Inference for categorical data

Load packages

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ---------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(infer)
```

Also load my own functions

In this lab, we will explore and visualize the data using the **tidyverse** suite of packages, and perform statistical inference using **infer**. The data can be found in the companion package for OpenIntro resources, **openintro**.

Let's load the packages.

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

```
## Loading required package: airports

## Loading required package: cherryblossom

## Loading required package: usdata
```

```
library(infer)
```

Also load my own functions

In this lab, we will explore and visualize the data using the **tidyverse** suite of packages, and perform statistical inference using **infer**. The data can be found in the companion package for OpenIntro resources, **openintro**.

Let's load the packages.

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

Also load my own functions

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

```
yrbss$text_while_driving_30d%>%
  unlist()%>%
  unique()
```

```
## [1] "0"              NA               "30"             "did not drive"
## [5] "1-2"            "3-5"            "20-29"          "10-19"
## [9] "6-9"
```

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

```
helmet_count<-yrbss%>%
  filter(helmet_12m=="never")%>%
  nrow()
text_count<-yrbss%>%
  filter(helmet_12m=="never")%>%
  filter(!is.na(text_while_driving_30d),text_while_driving_30d=="30")%>%
  nrow()
proportion<-text_count/helmet_count

cat("the proportion of people who do not wear helmets who have texted every day in the last thirty days
```

```
## the proportion of people who do not wear helmets who have texted every day in the last thirty days i
```

1. 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?

```
#(z*)*sqrt(p(1-p)/n)
p<-proportion
n<-helmet_count
margin_of_error<-round((1.96^2)*sqrt(p*(1-p)/n),2)
cat("the margin of error for this estimate is ",margin_of_error)
```

```
## the margin of error for this estimate is  0.01
```

1. 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.

My question will be, are tall people more physically active than average. I will filter for tall (based on quantile), then create a text indication for if they are above average for physical activity.

```
#find out who's tall
summary(yrbss$height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   1.270   1.600   1.680   1.691   1.780   2.110    1004
```

```
#filter for tall
tall<-yrbss%>%
  filter(!is.na(height),height>1.78)

#find mean for activity
summary(yrbss$physically_active_7d)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   0.000   2.000   4.000   3.903   7.000   7.000     273
```

```
#create text indication for second variable
tall <- tall %>%
  mutate(text_ind=ifelse(physically_active_7d> 4,"yes","no"))

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

```
## Warning: Removed 44 rows containing missing values.
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    0.571    0.612
```

1. 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?

   The relationship is parabolic. As population proportion increases, margin of error increases until it hits a maxima, then it decreases. Its maxima is .5.

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

   It is right skewed, as the lower tail is cut off by zero, and it looks fairly normal other than that.

3. 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.

   the center moves higher. The spread and shape seem to stay the same, but you seem to have less peaks at integers.

4. Now also change n. How does n appear to affect the distribution of $\hat{p}$?

As n increases, the spread decreases, which makes sense as the sample grows, you are more sure of the population mean.

1. 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.

Hypothesis,

H_0: people who sleep 10+ hours a night are not more likely to strength train every day.

H_A: people who sleep 10+ hours a night are more or less likely to strength train every day.

```r
#text indication for sleep every day
sleep_every_day<-yrbss%>%
  filter(!is.na(school_night_hours_sleep))%>%
  mutate(ten_plus=ifelse(school_night_hours_sleep=="10+","yes","no"))
#text indication for train every day
sleep_every_day<-sleep_every_day%>%
  mutate(seven_day=ifelse(strength_training_7d=="7","yes","no"))

table<-table(sleep_every_day$ten_plus,sleep_every_day$seven_day)

colnames(table)<-c("low_workout","max_workout")
rownames(table)<-c("low sleep", "max_sleep")

table
```

```
##
##             low_workout max_workout
##   low sleep        9949        1958
##   max_sleep         228          84
```

```r
#looks like people might be more likely to max sleep when they max workout
prop.table(table,1)
```

```
##
##             low_workout max_workout
##   low sleep   0.8355589   0.1644411
##   max_sleep   0.7307692   0.2692308
```

```r
#check chisq and pvalues
summary(table)
```

```
## Number of cases in table: 12219
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 23.986, df = 1, p-value = 9.705e-07
```

the p-value is lower than .05, so we reject the null hypothesis, and accept that people who max sleep are more likely to max work out.

1. 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 probablity 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.

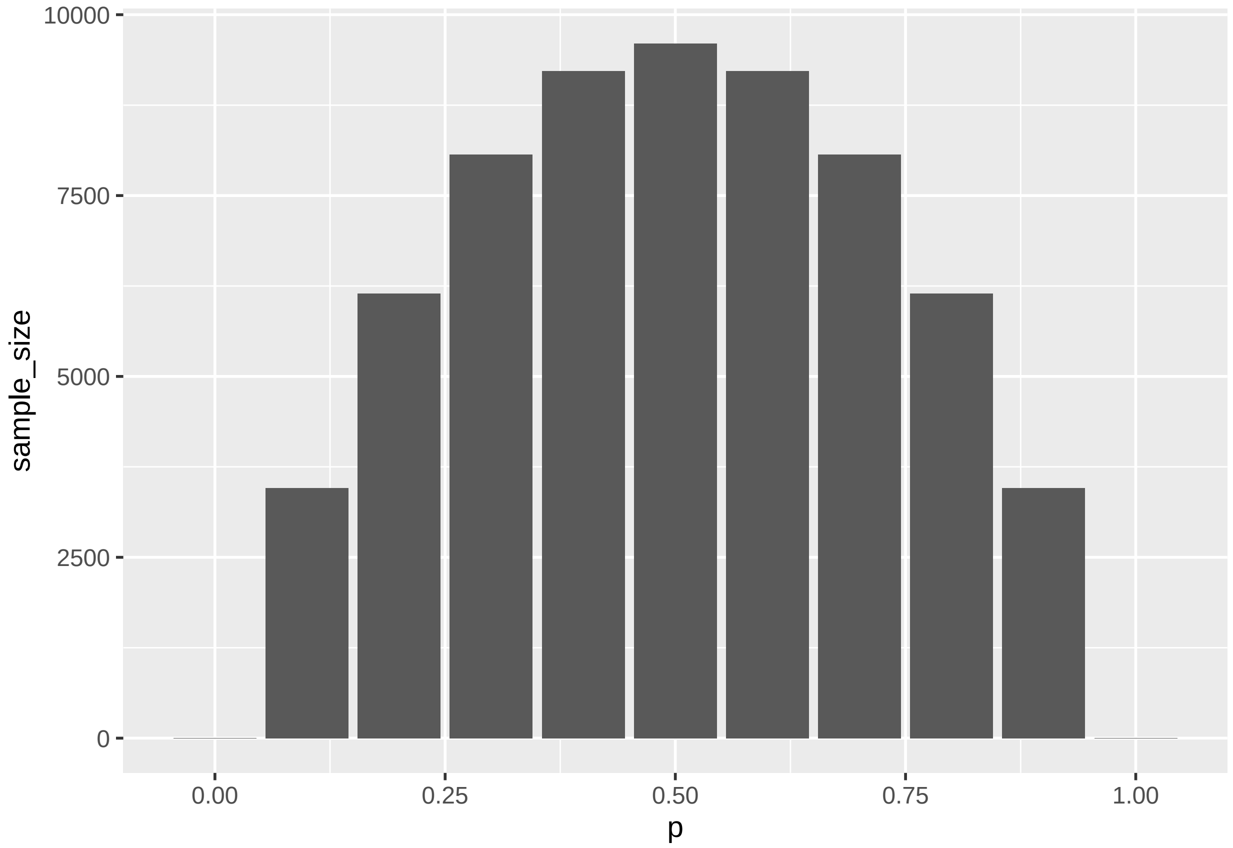
   the chance that you detect a change is 5% at a 5% significance level. (if you are 95% confident, 5% of the time you are wrong.)

2. 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.

```r
p <- seq(from = 0, to = 1, by = 0.1)

results<-lapply(p,FUN=function(x){n_size_margin_of_error(1.96,.01,x)})
results<-results%>%
  unlist()%>%
  as.vector()
df_res<-data.frame("p"=p,"sample_size"=results)

ggplot(data=df_res, aes(x=p, y=sample_size)) +
    geom_bar(stat="identity")
```


sample_size
p