Complete all *Exercises*, and submit answers to Questions on the Coursera platform. Note that the order of the choices in multiple choice questions may be different on the Coursera platform than the order in this document.

Lab 5: Inference for Categorical Data

Load the inference function:

```
source("http://bit.ly/dasi_inference")
```

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.wingia.com/web/files/richeditor/filemanager/Global_INDEX_of_Religiosity_and_Atheism_PR__6.pdf

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

Question 1 [MULTIPLE CHOICE] How many people were interviewed for this survey?

(a) A poll conducted by WIN-Gallup International surveyed 51,000 people from 57 countries

(b) A poll conducted by WIN-Gallup International surveyed 52,000 people from 57 countries

(c) A poll conducted by WIN-Gallup International surveyed 51,917 people from 57 countries

(d) A poll conducted by WIN-Gallup International surveyed 51,927 people from 57 countries

Question 2 [MULTIPLE CHOICE] Which of the following methods were used to gather information

(a) Face to face

(b) Telephone

(c) Internet

(d) All of the above

Question 3 [TRUE / FALSE] In the first paragraph, several key findings are reported. These percentages appear to be *sample statistics*

- TRUE
- FALSE

Question 4 [TRUE / FALSE] The title of the report is "Global Index of Religiosity and Atheism". To generalize the report's findings to the global human population, We must assume that the sample was a random sample from the entire population in order to be able to generalize the results to the global human population. This does seem to be a reasonable assumption.

- TRUE
- FALSE

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(url("http://www.openintro.org/stat/data/atheism.RData"))
```

Question 5 [MULTIPLE CHOICE] What does each row of Table 6 correspond to?

(a) Countries

(b) Individual Persons

(c) Religions

Question 6 [MULTIPLE CHOICE] What does each row of `atheism` correspond to?

(a) Countries

(b) Individual Persons

(c) Religions

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.

Question 7 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. [TRUE / FALSE] This percentage agrees with the percentage in Table 6.

- TRUE
- FALSE

```
us12 = subset(atheism, atheism$nationality == "United States" & atheism$year ==
    "2012")
```

Inference on proportions

As was hinted at in Exercise 4, 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*** 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(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*** 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?

> **Question 8** [MULTIPLE CHOICE] 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?
>
> (a) The margin of error for the estimate of the proportion of atheists in the US in 2012 is 0.05.
>
> (b) The margin of error for the estimate of the proportion of atheists in the US in 2012 is 0.025.
>
> (c) The margin of error for the estimate of the proportion of atheists in the US in 2012 is 0.0135.

> ***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.

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)
```

Question 9 [MULTIPLE CHOICE] Which of the following is false about the relationship between `p` and `me`.

(a) The `me` reaches a minimum at p = 0.

(b) The `me` reaches a minimum at p = 1.

(c) The `me` is maximized when p = 0.5.

(d) The most conservative estimate when calculating a confidence interval occurs when p is set to 1.

The question of atheism was asked by WIN-Gallup International in a similar survey that was conducted in 2005.[†] 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.

Question 10 [TRUE / FALSE] There is convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012.

- TRUE
- FALSE

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.

Question 11 [TRUE / FALSE] There is convincing evidence that the United States has seen a change in its atheism index between 2005 and 2012.

- TRUE
- FALSE

Question 12

[MULTIPLE CHOICE] 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.

(a) 0

(b) 1

(c) 1.95

(d) 5

Question 13 [MULTIPLE CHOICE] 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

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

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?

(a) 2401 people

(b) At least 2401 people

(c) 9604 people

(d) At least 9604 people

Hint: Refer to your plot of the relationship between p and margin of error. Do not use the data set to answer this question.

Final notes: If you want to turn off the exploratory analysis plot in the output of the `inference` function (you might want to do this in your projects to avoid repetitive graphs), you can simply add the argument `eda_plot = FALSE` to the function. Similarly, if you want to turn off the p-value plot, add `inf_plot = FALSE`. The defaults for both of these arguments are set to `TRUE`.

End of Lab Survey

The following questions are not graded, but your feedback is very much appreciated and immensely useful for the development of the course.

Question 14 [MULTIPLE CHOICE] This lab covered material that is covered in the class.

(a) Strongly Disagree

(b) Disagree

(c) Neutral

(d) Agree

(e) Strongly Agree

Question 15 [MULTIPLE CHOICE] The lab improved your understanding of these topics.

(a) Strongly Disagree

(b) Disagree

(c) Neutral

(d) Agree

(e) Strongly Agree

Question 16 [MULTIPLE CHOICE] The instructions were clear and it was easy to understand what was wanted.

(a) Strongly Disagree

(b) Disagree

(c) Neutral

(d) Agree

(e) Strongly Agree

Question 17 [MULTIPLE CHOICE] The data were relevant and interesting to me.

(a) Strongly Disagree

(b) Disagree

(c) Neutral

(d) Agree

(e) Strongly Agree

Question 18 [MULTIPLE CHOICE] The length of time took to complete lab.

(a) Less than 30 minutes

(b) Between 30 minutes and 1 hour

(c) Between 1 hour and 2 hours

(d) More than 2 hours