UNIVERSITÀ DEGLI STUDI ROMA TRE
Corso di Laurea in Matematica
ST410 - Statistica 1 - A.A. 2013/2014

II Esonero - 10 Gennaio 2014

1	2	3	4	5	6	7	8	Tot.

Avvertenza: Svolgere ogni esercizio nello spazio assegnato, senza consegnare altri fogli e accompagnando le risposte con spiegazioni complete, chiare ed essenziali. Scrivere il proprio nome su ogni foglio nello spazio predisposto. Non è consentito l'utilizzo di libri o appunti; è consentito l'uso di una calcolatrice non programmabile.

Io sottoscritto/a

COGNOME: .... **NOME:** ....

MATRICOLA: ....

dichiaro, sotto la mia responsabilità, di aver svolto l'odierna prova di esame in modo autonomo e senza l'ausilio di strumenti o materiale non permesso.

Roma, 10 Gennaio 2014

FIRMA

Esercizio 1. Sia $X_1, \ldots . X_m$ un campione casuale di ampiezza m di v.c. i.i.d., estratto da una densità $f(\cdot, \theta)$ dipendente da un parametro $\theta \in \mathbb{R}$. Sia T un generico stimatore di θ.

1. (2pt) Enunciare la disuguaglianza di Frechét – Rao – Cramér sulla varianza di T.
2. (4pt) Si supponga che f sia la densità di una Poisson di parametro θ (i.e., $f(n, \theta) = e^{-\theta} \frac{\theta^n}{n!}$ con $n \in \mathbb{N}$ e $\theta > 0$). Sia $T = \frac{1}{m} \sum_{i=1}^m X_i$ stimatore di θ. Dimostrare che T è corretto ed efficiente (in senso assoluto). [Si ricordi che una Poisson di paramentro θ ha media θ e varianza θ].
3. (1pt) Relativamente al punto precedente, è possibile trovare uno stimatore M di θ corretto ed efficiente (in senso assoluto) e con $M \neq T$? Nel caso se ne dia un esempio.

Soluzione:

1. Si veda, ad esempio, L. Pieraccini - Fondamenti di Inferenza Statistica (2a edizione) - Giappichelli Editore - pagina 288.
2. $E(T) = \frac{1}{m} \sum_{i=1}^m E(X_i) = \theta$, quindi T è corretto. Data l'indipendenza delle X_i, $Var(T) = \frac{1}{m^2} \sum_{i=1}^m \theta = \frac{\theta}{m}$.
   La funzione di verosimiglianza del campione L è
   $$L(x_1, \ldots, x_m; \theta) = e^{-m\theta} \frac{\theta^{x_1 + \cdots + x_m}}{x_1! \cdot \cdots \cdot x_m!},$$
   e il suo logaritmo
   $$\log(L) = -m\theta + \log(\theta) \cdot \sum_{i=1}^m x_i - \log(x_1! \cdot \cdots \cdot x_m!).$$
   La funzione L soddisfa le ipotesi della disuguaglianza di Frechét – Rao – Cramér, quindi per verificare che T è efficiente in senso assoluto bisogna solo calcolare il limite inferiore della disuguaglianza e controllare che sia pari a $Var(T)$.
   Essendo T non distorto, il limite inferiore della disuguaglianza è
   $$\frac{1}{E((\frac{\partial}{\partial\theta} \log(L(X_1, \ldots, X_m; \theta)))^2)}.$$
   dove $\frac{\partial}{\partial\theta} \log(L(X_1, \ldots, X_m; \theta)) = -m + \frac{1}{\theta} \sum_{i=1}^m X_i = -m + \frac{m}{\theta} T$ e quindi $\left(\frac{\partial}{\partial\theta} \log(L(X_1, \ldots, X_m; \theta))\right)^2 = m^2 + \frac{m^2}{\theta^2} T^2 - 2\frac{m^2}{\theta} T$.
   $$E\left(\left(\frac{\partial}{\partial\theta} \log(L(X_1, \ldots, X_m; \theta))\right)^2\right) = m^2 + \frac{m^2}{\theta^2} E(T^2) - 2\frac{m^2}{\theta} E(T).$$
   Siccome $E(T^2) = Var(T) + E(T)^2$ e $E(T) = \theta$, allora
   $$E\left(\left(\frac{\partial}{\partial\theta} \log(L(X_1, \ldots, X_m; \theta))\right)^2\right) = m^2 + \frac{m^2}{\theta^2}(Var(T) + \theta^2) - 2m^2 = \frac{m}{\theta} = \frac{1}{Var(T)}$$
3. Non è possibile: uno stimatore T di θ non distorto ed efficiente è unico. Si veda ad esempio L. Pieraccini - Fondamenti di Inferenza Statistica (2a edizione) - Giappichelli Editore - pagina 290.

Nome e Cognome:

Esercizio 2. Sia $X_1, \dots, X_n$ un campione casuale estratto da una densità di probabilità $f(\cdot, \theta)$ dipendente da un unico parametro incognito θ. Sia $T = f(X_1, \dots, X_n)$ uno stimatore di θ.

1. (3pt) Cosa vuol dire, per T, essere uno stimatore sufficiente di θ?
2. (2pt) Enunciare il criterio di fattorizzazione di Neyman – Fisher sulla caratterizzazione degli stimatori sufficienti.

Soluzione:

1. Si veda ad esempio L. Pieraccini - Fondamenti di Inferenza Statistica (2a edizione) - Giappichelli Editore - pagina 292.
2. Si veda ad esempio L. Pieraccini - Fondamenti di Inferenza Statistica (2a edizione) - Giappichelli Editore - pagina 293.

Esercizio 3. Sia $X_1, \ldots, X_4$ un campione casuale estratto da una normale di media incognita e deviazione standard σ. Sia $3.3, 0.3, 0.6, 0.9$ una realizzazione campionaria.

1. (2pt) Sulla base della realizzazione campionaria indicata, e supponendo che $\sigma = 3$, trovare un intervallo di confidenza al 90% per la media della popolazione.

2. (3pt) Sulla base della realizzazione campionaria indicata, e supponendo, questa volta, che σ è anch'esso incognito, trovare un intervallo di confidenza al 90% per la media della popolazione.

Soluzione:

1. Lo stimatore $Z := \frac{\overline{X}-\mu}{3/2}$ (dove $\overline{X}$ è la media campionaria) è distribuito come una normale standardizzata. Sia $z_{0.95}$ il quantile di ordine 95% di una normale standardizzata. Allora

$$P\left(-z_{0.95} \leq \frac{\overline{X}-\mu}{3/2} \leq z_{0.95}\right) = 0.9.$$

Con semplici passaggi algebrici si ricavano gli stimatori L_{inf} e L_{sup} del limite inferiore e superiore dell'intervallo:

$$L_{\text{inf}} = \overline{X} - \frac{3}{2}z_{0.95}, L_{\text{sup}} = \overline{X} + \frac{3}{2}z_{0.95}.$$

Sulla realizzazione campionaria $\overline{X} = 1.275$, perciò l'intervallo di confidenza cercato è

$$\left[1.275 - \frac{3}{2}z_{0.95}, 1.275 + \frac{3}{2}z_{0.95}\right].$$

2. Lo stimatore $T := \frac{\overline{X}-\mu}{S/2}$ (dove $\overline{X}$ è la media campionaria e S la radice della varianza campionaria) è distribuito come una t di student con 3 gradi di libertà. Sia $t_{0.95}$ il quantile di ordine 95% di una t di student con 3 gradi di libertà. Allora

$$P\left(-t_{0.95} \leq \frac{\overline{X}-\mu}{S/2} \leq t_{0.95}\right) = 0.9.$$

Con semplici passaggi algebrici si ricavano gli stimatori L_{inf} e L_{sup} del limite inferiore e superiore dell'intervallo:

$$L_{\text{inf}} = \overline{X} - \frac{S}{2}t_{0.95}, L_{\text{sup}} = \overline{X} + \frac{S}{2}t_{0.95}.$$

Sulla realizzazione campionaria $\overline{X} = 1.275$ e $S = 1.372$, perciò l'intervallo di confidenza cercato è:

$$\left[1.275 - 0.686t_{0.95}, 1.275 + 0.686t_{0.95}\right].$$

Nome e Cognome:

Esercizio 4. Si supponga di avere un dado a sei facce, numerate da 1 a 6, e che si voglia stabilire se è truccato o regolare sulla base di un unico lancio. Sia L la variabile casuale "punteggio del lancio del dado". A priori si sa che la densità di probabilità di L assume la seguente forma (in dipendenza di un unico parametro incognito $\theta \in \left[0, \frac{5}{6}\right]$):

$$f(x, \theta) = \begin{cases} \frac{1}{6} - \frac{\theta}{5} & \text{se } x = 1, 2, 3, 4, 5; \\ \frac{1}{6} + \theta & \text{se } x = 6. \end{cases}$$

Si vuole testare l'ipotesi nulla che il dado sia regolare $H_0 : \theta = 0$ in alternativa all'ipotesi $H_1 : \theta \neq 0$, i.e., all'ipotesi che il dado sia truccato. Sia l il valore del lancio del dado. Si vuole utilizzare la seguente regola di decisione D:

$$D : \begin{cases} l \leq 4 & \text{si conclude per } H_0; \\ l > 4 & \text{si conclude per } H_1. \end{cases}$$

1. (2pt) Nel caso in considerazione: cosa si intende per errore di prima specie del test d'ipotesi? Si determini la sua probabilità.
2. (2pt) Nel caso in considerazione: cosa si intende per regione di accettazione dell'ipotesi nulla? E per regione critica (o di rifiuto) dell'ipotesi nulla? Si determinino esplicitamente sia la regione di accettazione che la regione critica.
3. (2pt) Nel caso in considerazione: cosa si intende per funzione di potenza del test? La si determini esplicitamente.
4. (2pt) Nel caso in considerazione: elaborare un test, basato sempre su un unico lancio del dado e per testare l'ipotesi H_0 in alternativa a H_1, per il quale, però, la probabilità dell'errore di prima specie sia strettamente minore della probabilità calcolata al punto 1.

Soluzione:

1. Si veda ad esempio L. Pieraccini - Fondamenti di Inferenza Statistica (2a edizione) - Giappichelli Editore - pagina 401.

   La probabilità di commettere un errore di prima specie è la probabilità che si concluda per H_1 nonostante sia vera H_0, quindi, in questo caso, la probabilità (calcolata per $\theta = 0$) che L sia strettamente maggiore di 4: $P_{\theta=0}(L > 4) = P_{\theta=0}(L = 5) + P_{\theta=0}(L = 6) = \frac{1}{3}$.

2. Si veda ad esempio L. Pieraccini - Fondamenti di Inferenza Statistica (2a edizione) - Giappichelli Editore - pagina 401.

   La regione di accettazione è il sottoinsieme dello spazio campionario per cui H_0 viene accettata, la regione di rifiuto è il complementare (nello spazio campionario) della regione di accettazione. Nel caso in considerazione lo spazio campionario consiste nei 6 possibili risultati del lancio: $\{1, 2, 3, 4, 5, 6\}$. Data la regola di decisione, quindi, la regione di accettazione è il sottoinsieme $\{1, 2, 3, 4\}$, mentre la regione di rifiuto il sottoinsieme $\{5, 6\}$.

3. Si veda ad esempio L. Pieraccini - Fondamenti di Inferenza Statistica (2a edizione) - Giappichelli Editore - pagina 415.

La funzione di potenza del test è la funzione che associa ad ogni possibile valore del parametro θ la probabilità (sotto θ) che il punteggio del lancio cada nella regione di rifiuto dell'ipotesi nulla. Quindi la funzione di potenza è $\gamma(\theta) = P_\theta(L > 4) = \frac{1}{3} + \frac{4}{5}\theta$ per $\theta \in \left[0, \frac{5}{6}\right]$.

4. L'aver scelto 4 come valore critico per la regola di decisione non sembra essere stata la scelta più ragionevole. In effetti si vuole stabilire se il dado abbia una "preferenza" per il numero 6, dato che gli altri numeri sono comunque tutti equiprobabili, compreso il numero 5. Visto che ci basiamo su un unico lancio, intuitivamente saremo portati a considerare il dado truccato se esce 6 e non truccato se esce un altro numero. Perciò proviamo a modificare il test cambiando esclusivamente la regola di decisione, e optando per accettare H_0 anche se esce il numero 5. In questo caso la probabilità dell'errore di prima specie è $P_{\theta=0}(L = 6) = \frac{1}{6} < \frac{1}{3}$.

Nome e Cognome:

Esercizio 5. Ad un campione di 21 maschi e ad un campione di 3 femmine viene somministrato un test per la rilevazione dell'ansia. Attraverso il test si ottiene un punteggio tanto più alto quanto più è elevato il grado d'ansia del soggetto. Si supponga che il punteggio, sia per la popolazione dei maschi che per la popolazione delle femmine sia distribuito come una normale con varianza (incognita) σ^2 e media (incognita) e_m per i maschi e e_f per le femmine, eventualmente con $e_m \neq e_f$.

Si supponga di aver ottenuto i seguenti dati campionari:

- media del punteggio dei maschi: 44;
- devianza del punteggio dei maschi: 1029;
- media del punteggio delle femmine: 55;
- devianza del punteggio delle femmine: 78.

Si vuole testare l'ipotesi $H_0 : e_m = e_f$ in alternativa all'ipotesi $H_1 : e_m \neq e_f$.

1. (2pt) Si dica quale statistica T è opportuno usare, e (senza dimostrarlo) qual è la sua distribuzione.
2. (2pt) Si determini la regola di decisione (per accettare o rigettare H_0) se si vuole che la probabilità dell'errore di prima specie sia α.
3. (2pt) Si calcoli la statistica T sul campione. Sapendo che il quantile di ordine 0.995 di T è 2.819 e che il quantile di ordine 0.95 di T è 1.717 dire se H_0 si può rigettare con un livello di significatività del 10%; dire se H_0 si può rigettare con un livello di significatività dell'1%.

Soluzione:

1. Si veda ad esempio L. Pieraccini - Fondamenti di Inferenza Statistica (2a edizione) - Giappichelli Editore - pagina 432.

   Il problema proposto richiede di testare l'uguaglianza o meno della media dei punteggi per la popolazione dei maschi (utilizzando un campione di 21 individui) e per la popolazione delle femmine (utilizzando un campione di 3 individui). Si suppone che il punteggio nella popolazione maschile e nella popolazione femminile segua una legge normale con identica varianza. Dato che il campione maschile è indipendente dal campione femminile, si è nel caso classico del test sulla differenza tra due medie, perciò la statistica T da usare è

   $$\frac{\overline{X_1} - \overline{X_2}}{S\sqrt{\frac{1}{21} + \frac{1}{3}}},$$

   dove $\overline{X_1}$ è la media campionaria dei punteggi dei maschi, $\overline{X_2}$ la media campionaria dei punteggi delle femmine, S la radice di $\frac{D_1+D_2}{21+3-2}$, con D_1 devianza campionaria del punteggio dei maschi e D_2 devianza campionaria del punteggio delle femmine. Tale statistica T è distribuita come una t di Student con $21 + 3 - 2 = 22$ gradi di libertà.

2. Sia $q := t_{1-\frac{\alpha}{2}}$ il quantile di ordine $1 - \frac{\alpha}{2}$ di una t di student con 22 gradi di libertà. Allora $P_{e_m=e_f}(-q \leq T \leq q) = 1 - \alpha$ e quindi la probabilità dell'errore di primo tipo è esattamente α. Quindi, chiamato t il valore della statistica test T sul campione, la regola di decisione sarà semplicemente:

$$D : \begin{cases} |t| \leq q & \text{si conclude per } H_0; \\ |t| > q & \text{si conclude per } H_1. \end{cases}$$

3. Sul campione T assume valore $\cong -2.5124$. Essendo la distribuzione di T simmetrica rispetto all'asse delle ordinate, H_0 si rigetta con livello di significatività pari al 10% se il valore di T sul campione cade al di fuori dell'intervallo $[x, y]$, con x quantile di ordine 0.05 e y quantile di ordine 0.95. Per simmetria della distribuzione, $x = -y$. Dato che $-2.5124 < -1.717$ allora possiamo rigettare l'ipotesi nulla con un livello di significatività del 10%.

   Analogamente: $-2.5124 > -2.819$, quindi non possiamo rigettare H_0 con un livello di significatività dell'1%.

Nome e Cognome:

Esercizio 6. Le variabili casuali $Y_1, \dots, Y_n$ sono descritte dalla relazione

$$Y_i = \theta x_i^2 + U_i,$$

dove $x_1, \dots, x_n$ sono costanti fisse e $U_1, \dots, U_n$ sono v.c. i.i.d. $\sim N(0, \sigma^2)$

1. (2pt) Trovare lo stimatore dei minimi quadrati di θ.
2. (3pt) Trovare lo stimatore di massima verosimiglianza di θ.

Soluzione:

1. Si tratta di trovare il minimo, in funzione di θ, della funzione $q(\theta) := \sum_{i=1}^n (Y_i - \theta x_i^2)^2$. Siccome $q(\theta)$ è una parabola (in θ), allora il minimo si ottiene semplicemente trovando l'unico zero della derivata prima $q'(\theta)$.

   Ora: $q'(\theta) = \sum_{i=1}^n -2(Y_i - \theta x_i^2)(x_i^2)$, quindi

   $$q'(\theta) = 0 \Leftrightarrow \theta = \frac{\sum_{i=1}^n x_i^2 Y_i}{\sum_{i=1}^n x_i^4},$$

   e perciò $\frac{\sum_{i=1}^n \sum x_i^2 Y_i}{\sum_{i=1}^n x_i^4}$ è lo stimatore dei minimi quadrati di θ.

2. La funzione di verosimiglianza del campione di ampiezza n è:

   $$L(y_1, \dots, y_n; \theta | x_1, \dots, x_n) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{\frac{-\sum_{i=1}^n (y_i - \theta x_i^2)^2}{2\sigma^2}}.$$

   Per trovarne il massimo (in funzione di θ) è opportuno passare al logaritmo:

   $$\log(L) = -\frac{n}{2} \log(2\pi\sigma^2) - \sum_{i=1}^n (y_i - \theta x_i^2)^2.$$

   Quindi il massimo di $\log(L)$ (e perciò di L) si ha quando è minima la funzione $\sum_{i=1}^n (y_i - \theta x_i^2)^2$, i.e., lo stimatore di massima verosimiglianza di θ coincide con lo stimatore dei minimi quadrati di θ.

(eventuale continuazione esercizio n.:)