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: settembre 2022 - II
Corso di studi: _____ Data: 20/09/2022

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) In un'indagine sui consumi di nuove auto a benzina è stata osservata la distribuzione dei litri consumati per 100 km. I dati sono rappresentati raggruppati in intervalli di classi nella seguente tabella:

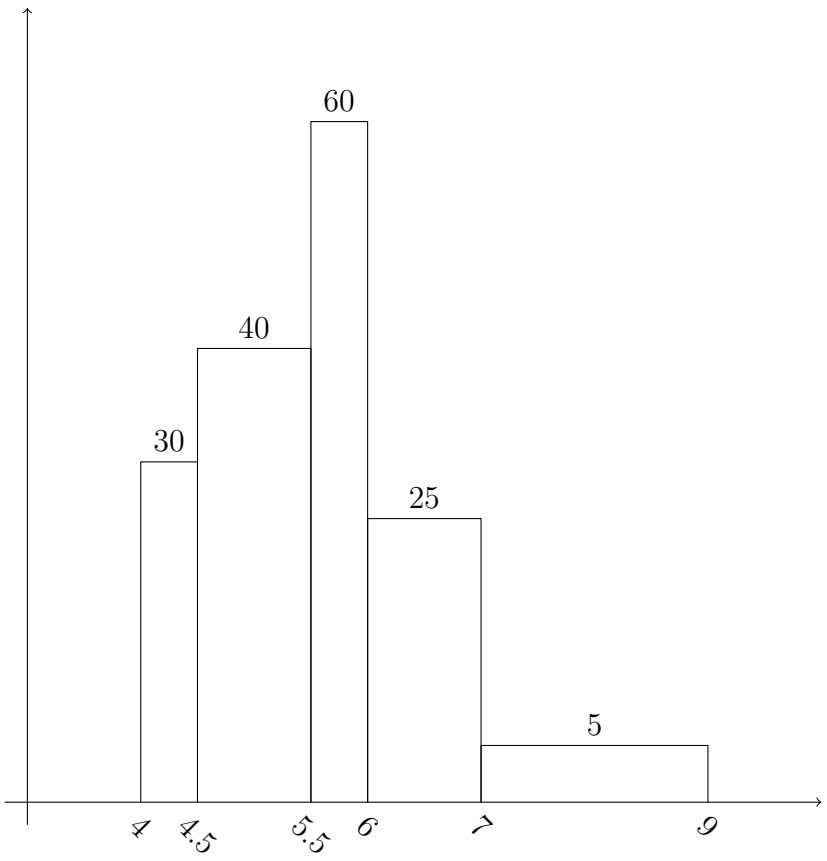
intervalli	frequenze assolute
$[4, 4.5)$	15
$[4.5, 5.5)$	40
$[5.5, 6)$	30
$[6, 7)$	25
$[7, 9)$	10

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

Soluzione. 1. Denotiamo con $I_1, \dots, I_5$ gli intervalli, $f_1, \dots, f_5$ le frequenze assolute. Abbiamo che $n = f_1 + \cdots + f_5 = 15 + 40 + 30 + 25 + 10 = 120$. Ricordiamo che le densità di frequenze assolute sono date da $d_j = f_j/|I_j|$ dove $|I_j| = b_j - a_j$ se $I_j = [a_j, b_j)$. Completiamo la tabella (scriviamo anche le frequenze relative per il punto 3. e le frequenze assolute cumulate per il punto 4.):

intervalli	f. assolute	densità f. ass.	f. relative	f. cumulate
$[4, 4.5)$	15	30	12.5%	15
$[4.5, 5.5)$	40	40	33.33%	55
$[5.5, 6)$	30	60	25%	85
$[6, 7)$	25	25	20.83%	110
$[7, 9)$	10	5	8.34%	120

Rappresentiamo l'istogramma:



2. La classe modale è l'intervallo $[5.5, 6)$ poiché è l'intervallo con la maggiore densità di frequenza relativa.
3. Ricordando che la media calcolata su un campione di dati $x_1, \ldots, x_n$ è

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

possiamo riscrivere la formula in termini dei valori assunti $v_1, \ldots, v_k$ utilizzando le frequenze assolute $f_1, \ldots, f_k$ e le frequenze relative $p_1, \ldots, p_k$

$$\overline{x} = \frac{1}{n} \sum_{j=1}^{k} f_j v_j = \sum_{j=1}^{k} p_j v_j \,.$$

Per approssimare la media sostituiamo ai valori v_j i valori centrali $\tilde{v}_j$ degli intervalli e le frequenze relative:

$$\overline{x} \simeq 12.5\% \cdot 4.25 + 33.33\% \cdot 5 + 25\% \cdot 5.75 + 20.83\% \cdot 6.5 + 8.34\% \cdot 8 = 5.6564 \,.$$

Ricordiamo che la varianza calcolata su un campione di dati $x_1, \ldots, x_n$ può essere calcolata mediante la formula:

$$s^2 = \frac{1}{n-1} \Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \Big).$$

Possiamo riscrivere la formula in termini dei valori assunti $v_1, \ldots, v_k$ utilizzando le frequenze assolute $f_1, \ldots, f_k$ e le frequenze relative $p_1, \ldots, p_k$:

$$s^2 = \frac{1}{n-1} \Big(\sum_{j=1}^{k} f_j v_j^2 - n\overline{x}^2 \Big) = \frac{n}{n-1} \Big(\sum_{j=1}^{k} p_j v_j^2 - \overline{x}^2 \Big).$$

Per approssimare la varianza sostituiamo ai valori v_j i valori centrali $\tilde{v}_j$ degli intervalli e le frequenze relative e la media approssimata calcolata nel punto precedente:

$$s^2 \simeq \frac{120}{119}\Big(12.5\% \cdot 4.25^2 + 33.33\% \cdot 5^2 + 25\% \cdot 5.75^2 + 20.83\% \cdot 6.5^2 + 8.34\% \cdot 8^2 - 5.6564^2\Big) \simeq 1.0077$$

da cui segue che la deviazione standard è approssimata da

$$s \simeq 1.0038.$$

4. Per calcolare un'approssimazione della mediana utilizziamo le frequenze cumulate $F_1, \dots, F_k$. La mediana divide l'insieme il campione di dati in due parti, quindi calcoliamo $\frac{n}{2} = \frac{120}{2} = 60$ e osserviamo che per l'intervallo $I_3 = [a_3, b_3) = [5.5, 6)$ si ha che

$$F_2 = 55 < 60 < 85 = F_3\,.$$

La mediana è allora approssimata da

$$Q_2 \simeq a_3 + \lambda_3(b_3 - a_3), \quad \lambda_3 = \frac{\frac{n}{2} - F_2}{F_3 - F_2} = \frac{60 - 55}{30} = \frac{1}{6}$$

quindi

$$Q_2 \simeq 5.5 + \frac{1}{6} \cdot 0.5 \simeq 5.58.$$

Esercizio 2. (8 punti) L'azienda per cui lavori offre ogni anno un corso di aggiornamento facoltativo. Il numero di persone che fa domanda per il corso è una variabile aleatoria distribuita con una legge di Poisson e, in media, 5 persone fanno domanda per seguire il corso. L'azienda deve decidere se offrire il corso in streaming oppure in presenza (e in tal caso deve organizzarsi per tempo per procurarsi un'aula). Se il numero di persone partecipanti è almeno 4 (compreso), il corso è offerto in presenza, altrimenti il corso è offerto in streaming online.

1. Qual è la probabilità che il corso venga offerto in presenza?

2. L'azienda viene a conoscenza del numero delle prime persone iscritte e sa che il corso verrà offerto in presenza. Deve quindi prenotare un'aula. Se l'azienda vuole che la probabilità di far sedere tutti i partecipanti sia almeno del 90%, sono sufficienti 6 posti a sedere? Se no, quanti ne servono?

3. La seguente affermazione è vera oppure falsa? "Grazie all'assenza di memoria possiamo affermare che la probabilità che il numero di partecipanti sia maggiore di 15 sapendo che il numero di partecipanti è maggiore di 10 è uguale alla probabilità che il numero di partecipanti sia maggiore di 5." (N.B.: non sono richiesti i calcoli, ma si deve motivare la risposta)

Soluzione. Consideriamo la variabile aleatoria

$$X = \text{"numero di partecipanti"} \sim \mathrm{P}(\lambda)\,.$$

Per determinare il parametro della legge di Poisson ricordiamo che $\mathbb{E}(X) = \lambda$ e la traccia spiega che $\mathbb{E}(X) = 5$, quindi $\lambda = 5$ e $X \sim \mathrm{P}(5)$, cioè

$$\mathbb{P}(\{X = k\}) = e^{-5}\frac{5^k}{k!}\,.$$

1. Ci viene chiesto di calcolare la probabilità che ci siano almeno 4 persone partecipanti

$$\begin{aligned}\mathbb{P}(\{X \geq 4\}) &= 1 - \mathbb{P}(\{X < 4\}) = 1 - \mathbb{P}(\{X = 0\}) - \mathbb{P}(\{X = 1\}) - \mathbb{P}(\{X = 2\}) - \mathbb{P}(\{X = 3\}) \\ &= 1 - e^{-5}\Big(1 + 5 + \frac{5^2}{2!} + \frac{5^3}{3!}\Big) \simeq 73.50\%.\end{aligned}$$

2. Poiché sappiamo che si è realizzato l'evento $\{X \geq 4\}$, ci viene chiesto di calcolare la probabilità condizionata

$$\mathbb{P}(\{X \leq 6\}|\{X \geq 4\}) = \frac{\mathbb{P}(\{4 \leq X \leq 6\})}{\mathbb{P}(\{X \geq 4\})} = \frac{e^{-5}\Big(\frac{5^4}{4!} + \frac{5^5}{5!} + \frac{5^6}{6!}\Big)}{73.50\%} \simeq 67.64\%.$$

Prenotare 6 posti non è sufficiente per ottenere il 90% di probabilità. Aggiungiamo un posto alla volta finché la probabilità non supera il 90%:

$$\mathbb{P}(\{X \leq 7\}|\{X \geq 4\}) = \frac{\mathbb{P}(\{4 \leq X \leq 7\})}{\mathbb{P}(\{X \geq 4\})} = \frac{e^{-5}\Big(\frac{5^4}{4!} + \frac{5^5}{5!} + \frac{5^6}{6!} + \frac{5^7}{7!}\Big)}{73.50\%} \simeq 81.85\%,$$

$$\mathbb{P}(\{X \leq 8\}|\{X \geq 4\}) = \frac{\mathbb{P}(\{4 \leq X \leq 8\})}{\mathbb{P}(\{X \geq 4\})} = \frac{e^{-5}\Big(\frac{5^4}{4!} + \frac{5^5}{5!} + \frac{5^6}{6!} + \frac{5^7}{7!} + \frac{5^8}{8!}\Big)}{73.50\%} \simeq 90.73\%,$$

quindi sono sufficienti 8 posti.
3. L'affermazione è falsa: le uniche variabili aleatorie discrete che godono dell'assenza di memoria sono quelle con distribuzione geometrica.

Esercizio 3. (7 punti) Sia (X_1, X_2) un vettore aleatorio con funzione di probabilità congiunta descritta dalla seguente tabella:

X_2 \ X_1	1	2	3
-1	$1/6$	a	b
1	b	$1/6$	a

dove $a, b \geq 0$.

1. Determinare i valori di a e b per cui $\mathrm{Cov}(X_1, X_2) = 0$.
2. Per i valori di a e b determinati nel punto 2., si ha che X_1 e X_2 sono indipendenti?
3. Si vuole osservare una successione di realizzazioni indipendenti del vettore aleatorio (X_1, X_2). Qual è la probabilità di dover attendere (strettamente) più di 10 osservazioni perché si verifichi $X_1 = 1$?

Soluzione. Ricordiamo che la somma dei valori assunti dalla funzione di probabilità congiunta deve essere 1:

$$1 = \frac{1}{6} + a + b + b + \frac{1}{6} + a \implies a + b + \frac{1}{6} = \frac{1}{2}.$$

1. Calcoliamo la covarianza, ricordando che $\mathrm{Cov}(X_1, X_2) = \mathbb{E}(X_1X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2)$. Partiamo dai valori attesi di X_1 e X_2 e notiamo che

$$\mathbb{E}(X_2) = -1 \cdot \Big(\frac{1}{6} + a + b\Big) + 1 \cdot \Big(b + \frac{1}{6} + a\Big) = 0$$

quindi $\mathrm{Cov}(X_1, X_2) = \mathbb{E}(X_1 X_2)$. Resta da calcolare

$$\mathbb{E}(X_1 X_2) = -1 \cdot \frac{1}{6} - 2 \cdot a - 3 \cdot b + 1 \cdot b + 2 \cdot \frac{1}{6} + 3 \cdot a = a - 2b + \frac{1}{6}.$$

Imponiamo che questa quantità si annulli come richiesto e mettiamo a sistema con la relazione trovata precedentemente:

$$\begin{cases} a + b + \frac{1}{6} = \frac{1}{2} \\ a - 2b + \frac{1}{6} = 0 \end{cases} \implies \begin{cases} a = \frac{1}{3} - b \\ 3b = \frac{1}{2} \end{cases} \implies \begin{cases} a = \frac{1}{6} \\ b = \frac{1}{6}. \end{cases}$$

2. Riscrivendo la tabella con i valori trovati

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

ci rendiamo conto che (X_1, X_2) è distribuito in modo uniforme. Per avere l'indipendenza di X_1 e X_2 deve valere che

$$\mathbb{P}(\{X_1 = h, X_2 = k\}) = \mathbb{P}(\{X_1 = h\})\mathbb{P}(\{X_2 = k\}) \quad \text{per ogni } h \in \{1, 2, 3\}, k \in \{-1, 1\}.$$

Da un lato abbiamo che, per qualunque h e k,

$$\mathbb{P}(\{X_1 = h, X_2 = k\}) = \frac{1}{6}\,.$$

D'altro canto, per qualunque h abbiamo che

$$\mathbb{P}(\{X_1 = h\}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

mentre per qualunque k

$$\mathbb{P}(\{X_2 = k\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}.$$

Segue che

$$\mathbb{P}(\{X_1 = h, X_2 = k\}) = \frac{1}{6} = \frac{1}{2}\frac{1}{3} = \mathbb{P}(\{X_1 = h\})\mathbb{P}(\{X_2 = k\}),$$

cioè X_1 e X_2 sono effettivamente indipendenti.
(Attenzione: per l'indipendenza è importante controllare la condizione su tutti i valori assunti, non è sufficiente controllare una sola coppia di valori, come $\mathbb{P}(\{X_1 = 1, X_2 = 1\}) = \mathbb{P}(\{X_1 = 1\})\mathbb{P}(\{X_2 = 1\})$.)
3. Consideriamo la variabile aleatoria

$$Y = \text{“prima volta che si osserva } X_1 = 1\text{”} \sim \mathrm{Geo}(p)$$

dove $p = \mathbb{P}(\{X_1 = 1\}) = \frac{1}{3}$. Ci viene chiesto di calcolare $\mathbb{P}(\{Y > 10\})$. Possiamo farlo direttamente, calcolando

$$\begin{aligned} \mathbb{P}(\{Y > 10\}) &= 1 - \mathbb{P}(\{Y \leq 10\}) \\ &= 1 - \mathbb{P}(\{Y = 1\}) - \mathbb{P}(\{Y = 2\}) - \mathbb{P}(\{Y = 3\}) - \mathbb{P}(\{Y = 4\}) - \mathbb{P}(\{Y = 5\}) \\ &\quad - \mathbb{P}(\{Y = 6\}) - \mathbb{P}(\{Y = 7\}) - \mathbb{P}(\{Y = 8\}) - \mathbb{P}(\{Y = 9\}) - \mathbb{P}(\{Y = 10\}) \\ &= 1 - p - (1-p)p - (1-p)^2 p - (1-p)^3 p - (1-p)^4 p \\ &\quad - (1-p)^5 p - (1-p)^6 p - (1-p)^7 p - (1-p)^8 p - (1-p)^9 p\,, \end{aligned}$$

e sostituendo il valore di p, ma questo è noioso e può portare a errori di conto. È meglio utilizzare il trucco della somma geometrica

$$s_{10} = \sum_{i=1}^{10}(1-p)^{i-1}p \implies (1-p)s_{10} = \sum_{i=1}^{10}(1-p)^i p = \sum_{i=2}^{11}(1-p)^{i-1}p = s_{10} + (1-p)^{10}p - p$$
$$\implies s_{10} = 1-(1-p)^{10}$$

da cui segue che

$$\mathbb{P}(\{Y > 10\}) = 1 - s_{10} = (1-p)^{10} \simeq 1.73\%.$$

Possiamo anche pensarla così: $\{Y > 10\}$ vuol dire che ci sono stati 10 insuccessi consecutivi. Questo evento ha probabilità $(1-p)^{10}$.

Esercizio 4. (7 punti) Un produttore sostiene che le sue batterie abbiano una durata di almeno 100 ore. Si sa che la deviazione standard per questo tipo di batterie è di $\sigma = 10$ ore. Un cliente, insospettito dall'affermazione del produttore, fa una prova: acquista e testa 40 campioni, osservando una media campionaria di 96.5 ore.

1. L'osservazione del cliente è significativa al 5% per destare sospetti sull'effettiva qualità delle batterie?

2. Qual è il più piccolo livello di significatività per cui i dati osservati permettono di contestare l'affermazione del produttore?

Soluzione. Stiamo considerando un campione casuale $X_1, \dots, X_n$ con $n = 40 > 30$. Non conosciamo la distribuzione delle X_i, ma è noto che la deviazione standard della popolazione è $\sigma = 10$. La media della popolazione μ è incognita e si effettua un test d'ipotesi su μ.
1. Poiché ci viene chiesto se i dati sono abbastanza significativi da rifiutare l'affermazione del produttore, il test d'ipotesi in questione è

$$H_0 : \mu \geq \mu_0 \text{ (affermazione del produttore)} \quad H_1 : \mu < \mu_0,$$

dove $\mu_0 = 100$. La regione critica è pertanto 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 campionaria calcolata sui dati, realizzazione della media campionaria (variabile aleatoria) $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, stimatore corretto della media. (Brevemente: se i dati mostrano una media campionaria esageratamente più piccola di μ_0, l'ipotesi nulla va rifiutata).
Ricordiamo che il livello di significatività $\alpha = 5\%$ del test è la probabilità di commettere un errore del I tipo. Supponiamo quindi che l'ipotesi nulla sia vera, cioè $\mu \geq \mu_0$. Non possiamo dire molto sulla statistica

$$\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}$$

poiché μ_0 non è la media della popolazione. Allora osserviamo che, essendo $\mu \geq \mu_0$, si ha che

$$\overline{X}_n \leq \mu_0 - \delta \implies \overline{X}_n \leq \mu - \delta$$

quindi

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

Per il Teorema del Limite Centrale e poiché il campione è numeroso ($n > 30$), possiamo approssimare $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ con una variabile aleatoria con legge normale standard

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

Consideriamo il quantile Gaussiano z_α definito da $\alpha = \mathbb{P}(\{Z \geq z_\alpha\})$. Scegliendo $\frac{\delta}{\sigma/\sqrt{n}} = z_\alpha$ abbiamo che

$$\mathbb{P}(\{(X_1, \ldots, X_n) \in R_C\}) \leq \mathbb{P}(\{Z < -z_\alpha\}) = \alpha,$$

cioè la probabilità di commettere un errore del I tipo è meno di α. Per calcolare la regione critica utilizziamo le tavole per calcolare il quantile Gaussiano

$$z_\alpha = z_{0.05} \simeq 1.645$$

da cui segue

$$\frac{\delta}{\sigma/\sqrt{n}} = z_\alpha \implies \delta = z_\alpha \frac{\sigma}{\sqrt{n}} = 1.645 \frac{10}{\sqrt{40}} \simeq 2.60.$$

Non resta che controllare se la realizzazione del campione è nella regione critica:

$$\overline{x}_n = 96.5$$

$$\mu_0 - \delta = 100 - 2.60 = 97.4$$

quindi $\overline{x}_n < \mu_0 - \delta$, cioè la realizzazione del campione è nella regione critica e l'ipotesi nulla va rifiutata.

2. Il più piccolo livello di significatività per cui i dati osservati permettono di rifiutare l'ipotesi nulla è il p-value del test

$$\begin{aligned} p - \text{value} &= \inf_\alpha \left\{ \overline{x}_n < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}} \right\} = \inf_\alpha \left\{ \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha \right\} = \inf_\alpha \left\{ \Phi\Big(\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} \Big) < \Phi(-z_\alpha) \right\} \\ &= \inf_\alpha \left\{ \Phi\Big(\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} \Big) < \alpha \right\} = \Phi\Big(\frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} \Big) = \Phi\Big(\frac{96.5 - 100}{10/\sqrt{40}} \Big) \\ &\simeq \Phi(-2.21) = 1 - \Phi(2.21) \simeq 1 - 0.9864 = 0.0136 = 1.36\%. \end{aligned}$$