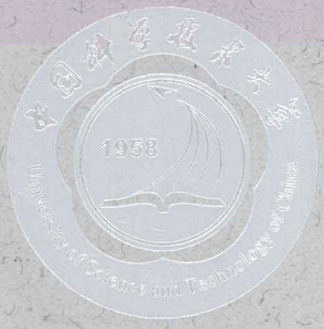
Frontiers of Biostatistics and Bioinformatics



Shuangge Ma
Yuedong Wang

中国科学技术大学出版社

当代科学技术基础理论与前沿问题研究丛书

Frontiers of Biostatistics and Bioinformatics

Shuangge Ma
Yuedong Wang

中国科学技术大学出版社

内 容 简 介

作为现代科学的一个重要分支，生物统计和生物信息学运用统计、数学和计算机科学的方法研究生物及医学问题。在本书中，我们收录了15篇文章介绍生物统计和生物信息学中统计方法的最新进展。由于篇幅有限，我们着重介绍了高维数据分析方法及其应用。和统计学的其他分支相比，高维数据分析方法及其应用还处在相对初级阶段，我们希望本书的出版能够推动此方向的研究。

Frontiers of Biostatistics and Bioinformatics
Shuangge Ma & Yuedong Wang

Published by University of Science and Technology of China Press
96 Jinzhai Road, Hefei, Anhui, P. R. China

图书在版编目(CIP)数据

生物统计及生物信息学前沿＝Frontiers of Biostatistics and Bioinformatics：英文/马双鸽，王跃东主编. —合肥：中国科学技术大学出版社，2009.2
（当代科学技术基础理论与前沿问题研究丛书；中国科学技术大学校友文库）
“十一五”国家重点图书
ISBN 978-7-312-02228-9

Ⅰ.生… Ⅱ.①马… ②王… Ⅲ.①生物统计—英文 ②生物信息论—英文
Ⅳ.Q-332 Q811.4

中国版本图书馆CIP数据核字(2009)第008246号

中国科学技术大学出版社出版发行
安徽省合肥市金寨路96号，230026
http://press.ustc.edu.cn
合肥晓星印刷有限责任公司印刷
全国新华书店经销
开本：710 mm×1000 mm 1/16 印张：18.5 字数：310千
2009年2月第1版 2009年2月第1次印刷
印数：1—1500册
定价：58.00元

总　　序

侯建国
(中国科学技术大学校长、中国科学院院士、第三世界科学院院士)

大学最重要的功能是向社会输送人才. 大学对于一个国家、民族乃至世界的重要性和贡献度，很大程度上是通过毕业生在社会各领域所取得的成就来体现的.

中国科学技术大学建校只有短短的五十年，之所以迅速成为享有较高国际声誉的著名大学之一，主要就是因为她培养出了一大批德才兼备的优秀毕业生. 他们志向高远、基础扎实、综合素质高、创新能力强，在国内外科技、经济、教育等领域做出了杰出的贡献，为中国科大赢得了“科技英才的摇篮”的美誉.

2008 年 9 月，胡锦涛总书记为中国科大建校五十周年发来贺信，信中称赞说：半个世纪以来，中国科学技术大学依托中国科学院，按照全院办校、所系结合的方针，弘扬红专并进、理实交融的校风，努力推进教学和科研工作的改革创新，为党和国家培养了一大批科技人才，取得了一系列具有世界先进水平的原创性科技成果，为推动我国科教事业发展和社会主义现代化建设做出了重要贡献.

据统计，中国科大迄今已毕业的 5 万人中，已有 42 人当选中国科学院和中国工程院院士，是同期（自 1963 年以来）毕业生中当选院士数最多的高校之一. 其中，本科毕业生中平均每 1000 人就产生 1 名院士和七百多名硕士、博士，比例位居全国高校之首. 还有众多的中青年才俊成为我国科技、企业、教育等领域的领军人物和骨干. 在历年评选的“中国青年五四奖章”获得者中，作为科技界、科技创新型企业界青年才俊代表，科大毕业生已连续多年榜上有名，获奖总人数位居全国高校前列. 鲜为人知的是，有数千名优秀毕业生踏上国防战线，为科技强军做出了重要贡献，涌现出二十多名科技将军和一大批国防科技中坚.

为反映中国科大五十年来人才培养成果，展示毕业生在科学研究中的最新进展，学校决定在建校五十周年之际，编辑出版《中国科学技术大学校友文库》，于2008 年 9 月起陆续出书，校庆年内集中出版 50 种. 该《文库》选题经过多轮严格的评审和论证，入选书稿学术水平高，已列为“十一五”国家重点图书出版规划.

入选作者中，有北京初创时期的毕业生，也有意气风发的少年班毕业生；有“两院”院士，也有 IEEE Fellow；有海内外科研院所、大专院校的教授，也有金融、IT 行业的英才；有默默奉献、矢志报国的科技将军，也有在国际前沿奋力拼搏的科研将才；有“文革”后留美学者中第一位担任美国大学系主任的青年教授，也有首批获得新中国博士学位的中年学者 …… 在母校五十周年华诞之际，他们通过著书立说的独特方式，向母校献礼，其深情厚意，令人感佩!

近年来，学校组织了一系列关于中国科大办学成就、经验、理念和优良传统的总结与讨论. 通过总结与讨论，我们更清醒地认识到，中国科大这所新中国亲手创办的新型理工科大学所肩负的历史使命和责任. 我想，中国科大的创办与发展，首要的目标就是围绕国家战略需求，培养造就世界一流科学家和科技领军人才. 五十年来，我们一直遵循这一目标定位，有效地探索了科教紧密结合、培养创新人才的成功之路，取得了令人瞩目的成就，也受到社会各界的广泛赞誉.

成绩属于过去，辉煌须待开创. 在未来的发展中，我们依然要牢牢把握“育人是大学第一要务”的宗旨，在坚守优良传统的基础上，不断改革创新，提高教育教学质量，早日实现胡锦涛总书记对中国科大的期待：瞄准世界科技前沿，服务国家发展战略，创造性地做好教学和科研工作，努力办成世界一流的研究型大学，培养造就更多更好的创新人才，为夺取全面建设小康社会新胜利、开创中国特色社会主义事业新局面贡献更大力量.

是为序.

2008 年 9 月

From the Editors

Biostatistics and bioinformatics apply techniques in mathematics, informatics, statistics, computer science and artificial intelligence to solve biological problems. Recent advancements in technology make it possible to gather data on the molecular level, which results in data with extremely high dimensionality. There is thus an urgent need for new and powerful statistical methods for analyzing such data. Although considerable progress has been made, high dimensional analysis still remains a young filed with many theoretical and practical problems yet to be addressed.

The primary goal of this volume is to create a platform for researchers to present the most recent developments in statistical methodologies for high dimensional data. We hope it can provide a window into the state-of-art research in high dimensional analysis, motivate interested researchers, and foster more research in this area. This volume does not, nor is it intended to, present a full picture of the vast area of high dimensional data analysis. Topics were selected according to personal research interests only.

Fifteen papers of extremely high quality were peer reviewed by experts and selected. They are loosely divided into four topics: (i) statistical genetics with the goal to identify quantitative trait loci associated with phenotype variations; (ii) statistical analysis of microarray data with the goal to identify differentially expressed genes under different disease stages or experimental conditions based on microarray experiments; (iii) computational biology with the goal to construct mathematical and statistical models that can best describe the molecular structure of cells; and (iv) general methodology.

As proud graduates of USTC, we are extremely honored to have this opportunity to edit this special volume as part of celebration to her 50th birthday. We are deeply exhilarated by positive and enthusiastic responses from such a successful group of researchers who have made significant contributions in various areas. All papers have at least one author who is an alumnus of USTC. We owe special thanks to the USTC press. Without their encouragements and support, this volume would not be possible. Finally, we would like to express our appreciation to the following reviewers, Jinbo Chen (University of Pennsylvania), Ruzong Fan (Texas A & M University), Anna Liu (University of Massachusetts), Xiao Song (University of Georgia), Huiliang Xie (University of Miami), Tiejun Tong (University of Colorado, Boulder), Momiao Xiong

(University of Texas, Houston), Hongyu Zhao (Yale University), Chuan Zhou (Vanderbilt University) and Zhu Ji (University of Michigan). Their insightful comments substantially improved quality of this volume.

Shuangge Ma & Yuedong Wang
Yale University and
University of California-Santa Barbara

Preface

It would not be a stretch to say that modern statistics is enjoying a youthful bloom in greater China. The Chinese have long been renowned for fundamental contributions to the foundations of science and mathematical thinking. We expect no less from future generations of Chinese statisticians.

At the dawn of the 20th century, statistics had but a small role on the stage of scientific practice; by the close of the century, the use of statistics was ubiquitous in virtually every area of scientific and practical interest, with statisticians having developed many procedures to address substantive problems arising in these various areas. In the last 25 years especially, proceeding in hand with the exponential growth in computing power, new technologies have emerged in various disciplines, leading to data sets that are orders of magnitude larger than those for which the earlier procedures were developed. This prompts questions which these earlier procedures are not well suited to address. Substantial statistical innovations are necessary to take on the task, and while statisticians are responding well to the challenge, much work remains to be done.

This volume, occasioned by the University of Science and Technology (USTC) 50th anniversary celebration, features 15 statistical papers by renowned statisticians and USTC alumni covering 4 topics in high dimensional data analysis and bioinformatics at the frontiers of modern statistical science: (i) statistical genetics, (ii) the analysis of microarray data, (iii) computational biology and statistical learning, and (iv) statistical methods for analyzing high dimensional data.

The three statistical genetics papers cover several important topics in the area. Cui, Zhang, Yang and Li address the problems of bias and power in linkage analysis with mixed affected sibling pair data. They propose three test procedures to address these issues, and show that all three perform satisfactorily. Chen reviews sequential importance sampling algorithms developed in population genetics, as well as a more recently proposed technique developed by Chen that incorporates resampling. Lin proposes a hierarchical Bayesian approach for detecting QTL using model selection techniques. The approach works well, provided the number of markers is not very large.

Four papers address the analysis of microarray data. By incorporating longitudinal information on gene expression, Hong develops a functional hi-

erarchical empirical Bayes approach for detecting TR and TDE genes from MTC gene expression experiments. Using a smoothness assumption on the gene expression trajectories, the gene expression profiles are modeled and approximated by well known basis function expansions. The paper by Lai addresses, in the context of FDR, the problem of estimating the fraction of null hypotheses that are true when a large number of tests are performed; using a nonparametric method, an upper bound on the fraction is obtained. Based on earlier work on normalization methods for microarray data, and information on non-replicating genes, Peng proposes new methods that lead to improved estimation of the intensity functions. In connection with the need for reliable variance estimation for gene expression microarray data, Tong and Wang review several statistical methods for estimating variances in the "large p, small n" context.

Four papers are addressed to the subject of computational biology and statistical learning. Feng, Xu, Zhang, Li, Xie and Wang study the problem of predicting protein subcellular locations using a machine learning type of approach. Through experiments and comparisons, the authors conclude that using the PSSM generated from PSI-BLAST as input and SVM as classifier leads to better predictive performance. A new learning method called LOCSVMPSI is proposed and recommended based on its even better performance. Chao and Jiang give a general review of methods for extracting information at the biomolecular sequence level, with emphasis on biological sequence alignment. Lin, Simmons, Beecher, Truoung and Young apply various classification methods, including RP, SVM, and RF, to identify an important set of metabolites for disease classification. Wang and Xi survey recent developments in the statistical modeling of chromatin Sequences. They argue that chromatin sequences trained by a previously proposed model called DHMM may have larger power in predicting the correct nucleosome positioning.

The final set of papers is concerned with high dimensional data analysis. Chen gives a comprehensive review of recent developments in finite mixture modeling. Guo and Dai propose an iterative procedure to fit a smoothing spline ANOVA model with heterogeneous variances. The method is then applied to a data set with a sample of epileptics. Zeng and Yu address the issue of bias arising in kernel estimation in longitudinal studies. They propose a bias-corrected procedure and derive its large sample properties. Zou derives a computable bound in evaluating the quality of the Gibbs sampler, in the context of estimating the posterior mode of the Lasso distribution.

The bound has direct implications for deriving the Lasso estimator. The papers in this collection illustrate both the types of challenges statisticians face and will continue to face and as well as the opportunities such challenges open up for new statistical work. On one hand, statisticians need to look

inward and develop better statistical methodologies and procedures, and on the other, to look outward and work with researchers in other disciplines to meet the new challenges they bring to the table. There is every reason to believe that statistics in the 21st century will continue to be as exciting an area as it was in the 20th century, and the need to train and develop a talented generation of younger statisticians will be great. We hope that USTC and its alumni will continue its work along these line and we trust that the next 50 years for USTC will be even more fruitful and exciting than the 50 past years that we celebrate today.

Shaw-Hwa Lo
Columbia University

Contents

Section I

Statistical Genetics

Linkage Analysis with Mixed Sample of Affected Full and Half Sib Pairs in the Presence of Uncertain Relationship

Wenquan Cui[1]*, Hong Zhang[1,2], Yaning Yang[1] and Zhaohai Li[3,4]

[1]Department of Statistics and Finance, University of Science and Technology of China, Hefei, Anhui, P.R. China.

[2]Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT, USA.

[3]Department of Statistics, George Washington University, Washington, DC, USA.

[4]Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD, USA.

Abstract

Affected sib pairs are widely used in human genetic linkage studies. Simple non-parametric tests, such as mean and proportion tests have been proposed for detecting linkage between the disease of interest and candidate markers. In practice, however, some of the relationship information of affected sib pairs (full sib pairs vs. half sib pairs) may be missing due to unavailability of parents, incompleteness of records, and some other reasons. Power loss is unavoidable if those sib pairs with uncertain relationship are excluded in analysis, while a small proportion of half sib pairs can lead to substantial increase of false positive rates if half sib pairs are misspecified as full sib pairs. Particularly, serious bias will be introduced into mean and proportion tests in the presence of uncertain relationship. In this article, we propose three test procedures, one is likelihood ratio test, the other two are modified mean and proportion tests, with biases and variances being carefully corrected. All the three tests are theoretically shown to be valid in the presence of sib pairs with uncertain relationship. Simulation studies are conducted to assess the performances of the proposed tests, and the results show that the proposed tests have correct type I error rates and satisfactory powers.

*Corresponding author

Furthermore, power loss is minor when the proportion of sib pairs with uncertain relationship is not too large, compared with the situation when true sibling relationships for all sib pairs are known for sure.

Key words: Affected Sib Sair; Linkage Analysis; Full Sib Pairs; Half Sib Pairs; Mixture Model; Identifiability.

1 Introduction

The affected sib-pair (ASP) design has been widely used in human genetic linkage studies (Suarez *et al.*, 1978; Blackwelder & Elston 1985; Holmans, 1993; Knapp *et al.*, 1994a, 1994b, 1998; Whittemore & Tu, 1998; Dudoit & Speed, 1999; Li & Gastwirth, 2003; Wang, 2004). The ASP linkage method (Penrose, 1935, 1953) was originally developed for detection of linkage between a genetic marker and a dichotomous trait locus by using full sib pairs (FSP). Based on the concept of identical-by-descent (IBD), nonparametric tests such as mean and proportion tests were proposed to detect the departure of the distribution of the ASP IBD sharing status from what is expected under the null hypothesis of no linkage. The ASP linkage method has been generalized to affected relative pair (Risch, 1990; Jung *et al.*, 2006) method for linkage studies. Considerable researches have focused on increasing the powers of these tests under different scenarios.

The ASP design ascertains affected full sib pairs. However, due to unavailability of parents or incomplete records for some of the sib pairs, half sib pairs (HSP) may enter the study (Göring & Ott, 1997). If a researcher does not recognize this issue and treat all the pairs as full sib pairs, then spurious conclusion could be reported since the classical mean, proportion or any other linear tests have inflated type I errors with the mixed samples of full and half sib pairs (Neale *et al.*, 2002). We call this type of data as mixed affected sib pair (MASP). Schaid *et al.* (2000) proposed a method for combining full sib and half sib pairs in a single test for quantitative trait when the true sibling information is known for all the pairs. For situations when the complete information of true sibling relationship is not available, Ehm & Wagner (1998) proposed a test based on identity by state (IBS) to detect errors in sib pair relationships, which computes the sum of difference of IBS with the expected IBS for each pair given true relationship of multiple marker; Göring & Ott (1997), Boehnke & Cox (1997) developed likelihood method to estimate the relationship by using information from multiple linked markers. However, these methods relies on linked marker information and not robust to the specification of the model assumptions. Also, estimation of relation-

ships by using linked marker information might cause genetic privacy concerns (Roche & Annas, 2006).

Currently available tests for ASP linkage studies, such as the mean and proportion tests, can not incorporate such mixed data in a single analysis. If the nonparametric IBD-based tests are applied to these data without proper corrections, the type I error rate is shown to be seriously inflated for moderately large sample size even if the proportion of half sib pairs is small. When the sibling relationship for every pair is unknown, the IBD distribution is a mixture of two discrete distributions, and the parameters involved are non-identifiable. However, if the true sibling relationship for a fraction of pairs is known, then the parameters are identifiable. In this situation, one can estimate the proportion of HSPs in the sample and adjust for the biases. The maximum likelihood estimate (MLE) of the parameters can be obtained via an E-M algorithm. In the common case of model parameters being weakly constrained, the likelihood ratio test (LRT) statistic has a limiting chi-square distribution under the null hypothesis of no linkage. We also propose to modify the classical mean and proportion tests by carefully correcting their biases and variances simultaneously. The two corrected tests are shown to be valid theoretically and confirmed by simulations. Simulation results also show that the power loss is tolerable up to 60% of pairs with uncertain relationship compared with full information data, showing the proposed tests are quite feasible in practice.

The rest of the article is organized as follows. In the next section, some issues related to the mixed samples are presented; then, the IBD sharing probabilities are shown to be non-identifiable and the classical mean and proportion tests have biases, especially when sample size is relatively large even if the proportion of half sib pairs is small; finally, the validity of three new tests in the presence of MASP are discussed. Simulation results are shown in Results Section. A brief summary and related issues are presented in Discussion Section. Appendices give algorithms for solving MLEs under both the null and alternative hypotheses, calculations of bias and variances for classical mean and proportion tests, and derivation of the modified mean and proportion tests.

2 Methods

2.1 Classical mean and proportion tests

For easy presentation, it is assumed that the marker is fully informative so that the IBD sharing status at the marker locus is observed without ambiguity. For the situation when the IBD is not available, a solution is given in Discussion Section.