STAT 511: Statistical Methods
Summer 2014 - Project 2 (60 pts. + 21 pts. BONUS)
Due Tuesday, July 29.

There are three parts to this project, Part I concerns confidence intervals and hypothesis testing. This includes the interpretation of confidence intervals, checking the assumptions in the methodology and determining which method is appropriate in a given circumstance. Part II (BONUS because of timing) is concerned with ANOVA. Part III concerns a bivariate data set concerning wine consumption versus death rates; how to describe it graphically and how to analyze it using linear regression (BONUS because of timing).

Note: All of the plots and figures in this project must be generated using computer software (statistical software, Excel or another graphing program). No hand drawn plots will be accepted. I would also strongly suggest that you use statistical software (or Excel) to do the requested calculations since I will be giving you raw data and not the summarized information. Therefore the required work is just the formulas used with the variables, then the numbers substituted in followed by the answer. If the formula includes a sum, you only need to include one set of numbers substituted in.

(30 pts.) Part I: Confidence Intervals and Hypothesis Testing

This part will explore the interpretation of confidence intervals and how they relate to hypothesis testing.

1) Generate 50 normal distributions with $\mu = 10$ and $\sigma = 2$ with a sample size of 40 each (http://www.random.org/gaussian-distributions/ or your software package). For each of these data sets, use statistical software to calculate the 90% confidence interval. Please provide the 50 confidence intervals and the number of intervals that contain the actual mean but no generated data. Is this number that you calculate the same as you would expect? Why or why not?

2) For the following problems, a) determine if you can perform a statistical analysis on it (are all of the assumptions met?), b) determine the appropriate methodology (one-sample mean: two-tailed, upper or lower bounds or one sample proportion,) and c) perform the methodology stated in part b) using both the appropriate confidence interval (with interpretation) and the appropriate hypothesis test (with all 7 steps). Note: if you cannot perform the statistical analysis, no further information is required. In each case, explain why the results are the same. The data set for this part is in project2.xlsx sheet CI.

i) The data in Sample 1 is the number of cycles to failure in seawater for beams subjected to certain bending and loading stress (in thousands). It is important that this number be greater than 300 for these beams to perform satisfactory in bridges that are subject to exposure to seawater. Does the data contradict this design specification at a 1% significance factor?

ii) the data in Sample 2 is the soil heat flux of eight plots covered with coal dust. The mean soil heat flux for plots covered only with grass is 34.0. You can assume that changing the covering does not affect the standard deviation and the population standard deviation is 3.3. The investigator wants to increase the soil heat flux and wants to know if the company should cover the plots with the more expensive coal dust. Should she change the methodology at a 5% significance factor?

iii) The data in Sample 3 is concerning the breaking strengths of cotton threads (g). According to the manufacturing process, this value should be 210 g. Does the data contradict this design specification to a 10% significance level?

iv) The data in sample 4 is the widths of a number of contact windows (μm) in certain CMOS circuit chips. The design specification of these widows in 3.5 μm. Does the data contradict the design specification to a 5% significance level.

v) BONUS (because of the timing, 7 pts.): Perform parts a), b) and c) above except that this is a two sample problem so you need to determine if the samples are paired or not. Is it harder to maintain your balance while you are concentrating? Eight elderly (both men and women) (Sample 5A) and eight young men (Sample 5B) were subjects in this experiment. Each subject stood barefoot on a "force platform" and was asked to maintain a stable upright position and to react as quickly as possible to an unpredictable noise by pressing a hand held button. The noise came randomly and the subject concentrated on reacting as quickly as possible. The platform automatically measured how much each subject swayed in millimeters in both the forward/backward and the side-to-side directions. At a 10% significance level, we are interested in if the side-by-side swaying is the same for both age groups.

Part II: ANOVA (BONUS because of the timing, 7 pts.)

A consumer organizations studied the effect of age of automobile owner on size of cash offer for a used car by utilizing 12 persons in each of three age groups (young, middle, elderly) who acted as the owner of a used car. A medium price, six-year-old car was selected for the experiment, and the “owners” solicited cash offers for this car from 36 dealers selected at random from the dealers in the region. Randomization was used in assigning the dealers to the “owners”. The offers (in hundred dollars) are included in the data file project2.xlsx sheet ANOVA.

1) Using a statistical software package, determine if age is important in determining the size of the cash offer solicited. Note that this is a hypothesis test so all 7 steps need to be included.

2) If age is important, then determine which age group(s) get the most money for the same car. Explain your reasoning. The work for this part includes the equations used for the comparison.

(30 pts.) Part III: Descriptive statistics and Linear Regression: Wine Consumption vs. Death Rates. Some people believe that drinking moderate amounts of wine is good for your heart.

Use any statistical software of your choice to analyze *Wine Consumption and Death Rate* data (project2.xlxs sheet WineConsumption) on the project webpage. Note that this is a bivariate data set, so each row is an (x,y) pair. Therefore, please do NOT sort the data in part e). If you choose, you may write this part as a research paper for a possibility of up to 3 bonus points. In addition to the information below, you will also need to write an introduction and include any references that you may use for the introduction and conclusion.

1) Generate two histograms; one of the wine consumption and the other of the death rate. Describe the shape of the distributions. How do the two histograms compare?

2) Create a summary statistics table including the minimum, maximum, the quartiles (first, second, and third), sample mean and sample standard deviation for both variables. You may use statistical software to do this but be sure to check to be sure that the standard deviation and quartiles are being calculated the way that we have defined them in class This part asks for the sample standard deviation and not the population standard deviation. The work required here are the equations with sample numbers.

3) Generate a comparative boxplot for wine consumption and death rate. Remember this has the outliers as separate points. Be sure to report the methodology that the statistical software uses to calculate the quartiles involved. Again, describe the shape of the distribution. How do the two plots compare? Do you see any outliers? If you do see any outliers, explain why they are outliers. The last explanation involves both calculations and words.

4) Write a summary of the data that you have presented so far. The key question is if there is a relationship between death rate and wine consumption. This should include any trends that you noticed from the visual presentation of the data including a comparison of the two shapes for each distribution from both the histogram and the boxplots. Be sure to include what the differences are between these two methodologies. This summary should be a short paragraph that includes at least three sentences.

5) Now, we are going to continue to determine if there is a relationship between the data via three methodologies, graphically, the covariance and correlation using the equations in chapter 5 and linear regression. Remember, to NOT sort the data because this is a bivariate set.

i) First, we are going to display this graphically. Create a scatter plot with the y-axis as the death rate and the x-axis as wine consumption. Note that the scatter plot is the visual representation only; no number should be submitted. From this information, do you think that there is a relationship between death rate and wine consumption (at least one sentence)?

ii) Calculate the covariance and correlation using the equations in Chapter 5. You can use a statistical software package to generate the values, however, be sure to double check to be sure the equations are what are presented in Chapter 5. Again, the equations with some numbers need to be submitted for the work. From this information, do you think that there is a relationship between death rate and wine consumption (at least one sentence)? Is the result here the same or different from part i)?

(BONUS because of timing, 4 points) iii) Using linear regression, perform a hypothesis test on the slope (β_1) (with all 7 steps). Is the result here the same or different from parts i) and ii)?

6) From the information that you have generated so far, do you believe that there is a relationship between deaths and wine consumption? Explain your answer including your results from d) and e). This should be a paragraph that includes at least four sentences.