Statistical inference with the GSS data

Setup

Load packages

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

Load data

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

Part 1: Data

The General Social Survey (GSS) gathers data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes.

The vast majority of GSS data is obtained in face-to-face interviews. Probability sampling (i.e. uses random sampling techniques) is used to create the sample. Because random sampling is used, the result of the analysis can be generalized to the population of interest.

Because random assignment is not used, the result of the analysis cannot be used for causal inference.

Part 2: Research question

Is the median family income in constant dollars different between year 2012 and 2000?

I am interested in knowing whether households in the United States have become richer over a 10 years period. Because the survey data in this sample is until 2012, so my research question is over a 10 years period between 2012 and 2002.

Part 3: Exploratory data analysis

```
family_income_2002_2012 <- gss %>%
  filter(year %in% c(2002, 2012)
         , !is.na(coninc)) %>%
  select(year, coninc)

family_income_2002_2012 <- family_income_2002_2012 %>%
  mutate(year = as.factor(year))

median_fam_inc_2002_2012 <- family_income_2002_2012 %>%
```

```
  group_by(year) %>%
  summarise(median_family_income = median(coninc))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
median_fam_inc_2002_2012
```

```
## # A tibble: 2 x 2
##   year  median_family_income
##   <fct>                <dbl>
## 1 2002                 36482
## 2 2012                 34470
```

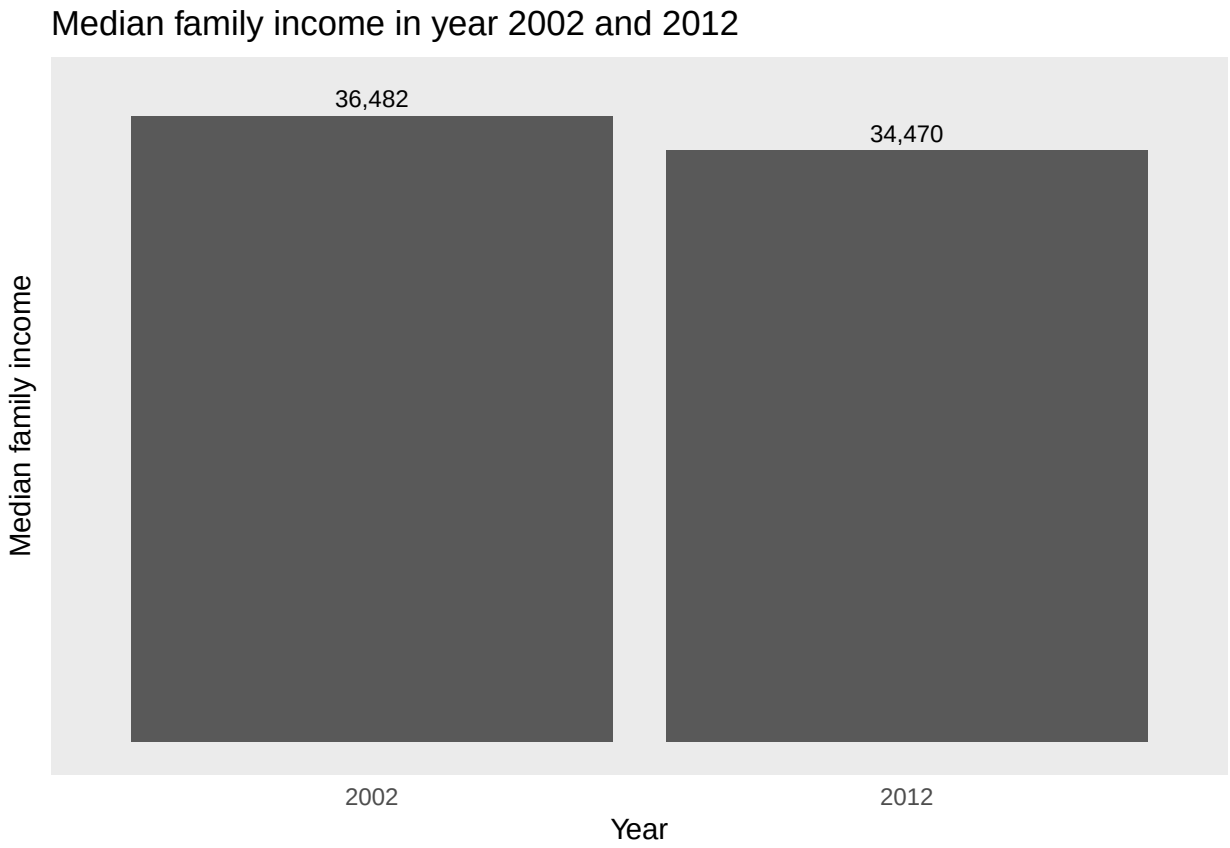
From the summary statistics, it shows that the median family income has decreased from year 2002 to 2012.

```
median_fam_inc_2002 <- median_fam_inc_2002_2012 %>%
  filter(year == 2002) %>%
  select(median_family_income)

median_fam_inc_2002 <- as.numeric(median_fam_inc_2002)

median_fam_inc_2012 <- median_fam_inc_2002_2012 %>%
  filter(year == 2012) %>%
  select(median_family_income)

median_fam_inc_2012 <- as.numeric(median_fam_inc_2012)

median_fam_inc_2012 - median_fam_inc_2002
```

```
## [1] -2012
```

It shows that the median family income has decreased by 2012.

```
ggplot(data = median_fam_inc_2002_2012
       , aes(x = year, y = median_family_income)
       ) +
  geom_col() +
  ggtitle('Median family income in year 2002 and 2012') +
  xlab('Year') +
  ylab('Median family income') +
  geom_text(aes(label = format(median_family_income, big.mark = ','))
            , vjust = -0.5
            , size = 3.3) +
  scale_y_continuous(labels = NULL
                     , lim = c(0, 38000)) +
  theme(axis.ticks = element_blank()
        , panel.grid = element_blank())
```



From this bar plot, it shows that there is little difference in the median family income between year 2012 and 2002.

Part 4: Inference

Hypotheses

The null hypothesis is that there is no difference in the median family income in constant dollars between year 2012 and 2002.

The alternative hypothesis is that there is difference in the median family income in constant dollars between year 2012 and 2002.

Conditions For Inference

I will check the conditions for inference for comparing two independent medians.

```
family_income_2002_2012 %>%
  group_by(year) %>%
  summarise(n = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 2
##   year      n
##   <fct> <int>
## 1 2002   2463
## 2 2012   1758
```

Because random sampling is used, and the sample size when sampling without replacement is less than 10% of population, the sampled observations are independent.

Because random sampling is used for each year, there is no dependence between the observations from 2012 and 2002, the two groups are independent of each other.

Because the sample size for each group is greater than 30, the sampling/bootstrap distribution will be nearly normal.

Method(s) To Be Used

I will conduct a hypothesis test to check the likelihood of the difference in the median family income to be 2012 or more extreme in either directions, if in fact that there is no difference in the median family income in constant dollars between year 2012 and 2002. I will use the default significance level of 0.05.

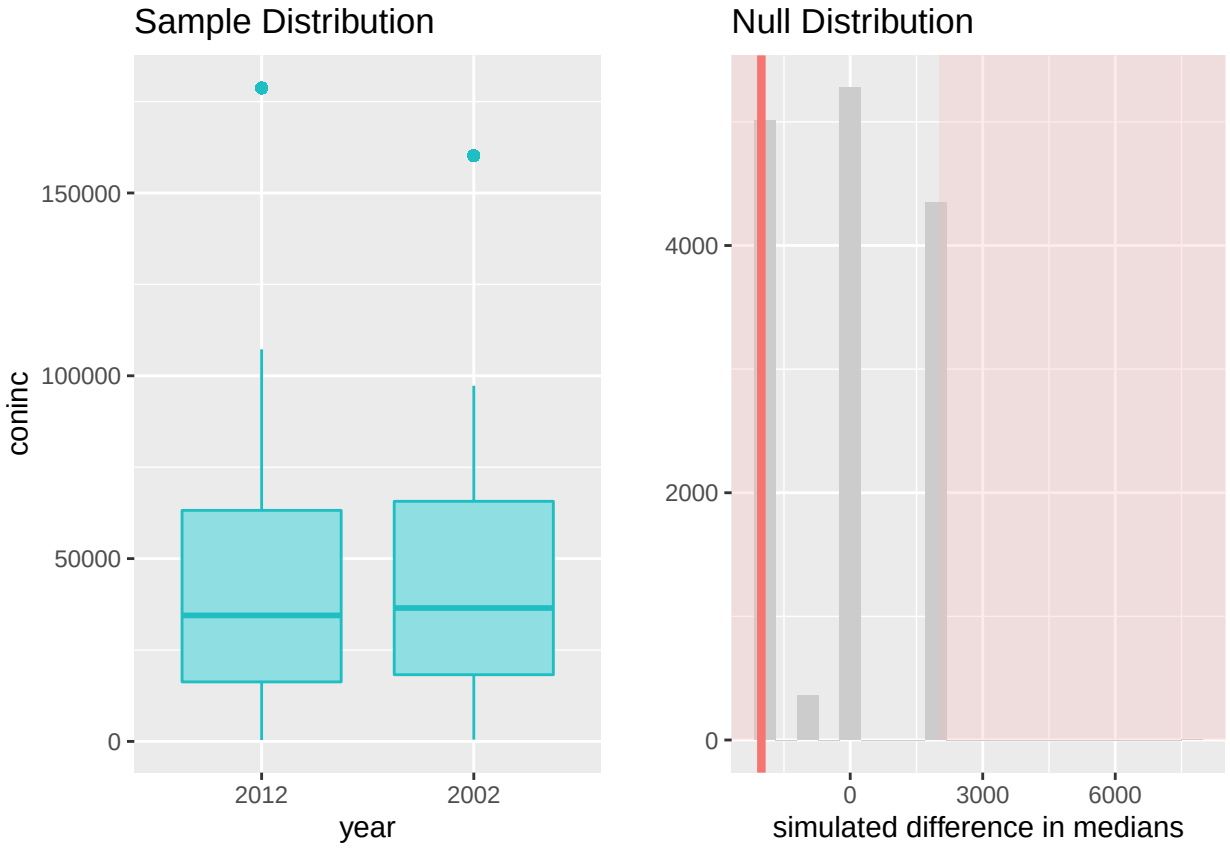
I will also calculate the 95% confidence interval to check whether the interval contains the null value, which is a difference of 0.

Because I am using median, I can only use the simulation method.

Hypothesis Test

```
inference(y = coninc, x = year, data = family_income_2002_2012, type = 'ht'
          , statistic = 'median', method = 'simulation', null = 0
          , alternative = 'twosided', order = c('2012', '2002'), seed = 1)
```

```
## Response variable: numerical
## Explanatory variable: categorical (2 levels)
## n_2012 = 1758, y_med_2012 = 34470, IQR_2012 = 46917
## n_2002 = 2463, y_med_2002 = 36482, IQR_2002 = 47427
## H0: mu_2012 =  mu_2002
## HA: mu_2012 != mu_2002
## p_value = 0.6683
```

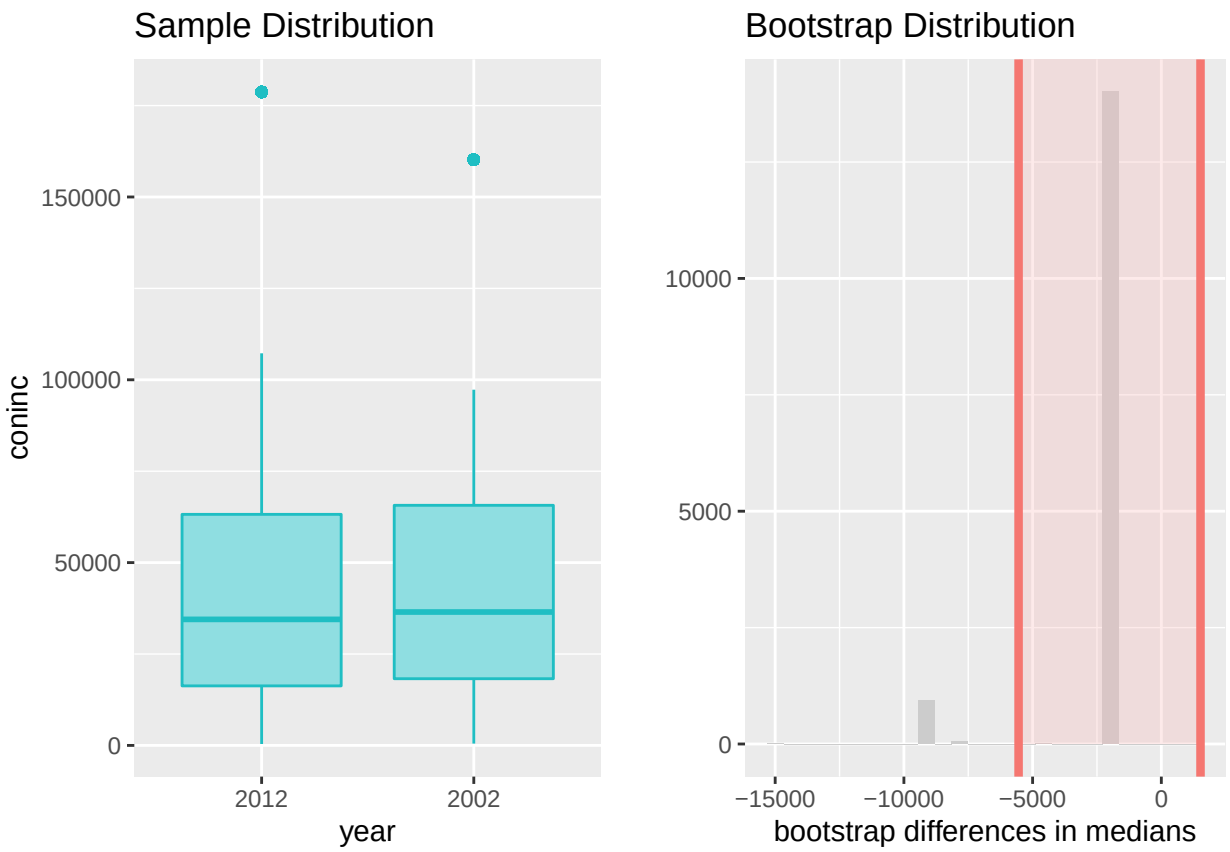


Because the p-value of 0.6683 is greater than the significance level of 0.05, I fail to reject the null hypothesis that there is no difference in the median family income in constant dollars between year 2012 and 2002. The data does not provide convincing evidence that there is difference in the median family income in constant dollars between year 2012 and 2002. The observed difference of -2012 is simply due to chance.

Confidence Interval

```
inference(y = coninc, x = year, data = family_income_2002_2012, type = 'ci'
          , statistic = 'median', method = 'simulation', order = c('2012', '2002')
          , boot_method = 'se', seed = 1)
```

```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_2012 = 1758, y_med_2012 = 34470, IQR_2012 = 46917
## n_2002 = 2463, y_med_2002 = 36482, IQR_2002 = 47427
## 95% CI (2012 - 2002): (-5545.2602 , 1521.2602)
```



The 95% confidence interval of the difference in the median family income in constant dollars between year 2012 and 2002 is -5545.2602 to 1521.2602, which contains the null value (i.e. difference of 0), therefore the result of the hypothesis test and the confidence interval agree with each other.