LAB DUE: Wed. Oct. 14th

Statistics
Fall 2009

Computer Lab - 3

This lab is somewhat different than the last lab in that instead of analyzing data that I give to you, you will generate data of your own. You will do this using a random number selection system found at an internet site. The samples that you generate will be of varying sizes. The analysis of these samples will help you to see first hand the issues surrounding how well particular samples will represent a defined population.

Generating samples:

You should generate 10 **random** samples of numbers. I want you to create samples of the following sizes: 5 samples of 4 values each (N = 4) and 5 samples with 16 values each (N = 16).

Go to http://www.dougshaw.com/sampling/. This web site will allow you to choose samples of particular sizes (in this case n=4 and n=16) from numbers out of a particular range. Set the range of between 1 and 100. (The default setting is 20 and 100). Make sure to click the non-repeating choice in the list just below the range. Once you click the "go" button the web site will generate a random sample. Copy down the set of numbers and continue generating new samples until you have all ten samples.

Using SPSS:

You should start the software as you have done in previous labs. You are automatically put into a new, blank data screen. Begin by defining the variables for the 10 samples. (See the handouts for labs one and two for additional instruction about defining variables) You will need 10 variables (10 columns) defined, one for each of the samples that you have generated. When defining these variables you should identify:

Variable name: a name for each of the ten samples that will allow you to keep track of them.

type: numeric with at least a column width of 3 (default is 8) and the number of decimal places as 0 (default is 2).

Label: a descriptive phrase that will help you to distinguish between your samples.

Once the variables have been defined you can enter your 10 sets of data. Remember that the first five samples (variables) should each have 4 randomly selected numbers. The remaining samples should have 16 randomly selected numbers.

Analysis:

Again you will use the **descriptive statistics** command, but this time, choose **descriptives** from the side menu to analyze these data. Select all ten variables to be analyzed and then click on the

options button. Select mean and standard deviation from the menu of statistics and then click on the **continue** button. When you are returned to the **descriptives** dialogue box click on the **OK** button. This will send your statistics to the output file. Notice that this command option does not include a chart option but we do not need that for this analysis.

Using your data:

At this point you can either print the output file (and the data file if you wish) or you can gather the necessary data directly from the screen and write it onto the lab report. I am not asking for a printout with this lab.

Answering the lab questions:

This lab is intended to demonstrate that although a population mean describes the population as a whole and that many of the random samples that are drawn will represent the population well, some random samples will look quite a bit different from the population mean. This is more likely to be true when drawing small samples rather than large samples. In order to see this I ask you to compute by hand two things using the values for your sample means. Hopefully, the instructions below will clarify what I am asking for.

Part c computations:

After looking at your five means for small samples (n = 4) and five means for large samples (n = 16) I would like you to for a brief time use the means as if they were data points themselves. I would like you to take the five means of samples 1 – 5 and find the mean (average) of these values. Then I would like you to use these same five values for the means and compute a standard deviation. An example appears below. I ask you to find the same two pieces of information (the mean of the means and the standard deviation of the means) for the samples of 16 (samples 6 – 10). The point of this is that finding a mean always gives information about what the central feature of a set of numbers is. So, by finding the "mean of the means" you can see how well, on average, lots of samples represent the population. By finding the "standard deviation of the means" you can see how much variability there is from one sample that could be drawn to another.

Example of part c computations:

Assume that the following represented my data for samples with n = 4:

20 45 52 60 78

I would use those five numbers to compute the mean

$$\frac{20 + 45 + 52 + 60 + 78}{5}$$

And I would use those five numbers to compute the standard deviation. I could do this using either the definitional formula or the computational formula.

$$\sum X = 255 \qquad \sum X^2 = 14813 \qquad S = \sqrt{\frac{14813 - \frac{255^2}{5}}{5}}$$