Building on Visualization Concepts

This week, we will be building on the previous concepts that we have discussed in this course for visualizing data. We are now going to put more of a focus on where data come from, and important concepts related to random sampling as a scientific tool for making inferences about larger populations. With a representative, well-designed random sample of units (people, households, businesses, etc.) from a well-defined target population of interest, we can make sound scientific conclusions about population features of interest (e.g., mean income), and we don't need a very large sample to do so! However, before we can understand why this is possible, we need to understand the idea of sampling variability, and the uncertainty associated with estimates computed using data collected from random samples. In order to understand sampling variability, we will need to introduce the concept of a sampling distribution, and this is where we are going to build on previous discussions of distributions and random variability.

A key distinction of this week's material from all previous weeks is that we will now be visualizing distributions of survey estimates based on many hypothetical random samples, rather than distributions of the values on a variable of interest for a given population. In reality, in any given study, we only get to work with one random sample of units. However, the important (and beautiful!) statistical concept of random sampling is that we can use that one sample to estimate features of the sampling distribution that would emerge if we selected multiple random samples using the same techniques. We don't actually need to draw many random samples, and then examine the distribution of estimates that would emerge after collecting data from every random sample; we just need to draw one, and use the information that we collect from the sample to estimate features of the sampling distribution. This provides us with a sense of the uncertainty in our estimate based on only one sample, and allows us to make conclusions about the value of our target parameter of interest (e.g., mean income) in the larger population.

In order to visualize the distribution of a variable of interest in some population, we need to collect data from as many units as possible in that population, and then plot the distribution of values (e.g., using a histogram). The beautiful property of random sampling is that we can estimate the features of the distribution of estimates that would emerge (again, the sampling distribution) if we selected many random samples using the same design, and we only need one sample to do so. We would not be able to plot (or even estimate!) the distribution of values on a variable of interest by only measuring one unit. This is the wonderful statistical luxury of designing good random samples. This will be the focus of our discussion this week.

Additional Deep-Dive Readings and Web Sites on Random Sampling and Sampling Distributions

- Kish, Leslie. (1965). Survey Sampling. Wiley.
- Lohr, Sharon. (1999). Sampling: Design and Analysis, Second Edition. Cengage Learning.
- Statistics How-To - Sampling Distribution: Definition, Types, Examples
- What is a Sampling Distribution?