STATISTICA per ING INF TEL [2L] I. Epifani, A.M. Pievatolo 01.03.10

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

Esercizio 1 Sia $X_1, \ldots, X_n$ un campione gaussiano di media incognita μ e varianza nota e pari a 900.

1. Costruite un test su μ tale che sia al più pari a 6% la probabilità di commettere l'errore di prima specie di ritenere che μ sia maggiore di 5500 quando effettivamente μ al più vale 5500.
2. Se abbiamo 100 osservazioni e la media campionaria $\overline{X}_{100}$ vale 5506.0, qual è il p-value del test costruito al punto 1.? Che decisione prendete?
3. Calcolate la probabilità di prendere la corretta decisione in base al test costruito al punto 1. se il "vero" valore di μ è 5505 e $n = 100$.
4. Potete raccogliere ulteriori osservazioni. Determinate il numero minimo di osservazioni da raccogliere affinché la potenza del test in $\mu = 5505$ aumenti almeno del 50%.

Soluzione Dobbiamo impostare un opportuno z-test unilatero sulla media μ incognita di un campione casuale $X_1, \ldots, X_n$ estratto da una popolazione di densità $\mathcal{N}(\mu, 900)$.

1. Dobbiamo impostare un test di significatività $\alpha = 6\%$ per l'ipotesi nulla $H_0 : \mu \leq 5500$ contro l'alternativa $H_1 : \mu > 5500$. Segue che rifiutiamo H_0 a favore di H_1 a livello 6% se $\sqrt{n}(\overline{X}_n - 5500)/30 \geq z_{94\%}$ o, equivalentemente, se $\overline{X}_n \geq 5500 + 1.555 \times 30/\sqrt{n}$, dal momento che $z_{94\%} \simeq 1.555$.

2. Se $n = 100$ e $\overline{X}_{100} = 5506.0$ allora il p-value del test costruito al punto 1. è dato da

$$P_{5500}\left(\overline{X}_{100} \geq 5506.0\right) = 1 - \Phi(2) \simeq 1 - 0.9778 = 2.28\%$$

Essendo $\alpha = 6\% > 2.28\%$ rifiutiamo $H_0 : \mu \leq 5500$ a favore di $H_1 : \mu > 5500$.

3. Per il problema di ipotesi $H_0 : \mu \leq 5500$ contro $H_1 : \mu > 5500$, la probabilità di prendere la corretta decisione se $\mu = 5505$ è data dalla potenza del test in $\mu = 5505$. Se $n = 100$ allora rifiutiamo $H_0 : \mu \leq 5500$ a favore di $H_1 : \mu > 5500$ se $\overline{X}_n \geq 5504.665$ e la potenza del test in $\mu = 5505$ è

$$P_{5505}\left(\overline{X}_{100} \geq 5504.665\right) = 1 - \Phi\left(\frac{5504.665 - 5505}{3}\right) = \Phi(0.111667) \simeq \Phi(0.11) = 0.5438$$

4. Aumentando del 50% la potenza del test passiamo da 0.5438 a $0.5438 \times 1.5 = 0.8197$. Stiamo quindi cercando il minimo n tale che

$$P_{5505}\left(\overline{X}_n \geq 5500 + 1.555\frac{30}{\sqrt{n}}\right) \geq 0.8157$$

cioè tale che

$$1 - \Phi\left(\frac{5500 - 5505}{30/\sqrt{n}} + 1.555\right) \geq 0.8157$$

che ha soluzione $n \geq (6(1.555 + 0.90))^2 = 14.73^2$ (abbiamo usato l'approssimazione $z_{0.8157} \simeq z_{0.8159} = 0.90$). Otteniamo così $n \geq 217$, cioè dobbiamo raccogliere ulteriori 117 osservazioni. ∎

Esercizio 2 Considerate la famiglia di densità

$$f(x;\vartheta)=\begin{cases}\frac{2x}{\vartheta} & \text{se } 0<x\le\vartheta\\ \frac{2(1-x)}{1-\vartheta} & \text{se } \vartheta<x<1\\ 0 & \text{altrove}\end{cases}$$

con $0<\vartheta<1$ parametro incognito.

1. Calcolate media e varianza di $f(x;\vartheta)$ per ogni $\vartheta\in(0,1)$.
2. Dato un campione casuale $X_1,\dots X_n$ estratto da una popolazione di densità $f(x;\vartheta)$, determinate uno stimatore T_n di ϑ con il metodo dei momenti.
3. Verificate se la successione di stimatori $(T_n)_n$ è non distorta e consistente in media quadratica. Inoltre, qual è la distribuzione asintotica della successione $(T_n)_n$? *Giustificare adeguatamente le risposte.*
4. Costruite un intervallo di confidenza asintotico bilatero di livello approssimativamente 90% per ϑ, se $n=169$ e $\sum_{j=1}^{169}x_j=68.0$.
5. Supponete ora di avere un'unica osservazione X_1. Tracciate il grafico di $\vartheta\mapsto f(x;\vartheta)$ e determinate lo stimatore di massima verosimiglianza di ϑ basato sull'unica osservazione X_1.

Soluzione

1.

$$\begin{aligned}
\mathrm{E}(X)&=\int_{-\infty}^{+\infty}xf(x;\vartheta)\,dx=\int_0^{\vartheta}x\frac{2x}{\vartheta}dx+\int_{\vartheta}^1x\frac{2(1-x)}{1-\vartheta}\,dx=\frac{\vartheta+1}{3}\\
\mathrm{E}(X^2)&=\int_{-\infty}^{+\infty}x^2f(x;\vartheta)\,dx=\int_0^{\vartheta}x^2\frac{2x}{\vartheta}dx+\int_{\vartheta}^1x^2\frac{2(1-x)}{1-\vartheta}\,dx=\frac{\vartheta^2+\vartheta+1}{6}\\
\mathrm{Var}(X)&=\mathrm{E}(X^2)-(\mathrm{E}(X))^2=\frac{\vartheta^2+\vartheta+1}{6}-\frac{(\vartheta+1)^2}{3^2}=\frac{\vartheta^2-\vartheta+1}{18}
\end{aligned}$$

2. Poiché $\mathrm{E}(X)=(\vartheta+1)/3$ lo stimatore cercato è $T_n=3\overline{X}_n-1$
3. T_n è non distorto per costruzione. È consistente in media quadratica in quanto è non distorto e

$$\mathrm{Var}(T_n)=\mathrm{Var}(3\overline{X}_n-1)=9\,\mathrm{Var}(\overline{X}_n)=\frac{9\,\mathrm{Var}(X_1)}{n}=\frac{(\vartheta^2-\vartheta+1)}{2n}\to 0,\ \text{per } n\to\infty.$$

Infine, per il Teorema centrale del limite la funzione di ripartizione asintotica di T_n è gaussiana di media $\mathrm{E}(T_n)=\vartheta$ e varianza $\mathrm{Var}(T_n)=(\vartheta^2-\vartheta+1)/(2n)$.

4. La f.d.r. asintotica di $(T_n-\vartheta)/\sqrt{(\vartheta^2-\vartheta+1)/(2n)}$ è $\mathcal{N}(0,1)$. Segue che un IC asintotico bilatero per ϑ di livello approssimativamente 90% ha estremi $T_n\pm1.645\times\sqrt{(T_n^2-T_n+1)}/\sqrt{2n}$. Con i dati forniti, $T_{169}=3\times68/169\simeq0.207$ e siamo 90%-confidenti (approssimativamente) che $0.1252<\vartheta<0.2888$.

5. La funzione $\vartheta\mapsto f(x_1;\vartheta)$ è la funzione di verosimiglianza corrispondente a un'unica osservazione x_1; tale funzione è massima per $\vartheta=x$ e quindi $\widehat{\vartheta}_1=X_1$ è lo stimatore di massima verosimiglianza di ϑ per un campione di una sola osservazione. ▮

Esercizio 3 È stato condotto uno studio su 200 pazienti affetti da retinopatia diabetica. Un occhio, casualmente scelto fra il destro e il sinistro, è trattato e l'altro è osservato senza trattamento. T_1 rappresenta il tempo che trascorre -a partire da un tempo 0- fino alla cecità dell'occhio trattato e T_2 di quello non trattato. I tempi T_1, T_2 sono entrambi espressi in anni e i dati raggruppati raccolti sono riportati nella seguente tabella:

$T_1 \ \backslash \ T_2$	$(0,6]$	$(6,7]$	$(7,\infty)$
$(0,6]$	20	20	40
$(6,7]$	20	20	10
$(7,10]$	15	10	15
$(10,\infty)$	5	10	15

1. Verificate con un opportuno test se una densità esponenziale si adatta ai dati del tempo T_1.
2. Stimate la probabilità p che la cecità sopraggiunga per l'occhio non trattato prima dei 7 anni e verificate con un test asintotico di significatività approssimativamente $\alpha = 10\%$ l'ipotesi nulla $H_0 : p = 0.55$ contro l'alternativa $H_1 : p \neq 0.55$.
3. Verificate con un opportuno test di significatività $\alpha = 5\%$ se i tempi T_1, T_2 sono indipendenti.

Soluzione Riportiamo la tabella dei dati arricchita delle numerosità marginali di T_1, T_2:

$T_1 \ \backslash \ T_2$	$(0,6]$	$(6,7]$	$(7,\infty)$	
$(0,6]$	20	20	40	$\boxed{80}$
$(6,7]$	20	20	10	$\boxed{50}$
$(7,10]$	15	10	15	$\boxed{40}$
$(10,\infty)$	5	10	15	$\boxed{30}$
	$\boxed{60}$	$\boxed{60}$	$\boxed{80}$	$\boxed{200}$

1. Impostiamo un test χ^2 di buon adattamento per H_0: "*T_1 ha densità esponenziale*" contro l'alternativa H_1 : "*T_1 non ha densità esponenziale*". Uno stimatore per θ ottenuto con il metodo dei momenti applicato ai dati raggruppati di T_1 è dato da $\hat{\theta} = (3 \times 80 + 6.5 \times 50 + 8.5 \times 40 + 10 \times 30)/200 = 1155/200 = 6.025$, dove $3, 6.5, 8.5, 10$ sono i valori centrali e $80, 50, 40, 30$ le numerosità delle rispettive classi. Sotto H_0 le probabilità stimate che T_1 appartenga ad ognuna delle 4 classi sono

$$\hat{p}_{01} = P_{H_0}(T_1 \leq 6) = 1 - \mathrm{e}^{-6/6.025} \simeq 0.6306, \ \hat{p}_{02} = P_{H_0}(6 < T_1 \leq 7) = \mathrm{e}^{-6/6.025} - \mathrm{e}^{-7/6.025} \simeq 0.0565,$$
$$\hat{p}_{03} = P_{H_0}(7 < T_1 \leq 10) = \mathrm{e}^{-7/6.025} - \mathrm{e}^{-10/6.025} \simeq 0.1227, \ \hat{p}_{04} = 1 - (0.6306 + 0.0565 + 0.1227) = 0.1902,$$

e la statistica di Pearson è

$$Q_1 = \sum_{i=1}^{4} \frac{(N_i - n\hat{p}_{0i})^2}{n\hat{p}_{0i}} = \sum_{i=1}^{4} \frac{N_i^2}{200\hat{p}_{0i}} - 200 = \frac{1}{200}\left(\frac{80^2}{0.6306} + \frac{50^2}{0.0565} + \frac{40^2}{00.1227} + \frac{30^2}{0.1902}\right) - 200 \simeq 160.843$$

(non vi è bisogno di raggruppare ulteriormente i dati in quanto $0.0565 \times 200 = 11.3 > 5$). Il p-value del test vale $1 - F_{\chi^2_{4-1-1}}(160.843) = \mathrm{e}^{-160.843/2} \simeq 0$. Quindi vi è una fortissima evidenza empirica contro l'ipotesi nulla di dati esponenziali.

2. $p = P(T_2 \leq 7)$ e $\hat{p} = (\#$ di pazienti con $T_2 \leq 7)/200 = (60 + 60)/200 = 0.6$. Avendo tante osservazioni (200), ed essendo $200 \times 0.55 = 110 > 5$, allora un test bilatero asintotico di livello $\alpha = 10\%$ per $H_0 : p = 0.55$ contro $H_1 : p \neq 0.55$ prevede di rifiutare H_0 se $\sqrt{200}|\hat{p} - 0.55|/\sqrt{0.55 \times 0.45} \geq z_{1-0.1/2} \simeq 1.645$. Nel nostro caso, $\sqrt{200}|0.6 - 0.55|/\sqrt{0.55 \times 0.45} \simeq 1.42$. Perciò accettiamo H_0.

3. Effettuiamo un test χ^2 di indipendenza fra T_1, T_2. La statistica test è

$$Q_2 = 200 \sum_{i=1}^{4} \sum_{j=1}^{3} \frac{N_{ij}^2}{N_{i.}N_{.j}} - 200 = 200\left(\frac{20^2}{80 \times 60} + \frac{20^2}{80 \times 60} + \frac{40^2}{80 \times 80} + \cdots + \frac{15^2}{30 \times 80}\right) - 200 \simeq 15.451$$

e, a livello 5% rifiutiamo l'ipotesi di indipendenza se $Q_2 > \chi^2_{(4-1)(3-1)}(95\%)$. Poiché $\chi^2_6 = 12.592$ e $15.451 > 12.592$, allora rifiutiamo l'ipotesi di indipendenza fra T_1, T_2.

∎