Esame Scritto: Appello Straordinario

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. 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.

1. Per ognuna delle seguenti affermazioni si determini se essa è VERA oppure FALSA, dimostrandola nel primo caso e fornendo un controesempio nel secondo.

   (a) Due variabili aleatorie indipendenti (di varianza finita) sono sempre scorrelate.
   VERO: come visto a lezione l'indipendenza di X, Y implica che $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, che è il caso con valori attesi nulli, da cui quello generale segue con passaggi elementari.

   (b) In un test statistico, il p-value è la probabilità che l'ipotesi nulla sia vera.
   FALSO: per definizione il p-value è la probabilità di esiti di coda della statistica su cui il test si basa, sotto l'ipotesi nulla. NON si può determinare "la probabilità che l'ipotesi nulla sia vera" nell'approccio frequentista alla statistica.

   (c) Una variabile aleatoria è discreta se e solo se assume una quantità finita di esiti.
   FALSO: è falsa la parte "solo se", perchè esistono variabili discrete con infiniti esiti di probabilità positiva, ad esempio le variabili di Poisson.

   (d) Una variabile aleatoria X con legge Gaussiana $N(0, \sigma^2)$ prende valori nell'intervallo $[-\sigma, \sigma]$ con probabilità di circa il 95%.
   FALSO: La probabilità cercata è $\Phi(1) - \Phi(-1) = 2\Phi(1) - 1 = 0.6826$.

   (e) In un test statistico, fissare un livello α è equivalente a decidere se il p-value è abbastanza alto da ritenere accettabile H_0).
   VERO: il p-value è per definizione il livello α sotto cui, fissati gli esiti dell'esperimento, si accetta H_0).

   (f) Se un campione statistico $X_1, \dots, X_n$ ha legge con momento primo finito, allora la media campionaria è uno stimatore corretto del valore atteso.
   VERO: infatti per linearità del valore atteso, e usando che le variabili del campione hanno lo stesso valore atteso,

$$\mathbb{E}[\frac{1}{n}(X_1 + \cdots + X_n)] = \frac{1}{n}(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = \mathbb{E}[X_1].$$

2. Il centro di calcolo acquista CPU da un produttore che ha due stabilimenti diversi, che chiamiamo α e β. Lo stabilimento α produce il 60% delle CPU acquistate e quello β il restante 40%. Il tempo di vita (cioè il tempo trascorso prima del primo guasto), misurato in anni, di una CPU prodotta da α ha distribuzione esponenziale di parametro 1/2, mentre il tempo di vita di una CPU prodotta da β ha distribuzione esponenziale di parametro 1/3. Il tempo di vita di una data CPU è indipendente da quello delle altre.

(a) Se una CPU viene dallo stabilimento α, qual è la probabilità che entro un anno si sia già guastata? Come cambia la risposta se la CPU viene dallo stabilimento β?
Sia T il tempo di vita della CPU. Se la CPU viene da α, la probabilità che si sia guastata entro un anno è

$$P(T \leq 1 \mid \alpha) = \int_0^1 \frac{1}{2} e^{-x/2}\, dx = 1 - e^{-1/2} = 0.393.$$

Se la CPU viene da β, questa probabilità è

$$P(T \leq 1 \mid \beta) = \int_0^1 \frac{1}{3} e^{-x/3}\, dx = 1 - e^{-1/3} = 0.283.$$

(b) Con l'ultimo acquisto il centro di calcolo ha comperato 4 CPU dallo stabilimento α. Qual è la probabilità che entro un anno se ne siano guastate già almeno 2? Come cambia la risposta se le 4 CPU vengono dal produttore β?
Sia X la v.a. che conta il numero di pezzi difettosi estratti. Sapendo che il lotto viene da α (cioè sotto $P(\cdot \mid \alpha)$), X ha distribuzione binomiale di parametri 4, 0.393. La probabilità cercata è quindi

$$\begin{aligned} P(X \geq 2 \mid \alpha) &= 1 - P(X = 0 \mid \alpha) - P(X = 1 \mid \alpha) \\ &= 1 - \binom{4}{0} \cdot (1 - 0.393)^4 - \binom{4}{1} \cdot 0.393 \cdot (1 - 0.393)^3 = 0.513. \end{aligned}$$

Se invece il lotto viene da β, X ha distribuzione binomiale di parametri 4, 0.283 e la probabilità cercata è

$$\begin{aligned} P(X \geq 2 \mid \beta) &= 1 - P(X = 0 \mid \beta) - P(X = 1 \mid \beta) \\ &= 1 - \binom{4}{0} \cdot (1 - 0.283)^4 - \binom{4}{1} \cdot 0.283 \cdot (1 - 0.283)^3 = 0.318. \end{aligned}$$

(c) Di una partita di 4 CPU acquistate un anno fa (tutte dallo stesso stabilimento), se ne sono già guastate almeno 2. Qual è la probabilità che questa partita di CPU provenisse dallo stabilimento α?
Usiamo il teorema di Bayes:

$$P(\alpha \mid X \geq 2) = \frac{P(X \geq 2 \mid \alpha) P(\alpha)}{P(X \geq 2 \mid \alpha) P(\alpha) + P(X \geq 2 \mid \beta) P(\beta)} = 0.708,$$

dove abbiamo usato $P(\alpha) = 0.6$.

3. Viene effettuato un sondaggio per stimare la percentuale di persone favorevoli a una nuova politica ambientale. Il sondaggio coinvolge 800 persone e di queste 420 si dichiarano favorevoli alla nuova politica.

(a) Fornire un intervallo di fiducia, di livello 0.95, per la percentuale di persone favorevoli (alla nuova politica ambientale).
Siamo nel caso di popolazione Bernoulli di parametro p, con campione grande. Usiamo

dunque l'approssimazione normale: detta $\bar{X}$ la frequenza relativa campionaria (delle persone favorevoli), l'intervallo di fiducia cercato è (con $1-\alpha = 0.95$)

$$[\bar{X} \pm q_{1-\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}] = [\bar{X} \pm 1.96\sqrt{\frac{\bar{X}(1-\bar{X})}{800}}]$$

Sostituendo il valore $\bar{x} = 420/800 = 0.525$, otteniamo l'intervallo numerico $[0.525 \pm 0.0346]$.

(b) I dati forniscono evidenza che più del 50% della popolazione sia favorevole (alla nuova politica ambientale)? Formulare un opportuno test di ipotesi di livello 0.01 e applicarlo ai dati del campione.
L'ipotesi nulla (non più del 50% della popolazione favorevole) è $H_0 : p \le 0.5 =: p_0$, contro $H_1 : p > 0.5$. Siamo quindi nel caso di test unilatero di livello $\alpha = 0.01$ per media di popolazione Bernoulli, con campione grande. Detta $\bar{X}$ la frequenza relativa campionaria, la regione critica è quindi

$$C = \{\sqrt{n}\frac{\bar{X}-p_0}{\sqrt{p_0(1-p_0)}} < -q_{1-\alpha}\} = \{\sqrt{800}\frac{\bar{X}-0.5}{0.5} > 2.33\}.$$

Sostituendo il valore $\bar{x} = 0.525$, troviamo

$$\sqrt{800}\frac{\bar{X}-0.5}{0.5} = 1.41 < 2.33$$

quindi accettiamo H_0: non c'è evidenza che più della metà della popolazione sia favorevole. Possiamo anche calcolare il p-value (chiamiamo Z una v.a. normale standard):

$$P(Z > 1.41) = 1 - \Phi(1.41) = 0.0793,$$

che risulta più alto del livello 0.01, a conferma che accettiamo H_0.

(c) Se vogliamo che l'intervallo di fiducia (per la percentuale di favorevoli), sempre a livello 95%, abbia semiampiezza al massimo del 3%, quando grande deve essere il campione?
La semi-ampiezza dell'intervallo di fiducia è $q_{1-\alpha/2}\sqrt{\bar{X}(1-\bar{X})/n}$. Poiché essa non è nota a priori (prima di effettuare le misurazioni), usando $\max_{x\in[0,1]} x(1-x) = 1/4$ possiamo stimare la semi-ampiezza con

$$q_{1-\alpha/2}\frac{1}{2\sqrt{n}}.$$

Imponiamo quindi che la quantità sopra sia ≤ 0.03 e troviamo

$$n \ge \left(\frac{q_{1-\alpha/2}}{2\cdot 0.03}\right)^2 = 1067.11,$$

quindi $n \ge 1068$.

Valori numerici utilizzabili:

$$\Phi(1) = 0.8413, \quad \Phi(2) = 0.9773, \quad \Phi(3) = 0.9987,$$
$$\Phi(0.70) = 0.7580, \quad \Phi(1.41) = 0.9207, \quad \Phi(2.82) = 0.9976,$$
$$q_{0.95} = 1.64, \quad q_{0.975} = 1.96, \quad q_{0.99} = 2.33, \quad q_{0.995} = 2.58.$$