STATISTICAL COMPUTING ACTIVITY

7b: Hypothesis Testing II

Purpose: This SCA is for your out-of-class use. It gives you yet 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`
- `prop.test`
- `binom.plot`

PROCEDURE

As usual, here is the procedure.

PART O: START-UP

Let us examine a new data file. The `fuf` data file consists of a sample of militarized international disputes (MIDs) from 1980 until 2001. Each record in the data set is a country involved in a specific MID. The variables are measured at the country level. The acronym "fuf" stands for "**f**irst **u**se of **f**orce" A country that uses force first in a MID is a fuffer. A country that has force used on it first is fuffed. One area of quantitative international relations seeks to determine the factors that influence a country's entrance into a MID.

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

```
##### Script for SCA7b
#####


### Preamble

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

# Read in data
dt = read.csv("http://rfs.kvasaheim.com/data/fuf.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: DURABILITY

The variable **`durable`** refers to the number of years since the last major transition in structural power in the country. For the United States, that was in 1809 when the US Supreme Court made it clear that the power of the Federal government is greater than that of any state.

It would be interesting to see if the average durability of fuffers is greater than that of non-fuffers. So, the full research hypothesis is "The average durability of fuffers is greater than that of non-fuffers." Again, ask the usual questions. Here, we are testing a hypothesis about two averages. That tells us that the parametric test we want to use is the two-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 $\mu_0 < \mu_1$, where μ_1 is the mean durability of fuffers and μ_0 is the mean durability of non-fuffers. That means the three hypotheses for this one-tailed test are

$$H_R : \mu_0 < \mu_1$$

$$H_0 : \mu_0 \geq \mu_1$$

$$H_A : \mu_0 < \mu_1$$

Before continuing our procedure, let us look at the data via the side-by-side box-and-whiskers plot to see if we can learn anything just by looking at the picture.

```
boxplot(durable~fuffer)
```

According to the box-and-whiskers plot, there seems to be little difference in mean durability between the two groups. Let us see if the statistical test bears that out.

1. The assumption of the two-sample t-procedure is that the data in each group comes from a Normal population. To test this, let us use the Shapiro-Wilk test. If the data in each group do not come from a Normal population, then the p-value of the Shapiro-Wilk test will be greater than 0.05.

```
shapiroTest(durable~fuffer)
```

2. Because the p-values are much less than 0.05 (each p-value << 0.0001), the assumption is violated. Thus, we cannot use the two-sample t-procedure.
3. The next most powerful procedure for the population mean is the Mann-Whitney procedure. It is highly robust to violations of its assumption, thus we will use it here.

```
wilcox.test(durable~fuffer, alternative="less")
```

4. Note the addition of the alternative. It reflects the inequality in the alternative hypothesis. Note that R uses the alphabet and number to determine which comes first. That is why you know "less" corresponds to the hypothesis $H_A : \mu_0 < \mu_1$.

5. We have a p-value that we can interpret. Here is how we get the confidence interval:

```
wilcox.test(durable~fuffer, conf.int=TRUE)
```

6. With all of that, we have the following conclusion:

> ***According to the Mann-Whitney procedure, we cannot reject the null hypothesis that the mean durability of fuffers is greater than that of non-fuffers. The p-value of 0.8982 is not less than $\alpha = 0.05$.***
>
> ***We are 95% confident that the average durability of non-fuffers exceeds that of fuffers by between -1 and 3 years, with a point estimate of 1 year.***

7. In other words, fuffers do not appear to have a higher durability than those they fuf.

Research Hypothesis 2: Durability II

In the previous example, we concluded that fuffers do not appear to have a higher durability than those they fuf. It may be interesting to see if there is a difference between those who are fuffed and those who are not. Remember, some countries may enter a conflict late. They would be neither a fuffer or the fuffed.

Thus, the full research hypothesis is "The average durability for fuffed states is the same as for unfuffed states." Here, the claim is $\mu_0 = \mu_1$, where μ_0 represents the average duration for states that were not fuffed, and μ_1 represents that for states who were fuffed. Here are the three hypotheses for this two-tailed test.

$$H_R : \mu_0 \neq \mu_1$$

$$H_0 : \mu_0 = \mu_1$$

$$H_A : \mu_0 \neq \mu_1$$

Now, let us ask our usual questions. How many populations? What population parameter? What is the optimal test? What are its assumptions? There are two populations. We are making a claim on the relationship between the population means. The optimal test is the t-test. The t-test requires Normality. Let us test Normality.

1. The box-and-whiskers plot indicate that the data are positively skewed. According to the Shapiro-Wilk test, the data do not come from a Normal population (both p-values << 0.0001). Thus, the t-test cannot be used.

2. The second test is the Mann-Whitney test. Again, it is robust to violations of its assumptions. As such, that will be our test choice.

```
wilcox.test(durable~fuffed)
wilcox.test(durable~fuffed, conf.int=TRUE)
```

3. And so, from these results, we can conclude

> ***According to the Mann-Whitney test, we cannot reject the null hypothesis that the mean durability for fuffed states is the same as for non-fuffed states (p-value = 0.6527).***
>
> ***In fact, we are 95% confident that the mean durability for fuffed states is between -1 and 3 years less than that of non-fuffed states, with a point estimate of 1 year.***

Research Hypothesis 3: Failed States Index, II

Ok. It seems as though durability is not a factor in whether or not a country fufs or is fuffed upon. Perhaps there is a difference in terms of major power status. That is, are major power more likely or less likely to be fuffers? A lot of theory suggests that they are more likely to be fuffers because they are more powerful and willing to use force in a conflict. I am not at all sure about this theory, so let us just test the hypothesis that they differ.

1. The claim is $p_0 \neq p_1$, where p_0 is the proportion of non-major powers who fuf and p_1 is the proportion of major powers who fuf. This leads to the three hypotheses:

$$H_R : p_0 \neq p_1$$

$$H_0 : p_0 = p_1$$

$$H_A : p_0 \neq p_1$$

2. Thus, we are dealing with two populations and proportions. The correct test is the proportions test. The proportions test requires knowing the number of successes and trials for each of the two groups. To get those numbers, use the **table** function.

```
table(majpower,fuffer)
```

3. Thus, the number of successes is 380 and 45, while the number of trials is 1351 and 273, respectively for non-major powers and major powers. Before we perform the test, however, let us look at the usual plot of this data. For a 2×2 matrix, the usual plot is the binomial plot.

```
binom.plot(x=c(380,45), n=c(1351,273), names=c("Non-Major Power","Major
    Power"))
```

4. I like the binomial test because it shows the actual estimates and 95% confidence interval for each estimate.

5. The code to perform the proportions test is

```
prop.test( x=c(380,45), n=c(1351,273) )
```

6. The test produces the results of the hypothesis test of $p_1 = p_2$ against the alternative $p_1 \neq p_2$. The p-value is much less than $\alpha = 0.05$, so we reject the null hypothesis in favor of the alternative. We have shown that major powers fuf at a different rate than other countries.

7. But, which group fufs more? In the output from the proportions test, we are also given the sample proportions: **prop 1** = 0.2812731; **prop 2** = 0.1648352. So, major powers fuf at a lower rate than other countries. In fact, we are 95% confident that major powers fuf at a rate between 6.4% and 16.9% lower than non-major powers, with a point estimate of 28.1% – 16.5% = 11.6%.

8. How did we know prop 1 belonged to the non-major powers? Look at the major powers variable (**majpower**). It has two levels: 0 and 1. Numerically speaking, 0 comes before 1. So, **prop 1** belongs to **majpower = 0** and **prop 2** belongs to **majpower = 1**.

9. From all of this, we can conclude

According to the proportions test, major powers fuf at a lower rate than non-major powers (p-value = 0.00008983). In fact, we are 95% confident that the major power fuf at a rate of between 6.4% and 16.9% lower than non-major powers, with a point estimate of 11.6%.

RESEARCH HYPOTHESIS 4: MAJOR POWER II

From the previous research analysis, we discovered that major powers tend to fuf less frequently than non-major powers. International relations scholars think in terms of revisionist states and status quo states. Revisionist states want to change the world, while status quo states want it to remain the same. The results in 3 support this theory.

However, does major power status also affect the probability a state is fuffed? The hypothesis seems to be that there is a significant difference. That is,

$$H_R : p_0 \neq p_1$$

$$H_0 : p_0 = p_1$$

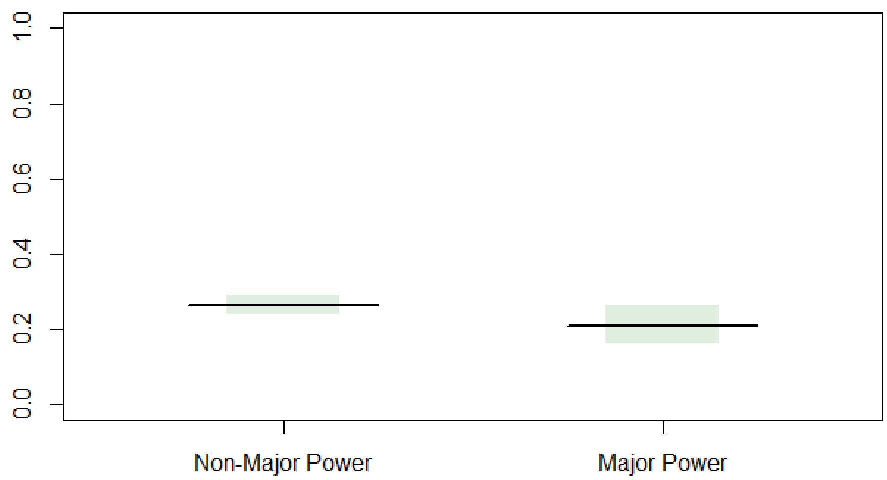
$$H_A : p_0 \neq p_1$$

Here, p_0 is the proportion of non-major power states who are fuffed, and p_1 is the proportion of major power states that are fuffed.

1. Make sure you create an appropriate plot. To do this, you need the numbers off the table

```
table(fuffed, majpower)
```

2. With those numbers, you can make this binomial plot.



3. It still looks like non-major powers are fuffed at a higher rate than major powers, but the margin is much narrower. Let us see what the statistics tell us.

According to the proportions test, there is not a significant difference in the rate that major powers and non-major powers are fuffed (p-value = 0.06491). We are 95% confident that the non-major powers get fuffed from -0.02% to 11.14% more frequently than major powers, with a point estimate of 5.5%.

Research Hypothesis 5: Autocratic Government

The previous four hypotheses have taught us something about the factors that may influence the decision to use force first and the result of waiting too long. This hypothesis examines the effect of the government being autocratic on the probability of fuffing another state.

The three hypotheses are

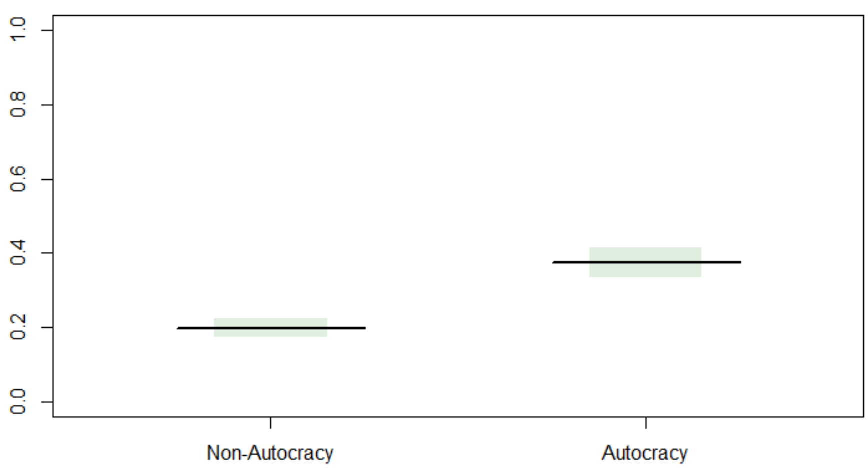
$$H_R : p_0 \neq p_1$$

$$H_0 : p_0 = p_1$$

$$H_A : p_0 \neq p_1$$

Here, p_0 is the proportion of non-autocratic states that fuf, and p_1 is the proportion of autocratic states that fuf. Again, this is comparing two population proportions. Thus, it will follow the same process as the previous two hypothesis tests.

Here is the binomial plot:



Here is the full conclusion.

According to the proportions test, there is a significant difference in the rate that autocracies and non-autocracies fuf (p-value << 0.0001). We are 95% confident that the autocracies fuf from 12.9% to 22.4% more frequently than non-autocracies, with a point estimate of 17.6%.

Make sure that your script produces that binomial plot and comes up with the right answers. Make sure that you can write out these conclusions. It is very important that you understand the wording.