

Module 5

Data Summary of Two Variables

Module Learning Outcomes

- Identify the use (or the role) of the two variables in a study.
- Provide a description of the scatter diagram.
- Provide a description of the correlation coefficient.
- Identify and apply properties of correlation coefficient.
- Compute table percentage, column percentage and row percentage using the two-way table.
- Draw a side-by-side bar graph (manually and using Excel) and provide a description of it.
- Calculate the expected frequency and provide an interpretation of its value.
- Calculate the chi-square statistic and provide a description of it.
- Use the most appropriate graphs to summarize any two-variate data.

5.1 Use of Variables

- In this unit, we investigate the relationship between two variables.
- Because of having two variables, we will need to know their **use** (or their **role**) in the study or analysis.
- In particular, we want to know which variable is used as the **dependent variable** (or **response variable**) and which one is used as the **independent variable** (or **explanatory variable**).
- Note: See Module #1 notes for reference.
- **MLO: Identify the use (or the role) of the two variables in a study.**

Data Summary with Two Variables

- The concept of data summary does not change here. We will still use a graph and a statistic.
- In this course, we will only study two combinations: A) two numerical variables, and B) two categorical variables.
- For two numerical variables, we have (1) scatter diagram and (2) correlation coefficient.
- For two categorical variables, we have (3) side-by-side bar graph and (4) χ^2 -statistic. More on Module #14.
- What if you only have one of each? Please read the last section of Module #14 as well.

5.2 Scatterplot & its Description

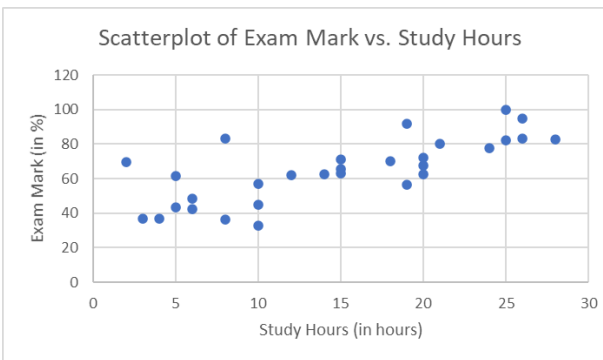
- The **scatterplot** (or **scatter diagram**) is the graph displaying the relation between the two numerical variables.
- By convention, the dependent variable is also called Y-variable (which goes to the y-axis of a plot) the independent variable is called X-variable (which goes to the x-axis).
- Each dot (or points) on the graph denotes a subject in the sample.
- There are four areas when describing scatterplot: Direction, Strength, Outliers, and Form.

Direction of the Relation

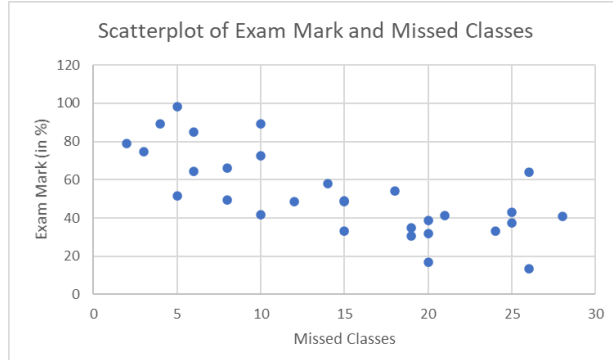
- We want to see if there is an apparent direction that the dots are concentrating.
- If dots are concentrating or spreading from lower left to upper right, we say that the relation has a positive direction.
- If dots are concentrating or spreading from upper left to lower right, we say that the relation has a negative direction.

DANA 4800 Notes: Module 5 – Data Summary of Two Variables

Positive Relation



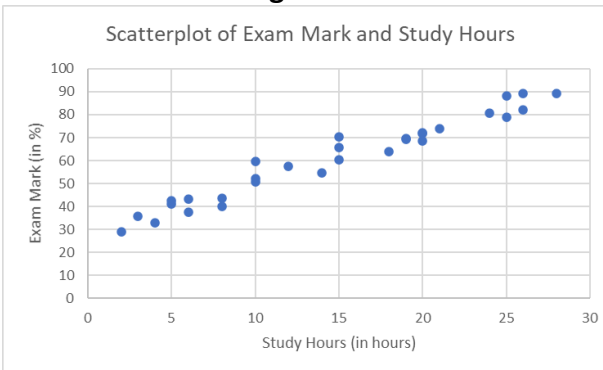
Negative Relation



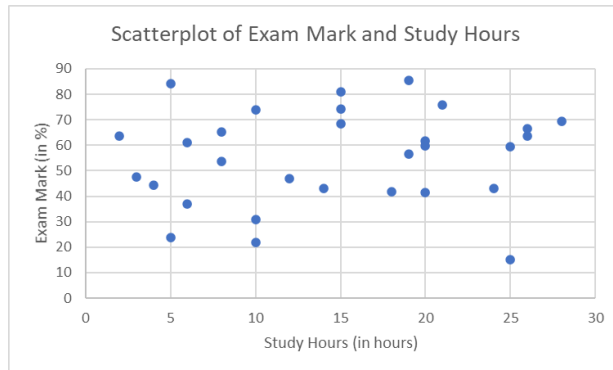
Strength of the Relation

- We want to get a basic concept of the strength by checking if dots are close to each other vertically.
- If dots are close (relatively) in the vertical direction, it has a (relatively) strong relation.
- If dots are far away (relatively) from each other in the vertical direction, it has a (relatively) weak relation.
- Note: A “stronger strength” or a strong relation in general gives you more “power” in making prediction (more later). You get a better prediction (i.e. more accurate) when the relation is stronger because less error is involved.

Strong Relation



Weak Relation

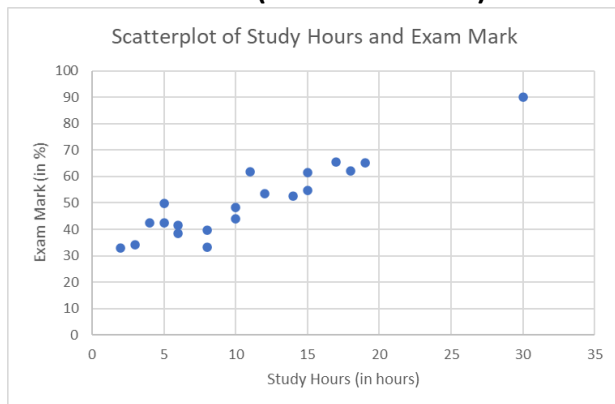


Outliers

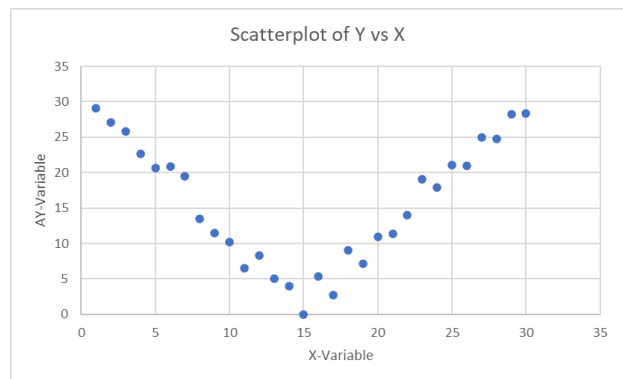
- We want to see if there are one or more dots that are located significantly far away (in any direction) from the rest of the dots.
- Unlike the univariate case (where we could use IQR method), we do not have any formula helping us determining outliers in two-dimensional case. But we get better doing so with experience.
- Note: When there are outliers in the scatter diagram, it is a common practice to provide two sets of description of it: one with the outliers included and another one without.

DANA 4800 Notes: Module 5 – Data Summary of Two Variables

Outliers (But Linear Form)



Curvi-linear Form



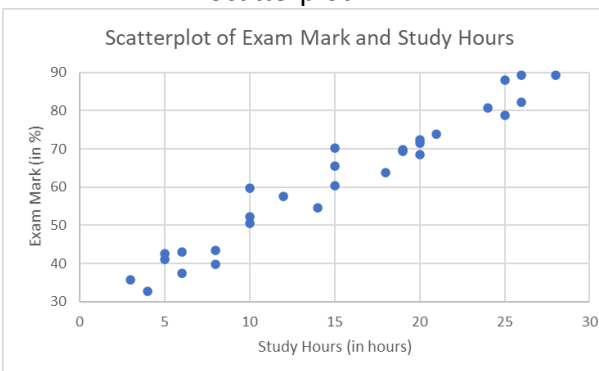
Form of the Relation

- Once we have determined that there are not any obvious outliers, we should now focus on the overall pattern. The two main forms you will see is a **linear relation** or a **curvilinear relation**.
- MLO: Provide a description of the scatter diagram.
- MLO: Use Excel to produce necessary outputs.

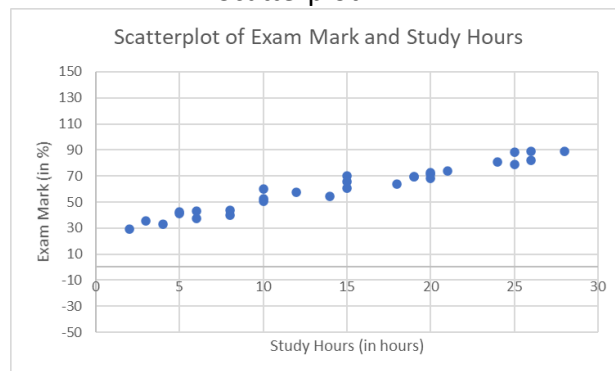
Pitfalls with Scatterplot

- Scatterplot could be very deceiving!
- Look at the two scatterplots below. They are from the same data set, but the y-axis on the right has been “compressed”. Because the y-axis (or the vertical axis) has been compressed, the absolute scale is larger. Hence, it looks as though points are closer to each other on the right than it is on the left. As a result, the right-hand scatterplot seems to show a stronger relation between the two variables.
- Because of this deceiving nature of scatterplot, it is not recommended to look at scatterplots alone (without also looking at some summary statistics).

Scatterplot #1



Scatterplot #2



5.3 Sample Correlation Coefficient

- The (sample) **correlation coefficient** (denoted as r) is the statistic used to show the linear correlation between two numerical variables.
- It is sometimes called the **coefficient of correlation**, or simply **correlation**.
- It is the only statistic to summarize two numerical variables in this course.
- Its formula is shown here for reference:

$$r = \frac{1}{n-1} \sum \left(\frac{X - \bar{X}}{s_X} \right) \left(\frac{Y - \bar{Y}}{s_Y} \right)$$

- The calculation of r is not required in this course and its value is always given.
- But make sure you understand its properties and what pitfalls to look for.
- The description of the correlation coefficient (note that it is a statistic) is consisted of the same three components: 1) **a single value**, from the 2) **sample** between the 3) **two variables**.
- **MLO: Provide a description of the correlation coefficient.**

5.4 Properties of Correlation Coefficient

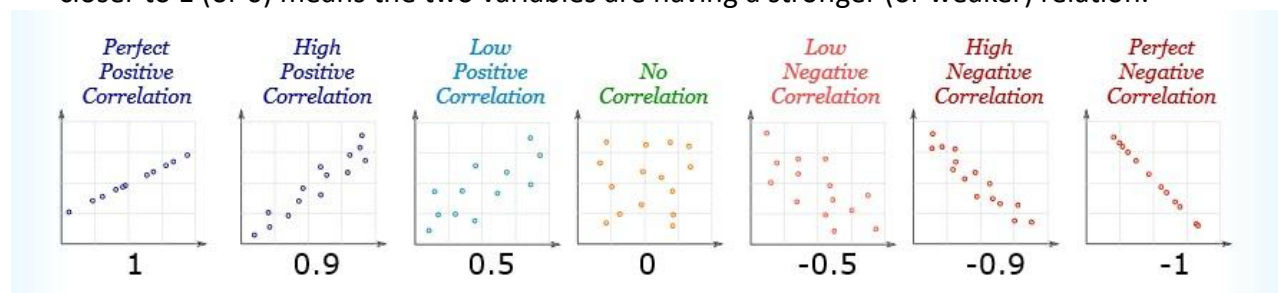
1. Possible values of correlation coefficient.

Values are always between -1 and +1, i.e. $-1.00 \leq r \leq +1.00$.

2. The **sign** and **magnitude** refer to different things.

The “sign” of the correlation coefficient tells you about the direction of the relation – positive (or negative) sign means the two variables are having a positive (or negative) relation.

The “magnitude” of the correlation coefficient tells you about the strength – magnitude closer to 1 (or 0) means the two variables are having a stronger (or weaker) relation.



For example, $r_1 = -0.95$ is a stronger relation than $r_2 = +0.80$.

3. Correlation coefficient has no units.
No matter what units we use for measurement (centimetres or inches), as long as the conversion is done properly, we would always get the same value of r .
4. It is all about linear relation or linear correlation.

DANA 4800 Notes: Module 5 – Data Summary of Two Variables

Correlation coefficient having a value of zero (or about zero) does not mean that there is no relation at all. It simply means that there is no linear relation, or no linear correlation.

- Correlation coefficient is extremely sensitive to outliers.

Important note: We never use the correlation coefficient alone and without referring to the scatter diagram first.

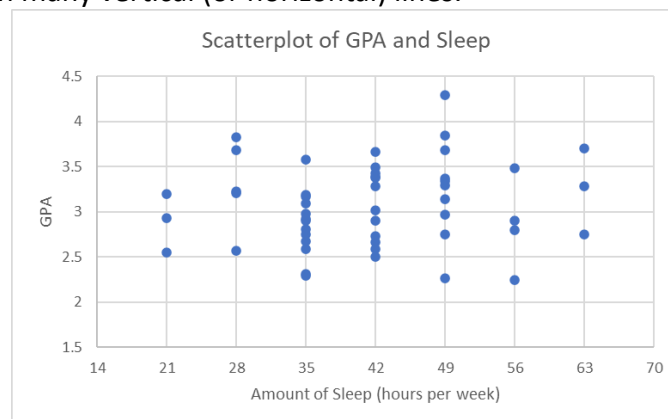
- Correlation does not imply causation.

Even when two numerical variables have a strong correlation, it does not always mean that “one causes the other”. In most cases, they are both influenced by a separate variable called the **lurking variable**.

For example, the number of shark attacks (on swimmers) on some beach is (positively) strongly correlated with the amount of ice-cream sales near the beach area. Does it make sense to conclude “the number shark attack has an effect on the increase of ice-cream sales” or “higher ice-cream sales leads to more shark attacks”?

- Correlation coefficient is only used for two numerical variables.

Keep in mind about rule #2 in determining the type of variable. If you do not have enough distinct values, the scatterplot would look very “discrete” – meaning that there would be points gathering in many vertical (or horizontal) lines.



Warning: For small sample size, it may not be an issue. But for larger data set, you could imagine that a single dot could represent multiple data points (because of the discrete nature of the variable). As a result, we cannot see what exactly the relation is by looking at the scatterplot.

- MLO: Identify and apply properties of correlation coefficient.**

5.5 Two-Way Table

- Before we can analyze the data, we first need to organize it in the **two-way table**.
- Suppose the independent variable has c classes (" c " for column) and the dependent variable has r classes (" r " for row).
- The **two-way table** is simply a table showing the number of subjects (or frequency) in each of the $r \times c$ **cells** or combinations.
- Terminology of the table includes **observed frequencies**, **column variable** and **column totals**, **row variable** and **row totals**, and the **overall total** (or sample size).
- Note: The convention is that the independent variable is usually put on the column variable and the dependent variable is put as the row variable.

Example

- Suppose we want to know if there is any difference between lunching behaviour among students in different PDD Terms?
- A random sample of 100 Langara PDD students was drawn and they were asked which term they are in (Term 1, 2, 3, or 4) and if they regularly lunch at Langara cafeteria (Yes, No). The results are summarized in the following two-way table.
- Because the variable Lunching at Langara is the dependent variable, it is the row variable. Hence, the numbers 42 and 58 are the **row totals**.
- In addition, the four numbers 34, 26, 26, and 14 are the **column totals** and 100 is the **overall total (or the sample size)**. The eight numbers inside the two-way table are the **observed frequencies**.

		Term 1	Term 2	Term 3	Term 4	
Lunching at Langara	Yes	12	8	8	4	42
	No	22	8	18	10	58
		34	26	26	14	100

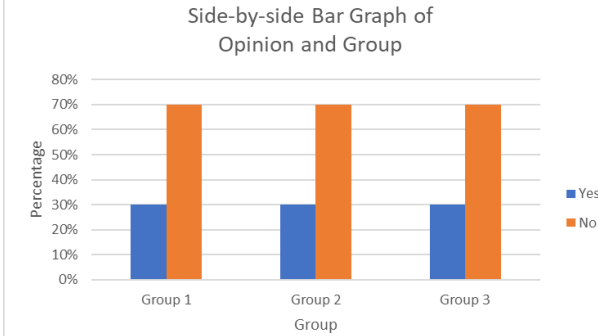
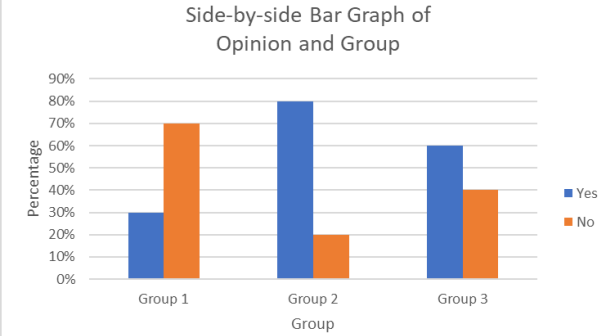
- MLO: Compute table percentage, column percentage and row percentage using the two-way table.

5.6 Side-by-Side Bar Graph

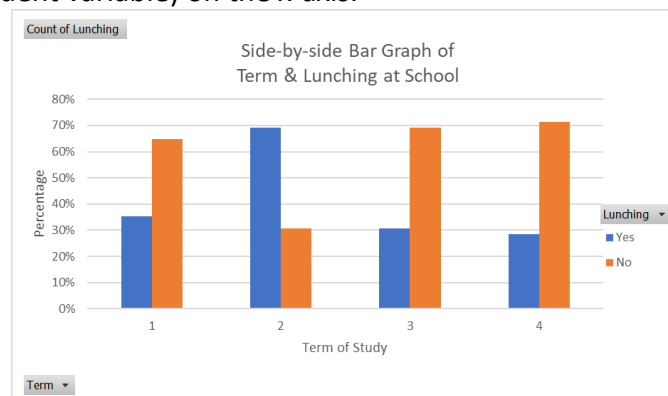
- The **side-by-side bar graph** is the graph displaying the relation between the two categorical variables.
- Observed frequencies from the two-way table are used to make this plot.
- There are two ways to present this graph: one with independent variable on the x-axis and the other with dependent variable on the x-axis.
- In this course, we will only put the independent variable on the x-axis.
- MLO: Draw a side-by-side bar graph (manually and using Excel).

Description of the Side-by-Side Bar Graph

- After making the plot, we always want to know what the relation between the two categorical variables is.
- When describing the side-by-side bar graph, we want to see if the groups of bars are similar or groups of bars look similar.
- The following shows two opposite looks.

	
<p>Description:</p> <p>The 3 groups of bars look alike, or the 3 groups of bars are similar.</p>	<p>Description:</p> <p>The 3 groups of bars are not identical, or the 3 groups of bars are not similar.</p>
<p>Conclusion:</p> <p>The two variables (opinion and group) are independent to each other.</p>	<p>Conclusion:</p> <p>The two variables (opinion and group) and not independent to each other.</p>

- Following the above example, here is the corresponding side-by-side bar graph with the Term (an independent variable) on the x-axis.



Here, the four groups of bars do not look alike. Hence, the two variables are not independent to each other.

- **MLO: Provide a description of the given side-by-side bar graph.**

5.7 Expected Frequency

- The **expected frequency** (denoted as e) can be calculated by

$$e = \frac{(\text{row total}) \times (\text{column total})}{\text{overall total}}$$
- It is like the “standard” in each cell and they are usually placed inside parentheses in the two-way table.
- It is typically interpreted as the number that is expected to be in a cell of the two-way table, when the two (categorical) variables are assumed to be independent to each other.
- Note: The results of the chi-square analysis will not be accurate when **one or more of the expected frequencies are under five.**
- When we have one or more cells that have expected frequency less than five, an *ad hoc* method will be used to combine adjacent rows or columns to form a smaller two-way table. If it happens in your project, please feel free to discuss it with me.
- Following the example from above, here is the two-way table with expected frequencies. Note that it is the convention that expected frequencies are put inside parentheses.

		Term 1	Term 2	Term 3	Term 4	
Lunching at Langara	Yes	12 (14.3)	8 (10.9)	8 (10.9)	4 (5.9)	42
	No	22 (19.7)	8 (15.1)	18 (15.1)	10 (8.1)	58
		34	26	26	14	100

Note that all expected frequencies are bigger than 5 here. So, the approximation is good enough.

- MLO: Calculate the expected frequency and provide an interpretation of its value.**

5.8 Chi-Square Statistic

- The **chi-square statistic** (denoted as χ^2) is the statistic used to learn about the relationship between the two categorical variables. In other words, how “independent” they are.
- It is calculated by

$$\chi^2 = \sum \frac{(f - e)^2}{e}$$

where f = observed frequencies and e = expected frequencies.
- Small values of χ^2 (the extreme is zero) mean the two categorical variables are independent to each other.
- In contrast, large values of χ^2 (no limit) mean the two categorical variables are not independent to each other.
- The description of the chi-square statistic has three components: 1) **a single value**, from the 2) **sample** between the 3) **two variables**.
- MLO: Calculate the chi-square statistic and provide a description of it.**

5.9 Other Graphs

- Here, we will look at how the dependent (numerical) variable is related to three different kinds of variables.
- 1) **Scatterplot**: a numerical dependent variable against a numerical independent variable.
- It is used when we want to see how the two variables are correlated.
- For example, we want to see how price (dependent variable) of used cars is related to the mileage (independent variable).
- 2) **Time Series Plot**: a numerical dependent variable against a special type of numerical independent variable – time.
- It is used when we want to see how the numerical variable varies with time.
- For example, we want to see how the regular gas price (dependent variable) varies with time in days (independent variable).
- 3) **Side-by-side Boxplot**: a numerical dependent variable against a categorical independent variable.
- It is used when we want to see how the numerical variable varies across different categories/groups.
- For example, we want to see how the SAT (Scholastic Aptitude Test) score (dependent variable) varies across different countries in North America (Canada, U.S. and Mexico).
- **MLO: Use the most appropriate graphs to summarize any two-variate data.**