Statistics 104 - Laboratory 4

06.04.14

Drawing samples from populations requires some method for random selection. In the two group activities today you will use different methods to randomly select a sample from a population. In the process you will also learn something about the variability introduced by random sampling. Answer the following questions on the separate group answer sheet.

1. **Activity 1:** Your group will need a sheet of 100 random rectangles for this activity. DO NOT study the sheet of random rectangles. The random rectangles represent a generic population of values. The areas of the rectangles could represent the number of people in households or the number of hours students spend on STAT 104 each week. What we wish to investigate in this activity is what can be learned about the area of the rectangles by looking at samples from the population.
2. **Each individual in the group** should look at the sheet of random rectangles for a few seconds and guess the mean area of all the rectangles on the sheet. The unit of measure is the background square. Thus rectangle 7 has area 4x3=12. Compare your guess with those of the other members in your group. Discuss with your group why you got different guesses. Record the individual guesses for each member in your group on the group answer sheet.
3. **Each individual** should now select five rectangles that, in her/his judgment, are representative of the rectangles on the sheet. Write down rectangle numbers and the areas on the lab answer sheet. Compute the mean of the five areas and record it on the group answer sheet. How does this compare with your guess? How does it compare with the means of the judgment samples for the others in your group? Again discuss with your group why you got different means.
4. We want to take a random sample from the 100 rectangles. We could cut out the rectangles, laminate them, put them in a bag, mix thoroughly and draw five, without replacement, from the bag. Do you think this would give you a random sample?
5. You should have a 10 sided die available to use. Roll the die ten times and record the results on your group answer sheet. Explain how you can use these rolls of the die to select a simple random sample of 5 rectangles.
6. Using your sampling plan, **each individual** should take their own simple random sample of 5 rectangles. Record the rectangle numbers and areas on your group answer sheet. What is the mean area of each sample? Are sample mean areas the same for all your samples? Why?
7. How do the random sample means compare to the judgment sample means and guesses?
8. Which do you think will be a better guess of the true mean area of the 100 rectangles, the random sample mean of 5 rectangles or the random sample mean of 20 rectangles? Explain briefly.
9. **Activity 2:** For this activity you will be using a tool called the Shiny Database Sampler to take a random sample of United States residents from US census data. The census data is the Public Use Microdata Sample (PUMS) which is a 3 million person subset of the entire Census data. For this activity we treat our samples as though they are selected from the full census records.

We are going to explore how these random sampling plans relate to the goals of a sample survey. The tool will allow you to define either a simple random sampling plan or a stratified random sampling plan. In the following two scenarios we will explore the advantages and disadvantages of these two sampling plans. Access the tool at <http://shiny1.stat.iastate.edu/karstenm/>.

Scenario 1: Suppose that our goal is to investigate the association between age and income for all US residents. We have a budget that allows us to survey around 2000 people. To collect our sample we decide to take a simple random sample of 2000 US residents and record the information about their age and income.

* 1. Is this study and example of an experiment or an observational study? Explain your answer.
  2. Your colleague Bob claims that we are wasting our budget to get only 2000 people using random sampling. He says that we could get 20000 responses to the survey if we invested that money into a mailing campaign in Minneapolis. Explain to them why the random selection is important.
  3. Another colleague, Jill, asks why we do not stratify by state when we take the sample so that each state is well represented. Explain why this idea would not create a representative sample to pursue our goal.

Now that we have decided on our sampling plan, let’s go collect our data. The Shiny Database Sampler needs to be told **4** pieces of information in order to collect census records the way you want. (1) Select the database called “Census”, (2) select the “simple random sample” option, (3) enter the random seed as today’s date, “0604”, you can think of this as telling the computer where to stop its random number generation so that we each get the same sample and (4) lastly tell it that we want “2000” random draws. Once you have drawn your samples you can take a peek at your data set in the main panel of the webpage. After you get a sense of what is in the data, download it to the desktop and open the file in JMP.

* 1. In this scenario we are looking at the association between variables age and total income (labeled in the data as age and inctot). Which of these is the explanatory variable? Which is the response?
  2. Use Analyze 🡪 Fit Y-by-X in JMP to make the scatterplot of income and age. Describe what this scatterplot tells us about the association these variables.
  3. Now use the option menu in JMP to fit the linear regression line to your sample data. Report and *interpret* the slope within the context of this situation.
  4. The correlation coefficient for this data, . Bob claims that this shows that age causes a moderately weak effect on income. What is wrong with his claim?

Scenario 2: Suppose now that our goal has changed. Now we wish to investigate the association between gender and income. We want to compare the median income of males and female. We still have a budget that allows us to survey around 2000 people. To collect our sample we decide to take a stratified random sample of 1000 residents from each gender.

* 1. Explain in general why collecting a stratified random sample is a better plan than a simple random sample for investigating this relationship.

Now that we have decided on our new sampling plan, let’s go collect our data. The Shiny Database Sampler will need to be told **5** pieces of information in order to collect census records the way you want this time. (1) Select the database called “Census”, (2) select the “**stratified** random sample” option, (3) keep the random seed as today’s date, “0604”, (4) select “sex” as strata variable and (5) lastly tell it that we want “1000” random draws from each gender.

It will take a few seconds to collect this data, it is sifting through millions of records and randomly selecting them from within state groups after all! Once you have drawn your samples you can take a peek at your data set in the main panel of the webpage. You will be able to answer the following questions using the summaries provided on the webpage.

You will notice that the summaries are all broken down by gender, but the genders are not given names, they are given a code numbers. This is done on the census to save computer storage space (saving a “1” is much smaller than “Male”, “2” is smaller than “Female”). All survey codes is available at [www.census.gov/govs/www/gidstate](http://www.census.gov/govs/www/gidstate) .

1. Report the mean and 5-number summary for the income of the sample from the males (male = 1).
2. Report the mean and 5-number summary for the income of the sample from the females (female = 2).
3. Compare the distribution of incomes for males and females using the values from parts i and j.
4. What would be a good way to **visually display** this information so aid in making these comparisons? Explain your answer. (hint: look at the second tab in the webpage to play around with the plots.)
5. Why might this information not be representative of the population that we may have intended? Hint: look at the age column in the sample you have collected.