Statistical inference with the GSS data

Setup

Load packages

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

Load data

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

Part 1: Data

From the introduction to the GSS Codebook,

> "Each survey from 1972 to 2004 was an independently drawn sample of English-speaking persons 18 years of age or over, living in non-institutional arrangements within the United States. Starting in 2006 Spanish-speakers were added to the target population."

So, independence between groups is ensured.

> "Full probability sampling was employed in half of the 1975 and 1976 surveys and the 1977, 1978, 1980, 1982-1991, 1993-1998, 2000, 2002, 2004, 2006, 2008, 2010, 2012, and 2014 surveys."

And, full probability sampling is employed in many years. Analysis will be conducted in these years in order to apply learned statistical inferences appropriately because the samples are completely random.

> "The median length of the interview has been about one and a half hours."

Observations are collected by interviews accompanied by questionaires.

Part 2: Research question

Question: As time goes, has US people's attitude to homosex changed?

Homosex is always a controversial topic in the US and other places in the world. Investigate this question will inform ourselves of the trend and how the opinions are changing.

Particularly, we want to know whether there is a big difference in the proprotion of people who accept homosex in the year 1982 and 2012.

Part 3: Exploratory data analysis

In the dataset, the opinions about homosex are one of the followings:

- Always Wrong
- Almst Always Wrg
- Sometimes Wrong
- Not Wrong At All
- Other

And we will consider **Not Wrong At All** as agree with homosex and other four choices as the opposite. The proportion of the first people is a key index of people's opinions to homosex.

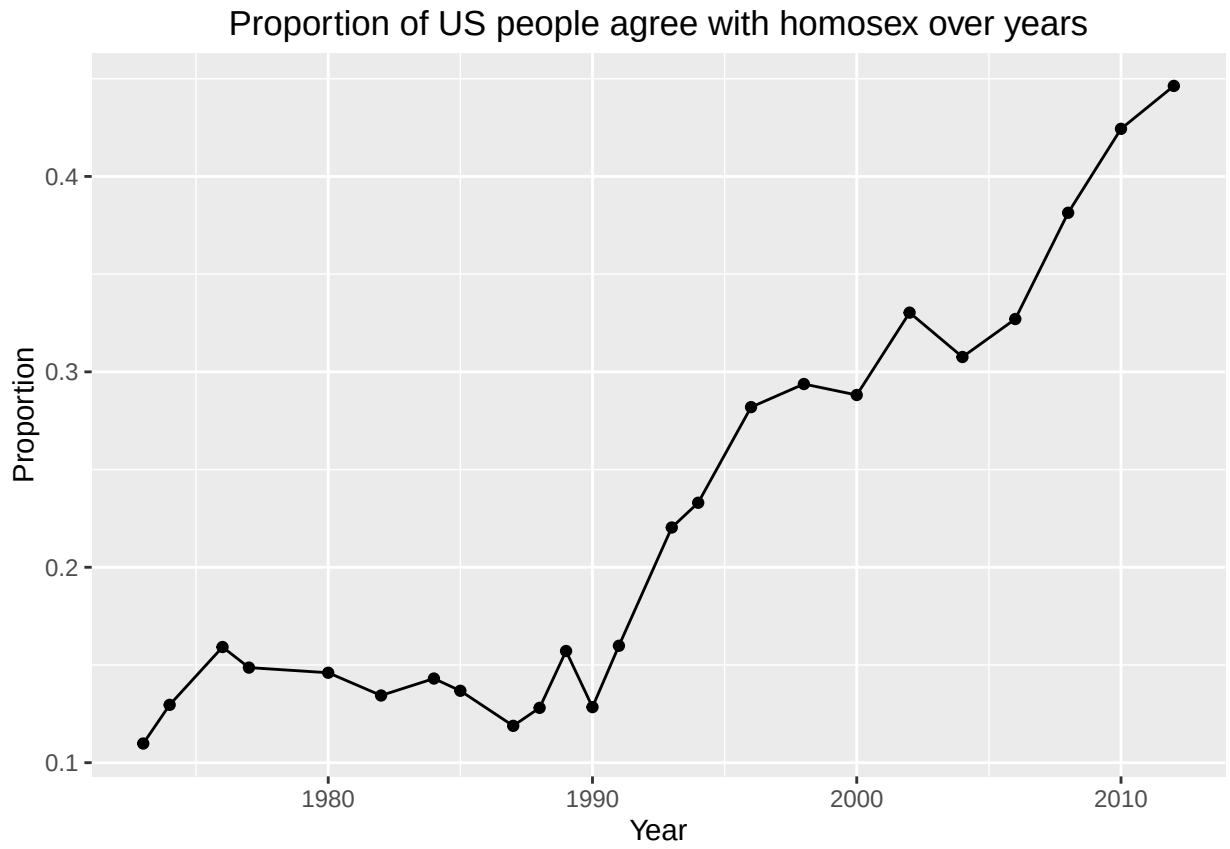
First, let's have a look at the trend of the proportion.

```
# Get a function to calculate the proportion
get_proportion <- function(x) {
  nume <- sum(x == "Not Wrong At All", na.rm = T)
  deno <- sum(!is.na(x), na.rm = T)
  return(nume / deno)
}

# aggregate the proportions by years
prop_aggr <- aggregate(gss$homosex, by = list(gss$year), get_proportion)
colnames(prop_aggr) = c("year", "proportion")
prop_aggr <- prop_aggr[!is.nan(prop_aggr$proportion), ]
head(prop_aggr)
```

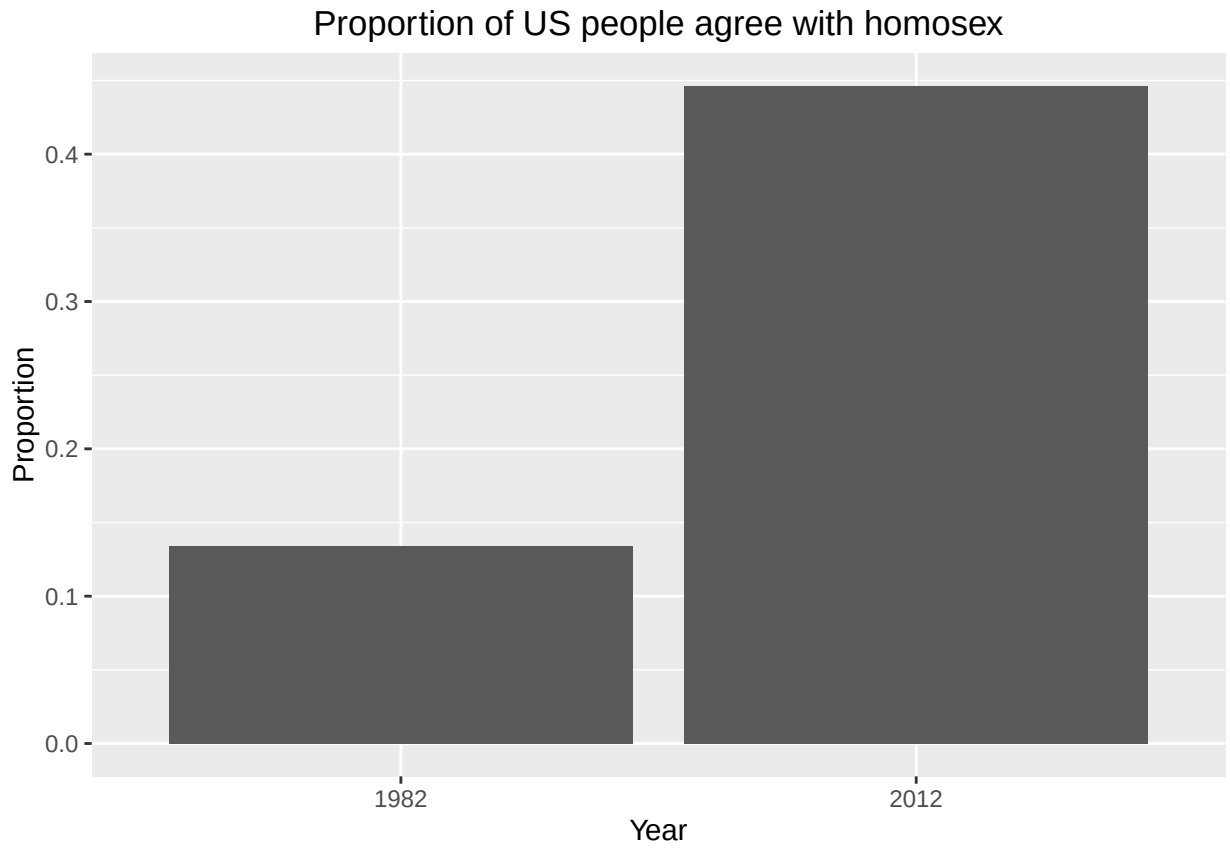
```
##   year proportion
## 2 1973  0.1098066
## 3 1974  0.1296034
## 5 1976  0.1591865
## 6 1977  0.1486579
## 8 1980  0.1460272
## 9 1982  0.1343874
```

```
# Now plot the trend
g <- ggplot(data = prop_aggr, aes(x = year, y = proportion))
g <- g + geom_point() + geom_line()
g <- g + xlab("Year") + ylab("Proportion")
g <- g + labs(title = "Proportion of US people agree with homosex over years")
g
```



In the above plot, there is a clear uptrend of the proportion of US people who have the opinion that homosex is **not wrong at all**. Particularly, we want to check seriously whether there is a difference of proportion in the year 1982 and 2012. First, let's look at these two years' comparison.

```
two_year_cmp <- prop_aggr[prop_aggr$year %in% c("1982", "2012"), ]
g <- ggplot(data = two_year_cmp, aes(x = factor(year), y = proportion))
g <- g + geom_bar(stat = "identity")
g <- g + labs(x = "Year", y = "Proportion")
g <- g + labs(title = "Proportion of US people agree with homosex")
g
```



The comparison between 1982 and 2012 shows that there is a difference in the proportion of US people agree with homosex.

```
gss1982 <- gss[gss$year == "1982", ]
gss2012 <- gss[gss$year == "2012", ]
n1982 <- sum(!is.na(gss1982$homosex), na.rm = T)
n2012 <- sum(!is.na(gss2012$homosex), na.rm = T)
two_year_cmp$n <- c(n1982, n2012)
two_year_cmp
```

```
##    year proportion    n
## 9  1982  0.1343874 1771
## 29 2012  0.4463277 1239
```

```
p_diff_hat <- two_year_cmp[2, 2] - two_year_cmp[1, 2]
p_diff_hat
```

```
## [1] 0.3119403
```

And we get

$$\hat{p}_{1982} = 0.1343874$$

$$\hat{p}_{2012} = 0.4463277$$

$$\hat{p}_{diff} = \hat{p}_{2012} - \hat{p}_{1982} = 0.3119403$$

But, is that difference statistically significant? We will use statistical inference to check it strictly.

Part 4: Inference

Hypothsis test

First, we need hypotheses:

H_0: There is no difference in the proportion of US people who think homosex is not wrong at all between year 1982 and 2012.

H_A: There is a difference.

We use p_{1982} and p_{2012} to denote the true proportion of US people who agree with homosex in year 1982 and 2012 respectively. Then our hypotheses can be conveyed mathematically:

$$H_0 : p_{2012} - p_{1982} = 0$$

$$H_A : p_{2012} - p_{1982} \neq 0$$

And we set our significance level as $\alpha = 0.05$.

I want to use hypothesis tests for comparing two proportions. First, let's check conditions.

```
two_year_cmp$success <- with(two_year_cmp, round(n * proportion))
two_year_cmp$failure <- with(two_year_cmp, n - success)
two_year_cmp
```

```
##    year proportion    n success failure
## 9  1982  0.1343874 1771     238    1533
## 29 2012  0.4463277 1239     553     686
```

```
# now compute pooled proportion
p_pool <- sum(two_year_cmp$success) / sum(two_year_cmp$n)
two_year_cmp$success_hat <- round(two_year_cmp$n * p_pool)
two_year_cmp$failure_hat <- round(two_year_cmp$n * (1 - p_pool))
two_year_cmp
```

```
##    year proportion    n success failure success_hat failure_hat
## 9  1982  0.1343874 1771     238    1533         465        1306
## 29 2012  0.4463277 1239     553     686         326         913
```

The numbers of success and failure computed are great or equal to 10. So the success-failure condition satisfies. The independence is also satisfied because of the full probability sampling. And we can use CLT for the two proportion comparison. In another word,

$$\hat{p}_{2012} - \hat{p}_{1982} = \hat{p}_{diff} \sim N(\mu, SE)$$

Under the H_0, we have $\mu = 0$, $SE = \sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}}$

```
se <- sqrt(p_pool * (1 - p_pool) * (1/two_year_cmp[1, "n"] + 1/two_year_cmp[2, "n"]))
se
```

```
## [1] 0.01630192
```

Now, I get all the elements I need to calculate our z-score. Use the formula to compute:

$$z^* = \frac{\hat{p}_{diff} - 0}{SE}$$

```
z_score <- p_diff_hat / se
z_score
```

```
## [1] 19.13519
```

Finally, we calculate our p-value.

```
p_value <- 2 * pnorm(z_score, lower.tail = FALSE)
print(paste("p-value: ", p_value), quote = FALSE)
```

```
## [1] p-value:  1.28605286758851e-81
```

The p-value is almost 0, less than 0.05 of course. Hence, we reject the null hypothesis in favour of the alternative hypothesis. In another word, the difference of the proportions of people who agree with homosex in year 1982 and 2012 are statistically significant. We can't consider the two proportions as the same. The proportion has changed over the time.

Confidence Interval

Next, we want to construct 95% confidence interval of p_{diff}. Because of the difference between procedure HT and CI, we need to recalculate the center and standard error of p_{diff}.

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

```
se <- with(two_year_cmp, sqrt(sum(proportion * (1 - proportion) / n)))
se
```

```
## [1] 0.01628297
```

```
p_diff_hat + c(-1, 1) * 1.96 * se
```

```
## [1] 0.2800257 0.3438550
```

The confidence interval is $(0.280, 0.344)$. We are 95% confident that the proportion of US people who agree with homosex in 2012 is 0.280 to 0.344 higher than that in 1982.

0 is not in the interval $(0.280, 0.344)$. HT and CI agree with each other, which is expected.

Conclusion

The proportions of US people who agree with homosex has significantly difference between 1982 and 2012. And the trend is more and more people arguing that homosex is not wrong at all.(Easy to get the information in the EDA phase).