# EDLD 710: Data Analysis for Problems of Practice

Jack B. Huber, Ph.D.

2022-08-31T08:46:51-07:00

# Table of contents

# Purpose of this Document

The purpose of this site[1] is to offer targeted advice and resources to help you analyze the quantitative data you have collected for your capstone project.

The chapter structure to the left illustrates the organization of this site. Each page attempts to identify and deal with important issues that come up in the process on analyzing quantitative data. This includes explanation, advice, and links to resources where available and appropriate.

These pages are works in progress. The intent is for them to evolve in adaptation to your needs. I will continue to edit them as I see fit; and if something is missing, please don't hesitate to reach out to me: huberj@gonzaga.edu.[2]

[1] This site was built in `R` (**R-base?**) with `R Markdown` (**R-rmarkdown?**).
[2] A little bit about me.

# Preparing your Data

The purpose of this page is to help you clean, organize, and otherwise prepare and enhance your data for statistical analysis and/or visualization.

---

## Best practices for Excel

In all likelihood the primary tool you'll use for working with your data is Excel - possibly by now the most commonly used tool for working with data in the world. What follows is a list of good practices that might make your Excel data easier to manage and collaborate with others. These tips will save you time wasted on fixing data and will help keep your data in format appropriate for analysis:

**Put *variables in columns* and *observations in rows***. Include a unique identifying number for each case. Be sure that each variable name is unique (no duplicate variable names).

**Put *variable names* in the *first row***. Variables must start with a letter. Do not include special characters (#, !, ?, %, etc.) or spaces in your variable names.

**Keep all your data "touching". *No empty rows or columns***. This is critically important for sorting. Empty columns or rows break the structural integrity of your data set and could allow you to sort a subsection of your data apart the rest of it.

**No merged cells**.

**Use a separate column for each piece of information**. Don't enter data such as "120/80" for blood pressure. Enter systolic blood pressure as one variable and diastolic blood pressure as another variable. Don't enter data as "A,C,D" or "BDF" if there are three possible answers to a question. Include a separate column for each answer.

**Decide on a "missingness" convention**. Missing data can cause a multitude of problems. To enter a missing data value either enter a blank or an "impossible" numeric code (for numbers) or an easily recognizable single digit character code for character (trying to avoid mixing numeric and character data). Be sure, if you use a missing value code, that it cannot be confused with a "real" data value.

**Use only one worksheet for your data; do analysis on a different worksheet**. If you decide to use multiple sheets for you data, follow the variable naming conventions for the tabs that name the sheets (keep the names simple and unique).

**Do not "stack" data on the same sheets**. For example, "treated" versus "non-treated" patients can be handled by column variable that has a code for Treated (yes/no).

**Dedicate one worksheet to your original, unedited raw data. Make a copy of it to do all your cleaning and analysis**. You might label this worksheet "Original" or "Raw data." This is important so that **when you make a mistake, you always have your original data to fall back on**.

**Dedicate one worksheet to your Clean / Working data**.

**Make the most of your variable labels**. On your worksheet of "Clean" (or "Working") data, make sure every column of data has a clear, concise, descriptive label. Here's what I do to take column labels to the next level:

- Ensure that the top row of my data includes a clear, concise label for each column of data.
- Bold the row.
- Add a fill color to the row.
- Freeze the row (Select the row, then View –> Freeze top row).
- Enable word wrap in the row.

**Apply a consistent format for your columns**. Data elements are different sizes. Names tend to be long while numerical values tend to be short. I don't like it when a column label is left-aligned but the data are right-aligned. I find these variations in visual formatting distracting. To deal with these distractions, I tend to:

- Apply all the column label formatting mentioned above.
- Fix all my column widths to 15.
- Left align columns (both column labels and data) for text (patient names, medication names, etc.).
- Center columns (both column labels and data) for numeric values.
- Right align columns for time data.

I find (and I think you will too) that enforcing a consistent format removes variable formatting as a distraction so I can see and focus on the data.

---

## Excel skills

Here are the skills you will most likely need and use:

- Sorting your data array on a column
- Filtering your data array based on specific values of one or more columns
- Fill down
- Pivot Table
- VLOOKUP function - to matching together related data from different sources
- Conditional formatting
- Basic calculations (SUM, AVG, COUNTIF, etc.)
- CONCATENATE function - to stitch together text and values from different data columns into a new column (which is sometimes helpful and necessary but is generally bad data practice to be avoided)

---

## Resources for data preparation

Tidy Data, by Hadley Wickham (a well-known data scientist), is a classic paper that defines what makes data clean (or "tidy") [@WickhamTidy]

The University of New Hampshire Library has an excellent research guide for using Excel, including data cleaning, data analysis, data visualization, and spreadsheet best practices.

Preparing Data in Excel, from the University of Nebraska Medical Center College of Public Health, has an excellent set of guidelines for working with Excel

Introduction to Excel is an excellent online module from the University of South Australia Research Methodologies and Statistics department.

Analysis Ready Datasets is an excellent resource from Harvard Medical School

---

## Granularity of data

Granularity of data means two related things worth your attention:

1. One is the *size of the data point*, which is to say what context it provides for other, smaller, data points. Put another way, what data points do you intend to count or summarize, and by which groups do you intend to compare these summaries?

2. The other is, essentially this question: What does a *row* in the spreadsheet mean? Because Excel counts rows, but what's contained in a row may not be what you intend to count.

Here are several different levels of granularity of Epic data:

**Patient level**. A patient has a unique ID number: the MRN. No two patients have the same MRN. When it comes to mining Epic data for the research project, the patient list is perhaps the most important: the resident needs a "patient list". When it comes to data mining, the patient "level" is context to more granular data in the sense that a patient can have multiple encounters - and thus multiple Encounter CSNs "within" the same Patient MRN.

**Encounter level**. The unique Epic ID number for the encounter is the CSN. The encounter is context to more granular data such as a treatment regimen of a particular medicine. Multiple drug administrations can occur "within" an encounter CSN.

**Medication administration level**. This is possibly the lowest level and the most granular data. In Caboodle, each administration of a medicine has a unique ID number and is time-stamped. My queries to date have been for counts of medicine administrations, or firsts, lasts, minimums and maximum doses within a hospital encounter or ICU stay.

**Lab results level**. Lab data is similar to medicine administration data because, again in Caboodle, each lab result is has its own unique ID number and is time-stamped. There can be a great many lab results within a hospital encounter. My queries to date have been for counts of lab results, or firsts, lasts, minimums and maximum values within a hospital encounter or ICU stay.

The resident data collection form is designed for patient level data; each row in the spreadsheet captures the experience of a hospital encounter. It can also be helpful to report the medicine administration and lab result data sorted chronologically by patient and encounter.

# Analyzing your Data

The purpose of this page is to give you tools to analyze your data using appropriate statistics. There are two important tasks:

1. Define the measurement levels of your variables.
2. Choose statistics appropriate for the measurement levels of your variables and the purpose of your analysis.

---

## Define the measurement levels of your variables

It is important to know the measurement level of your variables (De Muth 2009). How do you express the outcome by which to compare your pre- and post- samples? Is it…

- Percent of patients who achieve an initial therapeutic goal?
- Time to initial therapeutic level?
- Percent of patients who experience an adverse outcome (such as acute kidney injury)?
- Mortality rate (% surviving)? In such a case you would be comparing two proportions.
- "Time to…" a therapeutic level? In such a case you would be comparing two different quantities of time.

### Nominal measurement

Nominal measurement is categories. Each patient must fall into only one category, and the categories must be mutually exclusive and exhaustive. Here are some examples:

- Gender (Male/Female)
- Racial identity
- Marital status
- Control group/Experimental group
- Infected with COVID vs. not infected with COVID
- Disease presence
- Mortality

Outcomes are usually reported as frequency counts or percentages (in each category).

### Ordinal measurement

Ordinal measure also puts patients into categories, but the categories have an ascending or descending order: patients have *more* or *less* of somthing. But the differences between the categories is not necessarily the same. Here are some examples:

- Stages I-IV tumors
- 0-10 Apgar scores

A Stage IV tumor is more advanced than a Stage II tumor, but not necessarily by twice as much. A Stage III tumor is more advanced than a Stage I tumor, but not necessarily by three times as much.

For this reason, we cannot perform arithmetic or calculate means or other parametric statistics on ordinal values.

However, if your project has an ordinal level outcome on which you need to compare treatment groups, there are appropriate **nonparametric** statistics you can use to see which group is significantly *more of* this outcome than another. Examples include:

- **chi-square** $\chi 2$ statistics
- the **Mann-Whitney U** test
- the **Spearman's rho** test

### Interval and ratio level measurement

Finally, interval and ratio measurement means **continuous** data: patients fall somewhere on a continuum, like a temperature scale. As a result, variables measured using interval and ratio scales are often referred to as continuous variables. Here are some examples:

- Height
- Weight
- Cholesterol level
- Blood pressure
- Time

On variables like these there is relative positioning with no gaps or interruptions in the continuum.

The difference between interval and ratio scales is that ratio has a true zero value while interval does not.

On variables like these it is permissible to do arithmetic and to summarize them with the **mean** and **standard deviation** which, in turn, avail to you more commonly used advanced statistics like:

- t tests
- analyses of variance (ANOVA)
- correlation
- regression

Once you have a good feel for the measurement levels of your outcome and predictor variables, you can choose appropriate statistics. (**Simpson2015?**) offers two decision trees to help you make these choices:

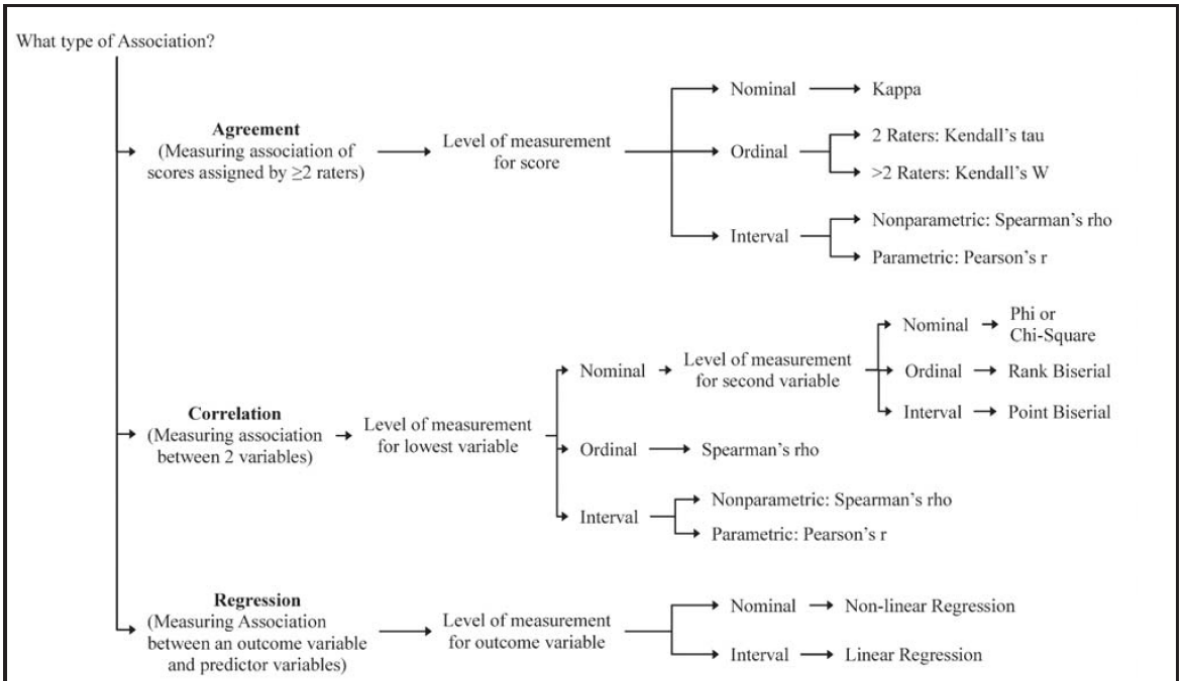


**Figure 1.** Decision tree to identify inferential statistics for an association.