STATISTICA per ING INF TEL [2L]   A. Pievatolo, scaglione [M-Z]   13.07.10

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

Esercizio 1 Abbiamo estratto un campione casuale $X_1, \dots, X_{25}$ dalla densità di probabilità

$$f(x,\theta) = \frac{1}{\theta} \mathrm{e}^{1-\frac{x}{\theta}} \mathbf{1}_{(\theta,\infty)}(x),$$

con θ parametro positivo incognito, e abbiamo ottenuto la seguente funzione di ripartizione empirica:

x	2.6	3.8	4.5	5.5	6.6	7.4	19.3
$\hat{F}_{25}(x)$	3/25	12/25	16/25	19/25	22/25	24/25	1

(1)

Indichiamo con $\overline{X}$ la media campionaria.

1. Calcolate $\mathrm{E}(\overline{X})$ e $\mathrm{Var}(\overline{X})$.
2. Costruite uno stimatore non distorto per θ partendo da $\overline{X}$ e calcolatene l'errore quadratico medio.
3. Usate la funzione di ripartizione empirica (1) per fornire una stima di θ (valore numerico dello stimatore trovato al punto precedente) e una stima del suo errore quadratico medio.
4. Determinate lo stimatore di massima verosimiglianza di θ e fornitene anche il valore numerico per il campione di dati in (1).

Soluzione

1. Poiché

$$\begin{aligned}
\mathrm{E}_\theta(X_1) &= \int_\theta^\infty x \times \frac{1}{\theta}\mathrm{e}^{1-\frac{x}{\theta}}\, dx = \ldots = \theta + \theta \mathrm{e}^1 \mathrm{e}^{-\theta/\theta} = 2\theta;\\
\mathrm{E}_\theta(X_1^2) &= \int_\theta^\infty x^2 \times \frac{1}{\theta}\mathrm{e}^{1-\frac{x}{\theta}}\, dx = \ldots = \theta^2 + 2\theta(2\theta) = 5\theta;^2\\
\mathrm{Var}_\theta(X_1) &= 5\theta^2 - (2\theta)^2 = \theta^2
\end{aligned}$$

allora $\mathrm{E}(\overline{X}) = 2\theta$ e $\mathrm{Var}(\overline{X}) = \dfrac{\theta^2}{25}$

2. Da $\mathrm{E}(\overline{X}) = 2\theta$ deduciamo che $\hat{\theta} = \overline{X}/2$ ha media θ e quindi è stimatore non distorto di θ. Inoltre, $MSE(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) = \mathrm{Var}(\overline{X}/2) = \theta^2/(4n) = \theta^2/100$.

3. Da $\hat{F}_{25}(x)$ ricaviamo le frequenze relative dei valori campionari: per esempio 2.6 ha frequenza 3/25, 3.8 ha frequenza $\hat{F}_{25}(3.8) - \hat{F}_{25}(2.6) = 9/25$, eccetera. Allora

$$\bar{x} = \frac{2.6 \times 3 + 3.8 \times 9 + 4.5 \times 4 + 5.5 \times 3 + 6.6 \times 3 + 7.4 \times 2 + 19.3 \times 1}{25} = 5.216$$

e quindi $\hat{\theta} = 5.216/2 = 2.608$ e la stima di $MSE(\hat{\theta})$ risulta $2.608^2/100 = 0.06801664$.

4. La funzione di verosimiglianza è

$$L_\theta(x_1, \ldots, x_{25}) = \frac{e^{25-\frac{\sum_i x_i}{\theta}}}{\theta^{25}} \mathbf{1}_{(\theta,+\infty)}(\min\{x_1, \ldots, x_n\})$$

il cui punto di massimo non vincolato è in $\bar{x}$. Poiché è sempre vero che $\bar{x} \geq \min\{x_1, \ldots, x_n\}$ e θ non può superare il più piccolo valore osservato, allora $\hat{\theta}_{ML} = \min\{x_1, \ldots, x_n\} = 2.6$ ∎

Esercizio 2 Un'azienda di succhi di frutta usa un macchinario che versa il succo nelle bottigliette in un processo produttivo continuo. Il macchinario lavora secondo gli standard quando versa in ciascuna bottiglietta esattamente 33 ml di succo. Periodicamente si selezionano 49 bottigliette, si misura la media campionaria $\bar{X}$ del contenuto delle 49 bottigliette e si conclude che il macchinario NON rispetta gli standard se $\bar{X}$ si scosta -in più o in meno- di 2.10 ml da 33 ml. Se ciò si verifica, si interrompe la produzione. Inoltre si sa che le misurazioni del contenuto di succo nelle bottigliette sono variabili aleatorie gaussiane con deviazione standard $\sigma = 6.4$ ml.

1. Formalizzate con il linguaggio della verifica di ipotesi la procedura di controllo del funzionamento del macchinario sopra descritta, cioè traducetela mediante un test di ipotesi.

2. Calcolate il livello di significatività α del test di ipotesi formalizzato al punto 1.

3. Fornite l'espressione analitica della funzione di potenza della procedura di verifica di ipotesi descritta.

4. Calcolate la probabilità di "NON interrompere la produzione" quando in realtà il macchinario versa mediamente 3.0 ml di succo IN MENO dei 33 ml regolamentari.

5. Determinate quante ulteriori bottigliette, oltre alle 49, bisogna controllare affinché la lunghezza dell'intervallo di confidenza bilatero al 97.86% del contenuto medio di succo sia minore o uguale di 2.10 ml.

Soluzione

1. Le 49 misure $X_1, \dots, X_{49}$ sono un campione casuale (variabili aleatorie i.i.d.) da una popolazione $\mathcal{N}(\mu; 6.4^2)$; il parametro incognito è la media $\mu \in \mathbb{R}$. Le ipotesi sono $H_0 : \mu = \mu_0 \equiv 33$ contro $H_1 : \mu \neq \mu_0$. La regola di interruzione data nel testo si traduce nella regione critica $\{ |\overline{X} - \mu_0| \geq 2.10 \}$, con $\overline{X} = \frac{1}{49}\sum_{j=1}^{49} X_j$.

2. Il livello cercato è

$$\alpha = P_{\mu=33}\left(\left|\overline{X} - 33\right| \geq 2.10\right) = 2\left[1 - \Phi\left(\frac{2.10}{6.4/\sqrt{49}}\right)\right] = 2\left[1 - \Phi\left(2.296875\right)\right] \simeq 2\left[1 - \Phi\left(2.30\right)\right] \simeq 2.14\%.$$

3. La funzione di potenza è la probabilità di rifiutare H_0 se $\mu \neq 33$, cioè

$$\pi(\mu) = P_\mu\left(\left|\overline{X} - 33\right| \geq 2.10\right) = 1 - P_\mu\left(-2.10 \leq \overline{X} - 33 \leq 2.10\right) =$$
$$= 1 - \left[\Phi\left(\frac{35.10 - \mu}{6.4/\sqrt{49}}\right) - \Phi\left(\frac{30.90 - \mu}{6.4/\sqrt{49}}\right)\right], \qquad \mu \neq 33.$$

4. Dobbiamo calcolare la probabilità di errore di secondo tipo quando $\mu = 30.0$; abbiamo

$$1 - \pi(30) = \Phi\left(\frac{35.1 - 30}{6.4/\sqrt{49}}\right) - \Phi\left(\frac{30.9 - 30}{6.4/\sqrt{49}}\right) = \Phi(5.578125) - \Phi(0.984375) \simeq 1 - 0.8375 = 16.25\%.$$

5. Gli estremi di un intervallo bilatero sono $\overline{x}_n \pm z_{.9893}\frac{6.4}{\sqrt{n}}$ e la lunghezza è

$$2 \times z_{0.9893}\frac{6.4}{\sqrt{n}} \simeq 2 \times 2.30 \times \frac{6.4}{\sqrt{n}} = \frac{29.44}{\sqrt{n}}$$

Noi vogliamo che questa quantità sia più piccola di 2.10 e otteniamo $\sqrt{n} \geq 29.44/2.10 \simeq 14.02$ da cui $n \geq 197$: dobbiamo controllare altre 148 bottigliette. ▮

Esercizio 3 Nell'ambito di uno studio statistico su *inaffidabilità* e *prestazioni* di un sistema software aperto di acquisti telematici abbiamo analizzato le richieste giornaliere arrivate nel corso del 2009 (365 giorni). (La *inaffidabilità* è la probabilità di fallimento di una richiesta e la *prestazione* la distribuzione dei tempi di risposta). Abbiamo così registrato 1) la percentuale giornaliera X di richieste fallite e 2) la media giornaliera Y dei tempi di risposta (espressi in minuti primi) delle richieste soddisfatte. I dati sono sintetizzati nella seguente tabella:

$X \setminus Y$	$(0, 5.0)$	$[5.0, 12.0)$	$[12.0, \infty)$	
$(0, 0.02]$	111	90	69	
$(0.02, 0.05]$	8	5	6	
$(0.05, 0.13]$	12	6	7	
$(0.13, 1]$	28	13	10	

1. Sulla base di questi dati, pensate che le caratteristiche di inaffidabilità e prestazioni del sistema software analizzato siano indipendenti? Rispondete impostando un opportuno test di ipotesi.

2. Verificate al 2% l'ipotesi che almeno il 50% delle richieste giornaliere siano soddisfatte (mediamente) in meno di 5 minuti contro l'alternativa che le richieste giornaliere soddisfatte in meno di 5 minuti siano meno del 50%.

3. Astraendo, la percentuale giornaliera X di richieste fallite può essere pensata come una variabile aleatoria continua. Verificate con un opportuno test se il modello beta dato da

$$f(x; \theta) = \frac{1}{\theta} x^{\frac{1}{\theta}-1} \mathbf{1}_{(0,1)}(x), \quad \theta > 0$$

si adatti ai dati forniti.

Soluzione Completiamo la tabella delle numerosità, calcolando le numerosità marginali di X e Y:

$X \setminus Y$	$(0, 5.0)$	$[5.0, 12.0)$	$[12.0, \infty)$	$N_{i\cdot}$
$(0, 0.02]$	111	90	69	270
$(0.02, 0.05]$	8	5	6	19
$(0.05, 0.13]$	12	6	7	25
$(0.13, 1]$	28	13	10	51
$N_{\cdot j}$	159	114	92	$n = 365$

1. Impostiamo un test χ^2 di indipendenza per verificare H_0: *"X e Y sono indipendenti"* contro H_1: *"X e Y non sono indipendenti"*. La statistica test è

$$Q_{\text{ind}} = \sum_{i=1}^{4} \sum_{j=1}^{3} \frac{\left(N_{ij} - \frac{N_{i\cdot} N_{\cdot j}}{365}\right)^2}{\frac{N_{i\cdot} N_{\cdot j}}{365}} = 365 \sum_{i=1}^{4} \sum_{j=1}^{3} \frac{{N_{ij}}^2}{N_{i\cdot} N_{\cdot j}} - 365$$

e nel nostro caso troviamo il valore $q_{\text{ind}} \simeq 4.517$. Dato che la statistica test sotto l'ipotesi nulla ha distribuzione limite chiquadro con $(4-1)(3-1) = 6$ gradi di libertà, la cui f.d.r. indichiamo con F_6, allora per il p-value, usando le tabelle, abbiamo $F_6(4.074) = 0.333$ e $F_6(5.348) = 0.500$. Il p-value è $1 - F_6(q_{\text{ind}}) \in (0.500, 0.667)$. Con un p-value così alto non possiamo rifiutare H_0 a nessun livello di significatività ragionevole e accettiamo l'indipendenza di inaffidabilità e prestazioni.

2. Chiamiamo p la "*percentuale di richieste giornaliere soddisfatte (mediamente) in meno di* 5 *minuti*", cioè $p = P(Y < 5)$. La stima di massima verosimiglianza o del metodo dei momenti di p è $\hat{p} = 159/365 \simeq 0.4356$. Il numero delle prove è grande e usiamo un test asintotico: a livello 2% rifiutiamo H_0 se $\frac{\hat{p}-0.5}{\sqrt{0.5\times0.5/365}} \leq z_{.02} = -z_{.98} \simeq -2.0537$. Poiché $\frac{\hat{p}-0.5}{\sqrt{0.5\times0.5/365}} \simeq -2.46 < -2.0537$ al livello 2% rifiutiamo H_0. Il p-value approssimato è $\Phi(-2.46) = 1 - \Phi(2.46) \simeq 0.69\%$.

3. Conduciamo un test χ^2 di buon adattamento per verificare H_0: "$X \sim f(x, \theta)$, *per qualche* $\theta > 0$" contro H_1: "$X \not\sim f(x,\theta)$". Innanzitutto stimiamo θ con il metodo dei momenti, usando i dati raggruppati su X:

$$M_c = (0.01 \times 270 + 0.035 \times 19 + 0.09 \times 25 + 0.565 \times 51)/365 = 34.43/365 \simeq 0.09433$$

e

$$\mathrm{E}_\theta(X) = \int_0^1 x \times \frac{1}{\theta} x^{\frac{1}{\theta}-1}\, dx = \frac{1}{1+\theta},$$

da cui otteniamo $\hat\theta = 1/M_c - 1 = 365/34.43 - 1 \simeq 9.60122$. Inoltre abbiamo:

$$\hat p_{0i} = \int_a^b \frac{1}{\hat\theta} x^{\frac{1}{\hat\theta}-1} = b^{1/\hat\theta} - a^{1/\hat\theta}, \qquad\qquad 0 < a < b < 1.$$

La statistica test è

$$Q_{\text{goodness}} = \sum_{i=1}^4 \frac{(N_{i\cdot} - 365\hat p_{0i})^2}{365\hat p_{0i}} = \sum_{i=1}^4 \frac{N_{i\cdot}^2}{365\hat p_{0i}} - 365$$

che sotto l'ipotesi nulla ha distribuzione limite chiquadro con $4-1-1=2$ gradi di libertà, cioè $\mathcal{E}(2)$; nel nostro caso troviamo il valore $q_{\text{goodness}} \simeq 9.649$ e per il p-value abbiamo p-value$\simeq \mathrm{e}^{-9.649/2} \simeq 0.008 = 0.8\%$: rifiutiamo H_0 per ogni $\alpha \geq 0.08\%$. Quindi c'è una forte evidenza empirica contro H_0 e concludiamo che il modello beta specificato non si adatta ai dati. ∎