Esame Scritto del Quarto 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.
Non è permesso l'uso di note, appunti, manuali o materiale didattico di alcun tipo. 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) Dato un campione aleatorio $X_1, \dots X_n$, il valore atteso delle X_i coincide con la media campionaria del campione $X_1, \dots X_n$.
   FALSO: Ad esempio, per l'esperimento lancio di una moneta equilibrata, e $X = 1$ se esce testa, $X = 0$ se esce croce, il valore atteso di X è $1/2$, ma la media campionaria del campione (testa, testa) è 1.

   (b) Se due eventi A, B sono indipendenti, lo sono anche i loro complementari A^c, B^c.
   VERO: se A, B sono eventi indipendenti, dalla legge di De Morgan e la definizione di indipendenza si deduce che

$$\begin{aligned} P\left(A^c \cap B^c\right) &= P\left((A \cup B)^c\right) = 1 - P(A \cup B) = 1 - P(A) - P(B) + P(A \cap B) \\ &= 1 - P(A) - P(B) + P(A)P(B) = (1 - P(A))(1 - P(B)) = P\left(A^c\right) P\left(B^c\right). \end{aligned}$$

   (c) Dato un campione aleatorio $X_1, \dots, X_n$ la cui legge dipende da un solo parametro $\theta \in \Theta \subseteq \mathbb{R}$, un intervallo di fiducia per il parametro θ è dato da una qualsiasi coppia di numeri $a, b \in \mathbb{R}$ in modo che $I = (a, b)$ contenga θ.
   FALSO: un intervallo di fiducia è dato da una coppia di variabili aleatorie, $I = [a(\omega), b(\omega)]$, e contiene il parametro θ con una certa probabilità, determinata dal livello di fiducia.

   (d) Una variabile aleatoria X ammette funzione di densità $f : \mathbb{R} \to \mathbb{R}$ se la probabilità che X assuma il valore $x \in \mathbb{R}$ è dato da $f(x)$.
   FALSO: se $f(x)$ è una funzione tale che

$$\mathbb{P}(X = x) = f(x),$$

   allora la condizione $\mathbb{P}(\Omega) = 1$ impone che f sia non nulla su al più numerabili valori, e X è dunque una variabile discreta.

   (e) Un p-value dell'ordine di 10^{-5} è da considerarsi evidenza statistica contro l'ipotesi nulla.

   VERO: se il livello $\alpha > \overline{\alpha} = 10^{-5}$, l'ipotesi nulla viene rifiutata dal test di regione critica di livello α, dunque un p-value così piccolo fa sì che l'ipotesi venga rifiutata per ogni livello α ragionevole.

(f) Per effettuare il test di Student sulla media di una popolazione è necessario conoscere la deviazione standard della popolazione.
FALSO: Il test di Student usa la deviazione standard campionaria.

2. Il Click Through Rate (CTR) di un annuncio pubblicitario è il rapporto tra il numero di click sull'annuncio e il numero di visualizzazioni dell'annuncio. Se un dato annuncio ha CTR pari a $p \in (0,1)$ la probabilità che una visualizzazione dell'annuncio porti a un click sull'annuncio stesso è pari a p. Supponiamo che un dato annuncio abbia CTR pari a 0.2.

(a) Su 5 visualizzazioni dell'annuncio, qual è la probabilità che ci siano almeno 2 click?
La distribuzione del numero X dei click è binomiale di parametri $n = 5$ e $p = 0.2$, dunque

$$\mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) = 1 - 1 \cdot 0.2^0 \cdot 0.8^5 - 5 \cdot 0.2^1 \cdot 0.8^4 = 0.26272.$$

(b) Su 10000 visualizzazioni dell'annuncio, qual è la probabilità che ci siano almeno 2100 click?
La distribuzione del numero X dei click è binomiale di parametri $n = 10000$ e $p = 0.2$. Dobbiamo ricorrere alla approssimazione Gaussiana fornita dal TCL,

$$\mathbb{P}(X \leq t) = \mathbb{P}\left(\frac{X - np}{\sqrt{np(1-p)}} \leq \frac{t - np}{\sqrt{np(1-p)}}\right) \simeq \Phi\left(\frac{t - np}{\sqrt{np(1-p)}}\right),$$

per cui sostituendo i dati

$$\begin{aligned}\mathbb{P}(X \geq 2100) = 1 - \mathbb{P}(X < 2100) &\simeq 1 - \Phi\left(\frac{2100 - 0.2 \cdot 10000}{\sqrt{0.2 \cdot 0.8 \cdot 10000}}\right) \\ &= 1 - \Phi(2.5) = 1 - 0.9938 = 0.0062.\end{aligned}$$

(c) Vogliamo che, con probabilità del 95%, su 10000 visualizzazioni vi siano almeno 3000 click. Come deve cambiare il CTR?
Ricorrendo al TCL come nel punto precedente, imponiamo

$$0.95 = \mathbb{P}(X \geq 3000) = 1 - \mathbb{P}(X < 3000) \simeq 1 - \Phi\left(\frac{3000 - p \cdot 10000}{\sqrt{p(1-p) \cdot 10000}}\right),$$

da cui prendendo il quantile abbiamo l'equazione approssimata

$$-1.64 \simeq q_{0.05} = \frac{30 - p \cdot 100}{\sqrt{p(1-p)}},$$

che conduce (quadrando e risolvendo l'equazione di secondo grado in p) alle soluzioni $p \simeq 0.2925$ e $p \simeq 0.3075$; la prima chiaramente non va bene, essendo minore di $0.3 = 3000/10000$ (essa corrisponde al quantile 1.64 e quindi alla probabilità 0.05), mentre la seconda verifica la condizione richiesta.

3. Una ditta di microchip dichiara che il tempo medio di elaborazione per i propri chip è di 130 ms. Su un campione di 225 chip, viene rilevato il tempo di elaborazione, trovando una media di 140 ms e una deviazione standard di 30 ms.

(a) Determinare un intervallo di fiducia di livello 95% per il tempo medio di elaborazione
Chiamiamo X_i il tempo di elaborazione dell'i-simo chip. Siamo nel caso di campione di taglia grande, con deviazione standard non nota. Applicando il TCL, troviamo l'intervallo di fiducia ($n = 225$, $\alpha = 0.95$, $\bar{X}_n$ e S_n sono rispettivamente media campionaria e deviazione standard campionaria)

$$[\bar{X}_n \pm \frac{S_n}{\sqrt{n}} q_{1-\alpha/2}] = [\bar{X}_n \pm \frac{S_n}{15} \cdot 1.96].$$

Sostituendo i valori $\bar{x}_n = 140$ e $s_n = 30$, otteniamo l'intervallo numerico $[140 \pm 3.92] = [136.08, 143.92]$.

(b) Sulla base dei dati del campione, l'affermazione della ditta (tempo medio pari a 130 ms) è plausibile? Effettuare un test statistico di livello 5%.
Dobbiamo effettuare un test bilatero sulla media m una popolazione di varianza non nota, nel caso di campione di taglia grande, con $H_0 : m = m_0 := 130$ e $H_1 : m \neq 130$. La statistica di test è

$$Z = \frac{\sqrt{n}}{S_n}(\bar{X}_n - m_0) = \frac{15}{S_n}(\bar{X}_n - 130),$$

con distribuzione approsimativamente gaussiana (per il TCL) sotto H_0. La regione di rigetto è $|Z| > q_{1-\alpha/2} = 1.96$. Applichiamo ora il test ai dati del problema: il valore assunto da Z per $\bar{x}_n = 140$ e $s_n = 30$ è $z = 5$, che è nella regione di rigetto: l'affermazione della ditta non è plausibile a livello 0.05.
Si può arrivare alla stessa conclusione semplicemente osservando che il valore assunto da Z cade nella regione di rigetto se e solo se $m_0 = 130$ non appartiene all'intervallo di fiducia trovato prima.

(c) Supponiamo che la distribuzione del tempo di elaborazione sia gaussiana. Vogliamo ora ottenere un intervallo di fiducia per il tempo medio di elaborazione, sempre di livello 95%, la cui semiampiezza sia non superiore a 1 ms. Quanto grande deve essere il campione? (si tenga conto che, cambiando il campione, possono variare sia la media campionaria sia la deviazione standard campionaria).
La semiampiezza dell'intervallo è

$$\frac{S_n}{\sqrt{n}} q_{1-\alpha/2} = \frac{S_n}{\sqrt{n}} \cdot 1.96.$$

Poiché essa dipende da S_n, che non è nota a priori, dobbiamo prima stimare S_n. Questo si può fare ad esempio nel modo seguente: essendo S_n^2 uno stimatore di σ^2, cerchiamo un intervallo di fiducia per σ^2, della forma $(0, a^2]$ e usiamo a come stima superiore per S_n. Per questo intervallo di fiducia, scegliamo il livello 0.95 (ma altre scelte più conservative sono possibili). L'intervallo di fiducia per σ^2 a livello $\beta = 0.05$ è

$$(0, \frac{(n-1)S_n^2}{\chi^2_{\beta,n-1}}] = (0, \frac{224 \cdot S_n^2}{190.36}]$$

e inserendo il valore numero $s_n = 30$, otteniamo l'intervallo $(0, 1059.05]$, da cui la stima superiore $a = 32.54$ per S_n. Ci aspettiamo quindi che, con alta probabilità, $\sigma \leq 32.54$. Imponiamo ora

$$\frac{a}{\sqrt{n}} \cdot 1.96 \leq 1$$

e troviamo $n \geq (a \cdot 1.96)^2 = 4067.7$. Quindi il valore minimo di n è 4068. Va poi verificato a posteriori, una volta avuti i dati, che S_n sia non superiore ad $a = 32.54$ (cosa molto probabile ma non certa).

Valori numerici utilizzabili:

$$\Phi(1.25) = 0.8944, \quad \Phi(2.5) = 0.9938,$$
$$q_{0.95} = 1.64, \quad q_{0.975} = 1.96,$$
$$t_{0.95,15} = 1.75, \quad t_{0.975,15} = 2.13,$$
$$t_{0.95,224} = 1.65, \quad t_{0.975,224} = 1.97,$$
$$\chi^2_{0.25,224} = 184.44, \quad \chi^2_{0.05,224} = 190.36, \quad \chi^2_{0.95,224} = 259.91, \quad \chi^2_{0.975,224} = 267.35$$