(Module 2) **Data Analysis** September 30, 2015

What is Data Analysis?

Now that you've piqued your interest in research, the first step is to learn how to both take and analyze data. We can't help you with the taking data part, that will be up to you, but as for analyzing it, we can help you out. Analyzing data in an efficient, effective, and responsible manner is arguably one of the most important skills in research. Anyone can take data, but if you cannot extract meaning from your data and provide evidence that your results are in fact results and not just a statistical anomaly, then you can't really move forward with your work. We'd argue that the three principle tenants of data analysis were mentioned above and here's why they are important:

- **Efficient**: In data collection, you will almost always have more data than you personally can handle, requiring some sort of analysis software or code. Efficiency in coding and organizing data will save you both time and an unbelievable amount of effort.
- **Effective**: To analyze data effectively refers to being able to focus on what portion of your data is important and what you need to do to extract the meaningful information relevant to your hypothesis/experiment.
- **Reliable**: Reliable data analysis refers to correctly considering error in your analysis. This means performing error analysis and conveying that to anyone reading your data/work. We will cover error analysis later, but in short this is what proves to everyone that your methods are trustworthy.
- **Fun?** Sometimes.

How to Analyze Data

Data analysis isn't so much a question of "Why should I do it" but more a question "How do I do it?.. No really someone tell me.." This is something you will probably say at some point in your research career because when it comes to analyzing data one of the hardest parts is figuring out where to start. Oftentimes research projects yield data with very specific needs. Sometimes you will have to develop your own methods to analyze your data, and sometimes you will have to search existing software to deal with complex data plotting, fitting, etc.. Our goal is to provide you with a brief list of useful data analysis software, when you use them, what you use them for, and give you the resources to learn more. Basically software speed dating. Before we go into this we would like to give you a brief overview of error analysis and emphasize the importance of this step in the analysis process. If you have taken a stats course you may be familiar with some of the concepts, but for the most part error analysis is something that is rarely touched until you get into research.

What is error and how do I make it go away?

Error: The error of an observed value is the deviation of the observed value from the true (unobservable) value of a quantity of interest. Reducing the error in an experiment is a matter of experimental repetition. The size of your error actually decreases as a function of $1/\sqrt{N}$ where N are the number of experiments done. Because we cannot perform infinite experiments there will always be some amount of experimental uncertainty.

This is the definition of error. Every experiment you perform will have some sort of error and it is your job to consider all of it in your results. The more times you conduct an experiment, i.e the larger your sample size is the less erroneous your data set will be, which is why reproducibility of results is necessary in science.

Another important facet of error analysis are **Error Bars** . These are a graphical representation of the variability of data on graphs and are used to indicate error/uncertainty in a measurement.

An important question you might ask yourself is "How do I know when my results are good enough?". This is a great question and the answer can vary depending on what it is you are studying, but oftentimes in research an accepted value is one with a "p value" p=.05. What does this mean? A "p value" is used to describe the probability of observing the results (or more extreme results) you did while assuming the "null hypothesis" to be true, meaning you assume that there is no correlation/association in your dataset. When p=.05 it means that there was a 5 percent chance of getting the data you did assuming there was no relation, which will usually be enough to argue that your result is "statistically significant". This is necessary for any peer reviewed work.

Before we hit you with a barrage of different software to help you in your data analysis adventures, we'd like to bring up **Fitting** data. When you take a multitude of data points you almost always want to fit your data to a mathematical function that has the best fir to a series of data points, potentially subject to known constraints. Fitting data is one of the most important and difficult aspects of the analysis process. When you fit data you are normally looking for some particular trend in your data or checking to see if the fit matches a standard curve relevant to your research.

Software

Github:

- *Related Topics*: All topics.
- *Functionality*: File sharing/storage. Widely used for publishing code and group coding efforts.
- *Resources*: https://guides.github.com/activities/hello-world/

Python:

- *Related Topics*: Very ubiquitous. Everything
- *Functionality*: Calculations, image analysis, data visualization.
- *Resources:* https://www.ocf.berkeley.edu/ ipasha/python/

MatLAB:

- *Related Topics*: Very Ubiquitous.
- *Functionality*: MatLAB and Python have similar functionality. Calculations, image analysis, data visualization.
- *Resources:* http://www.tutorialspoint.com/matlab/

Mathematica:

- *Related Topics*:(abstract) Mathematics/Engineering/Modeling
- *Functionality*: Designed for computationally heavy work and mathematical modeling. Can be used to perform calculations and fitting on data, as well
- *Resources:* https://www.cs.purdue.edu/homes/ayg/CS590C/www/mathematica/math.html

Origin 8:

- *Related Topics*: All topics (focus on fitting data)
- *Functionality*: "Souped up excel". Used for treating, plotting, fitting, and performing analysis on fitted data. This software is will allow you to easily generate fitted plots with error bars, resulting in "publication quality" plots.
- *Resources*: http://exphys.science.upjs.sk/sites/default/files/zaprog/OriginTutorials.pdf

MS Excel:

- *Related Topics*: All topics.
- *Functionality*: Most of you are probably familiar with the basic functionality of Excel. Excel can be used for calculations, plotting, as well as more advanced fitting and error analysis.
- *Resources*: http://www.excel-easy.com/data-analysis.html

Image J:

- *Related Topics*: Image Analysis
- *Functionality*: Image J is freeware that allows you to analyze images, perform image calculations, record intensities, create image stacks, and write scripts to perform various other functions.

- *Resources:* http://imagej.nih.gov/ij/

IDL:

- *Related Topics*: Data Analysis and Visualization
- *Functionality*: Not Free. Scripts can be run. Especially good for large amounts of data and for doing matrix manipulation as well as plotting
- *Resources:* http://www.pha.jhu.edu/ wonnell/TutStuff/IDLtutor.html

DS9

- *Related Topics*: Image Analysis, Data Visualization
- *Functionality*: Stand-alone application,free, user-configurable GUI. Used to analyze astronomical images.
- *Resources*: http://ds9.si.edu/doc/user/index.html

Contact

If you have any questions about what you heard today or about Research Workshop in general contact today's instructors at the emails below, or chat us on Facebook:

- MacCallum Robertson (**maccallumr@berkeley.edu**)
- Christopher Agostino (**cagostino@berkeley.edu**)

Next week's Research Workshop topic:

Simulations and Computer Aided Design (CAD)