Data606 Lab6 - Inference for categorical data

Lab completed by Chad Bailey

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*?

```
## Student response to exercise 1

## These percentages 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?

```
## Student response to exercise 2

## To be generalizeable to the gobal human population, the survey would have to
## be designed to randomly sample the population. True random sampling of world
## population is not possible from a techonogical, records, or cost perspective
## but a moderate approximation could be done. If Gallup International is a
## well-respected polling organization, it would be reasonable to assume they
## employed a random or approximately random process.
```

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?

```r
## Student response to exercise 3

## get the first few rows of the 'atheism' dataset
  head(atheism)
```

```
##   nationality    response year
## 1 Afghanistan non-atheist 2012
## 2 Afghanistan non-atheist 2012
## 3 Afghanistan non-atheist 2012
## 4 Afghanistan non-atheist 2012
## 5 Afghanistan non-atheist 2012
## 6 Afghanistan non-atheist 2012
```

```r
  nrow(atheism)
```

```
## [1] 88032
```

```r
## Each row of Table 6 corresponds to summary statistics for that country
## Each row of the data set 'atheism' corresponds to an individual's 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?

```r
## Student response to exercise 4

## subset and get proportions
    library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
    us12 <- atheism %>%
      filter(nationality == "United States" & year == "2012") %>%
      mutate(nationality = as.character(nationality)
             ,response = as.character(response))

    table(us12)
```

```
## , , year = 2012
## 
##                response
## nationality     atheist non-atheist
##   United States      50         952
```

```
prop.table(table(us12))
```

```
## , , year = 2012
## 
##                response
## nationality       atheist non-atheist
##   United States 0.0499002   0.9500998
```

```
## The percentages partially agreee with the values in Table 6. The dataset
## 'atheism' only gives the data in a binary 'atheist'/'non-atheist' scale
## whereas the table uses a four category scale. When the four categories are
## collapsed to just a binary scale the values do agree.
```

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.

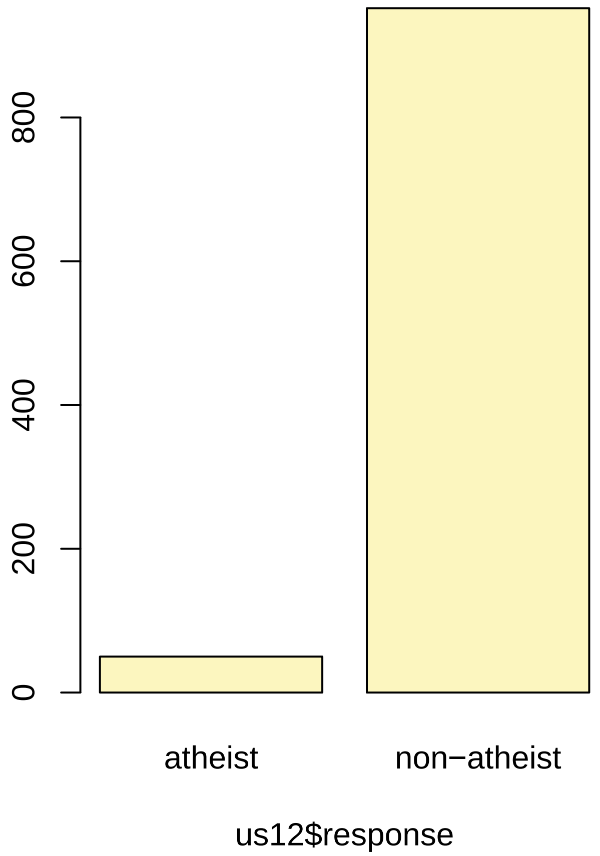
5. Write out the conditions for inference to construct a 95% confidence interval for the proportion of atheists in the United States in 2012. Are you confident all conditions are met?

```
## Student response to exercise 5

## The conditions for a confidence interval of a proportion are:
##  1. observations are independent
##  2. successes (p) and failures (1-p) are both greater than 10

## Yes all conditions have functionally been 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 the proportion of atheists in US in 2012?

```
## student response to exercise 6

round(sqrt(0.05*0.95/1002)*100, 1)
```

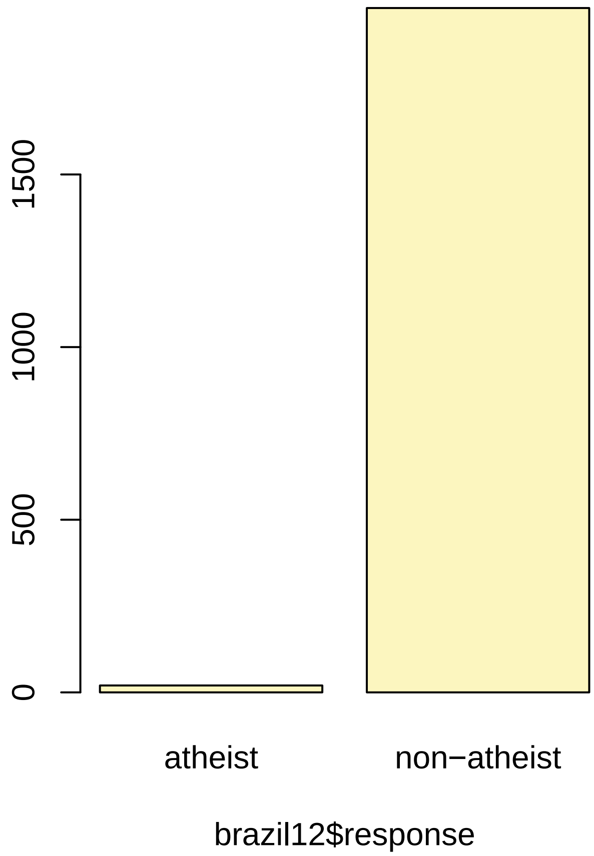
```
## [1] 0.7
```

```
## The margin of error for the estimate of the proportion of atheists in the US
## in 2012 based on the sample was 0.7%.
```

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.

```
## Student repsonse to exercise 7

## Brazil
brazil12 <- atheism %>% filter(nationality == 'Brazil' & year == 2012)
inference(brazil12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

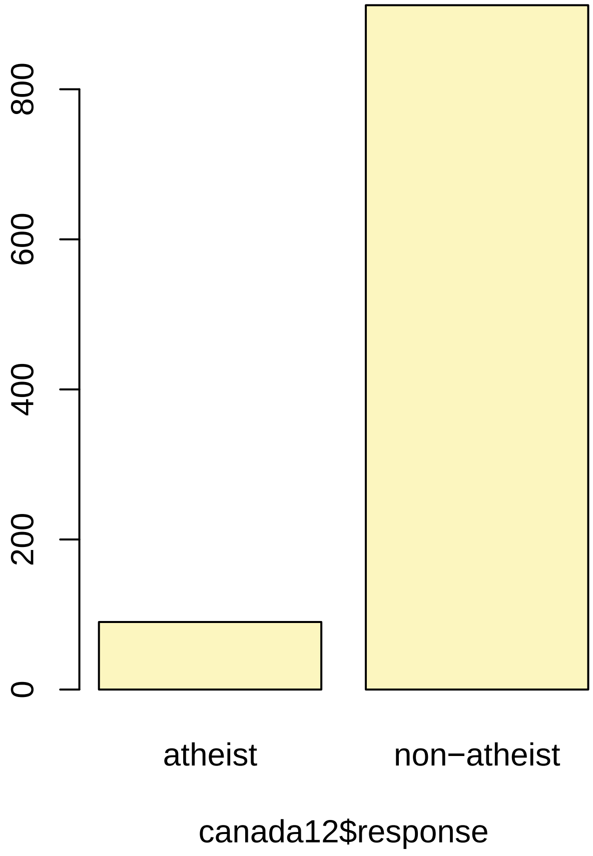
```
## Single proportion -- success: atheist
## Summary statistics:
```



```
## 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 )
```

```
## Canada
canada12 <- atheism %>% filter(nationality == 'Canada' & year == 2012)
inference(canada12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

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



```
## p_hat = 0.0898 ;  n = 1002
## Check conditions: number of successes = 90 ; number of failures = 912
## Standard error = 0.009
## 95 % Confidence interval = ( 0.0721 , 0.1075 )
```

```
## Based on the 2012 sample, the confidence intervals for Brazil and Canada would be
## (0.0056, 0.0143) and (0.0721, 0.1075) respectively
```

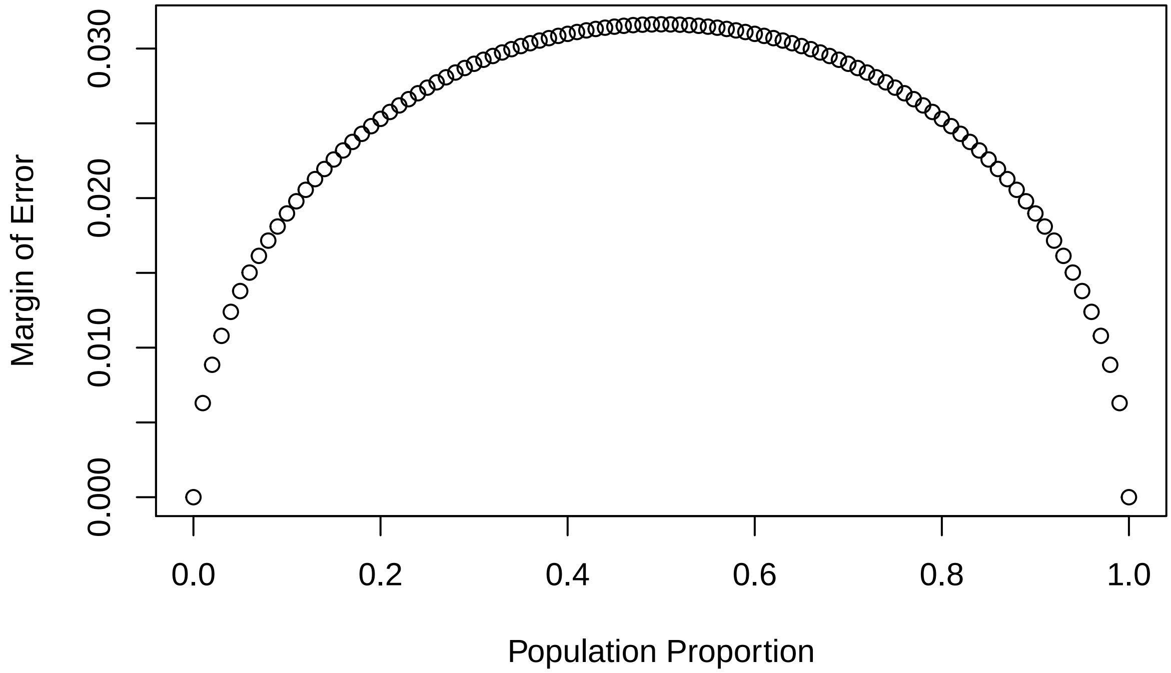
How does the proportion affect the margin of error?

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.

```r
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`.

```r
## student response to exercise 8

## The relationsihp of 'p' and 'me' is semi-circular in shape. As 'p' increases
## so does 'me' until 0.5. At 0.5 the graph inverts and as 'p' further increases
## 'me' descreases.
```

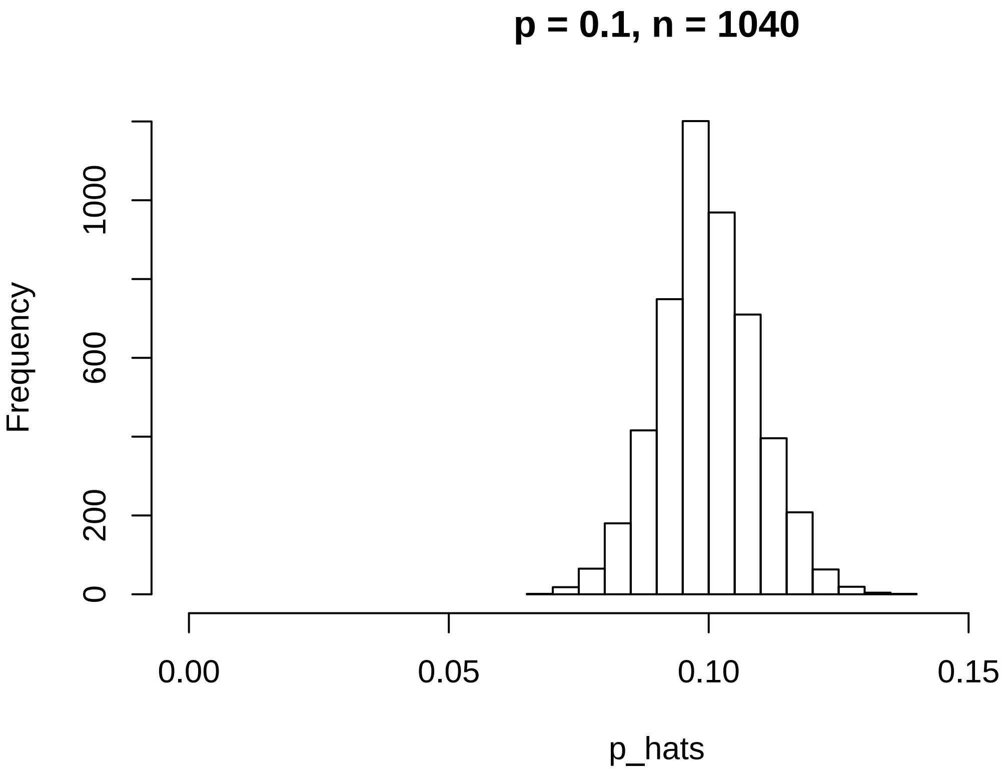
Success-failure condition

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.

```r
p <- 0.1
n <- 1040
p_hats <- rep(0, 5000)

for(i in 1:5000){
  set.seed(i); 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.

```
## Student response to exercise 9

mean(p_hats)
```

```
## [1] 0.09997
```

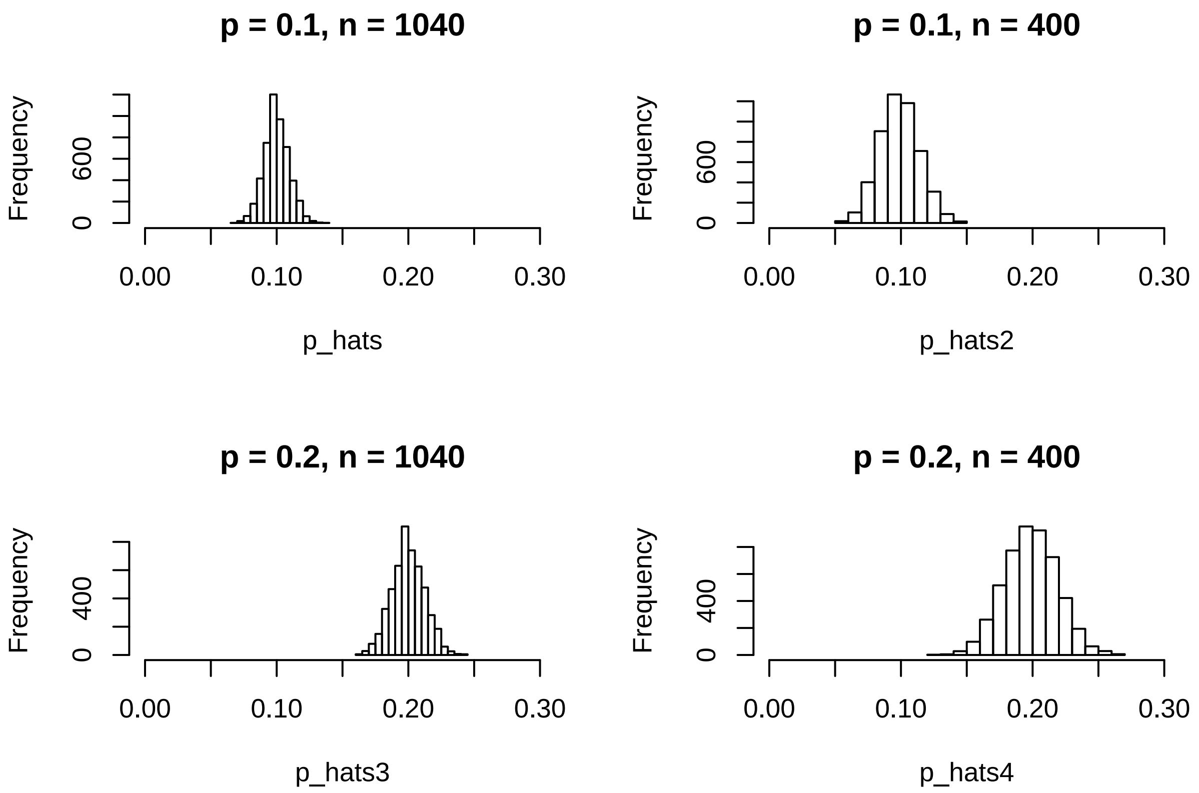
```
sd(p_hats)
```

```
## [1] 0.009324635
```

```
## The distirubtion is unimodal, nearly symmetrical, centered at 0.1, and has
## as standard deviation of 0.009
```

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?

```
## simulation1
  p <- 0.1
  n <- 1040
  p_hats <- rep(0, 5000)

  for(i in 1:5000){
    set.seed(i); samp <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
    p_hats[i] <- sum(samp == "atheist")/n
  }

  ## simulation2
  p <- 0.1
  n <- 400
  p_hats2 <- rep(0, 5000)

  for(i in 1:5000){
    set.seed(i); samp <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
    p_hats2[i] <- sum(samp == "atheist")/n
  }

  ## simulation1
  p <- 0.2
  n <- 1040
  p_hats3 <- rep(0, 5000)

  for(i in 1:5000){
    set.seed(i); samp <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
    p_hats3[i] <- sum(samp == "atheist")/n
  }
  ## simulation1
  p <- 0.2
  n <- 400
  p_hats4 <- rep(0, 5000)

  for(i in 1:5000){
    set.seed(i); samp <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
    p_hats4[i] <- sum(samp == "atheist")/n
  }

## plot the histograms
  par(mfrow = c(2, 2))
  hist(p_hats,  main = "p = 0.1, n = 1040", xlim = c(0, 0.3))
  hist(p_hats2, main = "p = 0.1, n = 400",  xlim = c(0, 0.3))
  hist(p_hats3, main = "p = 0.2, n = 1040", xlim = c(0, 0.3))
  hist(p_hats4, main = "p = 0.2, n = 400",  xlim = c(0, 0.3))
```



```
## As 'n' increases the variance decreases
## Changes in 'p' shift the center of the distribution
```

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 report does?

```
## Student response to exercise 11

## No, it is not sensible to report margin of errors for Ecuador based on the
## sample size of 400 and proportion of 0.02. This would not meet the required
## conditions that both p and (1-p) be greater 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.

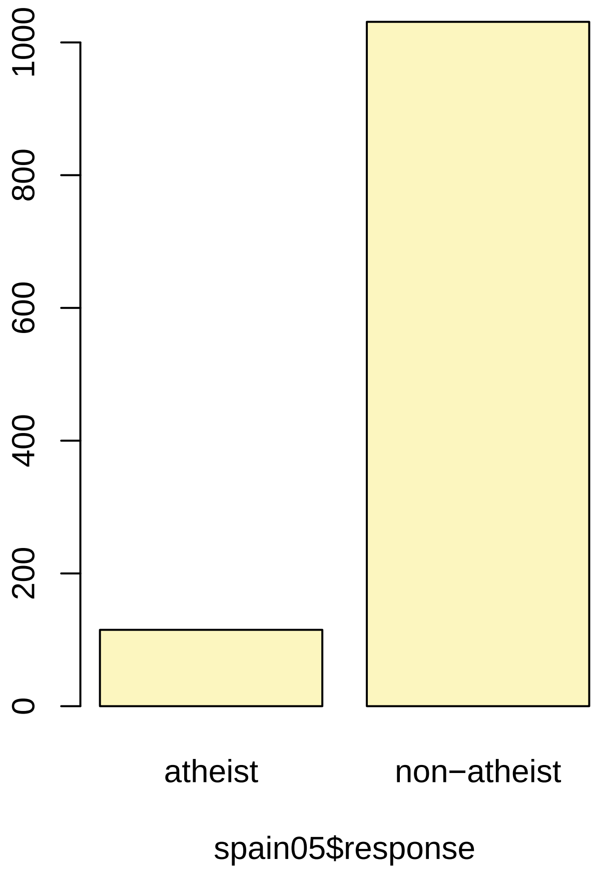
  a. Is there convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012?
  Hint: Create a new data set for respondents from Spain. Form confidence intervals for the true proportion of athiests in both years, and determine whether they overlap.

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

```
## Student response to 'On your own 1'

## 1.a1 NULL Hypothesis: Spain has not seen significant change in its
##                       atheism index from 2005 to 2015.
##      Alternate Hypothesis: Spain has seen a significant change in its
##                            atheism index from 2005 to 2012.
## 1.a2 The conditions for a valid confidence interval are:
##      (1) independent observations
##      (2) Both success (p) and failure (1-p) are 10 or greater




## 1.a3 get subsetted data sets for Spain-2005 and Spain-2012
  spain05 <- atheism %>%
            filter(nationality == 'Spain' & year == 2005) %>%
            mutate(nationality = as.character(nationality)
                   ,response = as.character(response))

  spain12 <- atheism %>%
            filter(nationality == 'Spain' & year == 2012) %>%
            mutate(nationality = as.character(nationality)
                   ,response = as.character(response))
```

```r
## 1.a4 calculate confidence intervals using function 'inference'

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

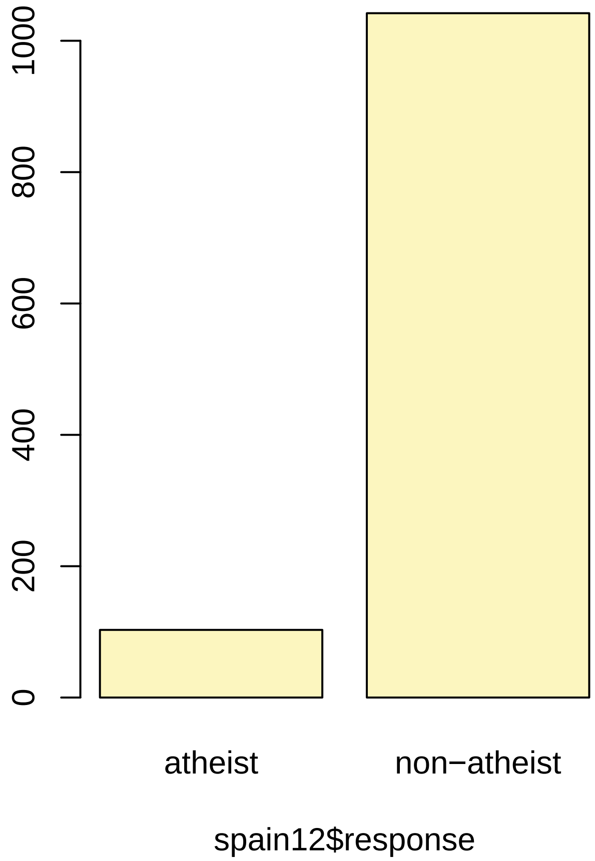
```
## Single proportion -- success: atheist
## Summary statistics:
```



```
## 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 )
```

```r
  inference(spain12$response, est = "proportion", type = "ci", method = "theoretical",
            success = "atheist")
```

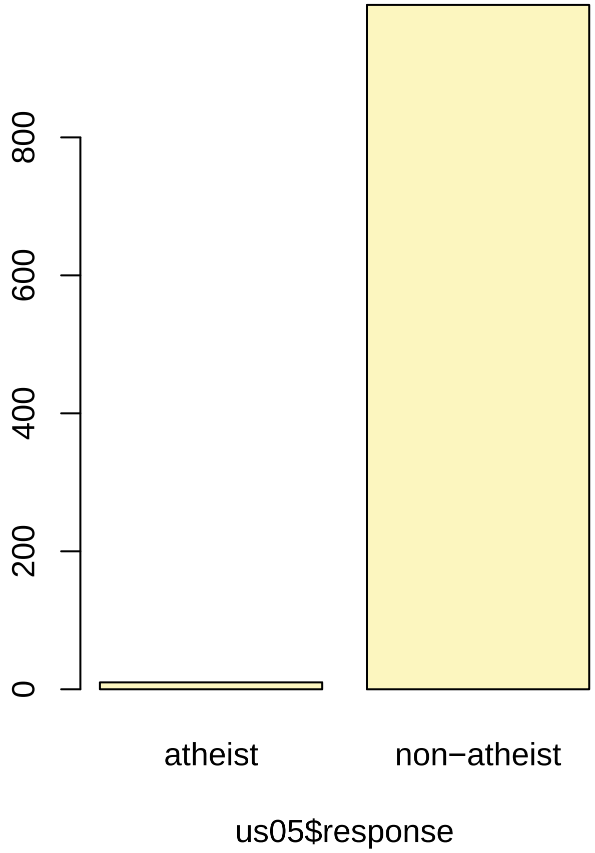
```
## Single proportion -- success: atheist
## Summary statistics:
```



```
## 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 )
```

```
## 1.a5 The confidence intervals for the two years overlap, therefore we fail
##      to rejec the NULL Hypothesis.


## 1.b1 NULL Hypothesis: US has not seen significant change in its
##                       atheism index from 2005 to 2015.
##      Alternate Hypothesis: US has seen a significant change in its
##                            atheism index from 2005 to 2012.
## 1.b2 The conditions for a valid confidence interval are:
##      (1) independent observations
##      (2) Both success (p) and failure (1-p) are 10 or greater



## 1.b3 get subsetted data sets for Spain-2005 and Spain-2012
  us05 <- atheism %>%
            filter(nationality == 'United States' & year == 2005) %>%
            mutate(nationality = as.character(nationality)
                   ,response = as.character(response))
  us12 <- atheism %>%
            filter(nationality == 'United States' & year == 2012) %>%
            mutate(nationality = as.character(nationality)
                   ,response = as.character(response))


## 1.b4 calculate confidence intervals using function 'inference'

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

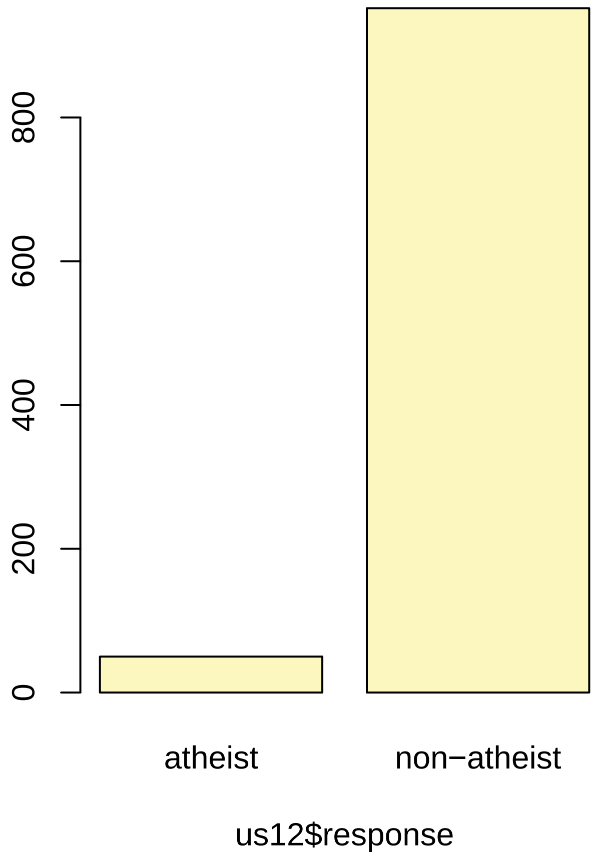
```
## Single proportion -- success: atheist
## Summary statistics:
```



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

```
  inference(us12$response, est = "proportion", type = "ci", method = "theoretical",
            success = "atheist")
```

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



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

```
## 1.b5 The confidence intervals for the two years do not overlap and so
##      we 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.

```
## Student response to 'On your own 2'

## By definition you would expect to have roughly 5% of cases to have a
## a change detected by a significance level of 0.05.
```

- 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.

```
## Student response to 'On you own 3'

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

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

```
## Taking the most conservative approach, scenario that success (p)
## and failure (1-p) are equally likely and so both are equal to 0.5,
## then the required minimum sample size would 9604 to ensure a margin of
## error no greater than 1% with 95% confidence.
```