Inference for categorical data

Tony Fraser

I will denote personal notes of confusion in red italics, like this.

There are a few places in this document where drop_na is used and it is very much throwing me. At this phase of my math career I will chock this up to me needing to walk before I can run. For real work, I do not think any of these drop_na lines should be in here.

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?

1. Insert your answer here

```
print(yrbss %>%
  group_by(text_while_driving_30d) %>%
  summarise(count = n()))
```

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

```
# or

print(table(yrbss$text_while_driving_30d))
```

```
          0           1-2         10-19         20-29           3-5
       4792           925           373           298           493
         30           6-9 did not drive
        827           311          4646
```

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

2 Insert your answer here

```
matches = nrow(
    yrbss %>%
  filter(text_while_driving_30d >=30 &  helmet_12m == "never")
)
print(sprintf("Proportion of non helmet wearing every day texters: %s%%",
                round((matches /nrow(yrbss)),4) * 100  ))
```

```
[1] "Proportion of non helmet wearing every day texters: 20.28%"
```

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   0.0772
```

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?

3. Insert your answer here

The margin of error here should be the difference between the confidence intervals, then split that difference in half.

I has a bunch of problems with this calculate method. In most of these code blocks I could not get calculate to run without first running drop_na on the column I used to build the indicator column.

```
ci <- no_helmet %>%
      drop_na(text_ind) %>% # <- 474 NA records.. 463 yes records..
      specify(response = text_ind, success = "yes") %>%
      generate(reps = 1000, type = "bootstrap") %>%
      calculate(stat = "prop") %>%
      get_ci(level = 0.95)

print(sprintf("MOE of this bootstrapped sample %s",
                round(((ci$upper_ci - ci$lower_ci)/2 ),4)
                ))
```

```
[1] "MOE of this bootstrapped sample 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. Interpret 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.

4. Insert your answer here

```
# Let's look at teenagers not getting enough sleep.
ci <- yrbss %>%
      ## no na records, no drop necessary.
      mutate(success_ind =
         ifelse(school_night_hours_sleep %in% c("10+", "9", "8"),
            "yes", "no")) %>%
      specify(response = success_ind, success = "yes") %>%
      generate(reps = 1000, type = "bootstrap") %>%
```

```
      calculate(stat = "prop") %>%
      get_ci(level = 0.95)

print(sprintf("Enough Teen Sleep Confidence Interval at 95%%: [%s,%s]",
                round(ci$lower_ci,4),
                round(ci$upper_ci,4)
                ))
```

```
[1] "Enough Teen Sleep Confidence Interval at 95%: [0.2695,0.2845]"
```

I'm going to write this out so I remember it. The resulting 95% confidence interval provides a range within which we are 95% confident that the true proportion of individuals (in the entire population from which yrbss was sampled) who get 8, 9, or 10+ hours of sleep on school nights. If we were to repeat this sampling process many times, obtaining different samples from the population each time and computing a 95% CI for each of them, we would expect about 95% of those intervals to contain the true population proportion.

```
## let's do the second one with the zero physical activity demographic
cpi <- yrbss %>%
      drop_na(physically_active_7d) %>%  # <- contains 273 na records.
      mutate(not_physically_active_ind =
                ifelse(physically_active_7d == 0,
                    "yes", "no")) %>%
      specify(response = not_physically_active_ind, success = "yes") %>%
      generate(reps = 1000, type = "bootstrap") %>%
      calculate(stat = "prop") %>%
      get_ci(level = 0.95)


print(sprintf("Has no physical activity confidence interval: [%s,%s]",
                round(cpi$lower_ci,4),
                round(cpi$upper_ci,4)
                ))
```

```
[1] "Has no physical activity confidence interval: [0.1569,0.1691]"
```

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:

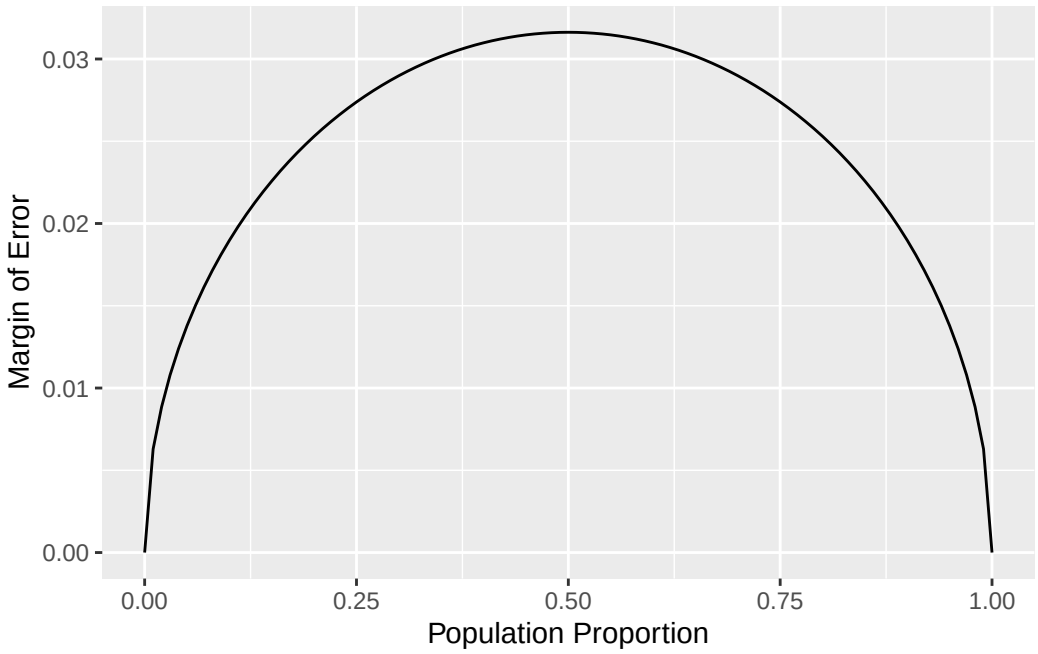
```
n <- 1000
```

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?

5. Insert your answer here

From this plot, we can infer that for a given sample size, the margin of error is maximized when the population proportion, p, is 0.5. This means that when the proportion is at its most uncertain (i.e., evenly split down the middle), the uncertainty (or margin of error) in our estimate is the highest."

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.

6. Insert your answer here

It seems to be normally distributed around about .12 with a spread of about .10, and a standard deviation that looks like about.02.

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.

7. Insert your answer here

It's like in question 5. when p_hat is in the middle, it has the most variance. The spread is much wider, with a standard deviation of perhaps .1 or so. When $p = .9$ it looks the same as .1 in terms of tightness of spread,. but the center of the distribution shifted to about .78

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

8. Insert your answer here

When n gets larger, the variance gets smaller and the distribution gets tighter.

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.

9 Insert your answer here

Ho With respect to people who strength train, there is no difference in proportion btw 10+ hour sleepers and less than 10 hour sleepers. *Ha* Those who sleep 10+ hour are more likely to strength train

```
yrbss_with_identifiers <- yrbss %>%
  drop_na(school_night_hours_sleep) %>%
  drop_na(strength_training_7d) %>%
      # <- Again with the drop na required!!! Which adjusts proportions!
  mutate(sleep_10plus_iden = ifelse(school_night_hours_sleep == "10+", "yes", "no"),
         strength_train_daily_iden = ifelse(strength_training_7d == 7, "yes", "no"))

infer_test <- yrbss_with_identifiers %>%
  specify(strength_train_daily_iden ~ sleep_10plus_iden, success = "yes") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in props", order = c("no", "yes"))

ci <- yrbss_with_identifiers %>%
  specify(strength_train_daily_iden ~ sleep_10plus_iden, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in props", order = c("no", "yes")) %>%
  get_ci(level = 0.95)

ci
```

```
# A tibble: 1 x 2
  lower_ci upper_ci
     <dbl>    <dbl>
1   -0.154  -0.0572
```

The negative upper and lower boundaries of the confidence interval suggest that individuals who get more sleep (10+ hours) have a higher likelihood of training daily than those who get less sleep. We should reject the null hypothesis Ho and accept Ha

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.

10. Insert your answer here

I think this question answers itself. The probability of making a Type 1 error is equivalent to the significance level (alpha) of the test. The significance level represents the probability of rejecting the null hypothesis when it is true. At a significance level of $\alpha = 0.05$ the probability of detecting a change by chance (i.e., making a Type 1 error) is 0.05 or 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.

11. Insert your answer here

I think we're supposed to solve for n. We have the MOE (.01) and Z (1.96). Not 100% sure how to do this, but in our plot the margin of error is highest when the proportion is close to .5, so that's the worst case scenario p which might want to be the one we use.

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

$$0.01 = 1.96 \times \sqrt{\frac{0.5 \times 0.5}{n}}$$

$$n = (1.96/0.01)^2 \times .5 \times .5$$

```
(1.96/.01)^2 * .5 * .5
[1] 9604
```

9,604 people!

I am not comfortable with this problem. As a guess I used the p population proportion as a fixed number of .5 because of the question it points me to that. I assume that is wrong and the solve from formula was right, but right or wrong Los Angeles and NYC are massive local governments. 9,604 is not even a rounding error to them. Or, what if said town was a total of twenty people and was governed by two. I also do not understand why 1300 people in a political poll is within a 3 percent margin of error.