Inferenza Statistica

Esame del 26 maggio 2012

Tempo a disposizione 2 ore.

1. La v.a. X, tempo di attesa tra due arrivi a uno sportello di un ufficio espresso in minuti, è distribuita secondo un'esponenziale di parametro λ.

a. E' noto che la probabilità che il tempo tra due arrivi sia maggiore di 5 minuti è 0.75; quanto vale λ?

c. Qual è la probabilità che in 1 minuto vi siano 2 arrivi?

b. Qual è la probabilità che occorrano più di 30 ore per osservare 100 arrivi?

Soluzione

a. $0.75 = P(X > 5) = 1 - F_X(5) = e^{-5\lambda}$, dove F_X è la funzione di ripartizione di una variabile casuale esponenziale. Quindi si ha $\lambda = -\log(0.75)/5 = 0.058$.

b. Se $Y_t = \{\text{numero di arrivi in } [0, t]\}$, si ha che Y_t è distribuita come una Poisson di parametro λt, che ha funzione di probabilità

$$P(Y_t = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \ k = 0, 1, 2, \ldots$$

Posto $t = 1$,

$$P(Y_1 = 2) = \frac{0.058^2}{2!} e^{-0.058} = 0.0016.$$

c. La probabilità cercata (tenendo conto che 30 ore = 1800 minuti) è $P(Y_{t=1800} < 100)$, dove $Y_{t=1800} \sim Po(\lambda t = 104.4)$. La distribuzione di Poisson di parametro ν può essere approssimata da una distribuzione normale di media e varianza ν, per ν elevato. Quindi, approssimativamente, se Z è distribuita secondo una normale standard, si ha

$$P(Y_{t=1800} < 100) \approx P\left(Z < \frac{100 - 104.4}{\sqrt{104.4}}\right) = \Phi(-0.43) = 1 - \Phi(0.43) = 0.33.$$

2. Si dispone di un campione casuale semplice di 3 unità dalla variabile casuale X che ha sia valore atteso che varianza pari a λ. Si considerino i seguenti due stimatori per λ:

$$T = (2X_1 + X_2 + 2X_3)/5 \qquad S = (X_1 + 2X_2 + X_3)/4$$

a. È vero che entrambi gli stimatori sono non distorti per λ?

b. Quale dei due è preferibile?

c. Sapreste indicare uno stimatore migliore?

Soluzione

a. Per valutare la correttezza di uno stimatore dobbiamo determinare il suo valore atteso. Possiamo considerare X_1, X_2, X_3 come variabili aleatorie indipendenti e identicamente distribuite come X, quindi $E(X_i) = E(X) = \lambda$, $\forall i$. Allora

$$E(T) = \frac{1}{5}(2\lambda + \lambda + 2\lambda) = \lambda$$

Analogamente per S, si ha che

$$E(S) = \frac{1}{4}(\lambda + 2\lambda + \lambda) = \lambda$$

ovvero T ed S sono entrambi stimatori non distorti per il parametro λ.

b. Un criterio è quello di confrontare l'errore quadratico medio (EQM) dei due stimatori, che per uno stimatore corretto coincide con la varianza dello stesso. Quindi lo stimatore preferibile sarà quello con varianza minore. Essendo $V(X_i) = V(X) = \lambda$, si ha

$$V(T) = \frac{1}{25}(4V(X_1) + V(X_2) + 4V(X_3)) = \frac{9}{25}\lambda$$

e analogamente

$$V(S) = \frac{1}{16}(V(X_1) + 4V(X_2) + V(X_3)) = \frac{3}{8}\lambda$$

Pertanto, essendo $V(T) < V(S)$, T ha EQM più basso.

c. E' sensato proporre come stimatore la media aritmetica del campione (che pesa egualmente le tre osservazioni)

$$H = \frac{1}{3}(X_1 + X_2 + X_3)$$

Per H si ha $E(H) = \lambda$; pertanto la sua varianza (che è uguale all'errore quadratico medio) è pari a $V(H) = 1/3\lambda$. Essendo $V(H) < V(T) < V(S)$, H è preferibile tanto a T quanto a S.

3. Un campione casuale di dimensione n viene tratto da una variabile Y distribuita normalmente con media μ e varianza nota σ^2. Si vuole verificare l'ipotesi che $H_0 : \mu = \mu_0$ contro $H_1 : \mu = \mu_1 > \mu_0$. Come è noto la regione critica ottima, fissato un determinato valore di α, è del tipo $\bar{y} \geq c$.

a. Si verifichi se il valore c è funzione o meno della differenza fra μ_1 e μ_0.

b. Si dimostri che la probabilità dell'errore di secondo tipo è funzione monotona della differenza fra μ_1 e μ_0.

c. Si immagini ora che si vogliano fissare le due probabilità di errore, α e β. Come dovrebbe essere scelto n?

Soluzione

a. Per il Lemma di Neyman-Pearson, se $L(\mu_0)$ è la funzione di verosimiglianza del campione sotto l'ipotesi H_0, allora la regione R^* del test più potente di ampiezza α per il test $H_0 : \mu = \mu_0$ contro $H_1 : \mu = \mu_1 > \mu_0$ è individuata da $L(\mu_1)/L(\mu_0) \geq k$ dove $P((Y_1, \ldots, Y_n) \in R^* | H_0) = \alpha$. Dunque, si ha

$$\begin{aligned}
\frac{L(\mu_1)}{L(\mu_0)} \geq k &\Rightarrow \exp\left\{-\frac{1}{2\sigma^2}\left[\sum(y_i - \mu_1)^2 - \sum(y_i - \mu_0)^2\right]\right\} \geq k \\
&\Rightarrow \sum y_i \geq \frac{2\sigma^2 \log k + n(\mu_1 - \mu_0)(\mu_1 + \mu_0)}{2(\mu_1 - \mu_0)} \\
&\Rightarrow \bar{y} = \frac{\sum y_i}{n} \geq \frac{1}{2}(\mu_1 + \mu_0) + \frac{\sigma^2 \log k}{n(\mu_1 - \mu_0)} = c
\end{aligned}$$

Ponendo $\alpha = P(\overline{Y} \geq c | H_0)$ ed essendo $\overline{Y} \sim N(\mu_0, \sigma^2/n)$, si ricava

$$\Phi\left(\frac{c - \mu_0}{\sigma/\sqrt{n}}\right) = 1 - \alpha \Rightarrow c = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}},$$

dove $z_{1-\alpha}$ è il quantile della variabile normale standard Z di ordine $1 - \alpha$. Pertanto c dipende solo da μ_0.

b. La probabilità dell'errore di secondo tipo β è data da

$$P(\overline{Y} \leq c|\mu = \mu_1) = P\left(Z \leq \frac{c-\mu_1}{\sigma/\sqrt{n}}\right) = \Phi\left(z_{1-\alpha} - \frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}\right),$$

dove si è sostituito il valore $c = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$. Posto $\delta = \mu_1 - \mu_0$, la probabilità dell'errore di secondo tipo è funzione monotona decrescente di δ (cioè della differenza fra μ_1 e μ_0), essendo $\Phi(z)$ una funzione monotona crescente di z.

c. Fissati $\alpha = \alpha_0$ e $\beta = \beta_0$, si ottiene

$$\beta_0 = \Phi\left(z_{1-\alpha_0} - \frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}\right),$$

la cui soluzione rispetto ad n, applicando la funzione $\Phi^{-1}(\cdot)$ ad ambo i membri, risulta essere

$$n = \sigma^2\left(\frac{z_{1-\alpha_0} - z_{\beta_0}}{\mu_1-\mu_0}\right)^2$$

dove si è posto $\Phi^{-1}(\beta_0) = z_{\beta_0}$; si può quindi porre n uguale al numero intero non minore del valore sopra ottenuto.

4. Un'azienda vuole valutare di quanto siano aumentate le vendite dei suoi prodotti a seguito di una campagna pubblicitaria. A tal fine, si osservano le vendite settimanali per 10 settimane prima della campagna pubblicitaria e 5 settimane dopo la campagna pubblicitaria. Si assume che le osservazioni fatte prima e dopo siano determinazioni indipendenti da distribuzioni gaussiane. Le vendite medie nei due periodi sono state 15340 e 21224 euro, si assume inoltre che la varianza è nota e in entrambi i casi pari a 3000^2.

a. Si fornisca un intervallo di confidenza al 95% per l'incremento della vendita settimanale media valutato dopo la campagna pubblicitaria.

b. Si dica quanto dovrebbe essere grande il campione relativo alle vendite post campagna pubblicitaria per ridurre l'ampiezza dell'intervallo di confidenza ad al più 5000 euro.

c. Dovendo comunicare il risultato dell'analisi alla filiale americana della ditta, si vuole l'intervallo di confidenza per la differenza tra le vendite medie espresse in dollari (per rispondere si assuma che un euro valga 1.5 dollari).

Soluzione

a. Consideriamo i due campioni $(X_1, \ldots, X_{n_1})$, $(Y_1, \ldots, Y_{n_2})$ relativi alle osservazioni prima e dopo la campagna pubblicitaria, provenienti da popolazioni normali con rispettive medie μ_1, μ_2 e varianza nota $\sigma^2 = 3000^2$. Posto $\alpha = 0.05$, l'intervallo di confidenza cercato è

$$\left((\bar{Y} - \bar{X}) - z_{1-\frac{\alpha}{2}}\sigma_{\bar{Y}-\bar{X}}, (\bar{Y} - \bar{X}) + z_{1-\frac{\alpha}{2}}\sigma_{\bar{Y}-\bar{X}}\right)$$

dove $\sigma_{\bar{Y}-\bar{X}} = \sqrt{\sigma^2/n_1 + \sigma^2/n_2}$. Sostituendo i valori osservati dai due campioni $\bar{x} = 15340$, $\bar{y} = 21224$, $n_1 = 10$, $n_2 = 5$, ed essendo $z_{1-\alpha/2} = 1.96$, si ottengono gli estremi dell'intervallo di livello 0.95, $5884 \pm 1.96 \cdot 1643.2$, per cui si trova l'intervallo per $\mu_2 - \mu_1$

$$(2663.33, 9104.67)$$

b. Posto $A = 5000$, si determina n_2 ponendo

$$5000 = 2(1.96)\,\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

da cui $n_2 = [(2500/(3000 \cdot 1.96))^2 - (1/10)]^{-1} = 12.38$. Pertanto occorre considerare un periodo di 13 settimane.

c. L'intervallo di confidenza per la differenza tra le vendite medie espresse in dollari si ottiene da $(2663.33 \cdot 1.5, 9104.67 \cdot 1.5)$ che corrisponde a $(3995\,\$, 13657\,\$)$.