

Tema d'esame di Statistica e analisi dei dati

Prova scritta del 6 luglio 2017

Esercizio 0

Tabella 1: Percentili della variabile aleatoria X

decimo	trentesimo	cinquantesimo	settantesimo	novantesimo
15	55	200		

Sia X una variabile aleatoria normale.

1. Completare la Tabella 1.
2. Tracciare un grafico indicativo ma il più preciso possibile della funzione di ripartizione di X .

Esercizio 1

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

Scaricate il file `BibliotecheQuartiere.txt`, che contiene dati relativi alle biblioteche rionali di Milano. (Fonte: open data del comune di Milano, <http://www.datiopen.it/it/catalogo-opendata/arte-cultura/>). Gli attributi del dataset che consideriamo sono:

- *Anno*: anno
- *Biblioteche*: nome della biblioteca
- *Libri*: numero di libri acquisiti nell'anno
- *Adulti.Iscritti*: numero di nuovi utenti maggiorenni
- *Ragazzi.Iscritti*: numero di nuovi utenti minorenni
- *Totale.Iscritti*: numero totale di nuovi utenti

1. Importare i dati (tenendo presente che il separatore di decimali per i numeri è la virgola e i valori sono separati dal carattere ";") e dire quanti casi sono presenti nel dataset.
2. Da quale anno a quale anno sono stati raccolti i dati?
3. Quante sono le biblioteche rionali presenti nel dataset? Elencarne i nomi.
4. 4.1. Tracciare il grafico che si ritiene più opportuno per descrivere il numero di ragazzi che si iscrivono in biblioteca all'anno. Il grafico deve avere il titolo "RAGAZZI" e sull'asse opportuno (a seconda del grafico che scegliete) deve apparire l'etichetta "numero iscrizioni annuali per biblioteca".
4.2. Tracciare un grafico analogo che descriva il numero di adulti che si iscrivono in biblioteca all'anno.
5. 5.1. Calcolare la media, la deviazione standard e il coefficiente di variazione del numero di ragazzi che si iscrivono in biblioteca all'anno.
5.2. Fare lo stesso per il numero di adulti.
5.3. Confrontare la variabilità del numero di iscrizioni di ragazzi rispetto a quella di adulti.

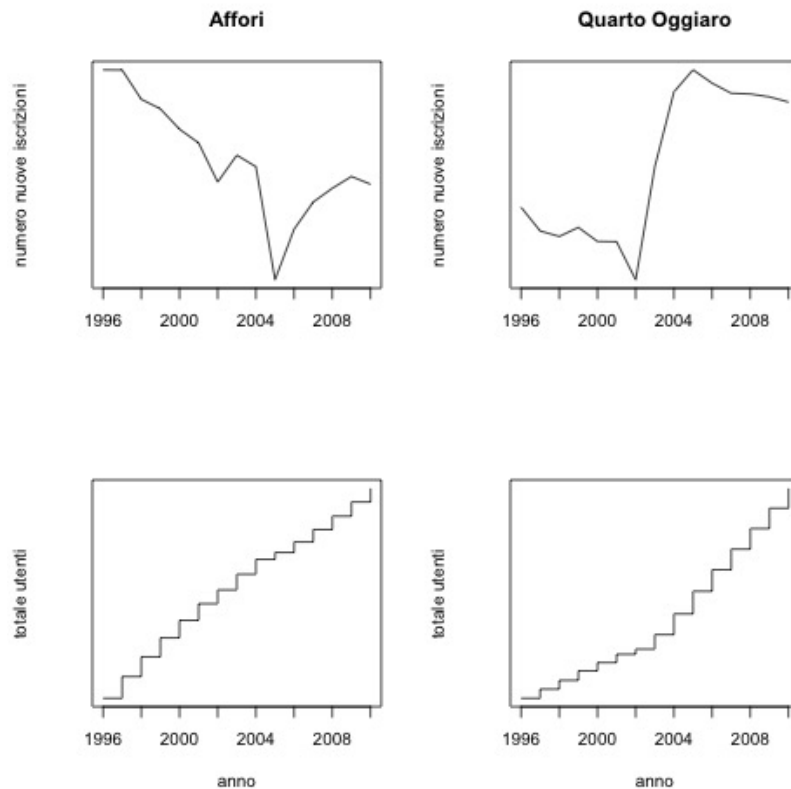


Figura 1: Numero di nuovi iscritti per anno e numero complessivo di utenti

Esercizio 2

Concentriamoci sull'anno 2000.

1. Tracciare, possibilmente nella stessa figura, il boxplot del numero di ragazzi che si sono iscritti e del numero di adulti che si sono iscritti a una biblioteca rionale di Milano (nell'anno 2000).
2. Utilizzare il risultato del comando `summary` per rispondere alle seguenti domande:
 - 2.1. nell'anno 2000 quale percentuale (circa) di biblioteche ha avuto più di 300 nuovi ragazzi iscritti?
 - 2.2. nell'anno 2000 quale percentuale (circa) di biblioteche ha avuto più di 950 nuovi iscritti?

Esercizio 3

Concentriamoci ora sul servizio Bibliobus.

1. Creare una variabile che contiene i soli casi del dataset che si riferiscono alla biblioteca Bibliobus.
2. Calcolare la tabella delle frequenze congiunte tra l'anno e il totale di nuovi iscritti al Bibliobus.
3. Tracciare il grafico di dispersione dei caratteri *Anno* e *Totale.Iscritti*. Siccome in questo caso i dati sono ordinati per anno crescente, rigenerare il grafico di dispersione collegando ciascun punto al successivo tramite una linea spezzata, al fine di evidenziare una tendenza.
4. Commentare, anche avvalendosi di strumenti formali, la seguente affermazione: "si può notare che nel corso degli anni c'è stato un incremento, seppur modesto, del numero di iscrizioni al servizio Bibliobus".

Prendiamo ora in considerazione le biblioteche Affori e Quarto Oggiaro.

5. In Figura 1 sono mostrati nella parte alta i grafici del numero totale di nuovi utenti in ciascun anno e nella parte bassa la funzione cumulativa, che indica quindi il totale degli utenti della biblioteca in ciascun anno. I grafici della parte bassa della figura sono in ordine giusto? Cioè ciascuno corrisponde al grafico delle frequenze soprastante? Giustificate la risposta.

Esercizio 4

1. Tracciare il boxplot oppure l'istogramma (se uno dei due grafici vi sembra più rappresentativo) del numero di libri acquisiti in un anno da una biblioteca.
2. I grafici del punto precedente rivelano la presenza di alcuni outlier. Questi valori sono tutti relativi alla **Biblioteca Centrale Sormani**. Tracciare il boxplot oppure l'istogramma (se uno dei due grafici vi sembra più rappresentativo) del numero di libri acquisiti in un anno da una biblioteca, escludendo però la **Biblioteca Centrale Sormani**.
3. Se si esclude la **Biblioteca Centrale Sormani** si vede che il numero di libri acquisiti annualmente da una biblioteca ha un andamento "a campana":
 - 3.1. determinare i parametri di tale distribuzione;
 - 3.2. utilizzare la tecnica del qqplot per controllare se anche il numero di libri acquisiti annualmente dalla sola **Biblioteca Centrale Sormani** segue una legge normale.

Esercizio 5

1. Si X una variabile aleatoria e supponiamo di aver stimato in precedenza la sua deviazione standard $\sigma = 90$. Sia X_1, \dots, X_n un campione casuale di taglia n estratto da X . Consideriamo la media campionaria $\sum_{i=1}^n X_i/n$ come stimatore del valore atteso μ_X di X .

1.1. Si esprima, esclusivamente in funzione di n , la deviazione standard di $\sum_{i=1}^n X_i/n$.

1.2. Si controlli che, per $n \rightarrow \infty$, l'espressione:

$$P(-100 < \sum_{i=1}^n X_i/n - \mu_X < 100) = 0.9$$

equivale all'espressione:

$$P(|Z| < 10/9 \cdot \sqrt{n}) = 0.9$$

dove Z è una variabile aleatoria normale standard.

1.3. Si controlli che l'espressione precedente equivale a:

$$P(Z < 10/9 \cdot \sqrt{n}) = 0.95$$

2. Supponiamo che a Milano, in un quartiere di recente edificazione, apra una nuova biblioteca rionale. Utilizziamo la media campionaria per stimare il numero atteso, chiamiamolo μ , di iscritti a questa nuova biblioteca per il prossimo anno. Sulla base delle osservazioni a disposizione (escludendo sempre la **Biblioteca Centrale Sormani**):
 - 2.1. dire quanto è numeroso il campione fornire una approssimazione della probabilità che nella stima di μ si compia un errore inferiore, per eccesso o per difetto, a 100 unità;
 - 2.2. se nella stima volessimo tollerare di compiere un errore al più di 100 unità con probabilità 0.9, quale sarebbe un numero sufficiente di osservazioni che dovremmo avere a disposizione?