Cognome : **Nome :** **Matricola:**

Modulo: Laboratorio di Metodi Matematici e Statistici: M–Z

Appello del 15 Settembre 2017

Esercizio 1

Secondo una recente rilevazione statistica a scala nazionale, il 30% delle donne lavoratrici dipendenti ha avuto una promozione negli ultimi 5 anni.
1) Si calcoli la probabilità che, estrendo a caso 15 donne dalla popolazione di riferimento, non più di 3 abbiano avuto una promozione negli ultimi 5 anni.
2) Con riferimento allo stesso campione del punto 15, si indichi il numero atteso di donne lavoratrici con una promozione negli ultimi 5 anni, e la relativa deviazione standard.
3) In un campione casuale di 120 donne lavoratrici dipendenti nel settore bancario, 40 hanno avuto almeno una promozione negli ultimi 5 anni. C'è abbastanza evidenza nel campione per sostenere, con un livello di significatività del 5%, che nel settore bancario negli ultimi 5 anni le donne hanno avuto maggiori probabilità di essere promosse che negli altri settori?
4) Il margine di errore nella stima della proporzione di dipendenti donne del settore bancario che negli ultimi 5 anni hanno avuto una promozione è maggiore nell'intervallo di confidenza al livello del 5% o in quello al livello del 2.5%? Si giustifichi la risposta senza ricorrere al calcolo.

Soluzione

1) Con riferimento al dato fornito dalla rilevazione nazionale, il numero di donne X con almeno una promozione in un campione casuale di 15 soggetti ha distribuzione Binomiale$(15, 0.30)$. Pertanto

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = \sum_{k=0}^{3} \binom{15}{k} 0.3^k (1 - 0.3)^{15-k} = 0.297$$

2) Per $X \sim Binom(15, 0.3)$ si ha che il numero atteso di donne con almeno una promozione negli ultimi 5 anni è $E(X) = 15 \times 0.3 = 4.5$ e $Var(X) = 15 \times 0.3 \times (1 - 0.3) = 3.15$ da cui $\sqrt{Var(X)} = 1.775$.
3) Sottoponiamo a verifica l'ipotesi nulla che nella popolazione delle donne dipendenti nel settore bancario la percentuale p di quelle che hanno avuto una promozione negli ultimi 5 anni coincida col dato nazionale: $H_0 \ : \ p = 0.3$ contro l'alternativa (unilatera) che sia $H_1 \ : \ p > p_0 = 0.3$. La stima campionaria di p è data da $\hat{p}_n = \frac{40}{120} = 0.333$. Sussitendo le condizioni per l'uso del test approssimato ($120 \times 0.3 \geq 5$ e $120 \times (1 - 0.3) \geq 5$) calcoliamo la statistica test e la confrontiamo con il valore critico $z_{0.05} = 1.645$ della densità gaussiana standard:

$$\frac{\hat{p}_n - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{0.333 - 0.30}{\sqrt{0.30 \times 0.70/120}} = 0.789$$

Siccome la statistica test non supera il valore critico, non c'è abbastanza evidenza nel campione per concludere al livello del 5% che le donne del settore bancario siano favorite rispetto alle altre donne.
4) Il margine d'errore nell'intervallo di confidenza (pari a metà della lunghezza dell'intervallo stesso) dipende in maniera direttamente proporzionale dal livello di significatività e, pertanto, all'aumentare del livello dell'intervallo (cioè, al diminuire di α) aumenta anche il margine d'errore.

Esercizio 2

Si vuole determinare la correlazione tra due fenomeni studiando le variabili quantitative continue X ed Y. Da un campione casuale di $n = 70$ dati $(x_i, \ y_i)$ si ottiene che:

$$X: \quad \bar{x}_n = 51.115, \qquad \sum_{i=1}^{70} x_i^2 = 233702$$

$$Y: \quad \bar{y}_n = 7.186, \qquad \sum_{i=1}^{70} y_i^2 = 9402.267$$

$$\sum_{i=1}^{70} x_i y_i = 40880.44$$

1) I due fenomeni sono correlati? Come e quanto?
2) Si discuta l'opportunità di adattare ai dati un modello di regressione lineare $Y_i = a + b\,x_i + \epsilon_i$ e, in ogni caso, si forniscano le stime dei parametri incogniti del modello.
3) Si completi la tabella seguente:

Variabile	Coefficiente	Dev. standard	Statistica t	$p - value$
Intercetta		1.10320	-7.318	
X		0.01909	15.635	

e si traggano le opportune conclusioni sulla significatività della regressione.
4) Quale dei grafici in Figura 1 rappresenta un grafico dei residui per un modello di regressione lineare che bene si adatti ai dati? Giustificare la risposta.

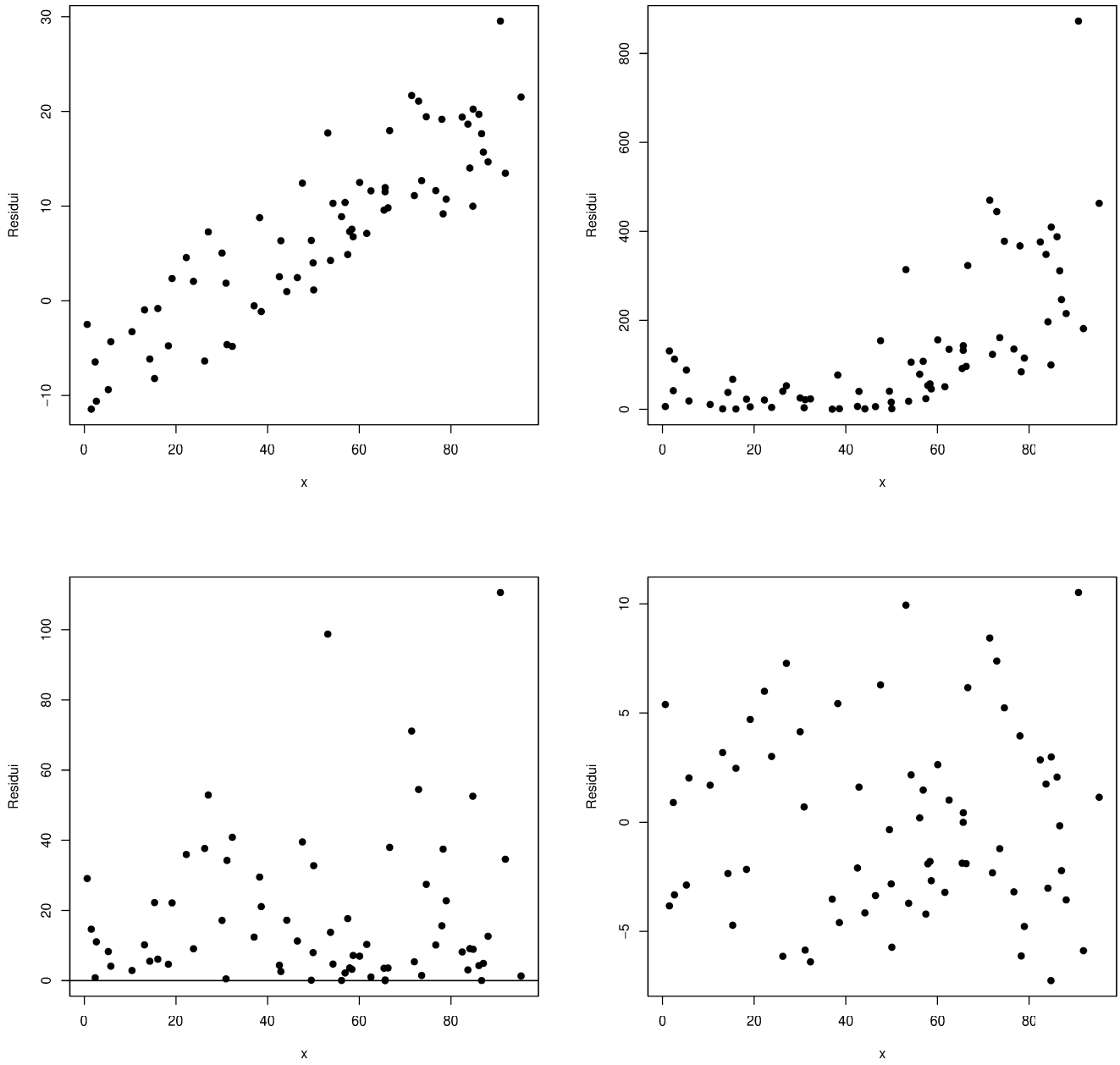

Figura 1: Esercizio 2

Soluzione
1) Cominciamo con il calcolo della covarianza:

$$\sigma_{xy} = cov(x,y) = \left(\frac{1}{70}\sum_{i=1}^{70} x_i y_i\right) - \bar{x}_n \bar{y}_n = \left(\frac{1}{70} \times 40880.44\right) - 51.115 \times 7.186 = 216.675$$

che ci dice che le variabili sono positivamente correlate. Calcoliamo le varianze σ_x e σ_y per poter poi determinare il coefficiente di correlazione lineare:

$$\sigma_x^2 = \left(\frac{1}{70}\sum_{i=1}^{70} x_i^2\right) - \bar{x}_n^2 = 725.817,\ \sigma_y^2 = 82.675$$

da cui

$$\rho_{x,y} = \frac{\sigma_{x,y}}{\sigma_x \sigma_y} = \frac{216.675}{\sqrt{725.817 \times 82.675}} = 0.884$$

che ci dice che la correlazione si può anche considerare lineare, avendo ottenuto un valore dell'indice abbastanza vicino ad 1.
2) L'indice di buon adattamento di un modello lineare ai dati, R^2 si ottiene come $\rho_{x,y}^2 = 0.781 > 0.70$ e, dunque, il modello lineare che andiamo a stimare mostrerà un buon adattamento ai dati.
La stima della pendenza è data da $\hat{b} = \sigma_{xy}/\sigma_x^2 = 216.675/725.817 = 0.298$ mentre la stima dell'intercetta è data da $\hat{a} = \bar{y}_n - \hat{b}\bar{x}_n = 7.186 - 0.298 \times 51.115 = -8.046$.
3) Per completare la tabella ci manca solo di calcolare il p–valore associato al valore assoluto delle statistiche test indicate. Le statistiche vanno confrontate con la densità gaussiana standard: siccome in valore assoluto entrambe le statistiche test superano il valore più grande indicato nella tabella usata a lezione per i valori critici della gaussiana standard (4.2649), possiamo approssimare i p–valori richiesti con 0. Pertanto, la tabella completa è data da

Variabile	Coefficiente	Dev. standard	Statistica t	$p - value$
Intercetta	**-8.046**	1.10320	-7.318	**0**
X	**0.298**	0.01909	15.635	**0**

e la regressione è statisticamente significativa, a qualunque livello di significatività, dato che il p–valore corrispondente alla variabile dipendente X è pari a 0.
4) L'unico grafico corretto è quello in basso a destra perchè è l'unico per il quale si possa dire chge la media dei residui sia zero o prossima a zero. Il grafico, inoltre, mostra valori dei residui sparsi attorno all'asse delle ascisse, più o meno ben distribuiti nei valori positivi e negativi, caratteristica tipica di un grafico dei residui per un modello lineare che ben si adatti ai dati.

Esercizio 3
Per valutare la pericolosità di una sostanza inquinante come possibile causa di infarto, sono stati rilevati in diversi paesi i valori del livello X dell'inquinante (in microgrammi) e il numero Y di infarti ogni 10000 abitanti. Si sono ottenuti i seguenti dati:

	$0 \le X < 5$	$5 \le X < 10$	$X \ge 10$
$0 \le Y < 2$	3	4	3
$2 \le Y < 5$	5	10	5
$Y \ge 5$	3	4	8

1) Si determinino le distribuzioni marginali delle variabili categorizzate (cioè, date per classi) X e Y e se ne determinino i valori medi.
2) Si costruiscano gli istogrammi delle due distribuzioni marginali per classi.
3) Sulla base del campione, c'è abbastanza evidenza per affermare che esiste una associazione tra l'inquinante e l'occorrenza dell'infarto?

Soluzione.
1) La tabella completa delle distribuzioni marginali è data da:

	$0 \le X < 5$	$5 \le X < 10$	$X \ge 10$	
$0 \le Y < 2$	3	4	3	10
$2 \le Y < 5$	5	10	5	20
$Y \ge 5$	3	4	8	15
	11	18	16	45

Si noti che sia per la variabile X sia per la variabile Y l'ultima classe è aperta ($X \ge 10$ o $Y \ge 5$) e, pertanto, si dovrà provvedere a chiuderle, per calcolare le medie, fissando un valore limite. Non avendo alcuna informazione, possiamo, per esempio, scegliere di chiudere l'ultima classe di X a 15 e l'ultima classe di Y a 10.
A questo punto possiamo usare i punti medi delle classi per calcolare le medie:

$$\bar{x}_n = 2.5 \times \frac{11}{45} + 7.5 \times \frac{18}{45} + 12.5 \times \frac{16}{45} = 8.056$$

e

$$\bar{y}_n = 1 \times \frac{10}{45} + 3.5 \times \frac{20}{45} + 7.5 \times \frac{15}{45} = 4.278$$

2) Si noti che le classi usate per la variabile y hanno ampiezze diverse e, pertanto, per il disegno dell'istogramma si dovranno usare le densità e non le frequenze. Faremo questa scelta anche per la variabile X. Inoltre, in entrambi i casi usiamo come valori limite per le classi aperte quelli stabiliti al punto 1). Usando le frequenze relative, per la variabile X si hanno le densità: $\frac{11/45}{5} = 0.0489$, $\frac{18/45}{5} = 0.08$, e $\frac{16/45}{5} = 0.071$. Per la variabile Y, invece: $\frac{10/45}{2} = 0.111$, $\frac{20/45}{3} 0.148$, e $\frac{15/45}{5} = 0.067$.
3) Si esegue un test χ^2 per l'ipotesi H_0 di indipendenza. La tabella delle frequenze attese è la seguente:

	$0 \le X < 5$	$5 \le X < 10$	$X \ge 10$	Totale
$0 \le Y < 2$	2.42	3.96	3.52	10
$2 \le Y < 5$	4.95	8.10	7.20	20
$Y \ge 5$	3.63	5.94	5.28	15
Totale	11	18	16	45

Il valore della statistica test è pertanto

$$\chi^2 = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}} = 3.4788$$

che va confrontato con i valori critici di una distribuzione $\chi^2(2 \times 2)$. Con riferimento alle tavole usate a lezione, vediamo che il valore della statistica test è inferiore al valore critico 9.4877 corrispondente al $\alpha = 0.05$ e, anzi, al valore 5.9886 corrispondente ad $\alpha = 0.20$, che significa un p–valore superiore a 0.20, quindi troppo alto per poter rifiurare l'ipotesi nulla di indipendenza tra la quantità di inquinante e l'occorrenza di infarti.