Randomization in Surveys and Causal Studies

Instructions for Lab #3

Statistics 111 - Probability and Statistical Inference

Lab Objective

To get more practice using Stata commands, and to illustrate the benefits of random sampling in surveys and causal studies.

Lab Procedures

1 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.

Download the file agpop1997_2007.dta from the course directory. 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 "." in the data. Remember that this is Stata's simple 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, which Stata does by default for most commands.

Questions:

Hints for questions 1 - 3: With smart `sort`s and `summarize`s, you can answer these questions very quickly.

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. What is the trend in total acres devoted to farmland in the state of North Carolina over the same period?

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

The Census of Agriculture is a census, so the data set 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. Let's use Stata to get some of these quantities.

`summarize [varname(s)]` will give basic summary stats (mean, standard deviation, minimum, maximum). `summarize [varname(s)], detail` will show many other statistics, including the 50th percentile (median). Write down the population means on scrap paper or copy to a text file for use in a later question.

Data Analysis Tip: You can use wildcards if you don't want to type the entire variable name. The asterisk (`*`) can take the place of 0, 1 or more characters. The question mark (`?`) can take the place of exactly 1 character. So if you wanted summary statistics for all variables in 2007 you could type `summarize *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 lab is to see if random sampling works in a real data set. So let's use Stata to take a random sample of 500 counties. If random selection truly gives a representative sample, the averages of the variables in the sample should be close to the averages of the variables in the whole population of 3,078 counties.

At first glance, it may seem that 500 counties can't represent 3,078 counties. Look at the ranges of some of the variables: ***acres2007*** ranges from 7 to 6,101,943 acres, and ***large2007*** stretches from 1 to 523. Can a sample reflect the characteristics of all these wide-ranging variables with only 500 out of 3,078 counties?

Question:

4. Take a random sample. Based on comparisons between the sample means and population means, does it seem that picking counties at random provides a representative sample? Talk to the TA or instructor about your conclusions, and any questions that you may have. After you talk to the TA or instructor, they will give you credit for answering this question.

   It's easy to take a simple random sample in Stata from a data file. `sample` will randomly take a certain number of observations from the dataset. You can either tell Stata the number of observations you want (500) or the percent of observations you want (i.e., 16%). `sample #` will take a random sample of #%, and `sample #, count` will take a random sample of # observations.

   Data Analysis Tip: When you `sample` data from Stata, the data in memory is deleted, and you are only left with the sampled data (the `.dta` file still exists unless you overwrite it, so don't worry too much). Also, if you ever want to be able to replicate the "random" sample of data, you can tell Stata to start from the same place, or seed, with the `set seed ###` command. As long as you use the same seed number, then any random sample you create will be the same random sample.

   The sample size 500 was chosen arbitrarily. Later in the semester, we'll learn a principled method of choosing sample sizes.

   Amazingly, random samples generally provide statistics that are very close to the population statistics. In fact, you would be hard pressed to get closer on all variables by any non-random method of selecting data. Feel free to try.

   If you want to try to check the results, simply load the data back in and take another sample.

   Data Analysis Tip: Here's a generic method for taking a random sample from a population. First, give each unit on the sampling frame a distinct number in the range 1 to N, where N is the total number of units on your sampling frame. Second, create a new data file in Stata and create a single column with numbers from 1 to N. Third, pick a random sample of these numbers from this file using `sample`. Finally, collect data for those units whose numbers were picked in the sample. Getting Stata to do this fourth step requires more advanced data techniques, specifically using `merge`.

   Generating a column of numbers in Stata that go from 1 to N is very simple. Stata has two system variables that can help with counting: `_n` will index the number of an observation in a dataset, and `_N` equals the total number of observations in a dataset. So, `generate index=_n` will create a new variable called ***index*** from 1 to N. You can then `keep` this variable and `save` it in a new dataset.

2 The benefits of random assignment of treatments in causal studies

What are the characteristics of youth 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.

Download the data set syc2.dta from the link. The data set is comprised of 22 variables for 2621 youths (verify this by typing `describe`). 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, didn't drink during the year before
	3 = no, doesn't 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:

5. Let's look at the characteristics of these youths before illustrating random assignment of treatments. For this question, use `tabulate` and/or `summarize`.

(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 percentages of youths who committed these two crime types on your lab report.

(b) Before looking at the data, guess the average age at first arrest for institutionalized youths (you don't need to write your guess on the lab report).

Now, what is the average age at first arrest in the data? Report the average age at first arrest on your lab report.

- For exploratory purposes only, get summary stats on age after excluding those with ***age*** of 99. Observe how both the mean and the standard deviation change.

(c) Before looking at the data, guess the percentage 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 percentage of these youths who drank alcohol in that year? Report the percentage on your lab report.

6. Can you conclude using these data alone that using alcohol increases the chance that youths will go to institutions? Explain your answer in three or less sentences.

7. Now let's randomly assign half the youths to one group, and half to another group. The way we will do this is by `generate`ing a uniform random variable that takes values between 0 and 1. We will then put everyone less than the median in one group, and everyone greater than or equal to the median in another group.

Creating a new variable in Stata is very simple using `generate` ***varname=exp*** where ***varname*** is the new variable you are creating and ***=exp*** is what generates the value of the variable.

In order to create a uniform random variable, we can use Stata's `runiform()` function, which by default creates numbers between 0 to 1: `generate random = runiform()`. Find the median. You can now `generate` another variable based off of this median, or simply condition on the median value. Because this is uniform over 0 to 1, we could also assume the median is 1/2, but let's find the precise median in this case.

Now, perform the same `summarize`s and `tabulate`s as in the previous question for each separate group. Are the percentages or means of ***crimtype***, ***numarr***, ***agefirst***, ***alcuse***, and ***everdrug*** in Group 1 and in Group 2 reasonably similar? Talk to the TA or instructor about your conclusions, and any questions that you may have. After you talk to the TA or instructor, they will give you credit for answering this question.

Notice that the percentages and means for these variables are very similar in the two groups. By assigning the youths to groups at random, we are able to get close balance on all these

variables. This is just one example of the power of statistics.

Data Analysis Tip: As mentioned above, one way to compare these two groups was to actually create a variable that has two values for each of the two groups. If you make this variable, you can then take advantage of Stata's `by` prefix. `by varname:` will perform a command for each subgroup in `by`. So, `by varname: summarize agefirst` will report two different tables of summary statistics, one for each of the groups in ***varname***. Note that you need to make sure the data is `sort`ed by ***varname*** first, or use `bysort` instead of `by`.