Politecnico di Milano
Temi d'esame di STATISTICA dell'AA 2006/2007
per allievi ING INF [2L], docente I. Epifani

1 STATISTICA per ING INF TEL [2L] I. Epifani, A. Guglielmi 25.06.07

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

Esercizio 1.1 Dobbiamo decidere se una variabile aleatoria continua abbia densità

$$f_a(x) = \frac{1}{\theta_a} x^{\frac{1}{\theta_a}-1}\mathbf{1}_{(0,1)}(x) \text{ oppure } f_b(x) = \frac{1}{\theta_b} x^{\frac{1}{\theta_b}-1}\mathbf{1}_{(0,1)}(x)$$

usando una sola osservazione X. I valori θ_a, θ_b sono fissati, entrambi positivi ed ovviamente diversi fra loro.

1. Scegliete ipotesi nulla e ipotesi alternativa in modo tale che sia pari ad α la probabilità di commettere l'errore di primo di tipo di decidere a favore di f_b quando il vero valore di θ è θ_a.
2. Costruite il test uniformemente più potente di livello α.
3. Calcolate il p-value dei dati se $X = 0.7$, $\theta_a = 1$ e $\theta_b = 5$.
4. Calcolate la potenza del test se $\alpha = 5\%$, $\theta_a = 1$ e $\theta_b = 5$.

SOLUZIONE Deduciamo dal testo dell'esercizio di dover impostare il test di Neyman-Pearson per verificare l'ipotesi nulla $H_0 : \theta = \theta_a$ contro l'alternativa $H_1 : \theta = \theta_b$.

1. $H_0 : \theta = \theta_a$ e $H_1 : \theta = \theta_b$.
2. Dal Lemma di Neyman-Pearson deriva che la regione critica di ampiezza α è

$$\begin{aligned}\mathcal{G} &= \left\{x \in (0,1) : \frac{L_a(x)}{L_b(x)} \leq \delta\right\} = \left\{x \in (0,1) : \frac{\theta_b}{\theta_a} x^{\frac{1}{\theta_a}-\frac{1}{\theta_b}} \leq \delta\right\} \\ &= \begin{cases} \{x \in (0,1) : x \leq k_1\} = \left\{x \in (0,1) : x \leq \alpha^{\theta_a}\right\} & \text{se } \theta_a < \theta_b \\ \{x \in (0,1) : x \geq k_2\} = \left\{x \in (0,1) : x \geq (1-\alpha)^{\theta_a}\right\} & \text{se } \theta_a > \theta_b \end{cases}\end{aligned}$$

   dal momento che k_1 è tale che

$$P_{\theta_a}(X \leq k_1) = \int_0^{k_1} \frac{1}{\theta_a} x^{\frac{1}{\theta_a}-1} = k_1^{\frac{1}{\theta_a}} = \alpha.$$

   e k_2 è tale che

$$P_{\theta_a}(X \geq k_2) = \int_{k_2}^1 \frac{1}{\theta_a} x^{\frac{1}{\theta_a}-1} = 1 - k_2^{\frac{1}{\theta_a}} = \alpha.$$

3. Essendo $\theta_a = 1 < 5 = \theta_b$, il p-value dei dati è $P_1(X \leq 0.7) = 0.7$.
4. $\pi = P_5(\mathcal{G}) = P_5(X \leq 0.05^1) = 0.05^{\frac{1}{5}} \simeq 0.549$. ∎

Esercizio 1.2 Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla popolazione di densità $f(x,\theta) = \frac{1}{\theta}(1+x)^{-\frac{1}{\theta}-1}\mathbf{1}_{(0,\infty)}(x)$, con θ parametro positivo incognito.

1. Determinate uno stimatore di θ usando il metodo di massima verosimiglianza.
2. Stabilite se lo stimatore di massima verosimiglianza di θ trovato al punto 1. sia efficiente.
3. Determinate la distribuzione di $Y_j = \ln(1+X_j)$ per ogni $j = 1, \ldots, n$ (ln indica il logaritmo in base naturale). Quindi deducete la distribuzione dello stimatore di massima verosimiglianza di θ trovato al punto 1.
4. Usate i risultati trovati ai punti 1. e 3. per calcolare un intervallo di confidenza esatto per θ unilatero della forma $(0, c)$ di livello $\gamma = 90\%$ per il campione di quattro osservazioni 0.11, 0.94, 0.48, 1.23 estratte da $f(x,\theta)$.

SOLUZIONE

1. La funzione di verosimiglianza del campione $X_1, \ldots, X_n$ è

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

da cui deriviamo che

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

e quindi $\frac{\partial}{\partial\theta}\ln L_\theta(x_1, \ldots, x_n) \geq 0$ se e solo se $\theta \leq \frac{\sum_{j=1}^{n}\ln(1+x_j)}{n}$. Segue che $\hat{\theta}_{ML} = \frac{\sum_{j=1}^{n}\ln(1+x_j)}{n}$.

2. Innanzitutto osserviamo che in virtù dell'Equazione (1) $\mathrm{E}(\frac{\partial}{\partial\theta}\ln L_\theta(X_1, \ldots, X_n)) = 0$ se e solo se $\mathrm{E}\left(\frac{\sum_{j=1}^{n}\ln(1+X_j)}{n}\right) = \theta$ e quindi $\hat{\theta}_{ML}$ è stimatore non distorto. Inoltre, introdotta la funzione $a(n,\theta) = n/\theta^2$, l'Equazione (1) può essere riletta come segue

$$P_\theta\left(\frac{\partial}{\partial\theta}\ln L_\theta(x_1, \ldots, x_n) = a(n,\theta)(\hat{\theta}_{ML} - \theta)\right) = 1, \quad \forall\theta \tag{2}$$

Infine, sappiamo dai risultati sul confine di Cramer Rao che l'Equazione (2) è condizione necessaria e sufficiente per l'efficienza dello stimatore.

3. Se $X_j \sim f(x,\theta)$, allora $Y_j = \ln(1+X_j) \sim Exp(\theta)$. Infatti, $P(Y > 0) = 1$, mentre, per ogni $y > 0$ abbiamo $F_Y(y) = P(\ln(1+X_j) \leq y) = F_X(\mathrm{e}^y - 1)$ e quindi, $f_Y(y) = \mathrm{e}^y f(\mathrm{e}^y - 1, \theta)\mathbf{1}_{(0,\infty)}(y) = \frac{1}{\theta}\mathrm{e}^{-\frac{y}{\theta}}\mathbf{1}_{(0,\infty)}(y)$.

   Lo stimatore $\hat{\theta}_{ML}$ è dunque una media campionaria di n variabili aleatorie i.i.d. esponenziali di parametro θ e quindi, per le proprietà della famiglia delle leggi gamma, $\hat{\theta}_{ML}$ ha densità $\Gamma(n, \frac{\theta}{n})$.

4. Se $\hat{\theta}_{ML} \sim \Gamma\left(n, \frac{\theta}{n}\right)$, allora $\frac{2n\hat{\theta}_{ML}}{\theta} \sim \chi^2_{2n}$ e $P\left(\frac{2n\hat{\theta}_{ML}}{\theta} > \chi^2_{2n}(10\%)\right) = 0.90$ Quindi un $IC(\theta)$ esatto unilatero di forma $(0,c)$ è dato da $\left(0, \frac{2n\hat{\theta}_{ML}}{\chi^2_{2n}(10\%)}\right)$. Con i dati a nostra disposizione abbiamo $\hat{\theta}_{ML} \simeq 0.49$, $\chi^2_8(10\%) = 3.49$ e $\left(0, \frac{2n\hat{\theta}_{ML}}{\chi^2_{2n}(10\%)}\right) = (0, 1.123)$. ∎

Esercizio 1.3 Per 51 minuti primi si è registrato il numero di *outlink* dalla pagina web `aaa` alla pagina web `bbb`. Questi dati costituiscono un campione casuale $X_1, \ldots, X_{51}$ la cui funzione di ripartizione empirica $\hat{F}_{51}$ è data da

k	0	1	2	3	4	5	6
$\hat{F}_{51}(k)$	6/51	20/51	36/51	44/51	46/51	48/51	1

Più precisamente, X_j rappresenta il numero di *outlink* da `aaa` a `bbb` nel minuto j.

1. Usate i dati forniti per costruire un intervallo di confidenza asintotico di livello 95% della probabilità che in un minuto non ci siano *outlink* da `aaa` a `bbb`.

2. Calcolate il numero medio per minuto di *outlink* da `aaa` a `bbb`.

In realtà, sulla base di precedenti analisi statistiche fatte su altre pagine web, si può modellare il numero di *outlink* al minuto da `aaa` a `bbb` come una variabile aleatoria di Poisson di parametro $\theta > 0$ e θ è incognito. In altri termini, da questo momento i dati assegnati costituiscono la realizzazione di un campione casuale $X_1, \ldots, X_n$ estratto dalla densità di Poisson di parametro θ incognito.

3. Determinate lo stimatore di massima verosimiglianza della probabilità $p = p(\theta)$ che non ci siano *outlink* da `aaa` a `bbb` in un minuto e calcolatene il suo valore sulla base dei dati forniti.

4. Proponete un test di livello α per verificare l'ipotesi nulla $H_0 : p = 0.1$ contro l'alternativa $H_1 : p > 0.1$. Se $\alpha = 2.5\%$, cosa decidete sulla base del test di ipotesi costruito?

SOLUZIONE

1. Sia $\hat{p}$ è la frequenza campionaria dell'evento "nessun outlink da `aaa` a `bbb`" in 51 minuti. Approssimativamente $\hat{p}$ ha f.d.r. gaussiana di media p e varianza $p(1-p)/51$. Allora un $IC(p)$ asintotico di livello 95% ha estremi $\hat{p} \pm z_{\frac{1+0.95}{2}} \sqrt{\frac{\hat{p} \times (1-\hat{p})}{51}}$. Con i dati assegnati, questo $IC(p)$ è $\frac{6}{51} \pm z_{\frac{1+0.95}{2}} \sqrt{\frac{\frac{6}{51} \times \frac{45}{51}}{51}} = (0.029, 0.206)$.

2. $\bar{X} = \sum_{k=1}^{6} k \left(\hat{F}_{51}(k) - \hat{F}_{51}(k-1) \right) = \frac{1 \times 14 + 2 \times 16 + 3 \times 8 + 4 \times 2 + 5 \times 2 + 6 \times 3}{51} = 2.078$

3. Lo stimatore ML del parametro θ di un modello statistico di Poisson è $\bar{X}$ e quindi quello ML di $p = p(\theta) = P_\theta(X = 0) = \mathrm{e}^{-\theta}$ è $p(\bar{X}) = \mathrm{e}^{-\bar{X}}$. Con i dati assegnati, la stima di massima verosimiglianza di p è $\hat{p} = \mathrm{e}^{-2.078} \simeq 0.125$.

4. Osserviamo che $p = 0.1$ se e solo se $\theta = \ln 10$; quindi, $H_0 : p = 0.1$ è equivalente a $H_0 : \theta = \ln 10$ e $H_1 : p > 0.1$ è equivalente a $H_0 : \theta < \ln 10$. Cioè ci siamo ridotti a costruire un test sulla media per una popolazione di Poisson.

   Prima soluzione Per queste ipotesi sulla media, possiamo usare la statistica test $\bar{X}$ e rifiutare H_0 se $\bar{X} \leq k$ con k tale che $P_{\ln 10}\left(\bar{X} \leq k\right) \leq \alpha$. Poiché sotto H_0, $\sum_{j=1}^{51} X_j$ è variabile di Poisson di parametro $51 \ln 10$, allora k risulta essere il quantile di ordine α , q_α, della f.d.r. di Poisson di parametro $51 \ln 10$ diviso per 51. Pertanto la regione critica del test è $\{\bar{X} \leq q_\alpha/51\}$. Inoltre, essendo $51 \ln 10 \simeq 117.43$ un numero "grande", possiamo approssimare la f.d.r. $Poisson(51 \ln 10)$ con la f.d.r. $\mathcal{N}(51 \ln 10, 51 \ln 10)$. Segue che un valore approssimato di q_α è dato da $q_\alpha = 51 \ln 10 - z_{1-\alpha}\sqrt{51 \ln 10}$. Se $\alpha = 2.5\%$, allora $q_{0.25} \simeq 96.19$ e $k \simeq 96.19/51 \simeq 1.89$. Siccome $\bar{X} = 2.078 > 1.89$, non possiamo rifiutare H_0 al livello $\alpha = 2.5\%$.

   Seconda soluzione La regione di rifiuto del test può essere ricavata per dualità dall'IC unilaterale asintotico per θ di livello $1-\alpha$ della forma $(0, c)$. Ricordiamo che $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X_1)/n = \theta/n$ può essere stimata da $\hat{\theta}/n$ e quindi asintoticamente $\dfrac{\bar{X} - \theta}{\sqrt{\hat{\theta}/n}}$ è una quantità pivotale che ha f.d.r. (asintotica) $\mathcal{N}(0,1)$.

   Segue che l'IC unilaterale asintotico per θ di livello $1-\alpha$ della forma $(0, c)$ è $(0, \bar{x} - z_{1-\alpha}\sqrt{\bar{x}/n}) = (0, 2.474)$. Poiché $\ln 10 \simeq 2.303 \in (0, 2.474)$, non possiamo rifiutare H_0 al livello $\alpha = 2.5\%$. ∎

Esercizio 1.4 L'AZT è stato il primo farmaco antiretrovirale approvato dalla "Food and Drug Administration" statunitense ad essere usato nella cura di pazienti affetti dal virus dell'HIV perché riduce l'attività virale. La dose standard di AZT è di 300 mg giornalieri. Alcuni studi sostengono che dosaggi giornalieri più alti non sono più efficaci, provocando invece fastidiosi effetti collaterali.

Per verificare se il dosaggio di $600mg$ giornaliero di AZT sia ugualmente efficace della dose standard nel controllo della malattia, si sono misurati i livelli di antigene p24 nel sangue di due gruppi di pazienti, il primo trattato con 300 mg giornalieri di AZT e il secondo con 600 mg, perchè livelli alti di antigene p24 indicano elevata replicazione virale, e quindi, "malattia conclamata". I risultati ottenuti sono:

$$300\,\text{mg}: 283, 284, 285, 286, 288, 289, 291, 295, 303 \text{ e } 600\,\text{mg}: 287, 292, 293, 296, 298, 310, 314.$$

1. Impostate un test di livello 1% per verificare l'ipotesi nulla H_0: "*i due dosaggi controllano la malattia allo stesso modo*" contro l'alternativa H_1: "*i due dosaggi non controllano la malattia allo stesso modo*". Le conclusioni cambiano a livello 10%?

Supponiamo ora che i due campioni dei livelli di antigene p24 siano gaussiani.

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

3. Costruite un test di ipotesi di livello $\alpha = 5\%$, che utilizzi l'ipotesi di gaussianità e le informazioni desunte dal punto 2., per verificare H_0: "*i due dosaggi controllano la malattia allo stesso modo*" contro l'alternativa H_1: "*i due dosaggi non controllano la malattia allo stesso modo*".

SOLUZIONE

1. I due gruppi di dati costituiscono due campioni casuali indipendenti, $X_1, \dots, X_{10}$ i.i.d. $\sim F$ e $Y_1, \dots, Y_7$ i.i.d. $\sim G$ (F, G sono f.d.r.) e le ipotesi da verificare sono

$$H_0 : F(x) = G(x) \text{ per ogni } x \text{ vs. } \; H_1 : F(x) \neq G(x) \text{ per almeno una } x.$$

   Se ordiniamo i valori misurati di p24 dal più piccolo al più grande, i ranghi R_X delle osservazioni del primo gruppo sono:

$\{X_i\}, \{Y_j\}$:	283	284	285	286	287	288	289	291	292	293	295	296	298	303	310	314
R_X:	1	2	3	4		6	7	8			11			14		

   Il valore osservato della statistica test T_X è 56, con $m = 9$ e $n = 7$ e rifiutiamo H_0 se $T_X \notin (w_{\frac{\alpha}{2}}, w_{1-\frac{\alpha}{2}})$. Con $\alpha = 0.01$, abbiamo $w_{\frac{\alpha}{2}} = w_{0.005} = 53$ e $w_{1-\frac{\alpha}{2}} = w_{0.995} = m(m+n+1) - w_{0.005} = 153 - 53 = 100$; quindi non possiamo rifiutare H_0 al livello di significatività dell'1%. Se invece $\alpha = 10\%$, allora $w_{\frac{0.1}{2}} = w_{0.05} = 61$ e $w_{1-5\%} = w_{0.95} = 153 - 61 = 92$ e cambiamo decisione, rifiutando H_0.

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

   I valori osservati delle varianze campionarie sono $s_X^2 = 40.25$ e $s_Y^2 = 97.286$, e dunque $s_X^2/s_Y^2 = 0.414$. Inoltre, rifiutiamo H_0 se $s_X^2/s_Y^2 \leq F_{m-1,n-1}(\frac{\alpha}{2})$ o $s_X^2/s_Y^2 \geq F_{m-1,n-1}(1 - \frac{\alpha}{2})$, dove $F_{m-1,n-1}(a)$ è il quantile di ordine a della f.d.r. di Fisher con $m-1$ gradi di libertà al numeratore e $n-1$ gradi di libertà al denominatore. Dalle tavole abbiamo $F_{m-1,n-1}(\alpha/2) = F_{8,6}(0.025) = 1/F_{6,8}(1 - 0.025) = 1/4.65 \simeq 0.215$ e $F_{m-1,n-1}(1 - \alpha/2) = F_{8,6}(0.975) = 5.60$. Pertanto, accettiamo l'ipotesi di uguaglianza delle varianze al livello di significatività del 5%.

3. Ora possiamo assumere che i due campioni gaussiani $(X_1, \dots, X_{10})$ e $(Y_1, \dots, Y_7)$ abbiano stessa varianza incognita σ^2. Impostando il t test per verificare $H_0 : \mu_X = \mu_Y$ *vs* $H_1 : \mu_X \neq \mu_Y$, rifiutiamo H_0 a livello $\alpha = 5\%$ se $T = |\bar{X} - \bar{Y}| / \sqrt{S_p^2(\frac{1}{m} + \frac{1}{n})} > t_{14}(0.975)$. Lo stimatore pooled di σ^2 vale $s_p^2 \simeq 64.69$, mentre $\bar{x} = 289.33$, $\bar{y} = 298.57$ e quindi $T = 2.279$ che è maggiore di $t_{14}(0.975) = 2.145$: nuovamente rifiutiamo l'ipotesi che il livello di antigene p24 sia lo stesso nei due gruppi. ∎

2 STATISTICA per ING INF TEL [2L] I. Epifani, A. Guglielmi 10.07.07

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

Esercizio 2.1 Il reddito mensile di una certa popolazione è una variabile aleatoria continua X con densità di Pareto data da

$$f(x,a,b) = \begin{cases} \frac{ab^a}{x^{a+1}} & \text{se } x > b \\ 0 & \text{se } x \le b, \end{cases} \qquad a > 2 \text{ e } b > 0.$$

I parametri a, b sono entrambi incogniti e per stimarli si analizzano i redditi $X_1, \ldots, X_{100}$ di un campione casuale di individui di questa popolazione, ottenendo un reddito medio campionario pari a $\bar{x} = 1500$ euro con varianza campionaria $s^2 = 750000$.

1. Calcolate i primi due momenti $\mu_1(a,b) = \mathrm{E}(X)$ e $\mu_2(a,b) = \mathrm{E}(X^2)$ della densità $f(x,a,b)$.
2. Determinate uno stimatore di a e uno di b usando il metodo dei momenti e calcolatene il loro valore sulla base dei valori forniti per $\bar{X}$ e S^2.
3. Calcolate (in funzione di a, b) la probabilità che un individuo estratto a caso da quella popolazione percepisca un reddito compreso tra 1900 e 3500 euro.
4. Proponete una stima della probabilità che un individuo estratto a caso da quella popolazione percepisca un reddito compreso tra 1900 e 3500 euro.

SOLUZIONE

1. $$\mu_1(a,b) = \int_b^\infty x \frac{ab^a}{x^{a+1}}\, dx = ab^a \int_b^\infty x^{-a}\, dx = \frac{ba}{a-1}$$
$$\mu_2(a,b) = \int_b^\infty x^2 \frac{ab^a}{x^{a+1}}\, dx = ab^a \int_b^\infty x^{-a+1}\, dx = \frac{b^2 a}{a-2}.$$

2. Risolviamo il seguente sistema in a, b:

$$\begin{cases} \frac{ba}{a-1} = M_1 \\ \frac{b^2 a}{a-2} = M_2 \end{cases} \text{ sse } \begin{cases} b = M_1(a-1)/a \\ b^2 a = M_2(a-2) \end{cases} \text{ sse } \begin{cases} b = M_1(a-1)/a \\ (a-1)^2 M_1^2 - M_2(a-2)a = 0 \end{cases} \text{ sse}$$
$$\begin{cases} b = M_1(a-1)/a \\ a^2(M_2 - M_1^2) - 2a(M_2 - M_1^2) - M_1^2 = 0 \end{cases} \text{ sse } \begin{cases} b = M_1(a-1)/a \\ a = 1 \pm \sqrt{1 + \frac{M_1^2}{M_2 - M_1^2}} \end{cases}$$

dove M_1 è la media campionaria e M_2 il momento secondo campionario. Poiché $a > 2$ escludiamo la soluzione $a = 1 - \sqrt{1 + \frac{M_1^2}{M_2 - M_1^2}}$. Segue che

$$\hat{a} = 1 + \sqrt{1 + \frac{M_1^2}{M_2 - M_1^2}} = 1 + \sqrt{1 + \frac{\bar{X}^2}{S^2 \times 99/100}}, \qquad \hat{b} = \frac{\hat{a} - 1}{\hat{a}} \bar{X}$$

sono gli stimatori dei momenti rispettivamente di a e b. Per quanto concerne i valori delle stime abbiamo: $\hat{a} \simeq 3.008$ e $\hat{b} \simeq 1001.33$.

3. Si ricava

$$\begin{aligned} P_{a,b}(1900 \le X \le 3500) &= \int_{1900}^{3500} \frac{ab^a}{x^{a+1}}\, dx = b^a \left(\frac{1}{1900^a} - \frac{1}{3500^a} \right) && \text{se } b \le 1900, \\ &= \int_b^{3500} \frac{ab^a}{x^{a+1}}\, dx = b^a \left(\frac{1}{b^a} - \frac{1}{3500^a} \right) && \text{se } 1900 < b \le 3500, \\ &= 0 && \text{se } b > 3500. \end{aligned}$$

4. Uno stimatore di $P_{a,b}(1900 \le X \le 3500)$ è dato dalla statistica $\hat{b}^{\hat{a}} \left(\frac{1}{1900^{\hat{a}}} - \frac{1}{3500^{\hat{a}}} \right)$ che vale approssimativamente 0.1224. ∎

Esercizio 2.2 Una compagnia di assicurazioni deve eseguire uno studio per stimare gli indennizzi pagati a seguito di "*danni provocati dai figli minori per uso di giocattoli*". Sospetta infatti che mediamente questi importi siano aumentati rispetto al triennio precedente in cui l'indennizzo medio era stato di 3500 euro. Per questo motivo viene analizzato un nuovo campione casuale di 16 incidenti del suddetto tipo per il quale si osserva una media campionaria degli indennizzi pari a 3525.438 euro. Inoltre, da studi precedenti è emerso che si può assumere che tali importi abbiano densità gaussiana con deviazione standard nota e pari a 50 euro.

1. Aiutate la compagnia di assicurazioni a decidere se il suo sospetto sia fondato o meno usando un opportuno test di ipotesi al livello $\alpha = 6\%$. Sulla base dei dati forniti, il sospetto è fondato?
2. Sulla base dei dati a disposizione, quanto siete confidenti che attualmente l'indennizzo medio abbia superato i 3506 euro?
3. Con riferimento al test di ipotesi costruito al punto 1., indicate per quali livelli α ritenete il sospetto fondato, con i dati a disposizione.
4. Determinate per quali valori dell'indennizzo medio la probabilità di errore di secondo tipo è inferiore o uguale al livello $\alpha = 6\%$ del test al punto 1.

SOLUZIONE Se X_i è l'indennizzo all'i-esimo incidente, $X_1, \ldots, X_{16}$ è un campione casuale estratto da una popolazione $\mathcal{N}(\mu, 2500)$ di media μ incognita.

1. Impostiamo il problema di verifica di ipotesi $H_0 : \mu \leq 3500$ contro $H_1 : \mu > 3500$. Quindi rifiutiamo se $\bar{X} \geq 3500 + z_{0.94}50/\sqrt{16} = 3500 + 1.555 \times 50/4 = 3519.438$; poiché $\bar{x} = 3525.438 > 3519.438$, rifiutiamo H_0, concludendo che il sospetto è fondato.
2. Siamo confidenti al 94% che attualmente l'indennizzo medio abbia superato i 3506 euro. Infatti, con i dati a disposizione, in virtù della dualità fra la verifica di ipotesi e gli intervalli di confidenza, segue dal punto 1. che $\{\mu : \ \bar{X} < \mu + z_{0.94}50/\sqrt{16}\} = \{\mu : \ \mu > \bar{x} - z_{0.94}50/\sqrt{16}\} = \{\mu : \mu > 3525.438 - 19.438\} = (3506, \infty)$ è un IC unilatero a una coda superiore per μ di confidenza $(100 - 6)\% = 94\%$.
3. Dobbiamo determinare il p-value dei dati $p = 1 - \Phi\left(\dfrac{3525.438 - 3500}{50/4}\right) = 1 - \Phi\left(2.03504\right) \simeq 1 - \Phi\left(2.04\right) \simeq 1 - 0.9793 = 0.0207$. Segue che, se $\alpha \geq 2.07\%$ accettiamo l'ipotesi H_1 che il sospetto sia fondato, mentre, per $\alpha < 2.07\%$ la rifiutiamo, con i dati a disposizione.
4. Dobbiamo determinare $\mu > 3500$ tale che $\beta(\mu) \leq 6\%$. Ma,

$$\beta(\mu) = P_\mu\left(\bar{X} < 3519.438\right) = \Phi\left(\frac{3519.438 - \mu}{50/4}\right)$$

e

$$6\% = \Phi(z_{0.06}) = \Phi(-z_{0.94}) \simeq \Phi(-1.555)$$

cosicché

$$\beta(\mu) \leq 6\% \text{ se e solo se } \frac{3519.438 - \mu}{50/4} \leq -1.555$$

In definitiva, $\beta(\mu) \leq 6\%$ per $\mu \geq 3538.876$. ∎

Esercizio 2.3 È stata effettuata un'indagine statistica che ha coinvolto i numerosissimi allievi del corso `XXX` della laurea triennale `AAA`. Agli allievi partecipanti al primo appello, che prevedeva domande a risposta multipla, è stato chiesto anche di valutare complessivamente il corso assegnando un voto da 0 a 10. Tutte le schede raccolte sono state corrette con il lettore ottico[1].

I dati sui voti riportati da ciascuno allievo, espressi in trentesimi, sono stati poi raggruppati secondo le categorie A, B, C e D con A che corrisponde a un voto in $[26, 30L]$, B a un voto in $[22, 25]$, C a un voto in $[18, 21]$ e D a un voto in $[0, 17]$, invece, i dati sui voti rassegnati al corso sono stati raggruppati nelle classi $[0, 4], [5, 7], [8, 10]$. I dati ottenuti sono riportati nella tabella seguente:

[1] Si tratta di un grosso ateneo con migliaia di matricole e tantissime sezioni parallele del corso `XXX`. Inoltre, la correzione mediante lettore ottico garantisce nella ricerca l'anonimato degli studenti e quindi che esprimano liberamente il loro giudizio sul corso, senza paura di essere riconosciuti

$Vcor \setminus Vstud$	D	C	B	A
[0,4]	68	115	157	50
[5,7]	100	130	215	65
[8,10]	20	22	43	15

dove $Vcor$ sta per il voto assegnato al corso e $Vstud$ per il voto riportato dallo studente all'appello.

1. Secondo voi, i risultati degli studenti all'esame dipendono dalle valutazioni del corso da parte degli studenti? Usate i precedenti dati e costruite un opportuno test di livello $\alpha = 5\%$.

2. Verificate l'ipotesi nulla che la distribuzione dei voti degli studenti in un appello sia uniforme discreta sull'insieme $\{A, B, C, D\}$. Usate i precedenti dati e costruite un opportuno test di livello $\alpha = 5\%$.

3. Usate i precedenti dati per costruire un intervallo di confidenza bilatero di livello approssimato $\gamma = 95\%$ per la percentuale (sull'intera popolazione studentesca) di bocciati a un appello *(Ovviamente, sono bocciati gli studenti che non raggiungono il* 18*).*

Soluzione

1. Deduciamo dalla tabella dei dati che in totale sono stati coinvolti nell'indagine statistica $n = 1000$ studenti. Completiamo la tabella con le numerosità marginali di ogni categoria per le variabili $Vcor$ (N_c) e $Vstud$ (N_s):

$Vcor \setminus Vstud$	D	C	B	A	N_c
[0,4]	68	115	157	50	390
[5,7]	100	130	215	65	510
[8,10]	20	22	43	15	100
N_s	188	267	415	130	1000

e impostiamo un test χ^2 di indipendenza fra le variabili $Vcor$ e $Vstud$. La statistica di Pearson Q_1 ha valore:

$$Q_1 = 1000 \left(\sum_{i=1}^{3} \sum_{j=1}^{4} \frac{N_{ij}^2}{N_{ci} N_{sj}} - 1 \right) \simeq 3.48$$

Asintoticamente Q_1 ha f.d.r. $\chi^2_{(3-1)(4-1)} = \chi^2_6$. Quindi, il p-value del test di Pearson è $1 - F_{\chi^2_6}(3.48) \geq 1 - F_{\chi^2_6}(3.455) = 1 - 25\% = 75\%$. Abbiamo usato le tavole dei quantili della f.d.r. χ^2_6 in rete. Concludiamo che ai "canonici" livelli di α (compreso $\alpha = 5\%$) non possiamo rifiutare l'ipotesi di indipendenza fra voto al corso e voto all'esame.

2. Impostiamo un test χ^2 di buon adattamento per verificare l'ipotesi $H_0 : P(V_{stud} = j) = 0.25 \ \forall j = A, B, C, D$ contro $H_1 : P(V_{stud} = j) \neq 0.25$ per qualche j. La statistica di Pearson Q_2 ha valore:

$$Q_2 = \sum_{j=A,B,C,D} \frac{N_{sj}^2}{250} - 1000 \simeq 183.03$$

Asintoticamente Q ha f.d.r. $\chi^2_{4-1} = \chi^2_3$ e rifiutiamo H_0 perché $183.03 > 7.815 = \chi^2_3(0.95)$.

3. La stima della percentuale p di essere bocciati all'esame è $\hat{p} = 188/1000 = 0.188$ e un intervallo bilatero di confidenza approssimativamente pari a 0.95 è dato da $\hat{p} \pm z_{0.975} \sqrt{\frac{\hat{p}(1-\hat{p})}{1000}} = 0.188 \pm 1.96 \sqrt{\frac{0.188 \times 0.812}{1000}} = (0.164, 0.212)$. ∎

Esercizio 2.4 Si vuole stabilire l'efficacia di un nuovo farmaco per aumentare il livello di sideremia (cioè della concentrazione di ferro nel plasma sanguigno). Per questo motivo, 75 donne iposideremiche sono state sottoposte a una cura di 30 giorni col nuovo farmaco. Quindi, sono stati misurati i livelli di sideremia, in microgrammi per decilitro (mcg/dl), prima e dopo la cura. Le pazienti su cui il farmaco è stato efficace (perché c'è stato aumento della sideremia) sono state 59 e non si sono avute ripetizioni nei valori osservati.

1. Costruite un test di livello $\alpha = 5\%$ per verificare l'ipotesi nulla che il farmaco non sia efficace contro l'alternativa che invece lo sia.

In realtà, di ogni donna si conosce l'esatto livello di sideremia prima e dopo la cura col nuovo farmaco e da questi valori si sono calcolate le seguenti utili statistiche:

$$\sum_{j=1}^{75} x_j = 4394.294,\ \sum_{j=1}^{75} x_j^2 = 303548,\ \sum_{j=1}^{75} y_j = 2959.761,\ \sum_{j=1}^{75} y_j^2 = 133417,$$
$$\sum_{j=1}^{75} x_j y_j = 187947.7,\ \sum_{j=1}^{75} (x_j - y_j - (\bar{x} - \bar{y}))^2 = 33631.13.$$

dove y_j è il livello della sideremia prima della cura e x_j quello dopo la cura. Inoltre, si può supporre che i dati accoppiati $\{(x_j, y_j), j = 1, \ldots, 75\}$ siano gaussiani.

2. Usando i dati forniti, verificate al livello 5% l'ipotesi nulla che il farmaco non sia efficace contro l'alternativa che invece lo sia.

3. Eseguite un test d'ipotesi di livello 1% per stabilire se c'è indipendenza tra i livelli di sideremia prima e dopo la cura, sulla base dei dati forniti.

SOLUZIONE Sia X il livello di sideremia nel sangue dopo la cura e Y quello prima della cura. I dati costituiscono un campione $(x_1, y_1), \ldots, (x_{75}, y_{75})$ da una f.d.r. congiunta H, con marginali F_X e F_Y. Il farmaco è efficace se il livello di sideremia nel sangue dopo la cura è più alto del livello prima della cura.

1. Impostiamo un test per verificare l'ipotesi

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

Notiamo che H_1 è equivalente a "X domina stocasticamente Y". Rifiutiamo H_0 se $T^+ > q^+_{1-\alpha}$, dove T^+ indica il numero di coppie (X_j, Y_j) nel campione casuale per cui X_j supera Y_j e $q^+_{1-\alpha}$ è il quantile di ordine $1-\alpha$ della f.d.r. $Bin(n, 1/2)$. Poiché $n = 75$ è "grande", un valore approssimato di $q^+_{1-\alpha}$ è

$$q^+_{1-\alpha} \simeq \frac{n}{2} + z_{1-\alpha}\frac{\sqrt{n}}{2} = \frac{75}{2} + z_{0.95}\frac{\sqrt{75}}{2} = \frac{75}{2} + 1.645\frac{\sqrt{75}}{2} \simeq 44.62.$$

Poiché $T^+ = 59 > 44.62$, rifiutiamo l'ipotesi H_0 che il farmaco non sia efficace nella cura dell'iposideremia.

2. Ora assumiamo che H indichi la f.d.r. gaussiana bivariata $\mathcal{N}(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$, dove ρ indica il coefficiente di correlazione lineare di X e Y. Impostiamo un test di verifica delle ipotesi $H_0 : \mu_X = \mu_Y$ vs $H_1 : \mu_X > \mu_Y$. Rifiutiamo H_0 a livello α se $t := (\bar{x} - \bar{y})/\sqrt{s^2_{X-Y}/n} \geq t_{n-1}(1-\alpha)$. Con i dati forniti abbiamo

$$n = 75,\ \bar{x} = 58.591,\ \bar{y} = 39.463,\ s^2_{X-Y} = \frac{\sum_{j=1}^{75}(x_j - y_j - (\bar{x} - \bar{y}))^2}{74} = 454.475 \text{ e } t_{74}(0.95) \simeq z_{0.95} = 1.645$$

Poiché $t = 7.77 > 1.645$, allora rifiutiamo l'ipotesi H_0 che il farmaco non sia efficace nella cura dell'iposideremia.

3. Impostiamo un test per le ipotesi $H_0 : \rho = 0$ contro $H_1 : \rho \neq 0$. Rifiutiamo H_0 se

$$|\frac{\sqrt{n-2}}{\sqrt{1-r^2}} r| \geq t_{n-2}(1 - \frac{\alpha}{2}),$$

dove $r = \dfrac{\sum(x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum(x_j - \bar{x})^2 \sum(y_j - \bar{y})^2}}$ è il coefficiente di correlazione campionario. Dai dati forniti ricaviamo

$$\sum(x_j - \bar{x})(y_j - \bar{y}) = \sum x_j y_j - n\bar{x}\bar{y} = 14534.45,$$
$$\sum(y_j - \bar{y})^2 = \sum y_j^2 - n(\bar{y})^2 = 16617.37,\ \sum(x_j - \bar{x})^2 = \sum x_j^2 - n(\bar{x})^2 = 46080.1,$$

da cui $r = 0.525$ e $r\sqrt{n-2}/\sqrt{1-r^2} = 5.270$. Poiché $t_{n-2}(1 - \frac{\alpha}{2}) = t_{73}(1 - \frac{0.01}{2}) \simeq z_{0.995} = 2.576 < 5.270$, rifiutiamo l'ipotesi H_0 che i livelli di sideremia prima e dopo la cura siano indipendenti a livello di significatività dell'1%. ∎

3 STATISTICA per ING INF TEL [2L] A. Barchielli (I. Epifani), A. Guglielmi 13.09.07

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

Esercizio 3.1 Per provare se una moneta è bilanciata viene adottata la seguente regola decisionale: si effettuano 100 lanci quindi, si accetta l'ipotesi che la moneta sia bilanciata se il numero di teste osservate è compreso nell'intervallo $[40, 60]$. Sia p la probabilità di ottenere testa in un lancio della moneta.

1. Specificate in funzione di p l'ipotesi che la moneta sia bilanciata.
2. Calcolate il valore approssimato della probabilità di rifiutare l'ipotesi che la moneta sia bilanciata, quando essa in realtà lo è (*errore di prima specie*) utilizzando la regola decisionale sopra descritta.
3. Calcolate il valore approssimato della probabilità di accettare l'ipotesi che la moneta sia bilanciata quando in realtà essa è truccata (*errore di seconda specie*) con $p = 0.7$, utilizzando la regola decisionale sopra descritta.

Con la precedente moneta si supponga di giocare il seguente gioco in due lanci. Nel primo lancio se esce testa il gioco termina in pareggio, se invece esce croce si rilancia la moneta e, nel secondo lancio, se esce testa si perde 1 €, se esce croce si vince 1€. Sia X la variabile aleatoria che indica la somma vinta nel gioco appena descritto.

4. Scrivete la funzione di densità di X.
5. Calcolate lo stimatore di massima verosimiglianza di p basato sul campione osservato $(1, 0, 1, -1, -1, 0)$ di 6 somme vinte.

SOLUZIONE Siano $Y_1, \dots, Y_n$ le v.a. di Bernoulli che descrivono il risultato degli $n = 100$ lanci, cioè $Y_1, \dots, Y_n$ i.i.d. $\sim Be(p)$, dove p è la probabilità di ottenere testa in un lancio.

1. La moneta è bilanciata se $p = p_0 := 1/2$.
2. Si tratta di trovare $\mathbb{P}(\text{“concludere che la moneta sia truccata”}|\text{“la moneta è bilanciata”})$.

   In base alla regola decisionale possiamo scrivere questa probabilità di errore di I specie come

   $$\alpha = \mathbb{P}\Big(\sum_1^n Y_i < 40, \sum_1^n Y_i > 60 \Big| p = 0.5\Big) = \mathbb{P}\Big(\sum_1^n Y_i \le 39 \Big| p = 0.5\Big) + 1 - \mathbb{P}\Big(\sum_1^n Y_i \le 60 \Big| p = 0.5\Big).$$

   Si osservi ora che nel nostro caso è verificata la regola empirica : $n > 50$, $np_0 > 5$, $n(1-p_0) > 5$; pertanto, dal TCL segue che la distribuzione di $\sum_1^n Y_i$ è approssimativamente $\mathcal{N}(np_0, np_0(1-p_0)) = \mathcal{N}(50, 25)$. Otteniamo dunque

   $$\alpha \simeq \Phi\Big(\frac{39-50}{5}\Big) + 1 - \Phi\Big(\frac{60-50}{5}\Big) = \Phi(-2.2) + 1 - \Phi(2) \simeq 0.0139 + 1 - 0.9772 = 0.0367$$

   senza correzione di continuità e

   $$\alpha \simeq \Phi\Big(\frac{39.5-50}{5}\Big) + 1 - \Phi\Big(\frac{60.5-50}{5}\Big) = \Phi(-2.1) + 1 - \Phi(2.1) = 2(1 - 0.9821) \simeq 0.0358$$

   con correzione di continuità. Il valore esatto è $\alpha = 0.0352$, calcolato sapendo che $\sum_1^n Y_i \sim Bin(n, p_0)$ quando p_0 è il “vero” valore di p.
3. Si tratta di trovare $\beta(0.7) = \mathbb{P}(40 \le \sum_1^n Y_i \le 60 | p = 0.7)$. Per il TCL, in questo caso $\sum_1^n Y_i \sim \mathcal{N}(70, 21)$ approssimativamente; pertanto, con correzione di continuità,

   $$\beta(0.7) = \Phi(-2.07) - \Phi(-6.66) \simeq \Phi(-2.07) \simeq 0.0192;$$

   il valore approssimato senza correzione di continuità è $\beta(0.7) \simeq \Phi(-2.18) \simeq 0.0146$. Il valore esatto è $\beta(0.7) = 0.0210$, calcolato sapendo che $\sum_1^n Y_i \sim Bin(n, 0.7)$ quando $p = 0.7$.

4. Indichiamo con T_i e C_i, $i = 1, 2$, i risultati (testa o croce) dell'i-esimo lancio della moneta; per l'indipendenza dei lanci si ottiene

$$\mathbb{P}(X = -1) = \mathbb{P}(C_1 \cap T_2) = p(1-p), \qquad \mathbb{P}(X = 0) = \mathbb{P}(T_1) = p, \qquad \mathbb{P}(X = 1) = \mathbb{P}(C_1 \cap C_2) = (1-p)^2.$$

5. Sia $X_1, \ldots, X_6$ un campione dalla densità al punto 4; la funzione di verosimiglianza sulla base dei dati forniti dal testo è, per $p \in [0, 1]$,

$$\begin{aligned} L(p; x_1, \ldots, x_6) &= \mathbb{P}(X_1 = 1)\mathbb{P}(X_2 = 0)\mathbb{P}(X_3 = 1)\mathbb{P}(X_4 = -1)\mathbb{P}(X_5 = -1)\mathbb{P}(X_6 = 0) \\ &= p^4(1-p)^6. \end{aligned}$$

Differenziando rispetto a p la funzione $\log L(p)$ si ottiene

$$\frac{4}{p} - \frac{6}{1-p} \geq 0 \quad \text{ovvero } 4(1-p) - 6p \geq 0 \quad \text{cioè } 10p - 4 \leq 0.$$

Pertanto, il punto $\hat{p} = \frac{4}{10} = 0.4$ è la stima di massima verosimiglianza di p.

∎

Esercizio 3.2 Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla popolazione di densità

$$f(x, \theta) = \begin{cases} \frac{2\theta^2}{x^3} & \text{se } x \geq \theta \\ 0 & \text{se } x < \theta, \end{cases}$$

con θ parametro positivo incognito.

1. Determinate gli stimatori di θ e di θ^2 usando il metodo di massima verosimiglianza.
2. Calcolate la funzione di ripartizione e la densità dello stimatore di massima verosimiglianza di θ.
3. Lo stimatore di massima verosimiglianza di θ è non distorto? È asintoticamente non distorto? Giustificate le risposte.
4. Ricavate lo stimatore di θ col metodo dei momenti.
5. Confrontate gli errori quadratici medi degli stimatori per θ di massima verosimiglianza e dei momenti, ricavati ai punti precedenti. Quale dei due stimatori è preferibile e perché?

Soluzione

1. La funzione di verosimiglianza del campione $X_1, \ldots, X_n$ è

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{2\theta^2}{x_i^3} \mathbf{1}_{(\theta, +\infty)}(x_i) = 2^n \Big(\prod_{i=1}^{n} x_i^{-3} \Big) \theta^{2n} \mathbf{1}_{(0, x_{(1)})}(\theta) \propto \theta^{2n} \mathbf{1}_{(0, x_{(1)})}(\theta),$$

dove $x_{(1)} = \min_{i=1,\ldots,n}(x_i)$. Pertanto, il punto di massimo di $L(\theta)$ è $x_{(1)}$. Quindi

$$\hat{\theta}_n = X_{(1)}, \quad \widehat{\theta^2_n} = (X_{(1)})^2.$$

2. La funzione di ripartizione di $\hat{\theta}_n$ è $F_{X_{(1)}}(x) = 1 - P(X_{(1)} > x) = 1 - \prod_1^n P(X_i > x) = 1 - (1 - F_{X_1}(x))^n$, dove $F_{X_1}(x) = \int_\theta^x 2\theta^2/u^3 du = 1 - \theta^2/x^2$ se $x > \theta$. Dunque

$$F_{X_{(1)}}(x) = \Big(1 - \frac{\theta^{2n}}{x^{2n}}\Big) \mathbf{1}_{(\theta, +\infty)}(x), \text{ da cui si ricava } f_{X_{(1)}}(x) = 2n \frac{\theta^{2n}}{x^{2n+1}} \mathbf{1}_{(\theta, +\infty)}(x).$$

3. Si ricava $\mathrm{E}(X_{(1)}) = \int_\theta^{+\infty} x \cdot 2n \frac{\theta^{2n}}{x^{2n+1}} dx = \frac{2n}{2n-1}\theta$. Dunque lo stimatore di massima verosimiglianza $\hat{\theta}$ è distorto, ma tuttavia

$$\lim_{n \to \infty} \hat{\theta}_n = \lim_{n \to \infty} \frac{2n}{2n-1}\theta = \theta, \quad \forall \theta > 0,$$

cioè lo stimatore $\hat{\theta}_n$ è asintoticamente non distorto.

4. Poiché $\mathrm{E}(X_1) = \mathrm{E}(X_{(1)})$ con $n = 1$, dal punto 3 si trova che $\mathrm{E}(X_1) = 2\theta$. Dall'equazione $\bar{X}_n = \mathrm{E}(X_1)$ si ricava lo stimatore dei momenti
$$\widetilde{\theta} = \frac{\bar{X}_n}{2}.$$
Si noti che $\widetilde{\theta}$ è stimatore non distorto per θ.

5. Per quanto riguarda lo stimatore $\widetilde{\theta}$ del metodo dei momenti, poiché esso è non distorto,
$$\mathrm{MSE}(\widetilde{\theta}) = \mathrm{Var}(\widetilde{\theta}) = \mathrm{Var}(\frac{\bar{X}_n}{2}) = \frac{\mathrm{Var}(X_1)}{4n} = +\infty,$$
perché $\mathrm{E}(X_1^2) = \int_\theta^\infty x^2 \frac{\theta^2}{x^3} dx = +\infty$.
Per quanto riguarda $\hat{\theta}$, invece, vale
$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \Big(\mathrm{E}(\hat{\theta}) - \theta\Big)^2 = \mathrm{Var}(X_{(1)}) + \Big(\frac{\theta}{2n-1}\Big)^2;$$
d'altro canto $\mathrm{E}(X_{(1)}^2) = \int_\theta^\infty x^2 2n \frac{\theta^{2n}}{x^{2n+1}} dx = \frac{n}{n-1}\theta^2$ se $n \geq 2$. Pertanto
$$\mathrm{MSE}(\hat{\theta}) = \frac{n}{(n-1)(2n-1)^2}\theta^2 + \frac{1}{(2n-1)^2}\theta^2 = \frac{\theta^2}{(n-1)(2n-1)}, \text{ se } n \geq 2.$$
In conclusione, se si sceglie lo stimatore in base all'errore quadratico medio, è preferibile $\hat{\theta}$ a $\widetilde{\theta}$, con un campione di ampiezza maggiore di 1. ∎

Esercizio 3.3 Una macchina dovrebbe tagliare del filo metallico in pezzi che hanno una lunghezza che può essere rappresentata da una variabile aleatoria X gaussiana con media 10.5 cm e deviazione standard 0.15 cm. Per verificare il corretto funzionamento del macchinario sono stati scelti a caso 16 pezzi da un lotto numeroso e le misure ottenute, espresse in cm ed ordinate, sono le seguenti:

$$10.1, 10.2, 10.2, 10.3, 10.3, 10.4, 10.4, 10.5, 10.5, 10.5, 10.6, 10.6, 10.7, 10.7, 10.8, 10.9.$$

Sulla base di questi dati, ci si propone di stabilire se c'è evidenza sperimentale che la macchina non funzioni correttamente.

1. Specificate ipotesi nulla e ipotesi alternativa.
2. Ricavate la funzione di ripartizione empirica sulla base del campione osservato.
3. Costruite un opportuno test per verificare le ipotesi specificate al punto 1, fissando un livello di significatività $\alpha = 10\%$. Cosa decidete sulla base del campione osservato?
4. Ricavate un intervallo di confidenza di livello 90% per $F_X(10.5)$.

Soluzione L'estrazione a caso da un campione numeroso garantisce che le misure ottenute costituiscono un campione i.i.d. $X_1, \ldots, X_{16}$ dalla funzione di ripartizione F_X.

1. Si tratta di verificare le ipotesi $H_0 : F_X(x) = F_0(x)$ per ogni x vs $H_1 : F_X(x) \neq F_0(x)$ per almeno una x, dove $F_0(x) = \Phi(\frac{x-10.5}{0.15})$ e Φ è la funzione di ripartizione di una gaussiana standard.
2. È facile ricavare che
$$F_n(x) = \begin{cases} 0 & \text{se } x < 10.1, \\ 1/16 & \text{se } 10.1 \leq x < 10.2, \\ 3/16 & \text{se } 10.2 \leq x < 10.3, \\ 5/16 & \text{se } 10.3 \leq x < 10.4, \\ 7/16 & \text{se } 10.4 \leq x < 10.5, \\ 10/16 & \text{se } 10.5 \leq x < 10.6, \\ 12/16 & \text{se } 10.6 \leq x < 10.7, \\ 14/16 & \text{se } 10.7 \leq x < 10.8, \\ 15/16 & \text{se } 10.8 \leq x < 10.9, \\ 1 & \text{se } x > 10.9. \end{cases}$$

3. Usiamo il test di *Kolmogorov-Smirnov*. È facile verificare dalla tabella

x	10.1	10.2	10.3	10.4	10.5	10.6	10.7	10.8	10.9
$F_n(x)$	0.0625	0.1875	0.3125	0.4375	0.6250	0.7500	0.8750	0.9375	1.
$F_0(x)$	0.0038	0.0228	0.0912	0.2525	0.5	0.7475	0.9088	0.9772	0.9962

   che la statistica $D_n = \max_x |F_n(x) - F_0(x)| = 0.2213$. Fissato il valore $\alpha = 0.1$ si ricava dalle tavole che $q_{n,1-\alpha} = q_{16,0.9} = 0.2947$. Poiché $D_n < q_{n,1-\alpha}$, non si rifiuta l'ipotesi nulla che la macchina funzioni correttamente.

4. Sappiamo che le statistiche $L(x) = \max\{F_n(x) - q_{n,\gamma}, 0\}$ e $U(x) = \min\{F_n(x) + q_{n,\gamma}, 1\}$, sono, rispettivamente, il limite inferiore e superiore di un intervallo di confidenza di livello $\gamma = 0.9$ per $F_0(x)$. Sostituendo i valori numerici otteniamo l'intervallo $[0.3303, 0.9197]$.

∎

4 STATISTICA per ING INF TEL [2L] A. Barchielli (I. Epifani), A. Guglielmi 27.09.07

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

Esercizio 4.1 Un apparato di inoculazione automatica è stato progettato per rimpiazzare le siringhe nella somministrazione di vaccini. L'apparato può essere predisposto per iniettare differenti quantità di siero, ma a causa di fluttuazioni casuali la quantità di siero realmente iniettata è distribuita normalmente con una media μ_0 nota e uguale alla quantità desiderata e con varianza σ^2 non nota. Si è stabilito che l'apparato sarebbe troppo pericoloso se σ fosse maggiore od uguale a 0.10. Un campione di 50 iniezioni fornisce un valore osservato $s_0^2 = (0.08)^2$ per la statistica $S_0^2 = \frac{1}{50}\sum_{i=1}^{50}\left(X_i - \mu_0\right)^2$; ci si pone il problema di verificare se l'apparato sia o non sia pericoloso.

1. Specificate ipotesi nulla e ipotesi alternativa in modo che l'errore di prima specie sia quello che si commetterebbe ritenendo l'apparato innocuo quando invece non lo è.
2. Proponete un test per verificare le ipotesi specificate al punto 1. Scrivete esplicitamente la regione di rifiuto per un livello di significatività $\alpha = 5\%$; cosa decidete sulla base dei dati circa la pericolosità dell'apparato di inoculazione?
3. Calcolate il p-value del test proposto al punto 2. Quali sono le conclusioni del test se $\alpha = 1\%$?
4. Sulla base della regione di rifiuto determinata al punto 2, fornite un intervallo di confidenza unilatero per σ^2 di livello $\gamma = 95\%$.

Soluzione

1. Poiché l'errore di prima specie è rifiutare l'ipotesi nulla quando essa in realtà è vera, H_0 deve tradurre l'ipotesi "l'apparato è pericoloso"; dunque
$$H_0 : \ \sigma \geq 0.10 \ \text{ contro } H_1 : \ \sigma < 0.10.$$
2. Si tratta di un test sulla varianza di una popolazione $\mathrm{N}(\mu_0, \sigma^2)$, con media μ_0 nota, sulla base di un campione di ampiezza $n = 50$, per verificare le ipotesi
$$H_0 : \ \sigma^2 \geq \sigma_0^2 \ \text{ contro } H_1 : \ \sigma^2 < \sigma_0^2,$$
con $\sigma_0^2 = (0.1)^2 = 0.01$. La statistica test in questo caso è
$$\frac{nS_0^2}{\sigma_0^2} \sim \chi_n^2,$$
dove $S_0^2 = 1/n\sum_{i=1}^n (X_i - \mu_0)^2$, e χ_n^2 indica la distribuzione chi-quadrato con n gradi di libertà. La regione di rifiuto per il test in esame è $G = \left\{(x_1, \ldots, x_n) : \frac{ns_0^2}{\sigma_0^2} \leq \chi_n^2(\alpha)\right\}$. Poiché $\frac{ns_0^2}{\sigma_0^2} = 32 < \chi_n^2(\alpha) = \chi_{50}^2(0.05) = 34.764$, ad un livello di significatività del 5%, rifiutiamo l'ipotesi nulla e accettiamo l'ipotesi che l'apparato inoculatore non sia pericoloso.
3. Il p-value per il test proposto al punto precedente è dato da $p = F_n(\frac{ns_0^2}{\sigma_0^2})$, dove $F_n(\cdot)$ è la funzione di ripartizione di una χ_n^2 e $n = 50$. Il valore esatto di p è 0.0223, ma dalla sola lettura della tavole otteniamo che
$$0.01 < p < 0.025.$$
Quindi sulla base dei dati, ad un livello di significatività dell'1%, non possiamo rifiutare l'ipotesi che l'apparato inoculatore sia pericoloso.
4. Sulla base della dualità fra test delle ipotesi e intervalli di confidenza, un intervallo di confidenza al livello $\gamma = 1 - \alpha$ per il parametro σ^2 è dato dall'insieme
$$\left\{\sigma_0^2 : \frac{nS_0^2}{\sigma_0^2} > \chi_n^2(\alpha)\right\} = \left\{\sigma_0^2 : 0 < \sigma_0^2 < \frac{nS_0^2}{\chi_n^2(\alpha)}\right\};$$
sulla base dei dati campionari e di $\chi_n^2(\alpha) = \chi_{50}^2(0.05) = 34.764$, otteniamo $[0, 0.0092]$ come intervallo di confidenza per σ^2 di livello 95%. ∎

Esercizio 4.2 Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla popolazione di densità $f(x;\theta) = \theta x^{\theta-1}\mathbf{1}_{(0,1)}(x)$, con θ parametro positivo incognito.

1. Determinate la distribuzione di $Y_i = -\log X_i$ per ogni $i = 1, \ldots, n$ (log indica il logaritmo in base naturale). Quindi deducete la distribuzione di $T = -1/n \sum_1^n \log X_i$.
2. Calcolate $\mathrm{E}(T)$ e $\mathrm{Var}(T)$.
3. Stabilite se T è stimatore efficiente per la caratteristica $\kappa(\theta) = 1/\theta$.
4. Usate i risultati trovati al punto 1 per calcolare un intervallo di confidenza esatto per $\kappa(\theta) = 1/\theta$ unilatero della forma $(c, +\infty)$ di livello $\gamma = 95\%$ sulla base del campione di osservazioni 0.77, 0.64, 0.53, 0.16, 0.99 estratte da $f(x;\theta)$.

Soluzione

1. Le v.a. $\{Y_j\}$ sono i.i.d. e positive quasi certamente (q.c.), con funzione di ripartizione

$$F_{Y_1}(y) = P(-\log X_1 \le y) = 1 - F_{X_1}(\mathrm{e}^y) = \int_{\mathrm{e}^{-y}}^1 \theta x^{\theta-1} dx = 1 - \mathrm{e}^{-\theta y}, \quad y > 0.$$

   Quindi $\{Y_j\}$ sono i.i.d. $\sim Exp(1/\theta)$ e quindi, per le proprietà della famiglia delle leggi gamma, $T = 1/n \sum_1^n Y_j = \bar{Y}$ ha densità $\Gamma(n, \frac{1}{n\theta})$.

2. Dalla distribuzione di T segue che

$$\mathrm{E}(T) = \mathrm{E}(\bar{Y}) = \mathrm{E}(Y_1) = \frac{1}{\theta}, \quad \mathrm{Var}(T) = \frac{\mathrm{Var}(Y_1)}{n} = \frac{1}{n\theta^2}.$$

3. T è uno stimatore non distorto per $\kappa(\theta) = 1/\theta$ a varianza finita da quanto visto al punto 2. La funzione di verosimiglianza del campione $X_1, \ldots, X_n$ è

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^n (\theta x_i^{\theta-1}) = \theta^n (\prod_1^n x_i)^{\theta-1}$$

   da cui deriviamo che $\log L(\theta; X_1, \ldots, X_n) = n \log \theta + (\theta - 1) \sum_{i=1}^n \log X_i$ e

$$\frac{\partial}{\partial\theta} \log L(\theta; X_1, \ldots, X_n) = \frac{n}{\theta} + \sum_{i=1}^n \log X_i = -n\Big(- \frac{\sum_{i=1}^n \log X_i}{n} - \frac{1}{\theta}\Big) = -n(T - \kappa(\theta)) \quad \text{q.c.}$$

   Dalla diseguaglianza di Fréchet-Cramer-Rao segue l'efficienza dello stimatore T.

4. Poiché $T \sim \Gamma\left(n, \frac{1}{n\theta}\right)$, allora $2n\theta T \sim \Gamma\left(\frac{2n}{2}, 2\right) = \chi^2_{2n}$ e $\Pr\left(2n\theta T \le \chi^2_{2n}(0.95)\right) = 0.95$. Quindi un IC per $1/\theta$ esatto unilatero di forma $(c, +\infty)$ è dato da $\left(\frac{2nT}{\chi^2_{2n}(0.95)}, +\infty\right)$. Con i dati a nostra disposizione abbiamo $t = 0.637$, $\chi^2_{10}(0.95) = 18.307$ e $\left(\frac{2nt}{\chi^2_{2n}(0.95)}, +\infty\right) = (0.348, +\infty)$. ∎

Esercizio 4.3 Su 250 transistor a giunzione estratti a caso da un lotto numeroso, 103 hanno avuto durata inferiore alle 30 ore, 78 l'hanno avuta tra le 30 e le 60, 33 tra le 60 e le 90 ore, e 36 oltre le 90 ore. Si vuole verificare l'ipotesi che la durata T di questi transistor abbia distribuzione esponenziale di media pari a 50 ore, sulla base dei dati a disposizione.

1. Calcolate la probabilità che la durata di un transistor sia inferiore a 30 ore, la probabilità che sia compresa tra 30 e 60 ore e la probabilità che sia superiore alle 90 ore, sotto l'ipotesi che T abbia distribuzione esponenziale di media 50.
2. Sulla base dei dati a disposizione, impostate un test per verificare l'ipotesi sopra descritta sulla distribuzione di T, specificando ipotesi nulla e ipotesi alternativa. Quali conclusioni raggiungete per un livello di significatività $\alpha = 7.5\%$?
3. Calcolate un intervallo di confidenza bilatero di livello (approssimato) $\gamma = 90\%$ per la percentuale, sull'intero lotto, di transistor con durata superiore alle 90 ore.

Soluzione

1. Sotto l'ipotesi che $T \sim Exp(50)$ (si noti che è una distribuzione assolutamente continua) si ha
$$\Pr(T < 30) = F_T(30) = 1 - \mathrm{e}^{-30/50} \simeq 0.451, \quad \Pr(30 < T < 60) = F_T(60) - F_T(30) \simeq 0.248,$$
$$\Pr(T > 90) = 1 - F_T(90) = \mathrm{e}^{-90/50} \simeq 0.165.$$
2. L'estrazione a caso da un campione numeroso garantisce che le durate dei transistor costituiscano un campione i.i.d. $(T_1, \ldots, T_{250})$ dalla funzione di ripartizione F_T.

   Si tratta di impostare un test per le ipotesi
$$H_0 : T \sim Exp(50) \text{ contro } H_1 : \ T \not\sim Exp(50).$$
   I dati a nostra disposizione sono raggruppati nelle $k = 4$ classi $A_1 = (0, 30]$, $A_2 = (30, 60]$, $A_3 = (60, 90]$, $A_4 = (90, +\infty)$; pertanto possiamo impostare un test χ^2 di adattamento per le ipotesi
$$H_0 : \ p_i = p_{0i}, i = 1, 2, 3, 4 \quad \text{vs.} \quad H_1 : \ p_i \neq p_{0i} \text{ per almeno una } i$$
   dove $p_{01} = \Pr_{H_0}(A_1) = 0.451$, $p_{02} = \Pr_{H_0}(A_2) = 0.248$ e $p_{04} = \Pr_{H_0}(A_4) = 0.165$ sono state calcolate al punto 1 e $p_{03} = 1 - (p_{01} + p_{02} + p_{04}) = 0.136$. La statistica test è
$$Q_n = \sum_{i=1}^{k} \frac{(N_i - np_{0i})^2}{np_{0,i}} \sim \chi^2_{k-1} = \chi^2_3$$
   per n sufficientemente grande e la regione di rifiuto è $\{(N_1, \ldots, N_4) : Q_n > \chi^2_{k-1}(1 - \alpha)\}$. In questo caso la regola empirica $n > 50$, $n \cdot \min(p_{0i}) > 5$ è verificata. Dalla tabella

A_i	$(0, 30]$	$(30, 60]$	$(60, 90]$	$(90, +\infty)$
p_{0i}	0.451	0.248	0.136	0.165
n_i	103	78	33	36
$np_{0,i}$	112.75	62	34	41.35

   si ricava
$$q_n = \frac{103^2}{112.75} + \frac{78^2}{62} + \frac{33^2}{34} + \frac{36^2}{41.25} - 250 = 5.670.$$
   Il valore esatto del p-value è $p = 1 - F_3(5.67) = 0.129$, dove F_3 indica la funzione di ripartizione della χ^2_3; dalle tavole si può dedurre che $0.8 < F_3(5.670) < 0.875$ e quindi che $0.125 < p < 0.2$. In conclusione, poiché $\alpha = 0.075 < p$, non si può rifiutare H_0 al livello di significatività del 7.5%.
3. Sia p_0 la percentuale, sull'intero lotto, di transistor a giunzione con durata superiore alle 90 ore. Si tratta di ricavare un IC asintotico per p_0 sulla base del campione aleatorio $(Y_1, \ldots, Y_{250})$, dove $Y_i = \mathbf{1}_{A_i}(T_i)$, $i = 1, \ldots, 250$. L'intervallo di confidenza asintotico è
$$\left(\bar{y} - z_{\frac{1+\gamma}{2}} \sqrt{\frac{\bar{y}(1-\bar{y})}{n}}, \bar{y} + z_{\frac{1+\gamma}{2}} \sqrt{\frac{\bar{y}(1-\bar{y})}{n}} \right) \simeq (0.107, 0.181),$$
   con $\bar{y} = 36/250 = 0.144$ e $z_{\frac{1+\gamma}{2}} = z_{0.95} = 1.645$. ∎

Esercizio 4.4 Il numero di gocce di cioccolato in un biscotto di tipo *"Crunch"* dell'azienda *"Mangiasano"* può essere rappresentato da una variabile aleatoria di Poisson $\mathcal{P}(\theta)$, dove θ è un parametro positivo. L'azienda produttrice afferma che il numero medio di gocce di cioccolato in un biscotto è 5, ma una associazione di consumatori sospetta che sia solo 2.5. In qualità di perito, l'associazione di consumatori vi chiede di aiutarla a decidere tra l'ipotesi nulla che l'azienda affermi il vero e l'ipotesi alternativa che il sospetto dell'associazione sia fondato.

1. Costruite il test uniformemente più potente di livello α sulla base di una sola osservazione x_1 per le ipotesi sopra specificate. Fornite esplicitamente la regione critica del test.
2. Se $\alpha = 1\%$ e $x_1 = 2$, quali sono le conclusioni del test costruito al punto 1?
3. Calcolate la probabilità di errore di seconda specie del test considerato se $\alpha = 1\%$.
4. Costruite il test uniformemente più potente di livello α sulla base di un campione aleatorio $(X_1, \ldots, X_n)$ da $\mathcal{P}(\theta)$ con $n > 1$ per le ipotesi sopra specificate. Fornite esplicitamente la regione critica del test, determinando la distribuzione della statistica test (cioè della funzione del campione aleatorio di ampiezza $n > 1$ che descrive la regione di rifiuto del test considerato).

Soluzione

1. Deduciamo dal testo dell'esercizio di dover impostare il test di Neyman-Pearson per verificare l'ipotesi nulla $H_0 : \theta = 5$ contro l'alternativa $H_1 : \theta = 2.5$ sulla base di un campione di ampiezza 1. Dal Lemma di Neyman-Pearson deriva che la regione critica di ampiezza α è

$$\mathcal{G} = \left\{ x = 0, 1, 2, \ldots : \frac{L_5(x)}{L_{2.5}(x)} \leq \delta \right\} = \left\{ x = 0, 1, 2, \ldots : \mathrm{e}^{-2.5} 2^x \leq \delta \right\} = \{ x = 0, 1, 2, \ldots : x \leq k \}$$

   e k è tale che

$$\Pr_{\theta=5}(X \leq k) = \sum_{j=0}^{k} \mathrm{e}^{-5} \frac{5^j}{j!} \leq \alpha.$$

2. Se $\alpha = 0.01$, la costante k in $\mathcal{G}$ è $k = 0$ perché $\Pr_{\theta=5}(X \leq 0) \simeq 0.007$ e $\Pr_{\theta=5}(X \leq 1) \simeq 0.04$. Dunque $\mathcal{G} = \{0\}$. Poiché $x_1 = 2 \notin \mathcal{G}$, non si può rifiutare l'ipotesi nulla che l'azienda produttrice dica il vero a livello di significatività dell'1%.

3. La probabilità di errore di seconda specie è

$$\Pr(X \notin \mathcal{G} | H_1 \text{ vera}) = \Pr_{\theta=2.5}(X > 0) = 1 - \Pr_{\theta=2.5}(X = 0) = 1 - \mathrm{e}^{-2.5} \frac{2.5^0}{0!} \simeq 0.918.$$

4. Ora dobbiamo impostare il test di Neyman-Pearson per verificare le ipotesi $H_0 : \theta = 5$ contro $H_1 : \theta = 2.5$ sulla base di un campione di ampiezza n. In questo caso, il Lemma di Neyman-Pearson fornisce la regione critica di ampiezza α

$$\mathcal{G} = \left\{ \frac{L_5(x_1, \ldots, x_n)}{L_{2.5}(x_1, \ldots, x_n)} \leq \delta \right\} = \left\{ \mathrm{e}^{-n2.5} 2^{\sum_1^n x_i} \leq \delta \right\} = \left\{ \sum_1^n x_i \leq k \right\}$$

   con k tale che $\Pr_{\theta=5}(\sum_1^n X_i \leq k) \leq \alpha$, dove $(X_1, \ldots, X_n)$ i.i.d. $\sim \mathcal{P}(\theta)$ quando θ è il "vero" valore del parametro. Ora la distribuzione di $\sum_1^n X_i$ si può facilmente ricavare dall'espressione della sua funzione generatrice dei momenti (f.g.m.)

$$m_{X_1 + \cdots + X_n}(t) = \prod_1^n m_{X_i}(t) = \Big(\exp\{\theta(\mathrm{e}^t - 1)\} \Big)^n = \exp\{n\theta(\mathrm{e}^t - 1)\},$$

   che è la f.g.m. di $\mathcal{P}(n\theta)$. ∎

5 STATISTICA per ING INF TEL [2L] I. Epifani, A. Guglielmi 25.02.08

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

Esercizio 5.1 Una grande catena di *fast-food* analizza mensilmente un campione casuale dei suoi famosi panini *big*, con lo scopo di controllare che mediamente pesino al più 600 grammi (gr). In caso contrario, si rendono necessarie delle operazioni di taratura dei macchinari di produzione. Nel mese di gennaio su un campione casuale di 25 panini *big* si è registrato un peso medio campionario pari a 620.0 gr.

Assumendo che il peso dei panini espresso in grammi sia una variabile aleatoria gaussiana, con deviazione standard nota e pari a 80, rispondete alle seguenti domande.

1. Impostate un opportuno test di ipotesi tale che sia al più pari ad $\alpha = 5\%$ la probabilità di commettere l'errore di prima specie di ritenere necessarie operazioni di taratura dei macchinari di produzione che in realtà sono già ben tarati. Alla luce dei dati di gennaio, l'operazione di taratura è necessaria?
2. Sulla base dei dati a disposizione, quanto siete confidenti che la media teorica dei panini *big* di gennaio sia maggiore di 593.68 gr?
3. Fornite analiticamente e rappresentate graficamente la funzione di potenza π del test al punto 1. ($\alpha = 5\%$).

Per il mese di marzo si è deciso di misurare un campione più numeroso di panini *big* al fine di aumentare la potenza del test, ma mantenendo una significatività pari ad $\alpha = 5\%$.

4. Se effettivamente il peso medio dei panini *big* è pari a 640 gr, quanti panini bisogna misurare affinché la potenza sia maggiore di 0.94?

SOLUZIONE Se X_i è il peso dell'i-esimo panino, $X_1, \ldots, X_n$ è un campione casuale estratto da una popolazione gaussiana $\mathcal{N}(\mu, 80^2)$ di media μ incognita.

1. Impostiamo il problema di verifica di ipotesi $H_0 : \mu \leq 600$ contro $H_1 : \mu > 600$. Quindi rifiutiamo se $\bar{X} \geq 600 + z_{1-\alpha}80/\sqrt{25} = 600 + 16z_{1-\alpha}$. Il p-value è dato da

$$\text{p-value} = 1 - \Phi\left(\frac{620-600}{16}\right) = 1 - \Phi(1.25) \simeq 1 - 0.894 = 0.106 = 10.6\%$$

   Ai consueti livelli di significatività (inferiori al 10%, quindi anche 5%) non possiamo rifiutare l'ipotesi nulla H_0 e concludiamo che non sono necessarie operazioni di taratura.

   Alternativamente: se $\alpha = 5\%$, $z_{1-\alpha} = 1.645$ e la regione critica risulta $\{\bar{X} \geq 626.32\}$. Dato che abbiamo trovato $\overline{x} = 620.0$, non possiamo rifiutare l'ipotesi nulla.

2. Siamo confidenti al 95% che i panini *big* superino mediamente i 593.68 gr. Infatti, con i dati a disposizione, in virtù della dualità fra la verifica di ipotesi e gli intervalli di confidenza, segue dal punto 1. che $\{\mu : \bar{x} < \mu + 16z_{1-\alpha}\} = \{\mu : \mu > \bar{x} - 16z_{1-\alpha}\} = (620.0 - 16z_{1-\alpha}, \infty)$ è un IC unilatero a una coda superiore per μ di confidenza $1-\alpha$. Ma, $620.0 - 16z_{1-\alpha} = 593.68$ se, e solo se, $1 - \alpha = 95\%$.

3. La funzione di potenza del test del punto 1. quando si campionano n panini e $\alpha = 5\%$ è data da

$$\pi(\mu) = P_\mu\left(\bar{X} \geq 600 + 1.645 \times \frac{80}{\sqrt{n}}\right) = 1 - \Phi\left(\frac{600-\mu}{80}\sqrt{n} + 1.645\right) = \Phi\left(\frac{\mu-600}{80}\sqrt{n} - 1.645\right), \ \mu > 600$$

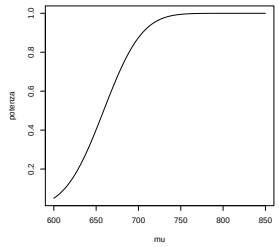
   Il grafico è nella figura che segue

4. La potenza π calcolata in $\mu = 640$ quando si campionano n panini e $\alpha = 5\%$ diventa $\pi(640) = \Phi\left(\sqrt{n}/2 - 1.645\right)$ che è maggiore di 0.94 se, e solo se, $\sqrt{n}/2 - 1.645 > 1.555$, cioè $n > 40.96$. In definitiva, $\pi(640) > 0.94$ per $n \geq 41$ panini. ∎

Esercizio 5.2 Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla densità

$$f(x, \theta) = \frac{1}{4}\mathrm{e}^{-|\frac{x}{2} - \theta|}, \qquad x \in \mathbb{R}$$

dove θ è un parametro reale 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).

Soluzione

1. Per ogni $\theta \in \mathbb{R}$ abbiamo

$$\begin{aligned}\mathrm{E}(\overline{X}) = \mathrm{E}(X_1) &= \int_{-\infty}^{\infty} \frac{x}{4} \mathrm{e}^{-|\frac{x}{2}-\theta|}\, dx = \frac{1}{4}\int_{-\infty}^{2\theta} x\mathrm{e}^{\frac{x}{2}-\theta}\, dx + \frac{1}{4}\int_{2\theta}^{\infty} x\mathrm{e}^{-\frac{x}{2}+\theta}\, dx \\ &= \frac{\mathrm{e}^{-\theta}}{2}\left[(x-2)\mathrm{e}^{x/2}\right]_{-\infty}^{2\theta} + \frac{\mathrm{e}^{\theta}}{2}\left[-(x+2)\mathrm{e}^{-x/2}\right]_{2\theta}^{+\infty} = \theta - 1 + \theta + 1 = 2\theta\end{aligned}$$

   (si noti che X_1 ha distribuzione simmetrica rispetto a 2θ e dunque $\mathrm{E}(X_1)$, che esiste, necessariamente coincide con 2θ) e

$$\mathrm{Var}(\overline{X}) = \frac{\mathrm{Var}(X_1)}{n} = \frac{8}{n}$$

   in quanto

$$\begin{aligned}\mathrm{Var}(X_1) = \mathrm{E}((X_1 - 2\theta)^2) &= \frac{1}{4}\int_{-\infty}^{2\theta} (x-2\theta)^2 \mathrm{e}^{\frac{x}{2}-\theta}\, dx + \frac{1}{4}\int_{2\theta}^{\infty} (x-2\theta)^2 \mathrm{e}^{-\frac{x}{2}+\theta}\, dx \\ &= 2\int_{-\infty}^{0} u^2 \mathrm{e}^u du + 2\int_0^{+\infty} u^2 \mathrm{e}^{-u} du = 4\int_0^{+\infty} u^2 \mathrm{e}^{-u} du = 8.\end{aligned}$$

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

$$\mathrm{MSE}(T) = \mathrm{Var}(T) = \mathrm{Var}(\frac{\overline{X}}{2}) = \frac{8}{4n} = \frac{2}{n}, \quad \forall\theta \in \mathbb{R}$$

3. La funzione di verosimiglianza è

$$L_\theta(x) = \begin{cases} \frac{1}{4}\mathrm{e}^{-\frac{x}{2}+\theta} & \text{se } \theta \le x/2 \\ \frac{1}{4}\mathrm{e}^{\frac{x}{2}-\theta} & \text{se } \theta > x/2 \end{cases}$$

   $L_\theta(x)$ è crescente in $(-\infty, x/2)$, decrescente in $(x/2, +\infty)$ e tende a zero sia per $\theta \to -\infty$ che per $\theta \to +\infty$, quindi $L_\theta(x)$ ha massimo assoluto in $\theta = x/2$. Segue che $\hat{\theta} = X_1/2$ è lo stimatore ML di θ[2]. ∎

[2]In generale si ha: con un numero n dispari di osservazioni, lo stimatore ML di θ è dato dalla mediana campionaria/2 (la mediana è quell'osservazione che lascia alla sua sinistra e alla sua destra metà delle osservazioni). Se, invece, n è pari ogni valore di θ che cade nell'intervallo avente come estremi la metà delle due osservazioni centrali (centrali una volta che tutte siano state ordinate dalla più piccola alla più grande) è stimatore ML.

Esercizio 5.3 Alcuni dei risultati di uno studio statistico sulla distribuzione dei redditi dei lavoratori dipendenti del quartiere `xxx` di Milano, condotto su un campione di 500 persone, sono sintetizzati nella seguente tabella:

A_k	$(0, 11588.0]$	$(11588.0, 13360.0]$	$(13360.0, 15287.0]$	$(15287.0, 18248.0]$	$(18248.0, \infty)$
N_k	175	105	95	65	60

dove A_k indica la classe di reddito annuo netto in euro e N_k il numero di dipendenti con reddito appartenente alla classe A_k. Inoltre, il reddito medio campionario è pari a 13381.1 euro e la varianza campionaria è 16×10^6.

1. Costruite un intervallo di confidenza bilatero di livello approssimato $\gamma = 95\%$ del reddito medio annuo dei lavoratori dipendenti del quartiere `xxx` di Milano.

2. Costruite un intervallo di confidenza bilatero di livello approssimato $\gamma = 95\%$ della percentuale di lavoratori dipendenti del quartiere `xxx` di Milano con reddito (annuo netto) superiore a 13360.0 euro.

Una variabile aleatoria continua X è detta *lognormale di parametri μ e σ* se il suo logaritmo naturale $\ln(X)$ ha densità $\mathcal{N}(\mu, \sigma^2)$.

3. Riportate le classi di reddito A_k in scala logaritmica.

4. Verificate con un opportuno test se il reddito dei lavoratori dipendenti del quartiere `xxx` di Milano può essere modellato con una variabile aleatoria X lognormale di parametri $\mu = 9.5$ e $\sigma = 0.11$.

SOLUZIONE Sia $X_1, \ldots, X_{500}$ il campione casuale dei 500 redditi.

1. Un IC bilatero di livello approssimato $\gamma = 95\%$ del reddito medio annuo dei lavoratori dipendenti del quartiere `xxx` di Milano ha estremi $\overline{x} \mp z_{\frac{1+95\%}{2}} \sqrt{\frac{s^2}{500}}$. Sostituendo i valori campionati abbiamo $(13381.1 - 1.96 \times 4 \times 10^2/\sqrt{5}, 13381.1 + 1.96 \times 4 \times 10^2/\sqrt{5}) = (13030.5, 13731.7)$.

2. La frequenza relativa campionaria di lavoratori dipendenti del quartiere `xxx` di Milano con reddito (annuo netto) superiore a 13360.0 ha valore $(95 + 65 + 60)/500 = 0.44$. Un IC bilatero di livello approssimato $\gamma = 95\%$ della percentuale di lavoratori dipendenti del quartiere `xxx` di Milano con reddito superiore a 13360.0 euro è dato da $\left(0.44 - 1.96 \times \sqrt{\frac{0.44 \times 0.56}{500}}, 0.44 + 1.96 \times \sqrt{\frac{0.44 \times 0.56}{500}}\right) = (0.396, 0.484)$.

3. Le classi A_k in scala logaritmica risultano

A_k	$(0, 11588.0]$	$(11588.0, 13360.0]$	$(13360.0, 15287.0]$	$(15287.0, 18248.0]$	$(18248.0, \infty)$
$\ln(A_k)$	$(-\infty, 9.36]$	$(9.36, 9.50]$	$(9.50, 9.63]$	$(9.63, 9.81]$	$(9.81, \infty)$

4. Considerato che abbiamo un numero "grande" (500) di dati raggruppati e che X ha densità lognormale di parametri $\mu = 9.5$ e $\sigma = 0.11$ se, e solo se, $W := \ln(X)$ ha densità $\mathcal{N}(9.5, 0.11^2)$, allora impostiamo un test asintotico chiquadrato di buon adattamento per verificare: $H_0 : W \sim F_0 = \mathcal{N}(9.5, 0.11^2)$ contro l'alternativa $H_1 : W \not\sim \mathcal{N}(0.95, 0.11^2)$. A tal fine, usiamo le classi di reddito A_k in scala logaritmica e calcoliamo le probabilità attese sotto H_0 di A_k:

A_k	$(0, 11588.0]$	$(11588.0, 13360.0]$	$(13360.0, 15287.0]$	$(15287.0, 18248.0]$	$(18248.0, \infty)$
$\ln(A_k)$	$(-\infty, 9.36]$	$(9.36, 9.50]$	$(9.50, 9.63]$	$(9.63, 9.81]$	$(9.81, \infty)$
F_0	$\Phi(\frac{9.36-9.5}{0.11}) = \Phi(-1.27)$	$\Phi(0)$	$\Phi(1.18)$	$\Phi(2.82)$	
θ_{0k}	$\Phi(-1.27) \simeq 0.102$	$\Phi(0) - \Phi(-1.27) \simeq 0.398$	0.381	0.117	0.002
N_k	175	105	95	65	60

Accorpiamo le ultime due classi dal momento che $500 \times 0.002 = 1$ affinché l'approssimazione asintotica χ^2 con $4 - 1 = 3$ gradi di libertà per la statistica di Pearson $Q_{500} = \sum_{k=1}^{4} \frac{(N_k - 500 \times \theta_{0k})^2}{500\theta_{0k}}$ funzioni. Il valore della statistica Q_{500} è

$$Q_{500} = \sum_{k=1}^{4} \frac{N_k^2}{500\theta_{0k}} - 500 = \frac{175^2/0.102 + 105^2/0.398 + 95^2/0.381 + 125^2/0.118}{500} - 500 \simeq 465.87$$

da cui deriva che il p-value è praticamente nullo. Concludiamo che c'è una forte evidenza sperimentale contro l'ipotesi nulla di redditi lognormali di parametri $\mu = 9.5$ e $\sigma = 0.11$. ∎

Esercizio 5.4 Le pratiche dei rimborsi missione degli afferenti all'istituto `xxx` sono affidate dal signor Antonio e dal signor Biagio. Per valutare l'efficienza dei due, il direttore ha calcolato il tempo che Antonio e Biagio hanno impiegato per espletare le ultime 6 pratiche loro assegnate (6 per ciascuno), ottenendo che Antonio ha impiegato rispettivamente 5.18, 13.43, 6.31, 3.18, 4.91, 11.07 ore e Biagio 5.50, 18.16, 8.14, 9.14, 14.24, 10.72 ore.

1. Alla luce di questi dati, credete che Antonio sia più veloce e quindi più efficiente di Biagio? Per rispondere usate un opportuno test di livello $\alpha = 10\%$.

Supponete ora che il tempo T, espresso in ore, che un (qualunque) dipendente amministrativo impiega per espletare una pratica di rimborso missione si possa modellare come una variabile aleatoria di densità

$$f(t,\theta) = \frac{1}{2\theta\sqrt{t}} \mathrm{e}^{-\frac{\sqrt{t}}{\theta}} \mathbf{1}_{(0,\infty)}(t), \quad \theta > 0$$

con θ incognito. (Il valore di θ è specifico di ogni dipendente; per esempio, potrebbe essere diverso per Antonio e Biagio, se Antonio e Biagio sono diversamente efficienti).

2. Calcolate $P_\theta(T > 7)$ e il suo stimatore di massima verosimiglianza, basato su un campione casuale $T_1, \ldots, T_n$ estratto da $f(t,\theta)$.
3. Stimate la probabilità che per espletare una pratica di rimborso missione nell'istituto `xxx` siano necessarie più di 7 ore (lavorative), nell'ipotesi che il 55% delle pratiche sia affidato al signor Antonio e il rimanente 45% al signor Biagio. (I dati dell'istituto su cui costruire la stima sono quelli sopra riportati).

Soluzione

1. Antonio è più veloce di Biagio se impiega minor tempo ad espletare le pratiche. Non avendo nessuna informazione sulla famiglia di densità da cui sono stati estratti i due campioni di dati, impostiamo il test di omogeneità unilatero non parametrico di Wilcoxon-Mann-Wintney per verificare l'ipotesi nulla H_0 :"*Antonio e Biagio sono ugualmente efficienti*" contro l'alternativa H_1 :"*Antonio è più veloce di Biagio*". Se F_A e F_B indicano le funzioni di ripartizione dei tempi di espletamento delle pratiche di Antonio e Biagio, rispettivamente, allora il problema di verifica di ipotesi diventa

$$H_0 : F_A(x) = F_B(x)\ \forall x \text{ e } H_1 : F_A(x) \geq F_B(x)\ \forall x \text{ (e } F_A(x) > F_B(x) \text{ per almeno qualche } x).$$

   Usiamo la statistica T_A data dalla somma dei ranghi dei tempi impiegati da Antonio per espletare le 6 pratiche: $T_A = 1 + 2 + 3 + 5 + 9 + 10 = 30$. Rifiutiamo H_0 a livello α se $T_A < w_{6,6}(\alpha)$; in questo caso risulta $30 < 31 = w_{6,6}(0.1)$ e quindi, a livello di significatività $\alpha = 0.1$, rifiutiamo H_0 e accettiamo l'ipotesi alternativa che Antonio sia più efficiente di Biagio.

2. $P_\theta(T > 7) = \displaystyle\int_7^\infty \frac{1}{2\theta\sqrt{t}} \mathrm{e}^{-\frac{\sqrt{t}}{\theta}}\, dt = -\,\mathrm{e}^{-\frac{\sqrt{t}}{\theta}}\Big|_7^\infty = \mathrm{e}^{-\frac{\sqrt{7}}{\theta}}.$

   Lo stimatore ML di θ è $\hat\theta = \dfrac{\sum_{j=1}^n \sqrt{T_j}}{n}$. Infatti, la verosimiglianza del campione, per $t_1, \ldots, t_n > 0$, è

$$L_\theta(t_1, \ldots t_n) = \left(\frac{1}{2\theta}\right)^n \frac{1}{\prod_{j=1}^n \sqrt{t_j}} \exp\left\{-\sum_{j=1}^n \frac{\sqrt{t_j}}{\theta}\right\}$$

   da cui deriviamo che

$$\frac{\partial}{\partial\theta} \ln L_\theta(t_1, \ldots t_n) = \frac{n}{\theta^2}\left(\frac{\sum_{j=1}^n \sqrt{t_j}}{n} - \theta\right)$$

   e quindi $\dfrac{\partial}{\partial\theta} \ln L_\theta(t_1, \ldots t_n) \geq 0$ se e solo se $\theta \leq \dfrac{\sum_{j=1}^n \sqrt{t_j}}{n}$.
   Segue che lo stimatore ML di $P_\theta(T > 7)$ è $P_{\hat\theta}(T > 7)$.

3. Risulta che $\hat\theta_A \simeq 2.63$ e $\hat\theta_B \simeq 3.26$ da cui otteniamo che una stima della probabilità di espletare una pratica di rimborso missione nell'istituto `xxx` in più di 7 ore (lavorative) è

$$0.55 P_{\hat\theta_A}(T > 7) + 0.45 P_{\hat\theta_B}(T > 7) = 0.55 \times \mathrm{e}^{-\frac{\sqrt{7}}{2.63}} + 0.45 \times \mathrm{e}^{-\frac{\sqrt{7}}{3.26}} \simeq 0.42. \quad \blacksquare$$