**Central Limit Theorem**

**Purpose and summary of procedures**

The purpose of today's lab is to use simulation to visualize the sampling distribution for the sample mean (as we have discussed in class). You should remember that the Central Limit Theorem (CLT) tells us that as our sample size gets larger, the sampling distribution of the sample mean converges to a normal distribution. Therefore, when we have a large sample size, we can say that the sampling distribution for the sample mean is approximately normal, regardless of the distribution from which we are sampling.

The procedure for today's lab is listed below.

1. Define the population. In our case, we begin with one "population" of 10000 incomes.
2. Define the population parameter of interest. As a simple example, you may be interested in mu, where mu represents the mean income of people in the population. (As an example, another possible interest is in the proportion of the population, p, who make more than a certain amount.)
3. Determine what statistic/estimate you would use to estimate the population parameter. For the first example above, this would be x-bar, the sample mean. (In the latter case, we would use p-hat, the sample proportion.)
4. Use S-Plus to take a random sample from the population of interest, then calculate the sample statistic for this sample. This is analogous to the "real-life" procedure of randomly selecting a sample of people, recording data about them, and calculating the mean of their heights or the proportion of people meeting a certain criterion.
5. After following the procedure above with many different samples, we can make a histogram of the sample means to see what the sampling distribution looks like.

The *Central Limit Theorem (CLT)* states that as the sample size grows larger, the sampling distribution for the sample mean, x-bar, converges to normal. The sample mean is an *unbiased* estimator (of the quantity we're interested in, mu), so their expected value of the sample mean is equal to the parameter which we are trying to estimate. This means that the sampling distribution of the sample mean must be centered around the population parameter *mu*.

1. How would you describe the shape of this population?
2. What is the mean income of this population?
3. What sampling statistic/point estimate would you use to estimate the mean of this population if you were given a random sample from the population?

#### Simulated sampling (sample means)

Begin by taking a small sample of 5 people from the population. To generate this random sample, use the following command (in the "Commands" window) to randomly select five different members of the population given in the first column of the dataset. (This assumes a dataset named "incomes".)

 sample(incomes[,1], replace=F, size=5)

1. What is the sample mean for your sample?
2. How far/close is it to the actual population mean?

##### Simulation with n=5

Now, we'd like to get an idea of what happens when we take multiple random samples of size 5. Use the following command to take a series of 1000 samples of size 5, find the corresponding sample means, and store the sample means in the second column.

for (i in 1:1000) incomes[i,2] \_ mean(sample(incomes[,1], replace=F, size=5))

Now make a histogram of all the sample means. To calculate the mean and standard deviation of your sample means you can issue the commands

mean(incomes[,2], na.rm=T)

sqrt(var(incomes[,2], unbiased=F, na.method="omit"))

1. Describe the shape of the histogram.
2. What is the center of the distribution of sample means?

Repeat the above steps for n=15, n=35.