Scrivete qui sotto i dati richiesti e riportateli anche su tutti i fogli protocollo che vi sono stati consegnati. Non potete tenere altri fogli bianchi rispetto a quelli che vi sono stati consegnati, né potete utilizzare calcolatrici portatili (avete a disposizione direttamente il computer!). Potete tenere sul banco, oltre alla penna, un libro e un quaderno di appunti per consultazione. Tutte le altre cose che avete con voi, **compresi telefoni e smartwatch**, devono essere appoggiate o appese lontano dai banchi, dove vi verrà indicato. Potete scegliere a quali domande rispondere scrivendo su carta e a quali scrivendo ed eseguendo codice, eventualmente arricchito da commenti o da celle di testo. A fine prova dovrete consegnare **tutti** i fogli che vi sono stati dati, e inviare la versione digitale di questo notebook, contenente il vostro svolgimento, utilizzando il sito https://upload.di.unimi.it.

Cognome e nome

Matricola

Esercizio 1

Sia X una variabile aleatoria distribuita secondo un modello di Poisson di parametro λ.

1. λ può assumere un valore maggiore di 1? Perché? E può assumere un valore negativo? Perché?

2. Esprimete il valore atteso e la deviazione standard di X in funzione di λ. Esprimete successivamente la varianza di X in funzione del valore atteso di quest'ultima.

3. Sia Y un'altra variabile aleatoria distribuita come X e da essa indipendente. Dato un generico **numero naturale** x, indichiamo con $p(x)$ la probabilità dell'evento $X + Y = x$. Esprimente $p(x)$ in funzione di λ e x.

4. Fissiamo, **solo in questo punto**, $\lambda = 0.9$. Utilizzando il computer visualizzate graficamente l'andamento di $p(x)$ al variare di x, scegliendo **opportunamente** il tipo di grafico da generare e motivando la scelta fatta. Come avete scelto il valore massimo per x?

5. Date n variabili aleatorie $X_1, \ldots X_n$ indipendenti e identicamente distribuite come X, indichiamo con S la somma $\sum_{i=1}^{n} X_i$. Che distribuzione segue S?

Esercizio 2

... descritta nell'Esercizio 1.

4. Fissiamo, **solo in questo punto**, $\lambda = 0.9$. Utilizzando il computer visualizzate graficamente l'andamento di $p(x)$ al variare di x, scegliendo **opportunamente** il tipo di grafico da generare e motivando la scelta fatta. Come avete scelto il valore massimo per x?

5. Date n variabili aleatorie $X_1, \dots X_n$ indipendenti e identicamente distribuite come X, indichiamo con S la somma $\sum_{i=1}^{n} X_i$. Che distribuzione segue S?

Esercizio 2

Per $n \in \mathbb{N}$ fissato, sia $X_1, \dots, X_n$ un campione aleatorio estratto da una popolazione modellata tramite la variabile aleatoria X descritta nell'Esercizio 1.

1. Fissiamo, **solo in questo punto**, $n = 3$. La variabile aleatoria $T = (X_1 + X_2 - X_3)$ è uno stimatore deviato per λ? Perché? Come cambiano (se cambiano) le risposte a queste due domande considerando $U = \frac{1}{5}(2X_1 + 2X_2 + X_3)$ al posto di T? Possiamo valutare se questi stimatori godono della proprietà di consistenza in media quadratica?

Concentriamoci, nel resto di questo esercizio, sulla variabile aleatoria $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

2. $\overline{X}$ è distribuita secondo uno dei modelli che avete studiato? Indipendentemente dalla risposta data alla domanda precedente, siete in grado di esprimere la probabilità $\mathrm{P}(\overline{X} = x)$ in funzione di λ?

3. Esiste un modello che approssima in modo ragionevole la distribuzione di $\overline{X}$? Motivate la vostra risposta, specificando anche quali valori assumono eventuali parametri del modello considerato.

4. $\overline{X}$ è uno stimatore non deviato per $\mu = \mathbb{E}(X)$? Calcolate il suo bias e il suo MSE e dite se esso gode anche della proprietà di consistenza in media quadratica.

5. $\overline{X}$ è uno stimatore non deviato per $\sigma^2 = \mathrm{Var}(X)$? Calcolate il suo bias e il suo MSE e dite se esso gode anche della proprietà di consistenza in media quadratica.

cambiano) le risposte a queste due domande considerando $U = \frac{1}{5}(2X_1 + 2X_2 + X_3)$ al posto di T? Possiamo valutare se questi stimatori godono della proprietà di consistenza in media quadratica?

Concentriamoci, nel resto di questo esercizio, sulla variabile aleatoria $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

2. $\overline{X}$ è distribuita secondo uno dei modelli che avete studiato? Indipendentemente dalla risposta data alla domanda precedente, siete in grado di esprimere la probabilità $\mathrm{P}(\overline{X} = x)$ in funzione di λ?

3. Esiste un modello che approssima in modo ragionevole la distribuzione di $\overline{X}$? Motivate la vostra risposta, specificando anche quali valori assumono eventuali parametri del modello considerato.

4. $\overline{X}$ è uno stimatore non deviato per $\mu = \mathbb{E}(X)$? Calcolate il suo bias e il suo MSE e dite se esso gode anche della proprietà di consistenza in media quadratica.

5. $\overline{X}$ è uno stimatore non deviato per $\sigma^2 = \mathrm{Var}(X)$? Calcolate il suo bias e il suo MSE e dite se esso gode anche della proprietà di consistenza in media quadratica.

6. $\overline{X}$ è uno stimatore non deviato per $\sigma = \sqrt{\mathrm{Var}(X)}$? Calcolate il suo bias e il suo MSE e dite se esso gode anche della proprietà di consistenza in media quadratica.

7. Supponiamo, **solo in questo punto** che $n = 1125$. Indicare, in funzione di ϵ e λ, la probabilità che l'errore (in valore assoluto) che si compie usando $\overline{X}$ per stimare λ sia minore o uguale a ϵ.

Esercizio 3

Collegatevi al sito upload.di.unimi.it, selezionate l'esame di *Statistica e analisi dei dati* per l'appello odierno e scaricate il file `rilevazioni.csv`. Questo file contiene le seguenti informazioni raccolte da un ipotetico fornitore di computer i cui modelli sono dotati di sensori interni che eseguono delle rilevazioni ogni

Esercizio 3

Collegatevi al sito upload.di.unimi.it, selezionate l'esame di *Statistica e analisi dei dati* per l'appello odierno e scaricate il file `rilevazioni.csv`. Questo file contiene le seguenti informazioni raccolte da un ipotetico fornitore di computer i cui modelli sono dotati di sensori interni che eseguono delle rilevazioni ogni ora:

- *blocchidanneggiati*: numero di blocchi dell'hard disk che si sono danneggiati nell'ora considerata;
- *temperatura*: media della temperatura interna del computer, misurata in gradi centigradi e riferita all'ora considerata;
- *raffreddamento*: valore binario che indica se il sistema di raffreddamento automatico addizionale è stato attivato almeno una volta nell'ora considerata (0: sistema non attivato, 1: sistema attivato).

In questo file il carattere `;` separa le colonne.

1. Scrivete ed eseguite del codice il cui output indichi quanti casi sono contenuti nel *dataset*, quanti casi contengono almeno un valore mancante e quali sono gli attributi per i quali è presente almeno un valore mancante.

2. Descrivete l'attributo *raffreddamento* utilizzando la rappresentazione grafica che ritenete più adeguata, motivando la scelta fatta e commentando i risultati ottenuti.

3. Valutate l'ipotesi che la temperatura influisca sull'accensione del sistema automatico di raffreddamento, commentando i risultati ottenuti. Quali strumenti avete utilizzato per valutare questa ipotesi? Perché?

4. Visualizzate graficamente l'attributo *blocchidanneggiati* (tenendo presente il tipo di questo attributo), commentando i risultati ottenuti e giustificando la scelta del tipo di grafico realizzato.

5. Generate una tabella che mostri le frequenze congiunte degli attributi *blocchidanneggiati* e *raffreddamento*. Indicando con R l'evento che si verifica quando il sistema di raffreddamento automatico è stato attivato durante un'ora e con D l'evento che si verifica quando in un'ora risultano danneggiati dei blocchi del disco fisso, usate la tabella generata per calcolare le seguenti probabilità: i. $\mathrm{P}(D)$, ii. $\mathrm{P}(R \cap D)$, iii. $\mathrm{P}(D|R)$, iv. $\mathrm{P}(R|\overline{D})$. Le probabilità che avete calcolato

4. Visualizzate graficamente l'attributo *blocchidanneggiati* (tenendo presente il tipo di questo attributo), commentando i risultati ottenuti e giustificando la scelta del tipo di grafico realizzato.

5. Generate una tabella che mostri le frequenze congiunte degli attributi *blocchidanneggiati* e *raffreddamento*. Indicando con R l'evento che si verifica quando il sistema di raffreddamento automatico è stato attivato durante un'ora e con D l'evento che si verifica quando in un'ora risultano danneggiati dei blocchi del disco fisso, usate la tabella generata per calcolare le seguenti probabilità: i. $\mathrm{P}(D)$, ii. $\mathrm{P}(R \cap D)$, iii. $\mathrm{P}(D|R)$, iv. $\mathrm{P}(R|\overline{D})$. Le probabilità che avete calcolato devono sommare a uno? Perché?

Esercizio 4

In questo esercizio ci concentreremo sui valori dell'attributo *blocchidanneggiati*.

1. Memorizzate in due variabili `raffreddamento_si` e `raffreddamento_no` i valori dell'attributo *blocchidanneggiati* che si riferiscono, rispettivamente, ai casi in cui il sistema di raffreddamento automatico è attivato oppure no.

2. Calcolate i decili empirici (i quantili di livello uguale a $i/10$ per $i = 0, \ldots, 10$) separatamente per i dati contenuti in `raffreddamento_si` e `raffreddamento_no` e disegnate i punti le cui ascisse e ordinate corrispondono, rispettivamente, ai valori ottenuti, raggruppati per livello. Usate il grafico ottenuto per confermare o confutare l'ipotesi che la distribuzione del numero di blocchi danneggiati sia dipendente dal funzionamento del sistema di raffreddamento automatico, motivando il vostro ragionamento. Integrate la vostra analisi con altri metodi che ritenete opportuni.

3. Supponendo che i valori contenuti in `raffreddamento_si` siano assimilabili a un campione estratto da una popolazione descritta da una variabile aleatoria X_{SI}, stimate il valore atteso di quest'ultima.

4. Confermate l'ipotesi che i valori contenuti in `raffreddamento_si` siano descritti da una distribuzione appartenente al modello di Poisson (suggerimento: potete utilizzare una tecnica analoga a quella descritta nel punto 2 di questo esercizio, sostituendo i quantili empirici di `raffreddamento_no` con dei quantili teorici opportunamente calcolati).

5. Ripetete le analisi fatte al punto precedente, considerando i dati contenuti in `raffreddamento_no`

3. Supponendo che i valori contenuti in `raffreddamento_si` siano assimilabili a un campione estratto da una popolazione descritta da una variabile aleatoria X_{SI}, stimate il valore atteso di quest'ultima.

4. Confermate l'ipotesi che i valori contenuti in `raffreddamento_si` siano descritti da una distribuzione appartenente al modello di Poisson (suggerimento: potete utilizzare una tecnica analoga a quella descritta nel punto 2 di questo esercizio, sostituendo i quantili empirici di `raffreddamento_no` con dei quantili teorici opportunamente calcolati).

5. Ripetete le analisi fatte al punto precedente, considerando i dati contenuti in `raffreddamento_no`

6. Indicate, se esiste, uno stimatore non deviato per: i. la varianza di X_{SI}, ii. la deviazione standard di X_{SI}, iii. il parametro λ_{SI}. Di quali proprietà godono gli stimatori che avete indicato?

7. Con quale probabilità potete garantire che la stima fatta al punto 3 disti in valore assoluto dal valore sconosciuto più di 0.1?

8. Stimate la probabilità che il sistema di raffreddamento automatico sia in funzione.

9. Sostituendo a λ_{SI} e λ_{NO} delle opportune stime che potete ricavare dai dati a disposizione, calcolate la probabilità che in un'ora si danneggino più di quattro blocchi, considerando separatamente i due casi nei quali rispettivamente il sistema di raffreddamento è in funzione oppure no.

10. Se in un'ora di funzionamento si danneggiano più di quattro blocchi, il sistema di rilevazione genera un allarme. Qual è la probabilità che questo allarme venga generato?

11. Se l'allarme descritto al punto precedente viene attivato più di sette volte in un anno, il produttore sostituisce gratuitamente l'intero computer. All'inizio dell'anno sono stati venduti 7500 computer: supponendo che l'emissione dell'allarme in un giorno sia indipendente da ciò che succede negli altri giorni, qual è il numero atteso di computer che il produttore dovrà sostituire a fine anno?

NON SCRIVETE NULLA QUI SOTTO