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: giugno 2022
Corso di studi: ______________________	Data: 20/06/2022

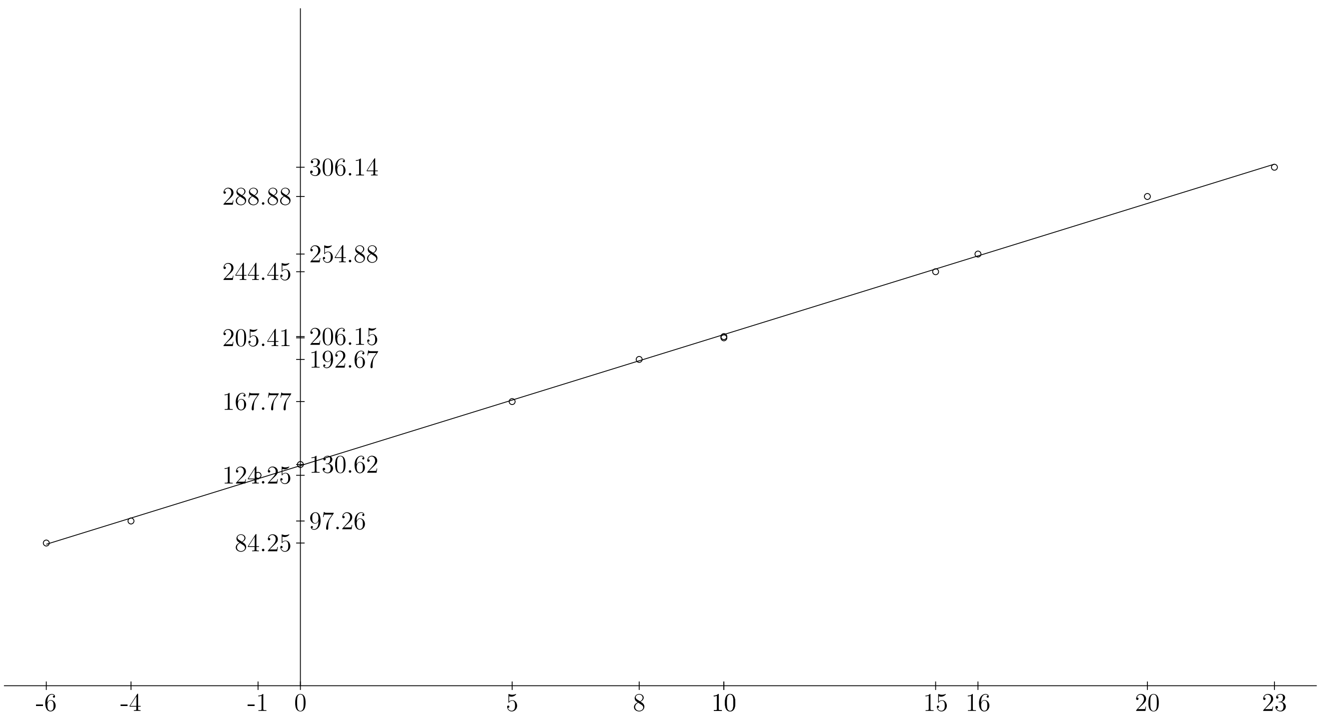
Tempo massimo: 2 ore.

Esercizio 1. Si pensa che la massa di vapore utilizzato al mese da un impianto chimico sia correlato alla temperatura ambiente media di quel mese. L'utilizzo e la temperatura dell'ultimo anno sono riportati nella tabella seguente:

mese	temperatura ($°C$)	vapore ($kg/1000$)
gen.	-6	84.25
feb.	-4	97.26
mar.	0	130.62
apr.	8	192.67
mag.	10	206.15
giu.	15	244.45
lug.	20	288.88
ago.	23	306.14
set.	16	254.88
ott.	10	205.41
nov.	5	167.77
dic.	-1	124.25

1. Rappresentare i dati in un diagramma a dispersione.
2. Calcolare e rappresentare la retta di regressione lineare.
3. Calcolare il coefficiente di correlazione.

Soluzione. 1. Segue il diagramma a dispersione.



2. Riportiamo i dati in una tabella:

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
-6	84.25	-505.50	36	7098.0625
-4	97.26	-389.04	16	9459.5076
0	130.62	0.00	0	17061.5844
8	192.67	1541.36	64	37121.7289
10	206.15	2061.50	100	42497.8225
15	244.45	3666.75	225	59755.8025
20	288.88	5777.60	400	83451.6544
23	306.14	7041.22	529	93721.6996
16	254.88	4078.08	256	64963.8144
10	205.41	2054.10	100	42193.2681
5	167.77	838.85	25	28146.7729
-1	124.25	-124.25	1	15438.0625

Utilizzando il metodo dei minimi quadrati si trovano i coefficienti α e β della retta di regressione lineare $y = \alpha x + \beta$. L'obiettivo è minimizzare:

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

Imponiamo che il gradiente rispetto ad a e b sia zero:

$$0 = \sum_{i=1}^{n} -2x_i(y_i - ax_i - b)$$
$$0 = \sum_{i=1}^{n} 2(y_i - ax_i - b)$$

da cui segue

$$\begin{aligned} a &= \frac{\sum_{i=1}^n x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^n x_i^2 - n\overline{x}^2} \\ b &= \overline{y} - a\overline{x}\,. \end{aligned}$$

Calcoliamo la media dei dati x_i e dei dati y_i:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i = \frac{1}{12}(-6-4+\cdots+5-1) = 8$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i = \frac{1}{12}(84.25+97.26+\cdots+167.77+124.25) \simeq 191.89\,.$$

Otteniamo

$$a = \frac{(-505.50-389.04+\cdots+838.85-124.25) - 12\cdot 8\cdot 191.89}{(36+16+\cdots+25+1) - 12\cdot 8^2} = \frac{26040.67-18421.44}{1752-768} \simeq 7.74$$

$$b = 129.95$$

quindi la retta di regressione lineare è

$$y = 7.74x + 129.95\,.$$

3. Per calcolare il coefficiente di correlazione possiamo usare le formule:

$$\rho = \frac{\sum_{i=1}^n x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^n x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^n y_i^2 - n\overline{y}^2}} = a\frac{\sqrt{\sum_{i=1}^n x_i^2 - n\overline{x}^2}}{\sqrt{\sum_{i=1}^n y_i^2 - n\overline{y}^2}} \simeq 7.74\frac{\sqrt{984}}{\sqrt{59029.33}} \simeq 0.9996703752\,.$$

Esercizio 2. Una cioccolateria produce due varietà di cioccolatini (fondenti oppure al latte). Vende confezioni assortite composte da 10 cioccolatini. I cioccolatini possono essere indipendentemente fondenti o al latte e, in media, ci sono 6 cioccolatini al latte in una confezione.

1. Compri una confezione di cioccolatini. Qual è la probabilità di trovare almeno 8 cioccolatini fondenti?

2. Compri e ricompri confezioni di 10 cioccolatini (ogni acquisto è indipendente dal successivo) finché non hai una confezione con un ugual numero di cioccolatini fondenti e al latte. In media, quante confezioni devi acquistare prima di avere una confezione con con un ugual numero di cioccolatini fondenti e al latte?

3. Nella stessa situazione del punto 2., qual è la probabilità di dover acquistare più di 10 confezioni per avere una confezione con con un ugual numero di cioccolatini fondenti e al latte?

Soluzione. Il numero di cioccolatini fondenti trovati in una confezione da 10 è distribuito come una variabile aleatoria binomiale $X \sim B(n,p)$ con parametri $n = 10$ e p tale che

$$np = \mathbb{E}(X) = \text{numero medio di cioccolatini fondenti} = 10 - 6 = 4\,,$$

da cui $p = 4/n = 4/10 = 2/5$. Quindi $X \sim B(10, 2/5)$.

1. Calcoliamo

$$\begin{aligned}
\mathbb{P}(\{X \geq 8\}) &= \mathbb{P}(\{X = 8\}) + \mathbb{P}(\{X = 9\}) + \mathbb{P}(\{X = 10\}) \\
&= \binom{10}{8}\left(\frac{2}{5}\right)^8\left(\frac{3}{5}\right)^2 + \binom{10}{9}\left(\frac{2}{5}\right)^9\left(\frac{3}{5}\right) + \binom{10}{10}\left(\frac{2}{5}\right)^{10} \\
&= \frac{10!}{8!2!}\left(\frac{2}{5}\right)^8\left(\frac{3}{5}\right)^2 + \frac{10!}{9!1!}\left(\frac{2}{5}\right)^9\left(\frac{3}{5}\right) + \left(\frac{2}{5}\right)^{10} \simeq 1.23\%\,.
\end{aligned}$$

2. Il numero di confezioni che si comprano fino al primo successo "è stata trovata una confezione con esattamente 5 cioccolatini fondenti" è una variabile aleatoria distribuita con una legge geometrica $Y \sim \text{Geo}(q)$ con parametro q dato dalla probabilità di successo, ovvero $q = \mathbb{P}(\{X = 5\}) = \binom{10}{5}\left(\frac{2}{5}\right)^5\left(\frac{3}{5}\right)^5 \simeq 20.07\%$. Il numero medio di confezioni da comprare è dato dal valore atteso

$$\mathbb{E}(Y) = \frac{1}{q} \simeq 4.98$$

quindi circa 5 confezioni.
3. Utilizzando il punto precedente, dobbiamo calcolare

$$\begin{aligned}
\mathbb{P}(\{Y > 10\}) = 1 - \mathbb{P}(\{Y \leq 10\}) = 1 - \sum_{k=1}^{10}\mathbb{P}(\{Y = k\}) = 1 - \sum_{k=1}^{10}(1-q)^{k-1}q \\
= 1 - 1 + (1-q)^{10} = (1-q)^{10} \simeq 10.64\%\,.
\end{aligned}$$

Ricordiamo che, posto

$$s_M = \sum_{k=1}^{M}(1-q)^{k-1}q\,,$$

si ha

$$(1-q)s_M = \sum_{k=1}^{M}(1-q)^k q = \sum_{k=2}^{M+1}(1-q)^{k-1}q = s_M + (1-q)^M q - q$$

da cui

$$s_M = 1 - (1-q)^M\,.$$

Esercizio 3. Sia (X_1, X_2) il vettore aleatorio con la seguente funzione di probabilità congiunta:

X_2 \ X_1	-1	0	1
-1	a	$2a$	b
1	b	a	$2a$

1. Calcolare la varianza di X_2.

2. Determinare a e b tali che $\text{Cov}(X_1, X_2) = 0$.

3. Per i valori a e b trovati nel punto 2., le variabili aleatorie X_1 e X_2 sono indipendenti?

Soluzione. 1. La varianza è data da

$$\text{Var}(X_2) = \mathbb{E}(X_2^2) - \mathbb{E}(X_2)^2.$$

Calcoliamo quindi

$$\mathbb{E}(X_2^2) = (-1)^2 \cdot \mathbb{P}(\{X_2 = -1\}) + 1^2 \cdot \mathbb{P}(\{X_2 = 1\}) = \mathbb{P}(\{X_2 \in \{-1, 1\}\}) = 1$$

e

$$\mathbb{E}(X_2) = -1 \cdot \mathbb{P}(\{X_2 = -1\}) + 1 \cdot \mathbb{P}(\{X_2 = 1\}) = -(3a + b) + (3a + b) = 0\,.$$

In conclusione $\mathrm{Var}(X_2) = 1$.
2. Ricordiamo che la covarianza è data da

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2)\,.$$

Il range di $X_1 X_2$ è l'insieme $\{-1, 0, 1\}$, quindi

$$\begin{aligned}
\mathbb{E}(X_1 X_2) &= -1 \cdot \mathbb{P}(\{X_1 X_2 = -1\}) + 0 \cdot \mathbb{P}(\{X_1 X_2 = 0\}) + 1 \cdot \mathbb{P}(\{X_1 X_2 = 1\}) \\
&= -\big(\mathbb{P}(\{X_1 = -1, X_2 = 1\}) + \mathbb{P}(\{X_1 = 1, X_2 = -1\})\big) \\
&\quad + \big(\mathbb{P}(\{X_1 = 1, X_2 = 1\}) + \mathbb{P}(\{X_1 = -1, X_2 = -1\})\big) \\
&= -2b + 3a\,.
\end{aligned}$$

Non è importante calcolare il valore atteso di X_1 poiché abbiamo già trovato nel punto 1. che $\mathbb{E}(X_2) = 0$.
Imponiamo

$$-2b + 3a = 0 \implies a = \frac{2}{3}b\,.$$

Sostituiamo nella relazione

$$6a + 2b = 1 \implies 4b + 2b = 1 \implies b = \frac{1}{6}\,.$$

Quindi $a = \frac{1}{9}$.

3. Calcoliamo, ad esempio,

$$\mathbb{P}(\{X_1 = 1\}) = b + 2a = \frac{1}{6} + \frac{2}{9} = \frac{7}{18}$$

$$\mathbb{P}(\{X_2 = 1\}) = b + 3a = \frac{1}{6} + \frac{3}{9} = \frac{7}{18} = \frac{1}{2}\,.$$

Poiché

$$\frac{2}{9} = \mathbb{P}(\{X_1 = 1, X_2 = 1\}) \neq \mathbb{P}(\{X_1 = 1\})\mathbb{P}(\{X_2 = 1\}) = \frac{7}{36}$$

le due variabili aleatorie non sono indipendenti.

Esercizio 4. Il contenuto di catrame (in mg) in sigarette prodotte da un'azienda si può supporre distribuito con legge normale. Dalle misurazione di 15 campioni di sigarette si ottengono i seguenti risultati:

$$6.9 \quad 7.4 \quad 7.3 \quad 6.6 \quad 7.0 \quad 6.7 \quad 7.1 \quad 6.2 \quad 7.2 \quad 6.6 \quad 6.9 \quad 6.5 \quad 7.2 \quad 7.7 \quad 7.5\,.$$

1. Determinare un intervallo di confidenza al 95% per la media del contenuto di catrame calcolata sui dati.

2. La realizzazione di un intervallo di confidenza al 97% sugli stessi dati (calcolata con lo stesso metodo del punto 1.) sarebbe più più o meno grande dell'intervallo trovato nel punto 1.? Motivare la risposta (N.B.: non è richiesto calcolare esplicitamente l'intervallo!)

Soluzione. 1. Abbiamo un campione casuale $X_1, \dots, X_n$, dove ciascuna X_i ha distribuzione normale $X_i \sim \mathcal{N}(\mu, \sigma^2)$, ma i parametri μ e σ^2 non sono noti.
Sia $\beta = 95\%$. Un intervallo di confidenza di livello β è $[U_n, V_n]$ dove gli estremi U_n e V_n sono variabili aleatorie tali che

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

Utilizzando la media campionaria e la varianza campionaria

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_i)^2$$

otteniamo che

$$\beta = \mathbb{P}(\{U_n \leq \mu \leq V_n\}) = \mathbb{P}\Big(\Big\{\frac{\overline{X}_n - V_n}{S_n/\sqrt{n}} \leq \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \leq \frac{\overline{X}_n - U_n}{S_n/\sqrt{n}}\Big\}\Big)\,.$$

La statistica

$$T_{n-1} := \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$$

ha distribuzione t-Student con $n-1$ gradi di libertà, poiché le X_i hanno distribuzione normale.
Segue che

$$\beta = \mathbb{P}\Big(\Big\{\frac{\overline{X}_n - V_n}{S_n/\sqrt{n}} \leq T_{n-1} \leq \frac{\overline{X}_n - U_n}{S_n/\sqrt{n}}\Big\}\Big) = 1 - \mathbb{P}\Big(\Big\{T_{n-1} < \frac{\overline{X}_n - V_n}{S_n/\sqrt{n}}\Big\}\Big) - \mathbb{P}\Big(\Big\{T_{n-1} > \frac{\overline{X}_n - U_n}{S_n/\sqrt{n}}\Big\}\Big)$$

da cui, ponendo $\alpha = 1 - \beta = 5\%$,

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

Decidiamo di equipartire α, ovvero

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

Queste uguaglianze si ottengono scegliendo

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

dove $t_{n-1,\alpha/2}$ è il quantile della t-Student con $n-1$ gradi di libertà tale che

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

Possiamo calcolarlo nel nostro caso utilizzando le tavole:

$$t_{n-1,\alpha/2} = t_{14,0.025} \simeq 2.145\,.$$

Quindi le variabili aleatorie

$$U_n = \overline{X}_n - \frac{S_n}{\sqrt{n}}2.145\,, \quad V_n = \overline{X}_n + \frac{S_n}{\sqrt{n}}2.145$$

costituiscono gli estremi di un intervallo di confidenza al 95%.
Calcoliamo la realizzazione di questo intervallo di confidenza sui dati. La realizzazione della media campionaria è

$$\overline{x}_{15} = \frac{1}{15}(6.9 + \cdots + 7.5) \simeq 6.9866666667$$

mentre la realizzazione della deviazione standard campionaria è

$$s_{15} = \sqrt{\frac{1}{14}(6.9^2 + \cdots + 7.5^2 - 15 \cdot \overline{x}_{15}^2)} \simeq \sqrt{\frac{1}{14}(734.6 - 732.2026666737)} \simeq 0.4138092492\,.$$

La realizzazione degli estremi è

$$u_{15} = \overline{x}_{15} - \frac{s_{15}}{\sqrt{15}}2.145 \simeq 6.7574839514 \simeq 6.8\,, \quad v_{15} = \overline{x}_{15} + \frac{s_{15}}{\sqrt{15}}2.145 \simeq 7.215849382 \simeq 7.2\,.$$

2. L'unica differenza nel calcolo della realizzazione dell'intervallo di confidenza al 97% sarebbe nel quantile della t-Student da utilizzare, che sarebbe $t_{14,0.015}$, dove $0.015 = 1.5\% = 3\%/2 = (1 - 97\%)/2$. Si ha che

$$\mathbb{P}(\{T_{14} > t_{14,0.015}\}) = 1.5\% < 2.5\% = \mathbb{P}(\{T_{14} > t_{14,0.025}\})\,.$$

Osservando il grafico della t-Student, ci rendiamo conto del fatto che che $t_{14,0.015} > t_{14,0.025}$. Quindi l'intervallo di confidenza al 97% è più grande di quello al 95%. In effetti, se la probabilità che gli estremi dell'intervallo contengano il parametro è più grande, l'intervallo deve essere più ampio.