# 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*?

   **The findings are sample statistics, as all people in the world were surely not polled for this study.**

2. The title of the report is "Global Index of Religiosity and Atheism". To generalize the report's findings to the global human population, what must we assume about the sampling method? Does that seem like a reasonable assumption?

   **We have to assume that the study was the result of a random sample and that those samples were representative of the larger global population. It seems like a reasonable assumption given the considerable sample sizes used (page 15).**

## 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?

**Each row of Table 6 corresponds with the results of sample within the country, while each row of the `atheism` data corresponds to a single response in the survey.**

To investigate the link between these two ways of organizing this data, take a look at the estimated proportion of atheists in the United States. Towards the bottom of Table 6, we see that this is 5%. We should be able to come to the same number using the `atheism` data.

4. Using the command below, create a new dataframe called `us12` that contains only the rows in `atheism` associated with respondents to the 2012 survey from the United States. Next, calculate the proportion of atheist responses. Does it agree with the percentage in Table 6? If not, why?

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

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

**Yes our calculated proportion agrees with the survey.**

## 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.

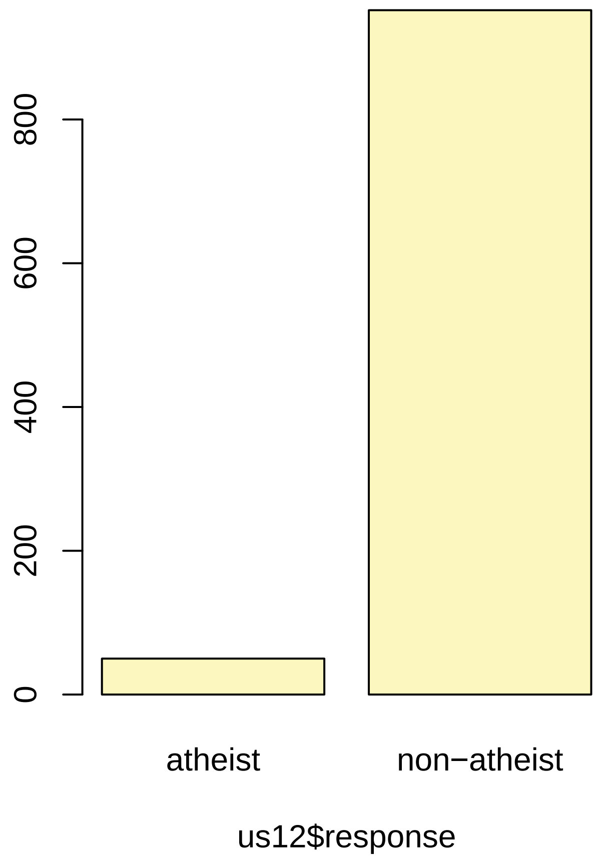
5. Write out the conditions for inference to construct a 95% confidence interval for the proportion of atheists in the United States in 2012. Are you confident all conditions are met?

**The conditions we need to construct a confidence interval are: independence and the success-failure condition. For independence, we assume that the samples are random, as each comes from a different person. For the success-failure condition, we check to see that there are at least 10 successes and 10 failures in the population by using the sample proportion as an estimator. Looking at the data we have 50 "successes" and 952 "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.

```r
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 in the R output is 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.

```r
# Get data for our countries
india12 <- subset(atheism, nationality == "India" & year == "2012")
austria12 <- subset(atheism, nationality == "Austria" & year == "2012")

# Check proportions for success-failure
pIndia <- nrow(india12[india12$response=="atheist",])
```

```r
qIndia <- nrow(india12[india12$response!="atheist",])
pIndia
```

```
## [1] 33
```

```r
qIndia
```

```
## [1] 1059
```

```r
pAustria <- nrow(austria12[austria12$response=="atheist",])
qAustria <- nrow(austria12[austria12$response!="atheist",])
pAustria
```

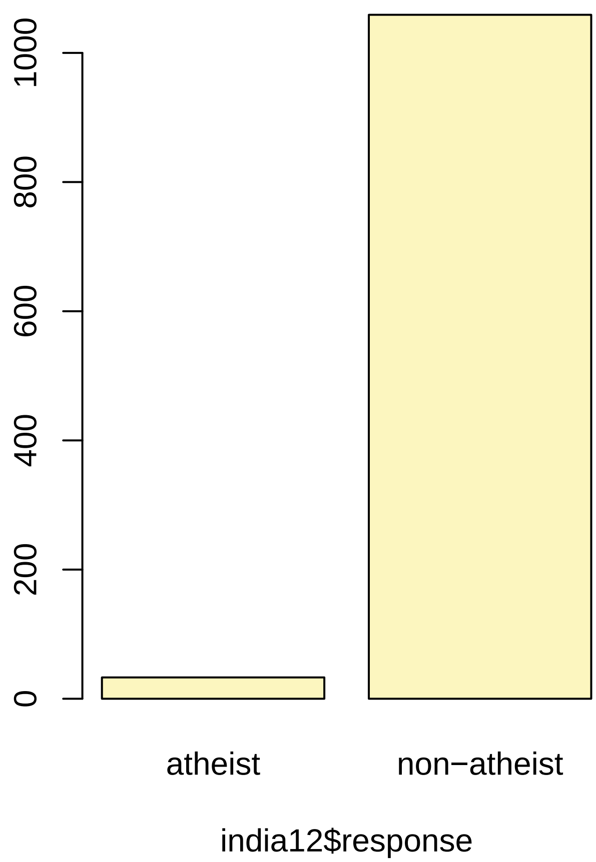
```
## [1] 100
```

```r
qAustria
```

```
## [1] 902
```

```r
inference(india12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

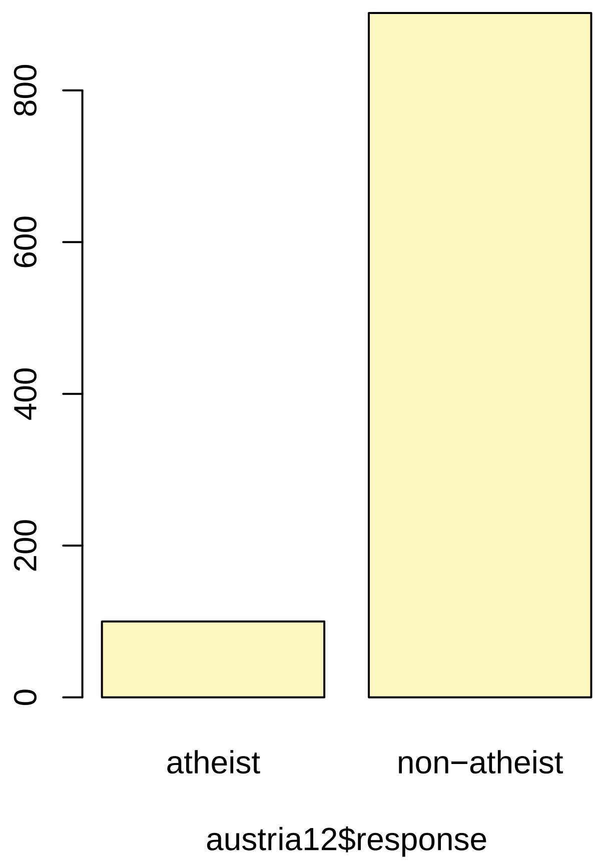
```
## Single proportion -- success: atheist
## Summary statistics:
```




```
## p_hat = 0.0302 ;  n = 1092
## Check conditions: number of successes = 33 ; number of failures = 1059
## Standard error = 0.0052
## 95 % Confidence interval = ( 0.0201 , 0.0404 )
```

```r
inference(austria12$response, est = "proportion", type = "ci",
          method = "theoretical", success = "atheist")
```

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




```
## p_hat = 0.0998 ;  n = 1002
## Check conditions: number of successes = 100 ; number of failures = 902
## Standard error = 0.0095
## 95 % Confidence interval = ( 0.0812 , 0.1184 )
```

**The conditions for inference are met (independence and number of successes)**

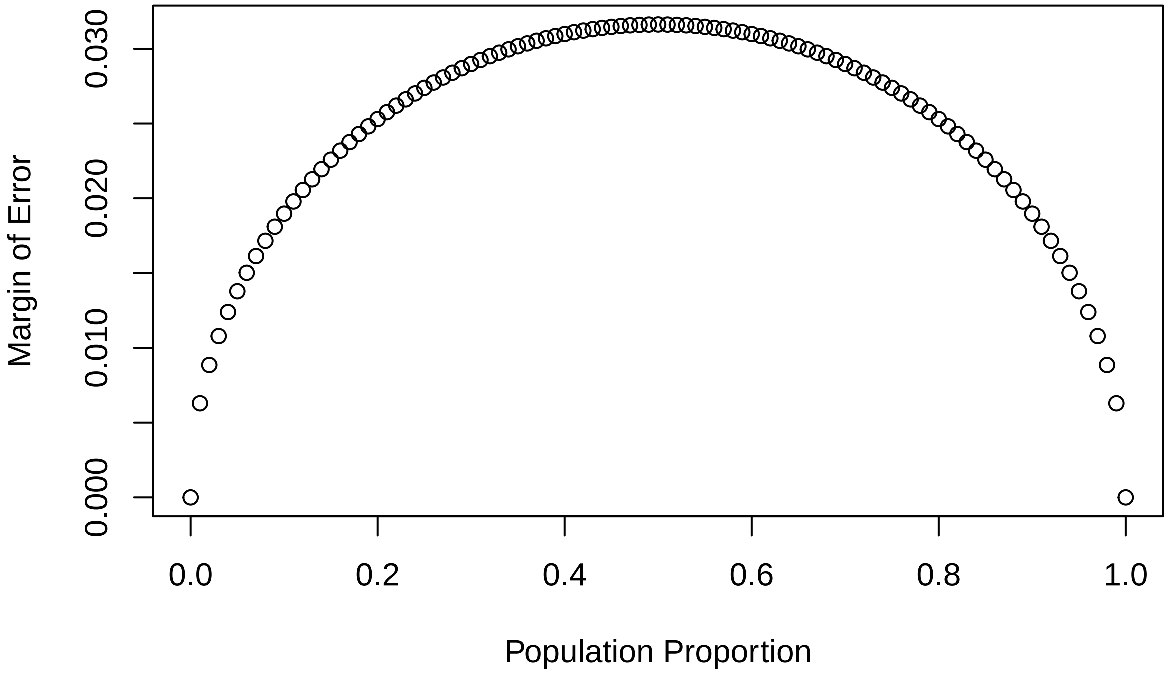
### How does the proportion affect the margin of error?

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 proportion moves from the extremes to 0.05 the margin of error increases.**

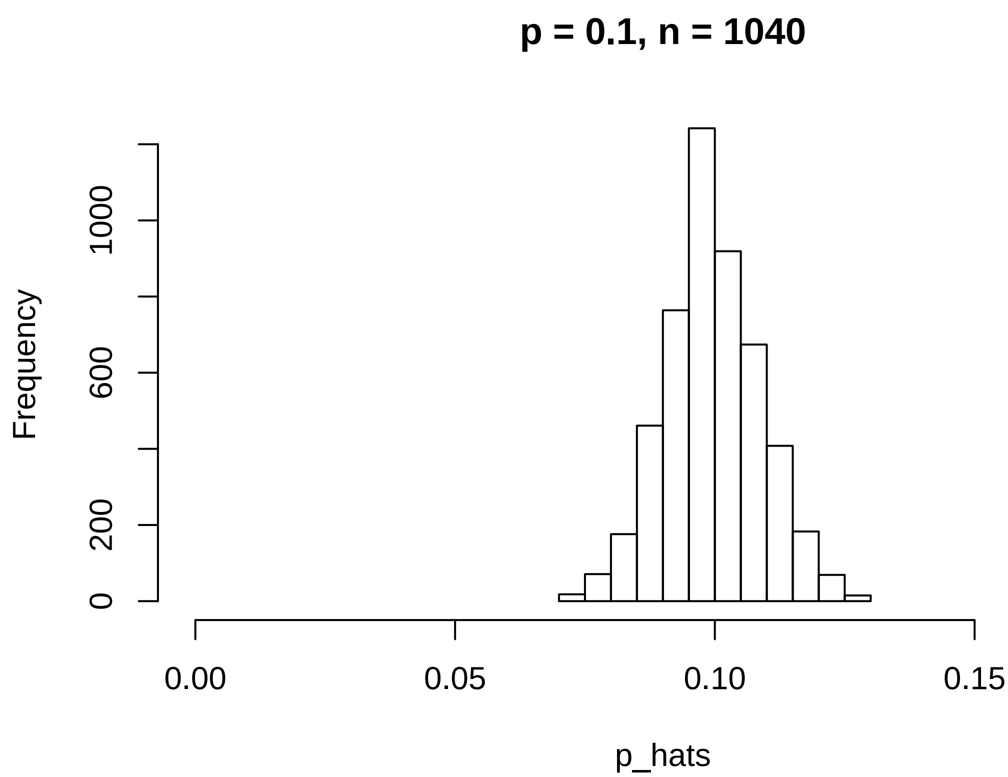
## Success-failure condition

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.

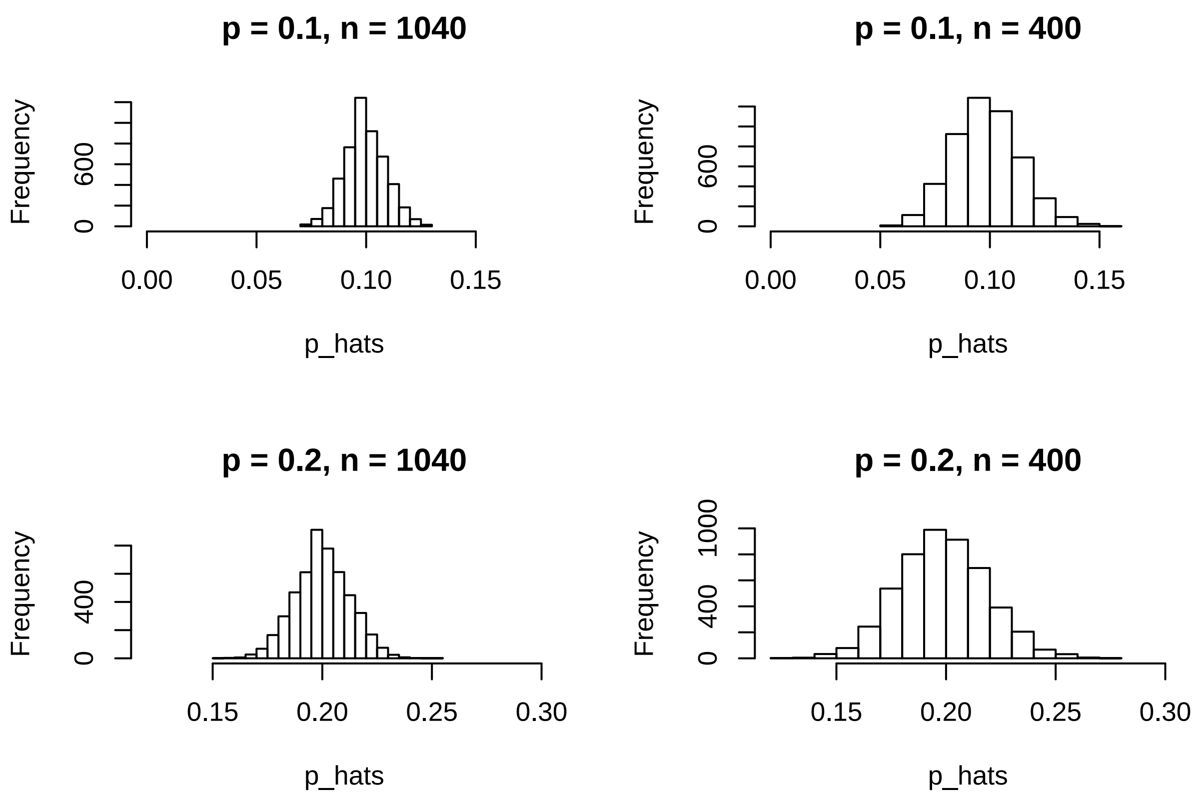
9. Describe the sampling distribution of sample proportions at $n = 1040$ and $p = 0.1$. Be sure to note the center, spread, and shape.
   *Hint:* Remember that R has functions such as `mean` to calculate summary statistics.

**The sampling distribution is nearly normal and mostly symmetric, centered at 0.09969.**

10. Repeat the above simulation three more times but with modified sample sizes and proportions: for $n = 400$ and $p = 0.1$, $n = 1040$ and $p = 0.02$, and $n = 400$ and $p = 0.02$. Plot all four histograms together by running the `par(mfrow = c(2, 2))` command before creating the histograms. You may need to expand the plot window to accommodate the larger two-by-two plot. Describe the three new sampling distributions. Based on these limited plots, how does $n$ appear to affect the distribution of $\hat{p}$?

How does $p$ affect the sampling distribution?

```r
par(mfrow = c(2, 2))

# Get plot from above
plot1 <- hist(p_hats, main = "p = 0.1, n = 1040", xlim = c(0, 0.18))

# Plot 2
p <- 0.1
n <- 400
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
}

plot2 <- hist(p_hats, main = "p = 0.1, n = 400", xlim = c(0, 0.18))

# Plot 3
p <- 0.2
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
}

plot3 <- hist(p_hats, main = "p = 0.2, n = 1040", xlim = c(0.12, 0.30))

# Plot 4
p <- 0.2
n <- 400
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
}

plot4 <- hist(p_hats, main = "p = 0.2, n = 400", xlim = c(0.12, 0.30))
```




**A larger proportion spreads out the distribution more, which aligns with the formula for margin of error. A smaller $n$ seems to make the margin of error greater and less-normal.**

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

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

11. If you refer to Table 6, you'll find that Australia has a sample proportion of 0.1 on a sample size of 1040, and that Ecuador has a sample proportion of 0.02 on 400 subjects. Let's suppose for this exercise that these point estimates are actually the truth. Then given the shape of their respective sampling distributions, do you think it is sensible to proceed with inference and report margin of errors, as the reports does?

**I would say it is fair to proceed with the inference**

---

## On your own

The question of atheism was asked by WIN-Gallup International in a similar survey that was conducted in 2005. (We assume here that sample sizes have remained the same.) Table 4 on page 13 of the report summarizes survey results from 2005 and 2012 for 39 countries.

- Answer the following two questions using the `inference` function. As always, write out the hypotheses for any tests you conduct and outline the status of the conditions for inference.

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

```
spain05 <- subset(atheism, nationality == "Spain" & year == "2005")
spain12 <- subset(atheism, nationality == "Spain" & year == "2012")
```

**First we check conditions for inference. We already trust that the study has independent responses so next we check to see if the success-failure condition is met for the data sets

```
pSpain05 <- nrow(spain05[spain05$response=="atheist",])
qSpain05 <- nrow(spain05[spain05$response!="atheist",])

pSpain12 <- nrow(spain12[spain12$response=="atheist",])
qSpain12 <- nrow(spain12[spain12$response!="atheist",])
```

**We have at least 10 successes and failures for both data sets, so we should be ok to assume normality of the sampling distributions.**

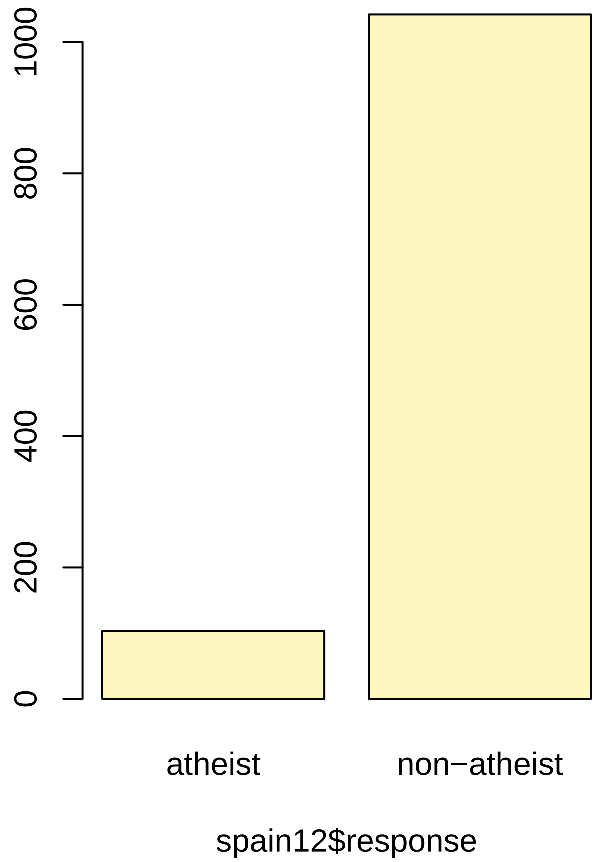
**Our hypothesis is that the sample mean of the 2012 distribution is the same as it was in 2005 and the difference is merely due to random chance. Said mathematically:**

$$H_0 : \mu_{2012} = \mu_{2005} \; ; H_A : \mu_{2012} \neq \mu_{2005}$$

Now we check the confidence intervals to see if there is overlap.

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

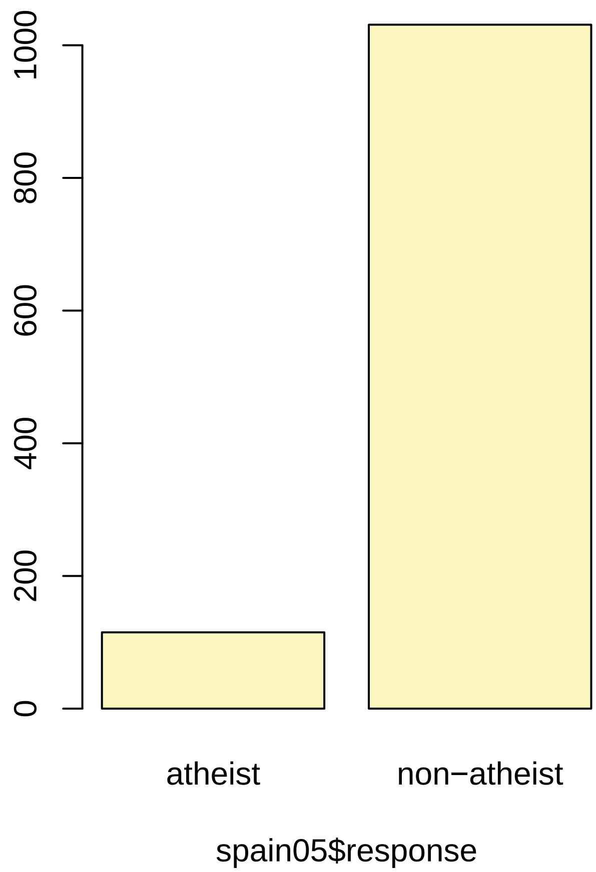
```
## Single proportion -- success: atheist
## Summary statistics:
```




```
## 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 )
```

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

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



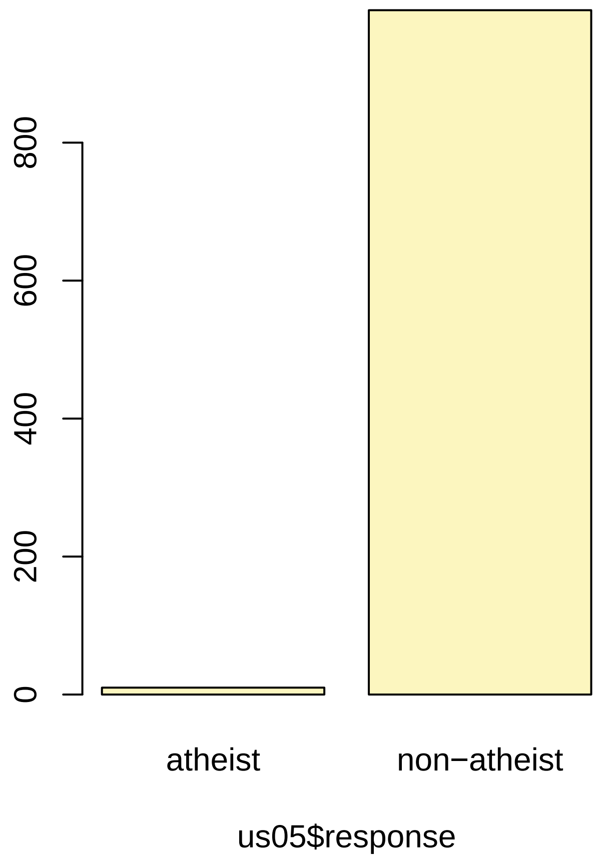

```
## p_hat = 0.1003 ;  n = 1146
## Check conditions: number of successes = 115 ; number of failures = 1031
## Standard error = 0.0089
## 95 % Confidence interval = ( 0.083 , 0.1177 )
```

**Looking at the confidence intervals, they overlap considerably. So, there is not enough compelling evidence to reject the null hypothesis and we continue to accept that the differences we are seeing in $\hat{p}$ are due to random chance and not indicative of a change in the population proportion ($p$).**

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

```
us05 <- subset(atheism, nationality == "United States" & year == "2005")
pUS05 <- nrow(us05[us05$response == "atheist",])

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 )
```

**Yes, there is convincing evidence of a change in population proportion. Comparing the confidence interval for 2005 with the one from 2012 we calculated above in #5, there is no overlap at the 95% confidence interval.**

- 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.
- 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.

**Since we do not know what to expect for $p$, we use the worst-case of 0.50 because, as shown above, that gives us the worst margin of error. We can then calculate the minimum sample size to achieve this maximum level of error:**

$$1.96^2 \times \frac{0.5^2}{n} < 0.01^2$$
$$3.8416 \times \frac{0.025}{0.0001} < n = 960.4$$

**We would need a sample size of at leaast 961 to ensure a margin of error no greater than 0.01.**