Statistical inference with the GSS data

Setup

Load packages

```
options(warn=-1) # turn off warnings
library(ggplot2)
library(dplyr)
library(statsr)
library(tidyr)
library(knitr)
library(colorspace)
```

Load data

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

Part 1: Data

The General Social Survey (GSS) data is composed of results from multiple surveys, which consist of demographic, behavioral, attitudinal, and special topic questions. Each survey corresponds to an independently drawn simple random sample from the population. Here, the population is adults (>= 18 years old) living in a non-institutional situation in the United States (US).

The data collection methodology is designed to represent the population, so the sample is generalizable to this population. There are, however, sources of bias, including voluntary response; some adults in the sample may choose to not participate, and this may correlate with responses to the survey questions.

Because this data was taken in the past and we are not actively collecting data for the study, our following analysis is strictly observational. We can not infer any causality among the variables in our analysis.

Information on the collection of this data was taken from the following GSS document online:

http://gss.norc.org/documents/codebook/GSS_Codebook.pdf

Part 2: Research question

Question: Did the proportion of adults in the United States who think their standard of living is better than their parent's was at the same age decrease between the years 1994 and 2012?

Time passing between generations can bring a number of changes that influnce the standard of living (economic changes, increasing population, etc.). By exploring the respondent's opinions on the matter, this question addresses whether the proportion of respondents who think their standard of living is better than their parent's was at the same age is decreasing over time, in particular, between the years 1994 and 2012 (the earliest and latest years for which this question was asked in the GSS).

For this research question, the following variables will be considered:

year: GSS year for the respondent parsol: Compared to your parents when they were the age you are now, do you think your own standard of living now is much better, somewhat better, about the same, somewhat worse, or much worse than theirs was?

Part 3: Exploratory data analysis

To begin the exploratory data analysis, we create a reduced dataframe with only the variables of interest, and we remove all non-complete cases (with NA values).

```
df <- gss[,c("year","parsol")]
df <- df[complete.cases(df),]
```

Next, we summarize the dataframe. Here, we see that the earliest year for which responses to the "parsol" question were recorded is 1994, and the latest year is 2012.

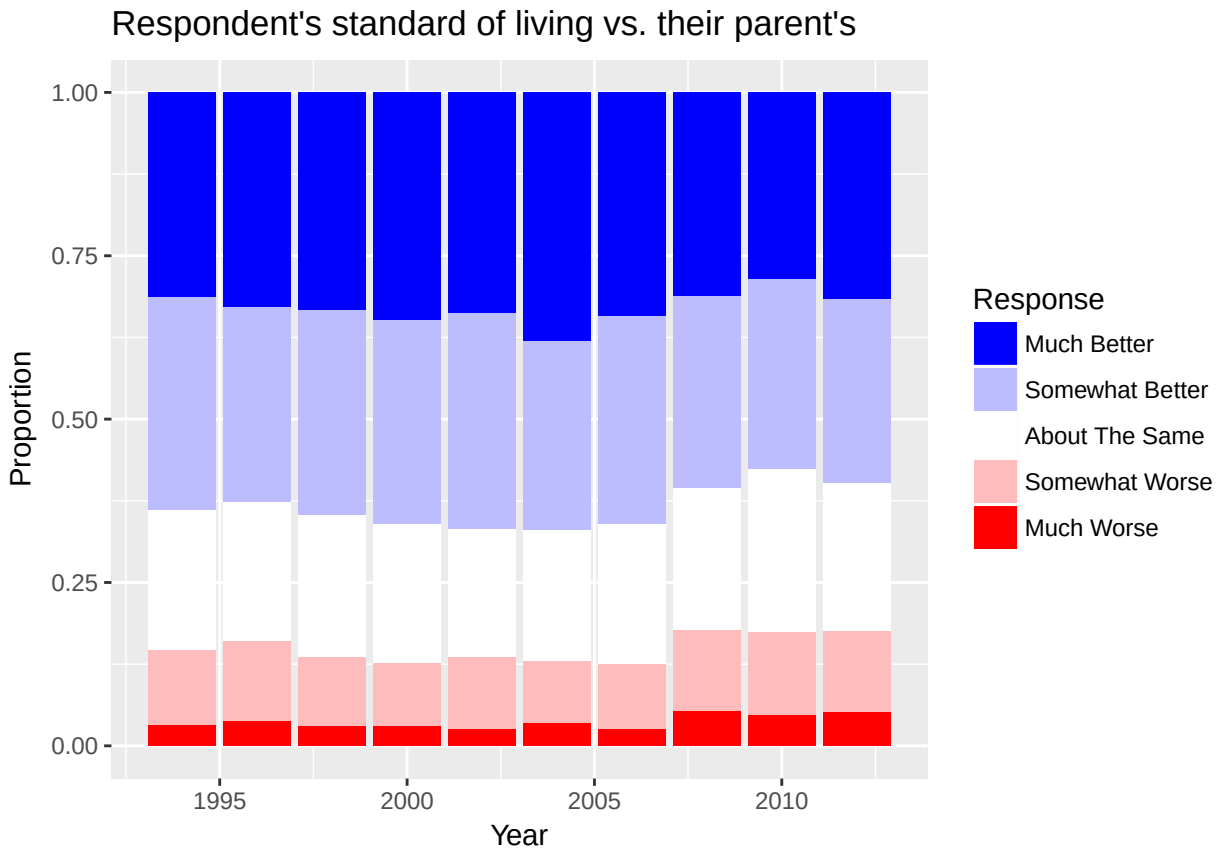
```
dfsum <- as.data.frame.matrix(addmargins(table(df)))
dfsum <- add_rownames(dfsum, "Year")
kable(dfsum,col.names = c("Year","Much Better","Somewhat Better","About The Same","Somewhat Worse","Much
```

Table 1: Number of occurances in dataframe

Year	Much Better	Somewhat Better	About The Same	Somewhat Worse	Much Worse	Sum
1994	460	478	313	170	47	1468
1996	619	564	401	229	74	1887
1998	626	589	408	199	57	1879
2000	646	579	396	178	57	1856
2002	303	298	175	99	24	899
2004	329	251	173	83	30	866
2006	671	624	421	196	52	1964
2008	417	394	290	166	72	1339
2010	389	394	341	173	64	1361
2012	413	368	296	162	69	1308
Sum	4873	4539	3214	1655	546	14827

Below, we visualize this data using a geometric bar plot. We can see that the proportion of "Much Better" and "Somewhat Better" responses decreased in the years leading up to 2012.

```
ggplot(df, aes(x= year, y = 1, fill = parsol)) +
    geom_bar(stat="identity", position = "fill") +
    scale_fill_manual(values = diverge_hsv(5)) +
    labs(title = "Respondent's standard of living vs. their parent's", fill = "Response",x = "Year", y =
```



Because we are interested only in the years 1994 and 2012, we filter out the other years from the dataframe, and change the "year" variable from a numeric type to a factor (catagorical type). We also introduce the new variable "better," which is defined to be "Better" if the response to the "parsol" question was "Much Better" or "Somewhat Better," and "Not Better" otherwise.

```
df <- df[df$year == 1994 | df$year == 2012,]
df$year <- factor(df$year)
df$better <- ifelse(df$parsol == "Much Better" | df$parsol == "Somewhat Better", "Better", "Not Better")
df <- df[c("year","better")]
```

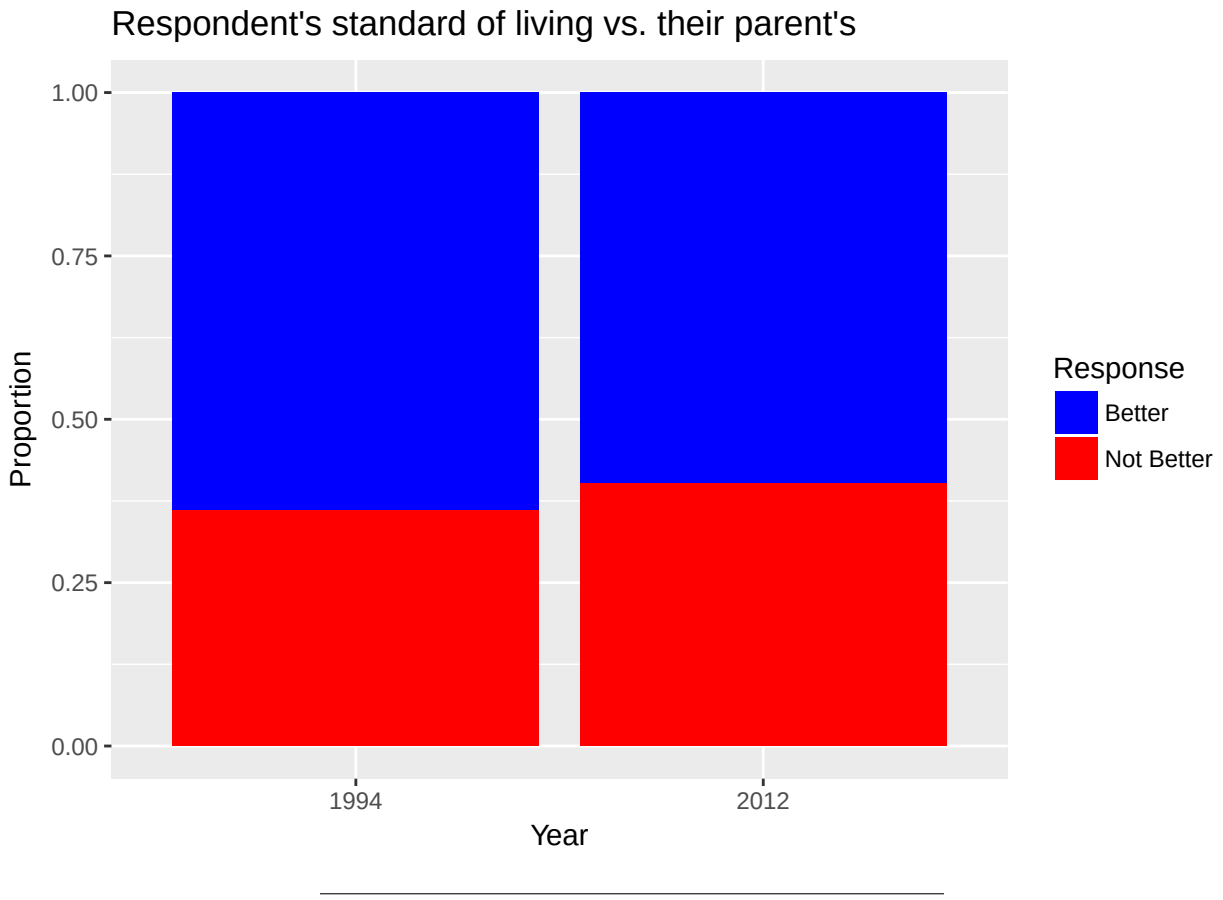
Next, we summarize this simplified dataframe. It is most convienent to inspect proportions of responses, so the table and barchart below compare the proportion of "Better" and "Not Better" responses between the years 1994 and 2012. This shows a decrease in the proportion of respondents who indicate that their standard of living is better than their parent's was at the same age, from 0.639 in 1994 to 0.597 in 2012. Below, we determine if this change is statistically significant by performing inferential statistics on the data.

```
dfsum2 <- table(df$year,df$better)
dfsum2 <- prop.table(dfsum2,1)
dfsum2 <- as.data.frame.matrix(dfsum2)
dfsum2 <- add_rownames(dfsum2, "Year")
kable(dfsum2,col.names = c("Year","Better","Not Better"),caption = "Proportion of occurances per year")
```

Table 2: Proportion of occurances per year

Year	Better	Not Better
1994	0.6389646	0.3610354
2012	0.5970948	0.4029052

```r
ggplot(df, aes(x= year, y = 1, fill = better)) +
    geom_bar(stat="identity", position = "fill") +
    scale_fill_manual(values = diverge_hsv(2))  +
    labs(title = "Respondent's standard of living vs. their parent's", fill = "Response",x = "Year", y =
```



Part 4: Inference

In this section we perform inferential statistics on the data to address the research question stated above. If the response "Better" corresponds to success, then the sample data has a proportion difference of

$\hat{p}_{1994} - \hat{p}_{2012} = 0.0418.$

We employ inferential statistics to determine if this difference is statistically significant. We begin by establishing a null hypothesis and one-sided alternative hypothesis (where $p_{...}$ is a population proportion):

Null hypothesis $H_0 : p_{1994} - p_{2012} = 0$ Alternative hypothesis $H_A : p_{1994} - p_{2012} > 0$

These hypotheses are appropriate for our research question, which asks if the proportion of respondents who indicated their standard of living is greater than their parent's was at the same age decreased from the year 1994 to the year 2012. If it did indeed decrease, then it would be true that $H_A : p_{1994} - p_{2012} > 0$, which is our alternative hypothesis.

To choose the appropriate inference method for the task at hand, we must explore the criteria for using theoretical and/or simulation methods for inferential statistics. Here, we have two catagorical variables, "year" (the explanatory variable) and "better" (the dependent variable). Thus, we will perform inference to compare proportions. To determine whether we use theoretical (via central limit theorem) or simulation-based methods,

we must first check the success-failure condition. Below, we tabulate the number of successes ("Better") and failures ("Not Better") for each year. Because all numbers are greater than 10, the success-failure condition is easily met and we may use a theoretical method of inference. Further, the total number of respondents (2776) is much less than 10% of the population, so we can safely assume independence of the responses.

```
dfsum3 <- as.data.frame.matrix(addmargins(table(df)))
dfsum3 <- add_rownames(dfsum3, "Year")
kable(dfsum3,col.names = c("Year","Better","Not Better","Sum"), caption = "Number of occurances in data
```

Table 3: Number of occurances in dataframe

Year	Better	Not Better	Sum
1994	938	530	1468
2012	781	527	1308
Sum	1719	1057	2776

Below, we perform both a hypothesis test and calculate a 95% confidence interval for the difference in proportions. For each, we calculate them independently, and compare against the results of the inference.R function.

Hypothesis Test

Here we perform a hypothesis test to determine if the decrease in the proportion of the population indicating their standard of living is better than their parent's at the same age is statistically significant. Because our null Hypothesis is $H_0 : p_{1994} - p_{2012} = 0$, we use a pooled proportion to calculate the standard error SE. Below, we calculate the p-value "by hand," then use the inference.R function. Because our alternative hypothesis is one-sided, we do not multiply our p-value by 2, and we choose the alternative="greater" option for the inference.R function.

We see that our results agree with those of the inference.R function; both produce a p-value of 0.0117. Using a confidence level of 0.05, we thus reject the null hypothesis and conclude that the observed inequality in the proportions is statistically significant. An equivalent interpretation is: there is a probability of 0.0117 of drawing an equivalently sized simple random sample with a difference in sample proportions being as or more extreme than that observed in this sample, given that the null hypothesis is true.

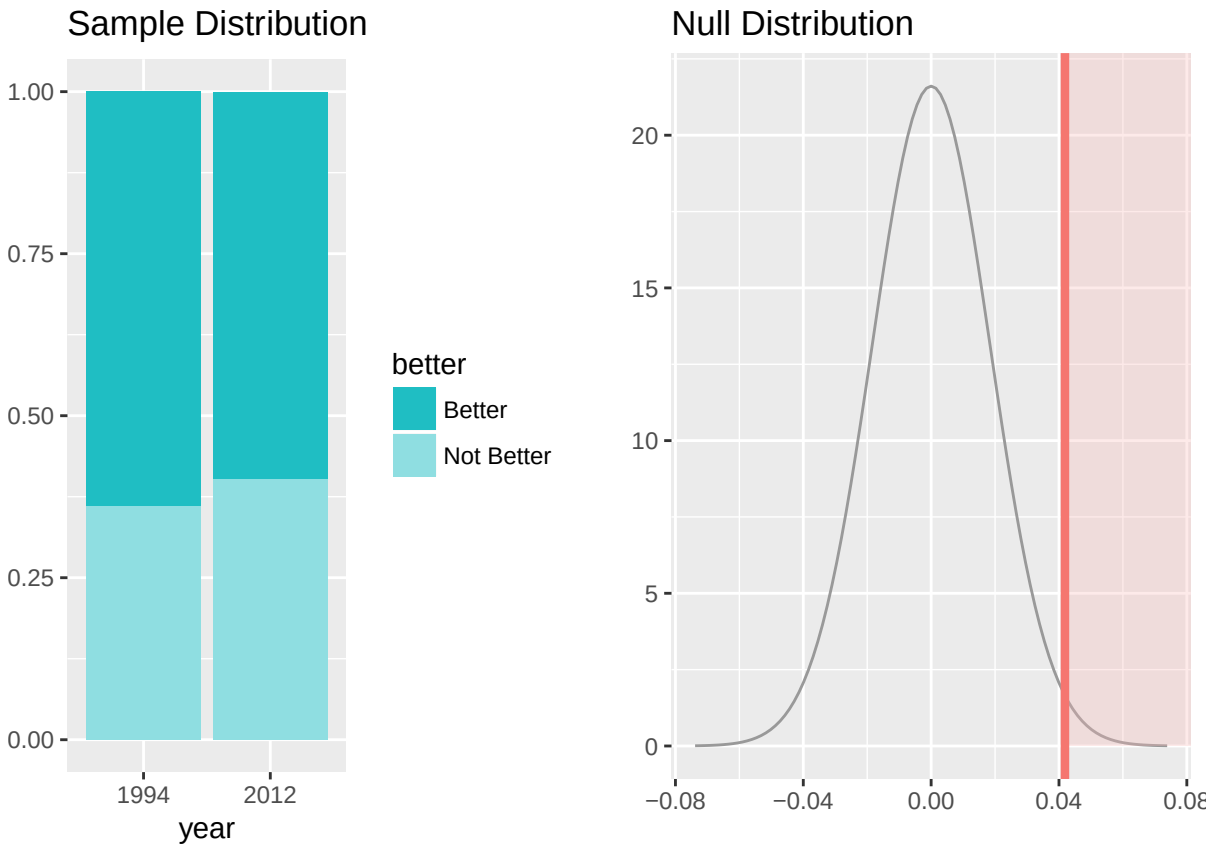
```
p1 = dfsum3[["Better"]][1]/dfsum3[["Sum"]][1]
p2 = dfsum3[["Better"]][2]/dfsum3[["Sum"]][2]
dp = p1-p2
p_pool = dfsum3[["Better"]][3]/dfsum3[["Sum"]][3]
n1 = dfsum3[["Sum"]][1]
n2 = dfsum3[["Sum"]][2]
SE = sqrt(p_pool*(1-p_pool)*(1/n1+1/n2))
zscore = dp/SE
pvalue = pnorm(zscore,lower.tail=FALSE)
cat("p-value:",pvalue)
```

```
## p-value: 0.01167116
```

```
inference(y = better, x = year, data = df, statistic = "proportion", type = "ht", alternative = "greate
```

```
## Response variable: categorical (2 levels, success: Better)
## Explanatory variable: categorical (2 levels)
## n_1994 = 1468, p_hat_1994 = 0.639
## n_2012 = 1308, p_hat_2012 = 0.5971
## H0: p_1994 =  p_2012
## HA: p_1994 > p_2012
```

```
## z = 2.2678
## p_value = 0.0117
```



Confidence Interval

Here we calculate a 95% confidence interval for the difference in the population proportions $p_{1994} - p_{2012}$. Because this is not a hypothesis test, we do not need to use the pooled proportion for calculating the standard error. Below, we calculate the standard error and the corresponding margin of error for a 95% confidence interval by hand, and we calculate the 95% confidence interval using the inference.R function.

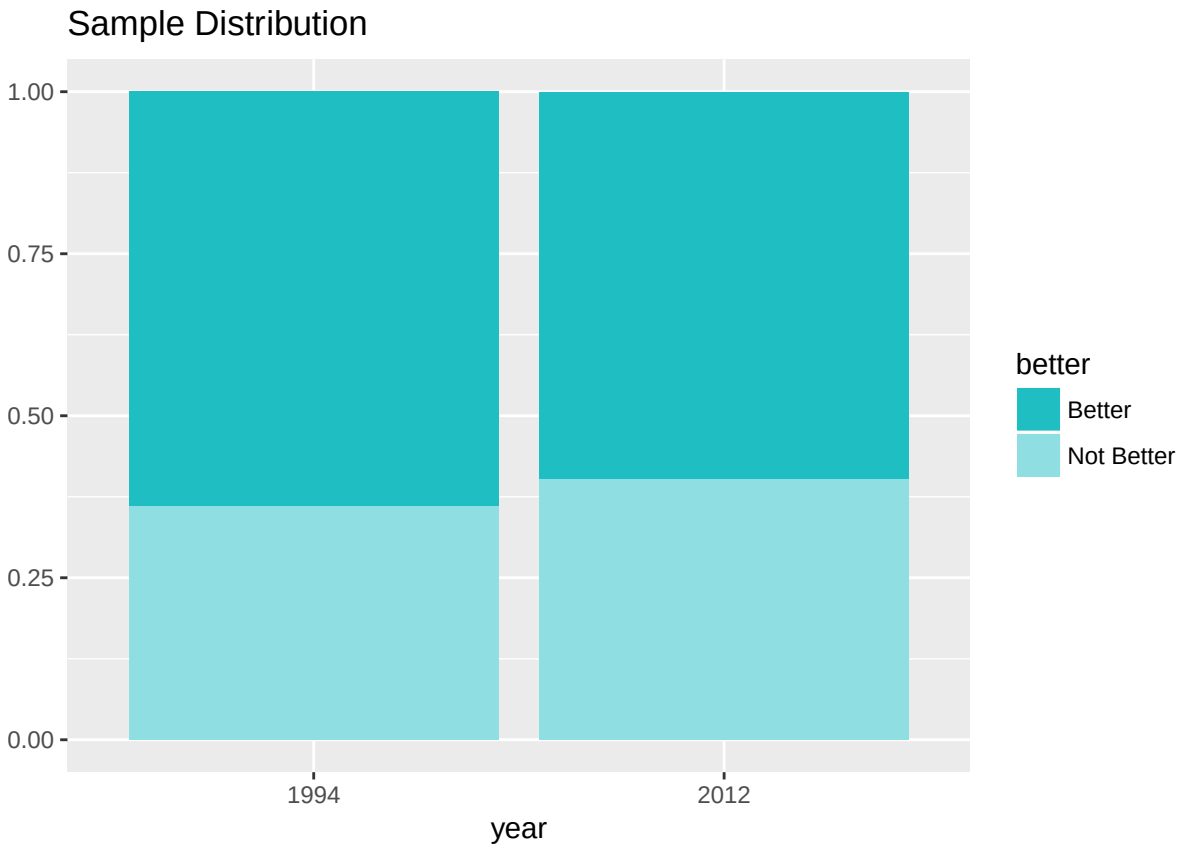
We see that our results agree with those of the inference.R function, obtaining a 95% confidence interval of $p_{1994} - p_{2012} \in (0.0057, 0.0781)$. Note that this result is consistent with that of the hypothesis test. The null hypothesis value of $p_{1994} - p_{2012} = 0$ is not contained in this 95% confidence interval, and we rejected the null hypothesis with the above hypothesis test with a confidence level of 0.05.

```
SE = sqrt(p1*(1-p1)/n1+p2*(1-p2)/n2)
ME = qnorm(.975)*SE
CIlower = dp-ME
CIupper = dp+ME
cat("95% CI: (",CIlower,",",CIupper,")")
```

```
## 95% CI: ( 0.005673057 , 0.0780665 )
```

```
inference(y = better, x = year, data = df, statistic = "proportion", type = "ci", alternative = "greate
```

```
## Response variable: categorical (2 levels, success: Better)
## Explanatory variable: categorical (2 levels)
## n_1994 = 1468, p_hat_1994 = 0.639
## n_2012 = 1308, p_hat_2012 = 0.5971
## 95% CI (1994 - 2012): (0.0057 , 0.0781)
```


Sample Distribution
1.00
0.75
0.50
0.25
0.00
1994
2012
year
better
Better
Not Better