

12 Giugno 2024

Esercizio 1

In questo esercizio considereremo la funzione

$$f(x) = \frac{6}{a^3}(ax - x^2)I_{[0,a]}(x)$$

dove I_A denota la funzione indicatrice dell'insieme A e $a > 0$ è un parametro della funzione. Indicheremo inoltre con X una variabile aleatoria avente f come funzione di densità di probabilità, per un valore ignoto di a .

1. Dimostrate che f è una funzione di densità di probabilità per ogni valore reale $a > 0$.
2. Calcolate il valore atteso di X esprimendola in funzione di a (suggerimento: ragionando su alcune proprietà geometriche, potete giustificare il risultato senza fare conti).
3. Indichiamo con F la funzione di ripartizione di X . Ricavate la forma analitica di $F(x)$, esprimendola in funzione di x e a .
4. Supponete, **solo in questo punto**, che $a = \frac{1}{2}$. Scrivete ed eseguite del codice che disegni il grafico della funzione F definita al punto precedente.
5. Il grafico che avete ottenuto al punto 4 potrebbe suggerire che X segue una distribuzione normale? Perché? Validate o refutate questa ipotesi.
6. Calcolate la varianza di X esprimendola in funzione di a .

Esercizio 2

In questo esercizio considereremo una popolazione X la cui distribuzione è la stessa dell'omonima variabile aleatoria introdotta nell'esercizio precedente, dove a rappresenterà un parametro incognito. Per $n \in \mathbb{N}$ fissato, X_1, \dots, X_n indicheranno delle variabili aleatorie che descrivono un campione estratto da X .

1. Dimostrate che la media campionaria è uno stimatore **distorto** per il parametro a .
2. Calcolate il bias e lo scarto quadratico medio di \bar{X} rispetto ad a , esprimendoli solo in funzione di n e a .
3. La media campionaria gode della proprietà di consistenza in media quadratica se la utilizziamo per stimare a ? Motivate la vostra risposta.
4. Dovendo scegliere se applicare il metodo plug-in o il metodo della massima verosimiglianza per ottenere uno stimatore non distorto per a , quale opzione risulta più agevole? Perché?
5. Applicate il metodo scelto al punto precedente e determinate uno stimatore T che sia non distorto per a .
6. Utilizzando il teorema centrale del limite, determinate la distribuzione approssimata dello stimatore T che avete ottenuto al punto 5.
7. Calcolate la probabilità dell'evento che si verifica quando l'errore (in valore assoluto) che si compie usando T per stimare a sia minore o uguale di 1, esprimendola in funzione di a , n e della funzione di ripartizione della distribuzione normale standard, giustificando i vostri passaggi e indicando eventuali approssimazioni che è necessario introdurre.

Esercizio 3

Collegatevi al sito upload.di.unimi.it, selezionate l'esame di Statistica e analisi dei dati per l'appello odierno e scaricate il file `risultati.csv`. Questo file contiene le seguenti informazioni raccolte da un ipotetico centro di formazione relativamente ai risultati che i propri studenti e le proprie studentesse hanno ottenuto nella tornata annuale di un test di idoneità organizzato a livello nazionale da un Ministero.

- matricola: numero di matricola;
- genere: genere (codificato come `'F'` oppure `'M'`);

- eta: età;
- punteggio: punteggio conseguito al test:
- tempo: tempo necessario per terminare il test, espresso in minuti

In questo file il carattere `,` separa le colonne.

1. Scrivete ed eseguite del codice che visualizzi su righe differenti il nome di ogni attributo unitamente al corrispondente numero di valori mancanti.
2. Di che tipo è l'attributo genere? Sulla base della risposta data, visualizzate la distribuzione di questo attributo fornendo sia una formulazione tabulare, sia un grafico, motivando le vostre scelte.
3. Considerate l'attributo punteggio, e ripetete l'analisi svolta al punto precedente, valutando se debba essere fatta nello stesso modo oppure se debbano essere utilizzati strumenti diversi.
4. Valutate l'ipotesi che vi sia una relazione tra gli attributi punteggio e tempo, specificando eventualmente il tipo e la forza della relazione determinata. Quali strumenti avete utilizzato per valutare questa ipotesi? Perché?
5. Gli esperti del centro di formazione sospettano che l'attributo punteggio dovrebbe sia ben descritto da una distribuzione analoga a quella studiata nell'Esercizio 1. Scegliete uno strumento che ha senso utilizzare per validare questa ipotesi e applicatelo, commentando i risultati ottenuti.

Esercizio 4

I valori dell'attributo punteggio nel dataset considerato al punto precedente sono espressi in una scala il cui valore massimo α non è stato reso noto, e il centro di formazione vuole stimare questo valore.

1. Sulla base della soluzione che avete proposto per l'Esercizio 2, calcolate una stima per α .
2. Utilizzare il risultato dell'Esercizio 2.7 per stimare la probabilità che la stima ottenuta al punto precedente comporti un errore (in valore assoluto) minore o uguale di 1.
3. Indichiamo con X la variabile aleatoria che descrive il punteggio ottenuto. Il test si considera sostenuto con successo se si ottiene un punteggio superiore a 21. Calcolate la frequenza di questo evento nel dataset considerato e confrontatela con la probabilità $P(X > 21)$, calcolata sostituendo al parametro α la corrispondente stima ottenuta nel punto 1 di questo esercizio, commentando i risultati ottenuti.
4. Ipotizzando che sussista indipendenza tra i punteggi ottenuti nel test da persone diverse che lo sostengono, supponiamo che tre studenti o studentesse del centro svolgano il test in una stessa tornata, e indichiamo con Y la variabile aleatoria che indica il numero di test superati. Dite quale distribuzione segue questa variabile aleatoria, e calcolate le seguenti probabilità, esprimendo anche l'evento corrispondente in termini di Y :
 - a. tre persone svolgono il test nella stessa tornata, e tutte e tre lo superano;
 - b. tre persone svolgono il test nella stessa tornata, ma solo una tra esse lo supera;
 - c. tre persone svolgono il test nella stessa tornata, e al più una tra esse lo supera.
5. Ipotizzando che sussista indipendenza tra punteggi ottenuti da una stessa persona in tempi diversi, indichiamo con Z la variabile aleatoria che indica il numero di bocciature al test prima di superarlo. Dite quale distribuzione segue Z , e calcolate la probabilità dei seguenti eventi, esprimendo anche l'evento corrispondente in termini di Z :
 - a. una persona supera il test al quarto tentativo;
 - b. una persona che non ha superato il test al secondo tentativo deve svolgerlo nuovamente almeno due volte prima di superarlo.