Spring 2018 STAT 350 Project (230 points)

Due Friday April 20, 2018

Objectives: Statistical Inference

Instructions

- Groups of 2 – 4 students are required.
- NO late work is accepted.
- Names of all students in the group with their sections (times) are required on the top of the first page. Please include your Blackboard name especially if you have both an international name and an American name. Remember that all students have to have the same instructor.
- Each student must submit his or her own statement of contribution (submitted separately). See below for details.
- Put all code in an appendix; no code is required in the main body. Be sure to clearly label which code is for which part. The output that is necessary to answer the questions is required to be in the main body of the project report. You will be graded on whether you have enough or too much output included. Different people in the group can use different software packages.
- 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. The ***statement of contribution*** should consist of what you did in the project and if there were any problems with any individuals in the group. A chart with some guidelines is listed at the end of this assignment. Please include all of your group mates' names at the top of this page. Please rate each member of your group as poor, good, or exemplary. People will often be good in some areas below average in others. Therefore, if someone is mostly good, that is the rating that you should give them. Please provide further explanations if a person is rated either poor (or unacceptable) or if the person is exemplary. This statement should not be shared with your group mates; therefore, it cannot be included in the body of the project. For the person who is submitting the report, you will need to add a separate attachment for the statement of contribution. Please ask your instructor on where the statement of contribution is to be submitted.

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 lab tutorials as needed.

Project: Statistical Inference

Throughout the semester we have learned some basic but useful statistical 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 are appropriate for each type of data.

In this project, you are to create and answer three ***related*** questions based on the US Demographic data that we have been using this semester.

Spring 2018 STAT 350 Project (230 points)

Due Friday April 20, 2018

By “related,” we mean that you should seek to understand one general situation and ask three questions pertaining to that situation. Another way of looking at this is that you should analyze the features of one variable X and its relationship to the other variables in the dataset. Hence, all inferences would involve X alone or X with another variable. Some examples of this are given in the table below.

You will then answer each question by performing statistical inference (please see subpart e) of "Grading for C, D, and E" below for a list of the inference procedures you are allowed to consider). Furthermore, you must use at least two DIFFERENT techniques. For example, it would not be acceptable to use three one-sample t-tests for your inference procedures; however, it is acceptable to use a one-sample t-test and two different two-sample t-tests. You may also use three different techniques like a one-sample t-test, a two-sample t-test, and ANOVA.

You are not required to use the question that you posed in Lab 1. All of the analyses must be different from what are asked in the labs. If you repeat anything that was previously asked, you will receive a **zero** on that part. The variables that you can **NOT** use are listed here:

Lab	**Variables**
6	One-sample: Average Test Score
7	Two-sample: Median Income NE vs. NC, Education Spending in both periods
8	ANOVA: Average Test Score vs. Region
9	Linear Regression: Average Test Score vs. Median Income

Note that percent college graduates (“PercentCollegeGraduates”), percent divorced males (“PercentMaleDivorce”), and percent divorced females (“PercentFemaleDivorce”) are proportions and so the inference techniques that have been discussed in class are not applicable for them.

The following are some examples of acceptable situations and the inferences that can be used to check them using a dataset that was used previously: the heights and weights of major league baseball players.

Situation	Inference 1	Inference 2	Inference 3
Heights of baseball players	one-sample t procedure: heights	two-sample t procedure: average heights in American League versus National League	ANOVA: heights versus positions of interest (at least 3) OR two-sample t procedure: heights of two different positions
Weights of baseball players	one sample t procedure: weights	ANOVA: weights versus positions of interest (at least 3) OR two-sample t procedure: weights of two different positions	Regression: heights versus weights

Spring 2018 STAT 350 Project (230 points)

Due Friday April 20, 2018

Grading and Content Information:

A. (15 points) Introduction and question. Decide on three related questions that can be answered via inference in the US Demographic data set. Please see the above for help on creating your questions. Once you decide on the questions, briefly explain why the answers to these questions are important. This part should consist of at least one paragraph with at least one reference. References may be from online sources as long as they are correctly cited.

B. (5 points) Data. Make a table of the variables that you are using with the following information: the variable name, brief description of the variable, and the type of variable (numeric/categorical).

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

D. (50 points) Inference 2. See below for what needs to be included.

E. (50 points) Inference 3. See below for what needs to be included.

F. (20 points) Conclusion. Write a final conclusion based on Parts C, D, and E. This should be a brief summary of what you have already written in the conclusions of Parts C, D, and E plus a final conclusion that encompasses all of the questions from Part A. You will not receive full credit unless you also discuss the practical significance in context without relying on statistical inferential methods. Please write your response so that it is understandable to someone who has not taken statistics.

Although Parts C, D, and E may be done separately, please work on Parts A and F as a group.

It is acceptable if the result of any of the inferences is “not significant.” You will need to explain in Part F how the “not significant” conclusion answers the question(s) that you pose in Part A.

In addition to the points mentioned above, you will be graded on two additional aspects worth of **40 points**. **10 points** will be for **organization and style**. These points will depend on whether the organization of the report is easy to read, the items are in the correct order, complete English sentences are used, and whether student names and sections are at the beginning of the report.

The other **30 points** will be for **group participation** as graded by your peers. The points will be based on the submitted **statement of contributions** of the members of the group. If you do not submit the statement, then you will lose all of these points. The number of points may change depending on yours and your group mates' statements. Because of these points, the final score on the project might be different for different members of the group.

Grading for Parts C, D, and E:

a) (5 pts.) Code: The code should be clearly labeled in the appendix. You may use different software packages for the different Parts.

b) (5 pts.) What statistical procedure should be used and why? Besides the technique itself, be sure to state whether you are performing an inference procedure for a one-sided or two-sided hypothesis with an explanation for the choice. Remember, this needs to be determined BEFORE you analyze the data.

c) (10 pts.) Determine if the appropriate assumptions are satisfied. Please provide all of the diagnostic graphs to show that the assumptions are met and explain your decision*. If the assumptions are not correct for your methodology and you still perform the analysis, you will lose 25 points.* If a transformation is needed, state that you have performed a transformation and explain why. The explanation for the transformation should include at a minimum the histogram of the original data. You may include additional graphs of the untransformed variables if necessary for your explanation. You will then need to provide all of the diagnostic graphs for the transformed variables. You may assume that the data set is from an SRS as you have been assuming this semester. This assumption must be explicitly stated.

d) (5 pts.) Graphically display the data as appropriate for your answer in step b) with an interpretation of the output. The point of this part is to understand and explore your data, not merely to check the assumptions needed for inference as you did in step c). Some of the graphs in step d) may have already been used in step c)； however, the description of the graphs will be different. To determine which graphs are appropriate, please see the appropriate labs.

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 (Chapters 8, 9, and 10): This includes one-sample, two-sample independent and two-sample paired t procedures. Each of these procedures is regarded as a different type of inference.
2) ANOVA (Chapter 11): Both the hypothesis test and the multiple comparison (if appropriate) need to be included.
3) Linear regression (Chapter 12): At least one inference needs to be included besides the equation of the line. Please see Lab 9 for possible inferences.

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

f) (5 pts.) A conclusion in words that relates to the context of the question. This should be a short paragraph explaining your conclusions of the Part and should be understandable to someone who has not taken a course in statistics.

Spring 2018 STAT 350 Project (230 points)

Due Friday April 20, 2018

	Unacceptable	**Poor**	**Good**	**Exemplary**
Contribution to Group's Tasks	• Chooses not to participate in the group • Shows no concern for goals • Impedes goal setting process • Chooses not to participate in problem-solving	• Participates inconsistently in the group and sometimes helps in the group work • Shows sporadic concern for goals and sometimes helps in goal setting • Offers suggestions occasionally to solve problems	• Participates in the group all or most of the time • Shows concern for goals and participates in goal setting all or most of the time • Offers suggestions to solve problems and sometimes encourages group participation	• Always leads in group activities • Always leads in setting goals • Involves the whole group in problem-solving
Completion of Personal Tasks	• Impedes others from completing their assigned tasks • Does not complete their assigned tasks	• Sometimes helps others to complete their assigned tasks • Completes assigned tasks some of the time.	• Sometimes helps others to complete their assigned tasks • Completes all of their assigned tasks	• Actively helps others to complete their assigned tasks • Thoroughly completes their assigned tasks
Group Interaction	• Discourages sharing • Does not participate in group discussions • Does not listen to others	• Shares ideas occasionally when encouraged • Occasionally encourages other group members to share • Listens to others sometimes	• Shares ideas all or most of the time and sometimes encourages group members to share • Listens and takes other's feelings into consideration all or most of the time	• Shares ideas all or most of the time • Actively encourages all group members to share their ideas • Listens attentively to others • Empathetic to other people's feelings and ideas