STATISTICAL COMPUTING ACTIVITY

6a: One-Population Confidence Intervals I

Purpose: This SCA introduces us to the analysis cycle, the process we should follow in doing statistical analysis. Here, because of where we are in the course, we are limited to calculating confidence intervals about a single population parameter… usually the mean. However, the process will remain the same for any statistical analysis you may do in your life.

Note that this is the last SCA with a narrative document. Future SCAs will consist of R scripts with comments throughout. Make sure you are able to make the leap. ☺

R Functions: We will see the following functions in R. Some are new; some are not.

- `factor`
- `hildebrand.rule`
- `table`
- `shapiroTest`
- `t.test`
- `binom.test`
- `normoverlay`
- `wilcox.test`
- `onevar.test`

PROCEDURE

As usual, here is the procedure.

PART O: START-UP

Here are the usual start-up steps. In this, I am assuming that you have started R from your working directory. If not, close R and restart from your working directory.

1. Load the usual helper functions.

```
source("http://rfs.kvasaheim.com/stat200.R")
```

2. The data file we will use is the `someCollegeClean` data set. Read it into R in the usual manner:

```
dt = read.csv("http://rfs.kvasaheim.com/data/someCollegeClean.csv")
```

3. Normally, you would want to just attach the data as-is. However, let us look at the data before attaching it (you only get one chance to attach it). The `summary` function will help here.

```
summary(dt)
```

4. Notice that one of the variable in the data set is **level**. This variable is the level that the student is at this college. It can take on levels **Freshman**, **Sophomore**, **Junior**, and **Senior**. It is an ordinal-level variable. However, without telling the computer it is ordinal and what that order is, there is no way for it to know this. So, that will be the fix we implement on the data.

5. The new function is **factor**. It takes (at least) two pieces of information. The first is the variable you want to order. The second is the order. Here is the line to run.

```
dt$level = factor(dt$level, levels=c("Freshman","Sophomore","Junior","Senior"))
```

   The first slot in the factor function is the variable I want to order. Here, that variable is **level** from the **dt** data set. Because I have not attached the **dt** data set, I had to use the **$** operator to tell R where to find the **level** variable. The second slot is the levels of the variable in their correct order. Make sure the spelling matches the levels in the current variable. Otherwise, there will be an error.

   Finally, I am taking what I just had R calculate and store it back in the **level** variable of the **dt** data set. When this line is run, R will interpret the **level** variable of the **dt** data set as an ordinal variable.

6. Now, summarize the data set to make sure that the change was made.

```
summary(dt)
```

7. Finally, attach the data set so we can avoid the **$** notation in the future.

```
attach(dt)
```

Now that the data are ordered and attached, let us look at several research questions that the data can answer for us.

Part I: The Research Questions

Now that we have the data imported into R, we can perform some statistical analysis rather easily. What follows is a series of research questions that we will use the data to properly answer.

Research Question 1: ACT Composite

The full research question is "What is the average ACT composite score at Some College (SC)?" Note that we are trying to estimate an average (mean) of a numeric variable. That tells us that the parametric test we want to use is the one-sample t-procedure. However, that procedure has assumptions that may or may not be met with the data. If any assumption is violated, we cannot use the t-procedure.

What is the assumption of the t-procedure?

1. The assumption of the one-sample t-procedure is that the data come from a Normal population. To test this, let us use the Shapiro-Wilk test. If the data do come from a Normal population, then the p-value of the Shapiro-Wilk test will be greater than 0.05.

```
shapiroTest(ACTcomposite)
```

2. Because the p-value of 0.1428 is greater than 0.05, the assumption is not violated. Thus, we can use the one-sample t-procedure to estimate the mean ACT composite score.

```
t.test(ACTcomposite)
```

3. The command gives a lot of information. Until we cover hypothesis testing, most of that information will be useless to you. However, it also gives the limits of a 95% confidence interval.
4. From the results of the one-sample t-procedure, we have the following conclusion:

> ***According to the one-sample t-procedure, we are 95% confident that the mean ACT composite score at SC is between 20.6 and 21.4, with a point estimate of 21.0.***

That is all there is to it:

1. Look at the research question to determine the parameter and the number of populations
2. Determine the appropriate parametric procedure
3. Test the assumptions of that procedure
4. If no assumption is violated, use the procedure
5. Interpret the results

Of course, sometimes the assumptions are violated. That leads to another test.

RESEARCH QUESTION 2: SAT MATH SCORE

The full research question is "What is the average SAT Math score at SC?" Note that we are again trying to estimate an average (mean) of a numeric variable. That tells us that the parametric test we want to use is the one-sample t-procedure. However, that procedure has assumptions that may or may not be met with the data. If any assumption is violated, we cannot use the t-procedure.

1. The assumption of the one-sample t-procedure is that the data come from a Normal population. To test this, let us use the Shapiro-Wilk test. If the data do come from a Normal population, then the p-value of the Shapiro-Wilk test will be greater than 0.05.

```
shapiroTest(math)
```

2. Because the p-value is much less than 0.05, the assumption is violated. Thus, we cannot use the one-sample t-procedure.

   The next most powerful procedure for the population mean is the Wilcoxon procedure. It is an example of a non-parametric test (does not assume Normality). Because the Wilcoxon procedure actually is a procedure for the population median, we can only use it to learn about the population mean when the data come from a symmetric distribution.

3. How have we tested for symmetry in data? That's right!!! We used the Hildebrand rule.

```
hildebrand.rule(math)
```

4. According to the Hildebrand rule (H = -0.045), the assumption of symmetry is not violated in this data. Thus, we can use the Wilcoxon procedure.

```
wilcox.test(math, conf.int=TRUE)
```

5. Again, the command gives a lot of information. Until we cover hypothesis testing, most of that information will be useless to you. However, it also gives the limits of a 95% confidence interval.

   From the results of the Wilcoxon procedure, we have the following conclusion:

   According to the Wilcoxon procedure, we are 95% confident that the mean SAT Math score at SC is between 625 and 635, with a point estimate of 630.

   Feel free to use this space to summarize the analysis steps.