BIOSTATISTICAL METHODS AND RESEARCH DESIGNS

Xihong Lin
Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

Keywords: Case-control study, Cohort study, Cross-Sectional Study, Generalized linear model, Longitudinal data, Study design, Statistical model, Statistical inference, Survival analysis.

Contents

Summary

This paper provides an overview of commonly used biostatistical research designs and methods. It helps the reader understand the biostatistical developments in the subject-matter areas that will be discussed in detail in the collection of chapters under the present topic. General biostatistical research strategies will be discussed. Several study designs will be reviewed. Commonly used statistical models and methods will be provided and illustrated using examples. Major statistical inference procedures will be described.

1. Introduction

The field of biostatistics has developed rapidly in the past thirty years. The scientific advances and their practical data analysis needs of a broad range of subject matter disciplines in health sciences play a central role in this rapid development. They motivate investigation and development of novel study designs and statistical methodology. Compared to the traditional statistical discipline, a unique feature of biostatistical research is its close connection with real world applications and biostatisticians work closely with scientific investigators. Such a nature is important for future developments of the field of biostatistics.

For example, survival analysis for censored time-to-event data is now a well-established field in biostatistics and is widely used in many fields of health sciences, especially in clinical trials. This field has experienced a tremendous growth in the last fifty years. The fundamental contributions include the Kaplan-Meier estimator of the survival probability, the Mantel log-rank test and the Cox proportional hazard model. Although censored survival data were sometimes collected in the mid 20th century, their use in medical research was rather limited due to the lack of attractive regression models and statistical software. The practical need to develop such a regression model was in high demand in late 1960's and early 1970's. The seminal paper by D. R. Cox in 1972 on the proportional hazard model and the partial likelihood method was a landmark of the field. The convenience of the estimation procedure of the proportional hazard model and its attractive parameter interpretation in terms of the relative risk widely opened the door of the field for its rapid developments and practical applications. The Cox paper is one of the most cited statistical papers in health sciences.

The five chapters under this topic describe in detail statistical designs and methods used in several areas of health sciences, including epidemiology, communicable diseases, nutritional epidemiology, laboratory and basic science research, and toxicology. These articles provide the reader a survey of biostatistical issues, designs, and methods that have been developed in these subject matter disciplines and the future research directions. These articles allow the reader not only to learn the statistical developments in these subject-matter areas but also to appreciate the fact that biostatistical research is well motivated, tailored and integrated with a scientific discipline. The detailed statistical designs and methods vary from one discipline to another. However there are many common biostatistical research principles and designs and methods that are shared by different disciplines.

This paper provides an overview of biostatistical research designs and methods. This overview aims at providing the reader with a background of biostatistical research principles and commonly used statistical designs and methods, and facilitating the reader with a better integration of different statistical methods across the five subject-matter disciplines. The structure of the paper is as follows. Section 2 describes general biostatistical research strategies. Section 3 discusses common study designs. Section 4 describes commonly used statistical methods. Section 5 discusses statistical inference procedures.

2. Biostatistical Research Strategies

2.1. Understanding Scientific Disciplines

Unlike the traditional field of statistics, advances in the field of biostatistics require biostatisticians to be engaged substantially in real world applications and collaborate closely with subject-matter researchers. Specifically, to make a significant biostatistical contribution in any scientific discipline, one needs to have a solid understanding of sciences in the particular discipline, including scientific background, scientific questions and quantities of primary interest, scientific terminology, data collection procedures, and practical needs of subject-matter researchers. This will allow one to develop an appropriate study design, formulate meaningful statistical models, generate ready interpretable statistical results, and effectively communicate with subject-matter investigators. It would also allow a biostatistician to identify new statistical problems and develop novel statistical methods.

It is common in a scientific discipline to first understand the population parameters of primary interest and develop scientific hypotheses of interest. For example, in epidemiology, which is a discipline investigating factors that cause a disease distribution, one is often interested in studying the relationship between a disease outcome, e.g., breast cancer, and an exposure variable, e.g, dietary intake. Such an association is commonly measured using an odds ratio or a relative risk. Denote by E the exposure status (1=Exposed and 0=Unexposed) and by D the disease status (1=Disease and 0=Non-disease). The odds ratio (*OR*) for disease and exposure is defined as the ratio of the odds of disease among exposed subjects and that among unexposed subjects

$$OR = \frac{Pr(D=1 \mid E=1)/Pr(D=0 \mid E=1)}{Pr(D=1 \mid E=0)/Pr(D=0 \mid E=0)}.$$

Denote by $\lambda(t)$ the incidence rate (hazard) of disease at a particular time or age t. The relative risk (*RR*) for disease and exposure is defined as the ratio of the hazard of disease among exposed subjects and that among unexposed subjects

$$RR = \frac{\lambda(t \mid E=1)}{\lambda(t \mid E=0)}.$$

No association between disease and exposure corresponds to $OR=1$ and $RR=1$. For rare diseases, the odds ratio is an approximation of the relative risk. Similarly, in communicable diseases research, which investigates the effects of the presence of the infectious agents in the host population, two key measures in this field are the transmission probability and the basic reproductive number.

2.2. Study Design and Data Collection

Scientific investigators are often interested in addressing particular scientific questions. These scientific questions can often be formulated by either estimating some population parameters of interest or by testing particular hypotheses that are constructed using some population parameters. For example, in a study of the association of dietary intake and breast cancer, one is interested in testing the null hypothesis that there is no association between dietary intake and breast cancer, i.e., the relative risk of dietary intake and the risk of breast cancer is one ($OR=1$ equivalently $RR=1$).

After establishing the goals of a study, appropriate study designs are developed to address the scientific hypotheses of interest. A survey of several commonly used study designs is provided in Section 3. Studies in health sciences research can be broadly classified into observational studies and randomized studies, the latter involves randomization, e.g., clinical trials are randomized studies. Study designs can be further classified into cross-sectional studies, cohort/follow-up studies, and case-control studies, and laboratory studies. The choice of a particular study design should be based on the aims of the underlying scientific investigation and practical feasibility. For example, for a rare disease, it is often not feasible to conduct a follow-up study and a case-control study is more feasible. It is important to understand the limitations of a chosen study design. For example, selection bias is a major problem in a case-control study.

The next stage is data collection. This first involves sample size calculations to ensure sufficient power is available for the scientific hypotheses. Such sample size calculations often require results obtained from either a pilot study or from the published literature. One also needs to carefully consider choices of outcome variables and covariates, sampling strategies and potential data issues such as ways to reduce missing data.

2.3. Statistical Data Analysis

Statistical data analysis often starts with exploratory analysis. This involves calculating descriptive statistics of outcomes and covariates, such as frequencies, means, and standard deviations. Correlation coefficients can be examined. Graphical displays such as scatter plots and histograms are often constructed to examine the distributions of variables and the crude relationships between variables. Such exploratory data analysis provides an opportunity for investigators to get familiar with the data and get a rough picture of the relationships of interest, and identify potential issues in the data such as data entry errors and missing data.

Formal statistical analysis requires developing statistical models that are suitable for the study design and the data structure under investigation. In Section 4, several commonly used statistical models are reviewed. They include linear and generalized linear models, models for survival data, and models for longitudinal data. A particular scientific discipline has its own nature and special issues. These common statistical models need to be tailored towards the particular needs of the discipline. Further, these standard models might not be sufficient for the needs of a particular discipline. It is common that specific statistical models are formulated and developed based on the scientific nature of a discipline. For example, in communicable diseases, one is interested in studying disease transmission instead of the association between an outcome and covariates. Statistical models in this field are rather specialized. In nutritional epidemiology, an important issue is that measures of nutrition intake are often subject to considerable measurement errors. Statistical methods for handling measurement errors are particularly useful.

2.4. Dissemination of Results

Dissemination of newly developed statistical methods is an important step in biostatistical research. This involves publishing the methodology and illustrating their applications in statistical journals and subject matter scientific journals, giving presentations in conferences and workshops, and developing user-friendly software. Software developments are especially critical in disseminating advanced statistical methods. For example, after the Cox proportional hazard model was implemented in commercial software, such as SAS, it became a standard practice for analyzing censored survival data and is widely used in many disciplines of health sciences.

3. Study Designs

Health sciences studies can often be divided into three broad design strategies: observational studies, randomized studies and other types of studies. The key difference between an observational study and a randomized study is whether sampling units are subject to randomization. Other types of studies include a hybrid of observational and randomized studies, and laboratory experiments.

3.1. Observational Studies

An observational study refers to a study in which investigators collect data by observing the natural course of an event. Many epidemiological studies, chronic disease studies and communicable disease studies fall in this category. Observational studies can be further categorized into cross-sectional studies, case-control studies and cohort studies.

-
-
-

Bibliography

Breslow, N. E. and Day, N. E. (1980). Statistical Methods in Cancer Research, Vol. 1: The Analysis of Case Control Studies. Lyon: International Agency For Research on Cancer. [This is a classical reference on case-control data.]

Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*. London: Chapman and Hall. [This is a classical reference for statistical methods handling measurement errors.]

Casella, G., and Berger, R. L. (2002). *Statistical Inference (2nd Edition)*. Pacific Grove: Duxbury Press. [This is an introductory textbook on statistical inference.]

Cox, D. R. (1972) Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, *Series B*, 34, 187-220. [This is the original paper that proposes the proportional hazard model.]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological*, 39 , 1-22. [This is an original article on the EM algorithm.]

Diggle, P. J., Liang, K. Y., Heagerty, P. and Zeger, S. L. (2002). *Analysis of Longitudinal Data (2nd Edition).* Oxford: Oxford University Press. [This textbook provides an overview of statistical methods for analyzing longitudinal data.]

Hosmer, D. W. and Lemeshow, S. (1989) *Applied Logistic Regression (2nd Edition).* New York: John Wiley & Sons. [This is an introductory textbook on logistic regression]

Hosmer D. W. and Lemeshow, S. (1999). *Applied Survival Analysis (2nd Edition)* New York: John Wiley & Sons. [This is an introductory textbook on survival analysis.]

Kalbfleisch. J. D. and Prentice, R. L. (2003) *The Statistical Analysis of Failure Time Data.* [This is an advanced and updated textbook on survival analysis.]

Little, R. J. A. and Rubin, D. B. (2003) *Statistical Analysis with Missing Data (2nd Edition).* New York: John Wiley & Sons. [This is a classical and updated reference on missing data.]

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (2nd Edition).* London: Chapman and Hall. [This textbook provides an overview of generalized linear models.]

Prentice, R. L., and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403-412. [This articles shows that case-control data can be analyzed prospectively using logistic regression.]

Weisberg, S. (1985). *Applied Linear Regression*. New York: Wiley. [This textbook is a classical reference for linear regression models.]

Biographical Sketch

Xihong Lin was born in Hei Long Jiang Province, People's Republic of China. She obtained her B.S. in Applied Mathematics from Tsinghua University, Beijing, China, in 1989. She obtained her M.S. in Statistics from the University of Iowa in 1991 and her Ph.D. in Biostatistics from the University of Washington in 1994. Dr. Lin was Assistant Professor (1994-1999), Associate Professor (1999-2002), and Professor (2002-present) of the Department of Biostatistics at the University of Michigan. She is Member of the Cancer Center and a Faculty Associate of the Institute of Social Research of the University of Michigan. Her main research interests are analysis of correlated data, such as longitudinal data, clustered data, spatial data, and multivariate survival data, nonparametric and semiparametric regression, measurement error and missing data. She has collaborated with investigators in epidemiology, environmental health, community-based research, cancer and neurology. She has served on Clinical Trial Data and Safety Monitoring Boards and review panels of the National Institute of Health and the National Science Foundation, USA. She has served on several statistical journal editorial boards and committees of national and international statistical associations. Dr Lin is currently the Coordinating Editor of *Biometrics*, the Journal of the International Biometric Society.