Politecnico di Milano
Temi d'esame di STATISTICA dell'AA 2008/2009
per allievi ING INF [2L], docenti I. Epifani, A.M. Pievatolo

© I diritti d'autore sono riservati. Ogni sfruttamento commerciale non autorizzato sarà perseguito.

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

Esercizio 1.1 Avete raccolto 100 misurazioni ripetute di una grandezza incognita μ e avete salvato le seguenti informazioni:

$$\sum_{i=1}^{100} x_i = 2371.2 \ , \quad \sum_{i=1}^{100} x_i^2 = 57682.1$$

con x_1 che rappresenta la prima misurazione, ..., e x_{100} la centesima. Nelle domande che seguono ipotizzate che le misurazioni siano gaussiane.

1. Costruite un test tale che sia al più pari a 2.5% la probabilità di commettere l'errore di prima specie di ritenere che la grandezza misuri più di 22.6 quando effettivamente è minore o uguale di 22.6. Sulla base delle statistiche fornite, che decisione prendete?
2. Calcolate il p-value del test.
3. Qual è il livello del seguente intervallo di confidenza per μ: $(22.960, \infty)$?

Abbiamo raccolto ulteriori 44 misurazioni di μ e, per le nuove 44 misurazioni, abbiamo ottenuto una media campionaria pari a 22.21 e un momento secondo campionario pari a 550.05.

4. Usate l'intero campione di 144 osservazioni per calcolare media e varianza campionarie.
5. Alla luce dei nuovi dati, e usando l'intero campione, qual è ora il livello dell'intervallo di confidenza $(22.960, \infty)$ di μ?

Soluzione Interpretiamo la grandezza incognita μ come il valore atteso incognito del campione casuale delle misurazioni aleatorie $X_1, \ldots, X_{100}$ che hanno comune distribuzione $\mathcal{N}(\mu, \sigma^2)$, con varianza σ^2 anche essa incognita. Pertanto, se μ è il vero valore della grandezza incognita, allora $\sqrt{n}(\bar{X} - \mu)/S \sim t_{99}$ e impostiamo un opportuno t-test unilatero sulla media.

1. Dobbiamo impostare un test di significatività $\alpha = 2.5\%$ per l'ipotesi nulla $H_0 : \mu \leq 22.6$ contro l'alternativa $H_1 : \mu > 22.6$. Segue che rifiutiamo H_0 a favore di H_1 a livello 2.5% se $10(\bar{X} - 22.6)/S \geq t_{99}(97.5\%)$ o, equivalentemente, se $22.6 \leq \bar{X} - 1.96S/10$, dal momento che $t_{99}(97.5\%) \simeq z_{97.5\%} \simeq 1.96$. Media e varianze campionarie valgono:

$$\bar{X} = \frac{\sum_{j=1}^{100} X_j}{100} = 23.712, \quad S^2 = \frac{\sum_{j=1}^{100} x_j^2 - 100 \times 23.712^2}{99} = 14.709$$

Quindi: $\bar{X} - 1.96S/10 = 23.712 - 1.96 \times \sqrt{14.709}/10 \simeq 22.960$ e rifiutiamo H_0 a livello 2.5%.

2. Il p-value del test è il più piccolo valore di α per cui si rifiuta H_0 ed è dato da

$$P_{22.6}\left(\frac{10(\bar{X} - 22.6)}{S} \geq 2.899\right) = 1 - F_{t_99}(2.899) \simeq 1 - \Phi(2.90) \simeq 1 - 0.9981 = 0.0019;$$

concludiamo che c'è una forte evidenza empirica contro l'ipotesi che μ valga al più 22.6.

3. Nel punto precedente abbiamo stabilito che con i dati forniti rifiutiamo un'ipotesi nulla del tipo : "$\mu \leq \mu_0$" a livello 2.5% se $\mu_0 \leq \bar{X} - 1.96S/10 = 22.960$. Per la dualità fra la verifica di ipotesi e gli intervalli di confidenza, allora siamo confidenti al 97.5% che $\mu > 23.712 - 1.96 \times \sqrt{14.709}/10 \simeq 22.960$.

4. $\bar{x}_{144} = \dfrac{100\bar{x}_{100} + 44\bar{x}_{44}}{144} = \dfrac{2371.2 + 22.21 \times 44}{144} \simeq 23.253$;

$$\sum_{j=1}^{144} x_j^2 = 57682.1 + 550.05 \times 44 = 81884.3 \text{ e quindi } s_{144}^2 = (81884.3 - 144 \times 23.253^2)/143 = 28.134$$

5. Poiché l'$IC(\mu)$ di livello $(1-\alpha)100\%$ è $(\bar{x}_{144} - z_{1-\alpha}s_{144}/\sqrt{144}, +\infty)$ e, per l'intero campione di 144 osservazioni, abbiamo:

$$\bar{x}_{144} - z_{1-\alpha}s_{144}/\sqrt{144} = 22.960 \quad \text{se e solo se} \quad z_{1-\alpha} = \frac{\sqrt{144}(23.253 - 22.960)}{\sqrt{28.134}} \simeq 0.663,$$

allora, $1 - \alpha \simeq \Phi(0.66) \simeq 0.7454$. Notate che il livello di confidenza si è abbassato che è ciò che ci aspettavamo dato che la media campionaria si è ridotta... ∎

Esercizio 1.2 Sia $X_1, \dots, X_n$ un campione casuale estratto da una popolazione di densità

$$f(x;\theta) = \frac{1}{4\theta^3} \exp\left\{ -\frac{|x|}{2\theta^3} \right\}$$

con θ parametro positivo incognito.

1. Determinate la distribuzione di $|X_1|$. Vi riconoscete una "densità notevole"?
2. Determinate lo stimatore di massima verosimiglianza $\hat{\kappa}$ della caratteristica $\kappa = 2\theta^3$ e la sua distribuzione (esatta).
3. Lo stimatore $\hat{\kappa}$ è non distorto per κ? È consistente per κ? Giustificate rigorosamente la risposta.
4. Sia $n = 4$. Verificate che il test più potente di livello 5% per il problema di ipotesi $H_0 : \theta = 2$ contro $\theta = 1$ ha regione critica $\mathcal{G} = \{(x_1, \dots, x_n) : \hat{\kappa} \le 5.466\}$.
5. Calcolate la potenza del test del punto 4. o stabilite un intervallo in cui cade.

Soluzione

1. Sia $Y = |X_1|$. Se $y \le 0$ allora $F_{Y,\theta}(y) = 0$; per $y > 0$:

$$F_{Y,\theta}(y) = P_\theta(|X_1| \le y) = P_\theta\left(-y \le X_1 \le y\right) = 2F_{X_1,\theta}(y) - 1$$

da cui segue che la densità di Y è

$$f_Y(y,\theta) = 2f(y,\theta)\mathbf{1}_{(0,+\infty)}(y) = \frac{2}{4\theta^3} \exp\left(-\frac{y}{2\theta^3}\right) \mathbf{1}_{(0,+\infty)}(y)$$

cioè Y ha densità esponenziale di parametro $\kappa = 2\theta^3$.

2. Per quanto stabilito al punto 1., senza perdere in generalità, possiamo lavorare con il campione casuale $Y_1, \dots, Y_n$ definito da $Y_i = |X_i|,\ i = 1, \dots, n$, con comune densità esponenziale di parametro κ. Si ottiene che

$$\hat{\kappa} = \bar{Y} = \frac{\sum_{j=1}^n |X_j|}{n}$$

Inoltre, dalle proprietà della famiglia delle densità gamma deduciamo che $\hat{\kappa} = \sum_{j=1}^n |X_j|/n \sim \Gamma(n, \kappa/n)$.

3. Le proprietà di non distorsione e consistenza di $\hat{\kappa}$ per la caratteristica κ discendono dalla rappresentazione di $\hat{\kappa}$ come media campionaria delle Y_j che hanno media teorica κ e varianza teorica (finita) κ^2 insieme al fatto che $\mathrm{E}(\bar{Y}) = \mathrm{E}(Y_1) = \kappa$ e $\mathrm{Var}(\bar{Y}) = \mathrm{Var}(Y_1)/n = \kappa^2/n \to 0$, per $n \to \infty$.

4. Il test più potente di livello 5% per il problema di ipotesi $H_0 : \theta = 2$ contro $\theta = 1$ è dettato dal Lemma di Neyman Pearson e ha regione critica della forma: $\mathcal{G} = \{(x_1, \dots, x_n) : L_2(x_1, \dots x_n)/L_1(x_1, \dots, x_n) \le \delta\}$. Abbiamo che

$$\frac{L_2(x_1, \dots x_n)}{L_1(x_1, \dots x_n)} = \frac{1}{8^n} \mathrm{e}^{\frac{7}{16}\sum_{j=1}^n |x_j|} = \frac{1}{8^n} \mathrm{e}^{\frac{7n}{16}\hat{\kappa}}$$

e quindi il rapporto di verosimiglianza L_2/L_1 è crescente in $\hat{\kappa}$. Segue che $\mathcal{G} = \{(x_1, \dots, x_n) : \hat{\kappa} \le q\}$ con q tale che $P_2(\mathcal{G}) = 5\%$. D'altro canto, se $\theta = 2$ e $n = 4$ allora $\hat{\kappa} \sim \Gamma(4, 2 \times 2^3/4) = \Gamma(4, 4)$ cosicché $q = 2\chi^2_8(5\%) = 2 \times 2.733 = 5.466$.

5. Se $\theta = 1$ e $n = 4$ allora $\hat{\kappa} \sim \Gamma(4, 2 \times 1^3/4) = \Gamma(4, 1/2)$ e

$$\pi = P_1(\hat{\kappa} \le 5.466) = P(4\hat{\kappa} \le 4 \times 5.466) = F_{\chi^2_8}(21.864) \in (F_{\chi^2_8}(20.090), F_{\chi^2_8}(21.955)) = (0.99, 0.995).$$

∎

Esercizio 1.3 L'azienda PPP qualche mese fa ha cambiato il suo centro di assistenza informatica passando dal centro di assistenza yyy al centro di assistenza xxx. I dati che seguono si riferiscono ai tempi di riparazione, espressi in giorni, degli ultimi 7 guasti di PC riparati da yyy ($y_j, j = 1, \ldots, 7$) e dei primi 6 guasti di PC riparati da xxx ($x_i, i = 1, \ldots, 6$):

y_j	6.99	8.60	10.50	10.60	12.33	14.13	16.49
x_i	11.24	7.07	9.32	8.37	5.64	5.87	

1. Secondo voi, cambiando centro di assistenza, i tempi di riparazione delle apparecchiature informatiche si sono ridotti? Per rispondere costruite un opportuno test di livello 5%.

2. Costruite la funzione di ripartizione empirica del campione dei tempi di riparazione del nuovo centro di assistenza xxx.

3. Secondo voi, il modello di Weibull descritto dalla densità di probabilità $f_0(x) = \dfrac{3x^2}{10^3} e^{-\left(\frac{x}{10}\right)^3} \mathbf{1}_{(0,\infty)}(x)$ si adatta ai 6 tempi di riparazione del nuovo centro di assistenza xxx? Per rispondere costruite un opportuno test di significatività 10%.

Soluzione Introduciamo le variabili aleatorie X, Y con $X =$ "tempo di riparazione di un guasto al PC eseguita dal centro di assistenza xxx" e $Y =$ "tempo di riparazione di un guasto al PC eseguita dal centro yyy".

1. I dati sono senza ripetizione, i campioni sono piccoli e nessuna ipotesi è formulata sul modello parametrico generatore di questi dati. Quindi, impostiamo il test di omogeneità non parametrico di Wilcoxon-Mann-Wintney per verificare l'ipotesi nulla H_0 :"I tempi di riparazione non si sono ridotti" contro l'alternativa H_1 : " I tempi di riparazione si sono ridotti", cioè H_0 : "$F_X(w) \leq F_Y(w), \forall w \in \mathbb{R}''$ contro H_1 : "$F_X(w) \geq F_Y(w), \forall w \in \mathbb{R}$ *e* $F_X(w) > F_Y(w)$, *per qualche* w''. Usiamo la statistica T_X data dalla somma dei ranghi dei tempi di riparazione $X_1, \ldots, X_6$: $T_X = 1 + 2 + 4 + 5 + 7 + 10 = 29$. Infatti tutte le osservazioni ordinate dalla più piccola alla più grande risultano:

Ranghi	1	2	3	4	5	6	7	8	9	10	11	12	13
y_j			6.99			8.60		10.50	10.60		12.33	14.13	16.49
x_i	5.64	5.87		7.07	8.37		9.32			11.24			

Confrontiamo ora 29 con il quantile di ordine $w_{0.05} = 30$ della statistica T_X con $m = 6$, $n = 7$. Essendo $29 < 30$, a livello 5% rifiutiamo l'ipotesi nulla H_0 che i tempi di riparazione NON si siano ridotti a favore dell'ipotesi che i tempi di riparazione del centro xxx siano inferiori a quelli del centro yyy.

2. La f.d.r. empirica $\hat{F}_6$ associata al campione dei 6 tempi di riparazione X è

x	$(-\infty, 5.64)$	$[5.64, 5.87)$	$[5.87, 7.07)$	$[7.07, 8.37)$	$[8.37, 9.32)$	$[9.32, 11.24)$	$[11.24, \infty)$
$\hat{F}_6$	0	1/6	1/3	1/2	2/3	5/6	1

3. Abbiamo un numero "piccolo" di dati (6) non raggruppati e il campione proviene da una f.d.r. 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 f.d.r. corrispondente a $f_0(x)$ che, per $x > 0$, è $F_0(x) = \displaystyle\int_0^x \frac{3t^2}{10^3} e^{-\left(\frac{t}{10}\right)^3}\, dt = 1 - e^{-\left(\frac{x}{10}\right)^3}$. Pertanto:

$F_0(x_i)$	0.164	0.183	0.298	0.444	0.555	0.758
$\lvert\hat{F}_6(x_i) - F_0(x_i)\rvert$	0.003	0.150	0.202	0.223	0.278	0.242
$\lvert\hat{F}_6(x_{i-1}) - F_0(x_i)\rvert$	0.164	0.016	0.035	0.056	0.112	0.075

Segue che $D_6 := \sup_{x \in \mathbb{R}} |\hat{F}_6(x) - F_0(x)| = 0.278$. Inoltre, dalle tavole dei quantili della statistica di Kolmogorov-Smirnov con $n = 6$ scopriamo che $q_{D_6}(1 - 0.1) = 0.4680$: quindi non rifiutiamo H_0 a livello 10%, cioè il modello di Weibull f_0 si adatta ai dati con significatività 10%. ∎

© I diritti d'autore sono riservati. Ogni sfruttamento commerciale non autorizzato sarà perseguito.

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

Esercizio 2.1 Per verificare la Qualità di Servizio di un Sistema Software Aperto di acquisti telematici, abbiamo deciso di registrare le proporzioni giornaliere $X_1, \dots, X_n$ di richieste fallite dei prossimi n giorni. Astraendo, possiamo considerare $X_1, \dots, X_n$ come un campione casuale estratto da una popolazione di 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.

1. Determinate lo stimatore dei momenti di θ.
2. Determinate lo stimatore di massima verosimiglianza di θ.
3. Verificate se lo stimatore di massima verosimiglianza di θ è efficiente.
4. Determinate lo stimatore di massima verosimiglianza della proporzione attesa $\mathrm{E}_\theta(X)$ di richieste giornaliere fallite. Quindi dimostrate che uno stimatore efficiente per $\mathrm{E}_\theta(X)$ NON esiste. (*Giustificate rigorosamente la risposta.*)
5. Monitorando il sistema software per $n = 4$ giorni registriamo i seguenti risultati: il primo giorno falliscono 4 su 480 richieste arrivate, il secondo 5 su 450, il terzo 3 su 300 e il quarto 6 su 500. Qual è la stima di θ basata sul metodo dei momenti? E quella basata sul metodo di massima verosimiglianza?

Soluzione

1. $\mathrm{E}_\theta(X_1) = \int_0^1 x \frac{1}{\theta} x^{\frac{1}{\theta}-1}\, dx = \frac{1}{\theta} \left[\frac{x^{1/\theta+1}}{1/\theta+1} \right]_0^1 = \frac{1/\theta}{1/\theta+1} = \frac{1}{\theta+1}$

ed $\mathrm{E}_\theta(X_1) = \bar{X}$ se e solo se $\theta = 1/\bar{X} - 1$; segue che $\widehat{\theta}_{mom} = 1/\bar{X} - 1$ è lo stimatore dei momenti di θ.

2. Studiamo la funzione di verosimiglianza del campione $X_1, \dots, X_n$:

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

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

$$\frac{\partial \log L_\theta(x_1, \dots, x_n)}{\partial \theta} = \frac{n}{\theta^2} \left(-\frac{\sum_{j=1}^n \log x_j}{n} - \theta \right) \geq 0 \tag{1}$$

se e solo se $\theta \leq -(1/n) \sum_{j=1}^n \log x_j$; infine osserviamo che $-\sum_{j=1}^n \log X_j > 0$ poiché $P(0 < X_j < 1) = 1$. Segue che $\widehat{\theta}_{ML} = -\frac{\sum_{j=1}^n \log X_j}{n}$ è stimatore ML per θ.

3. 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 $\widehat{\theta}_{ML}$ e θ. Ma (1) è CNS affinchè la varianza di $\hat{\theta}$ raggiunga il confine di Fréchet-Cramér-Rao, cioè affinché $\widehat{\theta}_{ML}$ sia stimatore efficiente.

4. Poiché $\kappa = \mathrm{E}_\theta(X_1) = 1/(1+\theta)$, allora, $\widehat{\kappa}_{ML} = 1/(1+\widehat{\theta}_{ML})$. Inoltre, deriviamo da (1) che

$$\frac{\partial \log L_\theta(x_1, \dots, x_n)}{\partial \theta} = \frac{n}{\theta^2} ((1 + \widehat{\theta}_{ML}) - (1+\theta)) = \frac{n}{\theta^2} \left(\frac{1}{\widehat{\kappa}_{ML}} - \frac{1}{\kappa} \right):$$

cioè $\partial \log L_\theta / \partial \theta$ è funzione lineare di $\widehat{\kappa}_{ML}^{-1}$, quindi non può esserlo di $\widehat{\kappa}_{ML}$: non essendo soddisfatta una CNS per l'efficienza, allora $\widehat{\kappa}_{ML}$ non è stimatore efficiente di κ. D'altro canto se uno stimatore efficiente per κ esiste, allora esso è necessariamente stimatore ML: rimane così stabilito che NON esiste nessun stimatore efficiente di κ.

5. Deriviamo dai dati forniti il seguente campione: $x_1 = 4/480$, $x_2 = 5/450$, $x_3 = 3/300$, $x_4 = 6/500$; quindi $\bar{x} = 373/36000 \simeq 0.0104$ e $\widehat{\theta}_{mom} = 35627/373 \simeq 95.515$; $\widehat{\theta}_{ML} \simeq 4.579$. ∎

Esercizio 2.2 Abbiamo registrato i seguenti tempi di riparazione, espressi in minuti, dei guasti occorsi a due tipi di macchine fotocopiatrici A e B:

macchina A:	32	84	37	42	78	62	59
macchina B:	39	111	55	106	90	87	

Uno studio passato di questo tipo di dati ci permette di dire che la distribuzione è per entrambi i guasti *lognormale*, cioè i logaritmi (naturali) dei tempi di riparazione hanno distribuzione gaussiana, e che le varianze dei logaritmi dei tempi di riparazione dei guasti occorsi alle macchine A e B sono uguali.

Sia σ^2 la comune varianza dei logaritmi dei tempi di riparazione dei guasti occorsi ad A e B

1. Proponete uno stimatore non distorto di σ^2 e calcolatene il valore in corrispondenza dei dati forniti.
2. Costruite un intervallo bilatero di confidenza $\gamma = 90\%$ per σ^2.
3. Verificate a livello 5% l'ipotesi nulla $H_0 : \sigma^2 = 0.20$ contro l'alternativa $H_1 : \sigma^2 \neq 0.20$.

Infine

4. Costruite un test per verificare l'ipotesi che le distribuzioni (entrambe lognormali) dei tempi di riparazione dei guasti occorsi ad A e B siano DEL TUTTO uguali contro l'alternativa che siano diverse.

Soluzione Per risolvere l'esercizio occorrono i campioni dei logaritmi dei tempi di riparazione dei guasti occorsi ad A e B, con le loro medie e varianze campionarie. Siano $X_1, \dots, X_7$ il campione dei logaritimi dei tempi di riparazione di A e $Y_1, \dots, Y_6$ il campione dei logaritimi di quelli di B:

$X = \ln A$:	3.466	4.431	3.611	3.738	4.357	4.127	4.078
$Y = \ln B$:	3.664	4.710	4.007	4.663	4.500	4.466	

Abbiamo:

$\bar{x} = 3.973,\ \bar{y} = 4.335,\ S_x^2 = \dfrac{\sum_{i=1}^7 x_i^2 - 7\bar{x}^2}{6} = \dfrac{111.305 - 7 \times 3.973^2}{6} \simeq 0.139,\ S_y^2 = \dfrac{\sum_{i=1}^6 y_i^2 - 6\bar{y}^2}{6} = \dfrac{113.604 - 6 \times 4.335^2}{5} \simeq$ 0.170.

1. Stimiamo σ^2 con la varianza pooled S_p^2 dei logaritmi dei tempi di riparazione di A e B:

$$S_p^2 = \frac{0.139 \times 6 + 0.170 \times 5}{7+6-2} \simeq 0.153, \quad (\sqrt{S_p^2} \simeq 0.391)$$

2. La variabile aleatoria $S_p^2(7+6-2)/\sigma^2$ è una quantità pivotale; infatti essa dipende dalle osservazioni e dal parametro σ^2, ma la sua distribuzione è χ^2_{11}, quindi è indipendente da tutti i parametri incogniti. Vale che

$$P^2_\sigma\left(\chi^2_{11}(5\%) < \frac{S_p^2(7+6-2)}{\sigma^2} < \chi^2_{11}(95\%)\right) = 0.90$$

quindi un IC(σ^2) di confidenza 90% è

$$\left(\frac{S_p^2(7+6-2)}{\chi^2_{11}(95\%)}, \frac{S_p^2(7+6-2)}{\chi^2_{11}(5\%)}\right) = \left(\frac{0.153 \times 11}{19.675}, \frac{0.153 \times 11}{4.575}\right) = (0.0856, 0.3679) \tag{2}$$

3. Poiché $0.20 \in (0.0856, 0.3679)$, allora, per la dualità fra IC e test di ipotesi, accettiamo $H_0 : \sigma^2 = 0.20$ contro l'alternativa $H_1 : \sigma^2 \neq 0.20$, per ogni livello $\alpha \leq 10\%$. Infatti l'IC(σ^2) della stessa forma di quello costruito in (2) ma di confidenza $\gamma_1 = 1 - \alpha_1$ con $\alpha_1 < 10\%$ contiene strettamente $(0.0856, 0.3679)$.

4. Le due distribuzioni lognormali dei tempi di riparazione di A e B sono interamente uguali se e solo se i logaritmi dei tempi di riparazione di A e B che costituiscono due campioni gaussiani indipendenti di stessa varianza hanno anche la stessa media. Dobbiamo dunque costruire un test t di confronto delle medie per campioni gaussiani indipendenti con comune varianza. Le ipotesi nulla e alternativa sono rispettivamente: $H_0 : \mu_X = \mu_Y$ contro $H_0 : \mu_X \neq \mu_Y$. La statistica test è $\left|\dfrac{\bar{X} - \bar{Y}}{\sqrt{S_p^2\left(\frac{1}{7} + \frac{1}{6}\right)}}\right| \simeq |-1.6635| = 1.6635$ e rifiutiamo H_0 per ogni $\alpha \geq 2(1 - F_{t_{11}}(1.6635))$. Dalle tavole della distribuzione t di student con 11 gradi di libertà abbiamo: $0.90 = F_{t_{11}}(1.363) < F_{t_{11}}(1.6635) < F_{t_{11}}(1.796)) = 0.95$; segue che il p-value cade nell'intervallo $(0.1, 0.2)$. Sicuramente accetto H_0 per ogni $\alpha \leq 10\%$ e la rifiuto per ogni $\alpha \geq 20\%$. (p-value effettivo= 0.1244). ∎

Esercizio 2.3 I pezzi meccanici cilindrici prodotti da una certa linea produttiva devono rispettare le seguenti specifiche di lunghezza e diametro: la lunghezza L, espressa in cm, deve cadere nell'intervallo $[19.70, 20.30]$ e il diametro D, in cm, nell'intervallo $[1.98, 2.26]$. Abbiamo controllato un lotto di 2000 pezzi e catalogato ogni pezzo sulla base di lunghezza e diametro, ottenendo i risultati riportati nella seguente tabella.

	$D < 1.98$	$1.98 \leq D \leq 2.26$	$D > 2.26$	
$L < 19.70$	26	150	8	
$19.70 \leq L \leq 20.30$	124	1320	160	
$L > 20.30$	10	186	16	

1. Verificate con un opportuno test se lunghezza e diametro dei pezzi meccanici sono indipendenti.
2. Verificate con un opportuno test se le lunghezze reali dei pezzi prodotti sono variabili aleatorie gaussiane di media 20 cm e varianza $0.018\,\mathrm{cm}^2$.
3. Costruite un intervallo di confidenza bilatero di livello approssimato $\gamma = 95\%$ per la percentuale (attesa) di pezzi meccanici cilindrici che rispettano le specifiche di lunghezza e diametro.

Soluzione Completiamo la tabella dei dati, calcolando nell'ultima riga e nell'ultima colonna le numerosità marginali degli intervalli di valori in cui cadono le variabili aleatorie L e D:

	$D < 1.98$	$1.98 \leq D \leq 2.26$	$D > 2.26$	
$L < 19.70$	26 (N_{11})	150	8	184 (N_{1L})
$19.70 \leq L \leq 20.30$	124	1320	160	1604
$L > 20.30$	10	186	16	212
	160 (N_{1D})	1656	184	2000 (n)

1. Avendo dati raggruppati e numerosi (2000 pezzi), possiamo impostare un test asintotico chiquadrato di indipendenza per verificare l'ipotesi H_0: "L, D sono indipendenti" contro l'alternativa H_1: "L, D non sono indipendenti". La statistica test è

$$Q_1 = \sum_{i,j=1}^{3} \frac{\left(N_{ij} - \frac{N_{iL}N_{jD}}{n}\right)^2}{\frac{N_{iL}N_{jD}}{n}} = 2000\left(\frac{26^2}{160 \times 184} + \frac{150^2}{1656 \times 184} + \cdots + \frac{16^2}{184 \times 212}\right) - 2000 \simeq 18.74$$

e rifiutiamo H_0 per ogni $\alpha \geq 1 - F_{\chi^2_4}(18.74)$; infatti avendo 2000 pezzi in totale e più di 5 in ogni casella, allora, la f.d.r. $\chi^2_{(3-1)(3-1)}$ è una buona approssimazione della f.d.r. di Q_1. Dalle tavole abbiamo: $1 - F_{\chi^2_4}(18.74) < 1 - F_{\chi^2_4}(18.467) = 1 - 99.9\%$, quindi rifiutiamo H_0 per ogni $\alpha \geq 0.1\%$: concludiamo che c'è forte evidenza empirica contro l'ipotesi di indipendenza stocastica di L, D.

2. Impostiamo un test asintotico chiquadrato di buon adattamento per verificare H_0 :"$L \sim \mathcal{N}(20, 0.018)$" contro l'alternativa: H_1 :"$L \not\sim \mathcal{N}(20, 0.018)$". Servono le seguenti quantità

	$L < 19.70$	$19.70 \leq L \leq 20.30$	$L > 20.30$
N_{iL}	184	1604	212
p_{0i}	0.0125	0.975	0.0125
np_{0i}	25	1950	25

dove: $p_{01} = P_{H_0}(L < 19.70) = \Phi\left(\frac{19.70-20}{\sqrt{0.018}}\right) = \Phi(-\sqrt{5}) \simeq 1 - \Phi(2.24) \simeq 0.0125$, $p_{03} = P_{H_0}(L > 20.30) = 1 - \Phi(\sqrt{5}) = p_{10}$ e $p_{02} = 1 - p_{01} - p_{02} = 0.975$. La statistica test è

$$Q_2 = \sum_{i=1}^{3} \frac{(N_{iL} - np_{0i})^2}{np_{0i}} = \frac{184^2}{25} + \frac{1604^2}{1950} + \frac{212^2}{25} - 2000 \simeq 2471.4$$

Essendo la numerosità del campione grande e $np_{01} > 5$ per ogni $i = 1, 2, 3$, allora una buona approssimazione della f.d.r. di Q_2 è la f.d.r. esponenziale di parametro 2 e rifiutiamo H_0 per ogni $\alpha \geq \mathrm{e}^{-2471.4/2} \simeq 0$: c'è una fortissima evidenza empirica contro H_0.

3. Sia $\theta 100\%$ la percentuale attesa di pezzi meccanici cilindrici che rispettano le specifiche di lunghezza e diametro. Segue dai risultati di inferenza asintotica per campioni casuali di va bernoulliane di parametro incognito θ che $\widehat{\theta} = 1320/2000 = 0.66$ è una stima di θ e un $IC(\theta)$ asintotico di confidenza 95% ha estremi $\widehat{\theta} \mp z_{97.5\%}\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{2000}}$. Pertanto un $IC(\theta 100\%)$ di confidenza 95% è $(63.9\%, 68.1\%)$. ∎

© I diritti d'autore sono riservati. Ogni sfruttamento commerciale non autorizzato sarà perseguito.

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

Esercizio 3.1 Sia X_1 un'unica osservazione estratta da una popolazione di densità

$$f(x,\theta) = [2\theta x + 2(1-\theta)(1-x)]\,\mathbf{1}_{(0,1)}(x),$$

con $\theta \in [0,1]$ parametro incognito.

1. Costruite il test più potente di livello $\alpha = 1.2\%$ per verificare l'ipotesi nulla $H_0 : \theta = 0.1$ contro $H_1 : \theta = 0.8$.
2. Calcolate la probabilità di prendere la corretta decisione, se effettivamente $\theta = 0.8$.
3. Se $X_1 = 1/4$, quanto vale il p-value del test? Cosa si può concludere?
4. Calcolate la funzione di verosimiglianza nei seguenti tre casi $X_1 = 1/4$, $X_1 = 1/2$, $X_1 = 5/6$, discutete esistenza e unicità dello stimatore di massima verosimiglianza di θ (sempre in questi tre casi) e, se ne esiste uno, determinatelo.

Soluzione

1. Dobbiamo costruire il test dettato dal Lemma di Neyman Pearson che ha regione critica della forma $\mathcal{G} = \Big\{x : \frac{L_{0.1}(x)}{L_{0.8}(x)} \le \delta\Big\}$. Abbiamo che $L_{0.1}(x) = f(x, 0.1) = 1.8 - 1.6x$, $L_{0.8}(x) = f(x, 0.8) = 0.4 + 1.2x$ e

$$\frac{L_{0.1}(x)}{L_{0.8}(x)} = \frac{1.8 - 1.6x}{0.4 + 1.2x}\ ,$$

che è funzione decrescente di x per $x \in (0,1)$, cosicché

$$\frac{L_{0.1}(x)}{L_{0.8}(x)} \le \delta \text{ con } \delta \text{ t.c. } P_{0.1}\Big(\frac{L_{0.1}(x)}{L_{0.8}(x)} \le \delta\Big) = 1.2\% \quad \text{se e solo se } x \ge k \text{ con } k \text{ t.c. } P_{0.1}(X_1 \ge k) = 0.012.$$

Ma, per $0 < k \le 1$, $P_{0.1}(X_1 \ge k) = \int_k^1 (1.8 - 1.6x)\,dx = 0.8k^2 - 1.8k + 1$ e $0.8k^2 - 1.8k + 1 = 0.012$ se e solo se

$$k = \frac{1.8 - \sqrt{1.8^2 - 4 \times 0.8 \times (1 - 0.012)}}{2 \times 0.8} = \frac{9 - \sqrt{1 + 80 \times 0.012}}{8} = 0.95$$

(Escludiamo la soluzione $k = \frac{1.8+\sqrt{1.8^2-4\times0.8\times(1-0.012)}}{2\times0.8}$, perché maggiore di 1). Concludiamo che la regione di rifiuto del test più potente di livello 1.2% per H_0 contro H_1 è

$$\mathcal{G} = \Big\{x \in (0,1) :\ x \ge 0.95\Big\}.$$

2. Dobbiamo calcolare la potenza del test: $\pi(0.8) = P_{0.8}(X_1 \in \mathcal{G}) = P_{0.8}(X_1 \ge 0.95) = \int_{0.95}^1 (0.4 + 1.2x)\,dx = 0.4 - 0.4 \times 0.95 + 0.6 - 0.6 \times 0.95^2 = 0.0785$.

3. Il p-value del test se $X_1 = 1/4$ è dato da

$$P_{0.1}(X_1 \ge \frac{1}{4}) = \int_{\frac{1}{4}}^1 (1.8 - 1.6x)\,dx = \frac{0.8}{4^2} - \frac{1.8}{4} + 1 = 0.6$$

Concludiamo che c'è forte evidenza empirica contro H_0.

4. Se $X_1 = 1/4$, allora $\theta \mapsto L_\theta(1/4) = (1.5 - \theta)$ è funzione strettamente decrescente sull'intervallo $[0,1]$: quindi $\widehat{\theta}_{ML}$ esiste e vale 0; se invece $X_1 = 5/6$, allora $\theta \mapsto L_\theta(5/6) = \dfrac{1+4\theta}{3}$ è funzione strettamente crescente sull'intervallo $[0,1]$: quindi $\widehat{\theta}_{ML}$ esiste e vale 1. Infine, se $X_1 = 1/2$, la funzione $\theta \mapsto L_\theta(1/2)$ è costante e pari a 1 sull'intervallo $[0,1]$: quindi ogni punto dell'intervallo $[0,1]$ massimizza L_θ. ∎

Esercizio 3.2 Abbiamo estratto a caso un campione di sette coppie di gemelli maschio, femmina, quindicenni che ricevono una paghetta settimanale di 40 euro a testa e abbiamo chiesto a ciascuno quanto ha speso nell'ultima settimana. I dati ottenuti, espressi in euro, sono sintetizzati nelle seguenti statistiche:

$$\sum_{j=1}^{7} f_j = 225.5, \ \sum_{j=1}^{7} m_j = 243.1, \ \sum_{j=1}^{7} f_j^2 = 8565.99, \ \sum_{j=1}^{7} m_j^2 = 10315.17, \ \sum_{j=1}^{7} f_j m_j = 8147.90$$

dove f_j fornisce quanto ha speso nell'ultima settimana la femmina e m_j il maschio.
Per rispondere a tutte le domande che seguono ipotizzate che i dati bivariati $(f_1, m_1), \dots, (f_7, m_7)$ siano gaussiani.

1. Chi spende di più fra maschi e femmine? Rispondete usando un opportuno test di livello $\alpha = 5\%$ tale che commettete un errore di primo tipo se concludete erroneamente che i maschi spendono più delle femmine.

2. Quanto spende un gemello maschio dipende da quanto spende la gemella femmina? Rispondete usando un opportuno test di livello $\alpha = 5\%$.

3. Determinate la probabilità che per una coppia di gemelli scelta a caso il gemello maschio spenda più della gemella femmina e stimate questa probabilità con uno stimatore che sia funzione delle statistiche precedenti.

Soluzione Introduciamo le variabili aleatorie F, M definite da: F = euro della paghetta settimanale spesi dalla gemella femmina, M = euro della paghetta settimanale spesi dal gemello maschio.

1. Per rispondere usiamo il campione dei 7 dati accoppiati (f_j, m_j), $j = 1, ..., 7$, provenienti da una popolazione bivariata gaussiana e costruiamo un test t di confronto delle medie μ_F, μ_M di questi dati gaussiani accoppiati. Dal testo ricaviamo che l'errore di prima specie consiste nel rifiutare l'ipotesi nulla che le femmine spendano almeno quanto i maschi, cioè $\mu_M \leq \mu_F$, quando questa ipotesi è vera. Pertanto, dobbiamo verificare $H_0 : \mu_M - \mu_F \leq 0$ vs $H_1 : \mu_M - \mu_F > 0$. Rifiutiamo H_0 a livello α se $T = \sqrt{7}(\bar{M} - \bar{F})/S_D > t_6(1-\alpha)$, dove $t_6(1-\alpha)$ è il quantile di ordine $1-\alpha$ della f.d.r. t di Student con 6 gradi di libertà e S_D^2 è la varianza campionaria del campione delle differenze $(m_j - f_j)$; abbiamo

$$\begin{aligned}
\bar{m} &= \frac{243.1}{7} \simeq 34.72857; \ \bar{f} = \frac{225.5}{7} \simeq 32.21429; \ \bar{m} - \bar{f} \simeq 2.51428 \\
S_F^2 &= \frac{\sum_{j=1}^{7} f_j^2 - 7\bar{f}^2}{6} = \frac{8565.99 - 7 \times 32.21429^2}{6} \simeq 216.9444 \\
S_M^2 &= \frac{\sum_{j=1}^{7} m_j^2 - 7\bar{m}^2}{6} = \frac{10315.17 - 7 \times 34.72857^2}{6} \simeq 312.1092 \\
\frac{\sum_{j=1}^{7} m_j f_j - 7\bar{f}\bar{m}}{6} &= \frac{8147.90 - 225.5 \times 243.7/7}{6} \simeq 49.54643 \\
S_D^2 &= S_F^2 + S_M^2 - 2 \times \frac{\sum_{j=1}^{7} m_j f_j - 7\bar{f}\bar{m}}{6} \simeq 429.96 \\
T &= \sqrt{7}\frac{\bar{M} - \bar{F}}{S_D} = \frac{\sqrt{7} \times 2.514}{\sqrt{429.96}} \simeq 0.3208
\end{aligned}$$

Se $\alpha = 0.05$ allora $t_6(1-\alpha) = 1.943$. Quindi al livello 5% non rifiutiamo H_0 e propendiamo a credere che le femmine spendano almeno quanto i maschi.

2. I dati bivariati sono gaussiani, impostiamo un test t di livello $\alpha = 5\%$ sul coefficiente ρ per il seguente problema di ipotesi: $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$. Il coefficiente di correlazione campionario R vale $R = \dfrac{49.54643}{\sqrt{216.9444 \times 312.1092}} \simeq$ 0.1904 La statistica test $\dfrac{R}{\sqrt{1-R^2}}\sqrt{n-2}$ vale 0.4337 e, sotto H_0, ha distribuzine t di student con 5 gradi di libertà. Poiché $|0.4337| < t_5(97.5\%) = 2.571$, allora NON rifiutiamo l'ipotesi H_0 che i gemelli maschio e femmina spendano in modo indipendente uno dall'altro. Inoltre, il p-value di questo test è $p = 2(1 - F_{t_5}(0.4337)) \simeq 2(1 - F_{t_5}(0.457)) = 2(1 - 66.7\%) = 66.6\%$: non c'è assolutamente evidenza empirica contro H_0.

3. Poiché M, F sono congiuntamente gaussiane, allora la differenza $M - F$ è va gaussiana e la probabilità cercata è data da:

$$P(M > F) = P(M - F > 0) = 1 - \Phi\left(\frac{0 - (\mu_M - \mu_F)}{\sqrt{\sigma^2_{M-F}}}\right)$$

dove σ^2_{M-F} indica la varianza di $M - F$. Poiché per un campione gaussiano lo stimatore ML della media è la media campionaria e quello della varianza (quando la media è incognita) è $(n-1)S^2/n$, allora lo stimatore ML di $P(M > F)$ è dato da $1 - \Phi\left(\frac{0-(\bar{M}-\bar{F})}{\sqrt{6S_D^2/7}}\right) \simeq 1 - \Phi(-0.13) = \Phi(0.13) \simeq 0.5517$. ∎

Esercizio 3.3 Abbiamo i seguenti dati raggruppati sui tempi di vita, espressi in anni, di 500 componenti di un certo tipo:

A_k	$(0, 1.28]$	$(1.28, 2.93]$	$(2.93, 4.18]$	$(4.18, 4.84]$	$(4.84, 6.52]$	$(6.52, 7.66]$	$(7.66, 9.32]$	> 9.32
N_k	4	41	55	50	125	120	105	0

dove A_k indica l'intervallo di vita, espresso in anni, e N_k il numero di componenti con durata che cade nell'intervallo A_k.

Ci chiediamo se la famiglia di densità *beta* data da

$$f(x, \beta) = \begin{cases} \frac{12x^2(\beta-x)}{\beta^4} & 0 < x < \beta \\ 0 & \text{altrove} \end{cases}, \qquad \beta > 0 \tag{3}$$

si adatta ai dati raggruppati forniti. A tal fine:

1. calcolate $\mathrm{E}_\beta(X)$ quando X ha densità $f(x, \beta)$ del tipo (3);
2. fornite uno stimatore di β usando i precedenti dati raggruppati;
3. calcolate la funzione di ripartizione di X quando X ha densità $f(x, \beta)$ del tipo (3);
4. valutate con un opportuno test la bontà di adattamento del modello beta in (3) ai dati raggruppati forniti. (*Se non siete riusciti a risolvere il punto 1., scegliete voi un valore per il parametro β ed eseguite un opportuno test.*)

Soluzione

1. $\mathrm{E}_\beta(X) = \int_0^\beta x \frac{12x^2(\beta - x)}{\beta^4}\, dx = \frac{3x^4(\beta - x)}{\beta^4}\Big|_0^\beta + \int_0^\beta \frac{3x^4}{\beta^4}\, dx = \frac{3x^5}{\beta^4}\Big|_0^\beta = \frac{3\beta}{5}$

2. Stimiamo β con il metodo dei momenti e usando la media campionaria (espressa in anni) dei dati raggruppati. Poiché nel primo intervallo cadono poche osservazioni lo accorpiamo con il secondo intervallo e per le nuove 6 classi i valori centrali c_j sono: $1.465, 3.555, 4.510, 5.680, 7.090, 8.489$. Segue che la media campionaria data da $M_c = \sum_j c_j N_j/500$ vale 5.878. Inoltre, $\mathrm{E}_\beta(X) = M_c$ se e solo se $\beta = 5M_c/3$; segue che $\widehat{\beta} = 5M_c/3 \simeq 5 \times 5.878/3 \simeq 9.797$ è una stima di β.

3.

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \int_0^x \frac{12t^2(\beta-t)}{\beta^4}\, dt = \frac{4x^3(\beta-x)}{\beta^4} + \frac{x^4}{\beta^4} = \frac{4x^3}{\beta^3} - \frac{3x^4}{\beta^4} & 0 < x < \beta \\ 1 & x \geq \beta \end{cases}$$

4. Dobbiamo impostare un test chiquadrato di buon adattamento per dati raggruppati per l'ipotesi nulla composta H_0 :"X ha densità beta del tipo (3)" contro l'alternativa che X non abbia densità del tipo (3).

Usando la f.d.r. calcolata al punto 3., sostituendovi il valore di β stimato: $\widehat{\beta} = 9.797$, otteniamo i valori "stimati" delle probabilità "teoriche": $p_1^{(0)}(\widehat{\beta}) = F_{H_0}(2.93) \simeq 0.084$, $p_2^{(0)}(\widehat{\beta}) = F_{H_0}(4.18) - 0.084 \simeq 0.211 - 0.084 = 0.127, \ldots$; in sintesi

A_k	$(0, 2.93]$	$(2.93, 4.18]$	$(4.18, 4.84]$	$(4.84, 6.52]$	$(6.52, 7.66]$	> 7.66
$p_i^{(0)}(\widehat{\beta})$	0.084	0.127	0.093	0.286	0.201	0.209
$np_i^{(0)}(\widehat{\beta})$	42	63.5	46.5	143	100.5	104.5

La statistica test di Pearson $Q = \sum_{i=1}^{5} \frac{(N_i - np_i^{(0)}(\widehat{\beta}))^2}{np_i^{(0)}(\widehat{\beta})}$ vale 7.66723 e la sua f.d.r. asintotica è $\chi^2_{6-1-1} = \chi^2_4$. Usando le tavole dei quantili della f.d.r. χ^2_4 deduciamo che il p-value del test cade nell'intervallo $(10\%, 12.5\%)$, essendo $\chi^2_4(87.5\%) = 7.214 < 7.66723 < 7.779 = \chi^2_4(90\%)$. Pertanto, ai consueti livelli di significatività (minori o uguali a 10%), rifiutiamo l'ipotesi di dati beta del tipo (3). ∎

4 STATISTICA per ING INF TEL [2L] I. Epifani, A.M. Pievatolo 26.09.09

© I diritti d'autore sono riservati. Ogni sfruttamento commerciale non autorizzato sarà perseguito.

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

Esercizio 4.1 Un ingegnere deve controllare la precisione di un multimetro. Per fare ciò misura 5 volte la differenza di potenziale tra i due poli di un generatore di potenziale μ_0 noto: $\mu_0 = 10^{-1}$V e ottiene:

$$X_1 = 0.997055 \times 10^{-1},\ X_2 = 0.996954 \times 10^{-1},\ X_3 = 0.999521 \times 10^{-1},\ X_4 = 1.00664 \times 10^{-1},\ X_5 = 1.00261 \times 10^{-1}$$

Inoltre, le misurazioni $X_1, X_2, \dots, X_5$ costituiscono un campione gaussiano di media 10^{-1}V e varianza σ^2 incognita.

L'ingegnere definisce il multimetro "preciso" se $\sigma \leq \sigma_0 = 0.5 \times 10^{-2}$V, altrimenti il multimetro non è preciso.

1. Determinate un intervallo di confidenza unilaterale del tipo (c, ∞) di livello 95% per σ.
2. Verificate con un test di ampiezza 5% l'ipotesi H_0: *"il multimetro è preciso"*, contro H_1: *"il multimetro non è preciso"*.
3. Determinate la funzione di potenza del test e fornitene un grafico qualitativo.
4. Fornite un valore approssimato della probabilità di errore di secondo tipo del test costruito al punto 2. in $\sigma = 0.6 \times 10^{-2}$, o eventualmente indicate l'intervallo dove questo valore cade.

Soluzione

1. Prima costruiamo un intervallo di confidenza (IC) per σ^2 e poi ne ricaviamo uno per σ. $S_0^2 = \sum_{j=1}^5 (X_j - \mu_0)^2/5$ è uno stimatore (non distorto) di σ^2 e $\frac{5S_0^2}{\sigma^2} \sim \chi^2_5$ è una quantità pivotale. Inoltre, per il nostro campione abbiamo S_0^2 vale 0.138165×10^{-8}. Quindi: $P_{\sigma^2}(\frac{5S_0^2}{\sigma^2} < q) = \gamma$ se e solo se $q = \chi^2_5(\gamma)$. Pertanto, con probabilità γ: $\frac{5S_0^2}{\chi^2_5(\gamma)} < \sigma^2$ e un IC di livello 95% per σ^2 è ad esempio:

$$\left(\frac{5 \times s_0^2}{\chi^2_5(0.95)}, \infty\right) = \left(\frac{5 \times 0.138165 \times 10^{-8}}{11.07}, \infty\right) = (0.062405 \times 10^{-8}, \infty).$$

Ricaviamo che un IC(σ) di livello 95% è $(\sqrt{0.062405 \times 10^{-8}}, \infty) = (0.24981 \times 10^{-4}, \infty)$

2. Dobbiamo verificare $H_0 : \sigma \leq \sigma_0 = 0.5 \times 10^{-2}$ contro $H_1 : \sigma > \sigma_0 = 0.5 \times 10^{-2}$. Alla luce del punto 1. siamo poco confidenti (solo 5%) che $\sigma \leq 0.24981 \times 10^{-4}$ e decidiamo di rifiutare H_0 se $\sigma_0 \leq 0.24981 \times 10^{-4}$. Poiché $\sigma_0 = 0.5 \times 10^{-2}$, allora accettiamo H_0 a livello 5%. Effettivamente $\sup_{\sigma \leq \sigma_0} P_\sigma(\sigma_0 \leq \sqrt{5S_0^2/\chi^2_5(0.95)}) = 5\%$: è un esempio di dualità fra VI e IC.

3.

$$\pi(\sigma) = P_\sigma\left(0.5 \times 10^{-2} \leq \sqrt{5S_0^2/11.07}\right) = P_{\sigma^2}\left(\frac{5S_0^2}{\sigma^2} \geq \frac{0.25 \times 10^{-4} \times 11.07}{\sigma^2}\right) =$$
$$= 1 - F_{\chi^2_5}\left(\frac{2.7675 \times 10^{-4}}{\sigma^2}\right), \quad \sigma > 0.5 \times 10^{-2}$$

che è funzione crescente di σ con asintoto orizzontale in 1 per $\sigma \to \infty$.

4. $\beta(0.6 \times 10^{-2}) = 1 - \pi(0.6 \times 10^{-2}) = F_{\chi^2_5}\left(\frac{2.7675 \times 10^{-4}}{0.36 \times 10^{-4}}\right) = F_{\chi^2_5}(7.6875) \simeq F_{\chi^2_5}(7.289) = 80\%$ ∎

Esercizio 4.2 Sia $X_1, \dots, X_n$ un campione casuale estratto da una popolazione di densità $\Gamma(2, \lambda)$, cioè

$$f(x, \lambda) = \frac{x}{\lambda^2} e^{-\frac{x}{\lambda}} \mathbf{1}_{(0,\infty)}(x),$$

con λ parametro positivo incognito.

1. Determinate uno stimatore $\widehat{\lambda}$ di λ con il metodo di massima verosimiglianza.
2. Calcolate la caratteristica $\kappa = P_\lambda(X_1 > 2)$ e stimatela con il metodo di massima verosimiglianza.
3. Verificate che l'informazione di Fisher del modello $\Gamma(2, \lambda)$ è $I(\lambda) = \dfrac{2}{\lambda^2}$ e che il confine inferiore di Frechét-Cramer-Rao per la varianza di uno stimatore (non distorto) della caratteristica $\kappa = P_\lambda(X_1 > 2)$ è dato da $\dfrac{8e^{-\frac{4}{\lambda}}}{n\lambda^4}$.
4. Qual è la distribuzione asintotica dello stimatore di massima verosimiglianza per κ, determinato al punto 2.? (Fornitene esplicitamente i parametri)
5. Determinate un intervallo di confidenza asintotico di livello approssimativamente 95% per $\kappa = P_\lambda(X_1 > 2)$, se $n = 200$ e $\widehat{\lambda} = 2.0$.

Soluzione

1. $L_\lambda(x_1, \dots, x_n) = \prod_{j=1}^n f(x_j, \lambda) = \frac{\prod_{j=1}^n x_j}{\lambda^{2n}} e^{-\frac{\sum_{j=1}^n x_j}{n}}$

e

$$\frac{\partial \log L_\lambda}{\partial \lambda} = -\frac{2n}{\lambda} + \frac{\sum_{j=1}^n x_j}{\lambda^2} = \frac{2n}{\lambda^2}\left(\frac{\bar{x}}{2} - \lambda\right) \geq 0 \text{ se e solo se } \frac{\bar{x}}{2} \leq \lambda$$

quindi $\widehat{\lambda} = \dfrac{\bar{X}}{2}$.

2. La caratteristica κ è data da

$$\kappa = P_\lambda(X_1 > 2) = \int_2^\infty \frac{x}{\lambda^2} e^{-\frac{x}{\lambda}}\, dx = -\frac{x}{\lambda} e^{-\frac{x}{\lambda}}\Big|_2^\infty - \int_2^\infty -\frac{1}{\lambda} e^{-\frac{x}{\lambda}}\, dx = e^{-2/\lambda}\left(1 + \frac{2}{\lambda}\right)$$

e un suo stimatore ML è

$$\widehat{\kappa} = e^{-2/\widehat{\lambda}}\left(1 + \frac{2}{\widehat{\lambda}}\right)$$

3.

$$I(\lambda) = \mathrm{E}\left[\left(\frac{\partial \log f(X_1, \lambda)}{\partial \lambda}\right)^2\right] = \mathrm{E}\left[\frac{4}{\lambda^4}\left(\frac{X_1}{2} - \lambda\right)^2\right] = \frac{4}{\lambda^4} \mathrm{Var}\left(\frac{X_1}{2}\right) = \frac{4}{\lambda^4} \times \frac{\lambda^2}{2} = \frac{2}{\lambda^2}$$

in quanto $X_1 \sim \Gamma(2, \lambda)$ cosicché $\mathrm{E}(X_1/2) = \lambda$ e $\frac{1}{4}\mathrm{Var}(X_1) = \frac{1}{4} \times 2\lambda^2$. Inoltre,

$$\kappa'(\lambda) = \frac{4e^{-\frac{2}{\lambda}}}{\lambda^3}$$

quindi

$$\frac{(\kappa'(\lambda))^2}{nI(\lambda)} = \frac{\frac{4^2 e^{-\frac{4}{\lambda}}}{\lambda^6}}{n\frac{2}{\lambda^2}} = \frac{8e^{-\frac{4}{\lambda}}}{n\lambda^4}$$

4. Segue dalle proprietà asintotiche degli stimatori ML di modelli regolari di caratteristiche $\kappa(\lambda)$ derivabili che la f.d.r. asintotica di $\widehat{\kappa}$ è $\mathcal{N}\left(\kappa, \frac{(\kappa'(\lambda))^2}{nI(\lambda)}\right)$.

5. $\widehat{k} = e^{-2/2}(1 + 2/2) = 2e^{-1}$ e

$$\left[\widehat{\frac{(\kappa'(\lambda))^2}{nI(\lambda)}}\right] = \frac{8e^{-4/2}}{200 \times 2^4} = \left(\frac{e^{-1}}{20}\right)^2.$$

Segue che un IC asintotico per κ di livello 95% ha estremi $2e^{-1} \pm 1.645 \times e^{-1}/20$: cioè, siamo 95%-confidenti (approssimativamente) che $.7055 < \kappa < 0.7660$. ▮

Esercizio 4.3 Su di un campione di 10 donne viene misurata la permeabilità di una membrana placentare al termine della gravidanza. Su di un secondo campione di 5 donne la misura è effettuata tra la 12a e la 26a settimana. Gli indici di permeabilità per i due campioni sono rispettivamente

$$(0.80, 0.83, 1.89, 1.04, 1.45, 1.38, 1.91, 1.64, 0.73, 1.46) \qquad \text{e}$$

$$(1.15, 0.88, 0.90, 0.74, 1.21)$$

A un valore più elevato dell'indice corrisponde una maggiore permeabilità.

1. Costruire un test di livello 5% per determinare se la permeabilità a termine è superiore a quella tra la 12a e la 26a settimana.
2. Fornire un'approssimazione del p-value del test.
3. Calcolare la funzione di ripartizione di una variabile aleatoria uniforme nell'intervallo $(0.5, 2)$.
4. Alla luce dei risultati del precedente test, stabilire se l'indice di permeabilità può avere densità uniforme nell'intervallo $(0.5, 2)$.

Soluzione

1. Indichiamo con X e Y le v.a. della permeabilità rispettivamente nel primo e nel secondo campione e con F e G le loro funzioni di ripartizione. Dobbiamo verificare l'ipotesi nulla che $F(x) = G(x)$ per ogni x contro $F(x) \leq G(x)$, con $F(x) < G(x)$ per qualche x. I ranghi del campione di 10 donne e del campione di 5 donne sono rispettivamente $(3, 4, 14, 7, 11, 10, 15, 13, 1, 12)$ e $(8, 5, 6, 2, 9)$, pertanto $T_X = 90$ e tenderemo a rifiutare l'ipotesi nulla per valori elevati di T_X, cioè per $T_X > w_{1-\alpha}$.

Dalle tavole troviamo che, con $m = 10$ e $n = 5$,

$$w_{0.95} = 10 \times (10 + 5 + 1) - w_{0.05} = 160 - 67 = 93$$

Poiché $T_X = 90 < 93$ non possiamo rifiutare l'ipotesi che la permeabilità nei due periodi della gravidanza sia la stessa.

2. Nelle tavole troviamo che con $m = 10$ e $n = 5$, $w_{0.9} = 160 - 69 = 91$. Pertanto $\Pr(T_X > 90) > \Pr(T_X > 91) = 0.10$ e il p-value è superiore al 10% (valore esatto 12.72%).

3. La f.d.r. uniforme nell'intervallo $(0.5, 2)$ è

$$F_0(x) = \begin{cases} 0 & \text{se } x \leq 0 \\ (x - 0.5)/1.5 & \text{se } 0.5 < x < 2 \\ 1 & \text{se } x \geq 2 \ . \end{cases}$$

4. Dato che le permeabilità si possono considerare provenienti dalla stessa distribuzione di probabilità, uniamo i due campioni in un unico campione di dimensione $n = 15$. Il nuovo campione (ordinato) $x_1, \ldots, x_{15}$ è

$$0.73, 0.74, 0.80, 0.83, 0.88, 0.90, 1.04, 1.15, 1.21, 1.38, 1.45, 1.46, 1.64, 1.89, 1.91$$

Utilizziamo il test di adattamento di Kolmogorov-Smirnoff per verificare l'ipotesi nulla $F = F_0$.

I valori assunti da F_0 nei punti campionari sono

$$0.15, 0.16, 0.20, 0.22, 0.25, 0.27, 0.36, 0.43, 0.47, 0.59, 0.63, 0.64, 0.76, 0.93, 0.94$$

Le differenze assolute tra la f.d.r. empirica e quella teorica nei punti campionari, cioè $|\hat{F}_n(x_i) - F_0(x_i)|$, per $i = 1, \ldots, n$ sono

$$0.08, 0.03, 0.00, 0.05, 0.08, 0.13, 0.11, 0.10, 0.13, 0.08, 0.10, 0.16, 0.11, 0.00, 0.06$$

Le differenze assolute $|\hat{F}_n(x_{i-1}) - F_0(x_i)|$, per $i = 1, \ldots, n$ sono

$$0.15, 0.09, 0.07, 0.02, 0.02, 0.06, 0.04, 0.04, 0.06, 0.01, 0.04, 0.09, 0.04, 0.06, 0.01$$

pertanto $D_n = 0.16$. Avendosi dalle tavole della statistica di Kolmogorov-Smirnoff con $n = 15$ che $q_{0.5} = 0.2659$, non rifiutiamo l'ipotesi di uniformità al livello del 50%. (p-value esatto 78.17%). ∎

© I diritti d'autore sono riservati. Ogni sfruttamento commerciale non autorizzato sarà perseguito.

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

Esercizio 5.1 Sia $X_1, \ldots, X_n$ un campione gaussiano di media incognita μ e varianza nota e pari a 900.

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

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

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

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

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

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

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

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

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

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

cioè tale che

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

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

Esercizio 5.2 Considerate la famiglia di densità

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

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

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

Soluzione

1.

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

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

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

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

4. La f.d.r. asintotica di $(T_n - \vartheta)/\sqrt{(\vartheta^2-\vartheta+1)/(2n)}$ è $\mathcal{N}(0,1)$. Segue che un IC asintotico bilatero per ϑ di livello approssimativamente 90% ha estremi $T_n \pm 1.645 \times \sqrt{(T_n^2 - T_n + 1)}/\sqrt{2n}$. Con i dati forniti, $T_{169} = 3 \times 68/169 - 1 \simeq 0.207$ e siamo 90%-confidenti (approssimativamente) che $0.1252 < \vartheta < 0.2888$.

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

Esercizio 5.3 È stato condotto uno studio su 200 pazienti affetti da retinopatia diabetica. Un occhio, casualmente scelto fra il destro e il sinistro, è trattato e l'altro è osservato senza trattamento. T_1 rappresenta il tempo che trascorre -a partire da un tempo 0- fino alla cecità dell'occhio trattato e T_2 di quello non trattato. I tempi T_1, T_2 sono entrambi espressi in anni e i dati raggruppati raccolti sono riportati nella seguente tabella:

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

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

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

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

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

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

e la statistica di Pearson è

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

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

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

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

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

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