STATISTICA per ING INF TEL [2L] A. Pievatolo, scaglione M-Z 29.06.10

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

Esercizio 1 Dobbiamo fare inferenza sui tempi di guasto di un sistema formato da 4 componenti connessi in serie, cosicché il sistema funziona se tutti i componenti funzionano. Sappiamo che i componenti funzionano in modo indipendente uno dall'altro, che sono tutti dello stesso tipo A e che i loro tempi di guasto $Y_1, \ldots, Y_4$ (espressi in ore) hanno densità esponenziale di parametro $\theta > 0$, cioè $f(y;\theta) = (1/\theta)\mathrm{e}^{-y/\theta}\mathbf{1}_{(0,\infty)}(y)$ con θ incognito.

Abbiamo così acquistato 280 componenti di tipo A e abbiamo costruito 70 sistemi in serie, tenendo ciascuno attivo fino alla rottura. Per il campione casuale $X_1, \ldots, X_{70}$ delle durate dei 70 sistemi abbiamo ottenuto $\sum_{j=1}^{70} X_j = 693.0$.

1. Verificate che la durata di un intero sistema ha densità esponenziale di parametro $\theta/4$, $\theta > 0$.
2. Determinate uno stimatore $\hat{\theta}_{ML}$ del parametro θ e $\widehat{\kappa}$ della probabilità κ che un sistema funzioni al più 12 ore, usando il metodo di massima verosimiglianza.
3. Verificate che la varianza di $\hat{\theta}_{ML}$ raggiunge il confine di Frechét-Cramer-Rao per la varianza di uno stimatore (non distorto) di θ ma che uno stimatore efficiente per κ non esiste (Giustificate rigorosamente la risposta).
4. Costruite un intervallo di confidenza bilatero di livello 90% per θ.
5. Verificate l'ipotesi nulla $H_0 : \kappa = 0.75$ contro l'alternativa $H_1 : \kappa \neq 0.75$, a una significatività $\alpha = 2.5\%$.

Soluzione

1. Il tempo di guasto X di un sistema ingegneristico ottenuto collegando in serie 4 componenti con tempi di guasto $Y_1, \ldots, Y_4$ i.i.d. $\sim Exp(\theta)$ è $X = \min\{Y_1, \ldots, Y_4\}$ e la sua f.d.r. F_X si ottiene nel seguente modo:

$$1 - F_X(x;\theta) = P(X > x;\theta) = P(\min\{Y_1, \ldots, Y_4\} > x;\theta) = \prod_{j=1}^{4} P(Y_j > x;\theta) = (1 - F_{Y_1}(x;\theta))^4 =$$
$$= [1 - (1 - \mathrm{e}^{-x/\theta})]^4 = \mathrm{e}^{-4x/\theta}$$

da cui deduciamo che la densità di X è $f(x;\theta) = \dfrac{4}{\theta}\mathrm{e}^{-4x/\theta}\mathbf{1}_{(0,\infty)}(x)$, cioè esponenziale di media $\theta/4$.

2-3. La funzione di verosimiglianza del campione casuale $X_1, \ldots, X_{70}$ è data da

$$L_\theta(x_1, \ldots, x_{70}) = \left(\frac{4}{\theta}\right)^n exp\left(-\frac{n4\overline{x}}{\theta}\right)$$

e quindi:

$$\frac{\partial \ln L_\theta(x_1, \ldots, x_{70})}{\partial\theta} = \frac{n}{\theta^2}(4\overline{x} - \theta) \tag{1}$$

da cui deduciamo che *a)* $\hat{\theta}_{ML} = 4\overline{X}$ e, per la diseguaglianza di FCR, *b)* $4\overline{X}$ è stimatore efficiente di θ (effettivamente $\mathrm{E}(4\overline{X}) = 4 \times (\theta/4) = \theta$ e abbiamo anche la non distorsione).

Per quanto riguarda la caratteristica κ definita come la probabilità che un sistema funzioni al più 12 ore, abbiamo che $\kappa = P(X \leq 12;\theta) = 1 - \mathrm{e}^{-4\times 12/\theta}$ e quindi $\widehat{\kappa}_{ML} = 1 - \mathrm{e}^{-4\times 12/\hat{\theta}_{ML}} = 1 - \mathrm{e}^{-12/\overline{x}}$.

Per la (1) non possiamo mai avere $\frac{\partial \ln L_\theta(x_1,\ldots,x_{70})}{\partial\theta} = a(n,\theta)(\widehat{\kappa}_{ML} - \kappa)$ per nessuna scelta della funzione $a(n,\theta)$; inoltre se uno stimatore efficiente di κ esiste allora necessariamente è ML. Considerato tutto ciò, segue che non solo $\widehat{\kappa}_{ML}$ non è stimatore efficiente di κ, ma anche che nessun possibile stimatore di κ è efficiente. Infine, sul nostro campione abbiamo: $\overline{x} = 9.9, \hat{\theta}_{ML} = 39.6$ e $\widehat{\kappa} = 1 - \mathrm{e}^{-1.21} \simeq 0.7018$.

4. Segue dalle proprietà della famiglia di distribuzione gamma che $\hat{\theta}_{ML} \sim \Gamma(70, \theta/70)$ cosicché $8\sum_{j=1}^{70} X_j/\theta \sim \chi^2_{140}$ da cui abbiamo:

$$P\left(\chi^2_{140}(5\%) < \frac{8\sum_{j=1}^{70} X_j}{\theta} < \chi^2_{140}(95\%)\right) = 90\%$$

e

$$P\left(\frac{8\sum_{j=1}^{70} X_j}{\chi^2_{140}(95\%)} < \theta < \frac{8\sum_{j=1}^{70} X_j}{\chi^2_{140}(5\%)}\right) = 90\%$$

Poiché i gradi di libertà sono numerosi:

$$\chi^2_{140}(95\%) \simeq \sqrt{280} \times 1.645 + 140 \simeq 167.5261$$

e

$$\chi^2_{140}(5\%) \simeq \sqrt{280} \times (-1.645) + 140 \simeq 112.4739.$$

Infine l'IC bilatero cercato per θ è $(33.0934, 49.2914)$.

5. Il problema di verifica dell'ipotesi $H_0 : \kappa = 0.75$ contro l'alternativa $H_1 : \kappa \neq 0.75$ è equivalente al problema di ipotesi su θ: $H_0 : \theta = -48/\log(1 - 0.75)$ contro l'alternativa $H_1 : \theta \neq -48/\log(1 - 0.75)$. Il valore $-48/\log(1 - 0.75)$ cade nell'IC precedentemente identificato ($-48/\log(1 - 0.75) \simeq 34.6247$) e per la dualità fra IC e VI accettiamo H_0 non solo a livello 10% ma anche per ogni $\alpha \leq 10\%$ e quindi anche al livello $\alpha = 2.5\%$ richiesto. ∎

Esercizio 2 [1] È stato condotto uno studio su come le abitudini alimentari delle donne si modifichino tra l'inverno e l'estate. Si è tenuto sotto osservazione un campione aleatorio di 12 donne durante i mesi di gennaio e luglio 2009, misurando fra le altre cose quale percentuale delle calorie da loro assunte provenisse dai grassi. I risultati ottenuti sono i seguenti.

Gennaio	30.5	28.4	40.2	37.6	36.5	38.8	34.7	29.5	29.7	37.2	41.5	37.0
Luglio	32.2	27.4	28.6	32.4	40.5	26.2	29.4	25.8	36.6	30.3	28.5	32.0

Assumiamo che il valori accoppiati formino un campione casuale da una distribuzione gaussiana bivariata.

1. Impostate un opportuno test per verificare, al livello del 5%, se la percentuale media di calorie ricavate dai grassi cambi nei mesi estivi rispetto a quelli invernali. L'errore di primo tipo è quello di ritenere che la percentuale media di calorie ricavate dai grassi sia strettamente minore nel mese di luglio rispetto a quella di gennaio, quando in realtà è vero il contrario. Abbiate cura di specificare *a*) le ipotesi nulla e alternativa, *b*) la regione critica e *c*) la decisione cui arrivate con la vostra procedura di verifica, calcolando anche il p-value del test.

2. Ricavate la stima di massima verosimiglianza della probabilità che, per una donna scelta a caso, la percentuale di calorie ricavate dai grassi nel mese di luglio sia minore della percentuale di calorie ricavate dai grassi nel mese di gennaio.

Soluzione Indichiamo con L la percentuale di calorie ricavate dai grassi nel mese di luglio e con G quella di gennaio, con μ_L, σ_L^2 media e varianza di L, con μ_G e μ_D rispettivamente le medie di L, G, con D la differenza $D = L - G$ e con μ_D e σ_D^2 media e varianza di D. Infine, siano $(L_1, G_1), \ldots, (L_{12}, G_{12})$ il campione casuale di dati accoppiati estratti dalla popolazione (L, G) e $D_1, \ldots, D_{12}$ quello delle differenze D.

1. Sotto ipotesi di normalità delle differenze: $D_1, \ldots, D_{12}$ i.i.d. $\sim N(\mu_D, \sigma_D^2)$, impostiamo il seguente t-test di confronto fra medie per dati gaussiani accoppiati: $H_0 : \mu_L \geq \mu_G$ versus $H_1 : \mu_L < \mu_G$ o, equivalentemente, $H_0 : \mu_D \geq 0$ versus $H_1 : \mu_D < 0$. La statistica test è $\sqrt{12}\overline{D}/\sqrt{S_D^2}$ che vale -2.338. Infatti, il campione delle differenze è:

$$(1.7, -1.0, -11.6, -5.2, 4.0, -12.6, -5.3, -3.7, 6.9, -6.9, -13.0, -5.0),$$

con media campionaria $\overline{D} = -4.308333$, varianza campionaria $S_D^2 \simeq 40.77356$ e $\sqrt{S_D^2} \simeq 6.385418$. Inoltre il p-value $\overline{\alpha}$ è

$$\overline{\alpha} = P_{\{\mu_D = 0\}}\left(\sqrt{12}\frac{\overline{D}}{\sqrt{S_D^2}} \leq -2.338\right) = F_{11}(-2.338) = 1 - F_{11}(2.338) \in (1\%, 2.5\%)$$

(in questo punto F_{11} rappresenta la f.d.r. t di student con 11 gradi di libertà). Segue che a livello 5% rifiutiamo H_0. Osservate che comunque non c'è una forte evidenza empirica contro H_0; per esempio, a livello 1% non la rifiutiamo.

2. Segue dall'ipotesi di dati gaussiani bivariati che la differenza D è gaussiana e la probabilità da stimare è:

$$P(L < G) = P(D < 0) = \Phi\left(\frac{0 - \mu_D}{\sqrt{\mathrm{Var}(D)}}\right).$$

Poiché per un campione gaussiano lo stimatore ML della media è la media campionaria e quello della varianza (quando la media è incognita) è $(n-1)S^2/n \simeq 37.37576$, allora lo stimatore ML di $P(D < 0)$ è dato da $\Phi\left(\frac{4.308333}{\sqrt{37.37576}}\right) \simeq \Phi(0.7) \simeq 0.74$. ∎

[1] Dati tratti da Ross S.M., Probabilità e statistica per l'ingegneria e le scienze, Apogeo 2008.

Esercizio 3 [2] I valori che seguono rappresentano le lunghezze in millimetri di un campione di 10 granelli presi da una grossa pila di polvere metallica:

$$2.2 \quad 3.4 \quad 1.6 \quad 0.8 \quad 2.7 \quad 3.3 \quad 1.6 \quad 2.8 \quad 2.5 \quad 1.9$$

1. Stabilite con un opportuno test se una densità lognormale si adatti ai dati forniti.
2. Stimate la percentuale di granelli nella pila la cui lunghezza è compresa fra 1.5 e 2.5 mm.

(*Vi ricordiamo che una variabile aleatoria X è detta* lognormale *di parametri μ, σ se il suo logaritmo naturale $\ln X$ è variabile aleatoria gaussiana di media μ e varianza σ^2.*)

Soluzione

1. Considerato che X è *lognormale* di parametri μ, σ se $Y = \ln X \sim \mathcal{N}(\mu, \sigma^2)$, usiamo un test di Lilliefors per la normalità dei dati logaritmici $Y_1, \ldots, Y_{10}$, con $Y_j = \ln X_j$, per $j = 1, \ldots, 10$. Infatti i dati sono continui e in numero esiguo e i parametri della distribuzione $\mathcal{N}(\mu, \sigma^2)$ non sono assegnati. I dati in scala logaritmica e ordinati dal più piccolo al più grande sono:

   $$y_i: \quad -0.2231 \quad 0.4700 \quad 0.4700 \quad 0.6419 \quad 0.7885 \quad 0.9163 \quad 0.9933 \quad 1.0296 \quad 1.1939 \quad 1.2238$$

   La media campionaria delle y_i vale $\overline{y} \simeq 0.7504$ e la deviazione standard campionaria $\sqrt{s_Y^2} \simeq 0.4351$ da cui otteniamo per $z_i := (y_i - \overline{y})/\sqrt{s_Y^2}$ i seguenti valori (ordinati e distinti) e la corrispondente funzione di ripartizione empirica (indicata con $\widehat{F}_{10}$):

z_i	-2.24	-0.64	-0.25	0.09	0.38	0.56	0.64	1.02	1.09
$\widehat{F}_{10}(z_i)$	0.1	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\Phi(z_i)$	0.0125	0.2611	0.4013	0.5359	0.6480	0.7123	0.7389	0.8461	0.8621
$\lvert\widehat{F}_{10}(z_i) - \Phi(z_i)\rvert$	0.0875	0.0389	0.0013	0.0359	0.0480	0.0123	0.0611	0.0539	0.1379
$\lvert\widehat{F}_{10}(z_{i-1}) - \Phi(z_i)\rvert$	0.0125	0.1611	0.1013	0.1359	0.1480	0.1123	0.0389	0.0461	0.0379

   Deduciamo dalla precedente tabella che la statistica test $D_{10} = \sup_{z \in \mathbb{R}} |\widehat{F}_{10}(z) - \Phi(z)|$ ha valore approssimativamente pari a 0.1611. Dalle tavole di Lilliefors abbiamo che il quantile di ordine $1-0.2$ della statistica di Lilliefors (sotto l'ipotesi H_0 che i dati in scala logaritmica siano gaussiani) è $q(1-0.2) = 0.2171$. Poiché $0.1611 < 0.2171$ allora accettiamo l'ipotesi di dati Y_i normali per ogni $\alpha \leq 20\%$: altrimenti detto, non c'è alcuna evidenza empirica contro la log-normalità dei dati X_i.

   (Usando il pacchetto R, "con meno approssimazioni nei conti" otteniamo $D_{10} = 0.1596$ con p-value= 0.6668

2. Avendo accettato l'ipotesi di log-normalità dei dati X_i la percentuale di granelli nella pila la cui lunghezza è compresa fra 1.5 e 2.5 mm è:

$$P(1.5 < X < 2.5) = P(\ln 1.5 < Y < \ln 2.5) = \Phi\left(\frac{\ln 1.5 - \mu}{\sigma} < \frac{Y - \mu}{\sigma} < \frac{\ln 2.5 - \mu}{\sigma}\right) = \\ = \Phi(\frac{\ln 2.5 - \mu}{\sigma}) - \Phi(\frac{\ln 1.5 - \mu}{\sigma})$$

   e una sua stima è data da

$$\Phi(\frac{\ln 2.5 - \overline{y}}{s}) - \Phi(\frac{\ln 1.5 - \overline{y}}{s}) \simeq \Phi(0.38) - \Phi(-0.79) \simeq 0.6485 - 0.214 = 43.45\% \quad \blacksquare$$

[2] Dati tratti da Ross S.M., Probabilità e statistica per l'ingegneria e le scienze, Apogeo 2008.