Inference for categorical data

Title: CUNY SPS MDS Data606_LAB6"

Author: Charles Ugiagbe

Date: “10/16/2021”

Load 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`.

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?

Solution 1:

```
stud_text <- yrbss %>%
  count(text_while_driving_30d, sort=TRUE)
stud_text
```

```
## # 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
```

As seen from the data frame above, 0 student didn't text while driving, 4646 students texted 1-2 days; 493 students texted within 3-5 days;311 students texted within 6-9 days;373 students texted within 10-19 days; 298 students texted within 20-29 days; 827 students texted for 30 days;

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

Solution 2:

```
data('yrbss', package='openintro')
no_helmet <- yrbss %>%
  filter(helmet_12m == "never")
```

```
no_helmet <- no_helmet %>%
  mutate(text_ind = ifelse(text_while_driving_30d == "30", "yes", "no"))
no_helmet %>% count(text_ind)
```

```
## # A tibble: 3 x 2
##   text_ind     n
##   <chr>    <int>
## 1 no        6040
## 2 yes        463
## 3 <NA>       474
```

```
no_helmet %>%
  count(text_ind) %>%
  mutate(p = n / sum(n))
```

```
## # A tibble: 3 x 3
##   text_ind     n      p
##   <chr>    <int>  <dbl>
## 1 no        6040 0.866
## 2 yes        463 0.0664
## 3 <NA>       474 0.0679
```

```
no_helmet %>%
  filter(text_ind == "yes" | text_ind == "no") %>%
  count(text_ind) %>%
  mutate(p1 = n / sum(n))
```

```
## # A tibble: 2 x 3
##   text_ind     n     p1
##   <chr>    <int>  <dbl>
## 1 no        6040 0.929
## 2 yes        463 0.0712
```

Proportion of those who texted for 30 days & no_helmet = 0.0712

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.

Exercise 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?

Solution 3:

```
set.seed(0803)
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   0.0654   0.0773
```

Margin Error = (Upper Tail - Lower tail)/2

```
UpperTail <- 0.0773
LowerTail <- 0.0654
ME <- (UpperTail - LowerTail)/2
ME
```

```
## [1] 0.00595
```

Exercise 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.

How does the proportion affect the margin of error?

Solution 4:

```
# 1st Category: the proportion of non-helmet wearers than are male

set.seed(0802)
no_helmet %>%
  specify(response = gender, success = "male") %>%
  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    0.569    0.592
```

Confidence interval = (56.9%, 59.2%) We are 95% confident that the proportion of non-helmet wearers and are male is between 40.9% and 43.2%

```
UpperTail <- 0.5922
LowerTail <- 0.56897
ME <- (UpperTail - LowerTail)/2
ME
```

```
## [1] 0.011615
```

Mean of Error for the proportion of Non-helmet wearers that are Male = 0.012

```
# the proportion of non-helmet wearers and are White

no_helmet %>%
  specify(response = hispanic, success = "hispanic") %>%
  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    0.265    0.286
```

Confidence interval = (26.5%, 28.6%)

We are 95% confident that the proportion of non-helmet wearers and are male is between 26.5% and 28.6%

```
UpperTail <- 0.2858
LowerTail <- 0.26546
ME <- (UpperTail - LowerTail)/2
ME
```

```
## [1] 0.01017
```

Margin of Error for the proportion of non-helmet wearers that are hispanic is 0.10

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
```

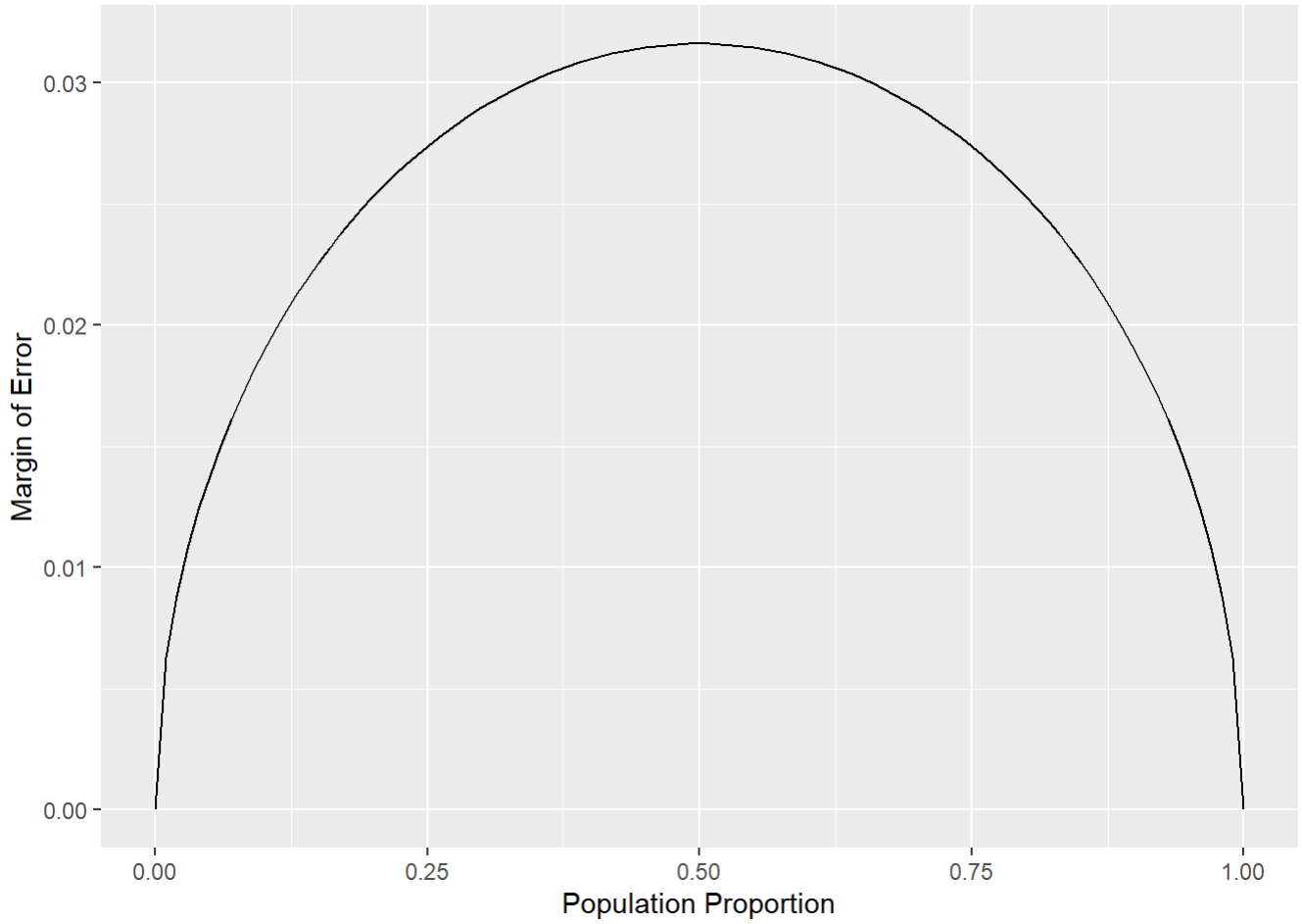
The first step is to make a variable `p` that is a sequence from 0 to 1 with each number incremented by 0.01. You can then create a variable of the margin of error (`me`) associated with each of these values of `p` using the familiar approximate formula ($ME = 2 \times SE$).

```
p <- seq(from = 0, to = 1, by = 0.01)
me <- 2 * sqrt(p * (1 - p)/n)
```

Exercise 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?

Solution 5:

```
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")
```



As the proportion p increases, margin of error me increases as well until it reaches its maximum at 50% and starts decreasing as the proportion passes 50%. Therefore, for a given sample size, the margin of error is maximized for p = 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.

Exercise 6 Describe the sampling distribution of sample proportions at $n = 300$ and $p = 0.1$. Be sure to note the center, spread, and shape.

Solution 6:

The sampling distribution is approximately normal, centered at p = 0.1, and the spread tends toward the mean. We can note that there is not much of the spread.

Exercise 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.

Solution 7:

As the value of P increase with n remaining constant, the shape of the distribution remains slightly normal; more spread is seen around the centre that is larger mean error (ME). Also, as the p increases above 50%, the spread of the distribution started decreasing. As i adjust the range of the min and max, the spread of the distribution decreases.

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

Solution 8:

As n increases, the distribution become more and more normal, that's, there is less spread and there is more data toward the center. In term of confidence interval, we would say that the standard error decreases (less spread), hence the margin of error decreases as well.

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 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.

Solution 9:

Chi-Square test would be a good option for testing relationship.

H0: There is no relation between to sleep 10+ hours per day and to strength train every day of the week

H1: There is a relation between to sleep 10+ hours per day and to strength train every day of the week

```
# New data frame with the needed conditions

sleep_more <- yrbss %>%
  mutate(sleep_ind = ifelse(school_night_hours_sleep == "10+", "yes", "no"),
         train_ind = ifelse(strength_training_7d == "7", "yes", "no") )
```

```
sleep_more %>%
  filter(sleep_ind == "yes" | sleep_ind == "no") %>%
  count(sleep_ind) %>%
  mutate(p_sleep = n / sum(n))
```

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

```
sleep_more %>%
  filter(train_ind == "yes" | train_ind == "no") %>%
  count(train_ind) %>%
  mutate(p_train = n / sum(n))
```

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

calculate the difference in proportions for the two categories

Difference = 0.0256 - 0.1681 = -0.1425

Since the difference is significant, we do independence test with infer

```
sleep_more %>%
  specify(sleep_ind ~ train_ind , success = "yes") %>%
  hypothesize( null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate ("diff in props") %>%
  get_ci(level = 0.95)
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1 -0.00697  0.00714
```

The value from the independence test is very small to almost negligible. we would say that there is not convincing evidence that those who sleep 10+ hours per day are more likely to strength train every day of the week

---

Exercise 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.

Solution 10:

We know that the probability of making type I error is α, the level of significance.

At α=0.05, I am willing to accept 5% chance that I am wrong when I reject the null hypothesis.

Thus, i would argue that the probability that I could detect a change at α=0.05 is 5%.

Exercise 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.

Solution 11:

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

Our goal is to find the smallest sample size n so that this margin of error is not greater than 1%. For a 95% confidence level.

The value z* corresponds to 1.96:

$ME = Z^* \times \sqrt{\frac{p(1-p)}{n}}$

$0.01 = 1.96 \times \sqrt{\frac{p(1-p)}{n}}$

Since the value of P is unknown, we assume a value of P when the margin of error is largest; that is when p is 0.5.

$0.01 = 1.96 \times \sqrt{\frac{0.5(1-0.5)}{n}}$

$0.01^2 = 1.96^2 \times \frac{0.5(1-0.5)}{n}$

$n = 1.96^2 \times \frac{0.5(1-0.5)}{0.01^2}$

$n = 9604$
