# Inference for categorical data

*Jimmy Ng*

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*? # JN: these are read as population parameters. They represent the population, and these numbers are inferred from the sample.
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? # JN: we must assume a random sampling method and that method should cover sufficient number of people that are representative enough of that significant culture. Randomly selecting 50000 individuals from 57 countries could be enough; however, the sampling methodology may heavily rely on convenient sampling, i.e. people living in rural area without internet, telecommunication or people who are illiterate may be excluded from the study. It is extremely difficult (and very expensive) to control for all these extraneous variables to come up with appropriate samples for this global study.

## 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? # JN: each row of table 6 corresponds to a single country. The dataframe of “atheism” has only three columns, i.e. nationality, response and year. Each row presumably represents a sample from the associated nationality and year and his/her response of whether he/she self-identified as an atheist (there’s only two responses – unique(atheism$response), i.e. [1] non-atheist 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")
prop.table(ftable(us12$response))
```

```
##    atheist non-atheist
##
##  0.0499002   0.9500998
```

# JN: yes it does agree with the result on table 6.

## Inference on proportions

As was hinted at in Exercise 1, Table 6 provides *statistics*, that is, calculations made from the sample of 51,927 people. What we'd like, though, is insight into the population *parameters*. You answer the question, "What proportion of people in your sample reported being atheists?" with a statistic; while the question "What proportion of people on earth would report being atheists" is answered with an estimate of the parameter.

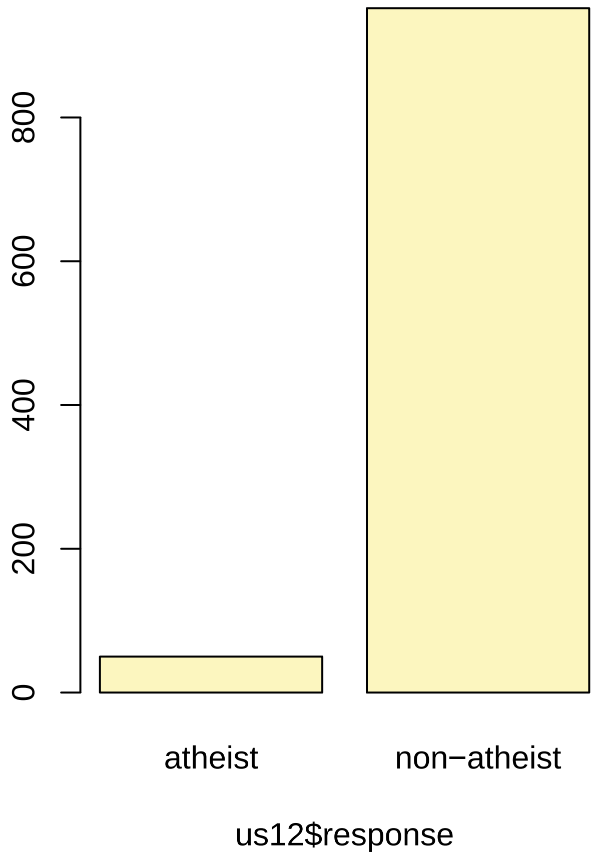
The inferential tools for estimating population proportion are analogous to those used for means in the last chapter: the confidence interval and the hypothesis test.

5. Write out the conditions for inference to construct a 95% confidence interval for the proportion of atheists in the United States in 2012. Are you confident all conditions are met? # JN: There are two conditions that must be met, i.e. first of all, observations are independent, and second, we must meet the success-failure condition. The data collection of the US population is based on simple random sample and consisted of less than 10% of the population, so that verifies independence. In addition, there were 1002 * 0.05 "successes" and 1002 * (1 - 0.05) "failures" in the sample, both easily greater than 10. Therefore, it verifies the success-failure condition as well.

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"`.

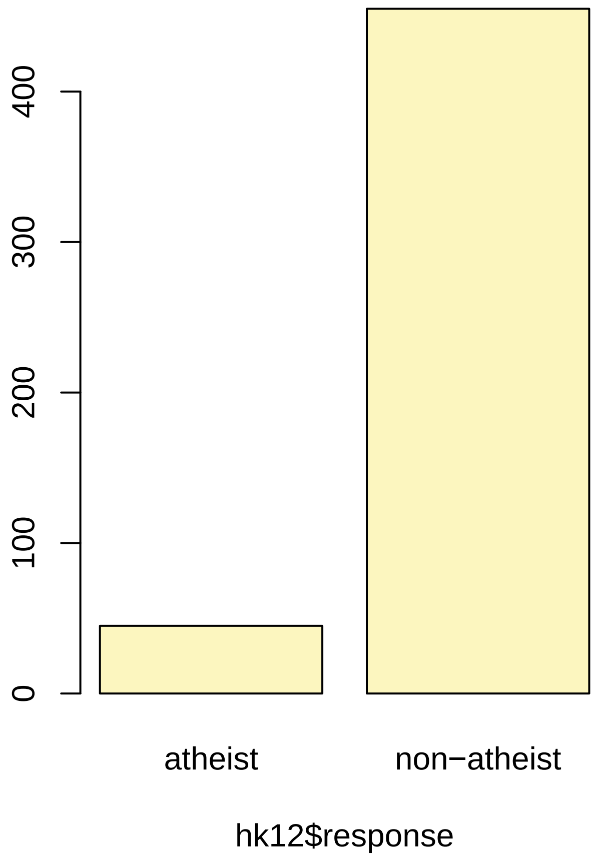
Although formal confidence intervals and hypothesis tests don't show up in the report, suggestions of inference appear at the bottom of page 7: "In general, the error margin for surveys of this kind is ± 3-5% at 95% confidence".

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? # JN: the standard error is equal to 0.0069. The margin of error for 95% confidence intervals is equal to z * se = 1.96 * 0.0069, which is approximately 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.

```
hk12 <- subset(atheism, nationality == "Hong Kong" & year == "2012")
j12 <- subset(atheism, nationality == "Japan" & year == "2012")

inference(hk12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

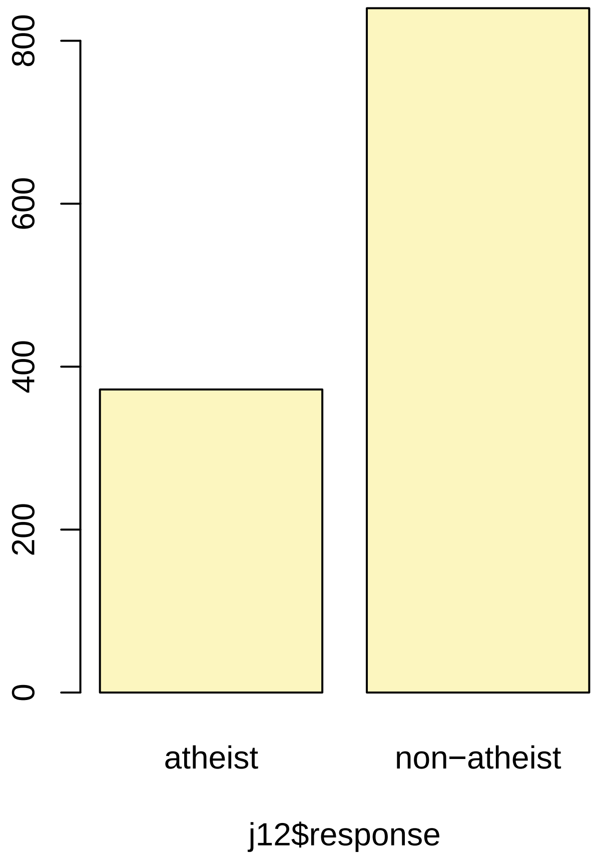
```
## Single proportion -- success: atheist
## Summary statistics:
```



```
## p_hat = 0.09 ;  n = 500
## Check conditions: number of successes = 45 ; number of failures = 455
## Standard error = 0.0128
## 95 % Confidence interval = ( 0.0649 , 0.1151 )
```

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

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



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

**JN: both countries have met the two conditions. First, their sampling were independent and the sample size for both were way less than (10% of) their population (Hong Kong had over 7 million, whereas Japan had over 100 million as of 2012); second, the "success-failure" condition also checked out. For Hong Kong, 500 * 0.09 and 500 * (1 - 0.09), both were easily greater than 10; for Japan, 1200 * 0.31 and 1200 * (1 - 0.31), both were also easily greater than 10. Therefore, the conditions checked out and we can move on to calculate the margin of error. For Hong Kong, the 95% confidence interval is equal to (0.0649, 0.1151) and the margin of error is 1.96 * 0.0128 = 0.025088. For Japan, the 95% confidence interval is equal to (0.281, 0.3329) and the margin of error is 1.96 * 0.0132 = 0.025872.**

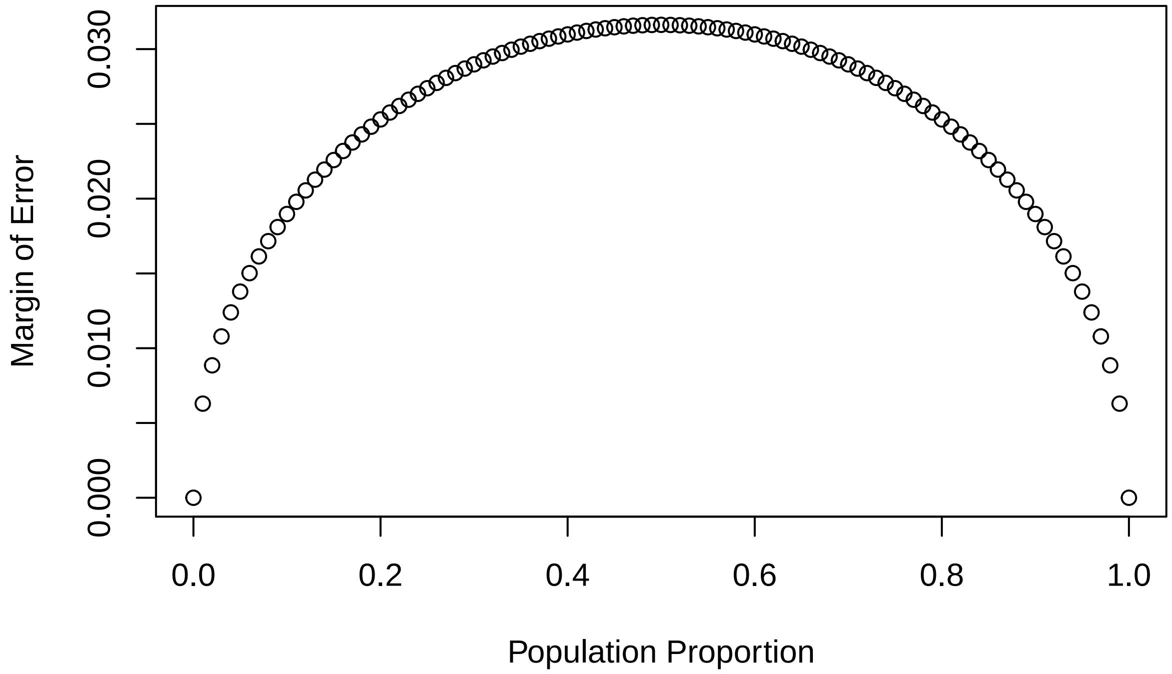
### How does the proportion affect the margin of error?

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`. # JN: The size of ME reaches its peak when the proportion reaches half. The size of ME rises along with the increase of p from 0, and then it reaches its peak when p is about half (p = 0.5). After that, it goes down hill and mirrors a symmetrical relationship as the first half. Essentially, it's a parabola.

### 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?

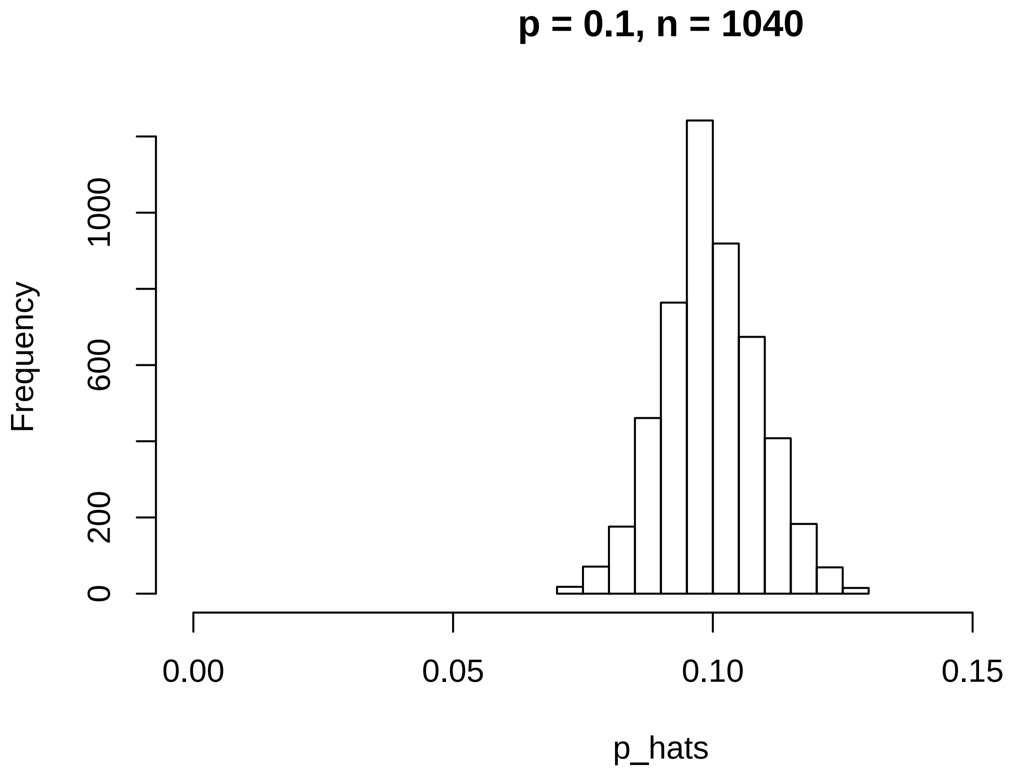
The short answer is: nothing. You could argue that we would be fine with 9 or that we really should be using 11. What is the "best" value for such a rule of thumb is, at least to some degree, arbitrary. However, when $np$ and $n(1-p)$ reaches 10 the sampling distribution is sufficiently normal to use 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. # JN: the sampling distribution of the sample proportion (0.1) resembles a bell curve - a sign of normal distribution. The center is at 0.1 and the curve looks symmetrical with equal spread on both sides. This distribution visually presents a collection of the mean of 5000 samples (each has n = 1040). The probability of "atheist" is constant throughout the 5000 samples and the sampling distribution can reflect that (as the center is equal to the constant, i.e. 0.1).

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
simulation <- function(n, p, iteration = 5000, seed = 1234){
        if(!require(purrr)){install.packages("purrr"); require(purrr)}
        if(!require(plyr)){install.packages("plyr"); require(plyr)}

        set.seed(seed)
        sim <- as.vector(rep(NA, iteration))
        sim <- map(1:iteration, function(x) sim[x] <- mean( sample(
                c(1, 0),
                n,
                replace = TRUE,
                prob = c(p, 1-p)) )
        )
        sim <- sim %>%
                plyr::ldply(., data.frame)
        sim <- as.vector(sim[,1])
}

n1 = 1040; p1 = 0.1
n2 = 400; p2 = 0.1
n3 = 1040; p3 = 0.02
n4 = 400; p4 = 0.02

parameters <- list(n = c(n1, n2, n3, n4), p = c(p1, p2, p3, p4))
output <- purrr::map2(parameters$n, parameters$p,
                     simulation)
```

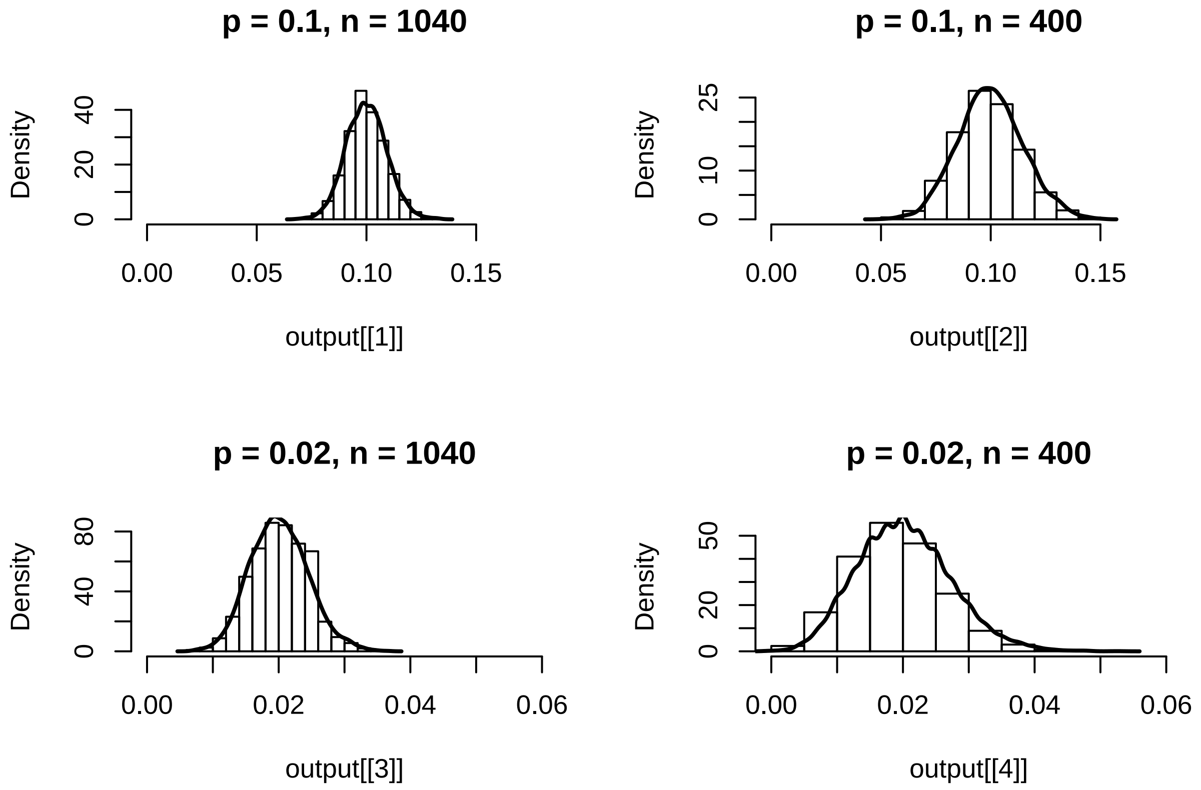
```
## Loading required package: purrr

## Loading required package: plyr

##
## Attaching package: 'plyr'

## The following object is masked from 'package:purrr':
##
##     compact
```

```r
par(mfrow = c(2, 2))
hist(output[[1]], main = "p = 0.1, n = 1040", xlim = c(0, 0.18), freq = F); lines(density(output[[1]]),
hist(output[[2]], main = "p = 0.1, n = 400", xlim = c(0, 0.18), freq = F); lines(density(output[[2]]),
hist(output[[3]], main = "p = 0.02, n = 1040", xlim = c(0, 0.06), freq = F); lines(density(output[[3]])
hist(output[[4]], main = "p = 0.02, n = 400", xlim = c(0, 0.06), freq = F); lines(density(output[[4]]),
```



# JN: The sampling distribution is affected by n; larger the n would better approximate the data toward a normal distribution. Smaller the n would fatten the distribution. The center of each sampling distribution would be the corresponding p value. Each sampling distribution is based off 5000 iterations varied by the size of n and proportion value. Seed is set to a constant of 1234 for each sampling.

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

11. If you refer to Table 6, you'll find that Australia has a sample proportion of 0.1 on a sample size of 1040, and that Ecuador has a sample proportion of 0.02 on 400 subjects. Let's suppose for this exercise that these point estimates are actually the truth. Then given the shape of their respective sampling distributions, do you think it is sensible to proceed with inference and report margin of errors, as the reports does? # JN: Yes, because in both cases they meet the "independence" and "success-failure" conditions.

---

## 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

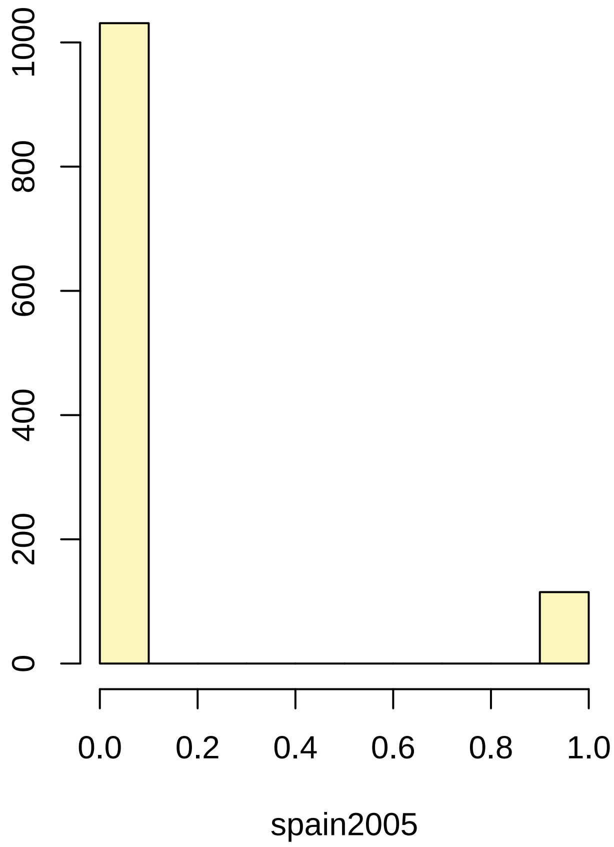
2012?

*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.

```
spain <- atheism %>%
        dplyr::filter(nationality == 'Spain')
spain$response <- ifelse(spain$response == "non-atheist", 0, 1)
spain2005 <- spain$response[spain$year == 2005]
spain2012 <- spain$response[spain$year == 2012]

inference(y = spain2005, est = "mean", type = "ci", null = 0,
          alternative = "twosided", method = "theoretical",
          conflevel = 0.95)
```

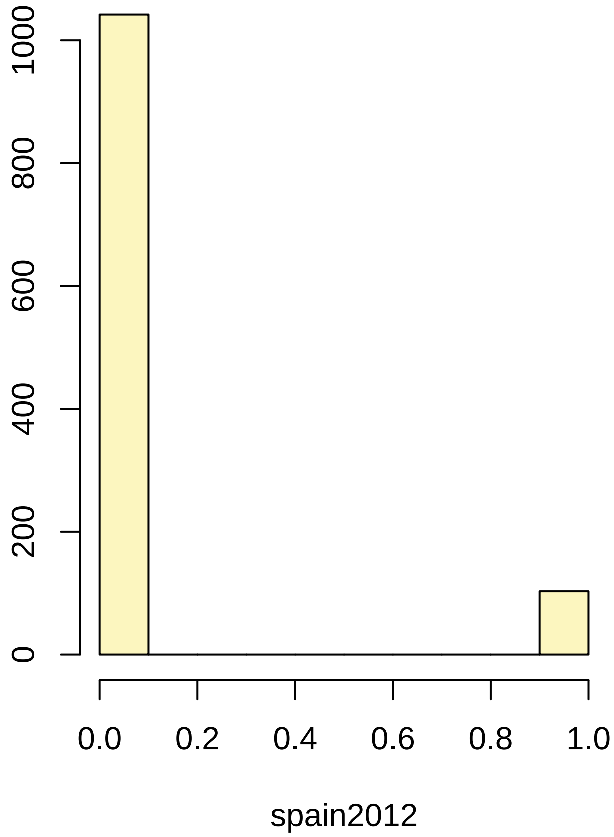
```
## Single mean
## Summary statistics:
```



```
## mean = 0.1003 ;  sd = 0.3006 ;  n = 1146
## Standard error = 0.0089
## 95 % Confidence interval = ( 0.0829 , 0.1178 )
```

```
inference(y = spain2012, est = "mean", type = "ci", null = 0,
          alternative = "twosided", method = "theoretical",
          conflevel = 0.95)
```

```
## Single mean
## Summary statistics:
```



```
## mean = 0.09 ;  sd = 0.2862 ;  n = 1145
## Standard error = 0.0085
## 95 % Confidence interval = ( 0.0734 , 0.1065 )
```

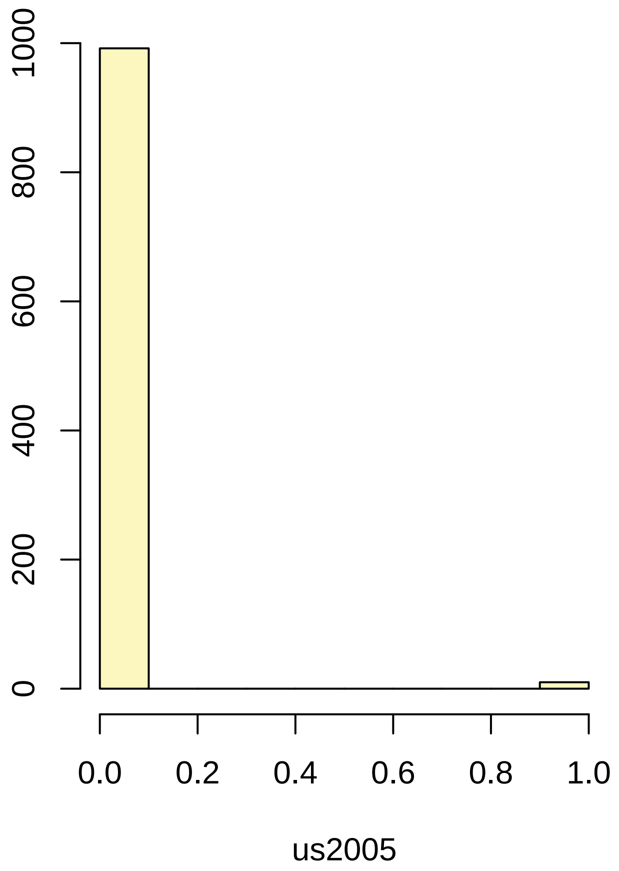
**JN: yes they do overlap. This overlap of the confidence intervals of 2005 and 2012 signals that the survey result in these two years are not significantly different.**

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

```
us <- atheism %>%
        dplyr::filter(nationality == 'United States')
us$response <- ifelse(us$response == "non-atheist", 0, 1)
us2005 <- us$response[us$year == 2005]
us2012 <- us$response[us$year == 2012]

inference(y = us2005, est = "mean", type = "ci", null = 0,
          alternative = "twosided", method = "theoretical",
          conflevel = 0.95)
```

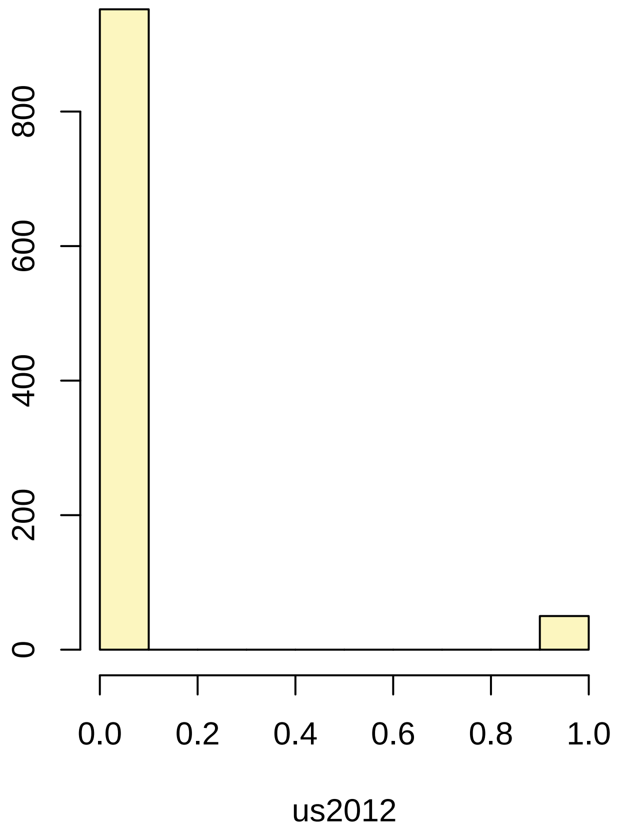
```
## Single mean
## Summary statistics:
```



```
## mean = 0.01 ;  sd = 0.0995 ;  n = 1002
## Standard error = 0.0031
## 95 % Confidence interval = ( 0.0038 , 0.0161 )
```

```
inference(y = us2012, est = "mean", type = "ci", null = 0,
          alternative = "twosided", method = "theoretical",
          conflevel = 0.95)
```

```
## Single mean
## Summary statistics:
```



```
## mean = 0.0499 ;  sd = 0.2178 ;  n = 1002
## Standard error = 0.0069
## 95 % Confidence interval = ( 0.0364 , 0.0634 )
```

# JN: no the confidence intervals do not overlap. It indicates that there's a significant change in the percent of self-identified atheists in the US in 2012. A significant jump from 1% in 2005 to 5% in 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. # JN: If indeed there's no change for all countries, we would still expect to find "significant" difference in about 5% of the countries due to chance. This "5%" is the alpha level (% of making type 1 error) that we have chosen, i.e. that's the percent of rejecting the null hypothesis when we actually shouldn't (type 1 error). The reason for why we should not reject the null hypothesis is because the observed difference is due to chance only.
- 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. # JN: In this case, we don't know p but we want to maximize the certainty of our margin of error. So, let's assume that our p is equal to 0.5. Since we want to have our margin of

error less than 1% with 95% confidence, we can do reverse engineering to figure out the size of n by plugging in the numbers,

# p = 0.5

# critical value = 1.96 (for 95% confidence)

# standard error (se) = sqrt((p * (1-p)) / n)

# margin of error (me) = 1.96 * se

# me = 1.96 * sqrt((p * (1-p)) / n)

# 0.01 = 1.96 * sqrt((0.5 * (1-0.5)) / n)

# n = 9604

# as a result, we should expect to get at least 9604+1 samples (make sure the me < 1%) to fulfill the required guidelines, i.e 1.96 * (sqrt((0.5)*(0.5)) / sqrt(9605)) = 0.009999479