# Statistical Inference with the GSS Data

*Date: 4/13/2020*

## Contents



### Part 1: Data

This project uses the **General Social Survey (GSS) Cumulative File 1972-2012** to conduct statistical analysis.[1]

The GSS samples data from contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes. The GSS contains a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. Among the topics covered are civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events. In short, the GSS is the single best source for sociological and attitudinal trend data covering the United States.[2]

In this project, we use the `gss` dataset[3] and rely on the `dplyr` and `ggplot2` packages.

```
# Read data
load("C:/Other/eLearning/Coursera/Inferential Statistics/Week 5/gss.Rdata")

# Load packages
library(dplyr)
library(ggplot2)
```

### Part 2: Research Question

Our research focuses on people's education level. Specifically, the question is: **Is there a statistically significant difference between the respondents' education level in 2002 and that in 2012 in the U.S.?**

The reasoning behind this research question is that education to a large extent determines the investment in human capital, and greatly affects the total factor productivity of a country. Therefore, it is meaningful to conduct research on this field.

Through this analysis, we would like to see if there is statistical evidence of education level difference as time pass by.

[1] Missing values from the responses have been removed and factor variables are created when appropriate by the course developers to facilitate analysis using R (source: project codebook).

[2] See "An Introduction to the General Social Survey".

[3] The modified `gss` dataset is provided by the course developers.

### Part 3: Exploratory Data Analysis

The variable of interest is *educ*, which indicates the highest year of school completed. It corresponds to the survey question "What is the highest grade in elementary school or high school that you finished and got credit for?"

The summary statistics of school year in 2002 is:

```
# Summary statistics of schooling year in 2002
summary(gss[gss$year==2002,]$educ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   12.00   13.00   13.36   16.00   20.00      12
```

The summary statistics of school year in 2012 is:

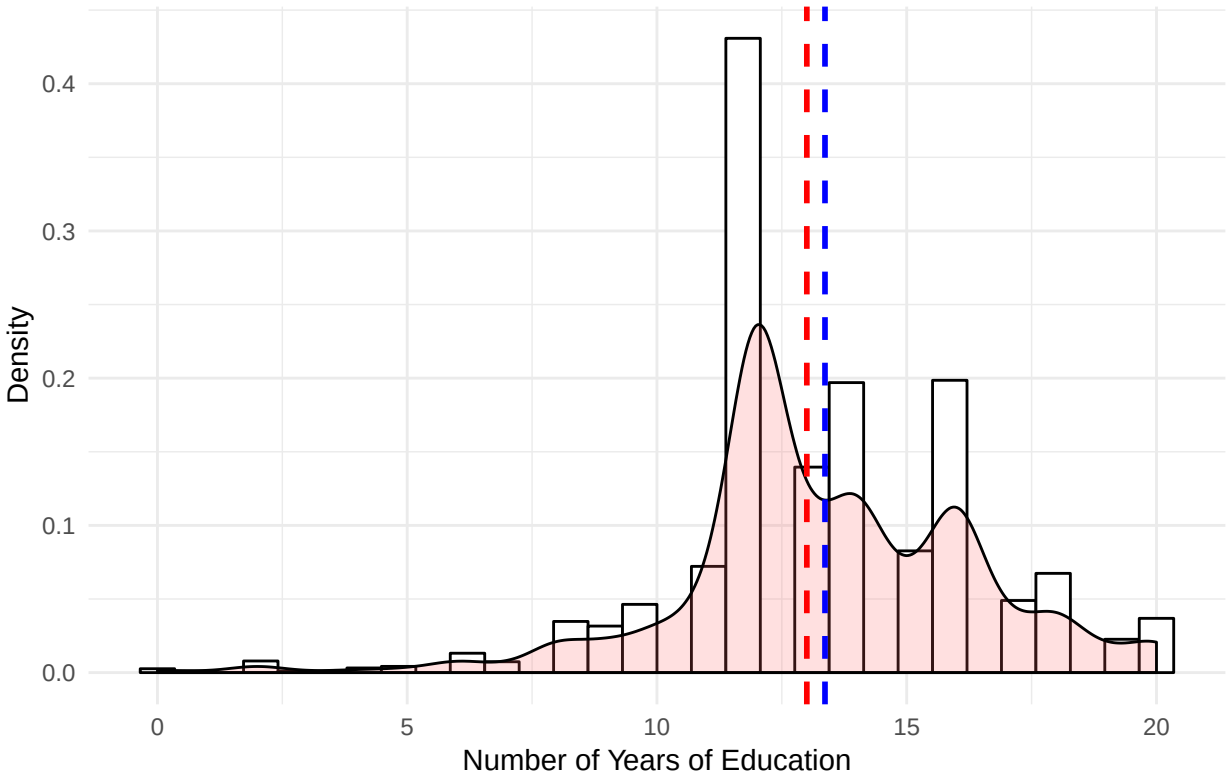
```
# Summary statistics of schooling year in 2012
summary(gss[gss$year==2012,]$educ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   12.00   13.00   13.53   16.00   20.00       2
```

According to the sample, it seems that there is a 0.17 (13.53-13.36) year of increase in the average number of years of education from 2002 to 2012. Nevertheless, whether the difference is statistically meaningful or not needs further investigation, which will be addressed in the following section.

Furthermore, we visualize the distribution of 2002 *educ* through a histogram overlaid with a density plot. 2002 *educ* is left skewed. Note that 12 null values of *educ* are dropped in order to draw the vertical lines that mark the median and the mean.

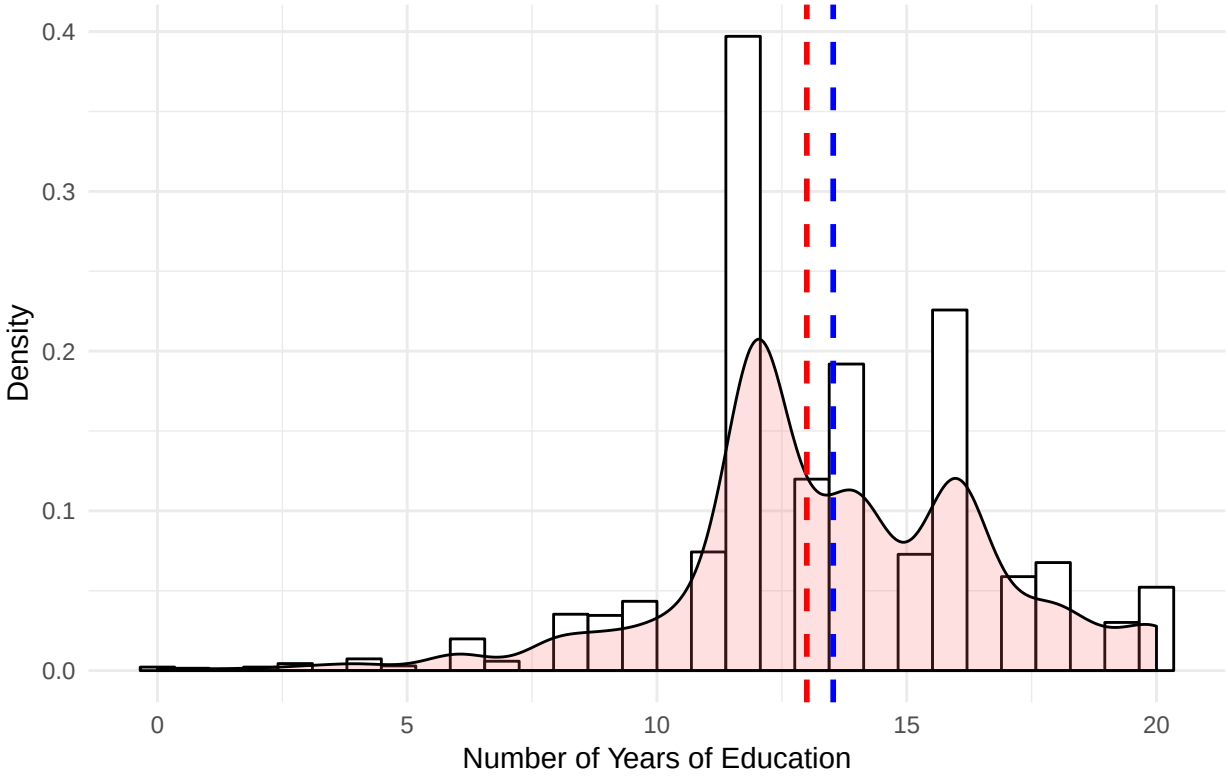
```
# Define a "histogram" function
histogram<-function(year){

  ggplot(gss[gss$year==year,][!is.na(gss[gss$year==year,]$educ),], aes(x=educ)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white")+
  geom_density(alpha=.2, fill="#FF6666")+
  geom_vline(aes(xintercept=mean(educ)),
             color="blue", linetype="dashed", size=1)+
  geom_vline(aes(xintercept=median(educ)),
             color="red", linetype="dashed", size=1)+
  xlab("Number of Years of Education")+
  ylab("Density")+
  labs(caption="Note: Red/Blue vertical line marks the median/mean value.
       The histogram is overlaid with the density plot.")+
  theme_minimal()
}

# Draw 2002 schooling year distribution
histogram(2002)
```




Note: Red/Blue vertical line marks the median/mean value.
The histogram is overlaid with the density plot.

Similarly, we visualize 2012 *educ*. It appears that the shape of the distribution has not changed too much, which makes sense.

```
# Draw 2012 schooling year distribution
histogram(2012)
```




Note: Red/Blue vertical line marks the median/mean value.
The histogram is overlaid with the density plot.

**Part 4: Inference**

In this section, we perform statistical inference addressed in Part 2. We hereby propose two hypotheses:

**H0: There is no difference in average schooling year between 2002 and 2012.**

**Ha: There is a difference in average schooling year between 2002 and 2012.**

In this case, we are comparing two independent means. We may assume that the within group independence (respondents sampled each year) and the between group independence (2002 vs. 2012) conditions are met, as people are picked randomly for the survey. In addition, the sample sizes are sufficiently large ($n$=2,765 for 2002 and $n$=1,974 for 2012). We proceed to conduct hypothesis test.

We first calculate test statistic. Below is the procedure:

```
# 2002 data
educ2002<-gss[gss$year==2002,]

# 2012 data
educ2012<-gss[gss$year==2012,]

# Test statistic
test_statistic<-(mean(educ2012$educ,na.rm=TRUE)-
                   mean(educ2002$educ,na.rm=TRUE))/
  sqrt(sd(educ2002$educ,na.rm=TRUE)^2/nrow(educ2002)+sd(educ2012$educ,na.rm=TRUE)^2/nrow(educ2012))

# Print
```

```r
print(paste("The test statistic is", test_statistic))
```

```
## [1] "The test statistic is 1.81569723291019"
```

The p-value is:

```r
# p-value
p_value<-(1-pt(test_statistic,df=min(nrow(educ2002)-1,nrow(educ2012)-1)))*2
#=pt(test_statistic,df,lower.tail=FALSE)*2
#=pt(-test_statistic,df) conservative method?

# Print
print(paste("The p-value is", p_value))
```

```
## [1] "The p-value is 0.0695685620867088"
```

p-value, which is about **0.07**, is smaller than 10% while greater than 5%. **Therefore, we reject the null hypothesis that there is no mean difference at 10% level, and we are 90% confident that the average schooling year in 2002 is different from that in 2012 in the whole U.S. population.**

### Part 5: Conclusion

In this project, we explore if there is statistical evidence that in the U.S., the average schooling year in 2012 is different from that in 2002. According to the result, we believe this is so at 90% confidence level.