Lab 6 - Inference for categorical data

CUNY MSDA DATA 606

Duubar Villalobos Jimenez mydvtech@gmail.com

April 2, 2017

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:

http://www.wingia.com/web/files/richeditor/filemanager/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:

From my point of view these are *sample statistics* derived from the data 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?

Answer:

- We must assume that the findings were simple randomly selected.
- The groups are independent of each other.
- Sample size is large enough.
- One sample is good enough.

The data

Turn your attention to Table 6 (pages 15 and 16), which reports the sample size and response percentages for all 57 countries. While this is a useful format to summarize the data, we will base our analysis on the original data set of individual responses to the survey. Load this data set into R with the following command.

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

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

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.

Answer:

Each row of Table 6 correspond to a country with response percentages related to religious degree.

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?

** Answer:**

```
library(plyr)
library(knitr)
us12 <- subset(atheism, nationality == "United States" & year == "2012")
pus12athe <- count(us12$response == 'atheist')
names(pus12athe) <- c("atheist", "total")
pus12athe$percent <- pus12athe$total / sum(pus12athe$total) * 100
kable(pus12athe)
```

atheist	total	percent
FALSE	952	95.00998
TRUE	50	4.99002

The percentage on here **agrees** with the percentage in 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?

The conditions are:

- The sample observations are independent (We can be certain this condition is met since the sample size of 1002 for the US is definitely less than 10% of the U.S. population).
- We assume that the samples are randomly selected.
- The success-failure conditions require **np >= 10** and **n(1-p) >= 10**.

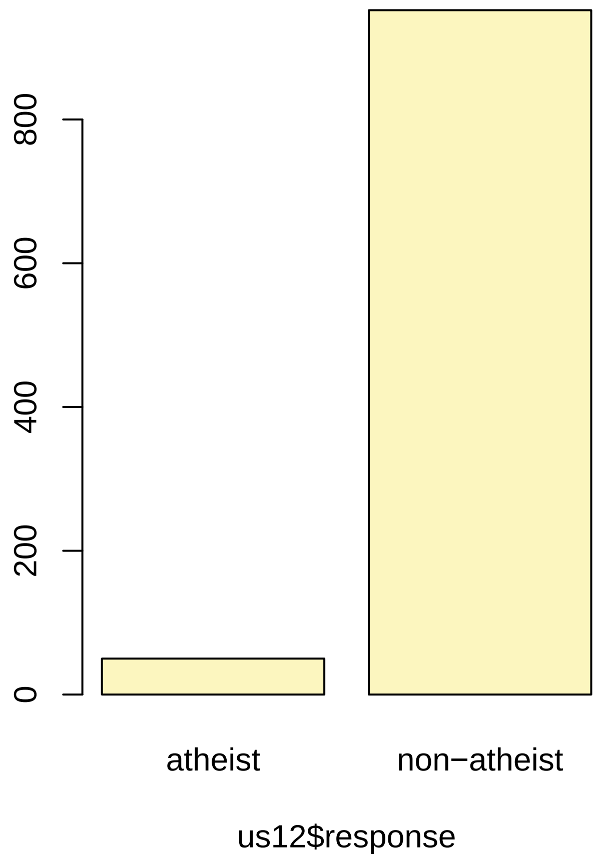
That is: np = 1002 * 0.05 = 50.1 and n(1-p) = 1002 * 0.95 = 951.9 so this condition is met.

If the conditions for inference are reasonable, we can either calculate the standard error and construct the interval by hand, or allow the `inference` function to do it for us.

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

```
## Warning: package 'BHH2' was built under R version 3.3.3
```

```
## 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:

Our **z** value for 95% confidence will be $z = 1.96$

From above we got **Standard error:** SE = 0.0069

```
z <- 1.96
SE <- 0.0069
ME <- z * SE
ME
```

```
## [1] 0.013524
```

The margin of error for US is 0.013524.

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:

The two countries are Brazil and Azerbaijan (Argentina as substitute).

The conditions for the inference for Brazil and Argentina are met.

- The sample size are much less than 10% of the total populations for each country.
- The sample sizes are big enough to satisfy the Success_failure condition.

However, the condition for Azerbaijan failed as the proportion for atheist is 0% (that is **np >= 10** condition is not met).

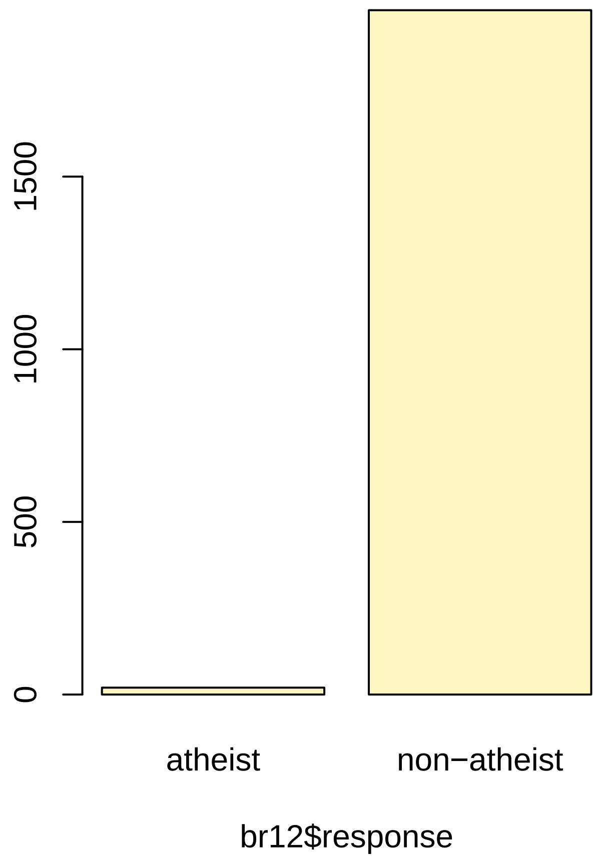
Country 1: Brazil

```
br12 <- subset(atheism, nationality == "Brazil" & year == "2012")
pbr12athe <- count(br12$response == 'atheist')
names(pbr12athe) <- c("atheist", "total")
pbr12athe$percent <- pbr12athe$total / sum(pbr12athe$total) * 100
kable(pbr12athe)
```

atheist	total	percent
FALSE	1982	99.000999
TRUE	20	0.999001

```
inference(br12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

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




```
## p_hat = 0.01 ;  n = 2002
## Check conditions: number of successes = 20 ; number of failures = 1982
## Standard error = 0.0022
## 95 % Confidence interval = ( 0.0056 , 0.0143 )
```

Our **z** value for 95% confidence will be $z = 1.96$

From above we got **Standard error:** $\text{SE} = 0.0022$

```r
z <- 1.96
SE <- 0.0022
ME <- z * SE
ME
```

```
## [1] 0.004312
```

The margin of error for Brazil is 0.004312.

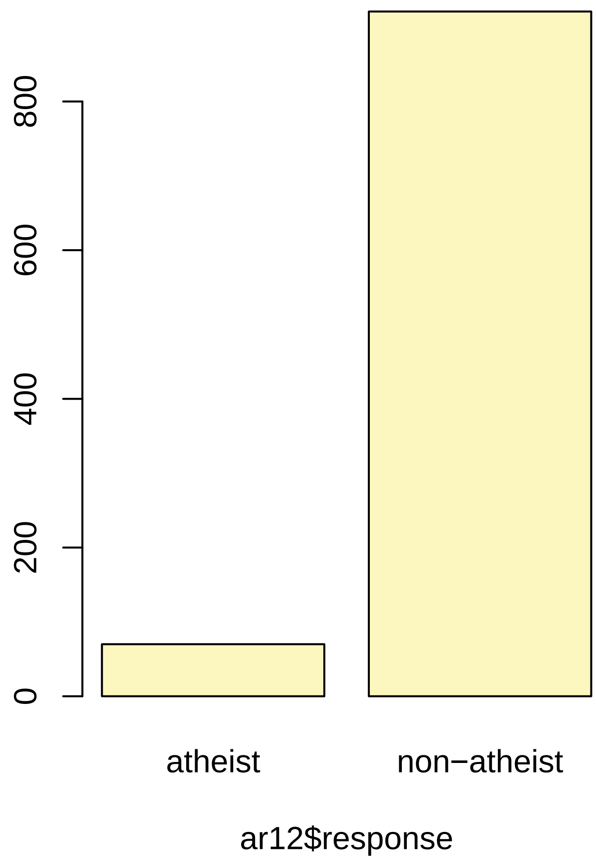
Country 2: Argentina

```r
ar12 <- subset(atheism, nationality == "Argentina" & year == "2012")
par12athe <- count(ar12$response == 'atheist')
names(par12athe) <- c("atheist", "total")
par12athe$percent <- par12athe$total / sum(par12athe$total) * 100
kable(par12athe)
```

atheist	total	percent
FALSE	921	92.936428
TRUE	70	7.063572

```r
inference(ar12$response, est = "proportion", type = "ci", method = "theoretical",
          success = "atheist")
```

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




```
## p_hat = 0.0706 ;  n = 991
## Check conditions: number of successes = 70 ; number of failures = 921
## Standard error = 0.0081
## 95 % Confidence interval = ( 0.0547 , 0.0866 )
```

Our **z** value for 95% confidence will be $z = 1.96$

From above we got **Standard error:** SE = 0.0081

```r
z <- 1.96
SE <- 0.0081
ME <- z * SE
ME
```

```
## [1] 0.015876
```

The margin of error for Argentina is 0.015876.

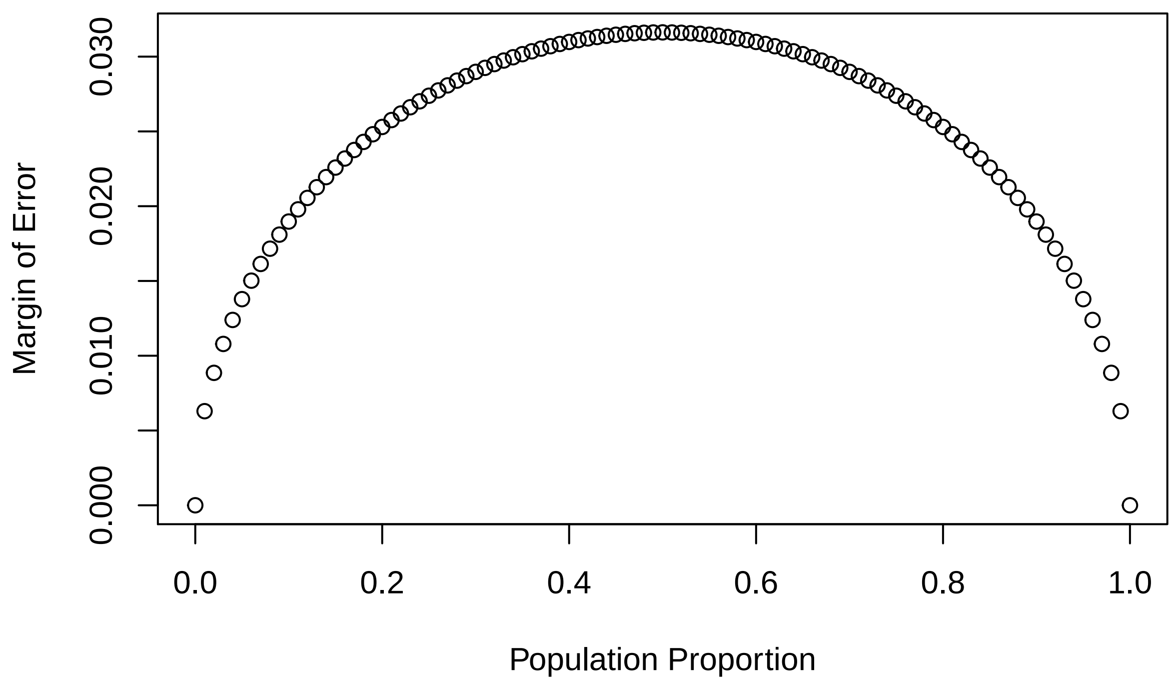
How does the proportion affect the margin of error?

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`.

Answer:

Based on the graph, the proportion of **0.50** is the proportion that will provide the largest margin of error possible.

That is, if we have **p = 0.5** and we do **p*(1 - p)** = 0.5 * (1 - 0.5) = 0.5 * 0.5 is the maximum value for the numerator, making it the biggest possible value for the **ME**.

Similar cases for **p = 0** and **p = 1** will be returning the minimums since **1 * 0 = 0**

In other words, the nearer p comes to 0.5 the bigger ME will be and the nearer p becomes to 0 or 1 the lower ME will be.

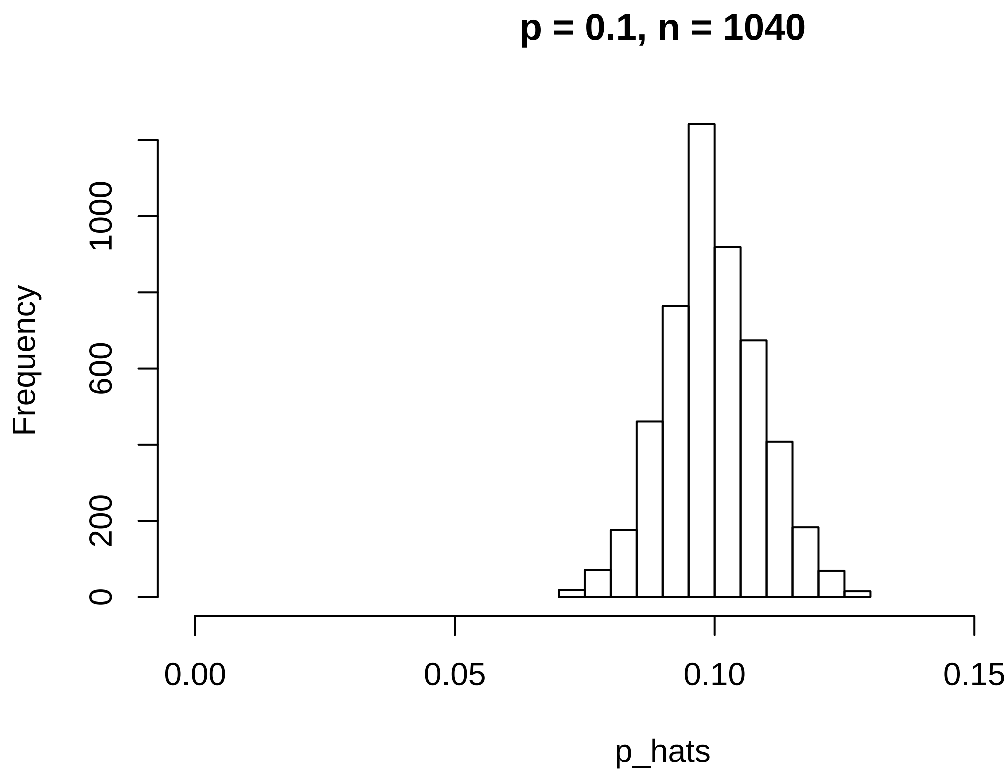
Success-failure condition

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.

Answer:

```
library(psych)
describe(p_hats)
```

```
##    vars    n mean   sd median trimmed  mad  min  max range skew kurtosis
## X1    1 5000  0.1 0.01    0.1     0.1 0.01 0.07 0.13  0.06 0.06    -0.09
##    se
## X1  0
```

```
summary(p_hats)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.07019 0.09327 0.09904 0.09969 0.10580 0.12980
```

The simulation to produce 5000 samples of size 1040 sample proportions, each follow normal distribution since they follow the conditions on Inference for proportions, some samples might have some outliers due to chance. The median and mean of the distribution are near identical at 0.1 with an standard deviation of 0.01. The sampling distribution of sample of proportions has a bell shape and follow a normal distribution.

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?

Answer:

a) $n = 400$ and $p = 0.1$

```
p <- 0.1
n <- 400
p_hats1 <- rep(0, 5000)

for(i in 1:5000){
  samp <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
  p_hats1[i] <- sum(samp == "atheist")/n
}
```

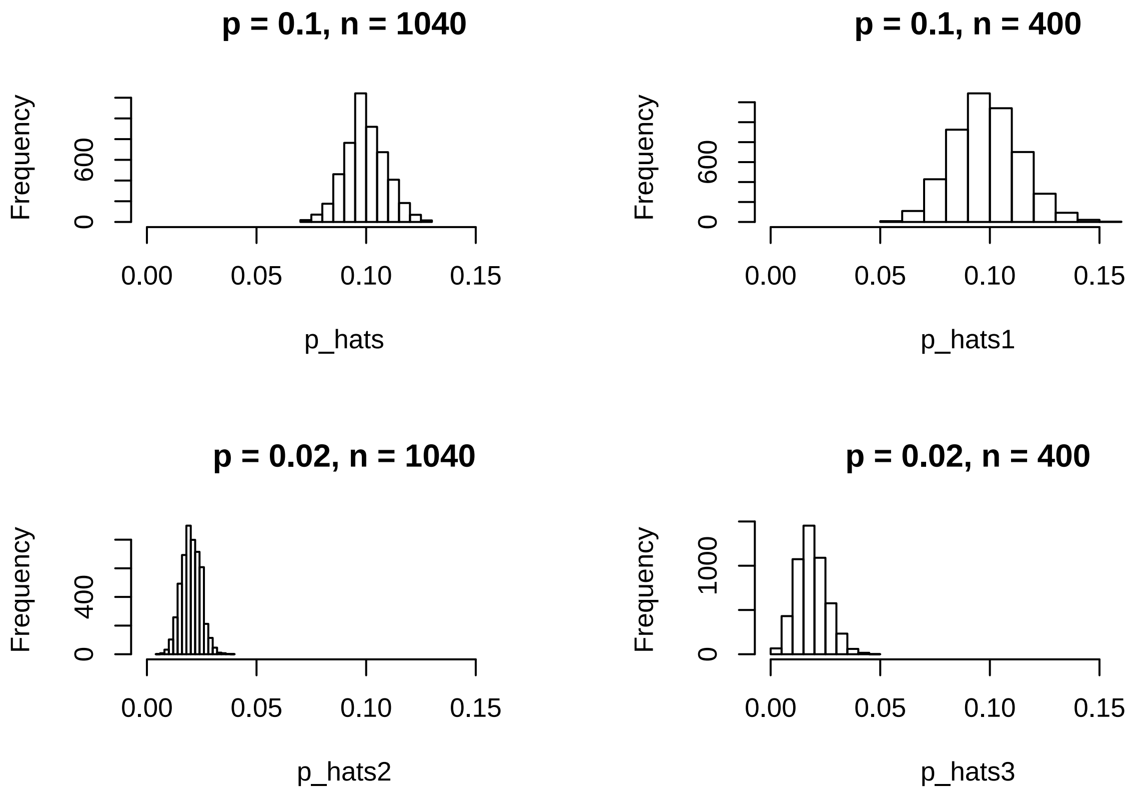
b) $n = 1040$ and $p = 0.02$

```
p <- 0.02
n <- 1040
p_hats2 <- rep(0, 5000)

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

c) $n = 400$ and $p = 0.02$

```
p <- 0.02
n <- 400
p_hats3 <- rep(0, 5000)

for(i in 1:5000){
  samp <- sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
  p_hats3[i] <- sum(samp == "atheist")/n
}
```

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

hist(p_hats, main = "p = 0.1, n = 1040", xlim = c(0, 0.18))
hist(p_hats1, main = "p = 0.1, n = 400", xlim = c(0, 0.18))
hist(p_hats2, main = "p = 0.02, n = 1040", xlim = c(0, 0.18))
hist(p_hats3, main = "p = 0.02, n = 400", xlim = c(0, 0.18))
```




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

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?

Answer:

By analyzing the conditions on inference proportions, we have a s follows:

```r
# Australia

n_au <- 1040
p_au <- 0.1

cond_au <- c(n_au * p_au >= 10, n_au * (1 - p_au) >= 10)

# Ecuador

n_ecu <- 400
p_ecu <- 0.02
```

```
cond_ecu <- c(n_ecu * p_ecu >= 10, n_ecu * (1 - p_ecu) >= 10)
```

Based on Australia's conditions are TRUE for **n * p** and TRUE **n * (1 - p)**.

Based on Ecuador's conditions are FALSE for **n * p** and TRUE **n * (1 - p)**.

Since one of the Ecuador's conditions are not met, we might be inclined to reject the results since one of the conditions is not met. However, in my opinion this could be counted as valid result since **n_ecu * p_ecu = 8** and the difference with the normal distribution should not be significant different.

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:

H0: There is no convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012; that is p2005 = p2012

HA: There is convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012; that is p2005 is not equal to p2012

2005

```
esp05 <- subset(atheism, nationality == "Spain" & year == "2005")
pesp05athe <- count(esp05$response == 'atheist')
names(pesp05athe) <- c("atheist", "total")
pesp05athe$percent <- pesp05athe$total / sum(pesp05athe$total) * 100
kable(pesp05athe)
```

atheist	total	percent
FALSE	1031	89.9651
TRUE	115	10.0349

2012

```
esp12 <- subset(atheism, nationality == "Spain" & year == "2012")
pesp12athe <- count(esp12$response == 'atheist')
names(pesp12athe) <- c("atheist", "total")
pesp12athe$percent <- pesp12athe$total / sum(pesp12athe$total) * 100
kable(pesp12athe)
```

atheist	total	percent
FALSE	1042	91.004367
TRUE	103	8.995633

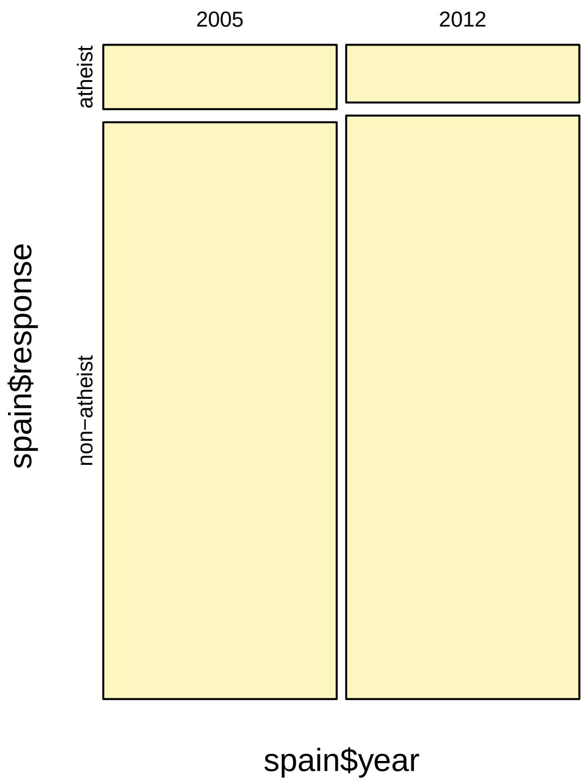
Inference

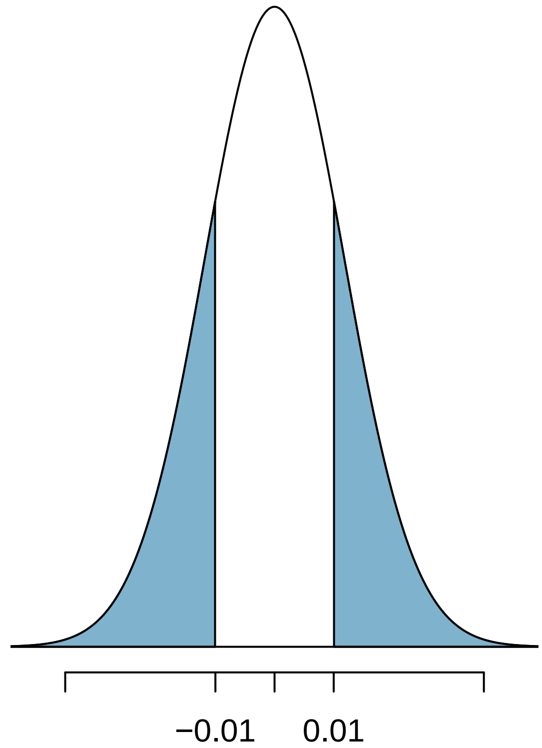
```
spain <- subset(atheism, nationality == "Spain" & year == "2005"  | nationality == "Spain" & year == "20

inference(y = spain$response, x = spain$year, est = "proportion",type = "ht", null = 0, alternative = "
```

```
## 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
```





Since our returned **p-value** is more that 0.05; we fail to reject our NULL hypothesis in favor of the alternative HA hypothesis; that is: There is No convincing evidence that Spain has seen a change in its atheism index 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:

H0: There is no convincing evidence that United States has seen a change in its atheism index between 2005 and 2012; that is p2005 = p2012

HA: There is convincing evidence that United States has seen a change in its atheism index between 2005 and 2012; that is p2005 is not equal to p2012

2005

```
us05 <- subset(atheism, nationality == "United States" & year == "2005")
pus05athe <- count(us05$response == 'atheist')
names(pus05athe) <- c("atheist", "total")
pus05athe$percent <- pus05athe$total / sum(pus05athe$total) * 100
kable(pus05athe)
```

atheist	total	percent
FALSE	992	99.001996
TRUE	10	0.998004

2012

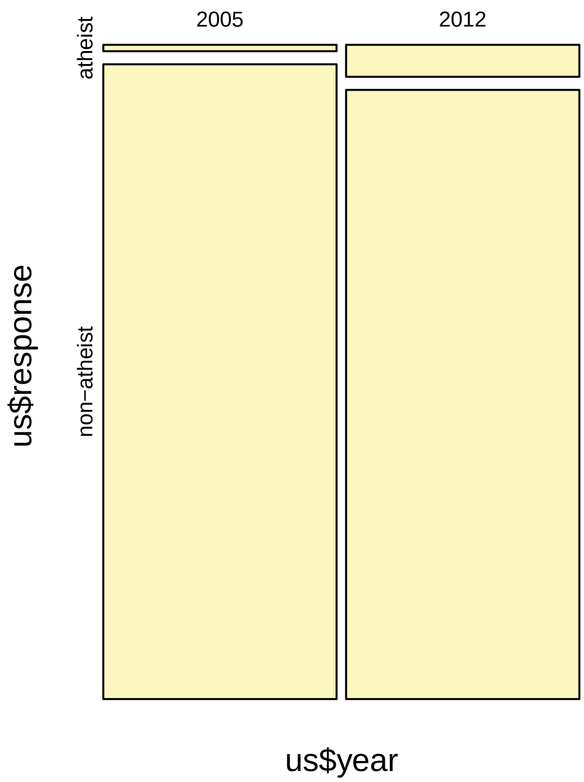
```
us12 <- subset(atheism, nationality == "United States" & year == "2012")
pus12athe <- count(us12$response == 'atheist')
names(pus12athe) <- c("atheist", "total")
pus12athe$percent <- pus12athe$total / sum(pus12athe$total) * 100
kable(pus12athe)
```

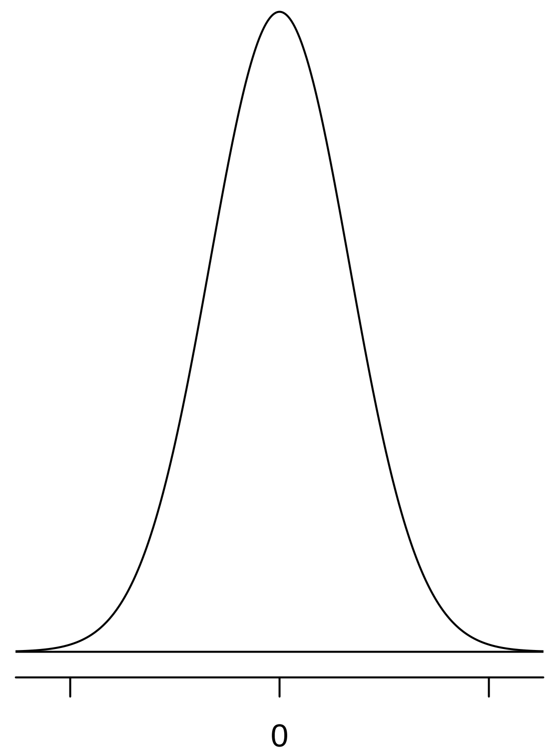
atheist	total	percent
FALSE	952	95.00998
TRUE	50	4.99002

Inference

```
us <- subset(atheism, nationality == "United States" & year == "2005"  | nationality == "United States"

inference(y = us$response, x = us$year, est = "proportion",type = "ht", null = 0, alternative = "twosid
```

```
## 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
```





Since our returned **p-value** equals 0; we reject our NULL hypothesis in favor of the alternative HA hypothesis; that is: There is convincing evidence that United States has seen a change in its atheism index 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.

Amswer:

A type 1 error is rejecting the null hypothesis when H0 is actually true.

Typically we do not want to incorrectly reject H0 more than 5% of the time; this resumes to a 0.05 significance level.

Since there are 39 countries listed in Table 4; all we need to do is to multiply 0.05 by 39 to estimate how many countries we would expect to detect a change in the atheism index simply by chance.

The result is 1.95, or about 2 countries would be expected to detect a change in atheism just 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:

There are two unknown variables in this question: **p** and **n**.

When we do not have and estimate for p we follow the guideline that the margin of error is largest when p is 0.5. As reference we typically use this as the worst case estimate if no other estimate is available. The estimate must have a margin of error no greater than 1%, this means that if use the formula $ME = z \cot SE$ with a 95% confidence interval we obtain $z = 1.96$; hence $ME = 1,96 * sqrt(p(1-p)/n))$; since our $ME \leq 0.01$.

$$SE = \sqrt{\frac{p \cdot (1-p)}{n}}$$

$$ME = z \cdot SE$$

$$ME = z \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$$

$$\frac{ME}{z} = \sqrt{\frac{p \cdot (1-p)}{n}}$$

$$\left(\frac{ME}{z}\right)^2 = \frac{p \cdot (1-p)}{n}$$

$$n = \frac{p \cdot (1-p)}{\left(\frac{ME}{z}\right)^2}$$

$$n = \frac{p \cdot (1-p) \cdot z^2}{ME^2}$$

```
p <-0.5 # Will generate the maximum ME
z <-1.96 # 95% Confidence interval
ME <- 0.01 # Marging of error no greater than 0.01
n <- p * (1 - p) * z ^ 2  / ME ^ 2
n
```

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

And based on the results we will need at least 9604 participants to ensure the sample proportion is within 0.01 of the true proportion with 95% confidence.