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 - I
Corso di studi: _____ Data: 07/09/2022

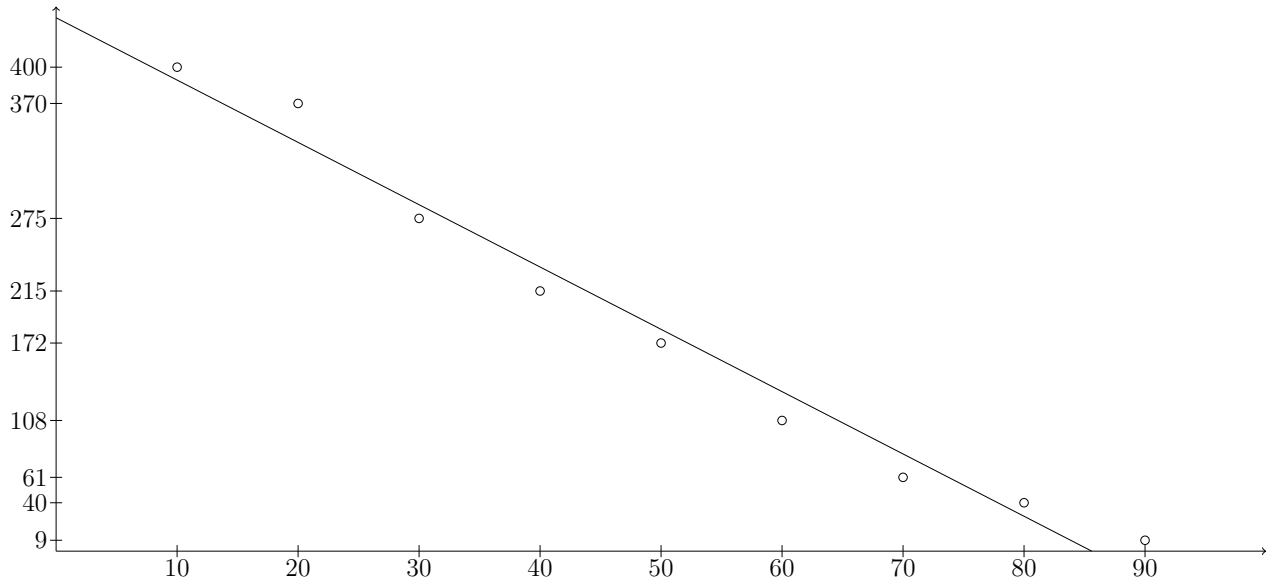
Tempo massimo: 2 ore.

Esercizio 1. (6 punti) Un'azienda produce un dispositivo elettronico da utilizzare in un intervallo di temperatura molto ampio. L'azienda sa che l'aumento della temperatura riduce il tempo di vita del dispositivo, e quindi viene eseguito uno studio in cui il tempo di vita è determinato in funzione della temperatura. Si trovano i seguenti dati:

temperatura in $°C$	tempo di vita in ore
10	400
20	370
30	275
40	215
50	172
60	108
70	61
80	40
90	9

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

Soluzione. 1. Segue lo scatterplot.



2. Riportiamo i dati in una tabella:

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
10	400	4000	100	160000
20	370	7400	400	136900
30	275	8250	900	75625
40	215	8600	1600	46225
50	172	8600	2500	29584
60	108	6480	3600	11664
70	61	4270	4900	3721
80	40	3200	6400	1600
90	9	810	8100	81

Utilizzando il metodo dei minimi quadrati si trovano i coefficienti a e b della retta di regressione lineare $y = ax + b$. 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:

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

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

Dalla seconda condizione segue

$$b = \overline{y} - a\overline{x}\,.$$

che sostituita nella prima dà

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

Calcoliamo la media dei dati x_i e dei dati y_i:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{9}(10 + 20 + \cdots + 80 + 90) = 50$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{9}(400 + 370 + \cdots + 40 + 9) \simeq 183.33\,.$$

Otteniamo

$$a = \frac{(4000 + 7400 + \cdots + 3200 + 810) - 9 \cdot 50 \cdot 183.33}{(100 + 400 + \cdots + 6400 + 8100) - 9 \cdot 50^2} = \frac{51610 - 82498.5}{28500 - 22500} = -\frac{30888.5}{6000} \simeq -5.15$$

$$b = 183.33 + 5.15 \cdot 50 = 440.83$$

quindi la retta di regressione lineare è

$$y = -5.15x + 440.83\,.$$

Per disegnarla, determiniamo due punti per cui passa, ad esempio $(0, 440.83)$ e $(85.6, 0)$.
3. Per calcolare il coefficiente di correlazione possiamo usare le formule:

$$\rho = \frac{\sum_{i=1}^n x_i y_i - n\overline{x}\,\overline{y}}{\sqrt{\sum_{i=1}^n x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^n y_i^2 - n\overline{y}^2}} = a\frac{\sqrt{\sum_{i=1}^n x_i^2 - n\overline{x}^2}}{\sqrt{\sum_{i=1}^n y_i^2 - n\overline{y}^2}} \simeq -5.15\frac{\sqrt{6000}}{\sqrt{162911}} \simeq -0.9883\,.$$

Esercizio 2. (7 punti) Una compagnia aerea ha osservato che su una certa tratta la probabilità che un passeggero che ha acquistato un biglietto non si presenti al momento dell'imbarco è del 5% (si supponga che i passeggeri siano indipendenti). L'aereo ha in tutto 96 posti, ma la compagnia prevede *overbooking* (sovrapprenotazione), quindi vende fino a 100 biglietti (supponiamo che la compagnia venda tutti i biglietti). Quindi non è detto che un posto a sedere sull'aereo sia garantito a tutti i passeggeri che hanno acquistato un biglietto e si presentano all'imbarco.

1. Qual è la probabilità che tutti i passeggeri che hanno acquistato il biglietto e si presentano all'imbarco abbiano un posto a sedere?

2. La compagnia ricava 200€ da ogni biglietto acquistato, mentre deve pagare un risarcimento di 600€ ai passeggeri che si sono presentati all'imbarco ma per cui non erano disponibili posti. Qual è il guadagno atteso per questo volo considerando risarcimenti dovuti?

Soluzione. 1. Consideriamo la seguente variabile aleatoria

$X =$ "numero di passeggeri che hanno acquistato il biglietto e si presentano all'imbarco" .

Osserviamo che X è una variabile aleatoria con legge binomiale con parametri $n = 100$ e $p = 0.95$, $X \sim \mathrm{B}(100, 0.95)$. Per convincerci di questo fatto, osserviamo che $X = X_1 + \cdots + X_{100}$ dove X_i sono le variabili aleatorie di Bernoulli indipendenti $X_i \sim \mathrm{Be}(p)$ tali che $X_i = 1$ se l'i-esimo passeggero che ha acquistato il biglietto si presenta all'imbarco e $X_i = 0$ se non si presenta all'imbarco.
In termini della variabile aleatoria X, l'evento "tutti i passeggeri che hanno acquistato il biglietto e si presentano all'imbarco hanno un posto a sedere" è $\{X \leq 96\}$ (cioè si presentano meno passeggeri dei posti disponibili). Allora calcoliamo

$$\begin{aligned}\mathbb{P}(X \leq 96) &= 1 - \mathbb{P}(X > 96) = 1 - \mathbb{P}(X = 97) - \mathbb{P}(X = 98) - \mathbb{P}(X = 99) - \mathbb{P}(X = 100)\\ &= 1 - \binom{100}{97}0.95^{97}0.05^3 - \binom{100}{98}0.95^{98}0.05^2 - \binom{100}{99}0.95^{99}0.05^1 - \binom{100}{100}0.95^{100}\\ &\simeq 74.21\%.\end{aligned}$$

2. Consideriamo la variabile aleatoria

$$Y = \text{“ricavo biglietti meno risarcimenti”} .$$

Possiamo scrivere Y in funzione della variabile aleatoria X. Se $X = x$, allora il guadagno dipende dal valore assunto x. Se $x \leq 96$, la compagnia non deve pagare alcun risarcimento, quindi ricava il massimo $200 \cdot 100$. Se invece $x \geq 97$, la compagnia deve pagare 600 a $x - 96$ passeggeri che non sono saliti sull'aereo, quindi guadagna $200 \cdot 100 - 600(x - 96)$. Quindi

$$Y = H(X) = \begin{cases} 200 \cdot 100 & \text{se } X \leq 96\,, \\ 200 \cdot 100 - 600(X - 96) & \text{se } X \geq 97\,. \end{cases}$$

Possiamo allora calcolare il valore atteso

$$\begin{aligned}
\mathbb{E}(Y) = \mathbb{E}(H(X)) &= \sum_{x=0}^{100} H(x)\mathbb{P}(\{X = x\}) = \sum_{x=0}^{96} H(x)\mathbb{P}(\{X = x\}) + \sum_{x=97}^{100} H(x)\mathbb{P}(\{X = x\}) \\
&= \sum_{x=0}^{96} 200 \cdot 100\ \mathbb{P}(\{X = x\}) + \sum_{x=97}^{100} (200 \cdot 100 - 600(x - 96))\mathbb{P}(\{X = x\}) \\
&= \sum_{x=0}^{100} 200 \cdot 100\ \mathbb{P}(\{X = x\}) - \sum_{x=97}^{100} 600(x - 96)\mathbb{P}(\{X = x\}) \\
&= 200 \cdot 100 \\
&\quad - 600\ \mathbb{P}(\{X = 97\}) - 600 \cdot 2\ \mathbb{P}(\{X = 98\}) - 600 \cdot 3\ \mathbb{P}(\{X = 99\}) - 600 \cdot 4\ \mathbb{P}(\{X = 100\}) \\
&= 20000 \\
&\quad - 600\Big(\binom{100}{97} 0.95^{97} 0.05^3 + 2\binom{100}{98} 0.95^{98} 0.05^2 + 3\binom{100}{99} 0.95^{99} 0.05^1 + 4\binom{100}{100} 0.95^{100}\Big) \\
&\simeq 20000 - 600 \cdot 41.91\% = 19748.54\,.
\end{aligned}$$

Esercizio 3. (7 punti) Sia X una variabile aleatoria assolutamente continua con la seguente densità

$$f(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} \quad x \in \mathbb{R}\,,$$

dove $\mu \in \mathbb{R}$ e $b > 0$ sono parametri da determinare.

1. Controllare che effettivamente $\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$. (Suggerimenti: effettuare un cambio di variabile per traslazione, spezzare l'integrale in due e riscalare la variabile).

2. Determinare μ e b tali che $\mathbb{E}(X) = 0$ e $\mathrm{Var}(X) = 1$. (Suggerimenti: per $\mathbb{E}(X)$, effettuare un cambio di variabile per traslazione e utilizzare il punto 1.; per $\mathrm{Var}(X)$, utilizzare il valore di μ trovato, spezzare l'integrale in due, riscalare la variabile, integrare per parti)

3. Per i valori trovati, calcolare la probabilità che $X \leq 1$ sapendo che si è verificato l'evento $X \geq 0$.

Soluzione. 1. Calcoliamo l'integrale utilizzando i suggerimenti

$$\begin{aligned}
\int_{\mathbb{R}} f(x)\,\mathrm{d}x = \int_{\mathbb{R}} \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}\,\mathrm{d}x &\overset{y=x-\mu}{=} \int_{\mathbb{R}} \frac{1}{2b} e^{-\frac{|y|}{b}}\,\mathrm{d}y = \int_0^{+\infty} \frac{1}{2b} e^{-\frac{y}{b}}\,\mathrm{d}y + \int_{-\infty}^0 \frac{1}{2b} e^{\frac{y}{b}}\,\mathrm{d}y \overset{z=-y}{=} \\
&= \int_0^{+\infty} \frac{1}{2b} e^{-\frac{y}{b}}\,\mathrm{d}y - \int_{+\infty}^0 \frac{1}{2b} e^{-\frac{z}{b}}\,\mathrm{d}z = 2\int_0^{+\infty} \frac{1}{2b} e^{-\frac{y}{b}}\,\mathrm{d}y = \int_0^{+\infty} \frac{1}{b} e^{-\frac{y}{b}}\,\mathrm{d}y \overset{s=y/b}{=} \\
&= \int_0^{+\infty} e^{-s}\,\mathrm{d}s = \Big[-e^{-s}\Big]_0^{+\infty} = 1\,.
\end{aligned}$$

2. Calcoliamo il valore atteso

$$0 = \mathbb{E}(X) = \int_{\mathbb{R}} x f(x)\,\mathrm{d}x = \int_{\mathbb{R}} \frac{1}{2b} x e^{-\frac{|x-\mu|}{b}}\,\mathrm{d}x \overset{y=x-\mu}{=} \int_{\mathbb{R}} \frac{1}{2b}(y+\mu)e^{-\frac{|y|}{b}}\,\mathrm{d}y$$
$$= \int_{\mathbb{R}} \frac{1}{2b} y e^{-\frac{|y|}{b}}\,\mathrm{d}y + \int_{\mathbb{R}} \frac{1}{2b}\mu e^{-\frac{|y|}{b}}\,\mathrm{d}y \overset{ye^{-\frac{|y|}{b}} \text{ è dispari}}{=} \mu \int_{\mathbb{R}} \frac{1}{2b} e^{-\frac{|y|}{b}}\,\mathrm{d}y \overset{\text{punto 1.}}{=} \mu\,,$$

quindi $\mu = 0$. Calcoliamo la varianza, usando il fatto che $\mathbb{E}(X) = 0$,

$$1 = \mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathbb{E}(X^2) = \int_{\mathbb{R}} x^2 f(x)\,\mathrm{d}x = \int_{\mathbb{R}} \frac{1}{2b} x^2 e^{-\frac{|x|}{b}}\,\mathrm{d}x$$
$$= \int_0^{+\infty} \frac{1}{2b} x^2 e^{-\frac{x}{b}}\,\mathrm{d}x + \int_{-\infty}^0 \frac{1}{2b} x^2 e^{-\frac{x}{b}}\,\mathrm{d}x \overset{\text{come in 1.}}{=}$$
$$= 2\int_0^{+\infty} \frac{1}{2b} x^2 e^{-\frac{x}{b}}\,\mathrm{d}x = \int_0^{+\infty} \frac{1}{b} x^2 e^{-\frac{x}{b}}\,\mathrm{d}x \overset{y=x/b}{=} \int_0^{+\infty} b^2 y^2 e^{-y}\,\mathrm{d}y \overset{\text{per parti}}{=}$$
$$= b^2\Big[-y^2 e^{-y}\Big]_0^{+\infty} + b^2 \int_0^{+\infty} 2y e^{-y}\,\mathrm{d}y = b^2\Big[-2y e^{-y}\Big]_0^{+\infty} + 2b^2 \int_0^{+\infty} e^{-y}\,\mathrm{d}y = 2b^2,$$

da cui segue $b = 1/\sqrt{2}$.
3. Utilizziamo la definizione di probabilità condizionata per calcolare

$$\mathbb{P}(\{X \le 1\}|\{X \ge 0\}) = \frac{\mathbb{P}(\{0 \le X \le 1\})}{\mathbb{P}(\{X \ge 0\})} = \frac{\int_0^1 f(x)\,\mathrm{d}x}{\int_0^{+\infty} f(x)\,\mathrm{d}x} = \frac{\int_0^1 \frac{1}{\sqrt{2}} e^{-\sqrt{2}x}\,\mathrm{d}x}{1/2} = 2\Big[-\frac{1}{2}e^{-\sqrt{2}x}\Big]_0^1$$
$$= 1 - e^{-\sqrt{2}}.$$

Esercizio 4. (8 punti) Si sa che la percentuale di titanio in una lega utilizzata nelle fusioni aerospaziali è distribuita con legge normale. Nelle domande seguenti per "esperimento statistico" intendiamo la misurazione della percentuale di titanio in 20 campioni selezionati casualmente.

1. Si fa un esperimento statistico e la deviazione standard calcolata sul campione risulta essere 0.37. Calcolare sui dati un intervallo di confidenza unilaterale sinistro (ovvero un limite superiore di confidenza) al 95% per la varianza. N.B.: derivare le formule.

2. È vero o falso che la varianza della popolazione appartiene all'intervallo calcolato nel punto precedente con il 95% di probabilità? Motivare la risposta.

3. Si ripetono tanti esperimenti statistici indipendenti. In media, dopo quanti esperimenti accade per la prima volta che la varianza della popolazione è fuori dall'intervallo di confidenza unilaterale sinistro al 95%?

Soluzione. L'esperimento statistico consiste nell'osservare un campione casuale $X_1, \ldots, X_n$ con $n = 20$, dove $X_i \sim \mathcal{N}(\mu, \sigma^2)$ e μ e σ sono incognite.
1. Un intervallo di confidenza unilaterale sinistro al 95% per la varianza della popolazione σ^2 è un intervallo della forma $(-\infty, V_n]$ dove V_n è una variabile aleatoria tale che

$$95\% = \mathbb{P}(\{\sigma^2 \le V_n\})\,.$$

Utilizzeremo la varianza campionaria

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

Ricordiamo che, poiché la popolazione è distribuita con legge normale, si ha che $\Xi_{n-1} = (n-1)S_n^2/\sigma^2$ è distribuita come una chi-quadro con $(n-1) = 19$ gradi di libertà. Allora

$$\begin{aligned} 0.95 = \mathbb{P}(\{\sigma^2 \leq V_n\}) &= \mathbb{P}\Big(\Big\{\frac{1}{V_n} \leq \frac{1}{\sigma^2}\Big\}\Big) = \mathbb{P}\Big(\Big\{\frac{(n-1)S_n^2}{V_n} \leq \frac{(n-1)S_n^2}{\sigma^2}\Big\}\Big) \\ &= \mathbb{P}\Big(\Big\{\frac{(n-1)S_n^2}{V_n} \leq \Xi_{n-1}\Big\}\Big)\,. \end{aligned}$$

Scegliamo $\frac{(n-1)S_n^2}{V_n} = \chi^2_{19,0.95}$, dove $\chi^2_{19,0.95} = 10.117$ (ottenuto dalle tavole) è il quantile della distribuzione chi-quadro tale che

$$0.95 = \mathbb{P}(\chi^2_{19,0.95} \leq \Xi_{n-1})$$

in modo che valga la condizione di intervallo di confidenza. Segue che

$$\frac{(n-1)S_n^2}{V_n} = \chi^2_{19,0.95} \implies V_n = \frac{(n-1)S_n^2}{\chi^2_{19,0.95}}\,.$$

I dati del problema forniscono una realizzazione della deviazione standard campionaria, e quindi della varianza campionaria

$$s_n^2 = 0.37^2 = 0.1369\,.$$

Utilizziamo questa per calcolare la realizzazione di V_n sui dati.

$$v_n = \frac{19 \cdot 0.1369}{10.117} = 0.2571$$

quindi $(-\infty, 0.2571]$ è un intervallo di confidenza unilaterale sinistro al 95% calcolato sui dati.
2. Se si considera l'intervallo $(-\infty, 0.2571]$ calcolato sui dati, l'affermazione è falsa! Non ha nemmeno senso calcolare la probabilità che $\sigma^2 \leq 0.2571$ perché σ^2 è un parametro e 0.2571 è un numero, non sono variabili aleatorie. Quello che si può dire è che σ^2 appartiene all'intervallo $(-\infty, V_n]$ dove $V_n = \frac{(n-1)S_n^2}{\chi^2_{19,0.95}}$ è una variabile aleatoria (non una sua realizzazione).
3. L'estremo dell'intervallo di confidenza V_n è una variabile aleatoria, e per definizione definizione di intervallo di confidenza si ha che σ^2 è fuori dall'intervallo $(-\infty, V_n]$ con probabilità $1 - 0.95 = 0.05$. Consideriamo la variabile aleatoria

$$Y = \text{“prima volta in cui } \sigma^2 > V_n\text{”}.$$

Poiché questa variabile aleatoria rappresenta il primo successo in una successione di prove indipendenti, ha una distribuzione geometrica. Il parametro della distribuzione è la probabilità di successo, quindi è 0.05. Il valore atteso di una variabile aleatoria con distribuzione geometrica è il reciproco del parametro, quindi

$$\mathbb{E}(Y) = \frac{1}{0.05} = 20\,.$$

In media, alla ventesima prova accadrà che σ^2 è fuori dall'intervallo di confidenza.