Example Manuscript Template for a Data Analysis Project

Andreas Handel

2/9/23

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

Globally, *Salmonella* is a significant public health threat. We evaluated the persistence of *Salmonella* on various surface materials (e.g., picnic tables and bird feeders). Our studies suggest that *Salmonella* can have prolonged persistence on shared spaces. However, persistence depends on surface material and precautionary measures should be adopted to reduce exposure.

# 2. Introduction (Why *Salmonella*?)

*Salmonella enterica* infections are a significant public health threat, responsible for over 93 million annual cases of human illness worldwide. In the United States alone, 1.35 million cases of salmonellosis and 420 deaths are reported annually. Most cases of human salmonellosis are caused by food-borne *Salmonella* strains associated with produce or undercooked meat. However, a rising subset of human infections are often associated with unidentified environmental exposures or contact with animals.

The American white ibis (Eudocimus albus), forms large nesting colonies in natural wetlands. However, due to habitat destruction in Everglades National Park, this gregarious member of the Pelecaniformes, has begun to urbanize parks with abundant anthropgenic food and water resources throughout South Florida. Infection with *Salmonella* spp. in American white ibis is well documented. It has been found that ibis can harbor diverse *Salmonella* strains and shed it at a higher prevalence than their natural counterparts. Studies have genetically matched the *Salmonella* shed by ibis to human salmonellosis cases in South Florida during the same time period ibis were sampled.

Similarly, in 2021, following an epidemiological investigation of 29 human cases of *Salmonella enterica* serovar Typhimurium (*S.* Typhimurium), the Centers for Disease Control and Prevention (CDC) documented a relationship between human cases and a concurrent avian salmonellosis outbreak. Of the 29 cases, 14 individuals were hospitalized and the outbreak was likely due to contact with bird feeders, sick or dead wild birds, or pets that had contact with wild birds.

Persistence of *Salmonella* can be affected by ambient temperature, presence/absence of a biofilm, and environmental nutrient conditions. Under ideal conditions (e.g. optimal temperature 35-43°C, pH 7-7.5, etc.), *Salmonella* is known to persist in the environment for extended periods of time. While *Salmonella* persistence has been examined on soil and poultry litter, there is a lack of understanding of the persistence of *Salmonella* on various surfaces including wood, plastic, and surfaces lined with antimicrobial coatings.

## 2.1 Description of data and data source

The data selected for my MADA project is persistence and prevalence data from two trial studies. The first trial was conducted in South Florida where picnic tables, under the same pavilion and exposed to conditions known to affect *Salmonella*, were selected, swabbed daily for a week and shipped back to the University of Georgia for *Salmonella* isolation. The second trial was conducted in Athens, Georgia, at Whitehall Forest in a modified shade house. Bird feeders of different materials (e.g., cedar, plastic, plastic coated, and wood coated) were hung in triplicate and seeded with an isolate of *S.* Typhimurium isolated from a songbird.

The data from the first trial in South Florida was collected by an undergraduate student of Dr. Sonia Hernandez and provided to me by Dr. Hernandez. All data from the Feeder Trial was collected by me, Kimberly Perez under the guidance of Dr. Sonia Hernandez.

The Picnic Table Trial data is composed of several worksheets embedded in one excel file. After viewing the file (and running it below), there seems to be 192 observations. I am not sure how the Picnic Table Trial was collected but hope to visit with Dr. Hernandez regarding data collection methods soon. Feeder Trial data was collected and noted upon confirmation of *Salmonella* from processed samples. There are 144 observations for the Feeder Trial which spanned nearly 2 months. For both trials, we were attempting to measure *Salmonella* persistence. The Picnic Table Trial collected data in an attempt to measure prevalence as well. (Please Note: Aforementioned data was imported under “Data import and cleaning” section.)

## 2.2 Questions/Hypotheses to be addressed

Given the uncertainty of *Salmonella’s* persistence on various surface materials, from my data I would like to answer: - What is the prevalence of *Salmonella* over time? - How long does *Salmonella* persist on different surface materials (e.g., wood, plastic, and those with an antimicrobial coating) exposed to various climatic factors (e.g., rain, sun, changing temperatures)

From these questions, I hope to better inform the public and public health practitioners on safer feeder handling methods and precautions to take when frequenting public locations where humans and wildlife (e.g., ibis) interact.

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

Coming soon… *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 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.*

The data from the first trial in South Florida was collected by an undergraduate student of Dr. Sonia Hernandez and provided to me by Dr. Hernandez. All data from the Feeder Trial was collected by me, Kimberly Perez under the guidance of Dr. Sonia Hernandez.

# 4. Data import and cleaning

## 4.1 Picnic Table Data

library(readxl)  
PTD<- read\_excel(here("data", "raw\_data", "Picnic\_Table\_Data.xlsx"))

New names:  
• `` -> `...13`

View(PTD)  
summary(PTD)

Sample ID Day Date Table   
 Length:192 Min. :1.00 Min. :2020-01-02 00:00:00 Min. :1   
 Class :character 1st Qu.:2.75 1st Qu.:2020-01-03 18:00:00 1st Qu.:1   
 Mode :character Median :4.50 Median :2020-01-05 12:00:00 Median :2   
 Mean :4.50 Mean :2020-01-05 12:00:00 Mean :2   
 3rd Qu.:6.25 3rd Qu.:2020-01-07 06:00:00 3rd Qu.:3   
 Max. :8.00 Max. :2020-01-09 00:00:00 Max. :3   
 Quadrant Sample Type (Pooled or Persistence)  
 Min. :1.00 Length:192   
 1st Qu.:1.75 Class :character   
 Median :2.50 Mode :character   
 Mean :2.50   
 3rd Qu.:3.25   
 Max. :4.00   
 Shippng Date Notes STAB TUBE NUMBER   
 Min. :2020-01-06 00:00:00 Length:192 Length:192   
 1st Qu.:2020-01-06 00:00:00 Class :character Class :character   
 Median :2020-01-07 00:00:00 Mode :character Mode :character   
 Mean :2020-01-07 06:00:00   
 3rd Qu.:2020-01-08 06:00:00   
 Max. :2020-01-09 00:00:00   
 ORIGIN OF FECES Original PP still intact? Pos orNeg ...13   
 Length:192 Length:192 Mode:logical Length:192   
 Class :character Class :character NA's:192 Class :character   
 Mode :character Mode :character Mode :character

## 4.2 Feeder Data

library(readr)  
FD<- read\_csv(here("data", "raw\_data", "Feeder Data.csv"))

New names:  
Rows: 144 Columns: 6  
── Column specification  
──────────────────────────────────────────────────────── Delimiter: "," chr  
(2): Collection\_Date, Feeder\_Type dbl (2): Feeder\_number, Absence\_0\_Presence\_1  
lgl (2): ...5, Absence (0)/Presence (1)  
ℹ Use `spec()` to retrieve the full column specification for this data. ℹ  
Specify the column types or set `show\_col\_types = FALSE` to quiet this message.  
• `` -> `...5`

View(FD)  
summary(FD)

Feeder\_number Collection\_Date Absence\_0\_Presence\_1 Feeder\_Type   
 Min. : 1.00 Length:144 Min. :0.0000 Length:144   
 1st Qu.: 3.75 Class :character 1st Qu.:0.0000 Class :character   
 Median : 6.50 Mode :character Median :0.0000 Mode :character   
 Mean : 6.50 Mean :0.1458   
 3rd Qu.: 9.25 3rd Qu.:0.0000   
 Max. :12.00 Max. :1.0000   
 ...5 Absence (0)/Presence (1)  
 Mode:logical Mode:logical   
 NA's:144 NA's:144

*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.*

## 4.3 Statistical analysis

*Explain anything related to your statistical analyses.*

# 5. Results

## 5.1 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.*

[Table 1](#tbl-summarytable) shows a summary of the data.

Note the loading of the data providing a **relative** path using the ../../ notation. (Two dots means a folder up). You never want to specify an **absolute** path like C:\ahandel\myproject\results\ because if you share this with someone, it won’t work for them since they don’t have that path. You can also use the here R package to create paths. See examples of that below.

Table 1: Data summary table.

| skim\_type | skim\_variable | n\_missing | complete\_rate | factor.ordered | factor.n\_unique | factor.top\_counts | numeric.mean | numeric.sd | numeric.p0 | numeric.p25 | numeric.p50 | numeric.p75 | numeric.p100 | numeric.hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| factor | Gender | 0 | 1 | FALSE | 3 | M: 4, F: 3, O: 2 | NA | NA | NA | NA | NA | NA | NA | NA |
| numeric | Height | 0 | 1 | NA | NA | NA | 165.66667 | 15.97655 | 133 | 156 | 166 | 178 | 183 | ▂▁▃▃▇ |
| numeric | Weight | 0 | 1 | NA | NA | NA | 70.11111 | 21.24526 | 45 | 55 | 70 | 80 | 110 | ▇▂▃▂▂ |

## 5.2 Basic statistical analysis

While I am unsure of proper analysis, I may start by calculating probabilities. It may also be helpful to run a regression or linear model. Other than this, I am unsure of how to start the analysis process for the Picnic Table data.

*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.*

[Figure 1](#fig-result) shows a scatterplot figure produced by one of the R scripts.

|  |
| --- |
| Figure 1: Height and weight stratified by gender. |

## 5.3 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.*

Example [Table 2](#tbl-resulttable2) shows a summary of a linear model fit.

Table 2: Linear model fit table.

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 149.2726967 | 23.3823360 | 6.3839942 | 0.0013962 |
| Weight | 0.2623972 | 0.3512436 | 0.7470519 | 0.4886517 |
| GenderM | -2.1244913 | 15.5488953 | -0.1366329 | 0.8966520 |
| GenderO | -4.7644739 | 19.0114155 | -0.2506112 | 0.8120871 |

# 6. Discussion

## 6.1 Summary and Interpretation

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

## 6.2 Strengths and Limitations

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

## 6.3 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.

# 7. 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>

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>