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.: 2021/2022
Nome: ____________	Docente: Gianluca Orlando
Matricola: ____________	Appello: febbraio 2023
Corso di studi: ____________	Data: 09/02/2023

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Viene esaminata in un campione la resistenza alla compressione del calcestruzzo quando miscelato con la cenere volante (una miscela di silice, allumina, ossido di ferro, e altri elementi). Vengono riportati i seguenti dati in megapascal:

22.4 50.2 30.4 14.2 28.9 30.5 25.8 18.4 15.3 21.1

1. Calcolare i quartili dell'insieme dei dati.
2. Determinare eventuali dati anomali o sospetti.
3. Disegnare un box-plot.

Soluzione. 1. Per prima cosa ordiniamo i dati:

14.2 15.3 18.4 21.1 22.4 25.8 28.9 30.4 30.5 50.2

Abbiamo $n = 10$ dati.

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

$$Q_1 = (1 - 0.75)x_2 + 0.75x_3 = 0.25 \cdot 15.3 + 0.75 \cdot 18.4 = 17.625\,.$$

Calcoliamo il secondo quartile: $(n+1)\frac{2}{4} = \frac{22}{4} = 5 + 0.5$. Allora

$$Q_2 = (1 - 0.5)x_5 + 0.5x_6 = 0.5 \cdot 22.4 + 0.5 \cdot 25.8 = 24.1\,.$$

Calcoliamo il terzo quartile: $(n+1)\frac{3}{4} = \frac{33}{4} = 8 + 0.25$. Allora

$$Q_3 = (1 - 0.25)x_8 + 0.25x_9 = 0.75 \cdot 30.4 + 0.25 \cdot 30.5 = 30.425\,.$$

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

$$IQR = Q_3 - Q_1 = 30.425 - 17.625 = 12.8\,.$$

I dati anomali apparterrebbero agli intervalli

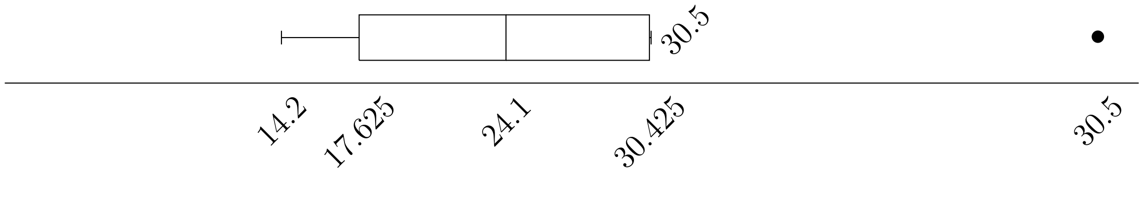
$$(-\infty, Q_1 - 3 \cdot IQR] \cup [Q_3 + 3 \cdot IQR, +\infty) = (-\infty, -20.775] \cup [68.825, +\infty)\,,$$

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

$$(Q_1 - 3 \cdot IQR, Q_1 - 1.5 \cdot IQR] \cup [Q_3 + 1.5 \cdot IQR, Q_3 + 3 \cdot IQR) = (-20.775, -1.575] \cup [49.625, 68.825)\,,$$

quindi 50.2 è un dato sospetto.

3. Segue il box-plot.



Esercizio 2. (7 punti) Alice e Bob fanno un gioco. Alice lancia consecutivamente un dado a 6 facce non truccato tante volte (i lanci sono indipendenti). Se entro il quarto lancio (compreso) esce un 6, vince Bob, altrimenti vince Alice.

1. Alice e Bob giocano una partita. Con che probabilità vince Alice?

Alice però è molto brava nel calcolo delle probabilità e si rende conto che c'è qualcosa che va a suo sfavore nelle regole di questo gioco... Decide allora di truccare il dado, all'insaputa di Bob. Dopo aver giocato alcune partite, Bob si rende conto che Alice vince con il 70% di probabilità. Inoltre, il 6 esce per la prima volta, in media, troppo tardi. Quindi l'accusa di aver barato.

2. Cosa ha fatto esattamente Alice al dado?

3. In media, dopo quale lancio esce il 6 con questo dado truccato?

Soluzione. 1. Il lancio di un dado non truccato è modellato da uno spazio di probabilità con spazio degli eventi elementari

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

in cui ciascun evento elementare ha probabilità $\frac{1}{6}$. La probabilità che esca un 6 in un lancio di un dado è quindi $\frac{1}{6}$. Chiamiamo "successo" l'evento "esce un 6".

Il gioco in questione è una sequenza di prove indipendenti, ciascuna con probabilità di successo $\frac{1}{6}$. La variabile aleatoria X (definita su uno spazio di probabilità differente) che fornisce la prima volta che viene osservato un successo in una sequenza di lanci indipendenti ha una distribuzione geometrica con parametro $\frac{1}{6}$, ovvero $X \sim \mathrm{Geo}(\frac{1}{6})$. Alice vince se il primo 6 non esce entro il quarto lancio (compreso), ovvero se esce dal quinto lancio in poi. La probabilità che si verifichi questo evento è (in questa formula $p = \frac{1}{6}$)

$$\begin{aligned}
\mathbb{P}(\{X \geq 5\}) &= \sum_{i=5}^{+\infty} \mathbb{P}(\{X = i\}) = \sum_{i=5}^{+\infty} (1-p)^{i-1} p = (1-p)^4 p \sum_{i=5}^{+\infty} (1-p)^{i-5} \\
&= (1-p)^4 p \sum_{j=0}^{+\infty} (1-p)^j = (1-p)^4 p \frac{1}{1-(1-p)} = (1-p)^4 \\
&= \left(1 - \frac{1}{6}\right)^4 = \left(\frac{5}{6}\right)^4 \simeq 48.23\%\,.
\end{aligned}$$

Si può anche ricordare direttamente la formula $(1-p)^4$, che va interpretata come: "I primi quattro risultati delle prove sono insuccessi".

In alternativa, poiché i tentativi interessanti sono pochi, si può calcolare direttamente la probabilità del complementare

$$\begin{aligned}\mathbb{P}(\{X \geq 5\}) &= 1 - \mathbb{P}(\{X \leq 4\}) = 1 - \mathbb{P}(\{X=1\}) - \mathbb{P}(\{X=2\}) - \mathbb{P}(\{X=3\}) - \mathbb{P}(\{X=4\}) \\ &= 1 - (1-p)^0 p - (1-p)^1 p - (1-p)^2 p - (1-p)^3 p \\ &= 1 - \frac{1}{6} - \left(\frac{5}{6}\right)\frac{1}{6} - \left(\frac{5}{6}\right)^2 \frac{1}{6} - \left(\frac{5}{6}\right)^3 \frac{1}{6} \simeq 48.23\% \,.\end{aligned}$$

Osserviamo che la probabilità di vittoria per Alice è leggermente meno del 50%, e questo è in accordo con la continuazione della traccia.

2. Alice ha truccato il dado in modo che il 6 esca con una probabilità diversa che chiamiamo p, diversa da $\frac{1}{6}$ e da determinare. Per capire esattamente con che probabilità esce il 6 con il dado truccato, usiamo l'informazione sulle partite vinte. Ci viene detto che Alice vince con il 70% di probabilità. Questo vuol dire, usando la formula del punto 1.,

$$0.7 = 70\% = \mathbb{P}(\{X \geq 5\}) = (1-p)^4 \implies p = 1 - (0.7)^{\frac{1}{4}} \simeq 8.53 \,.$$

Osserviamo che questa probabilità è meno di $\frac{1}{6} \simeq 16.67$, infatti Alice vuole che il 6 esca con una probabilità più bassa. Il dado truccato ha quindi una distribuzione non uniforme, ad esempio può essere fatto in modo che per gli eventi elementari dello spazio Ω introdotto inizialmente si abbia che

$$\mathbb{P}(\{6\}) = p \,, \quad \mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = \mathbb{P}(\{5\}) = \frac{1-p}{5} \,.$$

(Questa è solo una possibile configurazione, ce ne sono infinite per cui si ha la condizione richiesta.)

3. Ora la $X \sim \mathrm{Geo}(p)$. Sappiamo che $\mathbb{E}(X) = \frac{1}{p} \simeq 11.72$, quindi in media il 6 esce all'incirca dopo l'undicesimo lancio (molto di più rispetto al dado non truccato).

Esercizio 3. (7 punti) Il tempo necessario per un/a tecnico/a dell'assistenza per cambiare l'olio in un'auto è una variabile aleatoria. Se l'auto non presenta problemi, è distribuita uniformemente tra 10 e 20 minuti. Se l'auto presenta dei problemi, è distribuita uniformemente con media 20 minuti e varianza 12 min^2. In media, il 90% delle auto non presenta problemi.

1. Viene riparata un'auto. Sapendo che il tempo impiegato per cambiare l'olio è stato maggiore di 18 minuti, qual è la probabilità che l'auto avesse dei problemi?
2. Vengono riparate 2 auto (si assumano i tempi di cambio dell'olio per le due auto indipendenti). Qual è la probabilità che il cambio d'olio più rapido duri meno di 18 minuti?

Soluzione. Consideriamo la variabile aleatoria

$$X = \text{"tempo impiegato per cambio dell'olio"}$$

e la variabile aleatoria

$$Y = \text{"auto presenta problemi"} \,.$$

Per quanto riguarda la seconda variabile, possiamo utilizzare una legge di Bernoulli $Y \sim \mathrm{Be}(p)$ con $p = 90\%$ dove $\{Y = 1\}$ (successo) significa che l'auto non presenta problemi, mentre

$\{Y = 0\}$ (insuccesso) significa che l'auto presenta problemi. Infatti ricordiamo che $\mathbb{E}(Y) = p$ e ci viene detto che la media è 90%.

Per semplicità introduciamo le due variabili aleatorie ausiliarie

$$U_1 = \text{“tempo impiegato per cambio dell'olio se l'auto non presenta problemi”}$$

$$U_0 = \text{“tempo impiegato per cambio dell'olio se l'auto presenta problemi”}\,.$$

La traccia ci dice che $U_1 \sim \mathrm{U}(10, 20)$, ovvero, la densità è

$$f_{U_1}(x) = \begin{cases} \frac{1}{10}\,, & \text{se } 10 \leq x \leq 20\,, \\ 0\,, & \text{altrimenti}\,. \end{cases}$$

Per la seconda variabile si sa che $U_0 \sim \mathrm{U}(a, b)$, $\mathbb{E}(U_0) = 20$, $\mathrm{Var}(U_0) = 12$. Ricordiamo che per una variabile aleatoria con legge uniforme si ha che

$$\mathbb{E}(U_0) = \frac{a+b}{2}\,, \quad \mathrm{Var}(U_0) = \frac{(b-a)^2}{12}\,.$$

Imponiamo che

$$\begin{cases} \dfrac{a+b}{2} = 20 \\ \dfrac{(b-a)^2}{12} = 12 \end{cases} \implies \begin{cases} a+b = 40 \\ (b-a)^2 = 12^2 \end{cases} \implies \begin{cases} a+b = 40 \\ b-a = 12 \end{cases} \implies \begin{cases} 2b = 52 \\ 2a = 28 \end{cases} \implies \begin{cases} b = 26 \\ a = 14\,. \end{cases}$$

(Abbiamo usato che $b - a > 0$!) Quindi $X_2 \sim \mathrm{U}(14, 26)$, cioè

$$f_{U_0}(x) = \begin{cases} \frac{1}{12}\,, & \text{se } 14 \leq x \leq 26\,, \\ 0\,, & \text{altrimenti}\,. \end{cases}$$

Infine, sappiamo che

$$\mathbb{P}(\{X \geq x\}|\{Y = 1\}) = \mathbb{P}(\{U_1 \geq x\})\,, \quad \mathbb{P}(\{X \geq x\}|\{Y = 0\}) = \mathbb{P}(\{U_0 \geq x\})\,. \tag{1}$$

1. Ci viene chiesto di calcolare la probabilità condizionata

$$\mathbb{P}(\{Y = 0\}|\{X \geq 18\})\,.$$

Utilizziamo il Teorema di Bayes e (1):

$$\begin{aligned} \mathbb{P}(\{Y = 0\}|\{X \geq 18\}) &= \frac{\mathbb{P}(\{Y = 0\} \cap \{X \geq 18\})}{\mathbb{P}(\{X \geq 18\})} \\ &= \frac{\mathbb{P}(\{X \geq 18\}|\{Y = 0\})\mathbb{P}(\{Y = 0\})}{\mathbb{P}(\{X \geq 18\}|\{Y = 0\})\mathbb{P}(\{Y = 0\}) + \mathbb{P}(\{X \geq 18\}|\{Y = 1\})\mathbb{P}(\{Y = 1\})} \\ &= \frac{\mathbb{P}(\{U_0 \geq 18\})\mathbb{P}(\{Y = 0\})}{\mathbb{P}(\{U_0 \geq 18\})\mathbb{P}(\{Y = 0\}) + \mathbb{P}(\{U_1 \geq 18\})\mathbb{P}(\{Y = 1\})} \\ &= \frac{\int_{18}^{26} \frac{1}{12}\,\mathrm{d}x \cdot 10\%}{\int_{18}^{26} \frac{1}{12}\,\mathrm{d}x \cdot 10\% + \int_{18}^{20} \frac{1}{10}\,\mathrm{d}x \cdot 90\%} = \frac{8/12 \cdot 1/10}{8/12 \cdot 1/10 + 2/10 \cdot 9/10} \simeq 27.03\,. \end{aligned}$$

2. Abbiamo a che fare con due auto, quindi due tempi di cambio dell'olio X_1, X_2, variabili indipendenti e identicamente distribuite. Ci viene chiesto di calcolare la probabilità che il minimo tra X_1 e X_2 sia meno di 18 minuti. Possiamo calcolare la probabilità dell'evento

complementare, cioè che il minimo tra X_1 e X_2 sia maggiore di 18 minuti. Questo succede se sia X_1 è maggiore di 18 che X_2 è maggiore di 18. Quindi, usando l'indipendenza:

$$\begin{aligned}\mathbb{P}(\{\min\{X_1, X_2\} < 18\}) = 1 - \mathbb{P}(\{\min\{X_1, X_2\} \geq 18\}) &= 1 - \mathbb{P}(\{X_1 \geq 18\} \cap \{X_2 \geq 18\}) \\ &= 1 - \mathbb{P}(\{X_1 \geq 18\})\mathbb{P}(\{X_2 \geq 18\}) .\end{aligned}$$

In realtà abbiamo già calcolato le ultime probabilità nel punto precedente. Si ottengono con la formula della probabilità totale (sono uguali perché le variabili aleatorie sono identicamente distribuite)

$$\begin{aligned}&\mathbb{P}(\{X_1 \geq 18\}) = \mathbb{P}(\{X_2 \geq 18\}) \\ &= \mathbb{P}(\{U_0 \geq 18\})\mathbb{P}(\{Y = 0\}) + \mathbb{P}(\{U_1 \geq 18\})\mathbb{P}(\{Y = 1\}) = 8/12 \cdot 1/10 + 2/10 \cdot 9/10 = 24.67\% .\end{aligned}$$

Quindi

$$\mathbb{P}(\{\min\{X_1, X_2\} < 18\}) = 1 - (24.67\%)^2 = 93.91\% .$$

Esercizio 4. (8 punti) Il/la docente di Probabilità e Statistica vuole un'evidenza significativa del fatto che in questo anno accademico l'esame sia stato più difficile per gli studenti e le studentesse.[1] Negli anni accademici precedenti, la media dei voti era 24. In un appello di questo anno accademico, invece, sono stati registrati i seguenti voti:

21	26	23	25	18	24	18	28	23	26	20	20	18	20	21	20
30	25	19	23	26	26	26	26	29	28	25	18	21	26	18	27

1. I voti assegnati nell'ultimo appello sono significativi al 5% per concludere che la media è effettivamente più bassa rispetto agli anni precedenti? (N.B.: ricavare le formule)
2. In questo anno accademico si svolgono 8 appelli, come programmato. Dopo ciascun appello si registrano i voti come nel punto precedente. A volte si conclude che la media è più bassa rispetto agli anni precedenti, altre volte no. Assumendo che la media di questo anno accademico sia rimasta 24, qual è la probabilità di commettere un errore (strettamente) meno di 6 volte arrivando a conclusioni con l'analisi precedente?

Soluzione. 1. Abbiamo a che fare con un campione casuale $X_1, \ldots, X_n$ di variabili aleatorie indipendenti e identicamente distribuite con $n = 32$. Non conosciamo la legge delle variabili aleatorie, non conosciamo la media della popolazione $\mathbb{E}(X_i) = \mu$ e non conosciamo la varianza della popolazione $\mathrm{Var}(X_i) = \sigma^2$.

Per rispondere alla domanda impostiamo un test d'ipotesi:

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

dove $\mu_0 = 24$. Infatti, fino a prova contraria, la media della popolazione non è cambiata ed è sempre 24. Se i dati sono abbastanza significativi (media molto più bassa di 24), rifiutiamo l'ipotesi nulla a favore dell'ipotesi alternativa, che ci permette di stabilire con un certo livello di significatività che la media della popolazione è effettivamente più bassa.

Definiamo una regione critica

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

dove $\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ è la media calcolata sul campione.

[1]L'esercizio è inventato e ogni riferimento a persone o fatti realmente accaduti è puramente casuale.

Il livello di significatività α è la probabilità di commettere un errore del I tipo, ovvero di rifiutare l'ipotesi nulla quando è vera. Supponiamo che l'ipotesi nulla sia vera, cioè $\mu = \mu_0 = 24$. Allora

$$\begin{aligned}\alpha &= \mathbb{P}(\{(X_1,\dots,X_n) \in R_C\}) = \mathbb{P}(\{\bar{X}_n < \mu_0 - \delta\}) = \mathbb{P}(\{\bar{X}_n < \mu - \delta\}) = \mathbb{P}(\{\bar{X}_n - \mu < -\delta\}) \\ &= \mathbb{P}\Big(\Big\{\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < -\frac{\delta}{S_n/\sqrt{n}}\Big\}\Big)\,.\end{aligned}$$

Non avendo a disposizione la deviazione standard della popolazione, abbiamo usato lo stimatore $S_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$, dove $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ è la media campionaria. Poiché il campione è numeroso ($n \geq 30$), possiamo utilizzare teoremi di approssimazione. In questo caso utilizziamo il Teorema di Slutsky, che ci garantisce che la statistica $\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$ ha approssimativamente una distribuzione normale standard (più precisamente: converge in legge a una variabile aleatoria con distribuzione normale standard per $n \to +\infty$). Quindi

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \simeq Z \sim \mathcal{N}(0,1)\,.$$

Possiamo allora utilizzare i quantili gaussiani per determinare δ. Chiamiamo z_α il punto della retta reale tale che

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

Osserviamo che per la simmetria della gaussiana

$$\mathbb{P}(\{Z < -z_\alpha\}) = \mathbb{P}(\{Z \geq z_\alpha\}) = \alpha$$

Allora scegliendo $\frac{\delta}{S_n/\sqrt{n}} = z_\alpha$ abbiamo che

$$\mathbb{P}\Big(\Big\{\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} < -\frac{\delta}{S_n/\sqrt{n}}\Big\}\Big) \simeq \mathbb{P}(\{Z < -z_\alpha\}) = \alpha\,,$$

che è la condizione richiesta. Al posto di S_n, utilizzeremo la sua realizzazione sul campione dati $s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^n x_i^2 - n\bar{x}_n}$.

Abbiamo a disposizione tutti gli ingredienti e non ci resta che calcolare le quantità sui dati. Partiamo dal quantile gaussiano, per il quale utilizziamo la tabella della distribuzione normale standard:

$$\alpha = \mathbb{P}(\{Z \geq z_\alpha\}) \implies 1 - \alpha = \mathbb{P}(\{Z < z_\alpha\}) = \Phi(z_\alpha) \implies 0.95 = \Phi(z_{0.05}) \implies z_{0.05} \simeq 1.645\,.$$

Calcoliamo le realizzazioni della media campionaria e della deviazione standard campionaria sul campione di dati:

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{32}(21 + 26 + \cdots + 18 + 27) = 23.25\,.$$

$$\begin{aligned}s_n &= \sqrt{\frac{1}{n-1}\sum_{i=1}^n x_i^2 - n\bar{x}_n} = \sqrt{\frac{1}{31}(21^2 + 26^2 + \cdots + 18^2 + 27^2 - 32 \cdot 23.25^2)} \\ &= \sqrt{\frac{1}{31}(17712 - 17298)} \simeq 3.65\,.\end{aligned}$$

Segue che

$$\delta = s_n \frac{z_{0.05}}{\sqrt{32}} \simeq 3.65 \frac{1.645}{\sqrt{32}} \simeq 1.06\,.$$

Osserviamo che

$$\bar{x}_n = 23.25 > \mu_0 - \delta = 24 - 1.06 = 22.94\,,$$

quindi i dati non cadono nella regione critica e non sono abbastanza significativi da rifiutare l'ipotesi nulla. I dati non permettono di stabilire che l'esame è più difficile degli anni scorsi.

2. Abbiamo a che fare con 8 prove indipendenti. Consideriamo un "successo" un errore del I tipo, ovvero un rifiuto dell'ipotesi nulla assumendo che questa sia vera (ovvero che la media è rimasta 24). Abbiamo a che fare con una variabile aleatoria con legge binomiale $Y \sim \mathrm{B}(k, p)$ con $k = 8$ e $p = \alpha = 5\%$, perché la probabilità di commettere un errore del I tipo è esattamente la significatività del test. Ci viene chiesto di calcolare

$$\begin{aligned}\mathbb{P}(\{Y < 6\}) &= 1 - \mathbb{P}(\{Y \geq 6\}) = 1 - \mathbb{P}(\{Y = 6\}) - \mathbb{P}(\{Y = 7\}) - \mathbb{P}(\{Y = 8\}) \\ &= 1 - \binom{8}{6} 0.05^6 0.95^2 - \binom{8}{7} 0.05^7 0.95^1 - \binom{8}{8} 0.05^8 0.95^0 \simeq 99.99\%\,.\end{aligned}$$