CASO DI STUDIO #3

Esercizio A

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 nel file `cani.csv` 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),
- *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,
- *Antiaritmico*: indica se il cane assume un farmaco per l'aritmia,
- *Terapia*: indica il numero di farmaci somministrati,
- *PesoKg*: peso del cane (espressa in kg),

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

- *OndaEA*,
- *EDVI*,
- *Allodiast*.

Nel file `cani.csv` le colonne sono separate dal simbolo ";" e i numeri reali sono stati registrati con il simbolo "," come separatore dei decimali.

1. Quanti sono i cani seguiti dall'ambulatorio?
2. Quanti cani soffrono di ipertensione?
3. Consideriamo ora l'età dei pazienti.
   3.1. Tracciare un istogramma dell'età dei cani con i seguenti accorgimenti:
   - fissando a un anno l'ampiezza delle classi e
   - considerando gli intervalli chiusi a sinistra e aperti a destra.

Tabella 1: Tabella sintetica che descrive l'età dei pazienti.

Proprietà	**Indice**	**Valore**
Minimo	min	
Indice di centralità		
Indice di dispersione		
Massimo	max	

3.2. Descrivere l'età dei pazienti compilando la Tabella 1, in cui scegliere un opportuno indice di centralità e un opportuno indice di dispersione.

3.3. Quanti sono i pazienti di età compresa nell'intervallo tra i 12 e i 13 anni, estremo inferiore incluso ed estremo superiore escluso ?

3.4. Quanti anni ha il cane più anziano?

3.5. Qual è la fascia di età maggiormente rappresentata? Si risponda con un intervallo chiuso a sinistra e aperto a destra.

4. Consideriamo le variabili *MORTE* e *MC*.

4.1. Quanti cani sono deceduti?

4.2. Nell'inserire le informazioni riguardo a un cane deceduto, l'operatore ha sempre specificato se la morte è avvenuta per cause cardiache o per altre cause? Se la risposta è "no", in quanti casi (sempre relativamente ai cani deceduti) l'operatore ha omesso tale informazione?

4.3. Controllare che non ci siano nei dati incongruenze riguardo alla morte, ovvero che non ci siano casi per i quali il cane risulta vivo ma morto di morte cardiaca.

4.4. Quanti cani sono deceduti per cause cardiache?

4.5. Tra le morti avvenute, quale percentuale è stata per cause cardiache?

5. La variabile *GravitaIP* è un indice di gravità dell'ipertensione.

5.1. Si tratta di un carattere scalare, ordinale oppure nominale?

5.2. Quali valori può assumere?

5.3. Produrre la tabella delle frequenze relative di *GravitaIP*.

5.4. Tracciare un grafico opportuno per descrivere la gravità dell'ipertensione.

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

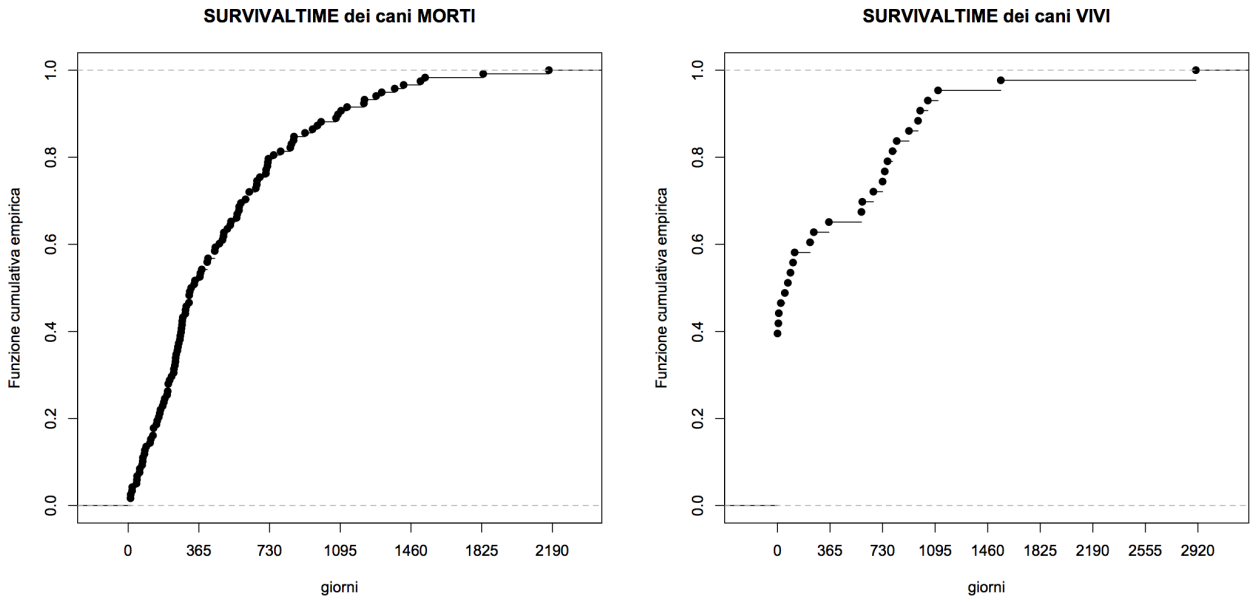
6.1. Produrre la tabella delle frequenze assolute del carattere *Antiaritmico*.

6.2. Quanti sono i cani che assumono un farmaco antiaritmico?

6.3. Il carattere *Antiaritmico* è categorico. Volendolo convertire in un carattere numerico, con quale valore numerico mettereste in corrispondenza valore *"SI"*? Con quale il *"NO"*?

6.4. Produrre la tabella delle frequenze assolute congiunte dei caratteri *Antiaritmico* e *MC*.

6.5. Quale percentuale dei cani morti per cause cardiache assumeva un farmaco antiaritmico?



(a) (b)

Figura 1: Funzione cumulativa empirica del tempo trascorso (a) dalla prima visita alla data della morte, (b) dalla prima visita a oggi.

7. Il carattere *SURVIVALTIME* (tempo di sopravvivenza) ci dice per quanti giorni il paziente è rimasto in vita a partire dalla prima visita presso l'ambulatorio. Come mostrato nei grafici di Figura 1, la distribuzione delle frequenze del tempo di sopravvivenza ha un aspetto molto diverso se si considera rispetto ai cani ancora in vita oppure a quelli morti. Potete rispondere alle seguenti due domande semplicemente ispezionando i grafici di Figura 1, considerando un anno costituito da 365 giorni.

7.1. Quale percentuale di cani tuttora vivi è in cura presso l'ambulatorio da *meno* di un anno?

7.2. Quale percentuale di cani deceduti è sopravvissuta *più* di 3 anni?

7.3. Tracciare un grafico opportuno per descrivere il tempo di sopravvivenza.

7.4. La Figura 2(a) mostra il boxplot del tempo di sopravvivenza dei cani del dataset. Completare il grafico con il valore numerico degli estremi della *scatola*.

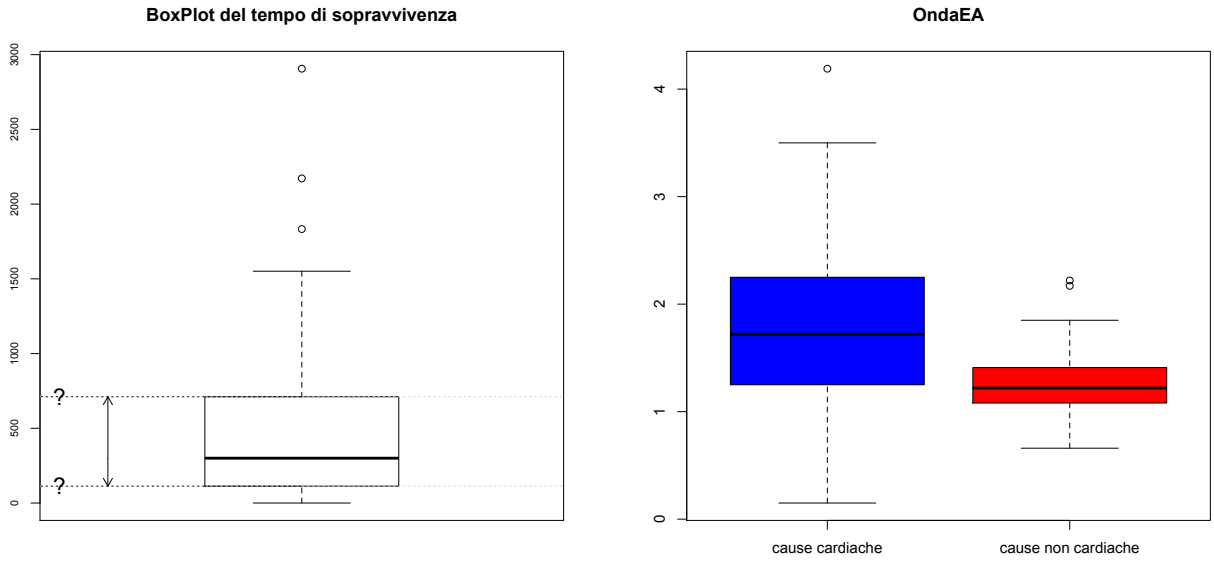
7.5. Quanti animali sono compresi all'interno della *scatola* (estremi inclusi)?

7.6. Tracciare un grafico opportuno, diverso dal boxplot, che descriva bene il tempo di sopravvivenza dei cani considerati.

7.7. Suggerite un modello teorico a voi noto che possa spiegare l'andamento aleatorio della variabile casuale X="Tempo di sopravvivenza dei cani che frequentano (o frequenteranno) l'ambulatorio". Giustificate la risposta.

7.8. Calcolate il tempo di sopravvivenza medio.

7.9. Calcolate la deviazione standard del tempo di sopravvivenza.



(a) (b)

Figura 2: (a) boxplot del tempo di sopravvivenza; (b) boxplot di *OndaEA*

8. Consideriamo il carattere *Allodiast*.

   8.1. Controllare se esso può essere considerato normale.

   8.2. Controllare che nell'intervallo di semi ampiezza 2 deviazioni standard e centrato sulla media risiede circa il 96% delle osservazioni per tale carattere.

9. I caratteri *EDVI* e *Allodiast* sono indipendenti? Motivare la risposta, anche con l'ausilio di un grafico.

Esercizio B

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*.

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 2(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 che stiamo considerando, 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.

9. Tracciare il grafico della curva ROC per il classificatore individuato nei punti precedenti, basato sul carattere *OndaEA*.