Compito A1- Soluzioni

Esercizio 1 (4 punti)

In una indagine statistica condotta presso 10 ristoranti si sono raccolti i dati riportati in tabella, dove il significato delle variabili è il seguente

Spesa: spesa a persona media (escl. bevande)
Coperti: capienza del ristorante (numero di posti)
Stelle: valutazione del critico di un noto quotidiano

	Tipo	Spesa	Coperti	Stelle	$Spesa^2$	$Coperti^2$	Spesa * Coperti
1	Etnico - Africano	35	58	3	1225	3364	2030
2	Tradizionale	25	65	3	625	4225	1625
3	Pizzeria	22	32	4	484	1024	704
4	Tradizionale	26	24	2	676	576	624
5	Lusso	70	12	5	4900	144	840
6	Self Service	9	138	2	81	19044	1242
7	Pizzeria	13	80	3	169	6400	1040
8	Giapponese	34	70	4	1156	4900	2380
9	Tradizionale	40	54	3	1600	2916	2160
10	Lusso	120	22	4	14400	484	2640
		394	555	33	25316	43077	15285

a) Si sintetizzi con un'opportuna statistica di posizione il carattere *Stelle* (1 punto)
Stelle è un carattere qualitativo ordinale. La sua distribuzione di frequenze è data da:

Stelle	Frequenze relative
2	0,20
3	0,40
4	0,30
5	0,10

La mediana è pari a 3.

b) Si calcoli la retta di regressione di *Spesa* su *Coperti* (3 punti)
Posti Y=Spesa e X=Coperti

$$Y = -0.5362X + 69.1610$$

N.B. M(X)=55.5, M(Y)=39.4, Var(X)=1227.45, Cov(XY)=-658.2.

Esercizio 2 (2 punti)

Perché l'indipendenza statistica implica l'indipendenza regressiva?
Due variabili X e Y si dicono regressivamente indipendenti se le medie delle distribuzioni condizionate dell'una a valori distinti dell'altra sono tutte eguali.
Se X e Y sono statisticamente indipendenti, le distribuzioni condizionate di Y (X) dati valori distinti di X (Y) sono tutte eguali, quindi avranno tutte la medesima media.

Esercizio 3 (4 punti)

Il 35% di una classe di matematica proviene dal Liceo Classico, mentre il rimanente 65% ha frequentato il Liceo Scientifico. Sapendo che la probabilità di superare l'esame al primo tentativo per il primo tipo di studenti è del 47%, mentre per chi ha frequentato lo Scientifico tale probabilità sale al 56%, si determini

a) la probabilità che uno studente scelto al caso all'interno della classe superi l'esame (2 punti)

Sia C l'evento "lo studente proviene dal Liceo Classico" e sia $\overline{C}$ l'evento "lo studente proviene dal Liceo Scientifico". Si ha P(C)=0.35 e P($\overline{C}$)=0.65.

Sia S l'evento "lo studente supera l'esame al primo tentativo". Si ha P(S|C)=0.47 e P(S|$\overline{C}$)=0.56.

La probabilità da determinare è data da:

P(S)= P(S|C)P(C)+ P(S|$\overline{C}$) P($\overline{C}$)=0.1645+0.364=0.5285.

b) la probabilità che uno studente che ha superato l'esame provenga dal Liceo Classico (2 punti).

$$P(C \mid S) = \frac{P(S \mid C)P(C)}{P(S)} = \frac{0.1645}{0.5285} = 0.3112.$$

Esercizio 4 (2 punti)

Sia Y una variabile aleatoria bernoulliana di parametro 0.3. Sia, inoltre, X una variabile aleatoria tale che la distribuzione di $X \mid Y = i \ \ (i = 0,1)$ è bernoulliana di parametro 0.4.

a) Si scriva la distribuzione del vettore aleatorio (X,Y). (1 punto)

Poiché le distribuzioni condizionate di X sono tutte uguali tra loro, le due variabili aleatorie sono indipendenti. Pertanto, tenendo conto della condizione di indipendenza, $p_{XY}(x,y) = p_X(x)\ p_y(y)$, si ha

Y / X	0	1	Totali
0	0.42	0.18	0.6
1	0.28	0.12	0.4
Totale	0.7	0.3	

b) Si calcoli la distribuzione di $Z = X - Y$. (1 punto)

Nella prossima tabella viene riportato, per ogni cella, il valore assunto dalla variabile aleatoria Z:

Y / X	0	1
0	**0**	**-1**
1	**1**	**0**

Pertanto la distribuzione di Z è la seguente

$$Z \equiv \begin{cases} -1 & 0 & 1 \\ 0.18 & 0.54 & 0.28 \end{cases}$$

Esercizio 5 (6 punti)

Un gruppo di studenti di Astronomia ha come obiettivo la misurazione della distanza fra due corpi celesti. Gli studenti sanno anche che il telescopio amatoriale in loro dotazione ha (come naturale) un errore di misurazione, che secondo il manuale delle istruzioni può essere così quantificato: ogni osservazione indipendente effettuata con il telescopio differisce dal dato reale per un errore casuale che si distribuisce normalmente con media 0 e varianza 16. Ne segue che la singola osservazione è una variabile casuale distribuita normalmente con media pari al valore reale della quantità che si vuole misurare, e varianza pari a 16.
A questo punto gli studenti decidono di effettuare 8 osservazioni indipendenti allo scopo di stimare la distanza fra i due corpi celesti, riportando i seguenti risultati:

40084
40118
40128
40076
40092
40102
40110
40125

a) In considerazione di quanto detto sull'errore di misurazione, si stimi con un intervallo di confidenza al 95% la distanza fra i due corpi celesti. (2 punti)

Le osservazioni compiuto costituiscono un campione *i.i.d* da una variabile Normale con valore atteso μ non noto e varianza eguale a 16. La distanza coincide con il valore atteso non noto.
L'intervallo di confidenza per μ a livello 0.95 è dato da: $\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2} \right) = \left(\overline{X} - \frac{4}{\sqrt{8}} 1.96, \overline{X} + \frac{4}{\sqrt{8}} 1.96 \right)$
L'intervallo stimato corrispondente è dato da:
$\left(40104{,}38 - \frac{4}{\sqrt{8}} 1.96, 40104{,}38 + \frac{4}{\sqrt{8}} 1.96 \right) = (40101.6, 40107.15)$.

b) Determinare il numero di osservazioni necessarie affinché l'ampiezza dell'intervallo di confidenza al 95% sia minore di 0,1 (2 punti)

La lunghezza dell'intervallo di confidenza di livello 0.95 è data da $l = 2 \cdot 1.96 \cdot 4/\sqrt{n}$, da cui:
$2 \quad 1.96 \quad 4/\sqrt{n} < 0{,}1 \Leftrightarrow n > 24586.24$. Si dovrebbero compiere 24587 osservazioni.

c) Sarebbe stato possibile rispondere al punto *b)* nel caso di popolazione con varianza non nota? Motivare la risposta. (2 punti)

Nel caso di varianza non nota, la varianza dovrebbe essere stimata a partire dalle realizzazioni campionarie. La lunghezza dell'intervallo di confidenza è data da $l = 2t^{n-1}_{1-\alpha/2} \frac{s_c}{\sqrt{n}}$. Non si potrebbe pervenire ad un risultato esatto.

Esercizio 6 (4 punti)

Sia X una popolazione normale con media incognita e varianza 9. Si vuole testare l'ipotesi nulla

$H_0 : \mu = 5$

contro

$H_1 : \mu = 3$

Dato un campione bernoulliano di ampiezza 12 e la regione di rifiuto:

$$R = \left\{ (x_1,...,x_{12}) : \sum_{i=1}^{12} x_i < 40 \right\}$$

si determinino:

a) la probabilità di errore di prima specie (2 punti)

$$P(R \mid H_0) = P\left(\sum_{i=1}^{12} X_i < 40 \middle| \mu = 5 \right) = P\left(Z < \frac{40-60}{10.3923} \right) = \phi(-1.92) = 0.0274$$

b) la probabilità di errore di seconda specie (2 punti)

$$P(\overline{R} \mid H_1) = P\left(\sum_{i=1}^{12} X_i \geq 40 \middle| \mu = 3 \right) = P\left(Z \geq \frac{40-36}{10.3923} \right) = 1 - \phi(0.38) = 0.352.$$

Esercizio 7 (3 punti)

Si considerino due popolazioni X e Y che rappresentano le durate (in minuti) di due differenti tipi di lampadine e si supponga che X e Y siano normalmente distribuite con valori attesi μ_X e μ_Y incogniti e varianze non note σ_X^2 e σ_Y^2 che supponiamo essere uguali.

Si considerano due campioni estratti da X e da Y, (X_1, X_2,……, X_{15}) e (Y_1, Y_2,….., Y_{12}) che hanno dato luogo alle realizzazioni:

$$\sum_{i=1}^{15} x_i = 5625 \qquad \sum_{i=1}^{15} (x_i - \overline{x})^2 = 216.000 \qquad \sum_{i=1}^{12} y_i = 4344 \qquad \sum_{i=1}^{12} (y_i - \overline{y})^2 = 144.000$$

Si scriva la regione di rifiuto per verificare l'ipotesi nulla $H_0 : \mu_X = \mu_Y$ contro l'ipotesi alternativa $H_1 : \mu_X \neq \mu_Y$ supponendo α=0.05 e si decida se accettare o non accettare H_0 sulla base della realizzazione osservata.

Sappiamo che la regione critica per il problema in questione ha la seguente forma:

$$R \equiv \left\{ (\underline{x}, \underline{y}) : \left| \frac{\overline{x} - \overline{y}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| > t_{1-\alpha/2}^{(n_1+n_2-2)} \right\}$$

Nel nostro caso, $n_1 = 15$, $n_2 = 12$ e $\alpha = 0.05$, dunque $t_{1-\alpha/2}^{(n_1+n_2-2)} = t_{0.975}^{(25)} = 2.060$. Pertanto la regione critica è

$$R \equiv \left\{ (\underline{x}, \underline{y}) : \left| \frac{\overline{x} - \overline{y}}{s_p \sqrt{\frac{1}{15} + \frac{1}{12}}} \right| > 2.060 \right\}.$$

In corrispondenza delle realizzazioni campionarie fornite, si ha inoltre

$$\bar{x} = \frac{5625}{15} = 375 \qquad \bar{y} = \frac{4344}{12} = 362 \qquad s_p^2 = \frac{216000 + 144000}{15 + 12 - 2} = 14400\,;$$

pertanto

$$\left| \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{15} + \frac{1}{12}}} \right| = \left| \frac{375 - 362}{\sqrt{14400} \sqrt{\frac{1}{15} + \frac{1}{12}}} \right| = 0.2797$$

e dunque l'ipotesi nulla viene accettata.

Esercizio 8 (2 punti)

Nell'ambito della regressione lineare, si riporti l'espressione analitica della scomposizione della devianza e si illustri l' indice ad essa collegato.

La somma totale dei quadrati si scompone nella somma della somma dei quadrati del modello e della somma dei quadrati dell'errore.

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y})^2 .$$

Il coefficiente di determinazione del modello R^2 è dato dal rapporto tra la somma dei quadrati del modello e la somma dei quadrati totale e rappresenta la percentuale della variabilità totale spiegata dal modello. R^2 varia tra 0 e 1, quanto più è vicino a 1 tanto migliore è il modello.

Compito A2-Soluzioni

Esercizio 1 (4 punti)

In una indagine statistica condotta presso 10 ristoranti si sono raccolti i dati riportati in tabella, dove il significato delle variabili è il seguente

Spesa: spesa a persona media (escl. bevande)
Coperti: capienza del ristorante (numero di posti)
Stelle: valutazione del critico di un noto quotidiano

	Tipo	Spesa	Coperti	Stelle	Spesa^2	Coperti^2	Spesa * Coperti
1	Etnico - Africano	40	45	4	1600	2025	1800
2	Tradizionale	22	78	3	484	6084	1716
3	Pizzeria	23	29	2	529	841	667
4	Tradizionale	42	34	2	1764	1156	1428
5	Lusso	68	18	4	4624	324	1224
6	Self Service	12	160	1	144	25600	1920
7	Pizzeria	15	56	3	225	3136	840
8	Giapponese	44	76	4	1936	5776	3344
9	Tradizionale	24	62	3	576	3844	1488
10	Lusso	115	32	5	13225	1024	3680
		405	590	31	25107	49810	18107

a) Si sintetizzi con un'opportuna statistica di posizione il carattere *Stelle* (1 punto)
Stelle è un carattere qualitativo ordinale. La sua distribuzione di frequenze è data da:

Stelle	Frequenze relative
1	0,10
2	0,20
3	0,30
4	0,30
5	0,10

La mediana è pari a 3.

b) Si calcoli la retta di regressione di *Spesa* su *Coperti* (3 punti)
Posti Y=Spesa e X=Coperti

$$Y=-0.3859X+63.2681$$

N.B. M(X)=59, M(Y)=40.5, Var(X)=1500, Cov(XY)=-578.8.

Esercizio 2 (2 punti)

Perché l'indipendenza statistica implica l'indipendenza regressiva?

Due variabili X e Y si dicono regressivamente indipendenti se le medie delle distribuzioni condizionate dell'una a valori distinti dell'altra sono tutte eguali.
Se X e Y sono statisticamente indipendenti, le distribuzioni condizionate di Y (X) dati valori distinti di X (Y) sono tutte eguali, quindi avranno tutte la medesima media.

Esercizio 3 (4 punti)

Il 30% di una classe di matematica proviene dal Liceo Classico, mentre il rimanente 70% ha frequentato il Liceo Scientifico. Sapendo che la probabilità di superare l'esame al primo tentativo per il primo tipo di studenti è del 39%, mentre per chi ha frequentato lo Scientifico tale probabilità sale al 73%, si determini

a) la probabilità che uno studente scelto al caso all'interno della classe superi l'esame (2 punti)

Sia C l'evento "lo studente proviene dal Liceo Classico" e sia $\overline{C}$ l'evento "lo studente proviene dal Liceo Scientifico". Si ha P(C)=0.3 e P($\overline{C}$)=0.7.

Sia S l'evento "lo studente supera l'esame al primo tentativo". Si ha P(S|C)=0.39 e P(S|$\overline{C}$)=0.73.

La probabilità da determinare è data da:

P(S)= P(S|C)P(C)+ P(S|$\overline{C}$) P($\overline{C}$)=0.117+0.511=0.628.

b) la probabilità che uno studente che ha superato l'esame provenga dal Liceo Classico. (2 punti)

$$P(C \mid S) = \frac{P(S \mid C)P(C)}{P(S)} = \frac{0.117}{0.628} = 0.1863.$$

Esercizio 4 (2 punti)

Sia Y una variabile aleatoria bernoulliana di parametro 0.4. Sia, inoltre, X una variabile aleatoria tale che la distribuzione di $X \mid Y = i \ \ (i = 0,1)$ è bernoulliana di parametro 0.7.

a) Si scriva la distribuzione del vettore aleatorio (X,Y). (1 punto)

Poiché le distribuzioni condizionate di X sono tutte uguali tra loro, le due variabili aleatorie sono indipendenti. Pertanto, tenendo conto della condizione di indipendenza, $p_{XY}(x,y) = p_X(x)\ p_y(y)$, si ha

Y / X	0	1	Totali
0	0.18	0.12	0.3
1	0.42	0.28	0.7
Totale	0.6	0.4	

b) Si calcoli la distribuzione di $Z = X - Y$. (1 punto)

Nella prossima tabella viene riportato, per ogni cella, il valore assunto dalla variabile aleatoria Z:

Y / X	0	1
0	**0**	**-1**
1	**1**	**0**

Pertanto la distribuzione di Z è la seguente

$$Z \equiv \begin{cases} -1 & 0 & 1 \\ 0.12 & 0.46 & 0.42 \end{cases}$$

Esercizio 5 (6 punti)

Un gruppo di studenti di Astronomia ha come obiettivo la misurazione della distanza fra due corpi celesti. Gli studenti sanno anche che il telescopio amatoriale in loro dotazione ha (come naturale) un errore di misurazione, che secondo il manuale delle istruzioni può essere così quantificato: ogni osservazione indipendente effettuata con il telescopio differisce dal dato reale per un errore casuale che si distribuisce normalmente con media 0 e varianza 16. Ne segue che la singola osservazione è una variabile casuale distribuita normalmente con media pari al valore reale della quantità che si vuole misurare, e varianza pari a 16.
A questo punto gli studenti decidono di effettuare 8 osservazioni indipendenti allo scopo di stimare la distanza fra i due corpi celesti, riportando i seguenti risultati:

34008
34021
34012
34037
34049
34024
34011
34052

a) In considerazione di quanto detto sull'errore di misurazione, si stimi con un intervallo di confidenza al 90% la distanza fra i due corpi celesti. (2 punti)

Le osservazioni compiuto costituiscono un campione *i.i.d* da una variabile Normale con valore atteso μ non noto e varianza eguale a 16. La distanza coincide con il valore atteso non noto.
L'intervallo di confidenza per μ a livello 0.95 è dato da:

$$\left(\overline{X}-\frac{\sigma}{\sqrt{n}}z_{1-\alpha/2},\overline{X}+\frac{\sigma}{\sqrt{n}}z_{1-\alpha/2}\right)=\left(\overline{X}-\frac{4}{\sqrt{8}}1.645,\overline{X}+\frac{4}{\sqrt{8}}1.645\right)$$

L'intervallo stimato corrispondente è dato da:

$$\left(34026{,}75-\frac{4}{\sqrt{8}}1.645,34026{,}75+\frac{4}{\sqrt{8}}1.645\right)=(34024.42,34029.08).$$

b) Determinare il numero di osservazioni necessarie affinché l'ampiezza dell'intervallo di confidenza al 90% sia minore di 0,05 (2 punti)

La lunghezza dell'intervallo di confidenza di livello 0.90 è data da $l=2\cdot 1.645\cdot 4/\sqrt{n}$, da cui:
$2\cdot 1.645\cdot 4/\sqrt{n}<0{,}05\Leftrightarrow n>69274{,}24$. Si dovrebbero compiere 69275 osservazioni.

c) Sarebbe stato possibile rispondere al punto *b*) nel caso di popolazione con varianza non nota? Motivare la risposta. (2 punti)

Nel caso di varianza non nota, la varianza dovrebbe essere stimata a partire dalle realizzazioni campionarie. La lunghezza dell'intervallo di confidenza è data da $l=2t^{n-1}_{1-\alpha/2}\frac{s_c}{\sqrt{n}}$. Non si potrebbe pervenire ad un risultato esatto.

Esercizio 6 (4 punti)

Sia X una popolazione normale con media incognita e varianza 9. Si vuole testare l'ipotesi nulla

$H_0 : \mu = 3$

contro

$H_1 : \mu = 5$

Dato un campione bernoulliano di ampiezza 12 e la regione di rifiuto:

$$R = \left\{ (x_1,...,x_{12}) : \sum_{i=1}^{12} x_i > 40 \right\}$$

si determinino:

a) la probabilità di errore di prima specie (2 punti)

$$P(R \mid H_0) = P\left(\sum_{i=1}^{12} X_i > 40 \middle| \mu = 3 \right) = P\left(Z > \frac{40-36}{10.3923} \right) = 1 - \phi(0.38) = 0.352$$

b) la probabilità di errore di seconda specie (2 punti)

$$P(\overline{R} \mid H_1) = P\left(\sum_{i=1}^{12} X_i \le 40 \middle| \mu = 5 \right) = P\left(Z \le \frac{40-60}{10.3923} \right) = \phi(-1.92) = 0.0274.$$

Esercizio 7 (3 punti)

Si considerino due popolazioni X e Y che rappresentano le durate (in minuti) di due differenti tipi di lampadine e si supponga che X e Y siano normalmente distribuite con valori attesi μ_X e μ_Y incogniti e varianze non note σ_X^2 *e* σ_Y^2 che supponiamo essere uguali.

Si considerano due campioni estratti da X e da Y, (X_1, X_2,……, X_{15}) e (Y_1, Y_2,….., Y_{12}), che hanno dato luogo alle realizzazioni:

$$\sum_{i=1}^{15} x_i = 2812 \qquad \sum_{i=1}^{15} (x_i - \overline{x})^2 = 54000 \qquad \sum_{i=1}^{12} y_i = 2172 \qquad \sum_{i=1}^{12} (y_i - \overline{y})^2 = 36000$$

Si scriva la regione di rifiuto per verificare l'ipotesi nulla $H_0 : \mu_X = \mu_Y$ contro l'ipotesi alternativa $H_1 : \mu_X \neq \mu_Y$ supponendo α=0.1 e si decida se accettare o non accettare H_0 sulla base della realizzazione osservata.

Sappiamo che la regione critica per il problema in questione ha la seguente forma:

$$R \equiv \left\{ (\underline{x}, \underline{y}) : \left| \frac{\overline{x} - \overline{y}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| > t_{1-\alpha/2}^{(n_1+n_2-2)} \right\}$$

Nel nostro caso, $n_1 = 15$, $n_2 = 12$ e $\alpha = 0.1$, dunque $t_{1-\alpha/2}^{(n_1+n_2-2)} = t_{0.95}^{(25)} = 1.708$. Pertanto la regione critica è

$$R \equiv \left\{ (\underline{x}, \underline{y}) : \left| \frac{\overline{x} - \overline{y}}{s_p \sqrt{\frac{1}{15} + \frac{1}{12}}} \right| > 1.708 \right\}.$$

In corrispondenza delle realizzazioni campionarie fornite, si ha inoltre

$$\bar{x} = \frac{2812}{15} = 187.4667 \qquad \bar{y} = \frac{2172}{12} = 181 \qquad s_p^2 = \frac{54000 + 36000}{15 + 12 - 2} = 3600;$$

pertanto

$$\left| \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{15} + \frac{1}{12}}} \right| = \left| \frac{187.4667 - 181}{\sqrt{3600} \sqrt{\frac{1}{15} + \frac{1}{12}}} \right| = 0.2783$$

e dunque l'ipotesi nulla viene accettata.

Esercizio 8 (2 punti)

Nell'ambito della regressione lineare, si riporti l'espressione analitica della scomposizione della devianza e si illustri l' indice ad essa collegato.
La somma totale dei quadrati si scompone nella somma della somma dei quadrati del modello e della somma dei quadrati dell'errore.

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y})^2 .$$

Il coefficiente di determinazione del modello R^2 è dato dal rapporto tra la somma dei quadrati del modello e la somma dei quadrati totale e rappresenta la percentuale della variabilità totale spiegata dal modello. R^2 varia tra 0 e 1, quanto più è vicino a 1 tanto migliore è il modello.