CASO DI STUDIO #1 completo

Esercizio 0

Considerate una generica variable aleatoria X che assume esclusivamente i valori -1 e 1. Indichiamo con p la probabilità $\mathrm{P}(X = 1)$.

1. Sia Z una variabile bernoulliana di parametro p. Esprimete, in funzione di p, il valore atteso e la varianza di Z.

2. Definiamo al variabile $X = 2Z - 1$.

   2.1. X è una variabile discreta o continua?

   2.2. Quali valori può assumere X?

   2.3. Esprimete il valore atteso di X in funzione di Z.

   2.4. Esprimete la varianza di X in funzione della varianza di Z.

   2.5. La varianza di X è superiormente limitata. Controllate che il valore massimo che essa può assumere è 1.

Esercizio 1

1. Indichiamo con $\overline{X}_{(n)}$ la media campionaria di un campione casuale $X_1, \ldots, X_n$ estratto dalla popolazione X studiata nell'esercizio precedente.

   1.1. Esprimete il valore atteso di $\overline{X}_{(n)}$ in funzione di p.

   1.2. Esprimete la varianza di $\overline{X}_{(n)}$ in funzione di $\mathrm{Var}(X)$.

   1.3. Esprimete la varianza di $\overline{X}_{(n)}$ in funzione di p e n.

2. Controllate che $T_n = \frac{1+\overline{X}_{(n)}}{2}$ è uno stimatore non distorto per il parametro p.

3. Tramite semplici passaggi algebrici controllate che:
   $P(|T_n - p| \leq 0.05) = P(|X_n - (2p - 1)| \leq 0.1)$.

4. Indicata con Φ la funzione di ripartizione della variabile normale standard, verificate che per $n \gg 1$ vale la seguente relazione:

$$\mathrm{P}(|T_n - p| \leq 0.05) \approx 2\Phi\left(\frac{0.1\sqrt{n}}{2 \cdot \sqrt{p(1-p)}}\right) - 1.$$

5. Si controlli che la funzione $\sqrt{p(1-p)}$ ha il massimo valore nel punto $p = 1/2$.

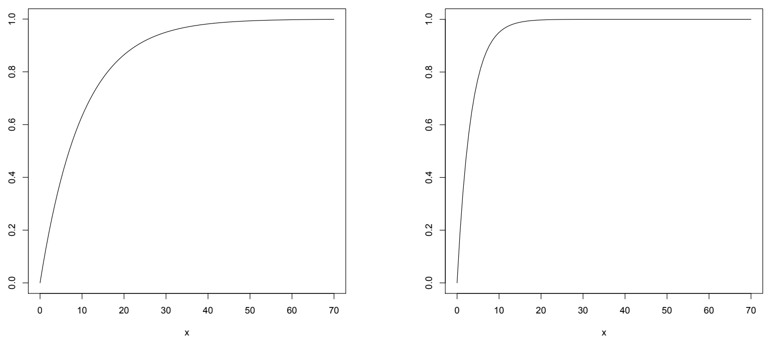
6. Si controlli che

$$\mathrm{P}(|T_n - p| \leq 0.05) \geq 2\Phi\big(0.1\sqrt{n}\big) - 1.$$

Esercizio 2

Sia G una variabile esponenziale di parametro ν.

1. Quali valori può assumere G?
2. Esprimete, in funzione di ν, la densità di probabilità f_G.
3. Fissato, solo in questo punto, $\nu = 0.1$, tracciate il grafico di f_G.
4. Di seguito In Figura 1 sono mostrati i grafici della funzione di ripartizione di due variabili esponenziali di parametri differenti. Quale dei due grafici corrisponde alla variabile di valore atteso *maggiore*? Giustificate la risposta.



(a) (b)

Figura 1: Funzione di ripartizione di due variabili esponenziali

5. Esprimete la deviazione standard σ_G di G in funzione del valore atteso $\mathrm{E}(G)$.
6. Esprimete, in funzione di ν, il valore atteso $\mathrm{E}(G)$.
7. Dato un campione casuale $G_1, \ldots, G_n$ estratto dalla popolazione esponenziale G studiata in questo esercizio, fornite uno stimatore per il parametro ν.

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 `carsharing.csv`. Questo file contiene le seguenti informazioni raccolte da un servizio di car sharing riguardo a singoli utilizzi dei veicoli della propria flotta:

- *CarIdentifier*: identificatore del veicolo;
- *TimeFrame*: fascia oraria in cui il veicolo è stato utilizzato;
- *RushHour*: indica se la fascia oraria corrisponde a un orario di punta, usando un'ovvia codifica binaria;
- *PremiumCustomer*: indica se l'utente che ha utilizzato il veicolo è iscritto al programma *Premium* (usando anche in questo caso una semplice codifica binaria);
- *Distance*: lunghezza del tragitto (espressa in km);
- *Time*: tempo impiegato a percorrere il tragitto (espresso in minuti).

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 file?
2. Analizziamo l'utilizzo del servizio di car sharing nelle diverse fasce orarie (carattere *TimeFrame*) e negli orari di maggior o minor traffico (carattere *RushHour*).
   - 2.1. Il carattere *TimeFrame* è nominale, ordinale o scalare? Giustificate la risposta.
   - 2.2. In quante fasce orarie è stata suddivisa una giornata?
   - 2.3. In quali fasce orarie il servizio di car sharing è stato maggiormente utilizzato?
   - 2.4. Calcolate la tabella delle frequenze congiunte di *TimeFrame* e *RushHour*.
   - 2.5. Leggendo la tabella calcolata al punto precedente determinate quali sono le fasce orarie che corrispondono all'ora di punta.
3. Consideriamo, solo in questo punto dell'esercizio, i clienti che hanno aderito al programma *Premium* (*Premium*=1).
   - 3.1. Quanti sono?
   - 3.2. Fornite una stima della distanza media percorsa in un tragitto da un cliente che ha aderito al programma *Premium*.
   - 3.3. Stimate la probabilità p che un nuovo cliente si iscriva al programma *Premium*.
   - 3.4. Quale stimatore avete utilizzato al punto precedente?
   - 3.5. Fornite un'approssimazione della probabilità di compiere nella stima di p un errore al più uguale a 0.05.
4. Ritorniamo a considerare il dataset completo e studiamo la distanza percorsa in ciascun utilizzo del servizio (carattere *Distance*).
   - 4.1. Il carattere *Distance* è nominale, ordinale o scalare? Giustificate la risposta.
   - 4.2. Tracciate il boxplot di tale carattere.
   - 4.3. In base all'aspetto del grafico ottenuto al punto precedente, determinate quali sono gli indici di centralità e di dispersione che meglio caratterizzano la distanza percorsa, calcolandone il valore.
   - 4.4. Riscontrate una relazione tra la distanza percorsa e il tempo impiegato? In caso affermativo, caratterizzate tale relazione. In ogni caso giustificate la vostra risposta mostrando un grafico.
   - 4.5. Calcolate l'indice di correlazione tra la distanza e il tempo. Il valore ottenuto supporta la risposta che avete dato al punto precedente?
   - 4.6. Tracciate, possibilmente nella stessa figura, il box plot della distanza nel caso di utilizzo dell'auto in orario di punta (*RushHour*=1) e in orario non di punta (*RushHour*=0).
   - 4.7. Ispezionando i due grafici ottenuti al punto precedente, dite se negli orari di punta sono privilegiati spostamenti "più brevi" oppure "più lunghi" rispetto agli orari non di punta, giustificando la risposta.
   - 4.8. Tracciate, possibilmente nella stessa figura, il box plot della distanza nel caso di utilizzo dell'auto da parte dei clienti che hanno aderito al programma *Premium* (*Premium*=1) e di quelli che non vi hanno aderito (*Premium*=-1).

4.9. Ispezionando i due grafici ottenuti al punto precedente, notate una grossa differenza nelle distanze percorse dai clienti dei due gruppi?

4.10. In Figura 2 è mostrato l'istogramma della distanza percorsa. In tale grafico si può individuare la presenza di due gruppi abbastanza distinti.

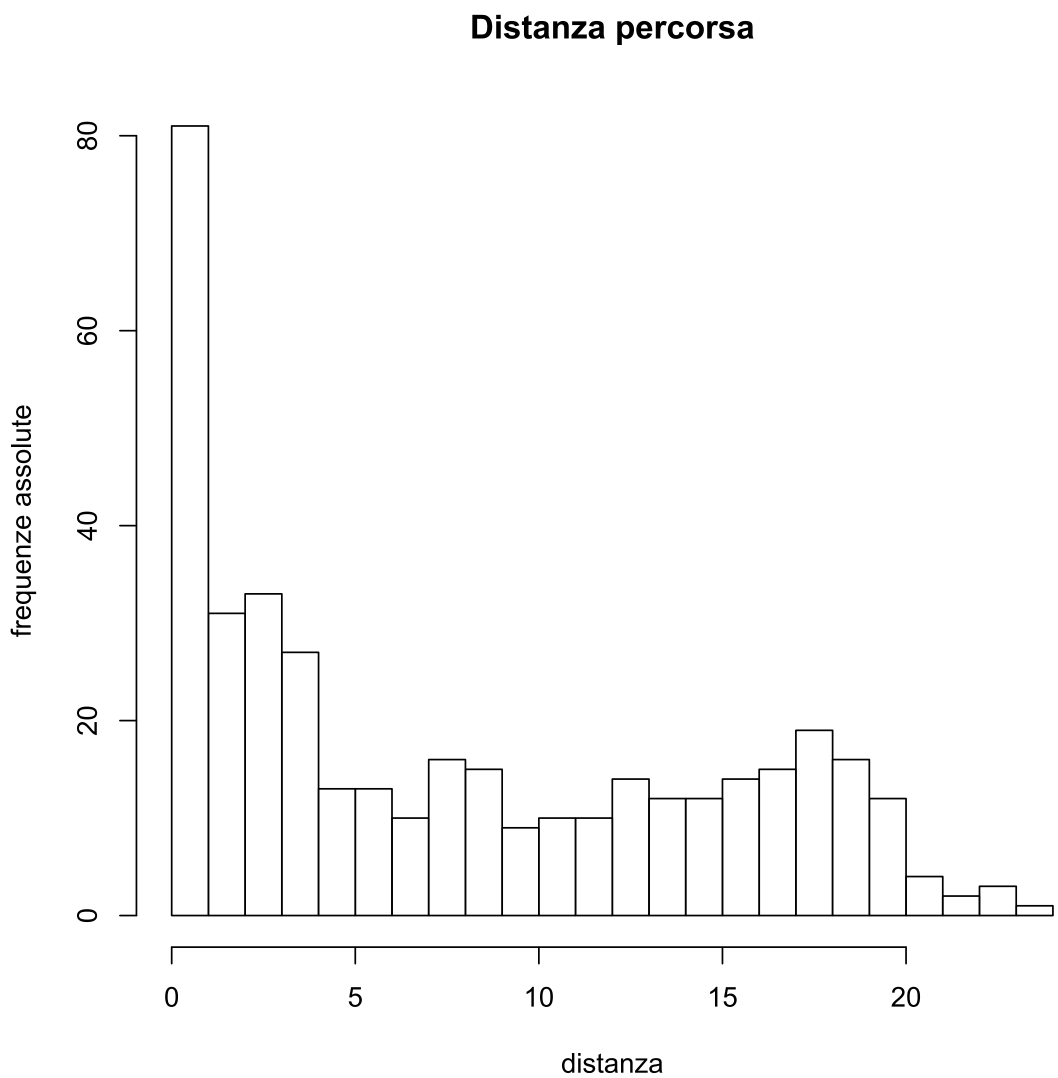


Figura 2: Istogramma della distanza percorsa

I due gruppi sono relativi al tipo di cliente (*PremiumCustomer*=1 oppure *PremiumCustomer*=-1) oppure all'orario di utilizzo del veicolo (*RushHour*=1 oppure *RushHour*=0)? In altri termini, la distanza percorsa dipende dal fatto che l'utente sia un cliente *Premium*/*non-Premium* oppure dal fatto che l'utilizzo è avvenuto in orario *Rush*/*non-Rush*? Suggerimento: per rispondere a questa domanda basta ispezionare i boxplot prodotti nei punti precedenti di questo esercizio.

4.11. Calcolate la distanza media nei due gruppi di orario (di punta/non di punta) e commentate l'istogramma di Figura 2 utilizzando queste due informazioni.

4.12. Sempre in riferimento ai due gruppi di orario (di punta/non di punta), calcolate la varianza *within groups* e la varianza *between groups*.

Esercizio 4

Analizziamo ora la distanza percorsa in ciascun utilizzo del servizio negli orari di punta (*Rush-Hour*=1)

1. Tracciate un grafico rappresentativo della distribuzione della distanza percorsa negli orari di punta.
2. È plausibile affermare che negli orari di punta la distanza segue una legge normale? Giustificate la risposta.
3. Stimate il valore atteso e la deviazione standard della distanza negli orari di punta.
4. Sapreste suggerire un modello probabilistico per la distanza percorsa negli orari di punta?
5. Le stime del valore atteso e della deviazione standard che avete appena calcolato sono compatibili con il modello che avete proposto? Giustificate la risposta.
6. Il modello probabilistico che avete proposto per descrivere la distanza percorsa negli orari di punta dovrebbe dipendere da un parametro: stimatene il valore.

Esercizio 5

Concentriamoci ora sulla distanza percorsa dai veicoli negli orari *non* di punta.

1. Tracciate un grafico opportuno che descriva la distanza percorsa negli orari *non* di punta.
2. È plausibile affermare che negli orari *non* di punta la distanza segue una legge normale? Giustificate la risposta.
3. Calcolate la media e la mediana della distanza negli orari *non* di punta e, alla luce di tali valori, commentate ulteriormente la risposta che avete dato al punto precedente.

Esercizio 6

Selezionate in una variabile chiamata `tragittibrevi` tutti i casi in cui il veicolo è stato utilizzato per percorrere un tragitto breve, cioè di lunghezza inferiore a 1.5 km.

1. Tracciate il grafico di dispersione della distanza e del tempo per i tragitti brevi.
2. Commentate il grafico che avete tracciato al punto precedente, possibilmente collegandolo al valore assunto dall'indice di variazione per il carattere *Time*.

Esercizio 7

1. Stimate la probabilità p che un'auto venga utilizzata in un orario di punta.
2. Quale stimatore avete utilizzato al punto precedente?
3. Qual è la numerosità del campione che avete a disposizione?
4. Fornite una minorazione della probabilità che nella stima di p abbiate compiuto un errore al più uguale a 0.05.

Esercizio 8

Utilizzando altre informazioni riguardo al servizio di carsharing (non presenti nel dataset che vi abbiamo fornito), si è stimato che:

(i) la probabilità che un'auto subisca un incidente è 0.15;

(ii) la probabilità che in un orario di punta un'auto subisca un incidente è 0.2.

Una data auto oggi non è disponibile perché ieri ha subito un incidente. Stimate la probabilità che l'incidente sia avvenuto in un orario di punta.