PS0700 Fall2022 Recitation Week 12

Jungmin Han

11/16/2022

1. Announcement

- The second mid-term's grade will be posted on 11/18.
- R assignment #2 is due by 11/20.
- R assignment #3 will be posted on Canvas on 11/18, which will be due by 11/30.
- R assignment #4 will be the last extra credit assignment, which will be posted on 12/2.
- An additional assignment for recitation make-up will be open on 11/18, which is due by 12/11.
- Research paper will be due by 12/11.

2. Review

2.1. Working with R Markdown files

- R Markdown files consist of two parts 'texts' and 'rcode.'
- To insert Rcode chucks, click **Insert** then choose R.
- You can run each code chunk by clicking the green icon above. RStudio executes the code and display the results inline with your file.
- After work, Clink **Knit** button to generate a document.

2.2. Creating Objects in R

- R stores information in the form of objects.
- We use the assignment operator (`<-`) to create an object in R
- All objects are shown in the environment (the upper-right window)

2.3. R symbolds and operators

- `<-` Assignment operator used to create new objects
- `#` used to comment code; R ignores everything that follows it
- `$` used to access an element inside an object, such as a variable inside a dataframe (`data$variable`)

2.4. R functions

- `read.csv()` reads CSV files (`read.csv("filename.csv")`)
- `read_dta()` reads dta files (`read_dta("filename.dta")`)
- `head()` shows the first six observations in a dataframe (`head(Data)`)
- `dim()` provides the dimensions of a dataframe (`dim(Data)`)
- `mean()` calculate the mean of a variable (`mean(Data$Variable)`)
- `==` evaluates whether two values are equal to each other.
- `ifelse` creates the values of a new variable based on an existing one
- `[ ]` Square brackets can be used to extract a selection of observations from a variable.

3. T-test

3.1. Load the World Value Survey 7 Data and assign it to the object `wvs7`.

- Because the type of this data is `rdata`, we are using `load()` function. Please put the location and name of data file (WVS_Cross-National_Wave_7_R_v2_0.rdata) into the function correctly.
- Once you successfully load the data, it will be shown in your environment with the original name of the data (WVS_Cross-National_Wave_7_v2_0). Now, we are assigning this to a new object, called `wvs`.

```r
setwd("/Users/jungminhan/Desktop/University_of_Pittsburgh/Teaching/Fall 2022/PS0700/ResearchPaper/Data"
load(file = "Wave 7/WVS_Cross-National_Wave_7_R_v2_0.rdata")
wvs7 <- `WVS_Cross-National_Wave_7_v2_0`
# We need ` ` in order to call the data's name
# because it has - (minus) sign in the middle of it.
```

3.2 Find two interseting variables and create a new dataset.

- Let's say I hypothesize that people who value `security` over `freedom` are likely to have a negative attitude toward immigration.
- Looking at the WVS7 codebook, I find two associated questions: *(1) Q121. How would you evaluate the impact of these people on the development of [your country]?* and *(2) Q150. Most people consider both freedom and security to be important, but if you had to choose between them, which one would you consider more important?*
- Now, I create a new dataset that includes these two variables, then naming them `immigration` and `freedom.vs.security` respectively.

```r
dat <- data.frame(wvs7$Q150, wvs7$Q121)
colnames(dat) <- c("freedom.vs.security", "immigration")
```

3.3 Data Cleaning

- By looking at the WVS7 codebook, what do you find? It seems that the numerical values below zero for both variables are not useful for our analysis. So, I remove these from my analysis by assigning `Not Applicable (NA)` to them.
- Also, I omit all the rows with `NA` from my new data by using `na.omit()` function.

```r
dat$freedom.vs.security[dat$freedom.vs.security < 0] <- NA
# 1: Freedom; 2: Security
dat$immigration[dat$immigration < 0] <- NA
# Higher number: Better for one's country
dat <- na.omit(dat)
```

3.4 Statistical Test: T-test

- Here, I am using `t.test()` function to test if the mean difference of attitudes toward immigration between two groups (security vs. freedom) statistically significant.
- As you see from the syntax below, you put your dependent variable first, put tilda (~), and put your independent variable.
- The result of the following **t-test** indicates that we reject the null hypothesis that they is no meaning difference between two groups in terms of their attitudes toward immigration.
- Here, t value is 9.7032, and this is larger than 1.96 standard error which uses with 0.05 significance level.
- Also, R output provides the exact probability of observing a value of `t` of the size (or greater) that we did observe if the null hypothesis value of 0 were true (here $p < 2.2e-16$). If this probability (p) is less than .05, we reject the null.

```
t.test(dat$immigration ~ dat$freedom.vs.security)
```

```
##
##  Welch Two Sample t-test
##
## data:  dat$immigration by dat$freedom.vs.security
## t = 9.7032, df = 40570, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  0.06687801 0.10073566
## sample estimates:
## mean in group 1 mean in group 2
##        3.026114        2.942307
```

4. Cross-tabulation

4.1. Load the World Value Survey 7 Data and assign it to the object wvs7.

- Because the type of this data is `rdata`, we are using `load()` function. Please put the location and name of data file (WVS_Cross-National_Wave_7_R_v2_0.rdata) into the function correctly.
- Once you successfully load the data, it will be shown in your environment with the original name of the data (WVS_Cross-National_Wave_7_v2_0). Now, we are assigning this to a new object, called `wvs`

4.2 Find two interseting variables and create a new dataset.

- Let's say I hypothesize that people who directly experienced civil war before value `security` over `freedom`.
- Looking at the WVS7 codebook, I find two associated questions: *(1) Q144 Have you been the victim of a crime during the past year?* and *(2) Q150. Most people consider both freedom and security to be important, but if you had to choose between them, which one would you consider more important?*
- Now, I create a new dataset that includes these two variables, then naming them `civil.war` and `freedom.vs.security` respectively.

```
dat <- data.frame(wvs7$Q150, wvs7$Q144)
colnames(dat) <- c("freedom.vs.security", "civil.war")
```

4.3 Data Cleaning

- By looking at the WVS7 codebook, what do you find? It seems that numerical values below zero for both variables are not useful for our analysis. So, I remove these from my analysis by assigning `Not Applicable (NA)` to them.
- Also, I omit all the rows with `NA` from my new data by using `na.omit()` function.

```
dat$freedom.vs.security[dat$freedom.vs.security < 0] <- NA
# 1: Freedom; 2: Security
dat$civil.war[dat$civil.war < 0] <- NA
# 1: Yes; 2: No
dat <- na.omit(dat)
```

4.4 Statistical Test: Cross-tabulation

- In order to run `CrossTable()` function, you need to install `gmodels` library and run it with `library()` function.

- Then, you put your dependent variable first and independent variable later. There are several arguments that you can choose to manage the information shown in your outcome. Here, I want to see (1) observed number, (2) expected number (`expected = TRUE`), and (3) percentage in a column (`prop.c = TRUE`) in each cell. Also, I want to see the chi-square statistics (`chisq = TRUE`). Other than these, I put `FALSE` for arguments (You can find more details by typing `?CrossTable` in the console or help window.).
- The result shows that people who previously experienced civil war are more likely to value *freedom* over *security*, which is opposite to my expectation. This direction can be examined by comparing the observed and expected cases in each cell. For example, in the first cell, the expected number of people who experienced civil war and value freedom is 2072.096, but we actually observe 2412. This indicates their positive relationship. The same logic can be applied to the other cells.
- Then, is this outcome statistically significant? To evaluate this, you need to look into p-value. Pearson's Chi-squared test tells us that `p =  9.106402e-21`, where p-value is far smaller than 0.05. So, we reject the null hypothesis saying that there is no relationship between two variables.

```r
# install.packages("gmodels")
library(gmodels)
CrossTable(dat$freedom.vs.security, dat$civil.war,
           expected = TRUE, prop.c = TRUE,
           prop.r = FALSE, prop.t = FALSE,
           prop.chisq = FALSE, chisq = TRUE)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |              Expected N |
## |           N / Col Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  75475 
## 
## 
##                         | dat$civil.war 
## dat$freedom.vs.security |         1 |         2 | Row Total | 
## ------------------------|-----------|-----------|-----------|
##                       1 |      2412 |     19968 |     22380 | 
##                         |  2072.096 | 20307.904 |           | 
##                         |     0.345 |     0.292 |           | 
## ------------------------|-----------|-----------|-----------|
##                       2 |      4576 |     48519 |     53095 | 
##                         |  4915.904 | 48179.096 |           | 
##                         |     0.655 |     0.708 |           | 
## ------------------------|-----------|-----------|-----------|
##            Column Total |      6988 |     68487 |     75475 | 
##                         |     0.093 |     0.907 |           | 
## ------------------------|-----------|-----------|-----------|
## 
## 
## Statistics for All Table Factors
## 
## 
## Pearson's Chi-squared test 
## ------------------------------------------------------------
## Chi^2 =  87.34687     d.f. =  1     p =  9.106402e-21
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  87.09009     d.f. =  1     p =  1.03689e-20
##
##
```