# Inference for categorical data

**Your reproducible lab report:** Before you get started, download the R Markdown template for this lab. Remember all of your code and answers go in this document:

```
download.file("https://dyurovsky.github.io/85309/post/rmd/lab7.Rmd",
              destfile = "lab7.Rmd")
```

## Atheism

In August of 2012, news outlets ranging from the Washington Post (http://www.washingtonpost.com/national/on-faith/poll-shows-atheism-on-the-rise-in-the-us/2012/08/13/90020fd6-e57d-11e1-9739-eef99c5fb285_story.html) to the Huffington Post (http://www.huffingtonpost.com/2012/08/14/atheism-rise-religiosity-decline-in-america_n_1777031.html) 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:

https://dyurovsky.github.io/85309/post/slides/gallup_atheism.pdf (https://dyurovsky.github.io/85309/post/slides/gallup_atheism.pdf)

Take a moment to review the report 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*? Explain your reasoning.

# 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.

```
atheism <- read_csv("https://dyurovsky.github.io/85309/data/lab7/atheism.csv")
```

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 2** Using the command below, create a new tibble 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 <- atheism %>%
  filter(nationality == "United States" & year == "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; 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 the confidence interval and the hypothesis test.

**Exercise 3** 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 go ahead and compute the 95% confidence interval. Remember what we need: a point estimate of the mean, a critical value, and a standard error.

```
mean <- us12 %>%
  summarise(atheist = mean(response == "atheist")) %>% # the mean of the sample
  pull()

z_star <- qnorm(.975) #The 97.5% percentile of the normal distribution

se <- sqrt((mean * (1 - mean)) / nrow(us12)) #the formula for the standard error

me <- se * z_star # margin of error is standard error times critical value

ci_95 <- c(mean - me, mean + me)
```

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

**Exercise 5** Using this same process to 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, and interpret the interval in context of the data.It may be helpful to create new data sets for each of the two countries first, and then use these data sets 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 a morning person? 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. $p$.

Since sample size is irrelevant to this discussion, let's just set it to some value ($n = 1000$) and use this value in the following calculations:

```
n <- 1000
```

The first step is to make a variable `p` that is a sequence from 0 to 1 with each number incremented 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 ($ME = 1.96 \times SE$).

```
p <- seq(0, 1, 0.01)
me <- 1.96 * sqrt(p * (1 - p) / n)
```

Lastly, we plot the two vectors against each other to reveal their relationship. To do so, we need to first put these variables in a data frame that we can call in the `ggplot` function.

```
me_tibble <- tibble(p = p, me = me)
ggplot(me_tibble, aes(x = p, y = me)) +
  geom_point() +
  labs(y = "Margin of Error", x = "Population Proportion")
```

**Exercise 6** Describe the relationship between `p` and `me`. Include the margin of error vs. population proportion plot you constructed in your answer. For a given sample size, for which value of `p` is margin of error maximized?

# Success-failure condition

We emphasized over and over 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 $np \geq 10$ 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. 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. Play around with the following code by changing the `n` and `p` parameters to investigate how the shape, center, and spread of the distribution of $\hat{p}$ changes as $n$ and $p$ change.

```
# Set parameters.
# If you change these, all of the sampling will change because we pass these into
# the sampling function
n <- 300
p <- .1

# A function that takes a number of samples(n) and a true success probability(p),
# and returns the proportion of success in a random sample of n trials.
one_sample <- function(n, p) {

  # We're again going to use the sample function, but this time we're going to
  # make things a little bit simpler by using the prob parameters. This lets us
  # say use p to assign a probability of sampling each of the values in x,
  # instead of just making an array with lots of copies of the values like we
  # did before.
  sampled_values <- sample(x = c(TRUE, FALSE), n,
                           prob = c(p, 1-p),
                           replace = TRUE)

  # return the proportion of successes
  mean(sampled_values)
}

# Take 5000 samples and look at the distribution
replicates <- tibble(sample = 1:5000,
                     value = replicate(5000, one_sample(n, p)))

ggplot(replicates, aes(x = value)) +
  geom_histogram(binwidth = .01) +
  xlim(c(0, 1)) + # set x-axis limits so we can see the whole 0-1 range
  ylim(c(0, nrow(replicates)))
                # set y-axis limits so we can see the whole 0-replicates range
```

**Exercise 7**

Describe the sampling distribution of sample proportions at $n = 300$ and $p = 0.1$. Be sure to note the center, spread, and shape.

Now, Keep $n$ constant and change $p$. How does the shape, center, and spread of the sampling distribution vary as $p$ changes.

Now also change $n$.How does $n$ appear to affect the distribution of $\hat{p}$?

# More Practice

The question of atheism was asked by WIN-Gallup International in a similar survey that was conducted in 2005. (We assume here that sample sizes have remained the same.) Table 4 on page 13 of the report summarizes survey results from 2005 and 2012 for 39 countries.

**Exercise 8** Is there convincing evidence that Spain has seen a change in its atheism index between 2005 and 2012? As always, write out the hypotheses for any tests you conduct and outline the status of the conditions for inference. If you find a significant difference, also quantify this difference with a confidence interval.
*Hint:* First create a new data set for respondents from Spain.

**Exercise 9** Is there convincing evidence that the US has seen a change in its atheism index between 2005 and 2012? As always, write out the hypotheses for any tests you conduct and outline the status of the conditions for inference. If you find a significant difference, also quantify this difference with a condfidence interval.

**Exercise 10** 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:* Review the definition of the Type 1 error.

**Exercise 11** 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. This question does not require using the dataset.