Politecnico di Milano
Temi d'esame di STATISTICA dell'AA 2009/2010
per allievi ING INF [2L], docente I. Epifani

Contents

1 29.06.10 STATISTICA per ING INF TEL [2L] I. Epifani, scaglione [A-Lz]

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

Esercizio 1.1 Dobbiamo fare inferenza sui tempi di guasto di un sistema formato da 4 componenti connessi in serie, cosicché il sistema funziona se tutti i componenti funzionano. Sappiamo che i componenti funzionano in modo indipendente uno dall'altro, che sono tutti dello stesso tipo A e che i loro tempi di guasto $Y_1, \ldots, Y_4$ (espressi in ore) hanno densità esponenziale di parametro $\theta > 0$, cioè $f(y;\theta) = (1/\theta)\mathrm{e}^{-y/\theta}\mathbf{1}_{(0,\infty)}(y)$ con θ incognito.

Abbiamo così acquistato 280 componenti di tipo A e abbiamo costruito 70 sistemi in serie, tenendo ciascuno attivo fino alla rottura. Per il campione casuale $X_1, \ldots, X_{70}$ delle durate dei 70 sistemi abbiamo ottenuto $\sum_{j=1}^{70} X_j = 693.0$.

1. Verificate che la durata di un intero sistema ha densità esponenziale di parametro $\theta/4$, $\theta > 0$.
2. Determinate uno stimatore $\hat{\theta}_{ML}$ del parametro θ e $\widehat{\kappa}$ della probabilità κ che un sistema funzioni al più 12 ore, usando il metodo di massima verosimiglianza.
3. Verificate che la varianza di $\hat{\theta}_{ML}$ raggiunge il confine di Frechét-Cramer-Rao per la varianza di uno stimatore (non distorto) di θ ma che uno stimatore efficiente per κ non esiste (Giustificate rigorosamente la risposta).
4. Costruite un intervallo di confidenza bilatero di livello 90% per θ.
5. Verificate l'ipotesi nulla $H_0 : \kappa = 0.75$ contro l'alternativa $H_1 : \kappa \neq 0.75$, a una significatività $\alpha = 2.5\%$.

SOLUZIONE

1. Il tempo di guasto X di un sistema ingegneristico ottenuto collegando in serie 4 componenti con tempi di guasto $Y_1, \ldots, Y_4$ i.i.d. $\sim Exp(\theta)$ è $X = \min\{Y_1, \ldots, Y_4\}$ e la sua f.d.r. F_X si ottiene nel seguente modo:

$$1 - F_X(x;\theta) = P(X > x;\theta) = P(\min\{Y_1, \ldots, Y_4\} > x;\theta) = \prod_{j=1}^{4} P(Y_j > x;\theta) = (1 - F_{Y_1}(x;\theta))^4 =$$
$$= [1 - (1 - \mathrm{e}^{-x/\theta})]^4 = \mathrm{e}^{-4x/\theta}$$

da cui deduciamo che la densità di X è $f(x;\theta) = \dfrac{4}{\theta}\mathrm{e}^{-4x/\theta}\mathbf{1}_{(0,\infty)}(x)$, cioè esponenziale di media $\theta/4$.

2-3. La funzione di verosimiglianza del campione casuale $X_1, \ldots, X_{70}$ è data da

$$L_\theta(x_1, \ldots, x_{70}) = \left(\frac{4}{\theta}\right)^n exp\left(-\frac{n4\overline{x}}{\theta}\right)$$

e quindi:

$$\frac{\partial \ln L_\theta(x_1, \ldots, x_{70})}{\partial \theta} = \frac{n}{\theta^2}\left(4\overline{x} - \theta\right) \tag{1}$$

da cui deduciamo che $a)$ $\hat{\theta}_{ML} = 4\overline{X}$ e, per la diseguaglianza di FCR, $b)$ $4\overline{X}$ è stimatore efficiente di θ (effettivamente $\mathrm{E}(4\overline{X}) = 4 \times (\theta/4) = \theta$ e abbiamo anche la non distorsione).

Per quanto riguarda la caratteristica κ definita come la probabilità che un sistema funzioni al più 12 ore, abbiamo che $\kappa = P(X \leq 12;\theta) = 1 - \mathrm{e}^{-4\times 12/\theta}$ e quindi $\widehat{\kappa}_{ML} = 1 - \mathrm{e}^{-4\times 12/\hat{\theta}_{ML}} = 1 - \mathrm{e}^{-12/\overline{x}}$.

Per la (1) non possiamo mai avere $\frac{\partial \ln L_\theta(x_1,\ldots,x_{70})}{\partial\theta} = a(n,\theta)(\widehat{\kappa}_{ML} - \kappa)$ per nessuna scelta della funzione $a(n,\theta)$; inoltre se uno stimatore efficiente di κ esiste allora necessariamente è ML. Considerato tutto ciò, segue che non solo $\widehat{\kappa}_{ML}$ non è stimatore efficiente di κ, ma anche che nessun possibile stimatore di κ è efficiente. Infine, sul nostro campione abbiamo: $\overline{x} = 9.9, \hat{\theta}_{ML} = 39.6$ e $\widehat{\kappa} = 1 - \mathrm{e}^{-1.2\bar{1}} \simeq 0.7018$.

4. Segue dalle proprietà della famiglia di distribuzione gamma che $\hat{\theta}_{ML} \sim \Gamma(70, \theta/70)$ cosicché $8\sum_{j=1}^{70} X_j/\theta \sim \chi^2_{140}$ da cui abbiamo:

$$P\left(\chi^2_{140}(5\%) < \frac{8\sum_{j=1}^{70} X_j}{\theta} < \chi^2_{140}(95\%)\right) = 90\%$$

e

$$P\left(\frac{8\sum_{j=1}^{70} X_j}{\chi^2_{140}(95\%)} < \theta < \frac{8\sum_{j=1}^{70} X_j}{\chi^2_{140}(5\%)}\right) = 90\%$$

Poiché i gradi di libertà sono numerosi:

$$\chi^2_{140}(95\%) \simeq \sqrt{280} \times 1.645 + 140 \simeq 167.5261$$

e

$$\chi^2_{140}(5\%) \simeq \sqrt{280} \times (-1.645) + 140 \simeq 112.4739.$$

Infine l'IC bilatero cercato per θ è $(33.0934, 49.2914)$.

5. Il problema di verifica dell'ipotesi $H_0 : \kappa = 0.75$ contro l'alternativa $H_1 : \kappa \neq 0.75$ è equivalente al problema di ipotesi su θ: $H_0 : \theta = -48/\log(1 - 0.75)$ contro l'alternativa $H_1 : \theta \neq -48/\log(1 - 0.75)$. Il valore $-48/\log(1 - 0.75)$ cade nell'IC precedentemente identificato ($-48/\log(1 - 0.75) \simeq 34.6247$) e per la dualità fra IC e VI accettiamo H_0 non solo a livello 10% ma anche per ogni $\alpha \leq 10\%$ e quindi anche al livello $\alpha = 2.5\%$ richiesto. ∎

Esercizio 1.2 [1] È stato condotto uno studio su come le abitudini alimentari delle donne si modifichino tra l'inverno e l'estate. Si è tenuto sotto osservazione un campione aleatorio di 12 donne durante i mesi di gennaio e luglio 2009, misurando fra le altre cose quale percentuale delle calorie da loro assunte provenisse dai grassi. I risultati ottenuti sono i seguenti:

Gennaio	30.5	28.4	40.2	37.6	36.5	38.8	34.7	29.5	29.7	37.2	41.5	37.0
Luglio	32.2	27.4	28.6	32.4	40.5	26.2	29.4	25.8	36.6	30.3	28.5	32.0

1. Impostate un opportuno test per verificare se la percentuale di calorie ricavate dai grassi cambi nei mesi estivi da quelli invernali, tenendo conto del fatto che si è disposti a commettere un errore di primo tipo al più pari al 5% che la percentuale di calorie ricavate dai grassi sia strettamente minore nel mese di luglio rispetto a quella di gennaio, quando in realtà è vero il contrario. Abbiate cura di specificare $a)$ le ipotesi nulla e alternativa, $b)$ la regione critica (o la statistica test e la regola di decisione) e $c)$ le assunzioni che state facendo sul modello statistico generatore dei dati. Ovviamente indicate anche la decisione cui arrivate con la vostra procedura di verifica.

Concentratevi ora sulle percentuali di calorie ricavate dai grassi nel mese di luglio e

2. costruite un opportuno test per verificare l'ipotesi nulla che la varianza della percentuale di calorie ricavate dai grassi nel mese di luglio sia al più pari a 10.0 contro l'alternativa che sia maggiore di 10.0, a un livello di significatività $\alpha = 5\%$, avendo cura di specificare le ipotesi che state assumendo sul modello statistico. Quindi

3. determinate la potenza del test costruito al punto 2. quando la varianza vale effettivamente 15.0.

SOLUZIONE Indichiamo con L la percentuale di calorie ricavate dai grassi nel mese di luglio e con G quella di gennaio, con μ_L, σ_L^2 media e varianza di L, con μ_G la media G, con F_L, F_G le f.d.r.rispettivamente di L, G, con D la differenza $D = L - G$ e con μ_D e σ_D^2 media e varianza di D. Infine, siano $(L_1, G_1), \ldots, (L_{12}, G_{12})$ il campione casuale di dati accoppiati estratti dalla popolazione (L, G) e $D_1, \ldots, D_{12}$ quello delle differenze D.

1. Dobbiamo verificare il seguente problema: H_0 : "L *tende a essere più grande di* G" contro H_1: "L *tende a essere più piccola di* G" con un test di significatività $\alpha = 5\%$. Inoltre, i dati sono pochi e possiamo eseguire solo test esatti.

   SOLUZIONE 1 PARAMETRICA. Sotto ipotesi di normalità delle differenze: $D_1, \ldots, D_{12}$ i.i.d. $\sim N(\mu_D, \sigma_D^2)$, impostiamo il seguente t-test di confronto fra medie per dati gaussiani accoppiati: $H_0 : \mu_L \geq \mu_G$ versus $H_1 : \mu_L < \mu_G$ o, equivalentemente, $H_0 : \mu_D \geq 0$ versus $H_1 : \mu_D < 0$. La statistica test è $\sqrt{12}\overline{D}/\sqrt{S_D^2}$ che vale -2.338. Infatti, il campione delle differenze è:

   $$(1.7, -1.0, -11.6, -5.2, 4.0, -12.6, -5.3, -3.7, 6.9, -6.9, -13.0, -5.0),$$

   con media campionaria $\overline{D} = -4.308333$, varianza campionaria $S_D^2 \simeq 40.77356$ e $\sqrt{S_D^2} \simeq 6.385418$. Inoltre il p-value $\overline{\alpha}$ è

   $$\overline{\alpha} = P_{\{\mu_D = 0\}}\left(\sqrt{12}\frac{\overline{D}}{\sqrt{S_D^2}} \leq -2.338\right) = F_{11}(-2.338) = 1 - F_{11}(2.338) \in (1\%, 2.5\%)$$

   (in questo punto F_{11} rappresenta la f.d.r. t di student con 11 gradi di libertà). Segue che a livello 5% rifiutiamo H_0. Osservate che comunque non c'è una forte evidenza empirica contro H_0; per esempio, a livello 1% non la rifiutiamo.

   SOLUZIONE 2 NON PARAMETRICA. Se invece assumiamo solo di avere un campione casuale di dati accoppiati $(L_1, G_1), \ldots, (L_{12}, G_{12})$, ma non ci impegniamo circa la forma della distribuzione, allora traduciamo H_0, H_1 in termini di dominanza stocastica nel seguente modo: $H_0 : F_L \leq F_G$ versus $H_1 : F_L > F_G$ ed impostiamo il corrispondente test unilatero non parametrico dei segni di Wilcoxon. La statistica test è $T^+ = n^0$ di coppie con $L > G$ che nel nostro caso ha valore 3 e il p-value $\overline{\overline{\alpha}}$ è

   $$\overline{\overline{\alpha}} = \sum_{j=0}^{3-1} \binom{12}{j} \frac{1}{2^{12}} = \frac{79}{40960} = 0.01928711 \simeq 2\% :$$

   analogamente al test parametrico otteniamo di rifiutare H_0 a livello 5%, ma non c'è forte evidenza empirica contro H_0.

[1] Dati tratti da Ross S.M., Probabilità e statistica per l'ingegneria e le scienze, Apogeo 2008.

2. Poniamoci sotto ipotesi di normalità del campione casuale $L_1, \ldots, L_{12}$, per il quale abbiamo $\overline{L} = 30.825$ e $S_L^2 \simeq 18.52205$. Rifiutiamo $H_0 : \sigma_L^2 \leq 10$ a favore di $H_1 : \sigma_L^2 > 10$ con significatività $\alpha = 5\%$ se $11S_L^2/10 \geq \chi^2_{11}(0.95)$. Poiché $11S_L^2/10 \simeq 20.37425$ e $\chi^2_{11}(0.95) \simeq 19.675$, allora rifiutiamo H_0 a livello 5%. (Il p-value vale circa 4%).

3. $\pi(15) = P_{15}\left(\frac{11S_L^2}{10} \geq 19.675\right) = P_{15}\left(\frac{11S_L^2}{15} \geq \frac{10 \times 19.675}{15}\right) = P_{15}\left(\frac{11S_L^2}{15} \geq 13.117\right) = 1 - F_{11}(13.117) \in (1 - F_{11}(13.701), 1 - F_{11}(12.414)) = (25\%, 33.3\%)$ (potenza esatta calcolata con R: $\pi(15) = 0.2858$). In questo punto 4. F_{11} sta per la f.d.r. chiquadrato con 11 gradi di libertà. ∎

Esercizio 1.3 [2] I valori che seguono rappresentano le lunghezze in millimetri di un campione di 10 granelli presi da una grossa pila di polvere metallica:

$$2.2 \quad 3.4 \quad 1.6 \quad 0.8 \quad 2.7 \quad 3.3 \quad 1.6 \quad 2.8 \quad 2.5 \quad 1.9$$

1. Stabilite con un opportuno test se una densità lognormale si adatti ai dati forniti.

2. Stimate la percentuale di granelli nella pila la cui lunghezza è compresa fra 1.5 e $2.5\,\mathrm{mm}$.

(*Vi ricordiamo che una variabile aleatoria X è detta* lognormale *di parametri μ, σ se il suo logaritmo naturale $\ln X$ è variabile aleatoria gaussiana di media μ e varianza σ^2.*)

SOLUZIONE

1. Considerato che X è *lognormale* di parametri μ, σ se $Y = \ln X \sim \mathcal{N}(\mu, \sigma^2)$, usiamo un test di Lilliefors per la normalità dei dati logaritmici $Y_1, \ldots, Y_{10}$, con $Y_j = \ln X_j$, per $j = 1, \ldots, 10$. Infatti i dati sono continui e in numero esiguo e i parametri della distribuzione $\mathcal{N}(\mu, \sigma^2)$ non sono assegnati. I dati in scala logaritmica e ordinati dal più piccolo al più grande sono:

y_i : -0.2231 0.4700 0.4700 0.6419 0.7885 0.9163 0.9933 1.0296 1.1939 1.2238

La media campionaria delle y_i vale $\overline{y} \simeq 0.7504$ e la deviazione standard campionaria $\sqrt{s_Y^2} \simeq 0.4351$ da cui otteniamo per $z_i := (y_i - \overline{y})/\sqrt{s_Y^2}$ i seguenti valori (ordinati e distinti) e la corrispondente funzione di ripartizione empirica (indicata con $\widehat{F}_{10}$):

z_i	-2.24	-0.64	-0.25	0.09	0.38	0.56	0.64	1.02	1.09
$\widehat{F}_{10}(z_i)$	0.1	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\Phi(z_i)$	0.0125	0.2611	0.4013	0.5359	0.6480	0.7123	0.7389	0.8461	0.8621
$\lvert\widehat{F}_{10}(z_i) - \Phi(z_i)\rvert$	0.0875	0.0389	0.0013	0.0359	0.0480	0.0123	0.0611	0.0539	0.1379
$\lvert\widehat{F}_{10}(z_{i-1}) - \Phi(z_i)\rvert$	0.0125	0.1611	0.1013	0.1359	0.1480	0.1123	0.0389	0.0461	0.0379

Deduciamo dalla precedente tabella che la statistica test $D_{10} = \sup_{z \in \mathbb{R}} |\widehat{F}_{10}(z) - \Phi(z)|$ ha valore approssimativamente pari a 0.1611. Dalle tavole di Lilliefors abbiamo che il quantile di ordine $1-0.2$ della statistica di Lilliefors (sotto l'ipotesi H_0 che i dati in scala logaritmica siano gaussiani) è $q(1-0.2) = 0.2171$. Poiché $0.1611 < 0.2171$ allora accettiamo l'ipotesi di dati Y_i normali per ogni $\alpha \leq 20\%$: altrimenti detto, non c'è alcuna evidenza empirica contro la log-normalità dei dati X_i.

(Usando il pacchetto R, "con meno approssimazioni nei conti" otteniamo $D_{10} = 0.1596$ con p-value$= 0.6668$

2. Avendo "accettato" l'ipotesi di log-normalità dei dati X_i la percentuale di granelli nella pila la cui lunghezza è compresa fra 1.5 e $2.5\,\mathrm{mm}$ è:

$$P(1.5 < X < 2.5) = P(\ln 1.5 < Y < \ln 2.5) = \Phi\left(\frac{\ln 1.5 - \mu}{\sigma} < \frac{Y - \mu}{\sigma} < \frac{\ln 2.5 - \mu}{\sigma}\right) = \\ = \Phi(\frac{\ln 2.5 - \mu}{\sigma}) - \Phi(\frac{\ln 1.5 - \mu}{\sigma})$$

e una sua stima è data da

$$\Phi(\frac{\ln 2.5 - \overline{y}}{s}) - \Phi(\frac{\ln 1.5 - \overline{y}}{s}) \simeq \Phi(0.38) - \Phi(-0.79) \simeq 0.6485 - 0.214 = 43.45\% \quad ∎$$

[2] Dati tratti da Ross S.M., Probabilità e statistica per l'ingegneria e le scienze, Apogeo 2008.

2 13.07.10 STATISTICA per ING INF TEL [2L] I. Epifani, scaglione [A-Lz]

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

Esercizio 2.1 Abbiamo estratto un campione casuale $X_1, \ldots, X_{25}$ dalla densità di probabilità

$$f(x, \theta) = \frac{1}{\theta} \mathrm{e}^{1-\frac{x}{\theta}} \mathbf{1}_{(\theta,\infty)}(x),$$

con θ parametro positivo incognito, e abbiamo ottenuto la seguente funzione di ripartizione empirica:

x	2.6	3.8	4.5	5.5	6.6	7.4	19.3
$\hat{F}_{25}(x)$	3/25	12/25	16/25	19/25	22/25	24/25	1

(2)

Indichiamo con $\overline{X}$ la media campionaria.

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.
3. Usate la funzione di ripartizione empirica (2) per fornire una stima di θ (valore numerico dello stimatore trovato al punto precedente) e una stima del suo errore quadratico medio.
4. Determinate lo stimatore di massima verosimiglianza di θ e fornitene anche il suo valore numerico per il campione di dati in (2).

SOLUZIONE

1. Poiché

$$\begin{aligned} \mathrm{E}_\theta(X_1) &= \int_\theta^\infty x \times \frac{1}{\theta} \mathrm{e}^{1-\frac{x}{\theta}}\, dx = \ldots = \theta + \theta \mathrm{e}^1 \mathrm{e}^{-\theta/\theta} = 2\theta; \\ \mathrm{E}_\theta(X_1^2) &= \int_\theta^\infty x^2 \times \frac{1}{\theta} \mathrm{e}^{1-\frac{x}{\theta}}\, dx = \ldots = \theta^2 + 2\theta(2\theta) = 5\theta; \\ \mathrm{Var}_\theta(X_1) &= 5\theta^2 - (2\theta)^2 = \theta^2 \end{aligned}$$

allora $\mathrm{E}(\overline{X}) = 2\theta$ e $\mathrm{Var}(\overline{X}) = \dfrac{\theta^2}{25}$

2. Da $\mathrm{E}(\overline{X}) = 2\theta$ deduciamo che $\hat{\theta} = \overline{X}/2$ ha media θ e quindi è stimatore non distorto di θ. Inoltre, $MSE(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) = \mathrm{Var}(\overline{X}/2) = \theta^2/(4n) = \theta^2/100$.

3. Da $\hat{F}_{25}(x)$ ricaviamo le frequenze relative dei valori campionari: per esempio 2.6 ha frequenza $3/25$, 3.8 ha frequenza $\hat{F}_{25}(3.8) - \hat{F}_{25}(2.6) = 9/25$, eccetera. Allora

$$\bar{x} = \frac{2.6 \times 3 + 3.8 \times 9 + 4.5 \times 4 + 5.5 \times 3 + 6.6 \times 3 + 7.4 \times 2 + 19.3 \times 1}{25} = 5.216$$

e quindi $\hat{\theta} = 5.216/2 = 2.608$ e la stima di $MSE(\hat{\theta})$ risulta $2.608^2/100 = 0.06801664$.

4. La funzione di verosimiglianza è

$$L_\theta(x_1, \ldots, x_{25}) = \frac{e^{25 - \frac{\sum_i x_i}{\theta}}}{\theta^{25}} \mathbf{1}_{(\theta,+\infty)}(\min\{x_1, \ldots, x_n\})$$

il cui punto di massimo non vincolato è in $\bar{x}$. Poiché è sempre vero che $\bar{x} \geq \min\{x_1, \ldots, x_n\}$ e θ non può superare il più piccolo valore osservato, allora $\hat{\theta}_{ML} = \min\{x_1, \ldots, x_n\} = 2.6$ ∎

Esercizio 2.2 Un'azienda di succhi di frutta usa un macchinario che versa il succo nelle bottigliette in un processo produttivo continuo. Il macchinario lavora secondo gli standard quando versa in ciascuna bottiglietta esattamente $33\,\mathrm{ml}$ di succo. Periodicamente si selezionano 49 bottigliette, si misura la media campionaria $\bar{X}$ del contenuto delle 49 bottigliette e si conclude che il macchinario NON rispetta gli standard se $\bar{X}$ si scosta -in più o in meno- di almeno $2.10\,\mathrm{ml}$ da $33\,\mathrm{ml}$. Se ciò si verifica, si interrompe la produzione. Inoltre si sa che le misurazioni del contenuto di succo nelle bottigliette sono variabili aleatorie gaussiane con deviazione standard $\sigma = 6.4\,\mathrm{ml}$.

1. Formalizzate con il linguaggio della verifica di ipotesi la procedura di controllo del funzionamento del macchinario sopra descritta, cioè traducetela mediante un test di ipotesi.
2. Calcolate il livello di significatività α del test di ipotesi formalizzato al punto 1.
3. Fornite l'espressione analitica della funzione di potenza della procedura di verifica di ipotesi descritta.
4. Calcolate la probabilità di "NON interrompere la produzione" quando in realtà il macchinario versa mediamente $3.0\,\mathrm{ml}$ di succo IN MENO dei $33\,\mathrm{ml}$ regolamentari.
5. Determinate quante ulteriori bottigliette, oltre alle 49, bisogna controllare affinché la lunghezza dell'intervallo di confidenza bilatero al 97.86% del contenuto medio di succo sia minore o uguale di $2.10\,\mathrm{ml}$.

SOLUZIONE

1. Le 49 misure $X_1, \ldots, X_{49}$ sono un campione casuale (variabili aleatorie i.i.d.) da una popolazione $\mathcal{N}(\mu; 6.4^2)$; il parametro incognito è la media $\mu \in \mathbb{R}$. Le ipotesi sono $H_0 : \mu = \mu_0 \equiv 33$ contro $H_1 : \mu \neq \mu_0$. La regola di interruzione data nel testo si traduce nella regione critica $\left\{ \left|\overline{X} - \mu_0\right| \geq 2.10 \right\}$, con $\overline{X} = \frac{1}{49}\sum_{j=1}^{49} X_j$.

2. Il livello cercato è

$$\alpha = P_{\mu=33}\left(\left|\overline{X} - 33\right| \geq 2.10\right) = 2\left[1 - \Phi\left(\frac{2.10}{6.4/\sqrt{49}}\right)\right] = 2\left[1 - \Phi\left(2.296875\right)\right] \simeq 2\left[1 - \Phi\left(2.30\right)\right] \simeq 2.14\%.$$

3. La funzione di potenza è la probabilità di rifiutare H_0 se $\mu \neq 33$, cioè

$$\pi(\mu) = P_\mu\left(\left|\overline{X} - 33\right| \geq 2.10\right) = 1 - P_\mu\left(-2.10 \leq \overline{X} - 33 \leq 2.10\right) = \\ = 1 - \left[\Phi\left(\frac{35.10 - \mu}{6.4/\sqrt{49}}\right) - \Phi\left(\frac{30.90 - \mu}{6.4/\sqrt{49}}\right)\right], \qquad \mu \neq 33.$$

4. Dobbiamo calcolare la probabilità di errore di secondo tipo quando $\mu = 30.0$; abbiamo

$$1 - \pi(30) = \Phi\left(\frac{35.1 - 30}{6.4/\sqrt{49}}\right) - \Phi\left(\frac{30.9 - 30}{6.4/\sqrt{49}}\right) = \Phi(5.578125) - \Phi(0.984375) \simeq 1 - 0.8375 = 16.25\%.$$

5. Gli estremi di un intervallo bilatero sono $\overline{x}_n \pm z_{.9893}\,\frac{6.4}{\sqrt{n}}$ e la lunghezza è

$$2 \times z_{0.9893}\,\frac{6.4}{\sqrt{n}} \simeq 2 \times 2.30 \times \frac{6.4}{\sqrt{n}} = \frac{29.44}{\sqrt{n}}$$

Noi vogliamo che questa quantità sia più piccola di 2.10 e otteniamo $\sqrt{n} \geq 29.44/2.10 \simeq 14.02$ da cui $n \geq 197$: dobbiamo controllare altre 148 bottigliette. ∎

Esercizio 2.3 Nell'ambito di uno studio statistico su *reliability* e *performance* di un sistema software aperto di acquisti telematici abbiamo analizzato le richieste giornaliere arrivate nel corso del 2009 (365 giorni). (La *reliability* è la probabilità di fallimento di una richiesta e la *performance* la distribuzione dei tempi di risposta). Abbiamo così registrato $1)$ la percentuale giornaliera X di richieste fallite e $2)$ la media giornaliera Y dei tempi di risposta (espressi in minuti primi) delle richieste soddisfatte. I dati sono sintetizzati nella seguente tabella:

$X \setminus Y$	$(0, 5.0)$	$[5.0, 12.0)$	$[12.0, \infty)$	
$(0, 0.02]$	111	90	69	
$(0.02, 0.05]$	8	5	6	
$(0.05, 0.13]$	12	6	7	
$(0.13, 1]$	28	13	10	

1. Sulla base di questi dati, pensate che le caratteristiche di reliability e performance del sistema software analizzato siano indipendenti? Rispondete impostando un opportuno test di ipotesi.
2. Verificate al 2% l'ipotesi che almeno il 50% delle richieste giornaliere siano soddisfatte (mediamente) in meno di 5 minuti contro l'alternativa che le richieste giornaliere soddisfatte in meno di 5 minuti siano meno del 50%. `Versione più semplice e diretta nei dati: Verificate al 2% l'ipotesi che per metà dell'anno le richieste giornaliere siano soddisfatte (mediamente) in meno di 5 minuti contro l'alternativa che per massimo per metà dell'anno le richieste giornaliere siano soddisfatte in meno di 5 minuti.`
3. Astraendo, la percentuale giornaliera X di richieste fallite può essere pensata come una variabile aleatoria continua. Verificate con un opportuno test se il modello beta dato da

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

si adatti ai dati forniti.

SOLUZIONE Completiamo la tabella delle numerosità, calcolando le numerosità marginali di X e Y:

$X \setminus Y$	$(0, 5.0)$	$[5.0, 12.0)$	$[12.0, \infty)$	$N_{i\cdot}$
$(0, 0.02]$	111	90	69	270
$(0.02, 0.05]$	8	5	6	19
$(0.05, 0.13]$	12	6	7	25
$(0.13, 1]$	28	13	10	51
$N_{\cdot j}$	159	114	92	$n = 365$

1. Impostiamo un test χ^2 di indipendenza per verificare H_0: *"X e Y sono indipendenti"* contro H_1: *"X e Y non sono indipendenti"*. La statistica test è

$$Q_{\text{ind}} = \sum_{i=1}^{4}\sum_{j=1}^{3} \frac{\left(N_{ij} - \frac{N_{i\cdot}N_{\cdot j}}{365}\right)^2}{\frac{N_{i\cdot}N_{\cdot j}}{365}} = 365 \sum_{i=1}^{4}\sum_{j=1}^{3} \frac{{N_{ij}}^2}{N_{i\cdot}N_{\cdot j}} - 365$$

e nel nostro caso troviamo il valore $q_{\text{ind}} \simeq 4.517$. Dato che la statistica test sotto l'ipotesi nulla ha distribuzione limite chiquadro con $(4-1)(3-1) = 6$ gradi di libertà, la cui f.d.r.indichiamo con F_6, allora per il p-value, usando le tabelle, abbiamo $F_6(4.074) = 0.333$ e $F_6(5.348) = 0.500$. Il p-value è $1 - F_6(q_{\text{ind}}) \in (0.500, 0.667)$. Con un p-value così alto non possiamo rifiutare H_0 a nessun livello di significatività ragionevole e accettiamo l'indipendenza di reliability e performance.

2. Chiamiamo p la "*percentuale di richieste giornaliere soddisfatte (mediamente) in meno di 5 minuti*", cioè $p = P(Y < 5)$. La stima di massima verosimiglianza o del metodo dei momenti di p è $\hat{p} = 159/365 \simeq 0.4356$. Il numero delle prove è grande e usiamo un test asintotico: a livello 2% rifiutiamo $H_0 : p \geq 0.5$ a favore di $H_1 : p < 0.5$ se

$$\frac{\hat{p} - 0.5}{\sqrt{0.5 \times 0.5/365}} \leq z_{.02} = -z_{.98} \simeq -2.0537.$$

Poiché $\frac{\hat{p}-0.5}{\sqrt{0.5\times 0.5/365}} \simeq -2.46 < -2.0537$ al livello 2% rifiutiamo H_0. Il p-value approssimato è $\Phi(-2.46) = 1-\Phi(2.46) \simeq 0.69\%$.

3. Conduciamo un test χ^2 di buon adattamento per verificare H_0: *"$X \sim f(x,\theta)$, per qualche $\theta > 0$"* contro H_1: "$X \not\sim f(x,\theta)$". Innanzitutto stimiamo θ con il metodo dei momenti, usando i dati raggruppati su X:

$$M_c = (0.01 \times 270 + 0.035 \times 19 + 0.09 \times 25 + 0.565 \times 51)/365 = 34.43/365 \simeq 0.09433$$

e

$$\mathrm{E}_\theta(X) = \int_0^1 x \times \frac{1}{\theta} x^{\frac{1}{\theta}-1}\, dx = \frac{1}{1+\theta},$$

da cui otteniamo $\hat{\theta} = 1/M_c - 1 = 365/34.43 - 1 \simeq 9.60122$. Inoltre abbiamo:

$$\hat{p}_{0i} = \int_a^b \frac{1}{\hat{\theta}} x^{\frac{1}{\hat{\theta}}-1} = b^{1/\hat{\theta}} - a^{1/\hat{\theta}}, \qquad 0 < a < b < 1.$$

La statistica test è

$$Q_{\text{goodness}} = \sum_{i=1}^{4} \frac{(N_{i\cdot} - 365\hat{p}_{0i})^2}{365\hat{p}_{0i}} = \sum_{i=1}^{4} \frac{N_{i\cdot}^2}{365\hat{p}_{0i}} - 365$$

che sotto l'ipotesi nulla ha distribuzione limite chiquadro con $4-1-1=2$ gradi di libertà, cioè $\mathcal{E}(2)$; nel nostro caso troviamo il valore $q_{\text{goodness}} \simeq 9.649$ e per il p-value abbiamo p-value$\simeq \mathrm{e}^{-9.649/2} \simeq 0.008 = 0.8\%$: rifiutiamo H_0 per ogni $\alpha \geq 0.08\%$. Quindi c'è una forte evidenza empirica contro H_0 e concludiamo che il modello beta specificato non si adatta ai dati. ∎

Write down clearly how the results are derived (not just the results!)

Esercizio 1. Siano $r > 0$ e $\theta > 0$ e sia $X_1, \ldots, X_n$ un campione di dimensione n estratto da una popolazione di densità Gamma di parametri (r, θ), cioè di comune densità

$$f_{(r,\theta)}(x) = \begin{cases} \frac{x^{r-1}}{\theta^r \Gamma(r)} \, \mathrm{e}^{-x/\theta} & x > 0, \\ 0 & x \leq 0; \end{cases}$$

Ricordate che avete densità Gamma, con media e varianza, e funzione Gamma sulla tabella delle distribuzioni notevoli. Supposto r NOTO (pensate ad un valore assegnato, ad esempio $r = 5$):

1. Determinare lo stimatore di massima verosimiglianza T_n di θ basato sul campione $X_1, \ldots, X_n$.
2. Determinare se T_n è uno stimatore non distorto e consistente in media quadratica per θ.
3. Dire, giustificando adeguatamente la risposta, se lo stimatore T_n è asintoticamente gaussiano e in caso affermativo determinarne la media e la varianza asintotica.

Exercise 1. Let $r > 0$ and $\theta > 0$ be two constants, and let $X_1, \ldots, X_n$ a random sample (having dimension n) from a population with distribution Gamma with parameters (r, θ); this means that $X_1, \ldots, X_n$ have the following probability density:

$$f_{(r,\theta)}(x) = \begin{cases} \frac{x^{r-1}}{\theta^r \Gamma(r)} \, \mathrm{e}^{-x/\theta} & x > 0, \\ 0 & x \leq 0; \end{cases}$$

(Remember that you can find the Gamma probability density- together with the Gamma function $\Gamma(r)$-, its mean and its variance, on your table of the main distributions from the lecture notes). Consider r KNOWN (for instance, consider $r = 5$):

1. Using the random sample $X_1, \ldots, X_n$, find the maximum likelihood estimator T_n of θ.
2. Is T_n an unbiased estimator for θ? Is it square mean consistent?
3. Is T_n asymptotically normal? Justify your answer. If yes, find its asymptotic mean and variance.

Esercizio 2. Due gruppi formati ciascuno da 21 programmatori, il gruppo xx e il gruppo yy devono consegnare un software. Per ciascun programmatore viene contato il numero medio giornaliero di righe di programma scritte, ottenendo i due campioni casuali $(x_1, \ldots, x_{21})$ e $(y_1, \ldots, y_{21})$ tra loro indipendenti. Di tali campioni abbiamo a disposizione le statistiche

$$\sum_{i=1}^{21} x_i = 2040.37 \qquad \sum_{i=1}^{21} x_i^2 = 227134 \qquad \sum_{i=1}^{21} y_i = 1886.661 \qquad \sum_{i=1}^{21} y_i^2 = 187065.7$$

e assumiamo che il numero medio giornaliero di righe scritte abbia distribuzione gaussiana.

1. Verificare, al livello del 10%, l'ipotesi che la varianza σ_X^2 di un programmatore del gruppo xx sia uguale alla varianza σ_Y^2 di un programmatore del gruppo yy, contro l'ipotesi che la prima sia maggiore della seconda.
2. Calcolare la funzione di potenza del test precedente nel punto $\sigma_X^2/\sigma_Y^2 = 1.25$.
3. Stimare con il metodo dei momenti la differenza tra il numero medio giornaliero di righe scritte da un programmatore del gruppo xx e quelle scritte da un programmatore del gruppo yy: $\Delta = \mathrm{E}(X) - \mathrm{E}(Y) = \mu_X - \mu_Y$. Verificare che lo stimatore sia non distorto e consistente. Proporre una stima per la varianza dello stimatore di Δ.

Exercise 2. Consider two groups xx and yy of programmers, having 21 programmers each one. The programmers have to prepare a software. The mean number of daily written lines (of program) of any programmer is counted, obtaining

two independent random samples $(x_1, \ldots, x_{21})$ e $(y_1, \ldots, y_{21})$. From these random samples we have the following statistics:

$$\sum_{i=1}^{21} x_i = 2040.37 \qquad \sum_{i=1}^{21} x_i^2 = 227134 \qquad \sum_{i=1}^{21} y_i = 1886.661 \qquad \sum_{i=1}^{21} y_i^2 = 187065.7$$

Assume that mean number of daily written lines is normally distributed.

1. Verify, at significance level 10%, the hypothesis that the variance σ_X^2 of a programmer of the group xx is equal to the variance σ_Y^2 of a programmer of the group yy, versus the hypothesis that the first one is greater then the second one.
2. Compute the power of the test considered at point 1., when $\sigma_X^2/\sigma_Y^2 = 1.25$.
3. Estimate, with the method of moments, the difference between the means of the mean number of daily written lines of a programmer of the group xx and a programmer of the group yy: $\Delta = \mathrm{E}(X) - \mathrm{E}(Y) = \mu_X - \mu_Y$. Verify that the estimator is unbiased and consistent. Propose an estimate for the variance of the estimator of Δ.

Esercizio 3. In una certa zona, i dati in possesso delle assicurazioni dicono che, in un anno, l'82% degli automobilisti non ha alcun incidente, il 15% ha esattamente un incidente, e il 3% ne ha 2 o più. Su un campione aleatorio di 440 automobilisti laureati in ingegneria, nell'ultimo anno 366 non hanno avuto incidenti, 68 ne hanno avuto uno, 6 ne hanno avuti 2 o più.

1. Si può concludere a livello $\alpha = 5\%$ che la sottopopolazione degli ingegneri presenta un profilo di rischio (ovvero la distribuzione di probabilità del numero di incidenti) diverso da quello generale della zona?
2. Si calcoli il p-value del test costruito al punto 1 e lo si commenti.

Exercise 3. In a certain area, data known by the car insurances say that, during one year, the 82% of the car drivers don't have any accident, the 15% of them have exactly one accident, while the 3% of them have 2 or more accidents. In a random sample of 440 car drivers having a degree in engineering, , during the last year, 366 of them have not had any accident, 68 have had 1 accident, 6 have had 2 or more accidents.

1. Can you deduce, at level $\alpha = 5\%$, that the subgroup of engineering present a risk profile (that is the probability distribution of the number of accidents) different from the general risk profile of that area?
2. Compute the p-value of the test you have formulated in point 1., and provide some comments.

Esercizio 4. Un edicolante ha venduto un *carnet* di 100 biglietti di una lotteria istantanea e 18 di questi sono risultati vincenti.

1. In base al numero di biglietti vincenti trovati nel carnet venduto, calcolare un intervallo di confidenza asintotico al 90% per la proporzione p di biglietti vincenti su tutti quelli prodotti.
2. La Divisione Lotterie aveva chiesto allo stampatore che i biglietti vincenti fossero il 10%. Proporre un test di ampiezza α per l'ipotesi nulla $H_0 : p \leq 10\%$ contro $H_1 : p > 10\%$.

Exercise 4. A newsagent has sold a *carnet* of 100 tickets of an instant lottery, and 18 of them have turned out to be winning tickets.

1. Compute an asymptotic 90% confidence interval for the proportion p of winnings tickets over the total number of tickets, on the base of the number of winning tickets found in the sold *carnet* .
2. The Division of Lotteries had requested to the tickets stamper that the winning tickets were the 10% of the total tickets. Propose a test of level α to verify the null hypothesis $H_0 : p \leq 10\%$ versus $H_1 : p > 10\%$.

SOLUZIONI STATISTICA 06/09/2010 STATISTICS SOLUTIONS

Esercizio 1.

1. La funzione di verosimiglianza, che è funzione del solo parametro θ essendo r noto, è

$$L(\theta; x_1, \dots, x_n)) = \prod_{i=1}^{n} \left(\frac{x_i^{r-1}}{\theta^r \Gamma(r)} \, \mathrm{e}^{-x_i/\theta} \right) = \left(\prod_{i=1}^{n} x_i^{r-1} \right) \mathrm{e}^{-\frac{1}{\theta} \sum_{i=1}^{n} x_i} \frac{1}{\theta^{nr} \Gamma(r)^n}.$$

   Passando al suo logaritmo naturale (log-verosimiglianza) si ottiene

$$l(\theta) = \ln \left(\prod_{i=1}^{n} x_i^{r-1} \right) - \frac{1}{\theta} \sum_{i=1}^{n} x_i - nr \ln(\theta) - n \ln \big(\Gamma(r) \big).$$

   Per calcolare il valore di θ che massimizza $l(\theta)$, calcoliamo la derivata prima e studiamone i punti di annullamento e il segno:

$$l'(\theta) = \frac{1}{\theta^2} \sum_{i=1}^{n} x_i - \frac{nr}{\theta} \geq 0,$$

   che dà $\theta \leq \hat{\theta} := \frac{1}{nr} \sum_{i=1}^{n} x_i$. Quindi $\hat{\theta}$ è il punto di massimo e lo stimatore di massima verosimiglianza di θ è $T_n = \frac{1}{nr} \sum_{i=1}^{n} X_i = \overline{X}_n / r$.

2. Usando media e varianza prese dalle tabelle si ha

$$\mathbb{E}(T_n) = \mathbb{E} \left(\frac{\overline{X}_n}{r} \right) = \frac{\mathbb{E}(X_1)}{r} = \frac{r\theta}{r} = \theta,$$

$$\mathrm{Var}(T_n) = \mathrm{Var} \left(\frac{\overline{X}_n}{r} \right) = \frac{\mathrm{Var} \left(\overline{X}_n \right)}{r^2} = \frac{\mathrm{Var}(X_1)}{nr^2} = \frac{r\theta^2}{nr^2} = \frac{\theta^2}{nr} \to 0.$$

   Quindi lo stimatore T_n è non distorto e consistente in media quadratica per θ.

3. Infine $T_n = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{r}$ è media campionaria di variabili aleatorie i.i.d., quindi, per il teorema centrale limite, è asintoticamente gaussiano, di media θ e varianza $\frac{\theta^2}{nr}$, calcolate al punto precedente.

Exercise 1.

1. The maximum likelihood function, which is only a function of the parameter θ since r is known, is

$$L(\theta; x_1, \dots, x_n)) = \prod_{i=1}^{n} \left(\frac{x_i^{r-1}}{\theta^r \Gamma(r)} \, \mathrm{e}^{-x_i/\theta} \right) = \left(\prod_{i=1}^{n} x_i^{r-1} \right) \mathrm{e}^{-\frac{1}{\theta} \sum_{i=1}^{n} x_i} \frac{1}{\theta^{nr} \Gamma(r)^n}.$$

   Considering its logarithm (log-likelihood) we obtain

$$l(\theta) = \ln \left(\prod_{i=1}^{n} x_i^{r-1} \right) - \frac{1}{\theta} \sum_{i=1}^{n} x_i - nr \ln(\theta) - n \ln \big(\Gamma(r) \big).$$

   To find the value of θ which maximizes $l(\theta)$, we compute the first derivative and we study its null points and its sign:

$$l'(\theta) = \frac{1}{\theta^2} \sum_{i=1}^{n} x_i - \frac{nr}{\theta} \geq 0,$$

   hence $\theta \leq \hat{\theta} := \frac{1}{nr} \sum_{i=1}^{n} x_i$. Then $\hat{\theta}$ is the maximum point, and the maximum likelihood estimator for θ is $T_n = \frac{1}{nr} \sum_{i=1}^{n} X_i = \overline{X}_n / r$.

2. Using the values of the mean and variance given in the tables, we have:

$$\mathbb{E}(T_n) = \mathbb{E}\left(\frac{\overline{X}_n}{r}\right) = \frac{\mathbb{E}(X_1)}{r} = \frac{r\theta}{r} = \theta,$$

$$\mathrm{Var}(T_n) = \mathrm{Var}\left(\frac{\overline{X}_n}{r}\right) = \frac{\mathrm{Var}\left(\overline{X}_n\right)}{r^2} = \frac{\mathrm{Var}(X_1)}{nr^2} = \frac{r\theta^2}{nr^2} = \frac{\theta^2}{nr} \to 0.$$

Hence the estimator T_n is unbiased and square mean consistent for θ.

3. $T_n = \frac{1}{n}\sum_{i=1}^{n}\frac{X_i}{r}$ is a sample mean of i.i.d. random variables, then, from the central limit theorem, it is asymptotically normal, with mean θ and variance $\frac{\theta^2}{nr}$ (from the calculation of point 2.).

Esercizio 2.

1. Utilizziamo il test F del rapporto tra le varianze campionarie con media incognita, la cui statistica test sotto H_0 ha distribuzione F di Fisher con 20 gradi di libertà al numeratore e al denominatore. Il valore osservato della statistica è

$$F = \frac{s_X^2}{s_Y^2} = \frac{\sum x_i^2 - 21\bar{x}^2}{\sum y_i^2 - 21\bar{y}^2} = \frac{28890.66}{17566.15} = 1.645.$$

L'ipotesi nulla viene rifiutata per valori alti della statistica e, con il livello dato, se essa supera $q_{20,20}(0.90) = 1.79$, vale a dire il 90^o percentile della F. Il valore osservato della statistica test non consente di rifiutare l'ipotesi nulla che le varianze siano uguali.

2. La funzione di potenza si ottiene calcolando la probabilità che la statistica test cada nella regione di rifiuto quando $\theta = \sigma_X^2/\sigma_Y^2 = 1.25$:

$$P_{\theta=1.25}(F > 1.79) = P_{\theta=1.25}\left(\frac{F}{1.25} > \frac{1.79}{1.25}\right) = P(F_{20,20} > 1.43)$$

dove con $F_{20,20}$ abbiamo indicato una F di Fisher con 20 gradi di libertà al numeratore e al denominatore. Nelle tavole troviamo che $P(F_{20,20} \leq 1.36) = 0.75$ e che $P(F_{20,20} \leq 1.47) = 0.80$, e quindi la funzione di potenza nel punto dato è compresa tra 0.20 e 0.25.

3. Osserviamo che $\Delta = \mathrm{E}(X - Y)$. Pertanto il suo stimatore con il metodo dei momenti si ottiene eguagliando il valore atteso al suo corrispettivo campionario:

$$\Delta = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) = \bar{X} - \bar{Y} = T.$$

dove n indica la dimensione del campione. Lo stimatore è non distorto, infatti:

$$\mathrm{E}(T) = \mathrm{E}(\bar{X}) - \mathrm{E}(\bar{Y}) = \mu_X - \mu_Y$$

e, per le assunzioni di indipendenza,

$$\mathrm{Var}(T) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n}$$

che converge a zero al crescere di n, dunque T è consistente. In corrispondenza del campione osservato, la stima di Δ è pari a $\bar{x} - \bar{y} = 7.319$.

Per quanto riguarda la varianza, secondo il test al punto 1, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, dunque $\mathrm{Var}(T) = 2\sigma^2/n$, che possiamo stimare usando lo stimatore della varianza *pooled*, $2s_p^2/n$, dove

$$s_p^2 = \frac{(\sum x_i^2 - n\bar{x}^2) + (\sum y_i^2 - n\bar{y}^2)}{2n - 2} = \frac{28890.66 + 17566.15}{40} = 1161.420$$

ottenendo infine il valore $2s_p^2/n = 2 \times 1161.420/21 = 110.6114$.

Exercise 2.

1. Let us consider the F test of the ratio between sample variances (when the mean is unknown), where the test statistic, under the null hypothesis H_0, has Fisher distribution with 20 degrees of freedom both at the numerator and at the denominator. The observed value of the test statistic is

$$F = \frac{s_X^2}{s_Y^2} = \frac{\sum x_i^2 - 21\bar{x}^2}{\sum y_i^2 - 21\bar{y}^2} = \frac{28890.66}{17566.15} = 1.645.$$

The null hypothesis is rejected for high values of the test statistic and, at the given level, if it is greater of $q_{20,20}(0.90) = 1.79$ (that is the 90^o percentile of the Fisher). Hence we can't reject the null hypothesis that the variances are equal.

2. The power is $\theta = \sigma_X^2/\sigma_Y^2 = 1.25$:

$$P_{\theta=1.25}(F > 1.79) = P_{\theta=1.25}\left(\frac{F}{1.25} > \frac{1.79}{1.25}\right) = P(F_{20,20} > 1.43)$$

where we have indicated by $F_{20,20}$ a Fisher distribution with 20 degrees of freedom at the numerator and at the denominator. Using the tables, we find that $P(F_{20,20} \leq 1.36) = 0.75$ and that $P(F_{20,20} \leq 1.47) = 0.80$, and then the power in the given point is between 0.20 and 0.25.

3. Notice that $\Delta = \mathrm{E}(X - Y)$. Then the method of moments estimator can be found:

$$\Delta = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) = \bar{X} - \bar{Y} = T.$$

where n is the sample dimension. The estimator is unbiased, in fact:

$$\mathrm{E}(T) = \mathrm{E}(\bar{X}) - \mathrm{E}(\bar{Y}) = \mu_X - \mu_Y$$

and, from the independence assumptions,

$$\mathrm{Var}(T) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{n}$$

which converges to zero for n increasing, then T is consistent. For the given sample, the estimate of Δ is equal to $\bar{x} - \bar{y} = 7.319$.

For what concerns the variance, from the test in 1. we have $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, and then $\mathrm{Var}(T) = 2\sigma^2/n$, which we can estimated with the *pooled* estimator, $2s_p^2/n$, where

$$s_p^2 = \frac{(\sum x_i^2 - n\bar{x}^2) + (\sum y_i^2 - n\bar{y}^2)}{2n - 2} = \frac{28890.66 + 17566.15}{40} = 1161.420,$$

obtaining finally the value $2s_p^2/n = 2 \times 1161.420/21 = 110.6114$.

Esercizio 3.

1. Le frequenze attese sono: $E_1 = 0.82 \times 440 = 360.8$, $E_2 = 0.15 \times 440 = 66$, $E_3 = 0.03 \times 440 = 13.2$. Le frequenze osservate sono: $N_1 = 366$, $N_2 = 68$, $N_3 = 6$.

   L'ipotesi nulla è che gli ingegneri abbiano lo stesso profilo di rischio di quello generale ($H_0 : p_1 = 0.82,\ p_2 = 0.15,\ p_3 = 0.03$) e l'alternativa che sia diverso (H_1: almeno uno dei tre p_j è diverso dai valori assegnati). Abbiamo una distribuzione discreta e un campione grande, per cui possiamo usare un test chi quadro di adattamento. La statistica test è

   $$Q = \sum_{i=1}^{3}\frac{(N_i - E_i)^2}{E_i} = \sum_{i=1}^{3}\frac{N_i^2}{E_i} - n = \frac{366^2}{360.8} + \frac{68^2}{66} + \frac{6^2}{13.2} - 440 \simeq 4.063.$$

   Sotto l'ipotesi nulla, la statistica test ha approssimativamente distribuzione chi quadrato con 2 gradi di libertà. La regione critica è $Q > \chi^2_{.95}(2) \simeq 5.991$. Dunque non rigetto l'ipotesi nulla che gli ingegneri abbiano lo stesso profilo di rischio di quello generale a livello 5%.

2. Il p-value è il più piccolo livello per cui per cui si rifiuta l'ipotesi nulla con i dati empirici ottenuti, cioè p-value $= \alpha^*$ con $\chi^2_{1-\alpha^*}(2) = 4.063$. Dalle tabelle si trova $\chi^2_{.80}(2) = 3.219$, $\chi^2_{.875}(2) = 4.159$, per cui p-value $\in (0.125; 0.200) = (12.5\%, 20\%)$. Un tale p-value è piuttosto alto e dunque non si rifiuta H_0 a nessuno dei livelli usuali.

Exercise 3.

1. The expected frequences are: $E_1 = 0.82 \times 440 = 360.8$, $E_2 = 0.15 \times 440 = 66$, $E_3 = 0.03 \times 440 = 13.2$. The observed frequencies are: $N_1 = 366$, $N_2 = 68$, $N_3 = 6$.
   Performing a goodness of fit chi-square test, the test statistic has approximately a chi-square distribution with 2 degrees of freedom under the null hypothesis. The value of the test statistic results to be:

$$Q = \sum_{i=1}^{3} \frac{(N_i - E_i)^2}{E_i} = \sum_{i=1}^{3} \frac{N_i^2}{E_i} - n = \frac{366^2}{360.8} + \frac{68^2}{66} + \frac{6^2}{13.2} - 440 \simeq 4.063.$$

   Since the the critic region at level 5% is $Q > \chi^2_{.95}(2) \simeq 5.991$, I can't reject the null hypothesis that the engineers have the same risk profile of the general one, at level 5%.

2. The p-value is the smallest level among the level such that I reject the null hypothesis on the base of the empirical data, that is p-value $= \alpha^*$ with $\chi^2_{1-\alpha^*}(2) = 4.063$. From the statistical tables we can find $\chi^2_{.80}(2) = 3.219$, $\chi^2_{.875}(2) = 4.159$, hence p-value $\in (0.125; 0.200) = (12.5\%, 20\%)$. This value gives an evidence in favor of the null hypothesis since I can't reject H_0 at the usual significance levels.

Esercizio 4.

1. Si tratta di un intervallo di confidenza asintotico al 90% per una proporzione che ha limiti di confidenza $T_\pm = \hat{p} \pm z_{.95}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, dove $\hat{p}$ è la proporzione empirica di successi. Per noi $n = 100$, $\hat{p} = 0.18$, $z_{.95} = 1.64485$. Si ha dunque $z_{.95}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \simeq 0.063$ e CI $= [0.117, 0.243]$.

2. Il test asintotico nel caso considerato ha regione critica $\left\{ \frac{\hat{p}-0.10}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \geq z_{.90} \right\}$. Abbiamo $\frac{\hat{p}-0.10}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \simeq \frac{0.08}{0.0384} \simeq 2.082$, $z_{.90} \simeq 1.282$. Dunque rifiutiamo l'ipotesi nulla al 10%.

Exercise 4.

1. It is an asympotic 90% confidence interval for the proportion p, which has limits $T_\pm = \hat{p} \pm z_{.95}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where $\hat{p}$ è is the empirical proportion of successes. Here $n = 100$, $\hat{p} = 0.18$, $z_{.95} = 1.64485$. It follows that $z_{.95}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \simeq 0.063$ and CI $= [0.117, 0.243]$.

2. In this case the asymptotic test has critic region $\left\{ \frac{\hat{p}-0.10}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \geq z_{.90} \right\}$. We have $\frac{\hat{p}-0.10}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \simeq \frac{0.08}{0.0384} \simeq 2.082$, $z_{.90} \simeq 1.282$. Hence we reject the null hypothesis at level 10%.

Write down clearly how the results are derived (not just the results!)
Nello svolgere gli esercizi fornire passaggi e spiegazioni: non bastano i risultati finali

Esercizio 1. Una grossa industria vuole confrontare le caratteristiche di due resine plastiche per costruire tubature. La resina 1 è meno costosa della resina 2, ma si sospetta che la resina 2 abbia un maggior allungamento medio percentuale (capacità del materiale di allungarsi senza superare la rottura). Un campione sperimentale di dimensione 120 della resina 1 ha fornito un'allungamento medio campionario pari a $\overline{x}_1 = 2.1$ e una deviazione standard campionaria pari a $s_1 = 0.51$, mentre un campione di dimensione 130 della resina 2 ha fornito un'allungamento medio campionario pari a $\overline{x}_2 = 2.5$ e una deviazione standard campionaria pari a $s_2 = 0.58$. Si assuma che le varianze effettive delle due resine siano uguali e che assumere la normalità dei dati sia una buona approssimazione.

1. Si fornisca una stima della varianza σ^2 delle due resine.
2. I dati a disposizione permettono di concludere a livello 1% che l'allungamento medio della resina 2 è maggiore di quello della resina 1? Si costruisca un opportuno test statistico.
3. Si costruisca un intervallo di confidenza di livello 98% per la differenza tra gli allungamenti medi delle due resine considerate.

Exercise 1. A big industry wants to compare the characteristics of two plastic resins to build pipelines. Resin 1 is less expensive than resin 2, but resin 2 is expected to have a greater mean specific expansion (ability to stretch the material without breaking). A random sample of dimension 120 of resin 1 has given an empirical mean specific expansion $\overline{x}_1 = 2.1$ and an empirical standard deviation $s_1 = 0.51$, while a random sample of dimension 130 of resin 2 has given an empirical mean specific expansion $\overline{x}_2 = 2.5$ and a standard deviation $s_2 = 0.58$. Let assume that the the two variances are equal and that, with a good approximation, the data can be considered to be normal.

1. Give an estimate of the common variance σ^2 of the two resins.
2. Do the data at disposal allow to conclude at significance level 1% that the mean specific expansion of the resin 2 is greater than the one of resine 1? You have to construct a suitable statistical test.
3. Construct a confidence interval of level 98% for the difference between the mean specific expansions of the two resins under consideration.

Esercizio 2. Gli orologi prodotti da un noto fabbricante presentano uno scostamento X (espresso in secondi alla settimana) dall'ora esatta che ha una distribuzione normale di media μ e varianza σ^2. Il fabbricante di orologi afferma che

a) non esiste alcuna evidenza del fatto che egli produca più orologi che rimangono indietro piuttosto che altri che vanno avanti (e viceversa) e che dunque si ha $\mu = 0$,

b) almeno il 95% dei suoi orologi non si discostano dall'ora esatta per più di 0.2 secondi alla settimana e che dunque si ha $\sigma^2 \leq \sigma_0^2 = 0.01041\ (\text{sec/sett})^2$.

Si consideri un campione di 15 orologi i cui scostamenti dall'ora esatta, in secondi per settimana, sono i seguenti (il segno negativo denota orologi che rimangono indietro):

$$0.09,\ -0.15,\ 0.05,\ -0.13,\ 0.15,\ -0.07,\ -0.02,\ 0.00,\ -0.22,\ -0.18,\ 0.00,\ 0.03,\ -0.03,\ -0.33,\ -0.03$$

1. Fornire stime non distorte di μ e σ^2, considerando μ e σ^2 incogniti.

Alcuni produttori concorrenti mettono in dubbio le affermazioni del nostro fabbricante sia sulla media che sulla varianza e progettano un test per verificare l'affermazione **b**.

2. Basandosi su di un campione casuale di n osservazioni della popolazione X, proporre un test di ampiezza α per verificare l'affermazione sulla varianza, considerando incognita la media.

3. Con i 15 dati osservati e con $\alpha = 5\%$, cosa si conclude riguardo alla varianza?

Altri produttori invece accettano l'affermazione del nostro fabbricante sulla media, ma non quella sulla varianza.

4. In questo caso, basandosi su di un campione casuale di n osservazioni della popolazione X, proporre un test di ampiezza α per verificare l'affermazione sulla varianza, supponendo la media uguale a quella dichiarata dal nostro fabbricante.

5. Con i 15 dati osservati e con $\alpha = 5\%$, cosa si conclude riguardo alla varianza in questo caso?

Exercise 2. The watches produced by a well-known manufacturer show a deviation X (expressed in seconds per week) from the exact time which has a normal distribution with mean μ and variance σ^2. The manufacturer says that

a) there is no evidence that he produces more watches which are slow rather than watches which are fast (and viceversa), so that one has $\mu = 0$,

b) at least the 95% of his watches do not deviate from the exact time for more than 0.2 seconds per week and, so, one has $\sigma^2 \leq \sigma_0^2 = 0.01041\ (\text{sec/week})^2$.

Let us consider a random sample of 15 watches having the following deviations from the exact time, in seconds per week (the minus sign is for watches that are slow):

$$0.09,\ -0.15,\ 0.05,\ -0.13,\ 0.15,\ -0.07,\ -0.02,\ 0.00,\ -0.22,\ -0.18,\ 0.00,\ 0.03,\ -0.03,\ -0.33,\ -0.03$$

1. Give unbiased estimates of μ and σ^2, considering μ and σ^2 unknown.

Some competitor manufacturers call into question the statements of our manufacturer, either on the mean, either on the variance; then they design an experiment to test the statement **b**.

2. Having a random sample of n observations of the population X, propose a test of size α to verify the statement on the variance, considering the mean unknown.

3. With the 15 observations given above and with $\alpha = 5\%$, what can you conclude about the variance?

Other producers instead accept the statement of our manufacturer on the mean, but not the one on the variance.

4. In this case, having a random sample of n observations of the population X, propose a test of size α to verify the statement on the variance, assuming the mean equal to the one stated by our manufacturer.

5. With the 15 observations given above and with $\alpha = 5\%$, what can you conclude about the variance in this case?

Esercizio 3. Sia $X_1, \ldots, X_n$ un campione estratto da una popolazione con densità di probabilità

$$f(x) = \begin{cases} \frac{1}{\theta}\, x^{\frac{1}{\theta}-1} & \text{se} \quad x \in [0,1] \\ 0 & \text{se} \quad x \notin [0,1] \end{cases} \qquad \theta > 0,$$

dove θ è un parametro incognito positivo. Verificare che

1. $\mathbb{E}[X_i] = \frac{1}{1+\theta}$ per ogni i,

2. Trovate lo stimatore $\hat{\theta}$ del metodo dei momenti di θ.

3. Trovate lo stimatore T di massima verosimiglianza di θ.

Exercise 3. Let $X_1, \ldots, X_n$ be a random sample extracted from a population with probability density

$$f(x) = \begin{cases} \frac{1}{\theta}\, x^{\frac{1}{\theta}-1} & \text{if} \quad x \in [0,1] \\ 0 & \text{if} \quad x \notin [0,1] \end{cases} \qquad \theta > 0,$$

where θ is an unknown positive parameter. Verify that

1. $\mathbb{E}[X_i] = \frac{1}{1+\theta}$ for all i,
2. Find the method of moments estimator $\hat{\theta}$ for θ.
3. Find the maximum likelihood estimator T for θ.

Esercizio 4.

1. In laboratorio cinque lavatrici vengono sottoposte a una prova di durata, ricavando i seguenti tempi al guasto, in migliaia di cicli di lavaggio:
$$1.662,\ 0.410,\ 0.682,\ 1.581,\ 2.813\,.$$
Con un opportuno test non parametrico di livello 5%, stabilire se la variabile aleatoria X del tempo al guasto possa provenire da una distribuzione con funzione di ripartizione
$$F_0(x) = 1_{(0,+\infty)}(x)\left(1 - \mathrm{e}^{-x^2}\right).$$
2. Vengono quindi installate altre cinque lavatrici presso altrettante abitazioni e vengono registrati i tempi al guasto anche in questo caso. Tuttavia, poiché l'uso quotidiano è ritenuto meno severo delle prove di laboratorio, si pensa che la variabile aleatoria del tempo al guasto per tale uso, indicata con Y, sia legata alla precedente variabile X dalla relazione $Y = \theta X$, con θ tipicamente maggiore di 1. I tempi al guasto osservati, in migliaia di cicli di lavaggio, sono ora:
$$(y_1, \ldots, y_5) = (1.420,\ 0.470,\ 0.343,\ 0.595,\ 2.053).$$
Il test al punto precedente ha confermato l'ipotesi formulata, perciò è possibile mostrare che $\mathbb{E}(X) \simeq 0.886227$. Sapendo ciò, ricavare una stima di θ con il metodo dei momenti e utilizzarla per calcolare la probabilità che una lavatrice installata in casa duri più di mille lavaggi.

Exercise 4.

1. In a laboratory, five washing machines are submitted to a duration test, obtaining the following times to the breakdown, in thousands of washing cycles:
$$1.662,\ 0.410,\ 0.682,\ 1.581,\ 2.813\,.$$
With a suitable non parametric test of significance level 5%, verify if the random variable X representing the time to the breakdown can come from the cumulative distribution function
$$F_0(x) = 1_{(0,+\infty)}(x)\left(1 - \mathrm{e}^{-x^2}\right).$$
2. Another set of 5 washing machines is installed in 5 habitations and the times to the breakdown are registered also in this case. However, as the daily usage is thought to be less severe than a test in a lab, the random variable giving the breaking time for such an usage, denoted by Y, is thought to be linked to the previous variable X by the relation $Y = \theta X$, with θ typically greater than 1. The observed breaking time in this case, in thousands of washing cycles, are:
$$(y_1, \ldots, y_5) = (1.420,\ 0.470,\ 0.343,\ 0.595,\ 2.053).$$
The test in the previous point has confirmed the null hypothesis, therefore it is possible to compute the mean value under this hypothesis and to get $\mathbb{E}(X) \simeq 0.886227$. Knowing this result, you must obtain an estimate of θ with the method of moments and you must use it to compute the probability that a washing machine installed in an habitation has a duration of more than 1000 washing cycles.

SOLUZIONI STATISTICA 20/09/2010 STATISTICS SOLUTIONS

Mettere più spiegazioni nelle soluzioni

Esercizio 1. Abbiamo due campioni normali indipendenti di varianza σ^2 incognita ed uguale e medie incognite μ_1 e μ_2.

1. Uno stimatore naturale e non distorto della varianza comune è la "pooled variance"
$$S_P^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{119 \times 0.51^2 + 129 \times 0.58^2}{119+129} = \frac{74.3475}{248} \simeq 0.3000.$$
2. $H_0 : \mu_1 = \mu_2$ contro $H_1 : \mu_2 > \mu_1$. Regione critica: $\{T > t_{1-\alpha}(n_1+n_2-2) \simeq z_{.99} \simeq 2.33\}$, dove la statistica test è $T = \dfrac{\overline{X}_2 - \overline{X}_1}{S_P\sqrt{1/n_1+1/n_2}}$. Abbiamo $\sqrt{s_P^2\,(1/n_1+1/n_2)} \simeq \sqrt{0.3 \times \left(\frac{1}{120}+\frac{1}{130}\right)} \simeq 0.06934$, $t \simeq \frac{0.4}{0.06934} \simeq 5.7689 > 2.33$; in conclusione rigetto H_0.
3. Abbiamo $z_{.99}\, s_P\sqrt{1/n_1+1/n_2} \simeq 2.33 \times 0.06934 \simeq 0.1616$; dunque
$$\text{IC}_{98\%}(\mu_2-\mu_1) = \left(\overline{x}_2 - \overline{x}_1 \pm z_{.99}\, s_P\sqrt{1/n_1+1/n_2}\right) \simeq (0.4 \pm 0.16) = (0.24; 0.56).$$

Exercise 1. We have two independent random samples with variance σ^2, which is equal in the two populations and unknown, and unknown means μ_1 e μ_2.

1. A natural unbiased estimator of the common variance is the "pooled variance"
$$S_P^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{119 \times 0.51^2 + 129 \times 0.58^2}{119+129} = \frac{74.3475}{248} \simeq 0.3000.$$
2. $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_2 > \mu_1$. Critical region: $\{T > t_{1-\alpha}(n_1+n_2-2) \simeq z_{.99} \simeq 2.33\}$, where the test statistics is $T = \dfrac{\overline{X}_2 - \overline{X}_1}{S_P\sqrt{1/n_1+1/n_2}}$. We have $\sqrt{s_P^2\,(1/n_1+1/n_2)} \simeq \sqrt{0.3 \times \left(\frac{1}{120}+\frac{1}{130}\right)} \simeq 0.06934$, $t \simeq \frac{0.4}{0.06934} \simeq 5.7689 > 2.33$; in conclusion we reject H_0.
3. We have $z_{.99}\, s_P\sqrt{1/n_1+1/n_2} \simeq 2.33 \times 0.06934 \simeq 0.1616$; therefore,
$$\text{CI}_{98\%}(\mu_2-\mu_1) = \left(\overline{x}_2 - \overline{x}_1 \pm z_{.99}\, s_P\sqrt{1/n_1+1/n_2}\right) \simeq (0.4 \pm 0.16) = (0.24; 0.56).$$

Esercizio 2.

1. Gli stimatori cercati sono media e varianza campionarie $\overline{X}$ e S^2; abbiamo $\overline{x} = \frac{1}{15}\sum_{i=1}^{15} x_i = -\frac{0.84}{15} = -0.056$, $s^2 = \frac{1}{14}\sum_{i=1}^{15}(x_i-\overline{x})^2 \simeq 0.01594$.
2. $H_0 : \sigma^2 \le \sigma_0^2$ contro $H_1 : \sigma^2 > \sigma_0^2$. Regione critica: $\left\{\frac{(n-1)S_n^2}{\sigma_0^2} > \chi^2_{1-\alpha}(n-1)\right\}$.
3. $\frac{14\,s^2}{\sigma_0^2} \simeq \frac{14 \times 0.01594}{0.01041} \simeq 21.4371$, $\chi^2_{.95}(14) \simeq 23.685$: non si può rifiutare H_0.
4. $H_0 : \sigma^2 \le \sigma_0^2$ contro $H_1 : \sigma^2 > \sigma_0^2$. Regione critica: $\left\{\frac{\sum_{i=1}^n X_i^2}{\sigma_0^2} > \chi^2_{1-\alpha}(n)\right\}$.
5. $\frac{\sum_{i=1}^{15} x_i^2}{\sigma_0^2} = \frac{0.2702}{0.01041} \simeq 25.956$, $\chi^2_{.95}(15) \simeq 24.995$: si rifiuta H_0.

Exercise 2.

1. The estimators we are searching for are the sample mean $\overline{X}$ and the sample variance S^2; we have $\overline{x} = \frac{1}{15}\sum_{i=1}^{15} x_i = -\frac{0.84}{15} = -0.056$, $s^2 = \frac{1}{14}\sum_{i=1}^{15}(x_i-\overline{x})^2 \simeq 0.01594$.

2. $H_0 : \sigma^2 \leq \sigma_0^2$ versus $H_1 : \sigma^2 > \sigma_0^2$. Critical region: $\left\{ \frac{(n-1)S_n^2}{\sigma_0^2} > \chi^2_{1-\alpha}(n-1) \right\}$.

3. $\frac{14\, s^2}{\sigma_0^2} \simeq \frac{14 \times 0.01594}{0.01041} \simeq 21.4371$, $\chi^2_{.95}(14) \simeq 23.685$: we cannot reject H_0.

4. $H_0 : \sigma^2 \leq \sigma_0^2$ versus $H_1 : \sigma^2 > \sigma_0^2$. Critical region: $\left\{ \frac{\sum_{i=1}^n X_i^2}{\sigma_0^2} > \chi^2_{1-\alpha}(n) \right\}$.

5. $\frac{\sum_{i=1}^{15} x_i^2}{\sigma_0^2} = \frac{0.2702}{0.01041} \simeq 25.956$, $\chi^2_{.95}(15) \simeq 24.995$: we reject H_0.

Esercizio 3.

1. $\mathbb{E}[X_i] = \dfrac{1}{\theta} \displaystyle\int_0^1 x^{1/\theta} \mathrm{d}x = \frac{1/\theta}{1 + 1/\theta} = \frac{1}{1+\theta}$.

2. Lo stimatore del metodo dei momenti si ottiene risolvendo l'equazione $\overline{X} = \frac{1}{1+\hat\theta}$, che dà $\hat\theta = \frac{1}{\overline{X}} - 1$.

3. Studiamo gli zeri e il segno della funzione di verosimiglianza: $L(\theta; x_1, \ldots, x_n) = \dfrac{1}{\theta^n} \displaystyle\prod_{i=1}^n x_i^{\frac{1}{\theta}-1}$,

$$\ln L(\theta) = -n \ln \theta + \left(\frac{1}{\theta} - 1 \right) \sum_{i=1}^n \ln x_i \,, \qquad \frac{\partial}{\partial \theta} \ln L(\theta) = -\frac{n}{\theta} - \frac{1}{\theta^2} \sum_{i=1}^n \ln x_i \,;$$

si ha $\frac{\partial \ln L(\theta)}{\partial \theta} \geq 0$ se e solo se $\theta \leq -\frac{1}{n} \sum_{i=1}^n \ln x_i$. Dunque lo stimatore di massima verosimiglianza esiste ed è dato da $T = \overline{Y} := \frac{1}{n} \sum_{i=1} Y_i$, dove $Y_i := -\ln X_i$.

Exercise 3.

1. $\mathbb{E}[X_i] = \dfrac{1}{\theta} \displaystyle\int_0^1 x^{1/\theta} \mathrm{d}x = \frac{1/\theta}{1 + 1/\theta} = \frac{1}{1+\theta}$.

2. The estimator of the method of moments is obtained by solving the equation $\overline{X} = \frac{1}{1+\hat\theta}$, which gives $\hat\theta = \frac{1}{\overline{X}} - 1$.

3. Let us study the sign and the zeros of the likelihood function: $L(\theta; x_1, \ldots, x_n) = \dfrac{1}{\theta^n} \displaystyle\prod_{i=1}^n x_i^{\frac{1}{\theta}-1}$,

$$\ln L(\theta) = -n \ln \theta + \left(\frac{1}{\theta} - 1 \right) \sum_{i=1}^n \ln x_i \,, \qquad \frac{\partial}{\partial \theta} \ln L(\theta) = -\frac{n}{\theta} - \frac{1}{\theta^2} \sum_{i=1}^n \ln x_i \,;$$

one has $\frac{\partial \ln L(\theta)}{\partial \theta} \geq 0$ if and only if $\theta \leq -\frac{1}{n} \sum_{i=1}^n \ln x_i$. Therefore, the maximum likelihood estimator exists and it is given by $T = \overline{Y} := \frac{1}{n} \sum_{i=1} Y_i$, where $Y_i := -\ln X_i$.

Esercizio 4.

1. Conduciamo un test di Kolmogorov-Smirnov sul campione dato. Il campione ordinato è

$$0.410,\ 0.682,\ 1.581,\ 1.662,\ 2.813,$$

a cui corrispondono i seguenti valori della funzione di ripartizione empirica $\hat F_n(x)$

$$0.2,\ 0.4,\ 0.6,\ 0.8,\ 1.0$$

e della funzione di ripartizione ipotizzata $F_0(x)$:

$$0.155,\ 0.372,\ 0.918,\ 0.937,\ 1.000\,.$$

Dunque $D_n = \sup_x |F_0(x) - \hat F_n(x)| = 0.518$. Dalle tavole otteniamo che per $n = 5$ si ha il quantile $q_{0.95} = 0.5633$, dunque non possiamo rifiutare l'ipotesi F_0 al livello del 5%.

2. Per quanto detto nel testo, possiamo affermare che

$$\mathbb{E}(Y) = \theta\mathbb{E}(X) \simeq 0.886227\,\theta$$

Allora la stima del metodo dei momenti di θ si ricava dall'equazione

$$0.886227\,\theta = \sum_{i=1}^{5} y_i/5 = 0.9762 \qquad \Longrightarrow \qquad \hat{\theta} = \frac{0.9762}{0.886227} \simeq 1.10152\,.$$

La probabilità cercata è

$$\mathbb{P}(Y > 1) = \mathbb{P}(\theta X > 1) = \mathbb{P}\left(X > \frac{1}{\theta}\right) = \mathrm{e}^{-\left(\frac{1}{\theta}\right)^2}.$$

Inserendo la stima di θ abbiamo infine

$$\mathrm{e}^{-\left(\frac{1}{\theta}\right)^2} \simeq \mathrm{e}^{-\frac{1}{1.10152^2}} \simeq 0.4386\,.$$

Exercise 4.

1. We perform a Kolmogorov-Smirnov test on the given sample. The ordered sample is

$$0.410,\ 0.682,\ 1.581,\ 1.662,\ 2.813,$$

the corresponding values of the empirical cumulative distribution function $\hat{F}_n(x)$ are

$$0.2,\ 0.4,\ 0.6,\ 0.8,\ 1.0$$

and the values of the cumulative distribution function $F_0(x)$ are

$$0.155,\ 0.372,\ 0.918,\ 0.937,\ 1.000\,.$$

Therefore, we have $D_n = \sup_x |F_0(x) - \hat{F}_n(x)| = 0.518$. From the statistical tables we get that for $n = 5$ the quantile is $q_{0.95} = 0.5633$, and, so, we cannot reject the hypothesis that the c.d.f. is F_0 at level 5%.

2. By what is said in the text, we can say that

$$\mathbb{E}(Y) = \theta\mathbb{E}(X) \simeq 0.886227\,\theta$$

Then, the method of moments estimate of θ comes out from the equation

$$0.886227\,\theta = \sum_{i=1}^{5} y_i/5 = 0.9762 \qquad \Longrightarrow \qquad \hat{\theta} = \frac{0.9762}{0.886227} \simeq 1.10152\,.$$

The probability we are searching for is

$$\mathbb{P}(Y > 1) = \mathbb{P}(\theta X > 1) = \mathbb{P}\left(X > \frac{1}{\theta}\right) = \mathrm{e}^{-\left(\frac{1}{\theta}\right)^2}.$$

By using the estimate of θ finally we get

$$\mathrm{e}^{-\left(\frac{1}{\theta}\right)^2} \simeq \mathrm{e}^{-\frac{1}{1.10152^2}} \simeq 0.4386\,.$$

5 28.02.11 STATISTICA per ING INF TEL [2L] I. Epifani, A. Pievatolo

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

Esercizio 5.1 Un punto vendita all'ingrosso ha appena iniziato a vendere in massima parte un articolo del valore di $p = 25$ euro. Negli $n = 78$ giorni di apertura dei primi tre mesi vengono raccolti i dati di vendita, ottenendo i dati $(x_1, \ldots, x_n)$ del numero di articoli venduti ogni giorno (si calcolano $m = 26$ giorni di apertura al mese). Si ipotizza che tali dati siano la realizzazione di un campione casuale da una Poisson di media θ.

1. Per ogni articolo venduto vengono fatturati p euro. Stimate in modo efficiente il fatturato medio giornaliero e ricavate in seguito uno stimatore efficiente del fatturato medio mensile. Calcolate la stima del fatturato medio mensile quando $\sum_{i=1}^{n} x_i = 1950$
2. Calcolate la varianza dello stimatore del fatturato medio mensile trovato al punto precedente e dimostrate quindi che lo stimatore è consistente al crescere del numero di giorni. Calcolate la stima di massima verosimiglianza di tale varianza.

L'esistenza del punto vendita è giustificata (in quanto a copertura delle spese di gestione) se il fatturato mensile supera i 15000 euro.

3. Calcolate un'approssimazione della probabilità che il fatturato mensile superi tale cifra e stimatene il valore quando $\sum_{i=1}^{78} x_1 = 1950$.

SOLUZIONE

1. Dobbiamo stimare $\mathrm{E}(p\sum_{i=1}^{m} X_i) = pm\theta = \kappa(\theta)$. Otteniamo prima lo stimatore di massima verosimiglianza di θ. La funzione di logverosimiglianza di θ è

$$\log(L_\theta(x_1, \ldots, x_n)) = -n\theta + \log(\theta)\sum_{i=1}^{n} x_i$$

e la sua derivata prima rispetto a θ è

$$\frac{\partial \log(L_\theta)}{\partial \theta} = -n + \frac{\sum_{i=1}^{n} x_i}{\theta} = \frac{n}{\theta}\left(\frac{\sum_{i=1}^{n} x_i}{n} - \theta\right)$$

e dunque, per il teorema di Frechet-Kramér-Rao, $\hat{\theta} = \bar{X}$ è stimatore efficiente di θ. Per la proprietà di invarianza dello stimatore di massima verosimiglianza,

$$T = \kappa(\hat{\theta}) = pm\bar{X}$$

è lo stimatore di massima verosimiglianza del fatturato medio mensile. Esso è anche efficiente in quanto trasformata lineare di uno stimatore efficiente. La stima del fatturato medio mensile è dunque $\kappa(\hat{\theta}) = 25 \times 26 \times 1950/78 = 16250$ euro.

2. Per l'ipotesi fatta che il campione è un campione casuale,

$$\mathrm{Var}(T) = (p^2m^2)\mathrm{Var}(\bar{X}) = (p^2m^2)\frac{\theta}{n}.$$

La varianza tende a zero al crescere di n e dunque lo stimatore è consistente essendo anche non distorto. Per quanto riguarda la stima di massima verosimiglianza di $\mathrm{Var}(T)$, applichiamo ancora la proprietà di invarianza e abbiamo:

$$(p^2m^2)\frac{\hat{\theta}}{n} = 25^2 \times 26^2 \times 25/78 = 25^2 \times 26 \times 25/3 \simeq 135416.67\ .$$

3. Dobbiamo approssimare $P(p\sum_{i=1}^{26} X_i > 15000)$. L'applicazione del teorema limite centrale, consente di ricavare che

$$P(p\sum_{i=1}^{26} X_i > 15000) = P(\sum_{i=1}^{26} X_i > 600) \simeq 1 - \Phi\left(\frac{600 - 26\theta}{\sqrt{26\theta}}\right)\ .$$

Sostituendo a θ la stima $\hat{\theta} = 1950/78 = 25$, otteniamo

$$1 - \Phi\left(\frac{600 - 26\hat{\theta}}{\sqrt{26\hat{\theta}}}\right) = 1 - \Phi\left(\frac{600 - 650}{\sqrt{650}}\right) \simeq 1 - \Phi(-1.96) = \Phi(1.96) \simeq 0.975 \,. \blacksquare$$

Esercizio 5.2 Per valutare l'effetto di un particolare ormone sulla produzione di latte bovino, si misurano le quantità settimanali di latte (in litri) prodotte da 8 mucche prima e dopo il trattamento ormonale. I risultati ottenuti sono nella seguente tabella:

prima del trattamento	151.9	157.9	149.5	147.0	161.3	150.9	160.6	147.9
dopo il trattamento	154.5	163.3	151.2	145.5	160.8	154.4	163.2	148.6

Sia X la produzione di latte di una mucca prima del trattamento e Y quella dopo il trattamento. Si assume che $(X_1, Y_1), \ldots, (X_8, Y_8)$ è un campione casuale da una distribuzione gaussiana bivariata. Le seguenti statistiche saranno utili nel corso dell'esercizio: $\sum x_i = 1227$, $\sum x_i^2 = 188420.5$, $\sum y_i = 1241.5$, $\sum y_i^2 = 192981$, $\sum x_i y_i = 190670.4$.

1. Verificate con un opportuno test di livello $\alpha = 0.001$ l'ipotesi alternativa che ci sia una dipendenza lineare positiva fra le due variabili X e Y.

2. In base al campione a disposizione, si può concludere che il trattamento aumenta la produzione media di latte?

Considerate ora solo i dati relativi alla produzione di latte dopo il trattamento ormonale.

3. Rifiutate o meno al livello $\alpha = 0.05$ l'ipotesi nulla che la deviazione standard del latte prodotto dopo il trattamento σ sia non superiore a 4.5 l.

4. Calcolate la potenza del test costruito al punto 3 quando effettivamente la deviazione standard vale 6.

SOLUZIONE I dati riportati in tabella sono la realizzazione di un campione bivariato accoppiato gaussiano di dimensione $n = 8$.

1. Impostiamo un test sul coefficiente di correlazione lineare di (X, Y) ρ per verificare

$$\mathrm{H}_0 : \ \rho \leq 0 \text{ contro } \mathrm{H}_1 : \ \rho > 0.$$

Stimiamo ρ con il coefficiente di correlazione lineare

$$R = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} = \frac{\mathrm{Cov}_{X,Y}}{\sqrt{S_X^2 S_Y^2}}$$

dove S_X^2 e S_Y^2 sono le varianza campionarie del latte prodotto prima e dopo il trattamento ormonale, e $\mathrm{Cov}_{X,Y}$ è la covarianza campionaria. Si ricordi che:

$$\mathrm{Cov}_{X,Y} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{n}{n-1} \left(\frac{\sum_{i=1}^n X_i Y_i}{n} - \bar{X}\bar{Y} \right).$$

Usando i dati riportati in tabella otteniamo: $\bar{x} = 153.375$, $\bar{y} = 155.1875$ e $\sum_{i=1}^8 x_i y_i / 8 = 23833.8$ e dunque $\mathrm{Cov}_{X,Y} \simeq 36.477$; inoltre $s_X^2 \simeq 32.768$ e $s_Y^2 \simeq 45.103$ e perciò $r \simeq 0.9488$. Rifiutiamo H_0 se

$$t := \frac{r}{\sqrt{1 - r^2}} \sqrt{n-2} \geq t_{n-2}(1 - \alpha)$$

Poiché $t \simeq 7.358$ e $t_6(1 - 0.001) = 5.208$, allora rifiutiamo l'ipotesi nulla e riteniamo che ci sia una relazione lineare positiva fra la produzione del latte prima e dopo il trattamento.

2. Consideriamo la variabile aleatoria che rappresenta la differenza fra la quantità di latte prodotta prima del trattamento meno la quantità prodotta dopo: $W = X - Y$; W ha distribuzione $\mathrm{N}(\mu_W, \sigma_W^2)$. Un campione di dimensione 8 per W si ricava, per sottrazione, dai dati in tabella: $(-2.6, -5.4, -1.7, 1.5, 0.5, -3.5, -2.6, -0.7)$. Bisogna verificare le ipotesi:

$$\mathrm{H}_0 \ : \mu_W = 0 \text{ contro } \mathrm{H}_1 \ : \mu_W < 0,$$

per le quali la statistica test è

$$U = \frac{\bar{W}}{\sqrt{S_W^2/n}} \overset{H_0}{\sim} t_{n-1}$$

Da $\bar{w} = -1.8125$ e $s_W^2 = 4.932679$, otteniamo che la realizzazione campionaria della statistica test è $u = -2.308243 \simeq -2.31$. Il p-value è $F_{t_7}(-2.31) = 1 - F_{t_7}(2.31)$ e con l'uso delle tavole per esso otteniamo $0.025 < p\text{-value} < 0.05$; il valore esatto è $F_{t_7}(-2.308243) = 0.02716360 \simeq 2.7\%$. Poiché non possiamo rifiutare H_0 per ogni $\alpha \leq 2.7\%$, allora concludiamo che le prove sperimentali sono a favore dell'aumento della produzione di latte dopo il trattamento ormonale.

3. Sia σ^2 la varianza di Y e verifichiamo

$$\mathrm{H}_0 : \ \sigma^2 \leq 4.5^2 \text{ contro } \mathrm{H}_1 : \ \sigma^2 > 4.5^2$$

Con $\sigma_0^2 = 4.5^2$, la statistica test per il problema in esame è:

$$(n-1)\frac{S_Y^2}{\sigma_0^2} \overset{H_0}{\sim} \chi_{n-1}^2,$$

e la regione critica di ampiezza α è costituita dalle osservazioni campionarie in:

$$G_2 = \left\{ (n-1)\frac{S_Y^2}{\sigma_0^2} \geq \chi_{n-1}^2(1-\alpha) \right\}.$$

Dai dati otteniamo che $(n-1)S_Y^2/\sigma_0^2 \simeq 15.59$ e dalle tavole $\chi_7^2(0.95) = 14.067$. Quindi rifiutiamo l'ipotesi nulla. (Per il p-value abbiamo p-value$= 1 - F_{\chi_7^2}(15.59) \simeq 0.029$.)

4. La potenza del test è

$$\begin{aligned}
\pi(\sigma^2) &= P_{\sigma^2}(\text{“Rifiutare ipotesi nulla”}), \qquad\qquad \sigma^2 > 4.5^2 \\
&= P_{\sigma^2}\left((n-1)\frac{S_Y^2}{\sigma_0^2} \geq \chi_{n-1}^2(1-\alpha) \right) \\
&= P_{\sigma^2}\left((n-1)\frac{S_Y^2}{\sigma^2} \geq \frac{\sigma_0^2}{\sigma^2}\chi_{n-1}^2(1-\alpha) \right) \\
&= P_{\sigma^2}\left(Q \geq \frac{\sigma_0^2}{\sigma^2}\chi_{n-1}^2(1-\alpha) \right),
\end{aligned}$$

con $Q \sim \chi_{n-1}^2$. Utilizzando i valori assegnati ($\sigma_0 = 4.5$ e $\sigma = 6$) abbiamo:

$$\pi(\sigma^2) \simeq P\left(Q \geq 7.913\right) = 1 - F_{\chi_7^2}(7.913) \in [1 - F_{\chi_7^2}(7.992), 1 - F_{\chi_7^2}(7.283)] = [0.333, 0.400] \quad \blacksquare$$

Esercizio 5.3 Un gruppo costituito da 5 programmatori deve consegnare un software. Per ciascun programmatore viene registrato il numero di righe di programma scritte giornalmente ottenendo il seguente campione di osservazioni:

$$x_1 = 50.97,\ x_2 = 71.58,\ x_3 = 340.29,\ x_4 = 112.06,\ x_5 = 76.44\ ;$$

l'i-esimo dato x_i riportato è la media aritmetica fatta in una prova di due giorni.

1. Stabilite con un opportuno test di livello 5% se la variabile aleatoria X del numero giornaliero di righe scritte abbia distribuzione gaussiana (normale).

Viene poi contattato un secondo gruppo di 6 programmatori e anche di ciascuno di questi viene registrato il numero giornaliero di righe di programma scritte ottenendo

$$y_1 = 65.30,\ y_2 = 187.48,\ y_3 = 111.84,\ y_4 = 2.60,\ y_5 = 23.67,\ y_6 = 42.60\ .$$

2. Sulla base dei dati forniti, il numero giornaliero di righe scritte da un programmatore del secondo gruppo segue lo stesso modello del primo gruppo? Per rispondere costruite un opportuno test di significatività 5%. Nella scelta del test tenete conto di quanto risposto al punto 1.

SOLUZIONE

1. Usiamo un test di Lilliefors per la normalità dei dati $X_1, \ldots, X_5$. Infatti i dati sono continui e in numero esiguo e i parametri della distribuzione $\mathcal{N}(\mu, \sigma^2)$ non sono assegnati.

La media campionaria delle x_i vale $\overline{x} \simeq 130.268$ e la deviazione standard campionaria $\sqrt{s_X^2} \simeq 119.4474$ da cui otteniamo per $z_i := (x_i - \overline{x})/\sqrt{s_X^2}$ i seguenti valori (ordinati e distinti) e la corrispondente funzione di ripartizione empirica (indicata con $\widehat{F}_5$):

z_i	-0.66	-0.49	-0.45	-0.15	1.76
$\widehat{F}_5(z_i)$	0.2	0.4	0.6	0.8	1.0
$\Phi(z_i)$	0.2546	0.3121	0.3264	0.4404	0.9608
$\lvert\widehat{F}_5(z_i) - \Phi(z_i)\rvert$	0.0546	0.0879	0.2736	0.3596	0.0392
$\lvert\widehat{F}_5(z_{i-1}) - \Phi(z_i)\rvert$	0.2546	0.1121	0.0736	0.1596	0.1608

Deduciamo dalla precedente tabella che la statistica test $D_5 = \sup_{z\in\mathbb{R}} |\widehat{F}_5(z) - \Phi(z)|$ ha valore approssimativamente pari a 0.3596. Dalle tavole di Lilliefors, sotto l'ipotesi H_0 che i dati siano gaussiani, abbiamo che la statistica test D_5 è compresa fra i quantili di ordine $95\% = 0.3427$ e $99\% = 0.3959$, allora a livello maggiore o uguale a 5% rifiutiamo l'ipotesi di normalità dei dati mentre la accettiamo a livello minore o uguale a 1%.

(Usando il pacchetto R, "con meno approssimazioni nei conti" otteniamo $D_5 = 0.3606$ con p-value$= 0.03236 \simeq 3.2\%$

2. A livello 5% abbiamo rifiutato l'ipotesi di normalità dei dati. Quindi è coerente usare un test non parametrico di omogeneità per campioni bivariati indipendenti, qualunque e senza ripetizione. Noi conosciamo il test di Wilcoxon Mann Wintney. Per verificare H_0: "X, Y *sono regolati dallo stesso modello probabilistico*" contro l'alternativa H_1 che non lo siano, la statistica test è la somma dei ranghi $R_1, \ldots R_5$ di X:

$$R_1 = 4, R_2 = 6, R_3 = 7, R_4 = 9, R_5 = 11, \quad T_X = \sum R_i = 37$$

e i quantili di riferimento sono: $w_{5,6}(2.5\%) = 19$ e $w_{5,6}(97.5\%) = 5(5+6+1) - 19 = 60 - 19 = 41$; confrontando il valore 37 con i quantili, deriviamo che a livello 5% non possiamo rifiutare l'ipotesi nulla che il numero giornaliero di righe scritte da un programmatore del secondo gruppo segua lo stesso modello del primo gruppo. ∎