**SEIS 631 Foundations of Data Analysis**

**Assignment 7**

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. We will 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, open the Global Index of Religiosity and Atheism article in the Assignment 7 folder.

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

|  |
| --- |
| **Q1**) 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*? |

|  |
| --- |
| **Q2**)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. 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 the **atheism.RData** into R.

|  |
| --- |
| **Q3)** 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, 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.

|  |
| --- |
| **Q4)** 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")**

## **Inference on proportions**: 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: the confidence interval and the hypothesis test.

|  |
| --- |
| **Q5)** 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 prop.test function to do it for us (ignoring the hypothesis testing output).

**prop.test(table(us12$response)[1], sum(table(us12$response)), correct=F)**

Note that the first argument gives the number of successes (50), and the second argument gives the total number sampled (1,002).

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”.

|  |
| --- |
| **Q6)** 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? |

|  |
| --- |
| **Q7)** Using the prop.test 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 prop.test 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 = 1.96SE = 1.96 \* 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 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)**

|  |
| --- |
| **Q8) Describe the relationship between p and me.** |

**Success-failure condition**

We have emphasized 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 and n(1-p) is greater than equal to 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 samples of size 1040 from a population with a true atheist proportion of 0.1. For each of the 5000 samples we compute the sample proportion 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**

**}**

**hist(p\_hats, main = "p = 0.1, n = 1040", xlim = c(0, 0.18))**

These commands build up the sampling distribution of **p\_hats** using a 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.

|  |
| --- |
| **Q9)** 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. |

|  |
| --- |
| **Q10)** 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 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 the sample proportion? How does p affect the sampling distribution? |

Once you’re done, you can reset the layout of the plotting window by using the command **par(mfrow = c(1,1))** or clicking on “Clear All” above the plotting window (if using RStudio). Note that the latter will get rid of all your previous plots.

|  |
| --- |
| **Q11)** 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? |

**Submission:**

Submit the following two files (no need to compress into a folder):

* **Pdf or word document** with answers to all the questions asked in the assignment.
* **.R file** with relevant code. It is useful to add comments within the code with a pound sign.