Preface

Current statistical inference problems in areas such as astronomy, genomics, and marketing routinely involve the simultaneous test of thousands, or even millions, of null hypotheses. These hypotheses concern a wide range of parameters, for high-dimensional multivariate distributions, with complex and unknown dependence structures among variables.

Motivated by these applications and the limitations of existing multiple testing methods, we have developed and implemented resampling-based single-step and stepwise multiple testing procedures (MTP) for controlling a broad class of Type I error rates, defined as tail probabilities and expected values for arbitrary functions $g(V_n, R_n)$ of the numbers of Type I errors V_n and rejected hypotheses R_n (Birkner et al., 2005a,b,c, 2006, 2007; Dudoit et al., 2004a,b, 2006; Keleş et al., 2006; van der Laan et al., 2004a,b, 2005; van der Laan and Hubbard, 2006; Pollard et al., 2005a,b; Pollard and van der Laan, 2004; Rubin et al., 2006). Our proposed procedures take into account the joint distribution of the test statistics and provide Type I error control in testing problems involving general data generating distributions (with arbitrary dependence structures among variables), null hypotheses (defined in terms of submodels for the data generating distribution), and test statistics (e.g., t-statistics, χ^2-statistics, F-statistics). A key ingredient of the procedures is the null distribution used in place of the unknown joint distribution of the test statistics. The results of a given MTP are reported in terms of rejection regions (i.e., cut-offs) for the test statistics, confidence regions for the parameters of interest, and adjusted p-values.

This book provides a detailed account of the theoretical foundations of our multiple testing methodology and discusses its software implementation in R (multtest package; Gentleman et al. (2004); Pollard et al. (2005b); R Development Core Team (2006); `www.bioconductor.org`; `www.r-project.org`) and SAS (`www.sas.com`). The proposed methods are applied to a range of testing problems in biomedical and genomic research, including: the identification of differentially expressed and co-expressed genes in high-throughput gene expression experiments, such as microarray experiments; tests of association be-

tween gene expression measures and biological annotation metadata, such as Gene Ontology (GO, `www.geneontology.org`) annotation; protein sequence analysis; and the genetic mapping of complex traits using single nucleotide polymorphisms (SNP).

Intended readership

Methodological Chapters 4–7 are intended for readers with advanced undergraduate or graduate statistical training, whereas introductory Chapters 1–3 and applications Chapters 8–13 are also aimed at readers with biological background.

Some of the material discussed in this book was taught in the Division of Biostatistics at the University of California, Berkeley: upper division undergraduate course *Introduction to Statistical Methods in Computational and Genomic Biology* (PB HLTH 143); MA/PhD graduate course *Biostatistical Methods: Applications of Statistics to Genetics and Molecular Biology* (PB HLTH 240D); and MA/PhD graduate course *Multiple Testing and Loss Function Based Estimation: Applications in Biological Sciences* (PB HLTH 246C).

Overview

Chapter 1 introduces a general statistical framework for *multiple hypothesis testing* and discusses in turn the main ingredients of a multiple testing problem, including: the data generating distribution; the parameters of interest; the null and alternative hypotheses; the test statistics; multiple testing procedures; rejection regions for the test statistics; errors in multiple hypothesis testing: Type I, Type II, and Type III errors; Type I error rates; power; unadjusted and adjusted p-values; and stepwise multiple testing procedures.

Chapter 2 concerns a key feature of our proposed multiple testing methodology: the *test statistics null distribution* used to obtain rejection regions for the test statistics, confidence regions for the parameters of interest, and adjusted p-values. Indeed, whether testing single or multiple hypotheses, one needs the (joint) distribution of the test statistics in order to derive a procedure that probabilistically controls Type I errors. In practice, however, the true distribution of the test statistics is unknown and replaced by a null distribution. The choice of a proper null distribution is crucial in order to ensure that (finite sample or asymptotic) control of the Type I error rate under the assumed null distribution does indeed provide the desired control under the true distribution. This issue is particularly relevant for large-scale testing problems, such as those described above in biomedical and genomic research, which concern high-dimensional multivariate distributions, with complex and unknown dependence structures among variables.

Common approaches use a data generating distribution, such as a permutation distribution, that satisfies the complete null hypothesis that all null hypotheses are true. Procedures based on such a data generating null distribution typically rely on the subset pivotality assumption, stated in Westfall and Young (1993, p. 42–43), to ensure that Type I error control under the data generating null distribution leads to the desired control under the true data generating distribution. However, subset pivotality is violated in many important testing problems, because a data generating null distribution may result in a joint distribution for the test statistics that has a different dependence structure than their true distribution. In fact, in most problems, there does not exist a data generating null distribution that correctly specifies the joint distribution of the test statistics corresponding to the true null hypotheses. Indeed, subset pivotality fails for two types of testing problems that are highly relevant in biomedical and genomic data analysis: tests concerning correlation coefficients and tests concerning regression coefficients (Chapter 8; Pollard et al. (2005a); Pollard and van der Laan (2004)).

To address the shortcomings of existing approaches, we have formulated a general characterization of a test statistics null distribution for which the multiple testing procedures of Chapters 3–7 provide proper Type I error control (Section 2.2). Our general characterization is based on the intuitive notion of *null domination*, whereby the number of Type I errors is stochastically greater under the test statistics' null distribution than under their true distribution. Null domination conditions lead to the explicit construction of two main types of test statistics null distributions. The first original proposal of Dudoit et al. (2004b), van der Laan et al. (2004a), and Pollard and van der Laan (2004), defines the null distribution as the asymptotic distribution of a vector of *null shift and scale-transformed test statistics*, based on user-supplied upper bounds for the means and variances of the test statistics for the true null hypotheses (Section 2.3). The second and most recent proposal of van der Laan and Hubbard (2006) defines the null distribution as the asymptotic distribution of a vector of *null quantile-transformed test statistics*, based on user-supplied marginal test statistics null distributions (Section 2.4).

Either test statistics null distribution (or consistent estimators thereof) may be used in any of the multiple testing procedures proposed in Chapters 3–7, as they both satisfy the key property of joint null domination for the test statistics corresponding to the true null hypotheses. The latest proposal of van der Laan and Hubbard (2006) has the additional advantage that the marginal test statistics null distributions may be set to the optimal (i.e., most powerful) null distributions one would use in single hypothesis testing (e.g., permutation marginal null distributions, Gaussian or other parametric marginal null distributions). Resampling procedures (e.g., non-parametric or model-based bootstrap) are provided to conveniently obtain consistent estimators of the null distribution and of the corresponding test statistic cut-offs, parameter confidence regions, and adjusted p-values.

We stress the generality of our proposed test statistics null distributions: Type I error control does not rely on restrictive assumptions such as subset pivotality and holds for general data generating distributions, null hypotheses, and test statistics. In particular, the proposed null distributions allow one to address testing problems that cannot be handled by existing approaches, such as tests concerning correlation coefficients and parameters in general regression models (e.g., linear regression models where the covariates and error terms are allowed to be dependent, logistic regression models, Cox proportional hazards models).

As detailed in Section 2.8, the following two important points distinguish our approach from existing approaches to Type I error control and the choice of a null distribution. Firstly, we are only concerned with Type I error control under the true data generating distribution. The notions of weak and strong control (and associated subset pivotality) are therefore irrelevant for our methods. Secondly, we propose a null distribution for the test statistics, and not a data generating null distribution. The latter practice does not necessarily provide proper Type I error control, as a data generating null distribution may result in a joint distribution for the test statistics that has a different dependence structure than their true distribution.

The simulation studies of van der Laan and Hubbard (2006), Pollard et al. (2005a), and Pollard and van der Laan (2004), demonstrate that the choice of null distribution can have a substantial impact on the Type I error and power properties of a given multiple testing procedure (Chapter 8). In particular, Pollard et al. (2005a) show that procedures based on our general non-parametric bootstrap null shift and scale-transformed test statistics null distribution typically control the Type I error rate "on target" at the nominal level. In contrast, comparable procedures, based on parameter-specific bootstrap data generating null distributions, can be severely anti-conservative (bootstrapping residuals for testing regression coefficients) or conservative (independent bootstrap for testing correlation coefficients). van der Laan and Hubbard (2006) further illustrate that, for finite samples, the new null quantile-transformed test statistics null distribution provides more accurate Type I error control and is more powerful than the original null shift and scale-transformed null distribution.

Chapter 3 presents an overview of basic multiple testing procedures for controlling the number of Type I errors (family-wise error rate and generalized family-wise error rate, in Sections 3.2 and 3.3, respectively) and the proportion of Type I errors among the rejected hypotheses (false discovery rate and tail probabilities for the proportion of false positives, in Sections 3.4 and 3.5, respectively). The different procedures are stated in terms of adjusted p-values as well as cut-offs for individual test statistics or unadjusted p-values. Summary tables are provided in Appendix A.

Chapter 4 proposes general *joint single-step common-cut-off* and *common-quantile procedures* for controlling Type I error rates $\Theta(F_{V_n})$, defined as arbitrary parameters of the distribution of the number of Type I errors V_n (Section 4.2; Dudoit et al. (2004b); Pollard and van der Laan (2004)). Such error rates include the generalized family-wise error rate (gFWER), $gFWER(k) = 1 - F_{V_n}(k) = \Pr(V_n > k)$, i.e., the chance of at least $(k+1)$ Type I errors, and, in particular, the usual family-wise error rate (FWER), $FWER = gFWER(0) = 1 - F_{V_n}(0) = \Pr(V_n > 0)$. In the special case of $gFWER(k)$ control, the procedures are based on the $(k+1)$st largest test statistic and $(k+1)$st smallest unadjusted p-value, respectively. For control of the FWER, the procedures reduce to the single-step maxT and minP procedures, based on the maximum test statistic and minimum unadjusted p-value, respectively. Adjusted p-values are derived in Section 4.3. Single-step common-cut-off and common-quantile procedures, based on consistent estimators of the test statistics null distribution, are shown to provide asymptotic control of the Type I error rate $\Theta(F_{V_n})$. General bootstrap procedures are supplied to conveniently obtain consistent estimators of the single-step common cut-offs and common-quantile cut-offs and of the corresponding adjusted p-values (Section 4.4). This chapter also establishes equivalence results between Θ-specific single-step multiple testing procedures and parameter confidence regions (Section 4.6) and addresses the issue of test optimality, i.e., the maximization of power subject to a Type I error constraint (Section 4.7; Rubin et al. (2006)).

Chapter 5 focuses on control of the family-wise error rate, $FWER = 1 - F_{V_n}(0)$, and provides *joint step-down common-cut-off maxT* and *common-quantile minP procedures*, based on maxima of test statistics and minima of unadjusted p-values, respectively (Sections 5.2 and 5.3; van der Laan et al. (2004a)). Two main types of results are derived concerning asymptotic control of the FWER. The more general theorems prove that the step-down maxT and minP procedures provide asymptotic control of the FWER, under general asymptotic null domination assumptions for the test statistics null distribution. Exact asymptotic control results are obtained by making additional asymptotic separation assumptions for the test statistics for the true and false null hypotheses. Step-up procedures are discussed in Section 5.4. Step-down maxT and minP procedures, based on consistent estimators of the test statistics null distribution, are shown to provide asymptotic control of the FWER. General bootstrap procedures are supplied to conveniently obtain consistent estimators of the step-down maxT and minP cut-offs and of the corresponding adjusted p-values (Section 5.5).

Chapter 6 proposes a new general and flexible approach to multiple hypothesis testing, the augmentation method, whereby a set of suitably chosen null hypotheses are added to the set of hypotheses already rejected by an initial MTP, in order to control a second target Type I

error rate (Dudoit et al., 2004a; van der Laan et al., 2004b). Specifically, given an initial gFWER-controlling procedure, this chapter provides (marginal/joint single-step/stepwise) *augmentation multiple testing procedures* (AMTP) for controlling generalized tail probability (gTP) error rates, $gTP(q,g) = \Pr(g(V_n, R_n) > q)$, for arbitrary functions $g(V_n, R_n)$ of the numbers of false positives V_n and rejected hypotheses R_n (Section 6.5). Simple augmentations of FWER-controlling procedures are treated in detail, for controlling tail probabilities for the number of false positives (gFWER), with $g(v,r) = v$, and tail probabilities for the proportion of false positives (TPPFP) among the rejected hypotheses, with $g(v,r) = v/r$ (Sections 6.2 and 6.5.3 for gFWER; Sections 6.3 and 6.5.4 for TPPFP). As shown in Section 6.5.2, the adjusted p-values for an augmentation multiple testing procedure are simply shifted versions of the ordered adjusted p-values for the initial MTP. Section 6.6 demonstrates that one can readily derive (conservative) procedures controlling generalized expected value (gEV) error rates, $gEV(g) = \mathrm{E}[g(V_n, R_n)]$, based on procedures controlling generalized tail probability error rates, $gTP(q,g) = \Pr(g(V_n, R_n) > q)$. Control of the false discovery rate (FDR), based on a TPPFP-controlling MTP, corresponds to the special case $g(v,r) = v/r$ (Section 6.4).

We stress the generality and important practical implications of the augmentation approach to multiple testing: any gFWER-controlling MTP immediately and trivially provides multiple testing procedures that control a wide variety of error rates, defined as tail probabilities $\Pr(g(V_n, R_n) > q)$ for arbitrary functions $g(V_n, R_n)$ of the numbers of false positives V_n and rejected hypotheses R_n. While existing approaches for controlling the proportion of false positives (e.g., TPPFP and FDR) typically assume either independence or specific dependence structures for the joint distribution of the test statistics, augmentation procedures can be derived for general data generating distributions (i.e., arbitrary joint distributions for the test statistics), null hypotheses, and test statistics. One can therefore build on the large pool of available FWER-controlling procedures to greatly expand the class of Type I error rates one can control (e.g., single-step and step-down maxT and minP procedures, summarized in overview Chapter 3 and discussed in detail in Chapters 4 and 5).

Chapter 7 builds on van der Laan et al. (2005) and proposes new *joint resampling-based empirical Bayes procedures* for controlling generalized tail probability error rates, $gTP(q,g) = \Pr(g(V_n, R_n) > q)$. The approach involves specifying: (i) a null distribution for vectors of null test statistics and (ii) a distribution for random guessed sets of true null hypotheses. By randomly sampling null test statistics and guessed sets of true null hypotheses, one obtains a distribution for a guessed g-specific function of the numbers of false positives and rejected hypotheses, for any given vector of cut-offs for the test statistics. Cut-offs can then be chosen to control tail probabilities for this distribution at a user-supplied level. This chapter also discusses empirical Bayes

q-value-based approaches to FDR control and connections to the frequentist step-up Benjamini and Hochberg (1995) procedure.

Chapter 8 presents simulation studies assessing the performance of the multiple testing procedures described in Chapters 1–7. The simulation studies focus on the choice of a test statistics null distribution in testing problems concerning correlation coefficients and regression coefficients in models where the covariates and error terms are allowed to be dependent (Pollard et al., 2005a).

Chapters 9–12 apply the proposed methodology to the following multiple testing problems in biomedical and genomic research: the identification of differentially expressed and co-expressed genes in high-throughput gene expression experiments (Chapter 9); tests of association between gene expression measures and biological annotation metadata, e.g., Gene Ontology annotation (Chapter 10); the identification of HIV-1 codon positions associated with viral replication capacity (Chapter 11); the genetic mapping of human obesity, based on tests of association between multilocus composite SNP genotypes and obesity-related phenotypes (Chapter 12).

The above testing problems share the following general characteristics: inference for high-dimensional multivariate distributions, with complex and unknown dependence structures among variables; broad range of parameters of interest, such as coefficients in general regression models relating possibly censored biological and clinical covariates and outcomes to genome-wide expression measures and genotypes; many null hypotheses, in the thousands or even millions; complex and unknown dependence structures among test statistics (e.g., directed acyclic graph (DAG) structure of GO terms in Chapter 10, Galois lattice for multilocus composite SNP genotypes in Chapter 12).

Due to their generality and flexibility, the multiple testing procedures of Chapters 1–7 are well-suited to address these and other high-dimensional testing problems arising in different areas of application of statistics. In particular, recall that the proposed procedures are designed to control a broad range of Type I error rates, for: general multivariate data generating distributions, with arbitrary dependence structures among variables; general null hypotheses, defined in terms of submodels for the data generating distribution; general test statistics, such as, t-statistics for tests of means, correlation coefficients, and coefficients in general regression models, and F-statistics for testing multiple-parameter null hypotheses.

Chapter 13 discusses the software implementation of the proposed multiple testing procedures in the R package multtest, released as part of the Bioconductor Project, an open-source software project for the analysis of biomedical and genomic data (Section 13.1; Gentleman et al. (2004); Pollard et al. (2005b); R Development Core Team (2006); `www.bioconductor.org`; `www.r-project.org`). This chapter also illustrates the implementation in SAS of a bootstrap-based single-step maxT procedure and gFWER- and

TPPFP-controlling augmentation multiple testing procedures (Section 13.2; Birkner et al. (2005b); SAS, Version 9, `www.sas.com`).

Appendix A contains summaries of basic definitions, notation, and multiple testing procedures.

Appendix B provides miscellaneous mathematical and statistical results used repeatedly throughout the book.

Appendix C supplies SAS code for some of the proposed multiple testing procedures.

Supplements

The book's website provides supplementary materials, such as, additional analyses, tables, and figures, articles, lecture notes, software, datasets, links, and errata (`www.stat.berkeley.edu/~sandrine/MTBook`; `www.springer-ny.com`).

The reader is referred to the National Center for Biotechnology Information (NCBI) website for online tutorials and other educational resources on genome biology (`www.ncbi.nlm.nih.gov/Education`). The supplements to *Nature Genetics* provide an overview of the biology, technology, and applications of microarray experiments (Phimister and Cohen (1999); Packer (2002); Packer and Axton (2005); `www.nature.com/ng/supplements`). The book edited by Speed (2003) discusses statistical methods for the analysis of microarray data.

Software packages (e.g., R package multtest), datasets, short course materials (e.g., lecture notes, computer labs), and documentation may be downloaded from the Bioconductor Project (Gentleman et al. (2004); `www.bioconductor.org`) and R Project (R Development Core Team (2006); `www.r-project.org`) websites. The monograph edited by Gentleman et al. (2005a) provides a survey of Bioconductor software packages and their applications to a range of problems in computational biology (`www.bioconductor.org/pub/docs/mogr`).

Technical reports are available from the *UC Berkeley Division of Biostatistics Working Paper Series* website (`www.bepress.com/ucbbiostat`).

Finally, our personal websites provide additional resources on multiple hypothesis testing (SD: `www.stat.berkeley.edu/~sandrine`; MJvdL: `www.stat.berkeley.edu/~laan`).

Acknowledgments

We warmly thank our former students and colleagues, Merrill D. Birkner, Sündüz Keleş, and Katherine S. Pollard, for most pleasant collaborations and their constructive comments on the following portions of the book: Chapters

8, 11, and 12, Section 13.2 (MDB); Chapter 10 (SK); Sections 9.3 and 13.1 (KSP).

We would also like to acknowledge the following colleagues and students for many stimulating discussions on multiple hypothesis testing: Simon E. Cawley, Yongchao Ge, Robert C. Gentleman, Torsten Hothorn, Jason C. Hsu, Alan E. Hubbard, Nicholas P. Jewell, Daniel Rubin, Juliet P. Shaffer, Phil Spector, and Terence P. Speed.

Much of our methodological work is motivated by statistical problems in biomedical and genomic research. We are most grateful to our collaborators in biology, computer science, and epidemiology for introducing us to fascinating new questions and for inspiring our research on multiple hypothesis testing: Alain Barrier, Jennifer C. Boldrick, Patrick O. Brown, Patricia A. Buffler, Matthew J. Callow, Karine Clément, Mélanie Courtine, Martyn T. Smith, and Jean-Daniel Zucker.

Last, but not least, we would like to thank students at UC Berkeley (PB HLTH 143, Spring 2004; PB HLTH 240D, Spring 2003 and 2005; PB HLTH 246C (formerly PB HLTH 243A), Fall 2003 and 2005) and participants in Bioconductor workshops for their valuable feedback on multiple testing methods and their software implementation.

Berkeley, CA
May 2007

Sandrine Dudoit
Mark J. van der Laan