Data_606_Lab_6_Inference for categorical data.rmd

Enid Roman

2022-10-18

```
set.seed(500)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(openintro)
```

```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
```

```
library(infer)
```

```
data('yrbss', package='openintro')
print(head(yrbss,2))
```

```
## # A tibble: 2 x 13
##     age gender grade hispa~1 race  height weight helme~2 text_~3 physi~4 hours~5
##   <int> <chr>  <chr> <chr>   <chr>  <dbl>  <dbl> <chr>   <chr>     <int> <chr>
## 1    14 female 9     not     Blac~     NA     NA never   0             4 5+
## 2    14 female 9     not     Blac~     NA     NA never   <NA>          2 5+
## # ... with 2 more variables: strength_training_7d <int>,
## #   school_night_hours_sleep <chr>, and abbreviated variable names 1: hispanic,
## #   2: helmet_12m, 3: text_while_driving_30d, 4: physically_active_7d,
## #   5: hours_tv_per_school_day
```

Exercise 1

The counts within each category for the amount of days these students have texted while driving within the past 30 days are:

4792 have reported 0 days.

4646 have reported did not drive.

925 have reported drive 1 to 2 days.

918 have reported NA days

827 have reported 30 days.

493 have reported 3 to 5 days.

373 have reported 10 to 19 days.

311 have reported 6 to 9 days.

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

298 have reported 20 to 29 days.

```
## # 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
```

Exercise 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)
```

6.64% (463/6977) (if counting the NA) is the proportion of people who have texted while driving every day in the past 30 days and never wear helmets. (See below)

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

```
no_helmet %>%
filter(!is.na(text_ind)) %>%
filter(helmet_12m == "never") %>%
filter(text_ind == "yes") %>%
nrow() / nrow(no_helmet)
```

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

```
no_helmet %>%
filter(text_ind != "") %>%
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   0.0775
```

Exercise 3

We are 95% confident that the proportion of non-helmet wearers that have texted while driving each day for the past 30 days based on this survey is between 6.57% and 7.75%.

```


1.96 * sqrt((0.0775*(1-.0775)/6977))
```

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 is .0063 or 6%

```
## [1] 0.006274162
```

Exercise 4

Hours TV Per School Day

Proportion of Interest of students who reported watching less than 1 hr of tv per school day is 15.96% (2168/13583).

We are 95% confident that the proportion of students who reported watching less than 1 hr of tv per school day is between 15.72% and 17.04%.

The margin of error for the estimate of the proportion of students who reported watching less than 1 hr of tv per school day is .0064 or 6%.

School Night Hours of Sleep

Proportion of Interest of students who reported less than 5hrs of sleep on school nights is 7.15% (965/12335).

We are 95% confident that the proportion of students who reported less than 5hrs of sleep on school nights is between 7.36% and 8.30%.

```
yrbss %>%
  count(hours_tv_per_school_day, sort=TRUE)
```

The margin of error for the estimate of the proportion of students who reported less than 5hrs of sleep on school nights is .0048 or 5%.

```
## # A tibble: 8 x 2
##   hours_tv_per_school_day     n
##   <chr>                   <int>
## 1 2                        2705
## 2 <1                       2168
## 3 3                        2139
## 4 do not watch             1840
## 5 1                        1750
## 6 5+                       1595
## 7 4                        1048
## 8 <NA>                      338
```

```
tv<- yrbss %>%
  filter(!is.na(hours_tv_per_school_day)) %>%
  mutate(tv_ind = ifelse(hours_tv_per_school_day == "<1", "yes", "no"))

tv %>%
  count(tv_ind)
```

```
## # A tibble: 2 x 2
##   tv_ind     n
##   <chr>  <int>
## 1 no     11077
## 2 yes     2168
```

```
tv %>%
    specify(response = tv_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    0.157    0.170
```

```
1.96 * sqrt((.1704*(1-.1704)/13245))
```

```
## [1] 0.006403232
```

```


yrbss %>%
  count(school_night_hours_sleep, sort=TRUE)
```

School Night Hours Sleep

```
## # A tibble: 8 x 2
##   school_night_hours_sleep     n
##   <chr>                    <int>
## 1 7                         3461
## 2 8                         2692
## 3 6                         2658
## 4 5                         1480
## 5 <NA>                      1248
## 6 <5                         965
## 7 9                          763
## 8 10+                        316
```

```
sleep <- yrbss %>%
  filter(!is.na(school_night_hours_sleep)) %>%
  mutate(sleep_ind = ifelse(school_night_hours_sleep == "<5", "yes", "no"))

sleep %>%
  count(sleep_ind)
```

```
## # A tibble: 2 x 2
##   sleep_ind     n
##   <chr>     <int>
## 1 no        11370
## 2 yes         965
```

```
sleep %>%
 specify(response = sleep_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   0.0736   0.0831
```

```
1.96 * sqrt((.0831*(1-.0831)/12335))
```

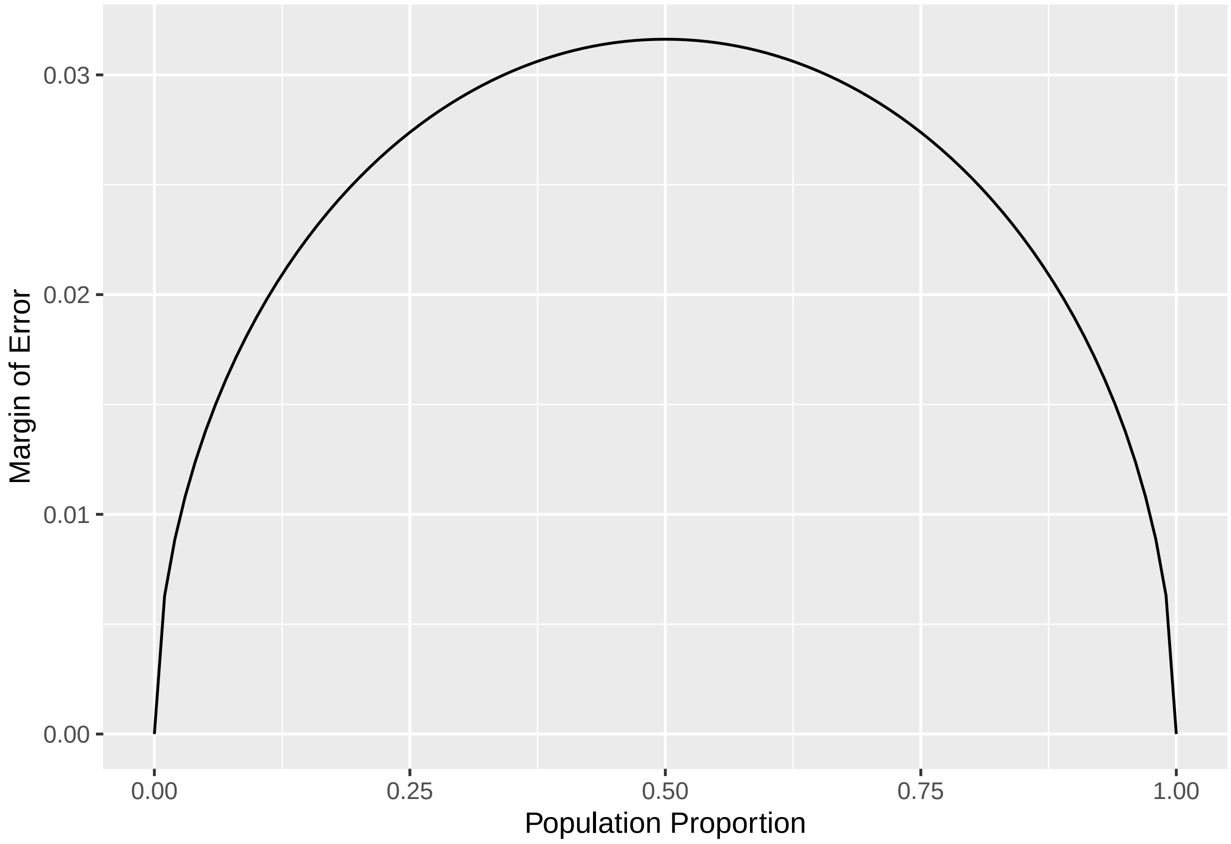
```
## [1] 0.004871335
```

```
n <- 1000
```

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

```
##   [1] 0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10 0.11 0.12 0.13 0.14
##  [16] 0.15 0.16 0.17 0.18 0.19 0.20 0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28 0.29
##  [31] 0.30 0.31 0.32 0.33 0.34 0.35 0.36 0.37 0.38 0.39 0.40 0.41 0.42 0.43 0.44
##  [46] 0.45 0.46 0.47 0.48 0.49 0.50 0.51 0.52 0.53 0.54 0.55 0.56 0.57 0.58 0.59
##  [61] 0.60 0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69 0.70 0.71 0.72 0.73 0.74
##  [76] 0.75 0.76 0.77 0.78 0.79 0.80 0.81 0.82 0.83 0.84 0.85 0.86 0.87 0.88 0.89
##  [91] 0.90 0.91 0.92 0.93 0.94 0.95 0.96 0.97 0.98 0.99 1.00
```

```
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 5

The relationship between p and me is the margin of error increases as the population proportion increases. Margin of error is greatest at the population of 50%. For a given sample size for which a value of p is margin of error maximized is at .5.

Exercise 6 (Not sure which app you were referring to)

```
p <- 0.1
n <- 300

(p*(1-p)/n)^.5
```

The distribution of sampling proportions with sampling size of 300 and p=0.1, center is at .01 and spread conformsis .08 to .11.

```
## [1] 0.01732051
```

```
.1-(p*(1-p)/n)^.5
```

```
## [1] 0.08267949
```

```
.1+(p*(1-p)/n)^.5
```

```
## [1] 0.1173205
```

Exercise 7

```
p <- 0.5
n <- 300

(p*(1-p)/n)^.5
```

Keeping the n constant and changing p the shape and center does change. The spread of the sampling distribution does vary as p changes. The distribution of sampling proportions with sampling size of 300 and p=0.5, center is at .03 and spread conformsis .07 to .12. Increasing the p does increase the spread.

```
## [1] 0.02886751
```

```
.1-(p*(1-p)/n)^.5
```

```
## [1] 0.07113249
```

```r
.1+(p*(1-p)/n)^.5
```

```
## [1] 0.1288675
```

Exercise 8

```r
p <- 0.5
n <- 400

(p*(1-p)/n)^.5
```

Keeping the n constant and changing p the shape and center does not change much. The spread of the sampling distribution does vary as p changes. The distribution of sampling proportions with sampling size of 300 and p=0.5, center is at .03 and spread conformsis .08 to .13. Increasing the n does increase the spread a little.

```
## [1] 0.025
```

```r
.1-(p*(1-p)/n)^.5
```

```
## [1] 0.075
```

```r
.1+(p*(1-p)/n)^.5
```

```
## [1] 0.125
```

Exercise 9

```r
sleep_less_than_10 <- yrbss %>%
  filter(school_night_hours_sleep != "10+")

sleep_less_than_10 %>%
  mutate(physical = ifelse(physically_active_7d == 7, "yes", "no")) %>%
  drop_na(physical) %>%
  specify(response = physical, success = "yes") %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "prop") %>%
  get_ci(level = 0.95)
```

Yes, there is convincing evidence that those who sleep 10+ hours per day are more likely to strength train every day of the week. The confidence interval range for sleep less than 10 is from 26.04% to 27.67%. The confidence interval range for sleep more than 10 is from 31.63% to 41.86%. Is a aprox. 5% to 15% difference.

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.260    0.277
```

```
sleep_10plus <- yrbss %>%
  filter(school_night_hours_sleep == "10+")

sleep_10plus %>%
  mutate(physical = ifelse(physically_active_7d == 7, "yes", "no")) %>%
  drop_na(physical) %>%
  specify(response = physical, 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    0.316    0.419
```

Exercise 10

```


mu <- 2.5
(cv <- qnorm(0.05,
             mean=0,
             sd=1,
            lower.tail=FALSE))
```

The probablity that you could detect a change (at a significance level of 0.05) simply by chance would be 62%.

```
## [1] 1.644854
```

```
pnorm(mu, mean=0, sd=1, lower.tail=FALSE)
```

```
## [1] 0.006209665
```

Exercise 11

With and estimate margin of error no greater than 1% with 95% confidence. I would have to sample 9694 people to ensure that you I am within the guidelines.

ME <- 1.96 * SE for 95% confidence margin

SE <- sqrt(p*(1 - p)/n)

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

ME^2 <- 1.96^2 * p*(1 - p)/n

```


ME <- 0.01
p <- 0.5

1.96^2 * p *(1 - 0.5)/ME^2
```

n <- 1.96^2 * p*(1 - p)/ME^2

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