Esame di Probabilità e Statistica [3231]

Esame di Calcolo delle Probabilità e Statistica [2959]

Corso di Studi di Ingegneria Gestionale (D.M.270/04) (L)

Dipartimento di Meccanica, Matematica e Management
Politecnico di Bari

Cognome: ______________________ Docente: Gianluca Orlando
Nome: ______________________ Appello: giugno 2023 - turno 2
Matricola: ______________________ Data: 26/06/2023

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Si studia il tempo di vita di una batteria per smartphone. I dati misurati (in anni) vengono raccolti in intervalli e si contano le osservazioni negli intervalli:

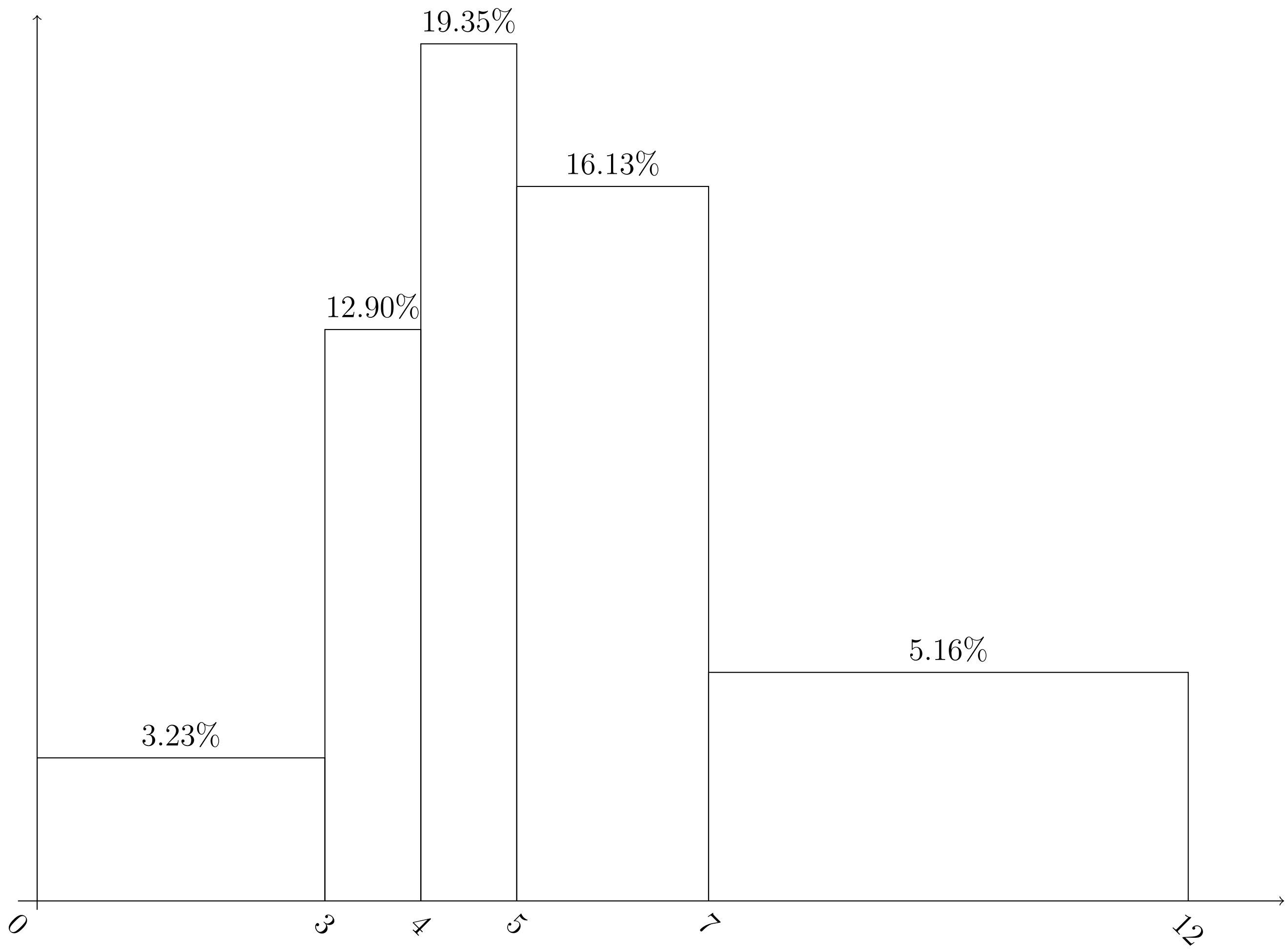
intervalli (anni)	frequenza assoluta
$[0, 3)$	3
$[3, 4)$	4
$[4, 5)$	6
$[5, 7)$	10
$[7, 12)$	8

1. Rappresentare un istogramma delle densità di frequenze relative.
2. Determinare la classe modale.
3. Calcolare un'approssimazione della media e della varianza dei dati.
4. Calcolare un'approssimazione della mediana dei dati.

Soluzione. 1. Calcoliamo le densità di frequenze relative dividendo le frequenze relative per l'ampiezza degli intervalli.

intervallo	freq. assolute	freq. relative	densità di freq. rel.	freq. cumulate
$[0, 3)$	3	9.68%	3.23%	3
$[3, 4)$	4	12.90%	12.90%	7
$[4, 5)$	6	19.35%	19.35%	13
$[5, 7)$	10	32.26%	16.13%	23
$[7, 12)$	8	25.81%	5.16 %	31

Rappresentiamo le densità di frequenze relative in un istogramma.



2. La classe modale è quella con maggiore densità di frequenza relativa, quindi è l'intervallo $[4, 5)$.

3. Per calcolare un'approssimazione della media utilizziamo le frequenze relative ottenute da $p_j = f_j/n$ dove $n = 31$ e i valori centrali $\tilde{v}_j$ degli intervalli

$$\begin{aligned}\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \simeq \frac{1}{n}\sum_{j=1}^{k} f_j\tilde{v}_j = \sum_{j=1}^{k} p_j\tilde{v}_j \\ &= 9.68\% \cdot 1.5 + 12.90\% \cdot 3.5 + 19.35\% \cdot 4.5 + 32.26\% \cdot 6 + 25.81\% \cdot 9.5 = 5.855\,.\end{aligned}$$

Calcoliamo un'approssimazione della varianza

$$\begin{aligned}s^2 &= \frac{1}{n-1}\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\Big) \simeq \frac{1}{n-1}\Big(\sum_{j=1}^{n} f_j\tilde{v}_j^2 - n\overline{x}^2\Big) = \frac{n}{n-1}\Big(\sum_{j=1}^{n} p_j\tilde{v}_j^2 - \overline{x}^2\Big) \\ &= \frac{31}{30}\Big(9.68\% \cdot 1.5^2 + 12.90\% \cdot 3.5^2 + 19.35\% \cdot 4.5^2 + 32.26\% \cdot 6^2 + 25.81\% \cdot 9.5^2 - 5.855^2\Big) \\ &\simeq 6.55\,.\end{aligned}$$

4. Per calcolare un'approssimazione della mediana dei dati, usiamo le frequenze cumulate. Troviamo l'intervallo I_j tale che $F_j \leq \frac{n}{2} = 15.5 < F_{j+1}$. Si tratta dell'intervallo $[5, 7)$. Approssimiamo la mediana con

$$Q_2 \simeq a_j + \lambda_j(b_j - a_j)$$

dove

$$\lambda_j = \frac{n/2 - F_j}{F_{j+1} - F_j} = \frac{15.5 - 13}{23 - 13} = 0.25\,.$$

Quindi

$$Q_2 \simeq 5 + 0.25(7-5) = 5.5\,.$$

Esercizio 2. (7 punti) Sia (X_1, X_2) un vettore aleatorio con probabilità congiunta descritta dalla seguente tabella:

$$\begin{array}{l|ccc} \quad\; X_1 & -1 & 0 & 1 \\ X_2 & & & \\ \hline 0 & a & \frac{1}{4} & \frac{1}{4} \\ 1 & \frac{1}{4} & b & c \end{array}$$

dove a, b, c sono parametri da determinare. Si assuma che

- $\mathbb{E}(X_2) = \frac{1}{2}$
- $\mathrm{Var}(X_1 + X_2) = \frac{1}{4}$ (Suggerimento: conviene ricordare la formula per la varianza della somma)

1. Determinare il valore di a utilizzando la condizione $\mathbb{E}(X_2) = \frac{1}{2}$.
2. Determinare il valore di b, c.
3. Stabilire se le variabili aleatorie X_1 e X_2 sono indipendenti per i valori di a, b, c trovati.

Soluzione. 1. Imponiamo che la somma delle probabilità nella tabella sia 1:

$$a + \frac{1}{4} + \frac{1}{4} + \frac{1}{4} + b + c = 1 \implies a + b + c = \frac{1}{4}\,.$$

Imponiamo che $\mathbb{E}(X_2) = \frac{1}{2}$:

$$\frac{1}{2} = \mathbb{E}(X_2) = 1 \cdot \left(\frac{1}{4} + b + c\right) \implies b + c = \frac{1}{4}\,.$$

Da queste due condizioni segue che

$$a + b + c = \frac{1}{4} \implies a + \frac{1}{4} = \frac{1}{4} \implies a = 0\,.$$

2. Imponiamo che $\mathrm{Var}(X_1 + X_2) = \frac{1}{2}$. Ricordiamo la formula per la varianza della somma di due variabili aleatorie:

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X_1, X_2)$$

Calcoliamo questi termini. Conviene calcolare prima

$$\mathbb{E}(X_1) = -1 \cdot \left(a + \frac{1}{4}\right) + 1 \cdot \left(\frac{1}{4} + c\right) = c\,.$$

Quindi

$$\begin{aligned} \mathrm{Var}(X_1) &= \mathbb{E}(X_1^2) - \mathbb{E}(X_1)^2 = (-1)^2 \cdot \left(a + \frac{1}{4}\right) + 1^2 \cdot \left(\frac{1}{4} + c\right) - c^2 \\ &= \frac{1}{2} + c - c^2 \end{aligned}$$

$$\mathrm{Var}(X_2) = \mathbb{E}(X_2^2) - \mathbb{E}(X_2)^2 = 1 \cdot \Big(\frac{1}{4} + b + c\Big) - \Big(\frac{1}{2}\Big)^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}\,.$$

$$\begin{aligned}\mathrm{Cov}(X_1, X_2) &= \mathbb{E}(X_1 \cdot X_2) - \mathbb{E}(X_1) \cdot \mathbb{E}(X_2) \\ &= -1 \cdot \frac{1}{4} + 1 \cdot c - \mathbb{E}(X_1) \cdot \mathbb{E}(X_2) = -\frac{1}{4} + c - \frac{1}{2}c = \frac{c}{2} - \frac{1}{4}\,.\end{aligned}$$

Segue che

$$\frac{1}{4} = \mathrm{Var}(X_1 + X_2) = \frac{1}{2} + c - c^2 + \frac{1}{4} + c - \frac{1}{2} = \frac{c}{2} - \frac{1}{4} = 2c - c^2 + \frac{1}{4}\,,$$

da cui $2c - c^2 = 0$, cioè $c(2-c) = 0$ e quindi (scartando $c = 2$ che non è ammissibile) $c = 0$, da cui

$$a = 0\,, \quad b = \frac{1}{4}\,, \quad c = 0\,.$$

La tabella completa è

X_2 \ X_1	-1	0	1
0	0	$\frac{1}{4}$	$\frac{1}{4}$
1	$\frac{1}{4}$	$\frac{1}{4}$	0

2. Dai conti del punto precedente abbiamo che $\mathrm{Cov}(X_1, X_2) = -\frac{1}{4}$. Le due variabili aleatorie non possono essere indipendenti perché la condizione di covarianza nulla è una condizione necessaria per l'indipendenza.

Esercizio 3. (8 punti) Il tempo necessario per un certo impiegato allo sportello delle poste per servire un cliente è distribuito con legge esponenziale. In media, il tempo in cui termina di servire il cliente è 7 minuti.

1. Qual è la probabilità che siano necessari più di 10 minuti per servire un cliente?
2. L'impiegato allo sportello inizia a servire un cliente alle 10:00. Passa del tempo, guardiamo l'orologio, sono le 10:10 e l'impiegato non ha ancora terminato di servire il cliente. Qual è la probabilità che il cliente finisca dopo le 10:20?
3. Ci sono in tutto 8 clienti da servire. Si assuma che i tempi necessari per servire ciascun cliente siano indipendenti. Qual è la probabilità che l'impiegato termini di servire almeno 3 clienti in più di 10 minuti?

Soluzione. Consideriamo $X \sim \mathrm{Exp}(\lambda)$ la variabile aleatoria che descrive il tempo allo sportello. Ricordiamo che $\mathbb{E}(X) = \frac{1}{\lambda}$. Quindi $\frac{1}{\lambda} = 7 \implies \lambda = \frac{1}{7}$.

1. Viene chiesto di calcolare

$$\mathbb{P}(\{X > 10\}) = e^{-\lambda 10} = e^{-\frac{10}{7}} = 23.97\%\,.$$

2. Viene chiesto di calcolare

$$\mathbb{P}(\{X > 20\} | \{X > 10\})\,.$$

Utilizzando l'assenza di memoria della legge esponenziale, abbiamo che

$$\mathbb{P}(\{X > 20\}|\{X > 10\}) = \mathbb{P}(\{X > 10\}) = 23.97\% \,.$$

3. Consideriamo l'evento "il cliente viene servito in più di 10 minuti". Questo evento ha probabilità

$$p = \mathbb{P}(\{X > 10\}) = 23.97\% \,.$$

Identifichiamo questo evento come un "successo". Consideriamo una variabile aleatoria con legge binomiale $Y \sim \mathrm{B}(n, p)$ con $n = 8$ e $p = 23.97\%$. Ci viene chiesto di calcolare

$$\mathbb{P}(\{Y \geq 3\}) = 1 - \mathbb{P}(\{Y \leq 2\}) = 1 - \mathbb{P}(\{Y = 0\}) + \mathbb{P}(\{Y = 1\}) + \mathbb{P}(\{Y = 2\})$$
$$= 1 - \binom{8}{0}(23.97\%)^0(1 - 23.97\%)^8 - \binom{8}{1}(23.97\%)^1(1 - 23.97\%)^7 - \binom{8}{2}(23.97\%)^2(1 - 23.97\%)^6$$
$$= 29.60\% \,.$$

Esercizio 4. (7 punti) Un produttore di acciaio INOX sostiene che il carico di rottura medio del materiale da lui prodotto è 730 $MPa = N/mm^2$. Si vuole stabilire se la media è in realtà più bassa. Si misura il carico di rottura su un campione casuale e si osservano i seguenti risultati:

732	723	729	715	712	721	730	724
749	710	708	691	723	693	734	749
733	715	708	742	728	720	734	708
725	762	729	722	705	730	712	698

La media calcolata sui dati di questo campione risulta essere 722.31. Si supponga che la distribuzione del carico di rottura abbia deviazione standard $\sigma = 20$.

1. I dati sono significativi al 5% per stabilire che la media è effettivamente più bassa di 730?
2. Qual è il più piccolo livello di significatività per cui i dati permettono di affermare che la media è più bassa di 730?

Soluzione. Si tratta di un test di ipotesi. Abbiamo un campione casuale $X_1, \ldots, X_n$ con $n = 32$, $\mathbb{E}(X_i) = \mu$ e $\mathrm{Var}(X_i) = \sigma^2$. La distribuzione della popolazione non è nota, la media μ non è nota, la deviazione standard è nota $\sigma = 20$. Il campione è numeroso.

Poniamo $\mu_0 = 730$. Fino a prova contraria, è vera l'ipotesi nulla

$$H_0 : \mu = \mu_0 \,,$$

e ci stiamo chiedendo se i dati sono abbastanza significativi da rifiutare l'ipotesi nulla a favore dell'ipotesi alternativa

$$H_1 : \mu < \mu_0 \,.$$

con livello di significatività $\alpha = 5\%$.

Poiché l'ipotesi alternativa è $H_1 : \mu < \mu_0$, i dati saranno significativi se la media è sufficiente più piccola di μ_0. La regione critica è allora della forma

$$R_c = \{(x_1, \ldots, x_n) \in R(X_1, \ldots, X_n) \ : \ \overline{x}_n < \mu_0 - \delta\} \,,$$

dove $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ è la media calcolata sulla realizzazione $x_1, \ldots, x_n$ del campione casuale.

Assumiamo l'ipotesi nulla H_0 vera, cioè $\mu = \mu_0$, ovvero $\mathbb{E}(X_i) = \mu_0$. Consideriamo la media campionaria $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ e utilizziamo la definizione di significatività per ottenere

$$\alpha = \mathbb{P}(\{(X_1,\dots,X_n) \in R_c\}) = \mathbb{P}(\{\overline{X}_n < \mu_0 - \delta\}) = \mathbb{P}\Big(\Big\{\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} < -\frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big)\,.$$

Non conosciamo la distribuzione della popolazione, ma il campione è numeroso ($n \geq 30$). Per il Teorema del Limite Centrale, si ha che $\frac{\overline{X}_n-\mu_0}{\sigma/\sqrt{n}} \overset{n\to+\infty}{\to} Z$ in legge, dove $Z \sim \mathcal{N}(0,1)$. Quindi

$$\alpha = \mathbb{P}\Big(\Big\{\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} < -\frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big) \simeq \mathbb{P}\Big(\Big\{Z < -\frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big) = \mathbb{P}\Big(\Big\{Z > \frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big)\,.$$

Introduciamo il valore z_α tale che

$$\mathbb{P}(\{Z > z_\alpha\}) = \alpha\,.$$

Allora, scegliendo

$$\frac{\delta}{\sigma/\sqrt{n}} = z_\alpha \implies \delta = \frac{\sigma}{\sqrt{n}} z_\alpha\,,$$

otteniamo la condizione che definisce il livello di significatività.

In conclusione, la regione critica è

$$R_c = \Big\{(x_1,\dots,x_n) \in R(X_1,\dots,X_n) \ :\ \overline{x}_n < \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha\Big\}\,,$$

e decidiamo come segue:

- Se $\overline{x}_n < \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha$, i dati sono sufficientemente significativi da rifiutare H_0. L'ipotesi nulla H_0 viene rifiutata (con livello di significatività α).
- Se $\overline{x}_n \geq \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha$, i dati non sono sufficientemente significativi da rifiutare. L'ipotesi nulla H_0 non viene rifiutata (con livello di significatività α).

In questo caso è anche possibile calcolare esplicitamente il p-value dei dati. Utilizzando il fatto che la funzione di distribuzione cumulativa della normale standard è strettamente crescente, otteniamo che

$$\begin{aligned}
p\text{-value} &= \inf\Big\{\alpha \ :\ \overline{x}_n - \mu_0 < -\frac{\sigma}{\sqrt{n}} z_\alpha\Big\} = \inf\Big\{\alpha \ :\ \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha\Big\} \\
&= \inf\Big\{\alpha \ :\ \Phi\Big(\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}\Big) < \Phi(-z_\alpha)\Big\} = \inf\Big\{\alpha \ :\ \Phi\Big(\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}\Big) < \alpha\Big\} \\
&= \Phi\Big(\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}\Big)\,.
\end{aligned}$$

Calcoliamo il p-value utilizzando la tavola della legge normale:

$$p\text{-value} = \Phi\Big(\frac{722.31 - 730}{20/\sqrt{32}}\Big) = \Phi(-2.175) = 1 - \Phi(2.175) = 1 - 0.9852 = 1.48\%\,.$$

Quindi:

1. L'ipotesi nulla è rifiutata con significatività 5%.
2. Il più piccolo livello di significatività per cui i dati permettono di rifiutare l'ipotesi nulla è 1.48%.