Resource File Project - Phase 3

#124124 - Mavis Yang, Bingzhong Zhou, Olivia Marsh, Jin Zhao

2022/4/7

```
rawDf <- read.csv("./examdata.csv")
```

Problem

Brief Overview

Dataset represents the math, reading, and writing scores of individual samples in relation to their gender, race, parental level of education, quality of lunch, and the completion of a test preparation course. Sample was taken from high school students in the US.

Research Objective

To understand different factors that could impact student scores on tests

Research Questions

1. What is the average test score of a child with parents who achieved some college-level education?
2. Does gender (female/male) influence the chance that a participant took the test preparation course?
3. Does a high reading score correlate to a high writing score?
4. How does a participant's math score compare with their parent's education level achieved? (High school, bachelor's degree, or associate's degree)
5. Will a participant who took the test preparation course score higher on the math test than a participant who did not take the test preparation course?

Plan

- **Population of interest:** High school students from the United states
- **Sampling Frame:** New York City High school district
- **Sample:** Elanor Roosevelt high school, Frank McCount High School, Emma Lazarus High school, Xavier High school, Stuyvesant high school, Regis High School

We chose to use **multi-stage sampling**. We first conducted **clustered sampling** by sorting the **sampling frame** into clusters based on high school, then randomly chose 6 high schools from the New York City high school district using **simple random sampling**. Then we used **random stratified sampling**, where we subdivided the students in the selected schools by grade (grades 9 to 12) and randomly selected 25% of students from each grade.

We chose **observational study design**, in the format of a **survey**. We used a **written survey** sample about demographic data (i.e. gender, ethnicity, parents' level of education, etc.), as well as data pertaining to their recent academics, including test scores in reading, writing, and math.

- Note: While we know that our data came from high school students in the US, we do not know if the data came from a specific school district, or which/how many schools the data was from. We also chose 25% from each grade arbitrarily (randomly), as we have no data about the age of each participant. We assume that the test scores/test preparation came from the schools themselves, and that we did not administer them (since there are no even numbers for us to work with, i.e. the test preparation was 36% complete rather than 50% in a random study).

Data

Numerical summary

```r
# Import raw dataset
rawDf <- read.csv("./examdata.csv")
```

Table

- *Type of data summarized:* Qualitative and univariate data
- *Objective:* How many of our participants are of group A race?

```r
table(rawDf$race.ethnicity)
```

```
##
## group A group B group C group D group E
##      89     190     319     262     140
```

Proportional Table

- *Type of data summarized:* Qualitative and univariate data
- *Objective:* What proportion of our sample did the preparation course?

```r
prop.table(table(rawDf$test.preparation.course))
```

```
##
## completed      none
##     0.358     0.642
```

Summary Table

- *Type of data summarized:* Quantitative and univariate data
- *Objective:* What was the mean math score?

```
summary(rawDf$math.score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   57.00   66.00   66.09   77.00  100.00
```

Graphical summary

Bar graph (Good Figure for Presentation)

- *Type of data summarized:* Univariate, bivariate and multivariate data. Usually at least one quantitative variable, and one qualitative variable.
- *Objective:* Does gender (female/male) influence the chance that a participant took the test preparation course?

```
ggplot(rawDf,
       aes(fill = test.preparation.course,
           x = gender)) +
  geom_bar(position = "fill") +
  xlab("Gender") + ylab("Test Preparation Course") +
  labs(title = "Proportion of Each Gender to Complete the Test Preparation Course")
```

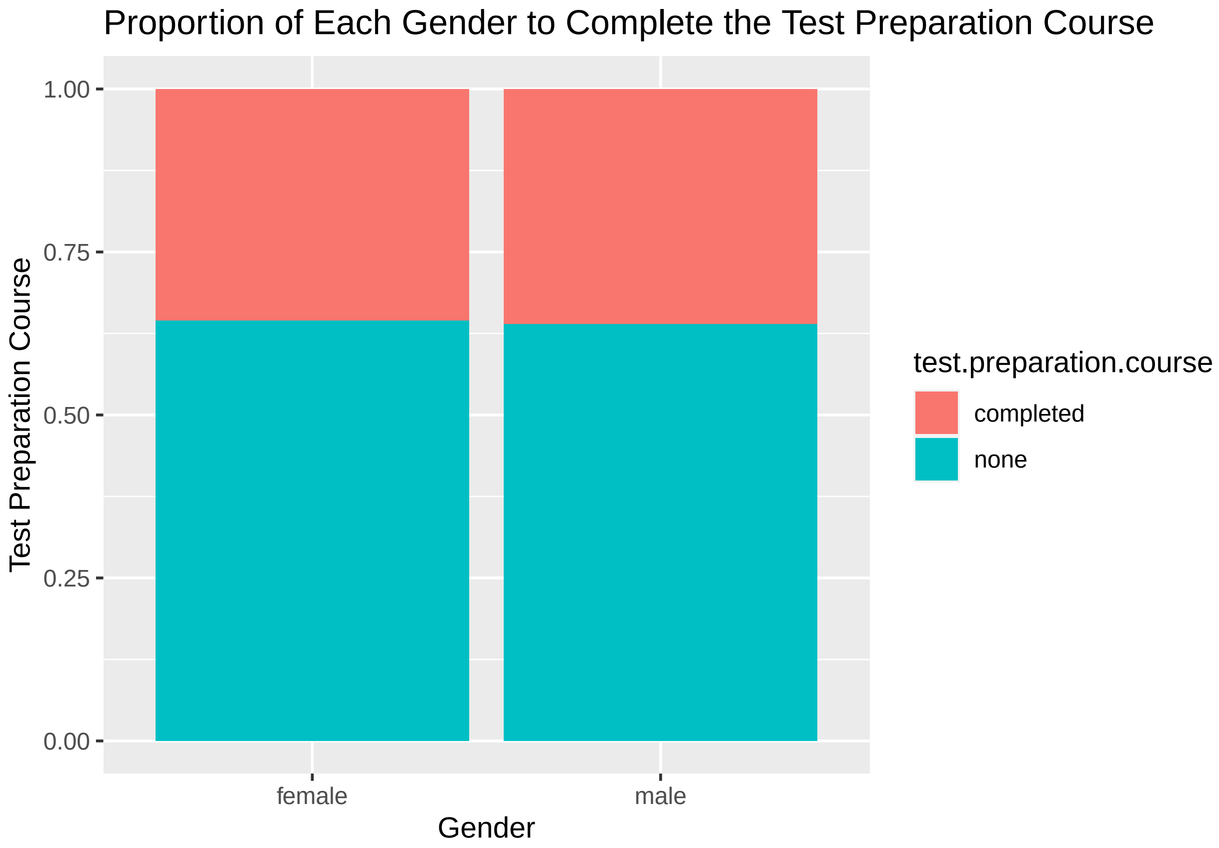


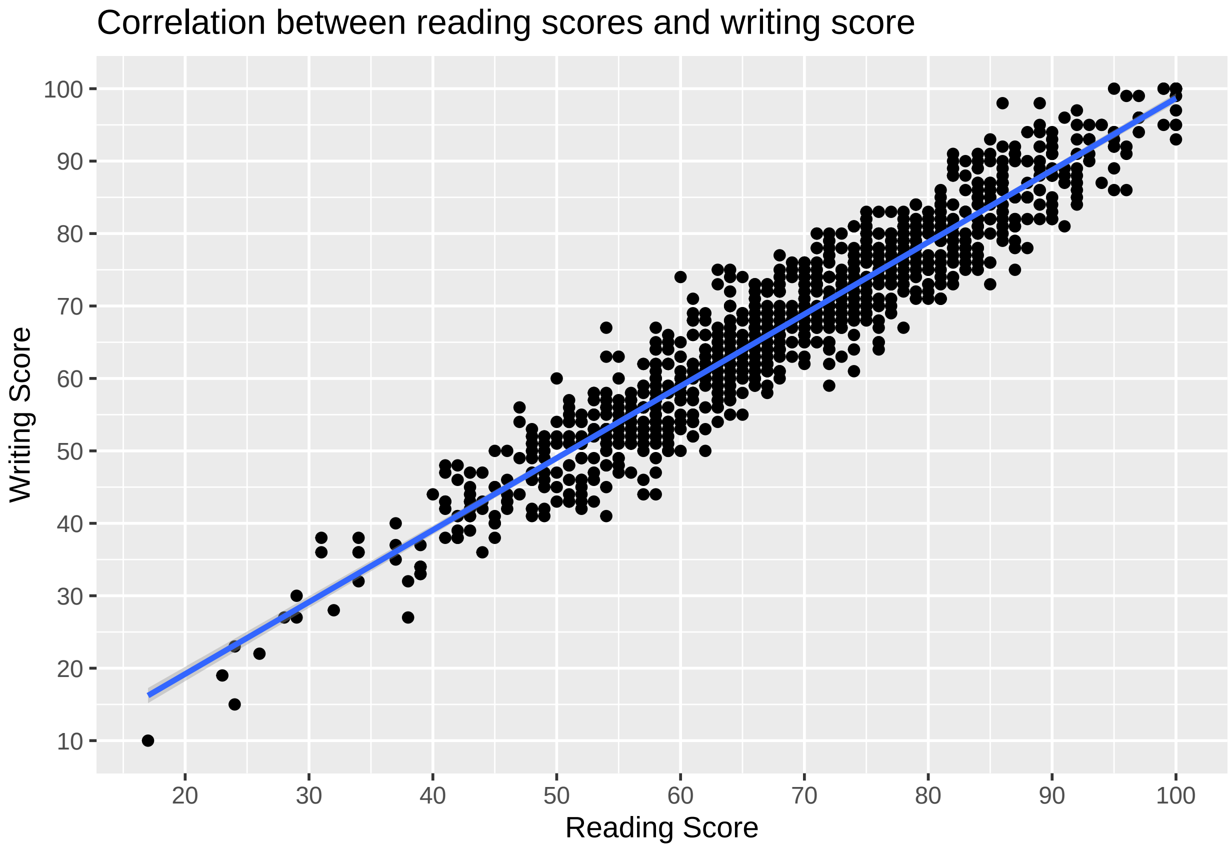
Figure 1. The proportion of females or males who completed the test preparation course (n=1000).

Scatterplot

- *Type of data summarized:* Bivariate quantitative data is summarized here.
- *Objective:* Does a high reading score correlate to a high writing score?

```
ggplot(rawDf,
       aes(x = reading.score, y = writing.score)) +
  geom_point(color = "black") +
  geom_smooth(method = lm) +
  labs(title = "Correlation between reading scores and writing score") +
  xlab("Reading Score") + ylab("Writing Score") +
  scale_x_continuous(breaks = seq(0, 100, by=10)) +
  scale_y_continuous(breaks = seq(0, 100, by=10))
```

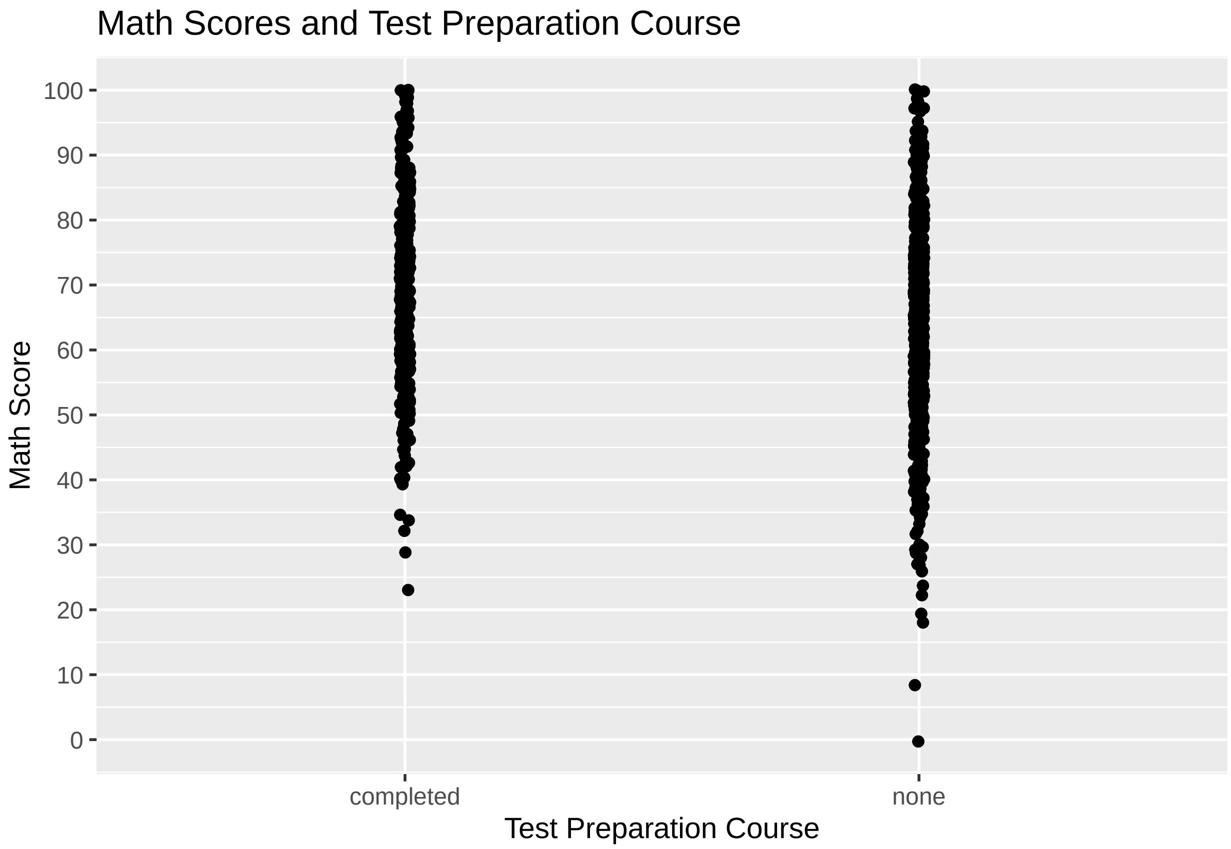
```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



Strip Chart

- *Type of data:* Univariate, bivariate and multivariate data that is quantitative and qualitative is summarized here.
- *Objective:* Will a participant who took the test preparation course score higher on the math test than a participant who did not take the test preparation course?

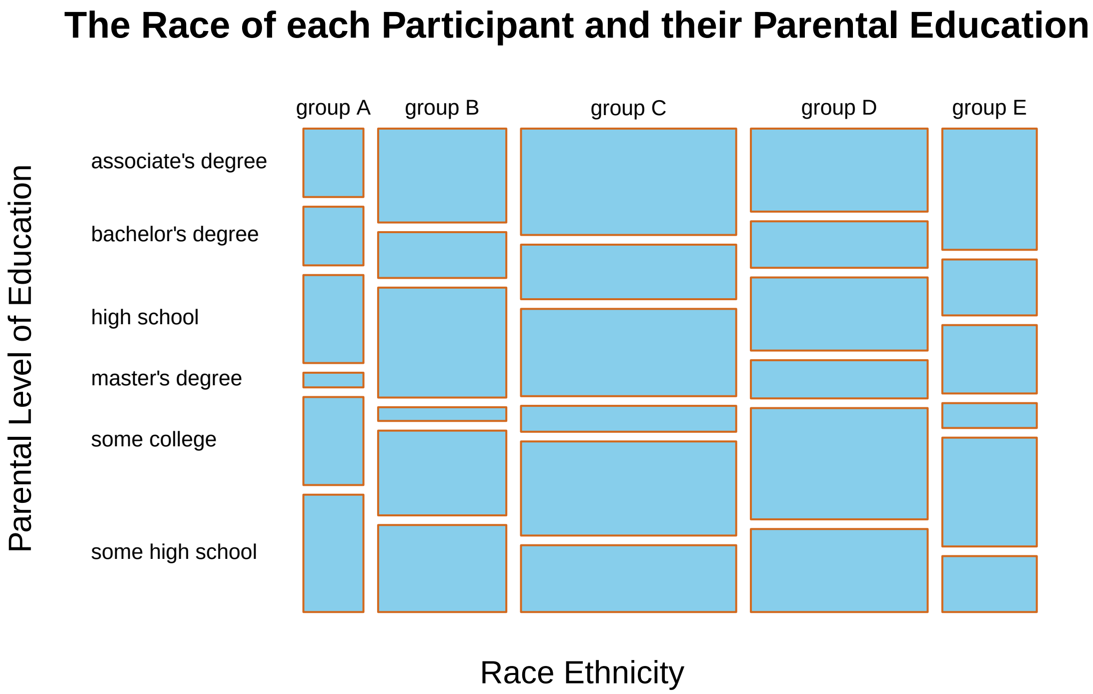
```r
ggplot(rawDf,
       aes(x = test.preparation.course, y = math.score)) +
  geom_jitter(position = position_jitter(0.01)) +
  scale_y_continuous(breaks = seq(0,100, by = 10)) +
  xlab("Test Preparation Course") + ylab("Math Score") +
  labs(title = "Math Scores and Test Preparation Course")
```



Mosaic Plot

- *Type of data summarized:* Bivariate and multivariate data. Usually qualitative and quantitative variables can be summarized here.
- *Objective:* Is there a correlation between the race of a participant and the level of education their parents achieved?

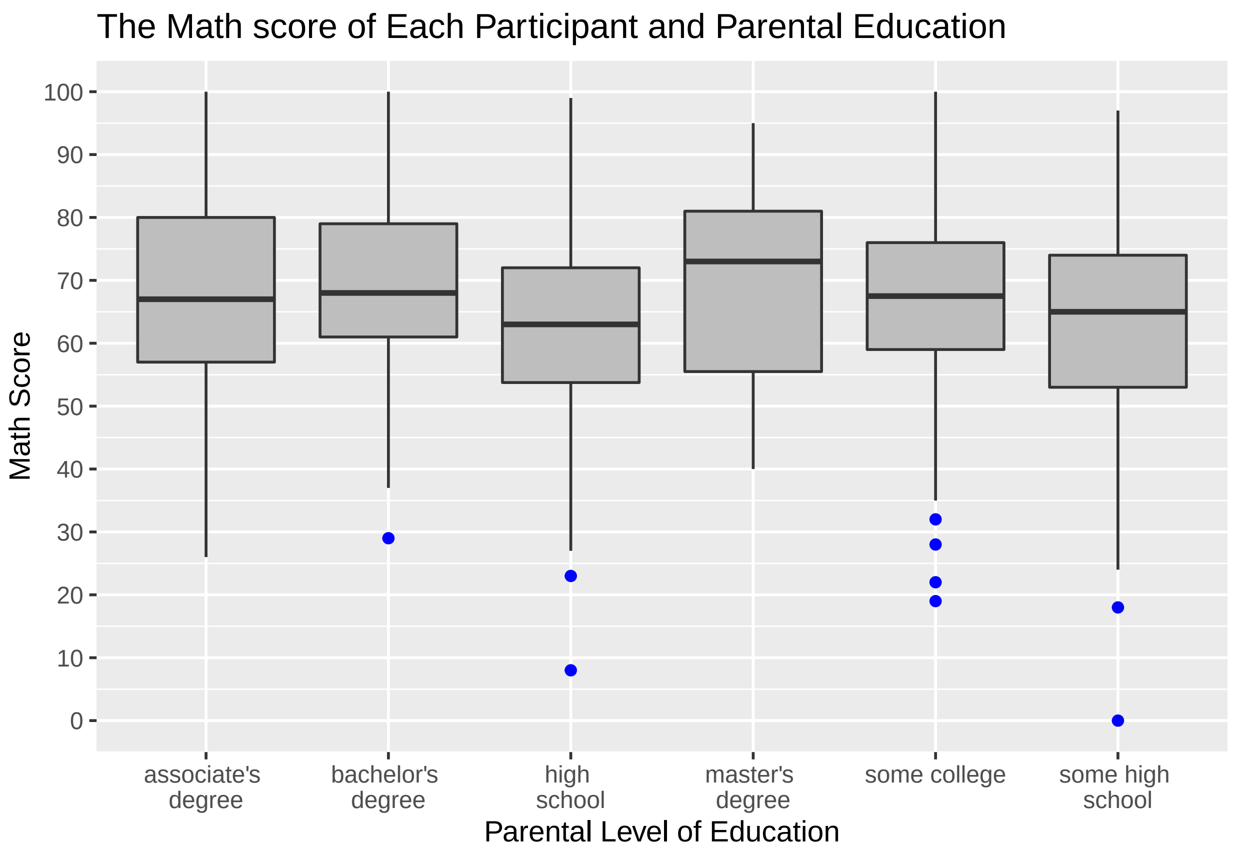
```r
table1 <- table(rawDf$race.ethnicity, rawDf$parental.level.of.education)
mosaicplot(table1, main = "The Race of each Participant and their Parental Education ",
           xlab = "Race Ethnicity",
           ylab = "Parental Level of Education",
           las = 1,
           color = "skyblue",
           border = "chocolate")
```



Boxplot

- *Type of data summarized:* Univariate, bivariate and multivariate data that is quantitative and qualitative is summarized here.
- *Objective:* How does a participant's math score compare with their parent's education level achieved? (High school, bachelor's degree, or associate's degree)

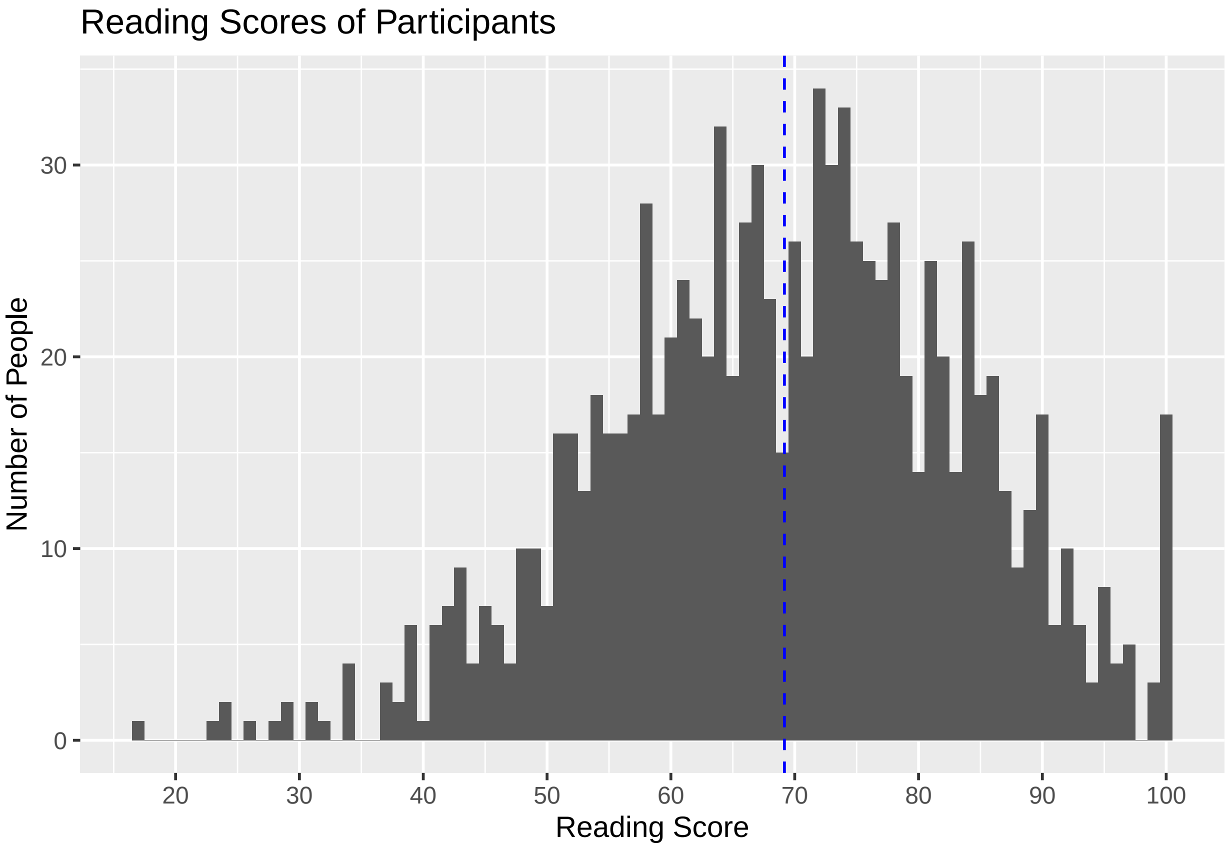
```
ggplot(rawDf,
       aes(x = parental.level.of.education, y = math.score)) +
  geom_boxplot(fill = "gray", outlier.color = "blue") +
  scale_x_discrete(labels = c("associate's
degree",
                              "bachelor's
degree",
                              "high
school",
                              "master's
degree", "some college",
"some high
school")) +
  labs(x = "Parental Level of Education",
       y = "Math Score", title = "The Math score of Each Participant and Parental Education") +
  scale_y_continuous(breaks = seq(0,100, by=10))
```



Histogram

- *Type of data summarized:* Bivariate, multivariate data that is quantitative and qualitative are summarized here.
- *Objective:* What was the distribution of reading scores among participants?

```
ggplot(rawDf,
       aes(x = reading.score)) +
  geom_histogram(binwidth = 1) +
  labs(x = "Reading Score", y = "Number of People", title = "Reading Scores of Participants") +
  geom_vline(aes(xintercept = mean(reading.score)),
             color = "blue",
             linetype = "dashed") +
  scale_x_continuous(breaks = seq(0,100, by=10)) +
  scale_y_continuous(breaks = seq(0,35, by=10))
```



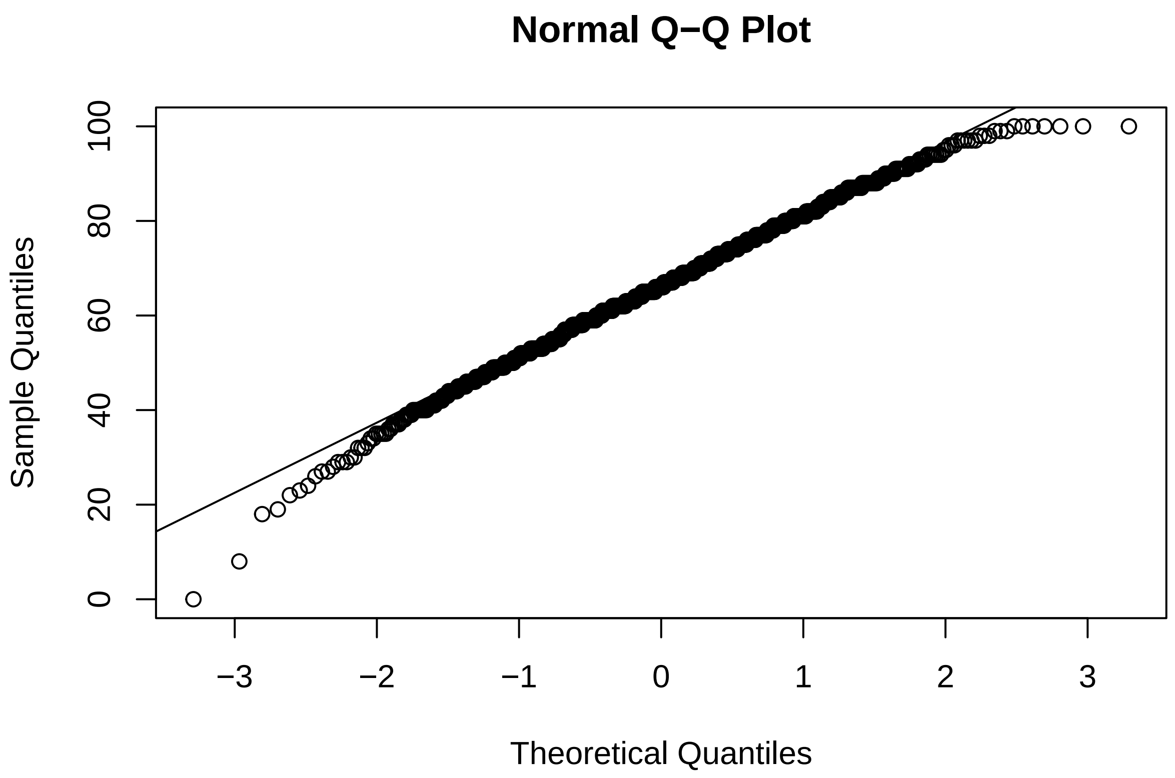
Analysis

"One sample inference"

1. T confidence interval for mean

- **Question:** What is the average score in Math for high schools students in America?
- **Conditions/assumptions:**
  - Sample data must be from a simple random sample; if this was true we would know that the mean of the sampling distributions of sample means is equal in value to the population mean, and the standard deviation of the sampling distribution of sample means is equal in value to the population standard deviation divided by the square root of the sample size
  - Sampling distribution of sample means can be modeled by a normal model; this ensures the confidence level was related to the multiplier
  - Standard deviation of the population distribution is known
- **R function:**

```
qqnorm(rawDf$math.score)
qqline(rawDf$math.score)
```



```
t.test(x=rawDf$math.score, conf.level=0.95)
```

```
##
##  One Sample t-test
##
## data:  rawDf$math.score
## t = 137.83, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  65.14806 67.02994
## sample estimates:
## mean of x
##    66.089
```

2. Large sample confidence interval for proportion

- **Question:** What proportion of students have a reduced lunch?
- **Conditions/assumptions:**
  - We needed a simple random sample, which was how our data was collected - each student had an equal chance of being selected for the study. The count must be described by a binomial model, which is true; a "success" would be the student has a reduced lunch, while a "failure" would be the student does not have a reduced lunch. And a normal approximation of the binomial model is reasonable, which is true when the product of our sample size and probability of success as well as the product of the sample size and probability of failure are both greater than 15, which is true in this case.

- **R function:**

```
# table()
# length()
# prop.test(x=, n=, conf.level=)
```

a. `table()`: This line of code is for computing our number of successes and failures
b. `length()`: This line of code is for computing our sample size
c. `prop.test(x=, n=, conf.level=)`: This line of code generates our large sample confidence interval. X is the number of successes, n is the sample size, and conf.level is the desired confidence level.

3. T test for mean

- **Question:** Do students with parents who have an above high school education level achieve higher math grades than the sample mean.
- **Conditions/assumptions:**
  - We need a simple random sample, which was how our data was collected (each students had an equal chance of being selected for the study).
  - T test for mean also require a population distribution that is Normal or sample size is larger than 50. Normality is tested using a Q-Q plot and the sample size is bigger than 50.
  - Observations from a sample of 1,000 students in American high schools are collected through simple random sampling and are independent samples of test preparation and math scores.
- **R function:**

```
# t.test(x=, mu=0, alternative= )
```

a. `x=` is asking for the vector of data (quantitative variable)
b. `mu=0` is the hypothesized value of population mean with the default being 0
c. `"alternative="` could take input “less”, “greater” and “two.sided”, it shows the direction of our “tail” hypothesis (less than or greater than the null)

4. Large sample test for proportion

- **Question:** What proportion of students completed the test preparation course?
- **Conditions/assumptions:**
  - We needed a simple random sample, which was how our data was collected - each student had an equal chance of being selected for the study. The count must be described by a binomial model, which is true; a “success” would be the student has completed the test preparation course, while a “failure” would be the student has not completed the test preparation course. And a normal approximation of the binomial model is reasonable, which is true when the product of our sample size and probability of success as well as the product of the sample size and probability of failure are both greater than 10, which is true in this case.
- **R function:** prop.test(x=, n=, p=, alternative=, correct=)

```
# table()
# length()
# prop.test(x=, n=, p=, alternative=, correct=)
```

a. `table()`: This line of code is for computing our number of successes and failures

b. `length()`: This line of code is for computing our sample size
c. `prop.test(x= , n= , alternative=", correct =  )`: This line of code generates our large sample test for proportion. X is the number of successes, n is the sample size, p is the value of our null hypothesis, alternative shows the direction of our "tail" hypothesis (less than or greater than the null), and correct is whether a continuity correction is used or not (set to FALSE during this inference procedure)

"Two sample inference"

1. T confidence interval for the difference in means

- **Question:** Is there a significant difference between reading scores and math scores for high school students in America?
- **Conditions/assumptions:**
  - Sample data must be from a simple random sample; if this was true we would know that the mean of the sampling distributions of sample means is equal in value to the population mean, and the standard deviation of the sampling distribution of sample means is equal in value to the population standard deviation divided by the square root of the sample size. Sampling distribution of sample means can be modeled by a normal model; this ensures the confidence level is related to the multiplier. Standard deviation of the population distribution is known
- **R function:**

```
# t.test(x= , y=, mu=, alternative=")
```

a. `t.test(x= , y=, mu=, alternative=")`: This generates the confidence interval for differences in means. x= is asking for the vector of data (quantitative variable). mu=0 is the hypothesized value of population mean with the default being 0. "alternative=" could take input "less", "greater" and "two.sided", it shows the direction of our "tail" hypothesis (less than or greater than the null)

2. T test for the difference in means

- **Question:** Is there a significant difference between reading scores and math scores for high school students in america?
- **Hypotheses:**
  - Null Hypothesis = Reading and math scores are equal.
  - Alternative hypothesis = Reading scores are greater or lesser than math scores
  - Let y represent mean reading scores, and x represent mean math scores
    - $H_0 = x = y$
    - $H_A = y > x < y$
- **Conditions/assumptions:**
  - Both samples are simple random samples We used simple random sampling to collect our data. Samples are independent (not paired data) from populations that are Normally distributed
  - Our data does not meet this requirement. The test scores are paired values, since all three values originate from one student. We can check with R to see if our data is Normally distributed using a QQplot. According to the QQ plot, reading and writing scores are BLANK distributed.
- **R function:**

```
t.test(x = rawDf$reading.score, y = rawDf$math.score,
       mu = 0, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  rawDf$reading.score and rawDf$math.score
## t = 4.6271, df = 1995.1, p-value = 3.947e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.774566 4.385434
## sample estimates:
## mean of x mean of y
##    69.169    66.089
```

3. Large sample confidence interval for difference in proportions

- **Question:** Does the gender of a student influence their chances of taking a test preparation course?
- **Conditions/assumptions:**
  - We needed a simple random sample, which was how our data was collected - each student had an equal chance of being selected for the study. The count must be described by a binomial model, which is true; a "success" would be the student has completed the test preparation course, while a "failure" would be the student has not completed the test preparation course. And a normal approximation of the binomial model is reasonable, which is true when the product of our first and second sample size and probability of success as well as the product of the first and second sample size and probability of failure are both greater than 5 (first and second referring to the two samples that we are subtracting from each other to make a "difference" in proportion).
- **R function:**

```
# table()
# length()
# prop.test(x= , n= , alternative=", correct= )
```

a. `table()`: This line of code is for computing our number of successes and failures
b. `length()`: This line of code is for computing our sample size
c. `prop.test(x= , n= , alternative=", correct= )`: This line of code generates the large sample confidence interval for differences in proportions. x is the number of successes, n is the sample size, alternative shows the direction of our "tail" hypothesis (less than or greater than the null), and correct is whether a continuity correction is used or not.

4. Large sample test for differences in proportions

- **Question:** Does the gender of a student influence their chances of taking a test preparation course?
- **Conditions/assumptions:**
  - We needed a simple random sample, which was how our data was collected - each student had an equal chance of being selected for the study. The count must be described by a binomial model, which is true; a "success" would be the student has completed the test preparation course, while a "failure" would be the student has not completed the test preparation course. And a normal approximation of the binomial model is reasonable, which is true when the product of our first and second sample size and probability of success as well as the product of the first and second sample size and probability of failure are both greater than 5 (first and second referring to the two samples that we are subtracting from each other to make a "difference" in proportion).

- **R function:**

```
# table()
# length()
# prop.test(x= , n= , alternative=", correct= )
```

a. `table()`: This line of code is for computing our number of successes and failures
b. `length()`: This line of code is for computing our sample size
c. `prop.test(x= , n= , alternative=", correct= )`: This line of code generates our large sample test for proportion. X is the number of successes, n is the sample size, alternative shows the direction of our "tail" hypothesis (less than or greater than the null), and correct is whether a continuity correction is used or not (set to FALSE during this inference procedure)

"Simple linear regression"

1. T test for slope

- **Question:** Is there a significant correlation between reading scores and writing scores for high school students in america?
- **Hypotheses:**
  - Null Hypothesis: A student's reading score has no correlation with its writing score.
  - Alternative Hypothesis: A student's reading score is correlated with its writing score.
  - Let x be a student's reading score, y be a student's mean writing score, and β be the slope of the line.
    - $H_0 = \beta = 0$
    - $H_A = \beta \neq 0$
- **Conditions/assumptions:**
  - If a relationship exists between the explanatory (X) and response (Y) variable, that relationship can be described by a line
    - A linear relationship between the 2 variables can be found by looking at the data in scatterplot
  - The observations of our response (Y) variable are independent
    - Student's writing score would be independent
  - The values of the response (Y) variable for a particular value of the explanatory (X) variable must come from a population distribution that is normally distributed
    - Normal distribution tested using qq plot and residual plot
  - The standard deviation of population distribution of response (Y) variable is the same for all Y at X
    - The standard deviation of student's reading score is the same as the standard deviation of student's writing score
- **Conditions for inference on β**
  - Linear relationship between explanatory and response
  - Response variable observations are independent
  - Population distribution of response variable y is Normally distributed for each value of x
  - Constant variance/standard deviation all y-values at respective x-values
- **R function:**

```
logwritingscore <- log(rawDf$writing.score)
regress <- lm(formula = logwritingscore ~ rawDf$reading.score)
summary(regress)
```

```
##
## Call:
## lm(formula = logwritingscore ~ rawDf$reading.score)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04525 -0.04931  0.00392  0.05584  0.26093
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.0729437  0.0141630  216.97   <2e-16 ***
## rawDf$reading.score 0.0161698  0.0002003   80.71   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.09245 on 998 degrees of freedom
## Multiple R-squared:  0.8671, Adjusted R-squared:  0.867
## F-statistic:  6514 on 1 and 998 DF,  p-value: < 2.2e-16
```

“Multi-sample inference”

1. One-factor ANOVA plus follow up analyses

- **Question:** Does taking a test preparation course influence math scores?
- **Conditions/assumptions:**
  - Observations from a sample of 1,000 students in American high schools are collected through simple random sampling and are independent samples of test preparation and math scores.
  - Each sample of test preparation course success/failure and math scores is a simple random sample from its population
  - Each sample of test preparation completion and math score observations come from Normally distributed populations
  - Standard deviation (or variance) of the population distribution is the same across the populations
- **R function:**

```
# aov(formula)
# newvariable<-aov(x~y)
# summary()
```

a. `aov(formula)`: This line of code computes a One-Way ANOVA test. The formula must have the vector where the response is stored ~ the vector where the explanatory variable is stored. The other defaults are fine.
b. `newvariable<-aov(x~y)`: Saves the ANOVA results as a new variable - makes the ANOVA data easier to work with
c. `summary()`: Displays the ANOVA results

Conclusion

One Sample Inference (Confidence Interval)

T confidence interval for mean

What is the average math score for high school students in America? - The mean math score for high school students in America is 65-67% (with 95% confidence, df=999)

Two Sample Inference (Hypothesis Test)

T test for the difference in means

Is there a significant difference between reading scores and math scores for high school students in america? - The mean difference in math scores for high school students in america ($\overline{x}$ = 66, n = 1000) was significant different for those in reading scores ($\overline{x}$ = 69, n = 1000)(t = 4.6271, df = 1995.1, P = 3.947e-06)

Simple Linear Regression

T test for slope

Is there a significant correlation between reading scores and writing scores for high school students in america? - There is a relationship between reading scores and writing scores for high school students in america, described by *log* writing score = 3.073 + 0.016*reading scores (t = 80.71. df = 998, $P < 2 \times 10^{16}$, $r^2 = 0.867$). The slope of this relationship is estimated to be

The helpful hints/reminder section

1. Use knit often when working in R Markdown to ensure the formatting is proper.
2. R is case sensitive, so double check which variables you are working with, and keep this in mind when naming new variables.
3. Add comments to complex sections of the code, so when coming back to it later, the purpose and function of the code can be quickly understood.

Dataset information

- Gender: represents the gender of the participant. Could be either male or female.
    - Categorical, nominal variable
- Race/ethnicity: represents the race/ethnicity of the participant. Could be one of four responses, group A (representing BLANK), group B (representing BLANK), group C (representing BLANK) or group D (representing BLANK)
    - Categorical, nominal variable
- Parental level of education: represents the highest form of education that either of the parents of the participant achieved. Could be high school, college, master's degree, or associate's degree, or any of those options with the prefix "some".
    - Categorical, ordinal variable

- Lunch: represents the price of lunch that the participant paid. Could be standard, or free/reduced.
    - Categorical, nominal variable
- Test preparation course: represents whether or not the participant completed an exam preparation course on the subject that will be tested on.
    - Categorical, ordinal variable
- Math score: represents the participants' percentage score on a math test.
    - Quantitative, interval variable, discrete
- Reading score: represents the participants' percentage score on a reading test.
    - Quantitative, interval variable, discrete
- Writing score: represents the participants' percentage score on a writing test.
    - Quantitative, interval variable, discrete

Reference

- Seshapanpu, J. (2018, November 9). Students performance in exams. Kaggle. Retrieved February 17, 2022, from https://www.kaggle.com/spscientist/students-performance-in-exams

Group Reflection

While this Project had some very specific requirements and structure, the overall theme was to work with some 'real' research questions and data, emulating (to a certain degree) the process of conducting research and working with data. What aspects of working on this Project and working/learning to use R were unexpected (could be positive and/or negative) for your group, based on your prior understanding/experience of research and working with data? Briefly explain the impact that these unexpected aspects have/will have on your future interactions with research and/or data

- The most unexpected aspect of working on this project was how difficult it was to coordinate a uniform structure for our answers. This project was lengthy, and we divided some of the questions and/or code to answer them individually, and it was definitely very noticeable in terms of style of answer. Especially in R, as we all had cultivated our own ways we were comfortable working with R, and so it was harder to work with other people's code when we would inevitably switch to check everything over. We all had to change the way we wrote code to make it more universal, and therefore easier to work with - since it wasn't practical to sit down with a group member and walk them through our code. This concept of universality is easily extrapolated for future interactions with research/data: to ensure that whoever comes along to work with our data/code after we use it is able to work with it efficiently, we'll have to record metadata, and leave notes as to why we did what we did. This lesson from the project will increase the reproducibility of our future interactions with research/data.

What do you think the chance is (as a value from 0% to 100%) that you will refer back to this Resource File in the future (e.g. for another course, for a job, out of interest, etc.)?

- 80%