Inference for categorical data

Richie Rivera

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.

```
set.seed(1994)
library(tidyverse)
library(openintro)
library(infer)
library(ggplot2)
```

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?

Insert your answer here

```
data("yrbss", package = "openintro")
#?yrbss

text_count <- yrbss |>
  group_by(text_while_driving_30d) |>
  summarise(observations = n())

text_count
```

```
## # A tibble: 9 x 2
##   text_while_driving_30d observations
##   <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
```

We can see here that a great deal of the students (827) have texted while driving everyday. That's no good!

End of your answer

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

Insert your answer here

```r
sample_pop <- yrbss |>
  summarise(total_observations = n())

risky_df <- yrbss |>
  filter(text_while_driving_30d == 30) |>
  filter(helmet_12m == "never")

risky_people <- risky_df |>
  summarise(risky_people = n())

daily_driving_texting <- yrbss |>
  filter(text_while_driving_30d == 30) |>
  summarise(daily_texters = n())

print(
  paste(
    round(100 * risky_people / daily_driving_texting, 2),
    "% of the sample population who text everyday also never wear helmets.",
    sep = ""
  )
)
```

```
## [1] "55.99% of the sample population who text everyday also never wear helmets."
```

```r
print(
  paste(
    round(100 * risky_people / sample_pop, 2),
    "% of the entire sample population text everyday and also never wear helmets."
  )
)
```

```
## [1] "3.41 % of the entire sample population text everyday and also never wear helmets."
```

It's very unlikely that someone texts everyday and never wears a helmet across the entire sample but if someone texts everyday then there is a much higher chance that they also never wear a helmet.

End of your answer

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   0.0652   0.0772
```

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?

Insert your answer here

$$MOE = Z * \sqrt{\frac{p(1-p)}{n}}$$

with that formula:

```r
z <- 1.96 # For 95% confidence
p <- nrow(risky_df) / nrow(yrbss)

moe <- z * sqrt(
  (p * (1 - p)) / nrow(yrbss)
)

moe
```

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

From this, we can see that the margin of error is 0.0031.

End of your answer

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.

Insert your answer here

What proportion of youths who text everyday are female?

First we'll look at male vs female where being a Female is marked as a success:

```r
daily_texters <- yrbss |>
  filter(text_while_driving_30d == 30)

daily_texters %>%
  drop_na(gender) %>% # Drop missing values
  specify(response = gender, success = "female") %>%
  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    0.399    0.468
```

From here we can see that we're 95% confident that the proportion of women youths who text and drive everyday is between 39.9% and 46.8%.

Secondly, let's see the proportion of youths who do not watch tv and are hispanic. In this, being hispanic is considered a successs:

```r
no_tv <- yrbss |>
  filter(hours_tv_per_school_day == "do not watch")

no_tv |>
  drop_na(hispanic) |>
  specify(response = hispanic, success = "hispanic") |>
  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    0.234    0.271
```

In this exmaple, we can see that we're 95% confident that the proportion of hispanic youths who watch no tv during the school week is between 23.3% and 27.3%

End of your answer

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:

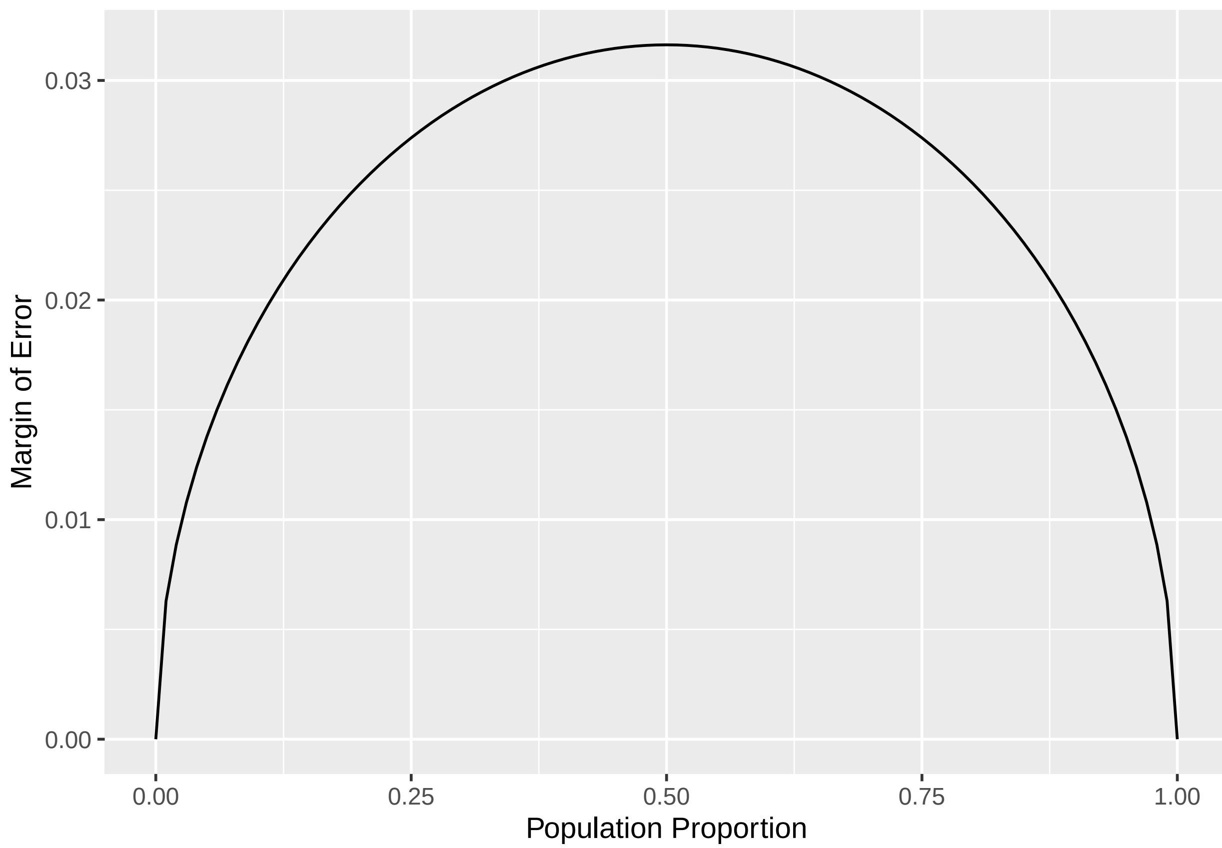
```
n <- 1000
```

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?

Insert your answer here

`me` seems to correlate positively with `p` up until 0.5. After 0.5 the trend reverses and `me` is inversely correlated with `p`.

End of your answer

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.

Insert your answer here

The histogram plotted seems normal centered around a p of 0.1 with a tight spreadthat appears to be mostly around $p \pm .05$

End of your answer

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.

Insert your answer here

As we saw above for when $p = 0.5$, the margin of error is greatest when $p = 0.5$, which we can visually see using the app. As we approach $p = 0.5$ we see that the spread incresases and then decreases as we move away from it.

End of your answer

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

Insert your answer here

As we increase n, we can see that the spread decreases, which makes sense as margin of error has an inverse relationship with n.

End of your answer

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.

Insert your answer here

For this, our H_0 is that those who sleep 10+ hours per day are equally likely to strength train every day of the week as the rest of the population.

```
no_sleepers <- yrbss |>
  drop_na(school_night_hours_sleep) |>
  drop_na(strength_training_7d) |>
  filter(school_night_hours_sleep != "10+") |>
```

```
  mutate(daily_workout = ifelse(strength_training_7d == 7, "yes", "no")) |>
  select(school_night_hours_sleep, daily_workout)

deep_sleepers <- yrbss |>
  drop_na(school_night_hours_sleep) |>
  drop_na(strength_training_7d) |>
  filter(school_night_hours_sleep == "10+") |>
  mutate(daily_workout = ifelse(strength_training_7d == 7, "yes", "no")) |>
  select(school_night_hours_sleep, daily_workout)

no_sleep_result <- no_sleepers |>
  specify(response = daily_workout, success = "yes") |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "prop") |>
  get_ci(level = 0.95)

nosleep_lci <- no_sleep_result$lower_ci
nosleep_uci <- no_sleep_result$upper_ci
```

For the population who do not sleep for 10+ hours a night, we are 95% confident that the proportion who also workout everyday is between 15.8% and 17.1%.

```
deep_sleep_result <- deep_sleepers |>
  specify(response = daily_workout, success = "yes") |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "prop") |>
  get_ci(level = 0.95)

deepsleep_lci <- deep_sleep_result$lower_ci
deepsleep_uci <- deep_sleep_result$upper_ci
```

For the group of youths who do sleep for 10+ hours, we are 95% confident that the proportion who also workout everyday is between 22.4% and 32.1%.

With these results, we have evidence to reject the null hypothesis as we are 95% confident that the group who sleeps 10+ hours a day are 5.3% more likely to strength train.

End of your answer

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.

Insert your answer here

By definition, with a significance level of 0.05 we will have a 5% chance to detect a change simply by chance.

End of your answer

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.

Insert your answer here

Using the formula:

$$MOE = Z * \sqrt{\frac{p(1-p)}{n}}$$

and acknowledging that moe is maximum when $p = 0.5$, we can solve for n by:

$$.01 = (1.96) * \sqrt{\frac{(0.5)(1-(0.5))}{n}}$$

Solving for n above, we get that the sample size will need to be at least 9604.

End of your answer