STATISTICA per ING INF TEL [2L] I. Epifani, A. Pievatolo 08.07.2011

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

Esercizio 1 Aldo si sta allenando al gioco delle freccette. Su 36 partite: 16 volte fa centro al 1° tentativo, 8 volte al 2°, 4 volte al 3°, 3 volte al 4°, 2 volte al 5° e 3 volte al 6°. Abbiamo quindi i dati di un campione casuale di 36 osservazioni provenienti da una popolazione geometrica.

1. Stimate con il metodo di massima verosimiglianza 1) la probabilità p che Aldo faccia centro in un singolo tentativo e 2) mediamente quanti tentativi sono necessari ad Aldo per fare centro. Determinatene le stime numeriche sulla base dei dati forniti.
2. Determinate la distribuzione asintotica dello stimatore ML del numero medio di tentativi nececessari ad Aldo per fare centro, avendo cura di specificarne tutti i parametri.
3. Verificate al 5% l'ipotesi nulla che la probabilità che Aldo faccia centro in un singolo tentativo sia $p = 1/2$, contro l'alternativa $p < 1/2$, usando un test per grandi campioni.
4. Determinate la potenza del test asintotico del punto 3, se $p = 1/4$.

Soluzione Abbiamo il campione $X_1, \ldots, X_n$ i.i.d. $\sim f(x, p)$ con

$$f(x,p) = p(1-p)^{x-1}\mathbf{1}_{\{1,2,\ldots\}}(x) \quad 0 < p < 1$$

e le caratteristiche media e varianza di $f(x,p)$ sono $\mu = \mu(p) = p^{-1}$ e $\sigma^2 = \sigma^2(p) = (1-p)/p^2 = p^{-2} - p^{-1}$.

1. La funzione di verosimiglianza è

$$L_p(x_1, \ldots, x_n) = \prod_{j=1}^{n} f(x_j, p) = \prod_{j=1}^{n} p(1-p)^{x_j - 1} = p^n (1-p)^{\sum_{j=1}^{n} x_j - n}/,$$

cosicché lo stimatore ML di p NON esiste se $\sum_j X_j = n = 36$ (in tal caso infatti $L_p(x_1, \ldots, x_n)$ raggiunge il sup in $p = 1$ che non appartiente all'intervallo $(0,1)$). Noi invece abbiamo $\sum_j X_j = 84 > n$, e in questo caso possiamo procedere a massimizzare $\log L_p$:

$$\frac{\partial}{\partial p} \log L_p(x_1, \ldots, x_n) = \frac{\partial}{\partial p}\left(n \log p + \left[\sum_{j=1}^{n} x_j - n\right]\log(1-p)\right) = -\frac{n}{1-p}\left(\bar{x} - \frac{1}{p}\right) \quad ;$$

$\bar{x}$ è la media campionaria e

$$-\frac{n}{1-p}\left(\bar{x} - \frac{1}{p}\right) \geq 0 \quad \text{se e solo se} \quad p \leq \frac{1}{\bar{x}}$$

Segue che $\widehat{p}_{ML} = 1/\overline{X}$ e $\widehat{\mu}_{ML}$ è la media campionaria $\overline{X}$. Le corrispondenti stime numeriche sono: $\overline{x} = 7/3 \simeq 2.33$ e $\widehat{p}_{ML} = 1/\overline{x} = 3/7 \simeq 0.4286$.

2. Abbiamo dal Teorema Centrale del Limite che asintoticamente: $\widehat{\mu}_{ML} = \overline{X} \sim \mathcal{N}(p^{-1}, (1-p) \times p^{-2}/n)$;

3. Innanzitutto osserviamo che $p = 1/2$ se e solo se $\mu = 2$ e $p < 1/2$ se e solo se $\mu > 2$; inoltre, abbiamo tante osservazioni (36). Per questo, impostiamo lo z-test asintotico sulla media costituito dalla tripletta:

$$\left(H_0 : \mu = 2 \text{ contro } H_1 : \mu > 2;\ X_1, \ldots, X_{36};\ \overline{X} \geq 2 + \frac{\sqrt{(1-1/2)/(1/2)^2}}{\sqrt{36}} z_{95\%}\right)$$

cosicché rifiutiamo H_0 se $\overline{X} \geq 2.3877$. Poiché $\overline{x} \simeq 2.33$ non possiamo rifiutare H_0.

4. $p = 1/4$ se e solo se $\mu = 4$ e $\pi(1/4) = P_{1/4}(\overline{X} \geq 2.3877) \simeq 1 - \Phi\left(\frac{\sqrt{36}(2.3877 - 4)}{\sqrt{(1-1/4)/(1/4)^2}}\right) = 1 - \Phi\left(-1.6123 \times \frac{6}{\sqrt{12}}\right) \simeq \Phi(2.79) \simeq 0.9974.$ ∎

Esercizio 2 Abbiamo due campioni indipendenti $X_1, \ldots, X_m$ i.i.d. $\sim \mathcal{E}(\theta)$ e $Y_1, \ldots, Y_n$ i.i.d. $\sim \mathcal{E}(\lambda)$ che rappresentano i tempi di guasto di m lampadine di tipo X e di n di tipo Y. Il nostro obiettivo è confrontare i tempi di guasto medi. A tal fine:

1. determinate la distribuzione di $\dfrac{2m\overline{X}}{\theta}$, $\dfrac{2n\overline{Y}}{\lambda}$ e $\dfrac{\overline{X}/\theta}{\overline{Y}/\lambda}$, dove $\overline{X}$ rappresenta il tempo medio (campionario) di guasto delle m lampadine di tipo X e $\overline{Y}$ quello delle n di tipo Y;
2. proponete una regione critica di livello $\alpha = 5\%$ per verificare $H_0 : \theta = \lambda$ contro $H_1 : \theta \neq \lambda$, basata sulla statistica test $\overline{X}/\overline{Y}$ (e ovviamente sul risultato del punto 1);
3. se $m = 6, n = 5$, $\overline{X} = 24.1$ e $\overline{Y} = 21.5$, accettate o rifiutate H_0 a livello 5%?
4. Calcolate $\mathrm{E}\left(\dfrac{\overline{X}}{\overline{Y}}\right)$ (se $n > 1$). La statistica $\overline{X}/\overline{Y}$ è stimatore asintoticamente non distorto di θ/λ?
5. Calcolate la varianza di $\overline{X}/\overline{Y}$ (se $n > 2$). La statistica $\overline{X}/\overline{Y}$ è stimatore consistente di θ/λ?

(*Hint per i punti 4 e 5: a) ricordate che $\overline{X}$ e $\overline{Y}$ sono indipendenti; b) vale che:* $\displaystyle\int_0^\infty \frac{1}{y^r} \times \frac{y^{n-1} \mathrm{e}^{-y/b}(1/b^n)}{\Gamma(n)}\, dy = \frac{1}{b^r(n-1)(n-2)\cdots(n-r)}$, $\forall n > r$ *e* $r = 1, 2, \ldots$)

Soluzione

1. $(2m\overline{X})/\theta \sim \chi^2_{2m}$, $(2n\overline{Y})/\lambda \sim \chi^2_{2n}$. Infine, per l'indipendenza dei due campioni: $\dfrac{\overline{X}/\theta}{\overline{Y}/\lambda} \sim F_{2m,2n}$.
2. Rifiutiamo $H_0 : \theta = \lambda$ a favore di $H_1 : \theta \neq \lambda$ a livello 5% se

$$\frac{\overline{X}}{\overline{Y}} \notin \left(\frac{1}{F_{2n,2m}(1 - 0.05/2)}, F_{2m,2n}(1 - 0.05/2) \right)$$

Infatti sotto H_0:

$$\frac{\overline{X}}{\overline{Y}} = \frac{\overline{X}/\theta}{\overline{Y}/\lambda} \sim F_{2n,2m} \; .$$

3. $24.1/21.5 \simeq 1.121$, $(F_{10,12}(0.975))^{-1} = (3.37)^{-1} \simeq 0.297$ e $F_{12,10}(0.975) = 3.62$ cosicché non possiamo rifiutare l'ipotesi nulla di uguaglianza delle medie dei tempi di vita dei due tipi di lampadine.

4/5. Deriva dall'indipendenza dei due campioni che

$$\mathrm{E}\left(\left(\frac{\overline{X}}{\overline{Y}}\right)^r\right) = \mathrm{E}\left(\overline{X}^r\right) \times \mathrm{E}\left(\frac{1}{\overline{Y}^r}\right) \; ;$$

inoltre:

$$\mathrm{E}\left(\overline{X}\right) = \mathrm{E}(X_1) = \theta \; ,$$

$$\mathrm{E}\left(\overline{X}^2\right) = \mathrm{Var}(\overline{X}) + \theta^2 = \theta^2\left(\frac{1}{m} + 1\right) \; ,$$

$$\mathrm{E}\left(\frac{1}{\overline{Y}^r}\right) = \frac{n^r}{\lambda^r(n-1)\cdots(n-r)} \; , \quad r = 1, 2 \text{ [infatti } \overline{Y} \sim \Gamma(n, \lambda/n) \text{ e l'}\textit{Hint} \text{ vale con } b = \lambda/n]$$

da cui otteniamo

$$\mathrm{E}\left(\frac{\overline{X}}{\overline{Y}}\right) = \frac{\theta n}{\lambda(n-1)} \to \frac{\theta}{\lambda}, \quad \text{per } n \to \infty \qquad [\text{quindi } \overline{X}/\overline{Y} \text{ è stimatore asintoticamente non distorto di } \theta/\lambda]$$

$$\mathrm{E}\left(\frac{\overline{X}^2}{\overline{Y}^2}\right) = \frac{\theta^2(1/m+1)n^2}{\lambda^2(n-1)(n-2)}$$

$$\mathrm{Var}\left(\frac{\overline{X}}{\overline{Y}}\right) = \frac{\theta^2}{\lambda^2}\left[\frac{(1/m+1)n^2}{(n-1)(n-2)} - \frac{n^2}{(n-1)^2}\right] \to 0, \quad \text{per } m, n \to \infty \qquad [\text{quindi } \overline{X}/\overline{Y} \text{ è stimatore consistente di } \theta/\lambda] \quad \blacksquare$$

Esercizio 3 [1]I dati che seguono costituiscono i tempi di vita (in centinaia di ore) di due tipi di valvole termoioniche:

$$\begin{array}{ll} \text{Tipo } X: & 32, 84, 37, 87 \\ \text{Tipo } Y: & 39, 55, 90 \end{array}$$

1. Verificate con un test non parametrico se i tempi di vita dei due tipi di valvole termoioniche siano regolati da uno stesso modello probabilistico.
2. Stabilite con un opportuno test la validità di una distribuzione lognormale per modellare tempi di vita di valvole termoioniche.
   (*Una variabile aleatoria* W *è detta* lognormale *di parametri* μ, σ *se il suo logaritmo naturale* $\ln W$ *è variabile aleatoria gaussiana di media* μ *e varianza* σ^2)
3. Stimate la probabilità che una valvola termoionica scelta a caso duri più di 5500 ore.

Soluzione

1. Usiamo un test non parametrico di omogeneità per campioni indipendenti, qualunque e senza ripetizione. Noi conosciamo il test di Wilcoxon-Mann-Whitney. Per verificare H_0: X, Y *sono regolati dallo stesso modello probabilistico* contro l'alternativa H_1 che non lo siano, la statistica test è la somma dei ranghi $R_1, \dots R_4$ di X: $T_X = 1 + 2 + 5 + 6 = 14$ con p-value $= 0.6286$ (esatto): è alto, quindi nessuna evidenza empirica contro H_0. Utilizzando le tavole, troviamo che il percentile associato alla massima probabilità rappresentata è $w_{0.1} = 12$, dunque dalle sole tavole possiamo concludere che il p-value approssimato del test è superiore a $2 \times 0.1 = 20\%$.
2. Avendo accettato l'ipotesi di omogeneità dei tempi di vita dei due tipi di valvole, consideriamo i 7 dati come un unico campione casuale estratto dalla stessa popolazione. Inoltre, considerato che W è *lognormale* di parametri μ, σ se $U = \ln W \sim \mathcal{N}(\mu, \sigma^2)$, che i dati sono continui e in numero esiguo e infine i parametri della distribuzione $\mathcal{N}(\mu, \sigma^2)$ non sono assegnati, allora usiamo un test di Lilliefors per la normalità dei dati logaritmici $U_1, \dots, U_7$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$, con $U_j = \ln X_j$, per $j = 1, \dots, 4$ e $U_j = \ln Y_j$, per $j = 5, \dots, 7$. I dati in scala logaritmica e ordinati dal più piccolo al più grande sono:

$$u_i: \quad 3.4657 \quad 3.6109 \quad 3.6636 \quad 4.0073 \quad 4.4308 \quad 4.4659 \quad 4.4998$$

La media campionaria delle u_i vale $\overline{u} \simeq 4.0206$ e la deviazione standard campionaria $\sqrt{s_U^2} \simeq 0.4471$ da cui otteniamo per $z_i := (u_i - \overline{u})/\sqrt{s_u^2}$ i seguenti valori (ordinati e distinti) e la corrispondente funzione di ripartizione empirica (indicata con $\widehat{F}_7$):

z_i	-1.24	-0.92	-0.80	-0.03	0.92	1.00	1.07
$\widehat{F}_7(z_i)$	1/7	2/7	3/7	4/7	5/7	6/7	1.0
$\Phi(z_i)$	0.1075	0.1788	0.2119	0.4880	0.8212	0.8413	0.8577
$\lvert\widehat{F}_7(z_i) - \Phi(z_i)\rvert$	0.0354	0.1069	0.2167	0.0834	0.1069	0.0158	0.1423
$\lvert\widehat{F}_7(z_{i-1}) - \Phi(z_i)\rvert$	0.1075	0.0359	0.0738	0.0594	0.2498	0.1270	0.0006

Deduciamo dalla precedente tabella che la statistica test $D_7 = \sup_{z \in \mathbb{R}} |\widehat{F}_7(z) - \Phi(z)|$ ha valore approssimativamente pari a $\simeq 0.2498$. 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 $q(1 - 0.2) = 0.2521$. Poiché $0.2498 < 0.2521$ allora accettiamo l'ipotesi di dati U_i normali per ogni $\alpha \leq 20\%$: altrimenti detto, non ricaviamo dai dati U_i alcuna prova empirica contro un modello lognormale per spiegare tempi di vita di valvole termoioniche.
3. Poiché non abbiamo rifiutato l'ipotesi di dati (tutti) lognormali, allora una stima della probabilità che una valvola termoionica duri più di 5500 ore è data da:

$$P(U > 55) = P(\ln U > \ln 55) = 1 - \Phi\left(\frac{\ln 55 - \overline{u}}{\sqrt{s_u^2}}\right) \simeq 1 - \Phi(-0.03) = \Phi(0.03) \simeq 0.512. \quad \blacksquare$$

[1]Dati estratti da Sheldon Ross, Probabilità e statistica, Ed Apogeo, 2004