Hao-6

```
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.
```

```
library(dplyr)
library(tidyr)
#setwd("C:/Users/bhao/Google Drive/CUNY/git/DATA606/Lab6")
setwd("~/Google Drive/CUNY/git/DATA606/Lab6")
load('more/atheism.RData')
```

Exercise1: 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?

These appear to be sample statistics. To be population parameters, the survey would have had to ask every last person within each country covered.

Exercise2: 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?

We would have to assume that the sample was representative of the global population. The survey covered 57 countries out of the ~200 or so that exist. Of course, regions within countries can be significantly different from one another as well. They survey also does not say anything about age, which can also be a relevant factor. Overall, it's very hard to say if this survey represents the entire human population well, but given the scope of the survey, it's probably as good as one can expect.

Exercise3: What does each row of Table 6 correspond to? What does each row of atheism correspond to?

Row 6 in the atheism data frame corresponds to an individual respondent in Afghanistan.

```
atheism[6,]
```

```
##   nationality    response year
## 6 Afghanistan non-atheist 2012
```

Exercise4: 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?

Table 6 actually shows 5% for 'convinced atheists' for the United States, and this matches data.

```
us12 <- subset(atheism, nationality == "United States" & year == "2012")
us12 %>% group_by(response) %>%
  summarise(n = n()) %>%
  mutate(freq = round(n / sum(n), 2))
```

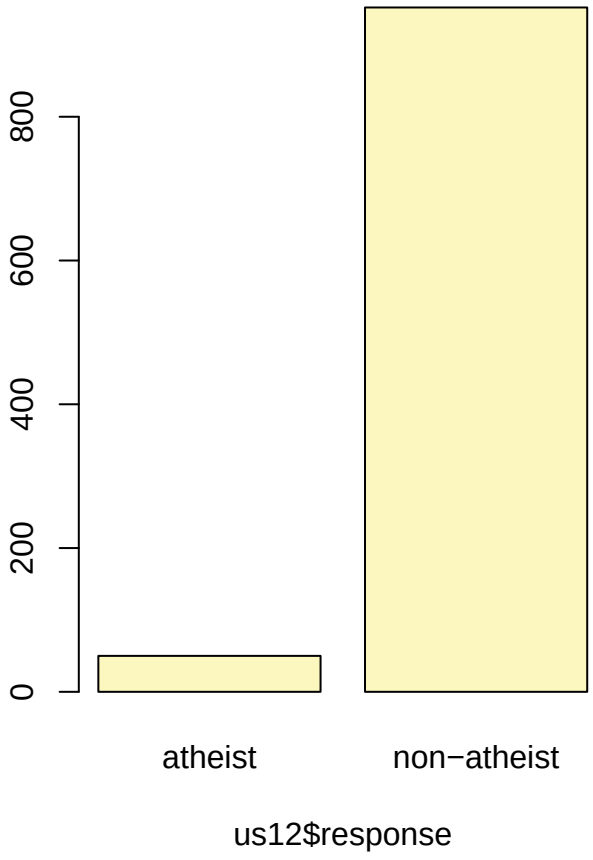
```
## # A tibble: 2 x 3
##      response     n  freq
##         <fctr> <int> <dbl>
## 1     atheist    50  0.05
## 2 non-atheist   952  0.95
```

Exercise5: 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 individual observations must be independent. A random sample from less than 10% of the population ensures independence. This condition is met.
- Size and skew. The more skewed the data, the larger the sample size need be in order to satisfy the conditions for inference. This condition is met as well.
- Success-failure condition. The sample sized must also be sufficiently large, which is checked using the success-failure condition. Both proportions well exceed n=10, so this condition is also 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 )
```

```
# manual calculation of standard error of proportion
p = us12 %>% group_by(response) %>%
  summarise(n = n()) %>%
  mutate(freq = round(n / sum(n), 2))
p = p[1, 3]
n = us12 %>% summarise(n = n())
n = n[1, 1]
se = sqrt(p * (1 - p) / n)
se
```

```
##          freq
## 1 0.006885143
```

Exercise6: 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?

The margin of error is more like +/- 1.35%

```
se * qnorm(0.025)
```

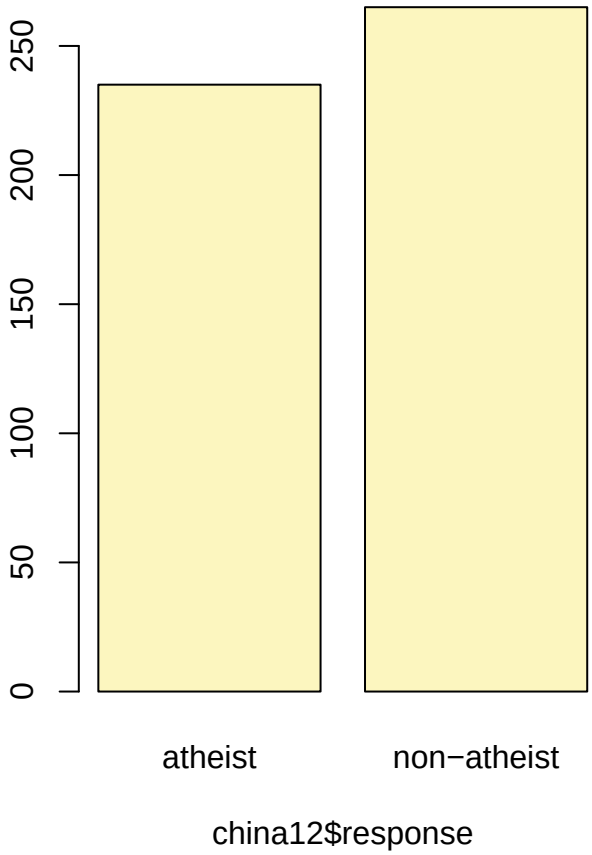
```
##          freq
## 1 -0.01349463
```

Exercise7: 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.

The first example is China. All of the conditions for inference are met. The confidence interval in this case is (0.4263 , 0.5137), and the standard error is 0.0223.

```
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 )
```

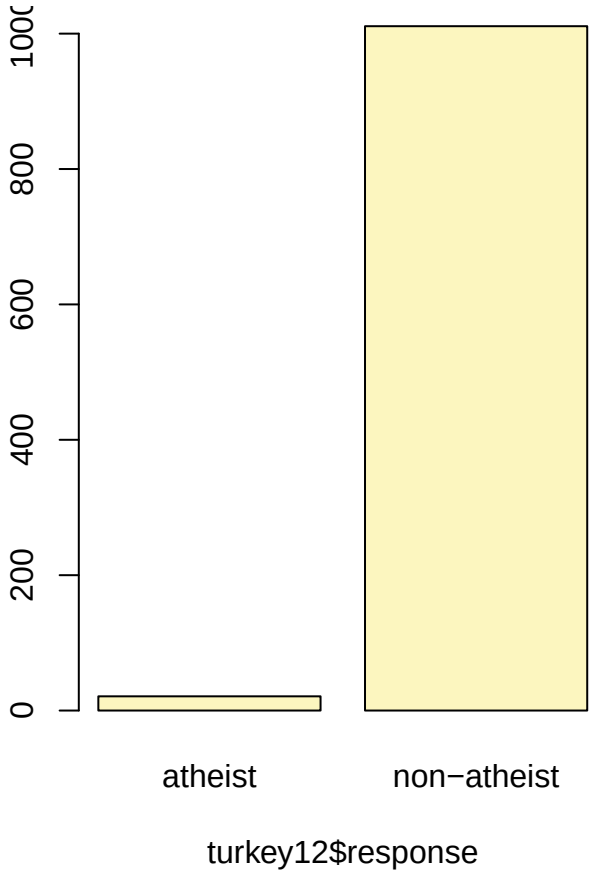
```
p = china12 %>% group_by(response) %>%
  summarise(n = n()) %>%
  mutate(freq = round(n / sum(n), 2))
p = p[1, 3]
n = china12 %>% summarise(n = n())
n = n[1, 1]
se = sqrt(p * (1 - p) / n)
se
```

```
##         freq
## 1 0.02232039
```

The second example is Turkey. All of the conditions for inference are met. The confidence interval in this case is (0.0117 , 0.029), and the standard error is 0.0044.

```
turkey12 <- subset(atheism, nationality == "Turkey" & year == "2012")
inference(turkey12$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 )
```

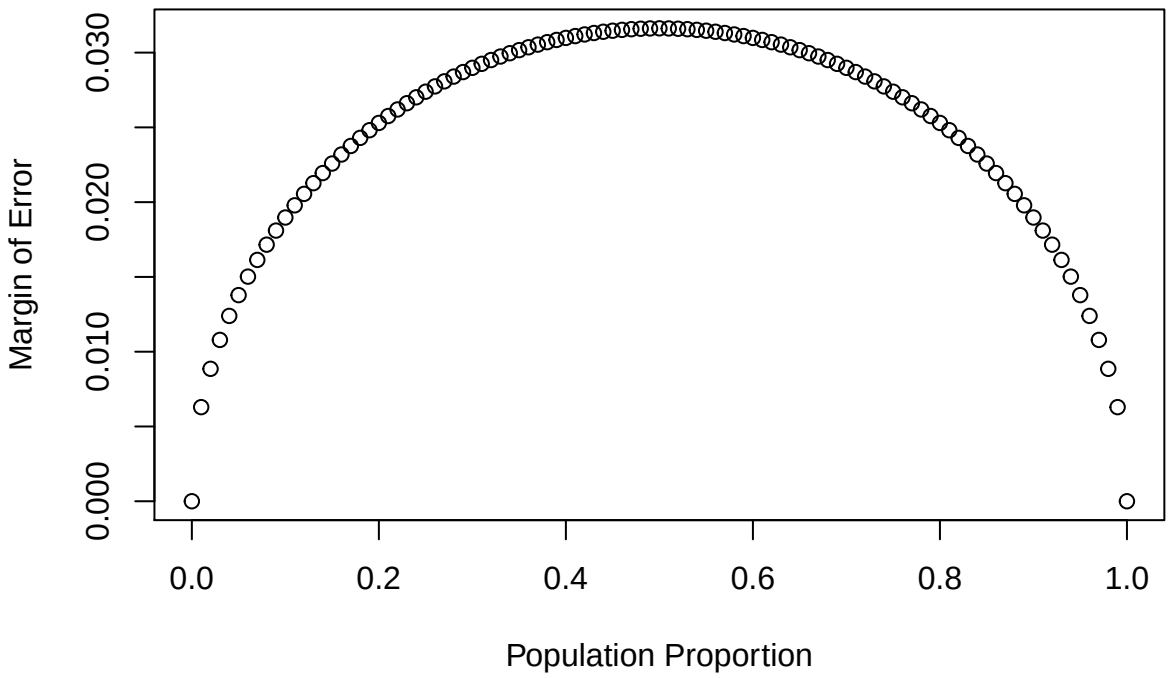
```
p = turkey12 %>% group_by(response) %>%
  summarise(n = n()) %>%
  mutate(freq = round(n / sum(n), 2))
p = p[1, 3]
n = turkey12 %>% summarise(n = n())
n = n[1, 1]
se = sqrt(p * (1 - p) / n)
se
```

```
##         freq
## 1 0.00435801
```

Exercise8: Describe the relationship between p and me.

The margin of error peaks when the population proportion equals 0.5 and then decreases on either side to 0.

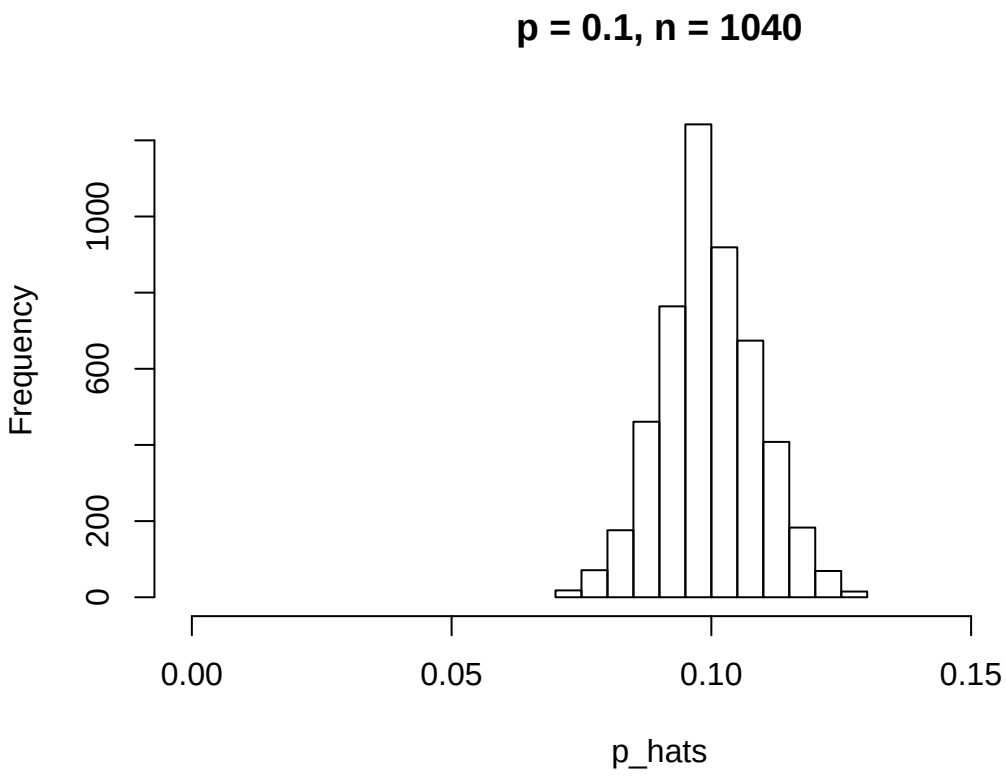
```
n <- 1000
p <- seq(0, 1, 0.01)
me <- 2 * sqrt(p * (1 - p)/n)
plot(me ~ p, ylab = "Margin of Error", xlab = "Population Proportion")
```




```
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))
```



Exercise9: Describe the sampling distribution of sample proportions at n=1040n=1040 and p=0.1p=0.1. Be sure to note the center, spread, and shape. Hint: Remember that R has functions such as mean to calculate summary statistics.

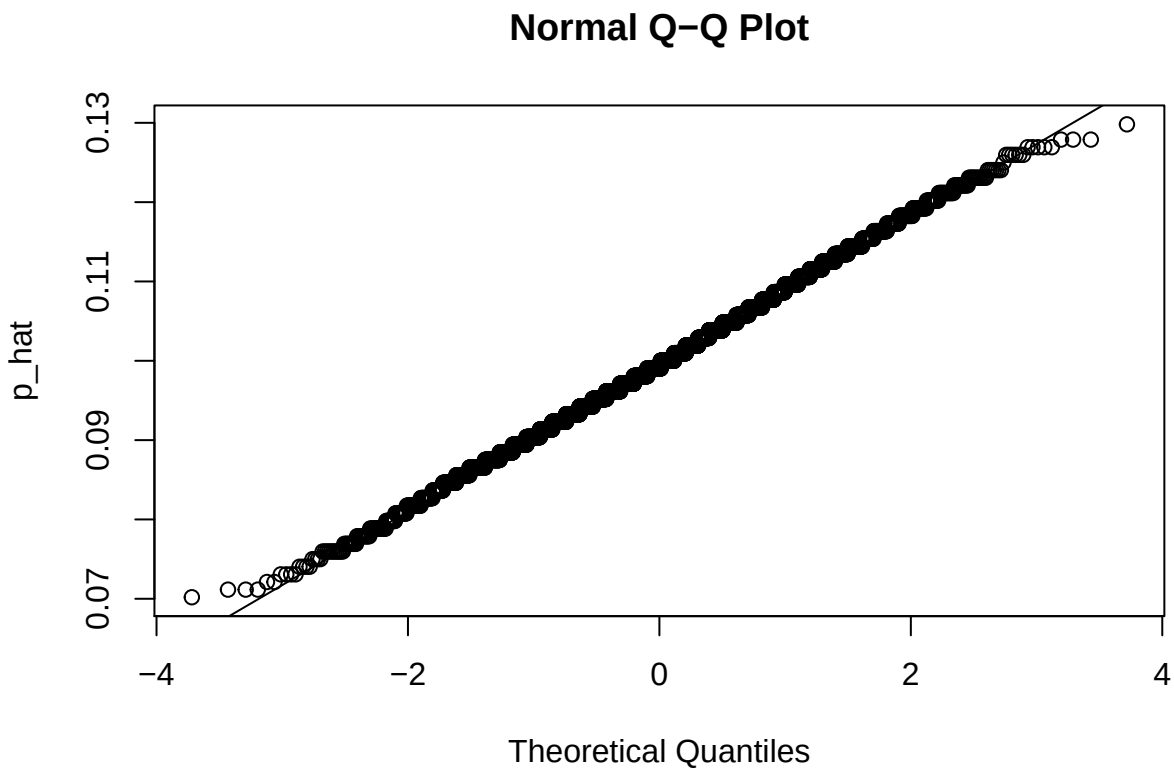
The mean is 0.1 with a sd of 0.01, and the qqplot indicates that the data are normally distributed.

```
psych::describeBy(p_hats)
```

```
## Warning in psych::describeBy(p_hats): no grouping variable requested
```

```
##    vars    n mean   sd median trimmed  mad  min  max range skew kurtosis
## X1    1 5000  0.1 0.01    0.1     0.1 0.01 0.07 0.13  0.06 0.06    -0.09
##    se
## X1  0
```

```
qqnorm(p_hats, ylab = 'p_hat')
qqline(p_hats)
```

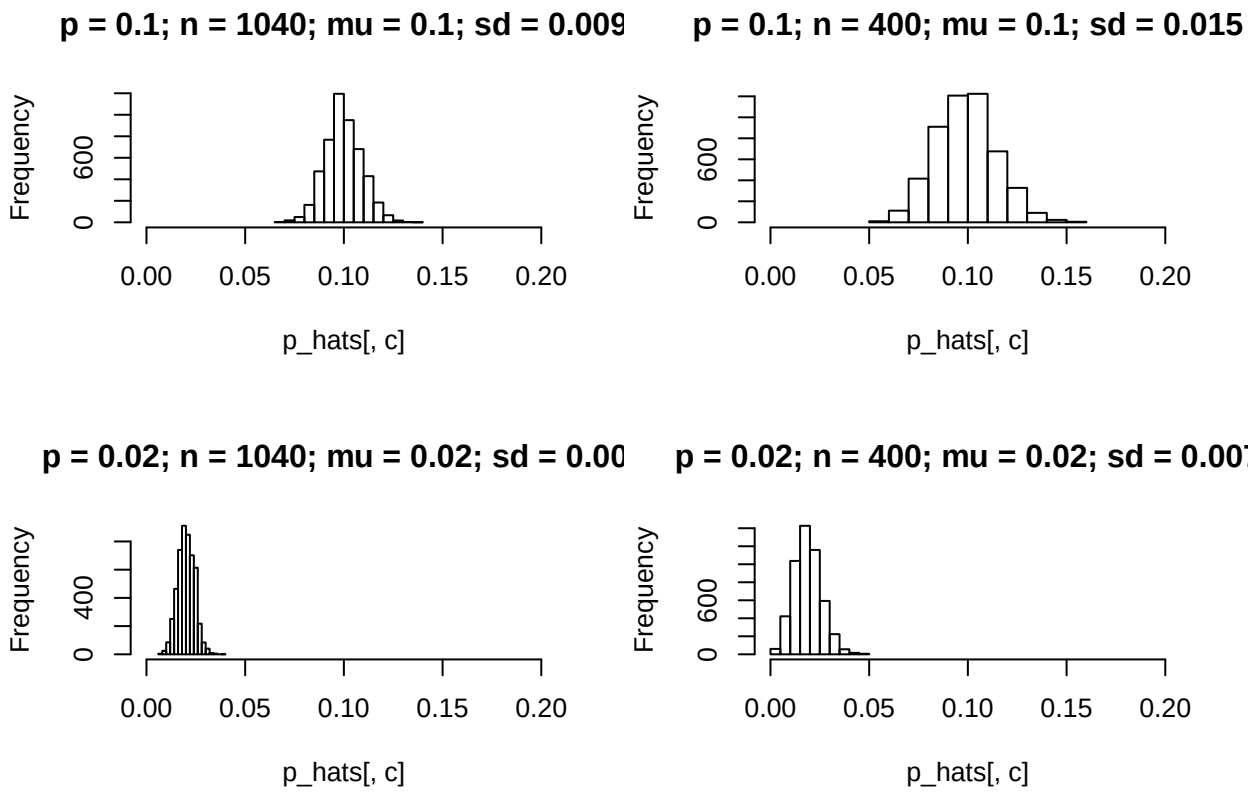


Exercise10: Repeat the above simulation three more times but with modified sample sizes and proportions: for n=400 and p=0.1, n=1040 and p=0.02, and n=400 and p=0.02. Plot all four histograms together by running the par(mfrow = c(2, 2)) command before creating the histograms. You may need to expand the 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 p^? How does p affect the sampling distribution?

As n decreases, the mean stays the same but the sd increases. When p decreases, both the mean and the sd decrease.

```
par(mfrow = c(2, 2))
p <- c(0.1, 0.1, 0.02, 0.02)
n <- c(1040, 400, 1040, 400)
p_hats <- data.frame(p_hats = rep(0, 5000), iter = c(1:4))

for (c in 1:4){
  for(i in 1:5000){
    samp <- sample(c("atheist", "non_atheist"), n[c], replace = TRUE, prob = c(p[c], 1-p[c]))
    p_hats[i, c] <- sum(samp == "atheist")/n[c]
  }
  hist(p_hats[, c], main = paste0("p = ", p[c], "; n = ", n[c],
                                  "; mu = ", round(mean(p_hats[, c]), 2),
                                  "; sd = ",  round(sd(p_hats[, c]), 3)), xlim = c(0, 0.2))
}
```




```
par(mfrow = c(1, 1))
```

Exercise11: 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?

As both n and p affect the standard error and thus the margin of error, then the report would more accurately report margin of error on a country by country basis.

OnYourOwn1a: 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. 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_0: $p_1 = p_2$
H_A: $p_1 <> P_2$

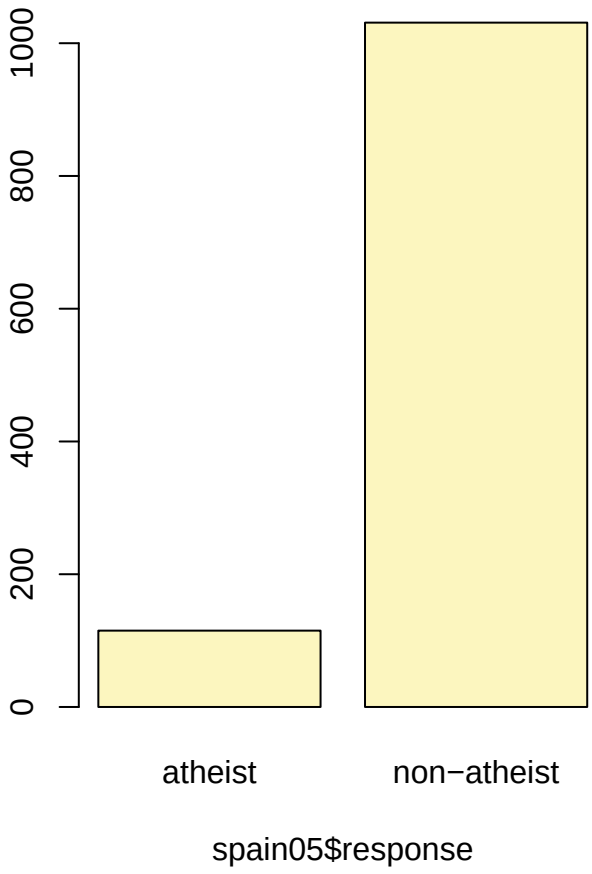
The conditions for inference are met. The confidence interval for spain05 and spain12 are (0.083 , 0.1177) and (0.0734 , 0.1065), respectively. Since they overlap, there is insufficient evidence to reject the null hypothesis that the two proportions are the same at a 5% significance level.

```
spain05 = atheism[atheism$nationality == 'Spain' & atheism$year == 2005, ]
spain12 = atheism[atheism$nationality == 'Spain' & atheism$year == 2012, ]

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

```
          success = "atheist")
```

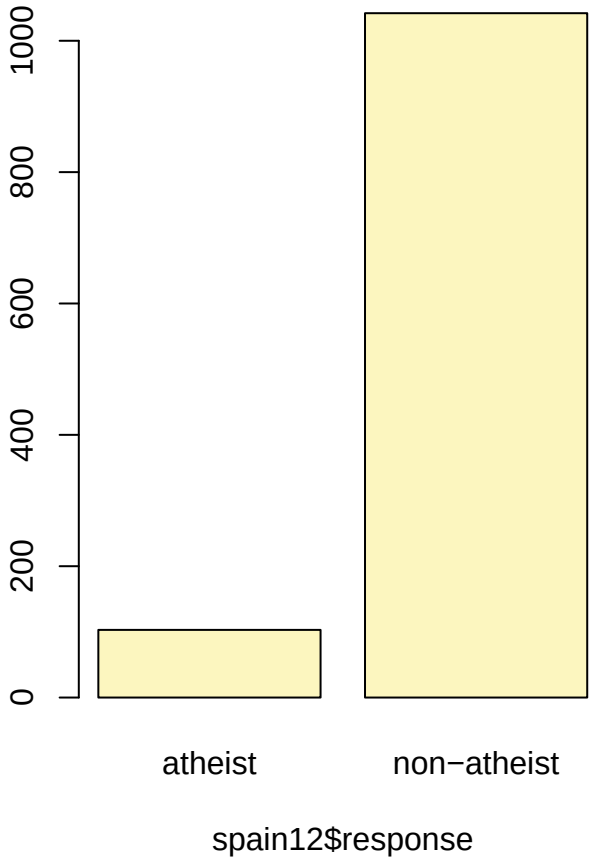
```
## Single proportion -- success: atheist
## Summary statistics:
```




```
## 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 )
```

```
inference(spain12$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 )
```

OnYourOwn1b: b. Is there convincing evidence that the United States has seen a change in its atheism index between 2005 and 2012?

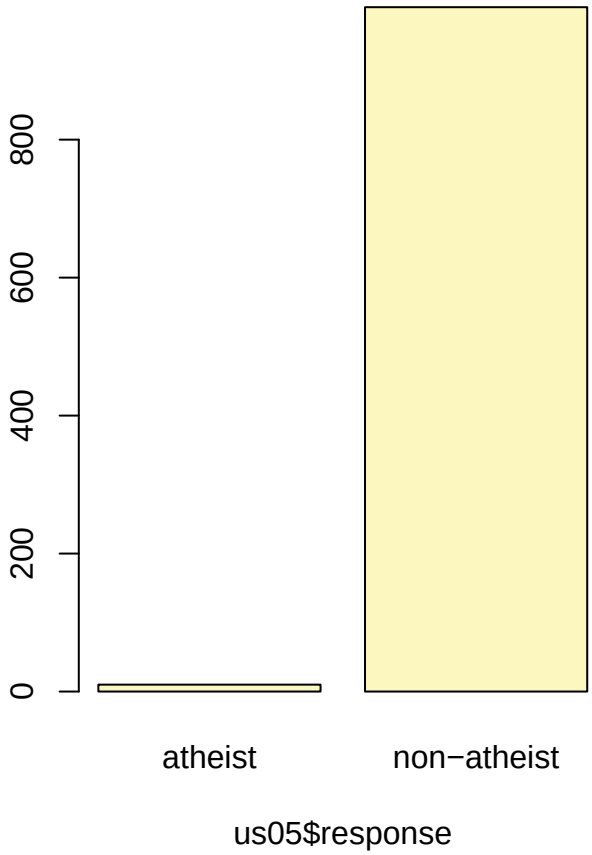
H_0: $p_1 = p_2$
H_A: $p_1 <> P_2$

Again, the conditions for inference are met (but just barely in the case of the 2005 data as it has exactly 10 successes). The confidence intervals for the 2005 and 2012 data are (0.0038 , 0.0161) and (0.0364 , 0.0634), respectively. Since the intervals do not overlap, there is sufficient evidence to reject the null in favor of the alternative hypothesis that the two proportions are in fact different at a 5% significance level.

```
us05 = atheism[atheism$nationality == 'United States' & atheism$year == 2005, ]
us12 = atheism[atheism$nationality == 'United States' & atheism$year == 2012, ]

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

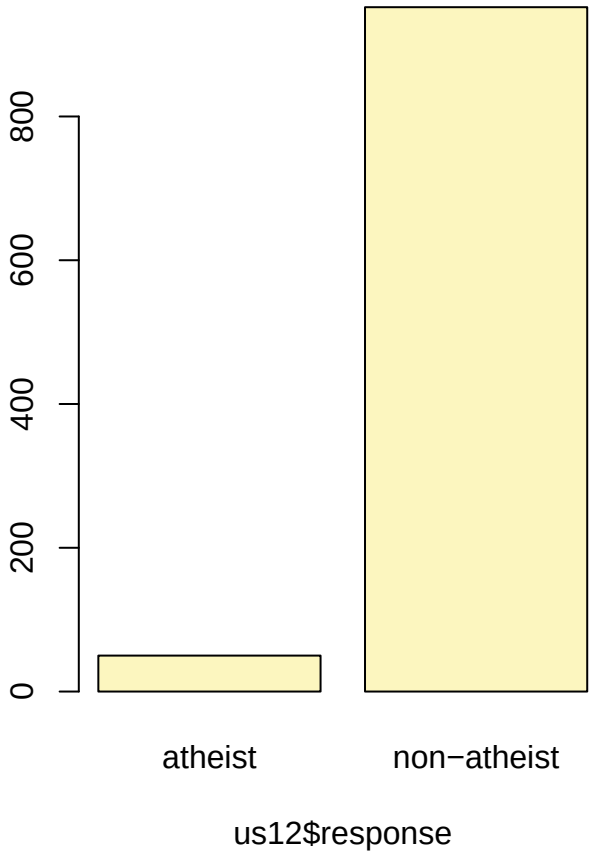
```
## Single proportion -- success: atheist
## Summary statistics:
```




```
## 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 )
```

```
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 )
```

OwnYourOwn2: 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.

A Type I error is rejecting the null hypothesis when it is in fact true. This matches the case above. The probability of a Type I error is the significance level, which is 5% in this particular case.

OnYourOwn3: 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.

Without knowing what to expect for p, we should assume the worst case scenario, i.e. the case where p results in the largest margin of error, which is 0.5. Solving for n results in having to sample 9604 people.

```
p = 0.5
me = 0.01
n = qnorm(0.975)^2 * p * (1 - p) / me^2
n
```

```
## [1] 9603.647
```