STATISTICAL COMPUTING ACTIVITY

6b: One-Population Confidence Intervals II

Purpose: This SCA give us more practice on 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.

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

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

PROCEDURE

As usual, here is the procedure.

PART O: START-UP

Here are the same start-up steps from SCA06a, with the change in the first line. In this, I am (as always) assuming that you have started R from your working directory.

```
##### Script for SCA6b
#####


### Preamble

# Import extra functionality
source("http://rfs.kvasaheim.com/stat200.R")

# Read in data
dt = read.csv("http://rfs.kvasaheim.com/data/someCollegeClean.csv")
summary(dt)
dt$level = factor(dt$level, levels=c("Freshman","Sophomore","Junior","Senior"))
attach(dt)
```

PART I: THE RESEARCH QUESTIONS

Now that we have the data imported into R, we can perform some statistical analysis rather easily. What follows continues the series of research questions we started in SCA06a.

RESEARCH QUESTION 3: SAT COMPOSITE SCORE

The full research question is "What is the average SAT composite score at SC?" Note that we are trying to estimate a single 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 (p-value < 0.0001), 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 = -0.065), 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 composite score at SC is between 1315 and 1335, with a point estimate of 1325.***

Research Question 4: Females at SC

The full research question is "What is the proportion of female students at SC?" As usual, ask yourself the following questions:

a) How many populations are involved? (One)

b) What population parameter am I looking for? (Proportion)

c) What is the optimal parametric test? (Binomial test)

d) What are its assumptions? ($np \geq 5$ and $n(1-p) \geq 5$)

e) Are any of those assumptions violated? (No)

Since, the assumptions are not violated, we can use the Binomial procedure to estimate the population proportion.

1. The Binomial procedure needs to know the number of successes and the number of trials. The number of successes is the number of females in the sample. The number of trials is the number of people in the sample. This information can be gleaned from using the **`table`** function.

```
table(gender)
```

2. From that command, we see there are 263 females in the sample of 643 people. The code to run the Binomial procedure is

```
binom.test( x=263, n=643 )
```

3. From the results, we again look for the confidence interval. From that, we can conclude

> ***According to the Binomial procedure, we are 95% confident that the proportion of students at SC who are female is between 37% and 45%, with a point estimate of 41%.***

RESEARCH QUESTION 5: HOME SCHOOLERS AT SC

The full research question is "What is the proportion of Home School students at SC?" As usual, ask yourself the following questions:

a) How many populations are involved? (One)

b) What population parameter am I looking for? (Proportion)

c) What is the optimal parametric test? (Binomial test)

d) What are its assumptions? ($np \geq 5$ and $n(1-p) \geq 5$)

e) Are any of those assumptions violated? (No)

Since, the assumptions are not violated, we can use the Binomial procedure to estimate the population proportion.

4. The Binomial procedure needs to know the number of successes and the number of trials. The number of successes is the number of home-schooled students in the sample. The number of trials is the number of people in the sample. This information can be gleaned from using the **`table`** function.

```
table(highschool)
```

5. From that command, we see there are 523 students home-schooled in the sample of 643 people. The code to run the Binomial procedure is

```
binom.test( x=523, n=643 )
```

6. From the results, we again look for the confidence interval. From that, we can conclude

> ***According to the Binomial procedure, we are 95% confident that the proportion of students at SC who are female is between 78% and 84%, with a point estimate of 81%.***

Research Question 6: Grade Point Averages

The full research question is "What is the average GPA of students at SC?" As usual, ask yourself: How many populations are involved? What population parameter am I looking for? What is the optimal parametric test? What are its assumptions? Are any of those assumptions violated? If so, what is the next-best test? What are its assumptions? Are any of those assumptions violated? If so, what is the next-best test? What are its assumptions? Are any of those assumptions violated? If so, etc.

There is just the one population (students at SC). The population parameter is the mean, μ. The parametric procedure is the one-sample t-procedure. It requires that the data come from a Normal distribution. According to the Shapiro-Wilk test, this assumption is violated (p-value << 0.0001). Thus, we ***cannot*** use the one-sample t-procedure.

The next-most optimal procedure is the Wilcoxon procedure. It requires that the data come from a symmetric distribution. According to the Hildebrand rule (H = -0.41), this assumption is also violated. Thus, we cannot use the Wilcoxon procedure, either.

That leaves the **non-parametric bootstrap** as our final option. We have done this many, many times in the past. I just have not officially given it a name. There are no assumptions to the non-parametric bootstrap, so we can always use it. Unfortunately, it is the least powerful procedure of any covered here. Thus, it should only be used as a last resort.

From the results, our conclusion is

> ***According to the non-parametric Bootstrap procedure, we are 95% confident that the average GPA of students at SC is between 2.80 and 2.97, with a point estimate of 2.88.***

Here is the entire code for this analysis:

```
### Research Question #6:
#   What is the median GPA of students at SC?

shapiroTest(gpa)
hildebrand.rule(gpa)

## Non-parametric bootstrapping

st = numeric()
for(i in 1:1e4) {
  x = sample(gpa, size=length(gpa), replace=TRUE)
  st[i] = mean(x)
}

quantile(st, c(0.025,0.975))
mean(st)
```

RESEARCH QUESTION 7: GRADE POINT AVERAGES II

The full research question is "What is the **median** GPA of students at SC?" Ask yourself the usual questions. The parameter we are estimating is the population median. The analysis process is very similar to that of the population mean. The t-test is preferred (Why??).

If that cannot be used, then we can use the Wilcoxon test, also called the Wilcoxon Signed Rank Test for a Median. The assumption of the Wilcoxon test is that the data are from a symmetric distribution. How do we test that?

Since the data are not from a symmetric distribution, we use the non-parametric bootstrap.

After running the non-parametric bootstrap, we have the following conclusion:

According to the non-parametric Bootstrap procedure, we are 95% confident that the median GPA of students at SC is between 3.00 and 3.33, with a point estimate of 3.28.

Here is the entire code for this analysis:

```
### Research Question #7:
#   What is the median GPA of students at SC?

shapiroTest(gpa)
hildebrand.rule(gpa)

# Non-parametric bootstrapping

st = numeric()
for(i in 1:1e4) {
  x = sample(gpa, size=length(gpa), replace=TRUE)
  st[i] = median(x)
}

quantile(st, c(0.025,0.975))
mean(st)
```

Again, I want to stress that you should know your data and what tests you are doing and why you are doing them. R will do just about whatever you want, including many things that you should not do.

Research Question 8: ACT Composite II

The full research question is "What is the **variance** of student GPAs at SC?" Ask yourself the usual questions. The parameter we are estimating is the population variance. We have only one method for estimating it, and that requires that the data come from a Normal distribution. That method is a Chi-Square test.

1. Since the Chi-Square test requires that the data come from a Normal distribution, we must check that first.

```
shapiroTest(ACTcomposite)
```

2. It passes! Because the p-value of the Shapiro-Wilk test, 0.1428, is greater than 0.05, the data come from a sufficiently Normal population. Thus, we can use the Chi-Square test.

```
onevar.test(ACTcomposite)
```

Finally, our conclusion is

> ***According to the Chi-Square test, we are 95% confident that the variance of student ACT composite scores at SC is between 22.19 and 27.63, with a point estimate of 24.69.***

Had the data not come from a Normal distribution, we would use the non-parametric bootstrap. We see that in the next research question.

RESEARCH QUESTION 9: GRADE POINT AVERAGES III

The full research question is "What is the **variance** of student GPAs at SC?" Ask yourself the usual questions. The parameter we are estimating is the population variance. We have only one method for estimating it, and that requires that the data come from a Normal distribution. That method is a Chi-Square test.

1. Since the Chi-Square test requires that the data come from a Normal distribution, we must check that first.

```
shapiroTest(gpa)
```

2. It fails. ☹ Because the p-value of the Shapiro-Wilk test is much less than 0.05, the data do not come from a sufficiently Normal population. Thus, we cannot use the Chi-Square test.
3. Our only option at this point is to use the non-parametric bootstrap. You should know it by now. From those results, our conclusion is

> ***According to the non-parametric Bootstrap procedure, we are 95% confident that the variance of student GPAs at SC is between 1.04 and 1.32, with a point estimate of 1.18.***

Here is the entire code for this analysis:

```
### Research Question #9:
#   What is the variance of student GPAs at SC?

shapiroTest(gpa)

# Non-parametric bootstrapping

st = numeric()
for(i in 1:1e4) {
  x = sample(gpa, size=length(gpa), replace=TRUE)
  st[i] = var(x)
}

quantile(st, c(0.025,0.975))
mean(st)
```

Conclusion: That is all for this SCA. Note that we examined one-population methods for estimating the mean, the proportion, the median, and the variance (and the standard deviation). Had we wanted, we could have also done analyses for interquartile ranges, half ranges, and mid ranges. Those would require we go directly to the non-parametric bootstrapping, since there are no general tests for those population parameters.