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.
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. L'infrazione di queste regole o la comunicazione con altri comportano l'annullamento del compito.
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) Dati tre eventi $A, B, C \subseteq \Omega$, se A è indipendente da B, B è indipendente da C e A è indipendente da C, allora A, B, C sono indipendenti tra loro.
   FALSO: basta considerare $\Omega = \{1, 2, 3, 4\}$ con probabilità discreta uniforme, $A = \{1, 2\}$, $B = \{2, 3\}$, $C = \{3, 4\}$. Ognuno di questi eventi ha probabilità $1/2$ e le intersezioni a coppie hanno probabilità $1/4$, dunque essi sono due a due indipendenti, e tuttavia la probabilità della loro intersezione (vuota) è 0, mentre se fossero indipendenti globalmente dovrebbe essere $1/8$.

   (b) Due variabili aleatorie X, Y con momento secondo finito sono scorrelate se e solo se $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.
   VERO: si è visto nel corso che $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$.
   RISPOSTE ERRATE: molti hanno risposto FALSO riportando la definizione di covarianza e assumento che il fatto che l'espressione proposta sia diversa indichi che essa non sia equivalente, quando invece a lezione è stato dimostrato rigorosamente il contrario.

   (c) Una variabile aleatoria Binomiale (per qualunque scelta dei parametri n, p) possiede tutti i momenti.
   VERO: la v.a. binomiale è limitata, quindi possiede tutti i momenti.
   RISPOSTE ERRATE: alcuni hanno sostenuto erroneamente che i momenti della Binomiale sono tutti uguali, ma questo è vero solo nel caso $n = 1$.

   (d) Il coefficiente di correlazione $\rho(X, Y)$ tra due variabili aleatorie, quando è ben definito, è un numero compreso tra 0 e 1.
   FALSO: $\rho(X, Y)$ può prendere valori nell'intervallo $[-1, 1]$, ad esempio se $Y = -X$ (e $\mathrm{var}(X) \neq 0$) vale $\rho(X, Y) = -1$.
   RISPOSTE ERRATE: molti hanno sostenuto che ρ sia una probabilità, il che non ha senso, ma non ha nemmeno senso sostenere (come fatto da altri candidati) che dal fatto che ρ non è una probabilità si possano dedurre informazioni sui valori che può assumere.

   (e) La densità di Student (con qualsiasi numero di gradi di libertà) è una funzione pari.
   VERO: è immediata conseguenza della relazione $F(t) + F(-t) = 1$ vista a lezione per la

funzione di ripartizione della variabile Student. Un modo per mostrarla è ricordare che la densità di Student corrisponde alla densità di una variabile

$$T = \frac{X_0}{\sqrt{X_1^2 + \cdots + X_n^2}}$$

in cui $(X_0, X_1, \ldots, X_n)$ sono Gaussiane standard indipendenti, ma poichè $-X_0$ ha la stessa legge (Gaussiana standard) di X_0 ed è indipendente dalle altre $X_1, \ldots, X_n$, allora $(-X_0, X_1, \ldots, X_n)$ ha la stessa legge di $(X_0, X_1, \ldots, X_n)$ e quindi

$$-T = \frac{-X_0}{\sqrt{X_1^2 + \cdots + X_n^2}}$$

ha la stessa legge di T. Sono state accettate le risposte che usavano come fatto noto la relazione di simmetria per i quantili.
RISPOSTE ERRATE: è palesemente FALSO che tutte le densità di probabilità siano pari, come anche che la parità della densità Student dipenda dal numero di gradi di libertà.

(f) Un intervallo di fiducia di livello $1-\alpha$ per un parametro θ non contiene il parametro con una probabilità che al massimo è α.
VERO: infatti per definizione di intervallo di fiducia $\mathbb{P}(\theta \in I) \geq 1-\alpha$, e quindi $\mathbb{P}(\theta \notin I) = 1 - \mathbb{P}(\theta \in I) \leq 1 - (1-\alpha) = \alpha$.

2. Un software antivirus ha due meccanismi di rilevamento di malware: il sistema A basato su *file signatures* e un sistema B basato sul comportamento. Il sistema A identifica correttamente il malware nel 95% dei casi in cui il malware è presente, ma ha anche un tasso di falsi positivi del 4% (cioè individua erroneamente un malware nel 4% dei casi in cui il malware è assente). Il sistema di rilevamento B identifica correttamente il comportamento del malware nel 91% dei casi in cui il malware è presente, ma ha anche un tasso di falsi positivi del 2%. Si sa che il malware è presente nel 3% di tutti i files da analizzare. Supponiamo inoltre che, per ogni dato file, gli esiti delle scansioni con A e con B siano indipendenti.

(a) In una scansione di un file con il sistema A, esso rileva un malware. Qual è la probabilità dell'effettiva presenza del malware?
Denotiamo con $+_A$, $-_A$ la rilevazione o meno del malware da parte di A e con S, N la presenza o assenza del malware). Per la formula di Bayes, abbiamo

$$\begin{aligned}\mathbb{P}(S \mid +_A) &= \frac{\mathbb{P}(+_A \mid S)\mathbb{P}(S)}{\mathbb{P}(+_A)} = \frac{\mathbb{P}(+_A \mid S)\mathbb{P}(S)}{\mathbb{P}(+_A \mid S)\mathbb{P}(S) + \mathbb{P}(+_A \mid N)\mathbb{P}(N)} \\ &= \frac{0.95 \cdot 0.03}{0.95 \cdot 0.03 + 0.04 \cdot 0.97} = 0.4235.\end{aligned}$$

(b) In una scansione di un file con A e con B, entrambi i sistemi rilevano un malware. Qual è la probabilità dell'effettiva presenza del malware?
Calcoliamo dapprima le probabilità di individuazione di malware dato il file, usando l'ipotesi di indipendenza:

$$\begin{aligned}\mathbb{P}(+_A \cap +_B \mid S) &= \mathbb{P}(+_A \mid S) \cdot \mathbb{P}(+_B \mid S) = 0.95 \cdot 0.91 = 0.8645, \\ \mathbb{P}(+_A \cap +_B \mid N) &= \mathbb{P}(+_A \mid N) \cdot \mathbb{P}(+_B \mid N) = 0.04 \cdot 0.02 = 0.0008.\end{aligned}$$

Quindi, sempre per Bayes, otteniamo

$$\mathbb{P}(S \mid +_A \cap +_B) = \frac{\mathbb{P}(+_A \cap +_B \mid S)\mathbb{P}(S)}{\mathbb{P}(+_A \cap +_B \mid S)\mathbb{P}(S) + \mathbb{P}(+_A \cap +_B \mid N)\mathbb{P}(N)} = 0.9709.$$

(c) Il sistema A viene testato su 400 file contenenti malware, identificando il malware con successo in 360 casi. Questa rilevazione è compatibile con l'affidabilità teorica (del sistema A) dichiarata nel testo?
Possiamo eseguire un test Z sulla probabilità p di identificare il malware quando è presente, nel caso grandi campioni, con $H_0 : p = p_0 := 0.95$, $H_1 : p \neq p_0$ (notiamo che $np_0(1-p_0) = 19 \geq 10$). Detta X la v.a. Bernoulli che indica l'identificazione o meno del malware, la statistica di test, con distribuzione circa Z sotto H_0, e la regione critica (a livello α) sono

$$Z = \sqrt{\frac{n}{p_0(1-p_0)}}(\bar{X}_n - p_0),$$
$$C = \{|Z| > q_{1-\alpha/2}\}.$$

Il p-value dei dati è ($\bar{x}_n = 360/400 = 0.9$)

$$\mathbb{P}(|Z| > \sqrt{\frac{n}{p_0(1-p_0)}}(\bar{x}_n - p_0)) = \mathbb{P}(|Z| > 4.59) \approx 0.$$

Dunque si rifiuta H_0 a ogni ragionevole livello: i dati non sono compatibili con l'affidabilità teorica dichiarata.

3. Nel primo articolo di Student (alias William Sealy Gosset), datato 1908, uno degli esempi riportati consiste in un esperimento per verificare l'effetto di un iptnotico (*hyoscyamine hydrobromide*, un'atropina ancora in uso oggi). Si era misurato su 10 pazienti l'effetto di due isomeri dell'atropina, in termini di ore di sonno in più:

- per l'isomero D si misurò una media campionaria di 0.75 ore di sonno in più, con deviazione standard campionaria di 1.70;
- per l'isomero L, media campionaria di 2.33 e deviazione standard campionaria 1.90;
- per la differenza tra isomero L e D (L-D), media campionaria di 1.58 e deviazione standard campionaria 1.17.

Dopo aver discusso le proprietà ipnotiche dei due isomeri separatamente, Student concludeva che l'isomero L era nettamente migliore.
Nel seguito, se necessario assumere che le distribuzioni dei dati siano gaussiane.

(a) Si effettuino test statistici a livello 10% per decidere se l'isomero D e quello L aumentino effettivamente in media le ore di sonno (ipotesi nulla che non ci sia aumento in media).
Siamo nel caso di test T sulla media di popolazione gaussiana, con varianza non nota. Sia X la v.a. che misura l'aumento di ore di sonno con l'isomero D, sia m_X il suo valore atteso. Testiamo $H_0 : m_X \leq m_0 := 0$ contro $H_1 : m_X > 0$. La statistica di test e la regione critica sono

$$T = \frac{\sqrt{n}}{S_n}(\bar{X}_n - m_0),$$
$$C = \{T > t_{1-\alpha,n-1} = 1.3830\}.$$

Il valore assunto da T con i dati $\bar{x} = 0.75$, $s = 1.70$ è 1.395, quindi rifiutiamo H_0 (anche se di poco): c'è evidenza, a livello 10% di un aumento medio con l'isomero D. Analogamente, per l'isomero L, il valore assunto dalla statistica è 3.878 e anche in questo caso rifiutiamo H_0: c'è evidenza di un aumento medio con l'isomero L.

(b) Si effettui un test statistico, a livello 10%, per decidere se l'isomero L sia mediamente più efficace di quello D, come dichiarato da Student (ipotesi nulla che non lo sia).
I due campioni $X_1, \dots X_n$, $Y_1, \dots Y_n$ relativi all'aumento di ore di sonno con i due isomeri sono accoppiati (poiché relativi alle stesse persone). Quindi siamo nel caso di test T per la media della differenza V di due campioni accoppiati, con varianza (della differenza) non nota. Testiamo $H_0 : m_V \leq 0$ contro $H_1 : m_V > 0$. La statistica di test e la regione critica sono (chiamando S_n^V la deviazione standard campionaria della differenza)

$$T_V = \frac{\sqrt{n}}{S_n^V}(\bar{V}_n - m_0),$$
$$C = \{T_V > t_{1-\alpha,n-1} = 1.3830\}.$$

Il valore assunto da T_V con i dati $\bar{v} = 1.58$, $s^v = 1.17$ è 4.270, quindi rifiutiamo H_0: c'è evidenza, a livello 10%, di una maggiore efficacia in media dell'isomero L.

(c) Si determini come cambiano le risposte alla domanda (a) se vengono applicati test per la media con varianza nota, assumendo che le deviazioni campionarie riportate siano invece date a priori. A livello $\alpha = 10\%$ l'uso dello Z-test (improprio, vista la numerosità piccola del campione) cambia l'esito dei test?
Con il test Z, la statistica di test e la regione critica diventano

$$Z = \frac{\sqrt{n}}{\sigma}(\bar{X}_n - m_0),$$
$$C = \{Z > q_{1-\alpha} = 1.64\}.$$

In questo caso, per l'isomero D cambia l'esito del test, che diventa l'accettazione di H_0, mentre non cambia per l'isomero L.