Statistical inference with the GSS data

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(gridBase)
library(gridExtra)
library(statsr) # For verification only
```

Load data

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

Part 1: Data

The GSS(General Social Survey) is personal interview survey from ramdom US household conducted by National Opinion Research Center (NORC). NORC at University of Chicago is an independant research institution that provide data and analysis in five main areas: Economics, Markets, and the Workforce; Education, Training, and Learning; Global Development; Health and Well-Being; and Society, Media, and Public Affairs.

Data is from randomly sampled subjects from the population of United States. This data is generalizable of US population and can be used only for observational study and not for causal reseaoning.

Part 2: Research question

Question: How do proportion of male and female in given GSS sample data compare with respect to attending higher education (Bachelors or above) ? Is there a difference in proportion of male and female with higher education (Bachelors or above) in United States ?

Background and motivation

According to Millennium Challenge Corporation (MCC) (https://www.mcc.gov/ (https://www.mcc.gov/))

- Gender and social inequality are known constraints to economic growth, and research shows that the benefits from poverty reduction initiatives are unequally shared with women.
- In some countries, productivity could increase by as much as 25% if the barriers that prevent women from entering the workforce were eliminated
- Studies find that increases in women’s education boost their wages and gender inequalities in education are associated with negative economic consequences.

While gender inequalities in education in general is more common in North African countries, in advanced economy

such as that of United States higher education is important for its citizen to participate in the economy. So our research focus on gender inequalities in **higher education**.

"We can't allow higher education to be a luxury in this country. It's an economic imperative that every family in America has to be able to afford." - President Barack Obama

---

Part 3: Exploratory data analysis

Research begins by performing exploratory data analysis on given dataset to determine if there is difference in percentage of Bachelors Degree and Graduate Degree between male and female subjects.

```
# Total female subjects
t_respF <- gss %>% filter(sex=="Female") %>% select(sex)
t_respF <- nrow(t_respF)

# Creating tables with female having education Bachelors or above
degree_tblF_cnt <- gss %>% filter(sex=="Female", degree=="Bachelor" | degree=="Graduate")
%>% select(degree) %>% group_by(degree) %>% tally()

# Adding a column in above table showing perecentage

degree_tblF_cnt <- degree_tblF_cnt %>% mutate(percent=n*100/ t_respF)

# Total male subjects
t_respM <- gss %>% filter(sex=="Male") %>% select(sex)
t_respM <- nrow(t_respM)

# Creating tables with male having education Bachelors or above
degree_tblM_cnt <- gss %>% filter(sex=="Male", degree=="Bachelor" | degree=="Graduate") %>
% select(degree) %>% group_by(degree) %>% tally()
# Adding a column in above table showing perecentage

degree_tblM_cnt <- degree_tblM_cnt %>% mutate(percent=n*100/ t_respM)
```

Snaphot of female subjects with education, bachelors or above

```
degree_tblF_cnt
```

```
## Source: local data frame [2 x 3]
##
##     degree     n   percent
##     <fctr> <int>     <dbl>
## 1 Bachelor  4180 13.097290
## 2 Graduate  1779  5.574181
```

Snaphot of male subjects with education, bachelors or above

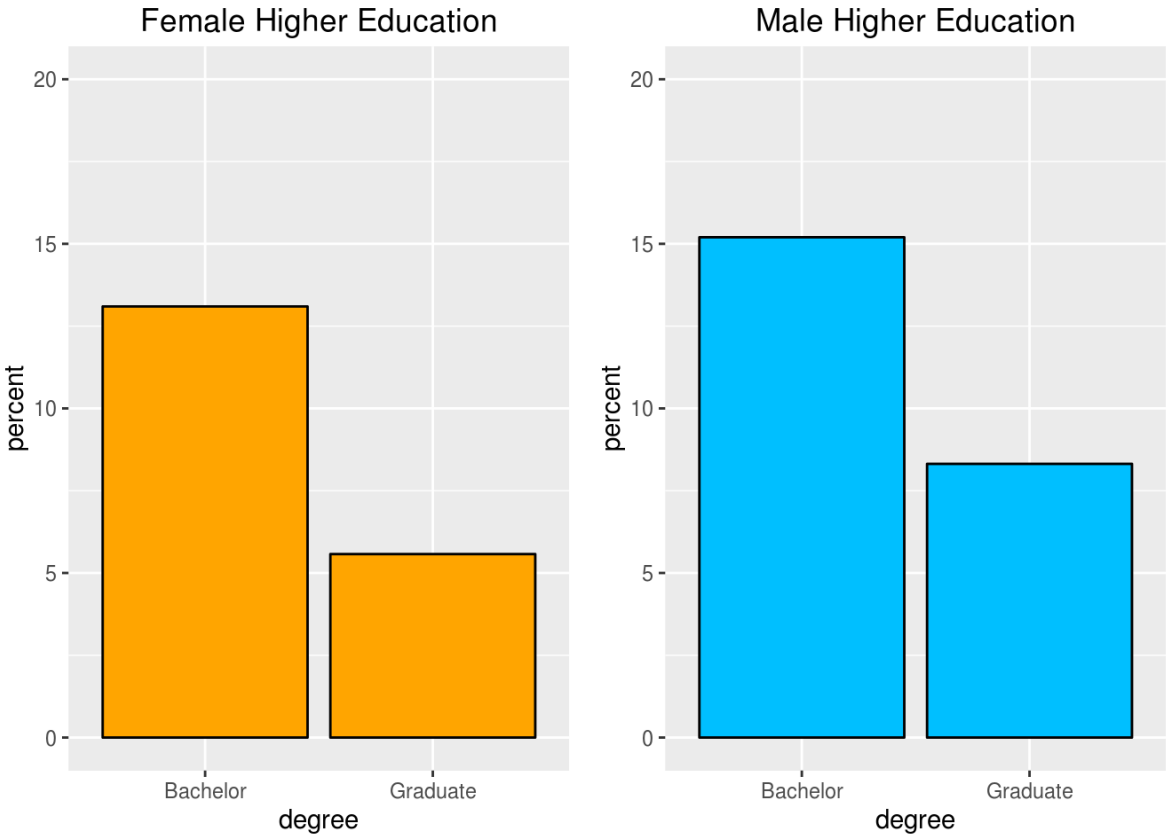
```
degree_tblM_cnt
```

```
## Source: local data frame [2 x 3]
##
##     degree     n   percent
##     <fctr> <int>     <dbl>
## 1 Bachelor  3822 15.199236
## 2 Graduate  2091  8.315438
```

Visual comparison of higher education among female and male in the sample

Note: The sample size of male and female are different so we used percentage for comparison

```
f_hEd <- ggplot(degree_tblF_cnt, aes(x=degree, y=percent)) + geom_bar(stat = "identity", c
olor="black",fill="orange")
m_hEd <- ggplot(degree_tblM_cnt, aes(x=degree, y=percent)) + geom_bar(stat = "identity", c
olor="black",fill="deepskyblue")
f_hEd <- f_hEd + scale_y_continuous(limits=c(0, 20)) + ggtitle("Female Higher Education")
m_hEd <- m_hEd + scale_y_continuous(limits=c(0, 20)) + ggtitle("Male Higher Education")
grid.arrange(f_hEd, m_hEd, ncol=2)
```



Conclusion: Over 2% more male subjects has bachelors degree and around 3% more male has Graduate degree compare to female subjects.

Part 4: Inference

In this research, we're going to label sex (male and female) as one of our categorical variables the explanatory variable and level of education (higher and lower), our response variable. So, our categorical variables that have two levels; success (higher education, Bachelors or above) and failure (Lower education). We are going to calculate proportion of successes in the two groups based on our sample, and were going to compare this to each other.

```
# Creating table of sex of respondants and their correponding education degree
ed_mf <- gss %>% filter(!is.na(sex), !is.na(degree)) %>% select(sex, degree)
# Classifying degree into level Bachelors and above Higher Education and everything else a
s Lower
ed_mf <- ed_mf %>% mutate(lev=ifelse(degree=="Bachelor" | degree=="Graduate", "High", "Low
"))
# Number of samples with High and Low Education for females
f_level <- ed_mf %>% filter(sex=="Female") %>% group_by(lev) %>% tally()
# Number of samples with High and Low Education for males
m_level <- ed_mf %>% filter(sex=="Male") %>% group_by(lev) %>% tally()
# Total number of subjects by sex
total <- ed_mf %>% group_by(sex) %>% tally()
# Extracting data from above tables for comparision tables
f_level_H <- f_level %>% filter(lev=="High") %>% select(n)
m_level_H <- m_level %>% filter(lev=="High") %>% select(n)

f_tot <- total %>% filter(sex=="Female") %>% select(n)
m_tot <- total %>% filter(sex=="Male") %>% select(n)
# Force R not to use exponential notation
options(scipen=10000)


male <- c(m_level_H$n , m_tot$n - m_level_H$n, m_tot$n, m_level_H$n/m_tot$n )
female <- c(f_level_H$n , f_tot$n - f_level_H$n, f_tot$n, f_level_H$n/f_tot$n)

#  Creating Comparision table
comp_table <- data.frame(male, female)

# For easy readabilty
rownames(comp_table) <- c("Higher Education", "No Higher education", "Total (n)", "p_hat")

comp_table
```

```
##                              male        female
## Higher Education     5913.0000000  5959.0000000
## No Higher education 18765.0000000 25414.0000000
## Total (n)           24678.0000000 31373.0000000
## p_hat                   0.2396061     0.1899404
```

Confidence Interval

We determine difference between 2 proportion

$$point\ estimate\ +- margin\ of\ error$$

$$(\hat{p}_{male} - \hat{p}_{female})\ +- z * SE_{(\hat{p}_{male} - \hat{p}_{female})}$$

We determine standard of error in our research as

$$SE = \sqrt{\frac{\hat{p}_{male} * (1 - \hat{p}_{male})}{n_{male}} + \frac{\hat{p}_{female} * (1 - \hat{p}_{female})}{n_{female}}}$$

Conditions for inference for comparing two independant proportion

1. Independence:

Within groups: Random population and less than 10% of population (US population 319 million approximately).

Between groups: Education level of one group of explanatory variable is independant of other.

So there is no dependency within or between groups.

2. Sample/skew:

```
# For males
m_tot$n*(m_level_H$n/m_tot$n) # Success
```

```
## [1] 5913
```

```
m_tot$n*(1-(m_level_H$n/m_tot$n)) #Failure
```

```
## [1] 18765
```

```
# For females
f_tot$n*(f_level_H$n/f_tot$n) # Success
```

```
## [1] 5959
```

```
f_tot$n*(1-(f_level_H$n/f_tot$n)) #Failure
```

```
## [1] 25414
```

From above output we can assume that the sampling distribution of the difference between two proportions is nearly normal.

Calculation

```
p_hat_male <- (m_level_H$n/m_tot$n) # Sample proportion of male success (Higher Education)
p_hat_female <- (f_level_H$n/f_tot$n) # Sample proportion of female success (Higher Educat
ion)

# Determining point estimate
pe <-p_hat_male-p_hat_female
pe
```

```
## [1] 0.04966573
```

```
# Determining Standard error
se <- sqrt(((p_hat_male*(1-p_hat_male))/m_tot$n)+((p_hat_female*(1-p_hat_female))/f_tot$n)
)
se
```

```
## [1] 0.003505311
```

```
# Determining confidence interval
ci <- c(pe-(1.96*se), pe+(1.96*se))
ci
```

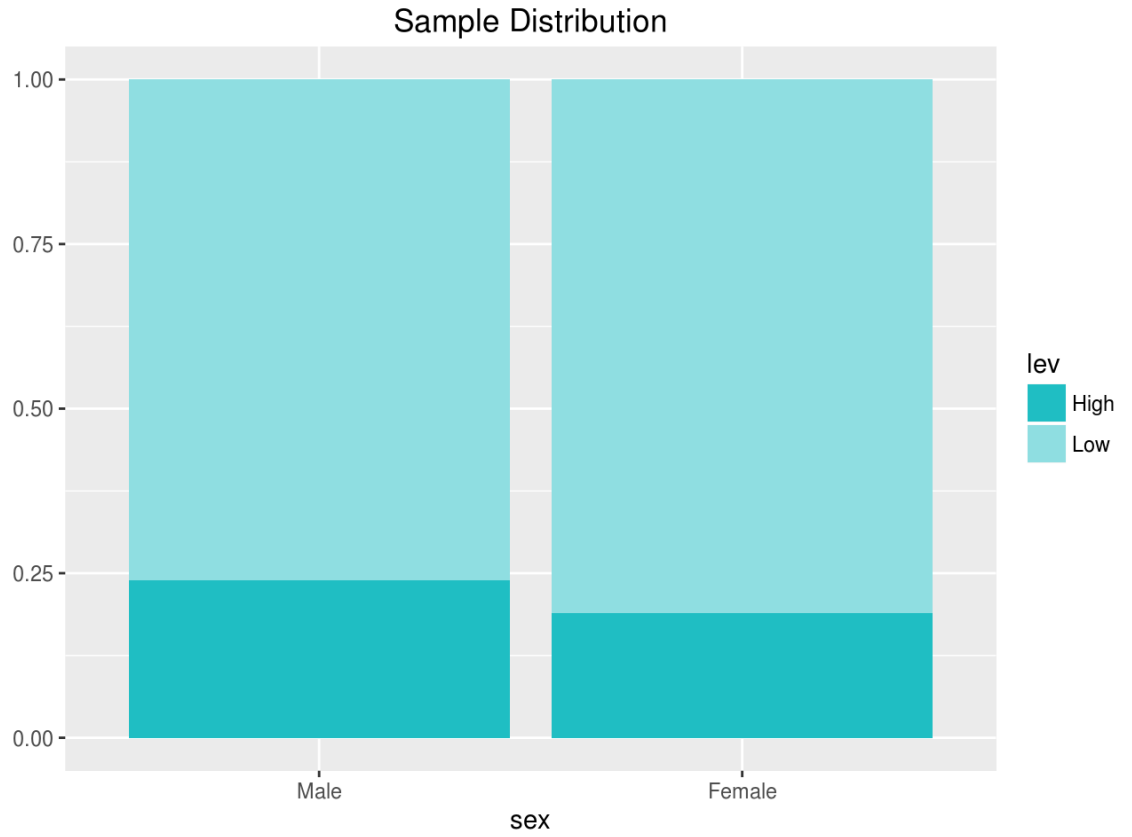
```
## [1] 0.04279532 0.05653614
```

Verification We will verify above calculation using statsr inference() function.

Note: statsr is not standard R library but a companion package for the Coursera Statistics with R specialization developed by Prof. Dr Mine Cetinkaya-Rundel, Duke University Department of Statistical Science.

```
inference(y = lev, x=sex, data = ed_mf,  statistic = "proportion", type = "ci",  method =
"theoretical", success = "High")
```

```
## Response variable: categorical (2 levels, success: High)
## Explanatory variable: categorical (2 levels)
## n_Male = 24678, p_hat_Male = 0.2396
## n_Female = 31373, p_hat_Female = 0.1899
## 95% CI (Male - Female): (0.0428 , 0.0565)
```



We see that our calculation matches with the ouput of statr inference() function.

Conclusion: From the above we are 95% confident that proportion of male who has higher education is 4.28% to 5.65% more than proprtion of female with higher education.

Hypothesis

Null Hypothesis: Proportion of male and female in population with higher education are equal

$$H_0 : p_{female} = p_{male}$$

Alternate Hypothesis: Proportion of male and female in population with higher education are not equal

$$H_A : p_{female} \neq p_{male}$$

We do not have proportion of higher education among male and female population of United States. So we will use **pooled proportion** .

We define pooled proportion in our research as :

$$\hat{p}_{pool} = \frac{number\ of\ male\ with\ higher\ education + number\ of\ female\ with\ higher\ education}{n_{male} + n_{female}}$$

```
p_hat_pool<-(m_level_H$n+f_level_H$n)/(m_tot$n+f_tot$n)
p_hat_pool
```

```
## [1] 0.2118071
```

Conditions for inference for hypothesis test

1. Independence:

Within groups: Random population and less than 10% of population (US population 319 million approximately).

Between groups: Education level of one group of explanatory variable is independant of other.

So there is no dependency within or between groups.

2. Sample/skew:

```
# For males
m_tot$n*p_hat_pool # Success
```

```
## [1] 5226.976
```

```
m_tot$n*(1-p_hat_pool) #Failure
```

```
## [1] 19451.02
```

```
# For females
f_tot$n*p_hat_pool # Success
```

```
## [1] 6645.024
```

```
f_tot$n*(1-p_hat_pool) #Failure
```

```
## [1] 24727.98
```

From above output we can assume that the sampling distribution of the difference between two proportions is nearly normal.

Conducting hypothesis test, at 5% significance level

We will determine Standard Error of Population as

$$SE = \sqrt{\frac{\hat{p}_{pool} * (1 - \hat{p}_{pool})}{n_{male}} + \frac{\hat{p}_{pool} * (1 - \hat{p}_{pool})}{n_{female}}}$$

Calculating Standard error

```
# Calculating p(1-p)/n1  and p(1-p)/n2
ml_s <- (p_hat_pool*(1-p_hat_pool))/m_tot$n
fl_s <- (p_hat_pool*(1-p_hat_pool))/f_tot$n
# Calculating Standard Error
se <- sqrt(ml_s + fl_s)
se
```

```
## [1] 0.003476524
```

Determining point estimate

```
prop_diff <- (m_level_H$n/m_tot$n ) -(f_level_H$n/f_tot$n )
prop_diff
```

```
## [1] 0.04966573
```

Parameters for calculating z-score and p-value

$$\alpha = 0.05$$

$$(\hat{p}_{male} - \hat{p}_{female}) \; is \; nearly \; N(mean = 0, SE = 0.003476524)$$

$$point \; estimate = 0.04966573$$

We will determine z score using

$$z = \frac{point \; estimate - 0}{SE}$$

Determining z score

```
z<-(prop_diff-0)/se
z
```

```
## [1] 14.28603
```

$$z \; score = 14.28603$$

This is very high test statistics much further than 3 standard deviation from null value. So we can safely conclude:

$$p \; value \; is \; almost \; 0$$

So we **reject** null hypothesis in favor of alternate hypothesis.

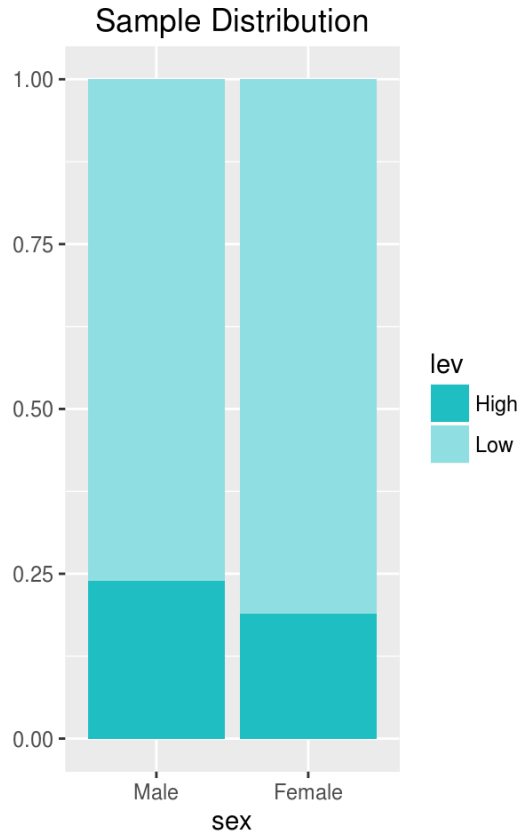
To avoid Type 1 error we will confirm above calculation using **statsr inference()** function.

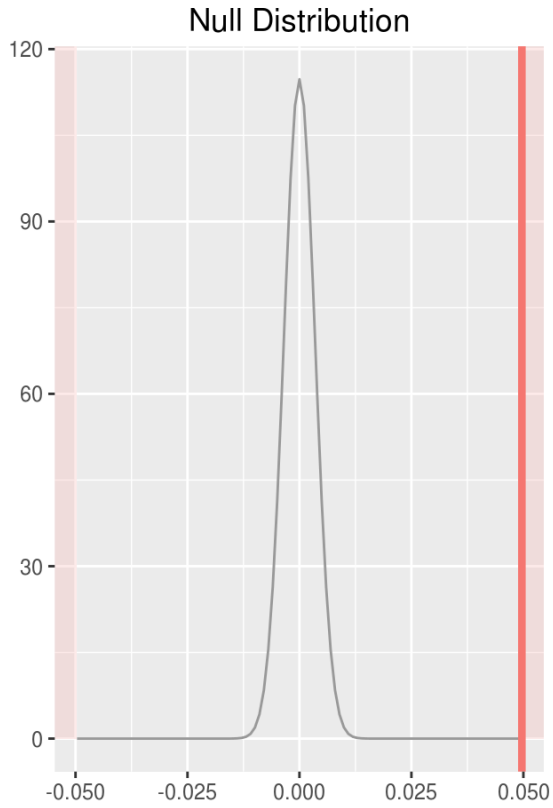
Confirming above calculation using inference {statsr}

Note: statsr is not standard R library but a companion package for the Coursera Statistics with R specialization developed by Prof. Dr Mine Cetinkaya-Rundel, Duke University Department of Statistical Science.

```
inference(y = lev, x=sex, data = ed_mf,  statistic = "proportion", type = "ht", alternativ
e="twosided", null=0, method = "theoretical", success = "High")
```

```
## Response variable: categorical (2 levels, success: High)
## Explanatory variable: categorical (2 levels)
## n_Male = 24678, p_hat_Male = 0.2396
## n_Female = 31373, p_hat_Female = 0.1899
## H0: p_Male =  p_Female
## HA: p_Male != p_Female
## z = 14.286
## p_value = < 0.0001
```





We see z score, p_hat_Male, p_hat_Female from inference() matches with that of our calculation. With p_value = < 0.0001
we **reject** null hypothesis in favor of alternate hypothesis.

Conclusion: Proportion of male and female with higher education in US are not equal. Probability of observed or more extreme outcome given null hypothesis is true is almost 0. In other word probability of difference in population proportion **simply by chance** is almost 0%.