Statistica inferenziale

Tuesday, 2 May 2023 13:24

Questa lezione è un parto. E' da ore che cerco di seguire questa lezione, però pecco sempre di concentrazione siccome sta prof spiega da far schifo.

- Introduzione
  - Da tanti dati noi vogliamo cercare di fare delle deduzioni plausibili
    Ed affinchè i dati sono plausibili:
    - I campioni devono essere selezionati in modo tale che racchiude tutte le categorie in maniera equa → Bisogna estrarre in o modo casuale oppure pescare da tutti i possibili sottogruppi
      Ed i campioni o sono infiniti, oppure estremamente molto grande.
- Stimatori
  - Sono delle variabili aleatorie indipendenti che assegniamo ad un gruppo di dati
    E tutte le variabili aleatorie hanno la stessa legge dipendenti ad un qualcosa, il parametro. Noi vogliamo fare le stime di questo parametro
  - Sono particolare funzioni del campione che ci servono per calcolare incognite
    Es. La media ha come stimatore la media campionaria
  - Devono essere:
    - Corretti
    - Non distorto
      Quando è non distorto può diventare consistente quando, all'aumentare di N ad infinito il nostro stimatore si avvicina sempre di più al suo valore originale

    E lo sono quando il suo valore medio è esattamente uguale al valore che vado a stimare, e la varianza campionaria è uno stimatore non distorto della media
    Se la nostra stima è Θ allora lo stimatore è $\widehat{\Theta}$
  - Statistica = Funzione del campione
    - Media campionaria
    - Varianza campionaria
  - Penso serviranno
    - $\sum \left(\frac{x_i - m}{\sigma}\right)^2 \sim X^2(n)$

- $\sum \left(\frac{x_i - \bar{x}_n}{o}\right)^2 \sim X^2(n-1)$

- Nuova v.a. T strudel (?)

  $$T = \frac{Z}{\sqrt{\frac{Y}{n}}} \to \begin{matrix} Z \sim N(0,1) \\ Y \sim X^2(n) \end{matrix}$$

  $E(t) = 0, Var[T] = 1$

  E la sua densità

  $$f_T(t) = c_n \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

  Il suo grafico è molto simile ad una $N(0,1)$ e per n grandi diventa una normale

  E detto questo, formula importante

  Dato $X_1 \dots X_n$ campione casuale estratto da $N(m, o^2)$

  $$\frac{\overline{x_n} - m}{\sqrt{\frac{s_n^2}{n}}} \sim t(n-1)$$

  E di questo ci importa calcolare i percentili

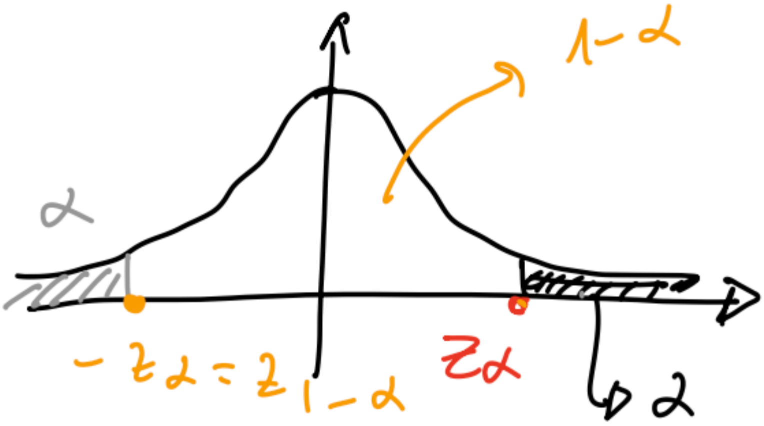
  $\alpha \in (0,1)$

  $P(X \leq q_\alpha) = \alpha$

  La probabilità che $P(X \leq$un valore) equivale a α

  Quindi, q_α che ora chiameremo z_α siccome la prof ha deciso così

  E' il punto nelle ascisse in cui dopo abbiamo alpha

  Ed ora diciamo che

  $P(Z > z_\alpha) = \alpha \to P(Z \leq z_\alpha) = 1 - \alpha$

  In qualche modo è possibile fare andare di la P

  $z_\alpha = 100(1-\alpha)$

  Ed in qualche modo a me ignoto

$z_\alpha = \phi^{-1}(1-\alpha)$
E si nota che $z_{1-\alpha} = -z_\alpha$
Quindi...
Supponiamo che dobbiamo trovare $\alpha = 0.025$
$P(Z < z_{0.025}) = \alpha$
$z_{0.025} = \phi^{-1}(1-0.025) = \phi^{-1}(0.975)$
Quindi ora bisogna trovare ϕ inverso
Aka dobbiamo trovare il valore nella tavola, e poi trovare con chi incrocia

z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750

Possiamo notare che 0.975 incrocia con 1.9 e 0.06
Quindi
$z_{0.025} = \phi^{-1}(0.975) = 1.96$
Legge: $z_\alpha = -z_{1-\alpha}$
$t_{\alpha,n} = -t_{1-\alpha,n}$

Per T strudel, usiamo la tabella T
Per X^2 idem la sua tabella

- Stima per intervalli / Intervalli di confidenza
  Abbiamo un campione $N(m, o^2)$
  Da ciò che avevamo detto ricordiamo:
  - $\overline{x_n} \to m \qquad n \to +\infty$

- $var(\overline{x_n}) = \frac{o^2}{n}, \qquad SD(\overline{x_n})$

Noi vogliamo costruire un intervallo aleatorio dentro cui la media cade con una probabilità alta α, e questo intervallo deve essere simmetrico: $(-\alpha, +\alpha)$

Ora da questo ne esce questa formula che date per buono:

$$P\left(\frac{|\overline{x_n} - m|}{\frac{o}{\sqrt{n}}} < \frac{E}{\frac{o}{\sqrt{n}}}\right) = 1 - \alpha$$

E qui noi abbiamo una N(0, 1), why?

$$\frac{\overline{x_n} - m}{\frac{o}{\sqrt{n}}} \sim N(0,1)$$

E quindi possiamo dire

$$\frac{|\overline{x_n} - m|}{\frac{o}{\sqrt{n}}} = |z|$$

$$P\left(|z| < \frac{E}{\frac{o}{\sqrt{n}}}\right) = 1 - \alpha$$

E siccome abbiamo $|z|$

Quando noi vogliamo trovare solamente z, per via delle regole di simmetria dobbiamo dividere per 2

$$P\left(z > \frac{E}{\frac{o}{\sqrt{n}}}\right) = \frac{\alpha}{2}$$

Ed a quanto pare abbiamo anche fatto il complementare (?)

E da questo ne esce

$$\frac{E}{\frac{o}{\sqrt{n}}} = z_{\frac{\alpha}{2}} \to E = z_{\frac{\alpha}{2}} * \frac{o}{\sqrt{n}}$$

E da qui possiamo sostituire

$E = |X_n - m|$

(A quanto pare questa è una formula?)

E ne esce

$$P\left(|X_n - m| < z_{\frac{\alpha}{2}} * \frac{o}{\sqrt{n}}\right) = 1 - \alpha$$

E quindi

Ora possiamo scrivere il nostro intervallo come

$$P\left(m \in \left(\overline{X_n} - z_{\frac{\alpha}{2}} * \frac{o}{\sqrt{n}}, \overline{X_n} + z_{\frac{\alpha}{2}} * \frac{o}{\sqrt{n}}\right)\right)$$

E con questo il nostro livello di confidenza è $100(1 - \alpha)\%$

E nel caso non si fosse capito, la confidenza è la probabilità con cui il parametro appartiene al parametro prima di fare le osservazioni

- Nota importante:

  $$z_{\frac{\alpha}{2}} * \frac{o}{\sqrt{n}} = errore$$

  E' possibile comprendere meglio tutto quanto leggendo gli esercizi pratici sotto

- Estremi inferiori e superiori di confidenza

  Noi abbiamo questa formula

  $$P\left(\frac{\overline{x_n} - m}{\frac{o}{\sqrt{n}}} < z_\alpha\right) = 1 - \alpha$$

  Riscriviamo in funzione di m

  $$P\left(m > \overline{x_n} - z_\alpha * \frac{o}{\sqrt{n}}\right) = 1 - \alpha$$

  E questo è l'estremo inferiore di confidenza al $100(1-\alpha)\%$ per la media di una popolazione normale con **varianza nota**

  Non c'ho capito un cazzo nemmeno io, lo prendo per buono

- Ora invece abbiamo la **varianza incognita** con campione normale

Allora, noi abbiamo

$$z = \frac{\overline{x_n} - m}{\frac{o}{\sqrt{n}}} = \frac{\overline{x_n} - m}{o}\sqrt{n} \sim N(0, 1)$$

Però a noi ci manca o

Ed ora...

$$T_{n-1} = \frac{\overline{x_n} - m}{\sqrt{S_n^2}} * \sqrt{n} \sim t(n-1)$$

In qualche modo ora i diventata una T di strudel

E grazie a questo il tutto diventa

$$P_{m,o}\left(\frac{|\overline{x_n} - m|}{\sqrt{S_n^2}} * \sqrt{n} < t_{n-1,\frac{\alpha}{2}}\right) = 1 - \alpha$$

E quindi l'intervallo...

$$= \left(\overline{x_n} \pm t_{n-1,\frac{\alpha}{2}}\sqrt{\frac{S_n^2}{n}}\right)$$

$+$ Confidenza, E aumenta però la stima è più affidabile

- Estremo superiore di confidenza

  $$\overline{x_n} + t_{n-1,\alpha} * \frac{S_n}{\sqrt{n}} : P\left(m < \overline{X_n} + t_{n-1,\alpha} * S_n\right)$$

  E noi qui diciamo che

$S_n = \sqrt{S_n^2}$

E gli intervalli di confidenza:

$$\left(-\infty, \overline{x_n} + t_{n-1,\alpha} * \frac{S_n}{\sqrt{n}}\right) u \left(\overline{x_n} - t_{n-1,\alpha} * \frac{S_n}{\sqrt{n}}, +\infty\right)$$

- Stime per grandi campioni

$$\frac{\overline{x_n} - m}{o} \sqrt{n} \sim N(0,1)$$

$$\frac{\overline{X_n} - m}{\sqrt{S_n^2}} \sqrt{n} \sim T_{n-1}$$

- Ricordi

  Se $X \sim Be(p) \to E(X) = p$

  $X_1 \dots X_n \sim Be(p)$

  $\overline{X_n}$ stimatore non distorto di p

  $\overline{x_n} = \hat{p}$ stima di p

  Questo viene usato se un certo carattere c'è o non c'è -> soddisfatto?

  Siccome p spesso è incognito, noi lo possiamo sostituire con la sua stima, cioè la media campionaria

  E quindi

$$\frac{\overline{X_n} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

$$Ed\ X_n \sim N\left(p, \frac{p(1-p)}{n}\right)$$

$$p(1-p) = \hat{p}(1-\hat{o}) = \overline{x_n}(1-\overline{x_n})$$

  Noi quindi possiamo interscambiare $\hat{p}, \overline{x_n}$

  E quindi, prendendo l'intervallo

$$\overline{X_n} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{x_n}(1-\overline{x_n})}{n}} = \hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

  Quindi, con questo possiamo risolvere l'ultimo esercizio

Esempio pratico:

- In un liceo

  100 studenti

  40 sono ragazze

  Fronire una stima delle proporzioni di ragazze frequentanti usando un stimatore non distorto.

  Siccome abbiamo ragazza-non ragazza

  $Be(p)$

La stima delle proporzioni si può usare usando la media, e la media è uno stimatore non distorto

$\hat{p} = \frac{40}{100} = \frac{2}{5}$

Ed ora dobbiamo fornire una stima della varianza del campione bernulliana

E siccome lo stimatore non è distorto, la stima non è distorta

$Var(X_i) = p(1-p)$

Questo siccome $X_i \sim Be(p)$

$o^2 = \hat{p}(1-\hat{p}) = \frac{2}{5} * \frac{3}{5} = \frac{6}{25}$

E qui ho usato $\overline{x_n}(1-\overline{x_n})$

- La statura dei piloti è distribuita secondo una normale $N(m, o^2)$

  Si vuole stimare m a seconda di 100 piloti

  E supponiamo o=0.1cm

  Aka $o^2 = 0.1^2$

  La media rilevata è 178.5

  $\overline{x_n} = \widehat{m} = 178.5$

  Iniziamo a calcolare il livello di confidenza

  $\left(\overline{x_n} - z_{\frac{\alpha}{2}} * \frac{o}{\sqrt{n}}, \overline{x_n} + z_{\frac{\alpha}{2}} * \frac{o}{\sqrt{n}}\right)$

  Iniziamo a trovare i valori

  $95\% = 100(1-\alpha)\% \rightarrow \alpha = 0.05$

  $z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = z_{0.025} = \phi^{-1}(1-\alpha) = \phi^{-1}(0.975) = 1.96$

  Quindi ora possiamo sostituire

  $\left(178.5 - 1.96 * \frac{0.1}{10}, 178.5 + 1.96 * \frac{0.1}{10}\right) = (177.3, 179.7)$

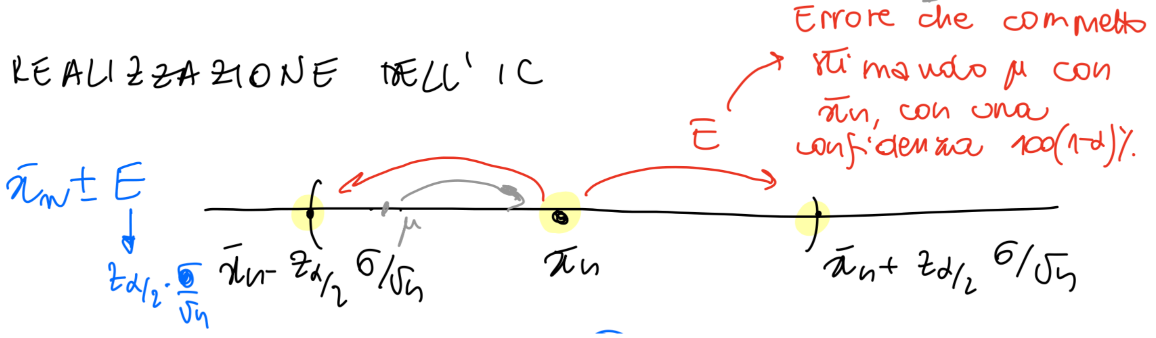
  Nota: per dimezzare l'ampiezza dobbiamo quadruplicare il campione

  $90\% \rightarrow \alpha = 0.1 \rightarrow z_{\frac{0.1}{2}} = 1.645$

  $95\% \rightarrow \alpha = 0.05 \rightarrow z_{\frac{0.5}{2}} = 1.960$

  $99\% \rightarrow \alpha = 0.01 \rightarrow z_{\frac{0.01}{2}} = 2.576$

  Graficamente l'intervallo esce così:



Ampiezza intervallo $= 2\mathrm{E} = 2z_{\frac{\alpha}{2}} * \frac{o}{\sqrt{n}}$

- E cresce se α diminuisce
- E diminuisce al crescere di n con α, o fissati

Per far diminuire di fattore $\frac{1}{2}$ dobbiamo moltiplicare n*4

- Vogliamo un grado di confisenza del 99%

Dati:

N=100

$\overline{x_n} = 178.5$

$o = 0.1$

Esecuzione:

$1 - \alpha = 0.99 \rightarrow \alpha = 0.01$

$z_{\frac{\alpha}{2}} = z_{0.005} = 2.578$

Intervallo:

$$\left(\overline{x_n} + z_{\frac{\alpha}{2}} * \frac{o}{\sqrt{n}}\right) = (176.9, 180.1)$$

Rifare i calcoli con n=500 (aka ora ho estratto 500 piloti)

Intervallo:

$$\left(\overline{x_n} + z_{\frac{\alpha}{2}} * \frac{o}{\sqrt{n}}\right) = \left(179.5 \pm 1.96 * \frac{0.1}{\sqrt{500}}\right) \simeq (177.9, 179.1)$$

Ora vogliamo trovare un certo n affinchè l'erroer sia uguale ad un certo E_o assegnato

$$z_{\frac{\alpha}{2}} * \frac{o}{\sqrt{n}} \rightarrow Errore = E_o$$

$$z_{\frac{\alpha}{2}} * \frac{o}{\sqrt{n}} = E_o \rightarrow \sqrt{n} = \frac{oz_{\frac{\alpha}{2}}}{E_o} = n = \left(\frac{oz_{\frac{\alpha}{2}}}{E_o}\right)^2$$

- Si vuole trovare n tale che $E \leq E_o, E_o$ assegnato

$$E = \sqrt{\frac{\overline{x_n}(1 - \overline{x_n})}{n}} z_{\frac{\alpha}{2}} \leq \frac{\sqrt{\frac{1}{4}}}{\sqrt{n}} = z_{\frac{\alpha}{2}} = \frac{1}{2\sqrt{n}} z_{\frac{\alpha}{2}}$$

E quindi ora possiamo ovviare il problema di $\overline{x_n}$

$$\frac{1}{2\sqrt{n}} z_{\frac{\alpha}{2}} \leq E_0 \rightarrow n \geq \left(\frac{z_{\frac{\alpha}{2}}}{2E_o}\right)^2$$

- Campione1 $= 130$

Favorevoli1 $= 75$

Campione2 $= 1056$

Favorevoli2 = 642

$Favorevoli2 = 642$

a. Costruire un intervallo C. al livello del 95% per la proporzione di elettori favorevoli

   Iniziamo a controllare che possiamo trasformarlo in una normale:
   $n \geq 30 \rightarrow si$
   $np > 5 \rightarrow si$
   Quindi possiamo continuare

   Siccome vogliamo grado di confidenza del 95%
   $1 - \alpha = 0.95 \rightarrow \alpha = 0.05$

   Primo campione:
   $$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{75}{130} \pm z_{0.025} \sqrt{\frac{\frac{75}{130} * \frac{55}{130}}{130}} = 0.5769 \pm 0.084$$
   $Ic \simeq (0.492, 0.662)$

   Secondo campione
   $$= \frac{642}{1056} \pm 1.96 \sqrt{\frac{\frac{642}{1056} * \frac{414}{1056}}{1056}}$$
   $Ic \simeq (0.578, 0.6384)$

b. Confrontare la precisione delle stime effettuate
   Per confrontare la precisione ci serve prima di tutto l'errore
   $Ic_1 \simeq (0.492, 0.662) \rightarrow 2E = (0.662 - 0.492) = 0.17$
   $Ic_2 \simeq (0.578, 0.638) \rightarrow 2E = 0.0589$

   Quindiiii
   Notiamo che il primo caso è nettamente molto meno preciso del secondo
   17% vs $5{,}89\%$

c. Noi ora vogliamo una ampiezza non superiore all'1%
   Aka E=0.05
   (Ricordo che l'ampiezza è 2E)

   Usando le formule di sopra che non ho compreso
   $$E \leq z_{\frac{\alpha}{2}} * \frac{1}{2\sqrt{n}}$$
   Noi vogliamo trovare una n tale che $E \leq E_0$