Inference for categorical data

Getting Started

Load packages

In this lab, we will explore and visualize the data using the **tidyverse** suite of packages, and perform statistical inference using **infer**. The data can be found in the companion package for OpenIntro resources, **openintro**.

Let's load the packages.

```
library(tidyverse)
library(openintro)
library(infer)
```

The data

You will be analyzing the same dataset as in the previous lab, where you delved into a sample from the Youth Risk Behavior Surveillance System (YRBSS) survey, which uses data from high schoolers to help discover health patterns. The dataset is called `yrbss`.

1. What are the counts within each category for the amount of days these students have texted while driving within the past 30 days?

Want the count of each category in our `texting_while_driving_30d` column using `dplyr`

```
yrbss  %>%
     count(text_while_driving_30d)
```

```
## # A tibble: 9 x 2
##   text_while_driving_30d     n
##   <chr>                  <int>
## 1 0                       4792
## 2 1-2                      925
## 3 10-19                    373
## 4 20-29                    298
## 5 3-5                      493
## 6 30                       827
## 7 6-9                      311
## 8 did not drive           4646
## 9 <NA>                     918
```

2. What is the proportion of people who have texted while driving every day in the past 30 days and never wear helmets?

Remember that you can use `filter` to limit the dataset to just non-helmet wearers. Here, we will name the dataset `no_helmet`.

```
data('yrbss', package='openintro')
no_helmet <- yrbss %>%
  filter(helmet_12m == "never")
```

```
no_helmet_and_text <- yrbss %>%
                        # first, filter to the cases where both clauses (texting and no helmet) are tru
                        filter(helmet_12m == "never",
                               !is.na(text_while_driving_30d),
                               text_while_driving_30d != "did not drive",
                               text_while_driving_30d != "0"
                              ) %>%
                        # Calculate the proportion of "yes" values against our total
                        summarise(pct_text_no_helmet = n() / nrow(yrbss))

no_helmet_and_text
```

```
## # A tibble: 1 x 1
##   pct_text_no_helmet
##                <dbl>
## 1              0.134
```

Also, it may be easier to calculate the proportion if you create a new variable that specifies whether the individual has texted every day while driving over the past 30 days or not. We will call this variable `text_ind`.

```
no_helmet <- no_helmet %>%
  mutate(text_ind = ifelse(text_while_driving_30d == "30", "yes", "no"))
```

Inference on proportions

When summarizing the YRBSS, the Centers for Disease Control and Prevention seeks insight into the population *parameters*. To do this, you can answer the question, "What proportion of people in your sample reported that they have texted while driving each day for the past 30 days?" with a statistic; while the question "What proportion of people on earth have texted while driving each day for the past 30 days?" 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.

```
# Replace
# no_helmet$text_ind <- replace_na(no_helmet$text_ind, "no")

# Filtering out NA values so our text_ind column is only yes or no vals
no_helmet <- no_helmet %>%
                    filter(!is.na(text_ind))

#Should work now
no_helmet %>%
  specify(response = text_ind, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   0.0649   0.0777
```

Note that since the goal is to construct an interval estimate for a proportion, it's necessary to both include the `success` argument within `specify`, which accounts for the proportion of non-helmet wearers than have consistently texted while driving the past 30 days, in this example, and that `stat` within `calculate` is here "prop", signaling that you are trying to do some sort of inference on a proportion.

3. What is the margin of error for the estimate of the proportion of non-helmet wearers that have texted while driving each day for the past 30 days based on this survey?

The margin of error, or half the width of a confidence interval, is 0.475 in this case, or 47.5%. This is because we have a confidence interval of 95%, leaving us to travel half that distance to reach the edge of our confidence interval. Our margin of error given our lower and upper CI bounds is $(0.0775 - 0.0649)/2 \approx 0.0063$

```
ci <- no_helmet %>%
  specify(response = text_ind, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)
margin_of_err <- (ci$upper_ci[1] - ci$lower_ci[1]) / 2

margin_of_err
```

```
## [1] 0.006077964
```

4. Using the `infer` package, calculate confidence intervals for two other categorical variables (you'll need to decide which level to call "success", and report the associated margins of error. Interpet 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.

Would like to see the proportion of teens that are sleeping 8 hours a night. Going to use the same bootstrapping and success parameters used above.

```
sleep_8hrs <- yrbss %>%
                  filter(!is.na(school_night_hours_sleep)) %>%  # First, want to filter out NA values a
                  mutate(sleep_enough = ifelse(school_night_hours_sleep=="8", "yes", "no")) %>%
                  specify(response = sleep_enough, success = "yes") %>%
                  generate(reps = 1000, type = "bootstrap") %>%
                  calculate(stat = "prop") %>%
                  get_ci(level = 0.95)

sleep_8hrs
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.211    0.226
```

This means we can be 95% confident the true proportion of teens that sleep 8 hours a night is in the range $(0.211, 0.225)$. This gives us a margin of error of 0.007

Also want to construct a confidence interval for our `hours_tv_per_school_day` to get an idea of the proportion of teens watching more than 5 hours of tv a day.

```
tv_over_5hrs <- yrbss %>%
        filter(!is.na(hours_tv_per_school_day)) %>%  # Filtering NAs again
        mutate(too_much_tv = ifelse(hours_tv_per_school_day=="5+", "yes", "no")) %>%
        specify(response = too_much_tv, success = "yes") %>%
        generate(reps = 1000, type = "bootstrap") %>%
        calculate(stat = "prop") %>%
        get_ci(level = 0.95)

tv_over_5hrs
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.115    0.126
```

We can be 95% confident the true proportion of teens who watch over 5 hours of TV per school day is within the range $(0.115, 0.126)$. This gives us a margin of error of 0.0055.

How does the proportion affect the margin of error?

Imagine you've set out to survey 1000 people on two questions: are you at least 6-feet tall? 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:

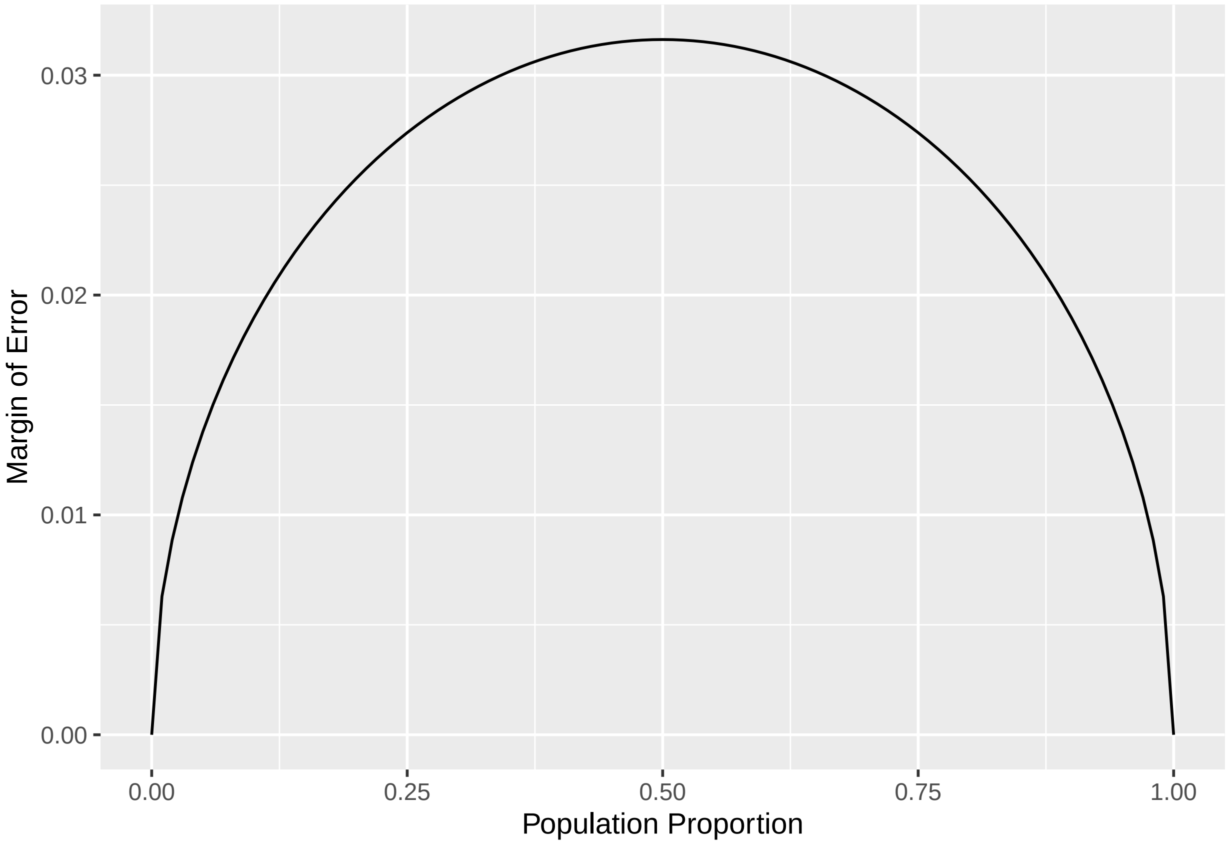
```
n <- 1000
```

The first step is to make a variable `p` that is a sequence from 0 to 1 with each number incremented by 0.01. You can then create a variable of the margin of error (`me`) associated with each of these values of `p` using the familiar approximate formula ($ME = 2 \times SE$).

```
p <- seq(from = 0, to = 1, by = 0.01)
me <- 2 * sqrt(p * (1 - p)/n)
```

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

```
dd <- data.frame(p = p, me = me)
ggplot(data = dd, aes(x = p, y = me)) +
  geom_line() +
  labs(x = "Population Proportion", y = "Margin of Error")
```



5. 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? Margin of error increases to a maximum at $p = 0.5$, and then decreases as `p` tends towards 1. The margin of error is maximized at a population proportion of 0.5.

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 $np \geq 10$ and $n(1-p) \geq 10$. This rule of thumb is easy enough to follow, but it makes you wonder: what's so special about the number 10?

The short answer is: nothing. You could argue that you would be fine with 9 or that you 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.

You can investigate the interplay between n and p and the shape of the sampling distribution by using simulations. Play around with the following app to investigate how the shape, center, and spread of the distribution of $\hat{p}$ changes as n and p changes.

6. Describe the sampling distribution of sample proportions at $n = 300$ and $p = 0.1$. Be sure to note the center, spread, and shape. At these parameter values, the distribution appears to be relatively normal (not skewed) and centered at $p = 0.1$ with a spread of about 0.03.

7. Keep n constant and change p. How does the shape, center, and spread of the sampling distribution vary as p changes. You might want to adjust min and max for the x-axis for a better view of the distribution. With n constant, p tends to still be non-skewed and centered at the selected value. The biggest change in the sampling distribution as p changes is the spread. Closer to 0 or 1, the distribution

is tighter (smaller spread). In the middle at $p = 0.5$, the distribution tends to be wider. This tracks with the me - p graph we plotted in `6`, where the margin of error peaks at $p \approx 0.5$.

8. Now also change n. How does n appear to affect the distribution of $\hat{p}$? As n increases, the sampling distribution remains centered in the same place and non-skewness. However, the spread of the sampling distribution decreases, and the distribution appears skinnier. This is because with larger sample sizes, more samples will have sample parameters closer to the true population proportion. * * *

More Practice

For some of the exercises below, you will conduct inference comparing two proportions. In such cases, you have a response variable that is categorical, and an explanatory variable that is also categorical, and you are comparing the proportions of success of the response variable across the levels of the explanatory variable. This means that when using `infer`, you need to include both variables within `specify`.

9. Is there convincing evidence that those who sleep 10+ hours per day are more likely to strength train every day of the week? 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.

We begin by defining our null and alternative hypotheses (H_0 and H_a, respectively)

- H_0: those who sleep 10+ hours per day are *no more likely* to strength train every day of the week
- H_a: those who sleep 10+ hours per day *are more likely* to strength train every day of the week

First, we'll construct a confidence interval of the sample proportion of those who train 7 days a week

```
yrbss %>%
        filter(!is.na(strength_training_7d)) %>%
        mutate(strength_ind = ifelse(strength_training_7d==7, "yes", "no")) %>%
        specify(response = strength_ind, success = "yes") %>%
        generate(reps = 1000, type = "bootstrap") %>%
        calculate(stat = "prop") %>%
        get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.162    0.175
```

We can be 95% confident that the true proportion of people that strength train 7 days a week is within $(0.162, 0.175)$. In order to see convincing evidence that sleeping 10+ hours a night changes the likelihood someone strength trains, we'll need to calculate the sample proportion of people that strength train 7 days a week, given that they sleep 10 hours

```
samp <- yrbss %>%
        filter(!is.na(school_night_hours_sleep), !is.na(strength_training_7d)) %>%
        mutate(sleep_ind = ifelse(school_night_hours_sleep == "10+", "yes", "no"),
               strength_ind = ifelse(strength_training_7d==7, "yes", "no")
               )

p_hat <- samp %>%
            filter(strength_ind == "yes") %>%
            summarise(p_hat = n() / nrow(samp))



p_hat
```

```
## # A tibble: 1 x 1
##    p_hat
##    <dbl>
## 1 0.167
```

Since our sample proportion $\hat{p} = 0.167$ is within our confidence interval of $(0.162, 0.175)$, we can not reject the null hypothesis that people who sleep more than 10 hours are not more likely to strength train 7 days a week.

10. Let's say there has been no difference in likeliness to strength train every day of the week for those who sleep 10+ hours. What is the probability that you could detect a change (at a significance level of 0.05) simply by chance? *Hint:* Review the definition of the Type 1 error.

Type 1 error - rejecting the null hypothesis when the null hypothesis is actually true. If we detect a change (at a significance level $\alpha = 0.05$), that means that we would be rejecting the null hypothesis that there is no change. This would correspond to a probability of 5%, because we would be detecting a 5% or less probability of getting a result at least as extreme as the one that leads us to rejecting the null hypothesis mistakenly.

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 a dataset.

Looking at our ME - p plot, the margin of error is at 1% $ME \leq 0.01$ when $p <0.05 $ or $p > 0.095$ (roughly speaking). Let's assume a population proportion of $p = 0.4$ for our calculation. Since we're using a 95% confidence interval, we can use a value of $z^* = 1.96$.

$ME = 1.96 * SE$ where $SE = \sqrt{\frac{p(1-p)}{n}}$

Rearranging, we get:

$n = \frac{p(1-p)}{(\frac{ME}{z^*})^2}$

Plugging in our values, we will allow a max value of the margin of error of 0.01 with $p = 0.4$ and $z^* = 1.96$

$n = \frac{0.4(1-0.4)}{(\frac{0.01}{1.96})^2}$

And calculating we get a sample size of

```r
samp_size <- (0.4 * (1 - 0.4)) / ((0.01 / 1.96) **2)
samp_size
```

```
## [1] 9219.84
```

$n = 9220$
