R Guide

This document is a guide to some helpful R commands. I recommend using scripts in R Studio to do things like homework. A script can be created by going to File > New File > R Script. It will appear as a blank document in the upper left quadrant. It will likely appear as a red tab called "Untitled". Scripts are nice because you can save them like any other document to look at later or reuse for other tasks or homeworks. If you only enter commands into the console, that code won't be saved unless you cut and paste it all.

This is an Rmarkdown document that will have both the commands entered into R and the output that would appear in the console.

Importing Data

There are two ways to import data. The first way is to use the "Import Dataset" option on the upper right of R Studio. The default options are usually fine so click import. In order to work with the data, you'll need to attach it like so:

```
attach(ny2016cces)
```

After attaching the dataset you can just use the variable names when working with a certain variable. To figure out the names of variables in the dataset:

```
names(ny2016cces)
```

```
##  [1] "educ"     "age"      "sex"      "natlecon" "favorssm" "ideology"
##  [7] "partyid"  "lgbt"     "income"   "presvote"
```

The second way involves setting up a working directory. The working directory is the folder in your computer where the dataset you need is located. Windows users need to include a second backslash because computers are weird. Then you would import the data according to the file name. I have also assigned it to the object "data". For my computer it would look like this:

```
setwd("C:\\Users\\Kevin\\Box Sync\\PSC 408\\data")
data <- read.csv("ny2016cces.csv")
```

This way is more flexible but there is a drawback. Instead of just calling a variable by its name, you would need to put the name of the dataset in front of the variable each time you use it. For example, if I wanted a summary of the lgbt variable I would need to use summary(data$lgbt) since I named the dataset "data".

Getting a Sense of the Data

First, look at the codebook for the data. This comes as a .txt file for the CCES data. Most of the data are numbers so the codebook will make sense of those numbers.

We might want to know some simple things. How many men are in the sample? What is the average income? What does the scale for party ID look like? How does education break down by sex? We can answer those questions with some quick commands:

```
table(sex)
```

```
## sex
##    0    1
## 1013 1088
```

```
mean(income)
```

```
## [1] 7.524988
```

```
table(partyid)
```

```
## partyid
##   1   2   3   4   5   6   7
## 768 348 193 187 144 207 254
```

```
table(educ, sex)
```

```
##     sex
## educ   0   1
##    1  13  13
##    2 184 253
##    3 299 324
##    4 287 316
##    5 230 182
```

The basic syntax for R is to have a command, like table, followed by some arguments, like education. What arguments does a certain command need? To figure that out, a handy command is to use ? before the commands name. A box of information about that specific command will appear in the lower right.

```
?table
```

Finding some p-values

To test some null hypotheses associated with the data we will need to run t tests or chi-square tests. If I wanted to test the null hypothesis that 10% of New Yorks population identifies as LGBT, I would use the following code:

```
t.test(lgbt, mu = .10)
```

```
##
##  One Sample t-test
##
## data:  lgbt
## t = 3.122, df = 2100, p-value = 0.001821
## alternative hypothesis: true mean is not equal to 0.1
## 95 percent confidence interval:
##  0.1083007 0.1363447
## sample estimates:
## mean of x
## 0.1223227
```

t.test is the command used in R to run a t.test. lgbt is used because that is the name of the variable in the CCES data set. Lastly, I set mu to .10 because that is instructing the t.test that I want the null hypothesis to be .10. The very low p-value listed in the output suggests that we can reject the null hypothesis. Note that the results also include a 95% confidence interval.

Maybe I want to argue that the null hypothesis is 10% or less of New Yorkers identify as LGBT. In order to test this hypothesis, I need to add another argument to the t.test command. From looking at the help for t.test I can see that I can specifiy an alternative hypothesis of "less" or "greater". Since the null is 10 or less, I shall use "greater".

```
t.test(lgbt, mu = .10, alternative = "greater")
```

```
##
##  One Sample t-test
```

```
##
## data:  lgbt
## t = 3.122, df = 2100, p-value = 0.0009103
## alternative hypothesis: true mean is greater than 0.1
## 95 percent confidence interval:
##  0.1105567       Inf
## sample estimates:
## mean of x
## 0.1223227
```

Again there is a low p-value suggesting that the null hypothesis can be rejected. So we might conclude that more than 10% of New York's population identifies as LGBT.

Maybe I want to test the null hypothesis that there is no difference in ideology scores between LGBT identifiers and non-identifiers. I change the arguments in the t.test command to indicate that I want to know about ideology based on the values of the lgbt variable:

```
t.test(ideology~lgbt)
```

```
##
##  Welch Two Sample t-test
##
## data:  ideology by lgbt
## t = 7.0581, df = 321.3, p-value = 1.045e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.6371168 1.1295594
## sample estimates:
## mean in group 0 mean in group 1
##        3.863883        2.980545
```

Low p-value and the fact that the confidence interval does not include 0 leads me to conclude that I can reject the null hypothesis of no difference in ideology between the two groups.

Are ideology and LGBT identification independent? We can use a chi-square test for independence to determine this. First, make a crosstab of ideology and lgbt and assign it to an object. Then, run a chi-square analysis on that object:

```
crosstab <- table(ideology, lgbt)
chisq.test(crosstab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  crosstab
## X-squared = 90.997, df = 6, p-value < 2.2e-16
```

The results from the chi-square test show a very low p-value. Thus, we can conclude that ideology and LGBT identification are not independent.