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: settembre 2023 - II
Matricola: ______________________ Data: 19/09/2023

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Il numero di veicoli venduti da un importante showroom di auto in un giorno è stato registrato per 10 giorni lavorativi. I dati sono i seguenti:

$$20 \quad 15 \quad 18 \quad 3 \quad 10 \quad 17 \quad 21 \quad 19 \quad 25 \quad 28$$

1. Determinare i quartili (esclusivi) dei dati.
2. Determinare eventuali dati anomali e sospetti.
3. Tracciare un box plot.

Soluzione. 1. Per prima cosa ordiniamo i dati:

$$3 \quad 10 \quad 15 \quad 17 \quad 18 \quad 19 \quad 20 \quad 21 \quad 25 \quad 28$$

Abbiamo $n = 10$ dati.

Calcoliamo il primo quartile: $\frac{n+1}{4} = \frac{11}{4} = 2 + 0.75$. Allora

$$Q_1 = (1 - 0.75)x_2 + 0.75x_3 = 0.25 \cdot 10 + 0.75 \cdot 15 = 13.75\,.$$

Calcoliamo il secondo quartile: $(n+1)\frac{2}{4} = 11\frac{2}{4} = 5 + 0.5$. Allora

$$Q_2 = (1 - 0.5)x_5 + 0.5x_6 = 0.5 \cdot 18 + 0.5 \cdot 19 = 18.5\,.$$

Calcoliamo il terzo quartile: $(n+1)\frac{3}{4} = 11\frac{3}{4} = 8 + 0.25$. Allora

$$Q_3 = (1 - 0.25)x_8 + 0.25x_9 = 0.75 \cdot 21 + 0.25 \cdot 25 = 22\,.$$

2. Per determinare i dati anomali e sospetti calcoliamo il range interquartile

$$IQR = Q_3 - Q_1 = 22 - 13.75 = 8.25\,.$$

I dati anomali apparterrebbero agli intervalli

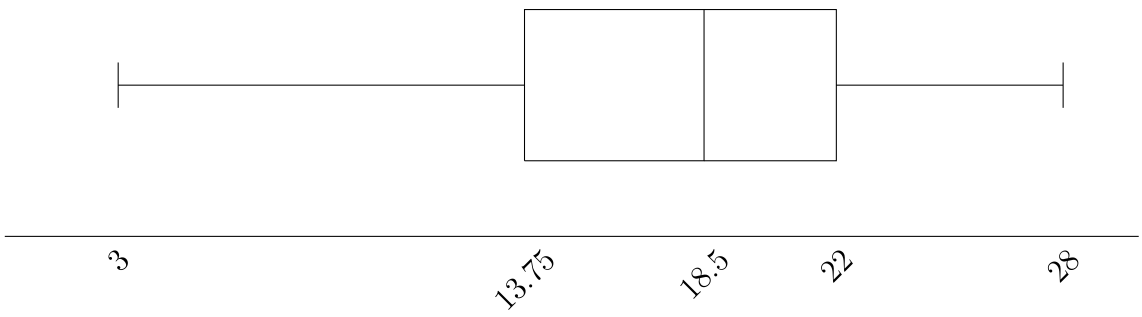
$$(-\infty, Q_1 - 3 \cdot IQR] \cup [Q_3 + 3 \cdot IQR, +\infty) = (-\infty, -11] \cup [46.75, +\infty)\,,$$

quindi non ci sono dati anomali. I dati sospetti appartengono agli intervalli

$$(Q_1 - 3 \cdot IQR, Q_1 - 1.5 \cdot IQR] \cup [Q_3 + 1.5 \cdot IQR, Q_3 + 3 \cdot IQR) = (-11, 1.375] \cup [34.375, 46.75)\,,$$

quindi non ci sono dati sospetti.

3. Segue il box-plot.



Esercizio 2. (7 punti) Si consideri un vettore aleatorio (X_1, X_2) avente funzione di probabilità congiunta descritta dalla seguente tabella:

X_2 \ X_1	1	2	3
1	a_{11}	a_{21}	a_{31}
2	a_{12}	a_{22}	a_{32}
3	a_{13}	a_{23}	a_{33}

con $a_{11}, a_{21}, a_{31}, a_{12}, a_{22}, a_{32}, a_{13}, a_{23}, a_{33} > 0$.

1. Calcolare il range dei possibili valori che possono essere assunti dalla variabile aleatoria $Y = \min\{X_1 + X_2, X_1 \cdot X_2\}$ (valore più piccolo tra $X_1 + X_2$ e $X_1 \cdot X_2$).

Si assuma che i valori nel range di Y siano equiprobabili.

2. Determinare i valori di a_{11}, a_{22}, a_{33}.

Si assuma che:

- $\mathbb{P}(\{X_1 = 2\} \cap \{X_2 = 1\}) = \frac{1}{12}$.
- $\mathbb{P}(\{X_1 = 3\} \cap \{X_2 = 1\}) = \mathbb{P}(\{W < \frac{1}{12}\})$ dove $W \sim \mathrm{U}(0, 1)$.
- $\mathbb{P}(\{X_1 = 3\} | \{X_2 = 2\}) = \frac{1}{4}$.

3. Determinare i valori rimanenti $a_{21}, a_{31}, a_{12}, a_{32}, a_{13}, a_{23}$.

4. Stabilire se X_1 e X_2 sono indipendenti.

Soluzione. 1. Per calcolare il range della variabile aleatoria Y, scriviamo tutti i possibili valori che possono essere assunti da Y.

$$X_1 = 1, X_2 = 1 \implies X_1 + X_2 = 2, X_1 \cdot X_2 = 1 \implies Y = 1\,,$$

$$X_1 = 1, X_2 = 2 \implies X_1 + X_2 = 3, X_1 \cdot X_2 = 2 \implies Y = 2\,,$$
$$X_1 = 2, X_2 = 1 \implies X_1 + X_2 = 3, X_1 \cdot X_2 = 2 \implies Y = 2\,,$$

$$X_1 = 1, X_2 = 3 \implies X_1 + X_2 = 4, X_1 \cdot X_2 = 3 \implies Y = 3\,,$$
$$X_1 = 3, X_2 = 1 \implies X_1 + X_2 = 4, X_1 \cdot X_2 = 3 \implies Y = 3\,,$$

$$X_1 = 2, X_2 = 2 \implies X_1 + X_2 = 4, X_1 \cdot X_2 = 4 \implies Y = 4\,,$$

$$X_1 = 2, X_2 = 3 \implies X_1 + X_2 = 5, X_1 \cdot X_2 = 6 \implies Y = 5\,,$$
$$X_1 = 3, X_2 = 2 \implies X_1 + X_2 = 5, X_1 \cdot X_2 = 6 \implies Y = 5\,,$$

$$X_1 = 3, X_2 = 3 \implies X_1 + X_2 = 6, X_1 \cdot X_2 = 9 \implies Y = 6\,,$$

Quindi $R(Y) = \{1, 2, 3, 4, 5, 6\}$ (osserviamo che ciascuno di questi valori ha probabilità positiva).

2. Scrivendo per quali valori di X_1, X_2 vengono assunti i valori di Y, otteniamo che

$$\begin{aligned}
\mathbb{P}(\{Y = 1\}) &= a_{11} \\
\mathbb{P}(\{Y = 2\}) &= a_{12} + a_{21} \\
\mathbb{P}(\{Y = 3\}) &= a_{13} + a_{31} \\
\mathbb{P}(\{Y = 4\}) &= a_{22} \\
\mathbb{P}(\{Y = 5\}) &= a_{23} + a_{32} \\
\mathbb{P}(\{Y = 6\}) &= a_{33}\,.
\end{aligned}$$

Poiché questi eventi sono equiprobabili, otteniamo che

$$a_{11} = a_{12} + a_{21} = a_{13} + a_{31} = a_{22} = a_{23} + a_{32} = a_{33} = b\,,$$

dove b è da determinare. Ricordiamo che è necessario avere

$$a_{11} + a_{12} + a_{21} + a_{13} + a_{31} + a_{22} + a_{23} + a_{32} + a_{33} = 1$$

da cui segue che

$$6b = 1 \implies b = \frac{1}{6}\,.$$

Aggiorniamo la tabella:

X_2 \ X_1	1	2	3
1	$1/6$	a_{21}	a_{31}
2	a_{12}	$1/6$	a_{32}
3	a_{13}	a_{23}	$1/6$

dove

$$a_{12} + a_{21} = a_{13} + a_{31} = a_{23} + a_{32} = \frac{1}{6}\,.$$

3. Imponiamo la prima condizione

$$\frac{1}{12} = \mathbb{P}(\{X_1 = 2\} \cap \{X_2 = 1\}) = a_{21}$$

e utilizziamo

$$a_{12} + a_{21} = \frac{1}{6} \implies a_{12} + \frac{1}{12} = \frac{1}{6} \implies a_{12} = \frac{1}{12}\,.$$

Imponiamo la seconda condizione, ricordando che

$$\mathbb{P}(\{W < \tfrac{1}{12}\}) = \int_0^{\frac{1}{12}} 1 \, \mathrm{d}x = \frac{1}{12}\,.$$

Otteniamo

$$\frac{1}{12} = \mathbb{P}(\{W < \frac{1}{12}\}) = \mathbb{P}(\{X_1 = 3\} \cap \{X_2 = 1\}) = a_{31}\,.$$

Utilizziamo

$$a_{13} + a_{31} = \frac{1}{6} \implies a_{13} + \frac{1}{12} = \frac{1}{6} \implies a_{13} = \frac{1}{12}\,.$$

Imponiamo l'ultima condizione:

$$\frac{1}{4} = \mathbb{P}(\{X_1 = 3\}|\{X_2 = 2\}) = \frac{\mathbb{P}(\{X_1 = 3\} \cap \{X_2 = 2\})}{\mathbb{P}(\{X_2 = 2\})} = \frac{a_{32}}{a_{12} + a_{22} + a_{32}} = \frac{a_{32}}{\frac{1}{12} + \frac{1}{6} + a_{32}}$$

$$\implies a_{32} = \frac{1}{4}\Big(\frac{1}{4} + a_{32}\Big) = \frac{1}{16} + \frac{1}{4}a_{32} \implies \frac{3}{4}a_{32} = \frac{1}{16} \implies a_{32} = \frac{1}{12}\,.$$

Utilizziamo

$$a_{23} + a_{32} = \frac{1}{6} \implies a_{23} + \frac{1}{12} = \frac{1}{6} \implies a_{23} = \frac{1}{12}\,.$$

Aggiorniamo la tabella:

X_2 \ X_1	1	2	3
1	1/6	1/12	1/12
2	1/12	1/6	1/12
3	1/12	1/12	1/6

3. Osserviamo che

$$\mathbb{P}(\{X_1 = 3\}) = a_{31} + a_{32} + a_{33} = \frac{1}{12} + \frac{1}{12} + \frac{1}{6} = \frac{1}{3}$$

e, invece,

$$\mathbb{P}(\{X_1 = 3\}|\{X_2 = 2\}) = \frac{1}{4}\,.$$

Quindi l'evento $\{X_2 = 2\}$ influenza il valore assunto da X_1. Le due variabili non sono indipendenti.

Esercizio 3. (8 punti) Alice e Bob fanno un gioco. Chiamano un numero verde e attendono che un/a operatore/trice risponda. Il tempo di attesa per la risposta è distribuito con legge esponenziale. Se la risposta avviene entro i 10 minuti vince Alice, altrimenti vince Bob. Supponiamo che il gioco sia equo (Alice vince con il 50% di probabilità).

1. In media, dopo quanto tempo risponde l'operatore/trice?

2. Calcolare la probabilità che la risposta avvenga tra 5 minuti e 10 minuti.

3. Alice e Bob chiamano il numero verde. Hanno aspettato 5 minuti e ancora nessuno ha risposto. Come viene aggiornata la probabilità che Bob vinca? Motivare la risposta.

Alice e Bob giocano 2 volte in sequenza. Appena qualcuno risponde, chiudono la chiamata e richiamano. Si assumano i tempi di attesa indipendenti.

4. Calcolare la varianza della durata dell'intera sequenza di chiamate.

5. Calcolare la probabilità che l'intera sequenza di chiamate duri più di 30 min. (Suggerimento: integrare per parti)

Soluzione. Consideriamo la variabile aleatoria

$$X = \text{“tempo di attesa per una chiamata”} \sim \text{Exp}(\lambda)\,.$$

Ci viene detto che il gioco è equo, quindi

$$\frac{1}{2} = \mathbb{P}(\{X < 10\}) = \int_0^{10} \lambda e^{-\lambda x}\ \mathrm{d}x = \Big[-e^{-\lambda x} \Big]_0^{10} = 1 - e^{-10\lambda}\,.$$

Da questa equazione possiamo ricavare λ:

$$e^{-10\lambda} = \frac{1}{2} \implies -10\lambda = \ln\left(\frac{1}{2}\right) = -\ln 2 \implies \lambda = \frac{\ln 2}{10} \simeq 6.93 \cdot 10^{-2}\,.$$

1. Calcoliamo
$$\mathbb{E}(X) = \frac{1}{\lambda} = \frac{10}{\ln 2} \simeq 14.43\,.$$

2. Calcoliamo
$$\begin{aligned}\mathbb{P}(\{5 \le X \le 10\}) &= \int_5^{10} \lambda e^{-\lambda x}\ \mathrm{d}x = \Big[-e^{-\lambda x} \Big]_5^{10} = e^{-5\lambda} - e^{-10\lambda} = e^{-5\frac{\ln 2}{10}} - e^{-10\frac{\ln 2}{10}} \\ &= e^{-\frac{\ln 2}{2}} - e^{-\ln 2} = \frac{1}{\sqrt{2}} - \frac{1}{2} \simeq 20.71\%\,.\end{aligned}$$

3. Utilizziamo l'assenza di memoria della legge geometrica per calcolare
$$\mathbb{P}(\{X > 10\} | \{X > 5\}) = \mathbb{P}(\{X > 5\}) = e^{-5\lambda} = \frac{1}{\sqrt{2}} \simeq 70.71\%\,.$$

4. Consideriamo due variabili aleatorie indipendenti $X_1, X_2 \sim \text{Exp}(\lambda)$. Il tempo dell'intera sequenza è $X_1 + X_2$. Poiché X_1 e X_2 sono indipendenti, abbiamo che

$$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2) = \frac{1}{\lambda^2} + \frac{1}{\lambda^2} = \frac{2}{\lambda^2} \simeq 416.27\,.$$

5. Il tempo totale $X_1 + X_2$ è distribuito con legge Gamma$(2, \lambda)$, quindi ha densità

$$f_{X_1+X_2}(x) = \begin{cases} \frac{\lambda^2}{\Gamma(2)} x^{2-1} e^{-\lambda x}\,, & \text{se } x > 0\,, \\ 0\,, & \text{altrimenti.} \end{cases}$$

Ricordiamo che $\Gamma(2) = (2-1)! = 1$. Possiamo allora calcolare, integrando per parti,

$$\begin{aligned}
\mathbb{P}(\{X_1 + X_2 > 30\}) &= \int_{30}^{+\infty} \lambda^2 x e^{-\lambda x} \, \mathrm{d}x \\
&= \Big[-\lambda x e^{-\lambda x} \Big]_{30}^{+\infty} + \int_{30}^{+\infty} \lambda e^{-\lambda x} \, \mathrm{d}x \\
&= 30\lambda e^{-30\lambda} + e^{-30\lambda} = (30\lambda + 1) e^{-30\lambda} \\
&= (\ln 8 + 1) \frac{1}{8} = 38.49\% \,.
\end{aligned}$$

Esercizio 4. (7 punti) Supponiamo di eseguire uno studio immunologico in un campione di individui, studiando la reazione all'antigene nel punto di inoculo. Viene misurato il diametro in mm dell'alone cutaneo in un campione, ottenendo i seguenti dati:

$$21 \quad 21 \quad 40 \quad 36 \quad 32 \quad 25 \quad 8 \quad 27$$

Si assuma che la distribuzione del diametro sia normale.

1. Si può stabilire con significatività del 5% che la media del diametro è diversa da 30 mm?
2. Si può stabilire con significatività del 5% che la varianza del diametro è superiore a 100 mm^2?

In entrambi i casi, derivare le formule utilizzate per rispondere.

Soluzione. 1. Si tratta di una popolazione $X \sim \mathcal{N}(\mu, \sigma^2)$ da cui è estratto un campione casuale $X_1, \dots, X_n$ con $n = 8$. La media μ e la varianza σ^2 sono incognite.

Sia $\mu_0 = 30$. Consideriamo il seguente test di ipotesi

$$H_0 : \mu = \mu_0 \,, \quad H_1 : \mu \neq \mu_0 \,,$$

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

Poiché l'ipotesi alternativa è $H_1 : \mu \neq \mu_0$, i dati saranno significativi se la media è sufficientemente distante da μ_0. La regione critica è allora della forma

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

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

Assumiamo l'ipotesi nulla H_0 vera, cioè $\mu = \mu_0$. Quindi $X_1, \dots, X_n \sim \mathcal{N}(\mu_0, \sigma^2)$. Non è nota σ^2, quindi verrà stimata dalla varianza campionaria $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2$, dove $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ è la media campionaria. Dalla definizione di livello di significatività:

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

Poiché $X_1, \dots, X_n \sim \mathcal{N}(\mu_0, \sigma^2)$ indipendenti, si ha che $T_{n-1} = \frac{\overline{X}_n - \mu_0}{S_n/\sqrt{n}} \sim t(n-1)$. Quindi, per la simmetria della t-Student

$$\begin{aligned}
\alpha &= \mathbb{P}\Big(\Big\{|T_{n-1}| > \frac{\delta}{S_n/\sqrt{n}}\Big\}\Big) = \mathbb{P}\Big(\Big\{T_{n-1} < -\frac{\delta}{S_n/\sqrt{n}}\Big\}\Big) + \mathbb{P}\Big(\Big\{T > \frac{\delta}{S_n/\sqrt{n}}\Big\}\Big) \\
&= 2\mathbb{P}\Big(\Big\{T_{n-1} > \frac{\delta}{S_n/\sqrt{n}}\Big\}\Big) \,,
\end{aligned}$$

da cui

$$\mathbb{P}\Big(\Big\{T_{n-1} > \frac{\delta}{S_n/\sqrt{n}}\Big\}\Big) = \frac{\alpha}{2}\,.$$

Introduciamo il valore $t_{n-1,\alpha/2}$ tale che

$$\mathbb{P}(\{T_{n-1} > t_{n-1,\alpha/2}\}) = \frac{\alpha}{2}\,.$$

Allora, scegliendo

$$\frac{\delta}{S_n/\sqrt{n}} = t_{n-1,\alpha/2} \implies \delta = \frac{S_n}{\sqrt{n}} t_{n-1,\alpha/2}\,,$$

otteniamo la condizione che definisce la significatività.

In conclusione, la regione critica è

$$R_c = \Big\{(x_1,\dots,x_n) \in R(X_1,\dots,X_n) \ : \ |\overline{x}_n - \mu_0| > \frac{s_n}{\sqrt{n}} t_{n-1,\alpha/2}\Big\}\,.$$

Calcoliamo

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{8}(21+21+40+36+32+25+8+27) = 26.25$$

e

$$s_n^2 = \frac{1}{n-1}\Big(\sum_{i=1}^n x_i^2 - n\overline{x}_n^2\Big) = \frac{1}{7}(21^2+21^2+40^2+36^2+32^2+25^2+8^2+27^2-8\cdot 26.25^2)$$
$$= 101.07\,.$$

Dalle tavole otteniamo

$$t_{n-1,\alpha/2} = t_{7,0.025} \simeq 2.365\,.$$

Quindi

$$|\overline{x}_n - \mu_0| = |26.25 - 30| = 3.75\,, \qquad \frac{s_n}{\sqrt{n}} t_{n-1,\alpha/2} = \frac{\sqrt{101.07}}{\sqrt{8}} 2.365 \simeq 8.41\,.$$

I dati non sono significativi al 5% per stabilire che la media è diversa da 30.

2. Sia $\sigma_0^2 = 100$. Consideriamo il seguente test di ipotesi

$$H_0 : \sigma^2 \le \sigma_0^2\,, \quad H_1 : \sigma^2 > \sigma_0^2\,,$$

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

Poiché l'ipotesi alternativa è $H_1 : \sigma^2 > \sigma_0^2$, i dati saranno significativi se la varianza è sufficientemente più grande di σ_0^2. La regione critica è allora della forma

$$R_c = \{(x_1,\dots,x_n) \in R(X_1,\dots,X_n) \ : s_n^2 > \delta\sigma_0^2\}\,,$$

dove $s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x}_n)^2$ e $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ sono la varianza e la media calcolata sulla realizzazione $x_1,\dots,x_n$ del campione casuale.

Assumiamo l'ipotesi nulla H_0 vera, cioè $\sigma^2 \le \sigma_0^2$. Dalla definizione di livello di significatività e utilizzando la varianza campionaria $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$,

$$\mathbb{P}(\{(X_1,\dots,X_n) \in R_c\}) = \mathbb{P}(\{S_n^2 > \delta\sigma_0^2\}) \le \mathbb{P}(\{S_n^2 > \delta\sigma^2\})$$
$$= \mathbb{P}\Big(\Big\{\frac{(n-1)S_n^2}{\sigma^2} > (n-1)\delta\Big\}\Big)\,.$$

Poiché $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ indipendenti, si ha che $Q_{n-1} = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2(n-1)$. Introduciamo il valore $\chi^2_{n-1,\alpha}$ tale che

$$\mathbb{P}(\{Q_{n-1} > \chi^2_{n-1,\alpha}\}) = \alpha\,.$$

Allora, scegliendo

$$(n-1)\delta = \chi^2_{n-1,\alpha} \implies \delta = \frac{\chi^2_{n-1,\alpha}}{n-1}\,,$$

otteniamo

$$\mathbb{P}(\{(X_1, \ldots, X_n) \in R_c\}) \leq \alpha\,,$$

(ovvero la probabilità di commettere un errore del I tipo è più piccola di α).

In conclusione, la regione critica è

$$R_c = \left\{(x_1, \ldots, x_n) \in R(X_1, \ldots, X_n) \;:\; s_n^2 > \frac{\chi^2_{n-1,\alpha}}{n-1}\sigma_0^2\right\},$$

Dalle tavole otteniamo

$$\chi^2_{n-1,\alpha} = \chi^2_{7,0.05} \simeq 14.067\,.$$

I dati forniscono i seguenti valori

$$s_n^2 = 101.07\,, \quad \frac{\chi^2_{n-1,\alpha}}{n-1}\sigma_0^2 = \frac{14.067}{7}100 \simeq 200.96\,.$$

I dati non sono significativi al 5% per stabilire che la varianza è più grande di 100.