Introduction

Psychology (Statistics) 484

Statistics, Ethics, and the Social and Behavioral Sciences

June 13, 2013

Quotations About Modeling

Statisticians, like artists, have the bad habit of falling in love with their models.
— George Box

. . . the statistician knows . . . that in nature there never was a normal distribution, there never was a straight line, yet with normal and linear assumptions, known to be false, he can often derive results which match to a useful approximation, those found in the real world.
— George Box

It is a capital mistake to theorise before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.
— Sir Arthur Conan Doyle (*A Scandal in Bohemia*, 1892)

Today's scientists have substituted mathematics for experiment, and they wander off through equation after equation, and eventually build a structure which has no relation to reality.
— Nikola Tesla

Week 1: Introduction to (and Review of) the Course

Required Reading:
SGEP (vii–xiv; 1–15) —
Preface
Preamble
Introduction
The (Questionable) Use of Statistical Models

Popular Articles —
Why I Wrote "The Crucible," Arthur Miller (*New Yorker*, October 10, 1996)
The Trauma Trap, Frederick C. Crews (*New York Review of Books*, October 14, 2009)
Health Care: Who Knows 'Best'? Jerome Groopman (*New York Review of Books*, February 11, 2010)

Suggested Reading:
1.1 Preamble: Additional Suggested Reading

Film: *The McCarthy Years* (114 minutes)

A Statistical Guide For the Ethically Perplexed

Our title is taken from the seminal work of the medieval Jewish philosopher Moses Maimonides, *The Guide for the Perplexed*.

This monumental contribution was written as a three-volume letter to a student and was an attempt by Maimonides to reconcile his Aristotelian philosophical views with those of Jewish law.

In an analogous way, this book tries to reconcile the areas of statistics and the behavioral (and related social and biomedical) sciences through the standards for ethical practice, defined as being in accord with the accepted rules or standards for right conduct that govern a discipline.

Preamble: Defining the Term "Ethical"

Two major characterizations of the term "ethical" need to be distinguished in this Preamble, if only because this book focuses on just one of them.

The definition not used is where "ethical" pertains to principles of morality and what is right or wrong in conduct; thus, we speak of an "ethical (or moral) dilemma" as a situation involving an apparent conflict between moral imperatives, where obeying one would result in transgressing another.

This is more than we wish to undertake, or even be capable of, in a book devoted to statistical literacy as an assistance to ethical reasoning.

The meaning of "ethical" adopted here is one of being in accordance with the accepted rules or standards for right conduct that govern the practice of some profession.

The professions we have in mind are statistics and the behavioral sciences, and the standards for ethical practice are what we try to instill in our students through the methodology courses we offer, with particular emphasis on the graduate statistics sequence generally required for all the behavioral sciences.

Our hope is that the principal general education payoff for competent statistics instruction is an improvement in people's ability to be critical and ethical consumers and producers of the statistical reasoning and analyses faced in various applied contexts over the course of their careers.

Thus, this book is not as much about the good uses of statistics, but more about the specious applications when either statistical ideas are being applied unethically, or some quantitative insight might otherwise help prevent statistical "blunders" by the chronically careless.

It may not be unethical to be ignorant of the principles that guide a particular profession, but it is unethical to be ignorant and act as if one is not.

As an example, it is unethical to use some statistical program blindly to do something that you don't understand, and know that you don't, but then proceed to interpret the results as if you really did.

It is best to keep the adage in mind that if you don't know how to do something, then you don't know how to do it on a computer.

Once individuals pass through a competently taught behaviorally oriented statistics sequence, certain statistical literacy skills should then be available to help guide their ethical application.

For someone in a clinical psychology program, for example, and who should now know better, it would be unethical to use projective instruments, such as the Rorschach Test or Draw-A-Person Test, in clinical practice for diagnosis (because of the prevalence of illusory correlations and unproven validity; see Lilienfeld, Wood, & Garb, 2000);

it is unethical to prefer clinical judgment for diagnosis when standard mechanical methods for using intake information exist (for example, by discriminant analysis methods; see Dawes, 1979; 2002; the latter article is entitled "The Ethics of Using or Not Using Statistical Prediction Rules in Psychological Practice and Related Consulting Activities");

it is ethically questionable to attempt equating intact treatment groups using analysis of covariance (due to Lord's Paradox; see Miller and Chapman, 2001);

it is unethical to argue for clinical improvement based on regression toward the mean; and so on.

Introduction: Historical Background

Generations of graduate students in the behavioral and social sciences have completed mandatory year-long course sequences in statistics, sometimes with difficulty and possibly with less than positive regard for the content and how it was taught.

Prior to the 1960s, such a sequence usually emphasized a cookbook approach where formulas were applied unthinkingly using mechanically operated calculators.

The instructional method could be best characterized as "plug and chug," where there was no need to worry about the meaning of what one was doing, only that the numbers could be put in and an answer generated.

It was hoped that this process would lead to numbers that could then be looked up in tables; in turn, *p*-values were sought that were less than the iconic .05, giving some hope of getting an attendant paper published.

The situation began to change for the behavioral sciences in 1963 with the publication of *Statistics for Psychologists* by William Hays.

For the first time, graduate students could be provided both the needed recipes and some deeper understanding of and appreciation for the whole enterprise of inference in the face of uncertainty and fallibility.

The Hays text is now in its fifth edition, with a shortened title of *Statistics* (1994); the name of Hays itself stands as the eponym for what kind of methodology instruction might be required for graduate students; that is, at the level of Hays, and "cover to cover."

Although now augmented by other sources for related computational work (e.g., by SAS, SPSS, or SYSTAT), the Hays text remains a standard of clarity and completeness.

In teaching graduate statistics, there are multiple goals:

(1) to be capable of designing and analyzing one's own studies, including doing the computational "heavy lifting" oneself, and the ability to verify what others attached to a project may be doing;

(2) to understand and consume other research intelligently, both in one's own area, and more generally as a statistically and numerically literate citizen;

(3) to argue for and justify analyses when questioned by journal and grant reviewers or others, and to understand the basic justification for what was done.

Graduate instruction in statistics requires the presentation of general frameworks and how to reason from these.

These frameworks can be conceptual:

(a) the Fisherian view that provided the evidence of success in the Salk Polio vaccine trials where the physical act of randomization lead to credible causal inferences;

(b) to the unification given by the notion of maximum likelihood estimation and likelihood ratio tests both for our general statistical modeling as well as for more directed formal modeling in a behavioral science subdomain, such as image processing or cognitive neuroscience.

These frameworks can also be based on more quantitatively formal structures:

(a) the general linear model and its special cases of multiple regression, analysis of variance (ANOVA), and analysis of covariance (ANCOVA), along with model comparisons through full and reduced models;

(b) the general principles behind prediction/selection/correlation in simple two-variable systems, with extensions to multiple-variable contexts;

(c) the various dimensionality reduction techniques of principal components/factor analysis, multidimensional scaling, cluster analysis, and discriminant analysis.

The (Questionable) Use of Statistical Models

The form of statistical practice most commonly carried out by those with a mathematical bent (and in contrast to those more concerned with simple manifest forms of data analysis and visualization), is through the adoption of a stochastic model commonly containing (unobserved) latent variables.

Here, some data-generating mechanism is postulated, characterized by a collection of parameters and strong distributional assumptions (for example, conditional independence, normality, or homogeneous variability).

Based on a given dataset, the parameters are estimated, and usually, the goodness of fit of the model assessed by some statistic.

We might even go through a ritual of hoping for nonsignificance in testing a null hypothesis that the model is true, generally through some modified chi-squared statistic heavily dependent on sample size.

The cautionary comments of Roberts and Pashler (2000) should be kept in mind that the presence of a good fit does not imply a good or true model.

Moreover, models with many parameters are open to the problems engendered by overfitting and of a subsequent failure to cross-validate. We provide the abstract of the Roberts and Pashler (2000) article, "How Persuasive Is a Good Fit? A Comment on Theory Testing":

Quantitative theories with free parameters often gain credence when they closely fit data. This is a mistake. A good fit reveals nothing about the flexibility of the theory (how much it cannot fit), the variability of the data (how firmly the data rule out what the theory cannot fit), or the likelihood of other outcomes (perhaps the theory could have fit any plausible result), and a reader needs all three pieces of information to decide how much the fit should increase belief in the theory. The use of good fits as evidence is not supported by philosophers of science nor by the history of psychology; there seem to be no examples of a theory supported mainly by good fits that has led to demonstrable progress.

A better way to test a theory with free parameters is to determine how the theory constrains possible outcomes (i.e., what it predicts), assess how firmly actual outcomes agree with those constraints, and determine if plausible alternative outcomes would have been inconsistent with the theory, allowing for the variability of the data.

David Freedman Critique

A model-based approach is assiduously avoided throughout this book.

It seems ethically questionable to base interpretations about a given dataset and the story that the data may be telling, through a model that is inevitably incorrect.

As one highly cherished example in the behavioral sciences, it is now common practice to frame questions of causality through structural equation or path models, and to perform most data analysis tasks through the fitting of various highly parameterized latent variable models.

In a devastating critique of this practice, David Freedman in a *Journal of Educational Statistics* article, "As Others See Us: A Case Study in Path Analysis" (1987, *12*, 101–128), ends with this paragraph:

My opinion is that investigators need to think more about the underlying social processes, and look more closely at the data, without the distorting prism of conventional (and largely irrelevant) stochastic models. Estimating nonexistent parameters cannot be very fruitful. And it must be equally a waste of time to test theories on the basis of statistical hypotheses that are rooted neither in prior theory nor in fact, even if the algorithms are recited in every statistics text without caveat.

Another Example

A current discussion of yet another attempt to use statistical models to infer causality from observational data is in the article by Gina Kolata, "Catching Obesity From Friends May Not Be So Easy" (*New York Times*, August 8, 2011).

Kolata reviews the criticisms of causal inferences made using social network models, particularly from the 2009 book by Christakis and Fowler, *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives—How Your Friends' Friends' Friends Affect Everything You Feel, Think, and Do*.

Their basic argument is that homophily, the tendency of individuals to associate and bond with similar others, can somehow be separated from contagion, the spread of some societal ill, such as obesity, through an explicit social network.

The quantitative argument rests on the presence of an asymmetry in the relationship of who is a friend of whom, based solely on parameter estimates in a statistical model.

Supposedly, this implies causality in who affects who in the direct sense of a contagious transmission. Thus, contagion through a network is causal for items such as depression, happiness, illegal drug use, smoking, and loneliness.

This can all be gleaned directly from observational data through the intermediary of fitting a statistical model; moreover, contagion effects can be cleanly separated from homophily.

Freedman's quotation given for path models is just as germane for these network models.

Leo Breiman's Two Cultures

Leo Breiman took on the issue directly of relying on stochastic models (or, as he might have said, "hiding behind") in most of contemporary statistics.

What Breiman advocates is the adoption of optimization in place of parameter estimation, and of methods that fall under the larger rubric of supervised or unsupervised statistical learning theory.

Currently, this approach is best exemplified by the comprehensive text, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed., 2009) (T. Hastie, R. Tibshirani, & J. Friedman).

We give the abstract from Leo Breiman's *Statistical Science* article, "Statistical Modeling: The Two Cultures" (2001, *16*, 199–215):

There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex datasets and as a more accurate and informative alternative to data modeling on smaller datasets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Our View on Models

The view of statistics to be followed in this book is to consider what linear regression models can or cannot do, or the implications of a basic sampling model, but we go no further than least squares treated as an algorithmic optimization process, and a suggestion to adopt various sample reuse methods to gauge stability and assess cross-validation.

Remembering the definition of a *deus ex machina*—a plot device in Greek drama whereby a seemingly insoluble problem is suddenly and abruptly solved with the contrived and unexpected intervention of some new character or god—we will not invoke any statistical *deus ex machina* analogues.

An ancedote told in some of our beginning statistics sequences reflects this practice of postulating a *deus ex machina* to carry out statistical interpretations. Three academics—a philosopher, an engineer, and a statistician—are walking in the woods toward a rather large river that needs to be crossed. The pensive philosopher stops, and opines about whether they really need to cross the river; the engineer pays no attention to the philosopher and proceeds immediately to chop down all the trees in sight to build a raft; the statistician yells to the other two: "stop, assume a boat."

Stochastic data models do have a place but not when that is only as far as it goes.

When we work solely within the confines of a closed system given by the model, and base all inferences and conclusions under that rubric alone (for example, we claim a causal link because some path coefficient is positive and significant), the ethicality of such a practice is highly questionable.

George Box has famously said that "essentially, all models are wrong, but some are useful"; or Henri Theil's similar quip: "It does require maturity to realize that models are to be used, but not to be believed."

Box was referring to the adoption of a model heuristically to guide a process of fitting data; the point being that we only "tentatively entertain a model," with that model then subjected to diagnostic testing and reformulation, and so on iteratively.

The ultimate endpoint of such a process is to see how well the fitted model works, for example, on data collected in the future.

Once again, some type of (cross-)validation is essential, which should be the sine qua non of any statistical undertaking.