Exercise 14-1: SAMPLING METHODS

EXERCISE BACKGROUND

This exercise demonstrates the application of some of the most common probability sampling methods. It is based on a set of (randomly generated and fictitious) data on reading scores for children in an area where we are considering an intervention to increase literacy. We are able to collect the following data:

- *Student_ID*: This is a six-digit code that uniquely identifies each student. Note that to preserve anonymity, we have not used names or other identifiable information
- *School:* A code which uniquely identifies the school attended by each student
- *Sex:* The sex of the student (M or F)
- *Score*: The student's score on a reading test administered during the survey, on a scale of 0-100
- *HH_Income*: weekly income of the student's household

The data-set presented includes data from all students in our area of interest. However, in reality, we only have the budget available to collect data from a sample of students included on this list. We would then like to use the average test score for this sample to learn about the average for the whole population of 200 students. How would we go about selecting that sample?

SAMPLING METHOD 1

The simplest approach to probability sampling would be to use a **simple random sample**. To do this, we would assign each student a random number, and then select the students with the X highest random numbers (where X is our sample size) for data collection. In this way, everyone has an equal probability of being selected. To do this in excel, we use the function *=rand()* but remember to copy and paste the values of the random number, as the random number will change each time a cell in the spreadsheet is changed.

Questions:

1. We use the average reading score for our sample to estimate the average test score of the whole group of 200. Do you think our estimate of the average test score will be better or worse as we increase our sample size?
2. Do you see any potential disadvantages to this approach?

SAMPLING METHOD 2

Typically, our data collection budgets limit the amount of travel we can do to collect data. Since simple random sampling may require us to visit a different school or community for each student

in our sample, our transportation budget may be excessive. Another approach commonly used to reduce travel costs is a **clustered random sample**. To use this method, we would randomly select a set number of schools (say 5 or 10), and collect data from all of the students in that school. That way, if we want a sample size of 50 students, we only visit 5 different schools or communities (assuming there are 10 students in our age range of interest per school) whereas simple random sampling may have required us to go to up to 50 different schools or communities. To do this in excel, we use the function *=rand()* to randomly assign a number to each school. We can then interview all of the students from the schools with the 5 (or 10…or whatever our sample size) lowest random numbers.

Questions:

1. Do you think our estimate of the average test score will be closer to the population average if we base it on 5 or 10 clusters?
2. Do you see any potential disadvantages to this approach?

SAMPLING METHOD 3

If we are most interested in comparing scores of boys versus girls, we need to ensure our sample includes sufficient numbers of each. **Stratified random sampling** allows us to ensure a certain percentage of important sub-groups (e.g. girls) in our sample. To do this, we first split our sampling frame into groups or strata, in this case, M and F. We then assign random numbers to all, and within each group select the lowest numbers for our sample. But, how many of each do we choose?

There are two methods of stratified random sampling:

- Proportional: We select the number of each subgroup to mirror their prevalence in the population. So, if girls make up 40% of the student population, we will ensure 40% of our sample is made up of girls.
- Disproportional: However, if the group of interest is relatively small, proportional random sampling may still yield a very small sample from that group, so we may *oversample* from that group.

To do this in excel, we first sort by the stratification variable (sex), use the function *=rand()* to randomly assign a number to each student, and then select the appropriate numbers of girls and boys based on their random number.

Questions:

1. If we use disproportional stratified random sampling, will our sample still be representative of the population?