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:

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*?

These percentages appear to be sample statistics since the data was collected from a 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 a random sample was taken and that each observation is independent. This seems like a reasonable assumption as long as each population of a country would be represented by a random sample that has a big enough sample size to represent every country.

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 correspond to a country and the results from the survey for each of them. Each row of atheism correspond to a single individual response in 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?

```
us12 <- subset(atheism, nationality == "United States" & year == "2012")
us12$nationality <- as.factor(as.character(us12$nationality))
us12_proportion <- prop.table(table(us12$nationality, us12$response))
us12_proportion
```

```
##
##                    atheist non-atheist
##   United States 0.0499002   0.9500998
```

The proportion of atheist responses in the United States is 0.0499002 and does agree with the percentage in Table 6 if we round.

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.

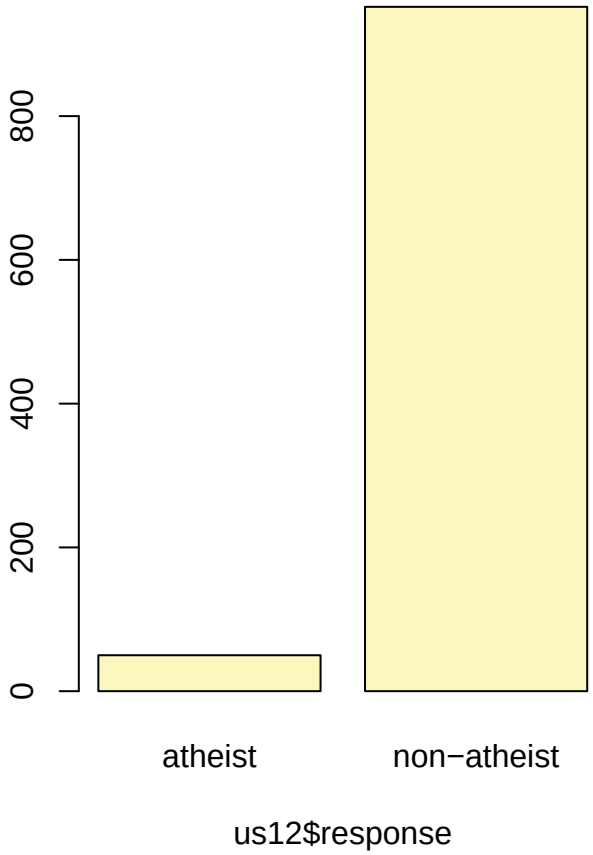
5. Write out the conditions for inference to construct a 95% confidence interval for the proportion of atheists in the United States in 2012. Are you confident all conditions are met?

The conditions for inference to construct a 95% confidence interval are each observation must be independent and the distribution is nearly normal or the sample size is large. I am confident all conditions are met since random samples were used and the sample is less than 10% of the population. This shows independence. Also the sample size is 1002, which is large enough to assume a 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:
```




```
## 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?

```
(0.0634 - 0.0364) / 2
```

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

The margin of error for the estimate of the proportion of atheists in US in 2012 is the upper bound of the confidence interval - the lower bound of the confidence interval / 2, which is equal to 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.

```
japan12 <- subset(atheism, nationality == "Japan" & year == "2012")
japan12$nationality <- as.factor(as.character(japan12$nationality))
japan12_proportion <- prop.table(table(japan12$nationality, japan12$response))
japan12_proportion
```

```
##
##           atheist non-atheist
##   Japan 0.3069307   0.6930693
```

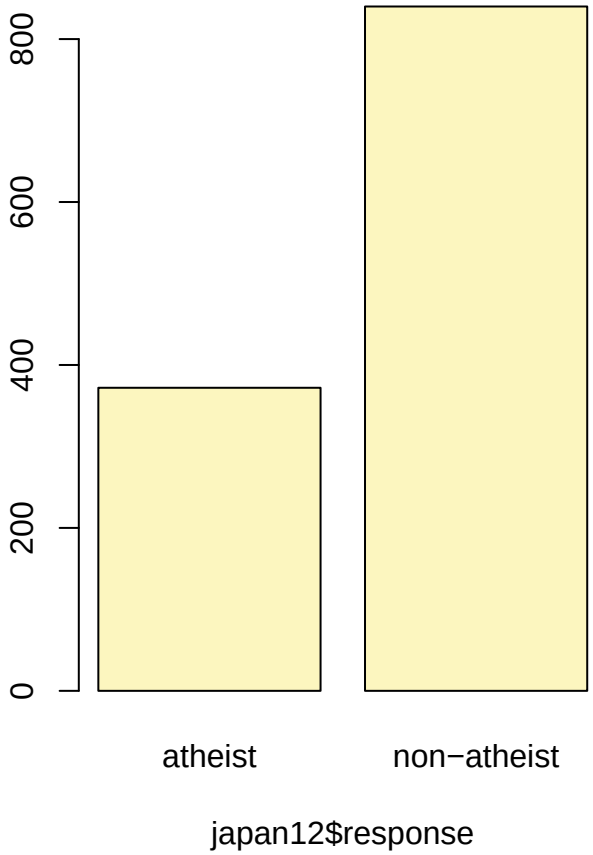
```
germany12 <- subset(atheism, nationality == "Germany" & year == "2012")
germany12$nationality <- as.factor(as.character(germany12$nationality))
germany12_proportion <- prop.table(table(germany12$nationality, germany12$response))
germany12_proportion
```

```
##
##             atheist non-atheist
##   Germany 0.1494024   0.8505976
```

The two countries that I chose are Japan and Germany. Both of them meet the conditions for inference since Japan has a sample size of 1200 and Germany has a sample size of 502. Both have a large enough sample size and the sample size is less than 10% of the population.

```
inference(japan12$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 )
```

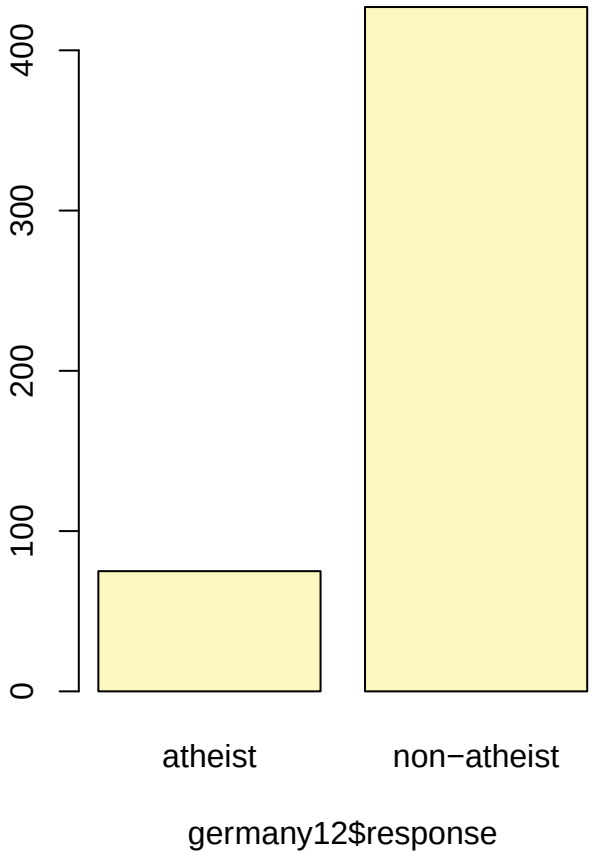
```
# margin of error for Japan
(0.3329 - 0.281) / 2
```

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

The margin of error for Japan is 0.02595.

```
inference(germany12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

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




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

```
# margin of error for Germany
(0.1806 - 0.1182) / 2
```

```
## [1] 0.0312
```

The margin of error for Germany is 0.0312.

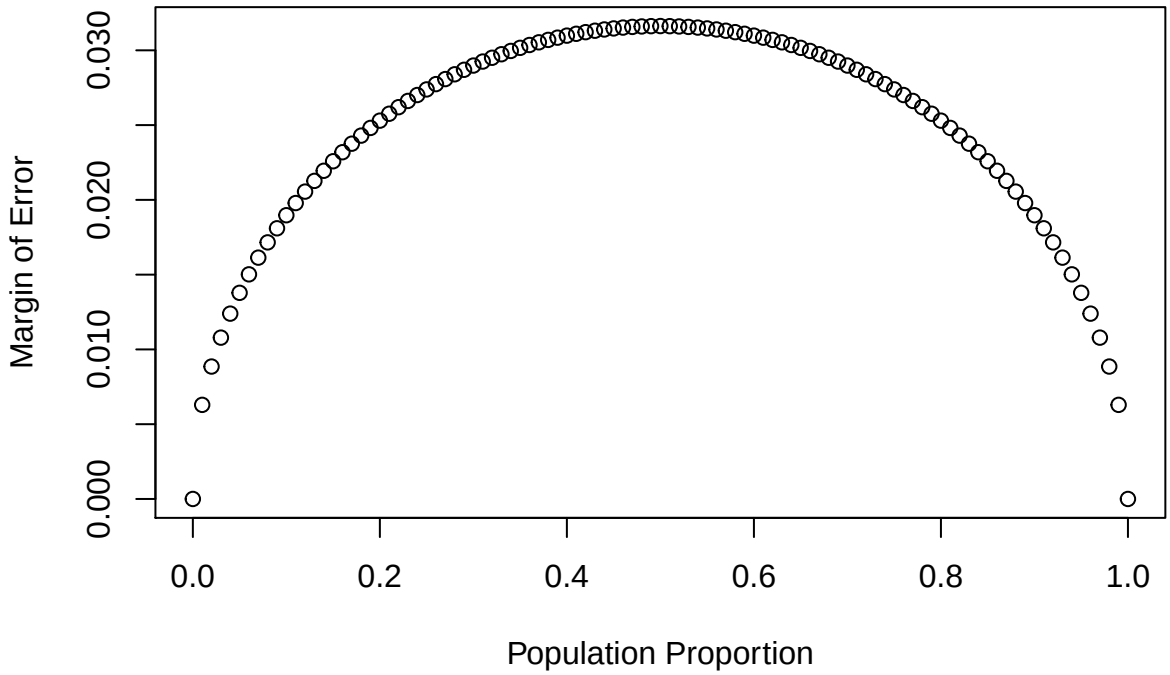
How does the proportion affect the margin of error?

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 relationship between p and me is an upside-down parabola from range 0 to 1, where the maximum point is p = 0.5, me = 0.030. As p increases from 0 to 0.5, me also increases. But when p increases from 0.5 to 1, me decreases.

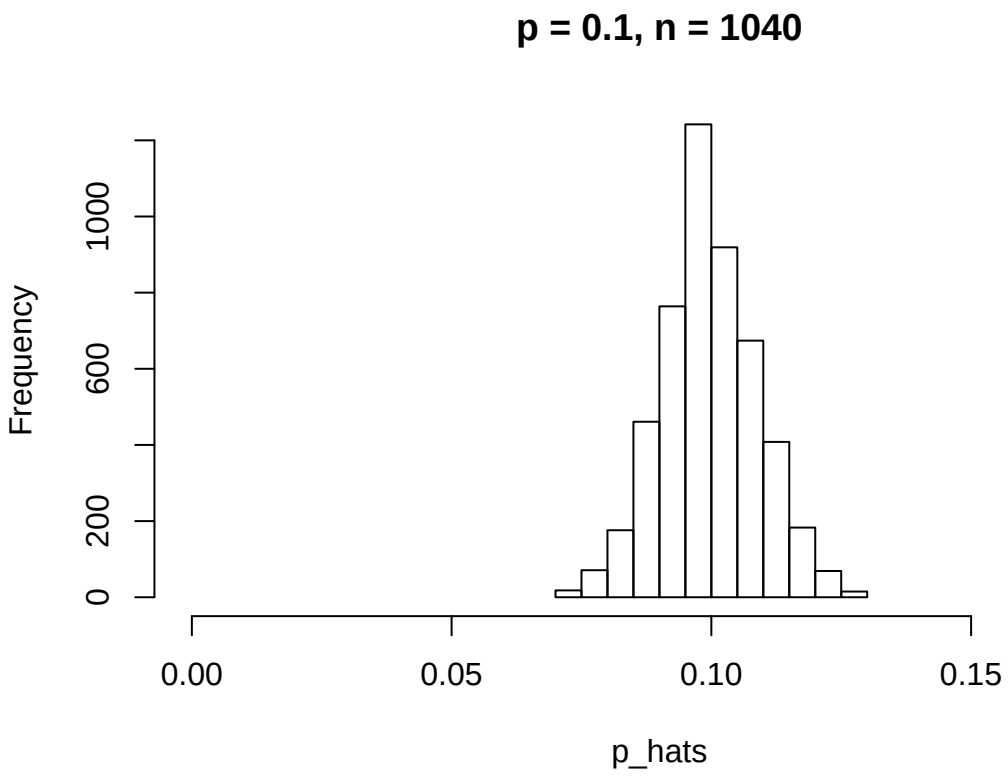
Success-failure condition

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.

```
summary(p_hats)
```

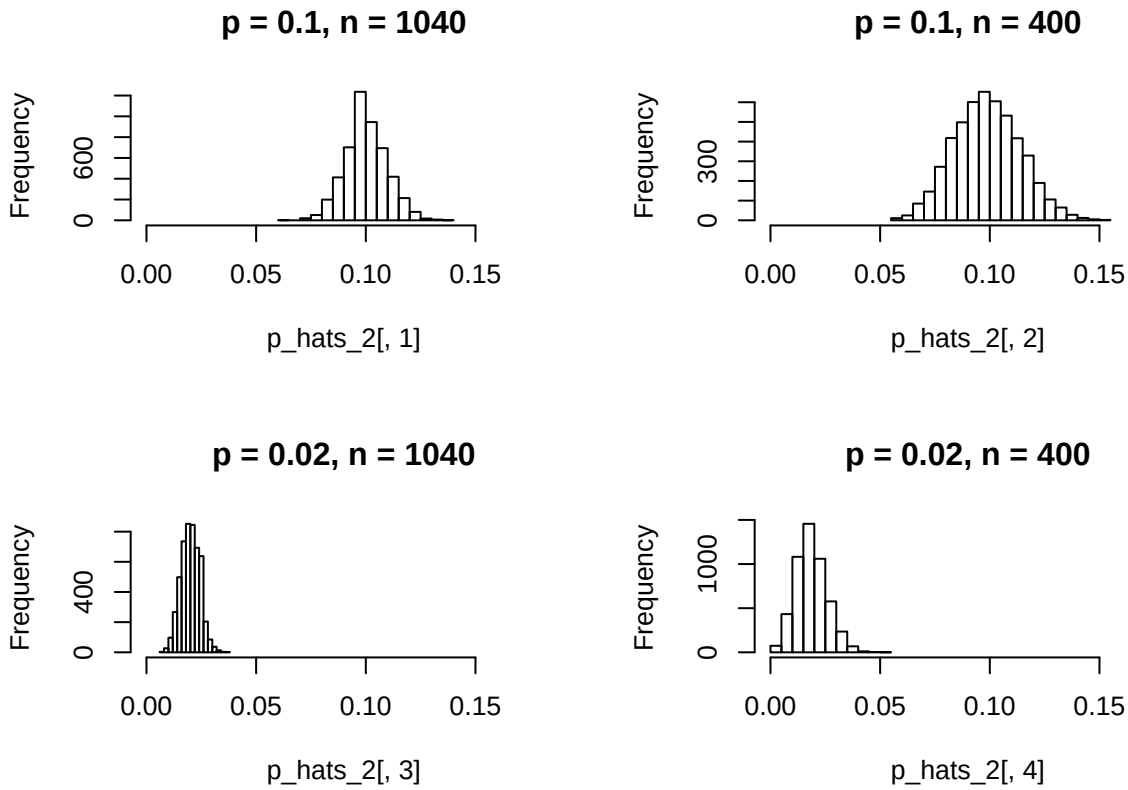
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.07019 0.09327 0.09904 0.09969 0.10577 0.12981
```

The sampling distribution of sample proportions n = 1040 and p = 0.1 is fairly normal and symmetric. The distribution is centered at mean = 0.09969.

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?

```
p_2 <- c(0.1, 0.1, 0.02, 0.02)
n_2 <- c(1040, 400, 1040, 400)
p_hats_2 <- data.frame(c(rep(0, 5000)), c(rep(0, 5000)), c(rep(0, 5000)), c(rep(0, 5000)))

for(i in 1:5000){
  samp <- sample(c("atheist", "non_atheist"), n_2[1], replace = TRUE, prob = c(p_2[1], 1-p_2[1]))
  p_hats_2[i, 1] <- sum(samp == "atheist")/n_2[1]

    samp <- sample(c("atheist", "non_atheist"), n_2[2], replace = TRUE, prob = c(p_2[2], 1-p_2[2]))
  p_hats_2[i, 2] <- sum(samp == "atheist")/n_2[2]

    samp <- sample(c("atheist", "non_atheist"), n_2[3], replace = TRUE, prob = c(p_2[3], 1-p_2[3]))
  p_hats_2[i, 3] <- sum(samp == "atheist")/n_2[3]

    samp <- sample(c("atheist", "non_atheist"), n_2[4], replace = TRUE, prob = c(p_2[4], 1-p_2[4]))
  p_hats_2[i, 4] <- sum(samp == "atheist")/n_2[4]

}
par(mfrow = c(2, 2))
hist(p_hats_2[, 1], main = "p = 0.1, n = 1040", xlim = c(0, 0.18))
hist(p_hats_2[, 2], main = "p = 0.1, n = 400", xlim = c(0, 0.18))
hist(p_hats_2[, 3], main = "p = 0.02, n = 1040", xlim = c(0, 0.18))
hist(p_hats_2[, 4], main = "p = 0.02, n = 400", xlim = c(0, 0.18))
```



All distributions except the last one appear to be fairly normal and symmetric. The last one is fairly normal but has some skew. N appears to affect the distribution of hat(p) by it affects the spread of the distribution. P affects the sampling distribution by it affects the center.

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

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

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?

It is not sensible to proceed with inference and report margin of errors for Ecuador because Ecuador has a sample proportion of 0.02 on 400 subjects, so the number of atheits in the sample is 8. This is too low of a sample to assume normal distribution. It is sensible to proceed with inferencce and report margin of errors for Australia because Australia has a sample proportion of 0.1 on a sample size of 1000, so the number of atheists in the sample is 104. This is a large enough sample to assume normal distribution.

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.

```
spain05 <- subset(atheism, nationality == "Spain" & year == "2005")
spain05$nationality <- as.factor(as.character(spain05$nationality))
table(spain05$nationality, spain05$response)
```

```
##
##         atheist non-atheist
##   Spain     115        1031
```

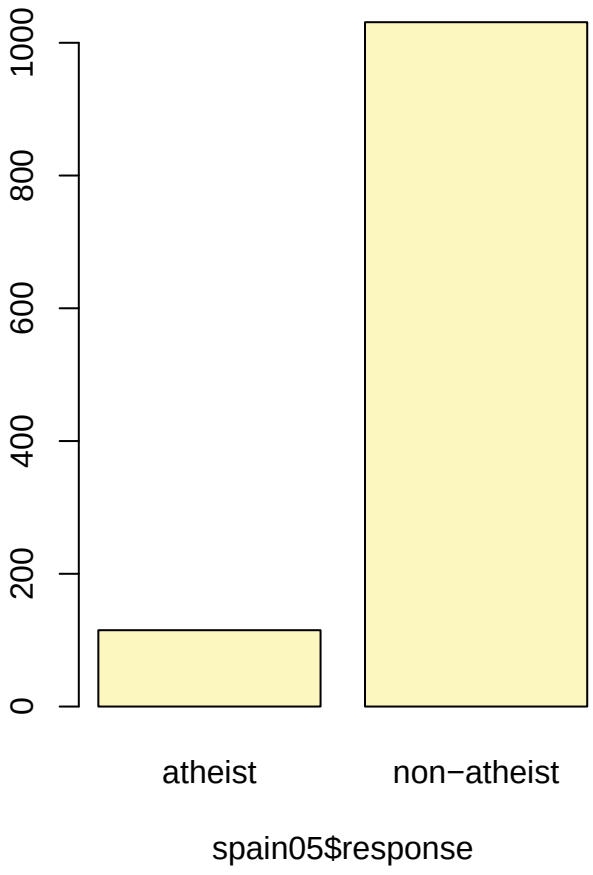
```
spain12 <- subset(atheism, nationality == "Spain" & year == "2012")
spain12$nationality <- as.factor(as.character(spain12$nationality))
table(spain12$nationality, spain12$response)
```

```
##
##         atheist non-atheist
##   Spain     103        1042
```

We can assume normal distribution for both year 2005 and year 2012 for Spain since both have a large enough sample size. Number of atheists in 2005 is 115 and number of atheists in 2012 is 103.

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

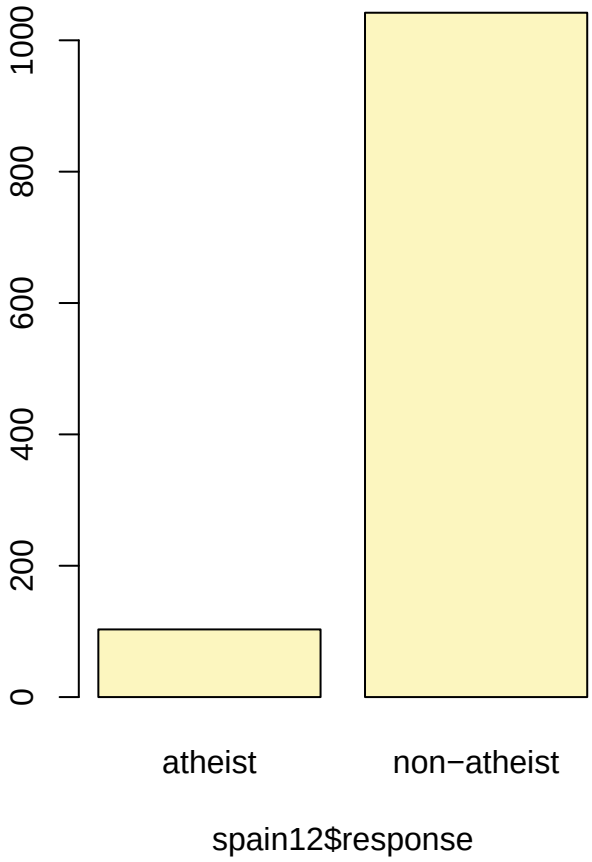
```
## Single proportion -- success: atheist
## Summary statistics:
```



```
## 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 )
```

There is an overlap for the confidence interval for the sample size for Spain in the year 2005 and the sample size for Spain in the year 2012. For the year 2005, the confidence interval is (0.083, 0.1177) and for the year 2012 the confidence interval is (0.0734, 0.1065). There is no convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012 since the confidence intervals overlap. This tells us that we cannot reject the null hypothesis and the change in atheism from the year 2005 to 2012 happened by chance.

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

```
us05 <- subset(atheism, nationality == "United States" & year == "2005")
us05$nationality <- as.factor(as.character(us05$nationality))
table(us05$nationality, us05$response)
```

```
##
##                 atheist non-atheist
##   United States      10         992
```

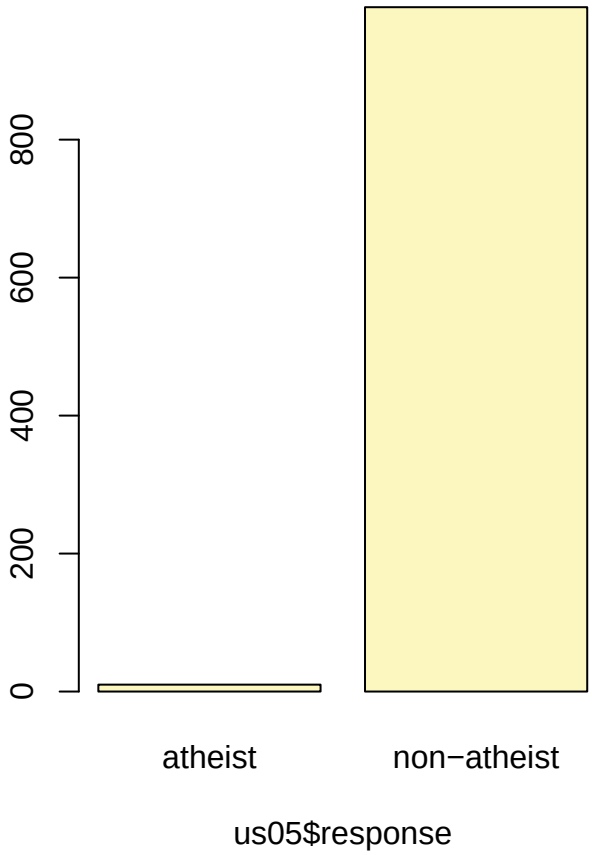
```
# Show year 2012 that we calculated above
table(us12$nationality, us12$response)
```

```
##
##                 atheist non-atheist
##   United States      50         952
```

We can assume normal distribution for year 2005 since the sample size is barely large enough. Number of atheists in the US for year 2005 is 10 and number of atheists in the US for the year 2012 is 50.

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

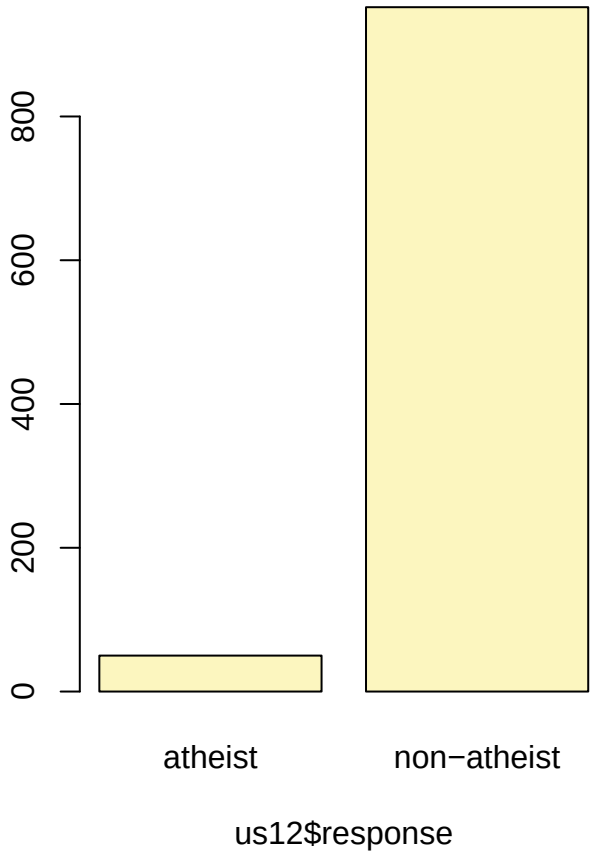
```
## Single proportion -- success: atheist
## Summary statistics:
```




```
## 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 )
```

```
# Let's reminds us of the inferene for year 2012
inference(us12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

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




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

There is no overlap for the confidence interval for the sample size for the US in the year 2005 and the sample size for the US in the year 2012. For the year 2005, the confidence interval is (0.0038, 0.0161) and for the year 2012 the confidence interval is (0.0364, 0.0634). There is convincing evidence that the US has seen a change in its atheism index between 2005 and 2012 since the confidence intervals do not overlap. This tells us we can reject the null hypothesis.

- 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.

```
round(0.05 * 39)
```

```
## [1] 2
```

If there has been no change in the atheism index in the countries listed in Table 4, we would expect to detect a change in two coutries at a significance level of 0.05 simply by chance.

- 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.

```
(0.5 * 0.5) / (0.01 / 1.96)^ 2
```

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

You would have to sample 9604 people to ensure you are within the guidelines. Since you do not know what to expect for p, you assume the worst case scenario for p, which is p = 0.5. You want a margin of error no greater than 1% or 0.01. Since you want a 95% confidence, you use z = 1.96.