MAT 4017 Project Assignment 1 – 50 pts

Due: Midnight, Thursday, March 30th

1. Overview

This assignment is to conduct an observational study on a population of your choosing. For this assignment, you may work in teams of up to two people, with each team submitting only one report. If you wish to work in a group, you have two options. First, you may form a group on your own. Second, I will maintain a list of people who want to be assigned into a group, and I will assign groups.

For the project, you are to perform an observational study about a quantitative variable on a population of interest. For the project, you will first identify a large population that you are interested in and one quantitative variable about individuals in the population you are interested in. Once you decide the population you are interested and the variable, you will seek for a large dataset. One easy way to get a large data is to download a publicly available dataset for the population. Here are an example web site that provides public health-related datasets:

https://www.cdc.gov/nchs/data_access/ftp_dua.htm?url_redirect=ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NSFG/

When you obtain the dataset from the web, someone, or other resources, you should provide the source of the dataset and acknowledge the help.

Next, you will analyze the population data without having the details outlined for you. In order to organize a statistical problem, you need to plan a four-step process (state, plan, solve, and conclude) and I give you examples of this process in this document. Based on your plan of the process, we will solve the statistical problem (analyze the population) using statistical tools you have learned from many of resources including class discussions. The statistical tools need to be decided based on the variable you choose. I will provide rough guideline for these tools but you are free to choose any available tools that work best for your dataset.

You will then perform experiments by taking a simple random sample of size approximately 40 from the population and computing the sample mean and standard deviation. Finally, you will report your findings in a short report. Many aspects of the project are fairly technical (selecting a quantitative variable, calculating variance and mean, calculating five number summary, etc.), so I also provide a Guidelines document to help you through some of the trick aspects if you need it.

2. The Four-Step Process for a Statistical Problem

The four-step process for data analysis is defined as follows:

- *State*: What is the practical question, in the context of the real-world setting?
- *Plan*: What specific statistical operations does this problem call for?
- *Solve*: Make the graphs and carry out the calculations needed for this problem.
- *Conclude*: Give your practical conclusion in the setting of the real-world problem.

Example:

How well have stocks done over the past generation? The Wilshire 5000 index describes the average performance of all U.S. stocks. The average is weighted by the total market value of each company's stock, so think of the index as measuring the performance of the average investor. The percent results on the Wilshire 5000 index for the years from 1971 to 2013 can be found in a given data file. What can you say about the distribution of yearly returns on stock?

State: How have returns on stocks behaved over the years?

Plan: We should examine the distribution through graphs and numerical summaries. Because this is a variable that changes over time, you should also look at a time plot.

Solve:

a. Provide a stemplot and time plot (a graph illustrating a return value for each year in time).

b. Provide the mean and standard deviation of the dataset, and the five-number summary (minimum value, quartile Q1, median, quartile Q3, maximum value).

Conclude: Provide a conclusion from the data analysis

3. Statistical Study

First, select a population that you can reasonably collect a simple random sample form. Then, illustrate the population using data visualization tools (histogram, stemplot, graphs, scatter plot, boxplot, etc.), and perform the theoretical computations (the mean and standard deviation, and five-number summary).

Second, perform your observational study by selecting and surveying a simple random sample of size approximately n = 40 from your population. Next, perform the theoretical computations (the mean and standard deviation, and five-number summary).

Lastly, report your results in the report. Begin the final report with an introduction section restating the population of interest, variable of interest, and data analysis you conduct.

(1) Question Studied – 5 points

Give a one paragraph description of the question you are interested in. In this paragraph, state the population you are interested in, the variable you will be studying and give your belief about the average value of the variable. (Though this is presented first, you may actually want to work the technical aspects section first.)

(2) Technical Aspects – 20 points

This section should be 6 short paragraphs (one to two sentences each) describing the following.

- Formal description of the population of interest.
- Formal definition of the variable you will study.

- Formal statement of the four-step process for data analysis.
- Your plan for selecting a simple random sample from your population.
- Formal statement of the data analysis using mean and standard deviation of the sampled data.

(3) **The summary of the observatory study – 25 pts**

1) Data: (a) Describe in 1 or 2 sentences how the data was collected and why (or why not) you believe that it is reasonable to consider it as coming from a simple random sample from you population of interest. (b) Provide a table of the population (or a separate excel file). (c) Give a graphical representation of your data in the form of a histogram or stem and leaf plot (or graphs).

2) Provide the four-step process of the data analysis: (a) state, (b) plan, (c) solve and (d) conclusion. In (c), provide the mean and standard deviation of the population, and the four-number summary.

3) Discussion of the shape of your data: Answer the question of whether or not it seems to have come from a population with an approximately normal distribution. Describe any skewness or outliers present or state that the data is symmetrical with no outliers.

4) Calculate and report the sample mean and standard deviation. (You do not need to show me the calculations – just give me the output from the software you are using.).

5) Report in one or two sentences whether or not you reject the null hypothesis and what this means about your population.

4. Software tools for data analysis

- The team decides the software for data analysis.
- Microsoft Excel, IBM SPSS, Python, R, MATLAB can be used for the analysis as your choice. There is NO discrimination among these software.

5. Submission

- Each team submits only one report through the course portal by midnight, Thursday, March 30 (with both students' names)
- The file format of the report must be either .pdf, .doc, or .docx
- Grace period:
    - 1 week delay: 20% penalty
    - 2 week delay: 40% penalty
    - Later than 2 week delay: 100% penalty

Additional Guidelines: If desired, consult the "Project Guidelines" posted in the course portal for additional pointers on completing the project. As usual, don't hesitate to contact me with questions. You can also get some help and guidelines from the TA.