Politecnico di Milano
Temi d'esame di STATISTICA dell'AA 2004/2005
per allievi ING INF [2L]. Proff. A. Barchielli, I. Epifani

1 STATISTICA per ING INF [2L] Proff. A. Barchielli, I. Epifani 30.06.05

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

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

Esercizio 1.1 Una grandezza incognita μ è stata misurata più volte usando due strumenti con precisioni diverse e note. Possiamo schematizzare la situazione dicendo che abbiamo due campioni casuali indipendenti $X_1, X_2, \dots, X_m$ da una popolazione gaussiana $\mathcal{N}(\mu; \sigma_X^2)$ e $Y_1, Y_2, \dots, Y_n$ da una popolazione gaussiana $\mathcal{N}(\mu; \sigma_Y^2)$; le deviazioni standard $\sigma_X > 0$ e $\sigma_Y > 0$ sono note e μ è da stimare. Usiamo come stimatore di μ una statistica della forma $T = a\overline{X}_m + (1-a)\overline{Y}_n$, dove a è un opportuno numero reale.

1. Si calcolino la media e la varianza di T.
2. Si trovi a in modo da minimizzare l'errore quadratico medio di T.
3. Fissato per a il valore trovato al punto precedente, quale è la distribuzione di T?

Siano ora $m = 10$, $n = 20$, $\sigma_X^2 = 0.1$, $\sigma_Y^2 = 0.2$ e sia a come trovato al punto 2; se non siete stati in grado di trovarlo, scegliete voi un valore ragionevole per a (diverso da 0 e da 1) in modo da poter proseguire.

L'esperimento ha fornito i seguenti valori per le due medie campionarie: $\overline{x}_m = 3.10$, $\overline{y}_n = 3.00$.

4. Si trovi un intervallo di confidenza (basato su T) al 95% per μ.

Esercizio 1.2 La prova d'esame del corso di CPS è composto da un test di dieci domande a risposta multipla e una parte a esercizi e si è ammessi alla parte ad esercizi solo se si risponde correttamente ad almeno cinque domande. L'80% degli allievi che sostengono la prova supera il test, ma il 40% (di questi ultimi) NON raggiunge la sufficienza negli esercizi.

Il docente di CPS ha il dubbio che il test non filtri adeguatamente la preparazione degli allievi. Per questo motivo, in un appello "affollato" da 120 allievi, a sorpresa il docente salta il test e ammette d'ufficio tutti alla parte ad esercizi. In quell'appello, risultano promossi 78 allievi su 120.

1. Determinate una stima puntuale della percentuale $\theta 100\%$ degli allievi che raggiungono la sufficienza negli esercizi, in una prova che non prevede test.
2. Verificate l'ipotesi nulla $H_0 : \theta = 64\%$ contro l'alternativa $H_1 : \theta > 64\%$, a livello approssimativamente $\alpha = 5\%$, proponendo un opportuno test.
3. Determinate una stima puntuale della caratteristica $\kappa = \kappa(\theta)$ data dalla percentuale $\kappa 100\%$ degli allievi bocciati al test che (se ammessi agli esercizi) raggiungerebbero la sufficienza.
4. Verificate l'ipotesi nulla $H_0 : \kappa = 20\%$ contro l'alternativa $H_1 : \kappa > 20\%$, a un livello approssimato $\alpha = 5\%$.
5. Calcolate la funzione di potenza del test di ipotesi su θ del punto 2. in $\theta = 66\%$ e la funzione di potenza del test di ipotesi su κ del punto 4 in $\kappa = 22.5\%$.

Esercizio 1.3 Si sono registrati i tempi di vita in ore di 30 pile AAA prodotte dall'azienda `bbb`. Questi dati costituiscono un campione casuale $X_1, \dots, X_{30}$ la cui funzione di ripartizione empirica $\hat{F}_{30}$ è data da

x	1	2	4	6	8	11	13	27	29	42
$\hat{F}_{30}(x)$	7/30	12/30	16 /30	20/30	23/30	26/30	27/30	28/30	29/30	1

Più precisamente, X_j rappresenta la durata della j-esima pila.

1. Determinate uno stimatore T_1 della probabilità che una pila AAA prodotta dall'azienda `bbb` duri più di 9 ore.
2. Costruite un intervallo di confidenza asintotico di livello 95% per la probabilità che una pila AAA prodotta da `bbb` duri più di 9 ore.
3. Calcolate la durata media campionaria delle 30 pile AAA prodotte da `bbb`.

In realtà, da precedenti analisi statistiche fatte su pile AAA ma prodotte da altre aziende, sappiamo che la durata X di una pila AAA ha distribuzione esponenziale di media $\theta > 0$. In altri termini, da questo momento i dati assegnati costituiscono la realizzazione di un campione casuale $X_1, \dots, X_{30}$ estratto dalla densità esponenziale di media θ incognita.

4. Determinate lo stimatore di massima verosimiglianza della probabilità che una pila AAA prodotta da `bbb` duri più di 9 ore, assumendo l'ipotesi che il modello sia esponenziale.

Indichiamo ora con $p = p(\theta)$ la probabilità che una pila AAA prodotta da `bbb` duri più di 9 ore. Vogliamo costuire un test di livello α per verificare l'ipotesi nulla $H_0 : p = 0.32$ contro l'alternativa $H_1 : p = 0.16$.

5. Dopo aver controllato che il problema di verifica di ipotesi $H_0 : p = 0.32$ contro $H_1 : p = 0.16$ è equivalente al seguente: $H_0 : \theta = 7.8986$ contro $H_1 : \theta = 4.9111$, proponete un test per verificare l'ipotesi nulla $H_0 : p = 0.32$ contro l'alternativa $H_1 : p = 0.16$. Se $\alpha = 1\%$, cosa decidete sulla base del test d'ipotesi costruito?

Esercizio 1.4 È stato condotto un esperimento per verificare se l'attività fisica diurna modifica il livello dell'ormone della crescita prodotto nelle ore di sonno notturno. L'esperimento ha coinvolto sei maschi sani e adulti, ai quali, durante le ore notturne di sonno, sono stati prelevati due campioni di sangue per mezzo di un catetere in vena. Il primo campione di sangue è stato prelevato dopo una giornata in assenza di attività fisica e il secondo dopo una giornata di faticosa attività fisica. Per ogni soggetto e per ogni campione di sangue è stato misurato (in mg/mL) il livello dell'ormone della crescita.

1. Se in 4 soggetti su 6 il livello di ormone della crescita nel primo campione di sangue è superiore al livello nel secondo, ritenete che l'attività fisica riduca il livello dell'ormone della crescita? Impostate ed eseguite un test d'ipotesi di livello 5%. Indicate le eventuali ipotesi necessarie alla conduzione del test.

Chiamiamo ora (x_j, y_j) la coppia che fornisce il livello dell'ormone della crescita dopo una giornata senza attività fisica e dopo una giornata di intensa attività fisica del soggetto j-esimo e assumiamo che il campione dei 6 dati accoppiati (x_j, y_j), $j = 1, ..., 6$, provenga da una popolazione bivariata gaussiana. Inoltre, delle 6 coppie di dati conosciamo le seguenti sintesi:

$$\sum_{i=1}^{6} x_i = 116.1, \quad \sum_{i=1}^{6} y_i = 101.9, \quad \sum_{i=1}^{6} x_i y_i = 2130.82, \quad \sum_{i=1}^{6} x_i^2 = 2992.81, \quad \sum_{i=1}^{6} y_i^2 = 1782.99 \tag{1}$$

2. Calcolate media e varianza campionarie del campione delle differenze $d_j = x_j - y_j$, $j = 1, \dots, 6$.
3. Eseguite un secondo test di ipotesi che usi i dati forniti in (1) sempre al fine di stabilire se l'attività fisica riduca il livello dell'ormone della crescita. Cosa concludete al livello 5%?

Soluzioni

Soluzione dell'Esercizio 1

1. Poiché $\mathrm{E}[\overline{X}_m] = \mathrm{E}[\overline{Y}_n] = \mu$ e quindi $\mathrm{E}[T] = a\,\mathrm{E}[\overline{X}_m] + (1-a)\,\mathrm{E}[\overline{Y}_n] = \mu$, allora T è uno stimatore non distorto di μ, qualunque sia a.

   Per l'indipendenza dei due campioni:

$$\mathrm{Var}[T] = a^2\,\mathrm{Var}[\overline{X}_m] + (1-a)^2\,\mathrm{Var}[\overline{Y}_n] = \frac{a^2\sigma_X^2}{m} + \frac{(1-a)^2\sigma_Y^2}{n}$$

2. Per gli stimatori non distorti l'errore quadratico medio si riduce alla varianza, per cui dobbiamo minimizzare rispetto ad a l'espressione appena trovata. La varianza di T è una parabola in a che va a $+\infty$ quando a diverge; dunque esiste un solo punto di stazionarietà ed è un minimo. Derivando rispetto ad a ed annullando, troviamo:

$$2a\frac{\sigma_X^2}{m} - 2(1-a)\frac{\sigma_Y^2}{n} = 0; \qquad a = \frac{m\sigma_Y^2}{n\sigma_X^2 + m\sigma_Y^2}\,, \qquad 1-a = \frac{n\sigma_X^2}{n\sigma_X^2 + m\sigma_Y^2}\,.$$

3. T è una combinazione lineare di variabili aleatorie gaussiane indipendenti; dunque segue una distribuzione gaussiana di media μ (come visto al punto 1.) e varianza σ_T^2 data da

$$\sigma_T^2 = \frac{a^2\sigma_X^2}{m} + \frac{(1-a)^2\sigma_Y^2}{n} = \frac{\sigma_X^2\sigma_Y^2}{n\sigma_X^2 + m\sigma_Y^2}$$

4. Partendo dalla quantità pivotale $T - \mu/\sigma_T \sim \mathcal{N}(0,1)$, otteniamo con i soliti ragionamenti l'intervallo di confidenza al 95% di estremi

$$\mu = T \pm z_{.975}\sigma_T, \qquad z_{.975} \simeq 1.960\,.$$

   Nel nostro caso si trova $a = 1-a = 0.5$, $t = (\overline{x}_m + \overline{y}_n)/2 = 3.05$, $\sigma_T^2 = 0.005$, $z_{.975}\sigma_T \simeq 0.139$ e risulta dunque $\mu \in (2.911, 3.189)$. ∎

Soluzione dell'Esercizio 2

1. La stima di massima verosimiglianza di θ, basata sul campione casuale delle 120 osservazioni bernoulliane corrispondenti ai risultati degli allievi nella prova senza test, è data dalla frequenza campionaria di allievi promossi $\hat{\theta} = 78/120 = 0.65$.

2. Dato che $n = 120$ è "grande" e $n\theta_0(1-\theta_0) = 120\times 0.64\times 0.36 = 27.648 > 5$, possiamo usare l'approssimazione asintotica della fdr Binomiale di parametri $(120, 0.64)$ con la fdr gaussiana $\mathcal{N}(76.8, 27.648)$. Equivalentemente, per il Teorema centrale del limite, possiamo approssimare la fdr di $\hat{\theta}$ con la fdr $\mathcal{N}(0.64, 0.00192)$. La regione critica di livello α dettata dal test del rapporto di verosimiglianza è $\mathcal{G} = \{\hat{\theta} > q(1-\alpha)\}$, dove $q(1-\alpha)$ è il quantile approssimato di ordine $1-\alpha$ della fdr $\mathcal{N}(0.64, 0.00192)$ e quindi $q(1-\alpha) = \sqrt{0.00192}\times z_{1-\alpha} + 0.64$. Per $\alpha = 5\%$, $q(0.95) = \sqrt{0.00192} * 1.645 + 0.64 \simeq 0.7121$, e accettiamo l'ipotesi nulla $H_0 : \theta = 0.64$. Notate che il p-value approssimato dei dati è $1 - \Phi\Big(\dfrac{0.65 - 0.64}{\sqrt{0.00192}}\Big) \simeq 0.4098$.

3. Introduciamo gli eventi: $E =$"Un allievo scelto a caso raggiunge la sufficienza negli esercizi" e $T =$"Un allievo scelto a caso supera il test". Allora

$$\begin{aligned} P(E) &= P(E|T)P(T) + P(E|T^c)P(T^c) \\ \theta &= (1-0.4)\times 0.8 + \kappa\times(1-0.8) \\ \kappa &= \frac{\theta - (1-0.4)\times 0.8}{0.2} = \frac{\theta - 0.48}{0.2} = 5\theta - 2.4 \end{aligned}$$

   Poiché abbiamo dati soltanto sui promossi ad un appello che non prevedeva test, allora lo stimatore di massima verosimiglianza di κ è dato da $\hat{\kappa} = \kappa(\hat{\theta}) = 5\times\hat{\theta} - 2.4$ e la stima è $\hat{\kappa} = 5\times 0.65 - 2.4 = 0.85 = 85\%$

4. L'ipotesi nulla $H_0 : \kappa = 20\%$ e l'alternativa $H_1 : \kappa > 20\%$ sono equivalenti rispettivamente alle ipotesi nulla $H_0 : \theta = 52\%$ e alternativa $H_0 : \theta > 52\%$. Inoltre, abbiamo dati soltanto sui promossi a un appello che non prevedeva la parte con domande a risposta multipla. Quindi rifiutiamo $H_0 : \kappa = 20\%$ a livello 5% se $\hat\theta > 1.645\sqrt{\dfrac{0.52 \times (1-0.52)}{120}} + 0.52 = 0.595$. Essendo $\hat\theta = 0.65$ effettivamente rifiutiamo $H_0 : \kappa = 20\%$.

5. Il valore della funzione di potenza del test al punto 2. in $\theta = 66\%$ è

$$P_{0.66}(\hat\theta > 0.7121) \simeq 1 - \Phi\left(\frac{0.7121 - 0.66}{\sqrt{\frac{0.66\times 0.34}{120}}}\right) \simeq 1 - \Phi\left(\frac{0.7121 - 0.66}{\sqrt{0.0018}}\right) \simeq 1 - \Phi(1.23) \simeq 1 - 0.8906 \simeq 0.11. \blacksquare$$

Mentre, $\kappa = 22.5\%$ se e solo se $\theta = 52.5\%$ e quindi, il valore della funzione di potenza del test al punto 4. in $\kappa = 22.5\%$ è

$$P_{0.525}(\hat\theta > 0.529) \simeq 1 - \Phi\left(\frac{0.595 - 0.525}{\sqrt{\frac{0.525\times(1-0.525)}{120}}}\right) \simeq 6.25\%.$$

Soluzione dell'Esercizio 3

1. La caratteristica da stimare è $p := P(X > 9)$ e uno stimatore "non parametrico" non distorto, consistente e asintoticamente gaussiano è $T_1 = 1 - \hat F_{30}(9) = 1 - \hat F_{30}(8) = 1 - 23/30 = 7/30 = 0.2\bar{3}$.

2. T_1 è la frequenza campionaria dell'evento "Una pila AAA prodotta da `bbb` scelta a caso dura più di 9 ore" su 30 pile osservate. Approssimativamente T_1 ha fdr gaussiana di media p e varianza $p(1-p)/30$. Allora un IC(p) asintotico di livello 95% ha estremi $\dfrac{7}{30} \pm z_{\frac{1+0.95}{2}}\sqrt{\dfrac{\frac{7}{30}\times\frac{23}{30}}{30}}$, cioè è dato da $(0.082, 0.385)$ (lunghezza $\simeq 0.303$).

   Viene anche usato l'intervallo "più prudente" $\dfrac{7}{30} \pm t_{.975}(29)\sqrt{\dfrac{\frac{7}{30}\times\frac{23}{30}}{29}}$, che fornisce $(0.073, 0.394)$ (lunghezza $\simeq 0.321$). La scelta "migliore" tra gli intervalli asintotici è però l'intervallo che si ottiene risolvendo due disequazioni…, ma che non è calcolabile in un compito; quest'ultimo intervallo fornisce $(0.118, 0.409)$ (lunghezza $\simeq 0.291$).

3. $$\bar X = \sum_{k\in\{1,2,4,6,8,11,13,27,29,42\}} k\left(\hat F_{30}(k) - \hat F_{30}(k-1)\right) = 7.5$$

4. Lo stimatore ML di θ di un modello statistico esponenziale di media θ è $\bar X$ e quindi lo stimatore ML di $p = p(\theta) = P_\theta(X > 9) = \displaystyle\int_9^\infty \frac{1}{\theta}\,\mathrm{e}^{-x/\theta}\,\mathrm{d}x = \mathrm{e}^{-9/\theta}$ è $p(\bar X) = \exp\left(-9/\bar X\right)$. Con i dati assegnati, la stima di massima verosimiglianza di p è $\hat p = \mathrm{e}^{-9/7.5} \simeq 0.301$.

5. Essendo entrambe le ipotesi nulla e alternativa semplici, possiamo costruire il test più potente, nella classe dei test di livello α, dettato dal Lemma di Neyman Pearson. Osserviamo poi che $p = 0.16$ se e solo se $\theta = -9/\log 0.16 \simeq 4.9111$ e $p = 0.32$ se e solo se $\theta = -9/\log 0.32 \simeq 7.8986$ quindi, $H_0 : p = 0.32$ è equivalente a $H_0 : \theta = 7.8986$ e $H_1 : p = 0.16$ è equivalente a $H_1 : \theta = 4.9111$. Abbiamo così ridotto il problema ad un test sulla media per una popolazione esponenziale. Il test di Neyman Pearson detta di rifiutare H_0 se $\bar X \le k$ con k tale che $P_{7.8986}\left(\bar X \le k\right) = \alpha$. Osservando che la media campionaria $\bar X$ di un campione casuale esponenziale di taglia 30 ha distribuzione $\Gamma(30, \frac{\theta}{30})$, allora $\frac{2\times 30\times \bar X}{7.8986}$ sotto H_0 ha distribuzione χ^2_{60} e il test di livello α sarà: rifiuta H_0 se $\frac{2\times 30\times\bar X}{7.8986} < \chi^2_{60}(\alpha)$. Possiamo approssimare la fdr χ^2_{60} con la fdr $\mathcal{N}(60, 120)$ e quindi, per $\alpha = 1\%$, $\chi^2_{60}(0.01) \simeq -\sqrt{120}z_{0.99} + 60 \simeq 50.81017$. Essendo $\frac{2\times 30\times 7.5}{7.8986} \simeq 56.97 > 50.81017$, accetteremo H_0.

   Alternativamente, poiché in questo caso $n = 30$ può essere considerato grande, si può procedere osservando che per il Teorema centrale del limite la fdr approssimata della media campionaria $\bar X$ di un numero n "grande" di variabili aleatorie i.i.d. di media 7.8986 e varianza 7.8986^2 è $\mathcal{N}\left(\theta_0, \frac{7.8986^2}{30}\right)$. $\blacksquare$

Soluzione dell'Esercizio 4 Siano X ed Y le v.a. che indicano, rispettivamente, il livello dell'ormone della crescita misurato dopo una giornata senza attività fisica e dopo una giornata di intensa attività fisica e siano F e G le rispettive fdr.

1. L'unica informazione che abbiamo riguarda il valore della statistica T^+ che indica il numero di coppie in cui X è maggiore di Y e vale 4. Possiamo quindi effettuare il test dei segni di Wilcoxon per dati accoppiati che verifichi se X ed Y provengano dalla stessa distribuzione oppure se X domini stocasticamente Y:

$$H_0 : \text{“}F(a) = G(a) \text{ per ogni } a\text{” vs } H_1 : \text{“}F(a) \leq G(a) \text{ per ogni } a \text{ e } F(a^*) < G(a^*) \text{ per qualche } a^*\text{”}$$

Se assumiamo che i 6 campioni accoppiati siano indipendenti e che le fdr marginali F, G siano continue, allora, sotto H_0, T^+ ha densità $\mathrm{Bin}(6, 1/2)$ e il test dei segni di Wilcoxon per dati accoppiati detta di rifiutare H_0 a livello α se $T^+ > q_{1-\alpha}$ dove $q_{1-\alpha}$ è il quantile di ordine $1-\alpha$ della fdr $\mathrm{Bin}(6, 1/2)$. Ovvero, rifiuteremo H_0 per ogni α tale che

$$\alpha > P\left(\mathrm{Bin}\left(6, \frac{1}{2}\right) > 4\right)) = \binom{6}{5}\frac{1}{2^6} + \binom{6}{6}\frac{1}{2^6} \simeq 11\%.$$

Essendo 5% minore del p-value del test (p-value=11%), allora a livello 5% concludiamo che l'attività fisica non riduce l'ormone della crescita.

2.

$$\begin{aligned}
\bar{d} &= \bar{x} - \bar{y} = \frac{116.1 - 101.9}{6} = 2.367 \\
s_D^2 &= \frac{\sum_{j=1}^{6}(d_j - \bar{d})^2}{5} = \frac{1}{5}\left(\sum_{j=1}^{6} x_j^2 + \sum_{j=1}^{6} y_j^2 - 2\sum_{j=1}^{6} x_j y_j - 6 \times (2.367)^2\right) \\
&= \frac{1}{5}(2992.81 + 1782.99 - 2 \times 2130.82 - 33.6161) \simeq 96.109
\end{aligned}$$

3. Supponendo che il campione provenga da una densità congiunta gaussiana, si può affrontare il problema con un test t sulla differenza delle medie di popolazioni gaussiane, ossia considerare il campione univariato delle differenze $D_j = X_j - Y_j$, $j = 1, \ldots, 6$ proveniente da una densità gaussiana di media $\mu_D = \mu_X - \mu_Y$ e varianza σ_D^2 incognite. Le precedenti ipotesi statistiche si traducono nelle seguenti ipotesi sulla media μ_D:

$$H_0 : \mu_D = 0 \text{ vs } H_1 : \mu_D > 0$$

e l'ipotesi H_0 è rifiutata a livello α se $T := \frac{\bar{D}-0}{S_D/\sqrt{6}} > t_5(1-\alpha)$, dove $t_5(1-\alpha)$ è il quantile di ordine $1-\alpha$ della fdr t di Student con 5 gradi di libertà. Sostituendo i valori trovati al punto 2. otteniamo $T = 0.5914$ e se $\alpha = 0.05$ allora $t_5(1-\alpha) = 2.015$. Quindi al livello 5%, anche con i dati in (1), accetteremo H_0. Usando le tavole, è facile verificare che il p-value del test è maggiore di 0.25 (con un pacchetto statistico otteniamo p-value $\simeq 0.29$). ∎

2 STATISTICA per ING INF [2L] Proff. A. Barchielli, I. Epifani 14.07.05

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

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

Esercizio 2.1 Sia $X_1, \dots, X_n$ un campione casuale estratto dalla popolazione gaussiana di densità

$$f(x, \theta) = \frac{\theta^{3/2}}{\sqrt{3\pi}} \mathrm{e}^{-\frac{\theta^3 x^2}{3}}, \qquad \theta > 0$$

dove θ è un parametro incognito.

1. Determinate lo stimatore di massima verosimiglianza di θ.

2. Costruite un test per verificare l'ipotesi nulla $H_0 : \theta = 1$ contro l'alternativa $H_1 : \theta \neq 1$ a livello $\alpha = 5\%$. Se avete rilevato il campione di quattro osservazioni: $-0.17,\ 0.71,\ 2.17,\ 1.00$, accettate o rifiutate H_0?

3. Se X_1, X_2, X_3, X_4 sono i.i.d. $\sim \mathcal{N}(0, 12)$, qual è la distribuzione di $\dfrac{\sum_{j=1}^4 X_j^2}{12}$? Scrivete l'espressione della densità e determinate esplicitamente la funzione di ripartizione.

4. Calcolate la potenza del test costruito al punto 2. in $\theta = 1/2$. (*Sugg. è utile quanto trovato al punto precedente*).

Soluzione dell'Esercizio 1 Osserviamo che $f(x, \theta) = \mathcal{N}\left(0, \dfrac{3}{2\theta^3}\right)$: dobbiamo fare inferenza su una caratteristica che dipende dalla varianza di una popolazione gaussiana con media nota e pari a zero. Chiamiamo σ^2 la varianza di questa densità gaussiana. La caratteristica su cui fare inferenza è $\theta = \left(\dfrac{3}{2\sigma^2}\right)^{1/3}$.

1. Lo stimatore di massima verosimiglianza di σ^2 nel modello gaussiano con media nota e pari a zero è $\widehat{\sigma^2} = \dfrac{\sum_{j=1}^n X_j^2}{n}$. Segue che quello di θ è $\hat{\theta} = \left(\dfrac{3n}{2\sum_{j=1}^n X_j^2}\right)^{1/3}$.

2. Se esprimiamo ipotesi nulla e alternativa in funzione della varianza σ^2 del modello gaussiano, esse diventano rispettivamente: $H_0 : \sigma^2 = 1.5$ e $H_1 : \sigma^2 \neq 1.5$ e quindi rifiutiamo $H_0 : \theta = 1$ a livello α se

$$\frac{\sum_{j=1}^4 X_j^2}{1.5} \notin \left(\chi_n^2\left(\frac{\alpha}{2}\right), \chi_n^2\left(1 - \frac{\alpha}{2}\right)\right)$$

Se $\alpha = 5\%$ e $n = 4$ allora $\chi_4^2(0.025) = 0.484$ e $\chi_4^2(0.975) = 11.1$. Inoltre, $\sum_{j=1}^4 X_j^2 = (-0.17)^2 + 0.71^2 + 2.17^2 + 1^2) = 6.2419$ e $\frac{\sum_{j=1}^4 X_j^2}{1.5} \simeq 4.16 \in (0.484, 11.1)$. Quindi accettiamo $H_0 : \theta = 1$.

3. Se $X_1, \dots, X_4$ i.i.d. $\sim \mathcal{N}(0, 12)$ allora $\frac{\sum_{j=1}^4 X_j^2}{12} \sim \chi_4^2$. La densità χ_4^2 è

$$f_4(x) = \frac{x}{4} \mathrm{e}^{-\frac{x}{2}} \mathbf{1}_{(0,\infty)}(x)$$

e la corrispondente funzione di ripartizione è

$$F_4(x) = \begin{cases} 0 & x \leq 0 \\ \int_0^x \frac{u}{4} \mathrm{e}^{-\frac{u}{2}}\, du = 1 - \frac{x}{2}\mathrm{e}^{-\frac{x}{2}} - \mathrm{e}^{-\frac{x}{2}} & x > 0 \end{cases}$$

4. $\theta = 1/2$ se e solo se $\sigma^2 = 12$ e se $\sigma^2 = 12$ allora $\frac{\sum_{j=1}^4 X_j^2}{12} \sim \chi_4^2$. La potenza del test in $\sigma^2 = 12$ è la

probabilità di rifiutare $\sigma^2 = 1.5$ quando $\sigma^2 = 12$, cioè:

$$\begin{aligned}
\pi(12) &= 1 - P_{\sigma^2}\left(0.484 < \frac{\sum_{j=1}^4 X_j^2}{1.5} < 11.1\right), \quad \sigma^2 = 12 \\
&= 1 - P_{12}\left(\frac{0.484 \times 1.5}{12} < \frac{\sum_{j=1}^4 X_j^2}{12} < \frac{11.1 \times 1.5}{12}\right) \\
&= 1 - P_{12}\left(0.0605 < \frac{\sum_{j=1}^4 X_j^2}{12} < 1.3875\right) = 1 - F_4(1.3875) + F_4(0.0605) \\
&= \frac{1.3875}{2}\mathrm{e}^{-\frac{1.3875}{2}} + \mathrm{e}^{-\frac{1.3875}{2}} + 1 - \frac{0.0605}{2}\mathrm{e}^{-\frac{0.0605}{2}} - \mathrm{e}^{-\frac{0.0605}{2}} \simeq 0.984 \ \blacksquare
\end{aligned}$$

Esercizio 2.2 Nella roulette francese regolare (primi 18 numeri pari neri, primi 18 numeri dispari rossi e lo zero verde), la probabilità θ di vincere puntando sul rosso è 18/37. Sospettiamo che la roulette dell'amico con cui giochiamo sempre sotto Natale sia truccata. Per questo motivo, una sera che si giocava, abbiamo contato il numero di "giochi" necessari perché uscisse rosso la prima volta e abbiamo annotato questo valore chiamandolo x_1. Poi abbiamo contato quante altre partite, dopo le prime x_1, abbiamo dovuto aspettare per vedere rosso per la seconda volta e abbiamo annotato questo numero x_2 e così via. Quella sera è uscito rosso n volte: la prima volta alla x_1-esima partita, la seconda volta alla partita numero $x_1 + x_2$, ..., l'n-esima volta alla partita numero $x_1 + \ldots + x_n$. Abbiamo quindi i dati di un campione casuale di n osservazioni $X_1, \ldots, X_n$ provenienti da una popolazione *geometrica di parametro* θ con $\theta \in (0, 1)$.

1. Determinate gli stimatori di massima verosimiglianza del parametro θ, della media e della varianza della densità geometrica (di parametro θ).
2. Verificate che lo stimatore di massima verosimiglianza della media è efficiente.
3. Provate ad argomentate perché lo stimatore di massima verosimiglianza della varianza non è efficiente.

Quella sera abbiamo registrato i seguenti dati:

$$(x_1, \ldots, x_{36}) = (1, 4, 1, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 4, 1, 1, 2, 3, 1, 1, 3, 2, 4, 1, 1, 2, 5, 2, 1, 6, 1, 1, 6, 3, 4)$$

4. Sulla base di questi dati, stimate il numero medio di giocate necessarie per ottenere rosso.
5. Verificate l'ipotesi nulla $H_0 : \theta = 18/37$ contro l'alternativa $H_0 : \theta \neq 18/37$, usando un test per grandi campioni.

Soluzione dell'Esercizio 2 Abbiamo il campione $X_1, \ldots, X_n$ i.i.d. $\sim f(x, \theta)$ con

$$f(x, \theta) = \theta(1-\theta)^{x-1}\mathbf{1}_{\{1,2,\ldots\}}(x) \quad 0 < \theta < 1$$

le caratteristiche media e varianza di $f(x, \theta)$ sono $\mu = \mu(\theta) = \theta^{-1}$ e $\sigma^2 = \sigma^2(\theta) = (1-\theta)/\theta^2 = \theta^{-2} - \theta^{-1}$.

1. La funzione di verosimiglianza è

$$L_\theta(x_1, \ldots, x_n) = \prod_{j=1}^n f(x_j, \theta) = \prod_{j=1}^n \theta(1-\theta)^{x_j - 1} = \theta^n (1-\theta)^{\sum_{j=1}^n x_j - n}$$

Per determinare lo stimatore di massima verosimiglianza (ML) di θ procediamo a massimizzare $\log L_\theta$:

$$\frac{\partial}{\partial\theta}\log L_\theta(x_1, \ldots, x_n) = \frac{\partial}{\partial\theta}\left(n\log\theta + \left[\sum_{j=1}^n x_j - n\right]\log(1-\theta)\right) = -\frac{n}{1-\theta}\left(\bar{x} - \frac{1}{\theta}\right) \qquad (2)$$

$\bar{x}$ è la media campionaria e

$$-\frac{n}{1-\theta}\left(\bar{x} - \frac{1}{\theta}\right) \geq 0 \quad \text{se e solo se} \quad \theta \leq \frac{1}{\bar{x}}$$

Segue che lo stimatore ML di θ è $\hat{\theta} = \bar{X}^{-1}$, lo stimatore ML della media μ è la media campionaria $\bar{X}$ e lo stimatore ML della varianza σ^2 è $\widehat{\sigma^2} = \bar{X}^2 - \bar{X} = \bar{X}(\bar{X}-1)$.

2. La media campionaria $\bar{X}$ è stimatore non distorto della media teorica $\mu = \theta^{-1}$, il modello geometrico è regolare e leggiamo nell'equazione (2) che la derivata del logaritmo della funzione di verosimiglianza è *"essenzialmente"* funzione lineare della differenza fra $\bar{X}$ e la caratteristica da stimare θ^{-1}. Ma (2) è condizione necessaria e sufficiente affinché la varianza di $\bar{X}$ raggiunga il confine di Cramer Rao $(\mu'(\theta))^2/(nI(\theta))$ dove $I(\theta)$ è l'Informazione di Fisher del modello geometrico.

Oppure:

$$\begin{aligned}
\mathrm{Var}_\theta(\bar{X}) &= \frac{\mathrm{Var}_\theta(X_1)}{n} = \frac{1-\theta}{\theta^2 n} \\
nI(\theta) &= E_\theta\left(-\frac{n}{1-\theta}\left(\bar{X}-\frac{1}{\theta}\right)\right)^2 = \frac{n^2}{(1-\theta)^2}\mathrm{Var}_\theta(\bar{X}) = \frac{n}{\theta^2(1-\theta)} \\
[\mu(\theta)']^2 &= \frac{1}{\theta^4} \\
\text{Confine di Cramer Rao} &= \frac{[\mu(\theta)']^2}{nI(\theta)} = \frac{1-\theta}{\theta^2 n} = \mathrm{Var}_\theta(\bar{X})
\end{aligned}$$

3. Se $\frac{\partial}{\partial\theta}\log L_\theta(x_1,\ldots,x_n)$ è *"essenzialmente"* funzione lineare di $\bar{X}$, allora non può contemporaneamente essere funzione quadratica di $\bar{X}$. Quindi $\bar{X}^2 - \bar{X}$ non è stimatore efficiente della caratteristica varianza.

4. La stima della media μ è $\bar{x} = (1/36)\times(1\times 18 + 2\times 8 + 3\times 3 + 4\times 4 + 5\times 1 + 6\times 2) = 19/9 \simeq 2.11$.

5. Il problema dato è equivalente a $H_0 : \mu = 37/18$ contro $H_1 : \mu \neq 37/18$. Costruiamo quindi un test sulla media di un campione casuale "numeroso" (36) e rifiutiamo H_0 per ogni α tale che

$$\frac{\left|\bar{x}-\frac{37}{18}\right|}{\sqrt{\left(\left(\frac{37}{18}\right)^2-\frac{37}{18}\right)\times\frac{1}{36}}} > z_{1-\frac{\alpha}{2}}$$

cioè il p-value del test è

$$\begin{aligned}
2\left[1-\Phi\left(\frac{\left|\bar{x}-\frac{37}{18}\right|}{\sqrt{\left(\left(\frac{37}{18}\right)^2-\frac{37}{18}\right)\times\frac{1}{36}}}\right)\right] &= 2\left[1-\Phi\left(\frac{\left|\frac{19}{9}-\frac{37}{18}\right|}{\sqrt{\left(\left(\frac{37}{18}\right)^2-\frac{37}{18}\right)\times\frac{1}{36}}}\right)\right] = \\
&= 2\left[1-\Phi\left(\frac{6}{\sqrt{19\times 37}}\right)\right] \simeq 2\left[1-\Phi\left(0.2263\right)\right] \simeq 2\left[1-\Phi\left(0.23\right)\right] \simeq 0.8181
\end{aligned}$$

Concludiamo che c'è una forte evidenza empirica ad accettare l'ipotesi nulla che la probabilità che esca rosso sulla roulette valga 18/37: il sospetto che la roulette del nostro amico sia truccata sembrerebbe infondato.

Oppure:

$\hat{\theta}$ è stimatore ML di θ; "36 è grande" e il modello geometrico è regolare. Dalle proprietà asintotiche degli stimatori ML in modelli regolari deriva che la fdr limite di $\sqrt{nI(\theta)}(\hat{\theta}-\theta)$ converge alla fdr $\mathcal{N}(0,1)$ per $n \to \infty$, cioè per n grande approssimativamente vale $\dfrac{\sqrt{n}(\hat{\theta}-\theta)}{\sqrt{\theta^2(1-\theta)}} \sim \mathcal{N}(0,1)$ se θ è il vero valore del parametro. Potremmo così impostare il seguente test per la media (asintotica) θ per grandi campioni: rifiutare $H_0 : \theta = 18/37$ contro $H_1 : \theta \neq 18/37$ a livello α se $\dfrac{\sqrt{n}\left|\hat{\theta}-\frac{18}{37}\right|}{\sqrt{(\frac{18}{37})^2(1-\frac{18}{37})}} > z_{1-\alpha/2}$. Il p-value di questo test è $2\left[1-\Phi\left(\frac{3}{19}\times\sqrt{\frac{37}{19}}\right)\right] \simeq 2\left[1-\Phi\left(0.2203\right)\right] \simeq 0.8259$. Notate che i due test sono diversi, quindi i p-value possono differire... ∎

Esercizio 2.3 In un esperimento di laboratorio viene misurata la temperatura ritenuta "più piacevole" da 8 donne e 10 uomini. Dalla classifica dei 18 esaminati (fatta in base alla temperatura per ciascuno di loro più piacevole) scopriamo che la persona meno freddolosa è un uomo (nella classifica occupa la prima posizione) e la più freddolosa è una donna (in ultima posizione c'è appunto una donna). Andando nel dettaglio, la tabella delle posizioni occupate da uomini e donne nella classifica è la seguente

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	u	d	u	u	u	u	d	u	u	u	d	u	d	d	u	d	d	d

(u=uomo, d=donna).

1. Impostate un test di verifica dell'ipotesi nulla H_0 :"*la temperatura più piacevole è la stessa per uomini e donne*" contro l'alternativa H_1 :"*Alle donne piace una temperatura più alta che agli uomini*", di livello $\alpha = 5\%$.

In realtà, i ricercatori del laboratorio ci hanno fornito anche i seguenti dati sui valori di temperatura (in gradi Celsius):

$$\sum_{j=1}^{8} x_j = 198.8, \qquad \sum_{j=1}^{8} x_j^2 = 4949.08, \qquad \sum_{j=1}^{10} y_j = 239.3, \qquad \sum_{j=1}^{10} y_j^2 = 5735.65$$

dove x_j è la temperatura più piacevole per la donna j-esima e y_j quella più piacevole per l'uomo j-esimo e ci hanno detto che i due campioni delle temperature possono essere assunti entrambi gaussiani.

2. Verificate con un test di significatività $\alpha = 5\%$ se i due campioni provengano da popolazioni con la stessa varianza.

3. Costruite un secondo test di ipotesi sempre di livello $\alpha = 5\%$, che usi le informazioni fornite dai ricercatori sui valori di temperature e le assunzioni di gaussianità, sempre per verificare H_0 :"*la temperatura più piacevole è la stessa per uomini e donne*" contro l'alternativa H_1 :"*Alle donne piace una temperatura più alta che agli uomini*".

Soluzione dell'Esercizio 3 Introduciamo le variabili aleatorie X, Y con $X =$ "temperatura più piacevole per le donne" e $Y =$ "temperatura più piacevole per gli uomini".

1. Se abbiamo a disposizione la sola classifica dei 18 esaminati in base alla temperatura "più piacevole", impostiamo il test di omogeneità non parametrico di Wilcoxon-Mann-Wintney e usiamo la statistica T_X data dalla somma dei ranghi delle temperature delle donne: $T_X = 2+7+11+13+14+16+17+18 = 98$. L'ipotesi nulla H_0 :"*la temperatura più piacevole è la stessa per uomini e donne*" si traduce in H_0 :"X, Y sono regolate dalla stessa fdr" e l'alternativa H_1 :"*Alle donne piace una temperatura più alta che agli uomini*" si traduce in H_1:"X domina stocasticamente Y". Quindi confrontiamo 98 con il quantile di ordine $w_{0.95}$ della statistica T_X con $m = 8$ e $n = 10$: $w_{0.95} = 8 \times (8 + 10 + 1) - w_{0.05} = 152 - 57 = 95$. Essendo $98 > 95$, allora a livello 5% accettiamo l'ipotesi che le donne stiano meglio a temperature più elevate rispetto agli uomini.

Alla luce delle nuove informazioni, il modello statistico di riferimento è quello di due campioni casuali indipendenti: $X_1, X_2, \ldots, X_8$ i.i.d. $\sim \mathcal{N}(\mu_X, \sigma_X)$ e $Y_1, Y_2, \ldots, Y_{10}$ i.i.d. $\sim \mathcal{N}(\mu_Y, \sigma_Y)$ entrambi con medie e varianze incognite.

2. Impostiamo un F-test per verificare $H_0 : \sigma_X^2 = \sigma_y^2$ contro $H_1 : \sigma_X^2 \neq \sigma_y^2$ e usiamo la statistica test S_X^2/S_Y^2:

$$\bar{X} = \frac{198.8}{8} \simeq 24.85, \quad \bar{Y} = \frac{239.3}{10} = 23.93 \text{ e } \frac{S_X^2}{S_Y^2} = 1.245 \quad \text{in quanto}$$

$$S_X^2 = \frac{\sum_{j=1}^{8} X_j^2 - 8\bar{X}^2}{7} = \frac{4949.08 - 4940.18}{7} \simeq 1.27, \quad S_Y^2 = \frac{\sum_{j=1}^{10} Y_j^2 - 10\bar{Y}^2}{9} = \frac{5735.65 - 5726.449}{9} \simeq 1.02$$

Sotto H_0, S_X^2/S_Y^2 ha densità F di Fisher con $(7,9)$ gradi di libertà e rifiuteremo H_0 a livello 0.05 se $S_X^2/S_Y^2 \notin (F_{7,9}(0.025), F_{7,9}(0.975)) = (1/4.82, 4.2) \simeq (0.2075, 4.2)$: ma $1.245 \in (0.2075, 4.2)$ quindi accettiamo l'ipotesi di varianze uguali.

3. Chiamiamo σ^2 la comune varianza; uno stimatore non distorto di σ^2 basato su entrambi i campioni è dato dalla varianza *pooled*:

$$S_p^2 = S_X^2 \frac{m-1}{m+n-2} + S_Y^2 \frac{n-1}{m+n-2} = 1.27 \times \frac{7}{16} + 1.02 \times \frac{9}{16} \simeq 1.13.$$

Sotto ipotesi di gaussianità e varianze incognite ma uguali, l'ipotesi nulla diventa $H_0 : \mu_X = \mu_Y$ contro l'alternativa $H_0 : \mu_X > \mu_Y$. Dobbiamo così impostare un t-test di confronto fra medie per l'alternativa unilatera del tipo $H_0 : \mu_X > \mu_Y$. La statistica test ha valore $\dfrac{24.85 - 23.93}{\sqrt{1.13(1/8 + 1/10)}} \simeq 1.8245$ e il quantile di ordine 0.95 della fdr t di student con 16 gradi di libertà vale 1.746: anche con questo test rifiutiamo H_0 e accettiamo l'ipotesi che le donne stiano meglio a temperature più elevate rispetto agli uomini. ∎

Esercizio 2.4 Recita la guida all'immatricolazione in rete: "*La prova d'ammissione per i corsi di studio delle Facoltà di Ingegneria ha carattere orientativo: gli studenti sono invitati a valutarne seriamente l'esito e l'indice attitudinale che ne consegue. ... La prova d'ammissione rappresenta la verifica di un'adeguata preparazione iniziale. Se il risultato segnala carenze formative in una o più aree -i cosiddetti debiti formativi- gli studenti dovranno provvedere a colmarle*".

Abbiamo i dati riguardanti un campione di 1366 studenti provenienti dal liceo scientifico, 118 dal liceo classico e 371 dagli istituti tecnici. Hanno raggiunto il punteggio minimo di sufficiente preparazione iniziale 805 studenti del liceo scientifico, 47 del liceo classico e 68 degli istituti tecnici.

1. Sulla base di questi dati, l'adeguatezza della preparazione iniziale è legata al tipo di scuola di provenienza? Rispondete costruendo un opportuno test di ipotesi.

Supponiamo ora che comunque sia andata la prova d'ammissione, tutti gli studenti di quel campione abbiano deciso di iscriversi a un corso di ingegneria.

2. Dopo aver fornito una stima puntuale della percentuale θ di matricole del prossimo AA 2005/2006 provenienti dagli istituti tecnici, determinate un intervallo asintotico di confidenza 90% per θ.

Soluzione dell'Esercizio 4

1. Organizziamo i dati in una tabella a doppia entrata, completandola con i dati mancanti:

scuola di provenienza \ *punteggio minimo*	raggiunto	NON raggiunto	
Istituti tecnici	68	$371 - 68 = 303$	371
Liceo classico	47	$118 - 47 = 71$	118
Liceo scientifico	805	$1366 - 805 = 561$	1366
	$68 + 47 + 805 = 920$	$1855 - 920 = 935$	$371 + 118 + 1366 = 1855$

e impostiamo un test χ^2 di indipendenza fra le variabili categoriche X =scuola di provenienza e Y = indicatore del punteggio minimo se raggiunto o no. La statistica di Pearson Q ha valore:

$$Q = 1855 \times \left(\frac{68^2}{371 \times 920} + \frac{47^2}{118 \times 920} + \frac{805^2}{1366 \times 920} + \frac{303^2}{371 \times 935} + \frac{71^2}{118 \times 935} + \frac{561^2}{1366 \times 935} - 1 \right) \simeq 197.2$$

Asintoticamente Q ha fdr $\chi^2_{(3-1)(2-1)} = \chi^2_2$, che coincide con la fdr esponenziale di parametro 2. Quindi, il p-value del test di Pearson è $P(Q > 197.2115) = \mathrm{e}^{-197.2115/2} \simeq 0$: c'è una forte evidenza empirica a rifiutare l'ipotesi H_0 : "X, Y sono indipendenti", ossia concludiamo che l'adeguatezza della preparazione dipende dalla scuola di provenienza. [Se vi siete limitati a usare le tavole dei quantili della fdr χ^2_2, avrete trovato p-value≤ 0.001]

2. La stima di massima verosimiglianza di θ basata sul campione casuale delle 1855 osservazioni bernoulliane corrispondenti agli studenti delle superiori che hanno sostenuto la prova di ingresso ad ingegneria è data dalla frequenza campionaria degli studenti provenienti dagli istituti tecnici $\hat{\theta} = 371/1855 = 20\%$ e un intervallo asintotico per θ di confidenza 90% ha estremi

$$\hat{\theta} \pm \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{1855}} z_{\frac{1+0.9}{2}} = 0.2 \pm \sqrt{\frac{0.2 \times 0.8}{1855}} \times 1.645$$

ed è quindi dato da $(18.47\%, 21.53\%)$. Dato che è una valutazione molto approssimata diciamo che l'intervallo è $(18\%, 22\%)$. ∎

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

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

Esercizio 3.1 Un chimico ha sviluppato un nuovo tipo di batterie per calcolatrici che, egli sostiene, dura "significativamente più a lungo" rispetto a quelle correntemente sul mercato.
Un'azienda intenzionata a produrle misura le durate di un campione di n batterie del nuovo tipo. L'azienda sa che la durata delle batterie per calcolatrici correntemente sul mercato è modellabile mediante una variabile aleatoria gaussiana di media 100.3min e deviazione standard 6.25min; inoltre, da verifiche preliminari, l'azienda può assumere che anche la durata del nuovo tipo di batterie sia gaussiana con deviazione standard 6.25min. Infine, l'azienda è disposta a commettere un errore di prima specie di probabilità al più 5% di produrre batterie del nuovo tipo effettivamente *non migliori di quelle correnti.*

1. Impostate un opportuno test di ipotesi per aiutare l'azienda a decidere se produrre o meno le nuove batterie; in particolare dovrete specificare: ipotesi nulla, ipotesi alternativa e regione critica.
2. Sono state misurate le durate di 16 batterie del nuovo tipo: se la durata media è risultata 104.6 minuti, cosa deciderà l'azienda usando il vostro test?
3. Sempre nell'ipotesi che si misurino le durate di 16 batterie del nuovo tipo, qual è la potenza del vostro test se effettivamente le nuove batterie avessero durata media di 105 minuti?
4. Determinate il numero minimo n di durate da rilevare affinché, se effettivamente le nuove batterie avessero durata media di 105 minuti, allora almeno il 95% delle volte si concluderebbe correttamente che le nuove batterie sono migliori delle correnti.

Soluzione Sia X la variabile aleatoria che modella la durata di una batteria di nuovo tipo. Allora $X \sim \mathcal{N}(\mu, 6.25^2)$ e traduciamo la congettura: *"le batterie del nuovo tipo sono non migliori di quelle correnti"* in termini della media μ come: "$\mu \leq 100.3$".

1. Dobbiamo impostare un test di verifica dell'ipotesi nulla H_0 :"$\mu \leq 100.3$" contro l'alternativa H_1 :"$\mu > 100.3$" di ampiezza 5%. Avendo a disposizione un campione casuale $x_1, \dots, x_n$ di dimensione n estratto dalla popolazione gaussiana di varianza nota e pari a 6.25^2, usiamo la seguente regione critica: $\mathcal{G} = \{(x_1, \dots, x_n) : \bar{x} \geq 1.645 \times 6.25/\sqrt{n} + 100.3\}$, dove $\bar{x}$ indica la media campionaria e 1.645 è il quantile di ordine 0.95 della fdr gaussiana standard.
2. Se $n = 16$ e $\bar{x} = 104.6$ allora consiglieremo all'azienda di produrre il nuovo tipo di batterie perché $1.645 \times 6.25/\sqrt{16} + 100.3 = 102.8703$ e $104.6 > 102.8703$. Il livello descrittivo del nostro test (p-value) è $1 - \Phi(\frac{104.6-100.3}{6.25/4}) = 1 - \Phi(2.752) \simeq 0.003$, quindi vi è una forte evidenza empirica a rifiutare l'ipotesi nulla che le nuove batterie siano non migliori di quelle correntemente prodotte.
3. Ci viene richiesto di calcolare la funzione di potenza $\pi(\mu)$ del test per $n = 16$ nel punto $\mu = 105$: con 16 osservazioni rifiutiamo H_0 se $\bar{X} \geq 1.645 \times 6.25/\sqrt{16} + 100.3 = 102.8703$ e
$$\pi(105) = P_{105}(\bar{X} \geq 102.8703) = 1 - \Phi\left(\frac{102.8703 - 105}{6.25/4}\right) = 1 - \Phi(-1.363) = \Phi(1.363) \simeq 0.9131 = 91.31\%$$
4. Dobbiamo determinare il minimo n tale che $P_{105}(\text{rifiutare } H_0) \geq 0.95$, ossia il minimo n tale che
$$P_{105}\left(\bar{X} \geq 1.645 \times 6.25/\sqrt{n} + 100.3\right) \geq 0.95$$
La precedente disequazione in n diventa:
$$1.645 + (100.3 - 105)\frac{\sqrt{n}}{6.25} \leq z_{0.05} = -z_{0.95} = -1.645$$
se e solo se
$$\sqrt{n} \geq \frac{3.29}{0.752}$$
se e solo se
$$n \geq 20. \quad \blacksquare$$

Esercizio 3.2 Abbiamo estratto un campione casuale $X_1, \dots, X_n$ dalla densità (di Raleigh) di parametro $\theta > 0$:

$$f(x, \theta) = \frac{x}{\theta^2} \exp\left\{-\frac{x^2}{2\theta^2}\right\} \mathbf{1}_{(0,\infty)}(x)$$

1. Determinate lo stimatore della caratteristica $\kappa(\theta) = \theta^2$ usando il metodo di massima verosimiglianza.
2. Determinate la densità di $Y = X^2$.
3. Chiamiamo $\hat{\kappa}_n$ lo stimatore di massima verosimiglianza di $\kappa(\theta) = \theta^2$: determinate come è distribuita la variabile aleatoria $Q_n = \dfrac{2n\hat{\kappa}_n}{\theta^2}$.
4. Discutete qualche proprietà dello stimatore di massima verosimiglianza $\hat{\kappa}_n$.
5. Fornite un intervallo di confidenza a due code per $\kappa(\theta)$ di livello 90% per $n = 10$ e $\hat{\kappa}_{10} = 0.0387$. Quindi deducete un intervallo di confidenza per θ, sempre bilatero e di livello 90%.

Soluzione

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

$$L_\theta(x_1, \dots, x_n) = \frac{\prod_{j=1}^n x_j}{\theta^{2n}} \exp\left\{-\frac{\sum_{j=1}^n x_j^2}{2\theta^2}\right\} \qquad \theta > 0$$

$$\log L_\theta(x_1, \dots, x_n) = -2n \log\theta + \log \prod_{j=1}^n x_j - \frac{\sum_{j=1}^n x_j^2}{2\theta^2}$$

$$\frac{\partial \log L_\theta(x_1, \dots, x_n)}{\partial\theta} = -\frac{2n}{\theta} + \frac{\sum_{j=1}^n x_j^2}{\theta^3} \geq 0 \text{ se e solo se } \theta^2 \leq \frac{\sum_{j=1}^n x_j^2}{2n}$$

Segue che $\hat{\kappa}_n := \frac{\sum_{j=1}^n X_j^2}{2n}$ è MLE per $\kappa(\theta) = \theta^2$

2. Se $y \leq 0$ allora $F_{Y,\theta}(y) = 0$; per $y > 0$:

$$F_{Y,\vartheta}(y) = P_\vartheta(X^2 \leq y) = P_\vartheta\left(X \leq \sqrt{y}\right) = F_{X,\vartheta}(\sqrt{y})$$

da cui segue che la densità di Y è

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

cioè $Y \sim \mathcal{E}(2\theta^2) = \Gamma\left(2/2,\ 2\theta^2\right)$

3. Segue dalle proprietà della distribuzioni gamma che $\dfrac{X_j^2}{\theta^2}$ ha densità gamma $\Gamma\left(2/2,\ 2\right) = \chi_2^2$. Essendo $Q_n = \dfrac{2n\hat{\kappa}_n}{\theta^2} = \dfrac{\sum_{j=1}^n X_j^2}{\theta^2}$ una somma di n va χ_2^2 indipendenti allora $Q_n \sim \chi_{2n}^2$.

4. Usando il punto 3., è facile dimostrare che lo stimatore $\hat{\kappa}$ ha distribuzione gamma $\Gamma\left(n,\ \frac{\theta^2}{n}\right)$, quindi:

a) $\mathrm{E}_\theta(\hat{\kappa}_n) = n \cdot \theta^2/n = \theta^2$: $\hat{\kappa}_n$ è stimatore non distorto di κ;

b) $\dfrac{\partial \log L_\theta(x_1, \dots, x_n)}{\partial\theta} = -\dfrac{2n}{\theta} + \dfrac{\sum_{j=1}^n x_j^2}{\theta^3} = \dfrac{2n}{\theta^3}(\hat{\kappa} - \kappa)$: quest'ultima è condizione necessaria e sufficiente affinché la varianza di $\hat{\kappa}$ raggiunga il confine di Cramer Rao $\dfrac{4\theta^2}{nI(\theta)}$ dove $I(\theta)$ è l'Informazione di Fisher del modello di Raleigh.

c) Ma il confine di Cramer Rao converge a 0 quando $n \to \infty$; quindi, in quanto efficiente, $\hat{\kappa}$ è anche stimatore consistente (in media quadratica) di κ. Infine,

d) $\hat{\kappa}_n$ ha fdr asintoticamente gaussiana di media κ e varianza κ^2/n.

5. Per il punto 3., Q_n è una quantità pivotale e possiamo costruire un intervallo bilatero per κ a code simmetriche di livello 0.90 nel seguente modo: Siano $q_1 = \chi_{2n}^2((1-\gamma)/2) = \chi_{20}^2(0.05) \simeq 10.9$, e $q_2 = \chi_{2n}^2((1+\gamma)/2) = \chi_{20}^2(0.95) \simeq 31.4$. (Abbiamo usato la tabella della fdr χ_{20}^2). Allora $P_\vartheta[q_1 \leq Q_n \leq q_2] = 0.90$. Ma, $\{q_1 \leq Q_n \leq q_2\} = \left\{\frac{2n\hat{\kappa}_n}{q_2} \leq \vartheta^2 \leq \frac{2n\hat{\kappa}_n}{q_1}\right\}$. Allora $P_\vartheta\left(\dfrac{2n\hat{\kappa}_n}{31.4} < \vartheta^2 < \dfrac{2n\hat{\kappa}_n}{10.9}\right) = 0.90$ e, l'intervallo di confidenza di κ di livello 90% è $\left(\dfrac{20 \times 0.0387}{31.4}, \dfrac{20 \times 0.0387}{10.9}\right) \simeq (0.0246, 0.0710)$ e quello di θ è $(0.1570, 0.2665)$. ∎

Esercizio 3.3 Sei settimane fa si è cominciato a quotare il titolo azionario `xxx`. Il suo prezzo iniziale Y_0 era di 1 euro, mentre il suo prezzo Y_k alla fine della k-esima settimana è stato

$$1.0304545,\ 0.9417645,\ 0.8693582,\ 0.8780954,\ 1.040811,\ 1.030454$$

È uso descrivere la dinamica dei prezzi azionari in termini dei rapporti $R_k = \dfrac{Y_k}{Y_{k-1}}$, $k \geq 1$, e un modello molto utilizzato suppone che gli $R_k, k = 1, \ldots$ costituiscano un campione di variabili aleatorie indipendenti e identicamente distribuite con densità "*lognormale*": una variabile aleatoria R è detta *lognormale di parametri* μ *e* σ se $\log R$ è gaussiana di media μ e varianza σ^2.

1. Stimate i parametri μ e σ della distribuzione lognormale dei rapporti fra i prezzi $R_k, k = 1, \ldots, 6$. Fornite una qualche giustificazione statistica delle stime proposte.
2. Determinate la funzione di ripartizione empirica dei logaritmi naturali dei rapporti $R_k, k = 1, \ldots, 6$.
3. Determinate una stima della probabilità che alla fine di una settimana il prezzo cresca rispetto alla settimana precedente.

In realtà, nutriamo qualche dubbio che il modello dei rapporti settimanali dei prezzi azionari `xxx` sia lognormale:

4. Implementate un opportuno test di significatività 5%, per verificare la bontà dell'adattamento dei rapporti settimanali dei prezzi R_k al modello lognormale di parametri $\mu = 0.02$ e $\sigma = 0.073$.

Soluzione Siano $W_k := \log R_k, k = 1, \ldots, 6$. Calcoliamo i sei valori $W_k, k = 1 \ldots, 6$:

$$0.03, -0.09, -0.08, 0.01, 0.17, -0.01$$

1. μ è la media teorica del campione casuale $W_1, \ldots, W_6$. Quindi uno stimatore non distorto, consistente ed efficiente è dato dalla media campionaria il cui valore è 0.005; $\dfrac{\sum_{k=1}^{6} \log R_k}{6} = \log \left(\prod_{k=1}^{6} \dfrac{Y_k}{Y_{k-1}} \right)^{\frac{1}{6}}$ è lo stimatore di massima verosimiglianza di μ. Stimiamo invece σ con la radice quadrata della varianza campionaria dello stesso campione $W_1, \ldots, W_6$ il cui valore è 0.094; oppure usiamo la radice quadrata di $\sum_j (W_j - \overline{W})^2/6 \simeq 0.086$: quest'ultimo valore è la stima di massima verosimiglianza di σ della distribuzione lognormale.

2. Ordiniamo i valori di W_k in ordine crescente: $-0.09, -0.08, -0.01, 0.01, 0.03, 0.17$. Segue che la funzione di ripartizione empirica delle sei osservazioni $w_1, \ldots, w_6$ (tutte distinte) è:

$$\hat{F}_6(w) = \begin{cases} 0 & x < -0.09 \\ \frac{1}{6} \simeq 0.166 & -0.09 \leq w < -0.08 \\ \frac{1}{3} \simeq 0.333 & -0.08 \leq w < -0.01 \\ \frac{1}{2} = 0.5 & -0.01 \leq w < 0.01 \\ \frac{2}{3} = 0.666 & 0.01 \leq w < 0.03 \\ \frac{5}{6} \simeq 0.833 & 0.03 \leq w < 0.17 \\ 1 & x \geq 0.17 \end{cases}$$

3. Ci chiedono di stimare $P(W > 0) = 1 - F_W(0)$' poiché $\hat{F}_6(0) = 0.5$, allora una stima di $P(W > 0)$ basata solo sulla funzione di ripartizione empirica è 0.5. Se il campione dei rapporti è effettivamente lognormale, allora la stima di massima verosimiglianza di $P(W > 0)$ è $1 - \Phi(-0.005/0.086) = \Phi(0.005/0.086) \simeq \Phi(0.058) \simeq 0.5231$.

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 Kolmogorov-Smirnov di livello 5% per verificare: $H_0 : W \sim F_0 = \mathcal{N}(0.02, 0.073^2)$ contro l'alternativa $H_1 : W \not\sim \mathcal{N}(0.02, 0.073^2)$. Pertanto,

$$\begin{cases} F_0(-0.09) = \Phi\left(\frac{-0.09-0.02}{0.073}\right) \simeq \Phi(-1.51) \simeq 0.066 \\ F_0(-0.08) \simeq \Phi(-1.37) \simeq 0.085 \\ F_0(-0.01) \simeq \Phi(-0.41) \simeq 0.341 \\ F_0(0.01) \simeq \Phi(-0.14) \simeq 0.444 \\ F_0(0.03) \simeq \Phi(0.14) \simeq 0.556 \\ F_0(0.17) \simeq \Phi(2.05) \simeq 0.980 \end{cases}$$

Rifiutiamo al livello α se $D_6 := \sup_{x \in \mathbb{R}} |\hat{F}_6(x) - F_0(x)| > q_{D_6}(1 - \alpha)$. Ma $\sup_{x \in \mathbb{R}} |\hat{F}_6(x) - F_0(x)| = \hat{F}_6(0.03) - F_0(0.03) \simeq 0.2777$, e dalle tavole dei quantili della statistica di Kolmogorov-Smirnov con $n = 6$ abbiamo $q_{D_6}(1 - 0.05) = 0.5193$: 0.2777 è minore di 0.5193 e accettiamo H_0. ∎

Esercizio 3.4 Abbiamo eseguito dieci volte il programma `xxx` sul calcolatore `yyy`. I tempi di esecuzione espressi in minuti e riportati nell'ordine di esecuzione sono i seguenti:

81.8 79.9 91.5 67.3 114.9 82.9 63.6 88.4 68.7 81.2

1. Si può concludere sulla base di questi dati che ci sia una sorta di "effetto memoria" per cui i tempi di esecuzione non sono indipendenti? Per rispondere usate un opportuno test con livello di significatività 10%.

Sappiamo che il tempo di esecuzione del programma `xxx` sul calcolatore `yyy` è modellabile come una variabile aleatoria X che ha densità

$$f(x,a,b) = \begin{cases} \frac{2(b-x)}{(b-a)^2} & a \leq x \leq b \\ 0 & \text{altrove} \end{cases} \qquad a, b > 0$$

Quindi: nel migliore dei casi sono sufficienti a minuti per eseguire il programma ma, nel peggiore ne saranno necessari b. Ora, noi non conosciamo né a né b; per questo vi chiediamo di

2. determinare degli stimatori di a e b usando il metodo dei momenti. (*Sugg: usare il fatto che* $\int_a^b 2x^2(b-x)/(b-a)^2\,dx = (b-a)^2/18 + [(2a+b)/3]^2$)

Soluzione

1. Usiamo il test di aleatorietà di Kendall a due code; infatti, a priori non ci aspettiamo né un andamento crescente né uno decrescente. Contiamo il numero di concordanze e discordanze; con i simboli degli appunti abbiamo:

tempi	81.8	79.9	91.5	67.3	114.9	82.9	63.6	88.4	68.7	81.2
C_i	4	5	1	5	0	1	3	0	1	
D_i	5	3	6	1	5	3	0	2	0	

e

$$C = \sum_{i=1}^{9} C_i = 20, \qquad D = \sum_{i=1}^{9} D_i = 25, \qquad T = C - D = -5\,.$$

Per un test di ampiezza α la regione di rifiuto è $\{|T| > q_{\mathrm{Ken};n}(1-\alpha/2)\}$. Noi abbiamo $n = 10$, $\alpha = 0.1$, $|T| = 5$ e, dalle tabelle, $q_{\mathrm{Ken};10}(.95) = 19$: non cadiamo nella regione critica, quindi accettiamo l'ipotesi nulla che i dati provengano da un campione casuale, cioè sulla base di questo campione non dovrebbe esserci "effetto memoria".

2. Calcoliamo i momenti primo e secondo della densità $f(x,a,b)$:

$$\mathrm{E}(X) = \int_a^b \frac{2x\,(b-x)}{(b-a)^2}\,dx = -x\left(\frac{b-x}{b-a}\right)^2\Big|_a^b + \frac{1}{(b-a)^2}\int_a^b (b-x)^2\,dx = \frac{2a+b}{3}$$

$$\mathrm{E}(X^2) = \int_a^b \frac{2x^2\,(b-x)}{(b-a)^2}\,dx = \frac{(b-a)^2}{18} + \left(\frac{2a+b}{3}\right)^2$$

Quindi risolviamo il seguente sistema in a, b:

$$\begin{cases} \frac{2a+b}{3} = M_1 \\ \frac{(b-a)^2}{18} + \left(\frac{2a+b}{3}\right)^2 = M_2 \end{cases} \text{sse} \begin{cases} \frac{2a+b}{3} = M_1 \\ \frac{(b-a)^2}{18} = M_2 - M_1^2 \end{cases} \text{sse} \begin{cases} \frac{2a+b}{3} = M_1 \\ b-a = 3\sqrt{2(M_2-M_1^2)} \end{cases} \text{sse} \begin{cases} \frac{3a+3\sqrt{2(M_2-M_1^2)}}{3} = M_1 \\ b = 3\sqrt{2(M_2-M_1^2)} + a \end{cases}$$

dove M_1 è la media campionaria e M_2 il momento secondo campionario. Segue che $\hat{a} = M_1 - \sqrt{2(M_2-M_1^2)}$ e $\hat{b} = M_1 + 2\sqrt{2(M_2-M_1^2)}$ sono gli stimatori dei momenti rispettivamente di a e b. Per quanto concerce i valori delle stime abbiamo: $M_1 = 82.02$, $M_2 = 6922.386$, $\sqrt{2(M_2-M_1^2)} = \sqrt{2(195.1056)} = 19.75376$ e quindi; $\hat{a} = 62.26624$ e $\hat{b} = 121.5275$. ∎

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

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

Esercizio 4.1 Vi chiedo di aiutarmi a decidere se un'unica osservazione assolutamente continua X abbia densità

$$f_a(x) = 2x\mathrm{e}^{-x^2}\mathbf{1}_{(0,\infty)}(x) \quad \text{oppure} \quad f_b(x) = 20x\mathrm{e}^{-10x^2}\mathbf{1}_{(0,\infty)}(x)$$

A tal fine dovete:

1. calcolare la funzione di ripartizione della densità f_a e quella della densità f_b, in un generico punto $x > 0$;
2. costruire il test uniformemente più potente di livello $\alpha \in (0,1)$ per verificare l'ipotesi nulla $H_0 : f = f_a$ contro l'ipotesi alternativa $H_1 : f = f_b$;
3. calcolare il p-value del test del punto 2. quando $X = 0.15$ e il p-value del test quando $X = 0.49$. Se $\alpha = 5\%$, cosa dovrò decidere quando $X = 0.15$? E quando $X = 0.49$?

Non contenta, vi chiedo ora di scambiare le ipotesi e verificare l'ipotesi nulla $H_0 : f = f_b$ contro l'ipotesi alternativa $H_1 : f = f_a$.

4. Se per verificare l'ipotesi nulla $H_0 : f = f_b$ contro l'ipotesi alternativa $H_1 : f = f_a$ usiamo un test di regione critica $\{X \geq k_2\}$, per quale valore di k_2 questo test ha ampiezza $\alpha \in (0,1)$?
5. Calcolate la potenza del test costruito al punto 4., cioè quello per verificare $H_0 : f = f_b$ contro l'ipotesi alternativa $H_1 : f = f_a$, quando $\alpha = 5\%$.
6. Se $\alpha = 5\%$ e $X = 0.49$, in base al secondo test del punto 4. accettate o rifiutate f_b? Più in generale, se $\alpha = 5\%$, per quali valori dell'unica osservazione X i test di ipotesi dei punti 2. e 4. portano alla stessa decisione?

Soluzione

1. Se $x > 0$, allora

$$F_a(x) = \int_0^x f_a(t)\,dt = \int_0^x 2t\mathrm{e}^{-t^2}\,dt = 1 - \mathrm{e}^{-x^2}$$

e

$$F_b(x) = \int_0^x f_b(t)\,dt = \int_0^x 20t\mathrm{e}^{-10t^2}\,dt = 1 - \mathrm{e}^{-10x^2}$$

2. Le ipotesi nulla $H_0 : f = f_a$ e alternativa $H_1 : f = f_b$ sono entrambe semplici. Dal Lemma di Neyman Pearson il test uniformemente più potente di livello α ha regione critica

$$\mathcal{G} = \left\{x > 0 : \frac{f_a(x)}{f_b(x)} \leq \delta\right\} = \left\{x > 0 : \frac{\mathrm{e}^{9x^2}}{10} \leq \delta\right\} = \{x : \ 0 < x \leq k_1\}$$

con k_1 tale che $P_a(X \leq k_1) = F_a(k_1) = \alpha$. Segue dal punto 1. che $F_a(k_1) = 1 - \mathrm{e}^{-k_1^2} = \alpha$ sse $k_1 = \sqrt{-\ln(1-\alpha)}$.

3. Il p-value del test è $P_a(X \leq$ "valore osservato di X"); se $X = 0.15$, il p-value è $P_a(X \leq 0.15) = F_a(0.15) = 1 - \mathrm{e}^{-0.15^2} \simeq 0.0222$; se invece $X = 0.49$, il p-value è $P_a(X \leq 0.49) = 1 - \mathrm{e}^{-0.49^2} \simeq 0.2135$. Essendo $0.0222 < 0.05 < 0.2135$, allora se $X = 0.15$ rifiuto f_a, se invece $X = 0.49$ rifiuto f_b.

4. Il test per verificare l'ipotesi nulla $H_0 : f = f_b$ contro l'ipotesi alternativa $H_1 : f = f_a$ di regione critica $\{X \geq k_2\}$ ha ampiezza $\alpha \in (0,1)$ se e solo se $P_b(X \geq k_2) = \alpha$; quindi k_2 deve essere tale che $1 - F_b(k_2) = \alpha$, da cui $k_2 = \sqrt{-(\ln\alpha)/10}$.

5. La potenza del test è data da

$$P_a\left(X \geq \sqrt{\frac{-\ln\alpha}{10}}\right) = 1 - F_a\left(\sqrt{\frac{-\ln\alpha}{10}}\right) = \alpha^{1/10} = 0.05^{0.1} \simeq 0.7411$$

6. Se $\alpha = 0.05$ allora $k_1 = \sqrt{-ln(1-0.05)} \simeq 0.2265$ e $k_2 = \sqrt{-(\ln 0.05)/10} \simeq 0.5473$; Allora, confrontando i due test, abbiamo che:
se $0 < x \leq 0.2265$ i due test concordano e portano entrambi a rifiutare f_a;
se $0.2265 < x < 0.5473$ i due test discordano: con quello al punto 2. rifiuto f_b e con quello al punto 4. la accetto;
infine, anche per $x \geq 0.5473$ i due test concordano ma portano a rifiutare f_b. ∎

Esercizio 4.2 Sia $X_1, \ldots, X_n$ un campione casuale estratto dalla funzione di densità discreta

$$f(x, \theta) = \frac{(\ln \theta)^x}{\theta(x!)} \mathbf{1}_{\{0,1,2,\ldots\}}(x), \quad \theta > 1$$

dove $\theta > 1$ è un parametro incognito e "ln" indica il logaritmo in base naturale.

1. Determinate lo stimatore $\widehat{\theta}$ di θ usando il metodo di massima verosimiglianza.
2. Calcolate l'informazione di Fisher $I(\theta)$ del modello statistico $\{f(x, \theta), \theta > 1\}$ e determinate lo stimatore di massima verosimiglianza di $I(\theta)$.
3. Qual è la funzione di ripartizione asintotica di $\widehat{\theta}$? (Fornite non solo la forma della fdr approssimata, ma determinatene anche i parametri esplicitamente).
4. Costruite un intervallo di confidenza a due code asintotico per θ di livello γ, usando i punti 1 e 3 precedenti. Quindi determinate numericamente l'intervallo quando $\gamma = 0.95$, $n = 100$ e $\sum_{j=1}^{100} x_j = 150.0$.
5. Siano $n = 100$ e $\sum_{j=1}^{100} x_j = 150.0$. Usate lo stimatore di massima verosimiglianza di θ per verificare asintoticamente l'ipotesi nulla $H_0 : \theta = 4$ contro l'ipotesi alternativa $H_1 : \theta \neq 4$, a un livello $\alpha = 5\%$.

SOLUZIONE

1. Riconosciamo nell'espressione di $f(x, \theta)$ la densità di Poisson di media $\ln \theta$; lo stimatore di massima verosimiglianza della media $\ln \theta$ è la media campionaria $\bar{X}$, da cui segue che lo stimatore di massima verosimiglianza di θ è dato da $\widehat{\theta} = \mathrm{e}^{\bar{X}}$.
2. Poiché
$$\frac{\partial}{\partial \theta} \ln(f(x, \theta)) = \frac{\partial}{\partial \theta} (-\ln \theta + x \ln \ln \theta - \ln(x!)) = -\frac{1}{\theta} + \frac{x}{\theta \ln \theta} = \frac{1}{\theta \ln \theta} (x - \ln \theta)$$
allora l'informazione di Fisher è
$$I(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial \theta} f(X_1, \theta)\right)^2\right] = \mathrm{E}\left[\left(\frac{1}{\theta \ln \theta}(X_1 - \ln \theta)\right)^2\right] = \frac{\mathrm{Var}_\theta(X_1)}{\theta^2 (\ln \theta)^2} = \frac{\ln \theta}{\theta^2 (\ln \theta)^2} = \frac{1}{\theta^2 \ln \theta}$$
Lo stimatore di massima verosimiglianza di $I(\theta)$ è $\widehat{I} = I\left(\widehat{\theta}\right) = \bar{X}^{-1} \mathrm{e}^{-2\bar{X}}$.
3. Segue dalle proprietà asintotiche degli stimatori di massima verosimiglianza che $\widehat{\theta} = \mathrm{e}^{\bar{X}}$ è asintoticamente non distorto, efficiente e gaussiano. Quindi, asintoticamente la media di $\widehat{\theta}$ è θ. Inoltre, la varianza è fornita dal confine inferiore di Cramer Rao: $1/(nI(\theta)) = \theta^2 \ln \theta / n$.
4. Abbiamo stabilito al punto 3. che asintoticamente $\sqrt{nI(\theta)}(\widehat{\theta} - \theta) \sim N(0, 1)$. Inoltre, per il punto 2., stimiamo la varianza $1/(nI(\theta))$ con $1/(n\hat{I}) = \bar{X}\mathrm{e}^{2\bar{X}}/n$. Possiamo procedere costruendo un intervallo di confidenza a due code simmetrico asintotico per la media di una popolazione gaussiana con varianza incognita e stimata con $1/(n\hat{I})$:
$$\hat{\theta} \pm z_{\frac{1+\gamma}{2}} / \sqrt{n\hat{I}} = \mathrm{e}^{\bar{x}} \pm z_{\frac{1+\gamma}{2}} \sqrt{\frac{\bar{x}\mathrm{e}^{2\bar{x}}}{n}}$$
Se $\gamma = 0.95$, $n = 100$ e $\sum_{j=1}^{100} X_j = 150.0$, allora un intervallo di confidenza asintotico per θ è $\mathrm{e}^{1.5} \pm 1.96 \times \sqrt{1.5}\mathrm{e}^{1.5}/10 \simeq (3.406, 5.558)$.
5. Possiamo usare l'intervallo di confidenza costruito al punto precedente: poiché $4 \in (3.406, 5.558)$, allora accettiamo l'potesi nulla $H_0 : \theta = 4$. Ciò corrisponde a costruire una regione critica basata sulla statistica test $\sqrt{n\hat{I}}(\widehat{\theta} - \theta_0)$ con $\theta_0 = 4$ che sotto H_0 ha fdr asintotica $N(0, 1)$.

   Invece di usare la statistica test $\sqrt{100\hat{I}}(\widehat{\theta} - \theta_0)$ potremmo usare $\sqrt{100I(4)}(\widehat{\theta} - 4)$, che è anche essa asintoticamente $N(0, 1)$, ma, a parità di numero di osservazioni (100), l'errore di approssimazione è più piccolo. Se usiamo la statistica $\sqrt{100I(4)}(\widehat{\theta} - 4)$, allora rifiutiamo $H_0 = 4$ a favore di $H_1 \neq 4$ se $|\hat{\theta} - 4| > 1.96 \times \sqrt{16 \ln 4/100} \simeq 0.923$, cioè se $\hat{\theta} \notin [3.077, 4.923]$. Comunque anche con la nuova statistica test arriviamo alla stessa conclusione di accettare $H_0 : \theta = 4$, dal momento che $\widehat{\theta} = \mathrm{e}^{1.5} \simeq 4.482$. ∎

Esercizio 4.3 Avete raccolto 136 misurazioni ripetute di una grandezza. Ma avete salvato solo le seguenti informazioni: esattamente 71 misurazioni sono minori o uguali della mediana della funzione di ripartizione gaussiana di media 20 e varianza 16 –da questo momento indicata come fdr $\mathcal{N}(20,16)$– e di queste 32 sono minori o uguali del quantile di ordine 0.25 della fdr $\mathcal{N}(20,16)$; poi, 31 misurazioni sono maggiori del quantile di ordine 0.75 della fdr $\mathcal{N}(20,16)$. Infine, 34 misurazioni hanno valore compreso fra la mediana e il quantile di ordine 0.75 della fdr $\mathcal{N}(20,16)$.

1. Esprimete la fdr $\mathcal{N}(\mu,\sigma^2)$ in termini della fdr $\mathcal{N}(0,1)$; esprimete il quantile di ordine a della fdr $\mathcal{N}(\mu,\sigma^2)$ in termini del quantile di ordine a della fdr $\mathcal{N}(0,1)$.
2. Ritenete che le vostre misurazioni provengano da un modello $\mathcal{N}(20,16)$? Per rispondere alla domanda costruite un opportuno test di livello $\alpha = 5\%$.
3. Calcolate un valore approssimato del p-value o, almeno, stabilite un intervallo dove tale p-value cade.

Soluzione

1. Sia $X \sim \mathcal{N}(\mu,\sigma^2)$ e $Z \sim \mathcal{N}(0,1)$. Allora $F_X(x) = P(X \le x) = P(Z \le (x-\mu)/\sigma) = \Phi((x-\mu)/\sigma)$, da cui la seguente equazione che lega i quantili q_a della fdr $\mathcal{N}(\mu,\sigma)$ e z_a della fdr gaussiana standard Φ:

$$F_X(q_a) = \Phi\left(\frac{q_a-\mu}{\sigma}\right)$$

In particolare, per ogni $a \in (0,1)$ il quantile di ordine a della fdr $\mathcal{N}(\mu,\sigma^2)$ è dato da $q_a = \sigma z_a + \mu$.

2. La fdr $\mathcal{N}(20,16)$ è simmetrica nella media, quindi, la media coincide con la mediana e la mediana è 20. Per quanto concerne invece i quantili di ordine 0.25 e 0.75 della fdr $\mathcal{N}(20,16)$, applicando il punto precedente abbiamo

$$q_{0.25} = 4z_{0.25} + 20 = 4 \times (-z_{1-0.25}) + 20 = -4z_{0.75} + 20 \simeq -4 \times 0.6745 + 20 = 17.302$$

e

$$q_{0.75} = 4z_{0.75} + 20 \simeq 4 \times 0.6745 + 20 = 22.698$$

È utile sintetizzare i dati raggruppati a nostra disposizione nella seguente tabella

classe:	$(-\infty, 17.302]$	$(17.302, 20]$	$(20, 22.698]$	$(22.698, +\infty)$
numerosità (N_j):	32	$71 - 32 = 39$	34	31
probabilità teorica:	0.25	0.25	0.25	0.25

Per rispondere al quesito impostiamo un test asintotico χ^2 di adattamento. Infatti $136 \times 0.25 = 34 > 5$ e l'approssimazione asintotica χ^2 con $4-1=3$ gradi di libertà per la statistica di Pearson $Q_{136} = \sum_{j=1}^{4}(N_j - 136 \times 0.25)^2/(136 \times 0.25)$ funziona. Il valore della statistica Q_{136} è

$$Q_{136} = \sum_{j=1}^{4} \frac{(N_j - 136 \times 0.25)^2}{136 \times 0.25} = \frac{\sum_{j=1}^{4} N_j^2}{136 \times 0.25} - 136 \simeq 1.1176$$

A livello $\alpha = 5\%$ rifiutiamo l'ipotesi nulla H_0: *"il modello è $\mathcal{N}(20,16)$"* se $Q_{136} > \chi^2_3(1-\alpha) = \chi^2_3(0.95) \simeq 7.81$. Essendo $1.1176 < 7.81$, accettiamo l'ipotesi nulla che le misurazioni seguano un modello $\mathcal{N}(20,16)$.

3. Il p-value, o livello descrittivo del test, è dato da p-value $= 1 - F_{\chi^2_3}(Q_{136}) \simeq 1 - F_{\chi^2_3}(1.1176) \in (1 - F_{\chi^2_3}(1.21), 1 - F_{\chi^2_3}(1.005)) \simeq (1-0.25, 1-0.2) = (0.75, 0.80)$: concludiamo che c'è una fortissima evidenza empirica ad accettare l'ipotesi nulla. (Calcolando il p-value con R otteniamo p-value= 0.7728158.) ∎

Esercizio 4.4 Abbiamo estratto a caso un campione di sette coppie sposate e abbiamo chiesto a ciascun marito e a ciascuna moglie *a*) quanto ha speso in euro per la spesa l'ultima volta che vi si è recato da solo (sola) e *b*) di esprimere con un numero compreso fra 0 e 1 la probabilità che l'altra(o) demandi a lui (lei) l'incombenza di fare la spesa. Le risposte sono state le seguenti:

coppia	1	2	3	4	5	6	7
marito	$(25.1, 0.356)$	$(21.5, 0.035)$	$(38.0, 0.506)$	$(64.3, 0.614)$	$(52.0, 0.330)$	$(16.1, 0.171)$	$(26.1, 0.447)$
moglie	$(16.0, 0.701)$	$(42.2, 0.903)$	$(56.5, 0.315)$	$(41.1, 0.220)$	$(19.0, 0.998)$	$(26.2, 0.355)$	$(24.5, 0.622)$

(dove, per esempio, per la prima coppia, alla riga dei mariti il dato accoppiato $(25.1, 0.356)$ indica che l'ultima volta che ha fatto la spesa da solo il marito della prima coppia ha speso 25.1 euro e che lui reputa che la probabilità che la moglie gravi lui dell'incombenza della spesa sia 0.356.[1])

1. Costruite un test di livello $\alpha = 5\%$ per verificare se i mariti siano oculati almeno quanto le mogli quando fanno la spesa contro l'alternativa che le mogli siano più oculate dei mariti, sotto ipotesi di gassianità delle spese di mogli e mariti.

2. Costruite un test di livello $\alpha = 1\%$ per verificare se sia più frequente che i mariti demandino alle mogli l'onere di fare la spesa (almeno nella percezione delle mogli), contro l'alternativa che sia più frequente che le mogli demandino questa incombenza ai mariti (almeno nella percezione dei mariti).

3. `[Allievi Prof Barchielli]` Guardate ora solo i dati sui mariti. Sulla base di questi dati, secondo voi i mariti che sarebbero incaricati più frequentemente dalle mogli di fare la spesa, spendono meno di quelli meno frequentemente incaricati? Rispondete costruendo un opportuno test di livello 10%.

Soluzione

Introduciamo le variabili aleatorie X_1, X_2, Y_1 e Y_2 definite da:
X_1 = spesa in euro fatta dal marito (non accompagnato dalla moglie),
X_2 = spesa in euro fatta dalla moglie (non accompagnata dal marito),
Y_1 = possibilità come percepita dal marito che la moglie demandi a lui l'onere della spesa e
Y_2 = possibilità come percepita dalla moglie che il marito demandi a lei l'onere della spesa.

1. Per rispondere usiamo il campione dei 7 dati accoppiati $(x_{1,j}, x_{2,j})$, $j = 1, ..., 7$, provenienti da una popolazione bivariata gaussiana e costruiamo un test t sulla differenza delle medie di popolazioni gaussiane. Ricordando che l'indice 1 si riferisce ai mariti e l'indice 2 alle mogli, dobbiamo verificare $H_0 : \mu_1 - \mu_2 \leq 0$ vs $H_1 : \mu_1 - \mu_2 > 0$ con μ_1 che indica la spesa media dei mariti e μ_2 quella delle mogli. Rifiutiamo H_0 a livello α se $T = \sqrt{7}(\bar{X}_1 - \bar{X}_2)/S_D > t_6(1-\alpha)$, dove $t_6(1-\alpha)$ è il quantile di ordine $1-\alpha$ della fdr t di Student con 6 gradi di libertà e S_D^2 è la varianza campionaria del campione delle differenze $(x_{1,j} - x_{2,j})$; abbiamo

$$\bar{X}_1 - \bar{X}_2 \simeq 34.728 - 32.214 = 2.514$$

$$\sum_{j=1}^{7}(X_{1,j} - X_{2,j})^2 = 2585.36$$

$$S_D^2 = \frac{\sum_{j=1}^{7}(X_{1,j} - X_{2,j})^2}{6} - \frac{7}{6}(\bar{X}_1 - \bar{X}_2)^2 = \frac{2585.36}{6} - \frac{7}{6} \times 2.514^2 \simeq 423.52$$

$$T = \sqrt{7}\frac{\bar{X}_1 - \bar{X}_2}{S_D} = \frac{\sqrt{7} \times 2.514}{\sqrt{423.52}} \simeq 0.3232$$

Se $\alpha = 0.05$ allora $t_6(1-\alpha) = 1.943$. Quindi al livello 5% accettiamo H_0 e propendiamo a credere che i mariti siano oculati almeno quanto le mogli nel fare la spesa.

2. Per rispondere usiamo il campione dei 7 dati accoppiati $(y_{1,j}, y_{2,j})$, $j = 1, ..., 7$. Non abbiamo elementi per ritenere che i dati siano stati generati da un modello gaussiano (anzi, i dati fra 0 e 1 sembrerebbero "uniformemente" distribuiti e non concentrati intorno a un polo di riferimento); d'altro canto, sono troppo pochi per implementare un test asintotico sulla media della differenza $Y_1 - Y_2$. Pertanto, impostiamo il test dei segni di Wilcoxon per dati accoppiati che verifichi l'ipotesi nulla H_0: "Y_1 *tende ad essere più piccola di* Y_2"

[1] Attenzione alla lettura dei dati sulle probabilità di essere demandato a fare la spesa: per ogni coppia avete due numeri compresi fra 0 e 1. Questi due numeri sono valutazioni soggettive di probabilità fatte da due persone diverse (marito e moglie); quindi, per qualche coppia potrebbero sommare a un numero maggiore di 1, oppure potrebbero essere entrambe "piccole", perché per esempio quella coppia usa andare a far la spesa spesso insieme.

contro l'alternativa H_1: "Y_1 *tende ad essere più grande di* Y_2". Calcoliamo la statistica T^+ = numero di coppie in cui Y_1 è maggiore di Y_2; essa vale 2. Il test dei segni di Wilcoxon per dati accoppiati detta di rifiutare H_0 a livello α se $T^+ > q_{1-\alpha}$ dove $q_{1-\alpha}$ è il quantile di ordine $1-\alpha$ della fdr $\mathrm{Bin}(7, 1/2)$. Ovvero, rifiuteremo H_0 per ogni α tale che

$$\alpha > P\left(\mathrm{Bin}\left(7, \frac{1}{2}\right) > 2\right) = 1 - \frac{1}{2^7}\left(\binom{7}{0} + \binom{7}{1} + \binom{7}{2}\right) = 1 - \frac{29}{2^7} = 1 - 0.2265625 = 0.7734475$$

cioè il p-value del test è 77.34%. Essendo 1% minore del p-value del test, a livello 1% non possiamo rifiutare l'ipotesi nulla; dunque non rifiutiamo l'ipotesi che sia più frequente che le mogli siano gravate dell'incombenza della spesa.

3. Consideriamo i dati accoppiati $(X_{1j}, Y_{1,j})$ e impostiamo il test di concordanza di Kendall di livello $\alpha = 10\%$ per il seguente problema:

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

Per eseguire il test è necessario calcolare il numero di concordanze e discordanze. A questo scopo riordiniamo le coppie per valori della variabile X_1 crescenti:

X_1 (spesa in euro)	16.1	21.5	25.1	26.1	38.0	52.0	64.3
Y_1 (probabibilità)	0.171	0.035	0.356	0.447	0.506	0.330	0.614
$segno[y_{[1,j]} - y_{[1,1]}]_{j>1}$		-1	$+1$	$+1$	$+1$	$+1$	$+1$
$segno[y_{[1,j]} - y_{[1,2]}]_{j>2}$			$+1$	$+1$	$+1$	$+1$	$+1$
$segno[y_{[1,j]} - y_{[1,3]}]_{j>3}$				$+1$	$+1$	-1	$+1$
$segno[y_{[1,j]} - y_{[1,4]}]_{j>4}$					$+1$	-1	$+1$
$segno[y_{[1,j]} - y_{[1,5]}]_{j>5}$						-1	$+1$
$segno[y_{[1,j]} - y_{[1,6]}]_{j>6}$							$+1$

Dalla tabella risulta $C = 17$ e $D = 4$, dunque $C - D = 13$. La regione critica del test di livello α è costituita dai campioni per cui $C - D > q_{1-\alpha}(C - D)$, dove $q_{1-\alpha}(C - D)$ rappresenta il quantile di ordine $1 - \alpha$ della statistica $C - D$ per $n = 7$. Per $\alpha = 0.1$ e $n = 7$: $q(0.90) = 9$. Avendo osservato $C - D = 13 > 9$, a livello $\alpha = 10\%$ propendiamo per l'ipotesi alternativa che quanto più i mariti ritengono sia probabile essere mandati a fare la spesa, tanto più spendono. ∎

5 STATISTICA per ING INF e TEL [2L] Proff. A. Barchielli, I. Epifani 08.03.06

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

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

Esercizio 5.1 Sia

$$10.42,\ 6.69,\ 11.18,\ 7.94,\ 9.65,\ 8.38,\ 11.80,\ 8.28,\ 10.84,\ 10.72$$

una realizzazione del campione casuale $X_1, \ldots, X_{10}$ estratto dalla densità gaussiana di varianza 2.25 e media incognita μ.

1. Determinate un intervallo di confidenza (IC) simmetrico di livello 90% per μ.

Siete disposti ad allargare il vostro campione per fare inferenza su μ.

2. Qual è il numero minimo di dati neccessari per essere fiduciosi almeno al 90% che l'errore che commettete stimando μ con la media campionaria non sia più grande di 0.26?

Sia n^* la risposta al punto 2.. Sulla base di quanto discusso al punto 2., decidete di raccogliere ulteriori $n^* - 10$ osservazioni $X_{11}, \ldots, X_{n^*}$. Sapete che la media campionaria delle nuove $n^* - 10$ osservazioni vale 8.96.

3. Aggiornate la stima puntuale di μ usando tutte le informazioni provenienti dal campione allargato $X_1, \ldots, X_{n^*}$; quindi determinate numericamente un IC simmetrico di livello 90% per μ.

Vogliamo ora studiare il problema di ipotesi $H_0 : \mu = 10$ *versus* $H_1 : \mu \neq 10$ e per questo estraiamo un campione di ampiezza $n = 20$ (totalmente nuovo) sempre dalla famiglia gaussiana $\mathcal{N}(\mu, 2.25)$. In corrispondenza di questo nuovo campione, $(9.307, 10.410)$ è l'IC simmetrico di μ di livello 0.95.

4. Sapere che $(9.307, 10.410)$ è l'IC simmetrico di μ di livello 0.95 basato su $n = 20$ osservazioni è sufficiente per scegliere quale ipotesi rifiutare a livello di significatività $\alpha = 1\%$? Se si, quale decisione prendete?

SOLUZIONE

1. Un IC simmetrico (nella media campionaria) di livello $\gamma 100\%$ della media per popolazione gaussiana con varianza nota è $IC(\mu) = \bar{X} \pm z_{\frac{1+\gamma}{2}} \sigma/\sqrt{n}$. Con i dati a nostra disposizione: $\bar{x}_{10} = 9.59$, $\sigma^2 = 2.25$, $n = 10$, $z_{\frac{1+\gamma}{2}} = z_{0.95} \simeq 1.645$ e quindi otteniamo $IC(\mu) = (8.81, 10.37)$.

2. Stiamo cercando il minimo numero di dati n tale che $P_\mu(|\bar{X}_n - \mu| \leq \epsilon) \geq \gamma$ con $\epsilon = 0.26$ e $\gamma = 0.9$. Ma

$$P_\mu(|\bar{X}_n - \mu| \leq \epsilon) \geq \gamma \iff 2\Phi\left(\frac{\sqrt{n}\epsilon}{\sigma}\right) - 1 \geq \gamma \iff n \geq \frac{\sigma^2 z^2_{\frac{1+\gamma}{2}}}{\epsilon^2}$$

   Con $\epsilon = 0.26, \sigma^2 = 2.25$ e $\gamma = 0.9$ (da cui $z_{\frac{1+\gamma}{2}} \simeq 1.645$), abbiamo $\sigma^2 z^2_{\frac{1+\gamma}{2}}/\epsilon^2 = 90.05$ e quindi $n^* = 91$.

3. $\bar{X}_{91} = (10\bar{X}_{10} + 81\bar{X}_{81})/91 = (10 \times 9.59 + 81 \times 8.96)/91 \simeq 9.03$;
   Poiché $P_\mu(|\bar{X}_{91} - \mu| \leq 0.26) \geq 0.90$, allora $IC(\mu) = (9.03 - 0.26, 9.03 + 0.26) = (8.77, 9.29)$.

4. Per ogni $\alpha < 5\%$ abbiamo $1 - \alpha > 95\%$ e quindi l'IC(μ) simmetrico nella media campionaria $\bar{X}$, di forma $(\bar{X} - \epsilon, \bar{X} + \epsilon)$ e di livello $(1-\alpha)100\%$, contiene $(9.307, 10.410)$. Essendo che $10 \in (9.307, 10.410)$, *a fortiori* 10 appartiene a ogni $IC(\mu)$ simmetrico di livello $(1-\alpha)100\% > 95\%$. Riassumendo: se $(9.307, 10.410)$ è un IC simmetrico di livello 0.95, allora accettiamo l'ipotesi nulla $H_0 : \mu = \mu_0$ contro l'alternativa $H_1 : \mu \neq \mu_0$ $\forall\ \alpha \leq 5\%, \forall \mu_0 \in (9.307, 10.410)$.

   Osservate che se la domanda fosse stata riferita a un valore di $\alpha > 5\%$, allora avremmo avuto bisogno di qualche ulteriore considerazione. Usando l'informazione che $(9.307, 10.410)$ è di forma $(\bar{X} - \epsilon, \bar{X} + \epsilon)$, deduciamo che la media campionaria è il valore centrale $\bar{x} = (9.307+10.410)/2 = 9.8585$ e quindi il p-value del test basato sulla statistica $\bar{X}$ è dato da

$$2\left[1 - \Phi\left(\frac{|9.8585 - 10|}{\sqrt{2.25/20}}\right)\right] \simeq 2(1 - \Phi(0.422)) \simeq 0.6745.$$

   Vi è una forte evidenza ad accettare H_0. ∎

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

$$f(x,\beta) = \frac{1}{2\beta} \exp\left(-\frac{|x|}{\beta}\right), \qquad \beta > 0$$

1. Determinate densità e media di $Y_i = |X_i|$, $i = 1, \dots .n$.
2. Determinate lo stimatore $\hat{\beta}$ di massima verosimiglianza di β e verificate che è non distorto.
3. Determinate k tale che $P_\beta\left(\frac{\hat{\beta}}{\beta} \geq k\right) = \gamma$, in funzione del numero di osservazioni n e del quantile di ordine $1-\gamma$ della funzione di ripartizione chiquadrato con $2n$ gradi di libertà.
4. Determinate un intervallo di confidenza a una coda inferiore per β di livello 90%, in corrispondenza della realizzazione campionaria $-1.095,\ 0.836,\ -0.256,\ 1.996,\ 4.675,\ 0.971$.
5. Proponete un test (sensato) di ampiezza 10% basato sullo stimatore $\hat{\beta}$ per verificare l'ipotesi nulla H_0 : $\beta \geq 4$ contro l'ipotesi alternativa $H_1 : \beta < 4$.
6. Sia $n = 6$. Calcolate un valore approssimato della potenza del test del punto 5. in $\beta = 1$, o, almeno, stabilite un intervallo dove tale potenza cade.

Soluzione

1. Se $y \leq 0$ allora $F_{Y,\beta}(y) = 0$; per $y > 0$:

$$F_{Y,\beta}(y) = P_\beta(|X| \leq y) = P_\beta\left(-y \leq X \leq y\right) = 2F_{X,\beta}(y) - 1$$

da cui segue che la densità di Y è

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

cioè $Y_1, \dots, Y_n$ sono variabili aleatorie i.i.d. $\sim \mathcal{E}(\beta) = \Gamma(2/2,\ \beta)$. Segue che $\mathrm{E}(Y_i) = \beta$.

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

$$L_\beta(x_1, \dots, x_n) = \frac{1}{(2\beta)^n} \exp\left\{-\frac{\sum_{j=1}^n |x_j|}{\beta}\right\}, \qquad \beta > 0$$

$$\ln L_\beta(x_1, \dots, x_n) = -n\ln 2 - n\ln\beta - \frac{\sum_{j=1}^n |x_j|}{\beta}$$

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

Segue che $\hat{\beta}_n := \frac{\sum_{j=1}^n |X_j|}{n} = \bar{Y}$ è MLE per β. Poiché $\mathrm{E}_\beta(\bar{Y}) = \mathrm{E}_\beta(Y_1) = \beta$, allora $\hat{\beta}$ è stimatore non distorto di β.

3. Segue dalle proprietà della famiglia delle densità gamma che $\hat{\beta}/\beta \sim \Gamma(2n/2,\ \beta/(n\beta))$ da cui otteniamo: $2n\hat{\beta}/\beta \sim \Gamma(2n/2,\ 2) = \chi^2_{2n}$. Allora

$$P_\beta\left(\frac{\hat{\beta}}{\beta} \geq k\right) = \gamma \iff P_\beta\left(\frac{2n\hat{\beta}}{\beta} \geq 2nk\right) = \gamma \iff F_{\chi^2_{2n}}(2nk) = 1-\gamma \iff k = \frac{\chi^2_{2n}(1-\gamma)}{2n}$$

4. La stima ML di β in corrispondenza del campione di 6 osservazioni assegnato è $\hat{\beta} = 9.829/6 \simeq 1.638$ e, per quanto dimostrato al punto precedente, vale che $P_\beta\left(\beta \leq \frac{12\hat{\beta}}{\chi^2_{12}(0.10)}\right) = 0.90$. Dalle tavole otteniamo che $\chi^2_{12}(0.10) = 6.304$. Quindi un IC a una coda inferiore di β di livello 90%, in corrispondenza della realizzazione campionaria assegnata, è $(0, 2 \times 9.829/6.304) \simeq (0, 3.118)$.

5. Basiamoci sulla statistica test $\hat{\beta}$, un test sensato potrebbe essere quello di rifiutare $H_0 : \beta \geq 4$ per accettare $H_1 : \beta < 4$, se $\hat{\beta}/4$ è "piccolo", cioè $\hat{\beta}/4 < k$, quantificando k in termini della ampiezza del test $\alpha = 10\%$:

$$10\% = \sup_{\beta \geq 4} P_\beta \left(\frac{\hat{\beta}}{4} < k \right) = \sup_{\beta \geq 4} P_\beta \left(\frac{\hat{\beta}}{4} \leq k \right) =$$
$$= \sup_{\beta \geq 4} P_\beta \left(\frac{2n\hat{\beta}}{\beta} \leq \frac{2n \times 4 \times k}{\beta} \right) = \sup_{\beta \geq 4} F_{\chi^2_{2n}} \left(\frac{2n \times 4 \times k}{\beta} \right) = F_{\chi^2_{2n}} \left(\frac{2n \times 4 \times k}{4} \right)$$

e quindi $k = \chi^2_{2n}(1-\gamma)/(2n) \simeq 6.304/12 = 0.525$.

6. La potenza del test nel punto $\beta = 1$ è

$$\pi(1) = P_1 \left(\frac{\hat{\beta}}{4} \leq 0.525 \right) = P_1 \left(\frac{12\hat{\beta}}{1} \leq 12 \times 4 \times 0.525 \right) = F_{\chi^2_{12}}(25.2)$$

D'altro canto, $F_{\chi^2_{12}}(25.2) \in (F_{\chi^2_{12}}(23.337), F_{\chi^2_{12}}(26.217)) = (0.975, 0.99)$. ▮

Esercizio 5.3 Nella Repubblica di `AILATI` oggi è giorno di elezioni. I partiti in lizza sono tre: *Beige*, *Crema* ed *Ecrù*. Vince chi prende più voti. Fuori da un seggio si sta svolgendo uno strano exit-poll: all'uscita a un campione di votanti uomini scelti a caso si chiede per chi hanno votato e vengono pesati. Le informazioni raccolte sono le seguenti: 246 intervistati hanno votato *Beige*, 200 *Crema* e i rimanenti *Ecrù*. Invece, per quanto riguarda il peso: 120 pesano meno di 72.0 Kg, 133 pesano più 90.0 Kg, 266 hanno poi un peso compreso fra 72.0 Kg e 79.0 Kg e 301 fra 79.0 Kg e 90.0 Kg. Inoltre, fra gli intervistati più magri di 72.0 Kg, 37 hanno votato *Beige* e 21 *Crema*; invece fra quelli di peso compreso fra 72.0 Kg e 79.0 Kg, 73 hanno votato *Beige* e 129 *Ecrù*. Infine, fra i votanti più grossi di 90.0 Kg, 32 hanno votato *Crema* e 63 *Ecrù*.

1. Stabilite se la caratteristica peso influenzi le preferenze politiche degli elettori.

2. Costruite un intervallo di confidenza bilatero asintotico di livello $\gamma = 0.95$ della percentuale di elettori del partito *Ecrù* più grassi di 79.0 Kg.

3. Verificate ad un livello $\alpha = 5\%$ l'ipotesi che la metà dei maschi adulti della Repubblica di `AILATI` siano più grassi di 79.0 Kg contro l'alternativa che i più grassi di 79.0 Kg siano più della metà.

SOLUZIONE Sia X il carattere corrispondente al partito votato e Y il carattere peso in Kg. Organizziamo i dati raggruppati forniti nella seguente tabella a doppia entrata:

$X \setminus Y$	< 72.0	$[72.0, 79.0]$	$(79.0, 90.0]$	> 90.0	$N_X =$
B	37	73			246
C	21			32	200
E		129		63	
	120	266	301	133	

Completiamo la tabella, ricostruendo i dati mancanti:

$X \setminus Y$	< 72.0	$[72.0, 79.0]$	$(79.0, 90.0]$	> 90.0	$N_X =$
$N_B =$	$N_{b1} = 37$	$N_{b2} = 73$	$N_{b3} = \boxed{98}$	$N_{b4} = \boxed{38}$	$N_{Xb} = 246$
$N_C =$	$N_{c1} = 21$	$N_{c2} = \boxed{64}$	$N_{c3} = \boxed{83}$	$N_{c4} = 32$	$N_{Xc} = 200$
$N_E =$	$N_{e1} = \boxed{62}$	$N_{e2} = 129$	$N_{e3} = \boxed{120}$	$N_{e4} = 63$	$N_{Xe} = \boxed{374}$
$N_Y =$	$N_{Y1} = 120$	$N_{Y2} = 266$	$N_{Y3} = 301$	$N_{Y4} = 133$	$N = \boxed{820}$

1. Impostiamo un test χ^2 di indipendenza fra la variabile categorica X e la variabile continua Y. La statistica di Pearson Q ha valore:

$$Q = N \sum_{j=1}^{4} \frac{N_{bj}^2}{N_{Xb} * N_{Yj}} + N \sum_{j=1}^{4} \frac{N_{cj}^2}{N_{Xc} * N_{Yj}} + N \sum_{j=1}^{4} \frac{N_{ej}^2}{N_{Xe} * N_{Yj}} - N \simeq 8.677.$$

   Asintoticamente $Q \sim \chi^2_{(3-1)(4-1)} = \chi^2_6$ e $P(\chi^2_6 > 8.677) = 1 - P(\chi^2_6 \leq 8.677) \simeq 1 - 0.81 = 0.19$: c'è una forte evidenza empirica ad accettare l'ipotesi H_0 : "X, Y sono indipendenti", cioè che il peso non influenzi le preferenze elettorali. (Avendo a disposizione soltanto la tavola del Pestman dei quantili della χ^2_6, osservo che $\chi^2_6(7.84) = 0.75 < \chi^2_6(8.677) < \chi^2_6(10.6) = 0.90$ e quindi $0.10 < p\text{-value} < 0.25$)

2. $(120 + 63)/374 = 183/374 \simeq 0.4893 \simeq 49\%$ fornisce una stima della percentuale di elettori del partito *Ecrù* più grassi di 79.0 Kg. Un intervallo di confidenza asintotico di livello 0.95 è dato da

$$183/374 \pm z_{\frac{1+0.95}{2}} \sqrt{\frac{183/374 \times (1 - 183/374)}{374}} \simeq (0.44, 0.54).$$

3. Sia $\theta = P($"un maschio adulto di AILATI pesa più di 79.0"$)$. Una stima $\hat{\theta}$ per θ è data da $\hat{\theta} = (301 + 133)/820 = 434/820 \simeq 0.5293$. Costruiamo un test per grandi campioni per verificare $H_0 : \theta = 0.5$ contro $H_1 : \theta > 0.5$. Rifiutiamo H_0 se $(\hat{\theta} - 0.5)/\sqrt{0.5 \times 0.5/820} > z_{0.95} = 1.645$. Essendo che $(434/820 - 0.5)/\sqrt{0.5 \times 0.5/820} \simeq 1.6762 > 1.645$, accettiamo H_1. ∎

Esercizio 5.4 Abbiamo raccolto 11 dati giornalieri sul peso medio delle uova grandi covate da galline vecchie allevate a terra nell'azienda avicola `aaa`. Di seguito sono riportati i pesi medi $X_i, \quad i = 1, \ldots, 11$ delle uova (espressi in grammi) nell'ordine del giorno di rilevazione:

65.0, 70.3, 71.2, 63.0, 64.6, 67.0, 62.8, 68.9, 71.3, 60.9, 66.4

1. Impostate un test di ampiezza 5% per verificare se i pesi delle uova in giorni diversi siano indipendenti.
2. Impostate un test di ampiezza 5% sulla varianza del peso medio giornaliero delle uova, specificando ipotesi nulla e alternativa in modo tale che costituisca errore di seconda specie acquistare uova il cui peso abbia varianza minore di 9.0 grammi^2. Assumete l'ipotesi di gaussianità dei dati.

Soluzione

1. Impostiamo un test di aleatorietà di Kendall di livello $\alpha = 5\%$ per l'ipotesi nulla di casualità del campione: $H_0 : X_1, \ldots, X_n \; iid \; \sim F$.

   Per ogni $i = 1, \ldots, 10$ occorre anzitutto calcolare C_i=numero di osservazioni successive a X_i maggiori di X_i e D_i=numero di osservazioni successive a X_i minori di X_i. Si ha:

i	1	2	3	4	5	6	7	8	9	10	11
X_i	65.0	70.3	71.2	63.0	64.6	67.0	62.8	68.9	71.3	60.9	66.4
$C_i, \; D_i$	6, 4	2, 7	1, 7	5, 2	4, 2	2, 3	3, 1	1, 2	0, 2	1, 0	

   Da ciò segue che $C = \sum_{i=1}^{10} C_i = 25, \;\; D = \sum_{i=1}^{10} D_i = 30$ e dunque il valore della statistica test $T = C - D$ è $T = -5$. Sia ora $q_{C-D}(p)$ il quantile di ordine p della statistica T. Per $n = 11$, $q_{C-D}(1 - 0.05) = 21$ e $q_{C-D}(1 - 0.05/2) = q_{C-D}(0.975) = 25$. Essendo $T = -5$, allora sia $T > -21$ sia $|T| < 25$. Quindi NON rifiutiamo l'ipotesi H_0 di dati indipendenti a livello $\alpha = 5\%$ sia contro l'alternativa unilatera: H_1:"c'è trend negativo", sia contro l'alternativa bilatera: H_1: "i dati non sono indipendenti".

2. Chiamiamo σ^2 la varianza del peso. Commettiamo errore di seconda specie se rifiutiamo l'ipotesi alternativa H_1 quando è vera; deduciamo dal testo che $H_1 : \sigma^2 < 9$. Quindi impostiamo un test di verifica dell'ipotesi nulla unilaterale $H_0 : \sigma^2 \geq 9$ (ma il test non cambia specificando $H_0 : \sigma^2 = 9$) contro l'alternativa $H_1 : \sigma^2 < 9$, nel caso di un campione casuale estratto da popolazione gaussiana con media incognita. Una regione critica di ampiezza 5% è

$$G = \{(x_1, \ldots, x_{11}) \in \mathbb{R}^{11} : \frac{10s^2}{9} \leq \chi^2_{10}(0.05)\}$$

   dove S^2 è la varianza campionaria. Con i dati a nostra disposizione abbiamo $\bar{x} \simeq 66.482$ e

$$s^2 = \frac{\sum_{j=1}^{11}(x_j - \bar{x})^2}{10} = \frac{11}{10}\Big(\frac{\sum_{j=1}^{11} x_j^2}{11} - \bar{x}^2\Big) \simeq \frac{11}{10}(4431.45 - 66.482^2) \simeq 12.753$$

   Poiché $10 \times 12.753/9 = 14.17 > 3.94 = \chi^2_{10}(0.05)$, accettiamo l'ipotesi nulla H_0. ∎