

Tema d'esame di Statistica e analisi dei dati

Prova scritta del 20 gennaio 2020

Esercizio 0

Dato un evento C in uno spazio campionario Ω , sia $P(C)$ la probabilità che C si verifichi. Data una partizione A_1, A_2, \dots, A_n di Ω , supponiamo che siano note, per ogni $i = 1, \dots, n$:

- le probabilità marginali dei singoli eventi A_i , e
- le probabilità condizionate che accada C dato che è accaduto A_i .

Esprimete, in funzione delle opportune probabilità marginali e condizionate, la probabilità che accada l'evento C .

Esercizio 1

Sia X una variabile casuale di valore atteso μ e varianza σ^2 . Indichiamo con X^* la corrispondente variabile standardizzata.

- Esprimete X^* in funzione di X , μ e σ .
- Controllate che il valore atteso di X^* è uguale a 0.
- Controllate che la varianza di X^* è uguale a 1.
- Supponiamo, **solo in questo punto**, che X segua una variabile uniforme discreta con punti di massa nell'insieme $\{1, 2\}$.
 - Tracciate il grafico della massa di probabilità di X .
 - Quali valori può assumere X^* ?
 - Tracciate il grafico della massa di probabilità di X^* .

Esercizio 2

Sia X una variabile casuale **normale** di parametri μ e σ^2 . Sia $k > 0$ un valore fissato.

- Esprimete, in funzione della funzione di ripartizione F_X , la probabilità che X assuma valori compresi tra $\mu - k$ e $\mu + k$.
- Controllate che $P(\mu - k \leq X \leq \mu + k) = P(|X^*| \leq k/\sigma)$.
- Indichiamo con Φ la funzione di ripartizione di una variabile normale standard. Controllate che $P(|X^*| \leq k/\sigma) = 2\Phi(k/\sigma) - 1$.

Esercizio 3

Sia $\bar{X}_{(n)}$ la media campionaria di un campione casuale X_1, \dots, X_n estratto da una popolazione normale X di valore atteso μ e di cui è disponibile una stima della varianza pari a $\hat{\sigma}^2 = 3$.

1. Fissati i valori $\epsilon = 0.25$ e $\alpha = 0.9$, determinate una condizione sufficiente per n affinché sia maggiore di α la probabilità $P(|\bar{X}_{(n)} - \mu| \leq \epsilon)$.
2. Proponete uno stimatore, chiamiamolo T_n , di μ .
3. Lo stimatore che avete proposto al punto precedente è non distorto? Giustificate la risposta.

Esercizio 4

Collegatevi al sito `upload.di.unimi.it`, selezionate l'esame di *Statistica e analisi 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;
- *cilindrata*: cilindrata (espressa in cavalli vapore);
- *peso*: peso, espresso in tonnellate;
- *test400metri*: tempo (espresso in secondi) impiegato per percorrere 400 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.

1. Quanti e quali sono i caratteri scalari presenti nel dataset?
2. Quante osservazioni contiene il dataset?
3. Qual è la moda del carattere *peso*?
4. Qual è il modello di auto più pesante?
5. Quali sono i valori possibili per il carattere *marce*?
6. Dei caratteri *marce* e *trasmissione* visualizzate:

- 6.1. la tabella delle frequenze congiunte **assolute**;
- 6.2. la tabella delle frequenze congiunte **relative**.

7. Utilizzate le informazioni contenute nelle tabelle prodotte al punto precedente per rispondere alle seguenti domande:

1. $P(M-H < v) = F_X(\mu, v)$
- Tracciate un grafico opportuno per controllare se è plausibile l'ipotesi che il carattere *test400metri* segua una legge normale (*Suggerimento*: se dovete riscontrare dei problemi nel generare questo grafico, verificate se i dati coinvolti non contengano valori mancanti).
 - Ritenete plausibile l'ipotesi che il carattere *test400metri* segua una legge normale? Giustificate la risposta avvalendovi dei risultati trovati nei due punti precedenti.

Esercizio 7

- Selezionate in una variabile chiamata *bolidi* i modelli di cilindrata superiore a 180. In questo esercizio ci occuperemo soltanto dei bolidi.
- Quanti sono i casi selezionati?
- Calcolate il primo e il terzo quartile dei consumi dei modelli selezionati.
- Tracciate il grafico che vi sembra più adatto a comunicare l'informazione che questi modelli hanno consumi alti. Commentate la vostra scelta.

Esercizio 8

Studiamo ora la relazione tra cilindrata, prestazioni e consumi.

- Tracciate un grafico opportuno per controllare se esiste una relazione tra i caratteri *cilindrata* e *test400metri*.
- Come varia il tempo necessario per percorrere 400 metri all'aumentare della cilindrata?
- Utilizzate un opportuno indice numerico per supportare la risposta che avete dato al punto precedente (*Suggerimento*: anche in questo caso, eventuali problemi nel calcolo dell'indice potrebbero essere legati alla presenza di valori mancanti).

Esercizio 9

Possiamo considerare i valori del carattere *test400metri* come la realizzazione campionaria di un campione casuale estratto dalla popolazione $X =$ "tempo necessario per percorrere 400 metri".

- Fornite una stima del valore atteso del tempo necessario per percorrere 400 metri.
- Qual è la taglia del campione che avete utilizzato per la stima?
- Lo stimatore che avete utilizzato è non distorto? Giustificate la risposta.
- Fornite una stima della deviazione standard di X .
- Nell'ipotesi che il tempo necessario per percorrere 400 metri segua una legge normale, calcolate la probabilità che, nella stima del valore atteso di X , si commetta un errore più uguale a 0.25, in eccesso o in difetto.

- 7.1. Quanti sono i modelli di auto a 5 marce con trasmissione automatica?
- 7.2. Quanti sono i modelli di auto a 5 marce con trasmissione manuale?
- 7.3. Qual è la percentuale di modelli che hanno 5 marce e trasmissione automatica?
- 7.4. Qual è la percentuale di modelli che hanno 5 marce e trasmissione manuale?
- 7.5. Qual è la percentuale di modelli che hanno 4 marce?
- 7.6. Tra i modelli che hanno trasmissione manuale, quale percentuale ha 4 marce?

Esercizio 5

Caratterizziamo ora i vari modelli di automobile rispetto al numero di marce, cioè dividiamo l'insieme delle nostre osservazioni in tre sottoinsiemi, in cui rispettivamente le auto hanno 3, 4 oppure 5 marce.

1. Compilate la Tabella 1 delle frequenze **relative** del carattere *marce*.

Tabella 1: Tabella delle frequenze relative del carattere *marce*

	modello a 3 marce	modello a 4 marce	modello a 5 marce
Frequenza relativa			

2. Qual è la probabilità che, sorteggiando un modello di auto dal dataset, quel modello sia a 4 marce?
3. In Tabella 2, per ogni categoria di auto (a 3, 4 oppure 5 marce) è mostrata la probabilità che l'auto abbia consumi alti (queste probabilità sono state stimate tramite un'analisi precedente). Fornite una stima della probabilità che, sorteggiando un modello di auto dal dataset, quel modello abbia consumi alti.

Tabella 2: Consumi rispetto al numero di marce.

	modello a 3 marce	modello a 4 marce	modello a 5 marce
probabilità di alti consumi	0.8	0.17	0.6

Esercizio 6

1. Tracciate l'istogramma del tempo impiegato a percorrere i 400 metri.
2. Per il tempo impiegato a percorrere i 400 metri calcolate:
 - 2.1. due indici di posizione centrale,
 - 2.2. due indici di dispersione.