

Tema d'esame di Statistica e analisi dei dati

Prova scritta del 20 settembre 2018

Esercizio 0

Considerate un esperimento che può avere due esiti possibili, che chiameremo *successo* e *insuccesso*. Sia X la variabile casuale che assume il valore 1 se eseguendo l'esperimento si ottiene un successo, e assume il valore 0 altrimenti. Indichiamo con $p \in (0, 1)$ la probabilità di ottenere un successo.

1. Che legge segue la variabile X ?
2. Esprimete, in funzione di p , la deviazione standard di X .
3. Eseguiamo $k \in \mathbb{N}$ prove dell'esperimento considerato in condizioni di indipendenza. Esprimete, in funzione di k e p , la probabilità di ottenere esattamente $x \in \mathbb{N}$ successi.
4. Consideriamo, solo in questo punto, il caso particolare $k = 50$ e $p = 0.3$, e la variabile casuale $Y = \text{"numero di successi ottenuti nelle } k \text{ prove sopra descritte"}$. Tracciate il grafico della funzione massa di probabilità di Y .
5. Fornite la definizione di *stimatore* per un parametro θ di una popolazione Z .

Sia X_1, \dots, X_n un campione casuale estratto dalla popolazione X descritta nei punti precedenti.

6. Proponete uno stimatore T_n non distorto per il parametro p .
7. Per $n \gg 1$ si controlli che:
$$P(-\epsilon < T_n - p < \epsilon) \approx 2\Phi(\epsilon \cdot \sqrt{n}/\sigma) - 1$$
dove Φ è la funzione di ripartizione di una normale standard e σ è la deviazione standard della popolazione X .

Esercizio 1

Il file `finanziamenti.csv` contiene alcune informazioni riguardo a progetti finanziati nell'anno corrente:

- *id*: identificatore del progetto,
- *TemaPrioritario*: codice che indica il tema prioritario del progetto,
- *FONTE*: area tematica del progetto,
- *CATEGORIA*: categoria del progetto,
- *CodiceCategoria*: codice numerico della *CATEGORIA*,

- 6.1. Selezionate i progetti di tipo A che hanno ricevuto un finanziamento provinciale compreso tra i 200 e i 1000 euro, estremo sinistro incluso, e salvate questa parte di dataset in una variabile chiamata `selezione_progetti_a`.
- 6.2. Tracciate un istogramma del finanziamento provinciale di tali progetti, imponendo che le classi abbiano ampiezza 100 euro.
- 6.3. Tracciate anche il boxplot per la medesima quantità.
- 6.4. Tra i due grafici appena prodotti, quale ritenete maggiormente informativo? Giustificate la risposta.
- 6.5. Relativamente a tali progetti, qual è stato l'importo medio finanziato dalla provincia? Quale la deviazione standard?
- 6.6. Quanti hanno ricevuto un finanziamento provinciale compreso tra i 500 e i 700 euro?
- 6.7. Esiste una evidente relazione tra finanziamento provinciale e spese sostenute. Descrivete tale relazione nel modo più dettagliato possibile, utilizzando un indice numerico e un metodo grafico.
- 6.8. Nella relazione avrete notato la presenza di almeno un progetto che si discosta notevolmente dall'andamento più generale. Eliminate tali progetti dall'insieme dei dati e rispondete nuovamente alle domande del punto precedente.

Esercizio 2

Ritorniamo al dataset completo. I dati ivi presenti costituiscono una fotografia della situazione contabile dopo un anno dall'assegnazione del finanziamento, in cui il valore dell'attributo *TotSpese* è mancante nel caso in cui non siano ancora state sostenute spese per il relativo progetto.

1. Quanti sono i progetti che non hanno ancora sostenuto spese?
2. Siamo interessati a stimare la probabilità che, a distanza di un anno dall'assegnazione del finanziamento, non si siano ancora sostenute spese. Fornite una stima di tale probabilità.
3. Consideriamo lo spazio campionario $\Omega = \mathbb{R}^+ \cup \{\text{NA}\}$ che codifica i valori possibili per l'attributo *TotSpese*. Tenuto conto del fatto che una variabile aleatoria è una funzione $X : \Omega \mapsto \mathbb{R}$, definite opportunamente X in modo da legarla all'evento "a un anno dall'assegnazione del finanziamento il progetto non ha sostenuto spese".
4. Che legge segue X ?
5. Fornite una stima del parametro di tale legge, precisando una sola cifra decimale.
6. Siamo disposti ad accettare di "sbagliare" nella stima con una probabilità del 95%. Qual è il margine di errore che dobbiamo tollerare?
7. Tale errore riguarda la prima cifra decimale oppure la seconda?
8. Pensando di finanziare nel prossimo futuro altri 50 progetti, date una stima della probabilità che, trascorso un anno dall'assegnazione del finanziamento, esattamente 10 di essi non abbiano ancora sostenuto spese.

- *FinProvincia*: entità del finanziamento da parte della provincia,
- *FinRegione*: entità del finanziamento da parte della regione,
- *TotSpese*: spese sostenute per il progetto.

Le colonne sono separate dal simbolo ";" e i numeri reali sono stati registrati con il simbolo "." come separatore dei decimali. Per accedere al file, collegatevi al sito upload.di.unimi.it, selezionate l'esame di *Statistica e analisi dei dati* per l'appello odierno e scaricate il file *finanziamenti.csv*.

1. Il carattere *CodiceCategoria* è nominale, ordinale o scalare? Giustificate la risposta.

2. Calcolate la tabella delle frequenze assolute del carattere *UNITA*.

3. Tracciate un grafico opportuno per descrivere il carattere *UNITA*.

4. La Figura 1 mostra la funzione di ripartizione empirica per un sottoinsieme delle osservazioni relativi al carattere *TotSpese*, in cui gli importi sono indicati in centinaia di migliaia di Euro. Leggendo esclusivamente il grafico:

4.1. indicate quale sottoinsieme di osservazioni è stato utilizzato;

4.2. specificate quale percentuale delle osservazioni visualizzate assume un valore compreso tra uno e due milioni di Euro.

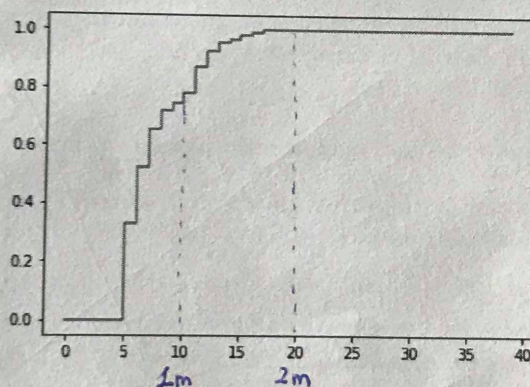


Figura 1: Il grafico della funzione di ripartizione empirica per un sottoinsieme delle osservazioni di *TotSpese*.

5. Prendiamo in considerazione la quota di finanziamento erogata dalla Provincia.

5.1. Create una variabile (chiamatela *progetti_a*, per indicare i progetti di tipo A) che contenga la parte di dataset relativa ai progetti per i quali la quota provinciale di finanziamento è minore di quella regionale, e un'altra (chiamata *progetti_b*, per indicare i progetti di tipo B) che contenga la parte di dataset relativa ai progetti restanti.

5.2. Quanti sono progetti di tipo A? Quanti sono progetti di tipo B?