CE5033 Statistical Methods and Data Mining

Assignment 1

Development Environment

For this assignment, you are required to use R as the primary software environment. While you are free to choose any development environment that supports R, such as RStudio, Jupyter Notebooks via an R kernel, or Google Colab with R configuration, it is crucial that all analyses and visualizations are performed using R. Please ensure that your submission clearly indicates the development environment you have utilized. Additionally, if you sought assistance from ChatGPT or any other tool, include the specific prompts or queries used to obtain help. This transparency will aid in understanding the context and nature of the assistance received during your assignment.

Submission Requirements

1. Deadline Adherence: Submit your assignment to ee-class before **2024-04-09 23:00**. Note that late submissions will not be accepted under any circumstances. Ensure timely submission to avoid exclusion from grading.
2. Format and Content: Submit your work as a **html** file, generated from an R Markdown document that includes all R code, visualizations, and comprehensive explanations. Include a section detailing the development environment used and any external assistance received, specifying the exact prompts or queries.
3. Naming Convention: Name your file as **"1122_ CE5033_ID_HW1. html"**.
4. If your submission does not meet these requirements, **5** points will be deducted from your total score.

General Evaluation Criteria

1. Methodological Justification: Rationale behind methodological choices, including statistical tests, visualization techniques, and bootstrap methods.
2. Analytical Insight: Depth of insights and interpretations derived from statistical analyses and visualizations.
3. Clarity and Coherence: Organization of the submission, clarity of writing, and logical flow of content.
4. Technical Proficiency: Accuracy in coding, data manipulation, and implementation of statistical techniques in R.

Task 1. Iris Dataset Analysis

The iris dataset features 150 observations of iris flowers, encompassing measurements of sepal length, sepal width, petal length, and petal width across three species: Setosa, Versicolor, and Virginica. Now you should do finish the following tasks.

1. Visualize the distribution of sepal length across the iris species. Employ an appropriate plot type, explain its suitability, and what insights it provides about sepal length variation among species.
2. Estimate a 95% confidence interval for the mean sepal length of Iris setosa using bootstrap methods. Conduct a minimum of 1000 bootstrap resamples, visualize the distribution of bootstrap means, and report the confidence interval. Discuss what this interval reveals about Iris setosa's mean sepal length.
3. Compare the average petal length between Iris versicolor and Iris virginica. Use bootstrap techniques to estimate the mean difference, performing at least 2000 resamples for each species. Visualize the distribution of mean differences and calculate a 95% confidence interval. Analyze the significance of the differences in average petal length between these species.

Task 2. Analyzing Vitamin C's Impact on Tooth Growth

The ToothGrowth dataset includes 60 observations on tooth growth (length) in guinea pigs following three different doses of Vitamin C (VC), administered either via orange juice or as ascorbic acid (a form of Vitamin C). Now you should investigate if the average tooth growth induced by Vitamin C differs from a standard growth level of 18.5 units. Formulate your hypotheses, select an appropriate statistical test, conduct the analysis, and interpret the findings.

Task 3. Effectiveness of Insecticide Sprays

The InsectSprays dataset records the effectiveness of different insecticide sprays against agricultural pests, with a count of insects killed by each of six spray types. Now you should compare the effectiveness of spray A to spray B. Formulate hypotheses and select a statistical approach for comparison. Analyze the data, justify your method selection, and interpret the findings.

Task 4. Chick Weight Changes Over Time

The ChickWeight dataset tracks the weights of chicks over time, with measurements taken at various ages for chicks on different diets. Now you should analyze weight changes from Day 0 to Day 21 for chicks on diet 1. Define suitable hypotheses, identify an appropriate statistical test for this comparison, explain your choice, and discuss the results.