Fall 2024 Stat 201 Project – Instructions

(+8 points) EARLY BONUS – Friday November 22nd by 11:59pm

DUE – Monday, November 25th by 11:59pm

OVERVIEW

You will pick one of the JMP data sets given on the assignment in Canvas to use for your project. You'll use that data set to complete the following visualizations, interpretations, and conclusions to create a data analysis report from start to finish. There is a data description file for each dataset to better understand the context of the data to assist in the interpretations of the results.

At the top of your Word document include:

- Your Name
- Your UTK Email (NetID@vols.utk.edu)
- Your Instructor's Name
- Your Section Number & Class Day/Time

For this assignment, you will use a random sample of 300 + the last 2 digits of your Student Id. For example, if your UTK Student ID ends in 23, then you would take a random sample of 323.

Note: When checking conditions, you are not to assume the population size to be the size of the original data set but rather the population of interest based on the context of the data set.

PROMPT

You have been working for **[Insert company of your choosing]** and have been tasked with analyzing **[Insert data you've chosen]**. You've been working for this company for 3 years and communicate more informally with the people you work with. The goal of your report is to analyze the data and communicate a statistical analysis in a way people who haven't taken this statistical class will understand. Use your imagination! You are allowed to construct a narrative and have fun with the report. You are not required to have fun but it can help. The goal is to practice using your skillsets in visualizing and communicating statistics (and have fun).

Remember, the Executive Summary and the Future Research will be constructed after all the related analysis has been completed and written up.

EXECUTIVE SUMMARY (15 Points)

The executive summary serves three key tasks. It orients the reader to the context of the research performed and how it ties into the overall mission of the company. Next, it summarizes key findings found throughout the report. Summarizing key findings is not just copying and pasting results. **Also, not all findings are key ones either.** Reporting non-key findings can clutter the summary making it difficult to discern was truly a key finding. We should try to take longer results of a key finding and condense them down but still give numeric backing for what was found. Finally, an action item should be given that can be implemented immediately. This decision should be based directly on the data we collected and a result of what was found. The goal of an executive summary is that someone can read your executive summary and then not read your report. It's a summary for executives who want all the information in a longer report delivered quickly and concisely with the decision of what to do as a result of the analysis.

Fall 2024 Stat 201 Project – Instructions

ONE SAMPLE Z (20 Points)

Pick a categorical variable with 2-levels (categories) you are interested in analyzing. Identify which level you will consider a "success" and why your company is interested in this level (category).

Create a bar chart and a frequency table of the variable you chose. Make sure your bar chart is in a horizontal display and bars have a value label (either count or percent) for each column. The frequency table should include the count and proportion (percent) for each including the totals. NOTE: Mac Users using "Dark Mode" may have problems adding the values to the visualization, search Stat 201 YouTube Channel or Discord Server for assistance. Include a screenshot of the visualization and provide a short descriptive write up of the variable making sure to cite the numbers in the graphic.

Next, create a confidence interval (Choose any % confidence other than 90%, 95%, or 99%) and include a separate screenshot of just the confidence interval.

Name and define the 3 conditions needed for a confidence interval for this variable type. Comment on whether your sample passes or fails each condition. Regardless of the outcome of your conditions, assume that all conditions have been met and continue to complete the analysis.

Interpret to the reader the confidence interval for only the chosen success level.

Next, use 0.25 as the hypothesized value (p_0) as the value your company believes to be the true proportion of what you have is considered a success for the population. Clearly explain to the reader the point of the test you are performing and why it is important to the company. Without using statistical notation, explain the null and alternative hypotheses of a "does not equals" hypothesis test.

Finally, using the confidence interval information, decide about whether to reject or fail to reject the null hypothesis and write the conclusion to the hypothesis test. Explain how the confidence interval assisted in the process of your decision about the null hypothesis.

ONE SAMPLE T (20 Points)

Pick a quantitative variable you are interested in analyzing.

Create a histogram making sure it has a count axis and is in a horizontal layout. Include a screenshot of the visualization including the summary statistics and quantiles sections.

Provide a short write up explaining the shape, and appropriate center and spread for the distribution. When interpreting values like mean, standard deviation, median or IQR, make sure interpretations are more than "The mean is ____" or "The median is ___". Interpretation should bring context to the data and not just cite values.

Next, create a confidence interval (Choose any % confidence other than 90%, 95%, or 99%) and include a separate screenshot of just the confidence interval.

Name and define the 3 conditions needed for a confidence interval for this variable type. Comment on whether your sample passes or fails each condition. Regardless of the outcome of your conditions, assume that all conditions have been met and continue to complete the analysis.

Interpret to the reader the confidence interval.

Fall 2024 Stat 201 Project – Instructions

SIDE-BY-SIDE BOXPLOTS (10 Points)

Using the quantitative variable and the categorical variable from the previous two sections, create side-by-side boxplots. Include a screenshot of the visualization including the 5-number summary.

Include a write-up describing the nature of the association of the two variables. Use the side-by-side boxplots to help you conclude about the relationship of the two variables. Provide supporting evidence for your decision using the numerical information from the output. That is, how do the similarities or differences among the levels of the categorical variable affect the quantitative variable help lead you to your conclusion.

REGRESSION (20 Points)

Select another quantitative variable that you may suspect has a linear relationship with the first quantitative variable you chose. Decide which variable will be the explanatory variable and which will be the response variable and explain your decision.

Create a scatterplot of the two variables. Identify the direction, form, and unusual features. Clearly note any outliers that may exist in the scatterplot. You will leave these data points in the analysis, but you must clearly identify any unusual features or outliers.

Perform a regression analysis on the two variables and provide a screenshot of the output. In addition, create the residual plot to evaluate the 4th condition for regression. Note: You only need to include the first residual plot (titled **Residual by Predicted Plot**).

Name and define the 4 conditions for regression as described from our book/notes and whether or not the condition passes or fails and what led you to that conclusion. Regardless of the outcome of your conditions, assume that all conditions have been met and continue complete the analysis.

In your write up include the interpretation of the slope, y-intercept, and RSquare values.

Also, evaluate if the regression model is statistically significant or not by using (and stating) the appropriate p-value.

FUTURE RESEARCH (15 Points)

Finally, give your ideas for future research. It's alright to mention briefly what the report covered, but don't use this section to mainly go through results you've already found. Your goal in this section is to talk about research that could be done in the future. Talk about new data to collect or ideas you think the company should investigate, the intent is that they will see your skills and innovative thinking as a benefit to the company and chose you to complete the research in the future. This section covers ideas that go beyond the current report.

FLOW AND NEATNESS (5 Points)

Flow of the report is gauged on the consistency of the results and the progression throughout the report in an effective communication style. Things like readability of the report and using complete sentences are keys to good flow. Read your report out loud, if it sounds choppy or disconnected that is a flag that the flow needs some help. Overall neatness is graded by scanning through the project and seeing how it looks. Most projects will obtain most, if not all of the 5 points. Projects that have graphics that are way too large or way too small will lose points. Projects with major issues like blank pages, poor screenshots, full page graphics, or other issues could lose all 5 points.

Note: Using AI to fabricate the initial report is not prohibited but should be reviewed. Output of AI programs often misses the context or flow of the concepts in a project such as this where you blend "regular conversation" with "statistical speak".