Inference for categorical data

Sean Connin

3/21/21

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)
library(magrittr)
```

Creating a reproducible lab report

To create your new lab report, in RStudio, go to New File -> R Markdown… Then, choose From Template and then choose `Lab Report for OpenIntro Statistics Labs` from the list of templates.

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` .

Exercise 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 category counts that I compiled are as follows:

Exercise 2	“0” = 4792
Exercise 3	“1-2” = 925
Exercise 4	“3-5” = 493
Exercise 5	“6-9” = 311
Exercise 6	“10-19” = 373
Exercise 7	“20-29” = 298
Exercise 8	“30” = 827
Exercise 9	“did not drive” = 4646

```
cnts = table(yrbss$text_while_driving_30d)
```

Exercise 10	What is the proportion of people who have texted while driving every day in the past 30 days and never wear helmets?

P = 0.7119 based on my calculations

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

```
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"))

(freq_table <- as.data.frame(table(no_helmet$text_ind)))  # --> yes = 463, no = 6040

(prop <- 463/sum(freq_table$Freq)) #-- > .07119
```

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 %>%
  specify(response = text_ind, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)

#lower_ci = 0.647, upper_ci = .0773
```

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.

Exercise 11 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?

I calculated a margin of error of .006

```
p <- .07119 # proportion of people who did not wear helmet and texted
n <- 6503 # total number of people sampled who did not wear a helmet


SE <- sqrt((p * (1 - p))/n)
z = 1.96 # for alpha = .05

me <- z*SE # ---> 0.006249887
```

Exercise 12 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.

Variable 1.Proportion of white men who are physically active more than 3 days per week

---

Exercise 13 P value = 0.6957

Exercise 14 Margin of Error = 0.0157

Exercise 15	Confidence Interval = (0.680, 0.711)
Exercise 16	Interpretation: We are 95% confident that the average proportion of white men who are physically active for more than 3 days a week is between 68-71.

Variable 2.Proportion of Black and African American males who are physically active more than 3 days per week

Exercise 17	P value = 0.6147
Exercise 18	Margin of Error = .02415
Exercise 19	Confidence Interval = (0.590, 0.639)
Exercise 20	Interpretation: We are 95% confident that the average proportion of black and African American men who are physically active for more than 3 days a week is between 59-64%.

```
#Var 1. proportion of white men who are physically active more than 3 days per week?

df <- yrbss

active <- df%>%select(c(race, gender, physically_active_7d ))%>%filter(race == "White", ge
         nder=="male")

active%<>%mutate(activity_cat = ifelse(physically_active_7d  > 3, "yes", "no"))

(active_tbl <- (table(active$activity_cat)))  # --> yes = 2301, no = 1006

# calc proportion by indexing table (for kicks)

(pwm <- active_tbl[["yes"]][1]/sum(active_tbl)) #---> 0.6957


# calculate confidence intervals

active %>%
  specify(response = activity_cat, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95) # -- > lower_ci = 0.680, upper_ci = .0711

# calculate margin of error

pwm  # proportion of white males who are active the majority of the week
n <- sum(active_tbl) # total number of white males

SE <- sqrt((pwm * (1 - pwm))/n)

z = 1.96 # for alpha = .05

(me_wm <- z*SE) # --> .01568


#Var 2. proportion of black men who are physically active more than 3 days per week?

bm_active <- df%>%select(c(race, gender, physically_active_7d ))%>%
  filter(race == "Black or African American", gender=="male")%>%
  mutate(activity_cat = ifelse(physically_active_7d  > 3, "yes", "no"))

bm_active_tbl <-table(bm_active$activity_cat)

pbm <- bm_active_tbl[["yes"]][1]/(bm_active_tbl[["yes"]][1]+bm_active_tbl[["no"]][1]) # --
         > 0.6147


# calculate confidence intervals

bm_active %>%
  specify(response = activity_cat, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
```

```
  get_ci(level = 0.95) # -- > lower_ci = 0.590, upper_ci = .639

# calculate margin of error

pbm  # proportion of white males who are active the majority of the week
n <- sum(bm_active_tbl) # total number of white males

SE <- sqrt((pbm * (1 - pbm))/n)
z = 1.96 # for alpha = .05

(me_bm <- z*SE) # --> .02415
```

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")
```

Exercise 21 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?

The margin of error increases with the population proportion (PP) until reaching an peak at the mid-point of the PP distribution (i.e. values 0.00-0.50)). It then decreases inversely with the population proportion (values 0.50 to 1.0). In this fashion, the margin of error is greatest at a population proportion of 0.50 (i.e., mid-point of the distribution).

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.

Exercise 22 Describe the sampling distribution of sample proportions at $n = 300$ and $p = 0.1$. Be sure to note the center, spread, and shape.

The SDSP is approximately normal (bell shaped) but quite jagged, with several minor peaks and an apparent maxiumum around 0.8. The range is approximately 0.3 to .13.

Exercise 23 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.

The peak decreases and range increases as P increases from 0.0 to 0.5. As P varies from 0.5 to 1.0 the peak increases and the range decreases. The shape of the distribution is tighter and taller moving either direction along the x axis from th 0.50 midpoint.

Exercise 24 Now also change n. How does n appear to affect the distribution of $\hat{p}$?

As n increases the shape of the distribution becomes tighter, taller, and less irregular.

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`.

Exercise 25 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.

Null Hypothesis: Those who sleep > 10 hrs per day are no more likely to strength train 7 days per week than those who do not.

Alternative Hypothesis: Those who sleep > 10 hrs per day are more likely to strength train 7 days per week than those who do not

```r
# Create columns for difference of proportions

sleep <- yrbss%>%
  select(c(strength_training_7d, school_night_hours_sleep))%>%
  mutate(ten = ifelse(school_night_hours_sleep == "10+", "yes", "no"))%>%
  mutate(seven = ifelse(strength_training_7d == "7", "yes", "no"))

# calculate confidence intervals for difference of proportions

sleep %>%
  specify(ten ~ seven, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in props", order = c("no", "yes")) %>%
  get_ci(level = 0.95)

# -- > lower_ci = 0.0282, upper_ci = 0.00981


sl_tbl <- table(sleep$ten) #-- no= 12019, yes =316

str_tbl <- table(sleep$seven) #-- no= 10322, yes = 2085


n1 <- sum(sl_tbl)
n2 <- sum(str_tbl)



p_sl <- sl_tbl[["yes"]][1]/n1 # --> 0.0256

p_str <- str_tbl[["yes"]][1]/n2 # --> 0.168

p_diff <- p_sl-p_str # --> -0.1424

p_pool = p_sl+p_str # -->-->0.1936

null <- 0

# Calculate SE for the diff of two means

SE <- sqrt((p_pool*(1-p_pool)/n1)+(p_pool*(1-p_pool)/n2))

# Calculate Z statistics

z <- p_diff-null/SE # --> -0.1424

pnorm(z, lower.tail = TRUE) #??? not sure how to deal
```

Exercise 26 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.

There would .05 or 5% probability of detecting a change and making a Type 1 error.

Exercise 27 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.

Setting p = 0.5, our minimum sample number within the guidelines provided = 9,604

```
#Margin of error is largest when p = 0.5, thus worst case scenario:

(n <- 1.96^2*((0.5*(1-0.5))/(.01^2)))
```
