Statistical Thinking for the 21st Century

Chapter 3

Summarizing data

I mentioned in the Introduction that one of the big discoveries of statistics is the idea that we can better understand the world by throwing away information, and that's exactly what we are doing when we summarize a dataset. In this Chapter we will discuss why and how to summarize data.

3.1 Why summarize data?

When we summarize data, we are necessarily throwing away information, and one might plausibly object to this. As an example, let's go back to the PURE study that we discussed in Chapter 1. Are we not supposed to believe that all of the details about each individual matter, beyond those that are summarized in the dataset? What about the specific details of how the data were collected, such as the time of day or the mood of the participant? All of these details are lost when we summarize the data.

We summarize data in general because it provides us with a way to *generalize* - that is, to make general statements that extend beyond specific observations. The importance of generalization was highlighted by the writer Jorge Luis Borges in his short story "Funes the Memorious", which describes an individual who loses the ability to forget. Borges focuses in on the relation between generalization (i.e. throwing away data) and thinking: "To think is to forget a difference, to generalize, to abstract. In the overly replete world of Funes,

Figure 3.1: A Sumerian tablet from the Louvre, showing a sales contract for a house and field. Public domain, via Wikimedia Commons.

there were nothing but details."

Psychologists have long studied all of the ways in which generalization is central to thinking. One example is categorization: We are able to easily recognize different examples of the category of "birds" even though the individual examples may be very different in their surface features (such as an ostrich, a robin, and a chicken). Importantly, generalization lets us make predictions about these individuals – in the case of birds, we can predict that they can fly and eat worms, and that they probably can't drive a car or speak English. These predictions won't always be right, but they are often good enough to be useful in the world.

3.2 Summarizing data using tables

A simple way to summarize data is to generate a table representing counts of various types of observations. This type of table has been used for thousands of years (see Figure 3.1).

Let's look at some examples of the use of tables, again using the NHANES dataset.

Let's have a look at a simple variable, called "PhysActive" in the dataset. This variable contains one of three different values: "Yes" or "No" (indicating whether or not the person reports doing "moderate or vigorous-intensity sports, fitness or recreational activities"), or "NA" if the data are missing for that individual. There are different reasons that the data might be missing; for example, this question was not asked of children younger than 12 years of age, while in other cases an adult may have declined to answer the question during the interview.

3.2.1 Frequency distributions

Let's look at how many people fall into each of these categories. We will do this by selecting the variable of interest (PhysActive) from the NHANES dataset, grouping the data by the different values of the variable, and then counting how many values there are in each group:

PhysActive	AbsoluteFrequency
No	2473
Yes	2972
NA	1334

This table shows the frequencies of each of the different values; there were 2473 individuals who responded "No" to the question, 2972 who responded "Yes", and 1334 for whom no response was given. We call this a *frequency distribution* because it tells us how frequent each of the possible values is within our sample.

This shows us the absolute frequency of the two responses, for everyone who actually gave a response. We can see from this that there are more people saying "Yes" than "No", but it can be hard to tell from absolute numbers how big the difference is. For this reason, we often would rather present the data using *relative frequency*, which is obtained by dividing each frequency by the sum of all frequencies:

$$relative\ frequency_i = \frac{absolute\ frequency_i}{\sum_{j=1}^{N} absolute\ frequency_j}$$

Table 3.1: Absolute and relative frequencies and percentages for PhysActive variable

PhysActive	AbsoluteFrequency	RelativeFrequency	Percentage
No	2473	0.454	45%
Yes	2972	0.546	55%

The relative frequency provides a much easier way to see how big the imbalance is. We can also interpret the relative frequencies as percentages by multiplying them by 100. In this example, we will drop the NA values as well, since we would like to be able to interpret the relative frequencies of active versus inactive people.

Computing this frequency distribution lets us see that 45% (0.454) of the individuals in the NHANES sample said "No" and 55% (0.546) said "Yes."