Stima di una proporzione

Lucio Demeio
Dipartimento di Ingegneria Industriale e Scienze Matematiche
Università Politecnica delle Marche

Il problema della stima di una proporzione sorge quando si vuole conoscere la percentuale di votanti per un determinato partito, la percentuale di pezzi difettosi tra quelli prodotti da una determinata ditta, etc. Come sempre, consideriamo un campione di rango n e siano $X_1, X_2, ..., X_n$ le variabili che lo compongono. Abbiamo visto precedentemente che, per la legge dei grandi numeri, la proporzione che vogliamo stimare concide, per n grande, con il parametro p della variabile di Bernoulli che sottintende alla proporzione. Stimare una proporzione è pertanto equivalente a stimare il parametro p di una variabile di Bernoulli. Abbiamo anche visto che lo stimatore di massima verosimiglianza per esso è la media campionaria, quindi

$$\hat{p} = \overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}. \tag{1}$$

La variabile S_n introdotta con il teorema del limite centrale è data dalla relazione

$$S_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \tag{2}$$

dove μ e σ^2 sono la media e la varianza della legge di Bernoulli. Sappiamo che

$$\mu = p \tag{3}$$
$$\sigma^2 = p\,(1-p) \tag{4}$$

e quindi

$$S_n = \frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}} \tag{5}$$

Adottando ora l'approssimazione normale per S_n, otteniamo per l'intervallo di confidenza di livello $1-\alpha$:

$$P\left(|S_n| \leq z_{\alpha/2}\right) = 1 - \alpha \tag{6}$$

ovvero

$$P\left(\left|\frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}}\right| \leq z_{\alpha/2}\right) = 1 - \alpha \tag{7}$$

$$P\left(\overline{X}_n - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \overline{X}_n + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha. \tag{8}$$

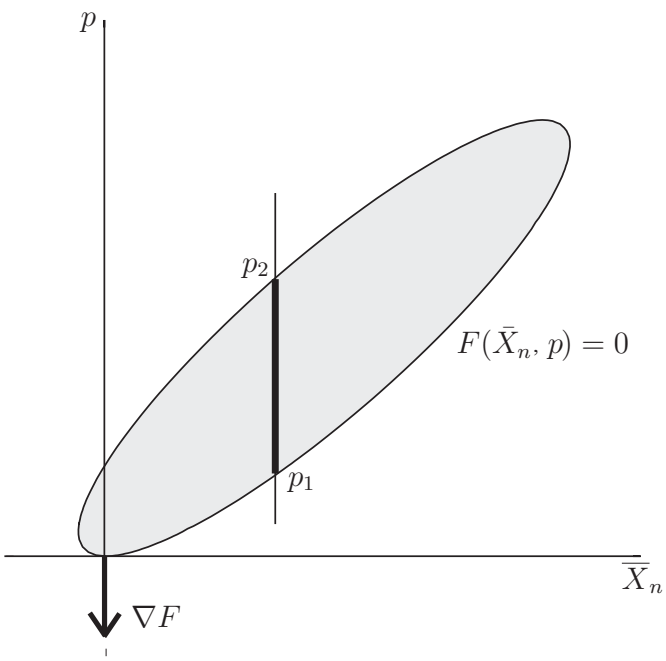

Figura 1: Ellisse di confidenza.

Da quest'ultima equazione si vede che gli estremi dell'intervallo di confidenza per p dipendono dallo stesso p, cioè dallo stesso parametro che si vuole stimare. Per ottenere una relazione in forma chiusa, riscriviamo l'equazione (7) nella forma

$$P\left[(\overline{X}_n - p)^2 - \frac{p(1-p)}{n}\, z^2_{\alpha/2} \leq 0\right] = 1 - \alpha \tag{9}$$

$$P\left[\left(1 + \frac{z^2_{\alpha/2}}{n}\right)\, p^2 - 2\,\left(\overline{X}_n + \frac{z^2_{\alpha/2}}{2\,n}\right)\, p + \overline{X}^2_n \leq 0\right] = 1 - \alpha \tag{10}$$

La disuguaglianza che compare come argomento della funzione di probabilità rappresenta l'interno di un'ellisse nel piano cartesiano $O(\overline{X}_n, p)$. L'ellisse ha equazione

$$F(\overline{X}_n, p) \equiv \left(1 + \frac{z^2_{\alpha/2}}{n}\right)\, p^2 - 2\,\left(\overline{X}_n + \frac{z^2_{\alpha/2}}{2\,n}\right)\, p + \overline{X}^2_n = 0 \tag{11}$$

e passa per l'origine, come si vede dal fatto che il punto $(\overline{X}_n = 0, p = 0)$ soddisfa l'equazione (11). Nell'origine, inoltre, l'ellisse è tangente all'asse $\overline{X}_n$; calcolando il gradiente ∇F otteniamo:

$$\frac{\partial F}{\partial \overline{X}_n} = -2\,p + 2\,\overline{X}_n \tag{12}$$

$$\frac{\partial F}{\partial p} = 2\,p\,\left(1 + \frac{z^2_{\alpha/2}}{n}\right) - 2\,\left(\overline{X}_n + \frac{z^2_{\alpha/2}}{2\,n}\right) \tag{13}$$

e si vede che $\nabla F(0,0) = (0, -z^2_{\alpha/2}/n)$, ortogonale alla direzione dell'asse $\overline{X}_n$ e rivolto verso l'esterno dell'ellisse.

Una volta ottenuto un valore per la media campionaria $\overline{X}_n$, l'intervallo di confidenza di livello $1-\alpha$ per p è dato dal segmento che la retta verticale di ascissa $\overline{X}_n$ stacca all'interno dell'ellisse (11) e che ha per estremi i valori p_1 e p_2, soluzioni dell'equazione (11) o, equivalentemente, dell'equazione

$$\left(n + z^2_{\alpha/2}\right)\, p^2 - 2\left(n\,\overline{X}_n + \frac{z^2_{\alpha/2}}{2}\right) p + n\,\overline{X}^2_n = 0.$$

Abbiamo pertanto:

$$p_{1,2} = \frac{n\,\overline{X}_n + z^2_{\alpha/2}/2 \pm \sqrt{(n\,\overline{X}_n + z^2_{\alpha/2}/2)^2 - (n + z^2_{\alpha/2})\,n\,\overline{X}^2_n}}{n + z^2_{\alpha/2}} \tag{14}$$

$$= \frac{n\,\overline{X}_n + z^2_{\alpha/2}/2 \pm \sqrt{z^4_{\alpha/2}/4 + n\,z^2_{\alpha/2}\,\overline{X}_n - n\,z^2_{\alpha/2}\,\overline{X}^2_n}}{n + z^2_{\alpha/2}} = \tag{15}$$

$$= \frac{n\,\overline{X}_n + z^2_{\alpha/2}/2 \pm z^2_{\alpha/2}\sqrt{z^2_{\alpha/2}/4 + n\,\overline{X}_n - n\,\overline{X}^2_n}}{n + z^2_{\alpha/2}} \tag{16}$$

Possiamo quindi affermare che il parametro p della legge di Bernoulli cade con probabilità $1-\alpha$ nell'intervallo $[p_1, p_2]$:

$$P(p_1 \le p \le p_2) = 1 - \alpha. \tag{17}$$

Notiamo che l'ampiezza dell'intervallo $[p_1, p_2]$ diminuisce all'aumentare di n: quanto più grande è la dimensione del campione tanto più precisa è la nostra stima. L'ampiezza dell'intervallo diminuisce anche al diminuire di $z_{\alpha/2}$, cioè al diminuire di α: quanto più piccola è la probabilità richiesta, tanto più piccolo è l'intervallo di confidenza. Questo comportamento è rappresentato nelle Figure 2 e **??**,

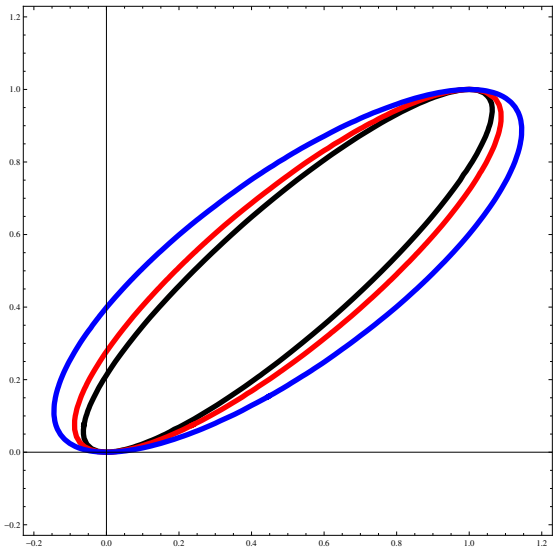

Figura 2: Ellisse di confidenza per $n = 10$ al variare del livello di fiducia: 90% (linea nera), 95% (linea rossa) e 99% (linea blu).

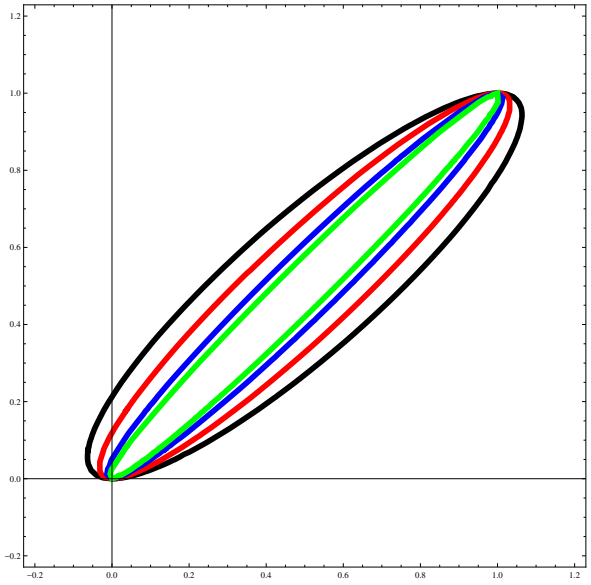

Figura 3: Ellisse di confidenza al livello di confidenza del 90% al variare del rango del campione: $n = 10$ (linea nera), $n = 20$ (linea rossa), $n = 50$ (linea blu) e $n = 100$ (linea verde).

Facciamo un esempio. Supponiamo di voler stimare la probabilità p di ottenere testa con una moneta. A tale scopo, usiamo il generatore di numeri casuali del calcolatore, che è in grado di generare sequenze di "0" e "1" con probabilità assegnata p. Vogliamo stimare p con il livello di fiducia del 95 %. In tal caso abbiamo $\alpha = 0.05$ (cioè 5 %) e $\alpha/2 = 0.025$ (cioè 2.5 %). Dalle tavole dei quantili della distribuzione normale abbiamo $z_{\alpha/2} = 1.960$. Nella tabella seguente riportiamo $\overline{X}_n$ e l'intervallo $[p_1, p_2]$ per alcuni valori di n.

n	$\overline{X}_n$	p_1	p_2
10	0.5	0.36	0.63
20	0.45	0.36	0.56
50	0.4	0.34	0.47
100	0.39	0.35	0.44
200	0.4	0.37	0.44
1000	0.4	0.39	0.52
10000	0.39	0.39	0.4

Ovviamente, generando un nuovo campione le medie campionarie cambiano e così pure gli intervalli. I numeri casuali generati in questa simulazione sono stati ottenuti da una distribuzione di Bernoulli con parametro $p = 0.4$.