DATA COLLECTION AND ANALYSIS IN BIOMETRICS

L. Billard
Department of Statistics, University of Georgia, Athens, GA USA

Keywords: Experimental design, split plots, repeated measures, survey sampling, clinical trials, case-control studies, longitudinal studies, time series, species abundance, data collection.

Contents

Summary

This chapter attempts to provide a brief overview of the statistical methodologies covered under Data Collection and Analysis in Biometrics while more details can be found in the related chapters. In particular, the articles consider statistical analytical methodologies and design issues for experimental design, sample surveys, clinical trials, time series and species abundance. This chapter also includes a brief introduction to related methodologies such as longitudinal data, case-control and cohort studies and their relationship and distinction from other seemingly related areas.

1. Introduction

Statistics is an integral thread in the fabric that constitutes science – science in all its manifestations, be this in its new research pursuits or its application in old or new investigations. As a science, statistics grew out of the need to develop methodologies to collect and to analyze data, to interpret these results and to draw conclusions as they pertained to the scientific undertaking that generated the original inquiries.

Historically, statistical science can be said to have gained its essential momentum in the 1800s. The American Statistical Association formed in 1839 out of concerns over the manner in which the decennial censuses were being conducted. Members were

scientists (in today's terminology, social scientists, biological scientists, engineers, medical practitioners, and so on, of all stripes and persuasions) who sought relevant data as aids in their own endeavors, with much of the early data coming from census efforts. Indeed, in what became the modern-day *Journal of the American Statistical Association* the first paper in the first volume is a study of water power, an important natural resource still today. The second paper investigated the importance of open park land in the health of the residents of a city. In Europe in 1838, the first volume of what became the *Journal of the Royal Statistical Society* contains an article on the welfare of children and one on the working conditions of employees. By the end of the 19th century, especially in the United Kingdom, the early work tended to revolve around biological issues, with particular emphasis on genetics. However, the common underlying tenets then, as now, were basic issues that dealt with the human condition and how statistics could and should be utilized to alleviate and advance the condition of society.

The published literature during the 19th century and the early part of the 20th century by and large focused on the substantive science *per se*, using and/or developing statistical thought as required to elicit new knowledge in that substantive field. Statistical ideas were usually presented as verbal arguments supported by arithmetic illustrations. Papers could be, and indeed often were seemingly, repositories of data with tables (sometimes pages and pages of tables) provided to elucidate the advances described. Of course, arithmetic is "mathematics" in some sense; but it was not until the 1910s that mathematics *per se* emerged as a primary tool. Its entry tended to revolve around the introduction of the concept of correlation – mostly in the context of economics and the social sciences in the US, and in biology in the UK with Fisher's derivation of the distribution of the sample correlation function playing a central role.

Correlation concepts quickly led to the notion of regression models, and the modern-day statistical world was clearly launched. Regression models, in and of themselves, is a vast subject covering an enormous range of types of models for a huge range of generalized and specialized applications, and is treated in many articles under different topics in this Theme. Two special directions relate to linear models and time series.

First, consider the multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_t X_t + e, \qquad (1)$$

where $\boldsymbol{X} = (X_1, \ldots, X_t)$ are t explanatory (and independent) variables, Y is the response variable (dependent on the values of X_i, $i = 1, \ldots, t$,), and e is the error term (with mean zero and variance σ^2) associated with that observational response value. In the completely general case, there are no special restrictions on the values that the X_i can take. Next, take a one-factor fixed effect factorial design (or completely randomized design) with t possible treatments. The model is usually written in the form of

$$Y = \mu + \tau_i + e, \qquad (2)$$

where μ is the overall mean, τ_i is the effect of the ith treatment for $i = 1, \ldots, t$, and where Y and e have the same definition as in Eq. (1).

Now if the explanatory variables X_i in (1) take specific values $X_i = 1$ if treatment i applies and $X_i = 0$ if not, for each $i = 1, \ldots, t$, then Eq. (2) is equivalent to Eq. (1) where now $\mu \equiv \beta_0$ and $\tau_i \equiv \beta_i, \quad i = 1, \ldots, t$. That is, the Eq. (2) for the factorial design model is a special case of the more general multiple regression model of (1). Corresponding equivalences can be made for other experimental design models, these falling today under the rubric of generalized linear models, a class of models which itself is a generalization of those in (1). Back in the early 1920s, these connections had not been formalized. However, at that time important radically new developments in experimental design were exploding on the statistical stage, primarily through agricultural experiments in the United Kingdom (at Rothamsted Experimental Station) and by the 1930s taken up equally vigorously by researchers in the US (at the Statistical Laboratory at Iowa State College, now University). Today design and its application go well beyond the specific field of agriculture in which it originated. Indeed, experimental design is itself a powerful tool in the toolkit for any life scientist. A further introduction to design is given in Section 2 below.

Correlation concepts also spawned developments in another direction. Consider the economist who sought to use a regression model (such as Eq. (1); let us set $t = 1$ for discussion purposes) to explain economic data where, e.g., the response variable Y may be a commodity price (wheat, say) and the explanatory variable X was time. Behind the use of (1) was an assumption that the different realizations of Y were independent of each other. For some economic data recording prices over time, this independence assumption was hard to sustain when it could be argued that "today's" price was in fact correlated in some way to "yesterday's" price. It was but one step to realize the explanatory variable was not X (time alone) but rather it was the value of the response variable at the previous time. That is, in its simplest form, Eq. (1) is replaced by

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + e_t, \tag{3}$$

and is called an autoregressive model. This mathematical formulation appeared in the early 1920s. Thus, began the area now called time series models; see Section 5.

While both experimental design and time series grew out of notions behind correlation and regression, they are fundamentally quite different statistical methodologies. The data (the Y's of Eq. (1) or (2)) from an experimental design might be tabulated in an organized and systematic manner, but the actual order in which the observations are measured is not important, since they are mutually independent. In contrast, time series model dependent data and so the order in which the observations occur is a crucial component; that is, the order is itself important and must be recorded along with the value of the observation itself.

Clinical trials (see Section 4) follow subjects (some with an intervention treatment, and some as controls) over time. This use of time however is statistically incidental and is not at all comparable to its use in time series. The goal of clinical trials is to study the effectiveness of these interventions. In some trials, the time it takes for the intervention (drug, say) to take effect may be the subject of interest, in which case it becomes the response variable. There can however be a regression component to the underlying analysis to model the impact of covariates or risk factors (equivalent terms for the

explanatory variables $\boldsymbol{X}$) on that intervention. More specific models such as logistic models, log-linear models, Poisson models, etc., are employed rather than the standard linear regression model. Similar comments apply to case-control studies and to longitudinal studies. The latter however often do follow subjects over time where time has the time series meaning. However, these studies typically involve considerably fewer time points than those required to obtain valid results from employing time series analytic methods; so alternative approaches have to be used.

Finally, the importance of designing the experiment properly can never be stressed too much. Data collected and assembled in a table can have arisen from a myriad of different ways. Yet, how the data arose, how the underlying experiment was run, determines how they are to be analyzed with different methods required and different conclusions drawn. Prior to all this, the experimenter must identify concisely what questions are to be answered from the study to be undertaken.

Regardless of the specifics of the questions or of the study designs, all experiments involve randomization and sampling to varying degrees for the results to be valid. Take a study on the electricity usage of city residents for which it has been determined that a sample size of 15 (say) is needed. If an interviewer goes door to door asking each householder in the one street what their usage is, the results may give a reasonable facsimile of electricity consumption for that residential area, but not at all for the city as a whole (unless a census of every household is undertaken – that is a different subject!). Sampling human populations brings its own difficulties especially when relying on the subject's ability to recall information.

Clinical trials, and especially case-control studies which overwhelmingly tend to use interview and questionnaire-based responses, are particularly vulnerable to these problems. In what was a startling surprise at the time, census investigators in 1943 discovered that, when returning to respondents after a first interview 8-10 days earlier, over 10% were now more than a year older (17% if a different person conducted the second interview). This inconsistency still exists as a persistent problem for human studies. In contrast, experiments based on non-human subjects do not have to deal with this faulty recall phenomenon with an attendant reduction in underlying sources of variation. The sampling process and its importance in obtaining meaningfully random samples of observations is covered in Section 3, with the specialized sampling methods required for species treated in Section 6.

Section 7 discusses some of the issues concerning data collection and Section 8 provides some concluding remarks.

-
-
-

Bibliography

The entries here are intended to serve as supplements to those already in the comprehensive bibliographic materials of the chapters under this topic.

Armitage, P. (1980). The Analysis of Data from Clinical Trials. *The Statistician* 28, 171-183. [Reviews methods for the analysis of clinical trials data, including a nice summary of data collection and protocol issues.]

Armitage, P. and David, H. A. (eds.) (1996). *Advances in Biometry*, John Wiley, New York. [This Volume celebrates 50 years of the International Biometric Society, with 21 articles from leading practitioners writing an historical overview and providing a current perspective on these 21 areas of biometry. The article on experimental design (by Bradley and Anderson) and agricultural and forestry (Freeman and Rily), case-control studies (by Breslow), longitudinal studies (by Ware and Liang), and clinical trials (by Pocock) are excellent overviews for topics covered under this chapter . There are also numerous articles, which provide excellent coverage of topics covered elsewhere under this Theme on Biometrics]

Billard, L. (1999). Sampling and the Census 2000. *Stats* 25, 6-11. [Though written against the backdrop of the sampling controversies surrounding the US' 2000 Census, this is a non-technical introduction to the lay reader about why samples are taken and what constitutes sampling as a science.]

Breslow, N. E. (1996). Statistics in Epidemiology: The Case-Control Study. *Journal of the American Statistical Association* 91, 14-28. [A review of case-control studies starting from Cornfield's odd's ratio, through the Mantel-Haenszel procedure, likelihood methods for relative risk regression to recent work on nested designs. There is an excellent discussion of the limitations and challenges of case-control studies. Breslow's paper formed the basis of the entry in the current article.]

Diggle, P. J., Liang, K. –Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data*, Clarendon Press, Oxford. [An introduction to the subject, both theoretical and practical, with numerous illustrative examples.]

Friedman, L. M., Furberg, C. D. and DeMets, D. L. (1998). *Fundamentals of Clinical Trials* (3rd ed.), Springer-Verlag, New York. [An easy to read conversational introduction to clinical trials covering the principles involved but without the mathematical theory.]

Jones, B. and Kenward, M. G. (1989). *Design and Analysis of Crossover Trials*, Chapman and Hall, London. [This is a hands-on text covering the basics of crossover designs and is particularly useful for the practitioner.]

Mantel, N. and Haenszel, W. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute* 22, 719-748. [Introduces the Mantel-Haenszel procedure, still used today. This classic paper is one of the most cited papers in the scientific literature.]

Piantadosi, S. (1997). *Clinical Trials: A Methodologic Perspective*, John Wiley, New York. [A comprehensive introduction to clinical trials replete with examples and extensive bibliography.]

Rosenberger, W. F. and Lachin, J. M. (2002). *Randomization in Clinical Trials Theory and Practice*, John Wiley, New York. [A detailed study of the many types, and their effects, of randomization processes used in clinical trials.]

Biographical Sketch

Lynne Billard was born in Toowoomba, Australia; she obtained her BS First Class Honors degree (1966) and PhD degree (1969) in Statistics from the University of New South Wales Australia. She has spent 13 years in administration including 9 years as Head of Statistics and 2 years as Associate to the Dean. She held academic positions in Australia, Britain, Canada and the USA. Currently, she is University Professor and Professor of Statistics at the University of Georgia where she has taught design of experiments to graduate students since 1980; she has also taught a wide range of courses (including time series, introductory statistics, and theoretical statistics) to both undergraduate and graduate students. She has over 150 publications mostly in the major journals including six books edited or co-edited, in

sequential analysis, AIDS and epidemics, time series, and inference, with applications in agriculture, biology, epidemiology, education and social sciences. She has been accorded many honors and awards including the 1990 American Statistical Association's (ASA) Award for Outstanding Statistical Application paper (shared), the ASA 1999 Wilks Award and the ASA's 2003 Founders Award, and the University of Georgia Creative Research Award. She has held numerous professional offices including International President 1994 and 1995 of the International Biometric Society and was the 1996 President of the American Statistical Association. She has served on the International Council of both the International Biometric Society and of the International Statistical Institute, and was the 1985 President of the Eastern North American Region of the International Biometric Society, and served on the executive committee as Program Secretary of the Institute of Mathematical Statistics. She has served on many US national committees including the Advisory Committee for DMS National Science Foundation, Panel on AIDS and Panel on Microsimulation Modeling of Social Welfare Policy both for the National Research Council, the National Academy of Sciences' Board of Mathematical Sciences, was Chair of the Conference Board of Mathematical Sciences, and numerous review panels for the National Institute of Health and the National Science Foundation as well as the UK Research Council. She currently serves on the US Secretary of Commerce Census Advisory Committee. She is a Fellow of the American Statistical Association, the American Association of the Advancement of Science, and the Institute of Mathematical Statistics and an elected Member of the International Statistical Institute.