Soluzioni Esame di Probabilità e Statistica [3231]

Soluzioni 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: ________________	A.A.: 2022/2023
Nome: ________________	Docente: Gianluca Orlando
Matricola: ________________	Appello: novembre 2022
Corso di studi: ________________	Data: 11/11/2022

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) I voti ottenuti dagli studenti e dalle studentesse a un appello dell'esame di Probabilità e Statistica[1] sono i seguenti:

26 8 28 5 27 30 26 18 28 26 25 20 22 30 20 21

1. Determinare i quartili.
2. Determinare eventuali dati anomali o sospetti.
3. Tracciare un box-plot.

Soluzione. 1. Ordiniamo i dati:

5 8 18 20 20 21 22 25 26 26 26 27 28 28 30 30

Denotiamo con $x_1, \dots, x_{16}$ i dati ordinati. L'ampiezza del campione è $n = 16$.

Calcolo di Q_1. Per trovare il primo quartile calcoliamo $\frac{n+1}{4} = \frac{17}{4} = 4 + \frac{1}{4} = 4 + 0.25$. Quindi

$$Q_1 = (1 - 0.25)x_4 + 0.25x_5 = 20\,.$$

Calcolo di Q_2. Per trovare il secondo quartile calcoliamo $(n+1)\frac{2}{4} = \frac{17}{2} = 8 + \frac{1}{2} = 8 + 0.5$. Quindi

$$Q_2 = 0.5x_8 + 0.5x_9 = 25.5\,.$$

Calcolo di Q_3. Per trovare il terzo quartile calcoliamo $(n+1)\frac{3}{4} = 17\frac{3}{4} = 12 + \frac{3}{4} = 12 + 0.75$. Quindi

$$Q_3 = (1 - 0.75)x_{12} + 0.75x_{13} = 27.75\,.$$

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

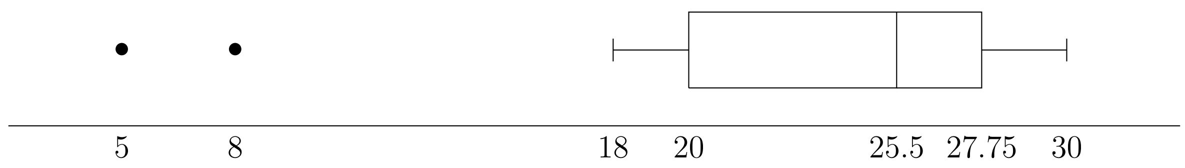
$$IQR = Q_3 - Q_1 = 27.75 - 20 = 7.75\,.$$

[1] I dati sono generati casualmente e non si riferiscono a fatti realmente accaduti.

I dati anomali sono più grandi di $Q_3 + 3IQR = 51$ o più piccoli di $Q_1 - 3IQR = -3.25$. Quindi non ci sono dati anomali.

I dati sospetti cadono tra $Q_3 + 1.5IQR = 39.375$ e $Q_3 + 3IQR = 51$ oppure tra $Q_1 - 3IQR = -3.25$ e $Q_1 - 1.5IQR = 8.375$. Quindi 5 e 8 sono dati sospetti.

3. Segue il box-plot:



Esercizio 2. (7 punti) Un ristorante studia i suoi clienti fissi e gli effetti della nuova campagna di marketing adottata dal ristorante tramite *ad* su un *social network*. Gli *ad* risultano efficaci su una certa proporzione p dei clienti. Lo studio porta a queste conclusioni:

- Per un cliente che è influenzato dagli *ad*, il numero di visite annue al ristorante è distribuito con una legge di Poisson con una media di 7 visite all'anno;
- Per un cliente che non è influenzato dagli *ad* (nella restante proporzione della popolazione $1 - p$), il numero di visite annue al ristorante è distribuito con una legge di Poisson con una media di 4 visite all'anno.

Rispondere ai seguenti quesiti:

1. Determinare (in funzione di p) la probabilità che il numero di visite annue di un cliente sia uguale a un dato numero $k = 0, 1, 2, \dots$.
2. Un'analisi mostra che il 50% dei clienti visita il ristorante almeno 5 volte all'anno. (Si legga: la probabilità che il numero di visite di un cliente sia maggiore o uguale a 5 è il 50%.) Ricavare in questo caso la proporzione p dei clienti che viene influenzata dagli *ad*.

Soluzione. Lo svolgimento è analogo a quella dell'esercizio sui farmaci svolto nella Lezione 13 il 12/04/2022.

Consideriamo le seguenti variabili aleatorie

$$X = \begin{cases} 1 & \text{se gli } ad \text{ sono efficaci sul cliente} \\ 0 & \text{se gli } ad \text{ non sono efficaci sul cliente} \end{cases}$$

$$Y = \text{“numero di visite annuali del cliente”}\,.$$

La variabile X è distribuita con una legge di Bernoulli con parametro p. La variabile Y ha una distribuzione dipendente dal valore assunto da X. Dalla traccia abbiamo che

$$\mathbb{P}(\{Y = k\}|\{X = 1\}) = e^{-7}\frac{7^k}{k!}$$

perché questa è una legge di Poisson con media 7 (ricordiamo che il parametro della legge di Poisson è uguale al valore atteso) e

$$\mathbb{P}(\{Y = k\}|\{X = 0\}) = e^{-4}\frac{4^k}{k!}$$

perché questa è una legge di Poisson con media 4.

1. Dal teorema della probabilità totale abbiamo che

$$\begin{aligned}\mathbb{P}(\{Y=k\}) &= \mathbb{P}(\{Y=k\}|\{X=0\})\mathbb{P}(\{X=0\}) + \mathbb{P}(\{Y=k\}|\{X=1\})\mathbb{P}(\{X=1\})\\ &= e^{-4}\frac{4^k}{k!}(1-p) + e^{-7}\frac{7^k}{k!}p\,.\end{aligned}$$

2. Utilizzando la formula ottenuta nel primo punto

$$\begin{aligned}&\mathbb{P}(\{Y\geq 5\})\\ &= 1-\mathbb{P}(\{Y\leq 4\}) = 1-\sum_{k=0}^{4}\mathbb{P}(\{Y=k\})\\ &= 1-(1-p)\sum_{k=0}^{4}e^{-4}\frac{4^k}{k!} + p\sum_{k=0}^{4}e^{-7}\frac{7^k}{k!}\\ &= 1-(1-p)e^{-4}\Big(1+4+\frac{4^2}{2!}+\frac{4^3}{3!}+\frac{4^4}{4!}\Big) - pe^{-7}\Big(1+7+\frac{7^2}{2!}+\frac{7^3}{3!}+\frac{7^4}{4!}\Big)\\ &\simeq 1-62.88\%(1-p)-17.30\%p = 1-62.88\%+62.88\%p-17.30\%p = 37.12\%+45.58\%p\end{aligned}$$

La traccia fornisce la seguente informazione

$$\mathbb{P}(\{Y\geq 5\}) = 50\%\,.$$

Possiamo allora imporre:

$$37.12\%+45.58\%p = 50\% \implies p\simeq 28.26\%\,.$$

Esercizio 3. (7 punti) L'ufficio informazioni di una compagnia ha due numeri verdi distinti. I tempi di attesa per parlare con gli operatori sono, per entrambi i numeri, variabili aleatorie distribuite con legge esponenziale con media 10 minuti. Inoltre i due tempi di attesa si possono considerare indipendenti. Avendo a disposizione due telefoni, decidi di chiamare contemporaneamente i due numeri.

1. Qual è la probabilità che qualcuno risponda dal primo numero dopo 5 minuti? E che qualcuno risponda dal secondo numero dopo 5 minuti?

2. Qual è la probabilità di attendere meno di 5 minuti fino alla risposta da uno dei due numeri (non importa quale dei due)?

3. (Domanda bonus: come se fosse un quesito teorico) Quanto tempo aspetterai in media fino alla risposta da uno dei due numeri?

Soluzione. Il tempo di attesa ai due sportelli è dato da due variabili aleatorie

$$X_1 = \text{“tempo di attesa al primo sportello”} \sim \mathrm{Exp}\Big(\frac{1}{10}\Big)\,,$$

$$X_2 = \text{“tempo di attesa al secondo sportello”} \sim \mathrm{Exp}\Big(\frac{1}{10}\Big)\,.$$

In entrambi i casi abbiamo utilizzato il fatto che per $X_1 \sim \mathrm{Exp}(\lambda)$ si ha $\mathbb{E}(X_1) = \frac{1}{\lambda}$.

1. La probabilità che qualcuno risponda dal primo numero dopo 5 minuti è

$$\mathbb{P}(\{X_1 \geq 5\}) = \int_5^{+\infty} \frac{1}{10} e^{-\frac{1}{10}x} \, dx = \Big[- e^{-\frac{1}{10}x} \Big]_5^{+\infty} = e^{-\frac{1}{10}5} = e^{-\frac{1}{2}} \simeq 60.65\% \, .$$

La stessa probabilità è $\mathbb{P}(\{X_2 \geq 5\})$ poiché X_1 e X_2 sono identicamente distribuite.

2. La risposta da uno dei due numeri entro i 5 minuti se almeno uno tra X_1 e X_2 è più piccolo di 5, ovvero si verifica il complementare dell'evento: entrambi X_1 e X_2 sono maggiori di 5. Utilizzando l'indipendenza:

$$\begin{aligned} \mathbb{P}(\{X_1 < 5\} \cup \{X_2 < 5\}) &= 1 - \mathbb{P}(\{X_1 \geq 5\} \cap \{X_2 \geq 5\}) = 1 - \mathbb{P}(\{X_1 \geq 5\})\mathbb{P}(\{X_2 \geq 5\}) \\ &= 1 - 60.65\% \cdot 60.65\% \simeq 63.22\% \, . \end{aligned}$$

3. Per calcolare il tempo medio di attesa integriamo i valori assunti dal tempo di attesa (il più piccolo tra il valore assunto da X_1 e da X_2) contro la densità di probabilità congiunta, che è data dal prodotto delle densità di probabilità essendo le due variabili aleatorie indipendenti:

$$\begin{aligned} &\int_0^{+\infty} \int_0^{+\infty} \min\{x_1, x_2\} \frac{1}{10} e^{-\frac{1}{10}x_1} \frac{1}{10} e^{-\frac{1}{10}x_2} \, dx_1 \, dx_2 = \\ &= \int_0^{+\infty} \int_0^{x_2} x_1 \frac{1}{10} e^{-\frac{1}{10}x_1} \frac{1}{10} e^{-\frac{1}{10}x_2} \, dx_1 \, dx_2 + \int_0^{+\infty} \int_{x_2}^{+\infty} x_2 \frac{1}{10} e^{-\frac{1}{10}x_1} \frac{1}{10} e^{-\frac{1}{10}x_2} \, dx_1 \, dx_2 \\ &= \int_0^{+\infty} \frac{1}{10} e^{-\frac{1}{10}x_2} \int_0^{x_2} \frac{1}{10} x_1 e^{-\frac{1}{10}x_1} \, dx_1 \, dx_2 + \int_0^{+\infty} \frac{1}{10} x_2 e^{-\frac{1}{10}x_2} \int_{x_2}^{+\infty} \frac{1}{10} e^{-\frac{1}{10}x_1} \, dx_1 \, dx_2 \\ &= \int_0^{+\infty} \frac{1}{10} e^{-\frac{1}{10}x_2} \Big[- 10 e^{-\frac{1}{10}x_1} - x_1 e^{-\frac{1}{10}x_1} \Big]_0^{x_2} \, dx_2 + \int_0^{+\infty} \frac{1}{10} x_2 e^{-\frac{1}{10}x_2} \Big[- e^{-\frac{1}{10}x_1} \Big]_{x_2}^{+\infty} \, dx_2 \\ &= \int_0^{+\infty} \Big(- e^{-\frac{1}{5}x_2} - \frac{1}{10} x_2 e^{-\frac{1}{5}x_2} + e^{-\frac{1}{10}x_2} \Big) \, dx_2 + \int_0^{+\infty} \frac{1}{10} x_2 e^{-\frac{1}{5}x_2} \, dx_2 \\ &= \int_0^{+\infty} \Big(- e^{-\frac{1}{5}x_2} + e^{-\frac{1}{10}x_2} \Big) \, dx_2 = \Big[5 e^{-\frac{1}{5}x_2} - 10 e^{-\frac{1}{10}x_2} \Big]_0^{+\infty} = -5 + 10 = 5 \, . \end{aligned}$$

Esercizio 4. (8 punti) Si vuole studiare l'effetto della delaminazione sulla frequenza naturale delle travi realizzate con laminati compositi. Si effettua il seguente esperimento statistico: si considera un campione di otto travi delaminate, le si sottopongono a carichi e se ne misurano le frequenze risultanti (in Hertz). Per un esperimento statistico si osservano i seguenti dati:

230.66 233.05 232.58 229.48 232.58 235.76 229.43 234.13

Si supponga che i dati provengano da una popolazione distribuita con legge normale.

1. Calcolare sui dati un intervallo di confidenza bilaterale al 90% sulla frequenza naturale media della popolazione.

2. Supponiamo di effettuare 20 esperimenti statistici indipendenti (ottenendo per ogni esperimento statistico nuovi valori) e di calcolare sui dati di ogni esperimento statistico un intervallo di confidenza bilaterale al 90% come nel punto precedente. Qual è la probabilità che almeno 4 volte la frequenza naturale media della popolazione sia fuori dall'intervallo di confidenza calcolato sui dati?

Soluzione. 1. Stiamo considerando un campione casuale $X_1, \ldots, X_n$, $n = 8$, estratto da una popolazione normale. Quindi $X_1, \ldots, X_n$ sono indipendenti e tutte distribuite con legge normale $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Osserviamo che la media della popolazione μ e la varianza σ^2 non sono note.

Un intervallo di confidenza bilaterale al $\beta = 90\%$ per la media della popolazione μ è un intervallo $[U_n, V_n]$ con estremi variabili aleatorie tali che

$$\beta = \mathbb{P}(\{U_n \leq \mu \leq V_n\}) \,.$$

Consideriamo la media campionaria $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ e la varianza campionaria $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_n - \bar{X}_n)^2$. Ricordiamo che, poiché $X_1, \dots, X_n$ hanno distribuzione normale, la variabile aleatoria $T_{n-1} = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$ è distribuita come una t-Student con $n-1 = 7$ gradi di libertà. Utilizziamo questo fatto per determinare gli estremi dell'intervallo di confidenza:

$$\begin{aligned}
\beta = \mathbb{P}(\{U_n \leq \mu \leq V_n\}) &= \mathbb{P}\Big(\Big\{\frac{\bar{X}_n - V_n}{S_n/\sqrt{n}} \leq \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \leq \frac{\bar{X}_n - U_n}{S_n/\sqrt{n}}\Big\}\Big) \\
&= \mathbb{P}\Big(\Big\{\frac{\bar{X}_n - V_n}{S_n/\sqrt{n}} \leq T_{n-1} \leq \frac{\bar{X}_n - U_n}{S_n/\sqrt{n}}\Big\}\Big) \\
&= \mathbb{P}\Big(\Big\{T_{n-1} \leq \frac{\bar{X}_n - U_n}{S_n/\sqrt{n}}\Big\}\Big) - \mathbb{P}\Big(\Big\{T_{n-1} < \frac{\bar{X}_n - V_n}{S_n/\sqrt{n}}\Big\}\Big) \\
&= 1 - \mathbb{P}\Big(\Big\{T_{n-1} > \frac{\bar{X}_n - U_n}{S_n/\sqrt{n}}\Big\}\Big) - \mathbb{P}\Big(\Big\{T_{n-1} < \frac{\bar{X}_n - V_n}{S_n/\sqrt{n}}\Big\}\Big)
\end{aligned}$$

da cui segue

$$\mathbb{P}\Big(\Big\{T_{n-1} > \frac{\bar{X}_n - U_n}{S_n/\sqrt{n}}\Big\}\Big) + \mathbb{P}\Big(\Big\{T_{n-1} < \frac{\bar{X}_n - V_n}{S_n/\sqrt{n}}\Big\}\Big) = 1 - \beta = \alpha = 10\% \,.$$

Decidiamo di equipartire la probabilità α, ovvero

$$\mathbb{P}\Big(\Big\{T_{n-1} > \frac{\bar{X}_n - U_n}{S_n/\sqrt{n}}\Big\}\Big) = \mathbb{P}\Big(\Big\{T_{n-1} < \frac{\bar{X}_n - V_n}{S_n/\sqrt{n}}\Big\}\Big) = \frac{\alpha}{2} \,. \tag{1}$$

Sia $t_{n-1,\alpha/2}$ il quantile della t-Student tale che

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

Allora scegliendo

$$\frac{\bar{X}_n - U_n}{S_n/\sqrt{n}} = t_{n-1,\alpha/2} \,, \quad \frac{\bar{X}_n - V_n}{S_n/\sqrt{n}} = -t_{n-1,\alpha/2}$$

abbiamo, per la simmetria della t-Student, che (1) è verificata. Quindi

$$U_n = \bar{X}_n - t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}} \,, \quad V_n = \bar{X}_n + t_{n-1,\alpha/2}\frac{S_n}{\sqrt{n}} \,.$$

Abbiamo tutti gli strumenti per calcolare la realizzazione degli estremi sui dati $x_1, \dots, x_n$ osservati. La realizzazione della media campionaria sui dati è

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{8}(230.66+233.05+232.58+229.48+232.58+235.76+229.43+234.13) \simeq 232.21 \,.$$

La realizzazione della deviazione standard sui dati è

$$\begin{aligned}
s_n &= \sqrt{\frac{1}{n-1}\Big(\sum_{i=1}^n x_i^2 - n\bar{x}_n^2\Big)} = \sqrt{\frac{1}{7}(431402.0807 - 431371.8728)} \\
&= \sqrt{4.315414285714285 7143} \simeq 2.08 \,.
\end{aligned}$$

Il quantile della t-Student è calcolato dalle tabelle

$$t_{n-1,\alpha/2} = t_{7,0.05} \simeq 1.895\,.$$

Concludiamo che la realizzazione degli estremi sui dati è

$$u_n = \bar{x}_n - t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}} \simeq 230.82\,, \quad v_n = \bar{x}_n + t_{n-1,\alpha/2}\frac{s_n}{\sqrt{n}} \simeq 233.60$$

e l'intervallo di confidenza calcolato sui dati è $[230.82, 233.60]$.

2. Consideriamo l'esperimento $j = 1, \dots, 20$ e la variabile aleatoria Y_j

$$Y_j = \begin{cases} 1 & \text{se la media della popolazione è fuori dall'intervallo di confidenza}\,, \\ 0 & \text{se la media della popolazione è nell'intervallo di confidenza}\,. \end{cases}$$

Osserviamo che Y_j è distribuita come una Bernoulli con probabilità di successo $p = 10\% = \frac{1}{10}$, poiché la popolazione è fuori dall'intervallo di confidenza con il 10% di probabilità. Le $Y_1, \dots, Y_{20}$ sono indipendenti (lo dice la traccia), quindi la variabile aleatoria

$$Y = \text{“numero di successi in 20 esperimenti statistici”} = Y_1 + \cdots + Y_{20}$$

è distribuita con legge binomiale $B(20, \frac{1}{10})$. Ricordando che $\mathbb{P}(\{Y = k\}) = \binom{n}{k} p^k (1-p)^{n-k}$, possiamo calcolare la probabilità di osservare più di 4 volte un successo (media della popolazione è fuori dall'intervallo di confidenza):

$$\begin{aligned}
&\mathbb{P}(\{Y \geq 4\}) = 1 - \mathbb{P}(\{Y \leq 3\}) = 1 - \mathbb{P}(\{Y = 0\}) - \mathbb{P}(\{Y = 1\}) - \mathbb{P}(\{Y = 2\}) - \mathbb{P}(\{Y = 3\}) \\
&= 1 - \binom{20}{0}\left(\frac{1}{10}\right)^0\left(\frac{9}{10}\right)^{20} - \binom{20}{1}\left(\frac{1}{10}\right)^1\left(\frac{9}{10}\right)^{19} - \binom{20}{2}\left(\frac{1}{10}\right)^2\left(\frac{9}{10}\right)^{18} - \binom{20}{3}\left(\frac{1}{10}\right)^3\left(\frac{9}{10}\right)^{17} \\
&\simeq 13.30\%\,.
\end{aligned}$$