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?

Insert your answer here

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

4792 have reported 0 days. 4646 have reported they did not drive. 925 have reported drive 1 to 2 days. 827 have reported 30 days. 493 have reported 3 to 5 days. 373 have reported 10 to 19 days. 311 have reported 6 to 9 days. 298 have reported 20 to 29 days. 918 have not answered the question.

```
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
```

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

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

```r
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`.

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

```r
no_helmet %>%
  count(text_ind)
```

```
## # A tibble: 3 x 2
##   text_ind     n
##   <chr>    <int>
## 1 no        6040
## 2 yes        463
## 3 <NA>       474
```

The proportion of students who texted while not wearing a helmet was 463/6503 (not counting any NAs).

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.

```r
set.seed(1125)
no_helmet %>%
  filter(!is.na(text_ind)) %>%
  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   0.0650   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?

Insert your answer here

The lower ci in my bootstrap was 0.065 and the upper ci was 0.077. That means we are 95% certain that the proportion of non-helmet wearers who have texted while driving in the past 30 days is within that range. The margin of error is (0.077-0.065)/2=0.006.

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

```r
set.seed(1125)
does_not_watch_tv <- yrbss |>
  mutate(does_not_watch_tv = ifelse(hours_tv_per_school_day  =="do not watch", "yes", "no"))

does_not_watch_tv |>
  dplyr::filter(!is.na(does_not_watch_tv)) |>
  specify(response = does_not_watch_tv, success = "yes") |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "prop") |>
  get_ci(level = 0.99)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.132    0.147
```

This means that we are 99% confident that the proportion of students who do not watch tv is in the range of 0.132-0.147, so 13.2%-14.7%.

```r
set.seed(1125)
no_work_out <- yrbss |>
  mutate(does_not_work_out = ifelse(strength_training_7d ==0, "yes", "no"))

no_work_out |>
  filter(!is.na(does_not_work_out)) |>
  specify(response = does_not_work_out, success = "yes") |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "prop") |>
  get_ci(level = 0.90)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.287    0.299
```

This means that we are 90% confident that the proportion of students who do not do any strength training is in the range of 0.287-0.299, so 28.7%-29.9%.

Since the second example is a lower confidence level the range can be much narrower as opposed to the first example.

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:

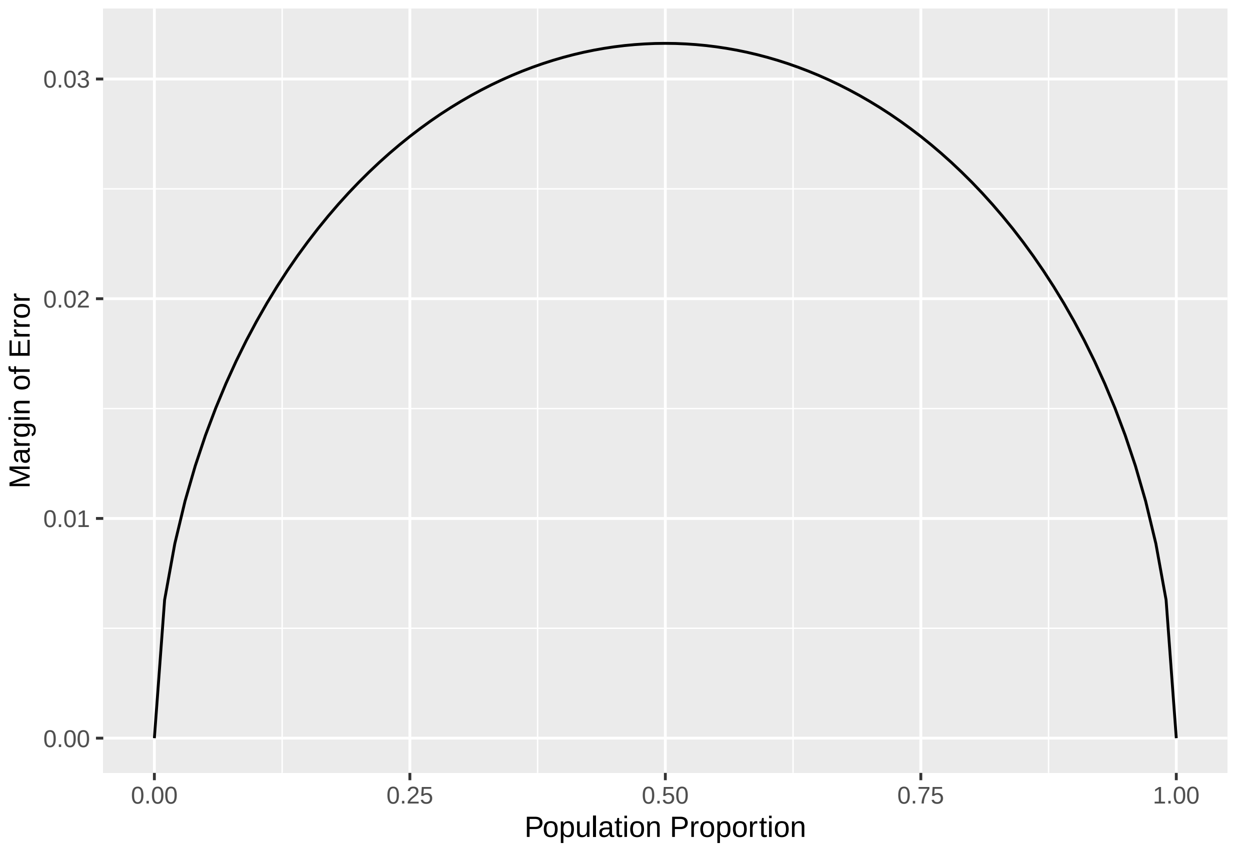
```
n <- 1000
```

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

The relationship between proportion (p) and Margin of Error (me), is parabolic with a maximum me at 0.50. The reason foe the points equidistant from 0.50 having the same me is due to it being a binomial variable and the two sides are always reverses of each other. So in essence, they are the same regardless pf which side you are looking at at that point.

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

At the given variables, the distribution was normal and almost a perfect bell curve with the center very left skewed at 0.10

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

At any given n, the greater the p, the more right skewed the distribution (and since it's a normal distribution, the center) becomes. Additionally, as noted above, the closer the p is to 0.50 the wider the spread is and points equidistant from 0.50 have the same spread.

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

Insert your answer here

As you increase the n, the distribution becomes more normal. This is due to the CLT. However, once you are dealing with a large enough n the difference becomes much less notiecable. Increasing n also decreases the width of the spread due to larger proportions having less variability between each other.

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

```
set.seed(1125)
sleepx10 <- yrbss %>%
  filter(school_night_hours_sleep != "10+")

sleepx10 %>%
  mutate(physical = ifelse(physically_active_7d == 7, "yes", "no")) %>%
  drop_na(physical) %>%
  specify(response = physical, 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    0.261    0.276
```

```
sleep10 <- yrbss %>%
  filter(school_night_hours_sleep == "10+")

sleep10 %>%
  mutate(physical = ifelse(physically_active_7d == 7, "yes", "no")) %>%
  drop_na(physical) %>%
  specify(response = physical, 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    0.316    0.422
```

There does indeed seem to be a correlation between sleeping 10 hours and working out every day. The confidence intervals for people not sleeping 10+ hours a night is 0.261-0.276 while the confidence intervals for students who were sleeping 10+ hours a night is 0.316-0.422. That means that we are 95% confident that the proportion of students sleeping 10+ hours every night and also working out every day is somewhere between 32% and 42%, while the proportion of people who don't sleep 10+ hours to be working out every day is 26% to 28%. Since these don't overlap at all this implies that there is indeed a relationship between sleeping 10+ hours and working out every day.

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

The probability of getting a false positive would be 5%.

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

```
me <- 0.01
p <- 0.5

(1.96^2 * p * (1 - p))/me^2
```

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

The total amount we would need to sample would be 9,604.