Assignment for Advances in Ecology, 597B

Due Tuesday, Sep.30,2008

I This exercise is a computer experiment that gives you a feel for the concept of **sampling variability**. This concept underlies basic statistical ideas, including fundamental concepts like confidence intervals, hypothesis testing. Note: You may also paste in all `R` code for this assignment from the file: `http://www.stat.psu.edu/~mharan/200H/labs/hw1.R`. Begin by loading the example data set from: `http://www.stat.psu.edu/~mharan/200H/labs/bulblife.txt`.
Do this using the command:

```
allbulbs=scan("http://www.stat.psu.edu/∼mharan/200H/labs/bulblife.txt")
```

This is a data set of the lifespan of 100,000 bulbs (in years).You are interested in finding out the population mean. If we treat this as the entire population, you do not need really need do any statistics — you would just take the average of all the bulbs and get your answer. In real life, however, you will only get to observe a small sample from the population and have to estimate the true mean based on this sample. We want to get an idea about how much our estimate would vary from sample to sample, even though we have only one sample to work with. Do the following:

1. First look at the distribution of the entire population (draw a histogram of it). What does this distribution look like? What is its mean? (this is *the true mean*). Use the commands:

   ```
   hist(allbulbs,main="histogram of bulb life spans’’)
   allbulbmean=mean(allbulbs)
   ```

2. Here, we will pretend we do not have access to the population and that we can only obtain one set of samples (this is what would happen in reality.) Find the mean based on just this sample. Use the commands:

   ```
   samp=sample(allbulbs,30)
   mean(samp)
   ```

   Report this estimate. This is your estimate of the true mean based on a single sample.

3. It is important to assess how much this estimate would vary from sample to sample to give us a sense of how confident we are about our estimates. How can we do this with just a single sample? Since we

have a random sample of size greater than 30, we can use standard statistical theory to produce an approximate sampling distribution for the sample mean based on this single sample. The approximate sampling distribution for the sample mean is:

$$\text{Normal}(\text{mean=true mean}, \text{standard deviation} = s/\sqrt{n}),$$

where s=sample standard deviation and n=sample size. Use the following command to find the standard error.

```
serror =sd(samp)/sqrt(30)
```

Report the estimated standard error. (If you type `serror`, you will see this estimate.) This provides you with an idea of how much the estimates would vary from sample to sample.

4. Now randomly draw a sample of size 30 from this population *10000 times*. Find the sample mean ($\bar{X}$) for each row of samples. Draw a histogram of the sample means. Use the commands:

   ```
   manysampmeans=rep(NA, 10000)
   for (i in 1:10000)
     {
       testsamp = sample(allbulbs,30)
       manysampmeans[i]=mean(testsamp)
     }
   hist(manysampmeans,main="histogram of sample means")
   ```

5. Now that you have 10000 sample means, find the mean and standard deviation of these 10000 sample means:

   ```
   mean(manysampmeans)
   sd(manysampmeans)
   ```

   Compare this mean and standard deviation to the mean and standard deviation you get for the normal distribution of question 3. Are they similar?

6. How does your approximate sampling distribution from question 3 compare to the histogram from question 4 ? You can overlay the two plots in the following way:

```
curve(dnorm(x,mean=allbulbmean,sd=serror), col="red",lwd=2)
hist(manysampmeans,main="histogram of sample means",freq=FALSE,add=TRUE)
```

The red plot is the distribution of sample means according to statistical theory, and the histogram is based on lots of (10000) sample means. Paste the plot into your report.
What you should note is the following: Statistical theory *using just 1 sample of size 30 from a population* gives you a pretty reasonable estimate of the variability of the sample mean over *all samples of size 30 from the population.* So it is possible to assess variability of your estimates based on just one sample of size 30.

7. Now repeat the above experiment for samples of size 500. You need to only report the following: (a) A plot like the one you have from question 6, overlaying the distribution estimated by statistical theory on the histogram of all sample means. (b) Do the two match up in this case as well? (c) Compare this plot to the plot from question 6. What do they have in common? What is different?
What you should note is that the variability of the estimates are smaller here than for question 6 — larger samples give better estimates!

II Statistical issues in your research. The following writup should be typed and no longer than 2 pages in length. Select a problem that is part of your current or recent research work (the research work must involve a dataset). If you do not have such research, discuss a closely related research problem, perhaps one that you anticipate working on in in the near future.

1. Explain clearly what the main research question or questions of interest are (do this in as quantitative a manner as possible).

2. Describe the data set clearly, include descriptions of each of the variables and any other information (e.g. how the data were collected) that may be relevant.

3. Explain what kind of statistical modeling approaches you would consider for analyzing the data set and answering your questions of interest. Do this in as much detail as feasible given the page limit and your knowledge of statistics. Answers can include descriptions of any regressions you plan on fitting to the data, exploratory data analysis (plots etc.) that you may choose to look at, and if there are issues that may be important but you are not equipped to address yet, list them

so we can look into statistical methods that may be useful in a future class. Make sure you address whether spatial dependence may be an issue, try to be as specific about what you think may be spatially dependent and why. Can you think of a context where hierarchical modeling may be important?

4. Discuss any concerns you may have regarding noise in the collected data, missing variables, information you wish you had but do not etc. Suggest possibilities (if you have any) for overcoming these problems (a new but feasible study, utilizing information from another source that may still inform you about variables you cannot collect etc.)

5. Be prepared to discuss this report in a short informal presentation in front of class.