STATISTICA per ING INF TEL [2L] I. Epifani, A. Pievatolo 06.09.10

Nello svolgere gli esercizi fornire passaggi e spiegazioni: non bastano i risultati finali.

Esercizio A Un database è formato da n blocchi di memoria, ciascuno dei quali contiene m record. Ogni volta che un record viene cambiato, viene riscritto anche il blocco nel quale esso si trova. Vogliamo analizzare il numero di blocchi riscritti e di record cambiati ogni settimana, sotto l'ipotesi che in una settimana ciascun record possa venire cambiato, indipendentemente dagli altri, con probabilità p.

1. Per $i = 1, \ldots, n$, sia
$$X_i = \begin{cases} 1 & \text{se il blocco } i \text{ viene riscritto} \\ 0 & \text{altrimenti} \end{cases}$$
e sia $\theta = \Pr(X_i = 1)$. In una data settimana osserviamo $X_1, \ldots, X_n$. Scegliendo l'appropriato modello generatore dei dati in base alle ipotesi fatte finora, trovare lo stimatore di massima verosimiglianza di θ e dimostrarne l'efficienza.
2. Mostrare che $\theta = 1 - (1-p)^m$.
3. Determinare lo stimatore di massima verosimiglianza di p e mostrare che non è efficiente.
4. Siano $m = 10$, $n = 100$ e $\sum_{i=1}^{100} X_i = 92$. Calcolare un intervallo di confidenza al 99.87% che abbia forma $(c, +\infty)$ per θ, quindi ricavarne uno per la percentuale dei record cambiati in una settimana.

Soluzione

1. Per l'ipotesi di indipendenza sul cambiamento dei record, possiamo affermare che anche i blocchi vengono riscritti l'uno indipendentemente dall'altro, perciò $X_1, \ldots, X_n$ costituisce un campione bernoulliano con parametro θ. Poiché
$$\frac{\partial \ln L_\theta}{\partial \theta} = \frac{n}{\theta(1-\theta)}(\bar{x} - \theta)$$
allora $\hat{\theta} = \bar{X}$ è lo stimatore efficiente di θ e coincide con lo stimatore di massima verosimiglianza.

2. Un blocco non viene riscritto se e solo se nessuno dei suoi m record viene cambiato. La probabilità che nessun record venga cambiato è pari a $(1-p)^m$, pertanto $\theta = 1 - (1-p)^m$.

3. Osserviamo che $p = 1 - (1-\theta)^{\frac{1}{m}} = \kappa(\theta)$. Per la proprietà di invarianza dello stimatore di massima verosimiglianza, $\hat{p} = \kappa(\hat{\theta})$. Per il punto 1, sono stimabili in modo efficiente soltanto le combinazioni lineari di θ, dunque lo stimatore ML di p non è efficiente.

4. Un intervallo di confidenza per p si può determinare partendo da un intervallo asintotico per θ e trasformando tale intervallo, oppure direttamente attraverso la distribuzione asintotica dello stimatore di massima verosimiglianza di p. In questo secondo intervallo, viene introdotta un'ulteriore approssimazione per il calcolo della varianza di $\hat{p}$, mentre invece quella di $\hat{\theta}$ è esatta.

Considerato che $z_{0.0013} = -3$, l'intervallo per θ ottenuto con il primo metodo è
$$\left(\bar{x} - 3\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}, 1\right) = (0.92 - 3 \times 0.027122932), 1) = (0.92 - 0.08138796, 1) = (0.8338612, 1).$$

Applicando la trasformazione $k(\theta)$, l'intervallo per p risulta:

$$\left(1-(1-0.8338612)^{\frac{1}{10}}, 1-(1-1)^{\frac{1}{10}}\right) = (0.1643063, 1)$$

Con il secondo metodo, sappiamo che la varianza di $\hat{p} = \kappa(\hat{\theta})$ è approssimabile con

$$\frac{\kappa'(\theta)^2}{nI(\theta)} = \frac{\theta(1-\theta)}{n}\frac{1}{m^2}(1-\theta)^{2/m-2} = \frac{\theta}{(1-\theta)^{1-2/m}nm^2}$$

che stimiamo inserendo $\hat{\theta}$ al posto di θ, ottenendo il valore

$$\frac{1}{100}\sqrt{\frac{0.92}{0.08^{0.8}}} = 0.02634256$$

che dà luogo all'intervallo di confidenza

$$\left(\hat{p} - 3 \times \sqrt{\frac{\kappa'(\theta)^2}{nI(\theta)}}, 1\right) = (0.2232004 - 3 \times 0.02634256, 1) = (0.1441727, 1).$$

Esercizio B La home page di un docente di statistica ha ricevuto, nelle cinque settimane dal 7/5/2010 al 10/6/2010, un numero di visite uniche settimanali pari a $(15, 12, 10, 13, 14)$. Le visite uniche nelle cinque settimane successive, cioè dall'11/6/2010 al 15/7/2010, sono state invece $(31, 24, 54, 17, 73)$. Le seconde cinque settimane corrispondono all'incirca al periodo che va dalla fine delle lezioni alla fine della sessione estiva di esami.

1. Con un opportuno test statistico non parametrico, stabilire se il sito del docente tende a ricevere più visite attorno alla sessione d'esame.

2. Verificare se nelle settimane ordinarie, cioè al di fuori dei periodi comprendenti le sessioni d'esame, il numero X di visite settimanali possa avere la seguente distribuzione di probabilità

$$\Pr(X = i) = \begin{cases} \frac{1}{6} & \text{se} \quad i = 10, 11, 12, 13, 14, 15 \\ 0 & \text{altrimenti} \end{cases}$$

   sulla base del seguente campione di 50 settimane ordinarie

N. di visite	10	11	12	13	14	15
N. di settimane	6	13	9	12	4	6

   In base al risultato della verifica, proporre un valore per il numero atteso di visite settimanali.

Soluzione

1. Facciamo l'ipotesi che le visite settimanali costituiscano dei campioni indipendenti. Sia X la v.a. del numero di visite nelle prime cinque settimane, con distribuzione F e sia Y la v.a. del numero di visite

nelle seconde cinque settimane, con distribuzione G. Per verificare che il numero di visite sia superiore durante la sessione d'esame, utilizziamo il test dei ranghi di Wilcoxon, formulando l'ipotesi

$$\begin{cases} H_0: & F = G \\ H_1: & F > G \end{cases}$$

I cinque valori di X sono tutti più piccoli del minimo valore di Y, pertanto $T_X = 15$, il minimo valore assumibile dalla somma dei ranghi. Il rifiuto dell'ipotesi nulla avviene per valori piccoli di T_X. Dalle tavole si ricava che, con $m = n = 5$, $w_{0.005} = 16$ e che $w_{0.001} = 15$, quindi l'ipotesi nulla viene rifiutata per ogni livello superiore allo 0.1%.

2. Impostiamo un test χ^2 di buon adattamento. La statistica test è pari a

$$Q = \frac{6^2 + 13^2 + 9^2 + 12^2 + 4^2 + 6^2}{50/6} - 50 = 57.84 - 50 = 7.84.$$

Il p-value associato al test è pari alla probabilità che una v.a. χ^2_5 sia superiore a 7.84, valore compreso tra $\chi^2_5(0.800) = 7.289$ e $\chi^2_5(0.875) = 8.625$, dunque il p-value si trova tra il 20% e il 12.5%. Non potendo rifiutare H_0, un valore del numero atteso di visite si ricava dalla media della distribuzione sotto H_0: $(10 + 11 + 12 + 13 + 14 + 15)/6 = 12.5$.

Esercizio C Due gruppi formati ciascuno da 21 programmatori, il gruppo xx e il gruppo yy devono consegnare un software. Per ciascun programmatore viene contato il numero medio giornaliero di righe di programma scritte, ottenendo i due campioni casuali $(x_1, \ldots, x_{21})$ e $(y_1, \ldots, y_{21})$ tra loro indipendenti. Di tali campioni abbiamo a disposizione le statistiche

$$\sum_{i=1}^{21} x_i = 2040.37 \quad \sum_{i=1}^{21} x_i^2 = 227134 \quad \sum_{i=1}^{21} y_i = 1886.661 \quad \sum_{i=1}^{21} y_i^2 = 187065.7$$

e assumiamo che il numero medio giornaliero di righe scritte abbia distribuzione gaussiana.

1. Verificare, al livello del 10%, l'ipotesi che la varianza σ^2_X di un programmatore del gruppo xx sia uguale alla varianza σ^2_Y di un programmatore del gruppo yy, contro l'ipotesi che la prima sia maggiore della seconda.
2. Calcolare la funzione di potenza del test precedente nel punto $\sigma^2_X/\sigma^2_Y = 1.25$.
3. Stimare con il metodo dei momenti la differenza tra il numero medio giornaliero di righe scritte da un programmatore del gruppo xx e quelle scritte da un programmatore del gruppo yy: $\Delta = \mathrm{E}(X) - \mathrm{E}(Y) = \mu_X - \mu_Y$. Verificare che lo stimatore sia non distorto e consistente. Proporre una stima per la varianza dello stimatore di Δ.

Soluzione

1. Utilizziamo il test F del rapporto tra le varianze campionarie con media incognita, la cui statistica test sotto H_0 ha distribuzione F di Fisher con 20 gradi di libertà al numeratore e al denominatore. Il valore osservato della statistica è

$$F = \frac{s^2_X}{s^2_Y} = \frac{\sum x_i^2 - 21\bar{x}^2}{\sum y_i^2 - 21\bar{y}^2} = \frac{28890.66}{17566.15} = 1.645.$$

L'ipotesi nulla viene rifiutata per valori alti della statistica e, con il livello dato, se essa supera $q_{20,20}(0.90) = 1.79$, vale a dire il 90^o percentile della F. Il valore osservato della statistica test non consente di rifiutare l'ipotesi nulla che le varianze siano uguali.

2. La funzione di potenza si ottiene calcolando la probabilità che la statistica test cada nella regione di rifiuto quando $\theta = \sigma_X^2/\sigma_Y^2 = 1.25$:

$$P_{\theta=1.25}(F > 1.79) = P_{\theta=1.25}\left(\frac{F}{1.25} > \frac{1.79}{1.25}\right) = P(F_{20,20} > 1.43)$$

dove con $F_{20,20}$ abbiamo indicato una F di Fisher con 20 gradi di libertà al numeratore e al denominatore. Nelle tavole troviamo che $P(F_{20,20} \leq 1.36) = 0.75$ e che $P(F_{20,20} \leq 1.47) = 0.80$, e quindi la funzione di potenza nel punto dato è compresa tra 0.20 e 0.25.

3. Osserviamo che $\Delta = \mathrm{E}(X - Y)$. Pertanto il suo stimatore con il metodo dei momenti si ottiene eguagliando il valore atteso al suo corrispettivo campionario:

$$\Delta = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) = \bar{X} - \bar{Y} = T.$$

dove n indica la dimensione del campione. Lo stimatore è non distorto, infatti:

$$\mathrm{E}(T) = \mathrm{E}(\bar{X}) - \mathrm{E}(\bar{Y}) = \mu_X - \mu_Y$$

e, per le assunzioni di indipendenza,

$$\mathrm{Var}(T) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n}$$

che converge a zero al crescere di n, dunque T è consistente. In corrispondenza del campione osservato, la stima di Δ è pari a $\bar{x} - \bar{y} = 7.319$.

Per quanto riguarda la varianza, secondo il test al punto 1, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, dunque $\mathrm{Var}(T) = 2\sigma^2/n$, che possiamo stimare usando lo stimatore della varianza *pooled*, $2s_p^2/n$, dove

$$s_p^2 = \frac{(\sum x_i^2 - n\bar{x}^2) + (\sum y_i^2 - n\bar{y}^2)}{2n - 2} = \frac{28890.66 + 17566.15}{40} = 1161.420$$

ottenendo infine il valore $2s_p^2/n = 2 \times 1161.420/21 = 110.6114$.