week4

kangzq

2021/8/26

Inference for categorical data

Inference for proportions

Sampling Variability and CLT for proportions

Define population proportion p (parameter) and sample proportion p_hat.

In the case of the proportion the CLT tells us that if:

- The observations in the sample are independent
- The sample size is sufficiently large (checked using the success/failure condition: $np \geq 10$ and $n(1-p) \geq 10$ – at least 10 successes and 10 failures.

Then the distribution of the sample proportion will be nearly normal and suitable for CLT.

If the CLT doesn't apply and the sample proportion is low (close to 0) the sampling distribution will likely be right skewed, if the sample proportion is high (close to 1) the sampling distribution will likely be left skewed. ### Confidence Interval for a Proportion

Notice that H0 should be decided by p0 and SE also need to be calculated, since the mean doesn't factor into the calculation of the standard error, while the proportion does.

When there is no additional information, we use $\hat{p}=0.5$ to estimate for the required sample size. - highest possible sample size.

Hypothesis Test for a proportion

H0: p = null value HA: p not null value

Estimating the difference Between Two Proportions & Hypothesis test for comparing 2 proportions

- confidence interval and hypothesis test when H0: p1 - p2 = some value other than 0
- hypothesis test when H0: p1 - p2 = 0, use p_pool (the overall rate of success = number of successes in group 1 + number of successes in group 2 (n1+n2)).

Simulation based inference for proportions and Chi-square testing

Small sample proportions

Simulation when CLT is not useful.

chi-square GOF (godness of fit) test

Use a chi-square test of goodness of fit to evaluate if the distribution of levels of a single categorical variable follows a hypothesized distribution. (How far the observed counts are from the expected counts.)

- H0: The distribution of observed counts follows the hypothesized distribution, and any observed differences are due to chance.
- HA: The distribution of observed counts does not follow the hypothesized distribution.

Calculate the expected counts for a given level (cell) in a one-way table as the sample size times the hypothesized proportion for that level.

Note that the chi-square statistic is always positive, and follows a right skewed distribution with one parameter: degrees of freedom (k-1).

Conditions:

1. The observations should be independent
2. Expected counts for each cell should be at least 5
3. Degrees of freedom should be at least 2 (if not, use methods for evaluating proportions)

`pchisq(chi, df, lower.tail = )` is useful for calculating p-value.

The Chi-square Independence Test

When evaluating the independence of two categorical variables where at least one has more than two levels, use a chi-square test of independence.

- H0: The two variables are independent.
- HA: The two variables are dependent.

Calculate the degrees of freedom for chi-square test of independence as df = (R−1) × (C−1), where R is the number of rows in a two-way table, and C is the number of columns.

Note that there is no such thing as a chi-square confidence interval for proportions, since in the case of a categorical variables with many levels, there isn't one parameter to estimate.

Use simulation methods when sample size conditions aren't met for inference for categorical variables.

- The t-distribution is only appropriate to use for means. When sample size isn't sufficiently large, and the parameter of interest is a proportion or a difference between two proportions, we need to use simulation.

1. for one categorical variable, generate simulated samples based on the null hypothesis, and then calculate the number of samples that are at least as extreme as the observed data.
2. for two categorical variables, use a randomization test.

Week4 Lab

In August of 2012, news outlets ranging from the Washington Post (http://www.washingtonpost.com/national/on-faith/poll-shows-atheism-on-the-rise-in-the-us/2012/08/13/90020fd6-e57d-11e1-9739-eef99c5fb285_story.html) to the Huffington Post (http://www.huffingtonpost.com/2012/08/14/atheism-rise-religiosity-decline-in-america_n_1777031.html) 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.

Getting Started

Load packages

In this lab we will explore the data using the `dplyr` package and visualize it using the `ggplot2` package for data visualization. The data can be found in the companion package for this course, `statsr`.

Let's load the packages.

```
library(statsr)
library(dplyr)
library(ggplot2)
library(varhandle)
library(janitor)
```

The survey

The press release for the poll, conducted by WIN-Gallup International, can be accessed here (https://www.scribd.com/document/136318147/Win-gallup-International-Global-Index-of-Religiosity-and-Atheism-2012).

Take a moment to review the report then address the following questions.

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.

```
data(atheism)
```

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.

Create a new dataframe called `us12` that contains only the rows in `atheism` associated with respondents to the 2012 survey from the United States:

```
us12 <- atheism %>%
  filter(nationality == "United States" , atheism$year == "2012")
```

```
# type your code for Question 7 here, and Knit
us12$nationality <- unfactor(us12$nationality)
us12$response <- unfactor(us12$response)
#Create a frequency distribution table (a table of counts for the categorical varia
         ble abhlth by year)
#Using the library janitor and the tabyl function
tabyl(us12, response) %>%
  adorn_totals(c('row', 'col'))
```

```
##     response    n   percent     Total
##      atheist   50 0.0499002   50.0499
##  non-atheist  952 0.9500998  952.9501
##        Total 1002 1.0000000 1003.0000
```

Inference on proportions

As was hinted earlier, Table 6 provides **sample statistics**, that is, calculations made from the sample of 51,927 people. What we'd like, though, is insight into the population **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.

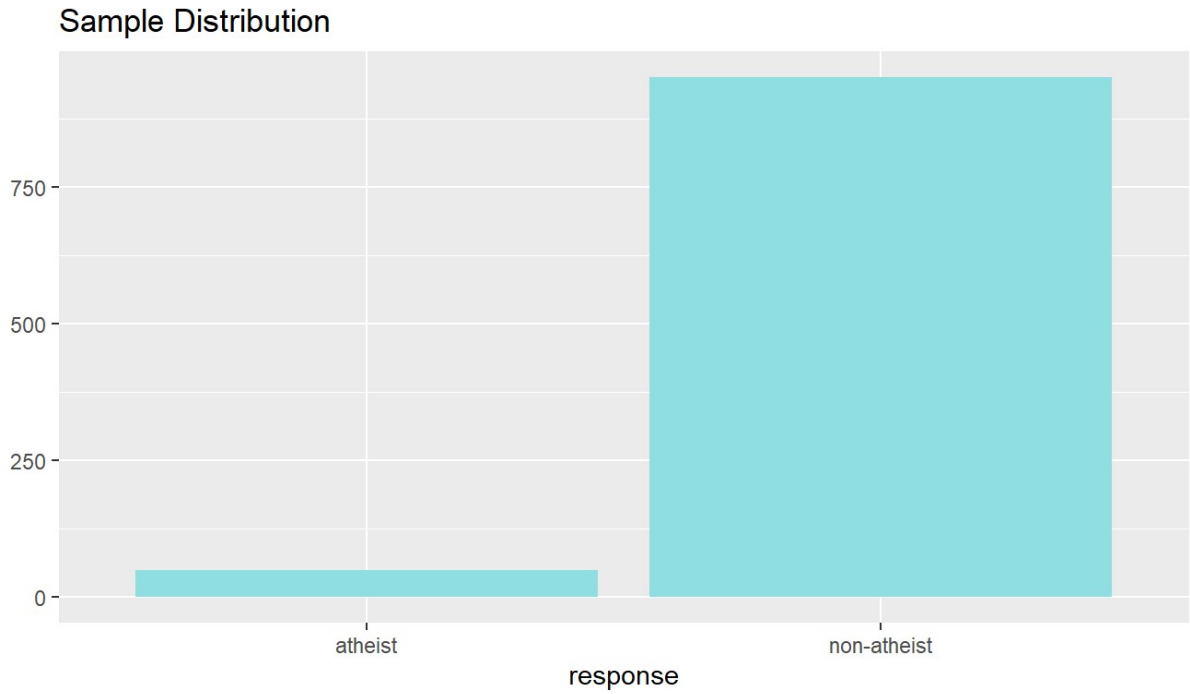
The inferential tools for estimating population proportion are analogous to those used for means in the last lab: the confidence interval and the hypothesis test.

Exercise: 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?

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(y = response, data = us12, statistic = "proportion", type = "ci", method
        = "theoretical", success = "atheist")
```

```
## Single categorical variable, success: atheist
## n = 1002, p-hat = 0.0499
## 95% CI: (0.0364 , 0.0634)
```



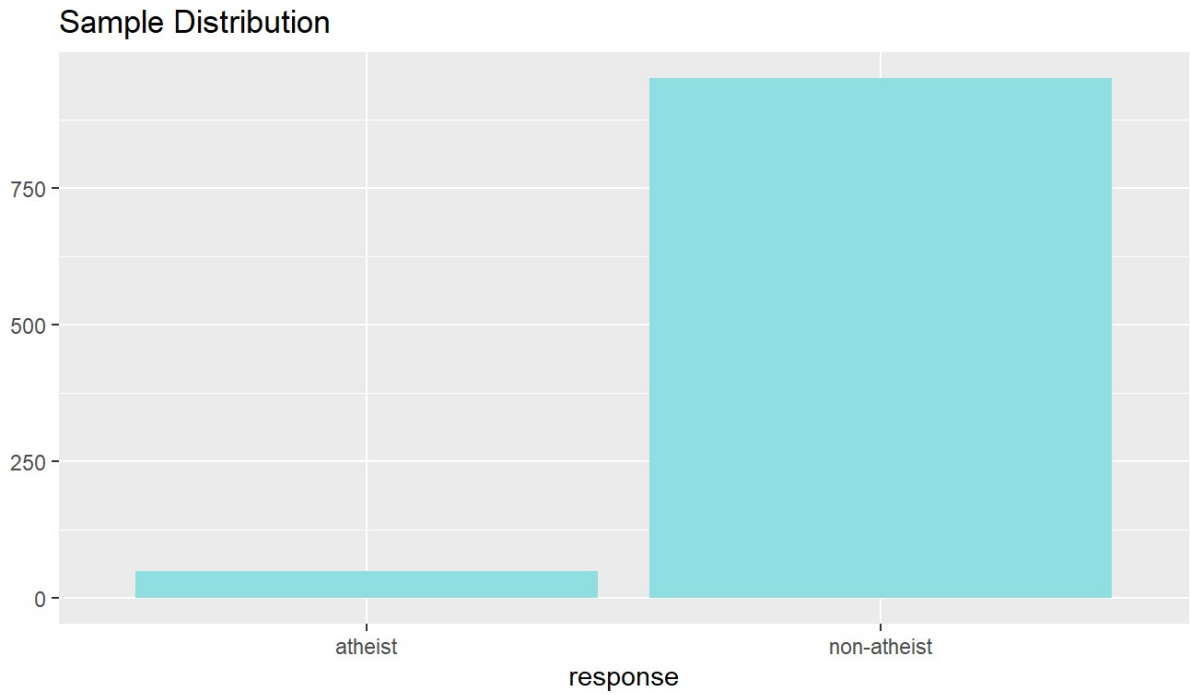
Note that since the goal is to construct an interval estimate for a proportion, it's necessary to specify what constitutes a `success'', which here is a response of atheist`.

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."

Exercise: Imagine that, after reading a front page story about the latest public opinion poll, a family member asks you, "What is a margin of error?" In one sentence, and ignoring the mechanics behind the calculation, how would you respond in a way that conveys the general concept?

```
# type your code for Question 8 here, and Knit
inference(y = response, data = us12, statistic = "proportion", type = "ci", method
        = "theoretical", success = "atheist")
```

```
## Single categorical variable, success: atheist
## n = 1002, p-hat = 0.0499
## 95% CI: (0.0364 , 0.0634)
```

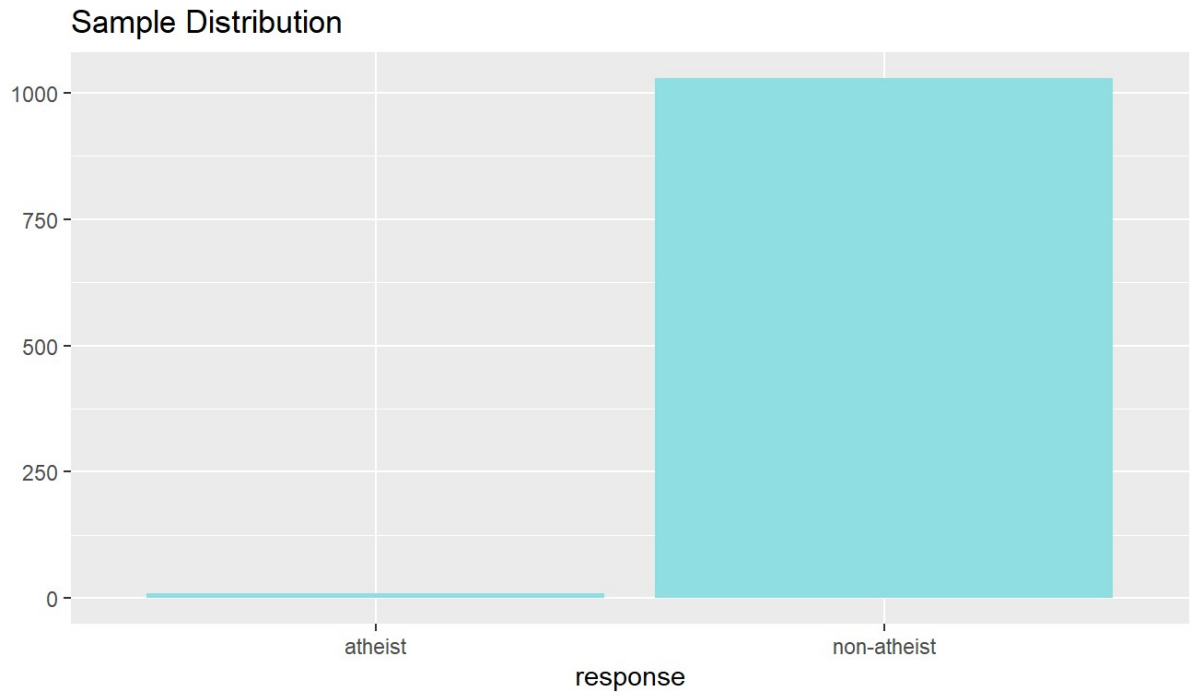


Exercise: Using the inference function, calculate confidence intervals for the proportion of atheists in 2012 in two other countries of your choice, and report the associated margins of error. Be sure to note whether the conditions for inference are met. It may be helpful to create new data sets for each of the two countries first, and then use these data sets in the `inference` function to construct the confidence intervals.

```
# type your code for the Exercise here, and Knit
#Subset the data to include India and China atheist data for 2012:
Romania12 <- atheism %>%
  filter(nationality == "Romania" , atheism$year == "2012")

China12 <- atheism %>%
  filter(nationality == "China" , atheism$year == "2012")

#As usual, unfactor all the variables that are factors to avoid any errors in the i
        nference function:
Romania12$nationality <- unfactor(Romania12$nationality)
Romania12$response <- unfactor(Romania12$response)

China12$nationality <- unfactor(China12$nationality)
China12$response <- unfactor(China12$response)
```

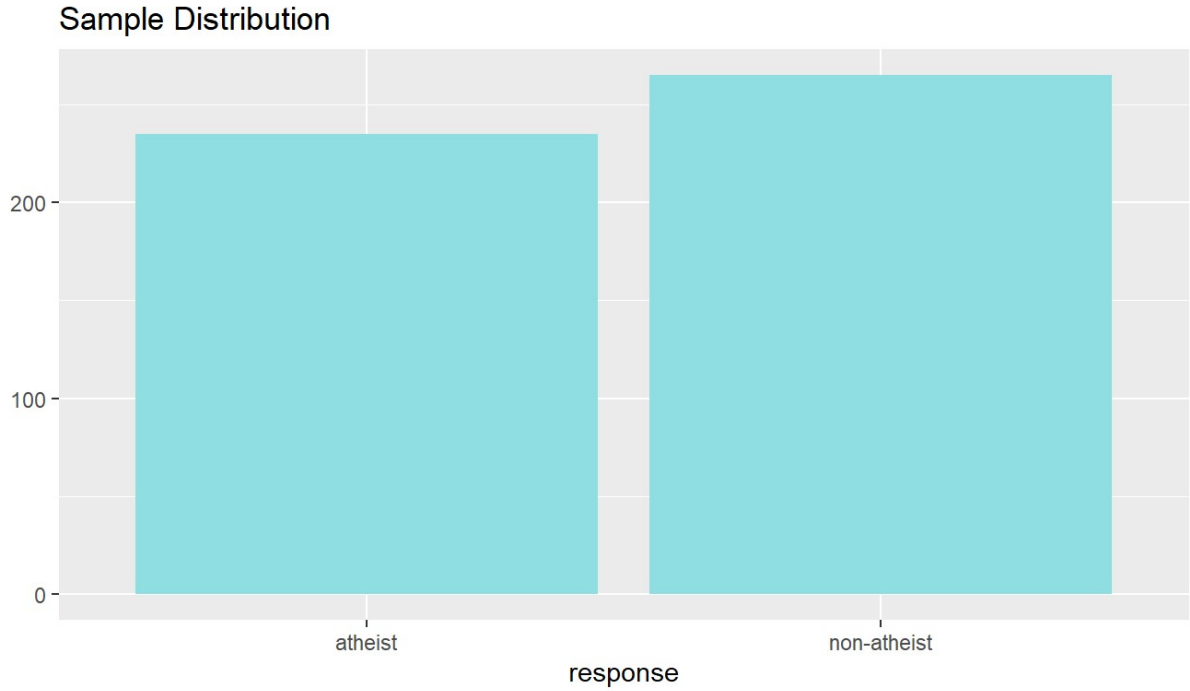
```
#Use the inference function for both countries
inference(y = response, data = Romania12, statistic = "proportion", type = "ci", me
        thod = "theoretical", success = "atheist")
```

```
## Single categorical variable, success: atheist
## n = 1039, p-hat = 0.0096
## 95% CI: (0.0037 , 0.0156)
```




```
inference(y = response, data = China12, statistic = "proportion", type = "ci", meth
         od = "theoretical", success = "atheist")
```

```
## Single categorical variable, success: atheist
## n = 500, p-hat = 0.47
## 95% CI: (0.4263 , 0.5137)
```



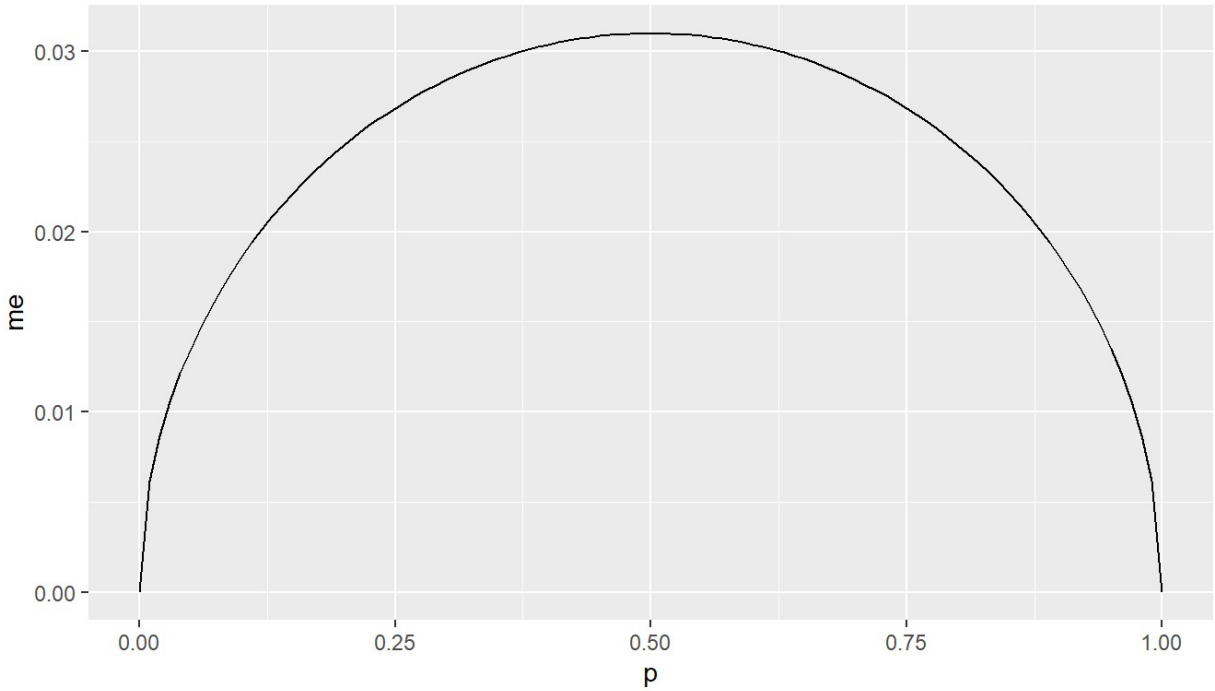
How does the proportion affect the margin of error?

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 ($ME = 1.96 \times SE$). Lastly, we plot the two vectors against each other to reveal their relationship.

```
d <- data.frame(p <- seq(0, 1, 0.01))
n <- 1000
d <- d %>%
  mutate(me = 1.96*sqrt(p*(1 - p)/n))
ggplot(d, aes(x = p, y = me)) +
  geom_line()
```



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

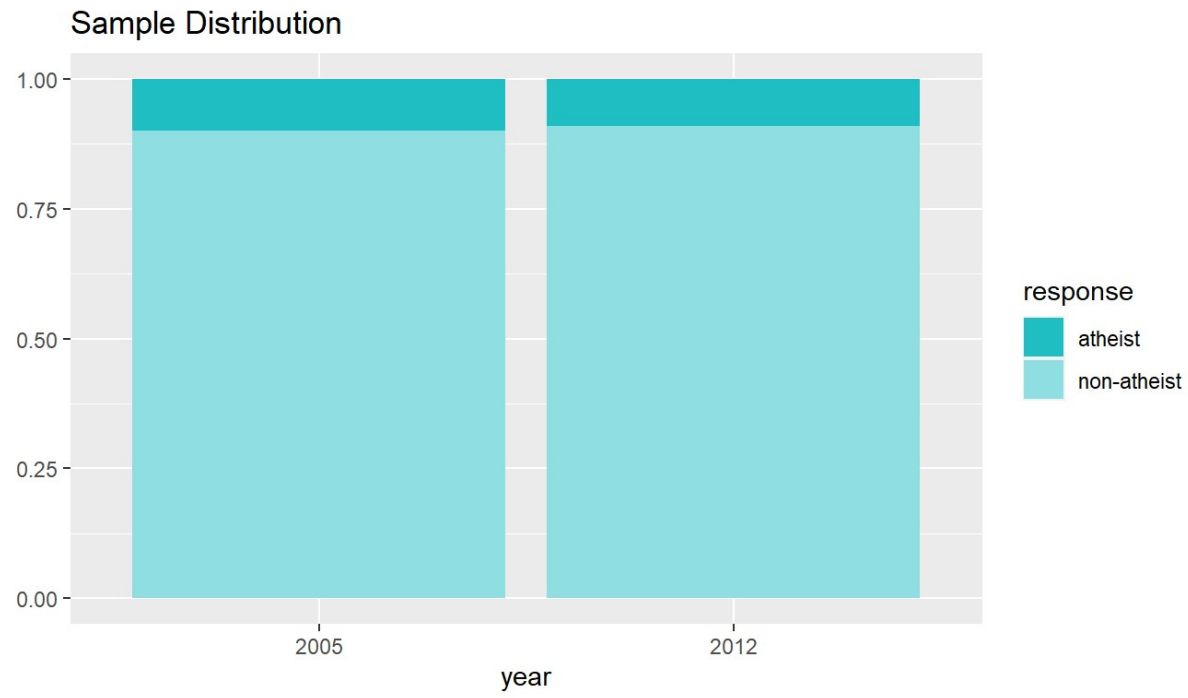
hypotheses for any tests you conduct and outline the status of the conditions for inference.

10. True / False: There is 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. Then use their responses as the first input on the `inference`, and use `year` as the grouping variable.
    1. True
    2. False

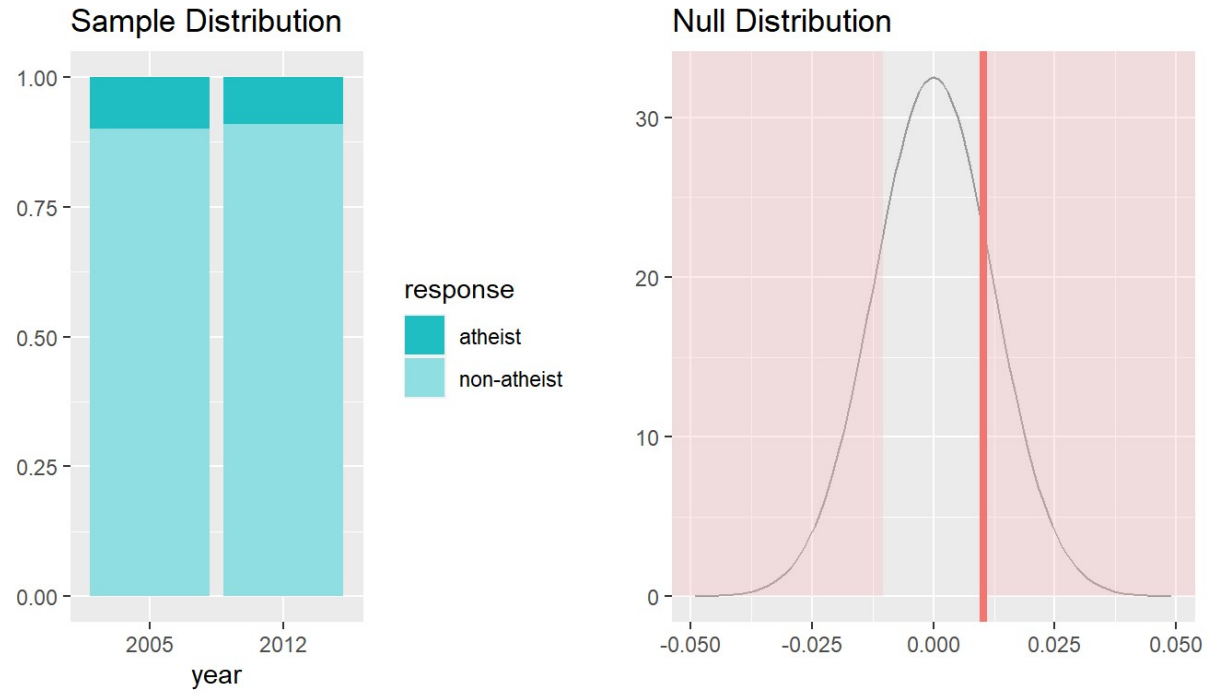
```
# type your code for Question 10 here, and Knit
Spain.atheism <- filter(atheism, atheism$nationality=="Spain")

Spain.atheism$year <- as.character(Spain.atheism$year)
Spain.atheism$response <- unfactor(Spain.atheism$response)

inference(y = response, x= year, data = Spain.atheism, statistic = "proportion", ty
         pe = "ci", method = "theoretical", success = "atheist")
```

```
## Response variable: categorical (2 levels, success: atheist)
## Explanatory variable: categorical (2 levels)
## n_2005 = 1146, p_hat_2005 = 0.1003
## n_2012 = 1145, p_hat_2012 = 0.09
## 95% CI (2005 - 2012): (-0.0136 , 0.0344)
```



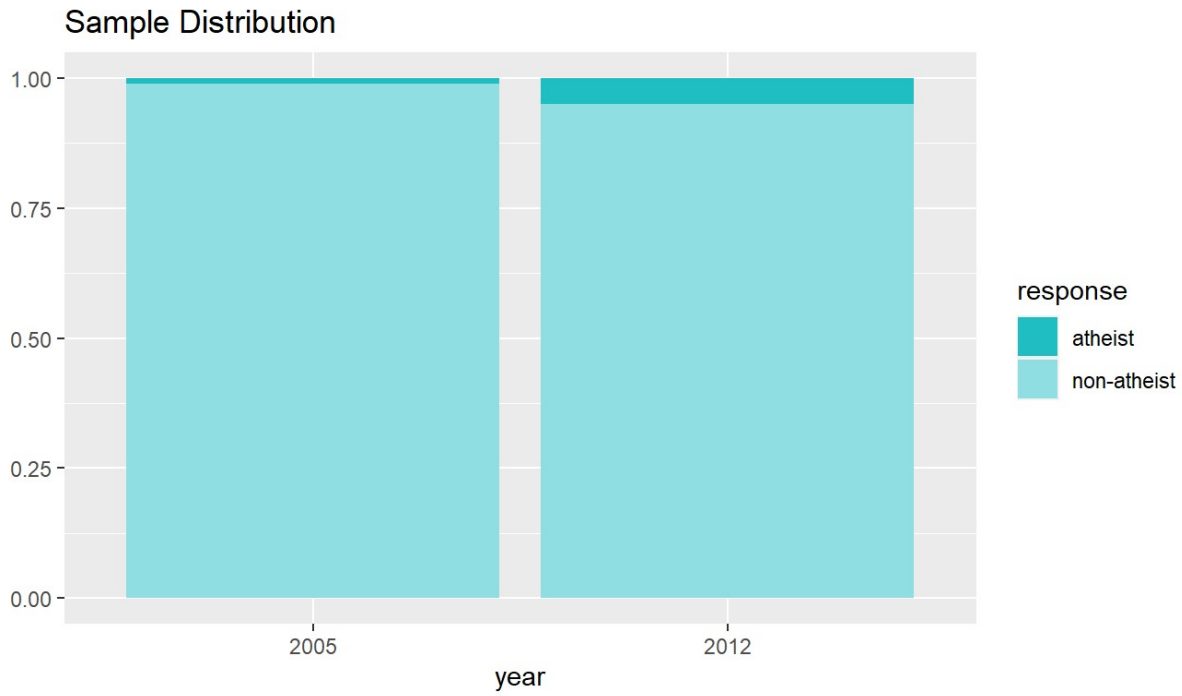

```
inference(y = response, x= year, data = Spain.atheism, statistic = "proportion", ty
         pe = "ht", null=0, alternative="twosided", method = "theoretical", success
         = "atheist")
```

```
## Response variable: categorical (2 levels, success: atheist)
## Explanatory variable: categorical (2 levels)
## n_2005 = 1146, p_hat_2005 = 0.1003
## n_2012 = 1145, p_hat_2012 = 0.09
## H0: p_2005 =  p_2012
## HA: p_2005 != p_2012
## z = 0.8476
## p_value = 0.3966
```




```
# type your code for Question 11 here, and Knit
US.atheism <- filter(atheism, atheism$nationality=="United States")

US.atheism$year <- as.character(US.atheism$year)
US.atheism$response <- unfactor(US.atheism$response)

inference(y = response, x= year, data = US.atheism, statistic = "proportion", type
         = "ci", method = "theoretical", success = "atheist")
```

```
## Response variable: categorical (2 levels, success: atheist)
## Explanatory variable: categorical (2 levels)
## n_2005 = 1002, p_hat_2005 = 0.01
## n_2012 = 1002, p_hat_2012 = 0.0499
## 95% CI (2005 - 2012): (-0.0547 , -0.0251)
```



12. 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: Type 1 error.

    1. 0
    2. 1
    3. 1.95
    4. 5

```
# type your code for Question 12 here, and Knit
```

13. 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.

    1. 2401 people
    2. At least 2401 people
    3. 9604 people
    4. At least 9604 people

```
# type your code for Question 13 here, and Knit
```