 **A Biostatistics Applications
Area in the Department of
Mathematics for a
PhD/MSPH Degree**

Patricia B. Cerrito
Department of Mathematics
Jewish Hospital Center for Advanced Medicine
pcerrito@louisville.edu

Biostatistics

The discipline of biostatistics focuses on the examination of healthcare data. Much of the data are collected through the process of randomized, clinical trials. All new medications are required to demonstrate safety and effectiveness through randomized, clinical trials before they will be approved for use by the Food and Drug Administration (FDA). For this reason, pharmaceutical companies employ many biostatisticians (and SAS programmers).

It is the responsibility of the biostatistician to design the clinical trial. The biostatistician must decide before any subjects are recruited just how many subjects are needed for the study. The statistical method must be clearly specified that is to be used to demonstrate both safety and effectiveness in the patients.

It is not always possible to design randomized studies. For example, it is not possible to test the effects of smoking by randomly assigning subjects to a “smoking” group and to a “no smoking group”. Observational studies must be used with the acknowledgement that they can be incomplete and incorrect if confounding factors are not taken into consideration. One of the most lucrative new areas of biostatistics research is to investigate health outcomes in clinical practice. This can only be done observationally as patients enter into treatment to determine “best practices”.

In order to accommodate this observational data, the field of medical informatics has been developed. Its purpose is to work with health information as it exists. The field is strongly advocating the development of electronic patient records. Since physicians are still using chart notes, it is also necessary to be able to analyze unstructured text data.

The field of data mining focuses on working with large datasets with hundreds and sometimes thousands of variables. The field has also developed tools to analyze text data. It is strongly recommended that students interested in biostatistics also learn something about data mining.

Biostatistics Courses

- **Math 665 Advanced Linear Statistical Models.** Distribution of quadratic forms, estimation and hypothesis testing in the general linear model, special linear models, applications.
- **Math 667 Methods of Classification.** Classification methods used in the industry to handle large databases. Logistic regression, structural equation modeling, multivariate analysis, data mining.
- **PHDA 602 Biostatistics-Decision Science Seminar.** Students are given an evaluation protocol for each semester and must turn in a written evaluation of the presentation. The protocols will vary according to the presentation topic, but each will focus on a critical component of research design or analysis.
- **PHDA 603 Biostatistics-Decision Science Public Health Practicum I.** A student is assigned to a health care agency and works with the staff of that agency on a policy issue facing that agency.
- **PHDA 605 Ethics and Bioethical Decision Making.** A study of ethical issues in contemporary bioethics. Ethical dilemmas in medical science will be analyzed for the philosophical assumptions, interplay of facts and values, the role of rules and principles, and the contextual factors involved. Such topics as abortion, elective death, genetic engineering, organ transplants, and health care reform will be explored.
- **PHDA 663 Decision Analysis.** This course teaches methods for making decisions in complex situations especially those involving conflicting values, uncertainty, or risk. Thinking from the early foundations in economics through current methods is covered. Included are methods of value or utility elicitation and probability assessment. Analysis methods covered include decision trees, conjoint measurement, and multiattribute utility theory. Also covered are findings from psychology on cognitive errors, which are common in decision making.
- **PHDA 680 Biostatistical Methods I.** A mathematically sophisticated presentation of principles and methods of: exploratory data analysis; statistical graphics; point estimation; interval estimation; hypothesis testing of means, proportions and counts; chi-square analysis; rate ratio; and Mantel-Haensel analysis. Matrix algebra is required. Data sets will be analyzed using statistical computer packages; examples will be drawn from the biomedical and public health literature. Emphasis will be placed on methods and models most useful in clinical research.
- **PHDA 681 Biostatistical Methods II.** A mathematically sophisticated introduction to: general linear models; regression; correlation; analysis of covariance; one and two-way analysis of variance; and multiple comparisons. Matrix algebra is required. Data sets will be analyzed using statistical computer packages; examples will be drawn from the biomedical and public health literature. Emphasis will be placed on methods and models most useful in clinical research.
- **PHCI 611 Introduction Clinical Epidemiology.** Comprehensive introduction to public health with emphasis on population-based approaches to health issues. Classical and clinical epidemiology will be presented. Covers health status indicators, including morbidity, mortality, vital statistics and measures of quality of life. Global applications of epidemiology and international health through investigations of the

leading causes of morbidity and mortality in developed, developing, and underdeveloped nations. Epidemiological concepts will be linked with computer exercises to reinforce learning and practical applications.

- **PHCI 631 Social and Behavioral Sciences in Health Care.** This course introduces public health students to social science perspectives and research on selected topics in health and health care. The course is organized into the following units: the sociology of knowledge and health behavior modeling; the social distribution of health, disease and utilization by social variables; social problems (e.g., violence and substance abuse) as public health concerns; health care industry and policy health behavior and the psychology of illness; international health and health care systems; and genetics and public health.
- **PHCI 651 Introduction to Environmental Health.** Lay a foundation for students to build upon their medical and scientific background in applying clinical skills in the resolution of real, in-the-field, community-based problems. Utilize problem-based training exercises involving both classroom (laboratory) and in-field epidemiological studies to active community-based environmental issues.
- **PHCI 662 Health Care Economics.** Provides a comprehensive groundwork in the economics of health care and the health care sector. The trainee will be able to effectively analyze issues in the health sector from an economic perspective and determine primary and secondary effects of changes in the health care market. In particular, this course examines the insurance market, the economics of hospital operations, physicians and the incentives of various payment schemes. Attention is given to the basic theory and techniques of cost-benefit, cost-effectiveness, and cost-utility analysis as well as methods for valuing outcomes including the use of quality-adjusted life years.

Use of Statistical Software

Spreadsheets have some statistical tools but are extremely limited and should not be used as statistical packages. Small statistical packages can be purchased for use on 1 desktop at cost < \$500. Their use is limited and can only perform relatively simple, routine statistical methods.

The two main statistical packages used in research are SPSS and SAS. SPSS was developed for use with the social sciences. It still sells primarily to an academic market. SAS now incorporates all the social science methodology of SPSS + database management. It was developed for use in the natural sciences. It is now the primary package used in the business market. **The ability to use SAS has now become a very marketable skill.** Businesses advertise for SAS experience. For example, FDA statistical guidelines were in part derived with SAS language. Pharmaceutical companies all use SAS as the statistical software package. Other statistical packages (such as S-plus) might be used as add-ons to SAS, but SAS remains the primary package.

One of the reasons that SAS remains so popular is that each year in early April, the SAS User's Group International holds an annual conference attracting approximately 4000 attendees. SAS developers interact with SAS users (mostly from the business environment) to improve the product. **Most of the innovations in using statistical methods are now coming from the business world rather than the academic world.**

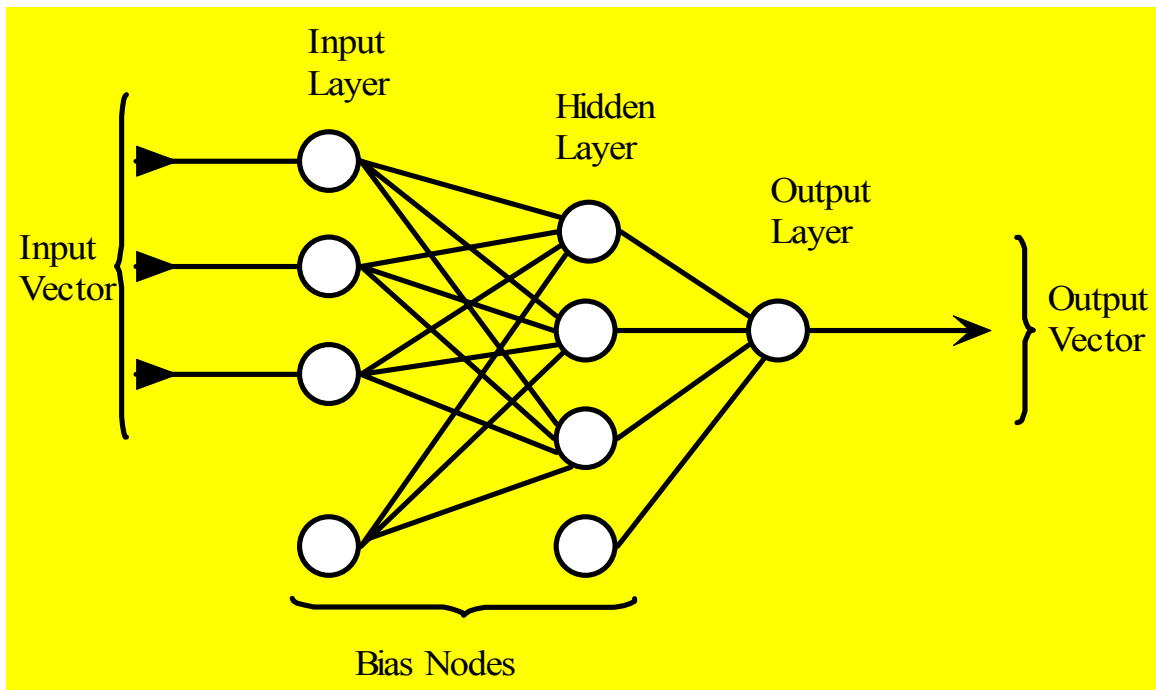
Data Mining Software

There are a number of data mining software packages, including Intelligent Miner by IBM. However, for good data mining software combined with good statistical software, there are two: Clementine by SPSS and Enterprise Miner by SAS. Enterprise Miner as the best integration of statistics with data mining. It contains methods that SPSS does not include: link analysis, text mining, memory-based reasoning. It is the data mining used in the business world.

Data Mining Techniques

- Artificial neural networks
- Rule induction
- Logistic regression
- Association rules
- Data visualization

Artificial Neural Networks are used to classify the observations into categories. The final result is a “black box” in that there is no statistical function that defines the classification rule.



Previous Student Research Projects in Biostatistics

Project	Company	Outcome
Examination and definition of compliance in treating diabetic patients	Norton Healthcare	Student currently employed in statistical analysis for a medical firm
Database development and analysis of data for leukemia clinical trial	Institute for Cellular Therapeutics	Internship led to full time employment in healthcare
Comparison of information provided to physicians via Medline and information provided to consumers via the web	University of Louisville	Currently in process
Analysis of quality-of-life data in cancer clinical trials	Schering-Plough	Completed. Student working full time as a Lan manager.
Analysis of clinical trial data to examine efficacy of pacemaker placement.	Guidant Technologies	Current in process.
Analysis of health insurance purchasing and use in physician practice	Department of Mathematics	Thesis completed. Student working full time at Army Corps of Engineers in web design and statistical analysis.
Analysis of Blue Cross/Blue Shield customer satisfaction survey	Datamax, Inc.	Student enrolled in PhD program in applied statistics
Analysis of data on burn victims to classify as high/low risk for infection	Family Practice	Student enrolled in PhD program in applied statistics
Analysis of method used to produce 3 dimensional prototypes	Departments of Mathematics/Industrial Engineering	Student employed full time at Price Waterhouse using data mining techniques**
Comparison of Hospital Outcomes Using Medpar Data	Jewish Hospital, Louisville	Degree completed.
Investigation of environmental hazards related to increased risk of asthma	Jewish Hospital, Louisville	Degree completed. Student enrolled in PhD program.
Screening of College Athletes for Symptoms of Asthma	Frazier Rehab Center, Louisville	Currently in process. Student displayed poster presentation at State Capital
Analysis of Survival for Melanoma Patients	Norton Hospital, Louisville	Current in Process. Student currently enrolled in Mathematics PhD Program with biostatistics application area**
Analysis of Patient Visits to Public Clinics in Mercer County, Kentucky	Kentucky Department of Health	Project completed. A second project is in process to link visits to services. Student is completing capstone project in MPH program.

PhD/MSPH required courses:

Courses from Mathematics	Courses from Biostatistics
<ul style="list-style-type: none">• Two from: Algebra I and II, Combinatorics I and II, Real Analysis I and II• Two from: Mathematical Modeling I and II, Applied Statistics I and II, Probability and Mathematical Statistics <p>Industrial Internship Dissertation and Qualifying Exams 24 Course Credits Internship, Practicum, and Masters Thesis give a combined 8 hours of course credit.</p>	<ul style="list-style-type: none">• Introduction Clinical Epidemiology• Social and Behavioral Sciences in Health Care• Introduction to Environmental Health• Health Care Economics• Biostatistics/Decision Science Seminar • Biostatistical Methods I and II or• Ethical Issues in Decision Making and Decision Analysis• Biostatistics Practicum• Masters Thesis <p>22 Course Credits</p>

Summary

- Statistics is a marketable and profitable field with a large number of possible fields of specialization.
- Biostatistics is particularly marketable.
- Biostatistics can be enhanced with data mining tools.