

Tema d'esame di Statistica e analisi dei dati

Prova scritta del 15 settembre 2017

Esercizio 0

Sia X una variabile aleatoria di valore atteso μ e varianza σ^2 . Siano A e B due variabili aleatorie indipendenti distribuite come X .

1. Calcolate il valore atteso di $A - B$.
2. Esprimete in funzione di σ^2 la varianza di $A - B$.
3. Supponiamo, solo in questo punto, che A e B siano variabili aleatorie normali. In questo caso che distribuzione segue $A - B$? Specificatene i parametri.

Sia $\{X_1, X_2, \dots, X_n\}$ un campione casuale estratto dalla popolazione X .

4. Si proponga uno stimatore per la varianza di X .
5. Lo stimatore proposto al punto precedente è non distorto? Si giustifichi la risposta.
6. Si proponga uno stimatore per la varianza di $A - B$.
7. Lo stimatore proposto al punto precedente è non distorto? Si giustifichi la risposta.

Esercizio 1

Collegatevi al sito upload.di.unimi.it e selezionate l'esame di *Statistica e analisi dei dati*.

Scaricate il file `sf-parks.csv`, che contiene la descrizione dei parchi pubblici della città di San Francisco (Fonte: open data della città di San Francisco, <https://datasf.org/opendata/>).

Gli attributi del dataset che consideriamo sono:

- *ParkName*: nome,
- *ParkType*: tipologia,
- *ParkServiceArea*: zona di appartenenza,
- *PSAManager*: responsabile della zona di appartenenza,
- *email*: indirizzo di posta elettronica del responsabile,
- *Number*: numero di telefono del responsabile,
- *Zipcode*: codice di avviamento postale,
- *Acreage*: estensione (in acri),
- *SupDist*: identificativo di zona,
- *ParkID*: identificativo univoco,
- *Address*: indirizzo,
- *Latitude*: latitudine,
- *Longitude*: longitudine,
- *Distance.from.downtown*: distanza dal centro città (in miglia).

1. Importate i dati (tenendo presente che i valori sono separati dal carattere "," e il separatore delle cifre decimali nei valori a virgola mobile è ".").

2. Elencate gli attributi numerici.
3. Dite quanti parchi sono presenti nel dataset.
4. Indicate quanti e quali sono i valori osservati per l'attributo *ParkServiceArea*.
5. Visualizzate il grafico della funzione cumulativa empirica per l'attributo *Acreage*.
6. Completate la seguente frase: “la metà dei parchi di San Francisco ha una estensione maggiore di acri”.
7. Qual è l'estensione media (espressa in acri) di un parco di San Francisco?
8. Quali osservazioni si possono trarre confrontando le risposte date alle domande 6 e 7?
9. Quanti parchi di San Francisco hanno una estensione minore di 50 acri?

Esercizio 2

1. Calcolate la media e la varianza di *Latitude* e *Longitude*.
2. Sfruttate i risultati dell'Esercizio 0 per fornire:
 - 2.1. una stima della varianza della differenza di *Latitude* tra due parchi di San Francisco;
 - 2.2. una stima della varianza della differenza di *Longitude* tra due parchi di San Francisco.
3. Utilizzate la tecnica del Q-Q plot per validare o confutare le seguenti ipotesi:
 - 3.1. il carattere *Latitude* è distribuito secondo una legge normale;
 - 3.2. il carattere *Longitude* è distribuito secondo una legge normale.
4. Che legge segue la differenza di *Longitude* tra due parchi di San Francisco? Specificatene i parametri.
5. Visualizzate il diagramma di dispersione per gli attributi *Latitude* e *Longitude* e calcolate l'indice di correlazione tra i medesimi attributi.
6. Utilizzate i risultati ottenuti per validare o confutare l'ipotesi che vi sia una forte dipendenza tra *Latitude* e *Longitude*.
7. Visualizzate il diagramma di dispersione per gli attributi *Distance.from.downtown* e *Acreage*.
8. Il grafico che avete generato al punto precedente evidenzia la presenza a San Francisco di alcuni parchi molto più estesi di altri (e di un parco molto più lontano dal centro della città). Memorizzate nella variabile `ParchiCitta` il sottinsieme del dataset che non contiene gli outlier identificati.
9. Visualizzate i boxplot di *Latitude* e *Longitude* del dataset iniziale e del sottinsieme `ParchiCitta`. Confrontateli commentando i risultati.

Esercizio 3

Ipotizziamo che la geografia della California sia fatta in modo tale che la differenza tra la *Latitude* di due punti nel suo territorio sia trascurabile. Ciò significa che è possibile esprimere approssimativamente la distanza tra due punti come la differenza tra i corrispondenti valori di *Longitude*. Chiamiamo *distanza longitudinale relativa* questa differenza (considereremo quindi, per semplicità, distanze con segni positivi e negativi) e misuriamo i suoi valori in gradi.

1. Consideriamo i parchi di San Francisco come un campione casuale estratto dalla popolazione dei parchi delle grandi città della California. Indicate con A e B le variabili aleatorie che corrispondono alle longitudini di due parchi della California, calcolate la probabilità $P(A - B < 0.1)$ che la distanza longitudinale relativa tra i due parchi sia minore di 0.1 gradi (potete approssimare la varianza della longitudine con la stima trovata nell'Esercizio 2).
2. Calcolate la probabilità che la differenza tra le due longitudini sia, in valore assoluto, minore di 0.1 gradi.