Inference for categorical data

Kayleah Griffen

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?

The counts for each category of the amount of days students have texted while driving within the past 30 days is below.

```
data('yrbss', package='openintro')
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")
```

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"))
```

The proportion of people who have texted while driving every day in the past 30 days and never wear helmets can be calculated by first calculating the total people who text every day while driving and do not wear a helmet.

```
text30_no_helmet <- no_helmet |>
  filter(text_ind == "yes") |>
  count(text_ind)

text30_no_helmet$n
```

```
## [1] 463
```

This is 463 people.

Then we can get the total number of people

```
total_people <- dim(yrbss)[1]
total_people
```

```
## [1] 13583
```

Total there were 13,583 people surveyed.

Now we can calculate the proportion of people who text every day and never where a helmet.

```
pt30nh <- text30_no_helmet$n/total_people
pt30nh
```

```
## [1] 0.03408673
```

The proportion of people who have texted while driving every day in the past 30 days and never wear helmets is 0.034 or 3.4%.

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.

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

no_helmet %>% filter(text_ind != "unknown") |>
  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   0.0655   0.0781
```

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 is the square root of the sample proportion times one minus the sample proportion, divided by the number of people in the population - all of this times a value for z, which depends on the confidence level. For 0.95 confidence the z is 1.96.

```
me <- 1.96*sqrt((pt30nh*(1-pt30nh))/total_people)
me
```

```
## [1] 0.003051546
```

The margin of error in the case is 0.003. Keep in mind for the confidence interval calculation NAs were removed, for my margin of error calculation NAs were not filtered out. NAs were treated as failures.

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. Interpret the interval in context of the data. It may be helpful to create new data sets for each of the two variables first, and then use these data sets to construct the confidence intervals.

Take a look at two of the other categorical variables.

```
yrbss |> count(physically_active_7d)
```

```
## # A tibble: 9 x 2
##   physically_active_7d     n
##                  <int> <int>
## 1                    0  2172
## 2                    1   962
## 3                    2  1270
## 4                    3  1451
## 5                    4  1265
## 6                    5  1728
## 7                    6   840
## 8                    7  3622
## 9                   NA   273
```

```
yrbss |> count(hours_tv_per_school_day)
```

```
## # A tibble: 8 x 2
##   hours_tv_per_school_day     n
##   <chr>                   <int>
## 1 1                        1750
## 2 2                        2705
## 3 3                        2139
## 4 4                        1048
## 5 5+                       1595
## 6 <1                       2168
## 7 do not watch             1840
## 8 <NA>                      338
```

Get the confidence intervals for very active.

```
very_active <- yrbss %>%
  mutate(active = replace_na(ifelse(physically_active_7d == "7", "yes", "no"), "unknown"))

va_ci <- very_active %>% filter(active != "unknown") |>
  specify(response = active, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)
va_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.265    0.280
```

The lower CI is 0.265 and upper CI is 0.279.

Calculate the margin of error for very active.

```r
very_active_count <- very_active |>
  filter(active == "yes") |>
  count(active)

very_active_proportion <- very_active_count$n/total_people

va_me <- 1.96*sqrt((very_active_proportion*(1-very_active_proportion))/total_people)
va_me
```

```
## [1] 0.007436836
```

The margin of error is 0.007.

Get the confidence intervals for no tv.

```r
no_tv <- yrbss %>%
  mutate(no_tv = replace_na(ifelse(hours_tv_per_school_day == "do not watch", "yes", "no"), "unknown"))

nt_ci <- no_tv %>% filter(no_tv != "unknown") |>
  specify(response = no_tv, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)
nt_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.133    0.145
```

The lower CI is 0.133 and upper CI is 0.145.

Calculate the margin of error for no tv.

```r
no_tv_count <- no_tv |>
  filter(no_tv == "yes") |>
  count(no_tv)

no_tv_proportion <- no_tv_count$n/total_people

nt_me <- 1.96*sqrt((no_tv_proportion*(1-no_tv_proportion))/total_people)
nt_me
```

```
## [1] 0.005755207
```

The margin of error is 0.0057.

In both cases, the computed margins of error make sense given the confidence intervals. The reason being, to get the upper and lower confidence interval you plus and minus the margin of error from the point estimate. So a quick check is if the margin of error is half of the distance between the upper and lower CI. And in both cases it is approximately true. Differences is likely regarding the null values.

```r
(va_ci$upper_ci - va_ci$lower_ci)/2
```

```
## [1] 0.007551653
```

```r
va_me
```

```
## [1] 0.007436836
```

```r
(nt_ci$upper_ci - nt_ci$lower_ci)/2
```

```
## [1] 0.005929596
```

```r
nt_me
```

```
## [1] 0.005755207
```

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:

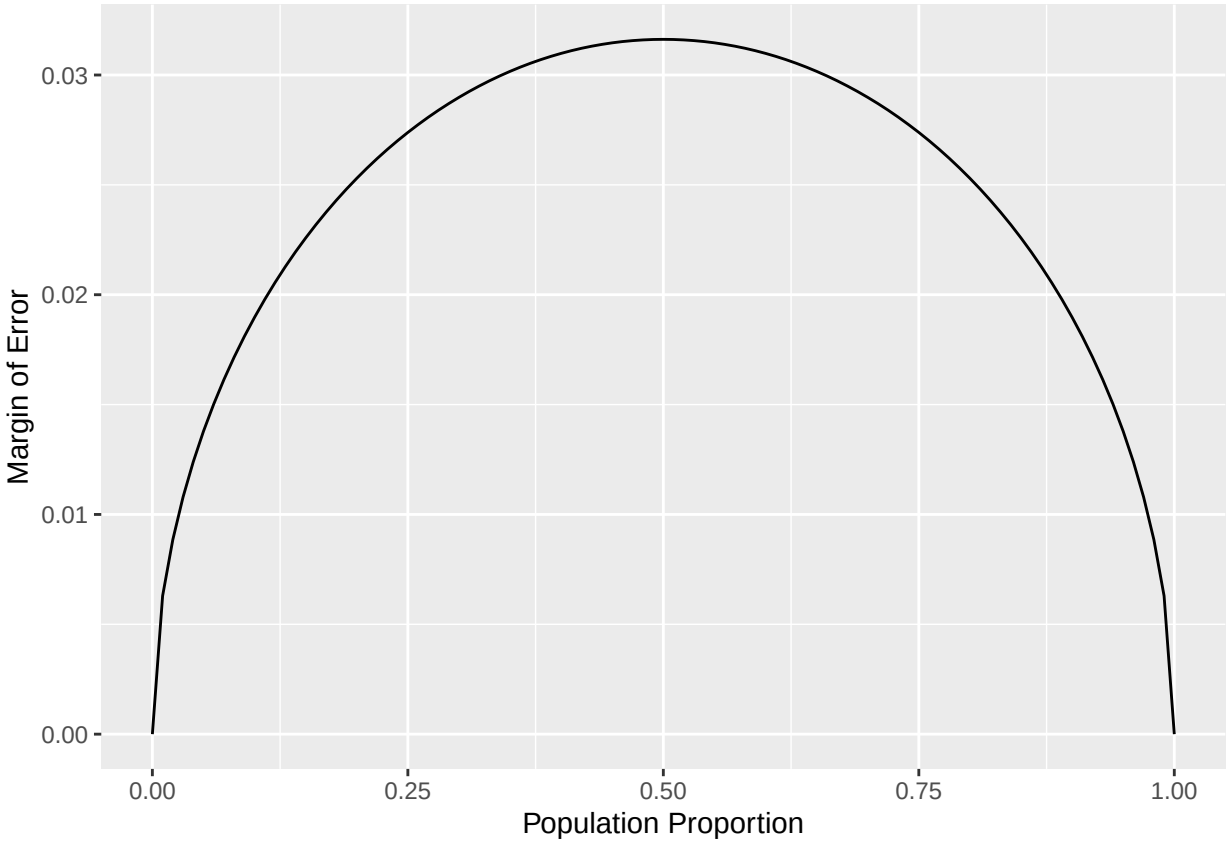
```r
n <- 1000
```

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 ($ME = 2 \times SE$).

```r
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.

```r
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?

The plot is above. There is a parabolic relationship between p and ME. At first as population proportion increases, margin of error increases - up until a critical point where the maximum margin of error occurs when population proportion is 0.5. After this as population proportion increases, margin of error decreases. For a given sample size - at p =0.5 the margin of error is maximized.

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.

The sampling distribution of sample proportions at n=300 and p = 0.1 is nearly normal. It's center is about 0.1, its spread goes from about 0.05 to 0.15, and its shape is basically normal.

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.

Keeping n constant and changing p, when p increases the shape stays basically normal, the center moves to whatever the p values is, and the spread is a maximum at p = 0.5 and decreases as you get closer to 0 or closer to 1.

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

A smaller n results in a larger spread, a greater n results in a narrower spread - so as n increases the SE decreases. The greater the n the more the distribution looks like a normal distribution - that is taller and narrower.

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.

The null hypothesis is that there IS NOT evidence that those who sleep 10+ hours per day are more likely to strength train every day of the week.

The alternative hypothesis is that there IS convincing evidence that those who sleep 10+ hours per day are more likely to strength train every day of the week.

```
yrbss_modified <- yrbss %>%
  mutate(sleep10 = replace_na(ifelse(school_night_hours_sleep == "10+", "yes", "no"), "unknown"),
         strength7 = replace_na(ifelse(strength_training_7d == "7", "yes", "no"), "unknown"))

ss_ci <- yrbss_modified %>% filter(sleep10 != "unknown" & strength7 !=  "unknown") |>
  specify(response = strength7, explanatory = sleep10, success = "yes")  |>
  generate(reps = 1000, type = "bootstrap")  |>
  calculate(stat = "diff in props", order = c("no", "yes")) |>
  get_ci(level = 0.95)
ss_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   -0.153  -0.0588
```

The null hypothesis was rejected because 0 is not in the range of the confidence interval.

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 is rejecting the null hypothesis when the null hypothesis is actually true. Type 2 error is failing to reject the null hypothesis when the alternative is actually true. In my case, the null hypothesis is there IS NOT a difference in likeliness to strength train every day for those who sleep 10+ hours. Type 1 error would be rejecting this null hypothesis when it is true. If there is no difference in likeliness to strength train every day of the week for those who sleep 10+ hours, the probability that you could detect a change - at a significance level of 0.05 - means you are using a 95% confidence interval. In this case, an error will happen whenever the point estimate is 1.96 standard errors away from the population parameter. This occurs about 5% of the time (2.5% for each tail).

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.

If you have no idea what to expect for p, you should assume that p = 0.5 because this is where the margin of error is maximized. For a margin of error no greater than 1% with 95% confidence you can then use the formula for margin of error but solve for n.

The margin of error (0.01) is the square root of the sample proportion (0.5) times one minus the sample proportion, divided by the number of people in the population - all of this times a value for z, which depends on the confidence level. For 0.95 confidence the z is 1.96.

So rearranging the formula n = p(1-p)/ ((ME/Z)^2)

```
n <- ((0.5)*(1-0.5))/((0.01/1.96)^2)
n
```

```
## [1] 9604
```

The number of people you would have to sample to ensure you are within the guidelines is 9604.
