Statistical inference with the GSS data

Setup

Load packages

```
#knitr::opts_chunk$set(fig.width=12, fig.height=12)

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

Load data

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

Part 1: Data

How is sample collected?

GSS collects data to understand trends in attitudes, behaviors, and attributes of American society.Most of the GSS data from 1972 is collected from face-to-face interviews. From 2002, these interviews came across a minor change. Personal interviews are changed into computer assisted. Whenever there is no possibility of doing in-person interview, survey is carried out through telephone.

How this sampling method effects the generalizability and casuality?

This is a observational study as data is being collected for certain period of time and stats are dervied from this. Hence, the results can establish a correlation but not make a casual statements. We can generalize the results at large as the sample selected here is obtained from random sampling.

Part 2: Research question

1990 is considered an important year in early history of internet. First web server was created and World Wide Web was founded. Considering this year as point of interest, Is there a relationship between level of education before 1990 and after 1990?

Part 3: Exploratory data analysis

For this test, columns needed for data set are educ and year

```r
# selecting only necessary columns
gss <- gss %>%  select("educ","year")

# checking sample data
head(gss)
```

```
##   educ year
## 1   16 1972
## 2   10 1972
## 3   12 1972
## 4   17 1972
## 5   12 1972
## 6   14 1972
```

Checking NA's in educ column:

```r
gss %>% select(educ) %>% is.na() %>% table()
```

```
## .
## FALSE  TRUE
## 56897   164
```

Cheking NA's in year column:

```r
gss %>% select(year) %>% is.na() %>% table()
```

```
## .
## FALSE
## 57061
```

There are No NA's in "year". Handling NA's in educ column by filling them with median of the column

```r
# filling NA's with median of the column and this is a categorical variable

gss$educ[is.na(gss$educ)]  <- median(gss$educ, na.rm = TRUE)
gss %>% select(educ) %>% is.na() %>% table()
```

```
## .
## FALSE
## 57061
```

NA's in the education column are resolved.

```r
# Total number of 'year' or unique items in 'year' columns

length(unique(gss$year))
```

```
## [1] 29
```

For the hypotheis that is framed above, we need 'year' variable to be rolled up into two levels. 'before-1990' & 'after 1990'

```r
gss$year <- ifelse(gss$year <= 1990, "before-1990","after-1990")
table(gss$year)
```

```
##
##  after-1990 before-1990
##       30796       26265
```

Exploring the education column data

```r
# Frequencies of educ column

gss %>% select("educ") %>% table() %>% sort()
```

```
## .
##     1     2     0     3     4     5     6    19     7    20    17     9
##    41   142   151   238   309   386   752   760   845  1157  1684  1920
##    18    15     8    10    11    13    14    16    12
##  1977  2513  2598  2635  3396  4742  6170  6988 17657
```

```r
# Maximun of educ column in gss data set

print(max(gss$educ))
```

```
## [1] 20
```

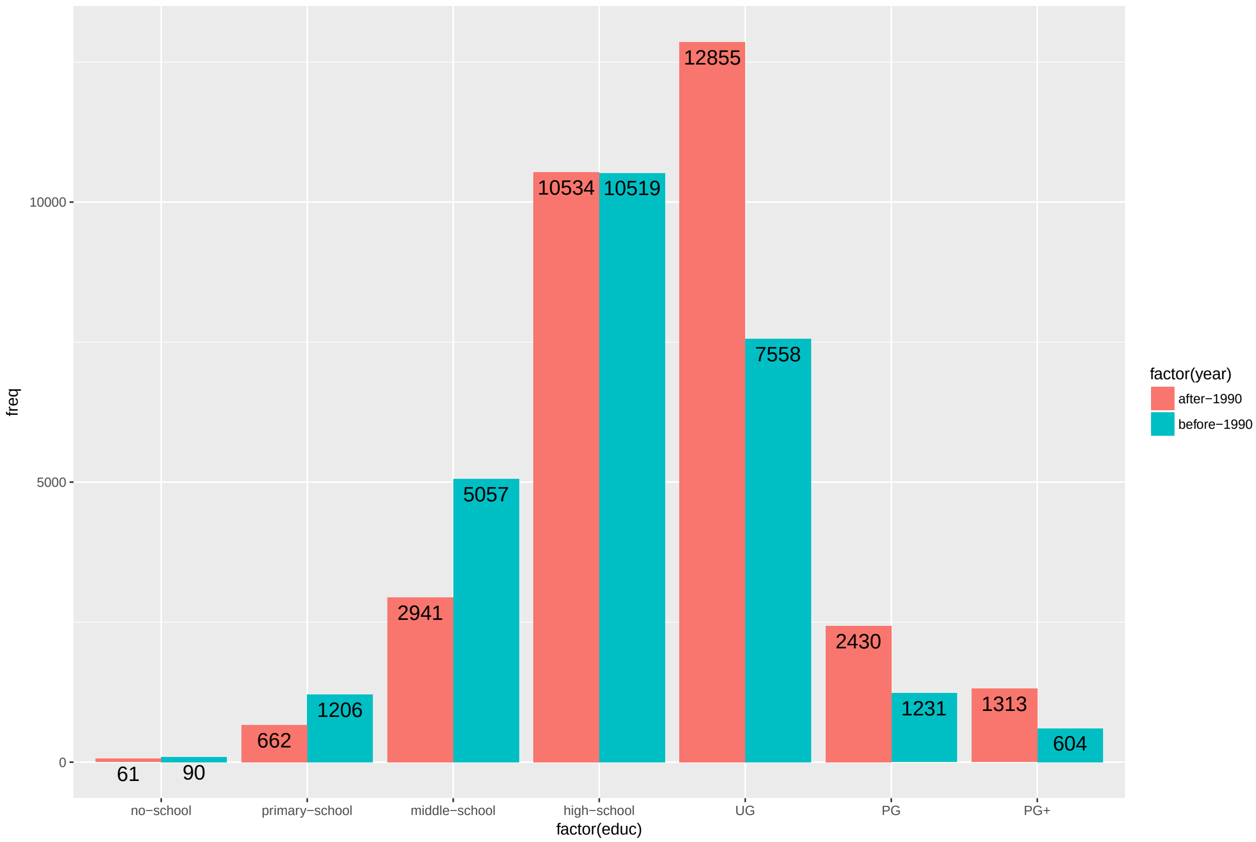
Here, there are 20 levels for education categorical variable. For making it more readable, I am categorzing these levels into categorizing education levels into no school, pre school, primary school, middle school , high school, UG, PG, PG+ higher education

```r
gss$educ <- factor(gss$educ)
levels(gss$educ) <- c("no-school","primary-school","primary-school","primary-school","primary-school","p
head(gss)
```

```
##            educ        year
## 1            UG before-1990
## 2 middle-school before-1990
## 3   high-school before-1990
## 4            PG before-1990
## 5   high-school before-1990
## 6            UG before-1990
```

```r
bar_plot <- gss %>%
        group_by(year, educ) %>%
        summarise(freq = n())

ggplot(bar_plot, aes(factor(educ), freq, fill = factor(year))) +
        geom_bar(stat = "identity", position = "dodge") +
        geom_text(aes(label = round(freq, 1)), position = position_dodge(0.9),
                  vjust = 1.5, color = "black", size = 5)
```



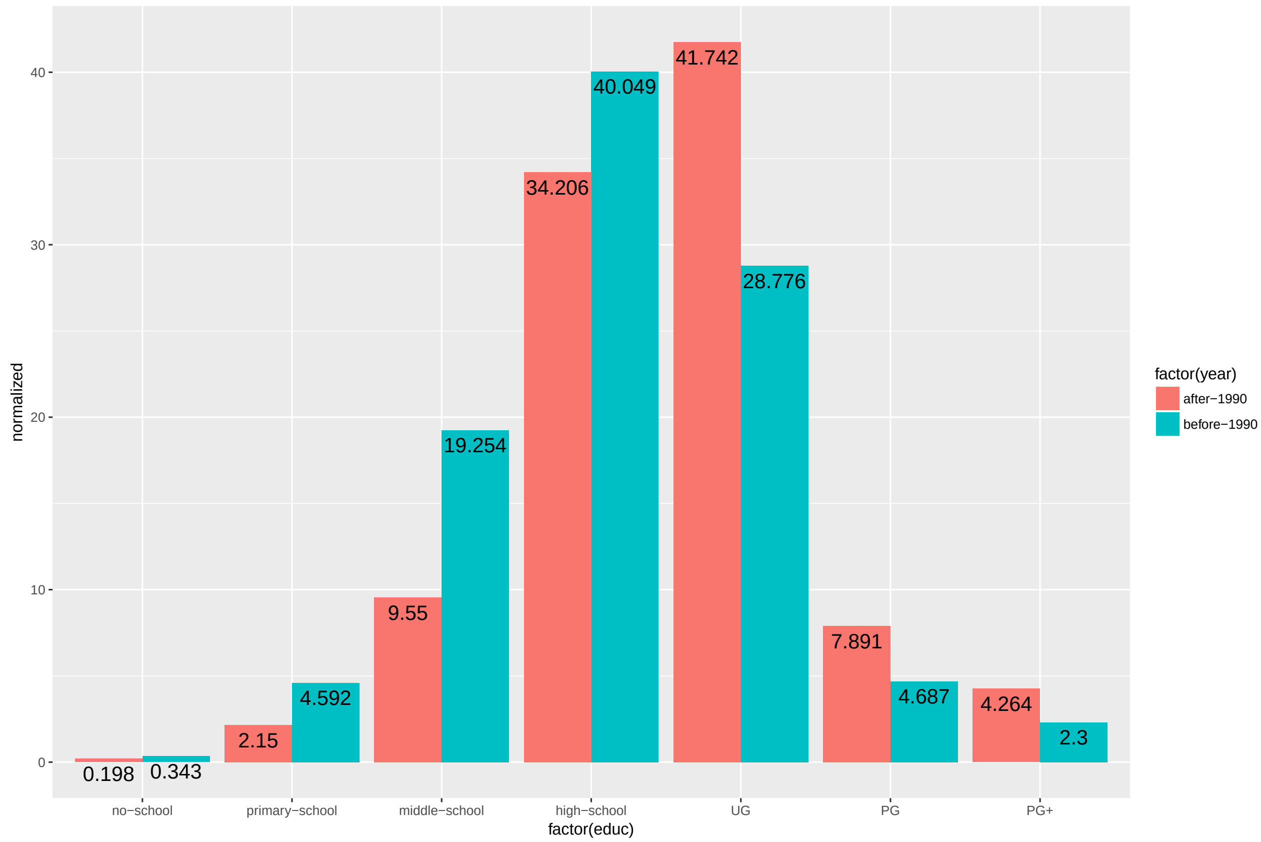

```r
head(bar_plot)
```

```
## # A tibble: 6 x 3
## # Groups:   year [1]
##   year       educ            freq
##   <chr>      <fct>          <int>
## 1 after-1990 no-school         61
## 2 after-1990 primary-school   662
## 3 after-1990 middle-school   2941
## 4 after-1990 high-school    10534
## 5 after-1990 UG             12855
## 6 after-1990 PG              2430
```

The graph above does suggest that, except for high-school level educated individuals, there is a significant difference in education levels after 1990 and before 1990. After 1990, education levels for number of individuals is almost 40-50% lower than the size of eudcation levels before 1990. The scenario is reversed when we compare education levels below high school. This might not give

```r
normalized_bar_plot <- bar_plot %>% mutate(normalized = 100 * freq/sum(freq))


ggplot(normalized_bar_plot, aes(factor(educ), normalized, fill = factor(year))) +
        geom_bar(stat = "identity", position = "dodge") +
        geom_text(aes(label = round(normalized, digits = 3)), position = position_dodge(0.9),
                      vjust = 1.5, color = "black", size = 5)
```




```
normalized_bar_plot
```

```
## # A tibble: 14 x 4
## # Groups:   year [2]
##    year        educ            freq normalized
##    <chr>       <fct>          <int>      <dbl>
##  1 after-1990  no-school         61      0.198
##  2 after-1990  primary-school   662      2.15
##  3 after-1990  middle-school   2941      9.55
##  4 after-1990  high-school    10534     34.2
##  5 after-1990  UG             12855     41.7
##  6 after-1990  PG              2430      7.89
##  7 after-1990  PG+             1313      4.26
##  8 before-1990 no-school         90      0.343
##  9 before-1990 primary-school  1206      4.59
## 10 before-1990 middle-school   5057     19.3
## 11 before-1990 high-school    10519     40.0
## 12 before-1990 UG              7558     28.8
## 13 before-1990 PG              1231      4.69
## 14 before-1990 PG+              604      2.30
```

If we look at the same graph changing the input from total values to average values, we see that the pattern do not change much except for the proportions in high school (6% difference)

Part 4: Inference

Framing Hypothesis

H0 (nothing changed) : Level of education did not change because of internet origin in 1990. The observed counts of level of education in years before 1990 and years after 1990 follow the same distribution.

HA (something changed) : Level of education did change because of internet origin in 1990. The observed counts of level of education in years before 1990 and years after 1990 do not follow the same distribution.

What type of hypothesis testing needs to be done?

As we changed year into categorical varibale with two categories (before-1990 & after-1990) and education into six categories, we can check if the distributions are similar using chi-square independence test. This test is perfect for our analysis because it is mainly used when working with categorical variables with at least one of them should have more than three levels.

Here, year is a categorical variables and education is a categorical varaible with more than two levels.Thus, we can use Chi-Square Independece Test

Checking Conditions

Evaluating conditions for the hypothesis test:

1. Independence : Sampled observations must be independent

- this is a random sample
- Is sample size less than 10% of American population?

```
# Total number of observations in gss data

str(gss)
```

```
## 'data.frame':    57061 obs. of  2 variables:
##  $ educ: Factor w/ 7 levels "no-school","primary-school",..: 5 3 4 6 4 5 5 5 4 4 ...
##  $ year: chr  "before-1990" "before-1990" "before-1990" "before-1990" ...
```

There are 57061 observations in the dataset. This is definelty lower than the total number of population of US

- checking if each case contributes to only one cell

```
# Total number of categories present in the education level column

table(gss)
```

```
##                 year
## educ             after-1990 before-1990
##   no-school              61          90
##   primary-school        662        1206
##   middle-school        2941        5057
##   high-school         10534       10519
##   UG                  12855        7558
##   PG                   2430        1231
##   PG+                  1313         604
```

Each observation will not fall into more than one category of education

2. Sample size : Each level has atleast 5

```r
head(gss)
```

```
##             educ        year
## 1             UG before-1990
## 2 middle-school before-1990
## 3   high-school before-1990
## 4             PG before-1990
## 5   high-school before-1990
## 6             UG before-1990
```

The minimum value among all the levels is 151, it is more than the minimum. So, this condition is satisfied

All the conditions are met. So, Chi Squared Independence Test can be used. Let's consider 0.05 to be significance level for this test.

Performing inference

```r
chisq.test(table(gss))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(gss)
## X-squared = 2408.7, df = 6, p-value < 2.2e-16
```

Interpreting results and Conclusion

Here, the p-value is very low thant. As the p-value is less than the significance level of 0.05, we reject the null hypothesis. Conclusion can be made that the observed proportions(after 1990) are significantly different from the expected proportions(before 1990) and they do not follow same distribution.

Reasoning for why CI is not also included?

CI is an estimated interval for a population parameter.At a defined probability, what is the range of values that we can come up with for population parameter to fall within it. This is used for estimating numerical data. Here, all we have is categorical variables. So, it cannot be used here.