Preface

THIS BOOK IS FOR THE CLASSROOM . . .

Reliable interpretation of genomics and genetic information has driven innovation in statistical methodology and its application to biological systems. Statisticians have met the need to test hundreds or thousands of hypotheses simultaneously with novel empirical Bayes methods that combine advantages of traditional Bayesian and frequentist statistics. In particular, techniques for estimating the local false discovery rate assign probabilities of differential gene expression, genetic association, etc. without requiring subjective prior distributions.

This book brings these methods to current and future scientists while keeping the mathematics at an elementary level. Readers will learn the fundamental concepts behind local false discovery rates, preparing them to analyze their own genomics data and to critically evaluate published genomics research. The book is suitable for researchers and for upper undergraduate students or beginning graduate students in the biological sciences. It gradually eases the readers into the mathematical equations needed. Exercises are provided at the end of each chapter to meet the needs of the classroom.

. . . AND ALSO FOR THE LAB

Researchers analyzing genomics data must use software that in most cases offers several choices of statistical methods designed for testing multiple hypotheses. Many of them are confused about

whether to use methods that correct p values or whether to use various methods of false discovery rates. It is difficult not only to choose the type of method for one's own research but also to evaluate the methods chosen by other researchers.

This book will address the problem by providing current and future researchers with a solid understanding of advantages and disadvantages of various methods of multiple hypothesis testing. In particular, the book will acquaint them with benefits of empirical Bayes methods of estimating local false discovery rates. Since those methods are not always available in software, they will also learn how to convert the output of most genomics hypothesis testing software to estimates of local false discovery rates.

To keep everything as simple and practical as possible, nuance and qualifications are routinely suppressed without denying their role in certain contexts. To the best of my knowledge, this is the only book on empirical Bayes methods and local false discovery rate estimation that does not require an advanced background in mathematics and statistics.

In addition, the book offers guidance through the minefield of current criticisms of p values. While much of the book concerns converting multiple p values to estimated local false discovery rates to bypass those criticisms, it also has a chapter on the calibration of a single p value. The calibration methods explained by the book aim at a balance between ignoring problems with significance testing at one extreme and rejecting all uses of the p value at the other extreme.

OVERVIEW

Special statistical methods are needed to analyze high-dimensional data from transcriptional microarrays and from other genome-scale measurement technology. Assuming readers have some knowledge of basic biological concepts, this book gradually introduces the mathematical concepts needed to apply statistical decision and inference theory to genomics data.

The statistical methods will be explained using the same microarray examples throughout the book. Due to such repeated

exposure, readers should become familiar with gene expression data, with the result that difficult statistical concepts can be understood without being obscured by unfamiliar biological concepts. The use of the same microarray data sets should also inculcate the awareness that the same data may be analyzed by more than one method.

Once the readers become familiar with the statistical methods through applications to microarray data, other data types will be covered. Related data types include genome-wide association data as well as quantitative proteomics and lipidomics data.

QUICK START FOR SELECTED TOPICS

Genetic Association

Example 1.4 has references to a website and to other parts of this book related to genetic association.

Local False Discovery Rate

Sections 3.2 and 3.4 provide introductions to local false discovery rates and their estimation. Section 5.3 states the reference class (reference set) problem. Section 8.2 explains a method of estimation called "Type II MLE" (Section 8.2.3); the prerequisites for that section are Sections 2.1 and 8.1.1.

DETAILED OUTLINE

Readers mastering the material of Chapter 1 ("Basic probability and statistics") will:

1. understand the concepts of gene expression, differential gene expression, probability mass functions, probability density functions, error models, probability distributions, normal distributions, tuples, random variables, hypothesis tests, null hypotheses, test statistics, significance levels, and p values;
2. be able to set up a contingency table from count data;

3. be able to compute a chi-square statistic on the basis of data from a contingency table;
4. understand how the above objectives are important for the analysis of gene expression data and genetic association data;
5. given genotype counts for cases and controls, be able to use the chi-square test to determine whether to reject the hypothesis of no genetic association at the 0.01 and 0.05 significance levels.

Readers mastering the material of Chapter 2 ("Introduction to likelihood") will:

1. understand the difference between a likelihood function and a probability function (§2.1);
2. be able to compute a likelihood ratio on the basis of prior probability, an error model, and data (§2.1);
3. know how to interpret a likelihood ratio as the strength of evidence for a hypothesis (§2.1);
4. understand the relationship between likelihood and the probability that a hypothesis is true (§2.3.1);
5. understand the relationship between probability and odds (§2.2);
6. be able to use likelihood to compute posterior odds on the basis of prior probability, an error model, and data (§2.3.1);
7. appreciate the importance of prior probability, especially with respect to its necessity in the interpretation of p values and likelihood ratios (§2.3.1).

Readers mastering the material of Chapter 3 ("False discovery rates") will:

1. understand the role of the likelihood ratio in defining the local false discovery rate (§3.2.2);

2. be able to estimate the nonlocal false discovery rate from a list of p values (§3.4);
3. understand the relationship between the LFDR and the nonlocal false discovery rate (§3.3);
4. understand the relationship between the LFDR and a posterior probability that a feature is affected (§3.2.1);
5. know where to find software to estimate the LFDR of each biological feature given a p value for each feature (§3.4).

Readers mastering the material of Chapter 4 ("Simulating and analyzing gene expression data") will:

1. understand the concept of simulating gene expression data and the relation between the posterior probability that a gene is differentially expressed and the local false discovery rate of that gene;
2. be able to estimate the posterior probability that a gene is differentially expressed given histograms of gene expression p values or fold change estimates;
3. be able to estimate the true mean expression level and its posterior distribution for each gene using a parametric empirical Bayes method (§4.3.3).

Readers mastering the material of Chapter 5 ("Variations in dimension and data") will:

1. know when to use a semi-parametric empirical Bayes method or a parametric empirical Bayes method in analyzing medium- to high-dimensional data;
2. be able to estimate the true mean expression level and its posterior distribution for each gene using the parametric empirical Bayes method of Section 4.3.3, even if some of the data are missing;

3. understand the concepts of a subclass and superclass and their relevance to the analysis of genomics data;
4. be able to determine whether to separate analysis of a p value subclass from analysis of a p value superclass for the purpose of estimating LFDRs (§5.3);
5. understand the biological meanings of *feature*, *null hypothesis*, and other terms needed to apply empirical Bayes methods of LFDR estimation to the analysis of genome-wide association data, QTL data, RT-PCR data, proteomics data, and metabolomics data.

Readers mastering the material of Chapter 6 ("Correcting bias in estimates of the false discovery rate") will:

1. understand why analyses of genomics data based on false discovery rates tend to be misleading;
2. be able to calibrate an estimated false discovery rate to correct that bias.

Readers mastering the material of Chapter 7 ("The $\mathcal{L}$ value: An estimated local false discovery rate to replace a p value") will:

1. know how to transform p values to estimates of the local false discovery rates of multiple hypotheses without relying on nonlocal false discovery rates;
2. know how to transform a single p value to an estimate of the posterior probability that the null hypothesis is true.

Readers mastering the material of Chapter 8 ("Maximum likelihood and applications") will:

1. understand the concept of maximum likelihood estimation;
2. understand the uses and limitations of p values and confidence intervals based on maximum likelihood estimation;

3. understand how maximum likelihood estimation can be used to estimate local false discovery rates;
4. understand the distinction between maximizing the likelihood function for an individual feature and maximizing the likelihood function for all features;
5. understand the analogy between maximum likelihood estimation and locating the highest point on a hill.

Finally, Appendix A extends a popular non-Bayesian method of multiple testing, and Appendix B provides concise and straightforward guidance to scientists.

FURTHER READING ETC.

A comprehensive review of empirical Bayes methods and false discovery rates would take us far beyond the scope of this book. For a more complete study of this area, I direct advanced readers to the works cited at the end of each chapter. The lists of references in the books cited give a much richer history of the domain.

Any supplementary material for this book will become available at http://davidbickel.com/genomics/ ("Statistical Inference to the Best Explanation of the Evidence"). Errata may also appear there, for I cannot control the book-wise error rate.

ACKNOWLEDGMENTS

The quotes that begin the chapters and the appendixes are from Doyle (1999). Chapter 6 develops material first released in Bickel (2016) and then in Bickel and Rahal (2019).

Now for the fun part, returning thanks to those who made the book possible. The efforts of David Grubbs, Karan Simpson, and Sherry Thomas at CRC Press made the publishing process early and efficient.

The University of Ottawa played several key roles. The book is much clearer as the result of feedback from multiple students. I thank the Department of Biochemistry, Microbiology, and Immunology and the Faculty of Medicine for granting a sabbatical to finish this project among others. The book was written in part at a writing workshop of the Centre for Academic Leadership that was organized by Françoise Moreau-Johnson.

Of course, the book is a product of my upbringing. My thrill for scientific debate comes from Bruce J. West, my PhD advisor. I deeply appreciate the sacrifices my parents made to provide for my education. The endless dice games with John and David and with my brothers Chris and Brain have left their mark.

Turning to my own family, I am grateful to my children for their encouragement to finish the book, and to the eldest for the beginning rows of Table 1.2. The flexibility and support of my wife helped me add new material just before the deadline. The poem in the dedication to her (Bickel, 2019a) is my translation of the *Prouerbia* 31:10–31 text found in Rahlfs and Hanhart (2006).

David R. Bickel
Ottawa
July 2019