Statistical inference with the GSS data

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(statsr)
library(knitr)
library(corrplot)
```

Load data

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

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

Part 1: Data

Dataset definition and objectives:

According to the GSS FAQ the General Social Survey, is is a sociological survey created and regularly collected since 1972 by the National Opinion Research Center at the University of Chicago Its basic purposes are to:

1) gather data on American society to:
    a) monitor and explain trends and constants in attitudes, behaviors, and attributes.
    b) examine the structure and functioning of society in general as well as the role of various sub-groups.
2) compare the United States (US) to other societies to:
    a) place American society in comparative perspective.
    b) develop cross-national models of human society.
3) make high-quality data easily accessible to scholars, students, and others with minimal cost and waiting.

Data collection process:

The GSS Wikipedia Page page states:

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.

Reservations:

Like any observational study, the GSS may suffer from several sources of bias: * non-response: This problem is handled using two-stage sub-sampling design, that means that if a selected person (or household) did not give a response it will be recorded and asked later for a response, this will reduce the rate of non-response significantly. * phrasing of the survey questions: the phrasing and vocabulary used in question can introduce a certain amount of bias, this problem is present in every survey, but there are a lot of studies that target this issue in GSS, and the phrasing is improved every year, so I think we can safely ignore this issue.

Conclusion:

GSS is an observational study, because it uses random sampling, BUT there is no random assignment and controlling so the results drawn from this data-set can be used to infer ASSOCIATION BUT NOT CAUSATION, and we can safely generalize the results to adults (18+) non-institutionalized living in the United States.

Part 2: Research question

There are several claims that race can affect access to education, this is a crucial question, because if true, that means that there is a certain type of discrimination in access to education in the US. Since this data-set was obtained via and observational study we can only assess association between those two variables.
Research quesion:
* *is there any association between race and level of education in the United States?*

Part 3: Exploratory data analysis

Let us start by exploring two variables we will use to do statistical inference and answer the question, let's extract the variables of interest:

```
gss %>%
  select(race, degree) %>%
  na.omit() ->
  degree_vs_race
dim(degree_vs_race)
```

```
## [1] 56051     2
```

exploring the research variables degree and race:

we will take a look at summary statistics for each variable:

Race:

the values taken by the variable:

```r
kable(unique(degree_vs_race$race),
       align = 'c')
```

x
White
Black
Other

We can see that the categorical variable race can take one one of three values: White, Black and Other which groups all the other racial groups other than White and Black.

summary statistics:

- Counts:

```r
kable(table(degree_vs_race$race),
       align = 'c')
```

Var1	Freq
White	45602
Black	7716
Other	2733

Here is the distribution of counts for the race variable, counts are useful, but proportions give us more insight into the distribution.

- Distribution of proportions:

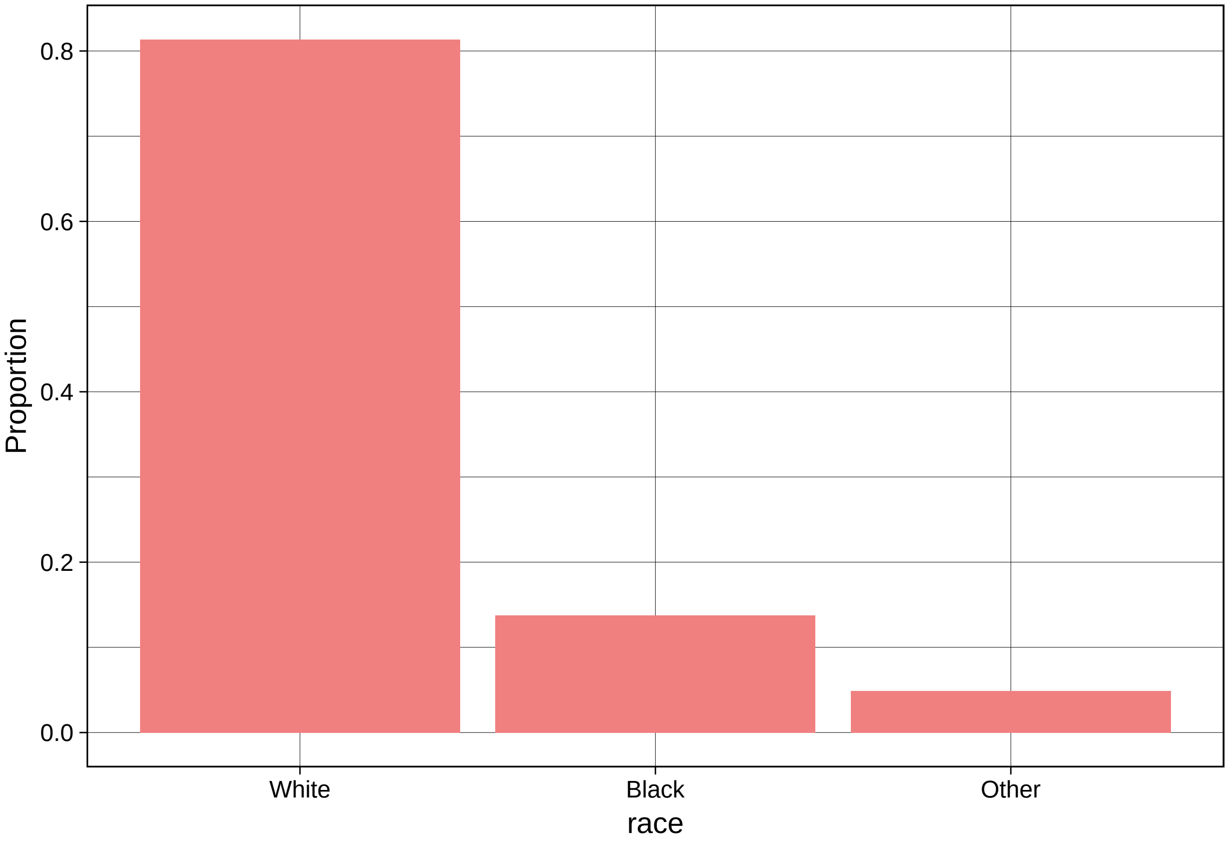
```r
kable( round( prop.table(table(degree_vs_race$race))*100, 2),
       align = 'c')
```

Var1	Freq
White	81.36
Black	13.77
Other	4.88

we can clearly see that most respondents are White, with 81.36%, then Black 13.77 and last Other with 4.88%.

- Plotting the distribution

```r
ggplot(degree_vs_race) +
  geom_bar(aes( x= race, y =  (..count..)/sum(..count..)), fill = 'lightcoral') +
  theme_linedraw() +
  labs(y = "Proportion")
```



When plotting the distribution of proportions we can observe the distribution clearly.

Degree:

the values taken by the variable:

```
kable(unique(degree_vs_race$degree),
      align = 'c')
```

x
Bachelor
Lt High School
High School
Graduate
Junior College

There are five categories, lt High School(lt for less than), High School, Junior College, Bachelor and Graduate

summary statistics:

- Counts:

```
kable(table(degree_vs_race$degree),
      align = 'c')
```

Var1	Freq
Lt High School	11822
High School	29287
Junior College	3070

Var1	Freq
Bachelor	8002
Graduate	3870

We can see that most people finish high school, followed by those who don't, to see the difference clearly let's see the proportions.

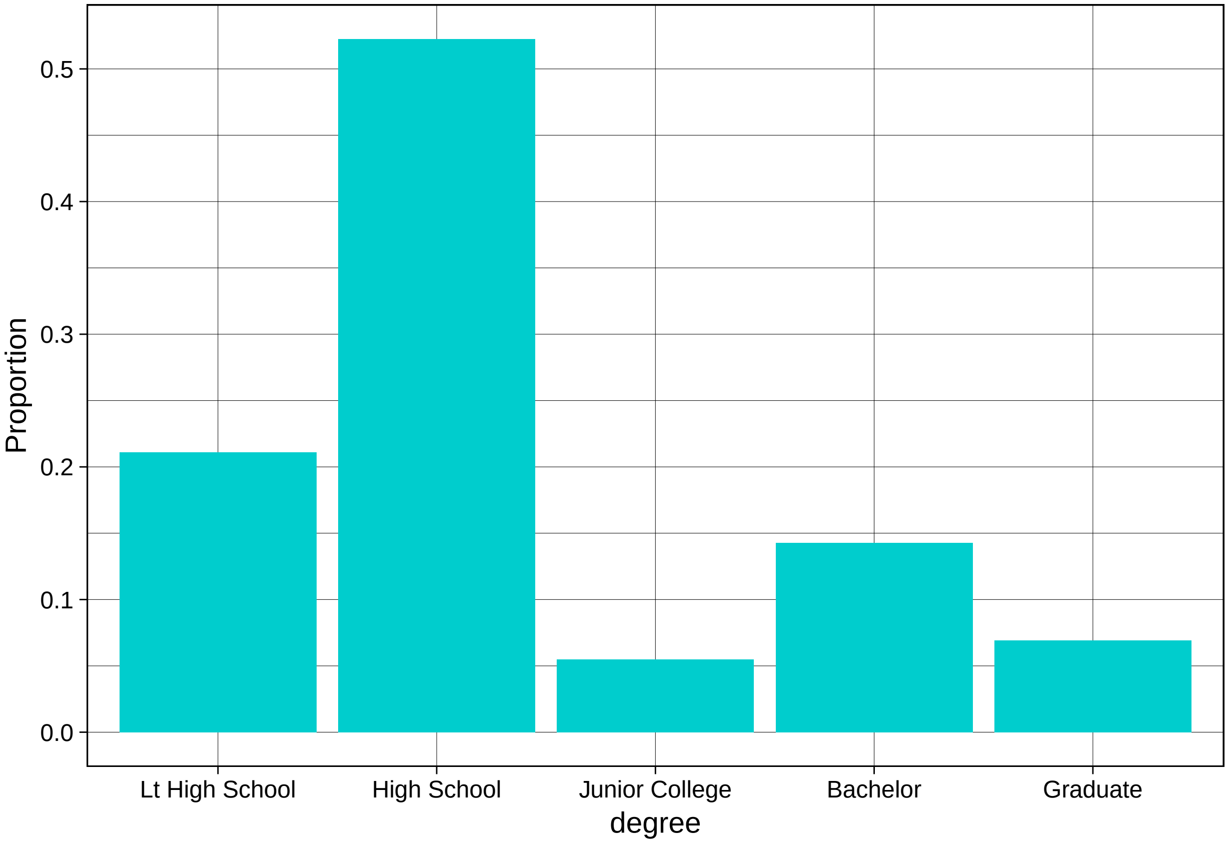
- Distribution of proportions:

```
kable( round( prop.table(table(degree_vs_race$degree))*100, 2),
       align = 'c')
```

Var1	Freq
Lt High School	21.09
High School	52.25
Junior College	5.48
Bachelor	14.28
Graduate	6.90
we can see that ha	lf the sample have finished high school, followed by 21% who don't then Bachelor's Degree holders, an

- Plotting the distribution:

```
ggplot(degree_vs_race) +
  geom_bar(aes( x= degree, y =  (..count..)/sum(..count..)), fill = 'cyan3') +
  theme_linedraw() +
  labs(y = "Proportion")
```



The plot illustrate the difference in the distribution of the degree variable.

exploring the relationship between the variables:

Now Let's dig deeper and get a closer look to the relationship between degree and race from an Exploratory Data Analysis point of view.

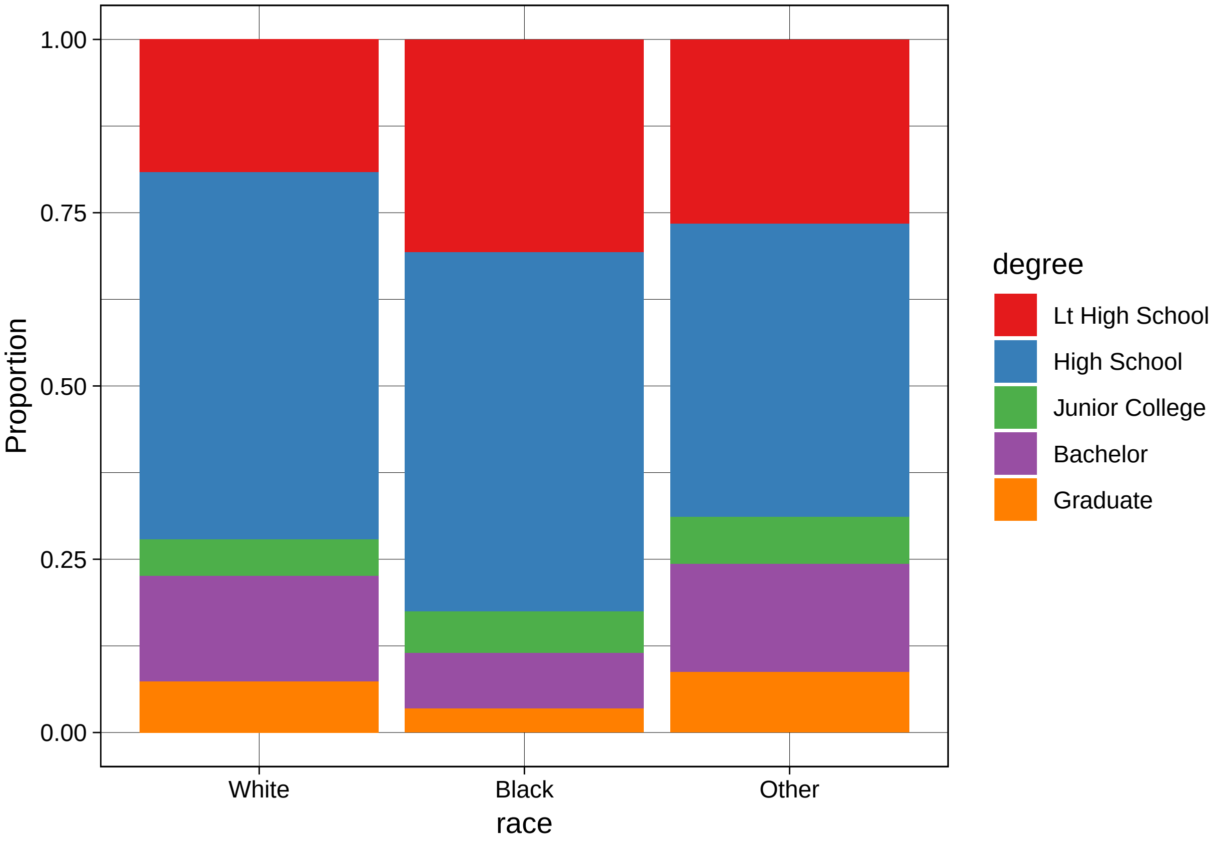
- Proportions:

```
kable( round( prop.table(table(degree_vs_race$degree, degree_vs_race$race), margin = 2)*100, 2),
       align = 'c')
```

	White	Black	Other
Lt High School	19.15	30.65	26.53
High School	52.91	51.85	42.33
Junior College	5.31	6.01	6.81
Bachelor	15.26	7.98	15.55
Graduate	7.37	3.50	8.78

The proportions table, summarizes the distribution of degrees inside each race category, we can't see the difference clearly, so let's plot this data to gain more insight.

```
ggplot(data = degree_vs_race) +
  aes(x = race, fill= degree) +
  geom_bar(position = 'fill')  +
  labs(y = "Proportion") +
  theme_linedraw() +
  scale_fill_brewer(palette = 'Set1')
```



We can see that in this sample, there is a difference between the Black category and other race categories(White and Other), Black has more less than High school observations, and less graduate and bachelor observations.

is this actually a trend and does this prove a difference in degree proportion distributions between different races, or is this just due to sampling chance? Let's go to the inference part to explore that.

Part 4: Inference

Before starting the inference we should state our hypotheses:

Hypotheses

H_0(null hypothesis): race and level of education are Independent of each other in the population of interest
H_A(alternative): there is an association between race and level of education in the population of interest

Checking conditions

- Independence:
- As we discussed in the introduction, random sampling was used in this observational study
- We are sampling WITHOUT REPLACEMENT from the US adult non-institutionalized population, and our sample size is: 56051

- There is no overlap in race categories nor degree levels, so each case contributes only to one cell in the table
- Sample size

The proportions tables from the EDA step, shows that there is no cell with less than 5 observations

Choice of the inference method

All the conditions are met, and we want to verify a Hypothesis of Independence between TWO CATEGORICAL variables, so we will use a Chi-Square test of independence.

Inference

We will perform a hypothesis test at a significant level of 0.05.

Let us start by building the table:

```
degree_vs_race_table <- table(degree_vs_race$race, degree_vs_race $degree)

kable(degree_vs_race_table)
```

	Lt High School	High School	Junior College	Bachelor	Graduate
White	8732	24129	2420	6961	3360
Black	2365	4001	464	616	270
Other	725	1157	186	425	240

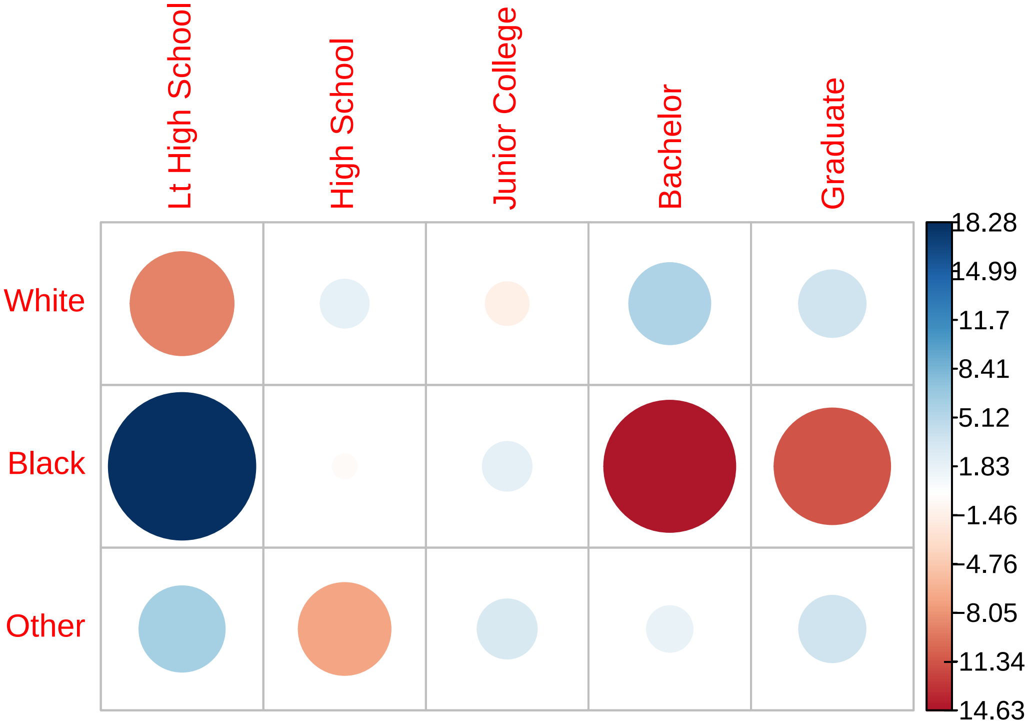
Let us perform the Chi-Square test:

```
Chi_square <- chisq.test(degree_vs_race_table)
Chi_square
```

```
##
##  Pearson's Chi-squared test
##
## data:  degree_vs_race_table
## X-squared = 931.06, df = 8, p-value < 2.2e-16
```

let’s see a way to visualize the result of this test:

```
corrplot(Chi_square$residuals, is.cor = FALSE)
```



For a given cell, the size of the circle is proportional to the amount of the cell contribution to the relationship.

For example we can see that the race black has a high positive correlation with the high school level of education, and a large negative correlation with both graduate and bachelor categories, which suggests that black people are more likely to drop from school at the high school level and are less likely to graduate or stay at college to reach a bachelor's degree level

Interpret results

We got a very small p-value from the Chi-squared test of independence, we got p-value < 2.2e-16, so at a significance level of 5% we will REJECT the null hypothesis in favor of the alternative hypothesis, and we will conclude that:

Education level and race are associated in our population.