Inference for categorical data

Jai Jeffryes

10/20/2019

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*?

ANSWER

These are sample statistics. It is impossible to query the entire population of China, for example.

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?

ANSWER

A generalization requires a representative sample from all regions of the world. The researchers sampled 57 countries. There are nearly 200 in the world. The results appear meaningful, but I would stop short of calling them global.

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.

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

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

ANSWER

- Table 6 is a frequency table containing statistical measures, each row summarizing responses from one country.
- One row of the atheism data frame is an observation, namely one sampled response. The data have been transformed, however. Responses consisting of "a religious person", "not a religious person," and "don't know / no response" have been merged into a single reponse identified as "non-atheist."

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_prop <- sum(us12$response == "atheist") / nrow(us12)
```

ANSWER

The proportion of atheist responses is: **0.0499002**. The tabular percentage differs because it is rounded and reported in integers.

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?

ANSWER

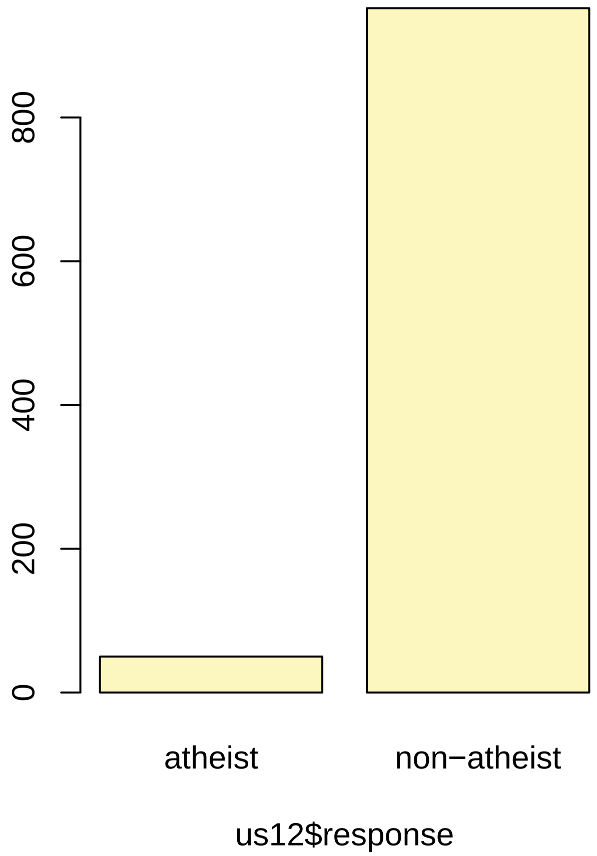
(5a.) The samples must be independent and representative of the population of interest. The WIN-Gallup poll says that the polls were random samples, and I trust this meets the requirement.

(5b.) The success-failure conditions must be met. We use $\hat{p}$ instead of p to calculate the confidence interval. Atheist: np = 1002 * 0.05 = 50.1. Not atheist: $n(1 - p)$ = 1002 * 0.95 = 951.9. Both values are at least 10, so we can use the normal distribution to model $\hat{p}$. We need at least 10 successes and 10 failures.

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?

ANSWER

The margin of error is the span of the confidence interval. The C.I. is (0.0364 , 0.0634). Its length is 0.027. Another way to figure it is from the original C.I. calculation. One side of interval is the Z score for 95% confidence (1.96) times the standard error (0.0069). Multiply that by 2 for both sides of the interval: 0.027048.

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.

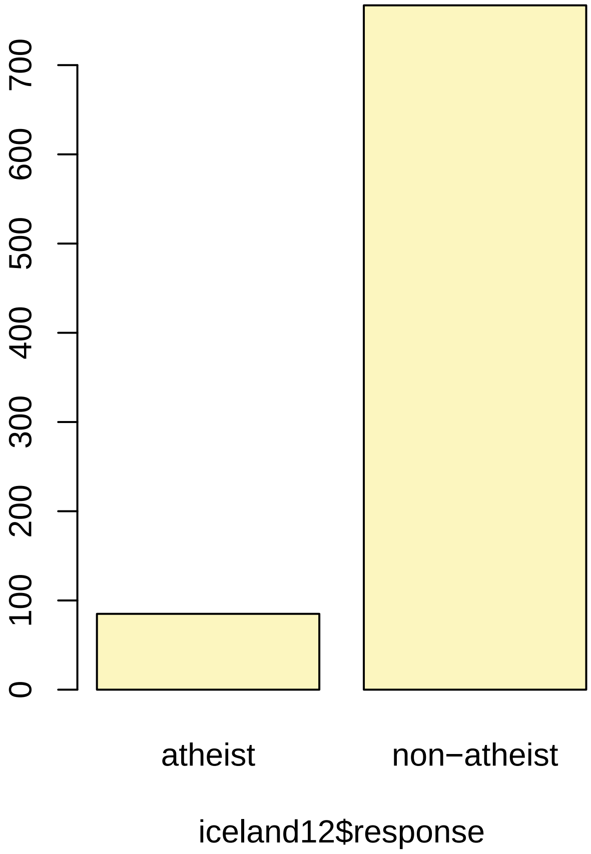
ANSWER

```
# Iceland.
iceland12 <- subset(atheism, nationality == "Iceland" & year == 2012)
# Count successes and failures. There are at least 10 of each.
table(iceland12$response)
```

```
##
##     atheist non-atheist
##          85         767
```

```
inference(iceland12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

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



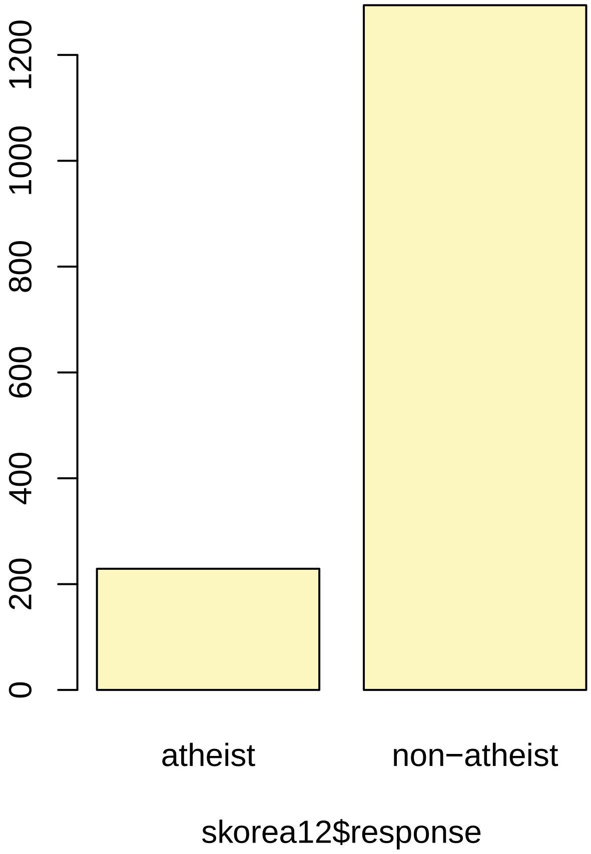
```
## p_hat = 0.0998 ;  n = 852
## Check conditions: number of successes = 85 ; number of failures = 767
## Standard error = 0.0103
## 95 % Confidence interval = ( 0.0796 , 0.1199 )
```

```
# South Korea.
skorea12 <- subset(atheism, nationality == "Korea, Rep (South)" & year == 2012)
# Count successes and failures. There are at least 10 of each.
table(skorea12$response)
```

```
##
##     atheist non-atheist
##         229        1294
```

```
inference(skorea12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

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



```
## p_hat = 0.1504 ;  n = 1523
## Check conditions: number of successes = 229 ; number of failures = 1294
## Standard error = 0.0092
## 95 % Confidence interval = ( 0.1324 , 0.1683 )
```

- Iceland margin of error: 0.0403.
- South Korea margin of error: 0.0359.

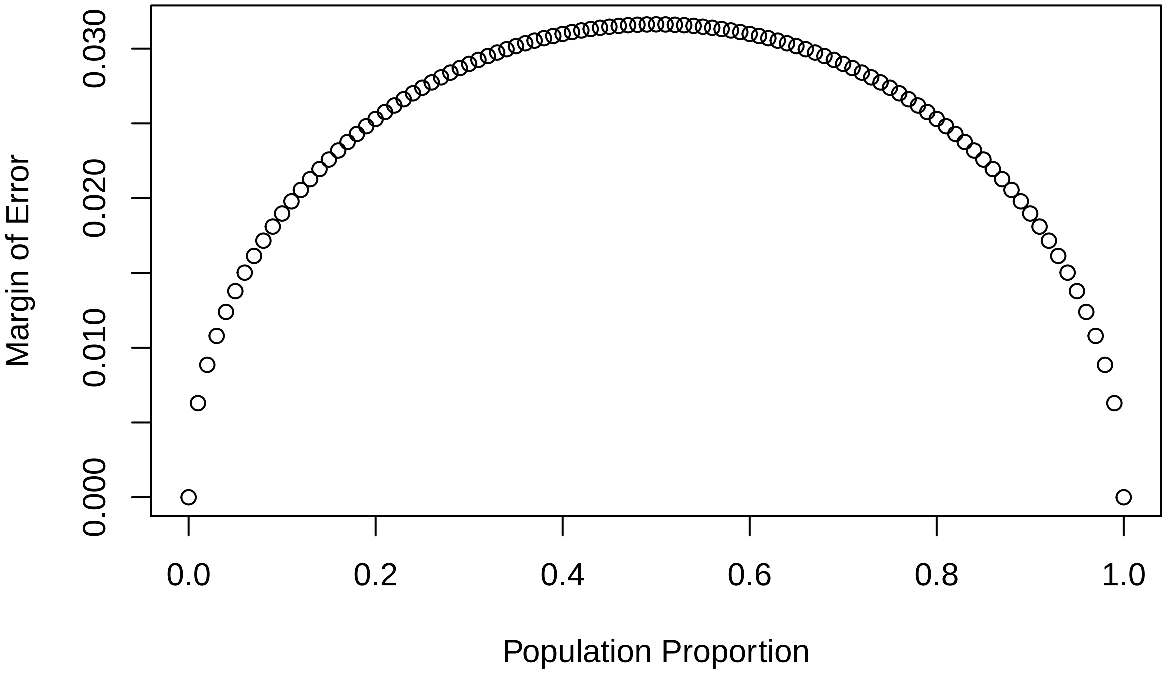
How does the proportion affect the margin of error?

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`.

ANSWER

Non-linear.

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?

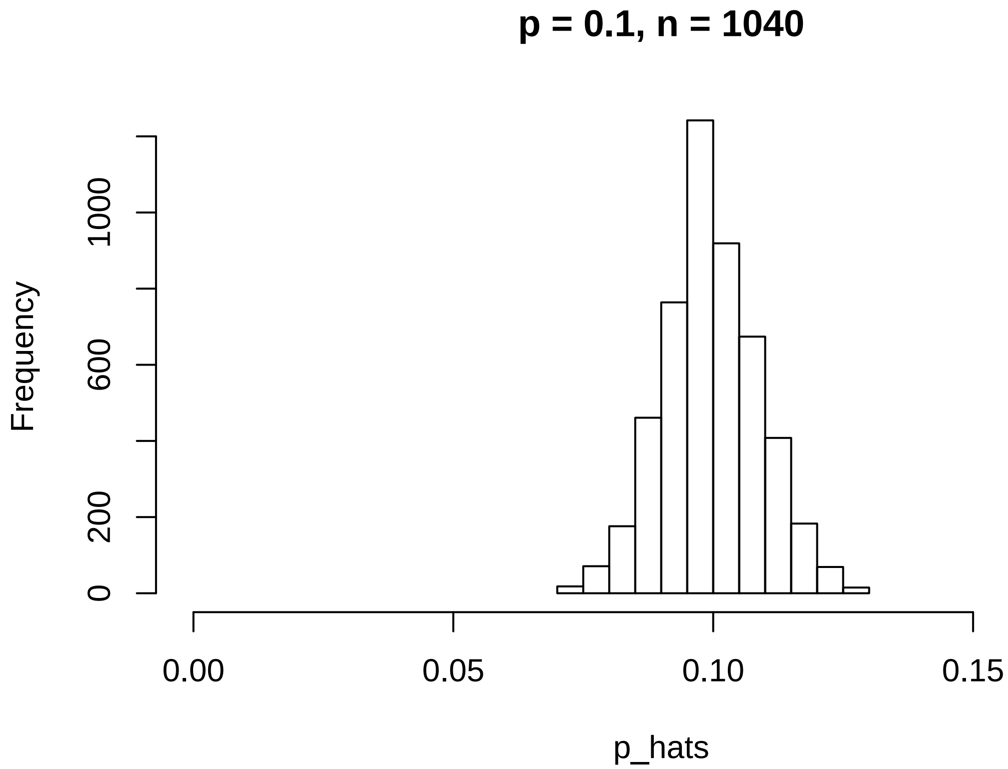
The short answer is: nothing. You could argue that we would be fine with 9 or that we really should be using 11. What is the "best" value for such a rule of thumb is, at least to some degree, arbitrary. However, when np and $n(1-p)$ reaches 10 the sampling distribution is sufficiently normal to use confidence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

ANSWER

It is a unimodal, symmetric distribution with a center near the population mean of .1. It's spread is about 0.6.

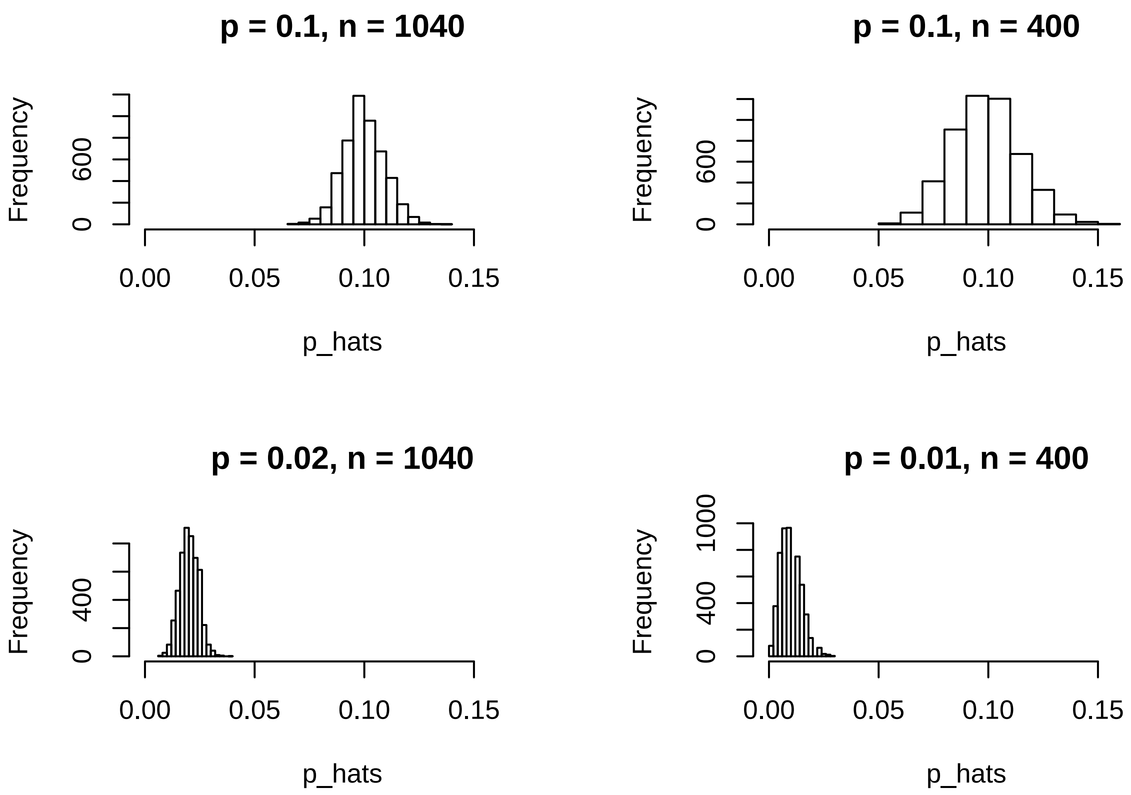
10. Repeat the above simulation three more times but with modified sample sizes and proportions: for $n = 400$ and $p = 0.1$, $n = 1040$ and $p = 0.02$, and $n = 400$ and $p = 0.02$. Plot all four histograms together by running the `par(mfrow = c(2, 2))` command before creating the histograms. You may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

```r
sim <- function(n, p) {
  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 = paste0("p = ", p, ", n = ", n), xlim = c(0, 0.18))
}

par(mfrow = c(2, 2))
sim(1040, .1)
sim(400, .1)
sim(1040, .02)
sim(400, .01)
```



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

- The higher the number of samples, n, the narrower the spread.
- The higher the proportion, p, the more symmetric 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?

SHORT ANSWER

Yes, because the reference to margin of error is at the level of the global population, not subsets by country.

LONG ANSWER

I had a strong reaction to this paper. I would kind of like to know if you think my impressions make sense or are off base.

The question is really undefined. The single mention of error margins is on page seven on the report, where it says that in general for surveys of this kind the margin is plus or minus 3% to 5%. It doesn't say what the margin is on this report. Moreover, though this lab question asks about inferences, it appears to me there is little in the way of clear inference in this report. The statistics appear more descriptive than inferential. Mostly, the exhibits present frequencies of survey answers along with descriptive statistics about their proportions in various regions and across time.

- Page 4 notes variations among believers. That looks descriptive.
- Page 5 correlates "religiosity" against income and against education. Those smell like hypotheses and inferences, but the evaluations don't appear to have statistical evidence and inference behind them. They appear to be motivated merely by a cursory reading of proportions on the tables.
- Page 6 headlines a "notable" decline in religiosity in 2005. No quantification is given for the word "notable." We see the percentage change in observations as presented in the table, but we don't have a technical measure of the significance of those so-called notable changes.

There is no clear hypothesis, null or alternative, stated anywhere in the report. Therefore, there is no inference. This report is pretty thin. Page 7 reveals its agenda in commentary by Jean-Marc Leger, President of WIN-Gallup International, commentary quite surprising and disappointing in its absence of statistical language, especially considering its source from a prominent producer of survey data. "Despite the immense impact [*sic*] of technology… the 21st century overwhelmingly espouses a religious faith…" This is not science, this is a religionist's puff-piece.

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 atheists in both years, and determine whether they overlap.

ANSWER

```r
# Spain.
spain05 <- subset(atheism, nationality == "Spain" & year == 2005)
spain12 <- subset(atheism, nationality == "Spain" & year == 2012)
# Count successes and failures. There are at least 10 of each.
table(spain05$response)
```

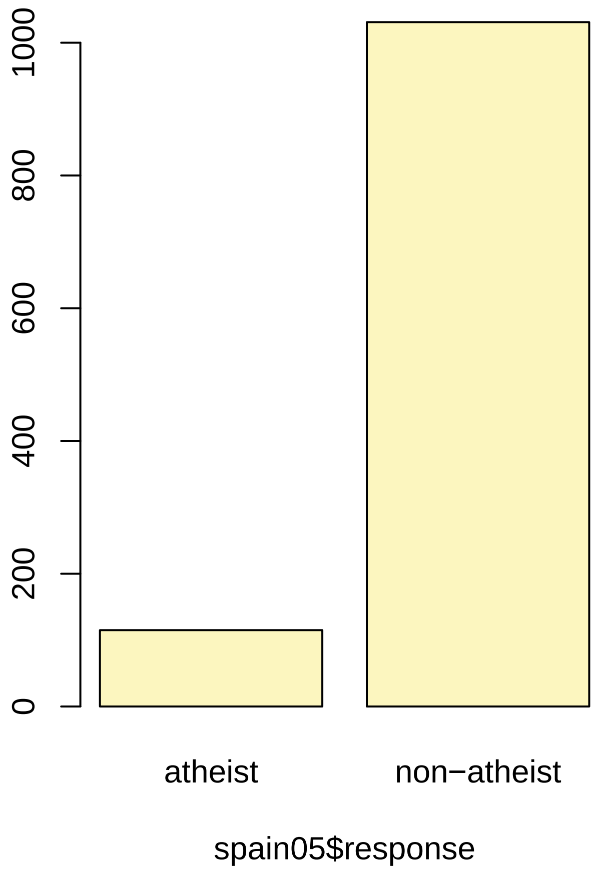
```
##
##     atheist non-atheist
##         115        1031
```

```r
table(spain12$response)
```

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

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

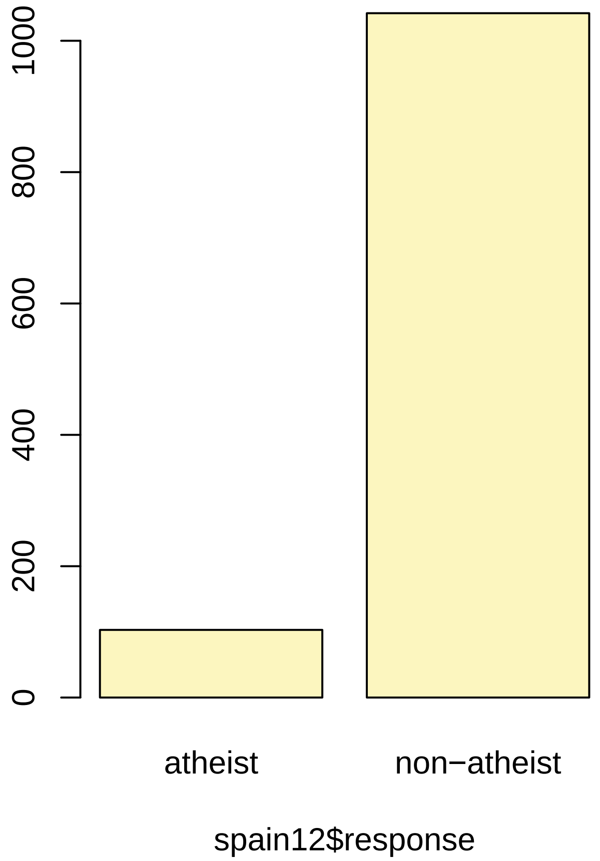
```
## Single proportion -- success: atheist
## Summary statistics:
```



```
## 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 )
```

a. Success conditions require at least 10 positive and negative results. Met for Spain's data in each year, 2005 and 2012.

b. H_0: There is no change in the proportion of atheists in Spain between years 2005 and 2012.

c. H_A: There is a change in the proportion of atheists in Spain between years 2005 and 2012.

d. At a confidence level of 95%, the confidence interval for $\hat{p}$ including the population proportion of atheists in 2005 is: (0.083 , 0.1177). In 2012 it is: (0.0734 , 0.1065).

e. Since the confidence intervals overlap, we do not reject the null hypothesis. There is no convincing evidence that the proportion of atheists in Spain changed 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?

ANSWER

```
# U.S.
us05 <- subset(atheism, nationality == "United States" & year == 2005)
# Count successes and failures. There are at least 10 of each.
table(us05$response)
```

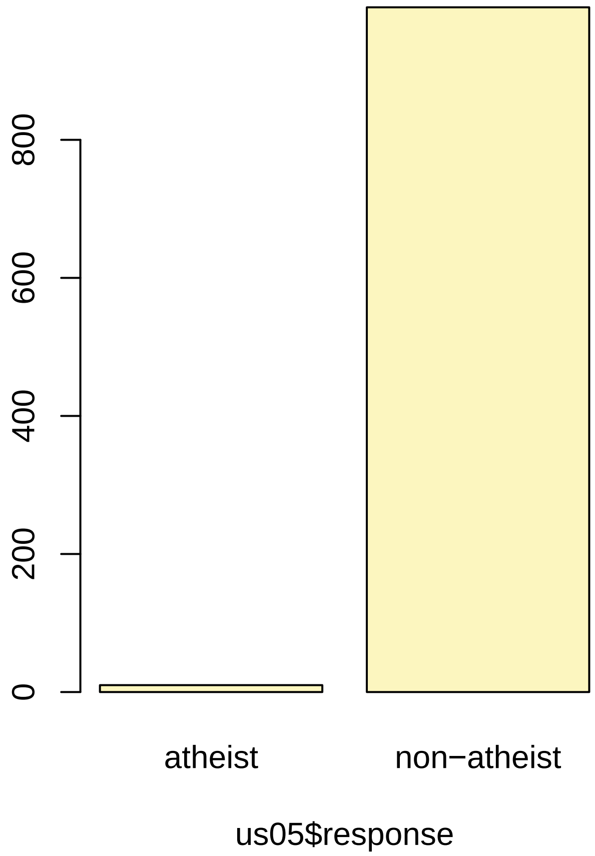
```
##
##     atheist non-atheist
##          10         992
```

```
table(us12$response)
```

```
##
##     atheist non-atheist
##          50         952
```

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

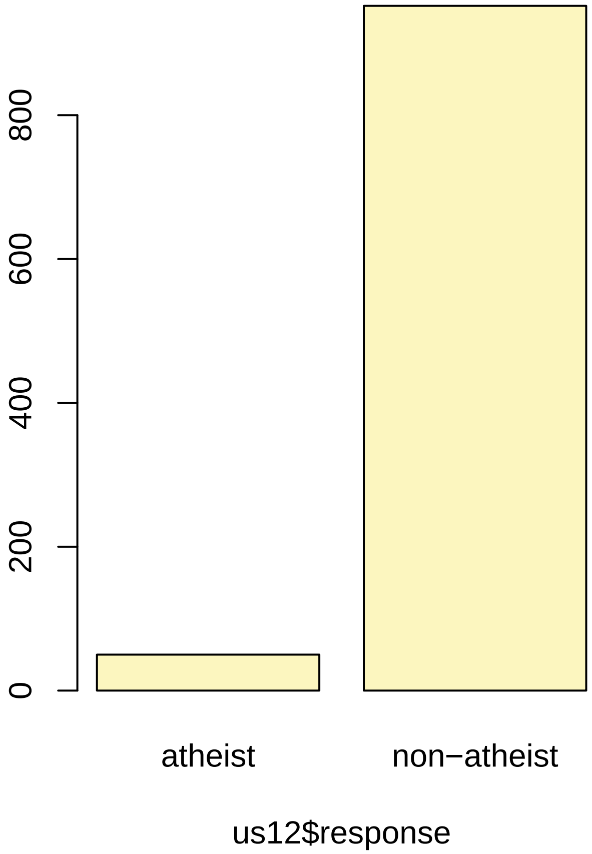
```
## Single proportion -- success: atheist
## Summary statistics:
```



```
## 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 )
```

a. Success conditions require at least 10 positive and negative results. Met for United States data in each year, 2005 and 2012.
b. H_0: There is no change in the proportion of atheists in the United States between years 2005 and 2012.
c. H_A: There is a change in the proportion of atheists in the United States between years 2005 and 2012.
d. At a confidence level of 95%, the confidence interval for $\hat{p}$ including the population proportion of atheists in 2005 is: (0.0038 , 0.0161). In 2012 it is: (0.0364 , 0.0634).
e. Since the confidence intervals do not overlap, we reject the null hypothesis. There is convincing evidence that the proportion of atheists in the United States 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.

ANSWER

A Type 1 error consists of rejecting H_0 when it is true. A significance level of .05 means that this error is expected to happen 5% of the time. The product of 39 countries times significance level .05 yields an expected 1.95 of the countries on Table 4 would detect a change 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.

ANSWER

We would use the worst case scenario of proportions, namely a proportion of 0.5. The margin of error for a sample proportion is: z* * sqrt((p * (1 - 0)) / n). For 95% confidence, the Z score is 1.96, and we want the margin of error to be less than 0.01. We'll rearrange the inequality for that to get n on one side of the equation.

```
n <- 1.96^2 * (0.5 * (1 - 0.5) / .01^2)
n
```

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

We need at least 9604 samples.