Introduction

This book is about statistical modeling. Some people define this term loosely as "fitting equations to data." That's close, but inexact. To do a bit better, let's take both words in turn.

Statistics is the study of variation among cases: economic growth rates, dinosaur skull volumes, student SAT scores, genes in a population, Congressional party affiliations, drug dosage levels, your choice of toothpaste versus mine . . . really any variable that can be measured.

A model is a metaphor, a description of a system that helps us to reason more clearly. Like all metaphors, models are approximations, and will never account for every last detail. In the words of the English statistician George Box: all models are wrong, but some are useful.

Aerospace engineers work with physical models—blueprints, simulations, mock-ups, wind-tunnel prototypes—to help them understand a proposed airplane design. Geneticists work with animal models—fruit flies, mice, zebrafish—to help them understand heredity. We will work with statistical models to help us understand variation.

Like the weather, most variation in the world exhibits some features that are predictable, and some that are unpredictable. Will it snow on Christmas day? It's more likely in Boston than Austin; that much we can anticipate. But even as late as Christmas eve, and even at the North Pole, nobody knows for sure.

The crucial thing about statistical models is that they describe both predictable and unpredictable variation. More than that, they allow us to partition observed variation into its predictable and unpredictable components—and not just in some loose allegorical way, but in a precise mathematical way that can, with perfect accuracy, be described as Pythagorean. (More on that later.)

This focus on the structured quantification of uncertainty is what distinguishes statistical modeling from ordinary evidence-based reasoning. It's important to know what the evidence says, goes this line of thinking. But it's also important to know what it

doesn't say. Sometimes that's the tricky part.

We will use statistical models for three purposes:

(1) *to explore* a large body of evidence, so that we might identify predictable features or trends amid random variation.

(2) *to test* our beliefs about cause-and-effect relationships among things in the world.

(3) *to predict* the future behavior of some system, and to say something useful about what remains unpredictable.

As Hippocrates enjoined doctors in the oath bearing his name: "Declare the past, diagnose the present, foretell the future."[1] These are the goals not merely of statistical modeling, but of the scientific method more generally.

[1] *Epidemics* Book I, section 11.

What modeling isn't. Many people assume that the job of a statistical modeler is to objectively summarize the facts, slap down a few error bars, and get out of the way. This view is mistaken. To be sure, statistical modeling demands a deep respect for facts, and for not allowing one's wishes or biases to change the story one tells *with* the facts. But the modeling process is inescapably subjective, in a way that should be embraced rather than ignored. Model-building requires not just technical knowledge of statistical ideas; it also requires care and judgment, and cannot be reduced to a flowchart, a table of formulas, or a tidy set of numerical summaries that wring every last drop of truth from a data set. There is almost never a single model that is obviously right. But there are definitely such things as good models and bad models, and learning to tell the difference is important. Just remember: calling a model good or bad requires knowing both the tool and the task. A shop-window mannequin is good for displaying clothes, but bad for training medical students about vascular anatomy.

Second, many people assume that statistical models must be complicated in order to do justice to the real world. Not always: complexity sometimes comes at the expense of explanatory power. We must avoid building models calibrated so tightly to past experience that they do not generalize to future cases. This idea—that theories should be made as complicated as they need to be, and no more so—is often called "Occam's Razor." A good model will be simple enough to understand and interpret, but not so simple that it does any major intellectual violence to the system being

modeled. All models of the world must balance these goals, and statistical models are no exception.

Finally, many people also assume that statistical modeling involves difficult, tedious mathematics. Happily, this isn't true at all. In fact, virtually all common statistical models are accessible to anyone with a high-school mathematics education, and these days all the tedious calculations are taken care of by computers. Modeling is pretty fun, once you get the hang of it.

Modeling then and now

On the time scale of important post-Enlightenment ideas, statistical modeling is middle-aged. An astronomer named Tobias Mayer was using something vaguely like linear modeling as early as 1750.[2] But most scholars credit two later mathematicians—Legendre, a Frenchmen; and Gauss, a German—with independently inventing the *method of least squares* some time between 1794 and 1805. That makes statistical modeling newer than the invention of calculus (credited jointly to Leibniz and Newton in the late 1600's), but older than the idea of evolution by natural selection (credited jointly to Darwin and Wallace over a period spanning the 1830's to the 1850's).

[2] Stephen M. Stigler, *The History of Statistics: The Measurement of Uncertainty before 1900*, pp. 16–25. Harvard University Press, 1986

For most of the nineteenth century, statistical modeling largely remained the concern of a highly specialized cadre of astronomers and geophysicists. But by our own age—one of fast, cheap computing and abundant data—it has become ubiquitous. In fact, the very same principle of least squares proposed by Legendre and Gauss remains, over two hundred years later, an important part of the day-to-day toolkit for solving problems in fields from aeronautics to zoology and everywhere in between.

But don't just take my word for it. The director of the White House Office of Management and Budget says so, too:

> The President has made it very clear that policy decisions should be driven by evidence—accentuating the role of Federal statistics as a resource for policymakers. Robust, unbiased data are the first step toward addressing our long-term economic needs and key policy priorities.[3]

[3] "Using Statistics to Drive Sound Policy." Office of Management and Budget Blog; May 8, 2009

So does the *Journal of the American Medical Association*, indirectly but pointedly, in its intimidating litany of statistical requirements:

> Numerical results should be accompanied by confidence intervals, if applicable, and exact levels of statistical significance.

> Evaluations of screening and diagnostic tests should include sensitivity, specificity, likelihood ratios, receiver operating characteristic curves, and predictive values.[4]

[4] *JAMA Instructions for Authors*, jama.ama-assn.org

Even the *New York Times* says so:

> For Today's Graduate, Just One Word: Statistics.[5]

[5] *New York Times* (Technology section); August 5, 2009

Of course, for all that, our political and cultural climate still exhibits a streak of distrust toward statistics. Why else would Churchill's brazen instructions to a young protégé sound so depressingly familiar?

> I gather, young man, that you wish to be a Member of Parliament. The first lesson that you must learn is that, when I call for statistics about the rate of infant mortality, what I want is proof that fewer babies died when I was Prime Minister than when anyone else was Prime Minister.[6]

[6] Quoted in *The Life of Politics* (1968), Henry Fairlie, Methuen, pp. 203–204

And why else would the famous remark, popularized by Twain and attributed to Disraeli, remain so apt, even a century later?

> Figures often beguile me, particularly when I have the arranging of them myself; in which case the remark attributed to Disraeli would often apply with justice and force: 'There are three kinds of lies: lies, damned lies, and statistics.'[7]

[7] *Chapters from My Autobiography*, North American Review (1907)

How do you tell the difference between "robust, unbiased evidence," misleading irrelevance, and cynical fraud? In considering this question, you will already have appreciated at least two good reasons to learn statistical modeling:

(1) To use data honestly and credibly in the service of an argument you believe in.

(2) To know how and when to be skeptical of someone else's damned lies.

For as John Adams put it,

> Facts are stubborn things; and whatever may be our wishes, our inclinations, or the dictates of our passion, they cannot alter the state of facts and evidence.[8]

[8] 'Argument in Defense of the Soldiers in the Boston Massacre Trials' (1770)