

# Tema d'esame di Statistica e analisi dei dati

Prova scritta del 12 febbraio 2019

## Esercizio 0

Siano  $A$  e  $B$  due eventi, e siano note le probabilità  $P(A)$  e  $P(B)$  che si verifichino rispettivamente  $A$  e  $B$ . Supponiamo inoltre di conoscere anche la probabilità  $P(A|B)$  che si verifichi  $A$  sapendo che si è verificato  $B$ .

1. Esprimete, in funzione di  $P(A)$ ,  $P(B)$  e  $P(A|B)$ , la probabilità  $P(B|A)$  che accada  $B$  sapendo che si è verificato  $A$ .
2. Considerate una variabile aleatoria  $X$  che assume esclusivamente i valori 0 e 1. Si indichi con  $p$  la probabilità  $P(X = 1)$ .
  - 2.1. Esprimete il valore atteso  $E(X)$  e la deviazione standard  $\sigma_X$  di  $X$  in funzione di  $p$ .
  - 2.2. In Figura 1 è mostrato il grafico della deviazione standard  $\sigma_X$  al variare di  $p$ .  
Per quali valori di  $p$  la deviazione standard di  $X$  assume il valore 0.3?
  - 2.3. Qual è il valore massimo che deviazione standard di  $X$  può assumere?

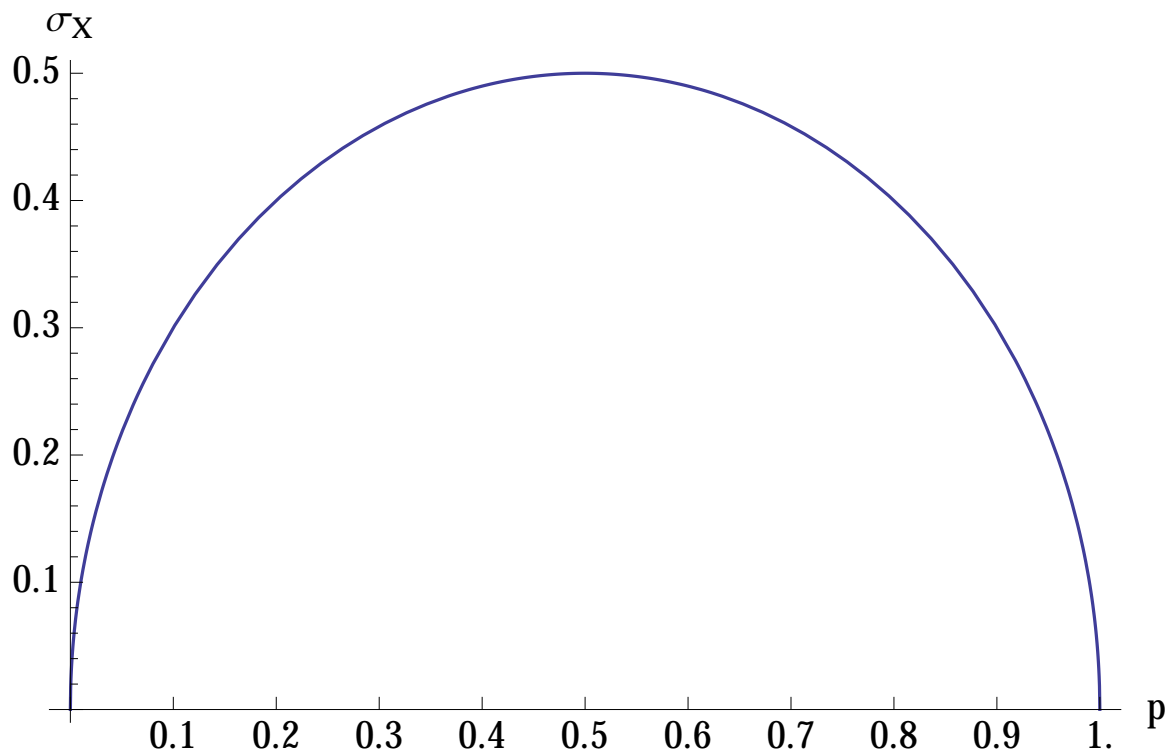


Figura 1: Deviazione standard di  $X$  al variare del parametro  $p$ .

- 2.4. Fissato, solo in questo punto,  $p = 0.5$ , tracciate a mano un grafico più dettagliato possibile della funzione di ripartizione di  $X$ .

## Esercizio 1

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

1. Esprimete, eventualmente in funzione di  $n$ , il valore assunto da  $\bar{X}_{(n)}$  nei seguenti casi:
  - 1.1. tutte le realizzazioni campionarie sono uguali a 0;
  - 1.2. esattamente due delle realizzazioni campionarie sono uguali a 1;
  - 1.3. tutte le realizzazioni campionarie sono uguali a 1.
2. Esprimete, in funzione di  $n$ , i valori che la variabile casuale  $\bar{X}_{(n)}$  può assumere.
3. La variabile casuale  $\bar{X}_{(n)}$  è uno stimatore non distorto di  $p$ ? Si giustifichi la risposta.
4. Indicata con  $\Phi$  la funzione di ripartizione della variabile normale standard, verificate che per  $n \gg 1$  vale la seguente relazione:

$$P(|\bar{X}_{(n)} - p| \leq \epsilon) \geq 2\Phi(2\epsilon\sqrt{n}) - 1.$$

## Esercizio 2

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.

In questo esercizio analizzeremo la distanza percorsa nei tragitti effettuati dagli utenti del servizio di carsharing (carattere *Distance*).

1. Il carattere *Distance* è nominale, ordinale o scalare? Giustificate la risposta.
2. 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).
3. 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.

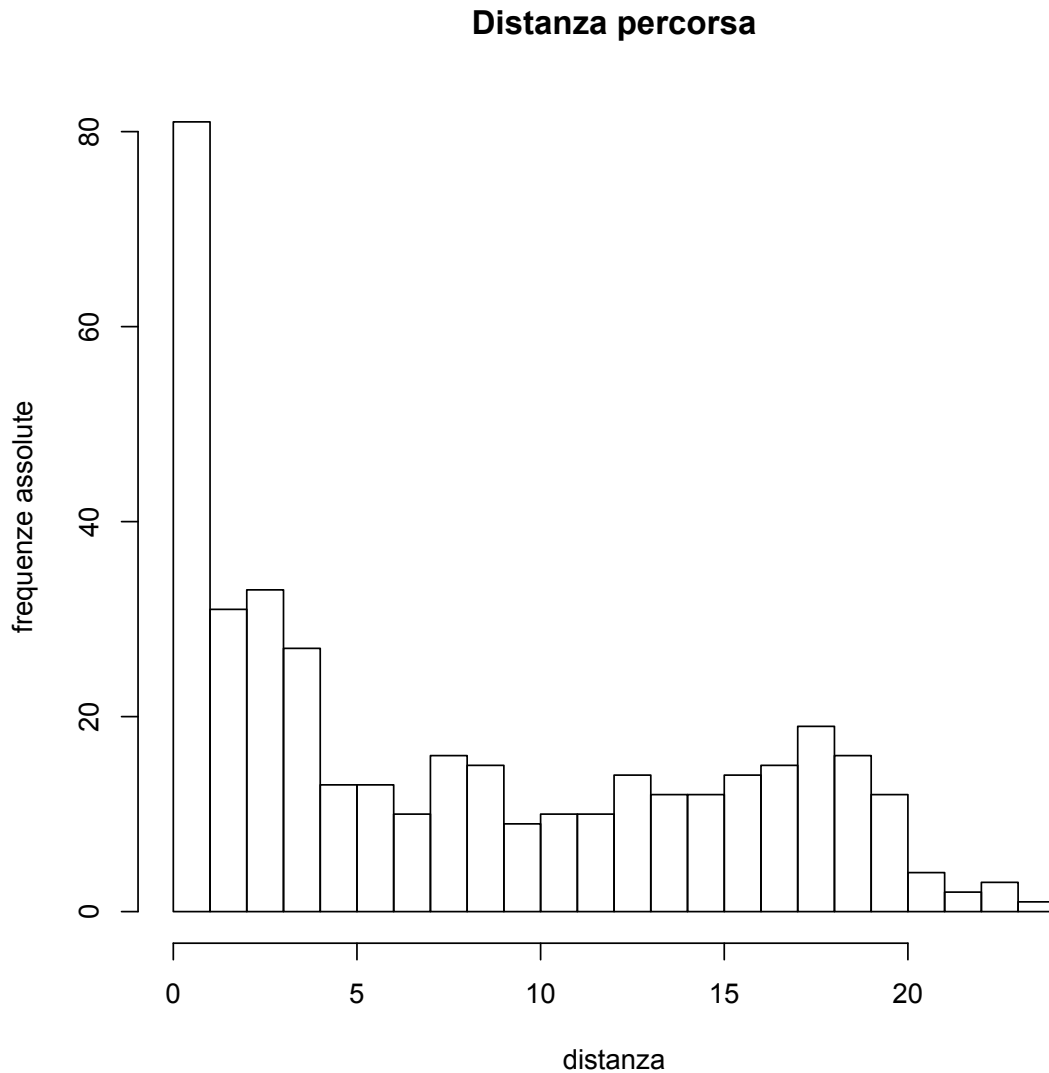


Figura 2: Istogramma della distanza percorsa

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

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

### Esercizio 3

1. Selezionate in una variabile chiamata `tragittibrevi` tutti i casi in cui il veicolo è stato utilizzato per percorrere un tragitto breve, inteso come una tratta la cui lunghezza è inferiore a 1.5 km".
2. Tracciate il grafico di dispersione della distanza e del tempo per i tragitti brevi.
3. Commentate il grafico che avete tracciato al punto precedente per concludere se, per i tragitti brevi, è riscontrabile una relazione tra la distanza e il tempo necessario per percorrerla.

## Esercizio 4

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.

## Esercizio 5

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 6

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.