STATISTICA per ING INF TEL [2L] I. Epifani, A. Pievatolo 01.03.2012

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

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

Sia

$$1.030,\ 0.914,\ 0.923,\ 1.010,\ 1.185,\ 0.990$$

la realizzazione di un campione casuale $X_1, \ldots, X_6$ estratto da una popolazione lognormale di parametri $\mu$ e $\sigma > 0$ entrambi incogniti.

1. Determinate uno stimatore $\hat{\mu}$ di $\mu$ e uno $\hat{\sigma}^2$ di $\sigma^2$.
2. Verificate l'ipotesi nulla $H_0 : \sigma \geq 0.4$ contro l'alternativa $H_1 : \sigma < 0.4$, a livello $\alpha = 2.5\%$.
3. Calcolate la potenza del test se $\sigma = 0.1$, o almeno fornite un intervallo in cui tale potenza cade.

In realtà, nutriamo qualche dubbio che le 6 osservazioni fornite provengano da una popolazione lognormale:

4. implementate un opportuno test di significatività 5%, per verificare la bontà di adattamento di un modello lognormale ai dati forniti.

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

$$\log X : \quad 0.03,\ -0.09,\ -0.08,\ 0.01,\ 0.17,\ -0.01$$

1. Gli stimatori ML di $\mu$ e di $\sigma^2$ nel caso di un campione gaussiano $Y_1, \ldots, Y_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$ sono dati dalle statistiche $\hat{\mu} = \bar{Y} = \sum_{j=1}^n \log X_j / n$ e $\sigma^2_{ML} = \sum_{j=1}^n (Y_j - \bar{Y})^2/n = \sum_{j=1}^n (\log X_j - \hat{\mu})^2/n$. Noi però preferiamo aggiustare lo stimatore ML della varianza in modo tale che risulti non distorto, per questo motivo usiamo $\hat{\sigma}^2 = S^2$. Con i dati forniti, abbiamo: $\hat{\mu} = 0.005$ e $\hat{\sigma} = \sqrt{0.0089} \simeq 0.094$.
2. Implementiamo il seguente test sulla varianza di popolazione gaussiana con media incognita per il problema: $H_0 : \sigma^2 \geq 0.16$ contro l'alternativa $H_1 : \sigma^2 < 0.16$. A livello 2.5% rifiutiamo $H_0$ se

$$S^2 \leq \frac{0.16 \times \chi^2_5(0.025)}{5} = \frac{0.16 \times 0.831}{5} \simeq 0.0266\ .$$

   Poiché $S^2 = 0.0089 < 0.0266$, allora rifiutiamo $H_0$ al 2.5%.

3. $$\pi(0.1^2) = \pi(0.01) = P_{\sigma^2=0.01}\left(S^2 \leq 0.0266\right) = P_{\sigma^2=0.01}\left(\frac{5S^2}{0.01} \leq \frac{5 \times 0.0266}{0.01}\right) =$$
$$= F_{\chi^2_5}(13.296) \in \left(F_{\chi^2_5}(12.833), F_{\chi^2_5}(15.086)\right) = (0.975, 0.99)\ .$$

4. Abbiamo un numero "piccolo" di dati (6) non raggruppati e il campione dei rapporti è ipotizzato provenire dalla fdr assolutamente continua lognormale: impostiamo il test di Lilliefors di livello 5% per verificare: $H_0 : Y \sim F_0 = \mathcal{N}$ contro l'alternativa $H_1 : Y \not\sim \mathcal{N}$. Pertanto,

$$\begin{cases} \hat{F}_0(-0.09) = \Phi\left(\frac{-0.09-0.005}{0.094}\right) \simeq \Phi(-1.01) \simeq 0.156 \\ \hat{F}_0(-0.08) \simeq \Phi(-0.90) \simeq 0.184 \\ \hat{F}_0(-0.16) \simeq \Phi(-0.41) \simeq 0.436 \\ \hat{F}_0(0.01) \simeq \Phi(0.05) \simeq 0.520 \\ \hat{F}_0(0.03) \simeq \Phi(0.27) \simeq 0.606 \\ F_0(0.17) \simeq \Phi(1.76) \simeq 0.961 \end{cases}$$

Inoltre

$$|\hat{F}_6(y_i) - \hat{F}_0(y_i)| = (0.011, 0.149, 0.064, 0.147, 0.227, 0.039)$$

e

$$|\hat{F}_6(y_{i-1}) - \hat{F}_0(y_i)| = (0.156, 0.017, 0.103, 0.020, 0.061, 0.128)$$

da cui otteniamo che $D_6 := \sup_{y\in\mathbb{R}} |\hat{F}_6(y) - \hat{F}_0(y)| = 0.227$. Infine, dalle tavole di Lilliefors, abbiamo $q_{D_6}(1 - 5\%) = 0.3245$. Poiché rifiutiamo al livello 5% se $D_6 > 0.3245$, essendo $D_6 = 0.227$ minore di 0.3245 , allora non possiamo rifiutare $H_0$. ∎

**Esercizio 2** Considerate la seguente tabella delle frequenze campionarie assolute di una variabile aleatoria $X$:

| Valori: | 0 | 1 | 2 | 3 | $\geq 4$ |
|---|---|---|---|---|---|
| Frequenze campionarie: | 8 | 40 | 20 | 20 | 12 |

1. Usando i precedenti 100 dati sperimentali, mostrare che la stima con il metodo dei momenti dei due parametri $m$ e $p$ di un modello binomiale $\mathbf{Bin}(m, p)$ restituisce $\hat{m} = 7$ e $\hat{p} = 0.263$.
2. I precedenti 100 dati sperimentali sono compatibili con l'ipotesi che la variabile aleatoria $X$ abbia densità Binomiale? Implementate un opportuno test e calcolate il $p$-value dei dati.
3. Stimate la caratteristica $\kappa := P(X = 0)$.
4. A quale livello di confidenza possiamo affermare che $\kappa := P(X = 0)$ è compresa nell'intervallo $[0.031, 0.129]$?

**Soluzione**

1.
$$M_1 = 1.88, \quad M_2 = 4.92, \quad S^2 \times 99/100 = 1.3856 \quad \text{quindi } \hat{p} = 0.263 \quad \hat{m} = 7$$

2. Avendo:

| Valori: | 0 | 1 | 2 | 3 | $\geq 4$ |
|---|---|---|---|---|---|
| Frequenze campionarie ($N_i$): | 8 | 40 | 20 | 20 | 12 |
| frequenze teoriche sotto $H_0$ ($n\hat{p_{0i}}$): | 11.81 | 29.50 | 31.58 | 18.78 | 8.33 |
| contingenze ($(N_i - n\hat{p_{0i}})^2/n\hat{p_{0i}}$): | 1.229 | 3.737 | 4.246 | 0.079 | 1.617 |

(dove le frequenze teoriche sotto $H_0$ e le contingenze sono approssimate alla seconda cifra decimale), otteniamo che 10.92 è un valore approssimato della statistica di buon adattamento $\chi^2$ di Pearson. Poiché $\chi^2_{5-1-2} = \mathcal{E}(2)$, allora il $p$-value è dato da:

$$p - \text{value} = \mathrm{e}^{-10.92/2} \simeq 0.00425 \simeq 0.4\%$$

Vi è quindi una forte evidenza empirica contro $H_0$: concludiamo che un modello binomiale non fornisce una buona spiegazione di questi dati.

3. $\hat{\kappa} = 8/100 = 0.08$;
4. Un IC bilatero asintotico per $\hat{\kappa}$ di confidenza approssimativamente pari a $\gamma$ è: $\hat{\kappa} \pm z_{\frac{1+\gamma}{2}} \sqrt{\frac{\hat{\kappa}(1-\hat{\kappa})}{n}} = (0.08 - 0.027 z_{\frac{1+\gamma}{2}}, 0.08 + 0.027 z_{\frac{1+\gamma}{2}})$; risolvendo per esempio la diseguaglianza $0.08 + 0.027 z_{\frac{1+\gamma}{2}} = 0.129$, otteniamo
$$\gamma = 2\Phi\left((0.129 - 0.08)/0.027\right) - 1 \simeq 0.9297042 \simeq 93\%. \quad ∎$$

**Esercizio 3** Dagli annuari degli anni scolastici 2000/2001 e 2003/2004 di un liceo abbiamo estratto a caso un campione di sette coppie di fratelli che hanno frequentato il quarto anno uno nel 2000/2001 e l'altro nel 2003/2004 e abbiamo calcolato per ciascuno di loro la media dei voti della pagella di fine anno. I dati ottenuti sono sintetizzati nelle seguenti statistiche:

$$\sum_{j=1}^{7} x_j = 46.8,\ \sum_{j=1}^{7} y_j = 49.5,\ \sum_{j=1}^{7} x_j^2 = 324.76,\ \sum_{j=1}^{7} y_j^2 = 389.31,\ \sum_{j=1}^{7} x_j y_j = 350.82$$

dove $x_j$ indica la media della pagella del fratello più vecchio che ha frequentato nel 2000/2001 e $y_j$ la media della pagella dell'altro.
Per rispondere a tutte le domande che seguono ipotizzate che i dati bivariati $(x_1, y_1), \ldots, (x_7, y_7)$ siano gaussiani.

1. In base a questi dati chi è più bravo, il fratello più giovane o più vecchio? Rispondete usando un opportuno test di livello $\alpha = 5\%$ tale che commettete un errore di primo tipo se concludete erroneamente che il più giovane è più bravo del più vecchio.

2. I rendimenti scolastici dei due fratelli sono indipendenti? Rispondete usando un opportuno test di livello $\alpha = 5\%$.

3. Determinate la probabilità che per una coppia di fratelli con differenza di età 3 anni (scelta a caso) il fratello più giovane sia più bravo del più vecchio. Quindi stimate questa probabilità con uno stimatore che sia funzione delle statistiche precedenti.

**Soluzione** Introduciamo le variabili aleatorie $X, Y$ definite da: $X$ = media dei voti della pagella di fine anno del fratello che ha frequentato nel 2000/2001 e $Y$ = media dei voti della pagella di fine anno del fratello che ha frequentato nel 2003/2004. Per rispondere a tutte le domande usiamo il campione dei 7 dati accoppiati $(x_j, y_j)$, $j = 1, ..., 7$, provenienti da una popolazione bivariata gaussiana

1. Costruiamo un test $t$ di confronto delle medie $\mu_X, \mu_Y$ di questi dati gaussiani accoppiati. Dal testo ricaviamo che l'errore di prima specie consiste nel rifiutare l'ipotesi nulla che il fratello più vecchio sia bravo almeno quanto il più giovane, quando questa ipotesi è vera. Pertanto, dobbiamo verificare:

$$H_0 : \mu_X - \mu_Y \geq 0 \text{ vs } H_1 : \mu_X - \mu_Y < 0\ .$$

Rifiutiamo $H_0$ a livello $\alpha$ se $T = \sqrt{7}(\bar{X} - \bar{Y})/S_D \leq -t_6(1-\alpha)$, 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 $(x_j - y_j)$; abbiamo

$$\bar{x} = \frac{46.8}{7} \simeq 6.686;\ \bar{y} = \frac{49.5}{7} \simeq 7.071;\ \bar{x} - \bar{y} \simeq -0.385$$
$$s_X^2 = \frac{\sum_{j=1}^{7} x_j^2 - 7\bar{x}^2}{6} = \frac{324.76 - 7 \times 6.686^2}{6} \simeq 1.978$$
$$s_Y^2 = \frac{\sum_{j=1}^{7} y_j^2 - 7\bar{y}^2}{6} = \frac{389.31 - 7 \times 7.071^2}{6} \simeq 6.546$$
$$\frac{\sum_{j=1}^{7} x_j y_j - 7\bar{x}\bar{y}}{6} = \frac{350.82 - 7 \times 6.686 \times 7.071}{6} \simeq 3.3138$$
$$s_D^2 = s_X^2 + s_Y^2 - 2 \times \frac{\sum_{j=1}^{7} x_j y_j - 7\bar{x}\bar{y}}{6} \simeq 1.8964$$
$$T = \sqrt{7}\frac{\bar{X} - \bar{Y}}{S_D} = \frac{\sqrt{7} \times (-0.385)}{\sqrt{1.8964}} \simeq -0.7397$$

Se $\alpha = 0.05$ allora $t_6(1-\alpha) = 1.943$. Quindi al livello 5% NON rifiutiamo $H_0$.

2. I dati bivariati sono gaussiani, impostiamo un test $t$ di livello $\alpha = 5\%$ sul coefficiente $\rho$ per il seguente problema di ipotesi: $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$. Il coefficiente di correlazione campionario $R$ vale $R = \frac{3.313}{\sqrt{1.978 \times 6.546}} \simeq 0.921$ La statistica test $\frac{R}{\sqrt{1-R^2}}\sqrt{n-2}$ vale e, sotto $H_0$, ha distribuzine $t$ di student con 5

gradi di libertà. Poiché $|5.286| > t_5(97.5\%) = 2.571$, allora rifiutiamo l'ipotesi $H_0$ che i rendimenti scolastici di due fratelli siano indipendenti.

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

$$P(X < Y) = P(X - Y < 0) = \Phi\left(\frac{0 - (\mu_X - \mu_Y)}{\sqrt{\sigma^2_{X-Y}}}\right)$$

dove $\sigma^2_{X-Y}$ indica la varianza di $X - Y$. Allora, stimiamo la caratteristica $P(X < Y)$ con $\Phi\left(\frac{0-(\bar{X}-\bar{Y})}{\sqrt{S^2_D}}\right) \simeq \Phi(0.28) \simeq 0.6103$. ∎