Lin-Lab6

Bin Lin

2016-10-29

Exercise 1: In the first paragraph, several key findings are reported. Do these percentages appear to be sample statistics (derived from the data sample) or population parameters?

They are sample statistics. It is impossible to ask every single person in the world if he or she is religious or not.

Exercise 2: The title of the report is “Global Index of Religiosity and Atheism”. To generalize the report’s findings to the global human population, what must we assume about the sampling method? Does that seem like a reasonable assumption?

The sampling method must include the countries that are very populous so that they can cover majority of the human population into the population parameters. In addition, the number of people they ask the questions are randomly selected from each individual countries and represent less than 10% of the nation’s population. In addition, each subjects’ responses are independent to each other. According to the report, they surveyed 57 countries, 51927 people. So the sample size is quite large. I think those are reasonable assumptions, so that the hypothesis test can be carried out.

```
#install.packages('Rcpp')
#install.packages('stringi')
#install.packages("inference")
library(inference)
```

```
## Loading required package: sandwich
```

```
library(IS606)
```

```
##
## Welcome to CUNY IS606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 3rd Edition. You can read this by typing
## vignette('os3') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='IS606') will list the demos that are available.
```

```
##
## Attaching package: 'IS606'
```

```
## The following object is masked from 'package:utils':
##
##     demo
```

```
#startLab('Lab6')
setwd('C:/Users/blin261/Documents/Lab6')
load("more/atheism.RData")
```

Exercise 3: What does each row of Table 6 correspond to? What does each row of atheism correspond to?

Each row represents the observation of the survey result for each individual country. The row of atheism tells us, the percentage of respondents (sample) are atheist for its corresponding country.

Exercise 4: Using the command below, create a new dataframe called us12 that contains only the rows in atheism associated with respondents to the 2012 survey from the United States. Next, calculate the proportion of atheist responses. Does it agree with the percentage in Table 6? If not, why?

Proportion of atheist responses = 50/1002 = 0.0499. It is very closed to 0.05 The proportion of atheist responses matches the percentage in Table 6.

```
us12 <- subset(atheism, nationality == "United States" & year == "2012")
summary(us12)
```

```
##         nationality          response        year
##  United States:1002   atheist    : 50   Min.   :2012
##  Afghanistan  :   0   non-atheist:952   1st Qu.:2012
##  Argentina    :   0                     Median :2012
##  Armenia      :   0                     Mean   :2012
##  Australia    :   0                     3rd Qu.:2012
##  Austria      :   0                     Max.   :2012
##  (Other)      :   0
```

```
50 / 1002
```

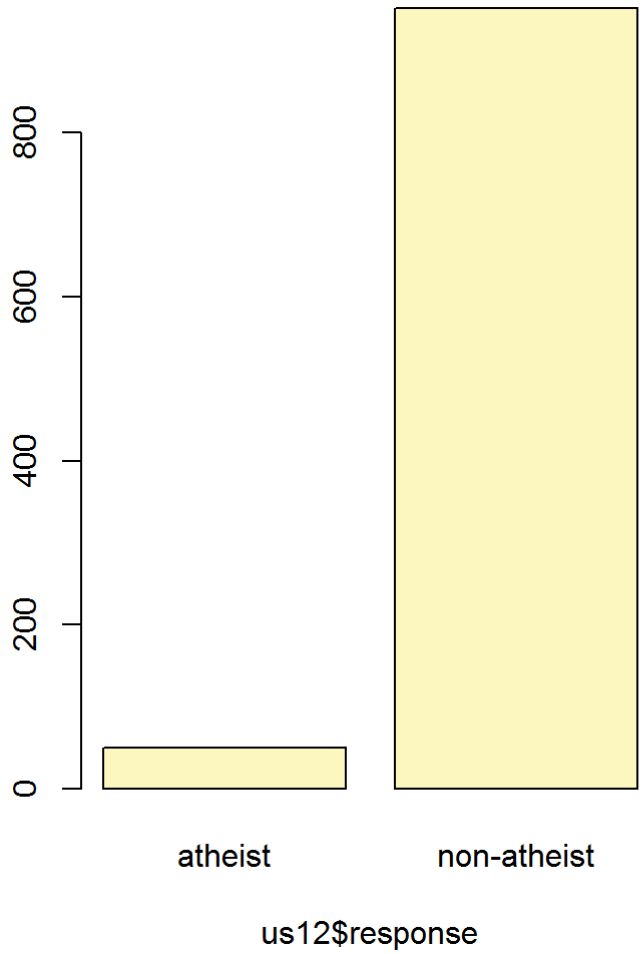
```
## [1] 0.0499002
```

Exercise 5: Write out the conditions for inference to construct a 95% confidence interval for the proportion of atheists in the United States in 2012. Are you confident all conditions are met?

Responders are randomly selected and represent less than 10% of population, therefore the test is independent within groups. Both success(atheist) and failure(non-atheist) are greater than 10. The conditions are all met.

```
inference(us12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

```
## Single proportion -- success: atheist
## Summary statistics:
```




```
## p_hat = 0.0499 ;  n = 1002
## Check conditions: number of successes = 50 ; number of failures = 952
## Standard error = 0.0069
## 95 % Confidence interval = ( 0.0364 , 0.0634 )
```

Exercise 6: Based on the R output, what is the margin of error for the estimate of the proportion of the proportion of atheists in US in 2012?

Margin of error equals 0.013524

```
standard_error <- 0.0069
qnorm(0.975)
```

```
## [1] 1.959964
```

```
margin_of_error <- 1.96 * standard_error
margin_of_error
```

```
## [1] 0.013524
```

Exercise 7: Using the inference function, calculate confidence intervals for the proportion of atheists in 2012 in two other countries of your choice, and report the associated margins of error. Be sure to note whether the conditions for inference are met. It may be helpful to create new data sets for each of the two countries first, and

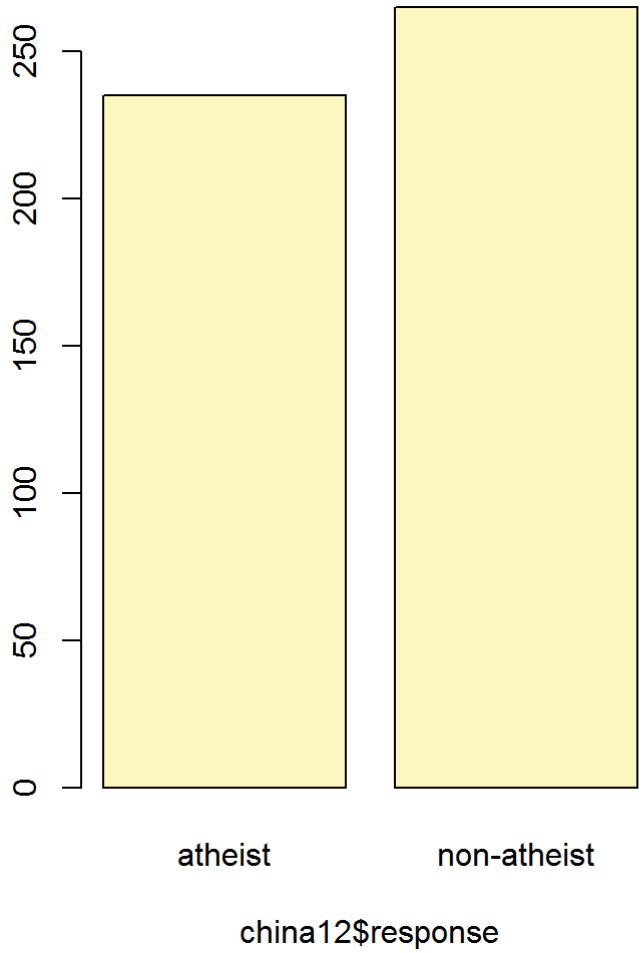
then use these data sets in the inference function to construct the confidence intervals.

I picked China and Japan. All conditions for inference are met. China: 95 % Confidence interval = (0.4263 , 0.5137) Standard error = 0.0223 Margin of error = 0.043708

Japan: 95 % Confidence interval = (0.281 , 0.3329) Standard error = 0.0132 Margin of error = 0.025872

```
china12 <- subset(atheism, nationality == "China" & year == "2012")
inference(china12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

```
## Single proportion -- success: atheist
## Summary statistics:
```



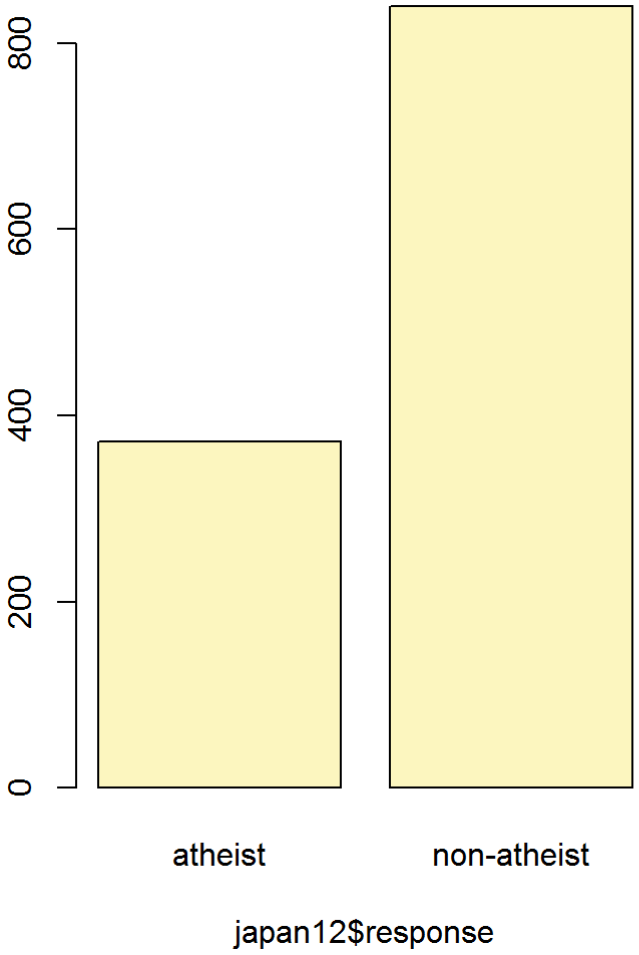

```
## p_hat = 0.47 ;  n = 500
## Check conditions: number of successes = 235 ; number of failures = 265
## Standard error = 0.0223
## 95 % Confidence interval = ( 0.4263 , 0.5137 )
```

```
standard_error <- 0.0223
moe <- 1.96 * standard_error
moe
```

```
## [1] 0.043708
```

```
japan12 <- subset(atheism, nationality == "Japan" & year == "2012")
inference(japan12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

```
## Single proportion -- success: atheist
## Summary statistics:
```

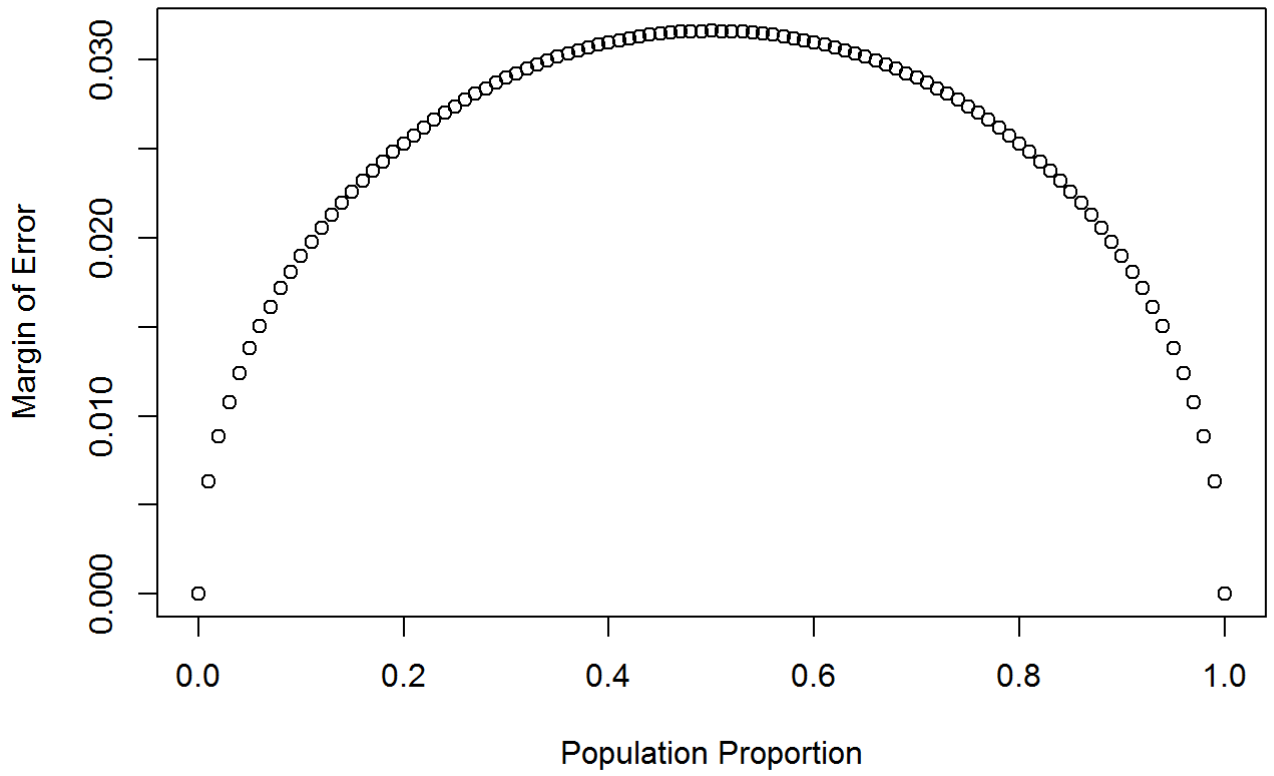



```
## p_hat = 0.3069 ;  n = 1212
## Check conditions: number of successes = 372 ; number of failures = 840
## Standard error = 0.0132
## 95 % Confidence interval = ( 0.281 , 0.3329 )
```

```
standard_error <- 0.0132
moe <- 1.96 * standard_error
moe
```

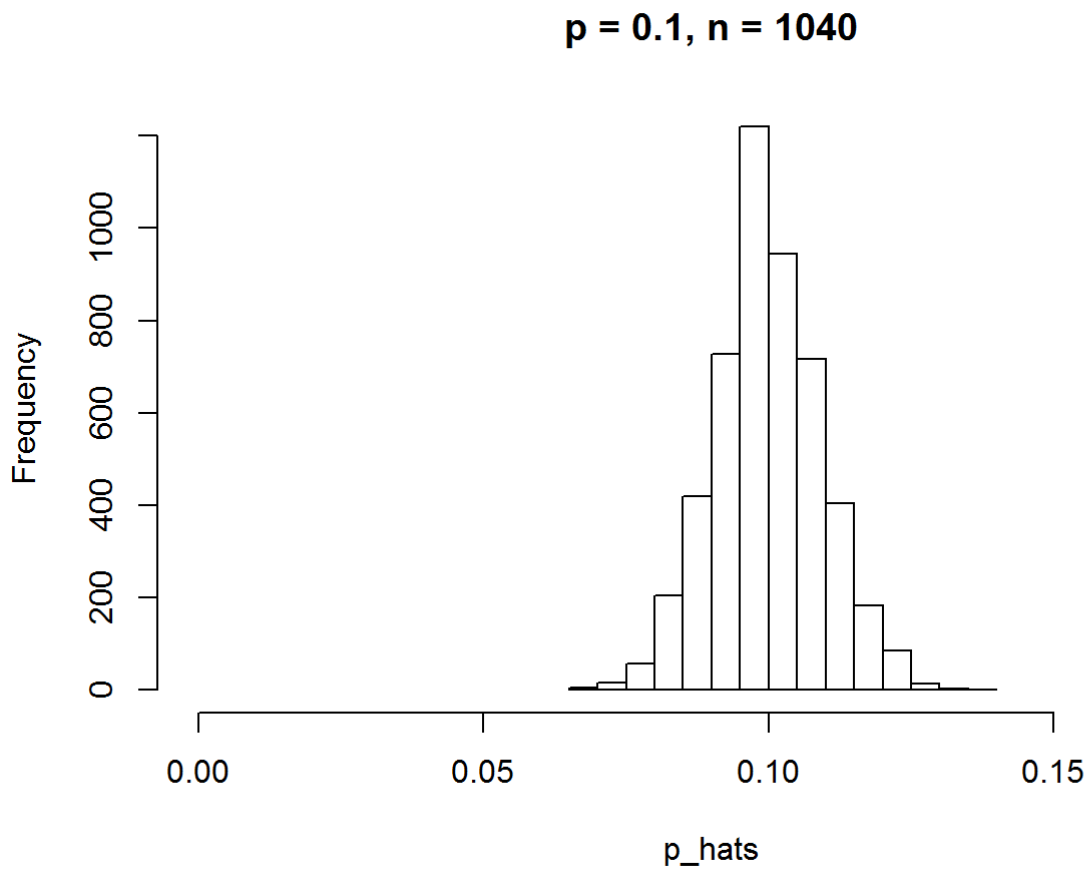
```
## [1] 0.025872
```

```
n <- 1000
p <- seq(0, 1, 0.01)
me <- 2 * sqrt(p * (1 - p)/n)
plot(me ~ p, ylab = "Margin of Error", xlab = "Population Proportion")
```



Exercise 8: Describe the relationship between p and me. It is unimodel, bell shaped graph with one peak at p = 0.5. Therefore, when p= 0.5, the margin of error is the highest. When p = 0 or 1, margin of error should be equal to 0.

```
set.seed(10)
p <- 0.1
n <- 1040
p_hats <- rep(0, 5000)

for(i in 1:5000){
  samp <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
  p_hats[i] <- sum(samp == "atheist")/n
}

hist(p_hats, main = "p = 0.1, n = 1040", xlim = c(0, 0.18))
```



Exercise 9: Describe the sampling distribution of sample proportions at n=1040 and p=0.1. Be sure to note the center, spread, and shape. Hint: Remember that R has functions such as mean to calculate summary statistics.

The sampling distribution is unimodel and bell-shapped with a peak centered at p = 0.1. The distribution approaches normal distribution. The range of the distribution is from 0.06731 to 0.13560. The median is 0.1 and the mean is 0.09997

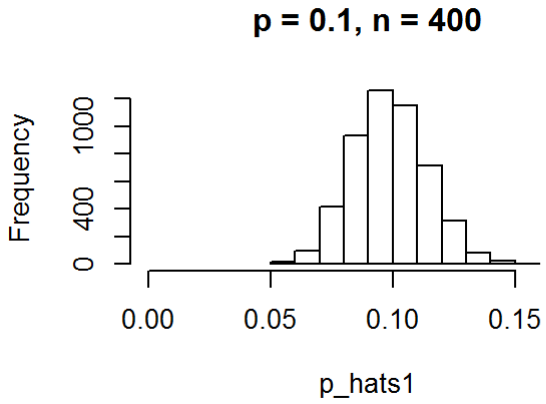
```
summary(p_hats)
```

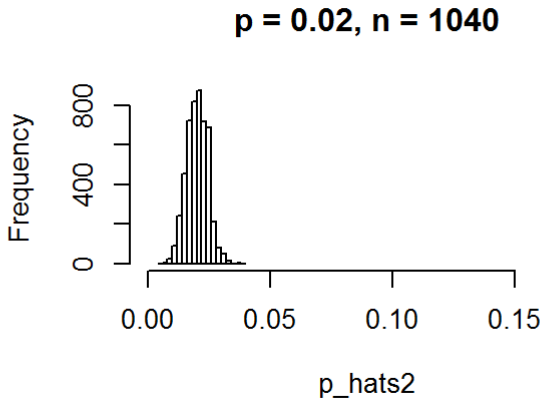
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06731 0.09327 0.10000 0.09997 0.10670 0.13560
```

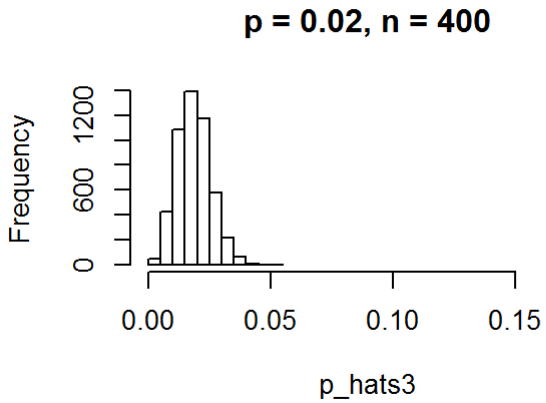
Exercise 10: Repeat the above simulation three more times but with modified sample sizes and proportions: for n=400n=400 and p=0.1p=0.1, n=1040n=1040 and p=0.02p=0.02, and n=400n=400 and p=0.02p=0.02. Plot all four histograms together by running the par(mfrow = c(2, 2)) command before creating the histograms. You may need to expand the plot window to accommodate the larger two-by-two plot. Describe the three new sampling distributions. Based on these limited plots, how does nn appear to affect the distribution of p^p? How does pp affect the sampling distribution?

If p stays the same, the smaller n will cause the distribution to be more spread out.However the distribution will still centered at p = 0.1. If we change p, we will shift the distribution along the x-axis.

```
set.seed(10)
p <- 0.1
n <- 400
p_hats1 <- rep(0, 5000)

for(i in 1:5000){
  samp1 <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
  p_hats1[i] <- sum(samp1 == "atheist")/n
}

p <- 0.02
n <- 1040
p_hats2 <- rep(0, 5000)

for(i in 1:5000){
  samp2 <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
  p_hats2[i] <- sum(samp2 == "atheist")/n
}

p <- 0.02
n <- 400
p_hats3 <- rep(0, 5000)

for(i in 1:5000){
  samp3 <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
  p_hats3[i] <- sum(samp3 == "atheist")/n
}

par(mfrow = c(2, 2))

hist(p_hats1, main = "p = 0.1, n = 400", xlim = c(0, 0.18))
hist(p_hats2, main = "p = 0.02, n = 1040", xlim = c(0, 0.18))
hist(p_hats3, main = "p = 0.02, n = 400", xlim = c(0, 0.18))
```







Exercise 11: If you refer to Table 6, you'll find that Australia has a sample proportion of 0.1 on a sample size of 1040, and that Ecuador has a sample proportion of 0.02 on 400 subjects. Let's suppose for this exercise that these point estimates are actually the truth. Then given the shape of their respective sampling distributions, do you think it is sensible to proceed with inference and report margin of errors, as the reports does? It is sensible to proceed with inference and report margin of errors for Australia. However, for Ecuador the date dose not meet the conditions for inference. Because np = 400 * 0.02 = 8 < 10

On your own The question of atheism was asked by WIN-Gallup International in a similar survey that was conducted in 2005. (We assume here that sample sizes have remained the same.) Table 4 on page 13 of the report summarizes survey results from 2005 and 2012 for 39 countries.

1. Answer the following two questions using the inference function. As always, write out the hypotheses for any tests you conduct and outline the status of the conditions for inference.

a. Is there convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012? Hint: Create a new data set for respondents from Spain. Form confidence intervals for the true proportion of athiests in both years, and determine whether they overlap.

H0: No change in Spain's atheism index HA: There is change

Pooled proportion = 0.0952 Standard error = 0.012 Test statistic: Z = 0.848 p-value = 0.3966

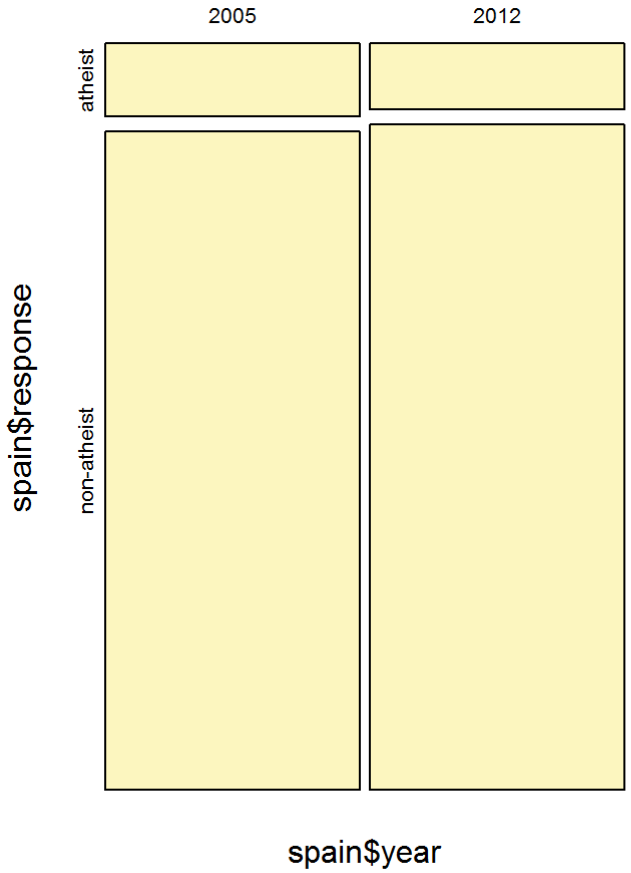
Since the p-value = 0.3966, which is greater than 0.05, therefore,we fail to reject the null hypothesis. The atheism index of Spain may or may not change from 2005-2012.

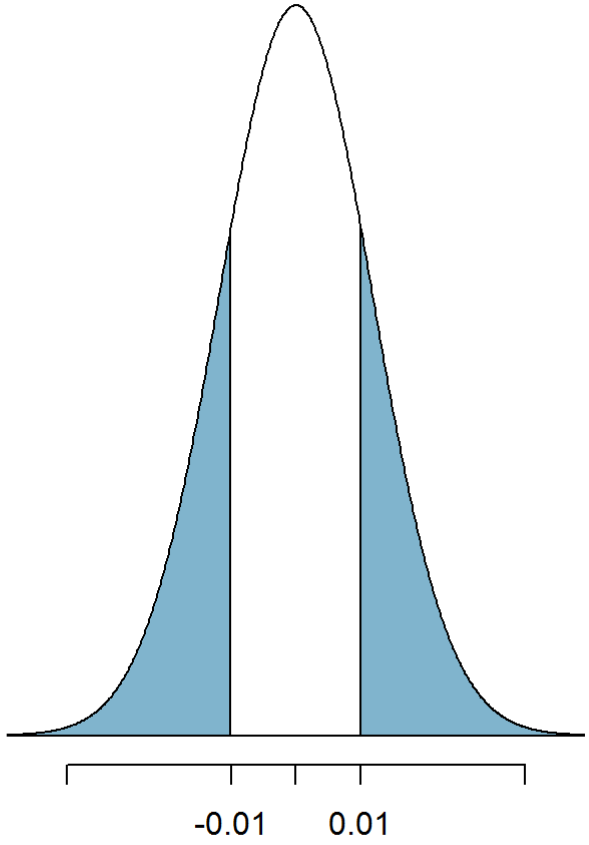
```
spain <- subset(atheism, atheism$nationality == "Spain")
inference(y = spain$response, x = spain$year, est = "proportion",type = "ht",null = 0, alternati
ve = "twosided", method = "theoretical", success = "atheist")
```

```
## Warning: Explanatory variable was numerical, it has been converted to
## categorical. In order to avoid this warning, first convert your explanatory
## variable to a categorical variable using the as.factor() function.
```

```
## Response variable: categorical, Explanatory variable: categorical
## Two categorical variables
## Difference between two proportions -- success: atheist
## Summary statistics:
##              x
## y             2005 2012  Sum
##   atheist      115  103  218
##   non-atheist 1031 1042 2073
##   Sum         1146 1145 2291
```

```
## Observed difference between proportions (2005-2012) = 0.0104
##
## H0: p_2005 - p_2012 = 0
## HA: p_2005 - p_2012 != 0
## Pooled proportion = 0.0952
## Check conditions:
##    2005 : number of expected successes = 109 ; number of expected failures = 1037
##    2012 : number of expected successes = 109 ; number of expected failures = 1036
## Standard error = 0.012
## Test statistic: Z =  0.848
## p-value =  0.3966
```





b. Is there convincing evidence that the United States has seen a change in its atheism index between 2005 and 2012?

H0: No change in US's atheism index HA: There is change Standard error = 0.008 Test statistic: Z = -5.243 p-value = 0

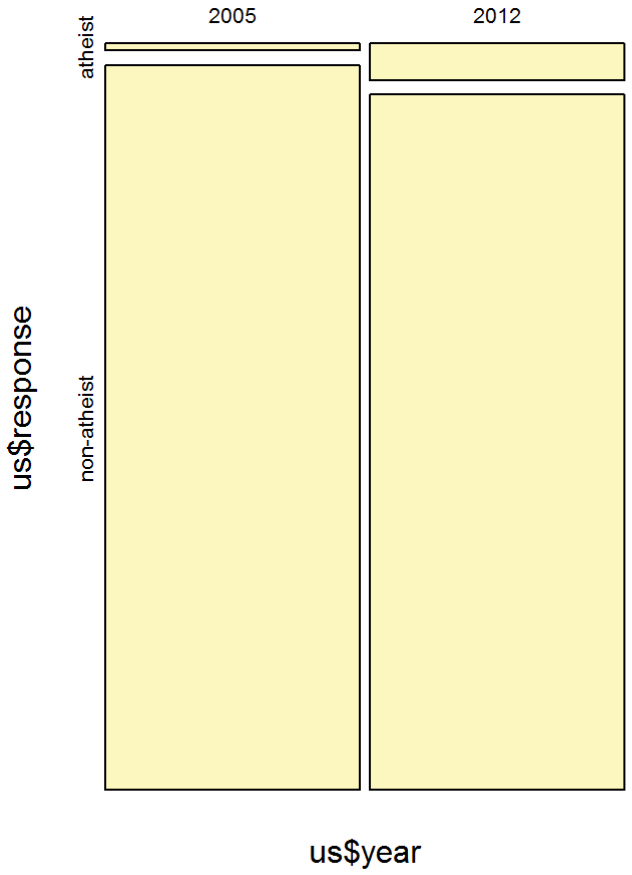
The p-value is 0, which is lower than 0.05. Therefore, we can reject the null hypothesis and claim that there is convincing evidence that the United States has seen a change in its athism index between 2005 and 2007

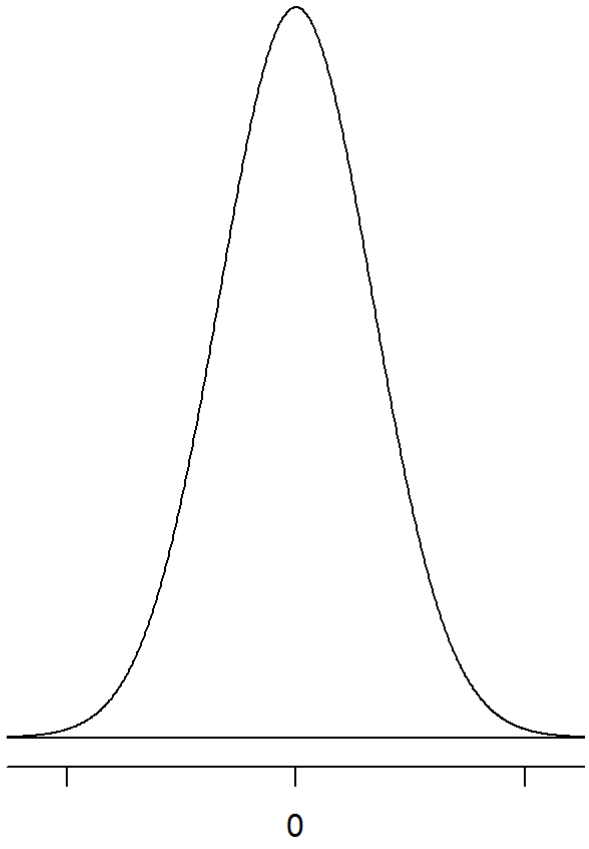
```
us <- subset(atheism, atheism$nationality == "United States")
inference(y = us$response, x = us$year, est = "proportion",type = "ht",null = 0, alternative =
"twosided", method = "theoretical", success = "atheist")
```

```
## Warning: Explanatory variable was numerical, it has been converted to
## categorical. In order to avoid this warning, first convert your explanatory
## variable to a categorical variable using the as.factor() function.
```

```
## Response variable: categorical, Explanatory variable: categorical
## Two categorical variables
## Difference between two proportions -- success: atheist
## Summary statistics:
##              x
## y             2005 2012  Sum
##   atheist       10   50   60
##   non-atheist  992  952 1944
##   Sum         1002 1002 2004
```

```
## Observed difference between proportions (2005-2012) = -0.0399
##
## H0: p_2005 - p_2012 = 0
## HA: p_2005 - p_2012 != 0
## Pooled proportion = 0.0299
## Check conditions:
##    2005 : number of expected successes = 30 ; number of expected failures = 972
##    2012 : number of expected successes = 30 ; number of expected failures = 972
## Standard error = 0.008
## Test statistic: Z =  -5.243
## p-value =  0
```






```
0.05*39
```

```
## [1] 1.95
```

2. If in fact there has been no change in the atheism index in the countries listed in Table 4, in how many of those countries would you expect to detect a change (at a significance level of 0.05) simply by chance? Hint: Look in the textbook index under Type 1 error.

Since the chance of causing the Type 1 error is simply the alpha, which is the significance level. Normally we set the alpha to be 0.05. Therefore the number of countries that we commit a Type I error will be 0.05 * 39 = 1.95 countries.

3. Suppose you're hired by the local government to estimate the proportion of residents that attend a religious service on a weekly basis. According to the guidelines, the estimate must have a margin of error no greater than 1% with 95% confidence. You have no idea what to expect for pp. How many people would you have to sample to ensure that you are within the guidelines? Hint: Refer to your plot of the relationship between pp and margin of error. Do not use the data set to answer this question.

According to the relationship plot between p and margin of error, the margin of error is highest when p = 0.5. And based on the question, we need to keep the margin of error is less than 0.01. If we do not know the value of p. In order to satified both conditions 100% of the time, we have to use p = 0.05. Once we use the formular: ME=1.96×SE=1.96× sqrt(p(1???p)/n), we should get 9604 people to recruit in our sample to ensure margin of error less than 0.01 and 95% confidence level.

```
z <- 1.96
p <- 0.5
margin_of_error <- 0.01

n <- 1.96^2 * p * (1-p) / (margin_of_error^2)
n
```

```
## [1] 9604
```