Stima puntuale e stima per intervalli

La stima puntuale è poco pratica perché

- **La probabilità di individuare il valore esatto è zero:** $P(T = \theta) = 0$
- **In ogni stima è insito un certo errore dovuto alle fluttuazioni campionarie e la stima puntuale non ne dà alcuna misura**

Ciò considerato ha più senso la stima per intervalli

Dicesi stima per intervalli di un parametro della v.c. la determinazione di due stimatori che contengano il valore teorico con una probabilità prefissata

$$L(X_1, X_2, \ldots, X_n); \quad U(X_1, X_2, \ldots, X_n)$$

$$P(L_n \le \theta \le U_n) \ge 1 - \alpha$$

(1-α) è il LIVELLO O COEFFICIENTE DI CONFIDENZA

Intervalli osservati

Rilevazione del peso di alcune monete di nuova coniazione con misure arrotondate al più vicino decimilligrammo

X_i	n_i
2.99	1
3.01	4
3.03	4
3.05	4
3.07	7
3.09	17
3.11	24
3.13	17
3.15	13
3.17	6
3.19	2
3.21	1
	100

Le stime intervallari più ovvie per il peso medio sono:

$$(-\infty, 3.21], \quad [2.99, \infty), \quad [2.99, 3.21]$$

Che sono, peraltro, applicabili a tutte le misure di centralità.

Il livello di confidenza è del 100%, ma la lunghezza dell'intervallo può risultare eccessiva se la variabilità delle misure è elevata

Statistiche di riferimento (pivotali)

Le stime intervallari si costruiscono utilizzando un particolare tipo di statistica T(X;θ) i cui valori siano funzione:

- ovviamente del campione casuale $(X_1, X_2, \ldots, X_n)$
- apparentemente del parametro da stimare: θ

La dipendenza da θ non è reale. Infatti, la distribuzione di T non dipende da θ, ma il parametro fà da perno (pivot) per determinare i limiti inferiore e superiore.

Le statistiche pivotali sono del tipo: $T(X;\theta) = \dfrac{T - \theta}{\sigma(T)}$

Intervalli di Tchebycheff

Supponiamo di ignorare del tutto la distribuzione dello stimatore T

In base alla disuguaglianza di Tchebycheff abbiamo: $P\left(|T - \theta| \le k\,\sigma(T)\right) \ge 1 - \dfrac{1}{k^2}$

Ponendo $k = \dfrac{1}{\sqrt{\alpha}}$ **si ha:** $P\left(|T - \theta| \le \dfrac{\sigma(T)}{\sqrt{\alpha}}\right) \ge 1 - \alpha$ **Pertanto:**

$$T - \frac{\sigma(T)}{\sqrt{\alpha}} \le \theta \le T + \frac{\sigma(T)}{\sqrt{\alpha}}$$

è un intervallo di lunghezza $\dfrac{2\sigma(T)}{\sqrt{\alpha}}$ **con un livello di confidenza di almeno l'(1-α)%**

La generalità di questo risultato lo rende molto grezzo e poco utile in pratica

Esempio

Per stimare il tempo di presenza davanti al televisore delle famiglie si è considerato un campione di ampiezza n=81 e si è trovato:

$$\bar{x} = 28.4, \quad s = 8.3$$

Un intervallo di confidenza non inferiore al 95% (α=0.05) è $P\left(|\bar{x} - \mu| \le \frac{\sigma}{\sqrt{0.05 * 81}}\right)$

Supponiamo inoltre che "n" sia tanto grande da consentire di sostituire "s" a "σ"

$$28.4 - \frac{8.3}{2.012} \le \mu \le 28.4 + \frac{8.3}{2.012} \quad \Rightarrow \quad 24.28 \le \mu \le 32.52$$

L'intervallo è piùttosto ampio. La sua lunghezza può essere ridotta in vari modi:

1. Riducendo il livello di confidenza (aumentando α)
2. Aumentando l'ampiezza del campione
3. Individuando la distribuzione dello stimatore

Intervalli di confidenza

In diversi casi è nota la distribuzione della statistica pivotale "T(X;θ)"

Tale distribuzione consente di determinare i limiti dell'intervallo

$$P(L_n \le T(X;\theta) \le U_n) = 1 - \alpha$$

I valori dei limiti si ottengono attraverso i quantili della distribuzione di T che deve essere poi espressa come segue

$$L_n(X_1, X_2, \ldots, X_n) \le \theta \le U_n(X_1, X_2, \ldots, X_n)$$

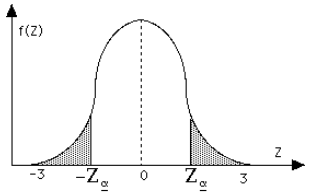
I due limiti sono due statistiche e quindi delle variabili casuali che includono il valore di θ con probabilità (1-α)

Intervalli di confidenza per la media/σ noto

La distribuzione campionaria di $\bar{x}$ è esprimibile con la standardizzata

$$Z = \sqrt{n}\left(\frac{\bar{x} - \mu}{\sigma}\right) \sim N(0,1)$$

Possiamo asserire che, con probabilità (1-α), la media campionaria sarà compresa nell'intervallo $\left[-Z_{\frac{\alpha}{2}},\ Z_{\frac{\alpha}{2}}\right]$



Quali sono i limiti dell'intervallo?

$$P\left(-Z_{\frac{\alpha}{2}} \le Z \le Z_{\frac{\alpha}{2}}\right) = 1 - \alpha \Rightarrow P\left(-Z_{\frac{\alpha}{2}} \le \sqrt{n}\left(\frac{\bar{x} - \mu}{\sigma}\right) \le Z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

$$= P\left(\bar{x} - Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Questo é il livello di confidenza

Dobbiamo dare un significato all'intervallo: $\bar{x} \pm Z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$

La lunghezza dell'intervallo è $\frac{2\sigma Z_{\frac{\alpha}{2}}}{\sqrt{n}}$

Esercizio

Si vuole ottenere una stima intervallare del consumo medio annuo di un certo bene il cui comportamento sia descritto da una v.c. normale N(*μ*,9)

a) Per un campione di ampiezza n=25 che fornisce la stima: $\bar{x} = 58.3$ **calcolare un stima intervallare al 90% per la media *μ***

In generale: $\bar{x} - Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ **(Per il valore di** $Z_{\frac{\alpha}{2}}$ **si usano le tavole dei percentili della normale)**

Nel particolare: $58.3 - 1.645\frac{3}{\sqrt{25}} \le \mu \le 58.3 + 1.645\frac{3}{\sqrt{25}} \Rightarrow 57.13 \le \mu \le 59.287$

b) Quando deve essere l'ampiezza del campione perché la lunghezza dell'intervallo sia pari ad uno?

$$\text{Lunghezza} = 2 * Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \Rightarrow n = \left[\left(2Z_{\frac{\alpha}{2}}\frac{\sigma}{L}\right)^2\right]$$

parte intera

Nell'esempio:

$$n = \left[\left(2 * 1.645\frac{3}{1}\right)^2\right] = 97$$

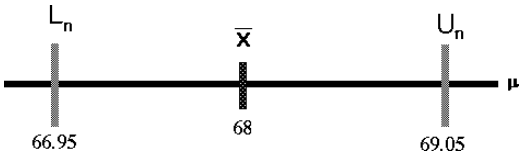
Interpretazione

Supponiamo di voler stimare l'incasso medio di una catena di negozi di cui è nota la deviazione standard: σ=3.2. Con un campione di n=36 si è ottenuto: $\bar{x} = 68$

Determiniamo l'intervallo di confidenza al 95% per l'incasso medio

Dalle tavole della normale si apprende che $Z_{\frac{0.05}{2}} = Z_{0.025} = 1.96$

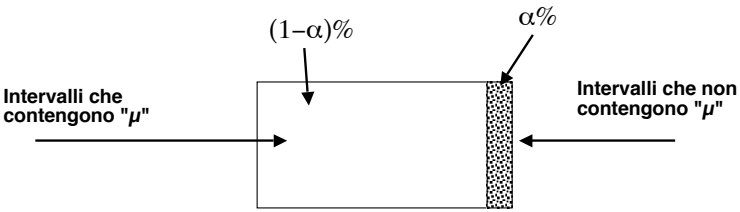
$$68 - 1.96 * \frac{3.2}{\sqrt{36}} \le \mu \le 68 + 1.96 * \frac{3.2}{\sqrt{36}} \quad \Rightarrow \quad 66.95 \le \mu \le 69.05$$



Interpretazione/2

GIUSTA: il 95% dI TUTTI I POSSIBILI intervalli, ognuno basato su di un campione diverso, costruiti con questo schema, conterrà il parametro vero (secondo l'interpretazione fisica della probabilità).

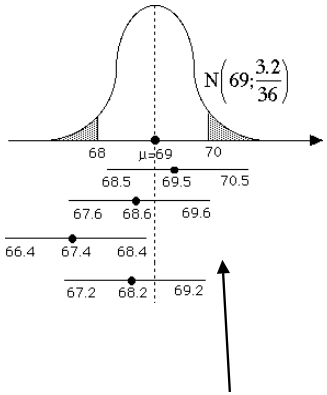
Tuttavia non è possibile essere sicuri che UN PARTICOLARE INTERVALLO contenga o no il parametro



SBAGLIATA: L'intervallo contiene il valore vero con una probabilità del 95% (in effetti "*μ*" è un parametro che non può variare a piacimento: o è incluso oppure no).

Approfondimento

Supponiamo che la media della popolazione sia 69 e che, ignorandola, si proceda alla estrazione di 5 diversi campioni di ampiezza 36



Ognuno di questi è una stima per intervalli da un diverso campione

La distribuzione di $\bar{x}$ **è tale che la probabilità che essa ricada tra 68-70 è del 95%**

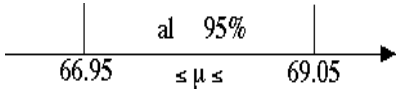
Poichè ignoriamo "*μ*" questo non è di grande aiuto. Sulla base di un campione di n=36 si effettua una stima per intervalli.

Il 95% di TUTTI I POSSIBILI INTERVALLI include "*μ*".

Non sappiamo però se il singolo intervallo contiene o non contiene *μ*

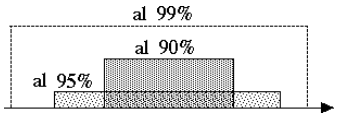
Se cambia il grado di confidenza ...

ESEMPIO: nella stima dell'incasso medio, al 95% si era ottenuto l'intervallo:



$$al\ 99\%\ si\ ha\ Z_{\frac{0.01}{2}} = 2.576 \Rightarrow 68 - 2.576 * \frac{3.2}{6} \le \mu \le 68 + 2.576 * \frac{3.2}{6} \Rightarrow 66.626 \le \mu \le 69.374$$

$$al\ 90\%\ si\ ha\ Z_{\frac{0.1}{2}} = 1.645 \Rightarrow 68 - 1.645 * \frac{3.2}{6} \le \mu \le 68 + 1.645 * \frac{3.2}{6} \Rightarrow 67.123 \le \mu \le 68.877$$



Se aumenta il grado di confidenza aumenta pure l'ampiezza dell'intervallo (infatti aumenta la "Z")

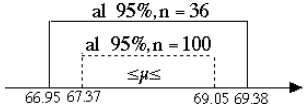
Più si vuole essere sicuri di non sbagliare, più generica deve essere l'asserzione

Se, a parità di (1-α), aumenta "n"...

Proviamo a calcolare un intervallo di confidenza al 95%, ma con n=100 nella stima dell'incasso medio dove

$$\bar{x} = 68;\ \sigma = 3.2$$

$$68 - 1.96 * \left(\frac{3.2}{10}\right) \le \mu \le 68 + 1.96 * \left(\frac{3.2}{10}\right) \Rightarrow 67.37 \le \mu \le 68.63$$



l'intervallo al 95% con n=100 è più stretto di quello con n=36.

A parità di livello di confidenza, un "n" più grande ci dà più informazioni e la nostra asserzione su "μ" deve essere più precisa.

Da notare però che la riduzione dell'intervallo procede in ragione di $\sqrt{n}$

Intervalli asimmetrici

Finora si sono scelti limiti equidistanti dalla media campionaria, ma nulla ci vieta di usare ad esempio

$$-Z_{\frac{2\alpha}{3}}, Z_{\frac{\alpha}{3}};\ -Z_{\frac{\alpha}{4}}, Z_{\frac{3\alpha}{4}}$$

Applichiamo il primo tipo all'esempio degli incassi medi con α=5%

$$-Z_{\frac{2\alpha}{3}} = -Z_{0,0333} = 1.83;\quad Z_{\frac{\alpha}{3}} = Z_{0,0167} = 2.13$$

$$68 - 1.83\left(\frac{3.2}{6}\right) \le \mu \le 68 + 2.13\left(\frac{3.2}{6}\right) \Rightarrow 67.02 \le \mu \le 69.14$$

La lunghezza di tale intervallo 2.12=69.14-67.02 è maggiore di quella con estremi Simmetrici: 1.76=68.88-67.12

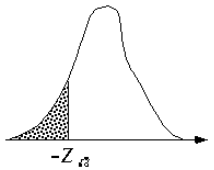
In generale, gli intervalli asimmetrici sono più lunghi di quelli simmetrici a parità di livello di confidenza.

Intervalli unidirezionali

Questi intervalli, utili quando è noto che un parametro non può scendere al di sotto (o salire al di sopra di una certa soglia) si costruiscono così

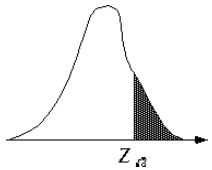
Unidirezionale superiore: $P\left(\mu \le \bar{x} + Z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$

Unidirezionale inferiore: $P\left(\mu \ge \bar{x} - Z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$



$$Z_{0.95} = 1.645 \Rightarrow \mu \le 68 + 1.645 * \frac{3.2}{\sqrt{36}} = 68.877$$

$$Z_{0.05} = -1.645 \Rightarrow\ \mu \ge 68 - 1.645 * \frac{3.2}{\sqrt{36}} = 67.12$$



(Per il valore di Z_α si usano le tavole dei percentili della normale

Esempio

La presenza di un composto nell'acqua minerale non deve superare l'ammontare di 0.005g per litro (ogni livello inferiore è accettabile, anche l'assenza completa.

In un campione casuale di prelievi si è riscontrato:

0.0045
0.0023
0.0038
0.0019
0.0033
0.0047
0.0051
0.0048
0.0017
0.0027
0.0045
0.0016
0.0063
0.0011

Sapendo che la deviazione standard dell'intera produzione è di 0.0016g determinare un intervallo di stima unidirezionale al 98% per la presenza media del composto

$$P\left(\mu \le 0.0035 + Z_{.02}\frac{0.0016}{\sqrt{15}}\right) = 0.98 \Rightarrow \mu \le 0.0035 + 2.0537 * \frac{0.0016}{\sqrt{15}}$$

$$\mu \le 0.0044$$

Poiché la stima si mantiene sotto il livello di soglia il processo appare regolare, almeno al livello di confidenza prescelto

Esercizio

Un avvocaticchio cerca una stima per intervalli degli incassi medi dello studio in cui è stato inserito dopo non poche pressioni. La Boss lo informa che gli scarti non sono diversi da quelli provinciali: σ= 400 mila lire.

Un campione di n=100 incassi ha prodotto una media di $\bar{x} = 700$ **mila lire**

a) Siccome non è brillante calcola un intervallo asimmetrico con

$$\alpha = 0.05; \quad \alpha_1 = \frac{\alpha}{4}; \quad \alpha_2 = \frac{3\alpha}{4}; \quad \alpha_1 + \alpha_2 = \alpha$$

Questi valori si trovano sulle tavole della normale

cosa otterrà?

$$\bar{x} - Z_{\alpha_1}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + Z_{\alpha_2}\frac{\sigma}{\sqrt{n}} \Rightarrow 700 - Z_{0.0125}\frac{400}{\sqrt{100}} \le \mu \le 700 + Z_{0.0375}\frac{400}{\sqrt{100}}$$

$$700 - 2.24\frac{400}{\sqrt{100}} \le \mu \le \bar{x} + 1.78\frac{400}{\sqrt{100}} \Rightarrow 610.4 \le \mu \le 771.2$$

b) La Boss gli dice bruscamente che Lei non lavora per meno di 650 mila lire e chiede un intervallo unilaterale superiore all'99%.

$$\mu \le \bar{x} + Z_{\alpha}\frac{\sigma}{\sqrt{n}} \Rightarrow \mu \le 700 + 2.326 * \frac{400}{\sqrt{100}} = 793.04$$

Se σ è incognito...

Lo sviluppo degli intervalli di confidenza ha finora assunto che σ fosse noto e ciò è molto irrealistico.

In genere "s" è stimato con una statistica sufficiente, centrata, consistente:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{x})^2}{n-1}}$$

La statistica pivotale utilizzata è

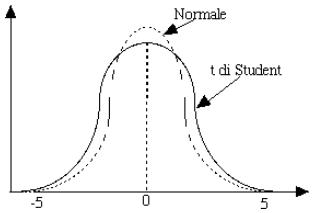
$$t_n = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \sqrt{n}\left(\frac{\bar{x} - \mu}{s}\right)$$

Questa v.c. è stata analizzata da W.S. Gosset, nel 1906, che firmò l'articolo con lo pseudonimo "STUDENT" ed è da allora nota come "La t di Student"

Perchè sia applicabile è necessario che il campione provenga dalla Normale

La t di Student

Questo modello è molto simile alla gaussiana, ma ha code più "spesse"



$$-\infty < t < \infty; \quad n > 2$$

$$E(t_n) = 0; \quad Var(t_n) = \frac{n}{n-2}$$

la varianza è superiore all'unità, ma si avvicina ad uno all'aumentare di "n"

L'elemento caratterizzante della t di Student sono "i gradi di libertà" cioè l'ampiezza campionaria ridotta di una unità: n-1. ("n-1" è il parametro della t di Student)

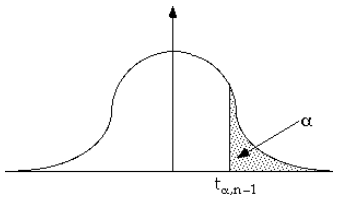
Per ogni grado di libertà esiste una t di Student, sebbene queste diventino poco distinguibili per n≥60.

Come per la Normale, anche per la "t" esistono le tavole (meno dettagliate, però)

Intervalli di confidenza per *μ* con σ ignoto

"σ" è stimato con la varianza campionaria "s"

$$P\left(|t_n| \le t_{\frac{\alpha}{2},n-1}\right) = 1 - \alpha$$



risolvendo rispetto a "*μ*"

$$\bar{x} - t_{\frac{\alpha}{2},n-1} * \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\frac{\alpha}{2},n-1} * \frac{s}{\sqrt{n}}$$

ESEMPIO

Una certa macchina produce confezioni il cui peso è descritto dalla gaussiana. Per un campione di ampiezza n=16 si è trovato $\hat{\mu} = 282; \ \hat{\sigma} = 8$

Determinare un intervallo di confidenza al 95%

Dalle tavole della t di Student si ottiene $t_{\frac{\alpha}{2},n-1} = t_{\frac{0.05}{2},15} = t_{0.025,15} = 2.131$

Per cui l'intervallo cercato è

$$282 - 2.131\left(\frac{8}{\sqrt{16}}\right) \le \mu \le 282 + 2.131\left(\frac{8}{\sqrt{16}}\right) \Rightarrow 272.4 \le \mu \le 286.26$$

Esempio

i risultati dell'analisi del contenuto medio in grammi di una data sostanza in 15 confezioni di prodotto sono stati:

{26.7, 28.8, 24.0, 24.9, 26.4, 25.9, 24.4, 21.7, 24.1, 25.9, 27.3, 26.9, 27.3, 24.8, 23.6}

Ottenere un intervallo di confidenza al 95% per il contenuto medio.

Abbiamo: $\hat{\mu} = \frac{\sum_{i=1}^{15} x_i}{n}; \quad \hat{\sigma}^2 = \frac{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}{n-1}$

$$\hat{\mu} = \frac{382.70}{15} = 25.51; \quad \hat{\sigma}^2 = \frac{(9810.17 - 15 * 25.51^2)}{15-1} = 3.4835$$

x	x^2
26.70	712.89
28.80	829.44
24.00	576.00
24.90	620.01
26.40	696.96
25.90	670.81
24.40	595.36
21.70	470.89
24.10	580.81
25.90	670.81
27.30	745.29
26.90	723.61
27.30	745.29
24.80	615.04
23.60	556.96
Σ=382.70	Σ=9810.17

l'intervallo cercato è:

$$\hat{\mu} - t_{\frac{\alpha}{2},n-1}\frac{\hat{\sigma}}{\sqrt{n}} \le \mu \le \hat{\mu} + t_{\frac{\alpha}{2},n-1}\frac{\hat{\sigma}}{\sqrt{n}} \Rightarrow 25.51 - t_{0.025,14} * 0.8994 \le \mu \le 25.51 + t_{0.025,14} * 0.8994$$

$$25.51 - 2.145 * 0.8994 \le \mu \le 25.51 + 2.145 * 0.8994 \Rightarrow 23.58 \le \mu \le 27.44$$

Intervalli unidirezionali

Le stime intervallari con σ incognito possono essere estese ai casi già segnalati

Unidirezionale superiore: $P\left(\mu \le \bar{x} + t_{\alpha,n-1} * \frac{s}{\sqrt{n}}\right) = 1 - \alpha$

Unidirezionale inferiore: $P\left(\mu \ge \bar{x} - t_{\alpha,n-1} * \frac{s}{\sqrt{n}}\right) = 1 - \alpha$

ESEMPIO

I prezzi (in migliaia di lire) di un certo prodotto rilevati in n=4 punti diversi sono stati: 260, 258, 252, 256. Quindi $\bar{x} = 256.6, \; s = 3.42$

Supponiamo che ci sia un prezzo imposto non minore di 250. Al 99%: $250 \le \mu \le 256.5 + 4.541 * \frac{3.42}{2} = 264.265$

Supponiamo che il prezzo non possa superare 265. Al 95% si ha: $256.5 - 2.353 * \frac{3.42}{2} = 252.476 \le \mu \le 265$

Considerazione

La stima intervallare della media si basa sulla normalità o su campioni grandi. Che succede se queste condizioni vengono meno?

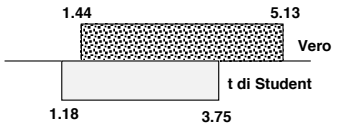
Consideriamo un campione ($X_1, X_2, \ldots, X_n$) I.I.D. dalla $f(X_i;\theta) = \frac{e^{-\frac{X_i}{\theta}}}{\theta}; \; X_i > 0$

Lo stimatore appropriato di θ è $\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$

La distribuzione di questa statistica può essere ricondotta alla χ^2

$$P\left(\frac{2n\bar{x}}{\chi^2_{\frac{\alpha}{2},2n}} \le \theta \le \frac{2n\bar{x}}{\chi^2_{1-\frac{\alpha}{2},2n}}\right) = 1 - \alpha$$

Per i valori: $n = 10, \; \bar{x} = 2.464, \; s = 1.9382$



Le decisioni basate sull'intervallo approssimato sono spesso errate

Generalizzazioni

Per campioni di ampiezza n>30, estratti da una normale con σ incognito si può usare la Normale invece della "t" di Student vista la somiglianza che ora si realizza tra i due modelli.

$$P\left(\bar{x} - Z_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + Z_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}\right) \ge 1 - \alpha$$

Per "n" abbastanza grande si può usare lo stesso intervallo anche per campioni non provenienti dalla normale (Teorema del limite centrale).

$$P\left(\bar{x} - Z_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + Z_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}}\right) \ge 1 - \alpha$$

Questi intervalli sono detti intervalli di confidenza per grandi campioni

Esempio

Un campione di n=56 tassi di disoccupazione ha dato luogo ad una media campionaria di $\hat{\mu} = 8.17$ ed una deviazione standard campionaria $\hat{\sigma}$=1.42

Si determini un intervallo di confidenza per grandi campioni con un livello del 95%

$$8.17 - Z_{\frac{0.05}{2}}\frac{1.42}{\sqrt{56}} \le \mu \le 8.17 + Z_{\frac{0.05}{2}}\frac{1.42}{\sqrt{56}} \Rightarrow 8.17 - 1.96\frac{1.42}{\sqrt{56}} \le \mu \le 8.17 + 1.96\frac{1.42}{\sqrt{56}}$$

$$8.17 - 0.37 \le \mu \le 8.17 + 0.37 \Rightarrow 7.80 \le \mu \le 8.54$$

L'uso della "t" di student avrebbe implicato

$$8.17 - t_{\frac{0.05}{2},55}\frac{1.42}{\sqrt{56}} \le \mu \le 8.17 + 2\frac{1.42}{\sqrt{56}} \Rightarrow 8.17 - 2\frac{1.42}{\sqrt{56}} \le \mu \le 8.17 + \quad 2\frac{1.42}{\sqrt{56}}$$

$$8.17 - 0.38 \le \mu \le 8.17 + 0.38 \Rightarrow 7.79 \le \mu \le 8.55$$

con differenze di scarsissimo rilievo

Generalizzazioni/2

Gli intervalli di confidenza per la media si modificano leggermente quando le unità sono estratte da una popolazione finita.

In questo caso:

$$E(\bar{x}) = \mu; \quad \sigma^2(\bar{x}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

La distribuzione di $\bar{x}$ tende inoltre ad essere ben approssimata dalla normale di modo che gli intervalli di confidenza per la media sono dati dalle formule:

$$\bar{x} - Z_{\frac{\alpha}{2}}\sigma\sqrt{\left[\frac{N-n}{n(N-1)}\right]} \le \mu \le \bar{x} + Z_{\frac{\alpha}{2}}\sigma\sqrt{\left[\frac{N-n}{n(N-1)}\right]}$$

Se "σ" non è noto può essere stimato con "s"

Se "n" è molto inferiore a "N" cioè n/N≤0.05 il fattore di correzione (N-n)/(N-1) può essere ignorato.

Esempio

In un quartiere residenziale sono presenti 63 condominî. Si vuole stimare il numero medio di residenti esaminando un campione casuale di n=7 condominî.

Ecco i risultati:

24 15 16 19 40 33 28 $\bar{x}$ **=25** s^2 **=86**

il fattore di correzione è: $\frac{N-n}{N-1} = \frac{63-7}{63-1} = \frac{56}{62} = 0.9032$

Supponiamo che il livello di confidenza (1-α) sia del 90%. Ecco gli estremi:

$$25 - 1.6449 * 9.2736 * \sqrt{\left[\frac{63-7}{7(63-1)}\right]} \le \mu \le 25 + 1.6449 * 9.2736 * \sqrt{\left[\frac{63-7}{7(63-1)}\right]}$$

$$19.522 \le \mu \le 30.478$$

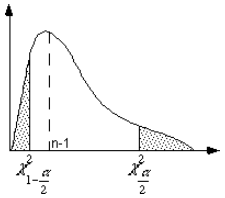
l'intervallo non è soddisfacente, ma nelle condizioni poste era difficile fare di meglio

Stima intervallare della varianza

Si ipotizza che il campione sia estratto dalla Normale. La statistica pivotale è:

$$C_n^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1} \qquad P\left(\chi^2_{1-\frac{\alpha}{2},n-1} \le (n-1)\frac{s^2}{\sigma^2} \le \chi^2_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$$

L'intervallo è pertanto: $\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}$



ESEMPIO

L'affitto mensile di un residence ha una base fissa. Varia il costo degli optional: 1.75, 2.25, 1.9, 2.3, 2.1, 1.7. Quindi $n = 6,\ \bar{x} = 2,\ (n-1)s^2 = 0.065$

L'intervallo al 99% è: $\frac{0.065}{16.750} = 0.0039 \le \sigma^2 \le 0.1578 = \frac{0.065}{0.412}$

Esercizio

Nel condurre un esperimento sui tempi di esplosione di alcuni detonatori si sono ottenuti i seguenti scarti dal valore di riferimento (arrotondati al millisecondo).

	11
	23
	25
	9
	2
	6
	-2
	-6
s=9.28	2
	8
	9
	19
	0
	2
	14
	-4

Il controllo della variabilità è essenziale per assicurare la simultaneità delle detonazioni e quindi l'efficacia dello scoppio.

Costruiamo un intervallo unidirezionale superiore $0<\sigma<K$ al 90% per la deviazione standard di questo esempio.

$$0.90 = P\left(\frac{16s^2}{\sigma^2} > 8.55\right) = P\left(\sigma^2 < \frac{16s^2}{8.55}\right) = P\left(\sigma < \frac{4s}{\sqrt{8.55}}\right) = P(\sigma < 1.368s)$$

$$0 < \sigma < 1.369 * 9.28 \Rightarrow 0 < \sigma < 12.704$$

Costruiamo un intervallo unidirezionale inferiore $\sigma>K$ al 99% per la deviazione standard

$$0.01 = P\left(\frac{16s^2}{\sigma^2} < 30.6\right) = P\left(\sigma^2 > \frac{16s^2}{30.6}\right) = P\left(\sigma > \frac{4s}{\sqrt{30.6}}\right) = P(\sigma < 0.723s)$$

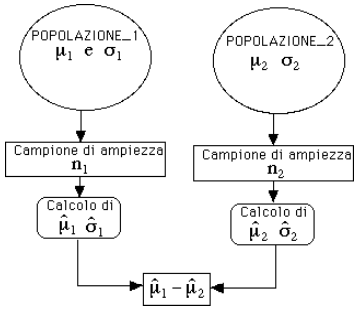
$$\sigma > 0.723 * 9.28 \Rightarrow \sigma < 6.709$$

Stima intervallare della differenza tra medie

in molte situazioni si è costretti a decidere in base a dei campioni sulla differenza tra le medie di due popolazioni.

ESEMPI

- **la nuova linea di produzione ha meno scarti di quella vecchia**
- **Un vaccino è più efficace di un altro**
- **Gli spot televisivi aumentano gli incassi più di quelli sui quotidiani**



La stima puntuale è immediata: la semplice differenza tra le medie, quella intervallare richiede qualche considerazione in più

Intervalli per (μ_1-μ_2)

Campioni indipendenti, varianze note e popolazioni normali

In questo caso si rientra nella combinazione lineare di v.c. incorrelate per cui

$$E(\hat{\mu}_1 - \hat{\mu}_2) = \mu_1 - \mu_2 \quad e \quad \sigma^2(\hat{\mu}_1 - \hat{\mu}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Se i campioni provengono dalla Normale possiamo porre

$$Z = \frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

ed i limiti dell'intervallo di confidenza saranno

$$(\hat{\mu}_1 - \hat{\mu}_2) - Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le (\mu_1 - \mu_2) \le (\hat{\mu}_1 - \hat{\mu}_2) + Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Esempio

La "Annamaria Filice & C." adopera una linea di "packaging" per il caffé tale da produrre confezioni con un dato peso medio e con un SQM di 6 grammi.

La titolare sta valutando una nuova linea di packaging che produce confezioni con lo stesso peso medio, ma con un SQM di 3 grammi

Supponiamo che il peso delle confezioni abbia in entrambi i casi distribuzione normale

Si considerano due campioni, che danno i seguenti risultati

$$\text{"Old"} \Rightarrow \begin{matrix} n_1 = & 25 \\ \hat{\mu}_1 = & 450 \end{matrix} \qquad \text{"New"} \Rightarrow \begin{matrix} n_2 = & 36 \\ \hat{\mu}_2 = & 510 \end{matrix}$$

Costruiamo un intervallo di confidenza al 95%

$$-60 - 1.96\sqrt{\frac{36}{25} + \frac{9}{36}} \le (\mu_1 - \mu_2) \le -60 + Z_{\frac{\alpha}{2}}\sqrt{\frac{36}{25} + \frac{9}{36}} \Rightarrow -62.55 \le (\mu_1 - \mu_2) \le -57.45$$

La differenza tra le medie campionarie è compresa tra due valori negativi. La nuova linea sembra preferibile alla vecchia

Intervalli per $(\mu_1-\mu_2)/2$

Campioni indipendenti, varianze incognite, campioni grandi (*Large samples*)

Se si ignora la v.c. di partenza ma gli "n" sono grandi al punto da rendere valido il T.L.C. gli intervalli si realizzano sostituendo le stime delle varianze

$$(\hat{\mu}_1-\hat{\mu}_2)-Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_2^2}{n_2}} \le (\mu_1-\mu_2) \le (\hat{\mu}_1-\hat{\mu}_2)+Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\sigma}_1^2}{n_1}+\frac{\hat{\sigma}_2^2}{n_2}}$$

N.B.

$$\hat{\sigma}_1^2=\frac{\sum_{i=1}^{n_1}(X_{i1}-\hat{\mu}_1)}{n_1-1};\quad \hat{\sigma}_2^2=\frac{\sum_{i=1}^{n_2}(X_{i2}-\hat{\mu}_2)}{n_2-1}$$

ESEMPIO

Nel confrontare due macchinari è risultato che quello di 1° tipo, su di un campione casuale di 40 pezzi ha funzionato in media 208 ore con un SQM di 26 ore; quello di 2° tipo, su di un campione di 50 pezzi è durato in media 192 ore con SQM di 22 ore.

Costruiamo una stima intervallare al 99% per la differenza tra le durate medie

$$16-2.576\sqrt{\frac{26^2}{40}+\frac{22^2}{50}} \le (\mu_1-\mu_2) \le 16+2.576\sqrt{\frac{26^2}{40}+\frac{22^2}{50}} \qquad \Rightarrow 2.72 \le (\mu_1-\mu_2) \le 29.28$$

Esercizio

Al fine di stimare la differenza nella vita media rispetto al vizio del fumo in una grande città industriale, sono stati scelti due campioni casuali: n_1=36 di persone che fumano ed n_2=44 di persone che non fumano, con i seguenti risultati:

	ampiezza	media	S.Q.M.
Non fumano	$n_1 = 36$	$\bar{x}_1 = 72$	$s_1 = 9$
Fumano	$n_2 = 44$	$\bar{x}_2 = 62$	$s_2 = 11$

Si determini un intervallo di confidenza al 90% per la differenza nella vita media tra le popolazioni:

Poiché: $\alpha = 0.1;\quad Z_{\frac{\alpha}{2}} = Z_{0.05} = 1.645$ **l'intervallo di confidenza è dato da:**

$$72-62-1.645\sqrt{\frac{9^2}{36}+\frac{11^2}{44}} \le (\mu_1-\mu_2) \le 72-62+1.645\sqrt{\frac{9^2}{36}+\frac{11^2}{44}} \Rightarrow 6.32 \le (\mu_1-\mu_2) \le 13.68$$

Possiamo perciò essere convinti che, al 90%, la vita media di chi non fuma supera quella di coloro che fumano per un margine che va dai 6.3 ai 13.7 anni

Intervalli per $(\mu_1-\mu_2)/3$

Campioni piccoli, popolazione normale, varianza comune, ma incognita

Non è possibile usare nessuno degli schemi precedenti

Se però le popolazioni sono normali ed hanno la stessa varianza gli intervalli possono ancora Essere costruiti.

Innanzitutto stimiamo la varianza delle popolazioni:

$$S_A^2=\frac{(n_1-1)\hat{\sigma}_1^2+(n_2-1)\hat{\sigma}_2^2}{n_1+n_2-2}$$

varianza campionaria aggregata

Si dimostra che uno stimatore non distorto della varianza della differenza tra due medie di due popolazioni normali con stessa varianza è

$$\hat{\sigma}^2(\hat{\mu}_1-\hat{\mu}_2)=S_A^2*\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$$

Continuazione

Nelle condizioni poste la differenza tra le due medie è distribuita secondo una t-student con gradi di libertà $r=n_1+n_2-2$

Ne consegue che l'intervallo di confidenza per la differenza tra le due medie è

$$(\hat{\mu}_1-\hat{\mu}_2)-t_{\frac{\alpha}{2},r}S_A*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le (\mu_1-\mu_2) \le (\hat{\mu}_1-\hat{\mu}_2)+t_{\frac{\alpha}{2},r}S_A*\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

ESEMPIO

Una speculatrice finanziaria valuta la differenza di rendimento tra obbligazioni statali e certe obbligazioni industriali. Scelti due campioni di ampiezza 12 da entrambe ottiene:

Tipo	Media	SQM	n
Statali	7.02	0.85	12
industriali	7.86	0.95	12

Si determini un intervallo di al 90% per la differenza tra le due medie

$$S_A^2=\frac{11*(0.95)^2+11*(0.85)^2}{12+12-2}=\frac{17.875}{22}=0.8125$$

$$r=12+12-2=22;\quad t_{\frac{0.1}{2},22}=1.717$$

$$(7.86-7.02)\pm 1.717*0.9014*\sqrt{\frac{2}{12}}$$

$$0.84\pm 0.63 \Rightarrow 0.21 \le (\mu_1-\mu_2) \le 1.47$$

Esercizio

Da due diverse popolazioni sono stati estratti due campioni: il primo di ampiezza 5 ed il secondo di ampiezza 4

{132, 145, 162, 166, 175} {137, 147, 158, 170};

Si determini un intervallo al 95% per la differenza tra le medie delle due popolazioni

In questo caso le varianze delle due popolazioni sono incognite e vanno stimate

Primo campione		Secondo campione	
×	$(x-156)^2$	×	$(x-153)^2$
132	576	137	256
145	121	147	36
162	36	158	25
166	100	170	289
175	361	612	606
780	1194	153	
156			

$\bar{x}_1 = 156; \quad \bar{x}_2 = 153; \qquad t_{0.025,7} = 2.365$

$$S_A = \frac{\sum_{i=1}^{4}(X_{i:} - \bar{x}_1)^2 + \sum_{i=1}^{5}(X_{i2} - \bar{x}_2)^2}{n_1 + n_2 - 2} = \frac{1194 + 606}{4 + 5 - 2} = 257$$

$$(\bar{x}_1 - \bar{x}_2) - t_{0.025,7} S_a \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le (\mu_1 - \mu_2) \le (\bar{x}_1 - \bar{x}_2) + t_{0.025,7} S_a \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(156 - 153) - 2.365\sqrt{257} * \sqrt{\frac{1}{5} + \frac{1}{4}} \le (\mu_1 - \mu_2) \le (156 - 153) - 2.365\sqrt{257} * \sqrt{\frac{1}{5} + \frac{1}{4}}$$

$$3 - 25.4 \le (\mu_1 - \mu_2) \le 3 + 25.4 \Rightarrow -22.4 \le (\mu_1 - \mu_2) \le 28.4$$

Intervalli per (μ_1-μ_2)/Variabili correlate

Gli intervalli precedenti presupponevano Cov(X_1,X_2)=0. Se I campioni sono indipendenti ciò è vero, ma spesso i campioni sono dipendenti e le variabili correlate *(matched pairs)*

Ufficio	*X*1	*X*2
1	200	220
2	150	170
3	55	65
4	350	360
5	250	240
6	120	140
7	85	95
8	20	40
9	195	180
$\hat{\mu}$	158	169

ESEMPIO

E' cambiata la responsabile di direzione. Occorre verificarne l'impatto sulle pratiche svolte nel mese prima e nel mese dopo.

E' accettabile l'ipotesi che le rilevazioni su uffici diversi siano indipendenti, ma non che lo siano quelle sullo stesso ufficio in tempi diversi.

$$E(X_1 - X_2) = \mu_1 - \mu_2$$

$$\sigma^2(X_1 - X_2) = \sigma^2(X_1) + \sigma^2(X_2) - 2Cov(X_1, X_2) = \sigma_D^2$$

In queste situazioni la correlazione è spesso positiva per cui:

$$\sigma_D^2 \le \sigma_1^2 + \sigma_2^2$$

Da notare che talvolta il campione è unico, cioè si rilevano due variabili sulle stesse unità e questo, da solo, può causare dipendenza e, verosimilmente, correlazione.

Variabili correlate/2

Per costruire gli intervalli usiamo la statistica pivotale: $\dfrac{(\mu_1 - \mu_2) - (\hat{\mu}_1 - \hat{\mu}_2)}{\hat{\sigma}_D / \sqrt{n}}$

che, se le v.c. sono normali, ha una distribuzione t-Student con "n-1" gradi di libertà:

$$(\hat{\mu}_1 - \hat{\mu}_2) - t_{\frac{\alpha}{2},n-1}\frac{\hat{\sigma}_D}{\sqrt{n}} \le (\mu_1 - \mu_2) \le (\hat{\mu}_1 - \hat{\mu}_2) + t_{\frac{\alpha}{2},n-1}\frac{\hat{\sigma}_D}{\sqrt{n}}$$

dove: $\hat{\sigma}_D = \sqrt{\hat{\sigma}_1 + \hat{\sigma}_2 - 2Cov(X_1, X_2)} = \sqrt{\dfrac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}}; \; d_i = X_{1i} - X_{2i}$

ESEMPIO

$$\hat{\sigma}_1 = 103.05; \quad \hat{\sigma}_2 = 98.75; \quad Cov(X_1, X_2) = 10095.83$$

n=9

$$-11 - 3.355 * \left(\frac{13.33}{3}\right) \le (\mu_1 - \mu_2) \le -11 + 3.355 * \left(\frac{13.33}{3}\right)$$

α=1%

$$-25.911 \le (\mu_1 - \mu_2) \le 3.911$$

Poiché i due estremi hanno segno diverso, non c'è una differenza significativa tra i due periodi nel senso che le pratiche evase NON sono significativamente aumentate.

Esercizio

Studio della concentrazione di beta-endorfine in alcuni atleti prima e dopo una maratona. Stima della differenza tra medie con 1-a=0.01

Atleta	Prima	Dopo	Differenza
1	4.3	29.6	-25.3
2	4.6	25.1	-20.5
3	5.2	15.5	-10.3
4	5.2	29.6	-24.4
5	6.6	24.1	-17.5
6	7.2	37.8	-30.6
7	8.4	20.2	-11.8
8	9.0	21.9	-12.9

$\hat{\mu}_1 - \hat{\mu}_2 = -19.2$

$\hat{\sigma}_D = 7.294$

$$-19.2 - 3.5\frac{7.294}{\sqrt{8}} \le \mu_1 - \mu_2 \le -19.2 + 3.5\frac{7.294}{\sqrt{8}} \Rightarrow -28.226 \le \mu_1 - \mu_2 \le -10.174$$

Gli estremi sono entrambi negativi per cui la differenza tra le medi deve essere considerata significativa.

Intervalli per la proporzione

Sia "H" la proporzione nel campione e "π" la proporzione nella popolazione.

Supponiamo che il numero di prove sia "elevato" al punto da far scattare il T.L.C.

$$P\left(-Z_{\frac{\alpha}{2}} \le \frac{H-\pi}{\sqrt{\frac{\pi*(1-\pi)}{n}}} \le Z_{\frac{\alpha}{2}}\right) \cong 1-\alpha$$

Quindi, l'intervallo di confidenza per "π" è: $H - Z_{\frac{\alpha}{2}}\hat{\sigma}_H \le \pi \le H + Z_{\frac{\alpha}{2}}\hat{\sigma}_H$

I limiti però dipendono "π".

Per superare l'ostacolo si usa la stima H ottenendo: $\hat{\sigma}_H = \sqrt{\frac{H*(1-H)}{n}}$

La stima intervallare per la proporzione diviene:

$$H - Z_{\frac{\alpha}{2}}\sqrt{\frac{H(1-H)}{n}} \le \pi \le H + Z_{\frac{\alpha}{2}}\sqrt{\frac{H(1-H)}{n}}$$

Esempio

In un campione casuale di n=200 persone si sono trovate 36 (H=0.18) d'accordo sull'università "a distanza". Determinare un intervallo al 95% per la proporzione

$$\pi = H \pm Z_{\frac{\alpha}{2}}\hat{\sigma}_H \Rightarrow \pi = 0.18 \pm Z_{\frac{0.05}{2}} * \sqrt{\frac{0.18*0.82}{200}} = 0.18 \pm 0.065$$

$$0.18 - 1.96*\sqrt{0.000738} \le \pi \le 0.18 + 1.96*\sqrt{0.000738}$$

$$0.127 \le \pi \le 0.233$$

Da notare che questo tipo di stima intervallare può essere usato solo se

$$n*H \ge 5 \quad e \quad n*(1-H) \ge 5$$

che in questo caso sono rispettate.

Esercizio

Prima di un ballottagio, l'aspirante sindaco, dott.ssa Domenica Fuoco, ha commissionato un sondaggio che la vede vincente per 59 voti su 112.

La percentuale, pur favorevole (52.7%), non è rassicurante. Si determini un intervallo al 95% per la reale percentuale di votanti che si schierano a favore della dott.ssa.

$$H - Z_{\frac{\alpha}{2}}\sqrt{\frac{H*(1-H)}{n}} \le \pi \le H + Z_{\frac{\alpha}{2}}\sqrt{\frac{H*(1-H)}{n}} \Rightarrow$$

$$0.527 - 1.96\sqrt{\frac{0.527*0.473}{112}} \le \pi \le 0.527 + 1.96\sqrt{\frac{0.527*0.473}{112}}$$

$$0.435 \le \pi \le 0.619$$

C'è una discreta possibilità che la percentuale vera sia inferiore al 50% per cui è bene che la dott.ssa Fuoco si dia da fare se non vuol perdere le elezioni.

Intervalli per la proporzione/2

Anche per la stima intervallare della proporzione è possibile dare stime unidirezionali

In particolare per un livello di confidenza pari 1-α

- **Unidirezionale inferiore** $\pi \ge H - Z_\alpha\sqrt{\frac{H(1-H)}{n}}$
- **Unidirezionale superiore** $\pi \le H + Z_\alpha\sqrt{\frac{H(1-H)}{n}}$

Esempio

Si desidera conoscere la percentuale di item difettosi in una linea di produzione di microprocessori che comunque non potrà mai superare il limite di garanzia del 5%.

Su 800 pezzi erano difettosi 27. Costruire un intervallo di confidenza unilaterale inferiore al 98% per i pezzi difettosi

$$H = \frac{27}{800} - 2.0537*\sqrt{\frac{\frac{27}{800}*\frac{773}{800}}{800}} = 2.06\% \qquad 2.06\% \le \pi \le 5\%$$

Esercizio

Una società di collocazione pubblicitaria segue un campione di 1000 famiglie che usa per stimare la proporzione di soggetti che guarda un dato programma.

Sul VideoTre, alle 21:00 del sabato è in programma il documentario:
"La Statistica. Ieri, oggi e domani".

Secondo i meter collegati ai televisori risultano sintonizzate sul programma 260 famiglie delle 1000 inserite nel panel.

1) Determinare una stima intervallare del tipo 0<π<K al 99.9%

$$0.999 = P\left(\frac{0.26-\pi}{\sqrt{\frac{0.26*0.74}{1000}}} \le 3.0902\right) = P(\pi \le 0.26 + 3.0902*0.0139) = P(\pi \le 0.3029) \Rightarrow 0 \le \pi \le 30.3\%$$

2) Determinare una stima intervallare del tipo 1>π>K al 95%

$$0.05 = P\left(\frac{0.26-\pi}{\sqrt{\frac{0.26*0.74}{1000}}} \ge 1.6449\right) = P(\pi \ge 0.26 + 1.6449*0.0139) = P(\pi \ge 0.2829) \Rightarrow 28.3\% \le \pi \le 100\%$$

Esempio

Margherita Biafora gestisce una agenzia di servizi per anziani. Al fine di stimare la Proporzione delle famiglie che prediligono un'assistente donna decide di interrogare in proposito 13 delle 76 famiglie che ha nella *mailing list*

Le telefonate danno la seguente proporzione: H=0.61

Se la Biafora decide per un intervallo al 99% otterrà:

$$\left[\ H - Z_{0.005}\sqrt{\frac{0.61*0.39}{13}*\frac{(76-13)}{(76-1)}},\ H + Z_{0.005}\sqrt{\frac{0.61*0.39}{13}*\frac{(76-13)}{(76-1)}}\ \right]$$

$$[\ 0.61 - 2.576*0.124,\ 0.61 + 2.576*0.124\ ] \Rightarrow [0.29, 0.93]$$

Che è chiaramente insoddisfacente. Per migliorarlo occorre ridurre l'affidabilità cioè Aumentare "α" oppure aumentare il numero di famiglie interrogate.

Modifiche per popolazioni finite

Stabiliamo una procedura per calcolare degli intervalli di confidenza quando il campione si forma da una popolazione finita di N unità.

In tale contesto la frazione campionaria di successi ha le seguenti proprietà

$$E(H) = \pi; \quad \sigma^2(H) = \frac{\pi(1-\pi)}{n} * \left(\frac{N-n}{N-1}\right)$$

e gli intervalli di confidenza per π sarà dato da:

$$\left[\ H - Z_{\frac{\alpha}{2}}\sqrt{\frac{H*(1-H)}{n}*\frac{(N-n)}{(N-1)}},\ H + Z_{\frac{\alpha}{2}}\sqrt{\frac{H*(1-H)}{n}*\frac{(N-n)}{(N-1)}}\ \right]$$

che si basa sulla distribuzione normale, ma che ora è meno plausibile

Differenza tra due proporzioni

L'interesse è centrato sulla differenza tra due percentuali: ad esempio la proporzione Di item difettosi du due diverse linee di produzione

Sia

- **X_1 il numero di successi nella prima serie di n_1 prove indipendenti**
- **X_2 il numero di successi nella seconda serie di n_2 prove indipendenti (tra di loro e da quelle della prima serie)**

$$H_1 = \frac{X_1}{n_1};\ \ H_2 = \frac{X_2}{n_2};\ \ (H_1 - H_2)\ \ stimatore\ \ di\ \ (\pi_1 - \pi_2)$$

Ipotesi di indipendenza

$$E(H_1 - H_2) = \pi_1 - \pi_2$$

$$\sigma^2(H_1 - H_2) = \sigma^2(H_1) + \sigma^2(H_2) - 2Cov(H_1, H_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2} - 2*0$$

Lo stimatore è centrato e consistente.

Differenza tra due proporzioni/2

Per n_1 ed n_2 grandi diventa applicabile il T.L.C. per cui la stima intervallare sarà

$$(H_1 - H_2) - Z_{\frac{\alpha}{2}}\sqrt{\frac{H_1*(1-H_1)}{n_1} + \frac{H_2*(1-H_2)}{n_2}} \le (\pi_1 - \pi_2) \le (H_1 - H_2) + Z_{\frac{\alpha}{2}}\sqrt{\frac{H_1*(1-H_1)}{n_1} + \frac{H_2*(1-H_2)}{n_2}}$$

Che è valido purché $n_1H_1, n_2H_2, n_1*(1-H_1), n_2*(1-H_2) \ge 5$

ESEMPIO

Prima del cambiamento del limite di velocità, su 100 auto scelte a caso 49 violarono il imite.
Dopo il cambiamento su un altro campione di 100 auto le violazioni furono 19.
Costruiamo un intervallo al 99% per la differenza tra le due proporzioni

$$0.30 - 2.576\sqrt{\frac{0.49*0.51}{100} + \frac{0.19*0.81}{100}} \le (\pi_1 - \pi_2) \le 0.30 + 2.576\sqrt{\frac{0.49*0.51}{100} + \frac{0.19*0.81}{100}}$$

$$0.30 - 0.16 \le (\pi_1 - \pi_2) \le 0.30 + 0.16 \Rightarrow 0.14 \le (\pi_1 - \pi_2) \le 0.46$$

La differenza è significativa dato che entrambi gli estremi hanno segno positivo

Esercizio

In un ente pubblico non economico sono state riscontrate 507 assenze dal lavoro per motivi di salute su un totale di 2698 dipendenti.

L'anno successivo le assenze furono 639 su 2724 dipendenti.

Ipotizzando che l'organico complessivo sia sostanzialmente rimasto immutato calcolare una stima intervallare al 99% per la differenza tra le due proporzioni.

$$\left(\frac{507}{2698} - \frac{639}{2724}\right) - 2.756\sqrt{\frac{\frac{507}{2698}*\frac{2191}{2698}}{2698} + \frac{\frac{639}{2724}*\frac{2085}{2724}}{2724}} \le (\pi_1 - \pi_2) \le \left(\frac{507}{2698} - \frac{639}{2724}\right) + 2.576\sqrt{\frac{\frac{507}{2698}*\frac{2191}{2698}}{2698} + \frac{\frac{639}{2724}*\frac{2085}{2724}}{2724}}$$

$$0.188 - 2.756*0.0111 \le (\pi_1 - \pi_2) \le 0.188 - 2.756*0.0111$$

$$15.7\% \le (\pi_1 - \pi_2) \le 21.9\%$$

Calcolo dell'ampiezza

La determinazione di "n" è essenziale in quanto se è troppo piccolo le stime possono risultare poco precise; se è troppo grande implicherebbe costi maggiori

Fissato il livello di confidenza si può stabilire l'n appropriato. L'errore è di definito come scarto assoluto tra stima e parametro da stimare

$$Errore = E = |\bar{x} - \mu| \quad \textit{(questo è un limite massimo)}$$

Può anche essere definito in modo proporzionale, ma lo schema non cambia

$$\gamma = \left|\frac{\bar{x} - \mu}{\mu}\right| \Rightarrow |\bar{x} - \mu| = \gamma\mu = E$$

Poiché l'errore è pari alla metà della lunghezza dell'intervallo otteniamo:

$$n(c.r.) = \left(\frac{Z_{\frac{\alpha}{2}}\sigma}{E}\right)^2 \quad ; \quad n(s.r.) = \frac{\left(\frac{Z_{\frac{\alpha}{2}}\sigma}{E}\right)^2 * N^2}{\left(\frac{Z_{\frac{\alpha}{2}}\sigma}{E}\right)^2 + (N-1)}$$

Esempio

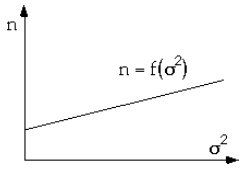
La direttrice dell'ufficio selezione personale intende assumere solo coloro che abbiano maturato una certa esperienza.

D'altra parte, non vuole correrre il rischio di eliminare, in base ad una soglia soggettiva, la maggior parte delle candidature.

Volendo stimare l'esperienza media con un errore di non più di 6 mesi, quante persone deve intervistare per ottenerne una stima intervallare al 99%?

Supponendo che la statistica: media campionaria abbia distribuzione normale possiamo porre

$$n = \frac{\left(\frac{Z_{\frac{0.01}{2}}*\sigma}{6^2}\right)^2}{6^2} = \frac{(2.576\sigma)^2}{36} = 0.1844\sigma^2$$



L'ampiezza dipende ora dalla varianza.
Se questa aumenta aumenta anche l'ampiezza.

Se la direttrice ritiene che $\sigma^2 = 144 \Rightarrow n \cong 27$

Calcolo dell'ampiezza/2

Se la popolazione è normale, ma σ è incognito si dovrebbe usare la Student. Lo scarto ammesso ora si esprimerebbe come:

$$|\bar{x} - \mu| = t_{\frac{\alpha}{2};n-1} \frac{s}{\sqrt{n}}$$

il che comporterebbe la seguente formula:

$$n = \left(\frac{s * t_{\frac{\alpha}{2};n-1}}{\bar{x} - \mu} \right)^2$$

che è però inutilizzabile: occorre la stima "s", ma il campione non è stato ancora ottenuto e poi "n"compare sia a sinistra che a destra dell'equazione.

Trovare "n" è perciò problematico. In genere si usano esperienze passate analoghe, indagini pilota o qualche stima intutitiva basata sulla relazione tra *range* e "σ".

$$Normale: \ \hat{\sigma} = \frac{X_{\max} - X_{\min}}{4}; \quad Uniforme: \ \hat{\sigma} = \frac{X_{\max} - X_{\min}}{3.4641}$$

Esempio

Si vuole introdurre un metodo di insegnamento della statistica su cassette.

Per accertare la validità del nuovo metodo si sottoporranno a controllo un gruppo di "n" studenti da scegliere casualmente dalla popolazione di iscritti.

L'ampiezza "n" deve essere tale che il voto medio del campione sia diverso da quello vero al più di due voti con un livello di confidenza del 90% (α=10%)

$$Errore = |\bar{x} - \mu| \le 2$$

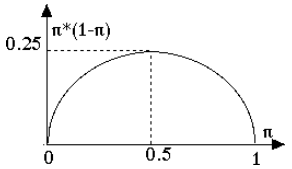
Poiché si tratta di voti possiamo stimare lo scarto con $\hat{\sigma} = \frac{30 - 17}{4} = \frac{13}{4} = 3.25$

A questo punto determiniamo l'ampiezza: $n \cong \left(\frac{Z_{\frac{\alpha}{2}} \sigma}{|E|} \right)^2 \Rightarrow \left(\frac{1.645 * 3.25}{2} \right)^2 = 7.15 \Rightarrow n = 7$

Ampiezza campionaria per le proporzioni

Si precisa il livello di confindenza: (1-α) e si stabilisce l'errore massimo accettabile

$$|H - \pi| = Z_{\frac{\alpha}{2}} * \sqrt{\frac{\pi * (1 - \pi)}{n}} \qquad \Rightarrow n = \left(\frac{Z_{\frac{\alpha}{2}}}{E} \right)^2 * H(1 - H)$$



"π" non è noto, però sappiamo che π*(1-π)≤1/4, cioé

$$\sqrt{\frac{\pi * (1 - \pi)}{n}} \le \sqrt{\frac{\frac{1}{2} * \frac{1}{2}}{n}} = \frac{1}{2\sqrt{n}}$$

Quindi, se si vuole costruire un intervallo di confidenza all' (1-α)% per "π" tale che la distanza sia pari, al più, ad un errore prefissato si calcola "n" in base alle seguenti formule

$$\Rightarrow n(c.r.) = \left(\frac{Z_{\frac{\alpha}{2}}}{2E} \right)^2; \quad n(s.r.) = \frac{\left(\frac{Z_{\frac{\alpha}{2}}}{2E} \right)^2 N}{\left(\frac{Z_{\frac{\alpha}{2}}}{2E} \right)^2 + (N - 1)}$$

Il presupposto è che la popolazione sia molto grande rispetto all'ampiezza del campione

Esempio

Prima di un ballottaggio, l'aspirante sindaco, dott.ssa Domenica Fuoco, ha commissionato un sondaggio che la vede vincente per 59 voti su 112.

La percentuale, pur favorevole (52.7%), non è rassicurante. Si determini un intervallo al 95% per la reale percentuale di votanti che si schierano a favore della dott.ssa.

$$H - Z_{\frac{\alpha}{2}} \sqrt{\frac{H * (1 - H)}{n}} \le \pi \le H + Z_{\frac{\alpha}{2}} \sqrt{\frac{H * (1 - H)}{n}} \Rightarrow$$

$$0.527 - 1.96 \sqrt{\frac{0.527 * 0.473}{112}} \le \pi \le 0.527 + 1.96 \sqrt{\frac{0.527 * 0.473}{112}}$$

$$0.435 \le \pi \le 0.619$$

C'è una discreta possibilità che la percentuale vera sia inferiore al 50% per cui è bene che la dott.ssa Fuoco si dia da fare se non vuol perdere le elezioni.

Esempio

L'assieditrice all'urbanistica, Ing. Lucia Berardi, intende modificare la destinazione di alcune aree incolte relativamente vicine al centro città.

Prima di portare la proposta in giunta attiva una indagine conoscitiva sul gradimento della modifica da parte dei cittadini.

Se l'errore che consente è del 3% al massimo, qual"'è l'ampiezza del campione che porta ad un intervallo di confidenza al 99% per la proporzione vera di favorevoli?

$$n(c.r.) = \left(\frac{2.5758}{2*0.03}\right)^2 \cong 1843; \quad n(s.r.) = \frac{\left(\frac{2.5758}{2*0.03}\right)^2 * 350}{\left(\frac{2.5758}{2*0.03}\right)^2 + 349} \cong 295$$

Per il campionamento s.r. si è ipotizzato una popolazione di N=350

Una tavola per l'ampiezza

L'impostazione precedente rischia di richiedere campioni di ampiezza più grande di quanto non impongano il richiesto livello di confidenza e/o il margine di errore.

La seguente tabella mostra le ampiezze campionarie necessarie per un margine di errore del {3%, 4%, 5%} e per i livelli di confidenza del 90%, 95% e 99%.

Valore vero di π	Margine di errore (±)	Livello di confidenza 90%	95%	99%
0.1 (od 0.9)	3%	271	385	664
"	4%	153	217	373
"	5%	98	139	239
0.3 (od 0.7)	3%	632	897	1548
"	4%	356	505	871
"	5%	228	323	557
0.5	3%	758	1068	1842
"	4%	423	601	1037
"	5%	271	385	664

Da notare che, come per la media, l'ampiezza si riduce in ragione di √n quando "n" aumenta

La tavola va usata in alternativa alla stima della varianza della frazione campionaria che impone di fatto che π=1/2.

Esempio

Sono molti i giochi di società che si basano sull'uso di dadi, ad esempio il Risiko. Per sfizio Mariella vuole alterare uno dei dadi in modo da aumentare la probabilità di ottenere il "6".

Per controllare se la sua operazione ha avuto successo, Mariella lancia il dado "truccato" 600 volte ottenendo il "6" ben 129 volte.

a) Stimate "π" cioè la probabilità di avere il "6" con il nuovo dado;

b) Pensate che la sua operazione abbia avuto successo?

---

a) $H = \frac{129}{600} = 21.5\%$

b) Se il dado non fosse truccato, ci si aspetterebbe di ottenere il "6" circa 100 volte: (600/6). Invece lo si ha 129 volte.

Certo, sappiamo che non tutti i 600 lanci di un dado danno sempre il "6" in 100 casi, ma 29 volte in più sono tante e l'operazione sembra riuscita

Non può però escludere che qualcuno dei suoi amici non si insospettisca e avanzi seri dubbi sui motivi affettivi e scaramantici della sua preferenza per quel dado

Ampiezza campionaria per la varianza

Ipotizziamo che il modello base sia gaussiano e che "n" sia grande perché la chi quadro sia ben approssimata dalla normale

$$\frac{s^2 - \sigma^2}{\sqrt{\frac{2\sigma^4}{n-1}}} \to N(0,1)$$

Fissato l'errore E=|s² - σ² | ed il livello di confidenza (1-α) avremo $n = 1 + 2\left[\frac{Z_{\frac{\alpha}{2}}\sigma^2}{E}\right]^2$

ESEMPIO

Supponiamo di disporre di un valore preliminare σ=5. Poniamo inoltre E=2, α=0.04

$$n = 1 + 2\left[\frac{2.0537 * 25\sigma^2}{2}\right]^2 = 1320$$

Che è molto maggiore dell'n=27 richiesto per stimare la media

Ampiezza campionaria per π_1-π_2

Occorrono due ipotesi semplificatrici

$$\pi_1 = \pi_2 = 0.5;$$
$$n_2 = cn_1; \quad c \ \ razionale$$

Fissato l'errore E=lπ_1 - π_2 l ed il livello (1-α) avremo $n_1 = \left(\frac{c+1}{c}\right)\frac{Z^2_{\frac{\alpha}{2}}}{4E^2}$

ESEMPIO

Si deve costruire un intervallo al 90% per la differenza tra due proporzioni. Si esige un un errore E inferiore o uguale a 0.05. Tenuto conto che n2=2n1 quanto deve essere la ampiezza complessiva del campione?

$$n_1 = \frac{3}{2}\left[\frac{1.6449}{2*0.05}\right]^2 = 406 \Rightarrow n_2 = 812 \Rightarrow n = 1218$$

Data l'ipotesi conservativa di π_1=π_2=0.5 questo "n" è -molto probabilmente- più alto del necessario