Chapter 6 - Inference for Categorical Data

Samuel I Kigamba

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*?

   They appear to be 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?

   The assumptions to be made are the sample selection was random and the observations were independent.

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")
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
```

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

   Each row of the table represents summary of responses/responsents by country with each response being an individual observation/response in year 2012.

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?

   Yes it agrees.

```
us12 <- subset(atheism, nationality == "United States" & year == "2012")

head(us12)
```

```
##         nationality    response year
## 49926 United States non-atheist 2012
## 49927 United States non-atheist 2012
## 49928 United States non-atheist 2012
## 49929 United States non-atheist 2012
## 49930 United States non-atheist 2012
## 49931 United States non-atheist 2012
```

```
sum(us12$response=='atheist')/length(us12$response =='atheist')
```

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

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?

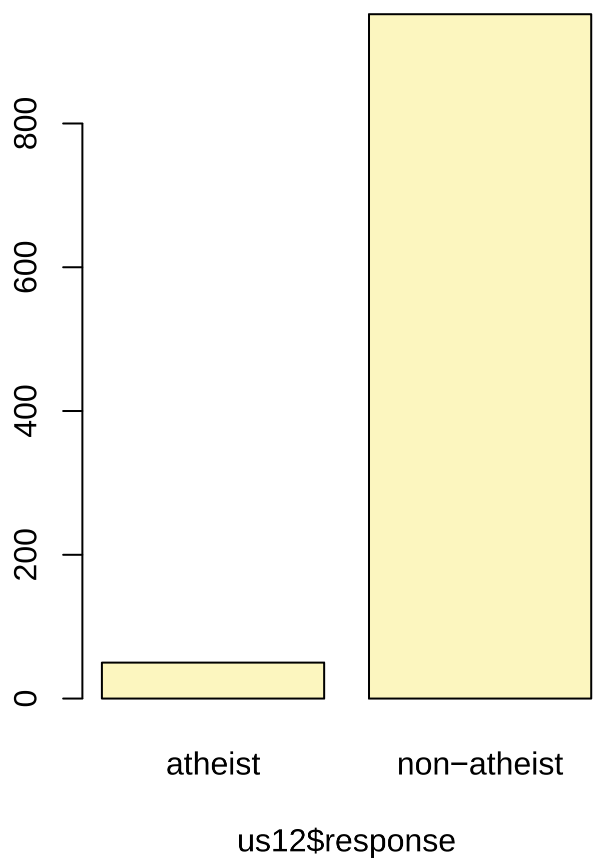
   We need a proportion sufficiently large enough to pass the success-failure condition n(p) >= 10 and n(1-p) >= 10 and made up of independent observations. Bothe conditions are met.

```
#install.packages('BHH2')
library(BHH2)
```

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?

   From the output the margin of error is 1.35%.

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.

Country: Japan

```
Japan12 <- subset(atheism, nationality == "Japan" & year == "2012")

head(Japan12)
```

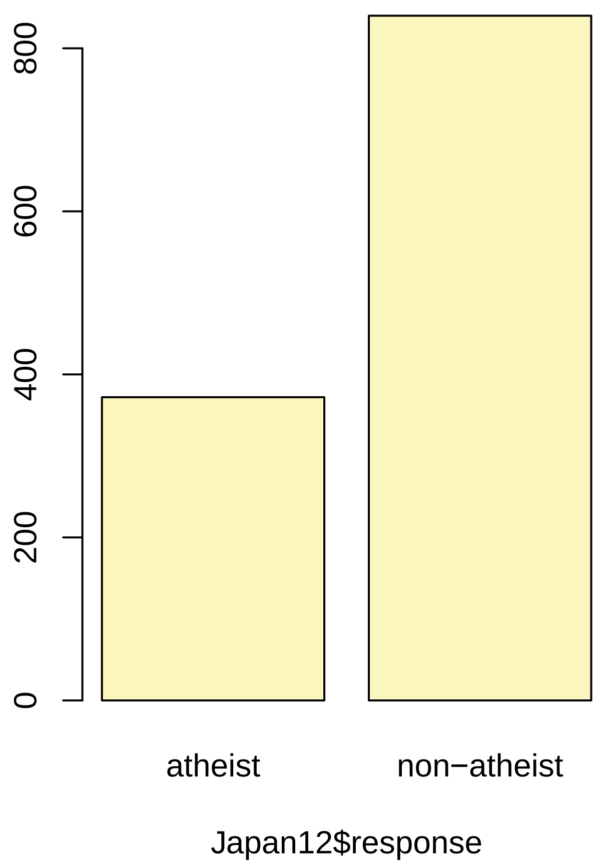
```
##       nationality    response year
## 25743       Japan non-atheist 2012
## 25744       Japan     atheist 2012
## 25745       Japan     atheist 2012
## 25746       Japan non-atheist 2012
## 25747       Japan     atheist 2012
## 25748       Japan non-atheist 2012
```

```
sum(Japan12$response=='atheist')/length(Japan12$response =='atheist')
```

```
## [1] 0.3069307
```

```
inference(Japan12$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 )
```

Margin of error is 2.59%

Country: Germany

```r
Germany12 <- subset(atheism, nationality == "Germany" & year == "2012")

head(Germany12)
```

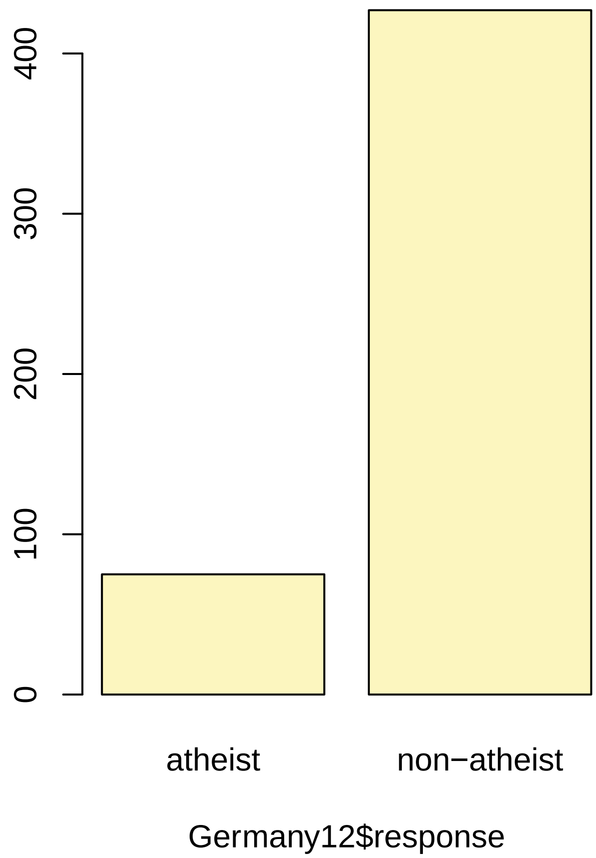
```
##       nationality    response year
## 18310     Germany     atheist 2012
## 18311     Germany non-atheist 2012
## 18312     Germany non-atheist 2012
## 18313     Germany non-atheist 2012
## 18314     Germany non-atheist 2012
## 18315     Germany non-atheist 2012
```

```r
sum(Germany12$response=='atheist')/length(Germany12$response =='atheist')
```

```
## [1] 0.1494024
```

```r
inference(Germany12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

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




```
## p_hat = 0.1494 ;  n = 502
## Check conditions: number of successes = 75 ; number of failures = 427
## Standard error = 0.0159
## 95 % Confidence interval = ( 0.1182 , 0.1806 )
```

Margin of error is 3.12%

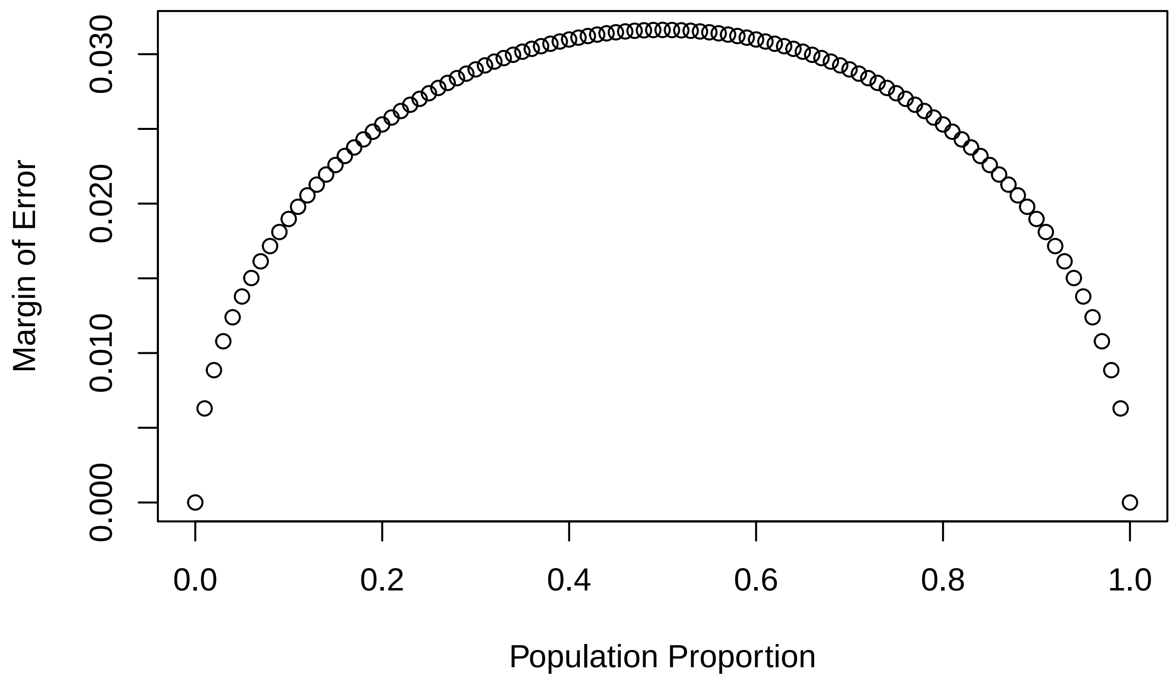
How does the proportion affect the margin of error?

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`.

```
A population proportion of 0.5 gives the highest margin of error. The further away the prop
```

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?

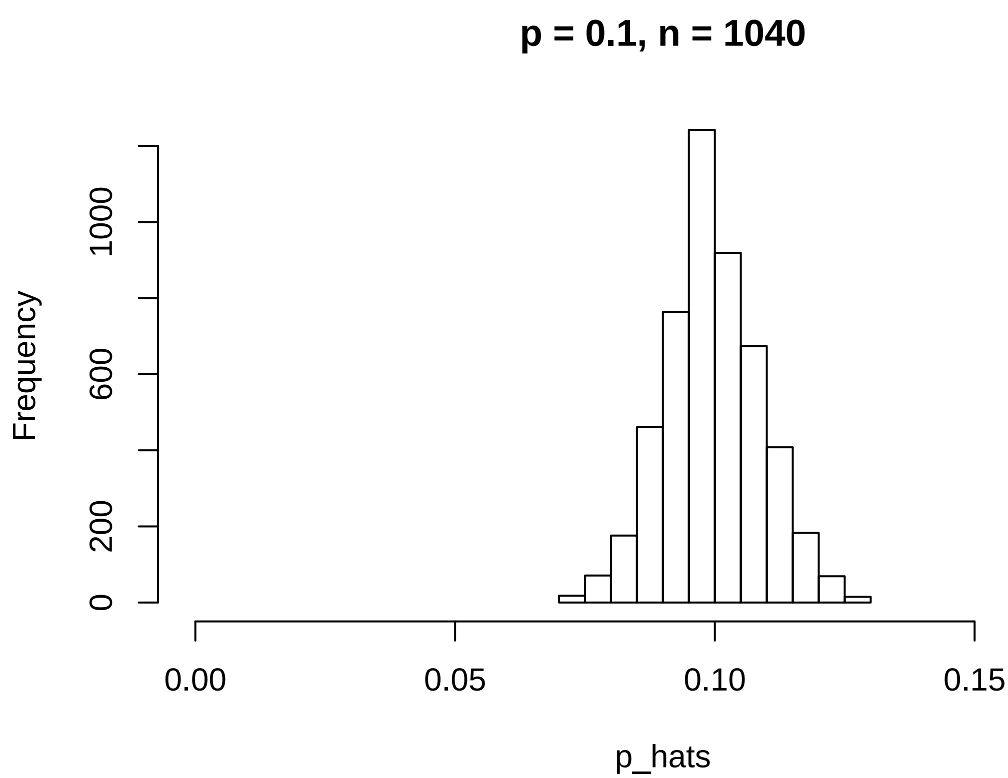
The short answer is: nothing. You could argue that we would be fine with 9 or that we really should be using 11. What is the "best" value for such a rule of thumb is, at least to some degree, arbitrary. However, when np and $n(1-p)$ reaches 10 the sampling distribution is sufficiently normal to use confidence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

```r
mean(p_hats)
```

```
## [1] 0.09969
```

```r
sd(p_hats)
```

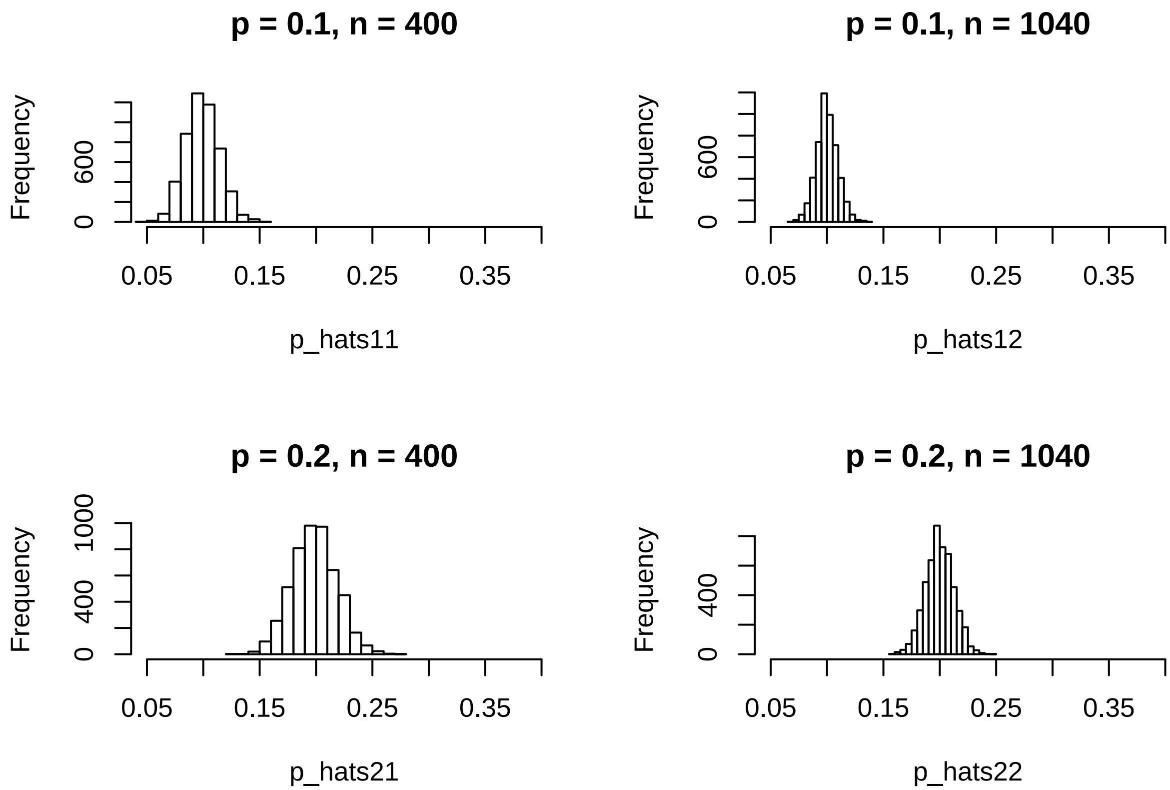
```
## [1] 0.009287382
```

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
p1 <- 0.1
p2 <- 0.2
n1 <- 400
n2 <- 1040
p_hats11 <- rep(0, 5000)
p_hats12 <- rep(0, 5000)
p_hats21 <- rep(0, 5000)
p_hats22 <- rep(0, 5000)

for(i in 1:5000){
  samp11 <- sample(c("atheist", "non_atheist"), n1, replace = TRUE, prob = c(p1, 1-p1))
  p_hats11[i] <- sum(samp11 == "atheist")/n1
  samp12 <- sample(c("atheist", "non_atheist"), n2, replace = TRUE, prob = c(p1, 1-p1))
  p_hats12[i] <- sum(samp12 == "atheist")/n2
  samp21 <- sample(c("atheist", "non_atheist"), n1, replace = TRUE, prob = c(p2, 1-p2))
  p_hats21[i] <- sum(samp21 == "atheist")/n1
  samp22 <- sample(c("atheist", "non_atheist"), n2, replace = TRUE, prob = c(p2, 1-p2))
  p_hats22[i] <- sum(samp22 == "atheist")/n2
}

par(mfrow = c(2, 2))
hist(p_hats11, main = "p = 0.1, n = 400", xlim = c(0.05, 0.4))
hist(p_hats12, main = "p = 0.1, n = 1040", xlim = c(0.05, 0.4))
hist(p_hats21, main = "p = 0.2, n = 400", xlim = c(0.05, 0.4))
hist(p_hats22, main = "p = 0.2, n = 1040", xlim = c(0.05, 0.4))
```



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?

    ```
    Australia meets the inference requirements (0.1 by 1040 = 104) while equador doesnt (0.02 by 400 =
    ```

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.

```
spn05 <- subset(atheism, nationality == "Spain" & year == "2005")
head(spn05)
```

```
##       nationality    response year
## 61188       Spain non-atheist 2005
## 61189       Spain non-atheist 2005
## 61190       Spain non-atheist 2005
## 61191       Spain non-atheist 2005
## 61192       Spain non-atheist 2005
## 61193       Spain non-atheist 2005
```

```
atheist_spn05 <- sum(spn05$response == "atheist")
non_atheists_spn05 <- sum(spn05$response != "atheist")
total_spn05 <- atheist_spn05 + non_atheists_spn05
atheist_spn05/total_spn05 *100
```

```
## [1] 10.0349
```

```
spn12 <- subset(atheism, nationality == "Spain" & year == "2012")
head(spn12)
```

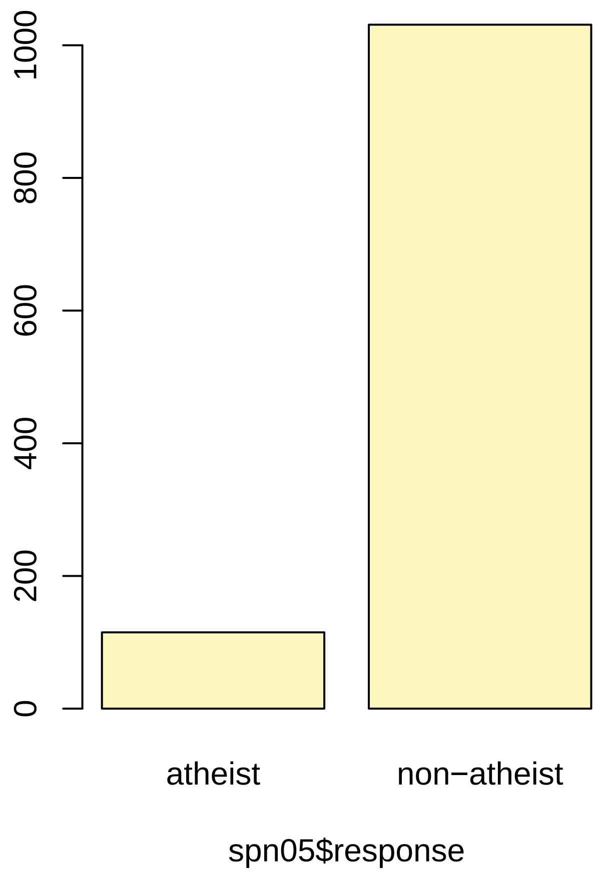
```
##       nationality    response year
## 45230       Spain non-atheist 2012
## 45231       Spain non-atheist 2012
## 45232       Spain non-atheist 2012
## 45233       Spain non-atheist 2012
## 45234       Spain non-atheist 2012
## 45235       Spain non-atheist 2012
```

```
atheist_spn12 <- sum(spn12$response == "atheist")
non_atheists_spn12 <- sum(spn12$response != "atheist")
total_spn12 <- atheist_spn12 + non_atheists_spn12
atheist_spn12/total_spn12 *100
```

```
## [1] 8.995633
```

```
par(mfrow = c(1, 2))
inference(spn05$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

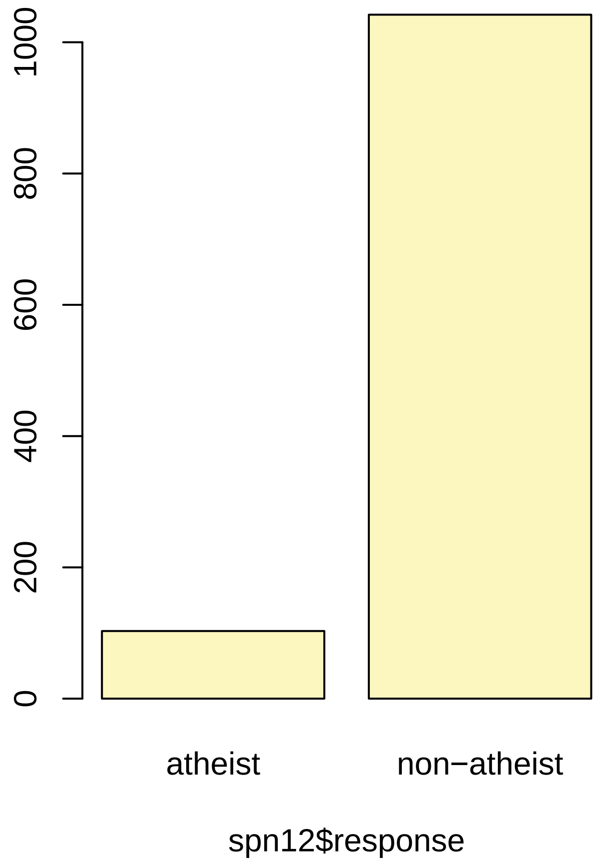
```
## Single proportion -- success: atheist
## Summary statistics:
```




```
## 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(spn12$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 )
```

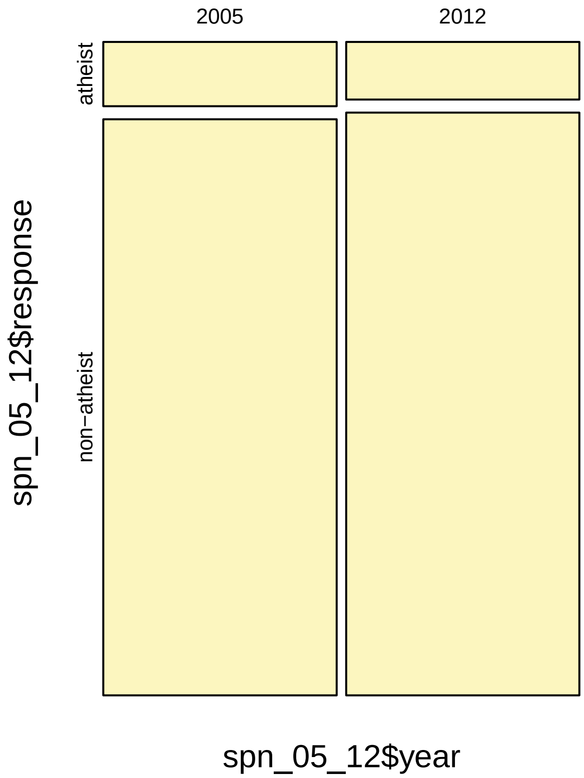
```
spn_05_12 <- subset(atheism, nationality == "Spain" & year == "2005"  | nationality == "Spain" & year ==
inference(y = spn_05_12$response, x = spn_05_12$year, est = "proportion",type = "ht", null = 0, alterna
```

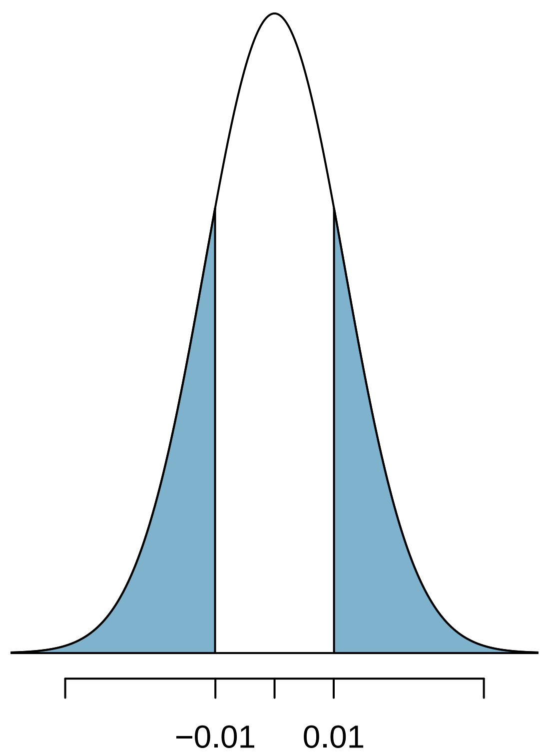
```
## Warning: Explanatory variable was numerical, it has been converted to
## categorical. In order to avoid this warning, first convert your explanatory
## variable to a categorical variable using the as.factor() function.
```

```
## 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
```






```
p_spn05 = 0.1003
n_spn05 = 1146
p_spn12 = 0.09
n_spn12 = 1145

PE_spn = p_spn12 - p_spn05

SE_spn = sqrt((p_spn05*(1-p_spn05)/n_spn05)+(p_spn12*(1-p_spn12)/n_spn12))
SE_spn
```

```
## [1] 0.01225854
```

Interval for difference between proportion in 2005 and 2012

```r
PE_spn + (1.96*SE_spn)
```

```
## [1] 0.01372674
```

```r
PE_spn - (1.96*SE_spn)
```

```
## [1] -0.03432674
```

Conclusion:

The observations are independent and are based on random distribution. The observations also pass the success failure condition.

Since the p-value is greater than .05, we fail to reject the null hypothesis and conclude there is no convincing evidence that there is a change in the atheism index in Spain from 2005 to 2012 since the confidence interval (95%) overlap.

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

```r
usa05 <- subset(atheism, nationality == "United States" & year == "2005")
head(usa05)
```

```
##         nationality    response year
## 67899 United States non-atheist 2005
## 67900 United States non-atheist 2005
## 67901 United States non-atheist 2005
## 67902 United States non-atheist 2005
## 67903 United States non-atheist 2005
## 67904 United States non-atheist 2005
```

```r
dim(usa05)
```

```
## [1] 1002    3
```

```r
atheist_usa05 <- sum(usa05$response == "atheist")
atheist_usa05
```

```
## [1] 10
```

```r
non_atheists_usa05 <- sum(usa05$response != "atheist")
head(non_atheists_usa05)
```

```
## [1] 992
```

```r
total_usa05 <- atheist_usa05 + non_atheists_usa05
total_usa05
```

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

```
atheist_usa05/total_usa05 *100
```

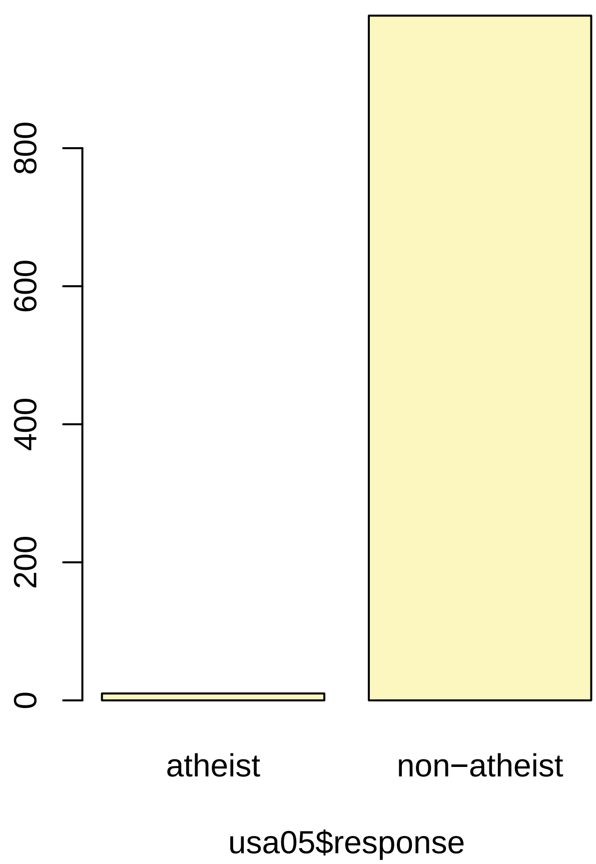
```
## [1] 0.998004
```

```
usa12 <- subset(atheism, nationality == "United States" & year == "2012")
atheist_usa12 <- sum(usa12$response == "atheist")
non_atheists_usa12 <- sum(usa12$response != "atheist")
total_usa12 <- atheist_usa12 + non_atheists_usa12
atheist_usa12/total_usa12 *100
```

```
## [1] 4.99002
```

```
par(mfrow = c(1, 2))
inference(usa05$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

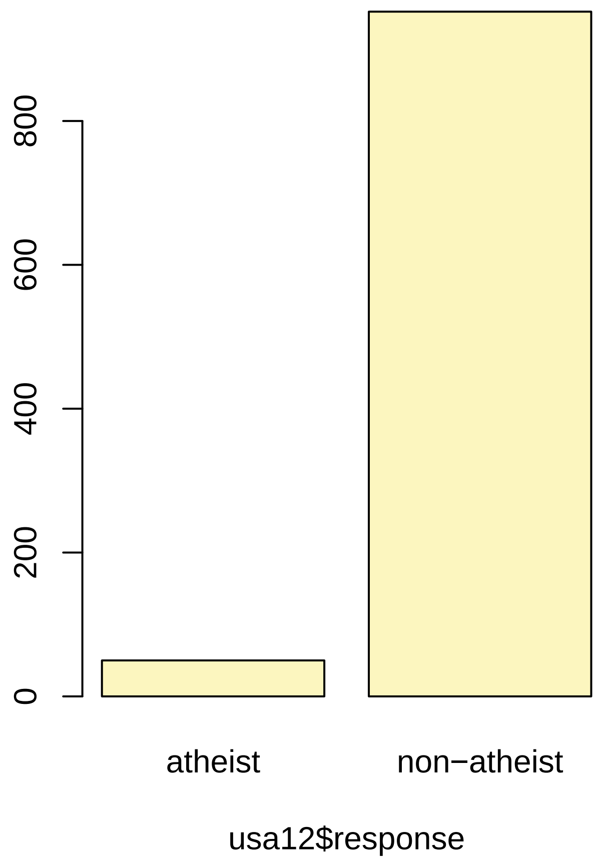
```
## Single proportion -- success: atheist
## Summary statistics:
```




```
## 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 )
```

```r
inference(usa12$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 )
```

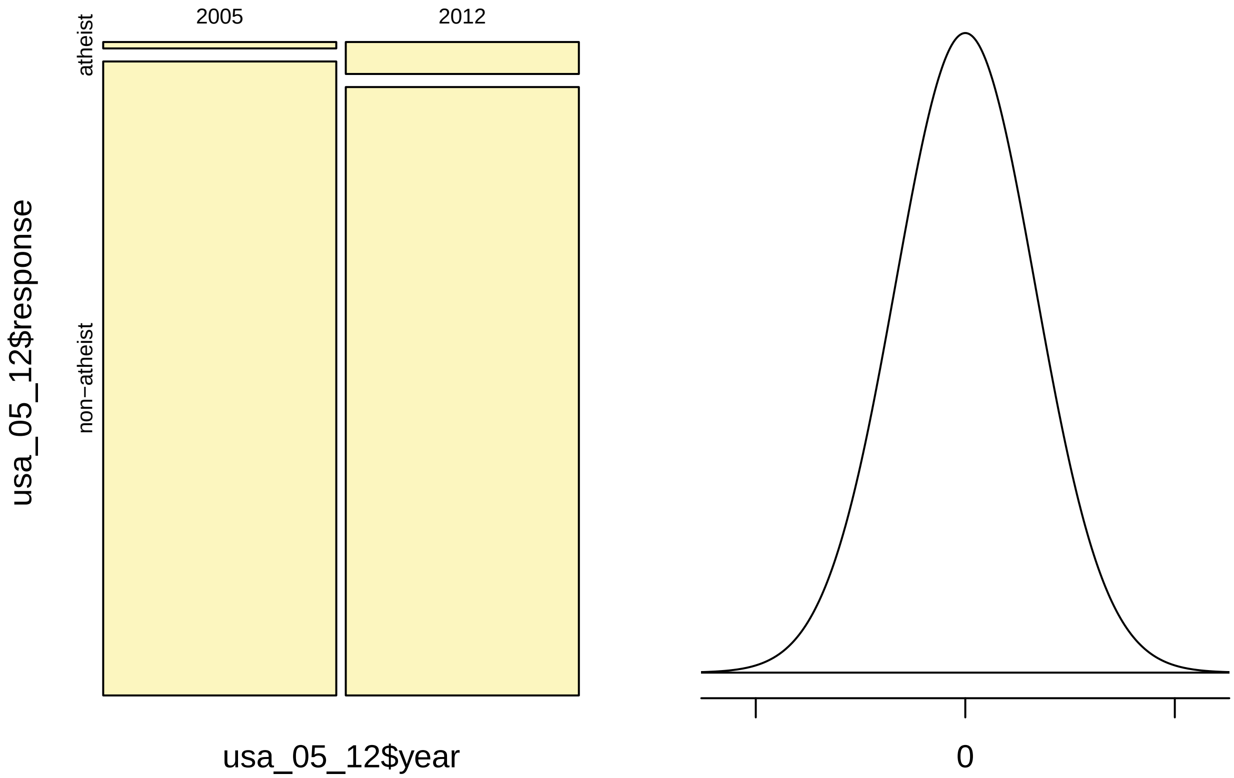
```r
usa_05_12 <- subset(atheism, nationality == "United States" & year == "2005"  | nationality == "United S
inference(y = usa_05_12$response, x = usa_05_12$year, est = "proportion",type = "ht", null = 0, alterna
```

```
## Warning: Explanatory variable was numerical, it has been converted to
## categorical. In order to avoid this warning, first convert your explanatory
## variable to a categorical variable using the as.factor() function.
```

```
## 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
```




```
p_usa05 = 0.01
n_usa05 = 1002
p_usa12 = 0.05
n_usa12 = 1002

PE_usa = p_usa12 - p_usa05
PE_usa
```

```
## [1] 0.04
```

```r
SE_usa = sqrt(((p_usa05*(1-p_usa05))/n_usa05)+((p_usa12*(1-p_usa12))/n_usa12))
SE_usa
```

```
## [1] 0.007568714
```

Interval for difference between proportion in 2005 and 2012

```r
PE_usa + (1.96*SE_usa)
```

```
## [1] 0.05483468
```

```r
PE_usa - (1.96*SE_usa)
```

```
## [1] 0.02516532
```

Conclusion:

Since p value is effectively 0, we can reject the Null Hypothesis and there is evidence that there is a change in the atheism index in the USA from 2005 to 2012. The control interval does NOT include 0 so we reject the null hypohtesis that p of 2012 - p of 2005 is zero.

- 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.

Response: Large point differences would detect a small change in the significance level.

A Type 1 error occures when we reject the null hypothesis when it is true. Since we are using a .05 significance level, we would expect 5% of countries to have a Type 1 error.

```r
paste("This would amount to ",.05*39,"countries.")
```

```
## [1] "This would amount to  1.95 countries."
```

- 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.

ME = z * SE = z * sqrt((p*(1-p))/n) SE = sqrt((p*(1-p))/n) n = (z/ME)^2 / (p*(1-p))

Usinging a p-value of 0.5, generate the maximum ME (refer to section above) of 0.1.

```r
p <- 0.5
z <- qnorm(0.975)
ME = 0.01
n = ( (z/ME)^2 )* ( p*(1-p) )
n
```

```
## [1] 9603.647
```

```
A population of 9604 is required to fulfil the requirements.
```