STA 104 Exam II Project, due by

10:00 PM, Wednesday, March 3^{rd} (PT) onto Gradescope

Read the following instructions carefully:

- **You may work in a group of two, or by yourself.**
- **You are not allowed to discuss the questions with anyone other than the instructor or TA and your group mate.**
- **Any outside help beyond that from the instructor or TA is considered plagiarism. This including asking a tutor, your classmates (for example, comparing answers), posting the questions to homework help sites, etc. Should we believe you have sought outside help, you will be reported to the Student Judicial Affairs office.**
- **You are allowed to use or modify your previous functions, or the instructors functions that are posted online.**
- **Do not share answers, or specific values for calculations, particularly on Piazza**.
- You may ask clarifying questions about code and general approach on Piazza, but do not give away any numerical answers. If you are concerned you may be giving something away, message me or the TA's directly.
- All R output should be **formatted**. You should **not** copy and paste directly from the R console.

Project Topic

For this project, we want to apply our class methods to a real dataset and solve two questions. We obtain provisional Covid-19 Death Counts by Sex, Age, and State from the Centers for Disease Control and Prevention (CDC) [1]. The raw dataset can be downloaded from the CDC website. Here, I provide cleaned datasets as `CovidA.csv` and `CovidB.csv`.

Question 1: K or more groups

The data is found in the file `CovidA.csv`, with the following columns:

Column 1: `Year`: Year in which death occurred (2020, 2021)

Column 2: `Month`: Month in which death occurred

Column 3: `State`: Jurisdiction of occurrence (50 States, District of Columbia, New York City, Puerto Rico)

Column 4: `Death`: Deaths involving COVID-19

We think State may affect the Covid-19 Death because each State has different covid regulations and infrastructure. Here you compare **4 States**, as specific as you can about your outcome.

Note:

- We assume each months are independent. (There is no time dependency.)
- First, you would subset the data which includes only your 4 states. (You can select District of Columbia, New York City, Puerto Rico if you like.) You should explain what 4 states you choose, why it is interesting, in introduction and present summary statistics regarding the 4 states.

Question 2: Tests for Independence

The data is found in the file `CovidB.csv`, with the following columns:

Column 1: `Year`: Year in which death occurred (2020, 2021)

Column 2: `Month`: Month in which death occurred

Column 3: `Sex`: Sex (Male or Female)

Column 4: `Age_Group`: Sex (Male or Female)

Column 5: `Death`: Deaths involving COVID-19

We are interested in the death counts between Age group and Sex at a certain time period. The goal is to determine if Age and Sex are independent variables regarding Deaths involving COVID-19. If they are not, specify what the dependence is.

Note:

- We assume each months are independent. (There is no time dependency.)
- The dataset has monthly death counts from March 2020 to January 2021. You can choose a certain period you interested or analyze all them. You should include what time period you work in introduction and present summary statistics regarding only the period.
- You may redefine your age group or choose certain age groups. In the case, you should have more than 3 age groups.
- The dataset you would import has frequencies of the death. It is different from our previous datasets in R handouts. You can use `xtabs()` to convert a data.frame to a table.

```
xtabs(Death ~ Sex + Age_Group, data = dataset)
```

[1] Data.CDC.gov `https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Sex-Age-and-S/9bhg-hcku`

3. The Report Format

Each question should be a short report. This means you write in full sentences, and have the following sections for each question, while being **as specific as you can** about your results. **There should not be any "copy and pasted" R code in this report. You must format the results you get from R.**

I. Introduction. State the question you are trying to answer, why it is a question of interest (why might we be interested in the answer), and what statistical technique you are going to use. **This must be a non-parametric technique**.

II. Summary of your data (and **only** the data you are using for the question). This should include things like plots (histograms, boxplots) including the interpretation of the plots, and summary values such as sample means and standard deviations. This is where you should justify which non-parametric technique you are using. An R handout is available online for graphing and summaries of various data types.

III. Analysis. Report back confidence intervals, test-statistics, and p-values, nulls and alternatives, etc. You may use tables here, but be sure that you organize your work. Remember to write your results in full sentences where possible.

IV. Interpretation. State your conclusion, and inference that you may draw from your corresponding tests or confidence intervals. These should all be in terms of your problem.

V. Conclusion. Summarize briefly your findings. Here you do not have to re-iterate your numeric values, but summarize all relevant conclusions. You may explain the limitation of this methods or mention any possible future works regarding the questions or data.

4. Details

Your report should be the following format:

i. Typed.

ii. A title page including your name/s, the name of the class, and the name of your instructor (me).

iii. Treat each question as a small, stand alone report.

iv. An appendix of your R code used to produce the results. Do not include in R code in the body of your report.

For example, your project should be put together in the following order:

```
Cover Page
Parts I-V for first question
Parts I-V for second question
Code appendix
```

Notice: your project will be graded as a group effort (if you have two people). This means that you are responsible for your own work, and your partners work. I will not assign two different grades to one project.

Rubric for Take Home Project

- Presence of content (40%)
    1. States statistical question of interest
    2. States why it is of interest
    3. States statistical technique to use
    4. Summary values and graphs
    5. Null/alternatives should be present or easily deducible from methods
    6. Results (test statistics, confidence intervals, p-values
    7. Interpretation of results
    8. Conclusion with respect to the statistical questions of interest
    9. R code present
- Accuracy of content (40%)
    1. Data is well-portrayed by summary values/graphs
    2. Technique answers the question of interest
    3. Technique is appropriate given the data
    4. Statistical methods are properly used (correct null/alternative, CI, p-values)
    5. Correct interpretation
    6. Conclusion follows from interpretation
    7. Conclusion satisfactorily addresses question of interest
    8. R code would generate the result
- Style of content (20%)
    1. Easy to find important information
    2. General style of report format is adhered to (okay if there are some deviations from I-VI format, as long as the order is logical)
    3. No raw R output in body of the report
    4. Graphs are clearly marked (labels, etc.)
    5. Some attempt at writing clearly (full sentences, etc.)