Lab Four: Randomization in Surveys and Causal Studies

STA 111: Probability & Statistical Inference

Lab Objective

To illustrate the benefits of random sampling in surveys and causal studies.

Lab Procedures

Part One: The benefits of randomization in surveys

In a survey, the sampled data should be representative of the target population. The simplest way to guarantee representative data is to collect data from randomly selected units in the population. We'll illustrate this using real data. Read in the data *"Agpop.txt"* from the course website by typing:

```
Agpop = read.table("http://www2.stat.duke.edu/~oma9/STA111_SummerI_2018/Labs/Agpop.txt",header=T)
```

This file is taken from the 2007 U.S. Census of Agriculture. It contains data on agricultural characteristics of all 3,078 counties in the United States. Variables include:

Variable	Description
state	State
statefips	FIPS code for the state
county	County
countycode	FIPS code for the county
acres2007	Number of acres devoted to farming in 2007
farms2007	Number of farms in 2007
large2007	Number of farms with more than 1,000 acres in 2007
small2007	Number of farms with fewer than 9 acres in 2007
.....2002	acres, farms, large, small for 2002
.....1997	acres, farms, large, small for 1997

For more information on the Census of Agriculture, including data from the census, you can visit the web site of the National Agricultural Statistics Service. Federal Information Processing Standards codes (FIPS codes) are a standardized set of numeric or alphabetic codes issued by the National Institute of Standards and Technology (NIST) to uniquely identify states and counties.

Data Analysis Tip: When looking at a data set for the first time, it is always a good idea to play around with it to get a feel for what it contains. For example, there are many instances of "NA" in the data. Remember that this is the standard notation in R for a missing data point. Missing data require special care, and you should seek out a professional statistician when you have lots of missing data. For this lab, we will simply ignore missing values.

Questions

1. What is the trend in total acres devoted to farmland in Durham County, NC, from 1997 to 2007? That is, did Durham become more or less agriculturally based over those 10 years? Report numbers to back up your claims.

2. Which state had the smallest number of farms in 2007?

3. Which state had the second largest number of farms in 2002?

   The Census of Agriculture is a census, so the dataset can be used to obtain quantities for the entire population. For example, we can calculate the total amount of acres devoted to farming in the whole United States, the total number of farms in the whole United States, etc.

4. Use R to get some basic summary stats (mean, standard deviation, minimum, maximum) for the number of farms in 1997, 2002 and 2007.

   Since we have the actual population means, there's no need to take random samples. There's no point in estimating numbers when you can know them exactly. However, our objective for this lab is to see if random sampling works in a real data set. So let's use R to take a random sample of counties. If random selection truly gives a representative sample, the summary of the variables in the sample should be close to the summary of the variables in the whole population of 3,078 counties. First take a sample of size 500 from the data. To do this, type: `n=500; AgpopSample = Agpop[sample(nrow(Agpop),n),]` so that *"AgpopSample"* is a new data frame containing your sample. The sample size 500 was chosen arbitrarily. Later in the semester, we'll learn a principled method of choosing sample sizes.

   There is a really nice package in R called dplyr (part of a collection of R packages for data science called tidyverse) which does data manipulation, subsetting, cleaning and so on in a much better way than we have been doing it in class. We will hopefully get to that later on.

5. Repeat part (4) above using this new data frame. Based on comparisons between the sample means and population means, does it seem that picking counties at random provides a representative sample? How would your answer change with a smaller sample size, for example, $n = 50$.

Part Two: The benefits of random assignment of treatments in causal studies

What are the characteristics of youths doing time? The 1987 Survey of Youth in Custody sampled juveniles and young adults in long-term, state-operated juvenile institutions. Residents of 206 facilities at the end of 1987 were interviewed about family background, previous criminal history, and drug and alcohol use. Read in the new data *"Syc2.txt"* from the course website by typing:

```
Syc2 = read.table("http://www2.stat.duke.edu/~oma9/STA111_SummerI_2018/Labs/Syc2.txt",header=T)
```

Oftentimes, categorical or factor variables are not treated as such by default when reading in new data. We need to ensure the categorical variables are treated accordingly. Type the following into your console to do that:

```
Syc2$crimtype = as.factor(Syc2$crimtype)
Syc2$alcuse = as.factor(Syc2$alcuse)
Syc2$everdrug = as.factor(Syc2$everdrug)
```

The data set is comprised of 22 variables for 2621 youths (verify this by using the `dim` function on the data frame). The variables we use are described below:

Variable	Description
crimtype	Most serious crime in current offense
	1 = violent (e.g. murder, rape, robbery, assault)
	2 = property (e.g., burglary, ;arceny, arson, fraud, motor vehicle theft)
	3 = drug (drug possession or trafficking)
	4 = public order (weapons violation, perjury, failure to appear in court)
	5 = juvenile-status offense (truancy, running away, incorrigible behavior)
	9 = missing
numarr	Number of times arrested
agefirst	Age at first arrest
	99 = missing
alcuse	Did the youth drink alcohol at all during the year before being sent to the institution?
	1 = yes
	2 = no, didnt drink during the year before
	3 = no, doesnt drink at all
	9 = missing
everdrug	Did the youth ever use illegal drugs?
	0 = no
	1 = yes
	9 = missing

The variables have missing data, filled in with 99s and 9s. Since the purpose of this lab is to see how well random assignment to treatments works, we'll feign ignorance and treat the 99s and 9s as if they are real values. Again, this is not good practice; contact a statistician for help when you encounter missing data in your research.

Questions

6. Let's look at the characteristics of these youths before illustrating random assignment of treatments. You may find the `summary` and `table` functions useful here.

   (a) Before looking at the data, guess what two types of crimes are most common among institutionalized youths (you don't need to write your guesses on the lab report). Now let's look at the data. What two types of crimes did most of these youths commit? Report the fraction (of the total) of youths who committed these two crime types on your lab report.

   (b) Before looking at the data, guess the fraction (of the total) of youths in institutions who drank alcohol in the year before being sent there (you don't need to write your guess on the lab report). What is the fraction of these youths who drank alcohol in that year? Report the fraction on your lab report.

   (c) Before looking at the data, guess the average number of times arrested (you dont need to write your guess on the lab report). Now, what is the average number of times arrested in the data? Report the average on your lab report.

   Now let's randomly assign half the youths to one group, and half to another group. The way we will do this is by generating a random sample from a uniform distribution with support between 0 and 1, so that each person has exactly one number between 0 and 1. We will then put everyone less than the median of all the numbers generated in one group, and everyone greater than or equal to the median in another group. To do this, type the following:

   ```
   Index = runif(nrow(Syc2))
   Group1 = Syc2[which(Index<median(Index)),]
   Group2 = Syc2[which(Index>=median(Index)),]
   ```

7. Now, perform the same summaries as in the previous question for each separate group. Are your answers for Group 1 and in Group 2 reasonably similar in comparison to your answers in the previous question? Why is this exercise useful? Think about a clinical trial where those in Group 1 are assigned a placebo drug and those in Group 2 are assigned a new drug. If we want to estimate the effectiveness of the drug, how will this exercise help us?

This ends the lab. Remember to turn in your lab reports on Sakai.