Inference for categorical data

Peter Gatica

03/21/2021

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)
```

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`.

```
head(yrbss,20)
```

```
## # A tibble: 20 x 13
##      age gender grade hispanic race  height weight helmet_12m text_while_driv~
##    <int> <chr>  <chr> <chr>    <chr>  <dbl>  <dbl> <chr>      <chr>
##  1    14 female 9     not      Blac~  NA      NA   never      0
##  2    14 female 9     not      Blac~  NA      NA   never      <NA>
##  3    15 female 9     hispanic Nati~   1.73   84.4 never      30
##  4    15 female 9     not      Blac~   1.6    55.8 never      0
##  5    15 female 9     not      Blac~   1.5    46.7 did not r~ did not drive
##  6    15 female 9     not      Blac~   1.57   67.1 did not r~ did not drive
##  7    15 female 9     not      Blac~   1.65  132.  did not r~ <NA>
##  8    14 male   9     not      Blac~   1.88   71.2 never      <NA>
##  9    15 male   9     not      Blac~   1.75   63.5 never      <NA>
```

```
## 10    15 male   10    not      Blac~   1.37   97.1 did not r~ <NA>
## 11    15 female 9     not      Blac~   1.68   69.8 did not r~ 0
## 12    15 female 9     not      Blac~   1.65   66.7 did not r~ did not drive
## 13    15 female 9     not      Blac~   1.63   67.1 did not r~ <NA>
## 14    16 male   9     not      Blac~   1.68   74.8 never      0
## 15    16 male   9     not      Blac~   1.85   74.4 did not r~ did not drive
## 16    15 male   9     not      Blac~   1.78   70.3 did not r~ 0
## 17    14 male   9     not      Blac~   1.73   73.5 never      did not drive
## 18    15 male   9     not      Blac~   1.83   67.6 never      0
## 19    14 male   9     not      Blac~   1.68   46.3 <NA>       0
## 20    16 male   9     not      Blac~   1.83   73.5 never      did not drive
## # ... with 4 more variables: physically_active_7d <int>,
## #   hours_tv_per_school_day <chr>, strength_training_7d <int>,
## #   school_night_hours_sleep <chr>
```

1. What are the counts within each category for the amount of days these students have texted while driving within the past 30 days?

Answer:

```
(yrbss %>% group_by(text_while_driving_30d) %>% summarise(count_cats = n()))
```

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

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

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

```
head(no_helmet <- yrbss %>%
  filter(helmet_12m == "never"),10)
```

```
## # A tibble: 10 x 13
##      age gender grade hispanic race  height weight helmet_12m text_while_driv~
##    <int> <chr>  <chr> <chr>    <chr>  <dbl>  <dbl> <chr>      <chr>
##  1    14 female 9     not      Blac~  NA       NA   never      0
##  2    14 female 9     not      Blac~  NA       NA   never      <NA>
##  3    15 female 9     hispanic Nati~   1.73   84.4 never      30
##  4    15 female 9     not      Blac~   1.6    55.8 never      0
##  5    14 male   9     not      Blac~   1.88   71.2 never      <NA>
```

```
##  6    15 male   9     not      Blac~   1.75   63.5 never      <NA>
##  7    16 male   9     not      Blac~   1.68   74.8 never      0
##  8    14 male   9     not      Blac~   1.73   73.5 never      did not drive
##  9    15 male   9     not      Blac~   1.83   67.6 never      0
## 10    16 male   9     not      Blac~   1.83   73.5 never      did not drive
## # ... with 4 more variables: physically_active_7d <int>,
## #   hours_tv_per_school_day <chr>, strength_training_7d <int>,
## #   school_night_hours_sleep <chr>
```

```
(no_helmet_count <- count(no_helmet))
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  6977
```

```
(count_nh_30d_text <- (no_helmet %>% filter(text_while_driving_30d == 30) %>% summarise(count_cats = n(
```

```
## # A tibble: 1 x 1
##   count_cats
##        <int>
## 1        463
```

Answer:

```
(prop_30dtxt_2_nohelmet <- round(count_nh_30d_text/no_helmet_count, 3))
```

```
##   count_cats
## 1      0.066
```

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 %>% select(helmet_12m,text_while_driving_30d) %>%
  mutate(text_ind = ifelse(text_while_driving_30d == "30", "yes", "no")))
```

```
## # A tibble: 6,977 x 3
##    helmet_12m text_while_driving_30d text_ind
##    <chr>      <chr>                  <chr>
##  1 never      0                      no
##  2 never      <NA>                   <NA>
##  3 never      30                     yes
##  4 never      0                      no
##  5 never      <NA>                   <NA>
##  6 never      <NA>                   <NA>
##  7 never      0                      no
##  8 never      did not drive          no
##  9 never      0                      no
## 10 never      did not drive          no
## # ... with 6,967 more rows
```

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)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   0.0657   0.0778
```

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.

1. 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?

   Answer:

$$ME = 1.96 \times SE = 1.96 \times \sqrt{\hat{p}(1-\hat{p})/n}\,.$$

$$1.96 \times \sqrt{.066(1-.066)/463}$$

```
(ME_30dtxt_2_nohelmet <- round(1.96 * (sqrt((prop_30dtxt_2_nohelmet * (1 - prop_30dtxt_2_nohelmet))/cou
```

```
##   count_cats
## 1     0.0226
```

1. 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 genders first, and then use these data sets to construct the confidence intervals.

```
yrbss_male <- yrbss %>%
  filter(gender == "male")
```

```
The counts within each category for the amount of days male students did strength
within the past 7 days.
```

```r
(yrbss_male %>% group_by(strength_training_7d) %>% summarise(count_actv_cats = n()))
```

```
## # A tibble: 9 x 2
##   strength_training_7d count_actv_cats
##                  <int>           <int>
## 1                    0            1346
## 2                    1             428
## 3                    2             564
## 4                    3             776
## 5                    4             609
## 6                    5             786
## 7                    6             317
## 8                    7            1493
## 9                   NA             631
```

```
What is the proportion of males who are physically active and have done strength training in
the past 7 days and ?

Male students that are physically active.
```

```r
active_males <- yrbss_male %>%
  filter(physically_active_7d > 0)
```

```
The count of physically active males.
```

```r
(num_phys_actv_count <- count(active_males))
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  5985
```

```
The count of physically active males that did strenght training in the last 7 days.
```

```r
(count_phact_7d_strtrn <- (active_males %>% filter(strength_training_7d == 7) %>% summarise(count_pa_ca
```

```
## # A tibble: 1 x 1
##   count_pa_cats
##           <int>
## 1          1441
```

It may be easier to calculate the proportion if you create a new variable that specifies whether the individual has strength training every day while being physically active over the past 7 days or not. We will call this variable `str_trn_ind`.

```r
(active_males <- active_males %>% select(physically_active_7d,strength_training_7d) %>%
  mutate(str_trn_ind = ifelse(strength_training_7d == "7", "yes", "no")))
```

```
## # A tibble: 5,985 x 3
##    physically_active_7d strength_training_7d str_trn_ind
##                   <int>                <int> <chr>
##  1                    4                    0 no
##  2                    5                    3 no
##  3                    7                    7 yes
##  4                    7                    7 yes
##  5                    7                    7 yes
##  6                    4                    3 no
##  7                    7                    7 yes
##  8                    7                    2 no
##  9                    7                    5 no
## 10                    7                    5 no
## # ... with 5,975 more rows
```

Calcualte the observed statistic.

```r
(p_hat <- round(active_males %>%
  specify(response = str_trn_ind, success = "yes") %>%
  calculate(stat = "prop"),3))
```

```
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1 0.258
```

Calculate the Confidence Interval

```r
round(active_males %>%
  specify(response = str_trn_ind, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95),3)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.247     0.27
```

Calculate the Margin of Error

$$ME = 1.96 \times SE = 1.96 \times \sqrt{\hat{p}(1-\hat{p})/n}\,.$$

$$ME = 1.96 \times \sqrt{.258(1-.258)/1441}$$

```r
(ME_phact_7d_strtrng <- round(1.96 * (sqrt((p_hat * (1 - p_hat))/count_nh_30d_text)),4))
```

```
##     stat
## 1 0.0399
```

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))
```

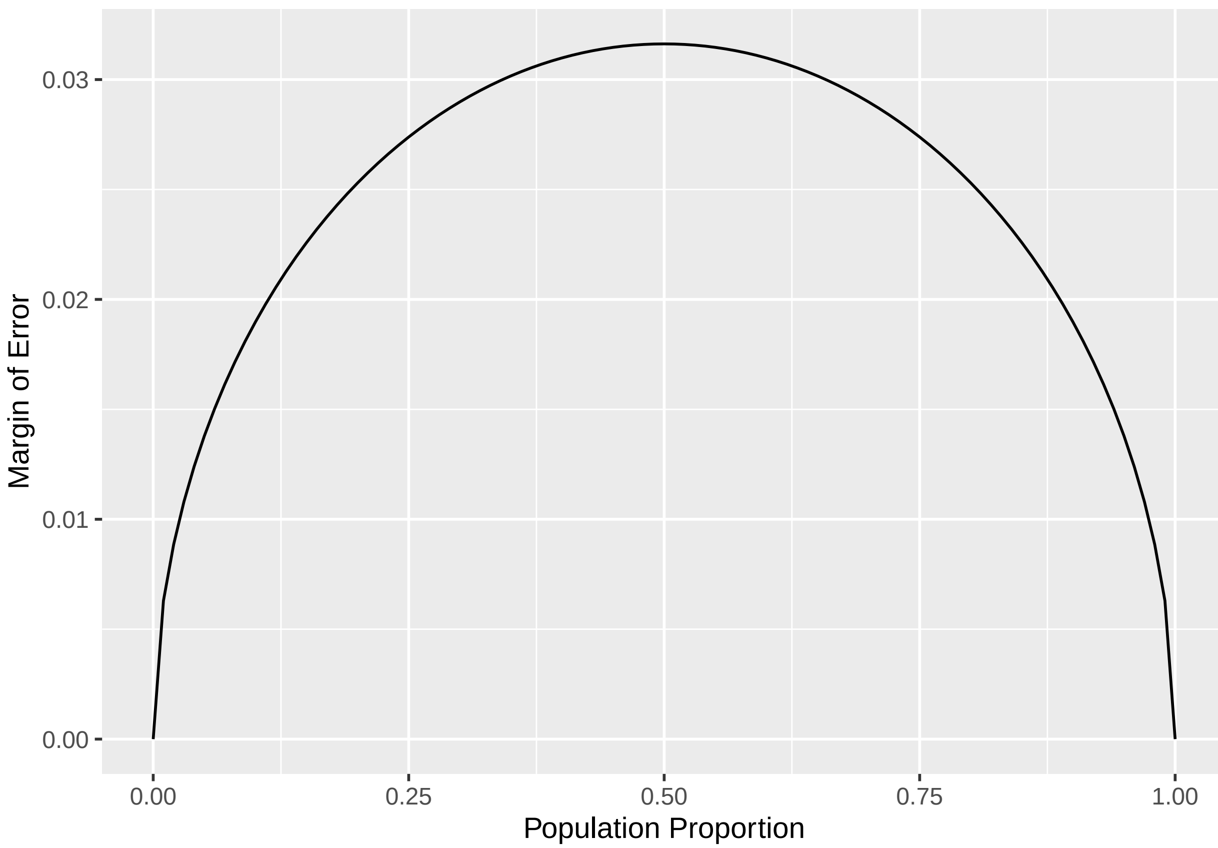
```
##   [1] 0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10 0.11 0.12 0.13 0.14
##  [16] 0.15 0.16 0.17 0.18 0.19 0.20 0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28 0.29
##  [31] 0.30 0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39 0.40 0.41 0.42 0.43 0.44
##  [46] 0.45 0.46 0.47 0.48 0.49 0.50 0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59
##  [61] 0.60 0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69 0.70 0.71 0.72 0.73 0.74
##  [76] 0.75 0.76 0.77 0.78 0.79 0.80 0.81 0.82 0.83 0.84 0.85 0.86 0.87 0.88 0.89
##  [91] 0.90 0.91 0.92 0.93 0.94 0.95 0.96 0.97 0.98 0.99 1.00
```

```
(me <- 2 * sqrt(p * (1 - p)/n))
```

```
##   [1] 0.000000000 0.006292853 0.008854377 0.010788883 0.012393547 0.013784049
##   [7] 0.015019987 0.016136914 0.017158088 0.018099724 0.018973666 0.019788886
##  [13] 0.020552372 0.021269697 0.021945387 0.022583180 0.023186203 0.023757104
##  [19] 0.024298148 0.024811288 0.025298221 0.025760435 0.026199237 0.026615785
##  [25] 0.027011109 0.027386128 0.027741665 0.028078461 0.028397183 0.028698432
##  [31] 0.028982753 0.029250641 0.029502542 0.029738863 0.029959973 0.030166206
##  [37] 0.030357866 0.030535226 0.030698534 0.030848015 0.030983867 0.031106269
##  [43] 0.031215381 0.031311340 0.031394267 0.031464265 0.031521421 0.031565804
##  [49] 0.031597468 0.031616451 0.031622777 0.031616451 0.031597468 0.031565804
##  [55] 0.031521421 0.031464265 0.031394267 0.031311340 0.031215381 0.031106269
##  [61] 0.030983867 0.030848015 0.030698534 0.030535226 0.030357866 0.030166206
##  [67] 0.029959973 0.029738863 0.029502542 0.029250641 0.028982753 0.028698432
##  [73] 0.028397183 0.028078461 0.027741665 0.027386128 0.027011109 0.026615785
##  [79] 0.026199237 0.025760435 0.025298221 0.024811288 0.024298148 0.023757104
##  [85] 0.023186203 0.022583180 0.021945387 0.021269697 0.020552372 0.019788886
##  [91] 0.018973666 0.018099724 0.017158088 0.016136914 0.015019987 0.013784049
##  [97] 0.012393547 0.010788883 0.008854377 0.006292853 0.000000000
```

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")
```



1. 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?

   Answer: The relationship of 'p' and 'me' is one of a normal distribution in that as population proportion 'p' increases from 0 the margin of error increases until it reaches a maximum of .03 at .50 of the population distribution and then decreases again as it approaches 1.0.

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.

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

   Answer: I appears to have close to a normal distribution with the center is at .10, an even spread on both sides of the center.

2. 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.

   Answer: The shape of the distribution gets closer to normal as the population proportion approached .50 as the plot above shows.

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

   Answer: As the sample size increases it appears that although the distribution is close to normal the spread seems to be more narrow.

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`.

1. 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.

2. 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.

3. 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.
