# Sheet 3 (Correlation, Regression, Review on Probability)

## 1. (Correlation)

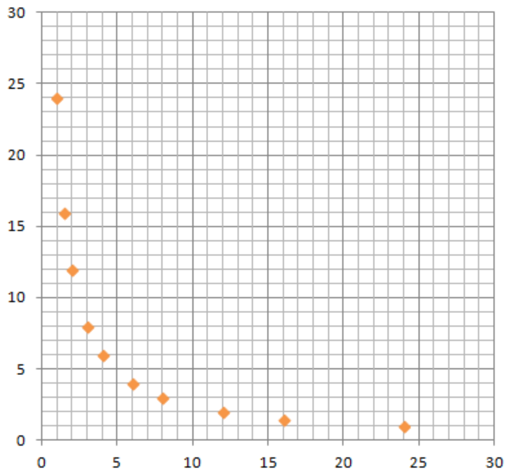
1. The following table gives the heights and weights of 10 friends:

| Name | Height (cm) | Weight (kg) |
|---|---|---|
| Albert | 180 | 87 |
| Beth | 176 | 55 |
| Cindy | 144 | 52 |
| David | 195 | 94 |
| Emily | 159 | 87 |
| Frank | 185 | 79 |
| Gary | 166 | 59 |
| Helen | 173 | 64 |
| Ida | 149 | 45 |
| Jeremy | 168 | 77 |

Which one of the following best describes the correlation between their heights and weights? (Hint: draw a scatter plot)

A. High positive correlation
B. Low positive correlation
C. No correlation
D. Low negative correlation

2. For the following scatterplot:



how could you describe the correlation?

A. There is no correlation.
B. There is positive linear correlation.
C. There is negative linear correlation.
D. There is correlation which is shown by the points best fitting curve.

3. Calculate the correlation coefficient for the following data:

| $x$ | $y$ |
|---|---|
| 12 | 1 |
| 8 | 7 |
| 5 | 4 |
| 3 | 6 |
| 2 | 4 |
| 0 | 2 |

A. -0.120
B. -0.2
C. -0.220
D. -0.320

4. Calculate the correlation coefficient for the following data:

| $x$ | $y$ |
|---|---|
| 1 | 9 |
| 2 | 6 |
| 4 | 4 |
| 6 | 12 |
| 7 | 8 |
| 10 | 3 |

A. -0.168
B. -0.2
C. -0.268
D. -0.368

5. Calculate the correlation coefficient and **comment on its value** for the following data:

| $x$ | $y$ |
|---|---|
| 1 | 4 |
| 2 | 6 |
| 3 | 5 |
| 4 | 7 |
| 5 | 9 |
| 6 | 11 |
| 7 | 12 |
| 8 | 17 |
| 9 | 19 |
| 10 | 20 |

A. 0.9
B. 0.932
C. 0.952
D. 0.972

6. Calculate the correlation coefficient and comment on its value for the following data:

| $x$ | $y$ |
|---|---|
| 18 | 1 |
| 16 | 3 |
| 15 | 5 |
| 11 | 6 |
| 12 | 9 |
| 10 | 11 |
| 8 | 10 |
| 4 | 12 |
| 2 | 11 |
| 0 | 15 |

A. -0.822
B. -0.9
C. -0.902
D. -0.922

7. The following table gives the heights and weights of 10 friends:

| Name | Height (cm) | Weight (kg) |
|---|---|---|
| Albert | 180 | 87 |
| Beth | 176 | 65 |
| Cindy | 144 | 52 |
| David | 195 | 94 |
| Emily | 159 | 87 |
| Frank | 185 | 79 |
| Gary | 166 | 59 |
| Helen | 173 | 64 |
| Ida | 149 | 45 |
| Jeremy | 168 | 77 |

Calculate the Pearson's correlation coefficient.
A. -0.9362
B. -0.7294
C. 0.7294
D. 0.9362

8. The following table gives the math scores and times taken to run 100 m for 10 friends:

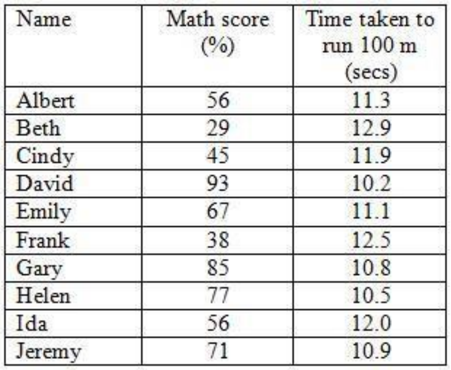
   Calculate the Pearson's correlation coefficient.

A. -0.9716

B. -0.9602

C. 0.9602

D. 0.9716



| Name | Math score (%) | Time taken to run 100 m (secs) |
|---|---|---|
| Albert | 56 | 11.3 |
| Beth | 29 | 12.9 |
| Cindy | 45 | 11.9 |
| David | 93 | 10.2 |
| Emily | 67 | 11.1 |
| Frank | 38 | 12.5 |
| Gary | 85 | 10.8 |
| Helen | 77 | 10.5 |
| Ida | 56 | 12.0 |
| Jeremy | 71 | 10.9 |

9. The following table gives the heights and math scores of 10 friends:

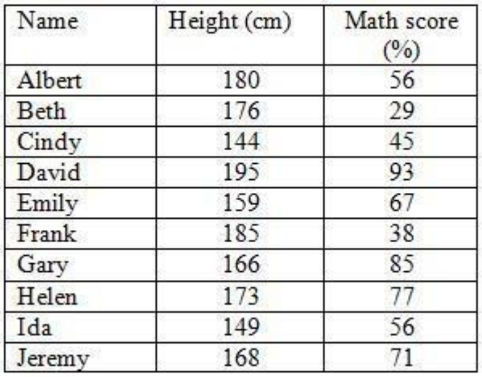
   Calculate the Pearson's correlation coefficient.

A. -0.2391

B. -0.2051

C. 0.2051

D. 0.2391



| Name | Height (cm) | Math score (%) |
|---|---|---|
| Albert | 180 | 56 |
| Beth | 176 | 29 |
| Cindy | 144 | 45 |
| David | 195 | 93 |
| Emily | 159 | 67 |
| Frank | 185 | 38 |
| Gary | 166 | 85 |
| Helen | 173 | 77 |
| Ida | 149 | 56 |
| Jeremy | 168 | 71 |

10. The opposite table gives the weights and waist sizes of 10 friends:

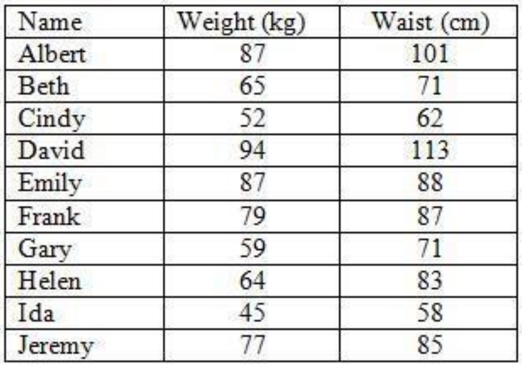
   Calculate the Pearson's correlation coefficient.

A. -0.9438

B. -0.9039

C. 0.9039

D. 0.9438



| Name | Weight (kg) | Waist (cm) |
|---|---|---|
| Albert | 87 | 101 |
| Beth | 65 | 71 |
| Cindy | 52 | 62 |
| David | 94 | 113 |
| Emily | 87 | 88 |
| Frank | 79 | 87 |
| Gary | 59 | 71 |
| Helen | 64 | 83 |
| Ida | 45 | 58 |
| Jeremy | 77 | 85 |

## 2. (Regression)

11. The evil Swindler has been collecting data on the effect of radiation exposure has on Captain Amazing's super powers. Here is the number of minutes of exposure to radiation, paired with the number of tons Captain amazing is able to lift.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Radiation exposure** | 4 | 4.5 | 5 | 5.5 | 6 | 6.5 | 7 |
| **Weight (tons)** | 12 | 10 | 8 | 9.5 | 8 | 9 | 6 |

a) Your job is to find the correlation coefficient to describe the strength of the relationship between the variables, and use the least squares regression to find the line of best fit. Sketch the scatter plot too.

b) Use the value of r to comment on the regression analysis.

c) If Swindler exposes Captain Amazing to radiation for 5 minutes, what weight do you expect Captain Amazing to be able to lift?

12. Last year, five randomly selected students took a math aptitude test before they began their statistics course. The Statistics Department has three questions.

| Student | xi | yi |
|---|---|---|
| 1 | 95 | 85 |
| 2 | 85 | 95 |
| 3 | 80 | 70 |
| 4 | 70 | 65 |
| 5 | 60 | 70 |

- Draw the scatter plot representing the data.
- What linear regression equation best predicts statistics performance, based on math aptitude scores?
- If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?
- How well does the regression equation fit the data? (hint: use the coefficient of determination to answer this question)

13. The following data was collected from an experiment. In the experiment, objects of different masses were placed on a horizontal surface and the force needed to make them start to move was recorded.

| *Mass*(kg) | 0.5 | 1.0 | 1.5 | 2.0 | 3.0 | 5.0 |
|---|---|---|---|---|---|---|
| *Force*(N) | 2.1 | 3.8 | 6.1 | 7.9 | 13.2 | 19.1 |

Estimate the force needed for a 2.5 kg mass.

14. Rafiq collected the following data on the height and shoe size of some pupils in his class:

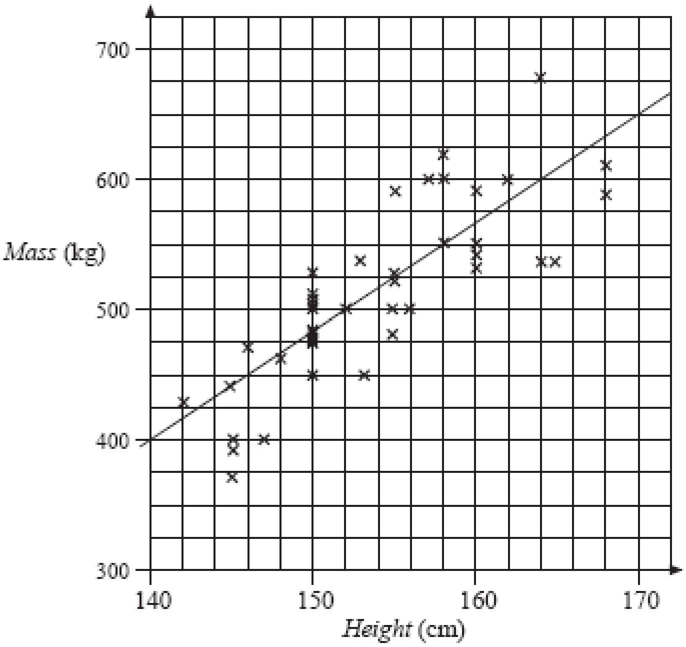
| *Shoe Size* | 6 | 4 | 8 | 5 | 9 | 10 | 4 | 5.5 |
|---|---|---|---|---|---|---|---|---|
| *Height* (cm) | 143 | 150 | 172 | 146 | 165 | 177 | 141 | 156 |

Draw a scatter plot and a line of best fit for the data.

(b) Estimate the height of a person with a shoe size of 7.5.

(c) Ian has a height of 170 cm. Estimate his shoe size.

15. The scatter diagram shows the heights and masses of some horses. The scatter diagram also shows a line of best fit.

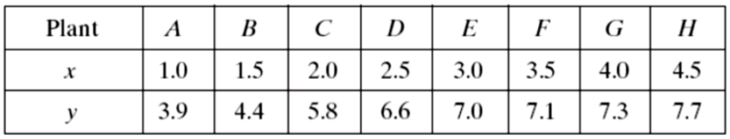


a) What does the scatter diagram show about the *relationship* between the height and mass of the horses?

b) The *height* of a horse is 163 cm. Use the line of best fit to estimate the mass of the horse.

c) A different horse has a mass of 625 kg. Use the line of best fit to estimate the height of the horse.

16. A biologist assumes that there is a linear relationship between the amount of fertilizer supplied to a tomato plant and the subsequent yield of tomatoes obtained. Eight tomato plants, of the same variety, were selected at random and treated, weekly, with a solution of $x$ grams of fertilizer dissolved in a fixed amount of water. The yield, $y$ kilograms, of potatoes was recorded.



| Plant | *A* | *B* | *C* | *D* | *E* | *F* | *G* | *H* |
|---|---|---|---|---|---|---|---|---|
| $x$ | 1.0 | 1.5 | 2.0 | 2.5 | 3.0 | 3.5 | 4.0 | 4.5 |
| $y$ | 3.9 | 4.4 | 5.8 | 6.6 | 7.0 | 7.1 | 7.3 | 7.7 |

a) Is the assumption of the biologist true ($\bar{x} = 2.75, \bar{y} = 6.225, S_x = 1.22, S_y = 1.4$)? What does the coefficient of determination $r^2$ indicate?

b) If the assumption is true, Calculate the least squares regression line of $y$ on $x$.

c) Estimate the yield of a plant treated weekly with 3.2 grams of fertilizer.

## 3. (Revision on Probability)

1. A researcher hypothesizes that it is possible to detect membrane proteins using the fraction of hydrophobic residues alone. To test this model, the researcher creates a library of 7500 proteins and scores each of these proteins based on their fraction of hydrophobic residues and whether they are membrane proteins. The results of this analysis are shown below:

| | Majority Hydrophobic | Majority Hydrophilic |
|---|---|---|
| Membrane Bound | 2911 | 961 |
| Cytosolic | 713 | 2915 |

Given this information, find the likelihood that a novel protein that is primarily hydrophobic is also a membrane protein.

2. A test for a rare disease claims that it will report a positive result for 99.5% of people with the disease, and will report a negative result for 99.9% of those without the disease. We know that the disease is present in the population at 1 in 100,000. Knowing this information, what is the likelihood that an individual who tests positive will actually have the disease?

3. Three jars contain colored balls as described in the table below:

| Jar # | Red | White | Blue |
|---|---|---|---|
| 1 | 3 | 4 | 1 |
| 2 | 1 | 2 | 3 |
| 3 | 4 | 3 | 2 |

   One jar is chosen at random and a ball is selected. If the ball is red, what is the probability that it came from the 2nd jar?

4. All tractors made by a company are produced on one of three assembly lines, named Red, White, and Blue. The chances that a tractor will not start when it rolls off of a line are 6%, 11%, and 8% for lines Red, White, and Blue, respectively. 48% of the company's tractors are made on the Red line and 31% are made on the Blue line. What fraction of the company's tractors do not start when they roll off of an assembly line?

Bonus question: What is the probability that a tractor came from the red company given that it was defective?

5. There are two urns containing colored balls. The first urn contains 50 red balls and 50 blue balls. The second urn contains 30 red balls and 70 blue balls. One of the two urns is randomly chosen (both urns have probability 50% of being chosen) and then a ball is drawn at random from one of the two urns. If a red ball is drawn, what is the probability that it comes from the first urn?

6. An economics consulting firm has created a model to predict recessions. The model predicts a recession with probability 80% when a recession is indeed coming and with

probability 10% when no recession is coming. The unconditional probability of falling into a recession is 20%. If the model predicts a recession, what is the probability that a recession will indeed come?

7. Alice has two coins in her pocket, a fair coin (head on one side and tail on the other side) and a two-headed coin. She picks one at random from her pocket, tosses it and obtains head. What is the probability that she flipped the fair coin?

8. In a certain population, 30% of the persons smoke and 8% have a certain type of heart disease. Moreover, 12% of the persons who smoke have the disease.

   a) What percentage of the population smoke and have the disease?
   b) What percentage of the population with the disease also smoke?
   c) Are smoking and the disease positively correlated, negatively correlated, or independent?

9. A company has 200 employees: 120 are women and 80 are men. Of the 120 female employees, 30 are classified as managers, while 20 of the 80 male employees are managers. Suppose that an employee is chosen at random.

   a) Find the probability that the employee is female.
   b) Find the probability that the employee is a manager.
   c) Find the conditional probability that the employee is a manager given that the employee is female.
   d) Find the conditional probability that the employee is female given that the employee is a manager.
   e) Are the events *female* and *manager* positively correlated, negatively correlated, or indpendent?