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.

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

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*?

   Although the percentages are written as if they are population parameters, "... Wester Europe, 14% of whose population says...", but realisiticly, they are sample statistics.

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 not be biased and must be representative of all people in the world. Due to the fact that each country represented generally has a sample size of ~1000, larger countries and small countries are not represented proportionally which could cause the global aggregate result to be skewed/biased depending on how the aggregates are computed.

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.

```
load("more/atheism.RData")
```

```
summary(atheism)
```

```
##             nationality          response          year
##  Pakistan          : 5409   atheist    : 5498   Min.   :2005
##  France            : 3359   non-atheist:82534   1st Qu.:2005
##  Korea, Rep (South): 3047                       Median :2012
##  Ghana             : 2995                       Mean   :2009
##  Macedonia         : 2418                       3rd Qu.:2012
##  Peru              : 2414                       Max.   :2012
##  (Other)           :68390
```

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

   Each row of Table 6 corresponds to a country level sample and associated proportions. The `atheism` data set contains individual cases across countries with a boolean, success/failure, data point on atheist or not, and the associated nationality and year of the response.

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?

```
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
```

```
atheist <- nrow(us12[us12$response == "atheist", ])
atheist
```

```
## [1] 50
```

```
total <- nrow(us12)
total
```

```
## [1] 1002
```

```
atheist / total
```

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

After rounding, the document's table 6 value for the US is the same.

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.

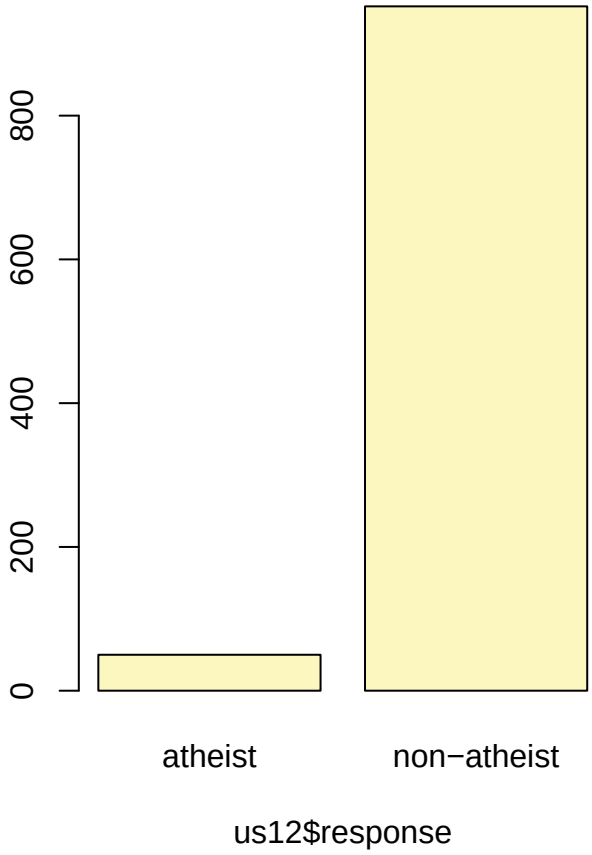
5. Write out the conditions for inference to construct a 95% confidence interval for the proportion of atheists in the United States in 2012. Are you confident all conditions are met?

    1. ***Observations are independent***: There are 1002 observations in the `us12` data set, which is well below 10% of the US population, so we accept that the observations are independent.
    2. ***Success-failure condition:*** $1002 \times 0.0499 = 50$ and $1002 \times 0.9501 = 952$ which are both greater than 10, therefore the condition is met.

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:
```




```
## 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?

```
lb <- 0.0364
ub <- 0.0634
diff <- ub - lb
moE <- diff / 2
moE
```

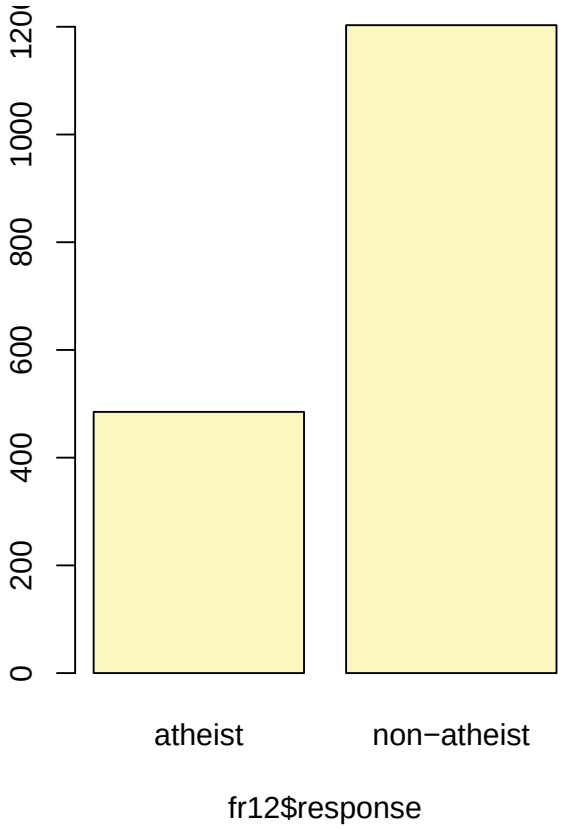
```
## [1] 0.0135
```

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.

First we examine France. As you will see, the success/failure counts are above 10. Again, we accept the cases are independent based on being less than 10% of the population of France.

```
fr12 <- subset(atheism, nationality == "France" & year == "2012")

inference(fr12$response, est = "proportion", type = "ci", method = "theoretical",
      success = "atheist")
```

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

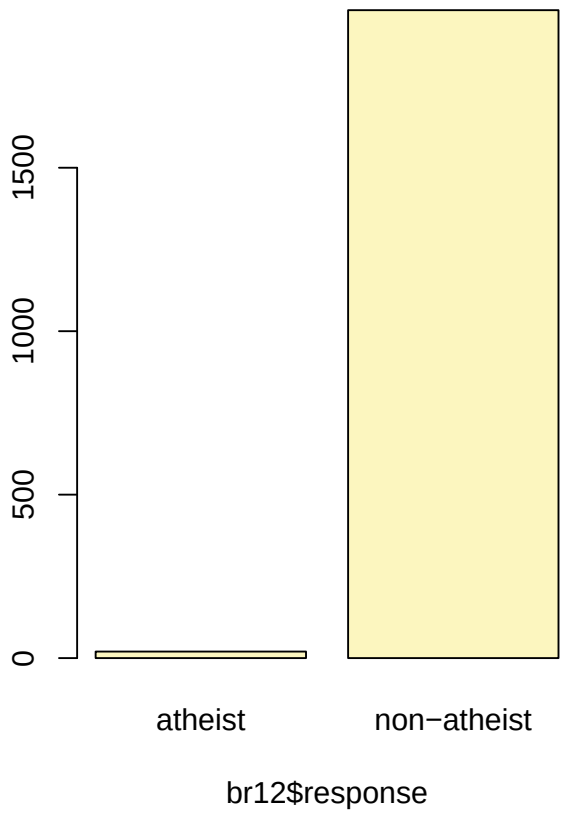



```
## p_hat = 0.2873 ;  n = 1688
## Check conditions: number of successes = 485 ; number of failures = 1203
## Standard error = 0.011
## 95 % Confidence interval = ( 0.2657 , 0.3089 )
```

Secondly, we examine Brazil As you will see, the success/failure counts are above 10. Again, we accept the cases are independent based on being less than 10% of the population of Brazil

```
br12 <- subset(atheism, nationality == "Brazil" & year == "2012")
#summary(sk12)
inference(br12$response, est = "proportion", type = "ci", method = "theoretical",
      success = "atheist")
```

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




```
## 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 )
```

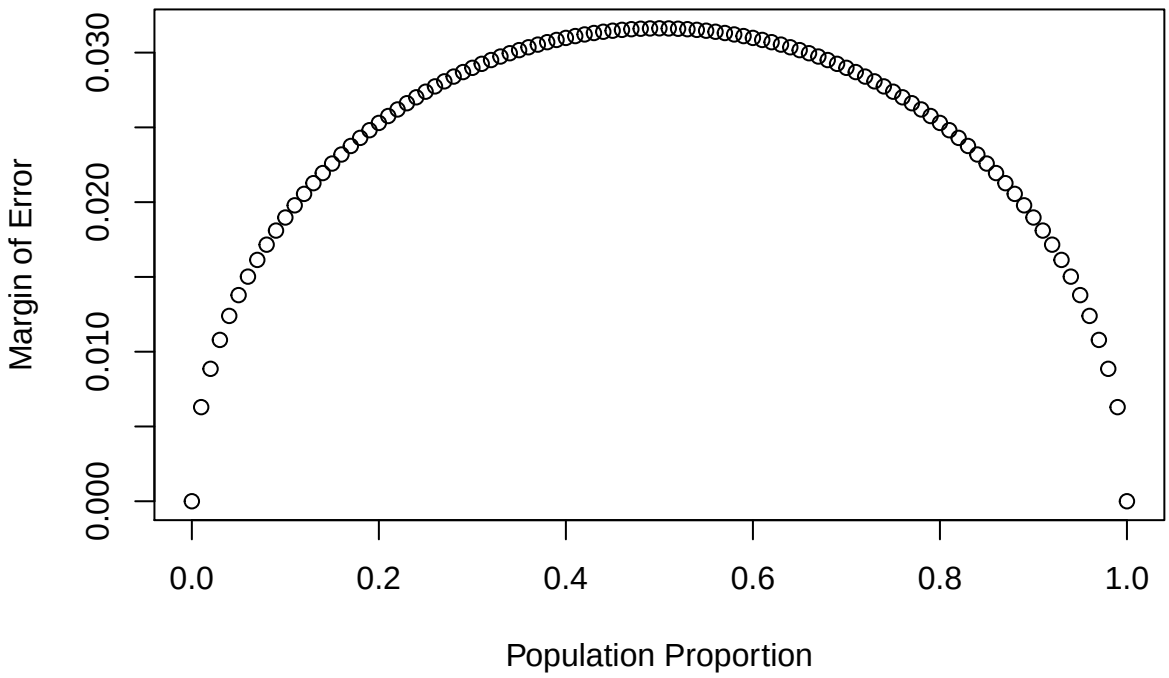
How does the proportion affect the margin of error?

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`.

As p increases from 0 - .5, me also increases to a maximum ~ 0.03. Between 0.5 - 1, as p increases me decreases back to 0.

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 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.

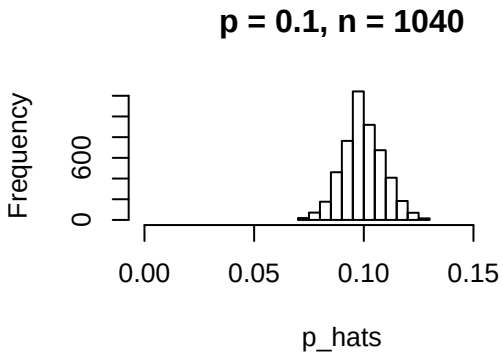
9. Describe the sampling distribution of sample proportions at $n = 1040$ and $p = 0.1$. Be sure to note the center, spread, and shape.
   Hint: Remember that R has functions such as `mean` to calculate summary statistics.

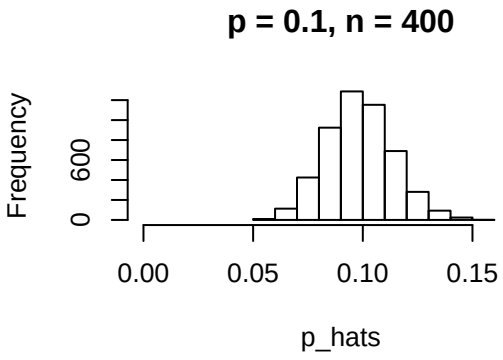
   The samplng distribution of sample proportions is centered near 0.1, the population proportion. It has a spread of ~0.025, is unimodel and is normally shaped

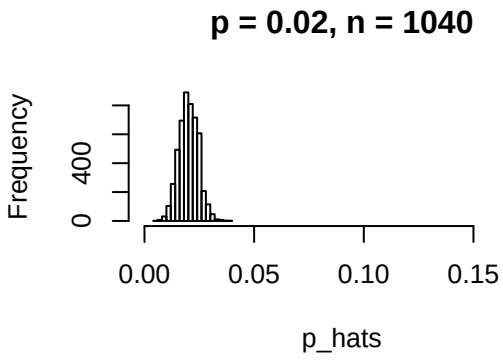
10. Repeat the above simulation three more times but with modified sample sizes and proportions: for $n = 400$ and $p = 0.1$, $n = 1040$ and $p = 0.02$, and $n = 400$ and $p = 0.02$. Plot all four histograms together by running the `par(mfrow = c(2, 2))` command before creating the histograms. You may 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?

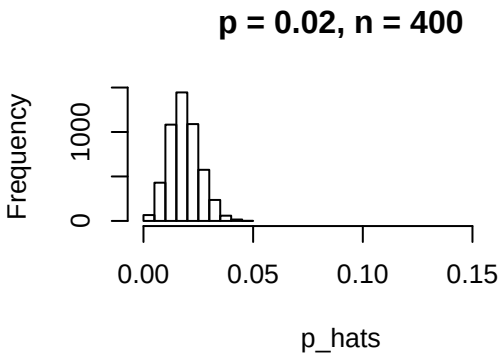
```
par(mfrow = c(2, 2))

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))

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
```

```r
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))
```









All 4 distributions are fairly normal in shape. The change in p shifts the location of the curve to center around the population value, as we would expect. The change in n affects the spread of the distribution. The greater the n, the smaller the spread.

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

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.

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?

```r
# Australia ME
p <- 0.1
n <- 1040
me <- 2 * sqrt(p * (1 - p)/n)
me
```

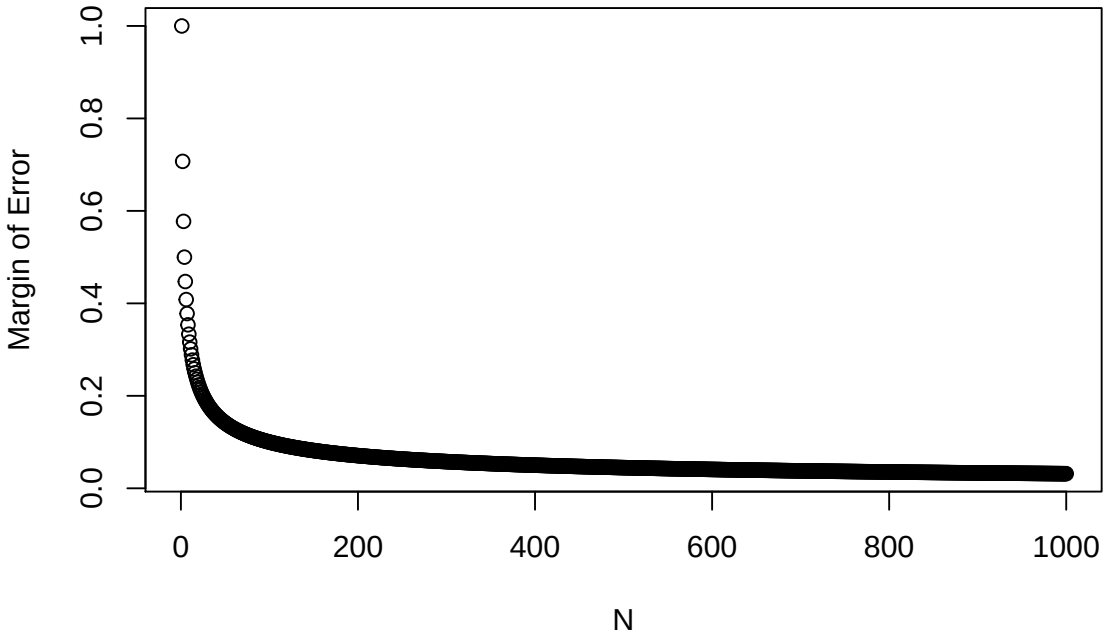
```
## [1] 0.01860521
```

```r
# Ecuador ME
p <- 0.02
n <- 400
me <- 2 * sqrt(p * (1 - p)/n)
me
```

```
## [1] 0.014
```

The margin's of error are interestingly similar. The following chart shows how n affects margin of error with a fixed proportion of 0.5. Interestingly, n >= 200, the affect is minimal, so the change from 400 to 1040 is not significant, but as shown in the chart above ME vs P, the change in p from 0.02 to 0.1 is more significant. Either way, reporting margin of error of +- 3-5% seems conservative to me at the individual country level after seeing this analysis. With that said, it does seem inaccurate.

```r
n <- seq(0, 1000, 1)
p <- 0.5
me <- 2 * sqrt(p * (1 - p)/n)
plot(me ~ n, ylab = "Margin of Error", xlab = "N")
```



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.

  $H_0 : \mu_{2005} - \mu_{2012} = 0$

  $H_a : \mu_{2005} - \mu_{2012} \neq 0$

  Both 2005, and 2012 have 1045+ cases which is less than 10% of the population. Both years also have success and failure counts > 10.

```
sp <- subset(atheism, nationality == "Spain")
sp2005 <- sp[sp$year == 2005,]
nrow(sp2005)
```

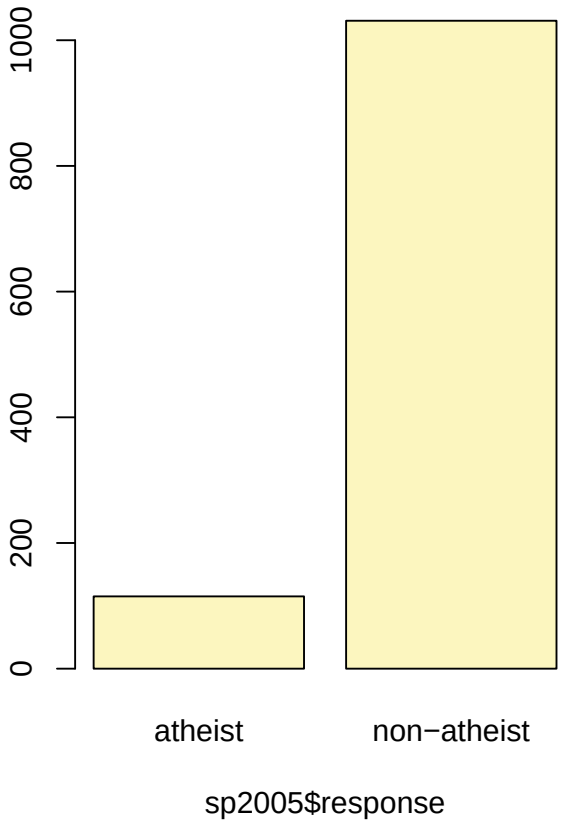
```
## [1] 1146
```

```r
sp2012 <- sp[sp$year == 2012,]
nrow(sp2012)
```

```
## [1] 1145
```

```r
inference(sp2005$response, est = "proportion", type = "ci", method = "theoretical",
      success = "atheist")
```

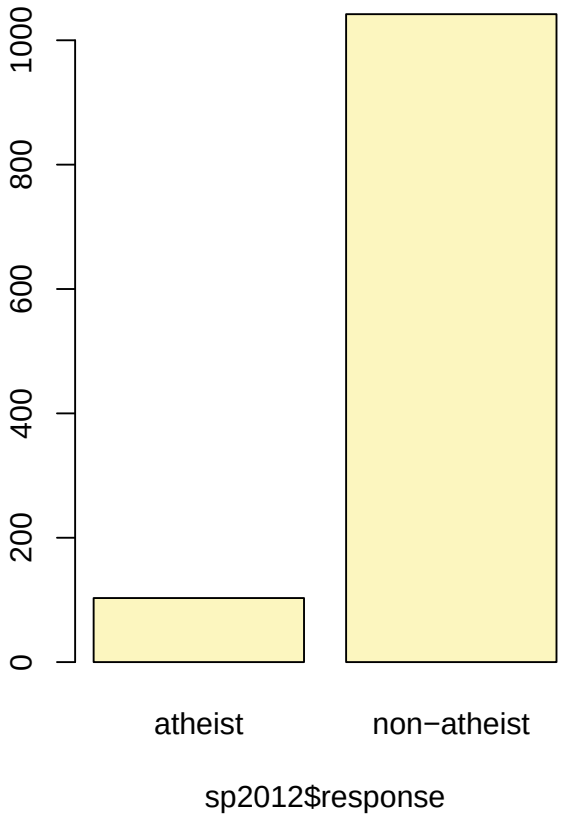
```
## Single proportion -- success: atheist
## Summary statistics:
```




```
## 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 )
```

```r
inference(sp2012$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 )
```

In 2005, Spain's atheist 95% confidence interval is 0.083 - 0.1177. In 2012, it is 0.0734 - 0.1065. Due to the overlap of confidence intervals, I conclude that there is not a statistically significant difference, but instead could be due to chance.

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

```
us <- subset(atheism, nationality == "United States")
us2005 <- us[us$year == 2005,]
nrow(us2005)
```

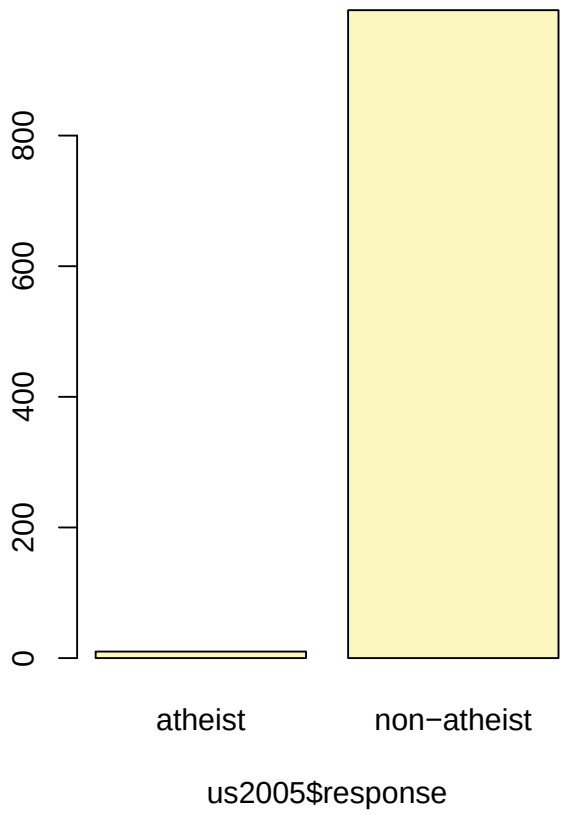
```
## [1] 1002
```

```
us2012 <- us[us$year == 2012,]
nrow(us2012)
```

```
## [1] 1002
```

```
inference(us2005$response, est = "proportion", type = "ci", method = "theoretical",
      success = "atheist")
```

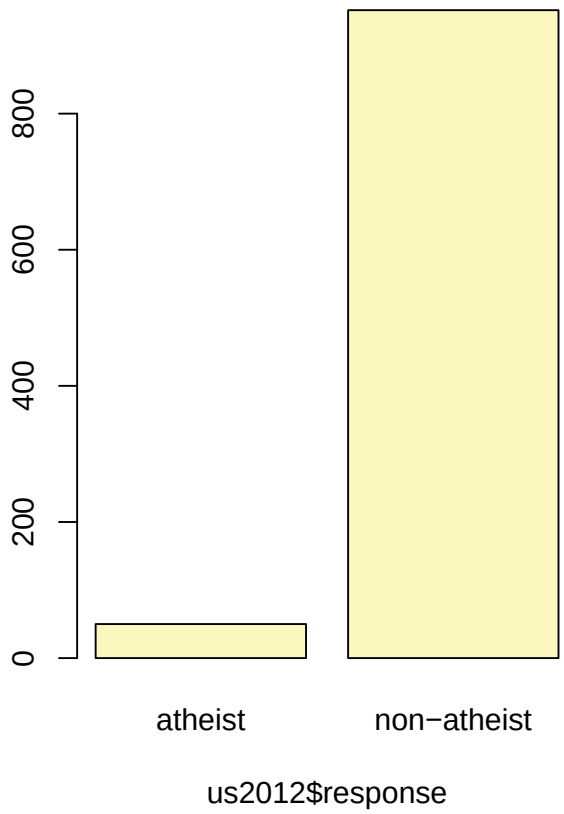
```
## Single proportion -- success: atheist
## Summary statistics:
```




```
## 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(us2012$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 )
```

In 2005, the 95% confidence interval of U.S. atheism is 0.0038 - 0.0161. In 2012, it is 0.0364 - 0.0634. In this case, there is no overlap, and I conclude that this is convincing evidence that the Unites States saw a change in its atheism index 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.

  When doing a proper hypothesis test at $\alpha = 0.05$, we would expect to correctly detect a change 95% of the time, and fail to detect a change 5% of the time.

- 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.

  The worst case p would for a margin of error would be 0.5, so we will use this as a basis to compute the required n value. At 95% confidence level, using Z=1.96, we would need to survey 9604 people.

```
p <- 0.5
me <- 0.01
```

```
n <- (p - p^2) / (me / 1.96)^2
n
```

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