Inference for categorical data

Alex Ptacek

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)
```

```
set.seed(606)
```

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

```
#Counts of each category of `text_while_driving_30d`
yrbss |>
  group_by(text_while_driving_30d) |>
  count() |>
  arrange(desc(n))
```

```
## # A tibble: 9 x 2
## # Groups:   text_while_driving_30d [9]
##   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

Answer: The proportion of people who have texted while driving every day in the past 30 days and never wear a helmet is 3.4%.

```
yrbss |>
  group_by(helmet_12m, text_while_driving_30d) |>
  summarise(n = n(), .groups = "drop") |>
  mutate(prop = round(n/sum(n), 3)) |>
  filter(helmet_12m == "never" & text_while_driving_30d == "30")
```

```
## # A tibble: 1 x 4
##   helmet_12m text_while_driving_30d     n  prop
##   <chr>      <chr>                   <int> <dbl>
## 1 never      30                        463 0.034
```

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   0.0650   0.0775
```

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

Answer: At a confidence interval (6.5%, 7.8%), the margin of error is 0.65%.

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

Answer: We can have 95% confidence that, if we take a sample from yrbss, the proportion of males in that sample would be between (50.3%, 52.1%) and the proportion of people who watch TV 5+ hours on a school day would be between (11.5%, 12.6%). Additionally, if `yrbss` is a random sample from the population, we can use those same intervals as an estimate for the population proportions for the respective variable with 95% confidence.

```
#Create new dataset that confirms if male or not
gender_male <- yrbss |>
  mutate(male_check = ifelse(gender == "male", "yes", "no"))

#Calculate a 95% confidence interval for the proportion of males
gender_male |>
  drop_na(male_check) |>
  specify(response = male_check, success = "yes") |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "prop") |>
  get_ci(level = .95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.503    0.521
```

```
#Create new data set that classifies people who watch 5+ hours of tv
five_plus_tv <- yrbss |>
  mutate(heavy_tv_usage = ifelse(hours_tv_per_school_day == "5+", "yes", "no"))

#Calculate 95% confidence interval for people who watch 5+ hours of tv
five_plus_tv |>
```

```
  drop_na(heavy_tv_usage) |>
  specify(response = heavy_tv_usage, success = "yes") |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "prop") |>
  get_ci(level = .95)
```

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

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:

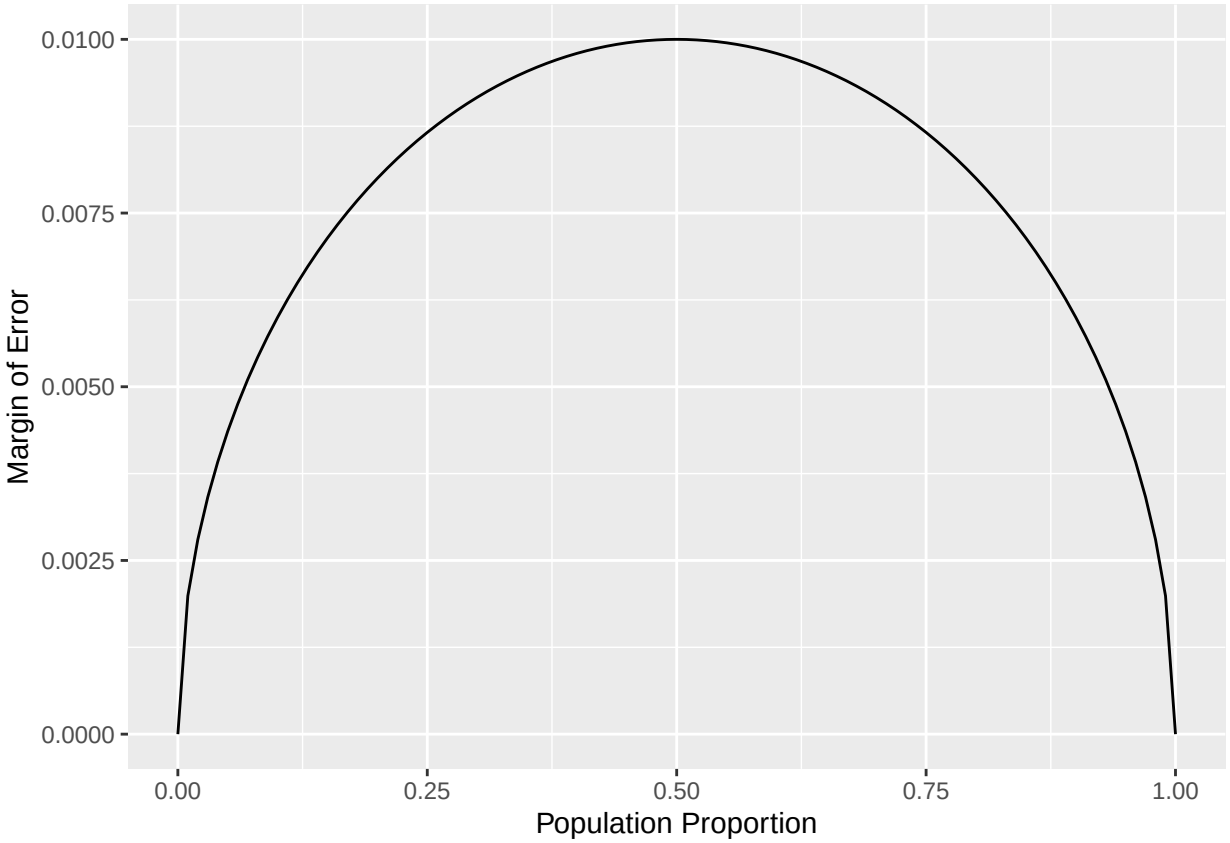
```
n <- 10000
```

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")
```

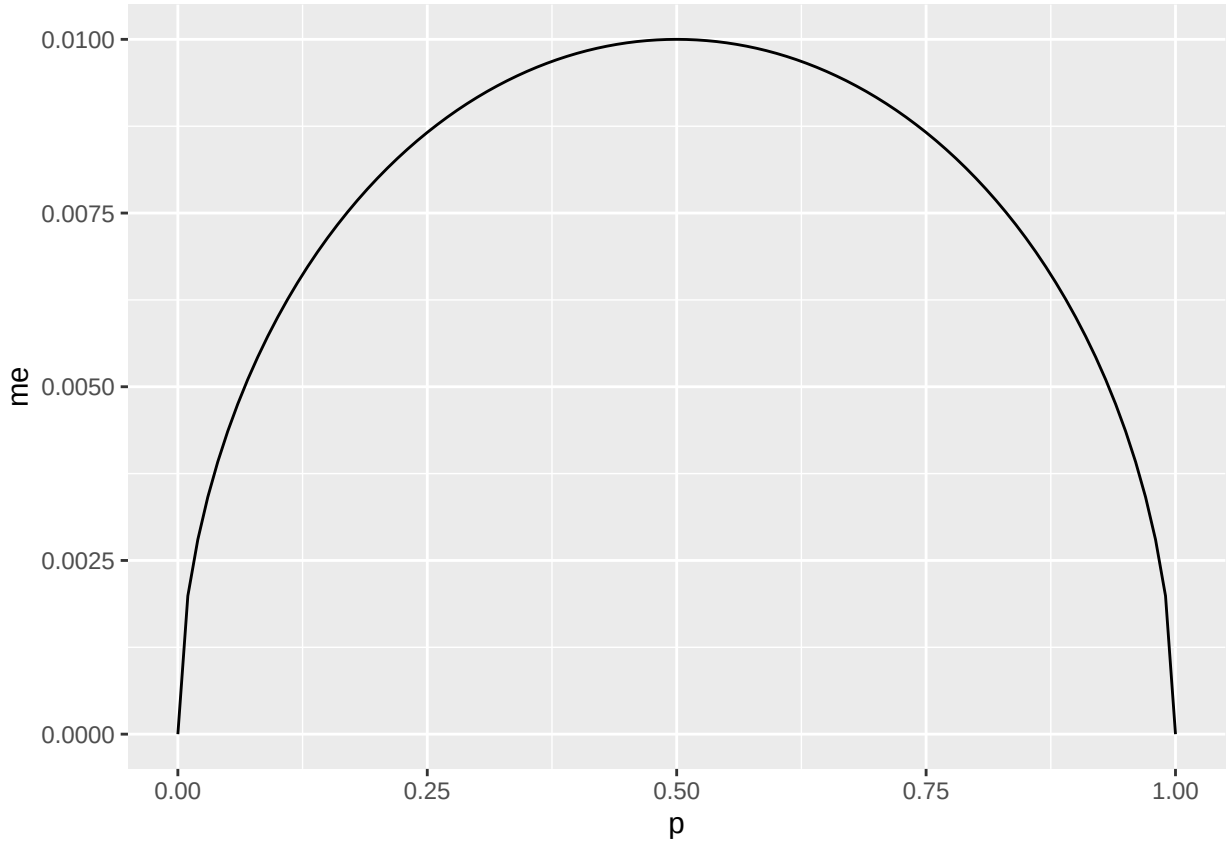


5. Describe the relationship between `p` and `me`. Include the margin of error vs. population proportion plot you constructed in your answer. For a given sample size, for which value of `p` is margin of error maximized?

Insert your answer here

Answer: As `p` approaches 0.50 in either direction, `me` increases. `me` is maximized when `p` = 0.50.

```
dd |> ggplot(aes(x = p, y = me)) +
  geom_line()
```



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

Answer: Under these conditions, the sampling distribution looks nearly normal and the center is about 0.1. The minimum p is about 0.05 and the maximum p is about 0.15

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

Answer: As **p** *approaches 0.5 from either direction, the spread of the sampling distribution increases. The center is always nearly equal to* **p**, *so it moves according to changes to* **p**. *The shape looks nearly normal regardless of changes to* **p**.

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

Insert your answer here

Answer: Increasing **n** *decreases the spread of the sampling distribution. When* **n** *is very low, the distribution is not very filled in, and therefore doesn't really resemble a normal distribution, and it's hard to determine a center. However, above a certain* **n** *(depending on* **p**), *changing* **n** *doesn't have too much effect on the shape or center of the distribution.*

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

Answer: First, I will test the independence of 10+ hours sleep and days training to see if there's a significant disproportion. The null hypothesis for this test is: of people who sleep 10+ hours, there is equal likelihood of them training 0-7 days (8 outcomes). The alternative hypothesis is that the distribution is uneven. The chi-square value for the distribution of days training is 206 and the p-value is near 0. This means we can be nearly 100% confident that people sleeping 10+ hours and the amount of days training are not independent. Since training 7 days is well above the expected proportion of .125 (1/8), we know that our sample of people who sleep 10+ hours are more likely to train 7 days. Next, we want to test if the population will also have a proportion for 7 days training above .125 with 95% confidence. The null hypothesis for this test is: of people who sleep 10+ hours, the amount who are training seven days a week is equal to or less than the amount of people who strength train any other amount days a week. The alternative hypothesis is that 7 days of training is higher than the expected proportion of 12.5%. Based on our sampling distribution, we can have 95% confidence that the proportion of 7 days training for a population that sleeps 10+ hours is between (21.8%, 32.1%). Since the minimum of this interval is greater that 12.5%, we have convincing evidence that people who sleep 10+ hours are more likely to strength train 7 days a week.

```
#Create dataset of sleep = 10+
most_sleep <- yrbss |>
  filter(school_night_hours_sleep == "10+")

#Chi_sq test to see if sleep 'explains' training days
most_sleep |>
  mutate(strength_training_7d = as.character(strength_training_7d)) |>
  drop_na(strength_training_7d) |>
  chisq_test(strength_training_7d ~ school_night_hours_sleep)
```

```
## # A tibble: 1 x 3
##   statistic chisq_df  p_value
##       <dbl>    <dbl>    <dbl>
## 1      206.        7 7.70e-41
```

```
#Add binary variable based on days training
most_sleep <- most_sleep |>
  mutate(train_ind = ifelse(strength_training_7d == 7, "yes", "no"))

#Calculate 95% CI of days training
most_sleep |>
  drop_na(train_ind) |>
  specify(response = train_ind, success = "yes") |>
  generate(reps = 1000, type = "bootstrap") |>
  calculate(stat = "prop") |>
  get_ci(level = .95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.224    0.314
```

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

Answer: If x_0:x_7 = .125 at a significance level of 0.05, there is a 5% probability that you can detect a change by chance.

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

Answer: To ensure a ME no greater that 1% with no knowledge of p, we have to calculate for p = .5, which is where ME will be highest, set z-score to 1.96 and solve for n. Our sample size must be at least 9,604.

```
(.5*.5)/((.01/1.96)^2)
```

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