The PhD Student's Guide To Processing And Presenting Important Data

Dr. John Hockey
Dr. Sandeep Gupta

Table of Contents

Introduction

Most PhD projects involve original research. And much of that research involves experiments, simulations, measurements of various kinds, and ultimately processing of that data to give meaningful results.

We remember during our PhDs, we had to process hundreds of millions of datapoints to get just one meaningful number, or worse still, to answer just one “yes or no” question – weeks of preparation, days of measurements, days of data processing and analysis boiling down to just one word, “Yes” or “No”. But that’s the nature of research – we’re interested in the final result.

But while we’re interested in that final result, we also need to be convinced that the final result is based on

sound principles. If we take 100 measurements, and randomly pick just one of those measurements to answer a particular question, is that a sound method? On the other hand, what if 95 of those measurements lie within a certain range while the remaining 5 lie well outside the range, is the conclusion drawn from those 95 measurements still valid?

If you go to all the trouble of planning a research campaign, conducting the campaign, analyzing the data, and finally writing a paper, don't you want to make sure that it will be accepted by the journal that you submit it to? We're also in favor of getting our works accepted (while we were doing my PhD, and in the many years since), but it's contingent on those assessing your work (the editor and the peer-reviewers) to be convinced that your work is rigorous, that it's *statistically significant*. If the results and conclusions drawn are not statistically significant, then the work doesn't mean much.

Likewise, when presenting at a conference, or to industry partners, they also want the results and conclusions to be statistically significant (or, “conclusive”).

Being able to show that your work is of high-quality and significant hinges on statistics and error analysis (also known as “uncertainty analysis”).

Every PhD student who deals with measurements and data, regardless of their field, must be up-to-date with fundamental statistics and error analysis. Presenting data and having someone ask you what the error in the measurements is, only to look down at your shoes and mumble that you don’t know, doesn’t fill the audience with confidence. On the other hand, being able to look the questioner in the eye and tell them that the error is less than the differences among the data fills the

audience, and yourself, with confidence in your work and results.

The best part about it is: stats and error analysis aren't hard! We know! We were as surprised as you are, but learning the stats and error analysis required for high-quality research is quite easy! Within a few hours, you can go from knowing very little about either sub-discipline, to knowing everything you need to ensure that your data, results, and conclusions are "bullet-proof". Standing up on stage and presenting your research, or submitting your thesis for examination, while knowing that your research is "bullet-proof" is a great confidence booster – you don't have to worry about looking like "an imposter" (to borrowing the term from Dr. Clance and Dr. Imes regarding the "Imposter Syndrome").

But there's one more benefit to understanding stats. To understand what that benefit is, let us ask you the

following question: How great would it be if you could make your good paper great? Or, being able to make your good paper two good papers? We bet you'd be happy if you could do that – we were too, the first time each of us managed to do both of those things.

One of the keys to strengthening your papers and multiplying them is your skills to properly analyze data. Understanding the stats skills that every researcher needs gives you the ability to multiply your papers and increase their strength. It's surprising how many people analyze their data once, write their paper and then move on – leaving so much more to be discovered – it's wasteful, but also heartbreaking to watch because the PhD student or researcher has to go back to do more work to get more results – there are still more results there waiting to be discovered!! Just re-analyze the data with different statistical approaches! You'll make your life so much easier! Why neglect the data – use all of it,

make your PhD life easier! Use stats to do that – use stats to work smarter!

This book will start with the fundamentals of statistics – the stuff you need to know as a PhD student conducting high-quality research – making your papers great and multiplying your publications. It will then move into error analysis, and how to present your data such that the errors (or uncertainties) are shown.

Implement these skills to really strengthen your PhD, research, and publication output.

Let's begin!

What Are Statistics

First thing's first, just so we're on the same page: What are statistics?

Statistics is the subject of dealing with, and effectively expressing, data. The data could be just 1 or 2 datapoints to billions of datapoints. The more data you have to sift through, make sense of, and present, the more useful statistics becomes.

One simple example of stats in action is if you have 100 datapoints, how do you express them succinctly? Sure, every time you want to express to someone the results of your research, you *could* regurgitate every single one of those 100 datapoints, listing every little detail about them: datapoint 1 is 4% larger than datapoint 2, but 3% less than datapoint 3, and 5% larger than datapoint 4…so

on, and so on…or…you could give 1 or 2 numbers that *sum up* those 100 datapoints succinctly. And, **that's really what statistics boils down to – being able to succinctly sum up the hundreds, thousands, millions, or even billions of datapoints with a few numbers so that people can understand what your work is about without having to go through all of those datapoints themselves.**

What's more, being able to succinctly sum up those 100 datapoints (or however many you have) helps you – you don't need to remember every single datapoint – all you need to remember are the stats, and you get all the information you need! If you really want to dig through each dataset, then stats can also help you identify which ones to dig through and which ones to leave.

Within the world of stats, there are a few quantities that come up again and again, primarily because of how

useful they are. These quantities include the average and percentages, among others. We see these quantities everywhere in our everyday life. For example, look up Shaquille O'Neal's free-throw shooting *percentage* and you'll see a relatively low number (around the 50% mark). You also see these quantities in hardcore research, such as: "14% of the bacteria cultures showed increased growth over the 24 hr period." So, these quantities, and stats in general, are highly universal. They're so universal because of how useful they are.

So, let's go through the most useful statistical quantities that you'll need in your research.

NOTE: Pages 13 to 30 have been omitted in this sample version.

Statistical Graphs

In your papers or conference presentations, or even when you just want to show someone your work, you might want to include some graphs. Sure, stating numbers is very useful and descriptive, but graphs are another way of succinctly describing your data. We are very visual creatures, and looking at a graph for 5 seconds can give us a lot of information.

But, there are many different graphs that you could use to display your data. The question is: which one should you use?

You could be like 99% of people and just use the “good ol’ faithful” scatter plot, but it is not always the best one to use. You see, the type of graph you use will largely dictate what information can be found. That’s not only a

bad thing for the reader, it's also a bad thing for yourself; if you only use one type of graph when processing and analyzing your data, then you will almost always miss some of the trends and conclusions – those additional trends and conclusions could give you another publication, or some profound insight into your niche. That could be the difference between just understanding what everyone else understands, and being a leader in your field.

Just as a side note, making graphs has always been one of my (Sandeep) strengths – I'm willing to bet that, no one has made more graphs than me during my time researching – I love to make graphs comparing all different quantities. I remember countless times during my early academic career where I would come into a meeting with hundreds of graphs and everyone would just look at me like I was crazy – they would be sitting there with one or two, and wonder how on earth I could make so many graphs. Long story short, those graphs led

to dozens of papers that wouldn't have been written if I hadn't made those graphs. While everyone else was struggling to extract enough information and find noteworthy trends for a single paper, I was having trouble limiting what I wanted to put into my many papers. Inevitably, there would be so much that I wanted to put in, that the data had to spill over into more papers – that's the power of making all different types of graphs. And statistical graphs are among the most powerful types of graphs – they show so much information that would be hidden otherwise.

So, let's go through some of the most powerful types of statistical graphs.

NOTE: Pages 34 to 135 have been omitted in this sample version.

Conclusion

Congratulations, you've made it to the end of the book. So you should feel confident in your abilities to better analyze and present your research, as well as make your experiments (and simulations) more accurate!

Some final points to sum up what we've covered:

- Remember to use several different types of graphs to tease out more trends. You'd be surprised what's hiding.
- Make sure to use stats to understand how significant trends and conclusions are, and use them to succinctly communicate your work.
- Remember to do an error and uncertainty analysis of your experiments – it helps you, and everyone else, trust your work. You'll be much more confident because of it.

- Reduce the uncertainty of your experiments by taking advantage of the “Standard Uncertainty Of The Mean” equation – simply repeat your experiment (even once), and you’ll dramatically reduce your uncertainty, and dramatically increase your faith in your answers!
- Be explicit about the Confidence Interval that you use – you can choose to use whichever one you want, as long as you express it in your work.

NOTE: Pages 137 to 138 have been omitted in this sample version.