Esame Scritto del Terzo Appello

Tempo a disposizione: 2 ore

Riportare il numero di matricola **all'inizio** di ogni foglio. La soluzione di ogni esercizio deve essere scritta in modo chiaro e ordinato, e può non essere valutata se la calligrafia è illeggibile. Le soluzioni degli esercizi devono essere riportate sul foglio protocollo nell'ordine proposto, la soluzione di ogni esercizio deve iniziare in una nuova pagina. Le soluzioni devono includere il procedimento dettagliato che porta alle risposte. Risposte corrette non adeguatamente motivate saranno penalizzate.
Non è permesso l'uso di note, appunti, manuali o materiale didattico di alcun tipo, al di fuori del formulario e delle tavole statistiche fornite assieme al compito. Non è permesso l'uso di dispositivi elettronici ad esclusiva eccezione di una calcolatrice non programmabile.
In ROSSO si riportano alcuni errori comuni riscontrati nella correzione.

1. Per ognuna delle seguenti affermazioni si determini se essa è VERA oppure FALSA, motivando rigorosamente le risposte.

   (a) La funzione di ripartizione di una variabile aleatoria di Poisson è una funzione continua.
   FALSO: la c.d.f. di una variabile discreta è una funzione a salti, con i salti corrispondenti ai punti in cui la funzione di massa non si annulla.
   ERRORE GRAVE: molti hanno scritto che *tutte* le funzioni di ripartizione sono continue.

   (b) Se un campione di variabili $X_1, \dots, X_n$ i.i.d. hanno momento secondo finito, allora $\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2$ è uno stimatore corretto della varianza $\sigma^2 = \mathrm{Var}(X_i)$.
   FALSO: come visto a lezione vale

   $$\mathbb{E}[\sum_{i=1}^n \left(X_i - \bar{X}\right)^2] = (n-1)\sigma^2,$$

   per cui per avere uno stimatore corretto è necessario dividere per $n-1$.

   (c) Se X è una variabile Gaussiana standard, la variabile aleatoria $Y = \frac{X+|X|}{2|X|}$ è Bernoulli di parametro $p = 1/2$.
   VERO: anzitutto osserviamo che siccome $\mathbb{P}(X=0)=0$ possiamo ignorare il problema della divisione per 0; detto questo se $X>0$ vale $Y=1$ e $\mathbb{P}(X>0)=1/2$, mentre se $X>0$ vale $Y=0$ e $\mathbb{P}(X<0)=1/2$.

   (d) Una variabile aleatoria con densità chi-quadro non assume mai valori negativi, a prescindere dal numero di gradi di libertà.
   VERO: la densità chi-quadro è la densità della somma di quadrati di variabili Gaussiane, dunque è un numero non negativo.
   ERRORI: non era sufficiente disegnare un grafico della densità.

   (e) Per effettuare il test chi-quadro sulla varianza σ^2 di un campione Gaussiano $N(m, \sigma^2)$ è necessario conoscere il valore di m.
   FALSO: la statistica test è $\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2$ e non fa uso del parametro m.

   (f) La regione critica dello Z-test bilatero corrisponde all'evento che un intervallo di fiducia per la media m del campione Gaussiano in considerazione contenga il parametro m.
   FALSO: La regione di accettazione si può riscrivere come segue:

   $$C^c = \left\{|\bar{X}_n - m_0| \leq \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right\} = \left\{m_0 \in \left[\bar{X}_n \pm \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}\right]\right\},$$

quindi essa corrisponde all'evento che un intervallo di fiducia *non* contenga il parametro.

2. Si consideri la funzione

$$f_\theta(x) = \begin{cases} 0 & x \le 0 \\ \frac{2x}{\theta} e^{-x^2/\theta} & x > 0 \end{cases}$$

dipendente da un paramentro $\theta > 0$.

(a) Si dimostri che, per ogni $\theta > 0$, f_θ è la funzione di densità di una variabile aleatoria, se ne tracci il grafico (per $\theta = 1$) e si calcoli la relativa funzione di ripartizione. (*Suggerimento:* usare la sostituzione $y = x^2/\theta$).
Si osserva facilmente che f_θ è una funzione non-negativa, e per $t \ge 0$ vale

$$F(t) = \int_0^t \frac{2x}{\theta} e^{-x^2/\theta} dx = \int_0^{t^2/\theta} e^{-y} dy = 1 - e^{-t^2/\theta}$$

in cui abbiamo sostituito $y = x^2/\theta$, e questo mostra che (mandando $t \to \infty$) vale $\int_{\mathbb{R}} f_\theta(x) dx = 1$ e determina la funzione di ripartizione (nulla per $t \le 0$).

(b) Si calcoli il momento secondo di una variabile aleatoria con densità f_θ.
Con la stessa sostituzione del punto precedente,

$$\mathbb{E}[X^2] = \int_0^\infty \frac{2x^3}{\theta} e^{-x^2/\theta} dx = \int_0^\infty \theta y e^{-y} dy = \theta.$$

(c) Si determini lo stimatore di massima verosimiglianza per θ e si dica se è corretto.
La verosimiglianza di dati $x_1, \ldots, x_n > 0$ è

$$L(\theta) = 2^n (x_1 x_2 \cdots x_n) \frac{\exp(-\frac{1}{\theta}(x_1^2 + \cdots + x_n^2))}{\theta^n}, \quad \theta > 0,$$

e imponendo $L' = 0$ (osservando che nella riga precedente la dipendenza da θ è solo nella frazione) si ottiene:

$$0 = \frac{\exp(-\frac{1}{\theta}(x_1^2 + \cdots + x_n^2))}{\theta^{n+2}} (x_1^2 + \cdots + x_n^2 - n\theta),$$

ossia

$$\hat{\theta} = \frac{x_1^2 + \cdots + x_n^2}{n}.$$

Se $X_1, \ldots, X_n$ è un campione di variabili con densità f_θ, dal punto precedente si ha

$$\mathbb{E}[\frac{X_1^2 + \cdots + X_n^2}{n}] = \mathbb{E}[X_1^2] = \theta,$$

per cui lo stimatore è corretto.

3. Viene condotto un trial clinico su una pomata che dovrebbe prolungare il tempo di reinfezione in individui che hanno contratto l'herpes labiale, che in pazienti non trattati è in media di un anno. Ai fini dell'analisi statistica dei dati si rappresenta il tempo di reinfezione di ogni paziente con una variabile aleatoria esponenziale di parametro λ da stimare.

(a) Il dottor Ambrosoli conduce il test su un campione (i.i.d.) di $N = 100$ pazienti che usano la pomata, ottenendo un tempo medio campionario di reinfezione di 1.1 anni, con una deviazione standard campionaria di 0.2 anni. Dire se è plausibile l'ipotesi nulla che la pomata in media non abbia effetto, ossia che il valore atteso del campione sia inferiore o uguale a un anno.

Siamo nel caso di test T sulla media di popolazione con varianza non nota, caso grandi campioni ($n = 100$). Detto X il tempo di reinfezione e m il suo valore atteso, testiamo $H_0 : m \leq m_0 := 1$ contro $H_1 : m_X > 1$. La statistica di test e la regione critica sono

$$\begin{aligned} T &= \frac{\sqrt{n}}{S_n}(\bar{X}_n - m_0), \\ C &= \{T > t_{1-\alpha,n-1} \approx z_{1-\alpha}\}. \end{aligned}$$

Il valore assunto da T con i dati $\bar{x} = 1.2$, $s = 0.2$ è 5, il cui p-value è $P(T > 5) \approx 1 - \Phi(5) \approx 0$, dunque rifiutiamo H_0 a ogni ragionevole livello: c'è evidenza che la pomata abbia effetto in media.

(b) Anche il dottor Brambilla conduce il test su pazienti che usano la pomata, ma nella sua clinica ha a disposizione solo 2 pazienti. Descrivendo il loro tempo di reinfezione con variabili i.i.d. esponenziali X_1, X_2, si determinino (in funzione del parametro incognito λ) la densità di probabilità della somma $Y = X_1 + X_2$, e il valore atteso $\mathbb{E}[Y]$.

Per linearità, il valore atteso di Y è

$$\mathbb{E}[Y] = \mathbb{E}[X_1] + \mathbb{E}[X_2] = \frac{2}{\lambda}.$$

Per calcolare la densità, usiamo la formula della convoluzione:

$$\begin{aligned} f_Y(z) &= \int_{\mathbb{R}} f_{X_1}(z-x) f_{X_2}(x)\, dx \\ &= \int_{\mathbb{R}} \lambda e^{-\lambda(z-x)} 1_{z-x>0} \lambda e^{-\lambda x} 1_{x>0}\, dx \\ &= \int_{\mathbb{R}} 1_{0<x<z} \lambda^2 e^{-\lambda(z-x)-\lambda x}\, dx \\ &= \lambda^2 z e^{-\lambda z} 1_{z>0}. \end{aligned}$$

Infine, la funzione di ripartizione $F_Y(z)$ è nulla per $z \leq 0$ e per $z > 0$ vale

$$\begin{aligned} F_Y(z) &= \int_0^z \lambda^2 x e^{-\lambda x}\, dx \\ &= -\lambda x e^{-\lambda x}\big|_{x=0}^{z} + \int_0^z \lambda e^{-\lambda x}\, dx \\ &= 1 - (1 + \lambda z) e^{-\lambda z}, \end{aligned}$$

dove abbiamo usato l'integrazione per parti.

(c) Nel caso del dottor Brambilla *non* è possibile applicare l'approssimazione Gaussiana e di conseguenza il T-test. Si determini una regione critica di livello α per il test di ipotesi

$$H_0) 1/\lambda = 1, \quad H_1) 1/\lambda > 1,$$

nella forma $C = \{\bar{X}_2 > d\}$ determinando d in funzione dei quantili della distribuzione di Y. Sapendo che per $\lambda = 1$ vale $F_Y(4.74) \simeq 0.95$, se i dati di Brambilla sono $x_1 =$

$1.2, x_2 = 1.1$ il test accetta o rifiuta H_0 a livello 5% ?
Imponiamo livello α e massima potenza, ricordando $\bar{X}_2 = 2Y$,

$$\alpha = \mathbb{P}_{\lambda=1}(\bar{X}_2 > d) = \mathbb{P}_{\lambda=1}(Y > 2d) = 1 - F_{Y,\lambda=1}(2d),$$

cioè $2d = r_{1-\alpha}$, dove r_β è il β-quantile della distribuzione Y (con densità dal punto b) nel caso $\lambda = 1$. Nel caso dei dati di Brambilla, la statistica test Y assume valore 2.3, mentre per $\alpha = 0.05$ si ha $d = 4.74/2 = 2.37$, per cui non possiamo rigettare H_0: Brambilla deve concludere che sul suo campione non si può escludere che l'aumento del tempo di reinfezione sia solo fluttuazione statistica.