

Esercizio 1

In questo esercizio considereremo la funzione

$$f(x) = Kx\mathbf{I}_{\{1,\dots,a\}}(x),$$

dove \mathbf{I}_A denota la funzione indicatrice dell'insieme A , $a \in \mathbb{N}$ è un parametro della funzione e K è una costante moltiplicativa. Nel risolvere questo esercizio vi saranno utili le seguenti formule note:

$$\sum_{i=1}^n i = \frac{n(n+1)}{2},$$

$$\sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6},$$

$$\sum_{i=1}^n i^3 = \frac{n^2(n+1)^2}{4}$$

1. Determinate il valore di K espresso come sola funzione di a , che rende f una funzione di massa di probabilità.
1. Nel resto di questo esercizio indicheremo con X una variabile aleatoria avente f come funzione di massa di probabilità, per un valore ignoto di a e nella quale K è sostituito con il valore determinato al punto precedente. Calcolate il valore atteso di X , esprimendolo in funzione di a .
1. Indichiamo con F la funzione di ripartizione di X . Ricavate la forma analitica di $F(x)$, esprimendola in funzione di x e a .
1. Supponete, **solo in questo punto**, che $a = 10$. Scrivete ed eseguite del codice che disegni il grafico della funzione f ottenuta nel punto 1.
1. Il grafico che avete ottenuto al punto 4 potrebbe suggerire che X segua una distribuzione uniforme? Perché? Valutate o refutate questa ipotesi.
1. 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 ignoto. 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 .

1. Calcolate il bias e lo scarto quadratico medio di \overline{X} rispetto ad a , esprimendoli **solo** in funzione di n e a .
1. La media campionaria gode della proprietà di consistenza in media quadratica se la utilizziamo per stimare a ? Motivate la vostra risposta.
1. Supponete, **solo in questo punto**, che $a = 10$. Scrivete ed eseguite del codice che mostri l'andamento dello scarto quadratico medio ottenuto al punto precedente al variare di n .
1. Applicando il metodo plug-in, determinate uno stimatore T che sia non distorto per a .
1. Utilizzando il teorema centrale del limite, determinate la distribuzione approssimata dello stimatore T che avete ottenuto al punto 5.
1. 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.
1. Di che tipo è l'attributo `tempo` ? Sulla base della risposta data, visualizzate la distribuzione di questo attributo, motivando la scelta dello strumento grafico utilizzato.
1. Considerate l'attributo `punteggio`, e ripetete l'analisi svolta al punto precedente, valutando se debba essere fatta nello stesso modo oppure se debba essere utilizzato uno strumento diverso.
1. Valutate l'ipotesi che vi sia una relazione tra gli attributi `punteggio` ed `eta`, specificando

eventualmente il tipo e la forza della relazione determinata. Quali strumenti avete utilizzato per valutare questa ipotesi? Perché?

1. 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 ed 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 .
1. 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.
1. 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 35. Calcolate la frequenza di questo evento nel dataset considerato e confrontatela con la probabilità $p = P(X > 35)$, calcolata sostituendo al parametro α la corrispondente stima ottenuta nel punto 1 di questo esercizio, commentando i risultati ottenuti.
1. Ipotizzando che sussista indipendenza tra i punteggi ottenuti nel test da persone diverse che lo sostengono, supponiamo che cinque 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. Considerate poi gli eventi seguenti, esprimendo ognuno di essi in termini di Y e calcolate la corrispondente probabilità, sostituendo al parametro α la stima che avete precedentemente ottenuto:
 - A. nessuna persona supera il test;
 - B. esattamente due persone superano il test;
 - C. almeno una persona supera il test.
1. Nelle stesse ipotesi del punto precedente, supponiamo che a livello nazionale vi siano 3000 persone che hanno svolto il test in una stessa giornata, e indichiamo con Z la variabile aleatoria che indica il numero di test superati. Dite quale distribuzione segue Z . Considerate poi gli eventi seguenti, esprimete ognuno di essi in termini di Z e calcolate la corrispondente probabilità, sostituendo al parametro α la stima che avete precedentemente ottenuto:
 - A. tra il 50% e il 60% dei partecipanti superano il test;
 - B. al più il 50% dei partecipanti supera il test.