

luglio_2024

July 2, 2024

0.1 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:

$$\begin{aligned}\sum_{i=1}^n i &= \frac{n(n+1)}{2}, \\ \sum_{i=1}^n i^2 &= \frac{n(n+1)(2n+1)}{6}, \\ \sum_{i=1}^n i^3 &= \frac{n^2(n+1)^2}{4}\end{aligned}$$

1. Determinate il valore di K espresso come sola funzione di a , che rende f una funzione di massa di probabilità.

Per essere una funzione di massa di probabilità, la funzione $f(x)$ deve soddisfare due condizioni: - $f(x) \geq 0$ per ogni x - La somma delle probabilità per tutti i possibili valori di x deve essere uguale a 1.

Nel caso specifico della funzione $f(x) = Kx\mathbf{I}_{1,\dots,a}(x)$, la prima condizione è soddisfatta per ogni $K \geq 0$, poiché la funzione indicatrice e x sono entrambi non negativi nell'intervallo considerato.

Per determinare il valore di K che soddisfa la seconda condizione, dobbiamo calcolare la somma dei valori di $f(x)$ per tutti i possibili valori di x . Poiché la funzione indicatrice $\mathbf{I}_{1,\dots,a}(x)$ assume valore 1 solo quando x è un intero tra 1 e a incluso, la somma si riduce a:

$$\sum_{x=1}^a Kx = K \sum_{x=1}^a x = K \frac{a(a+1)}{2}$$

Affinché $f(x)$ sia una funzione di massa di probabilità, questa somma deve essere uguale a 1:

$$K \frac{a(a+1)}{2} = 1$$

Risolviendo per K , otteniamo:

$$K = \frac{2}{a(a+1)}$$

Questo è il valore di K , espresso come funzione di a , che rende $f(x)$ una funzione di massa di probabilità.

2. 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 .

Il valore atteso di una variabile aleatoria discreta si calcola come la somma dei prodotti di ogni possibile valore della variabile per la sua probabilità. Nel nostro caso, la variabile aleatoria X ha funzione di massa di probabilità

$$f(x) = \frac{2}{a(a+1)} x \mathbf{I}_{1,\dots,a}(x)$$

dove abbiamo sostituito K con il valore determinato nel punto precedente. Il valore atteso di X è quindi:

$$E[X] = \sum_{x=1}^a x f(x) = \sum_{x=1}^a x \frac{2}{a(a+1)} x$$

Portando fuori dalla sommatoria i termini costanti e semplificando, otteniamo:

$$E[X] = \frac{2}{a(a+1)} \sum_{x=1}^a x^2 = \frac{2}{a(a+1)} \cdot \frac{a(a+1)(2a+1)}{6} = \boxed{\frac{2a+1}{3}}$$

3. Indichiamo con F la funzione di ripartizione di X . Ricavate la forma analitica di $F(x)$, esprimendola in funzione di x e a .

La funzione di ripartizione $F(x)$ di una variabile aleatoria discreta X è definita come la probabilità che X assuma un valore minore o uguale a x . In formule:

$$F(x) = P(X \leq x) = \sum_{i=1}^x f(i) = \sum_{i=1}^x \frac{2}{a(a+1)} \cdot i = \frac{2}{a(a+1)} \sum_{i=1}^x i = \frac{2}{a(a+1)} \frac{x(x+1)}{2} = \boxed{\frac{x(x+1)}{a(a+1)}}$$

4. 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]: import matplotlib.pyplot as plt
import numpy as np

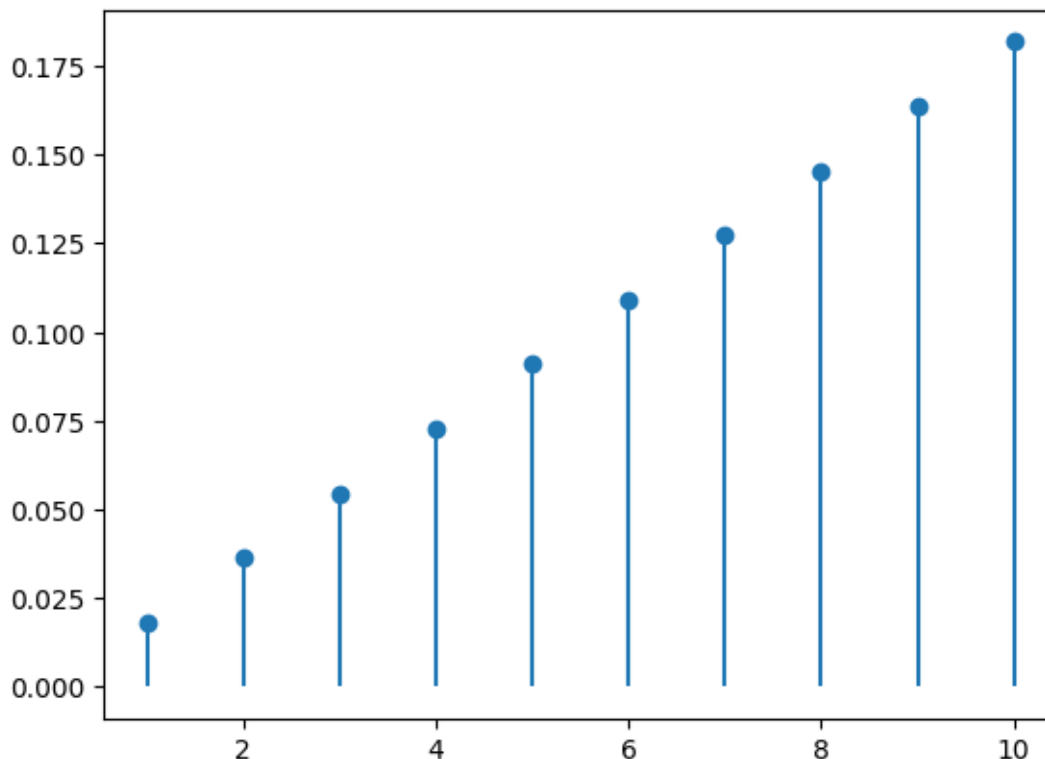
a = 10
K = 2 / (a * (a + 1))

def f(x):
    return K * x

x = np.arange(1, a + 1)
y = f(x)

plt.vlines(x, 0, y)
plt.plot(x, y, 'o')
```

```
plt.show()
```



5. Il grafico che avete ottenuto al punto 4 potrebbe suggerire che X segua una distribuzione uniforme? Perché? Valutate o refutate questa ipotesi.

Il grafico della funzione di massa di probabilità mostra che X **non** segue una distribuzione uniforme. La funzione $f(x)$ è una funzione lineare crescente, il che contraddice la caratteristica di una distribuzione uniforme, dove ogni valore all'interno dell'intervallo $\{1, \dots, a\}$ dovrebbe avere la stessa probabilità (e quindi ogni bastoncino dovrebbe avere la stessa altezza).

6. Calcolate la varianza di X , esprimendola in funzione di a .

Per calcolare la varianza, possiamo utilizzare la seguente formula:

$$\text{Var}(X) = E[X^2] - E[X]^2$$

Dobbiamo quindi calcolare sia $E[X]$ che $E[X^2]$: - $E[X]$ l'abbiamo calcolato al punto 2; - Calcolo di $E[X^2]$:

$$E[X^2] = \sum_{i=1}^a i^2 \cdot P(X=i) = \sum_{i=1}^a i^2 \cdot \frac{2}{a(a+1)} \cdot i = \frac{2}{a(a+1)} \sum_{i=1}^a i^3 = \frac{2}{a(a+1)} \cdot \frac{a^2(a+1)^2}{4} = \frac{a(a+1)}{2}$$

Infine, sostituiamo i valori trovati nella formula per la varianza:

$$\text{Var}(X) = E[X^2] - E[X]^2 = \frac{a(a+1)}{2} - \frac{(2a+1)^2}{9} = \frac{9a(a+1)}{18} - \frac{2(4a^2+4a+1)}{18} = \frac{9a^2+9a}{18} - \frac{8a^2+8a+2}{18} = \frac{a^2+9a-8a-2}{18} = \frac{a^2+a-2}{18}$$

0.2 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 .

Uno stimatore è considerato **non distorto** quando il suo valore atteso coincide con il parametro che deve stimare. Sappiamo che il valore atteso della media campionaria $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ è sempre uguale al valore atteso della popolazione, quindi:

$$E[\bar{X}] = E[X] = \frac{2a+1}{3} \neq a$$

Dal momento che $E[\bar{X}] \neq a$, 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 .

- bias:

$$b_a(\bar{X}) = E[\bar{X}] - a = \frac{2a+1}{3} - a = \frac{2a+1-3a}{3} = \boxed{\frac{1-a}{3}}$$

- scarto quadratico medio:

$$MSE_a(\bar{X}) = \underbrace{Var(\bar{X})}_n + b_a(\bar{X})^2 = \frac{a^2+a-2}{18n} + \left(\frac{1-a}{3}\right)^2 = \boxed{\frac{a^2+a-2}{18n} + \frac{(1-a)^2}{9}}$$

3. La media campionaria gode della proprietà di consistenza in media quadratica se la utilizziamo per stimare a ? Motivate la vostra risposta.

\bar{X} è consistente in media quadratica per a se $MSE_a(\bar{X}) \rightarrow 0$ per $n \rightarrow \infty$:

$$\lim_{n \rightarrow \infty} \underbrace{\frac{a^2+a-2}{18n}}_0 + \frac{(1-a)^2}{9} = \frac{(1-a)^2}{9}$$

Poiché questo limite non è zero, la media campionaria **non è consistente in media quadratica** per stimare a .

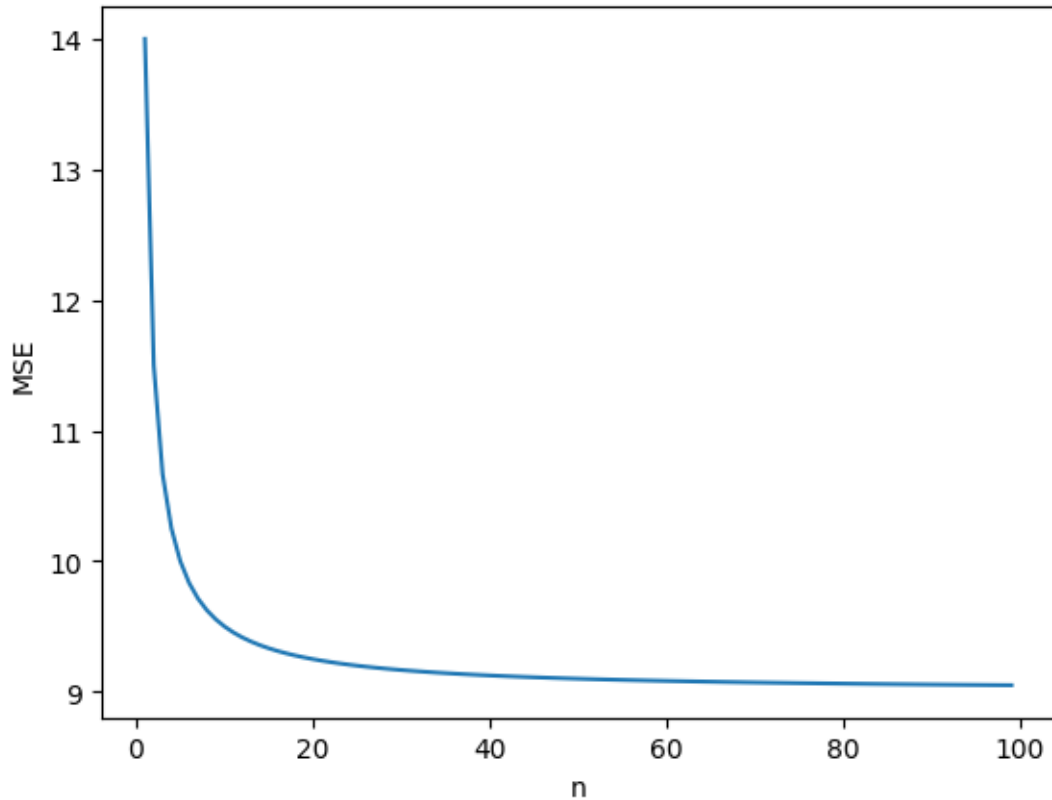
4. Supponete, **solo in questo punto**, che $a = 10$. Scrivete ed eseguite del codice che mostri l'andamento dello scarto quadratico medio ottenuto al punto precedente al variare di n .

```
[2]: import numpy as np

def MSE(x):
    return (a * (a - 1)) / (18 * x) + ((1 - a) ** 2) / 9
a = 10

x_mse = np.arange(1, 100)
y_mse = MSE(x_mse)
```

```
plt.plot(x_mse, y_mse)
plt.xlabel('n')
plt.ylabel('MSE')
plt.show()
```



5. Applicando il metodo plug-in, determinate uno stimatore T che sia non distorto per a .

$$E[X] = \frac{2a+1}{3} \Rightarrow a = \frac{3E[X] - 1}{2} = \frac{3E[\bar{X}] - 1}{2} \Rightarrow \boxed{T = \frac{3\bar{X} - 1}{2}}$$

6. Utilizzando il teorema centrale del limite, determinate la distribuzione approssimata dello stimatore T che avete ottenuto al punto 5.

Secondo il teorema centrale del limite, per n grande, la media campionaria \bar{X} è approssimativamente distribuita come una normale:

$$\bar{X} \sim N\left(\underbrace{\frac{2a+1}{3}}_{E(\bar{X})}, \underbrace{\sqrt{\frac{a(a-1)}{18n}}}_{\sqrt{Var(\bar{X})}}\right)$$

Dato che lo stimatore T è definito come: $T = \frac{3\bar{X}-1}{2}$, dobbiamo trasformare la distribuzione di \bar{X}

per ottenere la distribuzione di T : - valore atteso di T :

$$E[T] = E\left[\frac{3\bar{X} - 1}{2}\right] = \frac{1}{2}E[3\bar{X} - 1] = \frac{3}{2}E[\bar{X}] - \frac{1}{2} = \frac{3}{2}\frac{2a+1}{3} - \frac{1}{2} = \frac{2a+1-1}{2} = a$$

- varianza di T :

$$Var(T) = Var\left(\frac{3\bar{X} - 1}{2}\right) = \frac{9}{4}Var(\bar{X}) = \frac{9}{4}\frac{a^2 + a - 2}{18n} = \frac{a^2 + a - 2}{8n}$$

Pertanto, la distribuzione approssimata di T è:

$$T \sim N\left(a, \sqrt{\frac{a^2 + a - 2}{8n}}\right)$$

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.

$$P(|T - a| \leq 1)$$

Utilizziamo la distribuzione approssimata di T trovata al punto precedente. Standardizziamo:

$$\underbrace{\frac{T - a}{\sqrt{\frac{a^2 + a - 2}{8n}}}}_Z \sim N(0, 1)$$

Standardizzando a probabilità cercata diventa:

$$P\left(\left|\frac{T - a}{\sqrt{\frac{a^2 + a - 2}{8n}}}\right| \leq \frac{1}{\sqrt{\frac{a^2 + a - 2}{8n}}}\right) = P\left(|Z| \leq \frac{1}{\sqrt{\frac{a^2 + a - 2}{8n}}}\right) = P\left(-\frac{1}{\sqrt{\frac{a^2 + a - 2}{8n}}} \leq Z \leq \frac{1}{\sqrt{\frac{a^2 + a - 2}{8n}}}\right)$$

Questa è la probabilità che Z si trovi in un intervallo simmetrico intorno a zero. Utilizziamo la funzione di ripartizione della normale standard Φ :

$$P\left(-\frac{1}{\sqrt{\frac{a^2 + a - 2}{8n}}} \leq Z \leq \frac{1}{\sqrt{\frac{a^2 + a - 2}{8n}}}\right) = \Phi\left(\frac{1}{\sqrt{\frac{a^2 + a - 2}{8n}}}\right) - \Phi\left(-\frac{1}{\sqrt{\frac{a^2 + a - 2}{8n}}}\right) = 2\Phi\left(\frac{1}{\sqrt{\frac{a^2 + a - 2}{8n}}}\right) - 1$$

0.3 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 **tempo**? Sulla base della risposta data, visualizzate la distribuzione di questo attributo, motivando la scelta dello strumento grafico utilizzato.

L'attributo **tempo** è di tipo numerico discreto. Per visualizzare la distribuzione di un attributo discreto è meglio usare un grafico a barre (o a bastoncini):

3. 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.
4. 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?
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 ed applicatelo, commentando i risultati ottenuti.

0.4 Esercizio 4

I valori dell'attributo **punteggio** nel dataset considerato al punto precedente sono espressi in una scala il cui valore massimo a 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 a .
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 35. Calcolate la frequenza di questo evento nel dataset considerato e confrontatela con la probabilità $p = P(X > 35)$, calcolata sostituendo al parametro a 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 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 a la stima che avete precedentemente ottenuto:
 1. nessuna persona supera il test;
 2. esattamente due persone superano il test;

3. almeno una persona supera il test.
5. 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 a la stima che avete precedentemente ottenuto:
 1. tra il 50% e il 60% dei partecipanti superano il test;
 2. al più il 50% dei partecipanti supera il test.