

Tema d'esame di Statistica e analisi dei dati

Prova scritta del 13 giugno 2018

Esercizio 0

Sia X una variabile casuale continua. Indichiamo con q_1 , q_2 e q_3 rispettivamente il primo, il secondo e il terzo quartile di X .

1. Quanto vale la probabilità che X assuma valori $\leq q_2$? 50%
2. Quanto vale la probabilità che X assuma valori compresi tra q_1 e q_3 ? 25% , 75% , $0,5$
3. Facciamo l'ulteriore ipotesi che X sia una variabile normale di parametri μ e σ^2 .
 - 3.1. Si determini un valore reale positivo α tale che sia uguale a 0.5 la probabilità $P(\mu - \alpha \cdot \sigma \leq X \leq \mu + \alpha \cdot \sigma)$.
 - 3.2. Esprimere q_1 e q_3 in funzione di μ e σ .
 - 3.3. Fissati, solo in questo punto, $\mu = 1$ e $\sigma = 1$, tracciare un grafico indicativo della densità di probabilità di X . Su tale grafico si evidenzino q_1 , q_3 e la probabilità $P(q_1 \leq X \leq q_3)$.
 - 3.4. Controllare che la probabilità $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$ che X assuma valori in un intervallo di semiampiezza 2σ centrato su μ è circa uguale a 0.95.
4. Dato un campione casuale X_1, \dots, X_n estratto da una popolazione X normale di valore atteso μ e deviazione standard σ , fissati due valori $0 < \delta < 1$ e $\epsilon > 0$, indicata con \bar{X} la media campionaria $\frac{1}{n} \sum_{i=1}^n X_i$, e indicata con Φ la funzione di ripartizione della normale standard, si controlli che

$$P(|\bar{X} - \mu| \leq \epsilon) \geq \delta \text{ è vera se } \Phi\left(\frac{\epsilon\sqrt{n}}{\sigma}\right) \geq \frac{1+\delta}{2}.$$

Esercizio 1

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),
- **MORTE**: indica se il cane è ancora in vita oppure è deceduto,

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

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

- **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,

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, selezionare l'esame di *Statistica e analisi dei dati* per l'appello odierno e scaricare il file cani.csv.

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.

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

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

3.4. Quanti anni ha il cane più anziano?

3.5. Qual è l'età maggiormente rappresentata?

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? NO

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. 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.

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

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

Esercizio 2

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

Le osservazioni del carattere *SURVIVALTIME* costituiscono un campione casuale X_1, \dots, X_n estratto dalla popolazione $X = \text{"Tempo di sopravvivenza, espresso in giorni, dei cani che potenzialmente afferiscono all'ambulatorio"}$.

2. Calcolare una stima del tempo di sopravvivenza atteso.

3. Esprimere, in funzione di X_1, \dots, X_n , lo stimatore T_n utilizzato per eseguire la stima del punto precedente.

4. Tale stimatore è non distorto? Si giustifichi la risposta.

5. Esprimere, in funzione di n e della deviazione standard σ della relativa popolazione, la deviazione standard di T_n .

6. Calcolare una stima della deviazione standard del tempo di sopravvivenza.

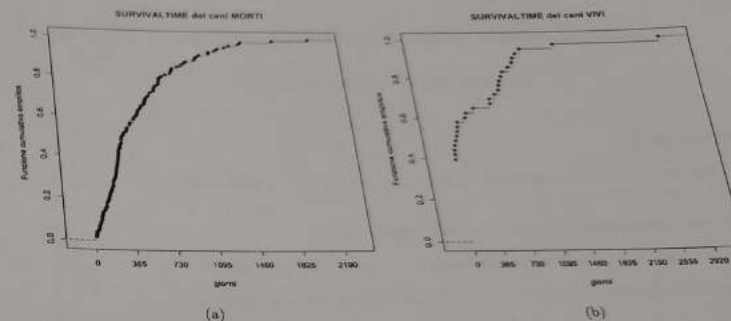


Figura 1: Funzione cumulativa empirica del tempo trascorso (a) dalla prima visita alla data della morte, (b) dalla prima visita a oggi.

- Determinare un numero di osservazioni sufficiente affinché, nella stima del tempo di sopravvivenza atteso, si compia un errore al più di due mesi (60 giorni) con probabilità almeno uguale a 0.9.
- Qual è la taglia n del nostro campione? Tale valore è sufficiente a garantire le condizioni richieste al punto precedente? Si giustifichi la risposta.
- Proporre uno stimatore del tempo di sopravvivenza atteso, espresso in anni.
- Lo stimatore proposto al punto precedente è non distorto? Si giustifichi la risposta.
- Calcolare una stima del tempo di sopravvivenza atteso, espresso in anni.

Esercizio 3

- I caratteri *EDVI* e *Allodiast* sono indipendenti? Motivare la risposta, anche con l'ausilio di un grafico.
- Con l'ausilio di uno o più grafici e del valore degli indici descrittivi che conoscete, commentate l'affermazione: "La variabile *Allodiast* segue una legge normale".
- Controllare se media e mediana della variabile *Allodiast* sono circa uguali.
- Verificare se le osservazioni di *Allodiast* sono coerenti con la proprietà al punto 3.4 dell'esercizio 0.