Statistical inference with the GSS data

## Setup

### Load packages

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

### 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")

## Part 1: Data

The survey randomly samples all household across the United States, eligible individual within each household are given equal oppotunity to answer the interview. Starting from 2004, GSS began to use a two-stage subsampling design for nonresponse. The second stage focused on gaining cooperation from people who did not response in the first stage. Because random sampling is employed, the results are likely to be able to generalize to the whole U.S. population, with cautions on person-level variables, since people from large households have lower chance to be selected. In addition, we can weight up the responses obtained from second-stage sampling beginning in 2004, in order to compensate for the non-response bias in the surveys before 2004. There is no random assignment, so from the survey we can only infer correlation but not causation. \* \* \*

## Part 2: Research question

What are some factors related to people’s political view? One can easily name a lot: sex, race, education level, social status… As an international student who come to California for graduate school, my personal experience is that my classmates are more on the liberal side — there are some riots in our campus when president Trump won the election. So I am curious to see if there is a significant correlation between U.S. citizen’s political opinion and their education level or geographical location.

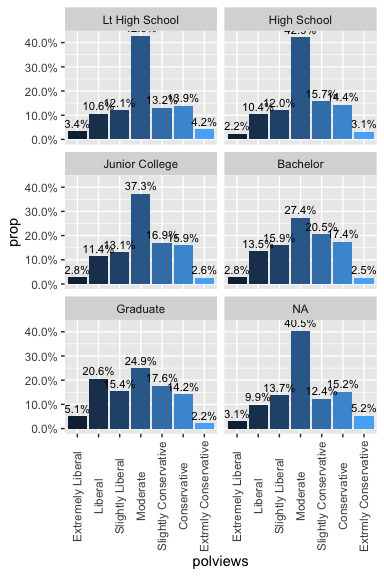
## Part 3: Exploratory data analysis

First select the three variables we are interested in: people’s highest degree, geographical location, and their political views. We will exclude rows with missing political views.

mydata <- gss %>% select(degree, region, polviews)  
mydata <- mydata %>% filter(!is.na(polviews))

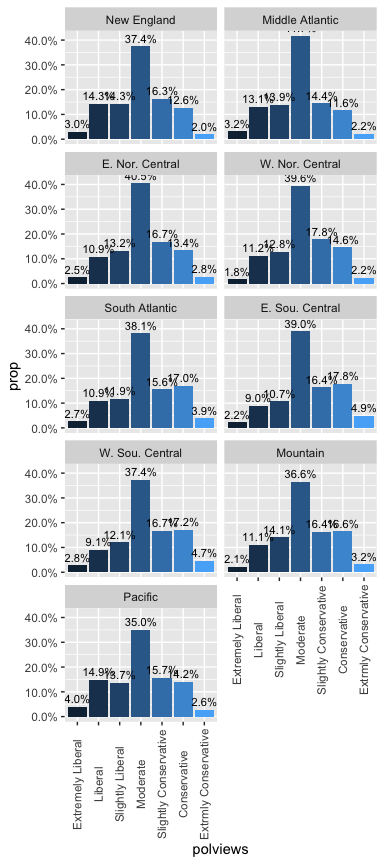
Next, let’s grouping people based on their education degree, and see what the distribution of political views looks like within each group. We find that the the distribution belongs to people with a graduate degree is more uniform compare to others. This indicates that people recieved higher education have more diverse political views: they tend to have a stronger political opinion – either liberal or conservative, while people with lower degree (like lower than high school.high school) tend to find themselves on the middle part of the political view spectrum.

ggplot(mydata, aes(x= polviews, group=degree)) +   
 geom\_bar(aes(y = ..prop.., fill = ..x..), stat="count") +  
 geom\_text(aes( label = scales::percent(..prop..),  
 y= ..prop.. ), stat= "count", vjust = -.5, size = 3) +  
 facet\_wrap(~degree, ncol=2) +  
 scale\_y\_continuous(labels = scales::percent) +  
 theme(axis.text.x=element\_text(angle=90,hjust=0.5,vjust=0.5), legend.position="none")



Similar to the analysis above, we look at how political view and geographical location are related. It seems people in the pacific are more likely to consider themselves as liberal versus people in the rest of the country. We will later use inference to findout whether there is indeed a significant difference between people’s political view and their location.

ggplot(mydata, aes(x= polviews, group=region)) +   
 geom\_bar(aes(y = ..prop.., fill = ..x..), stat="count") +  
 geom\_text(aes( label = scales::percent(..prop..),  
 y= ..prop.. ), stat= "count", vjust = -.5, size = 3) +  
 facet\_wrap(~region, ncol=2) +  
 scale\_y\_continuous(labels = scales::percent) +  
 theme(axis.text.x=element\_text(angle=90,hjust=0.5,vjust=0.5), legend.position="none")



## Part 4: Inference

#### Inference between education level and the extremeness of political view

The first hypothese test is on whether the extremeness of people’s political view is significantly different among people who recieved higher education and lower education. To describe “extremeness”, we define a new numerical variable “pol\_extrm” from the original variable “polviews”, larger value of this variable means the stronger political views.

mydata <- mydata %>%  
 mutate(pol\_extrm = NA)  
mydata$pol\_extrm[mydata$polviews == 'Moderate'] = 0  
mydata$pol\_extrm[mydata$polviews == 'Slightly Conservative' | mydata$polviews == 'Slightly Liberal'] = 1  
mydata$pol\_extrm[mydata$polviews == 'Conservative' | mydata$polviews == 'Liberal'] = 2  
mydata$pol\_extrm[mydata$polviews == 'Extrmly Conservative' | mydata$polviews == 'Extrmly Liberal'] = 3

We also group people into lower and higher education groups, higher education group includes graduate and bachelor degree, lower education group includes lower than high school, high school and junior college, missing values are filtered.

testdata <- mydata %>%  
 filter(!is.na(degree))  
testdata <- mydata %>%  
 mutate(education = ifelse(degree %in% c("Graduate", "Bachelor"), "high", "low"))  
testdata %>% group\_by(education) %>%  
 summarise(mean\_political\_view = mean(pol\_extrm, na.rm=TRUE), std=sd(pol\_extrm, na.rm=TRUE), count=n())

## # A tibble: 2 x 4  
## education mean\_political\_view std count  
## <chr> <dbl> <dbl> <int>  
## 1 high 1.11 0.834 10469  
## 2 low 0.893 0.902 37407

We see the mean extremeness of high and low education groups are different, we will use inference to determine whether this difference is due to chance or not. So our hypotheses are: : There is no significant difference in the average extremeness of political views between high and low education people. : There is some difference in the average extremeness of political views between high and low education people. We will use t-statistic for comparing two means, although we note because of the large sample size the t-distribution would be nearly normal.

xdiff = 1.1073586-0.8932207  
SE = sqrt(0.8344328^2/10469 + 0.9016376^2/37407)  
t = (xdiff)/SE  
message("t test statistic: ", t)

## t test statistic: 22.7959662447319

p\_value = pt(t, 10468, lower.tail = FALSE)  
message("p\_value: ", p\_value)

## p\_value: 1.32538002026575e-112

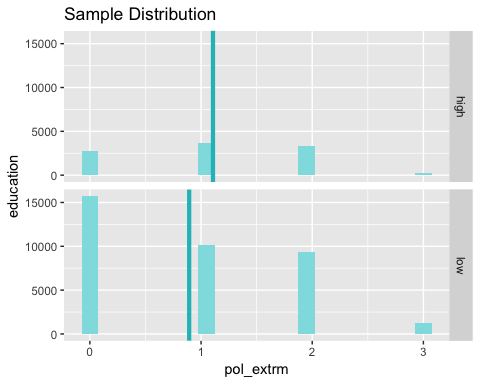
t\_ = qt(0.975, 10468)  
ci = c(xdiff-t\_\*SE, " ", xdiff+t\_\*SE)  
message("95% confidence interval: ", ci)

## 95% confidence interval: 0.195724506115611 0.232551293884389

We can also use the inference function to do hypothese test and compute confidence interval.

inference(y=pol\_extrm, x=education, data=testdata, type='ci', statistic="mean", method="theoretical", alternative="twosided")

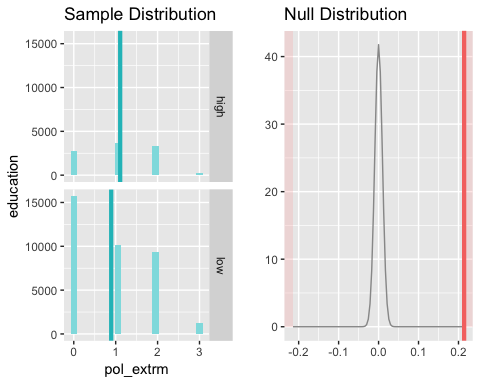
## Response variable: numerical, Explanatory variable: categorical (2 levels)  
## n\_high = 10097, y\_bar\_high = 1.1074, s\_high = 0.8344  
## n\_low = 36449, y\_bar\_low = 0.8932, s\_low = 0.9016  
## 95% CI (high - low): (0.1954 , 0.2329)



inference(y=pol\_extrm, x=education, data=testdata, type='ht', statistic="mean", method="theoretical", alternative="twosided")

## Warning: Missing null value, set to 0

## Response variable: numerical  
## Explanatory variable: categorical (2 levels)   
## n\_high = 10097, y\_bar\_high = 1.1074, s\_high = 0.8344  
## n\_low = 36449, y\_bar\_low = 0.8932, s\_low = 0.9016  
## H0: mu\_high = mu\_low  
## HA: mu\_high != mu\_low  
## t = 22.4154, df = 10096  
## p\_value = < 0.0001

 The resulting p-value is very small. Consistently, the confidence interval does not contain 0. Thus we reject the null hypothese that the average extremeness of political views are the same with low and high education groups.

## Inference between geographical location and the extremeness of political view

In this case, our hypothesis will be: Political views and people’s geographical location are independent. Political views and people’s geographical location are dependent. For simplicity we will pick two geographical location: pacific (where California belongs to) and W. Sou. Central (where Texas belongs), and see the distribution of political views is significantly different among these two regions.

#create contingency table  
testdata <- mydata %>%  
 filter(region %in% c("Pacific", "W. Sou. Central"))  
testdata$region <- factor(testdata$region)  
contingency\_tbl = table(testdata$region, testdata$polviews)  
message("contingency table")

## contingency table

contingency\_tbl

##   
## Extremely Liberal Liberal Slightly Liberal Moderate  
## W. Sou. Central 122 400 531 1645  
## Pacific 256 966 886 2265  
##   
## Slightly Conservative Conservative Extrmly Conservative  
## W. Sou. Central 734 758 207  
## Pacific 1017 917 171

#based on contingency table, add margins and compute the expected observations.  
m0 = margin.table(contingency\_tbl)  
m1 = margin.table(contingency\_tbl, margin=1)  
m2 = margin.table(contingency\_tbl, margin=2)  
expect\_table = m1 %o% m2 / m0  
message("expected values: row\_sum \* col\_sum / total\_sum")

## expected values: row\_sum \* col\_sum / total\_sum

expect\_table

##   
## Extremely Liberal Liberal Slightly Liberal Moderate  
## W. Sou. Central 152.8337 552.3036 572.924 1580.898  
## Pacific 225.1663 813.6964 844.076 2329.102  
##   
## Slightly Conservative Conservative Extrmly Conservative  
## W. Sou. Central 707.9675 677.2391 152.8337  
## Pacific 1043.0325 997.7609 225.1663

#compute test statistic and p\_value  
#chi = sum (Oi-Ei)^2/Ei  
chi = sum((contingency\_tbl-expect\_table)^2/expect\_table)  
message("chi square test statistic: ", chi)

## chi square test statistic: 140.465456696108

p\_value = pchisq(chi, (2-1) \* (7-1), lower.tail = FALSE)  
message("p\_value: ", p\_value)

## p\_value: 7.99334160899101e-28

We can also do the inference using function chisq.test(). The X-squared value agree with our manual calculation above, the p-value is a bit different (likely because it is too small, the function did not seek to achieve unnecessary numerical accuracy).

chisq.test(testdata$region, testdata$polviews)

##   
## Pearson's Chi-squared test  
##   
## data: testdata$region and testdata$polviews  
## X-squared = 140.47, df = 6, p-value < 2.2e-16

Because of the very small p-value, we will reject the null hypotheses that the political views and people’s geographical location are independent, and conclude that people’s political views vary when moving from pacific to W. Sou. Central area.