BMI 531 631 Final Project Fall 2020

**Final Project**

**BMI 531 631 Fall 2020**

**Overview**

Replicability, also referred to as repeatability or reproducibility, is key to scientific progress. While true replication would involve new data collection to determine whether the same effect or phenomena is observed in a new sample or population, another element of replication involves simply being able to take the original dataset and reproduce the reported results, following the reported methods used for data analyses. For your final project, you will *reproduce* and *extend* analyses from a published research article. There are several goals for this analysis replication project:

1. Give you hands-on practice manipulating and analyzing real (un-tidy) datasets;
2. Give you the chance to learn about new statistical methods and how to actually use them;
3. Give you solid experience in creating, documenting, and evaluating a reproducible analysis report;
4. Give you practice doing collaborative research, something you will do throughout your career;
5. Show you, in a way that no homework assignment can, the complexity and joy of doing real data analysis.

**Grading**

The final project will be turned in and graded in 5 phases, worth a total of 100 points toward your final grade. Everyone on your team will get the same grade.

| **Activity** | **Total points** |
| --- | --- |
| P0: Team and paper assigned | 0 |
| P1: Data quality review | 15 |
| P2: Data delivery | 15 |
| P3: EDA report | 20 |
| P4: Replication/extension report | 25 |
| P5: Presentation | 25 |

**P0: Pick a Group and a Paper**

This year we will divide the class into three groups by random assignment using the breakout rooms in Webex, unless someone has a preference to be together in a team. We will use one paper for all of the groups, so please do not share your work with other groups. The paper we will use this year can be found at the following link:

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0017820>

This paper compares outcomes from two different differential gene expression techniques (RNA Seq and microarrays). This paper was published out of our department in 2011. The data can be downloaded from this website and there is a data dictionary. Please study this paper first. If you are taking the Algorithms class in DMICE you will be learning about RNA Seq, as described in the paper.

**P1: Data Quality Review**

The goal of this project piece is to look at and explore your dataset for your final project. You will use the [Quartz Bad Data Guide](https://github.com/Quartz/bad-data-guide/) to help guide your data quality review. Please use this [.csv](file:///Users/chambest/Documents/OHSU/Class%20Instruction/BMi%20531%20631/Fall2020/Class%20PPTs/Final%20Project/project/P1-data-quality-review/quartz-bad-data-checklist.csv) file to indicate which, if any variable(s), in your dataset has any of the issues listed. If you indicate a variable with an issue, please include a brief paragraph (2-3 sentences tops) describing the issue further.

We will review this article and the analyses on November 5 in class after you have had time to read it.

**The report and .csv file will be due on November 6, 2020, one submission for each group.**

**P2: Data Delivery**

Please upload 4 things, as detailed in the [Ellis & Leek paper](https://peerj.com/preprints/3139.pdf):

* The raw data.
* A tidy data set (Wickham 2014).
* A code book describing each variable and its values in the tidy data set. [do this for all variables that will be part of your replication analyses]
* An explicit and exact recipe you used to go from 1 -> 2,3. [this must be an R script or R Markdown file]

**These four deliverables will be due November 13, 2020, one submission for each group.**

**P3: Exploratory data analysis (EDA) report**

This report should include:

* A discussion of issues uncovered in your data quality review. How did you resolve them? Or were you not able to? Could they impact your ability to replicate any downstream analyses?
* Descriptive statistics and/or plots presented in the paper that you know you can reproduce from the dataset available (these may be presented in the “methods” section of the article text or tables and can include sample, etc.)
* Any additional statistics or plots that should be part of a good EDA as covered in class
* A thoughtful discussion of any issues you identified in your EDA that could impact your ability to replicate any downstream analyses (i.e., Missing data? Or numbers of subjects don’t match? Basic statistics like means/sds are off? Key variables are missing? Unexpected values present in a variable that you can’t interpret?)
* During class **November 17** we will meet with each of the three groups individually to discuss your analysis plan, in breakout rooms.

**This report will be due November 30, 2020, one submission for each group.**

**P4: Replication/extension report**

Each team has already discussed the scope of your replication and extension in a one-on-one meeting with the Instructor.

You should attempt to replicate every result, table, and figure in the original paper that is relevant to that analysis, as well as any analyses reported in text.

Do not spend too much time on layout or formatting. So, if your figure has legends in a different place or uses different colors in your plot or if a table has a certain formatting, don’t worry about replicating every last detail- what is important is that you attempt to replicate the content.

Your extension involves going beyond the original published article, using the same data. Your extension must be well-reasoned, with a sound research question that is clearly stated, and must include at least 2 of 3 of the following possible *extension* pieces:

* A plot
* A data summary table
* An additional analysis

Your final report should be a document that summarizes your replication/extension project with code in R Markdown. This document will be highly structured. You will show all of your R code in chunks (you can do this easily by doing nothing at all! This keeps the default knitr global chunk setting of echo = TRUE for all chunks). Scripts that were used to import, clean, and tidy your data should be referenced in your R Markdown document using source() in a chunk, as in:

source("01-import.R") # part of your "data delivery"

source("02-clean.R") # part of your "data delivery"

source("03-tidy.R") # part of your "data delivery"

# etc as needed

We should be able to knit your R Markdown file with no errors after you upload a zip file for your project directory folder.

Your final report should have three sections:

1. Your exploratory data analysis (EDA) report
2. Your *replication* report
3. Your *extension* report

You will submit a zip file that contains:

* A knitted HTML(or .pdf) of your final replication/extension report
  + All R code chunks should be included and visible in the final report
* The .Rmd file used to generate that HTML(or pdf)
* All data(s) needed
* All scripts needed that are called in your .Rmd file

**This report will be due December 17, 2020, one submission for each group.**

**P5: Presentation**

At the end of the quarter, you and your partner(s) will present your project to the class. The goal of your presentation is to teach the class about your replication project, explain what you have learned, and reflect on the process. Each team will do a 20-minute presentation; here’s a rough outline (but you can tailor it to your specific project):

* Summary of the paper you replicated (including simple research questions, definition of key constructs, and brief overview of measurement methods) (3 minutes)
* Description of issues identified in your data quality review (3 minutes)
* EDA highlights, focusing on *why* you did each plot/data summary table, and what you learned from them (3 minutes)
* Description of your replication and any problems you had (5 minutes)
* Description of your extension project and insights gained (5 minutes)
* Three biggest takeaways from doing this project (about 1 minute)
* Leave some time for questions

You are welcome to use slides.

**The presentations will be given on December 17, 2020 the last day of class. The presentation PPT or PDF will be due to us by December 16, 2020. We will have an open class on December 15, 2020 to answer questions or practice presenting in the Webex environment.**