Statistical inference with the GSS data

Setup

Load packages

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

Load data

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

Part 1: Data

General information about data

The General Social Survey (GSS) is a nationally representative survey of adults in the United States conducted since 1972. The GSS collects data on contemporary American society in order to monitor and explain trends in opinions, attitudes and behaviors. The GSS has adopted questions from earlier surveys which allows researchers to conduct comparisons for up to 80 years.

The GSS contains a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. Among the topics covered are civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events.

Altogether the GSS is the single best source for sociological and attitudinal trend data covering the United States. It allows researchers to examine the structure and functioning of society in general as well as the role played by relevant subgroups and to compare the United States to other nations.

The dataset used in this analysis is slightly different from the original dataset, as this work is the final draft of the course from Duke University('Inferential Statstics'), whose professors modified the dataset to make it easier to work with:

"Unlike the full General Social Survey Cumulative File, we have removed missing values from the responses and created factor variables when appropriate to facilitate analysis using R. Our hope is that this will allow students to focus on statistical concepts without having to (initially) be concerned about some of the data management and interpretation issues associated with missing data and factor variables in R".

Causality and generabizability

These studies were carried out on the basis of a large set of data that were obtained through stratified random sampling or full probability random sampling - one adult is selected randomly from each sampled household. Also, the sampling frames are updated over time to highlight a population that can be representative of US population. Based on this, the results can be generalized to the entire population of the United States of America. However, the data were not obtained experimentally, because this is an observational study, and therefore we cannot establish causal relationships, but can only investigate the variables for the presence of correlation.

Collecting data in the way presented in the study could lead to a number of errors. They could be related to the fact that it was not taken into account that some people could give incorrect information or not give information at all. Also the information may have been recorded incorrectly. In addition, some of the records in the survey were incomplete, which could also affect the research results.

Part 2: Research question

The main issue of this study is the difference in attitudes towards abortion between men and women of different ages and classes. There are different opinions in different countries and regions about whether abortion should be banned or legalized. Religious beliefs, particularities of the respondents' mentality, cultural values and so on may be behind these differences. In connection with the increasing urgency of the issue of gender inequality, in this work we try to understand whether there is a relationship between the sex of the respondent and his opinion about abortion.

Part 3: Exploratory data analysis

To avoid missing values in our data, let's put the data filtered from missing records into a new dataset called **gss_females**.

```
gss_females <- gss %>%
  filter(!is.na(gss$sex), !is.na(gss$abany), !is.na(gss$class), !is.na(gss$age))
```

Let's do this also for a new dataset called **gss_males**.

```
gss_males <- gss %>%
  filter(!is.na(gss$sex), !is.na(gss$abany), !is.na(gss$class), !is.na(gss$age))
```

Now let's define the data for men and women separately in these datasets. To do this, let's filter the data by the **sex** variable. We will also select for the variable **abany** only the values "Yes" and "No", for the variable **class** we will select the values of four classes: "Lower Class", "Working Class", "Middle Class", "Upper Class".

```
gss_females <- gss_females %>%
  filter(gss_females$sex == 'Female', gss_females$abany %in% c('Yes', 'No'), gss_females$class %in% c('Lower (


gss_males <- gss_males %>%
  filter(gss_males$sex == 'Male', gss_males$abany %in% c('Yes', 'No'), gss_males$class %in% c('Lower Class',
```

We will also define a separate dataset for both genders, using elimination of missing values and the same class filtering.

```
gss_both <- gss %>%
  filter(!is.na(gss$sex), !is.na(gss$abany), !is.na(gss$class), !is.na(gss$age))

gss_both <- gss_both %>%
  filter(gss_both$abany %in% c('Yes', 'No'), gss_both$class %in% c('Lower Class', 'Working Class', 'Middle Cla
```

Let's take a look at an overview of these variables that we will be looking at.

```
summary(gss_both$sex)
```

```
##   Male Female
##  13602  17193
```

For the variable **sex** we see the number of surveyed men (13602) and women (17193).

Next, consider the variable **age** to find the minimum, maximum, and average ages among the respondents.

```
summary(gss_both$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   31.00   43.00   45.51   58.00   89.00
```

We can see that the minimum age is 18, which was stated in the dataset report, the maximum age is 89 and the average is around 45.

Let's also look at the number of representatives of different classes among the respondents.

```
summary(gss_both$class)
```

```
##   Lower Class Working Class  Middle Class   Upper Class      No Class
##          1647         14138         14012           998             0
```

Here we see that we are dealing with four classes, and among the respondents, the working and middle classes predominate.

Last but not least, let's look at the entries for the abortion-related variable. "Yes" is responsible for the fact that a pregnant woman should be able to have a legal abortion, and "No" for the opposite.

```r
summary(gss_both$abany)
```

```
##   Yes    No
## 12525 18270
```

It can be seen that the majority of the respondents (18270) are opposed to the legalization of abortion.

Let's do the same for the datasets for men and for women separately.

For the women we get:

```r
summary(gss_females$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   31.00   43.00   46.19   60.00   89.00
```

```r
summary(gss_females$class)
```

```
##   Lower Class Working Class  Middle Class   Upper Class      No Class
##          1036          7749          7878           530             0
```

```r
summary(gss_females$abany)
```

```
##   Yes    No
##  6841 10352
```

For the men we get:

```r
summary(gss_males$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   31.00   42.00   44.65   57.00   89.00
```

```r
summary(gss_males$class)
```

```
##   Lower Class Working Class  Middle Class   Upper Class      No Class
##           611          6389          6134           468             0
```

```r
summary(gss_males$abany)
```

```
##  Yes   No
## 5684 7918
```

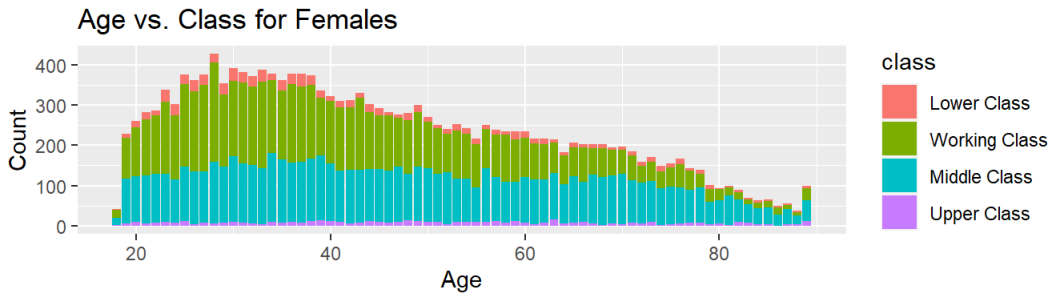
Now let's start looking at the data on opinions about abortion, grouping them by age and by grade. To begin with, let's see with the help of graphing how many people of different ages belong to one or another class among men and women.

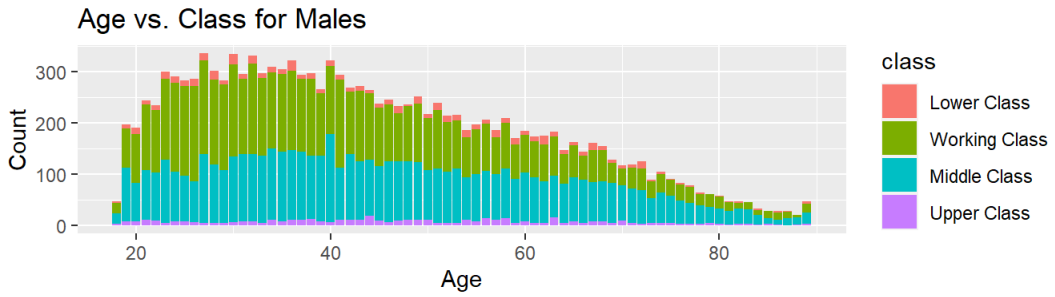
```r
plot_1_f <- ggplot(gss_females, aes(x = age, fill = class)) +
  geom_bar(stat = 'count') +
  labs(x = 'Age', y = 'Count') +
  ggtitle("Age vs. Class for Females"
)

plot_1_m <- ggplot(gss_males, aes(x = age, fill = class)) +
  geom_bar(stat = 'count') +
  labs(x = 'Age', y = 'Count') +
  ggtitle("Age vs. Class for Males")
```

Let's display the graphs together.

```r
grid.arrange(plot_1_f, plot_1_m)
```





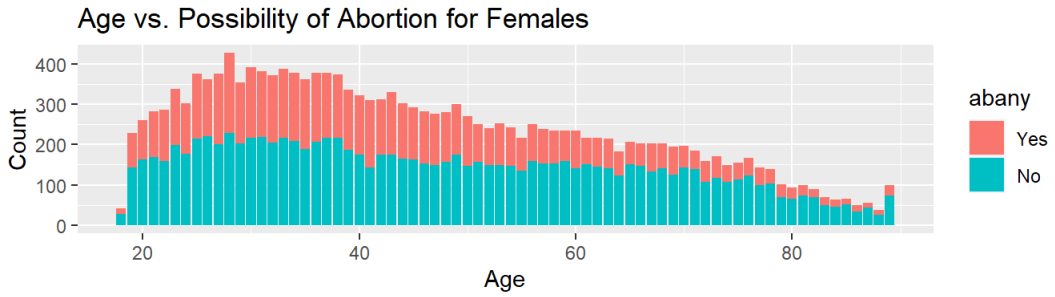
Now let's calculate the number of people by age who are for and against abortion.

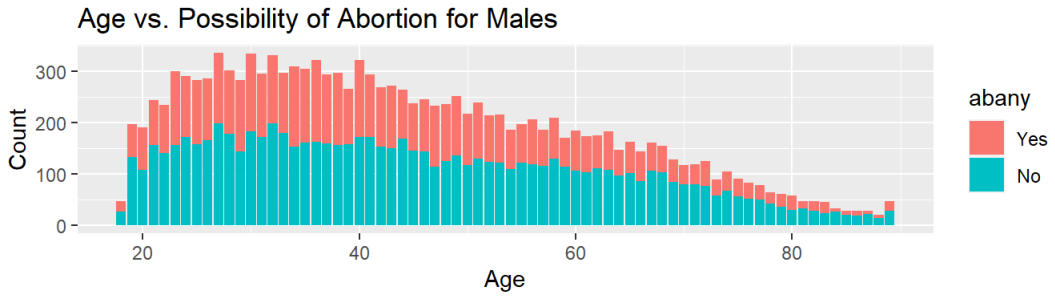
```
plot_2_f <- ggplot(gss_females, aes(x = age, fill = abany)) +
  geom_bar(stat = 'count') +
  labs(x = 'Age', y = 'Count') +
  ggtitle("Age vs. Possibility of Abortion for Females")

plot_2_m <- ggplot(gss_males, aes(x = age, fill = abany)) +
  geom_bar(stat = 'count') +
  labs(x = 'Age', y = 'Count') +
  ggtitle("Age vs. Possibility of Abortion for Males")
```

Let's display the graphs together.

```
grid.arrange(plot_2_f, plot_2_m)
```





Let's think about how the age of the respondents can be related to their attitude to abortion. Let's start with women. We will call women under 40 the first category, and women over 40 - the second.

First, it can be noted that the majority of women in the first category are in favor of legalizing abortion, while women in the second category are mostly against abortion. What could be the reason for this?

It is obvious that women under 40 have more sexual intercourse than after crossing this age threshold. Consequently, for them the problem of unwanted pregnancy does not lose its relevance, while for the second category it does. This expresses personal interest and indirect disinterest. Also, representatives of the second category of women may be against abortions due to the fact that they have outdated information about carrying out this, as it may seem to them, painful and leading to various disorders in the body procedure, which over the years and the development of medicine has become much safer for patients. In addition, women of the

second category have a different vision of the world, corresponding to the time in which they lived, and are more focused on the classical preservation and procreation, young girls nowadays are more interested in pursuing a career, mobility, free relationships, and so on. On the other hand, the representatives of the first category of women do not have as much experience as the second, and they may not think about some important things, for example, how they do not think about the fact that after a certain age they will not be able to fully function and have children. They may also know that in youth, under pressure from a partner, abortion is coerced, the legalization of which will simplify this phenomenon and reduce the partner's responsibility for having a child.

The situation is similar for men. We will also divide them into two groups. The first group is similarly interested mainly in building a career and in their freedom and mobility, as well as, for example, free relationships. The second category has the values of its time, experience, in some places relevant within its time, and is also tuned in to procreation.

Now let's analyze the situation by class. In this table, you can see the proportions of the various classes surveyed who are in favor of and against abortion.

```
prop.table(table(gss_females$abany, gss_females$class))
```

```
##
##       Lower Class Working Class Middle Class Upper Class   No Class
##   Yes  0.01884488    0.16646310   0.19839470  0.01419182 0.00000000
##   No   0.04141220    0.28424359   0.25981504  0.01663468 0.00000000
```

```
prop.table(table(gss_males$abany, gss_males$class))
```

```
##
##       Lower Class Working Class Middle Class Upper Class   No Class
##   Yes  0.01661520    0.17298927   0.20930745  0.01896780 0.00000000
##   No   0.02830466    0.29672107   0.24165564  0.01543891 0.00000000
```

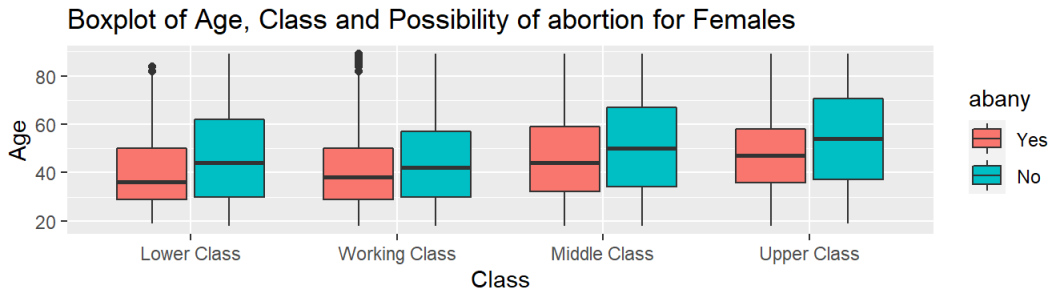
One can notice a tendency, as with an increase in the rank of the class, the attitude of men and women to abortion changes. The lower classes are mostly against abortion, both among women (4.1% of the total number of women - against, 1.9% - for) and among men (2.8% of the total number of men - against, 1.7% - for). Among the working classes, the difference is becoming less noticeable: among women, 28.4% are against, 16.6% are in favor, and among men, 29.6% are against, 17.2% are for. The difference is even less noticeable among the middle class: among women 26.0% are against, 19.8% are for, and among men 24.2% are against, 20.9% are for. The difference in the upper class for women is practically obliterated (16.6% are against and 14.2% are for), and the state of affairs with regard to men's opinion is changing (15.4% are against, 19.0% are for). This trend can be explained by the life priorities of various classes, such as career, family building, material condition, religious beliefs, and so on.

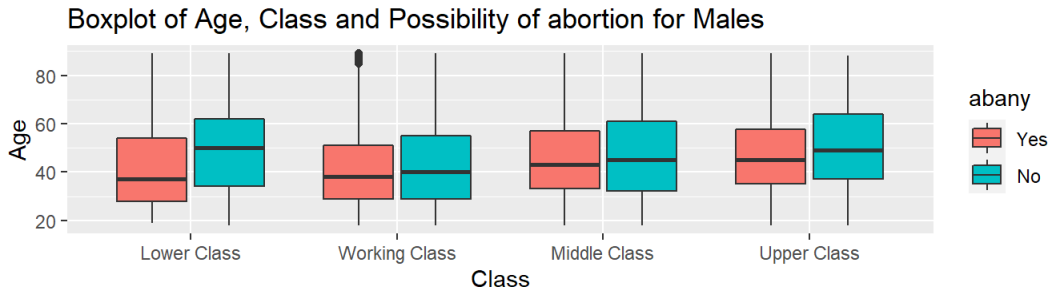
Let's combine the obtained results and construct graphs reflecting the age distribution of representatives of different classes and adherents of positive and negative opinions about abortion.

```
plot_3_f <- ggplot(gss_females, aes(x = class, y = age, fill = abany)) +
  geom_boxplot() +
  ggtitle("Boxplot of Age, Class and Possibility of abortion for Females") +
  labs(x = 'Class', y = 'Age')

plot_3_m <- ggplot(gss_males, aes(x = class, y = age, fill = abany)) +
  geom_boxplot() +
  ggtitle("Boxplot of Age, Class and Possibility of abortion for Males") +
  labs(x = 'Class', y = 'Age')

grid.arrange(plot_3_f, plot_3_m)
```





Part 4: Inference

Now let's create a sign to describe the number of men and women who are in favor of and against abortion.

```
test_data <- table(gss_both$sex, gss_both$abany)

test_data
```

```
##
##             Yes    No
##   Male     5684  7918
##   Female   6841 10352
```

If we want to estimate this in shares of the total number of respondents, then the following must be done:

```
prop.table(test_data)
```

```
##
##                Yes        No
##   Male   0.1845754 0.2571197
##   Female 0.2221465 0.3361585
```

It would also be interesting to know the share of supporters and opponents of abortion, not relative to the total number of respondents, but relative to the total number of respondents in the gender group.

```
males_yes = 5684 / (5684 + 7918)
females_yes = 6842 /(6842 + 10352)

males_no = 1 - males_yes
females_no = 1 - females_yes

percents_table <- data.frame(females_yes, males_yes, females_no, males_no)

percents_table
```

```
##   females_yes males_yes females_no  males_no
## 1   0.3979295 0.4178797  0.6020705 0.5821203
```

Now, finally, we will test our hypothesis about the dependence of a person's sex and attitudes towards abortion.

- The null hypothesis **\(H_0\)** will be a hypothesis expressing the absence of a relationship between the variables **sex** and **abany**.
- An alternative hypothesis **\(H_1\)** will represent the hypothesis that the variables **sex** and **abany** are independent.

We will use the chi-square test of independence to compare two categorical variables. The significance level in our case will be 5%. To apply this criterion, certain conditions must be met:

- Independence: Sampled observations must be independent.
    - random sample / assignment - condition meet, since respondants are randomly selected
    - if sampling without replacement, n <10% of the population - condition meet
    - each case only contributes to one cell in the table - condition meet
- Sample size. Each particular scenario must have at least 5 expected cases - condition meet.

Now let's run a test.

```
chisq.test(test_data)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  test_data
## X-squared = 12.487, df = 1, p-value = 0.0004097
```

As we can see, our p-value (0.0004097) is less than the specified significance level of 5% (0.05), which means we reject the null hypothesis, which said that there was no relationship between the two variables. Based on our test, we can say that the two studied variables are related, but we cannot prove a causal relationship in any way due to the stated method of conducting the research.

Part 5: Conclusion

Thanks to this study, the relationships between gender, class, age of the interviewed people and their attitude to abortion were studied. The hypothesis about the relationship between the sex of the interviewed person and his attitude to abortion was investigated, its result is the presence of a relationship between the studied groups of people. Gender influences attitudes towards abortion, which means that it is necessary to understand how gender adjustment should be taken into account in various campaigns to legalize / ban abortion. Causal relationships were not established in this work due to the specifics of the study.

It was found that with age the number of respondents supporting abortion decreases, and with an increase in the rank of the social class, on the contrary, it increases. This may be due to the level of the respondent's prosperity, his life priorities, experience, religious beliefs, culture, mentality and other criteria that need to be considered in more detail in subsequent studies.