Statistica e Laboratorio

8. Inferenza statistica: stima puntuale e stima intervallare

Paolo Vidoni

Dipartimento di Scienze Economiche e Statistiche
Università di Udine
via Tomadini 30/a - Udine
paolo.vidoni@uniud.it

https://elearning.uniud.it/

Sommario

Sommario

- **Introduzione**
- **Stime e stimatori**
- **Proprietà degli stimatori**
- **Intervalli di confidenza**

Introduzione

Le procedure di *stima puntuale* assegnano, sulla base delle informazioni contenute nel campione osservato e tenendo conto delle assunzioni sul modello generatore dei dati, un valore plausibile al parametro ignoto θ.

Nel caso della *stima intervallare*, si definisce un intervallo di valori plausibili per θ.

Esempio. *Controllo di qualità* (continua). Con riferimento al problema di controllo della qualità considerato in precedenza, si sono individuati 3 oggetti non conformi agli standard in un campione casuale semplice di $n = 40$ oggetti.

Poiché l'obiettivo è stimare la proporzione p di oggetti difettosi prodotti dal macchinario, è intuitivo considerare la proporzione campionaria osservata, che corrisponde a $\hat{p} = 3/40 = 0.075$.

Inoltre, 0.075 è il valore osservato $\bar{x}_n$ della variabile casuale media campionaria $\bar{X}_n = \sum_{i=1}^{40} X_i/n$, nel particolare campione selezionato. $\Diamond$

Sommario

Stime e stimatori

Dato un modello statistico parametrico per i dati campionari $x_1, \ldots, x_n$, riferiti al campione casuale (semplice) $X_1, \ldots, X_n$, con θ parametro ignoto, si definisce **stima per** θ un valore $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$, plausibile per il parametro θ, ottenuto a partire dai dati osservati nel campione.

Si definisce **stimatore per** θ la associata variabile casuale $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$, ottenuta come funzione delle variabili casuali che generano i dati campionati osservati.

Dunque, uno *stimatore* di θ è una opportuna **statistica campionaria** utilizzata per stimare θ, mentre la *stima* di θ è il suo valore osservato in corrispondenza ai dati campionari $x_1, \ldots, x_n$.

Si utilizzerà la notazione sintetica $\hat{\theta}$ sia per lo stimatore che per la stima di θ, poiché il significato appropriato è chiaro dal contesto. Spesso si scriverà $\hat{\theta}_n$ per evidenziare la numerosità n del campione.

Di solito θ è un parametro unidimensionale, e quindi $\hat{\theta}$ è una variabile casuale univariata.

Se si ripete l'esperimento, nelle medesime condizioni, si osserva un campione $x'_1, \ldots, x'_n$, usualmente diverso dal precedente. La stima che si ottiene sarà in genere diversa da quella basata sul primo campione, per effetto della variabilità campionaria.

Tutto ciò in accordo con il fatto che uno stimatore è una variabile casuale, una statistica campionaria, e la sua **distribuzione campionaria sotto** θ è informativa dell'incertezza insita nel procedimento di stima.

Esempio. *Successi.* In un esperimento casuale, si osserva il numero complessivo di successi x, in n prove indipendenti con uguale probabilità di successo $p \in (0, 1)$, ignota.

In questo caso, $X \sim Bi(n, p)$, $\theta = p$ e una stima naturale per p è data dalla frequenza relativa di successo osservata $\hat{p} = x/n$.

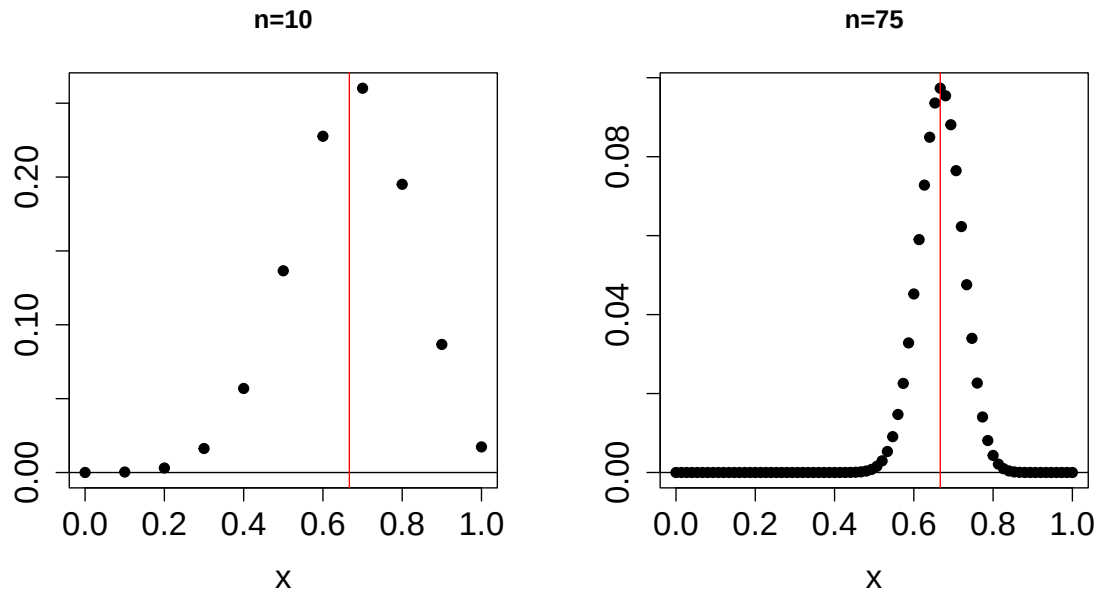
Lo stimatore $\hat{p} = X/n$ è una variabile casuale discreta tale che, qualunque sia $p \in (0, 1)$, $E(\hat{p}) = p$, $V(\hat{p}) = p(1-p)/n$ e, per $n \to +\infty$, $\hat{p} \xrightarrow{p} p$.

Oltre a questi aspetti parziali sulla distribuzione di probabilità dello stimatore, è nota la sua distribuzione campionaria sotto p.

In particolare, per ogni $x \in \{0, \ldots, n\}$,

$$P(\hat{p} = x/n) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Inoltre, per n sufficientemente elevato, vale l'approssimazione normale: $\hat{p} \dot{\sim} N(p, p(1-p)/n)$.



Si sono disegnati i grafici della funzione di probabilità di $\hat{p}$, per $n = 10$ e $n = 75$, sotto l'ipotesi che $p = 2/3$ (linea rossa). $\Diamond$

Esempio. *Bottiglie* (continua). Si considera il problema del controllo del contenuto effettivo delle bottiglie con la soluzione di glucosio da 500 ml.

Si estrae un campione casuale semplice di $n = 25$ bottiglie e si ipotizza che il contenuto di una generica bottiglia sia descritto da una variabile casuale con distribuzione $N(\mu, \sigma^2)$, con μ e σ^2 ignoti.

Per stimare il contenuto medio effettivo μ e la variabilità del processo produttivo σ^2 sulla base dei dati campionari si considerano, rispettivamente,

$$\bar{x}_{25} = \frac{1}{25}\sum_{i=1}^{25} x_i, \quad s_c^2 = \frac{1}{24}\sum_{i=1}^{25}(x_i - \bar{x}_{25})^2.$$

Dopo l'estrazione del campione, si ottengono le stime $\bar{x}_{25} = 498$ e $s_c^2 = 337.5$, che corrispondono alle determinazioni osservate delle variabili casuali media campionaria $\bar{X}_{25}$ e varianza campionaria corretta S_c^2, interpretabili come stimatori per μ e σ^2.

Se al posto di S_c^2 si considera la varianza campionaria S^2 come stimatore per σ^2, si ottiene il valore di stima $s^2 = 324$.

Le distribuzioni di probabilità delle statistiche campionarie, già analizzate in precedenza, sono informative sulla bontà della procedura inferenziale.

$\Diamond$

Per quanto riguarda la definizione di procedure inferenziali per la stima puntuale, si possono individuare due punti di fondamentale importanza, che verranno di seguito analizzati:

- la definizione di opportuni **metodi per definire stimatori**, con riferimento ad un certo parametro di interesse;
- l'analisi delle **proprietà di uno stimatore**, e la conseguente valutazione della sua efficacia nel definire valori di stima plausibili per il parametro ignoto.

Metodi di stima

Negli esempi considerati in precedenza è adombrato un metodo semplice per ottenere stimatori: il **metodo dell'analogia**.

In base a tale metodo, per stimare un certo *parametro della popolazione* si utilizza la corrispondente *quantità campionaria* (statistica campionaria).

Ad esempio, un valor medio si stima con una media campionaria, una varianza con una varianza campionaria (corretta), una mediana con una mediana campionaria, ecc.

Se le quantità a cui si applica il metodo dell'analogia sono momenti (ad esempio il valor medio, la varianza, ecc.), si parla di **metodo dei momenti**.

Si possono ricordare altri metodi più formali, come il **metodo della massima verosimiglianza** e il **metodo dei minimi quadrati**.

Accanto al metodo dell'analogia è utile considerare il **metodo di sostituzione** (**plug-in**), così specificato.

Se si è interessati alla stima di $\tau = g(\theta)$, funzione del parametro θ, e per θ è disponibile uno stimatore $\hat{\theta}$, allora uno stimatore per τ si ottiene sostituendo θ con $\hat{\theta}$ nella funzione $g(\cdot)$.

Quindi, $\hat{\tau} = g(\hat{\theta})$ è lo stimatore plug-in per τ.

Esempio. *Campione esponenziale*. Si consideri un campione casuale semplice costituito da variabili casuale $X_1, \ldots, X_n$ con distribuzione $Esp(\lambda)$, con λ ignoto.

Dal momento che $\mu = 1/\lambda$ e quindi $\lambda = 1/\mu$, si può pensare di stimare λ con $\hat{\lambda} = 1/\bar{X}_n$. Si noti che, in generale,

$$E(\hat{\lambda}) = E\left(\frac{1}{\bar{X}_n}\right) \neq \frac{1}{E(\bar{X}_n)} = \frac{1}{\mu} = \lambda,$$

mentre, per $n \to +\infty$, visto che $\bar{X}_n \xrightarrow{p} \mu$, si ha che

$$\hat{\lambda} = \frac{1}{\bar{X}_n} \xrightarrow{p} \frac{1}{\mu} = \lambda.$$

$\Diamond$

Sommario

Premessa

Dal momento che uno stimatore è una statistica campionaria, cioè una variabile casuale, la sua distribuzione di probabilità sotto θ sarà informativa sulla bontà del procedimento di stima.

In particolare, si vuole indagare se la distribuzione di probabilità campionaria dello stimatore è, in un certo senso, vicina al parametro che si vuole stimare.

In alcuni casi, come per la media campionaria e altre statistiche campionarie di uso comune, la distribuzione di probabilità è nota in modo esatto o approssimato.

L'obiettivo della stima non è l'*esattezza*, ma l'*accuratezza*, ossia che l'errore di stima sia usualmente piccolo, al variare del campione osservato.

Standard error

Come misura della precisione di uno stimatore $\hat{\theta}$ si può considerare, se θ è unidimensionale, l'**errore quadratico medio di stima** sotto θ,

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

La sua radice quadrata positiva $se(\hat{\theta}) = \sqrt{EQM(\hat{\theta})}$ è chiamata **errore standard** (**standard error**). Tali quantità dipendono dal parametro θ ignoto e dalla dimensione n del campione.

Non è sensato ricercare lo stimatore tale che $se(\hat{\theta}) = 0$ per ogni $\theta \in \Theta$, che esiste solo in casi banali e non interessanti.

Dati due stimatori $\hat{\theta}_1$ e $\hat{\theta}_2$ per θ, si preferisce $\hat{\theta}_1$ se ha errore standard uniformemente più piccolo, cioè se $se(\hat{\theta}_1) \leq se(\hat{\theta}_2)$ per ogni $\theta \in \Theta$.

È raro che esista uno stimatore con errore standard uniformemente minimo. È possibile individuarlo in alcune classi particolari di stimatori.

Poiché $se_\theta(\hat{\theta})$ dipende in genere da θ, che è ignoto, viene usualmente stimato sostituendo θ con la sua stima $\hat{\theta}$. Si ottiene l'**errore standard stimato (estimated standard error)**

$$\hat{se}(\hat{\theta}) = \sqrt{E(\hat{\theta} - \theta)^2} \mid_{\theta=\hat{\theta}},$$

che corrisponde alla valutazione di $se(\hat{\theta})$ con i dati del campione. Nelle procedure di stima puntuale si riporta usualmente $\hat{\theta}$ e $\hat{se}(\hat{\theta})$.

Si verifica facilmente che

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

dove la differenza $[E(\hat{\theta}) - \theta]$ corrisponde alla **distorsione** dello stimatore.

Quindi l'errore quadratico medio di stima, ed anche l'errore standard, tengono conto di due aspetti importanti per valutare la bontà di uno stimatore: la sua *variabilità* e la sua *distorsione*.

Proprietà degli stimatori

Un stimatore $\hat{\theta}$ è **non distorto** se

$$E(\hat{\theta}) = \theta, \quad \text{per ogni } \theta \in \Theta,$$

cioè se il suo valor medio coincide con il vero valore del parametro. In questo caso, $se(\hat{\theta})$ corrisponde allo *scarto quadratico medio* di $\hat{\theta}$.

In alcuni contesti si riesce a individuare uno stimatore **efficiente fra i non distorti**, cioè *non distorto* e che presenta *errore standard* (e quindi varianza) *uniformemente minimo* fra tutti gli stimatori non distorti per θ.

Se uno stimatore con forte distorsione è di fatto inutile, perché presenta nella generalità dei campioni un forte errore sistematico, può bastare, in pratica, la seguente richiesta più tenue.

Uno stimatore $\hat{\theta}$ è detto **asintoticamente non distorto** se al crescere della dimensione n del campione tende alla non distorsione, cioè se

$$\lim_{n \to \infty} E(\hat{\theta}_n) = \theta, \quad \text{per ogni } \theta \in \Theta.$$

Mentre la distorsione è una proprietà desiderabile per uno stimatore, la proprietà specificata di seguito costituisce un requisito necessario, poiché è ragionevole supporre che, all'aumentare dell'informazione campionaria, la conoscenza su θ diventi sempre più precisa.

Uno stimatore è detto **consistente** per θ se al crescere della dimensione n del campione produce realizzazioni vicine al vero valore del parametro con elevata probabilità, cioè se

$$\hat{\theta}_n \xrightarrow{p} \theta, \quad \text{per } n \to \infty.$$

Per quanto detto sulla convergenza in probabilità, la consistenza sussiste se lo stimatore è (almeno) *asintoticamente non distorto* e tale che $\lim_{n\to\infty} V(\hat{\theta}_n) = 0$.

In questo caso, anche l'*errore quadratico medio* $EQM(\hat{\theta})$ tende a zero al crescere della dimensione campionaria.

Il *metodo della massima verosimiglianza* fornisce stimatori con diverse proprietà desiderabili. Infatti sono usualmente consistenti, asintoticamente non distorti ed efficienti e con distribuzione approssimativamente normale. Inoltre, è semplice ottenere la loro varianza.

Esempio. *Campione gaussiano*. Sia $X_1, \ldots, X_n$, $n > 2$, un campione casuale semplice da una popolazione $N(\mu, \sigma^2)$. Si confrontano i seguenti stimatori per μ

$$\hat{\mu}_1 = \bar{X}_n, \quad \hat{\mu}_2 = \frac{X_1 + X_2}{2}.$$

Sono entrambi non distorti, ma $\hat{\mu}_1$ è più efficiente di $\hat{\mu}_2$ essendo

$$V(\hat{\mu}_1) = \frac{\sigma^2}{n} \leq \frac{\sigma^2}{2} = V(\hat{\mu}_2).$$

Inoltre, $\bar{X}_n$ è consistente ed ha errore standard $se(\bar{X}_n) = \sigma/\sqrt{n}$, mentre l'errore standard stimato è $\hat{se}(\bar{X}_n) = \hat{\sigma}/\sqrt{n}$, con $\hat{\sigma}$ un'opportuna stima per σ.

In presenza di valori anomali, la media campionaria può presentare problemi di robustezza. Si preferisce allora utilizzare la **media troncata**, ottenuta eliminando una opportuna percentuale (piccola) di dati estremi.

Come stimatori per σ^2 si possono considerare la varianza campionaria S^2 e la varianza campionaria corretta S_c^2.

Ricordando quanto detto in precedenza, S^2 è distorto, ma asintoticamente non distorto, mentre S_c^2 è non distorto. Entrambi soddisfano la proprietà della consistenza. $\Diamond$

Esempio. *Campione uniforme*. Sia $X_1, \ldots, X_n$, un campione casuale semplice da una popolazione $U(0, \theta)$, con $\theta > 0$ ignoto.

Poiché $X_i \sim U(0, \theta)$, $i = 1, \ldots, n$, si ha che $E(X_i) = \mu = \theta/2$ e $V(X_i) = \sigma^2 = \theta^2/12$.

Applicando il metodo di sostituzione, si ottiene lo stimatore $\hat{\theta} = 2\bar{X}_n$.

Uno stimatore alternativo per θ corrisponde a $\tilde{\theta} = \max\{X_1, \ldots, X_n\}$.

Con riferimento allo stimatore $\hat{\theta}$, si dimostra che è non distorto e consistente, infatti

$$E(\hat{\theta}) = 2E(\bar{X}_n) = 2\,\mu = 2\,\frac{\theta}{2} = \theta,$$

$$V(\hat{\theta}) = 4V(\bar{X}_n) = 4\,\frac{\sigma^2}{n} = 4\,\frac{\theta^2}{12\,n} = \frac{\theta^2}{3n}.$$

Inoltre, per la non distorsione, l'errore standard corrisponde a $se(\hat{\theta}) = \sqrt{V(\hat{\theta})} = \theta/\sqrt{3n}$. ◊

Esempio. *Controllo di qualità* (continua). Con riferimento al problema di controllo della qualità presentato in precedenza, si è individuato come stimatore per p la media campionaria $\hat{p} = \bar{X}_n$.

Tale stimatore è non distorto, consistente, asintoticamente normale, con errore standard $se(\hat{p}) = \sqrt{p(1-p)/n}$. Poiché su $n = 40$ oggetti si rilevano 3 difettosi, si ha che $\hat{p} = 3/40 = 0.075$ e $\hat{se}(\hat{p}) = \sqrt{0.075(1-0.075)/40} = 0.042$. ◊

Esempio. *Web*. Si considerano i dati $(0, 0, 3, 0, 1, 0, 0, 2, 1, 0, 0, 2)$, relativi ad osservazioni ripetute del numero di visite ad un sito web in un'ora.

I dati campionari sono analizzati come realizzazione di un campione casuale semplice $X_1, \ldots, X_{12}$, costituito da variabili casuali con distribuzione $P(\lambda)$, con $\lambda > 0$ ignoto.

Il metodo dei momenti individua come stimatore per λ la media campionaria $\hat{\lambda} = \bar{X}_n$.

Tale stimatore è non distorto, consistente, asintoticamente normale, con errore standard $se(\hat{\lambda}) = \sqrt{\lambda/n}$.

Poiché $n = 12$ e $\sum_{i=1}^{12} x_i = 9$, si ha che $\hat{\lambda} = 9/12 = 0.75$ e $\hat{se}(\hat{\lambda}) = \sqrt{0.75/12} = 0.25$. $\Diamond$

Sommario

Stima intervallare

Con le procedure di stima puntuale si ottiene un valore di stima che quasi certamente non coincide con il vero e ignoto valore del parametro θ.

Con la *stima intervallare* (*intervalli di confidenza*) si cerca di incorporare nel procedimento di stima una misura di accuratezza.

Il parametro ignoto non viene stimato con un punto, ma con un sottoinsieme più ampio dello spazio parametrico Θ, in genere un intervallo. Il parametro θ si considera unidimensionale.

Esempio. *Campione gaussiano* (continua). Sia $X_1, \ldots, X_5$ un campione casuale semplice da una popolazione $N(\mu, 16)$, con μ ignoto. Poiché si vuole fare inferenza su μ, si considera, come statistica campionaria, la media campionaria $\bar{X}_5 \sim N(\mu, 16/5)$.

La media campionaria standardizzata è una *quantità pivotale*, dal momento che ha distribuzione $N(0, 1)$, che non dipende dal parametro ignoto μ.

Ricordando quanto detto sui quantili di $N(0,1)$, si ha che, per ogni $\mu \in \mathbf{R}$,

$$P\left(-1.96 \leq \frac{\bar{X}_5 - \mu}{4/\sqrt{5}} \leq 1.96\right) = 0.95,$$

ovvero, esplicitando il parametro μ,

$$P\left(\bar{X}_5 - 1.96\frac{4}{\sqrt{5}} \leq \mu \leq \bar{X}_5 + 1.96\frac{4}{\sqrt{5}}\right) = 0.95.$$

Quindi risultano individuate due statistiche campionarie che definiscono l'intervallo casuale (aleatorio)

$$\left[\bar{X}_5 - 1.96\frac{4}{\sqrt{5}}, \bar{X}_5 + 1.96\frac{4}{\sqrt{5}}\right],$$

che contiene il vero valore del parametro μ, qualunque esso sia, con probabilità 0.95. Tale intervallo è detto *intervallo di confidenza* per μ, con *livello di confidenza* 0.95.

Se si osserva il campione $(169, 171, 174, 177, 179)$, si ottiene che $\bar{x}_5 = 174$ e l'intervallo di confidenza osservato è $[170.49, 177.51]$.

È sbagliato affermare che $[170.49, 177.51]$ contiene μ con probabilità 0.95. Infatti è l'intervallo casuale che contiene μ con probabilità 0.95 e quindi si può solo avere fiducia di aver osservato un campione che fornisce un intervallo numerico che contiene μ. $\Diamond$

Dato il campione casuale $X_1, \ldots, X_n$, un **intervallo di confidenza (stima intervallare)** per θ, con **livello di confidenza** $1 - \alpha \in (0, 1)$, è un **intervallo (casuale)** $[L, U]$, con $L = L(X_1, \ldots, X_n)$ e $U = U(X_1, \ldots, X_n)$ statistiche campionarie, tale che, per ogni $\theta \in \Theta$,

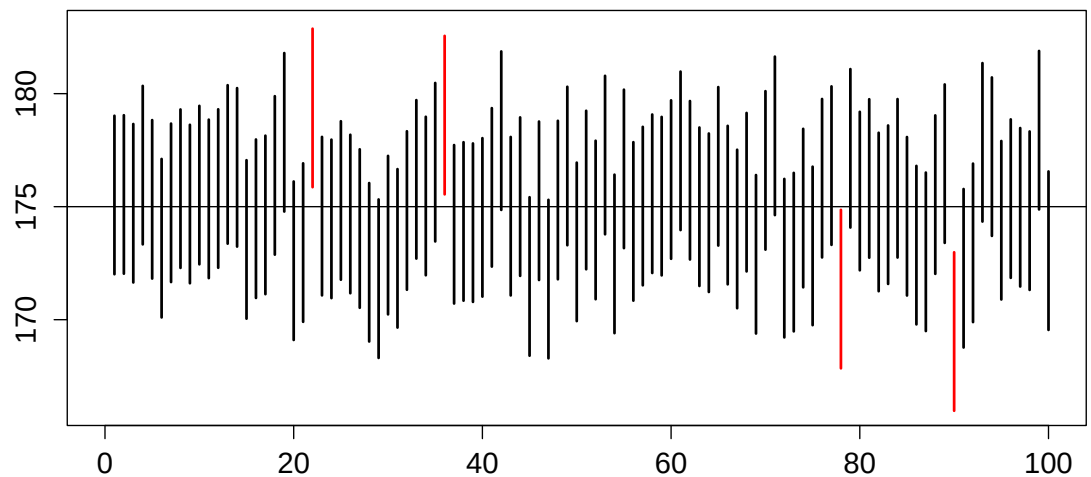
$$P(L \leq \theta \leq U) = 1 - \alpha\,.$$

Le statistiche L e U si chiamano **limite inferiore** e **limite superiore** di confidenza, mentre $U - L$ specifica la **lunghezza** dell'intervallo; $(U - L)/2$ si chiama **accuratezza** dell'intervallo di confidenza.

Considerando le osservazioni $x_1, \ldots, x_n$, si determina l'**intervallo di confidenza (osservato)** $[l, u]$ per θ, con l e u valori osservati per L e U.

L'interpretazione di $[l, u]$ è analoga a quella fornita nell'esempio precedente: se $1 - \alpha = 0.95$ e potessimo ripetere l'esperimento casuale un numero elevato di volte, gli intervalli ottenuti conterrebbero il vero valore di θ nel 95% dei casi.

Esempio. *Campione gaussiano* (continua). Si sono simulati 100 campioni di dimensione $n = 5$ da un modello $N(175, 16)$. Vengono riportati nel grafico sottostante i corrispondenti 100 intervalli di confidenza per μ di livello $1 - \alpha = 0.95$. Gli intervalli che non contengono il vero valore del parametro $\mu = 175$ sono indicati in rosso.



Nel seguito, non si svilupperà la teoria degli intervalli (regioni) di confidenza. Si forniranno delle indicazioni pratiche, utili per la costruzione di intervalli di confidenza per particolari problemi.

La procedura che verrà usualmente adottata per definire intervalli di confidenza per θ richiede

- la definizione di una opportuna **statistica campionaria** per fare inferenza su θ (in genere uno stimatore);
- la costruzione, a partire da tale statistica, di una **quantità pivotale**, la cui distribuzione di probabilità non dipende dal parametro ignoto θ.

In alcuni situazioni, si definiranno quantità pivotali *approssimate*, e quindi intervalli di confidenza con livello di confidenza *approssimato* $1 - \alpha$.

In genere, si considerano valori $1 - \alpha$ pari a $0.9, 0.95, 0.99$.

Intervallo di confidenza per la media di una popolazione normale

Sia $X_1, \ldots, X_n$ un campione casuale semplice da una popolazione $N(\mu, \sigma^2)$, con **varianza σ^2 nota**; si vuole determinare un intervallo di confidenza per μ con livello $1 - \alpha$.

Si considera la *media campionaria* $\bar{X}_n$ e, come *quantità pivotale*, la **media campionaria standardizzata** $\sqrt{n}(\bar{X}_n - \mu)/\sigma \sim N(0, 1)$.

Si cercano due costanti reali *a* e *b* tali che, per ogni $\mu \in \mathbf{R}$,

$$P\left(a < \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} < b\right) = 1 - \alpha,$$

ovvero, tali che

$$P\left(\bar{X}_n - b\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n - a\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Per le constanti a e b si scelgono usualmente i valori simmetrici $a = -z_{\alpha/2}$ e $b = z_{\alpha/2}$, con $z_{\alpha/2}$ *valore critico* tale che $P(Z \geq z_{\alpha/2}) = \alpha/2$, dove $Z \sim N(0, 1)$.

Quindi, un **intervallo di confidenza per μ con livello** $1 - \alpha$ è

$$\left[\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Se la varianza σ^2 è **ignota**, si utilizza al suo posto la varianza campionaria corretta S_c^2 e, come *quantità pivotale*, la **media campionaria studentizzata** $\sqrt{n}(\bar{X}_n - \mu)/S_c \sim t(n-1)$.

Quindi, un **intervallo di confidenza per μ con livello** $1 - \alpha$ è

$$\left[\bar{X}_n - t_{\alpha/2}\frac{S_c}{\sqrt{n}}, \bar{X}_n + t_{\alpha/2}\frac{S_c}{\sqrt{n}}\right].$$

Al posto di σ si considera S_c e al posto di $z_{\alpha/2}$ il *valore critico* $t_{\alpha/2}$ tale che $P(T \geq t_{\alpha/2}) = \alpha/2$, dove $T \sim t(n-1)$.

La densità t di Student ha code più pesanti di quelle della $N(0,1)$; l'*intervallo di confidenza* tende ad essere *più ampio* (meno preciso) del precedente (c'è maggiore incertezza poiché si stima σ con S_c).

Se la dimensione n del campione è sufficientemente elevata, si può approssimare $t_{\alpha/2}$ con $z_{\alpha/2}$.

Nel caso in cui σ^2 è nota, l'*intervallo si può riscrivere* come

$$\left[\hat{\mu} - z_{1-\alpha/2}\,\mathsf{se}(\hat{\mu}), \hat{\mu} + z_{1-\alpha/2}\,\mathsf{se}(\hat{\mu})\right] .$$

Se σ^2 è ignota, si considera l'errore standard stimato $\hat{se}(\hat{\mu})$ e si sostituisce $z_{\alpha/2}$ con $t_{\alpha/2}$.

In entrambi i casi l'**accuratezza** dell'intervallo di confidenza, che corrisponde a $z_{\alpha/2}\sigma/\sqrt{n}$, se σ^2 è nota, o a $t_{\alpha/2}S_c/\sqrt{n}$, se σ^2 è ignota, è tanto migliore quanto più n è elevata e σ (o S_c) è piccola.

Un aumento del livello di confidenza $1-\alpha$ si ottiene solo aumentando l'ampiezza dell'intervallo di confidenza.

Esempio. *Campione gaussiano con varianza nota*. Sia $X_1, \ldots, X_{50}$ un campione casuale semplice di dimensione $n = 50$ da una popolazione normale con μ ignoto e $\sigma^2 = 2$.

Nell'ipotesi che i risultati dell'indagine campionaria siano tali che $\sum_{i=1}^{50} x_i = 94.15$, si vuole determinare un intervallo di confidenza per μ con livello $1 - \alpha = 0.9$.

Poiché $\bar{x}_{50} = \sum_{i=1}^{50} x_i/50 = 1.883$, $\alpha/2 = 0.05$ e $z_{0.05} = 1.645$, si conclude che

$$\left[1.883 - 1.645\,\frac{\sqrt{2}}{\sqrt{50}}, 1.883 + 1.645\,\frac{\sqrt{2}}{\sqrt{50}}\right] = [1.554, 2.212]$$

è l'intervallo di confidenza cercato. ◊

Esempio. *Campione gaussiano con varianza ignota*. Sia $X_1, \ldots, X_{30}$ un campione casuale semplice di dimensione $n = 30$ da una popolazione normale con μ e σ^2 ignoti.

Nell'ipotesi che i risultati dell'indagine campionaria siano tali che $\sum_{i=1}^{30} x_i = 39.50$ e $\sum_{i=1}^{30} x_i^2 = 101.54$, si cerca un intervallo di confidenza per μ con livello $1 - \alpha = 0.95$.

In questo caso, $\bar{x}_{30} = \sum_{i=1}^{30} x_i/30 = 1.317$,

$$s^2 = \frac{1}{30}\sum_{i=1}^{30} x_i^2 - \bar{x}_{30}^2 = 1.651, \quad s_c^2 = \frac{30}{29}s^2 = 1.708.$$

Inoltre, $\alpha/2 = 0.025$ e, considerando un distribuzione t di Student con 29 gradi di libertà, $t_{0.025} = 2.045$. Quindi, si conclude che

$$\left[1.317 - 2.045\,\frac{\sqrt{1.708}}{\sqrt{30}}, 1.317 + 2.045\,\frac{\sqrt{1.708}}{\sqrt{30}}\right]$$

$$= [0.829, 1.805]$$

è l'intervallo di confidenza cercato. $\Diamond$

Intervallo di confidenza per la media di una popolazione qualsiasi

Sia $X_1, \ldots, X_n$ un campione casuale semplice da una popolazione non normale con media μ ignota. Si vuole determinare un intervallo di confidenza per μ con livello $1 - \alpha$

Per il teorema limite centrale, se la numerosità campionaria n è sufficientemente elevata, si determina, con relativa facilità, un **intervallo di confidenza per μ con livello $1 - \alpha$ approssimato**.

Si considera la *media campionaria* $\bar{X}_n$ e, come *quantità pivotale approssimata*, la **media campionaria standardizzata** $\sqrt{n}(\bar{X}_n - \mu)/\sigma \dot{\sim} N(0, 1)$.

Dal momento che σ non è noto, lo si può sostituire con un opportuno stimatore consistente $\hat{\sigma}$ e si ottiene l'intervallo di confidenza

$$\left[\bar{X}_n - z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right].$$

Si considerano i seguenti casi interessanti:

- se $X_i \sim Ber(p)$, $i = 1, \ldots, n$, allora l'intervallo di confidenza per $\mu = p$ si specifica con $\hat{p} = \bar{X}_n$ e $\hat{\sigma} = \sqrt{\hat{p}(1-\hat{p})}$;
- se $X_i \sim P(\lambda)$, $i = 1, \ldots, n$, allora l'intervallo di confidenza per $\mu = \lambda$ si specifica con $\hat{\lambda} = \bar{X}_n$ e $\hat{\sigma} = \sqrt{\hat{\lambda}}$.

Per distribuzioni non normali esistono anche metodi più accurati, generalmente implementati nei software statistici, che risultano utili per il calcolo di intervalli di confidenza per la media nel caso di piccoli campioni.

Esempio. *Farmaco*. Si vuole studiare l'efficacia di un farmaco per curare una determinata patologia. Si effettua una sperimentazione su 550 pazienti e si riscontra che in farmaco è efficace in 393 casi.

Si vuole determinare un intervallo di confidenza, con livello $1 - \alpha = 0.95$ (approssimato), per la frequenza relativa p dei casi in cui il farmaco è efficace, con riferimento all'intera popolazione.

Si può ragionevolmente pensare che il campione osservato, di dimensione $n = 550$, provenga da una popolazione $Ber(p)$, con p ignoto.

Quindi, poiché $\hat{p} = 393/550 = 0.715$, $\alpha/2 = 0.025$ e $z_{0.025} = 1.96$,

$$\left[0.715 - 1.96\frac{\sqrt{0.715 \cdot 0.285}}{\sqrt{550}}, 0.715 + 1.96\frac{\sqrt{0.715 \cdot 0.285}}{\sqrt{550}}\right]$$

$$= [0.677, 0.753]$$

è un intervallo di confidenza per p con livello di confidenza approssimato 0.95. $\Diamond$

Esempio. *Sito web*. Si vuole studiare il numero di accessi all'ora ad un sito web per il commercio elettronico. Si hanno i dati sugli accessi nelle ultime $n = 96$ ore, che risultano essere complessivamente 383.

Si vuole determinare un intervallo di confidenza, con livello $1 - \alpha = 0.90$ (approssimato), per il numero medio di accessi all'ora.

Si ipotizza che il campione osservato, di dimensione $n = 96$, provenga da una popolazione $P(\lambda)$, con media λ ignota.

Quindi, poiché $\hat{\lambda} = 383/96 = 3.989$, $\alpha/2 = 0.05$ e $z_{0.05} = 1.645$,

$$\left[3.989 - 1.645\sqrt{\frac{3.989}{96}}, 3.989 + 1.645\sqrt{\frac{3.989}{96}}\right]$$

$$= [3.65, 4.32]$$

è un intervallo di confidenza per λ con livello di confidenza approssimato 0.90. $\Diamond$

Intervallo di confidenza per la varianza di una popolazione normale

Sia $X_1, \ldots, X_n$ un campione casuale semplice da una popolazione $N(\mu, \sigma^2)$. Si vuole determinare un intervallo di confidenza per σ^2 con livello $1 - \alpha$. La media μ viene considerata ignota.

Si considera la *varianza campionaria* S^2 e, come *quantità pivotale*, la seguente **trasformata della varianza campionaria** $nS^2/\sigma^2 \sim \chi^2(n-1)$.

Per ogni $\sigma^2 \in \mathbf{R}^+$, si ha che

$$P\left(\chi^2_{1-\alpha/2} \leq \frac{nS^2}{\sigma^2} \leq \chi^2_{\alpha/2}\right) = P\left(\frac{nS^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{nS^2}{\chi^2_{1-\alpha/2}}\right) = 1 - \alpha,$$

dove $\chi^2_{1-\alpha/2}$ e $\chi^2_{\alpha/2}$ sono i *valori critici* tali che, rispettivamente, $P(Y \geq \chi^2_{1-\alpha/2}) = 1 - \alpha/2$ e $P(Y \geq \chi^2_{\alpha/2}) = \alpha/2$, essendo $Y \sim \chi^2(n-1)$.

Quindi, un **intervallo di confidenza per σ^2 con livello** $1-\alpha$ è

$$\left[\frac{nS^2}{\chi^2_{\alpha/2}}, \frac{nS^2}{\chi^2_{1-\alpha/2}}\right].$$

Il medesimo intervallo si può ottenere considerando la *varianza campionaria corretta* S_c^2 e, come *quantità pivotale*, la seguente **trasformata della varianza campionaria corretta** $(n-1)S_c^2/\sigma^2 \sim \chi^2(n-1)$. Più precisamente, si ottiene l'intervallo

$$\left[\frac{(n-1)S_c^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S_c^2}{\chi^2_{1-\alpha/2}}\right],$$

che coincide con il precedente.

I molti contesti applicativi, la varianza σ^2 è una quantità importante da studiare. Ad esempio, nel controllo di qualità, la *conformità del processo produttivo* è direttamente legata alla sua *variabilità*.

Esempio. *Campione gaussiano con varianza ignota* (continua). Sia $X_1, \ldots, X_{30}$ un campione casuale semplice di dimensione $n = 30$ da una popolazione normale con μ e σ^2 ignoti.

L'indagine campionaria fornisce i seguenti risultati $\sum_{i=1}^{30} x_i = 39.50$ e $\sum_{i=1}^{30} x_i^2 = 101.54$. Si vuole determinare un intervallo di confidenza per σ^2 con livello $1 - \alpha = 0.95$.

In questo caso, $\bar{x}_{30} = \sum_{i=1}^{30} x_i/30 = 1.317$ e

$$s^2 = \frac{1}{30} \sum_{i=1}^{30} x_i^2 - \bar{x}_{30}^2 = 1.651.$$

Inoltre, $\alpha/2 = 0.025$ e, considerando un distribuzione $\chi^2(29)$, si ha che $\chi^2_{1-0.025} = \chi^2_{0.975} = 16.047$ e $\chi^2_{0.025} = 45.722$. Quindi, si conclude che

$$\left[\frac{30 \cdot 1.651}{45.722}, \frac{30 \cdot 1.651}{16.047}\right] = [1.083, 3.087]$$

è l'intervallo di confidenza cercato. ◊