Statistical inference with the GSS data

Setup

Load packages

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

Load data

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

Part 1: Data

The General Social Survey (GSS) has been monitoring societal change and studying the growing complexity of American society since 1972. The GSS aims to gather data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes; to examine the structure and functioning of society in general as well as the role played by relevant subgroups; to compare the United States to other societies in order to place American society in comparative perspective and develop cross-national models of human society; and to make high-quality data easily accessible to scholars, students, policy makers, and others, with minimal cost and waiting.

GSS questions cover a diverse range of issues including national spending priorities, marijuana use, crime and punishment, race relations, quality of life, confidence in institutions, and sexual behavior.

Mode of data collection is computer-assisted personal interview (CAPI), face-to-face interview or telephone interview.

The dataset is composed of 57061 cases corresponding to an equal number of interviewed respondant. Each person may be considered as a single case and there are 114 variables (information) which were recoreded for each case.

This sample consists of all Americans which are randomly selected, generalizability to the entire American population can be made. In 1982, there was an oversample of black respondents, however recently there has not been oversampling.

However, as this is an observational study rather than an experiment, there was no random assignment which means that causality cannot be inferred.

Part 2: Research question

I am interested to know how people are getting along financially these days. So far as respondent and respondent's family are concerned, would they say that they are pretty well satisfied with their present financial situation?

I would like to know if there are any differences in financial satisfation by gender i.e. Is the proportion of males who are very satisfied with their financial situation different than the proportion of females who are very satisfied with their financial situation.

For this research question, I will take two variables into account: 1) sex: Respondent's sex 2) satfin: Satisfaction with financial situation

---

Part 3: Exploratory data analysis

```
total <- gss %>%
  group_by(sex) %>%
  summarise(total = n())

total
```

```
## # A tibble: 2 × 2
##      sex total
##   <fctr> <int>
## 1   Male 25146
## 2 Female 31915
```

In this dataset, there are 25146 males and 31915 females out of total 57061 respondents. These numbers will be used for the calculation of proportion later.

```
satisfied <- gss %>%
  group_by(sex) %>%
  filter(satfin == "Satisfied") %>%
  summarise(satisfied = n())

data <- data.frame(total, satisfied[ ,2])

data
```

```
##      sex total satisfied
## 1   Male 25146      6951
## 2 Female 31915      8393
```

Here, the total numbers of financially satisfied males and females have been calculated. There are 6951 satisfied males and 8393 satisfied females.

```
data <- data %>%
  mutate(prop_satisfied = satisfied / total)

data
```

```
##      sex total satisfied prop_satisfied
## 1   Male 25146      6951      0.2764257
## 2 Female 31915      8393      0.2629798
```

From the above calculation of proportion of financillay satisfied people of each gender, we can see that 27.6% males are happy with their financial situation while 26.2% of females are financially satisfied.

Part 4: Inference

HYPOTHESES: H0: p_male = P_female There is no difference in the proportion of financially satisfied males and females in the United States.

HA: p_male not equal to p_female There is a difference in the proportion of financially satisfied males and females in the United States.

CONDITIONS: 1. This is a randomly selected data. 2. Observations in the sample are independent. 3. Sample is sufficienly large as n > 30 4. Also np >= 10 and n(1-p) >= 10

All the conditions are met.

TECHNIQUE: For comparing two proportions, the technique of hypothesis test has been used.

A hypothesis test is being used over a confidence interval as there are two groups and I want to compare the two given our defined hypotheses.

The hypothesis test for comparing two proportions is ideal, as the data has been split by gender and then a proportion has been calculated.

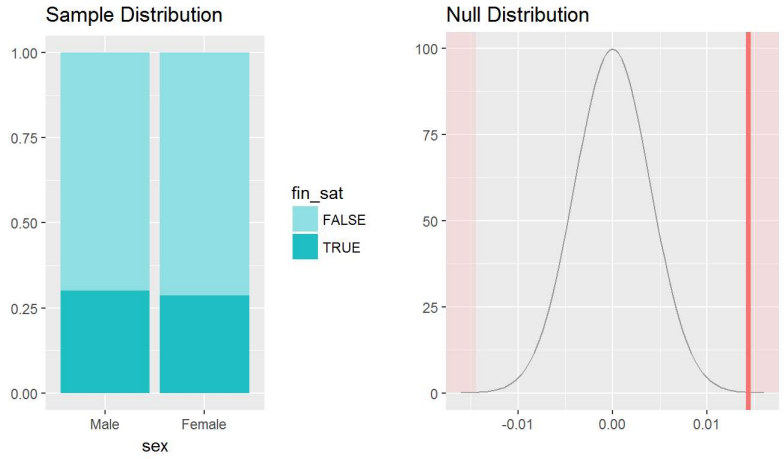
The significance level for this test has been used as 5% i.e. alpha = 0.05 (a = 0.05)

There are three options for financial satisfaction - Satisfied, More or Less Not at all Satisfied - but I am only interested about those individuals who are very well satisfied. Create a new dataset gss_2 with a new column for whether or not they were satisfied. I am left with only two options Satisfied or Unsatisfied.

```
gss_2 <- gss %>%
  mutate(fin_sat = satfin == "Satisfied")

inference(data = gss_2, x = sex, y = fin_sat, statistic = "proportion", null =0,type = "ht", method = "theoretical", success
= TRUE, alternative = "twosided", sig_level = 0.95)
```

```
## Response variable: categorical (2 levels, success: TRUE)
## Explanatory variable: categorical (2 levels)
## n_Male = 23126, p_hat_Male = 0.3006
## n_Female = 29328, p_hat_Female = 0.2862
## H0: p_Male =  p_Female
## HA: p_Male != p_Female
## z = 3.5978
## p_value = 3e-04
```



From the sample distribution and the exploratory analysis, it seems that proportion of males who are satisfied financially is slightly more than the proportion of females.

As per the calculations, the p_value is also very very small and much more smaller than the significance level of 5%, we reject the null hypothesis and accept the alternative, that means - "There is a difference in the proportion of financially satisfied males and females in the United States."