Lab6 - Inference for Categorical Data

Michael Y.

March 31st, 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*?

These are *sample statistics* as WIN-Gallup International pollsters indicate that they interviewed "more than 50,000 men and women [actually, 51,927] selected from 57 countries across the globe in 5 continents."

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?

To generalize, we must assume that the sampling method is random and proportionately covers the entire population (in this case, the entire globe.) This assumption is not reasonable because the the sampling covers only 57 countries, while there are approximately 200 countries in the world. Furthermore, the "Methods" section of the press release (22-24) indicates that the polling method differed by country, i.e., polling was conducted face-to-face in 35 countries, by telephone in 11 countries, and online in 11 countries. For telephone and online polling, one can only reach individuals who possess access to either a telephone or the internet, respectively. This could produce a non-representative sample in such countries, as many individuals are excluded from the sample. There is no discussion in the methodology section of the document as to what steps, if any, were taken by the individual pollsters (a separate individual or organization in each respective country) to ensure that the sample in their country was representative.

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")
describe(atheism)
```

```
##              vars     n    mean    sd median trimmed   mad  min  max range
## nationality*    1 88032   29.09 15.17     30   29.11 17.79    1   57    56
## response*       2 88032    1.94  0.24      2    2.00  0.00    1    2     1
## year            3 88032 2009.13  3.44   2012 2009.29  0.00 2005 2012     7
##                skew kurtosis   se
## nationality* -0.05    -1.00 0.05
## response*    -3.62    11.08 0.00
## year         -0.37    -1.87 0.01
```

```
byyear    <-table(atheism$year)
byyear
```

```
##
##  2005  2012
## 36105 51927
```

```
count2005 <- as.integer(byyear[1])
count2012 <- as.integer(byyear[2])
```

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

Each row of Table 6 corresponds to a country.

Each row of `atheism` correponds to an individual person. The data set contains 36105 responses from 2005 and 51927 responses from 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?

```
us12 <- subset(atheism, nationality == "United States" & year == "2012")
proportion <- length(us12$response[us12$response=="atheist"])  / length(us12$response)
proportion
```

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

Yes, up to rounding, both results are 5 percent.

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?

From the textbook, Page 275:

Conditions for the sampling distribution of $\hat{p}$ being nearly normal:

The sampling distribution for $\hat{p}$, taken from a sample of size n from a population with a true proportion p, is nearly normal when:

- **the sample observations are independent, and**

- **we expected to see at least 10 successes and 10 failures in our sample, i.e.** $np \geq 10$ **and** $n(1-p) \geq 10$ **.**

This is called the *success-failure condition.*

If these conditions are met, then the sampling distribution of $\hat{p}$ is nearly normal with mean p and standard error $SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

For confidence intervals, usually the sample proportion $\hat{p}$ is used to check the success-failure condition and compute the standard error.

Reminder on checking independence of observations:

If data come from a simple random sample and consist of less than 10% of the population, then the independence assumption is reasonable.

From the textbook, Page 276:

Observations are independent:

The poll is based on a simple random sample and consists of fewer than 10% of the population, which verifies independence.

Success-failure condition:

The sample size must also be sufficiently large, which is checked using the success-failure condition.

There were *50 "successes"* (atheist) and *952 "failures"* (non-atheist) in the sample of 1002 US residents. Both figures are greater than 10.

Constructing a confidence interval for a proportion:

• Verify the observations are independent and also verify the success-failure condition using $\hat{p}$ and n.

• If the conditions are met, the sampling distribution of $\hat{p}$ may be well approximated by the normal model.

• Construct the standard error using $\hat{p}$ in place of p and apply the general confidence interval formula.

```
### count the responses
USAtheist2012 <- sum(us12$response=='atheist')
USNonAtheist2012 <- sum(us12$response=='non-atheist')
UStotal2012 <- length(us12$response)

### Make sure we didn't miss anything
USAtheist2012+USNonAtheist2012==UStotal2012
```

```
## [1] TRUE
```

```
### compute the sample estimate of the proportion
pct_atheist_US <- USAtheist2012 / UStotal2012
pct_atheist_US
```

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

```
### compute the stderr
stderr_US  <- sqrt(USAtheist2012*USNonAtheist2012/UStotal2012**3)
stderr_US
```

```
## [1] 0.0068786291
```

```
### compute the margin of error
marginerror_US <- 1.96 * stderr_US
marginerror_US
```

```
## [1] 0.013482113
```

```
### compute the CI
lower_ci_US <- pct_atheist_US - marginerror_US
upper_ci_US <- pct_atheist_US + marginerror_US
ci_US       <- c(lower_ci_US, upper_ci_US)
```

The point estimate for percentage of atheists in the US in 2012 is 0.0499002 .

The manually computed standard error is 0.00687863 .

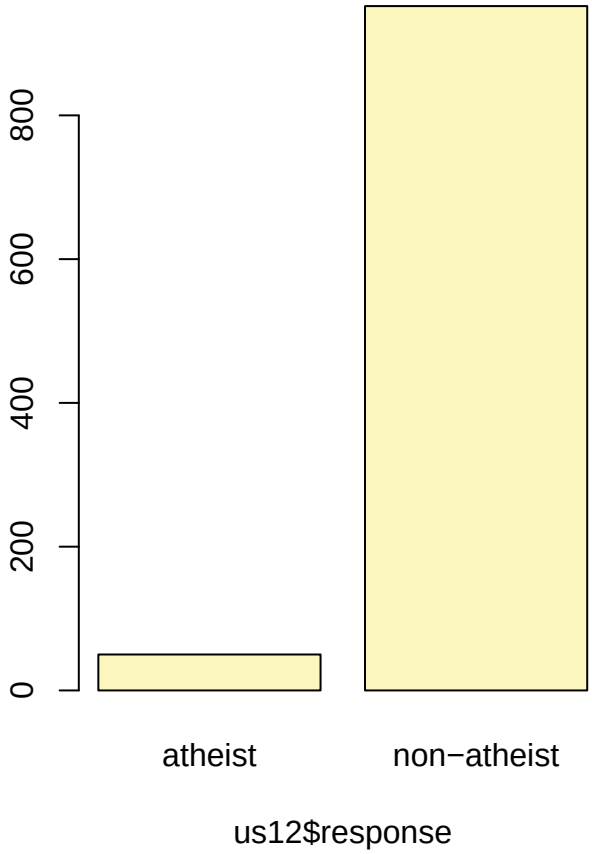
The margin of error is 0.01348211 .

The confidence interval is (0.03641809 , 0.06338231) .

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?

The margin of error is 0.01348211 .

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.

```
### get just the data for 2012
ath12 <- subset(atheism, year == "2012")

### find the countries
countries <- levels(ath12$nationality)
t(t(countries))
```

```
##       [,1]
##  [1,] "Afghanistan"
##  [2,] "Argentina"
##  [3,] "Armenia"
##  [4,] "Australia"
##  [5,] "Austria"
##  [6,] "Azerbaijan"
##  [7,] "Belgium"
##  [8,] "Bosnia and Herzegovina"
##  [9,] "Brazil"
## [10,] "Bulgaria"
## [11,] "Cameroon"
## [12,] "Canada"
## [13,] "China"
## [14,] "Colombia"
## [15,] "Czech Republic"
## [16,] "Ecuador"
## [17,] "Fiji"
## [18,] "Finland"
## [19,] "France"
## [20,] "Georgia"
## [21,] "Germany"
## [22,] "Ghana"
## [23,] "Hong Kong"
## [24,] "Iceland"
## [25,] "India"
## [26,] "Iraq"
## [27,] "Ireland"
## [28,] "Italy"
## [29,] "Japan"
## [30,] "Kenya"
## [31,] "Korea, Rep (South)"
## [32,] "Lebanon"
## [33,] "Lithuania"
## [34,] "Macedonia"
## [35,] "Malaysia"
## [36,] "Moldova"
## [37,] "Netherlands"
## [38,] "Nigeria"
## [39,] "Pakistan"
## [40,] "Palestinian territories (West Bank and Gaza)"
## [41,] "Peru"
## [42,] "Poland"
```

```
## [43,] "Romania"
## [44,] "Russian Federation"
## [45,] "Saudi Arabia"
## [46,] "Serbia"
## [47,] "South Africa"
## [48,] "South Sudan"
## [49,] "Spain"
## [50,] "Sweden"
## [51,] "Switzerland"
## [52,] "Tunisia"
## [53,] "Turkey"
## [54,] "Ukraine"
## [55,] "United States"
## [56,] "Uzbekistan"
## [57,] "Vietnam"
```

```
### see the totals
totals12 <- table(ath12$nationality)
t(t(totals12))
```

```
##
##                                      [,1]
##   Afghanistan                        1031
##   Argentina                           991
##   Armenia                             495
##   Australia                          1039
##   Austria                            1002
##   Azerbaijan                          509
##   Belgium                             527
##   Bosnia and Herzegovina             1000
##   Brazil                             2002
##   Bulgaria                           1006
##   Cameroon                            504
##   Canada                             1002
##   China                               500
##   Colombia                            606
##   Czech Republic                     1000
##   Ecuador                             404
##   Fiji                               1018
##   Finland                             985
##   France                             1688
##   Georgia                            1000
##   Germany                             502
##   Ghana                              1490
##   Hong Kong                           500
##   Iceland                             852
##   India                              1092
##   Iraq                               1000
##   Ireland                            1010
##   Italy                               987
##   Japan                              1212
##   Kenya                              1000
##   Korea, Rep (South)                 1523
##   Lebanon                             505
##   Lithuania                          1015
```

```
##   Macedonia                                        1209
##   Malaysia                                          520
##   Moldova                                          1085
##   Netherlands                                       509
##   Nigeria                                          1049
##   Pakistan                                         2704
##   Palestinian territories (West Bank and Gaza)      627
##   Peru                                             1207
##   Poland                                            525
##   Romania                                          1039
##   Russian Federation                               1000
##   Saudi Arabia                                      500
##   Serbia                                           1036
##   South Africa                                      202
##   South Sudan                                      1020
##   Spain                                            1145
##   Sweden                                            495
##   Switzerland                                       513
##   Tunisia                                           498
##   Turkey                                           1032
##   Ukraine                                          1013
##   United States                                    1002
##   Uzbekistan                                        500
##   Vietnam                                           500
```

```
### how many atheists?
atheists12 <- table(ath12$nationality[ath12$response=="atheist"])
t(t(atheists12))
```

```
##
##                                                  [,1]
##   Afghanistan                                      0
##   Argentina                                       70
##   Armenia                                         10
##   Australia                                      104
##   Austria                                        100
##   Azerbaijan                                       0
##   Belgium                                         42
##   Bosnia and Herzegovina                          40
##   Brazil                                          20
##   Bulgaria                                        19
##   Cameroon                                        15
##   Canada                                          90
##   China                                          235
##   Colombia                                        18
##   Czech Republic                                 300
##   Ecuador                                          8
##   Fiji                                            10
##   Finland                                         59
##   France                                         485
##   Georgia                                         10
##   Germany                                         75
##   Ghana                                            0
##   Hong Kong                                       45
##   Iceland                                         85
```

```
##   India                                          33
##   Iraq                                            0
##   Ireland                                       100
##   Italy                                          79
##   Japan                                         372
##   Kenya                                          20
##   Korea, Rep (South)                            229
##   Lebanon                                        10
##   Lithuania                                      10
##   Macedonia                                      12
##   Malaysia                                        0
##   Moldova                                        54
##   Netherlands                                    71
##   Nigeria                                        10
##   Pakistan                                       54
##   Palestinian territories (West Bank and Gaza)   25
##   Peru                                           36
##   Poland                                         26
##   Romania                                        10
##   Russian Federation                             60
##   Saudi Arabia                                   25
##   Serbia                                         31
##   South Africa                                    8
##   South Sudan                                    61
##   Spain                                         103
##   Sweden                                         40
##   Switzerland                                    46
##   Tunisia                                         0
##   Turkey                                         21
##   Ukraine                                        30
##   United States                                  50
##   Uzbekistan                                     10
##   Vietnam                                         0
```

```
### get the proportions
country_proportion <- atheists12 / totals12
t(t(country_proportion))
```

```
##
##                                                [,1]
##   Afghanistan                                  0.0000000000
##   Argentina                                    0.0706357215
##   Armenia                                      0.0202020202
##   Australia                                    0.1000962464
##   Austria                                      0.0998003992
##   Azerbaijan                                   0.0000000000
##   Belgium                                      0.0796963947
##   Bosnia and Herzegovina                       0.0400000000
##   Brazil                                       0.0099900100
##   Bulgaria                                     0.0188866799
##   Cameroon                                     0.0297619048
##   Canada                                       0.0898203593
##   China                                        0.4700000000
##   Colombia                                     0.0297029703
##   Czech Republic                               0.3000000000
```

```
##   Ecuador                                          0.0198019802
##   Fiji                                             0.0098231827
##   Finland                                          0.0598984772
##   France                                           0.2873222749
##   Georgia                                          0.0100000000
##   Germany                                          0.1494023904
##   Ghana                                            0.0000000000
##   Hong Kong                                        0.0900000000
##   Iceland                                          0.0997652582
##   India                                            0.0302197802
##   Iraq                                             0.0000000000
##   Ireland                                          0.0990099010
##   Italy                                            0.0800405268
##   Japan                                            0.3069306931
##   Kenya                                            0.0200000000
##   Korea, Rep (South)                               0.1503611293
##   Lebanon                                          0.0198019802
##   Lithuania                                        0.0098522167
##   Macedonia                                        0.0099255583
##   Malaysia                                         0.0000000000
##   Moldova                                          0.0497695853
##   Netherlands                                      0.1394891945
##   Nigeria                                          0.0095328885
##   Pakistan                                         0.0199704142
##   Palestinian territories (West Bank and Gaza)     0.0398724083
##   Peru                                             0.0298260149
##   Poland                                           0.0495238095
##   Romania                                          0.0096246391
##   Russian Federation                               0.0600000000
##   Saudi Arabia                                     0.0500000000
##   Serbia                                           0.0299227799
##   South Africa                                     0.0396039604
##   South Sudan                                      0.0598039216
##   Spain                                            0.0899563319
##   Sweden                                           0.0808080808
##   Switzerland                                      0.0896686160
##   Tunisia                                          0.0000000000
##   Turkey                                           0.0203488372
##   Ukraine                                          0.0296150049
##   United States                                    0.0499001996
##   Uzbekistan                                       0.0200000000
##   Vietnam                                          0.0000000000
```

```
### sort the proportions
sorted_proportions <- t(t((sort(country_proportion,decreasing = T))))
sorted_proportions
```

```
##
##                                                  [,1]
##   China                                          0.4700000000
##   Japan                                          0.3069306931
##   Czech Republic                                 0.3000000000
##   France                                         0.2873222749
##   Korea, Rep (South)                             0.1503611293
##   Germany                                        0.1494023904
```

```
##   Netherlands                                    0.1394891945
##   Australia                                      0.1000962464
##   Austria                                        0.0998003992
##   Iceland                                        0.0997652582
##   Ireland                                        0.0990099010
##   Hong Kong                                      0.0900000000
##   Spain                                          0.0899563319
##   Canada                                         0.0898203593
##   Switzerland                                    0.0896686160
##   Sweden                                         0.0808080808
##   Italy                                          0.0800405268
##   Belgium                                        0.0796963947
##   Argentina                                      0.0706357215
##   Russian Federation                             0.0600000000
##   Finland                                        0.0598984772
##   South Sudan                                    0.0598039216
##   Saudi Arabia                                   0.0500000000
##   United States                                  0.0499001996
##   Moldova                                        0.0497695853
##   Poland                                         0.0495238095
##   Bosnia and Herzegovina                         0.0400000000
##   Palestinian territories (West Bank and Gaza) 0.0398724083
##   South Africa                                   0.0396039604
##   India                                          0.0302197802
##   Serbia                                         0.0299227799
##   Peru                                           0.0298260149
##   Cameroon                                       0.0297619048
##   Colombia                                       0.0297029703
##   Ukraine                                        0.0296150049
##   Turkey                                         0.0203488372
##   Armenia                                        0.0202020202
##   Kenya                                          0.0200000000
##   Uzbekistan                                     0.0200000000
##   Pakistan                                       0.0199704142
##   Ecuador                                        0.0198019802
##   Lebanon                                        0.0198019802
##   Bulgaria                                       0.0188866799
##   Georgia                                        0.0100000000
##   Brazil                                         0.0099900100
##   Macedonia                                      0.0099255583
##   Lithuania                                      0.0098522167
##   Fiji                                           0.0098231827
##   Romania                                        0.0096246391
##   Nigeria                                        0.0095328885
##   Afghanistan                                    0.0000000000
##   Azerbaijan                                     0.0000000000
##   Ghana                                          0.0000000000
##   Iraq                                           0.0000000000
##   Malaysia                                       0.0000000000
##   Tunisia                                        0.0000000000
##   Vietnam                                        0.0000000000
```

```r
largest_country <- dimnames(sorted_proportions)[[1]][1]
largest_country
```

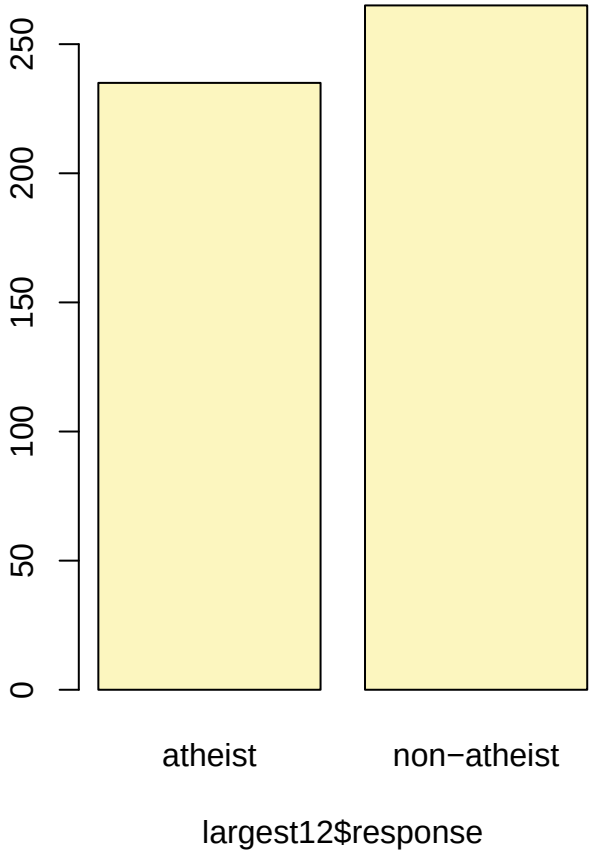
```
## [1] "China"
```

```
largest12 <- subset(atheism, nationality == largest_country & year == "2012")
summary(largest12, maxsum = 2)
```

```
##   nationality         response        year
##  China  :500   atheist    :235   Min.   :2012
##  (Other):  0   non-atheist:265   1st Qu.:2012
##                                  Median :2012
##                                  Mean   :2012
##                                  3rd Qu.:2012
##                                  Max.   :2012
```

```
inference(largest12$response, est = "proportion", type = "ci", method = "theoretical", success = "athei
```

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




```
## p_hat = 0.47 ;  n = 500
## Check conditions: number of successes = 235 ; number of failures = 265
## Standard error = 0.0223
## 95 % Confidence interval = ( 0.4263 , 0.5137 )
```

For *China* , the confidence interval is (0.4263 , 0.5137) .

As indicated above, the conditions have been checked - there are more than 10 successes and failures.

```r
second_country  <- dimnames(sorted_proportions)[[1]][2]
second_country
```

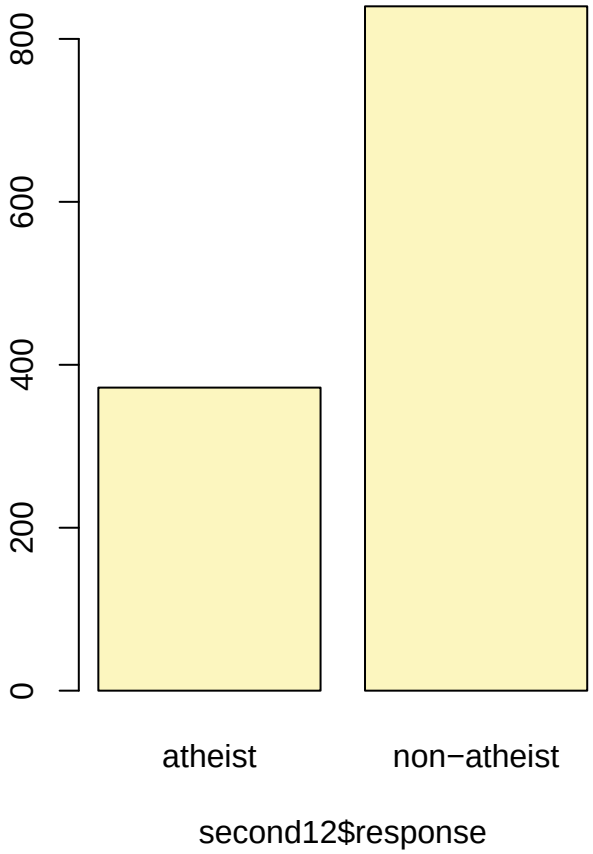
```
## [1] "Japan"
```

```r
second12 <- subset(atheism, nationality == second_country & year == "2012")
summary(second12, maxsum = 2)
```

```
##   nationality          response        year
##  Japan  :1212   atheist    :372   Min.   :2012
##  (Other):   0   non-atheist:840   1st Qu.:2012
##                                   Median :2012
##                                   Mean   :2012
##                                   3rd Qu.:2012
##                                   Max.   :2012
```

```r
inference(second12$response, est = "proportion", type = "ci", method = "theoretical", success = "atheis
```

```
## 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 )
```

For *Japan* , the confidence interval is (0.281 , 0.3329) .

As indicated above, the conditions have been checked - there are more than 10 successes and failures.

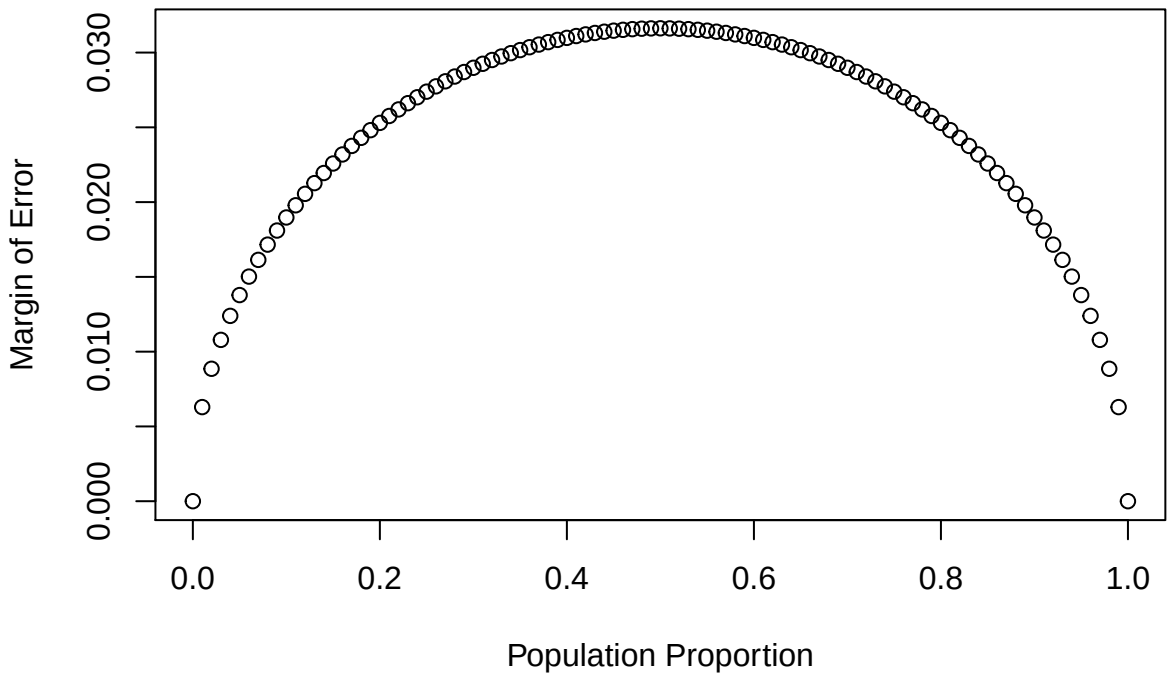
How does the proportion affect the margin of error?

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`.

The margin of error is an inverted (concave) curve with minima at p=0 and p=1, and a maximum at p=0.5 .

Success-failure condition

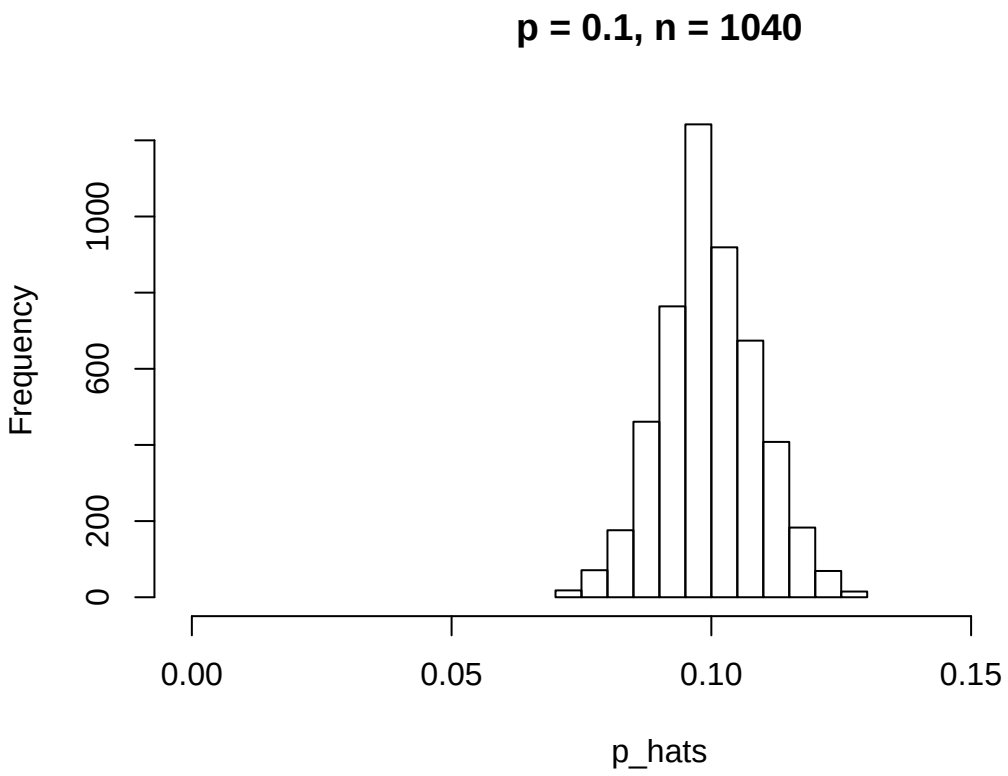
The textbook emphasizes that you must always check conditions before making inference. For inference on proportions, the sample proportion can be assumed to be nearly normal if it is based upon a random sample of independent observations and if both $np \geq 10$ and $n(1-p) \geq 10$. This rule of thumb 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
}
par(mfrow=c(1,1))
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.

```
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
```

```
phats_mean <- mean(p_hats)
phats_mean
```

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

```
phats_stdev <- sd(p_hats)
phats_stdev
```

```
## [1] 0.0092873823
```

The sampling distribution of sample proportions is unimodal with mean = 0.09969 and standard deviation 0.00928738 . It appears to be normal.

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?

```
hist_proportion <- function(p,n) {

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
}

titlestring = paste("p = ", toString(p) , " , n = " , toString(n))
histogram <- hist(p_hats, main = titlestring, xlim = c(0, 0.18))
#histogram <- hist(p_hats, main = titlestring)

mean_proportion <- mean(p_hats)
median_proportion <- median(p_hats)
stdev_proportion <- sd(p_hats)
skew_proportion <- skew(p_hats)
kurt_proportion <- kurtosi(p_hats)
shapiro_wilk <- shapiro.test(p_hats)
print(shapiro.test(p_hats))
if(shapiro_wilk$p.value>0.05) print("Normality PASSED") else print("Normality FAILED")

#plot(histogram)
return(c(mean_proportion, median_proportion, stdev_proportion, skew_proportion, kurt_proportion, shapir
}


par(mfrow=c(2,2))
result_10_1040 <- hist_proportion(.1,1040)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  p_hats
## W = 0.998131, p-value = 0.000010412
##
## [1] "Normality FAILED"
```

```
result_10_0400 <- hist_proportion(.1,400)
```

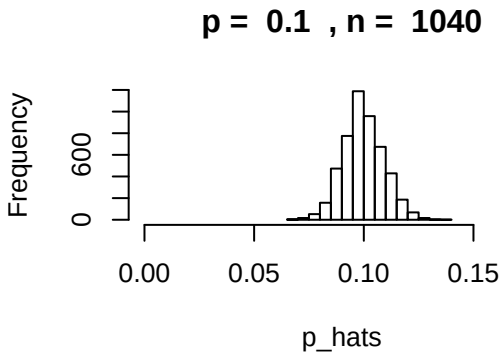
```
##
##  Shapiro-Wilk normality test
##
## data:  p_hats
## W = 0.995817, p-value = 0.00000000009431
##
```

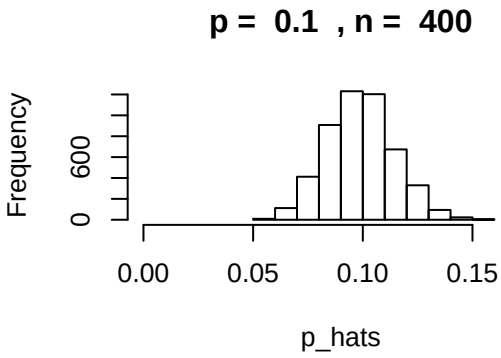
```
## [1] "Normality FAILED"
```

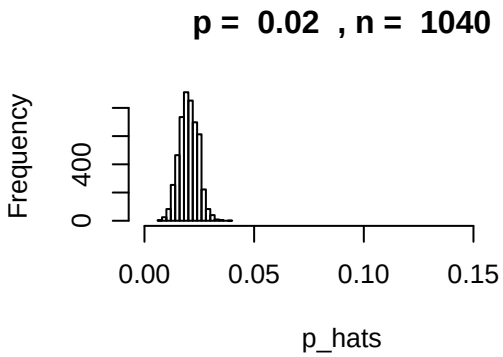
```
result_02_1040 <- hist_proportion(.02,1040)
```

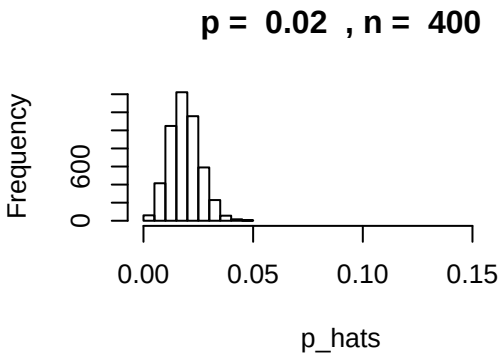
```
##
##  Shapiro-Wilk normality test
##
## data:  p_hats
## W = 0.992645, p-value = 0.0000000000000020646
##
## [1] "Normality FAILED"
```

```
result_02_0400 <- hist_proportion(.02,400)
```










```
##
##  Shapiro-Wilk normality test
##
## data:  p_hats
## W = 0.98102, p-value < 0.000000000000000222
##
## [1] "Normality FAILED"
```

Drawing a fresh sample for the *first* graph, where p=0.10 and n=1040, has

mean=0.09991096 , median=0.1, and stdev=0.00924425 .

The skew=0.10522501 and kurtosis=0.00104055.

The result of the Shapiro-Wilks test of Normality is a p-value of 0.00001041 .

#.

The *second* graph, where p=0.1 and n=400, has

mean=0.1001565 , median=0.1, and stdev=0.01525216 .

The skew=0.16836617 and kurtosis=-0.04503833.

The result of the Shapiro-Wilks test of Normality is a p-value of 0 .

The distribution is centered at the same place, but it is wider than the distribution where n=1040.

#.

The *third* graph, where p=0.02 and n=1040, has

mean=0.01994173 , median=0.02019231, and stdev=0.00423882 .

The skew=0.22425442 and kurtosis=0.02028649.

The result of the Shapiro-Wilks test of Normality is a p-value of 0 .

The distribution is centered at p=0.02, and is quite narrow.

#.

The *fourth* graph, where p=0.02 and n=400, has

mean=0.020053 , median=0.02, and stdev=0.00697545 .

The skew=0.37384267 and kurtosis=0.27338164.

The result of the Shapiro-Wilks test of Normality is a p-value of 0 .

The distribution is centered at p=0.02, but it is wider than that in the third graph. However, it is still considerably narrower than the earlier distributions.

Note that here the expected number of "successes" is only .02 * 400 = 8 . Because this is less than 10, the conditions for inference are not technically satisfied. This means that the results may be suspect.

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?

While the data from Australia indicates that about 104 respondents indicated that they are atheist, the data from Ecuador suggests that the number is 8. The conditions for inference require a minimum of 10 "successes" and 10 "failures" in each unit. This would appear to *rule out Ecuador from inference* as the relevant conditions are not met. Regarding Australia, the study indicates that the results were obtained over an interval of two days via online responses, which (in my opinion) renders the data suspicious.

I am not finding the portion of the report which "proceed[s] with inference and report[s] margin[s] of errors" as suggested in the question. Other than a single reference (on page 7-8) asserting that "In general the error margin for surveys of this kind is +/- 3-5% at 95% confidence level" there are no further claims which I could find in the report. (If there were, I would be sceptical for the reasons listed above, as well as those further qualms detailed below.)

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.

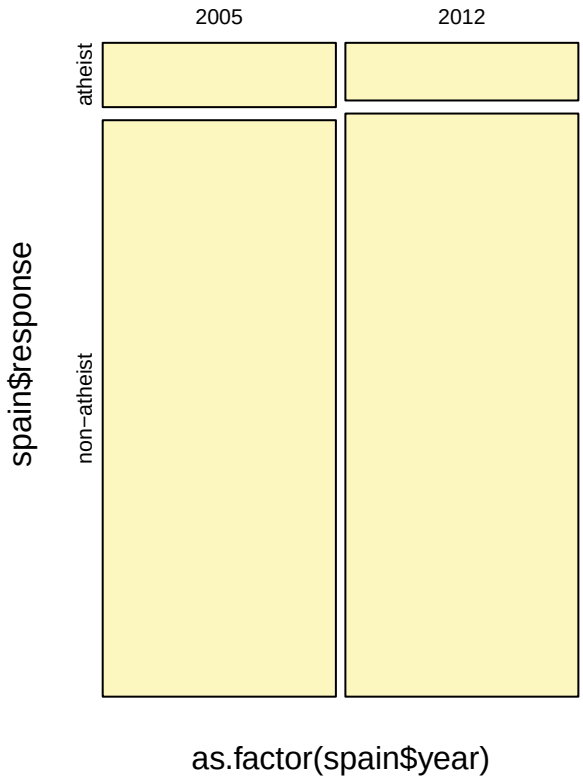
(12) Answer the following two questions using the `inference` function. As always, write out the hypotheses for any tests you conduct and outline the status of the conditions for inference.

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

```
spain         <- subset(atheism,nationality=="Spain")
spain_proportion    <- sum(spain$response=="atheist")/length(spain$response)
inference(y=spain$response, x=as.factor(spain$year), est="proportion", type="ci", method = "theoretical
```

```
## 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
##
## Check conditions:
##    2005 : number of successes = 115 ; number of failures = 1031
##    2012 : number of successes = 103 ; number of failures = 1042
## Standard error = 0.0123
## 95 % Confidence interval = ( -0.0136 , 0.0344 )
```

No, there is not convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012. Although the proportion changed from 10% to 9%, this drop of 1% is not statistically significant becauase the confidence interval for the difference between proportions includes zero. For the change to be significant, the confidence interval would have to not cross zero.

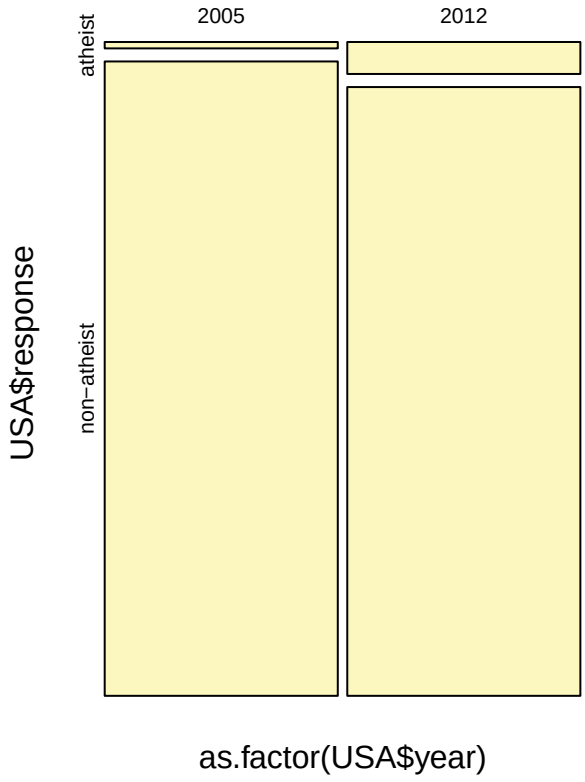
b. Is there convincing evidence that the United States has seen a change in its atheism index between 2005 and 2012?

```
USA         <- subset(atheism,nationality=="United States")
USA_proportion    <- sum(USA$response=="atheist")/length(USA$response)
USA_proportion
```

```
## [1] 0.02994012
```

```
inference(y=USA$response, x=as.factor(USA$year), est="proportion", type="ci", method = "theoretical", su
```

```
## 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
##
## Check conditions:
##    2005 : number of successes = 10 ; number of failures = 992
##    2012 : number of successes = 50 ; number of failures = 952
## Standard error = 0.0076
## 95 % Confidence interval = ( -0.0547 , -0.0251 )
```

The change in proportion in the US (from 1 percent to 5 percent) is significant because the confidence interval for the difference between proportions (-0.0547 , -0.0251) does not include zero. * Based on the data provided*** this would indicate that there is a significant change in the the level of atheism in the United States. However, this presumes that there isn't any problem with the sampling methodology. As discussed at top, certain of the techniques used by Gallup may not be defensible. In particular, we have to presume that their sample is indeed random and representative of the population. Unfortunately there is not sufficient detail given to confirm this.**

Also it is noteworthy that although there is a well-known US polling organization which also uses the "Gallup" name, the entity which organized this survey (Gallup International) is not the same firm. Furthermore, there appears to be a legal dispute between the two entities regarding rights to the "Gallup" name.

While the website of the US entity ("*Gallup Inc*", which did not conduct this poll) indicates that they have offices all over the world, the separate "*Gallup International*" organization appears to be a confederation of independent entities located in various countries around the globe. The survey from 2012 indicated that the US portion of the survey was conducted by some firm known as "TRiG". Online investigation indicates that there was such a firm known as "The Research Intelligence Group" which was a subsidiary of a Canadian firm known as "Leger,", which obtained the responses to the Canadian portion of the survey. Both such surveys (USA and Canada) were conducted "online" which makes one wonder how they obtained their sample and how they assessed the veracity of the response. It is noteworthy that neither "TRiG" nor "Leger" is listed as current affiliates of Gallup International. According to the Gallup International website, presently such surveys in Canada are now conducted by "Lightspeed Research" while in the United States, they now use surveymonkey.com !

Furthermore, although the Gallup International website contains a listing of "Members and Partners," such listing does not define which affiliate falls into each categorization, nor does it define what is the distinction between the two roles.

http://www.gallup-international.com/wp-content/uploads/2019/03/Members-and-Partners-List_2019.pdf

As *Gallup International* are now using an online website for United States surveys, and as their former US (and Canadian) affiliate from 2012 appears to have exited that role, this begs the question as to what entity bears responsibility for their US results. I grow extremely sceptical when looking at the results of this "Poll", especially in those countries where surveys were conducted online.

(13) 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.

There are 39 countries in table 4. At a significance level of 0.05, we would expect to detect a change in 5 percent of the countries simply by chance, which is 1.95 countries (of course, we would round to 2 in order to eliminate fractional countries.)

(14) 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.

```
Z <- qnorm(p = 1-.025)
p <- 0.5
ME <- 0.01
n <- (Z/ME)^2 * p * (1-p)
n
```

```
## [1] 9603.6471
```

```
roundup <- ceiling(n)
roundup
```

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

To achieve a margin of error no greater than 1% with 95% confidence, you would have to sample *9604* people.