Esercitazione 3 - Statistica (parte II)

Davide Passaretti

23/2/2017

Contents

1 Intervalli di confidenza

L'accezione frequentista (cioè classica) di intervallo di confidenza si basa sul *principio del campionamento ripetuto*. Secondo tale principio, un intervallo per un parametro θ con un livello di confidenza $1-\alpha$ indica che, qualora costruissimo lo stesso tipo di intervallo su un numero elevato di campioni, la percentuale di intervalli che conterrebbero θ sarebbe proprio $1-\alpha$. Dunque, va ricordato che:

- un intervallo di confidenza **si costruisce per un parametro della popolazione**. L'obiettivo è ottenere una stima di ciò che esiste nella popolazione, non di ciò che osservo nel mio campione;
- il livello di confidenza $1-\alpha$ **NON** indica la probabilità che il parametro sia compreso nell'intervallo, in quanto il parametro è una costante ignota appartenente alla popolazione, non una variabile casuale;
- dovendoci basare sul principio del campionamento ripetuto, non possiamo dire nulla riguardo l'intervallo che calcoliamo sul campione a disposizione, ma possiamo riporre fiducia nel metodo utilizzato per la costruzione di tale intervallo.

2 Intervalli su un campione

Gli intervalli di confidenza su un campione riguardano la stima intervallare di un parametro della popolazione (es: media, varianza o proporzione). Non essendovi alcun obiettivo di confronto tra parametri di due (o più) popolazioni, ci si avvarrà di un solo campione per la stima.

2.1 Intervallo di confidenza per la media

La media μ di una popolazione Normale è stimata tramite la media campionaria $\bar{x}$, seguendo il principio di naturalità. Se tale popolazione non è distribuita normalmente ma si ha a disposizione un campione non troppo piccolo ($n > 30$ in genere), si può sfruttare il **TLC** per la stima intervallare. Ciò è dovuto al fatto che la media campionaria è una combinazione di variabili casuali che converge asintoticamente ad una Normale.

2.1.1 Varianza nota

Nel caso σ^2 sia nota, allora la varianza della media campionaria è ottenibile come $\frac{\sigma^2}{n}$.
L'intervallo per μ è dunque il seguente:

$$\mathbf{IC}_{1-\alpha}(\mu) = \bar{x} \pm \underbrace{\mathcal{Z}_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}_{\text{margine di errore}}$$

Facciamo un esempio. Il seguente campione contiene i ricavi mensili di 9 imprese, distribuiti normalmente con varianza nota $\sigma^2 = 302500$:

39412.60	40654.48	39754.66	39756.74	39380.00	40299.44	39221.09	39350.41	39713.18

Calcolare l'intervallo di confidenza del 95% per la media.

La media campionaria è:

$$\bar{x} = \frac{1}{9} \sum_{i=1}^{9} x_i = 39726.96$$

I percentili della Normale standardizzata da utilizzare sono quelli che lasciano il 2.5% nella coda sinistra e il 2.5% nella coda destra. Essendo tali percentili uno l'opposto dell'altro, è sufficiente trovarne uno. Dovendo lavorare con la tavola della $\mathcal{Z}$, ci conviene il percentile destro, cioè quello corrispondente a $1 - \frac{\alpha}{2}$. In questo caso, dobbiamo cercare $\mathcal{Z}_{1-\frac{0.05}{2}} = \mathcal{Z}_{0.975} = 1.96$.

L'intervallo di confidenza per μ è il seguente:

$$\mathbf{IC}_{0.95}(\mu) = 39726.96 \pm 1.96 \frac{\sqrt{302500}}{\sqrt{9}} = 39726.96 \pm 359.33 = [39367.63, 40086.28]$$

2.1.2 Varianza incognita

Supponiamo ora di non avere conoscenza della varianza della popolazione. Dobbiamo dunque stimarla usando lo stimatore $s^2 = \frac{1}{n-1} \sum_i^n (x_i - \bar{x})^2$. Dobbiamo ricordare che, se $X_1, X_2, \ldots, X_n$ sono indipendenti e distribuite normalmente con stessa media e stessa varianza, allora:

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

Come conseguenza di ciò, la distribuzione di riferimento per la costruzione dell'intervallo non è più Normale, ma è una t di *Student* con $n-1$ gradi di libertà. Dunque l'intervallo è il seguente:

$$\mathbf{IC}_{1-\alpha}(\mu) = \bar{x} \pm \underbrace{t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}}_{\text{margine di errore}}$$

Ritorniamo all'esercizio con l'obiettivo di calcolare un intervallo di confidenza del 95% per la media dei ricavi. Ci stimiamo la varianza:

$$s^2 = \frac{1}{9-1}\sum_{i}^{9}(x_i - 39726.96)^2 = 225846.1$$

L'intervallo di confidenza per la media dei ricavi è il seguente:

$$\mathbf{IC}_{0.95}(\mu) = 39726.96 \pm 2.31\frac{\sqrt{225846.1}}{\sqrt{9}} = 39726.96 \pm 365.3 = [39361.66, 40092.25]$$

2.2 Intervallo di confidenza per la proporzione

La proporzione π di successi nella popolazione è stimata dalla proporzione di successi osservata nel campione, seguendo il principio di naturalità. Lo stimatore di π è dunque $\hat{\pi} = \mathscr{F}_n = \frac{\mathcal{Y}}{n}$, dove $\mathcal{Y} \sim \mathbf{Bin}(n, \pi)$.
Ne consegue che:

$$E(\mathscr{F}_n) = \frac{1}{n} \times n\pi = \underbrace{\pi}_{E(\mathbf{Ber}(\pi))}, \qquad Var(\mathscr{F}_n) = \frac{1}{n^2} \times n\pi(1-\pi) = \frac{1}{n} \times \underbrace{\pi(1-\pi)}_{Var(\mathbf{Ber}(\pi))}$$

Essendo $\mathscr{F}_n$ una combinazione lineare di variabili casuali identicamente distribuite (una media di v.c. Bernoulliane, nella fattispecie), tale distribuzione converge ad una Normale per il **TLC**. Per questo motivo, possiamo costruire un intervallo di confidenza per π sfruttando i percentili della v.c. Normale standardizzata. Essendo la varianza di $\mathscr{F}_n$ dipendente dal parametro ignoto π, ci è ignota anch'essa e viene stimata dal campione nel modo più naturale possibile:

$$\hat{Var}(\mathscr{F}_n) = \frac{\hat{\pi}(1-\hat{\pi})}{n}$$

L'intervallo di confidenza per π è quindi il seguente:

$$\mathbf{IC}_{1-\alpha}(\pi) \xrightarrow{\mathbf{TLC}} \hat{\pi} \pm \underbrace{\mathcal{Z}_{\frac{\alpha}{2}}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}}_{\text{margine di errore}}$$

Vediamo un esempio. In un sondaggio organizzato da un piccolo ente, si osservano 26 voti favorevoli in un campione casuale di 100 unità. **Calcolare l'intervallo di confidenza del 95% per la proporzione.**

$$\hat{\pi} = \frac{26}{100} = 0.26$$

$$\hat{Var}(\mathscr{F}_n) = \frac{0.26(1-0.26)}{100} = 0.001924$$

$$\mathbf{IC}_{0.95}(\pi) \xrightarrow{\mathbf{TLC}} 0.26 \pm 1.96\sqrt{0.001924} = 0.26 \pm 0.086 = [0.174, 0.346]$$

2.3 Determinazione della numerosità campionaria per μ e π

Per la determinazione della numerosità campionaria al fine di costruire intervalli per μ con una prefissata ampiezza, bisogna conoscere il margine di errore Δ (ovvero la semiampiezza dell'intervallo desiderato), il livello di confidenza $1 - \alpha$ e la varianza della popolazione σ^2.

$$\tilde{n}_\mu \geq \frac{\mathcal{Z}^2_{\frac{\alpha}{2}} \times \sigma^2}{\Delta^2}$$

Qualora il parametro oggetto di stima sia π, può essere sufficiente conoscere solo Δ e $1 - \alpha$, in quanto la varianza della Bernoulli $\pi(1 - \pi)$ può essere stimata pessimisticamente assegnando a π il valore di maggiore incertezza, cioè 0.5. La stima più pessimistica della varianza della Bernoulli è quindi 0.5(1 - 0.5) = 0.25.

$$\tilde{n}_\pi \geq \frac{\mathcal{Z}^2_{\frac{\alpha}{2}} \times 0.25}{\Delta^2}$$

Facciamo qualche esercizio. Prendiamo ancora la popolazione di imprese che hanno ricavi mensili distribuiti normalmente con varianza 302500 (esercizio 2.1.1) e vogliamo un intervallo di confidenza del 95% per la media dei ricavi.

a) Vogliamo determinare la numerosità campionaria minima che possa assicurare un margine di errore non superiore a 200 euro.

$$\tilde{n}_\mu \geq \frac{1.96^2 \times 302500}{200^2} = 29.05$$

Dunque necessitiamo di un campione di almeno 30 imprese per ottenere un intervallo di confidenza del 95% di ampiezza non superiore a 400 (cioè $2 \times \Delta$).

b) Se volessimo un intervallo di confidenza del 90% con lo stesso margine di errore, avremmo bisogno di un campione più ampio o più esiguo?

Con un livello di confidenza minore, ma la medesima ampiezza, possiamo prendere un campione più esiguo di imprese. Infatti:

$$\tilde{n}_\mu \geq \frac{1.645^2 \times 302500}{200^2} = 20.46$$

c) Se volessimo un intervallo di confidenza del 95% con un margine di errore non superiore ai 100 euro, avremmo bisogno di un campione più ampio o più esiguo?

Con il medesimo livello di confidenza, ma un'ampiezza desiderata minore (cioè maggiore accuratezza della stima intervallare), dobbiamo prendere un campione più ampio di imprese. Infatti:

$$\tilde{n}_\mu \geq \frac{1.96^2 \times 302500}{100^2} = 116.20$$

Poniamo ora il caso che vogliamo determinare la proporzione delle imprese collocate nel settore abbigliamento.

d) Vogliamo costruire un intervallo di confidenza del 99% per tale proporzione con un margine di errore non superiore a 0.07. Quanto dev'essere grande il campione?

Non avendo modo di risalire alla varianza della Bernoulli, né a una sua stima, useremo la stima pessimistica uguale a 0.25. Ricordiamo che il percentile da trovare sulla tavola è $\mathcal{Z}_{0.995}$.

$$\tilde{n}_\pi \geq \frac{2.58^2 \times 0.25}{0.07^2} = 338.51$$

2.4 Intervallo di confidenza per la varianza

Sia $X \sim \mathbf{N}(\mu, \sigma^2)$. Estraiamo un campione di numerosità n. Vogliamo costruire un intervallo di confidenza (di livello $1-\alpha$) per σ^2 nello scenario più realistico in cui ignoriamo μ, che quindi viene ad essere un *nuisance parameter* (parametro di disturbo). Sappiamo che lo stimatore di σ^2 è la varianza campionaria corretta s^2.

È possibile dimostrare che:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

Ne consegue (dopo un po' di passaggi algebrici) che:

$$\mathbf{IC}_{1-\alpha}(\sigma^2) = \left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},\, n-1}},\, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},\, n-1}}\right]$$

<u>**NB**</u>: si sta ragionando sull'area a destra del percentile considerato (in virtù delle tavole): per esempio, $\chi^2_{0.05,\, 10}$ è il percentile della v.c. Chi-quadro con 10 gradi di libertà che ha a destra un'area del 5%.

Riprendiamo il caso delle 9 imprese del paragrafo 2.1.
Abbiamo già calcolato la stima s^2, che è risultata uguale a 225846.1.

Costruire l'intervallo di confidenza del 90% per σ^2.

I percentili da considerare sono $\chi^2_{0.05,\, 9-1} = 15.507$ e $\chi^2_{0.95,\, 9-1} = 2.732$. L'intervallo richiesto è il seguente:

$$\mathbf{IC}_{0.90}(\sigma^2) = \left[\frac{(9-1) \times 225846.1}{15.507},\, \frac{(9-1) \times 225846.1}{2.732}\right] = [116510.8,\, 661181.5]$$

3 Intervalli su due campioni

Gli intervalli di confidenza su due campioni hanno come obiettivo il confronto tra i parametri di due popolazioni indipendenti. Date queste due popolazioni, X_A e X_B, estraiamo un campione di numerosità n_A da X_A ed un campione di numerosità n_B da X_B. Gli intervalli che prenderemo in considerazione sono:

- intervallo di confidenza per la differenza tra medie,
- intervallo di confidenza per la differenza tra proporzioni,
- intervallo di confidenza per il rapporto tra varianze.

3.1 Intervallo di confidenza per il confronto tra medie

Assumiamo $X_A \sim \mathbf{N}(\mu_A, \sigma^2_A)$ e $X_B \sim \mathbf{N}(\mu_B, \sigma^2_B)$.

La variabile casuale differenza tra le medie $\bar{X}_A$ e $\bar{X}_B$ è ancora Normale:

$$\bar{X}_A - \bar{X}_B \sim \mathbf{N}\left(\mu_A - \mu_B,\, \frac{\sigma^2_A}{n_A} + \frac{\sigma^2_B}{n_B}\right)$$

3.1.1 Varianze note

Se abbiamo informazione sulle due varianze, ci basta calcolare le due medie campionarie $\bar{x}_A$ e $\bar{x}_B$.
L'intervallo di confidenza per la differenza tra medie è il seguente:

$$\mathbf{IC}_{1-\alpha}(\mu_A - \mu_B) = (\bar{x}_A - \bar{x}_B) \pm \underbrace{\mathcal{Z}_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}_{\text{margine di errore}}$$

Vediamo un caso pratico. Vogliamo confrontare la media delle vendite giornaliere di due piccoli supermercati (A e B) tramite una stima intervallare costruita su un campione di 90 giorni per A e di 30 giorni per B (i giorni sono scelti casualmente nell'arco di un anno; quelli scelti per A sono indipendenti da quelli scelti per B e viceversa). Le varianze sono note: $\sigma_A^2 = 3600$, $\sigma_B^2 = 3900$. Le medie campionarie sono rispettivamente $x_A = 500$ € e $x_B = 600$ €. Vogliamo utilizzare un livello di confidenza del 90 %, dunque cerchiamo $\mathcal{Z}_{0.95}$.

$$\mathbf{IC}_{0.90}(\mu_A - \mu_B) = (500 - 600) \pm 1.645 \sqrt{\frac{3600}{90} + \frac{3900}{30}} = -100 \pm 21.45 = [-121.45, -78.45]$$

3.1.2 Varianze incognite ma supposte uguali

Nel caso in cui non avessimo conoscenza delle varianze delle due popolazioni, dovremmo stimarle dal campione. La stima di σ_A^2 è ovviamente la varianza campionaria corretta calcolata sul campione di X_A (che indichiamo con s_A^2). Analogamente, la stima di σ_B^2 è s_B^2. Dovendo stimare due varianze, la Normale diviene una t di *Student*. Sostituendo al posto delle varianze le loro stime, avremmo che la varianza stimata della differenza tra medie è:

$$\hat{Var}(\bar{X}_A - \bar{X}_B) = \hat{Var}(X_A) + \hat{Var}(X_B) = \frac{s_A^2}{n_A} + \frac{s_B^2}{n_B} \tag{1}$$

Tuttavia, questa stima rende complicato il calcolo dei gradi di libertà della t di *Student*. Per facilitare tale calcolo, assumiamo $\sigma_A^2 = \sigma_B^2 = \sigma^2$. In seguito a questa assunzione, abbiamo che:

- i gradi di libertà della t sono $n_A + n_B - 2$,
- la stima della varianza della differenza tra medie diviene:

  $$s_p^2 \left(\frac{1}{n_A} + \frac{1}{n_B} \right)$$

  dove s_p^2 è uno stimatore *pooled* (congiunto) di σ^2 e si ottiene come media pesata di s_A^2 e s_B^2:

  $$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$$

Supponiamo per esempio di non conoscere le varianze delle vendite giornaliere dei due supermercati. Assumiamo intanto che tali varianze siano uguali a σ^2. Dal campione di 90 giorni del supermercato A, otteniamo $s_A^2 = 3969$; dal campione di 30 giorni di B, otteniamo $s_B^2 = 4200$.
Qual è l'intervallo di confidenza del 90% per la differenza tra medie?

Il percentile di riferimento è $t_{0.95,\,90+30-2} = t_{0.95,\,118} \approx \mathcal{Z}_{0.95} = 1.645$. Dobbiamo calcolare s_p^2:

$$s_p^2 = \frac{(90 - 1)3969 + (30 - 1)4200}{90 + 30 - 2} = \frac{475041}{118} = 4025.77$$

Dunque l'intervallo richiesto è il seguente:

$$\mathbf{IC}_{0.90}(\mu_A - \mu_B) \approx (500 - 600) \pm 1.645 \sqrt{4025.77 \times \left(\frac{1}{90} + \frac{1}{30} \right)} = -100 \pm 22 = [-122, -78]$$

3.1.3 *Approfondimento*: varianze incognite e non necessariamente uguali

L'assunzione di varianze uguali è comoda per i calcoli, ma è il più delle volte irreale. Se volessimo procedere assumendo che le varianze siano diverse, la varianza della differenza tra medie verrebbe stimata come nell'equazione (1), vista nel paragrafo precedente. Non va quindi calcolato uno stimatore congiunto delle varianze, poiché esse devono essere mantenute differenti.

Lo svantaggio è che i gradi di libertà della t non sono determinabili con esattezza, ma sono approssimabili con un *broken number* (numero non intero), come mostrato su questo link (i gradi di libertà sono indicati con la lettera greca ν). Questo numero non intero rende l'uso della tavola della t piuttosto complicato.

3.2 Intervallo di confidenza per il confronto tra proporzioni

Partiamo da due popolazioni Bernoulliane $X_A \sim \mathbf{Ber}(\pi_A)$ e $X_B \sim \mathbf{Ber}(\pi_B)$.
Vogliamo un intervallo di confidenza (di livello $1-\alpha$) per $\pi_A - \pi_B$.

In virtù del **T**eorema del **L**imite **C**entrale, abbiamo che:

$$\hat{\pi}_A - \hat{\pi}_B \xrightarrow{\textbf{TLC}} \mathbf{N}\left(\pi_A - \pi_B,\ \frac{\pi_A(1-\pi_A)}{n_A} + \frac{\pi_B(1-\pi_B)}{n_B}\right)$$

Anche nel caso della differenza tra proporzioni assumiamo che le due popolazioni Bernoulliane abbiano la stessa varianza:

$$\pi_A(1-\pi_A) = \pi_B(1-\pi_B) = \pi(1-\pi)$$

Per stimare π, ci avvaliamo di uno stimatore *pooled* $\hat{\pi}_p$, uguale alla proporzione di successi totale nei due campioni:

$$\hat{\pi}_p = \frac{\sum_{i=1}^{n_A} X_{A,i} + \sum_{j=1}^{n_B} X_{B,j}}{n_A + n_B}$$

L'intervallo di confidenza è dunque il seguente:

$$\mathbf{IC}_{1-\alpha}(\pi_A - \pi_B) \xrightarrow{\textbf{TLC}} (\hat{\pi}_A - \hat{\pi}_B) \pm \underbrace{\mathcal{Z}_{\frac{\alpha}{2}} \sqrt{\hat{\pi}_p(1-\hat{\pi}_p)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}_{\text{margine di errore}}$$

Supponiamo di voler confrontare la proporzione di adesioni a due diverse (e tra loro indipendenti) campagne contro l'inquinamento. Per la campagna A osserviamo 220000 aderenti su un milione di persone scelte casualmente per partecipare alla campagna. Per la campagna B osserviamo 34500 aderenti su 150000 persone campionate casualmente.

Costruire un intervallo di confidenza del 99% per la differenza tra proporzioni.

In questo caso: $n_A = 1000000,\ \hat{\pi}_A = \frac{220000}{1000000} = 0.22,\ n_B = 150000,\ \hat{\pi}_B = \frac{34500}{150000} = 0.23.$
Il percentile da utilizzare è $\mathcal{Z}_{0.995} = 2.58$.

Lo stimatore congiunto di π è uguale a:

$$\hat{\pi}_p = \frac{220000 + 34500}{1000000 + 150000} = 0.2213$$

L'intervallo di confidenza richiesto è dunque il seguente:

$$\mathbf{IC}_{0.99}(\pi_A - \pi_B) \xrightarrow{\textbf{TLC}} -0.01 \pm 2.58\sqrt{0.2213(1-0.2213)\left(\frac{1}{1000000} + \frac{1}{150000}\right)} =$$

$$= -0.01 \pm 0.003 = [-0.013, -0.007]$$

3.3 Intervallo di confidenza per il confronto tra varianze

Siano $X_A \sim \mathbf{N}(\mu_A, \sigma_A^2)$ e $X_B \sim \mathbf{N}(\mu_B, \sigma_B^2)$. Vogliamo confrontare le varianze di queste due popolazioni tramite un intervallo di confidenza di livello $1 - \alpha$.

È possibile dimostrare che:

$$\frac{\left(\frac{s_A^2}{s_B^2}\right)}{\left(\frac{\sigma_A^2}{\sigma_B^2}\right)} \sim \mathbf{F}_{n_A-1,\, n_B-1}$$

Da ciò consegue che:

$$\mathbf{IC}_{1-\alpha}\left(\frac{\sigma_A^2}{\sigma_B^2}\right) = \left[\frac{\left(\frac{s_A^2}{s_B^2}\right)}{F_{\left(\frac{\alpha}{2},\, n_A-1,\, n_B-1\right)}},\ \frac{\left(\frac{s_A^2}{s_B^2}\right)}{F_{\left(1-\frac{\alpha}{2},\, n_A-1,\, n_B-1\right)}}\right]$$

Nota Bene:

- Si sta ragionando sull'area a destra del percentile considerato (in virtù delle tavole): per esempio, $F_{(0.05,\, 8,\, 2)}$ è il percentile della v.c. **F** di Fisher con 8 e 2 gradi di libertà che ha a destra un'area del 5%.
- Sulle tavole della **F** non abbiamo a disposizione i percentili nella coda di sinistra. Tuttavia, siamo in grado di risalire ad essi tramite la seguente relazione:

$$F_{\left(1-\frac{\alpha}{2},\, n_A,\, n_B\right)} = \frac{1}{F_{\left(\frac{\alpha}{2},\, n_B,\, n_A\right)}}$$

- Per non avere problemi nel calcolo dell'intervallo per $\frac{\sigma_A^2}{\sigma_B^2}$, si raccomanda di porre la varianza campionaria più alta al numeratore e quella più bassa al denominatore, in modo da avere $\frac{s_A^2}{s_B^2} \geq 1$.

Facciamo un esercizio. Si vuole confrontare la variabilità dei voti all'esame di Statistica di due diversi corsi di laurea (EA ed EC) tra loro indipendenti. A tal fine, si prende un campione di 20 studenti di EA e un campione di 15 studenti di EC. Dai campioni risulta che $s_{\text{EA}}^2 = 23$ e $s_{\text{EC}}^2 = 25$.

Calcolare un intervallo di confidenza del 90% per il rapporto tra varianze.

Per prima cosa, dobbiamo ricordare di mettere s_{EC}^2 al numeratore, in quanto ha un valore di variabilità stimata più alto. I percentili della **F** cui far riferimento sono:

- $F_{0.05,\, 15-1,\, 20-1} = 2.26$
- $F_{(0.95,\, 15-1,\, 20-1)} = \frac{1}{F_{0.05,\, 20-1,\, 15-1}} = \frac{1}{2.33} = 0.42$

L'intervallo di confidenza richiesto è il seguente:

$$\mathbf{IC}_{0.90}\left(\frac{\sigma_{\text{EC}}^2}{\sigma_{\text{EA}}^2}\right) = \left[\frac{\left(\frac{25}{23}\right)}{2.26},\ \frac{\left(\frac{25}{23}\right)}{0.42}\right] = [0.48,\ 2.61]$$