

# Tema d'esame di Statistica e analisi dei dati

Prova scritta del 2 luglio 2018

## Esercizio 0

1. In Figura 1 è mostrata la funzione di ripartizione di due variabili casuali  $X$  e  $Y$ . In Figura 2 è mostrata la funzione densità di probabilità delle medesime due variabili, ma non è precisato quale grafico competa a quale variabile. Si completi la figura con il riferimento alla variabile opportuna.
2. Quale delle due variabili ha valore atteso maggiore? Si giustifichi la risposta.
3. Il valore 2 a quale percentile di  $X$  corrisponde? A quale percentile di  $Y$  corrisponde?
4. Qual è il cinquantesimo percentile di  $X$ ? Qual è il cinquantesimo percentile di  $Y$ ?
5. Quanto vale la probabilità  $P(2 < X \leq 5)$ ? Quanto vale la probabilità  $P(2 < Y \leq 5)$ ?
6. Per la variabile  $Y$  si dica se la media è maggiore, minore oppure uguale alla mediana.

## Esercizio 1

Sia  $X$  una variabile casuale esponenziale di parametro  $\nu$ .

1. Si esprima, in funzione  $\nu$ , la funzione densità di probabilità di  $X$ .
2. Si esprimano, in funzione  $\nu$ , il valore atteso e la deviazione standard di  $X$ .

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

3. Si proponga uno stimatore, chiamiamolo  $T_n$ , del valore atteso di  $X$ .
4. Si proponga uno stimatore, chiamiamolo  $R_n$ , del parametro  $\nu$ .

## Esercizio 2

Presso un ambulatorio veterinario un gruppo di cani viene seguito regolarmente, e a oggi la situazione dello stato di salute dei pazienti a quattro zampe è documentata in un file che contiene le seguenti informazioni:

- *Cartella*: numero della cartella clinica,
- *IP*: indica se il paziente soffre di ipertensione,
- *GravitaIP*: gravità dell'ipertensione,
- *EtaAnni*: età (espressa in anni),

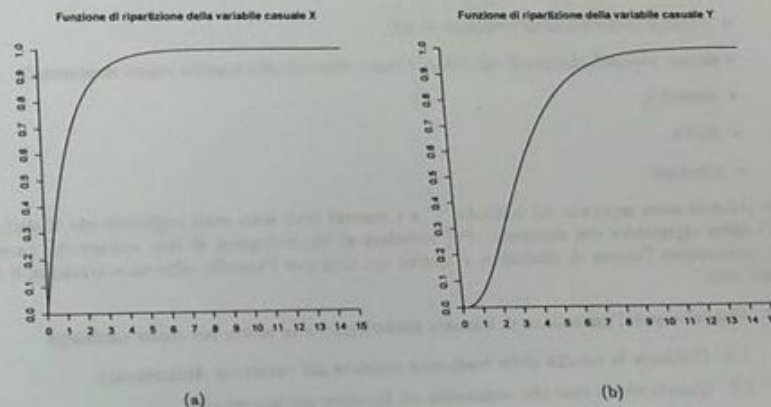


Figura 1: Funzione di ripartizione: (a) della variabile  $X$ , (b) della variabile  $Y$

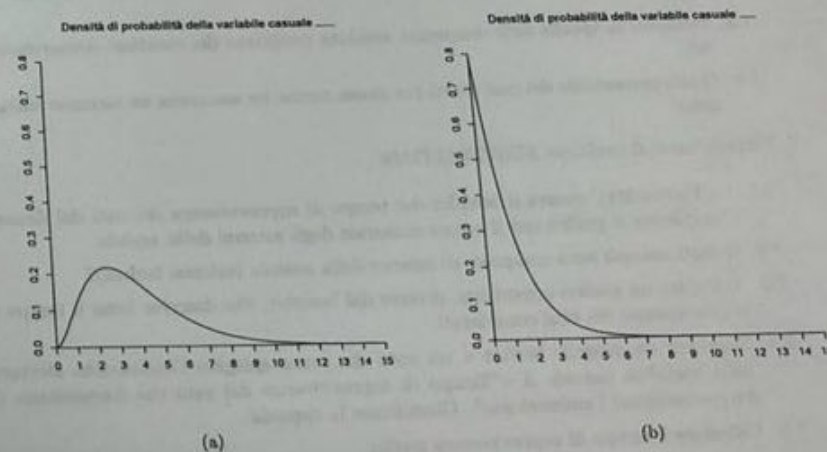


Figura 2: Funzioni densità di probabilità: which is which?

- *MORTE*: indica se il cane è ancora in vita oppure è deceduto,
- *MC*: indica se il cane è deceduto a causa di problemi cardiaci (morte cardiaca),
- *SURVIVALTIME*: tempo di sopravvivenza a partire dalla prima visita, espresso in giorni, cioè tempo intercorso tra la prima visita e il decesso oppure tempo intercorso tra la prima visita e oggi se il cane è ancora in vita,
- *Antiarritmico*: indica se il cane assume un farmaco per l'aritmia,
- *Terapia*: indica il numero di farmaci somministrati,

- *PesoKg*: peso del cane (espresso in kg),

e alcune variabili cliniche il cui valore è stato determinato tramite esami strumentali:

- *OndaEA*,
- *EDVI*,
- *Allodias*.

Le colonne sono separate dal simbolo ";" e i numeri reali sono stati registrati con il simbolo "." come separatore dei decimali. Per accedere al file collegarsi al sito [upload.di.unimi.it](http://upload.di.unimi.it), selezionare l'esame di *Statistica e analisi dei dati* per l'appello odierno e scaricare il file *cani.csv*.

1. Consideriamo l'assunzione di farmaci antiaritmici e la morte per cause cardiache.

- 1.1. Produrre la tabella delle frequenze assolute del carattere *Antiaritmico*.
- 1.2. Quanti sono i cani che assumono un farmaco antiaritmico?
- 1.3. Il carattere *Antiaritmico* è categorico. Volendolo convertire in un carattere numerico, con quale valore numerico mettereste in corrispondenza valore "SI"? Con quale il "NO"?
- 1.4. Produrre la tabella delle frequenze assolute congiunte dei caratteri *Antiaritmico* e *MC*.
- 1.5. Quale percentuale dei cani morti per cause cardiache assumeva un farmaco antiaritmico?

2. Consideriamo il carattere *SURVIVALTIME*.

- 2.1. La Figura 3(a) mostra il boxplot del tempo di sopravvivenza dei cani del dataset. Completare il grafico con il valore numerico degli estremi della scatola.
- 2.2. Quanti animali sono compresi all'interno della scatola (estremi inclusi)?
- 2.3. Tracciare un grafico opportuno, diverso dal boxplot, che descriva bene il tempo di sopravvivenza dei cani considerati.
- 2.4. Sugerite un modello teorico a voi noto che possa spiegare l'andamento aleatorio della variabile casuale  $X = \text{"Tempo di sopravvivenza dei cani che frequentano (o frequenteranno) l'ambulatorio"}$ . Giustificate la risposta.
- 2.5. Calcolate il tempo di sopravvivenza medio.
- 2.6. Calcolate la deviazione standard del tempo di sopravvivenza.
- 2.7. Il modello che avete suggerito dovrebbe presentare uno (o più) parametri. Fornitene una stima numerica.

## Esercizio 3

Create una variabile che contenga la parte di dataset relativa ai cani morti e considerando soltanto i casi in cui sia il carattere *MC*, sia il carattere *OndaEA* non siano mancanti. Nel presente esercizio le domande si riferiranno esclusivamente a questo sottoinsieme di casi.

1. L'*OndaEA* è un carattere scalare oppure ordinale?
2. Produrre il boxplot relativo al carattere *OndaEA*.

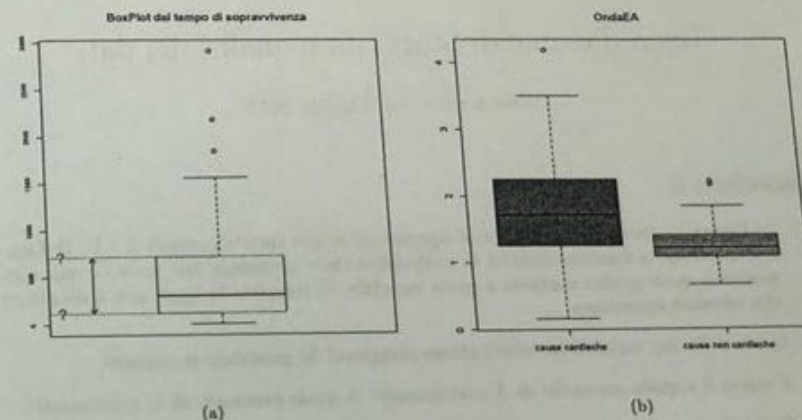


Figura 3: (a) boxplot del tempo di sopravvivenza; (b) boxplot di *OndaEA*

3. Il grafico ottenuto dovrebbe mostrare la presenza di un outlier. Determinare il valore di *OndaEA* per tale individuo.
4. L'outlier individuato è un cane morto per cause cardiache oppure no?

In Figura 3(b) sono messe a confronto le distribuzioni del carattere *OndaEA* nei due gruppi di cani deceduti per cause cardiache e per altre cause. Ci si convince facilmente del fatto che l'*OndaEA* appare molto diversa nei due gruppi, e ciò ci suggerisce che potremmo utilizzare l'*OndaEA* come criterio di discriminazione tra la morte per cause cardiache e quella per altre cause.

5. Si controlli che il terzo quartile, chiamiamolo  $s$ , dell'*OndaEA* relativamente ai cani deceduti per cause non cardiache è 1.41.
6. Quanti sono i cani deceduti per cause cardiache? Quanti per altre cause?
7. All'interno del dataset, quanti cani deceduti per cause cardiache avevano il valore di  $OndaEA \geq s$ ? E quanti cani deceduti per cause non cardiache avevano il valore di  $OndaEA < s$ ?
8. Utilizziamo il valore  $s$  trovato al punto 5. come soglia per un classificatore binario che discrimina tra morte cardiaca e morte non cardiaca: il classificatore classificherà come morte cardiaca i casi per i quali  $OndaEA \geq s$  e come morte non cardiaca i casi per i quali  $OndaEA < s$ .

Calcolare la sensibilità e la specificità di questo classificatore.