Lab6 - Inference for categorical data

Heleine Fouda

2023-10-15

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

```
library(openintro)
glimpse(yrbss)
```

```
## Rows: 13,583
## Columns: 13
## $ age                      <int> 14, 14, 15, 15, 15, 15, 15, 14, 15, 15, 15, 1~
## $ gender                   <chr> "female", "female", "female", "female", "fema~
## $ grade                    <chr> "9", "9", "9", "9", "9", "9", "9", "9", "9", ~
## $ hispanic                 <chr> "not", "not", "hispanic", "not", "not", "not"~
## $ race                     <chr> "Black or African American", "Black or Africa~
## $ height                   <dbl> NA, NA, 1.73, 1.60, 1.50, 1.57, 1.65, 1.88, 1~
## $ weight                   <dbl> NA, NA, 84.37, 55.79, 46.72, 67.13, 131.54, 7~
## $ helmet_12m               <chr> "never", "never", "never", "never", "did not ~
## $ text_while_driving_30d   <chr> "0", NA, "30", "0", "did not drive", "did not~
## $ physically_active_7d     <int> 4, 2, 7, 0, 2, 1, 4, 4, 5, 0, 0, 0, 4, 7, 7, ~
## $ hours_tv_per_school_day  <chr> "5+", "5+", "5+", "2", "3", "5+", "5+", "5+",~
## $ strength_training_7d     <int> 0, 0, 0, 0, 1, 0, 2, 0, 3, 0, 3, 0, 0, 7, 7, ~
## $ school_night_hours_sleep <chr> "8", "6", "<5", "6", "9", "8", "9", "6", "<5"~
```

```
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

```
higher_risk <- yrbss|>
  select (8:9)|>
  filter(!is.na(text_while_driving_30d) &     helmet_12m=="never")|>
  mutate(text = ifelse(text_while_driving_30d == 30, "everyday", "other")) |>
  count(text)
higher_risk
```

```
## # A tibble: 2 x 2
##   text          n
##   <chr>     <int>
## 1 everyday    463
## 2 other      6040
```

```
higher_riskp <- higher_risk$n/sum(higher_risk$n)
higher_riskp
```

```
## [1] 0.07119791 0.92880209
```

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")
no_helmet
```

```
## # A tibble: 6,977 x 13
##      age gender grade hispanic race                     height weight helmet_12m
##    <int> <chr>  <chr> <chr>    <chr>                     <dbl>  <dbl> <chr>
##  1    14 female 9     not      Black or African Americ~  NA      NA   never
##  2    14 female 9     not      Black or African Americ~  NA      NA   never
##  3    15 female 9     hispanic Native Hawaiian or Othe~   1.73   84.4 never
##  4    15 female 9     not      Black or African Americ~   1.6    55.8 never
##  5    14 male   9     not      Black or African Americ~   1.88   71.2 never
##  6    15 male   9     not      Black or African Americ~   1.75   63.5 never
##  7    16 male   9     not      Black or African Americ~   1.68   74.8 never
##  8    14 male   9     not      Black or African Americ~   1.73   73.5 never
##  9    15 male   9     not      Black or African Americ~   1.83   67.6 never
## 10    16 male   9     not      Black or African Americ~   1.83   73.5 never
## # i 6,967 more rows
## # i 5 more variables: text_while_driving_30d <chr>, physically_active_7d <int>,
## #   hours_tv_per_school_day <chr>, strength_training_7d <int>,
## #   school_night_hours_sleep <chr>
```

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"))
no_helmet
```

```
## # A tibble: 6,977 x 14
##      age gender grade hispanic race                     height weight helmet_12m
##    <int> <chr>  <chr> <chr>    <chr>                     <dbl>  <dbl> <chr>
##  1    14 female 9     not      Black or African Americ~  NA      NA   never
##  2    14 female 9     not      Black or African Americ~  NA      NA   never
##  3    15 female 9     hispanic Native Hawaiian or Othe~   1.73   84.4 never
##  4    15 female 9     not      Black or African Americ~   1.6    55.8 never
##  5    14 male   9     not      Black or African Americ~   1.88   71.2 never
##  6    15 male   9     not      Black or African Americ~   1.75   63.5 never
##  7    16 male   9     not      Black or African Americ~   1.68   74.8 never
##  8    14 male   9     not      Black or African Americ~   1.73   73.5 never
##  9    15 male   9     not      Black or African Americ~   1.83   67.6 never
## 10    16 male   9     not      Black or African Americ~   1.83   73.5 never
## # i 6,967 more rows
## # i 6 more variables: text_while_driving_30d <chr>, physically_active_7d <int>,
## #   hours_tv_per_school_day <chr>, strength_training_7d <int>,
## #   school_night_hours_sleep <chr>, text_ind <chr>
```

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   0.0647   0.0770
```

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

```
# ME = 1.96 * SE = 1.96 *  p(1 -p) /n
n <- 1000
p <- higher_risk$n[1] / sum(higher_risk$n)
ME <- 1.96 * sqrt(p*(1-p)/n)
ME
```

```
## [1] 0.01593864
```

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

```
glimpse(yrbss)
```

```
## Rows: 13,583
## Columns: 13
## $ age                      <int> 14, 14, 15, 15, 15, 15, 15, 14, 15, 15, 15, 1~
## $ gender                   <chr> "female", "female", "female", "female", "fema~
## $ grade                    <chr> "9", "9", "9", "9", "9", "9", "9", "9", "9", ~
## $ hispanic                 <chr> "not", "not", "hispanic", "not", "not", "not"~
## $ race                     <chr> "Black or African American", "Black or Africa~
## $ height                   <dbl> NA, NA, 1.73, 1.60, 1.50, 1.57, 1.65, 1.88, 1~
## $ weight                   <dbl> NA, NA, 84.37, 55.79, 46.72, 67.13, 131.54, 7~
## $ helmet_12m               <chr> "never", "never", "never", "never", "did not ~
## $ text_while_driving_30d   <chr> "0", NA, "30", "0", "did not drive", "did not~
## $ physically_active_7d     <int> 4, 2, 7, 0, 2, 1, 4, 4, 5, 0, 0, 0, 4, 7, 7, ~
## $ hours_tv_per_school_day  <chr> "5+", "5+", "5+", "2", "3", "5+", "5+", "5+",~
## $ strength_training_7d     <int> 0, 0, 0, 0, 1, 0, 2, 0, 3, 0, 3, 0, 0, 7, 7, ~
## $ school_night_hours_sleep <chr> "8", "6", "<5", "6", "9", "8", "9", "6", "<5"~
```

Let's first examine the TV time variable or hours_tv_per_school_day.

Proportion of Interest: Students who spend more than 5 hours watching TV

```
tv_time <- yrbss|>
 filter(!is.na(hours_tv_per_school_day)) |>
  mutate(five_hours_or_more = ifelse(hours_tv_per_school_day == "5+", "yes","no"))

tv_time |>
  count(five_hours_or_more) |>
  mutate(p = n / sum(n))
```

```
## # A tibble: 2 x 3
##   five_hours_or_more     n     p
##   <chr>              <int> <dbl>
## 1 no                 11650 0.880
## 2 yes                 1595 0.120
```

The 95% confidence interval for the probability of a student watching five or more hours of television is 0.114-0.125.

```
# Confidence interval
tv_time |>
  specify(response = five_hours_or_more, 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    0.115    0.126
```

Let's calculate the margin of error (ME):

```
# Margin of error
n <- nrow(tv_time)
p <- sum(tv_time$five_hours_or_more == "yes")/n
z <- 1.96 # z-score 95%
se <- z*sqrt((p*(1-p))/n)
me <- z*se
me
```

```
## [1] 0.01086369
```

The margin of error is 0.0108

Let's now calculate the sleep variable or school_night_hours_sleep

Proportion of Interest: Students who reported having at least 8 hours sleep per night

```
# students who sleep 8 hours or more
unique(yrbss$school_night_hours_sleep)
```

```
## [1] "8"   "6"   "<5"  "9"   "10+" "7"   "5"   NA
```

```
sleep_well <- yrbss |>
  filter(!is.na(school_night_hours_sleep)) |>
  mutate(eight_hours = ifelse(school_night_hours_sleep %in% c("8","9","10+"), "yes", "no"))

sleep_well |>
  count(eight_hours) |>
  mutate(p = n / sum(n))
```

```
## # A tibble: 2 x 3
##   eight_hours     n     p
##   <chr>       <int> <dbl>
## 1 no           8564 0.694
## 2 yes          3771 0.306
```

```
# Confidence interval
sleep_well |>
  specify(response = eight_hours, 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    0.297    0.314
```

The 95% confidence interval for the probability that a students gets at least 8 hours of sleep is between 0.297-0.313.

Let's now calculate the margin of error

```
# Margin of error
n <- nrow(sleep)
z <- 1.96 # z-score 95%
se <- z*sqrt((p*(1-p))/n)
me <- z*se
me
```

```
## [1] 0.2795687
```

The margin of error is 0.279

How does the proportion affect the margin of error?

Note: The margin of error changes with sample size,and is also affected by the proportion.

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:

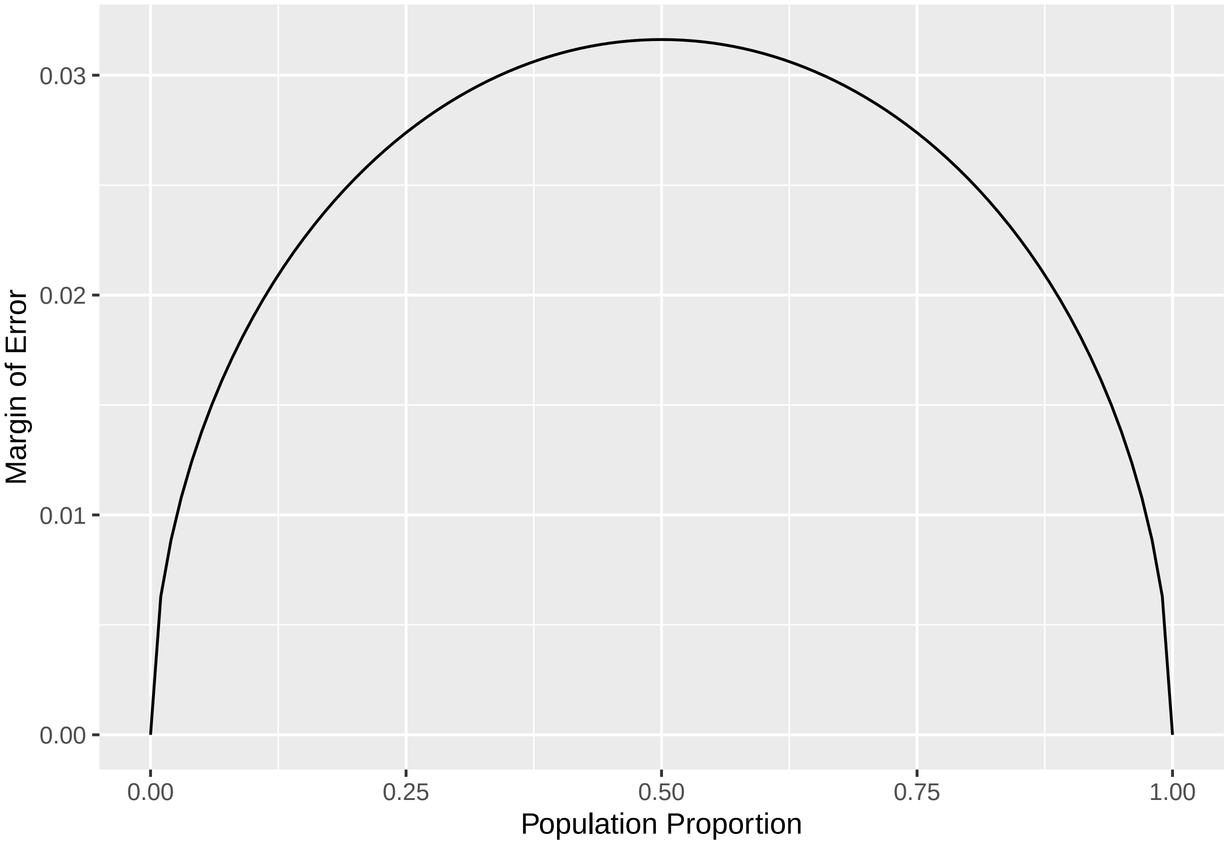
```
n <- 1000
```

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?

The me increase as the population proportion increases, and decreases as the population proportion decreases.Margin of error is higher at the population of 50%.

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.

The distribution appears normal, mostly bell-curved and symmetrical - sampling proportions are clustering around center with symmetrical tapering on either side. The center appears to be around 0.1, and spread from about 0.03 to 1.13. The standard deviation of the sample proportions is .017.

Simulation image: https://github.com/Heleinef/Data-Science-Master__Heleine/blob/main/photolab

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 proportion p increases, the center increases, the spread gets wider and stays consistent at first and then as p increases even more, the center becomes slimmer and there is less and less clustering around the center of range and less overall conformity to a normal distribution.

Simulation photo: https://github.com/Heleinef/Data-Science-Master__Heleine/blob/main/p__0.51

Simulation photo: https://github.com/Heleinef/Data-Science-Master__Heleine/blob/main/p__96

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

Insert your answer here As the sample size n increases, the spread decreases and symmetry increases. Also, one notes that more data cluster around the center forming a bell curve in accordance with the properties and characteristics of the normal distribution as defined by the CLT .

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

Null hypothesis (H0): Those who sleep 10+ hours per day are not more likely to strength train every day of the week.

Alternative hypothesis(HA): Those who sleep 10+ hours per day are more likely to strength train every day of the week.

```
# strength train
training <- yrbss |>
  filter(!is.na(strength_training_7d)) |>
  mutate(everyday = ifelse(strength_training_7d == "7", "yes", "no"))

training|>
  count(everyday) |>
  mutate(p = n / sum(n))
```

```
## # A tibble: 2 x 3
##   everyday     n     p
##   <chr>    <int> <dbl>
## 1 no       10322 0.832
## 2 yes       2085 0.168
```

```
# CI
 training|>
 specify(response = everyday, 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    0.162    0.174
```

Proportion of students training 7 days: 0.168 95% Confidence Interval: 0.161, 0.174 ME: 0.01289

```
# Margin of error
p <- sum(training$everyday == "yes") / sum(training$everyday == "yes"|training$everyday ==  "no")
n <- nrow(training)
z <- 1.96
se <- z*sqrt((p*(1-p))/n)

me<- z * se
me
```

```
## [1] 0.01289576
```

Now let's look at sleep 10+ hours

```
# 10 hours sleep or more
sleep_10 <- yrbss |>
  filter(!is.na(school_night_hours_sleep)) |>
  mutate(ten_or_more = ifelse(school_night_hours_sleep == "10+", "yes", "no"))

sleep_10 |>
  count(ten_or_more) |>
  mutate(p = n / sum(n))
```

```
## # A tibble: 2 x 3
##   ten_or_more     n      p
##   <chr>       <int>  <dbl>
## 1 no          12019 0.974
## 2 yes           316 0.0256
```

```
# CI
sleep_10 |>
 specify(response = ten_or_more, 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   0.0229   0.0286
```

```
# Margin of error
p <- sum(sleep_10$ten_or_more == "yes") / sum(sleep_10$ten_or_more == "yes"|sleep_10$ten_or_more ==  "n
n <- nrow(sleep_10)
z <- 1.96
se <- z*sqrt((p*(1-p))/n)

me<- z * se
me
```

```
## [1] 0.005464886
```

Proportion of students sleeping 10+ hours: 0.0256 95% Confidence Interval: 0.0229, 0.0286 ME: 0.005464

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.

Insert your answer here

The probability of making a Type I error is equal to the level of significance, so it is 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

```
E <- 0.01 #Margin of error
z <- 1.96 #z-score for 95% confidence

p <- 0.5 #Margin for error is highest at .5 of the population proportion

n <- z**2 * p*(1-p)/ME**2
n
```

```
## [1] 3780.503
```

I would need to sample at least 3781 people
