

13 Central Limit Theorem and the Lottery

This lab looks at digits randomly generated by statistics students, the lottery, and a computer, to explore the central limit theorem and various properties of numbers that are truly random.

About the data

Today's lab data set comes from the web site for the *Journal of Statistics Education* and consists of samples of size six taken without replacement from the integers $\{1, 2, 3, \dots, 42\}$. There are actually three data sets from three different sources, and in each case the sextuples are (in theory) random selections or samples. The observations in each sample are given in the order in which they were obtained or selected. The first 234 observations (coded with a 1) were obtained from university students in a large statistics class. After discussing the difficulties of being truly random in making selections, the 234 students were asked to act (once each) as a random generator for the Lotto 6 out of 42 game by writing down in any order six numbers selected from $\{1, 2, 3, \dots, 42\}$. The next 264 observations (coded with a 2) were the actual winning combinations for the Irish National Lottery 6 out of 42 Lotto game during the period from September 24, 1994 to March 8, 1997. The final 264 observations (coded with a 3) were obtained through computer simulation using the package S-Plus.

Today's Lab

In lab today we will explore various features of random data and see the Central Limit Theorem in action. Begin by loading the data set into Stata:

```
. use http://www.stat.ucla.edu/labs/datasets/random.dta
```

Then use the describe command to get a general feel for what is contained in the data set.

Question 1: What does a "2" for the type variable represent? Will the num4 variable value always be greater than the num3 variable within one observation? Explain.

Using the graph command that we've seen before, we can get an idea about how the numbers are distributed. Since there are three different ways the data have been collected, we are interested not only in the graph showing all observations combined, but also in the graphs of each type separately.

```
. graph num1, bin(15) xlabel  
. sort type  
. graph num1, bin(15) by(type) xlabel
```

Question 2: What does the bin() option do within the above graph commands?

Question 3: Looking at the graphs you just generated, what can you tell about the distribution of the first of the six numbers selected for each of the data collection types? Do they appear to be similar among all three types? Explain what you see in these graphs and discuss the distributions of one of the other num variables.

As you should recall from lecture and from your textbook, the Central Limit Theorem deals with how means are distributed, not with individual values. We can use a variation of the generate command called “egen” we can create a new variable that contains the mean of num1 through num6 for each observation.

```
. egen avg=rmean(num1 - num6)
```

Make sure that you have a space after “num1” and before “num6” or *Stata* will interpret your command as asking for the average difference of num1 minus num6 which is not what we want at all.

```
. by type: summarize avg
```

Verify that the egen command worked correctly by looking at the means displayed after typing in the summarize command. You should see a mean of 21.50379 for type 2 with a standard deviation of 4.654042. Compare your other summarize results with people near you in the lab before continuing.

Question 4: Before using Stata to make histograms of the new avg variable for each of the three types, what do you think the histograms should look like? What should their general shape be like, and why?

Now check to see if your guess about their shape is correct. Use the graph command with 20 bins, and overlay a normal curve to each graph.

```
. graph avg, by(type) bin(20) norm
```

Question 5: Do the three mean distributions look the same? Are they as you expected? Explain any interesting features in the graphs that you notice.

As we saw in our earlier graphs of the num1 value by type, the students tended to select small values for the first of the six numbers while the lottery numbers and the computer generated ones were more uniformly distributed over the 1-42 range. To further examine how well the students were able to randomly select numbers 1-42, we can look at the average increase from one number to the next. To do this we will generate five new variables.

The code for the first new variable, that shows the increase from num1 to num2 is shown here:

```
. generate inc1 = num2 - num1
```

Once you have inc1 through inc5 generated, we can use the egen command as we did before to find the average increase made in the series of six numbers for each of our 762 observations.

```
. egen avginc = rmean(inc1 - inc5)
```

Question 6: What should the average increase be if the six digits are selected randomly? How do you expect the average increases to be distributed?

Find the mean average increase by type using the summarize command and

create histograms of the new variable.

```
. by type: summarize avginc  
. graph avginc, bin(20) by(type) norm xlabel
```

Question 7: What do you find interesting in the different mean avginc values? Is this what you expected to see? What is the general shape of the distribution for avginc when the numbers come from the lottery or from a computer? How is the distribution shaped when the students generated the numbers?

Question 8: How good a job do you think the students did when trying to generate random numbers?

Assignment

Once you have completed the in-class portion of the lab, type in the clear command to clear the random.dta data file out of Stata. Then type in the edit command.

```
. clear  
. edit
```

The edit command is used to pull up a spreadsheet type window that enables you to enter your own data into Stata. Now you have a chance to try to generate random numbers yourself. Create six variables, and enter numbers 1 through 42 without repeating on any one line as randomly as you can. In the editor window, if you enter six values into the first line, and then click the mouse cursor to the first cell of the second line, Stata will understand that there are only six variables in the data set and you can then use the tab key to move from cell to cell. Create 40 observations, then do the following with your self-created random data.

Question 9: Look at a histogram of one of your six digits. Does it appear to be uniformly distributed over the numbers 1 through 42?

Question 10: Find the average of each line of six numbers and create a histogram of these averages. How is it shaped? Do your numbers look random so far?

Question 11: Generate a range variable that is the difference between your first and sixth numbers. What is the average range? What should this average be if your digits are random?

Question 12: How random do your digits now seem to be? What have you learned in this lab about randomness, and the Central Limit Theorem?