EDLD 710: Data Analysis for Problems of Practice

Jack B. Huber, Ph.D.

2022-08-31T08:25:40-07:00

Table of contents

Purpose of this Document

The purpose of this document is to assemble some advice and resources to support Swedish Pharmacy residents in their journey from project proposal to a completed research project for presentation and publication.

Planning the Project

The causal logic of the project

The resident researcher should begin the project with a basic sense of the design logic of the project. The point of this section is to introduce or reinforce a few basic concepts of research design as they apply to the resident research project.

The project aspires to make a case for a proposed improved medical treatment. This case rests on causal evidence that the proposed treatment is more clinically effective than the current mode of treatment. The most convincing case would demonstrate that the treatment alone caused the improvement and cast doubt that the improvement would have happened anyway. To thoughtfully plan data collection for the strongest causal evidence, and to anticipate and minimize challenges of rival explanations, is the purpose of research **design** (Campbell and Stanley 1963).

The ideal design would be based on random assignment of patients to conditions, as in a randomly controlled trial. The resident researcher would randomly assign patients to a control condition and various treatment conditions, then compare outcomes of all these groups following treatment, and the outcomes will differ at least slightly. Assume patients in the treatment conditions had better outcomes than patients in the control condition. The resident could attribute this difference to the treatment alone because in all other ways the groups would differ only by chance *by design*.

Random assignment is probably not an option for resident research project. Until that happens, the project falls under the category of **quasi-experiment** which means it is more vulnerable to **confounding explanations** of any differences in outcomes.

The resident research project based on retrospective chart review will culminate in a comparison between two groups: a pre-implementation group and a post-implementation group. Assume a set of data that show better outcomes for the post-implementation group. Great! Can we attribute that to implementation of an improved treatment or protocol, or for some other reason would the post-implementation group have fared better anyway?

One potentially confounding pre-existing difference between the two groups is *time*, or history. A recent resident project provides a good example. In this project, the time frame for the pre-implementation group was early 2020 and thus this group had a higher incidence of COVID-19 infection than the more recent post-implementation group. Were outcomes for the pre-implementation group less favorable because more of them were infected with COVID-19? The

proposed dosing protocol may very well have improved outcomes for the post-implementation group. The problem is there are competing explanations of the results. Pre-existing differences between the two groups in COVID-19 infection amounted to a *confounding* variable due to history.

The point here is not to teach a course in research design but to help the resident researcher clarify for this project:

- What are the primary outcomes to improve? Length of stay? Time-to-therapeutic level? These are *dependent* variables. They depend on, or are the effects of, other variables.
- What is the difference in treatment intended to cause the improvement in the post-implementation group? This is the independent variable.
- All other variables are control variables. They should differ only by chance. If there is a noticeable pre-existing difference, and the level of that factor in one group affects the outcome, it is a confound.

How many patients?

Possibly the most pressing question for resident research projects is: **How many patients do I need?**

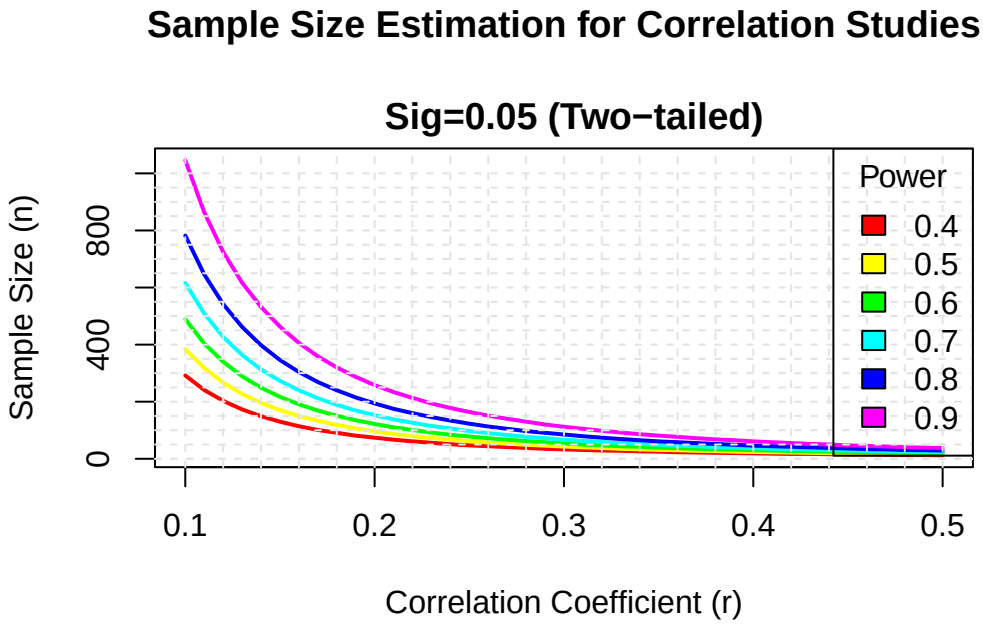
The resident research project will culiminate in a series of comparisons between the pre-implementation and post-implementation groups. Outcomes of the two samples will differ by at least some quantity. The resident researcher expresses this difference as an effect, like this:

[insert table about here]

Assume that this effect suggests more favorable outcomes for the post-implementation group. This effect raises several questions:

- How do we evaluate this effect?
- Could we attribute it to chance? (because it would be very unlikely for both groups to have exactly the same outcomes)
- Or is it larger than that?

In statistical terms, this is a question of **statistical power**. Power is the ability to isolate a treatment effect when it really does exist (Cohen 1988). Power is a function of effect size, sample size, and statistical significance. In order to decide on a number of patients we need to have a sense of what size of effect we want to reliably detect.



This a plot of power rates against these other variables. A large effect ($r \sim 0.5$) is detectable with a sample of any size. But only large samples have the power to detect a statistically significant ($p < .05$) small correlation ($r = 0.1$).

Collecting your Data

Data Sources

For collecting data from Epic, it might be helpful to have a sense of the landscape of its different databases. There are three primary databases:

Chronicles. This is the database that is collecting data from Hyperspace in real time. For reporting, Reporting Workbench pulls data directly from Chronicles, but it is otherwise not designed very well for historical reporting.

Clarity. This is the primary relational database for reporting Epic data. Through a nightly process known as ETL ("extract-transform-load"), Clarity extracts data from Chronicles and stores it in a thousand bazillion tables (think "spreadsheets"). To pull data for a report from Clarity is to identify the correct tables and fields and to write a SQL query to join the tables, apply the appropriate selection criteria, and report the appropriate fields.

Caboodle. This a relatively new relational database that functions essentially the same as Clarity but is designed to be much easier to use. Caboodle uses fewer tables derived from myriad Clarity tables which vastly simplifies the work and complexity of writing a SQL query. The down side is not all Clarity data are in Caboodle.

Granularity of data

Granularity means two related things. One is the size of the data point which is to say what context it provides for other, smaller, data points. The other is, essentially this question: What does a row in the spreadsheet mean? There are several different levels of granularity:

Patient-level data. 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 data. 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 data. 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 data. 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.

Advice for Data Collection and Management

Be proactive. As you begin to decide what data to collect, submit your project plan and requests for data in writing to me (Jack) as soon as possible. This is so I can have a bit of time to understand your project, do some discovery in Caboodle or Clarity, and note any questions. Then meet with me via Teams to get on the same page.

Devise a system for organizing your data. By this I mean version control. This data collection process is iterative. You will ask for data and I'll write a SQL query that produces an Excel workbook of data. That's one iteration. In all likelihood you'll need a revision or two. Upon receiving your feedback I'll edit my query and produce for you a new set of data to replace the first set. That's the second iteration. The more iterations, the more data, the more potential for multiple versions and data overload. And I may not be able to do more than a few iterations. I can work on some best practices to help us through this. Stay tuned.

There is no substitute for chart review. Some fraction of resident research data can come from Epic data mining, but it still needs to be validated with careful chart review. And some data which are very difficult to mine or to query into the correct format may have to come from chart review.

1 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.

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:

1. **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).
2. **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.
3. **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.
4. **No merged cells**.
5. **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.
6. **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.
7. **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).

8. **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).

9. **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**.

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

11. **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.

12. **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.

---

2 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)

---

3 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

---

4 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.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Houghton Mifflin Company.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Mahwah, New Jersey: Lawrence Erlbaum Associates.

De Muth, James E. 2009. "Overview of Biostatistics Used in Clinical Research." *American Journal of Health-System Pharmacy* 66: 70–81.

5 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.

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.

5.1 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).

5.2 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

5.3 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:

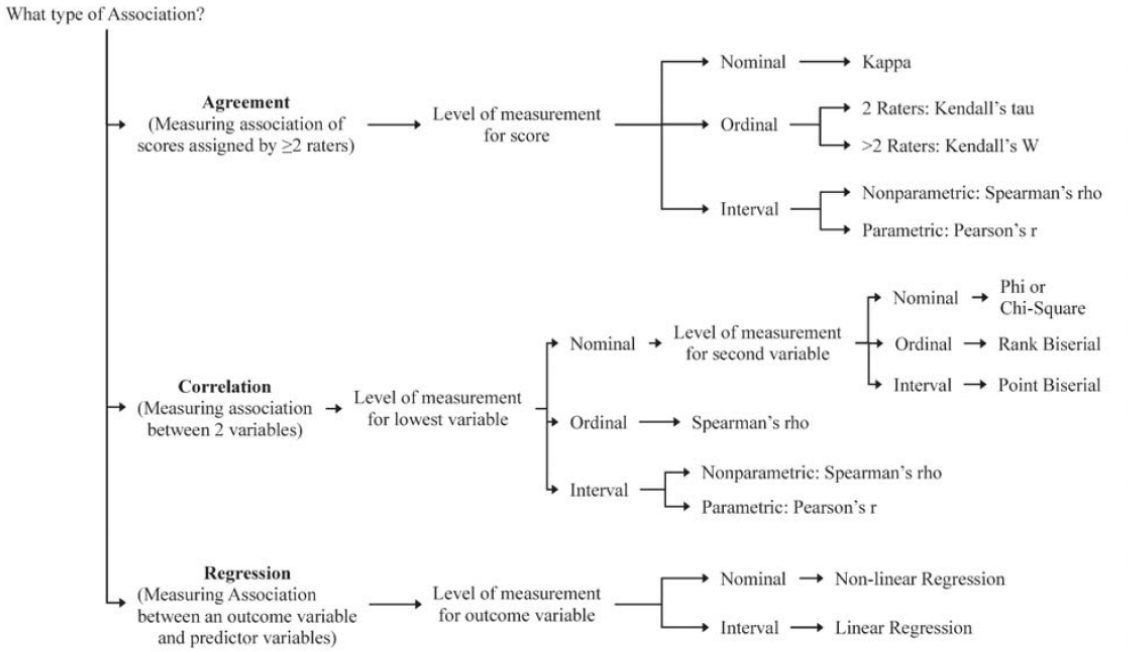


Figure 1. Decision tree to identify inferential statistics for an association.

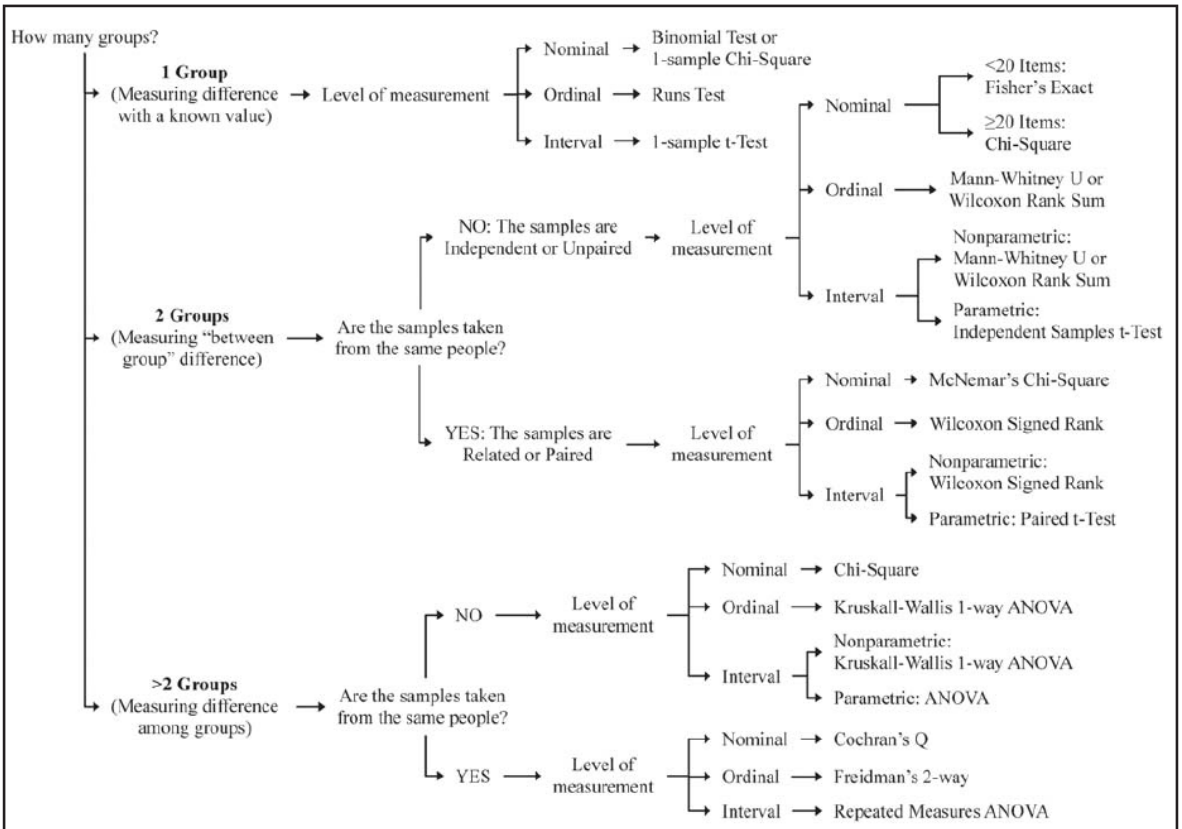


Figure 2. Decision tree to identify inferential statistics for measuring a difference.

6 Choose appropriate statistics

This section offers more information on several choice statistics. These are statistics I've used for recent resident projects and seen in the journals.

6.0.1 Chi-square $\chi 2$ tests

The chi-square $\chi 2$ test is a commonly used statistic for nominal/categorical data. We use it to examine the distribution of cases across **categories**. Essentially, it compares the distribution of cases you actually see to the distribution of cases you would expect to see from normal variation.

Here is one example of a chi-square $\chi 2$ test for a recent resident project. The question is whether gender (male/female) makes a statistically significant difference in whether patients need three or more dose changes of bivalirudin before they reach a therapeutic goal.

```
#d <- read.csv("data/bivalirudin.csv") # load data
#table_dosechgs_gender <- xtabs(~d$d_Male + d$DV_3DoseChanges, data=d) # crosstabulate
#knitr::kable(table_dosechgs_gender, align = "l")
#summary(table_dosechgs_gender) # calculate chi-square
```

The chi-square $\chi 2$ value of 1.9421 with one degree of freedom has a p-value of 0.1634. It is not statistically significant, suggesting that gender makes no significant difference in reaching therapeutic goal.

6.0.2 t tests

The t test is a commonly used statistic for comparing two groups on a continuous outcome. Here are some examples of t tests from a recent resident project:

propofol_stats.xlsx

6.1 External resources for data analysis

Here are a few links to external resources on data analysis and statistics.

- The R Psychologist, by @magnussonCohend - is an outstanding resource to better understand statistics
- Online Modules in Research Methods and Data Analysis at the University of South Australia
- Data Analysis from the University of New Hampshire

---

7 Software tools for data analysis

Of equal importance to the didactics of statistics are the brass tacks of software for working with statistics. Here are several software tools for analyzing data:

7.1 Excel-based tools

- EZAnalyze is a simple Excel add-in for data analysis. It includes menus for selecting various statistics from your variables.
- XLStat
- Data Analysis native add-in

7.2 Data analysis software

- R @R-base, with RStudio and R Markdown, is **free** open source software for data analysis, statistics, and data visualization. It is powerful and flexible but it does require ongoing learning of code because it is constantly evolving.

Here are several other robust software applications for data analysis available to you. One or more of them may be free for you as a Gonzaga student:

- JMP - JMP is a suite of software used for statistical analysis
- SAS - The SAS System is a comprehensive statistical software package from SAS Institute for data management, graphics, analysis, and presentation
- SPSS - IBM SPSS (Statistical Package for the Social Sciences) provides data and statistical analysis, file management capabilities, graphics and reporting features.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Houghton Mifflin Company.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Mahwah, New Jersey: Lawrence Erlbaum Associates.

De Muth, James E. 2009. "Overview of Biostatistics Used in Clinical Research." *American Journal of Health-System Pharmacy* 66: 70–81.

8 Reporting your Data

The purpose of this final page is to help you decide how best to package and present the results of your data analysis for a professional audience.

"Figures should be accurate, clear, and concise. As with tables, the figure with its title and legend should be understandable without undue reference to the text."[1]

8.1 1. Line graphs for over-time data

Line graphs are the appropriate way to show change over time in one or a few groups. Your x-axis (horizontal) should be time, and your y-axis should be the quantity by which you want to see change over time. You can use different lines for different groups.

8.2 2. Bar graphs for group comparisons.

The bar graph is the Swiss army knife of data visualization. It's useful because it is so versatile. Bar graphs use size to compare different quantities. Your y-axis is your quantity and on your x-axis you put your group categories.

8.3 3. Save the pies for dessert

Pie graphs are a great way to visualize proportions - parts that make up a whole. And, with a few nice colors, they're attractive. They're also simple.

But they can be a pain to create - to get right visually. They also lose their utility when you have more than a few categories. For this reason, the AMA discourages the use of pie graphs:

[1] From the AMA Style Guide, Section 4.2 on Figures

If you do want to use a pie graph, my advice to you is to keep it simple; use it only to show a few categories.

---

9 Best practices for tables

Although data visualization has become all the rage, there is a still a place at the table for tables (sorry - bad pun - I couldn't resist). Tables remain a concise way to present a sizable amount of quantitative data.

I encourage you to strive for your tables to meet this standard: "A properly designed and constructed table should be able to stand independently, without requiring undue reference to the text."[1]

I encourage you to review the section on tables in the APA Style Guide.

- What goes on the far left column?
- How do I format the main column headings?
- How do I report p-values?
- How do I report my data?
- How do I align things?

This article from (**MillerEtAl2020?**) illustrates how to present a table that reports the results of a series of chi-square tests of the differences between two groups (pre-implementation and post-implementation) on categorical outcomes.

Notice that the authors showed the independent variable (intervention group) as columns which enables us as readers to compare outcomes by reading left to right.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Houghton Mifflin Company.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Mahwah, New Jersey: Lawrence Erlbaum Associates.

De Muth, James E. 2009. "Overview of Biostatistics Used in Clinical Research." *American Journal of Health-System Pharmacy* 66: 70–81.

[1] From Section 4.1 Tables, in AMA Style Guide

References

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Houghton Mifflin Company.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Mahwah, New Jersey: Lawrence Erlbaum Associates.

De Muth, James E. 2009. "Overview of Biostatistics Used in Clinical Research." *American Journal of Health-System Pharmacy* 66: 70–81.