# DATA606_Lab6_RJM

## RJM

## 2019-12-29

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.

## The survey

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

*https://github.com/jbryer/DATA606/blob/master/inst/labs/Lab6/more/Global_INDEX_of_Religiosity_ and_Atheism_PR___6.pdf*

Take a moment to review the report then address the following questions.

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*? The percentages are sample statistics as the second paragraph confirms that they are derived from interviews with more than 50,000 men and women from 57 countries across the globe in five continents.

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 assumptions would be that the samples were random, observations were independent, and the sample size was reasonably large.

## 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("atheism.RData")
head(atheism, 5)
```

```
##   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
```

3. What does each row of Table 6 correspond to? What does each row of `atheism` correspond to?

Each row of the Table 6 corresponds to the sample size in a country and a summary of responses depicted as the percentage of people who are religious, not religious, atheist, or not responded. The rows in the 'atheism' correspond to indiviudal responses on the nationality, actualresponse, and the year of the survey.

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.

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? Yes, I agree as the percentage of the respondents claiming to be atheist comes to be 0.049 which is 5% if rounded up.

```
us12 <- subset(atheism, nationality == "United States" & year == "2012")
percent_atheist_us <- sum(us12$response=='atheist') / length(us12$response=='atheist')
head(us12)
```

```
##           nationality      response year
## 49926 United States non-atheist 2012
## 49927 United States non-atheist 2012
## 49928 United States non-atheist 2012
## 49929 United States non-atheist 2012
## 49930 United States non-atheist 2012
## 49931 United States non-atheist 2012
```

```
percent_atheist_us
```

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

### 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.

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? The conditions are normal distribution, random samples, and independent observations. A sample of 1002 from US's population could be easily inferred to be random and independent by the same logic as it is less than 10% of the population. There are at least 10 successful and failed observations for each of the category of atheist and non-atheist helping us to assume normal distribution.

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.

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

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



us12$response
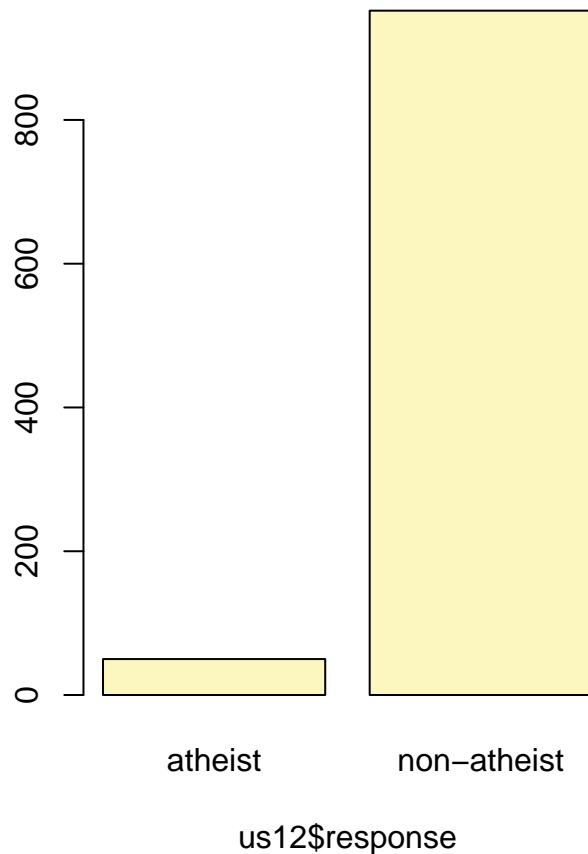
```
## p_hat = 0.0499 ;   n = 1002
## Check conditions: number of successes = 50 ; number of failures = 952
## Standard error = 0.0069
## 95 % Confidence interval = ( 0.0364 , 0.0634 )
```

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".

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?

```
se <-  0.0069 #standard error
z_score <- qnorm(0.975) # based on 95% CI
me <- se * z_score # margin of error
me # margin of error is 0.0135.
```

## [1] 0.01352375

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.

```
brazil12 <- subset(atheism, nationality == 'Brazil' & year == "2012")
head(brazil12,10)
```

```
##      nationality    response year
## 6595      Brazil non-atheist 2012
## 6596      Brazil non-atheist 2012
## 6597      Brazil non-atheist 2012
## 6598      Brazil non-atheist 2012
## 6599      Brazil non-atheist 2012
## 6600      Brazil non-atheist 2012
## 6601      Brazil non-atheist 2012
## 6602      Brazil non-atheist 2012
## 6603      Brazil non-atheist 2012
## 6604      Brazil non-atheist 2012
```

```
inference(brazil12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

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

brazil12$response
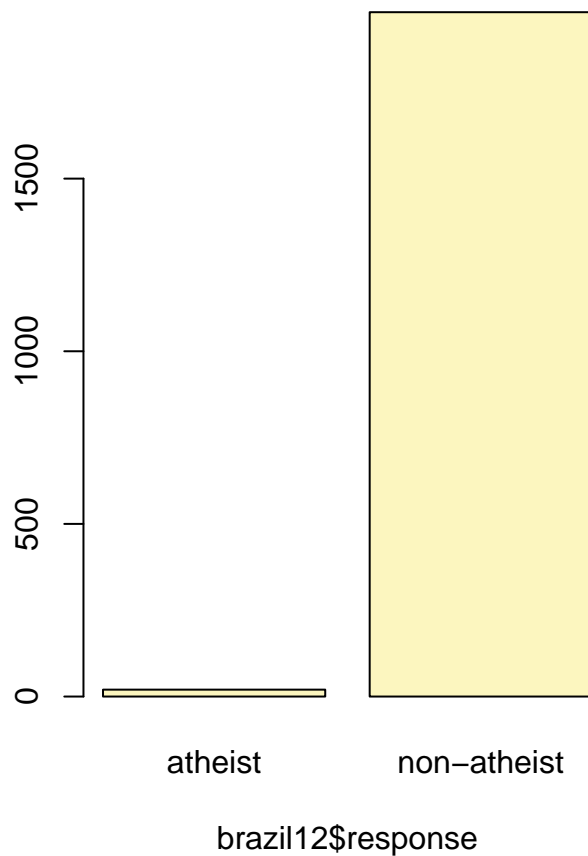
```
## p_hat = 0.01 ;  n = 2002
## Check conditions: number of successes = 20 ; number of failures = 1982
## Standard error = 0.0022
## 95 % Confidence interval = ( 0.0056 , 0.0143 )
```

```
se_brazil <-  0.0022 #standard error
z_score <- qnorm(0.975) # based on 95% CI
z_score
```

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

```
me_brazil <- se_brazil * z_score # margin of error
me_brazil # margin of error is 0.0043.
```

```
## [1] 0.004311921
```

```
# all the conditions of inference were met.
```

```
moldova12 <- subset(atheism, nationality == 'Moldova' & year == "2012")
head(brazil12,10)
```

```
##        nationality    response year
## 6595        Brazil non-atheist 2012
## 6596        Brazil non-atheist 2012
```

```
## 6597       Brazil non-atheist 2012
## 6598       Brazil non-atheist 2012
## 6599       Brazil non-atheist 2012
## 6600       Brazil non-atheist 2012
## 6601       Brazil non-atheist 2012
## 6602       Brazil non-atheist 2012
## 6603       Brazil non-atheist 2012
## 6604       Brazil non-atheist 2012
```

```r
inference(moldova12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

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



moldova12$response
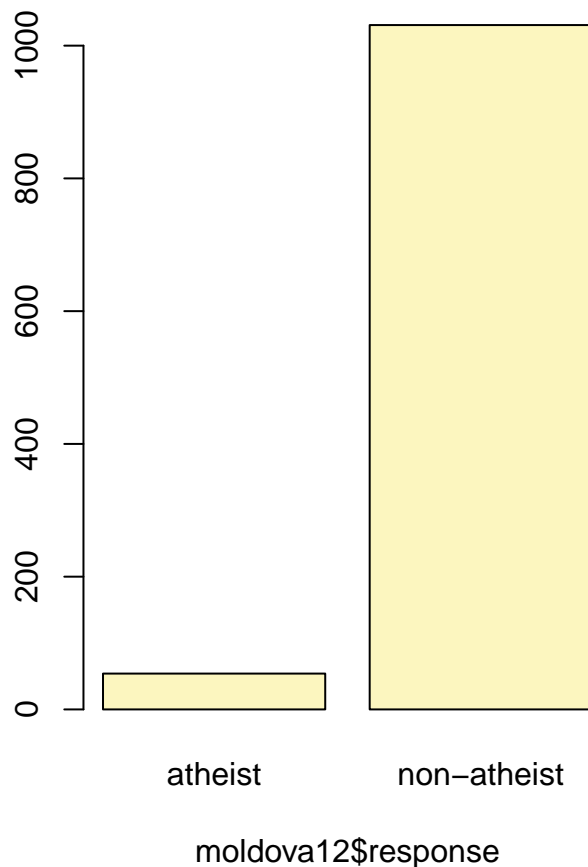
```
## p_hat = 0.0498 ;  n = 1085
## Check conditions: number of successes = 54 ; number of failures = 1031
## Standard error = 0.0066
## 95 % Confidence interval = ( 0.0368 , 0.0627 )
```

```r
se_moldova <-  0.0066 #standard error
z_score <- qnorm(0.975) # based on 95% CI
z_score
```

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

```
me_moldova <- se_moldova * z_score # margin of error
me_moldova # margin of error is 0.0129.
```

```
## [1] 0.01293576
```

```
# all the conditions of inference were met.
```

### 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: $SE = \sqrt{p(1-p)/n}$. This is then used in the formula for the margin of error for a 95% confidence interval: $ME = 1.96 \times SE = 1.96 \times \sqrt{p(1-p)/n}$. 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 ($ME = 2 \times SE$). 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")
```
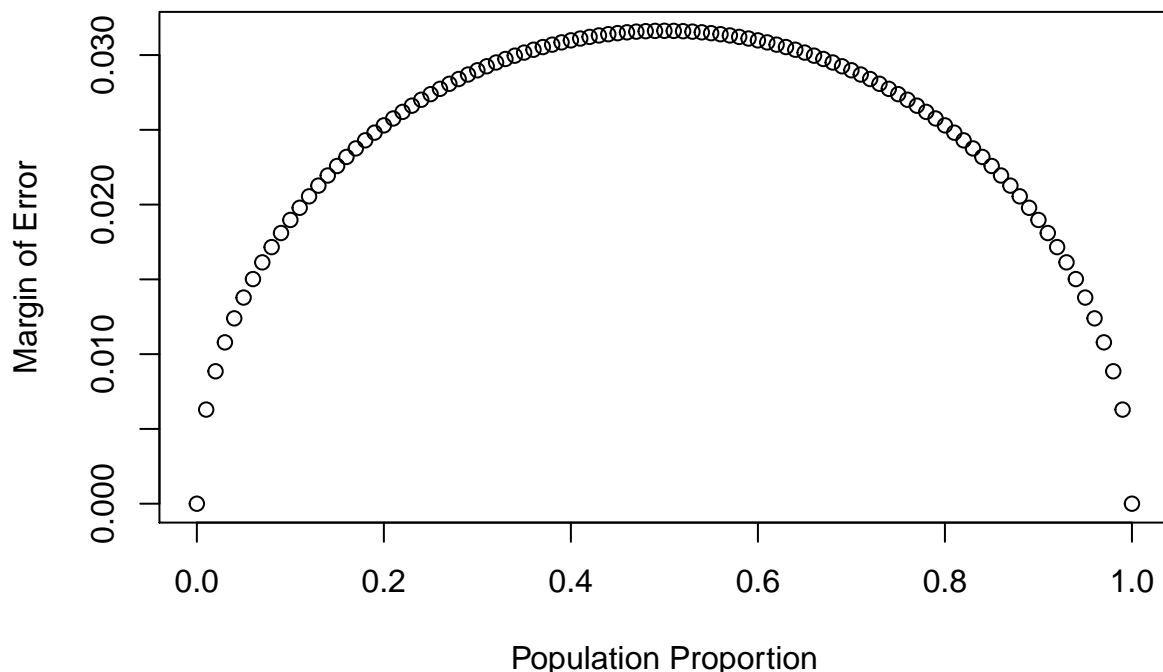


8. Describe the relationship between `p` and `me`. The me increased till the p reach 0.5 where it peaked to a little over 0.03. It started to decreased from there till it reached 0 at 1.0 p.

## 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 $np \geq 10$ and $n(1-p) \geq 10$. 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 $\hat{p}$ 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))
```
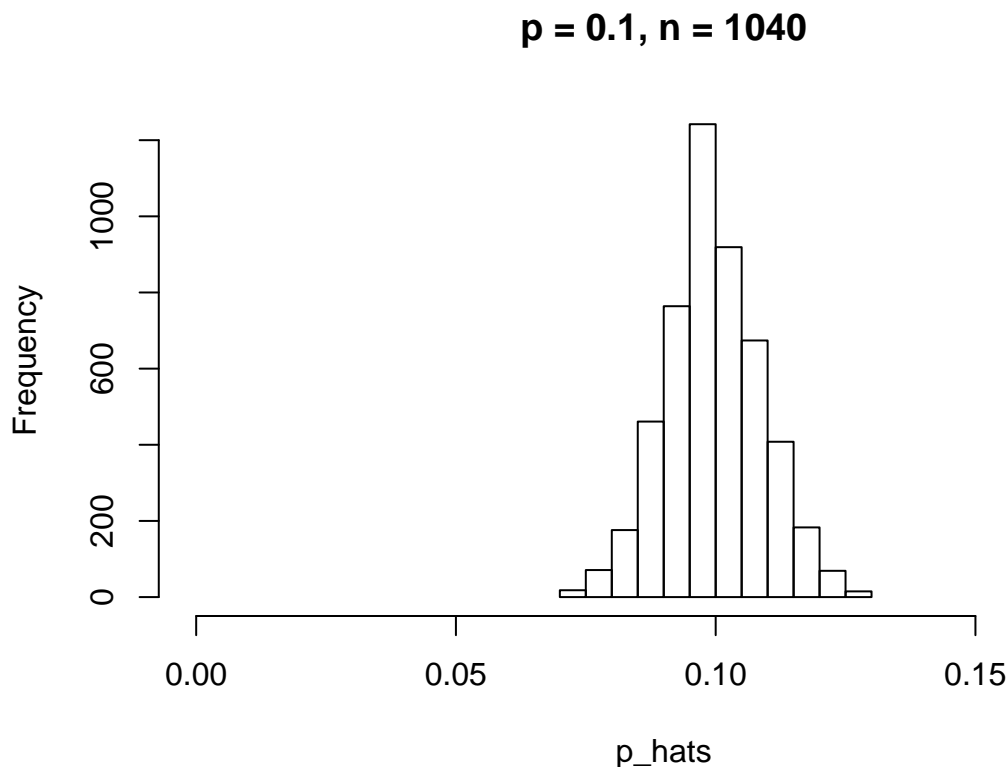


These commands build up the sampling distribution of $\hat{p}$ using the familiar `for` loop. You can read the sampling procedure for the first line of code inside the `for` loop as, "take a sample of size $n$ with replacement

8

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.

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.

```
library(Hmisc)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'lattice'
```

```
## The following object is masked from 'package:openintro':
##
##     lsegments
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survival'
```

```
## The following object is masked from 'package:openintro':
##
##     transplant
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

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

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:plyr':
##
##     is.discrete, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
describe(p_hats)
```

```
## p_hats
##         n  missing distinct     Info     Mean      Gmd      .05      .10
##      5000        0       62    0.999  0.09969  0.01049  0.08462  0.08750
##       .25      .50      .75      .90      .95
##   0.09327  0.09904  0.10577  0.11154  0.11538
##
## lowest : 0.07019231 0.07115385 0.07211538 0.07307692 0.07403846
## highest: 0.12500000 0.12596154 0.12692308 0.12788462 0.12980769
```

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 $\hat{p}$? How does $p$ affect the sampling distribution?

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.

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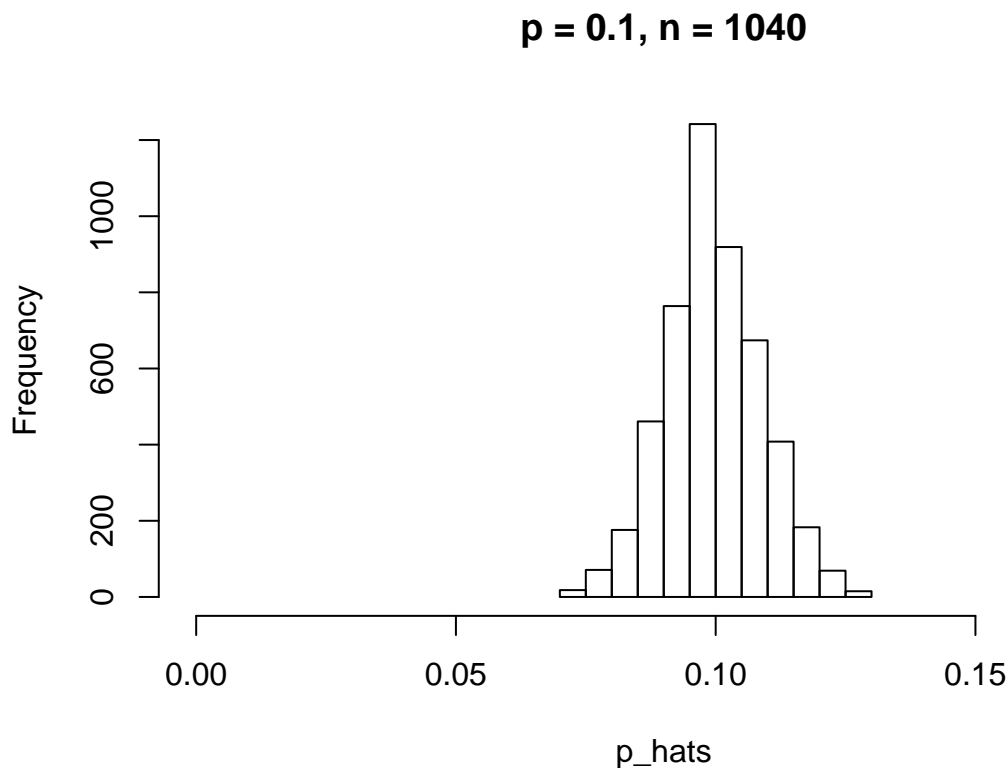
**p = 0.1, n = 1040**

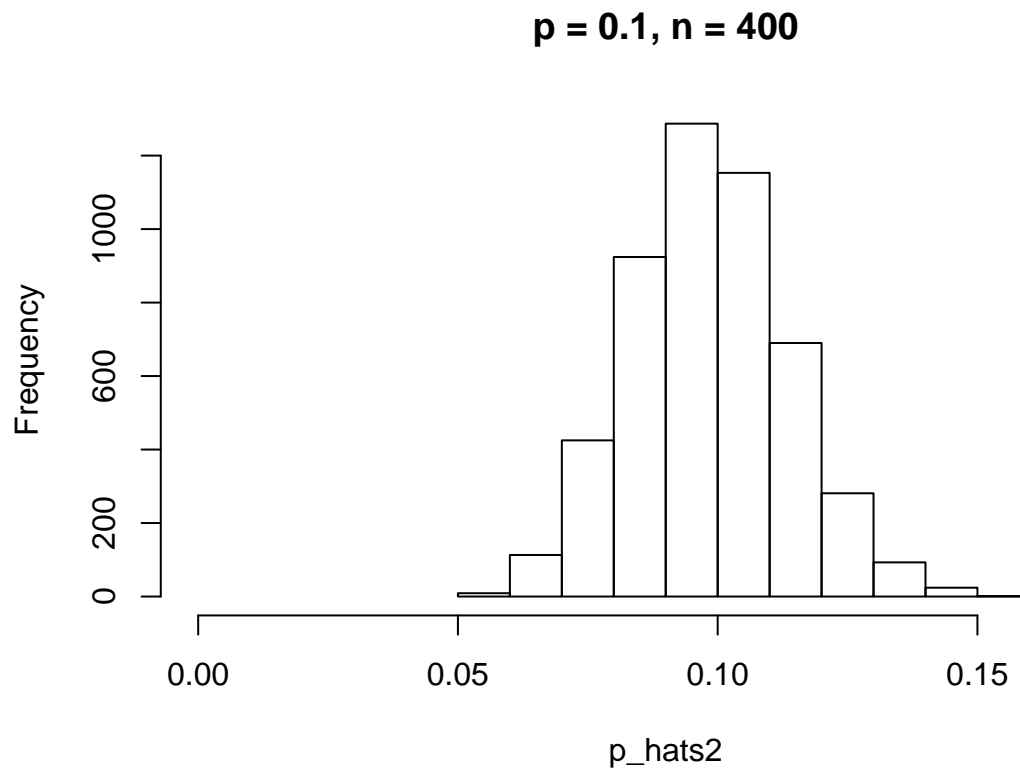

```
# when p = 0.1, n = 400
```

```
p2 <- 0.1
```

```
n2 <- 400
p_hats2 <- rep(0, 5000)

for(i in 1:5000){
  samp <- sample(c("atheist", "non_atheist"), n2, replace = TRUE, prob = c(p2, 1 - p2))
  p_hats2[i] <- sum(samp == "atheist") / n2
}

hist(p_hats2, main = "p = 0.1, n = 400", xlim = c(0, 0.18))
```
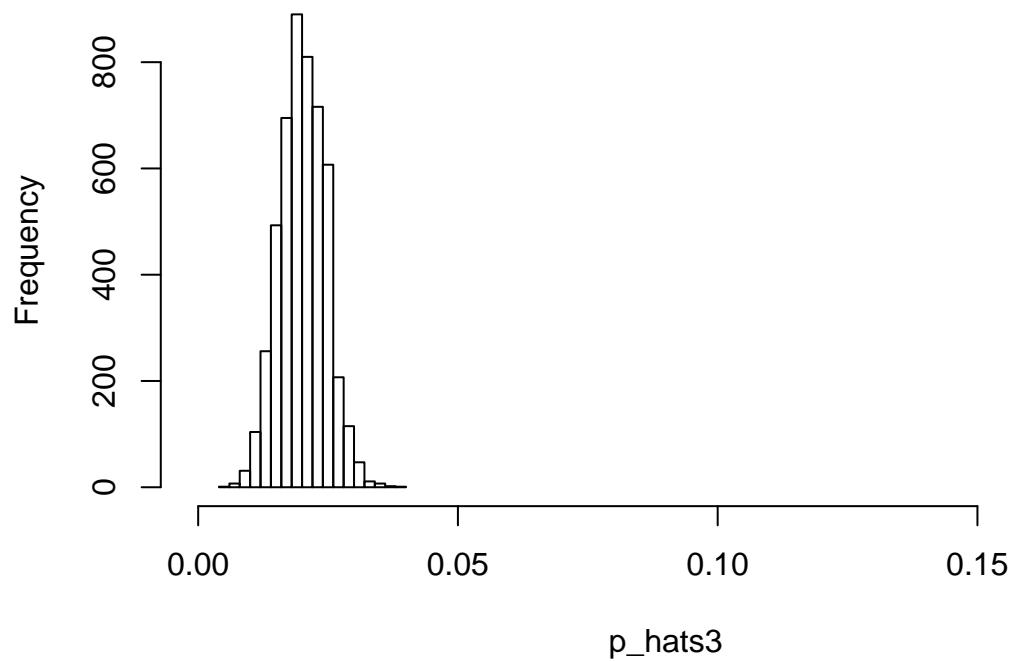
## p = 0.1, n = 400



```
# when p = 0.02, n = 1040

p3 <- 0.02
n3 <- 1040
p_hats3 <- rep(0, 5000)

for(i in 1:5000){
  samp <- sample(c("atheist", "non_atheist"), n3, replace = TRUE, prob = c(p3, 1 - p3))
  p_hats3[i] <- sum(samp == "atheist") / n3
}

hist(p_hats3, main = "p = 0.02, n = 1040", xlim = c(0, 0.18))
```
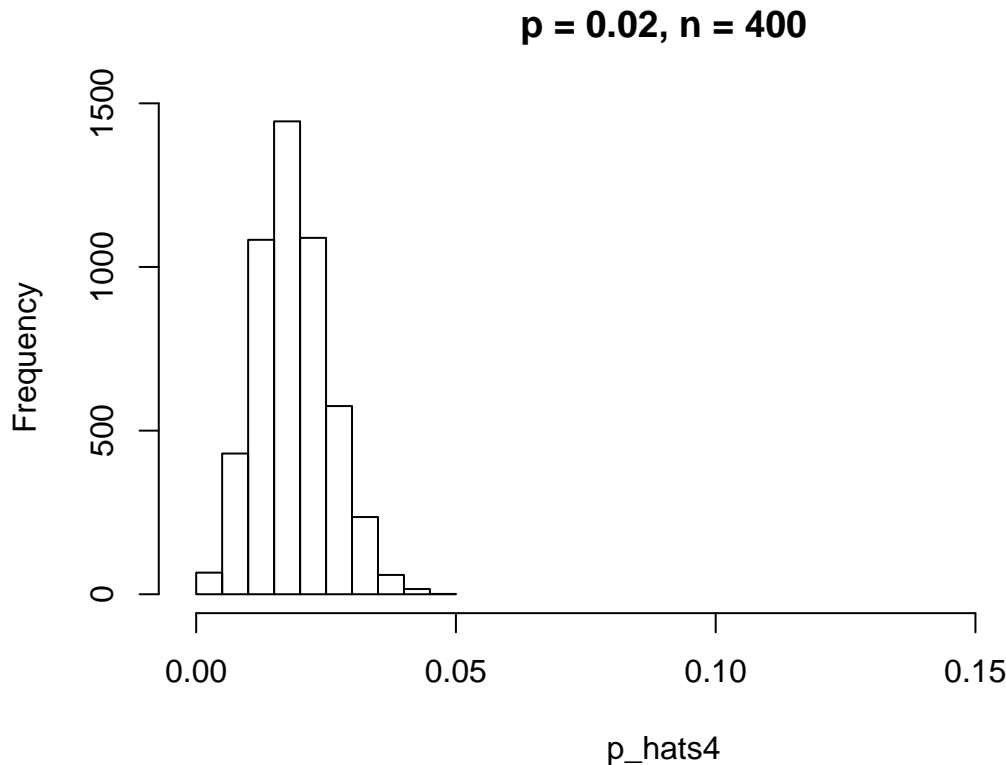
## p = 0.02, n = 1040



```
# when p = 0.02, n = 400

p4 <- 0.02
n4 <- 400
p_hats4 <- rep(0, 5000)

for(i in 1:5000){
  samp <- sample(c("atheist", "non_atheist"), n4, replace = TRUE, prob = c(p4, 1 - p4))
  p_hats4[i] <- sum(samp == "atheist") / n4
}

#fourth histogram
hist(p_hats4, main = "p = 0.02, n = 400", xlim = c(0, 0.18))
```

## p = 0.02, n = 400



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?

In case of Australia, it is sensible to proceed with inference as there are 104 (.1 x 1040) successes meeting the threshold of at least 10 success and failures each for normal distribution. Ecuador, on the other hand, has only 8 (.02 x 400) success and does not meet the threshold of at least 10 successes. Therefore, it not advisable to proceed with the inference in the case of Ecuador.

---

## 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.

  **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.

```
# H_o: There was no change between 2005 and 2012.
# H_a: There was a change between 2005 and 2012.

spain05 <- subset(atheism, nationality == "Spain" & year == "2005")
ath_spain05 <- count(spain05$response == "atheist")
non_ath_spain05 <- count(spain05$response == "non-atheist")
tot_spain05 <- ath_spain05 + non_ath_spain05
percent_ath_spain05 <- ath_spain05/tot_spain05 *100

ath_spain05
```

```
##       x freq
## 1 FALSE 1031
## 2  TRUE  115
```

```
non_ath_spain05
```

```
##       x freq
## 1 FALSE  115
## 2  TRUE 1031
```

```
percent_ath_spain05
```

```
##     x    freq
## 1 NaN 89.9651
## 2  50 10.0349
```

```
spain12 <- subset(atheism, nationality == "Spain" & year == "2012")
ath_spain12 <- count(spain12$response == "atheist")
non_ath_spain12 <- count(spain12$response == "non-atheist")
tot_spain12 <- ath_spain12 + non_ath_spain12
percent_ath_spain12 <- ath_spain12/tot_spain12 *100

ath_spain12
```

```
##       x freq
## 1 FALSE 1042
## 2  TRUE  103
```

```
non_ath_spain12
```

```
##       x freq
## 1 FALSE  103
## 2  TRUE 1042
```

```
percent_ath_spain12
```
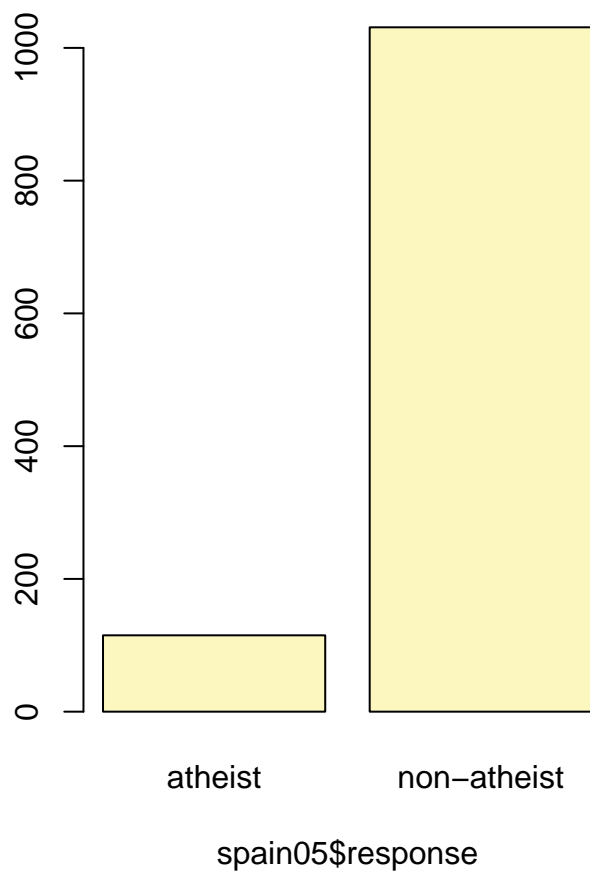
```
##     x      freq
## 1 NaN 91.004367
## 2  50  8.995633
```

```
inference(spain05$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

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



spain05$response

## 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 )

```
inference(spain12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

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

spain12$response
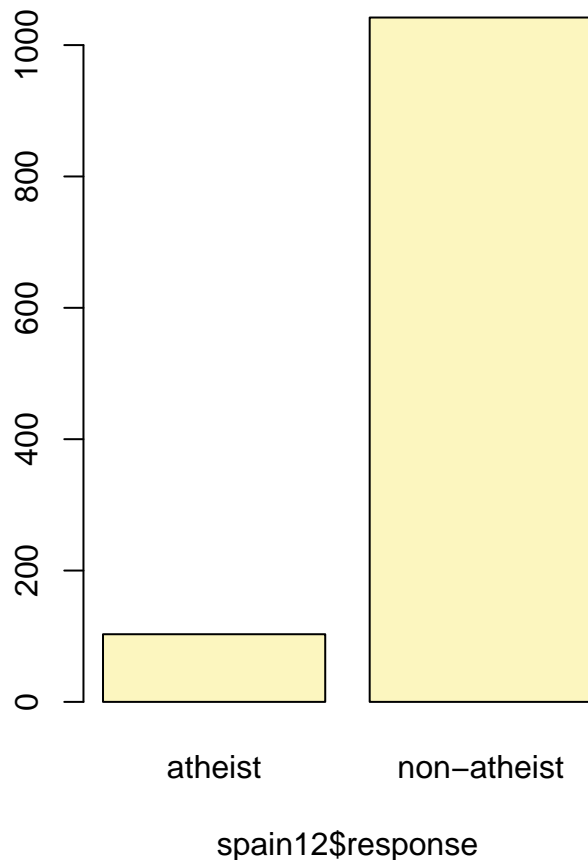
```
## p_hat = 0.09 ;  n = 1145
## Check conditions: number of successes = 103 ; number of failures = 1042
## Standard error = 0.0085
## 95 % Confidence interval = ( 0.0734 , 0.1065 )
```

```
combined_spain <- subset(atheism, nationality == "Spain" & year == "2005"  | nationality == "Spain" & y
```

```
inference(y = combined_spain$response, x = combined_spain$year, est = "proportion",type = "ht", null =
```

```
## 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
```
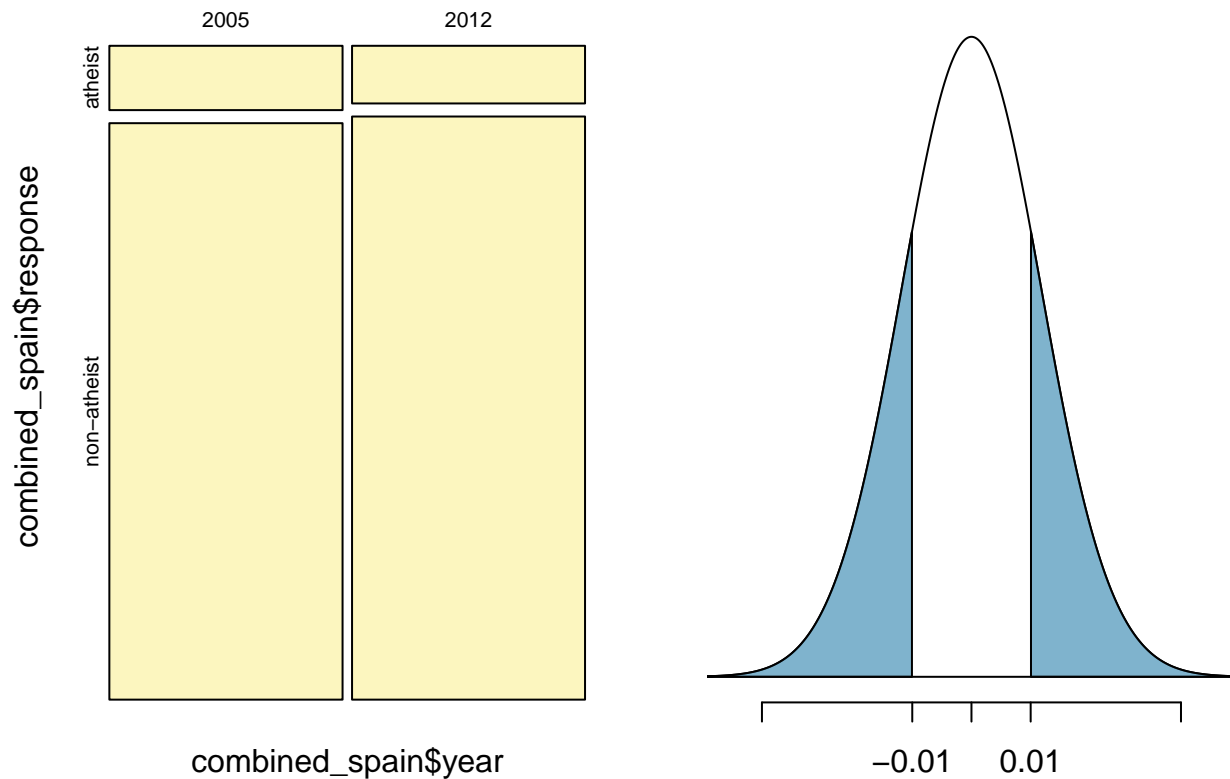


Since the p-value is greater than 0.05, we fail to reject the null hypothesis and accept that there has been the no significant change between the number of atheists in Spain between the years 2005 and 2012. We can also see that all the conditions for the inference have been met. Also, the CI for 2012 lies between the CI for 2005 indicating that we should not reject the null hypothesis.

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

```
# H_o: There was no change between 2005 and 2012.
# H_a: There was a change between 2005 and 2012.

US05 <- subset(atheism, nationality == "United States" & year == "2005")
ath_US05 <- count(US05$response == "atheist")
non_ath_US05 <- count(US05$response == "non-atheist")
tot_US05 <- ath_US05 + non_ath_US05
percent_US05 <- ath_US05/tot_US05 *100


ath_US05
```

```
##       x freq
## 1 FALSE  992
## 2  TRUE   10
```

non_ath_US05

```
##       x freq
## 1 FALSE   10
## 2  TRUE  992
```

percent_US05

```
##     x       freq
## 1 NaN 99.001996
## 2  50  0.998004
```

```
US12 <- subset(atheism, nationality == "United States" & year == "2012")
ath_US12 <- count(US12$response == "atheist")
non_ath_US12 <- count(US12$response == "non-atheist")
tot_US12 <- ath_US12 + non_ath_US12
percent_US12 <- ath_US12/tot_US12 *100
```

ath_US12

```
##       x freq
## 1 FALSE  952
## 2  TRUE   50
```
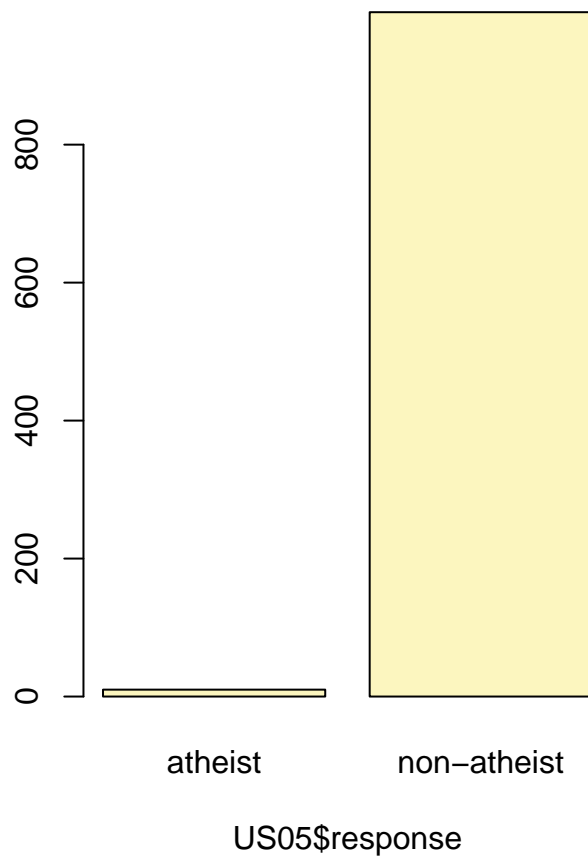
non_ath_US12

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

percent_US12

```
##     x      freq
## 1 NaN 95.00998
## 2  50  4.99002
```

```
inference(US05$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```
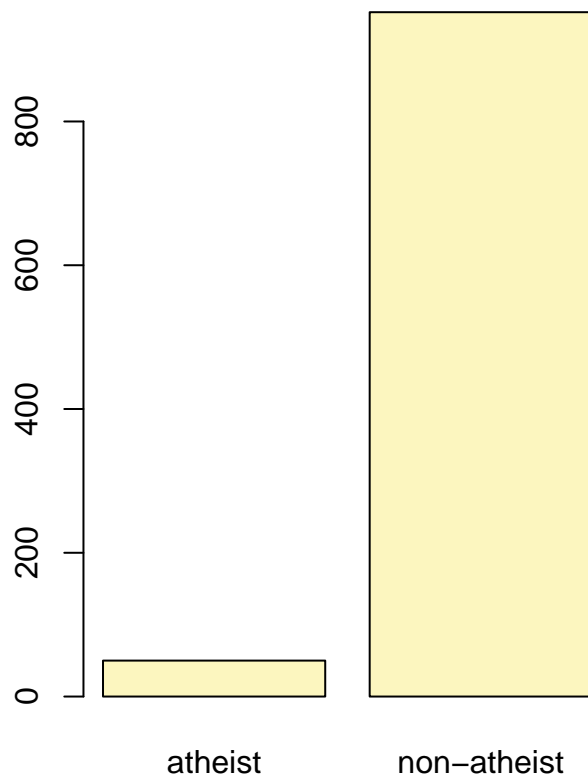
```
## Single proportion -- success: atheist
## Summary statistics:
```

US05$response

```
## 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 )
```

```r
inference(US12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

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

US12$response

```
## 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 )
```

```
combined_US <- subset(atheism, nationality == "United States" & year == "2005"  | nationality == "Unite
inference(y = combined_US$response, x = combined_US$year, est = "proportion",type = "ht", null = 0, alt
```

```
## 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
```
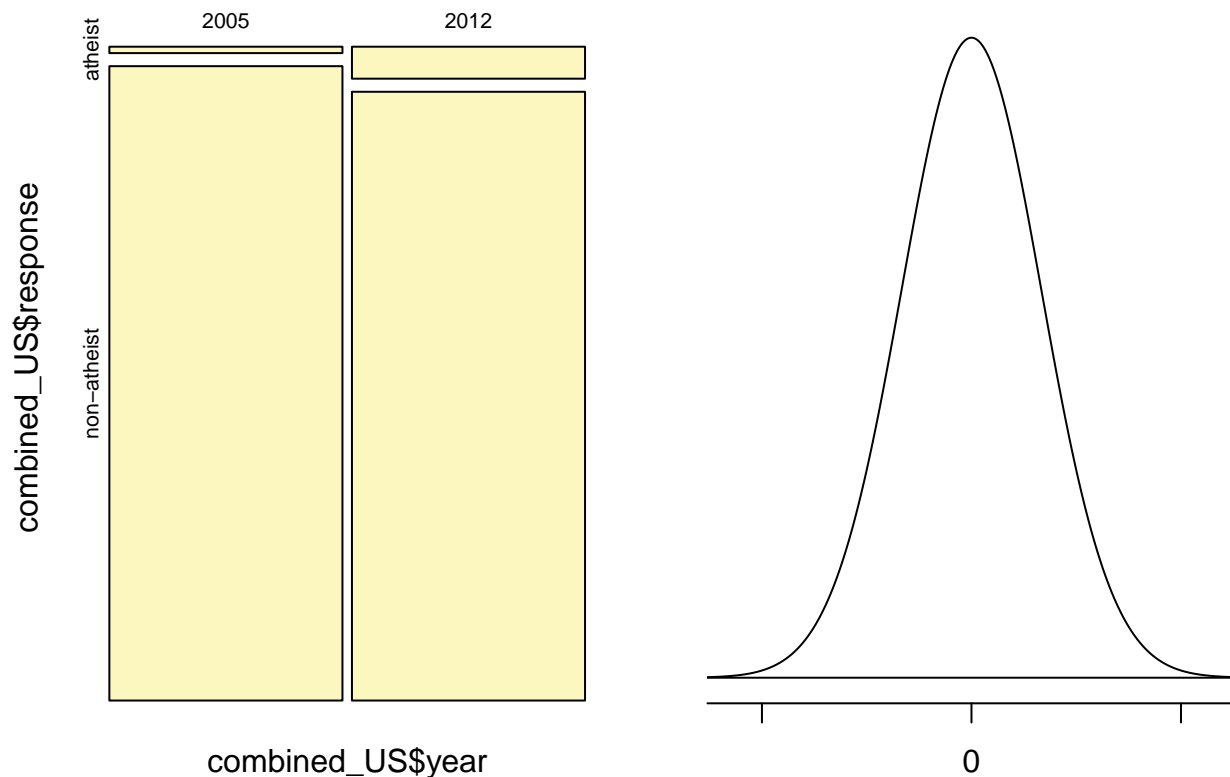


From the p-value approaching 0, we can see that the null hypothesis could be rejected and we can accept that there is evidence for a change in the atheism index in US between 2005 and 2012.

- 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. We would expect 5% of the countries to be included in the list of the countries with an evidence of change to atheism index by type 1 error if we use a significant level of 0.05.

- 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.

```
p <- .5
me <- .01
z_score <- qnorm(0.975)
```

```r
n <- round((p - p^2)/(me/z_score)^2, 0)
paste('We need to sample ', n, ' people to ensure that we are within the guidelines.', sep="")
```

```
## [1] "We need to sample 9604 people to ensure that we are within the guidelines."
```