Politecnico di Milano
Temi d'esame di STATISTICA dell'AA 2007/2008
per allievi ING INF [2L], docente I. Epifani

1 STATISTICA per ING INF TEL [2L] I. Epifani, A. Guglielmi 01.07.08

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

Esercizio 1.1 Sia X_1 un'unica osservazione estratta da una popolazione di densità

$$f(x;\theta) = 2\theta x(1-x^2)^{\theta-1}\mathbf{1}_{(0,1)}(x)$$

con θ parametro positivo incognito.

1. Costruite il test più potente di livello $\alpha = 4\%$ per verificare $H_0 : \theta = 1$ contro $H_1 : \theta = 10$.
2. Calcolate la probabilità di errore di secondo tipo del test costruito al punto 1.
3. Qual è la probabilità di prendere la corretta decisione, se effettivamente $\theta = 10$?
4. Se abbiamo osservato $X_1 = 0.5$, quanto vale il p-value? Cosa si può concludere?

SOLUZIONE

1. Dobbiamo costruire il test dettato dal Lemma di Neyman Pearson che ha regione critica della forma $\mathcal{G} = \left\{x_1 : \frac{L_1(x_1)}{L_{10}(x_1)} \leq \delta\right\}$. Abbiamo che $L_\theta(x_1) = f(x_1;\theta)$ e

$$\frac{L_1(x_1)}{L_{10}(x_1)} = \frac{2x_1}{20x_1(1-x_1^2)^9} = \frac{1}{10(1-x_1^2)^9},$$

cosicché

$$\frac{L_1(x_1)}{L_{10}(x_1)} \leq \delta \quad \Leftrightarrow \qquad (1-x_1)^2 \geq \left(\frac{1}{10\delta}\right)^{(1/9)} \Leftrightarrow \qquad x_1 \leq k$$

con k tale che $P_1(X_1 \leq k) = \alpha$.
Ma, per $0 < k \leq 1$, $P_\theta(X_1 \leq k) = \int_0^k 2\theta x(1-x^2)^{\theta-1}\,dx = 1-(1-k^2)^\theta$, da cui ricaviamo $k = \sqrt{\alpha}$; dunque la regione di rifiuto del test più potente di livello α per H_0 contro H_1 è

$$\mathcal{G} = \left\{x_1 \leq \sqrt{\alpha}\right\}.$$

Se $\alpha = 0.04$ allora $\mathcal{G} = \left\{x_1 \leq 0.2\right\}$.

2. L'errore di seconda specie si commette accettando H_0 quando essa è in realtà falsa. Pertanto, la probabilità di errore di seconda specie è

$$\beta = \beta(10) = P_{10}(X_1 \notin \mathcal{G}) = P_{10}(X_1 > 0.2) = 1 - P_{10}(X_1 \leq 0.2) = 1 - [1-(1-0.2^2)^{10}] = 0.96^{10} \simeq 0.665.$$

3. Dobbiamo calcolare la potenza del test: $\pi = 1-\beta = P_{10}(X_1 \in \mathcal{G}) = 1 - 0.96^{10} \simeq 0.335$.
4. Dobbiamo calcolare il più piccolo valore di α per cui si rifiuta H_0 quando $X_1 = 0.5$. Rifiutiamo H_0 a livello α se $0.5 \leq \sqrt{\alpha}$, ossia per ogni $\alpha \geq 0.5^2 = 0.25$; dunque il p-value è 0.25, un valore piuttosto alto. Concludiamo che non c'è evidenza empirica contro H_0. ∎

Esercizio 1.2 Una variabile aleatoria continua X è detta *lognormale di parametri μ e σ* se il suo logaritmo naturale $\log(X)$ ha densità $\mathcal{N}(\mu, \sigma^2)$. Comunque, qualora ne dobbiate aver bisogno, trovate l'espressione di una densità $f(x; \mu, \sigma^2)$ lognormale di parametri μ e σ a pagina 1 del Formulario.

Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla popolazione lognormale di parametri μ nota e uguale a 0, e $\sigma > 0$ incognito.

1. Determinate uno stimatore $\hat{\sigma}^2$ di σ^2 usando il metodo di massima verosimiglianza.
2. Discutete tutte le proprietà dello stimatore $\hat{\sigma}^2$ che conoscete, esatte e asintotiche.
3. Determinate un intervallo di confidenza bilatero di livello 95% per σ^2 per un campione di 55 osservazioni per il quale abbiamo ottenuto $\sum_{j=1}^{55} (\log X_j)^2 = 3644.0$.
4. Verificate l'ipotesi nulla $H_0 : \sigma = 8$ contro l'alternativa $H_1 : \sigma \neq 8$, a livello $\alpha = 3\%$.

SOLUZIONE Poichè X è lognormale di parametri $\mu = 0$ e $\sigma > 0$ se e solo se $\log(X) \sim \mathcal{N}(0, \sigma^2)$ e la trasformazione $\log(X)$ è biunivoca, allora, senza perdere in generalità, possiamo lavorare con il campione casuale dei dati trasformati $\log(X_1), \ldots, \log(X_n)$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$ e il campione casuale $\mathcal{N}(0, \sigma^2)$ è stato ampiamente studiato a lezione. Nel seguito, $Y_i = \log(X_i)$ per ogni i.

1. Lo stimatore ML di σ^2 nel caso di un campione gaussiano $Y_1, \ldots, Y_n$ i.i.d. $\sim \mathcal{N}(0, \sigma^2)$ è dato dalla statistica $S_0^2 = \sum_{j=1}^n Y_j^2/n$, e quindi lo stimatore cercato è $\hat{\sigma}^2 = \sum_{j=1}^n (\log X_j)^2/n$.
2. Per un campione gaussiano di varianza σ^2 abbiamo $\dfrac{nS_0^2}{\sigma^2} \sim \chi_n^2$. Segue dalle proprietà della famiglia delle distribuzioni gamma che $S_0^2 \sim \Gamma(n/2, 2\sigma^2/n)$ e quindi:

$$\begin{cases} \mathrm{E}(S_0^2) = \frac{n}{2} \times \frac{2\sigma^2}{n} = \sigma^2, \\ \mathrm{Var}(S_0^2) = \frac{n}{2} \times \frac{4\sigma^4}{n^2} = \frac{2\sigma^4}{n} \end{cases} \quad \text{da cui segue la non distorsione}$$

da cui segue la consistenza in media quadratica. Inoltre, asintoticamente la fdr asintotica di S_0^2 è $\mathcal{N}(\sigma^2, 2\sigma^4/n)$.

Per quanto riguarda l'efficienza, anzitutto si ricordi che il modello gaussiano verifica tutte le ipotesi di "regolarità" della disuguaglianza di Fréchet-Cramér-Rao. Dunque, procediamo a calcolare l'informazione di Fisher $I(\sigma^2)$ del modello gaussiano nel caso di media nulla. Abbiamo che

$$\begin{aligned} I(\sigma^2) = \mathrm{E}\left[\left(\frac{\partial \log(f(Y_1; \sigma^2))}{\partial \sigma^2}\right)^2\right] &= \mathrm{E}\left[\left(\frac{\partial \log(\frac{1}{\sqrt{\sigma^2}} \mathrm{e}^{-\frac{Y_1^2}{2\sigma^2}})}{\partial \sigma^2}\right)^2\right] = \mathrm{E}\left[\frac{1}{4\sigma^4}\left(\frac{Y_1^2}{\sigma^2} - 1\right)^2\right] \\ &= \frac{1}{4\sigma^4} \mathrm{E}\left[\left(\frac{Y_1^2}{\sigma^2} - \mathrm{E}\left(\frac{Y_1^2}{\sigma^2}\right)\right)^2\right] = \frac{1}{4\sigma^4} \mathrm{Var}\left[\frac{Y_1^2}{\sigma^2}\right] = \frac{2}{2\sigma^4} = \frac{1}{2\sigma^4} \end{aligned}$$

perché $Y_1^2/\sigma^2 \sim \chi_1^2$. Allora $1/(nI(\sigma^2)) = \mathrm{Var}(S_0^2)$. Quindi lo stimatore S_0^2 è anche efficiente.

3. Un intervallo esatto bilatero per σ^2 di livello di confidenza $1 - \alpha$ è

$$\left(\frac{nS_0^2}{\chi_n^2(1 - \frac{\alpha}{2})}, \frac{nS_0^2}{\chi_n^2(\frac{\alpha}{2})}\right) = \left(\frac{3644.0}{\chi_{55}^2(0.975)}, \frac{3644.0}{\chi_{55}^2(0.025)}\right) = \left(\frac{3644.0}{77.380}, \frac{3644.0}{36.398}\right) = (47.09, 100.12).$$

4. Per la dualità fra IC e VI, a livello $\alpha = 1 - 0.95$ accettiamo l'ipotesi nulla $H_0 : \sigma = 8$, in quanto questa ipotesi è equivalente a $H_0 : \sigma^2 = 64$ e $64 \in (47.09, 100.12)$. Continueremo ad accettare per ogni $\alpha < 5\%$ e quindi anche per $\alpha = 3\%$. ∎

Esercizio 1.3 Una delle densità tradizionalmente usate per modellare la distribuzione dei redditi è la densità di Pareto di parametri α, β data da

$$f(x;\alpha,\beta) = \begin{cases} \frac{\alpha\beta^\alpha}{x^{\alpha+1}} & \text{se } x > \beta \\ 0 & \text{altrove} \end{cases}, \qquad \alpha > 2, \beta > 0.$$

Nel modello di Pareto di parametri α, β con $\alpha > 2$ e $\beta > 0$, la media e il momento secondo sono

$$\mathrm{E}(X) = \frac{\alpha\beta}{(\alpha-1)}, \qquad \mathrm{E}(X^2) = \frac{\alpha\beta^2}{\alpha-2}.$$

Per verificare se un modello di Pareto ben si adatta ai redditi dei lavoratori dipendenti del quartiere OQ di Milano, abbiamo campionato i redditi di 500 lavoratori dipendenti che abitano nel quartiere e i dati ottenuti (espressi in migliaia di euro) sono sintetizzati nella seguente tabella:

A_k	$(0, 11.6]$	$(11.6, 13.4]$	$(13.4, 15.3]$	$(15.3, 18.3]$	$(18.3, 22.2]$	$(22.2, \infty)$
N_k	175	105	95	65	60	0

dove A_k indica la classe di reddito annuo netto, espresso in migliaia di euro, e N_k il numero di dipendenti con reddito appartenente alla classe A_k. Inoltre, il reddito medio campionario (espresso in migliaia di euro) e il momento secondo campionario dei dati raggruppati della precedente tabella valgono rispettivamente 12.00 e 169.64.

1. Determinate gli stimatori dei momenti di α e β sulla base dei valori di reddito medio campionario e momento secondo campionario dei dati raggruppati che vi sono stati forniti.
2. Determinate $P(x_1 < X \le x_2)$ quando X ha densità di Pareto $f(x;\alpha,\beta)$.
3. Valutate con un opportuno test la bontà di adattamento del modello di Pareto ai dati sui redditi dei lavoratori dipendenti del quartiere OQ di Milano. (*Se non siete riusciti a risolvere il punto 1., scegliete voi dei valori per i parametri α e β ed eseguite un opportuno test.*)

Soluzione

1. Siano M_{1c} la media campionaria e M_{2c} il momento secondo campionario dei dati raggruppati. Per ottenere degli stimatori dei momenti di α, β, risolviamo il sistema

$$\begin{cases} \frac{\alpha\beta}{(\alpha-1)} = M_{1c} \\ \frac{\alpha\beta^2}{\alpha-2} = M_{2c} \end{cases}.$$

Dalla prima equazione ricaviamo $\beta = (\alpha-1)M_{1c}/\alpha$ che, sostituita nella seconda, fornisce

$$M_{2c}\alpha(\alpha-2) = (M_{1c})^2(\alpha-1)^2$$

equivalente a

$$\left[M_{2c} - (M_{1c})^2\right]\alpha^2 - 2\left[M_{2c} - (M_{1c})^2\right]\alpha - (M_{1c})^2 = 0,$$

la cui unica soluzione maggiore di 2 è

$$\hat{\alpha} = 1 + \sqrt{1 + \frac{(M_{1c})^2}{M_{2c} - (M_{1c})^2}} = 1 + \sqrt{1 + 12.00^2/25.64} \simeq 3.57.$$

Inoltre,

$$\hat{\beta} = \frac{\hat{\alpha}-1}{\hat{\alpha}} M_{1c} = \frac{3.57-1}{3.57} \times 12.00 \simeq 8.64.$$

2. Si trova facilmente che la funzione di ripartizione di X è $F(x) = 0$ se $x \le \beta$ e

$$F(x) = \int_\beta^x \frac{\alpha\beta^\alpha}{s^{\alpha+1}}\, ds = 1 - \frac{\beta^\alpha}{x^\alpha}, \qquad \forall x > \beta,$$

da cui otteniamo che

$$P(x_1 < X \le x_2) = \begin{cases} 0 & \text{se } x_1 < x_2 \le \beta \\ \beta^\alpha\left(\frac{1}{x_1^\alpha} - \frac{1}{x_2^\alpha}\right) & \text{se } \beta < x_1 < x_2 \\ 1 - \frac{\beta^\alpha}{x_2^\alpha} & \text{se } x_1 \le \beta < x_2 \end{cases}$$

3. Dobbiamo impostare un test chiquadrato di buon adattamento per dati raggruppati per l'ipotesi nulla composta H_0 :"X è paretiana" contro l'alternativa che X non sia paretiana.

Sostituendo i valori di $\hat{\alpha}, \hat{\beta}$ nelle probabilità calcolate al punto 2., otteniamo i valori "stimati" delle probabilità "teoriche": $p_1^{(0)}(\hat{\alpha}, \hat{\beta}) = F_{H_0}(11.6) = 0.65$, $p_2^{(0)}(\hat{\alpha}, \hat{\beta}) = 0.14$, ...; in sintesi

A_k	$(0, 11.6]$	$(11.6, 13.4]$	$(13.4, 15.3]$	$(15.3, 18.3]$	$(18.3, \infty)$
$p_i^{(0)}(\hat{\alpha}, \hat{\beta})$	0.65	0.14	0.08	0.06	0.07
$np_i^{(0)}(\hat{\alpha}, \hat{\beta})$	325	70	40	30	35

.

Dunque, la statistica test di Pearson $Q = \sum_{i=1}^{5} \frac{(N_i - np_i^{(0)}(\hat{\alpha}, \hat{\beta}))^2}{np_i^{(0)}(\hat{\alpha}, \hat{\beta})}$ vale 221.05: rifiutiamo l'ipotesi di dati paretiani a qualunque livello del test, poiché i quantili della distribuzione $\chi^2_{5-1-2} = \chi^2_2 = \mathcal{E}(2)$, di ordine $1 - \alpha$, sono minori di 13.82 per ogni $\alpha > 0.001$. ∎

2 STATISTICA per ING INF TEL [2L] I. Epifani, A. Guglielmi 16.07.08

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

Esercizio 2.1 Per fare inferenza statistica sulla probabilità p che un allievo regolarmente iscritto a un appello d'esame non si presenti a sostenere l'esame, sono state controllate iscrizioni e presenze dell'appello del 25/06/07 di Statistica: esattamente 27 dei 128 allievi regolarmente iscritti sono risultati assenti.

1. Verificate l'ipotesi nulla $H_0 : p \leq 0.25$ contro l'alternativa $H_1 : p > 0.25$, con un test asintotico di livello $\alpha = 3\%$.
2. Determinate la potenza del test al punto 1. quando il vero valore di p è 0.35.
3. Costruite un intervallo di confidenza bilatero asintotico per p di confidenza $\gamma = 90\%$. Nell'approssimazione fermatevi alla seconda cifra decimale.

Vengono preparate le copie della traccia del compito, ovviamente una per ogni allievo, alla chiusura delle iscrizioni all'appello del 18 luglio[1], a cui risultano iscritti 100 allievi.

4. Sulla base dei risultati al punto 3., quanto deve essere fiducioso il docente che il numero medio di copie preparate inutilmente sia compreso fra 15 e 27?

Soluzione

1. Rifiuto $H_0 : p \leq 0.25$ a favore di $H_1 : p > 0.25$ se
$$\hat{p} \geq 0.25 + z_{97\%}\sqrt{\frac{0.25 \times (1 - 0.25)}{128}} \simeq 0.322,$$
dove $\hat{p} = 27/128 \simeq 0.211$; siccome ho tante osservazioni (128) e $n\hat{p} > 5$, allora $\alpha = 3\%$ è un livello approssimato di questo test. Risulta $\hat{p} = 0.211 < 0.322$, quandi accetto H_0 a livello approssimativamente pari al 3%.
2. $\pi(0.35) = P_{0.35}(\hat{p} \geq 0.322) \simeq 1 - \Phi\left(\frac{0.322 - 0.35}{\sqrt{\frac{0.35\times(1-0.35)}{128}}}\right) \simeq 1 - \Phi(-0.66) \simeq 0.74537.$
3. Un IC bilatero asintotico per p di confidenza γ ha estremi
$$IC(p) = \hat{p} \pm z_{\frac{1+\gamma}{2}}\sqrt{\frac{\hat{p} \times (1 - \hat{p})}{128}}.$$
Sostituendo i dati del campione e $\gamma = 10\%$ otteniamo $IC(p) = (0.15, 0.27)$.
4. Ci aspettiamo che mediamente non si presentino all'esame $100 \times p$ allievi dei 100 iscritti. Poiché $(0.15, 0.27)$ fornisce un IC per p di confidenza 90%, allora $(15, 27)$ è un IC per la caratteristica $\kappa(p) = 100 \times p$ sempre di confidenza 90%. Pertanto, il docente sarà fiducioso al 90% che gli allievi per cui avrà preparato le copie ma non si presenteranno, sprecando quindi inutilmente carta e toner, saranno compresi fra 15 e 27. ∎

[1] Data fittizia

Esercizio 2.2 Abbiamo estratto un campione casuale $X_1, \ldots, X_n$ da una popolazione con densità di parametro $\theta > 0$ data da

$$f(x, \theta) = \frac{3x^2}{\theta} \exp\left\{-\frac{x^3}{\theta}\right\} \mathbf{1}_{(0,\infty)}(x).$$

1. Determinate lo stimatore $\hat{\theta}_n$ della caratteristica θ usando il metodo di massima verosimiglianza.
2. Stabilite se lo stimatore $\hat{\theta}_n$ determinato al punto 1. è efficiente.
3. Determinate la densità di $Y = X^3$ e di $Q_n = \dfrac{2n\hat{\theta}}{\theta}$.
4. Fornite un intervallo di confidenza a due code per θ di livello 90% per $n = 10$ e $\hat{\theta} = 0.0387$.

SOLUZIONE

1. Studiamo la funzione di verosimiglianza del campione $X_1, \ldots, X_n$:

$$L_\theta(x_1, \ldots, x_n) = \frac{\prod_{j=1}^n (3x_j^2)}{\theta^n} \exp\left\{-\frac{\sum_{j=1}^n x_j^3}{\theta}\right\}, \qquad \theta > 0$$

$$\log L_\theta(x_1, \ldots, x_n) = -n \log \theta + \log \prod_{j=1}^n (3x_j^2) - \frac{\sum_{j=1}^n x_j^3}{\theta},$$

$$\frac{\partial \log L_\theta(x_1, \ldots, x_n)}{\partial \theta} = -\frac{n}{\theta} + \frac{\sum_{j=1}^n x_j^3}{\theta^2} = \frac{n}{\theta^2}\left(\frac{\sum_{j=1}^n x_j^3}{n} - \theta\right) \tag{1}$$

e

$$\frac{n}{\theta^2}\left(\frac{\sum_{j=1}^n x_j^3}{n} - \theta\right) \geq 0 \quad \text{ se e solo se } \quad \theta \leq \frac{\sum_{j=1}^n x_j^3}{n}.$$

Quindi $\hat{\theta} := \frac{\sum_{j=1}^n X_j^3}{n}$ è MLE per θ.

2. La densità $f(x, \theta)$ è "regolare" e leggiamo nell'equazione (1) che la derivata del logaritmo della funzione di verosimiglianza *"essenzialmente"* è funzione lineare della differenza fra $\hat{\theta}$ e il parametro da stimare θ. Ma (1) è condizione necessaria e sufficiente affinchè la varianza di $\hat{\theta}$ raggiunga il confine di Fréchet-Cramér-Rao $1/(nI(\theta))$ con $I(\theta)$ l'informazione di Fisher. Quindi $\hat{\theta}$ è stimatore efficiente.

3. Se $y \leq 0$ allora $F_{Y,\theta}(y) = 0$; per $y > 0$:

$$F_{Y,\theta}(y) = P_\theta(X^3 \leq y) = P_\theta\left(X \leq y^{1/3}\right) = F_{X,\theta}(y^{1/3}),$$

da cui abbiamo che la densità di Y è

$$f_Y(y, \theta) = f(y^{1/3}, \theta)\frac{1}{3}y^{-2/3}\mathbf{1}_{(0,+\infty)}(y) = \frac{1}{3}y^{-2/3}\frac{3}{\theta}y^{2/3} \exp\left\{-\frac{(y^{1/3})^3}{\theta}\right\} \mathbf{1}_{(0,+\infty)}(y) = \frac{1}{\theta} \exp\left\{-\frac{y}{\theta}\right\} 1_{(0,+\infty)}(y)$$

cioè $Y \sim \mathcal{E}(\theta) = \Gamma(2/2,\ \theta)$. Segue dalle proprietà della famiglia delle leggi gamma che $(2X_j^3)/\theta \sim \Gamma(2/2,\ 2) = \chi^2_2$. Essendo $Q_n = \displaystyle\sum_{j=1}^n 2X_j^3/\theta$ una somma di n va χ^2_2 indipendenti allora $Q_n \sim \chi^2_{2n}$.

4. Per il punto 3., Q_n è una quantità pivotale e possiamo costruire un intervallo bilatero per θ a code simmetriche di livello 0.90 nel seguente modo:
siano $q_1 = \chi^2_{2n}((1-\gamma)/2) = \chi^2_{20}(0.05) = 10.851$ e $q_2 = \chi^2_{2n}((1+\gamma)/2) = \chi^2_{20}(0.95) = 31.410$. (Abbiamo usato la tabella della fdr χ^2_{20}). Allora $P_\theta(q_1 < Q_n < q_2) = 0.90$. Ma, $\{q_1 < Q_n < q_2\} = \left\{2n\hat{\theta}/q_2 < \theta < 2n\hat{\theta}/q_1\right\}$ e quindi $P_\theta\left(2n\hat{\theta}/31.4 < \theta < 2n\hat{\theta}/10.9\right) = 0.90$. Segue che l'intervallo di confidenza per θ di livello 90% è $\left(\dfrac{20 \times 0.0387}{31.410}, \dfrac{20 \times 0.0387}{10.851}\right) \simeq (0.0246, 0.0713)$. ∎

Esercizio 2.3 Abbiamo misurato, in MegaPascal, i carichi di rottura a trazione di due tipi di filamenti di fibra di carbonio, uno di tipo *standard* e l'altro di un tipo *nuovo*, ottenuto da un processo di carbonizzazione, che dovrebbe produrre materiale più resistente, cioè con più alto carico di rottura.

Per verificare se effettivamente il nuovo materiale sia più resistente di quello standard, si raccolgono i valori dei carichi di rottura, in scala logaritmica, di 20 provini di filamenti di fibra di carbonio, di cui 13 di tipo *standard* e 7 di tipo *nuovo*. I dati sono:

$$\text{standard}: \ 5.97, 6.47, 6.01, 6.16, 6.27, 6.17, 6.36, 6.54, 6.19, 6.44, 6.24, 6.59, 6.13$$

$$\text{nuovo}: \ 6.25, 6.37, 6.45, 6.60, 6.21, 6.56, 6.28.$$

Deduciamo che i ranghi del campione dei 7 provini di tipo *nuovo* sono: 7, 9, 11, 13, 15, 18, 20.

1. Impostate un test di livello 10% per stabilire se i dati mostrano evidenza sperimentale a favore dell'ipotesi che il *nuovo* materiale sia più resistente di quello *standard*.

In realtà, possiamo supporre che i due campioni dei logaritmi dei carichi di rottura siano gaussiani. Inoltre, dai dati a disposizione troviamo

$$\sum_{j=1}^{7} x_j = 44.72\,, \quad \sum_{j=1}^{7} x_j^2 = 285.84\,, \quad \sum_{j=1}^{13} y_j = 81.54\,, \quad \sum_{j=1}^{13} y_j^2 = 511.9\,,$$

dove $\{x_j\}$ e $\{y_j\}$ indicano i carichi di rottura dei materiali *nuovo* e *standard*, rispettivamente.

2. Verificate con un test di significatività pari al 5% se i due campioni hanno la stessa varianza.

3. Costruite un test di ipotesi parametrico di livello $\alpha = 5\%$, che utilizzi l'ipotesi di normalità e le informazioni desunte dal punto 2., per stabilire se i dati mostrano evidenza sperimentale a favore dell'ipotesi che il *nuovo* materiale sia più resistente di quello *standard*.

SOLUZIONE

1. Non avendo nessuna informazione sulla famiglia di densità da cui sono stati estratti i due campioni di dati, impostiamo il test di omogeneità unilatero non parametrico di Wilcoxon-Mann-Whitney per verificare l'ipotesi nulla H_0 :"Il materiale *nuovo* ha una resistenza al più pari a quello *standard*" contro l'alternativa H_1 :"il materiale *nuovo* è più resistente di quello *standard*". Se F_X indica la funzione di ripartizione dei carichi di rottura dei filamenti di fibra di tipo *nuovo* e F_Y di quelli di tipo *standard*, allora il problema di verifica di ipotesi diventa

   $$H_0 : F_X(x) \geq F_Y(x) \ \forall x \text{ e } H_1 : F_X(x) \leq F_Y(x) \ \forall x \text{ (e } F_X(x) < F_Y(x) \text{ per almeno qualche } x).$$

   Usiamo la statistica T_X data dalla somma dei ranghi del campione dei 7 provini di tipo *nuovo*: $T_X = 7 + 9 + 11 + 13 + 15 + 18 + 20 = 93$. Rifiutiamo H_0 a livello 10% se $T_X > w_{7,13}(1-10\%) = 7(7+13+1) - w_{7,13}(10\%) = 147 - 57 = 90$; in questo caso risulta $93 > 90$ e quindi, a livello di significatività $\alpha = 0.1$, rifiutiamo H_0 e accettiamo l'ipotesi alternativa che il materiale *nuovo* sia più resistente di quello *standard*.

2. Poniamoci ora sotto ipotesi di normalità dei due campioni casuali indipendenti e impostiamo un test F per verificare $H_0 : \sigma_X^2 = \sigma_Y^2$ *vs* $H_1 : \sigma_X^2 \neq \sigma_Y^2$.

   I valori osservati delle varianze campionarie sono $s_X^2 \simeq 0.0235$ e $s_Y^2 \simeq 0.0382$ e dunque $s_X^2/s_Y^2 \simeq 0.615$. Inoltre, rifiutiamo H_0 se $s_X^2/s_Y^2 \notin \left(F_{m-1,n-1}(\frac{\alpha}{2}), F_{m-1,n-1}(1-\frac{\alpha}{2})\right)$, dove $F_{m-1,n-1}(a)$ è il quantile di ordine a della f.d.r. di Fisher con $m-1$ gradi di libertà al numeratore e $n-1$ gradi di libertà al denominatore. Dalle tavole abbiamo $F_{m-1,n-1}(\alpha/2) = F_{6,12}(0.025) = 1/F_{12,6}(0.975) = 1/5.37 \simeq 0.186$ e $F_{m-1,n-1}(1-\alpha/2) = F_{6,12}(0.975) = 3.73$. Poiché $0.615 \in (0.186, 3.73)$, allora accettiamo l'ipotesi di uguaglianza delle varianze al livello di significatività del 5%.

3. Ora possiamo assumere che i due campioni gaussiani $(X_1, \ldots, X_7)$ e $(Y_1, \ldots, Y_{13})$ abbiano stessa varianza incognita σ^2. Impostiamo allora un t test per verificare $H_0 : \mu_X \leq \mu_Y$ *vs* $H_1 : \mu_X > \mu_Y$; rifiutiamo H_0 a livello $\alpha = 5\%$ se $T = (\bar{X} - \bar{Y})/\sqrt{S_p^2(\frac{1}{m} + \frac{1}{n})} \geq t_{18}(0.95)$. Lo varianza pooled vale $s_p^2 \simeq 0.0333$, mentre $\bar{x} = 6.3886$, $\bar{y} = 6.2723$ e quindi $T \simeq 1.3595$ che è minore di $t_{18}(0.95) = 1.734$: questa volta, a livello 5%, non rifiutiamo H_0.

   Un'osservazione aggiuntiva: se sappiamo che effettivamente i due campioni di dati sono gaussiani con stessa varianza, allora, a parità di significatività, il test t è preferibile al test di Wilcoxon-Mann-Whitney perché usa più informazioni, cioè la normalità dei dati. Ma, attenzione che nel nostro caso i test al punto 1. e 3. non hanno la stessa significatività. ∎

‘

3 STATISTICA per ING INF TEL [2L] I. Epifani, A. Guglielmi 05.09.08

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

Esercizio 3.1 Per stabilire se una certa misurazione abbia densità uniforme continua sull'intervallo $(0,3)$ o uniforme continua sull'intervallo $(0,4)$, abbiamo effettuato un'unica misurazione X e abbiamo adottato la seguente regola decisionale: se $X \geq 2.7$, allora rifiutiamo l'ipotesi nulla H_0: *"la misurazione ha densità* $U(0,3)$*"* a favore dell'ipotesi alternativa H_1: *"la misurazione ha densità* $U(0,4)$*"*.

1. Calcolate il livello di significatività α di questo test.
2. Calcolate la potenza π di questo test.
3. Calcolate il $p-$value di questo test se la misurazione è $x = 2.4$.

SOLUZIONE

1. $\alpha = P_{U(0,3)}(X \geq 2.7) = \dfrac{3-2.7}{3-0} = \dfrac{0.3}{3} = 0.1.$
2. $\pi = P_{U(0,4)}(X \geq 2.7) = \dfrac{4-2.7}{4-0} = \dfrac{1.3}{4} = 0.325.$
3. Dobbiamo calcolare il più piccolo valore di α per cui si rifiuta H_0 quando $x = 2.4$. Rifiutiamo H_0 a livello α, per ogni α tale che $\alpha \geq P_{U(0,3)}(X \geq 2.4)$. Siccome $P_{U(0,3)}(X \geq 2.4) = (3-2.4)/(3-0) = 0.2$, allora il $p-$value vale 20% che è un valore alto. Concludiamo che non c'è evidenza empirica contro H_0. ∎

Esercizio 3.2 Il tempo di vita T, espresso in anni, dei sacchetti di plastica biodegradabili del supermercato *ZZ* ha densità

$$f(t,b) = \begin{cases} \frac{b}{2}\left(1-\frac{t}{2}\right)^{b-1} & \text{se } 0 < t < 2 \\ 0 & \text{altrimenti ,} \end{cases} \qquad \text{con } b > 0.$$

Il parametro positivo b è incognito e per stimarlo si registrano le durate di un campione di 75 sacchetti, $T_1, \ldots, T_{75}$, e si ottiene che $\displaystyle\sum_{j=1}^{75} \ln\left(1-\frac{t_j}{2}\right) = \ln\left[\prod_{j=1}^{75}\left(1-\frac{t_j}{2}\right)\right] \simeq -28.98.$

1. Calcolate la probabilità che un sacchetto estratto a caso dalla popolazione considerata duri meno di 15 mesi.
2. Usando il metodo di massima verosimiglianza stimate il parametro incognito b e la probabilità che un sacchetto duri meno di 15 mesi.
3. Verificate che l'Informazione di Fisher del modello statistico $\{f(t,b), b > 0\}$ è $I(b) = 1/b^2$. (*Suggerimento*: $\int_0^2 \left[\ln\left(1-\frac{t}{2}\right)\right]^k \frac{b}{2}\left(1-\frac{t}{2}\right)^{b-1} dt = \frac{(-1)^k k!}{b^k}$, $\forall k = 1, 2, \ldots$)
4. Determinate un intervallo di confidenza asintotico unilatero del tipo $(0,c)$ di livello 95% per il parametro b e uno sempre del tipo $(0,c)$ per la probabilità che un sacchetto duri meno di 15 mesi.

SOLUZIONE

1. Si tratta di calcolare la caratteristica $\kappa(b)$ data da

$$\kappa(b) = P\left(T < \frac{5}{4}\right) = \int_0^{\frac{5}{4}} \frac{b}{2}\left(1-\frac{t}{2}\right)^{b-1} dt = \left[-\left(1-\frac{t}{2}\right)^b\right]_0^{\frac{5}{4}} = 1-\left(\frac{3}{8}\right)^b.$$

2. La funzione di verosimiglianza del campione $t_1, \ldots, t_{75}$, per $t_1 \in (0,2), \ldots, t_{75} \in (0,2)$, è

$$L_b(t_1, \ldots, t_{75}) = \frac{b^{75}}{2^{75}}\left(\prod_{j=1}^{75}\left(1-\frac{t_j}{2}\right)\right)^{b-1}$$

da cui deriviamo che

$$\frac{\partial}{\partial b}\ln L_b(t_1,\ldots,t_{75}) = \frac{75}{b} + \ln\prod_{j=1}^{75}\left(1-\frac{t_j}{2}\right) = 75\left(\frac{1}{b} + \frac{\sum_{j=1}^{75}\ln\left(1-\frac{t_j}{2}\right)}{75}\right) \tag{2}$$

e quindi $\frac{\partial}{\partial b}\ln L_b(t_1,\ldots,t_{75}) \geq 0$ se e solo se $b \leq -\dfrac{75}{\sum_{j=1}^{75}\ln\left(1-\frac{t_j}{2}\right)}$. Segue che lo stimatore ML di b è $\hat{b}_{ML} = -\dfrac{75}{\sum_{j=1}^{75}\ln\left(1-\frac{t_j}{2}\right)}$ e di $\kappa(b)$ è $\hat{\kappa}_{ML} = 1-\left(\frac{3}{8}\right)^{\hat{b}_{ML}}$.
Con i dati forniti le stime hanno valore $\hat{b}_{ML} \simeq 2.59$ e $\hat{\kappa}_{ML} \simeq 0.92$.

3. $$\begin{aligned} I(b) &= \mathrm{E}[(\frac{\partial}{\partial b}\ln f(T,b))^2] = \mathrm{E}\left[\left(\frac{1}{b}+\ln\left(1-\frac{T}{2}\right)\right)^2\right] \\ &= \frac{1}{b^2} + \frac{2}{b}\,\mathrm{E}\left[\ln\left(1-\frac{T}{2}\right)\right] + \mathrm{E}\left[\left(\ln\left(1-\frac{T}{2}\right)\right)^2\right] \\ &= \frac{1}{b^2} + \frac{2}{b}\int_0^2 \ln\left(1-\frac{t}{2}\right)\frac{b}{2}\left(1-\frac{t}{2}\right)^{b-1}dt + \int_0^2\left[\ln\left(1-\frac{t}{2}\right)\right]^2\frac{b}{2}\left(1-\frac{t}{2}\right)^{b-1}dt \\ &= \frac{1}{b^2} + \frac{2}{b}\times\frac{(-1)}{b} + \frac{(-1)^2\times 2}{b^2} = \frac{1}{b^2}. \end{aligned}$$

4. Stiamo lavorando con un modello statistico "regolare"; quindi, segue dalle proprietà asintotiche degli stimatori ML di questo modelli che la funzione di ripartizione asintotica di $\hat{b}_{ML}$ è gaussiana di media b e varianza $1/(nI(b))$. Pertanto,

$$0.95 \simeq P\left(\sqrt{nI(b)}(\hat{b}_{ML}-b) > -1.645\right) = P\left(\sqrt{n}\left(\frac{\hat{b}_{ML}}{b}-1\right) > -1.645\right) = P\left(b < \frac{\hat{b}_{ML}}{1-1.645/\sqrt{n}}\right).$$

Dunque un IC di b unilatero asintotico di confidenza 95% è dato da $\left(0, \dfrac{2.59}{1-1.645/\sqrt{75}}\right) \simeq (0, 3.20)$.
Osservando che $\kappa(p) = 1-(3/8)^b$ è funzione crescente di b, otteniamo che $(0, 1-(3/8)^{3.20}) \simeq (0, 0.96)$ è un IC di $\kappa(b)$ unilatero asintotico di confidenza 95%.

Soluzione alternativa: La stima ML di $1/(nI(b))$ vale $2.59^2/75 \simeq 0.089$ e quindi $P\left(\frac{\hat{b}_{ML}-b}{\sqrt{0.089}} > -1.645\right) \simeq 0.95$. Segue che un (altro) *IC* di b unilatero del tipo $(0,c)$ asintotico di confidenza 95% è dato da $(0, \hat{b}_{ML}+1.645\times\sqrt{0.089}) \simeq (0, 3.08)$, mentre per $\kappa(b)$ è $(0, 0.95)$. ∎

Esercizio 3.3 Abbiamo raccolto dei dati su 200 allievi del corso di laurea in `xxx` che hanno superato l'esame di `LIN` e `SIN`, entrambi obbligatori, e li abbiamo sintetizzati come segue.

Tabella 1: # appelli sostenuti per superare `LIN` e `SIN`

`LIN\ SIN`	1	2	3 o più	
1	60	10	10	
2	30	15	15	
3 o più	10	20	30	

In Tabella 1[2] troviamo il numero di appelli sostenuti dagli allievi per superare gli esami `LIN` e `SIN`; per esempio, 20 è il numero di allievi che in `SIN` sono stati promossi al secondo appello sostenuto e che in `LIN` hanno dovuto faticare

[2]Dati del tutto fittizi, generati in R.

almeno 3 appelli. Poi, in (3) abbiamo le statistiche sui voti finali registrati, espressi in trentesimi:

$$\sum_{j=1}^{200} x_j = 4600, \sum_{j=1}^{200} x_j^2 = 107600, \sum_{j=1}^{200} y_j = 4900, \sum_{j=1}^{200} y_j^2 = 120850, \sum_{j=1}^{200} x_j y_j = 113704, \tag{3}$$

dove $\{x_j\}$ sono i voti di `LIN` e $\{y_j\}$ quelli di `SIN`.

1. Usate un opportuno test di livello $\alpha = 5\%$ per stabilire se ci sia dipendenza fra il numero di appelli necessari per superare gli esami `LIN` e `SIN`.

2. Stabilite se il voto medio registrato di `LIN` sia più basso di quello medio registrato per `SIN`. A tal fine, costruite un opportuno test tale che sia al più pari ad $\alpha = 5\%$ la probabilità di commettere l'errore di prima specie di ritenere il voto in `LIN` minore di quello in `SIN`, quando effettivamente è maggiore o uguale. Ipotizzate la normalità dei voti.

3. Stabilite se i voti riportati in `LIN` e `SIN` siano o no indipendenti, a livello $\alpha = 5\%$. Ipotizzate la normalità dei voti.

Soluzione

1. Impostiamo un test chi-quadrato di indipendenza fra le variabili L ed S che contano rispettivamente il numero di appelli necessari per superare `LIN` e quello per superare `SIN`. Le ipotesi nulla e alternativa sono rispettivamente H_0 : *"L e S sono indipendenti"*, H_1 : *"L e S non sono indipendenti"*. Completiamo la Tabella 1 con le numerosità marginali:

$L \setminus S$	1	2	3 o più	
1	60	10	10	80
2	30	15	15	60
3 o più	10	20	30	60
	100	45	55	

La statistica di Pearson ha valore

$$Q = 200 \left(\sum_{i=1}^{3} \sum_{j=1}^{3} \frac{N_{ij}^2}{N_{i\cdot} N_{\cdot j}} - 1 \right) \simeq 47.9.$$

Asintoticamente Q ha f.d.r. $\chi^2_{(3-1)(3-1)} = \chi^2_4$. Poichè risulta $47.9 > \chi^2_4(1 - 0.05) = 9.488$, rifiutiamo l'ipotesi H_0 di indipendenza fra le variabili L e S. Inoltre, il $p-$value del test risulta $1 - F_{\chi^2_4}(47.9) \leq 1 - F_{\chi^2_4}(18.467) = 0.1\%$: concludiamo che c'è una netta evidenza sperimentale contro l'ipotesi H_0.

2. Siano μ_X il voto medio in `LIN` e μ_Y quello in `SIN`. Dobbiamo impostare un t-test per dati gaussiani accoppiati di significatività $\alpha = 5\%$ per verificare $H_0 : \mu_X - \mu_Y \geq 0$ contro $H_1 : \mu_X - \mu_Y < 0$; quindi rifiutiamo H_0 se $t := \frac{\bar{x} - \bar{y}}{\sqrt{s^2_{X-Y}/200}} \leq -t_{199}(95\%)$. Con i dati forniti abbiamo

$$\bar{x} = 23,\ \bar{y} = 24.5,\ s_X^2 = 9.05,\ s_Y^2 = 4.02,\ \frac{\sum_{j=1}^{200}(x_j - \bar{x})(y_j - \bar{y})}{199} \simeq 5.05,$$

$$s^2_{X-Y} = s_X^2 + s_Y^2 - \frac{2}{199} \sum_{j=1}^{200} (x_j - \bar{x})(y_j - \bar{y}) = 9.05 + 4.02 - 2 \times 5.05 \simeq 2.97\ .$$

Poiché $t \simeq -12.31 < -1.645$, allora rifiutiamo l'ipotesi nulla che il voto in `LIN` sia mediamente maggiore o uguale di quello in `SIN`.

3. Verifichiamo le ipotesi sul coefficiente di correlazione lineare date da $H_0 : \rho = 0$ contro $H_1 : \rho \neq 0$, con il t-test che prescrive di rifiutare H_0 a livello 5% se $R\sqrt{n-2}/\sqrt{1-R^2} \geq t_{198}(97.5\%)$. Il coefficiente di correlazione campionario $R = \sum_j (X_j - \bar{X})(Y_j - \bar{Y}) / \sqrt{\sum_j (X_j - \bar{X})^2 \sum_j (Y_j - \bar{Y})^2}$ ha valore $r = 0.837$ e la statistica test $R\sqrt{n-2}/\sqrt{1-R^2}$ ha valore 21.523. Poiché $t_{198}(97.5\%) \simeq z_{97.5} \simeq 1.96$, allora rifiutiamo H_0 a livello $\alpha = 5\%$. In realtà, la statistica test ha un valore così elevato che rifiutiamo l'ipotesi di indipendenza a qualunque livello del test. ∎

4 STATISTICA per ING INF TEL [2L] I. Epifani, A. Guglielmi 22.09.08

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

Esercizio 4.1 Sia $X_1, \ldots, X_n$ un campione casuale estratto da una popolazione di densità

$$f(x;\theta) = \begin{cases} 3\theta^3 x^{-4} & \text{se } x > \theta \\ 0 & \text{altrove} \end{cases}$$

con θ parametro positivo incognito.

1. Determinate $\mathrm{E}(X_1)$ e $\mathrm{Var}(X_1)$.
2. Determinate uno stimatore di θ usando il metodo dei momenti.
3. Verificate se lo stimatore ottenuto al punto 2. è non distorto e consistente in media quadratica per θ.

SOLUZIONE

1. $\mathrm{E}(X_1) = \displaystyle\int_\theta^{+\infty} x 3\theta^3 x^{-4}\,dx = \int_\theta^{+\infty} 3\theta^3 x^{-3}\,dx = 3\theta^3 \frac{x^{-2}}{-2}\Big|_\theta^{+\infty} = \frac{3\theta}{2}$;
   $\mathrm{E}(X_1^2) = \displaystyle\int_\theta^{+\infty} x^2 3\theta^3 x^{-4}\,dx = \int_\theta^{+\infty} 3\theta^3 x^{-2}\,dx = -3\theta^3 x^{-1}\big|_\theta^{+\infty} = 3\theta^2$;
   $\mathrm{Var}(X_1) = \mathrm{E}(X_1^2) - [\mathrm{E}(X_1)]^2 = 3\theta^2 - \frac{9}{4}\theta^2 = \frac{3\theta^2}{4}$.
2. Dobbiamo risolvere l'equazione nell'incognita θ: $\mathrm{E}_\theta(X_1) = \dfrac{3\theta}{2} = \bar{X}$; la soluzione è $\theta = 2\bar{X}/3$. Segue che lo stimatore dei momenti di θ è $\hat{\theta} = 2\bar{X}/3$.
3. Poiché
   $$\mathrm{E}(\hat{\theta}) = \frac{2}{3}\,\mathrm{E}(\bar{X}) = \frac{2}{3}\,\mathrm{E}(X_1) = \frac{2}{3} \times \frac{3\theta}{2} = \theta$$
   allora $\hat{\theta}$ è stimatore non distorto di θ e l'errore quadratico medio coincide con la sua varianza data da
   $$\mathrm{Var}(\hat{\theta}) = \frac{4}{9}\,\mathrm{Var}(\bar{X}) = \frac{4}{9} \times \frac{\mathrm{Var}(X_1)}{n} = \frac{4}{9} \times \frac{\frac{3\theta^2}{4}}{n} = \frac{\theta^2}{3n}\ .$$
   Poiché $\lim_{n\to\infty} \theta^2/(3n) = 0\ \forall \theta > 0$, allora concludiamo che $\hat{\theta}$ è uno stimatore di θ consistente in media quadratica.
   ∎

Esercizio 4.2 Dobbiamo fare inferenza sul tempo medio μ di esecuzione di un programma, espresso in secondi (sec), e sappiamo che la varianza è nota e vale $\sigma^2 = 196$ sec^2. Per questo motivo, abbiamo fatto girare il programma 9 volte ottenendo un tempo medio campionario pari a 230 sec.
Rispondete alle domande seguenti ipotizzando la normalità dei dati.

1. Determinate un intervallo di confidenza bilatero di livello 94% per μ.
2. Determinate quante ALTRE volte bisogna far girare il programma affinchè la lunghezza dell'intervallo di confidenza si dimezzi.

Sia n_2 la risposta al punto precedente. Al fine di ottenere un intervallo due volte più preciso di quello al *punto 1.*, abbiamo fatto girare il programma altre n_2 volte e abbiamo ottenuto che il tempo medio campionario delle nuove n_2 esecuzioni vale 235 sec.

3. Fornite la stima puntuale di μ basata su tutte le $9 + n_2$ esecuzioni del programma; quindi ricalcolate l'intervallo di confidenza bilatero per μ di livello 94%.
4. Verificate l'ipotesi $H_0 : \mu = 237$ contro $H_1 : \mu \neq 237$ a livello $\alpha = 6\%$, usando come dati tutte le $9+n_2$ esecuzioni del programma.
5. Quanto vale la potenza del test al punto 4. se effettivamente il tempo medio di esecuzione del programma è pari a 241 sec?

Soluzione

1. Un $IC(\mu)$ bilatero di livello 94% ha estremi $\bar{X} \pm z_{\frac{1+0.94}{2}} \sqrt{\frac{\sigma^2}{n}}$ i cui valori numerici sono $230 \pm 1.88 \times 14/3$. Quindi risulta $IC(\mu) = (221.23, 238.77)$.

2. La lunghezza dell'$IC(\mu)$ simmetrico, con varianza nota e n dati gaussiani, è data da $\mathcal{L} = 2z_{\frac{1+0.94}{2}} \sqrt{\sigma^2/n}$ Essa risulta pari a metà del valore della lunghezza dell'$IC(\mu)$ al punto 1. se la numerosità n è tale che

$$\frac{2 \times z_{\frac{1+0.94}{2}} \times 14}{\sqrt{n}} = \frac{z_{\frac{1+0.94}{2}} \times 14}{3} \ .$$

La soluzione è $n = 36$. In definitiva, dobbiamo far girare il programma per altre $n_2 = 36 - 9 = 27$ volte.

3. Ovviamente continueremo a stimare μ con il tempo medio campionario di esecuzione del programma, $\bar{X}$, che sulle 36 esecuzioni è dato da

$$\bar{X}_{36} = \frac{9}{36} \times 230 + \frac{27}{36} \times 235 = \frac{935}{4} = 233.75 \ .$$

Il nuovo $IC(\mu)$ bilatero è $(233.75 - 1.88 \times 14/6, 233.75 + 1.88 \times 14/6) \simeq (229.36, 238.14)$.

4. Per la dualità fra IC e VI, a livello $\alpha = 1 - 0.94$ accettiamo l'ipotesi nulla $H_0 : \mu = 237$ contro $H_1 : \mu \neq 237$ a livello $\alpha = 6\%$, in quanto 237 cade nell'$IC(\mu)$ del punto 3.

5. Calcoliamo $\pi(241)$:

$$\begin{aligned}
\pi(241) &= P_{241}\left\{237 \notin \left(\bar{X} - z_{\frac{1+0.94}{2}} \frac{14}{6}, \bar{X} + z_{\frac{1+0.94}{2}} \frac{14}{6}\right)\right\} \\
&= 1 - P_{241}\left(\bar{X} - z_{\frac{1+0.94}{2}} \frac{14}{6} < 237 < \bar{X} + z_{\frac{1+0.94}{2}} \frac{14}{6}\right) \\
&= 1 - P_{241}\left(237 - z_{\frac{1+0.94}{2}} \frac{14}{6} < \bar{X} < 237 + z_{\frac{1+0.94}{2}} \frac{14}{6}\right) \\
&= 1 - P_{241}\left(\frac{237 - 241}{14/6} - z_{\frac{1+0.94}{2}} < \frac{\bar{X} - 241}{14/6} < \frac{237 - 241}{14/6} + z_{\frac{1+0.94}{2}}\right) \\
&\simeq 1 - \Phi(0.17) + \Phi(-3.59) = 1 - \Phi(0.17) + 1 - \Phi(3.59) \simeq 2 - 0.5675 - 0.9998 = 0.4327 \ . \quad \blacksquare
\end{aligned}$$

Esercizio 4.3 Per condurre uno studio sulla presenza femminile nel mondo accademico scientifico, vengono raccolti dei dati sul numero di donne fra il personale docente dei Dipartimenti di Matematica in Italia. I dati[3] relativi ai docenti del Dipartimento di Matematica del Politecnico di Milano, divisi nelle categorie *ricercatore*, *professore associato* e *professore ordinario* (R, PA e PO nel seguito) sono riportati nella tabella seguente:

Sesso\ Qualifica	R	PA	PO
F	11	20	6
M	21	15	27

e, in prima approssimazione, possiamo ritenere che costituiscano un campione estratto dalla popolazione dei docenti di tutti i Dipartimenti di Matematica italiani.

1. Secondo voi, la qualifica professionale dei docenti dei Dipartimenti di Matematica dipende dal sesso? Per rispondere, usate i dati del Politecnico e costruite un opportuno test di livello approssimato $\alpha = 5\%$.

2. Verificate l'ipotesi nulla che la distribuzione della qualifica professionale dei docenti dei Dipartimenti di Matematica sia uniforme discreta sull'insieme $\{R, PA, PO\}$. Usate i dati del Politecnico e costruite un opportuno test di livello approssimato $\alpha = 1\%$.

3. Usate i dati in tabella per costruire un intervallo di confidenza bilatero di livello approssimato $\gamma = 95\%$ per la percentuale (sull'intera popolazione dei docenti dei Dipartimenti di Matematica italiani) di professori ordinari di sesso femminile.

SOLUZIONE

1. Impostiamo un test χ^2 di indipendenza fra le variabili $Sesso$ e $Qualifica$. Le ipotesi nulla e alternativa sono rispettivamente H_0 : *"sesso e qualifica sono indipendenti"*, H_1 : *"sesso e qualifica non sono indipendenti"*. I docenti del Dipartimento di Matematica del Politecnico sono 100 e le numerosità marginali sono

Sesso\ Qualifica	R	PA	PO	
F	11	20	6	37
M	21	15	27	63
	32	35	33	100

La statistica di Pearson ha valore

$$Q_1 = 100 \left(\sum_{i=1}^{2} \sum_{j=1}^{3} \frac{N_{ij}^2}{N_{i\cdot} N_{\cdot j}} - 1 \right) \simeq 11.2$$

Inoltre, $N_{i\cdot} N_{\cdot j}/100 \geq 37 \times 31/100 = 11.84 > 5$ per ogni $i = F, M$ e $j = R, PA, PO$. Quindi possiamo approssimare la f.d.r. di Q_1 con la sua f.d.r. asintotica che è $\chi^2_{(2-1)(3-1)} = \chi^2_2$. Poiché $11.2 > 5.991 = \chi^2_2(1 - 0.05)$, allora rifiutiamo l'ipotesi H_0. Inoltre, il p-value del test risulta $1 - F_{\chi^2_2}(11.2) = \int_{11.2}^{\infty} (1/2)\mathrm{e}^{-x/2} dx = \mathrm{e}^{-11.2/2} \simeq 0.0037$: concludiamo che c'è una netta evidenza sperimentale contro l'ipotesi di indipendenza fra genere e qualifica professionale.

2. Impostiamo un test χ^2 di buon adattamento per verificare l'ipotesi H_0: "$P(Qualifica = j) = 1/3,\ \forall j = PO, PA, R$" contro H_1: *"$P(Qualifica = j) \neq 1/3$ per qualche j"*. La statistica di Pearson ha valore:

$$Q_2 = \sum_{j=PO,PA,R} \frac{N_{sj}^2}{100 \times (1/3)} - 100 = 0.14$$

ed ha f.d.r. asintotica $\chi^2_{3-1} = \chi^2_2$. Segue che a un livello approssimato $\alpha = 1\%$ non rifiutiamo H_0 perché $0.14 < 9.21 = \chi^2_2(0.99)$.

3. La frequenza relativa campionaria di professori ordinari donne è $\hat{p} = 6/100 = 0.06$ e un intervallo bilatero di confidenza approssimata $\gamma = 0.95$ è dato da

$$\left(\hat{p} - z_{0.975} \sqrt{\frac{\hat{p}(1-\hat{p})}{100}}, \hat{p} + z_{0.975} \sqrt{\frac{\hat{p}(1-\hat{p})}{100}} \right) = (0.013, 0.107). \quad \blacksquare$$

[3] Dati fittizi.

5 STATISTICA per ING INF TEL [2L] I. Epifani, A. Guglielmi 23.02.09

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

Esercizio 5.1 Un macchinario che produce Compact Disk (CD) è tarato in modo tale che il raggio di ogni CD, espresso in cm, sia una variabile aleatoria gaussiana R di media nominale $\mu_0 = 12\,\text{cm}$ e varianza incognita σ^2. Inoltre, il produttore di questo macchinario dichiara che non più del 2% dei CD prodotti con questo macchinario ha lunghezza del raggio distante almeno $0.308\,\text{cm}$ dalla media nominale.

La ditta `xxx`, produttrice di CD e intenzionata ad acquistare questo macchinario, teme che il produttore del macchinario non dichiari il vero. Per questo motivo controlla i raggi r_i di un campione casuale di 25 CD, ottenendo:

$$\sum_{i=1}^{25} r_i = 299.0\,\text{cm} \qquad \sum_{i=1}^{25} r_i^2 = 3576.3\,\text{cm}^2$$

Sia ora κ la percentuale, sull'intera popolazione di CD, di quelli che hanno lunghezza del raggio distante almeno $0.308\,\text{cm}$ dalla media nominale μ_0.

1. Determinate κ in funzione di σ^2. Quindi, stabilite per quali valori di σ^2 la percentuale κ è minore o uguale al 2%.
2. Fornite una stima puntuale di σ^2 e di κ.
3. Sulla base di questi dati, la ditta `xxx` può accusare il produttore del macchinario di non dichiarare il vero? Per rispondere impostate un opportuno test sulla varianza, specificando: ipotesi nulla, ipotesi alternativa e una regione critica di ampiezza 5%.
4. Fornite analiticamente e rappresentate graficamente la funzione di potenza del test costruito al punto 3.

Soluzione

1. Risulta $\kappa = P(|R-12| > 0.308) = 2\left(1 - \Phi\Big(\dfrac{0.308}{\sigma}\Big)\right)$, $\sigma > 0$. Inoltre, $\kappa \le 2\%$ sse $\Phi(0.308/\sigma) > 0.99$, cioè sse $0.308/\sigma \ge z_{0.99} \simeq 2.326$. Dunque $\sigma^2 \le (0.308/z_{0.99})^2 \simeq 0.0175$.
2. Essendo la media nota, stimiamo σ^2 con $S_0^2 = \sum_{i=1}^{25}(R_i - 12)^2/25$ che ha valore $s_0^2 = \sum_{i=1}^{25} r_i^2/25 + 12^2 - 2 \times 12 \sum_{=1}^{25} r_i/25 = 0.012$.
   Una stima puntuale di κ è data da $\hat\kappa = 2\left(1 - \Phi(0.308/\sqrt{s_0^2})\right) = 2\left(1 - \Phi(0.308/\sqrt{0.012})\right) = 2\left(1 - \Phi(2.81)\right) = 0.0049 \simeq 0.5\%$. Infatti, nel modello gaussiano con media nota, S_0^2 è lo stimatore ML di σ^2 e quindi $\hat\kappa = \kappa(S_0^2)$ è quello ML di κ.
3. Il produttore del macchinario dichiara il falso se la "vera" percentuale κ ha valore maggiore di 2%. Dobbiamo impostare quindi un test di verifica dell'ipotesi nulla (da confutare) $H_0 : \kappa \le 2\%$ contro l'alternativa $H_1 : \kappa > 2\%$. In termini di σ^2 le ipotesi diventano: $H_0 : \sigma^2 \le (0.308/z_{0.99})^2$ contro l'alternativa $H_1 : \sigma^2 > (0.308/z_{0.99})^2$. Una regione critica di ampiezza 5% è

   $$G = \left\{(x_1, \dots, x_{25}) \in \mathbb{R}^{25} : 25 s_0^2 \ge (0.308/z_{0.99})^2 \chi^2_{25}(0.95)\right\} = \{(x_1, \dots, x_{25}) \in \mathbb{R}^{25} : 25 s_0^2 \ge 0.659\}.$$

   Poiché $25 s_0^2 = 0.3 < 0.659$, allora il campione analizzato non cade nella regione critica e concludiamo che non si può rifiutare H_0 a livello di significatività del 5%, cioè la ditta `xxx` non può accusare il produttore del macchinario di non dichiarare il vero.
4. La funzione di potenza del test descritto al punto 3. è data da

   $$\pi(\sigma^2) = P_{\sigma^2}(G) = P_{\sigma^2}(25 S_0^2 \ge 0.659) = P_{\sigma^2}\left(\frac{25 S_0^2}{\sigma^2} \ge \frac{0.659}{\sigma^2}\right) = 1 - F_{\chi^2_{25}}\left(\frac{0.659}{\sigma^2}\right), \quad \sigma^2 > (0.308/z_{0.99})^2.$$

   $\pi(\sigma^2)$ è funzione crescente di σ^2, ha un asintotico orizzontale in 1 per $\sigma^2 \to \infty$ e tende a 5% per $\sigma^2 \to (0.308/z_{0.99})^2$. ∎

Esercizio 5.2 Abbiamo estratto un campione casuale $X_1, \ldots, X_n$ dalla funzione di densità discreta

$$f(x,\theta) = \theta^{3|x|}\left(1-2\theta^3\right)^{1-|x|}\mathbf{1}_{\{-1,0,1\}}(x),$$

dove θ, con $0 < \theta < 1/\sqrt[3]{2}$, è un parametro incognito.

1. Determinate gli stimatori di massima verosimiglianza $\widehat{\theta}$ di θ e $\widehat{\kappa}$ di $\kappa = 2\theta^3$.
2. Determinate $\mathrm{E}(|X_1|)$, $\mathrm{Var}(|X_1|)$ e $\mathrm{E}(\widehat{\kappa})$, $\mathrm{Var}(\widehat{\kappa})$.
3. Discutete qualche proprietà (esatta e asintotica) dello stimatore $\widehat{\kappa}$.
4. Costruite un intervallo di confidenza bilatero asintotico per κ di livello $\gamma = 0.95$, per un campione formato da 30 osservazioni uguali a -1, 55 uguali a 0 e 15 uguali a 1. Quindi deducetene uno (bilatero, asintotico, di livello $\gamma = 0.95$) per θ.

Soluzione

1. La verosimiglianza del campione è

$$L_\theta(x_1, \ldots x_n) = \theta^{3\sum_{j=1}^n |x_j|}\left(1-2\theta^3\right)^{n-\sum_{j=1}^n |x_j|}$$

da cui deriviamo che

$$\frac{\partial}{\partial\theta}\ln L_\theta(x_1, \ldots x_n) = \frac{3n}{\theta(1-2\theta^3)}\left(\frac{\sum_{j=1}^n |x_j|}{n} - 2\theta^3\right) \tag{4}$$

e quindi $\dfrac{\partial}{\partial\theta}\ln L_\theta(x_1, \ldots x_n) \geq 0$ se e solo se $\dfrac{\sum_{j=1}^n |x_j|}{n} \geq 2\theta^3$. Poiché $\sqrt[3]{\dfrac{\sum_{j=1}^n |X_j|}{2n}} \in (0, 1/\sqrt[3]{2})$, troviamo che gli stimatori ML di θ e κ sono dati da

$$\widehat{\theta} = \sqrt[3]{\frac{\sum_{j=1}^n |X_j|}{2n}} \qquad\qquad \widehat{\kappa} = \frac{\sum_{j=1}^n |X_j|}{n}\ .$$

Osserviamo che nei due casi estremi in cui tutte le osservazioni siano nulle ($x_j = 0,\ \forall j = 1, \ldots, n$), oppure tutte le osservazioni siano in modulo uguali a 1 ($|x_j| = 1,\ \forall j = 1, \ldots, n$), allora non esistono stimatori ML né di θ né di κ.

2. La variabile aleatoria $|X_1|$ assume solo i valori 0 o 1 e $P(|X_1| = 1) = P(X_1 = -1) + P(X_1 = 1) = 2\theta^3$, cioè X_1 ha densità di Bernoulli di parametro $\kappa = 2\theta^3$. Pertanto,

$$\mathrm{E}(|X_1|) = \kappa, \quad \mathrm{Var}(|X_1|) = \kappa(1-\kappa), \quad \mathrm{E}(\widehat{\kappa}) = \kappa, \quad \mathrm{Var}(\widehat{\kappa}) = \frac{\kappa(1-\kappa)}{n}\ .$$

3. Poiché $\mathrm{E}(\widehat{\kappa}) = \kappa$ e $\lim_{n\to\infty}\mathrm{Var}(\widehat{\kappa}) = \lim_{n\to\infty}\kappa(1-\kappa)/n = 0$, allora $\widehat{\kappa}$ è stimatore di κ non distorto e consistente in media quadratica. Inoltre, data la rappresentazione di $\dfrac{\partial}{\partial\theta}\ln L_\theta(x_1, \ldots x_n)$ in (4), segue che $\widehat{\kappa}$ è anche stimatore efficiente di κ, cioè la sua varianza raggiunge il confine di Cramer-Rao. Infine, per il teorema centrale del limite, $\sqrt{n}(\widehat{\kappa}-\kappa)/\sqrt{\kappa(1-\kappa)}$ ha funzione di ripartizione asintotica gaussiana standard.

4. Un intervallo di confidenza bilatero asintotico per κ è dato da

$$IC(\kappa) = \widehat{\kappa} \pm z_{\frac{1+\gamma}{2}}\sqrt{\frac{\widehat{\kappa}(1-\widehat{\kappa})}{n}}\ ,$$

dove z_a è il quantile di ordine a della fdr gaussiana standard. Se abbiamo 30 osservazioni uguali a -1, 55 uguali a 0 e 15 uguali a 1, allora la stima di κ vale $\widehat{\kappa} = (30+15)/(30+55+15) = 0.45$ e la realizzazione numerica dell'$IC(\kappa)$ asintotico di livello 0.95 proposto qualche riga prima è

$$IC(\kappa) = 0.45 \pm 1.96\sqrt{\frac{0.45\times 0.55}{100}} \simeq (0.352, 0.548).$$

Deduciamo che

$$\left(\sqrt[3]{\frac{0.352}{2}}, \sqrt[3]{\frac{0.548}{2}}\right) \simeq (0.560, 0.650)$$

è un IC (bilatero, asintotico, di livello $\gamma = 0.95$) per θ. ∎

Esercizio 5.3 Sia U una variabile aleatoria esponenziale di media $\lambda > 0$.

1. Determinate la funzione di ripartizione di $T = U^2$.
2. Determinate λ tale che la media di T sia pari a 128.

Vengono registrati i tempi di guasto, espressi in centinaia di ore, di 6 recipienti a pressione, ottenendo i seguenti risultati:

$$8.821 \quad 8.585 \quad 38.938 \quad 6.708 \quad 65.286 \quad 17.808.$$

3. Determinate la funzione di ripartizione empirica $\hat{F}_6$ associata al campione dei 6 recipienti a pressione.
4. Determinate una stima della probabilità che un recipiente a pressione non si guasti prima di 890 ore.
5. Usate un opportuno test di ipotesi per stabilire se il campione casuale dei tempi di guasto dei 6 recipienti a pressione provenga dalla distribuzione di T ottenuta al punto 1. con λ pari al valore determinato al punto 2.

SOLUZIONE Innanzitutto osserviamo che la variabile aleatoria U ha funzione di densità di probabilità esponenziale di parametro λ, cioè: $f_U(u,\lambda) = (1/\lambda)\mathrm{e}^{-u/\lambda}\mathbf{1}_{(0,\infty)}(u)$ e fdr $F_U(u,\lambda) = \left(1 - \mathrm{e}^{-u/\lambda}\right)\mathbf{1}_{(0,\infty)}(u)$.

1. Se $t \le 0$, allora $F_T(t,\lambda) = 0$; se, invece, $t > 0$, allora

$$F_T(t,\lambda) = P_\lambda(T \le t) = P_\lambda(U \le \sqrt{t}) = F_U(\sqrt{t},\lambda) = 1 - \mathrm{e}^{-\frac{\sqrt{t}}{\lambda}}. \tag{5}$$

2. Vale che $\mathrm{E}(T) = \mathrm{E}(U^2) = \mathrm{Var}(U) + [\mathrm{E}(U)]^2 = \lambda^2 + \lambda^2 = 2\lambda^2$. Pertanto $\mathrm{E}(T) = 128$ se e solo se $\lambda = 8$.

3. Ordiniamo le osservazioni in ordine crescente: 6.708, 8.585, 8.821, 17.808, 38.938, 65.286. Quindi, la fdr empirica $\hat{F}_6$ associata al campione dei 6 recipienti a pressione è

$$\hat{F}_6(t) = \begin{cases} 0 & t < 6.708 \\ \frac{1}{6} & 6.708 \le t < 8.585 \\ \frac{1}{3} & 8.585 \le t < 8.821 \\ \frac{1}{2} & 8.821 \le t < 17.808 \\ \frac{2}{3} & 17.808 \le t < 38.938 \\ \frac{5}{6} & 38.938 \le t < 65.286 \\ 1 & t \ge 65.286 \end{cases}.$$

4. Poiché i tempi del campione sono espressi in centinaia di ore, dobbiamo stimare la caratteristica $\kappa = P(T \ge 8.9) = 1 - \lim_{t \to 8.9^-} F(t)$. Usando $\hat{F}_6$, otteniamo $\widehat{\kappa} = 1 - \hat{F}_6(8.9) = 0.5$.

5. Abbiamo un numero "piccolo" di dati (6) non raggruppati e il campione proviene da una fdr continua. Impostiamo il test di Kolmogorov-Smirnov per verificare: $H_0 : X \sim F_0$ contro l'alternativa $H_1 : F \not\sim F_0$, dove F_0 è la fdr determinata in (5) con $\lambda = 8$. Pertanto:

$F_0(6.708)$	$F_0(8.585)$	$F_0(8.821)$	$F_0(17.808)$	$F_0(38.938)$	$F_0(65.286)$
$1 - \mathrm{e}^{-\frac{\sqrt{6.708}}{8}} \simeq 0.277$	0.307	0.310	0.410	0.542	0.636

.

Rifiutiamo H_0 al livello α se $D_6 := \sup_{x\in\mathbb{R}} |\hat{F}_6(x) - F_0(x)| > q_{D_6}(1-\alpha)$. Nel nostro caso abbiamo

$$\sup_{x\in\mathbb{R}} |\hat{F}_6(x) - F_0(x)| = \hat{F}_6(65.286) - F_0(65.286) = 0.364$$

e dalle tavole dei quantili della statistica di Kolmogorov-Smirnov, con $n = 6$, scopriamo che $q_{D_6}(1-0.2) = 0.4104$: ma $P_0(D_6 > 0.364) > P_0(D_6 > 0.4104)$; quindi il p-value del test è maggiore del 20%: c'è evidenza empirica ad accettare H_0.
Alternativamente, se fissiamo per esempio $\alpha = 5\%$, $d_6 = 0.364 < q_{D_6}(1-0.05) = 0.5193$ e dunque non rifiutiamo H_0 a livello del 5%. ∎