Inference for categorical data

```
library(tidyverse)
library(openintro)
library(infer)
```

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

```
data("yrbss")
```

```
count_t_n_d <- yrbss %>%
  group_by(text_while_driving_30d) %>%
  count()

count_t_n_d
```

```
## # A tibble: 9 x 2
## # Groups:   text_while_driving_30d [9]
##   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 values above outline the counts within each category for amount of days within the past 30 days students have texted while driving.

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

Exercise 2 Response

```
prop_text_hel <- yrbss %>%
filter(
  text_while_driving_30d == 30,
  helmet_12m == "never"
        )

count(prop_text_hel)/count(yrbss)
```

```
##            n
## 1 0.03408673
```

3.40% of people sampled texted while driving every day in the past 30 days and never wear helmets while riding a bike.

Note: the three chunks below are from the Lab 6 template.

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

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

```
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   0.0655   0.0774
```

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?

Exercise 3 Response To find the margin of error, you can use the equation

$$ME = z * \sqrt{p(1-p)/n}$$

Where: ME = margin of error, z = z-score, p = sample proportion and n = sample size

```
n <- nrow(no_helmet)
z <- 1.96 #Z-Score of 95% confidence interval
p <- count(prop_text_hel)/count(yrbss)
me <- z * sqrt(p * (1 - p)/n)
me
```

```
##             n
## 1 0.004257782
```

Based on the calculation above, the margin of error is .3%. Questions: (1) How can I change my code so I am not using the proportion calculation in the R chunk 'exercise-2-response' and instead using the no_helmet dataframe? (2) Why do we filter for no helmets instead of adding that logical statement into the ifelse() function?

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. **Exercise 4 Response The calculations below will explore the proportion of youth that were physically active for at least three days for 60+ minutes and did not watch TV.**

```r
no_TV <- c("<1", "do not watch")

prop_opposite_of_me <- yrbss %>%
mutate(resp_value = ifelse(physically_active_7d > 3 & hours_tv_per_school_day %in% no_TV, "yes", "no"))
    drop_na(resp_value)


prop_calc <- sum(prop_opposite_of_me$resp_value == "yes")/nrow(prop_opposite_of_me)

prop_opposite_of_me %>%
  specify(response = resp_value, 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    0.157    0.170
```

Setting a confidence level of 95%, the confidence interval for this sample is 0.157-0.169.

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:

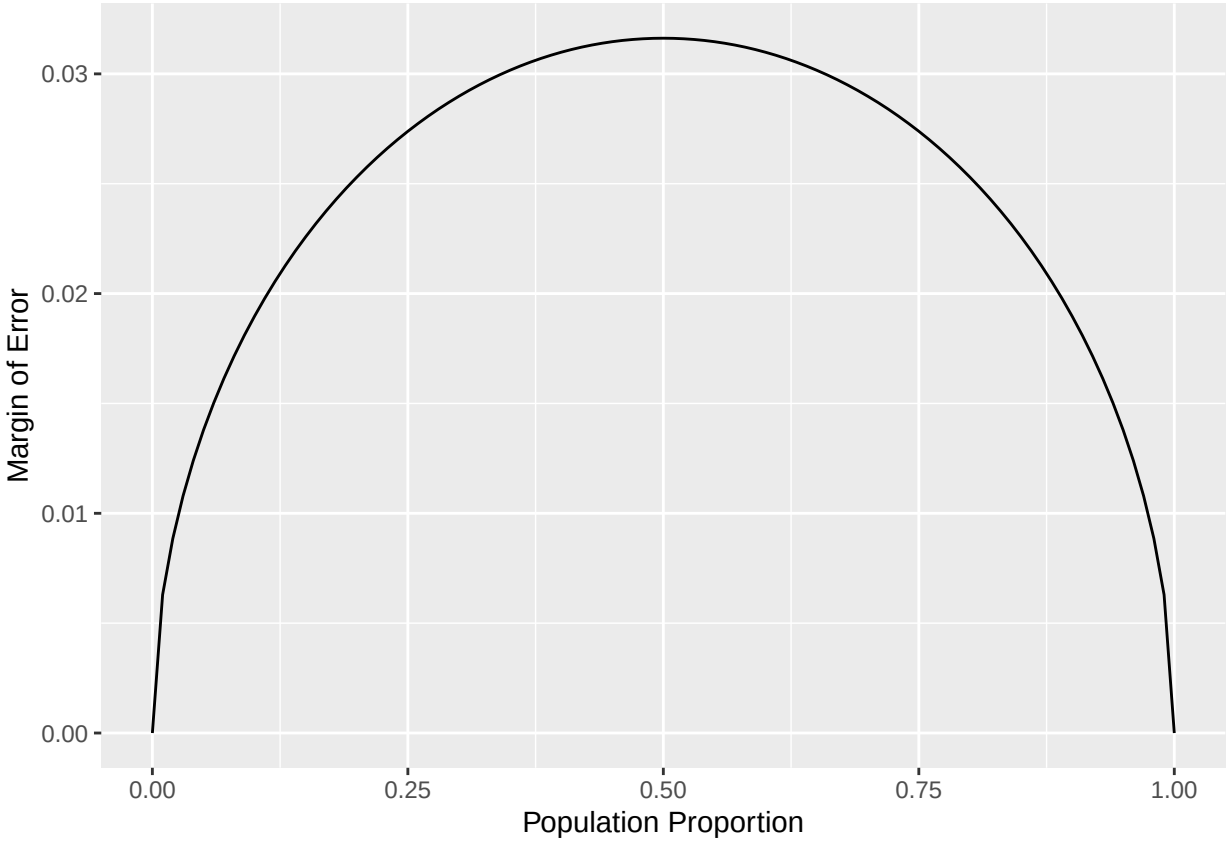
```r
n <- 1000
```

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 ($ME = 2 \times SE$).

```r
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.

```r
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? **Exercise 5 Response The margin of error and the population proportion are proportional to one another. This means that the margin of error increases as the proportion gets closer to .50, and decreases when the proportion is closer to 0 or 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.

6. Describe the sampling distribution of sample proportions at $n = 300$ and $p = 0.1$. Be sure to note the center, spread, and shape. **Exercise 6 Response The sampling distribution is normal and the center is at roughly .1. The spread is from .03-.135.**

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. **Exercise 7 Response As the proportion changes, the spread remains the same, but the center shifts to the population proportion value, p.**

8. Now also change n. How does n appear to affect the distribution of $\hat{p}$? **Exercise 8 Response When the sample size decreases, the spread is wider, however the proportion remains as the center. When the sample size increases, the spread decreases and the proportion remains the center. Roughly speaking, the distribution remains normal (as long as it is large enough, meaning larger than 50 observations).**

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. **Exercise 9 Response**

```
prop_from_yrbss <- yrbss %>%
  filter(school_night_hours_sleep == "10+",
         strength_training_7d == 7
         )

count(prop_from_yrbss)/count(yrbss)
```

```
##             n
## 1 0.006184201
```

```
sleep10hrs <- yrbss %>%
  filter(school_night_hours_sleep == "10+")

prop_strength_tren_sleep10 <- sleep10hrs %>%
  mutate(resp_value = ifelse(strength_training_7d == "7", "yes", "no"))

prop_strength_tren_sleep10 %>%
  drop_na(strength_training_7d) %>%
  specify(response = resp_value, 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    0.221    0.321
```

The null hypothesis is that students who sleep over 10+ hours do not strength train more than students who sleep less than 10 hours. The alternative hypothesis is that students who sleep more than 10 hours are more likely to strength train than students who do not.

Question: The proportion of those who strength train every day and sleep over 10 hours a night is ,618%. However, the bootstrap sample provided a confidence interval of .224-.321. Does this mean that we should reject the alternate hypothesis?

Note: I am finding this exercise difficult to calculate correctly. Should there be further analysis to support this hypothesis? What types of calculations should be completed to analyze if students that sleep more than 10 hours a night are more likely to strength train?

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. **Exercise 10 A Type 1 Error is rejecting the null hypothesis when it is actually true. On page 198 of the OpenIntro Statistics textbook, it describes that the significance level can be used to mitigate risk if making a Type I error would be detrimental. The probability that you could detect a change by chance if the significance level is 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.

Exercise 11 Response

```
me <- .01
z <- 1.96 #Z-Score of 95% confidence interval
p <- .05
n <- (z^2) * (p*(1-p))/(me^2)
n
```

```
## [1] 1824.76
```

The sample size should be at least 1825 residents.
