Statistical inference with the GSS data

Setup

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

Part 1: Data

Abstract for the General Social Survey file 1972-2012

Since 1972, the General Social Survey (GSS) has been monitoring societal change and studying the growing complexity of American society. 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.

Subset Variable List

- Case Identification and Year
- Respondent Background Variables
- Personal and Family Information
- Attitudinal Measures - National Problems
    - Social Problem Spending
- Personal Concerns
    - Personal Concerns
- Social Concerns
    - Confidence in Institutions
- Workplace and Economic Concerns
    - Job Security and Satisfaction
    - Class and Financial Needs
    - Standadr of Living
- Controversial Social Issues
    - Abortion
    - Family Planning, Sex and Contraception
    - Suicide
    - Violent Experiences
    - Media Exposure
    - **Race Part two**
- Obligations and Responsabilities
    - Government Responsibilities

Data Sampling Methodology

Wikipedia: "*The target population of the GSS is adults (18+) living in households in the United States. The GSS sample is drawn using an* ***area probability design that randomly selects respondents in households across the United States*** *to take part in the survey. Respondents that become part of the GSS sample are from a mix of urban, suburban, and rural geographic areas. Participation in the study is* ***strictly voluntary****. However, because only about a few thousand respondents are interviewed in the main study, every respondent selected is very important to the results.*

The survey is conducted ***face-to-face*** *with an in-person interview by NORC at the University of Chicago. The survey was conducted every year from 1972 to 1994 (except in 1979, 1981, and 1992). Since 1994, it has been conducted every other year. The survey takes about* ***90 minutes*** *to administer. As of 2014, 30 national samples with* ***59,599 respondents*** *and 5,900+ variables have been collected.*"

Random selection it is thus essential to external validity, or the extent to which the researcher can use the results of the study to generalize to the larger population.

Random assignment is an aspect of experimental design in which study participants are assigned to the treatment or control group using a random procedure.Random selection requires the use of some form of random sampling. random Sampling relies on the laws of probability to select a sample that can be used to make inference to the population; this is the basis of statistical tests of significance. Random assignment is central to internal validity, which allows the researcher to make causal claims about the effect of the treatment.

Random selection, 59.599 respondents, face-to-face interviews, interview lenght 90 min, voluntary participation.

Conclusion Generalization and Causality :

- Random Selection and sample size < 10% US population. The conclusions of this study can be generalized to the entire population of US. (Random Selection)
- No causality can be drawn from this study. There is not a control group to compare thus not causal claims about the reasons of the opinion change.

Limitations:

The voluntary participation can make that many people decide to do not participate or forget about it. Since only few thousands are interviwed every single answer has a crucial importance. The generalization of the results may be affected y this.
In face-to-face interviews the person asked has a tendence to answer what she/he thinks thes person answering wants to hear. Thus the answer is not completly free, more over when the question refers to such as controversial issues

Load data

Make sure your data and R Markdown files are in the same directory. When loaded your data file will be called `gss`. Delete this note when before you submit your work.

```
#load("gss.Rdata")
load("~/Documents/Stat_Duke_Univ/2ndCourse_Inferential_Stat_Intro/project_inference/gss.RData")
```

Part 2: Research question

Variables:

`year`, `race`, `polviews`

Codebook Produced April 17, 2014: http://sda.berkeley.edu/D3/GSS12/Doc/gs12.htm (http://sda.berkeley.edu/D3/GSS12/Doc/gs12.htm)

Q. **Does people with different race think differently about being self more or less Liberal or Conservative?**

`polviews` : Think of self as Liberal or Conservative.

- Question: We hear a lot these days about liberals and conservatives in the US. It is likely that the white US population normally more acommodate that black, latinos, etc it is also less liberal than the other races?

Part 3: Exploratory data analysis

1st approach. Differences in political views between 1980-2012

There is not data taking into account the variable `race` until 1974. I will compare 1980 and 2012

```
## Filter
my_data<-data.frame(year=gss$year, race=gss$race, polviews=gss$polviews)

#summary(my_data)
```

```
## Contingency Table Race - Polviews
tabl_race_pol=table(subset(my_data, select=c(race,polviews)))
tabl_race_pol
```

```
##        polviews
## race    Extremely Liberal Liberal Slightly Liberal Moderate
##   White               933    4204             4943    15080
##   Black               316    1022              924     2492
##   Other                81     356              314      922
##        polviews
## race    Slightly Conservative Conservative Extrmly Conservative
##   White                  6602         6198                 1196
##   Black                   759          627                  238
##   Other                   330          267                   72
```

** Proportion Tables**

```
# tot_pol=colSums(tabl_race_pol)
# tot_pol
# tot_race=rowSums(tabl_race_pol[,1:7])
# tot_race
# ----
# Proportion of white (or black or other) people from the sample that says to be Extr
emly liberal, Liberal,Slightly Liberal, Moderate, Slightly Conservative, Conservative
 or Extremly Conservative.
race_pol_f1 <- prop.table(tabl_race_pol,1)
# rowSums(prop.table(tabl_race_pol,1)) ##Verify Sums Rows are = 1

# For each political point of view proportion of white, black or other race
# race_pol_c1 <- prop.table(tabl_race_pol,2)
# a2=colSums(prop.table(tabl_race_pol,2)) ##Verify Sums Columns are = 1
```

```
round(race_pol_f1,3)
```

```
##        polviews
## race    Extremely Liberal Liberal Slightly Liberal Moderate
##   White             0.024   0.107            0.126    0.385
##   Black             0.050   0.160            0.145    0.391
##   Other             0.035   0.152            0.134    0.394
##        polviews
## race    Slightly Conservative Conservative Extrmly Conservative
##   White                 0.169        0.158                0.031
##   Black                 0.119        0.098                0.037
##   Other                 0.141        0.114                0.031
```

From the data set we can see that during the period 1972-2012 the sample's proportions are:

- **Liberal**:
  - $\hat{p}_{\text{white}} = 0.107$ proportion of white people in the sample who says to be liberal. Of all white people in the sample 11% of them are liberal
  - $\hat{p}_{\text{black}} = 0.160$ proportion of black people in the sample who says to be liberal. Of all black people in the sample 16% of them are liberal
  - $\hat{p}_{\text{black}} - \hat{p}_{\text{white}} = 0.053$. In our sample the portion of black people being liberal is 5% larger than the proportion of white people being liberal.

```
delta_p.liberal=pliberal_of_b-pliberal_of_w     ;    delta_p.liberal
```

```
## [1] 0.05287291
```

- **Conservative**
  - $\hat{p}_{\text{white}} = 0.169$ proportion of white people in the sample who says to be conservative
  - $\hat{p}_{\text{black}} = 0.119$ proportion of blcak people in the sample who says to be consrvative
  - $\hat{p}_{\text{black}} - \hat{p}_{\text{white}} = -0.050$

```
# delta_p.conservative=pconservative_of_b-pconservative_of_w     ;    delta_p.conser
vative
```

Part 4: Inference

Confidence Intervals for the Difference Between Two Population Proportions or Means

First:

Hypothesis testing no Confidence Interval

$$H_o^{lib} : p_{white} = p_{black} \quad or \quad H_o : p_{white} - p_{black} = 0$$

$$H_Alib : p_{white} \neq p_{black} \quad or \quad H_A : p_{white} - p_{black} \neq 0$$

Conditions:

```
N=table(subset(my_data, select=race))
N
```

```
##
## White Black Other
## 46350  7926  2785
```

- **Independence**:
  a. Random choice of people. YES
  b. Less than the 10% of all US population. YES (only ~1000)
- $\hat{p}_{black} - \hat{p}_{white} = 0.053$

Hypothesis Testing:

$$\hat{p} = 0.0499002$$

$$\text{Null} \quad \text{proportion} : p_0 \Rightarrow p_{hat} \simeq 0.5 \simeq 1 - p_0 \simeq 1000 * 0.5 \simeq 500 \geq 10$$

Confidence Interval Estimate Assuming that the data in gss follows the normal distribution, we want to find the 95% confidence interval estimate of the difference between the liberal proportion of black people and the liberal proportion of white people, each within their own ethnic group.

Standard Error of a Difference $\hat{p}_{Black} - \hat{p}_{White}$

When two samples are independent of each other,

Standard Error for a Difference between two sample summaries:

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{SE_1^2 + SE_2^2}$$

Where the Standard Error for proportion of the sample i is:

$$SE_i = \sqrt{\frac{p_i(1 - p_i)}{n_i}}$$

A **95% Confidence Interval** for the differences between these two proportions in the population is given by:

$$\hat{p_1} - \hat{p_2} \pm z^* SE_{\hat{p_1} - \hat{p_2}}$$

With $z^* = 1.96$ for a 95%confidence interval

Data about the interviewed does not start to be recorded until 1974. From know we will reject the data 1970-74

We are 95% confident that the proportion of black people that self-identifies as liberal is between 4.4% and 6.1% higher than white people.

```
nBW.liberal<-data.frame(n=N,p_lib=race_pol_f1[,2])
nBW.liberal<-nBW.liberal[1:2,2:3]
nBW.liberal<-nBW.liberal %>%
     mutate(SE_i=p_lib*(1-p_lib)/n.Freq)
SE=sqrt(sum(nBW.liberal$SE_i^2))
ci_95_lib=c(diff(nBW.liberal[,2])-1.96*SE,diff(nBW.liberal[,2])+1.96*SE)
ci_95_lib
```

```
## [1] 0.05283939 0.05290643
```

We are 95% confident that the proportion of black people that self-identifies as liberal is between 5.28% and 5.29% higher than white people.

Is there a statistical significance to say that the portion of liberals is higher than the portion of moderates

Between 1974-2012

`Polviews` Binaire variable: Liberal or Conservative (Success vs. Failure)

Polviews only 2 labels: Liberal or Not & Exclude NA values and Moderates.

Extremly Liberal, Liberal, Slightly Liberal = Liberal Extremly Conservative, Conservative, Slightly Conservative = Conservative

Until 1974 there is not data referred to the race. We *neglect the data from the years before 1974*.

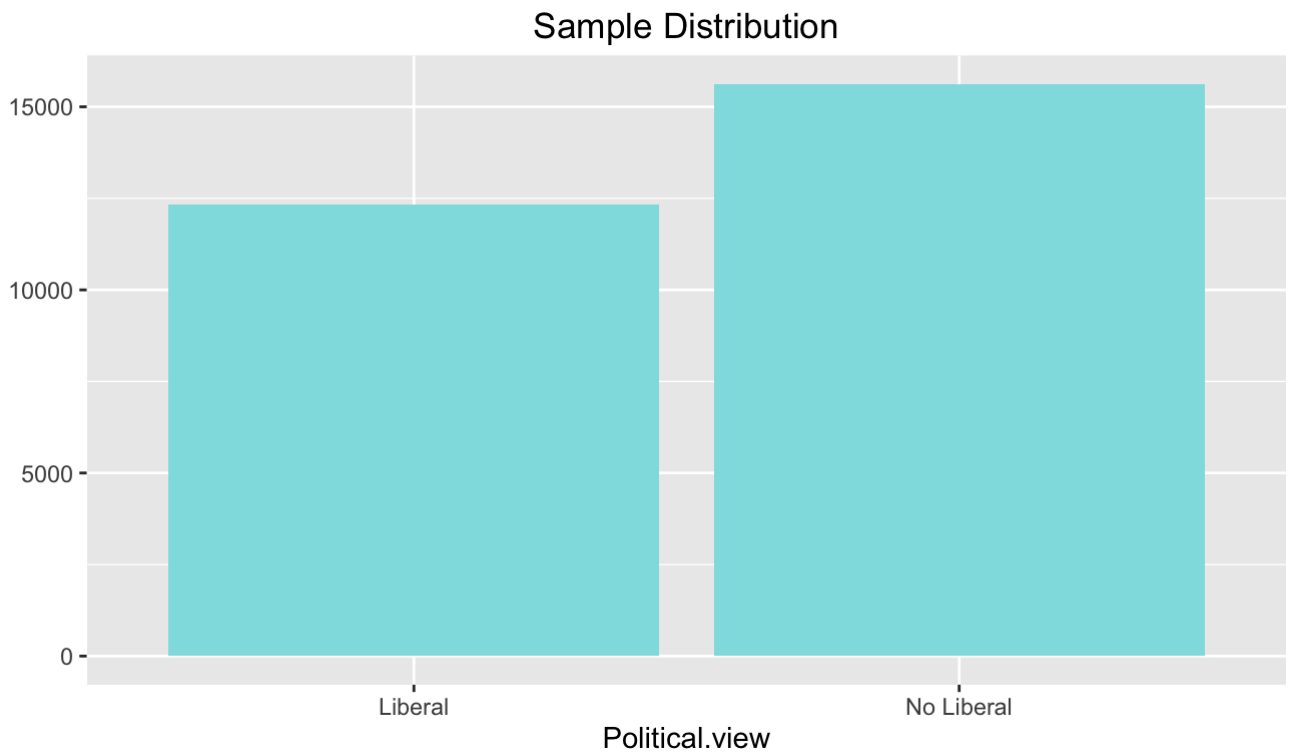
Moderates lyes in the center and we need a binare variable (Success or Failure). We *neglect the data from "Moderates"*.

```
y <- na.exclude(BW_LibCon)
BW_LibCon <- y %>% filter(polviews != "Moderate")
Lib_Not <-  BW_LibCon %>%
  mutate(Political.view = as.factor(ifelse (polviews != "Slightly Liberal" & polviews
 !="Liberal" & polviews !="Extremely Liberal" , "No Liberal","Liberal") ))
summary(Lib_Not)
```

```
##       year         race                         polviews
##  Min.   :1974   White:24076   Extremely Liberal    :1249
##  1st Qu.:1984   Black: 3886   Liberal              :5226
##  Median :1993                 Slightly Liberal     :5867
##  Mean   :1993                 Moderate             :   0
##  3rd Qu.:2002                 Slightly Conservative:7361
##  Max.   :2012                 Conservative         :6825
##                               Extrmly Conservative :1434
##     Political.view
##  Liberal   :12342
##  No Liberal:15620
##
##
##
##
##
```

```
inferencia<-inference(y=Political.view,data=Lib_Not, statistic="proportion",type="ci"
, method = "theoretical", success = "Liberal")
```

```
## Single categorical variable, success: Liberal
## n = 27962, p-hat = 0.4414
## 95% CI: (0.4356 , 0.4472)
```



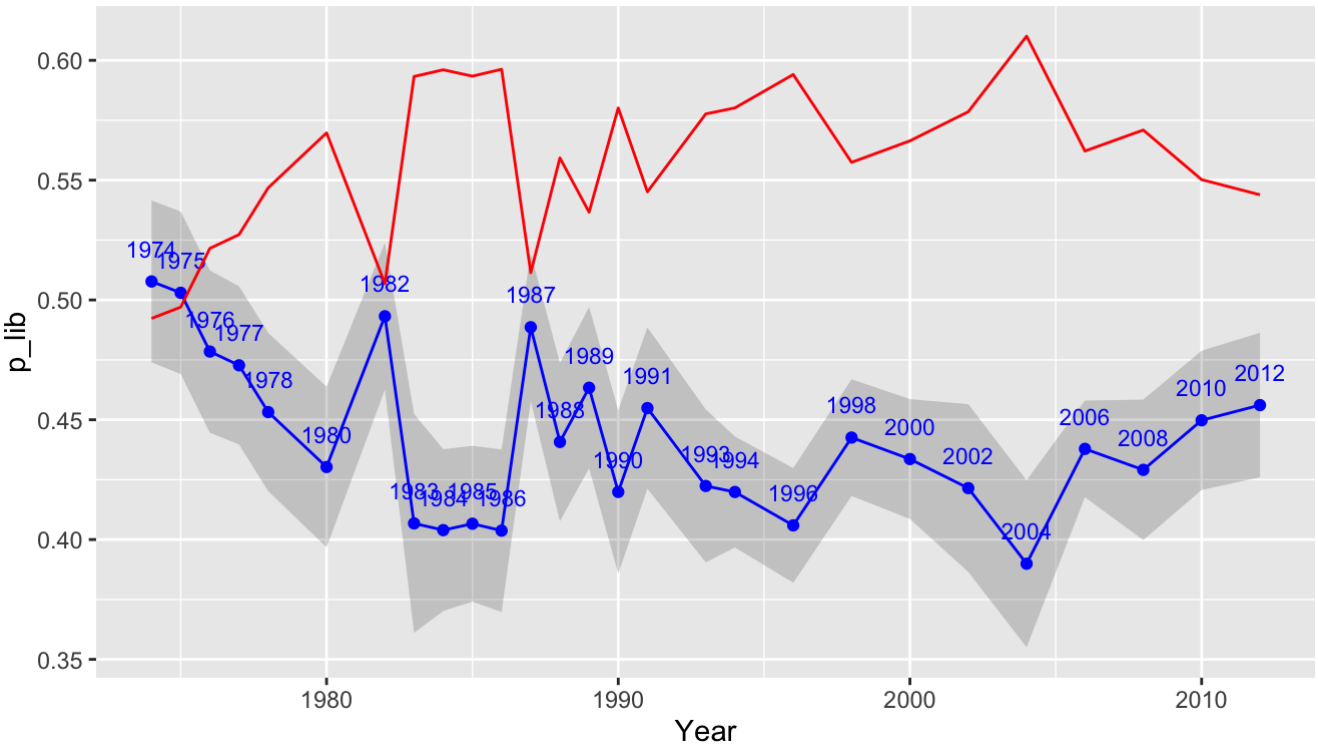

```
table_Lib_Not=table(subset(Lib_Not,select=c(year,Political.view)))
years<-unique(Lib_Not$year)
n.lib=table_Lib_Not[1:27]
n.cons=table_Lib_Not[28:54]
p_lib=n.lib/(n.lib+n.cons)
p_cons=n.cons/(n.lib+n.cons)
p_Lib_Not=as.data.frame(cbind(Year=years,p_lib,p_cons))
```

We are 95% confident that between 1974-2012 the proportion of people that self-identifies as liberal is between 43.56 % and 44.72 % of the total US population.

```
lib_y <- Lib_Not %>% filter(year == 1974)
Infer<-inference(y=Political.view,data=lib_y, statistic="proportion",type="ci", metho
d = "theoretical", success = "Liberal", show_eda_plot = "FALSE")
x <- c(Year=1974,SE=Infer$SE,ME=Infer$ME,CI=Infer$CI)
# the years the study is made are
years_vect<-unique(Lib_Not$year)
# there are years between 1974-2012 not included, as :1979,81,92,95+2i...2012
years_vect
```

```
for (Year in years_vect) {
  lib_y<-Lib_Not %>% filter(year == Year)
  Infer<-inference(y=Political.view,data=lib_y, statistic="proportion",type="ci",
     method = "theoretical", success = "Liberal",show_eda_plot = "FALSE")
  x2<-c(Year, Infer$SE, Infer$ME,Infer$CI)
  x<-data.frame(rbind(x,x2))
}
x<-x[2:28,]
```

```
# change atomic vector into a Data Frame
colnames(x)=c("Year","SE","ME","CI.min","CI.max")
CI_t=data.frame(cbind(x),p_lib,p_cons)
ggplot(CI_t,aes(x=Year,y=p_lib)) +
geom_ribbon(aes(ymin=CI.min,ymax=CI.max), alpha=0.2) +
geom_text(aes(label=Year), vjust=-1.5, colour="blue", size=3, angle=0) +
geom_line(colour="blue") +
geom_point(colour="blue") +
geom_line(aes(x=Year,y=p_cons),colour="red")
```




```
#geom_line(aes(x=Year,y=mean(p_lib)),colour="red") +
```

We can see that the amount of liberals decreases significantly at the beginning of the 80's

In the early 80's US was in a severe recession.
"The early 1980s recession describes the severe global economic recession affecting much of the developed world in the late 1970s and early 1980s. The United States and Japan exited the recession relatively early. The peak of the recession occurred in November and December 1982, when the nationwide unemployment rate was 10.8%, highest since the Great Depression."(Wikipedia)

Although this work can not conclude causalities we can especulate that this economical recession may have afected the US citizens political poit of views.