Question 1: Consider a study where data are collected on fasting blood sugar levels (`glucose`) and total cholesterol (`chol`) for two groups of individuals. Suppose that the sample correlation coefficient between `glucose` and `chol` is equal to 0.5 in each of the two groups. Let r be the sample correlation coefficient between `glucose` and `chol` when the two groups are combined. Indicate whether the following statements are true or false by filling in the corresponding circle. [5]

- r can be -0.5. TRUE ● FALSE ○
- r can be -1. TRUE ○ FALSE ●
- r can be $+1$. TRUE ○ FALSE ●
- r can be 0. TRUE ● FALSE ○
- r must be strictly less than 0.5. TRUE ○ FALSE ●

Question 2: Consider paired observations $(x_i, y_i), i = 1, 2, \ldots, n$. Let r be the sample correlation coefficient between x and y, and let the least squares regression line fit to this data be $y = \hat{a} + \hat{b}x$.

(a) Define the coefficient of determination for the line $y = \hat{a} + \hat{b}x$. [2]

Let $\hat{y}_i = \hat{a} + \hat{b}x_i$. Then, the the coefficient of determination R^2 is defined as

$$R^2 = \frac{T^2 - S^2}{T^2} = 1 - \frac{S^2}{T^2},$$

where

$$T^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

and

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

(b) Prove that the coefficient of determination equals r^2. [4]

Let $s_{xy} = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$, $s_x^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$, and $s_y^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$ (the proof is similar if we use $n-1$ as the denominator). We know that

- $r = \frac{s_{xy}}{s_x s_y} \implies s_{xy} = r s_x s_y$.
- $\hat{b} = \frac{s_{xy}}{s_x^2} = \frac{r s_x s_y}{s_x^2} = r\frac{s_y}{s_x}$.
- The regression line passes through $(\bar{x}, \bar{y})$, and so $\bar{y} = \hat{a} + \hat{b}\bar{x}$.
- $T^2 = \sum(y_i - \bar{y})^2 = n s_y^2$.

Therefore, we can write

$$\begin{aligned}
S^2 = \sum(y_i - \hat{y}_i)^2 &= \sum\left[(y_i - \bar{y}) - (\hat{y}_i - \bar{y})\right]^2 \\
&= \sum\left[(y_i - \bar{y}) - (\hat{a} + \hat{b}x_i - \hat{a} - \hat{b}\bar{x})\right]^2 \\
&= \sum\left[(y_i - \bar{y}) - \hat{b}(x_i - \bar{x})\right]^2 \\
&= \sum(y_i - \bar{y})^2 + \hat{b}^2\sum(x_i - \bar{x})^2 - 2\hat{b}\sum(x_i - \bar{x})(y_i - \bar{y}) \\
&= T^2 + \left(r^2\frac{s_y^2}{s_x^2}\right)\left(n s_x^2\right) - 2\left(r\frac{s_y}{s_x}\right)(n r s_x s_y) \\
&= T^2 + r^2T^2 - 2r^2T^2 \\
&= T^2(1 - r^2)
\end{aligned}$$

It follows that $r^2 = 1 - \frac{S^2}{T^2} = R^2$.

Question 3: Let the vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ represent n positive valued observations collected in a survey. Find the value of θ which minimizes the loss function [5]

$$\lambda(\boldsymbol{X} \mid \theta) = \sum_{i=1}^{n} \frac{(X_i - \theta)^2}{X_i}.$$

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\lambda(\boldsymbol{X} \mid \theta) = \sum_{i=1}^{n} \frac{2(X_i - \theta)}{X_i}(-1)$$

Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\lambda(\boldsymbol{X} \mid \theta) = 0 \implies \sum_{i=1}^{n}\left(\frac{2X_i}{X_i} - \frac{2\theta}{X_i}\right) = 0 \implies n - \theta\sum_{i=1}^{n}\frac{1}{X_i} = 0.$$

Thus, $\lambda(\boldsymbol{X} \mid \theta)$ is minimized at the harmonic mean

$$\theta = \frac{1}{\frac{1}{n}\sum \frac{1}{X_i}}.$$

Question 4: The following table gives frequency counts of the survival status of adult passengers on the ship Titanic, which sank in 1912.

Class	Sex	Survived	Frequency
1st	Male	No	118
2nd	Male	No	154
3rd	Male	No	387
1st	Female	No	4
2nd	Female	No	13
3rd	Female	No	89
1st	Male	Yes	57
2nd	Male	Yes	14
3rd	Male	Yes	75
1st	Female	Yes	140
2nd	Female	Yes	80
3rd	Female	Yes	76

(a) Compute the following summary statistics: [8]

The proportion of adult males in 2nd Class that survived.

$$\frac{\text{survived}}{\text{total}} = \frac{14}{14+154} = \frac{14}{168} \approx 0.083$$

The proportion of adult females that survived.

$$\frac{\text{survived}}{\text{total}} = \frac{140+80+76}{(4+13+89)+(140+80+76)} = \frac{296}{402} \approx 0.736$$

The odds of survival of a female passenger in 2nd Class.

$$\hat{p} = \frac{\text{survived}}{\text{total}} = \frac{80}{80+13} = \frac{80}{93} \implies \text{odds} = \frac{\hat{p}}{1-\hat{p}} = \frac{80}{13} \approx 6.154$$

The odds of survival of a female passenger in 3rd Class.

$$\text{odds} = \frac{\text{survived}}{\text{did not survive}} = \frac{76}{89} \approx 0.854$$

(b) Compute the odds ratio of survival of a male passenger in 2nd Class vs a male passenger in 3rd Class. Interpret the resulting statistic in terms of conditional probabilities. [4]

2nd class odds = Yes / No = 14 / 154.

3rd class odds = Yes / No = 75 / 387.

$$\text{odds ratio} = \frac{14/154}{75/387} = \frac{14 \times 387}{154 \times 75} \approx 0.469$$

This can be interpreted as an estimate of

$$\frac{P(A \mid B_1)/P(A^C \mid B_1)}{P(A \mid B_2)/P(A^C \mid B_2)},$$

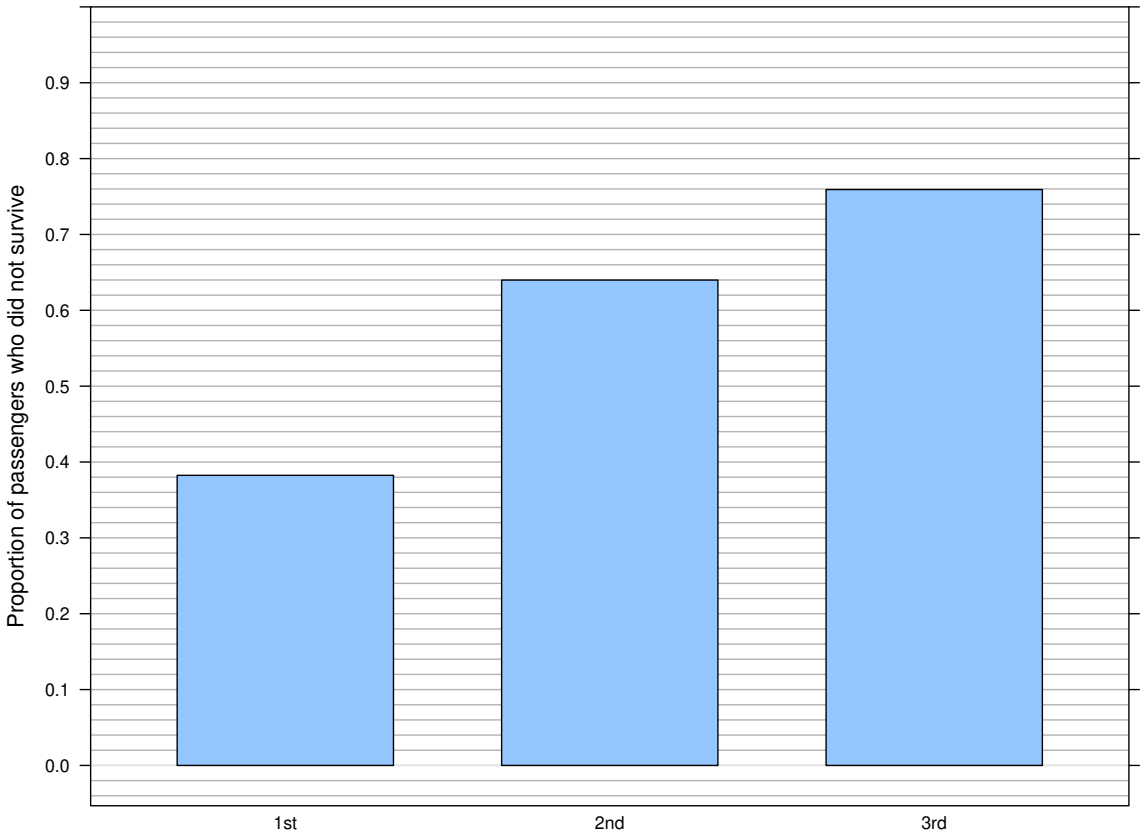
where

- A represents the event that a passenger survives
- B_1 represents the event that a passenger is a male in 2nd class
- B_2 represents the event that a passenger is a male in 3rd class

(c) Give the contingency table obtained by cross-tabulating survival status and Class. [3]

Class	Survived: No	Survived: Yes
1st	122	197
2nd	167	94
3rd	476	151

(d) Using the graphing area below, draw a barchart showing the *proportion* of adult passengers that did not survive in 1st Class, 2nd Class, and 3rd Class. [3]



Question 5: The following table gives data (x_i, y_i), for $i = 1, 2, \ldots, n$ where $n = 15$.

x	5	8	5	5	8	5	8	5	8	5	8	8	8	8	5
y	19	23	21	11	13	28	14	16	12	20	22	17	26	25	27

(a) Consider a line $y = a + bx$ which minimizes the sum of squared errors [3]

$$\sum_{i=1}^{n} \big(y_i - a - bx_i\big)^2.$$

Is this line unique? Briefly Justify your answer.

There are two distinct values of x, with the following corresponding values of y:

$x = 5$: $\boldsymbol{y}_1 = (19, 21, 11, 28, 16, 20, 27)$ with sample size $n_1 = 7$

$x = 8$: $\boldsymbol{y}_2 = (23, 13, 14, 12, 22, 17, 26, 25)$ with sample size $n_2 = 8$

The LSE line is the line that joins $(5, \bar{y}_1)$ and $(8, \bar{y}_2)$ as the total sum of squared errors (SSE) can be split into SSE_1 for $x = 5$ and SSE_2 for $x = 8$.

This line is unique because $\bar{y}_1$ and $\bar{y}_2$ are the unique minimizers of the SSE in each group.

(b) Compute one line $y = a + bx$ that minimizes the sum of squared errors. Briefly describe how you arrived at your solution. [3]

$\bar{y}_1 = 142/7 = 20.3, \bar{y}_2 = 152/8 = 19$, so the LSE line is the line joining $(5, 20.3)$ and $(8, 19)$. Expressed in the form $y = a + bx$, this line is

$$y = 22.46 - 0.43x$$

(c) Consider a line $y = a + bx$ which minimizes the sum of absolute errors [3]

$$\sum_{i=1}^{n} |y_i - a - bx_i|.$$

Is this line unique? Briefly Justify your answer.

The sum of absolute errors is similarly minimized by the line joining $(5, m_1)$ and $(8, m_2)$, where m_1 and m_2 are the medians of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ respectively. As $n_1 = 7$ is odd, $m_1 = 20$ is unique. However, $n_2 = 8$ is even, and the two middle observations are 17 and 22, so any $m_2 \in [17, 22]$ will minimize the sum of absolute errors in group 2.

So the minimizing line is not unique.

(d) Compute one line $y = a + bx$ that minimizes the sum of absolute errors. Briefly describe how you arrived at your solution. [3]

Choosing $m_2 = 20$ gives the line $y = 20$ as one possible LAD line. There are many other valid answers.

Question 6: Let $x_i = i$ and $y_i = i^2$ for $i = 1, 2, 3, 4$. Find all $\hat{\beta}$ that satisfy [4]

$$\sum_{i=1}^{4} \left|y_i - \hat{\beta} x_i\right| \leq \sum_{i=1}^{4} \left|y_i - \beta x_i\right| \text{ for all } \beta \in \mathbb{R}.$$

Briefly justify your answer.

We want to minimize

$$\sum_{i=1}^{4} |i^2 - \beta i| = \sum_{i=1}^{4} i\,|i - \beta| = 1\,|1 - \beta| + 2\,|2 - \beta| + 3\,|3 - \beta| + 4\,|4 - \beta|.$$

This is minimized when β is the median of $(1, 2, 2, 3, 3, 3, 4, 4, 4, 4)$, i.e., when $\hat{\beta} = 3$.