Lab 3: 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, 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.uvm.edu/~rsingle/stat211/data/extra/Gallup-International-Religion+Atheism-2012.pdf

Take a moment to look over the report then address the following questions.

Exercise 1 On pages 2 & 3, several key findings are reported. Are these percentages *sample statistics* (derived from the data sample) or *population parameters*?

Exercise 2 The title of the report is "Global Index of Religiosity and Atheism". To generalize the report's findings to the global human population, what must we assume about the sampling method? Does that seem like a reasonable assumption?

The data

Turn your attention to Table 6 (pages 14 and 15), 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.

```
download.file("http://www.openintro.org/stat/data/atheism.RData", destfile =
"atheism.RData")
load("atheism.RData")
```

Exercise 3 What does each row of Table 6 correspond to? What does each row of atheism correspond to? Try using dim(), head(), and/or tail() on the dataset to investigate.

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.

Exercise 4 Using the first command below, create a new data frame 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? Some of the other commands below can help reproduce the 5% proportion for the US.

```
us12 <- subset(atheism, atheism$nationality == "United States" & atheism$year
=="2012")
nrow(us12); sum(us12$response == "atheist")
tab1 <- table(us12$response); tab1; sum(tab1)
```

Inference on a single proportion

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.

> **Exercise 5** What are the conditions for inference to construct a 95% confidence interval for the proportion of atheists in the United States in 2012. Does it seem that these 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 prop.test() function. We will arbitrarily choose the less frequent outcome as a "success", which is a response of atheist in this case.

```
prop.test(x=50, n=1002, p=0.5, correct=FALSE)
prop.test(tab1, p=0.3, correct=F) #Note: now we have Ho:p=0.3, but the CI is the same
```

- x counts of "successes" (can be scalar, vector, or table/matrix)
- n counts of trials (with length matching x), ignored if x is a table/matrix
- p probability of success under Ho (length matching x) (default=0.5 for 1-sample tests)
- alternative The alternative hypothesis can be "less", "greater", or "two.sided". (default="two.sided").
- conf.level confidence level of the interval. (deafult=0.95)
- correct TRUE/FALSE for using Yate's continuity correction. (deafult=TRUE)

Although formal confidence intervals and hypothesis tests don't show up in the report, suggestions of inference appear at the bottom of page 6: "In general, the error margin for surveys of this kind is $\pm 3\% - 5\%$ at 95% confidence."

> **Exercise 6** Based on the R output, what is the margin of error for the estimate of the proportion of atheists in US in 2012? Remember that the margin of error (ME) is ½ of the width of the CI.

> **Exercise 7** Using the prop.test() function, calculate confidence intervals for the proportion of atheists in 2012 in another country of your choice, and find the associated margins of error. Be sure to note whether the conditions for inference are met. It may be helpful to create a new data set for this country first, and then use this data set in the prop.test() to construct the confidence intervals.

Comparing two proportions

The dataset has entries for 2005 as well as 2012. Table 4 on page 12 of the report summarizes survey results from 2005 and 2012 for 39 countries. Let's compare the proportion of atheists in these two years. †

```
us <- subset(atheism, atheism$nationality == "United States")
tab2 = table(us$response,us$year); tab2
```

We can test for a significant difference in proportions using either of the commands below:

```
prop.test(x=c(10,50), n=c(10+992,50+952), correct=FALSE)
chisq.test(tab2, correct=FALSE) #Note: the same statistic and p-value, but less detail
```

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

On your own

1. Using only the 2012 data, test the null hypothesis that the true proportion of atheists in France is 25% vs. the alternative that it is more than 25%. Write out the hypotheses, report the statistic and p-value, and state a conclusion at the 0.05 level. (Do not use the continuity correction). Also, provide a 95% confidence interval for the true proportion of atheists in France.
2. 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.) Write out the hypotheses, report the statistic and p-value, and state a conclusion at the 0.05 level. (Do not use the continuity correction)
3. If there really had been no change in the atheism index in any of the 39 countries listed in Table 4, in how many of the countries would you expect to detect a change simply by chance if you are testing at the 0.05 level of significance?

Extra/Optional: 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, you might think they would have the same margin of error. However, the margin of error not only changes with sample size, it is also affected by the proportion.

Think about the formula for the standard error: $SE = \sqrt{p\,(1-p)/n}$. This SE is 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 makes sense that the margin of error is dependent on the population proportion. We can visualize this relationship by creating a plot of ME vs. p. The code below creates this plot.

```
n <- 1000
p <- seq(0, 1, 0.01)
ME <- 1.96 * sqrt(p * (1 - p)/n)
plot(ME ~ p)
```

The first step makes 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. Lastly, we plot the two vectors against each other to show the relationship.

> **Exercise X** Describe the relationship between p and ME. For what value of p is the ME maximized?