HW #2

BUAN/MKT 6337.001/002

Important Note1: For all questions below assume the test level is $\alpha = 0.05$

Important Note2: For each question which asks you to do a hypothesis test, you need to (1) Write down your null and alternative hypothesis clearly and mention which test is appropriate in your word report, (2) Write a code in SAS to execute the test and (3) give the results of the test in your word report.

1. The United States Geological Survey provides data on earthquakes of historical interest. The SAS data set called EARTHQUAKES contains data about earthquakes with a magnitude greater than 2.5 in the United States and its territories. The variables are year, month, day, state, and magnitude.

(a) California and Alaska are the two states with the highest number of earthquakes in the country. Create a new data set that includes only these two states and use this data set to answer the following questions.

(b) You are interested in the following statistics for the magnitude of earthquake:

- Mean
- Median
- Standard deviation
- Minimum and maximum
- 25^{th} and 75^{th} percentiles

Create a table that shows the above statistics across different states within each year. In particular, your table must have years at the first column and it must break down the results across different states in the second column. In order to make the table short, further assume you are interested only in recent years and want to create a table that shows the desired statistics from 2002 to 2011.

(c) Modify you SAS code in (b) such that the results for each year is shown in a separate table.

(d) Now, assume you want to show the same results in part (b) but with the difference that years are shown is the first column and the states are shown in the top row.

(e) You are interested in how the magnitude of earthquakes is trending over time for each state. In one graph, plot two time series plots, side by side, which shows the trend of average magnitude of earthquakes over time for the two states.

(f) Test the following null hypothesis: "the average magnitude of earthquakes in California is equal to that of Alaska".

---

2. Suppose that at a local university the study guidelines for the College of Science and Math are to study two to three hours per unit per week. The instructor of the class, Orientation to the Statistics Major, takes these guidelines very seriously. He asks students to record their study time each week, and at the end of the term he compares their average study time per week to their term GPA. The SAS data set called STUDY_GPA contains student identification information, orientation course-section number, number of units enrolled, average time studied, and term GPA.

(a) Graph the histogram for hours of study. Use the start point=0 and bandwidth=5. Also, overlaid to this graph, display the plots for the kernel density and the best fitting normal curve. Using an eyeballing approach, can we say the hours of study follows a normal distribution?

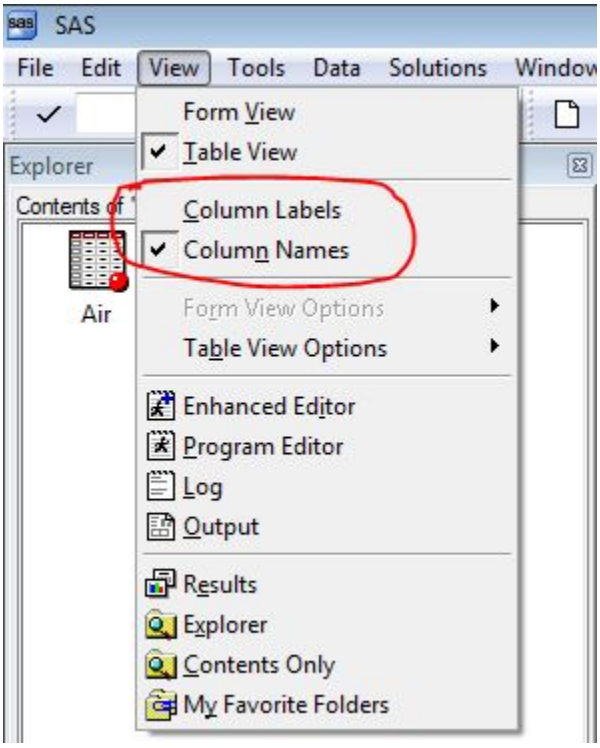
The next part focuses on the difference between causality and correlation. In most managerial situations, we are interested in causality rather than just correlation. Think carefully when you answer the next part. There may not be a definite correct answer when it comes to the causality argument. You will be given the credit as long as your reasoning is plausible.

(b) Conduct a hypothesis test to check whether there exists a significance correlation between units enrolled, hours of study and GPA for section 2. What is your conclusion? What variable you think may cause the other?

---

3. A study was conducted to see whether taking vitamin E daily would reduce the levels of atherosclerotic disease in a random sample of 500 individuals. Clinical measurements, including thickness of plaque of the carotid artery (taken via ultrasound), were recorded at baseline and at two subsequent visits in a SAS data set called VITE. Patients were divided into two strata according to their baseline plaque measurement.

(a) First, double-click on VITE file to check column names. If your column names are long (e.g., labels are shown instead of column names), then go the menu and change the setting to column names.



(b) Use proc content to see variable names as well as variable labels to better understand each variable. Note that the current data is in long format. We first want to transform the data to wide format so that we can conduct certain statistical analyses. See the following reference materials

to understand the difference between long and wide formats. Basically, long formats have repeated observations for a given person, whereas wide formats record those observations column-wise. Use proc transpose command to so that plague values for each visit (0, 1, 2) are recorded in 3 separate columns as opposed to 3 rows, by ID and treatment

a. Reference1:https://stats.idre.ucla.edu/sas/modules/how-to-reshape-data-long-to-wide-using-proc-transpose/
b. Reference2: https://data.library.virginia.edu/reshaping-data-from-wide-to-long/

(c) Assume there were no placebo group (i.e., treatment = 0) in your data set. Conduct a test to see whether there is a difference in plaque level before treatment and after the second visit? Interpret your results.

(d) Now, considering the fact that there is indeed a control group in your dataset, conduct a new test to check whether there is a difference in plaque level before treatment and after the second visit. Interpret your results.
(Hint: You need to make sure that the reduction in plaque level (if any) is indeed due to taking vitamin E. This is how having the control group in the study helps. In particular, you need to test whether the reduction in plaque level for treatment group is significantly less than that for control group)

(e) Which of the tests in part (c) and (d) is more reliable? Explain.
(Hint: This should be clear from the hint given above)

(f) One of the critical factors in randomizing the subjects in control and treatment groups is to make sure that the subject are perfectly randomized in all aspects. Using the last two columns (i.e., alcohol and cigarette usage) of the original (long format) data, conduct two tests to check whether subjects are randomized perfectly. If they are perfectly randomized, then we should not expect much difference in alcohol (or cigarette) consumption for control vs. treatment groups.
