Large-Scale Inference:

Empirical Bayes Methods for
Estimation, Testing, and Prediction

Bradley Efron
Stanford University

Contents

Prologue

At the risk of drastic oversimplification, the history of statistics as a recognized discipline can be divided into three eras:

1 The age of Quetelet and his successors, in which huge census-level data sets were brought to bear on simple but important questions: Are there more male than female births? Is the rate of insanity rising?
2 The classical period of Pearson, Fisher, Neyman, Hotelling, and their successors, intellectual giants who developed a theory of optimal inference capable of wringing every drop of information out of a scientific experiment. The questions dealt with still tended to be simple — Is treatment A better than treatment B? — but the new methods were suited to the kinds of small data sets individual scientists might collect.
3 The era of scientific mass production, in which new technologies typified by the microarray allow a single team of scientists to produce data sets of a size Quetelet would envy. But now the flood of data is accompanied by a deluge of questions, perhaps thousands of estimates or hypothesis tests that the statistician is charged with answering together; not at all what the classical masters had in mind.

The response to this onslaught of data has been a tremendous burst of statistical methodology, impressively creative, showing an attractive ability to come to grips with changed circumstances, and at the same time highly speculative. There is plenty of methodology in what follows, but that is not the main theme of the book. My primary goal has been to ground the methodology in familiar principles of statistical inference.

This is where the "empirical Bayes" in my subtitle comes into consideration. By their nature, empirical Bayes arguments combine frequentist and Bayesian elements in analyzing problems of repeated structure. Repeated structures are just what scientific mass production excels at, e.g., expression levels comparing sick and healthy subjects for thousands of genes at the same time by means of microarrays. At their best, the new methodolo-

gies are successful from both Bayes and frequentist viewpoints, which is what my empirical Bayes arguments are intended to show.

False discovery rates, Benjamini and Hochberg's seminal contribution, is the great success story of the new methodology. Much of what follows is an attempt to explain that success in empirical Bayes terms. FDR, indeed, has strong credentials in both the Bayesian and frequentist camps, always a good sign that we are on the right track, as well as a suggestion of fruitful empirical Bayes explication.

The later chapters are at pains to show the limitations of current large-scale statistical practice: Which cases should be combined in a single analysis? How do we account for notions of relevance between cases? What is the correct null hypothesis? How do we handle correlations? Some helpful theory is provided in answer, but much of the argumentation is by example, with graphs and figures playing a major role. The examples are real ones, collected in a sometimes humbling decade of large-scale data analysis at the Stanford School of Medicine and Department of Statistics. (My examples here are mainly biomedical, but of course that has nothing to do with the basic ideas, which are presented with no prior medical or biological knowledge assumed.)

In moving beyond the confines of classical statistics, we are also moving outside its wall of protection. Fisher, Neyman et al. fashioned an almost perfect inferential machine for small-scale estimation and testing problems. It is hard to go wrong using maximum likelihood estimation or a t-test on a typical small data set. I have found it very easy to go wrong with huge data sets and thousands of questions to answer at once. Without claiming a cure, I hope the various examples at least help identify the symptoms.

The classical era of statistics can itself be divided into two periods: the first half of the 20th century, during which basic theory was developed, and then a great methodological expansion of that theory in the second half. Empirical Bayes stands as a striking exception. Emerging in the 1950s in two branches identified with Charles Stein and Herbert Robbins, it represented a genuinely new initiative in statistical theory. The Stein branch concerned normal estimation theory, while the Robbins branch was more general, being applicable to both estimation and hypothesis testing.

Typical large-scale applications have been more concerned with testing than estimation. If judged by chapter titles, the book seems to share this imbalance, but that is misleading. Empirical Bayes blurs the line between testing and estimation as well as between frequentism and Bayesianism. Much of what follows is an attempt to say how well we can estimate a testing procedure, for example how accurately can a null distribution be esti-

mated? The false discovery rate procedure itself strays far from the spirit of classical hypothesis testing, as discussed in Chapter 4.

About this book: it is written for readers with at least a second course in statistics as background. The mathematical level is not daunting — mainly multidimensional calculus, probability theory, and linear algebra — though certain parts are more intricate, particularly in Chapters 3 and 7 (which can be scanned or skipped at first reading). There are almost no asymptotics. Exercises are interspersed in the text as they arise (rather than being lumped together at the end of chapters), where they mostly take the place of statements like "It is easy to see ..." or "It can be shown ...". Citations are concentrated in the **Notes** section at the end of each chapter. There are two brief appendices, one listing basic facts about exponential families, the second concerning access to some of the programs and data sets featured in the text.

I have perhaps abused the "mono" in monograph by featuring methods from my own work of the past decade. This is not a survey or a textbook, though I hope it can be used for a graduate-level lecture course. In fact, I am not trying to sell any particular methodology, my main interest as stated above being how the methods mesh with basic statistical theory.

There are at least three excellent books for readers who wish to see different points of view. Working backwards in time, Dudoit and van der Laan's 2009 *Multiple Testing Procedures with Applications to Genomics* emphasizes the control of Type I error. It is a successor to *Resampling-based Multiple Testing: Examples and Methods for p-Value Adjustment* (Westfall and Young, 1993), which now looks far ahead of its time. Miller's classic text, *Simultaneous Statistical Inference* (1981), beautifully describes the development of multiple testing before the era of large-scale data sets, when "multiple" meant somewhere between two and ten problems, not thousands.

I chose the adjective *large-scale* to describe massive data analysis problems rather than "multiple," "high-dimensional," or "simultaneous," because of its bland neutrality with regard to estimation, testing, or prediction, as well as its lack of identification with specific methodologies. My intention is not to have the last word here, and in fact I hope for and expect a healthy development of new ideas in dealing with the burgeoning statistical problems of the 21st century.