Andrew Sousa

Mentee: Andrea Boskovic

STAT 499

10 December 2023

Making More Out of Less: The Usefulness of Spatial Sampling and Horvitz-Thompson Estimator

The world we live in is big. Very big. With over 8 billion people and almost 200 million square miles of globe, getting insight and responses from everyone in a population is essentially impossible. In many cases, this would be where samples would come in. Instead of asking all 4000 people in a city, why don't we just ask 400? From this, random sampling has become a staple of research and data collection, allowing us to make assumptions about population parameters, or a true description of the population, from our sample statistic. However, in the scope of analyzing regions, a true random sample may not always be the best method.

Take the United States for example, where we want to know how much of the U.S. supports one political party. If we randomly chose people around the country, we might get only two from California, and three from smaller states like Arkansas, giving us a sample that is a vastly inaccurate representation of how the U.S. leans politically. So how do we create that accurate sample? This question takes us to what I've spent the last quarter learning about: spatial sampling.

In spatial sampling, we divide our geographic region into sections based on a certain characteristic, then select an equal proportion of observational units in each sector, called probability proportional to size. Sections with larger values, say more people, will have more units in the sample, having an impact on the study that's more natural. By doing this, we create a sample that is more representative of the population, while retaining a random system that's crucial to statistical analysis. To see this in practice, let's take a look back at the United States example. If we redo our study with the new spatial sampling system, we would see a much higher number of Californians than before, as California is home to over 10 percent of the U.S. population. Now that our sample closely resembles the distribution of the population, we can make estimations about our population parameter far more accurately.

To make such estimations, we can use an equation called the Horvitz-Thompson Estimator. The HTE takes in two main components: the observed value of some response, and the probability of that individual being chosen, generally called the "weight" of an individual. The value and weight are multiplied together to create the contribution of one response, and the contribution of every individual in the sample are added together for an estimation of some population characteristic. We can use the HTE in almost every scenario, and can also be used no matter the data type, whether it's categorical or numerical.

The effectiveness of a study is made by its design, and crucial decisions in our world can be made incorrectly if that design is poorly made. Using spatial sampling is an excellent way to ensure that our sample is a quality one, and from there, the Horvitz-Thompson Estimator can take us home, using our data to make the actual estimation. With these tools in conjunction, we

can make more decisions with less data, helping bring our world together, no matter how large it may be.

Works Cited

Brus, Dick J. "Spatial Sampling with R." *Chapter 2 Introduction to Probability Sampling*, 3 Sept. 2023, dickbrus.github.io/SpatialSamplingwithR/IntroProbabilitySampling.html.

"U.S. and World Population Clock." *United States Census Bureau*, www.census.gov/popclock/. Accessed 20 Nov. 2023.