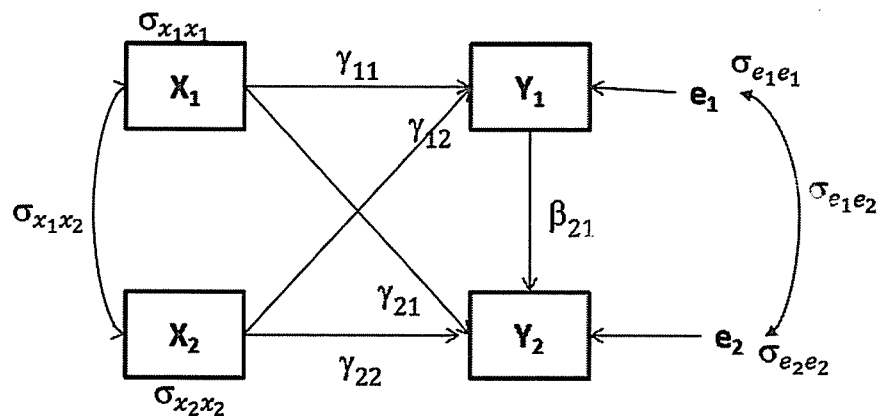


Figura 11.23: Sistema di equazioni strutturali.

In conclusione è bene chiarire che è possibile trovare altre terminologie che identificano il modello strutturale. In alcuni casi, infatti, si parla di "modello di equazioni simultanee" e "modello di equazioni lineari". Il primo termine fa riferimento alla contemporaneità dei nessi causali. Quanto alla seconda denominazione, essa deriva dal fatto che, nel caso più semplice, le equazioni del modello sono lineari. L'esistenza di forme di non linearità complica notevolmente l'analisi, specie quando si tratta di interazioni fra le variabili stesse.

Questa presentazione sulle equazioni strutturali è stata fatta in vista della loro risoluzione tramite l'approccio Lisrel che, come dice il nome stesso (*Li* sta per linear), è nato originariamente per trattare solo relazioni lineari tra le variabili. Successivi sviluppi hanno dimostrato l'applicabilità del suo approccio anche a relazioni non lineari (si veda Hayduc, 1987).

Capitolo 12

Inferenza

Nei capitoli precedenti è stato sviluppato il metodo statistico sotto l'ipotesi che il fenomeno reale fosse stato osservato su un collettivo di unità intese come repliche del fenomeno stesso. L'approccio utilizzato è quello descrittivo, ossia quello di desumere dai dati osservati le indicazioni di sintesi, di variabilità e di forma delle misure qualitative e quantitative delle caratteristiche specificate dallo studio del fenomeno in oggetto. In questo contesto le unità statistiche rappresentano la popolazione nella sua interezza. Tuttavia, se si fa mente locale sugli esempi e sui casi di studio trattati, ci si accorge che non sempre il fenomeno reale è esaustivamente osservabile, in quanto potrebbero esserci repliche non disponibili per ragioni diverse, che vanno dalla impossibilità materiale di osservare tutte le unità alla economicità dell'indagine.

In ogni caso va precisato che, sebbene le unità osservate siano solo una parte di quelle possibili, esse comunque danno indicazioni, sebbene circostanziate ai dati, su quanto si conosce del fenomeno reale. Ciò implica che unità aggiuntive potrebbero sostanzialmente modificare quanto desunto dall'analisi. Nell'approccio descrittivo, tuttavia, tutto è limitato alla disponibilità dei dati e nulla si dice sulla scelta delle unità, se non quanto desunto da un'attenta descrizione della fase di rilevazione. Le unità possono essere scelte usando un campionamento ragionato, cioè scegliendo opportunamente le unità più rappresentative, ma non mancano casi in cui l'analisi viene condotta solo su quanto si dispone, senza tener conto degli effetti prodotti dalla rilevazione

parziale dei dati.

Un approccio sostanzialmente diverso da quello sin qui analizzato, sia dal punto di vista filosofico che procedurale, è quello inferenziale. In questo caso si assume che il fenomeno reale sia costituito da un numero finito, infinito o illimitato di repliche. Il ricercatore ha l'obiettivo di ottenere indicazioni sulle caratteristiche della popolazione tramite l'estrazione casuale di un campione rappresentativo della popolazione stessa. In tal senso dette indicazioni assumono il significato di stima dei reali valori incogniti della popolazione. Tuttavia, come vedremo meglio in seguito, grazie all'introduzione della casualità nell'estrazione delle unità, l'attendibilità delle informazioni desunte può essere valutata per il tramite della misura della probabilità, che ci permette di effettuare una valutazione del processo di stima.

In termini essenziali, l'approccio inferenziale si articola in due parti:

- stima puntuale, che rappresenta la procedura implementata per ottenere un valore rappresentativo di un parametro della popolazione;
- stima intervallare, che rappresenta la procedura con cui il ricercatore, prefissata una probabilità, ricava un intervallo di valori in cui si può ritenere sia contenuto il parametro della popolazione.

È inoltre possibile distinguere due approcci all'inferenza, quali l'inferenza da popolazione finita e l'inferenza da modello.

12.1 Stima e stimatore

Da quanto abbiamo detto circa la stima puntuale, possiamo asserire che tale procedura fornisce, in sintesi, un valore numerico attendibile del parametro incognito della popolazione. Allora, è lecito chiedersi se essa possa essere considerata buona oppure no, e chiedersi come possa essere condotta un'opportuna valutazione del processo di stima. Per dare risposta a questa domanda, in genere, si rende necessario stabilire un termine di paragone, ossia stabilire un valore di riferimento ideale della bontà della stima. Infatti, il confronto con il valore

ideale o, addirittura, una misura della distanza da esso, fornirebbe un indiscusso termine di valutazione della stima ottenuta.

Prima di addentrarci in questo campo è, tuttavia, necessario introdurre il concetto di stimatore, operandone le opportune distinzioni dal concetto di stima.

Lo stimatore, contrariamente alla stima, che rappresenta un singolo valore, è una variabile casuale descritta da una specifica distribuzione di probabilità. Empiricamente, se fosse possibile, tale distribuzione di probabilità si potrebbe ottenere generando tutti i possibili campioni estraibili dalla popolazione. Per ciascun campione estratto si dovrebbe ricavare la stima del parametro, alcune stime ovviamente risulterebbero uguali e la frequenza relativa dei campioni che hanno generato lo stesso valore della stima indicherebbe la probabilità di estrarre un campione la cui stima è quella osservata. Organizzando una distribuzione di frequenza per tutte le stime ottenibili, si otterrebbe quello che in statistica si chiama stimatore. In generale, però, questo procedimento non è fattibile, per cui si rende necessario ricorrere a risultati teorici o di convergenza asintotica per desumere la forma distribuzionale dello stimatore. Per chiarire il concetto su esposto facciamo un semplice esempio. Supponiamo di disporre di una popolazione costituita da sole 4 unità, A, B, C, e D, e supponiamo, inoltre, che alle quattro unità sia associata una misura di un carattere, i cui valori risultano $A=1$, $B=2$, $C=3$ e $D=4$. Decidiamo di estrarre dalla popolazione tutti i possibili campioni con rimpiazzo di $n = 2$ unità; l'universo campionario, per quanto detto, risulta essere quello riportato nella tabella seguente, dove, nella prima parte è indicato lo spazio campionario delle unità, ed affianco è indicata la distribuzione congiunta delle due variabili casuali "primo estratto" e "secondo estratto".

Ω				X_1, X_2			
AA	AB	AC	AD	(1;1)	(1;2)	(1;3)	(1;4)
BA	BB	BC	BD	(2;1)	(2;2)	(2;3)	(2;4)
CA	CB	CC	CD	(3;1)	(3;2)	(3;3)	(3;4)
DA	DB	DC	DD	(4;1)	(4;2)	(4;3)	(4;4)

Supponiamo, inoltre, di essere interessati alla stima del parametro

media della popolazione calcolata sulle nostre quattro unità, ossia

$$\mu = \frac{(1+2+3+4)}{4} = 2,5$$

che, a soli fini didattici, assumiamo nota. Una stima diretta del parametro μ è la media campionaria, ossia la media della variabile casuale doppia X_1, X_2 . In altri termini, la stima del parametro incognito è una funzione della variabile casuale doppia X_1, X_2 espressa da:

$$h(X_1, X_2) = \frac{X_1 + X_2}{2} = \hat{\mu}$$

Da quanto detto, è immediato intuire che, essendo X_1, X_2 una variabile casuale, una sua funzione $\hat{\mu} = h(X_1, X_2)$ varierà al variare dei possibili valori di X_1, X_2 . Naturalmente, tale funzione di X_1, X_2 è anch'essa una variabile casuale, che chiameremo stimatore. Una volta estratto il campione, per esempio AC, la variabile casuale X_1, X_2 assumerà i valori x_1, x_2 pari a (1,3), e la funzione stimatore, calcolata nel punto (1,3), risulta:

$$\bar{x} = \frac{x_1 + x_2}{2} = \frac{1+3}{2} = 2.$$

Il valore così ottenuto rappresenta la stima del parametro. Ripetendo il procedimento di stima per ciascun campione si ottiene:

X_1, X_2				$h(x_1, x_2) = \frac{x_1+x_2}{2} = \bar{x}$			
(1;1)	(1;2)	(1;3)	(1;4)	1	1,5	2	2,5
(2;1)	(2;2)	(2;3)	(2;4)	1,5	2	2,5	3
(3;1)	(3;2)	(3;3)	(3;4)	2	2,5	3	3,5
(4;1)	(4;2)	(4;3)	(4;4)	2,5	3	3,5	4

Raccogliendo i valori comuni delle stime possiamo ricavare una distribuzione di frequenza che rappresenta lo stimatore della media

della popolazione:

$\hat{\mu}$	n_i	p_i
1	1	1/16
1,5	2	2/16
2	3	3/16
2,5	4	4/16
3	3	3/16
3,5	2	2/16
4	1	1/16
	16	1

È utile soffermarci sul fatto che l'esempio appena riportato è stato svolto esclusivamente a fini didattici, poiché, per quanto abbiamo già detto, i casi concreti sono formati da un numero infinito, illimitato o finito, ma comunque elevato, di repliche. Ciò rende improbabile la generazione di tutti i possibili campioni. Tuttavia, la conoscenza della distribuzione dello stimatore è necessaria al fine dell'inferenza statistica e, come vedremo più avanti, essa sarà desunta dalla teoria delle probabilità che, sotto definite ipotesi, garantisce la forma esatta dello stimatore. Una valutazione della bontà dello stimatore adottato scaturirà dallo studio di questa distribuzione.

In generale, diciamo che, dato un campione di dimensione n , indicato con x_1, x_2, \dots, x_n , ciascuna delle n osservazioni campionarie è una determinazione della n -pla di variabili casuali "primo estratto", "secondo estratto", ..., "n-esimo estratto", X_1, X_2, \dots, X_n , che ha distribuzione di probabilità $f(x_1, x_2, \dots, x_n)$. Una funzione $h(X_1, X_2, \dots, X_n)$ è ancora una variabile casuale con una sua specifica distribuzione di probabilità, ed è detta stimatore. Il valore dello stimatore ottenuto una volta noto il campione, $h(x_1, x_2, \dots, x_n)$, è la stima del parametro della popolazione.

Al fine di dare maggiore chiarezza indicheremo con le lettere dell'alfabeto greco i parametri della popolazione che vogliamo stimare: ad esempio, la media sarà indicata con la lettera μ , la varianza con la lettera σ^2 e, in generale, si indicherà con θ un qualsiasi parametro della popolazione. Indicheremo, invece, con lo stesso alfabeto, ma segnato con accento circonflesso, lo stimatore del parametro; quindi si

utilizzerà $\hat{\mu}$ per indicare lo stimatore della media, $\hat{\sigma}^2$ per indicare lo stimatore della varianza e $\hat{\theta}$ per indicare un qualsiasi stimatore.

Utilizzeremo, invece, le lettere dell'alfabeto romano in minuscolo, ed a volte soprasssegnate da una barra orizzontale, per indicare la stima dei parametri; ad esempio si utilizzeranno \bar{x} per indicare la stima della media e s^2 per indicare la stima della varianza. Infine, nella trattazione teorica degli argomenti, indicheremo con $\hat{\theta} = h(X_1, X_2, \dots, X_n)$ lo stimatore generico e con $\hat{\theta} = h(x_1, x_2, \dots, x_n)$ la stima del parametro.

12.2 Proprietà degli stimatori

Abbiamo detto che la procedura di stima deve essere valutata per stabilire la sua attendibilità. Una fase necessaria per raggiungere questo obiettivo è la verifica delle proprietà auspicabili di un stimatore. A questo fine distinguiamo le proprietà per campioni di qualsiasi numerosità da quelle asintotiche, ossia per grandi campioni.

In generale, le proprietà di uno stimatore sono ricavate dalle caratteristiche della rispettiva distribuzione e, in particolare, dallo studio della sua sintesi e delle sue variabilità e forma.

Correttezza

Questa proprietà può anche essere letta come la valutazione della sintesi della distribuzione degli stimatori. Essa consiste nel valutare il valore medio di $\hat{\theta} = h(X_1, X_2, \dots, X_n)$.

Formalmente, sia $X = f(x, \theta)$ un modello teorico dal quale è stato estratto il campione x_1, x_2, \dots, x_n , come determinazione della n -pla di variabili casuali X_1, X_2, \dots, X_n ; sia, inoltre, $\hat{\theta} = h(X_1, X_2, \dots, X_n)$ lo stimatore di un parametro incognito; allora $\hat{\theta}$ si dice corretto se

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

cioè se la media della distribuzione dello stimatore coincide con il parametro della popolazione.

In caso contrario, cioè quando $E(\hat{\theta}) \neq \theta$, diremo che lo stimatore è distorto. In questo caso, si pone l'obiettivo di eliminare, o almeno di ridurre, la distorsione attraverso metodi e tecniche che trascureremo

in quanto non rientrano nelle finalità di questo testo. La misura della distorsione (*bias*) può essere data quindi da:

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

Riprendiamo l'esempio didattico riportato sopra e verifichiamo se lo stimatore $\hat{\mu}$ è corretto per il parametro μ della popolazione. Si tratta di calcolare la media della variabile casuale $\frac{X_1 + X_2}{2}$ ossia $E(\frac{X_1 + X_2}{2})$.

Disponendo della distribuzione empirica, si tratta di calcolare la media dei valori nella tabella riportata sopra, ossia:

$$1 \cdot 1/16 + 1,5 \cdot 2/16 + 2 \cdot 3/16 + 2,5 \cdot 4/16 + \\ + 3 \cdot 3/16 + 3,5 \cdot 2/16 + 4 \cdot 1/16 = 2,5.$$

Da questi banali calcoli si ricava che la media dello stimatore risulta

$$E\left(\frac{X_1 + X_2}{2}\right) = E(\hat{\mu}) = 2,5.$$

Ricordando che la media della popolazione nell'esempio è $\mu = 2,5$, si verifica immediatamente che le due quantità coincidono, verificando così che la media campionaria è una stima corretta della media della popolazione. Naturalmente, la verifica empirica è risultata possibile perchè abbiamo supposto di conoscere la popolazione e l'universo campionario.

È altrettanto intuitivo che nei casi concreti ciò non è possibile; infatti, in pratica, si dispone di un solo campione. Ciò implica che la correttezza non può essere valutata dal punto di vista empirico. La verifica della correttezza di uno stimatore è, infatti, un procedimento teorico che va effettuato, di volta in volta, sulla base delle caratteristiche dello stimatore.

Consistenza

La consistenza è una proprietà asintotica, nel senso che essa si deve verificare al crescere della numerosità campionaria. Uno stimatore si dice consistente se, al crescere di n (numerosità del campione), la distorsione diventa sempre più piccola e tende a zero. Ciò può essere espresso da un punto di vista formale come segue: sia $X = f(x, \theta)$ un

modello teorico dal quale è stato estratto il campione x_1, x_2, \dots, x_n , come determinazione della n -pla di variabili casuali X_1, X_2, \dots, X_n ; sia inoltre $\hat{\theta} = h(X_1, X_2, \dots, X_n)$, uno stimatore di θ ; si dice che $\hat{\theta}$ è consistente se $\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1$, dove ε è una quantità piccola a piacere.

Efficienza

La proprietà dell'efficienza può essere letta come la misura della variabilità della distribuzione di probabilità dello stimatore. Una simile valutazione è estremamente importante perchè fornisce gli elementi per stabilire la dispersione delle stime intorno al parametro incognito della popolazione, sotto la condizione che lo stimatore sia corretto. Il prerequisito della correttezza di uno stimatore, ai fini della valutazione dell'efficienza, è essenziale, in quanto si intuisce che uno stimatore con una bassissima variabilità, ad esempio prossima allo zero, ma distorto, implica che la stragrande maggioranza dei campioni fornisce una stima diversa dal reale parametro incognito della popolazione. Ciò equivale a dire che si ha un'altissima probabilità di commettere un errore nell'assumere un qualsiasi campione come strumento di stima del parametro. Il grafico riportato nella figura 12.1 illustra il concetto appena esposto.

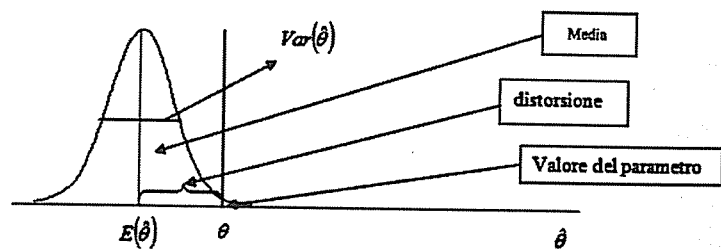


Figura 12.1: Distorsione dello stimatore

Da un punto di vista formale, il concetto dell'efficienza viene espresso in termini relativi, nel senso che si esprime l'efficienza di uno stimatore corretto rispetto ad un altro stimatore anch'esso corretto. Tra i due stimatori si dirà essere più efficiente quello con varianza minore.

Formalmente, sia $X = f(x, \theta)$ un modello teorico dal quale è stato estratto il campione x_1, x_2, \dots, x_n , come determinazione della n -pla di variabili casuali X_1, X_2, \dots, X_n ; inoltre siano $\hat{\theta}_1 = h(X_1, X_2, \dots, X_n)$ e $\hat{\theta}_2 = g(X_1, X_2, \dots, X_n)$ due stimatori, entrambi corretti, di θ , ossia $E(\hat{\theta}_1) = \theta$ e $E(\hat{\theta}_2) = \theta$; allora $\hat{\theta}_1$ si dice più efficiente di $\hat{\theta}_2$ se $VAR(\hat{\theta}_1) < VAR(\hat{\theta}_2)$.

Una rappresentazione grafica di tale concetto è riportata nella figura 12.2.

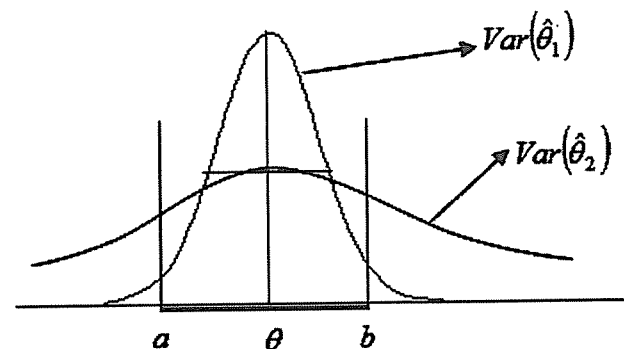


Figura 12.2: Confronto tra varianze di due stimatori

Infatti, si può vedere che, fissato l'intervallo $[a, b]$, contenente il parametro θ , l'area della distribuzione dello stimatore $\hat{\theta}_2 = g(X_1, X_2, \dots, X_n)$, calcolata in questo intervallo, è minore rispetto alla stessa calcolata per la distribuzione di $\hat{\theta}_1 = h(X_1, X_2, \dots, X_n)$. Ricordando che queste aree rappresentano la quantità di campioni che nell'universo forniscono stime appartenenti all'intervallo $[a, b]$, si intuisce il concetto e l'importanza dell'efficienza.

Va precisato che lo stimatore più efficiente ha un maggiore numero di campioni in un intervallo fissato intorno al parametro incognito della popolazione, così come, intuitivamente, può essere visto nel grafico riportato sopra.

A questo punto, nasce spontanea la domanda circa l'esistenza di un limite minimo teorico per la varianza di uno stimatore. L'utilità dell'introduzione di questo limite risiede nella possibilità di ricercare stimatori efficienti in senso assoluto, così come una misura della

distanza o un confronto con questo limite darebbero una valutazione dell'efficienza dello stimatore.

A questo proposito, esiste un risultato teorico fornito dal teorema di Cramer Rao valido solo per l'inferenza da modello. In tal caso, sia $\hat{\theta} = h(X_1, X_2, \dots, X_n)$ uno stimatore corretto di θ , si dimostra che:

$$\text{VAR}(\hat{\theta}) \geq \frac{1}{nE\left\{\left[\frac{\partial}{\partial \theta} \log f(x, \theta)\right]^2\right\}}$$

dove il secondo termine della disuguaglianza rappresenta il valore teorico della varianza minima di uno stimatore. Lo stimatore che raggiunge tale estremo inferiore si chiamerà stimatore "a varianza minima".

Stimatori con minimo errore quadratico medio

Un'utile proprietà di uno stimatore è quella dell'Errore Quadratico Medio (EQM) minimo. Questa proprietà è ampiamente utilizzata per valutare stimatori di parametri ottenuti attraverso processi simulativi, ossia quando si vuole tener conto sia della variabilità dello stimatore (efficienza) sia del suo valor medio (correttezza). Da un punto di vista formale, sia $X = f(x, \theta)$ un modello teorico dal quale è stato estratto il campione x_1, x_2, \dots, x_n , come determinazione della n -pla di variabili casuali X_1, X_2, \dots, X_n ; sia inoltre $\hat{\theta} = h(X_1, X_2, \dots, X_n)$ uno stimatore di θ ; si dice che $\hat{\theta}$ è uno stimatore con minimo errore quadratico medio, se il valore medio della differenza al quadrato tra il parametro e lo stimatore è minima, ossia:

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \text{Min.}$$

Dopo opportuni passaggi, sviluppando il quadrato e passando ai valori attesi, il MSE (*Mean Square Error*) può essere scritto anche come:

$$\text{MSE}(\hat{\theta}) = \text{VAR}(\hat{\theta}) + [B(\hat{\theta})]^2$$

da cui si evince che il MSE contiene sia il concetto di efficienza, sia il concetto di correttezza, attraverso la distorsione.

Sufficienza

Un'altra nota proprietà degli stimatori, per campioni di qualsiasi dimensione, è la sufficienza. Suddetta proprietà ha valenza prettamente teorica e la sua validità sarà chiarita più avanti, quando tratteremo l'inferenza da modello, per cui, in questa sede, ci limitiamo a fornirne soltanto alcuni principi essenziali. In generale, diciamo che uno stimatore $\hat{\theta} = h(X_1, X_2, \dots, X_n)$ è sufficiente per θ se esso riassume tutte le informazioni presenti nella variabile n -pla campionaria X_1, X_2, \dots, X_n .

In questo senso, lo stimatore si dice essere uno stimatore sufficiente per θ , se la sua distribuzione condizionale, dato $\hat{\theta}$, non dipende da θ (ciò significa che la distribuzione dello stimatore non è funzione del parametro).

Per verificare la sufficienza di uno stimatore si ricorre al teorema di fattorizzazione, di seguito presentato. Sia $X = f(x, \theta)$ un modello teorico dal quale è stato estratto il campione x_1, x_2, \dots, x_n , come determinazione della n -pla di variabili casuali X_1, X_2, \dots, X_n ; sia inoltre $\hat{\theta} = h(X_1, X_2, \dots, X_n)$ uno stimatore di θ ; allora $\hat{\theta}$ è sufficiente per θ se e solo se la distribuzione congiunta $f(x_1, x_2, \dots, x_n; \theta)$ della n -pla di variabili casuali X_1, X_2, \dots, X_n può essere fattorizzata in due funzioni:

$$f(x_1, x_2, \dots, x_n; \theta) = H[h(X_1, X_2, \dots, X_n); \theta]G(X_1, X_2, \dots, X_n)$$

dove la funzione $H(\cdot)$ dipende dal parametro θ e dalle variabili casuali X_1, X_2, \dots, X_n per il tramite dello stimatore $h(X_1, X_2, \dots, X_n)$, mentre la funzione $G(\cdot)$ dipende solo dalle variabili casuali X_1, X_2, \dots, X_n .

12.3 Inferenza da modello, un approccio basato sulla funzione di verosimiglianza

Nel capitolo relativo alle distribuzioni teoriche abbiamo detto che un fenomeno reale può essere surrogato da un modello teorico espresso da una funzione matematica. Abbiamo anche detto che i modelli teorici sono univocamente definiti dai parametri. Gli indici di sintesi, variabilità e forma sono ottenuti dal calcolo dei momenti che, a loro volta, dipendono dai parametri del modello. Sulla base di questi presupposti, in questo paragrafo svilupperemo i metodi di

stima dei parametri dei modelli teorici precedentemente trattati. In particolare, si tratterà dapprima la stima puntuale, quindi si passerà agli intervalli di confidenza, con lo scopo di effettuare una valutazione probabilistica delle stime dei parametri.

L'ambito disciplinare considerato è, dunque, quello dell'inferenza statistica da modello, detta anche parametrica, dove il costruito logico del metodo statistico consiste nell'estrarre un campione rappresentativo (casuale) dalla popolazione, sulla base del quale si procede alla deduzione (inferenza) sui parametri incogniti del modello.

In termini più generali, i modelli teorici possono essere visti come variabili casuali X con distribuzione di probabilità:

$$X \approx f(x, \theta)$$

dove $\theta = (\theta_1, \theta_2, \dots, \theta_s)$ sono i parametri incogniti da stimare.

Un modello teorico, quindi, dipende essenzialmente dai suoi parametri, che lo descrivono e lo definiscono. Tutto sta, allora, ad ottenere una buona stima di $\theta \in \Theta$, dove Θ indica l'insieme dei possibili parametri della distribuzione teorica $f(x, \theta)$. Schematicamente, la logica seguita si può riassumere come in figura 12.3.

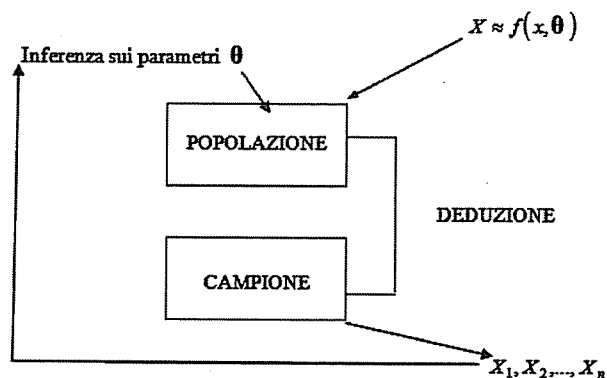


Figura 12.3: Inferenza sui parametri

Procediamo con un esempio. Supponiamo che un'azienda produttrice di lampadine voglia conoscere la durata media e la varianza del suo prodotto, allo scopo di fornirne le caratteristiche tecniche.

Possiamo ritenere che il fenomeno reale durata sia adeguatamente rappresentato da un modello teorico di tipo esponenziale il cui unico parametro da stimare è λ :

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

In questo caso, la popolazione composta da tutte le lampadine prodotte è completamente rappresentata dal modello teorico; quindi, estrarre un campione di n lampadine, su cui viene misurata la variabile casuale durata in ore, equivale a replicare n volte la v.c. durata. Se le estrazioni sono indipendenti, allora le repliche definiranno una nuova variabile casuale multipla costituita dalla n -pla X_1, X_2, \dots, X_n . In altri termini, avremo n variabili casuali, una per ciascuna estrazione. Se le estrazioni sono indipendenti l'una dall'altra, allora ciascuna v.c. avrà la stessa forma distribuzionale del modello decisionale da cui è stata estratta. Formalmente, si dirà che X_1, X_2, \dots, X_n sono n variabili casuali indipendenti ed identicamente distribuite (i.i.d.). Naturalmente, la nuova v.c. multipla, derivata dalla procedura campionaria, avrà a sua volta una distribuzione di probabilità, che deve essere definita e che dipenderà dai parametri del modello teorico.

Per meglio capire quanto abbiamo detto, facciamo ora un esempio banale finalizzato solo all'aspetto didattico. Supponiamo che la popolazione di riferimento sia quella definita nella figura 12.4.

POPOLAZIONE		A	B	C	D	
\bar{X}	1	2	3	4		$\mu = 2,5$

Figura 12.4: Popolazione di riferimento

In pratica, assumiamo che il modello teorico sia rappresentato dalla distribuzione discreta ed uniforme nell'intervallo $[1;4]$, ossia $X \approx U(1,4)$. È noto che i parametri della distribuzione uniforme sono gli estremi dell'intervallo, cioè 1 e 4, ed è altrettanto noto che la sintesi, cioè la media, corrispondente al momento primo, è data dalla semisomma degli estremi $\mu = (1 + 4)/2 = 2,5$. La rappresentazione grafica del modello teorico è, quindi, quella riportata nella figura 12.5.

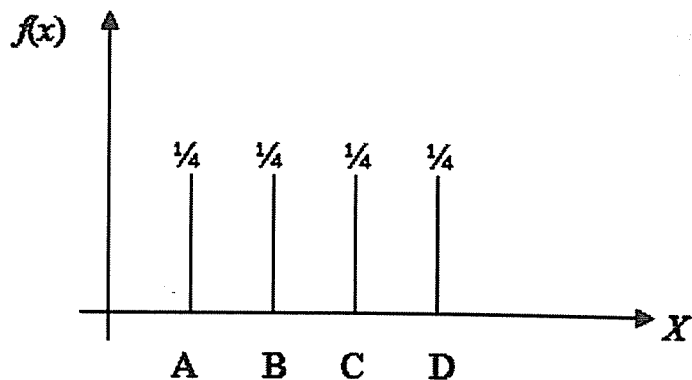


Figura 12.5: Modello teorico

Sempre al fine di comprendere la logica dell'inferenza da modello, supponiamo di estrarre tutti i possibili campioni con rimpiazzo di $n = 2$ unità dalla popolazione; come già visto in precedenza, l'universo campionario sarà:

Ω				X_1, X_2			
AA	AB	AC	AD	(1;1)	(1;2)	(1;3)	(1;4)
BA	BB	BC	BD	(2;1)	(2;2)	(2;3)	(2;4)
CA	CB	CC	CD	(3;1)	(3;2)	(3;3)	(3;4)
DA	DB	DC	DD	(4;1)	(4;2)	(4;3)	(4;4)

Segue che l'universo dei possibili campioni ha numerosità $N^n = 4^2 = 16$. Ogni campione può essere distinto in "primo estratto" e "secondo estratto". È immediato notare che, sia la variabile casuale "primo estratto" che quella "secondo estratto" sono distribuite identicamente al modello della popolazione, cioè

- v.c. primo estratto $X_1 \approx U(1,4)$
- v.c. secondo estratto $X_2 \approx U(1,4)$

V.C. primo estratto	Prob
1	$\frac{4}{16} = \frac{1}{4}$
2	$\frac{4}{16} = \frac{1}{4}$
3	$\frac{4}{16} = \frac{1}{4}$
4	$\frac{4}{16} = \frac{1}{4}$

v.c. secondo estratto	Prob
1	$\frac{4}{16} = \frac{1}{4}$
2	$\frac{4}{16} = \frac{1}{4}$
3	$\frac{4}{16} = \frac{1}{4}$
4	$\frac{4}{16} = \frac{1}{4}$

Dato il criterio di estrazione, dunque, le due variabili casuali sono indipendenti e identicamente distribuite. Esse sono indipendenti anche perché il sistema di estrazione che ha generato lo spazio campionario Ω è tale che il primo estratto e il secondo estratto non si influenzano reciprocamente. In generale, quindi, se il modello teorico è $X \approx f(x, \theta)$, allora il disegno sperimentale sarà tale che la n-pla campionaria X_1, X_2, \dots, X_n sarà definita dalle n repliche campionarie, ognuna con distribuzione di probabilità somigliante al modello teorico ipotizzato cioè:

$$X_1 \approx f(x_1, \theta), X_2 \approx f(x_2, \theta), \dots, X_n \approx f(x_n, \theta)$$

mentre la distribuzione di probabilità congiunta, $f(x_1, x_2, \dots, x_n; \theta)$, associata alla n-pla X_1, X_2, \dots, X_n , per il principio delle probabilità composte per eventi indipendenti, è data dal prodotto delle singole

densità di probabilità, ossia:

$$f(x_1, x_2, \dots, x_n; \theta) = f(x_1, \theta) \times f(x_2, \theta) \times \dots \times f(x_n, \theta) = \prod_{i=1}^n f(x_i, \theta).$$

Quest'ultima espressione ha la caratteristica di dipendere, oltre che dalla n-pla campionaria X_1, X_2, \dots, X_n , anche dai parametri $\theta = (\theta_1, \theta_2, \dots, \theta_s)$. Per meglio chiarire operativamente quanto appena detto, riprendiamo l'esempio della durata delle lampadine, espressa dal modello teorico esponenziale; in questo caso, l'esperimento campionario sarà tale che la n-pla campionaria X_1, X_2, \dots, X_n avrà distribuzione di probabilità congiunta

$$f(x_1, x_2, \dots, x_n, \lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum x_i}$$

dove si è posto nella espressione generale il modello esponenziale, la cui distribuzione è $f(x, \lambda) = \lambda e^{-\lambda x}$.

È immediato verificare che la funzione congiunta dipende sia dal parametro λ , che dalla variabile casuale multipla X_1, X_2, \dots, X_n . Tuttavia, una volta estratto il campione x_1, x_2, \dots, x_n , disporremo di una determinazione della variabile multipla X_1, X_2, \dots, X_n , il che equivale a dire che la funzione congiunta viene calcolata nel punto x_1, x_2, \dots, x_n . La distribuzione così ottenuta prende il nome di funzione di verosimiglianza ed in generale viene indicata con $L(\theta; x_1, x_2, \dots, x_n)$. Poichè il campione estratto è un vettore di numeri, ciò assicura che la funzione di verosimiglianza sarà funzione del solo parametro $\theta \in \Theta$.

Quindi, la distribuzione di probabilità congiunta ha due interpretazioni: la prima, prima di aver osservato il campione, come vera e propria distribuzione di probabilità; la seconda, dopo aver osservato il campione, come funzione del parametro θ che varia all'interno dello spazio dei parametri Θ .

Nel primo caso la distribuzione congiunta interpreta la probabilità che si verifichi la n-pla campionaria x_1, x_2, \dots, x_n ; mentre nel secondo caso, come abbiamo detto sopra, essa rappresenta la funzione di verosimiglianza ed in nessun modo rispetta le proprietà di una distribuzione di probabilità in quanto θ non è una variabile casuale.

Ritornando all'esempio della durata delle lampadine, una stima del parametro λ su base campionaria sarà data da quel valore di λ che massimizza la funzione

$$L(\lambda; x_1, x_2, \dots, x_n) = \lambda^n e^{-\lambda \sum x_i}.$$

Ciò equivale a cercare quel valore di λ tale che la distribuzione congiunta, ma condizionata al campione estratto, ha massima probabilità.

Questo metodo viene chiamato metodo della massima verosimiglianza. In termini generali, dunque, il metodo suddetto consiste nell'estrarre casualmente un campione di n unità da uno dei modelli teorici trattati nel corso delle distribuzioni teoriche, e nel massimizzare la funzione generale:

$$L(\theta; x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i, \theta).$$

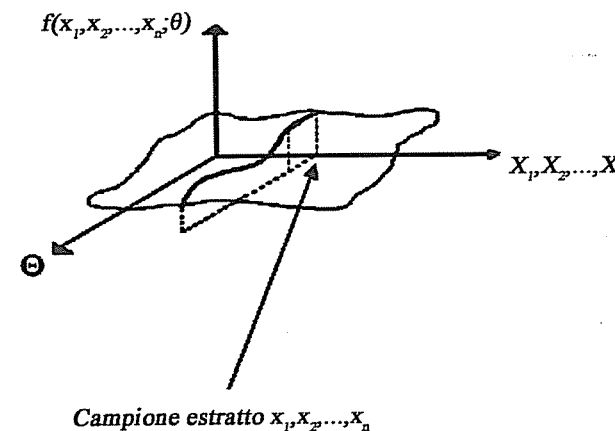


Figura 12.6: Processo di massimizzazione

Graficamente, il suddetto processo di massimizzazione, equivale a trovare il punto di massimo della funzione

$$L(\theta; x_1, x_2, \dots, x_n)$$

che, ricordiamo, varia nei parametri $\theta \in \Theta$, in un fissato campione x_1, x_2, \dots, x_n estratto, come illustrato in figura 12.6.

Come è noto dai corsi di analisi matematica, la ricerca del massimo di una funzione consiste nel risolvere il sistema delle derivate parziali e nel valutare la derivata seconda attraverso la matrice hessiana (o delle derivate seconde). Nel caso semplice di un solo parametro, si tratta di risolvere il seguente sistema:

$$\begin{cases} \frac{dL(\theta; x_1, x_2, \dots, x_n)}{d\theta} = 0 \\ \frac{d^2L(\theta; x_1, x_2, \dots, x_n)}{d\theta^2} < 0 \end{cases}$$

Il problema dell'inferenza statistica consiste nello stimare il parametro θ , conoscendo il campione di una popolazione. Distingueremo tali fasi in:

1. problema diretto, nel quale si estrae un campione dalla popolazione secondo uno dei vari disegni campionari;
2. problema inverso, ossia il problema di inferenza vero e proprio che, sulla base dei risultati campionari, fa risalire ai valori dei parametri della popolazione necessari per valutare il modello teorico oggetto di studio.

12.3.1 Inferenza classica

L'inferenza classica è anche detta inferenza basata sulla verosimiglianza. Essa si basa sul principio della stima dei parametri $\theta = (\theta_1, \theta_2, \dots, \theta_s)$, ossia sul procedimento con cui dal campione osservato si traggono informazioni per assegnare a $\theta = (\theta_1, \theta_2, \dots, \theta_s)$ un insieme di valori. Si distinguono due casi: nel primo si parlerà di stima puntuale, intendendo con essa il passaggio da un punto dello spazio campionario ad un punto dello spazio parametrico, che indichiamo con Θ . Nel secondo caso, invece, si farà riferimento al passaggio dallo spazio campionario ad un sottoinsieme (e, dunque, non più ad un punto) di Θ .

In generale, dunque, quando si parla di stima dei parametri si fa riferimento al procedimento (o metodo) con cui, sulla base delle informazioni tratte dal campione, si assegna al parametro incognito

della popolazione un valore o un insieme di valori. Nell'ambito della stima puntuale, il metodo che prendiamo in considerazione è quello della massima verosimiglianza, introdotto nel paragrafo precedente, e che, come sinteticamente ripetiamo, consiste nel massimizzare la funzione di verosimiglianza

$$L(\theta; x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i, \theta)$$

dato il campione.

Tra tutti i valori dei parametri $\theta \in \Theta$ si sceglie quello tale che

$$L(\theta; x_1, x_2, \dots, x_n)$$

sia massima.

Da un punto di vista analitico, si tratta di calcolare le derivate prime della funzione suddetta rispetto a θ ed eguagliarle a zero.

Esempio 1

Immaginiamo che ci siano delle elezioni politiche e che i cittadini siano chiamati a votare esprimendo la propria preferenza tra due soli candidati. La nostra variabile casuale sarà:

$$X = \begin{cases} \text{candidato A} = 1 \\ \text{candidato B} = 0 \end{cases}$$

Il candidato A ha probabilità $p(A) = \theta$ di essere eletto, e il candidato B ha probabilità $p(B) = 1 - \theta$ di essere eletto. Ciò vuol dire che questa variabile ha due sole possibili modalità, "candidato A" e "candidato B", che abbiamo codificato rispettivamente come "1" e "0". Il modello teorico rappresentativo di questo schema è la distribuzione bernoulliana la cui funzione è:

$$f(x; \theta) = \theta^x (1 - \theta)^{1-x}$$

Ipotizziamo di intervistare, attraverso un exit pool, un campione di n soggetti scelti casualmente dalla popolazione dei votanti; si ha che la funzione di verosimiglianza da massimizzare è:

$$L(\theta; x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i, \theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} =$$

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

dove x_1, x_2, \dots, x_n sono i risultati dello spoglio dell'exit pool. Per ragioni di calcolo è conveniente massimizzare il logaritmo di tale funzione che, essendo una funzione monotona della verosimiglianza, garantisce gli stessi punti di massimo e di minimo.

In sintesi, il problema si riduce nel massimizzare la funzione:

$$\log L(\theta; x_1, x_2, \dots, x_n) = \sum_{i=1}^n x_i \log \theta + (n - \sum_{i=1}^n x_i) \log(1-\theta)$$

Calcolando la derivata prima si ottiene:

$$\begin{aligned} \frac{d \log L(\theta; x_1, x_2, \dots, x_n)}{d\theta} &= \sum_{i=1}^n x_i \frac{1}{\theta} + (n - \sum_{i=1}^n x_i) \frac{1}{1-\theta} (-1) = \\ &= \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1-\theta} \end{aligned}$$

e ponendola uguale a zero si ha:

$$\frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1-\theta} = 0$$

per $\theta \neq 0$ e $\theta \neq 1$.

Risolviendo l'equazione rispetto alla variabile θ si ottiene:

$$\begin{aligned} \sum_{i=1}^n x_i (1-\theta) - (n - \sum_{i=1}^n x_i) \theta &= 0 \\ \sum_{i=1}^n x_i - \theta \sum_{i=1}^n x_i - n\theta + \theta \sum_{i=1}^n x_i &= 0 \\ \sum_{i=1}^n x_i - n\theta = 0 &\Rightarrow \hat{\theta} = \frac{\sum_{i=1}^n x_i}{n} \end{aligned}$$

Passando alla derivata seconda, si può verificare facilmente che essa è negativa; dunque si deduce che la stima di θ , data da $\hat{\theta} = \frac{\sum_{i=1}^n x_i}{n}$,

è un punto di ascissa in cui la funzione di verosimiglianza è massima. Formalmente, si ha:

$$\frac{d^2 \log L(\theta; x_1, x_2, \dots, x_n)}{d^2 \theta} = \frac{-\sum_{i=1}^n x_i}{\theta^2} - \frac{n - \sum_{i=1}^n x_i}{(1-\theta)^2}$$

e sostituendo $\hat{\theta} = \frac{\sum_{i=1}^n x_i}{n}$, otteniamo:

$$\begin{aligned} \frac{-\sum_{i=1}^n x_i}{\left(\frac{\sum_{i=1}^n x_i}{n}\right)^2} - \frac{n - \sum_{i=1}^n x_i}{\left(1 - \frac{\sum_{i=1}^n x_i}{n}\right)^2} &= -\sum_{i=1}^n x_i \frac{n^2}{\left(\sum_{i=1}^n x_i\right)^2} - (n - \sum_{i=1}^n x_i) \frac{n^2}{\left(n - \sum_{i=1}^n x_i\right)^2} = \\ &= -\frac{n^2}{\sum_{i=1}^n x_i} - \frac{n^2}{n - \sum_{i=1}^n x_i} = -\frac{n^2}{\sum_{i=1}^n x_i} - \frac{n^2}{n^2} + \frac{n^2}{\sum_{i=1}^n x_i} = -n \end{aligned}$$

Esempio 2

Come secondo esempio, riprendiamo il problema della durata delle lampadine. Sappiamo già che il modello teorico che riesce ad approssimare la durata è la funzione esponenziale, la cui espressione è $f(x; \lambda) = \lambda e^{-\lambda x}$ con $x > 0$. Supponiamo di aver estratto un campione di lampadine. Allora, la funzione di verosimiglianza è data da $L(\lambda; x_1, x_2, \dots, x_n) = \lambda^n e^{-\lambda \sum x_i}$. Passando al logaritmo si ottiene:

$$\log L(\lambda; x_1, x_2, \dots, x_n) = \log(\lambda^n e^{-\lambda \sum x_i}) = n \log \lambda - \lambda \sum_{i=1}^n x_i$$

Calcolando la derivata prima e eguagliandola a zero si ha:

$$\frac{d \log L(\lambda; x_1, x_2, \dots, x_n)}{d\lambda} = n \frac{1}{\lambda} - \sum_{i=1}^n x_i$$

$$n \frac{1}{\lambda} - \sum_{i=1}^n x_i = 0$$

per $\lambda \neq 0$. Risolvendo tale banale equazione di primo grado si ottiene $l = \frac{n}{\sum_{i=1}^n x_i}$. Passando alla derivata seconda si ha:

$$\frac{d^2 \log L(\lambda; x_1, x_2, \dots, x_n)}{d^2 \lambda} = -\frac{n}{\lambda^2}$$

e sostituendo $l = \frac{n}{\sum_{i=1}^n x_i}$, avremo che:

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

da cui risulta che $l = \frac{n}{\sum_{i=1}^n x_i}$ massimizza la funzione di verosimiglianza, essendo

$$-\frac{(\sum_{i=1}^n x_i)^2}{n} < 0$$

e, quindi, può essere assunta come stima del parametro λ .

Ci sono casi in cui non è possibile trovare una stima di massima verosimiglianza in forma chiusa, ossia non si riesce a trovare un risultato numerico sulla base del calcolo delle derivate.

In genere, questi modelli vengono risolti attraverso metodi di approssimazione, tipo Newton Rapsone che, per ragioni di semplicità, non saranno trattati in questa sede. In generale, nell'applicazione di tali metodi, si assegna un valore iniziale alla stima e poi si tenta di approssimare la funzione di verosimiglianza attraverso forme iterative.

In conclusione e a fini operativi, diciamo che per poter applicare il metodo della massima verosimiglianza abbiamo bisogno di due elementi fondamentali:

1. la distribuzione del carattere della popolazione, ossia il modello teorico, deve essere nota;
2. l'estrazione campionaria deve avere le caratteristiche di un campione probabilistico con osservazioni indipendenti.

12.4 Proprietà degli stimatori di massima verosimiglianza

Il metodo di stima di massima verosimiglianza fornisce, in sintesi, uno stimatore puntuale del parametro del modello decisionale studiato. Tuttavia, al variare dei possibili campioni, tale stima varia, definendo a sua volta una distribuzione di probabilità. In questo contesto, e per la trattazione degli argomenti che seguono, è necessario introdurre la funzione di log-verosimiglianza, definita da $\log L(\theta; X_1, X_2, \dots, X_n)$, la quale varia al variare delle realizzazioni della n-pla delle variabili casuali X_1, X_2, \dots, X_n , per cui lo stimatore di massima verosimiglianza è una v.c. espressa da tutti i valori che massimizzano

$$\log L(\theta; X_1, X_2, \dots, X_n)$$

al variare dell'universo campionario. Alla luce di questa considerazione è possibile derivare direttamente lo stimatore di massima verosimiglianza per mezzo della massimizzazione della funzione

$$\log L(\theta; X_1, X_2, \dots, X_n)$$

e la stima sarà ricavata sostituendo i valori osservati sul campione direttamente nello stimatore.

A questo punto ci possiamo interrogare circa le proprietà di cui gode uno stimatore di massima verosimiglianza. Senza addentrarci nelle dimostrazioni, diciamo che lo stimatore di massima verosimiglianza gode delle seguenti proprietà:

- è asintoticamente corretto, cioè al crescere della dimensione campionaria il valor medio dello stimatore converge al parametro della popolazione;
- è efficiente, infatti la varianza dello stimatore di massima verosimiglianza coincide con il limite inferiore di Cramer-Rao, facendo concludere che questo tipo di stimatore è anche efficiente in senso assoluto. Formalmente:

$$\text{var}(\hat{\theta}) = n \left\{ E \left[\frac{\partial}{\partial \theta} \log f(x, \theta) \right]^2 \right\}^{-1}$$

- è sufficiente, poichè, sotto le condizioni di validità per il procedimento di stima di massima verosimiglianza, vale il teorema di fattorizzazione di Fisher ossia: indicato con $X = f(x, \theta)$ un modello teorico dal quale è stato estratto il campione x_1, x_2, \dots, x_n , come determinazione della n-pla di variabili casuali X_1, X_2, \dots, X_n , allora $\hat{\theta} = h(X_1, X_2, \dots, X_n)$, è sufficiente per θ se e solo se esistono due funzioni non negative $g(\cdot)$ e $t(\cdot)$ tali che la funzione di verosimiglianza sia fattorizzabile in:

$$L(\theta; x_1, x_2, \dots, x_n) = g[h(x_1, x_2, \dots, x_n), \theta] t(x_1, x_2, \dots, x_n)$$

dove la funzione $g(\cdot)$ dipende dal parametro θ e dalle osservazioni campionarie solo attraverso la stima del parametro θ , mentre la funzione $t(\cdot)$ è una funzione del campione e non dipende dal parametro θ . Deduciamo questa proprietà con un esempio. Consideriamo ancora una volta il modello teorico di Poisson di parametro λ :

$$f(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Sia X_1, X_2, \dots, X_n , un campione i.i.d. estratto dal suddetto modello, allora la funzione di verosimiglianza, come abbiamo già visto, è:

$$L(\lambda; x_1, x_2, \dots, x_n) = \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!}$$

Per ottenere la stima procediamo con il metodo di massima verosimiglianza, da cui, come vedremo meglio nell'esempio riportato più avanti, risulta

$$l = \frac{n}{\sum_{i=1}^n x_i}$$

È immediato constatare che la funzione di verosimiglianza è fattorizzata in due funzioni: la prima

$$g[h(x_1, x_2, \dots, x_n), \lambda] = \lambda^{\sum_{i=1}^n x_i} e^{-n\lambda} = \lambda^{nl} e^{-n\lambda}$$

che dipende dal parametro λ e dalle osservazioni campionarie solo attraverso la stima del parametro λ ; la seconda

$$t(x_1, x_2, \dots, x_n) = \frac{1}{\prod_{i=1}^n x_i!}$$

che è funzione del campione e non dipende dal parametro; dunque, deduciamo che lo stimatore di massima verosimiglianza

$$\hat{\theta} = h(X_1, X_2, \dots, X_n)$$

è sufficiente per il parametro λ .

In generale, quindi, la sufficienza è strettamente legata alla funzione di verosimiglianza. Si può pertanto concludere che, se esiste uno stimatore sufficiente, esso sarà sicuramente uno stimatore di massima verosimiglianza. In altri termini, lo stimatore di massima verosimiglianza ha sempre la proprietà della sufficienza.

In sintesi possiamo dire che lo stimatore di massima verosimiglianza è ottimale. In particolare, si dimostra che esso è asintoticamente distribuito come una curva normale, potendo sostenere che se $\hat{\theta} = h(X_1, X_2, \dots, X_n)$ è uno stimatore di massima verosimiglianza, allora per $n \rightarrow \infty$

$$\hat{\theta} \xrightarrow{n \rightarrow \infty} N \left(\theta; \frac{1}{nE \left\{ \left[\frac{d}{d\theta} \log(x_1, x_2, \dots, x_n, \theta) \right]^2 \right\}} \right)$$

In termini pratici, diciamo che, per campioni grandi, lo stimatore di massima verosimiglianza ha distribuzione asintotica che converge alla distribuzione normale con media rappresentata dal parametro, e con varianza pari al limite inferiore di Cramer-Rao.

Un modo pratico di calcolare la varianza dello stimatore, partendo direttamente dalla funzione di verosimiglianza, è quello di passare attraverso l'informazione di Fisher. A questo scopo, si definisce informazione secondo Fisher, contenuta nella variabile casuale n-pla campionaria X_1, X_2, \dots, X_n , estratta dal modello $X \approx f(x, \theta)$,

l'espressione

$$I(\theta) = E \left[\frac{d}{d\theta} \log L(\theta; X_1, X_2, \dots, X_n) \right]^2$$

È possibile dimostrare che l'inverso dell'informazione di Fisher per il parametro θ , contenuta nel campione X_1, X_2, \dots, X_n , è uguale all'inverso del limite inferiore di Cramer-Rao, cioè:

$$I(\theta)^{-1} = \left\{ nE \left[\frac{d}{d\theta} \log f(x_1, x_2, \dots, x_n; \theta) \right]^2 \right\}^{-1}$$

In definitiva, abbiamo il seguente risultato generale per gli stimatori di massima verosimiglianza:

$$\hat{\theta} \xrightarrow{n \rightarrow \infty} N(\theta, [I(\theta)]^{-1})$$

Come vedremo successivamente, questo risultato sarà fondamentale per la costruzione di intervalli di confidenza e per la verifica di ipotesi sul parametro del modello teorico che si vuole studiare.

12.4.1 Massima verosimiglianza multiparametrica

Lo stimatore di massima verosimiglianza studiato nei paragrafi precedenti è limitato alla stima di un solo parametro. Vi sono, come abbiamo visto nel capitolo dei modelli teorici, casi in cui si rende necessario stimare più di un parametro. In questo caso, la procedura di stima deve essere estesa al caso multiparametrico.

Supponiamo, quindi, di avere un modello teorico $X \approx f(x, \theta)$, dove

$$\theta = (\theta_1, \theta_2, \dots, \theta_s)$$

rappresenta il vettore di parametri da stimare.

Si hanno, ad esempio, più parametri nel modello della curva normale, che ha parametri (μ, σ^2) . Supponiamo, come al solito, di estrarre dal modello $X \approx f(x, \theta)$ un campione indipendente x_1, x_2, \dots, x_n , allora la funzione di verosimiglianza sarà:

$$L(\theta; x_1, x_2, \dots, x_n) = \prod_{i=1}^n f(x_i, \theta_1, \theta_2, \dots, \theta_s)$$

dove

$$\theta = (\theta_1, \theta_2, \dots, \theta_s).$$

Ricordando che la funzione di log-verosimiglianza è monotona, per opportunità di calcolo si può ricorrere alla trasformata logaritmica, indicata con

$$\begin{aligned} \log L(\theta; X_1, X_2, \dots, X_n) &= \\ = \log L(\theta; X) &= \log \prod_{i=1}^n f(x_i, \theta_1, \theta_2, \dots, \theta_s) = \sum_{i=1}^n \log(f(x_i, \theta_1, \theta_2, \dots, \theta_s)) \end{aligned}$$

Uno stimatore di massima verosimiglianza per il vettore

$$\theta = (\theta_1, \theta_2, \dots, \theta_s)$$

può, allora, essere ottenuto a partire dalla soluzione del seguente sistema di derivate parziali:

$$\begin{cases} \frac{\partial L(\theta_1; X)}{\partial \theta_1} = 0 \\ \frac{\partial L(\theta_2; X)}{\partial \theta_2} = 0 \\ \vdots \\ \frac{\partial L(\theta_s; X)}{\partial \theta_s} = 0 \end{cases}$$

Sia $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_s)$ una soluzione del sistema, diremo che essa sarà uno stimatore di massima verosimiglianza del vettore dei parametri se la matrice delle derivate seconde (o hessiano) è definita negativa:

$$A = \begin{pmatrix} \frac{\partial^2 L}{\partial \theta_1^2} & \frac{\partial L}{\partial \theta_1 \theta_2} \cdots & \frac{\partial L}{\partial \theta_1 \theta_s} \\ \frac{\partial L}{\partial \theta_2 \theta_1} & \frac{\partial^2 L}{\partial \theta_2^2} \cdots \cdots & \frac{\partial L}{\partial \theta_2 \theta_s} \\ \frac{\partial L}{\partial \theta_s \theta_1} & \frac{\partial L}{\partial \theta_s \theta_2} \cdots & \frac{\partial^2 L}{\partial \theta_s^2} \end{pmatrix}$$

Passando al valore atteso e cambiando di segno si ricavano sulla diagonale principale le informazioni di Fisher per ogni parametro $E(-A) = I(\theta)$, da cui la matrice delle varianze degli stimatori di massima verosimiglianza è data da

$$\text{var}(\hat{\theta}) = I(\theta)^{-1}.$$

In definitiva, per n grande, abbiamo il seguente risultato generale per gli stimatori di massima verosimiglianza:

$$\hat{\theta} \xrightarrow{n \rightarrow \infty} \text{NMV}[\theta; I(\theta)^{-1}].$$

Esempio per un modello con un solo parametro

Supponiamo di voler stimare con il metodo della massima verosimiglianza il parametro λ del modello teorico di Poisson. Da quanto detto, si tratta di estrarre un campione indipendente dalla popolazione, il che significa replicare n volte la V.C. Poisson X_1, X_2, \dots, X_n di densità

$$f(x, \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

La funzione di verosimiglianza è allora data da:

$$L(\lambda; X_1, X_2, \dots, X_n) = \prod_{i=1}^n \lambda^{x_i} \frac{e^{-\lambda}}{x_i!} = \lambda^{\sum x_i} \frac{e^{-n\lambda}}{\prod x_i!}$$

e, passando al logaritmo, si ha:

$$\log L(\lambda; X_1, X_2, \dots, X_n) = \sum X_i \log \lambda - n\lambda - \log \prod X_i!$$

Calcolando la derivata prima rispetto a λ e eguagliandola a zero si ottiene:

$$\frac{d}{d\lambda} \log L(\lambda; X_1, X_2, \dots, X_n) = \frac{\sum X_i}{\lambda} - n = 0$$

da cui una stima del parametro, data da $l = \frac{\sum x_i}{n}$, si ottiene sostituendo allo stimatore i valori del campione estratto. Il calcolo della derivata seconda in questo punto verifica che $l = \frac{\sum x_i}{n}$ è un punto di massimo della funzione di verosimiglianza.

La varianza dello stimatore $\hat{\lambda} = \frac{\sum X_i}{n}$, da quanto sopra detto, può essere ricavata dall'informazione di Fisher, cioè tramite l'espressione

$$I(\lambda) = E \left[\frac{d}{d\lambda} \log L(\lambda; X_1, X_2, \dots, X_n) \right]^2$$

In particolare, si tratta di calcolare il momento secondo della derivata della verosimiglianza, ossia:

$$I(\lambda) = E \left[\frac{\sum X_i}{\lambda} - n \right]^2 = E \left[\frac{\sum X_i - n\lambda}{\lambda} \right]^2 = \frac{1}{\lambda^2} E \left[(\sum X_i - \sum \lambda)^2 \right] = \frac{1}{\lambda^2} E \left[\sum (X_i - \lambda)^2 \right] = \frac{1}{\lambda^2} \sum E (X_i - \lambda)^2$$

Ricordando che $E (X_i - \lambda)^2$ è la varianza del modello di Poisson che, come abbiamo visto, è uguale a λ , si ha :

$$I(\lambda) = \frac{1}{\lambda^2} \sum E (X_i - \lambda)^2 = \frac{1}{\lambda^2} \sum \lambda = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}$$

Essendo, infine, la varianza dello stimatore di massima verosimiglianza l'inverso dell'informazione di Fisher, si ha:

$$\text{var}(\hat{\lambda}) = \frac{\lambda}{n}$$

Sotto l'ipotesi di aver estratto un campione sufficientemente grande, si può asserire che:

$$\hat{\lambda} \approx N \left(\lambda; \frac{\lambda}{n} \right)$$

12.5 Stima per intervalli

Nel paragrafo precedente abbiamo introdotto le stime puntuali. L'uso di dette stime, tuttavia, non considera l'effetto dell'errore campionario, nel senso che difficilmente il valore della stima coincide con il valore del parametro. In questo paragrafo, svilupperemo la teoria statistica degli intervalli di confidenza. In termini pratici, si tenterà di costruire un intervallo che, con una prestabilita probabilità, contenga il parametro incognito.

Da un punto di vista formale, sia $\hat{\theta} = h(X_1, X_2, \dots, X_n)$, uno stimatore di θ del modello teorico $X = f(x, \theta)$, ottenuto da un campione casuale X_1, X_2, \dots, X_n ; allora, essendo lo stimatore una funzione dei dati campionari, esso sarà una variabile casuale con una definita distribuzione di probabilità. Nel caso in cui lo stimatore sia quello di massima verosimiglianza allora, per quanto visto sopra, esso sarà distribuito asintoticamente come una normale, con media il parametro del modello e varianza l'inverso dell'informazione di Fisher. Sotto la duplice condizione che lo stimatore $\hat{\theta} = h(X_1, X_2, \dots, X_n)$ sia corretto, cioè $E(\hat{\theta}) = \theta$, e che sia nota la sua distribuzione di probabilità, è facile, a partire da tale distribuzione, costruire un intervallo intorno

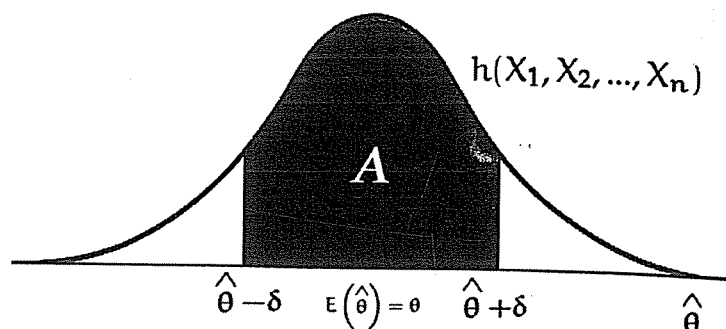


Figura 12.7: Intervallo di confidenza per il parametro θ .

al parametro θ , di ampiezza δ , così come viene mostrato nella figura 12.7.

Infatti, disponendo della distribuzione campionaria dello stimatore, è sempre possibile calcolare gli estremi di un intervallo di ampiezza δ , intorno al parametro θ . Gli estremi dell'intervallo, che abbiamo indicato con $\theta - \delta$ e $\theta + \delta$, si ricavano dalla distribuzione dello stimatore una volta fissata una probabilità A , soddisfacentemente ampia, a cui corrisponde la frequenza dei campioni che forniscono un sottoinsieme dello stimatore con al centro il suo valor medio $E(\hat{\theta})$. Quindi, una volta assicurata la correttezza dello stimatore

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

siamo certi che il sottoinsieme risultante contiene il parametro incognito del modello. Formalmente, indicando con A detta area, si ha:

$$A = \Pr(\theta - \delta \leq \hat{\theta} \leq \theta + \delta).$$

Si può immediatamente notare che l'intervallo di probabilità così ricavato contiene ai suoi estremi il parametro incognito, che non essendo noto non può essere definito. Inoltre, un simile intervallo non ha utilità statistica, in quanto fornisce un intervallo per la variabile casuale stimatore, contrariamente a quanto da noi richiesto,

ossia un intervallo intorno al parametro. Tuttavia, attraverso semplici operazioni algebriche, siamo in grado di ricavare l'espressione:

$$A = \Pr(\hat{\theta} - \delta \leq \theta \leq \hat{\theta} + \delta)$$

che prende il nome di intervallo di confidenza.

Analiticamente esso si ricava sottraendo la quantità θ a tutti i termini della disuguaglianza:

$$A = \Pr(-\delta \leq \hat{\theta} - \theta \leq \delta).$$

Poi, sottraendo nuovamente $\hat{\theta}$ a tutti i termini della disuguaglianza, si ottiene

$$A = \Pr(-\delta - \hat{\theta} \leq -\theta \leq \delta - \hat{\theta}).$$

Infine, cambiando di segno ed invertendo la disuguaglianza, si ha l'intervallo di confidenza cercato:

$$A = \Pr(\hat{\theta} - \delta \leq \theta \leq \hat{\theta} + \delta).$$

A questo punto, è necessario chiedersi qual è il significato statistico di questo nuovo intervallo e come lo stesso debba essere interpretato. A questo proposito, facciamo le seguenti considerazioni:

1. all'interno dell'intervallo abbiamo θ , che nel modello teorico è un numero e non una variabile. Quindi, dire che un numero varia all'interno di un intervallo vuol dire che l'incognita stessa può assumere tutti i valori dell'intervallo, dall'estremo inferiore a quello superiore compresi; ciò, tuttavia, è assurdo in quanto un numero, sebbene non noto, non può variare;
2. essendo δ l'unica grandezza che non varia, in quanto ricavabile una volta stabilita l'area di probabilità della distribuzione campionaria, A , possiamo concludere che l'intervallo è di ampiezza fissa;
3. $\hat{\theta}$ è la stima che attribuiamo al parametro dato il campione, ma sappiamo che essa varia al variare dei campioni. Ciò vuol dire che gli estremi non sono fissi, ma variabili secondo la forma della distribuzione campionaria; è chiaro che una volta assegnato il campione, gli estremi sono fissi.

In sintesi, l'intervallo $\hat{\theta} - \delta \leq \theta \leq \hat{\theta} + \delta$ è un intervallo di ampiezza fissa con estremi variabili a seconda del campione estratto.

Lo schema che segue mostra come, al variare dei campioni, l'intervallo di confidenza si muove lungo l'asse della variabile casuale dello stimatore.

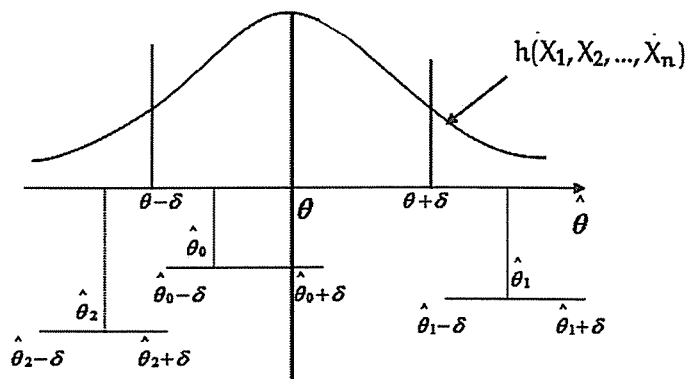


Figura 12.8: Variazioni dell'intervallo di confidenza.

Infatti, supponiamo di aver estratto un campione il cui valore stimato sia, così come riportato nel grafico, $\hat{\theta}_0$; l'intervallo di confidenza sarà di estremi $\hat{\theta}_0 - \delta$ e $\hat{\theta}_0 + \delta$. In questo caso, l'intervallo costruito avrà al suo interno il parametro θ .

Supponiamo, invece, di aver estratto un campione il cui valore della stima sia $\hat{\theta}_1$, allora l'intervallo di confidenza avrà estremi $\hat{\theta}_1 - \delta$ e $\hat{\theta}_1 + \delta$. In questo caso, invece, l'intervallo costruito non conterrà al suo interno il parametro θ .

Allo stesso modo, supponiamo di aver estratto un campione il cui valore della stima sia $\hat{\theta}_2$; l'intervallo di confidenza sarà di estremi $\hat{\theta}_2 - \delta$ e $\hat{\theta}_2 + \delta$. Anche in questo caso, l'intervallo non avrà al suo interno il parametro θ e così via per tutte le possibili stime. Ci rendiamo conto che tutte le stime, che sono all'interno dell'intervallo iniziale

$$A = \Pr(\hat{\theta} - \delta \leq \theta \leq \hat{\theta} + \delta)$$

forniranno un intervallo di confidenza tale da contenere il parametro θ della popolazione. Viceversa, tutte le stime che sono all'esterno di

tale intervallo forniranno un intervallo di confidenza che non conterrà il parametro della popolazione.

Possiamo allora chiederci: quanti campioni ci forniranno stime del primo tipo e quanti campioni ci forniranno stime del secondo tipo? Ricordando che ad ogni valore di $\hat{\theta}$ è associata una probabilità pari alla corrispondente ordinata, calcolata sulla distribuzione di probabilità dello stimatore, possiamo concludere che l'area A esprime la probabilità di estrarre un campione, la cui stima fornisce un intervallo di confidenza che contiene al suo interno il parametro incognito della popolazione, il che è diverso dal dire che A è la probabilità che il parametro θ cada all'interno dell'intervallo. Una volta estratto il campione allora lo stimatore fornisce la stima e l'intervallo assume estremi fissi.

Abbiamo detto che δ è l'ampiezza dell'intervallo e che dipende da A e viceversa. Ci rendiamo conto allora che l'ampiezza dell'intervallo di confidenza sarà variabile a seconda che la forma della distribuzione sia più o meno concentrata intorno al parametro. Infatti, a parità di A , per la curva più concentrata attorno al parametro, avremo un intervallo ridotto, e per la curva meno concentrata intorno al parametro, avremo un intervallo più ampio. Questo evidenzia l'importanza di ottenere stimatori efficienti. Riepilogando, gli elementi fondamentali per costruire intervalli di confidenza sono:

- conoscenza della distribuzione campionaria dello stimatore;
- stimatore corretto;
- livello di probabilità, chiamato di confidenza, desiderato;
- stime del parametro sulla base dei dati campionari.

Intervallo di confidenza per la media del modello teorico normale

Come esempio di un intervallo di confidenza, consideriamo il caso della stima del parametro media del modello teorico normale. Supponiamo, quindi, di disporre di un campione (x_1, x_2, \dots, x_n) , come determinazione della n -pla di v.c. campionarie (X_1, X_2, \dots, X_n) estratte dal modello $X \approx N(\mu, \sigma^2)$, e di voler costruire l'intervallo di confidenza per μ con una probabilità $A = 95\%$. Supponiamo di adottare come stimatore del parametro la media campionaria, $\hat{\mu} = \frac{\sum_{i=1}^n X_i}{n}$, la cui

stima si ottiene direttamente dai dati del campione estratto, attraverso l'espressione $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. E' facilmente dimostrabile che $\hat{\mu}$ è uno stimatore corretto di μ . Infatti:

$$E(\hat{\mu}) = E\left(\sum_{i=1}^n X_i\right) = \frac{1}{n} E\left(\sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} n\mu \Rightarrow E(\hat{\mu}) = \mu$$

La media campionaria $\hat{\mu}$ può anche essere vista come una combinazione lineare di variabili casuali normali¹, da cui si ricava il noto risultato che assicura che la forma distribuzionale dello stimatore $\hat{\mu}$, per una qualsiasi numerosità campionaria, è:

$$\hat{\mu} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \quad (12.5.0.1)$$

I quattro elementi precedentemente richiesti, utili per costruire l'intervallo di confidenza, sono, quindi:

- $\hat{\mu} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$;
- stimatore corretto;
- $A = 95\%$;
- $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$; σ^2 per ipotesi considerato noto.

Abbiamo già verificato che $\hat{\mu}$ è uno stimatore corretto del parametro μ . Dalla conoscenza della distribuzione dello stimatore possiamo procedere per ricavare il valore di δ ; per farlo, utilizziamo appunto la curva normale standardizzata che è stata opportunamente tabulata² e da cui è possibile ricavare i valori dell'ascissa per prefissati livelli di probabilità. In sintesi, dalla distribuzione dello stimatore della media campionaria standardizzata, la cui espressione è:

$$Z = \left(\frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}}\right) \quad \mu_z = 0, \sigma_z^2 = 1 \quad (12.5.0.2)$$

¹Maggiori dettagli sono riportati nel paragrafo 9.2.2.

²Si vedano le tavole della curva normale al paragrafo 15.2.1.

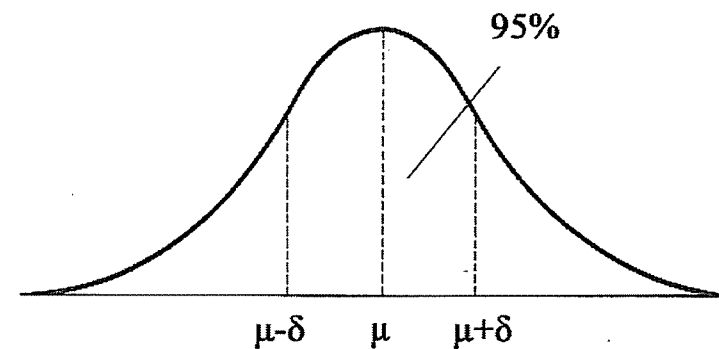


Figura 12.9: Intervallo di confidenza per la media

si ricavano i valori di δ in funzione del livello di probabilità A predefinito. In particolare, dalle tavole, si ricava che in corrispondenza di $A = 95\%$ si ha $\delta = 1,96$, da cui è possibile scrivere:

$$95\% = \Pr(-1.96 \leq Z \leq 1.96).$$

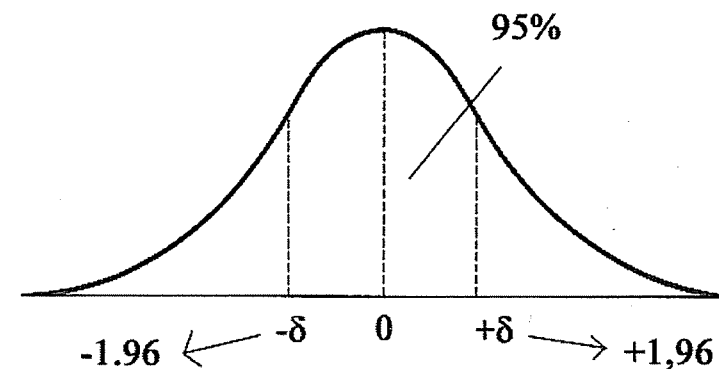


Figura 12.10: Intervallo di confidenza per la media, $\delta = 1.96$.

Sostituendo nella v.c. normale standardizzata la sua espressione si

ha:

$$95\% = \Pr \left(-1.96 \leq \frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq +1.96 \right).$$

Procedendo alle semplificazioni algebriche si ottiene:

$$95\% = \Pr \left(\hat{\mu} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu} + 1.96 \frac{\sigma}{\sqrt{n}} \right).$$

Sostituendo la stima della media $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, calcolata utilizzando i dati campionari, e \sqrt{n} otteniamo:

$$95\% = \Pr \left(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right) \quad (12.5.0.3)$$

dove σ^2 in questo caso è supposta nota.

Quest'ultima espressione è chiamata intervallo di confidenza per la media di una popolazione il cui modello teorico è quello di Gauss (curva normale).

Naturalmente, l'ipotesi della conoscenza del parametro σ^2 è raramente verificata. Quindi può accadere che σ^2 non sia noto. Allora, la variabile normale standardizzata

$$Z = \left(\frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}} \right) \quad (12.5.0.4)$$

non è definita dato che σ^2 non è noto. In questi casi si ricorre allo stimatore corretto di σ^2 , la cui espressione è:

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} \quad (12.5.0.5)$$

Infatti, si può dimostrare che $E(S^2) = \sigma^2$.

Ma se S^2 è uno stimatore, allora esso è anche una v.c. con una propria distribuzione di probabilità. Dalla teoria delle distribuzioni di probabilità si ricava il seguente risultato fondamentale:

$$\frac{(n-1)S^2}{\sigma^2} \approx \chi^2_{(n-1)} \quad (12.5.0.6)$$

che indica che la v.c. $\frac{(n-1)S^2}{\sigma^2}$ si distribuisce come un χ^2 con $(n-1)$ gradi di libertà³.

Inoltre, sempre dalla teoria delle distribuzioni di probabilità, si ricava la nota distribuzione della t-Student come il rapporto della v.c. normale standardizzata e la radice quadrata del χ^2 rapportato ai propri gradi di libertà. In particolare:

$$t = \frac{\left(\frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}} \right)}{\sqrt{\frac{(n-1)S^2}{\frac{\sigma^2}{n-1}}}} = \frac{\hat{\mu} - \mu}{\frac{S}{\sqrt{n}}} \quad (12.5.0.7)$$

dove $\frac{\hat{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}}$ è la v.c. normale standardizzata e $\sqrt{\frac{(n-1)S^2}{\frac{\sigma^2}{n-1}}}$ è una v.c. χ^2 in rapporto ai gradi di libertà.

Ricapitolando, possiamo concludere che, se (X_1, X_2, \dots, X_n) è una n-pla di v.c. campionarie estratte dal modello

$$X \approx \left(\mu, \frac{\sigma^2}{n} \right)$$

con μ e σ^2 non note, allora l'intervallo di confidenza per il parametro μ sarà:

$$A = \Pr(-\delta \leq t_{n-1} \leq +\delta)$$

Facendo ricorso alle tavole della t-Student in corrispondenza di $n-1$ gradi di libertà⁴, prestabilita la probabilità A , ricaviamo l'espressione

$$A = \Pr \left(\hat{\mu} - t_{\alpha, n-1} \frac{S}{\sqrt{n}} \leq \mu \leq \hat{\mu} + t_{\alpha, n-1} \frac{S}{\sqrt{n}} \right).$$

Analogamente a quanto visto sopra, sostituendo le stime $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ e $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$ si ottengono gli estremi dell'intervallo

³Per maggiori dettagli si consulti un testo specifico riportato in bibliografia. Mentre per il calcolo delle soglie del χ^2 in funzione dei gradi di libertà si consultino le tavole statistiche nella sezione χ^2 .

⁴I valori delle soglie della v.c. t-Student possono essere ricavati dalle tavole statistiche nella sezione t-Student.

di confidenza per la media, con σ^2 non noto e con probabilità prefissata A , la cui espressione è

$$A = \Pr \left(\bar{x} - t_{\alpha, n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha, n-1} \frac{s}{\sqrt{n}} \right).$$

Esempio

Supponiamo di aver estratto da un modello teorico distribuito normalmente, con media μ e varianza σ^2 incognite, un campione di numerosità 10, dal quale si sono ricavate le stime della media e della varianza, i cui valori sono $\bar{x} = 12$ e $s^2 = 16$.

Vogliamo trovare l'intervallo di confidenza per μ , con $A = 95\%$. Dalla lettura del problema si verifica facilmente che valgono le condizioni teoriche per applicare l'intervallo di confidenza sulla base della distribuzione t-Student, quindi possiamo scrivere

$$A = \Pr \left(\bar{x} - t_{\alpha, n-1} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha, n-1} \frac{s}{\sqrt{n}} \right).$$

Sostituendo i valori delle stime, e ricavando il valore di $\delta = 2.262$ dalle tavole della t-Student, con 9 gradi di libertà e $A = 95\%$, si ha:

$$95\% = \Pr \left(12 - 2.262 \frac{4}{\sqrt{10}} \leq \mu \leq 12 + 2.262 \frac{4}{\sqrt{10}} \right)$$

da cui, il risultato finale è:

$$95\% = \Pr (9.14 \leq \mu \leq 14.86)$$

È importante, a questo punto, richiamare un noto risultato asintotico della distribuzione t-Student; infatti, per grandi campioni, la v.c. t-Student converge in distribuzione a una curva normale con media zero e varianza 1, ossia:

$$T_{n-1} \xrightarrow{n \rightarrow \infty} N(0, 1)$$

Questo risultato risulta importante in quanto permette, in caso di grandi campioni, di calcolare i valori di δ direttamente sulle tavole della curva normale. Pertanto, l'intervallo di confidenza dello stimatore della media $\hat{\mu}$ diventa:

$$A = \Pr \left(\hat{\mu} - z_{\alpha} \frac{S}{\sqrt{n}} \leq \mu \leq \hat{\mu} + z_{\alpha} \frac{S}{\sqrt{n}} \right)$$

dove z_{α} è il valore di soglia della curva normale standardizzata in corrispondenza dell'area di probabilità A . Anche in questo caso, sostituendo i valori delle stime, otteniamo l'intervallo di confidenza cercato:

$$A = \Pr \left(\bar{x} - z_{\alpha} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha} \frac{S}{\sqrt{n}} \right)$$

12.6 Intervallo di confidenza per la varianza

Supponiamo di avere un modello teorico che si distribuisce normalmente, $X \approx N(\mu, \sigma^2)$, dal quale abbiamo estratto un campione *i.i.d* di n unità come determinazione della v.c. (X_1, X_2, \dots, X_n) . Siano inoltre

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

rispettivamente gli stimatori campionari corretti dei parametri μ e σ^2 .

Il problema è trovare l'intervallo di confidenza per il parametro σ^2 con un livello di probabilità pari ad A .

Il protocollo procedurale prevede innanzitutto di stabilire la distribuzione campionaria dello stimatore della varianza σ^2 .

A questo proposito, possiamo seguire due approcci: quello per grandi campioni, basato sulla verosimiglianza, e quello per piccoli campioni.

Approccio asintotico per grandi campioni

Da quanto detto nei paragrafi precedenti, se n è sufficientemente grande, si ha che, dalla funzione di log-verosimiglianza

$$\log L(\mu, \sigma^2; X_1, X_2, \dots, X_n)$$

è possibile ricavare gli stimatori per i parametri μ e σ^2 , e stabilire che la distribuzione degli stimatori è:

$$\begin{pmatrix} \hat{\mu} \\ S^2 \end{pmatrix} \xrightarrow{n \rightarrow \infty} N \left(\begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}; \begin{bmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{bmatrix} \right)$$

dove la matrice di varianza e covarianza è stata ricavata come inversa della matrice delle informazioni di Fisher $[I(\mu, \sigma^2)]^{-1}$.

Pertanto, la distribuzione campionaria asintotica per lo stimatore di massima verosimiglianza S^2 , dato μ , è:

$$S^2 \approx N\left(\sigma^2, \frac{2\sigma^4}{n}\right) \quad (12.6.0.8)$$

Sulla base di tale distribuzione, e procedendo in modo analogo a quanto detto sopra, si ricava l'intervallo di confidenza per σ^2 dopo aver sostituito le stime campionarie. In sintesi:

$$A = \Pr\left(s^2 - z_\alpha \frac{s^2\sqrt{2}}{\sqrt{n}} \leq \sigma^2 \leq s^2 + z_\alpha \frac{s^2\sqrt{2}}{\sqrt{n}}\right) \quad (12.6.0.9)$$

Approccio per piccoli campioni

Abbiamo già visto che lo stimatore della varianza ha la seguente forma $\frac{(n-1)S^2}{\sigma^2} \approx \chi_{n-1}^2$. A partire da esso possiamo scrivere $A = \Pr\left(\chi_{\frac{\alpha}{2}}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{1-\frac{\alpha}{2}}^2\right)$ il cui schema grafico è riportato in figura 12.11, e dove $\chi_{\frac{\alpha}{2}}^2$ e $\chi_{1-\frac{\alpha}{2}}^2$ rappresentano le soglie dell'intervallo che in precedenza sono state indicate con δ .

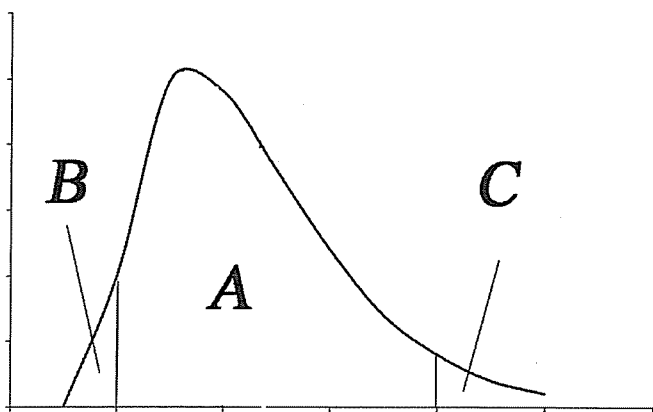


Figura 12.11: Intervallo di confidenza per la varianza.

Essendo la distribuzione del χ^2 asimmetrica, poniamo le aree delle code dell'intervallo pari a B e C. Vogliamo, inoltre, che le due aree B e C siano uguali, per cui, fissata la probabilità A, avremo $B = \frac{1-A}{2}$ e $C = \frac{1-A}{2} \Rightarrow B = \frac{\alpha}{2}$ e $C = \frac{\alpha}{2}$.

I valori di $\chi_{\frac{\alpha}{2}}^2$ sono tabulati in funzione dei gradi di libertà e in funzione di α .

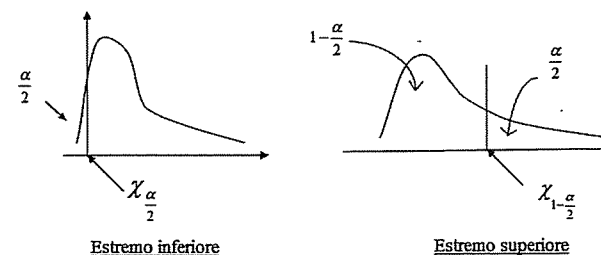


Figura 12.12: Estremi inferiore e superiore dell'intervallo di confidenza per la varianza.

Per quanto detto, l'intervallo di confidenza per il parametro σ^2 , ottenuto dopo una serie di passaggi algebrici e sostituendo le stime ricavate dal campione, diventa:

$$A = \Pr\left(\frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}}^2}\right) \quad (12.6.0.10)$$

12.7 Verifica delle ipotesi

Supponiamo che una ditta progetti lampadine la cui durata aspicata sia in media pari a un preciso valore. Ciò implica che il progettista ha una fondata idea sul valore del parametro del modello teorico "durata delle lampadine".

Il problema della verifica delle ipotesi consiste nello stabilire, sulla base di un campione estratto casualmente dalla popolazione delle lampadine, se si può confermare oppure smentire la supposizione progettuale iniziale. Si tratta di verificare un'ipotesi sul parametro del modello, sulla base di un campione casuale.

La pratica della verifica delle ipotesi è molto diffusa in diversi campi di ricerca: si applica nelle scienze mediche, per verificare se una terapia ha effetto oppure no; si applica al controllo della qualità, per verificare se il prodotto rispetta gli standard richiesti; in economia, per verificare se una politica economica ha raggiunto l'obiettivo desiderato, ecc. Ad esempio, immaginiamo di essere un'azienda che produce un semilavorato, il cui progetto prevede che gli ingranaggi siano distanti tra di loro 1 cm. Potrebbe, tuttavia, accadere che durante la produzione ci siano effetti di disturbo che alterano le caratteristiche del prodotto. È necessaria, allora, una tecnica statistica per assicurare l'idoneità del semilavorato. Un modo è vedere se la variabile "distanza tra gli ingranaggi" abbia un modello sottostante, così che una valutazione dei parametri permetterebbe di tenere sotto controllo, o addirittura ridurre, la difettosità.

Supponiamo di avere un modello teorico $x \approx f(x, \theta)$, e supponiamo inoltre che venga fatta un'affermazione sul parametro θ , di cui vogliamo verificare la veridicità. Da un punto di vista formale, possiamo scrivere:

$$\begin{cases} H_0 : \theta = \theta_0, & \text{detta ipotesi nulla} \\ H_1 : \theta \neq \theta_0, & \text{detta ipotesi alternativa} \end{cases} \quad (12.7.0.11)$$

L'ipotesi è detta *semplice* se è un'affermazione puntuale sul parametro (come, ad esempio, l'ipotesi nulla); è detta *composta* se si afferma un intervallo di valori per θ (come, ad esempio, l'ipotesi alternativa).

L'ipotesi è detta *semplice*, se è un'affermazione puntuale sul parametro (esempio: l'ipotesi nulla); è detta *composta*, se si afferma un intervallo di valori per θ (esempio: l'ipotesi alternativa). L'ipotesi alternativa espressa con $H_1 : \theta \neq \theta_0$ è detta *bidirezionale*, mentre quelle espresse con $H_0 : \theta > \theta_0$ o $H_0 : \theta < \theta_0$ sono dette *unidirezionali*.

Al solito per effettuare la verifica di ipotesi si estrae un campione casuale x_1, x_2, \dots, x_n come determinazione della v.c. X_1, X_2, \dots, X_n dal modello $X \simeq f(x, \theta)$. Sulla base dei valori campionari definiamo la funzione test che indichiamo con $t = t(X_1, X_2, \dots, X_n; \theta)$. Essendo t uno stimatore di θ , avrà una sua distribuzione chiamata *funzione test*. In particolare, se $\theta = \theta_0$, allora la distribuzione test sarà univocamente determinata dal parametro θ , in particolare se lo stimatore è uno stimatore corretto allora la funzione test sarà centrata in θ_0 .

Successivamente, la funzione test sarà divisa in due regioni:

- Regione *critica*, C , cioè l'insieme di tutti i campioni che forniscono una stima sostanzialmente diversa da θ_0 , facendo sospettare falsa l'ipotesi nulla H_0 ;
- Regione di *accettazione*, A , cioè l'insieme di tutti i campioni che forniscono una stima prossima a θ_0 , facendo sospettare vera l'ipotesi nulla H_0 .

Sostituendo i valori del campione estratto nella funzione test si ottiene il test statistico che indichiamo con $t = t(x_1, x_2, \dots, x_n; \theta_0) = t(x)$. Indicando con $d(x)$ la decisione assunta su base statistica si hanno due possibili decisioni, che in sintesi sono espresse come segue:

$$\begin{cases} d(x) = a_0 & \text{quando } t(x) \notin C \\ d(x) = a_1 & \text{quando } t(x) \in C \end{cases}$$

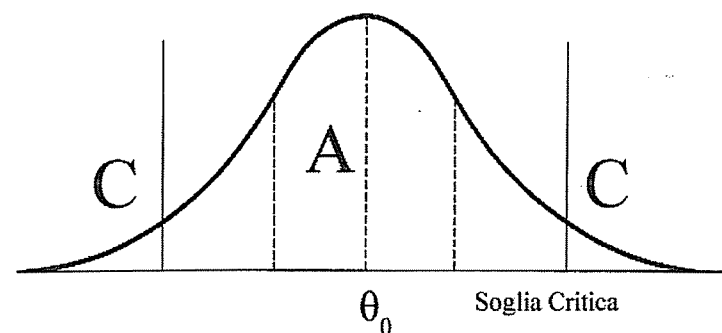


Figura 12.13: Rappresentazione grafica della verifica di ipotesi per $H_0 : \theta = \theta_0$.

Il test statistico si sviluppa pensando vera l'ipotesi nulla, allo scopo di fissare la distribuzione test in corrispondenza di una ipotesi semplice

$$H_0 : \theta = \theta_0.$$

Naturalmente, un valore sostanzialmente diverso dal valore sotto l'ipotesi nulla farà dubitare della sua veridicità.

Possiamo associare al test due decisioni, schematizzate in tabella 12.1, dove $d(x)$ è la decisione statistica.

Stati di natura del parametro	H_0	H_1
$\theta_0 = d(x) : t(x) \notin C$	$\theta \in \Theta_0$	$\theta \notin \Theta_0$
$\theta_1 = d(x) : t(x) \in C$	$\theta \notin \Theta_0$	$\theta \in \Theta_0$

Tabella 12.1: Decisioni associate ad un test per la verifica di ipotesi.

Naturalmente, se fissiamo anche l'ipotesi alternativa $H_1 : \theta = \theta_1$, allora si ottiene il grafico riportato nella figura 12.14.

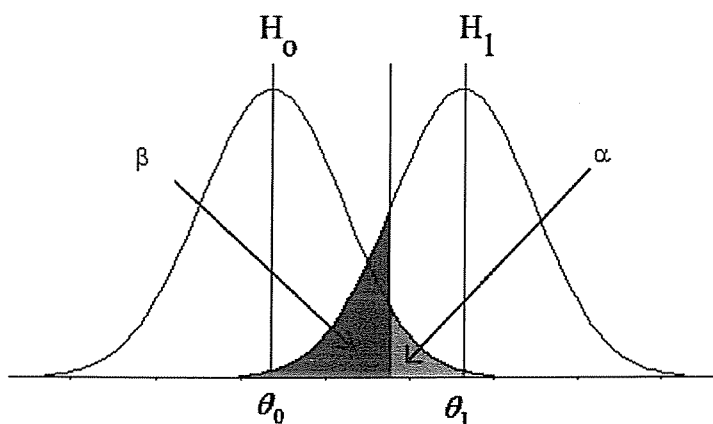


Figura 12.14: Rappresentazione della verifica di ipotesi fissati $H_0 : \theta = \theta_0$ e $H_1 : \theta = \theta_1$.

La prima curva nel grafico 12.14 (quella a sinistra) rappresenta il test statistico quando è vera l'ipotesi nulla, mentre l'altra curva rappresenta il test statistico quando è vera l'ipotesi alternativa.

Decisione	H_0	H_1
α_0	NO ERRORE	ERRORE DI II SPECIE
α_1	ERRORE DI I SPECIE	NO ERRORE

Tabella 12.2: Tipologie di errori nella verifica di ipotesi

Non si commette nessun errore se:

- la decisione è α_0 e l'ipotesi vera è H_0 ;
- la decisione è α_1 e l'ipotesi vera è H_1 .

Si commette un errore di prima specie α , se prendiamo la decisione α_1 , ma lo stato del parametro è θ_0 . Si commette un errore di seconda specie β , se prendiamo la decisione α_0 , ma lo stato del parametro è θ_1 . Tali errori sono schematizzati in tabella 12.2.

Una valutazione degli errori può essere fatta se si conosce la forma distribuzionale della funzione test, assegnata l'ipotesi. Infatti, stabilito il valore del parametro (ad esempio sotto l'ipotesi nulla $H_0 : \theta = \theta_0$), sarà univocamente determinato il valore di α . Allo stesso modo, se viene fissato il valore del parametro sotto $H_1 : \theta = \theta_1$, sarà univocamente determinato il valore di β . Tuttavia, va precisato che l'ipotesi alternativa, in genere, è composta, nel senso che assume valori in un intervallo continuo. Ciò implica che, se da un lato possiamo stabilire qual è l'errore connesso al rifiuto dell'ipotesi nulla, dall'altro non siamo in grado di valutare un equivalente valore dell'errore, nel caso di accettazione.

Formalmente, l'errore α può essere così espresso:

$$\alpha(\theta) = \Pr(t(x) \in C | \theta \in \Theta_0) \quad (12.7.0.12)$$

ed indica l'errore che si commette rifiutando l'ipotesi nulla quando questa è vera.

Allo stesso modo, l'errore β può essere così formalizzato:

$$\beta(\theta) = \Pr(t(x) \notin C | \theta \in \Theta_1) \quad (12.7.0.13)$$

e rappresenta l'errore che si commette accettando l'ipotesi nulla quando questa è falsa.

Poiché l'ipotesi nulla è l'ipotesi categorica, quindi il parametro θ_0 è noto a priori, la distribuzione di H_0 sarà sempre nota, e dunque $\alpha(\theta)$ sarà sempre noto. La stessa cosa non accade per l'ipotesi alternativa perché, per calcolare $\beta(\theta)$ bisogna, come si è detto sopra, stabilire il valore del parametro quando è vera H_1 .

Facendo variare i valori che possono essere assunti dal parametro nell'ipotesi H_1 , ricaviamo una curva che esprime il comportamento

dell'errore β . Questa funzione assume un ruolo di estrema importanza nella teoria dei test statistici per valutare la decisione assunta sul parametro θ_0 .

In particolare, si introduce la funzione "potenza del test", espressa dalla funzione $(1 - \beta(\theta_1))$, calcolata per ogni $\theta_1 \neq \theta_0$.

Quanto più H_1 è lontana da H_0 , tanto più β è piccolo e, dunque, la potenza del test $(1 - \beta)$ è grande; quanto più H_1 è vicina a H_0 , tanto più β è grande e, dunque, $(1 - \beta)$ è piccolo.

L'andamento della funzione di potenza del test è importante per la decisione statistica che prendiamo. Sappiamo che:

$$\beta(\theta) = \text{MAX}(1) \\ \beta(\theta) = \text{min}(0) \Rightarrow 0 \leq 1 - \beta(\theta) \leq 1.$$

Se il test cade nella zona critica si rifiuta l'ipotesi nulla, commettendo in questo caso un errore α che possiamo controllare. Il problema nasce quando il test non cade nella zona critica e si accetta l'ipotesi nulla. In questo caso, non sappiamo l'entità dell'errore β che potremmo commettere, diventa utile quindi studiare l'andamento del grafico della funzione di potenza del test.

Il grafico della funzione di potenza del nostro test ci dirà qual è l'andamento. Idealmente, la curva di potenza, come vediamo sotto, è una spezzata che vale α quando $\theta_1 = \theta_0$, mentre vale 1 per tutti i valori $\theta_1 \neq \theta_0$.

Quanto più la curva di potenza sarà vicina a questa curva ideale, tanto più il test sarà significativo.

Una rappresentazione grafica della funzione potenza del test, nel caso unidirezionale $\theta_1 > \theta_0$, è data in figura 12.15.

Proviamo a chiarire con un semplice esempio di verifica di ipotesi per il parametro μ di un modello di Gauss.

Immaginiamo che il fenomeno reale sotto osservazione sia un processo produttivo. Immaginiamo inoltre che i pezzi prodotti siano caratterizzati dalla lunghezza misurata in millimetri. In questo contesto è lecito supporre che il modello teorico che interpreta il fenomeno sia la curva normale la cui espressione sintetica è:

$$X = N(\mu, \sigma^2)$$

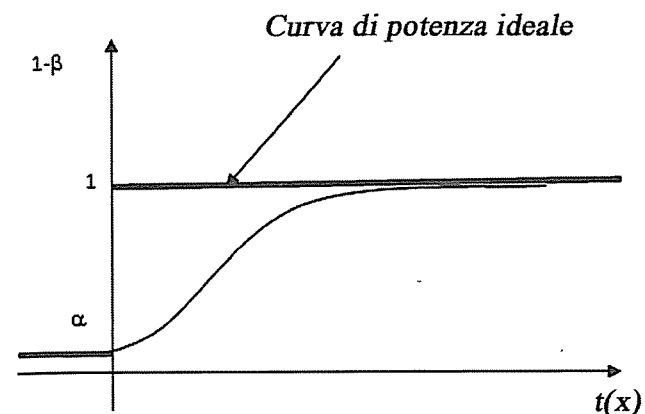


Figura 12.15: Curva di potenza ideale del test.

Nella fase di progettazione si stabilisce che i prodotti abbiano una lunghezza non superiore a 2000 mm e che la tolleranza, espressa attraverso lo scarto quadratico medio dei pezzi, non superi 50 mm. Il responsabile della qualità della fabbrica si pone di verificare se, nella fase di realizzazione dei prodotti, tali standard siano rispettati. In particolare pone l'attenzione solo sul parametro media dei prodotti e quindi ritiene, in questa fase, che il parametro varianza sia comunque rispettato cioè noto.

Procede quindi ad verifica di ipotesi statistica.

Dai dati di progetto ritiene che se tutto va per il verso giusto allora i prodotti devono avere una media della lunghezza pari a 2000 mm.

Quindi, sulla base di questa considerazione, l'ipotesi nulla è la seguente $H_0 : \mu = 2000$ mm. Potendoci essere fattori non progettati che alterano la produzione si pone il problema che in fase di produzione i pezzi possono avere una lunghezza media maggiore da quella di progetto e quindi non ammissibile in quanto si ammetterebbero solo prodotti con lunghezza non superiore a 2000 mm. Quindi stabilisce una ipotesi alternativa del tipo unidirezionale la cui espressione è $H_1 : \mu > 2000$ mm.

Dal magazzino dei prodotti preleva in modo casuale 100 pezzi la cui stima della media delle lunghezze risulta $\hat{\mu} = 2014$ mm.

È immediato verificare che i dati del problema rientrano nella teoria della verifica d'ipotesi sopra esposta.

Infatti dallo schema sperimentale si evince che il campione estratto può essere visto come una determinazione della $n - \text{pla}$ campionaria X_1, X_2, \dots, X_n con $n = 100$ estratta dal modello $X = N(\mu, \sigma^2)$. Pertanto, visto che il campione è sufficientemente grande, lo stimatore di massima verosimiglianza per il parametro media $\hat{\mu} = \frac{1}{n} \sum X_i$ la cui stima è 2014 mm, si distribuisce come una curva normale di espressione

$$\hat{\mu} = N\left(\mu, \frac{\sigma^2}{n}\right).$$

Sulla base di questi risultati possiamo costruire la funzione test sotto la validità dell'ipotesi nulla ossia

$$t(X) = t(X_1, X_2, \dots, X_n; \mu = 2000)$$

la cui espressione è:

$$t(X) = \frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma} = \frac{\sqrt{100}(\hat{\mu} - 2000)}{50}$$

Sostituendo i valori della stima si ottiene il test

$$t(x) = \frac{\sqrt{100}(2014 - 2000)}{50} = 2,8$$

sapendo che l'ipotesi alternativa è unidirezionale e considerando $\alpha = 0,05$ quale livello di significatività, dalla funzione test di forma normale standardizzata si delimitano due zone una critiche e l'altra di accettazione con soglia 1,65.

Possiamo dunque concludere che il test $t(x) = 2,8 > 1,65$ quindi cade nell'area critica di destra. Inducendo il responsabile della qualità a rifiutare l'ipotesi nulla con un rischio di commettere un errore di 1° specie (cioè di fare una affermazione errata) inferiore al 5 per cento. Pertanto segnalerà il problema al settore progettazione che prenderà le precauzioni necessarie.

Alternativamente senza fissare il valore di α , direttamente dalle tavole della curva normale standardizzata si può calcolare l'area della coda di destra delimitata dal test $t(x) = 2,8$ da cui si ricava che tale

area è uguale 0,003. In conclusione l'area critica minimale (p-value) è $\alpha > 0,003$ facendo sostenere la stessa decisione ma con un margine di errore di 1° specie addirittura inferiore.

Per tracciare la curva della potenza del test consideriamo il test statistico questa volta, però, in funzione del parametro μ_1 cioè sotto la validità dell'ipotesi alternativa $H_1 : \mu > 2000$ mm

$$t(\mu_1) = \frac{\sqrt{100}(2014 - \mu_1)}{50}$$

ora visto che il test cade nell'area critica di destra, prefissiamo un numero sufficiente di valori del parametro $\mu_1 > 2000$. Sulla base di ciascuno parametro assegnato calcoliamo il valore del test. Utilizzando le tavole della curva normale standardizzata, calcoliamo il valore della coda di destra dalla soglia ottenuta. Essi corrisponderanno alle ordinate della funzione di potenza $P(\mu_1) = 1 - \beta(\mu_1)$ calcolata per alcuni valori del parametro $\mu_1 \neq \mu_0$, i cui risultati sono riportati nella tabella 12.3. Sulla base di questi dati è possibile tracciare la curva di potenza come illustrato nel grafico 12.16.

μ_1	$t(\mu_1)$	$P(\mu_1)$
2000	2,8	0,003
2005	1,8	0,036
2010	0,8	0,212
2015	-0,2	0,579
2020	-1,2	0,885
2025	-2,2	0,986
2030	-3,2	0,999
2035	-4,2	1

Tabella 12.3: Valori della potenza del test.

12.8 Inferenza da popolazione finita

Quando le repliche del fenomeno reale sono un numero finito e la misura osservata è univocamente determinata su ciascuna re-

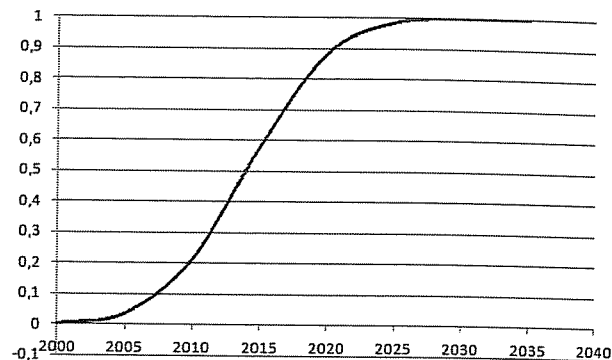


Figura 12.16: Curva di potenza del test

plica, parliamo di popolazione finita. In questo contesto, le N unità/repliche della popolazione sono indicate con u_1, u_2, \dots, u_N , dove i pedici $1, 2, \dots, N$ indicano le etichette, ossia l'identificativo delle unità della popolazione in oggetto. Naturalmente, ad ogni unità u_i è associata una variabile, ossia una misura del carattere di interesse, che indichiamo con y_i per $i = 1, 2, \dots, N$. Quindi, l'insieme di tutte le osservazioni può essere indicato mediante il vettore

$$\mathbf{y} = (y_1, y_2, \dots, y_N)^t.$$

Se la popolazione è finita, il vettore \mathbf{y} rappresenta un insieme fisso di costanti non note, quindi possiamo scrivere $\theta = \mathbf{y}$ per indicare i valori come un vettore di parametri incogniti.

L'estrazione casuale di un campione di dimensione n dalla popolazione è un sottoinsieme di etichette $c_0 = i_1, i_2, \dots, i_n$ (nel caso particolare di campionamento con ripetizione alcune di esse possono essere uguali). L'insieme di tutti i campioni estraibili dalla popolazione è indicato con C , mentre l'insieme di tutti i valori \mathbf{y} associati a ciascun campione è indicato con Y .

In altri termini, i dati campionari possono essere rappresentati dalle coppie formate dalle etichette e dai rispettivi valori, ad esse associati, ossia

$$d_0 = \{(i_1, y_{i_1}), (i_2, y_{i_2}), \dots, (i_n, y_{i_n})\} = \{(i, y_i) : i \in c_0\}.$$

Questa notazione può essere sintetizzata con l'espressione $d_0 = (c_0, \mathbf{y}_0)$, dove, come già detto, \mathbf{y}_0 è l'insieme dei dati osservati nel campione c_0 .

L'obiettivo dell'inferenza da popolazione finita è quello di stimare una funzione dei valori della popolazione $h(\mathbf{y})$ attraverso i valori osservati nel campione c_0 . Alcuni esempi di funzioni dei dati della popolazione che sono oggetto della stima sono il totale $h(\mathbf{y}) = \sum_{i=1}^N y_i = T$, la media $h(\mathbf{y}) = 1/N \sum_{i=1}^N y_i = \mu$, la varianza $h(\mathbf{y}) = 1/N \sum_{i=1}^N (y_i - \mu)^2 = \sigma^2$, e, in generale, qualsiasi indice di sintesi, di variabilità e forma incontrato nei capitoli precedenti.

La procedura casuale di estrazione del campione è chiamata disegno campionario, ed è specificata da una distribuzione di probabilità condizionata indicata con $p(c_0|\mathbf{y})$, ottenuta selezionando ogni campione $c_0 \in C$. I disegni campionari che non dipendono da \mathbf{y} sono indicati con $p(c_0)$ e si dicono *disegni convenzionali*. Rientrano in questa categoria il campionamento casuale semplice, con e senza ripetizione, il campionamento stratificato, ecc. Un esempio di disegno non convenzionale è il campionamento adattivo⁵.

12.8.1 Il campionamento casuale semplice

Sebbene non sia molto diffuso nella pratica delle indagini, il campionamento casuale semplice rappresenta il naturale punto di partenza per lo studio di tutti gli altri disegni campionari. Si consideri una popolazione di N unità dalla quale si debba estrarre un campione di n unità distinte. Il campionamento casuale semplice è la tecnica che attribuisce la stessa probabilità di selezione ad ogni insieme di n unità distinte della popolazione. Conseguenza è che anche ogni singola unità della popolazione ha la stessa probabilità di entrare a far parte del campione.

Nella selezione di un campione casuale è possibile scegliere se ogni unità possa entrare più di una volta nel campione. Se questa possibilità non è ammessa, il campionamento è detto senza ripetizione, altrimenti

⁵Per maggiori dettagli si consiglia di consultare S.K.T. Thompson, G.A.F. Seber, 1996 "Adaptive Sampling", Wiley NY. T. Di Battista (2003), "Resampling methods for estimating dispersion indices in random and adaptive design", Environmental and Ecological Statistics 10, Kluwer eds Academic Publishers (USA)

è detto con ripetizione. Nella pratica, l'estrazione con ripetizione viene adottata raramente. È intuitivo che, fissata la dimensione del campione, l'osservazione ripetuta di una o più unità rappresenta una perdita di informazione. Tuttavia, è anche evidente che la distinzione tra estrazione con e senza ripetizione perde gradualmente di importanza all'aumentare della dimensione della popolazione di rilevazione.

Il campionamento casuale semplice è uno dei principali metodi per ottenere campioni probabilistici. Esso può essere con ripetizione, se ad ogni estrazione si rimette l'elemento estratto nell'urna, o senza ripetizione in caso contrario.

Si consideri una popolazione di N elementi e si indichino con x_1, x_2, \dots, x_N i valori della variabile X in corrispondenza di ciascun elemento della popolazione. Ogni universo fornito di lista può essere inserito in un'urna che contiene N elementi, dalla quale si può procedere all'estrazione di un gruppo di $n < N$ elementi. Lo scopo di un campionamento è quello di estrarre un campione rappresentativo dell'intera popolazione, poiché i risultati di un'indagine statistica saranno tanto più attendibili quanto più il campione riesca a descrivere l'intera popolazione. Quando si effettua un *campionamento casuale semplice con rimpiazzo* la composizione dell'urna rimane invariata, poiché l'elemento estratto viene reinserto, quindi si potrebbero ottenere campioni che contengono più volte lo stesso elemento e si potrebbero perdere molte informazioni utili sulla popolazione totale. Per questo motivo è preferibile procedere al *campionamento casuale semplice senza rimpiazzo*.

Il più semplice dei disegni campionari è, appunto, quello casuale semplice senza rimpiazzo (*SRSWR*). Questo piano di campionamento prevede l'inserimento di tutte le unità della popolazione in un'urna e di estrarne $n < N$ da essa. In questo caso è immediato concepire una scelta casuale delle unità della popolazione. Alla prima estrazione avremo, quindi, N elementi della popolazione disponibili da estrarre, nella seconda estrazione $N - 1$ elementi e così via. Le unità possono essere estratte una alla volta o in blocco (ossia tutte insieme); in ogni caso l'universo dei campioni (cioè tutti i campioni possibili che hanno almeno una unità diversa) è dato dalle possibili combinazioni di N

elementi presi ad n a n . Formalmente si può scrivere⁶:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} \quad (12.8.1.1)$$

Ad esempio, se si volessero scegliere 3 bambini da una classe composta da 15, si avrebbero $\binom{15}{3} = \frac{15!}{3!12!} = 455$ possibilità diverse.

Data la casualità dell'estrazione, tutti i campioni hanno la stessa probabilità di essere estratti. Ricordiamo che la probabilità di un evento viene calcolata attraverso il rapporto tra casi favorevoli e casi possibili. Nel nostro caso specifico la probabilità di estrarre un campione su tutti i possibili campioni di una stessa numerosità è:

$$P(\text{di un campione}) = \frac{1}{\binom{N}{n}} \quad (12.8.1.2)$$

Questa espressione, infatti, presenta al numeratore il caso favorevole (campione di interesse) e al denominatore i casi totali (tutte le possibili combinazioni di campioni di numerosità n estraibili dall'intera numerosità della popolazione N).

Passando alla probabilità di inclusione di un'unità nel campione (*probabilità di inclusione del primo ordine*), per esempio la j -esima, essa è costituita da tutti i campioni che hanno l'elemento j -esimo ed $n - 1$ elementi degli $N - 1$ della popolazione, cioè:

$$\binom{N-1}{n-1} = \frac{(N-1)!}{(n-1)!(N-n)!} \quad (12.8.1.3)$$

Pertanto, la probabilità che la j -esima unità apparterrà al campione sarà data da:

$$P(u_i \in c) = \pi_j = \frac{\binom{N-1}{n-1}}{\frac{N!}{n!(N-n)!}} = \frac{n}{N} \quad (12.8.1.4)$$

Da ciò si deduce che tutte le unità hanno la stessa probabilità di appartenere al campione. Allo stesso modo è possibile calcolare

⁶Il simbolo $N!$ si legge N fattoriale. Il fattoriale è un'operazione algebrica in cui il numero a cui si applica il fattoriale viene moltiplicato per tutti i numeri interi precedenti ad esso fino a 1. Ad esempio $4! = 4 \times 3 \times 2 \times 1 = 24$.

la *probabilità del secondo ordine*, ossia la probabilità che un campione contenga contemporaneamente le unità u_i e u_j . Formalmente

$$P(u_i, u_j \in c) = \pi_{ij} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)} \quad (12.8.1.5)$$

quindi anche tutte le coppie di unità della popolazione hanno la stessa probabilità di essere incluse nel campione.

12.8.2 Stima del totale e della media

Il campionamento casuale semplice senza ripetizione è detto, per quanto visto sopra, a probabilità costanti; infatti, ogni unità ha la stessa probabilità di essere inclusa nel campione, per cui, nella trattazione che segue, ci limitiamo solo al vettore delle osservazioni campionarie $y = (y_1, y_2, \dots, y_N)^t$, trascurando le etichette, in quanto non necessarie.

Per ottenere uno stimatore corretto del totale, $h(y) = \sum_{i=1}^N y_i = T$, introduciamo lo stimatore di Horvitz-Thompson che, sebbene utilizzabile per ciascun tipo di disegno campionario con e senza ripetizione, meglio si presta per questo specifico disegno campionario.

Per un disegno campionario generico, esso ha la seguente espressione:

$$\hat{T}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i} \quad (12.8.2.1)$$

Ricordando che la probabilità di inclusione del primo ordine per il disegno campionario semplice con ripetizione è $\pi_i = \frac{n}{N}$, e, quindi, è costante per ogni unità, sostituendola nella espressione di sopra si ha:

$$\hat{T}_{HT} = \sum_{i=1}^n \frac{y_i}{\frac{n}{N}} = \frac{N}{n} \sum_{i=1}^n y_i \quad (12.8.2.2)$$

Il fattore $\frac{N}{n}$ prende il nome di **fattore di espansione**; la stima dell'ammontare totale di un carattere quantitativo è data, quindi, dall'ammontare totale campionario moltiplicato per il fattore di espansione.

Lo stimatore HT è corretto, nel senso che il suo valore atteso nell'universo dei campioni è uguale al parametro. Tuttavia, dalla trattazione

delle proprietà degli stimatori, abbiamo acquisito il concetto che, nonostante la correttezza, la maggior parte delle stime campionarie differirà, in più o in meno, dal parametro della popolazione.

In altre parole, le stime campionarie avranno una variabilità più o meno elevata intorno al valore centrale, rappresentato dal parametro della popolazione. È intuitivo che se questa variabilità è elevata allora sarà altrettanto elevata la probabilità che la stima di un campione casuale risulti anche molto diversa dal parametro della popolazione. Al contrario, se la variabilità è piccola, la distribuzione campionaria è non solo centrata, ma anche addensata sul parametro della popolazione e, di conseguenza, è alta la probabilità di selezionare casualmente campioni con stime prossime al parametro della popolazione. Il grado di addensamento della distribuzione campionaria intorno alla propria media è la proprietà dell'efficienza. Ribadiamo che essa esprime la precisione dello stimatore e si misura con un indice denominato *errore standard*. L'errore standard è pari alla radice quadrata della varianza della distribuzione campionaria delle stime.

Quindi, per avere una adeguata conoscenza del comportamento dello stimatore, è necessario esplicitare la sua varianza la cui espressione è:

$$\text{var}(\hat{T}_{HT}) = \sum_{i=1}^N \sum_{j>i} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \quad (12.8.2.3)$$

anche nota come espressione di *Sen-Yates-Grundy* per la varianza dello stimatore di *Horvitz-Thompson*.

Come si può notare, $\text{var}(\hat{T}_{HT})$ dipende dal vettore della variabile nella popolazione, $y = (y_1, y_2, \dots, y_N)^t$; infatti, l'estensione della sommatoria è $i = 1, 2, \dots, N$. Quindi, essa rappresenta un parametro della popolazione, che deve essere stimato. Uno stimatore corretto di $\text{var}(\hat{T}_{HT})$ è

$$\hat{\text{var}}(\hat{T}_{HT}) = \sum_{i=1}^N \sum_{j>i \text{ in } C} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \quad (12.8.2.4)$$

chiamato stimatore corretto di *Sen-Yates-Grundy* della varianza dello stimatore del totale ⁷ per campioni di dimensione prefissata.

Nel caso del campionamento casuale semplice senza ripetizione, riprendendo le probabilità di inclusione del primo e secondo ordine sopra esplicitate, e sostituendole nell'espressione 12.8.2.3, si ricava lo stimatore della varianza dello stimatore del totale HT. Senza addentrarci nei meandri dei calcoli, che lasciamo per esercitazione al lettore, esso ha seguente espressione:

$$\hat{v}ar(\hat{T}_{HT}) = \frac{N(N-n)}{n} \sigma^2 \quad (12.8.2.5)$$

dove $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2$ è la varianza della popolazione.

Ovviamente, nelle ricerche empiriche σ^2 non è nota e per questo deve essere stimata. Abbiamo già detto che un suo stimatore corretto è

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^n (y_i - \hat{\mu})^2$$

dove $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i$ è la media campionaria.

Sostituendo opportunamente σ^2 con s^2 si ottiene lo stimatore corretto della varianza dello stimatore totale cercato:

$$\hat{v}ar(\hat{T}_{HT}) = \frac{N(N-n)}{n} s^2$$

Infine, una stima numerica delle espressioni appena esplicitate è banalmente ricavabile dai dati del campione estratto, così come l'errore standard si calcola estraendo la radice quadrata di $\hat{v}ar(\hat{T}_{HT})$. Passando allo stimatore della media della popolazione è facile verificare che esso è dato dalla media campionaria, infatti:

$$\hat{\mu}_{HT} = \frac{\hat{T}_{HT}}{N} = \frac{1}{N} \frac{N}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n y_i \quad (12.8.2.6)$$

In modo analogo a quanto detto sopra, si può ricavare la varianza dello stimatore della media $\hat{v}ar(\hat{\mu}_{HT})$.

⁷I passaggi necessari per ricavare l'espressione della varianza dello stimatore HT sono, per ragioni di semplicità, trascurati. Si consiglia di consultare testi specifici sul campionamento riportati bibliografia.

In particolare, ricordando la relazione generale $var(ax) = a^2 var(x)$, dove a è una costante, possiamo scrivere:

$$\hat{v}ar(\hat{\mu}_{HT}) = var\left(\frac{\hat{T}_{HT}}{N}\right) = \frac{1}{N^2} var(\hat{T}_{HT})$$

essendo N una costante nota, corrispondente alla numerosità della popolazione.

Sostituendo, si ha l'espressione estesa della varianza dello stimatore della media campionaria.

Procedendo in modo analogo a quanto fatto per il totale, otteniamo prima lo stimatore e poi la stima di $\hat{v}ar(\hat{\mu}_{HT})$ che, per il caso del campionamento casuale semplice senza ripetizione, è:

$$\hat{v}ar(\hat{\mu}_{HT}) = \frac{N-n}{Nn} s^2 = \frac{N-n}{N(n-1)} \frac{\sum_{i=1}^n (y_i - \hat{\mu})^2}{n} \quad (12.8.2.7)$$

L'errore standard si ricava estraendo la radice quadrata della varianza. Quindi, dopo aver stimato l'errore standard del totale e/o della media, è possibile assumere la normalità della distribuzione dello stimatore ⁸; in particolare devono valere le condizioni $N; n \rightarrow \infty$, in modo che $(N-n) \rightarrow \infty$.

È possibile, per N e n sufficientemente elevati, costruire intervalli di confidenza centrati rispetto alla media ed al totale; cioè, è possibile individuare due valori, che chiameremo estremi dell'intervallo, che, con una prestabilita probabilità, riteniamo contenere al loro interno il parametro della popolazione. In particolare, sotto queste condizioni, gli intervalli di confidenza, al 95% di probabilità, per il totale e per la media della popolazione sono, rispettivamente:

$$\hat{T}_{HT} \pm 1.96 \sqrt{\hat{v}ar(\hat{T}_{HT})}$$

$$\hat{\mu}_{HT} \pm 1.96 \sqrt{\hat{v}ar(\hat{\mu}_{HT})}.$$

Consideriamo i casi in cui la variabile è rappresentata dalla presenza o assenza di una modalità qualitativa di una popolazione finita,

⁸Per ulteriori approfondimenti circa la normalità degli stimatori di Horvitz Thompson si consiglia la consultazione dell'articolo di Hájek (1961) riportato in bibliografia.

ossia:

$$y_i = \begin{cases} 1 & \text{se la } i\text{-esima unità ha la modalità oggetto d'indagine} \\ 0 & \text{se la } i\text{-esima unità non ha la modalità oggetto d'indagine} \end{cases}$$

Ad esempio, se in un collettivo di N studenti universitari siamo interessati a conoscere quanti sono quelli che hanno il diploma di liceo scientifico, allora $y_i = 1$ se l' i -esimo studente ha il diploma di liceo scientifico, mentre $y_i = 0$ in tutti gli altri casi. In questo modo il vettore delle osservazioni della popolazione $\mathbf{y} = (y_1, y_2, \dots, y_N)^t$ sarà un elenco di zeri ed uno. Il parametro totale $h(\mathbf{y}) = \sum_{i=1}^N y_i = T$ indica la frequenza assoluta della modalità d'interesse (numero di studenti con diploma di liceo scientifico), mentre il parametro media della popolazione $h(\mathbf{y}) = \frac{1}{N} \sum_{i=1}^N y_i = \mu$ rappresenta la frequenza relativa della stessa modalità, che moltiplicata per cento indica la percentuale (la percentuale degli studenti con diploma di liceo scientifico). La frequenza relativa può essere letta anche come la probabilità di estrarre dal collettivo degli studenti uno studente con diploma di liceo scientifico.

In questo caso gli stimatori di Horvitz-Thompson della frequenza assoluta e della frequenza relativa rimangono invariati.

Infatti, indicando con A la modalità di interesse del carattere, si ha che lo stimatore della frequenza assoluta è:

$$\hat{n}_A = \frac{N}{n} \sum_{i=1}^n y_i \quad (12.8.2.8)$$

e lo stimatore della frequenza relativa, invece, è:

$$\hat{f}_A = \frac{1}{n} \sum_{i=1}^n y_i \quad (12.8.2.9)$$

Gli stimatori corretti delle rispettive varianze sono:

$$\hat{\text{var}}(\hat{n}_A) = \frac{N(N-n)}{n-1} \hat{f}_A(1-\hat{f}_A) \quad (12.8.2.10)$$

$$\hat{\text{var}}(\hat{f}_A) = \frac{N-n}{N(n-1)} \hat{f}_A(1-\hat{f}_A) \quad (12.8.2.11)$$

Facciamo un esempio banale e utile ai soli fini didattici. Supponiamo di indagare la popolazione di una squadra di calcio composta da sole quattro unità, $N = 4$, quali Arbeloa, Buffon, Cavani e Del Piero. Supponiamo di essere interessati alla variabile "ingaggio stagionale in milioni di euro". I dati della popolazione, che qui supponiamo di conoscere (ma questo non accade nei casi concreti) sono riportati nella tabella 12.4.

Etichette	Unità u_i	Variabile ingaggio stagionale in milioni di euro
i_1	Arbeloa	0,8
i_2	Buffon	1,2
i_3	Cavani	0,9
i_4	Del Piero	1,8

Tabella 12.4: Esempio: dati della popolazione.

Disponendo dei dati della popolazione possiamo calcolare i parametri essenziali delle due variabili. Ad esempio, il totale degli ingaggi stagionali in milioni di euro, sostenuti dalla società di calcio in oggetto, è $T = 4,7$ milioni di euro. La media degli ingaggi di ciascun giocatore è, invece, $\mu = 1,175$ milioni di euro.

Supponiamo, per opportunità didattica, che non siano noti tali valori; quindi, proviamo a stimare questi parametri estraendo un campione casuale semplice senza rimpiazzo di $n = 2$ unità dalla popolazione. Dalla estrazione casuale sono risultate estratte le etichette i_2 e i_3 , corrispondenti alle unità Buffon e Cavani.

Gli stimatori di Horvitz-Thompson del totale e della media degli ingaggi sono:

$$\hat{T}_{HT} = \frac{N}{n} \sum_{i=1}^n y_i = \frac{4}{2}(1.2 + 0.9) = 4.2$$

$$\hat{\mu}_{HT} = \frac{\hat{T}_{HT}}{N} = 1.05.$$

Le stime dei rispettivi errori standard sono:

$$\begin{aligned}\sqrt{\widehat{\text{var}}(\hat{T}_{HT})} &= \sqrt{\frac{N(N-n)}{n} s^2} \\ &= \sqrt{\frac{4(4-2)}{2} [(1,2 - 1,05)^2 + (0,9 - 1,05)^2]} = 0,42\end{aligned}$$

$$\begin{aligned}\sqrt{\widehat{\text{var}}(\hat{\mu}_{HT})} &= \sqrt{\frac{N-n}{N(n-1)} \left(\sum_{i=1}^n (y_i - \hat{\mu})^2 \right) / n} \\ &= \sqrt{\frac{4-2}{4(1)} [(1,2 - 1,05)^2 + (0,9 - 1,05)^2 / 2]} = 0,27.\end{aligned}$$

Pur non valendo le condizioni di normalità, in quanto nell'esempio N e n sono molto piccoli, ma solo per esigenze didattiche costruiamo gli intervalli di confidenza al 95% di probabilità. Sotto tale ipotesi (di fatto non verificata) essi sono rispettivamente per il totale e la media degli ingaggi i seguenti:

$$\hat{T}_{HT} \pm 1,96 \sqrt{\widehat{\text{var}}(\hat{T}_{HT})} = 4,2 \pm 1,96(0,42)$$

da cui si desume che con probabilità del 95% la spesa totale per ingaggio della società di calcio è compresa tra 3,37 e 5,03 milioni di euro. Procedendo in modo analogo per la media si ha

$$\hat{\mu}_{HT} \pm 1,96 \sqrt{\widehat{\text{var}}(\hat{\mu}_{HT})} = 1,05 \pm 1,96(0,27)$$

Da cui si desume che con probabilità del 95% la spesa media per ingaggio della società di calcio è compresa tra 0,52 e 1,58 milioni di euro.

Come secondo esempio supponiamo che dai 450 studenti iscritti al primo a.c. 2011-12 di Scienze della Formazione è stato estratto un campione di 36 studenti a cui è stato rilevato il tipo di diploma di scuola secondaria di II grado, allo scopo di conoscere il numero e la percentuale sul totale di studenti che provengono dal Liceo Scientifico. Da quanto detto sopra si è introdotta la variabile indicatrice

$y_i = 1$ se la i -esima unità ha il diploma dello scientifico
 $y_i = 0$ se la i -esima unità ha un qualsiasi altro diploma

Sapendo che la somma degli uno nella sequenza delle osservazione campanarie è 12 possiamo facilmente ottenere le stime del totale e della frequenza relativa come segue:

$$\hat{n}_A = \frac{N}{n} \sum_{i=1}^n y_i = 450/36(12) = 150$$

e

$$\hat{f}_A = \frac{1}{n} \sum_{i=1}^n y_i = 12/36 = 0,33$$

Le stime delle varianze degli stimatori sono

$$\begin{aligned}\widehat{\text{var}}(\hat{n}_A) &= \frac{N(N-n)}{n-1} \hat{f}_A (1 - \hat{f}_A) = \\ &= \frac{450(450-36)}{36-1} 0,33(0,67) = 1176,88\end{aligned}$$

e

$$\widehat{\text{var}}(\hat{f}_A) = \frac{N-n}{N(n-1)} \hat{f}_A (1 - \hat{f}_A) = \frac{450-36}{450(36-1)} 0,33(0,67) = 0,006$$

Infine gli intervalli di confidenza al 95%, verificata la condizione di normalità degli stimatori sono:

$$\hat{n}_A \pm 1,96 \sqrt{\widehat{\text{var}}(\hat{n}_A)} = 150 \pm 1,96 \sqrt{1176,88}$$

$$\hat{f}_A \pm 1,96 \sqrt{\widehat{\text{var}}(\hat{f}_A)} = 0,33 \pm 1,96 \sqrt{0,006}$$

Quindi con probabilità del 95% il numero di studenti con diploma di Liceo Scientifico è compreso tra 82,77, che poniamo uguale a 80 (la frequenza assoluta non può essere decimale), e 217,23 che approssimiamo a 217. Mentre per la frequenza relativa l'intervallo al 95% sarà compresa tra 0,18 e 0,48.

12.9 Altri disegni campionari

12.9.1 Il campionamento stratificato

Il campionamento stratificato è molto utile quando lo studio di un fenomeno porta a diversificazioni nei risultati, sulla base delle caratteristiche delle unità statistiche appartenenti alla popolazione oggetto di studio. Supponiamo di voler studiare il numero di crediti acquisiti dagli studenti iscritti alla Laurea Triennale di Scienze della Formazione. Si capisce bene che non è possibile procedere ad un campionamento casuale senza rimpiazzo, poiché potrebbero entrare a far parte del campione studenti sia del I, sia del II, sia del III anno, in maniera del tutto casuale e ineguale, correndo il rischio che il campione sia formato, ad esempio, da troppi studenti del I anno, pochi del II e nessuno del III. In questo modo si corre il rischio di osservare un fenomeno che non rispecchia la realtà, a causa di una costruzione di un campione non rappresentativo della popolazione oggetto di studio. Si intuisce che il numero di crediti di un iscritto al I anno è molto inferiore al numero di crediti di un iscritto al II, così come quest'ultimo presenta un numero di CFU inferiori a quelli di un iscritto al III anno. Questa informazione sulla popolazione è disponibile dalla lista di iscritti alla Laurea Triennale, per cui è possibile applicare questo tipo di campionamento. L'aggettivo "stratificato" sta a significare proprio la creazione di strati all'interno della popolazione; nel nostro esempio gli strati possono essere costituiti dall'anno di corso a cui lo studente è iscritto. In questo modo si giunge ad un numero di strati (che indicheremo con G) pari a tre. Una volta fissato G , e il numero di unità statistiche da campionare per ogni strato, si estrae un campione casuale senza rimpiazzo di numerosità n_g da ogni strato g . Quindi, in simboli, da una popolazione di numerosità N , dalla quale si vuole estrarre un campione di n elementi, si creano G strati ognuno di numerosità N_g , da cui viene estratto un campione di numerosità n_g , per cui si ha:

$$\sum_{g=1}^G N_g = N$$

$$\sum_{g=1}^G n_g = n.$$

Proviamo ora a calcolare le probabilità di inclusione del primo ordine. Sappiamo che le n_g unità sono state estratte con un campionamento casuale senza rimpiazzo, per cui la probabilità di inclusione dell' i -esima unità appartenente al g -esimo strato è data da:

$$\pi_i^g = \frac{n_g}{N_g} \quad g = 1, 2, \dots, G$$

cioè π_i^g è simile a quella del campionamento casuale senza rimpiazzo, con l'unica differenza che qui viene calcolata per ogni strato.

Il campionamento stratificato proporzionale

Finora abbiamo trattato il problema della stratificazione senza tener conto della numerosità della popolazione per ogni strato. Nella maggior parte dei casi può accadere che ogni strato non abbia uguale numerosità; nel nostro esempio, non si ha, sicuramente, lo stesso numero di iscritti per ogni anno. Per fronteggiare questo problema si ricorre al campionamento stratificato proporzionale, che non è altro che un campionamento stratificato, con l'unica differenza che, in questo caso, la numerosità campionaria di ogni strato è proporzionale a quella della popolazione. Sapendo, ad esempio, che il numero di iscritti al I anno occupa il 50% della popolazione, quello degli iscritti al II anno il 30%, e che gli studenti iscritti al III anno sono il 20% della popolazione, il campionamento stratificato proporzionale garantisce che tra le n unità del campione ci sarà un 50% di studenti del I anno, un 30% di studenti del II anno e un 20% di studenti del III anno. In questo modo il campione rappresenterà in maniera migliore la popolazione di riferimento.

Formalmente si avrà che:

$$P_g = \frac{N_g}{N} \quad g = 1, \dots, G$$

che rappresenta la proporzione della numerosità dello strato g -esimo rispetto al totale della popolazione. Se si indica con:

$$p_g = \frac{n_g}{n} \quad g = 1, \dots, G$$

la proporzione campionaria dello strato g -esimo rispetto al totale del campione, si avrà, per definizione, che:

$$P_g = p_g \quad g = 1, \dots, G$$

poiché le proporzioni tra popolazione e campione rimangono invariate.

Alla luce di quanto detto finora, si può affermare che tra il campionamento stratificato e quello stratificato proporzionale ciò che cambia è solo la strategia di estrazione del campione, mentre la probabilità di inclusione del primo ordine è la stessa per entrambi i campionamenti, perché la logica alla base del calcolo di questa quantità è, per entrambi, la stessa:

$$\pi_i^g = \frac{n_g}{N_g} \quad g = 1, 2, \dots, G.$$

12.9.2 Il campionamento a grappolo

Il campionamento a grappolo è un'alternativa ai classici metodi di campionamento che considerano la popolazione di unità elementari. Infatti, mentre nei campionamenti descritti finora si considerano le unità statistiche elementari che, nel nostro esempio, sono gli studenti, nel campionamento a grappolo viene presa in considerazione la popolazione di unità complesse come, ad esempio, le famiglie, le scuole, gli ospedali, ecc. Questo significa che si applica il campionamento a grappolo quando si dispone di una lista della popolazione che contiene informazioni circa questo tipo di unità. Si pensi, ad esempio, agli ospedali: un ospedale è un'unità complessa e, se si volesse effettuare uno studio sulla degenza dei malati degli ospedali italiani, è immediato capire che non si dispone di una lista dei malati (che sono le unità elementari che voglio studiare), mentre si dispone di una lista di ospedali allocati sul territorio italiano. Altri esempi possono essere lo studio del rendimento degli alunni di una scuola; in questo caso i grappoli saranno le classi. Ancora, si può pensare all'indagine sulle famiglie che viene condotta trimestralmente dall'*Istat*; in questo caso i grappoli saranno le famiglie.

Quindi, l'applicazione del campionamento a grappolo è richiesta quando non si può giungere direttamente all'unità statistica elementare. Una volta identificate le unità complesse della popolazione e deciso

il numero di grappoli da estrarre, si può procedere con la selezione del campione attraverso un campionamento casuale senza rimpiazzo. Per non confondere le numerosità delle unità elementari (che finora abbiamo indicato con N e n) con quelle delle unità complesse, indichiamo con M la numerosità della popolazione complessa, ovvero il numero di grappoli presenti nella popolazione di interesse, e indichiamo con m la numerosità del campione composto dai grappoli. Poiché l'estrazione dei grappoli viene effettuata con il campionamento casuale senza rimpiazzo, si può ricondurre il ragionamento fatto con questa metodologia al campionamento a grappolo, per cui la probabilità di inclusione del primo ordine del k -esimo grappolo sarà uguale a:

$$\pi_k = \frac{m}{M} \quad k = 1, \dots, M.$$

La probabilità di inclusione del primo ordine per l'unità elementare è la medesima di quella del grappolo, poiché l'unità elementare verrà estratta solo se il grappolo a cui essa appartiene entrerà nel campione. Quindi la probabilità di inclusione del primo ordine va interpretata come rapporto tra la numerosità campionaria dei grappoli e la numerosità della popolazione dei grappoli.

12.9.3 Il campionamento sistematico

L'estrazione di un campione casuale semplice, o di un campione stratificato, di grandi dimensioni risulta oggi agevolata dall'utilizzo di calcolatori che ne consentono una adeguata programmazione. Diversamente, fino a non molti anni fa, tale operazione poteva risultare estremamente laboriosa implicando, per ogni unità da estrarre, il ricorso alla tavola dei numeri casuali.

Un metodo ideato per ridurre il lavoro sulle tavole, e tutt'oggi ancora molto utilizzato nonostante l'informatizzazione della maggior parte delle operazioni di rilevazione e, soprattutto, di selezione di campioni probabilistici, è rappresentato dal cosiddetto *campionamento sistematico*, che richiede l'utilizzo di un meccanismo casuale, come ad esempio la tavola dei numeri casuali, soltanto per la selezione della prima unità. Il campione, infatti, viene formato prendendo dal totale della popolazione una unità ogni k presenti nella lista, con k pari

al reciproco della frazione di campionamento. Quindi, ricordando che la *frazione di campionamento* è data dal rapporto tra la dimensione del campione n e quella della popolazione N ($f = \frac{n}{N}$), l'intervallo che dobbiamo prendere in considerazione per estrarre le unità del campione sarà rappresentato da $\frac{1}{f}$.

Questo schema, proprio per la sua semplicità, anche dal punto di vista della sua implementazione in un calcolatore, è ancora oggi molto utilizzato, soprattutto perché, ordinando la lista secondo un prestabilito criterio, esso consente di formare un campione basato su unità che provengono da ogni parte della lista e non solo da alcune parti, come potrebbe avvenire, per effetto della casualità, in altri schemi di campionamento. Se, ad esempio, si effettua una selezione sistematica da una lista di individui preliminarmente ordinati per età, lo schema garantisce la presenza nel campione di individui di tutte le età, proprio per la sua caratteristica di cogliere le unità gradualmente partendo dalla parte iniziale della lista e scorrendo lungo di essa.

A titolo di esempio concreto, si supponga di dover estrarre un campione di 1000 studenti iscritti ad un Ateneo, da una lista costituita dai 25000 studenti iscritti. Il reciproco della frazione di campionamento, $\frac{N}{n}$, è uguale a 25. Per formare il campione è sufficiente selezionare un numero casuale compreso tra 1 e 25 (estremi inclusi), che individua la prima unità estratta e, quindi, procedere selezionando le altre unità con una progressione aritmetica di ragione 25, fino all'esaurimento della lista. Se, ad esempio, il primo numero estratto fosse 10, il campione risulterebbe formato dalle unità della lista contrassegnate dai numeri d'ordine:

- 10 (studente che occupa le decima posizione nella lista);
- 10 + 25 (studente che occupa le trentacinquesima posizione nella lista);
- 10 + 25 · 2 (studente che occupa le sessantesima posizione nella lista);
- 10 + 25 · 3 (studente che occupa le ottantacinquesima posizione nella lista);
- e così via, fino al termine della lista.

Nell'esempio, volutamente molto semplice, la dimensione campionaria è tale da rendere intero il valore k . Nella pratica k , che prende il nome di *ragione* o *intervallo di selezione*, risulta spesso decimale. In questa situazione è possibile arrotondare k all'intero inferiore o superiore (in relazione alle consuete regole di arrotondamento) a prezzo di un piccolo cambiamento nella dimensione campionaria.

Nel campionamento sistematico, come in quello casuale semplice, ogni unità della popolazione ha la stessa probabilità di entrare a far parte del campione. Diversamente da quanto avviene nel campionamento casuale semplice, tuttavia, in quello sistematico non tutte le n -ple hanno la stessa probabilità di entrare a far parte del campione. Al contrario, fissato l'ordinamento della lista e stabilito di selezionare la prima unità tra le prime k , sono soltanto k le n -ple selezionabili, ciascuna, ovviamente, con probabilità $\frac{1}{k}$.

Il piano di campionamento è tale che:

$$p(c) = \frac{n}{N} = \frac{1}{k}$$

in quanto l'insieme dei possibili campioni C contiene soltanto $\frac{N}{n} = k$ elementi. Inoltre, sia la probabilità di inclusione semplice che quella congiunta sono uguali tra loro, e pari a $\frac{1}{k}$, poiché $\pi = \frac{1}{k}$ se $i \in c$ e $\pi = 0$ se $i \notin c$.

Si ribadisce che il ricorso al campionamento è necessario sia per la riduzione dei costi di un'indagine, sia quando le unità della popolazione sono illimitate e quindi non è possibile osservarle direttamente tutte.