Lab 7 - Inference for categorical data

Sanaz Saadatifar

3/16

Lab report

Load data

```
atheism <- read_csv("https://dyurovsky.github.io/85309/data/lab7/atheism.csv")
```

```
## Rows: 88032 Columns: 3
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): nationality, response
## dbl (1): year
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Exercise 1:

numbers reported in the first paragraph are sample statistics derived from a poll not population parameters which is nearly impossible to actually measure everyone.

Exercise 2:

```
us12 <- atheism %>%
  filter(nationality == "United States" & year == "2012")


us_atheist12 <- us12 %>%
  summarise(atheist_prop = mean(response == "atheist")) %>%
  pull()
```

we found that proportion of atheist was 0.049. that's very close to the 5% reported in the table.

Exercise 3:

it is ok to use normal distribution because all the condition listed below are met: observations need to be independent of each other which it is accurate here based on Gallup's description. at least 10 people in the sample need to be atheist, and t least 10 need to be non-atheist and that holds too.

Exercise 4:

```
#mean <- us12 %>%
#  summarise(atheist = mean(response == "atheist")) %>% # the mean of the sample
#  pull()
z_star <- qnorm(.975) #The 97.5% percentile of the normal distribution

se <- sqrt((us_atheist12 * (1 - us_atheist12)) / nrow(us12)) #the formula for the standard error

me <- se * z_star # margin of error is standard error times critical value

ci_95 <- c(us_atheist12 - me, us_atheist12 + me)

me
```

```
## [1] 0.01348187
```

```
ci_95
```

```
## [1] 0.03641833 0.06338206
```

the margin of error for the proportion of the atheist is 0.0134.

Exercise 5:

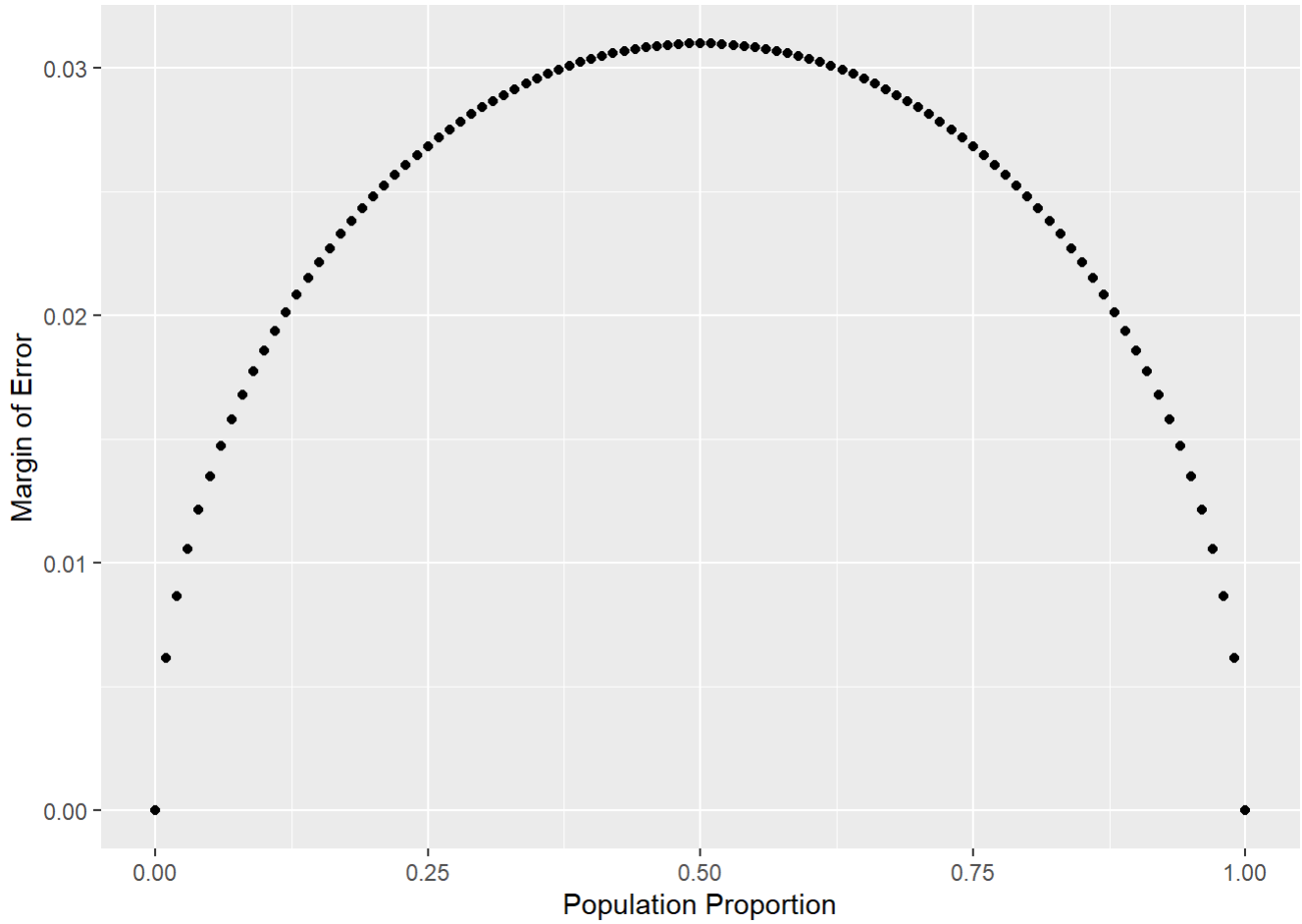
```
can12 <- atheism %>%
  filter(nationality == "Canada" & year == "2012")
can12_summary <- can12 %>%
  summarise(atheists = sum(response == "atheist"),
            non_atheist = sum(response == "non-atheist"),
            prop_atheist = atheists/(atheists+non_atheist))
can_prop <- pull(can12_summary, prop_atheist)



can_se <- sqrt((can_prop * (1 - can_prop)) / nrow(can12)) #the formula for the standard error

can_me <- can_se * z_star # margin of error is standard error times critical value

can_ci_95 <- c(can_prop - can_me, can_prop + can_me)




nigeria12 <- atheism %>%
  filter(nationality == "Nigeria" & year == "2012")
nigeria12_summary <- nigeria12 %>%
  summarise(atheists = sum(response == "atheist"),
            non_atheist = sum(response == "non-atheist"),
            prop_atheist = atheists/(atheists+non_atheist))
nigeria_prop <- pull(nigeria12_summary, prop_atheist)



nigeria_se <- sqrt((nigeria_prop * (1 - nigeria_prop)) / nrow(nigeria12)) #the formula for the s
         tandard error

nigeria_me <- nigeria_se * z_star # margin of error is standard error times critical value

nigeria_ci_95 <- c(nigeria_prop - nigeria_me, nigeria_prop + nigeria_me)
```

we picked Canada we checked to see if there is at least 10 atheist s and 10 non-atheist and assumed independence between sample people. so, its ok to use the central limit theorem. Canadian margin of error was 0.017. the 95% of confidence interval for Canada is (0.072, 0.107).

we next picked Nigeria we check to see if there are 10 sampled atheist and 10 non atheist, and again assumed independency. so that is ok to use central limit. Nigerian margin of error was 0.0058. the 95% of confidence interval for Nigeria is (0.0036, 0.015)

Exercise 6:

```
n <- 1000
p <- seq(0, 1, 0.01)
me <- 1.96 * sqrt(p * (1 - p) / n)
me_tibble <- tibble(p = p, me = me)
ggplot(me_tibble, aes(x = p, y = me)) +
  geom_point() +
  labs(y = "Margin of Error", x = "Population Proportion")
```

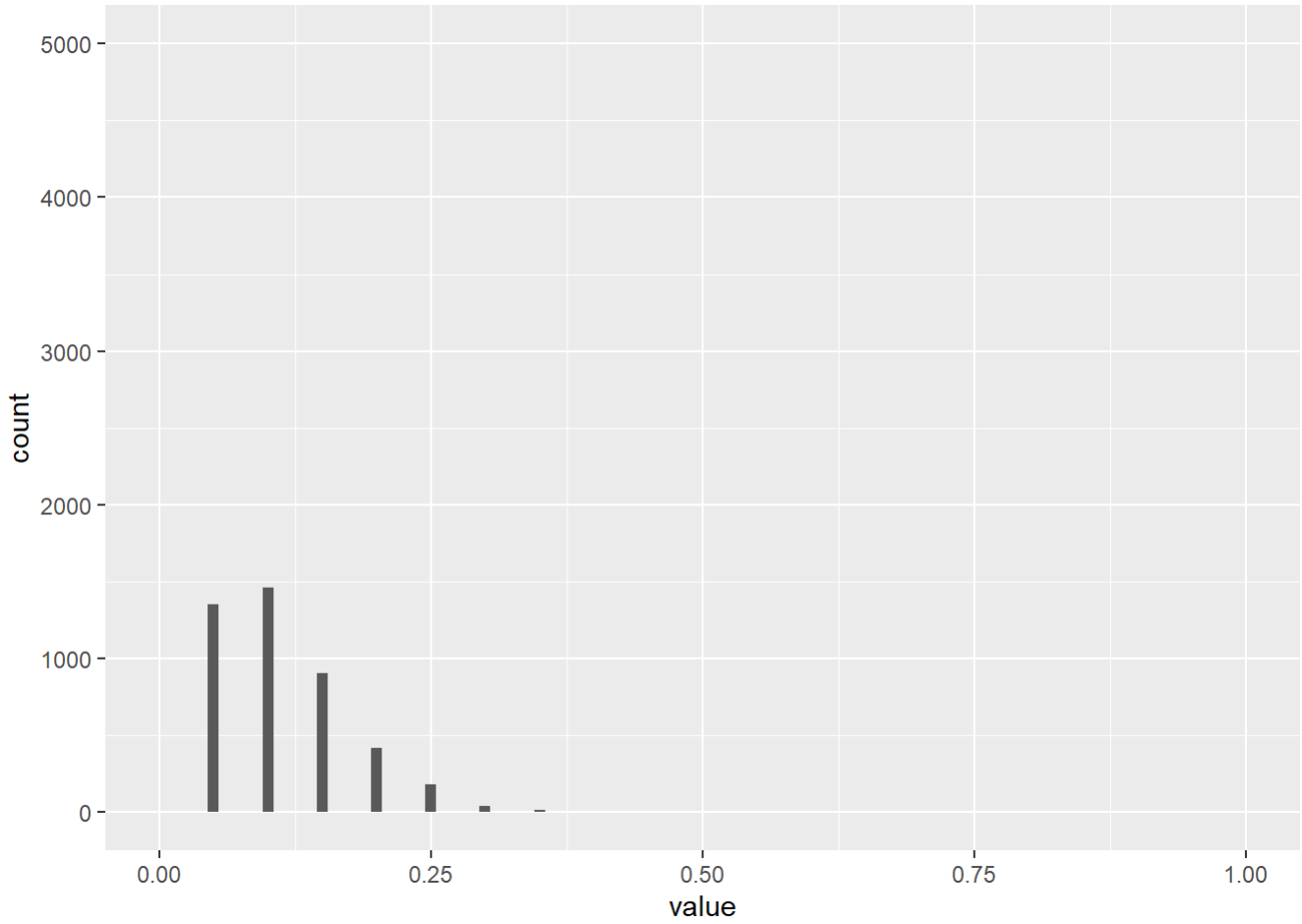


the margin of error depends on P(1-p) so it's the biggest when P is 0.5 an declines when p is smaller or greater than 0.5.

Exercise 7:

```r
# Set parameters.
# If you change these, all of the sampling will change because we pass these into
# the sampling function
n <- 20
p <- .1

# A function that takes a number of samples(n) and a true success probability(p),
# and returns the proportion of success in a random sample of n trials.
one_sample <- function(n, p) {

  # We're again going to use the sample function, but this time we're going to
  # make things a little bit simpler by using the prob parameters. This lets us
  # say use p to assign a probability of sampling each of the values in x,
  # instead of just making an array with lots of copies of the values like we
  # did before.
  sampled_values <- sample(x = c(TRUE, FALSE), n,
                           prob = c(p, 1-p),
                           replace = TRUE)

  # return the proportion of successes
  mean(sampled_values)
}

# Take 5000 samples and look at the distribution
replicates <- tibble(sample = 1:5000,
                     value = replicate(5000, one_sample(n, p)))

ggplot(replicates, aes(x = value)) +
  geom_histogram(binwidth = .01) +
  xlim(c(0, 1)) + # set x-axis limits so we can see the whole 0-1 range
  ylim(c(0, nrow(replicates)))
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```




```
# set y-axis limits so we can see the whole 0-replicates range
```

N would control the sored of the distribution and p would control the center. It will always center at p. but the variability and what the distribution would look like will depend on both p and n. so as p gets closer to 0.5, the distribution will get more variable, because P(1-P) is the maximum. And if we set a small sample size, we will start getting non-normal distributions, because the success-fail idea wont be applied(even if P is in edges).

Exercise 8:

```
SPAIN12 <- atheism %>%
  filter(nationality == "Spain" & year == "2012")
nrow(SPAIN12)
```

```
## [1] 1145
```

```
SPAIN_2005_prop <- .1
SPAIN_2012_prop <- .09
SPAIN_diffs = SPAIN_2012_prop - SPAIN_2005_prop
SPAIN_diffs
```

```
## [1] -0.01
```

```
SE_SPAIN = sqrt((SPAIN_2012_prop*(1-SPAIN_2012_prop)/nrow(SPAIN12))+(SPAIN_2005_prop*(1-SPAIN_20
         05_prop)/nrow(SPAIN12)))
SE_SPAIN
```

```
## [1] 0.0122528
```

```
pnorm(SPAIN_diffs, 0, SE_SPAIN)
```

```
## [1] 0.20721
```

```
CI_Spain <-c(SPAIN_diffs - (1.96*SE_SPAIN), SPAIN_diffs + (1.96*SE_SPAIN))
CI_Spain
```

```
## [1] -0.03401548  0.01401548
```

H0:There is not a difference between the proportion of US atheist population in 2005 and 2012 HA:There is a difference between the proportion of US atheist population in 2005 and 2012 There is no enough evidence to reject the null hypothesis because the P-value is 0.2 which is bigger than 0.05. also the CI does include 0.

Exercise 9:

```
US_2005_prop <- .01
US_2012_prop <- .05
nrow(us12)
```

```
## [1] 1002
```

```
US_diffs = US_2012_prop - US_2005_prop
US_diffs
```

```
## [1] 0.04
```

```
SE_US = sqrt((US_2012_prop*(1-US_2012_prop)/nrow(us12))+(US_2005_prop*(1-US_2005_prop)/nrow(us1
         2)))
SE_US
```

```
## [1] 0.007568714
```

```
pnorm(US_diffs, 0, SE_US)
```

```
## [1] 0.9999999
```

```
CI_US <-c(US_diffs - (1.96*SE_US), US_diffs + (1.96*SE_US))
CI_US
```

```
## [1] 0.02516532 0.05483468
```

H0:There is not a difference between the proportion of US atheist population in 2005 and 2012 HA:There is a difference between the proportion of US atheist population in 2005 and 2012 There is enough evidence to reject the null hypothesis because the P-value is 0.99 which is in the tope 2.5% percent. also the CI does not include 0.

Exercise 10:

```
Countries <- 39
0.05*Countries
```

```
## [1] 1.95
```

Based on the Type 1 error and the 0.05 criterion, in 5% of the times we make a false rejection. 5% of 39 countries is 1.95. so by chance we might detect a change in 1 or 2(maximum) countries. which 2 is very rare to happen.

Exercise 11:

```
p <- 0.5
SE <- .01
Z <- qnorm(.975)

#Z*sqrt(p*(1-p)/SampleSize) <= SE
#sqrt(p*(1-p)/SampleSize) <= SE/Z
#p*(1-p)/SampleSize <= (SE/Z)**2
#(p*(1-p))/((SE/Z)**2)<= SampleSize
SampleSize =(p*(1-p))/((SE/Z)**2) #SampleSize should be greater than this number
SampleSize
```

```
## [1] 9603.647
```

Based on the Z*sqrt(p*(1-p)/SampleSize) <= SE formula, we conclude that sample size should be equal or greater than (p*(1-p))/((SE/Z)**2). Since we have all the inputs except the P, we consider P as 0.5 since it is the when the margin of error is the biggest. So we should sample at least 9603.647 (9604 people).