Temi d'esame di STATISTICA dell'AA 2003/2004 per allievi ING INF [2L]. Proff. A. Barchielli, I. Epifani

1 STATISTICA per ING INF [2L] Proff. A. Barchielli, I. Epifani 29.06.04

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

Esercizio 1.1 Un informatico smemorato ha scritto un codice per la generazione di numeri casuali dalla densità continua

$$f(x,\theta) = \begin{cases} \theta(\theta+1)x^{\theta-1}(1-x) & x \in (0,1) \\ 0 & altrove \end{cases}, \qquad \theta > 0$$

ma non ricorda se il valore di θ sia 1 o 10. Pigro come pochi, genera un solo numero casuale x_1 e vi chiede di aiutarlo a decidere fra l'ipotesi nulla $H_0 : \theta = 1$ e l'ipotesi alternativa $H_1 : \theta = 10$.

1. Costruite un test uniformemente più potente di livello α per il precedente problema di ipotesi, sulla base di una sola osservazione x_1. Fornite esplicitamente la regione critica del test.

2. Se $\alpha = 2.5\%$ e $x_1 = 0.88$ cosa deciderà l'informatico sulla base del test costruito al punto 1.?

3. Sia $\alpha = 2.5\%$. Calcolate la probabilità di errore di secondo tipo (β) del test costruito al punto 1.

SOLUZIONE

1. Dal Lemma di Neyman Pearson:

$$\begin{aligned} \mathcal{G} &= \left\{ x_1 \in (0,1) : \frac{L_1(x_1)}{L_{10}(x_1)} \le \delta \right\} = \left\{ x_1 \in (0,1) : \frac{1 \cdot 2 \cdot x_1^{1-1}(1-x_1)}{10 \cdot 11 \cdot x_1^{10-1}(1-x_1)} \le \delta \right\} \\ &= \left\{ x_1 \in (0,1) : x_1^9 \ge \frac{1}{55\delta} \right\} = \left\{ x_1 \in (0,1) : x_1 \ge \left(\frac{1}{55\delta}\right)^{1/9} \right\} = \{ x_1 \in (0,1) : x_1 \ge 1 - \sqrt{\alpha} \} \end{aligned}$$

poiché δ è tale che

$$P_1(\mathcal{G}) = P_1\left(X_1 \ge \left(\frac{1}{55\delta}\right)^{1/9} \right) = \int_{(\frac{1}{55\delta})^{1/9}}^{1} f(x,1)\, dx = \int_{(\frac{1}{55\delta})^{1/9}}^{1} 2(1-x)\, dx = \left(1 - \left(\frac{1}{55\delta}\right)^{1/9}\right)^2 = \alpha$$

In definitiva: rifiuto H_0 a livello $\alpha \in (0,1)$ se $x_1 \ge 1 - \sqrt{\alpha}$.

2. Con $\alpha = 2.5\%$, rifiuto H_0 se $x_1 \ge 1 - \sqrt{2.5/100} \simeq 0.8419$: poiché $x_1 = 0.88 > 0.8419$ rifiuto H_0.

3. $\beta = P_{10}(\mathcal{G}^c) = P_{10}(X_1 < 1 - \sqrt{\alpha}) = \int_0^{1-\sqrt{\alpha}} 110x^9(1-x)\, dx = \int_0^{1-\sqrt{\alpha}} 110(x^9 - x^{10})\, dx = 11(1-\sqrt{\alpha})^{10} - 10(1-\sqrt{\alpha})^{11} = (1-\sqrt{\alpha})^{10}(1+10\sqrt{\alpha}) \simeq 0.4617$ ∎

Esercizio 1.2 Sia $X_1, \ldots, X_n$ un campione casuale estratto della funzione di densità discreta

$$p(x,\theta) = \begin{cases} \frac{\mathrm{e}^{-\theta^2}\theta^{2x}}{x!} & x = 0, 1, \ldots \\ 0 & altrove \end{cases} \qquad \theta \in \mathbb{R}$$

1. Determinate uno stimatore della caratteristica $\kappa(\theta) = \theta^2$ usando il metodo dei momenti.

Sia $\widehat{\kappa}$ lo stimatore individuato al punto 1.

2. Verificate se $\widehat{\kappa}$ è stimatore non distorto per $\kappa(\theta)$. Giustificate rigorosamente la risposta.
3. Calcolate la funzione di verosimiglianza del campione: $L_\theta(x_1, \ldots, x_n)$ e $\frac{\partial \log L_\theta(x_1,\ldots,x_n)}{\partial\theta}$.
4. Verificate se la varianza di $\widehat{\kappa}$ raggiunge il confine di Cramer-Rao.

SOLUZIONE

1. $p(x,\theta)$ è una densità di Poisson di parametro θ^2. $\kappa(\theta) = \theta^2$ è la media di questa densità e quindi $\widehat{\kappa} = \bar{X}$.
2. $\widehat{\kappa}$ è stimatore non distorto per θ^2 in quanto: $\mathrm{E}_\theta(\widehat{\kappa}) = \mathrm{E}_\theta(\bar{X}) = \mathrm{E}_\theta(X_1) = \theta^2 = \kappa(\theta)\ \forall\theta \in \mathbb{R}$.

3. $$L_\theta(x_1, \ldots, x_n) = \prod_{j=1}^{n} \frac{\mathrm{e}^{-\theta^2}\theta^{2x_j}}{x_j!} = \frac{\mathrm{e}^{-n\theta^2}\theta^{2\sum_{j=1}^n x_j}}{\prod_{j=1}^n x_j!}$$
$$\frac{\partial \log L_\theta(x_1, \ldots, x_n)}{\partial\theta} = \frac{\partial \log(\mathrm{e}^{-n\theta^2}\theta^{2\sum_{j=1}^n x_j})}{\partial\theta} = -2n\theta + 2\frac{\sum_{j=1}^n x_j}{\theta} = \frac{2n}{\theta}(\bar{x} - \theta^2)$$

4. Osserviamo che $\widehat{\kappa}$ è stimatore non distorto per $\kappa(\theta)$. Inoltre, deduciamo dal punto 3. che $\frac{\partial \log L_\theta(x_1,\ldots,x_n)}{\partial\theta} = \frac{2n}{\theta}(\bar{x} - \theta^2) = \frac{2n}{\theta}(\widehat{\kappa} - \kappa(\theta))$ per ogni realizzazione campionaria $(x_1, \ldots, x_n)$. Pertanto, scelta $a(\theta, n) = \frac{2n}{\theta}$, abbiamo:

$$P_\theta\left(\frac{\partial \log L_\theta(X_1, \ldots, X_n)}{\partial\theta} = a(\theta, n)(\widehat{\kappa} - \kappa(\theta))\right) = 1$$

Ma l'ultima è condizione necessaria e sufficiente affinché $\mathrm{Var}(\widehat{\kappa}) = \frac{[\kappa'(\theta)]^2}{nI(\theta)}$ dove $I(\theta)$ = Informazione di Fisher del modello.
Oppure:

$$\mathrm{Var}_\theta(\widehat{\kappa}) = \mathrm{Var}_\theta(\bar{X}) = \frac{\mathrm{Var}_\theta(X_1)}{n} = \frac{\theta^2}{n}$$
$$nI(\theta) = E_\theta\left(\frac{2n}{\theta}(\bar{X} - \theta^2)\right)^2 = \frac{4n^2}{\theta^2}\mathrm{Var}_\theta(\bar{X}) = 4n$$
$$[\kappa(\theta)']^2 = [2\theta]^2 = 4\theta^2$$
$$\text{Confine di Cramer Rao} = \frac{[\kappa(\theta)']^2}{nI(\theta)} = \frac{\theta^2}{n} = \mathrm{Var}_\theta(\hat{\kappa}) \qquad \blacksquare$$

Esercizio 1.3 Si pensa che il numero X degli arrivi giornalieri in una piccola località di montagna dal 16 giugno al 15 settembre possa essere modellata come una variabile aleatoria discreta con densità:

$$p_X(k = \begin{cases} 0.3 \cdot 0.7^k & k = 0, 1, 2, \dots \\ 0 & altrove \end{cases}$$

1. Verificate questa ipotesi sulla base del numero di arrivi (giornalieri e indipendenti) registrati dal 16 giugno al 15 settembre 2003 (92 giorni) e sintetizzati nella seguente tabella:

# di arrivi giornalieri $k =$	0	1	2	3	4	5	6 o più
# di giorni con k arrivi =	24	15	19	12	10	2	10

SOLUZIONE Sia $\theta_{0k} := P(\text{“}k \text{ arrivi in un giorno”}) = 0.3 \cdot 0.7^k, k = 0, 1, \dots$. Allora:

$k =$	0	1	2	3	4	5	6 o più
# di giorni con k arrivi=	24	15	19	12	10	2	10
$\theta_{0k} =$	0.3	0.21	0.147	0.1029	0.07203	0.050421	0.117649
$92 * \theta_{0k} =$	27.6	19.32	13.524	9.4668	6.62676	4.638732	10.82371

Accorpiamo le modalità 4 e 5 in una unica classe in quanto $92 * \theta_{05} < 5$ e l'approssimazione asintotica χ^2 con le 7 classi date nel testo non funziona. Chiamiamo A_k le nuove classi e sia $N_k = \#$ di giorni caratterizzati da un numero di arrivi $\in A_k$. Per le nuove classi abbiamo:

$A_k =$	$\{0\}$	$\{1\}$	$\{2\}$	$\{3\}$	$\{4, 5\}$	$\{6, 7, \dots\}$
$N_k =$	24	15	19	12	10+2 =12	10
$\theta_{0k} =$	0.3	0.21	0.147	0.1029	0.07203 + 0.050421	0.117649
$92 * \theta_{0k} =$	27.6	19.32	13.524	9.4668	6.62676 +4.638732 =11.26549	10.82371

$$Q_{92} = \sum_{k=1}^{6} \frac{(N_k - 92 \cdot \theta_{0k})^2}{92 \cdot \theta_{0k}} = \sum_{k=1}^{6} \frac{N_k^2}{92 \cdot \theta_{0k}} - 92 = 4.4412$$

Approssimativamente $Q_{92} \sim \chi^2_5$ e il p-value è pari a $P(Q_{92} > 4.4412) = 1 - F_{\chi^2_5}(4.4412) \simeq 1 - 0.5122 = 0.4878$: c'è una forte evidenza ad accettare l'ipotesi $p_X(x) = \begin{cases} 0.3 \cdot 0.7^k & k = 0, 1, 2, \dots \\ 0 & altrove \end{cases}$ ∎

Esercizio 1.4 Una compagnia di assicurazioni deve eseguire uno studio per stimare gli indennizzi pagati a seguito di incidenti automobilistici senza lesioni alle persone. Da studi precedenti è emerso che si può assumere che tali importi abbiano densità gaussiana con media μ incognita e deviazione standard nota e pari a 900 euro. Su un nuovo campione casuale di 100 incidenti del suddetto tipo è stato osservato un indennizzo medio pari a 5562 euro.

1. Determinate un intervallo di confidenza di livello 94% per il parametro μ.
2. Verificate l'ipotesi $H_0 : \mu = 5500$ contro $H_1 : \mu \neq 5500$ al livello $\alpha = 6\%$.
3. In realtà la compagnia assicurativa ritiene che l'intervallo di confidenza costruito al punto 1. non sia sufficientemente preciso. Decide quindi di condurre uno studio più vasto, cioè su un campione più numeroso. Determinate il numero minimo di casi da esaminare affinchè la lunghezza dell'intervallo di confidenza per μ non superi i 300 euro.

SOLUZIONE

1. $\bar{x} \mp z_{(1+\gamma)/2} \frac{\sigma}{\sqrt{n}} = 5562 \mp z_{0.97} \frac{900}{\sqrt{100}} \simeq 5562 \mp 1.88 \frac{900}{\sqrt{100}} = (5392.72, 5731.27)$.
2. $5500 \in (5392.72, 5731.27)$. Quindi a livello 6% accetto H_0
3. $\mathscr{L} = 2 \frac{\sigma}{\sqrt{n}} z_{\frac{1+\gamma}{2}} = 2 \frac{900}{\sqrt{n}} z_{0.97} \leq 300$ se e solo se $\sqrt{n} \geq 2 \frac{900}{300} z_{0.97} \simeq 11.285$ se e solo se $n \geq 11.285^2 \simeq 127.3512$. Segue che il numero minimo di casi da esaminare è $n^* = 128$. ∎

Esercizio 1.5 (Sezione Epifani) Si consideri la sequenza di numeri:

$$0.1421,\ 0.0519,\ 0.1049,\ 0.8168,\ 0.1921,\ 0.1019,\ 0.1549,\ 0.8668,\ 0.0521,\ -0.0380,\ 0.0149,\ 0.7268$$

Impostare un opportuno test per decidere, a livello di significatività $\alpha = 10\%$, se tale sequenza provenga da un campione casuale.

SOLUZIONE Contiamo il numero di concordanze e discordanze nel campione dato:

i	$C - D =$
0.1421	$+5-6$
0.0519	$+8-2$
0.1049	$+5-4$
0.8168	$+1-7$
0.1921	$+2-5$
0.1019	$+3-3$
0.1549	$+2-3$
0.8668	$+0-4$
0.0521	$+1-2$
-0.0380	$+2-0$
0.0149	$+1-0$
0.7268	$\boxed{30-36=-6=T}$

Impostiamo quindi un test di aleatorietà di Kendall a livello $\alpha = 10\%$ per l'ipotesi $H_0 : X_1, \ldots, X_n \ iid \sim F$. Sia $q_{C-D}(p)$ il quantile di ordine p della stastistica $T = C - D$. Per $n = 12$, $q_{C-D}(1 - 0.1) = 18$ e $q_{C-D}(1-0.1/2) = q_{C-D}(0.95) = 24$. Essendo $T = -6$, allora sia $T > -18$ sia $|T| < 24$. Quindi NON rifiutiamo l'ipotesi H_0 di dati indipendenti a livello $\alpha = 10\%$ sia contro l'alternativa unilatera: H_1:"c'è trend negativo", sia contro l'alternativa bilatera: H_1: "i dati non sono indipendenti". ∎

Esercizio 1.6 (Sezione Barchielli) Si vuole stabilire se il numero X di ore di lavoro sia collegato al numero Y di sigarette fumate durante l'orario di lavoro. A questo scopo si estrae un campione casuale da (X, Y) ottenendo:

$$(8.2, 3), \qquad (10, 4), \qquad (10.5, 7), \qquad (8, 5), \qquad (9.5, 6)$$

Effettuare un opportuno test di livello 5% per verificare se si possa ritenere che all'aumentare del numero di ore lavorate aumenti anche il numero di sigarette.

SOLUZIONE Introduciamo il test di concordanza/discordanza di Kendall per il seguente problema:

$$H_0 : \tau = 0 \ vs \ H_1 : \tau > 0$$

Per eseguire il test è necessario calcolare il numero di concordanze e discordanze. A questo scopo riordiniamo le coppie per valori di X crescenti:

$$(8, 5), \qquad (8.2, 3), \qquad (9.5, 6), \qquad (10, 4), \qquad (10.5, 7)$$

Costruiamo ora la tabella delle concordanze/discordanze:

$segno[y_{[j]} - y_{[1]}]_{j>1}$	-1	+1	-1	+1
$segno[y_{[j]} - y_{[2]}]_{j>2}$		+1	+1	+1
$segno[y_{[j]} - y_{[3]}]_{j>3}$			-1	+1
$segno[y_{[j]} - y_{[4]}]_{j>4}$				+1

Dalla tabella risulta $C = 7$ e $D = 3$, dunque $C - D = 4$. La regione critica del test è costituita dai campioni per cui $C - D > q_{0.95}(C - D)$, dove $q_{0.95}(C - D) = 6$ rappresenta il quantile di ordine 0.95 della statistica $C - D$. Concludiamo che non vi è evidenza statistica per rifiutare l'ipotesi nulla, quindi non possiamo affermare che vi sia una concordanza tra ore lavorate e sigarette fumate. ∎

2 STATISTICA per ING INF [2L] Proff. A. Barchielli, I. Epifani 15.07.04

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

Esercizio 2.1 La ditta Baltic Sea produce macchine per l'inscatolamento di caviale. A causa delle fluttuazioni casuali la quantità di caviale dosata dalla macchine è una variabile aleatoria X gaussiana con media nota $\mu = 30$ grammi e varianza incognita σ^2. Prima di procedere all'acquisto di una di queste macchine controllo le quantità x_i (misurate in grammi) di caviale contenute in un campione casuale di 100 scatolette ottenendo:

$$\sum_{i=1}^{100} x_i = 3006 \qquad\qquad \sum_{i=1}^{100} x_i^2 = 90711$$

Inoltre, sono disposto a commettere un errore di prima specie di probabilità al più pari ad α di acquistare una macchina che abbia un'imprecisione (=deviazione standard di X) effettiva maggiore o uguale di 2 grammi.

1. Fornite una stima puntuale della varianza σ^2.
2. Impostate un opportuno test sulla varianza, specificando: ipotesi nulla, ipotesi alternativa e una regione critica di ampiezza α.
3. Quale decisione prendete a livello $\alpha = 2.5\%$?

Soluzione

1. Essendo la media nota, stimo σ^2 con $S_0^2 = \frac{1}{100}\sum_{i=1}^{100}(X_i - 30)^2$ che ha valore $s_0^2 = \frac{\sum_{j=1}^{100} x_j^2}{100} + 30^2 - 2 * 30\frac{\sum_{j=1}^{100} x_j}{100} = 3.51$

2. Deduciamo dal testo di dover impostare un test di verifica dell'ipotesi nulla $H_0 : \sigma^2 \geq 4$ contro l'alternativa $H_1 : \sigma^2 < 4$, nel caso di un campione casuale estratto dalla popolazione $\mathcal{N}(30, \sigma^2)$. Una regione critica è $G = \left\{(x_1, \ldots, x_{100}) \in \mathbb{R}^{100} : \frac{100 * s_0^2}{4} \leq \chi^2_{100}(\alpha)\right\}$

3. Il valore della statistica è $\frac{100 * s^2}{4} = 87.75$ e il p-value è $P(\chi^2_{100} < 87.75)$. Asintoticamente χ^2_{100} ha fdr gaussiana di media 100 e varianza 200 e quindi, approssimativamente, p-value$\simeq \Phi\left(\frac{87.75-100}{\sqrt{200}}\right) = \Phi(-0.866) = 1 - \Phi(0.866) \simeq 0.193$. Essendo $2.5\% < 19.3\%$, (purtroppo) accetto H_0 a livello 2.5%. ∎

Esercizio 2.2 Il tempo di esecuzione del programma xxx sul calcolatore yyy è compreso fra 60 e 120 minuti primi. Idealmente, esso può essere modellato come una va X assolutamente continua con densità

$$f(x,\theta) = \begin{cases} \frac{\theta}{60^\theta}(x-60)^{\theta-1} & 60 \le x \le 120 \\ 0 & \text{altrove} \end{cases}$$

e $\theta > 0$.

1. Determinate in funzione di θ la caratteristica $\kappa(\theta)$ = "Probabilità che il calcolatore impieghi più di 90 minuti per eseguire il programma".

Su ciascuno di n calcolatori tutti del tipo yyy (e che lavorano indipendentemente uno dall'altro) lanciamo il programma xxx e, allo scadere dei 90 minuti, controlliamo se il programma è stato eseguito o no. Sia Y_i la va che vale 1 se il programma lanciato sull'i-esimo calcolatore è eseguito (nei 90 minuti) e vale 0 se non lo è, per $i = 1, \ldots, n$.

2. Quanto vale $P_\theta(Y_i = 1)$, $i = 1, \ldots, n$?

3. Verificate che $\widehat{\kappa} = 1 - \bar{Y}$ è lo stimatore di massima verosimiglianza di $\kappa(\theta)$ basato sul campione casuale $Y_1, \ldots, Y_n$.

4. (a) Verificate se $\widehat{\kappa}$ è stimatore non distorto e consistente in media quadratica per $\kappa(\theta)$ e (b) determinate la funzione di ripartizione asintotica di $\widehat{\kappa}$. Giustificate adeguatamente le risposte.

5. Alla luce di quanto ottenuto ai punti 1., 2. e 3., come stimereste θ sapendo soltanto che su $n = 15$ programmi lanciati 10 non sono stati eseguiti nei primi 90 minuti?

SOLUZIONE

1. $\kappa(\theta) = P_\theta(X > 90) = \int_{90}^{120} f(x,\theta)\, dx = \int_{90}^{120} \frac{\theta}{60^\theta}(x-60)^{\theta-1}\, dx = \frac{(x-60)^\theta}{60^\theta} \,|_{90}^{120} = 1 - 0.5^\theta$.

2. Sia X_i la va che modella il tempo di esecuzione del programma xxx lanciato sull'i-esimo calcolatore del tipo yyy. Allora $X_i \sim f(x,\theta)$ e
$P_\theta(Y_i = 1) = P_\theta(\text{"}i\text{-esimo programma è eseguito in 90 minuti"}) = P_\theta(X_i \le 90) = 1 - \kappa(\theta) = 0.5^\theta$.

3. $Y_1, \ldots, Y_n$ i.i.d. $\sim \mathbf{Be}(1 - \kappa(\theta))$ e $0 < 1 - \kappa(\theta) = 0.5^\theta < 1$, $\forall \theta > 0$. Chiamiamo $\kappa = \kappa(\theta)$ e studiamo la funzione di verosimiglianza del campione $Y_1, \ldots, Y_n$ in corrispondenza della realizzazione $y_1, \ldots, y_n$ in funzione di κ

$$L_\kappa(y_1, \ldots, y_n) = \prod_{i=1}^{n} (1-\kappa)^{y_i} \kappa^{(1-y_i)} = (1-\kappa)^{n\bar{y}} \kappa^{(n-n\bar{y})} \quad \kappa \in (0,1)$$

$$\frac{\partial \log(L_\kappa(y_1, \ldots, y_n))}{\partial \kappa} = \frac{\partial [n\bar{y}\log(1-\kappa) + n(1-\bar{y})\log(\kappa)]}{\partial \kappa} = \frac{n(1-\bar{y}) - n\kappa}{\kappa(1-\kappa)} \ge 0$$

se e solo se $\kappa \le (1 - \bar{y})$ cosicché $\widehat{\kappa} = 1 - \bar{Y}$ è MLE per κ. Esso esiste sempre a meno che non si verifichino i due casi estremi: tutte n le volte il programma è stato eseguito in meno di 90 minuti e quindi $\bar{y} = 1$ o tutte n le volte è stato eseguito in più di 90 minuti e quindi $\bar{y} = 0$. Infatti 0,1 non appartengono allo spazio in cui varia 0.5^θ, poiché $\theta > 0$.

4. $\widehat{\kappa} = \frac{\sum_{i=1}^n (1-Y_i)}{n}$ è la media campionaria delle variabili $1 - Y_1, \ldots, 1 - Y_n$ che hanno media $\mathrm{E}_\theta(1 - Y_1) = P_\theta(Y_1 = 0) = \kappa(\theta)$ e varianza $\mathrm{Var}_\theta(Y_1) = P_\theta(Y_1 = 0)(1 - P_\theta(Y_1 = 0)) = \kappa(\theta)(1 - \kappa(\theta)) \in (0, \infty)$. Quindi: $\mathrm{E}_\theta(\widehat{\kappa}) = E_\theta(1 - Y_i) = \kappa(\theta)$ $\forall \theta > 0$ e $\widehat{\kappa}$ è stimatore non distorto di $\kappa(\theta)$; inoltre: $\mathrm{Var}_\theta(\widehat{\kappa}) = \kappa(\theta)(1-\kappa(\theta))/n \to 0$ per $n \to \infty$, $\forall \theta > 0$: $\widehat{\kappa}$ è stimatore consistente in media quadratica per $\kappa(\theta)$. Infine, applicando il Teorema Centrale del Limite deduciamo la gaussianità asintotica di $\widehat{\kappa}$ nel senso che

$$\lim_{n\to\infty} P\left(\sqrt{n} \frac{\widehat{\kappa} - \kappa(\theta)}{\sqrt{\kappa(\theta)(1-\kappa(\theta))}} \le z \right) = \Phi(z) \qquad \forall z \in \mathbb{R}$$

5. $\widehat{\kappa} = 10/15 = 2/3$ e θ in termini del parametro $\kappa := P_\kappa(X > 90)$ è $\theta = \dfrac{\log(1-\kappa)}{\log(1/2)} = \log_{0.5}(1-\kappa)$. Quindi lo stimatore MLE di θ sulla base del campione $Y_1, \ldots, Y_{15}$ è $\hat{\theta} = \log_{0.5}(1/3) \simeq 1.585$. ∎

Esercizio 2.3 Le richieste di interventi che arrivano al pronto soccorso di una certa località balneare sono indipendenti. Si pensa che mediamente ci sia una richiesta ogni 3 ore e che il tempo (misurato in ore) intercorrente fra due richieste successive sia una va assolutamente continua con densità esponenziale. Per verificare quest'ipotesi statistica sono stati analizzati i tabulati delle richieste di intervento di domenica 11 luglio 2004. A partire da mezzanotte, sono arrivate in totale 6 richieste alle seguenti ore:

$$04:06:00,\ \ 06:00:00,\ \ 07:52:00,\ \ 13:19:00,\ \ 22:31:00,\ \ 22:46:00$$

1.1 Calcolate il tempo (espresso in ore) trascorso da mezzanotte (00:00:00 del 11/07/04) fino alla prima richiesta e gli altri cinque "intertempi" fra due richieste successive.

1.2 Sia F_0 la funzione di ripartizione di una va X assolutamente continua con densità esponenziale di media 3. Calcolate: $F_0(0.25), F_0(1.87), F_0(1.90), F_0(4.10), F_0(5.45), F_0(9.20)$.

2 Usando il campione degli intertempi ottenuto al punto 1.1 e quanto calcolato al punto 1.2, verificate il modello esponenziale dell'intertempo descritto prima, mediante un opportuno test di ipotesi di livello $\alpha = 5\%$.

SOLUZIONE

1.1

In minuti: 246, 114 ,112, 327, 552 , 15 =

in ore: 246/60, 114/60, 112/60, 327/60, 552/60 , 15/60 = 4.1, 1.9, 1.87, 5.45, 9.2, 0.25.

1.2. Se $x > 0$: $F_0(x) = 1 - \mathrm{e}^{-x/3}$ e quindi

$F_0(0.25)$	$F_0(1.87)$	$F_0(1.90)$	$F_0(4.1)$	$F_0(5.45)$	$F_0(9.20)$
0.07995559	0.4638457	0.46918055	0.74504460	0.83743326	0.95342385

2. Impostiamo il test di Kolmogorov-Smirnov di livello $\alpha = 5\%$ per verificare: $H_0 : X \sim F_0 = Exp(3)$ contro l'alternativa $H_1 : X \not\sim F_0$:

$$D_6 := \sup_{x\in\mathbb{R}} |\hat{F}_6(x) - F(x)| = F_0(1.87) - \hat{F}_6(0.25) = 0.4638437 - 1/6 \simeq 0.2972$$

dove $\hat{F}_6$ = fdr empirica del campione $4.1, 1.9, 1.87, 5.45, 9.2, 0.25$. Dalle tavole dei quantili della statistica di Kolmogorov-Smirnov con $n = 6$, abbiamo che $q_{D_6}(1 - 0.05) \simeq 0.5193$. Essendo $0.2972 < 0.5193$ accetto H_0. Con R:

```
    ks.test(c(4.10, 1.90, 1.87, 5.45, 9.20, 0.25), pgamma, 1,1/3)
   One-sample Kolmogorov-Smirnov test
data:  c(4.1, 1.9, 1.87, 5.45, 9.2, 0.25)
D = 0.2972, p-value = 0.6644
alternative hypothesis:  two.sided
```

Esercizio 2.4 I dati disponibili sul sito del Ministero dell'Istruzione, dell'Università e della Ricerca rivelano che nell'anno solare 2002, i laureati e diplomati presso le facoltà di ingegneria del Politecnico di Milano sono stati 3502 di cui 2967 uomini. Fra le donne, 176 erano fuori corso da un anno, 72 da due anni e 61 da tre anni. Invece, fra gli uomini, i fuori corso da un anno erano 733, quelli da due anni 531 e quelli da tre anni 279. Tutti gli altri erano fuori corso da quattro anni o più; nel 2002 non ci sono stati laureati in corso. Verificate sulla base di questi dati se uomini e donne impiegano (più o meno) lo stesso tempo per laurearsi in ingegneria.

SOLUZIONE Organizziamo i dati forniti dal testo nella seguente tabella a doppia entrata:

	fc da 1 anno	fc da 2 anni	fc da 3 anni	fc da 4 anni o più	
M	733	531	279		2967
D	176	72	61		
					3502

Completiamo la tabella con i dati mancanti: ($X = sesso$ e $Y =$ numero degli anni fuori corso al momento della laurea)

$X \setminus Y$	fc da 1 anno	fc da 2 anni	fc da 3 anni	fc da 4 anni o più	$N_X =$
M	$N_{M1} = 733$	$N_{M2} = 531$	$N_{M3} = 279$	$N_{M4} =$**1424**	2967
D	$N_{D1} =$176	$N_{D2} =$72	$N_{D3} =$61	$N_{D4} =$**226**	**535**
$N_Y =$	**909**	**603**	**340**	**1650**	$N =$3502

Impostiamo un test χ^2 di indipendenza fra le variabili categoriche $X = sesso$ e Y=numero di anni fuori corso al momento della laurea. La statistica di Pearson Q ha valore:

$$Q = N \sum_{j=1}^{4} \frac{N_{Mj}^2}{N_{XM} * N_{Yj}} + N \sum_{j=1}^{4} \frac{N_{Dj}^2}{N_{XD} * N_{Yj}} - N \simeq 21.953$$

Asintoticamente $Q \sim \chi^2_{(2-1)(4-1)} = \chi^2_3$ e $P(\chi^2_3 > 21.953) \simeq 6.681253 * 10^{-5} \simeq 0$: c'è una forte evidenza a rifiutare l'ipotesi H_0 : "X, Y sono indipendenti", ossia concludo che c'è dipendenza fra sesso e numero degli anni fuori corso. Avendo a disposizione soltanto la tavola del Pestman dei quantili della χ^2_3, osservo che $\chi_3(0.995) = 12.8 < 21.953$, quindi per qualunque livello $\alpha \geq 1 - 0.995 = 0.005$ rifiuto H_0.

Curiosità: nel 2002 il 32.89% delle donne si sono laureate soltanto un anno fuori corso mentre questa percentuale scende a 24.7% per gli uomini. ∎

Esercizio 2.5 Un laboratorio informatico ha elaborato un nuovo protocollo P_{new} per la trasmissione di dati. Per confrontare P_{new} con il vecchio protocollo P_{old} si procede a inviare un certo file per 7 volte da un server ad un altro usando il protocollo P_{new} e 6 volte usando il protocollo P_{old} e si misurano i tempi (in secondi) intercorrenti tra l'invio e la ricezione. I risultati ottenuti per P_{new} sono:

$$x_i: \quad 1.49, \quad 1.50, \quad 1.96, \quad 2.33, \quad 1.45, \quad 1.71, \quad 2.83$$

e quelli per P_{old} sono

$$y_i: \quad 1.85, \quad 3.47, \quad 4.44, \quad 1.75, \quad 2.16, \quad 3.93$$

Vi si chiede ora di usare questi dati al fine di stabilire se il nuovo protocollo P_{new} sia migliore del vecchio P_{old} dove, in modo naturale, un protocollo è ritenuto "migliore" di un altro se trasferisce i dati in meno tempo. Quindi:

- Costruite una opportuna strategia statistica (che usi i dati precedenti) per affrontare il problema ipotetico del confronto fra i protocolli P_A e P_B. In particolare abbiate cura di specificare *a*) le ipotesi statistiche da verificare, *b*) le regioni critiche e, se necessario, *c*) le condizioni che il modello statistico generatore dei dati deve soddisfare perché la vostra procedura trovi ragionevoli giustificazioni nella teoria dei test (parametrici e/o non parametrici) vista durante il corso.

SOLUZIONE Al laboratorio informatico che ha proposto il nuovo protocollo, piacerebbe dimostrare in modo convincente l'ipotesi che P_{new} sia migliore di P_{old}. Motivati da ciò, procederemo a verificare il seguente problema: H_0 : "P_{new} è non migliore di P_{old}" contro H_1: "P_{new} è migliore di P_{old}"; in questo modo, l'eventuale accettazione di H_1 sarebbe una conclusione forte.

Sol 1: Mi pongo sotto l'ipotesi che $(x_1, \ldots, x_7)$ = realizzazione di $\boldsymbol{X} = X_1, \ldots, X_7$ i.i.d. $\sim N(\mu_X, \sigma_X^2)$, $(y_1, \ldots, y_6)$ = realizzazione di $\boldsymbol{Y} = Y_1, \ldots, Y_6$ i.i.d. $\sim N(\mu_Y, \sigma_Y^2)$ e $\boldsymbol{X}, \boldsymbol{Y}$ indipendenti. Quindi imposto un F-test per il confronto di varianze di due popolazioni gaussiane $H_0 : \sigma_X^2 = \sigma_Y^2$ versus $H_1 : \sigma_X^2 \neq \sigma_Y^2$ di livello $\alpha = 5\%$:

$$\bar{x} = 1.895714 \quad \bar{y} = 2.93$$

$$S_X^2 = 0.2699952 \quad S_Y^2 = 1.344667 \quad \frac{S_X^2}{S_Y^2} = 0.2007897$$

$$F_{6,5}(0.975) = 6.98 \quad F_{6,5}(2.5/100) = 1/F_{5,6}(0.975) = 1/5.99 \simeq 0.167 \text{ e } 0.200787 \in (0.167, 6.98)$$

$\Rightarrow$ accetto l'ipotesi nulla di varianze uguali.

Imposto ora un test t per dati indipendenti e gaussiani per confrontare le medie, tenendo presente che a parità di variabilità, le ipotesi H_0, H_1 si traducono nel seguente modo: $H_0 : \mu_X \geq \mu_Y$ versus $H_1 : \mu_X < \mu_Y$:

$S_p^2 = 0.7584823$, $\dfrac{\bar{x} - \bar{y}}{s_p\sqrt{1/7 + 1/6}} = -2.1415$; gradi di libertà della $t = 7 + 6 - 2 = 11$ e $-t_{11}(1 - 5/100) \simeq -1.7958 > -2.0245$: Accetto l'ipotesi H_1 che P_{new} sia migliore di P_{old}.

Sol. 2 Mi pongo sotto l'ipotesi che $\boldsymbol{X} = X_1, \ldots, X_7$ i.i.d. $\sim F$, $\boldsymbol{Y} = Y_1, \ldots, Y_6$ i.i.d. $\sim G$ con $\boldsymbol{X}, \boldsymbol{Y}$ indipendenti e F e G assolutamente continue. Traduco H_0, H_1 in termini di dominanza stocastica nel seguente modo: $H_0 : F \leq G$ versus $H_1 : F > G$ ed imposto il corrispondente test unilatero non parametrico di Wilcoxon-Mann-Wintney: Ordino dalla più piccola alla più grande le osservazioni del campione riunito:

$$1.45x, 1.49x, 1.50x, 1.71x, 1.75y, 1.85y, 1.96x, 2.16y, 2.33x, 2.83x, 3.47y, 3.93y, 4.44y$$

e calcolo la statistica di Mann-Witney U= "somma dei ranghi di X" $-7 * 8/2 = 36 - 7 * 8/2 = 8$. Il quantile di U corrispondente a $m = 7$ e $n = 6$ di ordine 5% è $q_U(5\%) = 9$: Poiché $8 < 9$, sono nella regione critica ed accetto H_1 a livello 5%. ∎

3 STATISTICA per ING INF [2L] Proff. A. Barchielli, I. Epifani 16.09.04

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

Esercizio 3.1 Il reddito annuale (in opportune unità) X di un individuo di una certa popolazione è una variabile aleatoria assolutamente continua con densità di probabilità

$$f(x,\vartheta) = \frac{\vartheta}{x^{\vartheta+1}}\,\mathbf{1}_{(1,+\infty)}(x) \quad \vartheta > 0$$

Ci proponiamo di trovare una stima intervallare di ϑ, sulla base di un campione casuale X_1, $X_2, \ldots, X_n$ di X.

1. Determinate la densità di probabilità della variabile aleatoria $Y := 2\vartheta \ln X$. Vi riconoscete una "densità notevole"?
2. Determinate lo stimatore di massima verosimiglianza T_n di $\kappa(\vartheta) = 1/\vartheta$.
3. Come è distribuita la variabile aleatoria $Q_n := 2n\vartheta T_n$?
4. Sia $\alpha = 0.1$ e $n = 10$. Trovate (in modo semplice) due numeri q_1 e q_2 tali che $P_\vartheta[q_1 \le Q_n \le q_2] = 1-\alpha$. Quali tabelle sono necessarie per ottenere q_1 e q_2?
5. Fornite un intervallo di confidenza per ϑ, di livello 90% per $n = 10$ e $T_{10} = 22.0$.

Soluzione

1. Poiché $X \ge 1$, allora $Y = 2\vartheta \ln X \ge 0$ e, per $y \ge 0$:

$$F_{Y,\vartheta}(y) := P_\vartheta(Y \le y) = P_\vartheta\left(X \le \exp\left(\frac{y}{2\vartheta}\right)\right) = \int_1^{\exp\left(\frac{y}{2\vartheta}\right)} \frac{\vartheta}{x^{\vartheta+1}}\,\mathrm{d}x$$

da cui segue che la densità di Y è

$$f_{Y,\theta}(y) = \mathbf{1}_{(0,+\infty)}(y)\,\frac{1}{2\vartheta}\exp\left(\frac{y}{2\vartheta}\right)\vartheta\frac{1}{\left(\exp\left(\frac{y}{2\vartheta}\right)\right)^{\vartheta+1}} = 1_{(0,+\infty)}(y)\,\frac{1}{2}\,\exp\left(-\frac{y}{2}\right)$$

cioè $Y \sim \mathcal{E}(2) = \Gamma\,(2/2,\ 2) = \chi^2_2$

2. $\ln L_\vartheta(x_1,\ldots,x_n) = n\ln\vartheta - (\vartheta+1)\sum_{i=1}^n \ln x_i$ e $\frac{\partial}{\partial\vartheta}\,\ln L_\vartheta = \frac{n}{\vartheta} - \sum_{i=1}^n \ln x_i = 0$ se e solo se $\vartheta = \frac{n}{\sum_{i=1}^n \ln x_i}$. Inoltre: $\frac{\partial^2}{\partial\vartheta^2}\,\ln L_\vartheta = -\frac{n}{\vartheta^2} < 0\ \forall\vartheta$ e quindi: $T_n = \frac{1}{n}\sum_{i=1}^n \ln X_i$.
3. $Q_n = 2\vartheta\sum_{i=1}^n \ln X_i = \sum_{i=1}^n Y_i$ con $Y_1,\ldots,Y_n$ i.i.d. $\sim \chi^2_2$ per quanto stabilito al punto 1.. Segue che $Q_n \sim \chi^2_{2n}$.
4. Scelta "a code simmetriche": $q_1 = \chi^2_{2n}(\alpha/2) = \chi^2_{20}(0.05) \simeq 10.9$, $q_2 = \chi^2_{2n}(1-\alpha/2) = \chi^2_{20}(0.95) \simeq 31.4$. (Abbiamo usato la tabella della fdr χ^2_{2n}).
5. Poiché $\{q_1 \le Q_n \le q_2\} = \left\{\frac{q_1}{2nT_n} \le \vartheta \le \frac{q_2}{2nT_n}\right\}$, allora

$$P_\vartheta\left(\frac{\chi^2_{2n}(\alpha/2)}{2nT_n} < \vartheta < \frac{\chi^2_{2n}(1-\alpha/2)}{2nT_n}\right) = 1-\alpha$$

e, l'intervallo di confidenza cercato è $\left(\frac{10.9}{20\times 22}, \frac{31.4}{20\times 22}\right) \simeq (0.025, 0.071)$ ∎

Esercizio 3.2 Vogliamo fare inferenza sulla proporzione ϑ degli individui di tipo A presenti in una certa popolazione. Perciò procediamo con il seguente esperimento: effettuiamo n estrazioni casuali con reimmissione da questa popolazione e registriamo il numero di individui di tipo A ottenuti nelle n estrazioni. Sia X_n la variabile aleatoria definita da "numero di individui di tipo A ottenuti nelle n estrazioni".

1. Qual è la densità, la media e la varianza di X_n?

In particolare, siamo interessati a verificare l'ipotesi $H_0 : \vartheta = \vartheta_0$ contro $H_1 : \vartheta = \vartheta_1$ con $\vartheta_0 < \vartheta_1$ ($\vartheta_0, \vartheta_1 \in (0,1)$).

2. Verificate che la regione critica per il test determinato dal lemma di Neyman-Pearson e basato su X_n abbia la forma: $G = \left\{ \frac{X_n}{n} \geq t \right\}$.

Fissiamo ora i seguenti valori: $\vartheta_0 = 0.5$, $\vartheta_1 = 0.7$ e $\alpha = 5\%$.

3. Se $n = 30$, qual è approssimativamente la funzione di ripartizione di $\dfrac{X_n}{n}$ sotto H_0? E sotto H_1?

4. Assumete $n = 30$ e determinate t tale che "approssimativamente" la regione critica G abbia ampiezza 5%. Se $X_n = 18$, che decisione prendete?

5. Assumete $n = 30$ e calcolate "approssimativamente" la probabilità d'errore di seconda specie β.

Soluzione

1. $X_n \sim \mathbf{Bin}(n, \vartheta)$ e quindi $\mathrm{E}(X_n) = n\vartheta$ e $\mathrm{Var}(X_n) = n\vartheta(1-\vartheta)$.

2. La funzione di verosimiglianza basata sull'unica osservazione X_n è: $L_\vartheta(x) = \binom{n}{x}\vartheta^x(1-\vartheta)^{n-x}$, $\vartheta \in (0,1)$ e la regione critica dettata dal Lemma di Neyman Pearson ha forma: $G = \left\{ x = 0, \ldots, n : \frac{L_0(x)}{L_1(x)} \leq \delta \right\}$. Il rapporto di verosimiglianza $\frac{L_0(x)}{L_1(x)}$ è

$$\frac{L_0(x)}{L_1(x)} = \frac{\binom{n}{x}\vartheta_0^x(1-\vartheta_0)^{n-x}}{\binom{n}{x}\vartheta_1^x(1-\vartheta_1)^{n-x}} = \left(\frac{\vartheta_0}{\vartheta_1}\right)^{n\cdot x/n} \left(\frac{1-\vartheta_0}{1-\vartheta_1}\right)^{n(1-x/n)}$$

Sia $y := x/n$. Poiché supponiamo $\vartheta_0 < \vartheta_1$ le funzioni $y \mapsto \left(\frac{\vartheta_0}{\vartheta_1}\right)^{ny}$ e $y \mapsto \left(\frac{1-\vartheta_0}{1-\vartheta_1}\right)^{n(1-y)}$ sono entrambe decrescenti in $y = x/n$ e quindi L_0/L_1, in quanto prodotto di due funzioni decrescenti, è descrescente anche esso. Segue che $\frac{L_0(x)}{L_1(x)} \leq \delta$ se e solo se $x/n \geq t$, con un opportuno t funzione di δ. Invece di usare il ragionamento qui svolto, basato sulla monotonia, si può prendere $\ln L_0/L_1$ e "lavorare" su quest'ultimo.

3. Se $\vartheta = 0.5$ allora $30 \times 0.5 = 15 > 5$: è ragionevole approssimare la fdr di $\frac{X_{30}}{30}$ con la fdr $\mathcal{N}(0.5, \frac{1}{120})$. Analogamente, se $\vartheta = 0.7$ allora $30 \times (1-0.7) = 9 > 5$ e risulta ragionevole approssimare la fdr di $\frac{X_{30}}{30}$ con la fdr $\mathcal{N}(0.7, 0.007)$.

4. Imponendo $P_{\vartheta_0}(G) \simeq \alpha$, otteniamo per t:

$$P_{\vartheta_0}(G) = P_{\vartheta_0}\left(\frac{X_n}{n} \geq t\right) \simeq 1 - \Phi\left(\frac{\sqrt{n}(t-\vartheta_0)}{\sqrt{\vartheta_0(1-\vartheta_0)}}\right) = \alpha \text{ se e solo se } \frac{\sqrt{n}(t-\vartheta_0)}{\sqrt{\vartheta_0(1-\vartheta_0)}} \simeq \Phi^{-1}(1-\alpha) \implies$$

$$t = \vartheta_0 + \Phi^{-1}(1-\alpha)\sqrt{\frac{\vartheta_0(1-\vartheta_0)}{n}} = 0.5 + \Phi^{-1}(1-0.05)\sqrt{\frac{0.5\times 0.5}{30}} \simeq 0.5 + 0.1645 \times 0.5 \simeq 0.65$$

Essendo: $18/30 = 0.6 < 0.65$: NON possiamo rifiutare H_0.

5. $\beta = P_{\vartheta_1}\left(G^c\right) = P_{\vartheta_1}\left(\frac{X_n}{n} < t\right) \simeq \Phi\left(\frac{\sqrt{n}(t-\vartheta_1)}{\sqrt{\vartheta_1(1-\vartheta_1)}}\right) = \Phi\left(\sqrt{\frac{n}{\vartheta_1(1-\vartheta_1)}}\,(\vartheta_0 - \vartheta_1) + \Phi^{-1}(1-\alpha)\sqrt{\frac{\vartheta_0(1-\vartheta_0)}{\vartheta_1(1-\vartheta_1)}}\right) =$
$\Phi\left(\sqrt{\frac{30}{0.7\times 0.3}}(0.65 - 0.7)\right) \simeq \Phi(-0.5976) \simeq 0.275$ ∎

Esercizio 3.3 Due macchine A e B producono filo di rame, il cui diametro si è stabilito abbia un certo valore (assegnato) μ_0. Per controllare la qualità del processo vengono ispezionati 10 fili prodotti dalla macchina A e 15 prodotti dalla macchina B e per ogni filo viene registrato l'errore nella lunghezza del diametro (errore=valore misurato $-\mu_0$). Dalle misurazioni effettuate si ottiene: la somma dei quadrati degli errori nella lunghezza dei diametri è 0.017 mm^2 per i 10 fili prodotti dalla macchina A e 0.095 mm^2 per i 15 prodotti dalla macchina B.

Assumendo che gli errori nella lunghezza del diametro siano variabili aleatorie indipendenti, gaussiane e a media nota e uguale a zero:

1. fornite una stima puntuale della varianza σ_A^2 dell'errore nella lunghezza del diametro dei fili prodotti da A e della varianza σ_B^2 dell'errore nella lunghezza del diametro dei fili prodotti da B;

2. impostando un opportuno test di verifica di ipotesi di ampiezza $\alpha = 10\%$, potete ritenere che le due macchine abbiano la stessa precisione?

SOLUZIONE Indichiamo con X_j la lunghezza effettiva del diametro del j-esimo filo prodotto da A e con Y_j la lunghezza effettiva del diametro del j-esimo filo prodotto da B. Poiché per ipotesi gli errori hanno tutti media nulla, allora $\mathrm{E}(X_j) = \mathrm{E}(Y_j) = \mu_0 \ \forall j$, e

1. $S_{0A}^2 := \frac{\sum_{j=1}^{10}(X_j-\mu_0)^2}{10}$ è uno stimatore non distorto per σ_A^2 e $S_{0B}^2 := \frac{\sum_{j=1}^{15}(Y_j-\mu_0)^2}{15}$ è uno stimatore non distorto per σ_B^2. Sulla base dei dati a nostra disposizione: $s_{0A}^2 = 0.017/10 = 0.0017$ e $s_{0A}^2 = 0.095/15 \simeq 0.0063$.

2. Impostiamo il problema di verifica di ipotesi

$$H_0 : \sigma_A^2 = \sigma_B^2 \ \textit{versus} \ H_0 : \sigma_A^2 \neq \sigma_B^2$$

Fissato un livello di significatività α, una regione critica è

$$G = \left\{ \frac{S_{0A}^2}{S_{0B}^2} > F_{10,15}(1-\frac{\alpha}{2}) \text{ oppure } \frac{S_{0A}^2}{S_{0B}^2} < F_{10,15}(\frac{\alpha}{2}) \right\}$$

dove $F_{m,n}(\gamma)$ indica il quantile di ordine γ della fdr di Fisher con (m,n) gradi di libertà. Con $\alpha = 0.1$

$$F_{10,15}(1-\frac{\alpha}{2}) = F_{10,15}(0.95) = 2.54$$
$$F_{10,15}(\frac{\alpha}{2}) = \frac{1}{F_{15,10}(1-\frac{\alpha}{2})} = \frac{1}{2.85} \simeq 0.3508$$

Essendo $\frac{s_{0A}^2}{s_{0B}^2} = \frac{0.0017}{0.0063} \simeq 0.2698 < 0.3508$, rifiutiamo l'ipotesi nulla H_0: sulla base dei dati, con un livello di significatività del 10% accettiamo l'ipotesi alternativa che le due macchine abbiano precisione diversa. ∎

Esercizio 3.4 Il tempo di esecuzione del programma xxx sul calcolatore yyy è compreso fra 60 e 120 minuti primi. Vogliamo verificare se tale tempo possa essere modellato come una variabile aleatoria X assolutamente continua con densità

$$f(x) = \frac{1}{1800}(x-60)\mathbf{1}_{(60,120)}(x)$$

A tale fine, su ciascuno di 75 calcolatori tutti del tipo yyy e che lavorano indipendentemente uno dall'altro, viene lanciato il programma xxx e si registrano i tempi di esecuzione. I risultati sperimentali ottenuti sono i seguenti:

intervalli di tempo A_k	# di programmi il cui tempo di esecuzione cade in A_k
$A_1 = (60, 75)$	6
$A_2 = [75, 80)$	8
$A_3 = [80, 95)$	20
$A_4 = [95, 105)$	17
$A_5 = [105, 110)$	10
$A_6 = [110, 120)$	14

Sulla base dei dati raccolti, verificate con un opportuno test se la densità f fornisce un buon modello probabilistico per il tempo di esecuzione del programma xxx su un calcolatore del tipo yyy.

SOLUZIONE Avendo a disposizione solo dati raggruppati, effettuiamo un test χ^2 di buon adattamento.

Sia $\theta_{0,k} = P(X \in A_k) = \int_{A_k} \frac{1}{1800}(x-60)\mathbf{1}_{(60,120)}(x)\,dx$, per $k = 1, \ldots, 6$. Allora

intervalli di tempo A_k	# di programmi il cui tempo di esecuzione cade in A_k	θ_{0k}	$75\theta_{0k}$
$A_1 = (60, 75)$	6	0.0625	4.6875
$A_2 = [75, 80)$	8	0.0486	3.6450
$A_3 = [80, 95)$	20	0.2292	17.1900
$A_4 = [95, 105)$	17	0.2222	16.6650
$A_5 = [105, 110)$	10	0.1320	9.9000
$A_6 = [110, 120)$	14	0.3055	22.9125

Accorpiamo le classi A_1 ed A_2 dal momento che $75\theta_{01} < 5$ e $75\theta_{02} < 5$ e l'approssimazione asintotica χ^2 con le 6 classi non funziona.

Denotiamo con B_k le nuove classi. Abbiamo

intervalli di tempo B_k	# di programmi il cui tempo di esecuzione cade in $B_k (= N_k)$	θ_{0k}	$75\theta_{0k}$
$B_1 = (60, 80)$	14	0.1111	8.3325
$B_2 = [80, 95)$	20	0.2292	17.1900
$B_3 = [95, 105)$	17	0.2222	16.6650
$B_4 = [105, 110)$	10	0.1320	9.9000
$B_5 = [110, 120)$	14	0.3055	22.9125

La statistica di Pearson è

$$Q_{75} = \sum_{k=1}^{5} \frac{(N_k - 75\theta_{0k})^2}{75\theta_{0k}} = \sum_{k=1}^{5} \frac{N_k^2}{75\theta_{0k}} - 75 = 7.7887$$

Il p-value del test è pari a $P(Q_{75} > 7.7887) = 1 - F_{\chi^2_4}(7.7887) \simeq 0.09963$ per cui, ad esempio, se si effettuasse un test di livello $\alpha = 5\%$ si accetterebbe l'ipotesi mentre se si effettuasse un test di livello $\alpha = 10\%$ si rifiuterebbe.

∎

Esercizio 3.5 Sette insegnanti partecipanti ad un corso di aggiornamento in storia contemporanea hanno sostenuto la prova finale. Le età di questi insegnanti e i loro esiti nella prova (espressi in centesimi) sono i seguenti:

Età	24	31	38	45	46	28	30
Risultato della prova	68	85	84	92	90	65	86

Secondo voi (e sulla base di questi dati), gli insegnanti più anziani sono "più bravi" di quelli più giovani? Impostate un opportuno problema di verifica di ipotesi di livello $\alpha = 10\%$. La vostra risposta cambia per $\alpha = 1\%$?

SOLUZIONE Chiamiamo X la variabile aleatoria che indica l'età e Y quella che indica il risultato. Impostiamo il test di concordanza di Kendall di livello $\alpha = 10\%$ per il seguente problema:

$$H_0 : \tau = 0 \text{ (oppure } H_0 : \tau \leq 0) \textit{ versus } H_1 : \tau > 0$$

Per eseguire il test è necessario calcolare il numero di concordanze e discordanze. A questo scopo riordiniamo le coppie per valori della variabile X crescenti:

X (Età)	24	28	30	31	38	45	46
Y (Risultato	68	65	86	85	84	92	90
$segno\big[y_{[j]} - y_{[1]}\big]_{j>1}$		-1	$+1$	$+1$	$+1$	$+1$	$+1$
$segno\big[y_{[j]} - y_{[2]}\big]_{j>2}$			$+1$	$+1$	$+1$	$+1$	$+1$
$segno\big[y_{[j]} - y_{[3]}\big]_{j>3}$				-1	-1	$+1$	$+1$
$segno\big[y_{[j]} - y_{[4]}\big]_{j>4}$					-1	$+1$	$+1$
$segno\big[y_{[j]} - y_{[5]}\big]_{j>5}$						$+1$	$+1$
$segno\big[y_{[j]} - y_{[6]}\big]_{j>6}$							-1

Dalla tabella risulta $C = 16$ e $D = 5$, dunque $C - D = 11$. La regione critica del test di livello α è costituita dai campioni per cui $C - D > q_{1-\alpha}(C - D)$, dove $q_{1-\alpha}(C - D)$ rappresenta il quantile di ordine $1 - \alpha$ della statistica $C - D$ per $n = 7$. Per $\alpha = 0.1$ e $n = 7$: $q(0.90) = 9$. Avendo osservato $C - D = 11 > 9$, concludiamo che a livello $\alpha = 10\%$ propendiamo per l'ipotesi che gli insegnanti più vecchi siano più bravi di quelli più giovani. Diminuendo α, il quantile $q_{1-\alpha}(C - D)$ aumenta e quindi può succedere che per $\alpha < 10\%$, la decisione cambi. Effettivamente, $q_{1-0.01}(C - D) = 13 > 11 = C - D$: a livello $\alpha = 1\%$ accettiamo l'ipotesi H_0 che l'età non influisca sulla "bravura" degli insegnanti. ∎

4 STATISTICA per ING INF [2L] Proff. A. Barchielli, I. Epifani 15.02.05

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

Esercizio 4.1 Sospettiamo che due dadi perfettamente identici siano stati entrambi truccati in modo tale che, lanciandoli in coppia, la probabilità di ottenere come somma delle facce superiori il valore 7 sia pari a 1/11. Denotiamo con p la probabilità che la somma delle facce superiori di due dadi uguali e lanciati simultaneamente sia 7.

1. Se i due dadi sono regolari quanto vale p?

Sia p_0 il valore determinato al punto 1. Vogliamo verificare l'ipotesi nulla $H_0 : p = p_0$ contro l'alternativa $H_1 : p = 1/11$. Decidiamo di eseguire questa verifica nel seguente modo: lanciamo 144 volte la coppia di dadi e rifiutiamo H_0 se la somma dei due dadi è 7 al più per 14 volte.

2. Calcolate "approssimativamente" il livello di significatività α del test.

3. Calcolate "approssimativamente" la potenza π del test.

4. Calcolate "approssimativamente" la probabilità di errore di seconda specie β del test.

Soluzione

1. $p_0 = \dfrac{\#\{(i,j) :\ i+j=7 \text{ e } i,j=1,\dots,6\}}{\#\{(i,j) :\ i,j=1,\dots,6\}} = \dfrac{6}{36} = \dfrac{1}{6}$

2. Il livello di significatività del test α è la probabilità di rifiutare $H_0 : p = 1/6$ quando $p = 1/6$. Sia $\hat{p}$ la frequenza relativa dell'evento "somma della coppia di dadi = 7"; la regola è rifiutare $H_0 : p = 1/6$ contro $H_1 : p = 1/11$ se $\hat{p} \le \dfrac{14}{144}$, e quindi, per il Teorema Centrale del Limite, un valore approssimato di α è dato da

$$\alpha = P_{1/6}\left(\hat{p} \le \frac{14}{144}\right) \simeq \Phi\left(\sqrt{144}\,\frac{\frac{14}{144} - \frac{1}{6}}{\sqrt{\frac{1}{6}\times\frac{5}{6}}}\right) \simeq \Phi(-2.236) = 1 - \Phi(2.236) \simeq 1 - 0.9874 = 0.0126 = 1.26\%$$

Con la correzione di continuità, il valore approssimato di α risulta:

$$\alpha \simeq \Phi\left(\sqrt{144}\,\frac{\frac{14.5}{144} - \frac{1}{6}}{\sqrt{\frac{1}{6}\times\frac{5}{6}}}\right) \simeq \Phi(-2.124) = 1 - \Phi(2.124) = 1 - 0.98316 \simeq 1.68\%$$

In realtà, il valore esatto di α è 1.27%.

3. La potenza del test è la probabilità di accettare H_1 quando è vera: sempre per il Teorema Centrale del Limite:

$$\pi\left(\frac{1}{11}\right) = P_{1/11}\left(\hat{p} \le \frac{14}{144}\right) \simeq \Phi\left(\sqrt{144}\,\frac{\frac{14}{144} - \frac{1}{11}}{\sqrt{\frac{1}{11}\times\frac{10}{11}}}\right) \simeq \Phi(0.264) \simeq 0.604$$

Con la correzione di continuità otteniamo $\pi(1/11) \simeq \Phi\left(\sqrt{144}\,\frac{\frac{14.5}{144} - \frac{1}{11}}{\sqrt{\frac{1}{11}\times\frac{10}{11}}}\right) \simeq \Phi(0.41) \simeq 0.659$. In realtà, il valore esatto di $\pi(1/11)$ è 0.67.

4. La probabilità di errore di seconda specie β è la probabilità di rifiutare l'ipotesi alternativa quando è vera, quindi, un valore approssimato di $\beta = 1 - \pi(1/11)$ è $\beta \simeq 1 - 0.604 = 0.396$ senza la correzione di continuità e $\beta \simeq 1 - 0.659 = 0.341$ con la correzione di continuità. ∎

Esercizio 4.2 Il manufatto `aaa` è prodotto in un gran numero di stabilimenti. La proporzione X di manufatti difettosi (variabile da stabilimento a stabilimento) può essere modellata come una variabile aleatoria assolutamente continua con densità

$$f(x,\theta) = \begin{cases} \frac{1}{\theta} x^{\frac{1}{\theta}-1} & 0 < x < 1 \\ 0 & \text{altrove} \end{cases}$$

dove θ è un parametro positivo incognito.
Per stimare θ gli addetti al controllo di qualità scelgono a caso n stabilimenti.

1. Determinate uno stimatore di θ usando il metodo dei momenti.
2. Determinate uno stimatore di θ usando il metodo di massima verosimiglianza.
3. Determinate la densità di $Y = -\log X$.
4. Discutete le proprietà dello stimatore di massima verosimiglianza individuato al punto 2. (non distorsione, consistenza, efficienza, ...).
5. Gli addetti al controllo di qualità decidono di visitare 4 stabilimenti e di ispezionare 30 manufatti in ognuno di essi. Se trovano 2 pezzi difettosi nel primo, 3 nel secondo, 3 nel terzo e 1 nel quarto, qual è la stima di θ basata sul metodo dei momenti? E qual è quella basata sul metodo di massima verosimiglianza?

Soluzione

1. Sia $X_1, \ldots, X_n$ il campione casuale delle proporzioni di pezzi difettosi degli n stabilimenti e sia $\bar{X}$ la media campionaria di $X_1, \ldots, X_n$. Allora $\mathrm{E}_\theta(X) = \displaystyle\int_0^1 x \frac{1}{\theta} x^{\frac{1}{\theta}-1}\, dx = \frac{1}{\theta}\left[\frac{x^{1/\theta+1}}{1/\theta+1}\right]_0^1 = \frac{1/\theta}{1/\theta+1} = \frac{1}{\theta+1}$
ed $\mathrm{E}_\theta(X) = \bar{X}$ se e solo se $\theta = 1/\bar{X} - 1$; segue che $\widetilde{\theta} = 1/\bar{X} - 1$ (≥ 0) è lo stimatore dei momenti di θ.
2. Studiamo la funzione di verosimiglianza del campione $X_1, \ldots, X_n$:

$$L_\theta(x_1, \ldots, x_n) = \prod_{i=1}^n \left(\frac{1}{\theta} x_i^{\frac{1}{\theta}-1}\right) = \theta^{-n} \left(\prod_{i=1}^n x_i\right)^{\frac{1}{\theta}-1} \qquad \theta > 0$$

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

$$\frac{\partial \log L_\theta(x_1, \ldots, x_n)}{\partial\theta} = \frac{-n}{\theta} + \frac{-\sum_{j=1}^n \log x_j}{\theta^2} \geq 0 \text{ se e solo se } \theta \leq -\frac{\sum_{j=1}^n \log x_j}{n}$$

infatti $-\sum_{j=1}^n \log X_j > 0$ poiché $P(0 < X_j < 1) = 1$. Segue che $\widehat{\theta} = -\dfrac{\sum_{j=1}^n \log X_j}{n}$ è MLE per θ.
3. Dato che $P(0 < X < 1) = 1$ si ha $P(Y \leq y) = 0$ $\forall y \leq 0$. Per $y > 0$:

$$F_Y(y) = P(-\log X \leq y) = P(\log X \geq -y) = P(X \geq \mathrm{e}^{-y}) = 1 - F_X(\mathrm{e}^{-y})$$

e quindi $f_Y(y) = f_X(\mathrm{e}^{-y})\mathrm{e}^{-y}\mathbf{1}_{(0,\infty)}(y) = \frac{1}{\theta}\mathrm{e}^{-y(\frac{1}{\theta}-1)}\mathrm{e}^{-y}\mathbf{1}_{(0,\infty)}(y) = \frac{1}{\theta}\mathrm{e}^{-\frac{1}{\theta}y}\mathbf{1}_{(0,\infty)}(y)$: $Y \sim \mathcal{E}(\theta)$
4. Sia $Y_i = -\log X_i$. Allora $Y_i \sim \mathcal{E}(\theta)$ con $\mathrm{E}_\theta(Y_1) = \theta$ e lo stimatore di massima verosimiglianza $\widehat{\theta}$ coincide con la media campionaria di $Y_1, \ldots, Y_n$: $\widehat{\theta} = \bar{Y}$. Segue che $\widehat{\theta}$ è stimatore non distorto e consistente in media quadratica per θ. Inoltre, applicando il Teorema Centrale del Limite, deduciamo la gaussianità asintotica di $\widehat{\theta}$ nel senso che $\lim_{n\to\infty} P\left(\sqrt{n}\dfrac{\widehat{\theta}-\theta}{\theta} \leq z\right) = \Phi(z)$, $\forall z \in \mathbb{R}$. Infine, osserviamo che $\frac{\partial \log L_\theta(x_1,\ldots,x_n)}{\partial\theta} = \frac{n}{\theta^2}(\widehat{\theta} - \theta)$,
Pertanto, scelta $a(\theta, n) = n/\theta^2$, abbiamo:

$$P_\theta\left(\frac{\partial \log L_\theta(X_1, \ldots, X_n)}{\partial\theta} = a(\theta, n)(\widehat{\theta} - \theta)\right) = 1 \quad \forall\theta > 0$$

L'ultima è condizione necessaria e sufficiente affinché $\mathrm{Var}(\widehat{\theta})$ raggiunga il confine inferiore di Cramer Rao. Abbiamo così dimostrato che $\widehat{\theta}$ è stimatore efficiente per θ.
5. Abbiamo il campione delle quattro osservazioni: $x_1 = 2/30$, $x_2 = 3/30$, $x_3 = 3/30$ e $x_4 = 1/30$, in corrispondenza del quale $\widetilde{\theta} = 37/3 \simeq 12.33$ e $\widehat{\theta} \simeq 2.679$. Poiché $\widehat{\theta}$ è stimatore efficiente mentre $\widetilde{\theta}$ è distorto, allora $\widehat{\theta}$ è preferibile a $\widetilde{\theta}$. ∎

Esercizio 4.3 Un segnale di valore μ trasmesso dalla sorgente A viene raccolto dal ricevente B con un rumore additivo gaussiano di media nulla e varianza $\sigma^2 = 16$. Per ridurre l'errore, lo stesso segnale viene trasmesso 9 volte da A a B e la media campionaria dei segnali ricevuti è 9.00.

1. Quale fiducia avete che il segnale trasmesso da A fosse compreso fra 6.38 e 11.62?
2. B ha motivo di supporre che il segnale inviato dovesse essere 12. Verificate l'ipotesi nulla $H_0 : \mu = 12$ contro l'alternativa $H_1 : \mu \neq 12$, a livello di significatività $\alpha = 10\%$.
3. Determinate il p-value dei dati del test per l'ipotesi nulla $H_0 : \mu = 12$ contro l'alternativa $H_1 : \mu \neq 12$.

SOLUZIONE

1. Sia $(X_1, \ldots, X_9)$ il campione casuale delle 9 trasmissioni da A a B. Deriviamo dal testo che $X_i \sim \mathcal{N}(\mu, 16)$. Un intervallo di confidenza simmetrico per μ di livello γ è dato da $\bar{X} \mp z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}}$ ed è lungo $2z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}}$. L'intervallo $(6.38, 11.62)$ è lungo $11.62 - 6.38 = 5.24$. Risolvendo l'equazione in γ:

$$5.24 = 2z_{\frac{1+\gamma}{2}} \frac{\sigma}{\sqrt{n}} = 2z_{\frac{1+\gamma}{2}} \frac{4}{3} = \frac{8}{3} z_{\frac{1+\gamma}{2}}$$

otteniamo

$$z_{\frac{1+\gamma}{2}} = 1.965$$

Dalle tavole della fdr $\mathcal{N}(0,1)$ Φ, risulta che 1.965 è il quantile di ordine 0.9753 di Φ cioè $\frac{1+\gamma}{2} = 0.9753$, da cui $\gamma = 0.9753 \times 2 - 1 = 0.9506 (\simeq 0.95)$. Possiamo dire di avere il 95(.06)% di fiducia che il vero segnale fosse compreso fra 6.38 e 11.62.

2. Poiché $12 > 11.62$, per la dualità tra verifica delle ipotesi e intervalli di confidenza, rifiutiamo $H_0 : \mu = 12$ a favore di $H_1 : \mu \neq 12$ a livello $1 - 0.9506 = 4.94\%$. Essendo $10\% > 4.94\%$, rifiutiamo anche a livello 10%.
3. Usando la teoria dei test per popolazioni gaussiane, a livello α, rifiuteremo $H_0 : \mu = 12$ (a favore di $H_1 : \mu \neq 12$) se $\frac{|\bar{x} - 12|}{4/3} = \frac{|9 - 12|}{4/3} = 2.25 \geq z_{1-\alpha/2}$. Il p-value è il più piccolo livello per cui si rifiuta H_0 con i risultati empirici. Il p-value dei dati di questo test è $2\,(1 - \Phi(2.25)) = 0.02445 \simeq 2.45\%$: rifiutiamo H_0 per ogni $\alpha \geq 2.45\%$. ∎

Esercizio 4.4 I valori che seguono rappresentano i giorni di sopravvivenza di un campione di 6 topi affetti da cancro e curati con una terapia sperimentale:

$$29,\ 700,\ 1,\ 335,\ 15,\ 160$$

1. Determinate la funzione di ripartizione empirica $\hat{F}_6$ associata al campione dei 6 topi.
2. Determinate una stima della probabilità che un topo affetto da cancro sottoposto alla terapia viva più di 15 giorni.

Si pensa che la sopravvivenza dei topi malati di cancro e sottoposti alla terapia sperimentale possa essere modellata come una variabile aleatoria X assolutamente continua che ha densità di Weibull:

$$f_0(x) = \frac{1}{20\sqrt{x}} \mathrm{e}^{-\frac{\sqrt{x}}{10}} \mathbf{1}_{(0,\infty)}(x)$$

3. Usate un opportuno test con il 5% di livello di significatività, per stabilire se i dati forniti sui topi possano provenire dalla densità di Weibull $f_0(x) = \frac{1}{20\sqrt{x}} \mathrm{e}^{-\frac{\sqrt{x}}{10}} \mathbf{1}_{(0,\infty)}(x)$ ipotizzata.

Soluzione

1. Ordiniamo le osservazioni in ordine crescente: 1 15 29 160 335 700. Quindi:

$$\hat{F}_6(x) = \begin{cases} 0 & x < 1 \\ \frac{1}{6} \simeq 0.166 & 1 \le x < 15 \\ \frac{1}{3} \simeq 0.333 & 15 \le x < 29 \\ \frac{1}{2} = 0.5 & 29 \le x < 160 \\ \frac{2}{3} = 0.666 & 160 \le x < 335 \\ \frac{5}{6} \simeq 0.833 & 335 \le x < 700 \\ 1 & x \ge 700 \end{cases}$$

2. $P(X > 15) = 1 - F_X(15)$ e $\hat{F}_6(15) = \dfrac{1}{3}$. Quindi una stima di $P(X > 15)$ è $\dfrac{2}{3}$.

3. Ho un numero "piccolo" di dati (6) non raggruppati e il campione proviene da una fdr continua. Impostiamo il test di Kolmogorov-Smirnov di livello 5% per verificare: $H_0 : X \sim F_0$ contro l'alternativa $H_1 : F \not\sim F_0$, dove

$$F_0(x) = \int_0^x f_0(t)\, dt = \int_0^x \frac{1}{20\sqrt{t}} \mathrm{e}^{-\frac{\sqrt{t}}{10}}\, dt = \int_0^{\sqrt{x}} \frac{1}{20u} \mathrm{e}^{-\frac{u}{10}}\, 2u du = \int_0^{\sqrt{x}} \frac{1}{10} \mathrm{e}^{-\frac{u}{10}}\, du = 1 - \mathrm{e}^{-\frac{\sqrt{x}}{10}} \quad \forall x > 0$$

e $F_0(x) = 0$ se $x \le 0$. Pertanto:

$F_0(1)$	$F_0(15)$	$F_0(29)$	$F_0(160)$	$F_0(335)$	$F_0(700)$
$1 - \mathrm{e}^{-\frac{\sqrt{1}}{10}} \simeq 0.095$	0.321	0.416	0.718	0.840	0.929

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

$$\sup_{x \in \mathbb{R}} |\hat{F}_6(x) - F_0(x)| = F_0(160) - \hat{F}_6(29) = 0.718 - 0.5 = 0.218$$

e, dalle tavole dei quantili della statistica di Kolmogorov-Smirnov con $n = 6$, $q_{D_6}(1 - 0.05) = 0.5193$: 0.218 è minore di 0.5193 e accettiamo H_0. ∎

Esercizio 4.5 Quindici misure ripetute eseguite con lo stesso strumento e in modo indipendente hanno dato i seguenti risultati, riportati nell'ordine in cui sono stati ottenuti:

$$0.30 \quad 1.27 \quad -0.25 \quad -1.28 \quad -1.20 \quad -1.74 \quad 2.18 \quad 0.23 \quad -1.10 \quad 1.08 \quad 0.69 \quad 1.69 \quad 1.84 \quad 0.97 \quad 2.00$$

Lo sperimentatore sospetta però che ci sia stato un deterioramento dello strumento nel corso dell'esperimento che potrebbe aver distrutto l'indipendenza fra le misure.

1. Verificate l'ipotesi d'indipendenza delle misure al livello del 10%.
2. Determinate in modo approssimato il p-value del test, o piuttosto indicate un intervallo dove tale p-value cade.

SOLUZIONE Usiamo il test di aleatorietà di Kendall, a due code; a priori non ci aspettiamo né un andamento crescente né uno decrescente. Contiamo il numero di concordanze e discordanze; con i simboli degli appunti abbiamo:
$D_1 = 6,\ C_1 = 8;\ D_2 = 9,\ C_2 = 4;\ D_3 = 4,\ C_3 = 8;\ D_4 = 1,\ C_4 = 10;\ D_5 = 1,\ C_5 = 9;\ D_6 = 0,\ C_6 = 9;\ D_7 = 8,\ C_7 = 0;\ D_8 = 1,\ C_8 = 6;\ D_9 = 0,\ C_9 = 6;\ D_{10} = 2,\ C_{10} = 3;\ D_{11} = 0,\ C_{11} = 4;\ D_{12} = 1,\ C_{12} = 2;\ D_{13} = 1,\ C_{13} = 1;\ D_{14} = 0,\ C_{14} = 1$
e

$$D = \sum_{i=1}^{14} D_i = 34\,, \qquad C = \sum_{i=1}^{14} C_i = 71\,, \qquad T = C - D = 37$$

1. Per un test di ampiezza α la regione di rifiuto è $\{|T| > q_{\mathrm{Ken};n}(1-\alpha/2)\}$. Noi abbiamo $n = 15$, $\alpha = 0.1$, $|T| = 37$ e, dalle tabelle, $q_{\mathrm{Ken};15}(.95) = 33$: rifiutiamo l'ipotesi nulla che i dati provengano da un campione casuale e accettiamo l'ipotesi di non indipendenza dei dati.
2. Dalle tabelle vediamo che il nostro dato cade tra 33 e 39 corrispondenti ai quantili di ordine 0.95 e 0.975. Dunque il p-value cade fra 0.05 e 0.1. ∎

5 STATISTICA per ING INF [2L] Proff. A. Barchielli, I. Epifani 07.03.05

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

Esercizio 5.1 Sia $X_1, \dots, X_n$ un campione casuale estratto dalla popolazione gaussiana di densità

$$f(x, \theta_1, \theta_2) = \sqrt{\frac{\theta_1}{\pi}} \mathrm{e}^{-\theta_1 (x-\theta_2)^2} \qquad x \in \mathbb{R},\ \theta_1 > 0 \text{ e } \theta_2 \in \mathbb{R}$$

Entrambi i parametri θ_1, θ_2 sono incogniti.

1. Determinate uno stimatore di θ_1 usando il metodo di massima verosimiglianza.
2. Determinate un intervallo di confidenza a due code per θ_1 di livello $\gamma = 95\%$.
3. Avete ora a disposizione il campione di quattro osservazioni: $x_1 = -0.17$, $x_2 = 0.71$, $x_3 = 2.17$ e $x_4 = 1.00$ e dovete scegliere fra l'ipotesi nulla $H_0 : \theta_1 = 0.5$ e l'alternativa $H_1 : \theta_1 \neq 0.5$. Quale decisione prendete al livello $\alpha = 5\%$? (Giustificate rigorosamente la risposta).

SOLUZIONE Osserviamo che $f(x, \theta_1, \theta_2) = \mathcal{N}\left(\theta_2, \frac{1}{2\theta_1}\right)$. Quindi, praticamente, dobbiamo fare inferenza sul reciproco della varianza della popolazione gaussiana con media incognita

1. Chiamiamo σ^2 la varianza di questa densità gaussiana. La caratteristica da stimare è la funzione: $\theta_1 = (1/2\sigma^2)$. Lo stimatore di massima verosimiglianza di σ^2 nel modello gaussiano con media incognita è $\widehat{\sigma}^2 = \dfrac{\sum_{j=1}^n (X_j - \bar{X})^2}{n}$ e quindi quello di θ_1 è $\widehat{\theta}_1 = \dfrac{n}{2\sum_{j=1}^n (X_j - \bar{X})^2} = \dfrac{n}{2(n-1)S^2}$ dove S^2 è la varianza campionaria $S^2 = \dfrac{\sum_{j=1}^n (X_j - \bar{X})^2}{n-1}$.

2. Se (T_1, T_2) è un intervallo di confidenza per σ^2 di livello $\gamma = 95\%$, allora $\left(\dfrac{1}{2T_2}, \dfrac{1}{2T_1}\right)$ è un intervallo di confidenza per θ_1, sempre di livello $\gamma = 95\%$. Nel caso della popolazione $\mathcal{N}\left(\theta_2, \sigma^2\right)$, un intervallo di confidenza bilatero di livello $\gamma = 95\%$ per σ^2 è

$$\frac{(n-1)S^2}{\chi^2_{n-1}\left(\frac{1+0.95}{2}\right)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1}\left(\frac{1-0.95}{2}\right)}$$

Quindi l'intervallo cercato per θ_1 è

$$\frac{\chi^2_{n-1}(0.025)}{2(n-1)S^2} < \theta_1 < \frac{\chi^2_{n-1}(0.975)}{2(n-1)S^2}$$

Osservate che l'intervallo di confidenza trovato per θ_1 è funzione dello stimatore di massima verosimiglianza $\widehat{\theta}_1$. Infatti:

$$\left(\frac{\chi^2_{n-1}(0.025)}{2(n-1)S^2},\ \frac{\chi^2_{n-1}(0.975)}{2(n-1)S^2}\right) = \left(\chi^2_{n-1}(0.025)\,\frac{\widehat{\theta}_1}{n}, \chi^2_{n-1}(0.975)\,\frac{\widehat{\theta}_1}{n}\right)$$

3. $\chi^2_3(0.025) = 0.216$, $\chi^2_3(0.975) = 9.348$, $\bar{X} = 0.9275$, $S^2 = 0.933625$ quindi un intervallo di confidenza di livello 0.95 per θ_1 è $(0.039, 1.669)$. Poiché $0.5 \in (0.039, 1.669)$, per la dualità tra verifica delle ipotesi e intervalli di confidenza, accettiamo $H_0 : \theta_1 = 0.5$ a livello $1 - 0.95 = 5\%$. Alternativamente, scrivete le ipotesi come $H_0 : \sigma^2 = 1$, $H_1 : \sigma^2 \neq 1$ ed eseguite un test bilatero sulla varianza con media incognita. ∎

Esercizio 5.2 Abbiamo estratto il campione casuale $X_1, \ldots, X_n$ dalla densità esponenziale di parametro θ: $f(x,\theta) = \frac{1}{\theta}\mathrm{e}^{-x/\theta}\mathbf{1}_{(0,\infty)}(x)$, $\theta > 0$.

1. Determinate la densità della variabile aleatoria $\dfrac{2\sum_{i=1}^n X_i}{\theta}$.
2. Costruite un test uniformemente più potente di livello α per verificare l'ipotesi nulla $H_0 : \theta = 2$ contro l'alternativa $H_1 : \theta = 1.49$.
3. Sia $\alpha = 5\%$, $n = 3$ e $x_1 = 0.4$, $x_2 = 2.9$ e $x_3 = 1.2$. Sulla base del test costruito al punto 2., accettate o rifiutate H_0?
4. Sia $\alpha = 5\%$ e $n = 3$ come sopra. Calcolate la probabilità di errore di secondo tipo β del test costruito al punto 2.

Soluzione

1. Sia $Y_i = 2X_i/\theta,\ i = 1, \ldots, n$. Le variabili aleatorie Y_i sono i.i.d.. Determiniamo la comune densità a partire dalla fdr F_{Y_i}:

$$F_{Y_i}(y) = \begin{cases} 0 & y < 0 \\ P\left(\frac{2X_i}{\theta} \le y\right) & y \ge 0 \end{cases} = \begin{cases} 0 & y < 0 \\ F_{X_i}\left(\frac{\theta y}{2}\right) & y \ge 0 \end{cases}$$

Quindi

$$f_{Y_i}(y) = \frac{\theta}{2} f_{X_i}\left(\frac{\theta y}{2}\right)\mathbf{1}_{(0,\infty)}(y) = \frac{1}{2}\mathrm{e}^{-\frac{y}{2}}\mathbf{1}_{(0,\infty)}(y)$$

cioè $Y_1, \ldots, Y_n$ è un campione casuale dalla densità esponenziale di parametro 2, che coincide con la densità chiquadrato con due gradi di libertà: χ^2_2. La somma è variabile aleatoria $\Gamma(n,2) = \Gamma(2n/2,2) = \chi^2_{2n}$.

2. Dal Lemma di Neyman Pearson, segue che la regione critica del test uniformemente più potente di livello α per verificare $H_0 : \theta = 2$ contro $H_1 : \theta = 1.49$ è:

$$\begin{aligned} \mathcal{G} &= \left\{(x_1, \ldots, x_n) \in (0,\infty)^n : \frac{L_2(x_1, \ldots, x_n)}{L_{1.49}(x_1, \ldots, x_n)} \le \delta\right\} \\ &= \left\{(x_1, \ldots, x_n) \in (0,\infty)^n : \left(\frac{1.49}{2}\right)^n \mathrm{e}^{-\sum_{j=1}^n x_j(1/2 - 1/1.49)} \le \delta\right\} \\ &= \left\{(x_1, \ldots, x_n) \in (0,\infty)^n : \sum_{j=1}^n x_j \le k\right\} \end{aligned}$$

con k tale che $P_2(\sum_{j=1}^n X_j \le k) = \alpha$. Per il punto 2., se $\theta = 2$, allora $\sum_{j=1}^n X_j \sim \chi^2_{2n}$ e quindi $k = \chi^2_{2n}(\alpha)$. In definitiva, rifiuteremo H_0 se $\sum_{j=1}^n x_j \le \chi^2_{2n}(\alpha)$.

3. $(0.4 + 2.9 + 1.2) = 4.5 > 1.635 = \chi^2_{2\times 3}(0.05)$: accettiamo H_0.

4. La probabilità di errore di seconda specie β è la probabilità di rifiutare l'ipotesi alternativa quando è vera. Quindi:

$$\beta = P_{1.49}\left(\sum_{j=1}^3 X_j > 1.635\right) = P_{1.49}\left(\frac{2\times\sum_{j=1}^3 X_j}{1.49} > \frac{2\times 1.635}{1.49}\right) \simeq 1 - F_{\chi^2_6}(2.20) = 1 - 0.1 = 0.9!!!\ \blacksquare$$

Esercizio 5.3 Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla densità

$$f(x,\theta) = \begin{cases} \frac{2}{\theta}\left(1 - \frac{x}{\theta}\right) & \text{se } 0 < x < \theta \\ 0 & \text{altrove} \end{cases}$$

dove θ è un parametro positivo incognito. Indichiamo con $\overline{X}$ la media campionaria di $X_1, \ldots, X_n$.

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 (MSE).

Supponete ora di avere estratto una sola osservazione ($n = 1$).

3. Determinate lo stimatore di massima verosimiglianza di θ.
   Potrebbe esservi utile disegnare il grafico della funzione di verosimiglianza (x fissato, θ variabile).
4. Calcolate l'errore quadratico medio dello stimatore di massima verosimiglianza trovato al punto 3.

Soluzione

1.

$$\mathrm{E}(\overline{X}) = \mathrm{E}(X_1) = \int_0^\theta x \frac{2}{\theta}\left(1 - \frac{x}{\theta}\right)\, dx = \frac{1}{\theta}\left[x^2 - \frac{2x^3}{3\theta}\right]_0^\theta = \frac{\theta}{3} \qquad \forall \theta > 0$$

$$\mathrm{Var}(\overline{X}) = \frac{\mathrm{Var}(X_1)}{n} = \frac{\theta^2}{18n} \qquad \forall \theta > 0$$

in quanto

$$\mathrm{Var}(X_1) = \mathrm{E}(X_1^2) - \mathrm{E}(X_1)^2 \text{ e } \mathrm{E}(X_1^2) = \int_0^\theta x^2 \frac{2}{\theta}\left(1 - \frac{x}{\theta}\right)\, dx = \frac{2}{\theta}\left[\frac{x^3}{3} - \frac{x^4}{4\theta}\right]_0^\theta = \frac{\theta^2}{6}$$

2. Poiché $\mathrm{E}(\overline{X}) = \dfrac{\theta}{3}\ \forall \theta > 0$, allora $T = 3\overline{X}$ è stimatore non distorto per θ; T ha errore quadratico medio dato da:

$$\mathrm{MSE}(T) = \mathrm{Var}(T) = \mathrm{Var}(3\overline{X}) = 9\frac{\theta^2}{18n} = \frac{\theta^2}{2n} \quad \forall \theta > 0$$

3. La funzione di verosimiglianza è $L_\theta(x_1) = \frac{2}{\theta}\left(1 - \frac{x_1}{\theta}\right)$ per $\theta > x_1$ ($x_1 > 0$) e 0 altrove: $L_\theta(x_1)$ è crescente in $(x_1, 2x_1)$ e decrescente in $(2x_1, +\infty)$; $L_\theta(x_1)$ è concava in $(x_1, 3x_1)$ e convessa in $(3x_1, +\infty)$. $L_\theta(x_1)$ ha massimo assoluto in $\theta = 2x_1$, quindi $\hat{\theta} = 2X_1$ è lo stimatore di massima verosimiglianza di θ.

4.

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(2X_1) + (\mathrm{E}(2X_1) - \theta)^2 = 4\frac{\theta^2}{18} + \left(2\frac{\theta}{3} - \theta\right)^2 = \frac{3}{9}\theta^2 = \frac{\theta^2}{3}$$

Con una sola osservazione, lo stimatore di massima verosimiglianza è preferibile a quello individuato al punto 1. ∎

Esercizio 5.4 [1] Il numero π scritto in forma decimale contiene nelle prime 10002 posizioni dopo il punto decimale le cifre

$$0, \quad 1, \quad 2, \quad 3, \quad 4, \quad 5, \quad 6, \quad 7, \quad 8, \quad 9$$

rispettivamente

$$968, \quad 1026, \quad 1021, \quad 974, \quad 1014, \quad 1046, \quad 1021, \quad 970, \quad 948, \quad 1014$$

volte.

1. Sulla base di questi dati, ritenete che nella rappresentazione decimale di π, le cifre $0, 1, \ldots, 9$ dopo il punto decimale siano uniformemente distribuite? Per rispondere alla domanda usate un opportuno test e scegliete come livello di significatività del test $\alpha = 5\%$.

2. Determinate in modo approssimato il p-value dei dati del test, o piuttosto indicate un intervallo dove tale p-value cade.

SOLUZIONE Osserviamo che i dati provengono da un modello discreto. Per rispondere alla domanda usiamo un test di adattamento χ^2 di Pearson per verificare l'ipotesi nulla $H_0 : p_k = 0.1 \; \forall k = 0, \ldots, 9$ contro l'alternativa $H_1 : p_k \neq 0.1$ per qualche $k = 0, \ldots, 9$. La statistica di Pearson è data da

$$Q_{10002} = \sum_{k=1}^{10} \frac{(N_k - 10002 \cdot 0.1)^2}{10002 \cdot 0.1} = \sum_{k=1}^{10} \frac{N_k^2}{10002 \cdot 0.1} - 10002 = 9.367726 \simeq 9.368$$

(N_k indica il numero di volte in cui compare la cifra k nelle prime 10002 posizioni). Poiché $948 \times 0.1 = 94.8 > 5$, se H_0 è vera, approssimativamente $Q_{10002} \sim \chi^2_9$.

1. Rifiutiamo H_0 a livello 5% se $Q_{10002} \geq \chi_9(0.95) = 16.92$. Poiché $9.368 < 16.92$ accettiamo a livello 5% l'ipotesi che nella rappresentazione decimale di π la posizione delle cifre $0, \ldots, 9$ dopo il punto decimale sia casuale.

2. Il p-value è il più piccolo livello per cui si rifiuta H_0 con i risultati empirici. Il p-value dei dati di questo test è $P_0(Q_{10002} > 9.368) \simeq 1 - F_{\chi^2_9}(9.368)$. Dalle tabelle vediamo che il nostro dato cade tra 8.34 e 11.34 corrispondenti ai quantili della fdr χ^2_9 di ordine 0.5 e 0.75, rispettivamente. Dunque il p-value cade fra 0.25 e 0.5. (Usando il software R troviamo che p-value = 0.404021): praticamente non rifiutiamo mai H_0. ∎

[1] Da Bickel, Peter J., and Doksum, Kjell A. (1977), "Mathematical statistics: Basic ideas and selected topics", Holden-Day Inc (San Francisco)

Esercizio 5.5 (Sezione Epifani) Si sono registrati i minuti di funzionamento prima di rovinarsi di due tipi di isolanti elettrici A e B sottoposti a una forte differenza di potenziale ottenendo i seguenti risultati:

Tipo A:	162	88.5	122.3	125	132	66	211.9		
Tipo B:	34.6	54	116.4	49	77.3	121.3	127.8	120.2	49.8

Verificate l'ipotesi che i due campioni casuali di osservazioni provengano dalla stessa funzione di ripartizione contro l'alternativa che l'isolante elettrico di tipo B smetta di funzionare prima di quello di tipo A.
Scegliete come livello di significatività del test $\alpha = 2.5\%$ e supponete che i dati siano tutti generati da modelli assolutamente continui. La vostra risposta cambia se scegliete un livello di α superiore a 2.5%?

SOLUZIONE Sia $\boldsymbol{X} = (X_1, \ldots, X_7)$ il campione di 7 osservazioni sui tempi di vita degli isolanti di tipo A e $\boldsymbol{Y} = (Y_1, \ldots, Y_9)$ il campione di 9 osservazioni sui tempi di vita degli isolanti di tipo B. Ci poniamo nell'ipotesi che $X_1, \ldots, X_7$ i.i.d. $\sim F$, $Y_1, \ldots, Y_9$ i.i.d. $\sim G$ e $\boldsymbol{X}, \boldsymbol{Y}$ indipendenti e F e G assolutamente continue. Traduciamo ipotesi nulla e alternativa in termini di dominanza stocastica nel seguente modo:

$$H_0 : F = G \text{ versus } H_1 : F < G$$

e impostiamo il corrispondente test unilatero non parametrico di Wilcoxon-Mann-Wintney.

Ordiniamo dalla più piccola alla più grande le osservazioni del campione riunito:

$$34.6y \;\; 49.0y \;\; 49.8y \;\; 54y \;\; 66x \;\; 77.3y \;\; 88.5x \;\; 116.4y \;\; 120.2y \;\; 121.3y \;\; 122.3x \;\; 125x \;\; 127.8y \;\; 132x \;\; 162x \;\; 211.9x$$

e calcoliamo la statistica di Wilcoxon-Mann-Wintney:

$$U = \text{“somma dei ranghi di } X\text{”} - 7 \times \frac{8}{2} = 80 - 7 \times \frac{8}{2} = 52.$$

Il quantile di U corrispondente a $m = 7$ e $n = 9$ di ordine 2.5% è $q_U(2.5\%) = 13$ e quindi $q_U(1 - 2.5\%) = 7 \times 9 - 13 = 50$. Rifiutiamo $H_0 : F = G$ a favore di $H_1 : F < G$ se $U > q_U(1 - 2.5\%)$; poiché $52 > 50$, rifiutiamo H_0 a livello 2.5%, cioè sembrerebbe che gli isolanti elettrici di tipo B siano meno resistenti. Se $\alpha > 2.5\%$ allora $q_U(1 - \alpha\%) < q_U(1 - 2.5\%)$ e $q_U(1 - \alpha\%) < U$; quindi per $\alpha > 2.5\%$ la decisione non cambia: continuiamo a rifiutare H_0. ∎