# Data 606 - Lab 6

Cameron Smith

2020-10-11

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

**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?*

Please see output of code block below.

```
str(yrbss)
```

```
## tibble [13,583 x 13] (S3: tbl_df/tbl/data.frame)
##  $ age                     : int [1:13583] 14 14 15 15 15 15 15 14 15 15 ...
##  $ gender                  : chr [1:13583] "female" "female" "female" "female" ...
##  $ grade                   : chr [1:13583] "9" "9" "9" "9" ...
##  $ hispanic                : chr [1:13583] "not" "not" "hispanic" "not" ...
##  $ race                    : chr [1:13583] "Black or African American" "Black or African American" "
##  $ height                  : num [1:13583] NA NA 1.73 1.6 1.5 1.57 1.65 1.88 1.75 1.37 ...
##  $ weight                  : num [1:13583] NA NA 84.4 55.8 46.7 ...
##  $ helmet_12m              : chr [1:13583] "never" "never" "never" "never" ...
##  $ text_while_driving_30d  : chr [1:13583] "0" NA "30" "0" ...
##  $ physically_active_7d    : int [1:13583] 4 2 7 0 2 1 4 4 5 0 ...
##  $ hours_tv_per_school_day : chr [1:13583] "5+" "5+" "5+" "2" ...
##  $ strength_training_7d    : int [1:13583] 0 0 0 0 1 0 2 0 3 0 ...
##  $ school_night_hours_sleep: chr [1:13583] "8" "6" "<5" "6" ...
```

```
yrbss %>% count(text_while_driving_30d)
```

```
## # 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
```

### Exercise 2

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

Approximately 6.6%.

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

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

n1 <- nrow(no_helmet)
p1 <- nrow(filter(no_helmet, text_ind == "yes")) / n1

# Answer to the question, compared to all of sample population w/out helmets
p1
```

```
## [1] 0.0663609
```

### 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?*

.005, as indicated in the code block below.

```
no_helmet %>%
  specify(response = text_ind, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)
```

```
## Warning: Removed 474 rows containing missing values.
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1   0.0647   0.0780
```

```
# Standard error function (so I can re-use it)
se_function <- function(p, n) {
  sqrt((p*(1-p))/n)
}

se <- se_function(p1,n1)

# Margin of error calculation
1.96 * se
```

```
## [1] 0.005840733
```

**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.*

The categorical variables I chose are (1) # hours of tv watched per day; and (2) # of times physically active in the last 7 days.

The CI for those that do not watch TV is .855 to .866, and the margin of error is .006

The CI for those that are not physically active is .157 to .169, with a margin of error of .006.

```
# Confidence interval and margin of error for proportion that never watches tv
tv_watchers <- yrbss %>%
  mutate(no_tv = ifelse(hours_tv_per_school_day == "do not watch", "no", "yes"))

# Confidence interval
tv_watchers %>%
  specify(response = no_tv, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)
```

```
## Warning: Removed 338 rows containing missing values.
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.855    0.867
```

```
n2 <- nrow(tv_watchers)
p2 <- nrow(filter(tv_watchers, no_tv == "yes")) / n2
se2 <- se_function(p2,n2)
me2 <- 1.96 * se2

# Margin of error
me2
```

```
## [1] 0.006170766
```

```
# Confidence interval and margin of error for proportion that never exercises
physically_active <- yrbss %>%
  mutate(not_active = ifelse(physically_active_7d == "0", "yes", "no"))

# Confidence interval
physically_active %>%
  specify(response = not_active, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)
```

```
## Warning: Removed 273 rows containing missing values.
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.158    0.169
```

```
n3 <- nrow(physically_active)
p3 <- nrow(filter(physically_active, not_active == "yes")) / n2
se3 <- se_function(p3,n3)
me3 <- 1.96 * se3

# Margin of error
me3
```

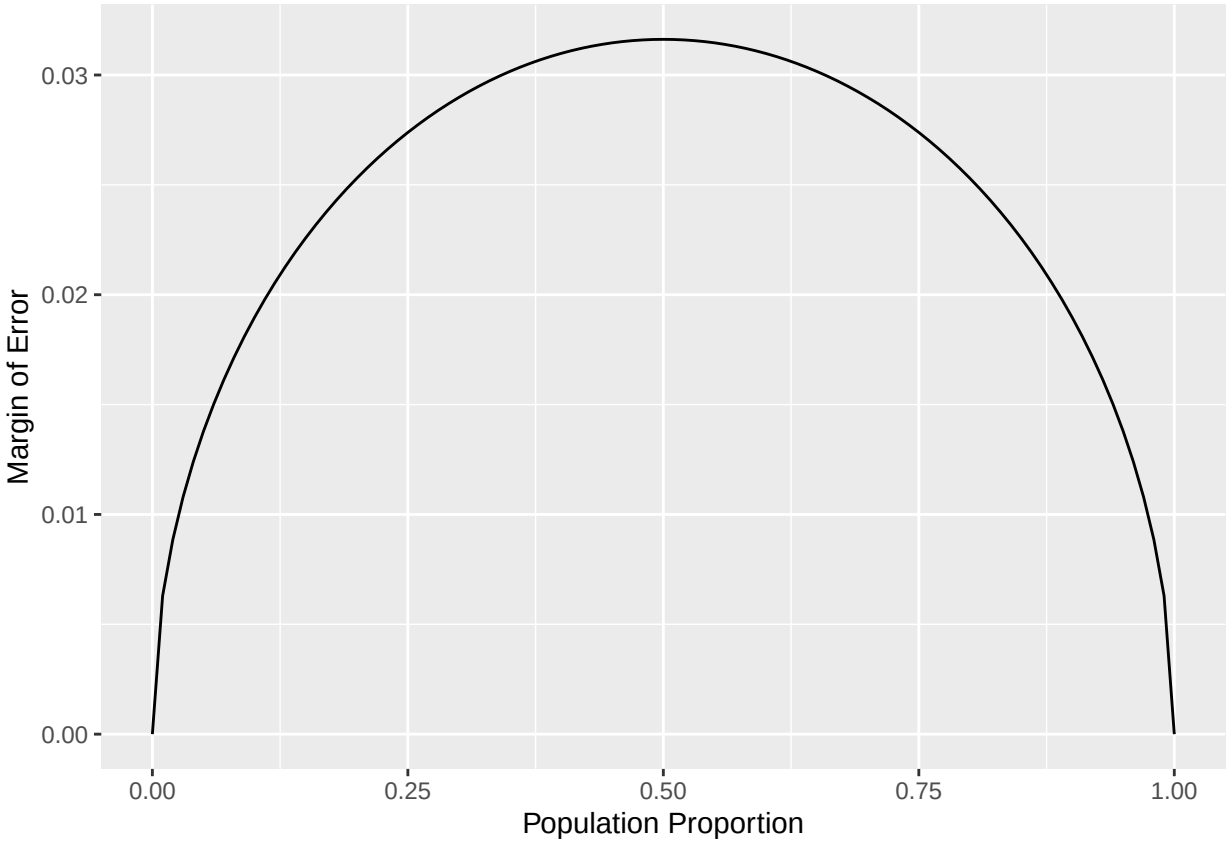
```
## [1] 0.006163882
```

### 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?*

As P increases, so does the ME up to .5 at which point it is maximized (50% of the population), and from there it is an inverse relationship of the same proportions.

```
n <- 1000
p <- seq(from = 0, to = 1, by = 0.01)
me <- 2 * sqrt(p * (1 - p)/n)
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")
```




**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.*

The distribution appears fairly normal. It is unimodal and not heavily skewed to the left or the right.

**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.*

As p increases the overall shape, center remain similar but the spread gets much narrower up until .5 when the spread is quite high and then it narrows down again from .6 to 1.0 (basically a reflection of the chart in exercise 5).

**Exercise 8**

*Now also change n. How does n appear to affect the distribution of p^?*

The distribution appears to become 'more normal', i.e. more uniform with a clear center and less outliers as n increases.

### 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.*

Conditions for inference:

- Independent, extended: data are independent w/in and between the 2 groups: YES
- At least 10 observed successes and failures in the two groups: YES

NOTE: The book seems to use a different formula (pooled proportion * n) rather than just n for this. I did this below, with the same result, though it is unclear to me why the formula in the lecture and that in the book are different.

Null hypothesis is that whether someone gets more or less than 10 hours of sleep does not impact the likelihood of whether they will strength train every day, i.e. p1 - p2 = 0.

```
strength_train <- yrbss %>%
  mutate(train_daily = ifelse(strength_training_7d == 7, "yes", "no")) %>%
  mutate(sleep_lots = ifelse(school_night_hours_sleep == "10+", "yes", "no"))
strength_train <- strength_train %>% select(train_daily, sleep_lots)

nsleepmore <- strength_train %>% filter(sleep_lots == "yes", train_daily == "yes") %>% nrow()
nsleepless <- strength_train %>% filter(sleep_lots == "no", train_daily == "yes") %>% nrow()

psleepmore <- nsleepmore / nrow(strength_train)
psleepless <- nsleepless / nrow(strength_train)

ppooled <- (nsleepmore + nsleepless) / nrow(strength_train)

# Standard error calculation
var1 <- ppooled * (1 - ppooled) / nsleepmore
var2 <- ppooled * (1 - ppooled) / nsleepless
se_both <- sqrt(var1 + var2)

# Compute a test statistic and p value
pt_estimate <- (psleepmore - psleepless)
Z <- (pt_estimate - 0) / se_both
P <- 2 * pnorm(Z)
P
```

```
## [1] 0.0005312568
```

```
# Conclusion
paste("The pvalue is", round(P,3), "and thus the null hypothesis can be rejected.")
```

```
## [1] "The pvalue is 0.001 and thus the null hypothesis can be rejected."
```

### 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.*

A type 1 error is known as a false positive, and given tat the significance level if .05 there is a 5% chance of a type 1 error.

### 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.*

We would need a sample of 9,604. See calculation below.

```
# Using .5 since we don't know any better, need to solve for n
# .01 = 1.96 * sqrt( (.5*.5) / n )
required_sample <- (1.96^2 * .5^2) / (.01^2)
required_sample
```

```
## [1] 9604
```