

Esercizio 0

1. Nella scheda riportata in fondo al testo sono mostrati in ordine sparso il boxplot, l'istogramma e la funzione cumulativa empirica di tre campioni e una tabella da compilare. Completate la tabella (da consegnare insieme allo scritto), inserendo in ciascuna riga vuota l'etichetta del grafico opportuno.
2. Dati due eventi A e B , scrivete la definizione della probabilità condizionata $P(B|A)$.

Sia X una variabile casuale normale di valore atteso μ e deviazione standard σ .

3. Dato il campione casuale X_1, \dots, X_n di taglia n estratto dalla popolazione X , proponete uno stimatore, chiamiamolo T_n , per il valore atteso μ .
4. Fissiamo, solo in questo punto, $n = 47$ e supponiamo che sia $\sigma = 2.5$. Usando lo stimatore T_n che avete proposto al punto precedente, calcolate la probabilità $P(|T_n - \mu| < 0.5)$.

Esercizio I

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

Scaricate il file `dati-ospedali.csv`, che contiene le descrizioni di alcune strutture ospedaliere italiane (Fonte: www.datiopen.it). Il file contiene una riga per ospedale e molti attributi. Gli attributi che prenderemo in considerazione sono:

- *grande.struttura*, vale 1 se l'ospedale è una grande struttura, 0 se l'ospedale è una piccola o media struttura;
- *Ingegneri.SSN*, numero di ingegneri dipendenti del Servizio Sanitario Nazionale che lavorano nell'ospedale;
- *Avvocati.SSN*, numero di avvocati dipendenti del Servizio Sanitario Nazionale che lavorano nell'ospedale;
- *Farmacisti.SSN*, numero di farmacisti dipendenti del Servizio Sanitario Nazionale che lavorano nell'ospedale;
- *Medici.SSN*, numero di medici dipendenti del Servizio Sanitario Nazionale che lavorano nell'ospedale.

1. Quanti sono gli ospedali presenti nel dataset?

2. Quanti sono in tutto gli attributi del dataset che descrivono un ospedale?

Consideriamo l'attributo *Medici.SSN*.

- 3.1. Visualizzate la funzione cumulativa empirica e un altro grafico che sia descrittivo della distribuzione di questo attributo.
- 3.2. Dai dati che abbiamo a disposizione si può osservare che, pur trattandosi di un attributo discreto, numero di medici presenti in una struttura ospedaliera può essere descritto abbastanza bene da una variabile normale. Convincetevi di ciò con un grafico e corredatelo di un commento.
- 3.3. Calcolate l'indice di posizione centrale e l'indice di variabilità più appropriati per l'attributo e giustificate la vostra scelta.

Consideriamo ora anche l'attributo *Farmacisti.SSN*.

- 1.1. Visualizzate il boxplot del numero di farmacisti rilevati negli ospedali osservati.

	Media	Dev.St.	Coeff.Variazione
MEDICI	406.6	160.7	7
FARMACISTI	6.4	2.9	7

4.3. Compilate la tabella 1 (trascrivete i valori sul foglio a quadretti).

4.3. Dalla tabella 1 si può osservare che, in media, ci sono più medici che farmacisti in un ospedale, e che la deviazione standard del numero di medici è maggiore della deviazione standard del numero di farmacisti. Calcolate e confrontate il coefficiente di variazione per i medici e per i farmacisti.

4.4. Notate una relazione tra il numero di medici e il numero di farmacisti presenti in un ospedale? In caso affermativo, di quale tipo di relazione si tratta? Supportate la vostra risposta con:

4.4.1. un grafico commentato

4.4.2. il valore di un indice numerico.

5. La tabella 2 mostra le frequenze assolute congiunte degli attributi *Avvocati.SSN* e *grande.struttura*. Utilizzate i valori della tabella per calcolare la tabella delle frequenze assolute marginali del numero di avvocati e quella della tipologia di ospedale.

Tabella 2: Frequenze congiunte di numero di avvocati per tipologia di ospedale

		Tipologia di ospedale	
		<i>grande struttura=0</i>	<i>grande struttura=1</i>
avvocati	0	2	0
	1	19	3
	2	3	1
	3	0	1

Esercizio 2

Consideriamo l'attributo bernoulliano *TantiMedici*, presente nel dataset, che abbiamo calcolato nel seguente modo: *TantiMedici* assume il valore 1 se nell'ospedale considerato lavorano più di 300 medici, assume il valore 0 se nell'ospedale considerato lavorano al più 300 medici. La tabella 3 mostra le frequenze congiunte della presenza di "*tanti medici*" (secondo la definizione appena data) e della tipologia di ospedale (*grande struttura: grande struttura=1*) o piccola/media struttura (*grande struttura=0*).

Vogliamo utilizzare un classificatore che classifica l'ospedale come grande oppure piccola/media struttura in funzione della presenza di "*tanti medici*": il classificatore classificherà come "*grande struttura*" un ospedale nel quale lavorano più di 300 medici, altrimenti lo classificherà come "*piccola o media struttura*".

1. È molto facile re-interpretare la tabella 3 per compilare la matrice di confusione riportata nella tabella 4: scrivetene i valori mancanti (sul foglio a quadretti).

Esercizio 3

Alla conferenza stampa ho rilasciato la seguente dichiarazione:

«Abbiamo a disposizione un campione di ospedali italiani sulla base del quale abbiamo stimato che il numero medio di ingegneri che lavorano in un ospedale sul territorio italiano è circa, con una deviazione standard di ingegneri.»

Tabella 3: Frequenze assolute congiunte di tipologia di ospedale e numerosità di medici

		Tipologia di ospedale	
		<i>grande struttura=0</i>	<i>grande struttura=1</i>
n. medici	<i>tanti medici = 0</i>	12	2
	<i>tanti medici = 1</i>	35	15

Tabella 4: Matrice di confusione per la classificazione della dimensione degli ospedali

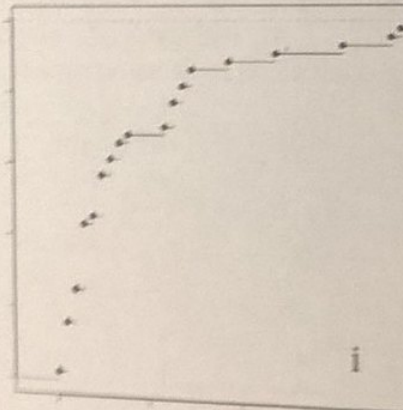
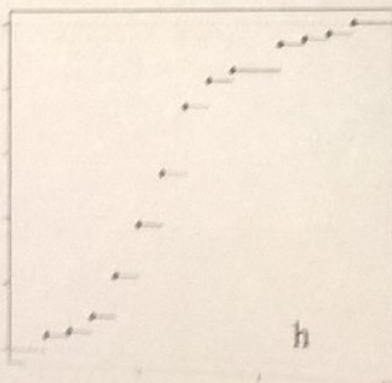
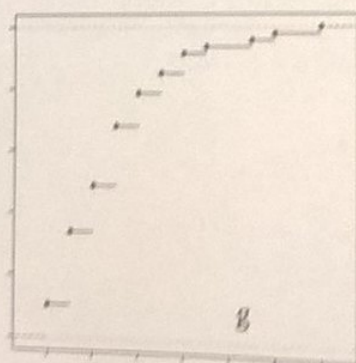
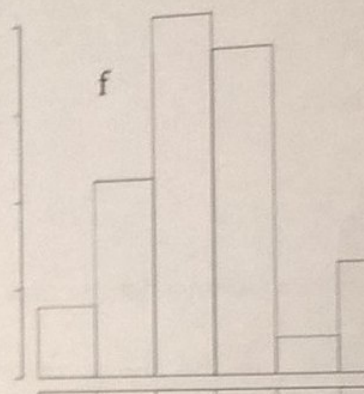
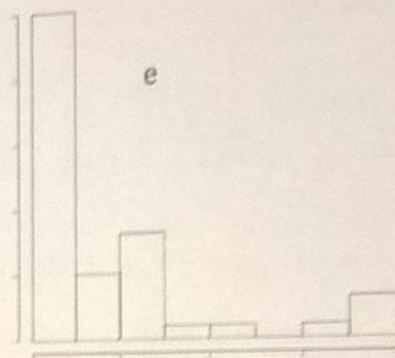
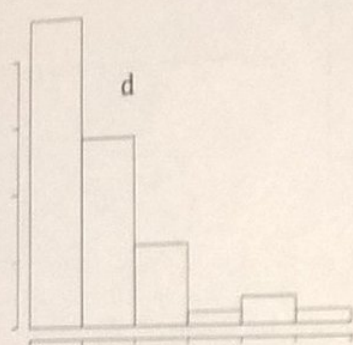
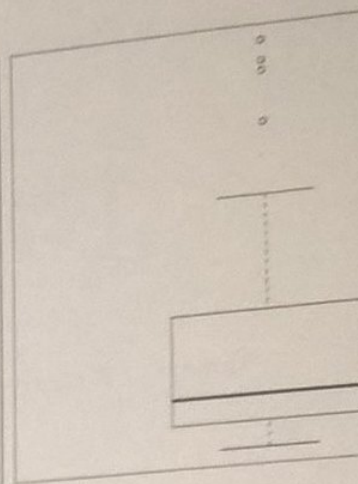
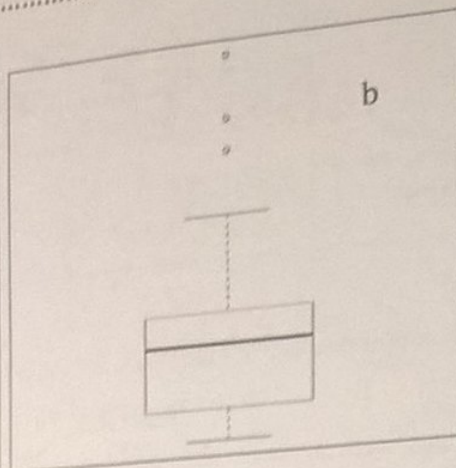
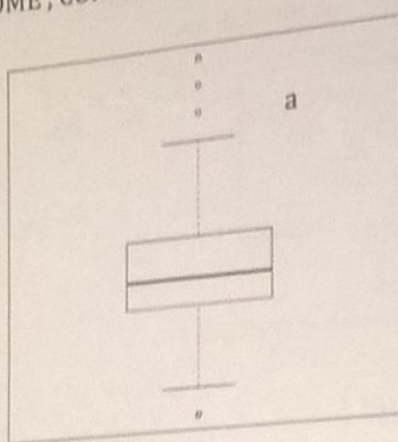
		ETICHETTA	
		NEGATIVI	POSITIVI
ESITO CLASS.	NEGATIVI	?	?
	POSITIVI	?	?

1. Si calcolino i valori da inserire nella dichiarazione sopra citata.

Al giornalista che mi ha chiesto cosa significa *circa* ho risposto che siamo sicuri almeno all'85% di aver sbagliato nella stima di tali valori medi al più di 1 unità.

2. Si controlli che, nel caso degli ingegneri, ho detto la verità. Si controlli cioè che, dato il campione casuale X_1, \dots, X_n estratto dalla popolazione $X = \text{"Numero di ingegneri che lavorano in un ospedale italiano"}$ e indicato con μ il valore atteso di X , vale la relazione $P(|\sum_{i=1}^n X_i/n - \mu| < 0.5) \geq 0.85$.

NOME, COGNOME E MATRICOLA



	BOXPLOT	FUNZIONE CUMULATIVA EMPIRICA	ISTOGRAMMA
VARIABILE 1	a		
VARIABILE 2	b		
VARIABILE 3	c		