**Assignment 3 – Descriptive Analysis and Hypothesis Testing**

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.   Examine this SAS data set including the variable labels and attributes. Create a scatter plot of year and magnitude for earthquakes that occurred in the year 2000 and beyond.

b.   Overlay a time series plot of the mean magnitude for each year on the same graph as part a). This line should appear in red on your graph.

c.   Include a legend for your graph that labels the time series plot as Mean. The legend should have no border and appear on the bottom right side of the plot.

d.   Earthquakes are classified by their magnitude. Overlay reference lines on your graph for light, moderate, strong, major, and great earthquakes defined at magnitudes of 4.0, 5.0, 6.0, 7.0, and 8.0, respectively. These lines should be labeled, dashed, and have 50% transparency.

e.  Make sure that all years appear on the X axis of your graph.

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.   Examine this SAS data set including the variable labels and attributes. Create box plots to compare the time studied between the two sections.

b.   Create a simple linear regression plot for time studied and GPA. Turn off the legend.

c.   Create a simple linear regression plot for time studied and GPA with a line for each section. Move the legend to the far right side of the plot.

d.   Add 95% confidence limits for the mean predicted values to your plot from part c). Adjust the transparency so that bands for both sections are visible on the plot.

e.   Add a comment to your program about any potential relationships that you see between the variables included on these three plots.

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.   Examine this SAS data set including the variable labels and attributes. Is there evidence to suggest that the true mean systolic blood pressure at baseline is significantly greater than 140 mm/Mg? Carry out an appropriate hypothesis test for each strata, and compare the resulting p-values to alpha = 0.05.

b.   You would like to test for differences in true mean plaque before treatment and at the second year visit. Examine the layout of the data set, and in a comment describe why this data set structure would not work for the hypothesis test that you need to use.

c.   Test for a difference in plaque as specified in part b) for the treatment group and also for the placebo group. Carry out appropriate hypothesis tests for each strata, and compare the resulting p-values to alpha = 0.05.

d.   To complete the analysis, compare the differences that you saw in mean plaque before and at the second year visit across the two treatment groups. Carry out appropriate hypothesis tests for each strata, and compare the resulting p-values to alpha = 0.05.

e.   Verify the assumption of normality for the tests in parts c) and d). Include plots as well as p-values.

f.    In a comment in your program, discuss the results from parts c) and d).

4. Researchers conducted a study of left handedness in a random sample of elementary school aged children and stored their results in a SAS data set called LEFTIES. They collected data on the hand preference for writing, cutting with scissors, and using a mouse. Data were also collected on the foot that the child used to kick a ball, as well as age and gender.

a.   Examine this SAS data set including the variable labels and attributes. Is there an association between writing hand and kicking foot preference? Carry out the appropriate hypothesis test and compare the resulting p-value to alpha = 0.05.

b.   Is there an association between writing hand and mousing hand preference? Carry out the appropriate hypothesis test and compare the resulting p-value to alpha = 0.05.

c.   Is there an association between writing hand and gender? Carry out the appropriate hypothesis test and compare the resulting p-value to alpha = 0.05.

d.   Create bar charts to accompany each of the tests in parts a) to c).

e.   In a comment in your program, discuss the results and any concerns that you might have about the expected cell count assumption for these tests.