Inferenza Statistica

Esame del 9 febbraio 2015

Tempo a disposizione 2 ore.

1. Il numero di chiamate al soccorso stradale che si verificano giornalmente segue una legge di Poisson. Tuttavia è noto che in media nei giorni feriali arrivano il doppio delle chiamate rispetto a sabato e domenica. In un campione di giorni si sono osservate le chiamate e in 65 giorni feriali si sono avute in totale 132 chiamate. Nei 15 giorni festivi le chiamate sono state in totale 56.

a. Si fornisca la stima di massima verosimiglianza delle chiamate medie giornaliere nei giorni festivi.

b. Si verifichi se lo stimatore ottenuto al punto **a.** è corretto.

c. Si fornisca una stima della probabilità che in un giorno feriale arrivino 2 o più chiamate.

Soluzione

a. Si consideri il campione $Y_1, \dots, Y_n$, dove $n = n_1 + n_2$, $n_1 = 65, n_2 = 15$; Y_i $(i = 1, \dots, n_1)$ sono variabili aleatorie di Poisson indipendenti di media 2λ, $n - n_1$ sono variabili aleatorie di Poisson indipendenti, ciascuna con valore atteso λ. La funzione di verosimiglianza per il campione $(y_1, \dots, y_n)$ è

$$f(y_1, \dots, y_n; \lambda) = L(\lambda) = \prod_{i=1}^{n_1} \frac{e^{-2\lambda}(2\lambda)^{y_i}}{y_i!} \prod_{i=n_1+1}^{n_1+n_2} \frac{e^{-\lambda}(\lambda)^{y_i}}{y_i!}$$
$$= \frac{(2\lambda)^{\sum_{i=1}^{n_1} y_i} e^{-2n_1\lambda}}{y_1! \cdots y_{n_1}!} \frac{\lambda^{\sum_{i=n_1+1}^{n_1+n_2} y_i} e^{-n_2\lambda}}{y_{n_1+1}! \cdots y_{n_1+n_2}!}.$$

Passando al logaritmo, si ha

$$l(\lambda) = \sum_{i=1}^{n_1} y_i \log(2\lambda) - 2n_1\lambda - \log c_1 + \sum_{i=n_1+1}^{n_1+n_2} y_i \log(\lambda) - n_2\lambda - \log c_2$$

dove $c_1 := y_1! \cdots y_{n_1}!$ e $c_2 := y_{n_1+1}! \cdots y_{n_1+n_2}!$ non dipendono da λ. Derivando si trova che

$$\frac{d}{d\lambda} l(\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n_1} y_i - 2n_1 + \frac{1}{\lambda} \sum_{i=n_1+1}^{n_1+n_2} y_i - n_2.$$

Infine, eguagliando a zero la funzione *score* e risolvendo rispetto a λ si ha

$$\hat{\lambda}_{MV} = \frac{\sum_{i=1}^{n_1} y_i + \sum_{i=n_1+1}^{n_1+n_2} y_i}{2n_1 + n_2}$$

b. $E(\hat{\lambda}_{MV}) = \dfrac{1}{2n_1 + n_2} \left(\sum_{i=1}^{n_1} E(Y_i) + \sum_{i=n_1+1}^{n_1+n_2} E(Y_i) \right) = \dfrac{1}{2n_1 + n_2} (2n_1\lambda + n_2\lambda) = \lambda.$

Pertanto lo stimatore trovato è corretto.

c. Si denoti con p la probabilità che in un giorno feriale arrivino 2 o più chiamate. Allora $p = P(Y \geq 2)$, $Y \sim Po(2\lambda)$. Lo stimatore di MV della probabilità p è

$$\hat{p}_{MV} = 1 - P(Y < 2) = 1 - P(Y = 0) - P(Y = 1) = 1 - e^{-2\hat{\lambda}_{MV}} - 2\hat{\lambda}_{MV} e^{-2\hat{\lambda}_{MV}},$$

in virtù della proprietà di equivarianza degli stimatori di MV. Essendo $\hat{\lambda} = \dfrac{132 + 56}{145} = 1.297$ la stima di massima verosimiglianza per λ, si ottiene $\hat{p} = 0.73$, dall'espressione sopra riportata.

2. Siano ${}_AY$ e ${}_BY$ le lunghezze dei chiodi prodotti da due diverse aziende, A e B. Si dispone di un campione casuale di 10 elementi dall'azienda A dal quale risulta:

$$10\,{}_A\bar{y} = \sum_i^{10} {}_Ay_i = 20; \quad \sum_i^{10} ({}_Ay_i - {}_A\bar{y})^2 = 20$$

Dalla seconda azienda si ha un campione di 6 elementi e si trova

$$6\,{}_B\bar{y} = \sum_i^{6} {}_By_i = 13.2; \quad \sum_i^{6} ({}_By_i - {}_B\bar{y})^2 = 20.5$$

Si voglia verificare l'ipotesi che le lunghezze medie delle due produzioni siano uguali contro l'alternativa che siano diverse.

a. Si supponga dapprima che varianze delle lunghezze delle viti nelle due popolazioni siano ignote. Sotto quali assunzioni si è in grado di verificare l'ipotesi data sopra? Nel caso sia possibile, condurre la verifica di ipotesi al livello $\alpha = 0.02$.

b. Per il caso al punto **a.** stabilire se il relativo p-value è superiore o inferiore a 0.1.

c. Se invece le varianze fossero e pari a 3 e 4.5, rispettivamente, quali assunzioni sarebbero necessarie per condurre la verifica di ipotesi di uguaglianza fra le medie? Sotto tali condizioni, si ottenga il p-value per la verifica di ipotesi data sopra.

Soluzione

a. Supponendo che le varianze σ_A^2 e σ_B^2 siano ignote, si assuma che i campioni di cui si dispone di ampiezza $n_A = 10$ e $n_B = 6$, rispettivamente, siano indipendenti e provenienti da popolazioni normali con varianze non note e uguali $\sigma_A^2 = \sigma_B^2 = \sigma^2$. Si vuole verificare l'ipotesi

$$H_0 : \mu_A - \mu_B = 0$$

contro l'alternativa $H_1 : \mu_A - \mu_B \neq 0$ al livello $\alpha = 0.02$. Le medie campionarie sono

$${}_A\bar{y} = \sum_i^{10} {}_Ay_i/10 = 2, \quad {}_B\bar{y} = \sum_i^{6} {}_By_i/6 = 2.2$$

Sotto l'ipotesi nulla H_0, si ha che la statistica T ha distribuzione t di Student con $n_A + n_B - 2 = 14$ gradi di libertà:

$$T = \frac{{}_A\bar{Y} - {}_B\bar{Y}}{S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} \sim t_{14}$$

dove

$$S_p = \sqrt{\frac{\sum_{i=1}^{10}({}_Ay_i - {}_A\bar{y})^2 + \sum_{i=1}^{6}({}_By_i - {}_B\bar{y})^2}{n_A + n_B - 2}}$$

è lo stimatore della varianza comune nell'ipotesi di omoschedasticità. Poiché si tratta di un test bilaterale, la zona di rifiuto del test basata sul valore osservato t di T è data da $R = \{t : |t| \geq t_{1-\alpha/2}\}$. Dai dati campionari si trova

$$t = \frac{2 - 2.2}{\sqrt{\frac{20+20.5}{14}\left(\frac{1}{10} + \frac{1}{6}\right)}} = -0.23.$$

ed essendo $|t| < t_{1-\alpha/2} = t_{0.99,14} = 2.624$, si decide di non rifiutare H_0 al livello del 2%.

b. Il p-value per il test al punto **a.** è calcolato come

$$\gamma = P(|T| \geq |t|\,|H_0) = 2(1 - P(T \leq 0.23))$$

con $T \sim t_{14}$. Dalle tavole si trova che $t_{1-0.1/2,14} = t_{0.95,14} \approx 1.76$, pertanto $P(T > 0.23)$ è molto maggiore di 0.05 e quindi il p-value dato da $2P(T > 0.23)$ è molto maggiore di 0.1 (se si ricorresse all'uso di `R`, si troverebbe $\gamma \approx 0.82$ confermando che i dati non danno evidenza contro H_0).

c. Assumendo che i due campioni provengano da popolazioni distribuite normalmente con medie μ_A, μ_B e varianze $\sigma_A^2 = 3$ e $\sigma_B^2 = 4.5$, si ha che, sotto H_0,

$$Z = \frac{{}_A\bar{Y} - {}_B\bar{Y}}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \sim N(0,1)$$

e dai dati si ottiene il valore

$$z = \frac{2 - 2.2}{\sqrt{\frac{3}{10} + \frac{4.5}{6}}} = -0.195.$$

Data la regione di rifiuto del test $R = \{z : |z| \geq 2.326\}$, si ha che $z \notin R$, e quindi l'ipotesi nulla non viene rifiutata al livello del 2%. Il p-value è dato da

$$\begin{aligned} P(|Z| \geq |z|\,|H_0) &= 2P(Z \geq 0.195) \\ &= 2(1 - \Phi(0.195)) \\ &= 0.85. \end{aligned}$$

3. Per verificare la durata delle lampadine prodotte da un'azienda si registrano i tempi di funzionamento (in giorni) per un campione casuale semplice di 190 lampadine. La tabella che segue riporta i dati ottenuti.

Durata in giorni	da 0 a 11	da 11 a 22	da 22 a 33	oltre 33
Frequenze	118	47	12	13

Sia inoltre noto che la somma delle durate di tutte le lampadine che durano oltre 33 giorni è pari a 537 giorni.

a. Si verifichi l'ipotesi che i tempi di durata delle lampadine sono distribuiti secondo una legge esponenziale (al livello $\alpha = 0.05$).

b. Si dica se il p-value per il test del punto **a.** è superiore a 0.2.

Soluzione

a. Si tratta di verificare l'ipotesi che la variabile Y che descrive i tempi di durata delle lampadine segua la distribuzione esponenziale:

$$H_0 : f(y; \lambda) = \lambda e^{-\lambda y}, \ \ y > 0$$

Occorre prima stimare λ. la stima di massima verosimiglianza di λ con i dati del campione osservato è il reciproco della media ponderata dei dati del campione. Posto $n = 190$, denotate con n_i le frequenze osservate e con $\tilde{y}_i$ i valori centrali in ogni classe, si ha

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{4} n_i\tilde{y}_i = \frac{118(11/2) + 47((11 + 22)/2) + 12((22 + 33)/2) + 537}{190} = 12.06053$$

Pertanto la stima di massima verosimiglianza di λ è $\hat{\lambda} = 1/12.06053 = 0.083$.
Le frequenze relative possono ora essere calcolate in corrispondenza di ciascuna classe. Ad esempio, per la classe $11 - 22$ si ha

$$P(11 \leq X \leq 22) = F(22) - F(11) = 1 - e^{-0.083\cdot 22} - (1 - e^{-0.083\cdot 11}) = 0.240$$

e la corrispondente frequenza attesa è $n \times 0.240 = 45.6$. I dati ottenuti sono riportati nella tabella seguente.

Durata in giorni	Freq. oss. n_i	Freq. relative p_{i0}	Freq. attese $n \cdot p_{i0}$
da 0 a 11	118	0.599	113.81
da 11 a 22	47	0.240	45.6
da 22 a 33	12	0.096	18.24
>33	13	0.065	12.35
Tot	190	1	190

Per condurre la verifica di ipotesi calcoliamo infine

$$\chi^2 = \frac{(118 - 113.81)^2}{113.81} + \frac{(47 - 45.6)^2}{45.6} + \frac{(12 - 18.24)^2}{18.24} + \frac{(13 - 12.35)^2}{12.35} = 2.37.$$

Posto $\alpha = 0.05$ l'ipotesi nulla non viene rifiutata poiché $\chi^2 = 2.37 < \chi^2_{1-0.05,2} = 5.99$ (i gradi di libertà sono (4-1)-1=2, essendo uno il parametro stimato con i dati del campione).

b. Il p-value per il test al punto **a.** è espresso da $P(Y \geq 2.37)$, dove $Y \sim \chi^2_2$. Dalle tavole si trova che $P(Y \geq 1.39) \approx 0.5$ e $P(Y \geq 2.77) \approx 0.25$, per cui la $P(Y \geq 2.37)$ sarà poco superiore a 0.25, e certamente compresa tra 0.25 e 0.5 (utilizzando `R` si trova che tale valore è 0.31).

4. I rendimenti annuali di tre titoli azionari X, Y e Z sono variabili aleatorie Gaussiane le cui medie sono rispettivamente 0.02, 0.01 e 0.05 e le varianze sono pari a 0.003, 0.001, 0.005. La matrice di correlazione è

$$\rho = \begin{bmatrix} 1 & 0.5 & 0.3 \\ 0.5 & 1 & 0.2 \\ 0.3 & 0.2 & 1 \end{bmatrix}$$

a. Calcolare la probabilità che un portafoglio di titoli costituito per il 60% da azioni X e per il 40% da azioni Y abbia rendimento totale positivo.

b. Ha più probabilità di rendimento positivo il portafoglio definito sopra oppure uno composto esclusivamente da azioni Z?

Soluzione

a. Si denoti con R il rendimento di un portafoglio costituito per il 60% da azioni X e per il 40% da azioni Y, pertanto $R = 0.6R_X + 0.4R_Y$, dove $R_X \sim N(0.02, 0.003)$ ed $R_Y \sim N(0.01, 0.001)$ rappresentano i rendimenti annuali dei titoli X e Y, rispettivamente.

Poiché R è combinazione lineare di v.a. normali, si ricava che R sarà distribuita normalmente con

$$E(R) = 0.6 \cdot E(R_X) + 0.4 \cdot E(R_Y) = 0.016,$$

$$V(R) = 0.6^2 \cdot V(R_X) + 0.4^2 \cdot V(R_Y) + 2 \cdot 0.6 \cdot 0.4 \cdot Cov(R_X, R_Y) = 0.00304,$$

essendo $Cov(R_X, R_Y) = \rho_{XY}\sqrt{V(R_X)V(R_Y)} = 0.0009$. La probabilità richiesta è data da

$$P(R > 0) = 1 - P\left(Z < \frac{0 - 0.016}{\sqrt{0.00304}}\right) = 1 - \Phi(-0.29) \approx 0.61,$$

dove con Z si denota una variabile casuale normale standardizzata e $\Phi(\cdot)$ è la sua funzione di ripartizione.

b. Un portafoglio composto esclusivamente da azioni Z ha rendimento positivo con probabilità

$$P(R_Z > 0) = 1 - \Phi\left(\frac{0 - E(R_Z)}{\sqrt{V(R_Z)}}\right) = 1 - \Phi\left(\frac{0 - 0.05}{\sqrt{0.005}}\right) = 1 - \Phi(-0.707) \approx 0.76 > 0.61.$$