Inference for categorical data

Souleymane Doumbia

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?

```
data(`yrbss`)

Count_twd30 <- yrbss %>%
  count(text_while_driving_30d)
Count_twd30
```

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

The code above answers the counts question for the amount of days these students have texted while driving within the past 30 days

2. 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`.

```
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"))
```

```
Proportion_noH_t30 <- no_helmet %>%
  count(text_ind) %>%
  mutate(p_prop = n / sum(n)) %>%
  filter(text_ind == "yes") %>%
  pull(p_prop) %>%
  round(2)
```

The proportion of people who have texted while driving every day in the past 30 days and never wear helmets is 7%.

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.

```
set.seed(999)
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   0.0649   0.0769
```

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?

Since CI is [0.0649,0.0769], Margin of error is: (0.0769 - 0.0649) / 2 = 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. 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.

    a) First choice Gender:

```r
set.seed(999)
gender_ci <- yrbss %>%
  mutate(gen_f_m = ifelse(gender == "female", "f", "m")) %>%
  drop_na(gen_f_m) %>%
  specify(response = gen_f_m, success = "f") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)
gender_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.480    0.496
```

```r
moe_gender <- round((gender_ci$upper_ci - gender_ci$lower_ci) / 2,4)
moe_gender
```

```
## [1] 0.0079
```

```
b) Second choice Hispanic:
```

```r
set.seed(999)
hispanic_ci <- yrbss %>%
  mutate(hisp_y_n = ifelse(hispanic == "hispanic", "yes", "no")) %>%
  drop_na(hisp_y_n) %>%
  specify(response = hisp_y_n, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)
hispanic_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.249    0.263
```

```
moe_hispanic <- round((hispanic_ci$upper_ci - hispanic_ci$lower_ci) / 2,4)
moe_hispanic
```

```
## [1] 0.0072
```

From the outputs of a and b we computed the margin error of the proportion of student who are female to be 0.0079, and 0.0072 to be the proportion of student who identified themselves as hispanic.

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 ($ME = 2 \times SE$).

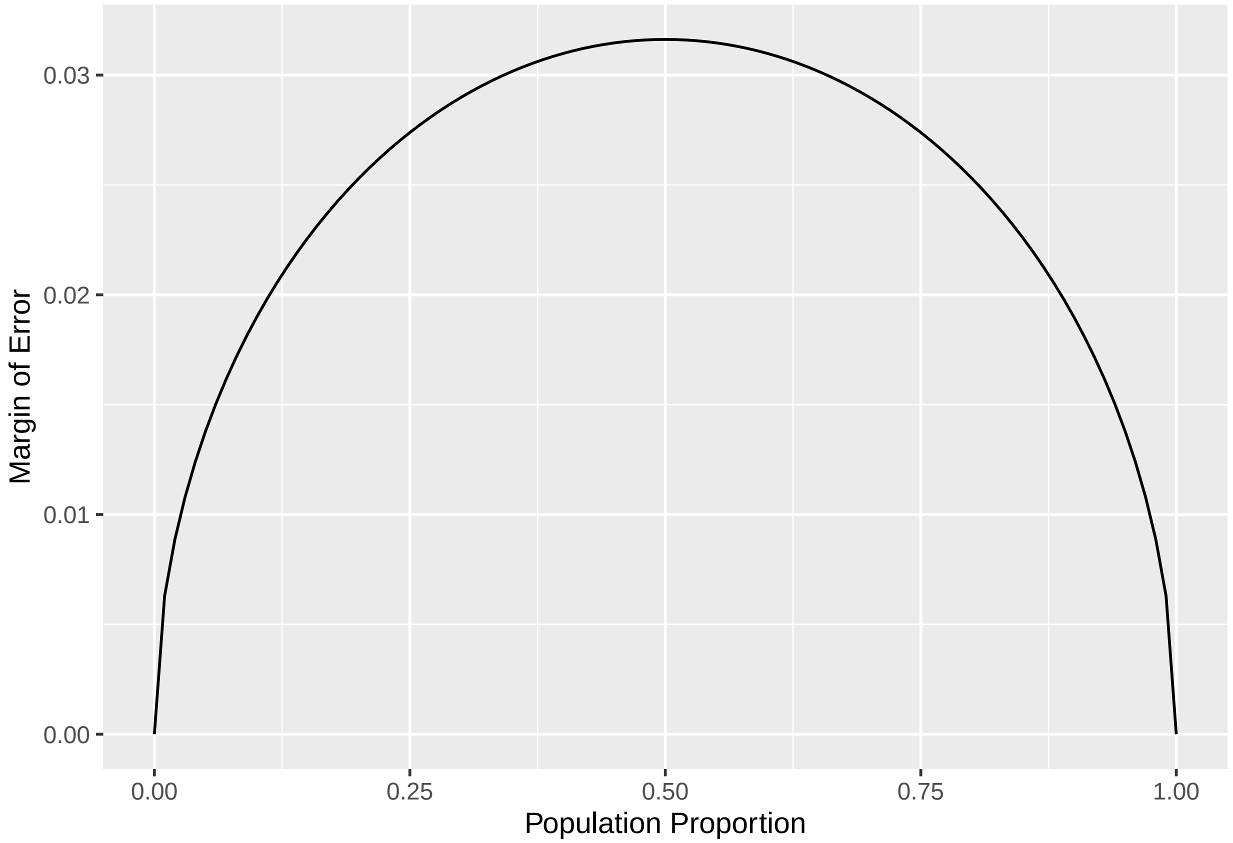
```
p <- seq(from = 0, to = 1, by = 0.01)
me <- 2 * sqrt(p * (1 - p)/n)
```

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)

p_vs_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?

```
p_vs_me
```



As p increases from 0 to 0.5, the margin of error also increases. However, as p moves from 0.5 to 1, the margin of error decreases. The margin of error is maximized when the population proportion is p = 0.5. The margin of error is minimized when the population proportion is either p = 0 or p = 1. The graph is symmetric about p = 0.5. This means the margin of error for a proportion of p is the same as for 1-p. For instance, the margin of error for a population proportion of 0.2 is the same as for a proportion of 0.8 for a given sample size. Therefore, for a given sample size, the margin of error is maximized when the population proportion is p = 0.5.

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 shape of the distribution is roughly bell-shaped or normal. This is consistent with the Central Limit Theorem, which states that for large sample sizes, the sampling distribution of the sample proportion will be approximately normally distributed. It is centered around 0.1, and the sample proportions are close to the population proportion of 0.1 implying that it has a narrow spread

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.

While changing p to be 0.3 (p = 0.3) and maintaining n = 300, we still get a bell-shaped or normal distribution, consistent with the Central Limit Theorem for large sample sizes. The center of the distribution has shifted to around 0.3. Also, the spread seems relatively narrow, indicating that the sample proportions are close to the population proportion. However, with p = 0.3, the distribution seems more symmetric than the case with p = 0.1.

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

Here, we kept p = 0.3, but changed n = 700. Comparing to the previous one (p = 0.3 and n = 300), both distributions maintain a bell-shaped or normal distribution. However, with a larger sample size, the distribution appears more symmetric. The spread of the distribution is narrower for n = 700 than n = 300. The center of the distribution remains around p = 0.3 for both cases

---

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.

```
set.seed(1111)

StrenghtT_7d <- yrbss %>%
  filter(strength_training_7d == 7) %>%
  mutate(sleep_hrs = ifelse(school_night_hours_sleep > 10, "over_10hrs", "10_or_less"))
```

```r
Prop_Stren7_Sleep10 <- StrenghtT_7d %>%
  count(sleep_hrs) %>%
  mutate(p_hat = n / sum(n)) %>%
  filter(sleep_hrs == "over_10hrs") %>%
  pull(p_hat) %>%
  round(2)
```

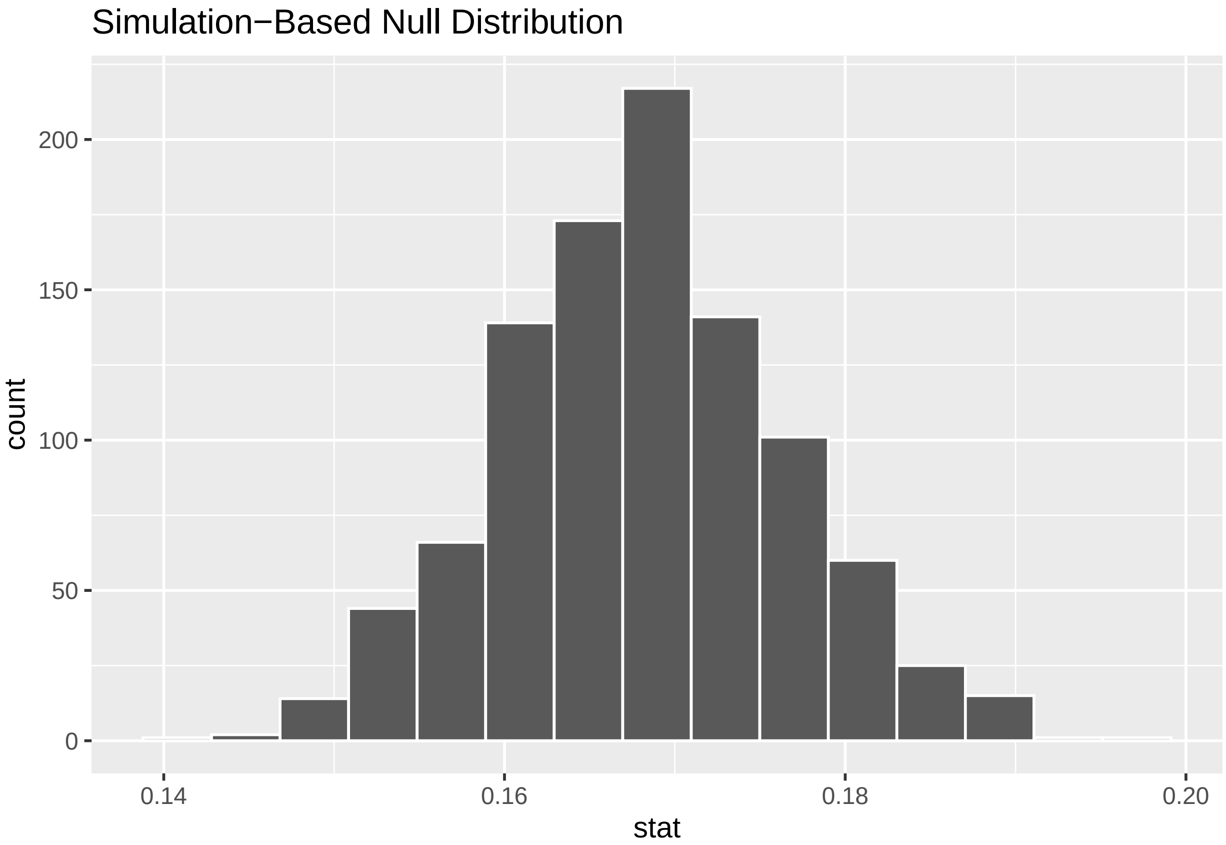
```r
overall_strength_training_prop <- yrbss %>%
  summarize(p_hat = mean(strength_training_7d == 7, na.rm = TRUE))

hypothesized_proportion <- overall_strength_training_prop$p_hat

results <- StrenghtT_7d %>%
  drop_na(sleep_hrs) %>%
  specify(response = sleep_hrs, success = "over_10hrs") %>%
  hypothesize(null = "point", p = hypothesized_proportion ) %>%
  generate(reps = 1000, type = "draw") %>%
  calculate(stat = "prop")

observe_stat <- Prop_Stren7_Sleep10

p_value <- mean(results$stat >= observe_stat)

p_value
```

```
## [1] 0
```

```r
ci <- get_ci(results, level = 0.95, type = "percentile")
ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.152    0.185
```

```r
results %>%
  visualize(obs_stat = results$stat, direction = "greater")
```



Hypotheses: Null Hypothesis (Ho): The proportion of individuals who sleep 10+ hours per day and strength train every day of the week is the same as the proportion of those who sleep less than 10 hours. Alternative Hypothesis (Ha): The proportion of individuals who sleep 10+ hours per day and strength train every day of the week is greater than the proportion of those who sleep less than 10 hours.

Conditions for Inference: Assuming the yrbss dataset represents a random sample; Assuming each student's response is independent of others; No extreme skewness observed from the visual distribution; sample size is laarge enough.

Test Results: p-value is 0, This gives very strong evidence against the null hypothesis in favor of the alternative. And, The 95% CI for the proportion of students who sleep 10+ hours and strength train every day of the week is (0.152, 0.185). This means we are 95% confident that the true proportion of students who meet these criteria lies within this interval.

Visualization: From the histogram, we observe that the simulated proportions based on the null hypothesis center around the hypothesized proportion.

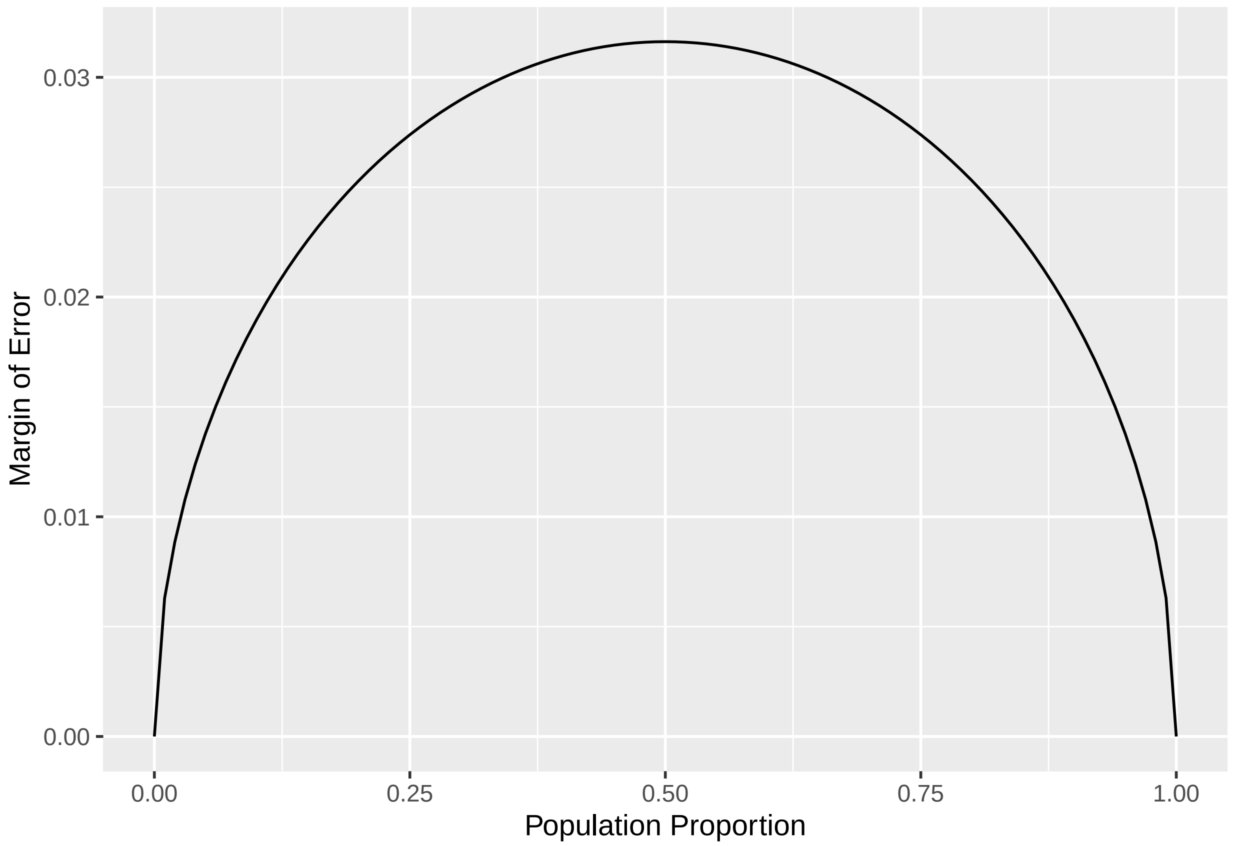
Conclusion: Given our computed p-value of 0, we have strong evidence against the null hypothesis. Our data suggests that the proportion of students who sleep 10+ hours per day and strength train every day of the week is statistically significantly different from the overall proportion of students who strength train every day of the week. The 95% confidence interval further quantifies this difference, indicating we are 95% confident that the true proportion of such students falls between 0.152 and 0.185.

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 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.

The probability of detecting a change, when in fact there isn't one, at a significance level of 0.05, 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.

```
p_vs_me
```



Margin of error = z * sqrt(p ***(1-p)/n), where z = 1.96 corresponding to 95% confidence level, p is the population proportion and n is the sample size to be determine. At p = 0.5, the margin of error seems to be at its peak, which is approximately 0.03. Since we want our margin of error to be no more than 0.01, by applying every value to its corresponding variable, we compute this inequation to solve the problem: 0.01 >= 1.96*** **sqrt(0.5(1-0.5)/n). This gives us n to be at least 97 to ensure that we are within the guidelines of a margin of error no greater than 1% with 95% confidence.**
