Teoremi limite

Introduzione

Uno degli aspetti fondamentali della statistica, che affronteremo in Statistica Inferenziale, è quello di estrarre informazioni sulla distribuzione di una v.a. X utilizzando un certo numero di osservazioni empiriche (misure) di X.

In questa lezione vedremo alcuni risultati teorici, rilevanti per l'analisi dei dati sperimentali, validi quando il numero di osservazioni di X tende all'infinito.

Esempio

Supponiamo di avere una popolazione costituita da individui di tipo A e di tipo B. Sia *p* la percentuale degli individui di tipo A e 1-*p* la percentuale degli individui di tipo B, e supponiamo che *p* sia incognito. Vogliamo determinare *p*.

1) Se la popolazione è piccola allora basta contare il numero di individui di tipo A e dividerlo per il numero totale di individui.

2) Se la popolazione è grande, o non interamente accessibile, allora possiamo solo ottenere una *stima* di *p*. A tale scopo estraiamo un campione casuale di dimensione *n* dall'intera popolazione e contiamo il numero N_A di individui di tipo A tra gli *n* estratti. Allora una stima $\overline{p}$ di *p* è:

$$\overline{p} = \frac{N_A}{n}$$

e questa stima migliora al crescere di *n*.

Esempio (cont.)

Qual è la differenza concettuale tra p e $\bar{p}$?

p è un numero. $\bar{p}$ è una *variabile aleatoria* in quanto varia al variare del campione estratto.

Per formalizzare questo concetto dobbiamo pensare alle n osservazioni del nostro campione casuale come le osservazioni di n v.a. *indipendenti* $X_1,\ldots,X_n$ tali che:

$$X_k = \begin{cases} 1 \text{ se l'individuo k - esimo è di tipo A} \\ 0 \qquad\qquad \text{altrimenti} \end{cases} \quad \text{per } k = 1,2,\ldots,n.$$

Poiché $P\{X_k=1\}=p$ allora X_k è una v.a. di Bernoulli $B(1, p)$ e $X_1,\ldots,X_n$ sono n v.a. *indipendenti* ed *identicamente distribuite* (i.i.d.) come una generica v.a. X di Bernoulli $B(1,p)$.

Esempio (cont.)

È evidente che il numero N_A di individui di tipo A nel nostro campione casuale è anch'esso una v.a.

$$N_A = X_1 + \ldots + X_n$$

e sappiamo da un teorema che N_A è Binomiale $B(n, p)$. Inoltre la nostra stima

$$\overline{p} = \frac{N_A}{n} = \frac{X_1 + \ldots + X_n}{n} = \overline{X}$$

è la media aritmetica delle X_k. Poiché X è $B(1,p)$, allora $E(X)=p$ e quindi il nostro problema si può scrivere come:

data una v.a. X con legge $B(1,p)$ stimare la media di X: $E(X)=p$.

La nostra soluzione intuitiva consiste nello stimare p con $\overline{p}$ (media aritmetica del campione). Vedremo dopo il perché.

Stime come variabili aleatorie

L'esempio mostra che se X è una v.a., la sua media $E(X)=\mu$ è un numero. La media aritmetica

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n}$$

calcolata su un campione casuale di v.a. $X_1,\ldots,X_n$ distribuite come X, utilizzata come stima di $E(X)$, è una v.a.

Inoltre, per stimare la varianza $\sigma^2=Var(X)$ potremo usare la v.a. varianza del campione

$$S_c^2 = \frac{1}{n}\sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2$$

Vedremo come cambia l'affidabilità di queste stime al crescere di n.

La legge dei grandi numeri

In modo intuitivo questo risultato asintotico afferma che l’affidabilità delle stime aumenta all’aumentare del numero delle osservazioni.

Quindi supporremo di avere a disposizione un numero grande a piacere di osservazioni di una v.a.

Convergenza di una successione di v.a.

Data una successione di v.a. Z_n con $n \in N$, diremo che per $n \to \infty$ essa converge ad un numero a, e scriveremo

$$Z_n \xrightarrow{n} a$$

quando

$$\lim_{n\to\infty} E(Z_n) = a \qquad \lim_{n\to\infty} Var(Z_n) = 0.$$

Nota che Z_n è una successione di v.a. mentre $E(Z_n)$ e $Var(Z_n)$ sono successioni di numeri.

La Legge dei Grandi Numeri

Se X_k è una successione di v.a. indipendenti, tutte con media μ e varianza σ^2, allora :

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{n} \mu \qquad (1)$$

$$S_c^2 = \frac{1}{n}\sum_{k=1}^{n} \left(X_k - \overline{X}\right)^2 \xrightarrow{n} \sigma^2. \qquad (2)$$

Infatti, per la linearità del valore atteso si ha :

$$E\left(\overline{X}\right) = E\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{E\left(X_1\right) + \ldots + E\left(X_n\right)}{n} = \frac{\mu + \ldots + \mu}{n} = \mu.$$

Inoltre, essendo le X_k indipendenti si ha :

$$Var\left(\overline{X}\right) = Var\left(\frac{X_1 + \ldots + X_n}{n}\right) = \frac{Var\left(X_1\right) + \ldots + Var\left(X_n\right)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Poiché $E\left(\overline{X}\right) \to \mu$ e $Var\left(\overline{X}\right) \to 0$ per $n \to \infty$ allora $\overline{X} \xrightarrow{n} \mu$.

Teorema Limite Centrale (intro.)

La LGN non ci consente di fare delle stime quantitative sulla attendibilità della stima di un parametro. Ad esempio non ci consente di dire qual è la probabilità che il valore stimato del parametro sia ad una certa distanza dal valore vero.

A tale scopo è necessario conoscere la distribuzione, o almeno una approssimazione, della v.a. utilizzata per stimare il parametro.

Il TLC afferma che se una v.a. *X* è la somma di un gran numero di piccole v.a. allora *X* è approssimativamente distribuita secondo una legge normale, indipendentemente dalla distribuzione delle singole v.a.

Teorema Limite Centrale

Sia X_k con $k \in N$ una successione di v.a. indipendenti con la stessa media μ e varianza $\sigma^2 < +\infty$.
Consideriamo la v.a.

$$S_n = \frac{\overline{X}_n - E\left(\overline{X}_n\right)}{\sqrt{Var\left(\overline{X}_n\right)}} = \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}}.$$

Allora la distribuzione di S_n tende alla distribuzione Normale standard per n che tende all'infinito.

Teorema Limite Centrale (cont.)

Osserva che S_n è una v.a. standardizzata: $E(S_n)=0$ e $Var(S_n)=1$. Inoltre per n abbastanza grande S_n è approssimativamente $N(0,1)$. Analogamente, $\bar{X}_n$ è asintoticamente $N(\mu, \sigma^2/n)$.

L'aspetto sbalorditivo del teorema è che nulla si conosce circa la distribuzione originaria delle v.a., se non che esse hanno varianza finita.

L'importanza del TLC per le applicazioni è che la media $\bar{X}_n$ di un campione casuale estratto da una qualsiasi distribuzione con varianza σ^2 finita e media μ è approssimativamente distribuita come $N(\mu, \sigma^2/n)$.

Osservazione

Quindi se X_k sono generiche v.a. con media μ e varianza σ^2 allora per *n* grande avremo:

$$P\left\{\overline{X} \le x\right\} = P\left\{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le \frac{x-\mu}{\sigma/\sqrt{n}}\right\} \approx \Phi\left(\frac{x-\mu}{\sigma/\sqrt{n}}\right)$$

$$\text{dove } \Phi \text{ è la cdf della } N(0,1).$$

Inoltre pur essendo N(0,1) la legge di una v.a. continua, il teorema resta valido anche se le X_k sono v.a. discrete.

Esempio

Ritorniamo all'esempio precedente e supponiamo che le proporzioni di individui di tipo A e B siano uguali, ossia p=1/2. Estraiamo un campione di n=100 individui e calcoliamo la probabilità di osservare un numero di individui di tipo A maggiore di 60.

Utilizzando la stessa notazione si ha che:

$$N_A = X_1 + \ldots + X_{100}$$

$$\text{è una v.a. Binomiale } B\left(100, \frac{1}{2}\right) \text{ per cui}$$

$$P\{N_A > 60\} = \sum_{k=61}^{100} \binom{100}{k} \left(\frac{1}{2}\right)^k \left(1-\frac{1}{2}\right)^{100-k} = \sum_{k=61}^{100} \binom{100}{k} \frac{1}{2^{100}}.$$

Questa quantità può essere calcolata con un computer.

Esempio (cont.)

Utilizziamo il TLC per ottenere un valore approssimato di $P\{N_A>60\}$. Osserva che N_A è la somma di n=100 v.a. indipendenti X_k ciascuna Bernoulli $B(1, ½)$. Inoltre:

$$\mu = E(X_k) = p = \frac{1}{2} \text{ e } \sigma^2 = Var(X_k) = p(1-p) = \frac{1}{4}.$$

Inoltre per l'indipendenza delle X_k :

$$E(N_A) = E(X_1 + \ldots + X_{100}) = E(X_1) + \ldots + E(X_{100}) = 100\mu = 50.$$

$$Var(N_A) = Var(X_1 + \ldots + X_{100}) = Var(X_1) + \ldots + Var(X_{100}) = 100\sigma^2 = 25.$$

Allora per il TLC $N_A \approx N(100\mu, 100\sigma^2)$ e quindi :

$$P\{N_A > 60\} = 1 - P\{N_A \le 60\} = 1 - P\left\{\frac{N_A - E(N_A)}{\sqrt{Var(N_A)}} \le \frac{60 - E(N_A)}{\sqrt{Var(N_A)}}\right\}$$

$$= 1 - P\left\{\frac{N_A - 50}{5} \le \frac{60-50}{5}\right\} \approx 1 - \Phi(2) = 1 - 0.97725 = 0.02275.$$

Esempio (cont.)

Il valore esatto di $P\{N_A>60\}$ è 0.0176 che è un po' diverso da quello trovato. La ragione sta nel fatto che abbiamo applicato il TLC a v.a. discrete. Come regola generale, se X è una v.a. a valori discreti, una migliore approssimazione normale del valore $P\{X\leq k\}$ la si ottiene calcolando $P\{X\leq k+1/2\}$ con k intero.

Quindi:

$$P\{N_A>60\}=1-P\{N_A\leq 60\}=1-P\{N_A\leq 60.5\}$$
$$=1-P\left\{\frac{N_A-50}{5}\leq\frac{60.5-50}{5}\right\}\approx 1-\Phi(2.1)=1-0.98214=0.01786.$$