Code ▾

DS606-Lab 6

George Cruz

2020-11-01

#Inference for categorical data

Hide

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

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?

Hide

```
data(yrbss)

text_while_driving <- yrbss %>%
  filter(text_while_driving_30d == 30)

summary(text_while_driving)
```

```
##       age           gender             grade             hispanic
##  Min.   :12.00   Length:827         Length:827         Length:827
##  1st Qu.:16.00   Class :character   Class :character   Class :character
##  Median :17.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :16.92
##  3rd Qu.:18.00
##  Max.   :18.00
##  NA's   :5
##      race               height          weight        helmet_12m
##  Length:827         Min.   :1.370   Min.   : 38.56   Length:827
##  Class :character   1st Qu.:1.630   1st Qu.: 58.97   Class :character
##  Mode  :character   Median :1.730   Median : 69.85   Mode  :character
##                     Mean   :1.721   Mean   : 72.66
##                     3rd Qu.:1.800   3rd Qu.: 81.65
##                     Max.   :2.110   Max.   :180.99
##                     NA's   :71      NA's   :71
##  text_while_driving_30d physically_active_7d hours_tv_per_school_day
##  Length:827             Min.   :0.000        Length:827
##  Class :character       1st Qu.:2.000        Class :character
##  Mode  :character       Median :5.000        Mode  :character
##                         Mean   :4.269
##                         3rd Qu.:7.000
##                         Max.   :7.000
##                         NA's   :16
##  strength_training_7d school_night_hours_sleep
##  Min.   :0.000        Length:827
##  1st Qu.:0.000        Class :character
##  Median :3.000        Mode  :character
##  Mean   :3.447
##  3rd Qu.:7.000
##  Max.   :7.000
##  NA's   :78
```

Hide

```
twd_counts <- yrbss %>%
  count(text_while_driving_30d) %>%
  mutate(p = n /sum(n))

twd_counts
```

```
## # A tibble: 9 x 3
##   text_while_driving_30d     n      p
##   <chr>                  <int>  <dbl>
## 1 0                       4792 0.353
## 2 1-2                      925 0.0681
## 3 10-19                    373 0.0275
## 4 20-29                    298 0.0219
## 5 3-5                      493 0.0363
## 6 30                       827 0.0609
## 7 6-9                      311 0.0229
## 8 did not drive           4646 0.342
## 9 <NA>                     918 0.0676
```

Exercise 2

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

Hide

```
no_helmet <- yrbss %>%
  filter(helmet_12m == "never") %>%
  count(text_while_driving_30d) %>%
  mutate(p = n /sum(n))

# no_helmet

percent(no_helmet$p[6], accuracy = 0.01)
```

```
## [1] "6.64%"
```

6.64% of those who drove while texting also reported no helmet.

Exercise 3

Hide

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

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

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   0.0655   0.0774
```

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?

Hide

```
p <- 463/6977
cv <- 1.96
n <- 6977

me <- 1.96 * sqrt((p*(1-p))/n)

me
```

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

ME = 0.00584

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.

Hide

```
good_sleep <- yrbss %>%
  mutate(slept_well = ifelse(school_night_hours_sleep > 5, "yes", "no") )

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

```
## Warning: Removed 1248 rows containing missing values.
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.769    0.784
```

Hide

```
total_slept <- good_sleep %>%
  filter(!is.na(slept_well)) %>%
  count(slept_well) %>%
  mutate( p = n/sum(n))

p <- 0.7761654
cv <- 1.96
n <- 12335

me <- 1.96 * sqrt((p*(1-p))/n)

me
```

```
## [1] 0.007355755
```

ME= 0.007355

Hide

```
no_tv <- yrbss %>%
  mutate(did_not_watch = ifelse(hours_tv_per_school_day == "do not watch", "yes", "no"
         ) )

no_tv %>%
  specify(response = did_not_watch, 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    0.133    0.145
```

Hide

```
total_no_tv <- no_tv %>%
  filter(!is.na(did_not_watch)) %>%
  count(did_not_watch) %>%
  mutate( p = n/sum(n))

p <- 0.1389203
cv <- 1.96
n <- 13245

me <- 1.96 * sqrt((p*(1-p))/n)

me
```

```
## [1] 0.005890262
```

ME = 0.00589

Exercise 5

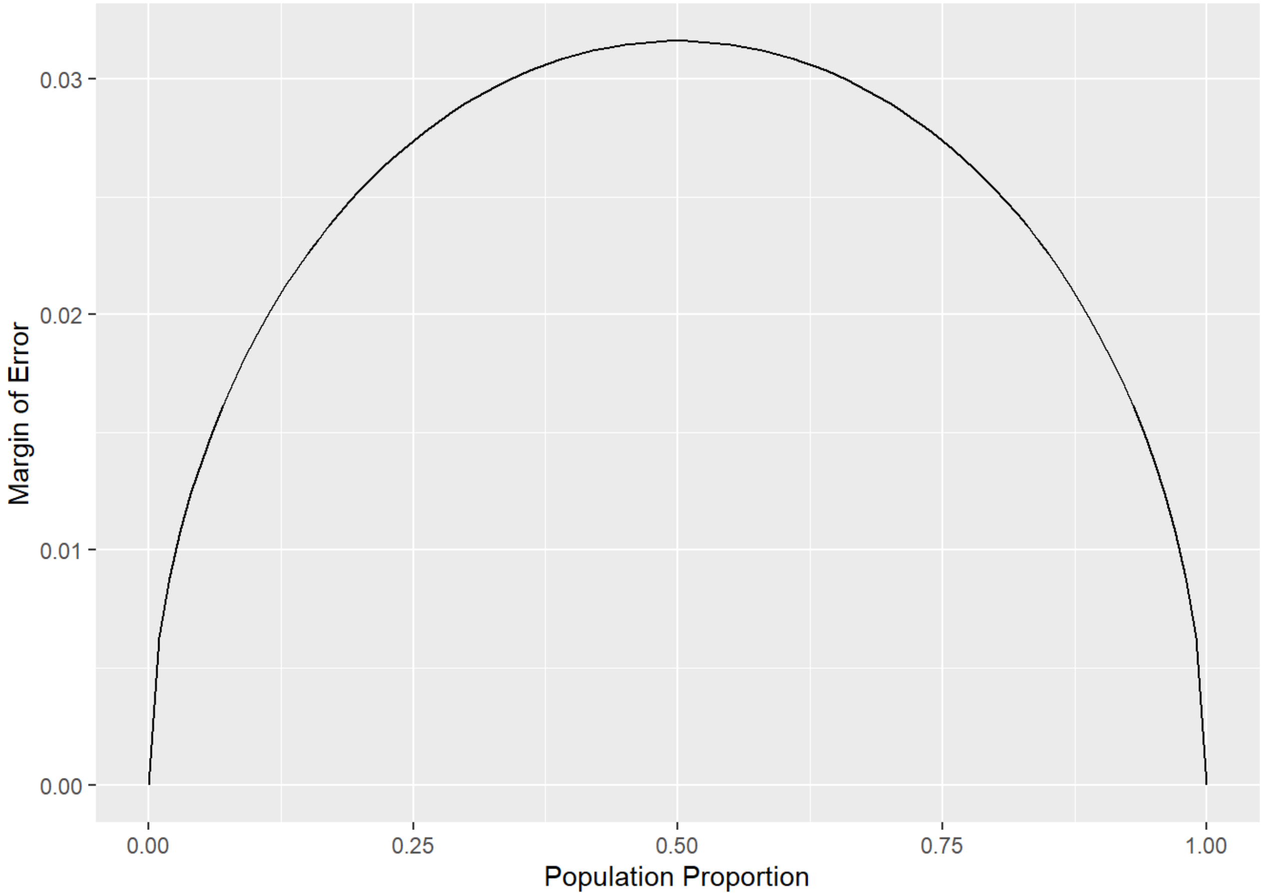
Describe the relationship between p and me. Include the margin of error vs. population proportion plot you constructed in your answer. For a given sample size, for which value of p is margin of error maximized?

Hide

```
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")
```



The margin of error appears to increase as the population proportion increases until the population proportion **reaches 50%**, from that point on, the margin of error decreases as the ME increases.

Exercise 6

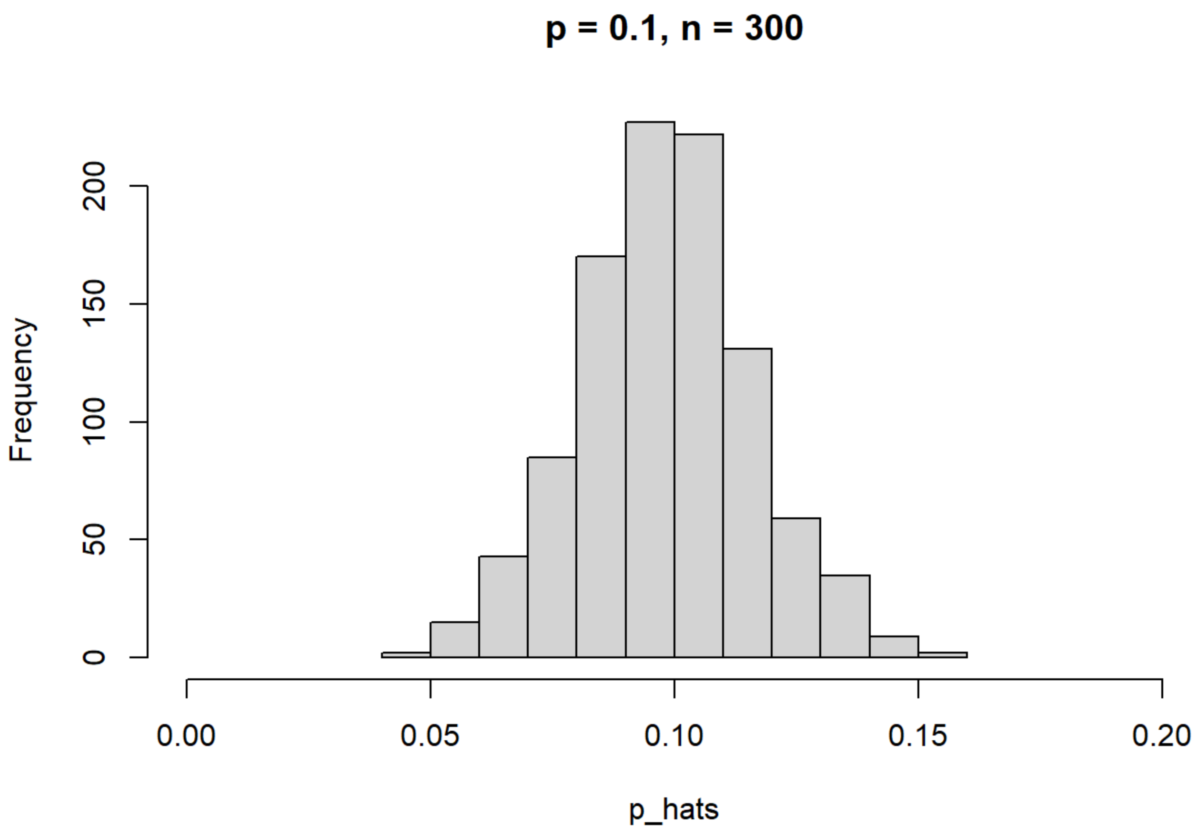
Describe the sampling distribution of sample proportions at n=300 and p=0.1. Be sure to note the center, spread, and shape.

Hide

```
p <- 0.1
n <- 300
p_hats <- rep(0, 1000)

for (i in 1:1000) {
    samp <- sample(c("A", "B"), n, replace = TRUE,
        prob = c(p, 1 - p))
    p_hats[i] <- sum(samp == "A")/n
}

hist(p_hats, main = "p = 0.1, n = 300", xlim = c(0, 0.2))
```



The sampling distribution appears normal, unimodal, centered at around 0.1 with values between 0.04 and 0.16.

Exercise 7

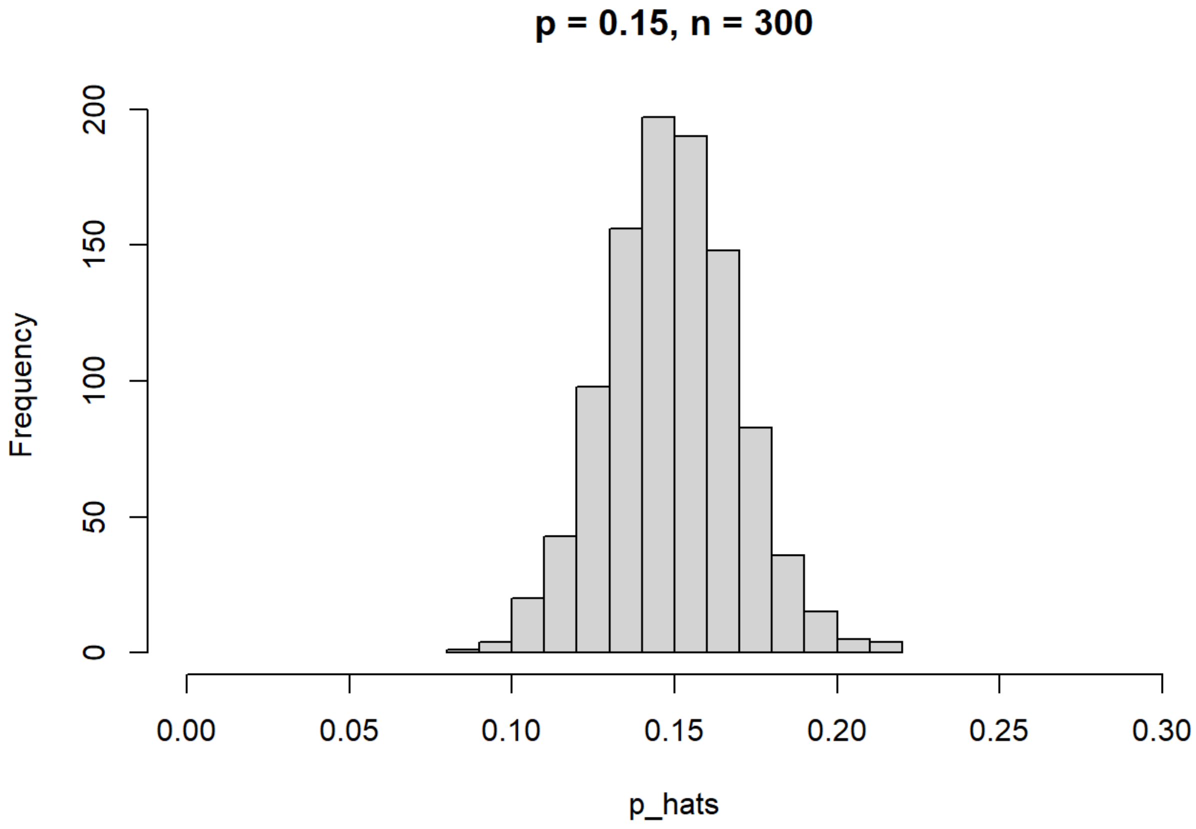
Keep n constant and change p. How does the shape, center, and spread of the sampling distribution vary as p changes. You might want to adjust min and max for the x-axis for a better view of the distribution.

Hide

```
p <- 0.15
n <- 300
p_hats <- rep(0, 1000)

for (i in 1:1000) {
    samp <- sample(c("A", "B"), n, replace = TRUE,
        prob = c(p, 1 - p))
    p_hats[i] <- sum(samp == "A")/n
}

hist(p_hats, main = "p = 0.15, n = 300", xlim = c(0, 0.30))
```



The spread and center of the distribution move with p. When I tried a lower p, the shape of the distribution changed slightly but still appeared normal and unimodal.

Exercise 8

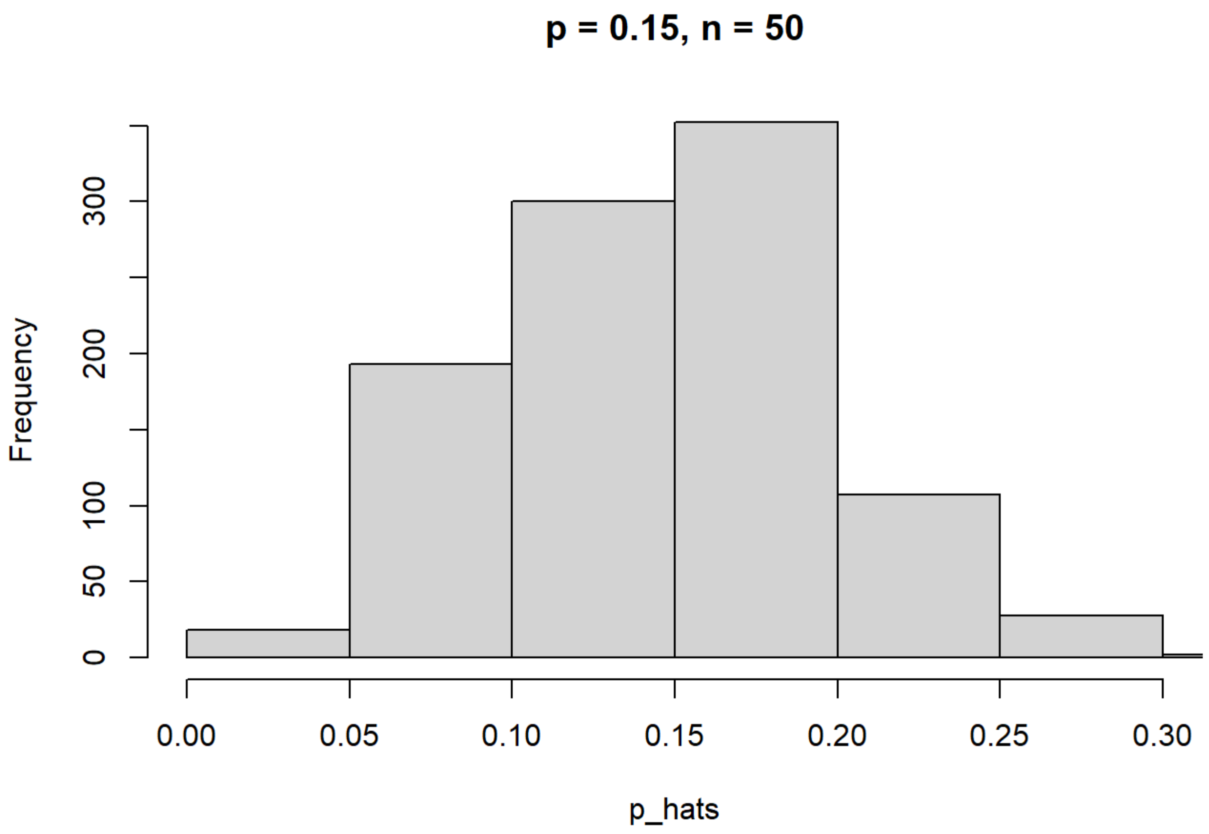
Now also change n. How does n appear to affect the distribution of p^?

Hide

```
p <- 0.15
n <- 50
p_hats <- rep(0, 1000)

for (i in 1:1000) {
    samp <- sample(c("A", "B"), n, replace = TRUE,
        prob = c(p, 1 - p))
    p_hats[i] <- sum(samp == "A")/n
}

hist(p_hats, main = "p = 0.15, n = 50", xlim = c(0, 0.30))
```



As n lowers, the distribution’s spread grows. If n increases the spread gets smaller.

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.

Hide

```
slept_10 <- yrbss %>%
  mutate(slept_10 = ifelse(school_night_hours_sleep == "10+", "yes", "no")) %>%
  filter(!is.na(slept_10))

total_slept <- slept_10 %>%
  count(slept_10) %>%
  mutate( p = n/sum(n))

p <- 0.02561816

slept_10 %>%
  specify(response = slept_10, 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   0.0230   0.0282
```

Hide

```
n <- nrow(slept_10)
z <- 1.96
se <- z*sqrt((p*(1-p))/n)

me_sleep<- z * se
me_sleep
```

```
## [1] 0.005464886
```

We’re 95% confident the mean of people that sleep 10 hours or more is between 0.0229 and 0.0285 with a margin of error of 0.0054.

Hide

```
strength_trained <- yrbss %>%
  mutate(trained = ifelse(strength_training_7d == 7, "yes", "no")) %>%
  filter(!is.na(trained))

total_trained <- strength_trained %>%
  count(trained) %>%
  mutate( p = n/sum(n))

p <- 0.1680503

strength_trained %>%
  specify(response = trained, 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    0.161    0.175
```

Hide

```
n <- nrow(strength_trained)
z <- 1.96
se <- z*sqrt((p*(1-p))/n)

me_trained<- z * se
me_trained
```

```
## [1] 0.01289576
```

We’re 95% confident the mean of people that strength trained everyday is between 0.1611 and 0.01744 with a margin of error of 0.0129.

There doesn’t seem to be conclusive evidence to state that those that slept over 10 hours are more likely to strength train as the numbers are very disimilar.

Exercise 10

There would be a 5% chance of detecting a change. A type 1 error is also known as a false positive and occurs when a researcher incorrectly rejects a true null hypothesis. This means that your report that your findings are significant when in fact they have occurred by chance.

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.

Hide

```
# margin of error is maximized when p = 0.5
p <- 0.5
me <- 0.01
z <- 1.96 #95% confidence cv

n <- (p ** 2)/((me/1.96)** 2)

n
```

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

I would need to sample 9,604 people.

…