Statistical inference with the GSS data

Setup

Load packages

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

Load data

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

Part 1: Data

The data set at hand is a cleaned subset of the original General Social Survey cumulative file that contains records from 1972 through 2012. The GSS aims to monitor and explain changes in the American society. In doing so, the GSS collects data from US households.
The data is observational, so we will not be able to infer causal relationships. While there are ways to infer causality from observational data, such as the disjunctive cause criterion, these were not covered in this course and will consequently not be used.

Since the GSS puts tremendous effort in the generalizability, utilizing random samples of less than 10% of the total population from multiple strata that were formed using national US census data, our findings can be generalized to the US population. However, slight bias is induced by the fact that responding to the survey is strictly voluntarily. Since the survey is conducted face-to-face at the University of Chicago, people who are not willing or not able to visit the university for the survey are likely underrepresented.

Part 2: Research question

We will examine the following research question:

Was there a difference between average years spent on education between firearm owners and citizens without firearms in the United States in the years 2010 through 2012?

Living in Europe, I have hardly any idea what are main causes for firearm ownership. In movies, it often appears that people with high education are less likely to have any firearms, so I would like to investigate this hypothesis using real world data.

In order to address the question, we will need to add another column to the data set that indicates firearm ownership.

Part 3: Exploratory data analysis

We start by selecting a relevant subset of the data. To account for newer trends, we will only consider data from 2010 through 2012. Also, to account for refused answers, we will only focus on those respondents who have either confirmed to have at least one firearm of any of the four types in the survey or have denied to have any of them.

```
df <- gss %>%
  filter(year >= 2010) %>%
  select(educ, owngun, pistol, shotgun, rifle)
summary(df)
```

```
##       educ           owngun         pistol        shotgun
##  Min.   : 0.00   Yes    : 834   Yes    : 528   Yes    : 517
##  1st Qu.:12.00   No     :1680   No     :1965   No     :1976
##  Median :13.00   Refused:  66   Refused:  78   Refused:  78
##  Mean   :13.49   NA's   :1438   NA's   :1447   NA's   :1447
##  3rd Qu.:16.00
##  Max.   :20.00
##  NA's   :7
##      rifle
##  Yes    : 500
##  No     :1993
##  Refused:  78
##  NA's   :1447
##
##
##
```

There are a few refused answers and NA values that we will have to get rid of. We will consider someone a firearm owner if he stated to own at least one of the four firearm types above. If he answered "No" to all of them, we will not consider him a firearm owner. In all other cases, we will remove the record from the data it does not allow us to classify people.

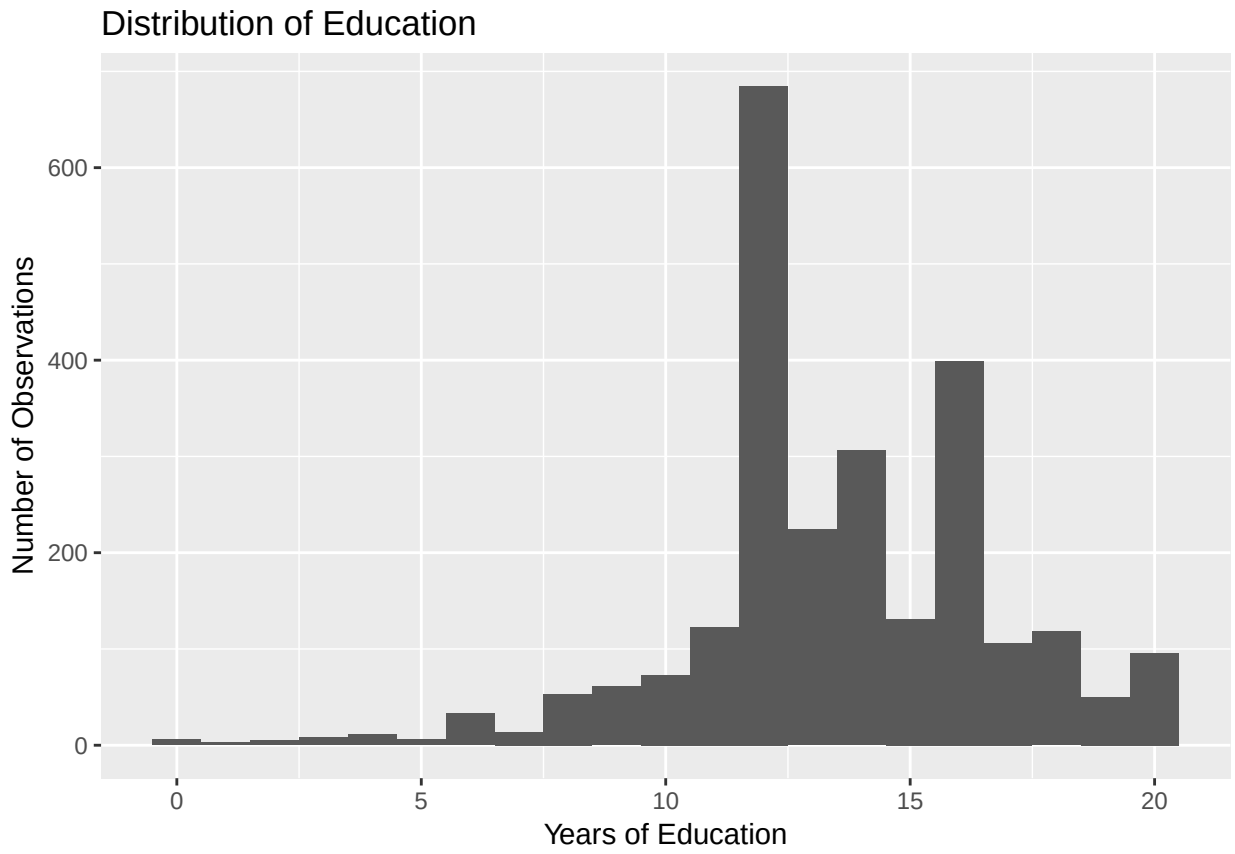
```
df <- df %>%
  filter((owngun == "Yes" | pistol == "Yes" | shotgun == "Yes" | rifle == "Yes") |
           (owngun == "No" & pistol == "No" & shotgun == "No" & rifle == "No")) %>%
  mutate(firearm_owner = owngun == "Yes" | pistol == "Yes" | shotgun == "Yes" |
           rifle == "Yes") %>%
  select(educ, firearm_owner) %>%
  filter(complete.cases(.))
summary(df)
```

```
##       educ       firearm_owner
##  Min.   : 0.00   Mode :logical
##  1st Qu.:12.00   FALSE:1677
##  Median :13.00   TRUE :834
##  Mean   :13.55
##  3rd Qu.:16.00
##  Max.   :20.00
```

Our new data frame contains only two variables, namely the years of education and whether the person owns a firearm, and has no missing values. To get started, we will consider the two variables of interest individually.

We will first look at the distribution education in the data.

```
ggplot(data = df, aes(x = educ)) +
  geom_histogram(binwidth = 1) +
  xlab("Years of Education") +
  ylab("Number of Observations") +
  labs(title ="Distribution of Education")
```



As we can see, there are peaks at 12 and 16 years of education, corresponding to high school and college education. The data is left skewed due to some respondents having 0 to 5 years of education.

Next, we will have a look at the distribution of firearm ownership.

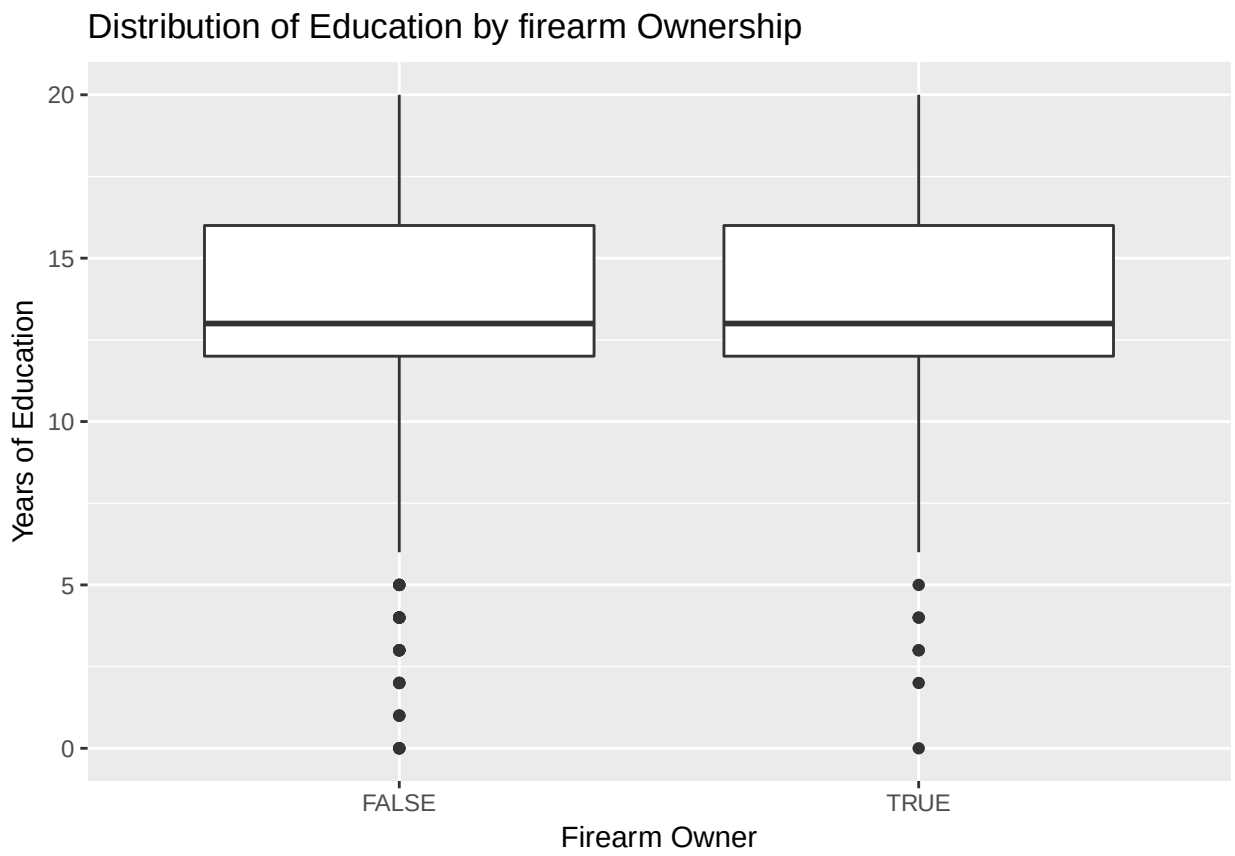
```
n <- dim(df)[1]
df %>%
  group_by(firearm_owner) %>%
  summarise(pct = n()/n)
```

```
## # A tibble: 2 x 2
##   firearm_owner       pct
##           <lgl>     <dbl>
## 1         FALSE 0.6678614
## 2          TRUE 0.3321386
```

About one in three of respondents in the sample own a firearm. Now, we will investigate the relationship of the two variables. A suitable kind of plot for this is the box plot.

```
ggplot(data = df, aes(x = firearm_owner, y = educ)) +
  geom_boxplot() +
  labs(title = "Distribution of Education by firearm Ownership") +
```

```
  xlab("Firearm Owner") +
  ylab("Years of Education")
```



It seems like the median and the lower and upper quartiles are identical. However, there might be a difference in the mean. Let us continue with calculating the group means:

```
educ_df <- df %>%
  group_by(firearm_owner) %>%
  summarise(mean_educ = mean(educ))
x_bar_1 = educ_df$mean_educ[2]
x_bar_2 = educ_df$mean_educ[1]
educ_df
```

```
## # A tibble: 2 x 2
##   firearm_owner mean_educ
##           <lgl>     <dbl>
## 1         FALSE  13.48778
## 2          TRUE  13.67506
```

The means are nearly identical at around 13.5 to 13.7 years of education. Next, we will examine whether this difference is significant or might be due to chance.

Part 4: Inference

This is a problem of comparing the mean of a numeric variable for two groups.

The hypothesis we are testing is whether the average population education differs between firearm owners and people that do not own a firearm. We do so by conducting a two-sided t-test at the 95% significance level. Later on, we will also provide a 95% confidence interval for the difference in group means.

Our hypotheses are as follows:

$H_0 : \mu = 0$, i.e. the true difference in average years of education between the two groups is 0.

$H_A : \mu \neq 0$, i.e. there is a non-zero difference in the true average years of education between the two groups.

Here, μ denotes the true difference in means between the two groups.

Before we can perform the hypothesis test, we have to check whether the conditions are met. In the introduction section, we already discussed independence within groups. Since the data is non-paired, it is also reasonable to assume that independence between groups is given. There might be a few individuals that have responded to the survey in multiple years between 2010 and 2012, but due to random sampling this proportion can be expected to be low. There is a bit of skew in the distribution of education, but with more than 2500 samples in our data, the Central Limit Theorem applies. This means we can assume the mean education per group to be nearly normally distributed. The difference in means is then a difference of two independent normally distributed random variables and as such normally distributed as well. However, since we do not know the population standard deviations of each group, we will have to estimate them using the sample standard deviations. Doing so, the random variable $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ follows a t-distribution under the null hypothesis, where $\bar{x}_i, s_i, n_i$ are the sample mean, the sample standard deviation and the sample size for group i respectively. The associated degrees of freedom are at least $\min(n_1 - 1, n_2 - 1)$ but the precise formula was not given in the lecture.

Normally, we would use the t.test function of R to take care of the t-test (and to automatically select the accurate degrees of freedom). However, for the sake of illustration, we will do so by hand and using the conservative formula for the degrees of freedom.

```
n_1 <- sum(df$firearm_owner == TRUE)
n_2 <- n - n_1
s_1 <- sd(df[df$firearm_owner == TRUE, ]$educ)
s_2 <- sd(df[df$firearm_owner == FALSE, ]$educ)
s_pooled <- sqrt(s_1^2/n_1 + s_2^2/n_2)
dof <- min(n_1 - 1, n_2 - 1)
t_val <- (x_bar_1 - x_bar_2)/s_pooled
p_val <- 2 * pt(t_val, df = dof, lower.tail = FALSE)
p_val
```

```
## [1] 0.14254
```

Since the p-value is about 14.25%, we fail to reject the null hypothesis. The p-value states that if the null hypothesis is true, we would observe a difference in means that is as least as extreme as our observed difference with a probability of 14.25%. There is no significant evidence for a difference in average education between firearm owners and people without firearms.

Next, we will calculate a confidence interval for the true difference in means. We would expect the confidence interval to contain 0 since we failed to reject the null hypothesis.

To do so, we take our sample difference in means as a point estimate and add and subtract a margin of error, i.e. the critical t-score multiplied by the standard error which is simply the pooled standard deviation we estimated above. The critical t-score is chosen such that 95% of the values of the t-distribution are within the interval $[-critical\_t\_score, critical\_t\_score]$.

```
critical_t_score <- qt(0.975, df = dof)
ci <- x_bar_1 - x_bar_2 +
  c(-1, 1) * critical_t_score * s_pooled
```

```
ci
```

```
## [1] -0.06316413  0.43773245
```

We are 95% confident that the true difference in means is between -0.063 and 0.438 years of education. If we were to collect survey data many times and built confidence intervals each time, we would expect 95% of these intervals to include the true difference average years of education between the two groups. Also, this confidence interval agrees with the result of the hypothesis test: Indeed, 0 is contained within that interval. Please note that the sign has to be interpreted in such a way that a positive sign is associated with more education for firearm owners while a negative sign is associated with less education for firearm owners (compared to non-owners).

Our findings do not show a significant difference in education between firearm owners and people without firearms. However, this does not necessarily mean there is no difference. A possible way to address the same question would be using more data in order to reduce the standard error. If the null hypothesis were in fact false, this would help in reducing the probability of doing a type II error. However, before blindly collecting more data, one should calculate how many samples would be needed in order to show the specific effect size at a specified significance level.

Just for the record, we will also consider the built-in t.test function:

```
t.test(x = df[df$firearm_owner == TRUE, ]$educ,
       y = df[df$firearm_owner == FALSE, ]$educ,
       alternative = "two.sided", mu = 0,
       paired = FALSE, var.equal = FALSE,
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  df[df$firearm_owner == TRUE, ]$educ and df[df$firearm_owner == FALSE, ]$educ
## t = 1.4678, df = 1903.6, p-value = 0.1423
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.06295934  0.43752766
## sample estimates:
## mean of x mean of y
##  13.67506  13.48778
```

As we can see, the degrees of freedom are actually 1903.6. However, both a t-distribution with 833 and with 1903.6 degrees of freedom resemble a standard normal distribution extraordinarily well, the difference is conclusions is minor. The p-value is only slightly different and so is the confidence interval. In a real decision setting, we should trust the more precise calculation of the built-in function with the correct degrees of freedom.