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: aprile 2023
Matricola: _____ Data: 03/04/2023

Tempo massimo: 2 ore.

Esercizio 1. (6 punti) La tabella seguente mostra i dati sul consumo medio annuo di vino pro capite e sul numero di morti dovute a malattie cardiache in un campione casuale di 10 paesi:

consumo di vino (in litri)	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7
morti	221	167	131	191	220	297	71	172	211	300

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

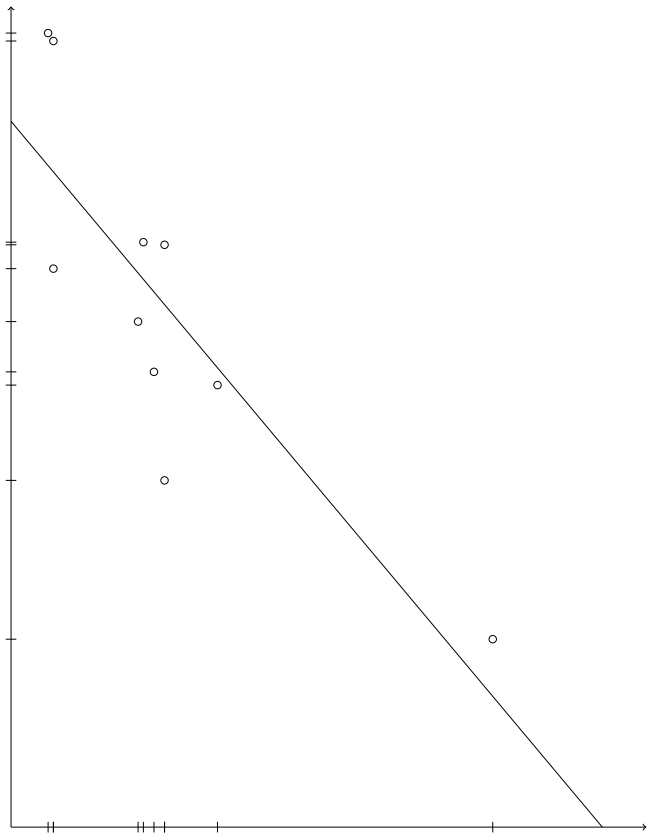

Figura 1: Scatterplot e retta di regressione lineare.

2. Denotiamo con $(x_1, y_1), \ldots, (x_n, y_n)$, $n = 10$, 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,

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

$$0 = \partial_b r(a, b) = -2 \sum_{i=1}^{n} (y_i - ax_i - b)$$

Dalla seconda equazione segue che

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

Sostituendo nella prima,

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

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

											somma
x_i	2.5	3.9	2.9	2.4	2.9	0.8	9.1	2.7	0.8	0.7	28.7
y_i	221	167	131	191	220	297	71	172	211	300	1981
x_i^2	6.25	15.21	8.41	5.76	8.41	0.64	82.81	7.29	0.64	0.49	135.91
y_i^2	48841	27889	17161	36481	48400	88209	5041	29584	44521	90000	436127
x_iy_i	552.5	651.3	379.9	458.4	638	237.6	646.1	464.4	168.8	210	4407

Pertanto $\overline{x} = 28.7/10 = 2.87$ e $\overline{y} = 1981/10 = 198.1$. Segue che

$$a = \frac{4407 - 10 \cdot 2.87 \cdot 198.1}{135.91 - 10 \cdot 2.87^2} = \frac{-1278.47}{53.541} \simeq -23.88\,,$$
$$b = 198.1 + 23.88 \cdot 2.87 \simeq 266.64\,,$$

ovvero, la retta di regressione lineare ha equazione

$$y = -23.88x + 266.64\,.$$

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

$$\rho_{x,y} = \frac{\sum_{i=1}^{n} x_iy_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}} = \frac{4407 - 10 \cdot 2.87 \cdot 198.1}{\sqrt{135.91 - 10 \cdot 2.87^2}\sqrt{436127 - 10 \cdot 198.1^2}}$$
$$= \frac{-1278.47}{\sqrt{53.541}\sqrt{43690.9}} \simeq -0.8359\,.$$

Esercizio 2. (7 punti) Un produttore di un componente elettronico sa che un componente prodotto è difettoso con una probabilità del 10% (si assumano i difetti dei componenti indipendenti tra loro).

1. Il produttore vende a un cliente una confezione con 20 componenti. Qual è la probabilità che la confezione contenga almeno 18 (18 incluso) componenti non difettose?

Il prezzo di vendita di una confezione da 20 pezzi è di 15€. Se il cliente riceve una confezione con almeno 18 componenti non difettose, non fa un reclamo. Altrimenti, il cliente fa un reclamo e chiede al produttore di inviare una nuova confezione (senza pagare nuovamente i 15€). Questa operazione si ripete finché il cliente non riceve una confezione con almeno 18 componenti sane.

2. In media, quante volte farà reclamo il cliente?

Per il produttore, il costo di produzione di una confezione da 20 pezzi è di 6€. Ogni volta che spedisce una confezione (la prima volta e per ogni eventuale reclamo), paga 2€ di costi di spedizione.

3. Qual è la probabilità che il produttore abbia una perdita per via dei ripetuti reclami dovuti ai difetti di una confezione?

Soluzione. 1. Identifichiamo come "successo" un componente non difettoso. Un successo ha probabilità 90%. Il numero di pezzi non difettosi in una confezione da 20 può essere modellato da una variabile aleatoria con legge binomiale $X \sim \mathrm{B}(n,p)$ con $n = 20$ e $p = 90\%$. Ci viene chiesto di calcolare

$$\begin{aligned}\mathbb{P}(\{X \geq 18\}) &= \mathbb{P}(\{X = 18\}) + \mathbb{P}(\{X = 19\}) + \mathbb{P}(\{X = 20\}) \\ &= \binom{20}{18}(90\%)^{18}(10\%)^2 + \binom{20}{19}(90\%)^{19}(10\%)^1 + \binom{20}{20}(90\%)^{20}(10\%)^0 \\ &= \frac{20\cdot 19}{2}(90\%)^{18}(10\%)^2 + 20(90\%)^{19}(10\%) + (90\%)^{20} = 67.69\%\,.\end{aligned}$$

2. Si sta effettuando una successione di prove in cui il "successo" è l'evento "la confezione contiene almeno 18 componenti sane", che ha probabilità 67.69%. Il primo successo (che corrisponde alla volta in cui il cliente smetterà di fare reclami) può essere modellato da una variabile aleatoria con legge geometrica $Y \sim \mathrm{Geo}(q)$ con $q = 67.69\%$. Il valore atteso di una variabile aleatoria con legge geometrica è

$$\mathbb{E}(Y) = \frac{1}{q} = \frac{1}{67.69\%} \simeq 1.48\,.$$

Poiché il numero di reclami è dato dal momento in cui osserviamo il successo meno 1, in media verranno fatti 0.48 reclami. (Se al primo tentativo arriva una confezione buona, ci saranno zero reclami, se al secondo tentativo arriva una confezione buona, ci sarà un reclamo, ecc.)

3. Consideriamo la variabile aleatoria Z che descrive il costo totale (costi di produzione e costi di spedizione). Poiché Y descritta nel punto 2. fornisce il numero di spedizioni effettuate, abbiamo che

$$Z = (6+2)\cdot Y = 8Y\,.$$

Si ottiene una perdita se il costo totale supera il ricavo, cioè $Z > 15$. Quindi dobbiamo calcolare

$$\begin{aligned}\mathbb{P}(\{Z > 15\}) = \mathbb{P}(\{8Y > 15\}) = \mathbb{P}(\{Y > 15/8\}) = \mathbb{P}(\{Y > 1.875\}) = \mathbb{P}(\{Y \geq 2\}) \\ = (1-q)^{2-1} = (1 - 67.69\%) = 32.31\%\,.\end{aligned}$$

Esercizio 3. (8 punti) Consideriamo una persona che sta svolgendo l'esame di Probabilità e Statistica. Se la persona ha studiato, il tempo (in minuti) che impiega a svolgere tutti gli esercizi del compito è distribuito con legge uniforme con media 90 min e varianza 12 min^2. Se la persona non ha studiato, il tempo (in minuti) che impiega a svolgere tutti gli esercizi del compito è distribuito con legge uniforme nell'intervallo $[90, 120]$. Il 70% delle persone che si presentano all'esame ha studiato.

1. Consideriamo una persona che sappiamo che non ha studiato. Con che probabilità impiegherà più di 100 minuti a svolgere il compito?

2. Consideriamo una persona che sappiamo che ha studiato. Con che probabilità impiegherà più di 90 minuti a svolgere il compito?

3. Consideriamo una persona che svolge l'esame (non sappiamo se ha studiato o se non ha studiato). Vediamo che ha terminato tutti gli esercizi del compito in meno di 95 minuti. Sapendo questo fatto, con che probabilità la persona ha studiato?

(I dati sono inventati.)

Soluzione. Consideriamo le seguenti variabili aleatorie:

$$X = \text{“tempo impiegato se la persona ha studiato”} \sim \mathrm{U}(a,b)$$

$$Y = \text{“tempo impiegato se la persona non ha studiato”} \sim \mathrm{U}(90,120)$$

$$T = \text{“tempo impiegato”}$$

$$S = \text{“la persona ha studiato”} \sim \mathrm{Be}(70\%)\,,$$

dove $S = 1$ (successo) quando la persona ha studiato e $S = 0$ quando la persona non ha studiato. Sappiamo che

$$\mathbb{P}(\{T \in E\}|\{S = 1\}) = \mathbb{P}(\{X \in E\})\,,$$
$$\mathbb{P}(\{T \in E\}|\{S = 0\}) = \mathbb{P}(\{Y \in E\})\,.$$

Per quanto riguarda la variabile aleatoria X utilizziamo il fatto che

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

Quindi, usando il fatto che $b - a > 0$,

$$\begin{cases} \dfrac{a+b}{2} = 90\,, \\ \dfrac{(b-a)^2}{12} = 12\,, \end{cases} \implies \begin{cases} a+b = 180\,, \\ b-a = 12\,, \end{cases} \implies \begin{cases} a = 84\,, \\ b = 96\,, \end{cases}$$

ovvero $X \sim \mathrm{U}(84,96)$.

1. Ci viene chiesto di calcolare

$$\mathbb{P}(\{Y \geq 100\}) = \int_{100}^{120} \frac{1}{120-90}\,\mathrm{d}x = \frac{120-100}{120-90} = \frac{20}{30} = \frac{2}{3}\,.$$

2. Ci viene chiesto di calcolare

$$\mathbb{P}(\{X \geq 90\}) = \int_{90}^{96} \frac{1}{96-84}\,\mathrm{d}x = \frac{96-90}{96-84} = \frac{6}{12} = \frac{1}{2}.$$

3. Ci viene chiesto di calcolare

$$\mathbb{P}(\{S = 1\}|\{T \leq 95\})\,.$$

Utilizziamo il Teorema di Bayes:

$$\begin{aligned} \mathbb{P}(\{S = 1\}|\{T \leq 95\}) &= \frac{\mathbb{P}(\{T \leq 95\}|\{S = 1\})\mathbb{P}(\{S = 1\})}{\mathbb{P}(\{T \leq 95\}|\{S = 1\})\mathbb{P}(\{S = 1\}) + \mathbb{P}(\{T \leq 95\}|\{S = 0\})\mathbb{P}(\{S = 0\})} \\ &= \frac{\mathbb{P}(\{X \leq 95\})\mathbb{P}(\{S = 1\})}{\mathbb{P}(\{X \leq 95\})\mathbb{P}(\{S = 1\}) + \mathbb{P}(\{Y \leq 95\})\mathbb{P}(\{S = 0\})} \\ &= \frac{\frac{95-84}{96-84}70\%}{\frac{95-84}{96-84}70\% + \frac{95-90}{120-90}30\%} \simeq 92.77\%\,. \end{aligned}$$

Esercizio 4. (7 punti) Uno studio statistico ha riferito che precedentemente gli/le adolescenti trascorrevano in media 3 ore al giorno con lo smartphone. Si vuole mostrare con

un'evidenza statistica che la media è diventata più alta. Ad alcuni/e adolescenti scelti casualmente è stato chiesto quante ore al giorno trascorrono con lo smartphone. I dati (in ore) sono i seguenti:

$$3.4 \quad 2.8 \quad 4.9 \quad 3.5 \quad 4.8 \quad 4.1 \quad 4.0 \quad 3.2 \quad 5.5 \quad 3.2 \quad 4.4 \quad 5.3 \quad 5.3 \quad 4.7 \quad 4.3\,.$$

(I dati sono inventati.) Si assuma che la popolazione abbia una distribuzione normale.

1. I dati sono significativi al 10% per stabilire che la media è davvero più alta?
2. In quale dei seguenti intervalli si posiziona il più piccolo livello di significatività per cui i dati portano a stabilire che la media è davvero più alta? $[0\%, 0.5\%)$, $[0.5\%, 1\%)$, $[1\%, 2.5\%)$, $[2.5\%, 5\%)$, $[5\%, 10\%)$, $[10\%, 100\%]$?

Soluzione. Si sta considerando un campione casuale $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ con $n = 15$. Sia la media μ che la varianza σ^2 della popolazione sono incognite. Fino a prova contraria, la media della popolazione è $\mu_0 = 3$. Grazie ai dati si può impostare un test di ipotesi unilaterale

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

ovvero ci si sta chiedendo se i dati sono abbastanza significativi per stabilire che la media della popolazione è, in realtà, maggiore di $\mu_0 = 3$.

Un livello di significatività α è la probabilità di commettere un errore del I tipo, ovvero di rifiutare l'ipotesi nulla quando questa è vera. Assumiamo allora che sia vera l'ipotesi nulla, ovvero che la media della popolazione sia $\mu = \mu_0 = 3$. Come regione critica per il rifiuto dell'ipotesi nulla considereremo un insieme della forma

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

dove $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ (ovvero, rifiutiamo l'ipotesi nulla se la realizzazione della media campionaria, stimatore corretto della media, sui dati del campione è sufficientemente lontana da μ_0).

Il livello di significatività α è allora, utilizzando la media campionaria $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ e la varianza campionaria $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ e il fatto che $\mu = \mu_0$,

$$\alpha = \mathbb{P}(\{\overline{X}_n > \mu_0 + \delta\}) = \mathbb{P}\Big(\Big\{\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} > \frac{\delta}{S_n/\sqrt{n}}\Big\}\Big)\,.$$

Poiché $X_1, \dots, X_n$ hanno distribuzione normale, $T_{n-1} = \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}$ è distribuita come una t-Student con $n-1$ gradi di libertà. Scegliendo $\frac{\delta}{S_n/\sqrt{n}} = t_{n-1,\alpha}$, ovvero $\delta = \frac{S_n}{\sqrt{n}} t_{n-1,\alpha}$, dove $t_{n-1,\alpha}$ è il quantile della t-Student, si ha effettivamente che

$$\mathbb{P}\Big(\Big\{\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} > \frac{\delta}{S_n/\sqrt{n}}\Big\}\Big) = \mathbb{P}(\{T_{n-1} > t_{n-1,\alpha}\}) = \alpha\,.$$

Quindi rifiutiamo l'ipotesi nulla se la realizzazione della media campionaria e della varianza campionaria sui dati verificano che

$$\overline{x}_n > \mu_0 + \frac{s_n}{\sqrt{n}} t_{n-1,\alpha}\,.$$

Svolgiamo prima il punto 2. Svolgiamo il test di ipotesi con livello di significatività $\alpha = 0.5\%$. Dalla tavola della t-Student otteniamo che

$$t_{n-1,\alpha} = t_{14,0.005} = 2.977\,.$$

Calcoliamo anche

$$\overline{x}_n = \frac{1}{15}(3.4+2.8+4.9+3.5+4.8+4.1+4.0+3.2+5.5+3.2+4.4+5.3+5.3+4.7+4.3) \simeq 4.23\,.$$

$$s_n = \sqrt{\frac{1}{14}(3.4^2 + 2.8^2 + \cdots + 4.7^2 + 4.3^2 - 15 \cdot 4.23^2)} \simeq \sqrt{0.7119} \simeq 0.8437\,.$$

Si ha che

$$\overline{x}_n = 4.23 > \mu_0 + \frac{s_n}{\sqrt{n}} t_{n-1,\alpha} \simeq 3 + \frac{0.8437}{\sqrt{15}} 2.977 \simeq 3.65\,.$$

Questo vuol dire che con significatività $\alpha = 0.5\%$ viene rifiutata l'ipotesi nulla a favore di quella alternativa (cioè si ritiene che la media sia più alta di 3). Il più piccolo livello di significatività è nell'intervallo $[0\%, 0.5\%)$ (i dati sono molto significativi), rispondendo al punto 2. A maggior ragione, con un livello di significatività più alto ($\alpha = 10\%$) l'ipotesi nulla verrà rifiutata.