Inference for categorical data

Victor H Torres

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`.

```
yrbss %>%
  count(text_while_driving_30d, sort=TRUE)
```

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

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 count is displayed with the method above: 4792 students did not text and drive. 4646 did not drive, etc.

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

```
no_fear <- yrbss %>%
  filter(helmet_12m=="never") %>%
  filter(!is.na(text_while_driving_30d)) %>%
  mutate(text_ind_everyday = ifelse(text_while_driving_30d == "30", "yes", "no"))

no_fear %>%
  count(text_ind_everyday)
```

```
## # A tibble: 2 x 2
##   text_ind_everyday     n
##   <chr>             <int>
## 1 no                 6040
## 2 yes                 463
```

The total of people who have texted while driving and never wore helmets is 464 which is a proportion of 7.66%. 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"))
```

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 %>%
  drop_na(text_ind) %>% # Drop missing values
  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   0.0778
```

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?

```
no_fear %>%
 specify(response = text_ind_everyday, 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   0.0650   0.0774
```

The margin of error ranges between 6.5% and 7.7% for the estimate of the proportion of non-helmet wearers.

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.

```
glimpse(yrbss)
```

```
## Rows: 13,583
## Columns: 13
## $ age                      <int> 14, 14, 15, 15, 15, 15, 15, 14, 15, 15, 15, 1~
## $ gender                   <chr> "female", "female", "female", "female", "fema~
## $ grade                    <chr> "9", "9", "9", "9", "9", "9", "9", "9", "9", ~
## $ hispanic                 <chr> "not", "not", "hispanic", "not", "not", "not"~
## $ race                     <chr> "Black or African American", "Black or Africa~
## $ height                   <dbl> NA, NA, 1.73, 1.60, 1.50, 1.57, 1.65, 1.88, 1~
## $ weight                   <dbl> NA, NA, 84.37, 55.79, 46.72, 67.13, 131.54, 7~
## $ helmet_12m               <chr> "never", "never", "never", "never", "did not ~
## $ text_while_driving_30d   <chr> "0", NA, "30", "0", "did not drive", "did not~
## $ physically_active_7d     <int> 4, 2, 7, 0, 2, 1, 4, 4, 5, 0, 0, 0, 4, 7, 7, ~
## $ hours_tv_per_school_day  <chr> "5+", "5+", "5+", "2", "3", "5+", "5+", "5+",~
## $ strength_training_7d     <int> 0, 0, 0, 0, 1, 0, 2, 0, 3, 0, 3, 0, 0, 7, 7, ~
## $ school_night_hours_sleep <chr> "8", "6", "<5", "6", "9", "8", "9", "6", "<5"~
```

```
unique(yrbss$physically_active_7d)
```

```
## [1]  4  2  7  0  1  5  3 NA  6
```

```r
forza <- yrbss |>
  filter(!is.na(physically_active_7d)) |>
  mutate(very_active = ifelse(physically_active_7d %in% c("1","2","3"), "yes", "no"))

forza |>
  count(very_active) |>
  mutate(p = n / sum(n))
```

```
## # A tibble: 2 x 3
##   very_active     n     p
##   <chr>       <int> <dbl>
## 1 no           9627 0.723
## 2 yes          3683 0.277
```

```r
forza |>
  specify(response = very_active, 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    0.269    0.284
```

The 95% confidence interval for the probability physically active in 7 days is between 0.269 and 0.284.

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:

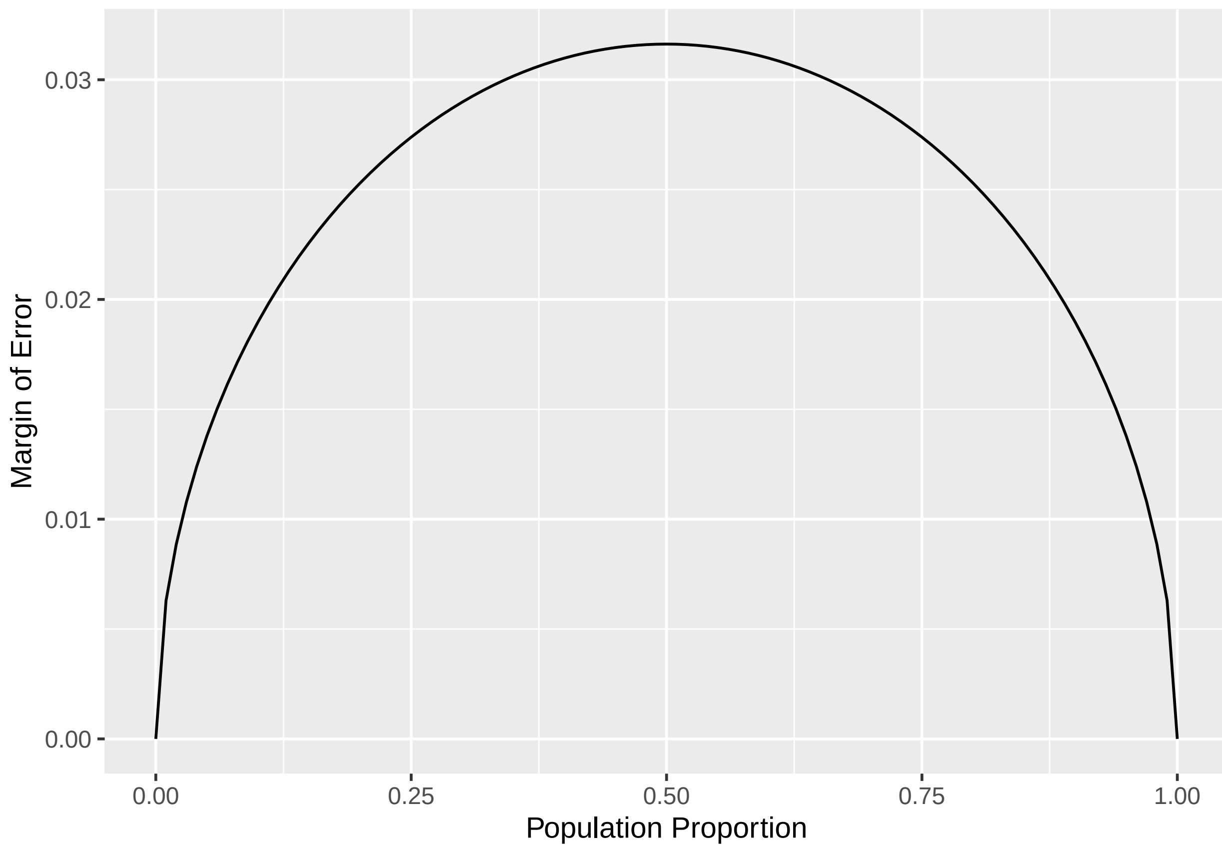
```r
n <- 1000
```

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?

```r
n <- 1000
p <- seq(from = 0, to = 1, by = 0.01)
me <- 2 * sqrt(p * (1 - p)/n)
```

We can se that the margin of error increases when the population proportion do the same.the margin of error reach the highest peak when the population reaches 50%..5 x .5 is the greatest that the numerator could ever be when calculating standard error.

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.

It seems that the center is at 0.10, the spread at 0.017, and the shape is at a normal distribution bell curve.

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.

When I change the P value to a smaller number, the data seems to maintain its normally distributed shape, however, it moves along the x-axis.

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

When I change N, the distribution of the data appears to maintain normality, and the center seems to be moving to the right.

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.

```
table(yrbss$strength_training_7d)
```

```
##
##    0    1    2    3    4    5    6    7
## 3632 1012 1305 1468 1059 1333  513 2085
```

```
workout <- yrbss |>
  filter(!is.na(strength_training_7d)) |>
  mutate(daily = ifelse(strength_training_7d == "7", "yes", "no"))

workout |>
  count(daily) |>
  mutate(p = n / sum(n))
```

```
## # A tibble: 2 x 3
##   daily     n     p
##   <chr> <int> <dbl>
## 1 no    10322 0.832
## 2 yes    2085 0.168
```

```
workout |>
 specify(response = daily, 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    0.161    0.174
```

```
table(yrbss$school_night_hours_sleep)
```

```
##
##   <5  10+    5    6    7    8    9
##  965  316 1480 2658 3461 2692  763
```

```
  hours_sleep <- yrbss |>
  filter(!is.na(school_night_hours_sleep)) |>
  mutate(ten_plus = ifelse(school_night_hours_sleep == "10+", "yes", "no"))

hours_sleep |>
  count(ten_plus) |>
  mutate(p = n / sum(n))
```

```
## # A tibble: 2 x 3
##   ten_plus     n      p
##   <chr>    <int>  <dbl>
## 1 no       12019 0.974
## 2 yes        316 0.0256
```

```
hours_sleep |>
 specify(response = ten_plus, 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   0.0229   0.0287
```

There is no evidence that those who sleep 10+ hours per day are more likely to strength train every day of the week.An alternative hypothesis can display evidence that those who sleep 10+ hours per day are more likely to strength train every day of the 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 probablity 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.

I can say that, there is a 95% confident level that the true proportion of those who sleep 10+ hours daily, that also strength train every day of the week it rages between 22.12% and 32.05%.

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.

Personally, I will select a sample of 0.10 or less than 10% of the local population.
