Last revised Oct 16, 2007

The Logic and Math of Causal Inquiry

A Guide for the Perplexed

Eyal Shahar

Preface (Do not skip, please)

Patricia O'Conner, my favorite maven of English grammar and writing, offers a piece of clever advice to authors who struggle with beginnings: start by summing it up. Which I will do. This book teaches you how to study causes through numbers, or more precisely, how to estimate the effect of a (presumed) causal variable from data. It is not a book about data analysis or statistical analysis even though both are amply used here. Nor is this a book about associations or correlations or predictions. CAUSES are the topic—plain causes free of semantic disguises behind which we like to hide them (risk factors, susceptibility factors, predictors, and the like.) That we cannot know a cause with certainty does not mean we should hide our ultimate interest in what it does.

You will find here widely used methods to study causal effects, explanations of their reasoning, and arguments against some prevailing practices. An outgrowth of a course, this book is a hybrid of theory and practice tied by the thread of real data, SAS code, and SAS printout.

What's in the book?

The book you are holding contains, among other topics, a discussion of causation and measures of effect, a harsh critique of P-values, a plea for estimation, and hand-held walk through contingency tables and regression models. This book covers most types of research material in epidemiology that came through my computer during the last 15 years but, of course, not every possible task is covered. You will also find here shameless uncertainties for numerous reasons: limits of methods, limits imposed by Nature, and limits of the author's mind and knowledge. Nobody is perfect.

This book does not shy away from challenging topics in philosophy of science and statistical thought even though I do not carry a membership card in either club. Degrees aside, early in my academic career, I sensed lifelessness of epidemiological practice without supplemental doses of philosophy of science and statistics—themselves linked disciplines. So I did a fair amount of reading and discovered a fascinating world behind the scientific method: zeal for truth and coherence, deductive logic, wishful induction, many faces of statistics, heated intellectual debates, arrogance, and even humbleness. At the end of some chapters, under the title "Suggested Reading", you will find reference to some of the readings that influenced me most—either because they were music to my ears or because they were not.

What background skills are assumed?

If you want to explore the beauty and frustration of scientific inquiry, you must meet the challenges of sound reasoning and quantitative work. In addition, I assume you have taken an introductory course in epidemiology or research methods, and you are familiar with the basic study designs (randomized trial, cohort study, case-control study, cross-sectional study). You should also have taken a course in statistics and have some understanding of the following terms: probability distribution, test statistic, *p*-value, confidence interval, and linear regression. Beyond basic algebra, exponentials, and logarithms, no high-level math is needed. Really!

Who may benefit from this book?

The typical reader may be a graduate student in epidemiology but students in other fields share the interest in cause-and-effect: medicine, environmental health, health services research, educational psychology, sociology—to name a few. Although each field has developed its own jargon and methods and explores unique topics, the conversion shouldn't be too difficult. For example, if your interest centers on the effect of parental education on average college grade, just replace my biomedical variables with the subject matter variables: *parental education* may replace *smoking status* and *average college grade* may replace *forced expiratory volume in one second*. In short, anyone who wants to dive into data with a causal inquiry in mind may find here practical and theoretical help—even experienced professionals.

How is this book organized?

I divided the book into two parts: The fist part (chapters 1 through 8) lays the theoretical foundation for the second, explaining ideas such as causal models, causal variables, measures of effect, causal parameters, effect modification, and three kinds of bias: confounding, selection, and information (chapters 1-7). Chapter 8 introduces key statistical ideas and explains why you won't find *p*-values or null hypothesis testing anywhere else. The second part of the book (chapters 9-21) contains both theory and practice: The theoretical side explains the rationale behind tabular methods and various regression models whereas the practical side teaches you how to estimate measures of effects (using SAS): the mean difference and the geometric mean ratio (chapters 9 through 11), various odds ratios (chapters 12 through 15), the proportion ratio (chapter 16), and the rate ratio (chapters 17 and 18). Chapter 19 addresses a common challenge to all those measures: how to estimate the effect of a continuous variable. Chapter 20 introduces a new and interesting method to handle confounders, and why I think it should be rejected from the realm of science. Miscellaneous topics are grouped in Chapter 21. My closing thoughts and various disclaimers are found in chapter 22.

Many textbooks are written to include self-contained topics; you may read selected chapters and fully follow the author's mind. This book, however, is telling an evolving story that starts on the first page and ends on the last page. If you try to read a chapter in isolation, you might raise an eyebrow occasionally ("Where is this coming from?"). So I really hope you will take the time to read the whole book. It's not Harry Potter but in my biased opinion it's human and interesting. Perhaps you can already sense that it's not another dry textbook by a faceless author.

Despite the breadth of topics, you are not expected to make chaotic jumps between chapters. Since good order is a prerequisite for the transfer of complex knowledge, I tried to build a logical structure. This is, perhaps, the biggest challenge for teachers of epidemiology and research methods because key concepts are intertwined in viciously circular ways. For example: to explain models of cause-and-effect, we would like to recruit the "probability ratio" as a teaching aid, but to explain "probability ratio" we need to understand models of cause-and-effect. An equally important challenge is following a famous quote of Albert Einstein: "Everything should be made as simple as possible but not simpler." If you find yourself lost in a chapter after reading it again and genuinely trying to follow, you may blame me. I hope you won't.

Why SAS?

It is possible to teach the content of this book without showing any analysis code or actual printout. I am convinced, however, that a live show is better than any substitute and that the concrete and real is grasped better than the general and the abstract. To look at a neatly transcribed printout is one thing; to look at an actual printout, even if censored, is another. SAS has been the software I have always used, but it shouldn't be too difficult to transport what you have learned here to another application.

What will you take home?

You should obtain enough information to test a causal theory in a dataset—one effect of one presumed cause—in the context of prior assumptions. By "enough information" I mean the mechanics of statistical models, interpretation of printout, and reasonable understanding of underlying theory and pitfalls. I hope that after reading this book you will become a worried scientist, a scientist who understands the insecure nature of scientific inquiry and the fallibility of human knowledge. That does not mean a paralyzed scientist fearing to draw inference from data, but one who recognizes that any inferential tower may collapse some day if a key assumption is successfully challenged.

You will take home, I hope, the following four messages. First, like all scientific knowledge, knowledge of causes remains conjectural forever, no matter how strongly we believe otherwise. Second, learning about causes should be equated with estimating the magnitude of their effects, not with declaring them causes. Third, every estimate of a causal effect embeds untestable assumptions; hence the first message. Four, there are better methods and poorer methods for causal inquiry: there are methods that rest on good reasoning, and there are methods that don't.

Keep in mind that good reasoning does not guarantee a hit on the truth, and that's true for every branch of science. We, scientists, constantly seek the Truth but never know whether we have reached it. If not our job, at least our profession is secured.