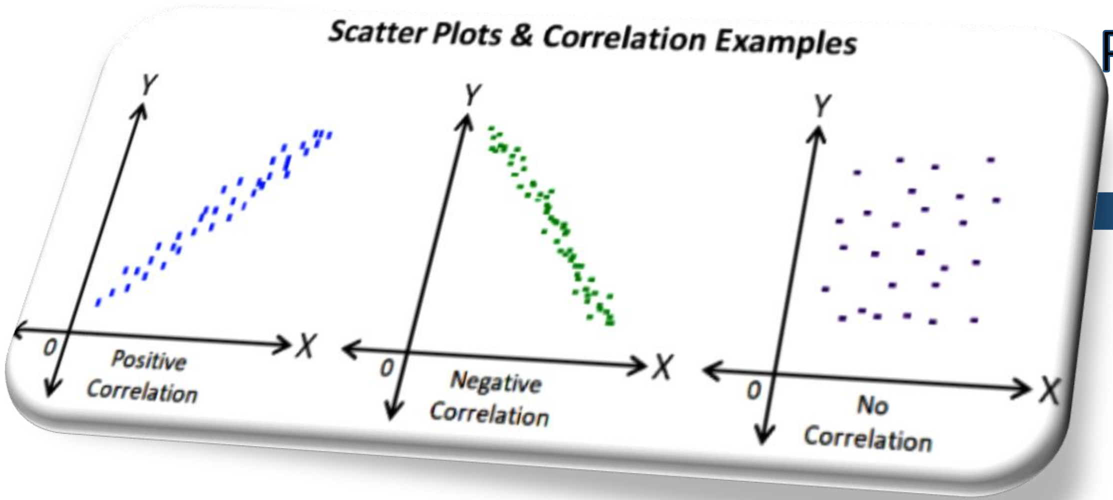
Introduction to Probability and Statistics

Topic (3): “Describing Bivariate Data”



Dr. Heba Ayyoub

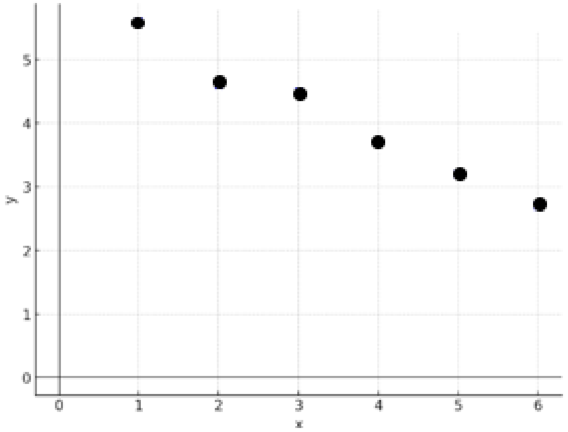
Philadelphia University

Solution to the Exercises for the Third Topic

Exercise (1): Consider this set of bivariate data:

x	1	2	3	4	5	6
y	5.6	4.6	4.5	3.7	3.2	2.7

a. Draw a scatterplot to describe the data.



b. Does there appear to be a relationship between x and y? If so, how do you describe it?

Negative relationship between x and y.

c. Calculate the correlation coefficient. Does the value of r confirm your conclusions in part **b**? Explain.

$$r = \frac{S_{xy}}{S_x S_y}$$

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	5.6	5.6	1	31.36
2	4.6	9.2	4	21.16
3	4.5	13.5	9	20.25
4	3.7	14.8	16	13.69
5	3.2	16	25	10.25
6	2.7	16.2	36	7.29
21	24.3	75.3	91	104

$$S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1} = \frac{75.3 - 6(3.5)(4.05)}{6-1} = -1.95$$

$$S_x^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{91 - 6(3.5)^2}{6-1} = 3.5$$

$$S_y^2 = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1} = \frac{104 - 6(4.05)^2}{6-1} = 1.117$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{21}{6} = 3.5$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{24.3}{6} = 4.05$$

$$\therefore r = \frac{S_{xy}}{S_x S_y} = \frac{-1.95}{\sqrt{3.5}\sqrt{1.117}} \approx -0.986$$

Conclusion: Very strong negative linear relationship between x and y.

This confirms the conclusion from the scatterplot.

d. Obtain the equation of the best fitting line.

$$\hat{y} = a + bx$$

$$b = \frac{S_{xy}}{S_x^2} = \frac{-1.95}{3.5} \approx -0.56$$

$$a = \bar{y} - b\bar{x} = 4.05 + 0.56(3.5) = 6.01$$

$$\therefore \hat{y} = 6.01 - 0.56x$$

e. Predict (estimate) y for x = 5 using the best fitting line?

$$\hat{y} = 6.01 - 0.56(5) = 3.21$$

f. Find the amount of error in the prediction of the best fitting line in part (**e**)?

$$e = |y - \hat{y}| = |3.2 - 3.21| = 0.01$$

Exercise (2): Data for the studying hours and final grades given in the following table.

Student	A	B	C	D	E	F
Studied Hours (x_i)	6	2	1	5	2	3
Grade (y_i)	82	63	57	88	68	75

1) Calculate the correlation coefficient and interpret the result.

$$r = \frac{S_{xy}}{S_x S_y}$$

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
6	82	492	36	6724
2	63	126	4	3969
1	57	57	1	3249
5	88	440	25	7744
2	68	136	4	4624
3	75	225	9	5625
19	433	1476	79	31935

$$S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1} = \frac{1476 - 6(3.17)(72.17)}{6-1} = 20.67$$

$$S_x^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{79 - 6(3.17)^2}{6-1} = 3.74$$

$$S_y^2 = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1} = \frac{31935 - 6(72.17)^2}{6-1} = 136.79$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{19}{6} = 3.17$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{433}{6} = 72.17$$

$$\therefore r = \frac{S_{xy}}{S_x S_y} = \frac{20.67}{\sqrt{3.74}\sqrt{136.79}} \approx 0.914$$

Conclusion: Very strong positive linear relationship between x and y (as studying hours increase the final grades increase).

2) Find the best fitting line.

$$\hat{y} = a + bx$$

$$b = \frac{S_{xy}}{S_x^2} = \frac{20.67}{3.74} \approx 5.53$$

$$a = \bar{y} - b\bar{x} = 72.17 - 5.53(3.17) = 54.64$$

$$\therefore \hat{y} = 54.64 + 5.53x$$

3) Predict (estimate) y for x = 5 using the best fitting line?

$$\hat{y} = 54.64 + 5.53(5) = 82.29$$

Exercise (3): Calculate the correlation coefficient for the number of absences and final grades given in the following table. Interpret the result.

Student	A	B	C	D	E	F	G
Number of Absences (x_i)	6	2	15	9	12	5	8
Grade (y_i)	82	86	43	74	58	90	78

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
6	82	492	36	6724
2	86	172	4	7396
15	43	645	225	1849
9	74	666	81	5476
12	58	696	144	3364
5	90	450	25	8100
8	78	624	64	6084
57	511	3745	579	38993

$$r = \frac{S_{xy}}{S_x S_y}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{57}{7} = 8.14$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{511}{7} = 73$$

$$S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1} = \frac{3745 - 7(8.14)(73)}{7-1} = -69.09$$

$$S_x^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{579 - 7(8.14)^2}{7-1} = 19.2$$

$$S_y^2 = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1} = \frac{38993 - 7(73)^2}{7-1} = 281.67$$

$$\therefore r = \frac{S_{xy}}{S_x S_y} = \frac{-69.09}{\sqrt{19.2}\sqrt{281.67}} \approx -0.94$$

Conclusion: Very strong negative linear relationship between x and y (as number of absences increase the final grades decrease).

Exercise (4): Calculations from a data set of pairs of n = 36 pairs of (x, y) values have provided the following results.

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 530.7$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 235.4$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = -204.3$$

Obtain the correlation coefficient.

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{-204.3}{36-1} = -5.84$$

$$S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{530.7}{36-1} = 15.16$$

$$S_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{235.4}{36-1} = 6.73$$

$$r = \frac{S_{xy}}{S_x S_y} = \frac{-5.84}{\sqrt{15.16}\sqrt{6.73}} = -0.58$$

Exercise (5): Evaluate the value of the correlation coefficient for the data with the following properties.

$$n = 30, \sum_{i=1}^{n} x_i = 680, \sum_{i=1}^{n} x_i^2 = 20154\,, \sum_{i=1}^{n} y_i = 996, \sum_{i=1}^{n} y_i^2 = 34670\,, \sum_{i=1}^{n} x_i y_i = 24844$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{680}{30} = 22.67$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{996}{30} = 33.2$$

$$S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1} = \frac{24844 - 30(22.67)(33.2)}{30-1} = 78.09$$

$$S_x^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{20154 - 30(22.67)^2}{30-1} = 163.31$$

$$S_y^2 = \frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1} = \frac{34670 - 30(33.2)^2}{30-1} = 55.27$$

$$\therefore r = \frac{S_{xy}}{S_x S_y} = \frac{78.09}{\sqrt{163.31}\sqrt{55.27}} \approx 0.82$$

Exercise (6): Find the equation of the regression line that best fits the following data (the best fitting line).

x	y	$x_i y_i$	x_i^2
0	1	0	0
1	5	5	1
2	3	6	4
3	9	27	9
4	7	28	16
10	25	66	30

$$S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1} = \frac{66 - 5(2)(5)}{5-1} = 4$$

$$S_x^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{30 - 5(2)^2}{5-1} = 2.5$$

$$\hat{y} = a + bx$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{10}{5} = 2$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{25}{5} = 5$$

$$b = \frac{S_{xy}}{S_x^2} = \frac{4}{2.5} = 1.6$$

$$a = \bar{y} - b\bar{x} = 5 - 1.6(2) = 1.8$$

$$\therefore \hat{y} = 1.8 + 1.6x$$

Then predict the value of y when x = 2.5.

$$\hat{y} = 1.8 + 1.6(2.5) = 5.8$$

Exercise (7): Find the equation of the regression line that best fits the following data (the best fitting line).

$$S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1} = \frac{77 - 5(5)(3.2)}{5-1} = -0.75$$

$$S_x^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{155 - 5(5)^2}{5-1} = 7.5$$

$$\hat{y} = a + bx$$

$$b = \frac{S_{xy}}{S_x^2} = \frac{-0.75}{7.5} = -0.1$$

$$a = \bar{y} - b\bar{x} = 3.2 + 0.75(5) = 6.95$$

$$\therefore \hat{y} = 6.95 - 0.1x$$

x	y	$x_i y_i$	x_i^2
8	1	8	64
1	4	4	1
4	2	8	16
7	6	42	49
5	3	15	25
25	16	77	155

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{25}{5} = 5$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{16}{5} = 3.2$$

Exercise (8): True or False

1) A scatterplot is used to display the relationship between two quantitative variables. **True**

2) A pie chart is most useful for showing the relationship between two variables. **False**