Lab 8: Inference for Categorical Data

In August of 2012, news outlets ranging from the Washington Post to the Huffington Post ran a story about the rise of atheism in America. The source for the story was a poll that asked people, “Irrespective of whether you attend a place of worship or not, would you say you are a religious person, not a religious person or a convinced atheist?” This type of question, very common in polling, 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.

Template for lab report & inference function

Write your report, or at least run the code and create the plots, as you go so that if you get errors you can ask your TA to help on the spot. Knit often to more easily determine the source of the error.

```
download.file("http://stat.duke.edu/~cr173/Sta102_Fa14/Lab/lab8.Rmd", destfile="lab8.Rmd", method="wget")
```

Also download the `inference` function if you haven’t done so last week.

```
source("http://bitly.com/dasi_inference")
```

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 read over Pages 2 through 8 of the report, then address the following questions.

> **Exercise 1** Exactly how many people were interviewed for this survey in 2012, and what different methods were used to contact them?
>
> **Exercise 2** The title of the report is “Global Index of Religiosity and Atheism”. In order for inference results based on these data to be generalizable to the global human population, the survey must reflect a truely random sample of that population. Does that seem like a reasonable assumption in this case?
>
> **Exercise 3** In the first paragraph on Page 2, several key findings are reported. Do these percentages appear to be *sample statistics* or *population parameters*?

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 dataset of individual responses to the survey. Load this dataset into R with the following command.

```
download.file("http://stat.duke.edu/~cr173/Sta102_Fa14/Lab/atheism.RData",
                 destfile = "atheism.RData", method="wget")
load("atheism.RData")
```

This lab was written for OpenIntro by Andrew Bray and Mine Çetinkaya-Rundel.

Exercise 4 What does each row of Table 6 correspond to? What does each row of the data set 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 arrive at the same number using the atheism data.

Exercise 5 Create a new dataframe called us12 that contains only the rows in atheism associated with respondents from the United States to the 2012 survey. Then, using the command below, calculate the proportion of atheist responses. Does it agree with the percentage in Table 6? If not, why?

```
us12 = subset(atheism,  atheism$nationality == "United States"
                      & atheism$year == "2012")
sum(us12$response == "atheist") / length(us12$response)
```

Note that the above piece of code first counts how many atheists there are in the sample, and then divides this number by the total sample size.

Inference on proportions

As was hinted at in Exercise 3, 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.

Exercise 6 Write out the conditions for inference to construct a 95% confidence interval for the proportion of atheists in the United States in 2012. How reasonable is each condition?

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

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 $\pm 3 - 5\%$ at 95% confidence".

Exercise 7 Based on the R output, what is the margin of error for the estimate of the proportion of atheists in US in 2012?

Exercise 8 Using the inference function, calculate a confidence interval for the proportion of atheists in 2012 in another countriy of your choice, and report the associated margin of error.
Hint: You will need to create a new subsetted dataset for the country of your choice for 2012, and then use this dataset in the inference function to construct the confidence intervals.

How does the proportion affect the margin of error?

Imagine you've set out to survey 100 classmates 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.

We can visualize this relationship by creating 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 the relationship.

```r
n = 100
p = seq(0, 1, 0.01)
me = 2*sqrt(p*(1 - p)/n)
plot(me ~ p)
```

Exercise 9 Describe the relationship between `p` and `me`.

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 on 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. However, when np and $n(1 - p)$ reaches 10 the sampling distribution is sufficiently normal to use confidence intervals and hypothesis tests that are based on that approximation.

We can investigate the interplay between n and p and the shape of the sampling distribution by using simulations. To start off, we simulate the process of drawing 5000 samples of size 1040 from a population with a true atheist proportion of 0.1. For each of the 5000 samples we compute $\hat{p}$ and then plot a histogram to visualize their distribution.

```r
p = 0.1
n = 1040

p_hats = rep(0, 5000)
for(i in 1:5000){
  samp = sample(c("atheist", "non_atheist"), n, replace = TRUE, prob = c(p, 1-p))
  p_hats[i] = sum(samp == "atheist")/n
}

hist(p_hats, main = "p = 0.1, n = 1040")
```

These commands build up the sampling distribution of `p_hats` using the familiar for loop. You can read the sampling procedure as, "take a sample of size n with replacement from the choices of atheist and non-atheist, with probabilities p and $1 - p$ respectively". Then, calculate the proportion of atheists in this sample, and record this value. Repeat this process 5,000 times to build the sampling distribution.

Exercise 10 Describe the shape of the sampling distribution when $n = 1040$ and $p = 0.1$.

Exercise 11 Is the success-failure condition met when $n = 1040$ and $p = 0.1$? How about when $n = 400$ and $p = 0.02$?

Exercise 12 Replicate the sampling distribution of sample proportions with $n = 400$ and $p = 0.02$. How does the shape of the new sampling distribution compare to the previous one?
Hint: You only need to change `n` and `p` in the above code. In order to make sure the plot is labeled correctly, also adjust the `main` argument in the `hist` function.

Exercise 13 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. Given the shape of their respective sampling distributions, do you think it is sensible to proceed with inference based on theoretical methods, and report margin of errors, as the reports does?

Additional Exercises

The question of atheism was asked by WIN-Gallup International in a similar survey that was conducted in 2005.[†] Table 4 summarizes the results from the 2005 and 2012 surveys.

Exercise 14 Answer the following two questions using the `inference` function.

(a) Is there convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012?
Hint: Create a new dataset for respondents from Spain. Then use their responses as the first input on the `inference`, and use year as the grouping variable.

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

Exercise 15 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?

Exercise 16 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, at 95% confidence level, must have a margin of error of no greater than 3%. 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: You do not need to use the dataset to answer this question.

[†] We assume here that sample sizes have remained the same.