Fall 2015 STAT 350 Project (150 points)

Due Friday April 15, 2016

Objectives: Statistical Inference

Instructions

- Groups of 2 – 4 students are required.
- Only PDF files are accepted.
- NO late work is accepted.
- Names of all students in the group with their section on top of the first page. Remember that all students have to have the same instructor.
- Statement of contribution for each student (submitted separately).
- Put all code in appendix; no code is required in the main body. Be sure to clearly label which code is for which part. However, the output that is necessary to answer the questions is required to be in the main body of the project. You will be graded on whether you have enough or too much output included.
- Your report should be in the same order as the questions posed. Clearly label each part.
- All discussion should be in complete English sentences.

Only one report should be submitted per group with each person submitting their own statement of contribution separately. Everything should be submitted on Blackboard. For the person who is submitting the report, you will need to add a separate attachment for the statement of contribution. The statement of contribution should consist of what each student did in the project and if there were any problems with the group as a whole. This statement should not be shared with your group mates, therefore, it can not be included in the body of the project.

If you have any question about the project, please ask on Piazza, ask during office hours, or discuss it with your instructor.

It is acceptable for different parts of the project to use different software packages. There will be no tutorials for this project, please refer to the appropriate tutorials for the individual parts.

Project: Statistical Inference

Throughout the semester we have learned some basic but useful statistics tools. With these tools, we can conduct analysis on some problems that we may be interested in. Since most data sets contain a large amount of different types of data, it is important to be able to determine which methods should be used to analyze it. In this project, you are to decide on a specific question based on the cleaned airline dataset from November, 2008 that we have been using this semester and use at least two DIFFERENT inferential methods to answer that question. You are not required to use the question that you posted in Lab 1. All of the analyses must be different from what is asked in the required labs. If you repeat anything that was previously asked, you will receive a 0 on that part. We will explain how to transform the data in Lab 6 and how to restrict which categories are used in Lab 7. For your information, the variables used in the labs are listed on Blackboard in the Project folder (under Assignments).

This list will be continually updated with newest addition be indicated in a different color. Please refer back to the page often.

Fall 2015 STAT 350 Project (150 points)

Due Friday April 15, 2016

The following are some examples of acceptable questions and the inferences that are used to check them using the dataset that was used last summer, the heights and weights of major league baseball players.

Question	Inference 1	Inference 2
How are the heights of the players related to position?	1-sample: heights	ANOVA: heights vs. position
Do different teams have the same heights of players?	1-sample: heights	ANOVA: heights vs. teams
Are the heights and weights of players related to their position (this one really has 3 inferences)	ANOVA: heights vs. position or ANOVA: weight vs. position	Linear Regression: heights vs. weights.
Do catchers weigh more than outfielders?	1-sample: weight	2-sample independent: weight of catchers and outfielders.

In the Airline dataset, there are more possibilities because there are more variables.

Grading and Content Information:

A. (5 pts.) Data: Make a table of each of the variables that you are using with a brief description of each variable and whether the variable is numeric or categorical.

B. (5 pts.) Decide on a question. Decide on a question that can be answered via inference. You need to use at least two different inference techniques to answer your chosen question. Please see above for help.

C. (50 pts.) Inference Method 1. See below for what needs to be included.

D. (50 pts.) Inference Method 2: See below for what needs to be included.

E. (10 pts.) Write a final conclusion based on Parts C and D. This should be a brief summary of what you have already written in the conclusions of parts C and D and a final answer to your question in part B.

Note that it is acceptable if the answer to your inference is 'not significant.' You will just need to explain in the part E how 'not significant' relates answers the question that you pose in Part B.

In addition to the points mentioned above, you will be graded on organization and style for an additional 30 points. These points will consist of whether the organization of the report is easy to read and the items are in the correct order, whether the student names are at the beginning of the report and if we receive the statement of contribution from each of the students. If not all members of the group participate equally in the project, this is where we will make those deductions.

Fall 2015 STAT 350 Project (150 points)

Due Friday April 15, 2016

Grading for Parts C and D:

a) (5 pts.) Code: The code should be clearly labeled in the appendix.
b) (5 pts.) What procedure should be used and why? Remember, this needs to be determined BEFORE you analyze the data.
c) (5 pts.) Graphically display the data as appropriate for your answer in part b) with an interpretation of the output. If a transformation is needed, this should be done before this step.
d) (10 pts.) Determine if the appropriate assumptions are correct. Please provide all graphs and explain your decision. If the assumptions are not correct for your methodology and you perform the analysis, you will lose 25 points. You may assume that the data set is SRS as you have been assuming for the rest of the semester.
e) (20 pts.) Perform the appropriate inference with a significance level of 0.05. This may consist of more than one step depending on the methodology in step b). The possible methodologies are

1) confidence interval AND hypothesis test (Chs. 8, 9, and 10): This includes both 1 sample, 2 – sample independent and paired. This methodology may be used more than once if you are using a different type of sample.
2) ANOVA (Ch. 11): Both the hypothesis test and the multiple comparison (if appropriate) need to be included.
3) Linear regression (Ch. 12): At least one inference needs to be included.

All confidence intervals, should include the interpretation. All hypothesis tests should consist of the 4 steps.

f) (5 pts.) A conclusion in words that relates to the context of the question. This should be a short paragraph in length and should be understandable to someone who has not taken a course in statistics which explains your conclusions of the part. This should answer part of the question that you posed in Part B.