Inferenza statistica

Inferenza statistica

Spesso l'informazione a disposizione deriva da un'**osservazione parziale** del fenomeno studiato.

In questo caso lo studio di un fenomeno mira solitamente a trarre, sulla base di ciò che si è osservato, **considerazioni di carattere generale**.

Per sua natura il processo di inferenza è soggetto ad **errore**, che può essere tenuto sotto controllo, o almeno quantificato, mediante criteri e tecniche di tipo statistico.

Fasi del processo di inferenza

- definizione del problema
- individuazione di un opportuno modello teorico
- estrazione del campione
- raccolta e analisi dei dati
- generalizzazione

Il modello

Generalmente descriviamo la distribuzione di un fenomeno mediante una opportuna **distribuzione di probabilità**.

La forma (il tipo) della **distribuzione** è assunta **nota**, mentre sono considerati **incogniti** i **parametri** della distribuzione.

In questo schema logico, i parametri (costanti caratteristiche del fenomeno studiato) sono l'oggetto di interesse del processo di inferenza (inferenza parametrica).

Il campionamento

Come selezionare il campione da osservare?

Possiamo distinguere:

- **campionamento ragionato**: il campione è scelto ad hoc in quanto rappresentativo della popolazione
- **campionamento casuale**: il campione è estratto mediante procedimenti di selezione casuale

Le origini delle indagini campionarie

1936: Elezioni presidenziali USA
Candidati: F.D. Roosvelt e A. Landon

- Indagine Literary Digest: 10 milioni di fac-simile di schede elettorali inviate a nominativi estratti dagli elenchi telefonici e dai registri automobilistici
- Risultato previsto: Roosvelt 41 % e Landon 59 %

- Indagine Gallup: alcune migliaia di interviste ad elettori estratti casualmente dall'intera popolazione
- Risultato previsto: Roosvelt 60 % e Landon 40 %

Le origini delle indagini campionarie

Risultato delle elezioni: Roosvelt 61 %

Gli errori del Literary Digest:

- ERRORE DI COPERTURA: le liste usate non erano complete gli elenchi usati non erano rappresentativi dell'intera popolazione ma solo dei ceti pi abbienti che tendevano a votare repubblicano
- AUTOSELEZIONE del CAMPIONE: Le caratteristiche socio-demografiche dei cittadini che risposero al sondaggio erano presumibilmente diverse da quelle di chi non rispose (istruzione, reddito, etc.)

Il campione ragionato

Il ricercatore cerca di costruire una buona **"immagine"** della popolazione sulla base di caratteristiche note e spera che il campione sia rappresentativo anche per le variabili oggetto di studio.

È usato molto di frequente per i sondaggi, rarissimamente (mai) in ambito sperimentale.

È uno strumento potente ma molto delicato (il rischio di introdurre distorsioni è elevato), inoltre è difficile quantificare l'errore.

Il campione casuale

Il campionamento dovrebbe essere sempre casuale, cioè ogni campione dovrebbe avere la stessa probabilità di essere scelto che hanno tutti gli altri possibili campioni della popolazione.

Soddisfare questo criterio di scelta equivale a fare una **estrazione probabilistica** (ovvero "casuale") del campione, che teoricamente si può realizzare nei modi seguenti

- **popolazione finita**: il campione viene estratto mediante etichettatura e sorteggio;
- **popolazione infinita**: le osservazioni campionarie (dati) derivano dalla ripetizione dell'esperimento casuale nelle medesime condizioni (esempio: lancio di una moneta ripetuto infinite volte)

NB: "casuale" o "a caso" non significa "a casaccio"

Matematicamente...

Definizione

Un campione casuale semplice di dimensione (o numerosità) N è una N-upla di v.c. $X_1, \ldots, X_N$ (i cui valori sono detti osservazioni o determinazioni campionarie o dati)

- *indipendenti*, cioè tali che per ogni scelta di intervalli $I_1, \ldots, I_N$ si ha

$$P(X_1 \in I_1, \ldots, X_N \in I_N) = P(X_1 \in I_1) \cdots P(X_N \in I_N),$$

- *identicamente distribuite*, cioè

$$X_i \sim X, \quad i = 1, \ldots, N$$

dove X è una distribuzione adottata come modello per la popolazione.

Sintesi dell'informazione campionaria

L'informazione campionaria può essere sintetizzata mediante gli indici sintetici già visti in statistica descrittiva. In particolare, possiamo definire:

- $m_N = \frac{1}{N}\sum_{i=1}^{N} X_i$ media campionaria
- $S_N^2 = \frac{1}{N-1}\sum_{i=1}^{N} |X_i - m_N|^2 = \frac{N}{N-1}\sigma^2$ varianza campionaria corretta

Entrambe le quantità sono v.c., e variano (cioè assumono valori) nell'universo dei campioni da cui selezioniamo in modo casuale.

Distribuzioni campionarie

Indicando con μ e σ^2 la media e la varianza della popolazione, cioè di X, è possibile dimostrare che

1. $\boxed{E[m_N] = \mu}$ cioè la media campionaria è uno stimatore non distorto della media della popolazione
2. $\boxed{\mathrm{Var}[m_N] = \sigma^2/N}$ (importantissimo per N grande!)
3. $\boxed{E[S_N^2] = \sigma^2}$ cioè la varianza campionaria corretta è uno stimatore non distorto della varianza della popolazione

Distribuzioni campionarie

Indicando con μ e σ^2 la media e la varianza della popolazione, cioè di X, è possibile dimostrare che

1. $\boxed{E[m_N] = \mu}$ cioè la media campionaria è uno stimatore non distorto della media della popolazione
2. $\boxed{\mathrm{Var}[m_N] = \sigma^2/N}$ (importantissimo per N grande!)
3. $\boxed{E[S_N^2] = \sigma^2}$ cioè la varianza campionaria corretta è uno stimatore non distorto della varianza della popolazione

Ricordiamo che μ e σ^2 sono incognite da determinarsi.
Dato un campione X_1, ..., X_N potremmo pensare di ottenerle da 1 e 3 calcolando i valori attesi. Purtroppo per fare questo calcolo occorrerebbe conoscere la distribuzione delle v.c. X_1, ..., X_N che è incognita al pari di quella della popolazione.

Distribuzioni campionarie

Indicando con μ e σ^2 la media e la varianza della popolazione, cioè di X, è possibile dimostrare che

1. $\boxed{E[m_N] = \mu}$ cioè la media campionaria è uno stimatore non distorto della media della popolazione
2. $\boxed{\mathrm{Var}[m_N] = \sigma^2/N}$ (importantissimo per N grande!)
3. $\boxed{E[S_N^2] = \sigma^2}$ cioè la varianza campionaria corretta è uno stimatore non distorto della varianza della popolazione

Ma la 2 dice che se il campione è sufficientemente grande allora $\mathrm{Var}[m_N]$ è molto piccola, cioè è piccola la probabilità che m_N si discosti molto dal proprio valore atteso $E[m_N] = \mu$

In pratica, se il campione è molto numeroso allora la media della popolazione μ si può stimare con la media campionaria m_N.

Esercizi

Da 12.6 a 12.9 del testo consigliato

Esempio - lancio di una moneta ripetuto infinite volte

Schematizziamo l'esperimento con una **successione** $(X_i)_{i\in\mathbb{N}}$ di v.c. che valgono 1 se esce testa e 0 se esce croce. Supponiamo che la moneta non sia truccata.

Nel caso del singolo lancio si ha

$$P(X_i = k) = \frac{1}{2}, \qquad k = 0, 1$$

quindi le X_i sono equidistribuite con distribuzione binomiale uniforme, e si ha

$$E(X_i) = \sum_{k=0}^{1} kP(X_i = k) = \frac{1}{2}$$

$$\mathrm{Var}(X_i) = \sum_{k=0}^{1}(k - \frac{1}{2})^2 P(X_i = k) = \frac{1}{2}\sum_{k=0}^{1}\frac{(2k-1)^2}{4} = \frac{1}{4}$$

Lancio ripetuto di una moneta

Un campione di dim. $N = 1$, è fatto da una sola v.c. X_1.

Si ha

$$m_1 = X_1, \quad E[m_1] = \frac{1}{2} = \mu$$

ma m_N non assume mai il valore $\mu = \frac{1}{2}$.

Lancio ripetuto di una moneta

Un campione di dim. $N = 2$, è costituito da 2 v.c. X_1 e X_2.
Si ha

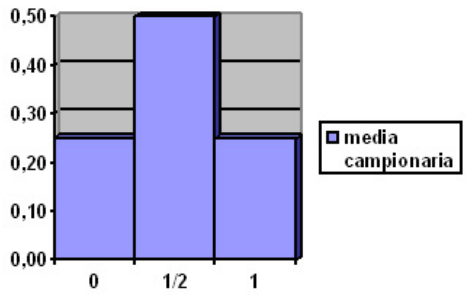
$$m_2 = \frac{X_1 + X_2}{2}, \quad E[m_2] = \frac{1}{2} = \mu$$

La distribuzione di probabilità di m_2, cioè

$$f(x) = P(m_2 = x) = P(\frac{X_1 + X_2}{2} = x)$$

risulta

x	coppie a media x	$f(x)$
0	(0, 0)	$f(0) = \frac{1}{4}$
$\frac{1}{2}$	(1, 0) (0, 1)	$f(\frac{1}{2}) = \frac{1}{2}$
1	(1, 1)	$f(1) = \frac{1}{4}$
	Tot. 4	



Si nota che

- la distribuzione della media non è più uniforme
- il risultato maggiormente probabile corrisponde esattamente alla media $\mu = 1/2$
- con un semplice calcolo $\mathrm{Var}[m_2] = 1/8$

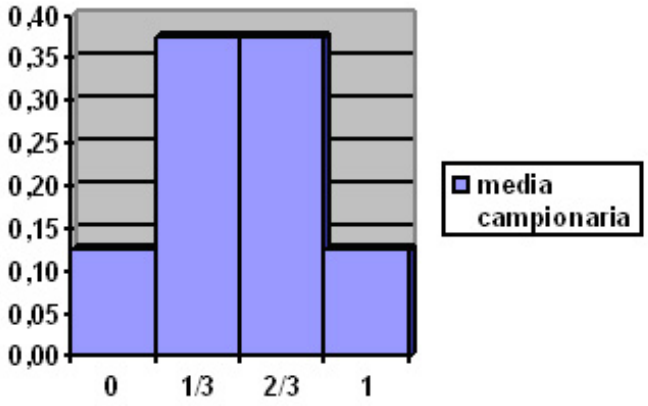
Lancio ripetuto di una moneta

Campione di dim. $N = 3$: X_1, X_2, X_3.

$$m_3 = \frac{X_1 + X_2 + X_3}{3}, \quad E[m_3] = \frac{1}{2} = \mu$$

Distribuzione di m_3: $f(x) = P(\frac{X_1+X_2+X_3}{3} = x)$

x	terne a media x	$f(x)$
0	(0, 0, 0)	$f(0) = \frac{1}{8}$
$\frac{1}{3}$	(1, 0, 0) (0, 1, 0) (0, 0, 1)	$f(\frac{1}{3}) = \frac{3}{8}$
$\frac{2}{3}$	(1, 1, 0) (1, 0, 1) (0, 1, 1)	$f(\frac{2}{3}) = \frac{3}{8}$
1	(1, 1, 1)	$f(1) = \frac{1}{8}$
	Tot. 8	



Osserviamo che $\mathrm{Var}[m_3] = 1/12$, quindi è diminuita la dispersione.

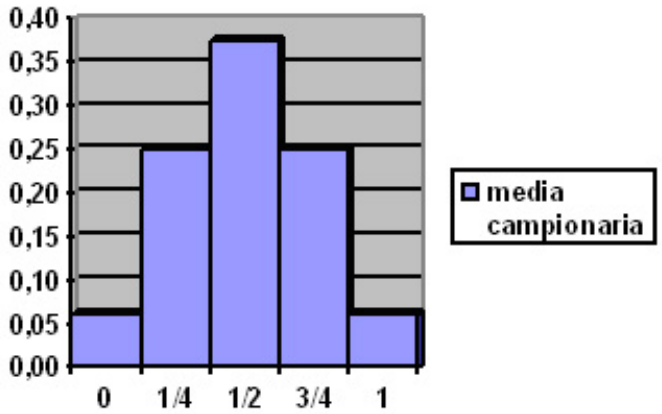
Lancio ripetuto di una moneta

Campione di dim. $N = 4$: X_1, X_2, X_3 e X_4.

$$m_4 = \frac{X_1 + X_2 + X_3 + X_4}{4}, \quad E[m_4] = \frac{1}{2} = \mu$$

Distribuzione di m_4:

x	quaterne a media x	$f(x)$
0	$\binom{4}{0} = 1$	$f(0) = \frac{1}{16}$
$\frac{1}{4}$	$\binom{4}{1} = 4$	$f(\frac{1}{4}) = \frac{1}{4}$
$\frac{1}{2}$	$\binom{4}{2} = 6$	$f(\frac{1}{2}) = \frac{3}{8}$
$\frac{3}{4}$	$\binom{4}{3} = 4$	$f(\frac{3}{4}) = \frac{1}{4}$
1	$\binom{4}{4} = 1$	$f(1) = \frac{1}{16}$
	Tot. 16	



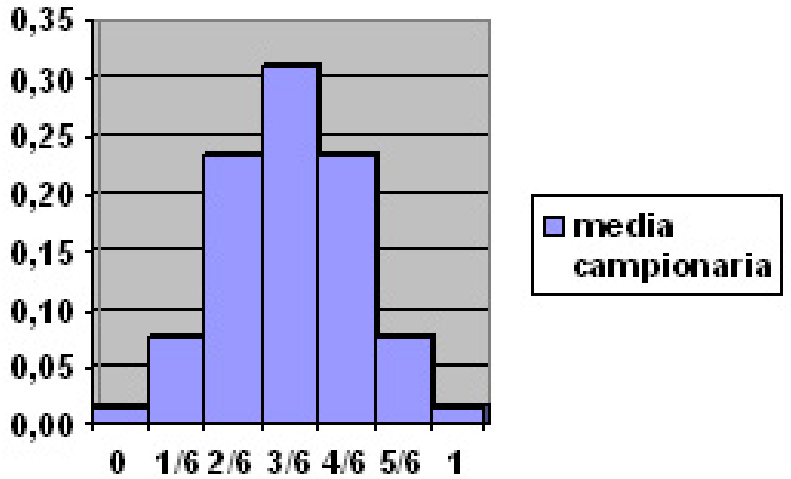
Osserviamo che

- la media camp. più probabile è la media $\mu = 1/2$
- $\mathrm{Var}[m_4] = 1/16$, quindi, come si nota anche dal grafico, la dispersione è ulteriormente diminuita
- la distribuzione comincia ad avere un andamento a campana.

Lancio ripetuto di una moneta

Il caso $N = 6$:

x	6-uple a media x	$f(x)$
0	$\binom{6}{0} = 1$	$f(0) = \frac{1}{2^6}$
$\frac{1}{6}$	$\binom{6}{1} = 6$	$f(\frac{1}{6}) = \frac{6}{2^6}$
$\frac{2}{6}$	$\binom{6}{2} = 15$	$f(\frac{2}{6}) = \frac{15}{2^6}$
$\frac{3}{6}$	$\binom{6}{3} = 20$	$f(\frac{3}{6}) = \frac{20}{2^6}$
$\frac{4}{6}$	$\binom{6}{4} = 15$	$f(\frac{4}{6}) = \frac{15}{2^6}$
$\frac{5}{6}$	$\binom{6}{5} = 6$	$f(\frac{5}{6}) = \frac{6}{2^6}$
1	$\binom{6}{6} = 1$	$f(1) = \frac{1}{2^6}$
	Tot. 2^6	



La tendenza ad assumere una forma a campana si accentua sempre di più al crescere di N. Si ha $\mathrm{Var}[m_N] = \frac{\sigma^2}{N} = \frac{1}{4N}$ quindi la campana diventa sempre più stretta.

Ciò significa che al crescere della dimensione del campione aumenta sempre più la probabilità che la media campionaria sia vicina ad $1/2$, cioè che testa e croce escano lo stesso numero di volte.

Esempio - lancio ripetuto di un dado

Indichiamo con X_i la variabile il cui valore coincide col numero uscito nel lancio i-esimo. Nel caso del singolo lancio si ha

$$P(X_i = k) = \frac{1}{6}, \qquad k = 1, ..., 6$$

quindi le X_i sono equidistribuite con distribuzione discreta uniforme, e si ha

$$E(X_i) = \sum_{k=1}^{6} kP(X_i = k) = \frac{1}{6}\sum_{k=1}^{6} k = \frac{1}{6}\frac{6(6+1)}{2} = \frac{7}{2},$$

$$\mathrm{Var}(X_i) = \sum_{k=1}^{6}(k - \frac{7}{2})^2 P(X_i = k) = \frac{1}{6}\sum_{k=1}^{6}\frac{(2k-7)^2}{4} = \frac{35}{12} \simeq 2,9$$

Lancio ripetuto di un dado

Un campione di dimensione $N = 1$ è costituito da una sola v.c. X_1 con distribuzione discreta uniforme. Si ha in tal caso

$$m_1 = X_1, \quad E[m_1] = E[X_1] = \frac{7}{2} = \mu$$

ma, evidentemente m_1, assumendo solo valori interi (k) non assume mai il valore $\mu = \frac{7}{2}$.

Lancio ripetuto di un dado

Nel caso di due lanci il risultato è espresso dalla variabile $X_1 + X_2$ che assume valori interi tra $x = 2$ e $x = 12$, ma questi non sono più equiprobabili. La situazione si può schematizzare nel modo seguente

x	coppie di somma x	$f(x) = P(X_1 + X_2 = x)$
2	$(1,1)$	$f(2) = 1/36$
3	$(1,2)$ $(2,1)$	$f(3) = 2/36$
4	$(1,3)$ $(2,2)$ $(3,1)$	$f(4) = 3/36$
5	$(1,4)$ $(2,3)$ $(3,2)$ $(4,1)$	$f(5) = 4/36$
6	$(1,5)$ $(2,4)$ $(3,3)$ $(4,2)$ $(5,1)$	$f(6) = 5/36$
7	$(1,6)$ $(2,5)$ $(3,4)$ $(4,3)$ $(5,2)$ $(6,1)$	$f(7) = 6/36$
8	$(2,6)$ $(3,5)$ $(4,4)$ $(5,3)$ $(6,2)$	$f(8) = 5/36$
9	$(3,6)$ $(4,5)$ $(5,4)$ $(6,3)$	$f(9) = 4/36$
10	$(4,6)$ $(5,5)$ $(6,4)$	$f(10) = 3/36$
11	$(5,6)$ $(6,5)$	$f(11) = 2/36$
12	$(6,6)$	$f(12) = 1/36$
	Tot. 36	

Lancio ripetuto di un dado

Quindi la distribuzione della media $\frac{X_1+X_2}{2}$ è

x	coppie a media x	$f(x) = P(\frac{X_1+X_2}{2} = x)$
1	$(1,1)$	$f(1) = 1/36$
3/2	$(1,2)$ $(2,1)$	$f(3/2) = 2/36$
2	$(1,3)$ $(2,2)$ $(3,1)$	$f(2) = 3/36$
5/2	$(1,4)$ $(2,3)$ $(3,2)$ $(4,1)$	$f(5/2) = 4/36$
3	$(1,5)$ $(2,4)$ $(3,3)$ $(4,2)$ $(5,1)$	$f(3) = 5/36$
7/2	$(1,6)$ $(2,5)$ $(3,4)$ $(4,3)$ $(5,2)$ $(6,1)$	$f(7/2) = 6/36 = 1/6$
4	$(2,6)$ $(3,5)$ $(4,4)$ $(5,3)$ $(6,2)$	$f(4) = 5/36$
9/2	$(3,6)$ $(4,5)$ $(5,4)$ $(6,3)$	$f(9/2) = 4/36$
5	$(4,6)$ $(5,5)$ $(6,4)$	$f(5) = 3/36$
11/2	$(5,6)$ $(6,5)$	$f(11/2) = 2/36$
6	$(6,6)$	$f(6) = 1/36$
	Tot. 36	

- la distribuzione della media non è uniforme
- la media campionaria più probabile coincide con $\mu = 7/2$

Lancio ripetuto di un dado

Continuando con un campione di dimensione $N > 2$, come abbiamo fatto nel caso dei lanci della moneta si noterebbe che la distribuzione della media comincia ad assumere una forma a campana.

Un esperimento simulato al computer di lancio di dadi si trova sul sito http://www.stat.sc.edu/∼west/javahtml/CLT.html

Il Teorema del Limite Centrale

Il fenomeno di convergenza della distribuzione delle medie ad una distribuzione Normale osservato negli esempi precedenti è formalizzato nel Teorema del Limite Centrale (TLC).

Esso afferma che, sotto opportune condizioni abbastanza generali (la più forte è l'indipendenza), **la distribuzione della somma (e quindi della media) di variabili casuali aventi tutte la medesima distribuzione, converge,** in un senso che andrebbe meglio precisato, **alla distribuzione Normale quando la numerosità tende ad infinito.**
Vale a dire, considerando la variabile m_N standardizzata,

$$\frac{m_N - \mu}{\sigma/\sqrt{N}} \to \mathcal{N}(0,1), \qquad \text{per } N \to \infty$$

Importanza del Teorema del Limite Centrale

Il TLC è importantissimo, perché ci consente di utilizzare la distribuzione Normale anche quando la popolazione non è distribuita normalmente, purché il campione sia sufficientemente grande.

Non esiste una regola per stabilire quando l'approssimazione basata sul TLC è buona: in alcuni casi anche poche osservazioni sono sufficienti, mentre in altri la numerosità campionaria deve essere dell'ordine delle centinaia.

Il Teorema del Limite Centrale

Le **applicazioni alla statistica** si basano sul seguente principio:

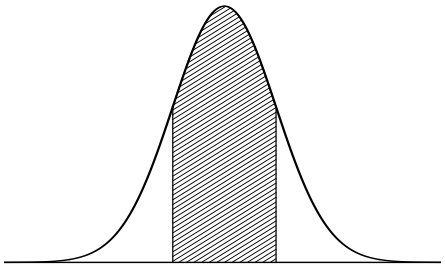
se X_1, X_2, ..., X_N sono v.c. che rappresentano i dati di un campione di dimensione N estratto da una popolazione con media (di popolazione) μ e varianza σ^2, la media campionaria m_N è distribuita, approssimativamente, come una variabile aleatoria gaussiana di media μ e varianza σ^2/N, cioè

$$m_N \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$$

formula, appunto, solo "approssimativamente" vera, perché in effetti m_N potrebbe anche essere discreta, come visto negli esempi precedenti

Servendoci delle stime per il calcolo della probabilità per un intervallo

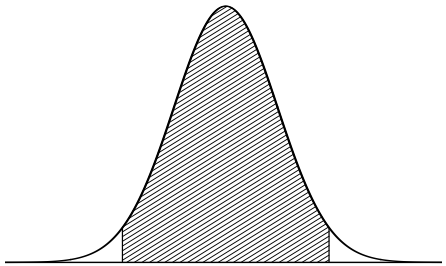
Alcune situazioni particolari - 1



$P(\mu - \sigma \leq X \leq \mu + \sigma) \simeq 0.67$

38

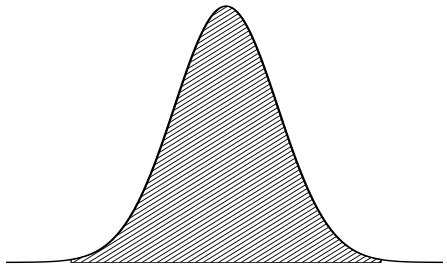
Alcune situazioni particolari - 2



$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \simeq 0.95$

39

Alcune situazioni particolari - 3



$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \simeq 0.997$

40

$$|m_N - \mu| \leq \frac{\sigma}{\sqrt{N}} \text{ con probabilità } 0.682$$

$$|m_N - \mu| \leq \frac{2\sigma}{\sqrt{N}} \text{ con probabilità } 0.954$$

$$|m_N - \mu| \leq \frac{3\sigma}{\sqrt{N}} \text{ con probabilità } 0.997$$

e la stima di μ con m_N diventa più accurata al crescere di N.

Stima e test delle ipotesi

Il problema di inferenza, cioè la formulazione di considerazioni di carattere generale a partire dalla sintesi dei dati campionari, può essere impostato in modi diversi.

- **Stima** sulla base dell'evidenza empirica: si assegna
  - un valore (stima **puntuale**)
  - un insieme di valori (stima **per intervallo**) al parametro di interesse
- **Test delle ipotesi**: si formulano ipotesi alternative sul valore del parametro di interesse e si valuta quale è maggiormente supportata dall'evidenza empirica

Stima puntuale

Il parametro incognito viene stimato mediante un'opportuna funzione dei dati campionari, detta stimatore.

Solitamente si usa:

- la media campionaria per stimare la media della popolazione
- la varianza campionaria per stimare la varianza della popolazione
- la frequenza relativa di successo per stimare la probabilità di successo

Stimatore e stima

La stima è il valore che lo stimatore assume nel campione osservato.

Lo stimatore è una v.c., la stima è un numero.

Mentre siamo in grado di valutare la qualità dello stimatore in base alle sue caratteristiche nell'universo dei campioni, non possiamo dire nulla della stima ottenuta in corrispondenza del singolo campione osservato.

In particolare, non siamo in grado, sulla base della sola stima (un numero), di valutare l'errore dovuto al campionamento.

Stima per intervallo

Il parametro viene stimato mediante un intervallo (detto intervallo di confidenza) i cui estremi dipendono dal campione estratto (sono casuali).

Un intervallo di confidenza è quindi un insieme di valori plausibili per il parametro incognito sulla base dell'evidenza empirica.

Se il campione è rappresentativo (ovviamente è impossibile saperlo), allora l'intervallo contiene il valore del parametro da stimare.

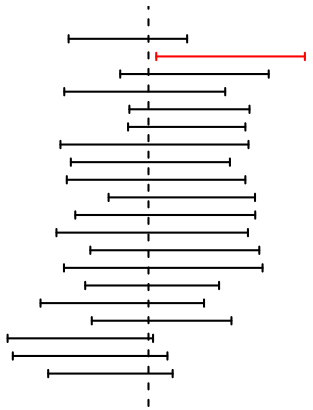
Stima per intervallo

Gli estremi dell'intervallo vengono individuati in modo tale che la probabilità di estrarre un campione che fornisce un risultato corretto (leggi l'intervallo contiene il valore del parametro) sia fissata pari a $1 - \alpha$ (livello di confidenza).

Attenzione: il livello di confidenza rappresenta il grado di affidabilità della procedura, non il grado di affidabilità del risultato corrispondente al singolo campione estratto. Generalmente si usa come livello di confidenza il 95% ($\alpha = 5\%$).

Ripetendo l'operazione di stima ...

su più campioni, potrebbe capitare la cosa seguente



Stima per intervallo della media

Indicando con μ e σ^2 la media e la varianza di X (incognite), una stima per intervallo del parametro μ può essere ottenuta sfruttando il fatto che:

$$\frac{m_N - \mu}{\sigma/\sqrt{N}} \to \mathcal{N}(0,1)$$

oppure

$$\frac{m_N - \mu}{S_N/\sqrt{N}} \sim t_{N-1}$$

dove t_{N-1} indica la distribuzione t di Student con $N-1$ gradi di libertà.

Solitamente la varianza della popolazione è incognita (mentre la varianza campionaria S_N è nota) e si deve quindi necessariamente ricorrere alla seconda espressione.

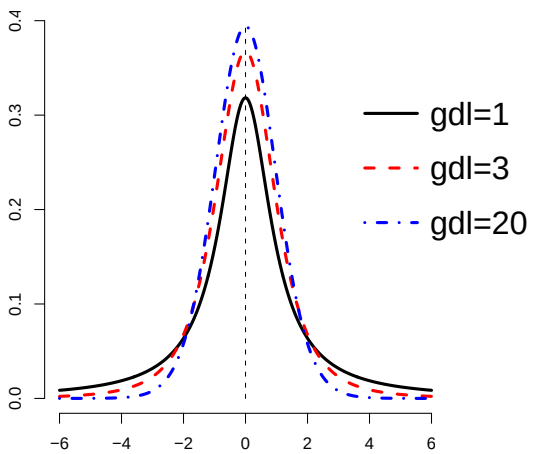
La distribuzione t di Student

W.S. Gossett (1876-1937), statistico inglese che si firmava "Student", ha mostrato che la variabile aleatoria

$$\frac{m_N - \mu}{S_N/\sqrt{N}}$$

ha una precisa distribuzione di probabilità detta *t* di Student

Densità della distribuzione t



La distribuzione t di Student

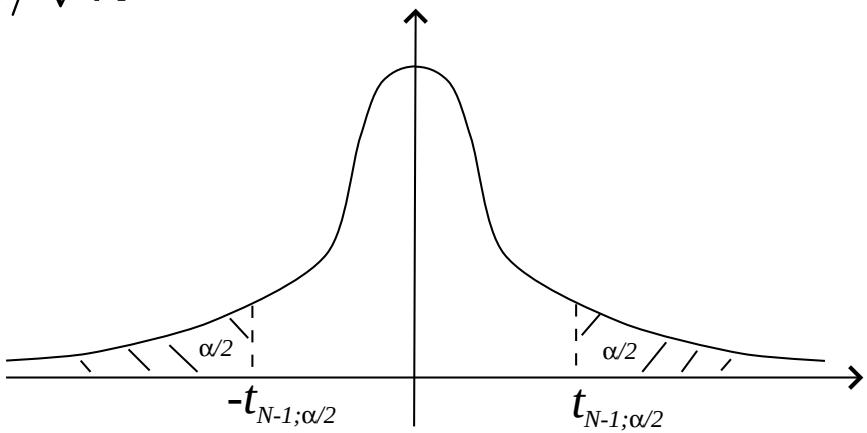
La distribuzione t di Student ha un andamento simile a quello della distribuzione Normale (campanulare simmetrico).

Rispetto alla Normale, la t ha le code più alte ("pesanti"), perché rappresenta una situazione di maggiore variabilità (incertezza), derivante dalla stima (soggetta quindi ad errore) della varianza della popolazione.

Le tavole della distribuzione t di Student consentono di trovare $t_{N-1;\alpha}$, ossia il valore che lascia sulla coda di destra un'area prefissata α.

L'intervallo che stima la media

Sapendo che $\frac{m_N - \mu}{S_N/\sqrt{N}} \sim t_{N-1}$ e che



$$P\left(\frac{m_N - \mu}{S_N/\sqrt{N}} \in [-t_{N-1;\alpha/2}, t_{N-1;\alpha/2}]\right) = 1 - \alpha,$$

allora l'intervallo di confidenza per la stima della media μ di una distribuzione a varianza incognita e livello di confidenza $1 - \alpha$ è

$$\mu \in [m_N - \frac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}, m_N + \frac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}]$$

Esempio - lunghezza media delle spighe di mais

Si vuole stimare per intervallo $(1 - \alpha = 95\%)$ *la lunghezza media della spiga di una nuova varietà di mais. I valori osservati sono riportati nella tabella che segue:*

X	
17.2	
20.1	
18.4	
16.3	
15.0	
14.8	
19.2	
16.7	
15.8	
17.8	
171.3	

Esempio - lunghezza media delle spighe di mais

Si vuole stimare per intervallo $(1 - \alpha = 95\%)$ *la lunghezza media della spiga di una nuova varietà di mais. I valori osservati sono riportati nella tabella che segue:*

X	
17.2	
20.1	
18.4	
16.3	
15.0	
14.8	
19.2	
16.7	
15.8	
17.8	
171.3	

Stima puntuale

$$m_{10} = \sum_{i=1}^{10} \frac{1}{10} x_i = 17.13$$

Stima per intervallo

$$[m_N - \frac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}, m_N + \frac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}] = ?$$

Esempio - lunghezza media delle spighe di mais

Si vuole stimare per intervallo $(1 - \alpha = 95\%)$ *la lunghezza media della spiga di una nuova varietà di mais. I valori osservati sono riportati nella tabella che segue:*

X	
17.2	
20.1	
18.4	
16.3	
15.0	
14.8	
19.2	
16.7	
15.8	
17.8	
171.3	

Stima puntuale

$$m_{10} = \sum_{i=1}^{10} \frac{1}{10} x_i = 17.13$$

Stima per intervallo

$$[m_N - \frac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}, m_N + \frac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}] = ?$$

$$S_{10}^2 = \frac{1}{10-1} \sum_{i=1}^{10} x_i^2 - \frac{10}{10-1} m_{10}^2$$

Esempio - lunghezza media delle spighe di mais

Si vuole stimare per intervallo $(1 - \alpha = 95\%)$ *la lunghezza media della spiga di una nuova varietà di mais. I valori osservati sono riportati nella tabella che segue:*

X	X^2
17.2	295.84
20.1	404.01
18.4	338.56
16.3	265.69
15.0	225.00
14.8	219.04
19.2	368.64
16.7	278.89
15.8	249.64
17.8	316.84
171.3	2962.15

Stima puntuale

$$m_{10} = \sum_{i=1}^{10} \frac{1}{10} x_i = 17.13$$

Stima per intervallo

$$[m_N - \frac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}, m_N + \frac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}] = ?$$

$$S_{10}^2 = \frac{1}{10-1} \sum_{i=1}^{10} x_i^2 - \frac{10}{10-1} m_{10}^2 = 3.0868$$

Esempio - lunghezza media delle spighe di mais

Si vuole stimare per intervallo $(1-\alpha = 95\%)$ *la lunghezza media della spiga di una nuova varietà di mais. I valori osservati sono riportati nella tabella che segue:*

X	X^2
17.2	295.84
20.1	404.01
18.4	338.56
16.3	265.69
15.0	225.00
14.8	219.04
19.2	368.64
16.7	278.89
15.8	249.64
17.8	316.84
171.3	2962.15

Stima puntuale

$$m_{10} = \sum_{i=1}^{10} \frac{1}{10} x_i = 17.13$$

Stima per intervallo

$$[m_N - \frac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}, m_N + \frac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}] = ?$$

$$S_{10}^2 = \frac{1}{10-1} \sum_{i=1}^{10} x_i^2 - \frac{10}{10-1} m_{10}^2 = 3.0868$$

$$t_{9;0.025} = 2.2622$$

Esempio - lunghezza media delle spighe di mais

Si vuole stimare per intervallo $(1-\alpha = 95\%)$ *la lunghezza media della spiga di una nuova varietà di mais. I valori osservati sono riportati nella tabella che segue:*

X	X^2
17.2	295.84
20.1	404.01
18.4	338.56
16.3	265.69
15.0	225.00
14.8	219.04
19.2	368.64
16.7	278.89
15.8	249.64
17.8	316.84
171.3	2962.15

Stima puntuale

$$m_{10} = \sum_{i=1}^{10} \frac{1}{10} x_i = 17.13$$

Stima per intervallo

$$[m_N - \frac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}, m_N + \frac{S_N}{\sqrt{N}} t_{N-1;\alpha/2}] = ?$$

$$S_{10}^2 = \frac{1}{10-1} \sum_{i=1}^{10} x_i^2 - \frac{10}{10-1} m_{10}^2 = 3.0868$$

$$t_{9;0.025} = 2.2622$$

$$[m_{10} - \frac{S_{10}}{\sqrt{10}} t_{9;0.025}, m_{10} + \frac{S_{10}}{\sqrt{10}} t_{9;0.025}] = [15.87; 18.39]$$

L'ampiezza dell'intervallo

L'ampiezza dell'intervallo è molto rilevante. Quanto più l'intervallo è stretto, tanto maggiore è il grado di precisione che caratterizza lo strumento statistico utilizzato.

L'ampiezza dell'intervallo

Nella **stima della media**, l'ampiezza dell'intervallo è pari a

$$\Delta = 2t_{N-1;\frac{\alpha}{2}} \frac{S_N}{\sqrt{N}}$$

NB: usando S_N, l'ampiezza dell'intervallo è una v.c., in quanto dipende dal campione estratto.

L'ampiezza dell'intervallo dipende quindi da

- α: al diminuire di α (al crescere del livello di confidenza $1-\alpha$) l'ampiezza dell'intervallo aumenta
- S_N: misura la variabilità del fenomeno studiato. Al crescere della variabilità, cresce anche l'incertezza e quindi l'ampiezza dell'intervallo aumenta
- N: al crescere di N aumenta la quantità di informazione disponibile e quindi l'ampiezza dell'intervallo diminuisce

Il dimensionamento del campione

In fase di pianificazione dello studio, è importante determinare la **numerosità campionaria** in modo tale che gli strumenti statistici utilizzati abbiano certe caratteristiche (per es. elevata precisione o bassa probabilità di errore).

Nel caso di **stima per intervallo**, l'obiettivo da raggiungere si individua fissando a priori un certo grado di precisione, ossia una certa ampiezza dell'intervallo.

Dimensionamento per la stima della media

Indicando con Δ^* l'ampiezza dell'intervallo prefissata, si ottiene

$$N = \left(\frac{2t_{N-1;\alpha/2}}{\Delta^*}\right)^2 S_N^2$$

Per calcolare il valore di N bisogna risolvere due **problemi:**

1. S_N^2 non è nota prima di estrarre il campione
2. $t_{N-1;\alpha/2}$ dipende da N (l'espressione non è in forma chiusa)

Soluzioni:

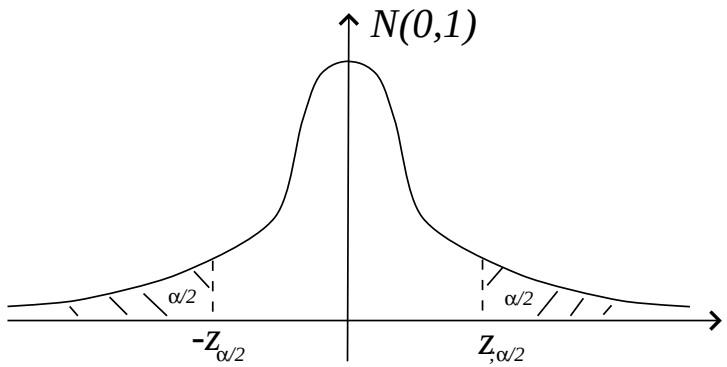
1. usare un valore presunto per S_N^2 (indicato con S^{*2}) derivandolo da studi precedenti, indagini pilota o valutazioni di esperti
2. usare un algoritmo iterativo, calcolando ripetutamente N usando di volta in volta i gradi di libertà ottenuti al passo precedente

L'algoritmo iterativo

L'algoritmo procede nel modo seguente:

1. $N_0 = \infty$ (inizializzazione)
2. $N_1 = \left(\frac{2t_{\infty;\alpha/2}}{\Delta^*}\right)^2 S^{*2} = \left(\frac{2z_{\alpha/2}}{\Delta^*}\right)^2 S^{*2}, \quad z_{\alpha/2} = \text{coda della } \mathcal{N}(0,1)$
3. $N_2 = \left(\frac{2t_{N_1-1;\alpha/2}}{\Delta^*}\right)^2 S^{*2}$
4. .....

terminando quando si ottiene lo stesso valore in due passi successivi.



Esempio

Calcolare il numero di osservazioni necessario per stimare con un intervallo di ampiezza pari a 1.5 *la lunghezza media della spiga di una nuova varietà di mais (livello di confidenza* 95%*). Su varietà simili si è osservata una varianza pari a* 3*.*

Esempio

Calcolare il numero di osservazioni necessario per stimare con un intervallo di ampiezza pari a 1.5 *la lunghezza media della spiga di una nuova varietà di mais (livello di confidenza* 95%*). Su varietà simili si è osservata una varianza pari a* 3*.*

$$N_0 = \infty \quad \Longrightarrow \quad t_{N_0;0.025} = z_{0.025} = 1.96$$

Esempio

Calcolare il numero di osservazioni necessario per stimare con un intervallo di ampiezza pari a 1.5 *la lunghezza media della spiga di una nuova varietà di mais (livello di confidenza* 95%*). Su varietà simili si è osservata una varianza pari a* 3*.*

$$N_0 = \infty \quad \Longrightarrow \quad t_{N_0;0.025} = z_{0.025} = 1.96$$

$$N_1 = \left(\frac{2t_{\infty;0.025}}{\Delta^*}\right)^2 S^{*2} = \left(\frac{2 \cdot 1.96}{1.5}\right)^2 \cdot 3 = 20.49 \simeq 20$$

Esempio

Calcolare il numero di osservazioni necessario per stimare con un intervallo di ampiezza pari a 1.5 *la lunghezza media della spiga di una nuova varietà di mais (livello di confidenza* 95%*). Su varietà simili si è osservata una varianza pari a* 3*.*

$$N_0 = \infty \quad \Longrightarrow \quad t_{N_0;0.025} = z_{0.025} = 1.96$$

$$N_1 = \left(\frac{2t_{\infty;0.025}}{\Delta^*}\right)^2 S^{*2} = \left(\frac{2 \cdot 1.96}{1.5}\right)^2 \cdot 3 = 20.49 \simeq 20$$

$$N_2 = \left(\frac{2t_{19;0.025}}{\Delta^*}\right)^2 S^{*2} = \left(\frac{2 \cdot 2.093}{1.5}\right)^2 \cdot 3 = 23.36 \simeq 23$$

Esempio

Calcolare il numero di osservazioni necessario per stimare con un intervallo di ampiezza pari a 1.5 *la lunghezza media della spiga di una nuova varietà di mais (livello di confidenza* 95%*). Su varietà simili si è osservata una varianza pari a* 3*.*

$$N_0 = \infty \quad \Longrightarrow \quad t_{N_0;0.025} = z_{0.025} = 1.96$$

$$N_1 = \left(\frac{2t_{\infty;0.025}}{\Delta^*}\right)^2 S^{*2} = \left(\frac{2 \cdot 1.96}{1.5}\right)^2 \cdot 3 = 20.49 \simeq 20$$

$$N_2 = \left(\frac{2t_{19;0.025}}{\Delta^*}\right)^2 S^{*2} = \left(\frac{2 \cdot 2.093}{1.5}\right)^2 \cdot 3 = 23.36 \simeq 23$$

$$N_3 = \left(\frac{2t_{22;0.025}}{\Delta^*}\right)^2 S^{*2} = \left(\frac{2 \cdot 2.0739}{1.5}\right)^2 \cdot 3 = 22.94 \simeq 23$$

Esempio

Calcolare il numero di osservazioni necessario per stimare con un intervallo di ampiezza pari a 1.5 *la lunghezza media della spiga di una nuova varietà di mais (livello di confidenza* 95%*). Su varietà simili si è osservata una varianza pari a* 3*.*

$$N_0 = \infty \quad \Longrightarrow \quad t_{N_0;0.025} = z_{0.025} = 1.96$$

$$N_1 = \left(\frac{2t_{\infty;0.025}}{\Delta^*}\right)^2 S^{*2} = \left(\frac{2 \cdot 1.96}{1.5}\right)^2 \cdot 3 = 20.49 \simeq 20$$

$$N_2 = \left(\frac{2t_{19;0.025}}{\Delta^*}\right)^2 S^{*2} = \left(\frac{2 \cdot 2.093}{1.5}\right)^2 \cdot 3 = 23.36 \simeq 23$$

$$N_3 = \left(\frac{2t_{22;0.025}}{\Delta^*}\right)^2 S^{*2} = \left(\frac{2 \cdot 2.0739}{1.5}\right)^2 \cdot 3 = 22.94 \simeq 23$$

La regola di arresto è soddisfatta e possiamo quindi fermarci.
Ripetendo il passo ancora una volta otterremmo lo stesso risultato.

Stima per intervallo di una probabilità

Se la popolazione è descritta mediante una distribuzione di Bernoulli (fenomeno dicotomico), il parametro da **stimare** è **la probabilità di successo p**.

Stima per intervallo di una probabilità

Se il campione è sufficientemente grande, per il TLC si ha

$$m_N \sim \mathcal{N}(\mu, \frac{\sigma^2}{N}) = \mathcal{N}(p, \frac{p(1-p)}{N})$$

cosa solo "approssimativamente" vera (m_N è discreta).

In modo analogo a quanto visto per la media della Normale, otteniamo il seguente **intervallo di confidenza per p** (livello di confidenza $1 - \alpha$)

$$p \in \left[m_N - z_{\alpha/2} \sqrt{\frac{m_N(1-m_N)}{N}}, m_N + z_{\alpha/2} \sqrt{\frac{m_N(1-m_N)}{N}} \right]$$

Stima per intervallo di una probabilità

Infatti, standardizzando si ha

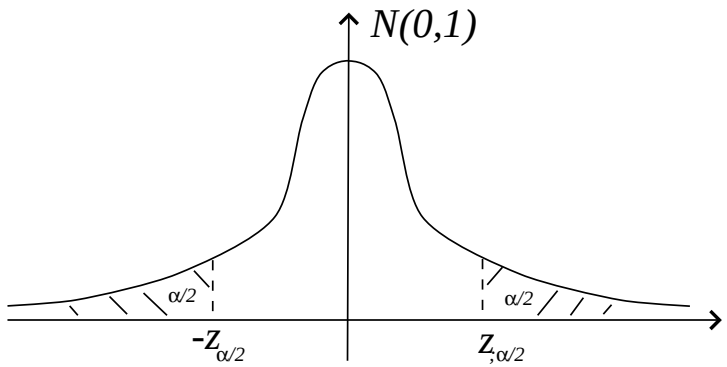
$$\frac{m_N - E[m_N]}{\sqrt{\mathrm{Var}[m_M]}} = \frac{m_N - p}{\sqrt{p(1-p)}}\sqrt{N} \sim \mathcal{N}(0,1),$$

quindi

$$P\Big(\frac{m_N - p}{\sqrt{p(1-p)}}\sqrt{N} \in I\Big) = P(\mathcal{N}(0,1) \in I).$$

D'altra parte

$$P(\mathcal{N}(0,1) \in I) = 1 - \alpha \quad \text{se} \quad I = [-z_{\alpha/2}, z_{\alpha/2}]$$



Stima per intervallo di una probabilità

Affinché $P(\frac{m_N - p}{\sqrt{p(1-p)}}\sqrt{N} \in I) = 1 - \alpha$ è quindi sufficiente che

$$\frac{m_N - p}{\sqrt{p(1-p)}}\sqrt{N} \in [-z_{\alpha/2}, z_{\alpha/2}],$$

cioè che

$$-z_{\alpha/2} \leq \frac{m_N - p}{\sqrt{p(1-p)}}\sqrt{N} \leq z_{\alpha/2}.$$

Per determinare un intervallo di confidenza per p è dunque sufficiente risolvere quest'ultimo sistema di disuguaglianze nell'incognita p. Il problema si semplifica sostituendo il denominatore $\sqrt{p(1-p)}$ con $\sqrt{m_N(1-m_N)}$ cioè sostituendo a p lo stimatore m_N per la stima della varianza.

Esempio

Sondaggio

100 persone vengono intervistate su come voteranno ad un referendum.

42 dichiarano di votare NO

58 dichiarano di votare SI

Determiniamo un'intervallo di confidenza al 95% *per la percentuale di SI al referendum.*

Sondaggio

100 persone vengono intervistate su come voteranno ad un referendum.

42 dichiarano di votare NO

58 dichiarano di votare SI

Determiniamo un'intervallo di confidenza al 95% *per la percentuale di SI al referendum.*

Si ha $m_{100} = 58/100 = 0.580$ e $\alpha = 0.05$). Dunque

$$p \in \left[m_N - z_{0.025}\sqrt{\frac{m_N(1-m_N)}{N}}, m_N + z_{0.025}\sqrt{\frac{m_N(1-m_N)}{N}}\right] =$$

$$= \left[0.58 - 1.96\sqrt{\frac{0.58 \cdot 0.42}{100}}, 0.58 + 1.96\sqrt{\frac{0.58 \cdot 0.42}{100}}\right] =$$

$$= \Big[0.580 - 1.960 \cdot 0.049, 0.580 + 1.960 \cdot 0.049\Big] = [0.484, 0.670]$$

Il risultato non da risposte conclusive sull'esito del referendum. Con un livello di confidenza del 99% si avrebbe $\bar{p} = [0.45, 0.71]$.

Sondaggi

1000 persone vengono intervistate su come voteranno ad un referendum.

420 dichiarano di votare NO

580 dichiarano di votare SI

Determiniamo un'intervallo di confidenza al 95% *per la percentuale di SI al referendum.*

Sondaggi

1000 persone vengono intervistate su come voteranno ad un referendum.
420 dichiarano di votare NO
580 dichiarano di votare SI
Determiniamo un'intervallo di confidenza al 95% per la percentuale di SI al referendum.

Si ha $m_{1000} = 580/1000 = 0.58$ e $\alpha = 0.05$. Pertanto

$$p \in \left[m_N - z_{0.025}\sqrt{\frac{m_N(1-m_N)}{N}}, m_N + z_{0.025}\sqrt{\frac{m_N(1-m_N)}{N}}\right] =$$

$$= \left[0.58 - 1.96\sqrt{\frac{0.58 \cdot 0.42}{1000}}, 0.58 + 1.96\sqrt{\frac{0.58 \cdot 0.42}{1000}}\right] =$$

$$= \Big[0.58 - 1.96 \cdot 0.016, 0.58 + 1.96 \cdot 0.016\Big] = [0.549, 0.611]$$

Sondaggi

In relazione all'esempio precedente, calcolare quanto deve essere grande N per essere sicuri al 99% che vinceranno i SI, se si osserva una frequenza del 58% di SI sul campione.

Esempio

Definizione del problema: Si vuole valutare l'effetto della conservazione in atmosfera modificata dell'insalata.

Raccolta dei dati: su 200 confezioni è stata rilevata la presenza di foglie avvizzite dopo 5 giorni trascorsi in un banco frigo. Si sono osservate 158 confezioni integre, mentre 42 presentano segni di degrado.

Individuazione del modello teorico: se $X = 1$ se la confezione è integra e $X = 0$ altrimenti, allora $X \sim BI(1, p)$ dove p rappresenta la probabilità che una confezione si mantenga integra.

Problema di inferenza: *determinare un intervallo di confidenza per p con livello di confidenza del* 95%.

Stima per intervallo di una probabilità

Riepilogo dati:

- numero totale di confezioni: $N = 200$
- confezioni integre: 158
- livello di confidenza 95%, quindi $\alpha = 0.05$

Stima intervallare

$$\left[m_N - z_{\alpha/2}\sqrt{\frac{m_N(1-m_N)}{N}}, m_N + z_{\alpha/2}\sqrt{\frac{m_N(1-m_N)}{N}}\right] = ?$$

Stima per intervallo di una probabilità

Riepilogo dati:

- numero totale di confezioni: $N = 200$
- confezioni integre: 158
- livello di confidenza 95%, quindi $\alpha = 0.05$

Stima intervallare

$$\left[m_N - z_{\alpha/2}\sqrt{\frac{m_N(1-m_N)}{N}}, m_N + z_{\alpha/2}\sqrt{\frac{m_N(1-m_N)}{N}}\right] = ?$$

$$m_N = \frac{1}{N}\sum_{i=1}^{N} X_i = \frac{1}{200}158 = 0.79 \quad \text{(stima puntuale)}$$

$$z_{\alpha/2} = z_{0.025} = 1.96,$$

$$\left[m_N - z_{\alpha/2}\sqrt{\frac{m_N(1-m_N)}{N}}, m_N + z_{\alpha/2}\sqrt{\frac{m_N(1-m_N)}{N}}\right] =$$

$$= \left[0.79 - 1.96\sqrt{\frac{0.79(1-0.79)}{200}}, 0.79 + 1.96\sqrt{\frac{0.79(1-0.79)}{200}}\right] = [0.7335, 0.8465]$$

Stima per intervallo della varianza

Supponendo che $X \sim \mathcal{N}(\mu, \sigma^2)$, una stima per intervallo del parametro σ^2 può essere ottenuta sfruttando il fatto che:

$$\frac{(N-1)S_N^2}{\sigma^2} \sim \chi^2_{N-1}$$

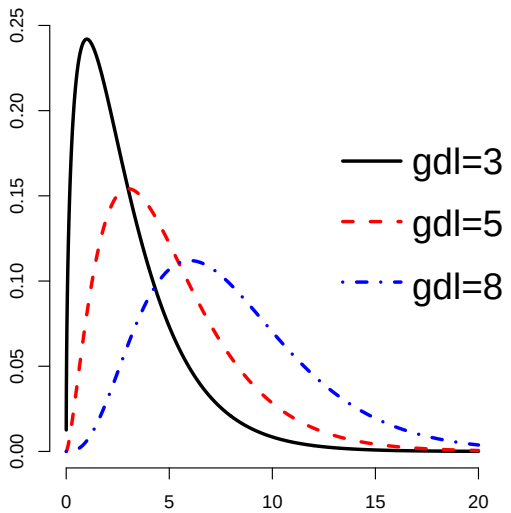
dove χ^2_{N-1} indica la distribuzione χ^2 (chi quadro) con $N-1$ gradi di libertà

La distribuzione χ^2

La v.c. χ^2 assume valori nell'intervallo $[0, +\infty[$ (come la varianza) ed ha distribuzione *asimmetrica*.

Le tavole della distribuzione χ^2 consentono di determinare $\chi^2_{N-1;\alpha}$, ossia il valore che lascia sulla coda di destra un'area prefissata α.

Densità della distribuzione χ^2



L'intervallo che stima la varianza

L'intervallo di confidenza per la stima della varianza a livello di confidenza $1 - \alpha$ ha la forma seguente:

$$\left[\frac{(N-1)S_N^2}{\chi^2_{N-1;\frac{\alpha}{2}}}, \frac{(N-1)S_N^2}{\chi^2_{N-1;1-\frac{\alpha}{2}}}\right]$$

Mentre l'intervallo per la media è simmetrico rispetto alla media campionaria, quello per la varianza è asimmetrico.

Esercizio

Nella tabella sono riportati i risultati ottenuti da un tecnico in 10 misurazioni della concentrazione di un certo principio attivo in una soluzione. *Stimare per intervallo* $(1 - \alpha = 95\%)$ *la varianza delle misure prodotte dal tecnico.*

X	X^2
14.8	219.04
14.7	216.09
14.8	219.04
15.0	225.00
14.6	213.16
14.7	216.09
14.5	210.25
14.8	219.04
14.8	219.04
14.7	216.09
147.4	2172.84

Stima per intervallo della varianza

Stima puntuale

$$N = 10, \quad m_N = \frac{1}{N}\sum_{i=1}^{N} x_i = 14.74,$$

$$S_N^2 = \frac{1}{N-1}\sum_{i=1}^{N} x_i^2 - \frac{N}{N-1} m_N^2 = 0.0182$$

Stima per intervallo

$$\chi^2_{9;0.025} = 19.0228, \quad \chi^2_{9;0.975} = 2.7004$$

$$\left[\frac{(N-1)S_N^2}{\chi^2_{N-1;\alpha/2}}, \frac{(N-1)S_N^2}{\chi^2_{N-1;1-\alpha/2}}\right] = \left[\frac{9 \cdot 0.0182}{\chi^2_{9;0.025}}, \frac{9 \cdot 0.0182}{\chi^2_{9;0.975}}\right] = [0.0086, 0.0607]$$

Esercizio - 12.10 del testo

Si sospetta che un campo di mais sia stato contaminato da semi transgenici oltre la soglia dello 0.1%. *Superata questa soglia è obbligatorio dichiarare la percentuale di OGM presente nelle farine ricavate dal mais.*
Viene analizzato un campione di 8000 *semi, di cui* 6 *risultano della varietà transgenica. A un livello di fiducia del* 95%, *qual'è l'intervallo di confidenza della frazione di semi transgenici sul totale della piantagione.*