Statistical inference with the GSS data

February 28, 2018

Objective

This document serves the purpose of a final evaluation of the 5 week Inferential Statistics intro course by Duke University. The data of interest is General Social Survey (**GSS**) dataset. It can be downloaded from here (https://d3c33hcgiwev3.cloudfront.net/_5db435f06000e694f6050a2d43fc7be3_gss.Rdata?Expires=1519948800&Signature=YTJ1pXeYJIIvTIBXm4umVd0y7kRFVC30PRqY0DCNBcCyxhMCeB235JB5igj7OWUT2hqxeK-R-uW8IPgeFpB5Yd0VeYFb92NvPQIa5B3tdyEQHIUZMZHdPRtNcrfx8voAJeD7jVBQ7E-HF5SOTX1lSQqLK0JyCNRq7-9IQVSKJKc_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A):

Setup

This section loads the required packages and data ### Load packages

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

Load data

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

Part 1: Data

The General Social Survey (GSS) is a sociological survey used to collect information and keep a historical record of the concerns, experiences, attitudes, and practices of residents of the United States.

Data Collection: The vast majority of GSS data is obtained in face-to-face interviews. Computer-assisted personal interviewing (CAPI) began in the 2002 GSS. Under some conditions when it has proved difficult to arrange an in-person interview with a sampled respondent, GSS interviews may be conducted by telephone.

Sampling: The GSS sample is drawn using an area probability design that randomly selects respondents in households across the United States. Respondents that become part of the GSS sample are from a mix of urban, suburban, and rural geographic areas.

Generalizability: The inferences made from this data set are generalizable because respondents are randomly selected. The results are generalizabile to the GSS Target population, which is Adults (18+) living in households in the United States. Residents of institutions and group quarters are out-of-scope.

Causality: This is an observational study with no random assignment. We cannot infer causation.

In addition, potential biases are associated with non-response because this is a voluntary in-person survey that takes approximately 90 minutes. Some potential respondents may choose not to participate.

Part 2: Research question

The question of my interest for this assignment is:

> **Is there a relationship between gun ownership and a respondent's region of residence?**

I would also lik focus my attention over the gun ownership before year 2001 and after year 2001.The choice for these years isn't random. It is based on a historical event that took place on 9/11. And I would also explore the variable "fear"

In order to answer this question I'll use the following variables:
* year - in which year the response was obtained
* region - in which region the response was obtained
* owngun - whether the respondent owns a gun or not
* fear - whether the respondent was afraid to walk at night in a neighborhood

I will remove the unnecessary variables and NAs from the dataset in order of better readability and simplicity to work with the dataset. I would also remove the answer "Refused" of the question regarding gun ownership for the same reasons.

Part 3: Exploratory data analysis

In this section we I will conduct the Exploratory data analysis

```
# check dimensions of the dataset 'gss'
dim(gss)
```

```
## [1] 57061   114
```

We have 57061 observations and 114 variables

Now I will reduce GSS dataset to just the columns of interest: 'year', 'region', owngun', 'fear'; remove the NAs from the dataset; remove the answer "Refused" to the question of gun ownership, segmenting the data by a new column that points out whether the interview was conducted before or after 2001

```
# making a new smaller dataset to work with
dff <- gss %>%
       select(year,region,owngun, fear) %>%
       na.omit() %>%
       filter(owngun != "Refused") %>%
       mutate(dev2001 = as.factor(ifelse(year >= 2001, "after 2001","before 2001")))
```

```
## Warning: package 'bindrcpp' was built under R version 3.3.3
```

```
# check dimensions of the new dataset 'dff'
dim(dff)
```

```
## [1] 33878     5
```

We have now 33878 observation and 5 variables. Next we make a summary statistic of our new dataset

```
# check the summary statistic of the 'dff' dataset
summary(dff)
```

```
##       year                  region         owngun       fear
##  Min.   :1973   South Atlantic :6544   Yes    :13944   Yes:13840
##  1st Qu.:1982   E. Nor. Central:6387   No     :19934   No :20038
##  Median :1991   Middle Atlantic:5018   Refused:    0
##  Mean   :1991   Pacific        :4535
##  3rd Qu.:2000   W. Sou. Central:3094
##  Max.   :2012   W. Nor. Central:2467
##                 (Other)        :5833
##         dev2001
##  after 2001 : 7582
##  before 2001:26296
##
##
##
##
##
```

We see that the observational study took place between 1973 and 2012. We see that more people don't own a gun and aren't afraid to walk at night in a neighbourhood. And we see that the observations before 2001 are 26296 and only 7582 after 2001.

```
# I'll make a table for the gun ownership in every of the 6 regions for the counts
df_table <- table(dff$region, dff$owngun)
df_table
```

```
##
##                    Yes   No Refused
##   New England      391 1167       0
##   Middle Atlantic 1330 3688       0
##   E. Nor. Central 2672 3715       0
##   W. Nor. Central 1208 1259       0
##   South Atlantic  2963 3581       0
##   E. Sou. Central 1354  920       0
##   W. Sou. Central 1492 1602       0
##   Mountain         971 1030       0
##   Pacific         1563 2972       0
```

```
# I'll delete the column with answer "Refused"
df_table <- df_table[,-3]
df_table
```

```
##
##                    Yes   No
##   New England      391 1167
##   Middle Atlantic 1330 3688
##   E. Nor. Central 2672 3715
##   W. Nor. Central 1208 1259
##   South Atlantic  2963 3581
##   E. Sou. Central 1354  920
##   W. Sou. Central 1492 1602
##   Mountain         971 1030
##   Pacific         1563 2972
```

```
# and now I'll make a table for the gun ownership in every of the 6 regions displayed in propor
tions for that region
prop.table(df_table,1)
```

```
##
##                          Yes        No
##   New England     0.2509628 0.7490372
##   Middle Atlantic 0.2650458 0.7349542
##   E. Nor. Central 0.4183498 0.5816502
##   W. Nor. Central 0.4896636 0.5103364
##   South Atlantic  0.4527812 0.5472188
##   E. Sou. Central 0.5954266 0.4045734
##   W. Sou. Central 0.4822237 0.5177763
##   Mountain        0.4852574 0.5147426
##   Pacific         0.3446527 0.6553473
```

We see that in E. Sou. Central Region the percentage of respondents who answered “yes” to the question whether they own a gun or not is the highest - nearly 60% and smallest in New England - 25%

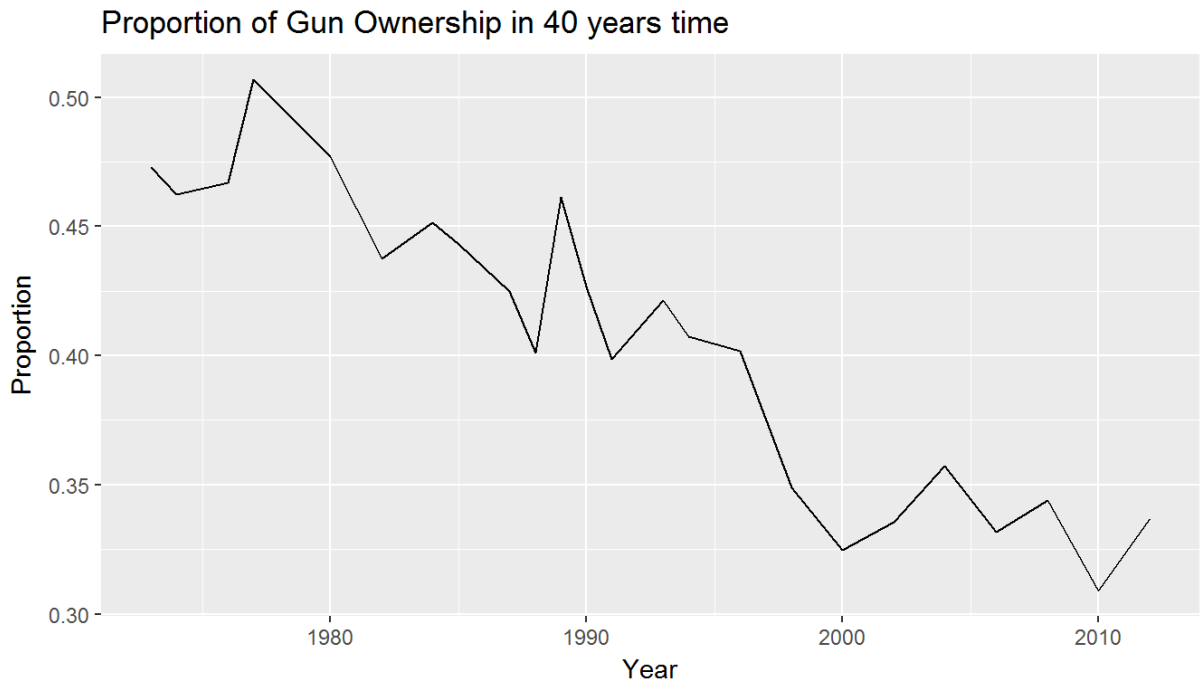
```
# I think there will be even more interesting to make a table of the proportion of positive res
ponses to the question of gun ownership segmented as well for the years before and after 2001 a
nd of course the proportion of fear is also curious to me
my_dff <- dff %>%
      group_by(region,dev2001) %>%
      mutate(own_g = ifelse(owngun == "Yes", 1,0)) %>%
      mutate(fear_c = ifelse(fear == "Yes", 1,0)) %>%
      summarise(propGun = mean(own_g),
                propFear = mean(fear_c))
      # arrange(desc(propGun, propFear))

my_dff
```

```
## # A tibble: 18 x 4
## # Groups:   region [?]
##              region     dev2001   propGun  propFear
##              <fctr>      <fctr>     <dbl>     <dbl>
##  1     New England  after 2001 0.2327044 0.3050314
##  2     New England before 2001 0.2556452 0.4024194
##  3 Middle Atlantic  after 2001 0.2191358 0.3518519
##  4 Middle Atlantic before 2001 0.2760751 0.4485912
##  5 E. Nor. Central  after 2001 0.3492787 0.3113136
##  6 E. Nor. Central before 2001 0.4362919 0.3824458
##  7 W. Nor. Central  after 2001 0.4421907 0.2657201
##  8 W. Nor. Central before 2001 0.5015198 0.3434650
##  9  South Atlantic  after 2001 0.3422195 0.3719777
## 10  South Atlantic before 2001 0.4889475 0.4662340
## 11 E. Sou. Central  after 2001 0.5154867 0.2986726
## 12 E. Sou. Central before 2001 0.6152580 0.4291987
## 13 W. Sou. Central  after 2001 0.4119948 0.4067797
## 14 W. Sou. Central before 2001 0.5053717 0.4585303
## 15        Mountain  after 2001 0.4086022 0.2706093
## 16        Mountain before 2001 0.5148995 0.3700624
## 17         Pacific  after 2001 0.2692308 0.3589744
## 18         Pacific before 2001 0.3685739 0.4812663
```

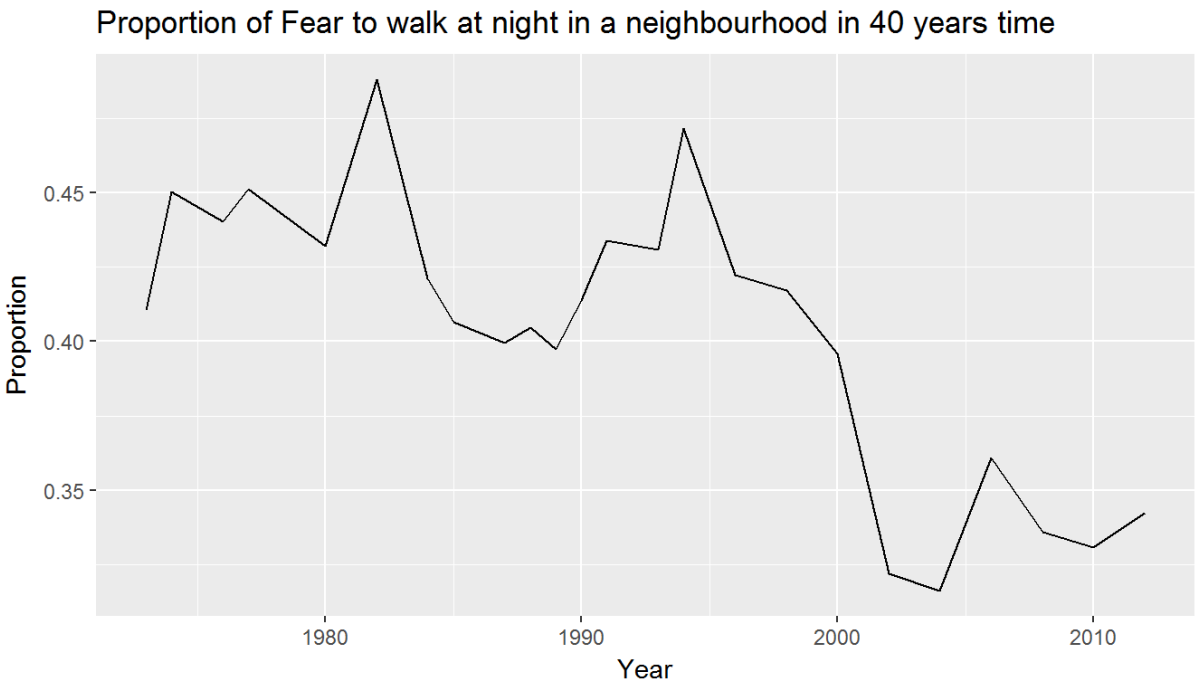
We can see here that people responded positvly to the question of gun ownership and were more afraid to walk at night in a neighbourhood before 2001 rather than after 2001.

Now, I would like to make a visual of the precentage of gun ownership for the period of this observational study and a visual of the percentage of people who are afraid to walk at night in a neighbourhood, again for the entire period of the study.

```
# making the plot of gunownership
gss %>%
      select(year,owngun) %>%
      na.omit() %>%
      group_by(year) %>%
      mutate(own_g = ifelse(owngun == "Yes", 1,0)) %>%
      summarise(propGun = mean(own_g)) %>%
      ggplot(aes(year, propGun)) +
      geom_line() +
      labs(x = "Year",y = "Proportion",title = "Proportion of Gun Ownership in 40 years time")
```

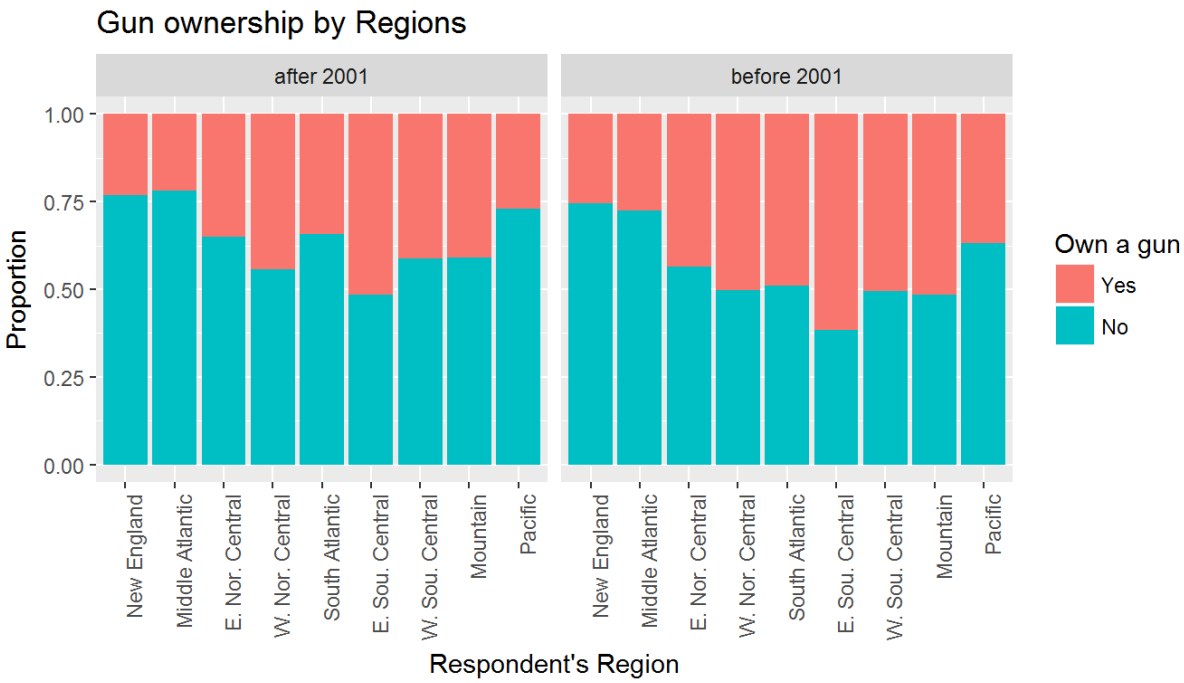



```
# making the plot of fear
gss %>%
      select(year,fear) %>%
      na.omit() %>%
      group_by(year) %>%
      mutate(feard = ifelse(fear == "Yes", 1,0)) %>%
      summarise(propF = mean(feard)) %>%
      ggplot(aes(year, propF)) +
      geom_line() +
      labs(x = "Year",y = "Proportion",title = "Proportion of Fear to walk at night in a neighb
ourhood in 40 years time")
```

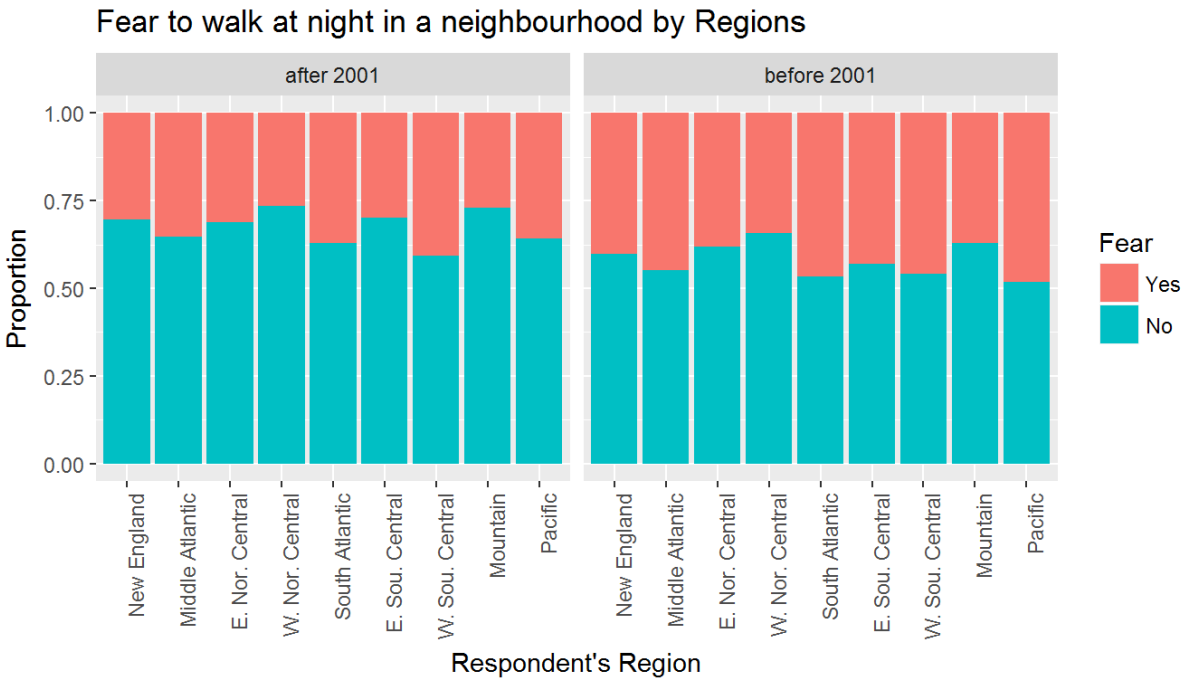


What we see here is a significant decline of the gun ownership in total of the 40 years of this observational study. We can also see the increase in gun ownership after year 2000, which is a historical moment I've mentioned in the beginning of this analyses. One other thing that cought my attention is the lowest percentage of gun ownership in year 2010. And what about fear? Fear found its peak after 1980, in about 1982,1983 at a level of a bit less than 50% = meaning that nearly half of the respondents from the study were afraid to walk at night during 1982/1983 year. We observe a decline in 2004 of a bit less than 32.5% of the people wer afraid.

```
# I would like to make a plot displaying gun ownership in every region segmented by the year be
fore and after 2001
ggplot(dff, aes(x = region,fill = owngun)) +
      geom_bar(position = "fill") +
      labs(x = "Respondent's Region",y = "Proportion",title = "Gun ownership by Regions") +
      scale_fill_discrete(name = "Own a gun") +
      facet_grid(~dev2001) +
      theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



```
# I would like to make a plot displaying fear to walk in a neighbourhood in every region segmen
ted by the year before and after 2001
ggplot(dff, aes(x = region,fill=fear)) +
      geom_bar(position = "fill") +
      labs(x = "Respondent's Region",y = "Proportion",title = "Fear to walk at night in a neigh
bourhood by Regions") +
      scale_fill_discrete(name = "Fear") +
      facet_grid(~dev2001) +
      theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



This visual comparisment shows us that there is yes indeed decline in gun ownership between the two segments of our data before/after 2001. What the second visual also shows us is the increase in fear of walking in a neighbourhood.

Part 4: Inference

In this section I will perform statistical inference. And try to answer the research question from section 2 -mIs there a relationship between gun ownership and a respondent's region of residence?

```
1. Define hypothesis
2. Choose statistical method
3. Check for conditions
4. Perform the inference tests
5. Interpret the results
```

1. Define hypothesis

The **null hypothesis** (H0): the region of the respondent independent to a positive answer of the question whether the respondent owns a gun. The **alternative hypothesis** (HA):the region of the respondent dependent to a positive answer of the question whether the respondent owns a gun.

2. Choose statistical method

In order to answer the research question /Is there a relationship between gun ownership and a respondent's region of residence?/ I will use chi-square test of independence. We use this test when comparing 2 categorical variables where one of the variables has more than 2 levels.

3. Check for conditions

The key conditions for the chi square test of independence are: Independence between observations. This is assumed to be true based on the sampling methodology used in the GSS, as it uses random sampling. Furthermore, the size of the sample is less than 10% of the population, and each result is only counted in one cell.
Sample size.

```
sum(df_table <= 10)
```

```
## [1] 0
```

We have at least 10 counts for each cell.

4. Perform the inference tests

Now we perform the inference calculation using the chi-square test.

```
# perform the chi-square test.
chisq.test(df_table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  df_table
## X-squared = 1229.9, df = 8, p-value < 2.2e-16
```

5. Interpret the results
Since the p-value is below alpha (0.05), we can conclude that there is sufficient evidence to reject H0 (null hypothesis). In context of the research question, this means, that region is related to gun ownership. But I can't make a causal statment since this is an observational data.
