Chapter 3 – Relationships between Categorical Variables

Introduction: An important field of exploration when analyzing data is the study of relationships between variables. A lot of thought has been put into determining which variables have relationships and the scope of that relationship. Is a person's diet related to having high blood pressure? Is the city a person lives in related to whether or not they have tuberculosis? Is being in a car accident related to texting while driving? These are all important questions that statisticians, data analysts and data scientists explore.

We can study relationships between two categorical variables like texting while driving and having a car accident. We can also study relationships between two quantitative variables like the weight of a person and their blood pressure. A third relationship we can study is the relationship between a categorical variable and a quantitative variable. For example, we can study the relationship between the type of job you have and your annual income. In this chapter, we will begin to explore the relationships between two categorical variables.

Remember, statistics is a deep well of mathematics and knowledge learned by years of study. There are much more advanced techniques for studying relationships, but we will be focusing on a basic introduction to the topic. You will find that a good understanding of this chapter will help tremendously when you go on to the more advanced techniques later on. For example, I find my advanced statistics students do not understand the Chi-Square distribution because they lack the foundational understanding of contingency tables and analyzing the differences between categories.

Note on Terminology: *When studying relationships between variables you will hear different words used to describe the relationship. The most common are "relationship", "association", or "correlation". "Correlation" is often used for describe a relationship between two quantitative variables, while "relationship" and "association" are used for two categorical variables or for a categorical - quantitative relationship study.*

In this chapter, we will be using the terms "relationship" and "association".

Note on Causation: *One of the most famous statements in statistics is that "correlation is not causation". Proving that one thing causes another is a much more complex kind of study and involves controlling confounding variables and experimental design. The main thing to remember is that just because there is a relationship, that does not prove causation. There may be many other factors involved.*
