Statistical inference with the GSS data

Setup

Load packages

```
#devtools::install_github("haleyjeppson/ggmosaic")
#install.packages('vcd')
#install.packages('pastecs')
#install.packages(c("ggplot2", "ggpubr", "tidyverse", "broom", "AICcmodavg"))

library(ggplot2)
library(dplyr)
library(statsr)
library(readr)
library(ggmosaic)
library(vcd)
library(pastecs)
library(skimr)
library(ggpubr)
library(tidyverse)
library(broom)
library(AICcmodavg)
library(car)
```

Load data

```
load("gss.Rdata")
```

Part 1: Data

How are the observations in the sample collected?

According to the official GSS website, the GSS is a “nationally representative” survey of adults in the United States. It utilizes an “area probability design that randomly selects respondents in households across the United States.”

Due to this, it’s safe to assume that the data collected is representative of the population of adults at large. However, it’s likely that due to the nature of data collection, causal connections cannot be inferred, but the results can be generalized.

Part 2: Research questions

Investigating the data

Potentially interesting variables to investigate further

1. Religiosity (relig) and Education (degree). Relationship between level of education and religious affiliation?
2. Religiosity (relig) and Political Party (partyid). Relationship between religiosity and political party? ***
3. Religiosity (relig), Political Party (partyid) and Education level (degree).
4. Is level of education (degree) and family income (coninc) associated? ***
5. News and overall happiness.

```
#### Investigate the dataset

#head(gss)
#colnames(gss)
#names(gss)
#str(gss)
#glimpse(gss)
#skim_without_charts(gss)
#summary(gss)
```

Analysis #1 - Religion and Party Affiliation (Chi-Square Test of Independence)

Part 3: Exploratory data analysis

Does there appear to be a relationship between religiosity and political party affiliation?

The parties contained in this analysis are Democrat, Independent, Republican and Other.

For the purposes of this analysis, anyone affiliated with a religion is considered religious. Those that are nearly or strongly leaning toward a particular political party will be considered associated with that party.

Initial Analysis and Cleaning

```
##### Keep specific columns for analysis.
trimmed_gss_analysis_one <- gss %>%
  select(c(relig, partyid))

##### Check for N/A's.
sapply(trimmed_gss_analysis_one, function(x) sum(is.na(x)))
```

```
##   relig partyid
##     233     327
```

```
##### Remove rows with N/A' (for this analysis, it was acceptable.)
trimmed_gss_analysis_one <- na.omit(trimmed_gss_analysis_one)

##### Recode data for easier analysis.
##### Combine religion and party affiliation values.

trimmed_gss_analysis_one <- trimmed_gss_analysis_one %>% mutate(
    relig = case_match(
      relig,
      "None" ~ "Not Religious",
      .default = "Religious"),
    partyid = case_match(
      partyid,
      "Strong Democrat"~"Democrat",
      "Not Str Democrat"~"Democrat",
      "Ind,Near Dem"~"Democrat",
      "Independent"~"Independent",
      "Ind,Near Rep"~"Republican",
      "Not Str Republican"~"Republican",
      "Strong Republican"~"Republican",
      "Other Party"~"Other"
    ),
    )

##### Rename column headers.
colnames(trimmed_gss_analysis_one) <- c("Religiosity","Political Party")

##### Create contingency table
(partyid_relig_n <- (table(trimmed_gss_analysis_one$Religiosity,
                                   trimmed_gss_analysis_one$'Political Party')))
```

```
## 
##                  Democrat Independent Other Republican
##   Not Religious     3031        1548   195       1306
##   Religious        24795        6891   663      18132
```

```
##### Export contingency table and trimmed data for external Chi-Sq Test of Independence in
        Excel.
write.csv(partyid_relig_n, "party_religion.csv", row.names = TRUE)

#trimmed_gss_analysis_one
```

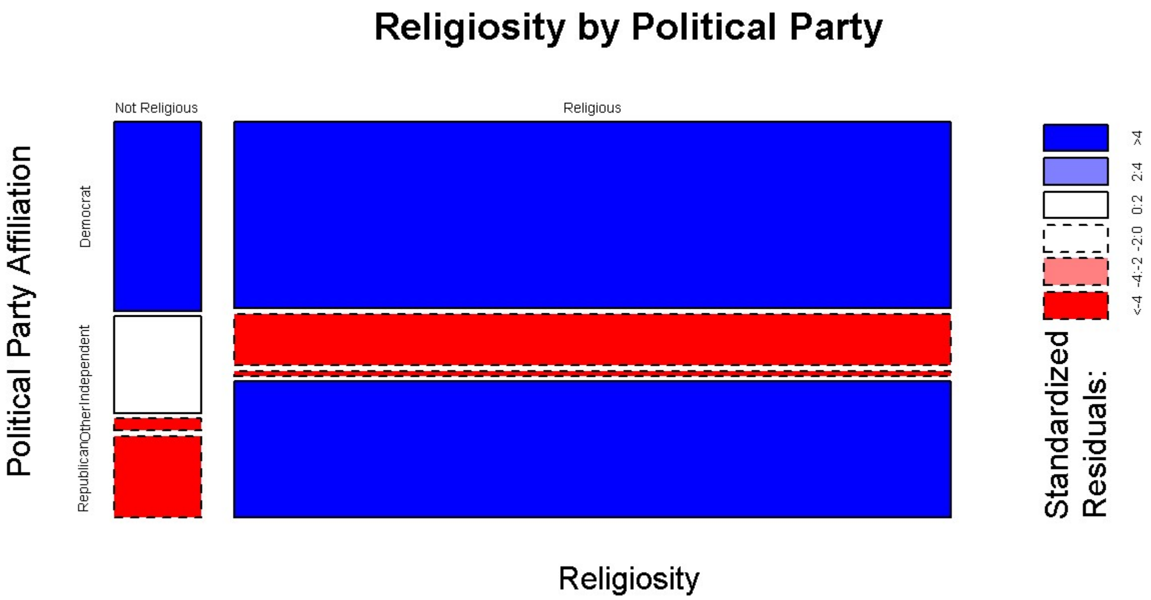
Plots

Mosaic Plot with Residuals

The following plot shows the the association between Political Party affiliation and religiosity. The data indicate that Democrats, Independents and Other see a higher proportion of Non-Religious members, in contrast to Republicans which see the opposite. Based on this, we can determine that political party and religiosity are dependent on one another.

The standardized residuals indicate the difference between the observed and expected values.

```
mosaicplot(partyid_relig_n, main='Religiosity by Political Party',
           col=TRUE,
           xlab = "Religiosity",
           ylab="Political Party Affiliation",
           shade = TRUE,
           off = 4,
           cex.axis = 0.45,
           margin = 1)
```



Summary Statistics

Contingency Tables

The summary statistics indicate a significant difference between the proportion of religious and non-religious individuals affiliated with a particular political party and those that are not

```
(partyid_relig_counts <- (addmargins(table(trimmed_gss_analysis_one$Religiosity,
                                   trimmed_gss_analysis_one$'Political Party'))))
```

```
## 
##                  Democrat Independent Other Republican   Sum
##   Not Religious      3031        1548   195       1306  6080
##   Religious         24795        6891   663      18132 50481
##   Sum               27826        8439   858      19438 56561
```

```
(partyid_relig_prop <- (prop.table(addmargins(table(trimmed_gss_analysis_one$Religiosity,
                                   trimmed_gss_analysis_one$'Political Party')))))
```

```
## 
##                     Democrat  Independent        Other   Republican
##   Not Religious 0.0133970404 0.0068421704 0.0008619013 0.0057725288
##   Religious     0.1095940666 0.0304582663 0.0029304645 0.0801435618
##   Sum           0.1229911069 0.0373004367 0.0037923658 0.0859160906
## 
##                          Sum
##   Not Religious 0.0268736408
##   Religious     0.2231263592
##   Sum           0.2500000000
```

Part 4: Inference

Hypotheses

Does there appear to be a relationship between religiosity and political party affiliation?

H0 (nothing is going on) Religion and Party affiliation are independent. Religiosity does not vary by Political Party.

HA (something is going on) Religion and party affiliation are dependent. Religiosity does vary by Political Party.

Evaluating the Hypothesis

1. Quantify how different the observed counts are from the expected counts.
2. Significant deviations from what is expected based on sampling variation (chance) alone provide strong evidence for the alternative hypothesis.

Conditions for the Chi-Square Test of Independence

Independence:

1. Sampled observations are independent.

- Random sampling was employed for the survey.

2. n < 10% of the population if sampling without replacement.

- Observations are less than 10% of the population.

3. Each case only contributes to one cell in the table.

- It is safe to assume that each individual surveyed reported once, although it’s not impossible someone took part more than once, and therefore is part of multiple cells. For the sake of this analysis, we can assume that each case has only contributed to one cell.

Sample Size:

1. Each cell must have at least 5 expected cases.

- True.

Test the hypothesis

What is the overall religiosity in the sample?

```
50481/55703
```

```
## [1] 0.9062528
```

Test the hypothesis that religiosity and political party affiliation are associated at the 5% significance level.

x^2 = 970.29 df = 3

```
(chiSq <- pchisq(970.29, 3, lower.tail = FALSE))
```

```
## [1] 5.012349e-210
```

```
if(chiSq < 0.05) {
  print("Reject the null hypothesis")
  } else {
    print("Fail to reject the null hypothesis")
  }
```

```
## [1] "Reject the null hypothesis"
```

Conclusion

The p_value of 5.012349e-210 is less than the 0.05 significance level. Therefore, we reject the null hypothesis in favour of the alternative; there is sufficient evidence indicating there is an association between Political Party affiliation and Religiosity.

Analysis #2 - Level of respondents education and household income (ANOVA and pairwise tests (theoretical only))

Part 3: Exploratory data analysis

Does there appear to be a relationship between the level of respondents education and their household income?

The levels of education contained in this analysis are: Limited HighSchool, HighSchool, Junior College, Bachelor, Graduate.

Initial Analysis and Cleaning

```
##### Keep specific columns for analysis.
trimmed_gss_analysis_two <- gss %>%
  select(c(coninc, degree))

##### Check for N/A's.
sapply(trimmed_gss_analysis_two, function(x) sum(is.na(x)))
```

```
## coninc degree
##   5829   1010
```

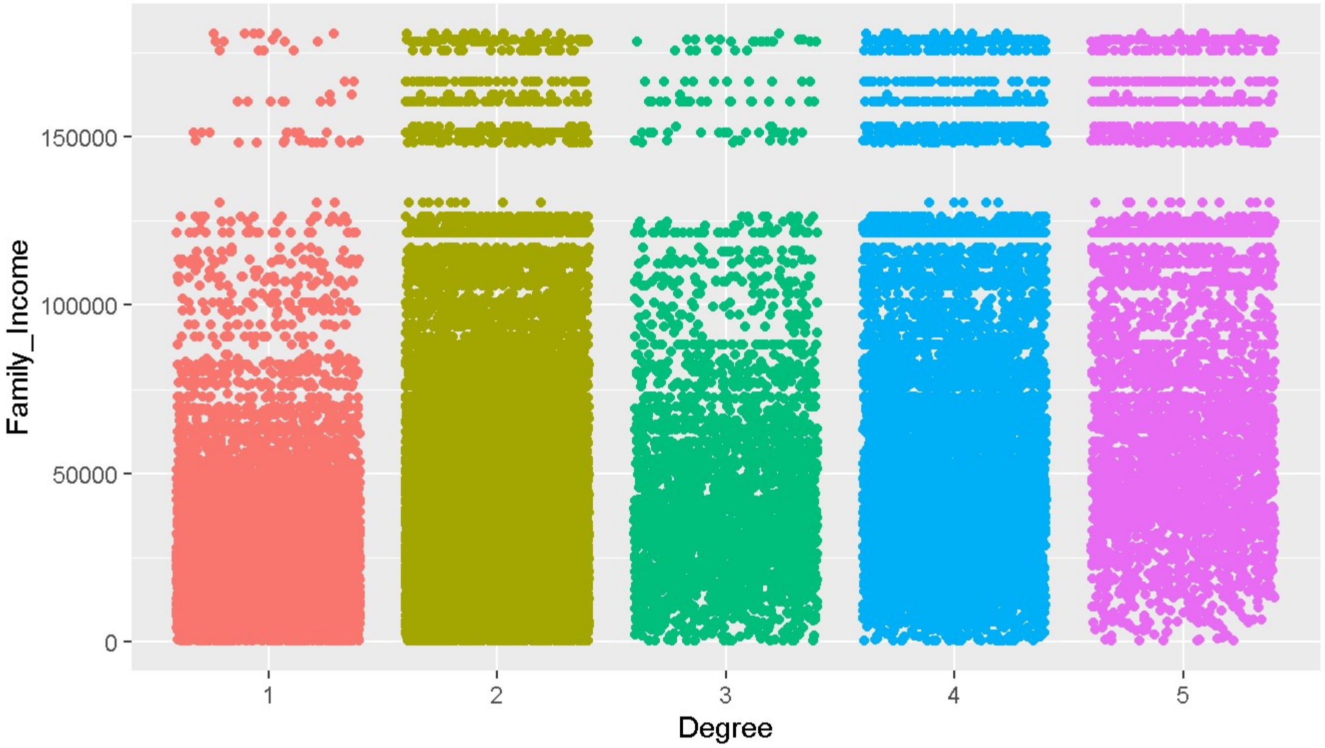
```
##### Replace missing values with the mean of the column.
#trimmed_gss_analysis_two$coninc[is.na(trimmed_gss_analysis_two$coninc)] <- #mean(trimmed_gss_analysis_two$coninc, na.rm=TRUE)

#### Omit rows with NA's (acceptable due to the large amount of data collected.)
trimmed_gss_analysis_two <- na.omit(trimmed_gss_analysis_two)

##### Rename column headers.
colnames(trimmed_gss_analysis_two) <- c("Family_Income","Degree")

##### Recode data for easier analysis.
trimmed_gss_analysis_two <- trimmed_gss_analysis_two %>% mutate(
    Degree = case_match(
      Degree,
      'Lt High School'~'1',
      'High School'~'2',
      'Junior College'~'3',
      'Bachelor'~'4',
      'Graduate'~'5',
    )
)

#### Export data for further analysis in Excel and additional formatting.
write.csv(trimmed_gss_analysis_two, "GSS_FamIncome_Degree_ANOVA.csv", row.names = TRUE)
```

Plots

Jitter Plot

The graph indicates a clear difference between various levels of respondent education and associated family earnings. The largest appears to be between Limited High School and High School; however, there are apparent differences among many of the degrees of education.

```
ggplot(trimmed_gss_analysis_two) +
  aes(x = Degree, y = Family_Income, color = Degree) +
  geom_jitter() +
  theme(legend.position = "none")
```



Part 4: Inference

Hypotheses

Does there appear to be a relationship between the level of respondents education and their household income?

H0: The average family income is the same across all degrees of education. (u1 = u2 = u3 = u4 = u5) HA: The average family income differs between at least one pair of degrees of education.

Evaluating the Hypothesis

Conditions for the ANOVA Test

```
res_aov <- aov(Family_Income ~ Degree,
  data = trimmed_gss_analysis_two
)
```

Independence

within: sampled observations must be independent of each other * random sample / assignment * each nj less than 10% of respective population * always important, but sometimes difficult to check

- Respondents were sampled randomly as part of the survey, therefore, it is safe to assume that the observations are independent of one another.
- The observations are less than 10% of the respective populations.

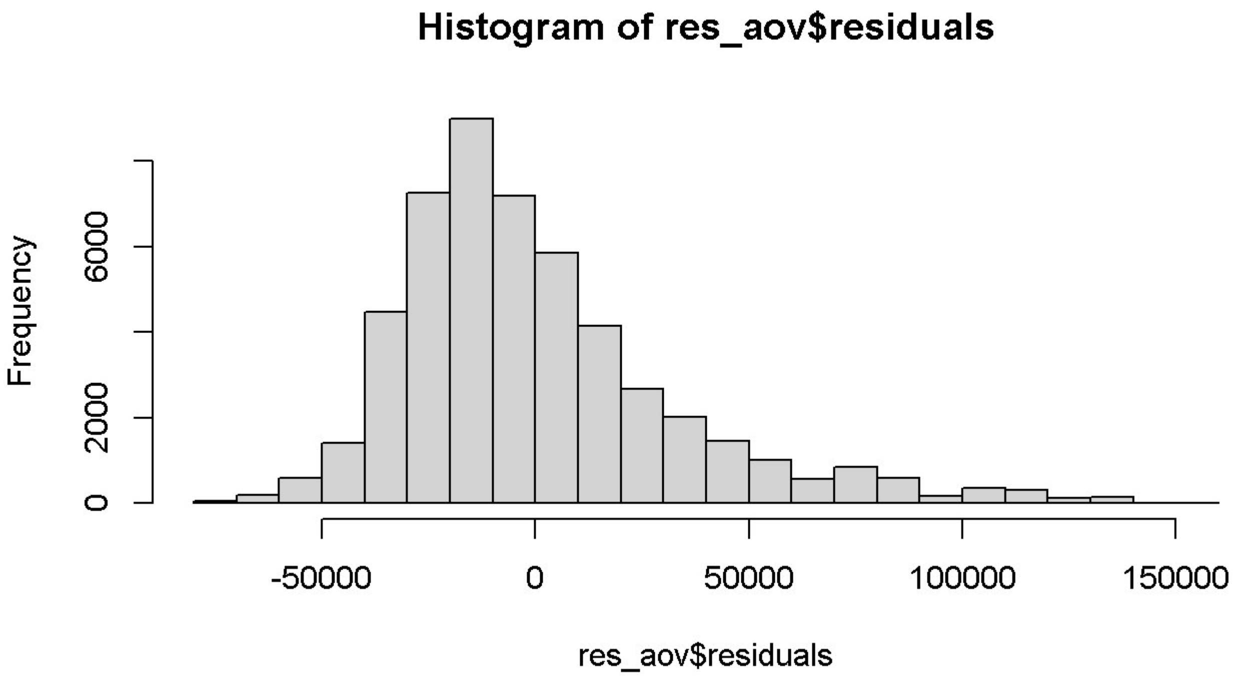
between: groups must be independent of each other * carefully consider whether the groups may be dependent

- Random sampling was employed, so it's safe to assume that each observation is independent, and therefore each group is also independent of one another as well as each observation is contained within one cell.

Approximately Normal

- distribution of response variable within each group should be approximately normal
- especially important when sample sizes are small
- Although there is skew in the data, the sample size is very large, therefore this shouldn't be an issue.

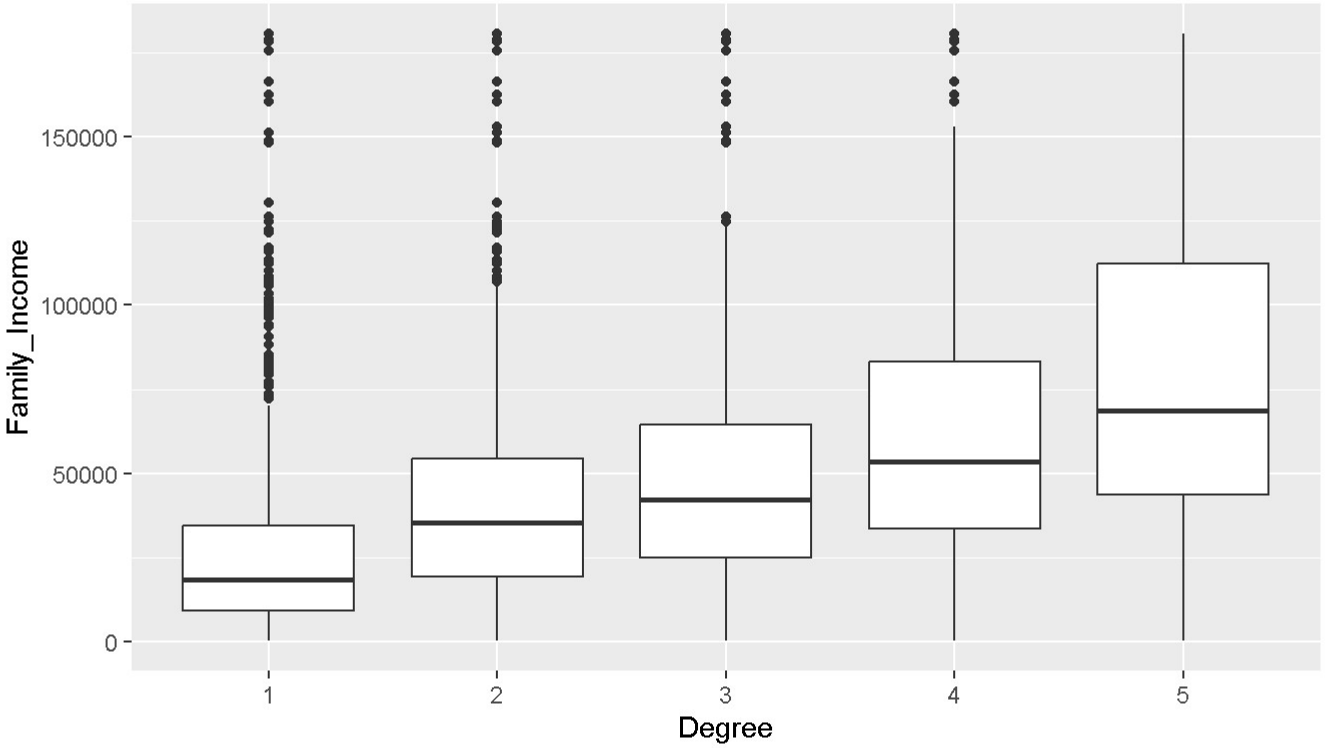
```
hist(res_aov$residuals)
```



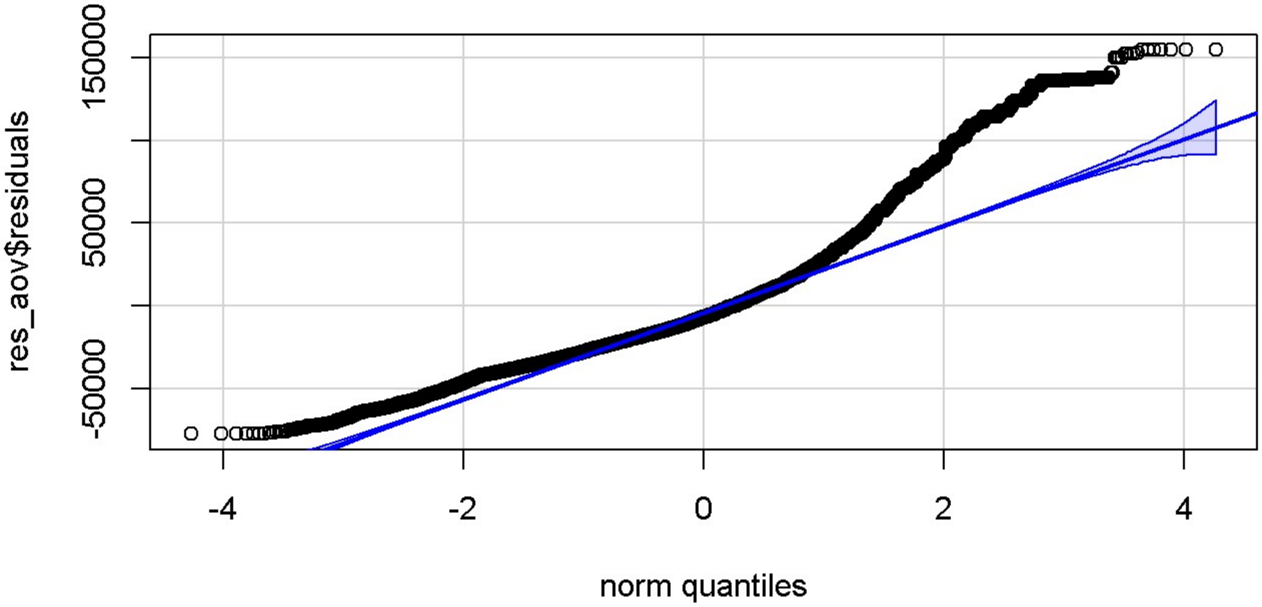
Constant Variance

- variability should be consistent across groups: homoscedastic groups
- especially important when sample sizes differ between groups
- Variance does not appear to be constant among all groups, therefore an alternative ANOVA test is required.

```
ggplot(trimmed_gss_analysis_two) +
  aes(x = Degree, y = Family_Income) +
  geom_boxplot()
```



```
qqPlot(res_aov$residuals,
  id = FALSE # id = FALSE to remove point identification
)
```



Test the hypothesis - ANOVA

How much variability is attributed to the explanatory variable? (Education)

```
1.08127E+13/6.53791E+13
```

```
## [1] 0.1653847
```

Approximately 16% of variability is due to the explanatory variable.

Test the hypothesis that Family Income and respondent Education are associated at the 5% significance level.

```
oneway.test(Family_Income ~ Degree,
  data = trimmed_gss_analysis_two,
  var.equal = FALSE # assuming unequal variances
)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  Family_Income and Degree
## F = 2340.8, num df = 4, denom df = 10598, p-value < 2.2e-16
```

```
pf(2340.8,4,10598,lower.tail=FALSE)
```

```
## [1] 0
```

Conclusion

- If p-value is small (less than α), reject H0.
- The data provide convincing evidence that at least one pair of population means are different from each other (but we can't tell which one).
- If p-value is large, fail to reject H0.
- The data do not provide convincing evidence that at least one pair of population means are different from each other, the observed differences in sample means are attributable to sampling variability (or chance).

The p_value of 0 is less than the 0.05 significance level. Therefore, we reject the null hypothesis in favour of the alternative; there is sufficient evidence indicating that at least one pair of population means are different from one another, however, we do not know which.

Test the hypothesis - Pairwise Testing

Modified Significance level given 5 levels: 0.02 SE for multiple pairwise comparisons: 382.773 Degrees of Freedom for multiple pairwise comparisons: 10267

1. Is there a difference between the average family income of Lt High School and High School respondents?

```
(lHS_HS <- 2 * pt(43.332, df = 10267, lower.tail = FALSE))
```

```
## [1] 0
```

```
if(lHS_HS < 0.02) {
  print("Reject the null hypothesis")
} else {
  print("Fail to reject the null hypothesis")
}
```

```
## [1] "Reject the null hypothesis"
```

2. Is there a difference between the average family income of Lt High School and Junior College respondents?

```
(lHS_JC <- 2 * pt(35.078, df = 2803, lower.tail = FALSE))
```

```
## [1] 8.33725e-224
```

```
if(lHS_JC < 0.02) {
  print("Reject the null hypothesis")
} else {
  print("Fail to reject the null hypothesis")
}
```

```
## [1] "Reject the null hypothesis"
```

3. Is there a difference between the average family income of Lt High School and Bachelor respondents?

```
(lHS_B <- 2 * pt(55.011, df = 7373, lower.tail = FALSE))
```

```
## [1] 0
```

```
if(lHS_B < 0.02) {
  print("Reject the null hypothesis")
} else {
  print("Fail to reject the null hypothesis")
}
```

```
## [1] "Reject the null hypothesis"
```

4. Is there a difference between the average family income of Lt High School and Graduate respondents?

```
lHS_G <- 2 * pt(-39.771, df = 3565, lower.tail = FALSE)

if(lHS_G < 0.02) {
  print("Reject the null hypothesis")
} else {
  print("Fail to reject the null hypothesis")
}
```

```
## [1] "Fail to reject the null hypothesis"
```

5. Is there a difference between the average family income of High School and Junior College respondents?

```
(HS_JC <- 2 * pt(12.255, df = 2803, lower.tail = FALSE))
```

```
## [1] 1.131081e-33
```

```
if(HS_JC < 0.02) {
  print("Reject the null hypothesis")
} else {
  print("Fail to reject the null hypothesis")
}
```

```
## [1] "Reject the null hypothesis"
```

6. Is there a difference between the average family income of High School and Bachelor respondents?

```
(HS_B <- 2 * pt(50.723, df = 7373, lower.tail = FALSE))
```

```
## [1] 0
```

```
if(HS_B < 0.02) {
  print("Reject the null hypothesis")
} else {
  print("Fail to reject the null hypothesis")
}
```

```
## [1] "Reject the null hypothesis"
```

7. Is there a difference between the average family income of High School and Graduate respondents?

```
(HS_G <- 2 * pt(61.497, df = 3565, lower.tail = FALSE))
```

```
## [1] 0
```

```
if(HS_G < 0.02) {
  print("Reject the null hypothesis")
} else {
  print("Fail to reject the null hypothesis")
}
```

```
## [1] "Reject the null hypothesis"
```

8. Is there a difference between the average family income of Junior College and Bachelor respondents?

```
(JC_B <- 2 * pt(19.144, df = 2803, lower.tail = FALSE))
```

```
## [1] 7.112095e-77
```

```
if(JC_B < 0.02) {
  print("Reject the null hypothesis")
} else {
  print("Fail to reject the null hypothesis")
}
```

```
## [1] "Reject the null hypothesis"
```

9. Is there a difference between the average family income of Junior College and Graduate respondents?

```
(JC_G <- 2 * pt(33.828, df = 2803, lower.tail = FALSE))
```

```
## [1] 1.176363e-210
```

```
if(JC_G < 0.02) {
  print("Reject the null hypothesis")
} else {
  print("Fail to reject the null hypothesis")
}
```

```
## [1] "Reject the null hypothesis"
```

10. Is there a difference between the average family income of Bachelor and Graduate respondents?

```
(B_G <- 2 * pt(21.036, df = 3565, lower.tail = FALSE))
```

```
## [1] 1.058915e-92
```

```
if(B_G < 0.02) {
  print("Reject the null hypothesis")
} else {
  print("Fail to reject the null hypothesis")
}
```

```
## [1] "Reject the null hypothesis"
```