Info 271b
Fall 2020
Lab 3

**Overview:**
This lab has a series of tasks using R. You will provide the answers to the questions below in a single document (your lab report). You should include any key output and/or graphics in this lab report. In addition, include your fully commented R script along with your submission (we should understand your answers without having to read the R script).

**Working on the Lab and Policy about Working with Others:**
You will write and comment your own final script to turn in. However, talking about and troubleshooting R scripts with other students is permitted. Your lab report should only be written by you and you should not discuss your answers to the questions with other students.

**Submission:**
Please submit the following:

- A PDF of your lab report named "lab3_LastName_FirstName.pdf"
- Your Rmd script named "lab3_LastName_FirstName.Rmd"
- Html output from Rmd file, named "lab3_ LastName_FirstName.html"

This lab is due **Tuesday, November 24th by 11:59 pm**. You should upload your three files (the lab report, Rmd and html files) to bCourses.

**Planning ahead would ensure you enjoy Thanksgiving week without having to worry about this lab submission at the last moment on Tuesday.**

To create a new Rmd file, go to File -> New File -> R Markdown. In RStudio, your open Rmd file will have a "Knit" option in the toolbar. When ready to output, select knit -> Knit to Html.

**Data Preparation, Analysis and Interpretation (100 Points)**

*Dataset (GSS)*:
Every other year, the General Social Survey collects responses to thousands of questions, covering a wide variety of topics. You will be using a subset of data from 2016, including a small number of variables. Please use the dataset under **bcourses -> Files -> Labs -> Lab 3 -> Lab3_GSS.Rdata**. This is a similar dataset to the one we've been using in class examples.

While some variables may be self-explanatory, others may not make sense until you look at the GSS codebook. An easy way to investigate a variable is to look it up in the GSS mnemonic index, located at: https://gssdataexplorer.norc.org/variables/vfilter

**Before you run a statistical test on a variable, you should always read its description in the codebook, in order to understand what different values mean.** For example, the codebook may explain if certain values stand for missing data. If this occurs, you should make sure those values are recorded as NA in R before proceeding.

Write a well-commented Rmd script to perform each of the following tasks, following the best practices described in class. For each new variable, look for obvious errors and make sure that appropriate values are coded as NA.

**Include all important output and answers to each question in your main lab report. If needed, you should also copy any graphics into the main lab report to make it easier for you to provide context for your answers. <u>We should be able to understand what you did and what your answer is for each item in your main lab report without hunting for things in your Rmd script.</u>**

For all hypothesis testing, please use 5% as your chosen level of significance.

You'll need the car, gmodels and lsr packages (see class code examples), so make sure you install them and load them into your library using *install.packages()* and *library().* You only need to run your *install.packages()* lines once – once installed, comment out or delete them in your .Rmd file after to avoid any knitting issues.

1. Task 1: Create a new factor variable, *degree_factor*, by recoding the values from *degree.* For *degree_factor*, referring to the codebook, set "Lt high school" and "High school" to "HIGH SCHOOL OR LESS", "Junior college" and "Bachelor" to "COLLEGE", and "Graduate" to "GRADUATE". All other values should be set to NA for your *degree_factor* variable. Use function *table()* to make sure you've created your *degree_factor* variable correctly.

   Next, also create a new factor variable, *gun_factor,* from the *gunlaw* variable. Your *gun_factor* variable should only have 3 categories – "Favor", "Oppose" and "Don't know." All other values should be set to NA for *gun_factor.* Note that while we often set "Don't know" values to NA, we retain "Don't know" values in this analysis because we are also interested in whether *degree_factor* is associated with respondents choosing "Don't know" for your *gun_factor* variable.

   Now that you have your new factor variables, conduct a Chi-square test of independence to examine the association between *degree_factor* and *gun_factor*.

   **A. What are the null and alternative hypotheses for your test?**

   - Null Hypothesis: The gun law attitude <u>is not related</u> to the degree level(that it is independent).

- Alternative Hypothesis: The gun law attitude is related to the degree level.

**B. What test statistic and p-value do you get?**

```
Number of cases in table: 908
Number of factors: 2
Test for independence of all factors:
        Chisq = 16.747, df = 4, p-value = 0.002164
        Chi-squared approximation may be incorrect
```

```
        Fisher's Exact Test for Count Data

data:  (table(df$degree_factor, df$gun_factor))
p-value = 0.0006739
alternative hypothesis: two.sided
```

p-value before the fisher's exact test: 0.002

p-value after the fisher's exact test: 0.0006, which is highly statistically significant. Chi-square value of 16.747.

**C. Is the Fisher's Exact Test necessary? Why or why not? If necessary, run Fisher's test and report the p-value from the test.**

- Yes, even though we have a big sample, but we have some expected values that are less than 5.

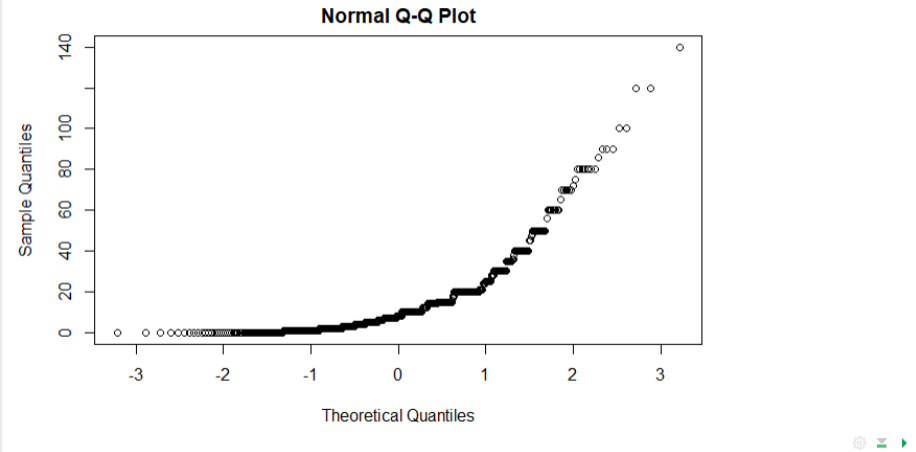
```
                    Don't know     Favor    Oppose
COLLEGE               3.579295 172.63216  73.78855
GRADUATE              1.474670  71.12445  30.40088
HIGH SCHOOL OR LESS   7.946035 383.24339 163.81057
```

**D. Conduct an appropriate effect size calculation for the Chi-square test of independence.**

```
```{r}
#run the approriate effect size test using Cramers'v
cramersV(df$degree_factor, df$gun_factor)
```
```

```
Chi-squared approximation may be incorrect[1] 0.09603083
```

Use the Cramer's V as the effect size test since we have more than 2 choices for each category, The Cramer's V value is 0.09, which is small.

**E. Examine the standardized residuals in the cross-table. What cells, if any, are especially practically significant (and how do you know)?**

| df$degree_factor | df$gun_factor Don't know | Favor | Oppose | Row Total |
|---|---|---|---|---|
| COLLEGE | 3 | 169 | 78 | 250 |
| | 3.579 | 172.632 | 73.789 | |
| | 0.094 | 0.076 | 0.240 | |
| | 1.200% | 67.600% | 31.200% | 27.533% |
| | 23.077% | 26.954% | 29.104% | |
| | 0.330% | 18.612% | 8.590% | |
| | -0.306 | -0.276 | 0.490 | |
| GRADUATE | 1 | 89 | 13 | 103 |
| | 1.475 | 71.124 | 30.401 | |
| | 0.153 | 4.493 | 9.960 | |
| | 0.971% | 86.408% | 12.621% | 11.344% |
| | 7.692% | 14.195% | 4.851% | |
| | 0.110% | 9.802% | 1.432% | |
| | -0.391 | 2.120 | -3.156 | |
| HIGH SCHOOL OR LESS | 9 | 369 | 177 | 555 |
| | 7.946 | 383.243 | 163.811 | |
| | 0.140 | 0.529 | 1.062 | |
| | 1.622% | 66.486% | 31.892% | 61.123% |
| | 69.231% | 58.852% | 66.045% | |
| | 0.991% | 40.639% | 19.493% | |
| | 0.374 | -0.728 | 1.031 | |
| Column Total | 13 | 627 | 268 | 908 |
| | 1.432% | 69.053% | 29.515% | |

- **Standard Residual measures the strength of the difference between observed and expected values;**
- For the graduate degree level, both favor and oppose has a high Standard Residual(because the difference between the observed value and expect value is large in both cells), which have a bigger practical effect contributed to the overall significance level, while the other cells has a smaller difference between these 2 values.

**F. Evaluate your hypothesis in light of your tests of statistical and practical significance. What, if anything, can you conclude from your results?**

- The p-value is less than the significance level of 5%, we can reject the null hypothesis and accept the alternative hypothesis that, there is a relationship between the two categorical variables (degree and gun law attitude). However, the effect size is pretty small, so this correlation might not be significant. We then explore more of the individual cells, that we find the Graduate degree level has a higher Standardize Residue which contribute to the overall statistical significance.

2. Task 2: Conduct a correlation analysis to examine the association between years of education (*educ*) and number of children (*childs*). Make sure to set any missing values for *educ* and *childs*, but keep any values that are grouped above some number (e.g., 8 or more children should be left in the data as value 8).

   **A. What are the null and alternative hypotheses for your test?**
   - Null: There is no correlation between the 2 categorical variables. (education and number of children)
   - Alternative: There is correlation between the 2 categorical variables.

   **B. What test statistic and p-value do you get?**

```
        Pearson's product-moment correlation

data:  df2$educ and df2$childs
t = -7.6453, df = 907, p-value = 5.306e-14
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3061775 -0.1839724
sample estimates:
       cor
-0.2460525
```

   T-test of -7.6453 suggests a negative relation between 2 categories, with a highly significant p-value.

   **C. What is the practical significance of your correlation?**

   - The correlation coefficient is small, with a negative direction, that when the education increase, the number of child's variable decrease.

   **D. Evaluate your hypothesis in light of your tests of statistical and practical significance. What, if anything, can you conclude from your results?**

   - We have a highly statistically significant p-value, reject the null, accept that there's a relation between education level and number of children; however, since we have lots of data input, it's easy to spot relationship with a statistical significance, by looking at the correlation coefficient, it points out there is a small negatively relationship between the 2 factors.

Task 3: Create a new factor variable, *exp_factor*, from the *expdesgn* variable (knowledge about experiment design). For *exp_factor*, we only want two categories – "No Control" and "Control." Set "All 1000 get the drug" to "No Control" and "500 get the drug 500 dont" to "Control". Set all other values to NA for *exp_factor.* Last, make sure to set appropriate missing values for our metric variable, *wwwhr*.

Check assumptions and conduct an appropriate t-test to examine the association between *wwwhr* (hours per week using Internet outside email) and *exp_factor* (whether someone understands experimental controls or not).

**A. What assumptions did you test and did this affect the t-test that you conducted?**

- Assumption of normality： with a large sample we can use the CLT to assures that the sample means will approach normality regardless of the underlying data(which based on the highly statistical p-value from the Shapiro-Wilk normality test as well as the QQ plot, the distribution is not normal), so we are not going to be concerned about the normality of our metric variable.

```
        Shapiro-Wilk normality test

data:  df3$wwwhr
W = 0.70613, p-value < 2.2e-16
```

```
```{r}
qqnorm(df3$wwwhr)
```
```




Test the assumption of equal variances with Levene's Test; slightly significant difference between the group variances which means reject the null, that there are differences between the group variances. Therefore, we will run the Welch's t - test for unequal variances.

```r
leveneTest(df3$wwwhr ~ df3$exp_factor, df3, center = mean)
```

```
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value   Pr(>F)
group   1  7.9638 0.004899 **
      747
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**B. What are the null and alternative hypotheses for your test?**

- Null: People who choose the Control option spend more time online.
- Alternate: People who choose the Control option do not spend more time online.

**C. What test statistic and p-value do you get?**

```
        Welch Two Sample t-test

data:  df3$wwwhr by df3$exp_factor
t = 3.3301, df = 261.12, p-value = 0.0004967
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 2.332416      Inf
sample estimates:
   mean in group Control mean in group No Control
                15.03722                 10.41221
```

- Since it’s a one tailed test with a statistically significant p-value, we can look at the group mean of hours spend online, found out that the mean is higher in the Control group than the No control group. T is positive, therefore the relationship between the Control and the number of hours spend online is positive, which support our null hypothesis.
- however, since we have a highly statistical p-value, we reject the null hypothesis.

**D. Calculate Cohen’s d and effect size correlation *r*. Are these effect sizes large?**

```r
t <- ind.t.test$statistic[[1]]
df <- ind.t.test$parameter[[1]]
```

```r
r <- sqrt(t ^ 2/(t^2 + df))
```

```r
round(r, 3)
```

```
[1] 0.202
```

```r
cohensD(df3$wwwhr ~ df3$exp_factor)
```

```
[1] 0.2529058
```

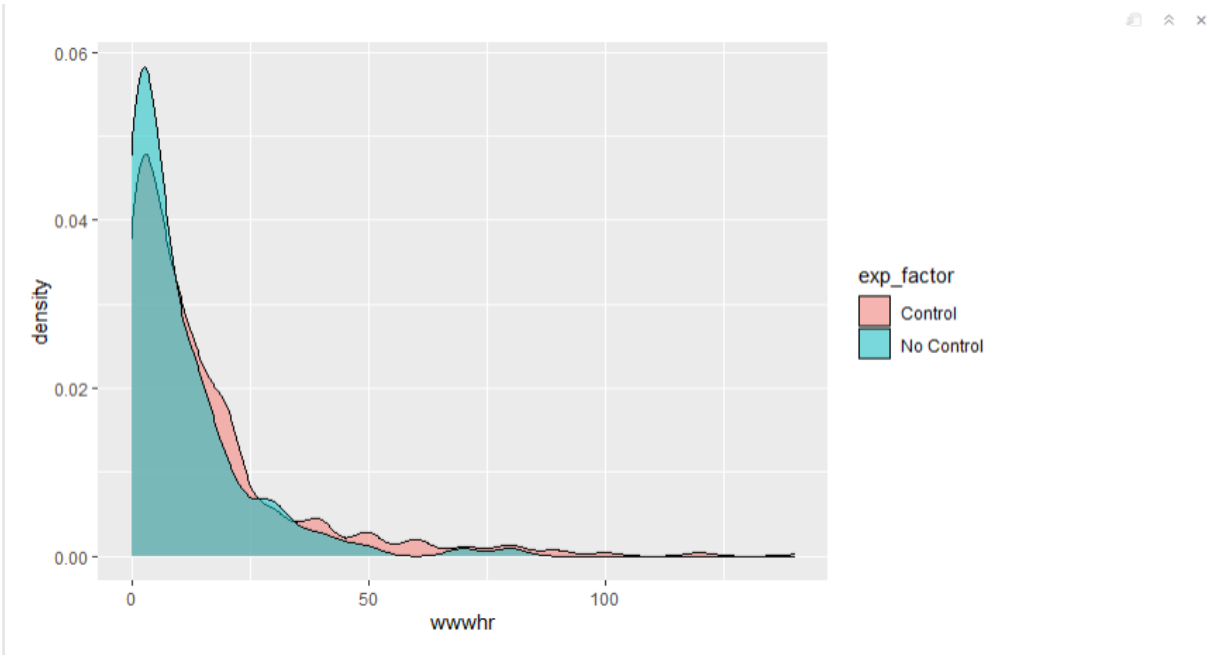
Small effect size based on the effect size calculation and Cohen's d, therefore not that much of a difference between the 2 groups.

**E. Evaluate your hypothesis in light of your tests of statistical and practical significance. What, if anything, can you conclude from your results?**

Even by looking at the mean, we see a slightly higher mean in the Control Group in their time spend online, but since we have a highly significant p-value so we reject the null hypotheses that people choose to use control group for the experiment spend more time online;

By calculating the effect size, we see a rather small effect size, which points to a not that much of a difference time spend online between these 2 groups.

By plotting the density graphs for both groups, we see lots of overlap which supports the small effect size between groups.


0.06
0.04
0.02
0.00
density
0
50
100
wwwhr
exp_factor
Control
No Control