

Chapter 3 Notes (Displaying & Describing Categorical Data)

- Thus far, we have practiced identifying the key components of a data set, such as the observational units, the variables measured, and what *types* of variables these are.
- Specifically, we learned to distinguish between **categorical** and **quantitative** variables. Chapter 3 considers a variety of techniques for displaying data from categorical variables.
- An initial exploration of *any* data is often referred to as an **exploratory data analysis (EDA)**. The first thing you should always do as part of an EDA is DRAW A PICTURE!!!

How to Summarize Data from ONE Categorical Variable

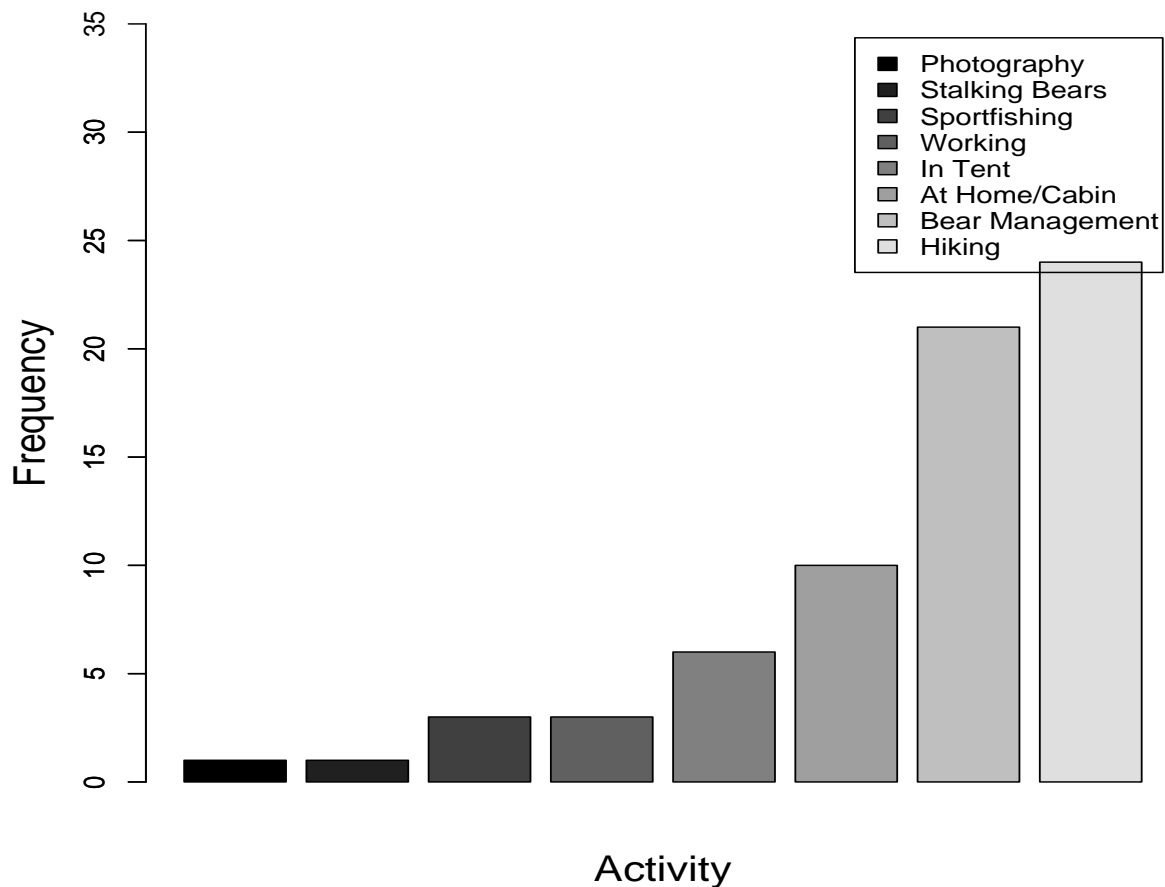
The most basic way to represent data from a categorical variable is through the use of a **frequency table** (table of counts) or **relative frequency table** (table of percentages or relative frequencies).

Example: The following data were taken from a paper titled “Efficacy of Bear Deterrent Spray in Alaska” published in the Journal of Wildlife Management in 2008. They looked at a sample of bear spray incidents occurring in Alaska between 1985 and 2006. One variable they include in their paper is the primary activity of persons involved in the bear spray incident. We can use SPSS to construct a frequency and relative frequency table for this activity variable. Record these tables below.

- What type of information is acquired from either of these tables?
The relative frequency with which the categories occur is known as the **distribution** of the variable.
- These tables provide a simple summary of the 67 data values, but how might this information be displayed in a graph?

Bar Graphs: A simple visual display for the distribution of a categorical variable is a bar graph. To illustrate how a bar graph is drawn, consider again the data on bear spray incidents. A bar graph representing these data is given at the top of the next page.

Bear Activity During Bear Spray Incident

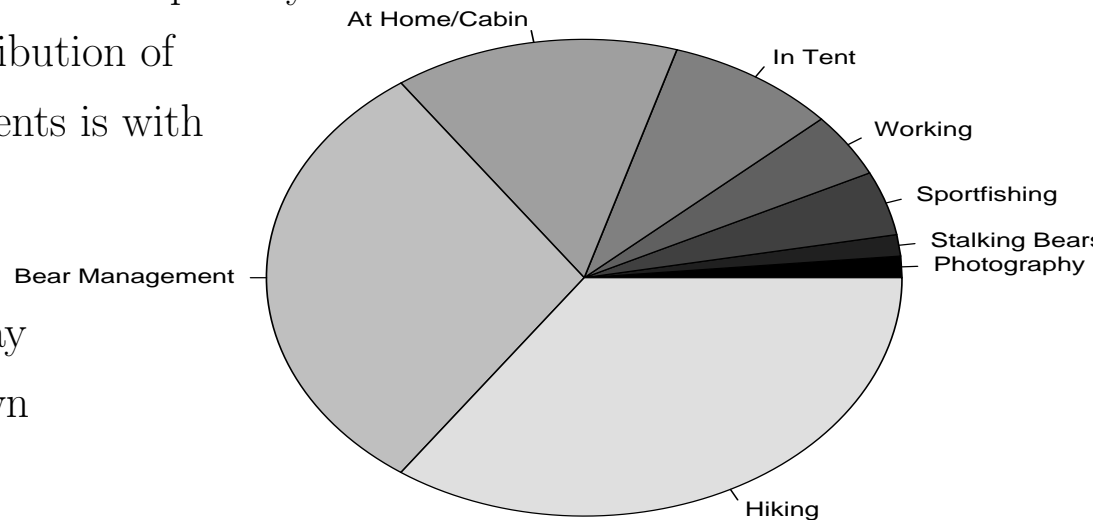


Notes:

- The bars are all the same width so that we need only examine the *heights* of the bars to compare the relative frequencies each incident category.
- The bars have space between them to indicate that they are truly separate categories that can be given in any order.
- In this example, there is not any natural ordering within the categories. However, it looks nice to place the bars in order of increasing or decreasing frequency.
- The bar graph shown is called a frequency bar graph. How would the bar graph change if we used percentages (relative frequencies) instead of the frequencies on the y-axis?

Pie Charts: A second simple way to display the distribution of bear spray incidents is with a pie chart.

A pie chart for the bear spray incidents is shown to the right.



Notes:

Bear Activity During Bear Spray Incident

- Does the pie chart give you better, worse, or equivalent information as the bar chart?
- When would you likely prefer a bar chart to a pie chart?

How to Examine TWO Categorical Variables and their Relationship

- Much more interesting than examining the distribution of a single categorical variable is examining the relationship between 2 such variables. For example, we might want to characterize any relationship between:
 1. the passenger class and survival on the Titanic,
 2. obesity level and physical activity level, or
 3. gender and university admissions.
- When we want to look at two categorical variables simultaneously, we can expand the idea of a frequency table where the

rows represent one variable and the columns the second variable. This yields a “two-way” table known as a **contingency table**, as shown in the example to follow. All analysis done on categorical variables involves nothing more than computing proportions and summing values.

Example: A 2000 study published in the journal *Circulation* examined the relationship between anger and coronary heart disease. Researchers used 8,474 total people, all of whom had normal blood pressure at the onset of the study. All participants took the Spielberger Trait Anger Scale test to measure proneness to anger and were followed for 4 years. The two-way table below contains counts of the number of participants who fell into each coronary heart disease (CHD)/anger category listed on the sides of the table. Totals for each level of CHD and anger level are given in the last column and row respectively. So, for example, of the 8,474 people in the study, 110 had CHD and were classified as having moderate anger.

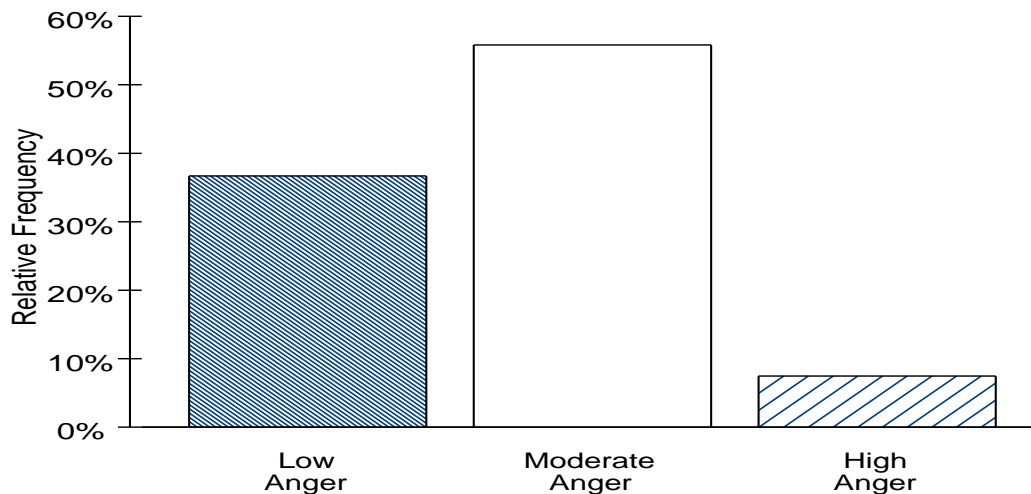
Coronary Heart Disease	Anger Level			Total
	Low	Moderate	High	
Presence	53	110	27	190
Absence	3057	4621	606	8284
Total	3110	4731	633	8474

- Virtually all aspects of the relationship between the variables Coronary Heart Disease and Anger Level can be found using proportions. To see this consider the following questions.

1. What proportion of those surveyed were classified as having low anger?

 2. What proportion of those surveyed were of moderate anger?

 3. It is simple to verify also that 7.47% of those surveyed are of high anger.
- These three percentages (36.70%, 55.83%, 7.47%) taken collectively comprise the **marginal distribution** of the variable Anger Level. Using the word “marginal” means we are only considering one of the variables (in this case: anger level), not the relationship between the two. So the marginal distribution gives the percentages of cases falling into each category of a single categorical variable.
 - Marginal distributions of categorical variables are best represented using bar graphs as was illustrated earlier. The marginal distribution for anger level is shown with a relative frequency bar graph at the top of the next page.



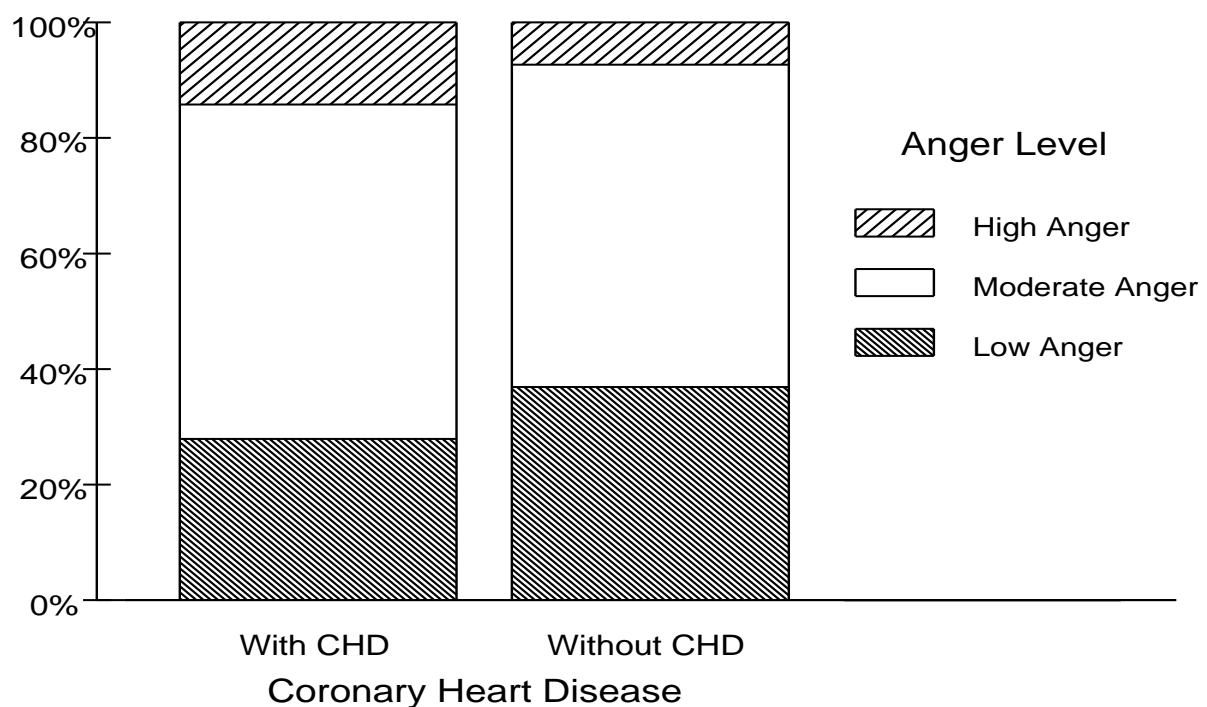
Relationships Between Categorical Variables: Consider now the following question: “What proportion *of those with coronary heart disease* are identified as having moderate anger?”

- The italicized part of this question is to draw your attention to the fact that the proportion sought is only relative to those with coronary heart disease. Since there are 190 people with coronary heart disease and 110 of them were identified as having moderate anger, then $110/190 = .5789$ or 57.89% of those with CHD are identified as having moderate anger.
- Suppose we wanted to find the proportion in each of the anger groups among those with coronary heart disease. These proportions are: .2789, .5789, and .1421 respectively. Because we are finding these proportions under the *condition* that the respondents have coronary heart disease, these values comprise the **conditional distribution** of the anger level variable for those with coronary heart disease.

Problem:

- (a) Find the conditional distribution of the anger level variable for those without coronary heart disease.

- Probably the best way to visually represent these conditional distributions between categorical variables on one graph is through what is known as a **segmented bar graph**. A segmented bar graph for these data is shown below.



Does there appear to be a relationship between anger level and presence of coronary heart disease? Describe this relationship in a sentence or two.