Università di Firenze
Corso di Laurea in Statistica
Statistica

Leonardo Grilli

Intervalli di confidenza (intervalli fiduciari)

Cicchitelli cap. 17

A.A. 2010/2011

Argomenti

- Stima per intervallo (IC: Intervalli di Confidenza):
  - IC per la media - varianza nota a priori
  - Proprietà degli IC:
    - Ampiezza
    - Livello di confidenza nominale vs effettivo
    - Ruolo dell'ipotesi di Normalità
  - IC per la media - varianza non nota a priori
  - IC per la proporzione
  - Determinazione della dimensione campionaria
  - IC per la varianza

Perché una stima per intervallo?

- L'inferenza statistica consiste nell'usare statistiche (= quantità calcolate nel campione) per stimare parametri incogniti della popolazione
- Come ogni processo induttivo, l'inferenza statistica porta a **conclusioni incerte**: infatti, in generale la stima non coincide con il parametro obiettivo (anche se una buona stima non dovrebbe esserne troppo lontana)
- La peculiarità dell'inferenza statistica è quella di ***quantificare l'incertezza associata al processo induttivo***
- La quantificazione dell'incertezza è essenziale per la corretta interpretazione di una stima

Stima per intervallo

- Obiettivo: quantificare l'incertezza derivante dalla variabilità campionaria, cioè il fatto che la stima varia a seconda del campione estratto
- Un modo per quantificare l'incertezza è quello di associare alla stima puntuale (es. $\bar{X}$ = 10) un intervallo, es. [8, 12], detto **intervallo di confidenza (IC)** o **intervallo fiduciario**, che contenga il parametro da stimare con una probabilità controllata, detta **livello di confidenza** e indicata con $1-\alpha$
- A parità di livello di confidenza, quanto più corto è l'intervallo tanto minore è l'incertezza (l'intervallo è più informativo)

Stimatore per intervallo

- Dato un campione casuale di dimensione *n* da una v.a. *X* avente distribuzione regolata dal parametro θ

  e date due statistiche campionarie

$$L_1 = \ell_1(X_1,\ldots,X_n)$$
$$L_2 = \ell_2(X_1,\ldots,X_n) \qquad L_1 < L_2$$

  l'intervallo aleatorio (L_1,L_2) si dice **stimatore per intervallo** se include il parametro θ con una probabilità che non dipende da θ:

$$P(L_1 < \theta < L_2) = 1-\alpha \quad \forall\theta$$

Confidenza vs ampiezza

- La qualità dell'informazione sul parametro θ che si ottiene da uno stimatore per intervallo dipende da due aspetti:
  - Il **livello di confidenza** (probabilità di copertura) $1-\alpha$, o in modo equivalente, la **probabilità di errore** (non copertura) α
  - L'**ampiezza dell'intervallo** $A = L_2-L_1$
- I due aspetti sono legati da una relazione inversa: una riduzione di α comporta un aumento di $A = L_2-L_1$
- A parità di α, più piccolo è A più informativo è l'intervallo

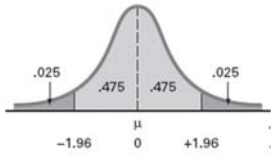
IC per μ (X~Normale, σ^2 nota) /1

- Consideriamo un carattere X~N(μ, σ^2), dove la media μ è ignota, mentre la varianza σ^2 è nota
- La derivazione dell'IC è più semplice quando la varianza è nota, anche se ciò accade raramente (tipicamente occorre disporre di dati raccolti in passato e assumere che la varianza non cambi nel tempo; successivamente abbandoneremo questa ipotesi e deriveremo l'IC per μ quando σ^2 è ignota)
- Supponiamo che $X_1,\ldots, X_n$ sia un campione casuale da X (cioè elementi campionari indipendenti e distribuiti come X)
- Il migliore stimatore della media della popolazione μ è la media campionaria
- X distrib. Normale → media campionaria distrib. Normale

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Poiché la media campionaria standardizzata ha distribuzione Normale standard, l'intervallo centrato su 0 di probabilità $1-\alpha$ (dove α è un valore fissato dall'analista, vedremo poi con quali criteri) è quello che lascia $\alpha/2$ sulle code, cioè l'intervallo $[-z_{\alpha/2}, z_{\alpha/2}]$, dove $z_{\alpha/2}$ è il valore z (*valore critico*) che lascia a destra una probabilità $\alpha/2$ (e quindi a sinistra un'area pari a $1-\alpha/2$). Ad esempio, $z_{0.05/2}=1.96$, $z_{0.10/2}=1.65$

$$1-\alpha = P\left(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le +z_{\alpha/2}\right)$$
$$= P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \bar{X}-\mu \le +z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$
$$= P\left(-\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le -\mu \le -\bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$
$$= P\left(\underbrace{\bar{X}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{L_1} \le \mu \le \underbrace{\bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{L_2}\right)$$



Si trasforma la disuguaglianza in modo che al centro compaia μ

IC per μ (X~Normale, σ^2 nota) /2

- In sintesi: con un campione casuale da $X \sim N(\mu, \sigma^2)$, con σ^2 nota, si ha

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Variabile aleatoria — Quantità note — Quantità fissa ma incognita

- L'intervallo $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ è un intervallo aleatorio perché è centrato sulla media campionaria, che varia da campione a campione
- *Prima di estrarre il campione* vi è una probabilità $1-\alpha$ che tale intervallo racchiuda la media della popolazione μ ; infatti, nell'universo dei possibili campioni, $(1-\alpha)100\%$ dei campioni portano ad un intervallo che include μ (c. "vincenti") e $\alpha 100\%$ ad un intervallo che non include μ (c. "perdenti")

IC per μ (X~Normale, σ^2 nota) /3

- In pratica si estrae un solo campione, sul quale si calcola la media. Pertanto, *dopo l'estrazione del campione*, la variabile aleatoria $\bar{X}$ diventa un numero $\bar{x}$ e di conseguenza l'intervallo aleatorio

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

diventa un intervallo con estremi determinati

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Intervallo di confidenza al livello $1-\alpha$

- Ad es. se $\sigma=40$ e $n=100$, scegliendo $\alpha=0.05$ si ha $z_{\alpha/2}=1.96$ e quindi il semi-intervallo è $1.96\times 40/\sqrt{100} = 7.84$ (e la lunghezza complessiva dell'intervallo è $2\times 7.84=15.68$). Dunque, prima di estrarre il campione si ha un intervallo aleatorio (al livello 95%) $\bar{X} \pm 7.84$ mentre una volta estratto il campione l'intervallo è determinato: ad es. se nel campione estratto la media è 250, l'intervallo di confidenza al livello 95% è 250 ± 7.84 = [242.16, 257.84]

Significato del livello di confidenza

- Una volta estratto il campione vi sono dunque due possibilità
  - Si è estratto un campione "vincente" (ve ne sono $(1-\alpha)100\%$) per cui l'intervallo appena calcolato include la media della popolazione μ
  - Si è estratto un campione "perdente" (ve ne sono $\alpha 100\%$) per cui l'intervallo appena calcolato non include la media della popolazione μ
- Poiché la media della popolazione μ è ignota non si può sapere se il campione estratto è "vincente" (= include μ) o "perdente" (= non include μ)
- Tutto quello che si può dire è che, prima di estrarre il campione, questa procedura porta
  - con probabilità $(1-\alpha)$ ad un intervallo che include la media della popolazione μ
  - con probabilità α ad un intervallo che non include la media della popolazione μ
- Fissando un livello di confidenza $(1-\alpha)$ alto (= fissando α ad un livello basso) si ottiene una procedura che con elevata probabilità fa la cosa giusta
- Nella singola applicazione si può essere sfortunati, per cui l'intervallo calcolato non include μ (anche se non lo sapremo mai che non include μ), ma nel lungo periodo questa procedura funziona bene perché in circa $(1-\alpha)100\%$ delle applicazioni l'intervallo include μ

Esempio /1

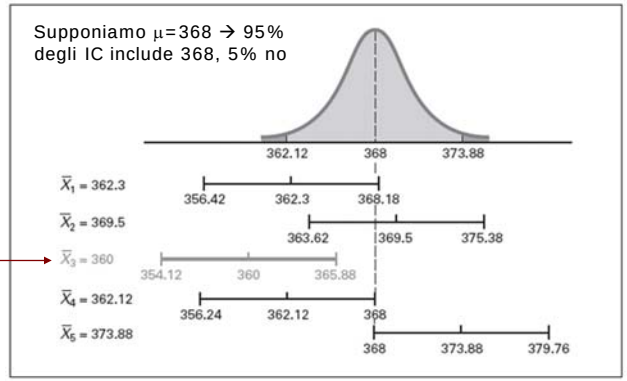
- Processo industriale per il riempimento delle scatole di cereali
- Il carattere di interesse è X = "peso in grammi dei cereali nella scatola"
- Il parametro di interesse è μ = "peso medio in grammi dei cereali nella scatola"
- Si assume che la distribuzione del peso sia Normale con una deviazione std nota dall'esperienza $\sigma=15$: in simboli, $X \sim N(\mu, 15^2)$
- Si assume di disporre di un campione casuale di $n=25$ scatole
- Con un livello di confidenza 95% ($\alpha=0.05$) si ottiene l'intervallo aleatorio

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \Rightarrow \bar{X} \pm 1.96\frac{15}{\sqrt{25}} \Rightarrow \bar{X} \pm 5.88$$

Esempio /1

$$\bar{X} \pm 5.88$$

- se la media risulta 362.3 si ottiene 362.3±5.88 = [356.42, 368.18]
- se la media risulta 369.5 si ottiene 369.5±5.88 = [363.62, 375.38]
- ….



Ampiezza dell'IC /1

Quando σ è nota l'ampiezza dell'IC è fissa (non varia da campione a campione)

$$A = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- I fattori che influenzano l'ampiezza dell'IC sono:
  - La deviazione standard del carattere nella popolazione σ (fattore non controllabile in alcun modo dall'analista): all'aumentare di σ aumenta la lunghezza dell'IC
  - L'ampiezza campionaria *n* (fattore controllabile in fase di progettazione dell'indagine): all'aumentare di *n* diminuisce la lunghezza dell'IC (in proporzione alla radice quadrata di *n*)
  - Il livello di confidenza $1-\alpha$ (fattore controllabile dall'analista) che determina il valore critico $z_{\alpha/2}$: all'aumentare del livello di confidenza $1-\alpha$ cresce il valore critico $z_{\alpha/2}$ e quindi aumenta la lunghezza dell'IC

Ampiezza dell'IC /2

- Assumiamo per il momento che l'ampiezza campionaria *n* sia data (vedremo più avanti come scegliere *n* in fase di progettazione dell'indagine), per cui l'unico fattore controllabile che determina l'ampiezza (lunghezza) dell'IC è il livello di confidenza $1-\alpha$
- Di per sé il livello di confidenza dovrebbe essere il più alto possibile; tuttavia vi è un trade-off, perché un aumento del livello di confidenza comporta un incremento della lunghezza dell'IC, cioè una minore precisione
- L'incremento è tanto maggiore quanto più il livello di confidenza si avvicina al 100%: per questo motivo di solito il livello di confidenza viene fissato al 95% (talvolta al 90% o al 99%)
- Ad es. se σ=255.102 e *n*=100, la lunghezza dell'IC è

liv.confidenza	80.0%	90.0%	95.0%	99.0%	99.9%
z critico	1.28	1.64	1.96	2.58	3.29
lunghezza IC	65	84	100	131	168

Quando il livello di confidenza supera il 95% piccoli incrementi comportano notevoli incrementi della lunghezza

Robustezza

- Cosa accade se la distribuzione del carattere nella popolazione non è esattamente Normale?
- Il quesito ha una notevole rilevanza pratica perché l'ipotesi di Normalità non è sempre adeguata (ad es. non lo è nel caso di asimmetria o code pesanti) e inoltre non è facilmente verificabile
- Una proprietà desiderabile per un metodo di stima è la robustezza: un metodo si dice **robusto** quando fornisce risultati approssimativamente validi anche quando alcune delle ipotesi di base sono violate
- Nel caso dell'IC per la media si tratta di valutare se i risultati sono approssimativamente validi anche quando il carattere non ha una distribuzione Normale

Livello effettivo vs livello nominale /1

- Ruolo dell'ipotesi di Normalità: se $X \sim N$ allora la standardizzata della media campionaria ha distribuzione $N(0,1)$ e quindi il valore critico $z_{\alpha/2}$ va letto sulla tavola della $N(0,1)$
- In mancanza di Normalità il valore critico andrebbe letto sulla tavola di un'altra distribuzione → il valore critico della $N(0,1)$ è un numero sbagliato
- La conseguenza è che l'IC ha un ***livello di confidenza effettivo*** diverso dal ***livello di confidenza nominale***
  - Ad es. si vuole costruire un IC al livello 95% (livello nominale), ma in mancanza di Normalità il valore critico 1.96 che si legge sulla tavola della $N(0,1)$ è sbagliato e quindi l'IC ha un diverso livello di confidenza effettivo, cioè la percentuale di campioni per i quali l'IC include la media della popolazione non è 95%. In caso di discrepanza, quasi sempre il livello effettivo è inferiore al livello nominale (es. 93.2%, 88.3%, …)

Livello effettivo vs livello nominale /2

- Nei casi di discrepanza tra livello effettivo e livello nominale accade quasi sempre che il livello **effettivo** sia inferiore a quello **nominale** (= l'IC include il parametro di interesse con una probabilità inferiore a quella programmata)
- In pratica ovviamente non si conosce il valore del parametro di interesse e si dispone di un solo campione → non vi è modo di rilevare la discrepanza tra livello effettivo e livello nominale
- Nella singola applicazione la discrepanza tra livello effettivo e livello nominale potrebbe avere conseguenze trascurabili (infatti l'IC "sbagliato" potrebbe comunque includere il parametro di interesse)
- Tuttavia se il livello effettivo è sostanzialmente inferiore a quello nominale si sta impiegando una procedura tarata male e questo a lungo andare (cioè applicandola ripetutamente) porta a commettere un errore sistematico

Ipotesi di Normalità /1

- Ricordiamo il ruolo dell'ipotesi di Normalità:
  - $X \sim N$ → la standardizzata della media campionaria ha distribuzione $N(0,1)$ → il valore critico $z_{\alpha/2}$ va letto sulla tavola della $N(0,1)$
- In realtà il calcolo del valore critico è basato sulla Normalità della media campionaria ($\bar{X} \sim N$), non sulla Normalità del carattere ($X \sim N$)
- La Normalità del carattere implica la Normalità della media campionaria: tuttavia, se valgono le condizioni del Teorema Limite Centrale (TLC) la distribuzione della media campionaria è approssimativamente Normale qualunque sia la distribuzione del carattere → in tal caso il valore critico $z_{\alpha/2}$ letto sulla tavola della $N(0,1)$ è approssimativamente corretto e quindi il livello di confidenza nominale $(1-\alpha)100\%$ è circa uguale al livello effettivo

Ipotesi di Normalità /2

- Dunque l'IC per μ può essere usato anche quando il carattere ha una **qualunque distribuzione** diversa dalla Normale **purché vi siano le condizioni del TLC** (per caratteri quantitativi: almeno $n=25$ osservazioni, ma già $n=10$ può bastare se la distribuzione è simmetrica unimodale)
- Quando l'ampiezza campionaria n è molto piccola (meno di 10 unità) usare l'IC per μ è rischioso perché:
  1. vi è poca evidenza empirica per verificare se il carattere ha distribuzione Normale, e
  2. le osservazioni sono poche per poter invocare con fiducia il TLC

"Air Traffic Control, I'm going 400 miles an hour and need to land this thing on a floating narrow runway, so I'd prefer something better than 95% confidence in those coordinates."

www.causeweb.org

IC per μ (X~Normale, σ^2 non nota) /1

- Quando si costruisce un IC per la media μ la deviazione std σ non è di diretto interesse, ma è comunque un ingrediente necessario perché entra nell'espressione dell'IC (in questo caso σ è un ***parametro di disturbo***)
- Nella maggior parte delle applicazioni la deviazione std σ non è nota e quindi per poter determinare l'IC per μ occorre rimpiazzare σ con una sua stima
- Per costruire l'IC per la media μ la dev.std σ viene stimata con la dev.std. campionaria

$$S = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2}{n-1}}$$

IC per μ (X~Normale, σ^2 non nota) /2

- Per costruire un IC per la media μ quando la deviazione std σ non è nota si rimpiazza il valore ignoto σ con una sua stima, la deviazione std campionaria *S* (la radice quadrata della varianza campionaria) → l'intervallo aleatorio diventa

$$\bar{X} \pm valorecritico_{\alpha/2} \frac{S}{\sqrt{n}}$$

- Una prima conseguenza della sostituzione di σ con *S* è che la lunghezza dell'IC diviene aleatoria, cioè cambia da campione a campione (la lunghezza può essere determinata solo dopo aver osservato i valori campionari e calcolato *S*)

IC per μ (X~Normale, σ^2 non nota) /3

- Un'altra conseguenza della sostituzione di σ (una quantità fissa) con *S* (uno stimatore, che assume valori diversi a seconda del campione estratto) è l'introduzione di una ulteriore fonte di incertezza → a parità di livello di confidenza l'IC si allunga per tener conto dell'aumentata incertezza

Da un punto di vista tecnico, la media campionaria standardizzata ha distribuzione

Normale standard quando *s* è nota $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

***t* di Student con *n*–1 gdl** quando *s* è ignota e viene sostituita dalla deviazione std campionaria *S* $T = \frac{\bar{X} - \mu}{S / \sqrt{n}} \sim t_{n-1}$

IC per μ (X~Normale, σ^2 non nota) /4

- La ***t di Student*** ha code più pesanti della Normale standard → per ogni data probabilità α da lasciare sulla coda destra il valore critico sulla t è più grande (= spostato verso destra) rispetto alla Normale standard
- La differenza nei valori critici è rilevante quando il numero di gdl è piccolo e tende a zero al crescere del numero di gdl
- Nel caso dell'IC per μ si usa
  - quando σ è nota → valore critico z della Normale standard
  - quando σ non è nota → valore critico t della t di Student con gdl=n–1

Esempio: valori critici che lasciano sulla coda destra α=0.025

n	gdl	*t*	*z*	*t/z*
3	2	4.30	1.96	2.20
5	4	2.78	1.96	1.42
10	9	2.26	1.96	1.15
20	19	2.09	1.96	1.07
30	29	2.05	1.96	1.04
40	39	2.02	1.96	1.03
50	49	2.01	1.96	1.03
75	74	1.99	1.96	1.02
100	99	1.98	1.96	1.01
120	119	1.98	1.96	1.01
200	199	1.97	1.96	1.01
500	499	1.96	1.96	1.00

L'IC per μ è più lungo quando σ non è nota (in quanto il valore critico è più grande: questo riflette l'incertezza addizionale causata dalla necessità di stimare σ); la differenza di lunghezza si riduce al crescere dell'ampiezza campionaria n (infatti quanto più grande è n tanto più lo stimatore S è preciso)

IC per μ (X~Normale, σ^2 non nota) /5

- Dunque, quando $X \sim N(\mu,\sigma^2)$ con μ e σ^2 entrambi ignoti e si dispone di un campione casuale di X di ampiezza n, l'intervallo aleatorio che include μ nel $(1-\alpha)100\%$ dei campioni è

$$\bar{X} \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}$$

- Una volta estratto il campione e calcolate media e deviazione standard, l'intervallo risulta determinato

$$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

Intervallo di confidenza al livello $1-\alpha$

Il valore critico $t_{n-1,\alpha/2}$ è il valore che, nella distribuzione t di Student con $n-1$ gdl, lascia a destra $\alpha/2$ (e a sinistra $1-\alpha/2$)

- Se la distribuzione del carattere è Normale il livello di confidenza nominale $1-\alpha$ è esatto (= coincide con il livello effettivo)
- Anche se la distrib. del carattere non è Normale, di solito con un campione di n=30 il livello nominale $1-\alpha$ è simile al livello effettivo

Esempio

- Per controllare il processo produttivo di una falegnameria vengono esaminate 10 tavole, il cui spessore medio è di 10.05 mm con deviazione standard campionaria di 0.05 mm
- Si assume che lo spessore abbia distribuzione Normale e che le 10 tavole siano un campione casuale
- I gdl sono 10–1=9 → con un livello del 95% il valore critico della t_9 è 2.2622 per cui

$$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} = 10.05 \pm 2.2622 \frac{0.05}{\sqrt{10}}$$
$$= 10.05 \pm 0.036$$
$$= [10.014, 10.086]$$

- Con un elevato livello di fiducia (95%) si può dire che lo spessore medio delle tavole che escono dal processo produttivo è compreso tra 10.014 mm e 10.086 mm

IC per la proporzione /1

- Quando il carattere di interesse è *dicotomico* (X =1 presenza della caratteristica in oggetto, X =0 assenza della caratteristica in oggetto) la sua distribuzione è (necessariamente) di tipo Bernoulli e l'unico parametro è la *probabilità* o *proporzione di successo* p : $X \sim Be(p)$
- Uno stimatore non distorto di p è il suo corrispondente nel campione, cioè la ***proporzione campionaria*** $\hat{P}$

$$\hat{P} = \frac{\text{numero unità campionarie con la caratteristica di interesse}}{\text{ampiezza campionaria}} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$$

Se $n\hat{P}(1-\hat{P}) > 9 \Rightarrow$ per il TLC la distribuzione di $\hat{P}$ è approssimativam. Normale

$$\hat{P} \overset{approx}{\sim} N\left(p, \frac{p(1-p)}{n}\right) \qquad \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} \overset{approx}{\sim} N(0,1)$$

IC per la proporzione /2

$$1-\alpha \approx P\left(-z_{\alpha/2} \le \frac{\hat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} \le z_{\alpha/2}\right) \approx P\left(-z_{\alpha/2} \le \frac{\hat{P}-p}{\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}} \le z_{\alpha/2}\right)$$

≈ perché la standardizzata della proporzione campionaria ha distribuzione *approssimativamente* Normale standard

≈ perché al denominatore il parametro p è stato sostituito dallo stimatore $P^{\wedge}$ (per n grande ciò comporta una differenza trascurabile)

Prendendo l'ultima disuguaglianza a destra e trasformandola in modo che p compaia al centro si ottiene

$$P\left(\hat{P}-z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \le p \le \hat{P}+z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right) \approx 1-\alpha$$

Intervallo aleatorio per p al livello approssimato $(1-\alpha)$

IC per la proporzione /3

- In pratica si estrae un solo campione, sul quale si calcola la proporzione. Pertanto, *dopo l'estrazione del campione*

  la v.a. $\hat{P}$ assume il valore $\hat{p}$ → l'intervallo aleatorio $\hat{P} \pm z_{\alpha/2}\frac{\sqrt{\hat{P}(1-\hat{P})}}{\sqrt{n}}$

  diventa un intervallo di confidenza con estremi determinati

$$\hat{p} \pm z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

Intervallo di confidenza al livello $1-\alpha$

- L'intervallo per p ha la stessa struttura di quello per μ ma con
  - $\hat{p}$ al posto di $\bar{x}$ (media delle variabili campionarie 0-1)
  - $\sqrt{\hat{p}(1-\hat{p})}$ al posto di σ (stima della dev.std. Bernoulli)

Si usa il valore critico z della Normale standard come se la deviazione std fosse è nota a priori, mentre in realtà viene stimata (l'uso di z è un'approssimazione perché per i dati dicotomici non esiste un analogo della t di Student)

Esempio

- In una indagine di mercato 90 persone su 225 intervistate (40%) ricordano la pubblicità di un certo prodotto
- Assumendo che le risposte delle 225 persone intervistate siano un campione casuale, l'IC al livello 95% per la proporzione nella popolazione di persone che ricordano la pubblicità è

$$\hat{p} \pm z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} = 0.4 \pm 1.96\frac{\sqrt{0.4(1-0.4)}}{\sqrt{225}}$$
$$= 0.4 \pm 1.96\frac{0.4899}{15}$$
$$= 0.4 \pm 0.0640$$
$$= [0.3360, 0.4640] \quad \textit{ovvero} \quad [33.60\%, 46.40\%]$$

- Con un elevato livello di fiducia (95%) si può dire che la proporzione nella popolazione di persone che ricordano la pubblicità è compresa tra il 33.60% e il 46.40%

Scelta della dimensione campionaria /1

- Uno dei fattori che influenzano l'ampiezza (lunghezza) di un IC, $A = L_2 - L_1$, è la dimensione campionaria n: all'aumentare di n la lunghezza si riduce
- Dati già raccolti → la dimensione campionaria è determinata e quindi è un fattore non controllabile dall'analista
- Dati non ancora raccolti (fase di progettazione) → la dimensione campionaria è un fattore da determinare
- Uno dei criteri per fissare n è quello di scegliere la lunghezza desiderata dell'IC e determinare il valore di n che porta alla lunghezza desiderata
- Di solito invece che in termini di lunghezza totale dell'intervallo si imposta il problema in termini di lunghezza del semi-intervallo, cioè la distanza tra la stima puntuale e uno degli estremi, detta anche *errore di campionamento* o *margine di errore* e indicata con d
  - Nell'esempio relativo a dati dicotomici si è ottenuto un IC di 0.4 ± 0.064: dunque l'errore di campionamento è $d = 0.064$, mentre la lunghezza totale è $0.064 \times 2 = 0.128$

Scelta della dimensione campionaria /2

Tipo di IC	*Errore di campionamento*
IC per μ, σ nota	$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
IC per μ, σ ignota	$t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$
IC per p	$z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$

L'errore di campionamento ha la struttura

$$(valore\ critico)\times\frac{(deviazione\ std)}{\sqrt{n}}$$

L'ampiezza campionaria n entra come radice quadrata al denominatore

- Il valore critico è determinato dal livello di confidenza $(1-\alpha)$ scelto (in realtà il valore critico $t_{n-1,\alpha/2}$ dipende anche da n ma per semplificare il problema si approssima con $z_{\alpha/2}$ in quanto il valore di n che porta all'errore di campionamento desiderato è usualmente abbastanza elevato da rendere del tutto trascurabile la differenza tra i due valori critici)
- Nel caso dell'IC per μ con σ nota gli altri fattori oltre a n che influenzano l'errore di campionamento sono noti a priori (= prima di estrarre il campione)

Scelta della dimensione campionaria /3

- Nel caso dell'IC per μ con σ ignota, per calcolare la dimensione campionaria n occorre ipotizzare la dev.std. campionaria s
  - Per pianificare un'indagine su larga scala talvolta si effettua un'indagine preliminare, detta *indagine pilota*, che fornisce una stima di s
- Nel caso dell'IC per p per calcolare la dimensione campionaria n occorre ipotizzare la proporzione campionaria (o stimarla con un'indagine pilota)
  - Ipotizzare $\hat{p}=0.5$ corrisponde allo *scenario peggiore* (ricorda: la dev.std. di una v.a. Bernoulli è massima quando la probabilità di successo è al 50%)
  - Il valore di n determinato nello scenario peggiore garantisce che la lunghezza dell'IC non superi in nessun caso quella desiderata
  - Calcolare n nello scenario peggiore non è una buona idea quando tale scenario è irrealistico (cioè quando si sa che la proporzione è ben lontana dal 50%): significa selezionare un campione molto più grande del necessario (spreco di risorse)

Scelta della dimensione campionaria /4

- Fissato un livello di confidenza $(1-\alpha)$ e scelto l'errore di campionamento desiderato d, il valore di n si ottiene risolvendo la seguente espressione

$$d=(valore\ critico)\times\frac{(deviazione\ std)}{\sqrt{n}}$$

da cui si ottiene

$$n=(valore\ critico)^2\times\frac{(deviazione\ std)^2}{d^2}$$

(da approssimare con il numero intero successivo)

Esempi

$$n=(valore\ critico)^2\times\frac{(deviazione\ std)^2}{d^2}$$

- livello di confidenza 95% → valore critico = 1.96
- Esempio: il carattere X è un peso espresso in grammi e si stima (o si ipotizza) che la deviazione std sia 15 grammi; allora la dimensione campionaria n che porta ad un IC di ±3 grammi intorno alla media campionaria (cioè d=3) è

$$n=1.96^2\times\frac{15^2}{3^2}=96.04\rightarrow 97$$

- Esempio: il carattere X è l'indicatore 0-1 della caratteristica "cliente soddisfatto" e si stima (o si ipotizza) che p sia 0.75 per cui la deviazione std è 0.4330; allora la dimensione campionaria n che porta ad un IC di ±5 punti percentuali intorno alla proporzione campionaria (cioè d=0.05) è

$$n=1.96^2\times\frac{0.4330^2}{0.05^2}=288.12\rightarrow 289$$

IC per la varianza (X~Normale)

- Se la popolazione generatrice è Normale, la varianza campionaria ha una distribuzione campionaria legata alla Chi-quadrato

$$X \sim N(\mu,\sigma^2) \quad X_1,\ldots,X_n \text{ iid-}X$$

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

$$\Rightarrow \quad V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

da cui segue

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

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

Attenzione: a differenza della Normale e della *t* di Student, il valore critico *inferiore* della Chi-quadrato non è il negativo del valore critico *superiore*

Esempio

- Obiettivo: valutare le variazioni di temperatura di una caldaia tramite un IC al 95% (α =0.05)
- Campione casuale di n=25 temperature, varianza campionaria s^2=100
- Si assume che X abbia distribuzione Normale

$$P\left(\frac{24\times100}{\chi^2_{24,0.025}} < \sigma^2 < \frac{24\times100}{\chi^2_{24,0.975}}\right) = 0.95$$

39.36 — 12.40

$$P\left(60.97 < \sigma^2 < 193.53\right) = 0.95$$

Attenzione: l'IC per la varianza è poco robusto rispetto alla violazione dell'ipotesi di Normalità → prima di riportare l'IC è bene verificare la Normalità della distribuzione delle temperature