STT 3820 Mid-Term Project Fall 2021 (Due Tuesday, Oct 26, by 4:00 p.m.)

This project asks you to analyze two sets of data: one deals with average student debt (described in **StudentDebt.doc**), and the other deals with marriage and divorce percentages. All data and description files are under the directory Mid Term Project. For each question, you should include requested graphs, all steps for appropriate hypothesis testing, any computer output/code, and complete sentences to describe your output and conclusions. Points for each question are in ( ). Calculations/graphs/etc. are to be done in R. You will have to find resources on the internet for graphics (e.g.: READ: http://www.statmethods.net/advgraphs/parameters.html BEFORE you start!). Grades will be awarded using Group Evaluations.

Average Student Debt:

MeanStudentLoanDebtByCategory.csv (READ: http://www.statmethods.net/advgraphs/parameters.html BEFORE you start!)

1) Create scatterplots with connect lines that show how mean student debt has changed from 1989 to 2010. Be sure to overlay all possible columns on a single graph:

A) with all highest level of education completed's debts as the y's. (2) Describe any obvious patterns or relationships. (2)
B) with all annual income categories' debts as the y's. (2) Describe any obvious patterns or relationships. (2)

2) Create comparative boxplots

A) with all highest level of education completed's debts as the y's. (2) Describe any obvious patterns or relationships. (2)
B) with all annual income categories' debts as the y's. (2) Describe any obvious patterns or relationships. (2)

3) Are the time series plots or the boxplots a more appropriate way to display the given data? (1)

4) Use a t-test to test whether the mean debt for college graduates is over $20,000 for the entire time period at a significance level of 5%. $(H_0: \mu = value)$ Be sure to show all FOUR steps of the hypothesis test. (And, yes…the time component IS a problem in the data, but we're going to use it anyway.) (11)

5) Repeat the test on the ***From1998*** variable. Does the conclusion change? (11) If so, how? (2)

6) Would it be more appropriate to use the difference of means test or a mean of differences test to compare debts in this data set? (1) Which met assumption is the most important in making that decision? (1)

Marriage and Divorce:

MarDiv2009.xls (Note that totals are in THOUSANDS!!!)

7) What races are represented by the data? (2)

9) How many people does the number in column E row 8 of the White sheet represent? (2)

A) In that column, how many men have never been married? (2)

B) Of the Ever married Men (column B), what percent of those have been married Once, Twice, and 3 or more times? (Be sure to use rules of conditional probability…should add up to 100%.) (6)

10) Testing differences in Race behavior for **Men**: [Hint: difference of proportions]

A) Is there evidence that there is a difference in the percentages of Black and White-NonHisp Ever Divorced **among the 35-39 year old <u>men</u>**? (11) If so, how much (95%CI)? (2)

B) Same question, Asian and Hispanic men. (13)

11) Statistics are often quoted that "more than 50% of marriages end in divorce." Using the ever divorced out of ever married as phat, is there evidence that the percent of people who have been married and then divorced is less than 50% for either gender of White Non-Hisp, Hispanic, Black, or Asian? (8 tests; show everything for White Non-Hisp Males (5), give test stats, p-values, and conclusions only for last 7 (14)…remember we are looking at the ever divorced out of ever married.) [Hint: one proportion tests multiple times…a function would be good.]

12) Using what you know of probability distributions, propose where that "50%" statistic could have originated and what a more reasonable statistic might be to express the relationship between marriage and divorce. (2)