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 are sample statistics since they were derived from 50,000 men and women selected from 57 countries across the globe. The population in this report is every human on earth.

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?

To generalize the report's findings to the global human population, you'd have to assume the following:

- **The sample was established based on random sampling**
- **The sample size is large enough (greater than 30)**
- **The observations are independent from one another**

It states that a probability sample was used, therefore, this confirms that the sample was created from random selection. Because of this, and the other factors above, we can assume that the findings are generalizable.

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 corresponds to the total unweighted sample of respondents per country. This isn't the full dataset of respondents, but an unweighted sample from each country. Each row of the 'atheism' data frame corresponds to an individual response for every respondent in the study. While Table 6 breaks down respondents into country, atheism has each individual response on its own row (88,032 in total).

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")

atheist_us12 <- subset(us12, response == "atheist")

percentage_check_t6 <- length(atheist_us12$nationality) / length(us12$nationality)

percentage_check_t6
```

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

After calculating the proportion, I've arrived at approximately 4.99%, which is very close to what is reported in Table 6 – at 5%. The reason why they disagree slightly is due to Table 6 being a sample of the entire 'atheism' dataset. Table 6 has 51,927 observations, while 'atheism' has 88,032 observations.

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 intervale for the proportion of atheists in the United States in 2012 are the following:

- **The data needs to come from a random sample.**
- **There needs to be at least 10 expected successes and 10 expected failures.**
- **The individual observations need to be independent from one another.**
- **The sample size of the population should not be more than 10% of the population.**

Since we know this is from a random sample, this part is met. The individual observations are also independente from one another, and a sample size of 1002 is definitely less than 10% of the population of the United States. In order to test the final condition, we can do the following calculation:

$np > or = 10 and n(1-p) > or = 10$

```
x <- 1002 * 0.05
y <- 1002 * (1 - 0.05)

sum <- x + y
x
```

```
## [1] 50.1
```

```
y
```

```
## [1] 951.9
```

```
sum
```

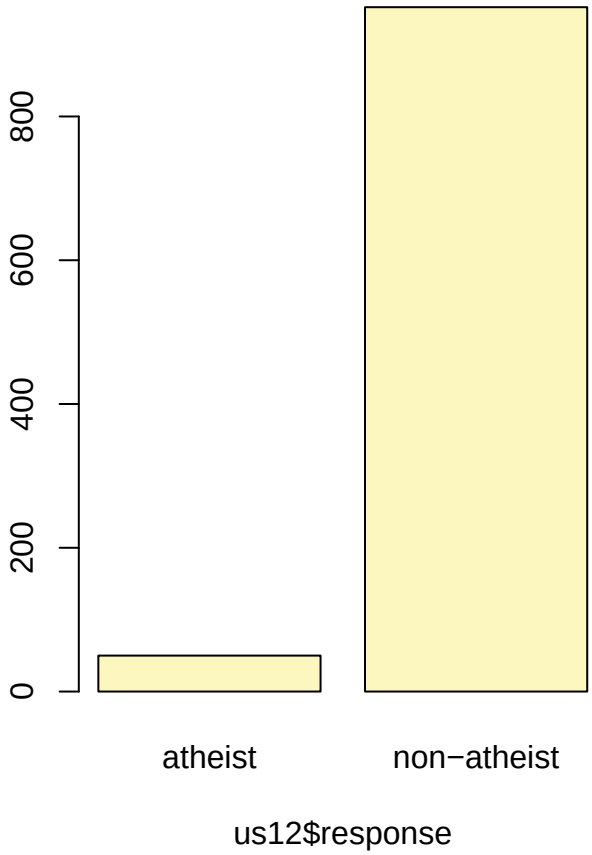
```
## [1] 1002
```

Since both of these are greater than or equal to 10, all conditions are 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?

The margin of error for the estimate of the proportion of atheists in the US in 2012 is 0.014:

```
moe <- 1.96 * (0.0069)
moe
```

```
## [1] 0.013524
```

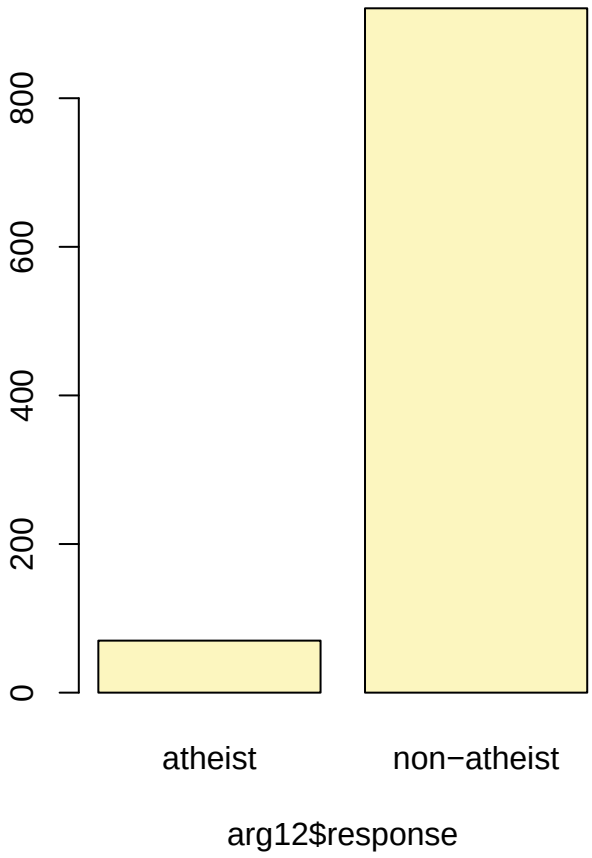
7. Using the `inference` function, calculate confidence intervals for the proportion of atheists in 2012 in two other countries of your choice, and report the associated margins of error. Be sure to note whether the conditions for inference are met. It may be helpful to create new data sets for each of the two countries first, and then use these data sets in the `inference` function to construct the confidence intervals.

The two other countries that I chose to calculate confidence intervals are Argentina and Spain.

```
arg12 <- subset(atheism, nationality == "Argentina" & year == "2012")
spain12 <- subset(atheism, nationality == "Spain" & year == "2012")


inference(arg12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

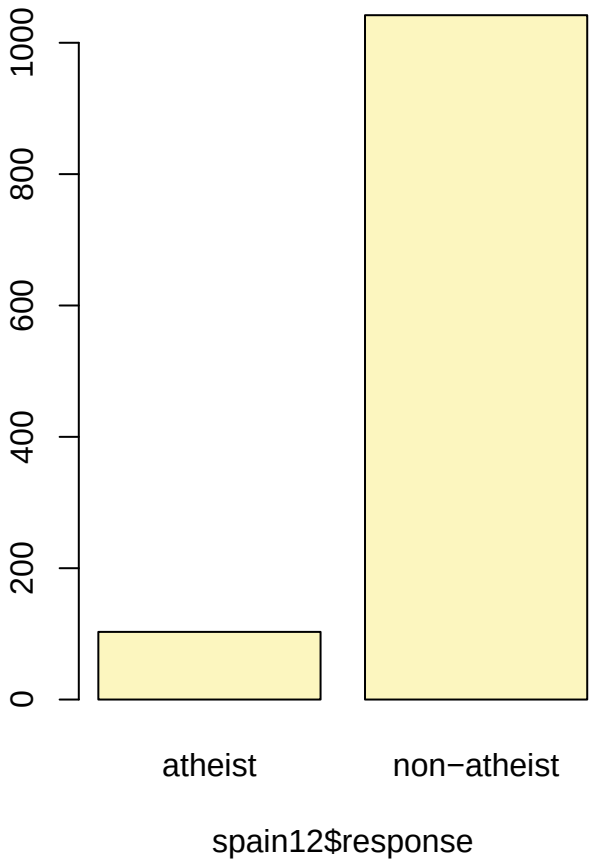
```
## Single proportion -- success: atheist
## Summary statistics:
```




```
## p_hat = 0.0706 ;  n = 991
## Check conditions: number of successes = 70 ; number of failures = 921
## Standard error = 0.0081
## 95 % Confidence interval = ( 0.0547 , 0.0866 )
```

```
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 )
```

```
moe_arg <- 1.96 * (0.0081)
moe_arg
```

```
## [1] 0.015876
```

```
moe_spain <- 1.96 * (0.0085)
moe_spain
```

```
## [1] 0.01666
```

The confidence interval for Argentina is (0.0547, 0.0866). The margin of error is +/- 0.159.

The confidence interval for Spain is (0.0734, 0.1065). The margin of error is +/- 0.0167.

The conditions for inference are met for both of these countries. However, if I were to try and find the confidence interval and margin of error for Vietnam, I would not have 10 or more successes and failures to utilize for this calculation (in this example, respondents that are atheist).

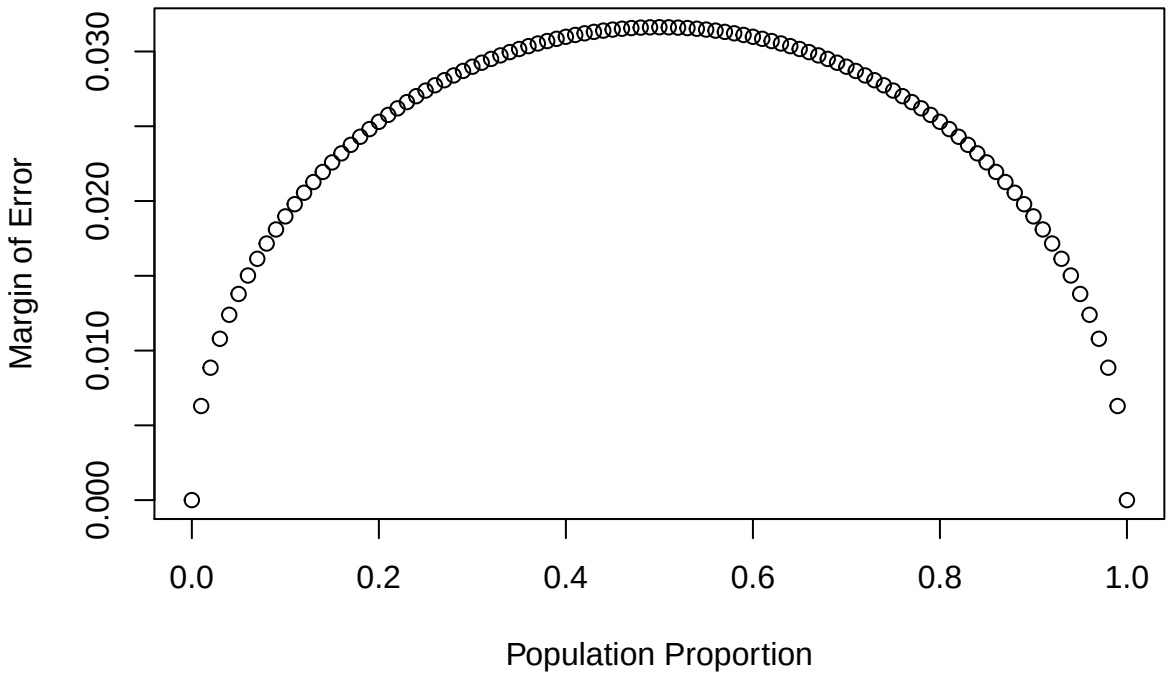
How does the proportion affect the margin of error?

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 the population proportion moves closer to an even split, the margin of error increases. The margin of error is the highest when the population proportion is split evenly, and is the lowest when the population is split entirely one way over the other.

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?

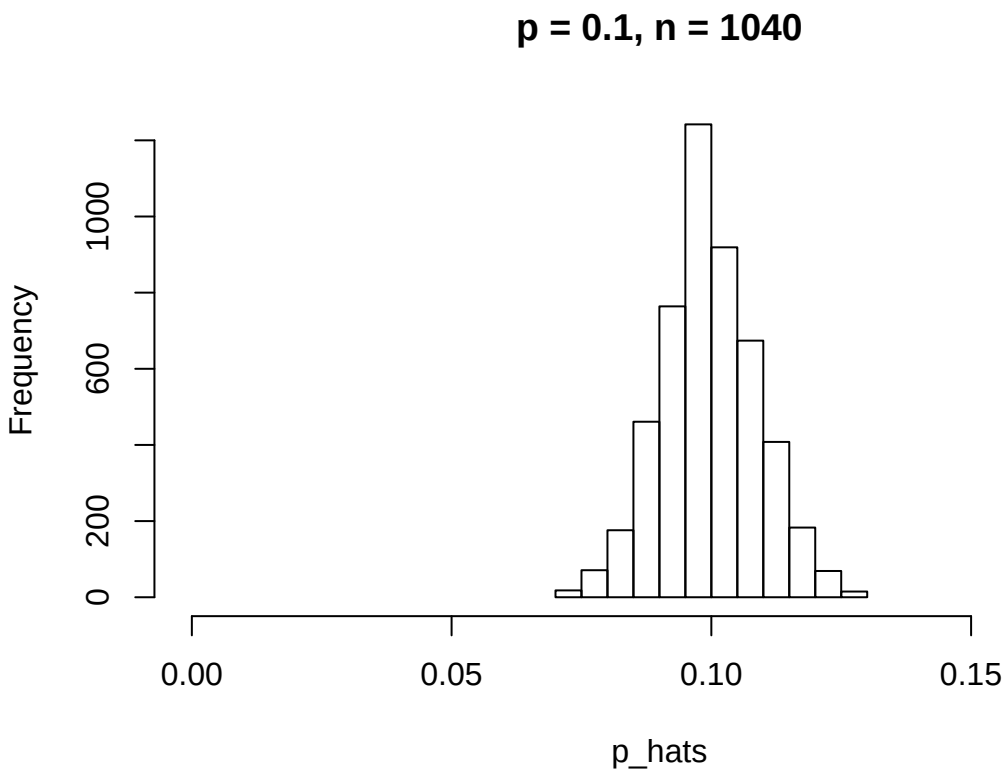
The short answer is: nothing. You could argue that we would be fine with 9 or that we really should be using 11. What is the "best" value for such a rule of thumb is, at least to some degree, arbitrary. However, when np and $n(1-p)$ reaches 10 the sampling distribution is sufficiently normal to use confidence 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.

```
library(psych)
summary(p_hats)
```

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

```
describe(p_hats)
```

```
##    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
```

The center of the distribution is right around 0.1 (which is the probability). The distribution is unimodal and approaching normal. There isn't much spread, and the mean and median are very close to one another with a standard deviation of 0.01.

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?

```
# simulation 2
p_2 <- 0.1
n_2 <- 400
p_hats2 <- rep(0, 5000)

for(i in 1:5000){
  samp_2 <- sample(c("atheist", "non_atheist"), n_2, replace = TRUE, prob = c(p_2, 1-p_2))
  p_hats2[i] <- sum(samp_2 == "atheist")/n_2
}
```

```
# simulation 3
p_3 <- 0.02
n_3 <- 1040
p_hats3 <- rep(0, 5000)

for(i in 1:5000){
  samp_3 <- sample(c("atheist", "non_atheist"), n_3, replace = TRUE, prob = c(p_3, 1-p_3))
  p_hats3[i] <- sum(samp_3 == "atheist")/n_3
}
```

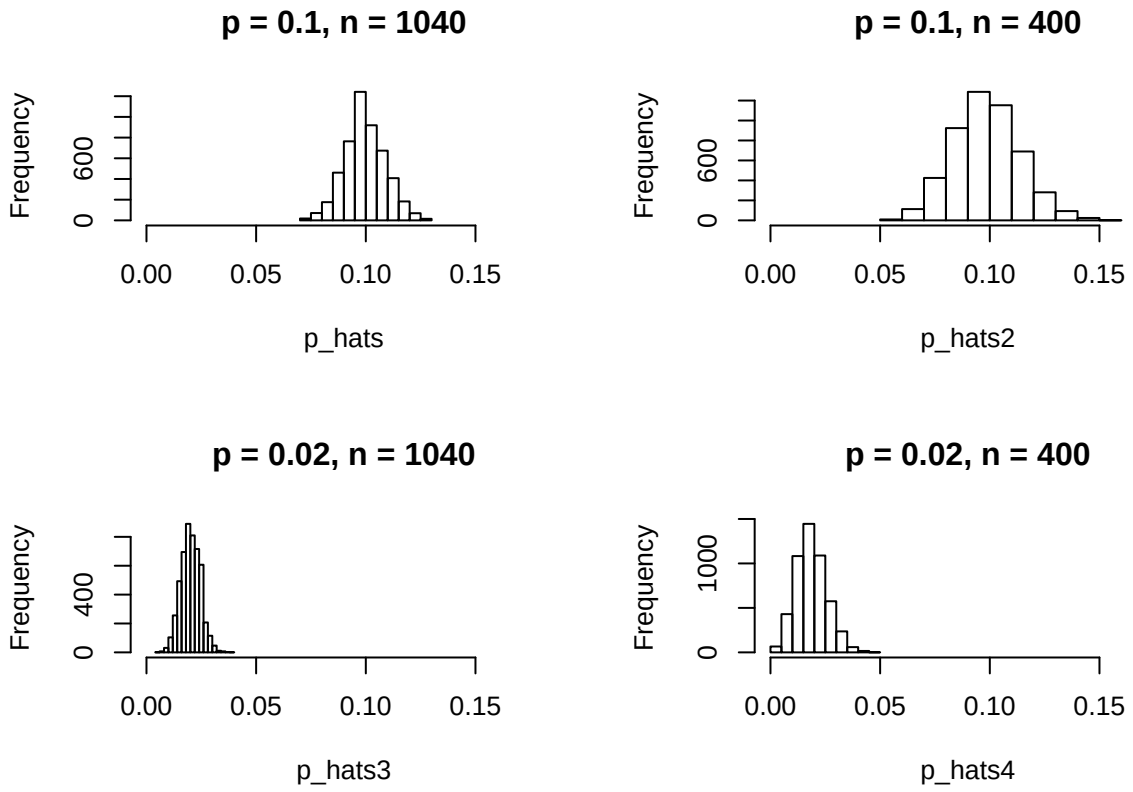
```r
# simulation 4
p_4 <- 0.02
n_4 <- 400
p_hats4 <- rep(0, 5000)

for(i in 1:5000){
  samp_4 <- sample(c("atheist", "non_atheist"), n_4, replace = TRUE, prob = c(p_4, 1-p_4))
  p_hats4[i] <- sum(samp_4 == "atheist")/n_4
}
```

```r
# plot of all 4 histograms

par(mfrow = c(2, 2))

hist(p_hats, main = "p = 0.1, n = 1040", xlim = c(0, 0.18))
hist(p_hats2, main = "p = 0.1, n = 400", xlim = c(0, 0.18))
hist(p_hats3, main = "p = 0.02, n = 1040", xlim = c(0, 0.18))
hist(p_hats4, main = "p = 0.02, n = 400", xlim = c(0, 0.18))
```



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.

```r
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?

```
aust1 <- 1040 * 0.1
aust2 <- 1040 * (1 - 0.1)

aust1
```

```
## [1] 104
```

```
aust2
```

```
## [1] 936
```

It is sensible to proceed with inference and report margin of errors for Australia, since it meets both conditions of $np > or = 10 and n(1-p) > or = 10$**.**

```
ecua1 <- 400 * 0.02
ecua2 <- 400 * (1 - 0.02)

ecua1
```

```
## [1] 8
```

```
ecua2
```

```
## [1] 392
```

However, it is not sensible to proceed with inference and report margin of errors for Ecuador, since it does not meet both conditions of np > or = 10 and n(1-p) > or = 10. n(1-p) = 8 which is less than 10.

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.

The conditions for inference to construct a 95% confidence interval for the proportion of atheists in Spain in 2005 and in 2012 are the following:

- **The data needs to come from a random sample.**
- **There needs to be at least 10 expected successes and 10 expected failures.**
- **The individual observations need to be independent from one another.**
- **The sample size of the population should not be more than 10% of the population.**

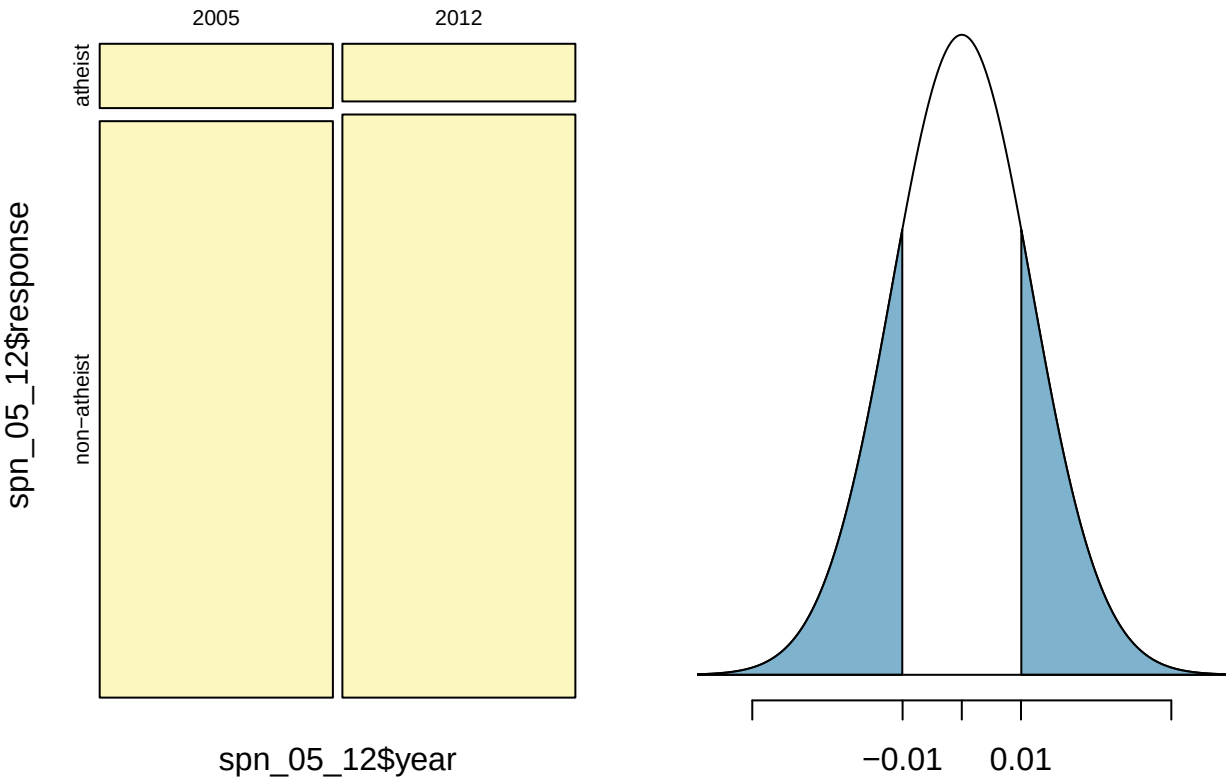
Since the proportions meet this criteria, we are able to proceed to forming confidence intervals.

- **NULL HYPOTHESIS (H0): there isn't a difference in the proportion of atheists in Spain from 2005 to 2012.**
- **ALTERNATIVE HYPOTHESIS (HA): there is a difference in the proportion of atheists in Spain from 2005 to 2012.**

```
spn_05_12 <- subset(atheism, nationality == "Spain" & year == "2005"  |
                      nationality == "Spain" & year == "2012")
inference(y = spn_05_12$response, x = spn_05_12$year, est = "proportion",
          type = "ht", null = 0, alternative = "twosided", method = "theoretical", success = "atheist")
```

```
## Response variable: categorical, Explanatory variable: categorical
## Two categorical variables
## Difference between two proportions -- success: atheist
## Summary statistics:
##              x
## y             2005 2012  Sum
##   atheist      115  103  218
##   non-atheist 1031 1042 2073
##   Sum         1146 1145 2291

## Observed difference between proportions (2005-2012) = 0.0104
##
## H0: p_2005 - p_2012 = 0
## HA: p_2005 - p_2012 != 0
## Pooled proportion = 0.0952
## Check conditions:
##    2005 : number of expected successes = 109 ; number of expected failures = 1037
##    2012 : number of expected successes = 109 ; number of expected failures = 1036
## Standard error = 0.012
## Test statistic: Z =  0.848
## p-value =  0.3966
```



Since the p-value between these two proportions is greater than 0.05, we cannot reject the null hypothesis. Therefore, we can conclude that there isn't convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012.

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

The conditions for inference to construct a 95% confidence interval for the proportion of atheists in Spain in 2005 and in 2012 are the following:

- **The data needs to come from a random sample.**
- **There needs to be at least 10 expected successes and 10 expected failures.**
- **The individual observations need to be independent from one another.**
- **The sample size of the population should not be more than 10% of the population.**

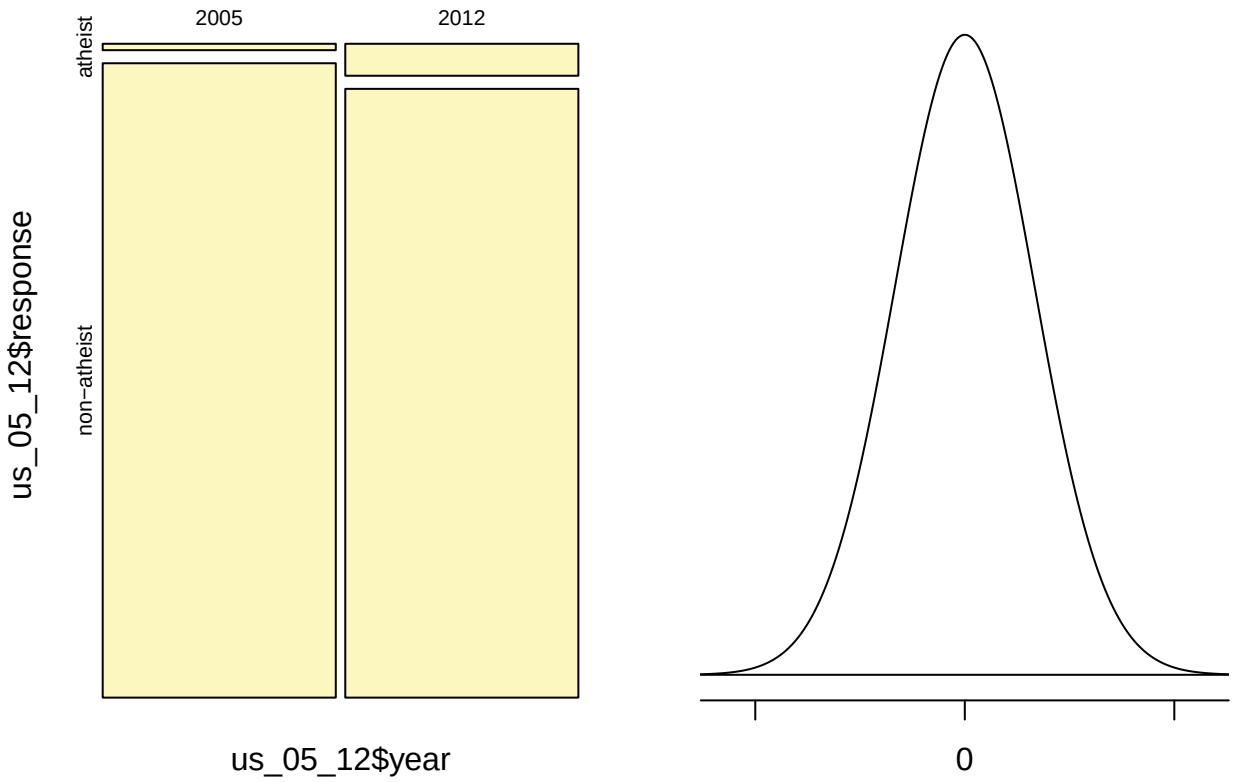
Since the proportions meet this criteria, we are able to proceed to forming confidence intervals.

- **NULL HYPOTHESIS (H0): there isn't a difference in the proportion of atheists in the U.S. from 2005 to 2012.**
- **ALTERNATIVE HYPOTHESIS (HA): there is a difference in the proportion of atheists in U.S. from 2005 to 2012.**

```r
us_05_12 <- subset(atheism, nationality == "United States" & year == "2005"  |
                     nationality == "United States" & year == "2012")
inference(y = us_05_12$response, x = us_05_12$year, est = "proportion",
          type = "ht", null = 0, alternative = "twosided", method = "theoretical", success = "atheist")
```

```
## Response variable: categorical, Explanatory variable: categorical
## Two categorical variables
## Difference between two proportions -- success: atheist
## Summary statistics:
##              x
## y             2005 2012  Sum
##   atheist       10   50   60
##   non-atheist  992  952 1944
##   Sum         1002 1002 2004

## Observed difference between proportions (2005-2012) = -0.0399
##
## H0: p_2005 - p_2012 = 0
## HA: p_2005 - p_2012 != 0
## Pooled proportion = 0.0299
## Check conditions:
##    2005 : number of expected successes = 30 ; number of expected failures = 972
##    2012 : number of expected successes = 30 ; number of expected failures = 972
## Standard error = 0.008
## Test statistic: Z =  -5.243
## p-value =  0
```



From the p-value above, which is less than 0.05, we can reject the null hypothesis and accept the alternative hypothesis that there is convincing evidence the U.S. has seen a change in its atheism index from 2005 to 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.

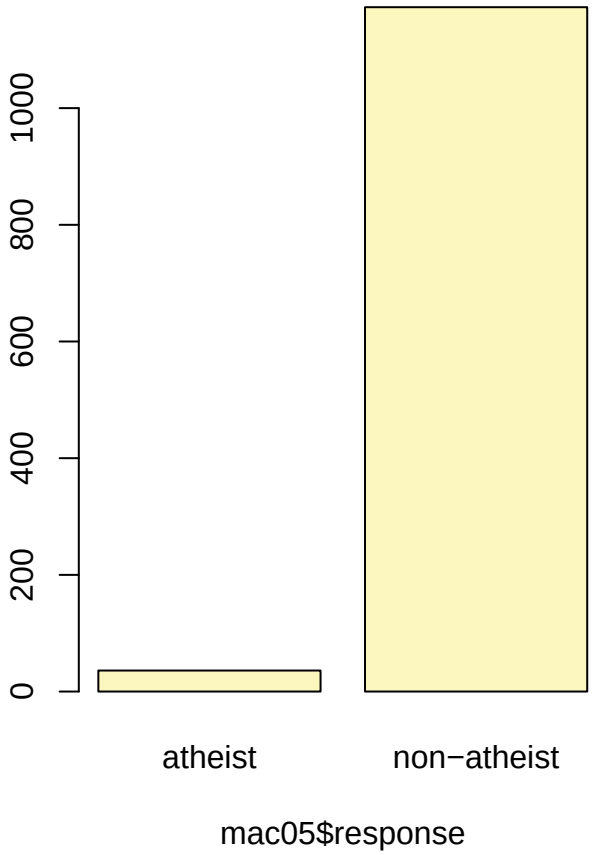
If there are large differences in the point estimates from 2005 to 2012 and a small margin of error due to a large sample size, then I'd expect to detect a change at at significance level of 0.05 simply by chance (about 11 out of the 39 countries listed with a point difference of +/- 4% or greater). However, if the statistical tests do show that there is a significant difference and there actually isn't a change in the atheism index, then this would be an example of a Type 1 error.

An example of this would be Macedonia (point estimate difference of 5%):

```
mac05 <- subset(atheism, nationality == "Macedonia" & year == "2005")
mac12 <- subset(atheism, nationality == "Macedonia" & year == "2012")

inference(mac05$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

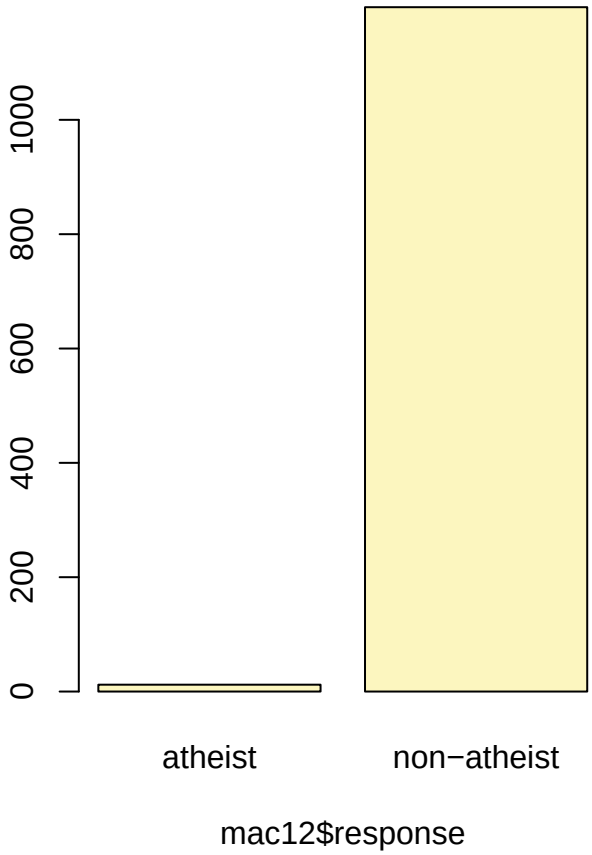
```
## Single proportion -- success: atheist
## Summary statistics:
```




```
## p_hat = 0.0298 ;  n = 1209
## Check conditions: number of successes = 36 ; number of failures = 1173
## Standard error = 0.0049
## 95 % Confidence interval = ( 0.0202 , 0.0394 )
```

```
inference(mac12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

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




```
## p_hat = 0.0099 ;  n = 1209
## Check conditions: number of successes = 12 ; number of failures = 1197
## Standard error = 0.0029
## 95 % Confidence interval = ( 0.0043 , 0.0155 )
```

```
p_mac05 = 0.0298
n_mac05 = 1209
p_mac12 = 0.0099
n_mac12 = 1209

PE_mac = p_mac12 - p_mac05

SE_mac = sqrt(((p_mac05*(1-p_mac05))/n_mac05)+((p_mac12*(1-p_mac12))/n_mac12))
SE_mac
```

```
## [1] 0.005658751
```

```
PE_mac + (1.96 * SE_mac)
```

```
## [1] -0.008808848
```

```
PE_mac - (1.96 * SE_mac)
```

```
## [1] -0.03099115
```

If there wasn't actually a change in the atheism index, then Macedonia would be an example of a Type 1 error.

- 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.

By referring back to the plot of the relationship between p and margin of error, and given that we do not know what to expect from p, we'll have to be safe and calculate a sample size reflecting a proportion of an even split (50/50), since this would yield the highest margin of error for a population.

We also have the following equations to work with:

$Marginof Error = z * \frac{\sqrt{(p*(1-p)}}{n}$

And solving for N:

$N = \frac{z^2 \times p(1-p)}{ME^2}$

```
# a 50/50 split would be a probability of 0.5
prob <- 0.5

# since we want a margin of error no greater than 0.1, we'll set it as such
ME <- .01

z <- qnorm(0.975)

# then, given a 95% confidence interval, we can use the following equation to calculate for n
n <- ((1.96^2)*prob*(1-prob))/((ME)^2)
n <- round(n, digits = 0)
n
```

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

It looks like, to be safe, you'd need to sample at least 9,604 individuals to ensure that you are within the guidelines.