

Tema d'esame di Statistica e analisi dei dati

Prova scritta del 10 giugno 2019

Esercizio 0

1. Sia Y una variabile casuale uniforme discreta che assume s valori distinti.

Fissiamo, solo in questo punto, $s = 5$.

- 1.1. Tracciate a mano il grafico della funzione di massa di probabilità di Y . Siate attenti a evidenziare sul grafico tutte le informazioni importanti.

- 1.2. Tracciate a mano il grafico della funzione di ripartizione di Y . Siate attenti a evidenziare sul grafico tutte le informazioni importanti.

2. Sia X una variabile bernoulliana di parametro p .

Fissiamo, solo in questo punto, $p = 0.8$.

- 2.1. Tracciate a mano il grafico della funzione di massa di probabilità di X . Siate attenti a evidenziare sul grafico tutte le informazioni importanti.

- 2.2. Tracciate a mano il grafico della funzione di ripartizione di X . Siate attenti a evidenziare sul grafico tutte le informazioni importanti.

- 2.3. Esprimete il valore atteso e la varianza di X in funzione di p .

Esercizio 1

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

1. Esprimete, in funzione di p , il valore atteso di $\bar{X}_{(n)}$.
2. Esprimete, in funzione di p e n , la varianza di $\bar{X}_{(n)}$.
3. Proponete uno stimatore, chiamiamolo T_n , del valore atteso di X .
4. Lo stimatore che avete proposto al punto precedente è non distorto? Giustificate la risposta.
5. Proponete uno stimatore non distorto, chiamiamolo U_n , della varianza di X .

Esercizio 2

Siano $0 < \delta < 1$ e $\epsilon > 0$ due valori prefissati.

1. Controllate che la relazione:

$$P(|\bar{X}_{(n)} - p| \leq \epsilon) \geq 1 - \delta$$

può essere riscritta nel seguente modo, dove $X_{(n)}^*$ è la media campionaria standardizzata:

$$P\left(|X_{(n)}^*| \leq \frac{\epsilon}{\sqrt{p \cdot (1-p)}} \cdot \sqrt{n}\right) \geq 1 - \delta$$

2. Ipotizzate ora che sia $n \gg 1$.

Sfruttate il teorema del limite centrale e il noto fatto che la varianza di una variabile bernoulliana non può superare il valore $1/4$ per controllare che la seguente relazione:

$$P(|\bar{X}_{(n)} - p| \leq \epsilon) \geq 1 - \delta$$

equivale, fatte le dovute approssimazioni, a:

$$\Phi(2\epsilon\sqrt{n}) \geq 1 - \delta/2,$$

dove Φ è la funzione di ripartizione della distribuzione normale standard.

3. Fissati $\delta = 0.05$ e $\epsilon = 0.01$, determinate un valore numerico n_0 tale che, per $n > n_0$, sia soddisfatta la relazione del punto precedente $\Phi(2\epsilon\sqrt{n}) \geq 1 - \delta/2$. *Suggerimento: per fare i conti usate lo strumento di calcolo che avete a disposizione.*

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 `impiantitermici.csv`. Questo file contiene le seguenti informazioni raccolte dal Comune di Milano riguardo agli impianti termici installati negli edifici della città:

- `IDENTIFICATIVO_IMPIANTO`: identificatore dell'impianto;
- `GENERATORI_NUMERO`: numero di generatori dell'impianto;
- `EDIFICIO_CATEGORIA`: categoria catastale dell'edificio;
- `GENERATORE_POTENZA`: potenza del generatore, espressa in kW;
- `GENERATORE_COMBUSTIBILE`: tipo di combustibile utilizzato;
- `GENERATORE_DATA_INST`: anno di installazione;
- `RAPPORTO_DI_CONTROLLO_DATA`: anno in cui è stato fatto il controllo dell'impianto;
- `RAP_DI_CONTROLLO_ESITO`: esito del controllo;
- `ISPEZIONE_DATA`: anno in cui è stata effettuata l'ispezione dell'impianto;
- `ISPEZIONE_ESITO`: esito dell'ispezione.

In questo file il carattere ";" separa le colonne e i numeri reali sono stati registrati usando il carattere "." come separatore dei decimali.

1. Quanti casi contiene il dataset?
2. Il carattere *GENERATORE_POTENZA* è nominale, ordinale o scalare?
3. Quanti impianti sono stati installati prima del 1940?
4. Memorizzate in una variabile chiamata *selezione* i casi in cui la potenza dell'impianto di riscaldamento è tra 15 e 35 kW (estremi esclusi) e il combustibile utilizzato è il *GAS NATURALE*.
5. Quanto vale in questa selezione l'eterogeneità del carattere *GENERATORE_COMBUSTIBILE*? (Suggerimento: non è necessario fare conti)
6. Nella selezione effettuata quante sono le diverse categorie di edifici?
7. Nella selezione effettuata qual è la moda del carattere *EDIFICIO_CATEGORIA*?
8. Nella selezione effettuata quanti sono gli edifici di categoria *E1*?
9. Quale percentuale della selezione effettuata corrisponde agli edifici considerati nel punto precedente?

Esercizio 4

Memorizzate in una variabile chiamata *selezione* gli impianti installati dopo l'anno 2000 e prima del 2019, e utilizzate solo questi dati per rispondere alle domande del presente esercizio.

1. Tracciate un grafico del carattere *GENERATORE_DATA_INST* che convinca del fatto che, a partire dall'anno 2001 e fino al 2018, ogni anno è stato installato circa lo stesso numero di impianti.
2. Quanto vale in questa selezione l'eterogeneità del carattere *GENERATORE_DATA_INST*? (Suggerimento: non è necessario fare conti)
3. Calcolate la tabella delle frequenze relative del carattere *ISPEZIONE_ESITO*.
4. Calcolate la tabella delle frequenze relative del carattere *ISPEZIONE_ESITO* facendovi comparire anche i valori mancanti.
5. Qual è la percentuale di impianti ispezionati che ha ottenuto un esito positivo dell'ispezione?
6. L'impianto termico del mio condominio è stato installato qualche anno fa, ricordo che era un anno successivo al 2000. Non è ancora stato ispezionato. Si fornisca una stima della probabilità p che, quando verrà ispezionato, l'impianto ottenga un esito positivo.
7. Qual è la taglia del campione che avete utilizzato nella stima del punto precedente?
8. Quale stimatore T_n avete utilizzato per stimare p ?
9. Si fornisca una condizione sufficiente per la taglia n del campione affinché non sia inferiore a 0.95 la probabilità $P(|T_n - p| \leq 0.01)$.