Essays on Data Analysis

Roger D. Peng

Essays on Data Analysis

Roger D. Peng

This book is for sale at
http://leanpub.com/dataanalysisessays

This version was published on 2021-11-17

This is a Leanpub book. Leanpub empowers authors and publishers with the Lean Publishing process. Lean Publishing is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

Also By Roger D. Peng

R Programming for Data Science

The Art of Data Science

Exploratory Data Analysis with R

Executive Data Science

Report Writing for Data Science in R

Advanced Statistical Computing

The Data Science Salon

Conversations On Data Science

Mastering Software Development in R

Tidyverse Skills for Data Science in R

Contents

1. The Question 1

1. The Question

In 2011 I was teaching the course "Methods in Biostatistics" to graduate students in public health at Johns Hopkins University. The students in this course were getting Master's and PhD degrees in the variety of disciplines that make up public health and were all very bright and very motivated to learn. They would need the skills taught in my course to complete their thesis research and to do research in the future. I taught the final two terms of a year-long sequence. The topics covered in these two terms could broadly be classified as "regression modeling strategies", including linear regression, generalized linear models, survival analysis, and machine learning.

I distinctly remember that at the end of my second-to-last lecture for the entire school year, one student came up to me at the end to ask a question. She was Michelle[1], and at this point in the term I already knew she was going to ace the course. She was one of the best in the class this year. She came up to me and said, "This entire year, I feel like I've learned so many tools and techniques. But I still don't know, when I open a new dataset, **what should I do?**"

Data analysis is often taught in the negative. Don't do this, don't do that, that's a bad idea. It's rarely taught in the affirmative. You should always do this, you should definitely do that. The reason is because it's not possible to do so. Affirmative statements like that do not hold for all possible data analysis scenarios. To make use of a common phrase uttered by data analysts the world over, "It depends."

After she asked me this question I let out a nervous laugh. The truth is, I *hadn't* told her what to do. What I and the professors that came before me had done was given her a list of tools

[1] Not her real name.

and descriptions of how they worked. We taught them about when some tools were appropriate for certain kinds of data and when they were not. But we had not outlined a sequence of steps that one could take for any data analysis. The truth is, no such sequence exists. Data analysis is not a rote process with a one-size-fits-all process to follow.

But then, how is it that all these people out there are doing data analysis? How did *they* learn what to do?

When Michelle asked that question I still had one more lecture to go in the term. So I threw out whatever it was I was going to talk about and replaced it with a lecture naively titled “What to Do.” When I revealed the title slide for my last lecture, I could tell that people were genuinely excited and eager to hear what I had to say. It seems Michelle was not the only one with this question.

I’ve tried to dig up those lecture slides but for whatever reason I cannot find them. Most likely I deleted them out of disgust! The truth is, I was unable to articulate what exactly it was that I did when I analyzed data. I had honestly never thought about it before.

At Johns Hopkins University in the Department of Biostatistics, we don’t have a course specifically titled “Data Analysis”. Data analysis isn’t taught in courses. We teach it using a kind of apprenticeship model. As an advisor, I watch my students analyze data one at a time, tell them what they could improve, steer them when they go wrong, and congratulate them when I think they’ve done something well. It’s not an algorithmic process; I just do what I think is right.

Some of what I think is right comes from my training in statistics. There we learned about many of the important tools for data analysis; the very same tools I was teaching in my Methods in Biostatistics course when Michelle asked her question. The guiding principle behind most of graduate education in statistics goes roughly as follows:

> If I teach you everything there is to know about a tool, using mathematics, or simulation, or real data

> examples, then you will know when and where it is appropriate to use that tool.

We repeat this process for each tool in the toolbox. The problem is that the end of the statement doesn't follow from the beginning of the statement. Just because you know all of the characteristics of a hammer doesn't mean that you know how to build a house, or even a chair. You may be able to infer it, but that's perhaps the best we can hope for.

In graduate school, this kind of training is arguably okay because we know the students will go on to work with an advisor who will in fact teach them data analysis. But it still raises the question: Why can't we teach data analysis in the classroom? Why must it be taught one-on-one with an apprenticeship model? The urgency of these questions has grown substantially in recent times with the rise of big data, data science, and analytics. Everyone is analyzing data now, but we have very little guidance to give them. We still can't tell them "what to do".

One phenomenon that I've observed over time is that students, once they have taken our courses, are often very well-versed in data analytic tools and their characteristics. Theorems tell them that certain tools can be used in some situations but not in others. But when they are given a real scientific problem with real data, they often make what we might consider elementary mistakes. It's not for a lack of understanding of the tools. It's clearly *something else* that is missing in their training.

The goal of this book is to give at least a partial answer to Michelle's question of "What should I do?" Perhaps ironically, for a book about data analysis, much of the discussion will center around things that are *outside* the data. But ultimately, that is the part that is missing from traditional statistical training, the things outside the data that play a critical role in how we conduct and interpret data analyses. Three concepts will emerge from the discussion to follow. They are **context**, **resources**, and **audience**, each of which will get their own chapter. In addition, I will discuss an expanded picture of what data analysis is and how previous models have omitted key information, giving us only a partial view of the process.

Data analysis should be thought of as a separate field of study, but for too long it has been subsumed by either statistics or some other field. As a result, almost no time has been spent studying the process of data analysis. John Tukey, in an important (but somewhat rambling) article published in 1962 titled “The Future of Data Analysis”, argued similarly that data analysis should be thought of as separate from statistics. However, his argument was that data analysis is closer to a *scientific* field rather than a *mathematical* field, as it was being treated at the time.

In my opinion, Tukey was right to say that data analysis wasn’t like mathematics, but he was wrong to say that it was like science. Neither mathematics nor science provides a good model for how to think about data analysis, because data analysis is a process that is neither deductive or inductive. Rather, it is a process that solves a problem given the data available. It is the cycling back and forth of proposing solutions and asking further questions that is characteristic of the data analytic process and often results in the problem space expanding and contracting over time. There is tremendous ambiguity in most data analyses and accepting that ambiguity while also providing a useful result is one of the great challenges of data analysis.

Ultimately, what the data analyst is left with at the end is not truth (as with science) or logical certainty (as with mathematics), but a *solution*; a solution to a problem that incorporates the data we have collected. Whether we discover some fundamental law is not important to the data analyst (although it may be very important to the scientist!). This is not to say that data analysts should be disinterested in the topics on which they are working. It is just to say that their process leads to a different result than the scientists’ process. The data analyst’s process leads to something that other people can *use*.

One book will not answer the complex question that Michelle raised back in 2011, but if I had to answer her now, I might say, “Think of data analysis as a design process. You are designing a solution to a problem that has been handed to you with a dataset attached. This problem has many components, constraints, and possible solutions and you will need to put it all together into something that is coherent and useful.”

Here we go!