

# Tema d'esame di Statistica e analisi dei dati

Prova scritta del 17 febbraio 2016

## Esercizio 0

Consideriamo le condizioni meteorologiche a Milano e per tutta la prossima settimana, a partire da lunedì. In particolare, dato  $i = 1, \dots, 7$  indichiamo con  $X_i$  la variabile aleatoria che assume il valore 0 se nell' $i$ -esimo giorno della settimana non ci saranno precipitazioni, mentre varrà 1 se in quel giorno dovesse piovere. Ipotizziamo inoltre che le previsioni meteorologiche per la settimana entrante ci permettano di dire che la probabilità che piovra in un generico giorno sia uguale a un valore fissato  $p \in (0, 1)$ . Assumiamo anche che il fatto che piovra (o che non piovra) in un dato giorno non influenzi le condizioni meteo nei giorni rimanenti.

1. Scrivere in funzione di  $p$  la legge di probabilità di  $X_i$  per un generico  $i$ .

Indichiamo rispettivamente con  $E_1$  ed  $E_2$  gli eventi “piove in almeno un giorno infrasettimanale” e “non piove nel fine settimana”.

2. Esprimere, in funzione di  $p$ , le probabilità di  $E_1$  e di  $E_2$ .
3. Gli eventi  $E_1$  e  $E_2$  sono indipendenti? Sono mutuamente esclusivi?

Fissiamo, solo in questo esercizio,  $p=0.4$ .

4. Calcolare la probabilità degli eventi  $E_1$ ,  $E_2$ ,  $E_3$  = “piove in almeno un giorno infrasettimanale ma non durante il fine settimana” ed  $E_4$  = “piove in almeno un giorno infrasettimanale oppure non piove nel fine settimana”.

## Esercizio I

Indichiamo con  $B$  la variabile aleatoria che conta il numero di giornate piovose in una settimana.

1. Quali sono i punti di massa di  $B$ ?
2. Scrivere in funzione di  $p$  la legge di probabilità di  $B$ .
3. Sono sufficienti 7 giorni di osservazione per poter affermare che è maggiore di 0.4 la probabilità di compiere nella stima di  $p$  un errore inferiore a 0.25? (Si ricordi che  $p(1-p)$  non può mai superare il valore  $1/4$ ).

## Esercizio II

Collegatevi al sito [upload.di.unimi.it](http://upload.di.unimi.it) e selezionate l'esame di *Statistica e analisi dei dati*; da qui scaricate il file `DATI AMBIENTE.txt`, che contiene informazioni sui valori della concentrazione media di vari inquinanti nell'aria registrati durante l'anno 2015 da diverse centraline presenti nelle province della Lombardia (Fonte: Regione Lombardia). Il file contiene una riga per centralina, e registra le informazioni seguenti:

- la provincia;
- la concentrazione di inquinante per le seguenti sostanze:

– C6H6: benzene,	– O3: ozono,	– As: arsenico,
– SO2: anidride solforosa,	– PM10: polveri sottili $10\mu$ ,	– Ni: nichel,
– CO: monossido di carbonio,	– PM2.5: polveri sottili $2.5\mu$ ,	– Cd: cadmio,
– NO2: diossido di azoto,	– Pb: piombo,	– BaP: benzoapirene;
- per l'ozono:
  - il numero di giorni nei quali è stata superata la soglia di tolleranza,
  - il numero di giorni nei quali per un dato inquinante è stata superata la soglia di allarme.

Tabella 1: Indici centrali e di dispersione

	<b>C6H6</b>	<b>SO2</b>	<b>CO</b>
media			
mediana			
deviazione standard			
distanza interquartile			

1. Quante centraline sono presenti in Lombardia?
2. Come sono distribuite le centraline nelle diverse province della Lombardia? Rispondere con un grafico.
3. Quale provincia è meno rappresentata? Qual è la più rappresentata?
4. Calcolare l'indice di eterogeneità per il carattere PROVINCIA e commentate il risultato ottenuto (che cosa mi dice sul carattere PROVINCIA il valore di tale indice?).

Consideriamo la concentrazione degli inquinanti C6H6, SO2 e CO.

5. Per ciascuno di essi indicare in Tabella 1 media, mediana, deviazione standard e distanza interquartile.
6. Mettere a confronto i tre inquinanti tramite un'opportuna modalità grafica, commentando il grafico ottenuto.

Degli elementi cadmio e benzoapirene abbiamo perso i dati ma ci sono rimasti i grafici della funzione cumulativa empirica e della frequenza relativa, illustrati in Figura 1.

7. Completare la figura con le corrispondenze appropriate.

Consideriamo la concentrazione di ozono: O3.

8. Calcolare la tabella delle frequenze relative della concentrazione annuale.
9. Tracciare il grafico della funzione cumulativa empirica di O3.
10. Quante centraline hanno registrato una concentrazione uguale a 46 mg/mcubo?
11. Quante centraline hanno registrato una concentrazione compresa tra i 45 e i 50 mg/mcubo, estremi inclusi?
12. Indicare quali valori di O3 sono “particolarmente” elevati rispetto agli altri, motivando la risposta con un grafico.

Consideriamo la variabile casuale  $X$  = numero dei giorni in un anno nei quali viene superata la soglia di tolleranza per l'ozono. Abbiamo a disposizione il campione casuale  $X_1, \dots, X_n$  di taglia  $n = 150$  estratto dalla popolazione  $X$  (alla centralina  $i$  è associata la variabile  $X_i$ ) e memorizzato nel file scaricato.

13. Si fornisca una stima del valore atteso e della deviazione standard di  $X$ .
14. Ci sono buoni motivi per pensare che  $X$  segua la legge binomiale? Si giustifichi la risposta.

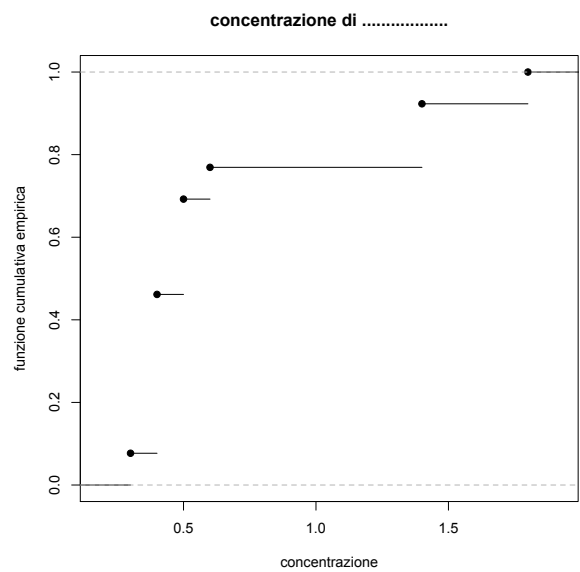
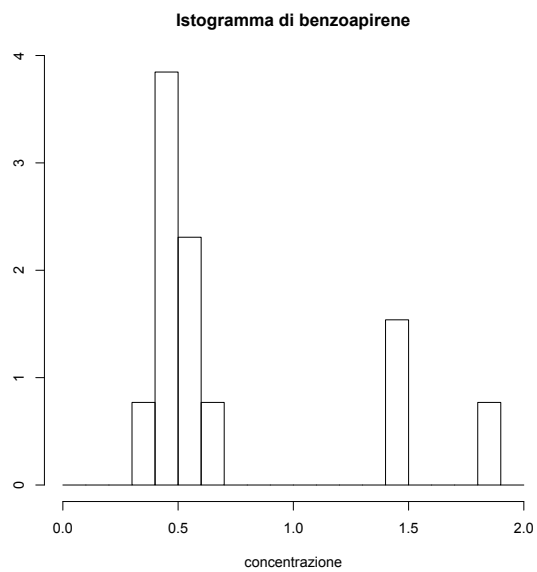
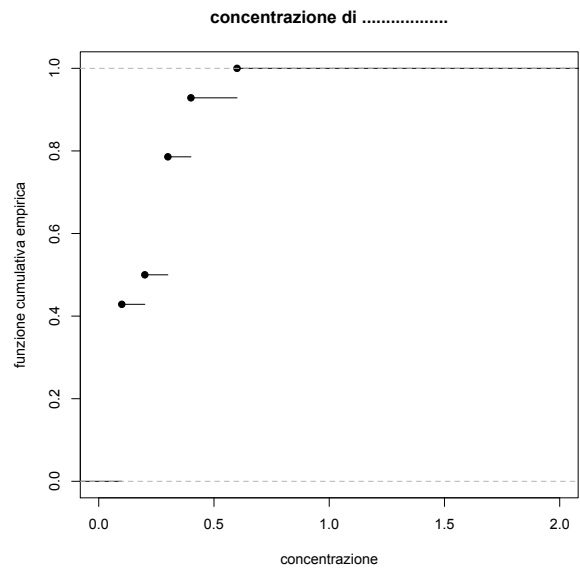
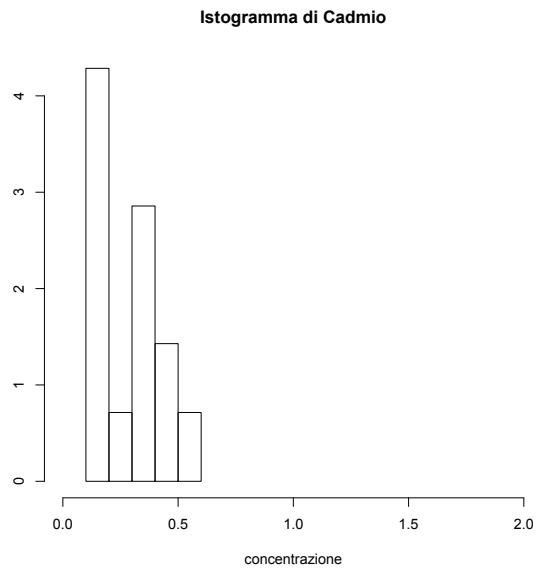


Figura 1: Cadmio e Benzoapirene: *which is which?* Quale funzione cumulativa empirica corrisponde a quale distribuzione di frequenza?