Math 156 Final Project Handout
Emma Kearney & Alex Bimm

Abstract
This project is a direct continuation of the Midterm Project. The dataset is a subset of the American Gut Project dataset, a publicly available online dataset for use in gut microbiome research. The data used for this project is a small fraction of the full dataset, with only a handful of the variables recorded for each sample. It consists of bacterial counts of a few families of bacteria and poll data focused on mental health and general statistics. The techniques used within this project will later be expanded upon and applied to the full dataset within Alex's senior thesis. The overall goal of the thesis is to explore and expand upon the theory that gut bacteria composition plays a role in neurotransmitter regulation and thus could be involved in mental disorders such as ADHD.

Summary of Findings
This analysis builds on the analysis conducted in the Midterm Project using a variety of new statistical tools. First, the script (final-project.R) explores the distribution of BMI (body mass index) which plausibly follows a gamma distribution by inspection. We use bootstrapping to estimate the sampling distribution of BMI, taking 10,000 samples of size 13,620 (the length of the BMI vector). From the mean and standard deviation estimates, we can calculate the appropriate parameters for a gamma distribution. Overlaying this gamma distribution to the bootstrap distribution of BMI means reveals the gamma distribution to be a strong fit. We also use both linear and logistic regressions to better understand the relationship between the variables of the dataset. We fit a linear regression to the age and BMI of patients in the sample, hypothesizing that BMI will increase with age. The regression indicates a weak positive relationship between these variables, suggesting that aging a year increases BMI by approximately 0.09 kg/meters^2. We also run a logistic regression to explore the relationship between BMI and Bulimia (indicator variable). The regression demonstrated interesting results, suggesting that a 1 kg/meter^2 increase in patient BMI effects a 0.021 decrease in the log-odds ratio of success:failure in having bulimia. We might expect individuals with higher BMIs to suffer from eating disorders, but it appears this might not be the case. To better affirm the findings of these regressions, it would be ideal to explore time-series data of individuals. We also explore whether levels of Bacteroidaceae are significantly different between males and females. The script demonstrates the process troubleshooting, first using the student t distribution but later finding it is highly inconsistent with a normal distribution, invalidating this method. Instead, we use bootstrapping to determine if there is a significant difference in means, taking 10,000 samples of the male subset and female subset respectively, then calculating the difference in their means. Zero is not included in the 95% confidence interval and it appears females have a statistically significant higher count of Bacteroidaceae than males.

RShiny Portion:
The RShiny portion of the project (app.R) is broken up into four main components. The first segment is Principle Component Analysis (PCA). Using a mixture of built in tools and plotting packages, this section provides a visual interpretation of the "bacterial space" and how various factors change across it. The second portion of the project is regression, which includes both linear and logistic types. This allows for the comparison of factors and numeric variables, as well as a comparison of the effectiveness of these two types of regression analysis. Thirdly, the RShiny explores Student t Confidence Intervals, which vary greatly in effectiveness across the numerical columns. Specifically, the highly skewed bacterial counts often end up defying the confidence intervals' expectations. Finally, the dataset is used to showcase the process of updating Bayesian Priors, taking samples from the large dataset and watching the posteriors converge upon the correct values for the entire dataset.

RShiny Link: https://abimm117.shinyapps.io/156FinalProject/

Bonus Points:

1. Use of regression, Student t, or Bayesian methods (not counted above) (2 points)
    1. All three are used within this project, as well as the Beta Distribution within the Bayesian Methods segment.
2. Calculation and display of a logistic regression curve
    1. Several logistic regression curves can be calculated within the regression segment. The script also includes a comparison of logistic and linear regression methods.
3. A data set with lots of columns, allowing comparison of many different variables.
    1. 15 columns in total, 5 factors, 10 numeric.
4. A graphical display that is different from those in the textbook or in the class scripts.
    1. PCA was discussed in class and shown with the COVID data script but I feel the visualization may still fit within this category.
5. Nicely labeled graphics, with good use of color, line styles, etc., that tells a convincing story.
6. Presentation is done with a Shiny app (2 points)
7. Interesting use of styles in a Shiny app.
8. Clicking or brushing on plots in a Shiny app.
    1. Bayesian Methods graph is interactive.
9. Appropriate use of bootstrap techniques (2 points)
    1. Utilized within the main script.