# 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*?

They appear to be sample statistics since the percentages were calculated based upon the responses given by individuals either chosen to or self-choosing to participate in the survey.

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?

We must assume that the sampling method results in independent observations and that there are at least 10 "successes" and 10 "failures" within the sample. If data come from a simple random sample and consist of less than 10% of the population, then the independence assumption is reasonable. Since the sample size was a little over 50,000, we can be confident that the sample size is less than 10% of the world's population.

## 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")
```

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

Each row of Table 6 provides a summary of the survey responses/observations recorded for a given nationality.

Each row of "atheism" represents a single survey observation/response, with each observation recording the respondent's nationality, their survey response, and the year the survey response was provided.

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")

# sum responses given to "atheist" vs. "non-atheist" survey question
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
```

R's **summary** function can be used to quickly sum up the "atheist" and "non-atheist" responses for respondents from the United States. As shown above, 952/1002 respondents responded "non-atheist" while 50/1002 respondents responded "atheist".

So we have 50/1002 = .0499 = .05 of US respondents responding as "atheist"

```
50/1002
```

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

As such, the proportion we've calculated from our data set agrees with the percentage shown in Table 6.

## 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?

Given: n = 1002; ^p = .05

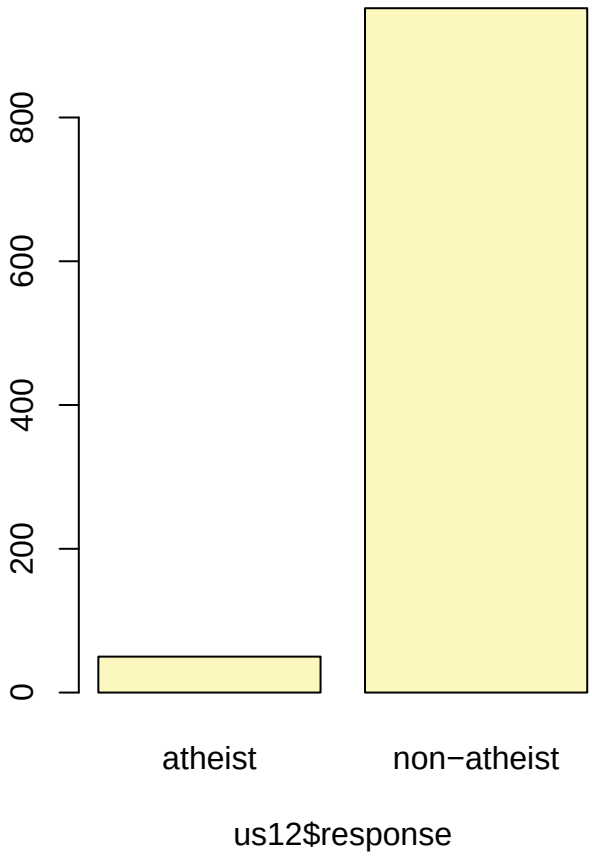
Condition 1: **Independence:** We have no indication that the respondents were randomly sampled. In fact, since large percentages of the responses were collected either online via the web or over the phone, we have reason to doubt that the responses are truly random. On the other hand, the sample size (1002) is definitely less than 10% of the population of the USA. If we are to proceed with any further analysis we'll have to assume independence despite the questions we have about the researchers' sampling techniques.

Condition 2: **Success-Failure:** 50 people responded correctly (successes) while 952 people responded incorrectly (failures). Both amounts are greater than 10 so we may proceed.

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 )
```

Based on the output of the **inference** function, we have a 95% confidence interval of ( 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?

To determine the margin of error, we calculate 1/2 the difference of the bounds of the confidence interval:

```
(.0634 - .0364)/2
```

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

The calculation shows the margin of error to be 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.

The two countries we'll look at are Japan and France.

**Japan**

We'll start with Japan:

```
j12 <- subset(atheism, nationality == "Japan" & year == "2012")

# sum responses given to "atheist" vs. "non-atheist" survey question
summary(j12)
```

```
##       nationality          response        year
##  Japan      :1212   atheist    :372   Min.   :2012
##  Afghanistan:   0   non-atheist:840   1st Qu.:2012
##  Argentina  :   0                     Median :2012
##  Armenia    :   0                     Mean   :2012
##  Australia  :   0                     3rd Qu.:2012
##  Austria    :   0                     Max.   :2012
##  (Other)    :   0
```

We again use R's **summary** function to quickly sum up the "atheist" and "non-atheist" responses for respondents. As shown above, 840/1212 respondents responded "non-atheist" while 372/1212 respondents responded "atheist".

So we have 372/1212 = .307 of Japanese respondents responding as "atheist"

```
372/1212
```

```
## [1] 0.3069307
```

Now we'll state the conditions for inference:

Given: n = 1212; ^p = .307

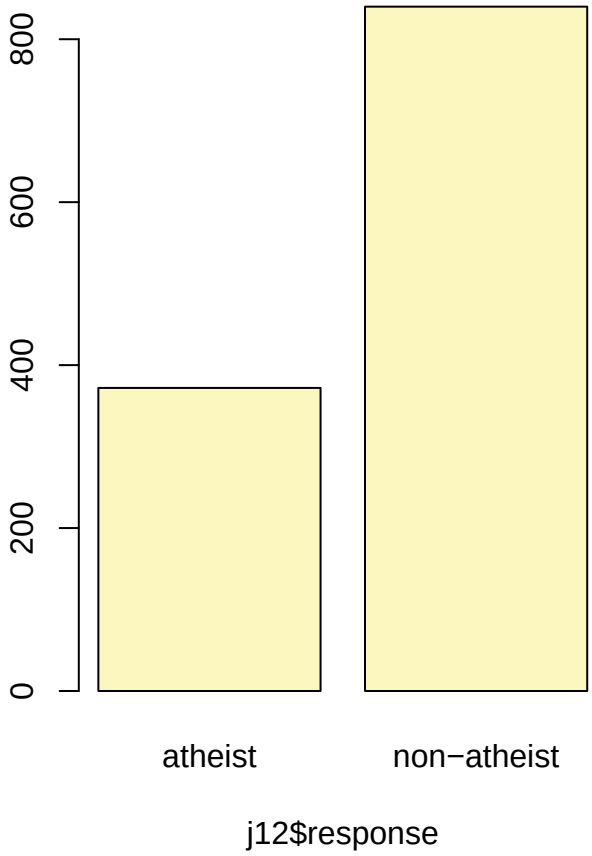
Condtion 1: **Independence:** We have no indication that the respondents were randomly sampled. In fact, since large percentages of the responses were collected either online via the web or over the phone, we have reason to doubt that the responses are truly random. On the other hand, the sample size (1212) is definitely less than 10% of the population of Japan. If we are to proceed with any further analysis we'll have to assume independence despite the questions we have about the researchers' sampling techniques.

Condition 2: **Success-Failure:** 372 people responded correctly (successes) while 840 people responded incorrectly (failures). Both amounts are greater than 10 so we may proceed.

Now we can calculate our confidence interval using the **inference** function:

```
inference(j12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

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




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

So our 95% confidence interval is (0.281, 0.3329)

To determine the margin of error, we calculate 1/2 the difference of the bounds of the confidence interval:

```
(.3329 - .281)/2
```

```
## [1] 0.02595
```

The calculation shows the margin of error to be 0.026 for Japanese respondents.

**France**

```
f12 <- subset(atheism, nationality == "France" & year == "2012")

# sum responses given to "atheist" vs. "non-atheist" survey question
summary(f12)
```

```
##       nationality          response         year
##  France     :1688   atheist    : 485   Min.   :2012
##  Afghanistan:   0   non-atheist:1203   1st Qu.:2012
##  Argentina  :   0                     Median :2012
##  Armenia    :   0                     Mean   :2012
##  Australia  :   0                     3rd Qu.:2012
##  Austria    :   0                     Max.   :2012
##  (Other)    :   0
```

We again use R's **summary** function to quickly sum up the "atheist" and "non-atheist" responses for respondents. As shown above, 1203/1688 respondents responded "non-atheist" while 485/1688 respondents responded "atheist".

So we have 485/1688 = .287 of French respondents responding as "atheist"

```
485/1688
```

```
## [1] 0.2873223
```

Now we'll state the conditions for inference:

Given: n = 1688; ˆp = .287

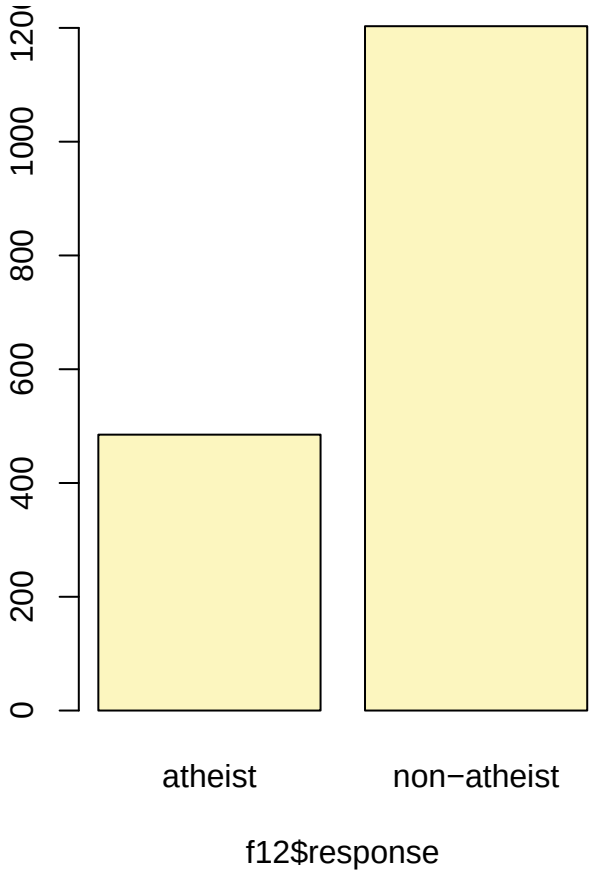
Condition 1: **Independence:** We have no indication that the respondents were randomly sampled. In fact, since large percentages of the responses were collected either online via the web or over the phone, we have reason to doubt that the responses are truly random. On the other hand, the sample size (1688) is definitely less than 10% of the population of France. If we are to proceed with any further analysis we'll have to assume independence despite the questions we have about the researchers' sampling techniques.

Condition 2: **Success-Failure:** 485 people responded correctly (successes) while 1203 people responded incorrectly (failures). Both amounts are greater than 10 so we may proceed.

Now we can calculate our confidence interval using the **inference** function:

```
inference(f12$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 )
```

So our 95% confidence interval is (0.2657, 0.3089)

To determine the margin of error, we calculate 1/2 the difference of the bounds of the confidence interval:

```
(.3089 - .2657)/2
```

```
## [1] 0.0216
```

The calculation shows the margin of error to be 0.0216

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

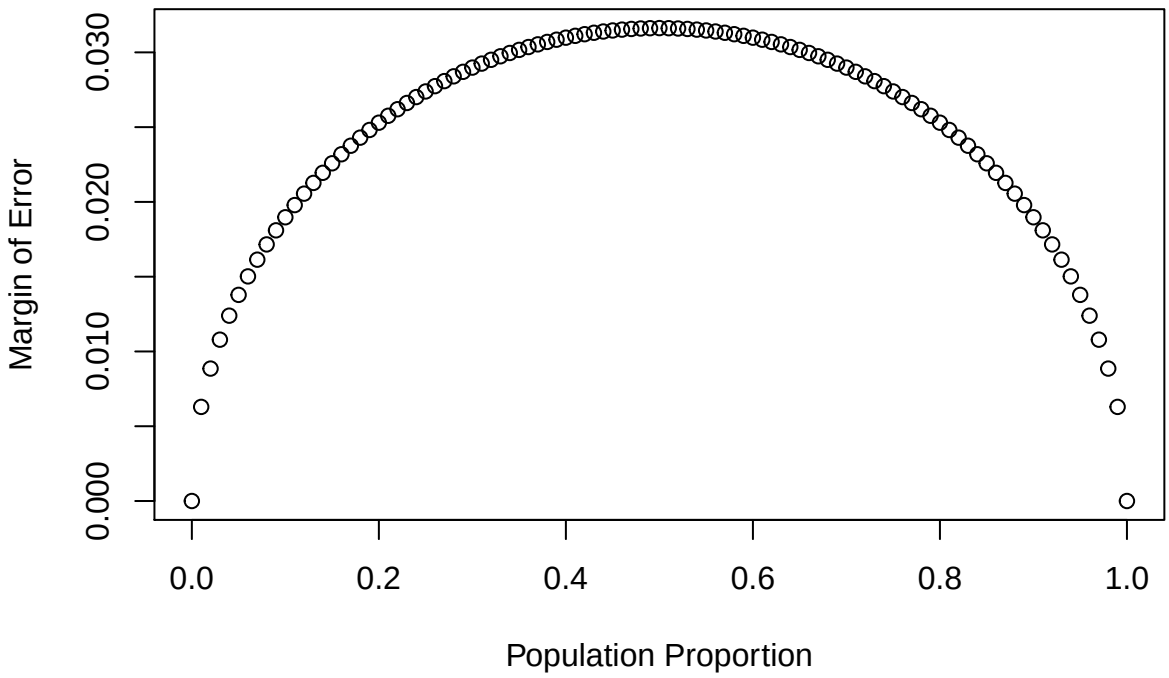
Imagine you've set out to survey 1000 people on two questions: are you female? and are you left-handed? Since both of these sample proportions were calculated from the same sample size, they should have the same margin of error, right? Wrong! While the margin of error does change with sample size, it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

# create a vector w/ values betw 0 and 1 stepped by .01
p <- seq(0, 1, 0.01)

# calculate the margin of error: 2 * standard error
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`.

At p = 0, ME = 0. As p approaches 0.5, ME increases. ME is maximized when p = 0.5. As p surpasses 0.5 and approaches 1, it decreases in a manner that mirrors its increase as it rose from 0 to 0.5.

Intuitively this makes sense: if the proportion of a population is in fact 0.5, we have less clarity about the certainty of our conclusions since the probability of any sample being either a "success" or a "failure" would be about equal. On the other hand, if the proportion is either less than or greater than 0.5, we have a higher

level of confidence in our conclusions since we'd have less of a chance of misinterpreting our results due to ambiguity.

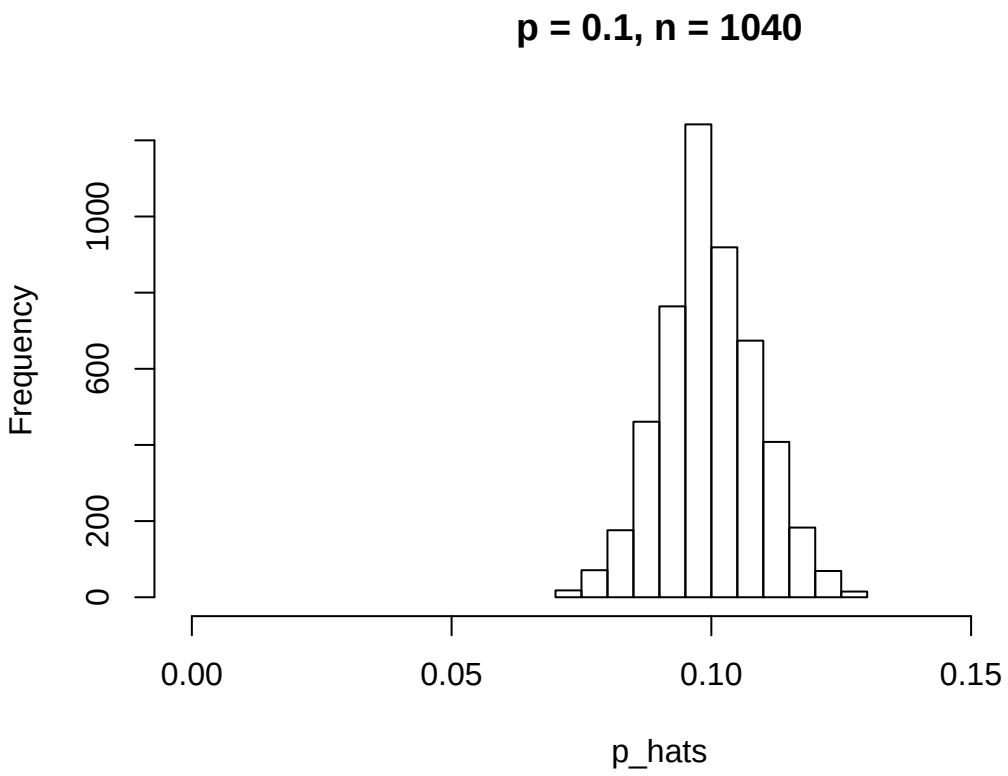
## Success-failure condition

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

# rep() function creates a vector of 5000 0's
p_hats <- rep(0, 5000)

for(i in 1:5000){
  # take a sample of size n with proportion p of "atheist"
  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.

We can use R's native **summary** and **IQR** functions to find the numeric values associated with the center and spread of the sampling distributions of sample proportions at n = 1040 and p = 0.1:

```
summary(p_hats)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.07019 0.09327 0.09904 0.09969 0.10580 0.12980
```

```
IQR(p_hats)
```

```
## [1] 0.0125
```

The sampling distribution of sample proportions as a mean and median of 0.10 and an interquartile range (IQR) of .0125. Per the histogram, the distribution appears to be approximately normal in shape.

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.

**N = 400, p = 0.1**

```
p <- 0.1
n <- 400

# rep() function creates a vector of 400 0's
p_hats2 <- rep(0, 5000)

for(i in 1:5000){
  # take a sample of size n with proportion p of "atheist"
  samp <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
  p_hats2[i] <- sum(samp == "atheist")/n
}
```

**N = 1040, p = .02**

```
p <- 0.02
n <- 1040

# rep() function creates a vector of 400 0's
p_hats3 <- rep(0, 5000)

for(i in 1:5000){
  # take a sample of size n with proportion p of "atheist"
  samp <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
  p_hats3[i] <- sum(samp == "atheist")/n
}
```

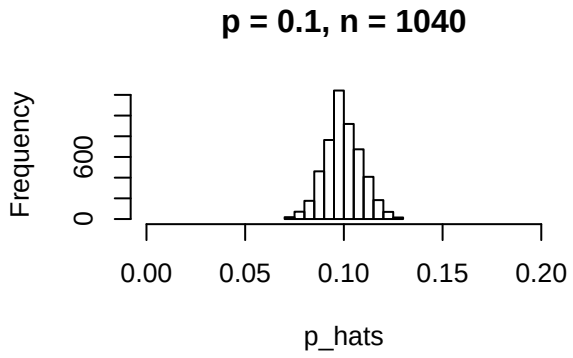
**N = 400, p = .02**

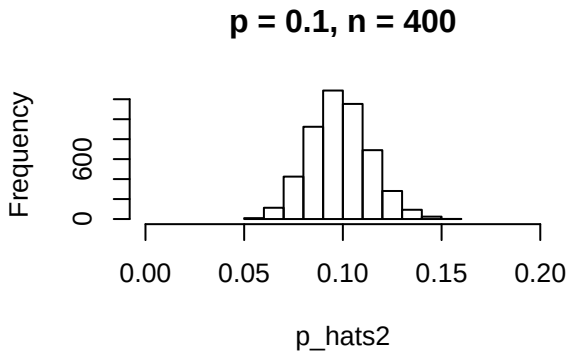
```
p <- 0.02
n <- 400

# rep() function creates a vector of 400 0's
p_hats4 <- rep(0, 5000)

for(i in 1:5000){
  # take a sample of size n with proportion p of "atheist"
  samp <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
  p_hats4[i] <- sum(samp == "atheist")/n
}
```

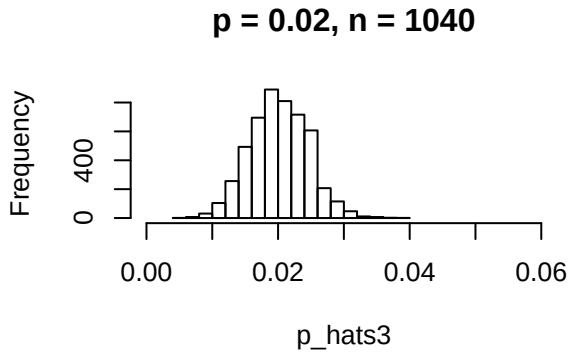
**Plot all 4 histograms together**

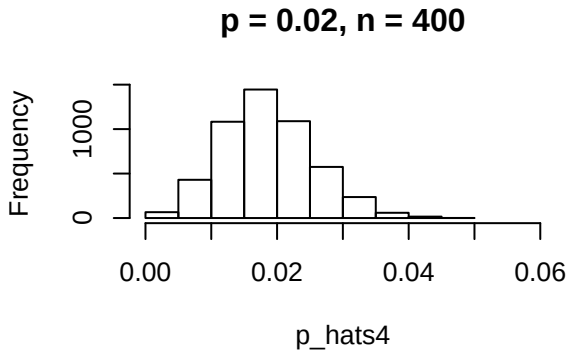
```
par(mfrow = c(2, 2))

hist(p_hats, main = "p = 0.1, n = 1040", xlim = c(0, 0.20))
```

```r
hist(p_hats2, main = "p = 0.1, n = 400", xlim = c(0, 0.20))
hist(p_hats3, main = "p = 0.02, n = 1040", xlim = c(0, 0.06))
hist(p_hats4, main = "p = 0.02, n = 400", xlim = c(0, 0.06))
```




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

**Describe the 3 new sampling distributions:**

**N = 400, p = .1**

```r
summary(p_hats2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.05250 0.09000 0.10000 0.09976 0.11000 0.15500
```

```r
IQR(p_hats2)
```

```
## [1] 0.02
```

The sampling distribution of sample proportions as a mean and median of 0.10 and an interquartile range (IQR) of .02. Per the histogram, the distribution appears to be approximately normal in shape.

**N = 1040, p = .02**

```
summary(p_hats3)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.005769 0.017310 0.020190 0.019950 0.023080 0.039420
```

```
IQR(p_hats3)
```

```
## [1] 0.005769231
```

The sampling distribution of sample proportions as a mean and median of 0.20 and an interquartile range (IQR) of .0057. Per the histogram, the distribution appears to be approximately normal in shape.

**N = 400, p = .02**

```
summary(p_hats4)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.01500 0.02000 0.01988 0.02500 0.04750
```

```
IQR(p_hats4)
```

```
## [1] 0.01
```

The sampling distribution of sample proportions as a mean and median of 0.20 and an interquartile range (IQR) of .01. Per the histogram, the distribution appears to be approximately normal in shape.

***Based on these limited plots, how does $n$ appear to affect the distribution of $\hat{p}$? How does $p$ affect the sampling distribution?***

Yes, $n$ does appear to affect the distribution of $\hat{p}$. The larger $n$ is, the narrower the distribution of $\hat{p}$ becomes. This can be seen in the histograms we've plotted as well as in the respective IQR's we've calculated. For example, the IQR for $(p = 0.02, n = 1040) = .0057$ while the IQR for $(p = 0.02, n = 400) = .01$. The larger the value for the IQR, the wider the spread of the distribution is.

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?

For Australia we have a sample proportion of 0.1 for a sample size of 1040. This tells us we have 104 respondents who answered the questions correctly. 1040 is definitely less than 10% of the population of Australia, and 104 is more than 10, so it seems sensible to proceed with inference and report margin of error for that country.

For Ecuador we have a sample proportion of 0.02 for a sample size of 400. This tells us we have 8 respondents who answered the questions correctly. 400 is definitely less than 10% of the population of Ecuador. However, $.02 * 400 = 8$ is less than 10. So should the researchers have proceeded with inference for Ecuador? It seems reasonable that they did since the "$np >= 10$" rule is rather arbitrary to begin with. The distribution appears to be nearly normal according to the histogram we plotted above, and that is the condition the "$np >= 10$" rule is meant to address.

---

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

---

**a)**

**Hypothesis:**

H0: P2012 = p2005 => Spain's atheism index **has not** changed between 2005 and 2012

H1: P2012 != p2005 => Spain's atheism index **has** changed between 2005 and 2012

To start, we'll create separate subsets of data for Spain for the years 2005 and 2012:

```
# create subset for 2005 data
s05 <- subset(atheism, nationality == "Spain" & year == "2005")

# sum responses given to "atheist" vs. "non-atheist" survey question
summary(s05)
```

```
##       nationality          response         year
##  Spain      :1146   atheist    : 115   Min.   :2005
##  Afghanistan:   0   non-atheist:1031   1st Qu.:2005
##  Argentina  :   0                     Median :2005
##  Armenia    :   0                     Mean   :2005
##  Australia  :   0                     3rd Qu.:2005
##  Austria    :   0                     Max.   :2005
##  (Other)    :   0
```

```
# create subset for 2012 data
s12 <- subset(atheism, nationality == "Spain" & year == "2012")

# sum responses given to "atheist" vs. "non-atheist" survey question
summary(s12)
```

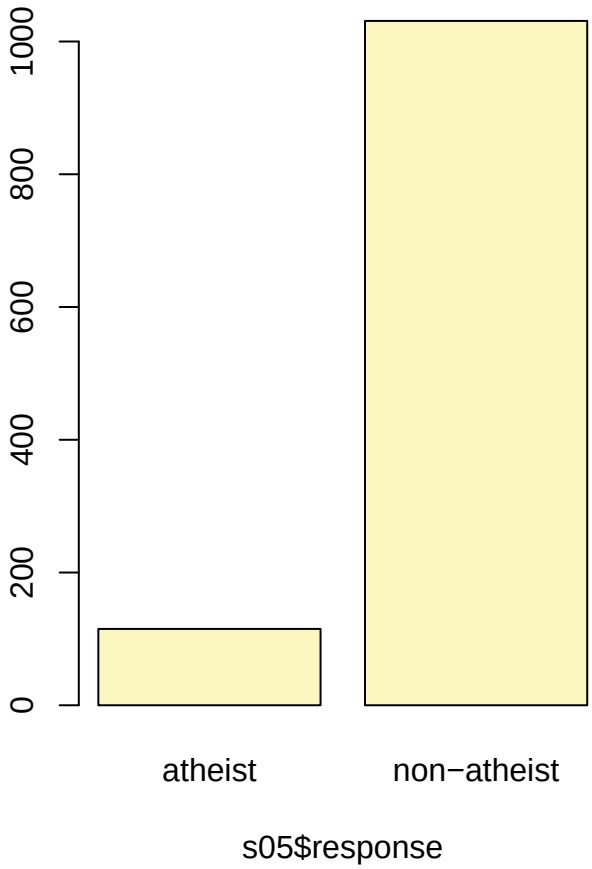
```
##       nationality          response         year
##  Spain      :1145   atheist    : 103   Min.   :2012
##  Afghanistan:   0   non-atheist:1042   1st Qu.:2012
##  Argentina  :   0                     Median :2012
##  Armenia    :   0                     Mean   :2012
```

```
##  Australia  :   0                    3rd Qu.:2012
##  Austria    :   0                    Max.   :2012
##  (Other)    :   0
```

Now we can calculate our confidence interval for Spain for 2005 using the **inference** function:

```
inference(s05$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

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



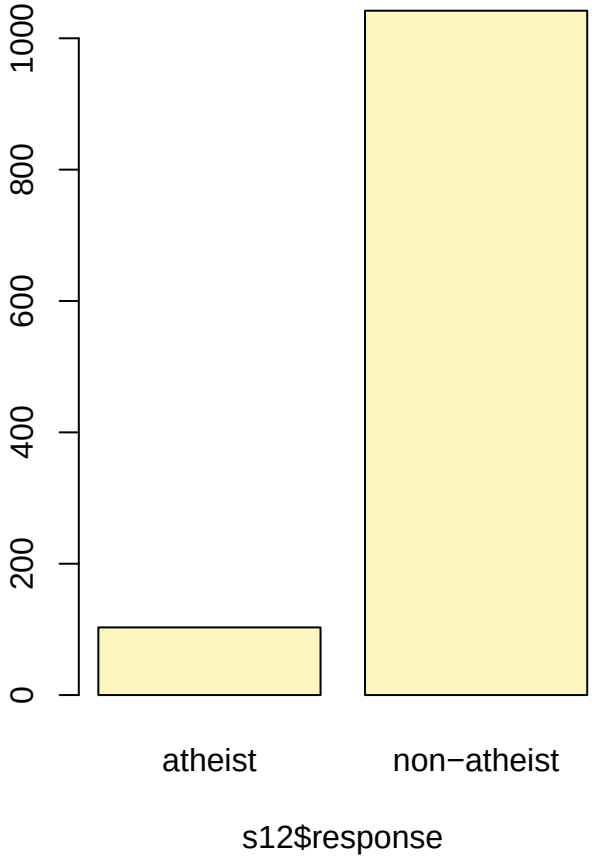

```
## p_hat = 0.1003 ;  n = 1146
## Check conditions: number of successes = 115 ; number of failures = 1031
## Standard error = 0.0089
## 95 % Confidence interval = ( 0.083 , 0.1177 )
```

So our 95% confidence interval for Spain for 2005 is (0.083, 0.1177)

For 2012:

```
inference(s12$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 )
```

So our 95% confidence interval for Spain for 2012 is (0.0734, 0.1065)

Therefore, the confidence intervals do overlap. When overlap occurs, we must fail to reject the null hypothesis that Spain's atheism index has not changed between 2005 and 2012.

---

**b)**

**Hypothesis:**

H0: P2012 = p2005 => The US's atheism index **has not** changed between 2005 and 2012

H1: P2012 != p2005 => The US's atheism index **has** changed between 2005 and 2012

To start, we'll create separate subsets of data for the US for the years 2005 and 2012:

```r
# create subset for USA 2005 data
us05 <- subset(atheism, nationality == "United States" & year == "2005")

# sum responses given to "atheist" vs. "non-atheist" survey question
summary(us05)
```

```
##         nationality          response        year
##  United States:1002   atheist    : 10   Min.   :2005
##  Afghanistan  :   0   non-atheist:992   1st Qu.:2005
##  Argentina    :   0                     Median :2005
##  Armenia      :   0                     Mean   :2005
##  Australia    :   0                     3rd Qu.:2005
##  Austria      :   0                     Max.   :2005
##  (Other)      :   0
```

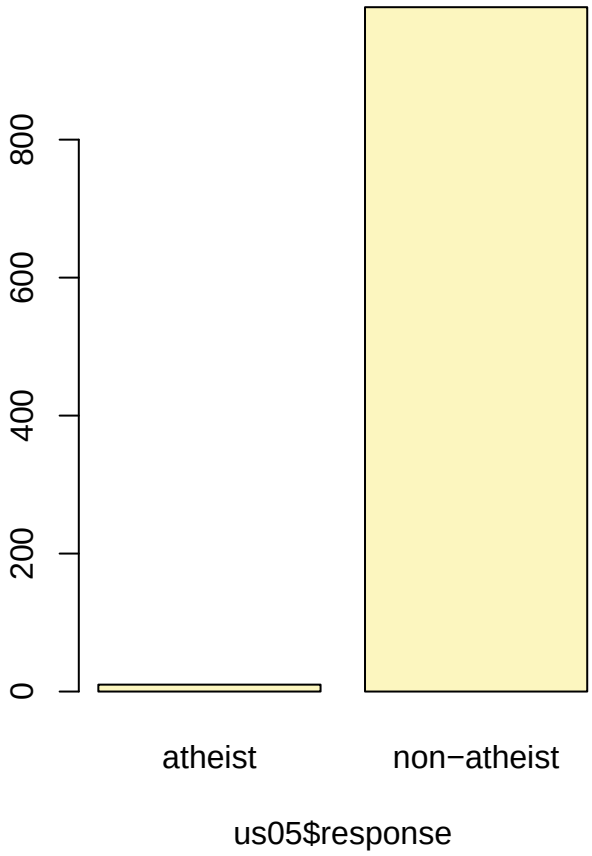
```r
# create subset for USA 2012 data
us12 <- subset(atheism, nationality == "United States" & year == "2012")

# sum responses given to "atheist" vs. "non-atheist" survey question
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
```

Now we can calculate our confidence interval for the USA for 2005 using the **inference** function:

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

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



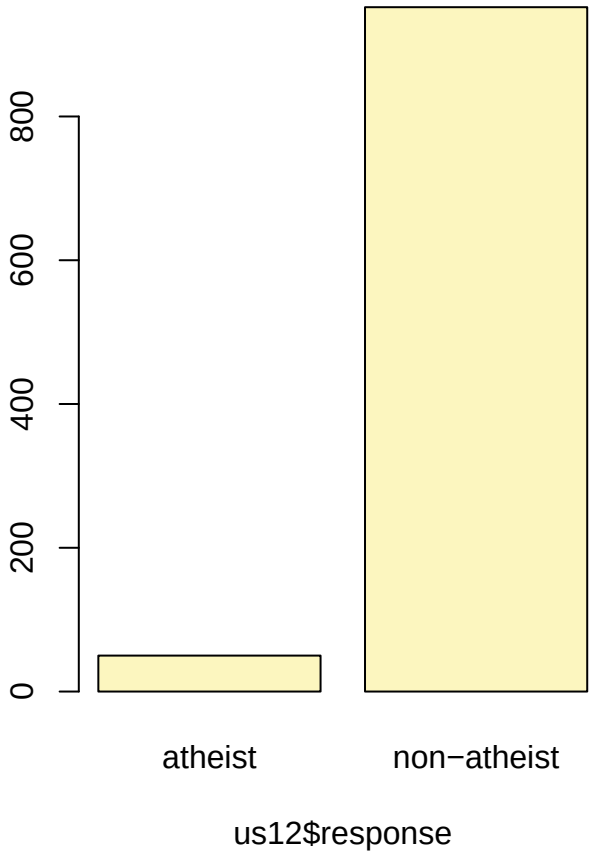

```
## p_hat = 0.01 ;  n = 1002
## Check conditions: number of successes = 10 ; number of failures = 992
## Standard error = 0.0031
## 95 % Confidence interval = ( 0.0038 , 0.0161 )
```

So our 95% confidence interval for the USA for 2005 is (0.0038, 0.0161)

For 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 )
```

So our 95% confidence interval for the USA for 2012 is (0.0364, 0.0634)

Therefore, the confidence intervals **do not** overlap. Since no overlap occurs, we must reject the null hypothesis since there is convincing evidence that the US's atheism index has changed 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.

There are 57 countries for which an atheism index was estimated. A Type 1 error occurs when the null hypothesis is true and we reject it. So if our null hypothesis is: "There is no change in the atheism index between 2005 and 2012" and we apply a .05 significance level, we would expect to detect a change in the atheism index for 5% of the 57 countries, which is 3 countries (rounding up the result of $.05 * 57$)

```
.05 * 57
```

```
## [1] 2.85
```

- 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.

Margin of error <= .01, 95% confidence interval

Since we don't know anything about $p$, we'll have to assume $p = 0.5$ to be conservative since the margin of error is largest when $p = 0.5$.

To find the appropriate sample size we plug our values for ME and p, and the critical value for a 95% confidence level (1.96) into the equation:

$ME = Criticalvalue * sqrt((p * (1 - p))/n)$

and solve for n.

$.01 = 1.96 * sqrt((0.5 * (1 - 0.5))/n) =>$

$(.01)^2 = 1.96^2 * (0.5 * (1 - 0.5))/n =>$

```
## [1] 1e-04
```

```
## [1] 3.8416
```

$.0001 = 3.8416 * (.25/n) =>$

$n * .0001 = 3.8416 * .25 =>$

$n = (3.8416 * .25)/.0001 =>$

```
3.8416 * .25 / .0001
```

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

$n = 9604$