THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

WEEK 2

2.2 CATEGORICAL VARIABLES by Chris Wild

Welcome. Today we're going to look at data on a categorical variable. This is a small part of our NHANES data set. The full data has 10,000 people and over 70 variables.

Whenever I get a new data set, I try to familiarise myself with what I've got before I do any real analysis. I look for things that might surprise me and might indicate errors in the data. One of the first things I do is to look at each variable in turn. We can't just scan down 10,000 rows of data. Large tables of data are incomprehensible. We need tools like graphs and summaries to convert them into things we can see and understand.

In Week 1, we made a distinction between numeric and categorical variables. This distinction is important because the type of a variable determines how we look at it. Numeric variables are those we can think of as measurements, like Age and Weight here. Categorical variables, like Gender, Age Decade, Race3, Education, and Marital Status give us group membership labels.

A categorical variable like Marital Status divides our people up into categories or groups. I see "Married" and "Widowed". Recall that NA is a code for missing. All we can do with group membership on its own is to count how many people fall into each group and compare these counts. This lets us answer questions like, "What are the most common categories of Marital Status?" Generally, we convert our counts into percentages so that we can compare groups of different sizes.

Next we will look at the data on race. In 2011, NHANES changed their race categories to separate out Asian. It had previously been included in "Other." The name of the new variable is Race3. These entries suggest that white will be the most common category.

Here is a table giving the counts and percentages for Race3. The percentages are plotted here, using a bar chart. The fact that white is by far the most common category jumps out at us from the graph. From the scale, we see that over 60% of our people have been classified White. The next most common category is Black, at about 12%, closely followed by Mexican, Hispanic, and so on.

The bar chart is our standard tool for plotting categorical data. You might ask, "Why bar charts in preference to pie charts or stacked bar charts?" Pies and stacked bars have the advantages of conveying the sense of being parts of a whole. However, they're not good for discovery because they're not good at letting you see changes and differences. And spotting changes is the lifeblood of discovery.

Let's go back to Marital Status. What we see first is differing bar heights. Higher bars indicate more people. With a little more effort, we can see that the Married bar is about 2 and 1/2 times as high as the Never Married bar. So about 2 and 1/2 times as many people reported their Marital Status as Married than as Never Married. Similarly, there about half as many people in the Divorced category as in the Never Married category, and so on. We can see that the difference in bar height between Divorced and Live with partner is fairly small compared to the height of either bar.

As soon as we've been educated to pay attention to these things, we can take in this sort of information very quickly and almost subconsciously. We can also use the scale to link these things with numbers. So Married looks like about 55%, Never Married about 20%, Divorced about 10%, and so on. We get useful impressions of what the data are saying much more quickly from the graph than from a table of numbers. We look first at the graph to get an impression of what's happening, and then consult the table if we want more accuracy.

When you looked at Marital Status, did you start to think, "I wonder how these patterns change with age, or with race, or region?" Or is that Never Married group inflated by classifying children as never married? These are places we want to go with our exploration and as quickly as possible. But first, we have to graduate from Boot Camp.

I'll leave you with these questions to remind you of the ideas we've just covered.

QUESTIONS

- What is the basic graph that we use for displaying the data for a categorical variable?
- Why do we use this in preference to pie charts and segmented bar charts?
- What features of the data does a bar chart highlight? (What sorts of things can we easily see?)