STAT 6337
Advanced Statistical Methods I (Fall 2024)
Project 1

This project is individual work. So do not consult with anybody in or out of class. You can ask me or TA questions if something is not clear.

Complete this page below and attach with your project. Your project will not be graded without it.

This project is entirely my work. I have not discussed about this project with anybody in or out of class. I understand and have complied with the academic integrity policies written in the *Handbook of Operating Procedures* of UT Dallas https://policy.utdallas.edu/utdsp5003**.**

YOUR NAME _____

DATE _____

YOUR SIGNATURE (NOT just typed name) _____

Stat 6337: Advanced Statistical Methods (Fall 2024)

Project# 1

Notes:

- You are supposed to work on this project entirely on your own. So, do not consult with anyone within or outside the class.
- You are welcome to ask me or TA questions. However, first try to find the answer on your own. Don't be afraid to google! It is a necessary skill for becoming a successful programmer.

1. The Framingham Heart Study is a multi-generational, longitudinal study to assess risk factors for cardiovascular disease (CVD). Details of the study, its design, and major accomplishments of this one of the longest running study in the US can be found at `https://www.framinghamheartstudy.org/`.

   This dataset is a subset of the data collected in the Framingham Heart Study. It includes n=4434 participants who completed one of the regularly scheduled examinations from 1956 – 1968. The following table shows variable names, as they appear in the Excel worksheet, along with brief descriptions and coding details for each variable.

Variable Name	Description	Coding
AGE	Age at exam, in years	32-70
TOTAL CHOL	Total cholesterol, mg/dL	107-696
SBP	Systolic blood pressure, mmHg	83.5-295
DBP	Diastolic blood pressure, mmHg	48-142.5
BMI	Body mass index, kg/meters2	15.54-56.8
CIGS PER DAY	Number of cigarettes smoked per day	0-70
GLUCOSE	Serum glucose mg/dL	40-394
HEART RATE	Heart rate, beats/minute	44-143
CVD	Cardiovascular disease over 24 year follow-up	0=no, 1=yes
HYPERTENSION	Hypertension over 24 year follow-up	0=no, 1=yes

   Note: The dataset is comma separated (.csv extension). To read it in SAS, use the options DSD and FIRSTOBS = 2 in the INFILE statement (e.g., INFILE FHS DSD FIRSTOBS = 2;)

   (a) Is hypertension associated with CVD? Carry out an appropriate test. Include the appropriate hypotheses, test statistic value, p-value, and conclusion.

   (b) Compute summary statistics for the variable CIGS PER DAY and make histogram and boxplot and comment on its distribution. Is there evidence that this variable is associated with CVD, i.e., its values are different in the two groups of CVD? Carry out two tests — one parametric and one non-parametric to investigate this. For each test, include the appropriate hypotheses, test statistic value, p-value, and conclusion.

   (c) Plot BMI vs GLUCOSE with different plotting symbols for the two CVD groups. Describe the relationship between BMI and GLUCOSE. Does CVD appears to have an effect on these variables?

(d) Is there evidence that the average SBP of this population is more than 125? Carry out an appropriate test. Include the appropriate hypotheses, test statistic value, p-value, and conclusion.

(e) Create a new 4-category variable using SBP by dividing SBP values into four parts using its quartiles. Test if this new categorical variable is associated with hypertension. Include the appropriate hypotheses, test statistic value, p-value, and conclusion.

(f) Check for normality of TOTAL CHOL using appropriate summaries, graphs, and tests (including goodness of fit test).

2. Consider the dogs data. In this study, 19 dogs were intitally given an anesthetic drug pentobarbitol. Each dog was then administered carbon dioxide (CO_2) at two pressure levels (high and low). Next, halothane (H) was added, and the administration of CO_2 was repeated at two pressure levels. The response, milliseconds between heartbeats (measuring anesthetic effect), was measured for each of four treatments:
Treatment 1: high CO_2 pressure without H
Treatment 2: low CO_2 pressure without H
Treatment 3: high CO_2 pressure with H
Treatment 4: low CO_2 pressure with H

   Is there a difference in mean response between any two treatments? Conduct appropriate parametric and non-parametric tests comparing the first two treatments and repeat the tests for other combinations of two treatments. For each test, use significance level 0.004. Which treatments differ from each other and how? Can you provide a justification for using this significance level?

Some useful links:

- `http://support.sas.com/documentation/cdl/en/procstat/63104/HTML/default/viewer.htm#procstat_univariate_sect016.htm`
- `https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_intronpar_sect006.htm`
- `http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_ttest_a0000000115.htm`
- `https://support.sas.com/documentation/cdl/en/lrcon/62955/HTML/default/viewer.htm#a000780367.htm`
- `http://support.sas.com/documentation/cdl/en/procstat/63104/HTML/default/viewer.htm#procstat_freq_sect027.htm`
- `http://www.ssc.wisc.edu/sscc/pubs/4-8.htm`

Submit your report with the following components (in this order):

- at most 3 pages of typed answers;
- relevant parts of SAS output with relevant numbers highlighted (label each part of the question);
- your SAS code including brief typed comments of main steps (each part labelled).