Lab 6: 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 look at the atheism survey and explore what's at play when making inference about population proportions using categorical data.[1]

The Survey

To access the press release for the poll, conducted by WIN-Gallup International, click 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 and then address the following questions.

Exercise 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*?

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 15 and 16), which reports the sample size and response percentages for all 57 countries. Although 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. The CSV file containing the data is named **atheism_sas.csv**.

Import the file **atheism_sas.csv** from the folder where you saved it. Output it into a SAS data set.

✎   If you are using SAS University Edition, you need to ensure that interactive mode is turned off. To do this, click the button to the right of **Sign Out** in the upper right corner of the window and then click **Preferences**. In the Preferences window, on the General tab, the bottom check box (located next to the text **Start new programs in interactive mode**) should *not* be selected. If the box is selected, you need to clear it and save your change.

```
filename atheism url
        'http://www.openintro.org/stat/data/atheism_sas.csv';

proc import datafile=atheism
            out=work.atheism
            dbms=csv
            replace;
    getnames=yes;
run;
```

PROC IMPORT can import data files of types other than SAS into SAS. The OUT= option names the output data set. Without a library name, the SAS data set is saved in the **Work** library and is deleted when the SAS session is ended. The DBMS= option tells PROC IMPORT that the file being read is a comma-separated values file. The REPLACE option enables PROC IMPORT to replace a data set with the same name in the **Work** library. Because the **atheism_sas.csv** file has a first row containing column names, GETNAMES=YES is used to retrieve those names for use as variable names in the SAS data set.

Exercise 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, look at the estimated proportion of atheists in the United States. Toward 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 DATA step below, create a new data set named **us12** that contains only the rows in **atheism** that are 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?

```
data work.us12;
   set work.atheism;
   if nationality="United States" and year=2012;
run;
```

A new data set, **us12**, is created by reading in the **atheism** data set using the SET statement. The IF statement requests that only observations (rows) in the **atheism** data set that meet certain conditions be output into the **us12** data set. These conditions are that the nationality is United States and that the year is 2012.

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. 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 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?

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

```
ods graphics;
proc freq data=work.us12;
   tables response / binomial(level="atheist") plots=freq;
run;
```

The TABLES statement names the variable to analyze and requests a frequency table of all values in the variable **response**. Options for the TABLES statement follow the /. The option BINOMIAL requests a table of statistics for a binomial proportion. This is what we want. Binomial proportions are appropriate for only dichotomous (two-valued) categorical variables. We then have two choices for level to analyze. In this case, we choose **level='atheist'**. We can also ask for a frequency plot to accompany the frequency table.

Confidence intervals and inferences are reported by default. Confidence limits are calculated using both the *asymptotic* method, which uses the standard errors described in this text, and the *exact* method, which uses a different approach. ASE stands for Asymptotic Standard Error and is used for calculating the asymptotic confidence limits. The confidence level is 95%, but you can choose any other level of confidence using an ALPHA= option after the / in the TABLES statement. For example, for a 99% CI, you would first calculate 1-0.99 to get the appropriate alpha level. The option would then read ALPHA=0.01.

The two-sided *p*-value reported in the output is the *p*-value for the test of the null hypothesis that the population proportion is 0.50. You can use a P= option in the parentheses after the BINOMIAL option to change that null hypothesis value. The one-sided *p*-value is half the value of the two-sided *p*-value and is used for null hypotheses of the form $p <= P$ or $p >= P$, where P is the hypothesized value (0.50, by default). The *p*-value is valid only if the sample *p* is on the opposite side of the null hypothesis (for example, if $p > P$, when the null hypothesis is $p <= P$). Otherwise, the *p*-value is 1 minus the reported *p*-value.

Exercise 6: Based on the SAS output, what is the margin of error for the estimate of the proportion of atheists in the US in 2012?

Exercise 7: Using PROC FREQ, 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 might be helpful to create new data sets for each of the two countries first, and then use these data sets in PROC FREQ to construct the confidence intervals.

How Does the Proportion Affect the Margin of Error?

Imagine you've set out to survey 1000 people on two questions: Are you female? Are you left-handed? Because both of these sample proportions were calculated from the same sample size, they should have the same margin of error, right? Wrong! Although 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: $SE = p(1 - p)/n$. This is then used in the formula for the margin of error for a 95% confidence interval: $ME = 1.96 \times SE = 1.96 \times p\,(1 - p)/n$. Because 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* versus *p*.

The first step is to make a variable **p** that is a sequence from 0 to 1 with each number separated by 0.01. We can then create a variable of the margin of error (**me**) associated with each of these values of **p** using the familiar approximate formula ($ME = 2 * SE = 2 * \sqrt{p * (1 - p)/n}$). Lastly, we plot the two variables against each other to reveal their relationship.

```
data work.proportions;
   do p=0.01 to 0.99 by 0.01;
      me=2 * sqrt(p * (1 - p)/1000);
         output;
   end;
run;

proc sgplot data=work.proportions;
   scatter x=p y=me;
run;
```

✏️   Without the OUTPUT statement in the DO/END loop, one observation would be created the first time through and then that observation and its values would be overwritten again and again until **p** reached the value 1.

Exercise 8: 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. What is the "best" value for such a rule of thumb is, at least to some degree, arbitrary.

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 random 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.

```
data work.multiple_samples;
   call streaminit(27513);
   do sample=1 to 5000;
      do obs=1 to 1040;
      if rand("UNIFORM") < 0.1 then atheist=1;
         else atheist=0;
         output;
      end;
   end;
run;

proc means data=work.multiple_samples noprint nway;
   class sample;
   var atheist;
   output out=work.means mean=;
run;
```

```
ods select histogram;
proc univariate data=work.means;
   histogram atheist;
   inset mean std;
   title "5000 samples, p=0.01, n=1040";
run;
```

In order to make anything random in SAS, you need to start with a seed for a pseudo-random number generator. The CALL STREAMINIT statement does this. The seed within the parentheses can be anything you want. The RAND statement generates samples randomly from the uniform distribution, whose range is 0 to 1. Here, the variable **atheist** is assigned the value 1 with a 10% probability and the value of 0 with a 90% probability. There are two DO loops and they are nested. The inner DO loop creates a sample of 1040 observations, and the outer DO loop creates 5000 of these samples.

PROC MEANS calculates summary statistics for variables in the VAR statement. Because the variable **atheist** is coded 0 and 1, the mean of the variable is the proportion of 1 values (atheists) in the sample. The CLASS statement calculates this value separately for each value of sample. The OUTPUT statement creates a data set named means to store the 5000 mean values of atheist.

PROC UNIVARIATE is used to generate a histogram of the mean values calculated in PROC MEANS. The INSET statement inserts a text box with summary statistics for the distribution. The ODS SELECT statement chooses only the histogram for display.

> ✎ PROC SGPLOT could have been used to generate the histogram, but it would not produce an inset.

Exercise 9: Describe the sampling distribution of sample proportions at $n = 1040$ and $p = 0.1$. Be sure to note the center, spread, and shape.

Exercise 10: Replicate 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 three histograms. 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?

When you're done, you can reset the titles by submitting the command TITLE.

Exercise 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?

On Your Own

The question of atheism was asked by WIN-Gallup International in a similar survey that was conducted in 2005[2]. Table 4 on page 13 of the report summarizes survey results from 2005 and 2012 for 39 countries.

1. Answer the following two questions using the inference function. As always, write out the hypotheses for any tests that 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. 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?

2. 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.

3. 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.

4. What concepts from the textbook are covered in this lab? What concepts, if any, are not covered in the textbook? Have you seen these concepts elsewhere (for example, lecture, discussion section, previous labs, or homework problems)? Be specific in your answer.

---

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