Esame Scritto del Secondo 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. Se l'esercizio lo richiede, evidenziare il risultato numerico nella soluzione. 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.
È permesso l'uso di note, appunti, manuali e materiale didattico. Non è permesso l'uso di dispositivi elettronici ad esclusiva eccezione di una calcolatrice non programmabile. L'infrazione di queste regole o la comunicazione con altri comportano l'annullamento del compito.

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

   (a) Data X v.a. discreta con $\mathbb{P}(X = 2) = \mathbb{P}(X = -1) = 1/2$, allora $\mathbb{E}[X^2] = 5/2$.
   VERA: $\mathbb{E}[X^2] = 2^2 \cdot 1/2 + (-1)^2 \cdot 1/2 = 5/2$.

   (b) La mediana dei dati $(x_1, \dots x_n)$ è il valore più frequente nei dati.
   FALSA: la moda è il valore più frequente, la mediana è il dato intermedio (se n è dispari) o la media dei due dati intermedi (se n è pari).

   (c) Se X è una variabile con densità f (ad esempio una v.a. esponenziale o gaussiana), allora $f(x) = \mathbb{P}(X = x)$ per ogni $x \in \mathbb{R}$.
   FALSA: Se X ha densità, $\mathbb{P}(X = x) = 0$ per ogni x.

   (d) Dati A, A', B eventi con $\mathbb{P}(B) > 0$, allora
   $$\mathbb{P}(A \cup A' \mid B) = \mathbb{P}(A \mid B) + \mathbb{P}(A' \mid B) - \mathbb{P}(A \cap A' \mid B)$$
   VERA: basta applicare la formula per la probabilità dell'unione alla probabilità condizionata $\mathbb{P}(\cdot \mid B)$.

   (e) Dato un campione $X_1, X_2, \dots$, uno stimatore corretto di un parametro θ è anche consistente (per la precisione, una famiglia di stimatori corretti $(U_n)_n$ di θ è anche consistente).
   FALSA: prendendo $\theta = m$ il valore atteso, $U_n = X_1$ è corretto ma non consistente (non converge in probabilità a m se X_1 non è costante).

   (f) Siano (x_i, y_i), $i = 1, \dots n$, i dati relativi a peso x_i in kg e altezza y_i in m (metri) di n individui. Siano ora $\tilde{y}_i = 100 \cdot y_i$ le altezze espresse in cm (centimetri). La retta di regressione $\tilde{y} = a + bx$ dei dati $(x_i, \tilde{y}_i)$ è la stessa $y = a + bx$ dei dati (x_i, y_i).
   FALSA: ad esempio, il coefficiente angolare $\tilde{b}$ dei dati $(x_i, \tilde{y}_i)$ e quello b dei dati (x_i, y_i) sono legati dalla relazione
   $$\tilde{b} = \frac{\mathrm{cov}(x, \tilde{y})}{\mathrm{var}(x)} = \frac{100 \cdot \mathrm{cov}(x, y)}{\mathrm{var}(x)} = 100 \cdot b,$$
   dove abbiamo usato che $\mathrm{cov}(x, cy) = c\,\mathrm{cov}(x, y)$.

2. Il numero di domande processate da un server in un minuto di un giorno feriale segue una distribuzione di Poisson di parametro 1. Supponiamo inoltre che il numero di domande processate in un dato minuto sia indipendente dal numero di domande processate in altri momenti.

   (a) Calcolare la probabilità che, in due dati minuti consecutivi (di un giorno feriale), il server processi almeno 2 domande.
   Per ogni i, sia X_i il numero di domande processate nell'i-simo minuto, le X_i sono v.a. Poisson di parametro 1 indipendenti. Per riproducibilità, $X_1 + X_2$ ha distribuzione Poisson di parametro $1 + 1 = 2$. La probabilità cercata è

$$\begin{aligned}\mathbb{P}(X_1 + X_2 \geq 2) &= 1 - \mathbb{P}(X_1 + X_2 = 0) - \mathbb{P}(X_1 + X_2 = 1)\\ &= 1 - e^{-2}(\frac{2^0}{0!} + \frac{2^1}{1!}) = 0.594.\end{aligned}$$

   (b) Calcolare almeno in modo approssimato la probabilità che, in un dato giorno (cioè 1440 minuti) feriale, il server processi almeno 1400 domande.
   Ricordiamo che una v.a. Poisson di parametro λ ha media $m = \lambda$ e varianza $\sigma^2 = \lambda$. Con la notazione del punto precedente, per il TCL (il numero di v.a. i.i.d. $n = 1440$ si può considerare grande), $\sum_{i=1}^{1440} X_i$ ha distribuzione approssimativamente $N(nm, n\sigma^2) = N(1440, 1440)$ (equivalentemente, $\sqrt{n}(\bar{X}_n - nm)/\sigma$ ha distribuzione approssimativamente $N(0,1)$). Quindi la probabilità cercata è circa (con Z v.a. $N(0,1)$)

$$\mathbb{P}(\sum_{i=1}^{1440} X_i \geq 1400) \approx \mathbb{P}(Z \geq \frac{1400 - 1440}{\sqrt{1440}} = -1.05) = \Phi(1.05) = 0.853.$$

In un giorno festivo, il numero di domande processate dallo stesso server in un minuto segue una distribuzione di Poisson di parametro 2. (Come prima, supponiamo che il numero di domande processate in un dato minuto sia indipendente dal numero di domande processate in altri momenti.)

   (c) Individuare almeno in modo approssimato la distribuzione del numero complessivo di domande tra domenica (festivo) e lunedì (feriale).
   Siano X_i, Y_j il numero di domande processate rispettivamente nell'i-simo minuto di lunedì e nel j-simo minuto di domenica. Per il TCL, come nel punto precedente, $\sum_{j=1}^{1440} Y_j$ ha distribuzione approssimativamente $N(1440 \cdot 2, 1440 \cdot 2) = N(2880, 2880)$. Inoltre, essendo le X_i e le Y_j indipendenti, anche $\sum_{i=1}^{1440} X_i$ e $\sum_{j=1}^{1440} Y_j$ sono indipendenti. Quindi, per riproducibilità delle v.a. gaussiane, il numero totale di domande processate

$$\sum_{i=1}^{1440} X_i + \sum_{j=1}^{1440} Y_j$$

   ha distribuzione approsimativamente $N(1440{+}2880, 1440{+}2880) = N(4320, 4320)$.

3. Sia $X_1, \dots X_n$ un campione i.i.d. di densità

$$f_\theta(x) = e^{\theta - x} 1_{[\theta,+\infty)}(x) = \begin{cases} e^\theta \cdot e^{-x} & \text{se } x \geq \theta \\ 0 & \text{se } x < \theta \end{cases}$$

con $\theta > 0$ parametro non noto.

(a) Verificare che la funzione di ripartizione F_θ delle X_i è

$$F_\theta(x) = \begin{cases} 1 - e^{\theta - x} & \text{se } x \geq \theta \\ 0 & \text{se } x < \theta \end{cases}$$

La funzione F_θ definita sopra è continua su $\mathbb{R}$ e C^1 a tratti (sui tratti $(-\infty, \theta)$ e $(\theta, +\infty)$), perciò ammette densità data da

$$F'_\theta(x) = e^{\theta - x} 1_{[\theta,+\infty)}(x),$$

che è proprio la densità f_θ. Alternativamente, si può verificare che $F_\theta(x) = \int_{-\infty}^{x} f_\theta(y)\, dy$ per ogni x.

(b) Calcolare $\mathbb{E}[e^{aX_1}]$ al variare di $a > 0$; per quali valori di $a > 0$, $\mathbb{E}[e^{aX_1}]$ è finito?
La v.a. e^{aX_1} è positiva, quindi ammette valore atteso (eventualmente infinito) dato da

$$\mathbb{E}[e^{aX_1}] = \int_{\mathbb{R}} e^{ax} f_\theta(x)\, dx = e^\theta \int_\theta^\infty e^{(a-1)x}\, dx$$
$$= \begin{cases} \frac{e^\theta}{a-1} e^{(a-1)x} \,\Big|_{x=\theta}^{\infty} = \frac{e^{-a\theta}}{1-a} & \text{se } a < 1, \\ +\infty & \text{se } a \geq 1, \end{cases}$$

dove abbiamo usato che $\int_b^\infty e^{cx}\, dx$ è finito se e solo se $c < 0$ (come si verifica facilmente dalla primitiva di e^{cx}).

(c) Calcolare, se esiste, lo stimatore di massima verosimiglianza di θ.
Dobbiamo massimizzare (in θ) la funzione di verosimiglianza

$$L(\theta; x_1, \dots x_n) = e^{n\theta} e^{-\sum_{i=1}^n x_i} 1_{x_i \geq \theta\ \forall i}.$$

Notiamo che $L(\theta; x_1, \dots x_n)$ vale 0 se $\theta > \min_i x_i$, mentre vale $e^{n\theta}$ moltiplicato per una costante (in θ) se $\theta \leq \min_i x_i$, in particolare è crescente su $(0, \min_i x_i)$. Di conseguenza la funzione L ha punto di massimo in $\min_i x_i$. Quindi

$$\hat{\theta} = \min_{i=1,\dots n} X_i$$

è l'unico stimatore di massima verosimiglianza.