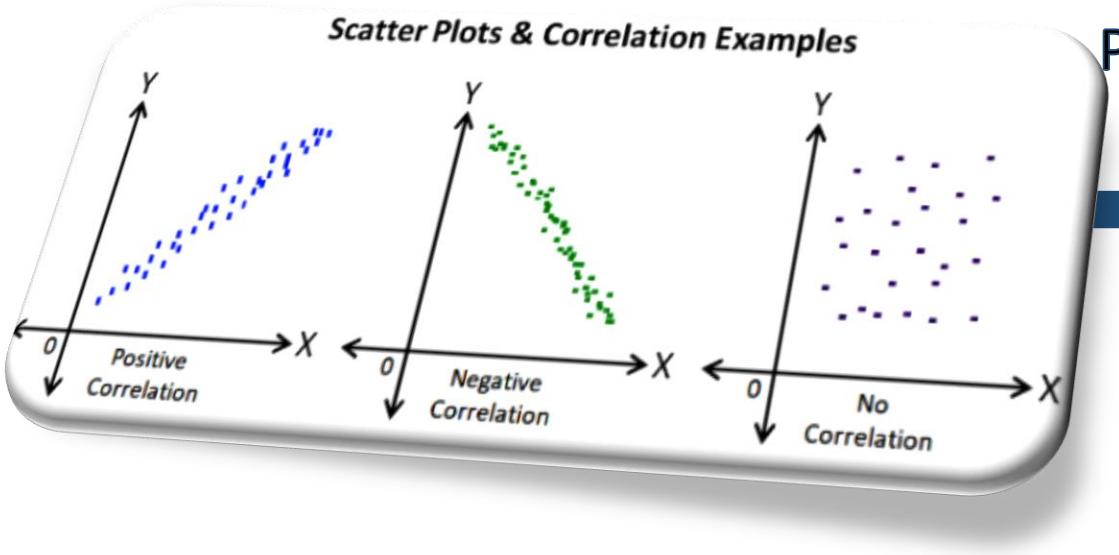
Introduction to Probability and Statistics

Topic (3): “Describing Bivariate Data”

Dr. Heba Ayyoub

Philadelphia University



Topic (3): “Describing Bivariate Data”

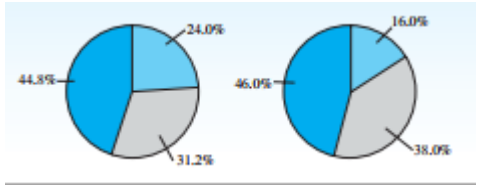
1 Bivariate data

Data consists of observations for two variables on the same experimental unit.

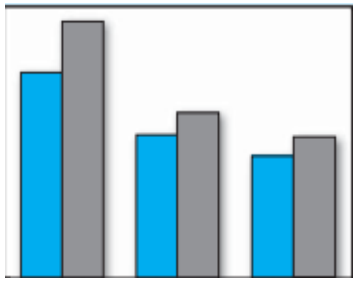
2 Graphs for Qualitative variable

Graphs for bivariate data:

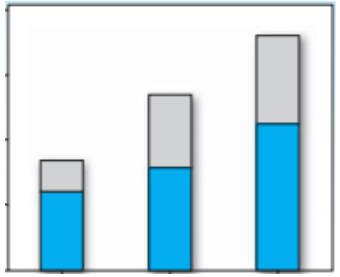
1) Side-by-side pie charts.



2) Side-by-side bar charts.



3) Stacked bar chart.



3 Scatterplots for Two Quantitative Variables

- Scatter plots show relationships between two variables.
- Each point represents one observation.
- Used to identify correlations and trends.

Example (1): Draw a scatterplots for the following bivariate data.

x	y
17	18
9	15
18	17
14	15
7	9

4 Numerical Measures for Quantitative Bivariate Data

❖ Correlation coefficient

The Correlation coefficient is used to measure the strength of the relationship between x and y.

$$r = \frac{S_{xy}}{S_x S_y}, \qquad -1 \le r \le 1$$

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

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

$$S_x = \sqrt{S_x^2}$$

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

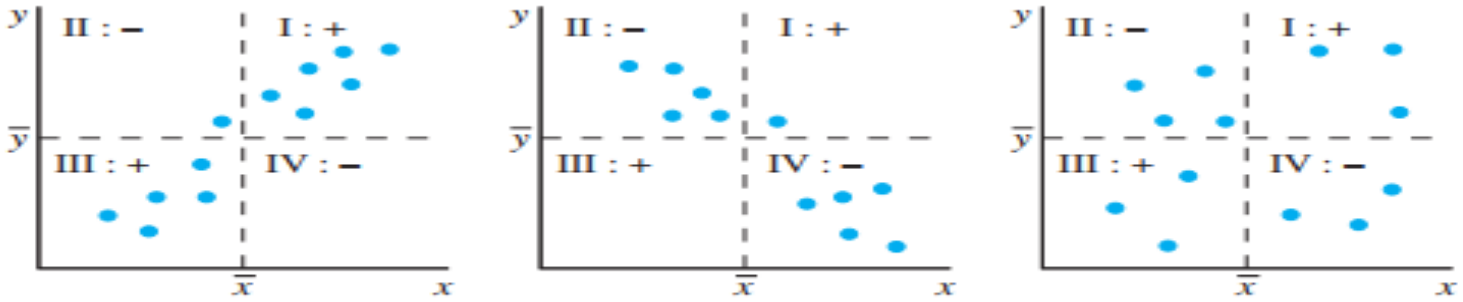
$$S_y = \sqrt{S_y^2}$$

S_{xy}: Covariance between x and y.

S_x^2: Variance of x.

S_y^2: Variance of y.

Note: $\sum_{i=1}^{n} x_i y_i \neq \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} x_i^2 \neq (\sum_{i=1}^{n} x_i)^2$.



(a) Positive pattern (b) Negative pattern (c) No pattern

- $r = 0$ means no linear relationship between the two variables x and y.
- r near zero means a weak linear relationship between x and y.
- r close to -1 or 1 means a strong linear relationship between x and y.

The sign of r provides important information about the direction of association.

- If r is positive, then as x increases, y increases linearly.
- If r is negative, then as x increases, y decreases linearly.

Example (2): Describe the relationship for the following correlations:

1) $r = 0.9$
2) $r = -0.8$
3) $r = 0.4$

❖ Regression

Best fitting line: is a straight line that most closely approximates the relationship between two variables in a set of data points.

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

where

x: Independent variable (the input value).

y: Dependent variable (the value you're predicting).

$a = \bar{y} - b\bar{x}$, a: intercept (the value of y when $x = 0$).

$b = \frac{S_{xy}}{S_x^2} = r\frac{S_y}{S_x}$, b: Slope (the rate of change in y for each unit change in x).

- The best-fitting line helps to make predictions, understand the trend of the data, and summarize the relationship between the variables.

Example (3): The following table represents the grades of 12 students in the first exam (x) and the second exam (y) in introduction to probability and statistics course.

#	x_i	y_i			
1	18	20			
2	14	11			
3	10	14			
4	15	16			
5	7	10			
6	12	10			
7	13	17			
8	8	11			
9	9	12			
10	17	20			
11	15	18			
12	12	12			

1) Draw a scatterplot and describe the relation between x and y?

2) Calculate the correlation coefficient and describe the relation between x and y.

3) Obtain the equation of the best fitting line.

4) If the grade of a student in first exam is 13. Predict (estimate) his grade in the second exam?

5) Find the amount of error in the prediction of the best fitting line in part (4)?

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?

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

d. Obtain the equation of the best fitting line.

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

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

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

1) Calculate the correlation coefficient and interpret the result.

2) Find the best fitting line.

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

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

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.

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

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

Exercise (6): Find the equation of the regression line that best fits the following data (best fitting line). Then predict the value of y when x = 2.5.

x	y
0	1
1	5
2	3
3	9
4	7

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

x	y
8	1
1	4
4	2
7	6
5	3

Exercise (8): True or False

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

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