Statistica inferenziale

Stima di parametri

Introduzione

In questa lezione affronteremo il problema di determinare una stima di qualche parametro θ incognito della distribuzione di una v.a. X a partire da n osservazioni $X_1,\ldots,X_n$ della variabile stessa. Per esempio, se X:$N(\mu,\sigma^2)$ allora il parametro θ incognito è la coppia di parametri $\theta=(\mu,\sigma)$.

Questo è un problema classico di statistica inferenziale in quanto si vogliono studiare le proprietà di una v.a. partendo da un finito e limitato di osservazioni empiriche.

Esempio: stima di una media

Si vuole determinare il livello di espressione medio del gene P53 nel colon in soggetti sani. In questo caso il livello del gene è rappresentato da una v.a. X e vogliamo stimare il parametro $\mu=E(X)$. A tale scopo consideriamo n soggetti sani, misuriamo il livello di espressone del gene nel colon ed utilizziamo la media di queste misure come stima di μ.

Questo è equivalente a prendere n v.a. $X_1,\ldots,X_n$ i.i.d. come X, ossia con la stessa media di X, e a calcolare la v.a. $\bar{X}$. Prenderemo come stima di μ il valore di $\bar{X}$ ricavato dal campione in esame fidandoci del fatto che per n abbastanza grande, grazie alla LGN, tale stima è vicina al valore vero μ.

Esempio: stima di una distribuzione discreta

Sia X una v.a. discreta che assume valori k=1,2… e si vuole determinare la distribuzione teorica $p_k=P\{X=k\}$.

Empiricamente, consideriamo n osservazione di X e supponiamo che N_k volte esce il valore k. Allora una stima di p_k è N_k/n. Questo equivale a considerare n v.a. $X_1,\ldots,X_n$ i.i.d. come X, nel senso che per ogni j=1,…,n: $p_k=P\{X_j=k\}$. Per un fissato valore k che può assumere una v.a. consideriamo:

$$Y_j(k)=\begin{cases}1 \text{ se } X_j=k\\ 0 \text{ altrimenti}\end{cases} \text{ per } j=1,2,\ldots,n.$$

Nota che $Y_1(k),\ldots,Y_n(k)$ sono v.a. i.i.d. $B(1,p_k)$, in quanto $P\{Y_j(k)=1\}=P\{X_j=k\}=p_k$. e quindi $E[Y_j(k)]=p_k$. Poniamo:

Esempio: stima di una distribuzione discreta

$$N_k = \sum_{j=1}^{n} Y_j(k) = \text{Numero delle } X_j \text{ che valgono } k.$$

Inoltre:

$$\frac{N_k}{n} = \frac{1}{n}\sum_{j=1}^{n} Y_j(k) = \overline{Y}(k).$$

Osserva che $Y_1(k),\ldots,Y_n(k),\ldots$ è una successione di v.a. con stessa media p_k allora per la LGN risulta che $\overline{Y}(k) \xrightarrow{n} p_k$. Quindi per grandi valori di n possiamo stimare i valori teorici p_k con i valori empirici N_k / n.

Definizioni

Diremo che n v.a. $X_1,...,X_n$ costituiscono un campione casuale della v.a. X se esse sono indipendenti ed identicamente distribuite con la stessa legge di X.

Si chiama statistica qualunque v.a. T funzione del campione, cioè $T = t(X_1,...,X_n)$.

Dato un campione $X_1,...,X_n$ di v.a. con legge dipendente dal parametro θ, chiameremo stimatore del parametro θ una statistica T che sia funzione del campione dato. Diremo che T è uno stimatore non distorto di θ se $E(T) = \theta$, e che esso è uno stimatore consistente se T converge a θ per $n \to \infty$.

Esempi di stimatori

Quindi uno stimatore è una qualunque funzione del campione.

Un esempio di stimatore del parametro $\theta=\mu$ è la media aritmetica $\bar{X}$. La LGN ci garantisce che $\bar{X}$ è uno stimatore consistente della media.

La varianza campionaria S^2_C è un esempio di stimatore della varianza. La LGN afferma che esso è uno stimatore consistente della varianza; esso è comunque distorto e per tale motivo richiede una piccola correzione.

Notazione

Nel seguito indicheremo con $X_1,\ldots,X_n$ un campione di n v.a. indipendenti ed identicamente distribuite, tutte con media μ e varianza σ^2.

Ricordiamo che la media del campione è : $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$

mentre la varianza del campione è : $S_C^2 = \frac{1}{n}\sum_{k=1}^{n}\left(X_k - \overline{X}\right)^2 = \overline{X^2} - \overline{X}^2.$

Introduciamo ora la varianza corretta :

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

E' facile verificare che :

$$S^2 = \frac{n}{n-1}S_C^2 = \frac{n}{n-1}\left(\overline{X^2} - \overline{X}^2\right).$$

Teorema

$\overline{X}, S_C^2$ e S^2 sono stimatori consistenti di μ e σ^2 rispettivamente. $\overline{X}$ e S^2 sono stimatori non distorti di μ e σ^2. S_C^2 è uno stimatore distorto di σ^2.

Infatti, la consistenza di $\overline{X}$ e S_C^2 segue dalla LGN. Inoltre, poiché $S^2 = \frac{n}{n-1} S_C^2$ allora anche S^2 è consistente. Calcoliamo:

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

e quindi $\overline{X}$ è uno stimatore non distorto di μ. Calcoliamo:

$$E\left(S_C^2\right) = E\left(\overline{X^2} - \overline{X}^2\right) = E\left(\overline{X^2}\right) - E\left(\overline{X}^2\right) =$$

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

Ricorda la proprietà : $Var(X) = E(X^2) - E(X)^2$. Allora :

$Var(X_k) = E(X_k^2) - E(X_k)^2 \Rightarrow E(X_k^2) = Var(X_k) + E(X_k)^2$ ossia

$E(X_k^2) = \sigma^2 + \mu^2$.

Per la stessa proprietà possiamo scrivere :

$Var(\overline{X}) = E(\overline{X}^2) - E(\overline{X})^2 \Rightarrow E(\overline{X}^2) = Var(\overline{X}) + E(\overline{X})^2 \Rightarrow$

$E(\overline{X}^2) = Var\left(\frac{X_1 + ... + X_n}{n}\right) + \mu^2$. Poichè le X_k sono indipendenti

$$E(\overline{X}^2) = \frac{1}{n^2}\left[Var(X_1) + ... + Var(X_n)\right] + \mu^2 = \frac{n\sigma^2}{n^2} + \mu^2 = \frac{\sigma^2}{n} + \mu^2.$$

Allora, dalla pagina precedente :

$$E(S_C^2) = \frac{1}{n}\sum_{k=1}^{n} E(X_k^2) - E(\overline{X}^2) = \left(\frac{1}{n}\sum_{k=1}^{n} \sigma^2 + \mu^2\right) - \left(\frac{\sigma^2}{n} + \mu^2\right)$$

$$= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2.$$

Allora S_C^2 è uno stimatore distorto di σ^2, mentre S^2 è non distorto in quanto :

$$E(S^2) = E\left(\frac{n}{n-1} S_C^2\right) = \frac{n}{n-1} E(S_C^2) = \frac{n}{n-1}\frac{n-1}{n}\sigma^2 = \sigma^2.$$

Osservazione

Nota che S^2_C è la varianza del campione, mentre S^2 è uno stimatore della varianza di una legge di distribuzione. Esempio

6.03	5.95	7.26	5.27	5.44	3.84	3.94	3.62	3.30	5.36
4.18	3.80	5.42	4.39	4.92	4.93	3.89	5.14	5.70	4.89

Consideriamo un campione costituito da $n = 20$ misure di una v.a. X. Allora :

$$\overline{X} = \frac{1}{20}\sum_{k=1}^{20} X_k = 4.86$$

$$S_C^2 = \overline{X^2} - \overline{X}^2 = 0.94 \qquad S^2 = \frac{n}{n-1}\left(\overline{X^2} - \overline{X}^2\right) = 0.99$$

Le due stime sono differenti e la differenza diminuisce al crescere di n. Quando dobbiamo stimare la varianza si usa S^2 perché è uno stimatore non distorto.

Stima per intervalli

E’ evidente che la probabilità che il valore stimato di un parametro coincida con il suo valore vero è zero. Quindi invece di stimare il valore di un parametro è possibile, partendo dalle osservazioni del campione, determinare gli estremi di un intervallo che contiene, con una certa probabilità, il valore vero del parametro.

Nella stima per intervalli cerchiamo di determinare utilizzando il campione gli estremi di un intervallo che contiene, con una certa probabilità prefissata, il valore vero del parametro.

Intervalli di confidenza

Diremo che le due v.a. $T_1 = t(X_1,\ldots,X_n)$ e $T_2 = t(X_1,\ldots,X_n)$ sono gli estremi di un intervallo di confidenza $[T_1, T_2]$ di livello α (con $0 < \alpha < 1$) per θ quando

$$P_\theta\{T_1 \le \theta \le T_2\} = 1 - \alpha$$

dove il simbolo P_θ indica che la probabilità è calcolata supponendo che il valore del parametro incognito sia θ.

Osservazione

Per un determinato α l'intervallo di confidenza non è unico. Allora si preferisce prendere T_1 e T_2 in maniera simmetrica rispetto al parametro, cioè tali che:

$$P_\theta\{\theta < T_1\} = P_\theta\{T_2 < \theta\} = \frac{\alpha}{2}$$

In generale gli estremi di un intervallo di confidenza assumono la forma $T \pm \Delta$ dove T è uno stimatore di θ e 2Δ è l'ampiezza dell'intervallo.

L'ampiezza dell'intervallo dipende dal livello α. In generale si sceglie α piccolo (α=0.05 o α=0.01) in modo che la probabilità che l'intervallo contenga il valore vero del parametro sia grande (0.95 o 0.99 rispettivamente).

Intervallo di confidenza per la media μ

Sia $X_1,\ldots,X_n$ un campione casuale della v.a. $X : N(\mu,\sigma^2)$ e supponiamo che σ^2 sia nota. Vogliamo stimare μ.
Utilizziamo $\overline{X}$ che sappiamo essere uno stimatore consistente e non distorto di μ. Consideriamo la v.a.

$Y = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ che sappiamo essere distribuita secondo la legge $N(0,1)$.

Abbiamo dimostrato che $P\left\{|Y| \geq \varphi_{1-\frac{\alpha}{2}}\right\} = \alpha$ da cui si ha :

$$1-\alpha = P\left\{|Y| \leq \varphi_{1-\frac{\alpha}{2}}\right\} = P\left\{\left|\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right| \leq \varphi_{1-\frac{\alpha}{2}}\right\} = P\left\{\left|\overline{X}-\mu\right| \leq \frac{\sigma}{\sqrt{n}}\varphi_{1-\frac{\alpha}{2}}\right\}$$

$$= P\left\{-\frac{\sigma}{\sqrt{n}}\varphi_{1-\frac{\alpha}{2}} \leq \mu-\overline{X} \leq \frac{\sigma}{\sqrt{n}}\varphi_{1-\frac{\alpha}{2}}\right\}$$

$$= P\left\{\overline{X}-\frac{\sigma}{\sqrt{n}}\varphi_{1-\frac{\alpha}{2}} \leq \mu \leq \overline{X}+\frac{\sigma}{\sqrt{n}}\varphi_{1-\frac{\alpha}{2}}\right\}$$

Intervallo di confidenza per la media μ

Quindi nel caso di varianza nota, l'intervallo di confidenza di livello α della media prende la forma :

$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}} \varphi_{1-\frac{\alpha}{2}}, \overline{X} + \frac{\sigma}{\sqrt{n}} \varphi_{1-\frac{\alpha}{2}} \right]$$

che può anche essere scritto come :

$$\overline{X} \pm \frac{\sigma}{\sqrt{n}} \varphi_{1-\frac{\alpha}{2}}.$$

Nota che l'ampiezza dell'intervallo è direttamente proporzionale a σ ed inversamente proporzionale ad *n*. Inoltre gli intervalli di confidenza per la media hanno sempre il centro in $\overline{X}$.

Esempio

Vogliamo studiare il livello medio *X* di un dato enzima in una certa popolazione umana. Consideriamo allora un campione costituito da n=10 individui, misuriamo in ciascuno il livello di tale enzima ed otteniamo il valore medio di $\bar{x}$=22. Supponiamo che *X* sia normalmente distribuito con varianza σ^2=45. Determiniamo un intervallo di confidenza di μ di livello α=0.05. Allora:

$$\left[22-\sqrt{\frac{45}{10}}\varphi_{1-\frac{0.05}{2}},22+\sqrt{\frac{45}{10}}\varphi_{1-\frac{0.05}{2}}\right]=$$

$$\left[22-2.1213\varphi_{0.975},22+2.1213\varphi_{0.975}\right]=\left[22-2.1213\cdot 1.96,22+2.1213\cdot 1.96\right]=$$

$$\left[22-2.1213\cdot 1.96,22+2.1213\cdot 1.96\right]=\left[17.84,26.16\right].$$

Questo è l'intervallo di confidenza al 95% della media, ossia se prendo 100 campioni di dimensione 10, circa 95 volte la media del campione cadrà nell'intervallo [17.84,26.16].

Intervallo di confidenza per la media μ

Sia $X_1,...,X_n$ un campione casuale della v.a. $X : N(\mu,\sigma^2)$ e supponiamo che σ^2 non sia nota. Vogliamo stimare μ.
In questo caso prima dobbiamo stimare la varianza. A tale scopo utilizziamo S^2 che sappiamo essere uno stimatore consistente e non distorto di σ^2. Consideriamo la v.a.
$T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}}$ che sappiamo essere distribuita come una t di Student $t(n-1)$ con $n-1$ gradi di libertà. In precedenza abbiamo dimostrato che $P\left\{|T| \geq t_{1-\frac{\alpha}{2}}(n)\right\} = \alpha$ e quindi ripetendo gli stessi passaggi possiamo affermare che l'intervallo di confidenza a livello α della media è

$$\left[\overline{X} - \frac{S}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1), \overline{X} + \frac{S}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)\right].$$

Esempio

Consideriamo il livello di arsenico *X* secreto nell'urina da 16 soggetti sani (milligrammi per giorno). Nell'ipotesi che X sia gaussiana, costruiamo l'intervallo di confidenza al 99% della media della popolazione.

.007	.03	.025	.008	.03	.038	.007	.005	.012	.006	.01	.032	.006	.009	.014	.011

$\bar{x}$=0.0156; s=0.0112;

$$\left[0.0156-\frac{0.0112}{4}t_{1-\frac{0.01}{2}}(15),0.0156+\frac{0.0112}{4}t_{1-\frac{0.01}{2}}(15)\right]=$$

$$\left[0.0156-\frac{0.0112}{4}t_{0.995}(15),0.0156+\frac{0.0112}{4}t_{0.995}(15)\right]=$$

$$\left[0.0156-\frac{0.0112}{4}\cdot 2.95,0.0156+\frac{0.0112}{4}\cdot 2.95\right]=$$

$$[0.00734,0.02386].$$

Osservazione

Le formule per gli intervalli di confidenza sono state ricavate nell'ipotesi di campioni gaussiani. In virtù del TLC esse rimangono approssimativamente valide anche quando non si conosce la legge di distribuzione dei dati, purché n sia abbastanza grande ($n \geq 20$).

Intervallo di confidenza per la varianza σ^2

Sia $X_1,...,X_n$ un campione casuale della v.a. $X : N(\mu,\sigma^2)$ e supponiamo di voler stimare σ^2.
Utilizziamo S^2 che sappiamo essere uno stimatore consistente e non distorto di σ^2.

Consideriamo la v.a. $Z = (n-1)\dfrac{S^2}{\sigma^2}$ che sappiamo essere $\chi^2(n-1)$. Sappiamo che:

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

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

Quindi l'intervallo di confidenza di livello α è:

$$\left[\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right].$$

Esempio

In uno studio sull'effetto della dieta sul colesterolo LDL, è stato misurato il livello del colesterolo nel plasma di 12 soggetti:

6,0	6,4	7,0	5,8	6,0	5,8	5,9	6,7	6,1	6,5	6,3	5,8

$s = 0.391868.$

Determiniamo l'intervallo al 95% della varianza :

$$\left[\frac{11 \cdot 0.391868}{\chi^2_{0.975}(11)}, \frac{11 \cdot 0.391868}{\chi^2_{0.025}(11)}\right] = \left[\frac{11 \cdot 0.391868}{21.92}, \frac{11 \cdot 0.391868}{3.8158}\right] =$$

$$[0.1966, 1.1297].$$