Homework 3

Spring 2023

Your Name Here

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.

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

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.)

3. Suppose

$$X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \mu_1, \sigma_0$$

independently of

$$Y_1, Y_2, \ldots, Y_m \overset{i.i.d.}{\sim} \mu_2, \sigma_0.$$

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

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.

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)

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.