STAT 311 Project Group 38

Kyle Mumma, Jasmine Joy Palaganas, Sofyar Satrio Utomo

12/15/2020

Setup

```
library(knitr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
knitr::opts_chunk$set(echo = TRUE)
SHOW_SOLUTIONS = TRUE
```

Read in the Data

Read in the data and set all categorical variables to factors.

```
P.df <- read.csv("Stat311Au20GroupProjectSurveyDataProcessed.csv",
                 header=TRUE, as.is=TRUE)
P.df$Ethnicity <- as.factor(P.df$Ethnicity)
P.df$Gender <- as.factor(P.df$Gender)
P.df$Religion <- as.factor(P.df$Religion)
P.df$Study <- as.factor(P.df$Study)
P.df$Origin <- as.factor(P.df$Origin)
P.df$SES <- as.factor(P.df$SES)
P.df$PoliticalAff <- as.factor(P.df$PoliticalAff)
P.df$Education <- as.factor(P.df$Education)
P.df$DataPlan <- as.factor(P.df$DataPlan)
P.df$SubUse <- as.factor(P.df$SubUse)
P.df$Diet <- as.factor(P.df$Diet)
P.df$SIFStudy <- as.factor(P.df$SIFStudy)
```

Problem 1: The Data Set

The data comes from the responses of students attending the University of Washington in the course, STAT 311, in the Fall quarter. It was collected in the form of a survey given as an optional assignment with extra credit to these students. The data set was then processed in the form of an excel spreadsheet with 178 observations and 23 variables. Each row represents a single student in the class. While each column represents a different variable or the response to each question in the survey. Answers took a variety of forms. There were both forms of numerical and categorical variables. There were questions that asked for whole integers such as age. A specific example of a discrete numerical variable was an instance where a student could respond with average daily screen time, "rounded to the nearest quarter hour". Whereas, the "Ethnicity" variable was recorded as a number, but each number represented a different ethnicity, making it a nominal categorical variable. Other multiple choice questions were also represented this way. Some other categorical variables include religion, field of study, and political affiliation. The survey also calculates a "climate sentiment score" based on answers to a variety of sub questions and reports them as a whole integer in the data set. There are also questions that ask to "select all that apply". These questions are processed as a sum of all of the options chosen.

Problem 2: Sampling

Part A: Type of Sample

Self-selected convenience sampling was used to obtain this data set. The survey was only given to students in this specific class and it was optional.

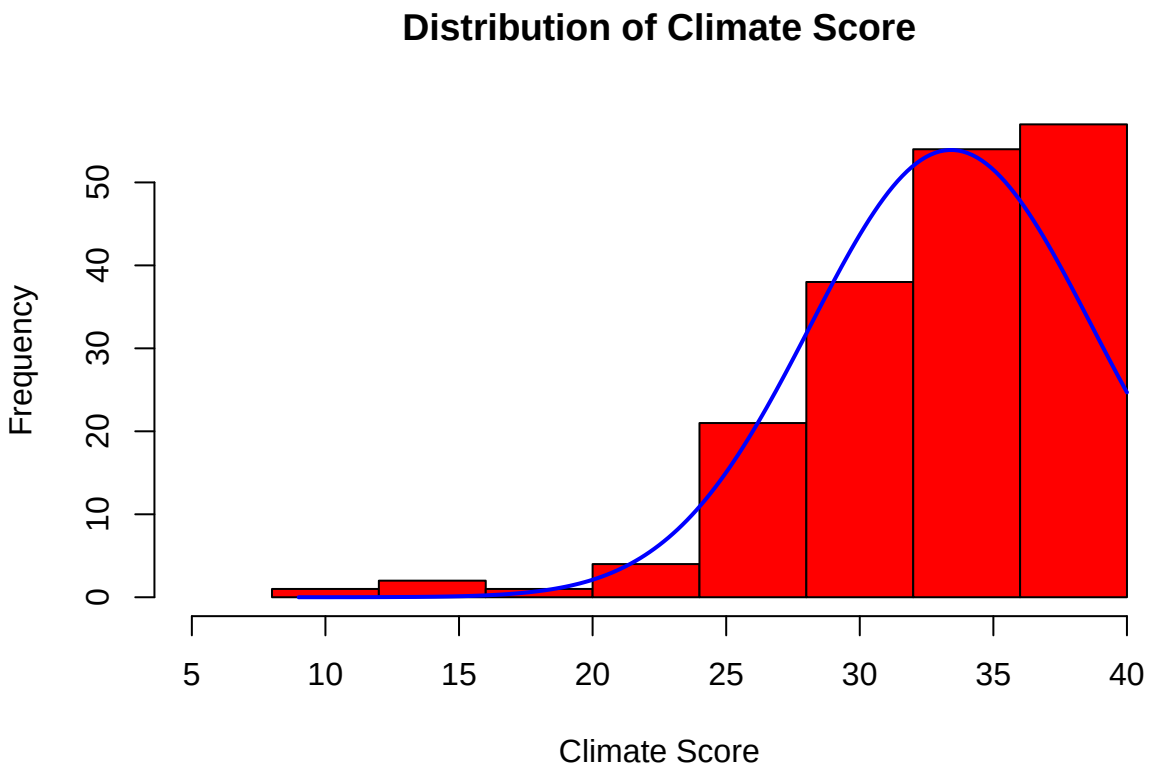
Part B: Represented Population

This sample could represent a couple of possible populations. Depending on the scale of the research question, We could see this sample help us interpret the populations of university students. This is because the data came from responses from students attending the University of Washington. However, to be even more accurate, the project could be a study of the general population of students at this school. Ultimately, it makes the most sense for this sample to represent students in their late teens and 20s, and to be current college or university students because many of the questions are most applicable to students. Questions about area of study, social media, and sleep and screen time, seem to pertain strongly to these populations.

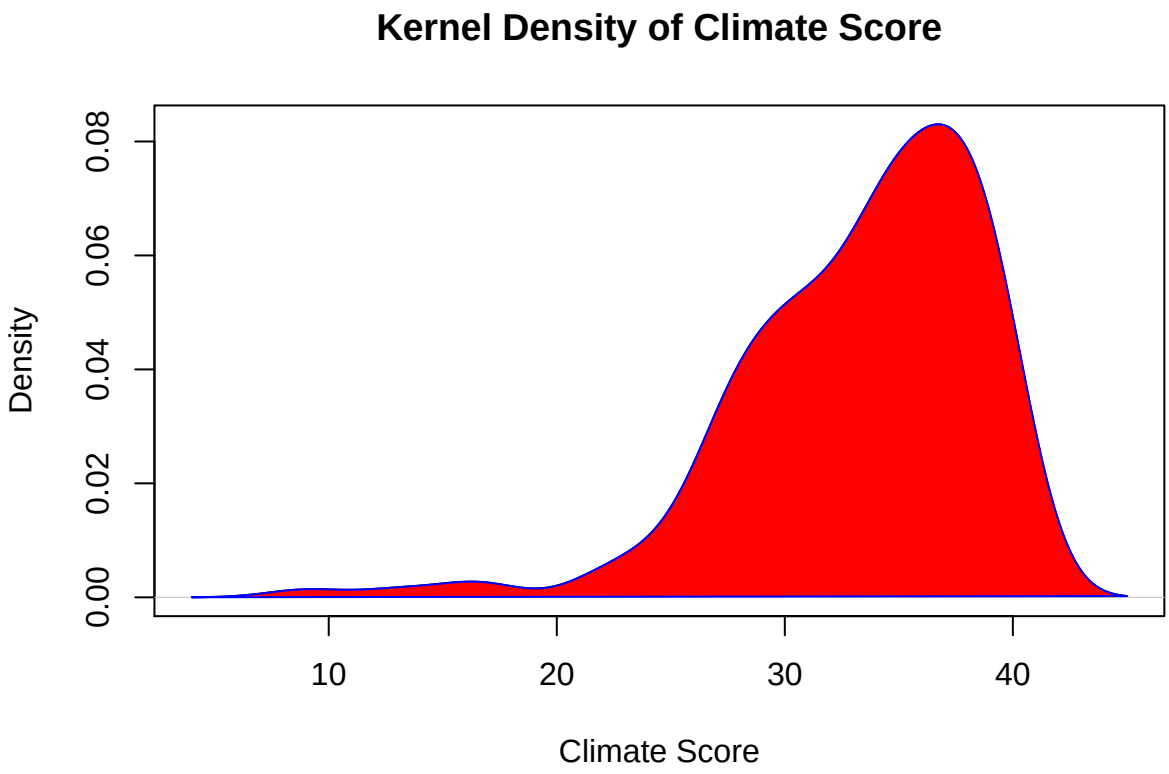
Problem 3: Explore Climate Score

Part A: Climate Score Overall

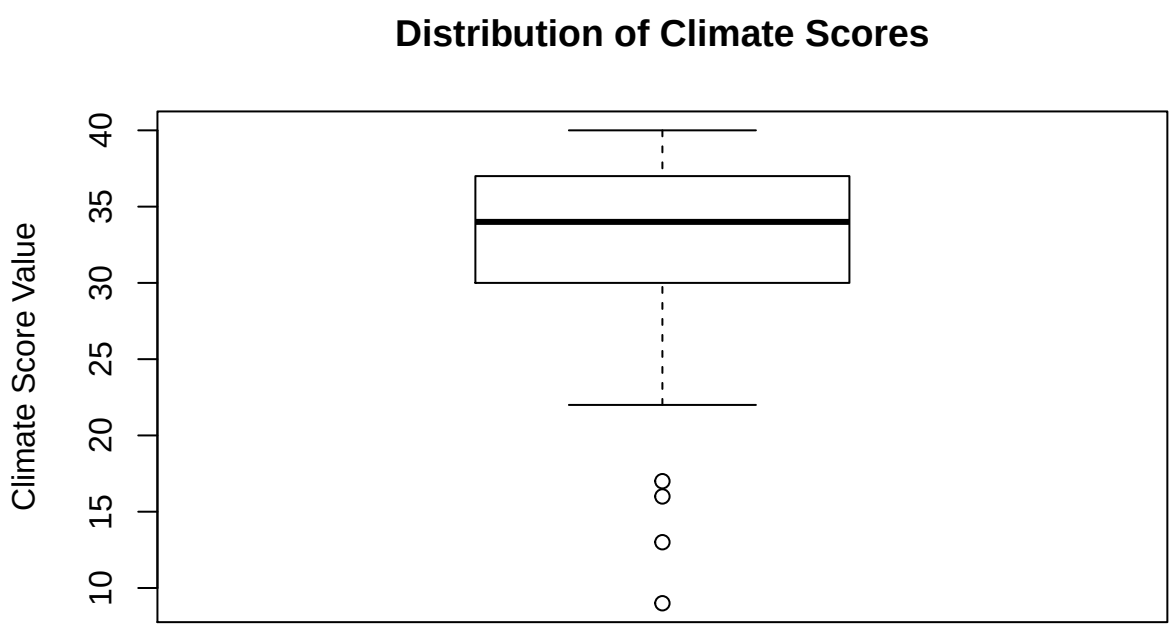
```
hplot1 <- hist(P.df$ClimateScore, breaks = c(8, 12, 16, 20, 24, 28, 32, 36, 40), xlab = "Climate Score"
     main = "Distribution of Climate Score", xlim = c(5, 40), col = "red")
xfit<-seq(min(P.df$ClimateScore), max(P.df$ClimateScore), length=178)
yfit<-dnorm(xfit, mean=mean(P.df$ClimateScore), sd=sd(P.df$ClimateScore))
yfit <- yfit*diff(hplot1$mids[1:2])*length(P.df$ClimateScore)
lines(xfit, yfit, col="blue", lwd=2)
```




```
kdplot1 <- density(P.df$ClimateScore)
plot(kdplot1, main="Kernel Density of Climate Score", xlab = "Climate Score")
polygon(kdplot1, col="red", border="blue")
```



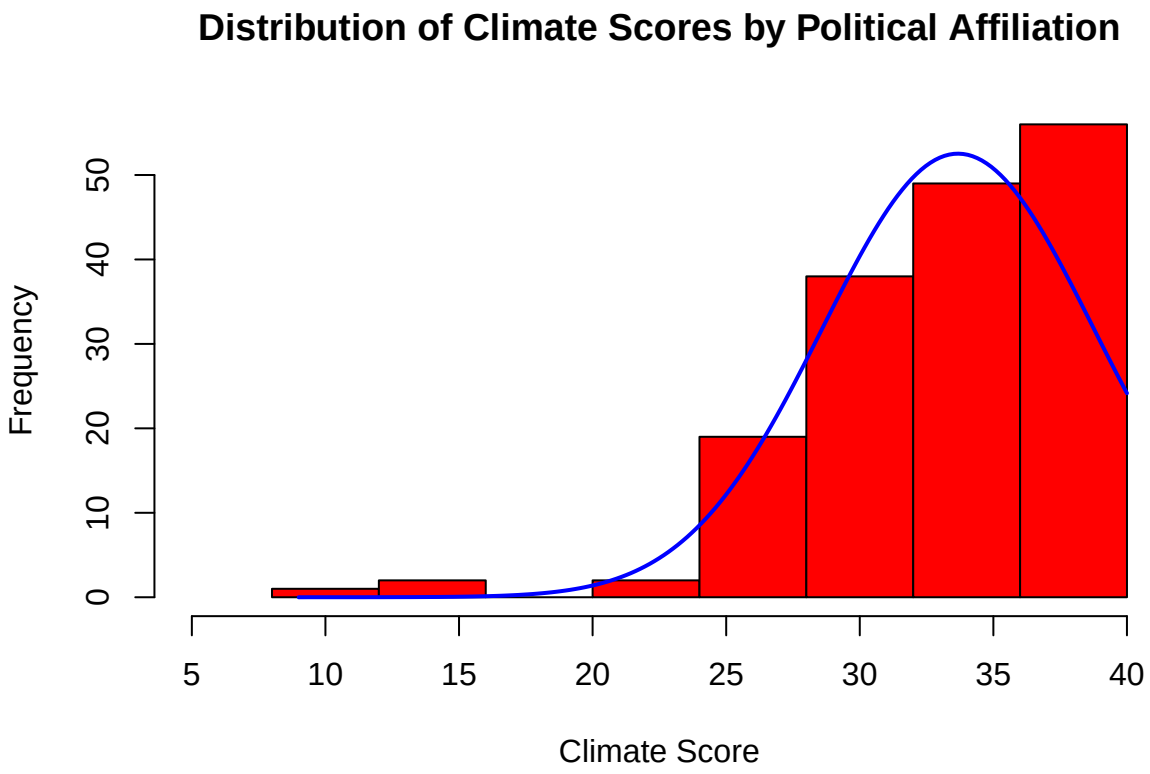

```
boxplot(P.df$ClimateScore, main = "Distribution of Climate Scores", ylab = "Climate Score Value")
```



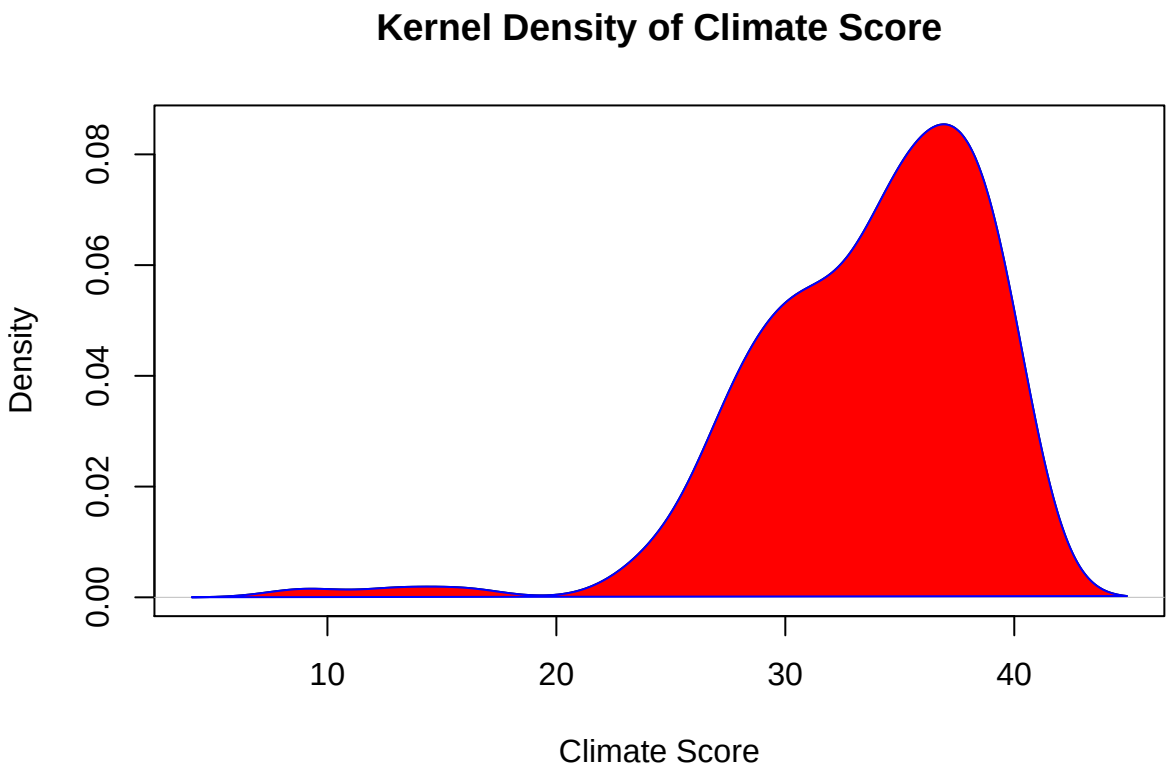
The distribution of the climate score is skewed to the left, with median at 34 and mean at 33.42. It has a standard deviation of 5.27 and it also has 4 outliers.

Part B: Climate Score by Democrat and No Affiliation

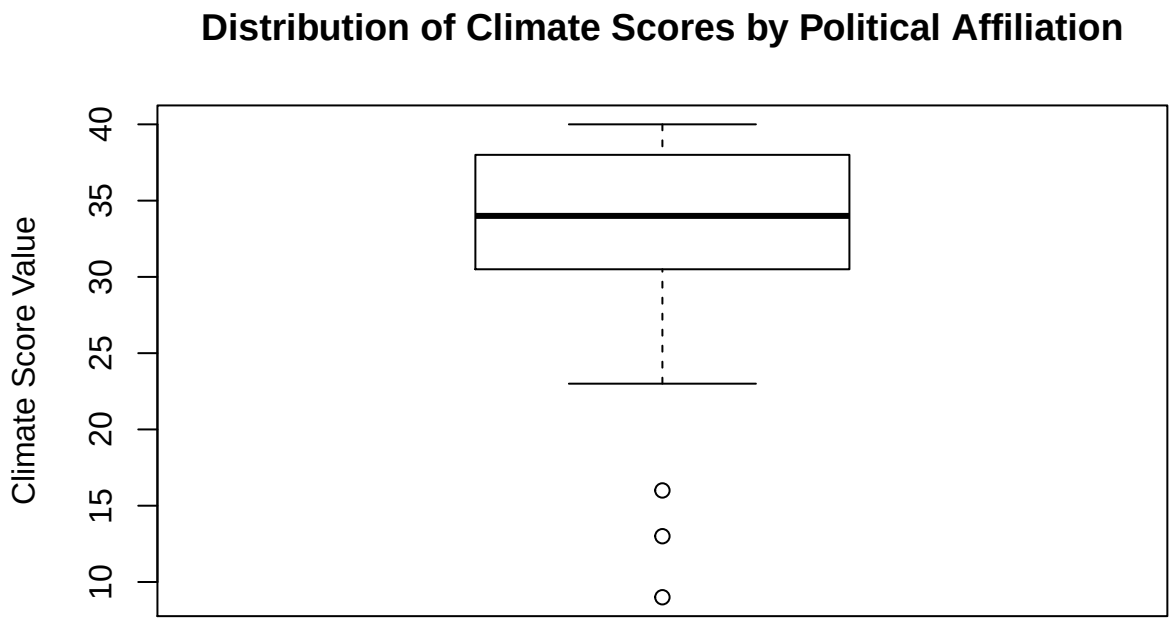
```
non_rep <- P.df %>% select(PoliticalAff, ClimateScore) %>%
  filter(PoliticalAff != "Republican")

hplot2 <- hist(non_rep$ClimateScore, breaks = c(8, 12, 16, 20, 24, 28, 32, 36, 40), xlab = "Climate Sco
     main = "Distribution of Climate Scores by Political Affiliation", xlim = c(5, 40), col = "red")
xfit<-seq(min(non_rep$ClimateScore), max(non_rep$ClimateScore), length=178)
yfit<-dnorm(xfit, mean=mean(non_rep$ClimateScore), sd=sd(non_rep$ClimateScore))
yfit <- yfit*diff(hplot2$mids[1:2])*length(non_rep$ClimateScore)
lines(xfit, yfit, col="blue", lwd=2)
```




```
kdplot2 <- density(non_rep$ClimateScore)
plot(kdplot2, main="Kernel Density of Climate Score", xlab = "Climate Score")
polygon(kdplot2, col="red", border="blue")
```




```
boxplot(non_rep$ClimateScore, main="Distribution of Climate Scores by Political Affiliation", ylab="Cli
```



The distribution of the filtered data is still very similar with the one earlier. The main notable difference is that it now has 3 outliers instead of 4.

Problem 4: Political Affiliation and SIF Study

Part A: Contingency Table for Political Affiliation

We have included the subset code in the R block below to get you started. You will need to form other subsets of data for other questions. Refer to R Assignment 2 Tutorial, R Assignment 2 Solutions and R Assignment 3 Tutorial for examples.

```
PAff.D <- filter(P.df, PoliticalAff == "Democrat" |  PoliticalAff =="No affiliation")

PAff.D <- filter(PAff.D, SIFStudy == "Yes" | SIFStudy == "No")
PAff.D$PoliticalAff <- factor(PAff.D$PoliticalAff)
PAff.D$SIFStudy <- factor(PAff.D$SIFStudy)

tabDNA <- table(PAff.D$PoliticalAff, PAff.D$SIFStudy)
tabDNA
```

```
##
##                  No Yes
##   Democrat        6  39
##   No affiliation 12  19
```

Part B: CI For Political Affiliation Difference

```
# Large number test. (Yes, as a success)
tabAll <- table(P.df$PoliticalAff, P.df$SIFStudy)
tabAll
```

```
##
##                  I prefer not to answer I'm not sure No Yes
##   Democrat                            0           22  6  39
##   Independent                         0            5  4   4
##   No affiliation                      4           43 12  19
##   Other                               1            4  2   2
##   Republican                          0            4  6   1
```

```
sum(tabAll[1,]) * tabAll[1,4]/sum(tabAll[1,]) # n1p1
```

```
## [1] 39
```

```
sum(tabAll[1,]) * (1- (tabAll[1,4]/sum(tabAll[1,]))) # n1q1
```

```
## [1] 28
```

```
sum(tabAll[3,]) * tabAll[3,4]/sum(tabAll[3,]) # n2p2
```

```
## [1] 19
```

```
sum(tabAll[3,]) * (1- (tabAll[3,4]/sum(tabAll[3,]))) # n2q2
```

```
## [1] 59
```

The large sample condition holds so we proceed with the confidence interval.

```
# CI construct
prop.test(c(tabAll[1,4], tabAll[3,4]), c(sum(tabAll[1,3:4]), sum(tabAll[3,3:4])),
                                    conf.level=0.95,
                                    correct=FALSE)$conf.int
```

```
## [1] 0.05561099 0.45191589
## attr(,"conf.level")
## [1] 0.95
```

We are 95% confident that the difference in the population proportion of democrats versus those with no political affiliation that favor a three-year SIF study in Seattle falls between about 0.056 and 0.452. Since zero is not contained in the interval, there may be be difference between the proportions of the proportion of the two groups.

Part C: Test for Political Affiliation

```r
# null hypothesis = the proportion of students in favor of a three-year SIF study in Seattle is greater

prop.test(c(tabAll[1,4], tabAll[3,4]), c(sum(tabAll[1,3:4]), sum(tabAll[3,3:4])),
          alternative = "less", correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(tabAll[1, 4], tabAll[3, 4]) out of c(sum(tabAll[1, 3:4]), sum(tabAll[3, 3:4]))
## X-squared = 6.5395, df = 1, p-value = 0.9947
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000  0.4200582
## sample estimates:
##    prop 1    prop 2
## 0.8666667 0.6129032
```

The null hypothesis is p-dem > p-no. The alternative hypothesis is p-dem <= p-no. The p-value is 0.99. Since p = 0.99 > alpha = 0.05, we fail to reject the null hypothesis. There is insufficient evidence to conclude that the proportion of students in favor of a three-year SIF study in Seattle is greater for students that identify as democrats versus no political affiliation.

Problem 5: Inference for Climate Sentiment

```r
meanConfidenceInterval <- function(sample, alpha) {
  x_bar <- mean(sample)
  me <- qnorm(alpha / 2) * (sd(sample) / sqrt(length(sample)))
  lowerBound <- x_bar + me
  upperBound <- x_bar - me
  return(c(lowerBound, upperBound))
}
```

Part A: 90% CI For Mean Climate Sentiment Score

```r
meanConfidenceInterval(P.df$ClimateScore, .1)
```

```
## [1] 32.76618 34.06528
```

We are 90% confident that the population mean for climate sentiment is between about 32.766 and 34.065

Part B: 99% CI For Mean Climate Sentiment Score

```r
meanConfidenceInterval(P.df$ClimateScore, .01)
```

```
## [1] 32.39854 34.43293
```

We are 99% confident that the population mean for climate sentiment is between about 32.399 and 34.433

Part C: Different Bounds

The lower and upper bounds in part A are tighter (smaller range) than the lower and upper bounds in part B. This is because in part A we are less confident that the mean is within the given bounds and therefore we can have a tighter range. In part b we need to be more confident that the mean is within the given bounds and therefore we need a larger range.

Part D: Test for Mean Climate Sentiment Score

```
sample <- P.df$ClimateScore
z <- (mean(sample) - 32.4) / (sd(sample) / sqrt(length(sample)))
p_val <- 2*pnorm(z, mean=0, sd=1, lower.tail=FALSE)
p_val
```

```
## [1] 0.01010778
```

```
p_val <= .05
```

```
## [1] TRUE
```

H_0: mu = 32.4 H_a: mu != 32.4 test statistic: x_bar = 33.41573 p-value: 0.01010778 decision: reject the null hypothesis

Because 0.010 <= .05, there is sufficient evidence to support the claim that the population mean for climate sentiment is different than 32.4.

Part E: Type of Error

For the hypothesis test in part D we risk making a type 1 error. It is possible that we could reject the hypothesis that the population mean for climate sentiment is equal to 32.4, when in reality it actually is true.

Problem 6: Climate Sentiment and Political Affiliation/Origin

Part A: 99% CI For Difference in Mean Climate Sentiment Scores

```
diffMeansCI <- function(x1, x2, alpha) {
  x1_bar <- mean(x1)
  x2_bar <- mean(x2)
  s_1 <- sd(x1)
  s_2 <- sd(x2)
  n_1 <- length(x1)
  n_2 <- length(x2)
  me <- qnorm(alpha / 2) * sqrt((s_1^2 / n_1) + (s_2^2 / n_2))
  x <- x1_bar - x2_bar
  return(c(x + me, x - me))
}

dems <- filter(P.df, PoliticalAff=="Democrat")
non_dems <- filter(P.df, PoliticalAff!="Democrat")

diffMeansCI(dems$ClimateScore, non_dems$ClimateScore, .01)
```

```
## [1] 3.503038 6.747870
```

We are 99% confident that the difference in population mean climate scores between democrats and non-democrats falls between about 3.503 and 6.748

Part B: Test For Mean Climate Sentiment Score by Political Affiliation

```
dems <- filter(P.df, PoliticalAff=="Democrat")
non_dems <- filter(P.df, PoliticalAff!="Democrat")
s1 <- sd(dems$ClimateScore)
s2 <- sd(non_dems$ClimateScore)
n1 <- length(dems$ClimateScore)
n2 <- length(non_dems$ClimateScore)
se <- sqrt((s1^2/n1) + (s2^2/n2))
z <- (mean(dems$ClimateScore) - mean(non_dems$ClimateScore)) / se
p_val <- pnorm(z, mean=0, sd=1, lower.tail=FALSE)
p_val
```

```
## [1] 2.018803e-16
```

```
p_val <= .01
```

```
## [1] TRUE
```

mu_1 = population mean for climate score of democrats

mu_2 = population mean for climate score of non-democrats

H_0: mu_1 - mu_2 = 0 H_a: mu_1 - mu_2 > 0

test statistic: xbar_1 - xbar_2 = 5.125454

p-value = 2*10^-16

decision: reject the null hypothesis

Because 2*10^-16 <= .01, there is sufficient evidence to support the claim that the population mean climate sentiment score for students that identify as democrats is greater than the population mean score for students that identify as something other than democrat.

Part C: Test for Mean Climate Sentiment Score by Origin

```
urb <- filter(P.df, Origin=="Urban")
sub <- filter(P.df, Origin=="Suburban")
s1 <- sd(urb$ClimateScore)
s2 <- sd(sub$ClimateScore)
n1 <- length(urb$ClimateScore)
n2 <- length(urb$ClimateScore)
se <- sqrt((s1^2 / n1) + (s2^2 / n2))
z <- (mean(urb$ClimateScore) - mean(sub$ClimateScore)) / se
p_val <- 2*pnorm(z, mean=0, sd=1)
p_val
```

```
## [1] 0.04459196
```

```
p_val <= .01
```

```
## [1] FALSE
```

mu_1 = population mean for climate score of students raised in urban environment mu_2 = population mean for climate score of students raised in suburban environment

H_0: mu_1 - mu_2 = 0 H_a: mu_1 - mu_2 != 0

test statistic = xbar_1 - xbar_2 = -1.471528

p-value: 0.04459196

decision: fail to reject to null hypothesis

Because 0.045 > .01, there is insufficient evidence to support the claim that the population mean climate sentiment score for students raised in urban environment is different than the population mean score for students raised in a suburban environment.

Extra Credit

```
P.df[,7]
```

```
##   [1] 5 4 4 4 4 4 4 4 4 1 2 3 3 4 4 2 4 4 3 3 4 4 4 2 4 3 4 4 4 4 4 4 2 4 4 4 4
##  [38] 4 4 4 3 4 4 4 4 3 3 3 4 4 4 3 5 4 3 4 4 4 4 4 4 4 4 3 4 4 4 3 4 4 4 4 4 4
##  [75] 4 4 3 3 3 4 4 3 4 3 4 3 4 4 4 4 3 4 4 4 3 4 3 4 4 4 4 4 4 3 4 4 4 4 4 4 4
## [112] 4 4 4 4 2 2 5 4 4 4 3 4 1 4 3 4 3 4 4 3 4 3 4 4 4 4 3 3 3 3 3 4 4 3 4 4 3
## [149] 4 4 3 4 3 4 4 3 3 3 4 3 4 4 4 4 4 4 4 4 2 3 2 4 3 4 4 4 3 3 4
## Levels: 1 2 3 4 5
```

It is difficult to eliminate groups down to only two categories because it feels like we are disregarding underrepresented groups. We think it would be best to combine students in the SES (socioeconomic status) category down to the labels, “Struggled” and “Little to No Struggle”. The original categories were somewhat difficult to interpret or measure so it made sense to just simplify it down to a binary variable. It is a fairly easy question to answer whether a family had significant struggle or not, rather than what specific Ethnicity a student is or what area of study they are in. Those groups are more difficult to combine. Here, we would combine the students who answered with 1s or 2s which originally represented the responses, “My family had enough money to take care of things fine” and “My family was able to make ends meet, but with some difficulties” into the “Little to No Struggle” category. Then, we could combine the students who answered with 3s, 4s, and 5s which represented the responses “My family had to struggle hard to make ends meet”, “My family was not able to make ends meet, despite struggling hard”, “Most of the time I was growing up my family was very poor or on welfare”, respectively, into the “Struggled” category. Meanwhile, no one answered with “Prefer not to answer”, so that would be completely omitted from the binary variable.