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: luglio 2023 - turno II
Matricola: ______________________ Data: 20/07/2023

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) I seguenti dati mettono in relazione il numero di unità di un bene che sono state ordinate in funzione del prezzo del bene in sei diverse località:

numero ordini	88	112	123	136	158	172
prezzo	50	40	35	30	20	14

1. Rappresentare i dati in uno scatterplot.
2. Determinare (derivando le formule) la retta di regressione lineare e rappresentarla.
3. Determinare il coefficiente di correlazione lineare.

Soluzione. Soluzione. 1. Segue lo scatterplot (con la retta di regressione lineare):

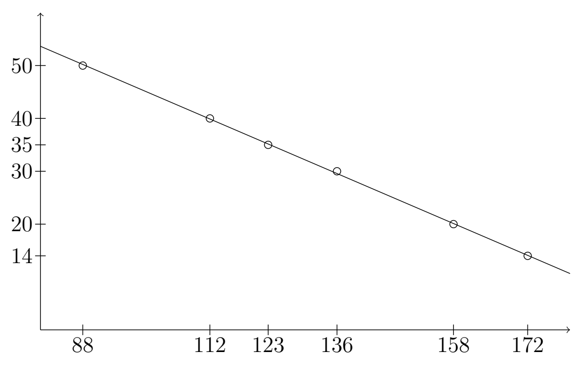

Figura 1: Scatterplot e retta di regressione lineare.

2. Denotiamo con $(x_1, y_1), \dots, (x_n, y_n)$, $n = 6$, i dati del campione. Cerchiamo la retta di equazione

$$y = ax + b$$

che meglio approssima i dati, utilizzando il metodo dei minimi quadrati. Vogliamo minimizzare l'errore

$$r(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2 .$$

Imponiamo che il gradiente rispetto ad (a,b) sia nullo, ovvero,

$$\begin{aligned} 0 = \partial_a r(a,b) &= -2\sum_{i=1}^{n}(y_i - ax_i - b)x_i = -2\sum_{i=1}^{n}(x_i y_i - ax_i^2 - bx_i)\,, \\ 0 = \partial_b r(a,b) &= -2\sum_{i=1}^{n}(y_i - ax_i - b) \end{aligned}$$

Dalla seconda equazione segue che

$$nb = \sum_{i=1}^{n}(y_i - ax_i) \implies b = \overline{y} - a\overline{x}\,.$$

Sostituendo nella prima,

$$\begin{aligned} \sum_{i=1}^{n}(x_i y_i - ax_i^2 - bx_i) = 0 &\implies \sum_{i=1}^{n}(x_i y_i - ax_i^2 - x_i\overline{y} + a\overline{x}x_i) = 0 \\ &\implies a\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\Big) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y} \\ &\implies a = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\,. \end{aligned}$$

Completiamo la tabella con i valori necessari a calcolare a e b:

							somma
x_i	88	112	123	136	158	172	789
y_i	50	40	35	30	20	14	189
x_i^2	7744	12544	15129	18496	24964	29584	108461
y_i^2	2500	1600	1225	900	400	196	6821
$x_i y_i$	4400	4480	4305	4080	3160	2408	22833

Pertanto $\overline{x} = 789/6 = 131.5$ e $\overline{y} = 189/6 = 31.5$. Segue che

$$a = \frac{22833 - 6\cdot 131.5\cdot 31.5}{108461 - 6\cdot 131.5^2} = \frac{-2020.5}{4707.5} \simeq -0.43\,,$$
$$b = 31.5 + 0.43\cdot 131.5 \simeq 88.05\,,$$

ovvero, la retta di regressione lineare ha equazione

$$y = -0.43x + 88.05\,.$$

3. Per calcolare il coefficiente di correlazione lineare usiamo la formula

$$\begin{aligned} \rho_{x,y} &= \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}} = \frac{22833 - 6\cdot 131.5\cdot 31.5}{\sqrt{108461 - 6\cdot 131.5^2}\sqrt{6821 - 6\cdot 31.5^2}} \\ &= \frac{-2020.5}{\sqrt{4707.5}\sqrt{867.5}} \simeq -0.9998\,. \end{aligned}$$

Esercizio 2. (7 punti) La prova scritta di un esame molto difficile è costituita da 4 esercizi. Uno studente si reca all'esame senza aver studiato, sperando di copiare. La probabilità di riuscire a copiare bene un singolo esercizio dipende da quanto è stanco il professore.

- Se il professore è poco stanco, la probabilità di copiare bene un singolo esercizio è del 2%.
- Se il professore è molto stanco, la probabilità di copiare bene un singolo esercizio è del 5%.
- Il professore è molto stanco con il 20% di probabilità.

Si assuma che la copia di un esercizio non influenzi la copia di un altro esercizio. Lo studente supera la prova se riesce a copiare bene almeno 2 esercizi.

1. Il professore annuncia all'inizio della prova: "Oggi sono molto stanco!". Con che probabilità lo studente supera la prova?
2. Il professore non dice nulla all'inizio della prova e non sappiamo se è poco stanco o molto stanco. Finisce l'esame e lo studente supera la prova! Con che probabilità il professore era molto stanco?

Soluzione. Consideriamo le variabili aleatorie

$$X = \text{“numero di esercizi copiati bene”}$$

e poiché dobbiamo contare il numero di successi in 4 prove indipendenti (la copia di un esercizio non influenza la copia di un altro esercizio)

$$Y = \text{“numero di esercizi copiati bene se il professore è poco stanco”} \sim \mathrm{B}(4, 2\%)$$

$$Z = \text{“numero di esercizi copiati bene se il professore è molto stanco”} \sim \mathrm{B}(4, 5\%)\,.$$

Ricordiamo che per una binomiale $\mathrm{B}(n,p)$, la probabilità di assumere il valore k è $\binom{n}{k}p^k(1-p)^{n-k}$. Inoltre, introduciamo

$$W \sim \mathrm{Be}(20\%)\,,$$

dove $W = 1$ se il professore è molto stanco e $W = 0$ se il professore è poco stanco.

Dalle ipotesi del problema,

$$\mathbb{P}(\{X = k\}|\{W = 0\}) = \mathbb{P}(\{Y = k\})\,,$$
$$\mathbb{P}(\{X = k\}|\{W = 1\}) = \mathbb{P}(\{Z = k\})\,.$$

1. Sappiamo che si è realizzato l'evento $\{W = 1\}$. Allora dobbiamo calcolare

$$\begin{aligned}&\mathbb{P}(\{X \geq 2\}|\{W = 1\}) = \mathbb{P}(\{Z \geq 2\}) = \mathbb{P}(\{Z = 2\}) + \mathbb{P}(\{Z = 3\}) + \mathbb{P}(\{Z = 4\})\\ &= \binom{4}{2}(5\%)^2(95\%)^2 + \binom{4}{3}(5\%)^3(95\%)^1 + \binom{4}{4}(5\%)^4(95\%)^0\\ &= 1.40\%\,.\end{aligned}$$

2. Sappiamo che lo studente ha superato la prova, cioè si è realizzato l'evento $\{X \geq 2\}$. Calcoliamo allora

$$\begin{aligned}\mathbb{P}(\{W = 1\}|\{X \geq 2\}) &= \frac{\mathbb{P}(\{W = 1\} \cap \{X \geq 2\})}{\mathbb{P}(\{X \geq 2\})}\\ &= \frac{\mathbb{P}(\{X \geq 2\}|\{W = 1\})\mathbb{P}(\{W = 1\})}{\mathbb{P}(\{X \geq 2\}|\{W = 1\})\mathbb{P}(\{W = 1\}) + \mathbb{P}(\{X \geq 2\}|\{W = 0\})\mathbb{P}(\{W = 0\})}\\ &= \frac{\mathbb{P}(\{X \geq 2\}|\{W = 1\})\mathbb{P}(\{W = 1\})}{\mathbb{P}(\{X \geq 2\}|\{W = 1\})\mathbb{P}(\{W = 1\}) + \mathbb{P}(\{X \geq 2\}|\{W = 0\})\mathbb{P}(\{W = 0\})}\\ &= \frac{\mathbb{P}(\{Z \geq 2\})\mathbb{P}(\{W = 1\})}{\mathbb{P}(\{Z \geq 2\})\mathbb{P}(\{W = 1\}) + \mathbb{P}(\{Y \geq 2\})\mathbb{P}(\{W = 0\})}\\ &= \frac{1.40\% \cdot 20\%}{1.40\% \cdot 20\% + 0.23\%80\%} \simeq 60.34\%\,,\end{aligned}$$

dove abbiamo usato che

$$\begin{aligned}
&\mathbb{P}(\{Y \geq 2\}) = \mathbb{P}(\{Y = 2\}) + \mathbb{P}(\{Y = 3\}) + \mathbb{P}(\{Y = 4\}) \\
&= \binom{4}{2}(2\%)^2(98\%)^2 + \binom{4}{3}(2\%)^3(98\%)^1 + \binom{4}{4}(2\%)^4(98\%)^0 \\
&= 0.23\% \,.
\end{aligned}$$

Esercizio 3. (8 punti) Devi viaggiare su un treno che percorre il seguente tragitto:

$$A \to B \to C \,.$$

Assumiamo che:

- nella tratta $A \to B$ il treno accumuli un ritardo medio di 2 min e che questo ritardo sia distribuito con legge esponenziale;
- nella tratta $B \to C$ il treno accumuli un ritardo medio di 2 min e che questo ritardo sia distribuito con legge esponenziale;
- il ritardo accumulato nella tratta $A \to B$ e il ritardo accumulato nella tratta $B \to C$ siano indipendenti.

Rispondere ai seguenti quesiti:

1. Qual è la probabilità che nella tratta $A \to B$ il treno accumuli un ritardo compreso tra 4 min e 12 min?
2. In media, qual è il ritardo accumulato nell'intero viaggio $A \to B \to C$? Motivare la risposta.
3. Qual è la varianza del ritardo accumulato nell'intero viaggio $A \to B \to C$? Motivare la risposta.
4. Considerare la probabilità che il ritardo accumulato nell'intero viaggio $A \to B \to C$ sia maggiore di 10 min. Stabilire se questa probabilità è maggiore o minore di 5%.

Soluzione. Consideriamo le due variabili aleatorie

$$X_1 = \text{“ritardo accumulato nella tratta } A \to B\text{”} \sim \mathrm{Exp}(\lambda)\,,$$

$$X_2 = \text{“ritardo accumulato nella tratta } B \to C\text{”} \sim \mathrm{Exp}(\lambda)\,.$$

In entrambi i casi si ha che $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \frac{1}{\lambda} = 2$ quindi $\lambda = \frac{1}{2}$.

Ricordiamo che la densità di una legge esponenziale con parametro λ è $\lambda e^{-\lambda x}$.

1. Viene chiesto di calcolare

$$\begin{aligned}
\mathbb{P}(\{4 \leq X_1 \leq 12\}) &= \int_4^{12} \lambda e^{-\lambda x}\, \mathrm{d}x = \Big[-e^{-\lambda x} \Big]_4^{12} \\
&= e^{-10\lambda} - e^{-30\lambda} = e^{-\frac{4}{2}} - e^{-\frac{12}{2}} = e^{-2} - e^{-6} \simeq 13.29\% \,.
\end{aligned}$$

2. Il ritardo accumulato nell'intero viaggio è $X_1 + X_2$. Per la linearità del valore atteso si ha che $\mathbb{E}(X_1 + X_2) = \mathbb{E}(X_1) + \mathbb{E}(X_2) = \frac{1}{\lambda} + \frac{1}{\lambda} = \frac{2}{\lambda} = 4$.

3. Poiché X_1 e X_2 sono indipendenti, si ha che

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = \frac{1}{\lambda^2} + \frac{1}{\lambda^2} = \frac{2}{\lambda^2} = 2 \cdot 4 = 8\,.$$

4. Dobbiamo calcolare

$$\mathbb{P}(\{X_1 + X_2 \geq 10\})\,.$$

Occorre ricordare che la somma di esponenziali indipendenti è distribuita come una legge Gamma, ovvero $X_1 + X_2 \sim \mathrm{Gamma}(2, \lambda)$, che ha densità

$$\frac{\lambda^2}{\Gamma(2)} x e^{-\lambda x} = \lambda^2 x e^{-\lambda x}\,,$$

dove abbiamo usato il fatto che

$$\Gamma(n) = (n-1)\Gamma(n-1) \implies \Gamma(2) = 1 \cdot \Gamma(1) = 1\,.$$

Quindi, integrando per parti,

$$\begin{aligned}\mathbb{P}(\{X_1 + X_2 \geq 10\}) &= \int_{10}^{+\infty} \lambda^2 x e^{-\lambda x}\,\mathrm{d}x = \Big[-\lambda x e^{-\lambda x} \Big]_{10}^{+\infty} + \int_{10}^{+\infty} \lambda e^{-\lambda x}\,\mathrm{d}x \\ &= 10\lambda e^{-10\lambda} + e^{-10\lambda} = (10\lambda + 1)e^{-10\lambda} = (10/2 + 1)e^{-5} = 6e^{-5} \simeq 4\%\,,\end{aligned}$$

quindi è minore del 5%.

In alternativa, si può ricordare che $\Gamma(\frac{n}{2}, \frac{1}{2}) = \chi^2(n)$, quindi $\Gamma(2, \frac{1}{2}) = \chi^2(4)$. Utilizzando la tavola della chi-quadro troviamo che una $Q_4 \sim \chi^2(4)$ soddisfa

$$\mathbb{P}(\{Q_4 > 9.488\}) = 5\%\,,$$

quindi la probabilità di essere maggiore di 10 è più piccola del 5%.

Esercizio 4. (7 punti) Un'auto è pubblicizzata come avente un consumo medio di benzina inferiore a 4.3 ℓ/100km (litri per 100 km) nella guida in autostrada. Si misura il consumo in un campione e si ottengono i seguenti dati (ℓ/100km):

3.4	3.3	5.5	5.0	3.2	3.9	3.8	4.1	5.3	5.3
3.7	5.7	5.5	5.6	5.7	5.1	4.9	4.5	3.1	5.6
3.1	4.8	5.5	5.1	6.4	3.5	5.6	6.4	4.4	4.6
4.8	3.8	4.2	5.4	3.0	4.2	5.7	6.0	4.3	4.4

La media calcolata sul campione risulta essere 4.7 ℓ/100km. Si sa che la deviazione standard dei consumi è 1.0 ℓ/100km.

1. I dati sono significativi al 10% per non credere alla pubblicità?

2. Assumiamo che la pubblicità dica il vero. Vengono misurati i dati di campioni della stessa ampiezza di quello precedente tante volte e ogni volta viene effettuato un test come nel punto 1. In media, dopo quanti tentativi osserveremo dei dati significativi al 10% che non ci fanno credere alla pubblicità?

Soluzione. 1. Abbiamo una popolazione $X\colon \Omega \to \mathbb{R}$ (non necessariamente normale) con $\mathbb{E}(X) = \mu$ e $\mathrm{Var}(X) = \sigma^2$ e un campione casuale $X_1, \ldots, X_n$ estratto dalla popolazione con $n = 40 \geq 30$. La varianza $\sigma^2 = 1$ è nota. Dobbiamo impostare il seguente test di ipotesi

$$H_0 : \mu \leq \mu_0 \, , \quad H_1 : \mu > \mu_0 \, ,$$

con livello di significatività $\alpha = 10\%$, dove $\mu_0 = 4.3$.

Poiché l'ipotesi alternativa è $H_1 : \mu > \mu_0$, i dati saranno significativi se la media è sufficientemente più grande 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 $\mu \leq \mu_0$.

Se $\mu \leq \mu_0$, non è detto che $\mathbb{E}(X_i) = \mu_0$. Il valore della media della popolazione è μ, quindi $\mathbb{E}(X_i) = \mu$. Tuttavia, poiché $\mu \leq \mu_0$

$$\overline{X}_n > \mu_0 + \delta \implies \overline{X}_n > \mu + \delta \, .$$

Utilizziamo questo per il calcolo della regione critica.

Utilizzando la media campionaria $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$,

$$\begin{aligned} \mathbb{P}(\{(X_1, \ldots, X_n) \in R_c\}) &= \mathbb{P}(\{\overline{X}_n > \mu_0 + \delta\}) \leq \mathbb{P}(\{\overline{X}_n > \mu + \delta\}) \\ &= \mathbb{P}\Big(\Big\{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} > \frac{\delta}{\sigma/\sqrt{n}}\Big\}\Big) \, , \end{aligned}$$

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

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

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 (approssimativamente)

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

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

In conclusione, la regione critica è

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

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

Calcoliamo $z_{0.1} \simeq 1.285$ utilizzando le tavole e quindi

$$\mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha = 4.4 + \frac{1}{\sqrt{40}} 1.285 \simeq 4.6\,.$$

Poiché

$$\overline{x}_n = 4.7 > 4.6\,,$$

i dati sono significativi al 10% per rifiutare l'ipotesi nulla.

2. Consideriamo una legge geometrica $Y \sim \mathrm{Geo}(p)$ dove p è la probabilità di un successo, ovvero la probabilità di rifiutare l'ipotesi nulla quando questa è vera. Quindi $p = \alpha = 10\%$. Per la legge geometrica si sa che

$$\mathbb{E}(Y) = \frac{1}{p} = 10\,.$$