STATISTICAL COMPUTING ACTIVITY

7a: Hypothesis Testing I

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 can now calculate — and interpret — p-values. Again, let the computer do the work, you need to focus on the proper test and the proper interpretation of the results.

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

- `shapiroTest`
- `wilcox.test`
- `boxplot`
- `hildebrand.rule`
- `aggregate`

PROCEDURE

As usual, here is the procedure.

PART 0: START-UP

Let us examine a new data file. The `clf` data file consists of a sample of countries from the world. For each country in the file, five variables are measured: year the data is taken, the country's 'failed stats index,' the production ratio, whether the country has nationalized its petroleum production, and whether the country has nationalized its petroleum production.

So, with that in mind, here are the usual start-up steps.

```
##### Script for SCA7a
#####
##### February 20, 2017
#####


### Preamble

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

# Read in data
dt = read.csv("http://rfs.kvasaheim.com/data/clf.csv")
attach(dt)
```

PART I: THE RESEARCH HYPOTHESES

Now that we have the data imported into R, we can perform some statistical analysis rather easily. When dealing with exploratory analysis, we only report confidence intervals. When doing hypothesis testing, both the p-value and the confidence interval are reported.

RESEARCH HYPOTHESIS 1: PRODUCTION RATIO

The full research hypothesis is "The average production ratio is 100?" Note that we are testing a hypothesis about 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.

Note, too, that the claim is $\mu=100$. That means the null hypothesis is $\mu=100$ and the alternative hypothesis is $\mu\neq100$. The test is a **two-tailed test**.

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(productionRatio)
```

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(productionRatio)
```

4. According to the Hildebrand rule (H = 0.34), the assumption of symmetry is violated in this data. Thus, we cannot use the Wilcoxon procedure either. Thus, we must use non-parametric bootstrapping

```
st=numeric()

for(i in 1:1e4) {
  x = sample(productionRatio, size=58, replace=TRUE)
  st[i] = mean(x)
}

hist(st)
mean(st)

2*mean(st>=100)                # The estimated p-value
quantile(st, c(0.025,0.975))   # The estimated conf int
```

5. Note that there are a couple things with this script that are new. They are here to estimate the p-value. The first thing is the line labeled "`The estimated p-value`." The second is "`mean(st)`." The mean line is required so that we know the direction of the inequality in the p-value line. Since the sample mean is less than our hypothesized mean, the values that are more extreme are those **greater than** 100.
6. From the results of the non-parametric bootstrap procedure, we have the following conclusion:

> ***According to the non-parametric bootstrap procedure, we cannot reject the null hypothesis that the mean production ratio is 100. The p-value of 0.06 is not less than $\alpha = 0.05$. Additionally, we are 95% confident that the mean production ratio is between 51 and 102, with a point estimate of 74.***

7. Here is a box-and-whiskers plot for this data.

```
boxplot(productionRatio)
```

8. From looking at this graphic, we can definitely see that the variable is neither symmetric nor Normal. Thus, it is not a surprise the two statistical tests returned those results.
9. Also, we can see that most of the values are less than 100. Thus, it makes sense that the null hypothesis was rejected.

Research Hypothesis 2: Failed States Index

The full research hypothesis is "The mean failed states index is 60." Again, the claim is $\mu=60$, which means the null hypothesis is $\mu=60$ and the alternative hypothesis is $\mu\neq60$. Written correctly:

$$H_R : \mu=60$$

$$H_0 : \mu=60$$

$$H_A : \mu\neq60$$

Now, let us ask our usual questions. How many populations? What population parameter? What is the optimal test? What are its assumptions?

There is still just one test. We are making a claim on the population mean. The optimal test is the t-test. The t-test requires Normality. Let us test Normality.

1. Before testing Normality, let us again look at the variable using a box-and-whiskers plot.

```
boxplot(fsi)
```

2. The data look to be negatively skewed. According to the Shapiro-Wilk test, the data do not come from a Normal population (p-value = 0.01817). Thus, the t-test cannot be used.
3. The second test is the Wilcoxon test. It requires symmetry. According to the Hildebrand rule, the data do not come from a symmetric distribution (H = -0.209). Thus, we cannot use the Wilcoxon test.
4. We, again, must use non-parametric bootstrapping. Here is the code:

```
st=numeric()
for(i in 1:1e4) {
  x = sample(fsi, size=58, replace=TRUE)
  st[i] = mean(x)
}

mean(st)

2*mean(st<=60)                  # The estimated p-value
quantile(st, c(0.025,0.975))    # The estimated conf int
```

   Note the differences between this code and the previous one. Note especially the difference on the p-value line. Why is the direction now "<"? The sample mean is greater than our hypothesized population mean. Thus, the "part that is more extreme" is to the lower side of 60.

5. And so, from the results, we can conclude

> ***According to the non-parametric bootstrap, we can reject the null hypothesis that the mean failed states index is 60 (p-value = 0.0016). In fact, we are 95% confident that the mean failed states index is between 63.7 and 76.3, with a point estimate of 70.1.***

RESEARCH HYPOTHESIS 3: FAILED STATES INDEX, II

1. In the previous example, we used the non-parametric bootstrap to test a hypothesis about a population mean from our data. Note that the bootstrap has a lower power than the t-test and the Wilcoxon test. Thus, we really do want to use those other tests when we can.
2. One option that may help is to transform the data. Transforming the data may produce a distribution that is sufficiently symmetric or is sufficiently Normal. The transformation function choice can sometimes be determined from the histogram. Most of the time, however, that function may not even exist. Here, we illustrate the process.
3. From the previous part, the failed states index was neither Normal or symmetric. Let us apply the "square" transformation to see if the resulting variable is either symmetric or Normal.

```
fsi2 = fsi^2
```

4. Now, let us perform the Shapiro-Wilk test on the new **fsi2** variable to see if it is sufficiently Normal.

```
shapiroTest(fsi2)
```

5. According to that test, the new variable is sufficiently Normal (p-value=0.1458). Because of this, we can use the t-test on the transformed variable.

```
t.test(fsi2, mu=3600)
```

6. Our null hypothesis is that μ=60, where μ is the mean of the failed states index variable. This is equivalent to $\mu^2 = 3600$, where μ^2 is the mean of the square of the failed stated index variable.
7. According to the t-test, we reject the null hypothesis (p-value = 0.000031). Thus, μ^2 is not 3600. Since μ^2 is not 3600, we know μ is not 60.
8. Also, we are given that the 95% confidence interval for μ2 is from 4671 to 6375. This means the 95% confidence interval for μ is from 68.3 to 79.8.
9. From all of this, we can conclude

> ***According to the t-test, based on the square of the failed states index, we can conclude that the mean failed states index is not 60 (p-value = 0.000031). In fact, we are 95% confident that the mean failed states index is between 68.3 and 79.8, with a point estimate of 74.3.***

RESEARCH HYPOTHESIS 4: PRODUCTION RATIO II

Let us now do some two-sample tests. Here, let our claim be that the mean production ratio for states who have nationalized their production is less than that for states who have not nationalized (are private). In symbols, this claim is $\mu_n < \mu_p$. That means the three hypotheses are

$$H_R : \mu_n < \mu_p$$

$$H_0 : \mu_n \geq \mu_p$$

$$H_A : \mu_n < \mu_p$$

Here are the usual questions. How many populations? What parameter? What test? What assumptions?

In this case, we are comparing ***two*** populations. The first population is the nationalized states. The second population is the privatized states. The population parameter remains the mean. The best test is the two-sample t-test. It assumes ***each group*** of measurements comes from a Normal population.

1. First, let us look at a box-and-whiskers plot of the data

```
boxplot(productionRatio~nationalized)
```

2. Notice that the median of the privatized sample is lower than that of the nationalized group. Both have an outlier. From the box-and-whiskers plot, it definitely does not look as though either group is Normal.
3. The Shapiro-Wilk test is still the statistical test for determining Normality. However, the version is slightly different. Here is its use:

```
shapiroTest(productionRatio~nationalized)
```

4. According to the Shapiro-Wilk test, neither population is Normally distributed. Thus, we cannot use the two-sample t-test. The two-sample Mann-Whitney test is the alternative to the two-sample t-test. It is robust to violations of its assumption. Because of this, we will use it whenever we cannot use the two-sample t-test.

```
wilcox.test(productionRatio~nationalized)
wilcox.test(productionRatio~nationalized, conf.int=TRUE)
```

5. Both provide the p-value, but only the second provides the confidence interval. Here is the conclusion.

> ***Because the p-value (1.302×10^{-6}) is less than alpha, we reject the null hypothesis. According to the Mann-Whitney test, the average production ratio differs between those countries who have nationalized their petroleum industry and those who have not.***
>
> ***A 95% confidence interval for how much higher the production ratio of nationalized states over privatized states is from 30.6 to 124.2, with a point estimate of 49.4.***

6. How did I know that the nationalized states had a higher average production than non-nationalized states? I had R calculate the mean for each group:

```
aggregate(productionRatio, by=list(nationalized), FUN=mean)
```

7. This is a very handy function to keep in mind.