



www.sce.unimore.it

Scienze della Comunicazione
e dell'Economia

STIMA PUNTUALE E PER INTERVALLO

Legacy Edition
Copyright 25 ottobre 2012

Luca La Rocca
luca.larocca@unimore.it

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA



Introduzione

Stima puntuale ed errore standard

Stima per intervallo

Introduzione

Stima puntuale ed errore standard

Stima per intervallo

Supponiamo di essere interessati alla media μ di un carattere quantitativo X in una popolazione la cui numerosità N è “troppo grande” per effettuare un’indagine totale: sulla base di un **campione casuale semplice**, di numerosità $n \ll N$, cosa possiamo dire su μ ?

Prima di raccogliere i dati, i valori di X nel campione sono numeri aleatori: X_1, \dots, X_n indipendenti e identicamente distribuiti con

$$\begin{aligned}\mathbb{E}[X_1] &= \mu \\ \text{sd}(X_1) &= \sigma\end{aligned}$$

dove σ è la deviazione standard di X nella popolazione di interesse. . .



... dopo avere raccolto i dati i valori di X nel campione sono quantità note: le modalità osservate x_1, \dots, x_n (distribuzione unitaria).

Un caso particolare, spesso di interesse, è quello in cui il carattere X è **dicotomico**; in questo caso

- ▶ denotiamo la media di popolazione con ψ , per ricordarci che in realtà si tratta di una proporzione di popolazione;
- ▶ la deviazione standard di popolazione è funzione della media di popolazione:

$$\sigma = \sqrt{\psi(1 - \psi)}$$

Per fissare le idee, consideriamo i seguenti esempi:

- ▶ misure ripetute della lunghezza di un pezzo meccanico (in questo caso σ può essere una caratteristica nota dello strumento di misura);
- ▶ indagine sul reddito dei membri di una certa comunità;
- ▶ indagine sull'opinione dei cittadini (favorevole/contrario) nei confronti di una data proposta di legge.

Introduzione

Stima puntuale ed errore standard

Stima per intervallo

Sembra naturale, “per analogia”, **stimare puntualmente** la media di popolazione μ con la media campionaria

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

Per esempio, se $n = 4$ misure della lunghezza (in micron) di un pezzo meccanico danno come risultato

$$\begin{aligned} x_1 &= 50000.92 & x_2 &= 49998.70 \\ x_3 &= 49998.89 & x_4 &= 50000.47, \end{aligned}$$

la media campionaria $\bar{x} = 49999.74$ sarà la nostra **stima puntuale** (per analogia) della vera lunghezza μ .

A parte la giustificazione per analogia, è \bar{x} una buona stima di μ ?

Nell'approccio frequentista si risponde a questa domanda studiando la distribuzione campionaria dello **stimatore puntuale**

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}$$

di cui \bar{x} è realizzazione per il campione selezionato: tale distribuzione (la distribuzione di \bar{X} come numero aleatorio prima di raccogliere i dati) riflette la variabilità della stima da campione a campione.

Circa la bontà della media campionaria \bar{X} come stimatore puntuale della media di popolazione μ si può dire che:

- ▶ \bar{X} è **corretto**, cioè $\mathbb{E}[\bar{X}] = \mu$, quale che sia $\mu \in \mathbb{R}$ (si dice anche che \bar{X} è **non distorto**);
- ▶ \bar{X} è **consistente**, cioè $\bar{X} \xrightarrow{\mathbb{P}} \mu$, per $n \rightarrow \infty$, quale che sia $\mu \in \mathbb{R}$ (in virtù della legge dei grandi numeri);

dove \mathbb{P} ed \mathbb{E} sono relativi alla distribuzione campionaria di \bar{X} .

Dunque ci aspettiamo che, tipicamente, il valore stimato \bar{x} sia “prossimo” al valore incognito μ . . . tanto più “prossimo” quanto più elevata è la numerosità campionaria. . . ma quanto “prossimo” per un dato valore di n ?

Possiamo misurare la variabilità di \bar{X} mediante il suo **errore standard**

$$se = sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

che ci darà un valore tipico del valore assoluto dell'**errore di stima** $\bar{x} - \mu$ (il quale non avrà un segno privilegiato in quanto \bar{X} è corretto).

Vediamo che se decresce al crescere di n (a conferma della consistenza) e in particolare decresce come “uno su radice di piccola enne”; inoltre se σ è nota (es. $\sigma = 2$ micron) possiamo valutare numericamente

$$se = \frac{2}{\sqrt{4}} = 1$$

nell'esempio della misura della lunghezza di un pezzo meccanico.

La valutazione numerica di se fornisce una **stima con errore standard** della vera lunghezza μ :

$$\mu = \bar{x} \pm se = 50000 \pm 1$$

laddove con una sola osservazione ($x_1 = 50000.92$) si troverebbe

$$\mu = x_1 \pm \sigma = 50001 \pm 2;$$

quadruplicando il numero di misure si **dimezza** l'errore standard.

E se invece σ non è nota?

Possiamo stimare σ mediante il suo analogo campionario:

$$sd_x = \sqrt{\frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}}$$

Sarà sd_x una buona stima di σ ? Possiamo dire che lo stimatore

$$sd_X = \sqrt{\frac{(X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n}}$$

è consistente, ma distorto: $\mathbb{E}[sd_X] \neq \sigma$.

In particolare $\mathbb{E}[sd_X] < \sigma$, ovvero stimiamo per difetto...

... inoltre $\mathbb{E}[sd_X^2] < \sigma^2$, di modo che sd_X^2 sarà una sottostima della varianza di popolazione σ^2 .

Poiché siamo in grado di calcolare $\mathbb{E}[sd_X^2] = \frac{n-1}{n}\sigma^2$ ci conviene usare

$$S = \sqrt{\frac{(X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n-1}} = \sqrt{\frac{n}{n-1}} \times sd_X$$

come stimatore di σ (al posto di sd_X) in modo da avere $\mathbb{E}[S^2] = \sigma^2$.

Avremo ancora $\mathbb{E}[S] < \sigma$, ma avremo ridotto “al meglio delle nostre possibilità” l’ammontare della distorsione...

... pertanto la nostra stima della deviazione standard di popolazione non sarà la deviazione standard campionaria, ma quella che chiameremo **deviazione standard "corretta"**:

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}} = \sqrt{\frac{n}{n-1}} \times sd_x.$$

Si noti che:

- ▶ alla luce di quanto visto è il quadrato di s a essere (il valore fornito da uno stimatore) corretto, non s stessa;
- ▶ ai fini pratici $s \simeq sd_x$, se n non è troppo piccolo (diciamo $n \geq 10$).

Se per esempio, in un'indagine sul reddito dei membri di una certa comunità, si seleziona un campione casuale semplice di $n = 9$ membri della comunità e si osservano i seguenti redditi (in migliaia di euro)

$$\begin{array}{lll} x_1 = 25.5 & x_2 = 21.0 & x_3 = 40.2 \\ x_4 = 15.1 & x_5 = 22.2 & x_6 = 16.6 \\ x_7 = 18.8 & x_8 = 20.0 & x_9 = 19.3 \end{array}$$

allora stimeremo il reddito medio μ relativo all'intera comunità mediante il reddito medio osservato $\bar{x} = 22.08$ e la deviazione standard di popolazione σ mediante la deviazione standard "corretta" $s = 7.44$.

Si noti che $sd_x = 7.01 = \sqrt{\frac{8}{9}} \times s$ di modo che $\frac{s - sd_x}{s} \simeq 6\%$.

Stimare la deviazione standard di popolazione σ ci consente di calcolare l'**errore standard stimato**

$$\widehat{se} = \frac{s}{\sqrt{n}} = \frac{7.44}{\sqrt{9}} = 2.48$$

di modo che una **stima con errore standard** (stimato) del reddito medio di popolazione è data da

$$\mu = \bar{x} \pm \widehat{se} = 22.1 \pm 2.5$$

migliaia di euro: anche in questo caso riusciamo a valutare l'ordine di grandezza dell'errore di stima.

Nel caso particolare in cui X sia un **carattere dicotomico**, per esempio

$$X_i = \begin{cases} 1 & \text{se l}'i\text{-esimo intervistato è favorevole} \\ 0 & \text{se l}'i\text{-esimo intervistato è contrario} \end{cases}$$

con riferimento a una data proposta di legge, la media di popolazione (parametro di interesse) è la **proporzione di individui favorevoli**

$$\psi = \mathbb{E}[X_1] = \mathbb{P}\{X_1 = 1\}$$

nell'intera popolazione, mentre la media campionaria (stima puntuale per analogia) è la proporzione campionaria della modalità 1:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{n_1}{n}$$

La peculiarità di questo scenario, come già visto, è che la deviazione standard di popolazione è funzione del parametro di interesse:

$$\sigma = \sqrt{\psi(1 - \psi)}$$

Pertanto l'errore standard di \bar{X} varrà

$$se = \sqrt{\frac{\psi(1 - \psi)}{n}}$$

e converrà stimarlo sostituendo ψ con \bar{x} :

$$\widehat{se} = \sqrt{\frac{\bar{x}(1 - \bar{x})}{n}}$$

Per esempio, se la proporzione di intervistati favorevoli è

$$\bar{x} = 48\%$$

in un campione casuale semplice di $n = 1600$ cittadini (popolazione di interesse) troveremo la seguente **stima con errore standard** (stimato):

$$\psi = \bar{x} \pm \hat{se} = 48\% \pm 1.25\%$$

dal momento che

$$\hat{se} = \sqrt{\frac{0.48 \times 0.52}{1600}} = 0.0125$$


In generale, se stimiamo il parametro θ con lo stimatore T , in presenza di un parametro di disturbo ν per il quale disponiamo dello stimatore U , possiamo valutare l'ordine di grandezza dell'errore di stima $t - \theta$ con la radice quadrata dell'**errore quadratico medio** (mean square error)

$$MSE_{\theta,\nu}(T) = \mathbb{E}_{\theta,\nu}[(T - \theta)^2],$$

avendo cura, in pratica, di rimpiazzare θ e ν con le loro stime t e u ; si avrà un **trade-off tra correttezza e variabilità**, in quanto

$$MSE_{\theta,\nu}(T) = \text{Bias}_{\theta,\nu}(T)^2 + \text{Var}_{\theta,\nu}(T),$$

dove $\text{Bias}_{\theta,\nu}(T) = \mathbb{E}_{\theta,\nu}[T] - \theta$ è la distorsione di T (nulla per T corretto)¹

¹Nel caso della media si ha $\text{Bias}_{\mu,\sigma}(\bar{X}) \equiv 0$ e $\text{Var}_{\mu,\sigma}(\bar{X}) = \sigma^2/n$. 

Diremo che T_1 è **più efficiente** di T_2 quando

$$MSE_{\theta, \nu}(T_1) \leq MSE_{\theta, \nu}(T_2)$$

per ogni coppia θ, ν e la disuguaglianza è stretta per qualche coppia θ, ν ; se si confrontano stimatori corretti, basterà considerare le varianze²

Avremo che T_n è consistente (in media quadratica) quando

$$\lim_{n \rightarrow \infty} MSE_{\theta, \nu}(T_n) = 0$$

per ogni coppia θ, ν ; questo implica la correttezza asintotica di T_n .

²La media campionaria \bar{X} è lo **stimatore corretto a varianza minima** della media μ di X_1 quando X_1 abbia **distribuzione normale**.



Introduzione

Stima puntuale ed errore standard

Stima per intervallo

Stima di una media con deviazione standard nota

Stima di una media con deviazione standard incognita

Stima di una proporzione

Determinazione della numerosità campionaria

Stima di una deviazione standard

Il difetto principale degli stimatori puntuali è che tipicamente danno il valore vero del parametro con probabilità (praticamente) nulla; per es.

- ▶ se $\psi = 0.47$ e $n = 10$, ben che vada $\bar{x} = 0.5$ (o $\bar{x} = 0.4$);
- ▶ il teorema del limite centrale ci dice che, per n “grande”, la distribuzione campionaria di \bar{X} è approssimativamente normale, di modo che

$$\mathbb{P}\{\bar{X} = \mu\} \simeq 0.$$

Per questo abbiamo voluto, sin da subito, corredare le nostre stime puntuali con il loro errore standard (deviazione standard campionaria); questo quantifica l'**inaffidabilità di uno stimatore puntuale**, senza però darne una valutazione probabilistica. . .

... se vogliamo giungere a una valutazione probabilistica del procedimento di stima, ci converrà sostituire la stima puntuale con una **stima per intervallo**:

$$l(x_1, \dots, x_n) \leq \mu \leq u(x_1, \dots, x_n)$$

dove $l(x_1, \dots, x_n)$ e $u(x_1, \dots, x_n)$ sono opportune statistiche, ovvero funzioni dei dati, tali da garantire

$$l(X_1, \dots, X_n) \leq \mu \leq u(X_1, \dots, X_n)$$

con “buona probabilità” (es. 95%) vale a dire per “molti campioni”.

Per esempio, con riferimento alle misure ripetute della lunghezza di un pezzo meccanico, si trova

$$\bar{X} - 2 \times se \leq \mu \leq \bar{X} + 2 \times se$$

con probabilità 95% (la distribuzione campionaria di \bar{X} è normale perché normale è, con buona approssimazione, ogni singola misurazione).

Se osserviamo $\bar{x} \simeq 50000$ e sappiamo (dalle specifiche dello strumento, “correggendo” per $n = 4$) che $se = 1$, otteniamo la stima per intervallo

$$49998 \leq \mu \leq 50002$$

ovvero un intervallo di valori cui confidiamo appartenga quello vero (perché ciò accade nella “quasi totalità” dei possibili campioni).



La coppia di numeri aleatori

$$\begin{aligned}L &= l(X_1, \dots, X_n) = \bar{X} - 2 \times se \\U &= u(X_1, \dots, X_n) = \bar{X} + 2 \times se\end{aligned}$$

si dice **intervallo di confidenza** per μ .

Si dice invece **intervallo di confidenza stimato** la coppia di numeri

$$\begin{aligned}l(x_1, \dots, x_n) &= \bar{x} - 2 \times se = 49998 \\u(x_1, \dots, x_n) &= \bar{x} + 2 \times se = 50002\end{aligned}$$

ovvero la nostra stima per intervallo.

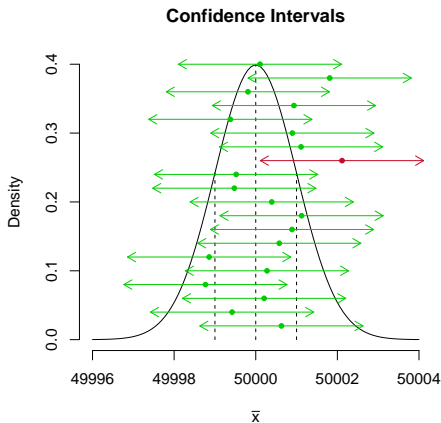
Il **livello di confidenza** $1 - 2\alpha$ dell'intervallo (L, U) è la probabilità

$$\mathbb{P}\{L \leq \mu \leq U\} = 1 - 2\alpha$$

che esso contenga il valore vero del parametro μ , vale a dire la proporzione di campioni in cui la stima per intervallo è giusta (es. $1 - 2\alpha = 0.95$, corrispondente ad $\alpha = 0.025$).

Infatti, per l'intervallo di confidenza stimato, delle due l'una:

- ▶ $l(x_1, \dots, x_n) \leq \mu \leq u(x_1, \dots, x_n)$ nel qual caso la stima è giusta;
- ▶ $\mu < l(x_1, \dots, x_n)$ oppure $u(x_1, \dots, x_n) < \mu$ nel qual caso la stima è sbagliata. . .



In pratica non possiamo sapere se una data stima per intervallo è giusta o sbagliata (dovremmo conoscere il valore del parametro, nel qual caso la stima non ci servirebbe).

Sappiamo tuttavia che il nostro **stimatore per intervallo** fornisce una stima giusta nella “quasi totalità” dei campioni che avremmo potuto selezionare e pertanto confidiamo (al livello di confidenza prescelto, es. 95%, in generale $1 - 2\alpha$) che la stima ottenuta sia giusta.

Se malauguratamente abbiamo selezionato un campione “sfortunato”, ci ritroviamo con una stima sbagliata e la nostra confidenza in tale stima è mal riposta.

Si noti che ci piacerebbe affermare

$$l(x_1, \dots, x_n) \leq \mu \leq u(x_1, \dots, x_n),$$

es. (misure ripetute della lunghezza di un pezzo meccanico)

$$49998 \leq \mu \leq 50002,$$

con probabilità 95, ma non possiamo...

...infatti, per farlo, dovremmo considerare μ come un numero aleatorio (invece che come una quantità deterministica, seppure incognita); in tal modo entreremmo nel reame della **statistica bayesiana**.

La determinazione di un intervallo di confidenza (stimato) per un parametro di interesse (es. media di popolazione) dipende dalla **distribuzione campionaria** della statistica usata come stimatore.

A tal fine, per quanto riguarda la stima della media di una popolazione, converrà distinguere i seguenti casi:

- ▶ caso non dicotomico (stima di μ)
 - ▶ σ nota
 - ▶ σ incognita
- ▶ caso dicotomico (stima di ψ)
 - ▶ campione “grande”
 - ▶ campione “piccolo”

Si noti come l'esempio delle misure ripetute della lunghezza di un pezzo meccanico rientri nel primo caso: stima di μ con σ nota.



Introduzione

Stima puntuale ed errore standard

Stima per intervallo

Stima di una media con deviazione standard nota

Stima di una media con deviazione standard incognita

Stima di una proporzione

Determinazione della numerosità campionaria

Stima di una deviazione standard

Vogliamo stimare μ e σ è nota.

Se

*per la distribuzione del carattere nella popolazione
(la distribuzione di X_1) vale l'**approssimazione normale**,*

oppure

*il campione è "grande" (diciamo $n \geq 30$) cosicché vale il
teorema del limite centrale,*

la distribuzione campionaria di $\frac{\bar{X}-\mu}{se}$, dove $se = \sigma / \sqrt{n}$ è noto,
è la distribuzione **normale standard**.

Pertanto

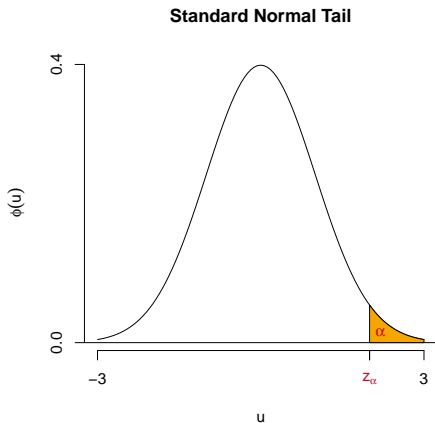
$$\bar{X} - z_{\alpha} \times se \leq \mu \leq \bar{X} + z_{\alpha} \times se$$

con probabilità $1 - 2\alpha$, se prendiamo z_{α} tale che

$$\Phi(z_{\alpha}) = 1 - \alpha,$$

dove Φ è la funzione di ripartizione normale standard; il valore z_{α} si dice **coefficiente di confidenza** al livello $1 - 2\alpha$.

Per esempio, se vogliamo un livello di confidenza pari al 95%, prenderemo $\alpha = (1 - 0.95)/2 = 0.025$ e troveremo (con un computer o usando una tavola) $z_{\alpha} = 1.96 \simeq 2 \dots$



... questo è quello che abbiamo fatto nell'esempio delle misure ripetute delle lunghezze di un pezzo meccanico.

Nello stesso esempio, se vogliamo un livello di confidenza pari al 90%, prenderemo $\alpha = (1 - 0.9) / 2 = 0.05$ e troveremo $z_{\alpha} = 1.64$, cosicché

$$49999.74 - 1.64 \times 1 \leq \mu \leq 49999.74 + 1.64 \times 1$$

ovvero

$$49998.10 \leq \mu \leq 50001.38$$

a meno che non ci abbia detto male (accade in un campione su dieci); si ricordi che $\sigma = 2$ e $n = 4$ forniscono $se = \sigma / \sqrt{n} = 1$, mentre $\bar{x} = 49999.74$ (recuperando le due cifre decimali).

Si noti che:

- ▶ poiché n è “piccolo”, la procedura descritta è valida in quanto vale l’approssimazione normale per X_1 (singola osservazione);
- ▶ se n fosse “grande” (diciamo $n \geq 30$) la procedura descritta sarebbe valida anche se non valesse l’approssimazione normale per X_1 (grazie al teorema del limite centrale);
- ▶ se avessimo n “piccolo”, ma non valesse l’approssimazione normale per la singola osservazione, saremmo nei guai. . .

. . . il che non sorprende, visto che avremmo pochi dati e nessun modello per spiegarli (dovremmo cercare un modello alternativo a quello normale).

Introduzione

Stima puntuale ed errore standard

Stima per intervallo

Stima di una media con deviazione standard nota

Stima di una media con deviazione standard incognita

Stima di una proporzione

Determinazione della numerosità campionaria

Stima di una deviazione standard

Vogliamo stimare μ e σ è incognita.

Se

*per la distribuzione del carattere nella popolazione
(la distribuzione di X_1) vale l'**approssimazione normale**,*

oppure

*il campione è "grande" (diciamo $n \geq 30$) cosicché vale il
teorema del limite centrale,*

la distribuzione campionaria di $\frac{\bar{X} - \mu}{\widehat{SE}}$, dove $\widehat{SE} = S / \sqrt{n}$, è la distribuzione **t di Student** con $n - 1$ gradi di libertà.

Si noti che una *t* di Student con "molti" gradi di libertà (diciamo $n \geq 101$) è praticamente indistinguibile dalla normale standard.



Pertanto

$$\bar{X} - t_\alpha \times \widehat{SE} \leq \mu \leq \bar{X} + t_\alpha \times \widehat{SE}$$

con probabilità $1 - 2\alpha$, se prendiamo t_α tale che

$$\mathbb{P}\{T \leq t_\alpha\} = 1 - \alpha,$$

dove T è un numero aleatorio che segue la distribuzione t di Student con $n - 1$ gradi di libertà.

Anche in questo caso l'uso di un opportuno **coefficiente di confidenza** (qui t_α ricavato per esempio da una tavola delle distribuzioni t di Student) permette di trasformare la nostra stima con errore standard in un intervallo di confidenza (al livello $1 - 2\alpha$).

Per esempio, se abbiamo osservato i redditi di $n = 9$ individui di una certa comunità, trovando

$$\bar{x} = 22.1 \quad \text{migliaia di euro}$$

$$s = 7.44 \quad \text{migliaia di euro,}$$

supponendo una distribuzione normale per il reddito nella comunità (supposizione “ardita” perché le distribuzioni di reddito sono tipicamente asimmetriche a destra) un intervallo di confidenza (stimato) al livello 99% per il reddito medio degli individui appartenenti alla comunità in questione si ottiene come segue. . .

- ▶ ... si prende $\alpha = 0.005$ (di modo che $1 - 2\alpha = 0.99$);
- ▶ si trova $t_\alpha = 3.36$ (es. da una tavola delle distribuzioni t di Student, selezionando il valore corrispondente a $n - 1 = 8$ gradi di libertà);
- ▶ si calcola l'errore standard stimato

$$\begin{aligned}\widehat{se} &= \frac{s}{\sqrt{n}} = \frac{7.44}{\sqrt{9}} = 2.48 \\ &= \frac{sd_x}{\sqrt{n-1}} = \frac{7.01}{\sqrt{8}} = 2.48;\end{aligned}$$

- ▶ si conclude

$$\begin{aligned}\bar{X} - t_\alpha \times \widehat{se} &\leq \mu \leq \bar{X} + t_\alpha \times \widehat{se} \\ 22.1 - 3.36 \times 2.48 &\leq \mu \leq 22.1 + 3.36 \times 2.48 \\ 13.77 &\leq \mu \leq 30.43\end{aligned}$$

al livello di confidenza 99%.

Si noti il ruolo giocato dalla distribuzione t di Student, confrontando $t_\alpha = 3.36$ con $z_\alpha = 2.56$: l'intervallo di confidenza è più ampio di quanto non sarebbe se adoperassimo la distribuzione normale standard.

Se il campione è “piccolo” e non vale l'approssimazione normale, siamo nel caso in cui occorre cercare un modello alternativo.

Se il campione è “molto grande” (come nell'esempio seguente) conviene prendere direttamente $t_\alpha = z_\alpha$; in pratica è come se σ fosse nota. . .

... supponiamo che la **spesa media settimanale per alimenti** in un campione casuale semplice di $n = 101$ famiglie, selezionate fra quelle residenti in una certa area geografica, sia pari a

$$\bar{x} = 315 \text{ euro}$$

con deviazione standard “corretta”

$$s = 82 \text{ euro.}$$

Un intervallo di confidenza (stimato) al livello 90% per la spesa media settimanale delle famiglie residenti nell'area in esame si ottiene come segue...

- ▶ ... si prende $\alpha = 0.05$ (di modo che $1 - 2\alpha = 0.9$);
- ▶ si trova $t_\alpha = z_\alpha = 1.64$ (es. da una tavola della normale standard);
- ▶ si calcola l'errore standard stimato

$$\widehat{se} = \frac{s}{\sqrt{n}} = \frac{82}{\sqrt{101}} = 8.16;$$

- ▶ si conclude

$$\begin{aligned} 315 - 1.64 \times 8.16 &\leq \mu \leq 315 + 1.64 \times 8.16 \\ 301.6 &\leq \mu \leq 328.4 \end{aligned}$$

al livello di confidenza 90%.

Introduzione

Stima puntuale ed errore standard

Stima per intervallo

Stima di una media con deviazione standard nota

Stima di una media con deviazione standard incognita

Stima di una proporzione

Determinazione della numerosità campionaria

Stima di una deviazione standard

Vogliamo stimare ψ .

Se il campione è “grande”, la distribuzione campionaria di $\frac{\bar{X} - \psi}{\widehat{SE}}$, dove $\widehat{SE} = \sqrt{\bar{X}(1 - \bar{X})/n}$, è la distribuzione **normale standard** (teorema del limite centrale).

In questo caso campione “grande” vuol dire:

- ▶ in teoria almeno $n\psi \geq 5$ e $n(1 - \psi) \geq 5$;
- ▶ in pratica almeno $n_1 = n\bar{x} \geq 5$ e $n_0 = n(1 - \bar{x}) \geq 5$.

Se invece il campione è “piccolo” ($n_1 < 5$ o $n_0 < 5$) la questione è più delicata: si può ricorrere alla **distribuzione esatta** di $n\bar{X}$ (distribuzione binomiale) come descritto per esempio da Borra & Di Ciaccio (2008, pag. 332).



Un intervallo di confidenza (stimato) per ψ , basato su un “grande” campione, si trova come

$$\bar{x} - z_{\alpha} \times \widehat{se} \leq \psi \leq \bar{x} + z_{\alpha} \times \widehat{se},$$

dove z_{α} si ricava da una tavola della distribuzione normale standard (o mediante un software statistico) in modo da garantire il livello di confidenza desiderato ($1 - 2\alpha$).

Se per esempio, con riferimento a una **proposta di legge**, si hanno

$$\begin{aligned} n_1 &= 768 \gg 5 && \text{intervistati favorevoli} \\ n_0 &= 832 \gg 5 && \text{intervistati contrari} \end{aligned}$$

di modo che $\bar{x} = 48\%$ e

$$\widehat{se} = \sqrt{\frac{\bar{x}(1 - \bar{x})}{n}} = \sqrt{\frac{48 \times 52}{1600}}\% = 1.25\%$$

troveremo un intervallo di confidenza al livello 80% prendendo $\alpha = 0.10$ e quindi $z_\alpha = 1.28$:

$$\begin{aligned} \bar{x} - z_\alpha \times \widehat{se} &\leq \psi \leq \bar{x} + z_\alpha \times \widehat{se} \\ 0.48 - 1.28 \times 0.0125 &\leq \psi \leq 0.48 + 1.28 \times 0.0125 \\ 46.4\% &\leq \psi \leq 49.6\% \end{aligned}$$

Quindi, al livello di confidenza 80%, escludiamo vi sia una maggioranza di elettori favorevole alla proposta di legge.

Introduzione

Stima puntuale ed errore standard

Stima per intervallo

Stima di una media con deviazione standard nota

Stima di una media con deviazione standard incognita

Stima di una proporzione

Determinazione della numerosità campionaria

Stima di una deviazione standard

Supponiamo di volere stimare una media di popolazione μ e, per semplicità, di conoscere la corrispondente deviazione standard σ (una stima essendo in pratica ugualmente utile).

Per ridurre l'errore standard sotto una soglia δ possiamo scegliere n in modo che si abbia

$$\frac{\sigma}{\sqrt{n}} \leq \delta$$

ovvero prendere $n \geq (\sigma/\delta)^2$... ovviamente intero!

Per esempio, con riferimento alle misure ripetute della lunghezza di un pezzo meccanico, per ridurre l'errore standard a meno di mezzo micron prenderemo $n \geq (2/0.5)^2 = 16$.

Introduzione

Stima puntuale ed errore standard

Stima per intervallo

Stima di una media con deviazione standard nota

Stima di una media con deviazione standard incognita

Stima di una proporzione

Determinazione della numerosità campionaria

Stima di una deviazione standard

Vogliamo stimare σ e supponiamo senz'altro che μ non sia nota.

Se

*per la distribuzione del carattere nella popolazione
(la distribuzione di X_1) vale l'**approssimazione normale**,*

oppure

*il campione è "grande" (diciamo $n \geq 30$) cosicché vale il
teorema del limite centrale e la distribuzione di popolazione ha
code della stessa "pesantezza" della normale (tecnicamente se
 $\mathbb{E}[(X_1 - \mu)^4] = 3\sigma^4$ come per la distribuzione normale),*

la distribuzione campionaria di $\frac{(n-1)S^2}{\sigma^2}$ è la distribuzione **chi-quadrato**
con $n - 1$ gradi di libertà.



Pertanto

$$\frac{(n-1)S^2}{\chi_\alpha^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\alpha}^2}$$

con probabilità $1 - 2\alpha$, se prendiamo χ_α^2 e $\chi_{1-\alpha}^2$ in modo che

$$\begin{aligned}\mathbb{P}\{X^2 \geq \chi_{1-\alpha}^2\} &= 1 - \alpha \\ \mathbb{P}\{X^2 \geq \chi_\alpha^2\} &= \alpha\end{aligned}$$

dove X^2 è un numero aleatorio che segue la distribuzione chi-quadrato con $n - 1$ gradi di libertà.

Si noti che, poiché le distribuzioni chi-quadrato non sono simmetriche, servono **due percentili** (uno per ogni coda).



Nell'esempio della **spesa media settimanale per alimenti** rilevata in un campione casuale semplice di $n = 101$ famiglie avevamo registrato una deviazione standard "corretta"

$$s = 82 \text{ euro.}$$

Troveremo un intervallo di confidenza (stimato) al livello 90% per la deviazione standard di popolazione come

$$s\sqrt{\frac{n-1}{\chi_{\alpha}^2}} \leq \sigma \leq s\sqrt{\frac{n-1}{\chi_{1-\alpha}^2}}$$

dove...

... $\alpha = 0.05$, in modo che $1 - 2\alpha = 0.9$,

$$\chi_{1-\alpha}^2 = 77.9294$$

$$\chi_{\alpha}^2 = 124.3421,$$

come indicato da un software statistico o da una tavola dei valori critici delle distribuzioni chi-quadrato, e $n - 1 = 100$; sarà dunque

$$74 = 82 \times \sqrt{\frac{100}{124}} \leq \sigma \leq 82 \times \sqrt{\frac{100}{77.9}} = 93$$

al livello di confidenza 90%.

In caso di code di “pesantezza” diversa dalla normale la distribuzione di S^2 dipende da un parametro di disturbo (non approfondiamo).

 BORRA, S. & DI CIACCIO, A. (2008).

Statistica: Metodologie per le Scienze Economiche e Sociali
(Seconda Edizione).

McGraw-Hill, Milano.