Homework 3

Spring 2023

Tanner Huck

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.

Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Norm(\mu_0, \sigma_0)$, where both parameters are unknown. Our goal is to find the smallest n such that we grantee the expected width of a 95% confidence interval for σ_0^2 is no greater than the true value of σ_0^2. Since each X_i is from a normal distribution, we know that the statistic $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a chi-squared distribution with $n-1$ degrees of freedom. Hence $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. Then we can create the 95% confidence interval for σ_0^2,

$$[\frac{(n-1)S^2}{\chi^2_{n-1,0.975}}, \frac{(n-1)S^2}{\chi^2_{n-1,0.025}}]$$

Then we want the expected width of this confidence interval to be less than or equal to true value of σ_0^2,

$$E[\frac{(n-1)S^2}{\chi^2_{n-1,0.025}} - \frac{(n-1)S^2}{\chi^2_{n-1,0.975}}] \leq \sigma_0^2$$

Then finding the smallest n for this to hold true,

$$E[(n-1)S^2(\frac{1}{\chi^2_{n-1,0.025}} - \frac{1}{\chi^2_{n-1,0.975}})] \leq \sigma_0^2$$

by linearity of expectation

$$(n-1)(\frac{1}{\chi^2_{n-1,0.025}} - \frac{1}{\chi^2_{n-1,0.975}})E[S^2] \leq \sigma_0^2$$

$$(n-1)(\frac{1}{\chi^2_{n-1,0.025}} - \frac{1}{\chi^2_{n-1,0.975}})\sigma_0^2 \leq \sigma_0^2$$

$$(n-1)(\frac{1}{\chi^2_{n-1,0.025}} - \frac{1}{\chi^2_{n-1,0.975}}) \leq 1$$

$$(n-1)(\frac{1}{\chi^2_{n-1,0.025}} - \frac{1}{\chi^2_{n-1,0.975}}) - 1 \leq 0$$

Then using uniroot to find n,

```
fun <- function(x) {
  (x-1)*(1/qchisq(0.025, x-1) - 1/qchisq(0.975, x-1)) - 1
}
smallest_n <- ceiling(uniroot(fun, lower=2, upper=100)$root)
```

Hence the smallest n such 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 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
- 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.

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

```r
df <- read.csv("resume.csv")

table <- df %>%
  tabyl(race, call) %>%
  adorn_totals(c("row")) %>%
  adorn_percentages("row") %>%
  adorn_pct_formatting(digits = 2) %>%
  adorn_ns()

table
```

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

This table shows us the number of callbacks, no callbacks, and total observations sorted by race. The row wise percentages are also shown, for example the 93.55% percentage represents the number of people who did not get callbacks who are also black.

This table can show us that black people have a slightly high chance of not reciving a call back. This suggests that there may in fact be some racial discrimination is the labor market.

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.)

```r
white_rate <- 2278 / (2278 + 157)
black_rate <- 2200 / (2200 + 235)

sig <- sqrt(white_rate*(1-white_rate)/2435 + black_rate*(1-black_rate)/2435)

lower <- (white_rate - black_rate) - qnorm(0.975) * sig
upper <- (white_rate - black_rate) + qnorm(0.975) * sig
```

The 95% confidence interval for the difference in callback rates between black and white applicants is [0.02, 0.05]. Since zero is not in this interval, we conclude that there is enough evidence against the null hypothesis that, there is discrimination in the labor market. Hence, with 95% there is a significant difference between the callback rates for whites vs blacks in the labor market.

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 $Var\left[S_p^2\right]$.

First starting with $Var\left[S_p^2\right]$ and substituting in the given equation for S_p^2,

$$
\begin{aligned}
Var[S_p^2] &= Var[cS_1^2 + (1-c)S_2^2] \\
&\text{by nonlinearity of varience} \\
&= c^2 Var[S_1^2] + (1-c)^2 Var[S_2^2] \\
&\text{since } S_1^2 \text{ and } S_2^2 \text{ are independent and both } S_1^2 \text{ and } S_2^2 \text{ are unbiased estimators of } \sigma_0^2 \\
&\text{based on the } X\text{'s and } Y\text{'s respectivily,} \\
&= c^2 \frac{2\sigma_0^4}{n-1} + (1-c)^2 \frac{2\sigma_0^4}{m-1}
\end{aligned}
$$

The last line above follows from the variance of a chi squared distribution being equal to 2 times he degrees of freedom,

$$
\begin{aligned}
Var(\frac{(n-1)S^2}{\sigma^2}) &= 2(n-1) \\
\frac{(n-1)^2}{\sigma^4} Var(S^2) &= 2(n-1) \\
Var(S^2) &= \frac{2\sigma^4}{(n-1)}
\end{aligned}
$$

Then to find the minimum of this we will find the derivative and set it equal to 0, to solve for the critical point c.

$$
\begin{aligned}
\frac{d}{dc}\left(c^2 \frac{2\sigma_0^4}{n-1} + (1-c)^2 \frac{2\sigma_0^4}{m-1}\right) &= 2c\frac{2\sigma_0^4}{n-1} - 2(1-c)\frac{2\sigma_0^4}{m-1} \\
&= \frac{4c\sigma_0^4}{n-1} - \frac{4(1-c)\sigma_0^4}{m-1} \\
&= 4\sigma_0^4(\frac{c}{n-1} - \frac{1-c}{m-1}) \\
&\text{then setting it equal to 0 and solving for c,} \\
0 &= 4\sigma_0^4(\frac{c}{n-1} - \frac{1-c}{m-1}) \\
0 &= (\frac{c}{n-1} - \frac{1-c}{m-1}) \\
\frac{1-c}{m-1} &= \frac{c}{n-1} \\
(n-1)(1-c) &= c(m-1) \\
n - 1 - nc + c &= cm - c \\
n - 1 &= cm + nc - 2c \\
\frac{n-1}{n+m-2} &= c
\end{aligned}
$$

Then using the second derivative test,

$$\frac{d}{dc}4\sigma_0^4(\frac{c}{n-1}-\frac{1-c}{m-1}) = 4\sigma_0^4\frac{d}{dc}(\frac{c}{n-1}-\frac{1-c}{m-1})$$
$$= 4\sigma_0^4(\frac{1}{n-1}+\frac{1}{m-1})$$

Here we can see that the second derivative is always positive. This is because both n and m were defined to be at least larger than 2 and varience is always non negative, hence a product of two posities yields a positive. Thus we know that our critical point is a global minimum, meaning $c = \frac{n-1}{n+m-2}$ will minimize $Var\left[S_p^2\right]$.

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.

To compare the performance of fourth grade reading and math tests for students in small classes vs regular sized classes, we can use an independent sample t-test. Here the Null hypothesis is that there is no difference in math or reading scores based on class sizes, $H_0 : \mu_{sizenormal} - \mu_{sizesmall} = 0$. The alternative hypothesis's is that there is a difference in math and reading scores based on class sizes $H_1 : \mu_{sizenormal} - \mu_{sizesmall} \neq 0$. (Two separate null and alternative hypothesis's, one for math and one for reading).

```
STAR %>% group_by(kinder) %>%
summarize(sd_math = sd(g4math, na.rm = TRUE),
         sd_reading = sd(g4reading, na.rm = TRUE))
```

```
## # A tibble: 3 x 3
##   kinder      sd_math sd_reading
##   <chr>         <dbl>      <dbl>
## 1 regular        41.0       53.2
## 2 regular-aid    44.7       52.4
## 3 small          43.6       51.5
```

```r
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
```

```r
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
```

Before running the t_tests, we can calculate the standard deviation of the math and reading scores separated by class size. From the table we can see that the stand deviation for math for the regular seize is about 41 whereas it is about 43 for the small size. Since these are similar enough, we can use the " var.equal = TRUE" option for our t_test. The same conclusion can be made for the reading scores.

These tests results give us a 95% confidence interval of about $[-4.51, 3.84]$ for the math scores and $[-1.71, 8.72]$ for the reading scores. For both of these intervals, we contain 0. This tells us that we do not have enough evidence to reject the null hypothesis that the average scores are the same. Hence we conclude that students do not preform better ion math or reading based on their class size.

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. Call the new variable `race_desc`. Be sure to print the frequency distribution of `race_desc`. (Show your code for this part)

```r
STAR_race <- STAR %>%
  filter(!is.na(race)) %>%
  mutate(race_desc = case_when(race == 1 ~ "White",
                               race == 2 ~ "Black",
                               race == 3 ~ "Asian",
                               race == 4 ~ "Hispanic",
                               race == 5 ~ "Native American",
                               race == 6 ~ "Other"))
STAR_race %>% tabyl(race_desc) %>% adorn_pct_formatting(digits = 2)
```

```
##        race_desc    n percent
##            Asian   14   0.22%
##            Black 2058  32.55%
##         Hispanic    5   0.08%
##  Native American    2   0.03%
##            Other    9   0.14%
##            White 4234  66.97%
```

From this table we can see that there is a much larger percentage of whites and blacks compared to the other races.

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.

```
STAR_race %>% filter(kinder == "small") %>% group_by(race_desc) %>%
summarize(sd_math_small = sd(g4math, na.rm = TRUE),
         sd_reading = sd(g4reading, na.rm = TRUE))
```

```
## # A tibble: 6 x 3
##   race_desc       sd_math_small sd_reading
##   <chr>                   <dbl>      <dbl>
## 1 Asian                   NA          NA
## 2 Black                   43.2        44.5
## 3 Hispanic                 6.36       31.8
## 4 Native American         NA          NA
## 5 Other                   NA          NA
## 6 White                   43.4        51.4
```

```
STAR_race %>% filter(kinder == "regular") %>% group_by(race_desc) %>%
summarize(sd_math_small = sd(g4math, na.rm = TRUE),
         sd_reading = sd(g4reading, na.rm = TRUE))
```

```
## # A tibble: 5 x 3
##   race_desc       sd_math_small sd_reading
##   <chr>                   <dbl>      <dbl>
## 1 Asian                    17.0       41.5
## 2 Black                    39.6       55.0
## 3 Native American          NA         NA
## 4 Other                    NA         NA
## 5 White                    41.1       51.1
```

```
STAR_race_small <- STAR_race %>% filter(kinder == "small")
STAR_race_small %>% infer::t_test(formula = g4math ~ race_desc,
                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
```

```
STAR_race_small %>% infer::t_test(formula = g4reading ~ race_desc,
                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
```

```
STAR_race_normal <- STAR_race %>% filter(kinder == "regular")
STAR_race_normal %>% infer::t_test(formula = g4reading ~ race_desc,
                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
```

```
STAR_race_normal %>% infer::t_test(formula = g4reading ~ race_desc,
                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
```

Once agian, before running the t_tests, we can calculate the standard deviation of the math and reading scores separated by class size for both whites and blacks. From the first table we can see that for small class sizes, the standard deviation for math scores for both whites and blacks is about 43. Additionally the standard deviations for reading scores is a bit farther apart, but still relatively close together. Since these are similar enough, we can use the " var.equal = TRUE" option for our t_test. The same conclusion can be made in the second table with the math and reading scores for blacks and whites in regular sized classes.

We can set up a test for the performance of fourth grade reading and math tests for students based on race (and separated by class size) similarly to part a. Our null hypotheses is that there is no difference in scores based on race, and the alternative is that there is a difference. The code to compute independent sample t-tests is shown above. You can find the exact confidence intervals above as well, but the conclusion for each confidence interval is similar. In each interval there is an absence of 0, meaning the intervals are entirely positive or negative. This shows us that we do have enough evidence to reject the null hypothesis that the average scores are the same. Hence we conclude that students do preform better on math or reading based on their race, in both normal and small class sizes.