Overview

This is the first of 3 unit projects that you will complete this semester. All unit projects are group projects. Once established, you will work with the same team of 4 -5 people for the entire semester. Group members are expected to contribute equally to the data analysis and project write-up. You may set up any communications and/or meeting system needed to achieve group cooperation and input.

You will continue to work with the dataset that you chose for Unit 1 Project.

Unit Project 2 Instructions

For Unit Project 2, you will write a follow-up briefing memo to your supervisor, Dr. Random. Dr. Random has asked that you perform inferential analysis with the variables in your data set. Your will conduct a series of estimates and hypothesis tests. You will then conclude with a set of recommendations.

The format for your briefing memo is similar to Unit1. Instructions are included in Step B of these instructions. Your memo, including all written analysis, R-code, and output from your code will be submitted as **an R Markdown notebook file with HTML output. Your project team will submit both the R Markdown notebook file and the HTML file that is generated from the template. There will be only one submission per team.**

EACH PERSON IN YOUR GROUP IS RESPONSIBLE FOR RUNNING AND WRITING UP ONE OF THE DATA ANALYSES. The name of the responsible person should be listed with the analysis.

Here is a list of the analyses that will be included in your briefing memo. **EACH TEAM MEMBER IS INDIVIDUALLY RESPONSIBLE FOR ONE ANALYSIS.** You will decide among yourselves how to divide the work. Each of you will use R for calculations and graphs. You are welcome to use the same variables and same assignments that you used for Unit 1. Or you can change the assignments. *The individual responsible for the code and analysis must be clearly indicated.*

Format of Briefing Memo (Using the R Markdown notebook template)

1. Header (Group Responsibility): Create a header for your memo entitled "Briefing Memorandum", followed by "To:", "From:", and "Subject:" lines

2. Background/Recap (Group Responsibility): Provide a brief section (1 - 2 paragraphs, please) that 1) recaps the key features of your dataset, 2) states your task as a follow-up of your previous data exploration. You may produce a table here, if you wish, but it is not required.

3. Analysis (Individual Responsibility) *Each team member is individually responsible for the written analysis in the team report which corresponds to his/her designated statistical analysis, selected from the Table below (Analyses A1 – A5).*   The written analysis should include any summary numerical and graphical output for the analysis, neatly displayed and labelled. Be sure to respond to the questions given with each analysis in Table A. **Use full sentences.** Remember, you are explaining your results to Dr. Random. *As you discuss each analysis, be sure to interpret what you are finding in the context of your particular data situation.*

4. Recommendations (Group Responsibility): In this section, you will briefly summarize the conclusions of your estimates and tests. What statistically significant findings did you have? What limitations are there in your findings? Discuss if/how this is relevant in the context of your data – does it give us new insights ? Help us to better understand some phenomenon? Or not?

Individual Analyses (Part 3 of Briefing Memo)
Table A presents the list of the analyses that will be included in your briefing memo. EACH TEAM MEMBER IS INDIVIDUALLY RESPONSIBLE FOR ONE ANALYSIS. You will decide among yourselves who will do what.   Each person on the team will develop the listed R analysis and discussion/interpretation text for their assigned analysis. The R code and discussion/interpretation text will be incorporated into the SINGLE R Markdown file for the entire team , properly annotated and commented where necessary. *The individual responsible for the code and analysis must be clearly indicated on the template.*

Table A   ***Please conduct all analyses using classical methods, even if the conditions are not met. Discuss the limitations and validity of your analysis in your write-up. We will learn more appropriate methods in Unit 3.***

	Analysis Using R	**Discussion/Interpretation**
A1. Inference with One Quantitative Variable	a. **Confidence Interval:** Choose a quantitative variable and construct a 95% confidence interval for the parameter. **Be sure to eliminate all missing or NA values before doing your analysis. DO NOT REPLACE THEM WITH ZEROS.** • Show whether the variable meets the CLT conditions. • Display both appropriate numerical summaries and graphs to justify your statement. b. **Hypothesis Test:** For the same quantitative variable, perform a hypothesis test for the parameter *at a significance level consistent with your confidence interval*. • Determine whether your data meets the conditions to perform a test using the normal distribution.	• Clearly identify the parameter and give a correct interpretation for the confidence interval. • Discuss the limitations of your CI analysis. • For your HT, y**ou will need to find a baseline for comparison from an external source.** (For example, if your parameter is the average height of student, you will need to find an external source that gives an average height for students that you can use.) **Cite the source in your analysis, with a link or reference citation. IN WORDS, STATE A CREDIBLE null and alternative hypothesis**. It has to make sense for your data. • In your written analysis, indicate whether you are using a one-tail or two tail test (and why). Cite your test statistic, p-value,

	• Display both appropriate numerical summaries and graphs to justify your statement, as appropriate.	and state whether or not you reject the null hypothesis. • Then give a conclusion in the context of the problem. • Do the results of the confidence interval corroborate the conclusions of the hypothesis test? Explain why/how or why not. • Discuss any limitations to your analysis due to sampling bias, outliers in the data, or other issues that you notice.
A2. Inference with One Categorical Variable	a. **Confidence Interval:** Choose a categorical variable and construct a 90% confidence interval for the parameter. **Be sure to eliminate all missing or NA values before doing your analysis. DO NOT REPLACE THEM WITH ZEROS.** • Show whether the variable meets the CLT conditions. • Display both appropriate numerical summaries and graphs to justify your statement. **b. Hypothesis Test:** For the same categorical variable, perform a hypothesis test for the parameter *at a significance level consistent with your confidence interval*.	• Clearly identify the parameter and give a correct interpretation for the confidence interval. • Discuss the limitations of your CI analysis. • **For HT, you will need to find a baseline for comparison from an external source.** (For example, if your parameter is the proportion of smokers at the university, you will need to find an external source that gives an estimate of the proportion of smokers that you can use.) **Cite the source in your analysis, with a link or reference citation**. **IN WORDS, STATE A CREDIBLE null and alternative**

	• Determine whether your data meets the conditions to perform a test using the normal distribution. • Display both appropriate numerical summaries and graphs to justify your statement, as appropriate.	**hypothesis**. It has to make sense for your data. • In your written analysis, indicate whether you are using a one-tail or two tail test (and why). Cite your test statistic, p-value, and state whether or not you reject the null hypothesis. • Then give a conclusion in the context of the problem. • Do the results of the confidence interval corroborate the conclusions of the hypothesis test? Explain why/how or why not. • Discuss any limitations to your analysis due to sampling bias, outliers in the data, or other issues that you notice.
A3. Inference with Two Categorical Variables	a. **Confidence Interval:** Choose two categorical variables and construct a 99% confidence interval for the **difference in proportion**s. **Be sure to eliminate all missing or NA values before doing your analysis. DO NOT REPLACE THEM WITH ZEROS.** • You may need to create a new variable that reclassifies the original so that it only has	• Explain which variable is the explanatory variable and which is the response. ***Your choice must make logical sense in the context of the dataset.*** • Clearly identify the parameter and give a correct interpretation for the confidence interval. • Discuss the limitations of your CI analysis.

	two categories. If so, be sure to explain your reclassification(s). • Show whether the response variable meets the CLT conditions. • Display both appropriate numerical summaries and graphs to justify your statement. • Clearly identify the parameter and give a correct interpretation for the confidence interval. **b. Hypothesis Test:** For two categorical variable, perform a hypothesis test for the difference in the proportions ***at a significance level consistent with your confidence interval***. • Determine whether your data meets the conditions to perform a test using the normal distribution. • Display both appropriate numerical summaries and graphs to justify your statement, as appropriate.	• <mark>**IN WORDS, STATE A CREDIBLE null and alternative hypothesis**</mark>. It has to make sense for your data. • In your written analysis, indicate whether you are using a one-tail or two tail test (and why). Cite your test statistic, p-value, and state whether or not you reject the null hypothesis. • Then give a conclusion in the context of the problem. • Do the results of the confidence interval corroborate the conclusions of the hypothesis test? Explain why/how or why not. • Discuss any limitations to your analysis due to sampling bias, outliers in the data, or other issues that you notice.
A4. **Inference with One Categorical Variable and a Quantitative Variable**	a. **Confidence Interval:** Choose one categorical variable and one quantitative variable and construct a 95% confidence interval for the **difference in means**. <mark>**Be sure to eliminate all missing or NA values before**</mark>	• Identify the samples/groups in your analysis and explain why/how they are independent (not paired). • Explain which variable is the explanatory variable and which is the response. ***Your***

==doing your analysis. DO NOT REPLACE THEM WITH ZEROS.==

- You may need to create a new variable that reclassifies the original so that it only has two categories. If so, be sure to explain your reclassification.
- Show whether the response variable meets the CLT conditions.
- Display both appropriate numerical summaries and graphs to justify your statement.

b. Hypothesis Test: For the same categorical variable and quantitative variable, perform a hypothesis test for the difference in the means *at a significance level consistent with your confidence interval*.
- Determine whether your data meets the conditions to perform a test using the normal distribution.
- Display both appropriate numerical summaries and graphs to justify your statement, as appropriate.

choice must make logical sense in the context of the dataset.
- Clearly identify the parameter and give a correct interpretation for the confidence interval.
- Discuss the limitations of your CI analysis.
- ==IN WORDS, STATE A CREDIBLE null and alternative hypothesis==. It has to make sense for your data.
- In your written analysis, indicate whether you are using a one-tail or two tail test (and why). Cite your test statistic, p-value, and state whether or not you reject the null hypothesis.
- Then give a conclusion in the context of the problem.
- Do the results of the confidence interval corroborate the conclusions of the hypothesis test? Explain why/how or why not.
- Discuss any limitations to your analysis due to sampling bias, outliers in the data, or other issues that you notice.

| A5. **Inference with Two Paired Means** | a. **Confidence Interval:** Choose two quantitative variables that logically can be paired (e.g. before/after, different geographic regions, ratings from different sources, etc.) and construct a 95% confidence interval for the **difference in paired means**. ==**Be sure to eliminate all missing or NA values before doing your analysis. DO NOT REPLACE THEM WITH ZEROS.**==

 • Show whether the response variable meets the CLT conditions.
 • Display both appropriate numerical summaries and graphs to justify your statement.

 b. Hypothesis Test: For the same two quantitative variables, perform a hypothesis test for the difference in the means *at a significance level consistent with your confidence interval*.
 • Determine whether your data meets the conditions to perform a test using the normal distribution.
 • Display both appropriate numerical summaries and graphs to justify your statement, as appropriate. | • Identify the samples/groups in your analysis and explain why they are paired (not independent.)
 • Explain which variable is the explanatory variable and which is the response. *Your choice must make logical sense in the context of the dataset.*
 • Clearly identify the parameter and give a correct interpretation for the confidence interval.
 • Discuss the limitations of your CI analysis.
 • ==**IN WORDS, STATE A CREDIBLE null and alternative hypothesis**==. It has to make sense for your data.
 • In your written analysis, indicate whether you are using a one-tail or two tail test (and why). Cite your test statistic, p-value, and state whether or not you reject the null hypothesis.
 • Then give a conclusion in the context of the problem.
 • Do the results of the confidence interval corroborate the conclusions of the hypothesis test? Explain why/how or why not. |

		• Discuss any limitations to your analysis due to sampling bias, outliers in the data, or other issues that you notice.