Example Manuscript Template for a Data Analysis Project

Xueyan HU

2024-02-02

The structure below is one possible setup for a data analysis project (including the course project). For a manuscript, adjust as needed. You don’t need to have exactly these sections, but the content covering those sections should be addressed.

This uses MS Word as output format. [See here](https://quarto.org/docs/output-formats/ms-word.html) for more information. You can switch to other formats, like html or pdf. See [the Quarto documentation](https://quarto.org/) for other formats.

# 1. Summary/Abstract

In this project, I will be concentrating on analyzing my own data gotten from the experiment that I carried on by my own. And I will be trying to figure out if there will be any relation between interested variables and observations.

# 2. Introduction

## 2.1 General Background Information

The core idea of this research comes from the following facts. First, sprouts are usually considered nutritious but they are consumed in raw mainly for foods like salad and sandwiches. Therefore, there have been a lot of foodborne pathogen outbreaks occurred across the states in the decades (Miyahira & Antunes, 2021), which means that there is an urgent need to reduce the contamination of sprouts with foodborne pathogens.

At the meantime, a recently study has revealed that a newly derived chemical could enhence the immune resistance of some certain plants to the pathogens including bacteria and virus but not by killing them directly(Manosalva et al., 2015).

Based on the facts above, a hypothesis is raised that if this chemical can also be applied on seedling seeds to control the level of contamination during sprouting process.

## 2.2 Description of data and data source

The data is obtained from experiment that I did it by myself, which has 160 observations and several variables. The data is the result of bacterial populations of foodborne pathogens on sprouts that are collected from different types of plants at different germinating time points under different treatments for the seeds

## 2.3 Questions/Hypotheses to be addressed

I would like to figure out the effect of different treatments on controlling bacterial populations as well as the factors also may be related to the effectiveness, including seed type for sprouts, bacterial strain type infecting sprouts, etc. So the treatment type will be the most important predictor that I will focus on. And I am willing to explore if there is any treatment that could reduce bacterial population level on seedling sprouts.

To cite other work (important everywhere, but likely happens first in introduction), make sure your references are in the bibtex file specified in the YAML header above (here dataanalysis\_template\_references.bib) and have the right bibtex key. Then you can include like this:

Examples of reproducible research projects can for instance be found in (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, Shen, & Handel, 2020).

# 3. Methods

*Describe your methods. That should describe the data, the cleaning processes, and the analysis approaches. You might want to provide a shorter description here and all the details in the supplement.*

## 3.1 Schematic of workflow

## 3.2 Data aquisition

*As applicable, explain where and how you got the data. If you directly import the data from an online source, you can combine this section with the next.*

## 3.3 Data import and cleaning

*Write code that reads in the file and cleans it so it’s ready for analysis. Since this will be fairly long code for most datasets, it might be a good idea to have it in one or several R scripts. If that is the case, explain here briefly what kind of cleaning/processing you do, and provide more details and well documented code somewhere (e.g. as supplement in a paper). All materials, including files that contain code, should be commented well so everyone can follow along.*

## 3.4 Statistical analysis

*Explain anything related to your statistical analyses.*

# 4. Results

## 4.1 Exploratory/Descriptive analysis

*Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.*

## 4.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

These papers (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, et al., 2020) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# 6. References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>

Manosalva, P., Manohar, M., Reuss, S. H. von, Chen, S., Koch, A., Kaplan, F., … Klessig, D. F. (2015). Conserved nematode signalling molecules elicit plant defenses and pathogen resistance. *Nature Communications*, *6*(1), 7795. <https://doi.org/10.1038/ncomms8795>

McKay, B., Ebell, M., Billings, W. Z., Dale, A. P., Shen, Y., & Handel, A. (2020). Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery. *Open Forum Infectious Diseases*, *7*(11), ofaa494. <https://doi.org/10.1093/ofid/ofaa494>

McKay, B., Ebell, M., Dale, A. P., Shen, Y., & Handel, A. (2020). Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients. *Proceedings. Biological Sciences*, *287*(1927), 20200496. <https://doi.org/10.1098/rspb.2020.0496>

Miyahira, R. F., & Antunes, A. E. C. (2021). Bacteriological safety of sprouts: A brief review. *International Journal of Food Microbiology*, *352*, 109266. <https://doi.org/10.1016/j.ijfoodmicro.2021.109266>