Statistical inference with the GSS data

Author: Jorge Álvarez de la Fuente

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(statsr)
```

Load data

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

Part 1: Data

The General Social Survey (GSS) is a sociological survey that has been monitoring and studying the growing complexity of American society since 1972, to gather data in order to track and explain trends and constants in attitudes, behaviors and attributes on contemporary American society. It uses full-probability, personal interview survey designed to monitor changes in both social characteristics and attitudes and is being conducted in the United States. The GSS questions cover a wide and diverse range of topics such as civil liberties, morality, national spending priorities, crime and violence, among others. In addition to this, the National Data Program for the Social Sciences has carried out an extensive range of methodological research designed both to advance survey methods to ensure that the GSS data are of the highest possible quality.

The target population of the GSS is adults aged 18 and older living in the United States. The GSS randomly selects respondents in households across the United States from a mix of urban, suburban and rural geographic areas. Participation in the study is voluntary and the survey is conducted face-to-face with an in-person interview by NORC at the University of Chicago.

Below the gss data dimension is presented:

```
#dimension of data
dim(gss)
```

```
## [1] 57061   114
```

The gss data contains 57,061 observations and 114 variables, which means the observations are less than 10% of entire US population and therefore independence can be assumed. The results can be generalized to the entire population of citizens aged 18 and over in the United States. Random sampling was used as part of the data collection process, although there is no random assignment since this is an observational study. Given that there is no random assignment to an experiment, then casual relationships (causation) cannot be inferred, only association can be inferred.

Part 2: Research question

Research Question: Is there any noteworthy association between subjective class identification and confidence in executive branch of federal government? Does the genre have any significant impact on confidence in executive branch of federal government or otherwise the observed effect is due to chance?

Variables used:

1. class: Subjective class identification
2. sex: Respondents sex
3. confed: Confidence in executive branch of federal government.

All variables are categorical.

This question aims to measure the trust of US citizens in the federal government and find whether genre and social class have a meaningful impact on US citizens opinions regarding their confidence in institutions. This is an interesting question nowadays because the financial and economic crisis of 2008 along with the current covid-19 outbreak crisis are leading to a significant loss of trust in government institutions and politicians. According to OECD in 2013 in its report “Government at a Glance 2013” by 2012 on

average only four out of ten people in OECD member countries expressed confidence in their government. The following exploratory data analysis aims to find whether there is a significant loss of trust in government institutions in the United States as well.

Part 3: Exploratory data analysis

Table 1 and Table 2 below show the proportion contingency tables to summarize the relationship between the categorical variables class and confed and sex and confed. A contingency table is a special type of frequency distribution table, where two variables are shown simultaneously.

```
#This code filters "NA" and "No Class" values from gss database
gss2 <- gss %>%
  filter(!is.na(confed), class != "No Class") %>%
  select(class, confed)

#This code prints the proportion table of the contingency table excluding "No Class" category in class variabl

prop.table(ftable(gss2$class, gss2$confed, exclude = c("No Class"))) %>% round(.,3)
```

```
##                A Great Deal Only Some Hardly Any
##
## Lower Class           0.008     0.026      0.025
## Working Class         0.068     0.238      0.152
## Middle Class          0.083     0.238      0.130
## Upper Class           0.007     0.016      0.009
```

Table 1: Proportion contingency table for the variables “class” and “confed”

The results in Table 1 show that most of respondents of the different social classes declare to have only some confidence in federal government.

```
#This code filters "NA" values from gss database
gss3 <- gss %>%
  filter(!is.na(confed)) %>%
  select(sex, confed)

#This code prints the proportion table of the contingency table for genres
prop.table(table(gss3$sex, gss3$confed)) %>% round(.,3)
```

```
##
##          A Great Deal Only Some Hardly Any
##   Male           0.079     0.220      0.146
##   Female         0.089     0.299      0.167
```

Table 2: Proportion contingency table for the variables “class” and “confed”

As the results shown in Table 2, respondents of both genres admit to have only some confidence in federal government. It can be confirmed, then, that the majority of respondents admit to have only some confidence in federal government.

Figures 1 and 2 below show plots linking the explanatory variables “social classes” and “sex” with the response variable “confidence in federal government” in proportions

```
#This code converts the data shown in Table 1 to a dataframe excluding "No Class" category in class variable
confedtable1 <- data.frame(prop.table(ftable(gss2$class, gss2$confed, exclude = c("No Class"))))

names(confedtable1) <- c("Class", "Confidence", "Respondents")

#Bar plot of proportion of confidence in federal government by social class
ggplot(data = confedtable1, aes(x = factor(Class), y = Respondents, fill = Confidence)) +
  geom_bar(stat = "identity", position = "stack") + xlab("Social Class") + ylab("Proportion of Respondents") +

  scale_fill_brewer(palette="YlGnBu", name = "Confidence Level")
```

Figure 1: Bar plot for the variables "class" and "confed"

The results as per Figure 1 show a pattern in all classes in which the majority of respondents admit to have only some confidence in the federal government. A significant number of respondents admit to have hardly any confidence in the federal government as well and only a small proportion of respondents admit to have a great deal of confidence in the federal government.

```
#This code converts the data shown in Table 2 to a dataframe
confedtable2 <- data.frame(prop.table(table(gss3$sex, gss3$confed)))

names(confedtable2) <- c("Sex", "Confidence", "Respondents")

#Bar plot of proportion of confidence in federal government by genre
ggplot(data = confedtable2, aes(x = factor(Sex), y = Respondents, fill = Confidence)) +
  geom_bar(stat = "identity", position = "stack") + xlab("Genre") + ylab("Proportion of Respondents") +
  scale_fill_brewer(palette="YlGnBu", name = "Confidence Level")
```

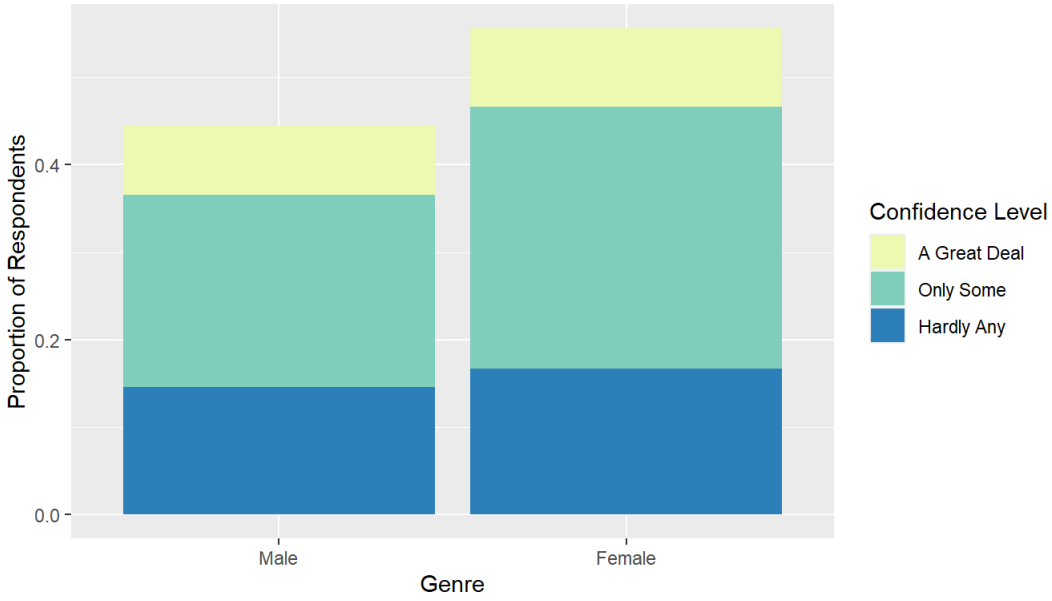


Figure 2: Bar plot for the variables "sex" and "confed"

Analogously, the results as per Figure 2 show a pattern in which both genres males and females admit to have only some confidence in the federal government, followed by hardly any confidence in the federal government as well and only a small proportion of great deal of confidence in the federal government.

The results of the present exploratory data analysis on the database GSS confirm that the majority of the population in the US are experiencing a loss of trust in federal government and institutions nowadays as stated by the report "Government at a Glance 2013" published by OECD. These results may be attributed to the financial and economic crisis of 2008. It would be interesting to conduct the same exploratory data analysis on the most recent GSS data to confirm whether this trend continues as the result of covid-19 outbreak.

Part 4: Inference

Hypothesis Test:

A hypothesis test for a difference of proportions of those respondents that do not trust federal government and institutions is conducted below as follow:

H0 (null hypothesis): The proportion of males who show only some confidence in federal government equals the proportion of females who show only some confidence in federal government.

H0: p1 - p2 = 0

*Where p1 = proportion of males who show only some confidence in federal government and p2 = the proportion of females who show only some confidence in federal government.

HA (alternative hypothesis): The proportion of males who show only some confidence in federal government is different from the proportion of females who show only some confidence in federal government.

HA: p1 - p2 ≠ 0

*Where p1 = proportion of males who show only some confidence in federal government and p2 = the proportion of females who show only some confidence in federal government.

Checking Conditions:

The conditions for inference using Central Limit Theorem are independence and sample size.

1. Independence: Since random sampling was used collecting GSS data and besides the number of observations (19,535) considered is less than 10% of the US population (328 million people), independence can be assumed.

The number of observations used in the hypothesis test is shown below:

```
#This code filters the confidence in federal government by "Only Some"
gss4 <- gss3 %>%
  filter(confed == "Only Some", !is.na(sex)) %>%
  select(sex, confed)
```

```
#Number of observations in the hypothesis test used
dim(gss4)
```

```
## [1] 19535     2
```

2. Sample size and success/failure condition: Given p = 0.22 (the lowest proportion among males and females who show only some confidence in federal government) and the number of observations 37,637, the success/failure condition gives the following result:

19,535 * 0.22 = 4,298. Then: n*p ≥ 10

19,535 * (1-0.22) = 15,238. Then: n*(1-p) ≥ 10

Given that we have 10 or more successes and 5 or more failures, a normal distribution can be used and sample size or success/failure condition is meet.

Both conditions of independence and sample size or success/failure condition are met so GSS data follows a nearly normal distribution.

Method used:

A Chi-Square test of independence is used. The Chi-Square test of independence is used to determine if there is a significant relationship between two categorical variables, as is this case with the variables: sex and confed. The chi-square test evaluates whether there is a significant association between the categories of the two variables. Chi-square statistic can be easily computed in R using the function chisq.test()

Inference Performance:

The hypothesis test is performed using the chisq.test function available in R. This function needs the previous proportion contingency table as input and gives the chi-square value, degrees of freedom and p-value as outputs. Additionally, inference function with Central Limit Theorem (CLT) based using a confidence level of 0.95 (95%) is used for hypothesis test to calculate the p-value and confirm the previous one calculated with chisq.test() function. The results are shown below:

```
#This code performs the Chi-Square test for the contingency table "sex" and "confed"
chisq.test(table(gss3$sex, gss3$confed))
```

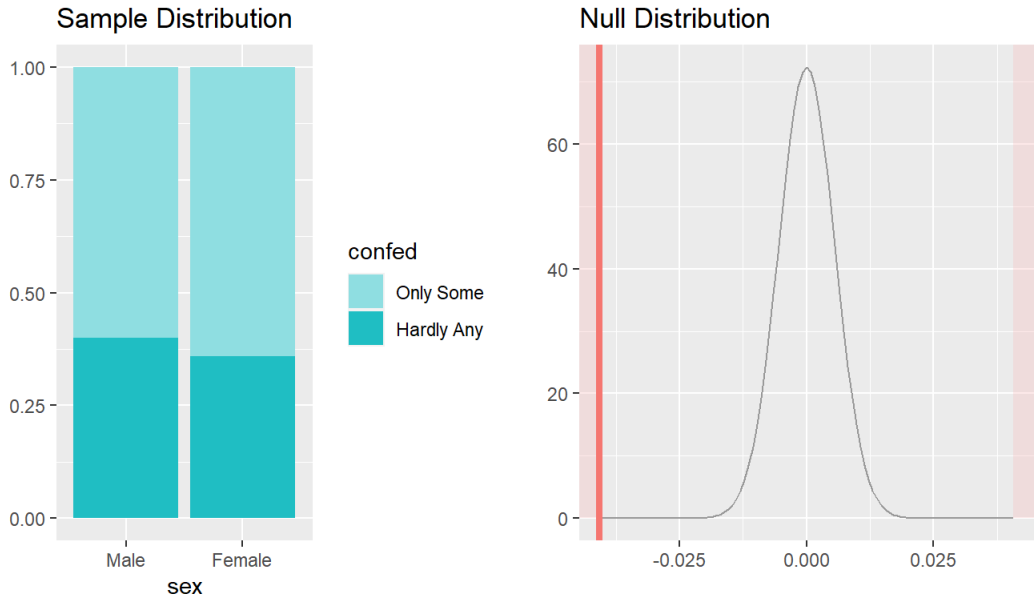
```
##
##  Pearson's Chi-squared test
##
## data:  table(gss3$sex, gss3$confed)
## X-squared = 72.322, df = 2, p-value < 2.2e-16
```

```
#This code filters confidence in federal government by "Only Some" and "Hardly Any" categories
gss5 <- gss3 %>%
  filter(xor(confed == "Only Some", confed == "Hardly Any")) %>%
  select(sex, confed)
```

```
#This code performs the hypothesis test over gss3 data to estimate the p-value for proportion
#of males and females that shown "Only Some" confidence in federal government
inference(y = confed, x = sex, data = gss5, type = "ht", null = 0, statistic = "proportion",
          success = "Only Some", method = "theoretical", alternative ="twosided", conf_level = 0.95)
```

```
## Warning: Ignoring null value since it's undefined for chi-square test of
## independence
```

```
## Response variable: categorical (2 levels, success: Only Some)
## Explanatory variable: categorical (3 levels)
## n_Male = 13762, p_hat_Male = 0.6009
## n_Female = 17556, p_hat_Female = 0.6417
## H0: p_Male =  p_Female
## HA: p_Male != p_Female
## z = -7.3848
## p_value = < 0.0001
```



Results interpretation:

The output of chisq.test() function gives a Chi-squared value of 72.322, df = 2 and p-value = 2.2e-16. The inference function gives a result of p-value < 0.0001. Since p-value < 0.05 for a confidence level of 95% (significance level $\alpha = 0.05$), we reject the null hypothesis H0 and conclude that there is a noteworthy difference between the proportion of males that admit to have only some confidence in federal government and the proportion of females that admit to have only some confidence in federal government.

However, is noteworthy that being an observational study, causal relationship between these two variables “sex” and “confed” cannot be inferred from the analysis. We can only infer that they are significantly associated with each other.