Homework 3

Spring 2023

Jaiden Atterbury

Instructions

- This homework is due in Gradescope on Wednesday April 26 by midnight PST.
- Please answer the following questions in the order in which they are posed.
- Don't forget to knit the document frequently to make sure there are no compilation errors.
- When you are done, download the PDF file as instructed in section and submit it in Gradescope.

Exercises

1. (Expected length) Suppose $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Norm(\mu_0, \sigma_0)$ where both parameters are unknown. Find the smallest n that will guarantee that the expected width of a 95% confidence interval for σ_0^2 is no greater than the true value of σ_0^2.

Setup:
Since $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Norm(\mu_0, \sigma_0)$, it follows that $\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$. Thus the confidence interval for σ_0^2 takes the form $\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right]$. Where χ^2_p represents the $1-p$th quantile of the chi square distribution with $n-1$ degrees of freedom. In particular, the width of a 95% confidence interval for σ_0^2 is $\frac{(n-1)S^2}{\chi^2_{0.975}} - \frac{(n-1)S^2}{\chi^2_{0.025}}$. Furthermore, the **expected** width of a 95% confidence interval for σ_0^2 is thus $E\left[\frac{(n-1)S^2}{\chi^2_{0.975}} - \frac{(n-1)S^2}{\chi^2_{0.025}}\right]$. Lastly, to find the smallest n that will guarantee that the expected width of a 95% confidence interval for σ_0^2 is no greater than the true value of σ_0^2, we must find the value of n that satisfies the following statement: $E\left[\frac{(n-1)S^2}{\chi^2_{0.975}} - \frac{(n-1)S^2}{\chi^2_{0.025}}\right] \leq \sigma_0^2$.

Calculation:

$$
\begin{aligned}
\sigma_0^2 &\geq E\left[\frac{(n-1)S^2}{\chi^2_{0.975}} - \frac{(n-1)S^2}{\chi^2_{0.025}}\right] \\
&\geq E\left[\frac{(n-1)S^2}{\chi^2_{0.975}}\right] - E\left[\frac{(n-1)S^2}{\chi^2_{0.025}}\right] \quad \text{(Linearity of Expectation)} \\
&\geq \frac{n-1}{\chi^2_{0.975}} E\left[S^2\right] - \frac{n-1}{\chi^2_{0.025}} E\left[S^2\right] \quad \text{(Linearity of Expectation)} \\
&\geq \left(\frac{n-1}{\chi^2_{0.975}} - \frac{n-1}{\chi^2_{0.025}}\right) E\left[S^2\right] \\
&\geq \left(\frac{n-1}{\chi^2_{0.975}} - \frac{n-1}{\chi^2_{0.025}}\right) \sigma_0^2 \quad \text{(Since } E\left[S^2\right] = \sigma_0^2\text{)} \\
1 &\geq \left(\frac{n-1}{\chi^2_{0.975}} - \frac{n-1}{\chi^2_{0.025}}\right) \quad \text{(Since } \sigma_0^2 > 0\text{)} \\
0 &\geq \left(\frac{n-1}{\chi^2_{0.975}} - \frac{n-1}{\chi^2_{0.025}}\right) - 1
\end{aligned}
$$

With this expression in terms on n we can use uniroot to solve for the value of n that satisfies this inequality.

```
# Define a function for the above expression:
find_n <- function(n) {
 (n - 1) / (qchisq(p = 0.025, df = n - 1)) - (n - 1) / (qchisq(p = 0.975, df = n - 1)) - 1
}

# Use uniroot to find the root of the above function:
value_of_n <- uniroot(f = find_n, lower = 2, upper = 100)$root
```

Conclusion:
As computed above the value of n that satisfies the inequality is 38.5621831. Therefore, the smallest n that will guarantee that the expected width of a 95% confidence interval for σ_0^2 is no greater than the true value of σ_0^2 is $n = 39$.

2. (Racial discrimination in the Labor Market) Does racial discrimination exist in the labor market? Or, should racial disparities in the unemployment rate be attributed to other factors such as racial gaps in educational attainment? To answer this question, two social scientists conducted the following experiment. In response to newspaper ads, the researchers sent out resumes of fictitious job candidates to potential employers. They varied only the names of the job applicants while leaving the other information in the resumes unchanged. For some resumes, stereotypically blacksounding names such as Lakisha Washington or Jamal Jones were used, whereas other resumes contained typically white-sounding names such as Emily Walsh or Greg Baker. The researchers then compared the callback rates between these two groups of resumes and examined whether the resumes with typical black names received fewer callbacks than those with stereotypically white names. The positions to which the applications were sent were either in sales, administrative support, clerical, or customer services.

   The data are in the file `resume.csv`. Each row represents a fictitious job applicant. For example, the second observation contains a resume of Kristin who is a white female who did not receive a callback.

a. Create a table (`tabyl`) summarizing the race of the applicant and whether or not they received a callback[1]. Your table

- should have the information for each race on different rows
- should show the total (`adorn_totals`) for the callback

[1] see the file 'HW3table.png' for an example of what your table should look like

- should show the row-wise percentages (`adorn_percentages`) for each of the cells using `adorn_percentages` (that is, what fraction of the row is in the cell)
- should have the percentages formatted to 2 digits (`adorn_pct_formatting`)
- should also have the frequencies (n) in each cell reported (`adorn_ns`)

  Show the code, output and also write a couple of sentences summarizing the data.

Creating the table:

```
# Read in the data:
resume_df <- read.csv("resume.csv")

# Create the tabyl:
resume_df %>% tabyl(var1 = race, var2 = call) %>%
                    adorn_totals(where = c("row")) %>%
                    adorn_percentages("row") %>%
                    adorn_pct_formatting(digits = 2) %>%
                    adorn_ns()
```

```
##   race             0           1
##  black 93.55% (2278) 6.45% (157)
##  white 90.35% (2200) 9.65% (235)
##  Total 91.95% (4478) 8.05% (392)
```

Context:
The above table was constructed from experimental data obtained from two social scientists who were trying to answer the question; does racial discrimination exist in the labor market? In this experiment the researchers sent out resumes of fictitious job candidates to potential employers. In these applications, the researchers varied only the names of the job applicants while leaving the other information in the resumes unchanged. For some resumes, stereotypically black sounding names were used, whereas other resumes contained stereotypically white sounding names. The main goal of this data was to see whether the resumes with stereotypical black sounding names received fewer callbacks than those with stereotypically white sounding names.

Summary of the data:
As can be seen in the above table there is an evident difference between the callback rates for applicants with stereotypically white sounding names compared to those with stereotypically black sounding names. In particular, the applicants with stereotypically white sounding names had 78 more callbacks than the applicants with stereotypically black sounding names. This relation obviously holds in the reverse order as the applicants with stereotypically white sounding names had 78 less rejections than the applicants with stereotypically black sounding names. But is this evident different significant enough to claim there is evidence of discrimination in the hiring processes of certain positions?

b. Is there evidence of discrimination? Calculate a 95% confidence interval for the difference in callback rates for black and white applicants. Please state your interval clearly and then write your conclusion in context. (You may use R as a calculator.)

Setup:

Although there is an evident difference between the number of callbacks/callback rates between the two different groups, we have yet to see if this difference is significant, or if it could just have happened by random chance. With that said, we will run a two sample proportion test to analyze this data and see if there truly is evidence of discrimination. In particular, we will test the hypotheses $H_0 : \pi_W - \pi_B = 0$ versus $H_1 : \pi_W - \pi_B \neq 0$, where π_W represents the callback rates for applicants with stereotypically white sounding names, and π_B represents the callback rater for applicants with stereotypically black sounding names.

Calculation:

```r
# Run a two sample proportion test to find the confidence interval:
prop_int <- mosaic::prop.test(call ~ race,
                              data = resume_df,
                              alternative = "two.sided",
                              conf.level = 0.95,
                              correct = FALSE)$conf.int
```

Conclusion:
As computed from the above two sample proportion tests, the 95 confidence interval for the difference between the two callback rates was [0.0167777, 0.047288]. Since 0 is not in this interval, we have significant evidence at the 5% level that there is a difference between the callback rates for applicants with stereotypically white sounding names and applicants with stereotypically black sounding names. In particular, we have significant evidence at the 5% level that applicants with stereotypically white sounding names have higher callback rates than applicants with stereotypically black sounding names. Hence, there is evidence that discrimination may be present in the hiring processes of certain sales, administrative support, clerical, or customer service positions.

3. Suppose
$$X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Norm(\mu_1, \sigma_0)$$
independently of
$$Y_1, Y_2, \ldots, Y_m \overset{i.i.d.}{\sim} Norm(\mu_2, \sigma_0).$$
Let S_1^2 be the usual unbiased estimator of σ_0^2 based on the X's, that is,
$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$
Similarly S_2^2 is the unbiased estimator of σ_0^2 based on the Y's.

Suppose we want to create a combined estimator - let's call it S_p^2 - of σ_0^2 by considering a *weighted average* of S_1^2 and S_2^2. In other words:
$$S_p^2 = cS_1^2 + (1-c)S_2^2$$
for some $0 < c < 1$. Show that $c = \frac{n-1}{n+m-2}$ will minimize $Var\left[S_p^2\right]$.

Setup:
If we suppose that the sample $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Norm(\mu_1, \sigma_0)$ is drawn independently from the sample $Y_1, Y_2, \ldots, Y_m \overset{i.i.d.}{\sim} Norm(\mu_2, \sigma_0)$. Then it follows that S_1^2 is the usual unbiased estimator of σ_0^2 based on the X's, and similarly S_2^2 is the usual unbiased estimator of σ_0^2 based on the Y's. Furthermore, if we let S_p^2 be a combined estimator of σ_0^2, then by considering a weighted average of S_1^2 and S_2^2, we obtain $S_p^2 = cS_1^2 + (1-c)S_2^2$ for some $0 < c < 1$. To show $c = \frac{n-1}{n+m-2}$ is the value of c which minimizes $Var[S_p^2]$, we will use the second derivative test. First, we will use properties of variance and manipulate the equation $S_p^2 = cS_1^2 + (1-c)S_2^2$:

$$\begin{aligned} Var\left[S_p^2\right] &= Var\left[cS_1^2 + (1-c)S_2^2\right] \\ &= Var\left[cS_1^2\right] + Var\left[(1-c)S_2^2\right] \quad \text{(Independence of the } X_i\text{'s and } Y_i\text{'s)} \\ &= c^2 Var\left[S_1^2\right] + (1-c)^2 Var\left[S_2^2\right] \quad \text{(Non-linearity of Variance)} \end{aligned}$$

Now we can think of the equation for $Var[S_p^2]$ as a function of c, then with this we can take derivatives to find the critical value and show that it's a minimum. Let $f(c) = Var[S_p^2]$:

$$\begin{aligned} f(c) &= c^2 Var\left[S_1^2\right] + (1-c)^2 Var\left[S_2^2\right] \\ f'(c) &= 2cVar\left[S_1^2\right] - 2(1-c)Var\left[S_2^2\right] \end{aligned}$$

Setting $f'(c) = 0$ we can find the critical value:

$$\begin{aligned}
0 &= 2cVar\left[S_1^2\right] - 2(1-c)Var\left[S_2^2\right] \\
(1-c)Var\left[S_2^2\right] &= cVar\left[S_1^2\right] \\
\frac{1-c}{c} &= \frac{Var\left[S_1^2\right]}{Var\left[S_2^2\right]} \\
\frac{1}{c} - 1 &= \frac{Var\left[S_1^2\right]}{Var\left[S_2^2\right]} \\
\frac{1}{c} &= \frac{Var\left[S_1^2\right]}{Var\left[S_2^2\right]} + 1 \\
c &= \frac{1}{\frac{Var\left[S_1^2\right]}{Var\left[S_2^2\right]} + 1}
\end{aligned}$$

However, before we can get an answer for the critical value in terms of n and m, we must show what $Var[S^2]$ is when $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Norm(\mu_0, \sigma_0)$. When $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Norm(\mu_0, \sigma_0)$, it follows that the statistic $\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$. Furthermore, the variance of a random variable that has a χ^2_{n-1} distribution is $2(n-1)$. Hence it follows that $Var\left[\frac{(n-1)S^2}{\sigma_0^2}\right] = 2(n-1)$. Through the use of the non-linearity of variance we can see that $Var[S^2] = \frac{2\sigma_0^4}{n-1}$ when $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Norm(\mu_0, \sigma_0)$. Hence $Var[S_1^2] = \frac{2\sigma_0^4}{n-1}$ and $Var[S_2^2] = \frac{2\sigma_0^4}{m-1}$. Plugging this back into the critical point expression we obtain:

$$\begin{aligned}
c &= \frac{1}{\frac{Var\left[S_1^2\right]}{Var\left[S_2^2\right]} + 1} \\
&= \frac{1}{\frac{\frac{2\sigma_0^4}{n-1}}{\frac{2\sigma_0^4}{m-1}} + 1} \\
&= \frac{1}{\frac{\frac{1}{n-1}}{\frac{1}{m-1}} + 1} \\
&= \frac{1}{\frac{m-1}{n-1} + 1} \\
&= \frac{1}{\frac{m-1}{n-1} + \frac{n-1}{n-1}} \\
&= \frac{1}{\frac{m+n-2}{n-1}} \\
&= \frac{n-1}{m+n-2}
\end{aligned}$$

Now, in order to show that $c = \frac{n-1}{m+n-2}$ minimizes $Var[S_p^2]$, we must show that $f''(\frac{n-1}{m+n-2}) > 0$:

$$\begin{aligned}
f''(c) &= 2Var[S_1^2] + 2Var[S_2^2] \\
&= \frac{4\sigma_0^4}{n-1} + \frac{4\sigma_0^4}{m-1}
\end{aligned}$$

Since $m, n > 1$ and $\sigma_0 > 0$ it follows that $f''(c) > 0,\ \forall c \in \mathbb{R}$. Hence, $f''(\frac{n-1}{m+n-2}) > 0$ and $c = \frac{n-1}{m+n-2}$ is thus the argmin for $Var[S_p^2]$.

4. The STAR (Student-Teacher Achievement Ratio) Project is a four year longitudinal study examining the effect of class size in early grade levels on educational performance and personal development.5 A longitudinal study is one in which the same participants are followed over time. This particular study lasted from 1985 to 1989 involved 11,601 students. During the four years of the study, students were randomly assigned to small classes, regular-sized classes, or regular-sized classes with an aid. In all, the experiment cost around $12 million. Even though the program stopped in 1989 after the first kindergarten class in the program finished third grade, collection of various measurements (e.g., performance on tests in eighth grade, overall high school GPA) continued through the end of participants' high school attendance.

   We will analyze just a portion of this data to investigate whether the small class sizes improved performance or not. The data file name is `STAR.csv`. The names and descriptions of variables in this data set are displayed in the codebook shown below. Note that there are a fair amount of missing values in this data set. For example, missing values arise because some students left a STAR school before third grade or did not enter a STAR school until first grade.

`race`	student's race (White = 1, Black = 2, Asian = 3, Hispanic= 4, Native American = 5, Others = 6)
`classtype`	type of kindergarten class (small = 1, regular = 2, regular with aid = 3)
`g4math`	total scaled score for math portion of fourth grade standardized test
`g4reading`	total scaled score for reading portion of fourth grade standardized test
`yearssmall`	number of years in small classes
`hsgrad`	high school graduation (did graduate = 1, did not graduate= 0)

   a. How does performance on fourth grade reading and math tests for those students assigned to a small class in kindergarten compare with those assigned to a regular-sized class? Do students in the smaller classes perform better? Give a brief substantive interpretation of the results. Show **tidy** output from `t_test` along with your code.

Data frame to analyze equal variance assumption:

```
STAR %>%
  filter(kinder != "regular-aid") %>%
  group_by(kinder) %>%
  summarise(sd_math = sd(g4math, na.rm = TRUE),
            sd_reading = sd(g4reading, na.rm = TRUE))
```

```
## # A tibble: 2 x 3
##   kinder  sd_math sd_reading
##   <chr>     <dbl>      <dbl>
## 1 regular    41.0       53.2
## 2 small      43.6       51.5
```

Since there is relatively no difference between the sample standard deviations across the two different classes for both test types, in the following t-tests we will use `var.equal = TRUE` assumption.

Setup for math scores:
To assess if performance on fourth grade math tests for students assigned to a small class in kindergarten is different than those assigned to a regular-sized class, we will run a two sample t test to test the competing hypotheses: $H_0 : \mu_S - \mu_A = 0$ versus $H_1 : \mu_S - \mu_A \neq 0$. Where μ_S is the average math test score for students in the small class size group, and μ_A is the average math test score for students in the average class size group.

We can run this test using the `t_test` function from the `infer` package in R.

t test for math:

```r
# Run two sample t test for average math scores:
STAR %>% infer::t_test(formula = g4math ~ kinder,
                       order = c("small", "regular"),
                       var.equal = TRUE)
```

```
## # A tibble: 1 x 7
##   statistic  t_df p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1    -0.158  1580   0.874 two.sided     -0.336    -4.51     3.84
```

Interpretation of results:
As we can see from the above two sample t test, the 95% confidence interval is $[-4.510427,\ 3.837941]$. Since 0 is included in this interval, we fail to reject the null hypothesis that $\mu_S - \mu_R = 0$. Thus we have no evidence that the average math score from individuals in the small class size group is any different than the average math score from individuals in the regular class size group. Hence we have no evidence that students in the smaller classes perform better on math tests than students in average class sizes.

Setup for reading scores:
To assess if performance on fourth grade reading tests for students assigned to a small class in kindergarten is different than those assigned to a regular-sized class, we will run a two sample t test to test the competing hypotheses: $H_0 : \mu_S - \mu_A = 0$ versus $H_1 : \mu_S - \mu_A \neq 0$. Where μ_S is the average reading test score for students in the small class size group, and μ_A is the average reading test score for students in the average class size group.

We can run this test using the `t_test` function from the `infer` package in R.

t test for reading:

```r
# Run two sample t test for average reading scores:
STAR %>% infer::t_test(formula = g4reading ~ kinder,
                       order = c("small", "regular"),
                       var.equal = TRUE)
```

```
## # A tibble: 1 x 7
##   statistic  t_df p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1      1.32  1560   0.188 two.sided       3.50    -1.71     8.72
```

Interpretation of results:
As we can see from the above two sample t test, the 95% confidence interval is $[-1.71492,\ 8.717385]$. Since 0 is included in this interval, we fail to reject the null hypothesis that $\mu_S - \mu_R = 0$. Thus we have no evidence that the average reading score from individuals in the small class size group is any different than the average reading score from individuals in the regular class size group. Hence we have no evidence that students in the smaller classes perform better on reading tests than students in regular class sizes.

b. Next, we examine whether the STAR program reduced the achievement gaps across different racial groups. Begin by re-coding the `race` variable by changing integer values to their corresponding informative labels. Be sure to print the frequency distribution of `race`. (Show your code for this part)

Recoding the STAR data:

```r
# Check to see if there are missing values in the race column:
num_nas <- sum(is.na(STAR$race))

# Recode the STAR data by race, removing the NA rows:
STAR <- STAR %>%
        filter(!is.na(race)) %>%
        mutate(race_new = case_when(race == 1 ~ "White",
                                    race == 2 ~ "Black",
```

```
                                race == 3 ~ "Asian",
                                race == 4 ~ "Hispanic",
                                race == 5 ~ "Native American",
                                race == 6 ~ "Others"))
```

Frequency table for the data:

```
# Output the frequency table for the race variable:
STAR %>% tabyl(race_new)
```

```
##         race_new    n      percent
##            Asian   14 0.0022144891
##            Black 2058 0.3255298956
##         Hispanic    5 0.0007908890
##  Native American    2 0.0003163556
##           Others    9 0.0014236001
##            White 4234 0.6697247706
```

As can be seen from the above frequency tables, the most frequent races appearing in the STAR experiment are white students, with a frequency of 4234, and black students, with a frequency of 2058. The other races only had 30 participants in this experiment, and there were also 3 rows with NA values.

c. Compare the average reading and math test scores between white and black students among those students who were assigned to regular classes with no aid. Conduct the same comparison among those students who were assigned to small classes. Give a brief substantive interpretation of the results of your analysis. Show **tidy** output from `t_test` along with the code.

Data frame to analyze equal variance assumption:

```
STAR %>%
  filter(kinder != "regular-aid") %>%
  filter(race_new == "White" | race_new == "Black") %>%
  group_by(kinder, race_new) %>%
  summarise(sd_math = sd(g4math, na.rm = TRUE),
            sd_reading = sd(g4reading, na.rm = TRUE))
```

```
## # A tibble: 4 x 4
## # Groups:   kinder [2]
##   kinder  race_new sd_math sd_reading
##   <chr>   <chr>      <dbl>      <dbl>
## 1 regular Black       39.6       55.0
## 2 regular White       41.1       51.1
## 3 small   Black       43.2       44.5
## 4 small   White       43.4       51.4
```

Since there is relatively no difference between the sample standard deviations across the two different classes and two different races for both test types, in the following t-tests we will use `var.equal = TRUE` assumption.

Setup for math test scores for white and black students assigned to regular class sizes:
To assess if performance on fourth grade math tests for students assigned to regular class sizes in kindergarten is different for white and black students, we will run a two sample t test to test the competing hypotheses: $H_0 : \mu_W - \mu_B = 0$ versus $H_1 : \mu_W - \mu_B \neq 0$. Where μ_W is the average math test score for white students in the average class size group, and μ_B is the average math test score for black students in the average class size group.

We can run this test using the `t_test` function from the `infer` package in R.

t test for math scores:

```r
# Filter the data to only get students in the regular class size group:
star_regular <- STAR %>%
  filter(kinder == "regular")

# Run two sample t test for average math scores:
star_regular %>% infer::t_test(formula = g4math ~ race_new,
                        order = c("White", "Black"),
                        var.equal = TRUE)
```

```
## # A tibble: 1 x 7
##   statistic  t_df p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1      3.24   836 0.00125 two.sided       12.9     5.07     20.7
```

Interpretation of results:
As we can see from the above two sample t test, the 95% confidence interval is [5.072817, 20.68339. Since 0 is not included in this interval, we reject the null hypothesis that $\mu_W - \mu_B = 0$ at the 5% level. Thus we have evidence that the average math score for white students in the regular class size group is different than the average math score from black students in the regular class size group. Hence we have evidence that white students in the regular class size group perform better on math tests than black students in the regular class size group do.

Setup for reading test scores for white and black students assigned to regular class sizes:
To assess if performance on fourth grade reading tests for students assigned to regular class sizes in kindergarten is different for white and black students, we will run a two sample t test to test the competing hypotheses: $H_0 : \mu_W - \mu_B = 0$ versus $H_1 : \mu_W - \mu_B \neq 0$. Where μ_W is the average reading test score for white students in the average class size group, and μ_B is the average reading test score for black students in the average class size group.

We can run this test using the `t_test` function from the `infer` package in R.

t test for reading scores:

```r
# Run two sample t test for average reading scores:
star_regular %>% infer::t_test(formula = g4reading ~ race_new,
                        order = c("White", "Black"),
                        var.equal = TRUE)
```

```
## # A tibble: 1 x 7
##   statistic  t_df  p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>    <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1      7.11   830 2.58e-12 two.sided       35.8     25.9     45.6
```

Interpretation of results:
As we can see from the above two sample t test, the 95% confidence interval is [25.88231, 45.63965]. Since 0 is not included in this interval, we reject the null hypothesis that $\mu_W - \mu_B = 0$ at the 5% level. Thus we have evidence that the average reading score for white students in the regular class size group is different than the average reading score from black students in the regular class size group. Hence we have evidence that white students in the regular class size group perform better on reading tests than black students in the regular class size group do. Furthermore, the sheer magnitude that the whole interval is bounded away from 0 shows that white students are performing much better on these tests than black students.

Setup for math test scores for white and black students assigned to small class sizes:
To assess if performance on fourth grade math tests for students assigned to small class sizes in kindergarten is different for white and black students, we will run a two sample t test to test the competing hypotheses: $H_0 : \mu_W - \mu_B = 0$ versus $H_1 : \mu_W - \mu_B \neq 0$. Where μ_W is the average math test score for white students in the small class size group, and μ_B is the average math test score for black students in the small class size group.

We can run this test using the `t_test` function from the `infer` package in R.

t test for math scores:

```
# Filter the data to only get students in the small class size group:
star_small <- STAR %>%
  filter(kinder == "small")

# Run two sample t test for average math scores:
star_small %>% infer::t_test(formula = g4math ~ race_new,
                        order = c("White", "Black"),
                        var.equal = TRUE)
```

```
## # A tibble: 1 x 7
##   statistic  t_df p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>   <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1      3.11   734 0.00195 two.sided       13.7     5.04     22.3
```

Interpretation of results:
As we can see from the above two sample t test, the 95% confidence interval is [5.04235, 22.32899]. Since 0 is not included in this interval, we reject the null hypothesis that $\mu_W - \mu_B = 0$ at the 5% level. Thus we have evidence that the average math score for white students in the small class size group is different than the average math score from black students in the small class size group. Hence we have evidence that white students in the small class size group perform better on math tests than black students in the class class size group do.

Setup for reading test scores for white and black students assigned to small class sizes:
To assess if performance on fourth grade reading tests for students assigned to small class sizes in kindergarten is different for white and black students, we will run a two sample t test to test the competing hypotheses: $H_0 : \mu_W - \mu_B = 0$ versus $H_1 : \mu_W - \mu_B \neq 0$. Where μ_W is the average reading test score for white students in the small class size group, and μ_B is the average reading test score for black students in the small class size group.

We can run this test using the `t_test` function from the `infer` package in R.

t test for reading scores:

```
# Run two sample t test for average reading scores:
star_small %>% infer::t_test(formula = g4reading ~ race_new,
                        order = c("White", "Black"),
                        var.equal = TRUE)
```

```
## # A tibble: 1 x 7
##   statistic  t_df      p_value alternative estimate lower_ci upper_ci
##       <dbl> <dbl>        <dbl> <chr>          <dbl>    <dbl>    <dbl>
## 1      5.68   720 0.0000000195 two.sided       29.2     19.1     39.3
```

Interpretation of results:
As we can see from the above two sample t test, the 95% confidence interval is [19.12367, 39.32589]. Since 0 is not included in this interval, we reject the null hypothesis that $\mu_W - \mu_B = 0$ at the 5% level. Thus we have evidence that the average reading score for white students in the small class size group is different than the average reading score from black students in the small class size group. Hence we have evidence that white students in the small class size group perform better on reading tests than black students in the small class size group do. Furthermore, the sheer magnitude that the whole interval is bounded away from 0 shows that white students are performing much better on these tests than black students.

Conclusions:
In all four cases, using a two sample t test we have shown that white students perform better on reading and math tests than their fellow black students do. In the case of reading scores, in both class sizes, white

students performed much better on these tests than their fellow black peers. With this said, it is evident that their is some disconnect between the teachers and the black students in the class, as it seems as if black students aren't learning well from the current teaching style of these schools. At the very least this difference needs to be addressed, and furthermore it is apparent that a change in regime or style is needed in order to decrease the gap between these scores, and increase these scores in general.