Inference for Numerical Data

Priya Shaji

3/26/2019

## Inference for categorical data

In August of 2012, news outlets ranging from the Washington Post to the Huffington Post ran a story about the rise of atheism in America. The source for the story was a poll that asked people, “Irrespective of whether you attend a place of worship or not, would you say you are a religious person, not a religious person or a convinced atheist?” This type of question, which asks people to classify themselves in one way or another, is common in polling and generates categorical data. In this lab we take a look at the atheism survey and explore what’s at play when making inference about population proportions using categorical data.

### Libraries used

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(plyr)

## -------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.  
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
## library(plyr); library(dplyr)

## -------------------------------------------------------------------------

##   
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

library(data.table)

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

library(knitr)  
library(DATA606)

## Loading required package: shiny

## Loading required package: openintro

## Please visit openintro.org for free statistics materials

##   
## Attaching package: 'openintro'

## The following objects are masked from 'package:datasets':  
##   
## cars, trees

## Loading required package: OIdata

## Loading required package: RCurl

## Loading required package: bitops

## Loading required package: maps

##   
## Attaching package: 'maps'

## The following object is masked from 'package:plyr':  
##   
## ozone

## Loading required package: ggplot2

##   
## Attaching package: 'ggplot2'

## The following object is masked from 'package:openintro':  
##   
## diamonds

## Loading required package: markdown

##   
## Welcome to CUNY DATA606 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='DATA606') will list the demos that are available.

##   
## Attaching package: 'DATA606'

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

### The survey

To access the press release for the poll, conducted by WIN-Gallup International, click on the following link:

<http://www.wingia.com/web/files/richeditor/filemanager/Global_INDEX_of_Religiosity_and_Atheism_PR__6.pdf>

### The data

Turn your attention to Table 6 (pages 15 and 16), which reports the sample size and response percentages for all 57 countries. While this is a useful format to summarize the data, we will base our analysis on the original data set of individual responses to the survey. Load this data set into R with the following command.

#download.file("http://www.openintro.org/stat/data/atheism.RData", destfile = "atheism.RData")  
load("/Users/priyashaji/Documents/cuny msds/Spring'19/data 606/Labs/Lab6/atheism.RData")

To investigate the link between these two ways of organizing this data, take a look at the estimated proportion of atheists in the United States. Towards the bottom of Table 6, we see that this is 5%. We should be able to come to the same number using the atheism data.

### Inference on proportions

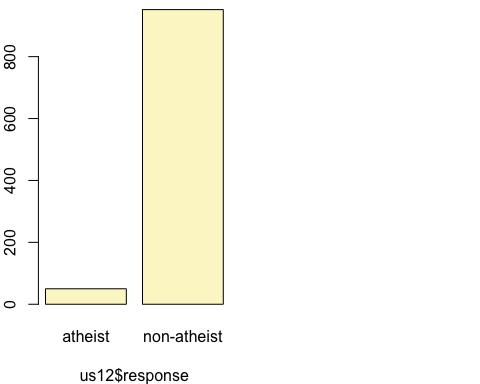
As was hinted at in Exercise 1, Table 6 provides statistics, that is, calculations made from the sample of 51,927 people. What we’d like, though, is insight into the population parameters. You answer the question, “What proportion of people in your sample reported being atheists?” with a statistic; while the question “What proportion of people on earth would report being atheists” is answered with an estimate of the parameter.

The inferential tools for estimating population proportion are analogous to those used for means in the last chapter: the confidence interval and the hypothesis test.

If the conditions for inference are reasonable, we can either calculate the standard error and construct the interval by hand, or allow the inference function to do it for us.

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

Note that since the goal is to construct an interval estimate for a proportion, it’s necessary to specify what constitutes a “success”, which here is a response of “atheist”.

Although formal confidence intervals and hypothesis tests don’t show up in the report, suggestions of inference appear at the bottom of page 7: “In general, the error margin for surveys of this kind is ± 3-5% at 95% confidence”.

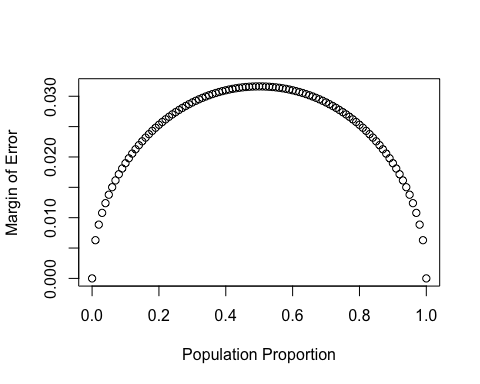
### How does the proportion affect the margin of error?

Imagine you’ve set out to survey 1000 people on two questions: are you female? and are you left-handed? Since both of these sample proportions were calculated from the same sample size, they should have the same margin of error, right? Wrong! While the margin of error does change with sample size, it is also affected by the proportion.

Think back to the formula for the standard error: SE=p(1−p)/√‾‾n . This is then used in the formula for the margin of error for a 95% confidence interval: ME=1.96×SE=1.96×p(1−p)/√‾‾n . Since the population proportion p is in this ME formula, it should make sense that the margin of error is in some way dependent on the population proportion. We can visualize this relationship by creating a plot of ME vs. p .

The first step is to make a vector p that is a sequence from 0 to 1 with each number separated by 0.01. We can then create a vector of the margin of error (me) associated with each of these values of p using the familiar approximate formula (ME=2×SE ). Lastly, we plot the two vectors against each other to reveal their relationship.

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")



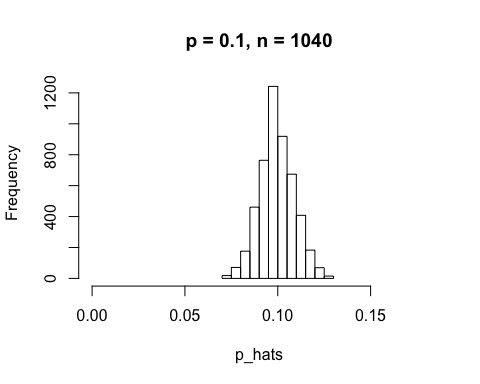
### Success-failure condition

The textbook emphasizes that you must always check conditions before making inference. For inference on proportions, the sample proportion can be assumed to be nearly normal if it is based upon a random sample of independent observations and if both np≥10 and n(1−p)≥10 . This rule of thumb is easy enough to follow, but it makes one wonder: what’s so special about the number 10?

The short answer is: nothing. You could argue that we would be fine with 9 or that we really should be using 11. What is the “best” value for such a rule of thumb is, at least to some degree, arbitrary. However, when np and n(1−p) reaches 10 the sampling distribution is sufficiently normal to use confidence intervals and hypothesis tests that are based on that approximation.

We can investigate the interplay between n and p and the shape of the sampling distribution by using simulations. To start off, we simulate the process of drawing 5000 samples of size 1040 from a population with a true atheist proportion of 0.1. For each of the 5000 samples we compute p̂ and then plot a histogram to visualize their distribution.

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))



These commands build up the sampling distribution of p̂ using the familiar for loop. You can read the sampling procedure for the first line of code inside the for loop as, “take a sample of size n with replacement from the choices of atheist and non-atheist with probabilities p and 1−p , respectively.” The second line in the loop says, “calculate the proportion of atheists in this sample and record this value.” The loop allows us to repeat this process 5,000 times to build a good representation of the sampling distribution.

### Exercises

#### 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?

Answer 1)

These appear to be sample statistics since the numbers are calculated from the sample of 50,000 people, and not directly from the human population of 7.55 billion people.

#### 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?

Answer 2)

The sample would have to be randomly selected from people across the Earth. They claim that they surveyed 50,000 people from 57 Countries on 5 continents. There are 233 countries in the world according to <https://en.wikipedia.org/wiki/List_of_countries_by_population_(United_Nations)>, but the majority of the currently 7.55 billion people are concentrated in the 13 most populous countries. Cases also need to be independent, so you need to sample less than 10% of the population, which is also true. We also need at least 10 ‘successes’ and ‘failures’, in this case ‘atheists’ and ‘non-atheists’, this condition is also true. It seems reasonable to generalize these data to the global human population.

#### Exercise 3

What does each row of Table 6 correspond to? What does each row of atheism correspond to?

Answer 3)

Each row in table 6 is a country where the survey was conducted. Each row in the ‘atheism’ table is a person who was interviewed for the survey.

head(atheism)

## nationality response year  
## 1 Afghanistan non-atheist 2012  
## 2 Afghanistan non-atheist 2012  
## 3 Afghanistan non-atheist 2012  
## 4 Afghanistan non-atheist 2012  
## 5 Afghanistan non-atheist 2012  
## 6 Afghanistan non-atheist 2012

#### 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?

Answer 4)

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

us12\_count <- count(us12$response == 'atheist')  
us12\_count

## x freq  
## 1 FALSE 952  
## 2 TRUE 50

names(us12\_count) <- c("atheist", "total")  
us12\_count$percent <- us12\_count$total / sum(us12\_count$total) \* 100  
kable(us12\_count)

|  |  |  |
| --- | --- | --- |
| atheist | total | percent |
| FALSE | 952 | 95.00998 |
| TRUE | 50 | 4.99002 |

This rounds to 5%, which is consistent with table 6.

#### 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?

Answer 5)

Cases are independent is true, 51,927 is less than 10% the world population. We also have at least 10 ‘successes’ and 10 ‘failures’, that is atheists and non-atheists or visa-versa. The conditions for inference have been met.

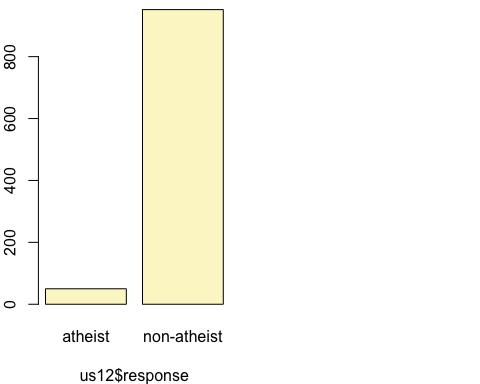
#### Exercise 6

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

Answer 6)

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 )

z <- 1.96  
SE <- 0.0069   
ME <- z \* SE  
#The margin of error for US is  
ME

## [1] 0.013524

1.35% margin of error.

#### 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.

Answer 7)

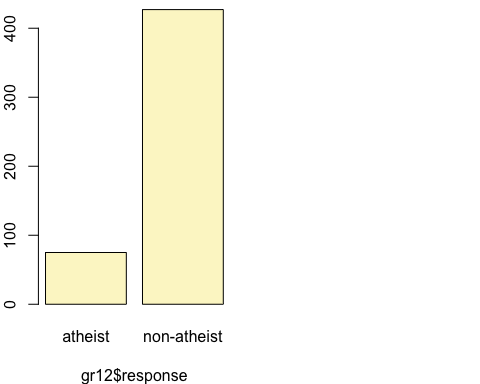
1. With Germany:

gr12 <- subset(atheism, nationality == "Germany" & year == "2012")   
gr12\_count <- count(gr12$response == 'atheist')  
names(gr12\_count) <- c("atheist", "total")  
gr12\_count$percent <- gr12\_count$total / sum(gr12\_count$total) \* 100  
kable(gr12\_count)

|  |  |  |
| --- | --- | --- |
| atheist | total | percent |
| FALSE | 427 | 85.05976 |
| TRUE | 75 | 14.94024 |

inference(gr12$response, est = "proportion", type = "ci", method = "theoretical",   
 success = "atheist")

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



## p\_hat = 0.1494 ; n = 502   
## Check conditions: number of successes = 75 ; number of failures = 427   
## Standard error = 0.0159   
## 95 % Confidence interval = ( 0.1182 , 0.1806 )

The sample size is small enough for the cases to be independent, and the success-failure criteria is met.

z value for 95% confidence will be z=1.96 From above we got Standard error: SE = 0.0159

z <- 1.96  
SE <- 0.0159   
ME <- z \* SE  
#The margin of error for Germany is  
ME

## [1] 0.031164

3.11% margin of error.

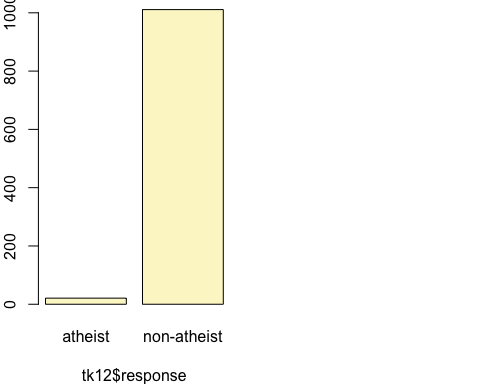
With Turkey:

tk12 <- subset(atheism, nationality == "Turkey" & year == "2012")  
ptk12\_count <- count(tk12$response == 'atheist')  
names(ptk12\_count) <- c("atheist", "total")  
ptk12\_count$percent <- ptk12\_count$total / sum(ptk12\_count$total) \* 100  
kable(ptk12\_count)

|  |  |  |
| --- | --- | --- |
| atheist | total | percent |
| FALSE | 1011 | 97.965116 |
| TRUE | 21 | 2.034884 |

inference(tk12$response, est = "proportion", type = "ci", method = "theoretical",   
 success = "atheist")

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



## p\_hat = 0.0203 ; n = 1032   
## Check conditions: number of successes = 21 ; number of failures = 1011   
## Standard error = 0.0044   
## 95 % Confidence interval = ( 0.0117 , 0.029 )

The sample size is small enough for the cases to be independent, and the success-failure criteria is met.

z value for 95% confidence will be z=1.96 From above we got Standard error: SE = 0.0044

z <- 1.96  
SE <- 0.0044   
ME <- z \* SE  
#The margin of error for Turkey is  
ME

## [1] 0.008624

#### Exercise 8

Describe the relationship between p and me.

Answer 8)

Based on the graph below the proportion of 0.50 is the proportion with the largest margin of error possible. the plot follows a bell curve. When the proportions move away from 0.50 and get closer to the extremes of 0.00 or 1.00, the margin of error decreases. There is an inverse Correlation between p and me as they move in opposite directions.

#### 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.

Answer 9)

summary(p\_hats)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.07019 0.09327 0.09904 0.09969 0.10577 0.12981

IQR(p\_hats)

## [1] 0.0125

sd(p\_hats)

## [1] 0.009287382

From this code we can determine two different measures of center and spread, median: 0.09904, IQR: 0.0125 mean: 0.09969, sd: 0.009287382

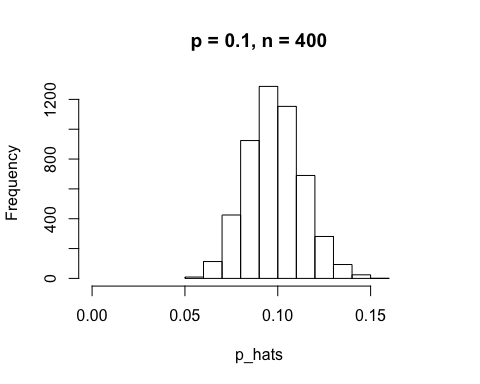
The shape of the distribution is bell-shaped.

#### Exercise 10

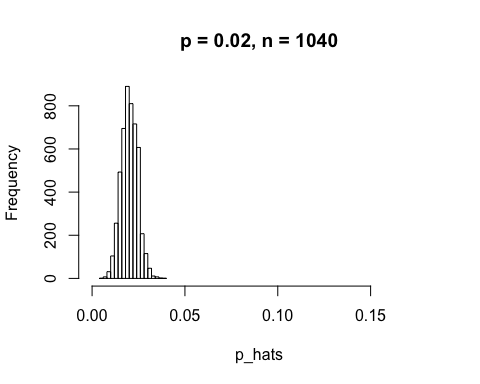
Repeat the above simulation three more times but with modified sample sizes and proportions: for n=400 and p=0.1 , n=1040 and p=0.02 , and n=400 and p=0.02 . Plot all four histograms together by running the par(mfrow = c(2, 2)) command before creating the histograms. You may need to expand the plot window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

Answer 10)

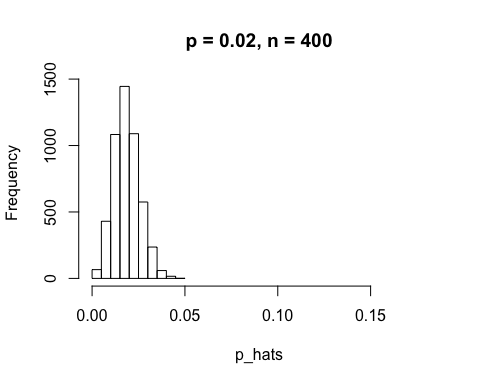
p <- 0.1  
n <- 400  
p\_hats <- rep(0, 5000)  
  
for(i in 1:5000){  
 samp <- sample(c("atheist", "non\_atheist"), n, replace = TRUE, prob = c(p, 1-p))  
 p\_hats[i] <- sum(samp == "atheist")/n  
}  
  
hist(p\_hats, main = "p = 0.1, n = 400", xlim = c(0, 0.18))



p <- 0.02  
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 = 0.02, n = 1040", xlim = c(0, 0.18))



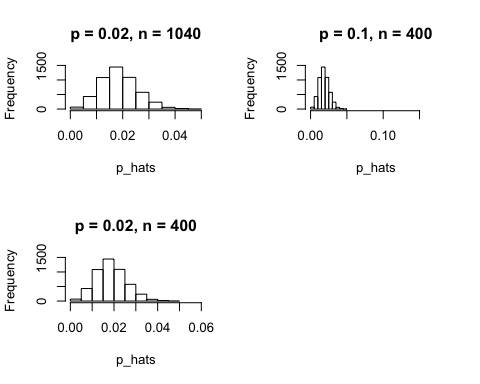
p <- 0.02  
n <- 400  
p\_hats <- rep(0, 5000)  
  
for(i in 1:5000){  
 samp <- sample(c("atheist", "non\_atheist"), n, replace = TRUE, prob = c(p, 1-p))  
 p\_hats[i] <- sum(samp == "atheist")/n  
}  
  
hist(p\_hats, main = "p = 0.02, n = 400", xlim = c(0, 0.18))



Decreasing p or increasing n makes the distribution more narrow.

Once you’re done, you can reset the layout of the plotting window by using the command par(mfrow = c(1, 1)) command or clicking on “Clear All” above the plotting window (if using RStudio). Note that the latter will get rid of all your previous plots.

par(mfrow = c(2,2))  
hist(p\_hats, main = "p = 0.02, n = 1040", xlim = c(0.0001, 0.05))  
hist(p\_hats, main = "p = 0.1, n = 400", xlim = c(0, 0.18))  
hist(p\_hats, main = "p = 0.02, n = 400", xlim = c(0.0001, 0.06))



#### 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?

Answer 11)

Based on Australia’s conditions are TRUE for n \* p and TRUE n \* ( 1 - p). Based on Ecuador’s conditions are FALSE for n \* p and TRUE n \* ( 1 - p).

I would feel comfortable proceeding with the inference and me for Australia as the data met the conditions for the sampling distribution of p. However, Ecuador data did not meet the success-failure condition as np is only 8, which is smaller than the 10 needed.

# Australia  
  
n\_au <- 1040  
p\_au <- 0.1  
  
cond\_au <- c(n\_au \* p\_au >= 10, n\_au \* (1 - p\_au) >= 10)  
cond\_au

## [1] TRUE TRUE

# Ecuador  
  
n\_ecu <- 400  
p\_ecu <- 0.02  
  
cond\_ecu <- c(n\_ecu \* p\_ecu >= 10, n\_ecu \* (1 - p\_ecu) >= 10)  
cond\_ecu

## [1] FALSE TRUE

### 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.

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.

#### Ques\_1

1. 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.

Answer a)

H0: There is no convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012 p2005 = p2012 HA: There is convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012 p2005 is not equal to p2012

Spain in 2005:

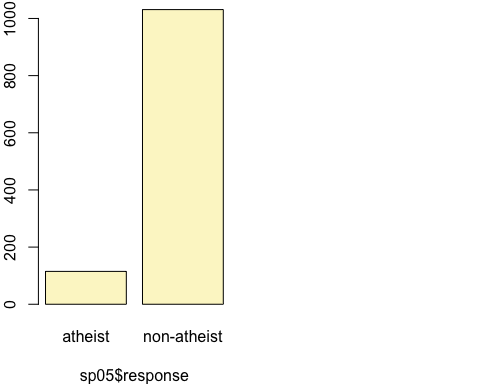
sp05 <- subset(atheism, nationality == "Spain" & year == "2005")   
sp05\_count <- count(sp05$response == 'atheist')  
names(sp05\_count) <- c("atheist", "total")  
sp05\_count$percent <- sp05\_count$total / sum(sp05\_count$total) \* 100  
kable(sp05\_count)

|  |  |  |
| --- | --- | --- |
| atheist | total | percent |
| FALSE | 1031 | 89.9651 |
| TRUE | 115 | 10.0349 |

Inference for spain in 2005

inference(sp05$response, est = "proportion", type = "ci", method = "theoretical",   
 success = "atheist")

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



## p\_hat = 0.1003 ; n = 1146   
## Check conditions: number of successes = 115 ; number of failures = 1031   
## Standard error = 0.0089   
## 95 % Confidence interval = ( 0.083 , 0.1177 )

The sample size is small enough for the cases to be independent, and the success-failure criteria is met.

z value for 95% confidence will be z=1.96 From above we got Standard error: SE = 0.0089

z <- 1.96  
SE <- 0.0089   
ME <- z \* SE  
#The margin of error for Spain in 2005 is  
ME

## [1] 0.017444

1.74% margin of error.

Spain in 2012:

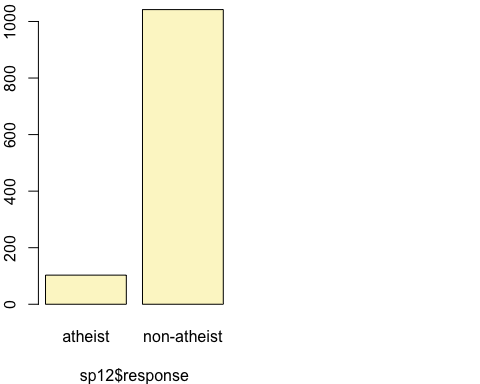
sp12 <- subset(atheism, nationality == "Spain" & year == "2012")   
sp12\_count <- count(sp12$response == 'atheist')  
names(sp12\_count) <- c("atheist", "total")  
sp12\_count$percent <- sp12\_count$total / sum(sp12\_count$total) \* 100  
kable(sp12\_count)

|  |  |  |
| --- | --- | --- |
| atheist | total | percent |
| FALSE | 1042 | 91.004367 |
| TRUE | 103 | 8.995633 |

Inference for spain in 2012

inference(sp12$response, est = "proportion", type = "ci", method = "theoretical",   
 success = "atheist")

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



## p\_hat = 0.09 ; n = 1145   
## Check conditions: number of successes = 103 ; number of failures = 1042   
## Standard error = 0.0085   
## 95 % Confidence interval = ( 0.0734 , 0.1065 )

The sample size is small enough for the cases to be independent, and the success-failure criteria is met.

z value for 95% confidence will be z=1.96 From above we got Standard error: SE = 0.0085

z <- 1.96  
SE <- 0.0085   
ME <- z \* SE  
#The margin of error for Spain in 2012 is  
ME

## [1] 0.01666

1.66% margin of error.

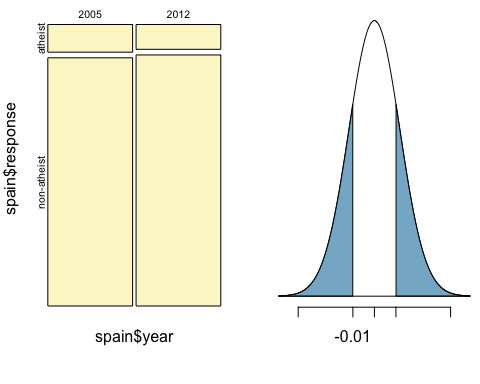
The 95% confidence intervals for both years do overlap for Spain. Therefore there is not convincing evidence that the proportion of Atheists have changed between 2005 and 2012.

Hypothesis test

spain <- subset(atheism, nationality == "Spain" & year == "2005" | nationality == "Spain" & year == "2012")  
  
suppressWarnings(inference(y = spain$response, x = spain$year, est = "proportion",type = "ht", null = 0, alternative = "twosided", method = "theoretical", success = "atheist"))

## 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



The returned p-value is more than 0.05; because of that it is fail to reject our NULL hypothesis in favor of the alternative HA hypothesis; that is: There is No convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012.

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

Answer b)

The observed p-value is 0, which is less than the significance level 0.05. Therefore, we reject the null hypothesis that there is no convincing evidence that US has seen a change in its atheism index between 2005 and 2012.

H0: There is no convincing evidence that US has seen a change in its atheism index between 2005 and 2012 p2005 = p2012 HA: There is convincing evidence that US has seen a change in its atheism index between 2005 and 2012 p2005 is not equal to p2012

US in 2005:

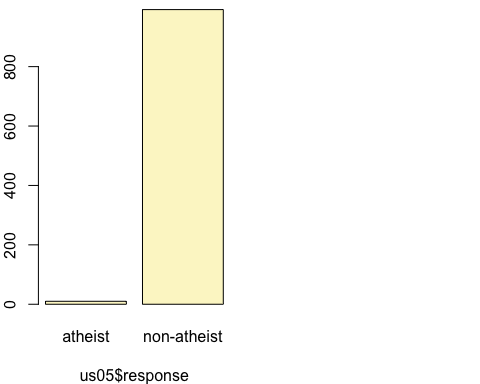
us05 <- subset(atheism, nationality == "United States" & year == "2005")   
us05\_count <- count(us05$response == 'atheist')  
names(us05\_count) <- c("atheist", "total")  
us05\_count$percent <- us05\_count$total / sum(us05\_count$total) \* 100  
kable(us05\_count)

|  |  |  |
| --- | --- | --- |
| atheist | total | percent |
| FALSE | 992 | 99.001996 |
| TRUE | 10 | 0.998004 |

Inference for spain in 2005

inference(us05$response, est = "proportion", type = "ci", method = "theoretical",   
 success = "atheist")

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



## p\_hat = 0.01 ; n = 1002   
## Check conditions: number of successes = 10 ; number of failures = 992   
## Standard error = 0.0031   
## 95 % Confidence interval = ( 0.0038 , 0.0161 )

The sample size is small enough for the cases to be independent, and the success-failure criteria is met.

z value for 95% confidence will be z=1.96 From above we got Standard error: SE = 0.0031

z <- 1.96  
SE <- 0.0031   
ME <- z \* SE  
#The margin of error for US in 2005 is  
ME

## [1] 0.006076

US in 2012:

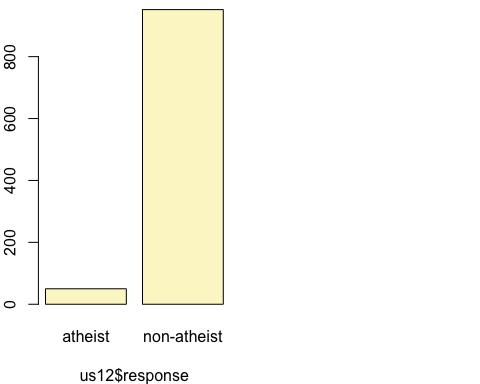
us12 <- subset(atheism, nationality == "United States" & year == "2012")   
us12\_count <- count(us12$response == 'atheist')  
names(us12\_count) <- c("atheist", "total")  
us12\_count$percent <- us12\_count$total / sum(us12\_count$total) \* 100  
kable(us12\_count)

|  |  |  |
| --- | --- | --- |
| atheist | total | percent |
| FALSE | 952 | 95.00998 |
| TRUE | 50 | 4.99002 |

Inference for spain in 2012

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 )

The sample size is small enough for the cases to be independent, and the success-failure criteria is met.

z value for 95% confidence will be z=1.96 From above we got Standard error: SE = 0.0069

z <- 1.96  
SE <- 0.0069   
ME <- z \* SE  
#The margin of error for US in 2012 is  
ME

## [1] 0.013524

1.35% margin of error.

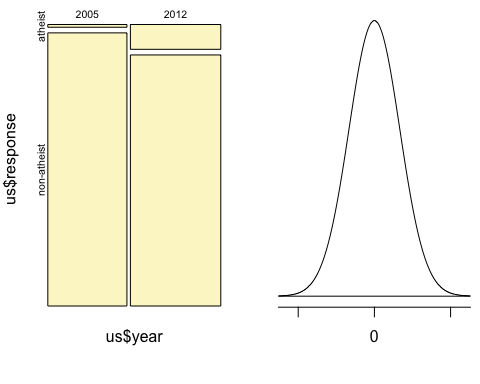
The 95% confidence intervals for both years do overlap for Spain. Therefore there is not convincing evidence that the proportion of Atheists have changed between 2005 and 2012.

Hypothesis test

us <- subset(atheism, nationality == "United States" & year == "2005" | nationality == "Spain" & year == "2012")  
  
suppressWarnings(inference(y = us$response, x = us$year, est = "proportion",type = "ht", null = 0, alternative = "twosided", method = "theoretical", success = "atheist"))

## Response variable: categorical, Explanatory variable: categorical  
## Two categorical variables  
## Difference between two proportions -- success: atheist  
## Summary statistics:  
## x  
## y 2005 2012 Sum  
## atheist 10 103 113  
## non-atheist 992 1042 2034  
## Sum 1002 1145 2147

## Observed difference between proportions (2005-2012) = -0.08  
##   
## H0: p\_2005 - p\_2012 = 0   
## HA: p\_2005 - p\_2012 != 0   
## Pooled proportion = 0.0526   
## Check conditions:  
## 2005 : number of expected successes = 53 ; number of expected failures = 949   
## 2012 : number of expected successes = 60 ; number of expected failures = 1085   
## Standard error = 0.01   
## Test statistic: Z = -8.279   
## p-value = 0



The observed p-value is 0, which is less than the significance level 0.05. Therefore, we reject the null hypothesis that there is no convincing evidence that US has seen a change in its atheism index between 2005 and 2012.

#### Ques\_2

1. 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.

Answer 2)

A type 1 error is rejecting the null hypothesis when H0 is actually true. We typically do not want to incorrectly reject H0 more than 5% of the time. This corresponds to a significance level of 0.05. Since there are 39 countries in Table 4 that summarizes survey results from 2005 to 2012 we will need to multiply 0.05 by 39 to estimate how many countries we would expect to detect a change in the atheism index simply by chance. the result is 1.95, or 2 countries.

#### Ques\_3

1. 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 p. 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 p and margin of error. Do not use the data set to answer this question.

Answer 3)

There are two unknown variablies in this question: p and n. When we do not have and estimate for p we follow the guideline that the margin of error is largest when p is 0.5. So we typically use this worst case estimate if no other estimate is available. The estimate must have a margin of error no greater than 1%. We use the formula ME = zSE = 1,96 sqrt(p(1-p)/n)) <=0.01. Based on the calculation we would need at least 9604 participants to ensure the sample proportion is within 0.01 of the true porportion with 95% confidence.

P <-0.5  
Z.alpha <-1.96  
ME <-0.01  
N<- Z.alpha^2\*P\*(1-P)/ME^2  
N

## [1] 9604

And based on the results we will need at least 9604 participants to ensure the sample proportion is within 0.01 of the true proportion with 95% confidence.