The Graduate Student's Guide to Statistics

A clear description of how to do statistics for your thesis or dissertation

Jerome V. Braun, Ph.D.

The Graduate Student's Guide to Statistics

A clear description of how to do statistics for your thesis or dissertation

Jerome V. Braun, Ph.D.

This book is for sale at http://leanpub.com/the-graduate-students-guide-to-statistics

This version was published on 2014-11-02

Contents

Preface

Graduate students get advice and guidance that varies across the board. I have seen some graduate students with amazing advisors and committee members that are obviously committed to the student's success, and who provide a lot of concrete advice and help to get them going and shepherd them through the whole process. I have seen other graduate students who were given next to no advice or advice that was frankly awful, and who were let to languish or even to fail. And, of course, everything in between.

What became obvious over time was that graduate students in general did not have an understanding of the big picture of how statistical analysis fit into their research program. The statistical analysis is only one small component of a larger set of activities.

This book is meant to be a practical guide to the big picture *and* the small details of statistical analysis. While there are too many specific statistical methodologies to cover even a tiny fraction of them, the framework is very similar from case to case.

Acknowledgements

Though my limitations are my own, I have learned a lot from my colleagues over the years.

Charles Du Mond at Syntex Corporation first introduced me to the culture of clinical biostatistics, as well as to the art of protocol review.

David Giltinan at Genentech, Inc. not only has an encyclopedic knowledge of clinical and nonclinical biostatistical techniques, but also made it a point to know the business behind the statistics. His dedication and drive set an expectation level that inspired striving for excellence.

Neil Willits at the University of California at Davis Statistical Laboratory is a consulting statistician that truly set the bar for me on sheer technical ability and adaptability. The more experience I obtain the more I appreciate Neil's decades of skill and experience.

Of course, without Hans Georg-Muller I would not have finished graduate school myself! Hans opened his home to me many times to work with him on my dissertation. In retrospect, Hans gave me a great model of a good advisor, able to hit all chords as necessary.

About the Author

I am an independent statistical consultant, and have consulted with well over 800 clients over the years in academia, industry, and government. These clients spanned a wide variety of industries and fields. I have experience in the corporate setting and with contract research organizations.

For about six years I worked at the Statistical Laboratory at the University of California at Davis and helped many hundreds of graduate students with their research, in addition to working with outside clients and with University faculty and staff on many hundreds of other projects.

Introduction

This Book Is For Graduate Students

You are either a Master's student or a Doctoral student, and have to write a thesis or dissertation.

You probably have to present your results in an oral examination or a defense, or in departmental seminars.

You may need to publish one or several academic articles, that will serve as your dissertation.

This book is for you. It is designed to help you with the statistical design, analysis, and writing necessary for you to graduate.

The Absolute Goal

The absolute goal is to graduate and get the degree.

Practically, this means that your job is to get the necessary signatures from your thesis or dissertation committee so you can get your degree, graduate, and move on with life! Also, you want to graduate in the shortest possible time frame.

Some Problem Personality Types

With enough experience, you start to see the same problems cropping up. Although this list and the descriptions may change after further reflection, here are some problem personality types that seem to have difficulty getting to the end of graduate school:

The *perfectionists*, who think that everything has to be black and white and perfect before they graduate. Nope. These are the absolute worst to deal with, they always have a reason to *not* graduate.

The *idealists*, who think that their work has to be ground-breaking before they graduate. Nope. Not everyone gets to be that lucky. If you are, awesome; if not, save it for post-graduate work.

The *procrastinators*, who seem to think that the work is going to be too hard and that graduation will *happen* simply by "hanging in there" and keeping warm and breathing. Nope. You have to get off your duff and start performing the tasks.

The *starry-eyed students*, who are now thoroughly disillusioned with academia and/or their mentors and advisors and academia in general, and are ready to give up after four or more years of work. Nope. It is your life, of course, but you will probably have more options if you finish the degree out.

Time

Trust me, you can probably do things differently and save a lot of time and pain on the way to getting your thesis or dissertation approved. It is too easy to become comfortable with a lack of progress and end up taking a year longer than necessary.

Look, you are already planning on spending at least four or five years of your life on this. An extra year on top of that represents at least 20% more time than planned; it might be 4% of your total life when you start. No one can afford to lose an extra year of earning power in this day and age.

If you are nearly done, then why stretch out the pain any more than necessary? I've watched many students simply allow themselves to go an extra half or full year because they couldn't get themselves to do the work now for whatever reason. The work didn't go away, it wasn't that hard, and they had to do it anyway—so what did they gain?

Topic Discussed in This Book

What I try to do in this book is give you a way to wrap your head around the statistical analysis part of your research. I think that will prevent you from getting stuck at that phase. Also, the earlier you start thinking about the statistical analysis, the more trouble and pain you will save yourself from experiencing in the first place!

With that in mind, this book covers:

- Writing the manuscript.
- Dealing with review comments.
- Creating a statistical analysis plan
- Handling communications with your advisor and committee.
- Presenting your results in seminars.
- Designing studies to avoid complication.
- Handling situations such as when you have no results.
- Handling potentially sticky ethical situations.

I also cover many specific technical issues, such as:

- Working with statistical packages or languages.
- Setting up data structures.
- Using linear models.
- Understanding the general flow of statistical analysis.
- Using data transformations.
- Working with a variety of specific statistical methods.

Triage

What if you are in trouble now? What to read and what to do?

End of Normative Time

If your study has already been long completed, your normative time clock is ticking, you are on filing fee, and your advisor and committee are breathing down your neck for your manuscript, then you are in trouble. You are about 25% of the graduate students that I have seen.

Where are you in the process of writing?

If you have barely started writing, then your problem isn't statistics but some other character flaw. Let the character flaw go for now for future work and start writing a rough draft so you can get to the next paragraph. See Writing the Manuscript.

If you have a Rough Draft started and are stuck on statistics, it would probably be best to get some competent outside advice on the statistical analysis. You probaby aren't getting good advice from your committee nor do you have the necessary skill, or you wouldn't be stuck here! Your goal is to get the most streamlined analysis specified that you can, then carry out that analysis and complete the Rough Draft. See Consulting with a Statistician.

Middle of Normative Time

If you have time to go but haven't just started, then chances are you are already committed to a study or line of research and it is already underway. About 60% of the students I have seen were in this phase. In the worst case, you have a poorly designed study that will be very difficult to analyze or that produces data that do not answer your research questions.

You need to plan your work better. See the section on how to write a Statistical Analysis Plan. Go forward from there.

Also, if you have not started writing, then I suggest writing a rough draft of your Introduction, Background, Methods, and References sections. See Writing the Mansucript.

Start of Normative Time

You are in the ideal world. You have not started your research yet, you may not have even picked a topic, and everything can be planned. You are among perhaps 15% of students I have seen. Everyone should be so smart!

In this case, you will probably want to start with reading the section the Research Proposal.

A Note to the Reader

Choices of Statistical Methodology

I am an applied statistician with a lot of experience.

To be honest, though, I would consider myself to be a Bayesian at heart, having been converted by Jaynes's beautiful exposition of probability theory.

It would be unrealistic in the extreme to attempt to be a statistical consultant and force everything into the Bayesian framework, even though in reality that framework is there whether wanted or not.

Working from first principles is always best, and I actually find it leads to superior consulting results. However, at the end of the day, the statistical analysis has to get done. So, at some point we have to leave the ivory tower and pick an actual method that can *actually* be implemented on our data, that will be *understandable* to our audience, and that our audience will *accept*.

If you are simply trying to get your analysis done, you probably have had the experience of facing a bewildering array of choices in terms of the statistical method to use. In the worst case, you might not completely understand the terminology or the "decisions" that the statistical package is making for you. But, that also is your choice.

So, choices of statistical methodology are always present. You may be comforted by this empirical observation:

> If the statistical methods chosen are at least reasonable for the data at hand, then the results probably would not change if you did the "right statistical analysis".

Criticism of Statistical Methodology

For some reason, statistics seems to attract a certain personality type. This personality type leans heavily into things like knee-jerk criticism, extreme nit-picking, and ultimately into a seeming need to create a black and white picture out of a world that is multi-colored. This can be pronounced.

Indeed, half of the field of academic statistics seems to be in setting some arbitrary preconditions for the world, then showing that some choice of statistical method is best in that world. Nothing to do with reality.

My point is this: You may disagree with me on my choice of statistical methodology and the way that I suggest to do things. That is very reasonable.

I probably agree with your criticism, actually! In fact, some of what I say to do is not necessarily "correct" by some standards that can be imposed.

But, a graduate student is *not* going to get anywhere trying to buck the tide, unless that is the goal of the research. That student is going to be much better off simply applying common statistical methods the same way as is usually done in the field.

Academic statisticians are very likely to turn their noses up at some of my recommendations. But, most of them could not analyze their way out of a paper bag, let alone provide actual timely help that gets the job done.

That having been said, I change what I do when I realize that I am wrong or that there is a better way.

Practical Ethics

Part of being a statistician is being able to say "No."

There is tremendous pressure on everyone in any area, whether research, government, or industry, to "lean" on the data and get "good results". This can range from subtle choices in statistical methods to outright data tampering.

What's the point? If you're a scientist, then this will lead to your work be non-replicable or to loss of reputation in the worst case. If you're in industry, then this can cost millions of dollars. If you're in government, again it can cost millions of dollars, now in taxpayer money. If you're in the legal system it may result in losing the case.

This also goes for "interpreting" the results. I've lost count of the number of times that people have willfully misinterpreted their results. I know it's willful since I just got through explaining it very clearly to them. Please save the wishful thinking for some other activity.

> **Horror story**: Leaving out the details for obvious reasons, I know of a multimillion dollar exercise in drug development that was predicated on a *single* incorrectly analyzed mouse study. This came to light when the study was replicated by an in-house scientist.
>
> The replication was unsuccessful. Well, not entirely unsuccessful. When the scientist "replicated" the *incorrect* analysis on the data, the results matched the first study. That drug was well on its way to Phase II by then, where it failed, of course.

If you are being pressured to analyze the data to a predetermined result, then it might actually help to visit a statistician. The statistician can take the heat for saying "No", can make the arguments in an objective manner, and might even have a constructive alternative.