Inferenza Statistica

Esame del 5 novembre 2012

Tempo a disposizione 2 ore.

1. Sia dato un campione casuale semplice di 5 unità da una popolazione normale di media μ e varianza σ^2.

a. Quale è la distribuzione della statistica campionaria $W = \frac{5(\overline{Y}-\mu)^2}{S^2}$, dove $\overline{Y}$ è la media campionaria e S^2 è la varianza campionaria (corretta)?

b. Per quale valore di c si ha $P(-c \leq S/(\overline{Y}-\mu) \leq c) = 0.95$?

Soluzione

a. Si consideri il campione $Y_1, \ldots, Y_5$ con $Y_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, 5$. È noto che $\overline{Y} \sim N(\mu, \sigma^2/5)$ e quindi

$$Z = \frac{\overline{Y}-\mu}{\sigma/\sqrt{5}} = \frac{\sqrt{5}(\overline{Y}-\mu)}{\sigma} \sim N(0,1)$$

Segue che $Z^2 \sim \chi^2_1$. Inoltre, poiché il campione proviene da una popolazione normale, risulta $X = \frac{4S^2}{\sigma^2} \sim \chi^2_4$. Quindi si ottiene

$$W = \frac{5(\overline{Y}-\mu)^2}{S^2} = \frac{Z^2}{X/4} \sim F(1,4)$$

essendo il rapporto tra due chi-quadrato indipendenti divise per i loro gradi di libertà.

b. Occorre determinare c tale che

$$0.95 = P\left(-c \leq \frac{S}{\overline{Y}-\mu} \leq c\right) = P\left(\frac{S^2}{(\overline{Y}-\mu)^2} \leq c^2\right) = P\left(\frac{S^2}{5(\overline{Y}-\mu)^2} \leq \frac{c^2}{5}\right) = P\left(\frac{1}{W} \leq \frac{c^2}{5}\right)$$

Da $W \sim F(1,4)$ si ricava $1/W \sim F(4,1)$ e quindi occorre determinare c tale che

$$\frac{c^2}{5} = f_{0.95,4,1}$$

dove $f_{0.95,4,1}$ è il quantile 0.95 della F di Fisher con 4 ed 1 gradi di libertà che soddisfa $P(F \geq f_{0.95,4,1} = 0.05$; dalle tavole si ricava $f_{0.95,4,1} = 224.5$. Pertanto $c = \sqrt{224.5 \cdot 5} \approx 33.5$.

2. Si supponga per Y una distribuzione bernoulliana $Be(p)$. Di solito è di grande interesse la quantità $logit(p) = \log \frac{p}{1-p}$.

a. Sapreste fornire uno stimatore di massima verosimiglianza per $logit(p)$ a partire da un campione di n elementi tratti da Y?

b. Per un campione numeroso da Y fornire l'espressione approssimata per la varianza dello stimatore ottenuto al punto sopra.

Soluzione

a. La funzione di log-verosimiglianza del campione è

$$\log f(y_1, \ldots, y_n; p) = l(p) = \log\left[\prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}\right]$$
$$= \sum_i y_i \log(p) + \left(n - \sum_i y_i\right) \log(1-p)$$

Lo stimatore di massima verosimiglianza per p si ottiene eguagliando a zero la funzione *score*

$$\frac{1}{p}\sum_i y_i - \left(n - \sum_i y_i\right)\frac{1}{1-p} = 0.$$

Risolvendo rispetto a p si trova $\hat{p} = \frac{\sum_i y_i}{n} = \bar{y}$. Per la proprietà di equivarianza degli stimatori di massima verosimiglianza

$$\widehat{logit(p)} = \log\frac{\hat{p}}{1-\hat{p}}$$

b. Posto $\psi = g(p) = \log\frac{p}{1-p}$, dal punto precedente si è ottenuto $\hat{\psi}_{MV} = \log\frac{\overline{Y}}{1-\overline{Y}}$. La distribuzione campionaria asintotica per ψ si ricava considerando la distribuzione asintotica dello stimatore di MV, $\hat{p}$, e applicando il metodo delta per ricavare la varianza asintotica

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

si ha

$$\widehat{logit(p)} = \hat{\psi}_{MV} \dot{\sim} N\left(\psi, [g'(p)]^2\frac{p(1-p)}{n}\right)$$

dove

$$g'(p) = \frac{d}{dp}\log\frac{p}{1-p} = \frac{1}{p(1-p)}$$

Pertanto l'espressione approssimata per la varianza dello stimatore di massima verosimiglianza del parametro ψ è

$$\frac{1}{p^2(1-p)^2}\cdot\frac{p(1-p)}{n} = \frac{1}{np(1-p)}$$

e la sua distribuzione asintotica è

$$\widehat{logit(p)} = \hat{\psi}_{MV} \dot{\sim} N\left(\log\frac{p}{1-p}, \frac{1}{np(1-p)}\right).$$

3. Un'apparecchiatura elettronica ha una componente il cui tempo di rottura (in giorni) è assimilabile ad una variabile casuale esponenziale X di media λ. Quando si verifica una rottura la componente viene sostituita. In magazzino sono conservati 19 componenti di riserva.

a. Supponendo $\lambda = 20$, si determini la probabilità che la componente duri oltre 25 giorni essendo noto che per i primi 10 giorni la componente non si è rotta.

b. Determinare la probabilità che l'apparecchiatura venga mantenuta in funzione per più di 1 anno dopo che la prima componente viene sostituita.

c. Determinare la probabilità che in 10 giorni si rompa più di una componente.

Soluzione

a. Il tempo di rottura X è tale che $X \sim Esp(1/20)$. La probabilità cercata è $P(X > 25|X > 10)$. Per una variabile aleatoria esponenziale vale la proprietà di assenza di memoria, ne segue che

$$P(X > 25|X > 10) = P(X > 15) = e^{-15/20} = 0.47$$

b. Sia $Y = \sum_{i=1}^{19} X_i$, dove X_i è la durata dell'i-esima componente distribuita secondo un'esponenziale di parametro $\theta = 1/20$. 19 componenti sono sufficienti per un anno se la somma delle loro durate supera 365 giorni. La probabilità che l'apparecchiatura venga mantenuta in funzione per oltre un anno è $P(Y > 365)$. La somma di $n = 19$ esponenziali di parametro $\theta = 1/20$ è distribuita secondo una gamma di parametri n e θ.

Per il Teorema del limite centrale possiamo approssimare la distribuzione di Y con una normale di media $n/\theta = 19 \times 20$ e varianza $n/\theta^2 = 19 \times 20^2$. Otteniamo così

$$P(Y > 365) \approx 1 - \Phi\left(\frac{365 - 19 \cdot 20}{20\sqrt{19}}\right) = 0.5683$$

c. Sia V il numero aleatorio di guasti in 10 giorni che richiedono la sostituzione della componente con un pezzo di ricambio e T il tempo fino alla seconda rottura. Allora $Y \sim Po(1/2)$ e $T \sim Erlang(2, 1/20)$. La probabilità cercata è la probabilità che i guasti in 10 giorno siano almeno due:

$$P(T \leq 10) = P(Y \geq 2) = 1 - [P(Y = 0) + P(Y = 1)] = 1 - \sum_{j=0}^{1} \frac{(1/2)^j}{j!} e^{-1/2} = 0.09$$

4. La lunghezza in cm del fusto di piante adulte Y è distribuita normalmente con varianza 64. Si ottiene un campione di 16 alberi e si decide di rifiutare l'ipotesi nulla $H_0 : \mu = 55$, contro l'ipotesi alternativa che la lunghezza media sia superiore a 55 cm, utilizzando quale regione di accettazione $A = \{(y_1, \dots, y_n) : \bar{y} \leq 57.5\}$.

a. Quanto vale il livello di significatività del test, α?

b. Supponendo di non conoscere $\bar{y}$ e sapendo che il livello di significatività osservato è risultato pari a 0.37, qual era la media del campione?

c. Per quale valore di μ in H_1 la potenza del test risulta pari a 0.8?

Soluzione

a. La regione di rifiuto è $\bar{y} > 57.5$, dove $\bar{y}$ è la media campionaria ottenuta da un campione di numerosità 16. Quindi si ricava

$$\begin{aligned} \alpha &= P(\overline{Y} > 57.5 | \mu = \mu_0) \\ &= P\left(Z > \frac{57.5 - \mu_0}{\sigma/\sqrt{n}}\right) \\ &= 1 - \Phi\left(\frac{57.5 - 55}{8/\sqrt{16}}\right) = 1 - \Phi(1.25) = 0.1 \end{aligned}$$

b. Il livello di significatività osservato (p-value) è $0.37 = P(\overline{Y} \geq \bar{y} | H_0)$, quindi

$$0.37 = P\left(Z \geq \frac{\bar{y} - 55}{8/\sqrt{16}}\right) = 1 - \Phi\left(\frac{\bar{y} - 55}{2}\right)$$

da cui si ricava

$$\frac{\bar{y} - 55}{2} = \Phi^{-1}(1 - 0.37) = \Phi^{-1}(0.63) = 0.33$$

dove il valore di $z_{0.63} = \Phi^{-1}(0.63)$ si ricava dalla tavola della funzione di ripartizione della distribuzione normale standard. Quindi si ottiene $\bar{y} = 55.66$.

c. La potenza del test è la probabilità di rifiutare H_0 se effettivamente è vera l'ipotesi H_1, ovvero $P(\overline{Y} > 57.5|H_1)$. Segue che

$$0.8 = P(\overline{Y} > 57.5|H_1) = 1 - P\left(Z \leq \frac{57.5 - \mu_1}{8/\sqrt{16}}\right) = 1 - \Phi\left(\frac{57.5 - \mu_1}{2}\right)$$

$$\frac{57.5 - \mu_1}{2} = \Phi^{-1}(1 - 0.8) \Rightarrow \mu_1 = 57.5 - 2z_{0.2} = 59.2.$$