Module 3
Data Summary for
One Categorical Variable

Module Learning Outcomes

- MLO: Choose the most appropriate graph to summarize one categorical variable and provide a description of the graph.
- MLO: Use Excel to produce bar graphs and pie charts.
- MLO: Calculate sample proportions in both decimals and percentages.
- MLO: Provide a description of key parameters and key statistics.
- MLO: Perform Data Summary for one variable by using only one graph and one statistic (or one set of statistics) and justify the choice.

3.1 General Data Summary

- The terms **Data Summary** and **EDA (Exploratory Data Analysis)** are nearly interchangeable in real life. Both refer to a set of procedures that let you understand the data set better, by looking at two key components: **Graphs** and numbers (as in **Statistics**).
- In this course, you are expected to choose only one graph out of all possibilities and be ready to defend your choice.
- You are also expected to choose only one statistic (or one set of statistics, in numerical case; more later) and be ready to provide solid reasons to support the choice.

Type of Variables – Revisit

- Being able to differentiate numerical variables from categorical variables is the key to do well in data summary or analysis. Please refer to Module #1 for details.

3.2 Summarizing a Single Categorical Variable Using Graphs

- There are only two graphs for one categorical in this course: 1) **bar graph**, and 2) **pie chart**.
- Please review the scale of measurement (in Module #1) before reading ahead.
- When the data is in ordinal scale of measurement, bar graphs are preferred over pie charts. It is because bar graphs shows the order, if any, in a linear (from left to right) fashion.
- When the data is in nominal scale of measurement, either graph would fit the bill.
- There is not a right or wrong way of describing these two graphs, but we usually want to describe any anomaly or something surprising.
- When describing bar graphs (ordinal data), we may also want to look at the number of classes and see if there is any trend (like increasing or decreasing trend) from left to right.
- Note: Only vertical bar graphs (i.e. bar graphs with vertical bars) will be used in this course.
- See this funny 11-second clip on YouTube about the two graphs.
- MLO: Choose the most appropriate graph to summarize one categorical variable and provide a description of the graph.
- MLO: Use Excel to produce bar graphs and pie charts.

3.3 Misleading Graphs

- "There are lies, damn lies and there are statistics." The origin of this is unknown, but most believed that it was from Benjamin Disraeli, a former British Prime Minister.
- Many people think that (the discipline of) statistics is used to manipulate data or distort facts.
- To avoid all that (or become a good person using statistics), you will need to learn how statistics are poorly presented first. Just like police officers must think like criminals before they could get them.

- In other words, you are not taught how to play around with data in this course, but to be careful when reading other people's graphs.
- There are many ways to "twist facts" in graphs, but the most prominent two are 1) scaling on the y-axis and 2) scaling on the "bars" in bar graphs or the "pies" in pie charts.
- Believe it or not, most misleading graphs are categorical graphs – bar graphs or pie charts.
- The best graph, in my opinion, should be able to give an instant pictorial idea about the data and is "no frills". And you should not need to ask viewer to lean in and read the details – that should be left to the written documentation or oral presentation.

3.4 Summarizing a Single Categorical Variable Using Statistics

- The main statistic for summarizing a single categorical variable is **sample proportion** (or simply proportion), although percentage is widely acceptable too.
- The calculation of **sample proportion** ($\bar{p}$, says "p bar") is $\bar{p} = X/n$, where X is the number of subjects having certain feature and n is the sample size.
- The calculation of **sample percentage** (not an official name) is just multiplying sample proportion by 100%.
- Note: Make sure you are comfortable with the conversion between decimals (or proportions) and percentages.
- MLO: Calculate sample proportions in both decimals and percentages.

3.5 Parameters vs. Statistics

- A **parameter** is defined as a number that describes a variable (or multiple variables later) of the **population**.
- Likewise, a **statistic** is defined as a number that describes (or summarizes) a variable (or multiple variables later) of the **sample**.
- A very simple mnemonic (or an easy way to memorize things) is P vs. S (**P**arameter from **P**opulation and **S**tatistic from **S**ample).
- Values of parameters are always unknown to us, but we should be able to describe it in words.
- The description of parameter has three main parts: (1) a single value, (2) of a variable, (3) from the population.
- Values of statistics vary from sample to sample and we can describe it in words too.
- The description of statistic also has three main parts: (1) a single value, (2) of the variable, (3) from the sample.
- One basic use of (the discipline of) statistics is to use statistics to estimate unknown parameters.
- MLO: Provide a description of key parameters and key statistics.

3.6 Data Summary for One Categorical Variable

- **Graph**: Choose between bar graph and pie chart.
- Criteria: Scale of Measurement and/or number of classes
- **Statistic**: There is only one statistic with two different presentation – proportions vs. percentages.
- By and large, percentages are used in speaking and writing. But we will use proportions (or decimals) when using formulas.
- MLO: Perform Data Summary for one variable by using only one graph and one statistic (or one set of statistics) and justify the choice.