Exercise 3

1. The results of 10 students in a test are given below, where L indicates that a student obtained a score below 55, and the two H's (not necessarily the same) indicate scores above 90. The maximum possible score is 100.

   60 80 L 70 H 83 59 H 70 65.

   Which of the following statements must be true about the scores given above?

   Select all that apply.

   (A) The median score is 70.

   (B) There is only one mode among them.

   (C) The range is at least 35.

   (D) The mean score is greater than 72.2.

   (A) and (C) are true. After reordering the 10 scores, we have:

   $$< 55 \quad 59 \quad 60 \quad 65 \quad 70 \quad 70 \quad 80 \quad 83 \quad > 90 \quad > 90.$$

   The median is the average of the 5th and 6th score, i.e. 70. Here, 70 is one mode. However, we may have two modes, if both scores above 90 are the same. Hence, the statement on mode may not be true.

   The range is greater than $90 - 55 = 35$. Hence, the statement on range is true. Here,

   $$72.2 = \frac{55 + 59 + 60 + 65 + 70 + 70 + 80 + 83 + 90 + 90}{10},$$

   It is possible that the scores corresponding to L, H and H are for example, 50, 91, 91 respectively. Then the mean score will be 71.9. Hence the statement on mean may not be true.

2. Outliers are observations that fall well above or below the overall bulk of the data. Consider a set of 50 (univariate) data points with a single outlier. Suppose the outlier is removed from the data set, which of the following is/are always true? Select all that apply.

   (A) The removal will cause the mean to decrease.

   (B) The removal will cause the interquartile range to decrease.

   (C) The removal will cause the standard deviation to decrease.

   (D) The removal will cause the range to change.

   (C) and (D) are true. The mean will increase if the outlier falls below the bulk of the data. Interquartile range depends on the value of Q_1 and Q_3. When an outlier is removed, suppose we assume the outlier falls above the bulk of the data, the values of Q_1 and Q_3 can either remain the same or become smaller. Depending on which happens, and also the magnitude of the changes in Q_1 and Q_3, the interquartile range can increase, remain the same or decrease. The same argument applies if the outlier falls below the bulk of the data. For example,

   - the IQR **decreases** when 60 (the outlier) is removed from 4, 6, 6, 7, 11, 60.
   - the IQR **increases** when 60 (the outlier) is removed from 1, 5, 6, 7, 7, 60.
   - the IQR **remains unchanged** when 60 (the outlier) is removed from 2, 2, 2, 4, 4, 60.

   The standard deviation will decrease if the outlier is removed. The range is the difference between the maximum and minimum values, so the removal will cause a decrease in the range.

3. Suppose that there are 76 pairs of siblings living in a particular block in Ang Sua, where the older sibling is always heavier than the younger sibling. Consider a scatter plot using the younger sibling's weight to predict the older sibling's weight, where each point in the scatter plot represents the weights of a pair of two siblings in the block. Which of the following statements must be true?

(I) There is a positive association between the older and younger siblings' weights.

(II) All the points lie above the line $y = x$ in the scatter plot.

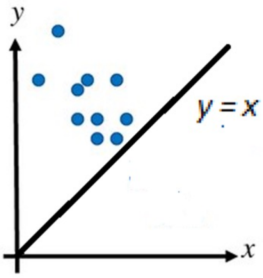
(A) Only (I).

(B) Only (II).

(C) Neither (I) nor (II).

(D) Both (I) and (II).

Answer is (B). Since each older sibling is heavier than each younger sibling, for each point (x, y), we must have $y > x$. Thus, all the points will lie above the line $y = x$ in the scatter plot. With only this condition that all the points lie above the line $y = x$, it is not possible to determine the direction of association between the two variables. For example, y can be negatively associated with x as shown in the plot below.



4. The Registry of Marriages is interested to see the relationship between the ages of husbands and wives in City X. They randomly sampled 1000 pairs of husbands and wives from the population of City X and obtained data of their ages (in years). Looking through the data, they found that men always marry women who are younger than them. Based only on the information given above, which of the following statements must be true?

(I) The average age of the husbands is more than the average age of the wives.

(II) The standard deviation of husband's age is more than the standard deviation of wife's age.

(A) Only (I).

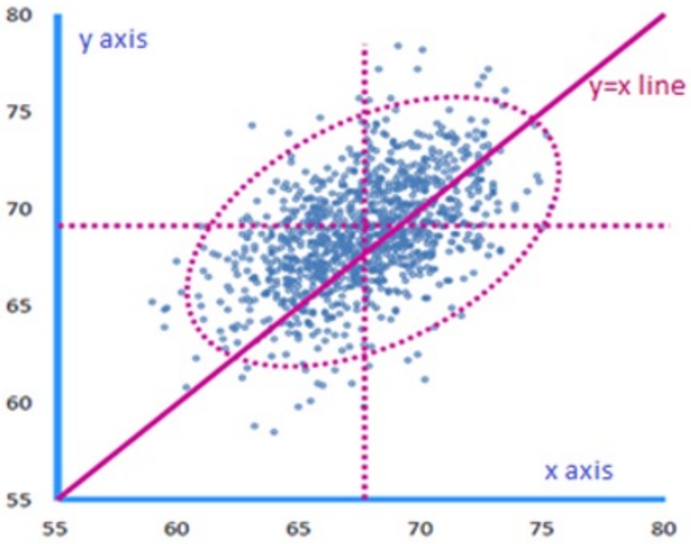
(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

Answer is (A). Since all the husbands are older than their respective wives, the average age of the husbands would be greater than the average age of the wives, and thus statement (I) is correct. On the other hand, we do not have information about the spread of the husband's age relative to the spread of the wife's age, so we cannot be sure if statement (II) is correct.

5. In the scatter plot below, the dotted straight lines mark the average values of X and Y.



Which of the following statements is/are correct?

(I) The line $Y = X$ cuts through the data points in half, with 50% of the data points on either side of the line.

(II) The average of Y is larger than the average of X.

(A) Only (I).

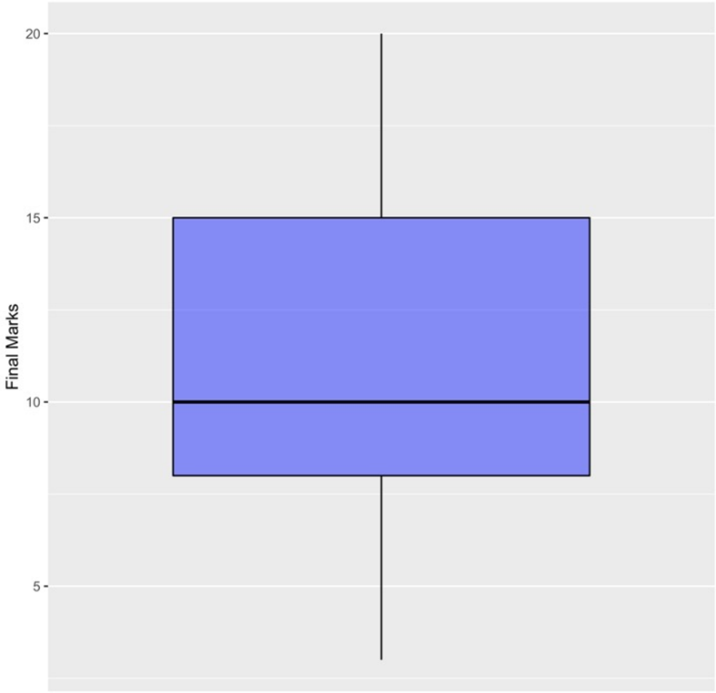
(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

Answer is (B). The intersection of the lines corresponding to the average values of X and Y lie above the $Y = X$ line. This shows that the average of Y is higher than the average of X. The $Y = X$ line does not cut through the data points in half. Visually, we can see that there are many more points that are above the line than below it.

6. The following boxplot shows the final examination marks of students from class A.



The passing mark for the final examination is 12 out of 20. Suppose that the boxplot of the final examination marks for class B is the same as the boxplot for class A. What can be said about the final examination marks of students from class B? Select all the correct statements.

(A) The proportion of students in class B who passed the examination must be the same as that for class A.

(B) At least 50% of the students from class B failed the final examination.

(C) The standard deviation of the students' marks in class B is equal to the standard deviation of those in class A.

(D) Based on the boxplot, there are no outliers in the grades of students from class B.

(E) The average of the students' marks for class B must be equal to that for class A.

(B) and (D) are correct. The boxplot above does not tell us anything about the percentage of students who obtained at least 12 out of 20 in the final examination.

The median of the final marks is 10, which implies that at least 50% of students failed the examination.

The boxplot does not tell us anything about the average and standard deviation of the final examination marks.

It is clear from the boxplot that there are no outliers.

7. The five-number summary of a numerical variable with 47 values is:

Min	Q_1	Median	Q_3	Max
12.0	15.0	16.5	18.0	24.0

Which of the following statements must be true? Select all that apply.

(A) There are no outliers in the data.

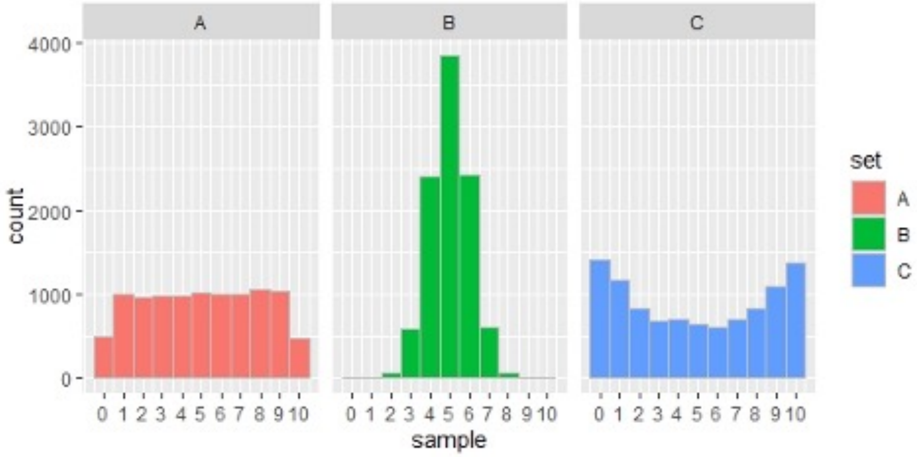
(B) There is at least one low outlier in the data.

(C) There is at least one high outlier in the data.

(D) There are both low and high outliers in the data.

Only (C) is true. The IQR is $18 - 15 = 3$. Since $24 > 18 + 1.5 \times 3 = 22.5$, 24 is a high outlier and thus there is at least one high outlier. As $15 - 1.5 \times 3 = 10.5$ and there are no values smaller than 12, there are no low outliers.

8. Consider data sets A, B and C, each consisting of 10,000 numbers with mean 5. The histograms for A, B and C are shown below.



Order the data sets according to the values of their standard deviations, from the smallest to the largest.

(A) A, B, C.

(B) A, C, B.

(C) B, A, C.

(D) B, C, A.

(E) C, A, B.

(F) C, B, A.

Answer is (C). Note that B has a smaller standard deviation compared to A, since more of the values are closer to the mean compared to A. C has a larger standard deviation compared to A, since more of the values are further from the mean compared to A. Thus the required order is B, A, C.

9. Suppose that the following are 10 data points for a numerical variable X:

$$3, \quad 50, \quad r, \quad 8, \quad 20, \quad 1, \quad 32, \quad 58, \quad 10, \quad 138,$$

where r is an unknown whole number and $r \neq 138$. Based on the definition of an outlier for a boxplot, if 138 is the only outlier in this data set, the **maximum** possible value of r is ____________________.

Answer is 133. Firstly, in order to have 138 as the only outlier, we must have $r < 138$. Then since we want to find the maximum possible value of r, when we order the other 9 points from the smallest to the largest value, we can try to let r be a number between the second largest number and 138. That is,

$$1, \quad 3, \quad 8, \quad 10, \quad 20, \quad 32, \quad 50, \quad 58, \quad r, \quad 138.$$

We can see that $Q_1 = 8$, $Q_3 = 58$ and IQR$= 50$. Thus, any number greater than $Q_3 + 1.5 \times$ IQR $= 58 + 1.5 \times 50 = 133$ will be an outlier. This means that the maximum possible value of r is 133.

10. Of the five values below, which would be that of a correlation coefficient with the strongest correlation?

(A) -1.4.

(B) -0.9.

(C) 0.

(D) 0.3.

(E) 0.7.

Answer is (B). A correlation coefficient always lies between -1 and 1 (inclusive). The higher the magnitude of the correlation coefficient, the stronger the correlation.

11. A researcher predicts the total number of bacteria in an experiment (denoted by y) using simple linear regression on $\ln y$ vs x. The regression equation is given by

$$\ln y = 0.5x + 2.5,$$

where the logarithm here is taken to base e, and x is the number of hours since 12PM. Which of the following statements is/are always correct?

(I) According to the researcher's model, there will be an average of 33 (rounded to the nearest integer) bacteria in 2 hours.

(II) The average number of bacteria is predicted to increase by the same amount every hour.

(A) Only (I).

(B) Only (II).

(C) Neither (I) nor (II).

(D) Both (I) and (II).

Answer is (A). When $x = 2$, $\ln y = 3.5$. Thus, we obtain $y = e^{3.5} = 33$. Statement (II) is not correct. We can verify this explicitly. For instance, from $x = 0$ to $x = 1$, the predicted average number of bacteria increases by 8. From $x = 1$ to $x = 2$, the predicted average number increases by 13.

12. In a study of 100 mother-daughter pairs, their heights were measured and plotted in a scatter diagram - mothers' heights at the horizontal axis, and their respective daughters' heights at the vertical axis. The horizontal and vertical axes are drawn on the same scale. All the data points are above the 45-degree line passing through the origin. Which of the following statements must be true?

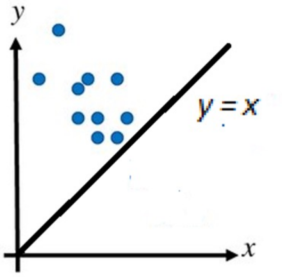
(A) The correlation between mothers' height and daughters' height is negative.

(B) The correlation between mothers' height and daughters' height is positive.

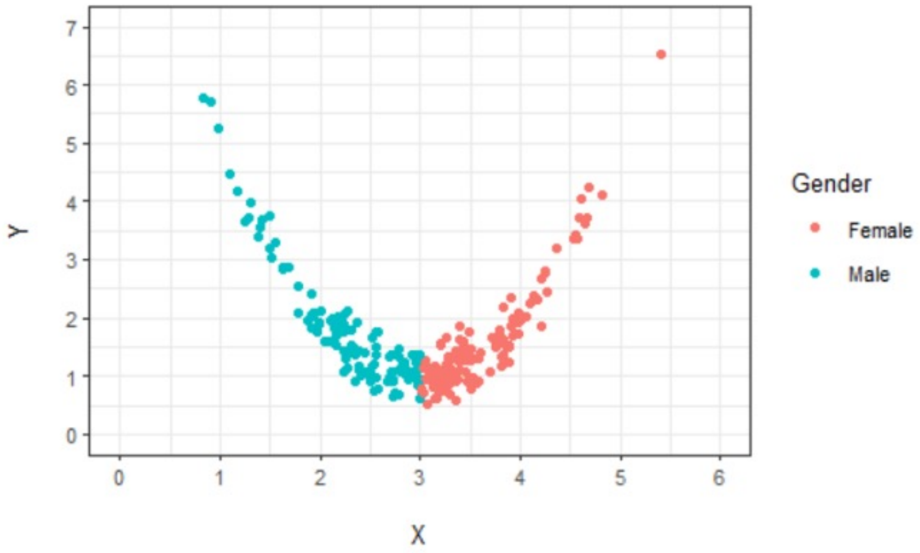
(C) The correlation between mothers' height and daughters' height is zero.

(D) None of the other given options is correct.

Answer is (D). Just because the scatter plot points are all above the 45-degree ($y = x$) line, it does not imply that the correlation must be positive/negative/zero. For example, the scatter plot points in the figure below indicates that the correlation is negative. The points can be easily rearranged, while keeping them above the 45-degree line, so that the correlation is positive or zero.



13. A researcher examined the relationship between variables X and Y among 250 male and female subjects. He graphed the relationship in the scatter plot shown below. Let r be the correlation coefficient for all 250 subjects, r_1 be the correlation coefficient among male subjects only and r_2 be the correlation coefficient among female subjects only.



Which of the following correctly describes the relationship between r, r_1 and r_2?

(A) $r_1 < r < r_2$.

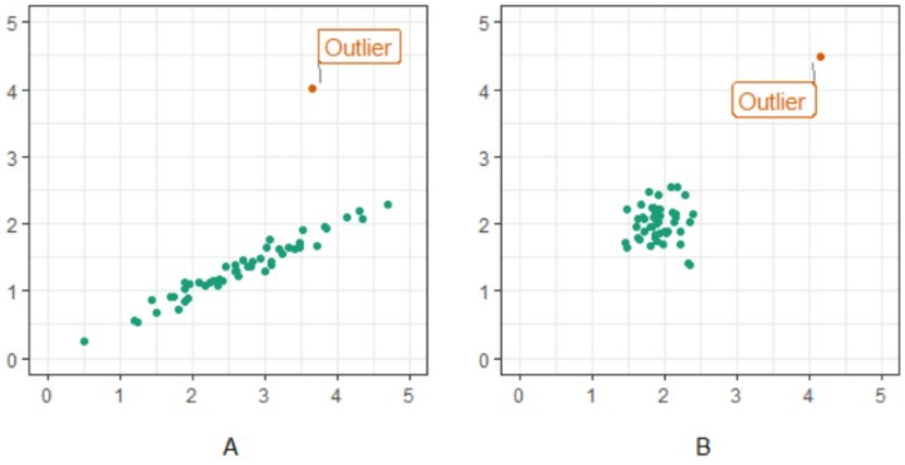
(B) $r_1 > r > r_2$.

(C) $r > r_1 > r_2$.

(D) $r < r_1 < r_2$.

Answer is (A). The correlation coefficient for all subjects is closer to zero when compared to either r_1 or r_2. The correlation coefficient for males only is negative, while the correlation coefficient for females only is positive.

14. Consider plots A and B shown below. What will happen to the correlation coefficients for both plots after removing the outliers indicated?



(A) The correlation coefficient in plot A will increase and correlation coefficient in plot B will decrease.

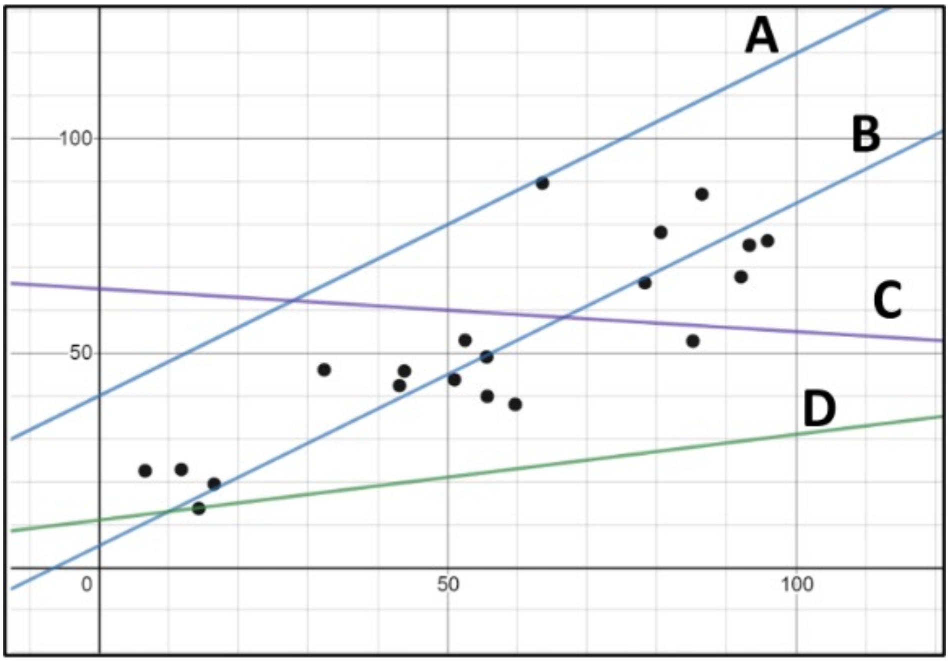
(B) The correlation coefficient in plot A will decrease and correlation coefficient in plot B will increase.

(C) The correlation coefficients in plots A and B will both increase.

(D) The correlation coefficients in plots A and B will both decrease.

Answer is (A). For plot A, removing the outlier will bring the correlation coefficient closer to 1 than before, so the correlation coefficient increases. For plot B, removing the outlier will bring the correlation coefficient, previously positive, closer to 0. Thus the correlation coefficient decreases.

15. A researcher examined the relationship between variables X and Y among 20 male subjects, and he graphed a scatter plot as shown below. One of the lines in the graph (A, B, C or D) is the actual best-fit regression line. Which one is it?



(A) Line A.

(B) Line B.

(C) Line C.

(D) Line D.

Answer is (B). The best-fit line is the line that on average, is closest to all the points in terms of the Y values, so that the overall distance from the data points to the line is minimised. Since line A is above all the data points, it cannot be the line of best-fit. For the same reason, since all the data points are either on or above line D, line D cannot be the line of best-fit. There is clearly a positive correlation between X and Y, so the line of best-fit must have a positive gradient, thus line C cannot be the line of best-fit.

16. A researcher is interested in the correlation between the amount of time an individual spends on social media and the individual's level of happiness. Suppose that she observed that the correlation coefficient r_1 for males only is 0.8, and that the correlation coefficient r_2 for females only is also 0.8. Which of the following statements must be true for r, the correlation coefficient when the data for males and females are combined?

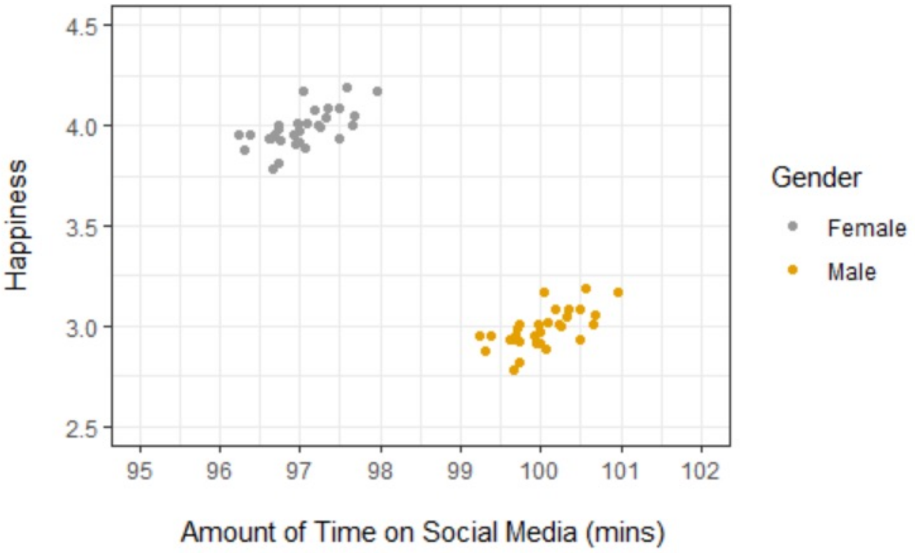
(A) $0 \leq r \leq 0.8$.

(B) $r = 0.8$.

(C) $0.8 < r \leq 1$.

(D) None of the other given options is correct.

Answer is (D). It is possible that the correlation coefficient in the combined data set is negative (see example below), so none of the other three options is correct.

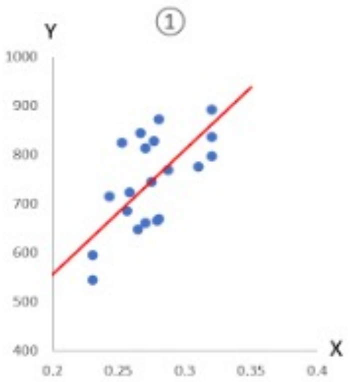


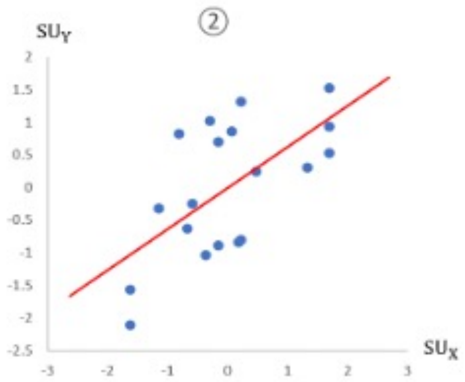
17. Let X and Y denote two variables measured on 19 subjects. Let the mean and standard deviation for the X values be written as $\overline{X}$ and s_X; similarly, the mean and standard deviation for the Y values are written as $\overline{Y}$ and s_Y. Plot "1" on the left shows the scatter plot for the 19 points (X, Y).

After converting the values of X and Y to standard units using the formulas:

$$SU_X = \frac{X - \overline{X}}{s_X} \qquad SU_Y = \frac{Y - \overline{Y}}{s_Y},$$

the scatter plot for the 19 points (SU_X, SU_Y) is shown as plot "2" on the right. The lines on both plots are the respective linear regression lines.





Which of the following statements is correct?

(A) The correlation coefficients in both plots are the same.

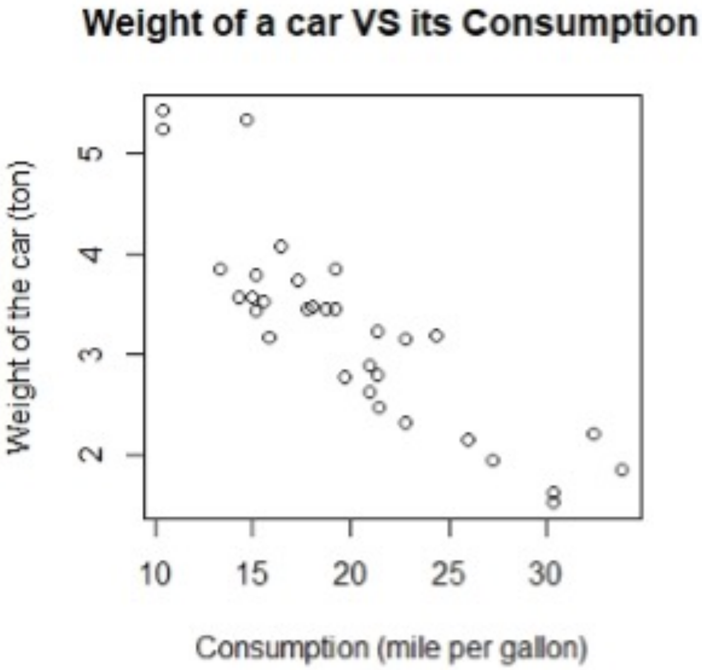
(B) The correlation coefficient decreased after conversion of X and Y to standard units.

(C) The correlation coefficient increased after conversion of X and Y to standard units.

(D) We do not have sufficient information to determine if the correlation coefficient has increased, decreased or remained the same after conversion of X and Y to standard units.

Answer is (A). Conversion to standard units involves (1) subtracting the same number ($\overline{X}$ or $\overline{Y}$) from all data points and (2) dividing all data points by the same positive number (s_X or s_Y). Both of these will not change the correlation coefficient.

18. Based on the scatter plot shown below, which of the following is closest to the equation for the regression line? Here, W is the weight of the car and C is the consumption.



(A) $W = 3 - 0.8C$.

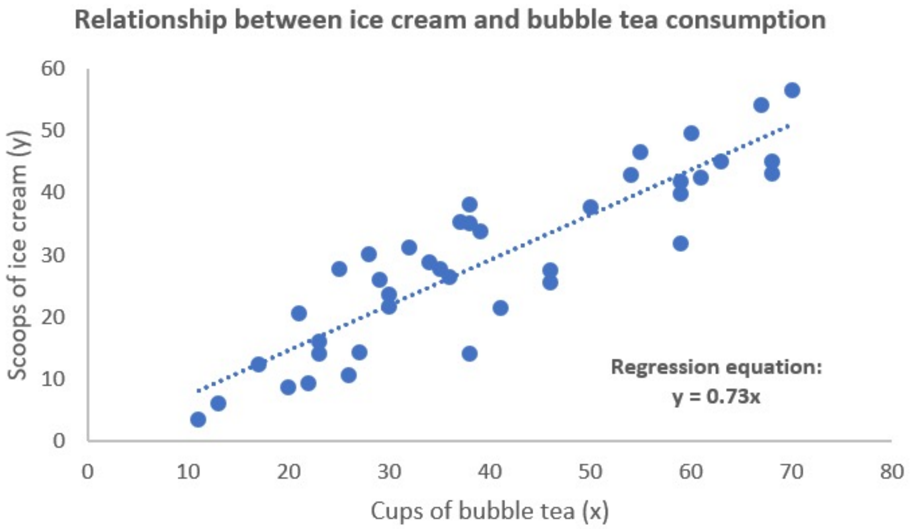
(B) $W = 5 - 0.8C$.

(C) $W = 3 + 0.8C$.

(D) $W = 5 + 0.8C$.

Answer is (B). The regression line should pass through the cloud of points in the scatter diagram. So its slope should be negative. Also, from the scatter plot, its y-intercept is more likely to be 5 than 3. Hence $W = 5 - 0.8C$ is the correct answer.

19. A group of students wanted to investigate the relationship between bubble tea and ice cream consumption. They conducted a survey to find out the number of cups of bubble tea and scoops of ice cream consumed in a year. From the data collected, they drew a scatter plot and fitted a linear regression line to the data with the equation of $y = 0.73x$.



Based on the information given in the scatter plot, which statement(s) must be true?

(I) The correlation between bubble tea and ice cream consumption is 0.73.

(II) People who consume 100 cups of bubble tea are predicted to consume 73 scoops of ice cream on average.

(A) Only (I).

(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

Answer is (D). The regression line equation tells us about the slope and intercept of the best-fit line. In this case, the slope is 0.73 and the intercept is 0. The slope of the regression line does not tell us the strength of the correlation so statement (I) is incorrect. There is no information about people who consume more than 70 cups of bubble tea from the current data. Hence we cannot use the regression line to predict the average number of scoops of ice cream consumed for someone who consumed 100 cups of bubble tea. Hence statement (II) is also incorrect.

20. There are two primary six classes in a tuition center. Class A and Class B each has 100 students and all students sat for a mathematics midterm test as well as a final examination. In Class A, every student scores 1 point higher in the final examination than in the midterm. In Class B, every student scores 1 point lower in the final examination than in the midterm. For the midterm test, the average score is 50 and standard deviation is 20 for both classes A and B. Which of the following statements is/are correct?

(I) The correlation coefficient between the final examination score and the midterm score in Class A is 1.

(II) The correlation coefficient between the final examination score and the midterm score in Class B is -1.

(A) Only (I).

(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

Answer is (A). Note that the relationship between the midterm score and final examination score is deterministic. Furthermore, since standard deviation is 20, not all students scored the same midterm/final examination mark. For Class A, imagine the points (49,50), (50,51), (52,53) and so on, each (x, y) representing the midterm score x and final exam score y. Clearly, these points lie on the straight line $y = x + 1$ with positive gradient. So the correlation coefficient between final examination score and the midterm score in Class A is 1. Similarly, for Class B, imagine the points (49,48), (50,49), (52, 51) and so on. These points also lie on a straight line $y = x - 1$ with positive gradient. The correlation coefficient between final examination score and the midterm score in Class B is also 1.

21. There are two primary six classes in a tuition center. Class A and Class B each has 100 students and all students sat for a mathematics midterm test as well as a final examination. In Class A, every student scores 1 point higher in the final examination than in the midterm. In Class B, every student scores 1 point lower in the final examination than in the midterm. For the midterm test, the average score is 50 and standard deviation is 20 for both Classes A and B. Suppose now Class C is formed by combining all the students from Classes A and B. Which of the following statements is/are correct?

    (I) The correlation coefficient of Class C is smaller than the correlation coefficient of Class A.

    (II) The correlation coefficient of Class C is larger than the correlation coefficient of Class B.

    (A) Only (I).

    (B) Only (II).

    (C) Both (I) and (II).

    (D) Neither (I) nor (II).

    Answer is (A). Note that the relationship between the midterm score and final examination score is deterministic. Furthermore, since standard deviation is 20, not all students scored the same midterm/final examination mark. The correlation coefficient of Class A is 1. The regression line for Class A has equation $y = x + 1$ where y is the final examination score and x is the midterm score. The correlation coefficient for Class B is also 1 and the regression line for Class B is $y = x - 1$. Statement (II) is incorrect since the correlation coefficient of Class C cannot be larger than 1. We know that the correlation coefficient of Class C is less than 1 since the points (on two parallel lines) will not fall on a single straight line. Thus statement (I) is correct.

22. Which of the following is/are true about a non-zero correlation coefficient? Select all that apply.

    (A) The correlation coefficient does not change when we add 5 to all the values of one variable.

    (B) The correlation coefficient is positive when the slope of the regression line is positive.

    (C) The correlation coefficient does not change when we multiply all the values of one variable by 2.

    (D) A correlation of -0.3 is stronger than a correlation of -0.8.

    (A), (B) and (C) are correct. The correlation coefficient r does not change when we add or multiply all the values of one variable by a positive number. Since $m = r\left(\frac{S_y}{S_x}\right)$, r and m will have the same sign. Only (D) is incorrect, as a correlation of -0.8 is stronger (since it is closer to -1) than a correlation of -0.3.

23. For the 40 students in a class, the results of their second English test are plotted against the results of their first English test. It was found that for each student, the result of the second test is better than that of the first test (i.e., the second test score is higher). Which of the following must be true about the relationship between the students' second test results and their first test results?

    (I) If student A scores better than student B in the first test, then student A also scores better than student B in the second test.

(II) The correlation between students' second test results and first test results is positive.

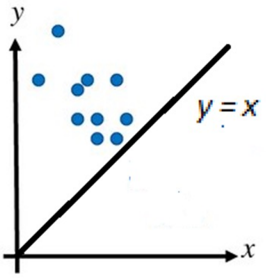
(A) Only (I).

(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

Answer is (D). Let x denote the variable representing the test score of a student's first English test, and y denote the variable representing the test score of a student's second English test. Then all we know from the results stated is that the y value is greater than the x value for each student. Statements (I) and (II) may not be true. We can easily plot a downward sloping graph where all the points lie above the line $y = x$ to see that both statements do not hold (see example below).



24. A student produces a scatter plot whereby the y values range from 1 to 3, and he observes a linear association, with regression equation $y = 1.5x + 5$. Which of the following statements is/are correct?

(I) The correlation coefficient must be positive.

(II) The average y value of the data set collected is 2.

(A) Only (I).

(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

Answer is (A). Since the regression line has a positive slope, the correlation coefficient must be positive and thus statement (I) is correct. While we know that the y values ranges from 1 to 3, it does not mean that the average value of y is 2, so statement (II) is not correct.

25. There is a weak positive linear association between numerical variables X and Y, where X ranges from 0 to 5 (inclusive). Based on the data from X and Y, the regression line is given by the equation $Y = 0.25X + 2$. Which of the following statements must be true? Select all that apply.

(A) We can obtain the exact value of Y when $X = 4$.

(B) The predicted average value of Y is 4 when $X = 8$.

(C) The correlation coefficient is 0.25.

(D) The equation corresponds to a non-deterministic relationship between X and Y.

Only (D) is correct. A weak association between 2 variables corresponds to a non-deterministic relationship since one particular X value can correspond to many Y values. The regression equation only gives us the predicted average of the Y values for a given X value and not the exact Y value. We cannot draw any conclusions of the predicted average values of Y for any X beyond the range of X values in the data. Lastly, the correlation coefficient is generally not the gradient of the regression line.

26. A researcher wanted to find the correlation between heights of 100 father-and-son pairs. After collecting and analysing his data, he realised that the device he had been using to measure height suffered from significant bias causing every measurement to be too high by 10cm. He then corrected the values of all his analyses. After the correction was done, which of the following will change? Select all that apply.

    (A) The correlation coefficient between the heights of father-and-son pairs.

    (B) The standard deviations of son's height and father's height.

    (C) The average son's height and the average father's height.

    Only (C) is correct. The measurements are systematically off the mark in the same direction. The correlation coefficient does not change by adding or subtracting a number to all values of a variable. The spread or the standard deviation will not change either. Only the average of the heights will change.

27. The relationship between the number of glasses of beer consumed daily (x) and blood alcohol content in percentage (y) was studied in young adults. The equation of the regression line is $y = -0.015 + 0.02x$ for $1 \leq x \leq 10$. The legal limit to drive in Singapore is having a blood alcohol content below 0.08%. Des, a young adult, had just finished 5 glasses of beer. After that, he wanted to take his car out for a drive. Is it legal for him to drive in Singapore?

    (A) Yes.

    (B) No.

    (C) Unable to determine.

    Answer is (C). The regression line only provides the predicted average blood alcohol content for someone who drank 5 glasses of beer, which is 0.085%. Although the value is in the illegal range, Des' blood alcohol content may have been below average, and not have hit 0.08%.

28. You are given that the variables X and Y are negatively correlated. Which of the following statements must be true? Select all that apply.

    (A) If we multiply all the values of X and Y by -1, then the correlation coefficient between X and Y will change.

    (B) The gradient of the regression line for Y vs X is the same as the gradient of the regression line for X vs Y.

    (C) If we remove an outlier from the data set, the correlation coefficient will change.

    (D) If we add 6 to each value of X and subtract 3 from each value of Y, the correlation coefficient does not change.

    (E) If there are only 2 points, the correlation coefficient between X and Y must be -1.

    (D) and (E) are correct. Multiplying all values of X or Y (but not both variables) by -1 will change the sign of the r value between X and Y. Hence if we multiply all values of X and Y by -1, there will be no change in the r value. As the gradient of the regression line for Y vs X is $r\left(\frac{S_Y}{S_X}\right)$ but the gradient of the regression line for X vs Y is $r\left(\frac{S_X}{S_Y}\right)$, we can see that the two gradients are not the same unless $S_Y = S_X$. If all the (X, Y) points lie on a negatively sloped straight line including the outlier(s), the r value remains at -1 after any outlier is removed. Adding any number to or subtracting any number from X or Y will not change the correlation coefficient between X and Y. If there are only 2 points, since X and Y are negatively correlated, we can connect them with a negatively sloped straight line and hence r must be -1.

29. 4 students take a midterm examination and a final examination. The minimum and maximum midterm scores are 20 and 40 respectively. The minimum and maximum final scores are 60 and 80 respectively. All midterm and final scores of the 4 students are plotted on a scatter diagram: the midterm scores on the horizontal x-axis, and the final scores on the vertical y-axis.

Consider the following statements:

(I) All points in the scatter plot lie above the line $y = x$.

(II) The correlation coefficient between the midterm scores and final scores **must be** nonzero.

Which of the above statements is/are true?

(A) Only (I).

(B) Only (II).

(C) Both (I) and (II).

(D) Neither (I) nor (II).

Answer is (A). The possible range of values for the midterm scores is between 20 and 40. The possible range for the final scores is between 60 and 80. This gives a region of the scatter plot where the points could lie, and the entire region lies above the line $y = x$. Thus statement (I) is true.

It is possible that the correlation coefficient is 0. The points are only restricted to the region mentioned above. Within the region, the arrangement of points in the scatter plot can take on any patten. For example, suppose that the 4 students obtained midterm and final scores as follows:

x (midterm)	y (final exam)
20	60
20	80
40	60
40	80

Calculating the correlation coefficient r shows that $r = 0$. Hence it is not necessarily true that correlation coefficient is always nonzero, so statement (II) is not true.

30. Suppose that there are 40 male students in a class and each student scored 5 less marks for his maths test than what he scored for his science test. What can we say about their maths and science test marks? Select all that apply.

(A) The interquartile range of science test marks is higher than that for maths test marks.

(B) If student A scored a higher mark for the maths test than student B, then he must have scored a higher mark than student B for the science test.

(C) The science test marks and maths test marks are perfectly negatively correlated.

(D) The standard deviation of maths test marks is equal to that of science test marks.

(B) and (D) are correct. Since quartile 1 and quartile 3 of the maths test marks decrease by the same amount (5 marks) as compared to quartile 1 and quartile 3 of the science test marks, there is no difference in the interquartile ranges of the maths and science test marks.

As standard deviation does not change when we subtract or add a number to every data point in a data set, the standard deviations of the maths and science test marks are equal.

Suppose that we let x and y denote the maths and science test marks of the students, respectively. Then we see that $y = x + 5$, that is, there is a perfect positive correlation between the science and maths test marks. In addition, y increases as x increases. Thus, if student A scored a higher mark for maths than student B, then he must have scored a higher mark than student B for the science test as well.