

Tema d'esame di Statistica e analisi dei dati

Prova scritta del 16 gennaio 2019

Esercizio 0

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

1. Quanto vale la probabilità $P(X = -1)$? Scrivete, sotto forma di tabella, la funzione di massa di probabilità di X .

2. Calcolate, in funzione di p , il valore atteso $E(X)$.

3. Esprimete p in funzione di $E(X)$.

4. Quali valori può assumere la variabile aleatoria $Y = g(X) = X^2$?

5. Quanto vale $E(Y)$?

6. Calcolate, in funzione di p , la varianza di X .

7. Scrivete la trasformazione $h: \mathbb{R} \rightarrow \mathbb{R}$ da applicare a X per ottenere una variabile bernoulliana $Z = h(X)$.

8. Fissiamo, solo in questo punto, $p = 0.7$.

8.1. Disegnate a mano due grafici qualitativi che descrivano rispettivamente le funzioni di massa di probabilità di X e di Z .

8.2. Disegnate a mano due grafici qualitativi che descrivano rispettivamente le funzioni di ripartizione di X e di Z .

Esercizio 1

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

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

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

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

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

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

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

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. Esprimete la deviazione standard σ_G di G in funzione del valore atteso $E(G)$.

5. Di seguito 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.

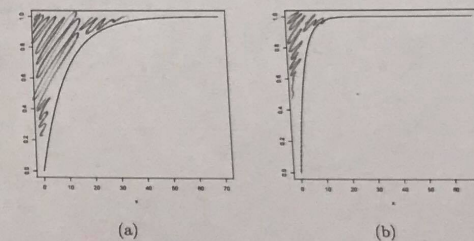


Figura 1: Funzione di ripartizione di due variabili esponenziali

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.~~ Tracciate il boxplot di tale carattere.
 - ~~4.2.~~ 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.
 - ~~4.3.~~ Ricontrate 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.
 - ~~4.4.~~ Calcolate l'indice di correlazione tra la distanza e il tempo. Il valore ottenuto supporta la risposta che avete dato al punto precedente?
5. Analizziamo ora la distanza percorsa in ciascun utilizzo del servizio negli orari di punta ($RushHour=1$).
- ~~5.1.~~ Tracciate un grafico rappresentativo della distribuzione della distanza percorsa negli orari di punta.
 - ~~5.2.~~ È plausibile affermare che negli orari di punta la distanza segue una legge normale? Giustificate la risposta.
 - ~~5.3.~~ Stimate il valore atteso e la deviazione standard della distanza negli orari di punta.
 - ~~5.4.~~ Sapreste suggerire un modello probabilistico per la distanza percorsa negli orari di punta?
 - ~~5.5.~~ Le stime del valore atteso e della deviazione standard che avete appena calcolato sono compatibili con il modello che avete proposto? Giustificate la risposta.