

# Tema d'esame di Statistica e analisi dei dati

Prova scritta del 7 febbraio 2020

## Esercizio 0

Sia  $X$  una variabile casuale che segue una legge bernoulliana di parametro  $p$ .

1. Quali valori può assumere  $X$ ?
2. Quali valori può assumere il parametro  $p$ ?
3. Esprimete, in funzione di  $p$ , il valore atteso  $E(X)$ .
4. Completate la Figura ??(a) con il grafico di  $E(X)$  al variare di  $p$ , evidenziando in tale grafico tutte le informazioni che ritenete rilevanti.
5. Quali valori può assumere  $E(X)$ ? Giustificate la risposta.
6. Esprimete, in funzione di  $p$ , la varianza  $\text{Var}(X)$ .
7. Completate la Figura ??(b) con il grafico di  $\text{Var}(X)$  al variare di  $p$ , evidenziando in tale grafico tutte le informazioni che ritenete rilevanti.
8. Quali valori può assumere  $\text{Var}(X)$ ? Giustificate la risposta.

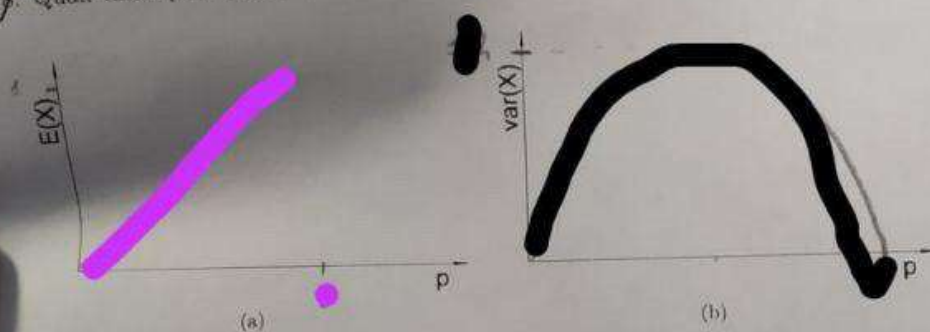


Figura 1: Grafici da completare

## Esercizio 1

Sia  $\bar{X}_{(n)}$  la media campionaria di un campione casuale  $X_1, \dots, X_n$  estratto da una popolazione bernoulliana di parametro  $p$ .

1. Esprimete  $E(\bar{X}_{(n)})$  in funzione di  $p$ .
2. Esprimete  $\text{Var}(\bar{X}_{(n)})$  in funzione di  $n$  e  $p$ .

✓ Controllate che  $\text{Var}(\bar{X}_{(n)}) \leq \frac{1}{4n}$ .

Sia  $n$  un valore abbastanza piccolo da non poter applicare l'approssimazione normale.

✓ Controllate che, per ogni  $\epsilon > 0$ , vale la disuguaglianza:  $P(|\bar{X}_{(n)} - p| \leq \epsilon) \geq 1 - \frac{1}{4n\epsilon^2}$ .

## Esercizio 2

Anche in questo esercizio  $\bar{X}_{(n)}$  è la media campionaria di un campione casuale  $X_1, \dots, X_n$  estratto da una popolazione bernoulliana di parametro  $p$ .

✓  $\bar{X}_{(n)}$  è stimatore non distorto di  $p$ ? Giustificate la risposta.

✓ Esprimete  $1 - p$  in funzione di  $E(X)$ .

✓ Determinate uno stimatore  $S_{(n)}$  del parametro  $\theta = 1 - p$ .

4. Lo stimatore trovato al punto precedente è non distorto? Giustificate la risposta.

## Esercizio 3

Collegatevi al sito [upload.di.unimi.it](http://upload.di.unimi.it), selezionate l'esame di *Statistics e analysis dei dati per l'appello odierno* e scaricate il file `mtcars.txt`. Questo file contiene, tra le altre, le seguenti informazioni riguardo al design e alle prestazioni di diversi modelli di automobili (fonte: Motor Trend US magazine, 1974):

- *modello*: identificatore univoco;
- *consumo*: espresso in km/l;
- *cilindri*: numero dei cilindri del motore;
- *cilindrata*: cilindrata (espressa in cavalli vapore);
- *peso*: peso, espresso in tonnellate;
- *test100metri*: tempo (espresso in secondi) impiegato per percorrere 100 metri partendo da fermo;
- *trasmissione*: tipo di trasmissione (0 se si tratta di trasmissione automatica, 1 se si tratta di trasmissione manuale);
- *marce*: numero di marce, senza contare la retromarcia.

In questo file il carattere di tabulazione ("`\t`") separa le colonne e i numeri reali sono stati registrati usando il carattere "`,`" come separatore dei decimali.

✓ Quanti casi contiene il dataset?

✓ Tracciate il boxplot del carattere *cilindrata*.

✓ Qual è o quali sono i modelli di auto che possono essere considerati degli outlier rispetto alla *cilindrata*?

✓ Calcolate i quartili del carattere *cilindrata*.

✓ Calcolate la distanza interquartile del carattere *cilindrata*.

✓ Tracciate un grafico, diverso dal boxplot, che secondo voi ben rappresenta la distribuzione delle *cilindrata*. Giustificate la vostra scelta.

## Esercizio 4

1. Tracciate un grafico per controllare se c'è una relazione tra il numero di cilindri e la cilindrata dell'auto.
2. Ispezionando il grafico ottenuto al punto precedente, individuate una relazione tra il numero di cilindri e la cilindrata dell'auto?
3. Utilizzate il valore di un appropriato indice numerico a supporto della vostra risposta al punto precedente.

## Esercizio 5

1. Calcolate la media del carattere *cilindrata*.
2. Calcolate la deviazione standard del carattere *cilindrata*.
3. Generate un campione casuale di 32 elementi estratto da una popolazione normale di valore atteso e deviazione standard uguali alla media e alla deviazione standard trovati nei due punti precedenti. Salvate tale campione nella variabile chiamata `valoriSimulati` (suggerimento: per generare un campione casuale di una popolazione normale si può usare il metodo `rvs` su un oggetto della classe `norm` in python e la funzione `rnorm` in R).
4. Tracciate il diagramma di dispersione tra la cilindrata e i valori simulati.
5. Ordinate in ordine crescente il campione `valoriSimulati` e salvate il risultato in una variabile chiamata `valoriSimulatiSorted`.
6. Ordinate i valori del carattere *cilindrata* in ordine crescente e salvateli in una variabile chiamata `cilindrataSorted`.
7. Tracciate il diagramma di dispersione tra i valori `valoriSimulatiSorted` e `cilindrataSorted`.
8. Rispondete alle seguenti domande, giustificando le vostre risposte e tendendo presente che i due grafici ottenuti ai punti precedenti sono uno strumento essenziale per poter formulare tali risposte.
  - (i) La cilindrata segue una legge normale?
  - (ii) Tra la cilindrata e i valori simulati esiste una relazione lineare?
9. Perché nel ragionamento fatto ai punti precedenti è importante che la variabile `valoriSimulati` sia basata su un campione di 32 elementi?

## Esercizio 6

Consideriamo ora il carattere *trasmissione*.

1. Tracciate un grafico opportuno per descrivere il carattere *trasmissione*. Giustificate la scelta fatta.
2. Consideriamo i valori osservati per il carattere *trasmissione* come la realizzazione campionaria di un campione estratto dalla popolazione  $X = \text{"tipo di trasmissione"}$ . Che legge segue la variabile casuale  $X$ ? Giustificate la risposta.

Supponiamo che il campione casuale a disposizione sia ben rappresentativo della popolazione di auto in circolazione (nel periodo di fine anni '70).

6. Stimare il valore atteso di  $X$ .
7. Lo stimatore  $T_n$  che avete utilizzato al punto precedente è non distorto? Giustificate la risposta.
8. Qual è la taglia  $n$  del campione che avete utilizzato per calcolare la stima del valore atteso di  $X$ ?
9. Calcolate la tabella delle frequenze assolute del carattere *trasmissione*.
10. Stimare la probabilità che un'auto in circolazione in quegli anni avesse trasmissione **manuale**.
11. Lo stimatore che avete utilizzato al punto precedente è non distorto? Giustificate la risposta.
12. Stimare la probabilità che un'auto in circolazione in quegli anni avesse trasmissione **automatica**.
13. Lo stimatore che avete utilizzato al punto precedente è non distorto? Giustificate la risposta.
14. Fissato  $\alpha = 0.85$ , determinate l'errore massimo commesso con probabilità maggiore o uguale a  $\alpha$ , per eccesso o per difetto, nella stima del valore atteso di  $X$ . In altre parole trovate un valore  $\epsilon$  tale che  $P(|T_n - E(X)| \leq \epsilon) \geq \alpha$ .